Mechanisms of NKG2D ligand regulation

Michael McCarthy

Balliol College

Nuffield Department of Medicine

University of Oxford

A thesis submitted for the degree of

DPhil in Clinical Medicine

Trinity term, June 2013

Candidate: Michael McCarthy
Title: Mechanisms of NKG2D ligand regulation
College: Balliol College, University of Oxford
Degree: DPhil in Clinical Medicine
Submitted: Trinity term, June 2013

Abstract

Background: The NKG2D ligands are a set of cell surface proteins, the expression of which can make cells susceptible to immunity mediated by NKG2D receptor expressing cells, which include NK cells, $CD8^+$ $\alpha\beta$ T cells and $\gamma\delta$ T cells. The NKG2D ligands are known to be expressed in distinct settings, including viral infection, cancer, T cell activation, and cellular proliferation, settings also tightly associated with Warburg metabolism. The molecular events which determine NKG2D ligand expression status are unknown.

Aims: We aim to enhance understanding of the deterministic molecular events that control NKG2D ligand expression. Specifically, to explore the relationship between Warburg metabolism and NKG2D ligand expression in cell line and physiological models, and second, to identify open chromatin elements at NKG2D ligand loci, and develop computational methods to analyse this data.

Methods: We use a range of molecular biology techniques to delineate the role of glucose metabolism in NKG2D ligand expression in a HEK293T cell model. We develop a physiological CMV-primary fibroblast model of NKG2D ligand induction to validate our key findings. We adapt, optimise and validate a DNaseI-seq protocol, to define open chromatin sites at the NKG2D ligand loci. We develop a data analysis 'pipeline', including our own peak-finding software ("PeakHunter"), to identify open chromatin sites in the data.

Key results: Glucose drives NKG2D ligand expression. This effect requires cellular uptake and metabolism of glucose. Purine nucleotides are a key glucose metabolite for this effect, and purine nucleosides are *sufficient* to induce NKG2D ligand expression in our HEK293T model. We have identified the open chromatin sites at the NKG2D loci in MCF7 breast cancer cells, and optimised and validated this protocol. Finally we have developed "PeakHunter" – a multifunctional software tool for mapped DNaseI-seq data analysis.

Conclusions: Glucose and its contribution to purine metabolism play a central role in the induction of NKG2D ligand expression in physiological settings. The influence of glucose leads to significant alterations in cellular NKG2D-dependent immunogenicity. PeakHunter is a useful tool for analysis of mapped DNaseI-seq data.

Acknowledgments

My dream for some time was to do a PhD at Oxford, and you get what you deserve, or so it is said. It has been quite a Journey, one that has changed the way I think about the world, and about the people that exist in it (please excuse my prepositions). I suppose for this experience alone it has been worth the effort. It wouldn't have been possible without the following people:

Prof. Chris O'Callaghan, who agreed to support my application for funding to the Medical Research Council, UK, and subsequently supported my application to be admitted to the University of Oxford to undertake the research necessary to complete this DPhil on time, if not under budget. He also demonstrated great and seemingly unreserved patience in allowing me the time in his laboratory to get to this point. I'd like to thank Dr. Katherine Bull for being an intellectual force of nature, and a source of computational help and general awe and inspiration. Tanya Cheetham (Crockford) and Sophia Bennett are also due thanks, for being a source of reason, wisdom (mostly Tanya), entertainment (mostly Sophia), and reagents/consumables (both; often in exchange for one or two sheets of blue role; I think I got the better deal). Tanya also, for running the coffee club, which is more important than *you* might think, and Tanya, Sophia, and Dr. Greg Crawford for putting up with all my flow cytometry questions, specifically and repeatedly: "But how does compensation *actually* work?". I'd like to thank members of the O'Callaghan group for ongoing suggestions, and patient availability, and Megan McGregor for her patience and good will. Dr Gerald Moncayo for sharing his original observations about the effect of glucose on NKG2D ligand expression. Dr. Lina Sciesielski, for being a remarkably astute and lovely person. I'd like to thank Dr. Delphine Baup for being a good friend, in the face of all reason and rational thought, more than I deserve; I don't know where you get your patience from.

I would also like to acknowledge Prof. Christopher Hitchens, for his ongoing forthright and uncompromising opinions (via YouTube). Dr. Jim Hughes of the MRC Molecular Haematology Unit (WIMM) gave a lot of advice, email support and face to face meetings on topics of wet and computational DNaseI-seq, despite

not having any obligation or responsibility to do so; he really knew what he was talking about (hence all those publications of his), and it made a big difference to me. Dr. Linda Hughes was very kind in answering bioinformatic questions. Dr. Gerton Lunter, also, for asking Linda to help me with my bioinformatic questions. I'm grateful to Dr. Mona Hosseini and Maria Suciu for letting me watch and ask yet more questions while they carried out their critical DNaseI-seq experiments. My Balliol advisors, Dr. Piers Nye and Prof Robin Choudhury were patient and supportive at all times. I'd like to thank Dr. Silvie Blažková for believing in me throughout. My grandmother, Patricia Harrison, who thought that being a doctor meant that you could identify a tablet by its shape and colour, and putting me to the test; I always said "Ah! That's a really good one!"; I'm quite sure it did as good a job as any other. Similarly, other ancestors, who must, at the very least by definition, accept equal responsibility for my arrival at this point. My parents, Donal and Liz, who said "Do what you think you should", whose names are now recorded for posterity in the Bodleian Library, Oxford; my more clever siblings, Eamonn and Jill, who let me get on with it, and never complained.

Michael McCarthy,
June 2013,
Oxford.

Es gibt drei Wege, Wissen zu erwerben:
Erstens durch Nachdenken, das ist der edelste;
Zweitens durch Nachahmen, das ist der leichteste;
Und drittens durch Erfahrung, das ist der bitterste.

Confucius (551–479 BCE)

Contents

List of Figures

List of Tables

Abbreviations

AICAR 5-Aminoimidazole-4-carboxamide 1-β-D-ribofuranoside

APRT amidophosphoribosyl transferase

ATCC American tissue culture collection

ATCC American type culture collection

BSA Bovine serum albumin

CCDS Consensus coding sequence database

cDNA Complementary DNA

CFDA SE 5(6)-Carboxyfluorescein diacetate N-succinimidyl ester

CFSE Carboxy fluorescein succinimidyl ester

ChREBP Carbohydrate response element binding protein

CMV Cytomegalovirus

DAP10 DNAX-activation protein 10

DEPC Diethylpyrocarbonate

dH_2O deionised (MilliQ) water

DMEM Dulbecco's minimal essential medium

DMSO Dimethylsulphoxide

DNA Deoxyribonucleic acid

DNaseI Deoxyribonuclease I

dNTP Deoxynucleotide triphosphate

DON 6-Diazo-oxo-norleucine

dTTP Deoxythymidine triphosphate

EBV Epstein Barr virus

EDTA Ethylene diamine tetra acetic acid

eGFP Enhanced green fluorescent protein

EtBr Ethidium bromide

FBP Fructose 1,6-bisphosphate

FC Flow cytometry

FCS Fetal calf serum

FH Fumarate hydratase

FSc Forward scatter

GAPDH Glyceraldehyde 3-phosphate dehydrogenase

gDNA Genomic DNA

GPI Glycosylphosphotidylinositol

HCl Hydrochloric acid

HDACI Histone deacetylase inhibitors

HEK Human embryonic kidney

HGPRT Hypoxanthine/guanosine phosphoribosyl transferase

HIF Hypoxia inducible factor

HLRCC Hereditary leiomyomatosis and renal cell carcinoma

HPV Human papilloma virus

IBMX 3-isobutyl 1-methylxanthine

IFNγ Interferon-γ

IHC Immunohistochemistry

IMP Inosine monophosphate

IMPDH Inosine monophosphate dehydrogenase

ITAM Immunoreceptor tyrosine-based activation motif

MCA Methylcholanthrene

MCS Multiple cloning site

MEF Mouse embryonic fibroblast

MFI Mean fluorescence intensity

MICA MHC class I chain-related peptide A

MMLV Moloney murine leukemia virus

MOI Multiplicity of infection

Na-MOPS Sodium 4-morpholinepropanesulfonic acid

NCBI National Centre for Biotechnology Information

NK cell Natural killer cell

NKG2D Natural killer group 2D

PBS Phosphate buffered saline

PBSA Phosphate-buffered saline with 0.03% azide

PCNA Proliferating cell nuclear antigen

PCR Polymerase chain reaction

PEI Polyethylenimine

PFA Paraformaldehyde

PI Propidium iodide

PMA phorbol-12-myristate-13-acetate

PMSF Phenylmethanesulfonyl fluoride

PPADS pyridoxalphosphate-6-azophenyl-2',4'-disulfonic acid

PPP Pentose phosphate pathway

PRPP Phosphoribosyl pyrophosphate

RAE Retinoic acid early transcript

RBP RNA binding protein

RNA Ribonucleic acid

RNase Ribonuclease

RPB Reads per base

RPMI Roswell Park Memorial Institute

RTPCR Reverse transcription polymerase chain reaction

RTPCR Reverse transcription polymerase chain reaction

SCID Severe combined immunodeficiency

SDH Succinate dehydrogenase

SDS Sodium dodecyl sulphate

SRA Short Reads Archive

SSc Side scatter

TAE Tris base, Acetic acid, and EDTA

TF Transcription factor

TLR Toll like receptor

TNFα Tumour necrosis factor-α

TNFβ Tumour necrosis factor-β

UCSC University of California, Santa Cruz

vpr viral protein R

ZMP AICAR 5-monophosphate; P-AICAR

1

Introduction

Contents

1.1 Preamble

1.1.1 A brief overview

The NKG2D ligands are a family of cell surface proteins that are expressed principally on cancer cells and infected cells. The importance of NKG2D ligands in immunity lies in their

ability to access the cytotoxic, cytokine-secreting, and co-stimulating effector functions of a range of immune effector cells, including NK cells, $\gamma\delta$ T cells and CD8$^+$ $\alpha\beta$ T cells through the activating receptor, NKG2D **(Fig 1.1)**. In doing so, they provide a key immune mechanism through which unhealthy cells can directly invoke a robust and immediate immune response from these widely distributed cell types.

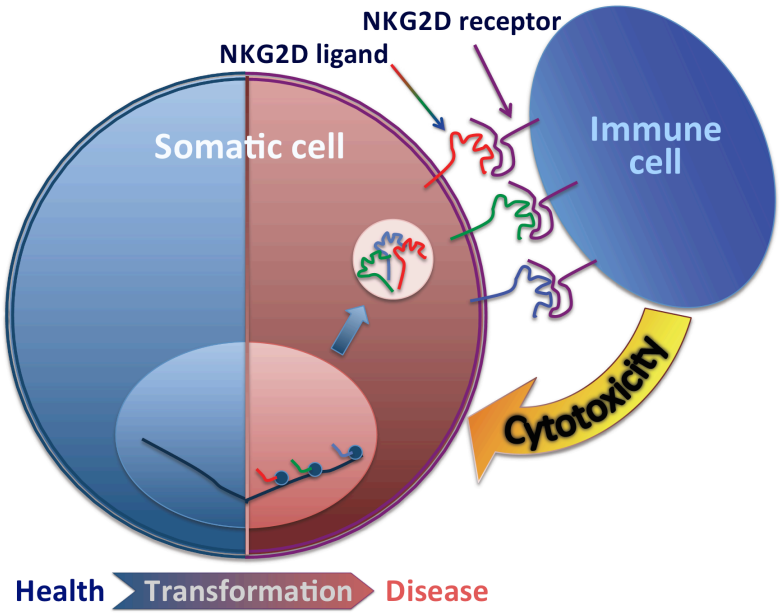

Figure 1.1: The NKG2D ligand-NKG2D receptor system. The NKG2D ligand-NKG2D receptor system provides a means by which somatic cells can invoke on themselves a robust cytotoxic innate immune response, without the need for adaptive immunity, or abnormal or foreign antigens. Healthy cells generally do not express NKG2D ligands. However, in certain typical situations, such as viral infection or malignant transformation, cell surface expression of NKG2D ligand proteins is induced. This allows recognition of NKG2D ligand-expressing target cells by NKG2D receptor-expressing immune cells, including NK cells, $\gamma\delta$ T cells and CD8$^+$ $\alpha\beta$ T cells. Successful NKG2D receptor activation can lead to cytotoxicity, cytokine secretion or co-stimulation, depending on the immune context.

While no naturally occurring human disease model of NKG2D receptor deficiency is known to exist, dysfunction of the NKG2D ligand-NKG2D receptor signalling system has been implicated in a wide range of disease processes in cellular and animal models, from cancer to autoimmunity and organ transplantation **(Fig 1.2)**. The ability to manipulate the expression of NKG2D ligands on somatic cells and the immune response they orchestrate has wide-ranging clinical application. Therefore it is useful to understand how this process is controlled.

Arguably, our understanding of how the expression of any of the ~21,000 known human genes is "regulated" is at best limited. In this context, it is unsurprising that the molecular events that occur within somatic cells to determine whether or not NKG2D ligands are expressed at the cell surface remain elusive, despite almost 20 years of intensive research.

1.1.2 Specific thesis aims

In carrying out this research, I aim to arrive at a better understanding of some of the deterministic molecular mechanisms that control cell surface NKG2D ligand expression status. Within this overall aim, the research is organised in two parts. First, I explore a specific hypothesis: the metabolic transformation ('Warburg effect') that occurs when cell surface NKG2D ligand expression is induced plays a central role in controlling cell surface NKG2D ligand expression **(Chapters 3, 4 & 5)**. Second, based on my observation **(Chapter 5)** that physiological NKG2D ligand induction has a transcriptional component, I pursue a strategy for the identification of open chromatin sites at NKG2D ligand loci **(Chapters 6 & 7)**.

Figure 1.2: Potential applications of NKG2D ligand manipulation. NKG2D ligands are of potential importance across a broad range of immune-related health settings, from immunity, to autoimmunity and alloimmunity.

In this chapter, I will build the case for the importance of this topic, by describing the NKG2D ligand system, the key role that NKG2D signalling plays in immunity and disease, the limitations in our current knowledge of how this system is regulated, and the potential clinical benefit that can arise from a deeper understanding of this process. I will then describe the

characteristics of 'Warburg metabolism', how and why transition from quiescent to Warburg metabolism is thought to occur, and why this metabolic switch may be of central importance to the regulation of NKG2D ligand expression. Background information specific to individual chapters is presented within the introduction to those chapters.

1.2 The NKG2D signalling system

1.2.1 The immune context of NKG2D

Traditionally, immunity is divided into innate and adaptive responses. Innate immunity is typically characterised by ancient origins, immediate availability, generic responses to foreign, fixed molecular patterns, limited potency, and the absence of memory (enhanced or more efficient response upon repeated exposure). Adaptive immunity is characterised by greater evolutionary complexity, delayed onset of response, antigen specificity, greater potency and immune memory. As research into mammalian immunity and immune system phylogenetics proceeds, a more nuanced picture of immunity than tradition suggests is evolving continuously (reviewed by (1)).

The innate immune features of the NKG2D response are clear. Natural killer-like signalling has been observed in primitive immune systems, such as in the bony fish, *Paralabidochromis chilotes* (2). In the present day, functional NKG2D receptor-ligand systems have survived evolution, and are evident in all mammals that have been studied, including non-human primates (3, 4, 5, 6), pigs (7), cattle (8) and rodents (9). Like other innate immune response mechanisms, the NKG2D receptor recognises a restricted molecular pattern (the NKG2D ligands), and is immediately available without the complex activation procedures associated with adaptive immunity. While some evidence has been put forward for a primitive immune memory associated with natural killer (NK) cells (a stronger response upon reactivation by cytokines; (10, 11)), this has not been associated with NKG2D signalling.

However, the NKG2D response can be seen as more complex than other forms of innate

immunity. In contrast to the conventional understanding of innate immunity, the NKG2D receptor is not activated by foreign, or mutated self antigens. Instead, the NKG2D receptor recognises normal self-antigens which have evolved to signify danger or stress, bearing some resemblance to the antigen presentation of adaptive immunity, but without actually presenting peptide, in spite of its major histocompatibility complex (MHC)-like structure. Further, the cytotoxicity that NKG2D receptor activation invokes is a complex effector function, more typically activated through humoral or cellular adaptive immunity. In these respects, the NKG2D signalling system might plausibly be viewed as a primitive form of adaptive immunity.

1.2.2 Historical overview of the NKG2D system.

The NKG2 gene family was first discovered in 1990, as part of a study aimed at identifying natural killer (NK) cell activating mechanisms. Natural killer cells were described gradually over the late 1960s and 1970s, when it was noticed that a population of lymphocytes existed that was capable of antibody dependent cytotoxic responses (12, 13). While Fc receptor (CD16) dependent activation was a received mechanism of NK cell activation, Houchins *et al* set out to identify differences between NK cells and T cells using a cDNA library subtraction approach. The NKG2 gene family was characterised in 1990 (14), and NKG2D itself was identified as a probable NK cell surface receptor by the same group in 1991 (15). The function of NKG2D wouldn't be uncovered for another 8 years.

In 1994, Bahram *et al* identified two genes from human keratinocyte and lung fibroblast cDNA libraries with homology to the human major histocompatibility complex (MHC) genes, called MHC class I polypeptide-related sequence A (MICA), and MICB (16). Groh *et al* noted in 1998 that the transfection of MICA or MICB cDNA into C1R cells (EBV-transformed B lymphoblastic cell line) or Daudi cells (Burkitt's lymphoma cell line) lead to enhanced $\gamma\delta$ T cell cytotoxicity (17). The following year, Bauer *et al* established that MICA was inducing cytotoxic responses in $\gamma\delta$ T cells and NK cells through its interaction with NKG2D on these

effector cells (18).

Table 1.1: NKG2D ligand discovery.

Ligands	Species	Year	Reference
MICA, MICB	human	1994	Bahram *et al* (16)
ULBP1-3	human	2001	Cosman *et al* (19)
ULBP4 (RAET1E; LETAL)	human	2003 & 2008	Chalupny *et al* (20) & Cao *et al* (21)
ULBP5 (RAET1G)	human	2004	Bacon *et al* (22)
ULBP6 (RAET1L)	human	2009	Eagle *et al* (23)
RAE1 (Raet1) α, β & γ	mouse	1996	Numura *et al* (24)
RAE1δ & H60a	mouse	2000	Cerwenka *et al* (25)
RAE1ϵ	mouse	2001	Girardi *et al* (26)
Mult1	mouse	2002	Carayannopoulos *et al* (27)
H60b & c	mouse	2008	Takada *et al* (28)

The discovery of other ligands for the NKG2D receptor followed gradually **(Table 1.1)**. Shortly after NKG2D was established as the activating receptor for MICA, a distinct second group of human ligands for NKG2D were found. Cosman *et al* used a cDNA library screening approach to identify cellular ligands for unique long-16 (UL16), a cytomegalovirus (CMV) protein thought to act to suppress the immune response to CMV. They identified two cDNAs that caused recombinant UL16 to bind to human cells. One of these cDNAs was MICB, already known as an NKG2D ligand, and the second was named UL16 binding protein 1 (ULBP1) (19). Screening of expressed sequence tag (EST) databases with the ULBP1 sequence identified two further ULBP proteins, named ULBP2 and ULBP3. Each of these proteins was found to bind NKG2D and activate cytotoxic responses in immune effector cells. A bioinformatic search for ULBP-related sequences subsequently identified ten genes (29), of which three have now been shown to express functional proteins: ULBP4 (RAET1E; LETAL) and its splice variants (20, 21), ULBP5 (RAET1G) and its splice variants (21, 22, 30), and ULBP6 (RAET1L) (23).

1.2.3 A concise physical description of the NKG2D system

This subject is reviewed in detail by Champsaur *et al* (31).

NKG2D. Human NKG2D (CD314) is a 316 amino acid integral membrane type 2 homodimeric protein **(Fig 1.3)**. Structurally it is classified as a C-type lectin-like protein although it isn't known to bind carbohydrates or calcium. Its crystal structure has been

determined (32). Unlike some other activating immune receptors, NKG2D does not contain an immunoreceptor tyrosine-based activation motif (ITAM). Instead, intracellular signalling from NKG2D is dependent on the recruitment of the adapter molecule DNAX-activation protein 10 (DAP10) (33). The association between NKG2D and DAP10 is maintained via an interaction between an arginine residue in the transmembrane portion of NKG2D, and an aspartate residue in the intramembranous portion of DAP10. DAP10 also lacks an ITAM, and signalling from DAP10 is dependent on the recruitment of downstream mediators such as PI3-K (34) and Grb2-Vav1 (35) via a YINM signalling motif in DAP10.

In addition to interacting with DAP10, mouse NKG2D also binds to DAP12, an ITAM containing adapter molecule (36). This allows direct ITAM-dependent signalling upon activation of NKG2D.

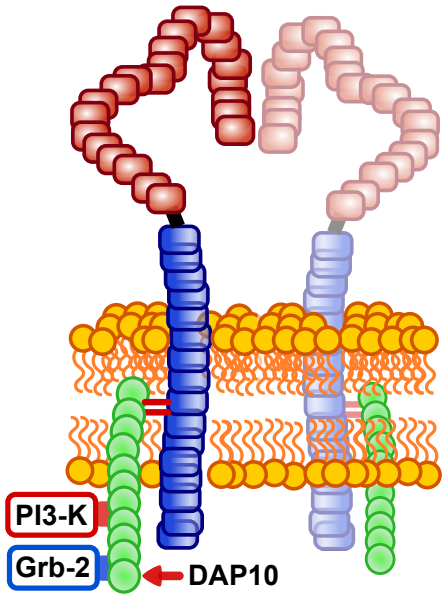

Figure 1.3: The NKG2D receptor. The NKG2D activating receptor is a homodimeric type 2 integral membrane protein. Intracellular signalling from NKG2D is dependent on an intramembranous interaction between an arginine residue in the transmembrane portion of NKG2D and an aspartate residue in the adaptor molecule DNAX-activation protein 10 (DAP10) (green circles). Downstream from DAP10, recruitment of the signalling molecules phosphoinositide 3-kinase (PI3-K) and growth factor receptor-bound protein 2 (Grb-2) via a YINM (Tyr-Ile-Asn-Met) motif is important for induction of cytotoxicity.

NKG2D has a comparatively wide cellular distribution for an immune activating receptor. Bauer *et al* described the cellular distribution of NKG2D in human immune cell types in 1999

(18). They found NKG2D receptor expression in $\gamma\delta$ T cells, CD8$^+$ $\alpha\beta$ T cells, and NK cells. Subsequently, NKG2D receptor expression has been described on human iNKT cells (37), some CD4$^+$ $\alpha\beta$ T cells (38).

Human NKG2D ligands. As described above, humans have eight NKG2D ligands, each capable of immune cell activation through interaction with the NKG2D receptor. These are MICA, MICB, and ULBPs 1-6 **(Fig 1.4)**. The *MICA* and *MICB* genes are unusually highly polymorphic. There are currently 91 known alleles of MICA, and 40 known alleles of MICB (hla.alleles.org; last updated April 2013). The existence of so many functional protein coding alleles suggests that a selective pressure for MICA/MICB diversity exists, although the reasons for this are unclear. While no NKG2D-null human phenotype has been described, a MICA-MICB null genotype is known to occur in ∼3.7% of Asian populations (39), without any reported immune or disease phenotype. MICA and MICB are monomeric type 1 integral membrane proteins, with three structural extracellular domains referred to as α1, α2, and α3. MICA and MICB do not bind β_2-microglobulin, and do not present peptide. The crystal structures for MICA (32, 40) and MICB (41, 42) have been published. Interestingly, the *MICA* and *MICB* genes are not present in rodents.

The human UL16-binding proteins, ULBP1-6, are distinct from MICA and MICB in characteristically possessing α1 and α2 domains, without an α3 domain **(Fig 1.4)**. ULBP4 and 5 are integral membrane proteins, while ULBP1-3 and ULBP6 are anchored to the cell surface membrane via a glycosylphosphatidylinositol (GPI) linkage. The ULBPs are considerably less polymorphic than MICA and MICB (43). The crystal structure of ULBP3 has been published (44).

Mouse NKG2D ligands. The *MICA* and *MICB* genes are absent in mice. Instead, mice have a range of ULBP homologues that began to be identified in 1996. The first group of genes discovered, RAE1 α, β and γ (retinoic acid early transcript) were identified in mouse carcinoma cells as GPI-linked proteins (24). RAE1δ and H60a were described in 2000 (25), and RAE1ϵ was found in 2001 (26). Mult1 was identified in 2002 (27) and finally, H60b and

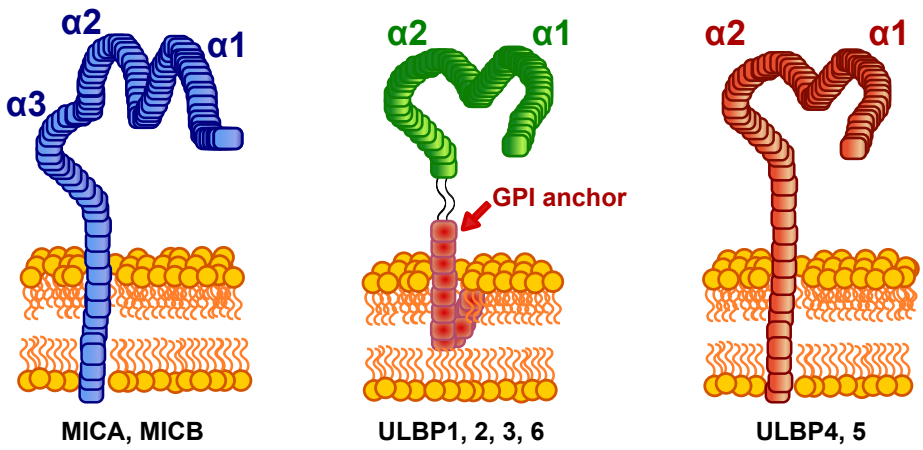

Figure 1.4: The NKG2D ligands. The human NKG2D ligands consist of a set of eight cell surface proteins. MICA and MICB are transmembrane proteins with three extracellular domains, referred to as $\alpha 1$, $\alpha 2$, and $\alpha 3$. Structural studies suggest that the 'peptide-binding grove' between the $\alpha 1$ and $\alpha 2$ domains is too small to present a peptide, unlike the classical major histocompatibility complex (MHC) molecules. MICA and MICB also do not bind to β_2-microglobulin. ULBP1-3 and ULBP6 are glycosylphosphotidylinositol (GPI)-linked proteins, lacking a transmembrane domain themselves. These ligands have two extracellular domains responsible for interaction with NKG2D called $\alpha 1$ and $\alpha 2$. ULBP4 and 5 exist as transmembrane proteins, also containing $\alpha 1$ and $\alpha 2$ extracellular domains.

c where uncovered in 2008 by Takada *et al* (28).

Each of the mouse NKG2D ligands has two extracellular domains, referred to as $\alpha 1$ and $\alpha 2$. Like the human ULBP4 and 5, Mult1 and H60b and c are transmembrane proteins. The RAE1 genes and H60a are GPI-linked proteins.

1.2.4 The physiological functions of the NKG2D system

1.2.4.1 The NKG2D ligands: all of a kind?

Most higher mammals have functioning NKG2D-receptor-ligand systems that consist of one NKG2D receptor capable of binding and being activated by a range of NKG2D ligands. It would be reasonable to ask: "why are there so many functional ligands for a single receptor?". Studies on mouse NKG2D ligands suggest that different ligands do indeed have distinct NKG2D receptor binding characteristics (45, 46). While it is clear that binding characteristics are different between ligands, the functional consequences of these differences are unclear. No published data is available comparing binding characteristics or functional outcomes for the human NKG2D ligand-receptor interactions.

Similarly, most NKG2D ligand studies (which will be described below) focus on different combinations of NKG2D ligands. As such, the patterns of NKG2D ligand induction in response to different stimuli, or indeed in different physiological contexts, have not been resolved. In any given situation where NKG2D ligand expression is induced (e.g. cancer, CMV infection), it is common for more than one NKG2D ligand to be expressed, but consistent patterns of expression between individuals have not been identified.

1.2.4.2 Cellular consequences of NKG2D receptor activation

Cytotoxicity. Cellular cytotoxicity is the most commonly described NKG2D effector function. NKG2D has become known as a 'dominant' activating receptor for cellular cytotoxicity on NK cells. This title is in reference to the observation that both MICA/MICB (18) and ULBPs (19) acting through the NKG2D receptor on NK cells can overcome the inhibitory signalling between target cell HLA class I antigens, and NK cell KIR (killer-cell immunoglobulin-like receptor) receptors, to result in target cell cytotoxicity. Despite these observations, it is probable that for NK cell cytotoxicity, activation is dependent on a broader balance of inhibitory and activating signalling (reviewed by Lanier (47)). It is certainly the case that freshly isolated human NK cells exhibit NKG2D-dependent cytotoxicity toward NKG2D ligand expressing targets *in vitro/ex vivo* (19, 48).

It is similarly clear that NKG2D-dependent cytotoxicity is seen in $\gamma\delta$ T cells exposed to NKG2D ligand-bearing target cells (18, 49, 50). Regarding $CD8^+$ $\alpha\beta$ T cells, TCR-independent 'innate' cytotoxicity has been shown to occur between NKG2D receptor-expressing primary $CD8^+$ $\alpha\beta$ T cells and NKG2D ligand expressing target cells (51, 52). However, this was in the context of cytokine support *in vitro*, and the extent to which these observations are relevant *in vivo* is unclear. $CD4^+$ $NKG2D^+$ $\alpha\beta$ T cells have also been shown to have NKG2D-dependent cytotoxic activity in one report (53). However, the characteristics and immune role of this $CD4^+$ $NKG2D^+$ $\alpha\beta$ T cell population is not clearly understood (54, 55).

Cytokine production. The extent to which NKG2D receptor activation induces cy-

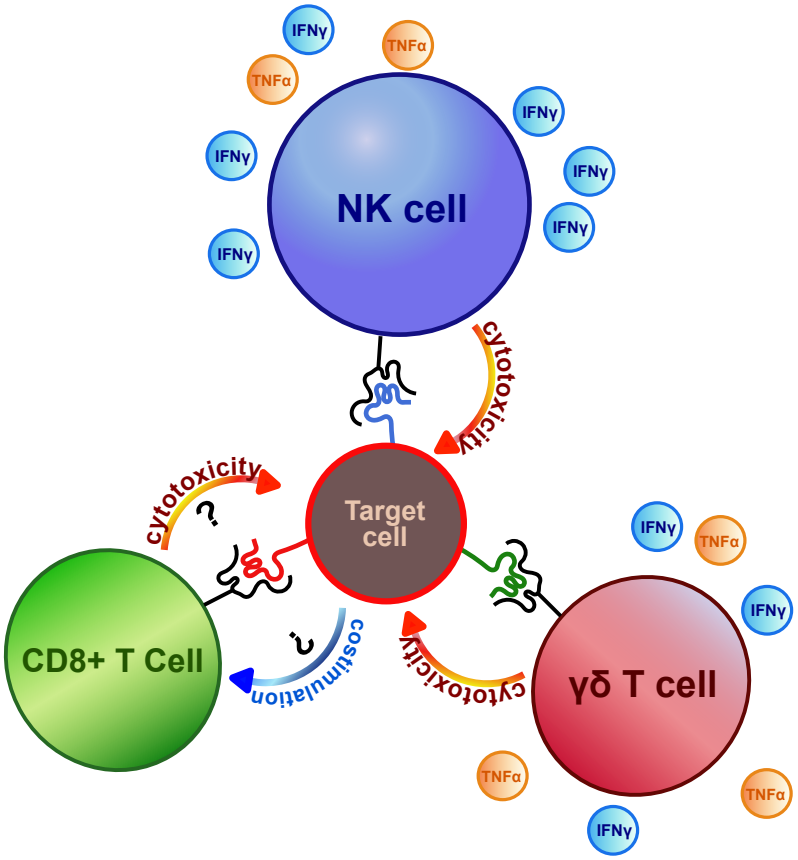

Figure 1.5: The typical effector functions of NKG2D. The most intensely studied role of NKG2D signalling is NK cell activation. Current understanding of NK cell activation suggests that a balance of inhibitory and activating signal is important for activation. Nevertheless, there is clear evidence that target cell expression of NKG2D ligands is capable of inducing an NK cell cytotoxic response. This is also the case with $\gamma\delta$ T cells. There is good evidence that NKG2D receptor ligation induces cytokine secretion in both cell types. IFNγ and TNFα are the two most consistently demonstrated cytokines produced following NKG2D receptor ligation. The role of NKG2D signalling in $CD8^+$ $\alpha\beta$ T cells is less clear. TCR-independent cytotoxicity has been shown *in vitro* with cytokine support, but whether this occurs *in vivo* is unclear. Similarly there are conflicting reports surrounding the ability of NKG2D ligation to provide co-stimulation in the context of TCR-dependent activation.

tokine secretion is less well understood, and the systematic measurement of cytokine output following NKG2D activation is rare. However, there is evidence to suggest that activation of the NKG2D receptor on NK cells with recombinant NKG2D ligands leads to secretion of the prototypical NK cell cytokines, IFNγ (56) and TNFα (56, 57). Cytokine secretion by NK cells in this context appears to by synergistically enhanced by co-ligation of other NKG2D receptors, such as 2B4, LFA-1 or CD16 (58, 59). Separate studies have also demonstrated IFNγ and TNFα secretion by $\gamma\delta$ T cells following NKG2D receptor ligation (49, 50). Cytokine production by NKG2D-stimulated CD8$^+$ $\alpha\beta$ T cells is less well described.

T cell co-stimulation. The ability of NKG2D to provide co-stimulatory signalling during the TCR-dependent activation of CD8$^+$ $\alpha\beta$ T cells is somewhat controversial. The idea that the NKG2D receptor may be capable of delivering co-stimulation comes from the observations that the NKG2D receptor is clearly stimulatory in other settings (NK cells, $\gamma\delta$ T cells), that it is constitutively expressed on human CD8$^+$ $\alpha\beta$ T cells, and that its signalling adaptor, DAP10 **(Fig 1.3)** has a similar intracellular signalling motif (YINM) to that of the prototypical co-stimulatory molecule, CD28 (YMNM). However, the evidence to support a co-stimulatory role for NKG2D is conflicting. Evidence in favour of co-stimulation comes from three studies, which show that both peptide-specific TCR-dependent T cell activation and non-peptide specific TCR-dependent activation of T cells to produce cytokines and cytotoxic responses can be co-stimulated by NKG2D stimulation (60, 61, 62). However, when NKG2D stimulation was directly compared to CD28 stimulation, no NKG2D co-stimulatory activity was demonstrated (63). A more recent study has suggested that NKG2D may play a co-stimulatory role for tissue resident effector CD8$^+$ $\alpha\beta$ T cells, while CD28 is more important in the case of naïve CD8$^+$ $\alpha\beta$ T cells (64). Given the complexities known to be involved in T cell activation, it is difficult to judge how relevant these results are *in vivo*, although, given NKG2D's constitutive presence on CD8$^+$ $\alpha\beta$ T cells, it is likely that a significant *in vivo* function of some description does exist.

1.2.4.3 Physiological NKG2D ligand induction

To gain a better insight into which molecular processes might control NKG2D ligand induction, I searched the literature to identify studies demonstrating differential NKG2D ligand expression in physiologically relevant settings. Studies carried out on primary cells, or *ex vivo* tissue samples were at the core of this search. Below is a description of the key physiological settings in which NKG2D ligand expression is seen.

Infection. One of the most widely explored models of NKG2D ligand induction in primary cells is the infection of human fibroblasts with laboratory strains of cytomegalovirus, principally strain AD169. The expression of MICA, MICB and ULBPs 1-3 in fibroblasts is induced by cytomegalovirus infection both at the level of mRNA (65, 66) and protein (60, 65, 67, 68). A second human herpes virus, Epstein Barr virus (EBV) has been shown to induce ULBP1 (69) and ULBP4 (50) at the cell surface in infected B cells, but examination of other NKG2D ligands suggest that any effect of EBV on NKG2D ligand induction is much more limited than that of CMV.

HIV virus infection induces mRNA and cell surface protein expression of MICA, MICB and ULBP1-3 in primary $CD4^+$ $\alpha\beta$ T cells (70, 71). The HIV viral protein R (vpr), which plays a role in nuclear import of the HIV pre-integration complex, and in initiating G2 phase cell cycle arrest in dividing cells, is important for the induction of NKG2D ligand expression (71, 72).

Several isolated studies have examined the influence of other infectious agents on NKG2D ligand expression. Influenza virus (73) and adenovirus (74) have been found to induce ULBP1-3 or MICA/B expression in primary immature dendritic cells and fibroblasts respectively.

Toll-like receptor ligation. A second method of inducing NKG2D ligand expression in primary cells is through ligation of Toll-like receptors (TLRs). TLRs are an important part of the innate immune system, representing a group of proteins that bind to simple 'molecular patterns', such as foreign DNA, RNA and lipopolysaccharide (for review see Takeda *et al*

(75)). The TLRs, originally described in drosophilla, were first found to play a role in human immunity in 1997 (76).

It has been firmly established that ligation of TLR3 (with polyI:C) or TLR4 (with lipopolysaccharide) on primary human macrophages/monocytes induces MICA expression at both mRNA and protein levels (77, 78, 79). One study using higher concentrations of lipopolysaccharide to stimulate TLR4 on primary macrophages found a wider effect on NKG2D ligand expression, with induction of cell surface MICA, MICB and ULBP1-3 expression (80). Again, the molecular mechanisms underlying this effect are unclear.

Cellular activation/proliferation. Primary cell activation and proliferation has been associated with induction of cell surface NKG2D ligand expression. The most intensely studied example of this is T cell activation, where NKG2D ligand expression has been found on proliferating activated T cells. In 2002, Molinero *et al* demonstrated the up-regulation of MICA on anti-CD3 stimulated or anti-CD28/phorbol myristate acetate (PMA)-stimulated $CD4^+$ and $CD8^+$ T cells, measured by flow cytometry and western blot (81). The authors used a polyclonal rabbit serum, raised against a MICA peptide, for their experiments, which raises some questions about the validity of these results. However, their findings have since been validated by several authors. Rabinovich *et al* demonstrated the up regulation of NKG2D ligands on TCR transgenic ovalbumin stimulated mouse T cells, measured by flow cytometry using an NKG2D receptor-Fc fusion protein. RTPCR suggested that H60a was responsible for some of this effect (82). Cereboni *et al* demonstrated up regulation of MICA, MICB and ULBP1-3 on $CD4^+$ and $CD8^+$ T cells using several methods of T cell activation, including PMA/ionomycin stimulation, allogeneic cells, and peptide specific T cell activation (83, 84). In these experiments, they also presented evidence to suggest that the up-regulation of NKG2D ligands on activated, proliferating T cells made these cells themselves targets for NK cell cytotoxicity (84). This insight, that excessive or aberrant cellular proliferation confers susceptibility to NK cell cytotoxicity through the up-regulation of NKG2D ligands is interesting.

T cell activation is a good model of cell proliferation, as the timing of activation can be controlled and known precisely. Measuring NKG2D ligand expression in primary cells growing in culture is less useful, as concepts of proliferation, confluence and contact inhibition of proliferation, are not as precise. Nevertheless, NKG2D ligand expression on the surface of proliferating primary fibroblasts was described in 2006 (85). Zou *et al* showed that MICA protein and mRNA levels decreased progressively with increasing confluence, and that chemical interference with cell-cell contact prevented the down-regulation of MICA.

Cancer. Increased cell surface NKG2D ligand expression has been found in a wide range of solid and haematological malignancies. An increasing number of published surveys demonstrate this **(Table 1.2)**. To date, increased cell surface NKG2D ligand expression has been found in pancreatic cancer, colorectal cancer, hepatocellular cancer, ovarian cancer, thyroid cancer, melanoma, prostate cancer, astrocytomas and in a range of haematological malignancies. While it is reasonable to make this conclusion based on the evidence presented, some points are worth considering. Most of this data is generated from immunohistochemical staining of fixed and embedded tumour blocks. This method is at best semi-quantitative, and accurate quantitation of NKG2D ligand expression levels is not possible. Additionally the 'control' samples used in each of these studies is less than ideal. In most studies where control data was included, NKG2D ligand expression on cancer cells was measured in comparison to NKG2D ligand expression in 'surrounding healthy tissue'. It isn't entirely clear that this is a reasonable control, as it is difficult to distinguish surrounding healthy tissue that is of the same epithelial type as the cancer. Furthermore, it isn't clear what NKG2D ligand expression levels are 'normal' for healthy epithelial tissues of different histological origins.

Also of interest, in the colorectal cancer case series of 462 cases, when the cancers were classified by stage from stage I to stage IV, it was noted that NKG2D ligand expression was higher in early stage cancers, and decreased with increasing stage (86). While it isn't possible to make any causal inferences from this data, it is tempting to speculate that this may be due to immune selective pressure.

Table 1.2: Studies examining NKG2D ligand expression in cancer.

Cancer	Cases	Key ligands	Method	Reference
Pancreatic adenocarcinoma	103	MICA	IHC	Duan *et al* (87)
Colorectal adenocarcinoma	462	MICA/B, ULBP1-4	IHC	McGilvray *et al* (86)
Hepatocellular carcinoma	10 + 10	MICA/B	RTPCR, IHC	Jinushi *et al* (88)
Ovarian cancer	82	MICA/B, ULBP2	IHC	Li *et al* (89)
Ovarian cancer	300	MICA, ULBP1-5	IHC	McGilvray *et al* (90)
Thyroid cancer	47	MICA/B	IHC	Xu *et al* (91)
Melanoma	40	MICA/B	IHC	Vetter *et al* (92)
Uveal melanoma	9	MiCA/B	IHC	Vetter *et al* (93)
Prostate adenocarcinoma	165	MICA/B	IHC	Wu *et al* (94)
Multiple myeloma/MGUS	60 + 10	MICA	IHC, FC	Jinushi *et al* (95)
Haematological malignancy	25	MICA/B, ULBP1-3	FC	Salih *et al* (96)
Chronic myeloid leukemia	9	MICA/B, ULBP1-2	FC	Boissel *et al* (97)
Glioma	5	MICA/B, ULBP1-3	RTPCR, FC	Friese *et al* (98)

IHC - Immunohistochemistry; FC - Flow cytometry; RTPCR - Reverse transcription PCR.

Cellular 'stress'. Many studies have examined the effect of generic cell stressors on NKG2D ligand expression in cell lines. Fewer studies have examined the effects of these agents on NKG2D ligand expression in the primary cell setting. Ionizing radiation was found to induce MICA, MICB and ULBP1-3 cell surface expression on a range of cell lines (99). The relevance of ionising radiation for NKG2D ligand induction in primary cells was demonstrated in 2005, by treating primary human fibroblasts with ionising radiation (100).

The effect of several DNA damaging agents on NKG2D ligand expression in primary fibroblasts has also been examined. The antibiotic, mitomycin C (which acts to cross-link DNA), the pyrimidine analogue, cytarabine, and the alkylating agent, cisplatin were each found to induce ULBP1-3 cell surface expression on primary fibroblasts (100).

1.2.5 The role of the NKG2D system in disease

I have outlined above the evidence that suggests NKG2D ligand expression is induced in a range of physiological settings, and also the evidence that describes the cellular consequences that increased NKG2D ligand expression can have (delivery of NK cell cytotoxicity, cytokine secretion, etc.). Perhaps NKG2D signalling is a vestigial arm of the immune response, an evolutionary intermediate between innate and adaptive immune responses that is insignificant today. Is there any evidence that the NKG2D signalling system plays a role in disease

processes, through over-activity or under-activity?

1.2.5.1 Autoimmune disease

A large amount of research that has been carried out into the pathogenesis of autoimmune disease focuses on the role of adaptive cellular and humoral immunity. Since the NKG2D receptor-NKG2D ligand signalling system was described ∼15 years ago, the availability of evidence to suggest a role for NKG2D in autoimmune disease pathogenesis is increasing. While it is probably too early to come to conclusions about the extent of NKG2D's involvement in any given autoimmune disease, there is reasonable evidence to conclude that NKG2D does play a role in autoimmunity. Whether this is predominantly a direct role (i.e. NKG2D receptor bearing immune cells acting against NKG2D ligand expressing host tissues) or an indirect role (e.g. failure to negatively regulate autoimmune T cell activity) is not fully clear. Teasing out a specific role for NKG2D in these processes is challenging, given that NKG2D is expressed on most immune cells implicated in autoimmune settings.

Rheumatoid arthritis. Natural killer cells, but not specifically NKG2D, have been observed in the synovial fluid of patients with rheumatoid arthritis soon after natural killer cells were first described (101). This NK cell activity has been linked to osteoclast formation and bone erosion in rheumatoid arthritis (102), and depletion of NK cells from mice with collagen induced arthritis significantly reduced bone erosion and synovitis (102).

In 2003, Groh *et al* demonstrated MICA/B expression by immunohistochemistry in a synovium section from a patient with rheumatoid arthritis (103). The cells staining positive for MICA/B histologically resembled synoviocytes. They also identified a population of $CD4^+$ $CD28^-$ $NKG2D^+$ T cells, present in both peripheral blood and synovial tissue. The production of IFNγ and TNFα by these cells was co-stimulated by NKG2D activation. While this study provides circumstantial evidence of a role for NKG2D in the pathogenesis of rheumatoid arthritis, it is unclear whether the pattern of synovial MICA/B expression is different in healthy joints, and also whether the $CD4^+$ $CD28^-$ $NKG2D^+$ T cells found in the

rheumatoid patients are abnormal.

Finally, Andersson *et al* showed that blocking NKG2D signalling through intraperitoneal administration of an anti-NKG2D antibody improved clinical outcomes and the histological appearance of joints in collagen inducted arthritis, a mouse model of rheumatoid arthritis. While it is likely that the pathogenesis of rheumatoid arthritis and collagen induced arthritis differ, this study provides the strongest evidence available that NKG2D signalling plays a significant role in rheumatoid arthritis.

Multiple sclerosis. Two studies have looked for NKG2D ligand expression in multiple sclerosis. Saikali *et al* assessed MICA/B expression in snap-frozen post mortem samples of multiple sclerosis (MS) plaques or white matter from non-MS brains. They found high levels of oligodendrocyte MICA/B expression at plaque edges, and little or no expression outside of plaques or in non-MS brains (104). Similarly, Fernandez-Morera *et al* used immunohistochemistry to demonstrate MICA/B expression in multiple sclerosis plaques, and also demonstrated a correlation between serum soluble MICB levels, and disease activity in multiple sclerosis patients (105). There is no evidence currently to support a causative role for NKG2D signalling in multiple sclerosis, but certainly enough evidence to warrant further investigation.

Inflammatory bowel disease. Inflammatory bowel disease describes two clinically distinct idiopathic immune-related illnesses of the gastrointestinal tract, ulcerative colitis, and Crohn's disease. Identifying a role for NKG2D signalling in inflammatory bowel disease is complicated by the fact that NKG2D ligands are normally expressed on intestinal epithelial cells (106). There are two types of evidence implicating NKG2D in the pathogenesis of inflammatory bowel disease: studies that describe a change in the level of expression, distribution, or function of the NKG2D receptor or it's ligands, and studies demonstrating the impact of NKG2D blockade on experimental mouse colitis.

La Scaleia *et al* examined the mucosal histology and peripheral blood cell phenotypes in 71 cases of paediatric inflammatory bowel disease, and 51 normal controls. Using immuno-

histochemical methods, they found relative stability in the level of NKG2D ligand expression in intestinal epithelial cells, but increased NKG2D ligand expression (MICA/B, ULBP1-2) in 'leucocytes' infiltrating active colitic lesions compared to inactive lesions or healthy controls (107). Similarly, the proportion of NKG2D expressing T cells was increased in active lesions, and NKG2D$^+$ cells were more likely to produce IFNγ than NKG2D$^-$ cells. The relationship with IFNγ production was stronger in CD4$^+$ cells than in CD8$^+$ cells.

Allez *et al* carried out a similar study in adult patients with Crohn's disease. They found that MICA expression was higher on intestinal epithelial cells in active colitis lesions, compared to inactive lesions or epithelium from healthy patients. Interestingly, they also isolated a population of CD4$^+$ NKG2D$^+$ T cells which were increased in active colitis lesions, and readily produced IFNγ when exposed to MICA expressing C1R target cells. In the presence of IL-2 *in vitro*, the CD4$^+$ NKG2D$^+$ T cells were also capable of TCR-independent cytotoxic responses to MICA-expressing C1R target cells.

Taken together, these studies suggest an important role for CD4$^+$ NKG2D$^+$ T cells in inflammatory bowel disease. It remains to be clarified whether cytokine production (IFNγ) or cytotoxicity is the key effector function of these cells, and whether these lymphocytes are activated predominantly by intestinal epithelial cells expressing increased cell surface NKG2D ligands, or 'leucocytes'/antigen presenting cells expressing increased cell surface ligands.

Two independent studies have examined the role of NKG2D signalling in experimental mouse colitis. The mouse models were generated by adoptive transfer of CD4$^+$ CD25$^-$ (108) or CD4$^+$ CD45b^{high} T cells to SCID mice. Following adoptive transfer these mice develop a colitis resembling inflammatory bowel disease. Both studies treated mice with intraperitoneal anti-NKG2D (CX5) or isotype control antibodies. Both studies found that NKG2D blockade prevented the development of colitis in these mice, providing evidence that NKG2D signalling plays an important role in the maintenance, if not initiation of colitis.

Diabetes. Non-obese diabetic (NOD) mice were bred in Japan in a process of selection of mice with diabetic characteristics (109), and were soon found to have an immune-mediated

diabetes (110). The NOD mouse has since been studied as a model of type I, or autoimmune-mediated diabetes in humans. Ogasawara *et al* found that pancreatic islet cells in these mice expressed NKG2D ligands at the mRNA and cell surface protein level, in contrast to pancreatic islet cells from healthy mice where no expression was found (111). NKG2D blockade, through the intraperitoneal administration of anti-NKG2D (CX5) prevented the development of diabetes in these mice. In a similar experiment, Joo *et al* administered a pancreatic islet targeted plasmid encoding a soluble version of the mouse NKG2D ligand, RAE-1 γ, recapitulating a mechanism of NKG2D immune evasion observed in some cancer cells and placental tissue. Mice treated with the soluble RAE-1 γ vector did not develop diabetes, suggesting that the interaction between epithelial (islet) cell NKG2D ligands and immune cell NKG2D is important for the development of diabetes in NOD mice.

Some studies in humans with type 1 diabetes (112), and late onset autoimmune diabetes (113), have noted altered frequencies of NK cells in the peripheral blood of these patients compared to healthy controls, but the wider role of NKG2D in autoimmune diabetes in humans hasn't yet been studied.

1.2.5.2 Transplant biology

Solid organ transplantation is an immunologically complex event. A central mechanism of solid organ graft rejection is the development of an adaptive immune response to non-self antigens expressed within the graft. Because human leucocyte antigens (HLA) are the most polymorphic human genes, and therefore most likely to generate recipient anti-donor cellular and humoral adaptive immune responses, it is standard practice to match donor and recipient HLA-A, -B and HLA-DR antigens.

The role of donor NKG2D ligand expression in the grafted organ in graft immune-mediated rejection is not fully understood, and is complicated by the fact that NKG2D ligands themselves activate and co-stimulate immunity, and are also (notably MICA and MICB) polymorphic molecules and thus alloantigens in their own right.

Immunohistochemical evidence for protein-level expression of NKG2D ligands in human solid organ transplants comes from a series of graft biopsies reported in 2002 (114). Hankey *et al* found a positive correlation between histological evidence of transplant rejection and the expression of NKG2D ligands in the graft. A later study examined mouse NKG2D ligand expression at the mRNA level in transplanted tissues (115). They found a biphasic pattern of NKG2D ligand mRNA expression, with early stage (3-5 days) increase in mRNA levels observed in syngeneic (genetically identical) and allogeneic transplants, and a later (5-9 days) increase in mRNA levels that was only seen in allogeneic tissue grafts with functioning adaptive immunity. These results suggest the possibility that NKG2D ligand expression in allogeneic transplanted tissue is induced by the adaptive immune response to the transplanted tissue.

There is little direct evidence that functional NKG2D signalling plays an important role in transplant rejection. However there is increasing evidence that an adaptive humoral immune response to the polymorphic MICA/B antigens does play a role in this process. The presence of both pre-existing (116), and post-transplant induced (117, 118) antibodies against MICA in solid organ transplant recipients have been associated with increased rates of transplant immune rejection. Further prospective studies would be necessary to identify the size of this effect, and to work out the immunological role of NKG2D in the process of transplant rejection. Nevertheless, convincing evidence suggests that these ligands are expressed on solid organs transplants, and immune reaction to this expression is harmful to graft survival. In this context, the ability to suppress NKG2D ligand expression would in principal be of clinical use in improving graft survival rates.

1.2.5.3 A role in cancer immunology?

There is a lot of good evidence that NKG2D ligands are expressed in a wide range of cancers **(Table 1.2)**, but is there any evidence that this expression is immunologically relevant? The stage-dependent expression of NKG2D ligands in colorectal cancer (86) would be con-

sistent with a hypothesis of 'immune-editing', the selective removal by the immune system of immunogenic NKG2D ligand$^{[HIGH]}$ cancer cells over time. Indeed, evidence for NKG2D ligand-dependent immunoediting of carcinogen induced tumours exists in mice (119). However, no human disease model of NKG2D receptor deficiency is known to occur, making direct evidence for an NKG2D role in immunosurveillance difficult to find.

Some evidence that NKG2D does play an active part in immunosurveillance of cancer comes from animal models. Several authors have demonstrated that mice have a greater capacity for immune rejection of cell-line based tumours that express NKG2D ligands at the cell surface, compared to cell-line tumours that do not express NKG2D ligands (120, 121, 122). This ability to reject NKG2D-expressing tumours appears in these publications to be most heavily dependent on the presence of cytotoxic-competent NK cells. A later study addressed this question in native carcinogen induced tumours, by giving mice carcinogenic doses of methylcholanthrene (MCA), either with or without routine administration of an NKG2D-blocking antibody or an isotype control. They found that mice treated with the anti-NKG2D antibody developed more tumours (principally fibrosarcoma), and died sooner than mice with intact NKG2D signalling.

Two NKG2D receptor mouse knockout models have since been described (123, 124). The first of these addressed the question of NKG2D-mediated tumour immunity. They crossed NKG2D knockout mice onto two distinct spontaneous mouse cancer models: the TRAMP model of prostate carcinoma (transgenic adenocarcinoma of the mouse prostate (125)) and an Eμ-myc mutant model of B cell lymphoma. Both of these mice develop spontaneous autologous tumours, without the need to administer carcinogens. In both cases, mice deficient in NKG2D developed spontaneous tumours faster, and with a greater total tumour mass. Additionally, mice died sooner in the absence of NKG2D.

Taken together, there is good evidence from *in vivo* animal models to suggest that the high expression of NKG2D ligands known to occur on a wide range of human cancers is likely to be playing a significant immunological role.

1.2.6 The regulation of NKG2D ligand expression

1.2.6.1 A brief introduction

Over the previous sections, we have described what the NKG2D ligand system is, how it functions, where ligands are expressed, and why its study is of clinical relevance. The principal aim of this thesis is to improve our understanding of how NKG2D ligand cell surface protein expression is switched on or off. What is already known about this process? In this section I will attempt to summarise the existing evidence that describes known features of the regulation of NKG2D ligand expression. This topic is also the subject of several helpful review articles (31, 126, 127). Here I will aim to focus on events that influence NKG2D ligand expression for which there is some consistent or reproduced evidence.

1.2.6.2 Pre-transcriptional regulation

This category of 'pre-transcriptional' regulation includes cellular processes that are known to influence NKG2D ligand expression, without having a direct connection to the transcriptional or post-transcriptional fates of NKG2D ligands. In years to come, it may turn out that some of these cellular processes in fact act after transcription, but in the absence of certainty about the point of action, and the usefulness of the relevant studies to understand what is known of NKG2D ligand regulation, they are included in this category.

TLR signalling. The toll-like receptors (TLRs), briefly mentioned previously, are a group of proteins, that represent prototypical pattern recognition receptors (reviewed by Takeda *et al* (75)). They are a group of germ-line encoded proteins, widely expressed in cells, that bind to a range of well defined foreign molecular signatures, such as lipopolysaccharide, bacterial glycolipids and lipopeptides, DNA, RNA, and fungal molecules. Binding of TLR to its cognate ligand leads to activation of signalling pathways which are aimed at priming the innate immune system for defence against these foreign substances. The signalling pathways downstream of the TLRs is relatively well defined.

TLR ligation was first found to induce NKG2D ligand expression in 2004, when Hammerman *et al* noted that mouse peritoneal macrophages, following stimulation with LPS (TLR4 agonist) and PolyI:C (a structural analogue of double-stranded RNA; TLR3 agonist) in particular lead to cell surface expression of RAE1 molecules, a process dependent on the TLR signalling adapter molecule MyD88. Mouse NKG2D ligand expression is also induced by TLR4 activation in renal epithelial cells an effect that also depends on the MyD88 adapter molecule (128).

In 2003, Schreiner *et al* showed that primary human myoblasts, when exposed to TLR3 agonist, PolyI:C, up-regulate the expression of MICA, MICB and ULBP1-3 at the mRNA level, and this leads to increased NKG2D receptor-dependent susceptibility to NK cell cytotoxicity. Evidence for the impact of TLR stimulation on NKG2D ligand expression in human monocyte-derived dendritic cells is mixed. While Ebihara *et al* have reported evidence for induction of ULBP and MICA expression in response to poly I:C or LPS stimulation (129), Kloss *et al* describe an effect on MICA only (77). This effect on cell surface MICA expression is also confirmed by Eissmann *et al* (78).

Together these studies provide convincing evidence that the ligation of TLRs, most notably TLR3 and TLR4, induces cell surface, and mRNA level expression of NKG2D ligands in a MyD88-dependent manner, although the pattern of expression, or tissue restrictions, are not entirely clear.

Histone deacetylase inhibition. The histone deacetylase enzymes are a large and growing family of proteins. They were named after their capacity to deacetylate histone proteins, in opposition to the histone acetyl transferase enzymes. In this context, the acetylation of histones is thought to reduce the packing of chromatin, leading to a more open chromatin formation, and the histone deacetylase enzymes reverse this effect. Increasingly, the histone deacetylase enzymes are found to deacetylate other non-histone proteins with functional consequences. Histone deacetylase inhibitors (HDACI) then can be expected to have widespread changes on cell physiology, not least an increased tendency to gene expression in general

(reviewed by Bolde *et al* (130)).

Importantly, studies demonstrating NKG2D ligand induction by HDACI are restricted to the setting of malignancy or in cell lines. It is unclear how healthy quiescent cells might respond. Armeanu *et al* demonstrated increased cell surface and mRNA expression of MICA and MICB, leading to increased NKG2D-dependent cytotoxicity, when they treated hepatoma cell lines with the HDAC inhibitor, sodium valproate. In 2007, the HDACI trichostatin A was found to increase MICA and MICB protein and mRNA expression in leukemic cells and cell lines (131). These findings were replicated by Andresen *et al* using Jurkat cells with the HDACI FR901228 (132), and again in HeLa and HepG2 cells using sodium valproate and sodium butyrate in 2009 (133).

Taken together, these studies strongly suggest that HDAC inhibitors do influence MICA and MICB expression at protein and mRNA levels, in malignant cells and cell lines, although like most of these studies, the mechanism is not clear.

Growth factor driven NKG2D ligand regulation. The idea that cellular growth and proliferation can give rise to cell surface NKG2D ligand regulation comes from several distinct sources. First, the observation that proliferating primary fibroblasts grown *in vitro* express NKG2D ligands at the cell surface and that this expression is inhibited with confluence or contact inhibition was originally reported in 2006 (85). We have also made this observation in our laboratory. However, the idea of proliferation and contact inhibition in this sense is somewhat vague, and therefore not very useful.

Second, and in contrast, the induction of primary T cell proliferation is more clear cut, beginning at a specific time, in response to a defined stimulus. And NKG2D ligands are induced at the cell surface and mRNA levels in this context (81, 83, 134). Again however, similar to previous situations, the mechanism underlying this effect on NKG2D ligand expression is unclear.

Third, evidence specifically of growth factor-induced NKG2D ligand induction comes from two distinct studies in cell lines. Okita *et al* found that in breast cancer cell lines, signalling

through the growth factor receptor-tyrosine kinase, her3, lead to induction of cell surface MICA and MICB with significant increases in NK cell mediated cytotoxicity (135). Lastly, chronic myeloid leukaemia is a slowly progressive malignancy of myeloid lineage leukocytes, defined by the presence of a chromosomal translocation resulting in the fusion of two genes, the 'break point cluster' gene, *bcr*, and the oncogenic tyrosine kinase gene, *abl*. This *bcr-abl* fusion gene results in the synthesis of a constitutively active tyrosine kinase activity, driving the CML malignancy. An inhibitor of the bcr-abl tyrosine kinase activity, imatinib, has revolutionised the clinical management of CML. Furthermore, NKG2D ligand-expressing CML cells stop expression of NKG2D ligands in the presence of imatinib, resulting in a reduction in cytotoxicity following treatment (97).

To summarise these experiments, it appears that inducing cells to proliferate, including primary cells, leads to cell surface NKG2D ligand expression, and blocking a malignant cell's primary driver of proliferation limits the expression of NKG2D ligands. Again, however, it isn't clear from the literature at which level of gene expression this effect occurs.

DNA repair pathways. That DNA repair pathways should control NKG2D ligand expression is an attractive hypothesis; NKG2D ligands are most reproducibly expressed on cancer cells, proliferating cells, and virus-infected cells, settings where activation of the DNA damage response is probably likely. An increasing number of publications suggest that NKG2D ligand expression is dependent on signalling via the ATM/ATR kinases, activated by the DNA damage response. On careful inspection, the evidence isn't overwhelmingly conclusive.

Three more recent publications treat the subject relatively lightly. Eissmann *et al* find that the ATM/ATR inhibitor caffeine, and the ATM-specific inhibitor KU55933 both limit MICA protein production, without altering mRNA production, in human macrophages exposed to TLR4 agonist, lipopolysaccharide, concluding a post-transcriptional role (78). Ward *et al* describe the ability of caffeine, but not KU55933 to prevent the expression of MICA protein in human $CD4^+$ T cells infected with CMV(71). Similarly, Cerboni *et al* demonstrate

that caffeine reduces MICA proteins expression on PHA-stimulated human T cell(83).

Gasser *et al* addressed this issue in more detail, working on mouse fibroblasts. They noticed that the antibiotic, aphidicolin, which inhibits DNA replication, lead to cell surface NKG2D ligand induction (100), measured by flow cytometry on NKG2D tetramer-stained cells. Caffeine prevented this response to aphidicolin, as did siRNA targeting *ATR*, and induction of a conditional mouse *ATR* knockout model.

In summary, there is convincing evidence that caffeine can limit NKG2D ligand cell surface expression in a range of situations, and aphidicolin induced NKG2D ligand expression is dependent on ATM/ATR signalling. However, it still isn't clear that the DNA damage response or ATM/ATR signalling are central, or even involved in physiological NKG2D ligand induction. The level at which these effects are active in affecting NKG2D ligand expression is also unclear.

1.2.6.3 Peri-transcriptional regulation

A handful of publications provide evidence regarding the transcriptional machinery that governs, or is capable of influencing NKG2D ligand expression. The core promoter and proximal promoter regulatory DNA elements of the MICA and MICB genes were described in 2007 by Venkataraman *et al* (136). They also demonstrated evidence of heat shock factor 1 (HSF1) binding to the MICB promoter, and carried out luciferase assays which suggested that infection of primary fibroblasts with CMV increased transcription from a DNA element encoding the MICA and MICB proximal and core promotors.

In 2007 Rodrìguez-Rodero *et al* demonstrated binding of the transcription factors Sp1, Sp3, Sp4 and NF-Y to the MICA and MICB promoters through a combination of EMSA and super-shift assays. Similarly, the transcription factor NF-κB has been found to bind and affect transcription at the MICA proximal promoter (137). Other transcription factors, including Stat3 (MICA down-regulation; (138)) and Hif1-α (increased MICA and MICB expression; (139)) have been functionally implicated in NKG2D ligand regulation.

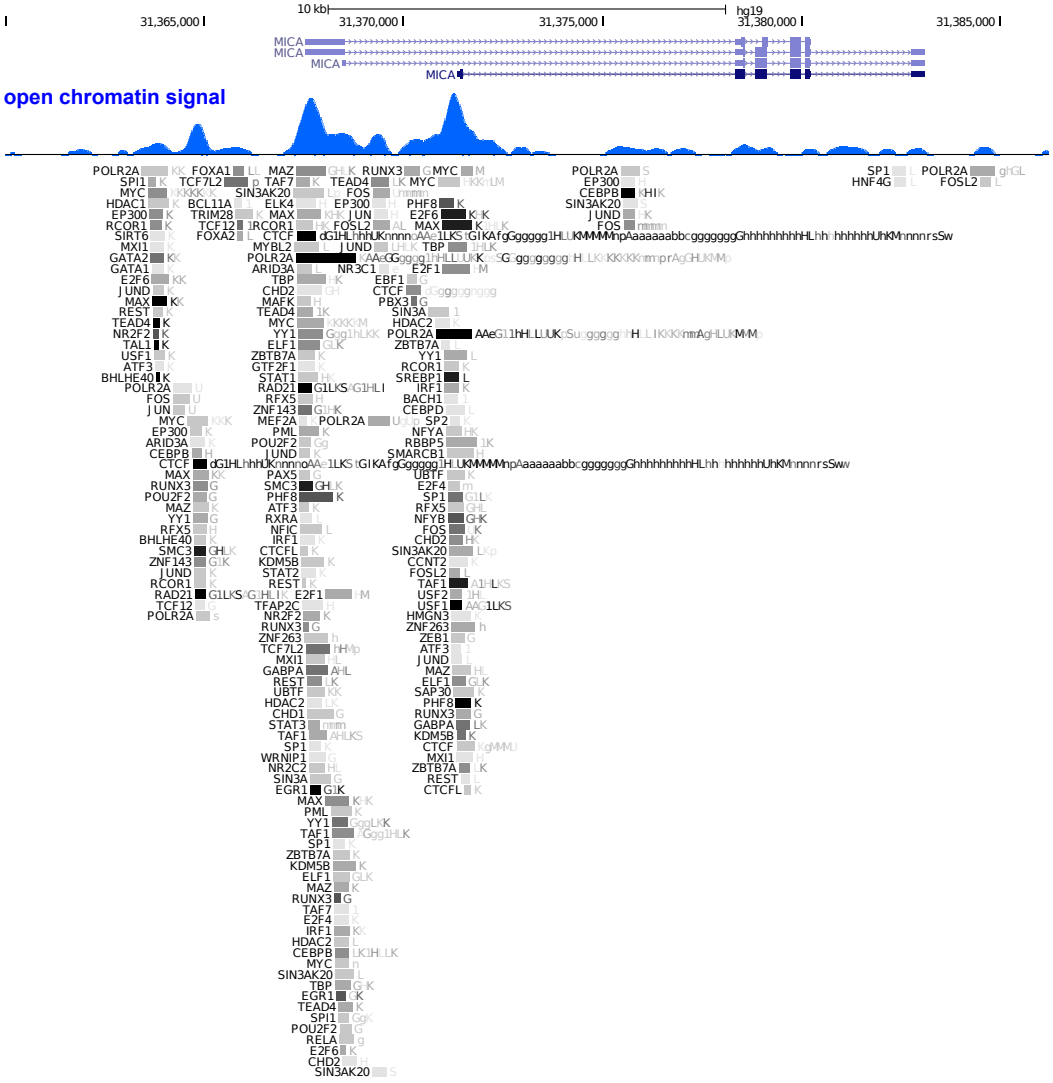

Figure 1.6: ENCODE transcription factor ChIP-seq at the MICA promoter. Transcription factor binding sites identified by ChIP-seq by the ENCODE consortium surrounding the MICA proximal and core promoters. The data is displayed on an image from the UCSC genome browser. The open chromatin signal is DNaseI-seq from MCF7 cells, and transcription factor ChIP-seq data is from many different cell types.

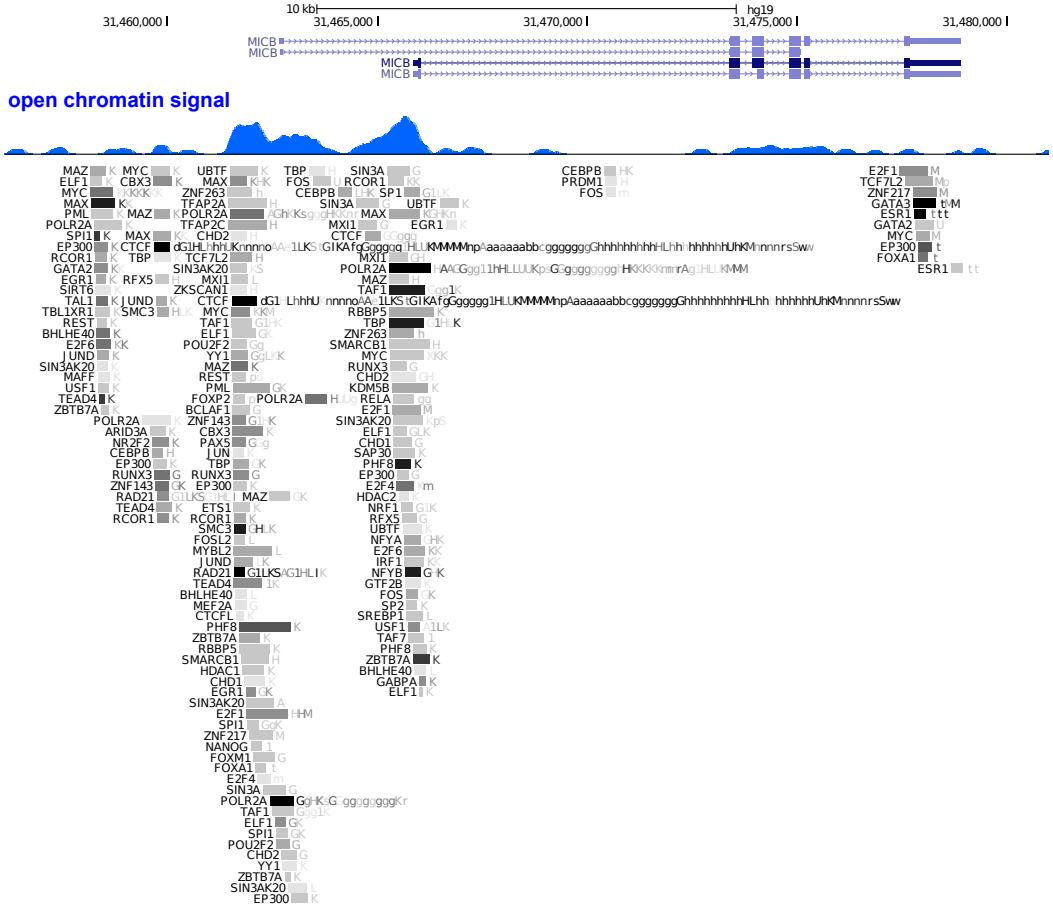

Figure 1.7: ENCODE transcription factor ChIP-seq at the MICB promoter. ENCODE consortium transcription factor ChIP-seq data surrounding the MICB core and proximal promoters displayed on the UCSC genome browser. The open chromatin signal is shown is DNaseI-seq from MCF7 cells The transcription factor ChIP-seq data is from many different cell types.

While these studies are certainly useful, whether any of these factors are relevant to physiological regulation of NKG2D ligand expression is unclear. Little work has been carried out on promotor architecture of the *ULBP* genes. In addition, transcription factor ChIP-seq, a high throughput method for identifying genome-wide binding sites of specific transcription factors indicate that transcriptional influences on MICA **(Fig 1.6)** and MICB **(Fig 1.7)** expression may be more complex than indicated by the traditional studies described above. Transcription factor ChIP-seq signals at the ULBP locus are equally complex **(Fig 1.8)**.

1.2.6.4 Post-transcriptional regulation

Potential mechanisms of post-transcriptional gene regulation include changes in mRNA stability or degradation, rates of translation, protein stability and protein transport. Several post-transcriptional regulatory processes affecting NKG2D ligand expression have been described to date.

miRNA. Evidence for regulation of NKG2D ligand expression by endogenous microRNAs is increasing. miRNAs are a group of non-protein-coding genes with a sequence-specific ability to bind to target sequences on expressed mRNAs (typically located in the 3'UTR), with the outcome of suppressing translation or promoting degradation of the target mRNA (reviewed by Bartel (140)). Stern-Ginossar *et al* have described the existence of a range of endogenous human microRNAs capable of controlling both MICA and MICB protein expression (141). Eissmann *et al* have also demonstrated differential expression of some of these microRNAs in a macrophage-LPS (TLR4) model of MICA induction (78). Evidence implicating specific miRNAs in the regulation of ULBP2 expression has also been reported (142).

The proteasome. The proteasome is a cytoplasmic structure responsible for the degradation of proteins that have been tagged for destruction. Following successful translation, the final amount of functional protein produced is also dependent on the rate of protein degradation. It is plausible that NKG2D ligands are translated constitutively, and immediately degraded. In this case, a reduction in the rate of degradation would produce functional

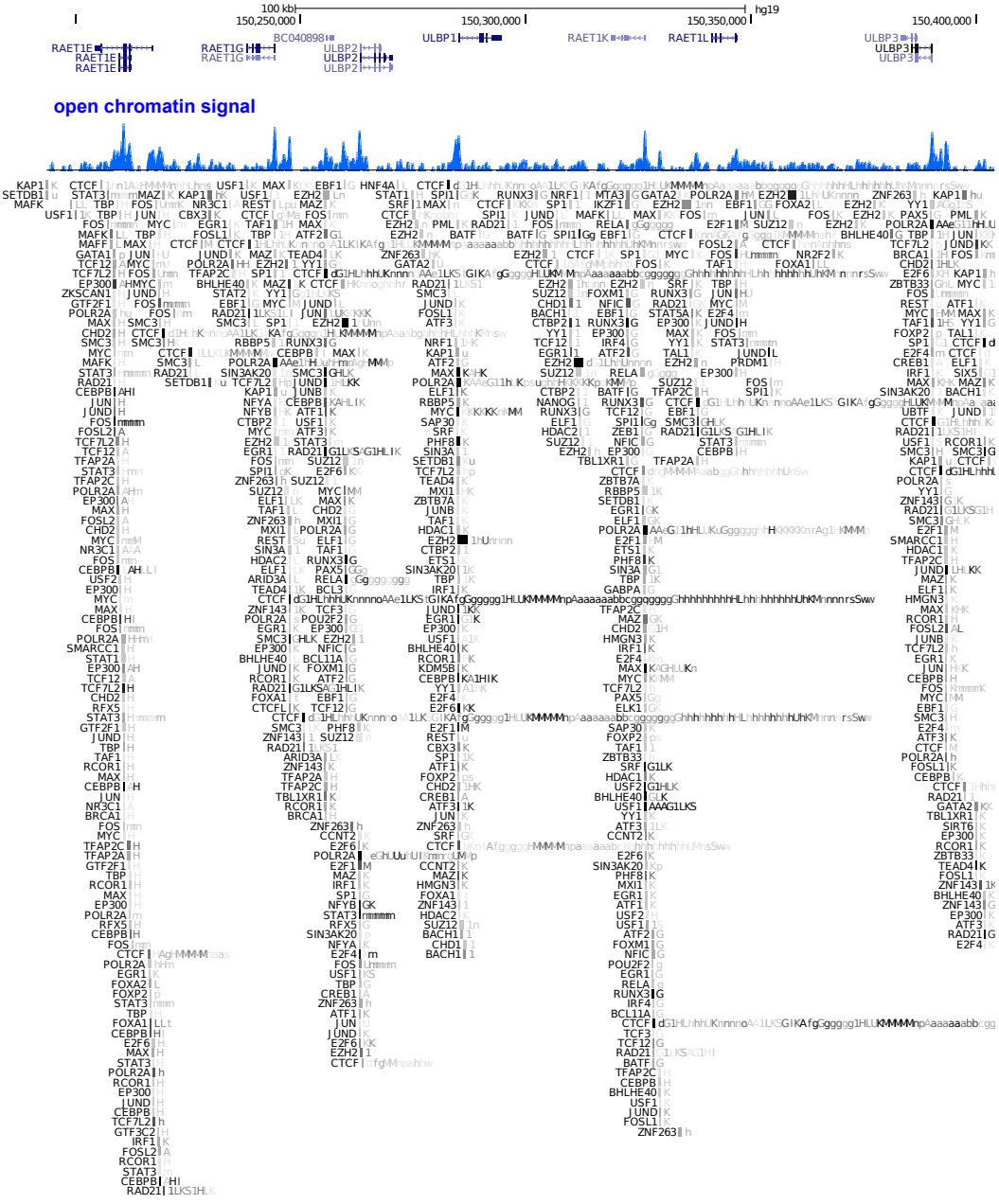

Figure 1.8: ENCODE transcription factor ChIP-seq at the ULBP locus. ENCODE consortium transcription factor ChIP-seq data covering the ULBP locus on chromosome 6, accessed via the UCSC genome browser. The open chromatin signal is shown is DNaseI-seq from MCF7 cells The transcription factor ChIP-seq data is from many different cell types.

expression. Two studies in cell lines have demonstrated increased NKG2D ligand expression in response to treatment with proteasome inhibitors. These studies are also complicated by the observation that proteasome inhibitors also increase mRNA transcript levels (143, 144). The absence of evidence for such mechanisms in a primary cell setting makes this data less convincing as a candidate mechanism for physiological regulation.

Cell surface proteolytic cleavage. It has been clearly established that in the setting of NKG2D ligand expression in tumours, various proteases reduce or eliminate surface expression by cleavage of MICA directly at the tumour cell surface (96, 145, 146). This is consistent with data showing that patients with malignant tumours have higher levels of soluble serum MICA than healthy individuals (147). It has not been demonstrated that the use of protease inhibitors on primary cells induces MICA expression, and the evidence suggests that healthy individuals do not have high soluble serum MICA/NKG2D ligand levels. Together this indicates that tumour shedding of MICA is represents a likely tumour 'immune evasion' mechanism, as opposed to a mechanism of physiological regulation of expression.

Glycosylation. A recent publication suggested that the glycosylation of MICA controlled cell surface expression of MICA in Jurkat cells, by altering transport of MICA to the cell surface. Inhibiting the glycosylation of MICA reduced cell surface expression (148). This result is consistent with earlier work showing that the complex cell adhesion molecule CEACAM1 can influence cell surface NKG2D ligand expression on mouse cells by effectively deglycosylating the ligands (149). This is a plausible mechanism of regulation of expression although further work is required to demonstrate its physiological relevance.

1.2.6.5 Gene regulation summary

In spite of all the studies outlined above aimed at identifying regulatory mechanism of NKG2D ligand expression, it remains unclear how NKG2D ligand expression is induced in physiological settings. Most studies approach this question by studying cell lines, where expression of NKG2D ligands is already 'on'. The molecular events responsible for the induction have

happened. At this point in the process, all the steps in the chain of expression are active, and the opportunity to isolate the mechanistic regulatory events has passed. This is underlined by the clear uncertainty as to whether regulation in physiological situations is pre-transcriptional, peri-transcriptional, or post-transcriptional.

1.3 Cell metabolism

1.3.1 What is 'Warburg metabolism'?

In the 1920s, Otto Warburg (150), who was awarded the Nobel Prize in Physiology or Medicine in 1931 for his work, described a distinctive metabolic phenotype that he associated with rapidly growing tumours (151). Warburg noticed that these tumours consumed greater amounts of glucose, produced higher quantities of lactic acid, and exhibited reduced rates of oxidative metabolism of glucose in the tricarboxylic acid (TCA) cycle in the mitochondrion, compared to "normal" tissues **(see Fig 1.3)**. While metabolism did not feature in Hanahan and Weinberg's survey, 'The Hallmarks of Cancer', published in 2000 (152), metabolism and Warburg were described as an 'emerging' hallmark of cancer in the 2011 update (153).

This form of metabolism, variably described in the literature as 'Warburg metabolism', the 'Warburg effect', 'nucleogenic metabolism' or 'aerobic glycolysis', has been the source of extensive research and debate over the past century. However, a clear understanding of its origins and regulation is still beyond our reach. Below, I will describe our rationale for pursuing a link between Warburg metabolism and NKG2D ligand expression, explore existing literature on Warburg metabolism and its regulation as a basis for my experiments, and describe why perhaps 'high-energy metabolism' may be a more appropriate term to describe this metabolic phenotype.

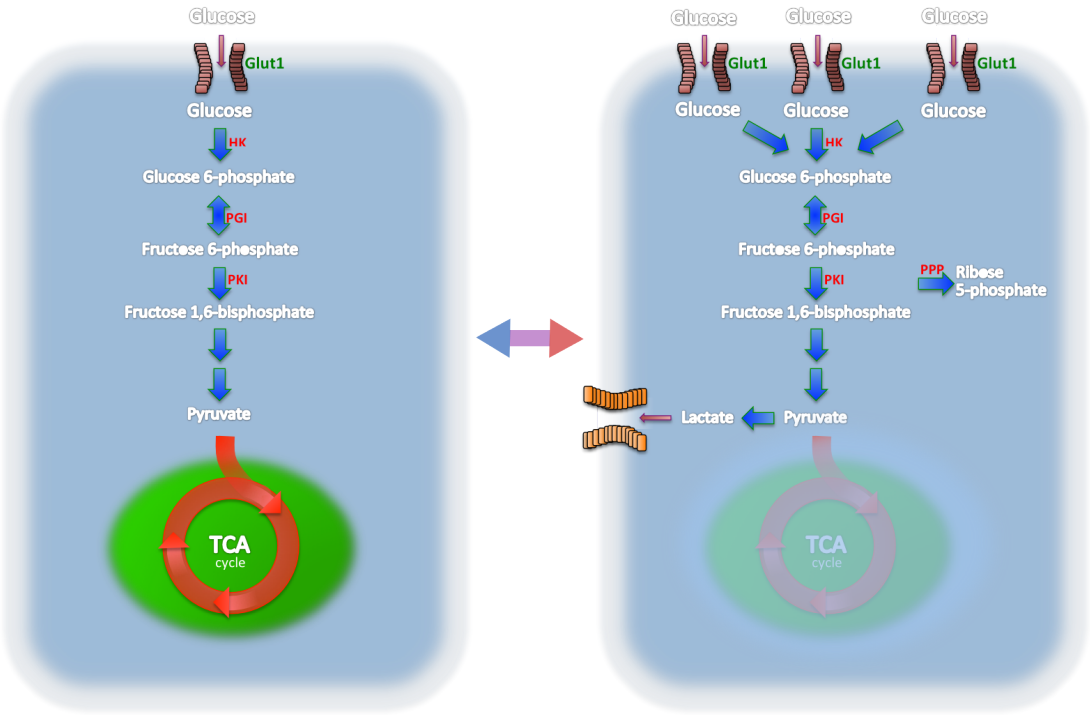

Figure 1.9: Warburg metabolism. The modern understanding of 'Warburg metabolism' has changed little since Warburg's original description. Quiescent or 'normal' cells (left) use small quantities of glucose from their environment. This glucose is metabolised to pyruvate in the cytoplasm. Pyruvate is transferred to the mitochondrion, where it is metabolised oxidatively to generate ATP from ADP via the electron transport chain. This generates ~36 molecules of ATP per molecule of glucose. In 'Warburg metabolism' (right), cells take up much greater amounts of glucose from their surroundings. The amount of pyruvate passing to the mitochondrion is reduced as is tricarboxylic acid cycle activity. Instead, pyruvate is converted to lactate, and extruded from the cell. Metabolism of glucose in this way leads to the generation of ~2 molecules of ATP per molecule of glucose. The increased 'flux' of glucose through this pathway is thought to result in the increased availability of biosynthetic intermediates for cellular biosynthetic pathways, such as the pentose phosphate pathway. TCA - Tricarboxylic acid; PPP - Pentose phosphate pathway; HK - Hexokinase; PGI - Phosphoglucose isomerase; PKI - Phosphofructokinase.

1.3.2 NKG2D ligands and Warburg metabolism

Our interest in exploring the relationship between metabolism and NKG2D ligand expression began with unpublished data from our laboratory, showing that culturing cell lines in medium containing different concentrations of glucose affects NKG2D ligand expression. While this simple observation is of interest in its own right, in a setting where little is known of NKG2D ligand regulatory mechanisms, reflection on existing literature suggests this result may have a greater importance **(Table 1.3)**.

As described previously, NKG2D ligand expression has been documented in several distinct settings, including many cancers, activated T cells, proliferating fibroblasts, viral infections (including CMV infection), toll-like receptor ligation, and in gastrointestinal epithelium. Following Warburg's first description of altered metabolic profiles in malignant cells in 1924, Munyon and Merchant showed a similar patten of glucose usage and lactate production in proliferating fibroblasts in 1959 (154). In 1976, Wang *et al* reported that concanavalin A stimulated mouse lymphocytes exhibit a discreet burst of glucose utilisation and lactate production, which closely tracks new RNA and DNA synthesis in these cells; the lymphocytes resume their normal metabolic patterns ∼80 hours after concanavalin A stimulation.

In 1984, Landini (155) published a short report demonstrating that glucose uptake was increased following infection of confluent human fibroblasts with Towne strain human cytomegalovirus (hCMV). Recent studies have confirmed this, showing *SLC2A4* (glut4) up regulation, increased glucose consumption and lactate production (156, 157). In 2010, Krawczyk *et al* (158) demonstrated increased aerobic glycolysis and decreased oxidative phosphorylation in response to lipopolysaccharide (LPS) stimulation of toll-like receptor 4 on mouse bone-marrow derived dendritic cells. Conversely, inhibition of glycolysis in these cells impaired activation.

Increasingly, there is improving evidence to support a transition to a metabolic phenotype closely resembling that described by Warburg in the early 1920s, in cells in which NKG2D

Table 1.3: Studies demonstrating NKG2D ligand expression and metabolic changes.

Setting	NKG2D references	Warburg references
Cancer	Duan *et al* (87); McGilvray *et al* (86, 90); Li *et al* (89); Vetter *et al* (92); Wu *et al* (94)	Warburg *et al* (159); reviewed by Cairns *et al* (160)
Fibroblast proliferation	Zou *et al* (85)	Munyon and Merchant (154)
Lymphocyte proliferation	Cerboni *et al* (83)	Wang *et al* (161)
Toll-like receptor ligation	Ebihara *et al* (129)	Krawcyzk *et al* (158)
Cytomegalovirus infection	Groh *et al* (60); Rölle *et al* (67); Eagle *et al* (66)	Landini (155); Yu *et al* (156)

ligand expression is known to occur. With improving metabolomics methodology, it will become easier to examine these cellular metabolic changes in greater detail in the future. Currently, there are no situations described in which NKG2D ligand induction has been demonstrated, and these metabolic changes have been shown not to occur. This striking relationship between NKG2D ligand induction and metabolic transition to 'aerobic glycolysis' provides a clear impetus to answer the question, 'Why does glucose availability alter cell surface NKG2D ligand expression levels?'.

1.3.3 Warburg metabolism - historical origins

It might be said that most revolutions in scientific thinking arise from innovations in methodology that allow new perspectives to be explored. This was the case when Warburg described oxidative metabolism of glucose in cancer tissues in the early 1920s. Warburg developed a new manometric method of measuring oxygen consumption by placing thin tumour slices on a glass slide within a sealed chamber attached to a manometer (for description see Koppenol *et al* (150)). Measurements of pressure changes due to oxygen consumption or carbon dioxide production were used to calculate rates of oxidative glucose metabolism and lactate production per dry weight of tissue examined. Warburg's key unexpected observations were that even in the presence of oxygen, cancer tissue produced lactic acid, and that a much greater amount of glucose was consumed per gram of tissue compared to non-malignant tissue (151). This was contrary to contemporary thought, in that the established 'Pasteur effect' proposed that, in the presence of adequate oxygen, glucose is metabolised oxidatively in the mitochondrion, and no lactate is produced. This experiment began a century of research that still has

not managed to confidently explain Warburg's observations.

In 1925, Cori and Cori (162) attempted to replicate these findings *in vivo*. To do this, they implanted Rous chicken sarcoma cells unilaterally into the wing musculature of Plymouth Rock chickens, and allowed the tumours to grow until they reached the size of a "hen's egg". Venous blood was taken from both wings, and glucose and lactate concentrations in the blood were measured. They found that venous blood from the tumour bearing wing had lower glucose and higher lactate than venous blood from the healthy wing.

Warburg *et al* carried out similar *in vivo* experiments, published in 1927, using implanted Jensen rat sarcoma cells (159). Instead of comparing glucose and lactate concentrations in venous blood from healthy or tumour-bearing sites, they compared glucose and lactate concentrations in arterial blood supplying healthy or malignant tissue, with concentrations in the venous blood draining those tissues. They found a much greater drop in glucose concentrations and increase in lactate concentrations across tumour sites compared to healthy tissue sites.

While the existence of this general metabolic phenotype in cancer cells has been repeatedly confirmed since these early experiments, Warburg made two major assumptions about this type of metabolism, which he maintained into the 1950s (163), and still influence thinking today: first, that aerobic glycolysis is restricted to malignant tissue and second, that it is an irreversible process.

As early as the 1920s, some doubt was placed on these assumptions, and attempts to use these metabolic profiles as a marker of malignancy were not particularly successful. In 1925, Murphy and Hawkins (164) examined metabolic phenotypes using Warburg's method in a range of settings, including rat placenta, liver, spleen and implanted Flexner-Jobling carcinoma, implanted chicken sarcoma (Chicken tumour 1 and Chicken tumour 9), and tumours developed in a strain of mice that has a "high tumour instance". While there was a trend towards increased levels of aerobic glycolysis in tumour samples, some spleen and placenta samples fell within the malignant 'group', while variability was noted in tumour samples,

with some tumours behaving metabolically as healthy tissue.

Crabtree examined metabolism in fowlpox lesions in pigeons, and vaccine lesions in rabbits, again using Warburg's methods for measurement of aerobic glycolysis and oxidative glucose metabolism. Crabtree also found that metabolic profiles in some of these tissues matched those found in malignant tissue.

In 1930, Jackson *et al* (165) examined a case series of 71 human lymph node sections from patients with histologically established diagnoses, in the hope of discovering whether lymphomas were of malignant of infectious origin. The 71 lymph nodes fell into seven disease categories: healthy, benign tumours, non-hodgkin lymphoma, hodgkin's lymphoma, tuberculosis, sarcoma and carcinoma. They found that while the mean rates of aerobic glycolysis in healthy nodes, benign tumours, and sarcomas were consistently low, tuberculous, and lymphomatous nodes had mean rates in the 'malignant group', while lymph nodes containing carcinomas had the highest rates of aerobic glycolysis. Nevertheless, in all groups, variation was wide, suggesting that the specificity of the rate of aerobic glycolysis for infectious or malignant tissue is low.

In summary, Warburg's key contribution to metabolic thinking was that some cells, even in the presence of oxygen, produce large quantities of lactate ('aerobic glycolysis'). Attempts to derive metabolic equations to define metabolic profiles as specific to disease processes however, failed. It was also clear at the end of this early period of intensive research on the topic, that aerobic glycolysis wasn't unique to malignant tissue, also being evident in some infected and healthy tissue types.

1.3.4 Current understanding of Warburg metabolism

Following the initial flurry of publications relating to aerobic glycolysis in the 1920s, a resurgence of research interest in the topic developed in the 1980s, and has peaked over the past five years **(Fig 1.10)**. In particular, interest in aerobic glycolysis has centred on its role in cancer cell metabolism. Much of this research is guided by one of Warburg's original aims:

to use the tumour's 'dependence' on aerobic glycolysis to target anti-cancer treatments. A side-effect of this line of research has been to reinforce the general presumption that aerobic glycolysis is an irreversible process, and a distinctive characteristic of malignancy. This has also largely limited our understanding of these metabolic changes to those changes found in genetically and metabolically heterogeneous cancer cell populations. Nevertheless, below is described the evidence underlying two recurrent themes in the research aimed at identifying the drivers of 'Warburg metabolism': defective mitochondrial oxidative phosphorylation, and altered glycolytic enzyme function.

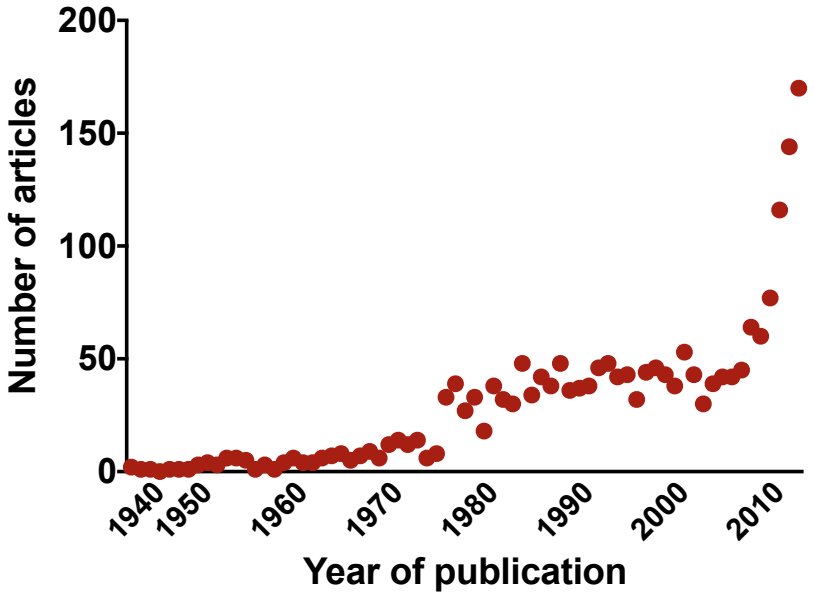

Figure 1.10: PubMed trends in 'aerobic glycolysis' Research. The recent resurgence in interest in aerobic glycolysis is indicated by a spike in the number of publications on the topic registered with NCBI's PubMed literature database.

1.3.4.1 Defective mitochondrial oxidative phosphorylation

Defects in aerobic oxidative phosphorylation, which result in increased glycolysis, underpin the Warburg effect, and have a primary causal effect in the development of cancer. This was the view of Warburg himself, who supported this argument principally with the observations that a range of "respiratory poisons", including hypoxia itself, arsenic oxide, hydrogen sulphide, and thioacetamide, each both inhibit the tricarboxylic acid cycle and are carcinogenic.

Warburg's idea that hypoxia leads to irreversible defects in oxidative phosphorylation and eventually cancer informed the work of Goldblatt and Cameron (166), who demonstrated this principle experimentally. They isolated fibroblasts from rat myocardium, and exposed them to intermittent hypoxia over several months. This eventually generated sub clones of fibroblasts exhibiting aerobic glycolysis, that were described as histologically malignant cells by blinded histopathologists, and displayed malignant behaviour when transplanted into rats.

Further evidence to support the role of impaired oxidative phosphorylation in tumour development was developed in the 1970s, beginning with a paper by Edwards *et al* (167). They noticed that cells in the carotid body (a collection of oxygen sensitive cells near the carotid artery bifurcation) underwent hypertrophy and hyperplasia in animals living at high altitude, compared to those living at low altitude. Arias-Stella and Valcarcel demonstrated a similar disparity in carotid body size in humans living at high altitude 1976 (168). However, none of these studies examined glucose metabolism in these hyperplastic cells.

Direct evidence implicating defective mitochondrial metabolism in tumorigenesis was published in 2000. Baysal *et al* mapped the gene responsible for hereditary paraganglioma to succinate dehydrogenase complex subunit D (*SDHD*). SDHD is an important enzyme in the mitochondrial tricarboxylic acid cycle. Based on the histological similarity with the carotid body tumours found in patients living at high altitude, they hypothesized that SDHD may be a component of the oxygen sensing mechanism in these cells, but did not characterise mutation bearing cells metabolically. Subsequently, cancer-causing mutations in other succinate dehydrogenase complex subunit genes, *SDHB* (169) and *SDHC* (170) were described, again without metabolic characterisation.

In 2002, Tomlinson *et al* (171) showed that germ-line mutations in the fumarate hydratase gene were responsible for the hereditary leiomyomatosis and renal cell carcinoma (HLRCC) syndrome. Cellular models of fumarate hydratase deficiency have since been metabolically characterised. Ashrafian *et al* (172) explored the genetic and metabolic impact of fumarate hydratase (FH) deficiency on mouse embryonic fibroblasts derived from a *FH* conditional

knockout mouse. They found enhanced expression of *hif1-α* target genes, glycolytic genes, and changes in metabolite concentrations, consistent with aerobic glycolysis, including increased lactate production, and nucleotide concentration. The *FH* deficient cells also exhibited spontaneous aneuploidy.

Yang *et al* (173) carried out similar experiments in 2012. They generated a homozygous *FH* deficient cell line from a renal lesion excised from a 26 year old patient with HLRCC. Analysing mitochondrial microarrays (including genomic and mitochondrial DNA mitochondrion associated genes) they found increased expression of glycolytic genes, with decreased expression of several TCA cycle genes. This was also associated with decreased oxygen consumption in the HLRCC cell line.

This evidence combined suggests that defects in mitochondrial function can lead to aerobic glycolysis, and perhaps even an increased propensity for malignant transformation, as Warburg had proposed. However, while gene expression data is informative, it fails to account for changes in gene function, which ultimately determine the shifts in metabolic patterns that comprise the 'Warburg effect'. A second challenge is in defining what "normal" cellular metabolism looks like. The changes described by Yang *et al* (173) in their *FH* deficient cell line was in comparison with an ATCC derived HPV16 transformed human kidney cell line. It is quite possible that these transformed cells also exhibit aerobic glycolysis. Similarly, the proliferating wild type mouse embryonic fibroblasts (MEFs) used by Ashrafian *et al* (172) as a control for their *FH* deficient MEF experiments also probably exhibit an altered metabolic profile compared to quiescent cells. This fundamental problem is difficult to overcome.

1.3.4.2 Altered glycolytic enzyme function

While reasonably strong evidence now shows that defects in mitochondrial oxidative phosphorylation have the capacity to produce cells with an increased capacity for aerobic glycolysis, there is no shortage of experimenters who suggest that primary defects or alterations of glycolytic enzyme function is primarily responsible for the Warburg phenotype, and research

interest in these enzymes has intensified over the past 10 years **(Fig 1.11)**.

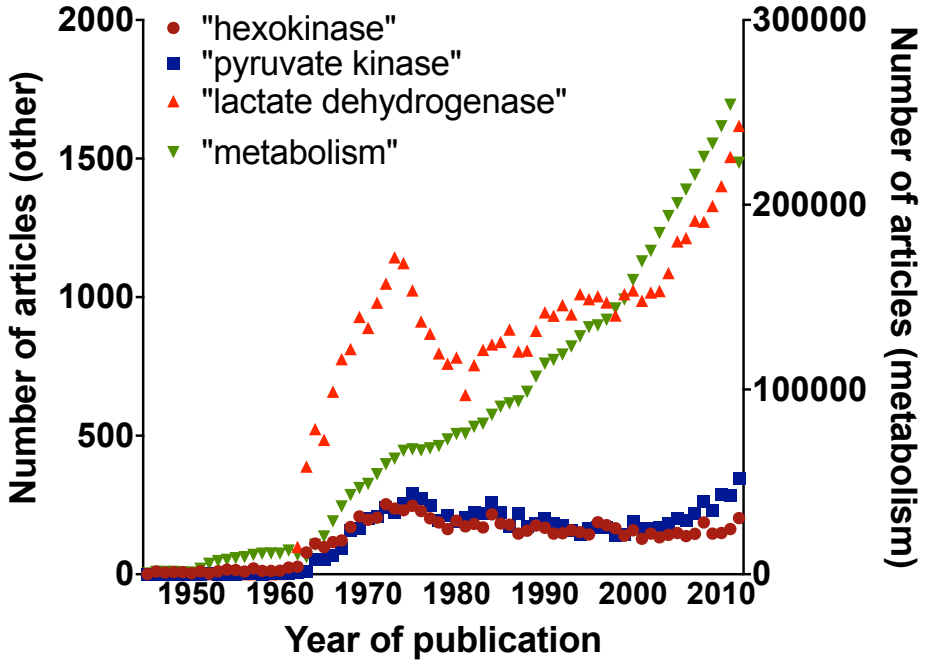

Figure 1.11: PubMed trends in glycolysis research. Research interest in key glycolytic enzymes was gauged by searching the NCBI's PubMed publication database for the terms indicated. While the number of research publications found with the keyword "metabolism" has increased steadily since the 1950s, the number of publications relating to individual glycolytic enzymes remained stable until 2000, and has been increasing over the past 10 years.

Hexokinase. In a review article published in 2009, Mathupala *et al* (174) assert that the application of ^{18}FDG-PET scanning to cancer imaging demonstrates

> *"...how a simple but pivotal discovery by basic biomedical scientists* [that some hexokinase isoforms metabolise glucose faster than others (175, 176)] *working independently of physicians can lead to clinical applications of profound utility to the entire world."*

While it is improbable that the absence of their earlier cited work would have impeded the development of ^{18}FDG-PET scanning, there is good evidence to suggest that changing hexokinase isoforms can influence the rate of glycolysis.

There are four hexokinase genes in humans: hexokinases 1-3 and glucokinase (hexokinase 4). Hexokinase carries out a key step in glucose metabolism: the phosphorylation of glucose upon transport into the cell. Of these four isozymes, it is most frequently proposed that

hexokinase 2 is critical for Warburg metabolism. In mice, hexokinase 2 mRNA is normally strongly expressed in myocardium, skeletal muscle and adipose tissue, and to a lesser extent in spleen, ovaries/testes and lung (177). Hexokinase 2 knockout is embryonic lethal in mice (178), suggesting a role in embryonic development also. In 1977, Bustamente and Pederson (175) demonstrated that hexokinase activity in a hepatoma cell line was ∼20 times higher than that in normal liver, and that 50% of this hexokinase activity was present in the mitochondrial fraction of cell line lysates. Following the isolation of human hexokinase 2 in 1993 (179), it became clear that hexokinase 2 was responsible for the mitochondrion-associated high hexokinase activity observed in previous experiments (reviewed by Mathupala *et al* (174)).

Evidence also exists that hexokinase 2 is 'aberrantly' expressed in human cancers. This evidence takes the form of small case series assessing the expression of hexokinase 2 in glioblastoma multiforme (GBM; 8 of 12 GBM, 0 of 3 normal brain) (180), cervical carcinoma (35 of 45 carcinoma; 8 of 12 control) (181) and gastric carcinoma (43 out of 257 consecutive cases) (182). Recent cellular studies have shown how expression of hexokinase 2 contributes to generating a Warburg metabolic profile. Wolf *et al* (180) expressed hexokinase 2 in the hexokinase 2-negative GBM cell line, U343, and found a 10% increase in lactate production, along with an increase in proliferation measured by BrdU incorporation. Similarly, shRNA mediated knockdown of hexokinase 2 in the hexokinase positive GBM cell line U87 caused decreased lactate production and increased oxygen consumption.

In summary, while it is clear that the hexokinase 2 isoform is expressed aberrantly in some cancers, and phosphorylates glucose more rapidly, it is also clear that it isn't necessary for either Warburg metabolism, or the development of cancer. Its role in ^{18}FDG-PET uptake in cancer is also debatable (reviewed by Smith (183)).

Pyruvate kinase. Humans have two pyruvate kinase genes, pyruvate kinase liver and red blood cell (*PKLR*), and pyruvate kinase muscle (*PKM*). The PKM gene has two expressed splice isoforms, called *PKM1* and *PKM2*. The existence of two PKM isoforms, isolated from rat adipose tissue, was first reported by Pogson in 1968 (184), and these isoforms were

functionally characterised in 1972 by Imamura *et al* (185). From an early stage it was found that pyruvate kinase exists primarily as a homotetramer (186), although heterotetrameric forms are known to exist and function (187). While today it is clear that homotetrameric forms and heterotetrameric forms of pyruvate kinase derive from just two genes, in the early days of pyruvate kinase research, nomenclature and function of different pyruvate kinase forms was somewhat obscured. Nevertheless, work by Marie *et al* (188) in 1976 demonstrates that the *PKM2* splice isoform is widely distributed in healthy adult human tissue, including expression in the spleen, testes, uterus, adrenal glands, stomach, prostate and lung tissues.

A study of rat hepatomas in 1965 by Shonk *et al* (189) suggested that rapidly growing tumours had much higher pyruvate kinase activity than slowly growing tumours. In 1968, Lo *et al* (190) showed that rat liver tumours express a different isoform of pyruvate kinase compared to healthy liver tissue. These isoform differences were resolved in a series of papers by Imamura *et al* in 1972 (185, 191). Again using rat tissues, they found that the M2 isoform was strongly expressed in rat tumours, and that the K_m of purified PKM1, PKM2 and PKL for phosphoenolpyruvate were 7.5 µM, 40 µM and 83 µM respectively.

While these early studies suggested that the excess expression of the M2 pyruvate kinase isoform in tumours facilitated the rapid consumption of glucose, more recent studies do not support that view. There is reasonable evidence that the M2 isoform is over-represented in human cancers, including colorectal cancer (192), oesophageal cancer (193), and glioma (194). However, studies carried out over the past 40 years suggest that the PKM2 isoform may confer a proliferative advantage on proliferating and cancer cells because it is *less* active *in vivo*. As early as 1972, it was noted that the activity of PKM2 and PKL are enhanced allosterically by fructose 1,6-bisphosphate (FBP), while the activity of PKM1 was unaffected (185). In 1989, Kato *et al* showed by gel-filtration chromatography that purified human PKM2 was tetrameric in the presence of FBP, and dimeric in its absence. Assessment of enzymatic activity of purified fractions showed that while the tetrameric form had high activity, the dimeric form had less enzymatic activity (195, 196). In contrast, in a separate study, PKM1

exhibited high enzymatic activity even in its purified monomeric form (197). The sequence of the PKM1 and PKM2 isoforms differs by a single exon, and structural studies demonstrate that these sequence differences are near the FBP binding site (198). The importance of this site for allosteric regulation is also supported by functional studies (199).

The final twist in the PKM2 tale is that several regulatory proteins have been described, which displace FBP and result in the conversion of active tetrameric PKM2 to a less active trimeric or dimeric forms. This is believed to occur in cancer cells, and proliferating cells to support 'Warburg metabolism', by leading to the increased availability of glycolytic intermediates for biosynthetic processes. In 1988, Presek *et al* demonstrated that infection of chicken embryonic cells with a Rous sarcoma virus lead to phosphorylation of PKM2, and that phosphorylated PKM2 had lower enzymatic activity compared to unphosphorylated PKM2 (200). A more detailed recent study has confirmed these earlier findings, demonstrating that phosphorylation of PKM2 at tyrosine^{105} is mediated by FGFR1, inhibiting tetramer formation, and reducing enzymatic activity. A similar mechanism has been proposed by Christofk *et al* (201). They have demonstrated that the interaction of certain peptides containing a phosphotyrosine residue with PKM2 leads to displacement of FBP and reduced PKM2 enzymatic activity and cell growth. Functional studies also suggest that small molecule activators of PKM2 enzymatic activity, or replacement of PKM2 with PKM1 lead to reduced tumour growth in a xenograft model (202).

Combined, these studies suggest that when pyruvate kinase activity is reduced, cells have a proliferative advantage, probably conferred by the increased availability of glycolytic intermediates. This switch becomes possible by changing from the constitutively active PKM1 splice isoform, to the regulatable PKM2 isoform. The enzymatic activity of the PKM2 isoform can be reduced by either phosphorylation or phosphotyrosine binding to the FBP binding site, thereby limiting PKM2 tetramer formation **(Fig 1.12)**. Based on this, some argue that Warburg metabolism is controlled by the presence or absence of the PKM2 isoform, and its enzymatic regulation.

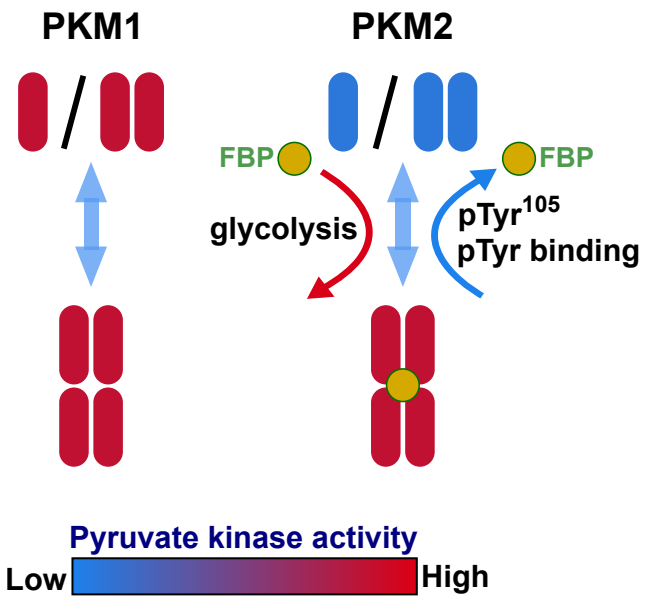

Figure 1.12: Regulation of pyruvate kinase activity. The *PKM2* splice isoform has regulatable pyruvate kinase activity. The PKM protein (left) has high pyruvate kinase activity in both tetrameric and mono- and dimeric forms. The PKM2 isozyme (right) has low enzymatic activity in mono- and dimeric forms. Binding of the glycolytic intermediate fructose 1,6-bisphosphate (FBP) to monomeric PKM2 leads to tetramer formation, and high enzymatic activity. However, phosphorylation of tyrosine^{105} at the FBP binding site, or binding of certain phosphotyrosine containing peptides can displace FBP, leading to dismantling of the tetramer, and decreased pyruvate kinase enzymatic activity. This is thought to increase the availability of glycolytic intermediates upstream of pyruvate, generating a 'biosynthetic milieu'. However, experiments in cell lines suggest that the replacement of PKM2 with PKM1 doesn't completely abolish the Warburg metabolic phenotype. FBP - Fructose 1,6-bisphosphate; pTyr - phosphotyrosine.

However, it is clear that expression of PKM2 in many normal healthy human tissues is widespread (188). Recent proteomic analysis of cancers and matched healthy human tissues suggests that PKM2 expression is as prevalent in healthy tissues as in cancer cells (203). Further, forcing an isozyme switch from PKM2 to PKM1 expression in cell lines only has a modest effect on cell proliferation *in vitro* (204). And this is in a setting where isozyme switched cells do not have the opportunity to genetically adapt to the enzymatic change. On balance, the evidence suggests that the predominance in a cell of dimeric PKM2 with reduced pyruvate kinase activity does facilitate biosynthetic processes and contribute to cell proliferation. However, it is improbable that this is the sole determinant of the Warburg metabolic phenotype.

1.3.5 Aerobic glycolysis - 'High energy metabolism'?

1.3.5.1 Problems with existing models

It is clear that healthy primary human cells display reversible 'Warburg metabolism' in the course of normal, physiological processes (161). This evidence is in direct conflict with Warburg's view expressed 20 years earlier, that this is an irreversible process, driven by defective mitochondrial oxidative phosphorylation (163). Additionally, with some exceptions, the enzymatic changes and isoform switches discussed above, are unlikely to have evolved to support metabolism when cells become malignant. Instead, these are healthy, non-mutated genes whose evolution is more likely explained by roles in physiological cellular metabolism. Aberrant regulation of these enzymes may confer a survival advantage and increase the possibility of cancer, but why do these metabolically efficient enzymes and isozymes exist, if not to support physiological metabolism. In fact the metabolic changes long associated with Warburg metabolism are increasingly thought to be primarily physiological events, not least those which occur in support of T cell activation (reviewed by Fox *et al* (205)). Considering 'Warburg metabolism' as a physiological reversible process, rather than an aberrant characteristic

of tumour cells is more likely to lead to progress in understanding its origins and regulation.

A second puzzling aspect, is the production of lactate by these cells. Warburg ascribed lactate production to defective mitochondrial oxidative phosphorylation. To generate enough energy from glucose, without access to oxidative metabolism in the defective mitochondrion, uptake of excessive quantities of glucose was necessary, and the only metabolic route left to pyruvate was conversion to lactate and elimination from the cell. This view is still maintained by some authors (206). Yet, this opinion isn't universally held, and evidence that at least certain tumours have normal mitochondrial capacity for oxidative phosphorylation is clear ((207); reviewed by Gogvadze *et al* (208)). If these cells are capable of oxidatively metabolising the available pyruvate, why is this pyruvate converted to lactate? If instead, glucose is taken up by the cell to provide glycolytic intermediates for biosynthesis, why is so much glucose taken up that excess lactate is produced? And in the case of pyruvate kinase, which is said to regulate Warburg metabolism by limiting the delivery of pyruvate to the mitochondrion, by providing a hypo-functional, regulatable pyruvate kinase isoform, it is clearly the case that the expressed pyruvate kinase isoforms must have activity, as pyruvate must first be formed before pyruvate can be converted to lactate, and extruded from the cell. These are issues that do not fit comfortably with current ideas of 'Warburg metabolism' and its regulation.

1.3.5.2 A speculative hypothesis of Warburg-NKG2D ligand regulation

An alternative, if speculative, hypothesis is that 'Warburg metabolism' is a physiological cellular process, that represents a high energy, biosynthetic cellular state. Cells can only embark on biosynthetic processes with the support of this high energy state. If the energetic capacity and biosynthetic capacity are viewed as a continuum in this way these metabolic features of 'Warburg metabolism' begin to make more sense. To imagine how this model might work, consider the process of T cell activation. A resting naïve quiescent T cell exists in a 'low energy state'. Glucose uptake into the cell is limited by the type and expression level of its

GLUT transporters, and most of the glucose taken in is metabolised oxidatively in the mitochondrion, leading to basal ATP production. When the T cell is activated, glucose uptake is enhanced, by increased GLUT expression. Some changes in glycolytic enzyme isoforms may occur to enhance glycolytic capacity. This glucose is metabolised by oxidative phosphorylation which increases cellular ATP levels to a 'high energy state'. Once this 'high energy state' is reached, ATP, or some other energetic end product, allosterically limits further increases in the flux of pyruvate through oxidative phosphorylation. Instead, glycolytic intermediates are now available as biosynthetic substrates. In this scenario, Warburg metabolism is a high energy - biosynthetic state, and not an energy poor, or energy neutral biosynthetic state. NKG2D ligands, through some mechanism yet to be described, have evolved to keep these high-energy biosynthetic competent cells checked by the immune system, as uncontrolled high energy biosynthetic capacity represents a threat in a multicellular organism. Viruses then have evolved to tap into this physiological high energy-biosynthetic system, to replicate themselves using the host cellular machinery. Inadvertently, the viruses, in triggering this state, also trigger NKG2D ligand expression, and so in turn evolve NKG2D-evasion mechanisms.

Each of the enzymatic changes, or defects in oxidative phosphorylation described in **Sections 1.3.4.1 & 1.3.4.2** can be viewed in this context as playing a Darwinian role in cancer development. Cancer cells must have biosynthetic capacity to develop as a threat to the host. Metabolic enzymes and enzyme isoforms that can provide this biosynthetic function in distinct physiological settings exist. Aberrant regulation, or mutation leading to the expression of one or more of these enzymes outside its normal physiological setting confers a bioenergetic survival advantage to such a cell, increasing the probability that progenitors of this cell will acquire further mutations or malignant traits. In this cancer-Warburg model, no single enzyme or isoform is essential to the process, every little helps

1.4 Approaches to studying gene regulation

1.4.1 How are genes regulated?

Gene regulation is a complex phenomenon, and like many complex phenomena, it is difficult to unravel. I would argue that it is not yet clear how the expression of any gene is 'regulated'. In order for a gene to be functionally expressed as a protein, a remarkable sequence of events must occur (illustrated schematically, **Fig 1.13**). For example, the gene promotor must be accessible, and the regulatory-sequence appropriate transcription factor(s) must be present in the cell, and available at the right genomic location. Similarly, this must occur at relevant enhancer sites and with appropriate 'enhancing' transcription factors. Appropriate epigenetic changes (which are still poorly understood) may be necessary. RNA polymerase and its machinery must be in place, and in elongating form. RNA transcripts must be appropriately spliced, and stabilised, or at least not degraded for long enough to allow the mRNA to be translated. Following translation, protein stability and turnover also affect the final functional protein levels as can parameters of compartmentalisation or transport within the cell.

So what then is a regulatory mechanism? If a physiological stimulus reproducibly brings about gene expression, something must physically cause this to happen, at the molecular level. At one extreme, stimulation of a non-expressing cell might result in cascading networks of protein kinase signalling, accumulating in the unravelling of an entire segment of chromatin, the activation and translocation of transcription factors, and RNA synthesis machinery, transcription, stabilisation, translation and transport of a protein to its functional location in the cell. At the other extreme, a post translational modification might be made to a constitutively synthesised and degraded protein, which then restricts constitutive proteasomal degradation, resulting in protein expression. Or anything in between. Or any combination of these events.

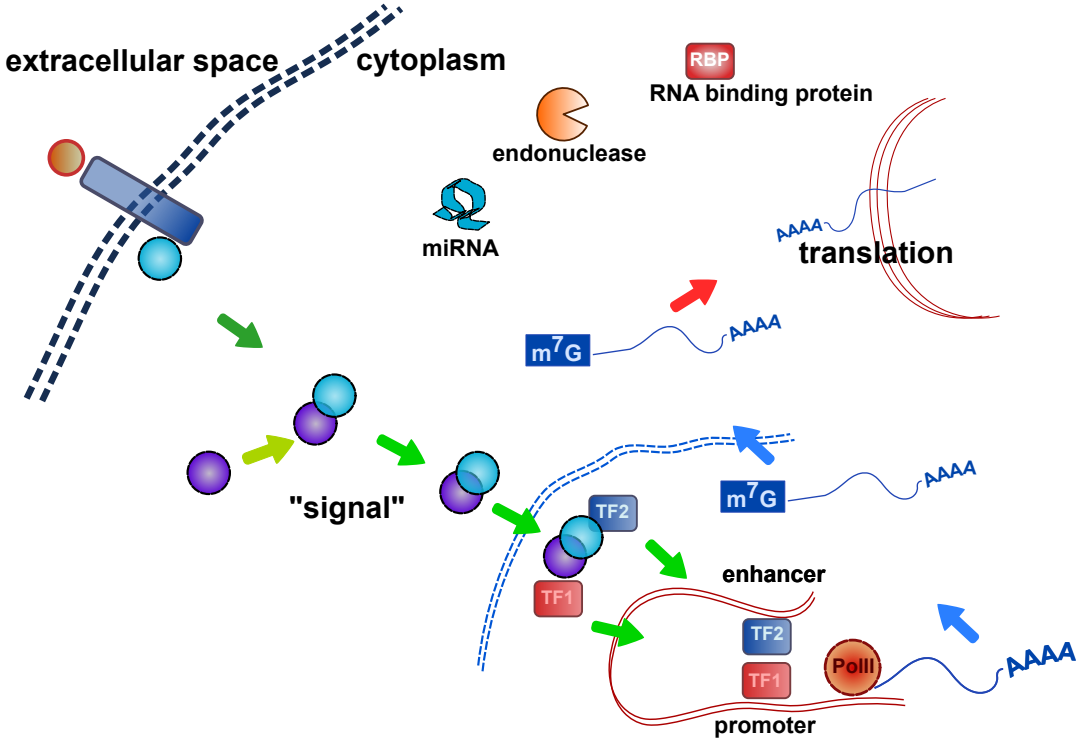

Figure 1.13: Schematic gene regulation. The 'regulation' of gene expression is a complex processes, and requires many different components to come together successfully to bring about functional protein expression. For example, an extracellular stimulus binds a cell surface receptor, activating a cytoplasmic signalling cascade, eventually reaching the nucleus, where appropriate transcription factors are 'activated' to bind in a sequence-specific manner to appropriate enhancers and promoters. The transcribed gene must be spliced, and stabilised sufficiently to be transported from the nucleus, possibly subject to the influence of stabilising RNA binding proteins and destabilising miRNA or endonucleases before translation occurs. Regulation can occur at *any* point in such a pathway. In physiological settings, the resting or 'off' state may be 'completely off', requiring all components of this pathway to be triggered for gene expression. Alternatively, the 'off' state may involve complete transcription of mRNA, that is rapidly degraded in the cytoplasm, where transition to the 'on' state simply requires the expression of an appropriate stabilising RNA binding protein. Or any other combination of events. Identifying these deterministic molecular events, though challenging, should be the ultimate goal of gene regulation studies.

1.4.2 An approach to NKG2D ligand gene regulation

We have chosen a two-sided approach to try to identify NKG2D ligand regulatory mechanisms.

Increased glucose metabolism is a property common to all cells known to express NKG2D ligands, and we assess the hypothesis that this increased glucose metabolism is a regulatory event in the process of NKG2D ligand expression. We initially aim to establish the importance of glucose metabolism to NKG2D ligand expression and cellular immunogenicity in a range of cell lines, before proceeding to focus our research on a HEK293T cell model of Warburg metabolism. Focussing on a single cell model allows us to overcome the genetic, metabolic, and proteomic heterogeneity that exists between different cancer cell lines. Key observations made in our HEK293T cell model are subsequently assessed in a physiological model of NKG2D ligand (and Warburg metabolism) induction: CMV infection of primary human fibroblasts.

The second approach is to take advantage of a recently developed high throughput method to identify open chromatin sites throughout the genome, DNaseI-seq. We chose this method because the NKG2D ligand loci are spread over an approximate genomic space of 300 kb. Traditional methods of identifying open chromatin sites over this range would be impractical. To improve the accuracy of data analysis, we also developed several of our own software tools **(Chapter 9)**. The work presented forms a strong basis upon which to build a longer term strategy to carry out similar experiments in a physiological model.

2

Materials and Methods

Contents

2.1 Basic molecular biology

2.1.1 Polymerase chain reaction

For a typical polymerase chain reaction (PCR), primers were custom synthesised by either Sigma Aldrich, or Integrated DNA Technologies (IDT). Stock primers were diluted in sterile deionised water (dH_2O) to a concentration of 100 µM, and diluted to a 10 µM working concentration as required. For methods other than DNA cloning, BioTaq DNA polymerase was used (Bioline, BIO-21040), as shown in **Table 2.1**.

Table 2.1: Polymerase chain reaction recipe.

Component	1× concentration
DNA	to 5 µl
10× buffer	2.5 µl
dNTP(10 µM)	2.5 µl
50 mM $MgCl_2$	1 µl
Forward primer (10 µM)	1 µl
Reverse primer (10 µM)	1 µl
Biotaq	0.5 µl
dH_2O	to 25 µl

The PCR reaction was carried out on a heat cycling block. Typical cycling conditions are shown in **Table 2.2**. The number of cycles and annealing temperatures were adjusted as appropriate.

Table 2.2: Polymerase chain reaction cycling conditions.

Stage	Temperature	Duration
Denaturation:	95 °C	5 min
20-35 cycles of:		
Denaturation:	95 °C	30 s
Annealing:	55-65°C	30 s
Elongation:	72 °C	1 min
Final elongation:	72 °C	5 min
Storage:	4 °C	∞

2.1.2 Agarose gel electrophoresis

Agarose gel electrophoresis was carried out using agarose purchased from Sigma Aldrich (A9539). A 50× stock TAE buffer solution was made as follows:

- 242 g Tris base
- 57.1 ml Acetic acid
- 37.2 g (0.1 M) EDTA ($Na_2EDTA \cdot 2H_2O$)
- dH_2O to 1 l

For typical agarose gel electrophoresis, a 1% gel was made by adding 1 gram of agarose to 100 ml 1× TAE buffer. This mixture was boiled by microwave to melt the agarose and allowed to cool. While cooling, 5 µl of a 10 mg ml^{-1} solution of ethidium bromide per 100 ml melted agarose was added to the solution. The gel was then mixed by swirling and allowed to set. A 6× DNA loading buffer was made using:

- 10 mM Tris-HCl pH 7.6
- 0.03% bromophenol blue
- 60% glycerol
- 60 mM EDTA
- in dH_2O

6× DNA loading buffer was added to samples to a final concentration of 1×, and after mixing, samples were loaded to the agarose gel, with a 100bp or 1000bp ladder as appropriate. A voltage of 50-150V was applied to the gel, depending on the distance between the electrodes and the size of the DNA fragments to be separated. Ethidium bromide bands were visualised under UV light.

2.1.3 RNA extraction

RNA extraction was carried out using Trizol and the Purelink RNA mini kit (Invitrogen, 12183018A), largely according to the manufacturers instructions. For adherent cells, the culture medium was removed from the plate, and 1ml of Trizol per 10 cm^2 of culture surface area was added to the plate. The cells were resuspended in trizol, and transferred to RNase-free 1.7 ml microcentrifuge tubes. In some instances, trizol samples were stored at −80 °C at this point for future extraction.

After standing at room temperature in trizol for 5 minutes, 200 µl of chloroform was added to each sample. These samples were shaken by hand for 15 seconds each to mix, and then incubated at room temperature for 3 minutes. The samples were then centrifuged at 12,000 g at 4 °C for 15 minutes. Following centrifugation, 700 µl of the clear aqueous upper phase was removed to a clean RNAse-free 1.7 ml microcentrifuge tube. An equal volume of 70% ethanol made in DEPC (diethylpyrocarbonate) treated water was added to each sample, for a final ethanol concentration of 35%. The samples were then added to the spin columns, 800 µl at one time, and centrifuged at 12,000 g for 15 seconds. The flow through was discarded and centrifugation step repeated until the entire sample was processed.

Where RNA was DNaseI digested, this was carried out using the Purelink DNaseI set (Invitrogen, 12185010), according to the kit instructions. Briefly, the samples were first washed on the spin-column with 350 µl buffer I, and a working dilution of DNaseI was added to the column. The DNaseI reaction was allowed to proceed for 20 minutes at room temperature. The DNaseI was removed by washing from the column, with 350 µl wash buffer I. The column was then washed twice with 500 µl wash buffer II, before spin drying at 12,000 g, and elution to 50 µl RNase-free water.

Where DNaseI digestion was not carried out, samples were instead washed with 700 µl wash buffer I, and wash buffer II as described.

RNA was quantified using a NanoDrop ND-1000 spectrophotometer. RNA samples were stored at −80 °C.

2.1.4 cDNA synthesis

RNA samples used for cDNA synthesis were first diluted to an approximate concentration of ∼200-300 ng µl^{-1}, and the concentration was measured again. On the basis of the diluted measurement, 1-2 µg of RNA was taken from each sample and diluted to a total volume of 10 µl with RNase free water. A dNTP (10 µM)–primer master mix was made, comprised of 1 µl dNTP mix and 0.5 µl primer for each sample. The primer used depended on the purpose

of cDNA synthesis. For selective cDNA synthesis from polyadenylated (full length) mRNA transcripts, 18-dTTP (SO131, Fermentas; 100 µM/0.5 µg µl^{-1}) was used as the primer. Where random cDNA synthesis from the total RNA pool was required, random hexamers (79236, Qiagen; 0.4 µg µl^{-1}) were used as the primer.

This mixture of RNA, dNTP, and primer was heated to 70 °C to anneal primers to the RNA, and then cooled on ice for 5 minutes. For the reverse transcription reaction, a master mix (1×) consisting of 0.25 µl RNasin (N2511, promega), 0.25 µl BioScript MMLV reverse transcriptase (BIO-27036, Bioline) 4 µl of MMLV reaction buffer (5x), and 4 µl of RNase free water. 8.5 µl of master mix was added to each RNA sample, which was then incubated at 42 °C for 60 minutes. The reaction was stopped by heat inactivation at 70 °C for 10 minutes, and then diluted 10 fold with 180 µl of dH_2O. cDNA was stored at −20 °C.

2.1.5 Real-time PCR

The Fast SYBR Green master mix was used for all qPCR reactions (Applied Biosystems, 4385612). The reaction mix for each sample comprised 9 µl DNA, 0.5 µl of each primer (at a typical concentration of 10 µM), and 10 µl of the Fast SYBR Green master mix.

For qPCR reactions measuring cDNA, the 9 µl DNA typically was made up of 4 µl cDNA and 5 µl dH_2O. For qPCR reaction measuring genomic DNA (DNaseI digestion validation experiments), the genomic DNA concentration was initially estimated by NanoDrop spectrophotometry, and 1 µg of gDNA was diluted in 900 µl of dH_2O, leaving a gDNA concentration of 10 ng per 9 µl dH_2O.

Samples were run as technical duplicates or triplicates, on the Applied Biosystems StepOnePlus real time PCR machine. The default cycling conditions were used (initial denaturation at 95 °C for 20 seconds, followed by 40 cycles of 95 °C for 3 seconds, and 60 °C for 30 seconds). A melt curve was routinely performed, to check for the presence of PCR products of different melting temperature.

Amplification efficiency was typically calculated for each primer pair on each real-time

qPCR plate. To calculate amplification efficiency, serial cDNA/gDNA dilutions of between 2-5 fold were made to generate serial target dilution, and amplification efficiency was calculated by plotting the resulting C_t value against $\log_{10}$ of the dilution factor. The slope of the linear best fit line was calculated, and from the slop, the percentage amplification efficiency was calculated as:

$$\left(10^{\left(\frac{-1}{slope}\right)}\right) - 1 \quad \times \quad 100\%$$

Typically, an amplification efficiency of between 90-110% was deemed acceptable. Unless stated, the Pfaffl method was used to calculate fold change in these experiments (209).

2.1.6 Primers

The primers used are shown in **Table 2.3**.

2.1.7 Agarose gel–ethidium fluorescence assessment

For some experiments, relative quantification of PCR products was done by measuring ethidium bromide fluorescence by plate-reader. To ensure product specificity, PCR products were first gel extracted and sequenced by Sanger sequencing with both PCR primers. For relative quantification of PCR products, a standard PCR reaction was carried out using BioTaq DNA polymerase (Bioline, BIO-21040) and typical reaction conditions described in **Tables 2.1** & **2.2**. Samples were run in biological triplicates. Typically a low number of PCR cycles (∼20) was used. For each amplicon, an amplification curve, consisting of serial dilutions of one or more of the DNA samples, was run on each plate, to ensure that the PCR conditions and primers were able to discriminate relative differences in template concentration.

Low melting point agarose was purchased from Sigma (A9414), and a 1% agarose gel in 1× TAE buffer, containing $0.5\,\mu g\,ml^{-1}$ ethidium bromide was allowed to set. The PCR reactions were run on the gel at ∼3-5 $V\,cm^{-1}$ for 15 minutes. Using a ruler and scalpel,

Table 2.3: Primers.

Purpose	Name	Target	Size	Forward seq	Reverse seq	Ref.
Unspliced RTPCR	CO539/CO4057	ULBP1	171bp	AGGCCAAAGCCTTTGCTTCT	TGTCTGCTCCCCTGGGCCATTTT	
	CO541/CO4058	ULBP2	195bp	CAGTCACACCTGTCAGTCC	CTCTAACCTACCACTGTATCTGT	
	CO4144/CO543	ULBP3	175bp	CTGCTCCCTTCTGTCCTTGGTCT	AGGTCTTATCTATGGGTCAC	
	CO537/CO4180	ULBP4	585bp	TATCCCTGACTTCTAGCCCT	CCTGACTGTCTGGGATGAGATAC	
	CO324/CO4147	ULBP5	255bp	AGCCCCCGCGTTCCTTCTA	CTCGGAGGATAAGTAGCCTCT	
	CO326/CO4182	ULBP6	464bp	CATCCCAGCTTTGCTTCTGTGC	GTCTGTGATTGTATCCCTGTGGT	
	CO4173/CO4174	MICA	302bp	TCTGAAATCCGGGGGTGGCCATC	AAGGCTTCCTCTCCTGCCTGACA	
	CO4175/CO4176	MICB	302bp	TCTGAAATCCGGGGGTGGCCATC	AGGGCTTCCTCTCCTGCCTAACT	
	CO3930/CO3931	ACTB	207bp	AGAAAATCTGGCACCACACC	CGGCAGAAGAGAGAACCAGT	
GLUT RTPCR	CO3062/CO3063	GLUT1	207bp	GTGCAGCAGCCTGTGTATGC	GGCCACGATGCTCAGATAGG	
	CO3066/CO3067	GLUT2	192bp	CTGTGCTGGGTTCCTTCCAG	AAGGGGTTGGTTTTGGGTTC	
	CO3070/CO3071	GLUT3	196bp	TACAGCGATGGGGACACAGAAG	CCAGAGAGACGTGAGCAGCAC	
	CO3074/CO3075	GLUT4	198bp	CCCTGGTCCTTGCTGTGTTC	AAAAGATGGCCACGGAGAGG	
	CO3078/CO3079	GLUT5	396bp	GAATTCATGGAAGACTT	GCCATCTACGTTTGCAA	(210)
GLUT5 cloning	CO3076/CO3077	GLUT5	1531bp	TTAATAAGCTTGCCATGGAGCAACAGGATCAGAG	TTAATCTCGAGTCACTGTTCCGAAGTGACAGGTG	
eGFP cloning (pTripz)	CO3950/CO3951	eGFP	720bp	TATATT CTCGAG ATGGTGAGCAAGGGC	TATATT ACGCGT TTACTTGTACAGCTC	
CMV UL123 RTPCR	CO3592/CO3593	UL123	101bp	ATTTTCTGGGCATAAGCCATAATC	GCCTTCCCTAAGACCACCAAT	(211)
β-actin RTPCR	CO1851/CO1852	ACTB	382bp	TCCAGCCTTCCTTCCTGGGCAT	GTCAAGAAAGGGTGTAACGCAACT	

the PCR products of interest were excised from the gel, along with adjacent empty lanes to measure background fluorescence. The excised bands were individually melted in 150 µl 1× TAE buffer on a heating block at 70 °C, and transferred to a black 96-well flat-bottomed plate. The plate was cooled at 4 °C to re-solidify the agarose. A Tecan Safire2 plate reader was used to measure the ethidium bromide fluorescence in each well at excitation/emission wavelengths of 493/620nm. The mean fluorescence and 95% confidence interval was calculated for each sample. The background fluorescence was subtracted by calculating the difference between two means.

2.1.8 Sanger sequencing

Sanger sequencing was used to confirm the sequence of key PCR products or plasmid design features/inserts. The BigDye v3.1 sequencing master mix and buffers were purchased locally through the Department of Zoology, University of Oxford. Typically, a 20 µl sequencing reaction was carried out on a 96-well PCR plate **(Table 2.4)**.

Table 2.4: BigDye v3.1 sequencing reaction recipe.

Component	1× concentration
Template	PCR product: 1- 10 ng Plasmid: 100 - 300 ng
BigDye v3.1 master mix	1 µl
BigDye 10× buffer	2 µl
Seq primer (3.2 µM)	1 µl
dH_2O	to 20 µl

The prepared reaction mixture was thermo-cycled under the conditions shown in **Table 2.5**. To clean up the dye terminated PCR products, the sequencing plate was centrifuged at 930g to collect all the sample at the bottom of the well. The DNA fragments were precipitated by adding 2 µl of 3 M sodium acetate (NaAc; pH 5.2) and 50 µl of 100% ethanol to each well. The plate was sealed with an adhesive sheet, and the sample was mixed by inversion and incubated at room temperature for 15 minutes. The DNA fragments were pelleted by centrifugation at 3000g for 30 minutes at 4 °C. To remove the liquid from the DNA pellet,

the plate seal was removed, and the plate was inverted onto a paper towel followed by pulse centrifugation of the inverted plate to 185g. The DNA fragments were washed with 70 µl of 70% ethanol per well, the plate was re-sealed, and centrifuged upright at 1650g for 15 minutes at 4 °C. Again, the plate was inverted onto a paper towel and centrifuged at 185g for 1 minute. The seal was again removed, and the DNA pellets were air dried at room temperature for 10 minutes. The plate was then re-sealed and stored at −20 °C prior to sequencing. Sequencing was carried out by the Department of Zoology, University of Oxford.

Table 2.5: Sanger sequencing reaction cycling conditions.

Stage	Temperature	Duration
Denaturation:	96 °C	1 min
25 cycles of:		
Denaturation:	96 °C	30 s
Annealing:	50 °C	30 s
Elongation:	60 °C	1 min
Storage:	4 °C	∞

2.1.9 Production of transformation competent cells

Transformation competent DH5α cells were used to propagate existing plasmids, and to isolate newly ligated plasmids. Two separate buffers were used in their production **(Tables 2.6** and **2.7)**. DH5α *E. coli* cells were streaked onto an ampicillin-free agar plate, and cultured at 37 °C overnight. A single DH5α colony was picked, and used to spike 5 ml of liquid broth, which was then transferred to a shaking incubator at 37 °C for 4 hours. This 5 ml culture was transferred to a 2 l conical flask containing 250 ml liquid broth, and placed in a shaking incubator at 37 °C until the optical density at 600nm, compared to sterile liquid broth, reached +0.6. At this point, the bacterial culture was transferred to 250ml conical centrifuge tubes, and centrifuged at 1,100 g for 5 minutes at room temperature to form a bacterial pellet. The liquid broth was discarded, and the pellet was resuspended in 50 ml of ice cold buffer A on ice, and transferred to 50 ml centrifuge tubes. The bacteria were centrifuged again at 1,100 g for 5 minutes at 4 °C, the excess liquid was discarded, and the

bacterial pellet was resuspended in 10 ml of ice cold buffer B. This suspension was transferred to 1.5 ml micro centrifuge tubes in 200 µl aliquots, before being snap frozen in liquid nitrogen, and stored at −80 °C.

Table 2.6: Transformation competence buffer A.

Component	Concentration	For 500 ml
Potassium acetate (KOAc)	30 mM	1.47 g
Manganese chloride ($MnCl_2$)	50 mM	4.95 g
Potassium chloride (KCl)	100 mM	3.72 g
Calcium chloride ($CaCl_2$)	100 mM	0.55 g
Glycerol	15% by vol	75 ml
dH_2O	–	to 500 ml

Table 2.7: Transformation competence buffer B.

Component	Concentration	For 500 ml
Sodium-MOPS (NaMOPS; pH7)	10 mM	5 ml of 1 M
Potassium chloride (KCl)	10 mM	0.37 g
Calcium chloride ($CaCl_2$)	75 mM	4.16 g
Glycerol	15% by vol	75 ml
dH_2O	–	to 500 ml

2.1.10 Standard gene cloning

The strategy for cloning primer design is illustrated in **Fig 2.1**. The Consensus Coding Sequence database (CCDS) was used to identify the reference gene coding sequence when designing primers and restriction assays (212). The CCDS database was accessed through the NCBI's Gene resource. To design the primers, the CCDS coding sequence from ATG to Stop was noted. For forward primers, the first ∼15 base pairs of the coding sequence were prefixed by a GCC kozak sequence (to ensure efficient translation), the restriction site of choice, and a TTAAT handle. For the reverse primer, the reverse complement of the final ∼15 base pairs of the gene were derived, and also prefixed by a TTAAT handle. For eGFP fusion proteins, the stop codon was omitted from the reverse primer.

For the cloning PCR reaction, the proofreading DNA polymerase, *pfu*, was used, and templates were amplified from oligo dT primed cDNA. The PCR mix for a typical reaction

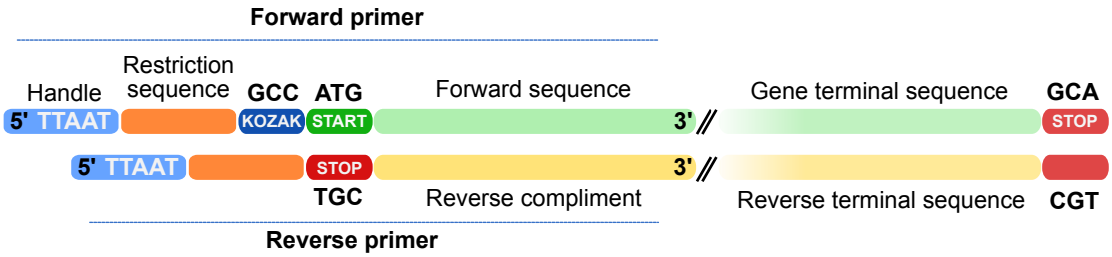

Figure 2.1: Design of primers for expression cloning. The primers used to clone target genes for expression were designed as depicted above. For forward primers, the primer starts with a TTAAT handle, followed by the restriction site of choice, and a GCC kozak sequence immediately before the start codon and first ∼15 bases of the target gene. The reverse primer was similarly designed, starting with a TTAAT handle, then restriction site and the reverse complement of the stop codon and final ∼15 bases of the target gene. For fusion proteins, the stop codon sequence was omitted. Once designed, primers were checked using the NCBI's primer blast service, and UCSC's in silica PCR.

is shown in **Table 2.8**

Table 2.8: Pfu polymerase PCR recipe.

Component	1× concentration
cDNA	10 µl
10× buffer	5 µl
dNTP(10 µM)	5 µl
50 mM $MgSO_4$	3 µl
Forward primer (10 µM)	2.5 µl
Reverse primer (10 µM)	2.5 µl
pfu polymerase	0.5 µl
dH_2O	to 50 µl

The cycling conditions were typically carried out as described in **table 2.9**.

Table 2.9: Pfu polymerase cycling conditions.

Stage	Temperature	Duration
Denaturation:	95 °C	5 min
20-35 cycles of:		
Denaturation:	95 °C	30 s
Annealing:	55-65°C	30 s
Elongation:	72 °C	2 min kb^{-1}
Final elongation:	72 °C	5 min
Storage:	4 °C	∞

To confirm successful amplification of the target gene, 10 µl of each PCR product was run on a 1% agarose gel. The remaining 40 µl was cleaned up using the Qiagen QIAquick PCR purification kit (#28104) according to the manufacturers protocol, eluting to 30 µl of dH_2O.

The cleaned up PCR reaction, and 2 µl of the parent plasmid (∼300 ng µl^{-1}) were digested with appropriate restriction enzymes, with the most suitable manufacturer recommended reaction buffer, at 37 °C for 3 hours. Bothe the digested plasmid and PCR product were gel purified to remove excess reactants and digested fragments using the Qiagen QIAquick gel extraction kit (#28704) according to the manufacturers protocol, eluting to 30 µl of dH_2O. The insert was ligated to the plasmid in a 10 µl ligation reaction consisting of 6 µl of gel purified insert, 2 µl of gel purified cut plasmid, 1 µl of T4 DNA ligase, and 1 µl of T4 DNA ligase reaction buffer, for 1 hour at room temperature.

To transform the ligated plasmid, competent DH5α cells were thawed from storage in −80 °C on ice, and 50 µl of DH5α cells were added to the ligated plasmid. The reaction was mixed by pipetting, before being heat shocked at 42 celsius for 1 minute. A total of 50 µl of antibiotic free liquid broth medium was added to this mixture, which was then placed in a shaking incubator at 37 °C for 30 minutes. The mixture was then spread on an agar plate containing the appropriate selection antibiotic. The plate was inverted and placed in an incubator at 37 °C overnight.

The following day, four colonies were picked from each plate, and again cultured over night in 5 ml of liquid broth medium containing an appropriate selecting antibiotic. Plasmids were isolated from the resulting cultures using the Fermentas GeneJET plasmid mini prep kit (#K0503), according to the manufacturers protocol, eluting to 50 µl of dH_2O. Plasmids were initially screened by restriction digestion, using one restriction site within the insert, and one within the plasmid. Where appropriate, the insert sequences were confirmed with Sanger sequencing.

2.1.11 Transient expression plasmid transfection

Transient transfections were typically carried out with polyethylenimine (PEI). A 1 mg ml^{-1} stock solution of PEI was made, and the pH was adjusted to pH7 with HCl. The solution was sterile filtered at 0.22 µm. Cells were plated in fresh medium the day before transfection. To

transfect 1 well of a 6-well plate, 2 µl of this PEI solution was added to 200 µl of serum free medium, the solution was mixed by flicking, and allowed to incubate at room temperature for 5 minutes. 1 µg of plasmid DNA was added to the PEI/medium mixture, mixed by flicking and incubated at room temperature for 10 minutes. The total volume was added drop-wise to the well. For replicate transfections, a master mix of the transfection mixtures described was used. Following transfection, cells were cultured for 24-48 hours following transfection before use. When cells were cultured for more than 24 hours following transfection, the culture medium was changed at 24 hours.

2.1.12 Generation of eGFP tagged expression plasmid

For experiments where N-terminal eGFP fusion proteins were expressed, and eGFP-pCDNA3.1 expression plasmid was used. To make the eGFP-pCDNA3.1 expression plasmid, an optimised multiple cloning site (MCS; **Fig 2.2**) was first cloned into pCDNA3.1 between the HindIII and XhoI restriction sites, generating the plasmid shown in **Fig 2.3**. The coding sequence for eGFP was amplified from pEGFP-N1 and was cloned into the modified pCDNA3.1 vector between the XhoI and XbaI sites **(Fig 2.4)**. Target genes were cloned, as described above, omitting the terminal stop codon in the reverse primer. This gene was cloned in-frame into the MCS in the modified pCDNA3.1 vector, generation a target gene-eGFP fusion gene. Clones were checked by restriction digestion or Sanger sequencing as appropriate.

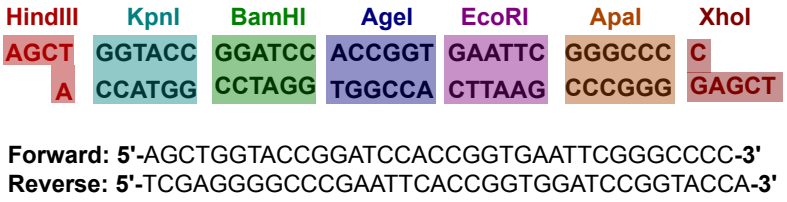

Figure 2.2: Multiple cloning site for pCDNA3.1. To facilitate the production of eGFP fusion proteins, the restriction sites in the existing pCDNA3.1 expression vector were concatenated into the new MCS shown above. The MCS oligomers were added in equimolar amounts and annealed by placing on a heating block at 95 °C for 10 minutes, switching off the heating block and allowing to cool gradually overnight.The coding sequence for eGFP was cloned between XhoI, and the existing XbaI site downstream.

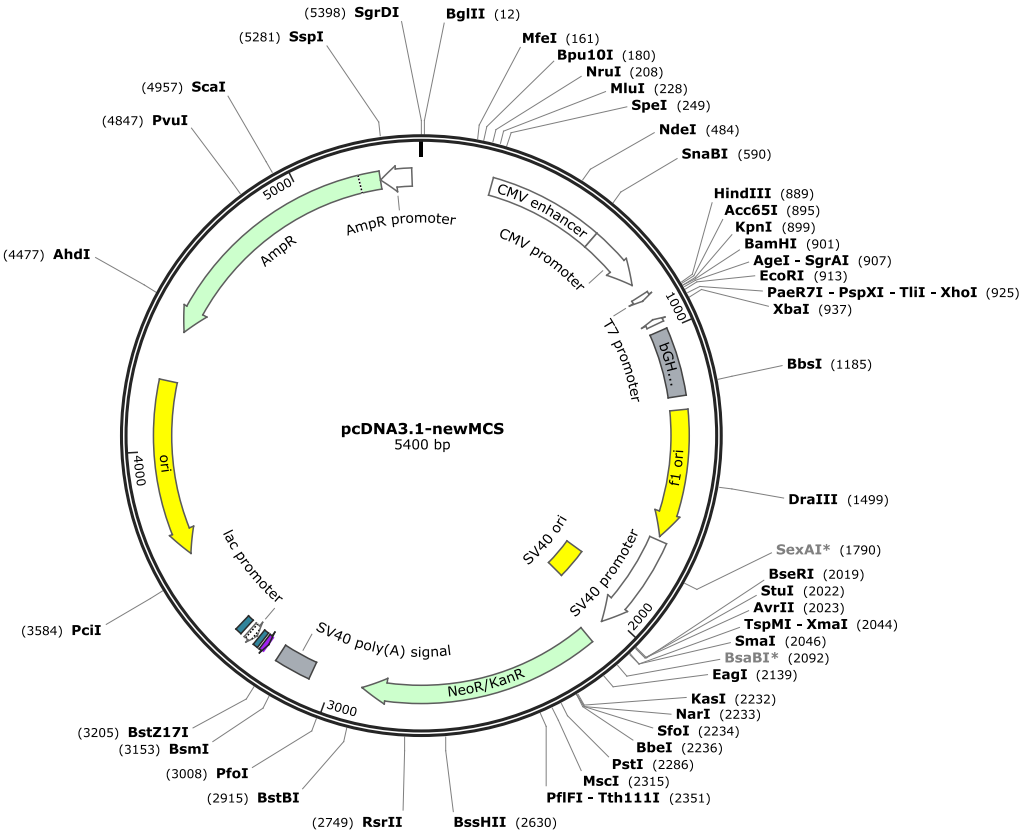

Figure 2.3: pCDNA3 map with new multiple cloning site. A new pCDNA3.1 expression vector with a compact multiple cloning site (MCS) was made, by excising the existing MCS between HindIII and XhoI, and ligating in the new MCS as shown in **Fig 2.2**. Map image source: `www.snapgene.com/resources`

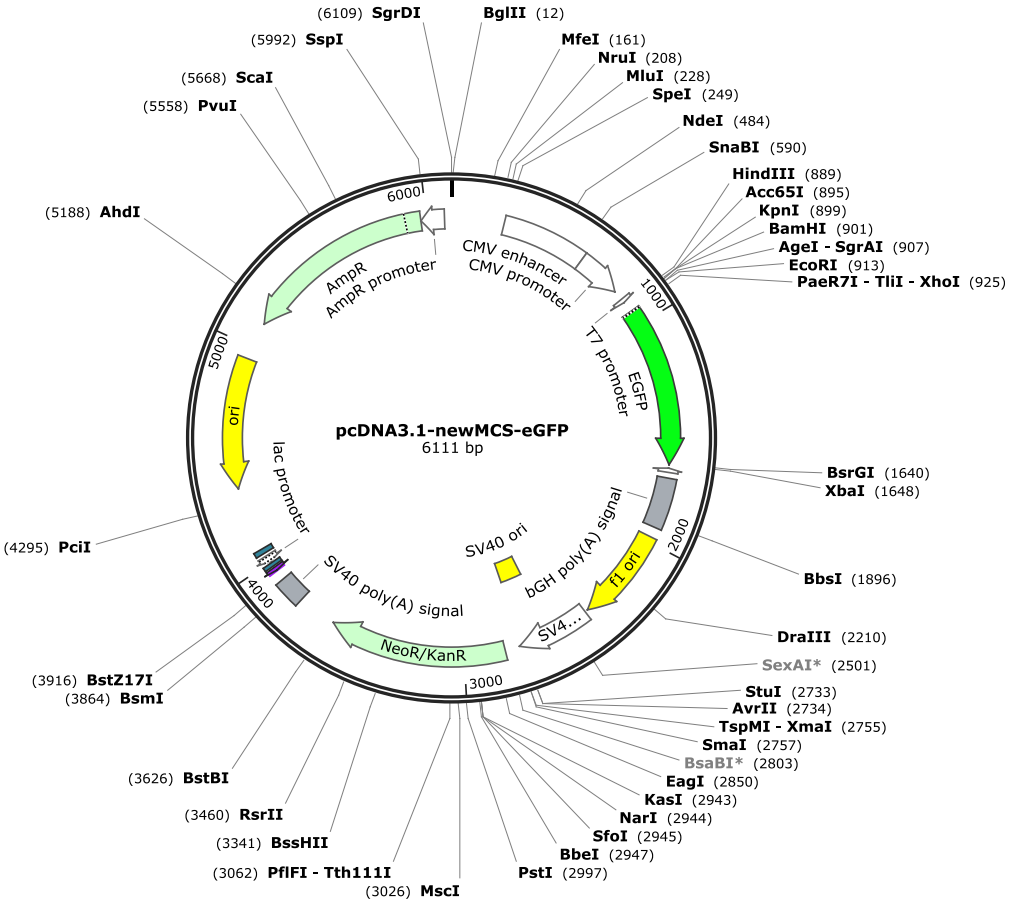

Figure 2.4: pCDNA3-newMCS-eGFP plasmid map. The coding sequence for eGFP was cloned into the newly generated expression plasmid between the XhoI site at the end of the new MCS, and the XbaI site present in the original pCDNA3.1 backbone. The final plasmid consisted of an eGFP expression vector under the control of a CMV promoter, with a compact MCS suitable for gene fusion. Map image source: `www.snapgene.com/resources`

2.1.13 Generation of a doxycycline inducible eGFP expression plasmid

For several experiments we wanted to have a selectable, doxycycline-inducible, lentiviral cDNA expression vector. The pTripZ empty vector (kindly donated by the Ratcliffe group, CCMP) was designed as an inducible microRNA expression vector. It includes a 'cassette' for shuttling in microRNA constructs in front of the tet promoter. TurboRFP is encoded downstream of the microRNA cassette, and is expressed in line with the microRNA. pTripZ also encodes the puromycin resistance gene, whose translation is dependent on an IRES, and whose transcription is driven by the tetracycline promotor. The vector includes the tetracycline transactivator protein **(Fig 2.5)**.

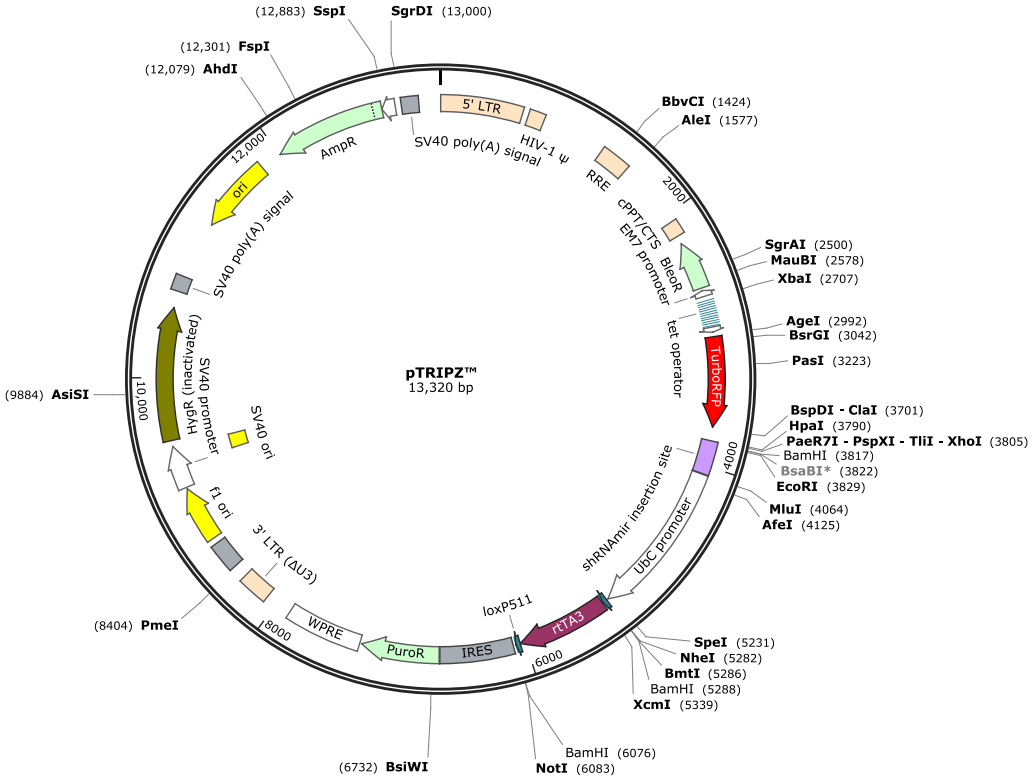

Figure 2.5: pTripZ empty vector. The original pTripZ was designed to express microRNA sequences under tetracycline control, in line with turboRFP. We converted this to a doxycycline-inducible cDNA expression plasmid by inserting a new MCS between the AgeI and MluI sites, replacing the turboRFP and microRNA expression cassette in the process. Map image source: `www.snapgene.com/resources`

To convert this to a cDNA expression vector, we excised turboRFP and the microRNA cassette by digesting the plasmid with AgeI and MluI. To replace the excised DNA, we made a new multiple cloning site **(Fig 2.6)** from two oligomers, which were mixed in equimolar

amounts and annealed by cooling from 95 °C on a heating block overnight. This new MCS was ligated between the AgeI and MluI sites. We generated an inducible eGFP vector, by cloning the eGFP coding sequence from pEGFP-N1 between the XhoI and MluI sites of the new pTripZ vector **(Fig 2.7)**.

Age	ClaI	EcoRI	BstBI	XhoI	AseI	HindIII	MluI
CCGGT	ATCGAT	GAATTC	TTCGAA	CTCGAG	ATTAAT	AAGCTT	A
A	TAGCTA	CTTAAG	AAGCTT	GAGCTC	TAATTA	TTCGAA	TGCGC

Forward: 5'-CCGGTATCGATGAATTCTTCGAACTCGAGATTAATAAGCTTA**-3'**
Reverse: 5'-CGCGTAAGCTTATTAATCTCGAGTTCGAAGAATTCATCGATA**-3'**

Figure 2.6: Multiple cloning site for pTripZ. The MCS was designed to replace the turboRFP coining sequence and microRNA cassette from the original pTripZ empty vector. It concatenates the available unique restriction sites occurring in the excised stretch of DNA.

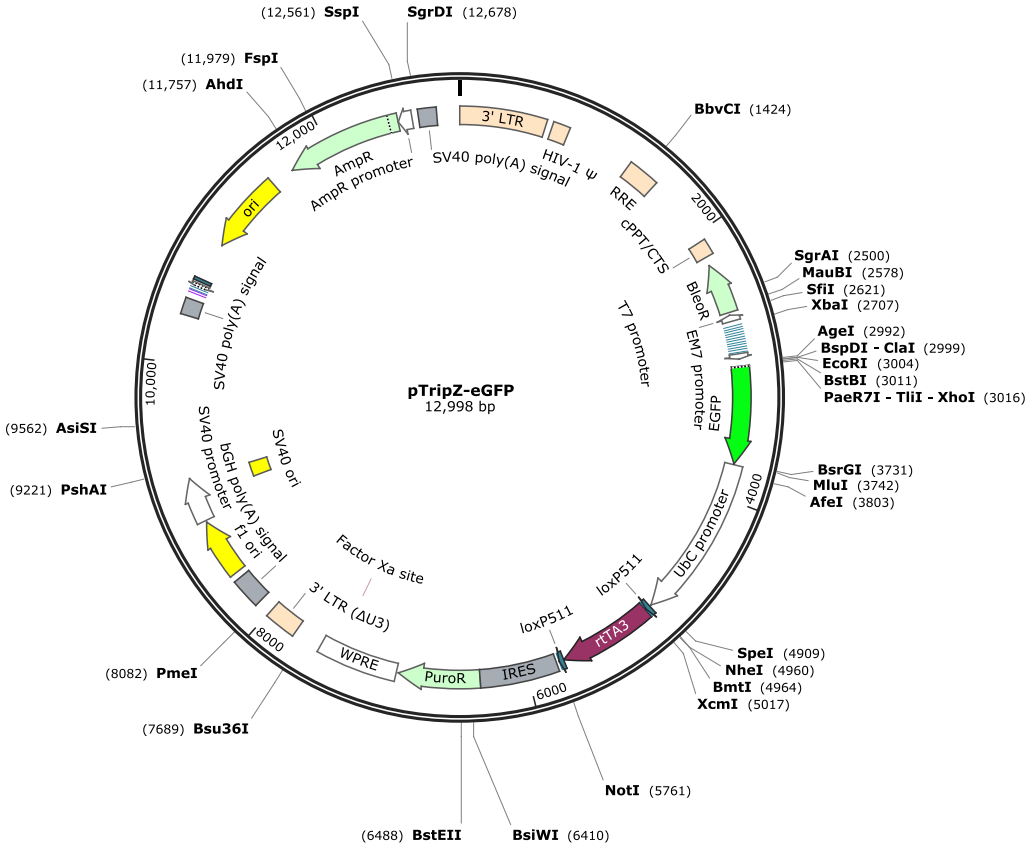

Figure 2.7: pTripZ-eGFP vector. The final pTripZ-eGFP expression plasmid is encodes eGFP under the control of a tetracycline enhancer and minimal CMV promoter, the tetracycline transactivator sequence, and the puromycin resistance gene for selection, driven by tetracycline and translated on an IRES sequence. Map image source: `www.snapgene.com/resources`

2.1.14 Plasmids and constructs

The key plasmids used or constructed are shown in **Table 2.10**.

2.1.15 Measurement of glucose, lactate and osmolality

While commercial kits and instruments are available for the measurement of glucose, lactate and osmolality, the Department of Biochemistry at the John Radcliffe Hospital, Oxford, offered a cost effective service with certifiable levels of accuracy. All samples were filtered at 0.2 µm, immediately stored at −80 °C, and transferred to the Department of Biochemistry on dry ice.

Glucose measurement was carried out using the glucose hexokinase II reagent kit (#04903429; Bayer Healthcare, Newbury, UK) **(Fig 2.8a)**. A solution containing ATP and NAD are added to the sample, and the absorbance of NADH at 340 nm is measured on the ADVIA 2400 clinical chemistry system. A second solution containing hexokinase II, and glucose 6-phosphate dehydrogenase is then added. The glucose in the sample is converted to glucose 6-phosphate, which in turn is converted to 6-phospho-glucono-1,5-lactone in a reaction that reduces NAD to NADH. The change in absorbance at 340 nm due to the production of NADH is proportional to the glucose concentration in the sample.

The lactate concentration is measured similarly using the lactate reagent kit (#07109944; Bayer Healthcare) **(Fig 2.8b)**. This two-step enzymatic reaction results in the production of a purple dye, which changes the absorbance of the solution at 545 nm. This change in absorbance is proportional to the lactate concentration in the original sample.

The osmolality was measured on a VITECH 3320 micro osmometer. This instrument measures the temperature at which the sample freezes, which is proportional to the osmolality of the sample.

Table 2.10: Plasmids and constructs

Number	Parent	Insert	Notes
pOC1220	pTripZ	–	Empty parental vector.
pOC1221	pTripZ	new MCS	CO3909/CO3910 annealed, ligated between AgeI-MluI.
pOC1222	pTripZ (pOC1221)	kozak-eGFP	XhoI-MluI.
pOC1224	pMD2.G	–	lentivirus core envelope.
pOC1225	psPAX2	–	lentivirus core packaging.
pOC1226	pCDNA3	NLRC5	myc-tagged NLRC5 expression, Addgene.
pOC1227	pCDNA3.1	new MCS	CO4347/CO4348 annealed, ligated between HindIII-XhoI.
pOC1229	pCDNA3.1 (pOC1227)	kozak-GLUT5	XhoI-XbaI.

MCS - Multiple cloning site.

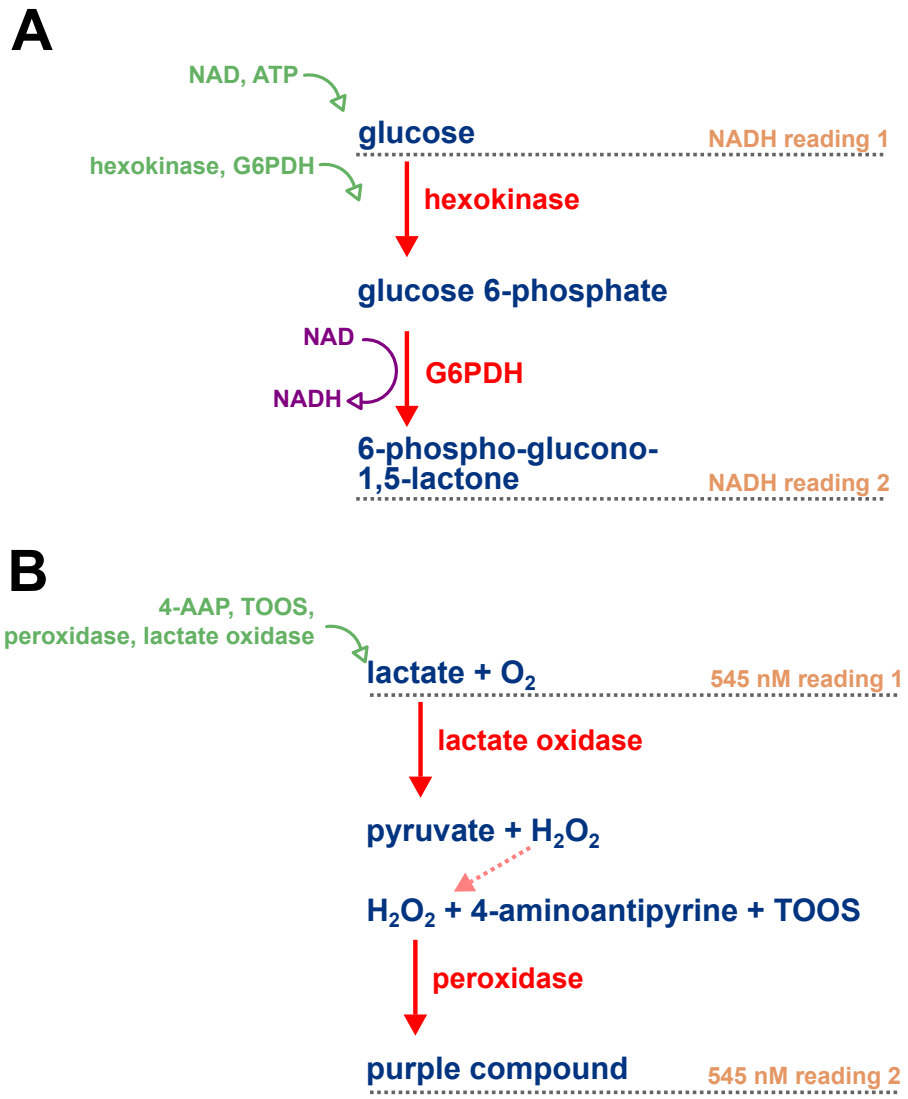

Figure 2.8: Measurement of glucose and lactate. Glucose and lactate concentrations were measured by the Department of Biochemistry at the John Radcliffe Hospital Oxford on an ADVIA 2400 clinical chemistry system. Both measurements are based on enzymatic reactions. (A) The amount of glucose in a sample is quantified by measuring the change in absorbance at 340 nm caused by the change in NADH concentration. NADH is produced when G6PDH acts on glucose 6-phosphate, the concentration of which is dependent on the amount of glucose present in the sample. (B) Lactate oxidase converts lactate present in the sample to pyruvate, with the production of H_2O_2. H_2O_2 reacts with 4-AAP and TOOS to generate a purple compound in the presence of peroxidase. The purple compound causes a change in the absorbance at 545 nm, which is proportional to the amount of lactate in the sample. NAD - Nicotine adenine dinucleotide; ATP - Adenosine triphosphate; G6PDH - Glucose 6-phosphate dehydrogenase; 4-AAP - 4-aminoantipyrine; TOOS - N-ethyl-N-(2Hydroxy-3-sulphopropyl) m-toluidine.

2.1.16 Chemicals and reagents

Glucose (#49152), mannitol (#17311), fructose (#F3510), mannose (#M6020), pyruvate (#P5280), azaserine (#11430), 6 diazo-oxo-norleucine (#D2141), fetal calf serum (#F7524), bovine serum albumin (#A3059), 2-deoxyglucose (#D8375), phloretin (#P7912), cytochalasin B (#C6762-1mg), CFDA-SE (#21888), caffeine (#C0750), all nucleosides and nucleobases, and 50x hypoxanthine-thymidine solution (A9666) were purchased from Sigma-Aldrich, Dorset, UK. AICAR (#9944S) was purchased from New England Biolabs (NEB; Ipswich, MA, USA). Adenosine Kinase Inhibitor (#116890) was purchased from Merck-Millipore, Watford, UK, and Iodotubercidin was purchased from Abcam, Cambridge, UK. RPMI (#E15-039), standard DMEM (#E15-011), L-glutamine (#M11-004) and penicillin/streptomycin (#P11-010) were purchased from PAA laboratories, Somerset, UK. DMEM without glucose (#11966-025) was purchased from Invitrogen. For general molecular biology purposes the GeneJET miniprep kit (#K0503) was purchased from Fermentas, Germany, and the QIAquick PCR purification (#28104) and gel extraction (#28704) kits were obtained from Qiagen, West Sussex, UK. Blue role (#CMC-717-022U) for wiping up spills was purchased from Fisher. Maxipreps were carried out using a VWR maxiprep kit (#OMEGD6922-02). Primers were purchased from Sigma or IDT. Biotaq (#BIO-21040) for standard PCR was purchased from BioLine. Restriction enzymes were obtained from either NEB, UK or Fermentas, Germany. For cloning, pfu polymerase (#EP0501, Fermentas) was used to amplify target cDNA according to the manufacturers protocol. For overexpression experiments, the target cDNA was cloned into pCDNA3.1 under the control of the CMV promoter. Transient transfections were carried out using genejuice (#70967-3, Merck Bio), or PEI according to their provided protocol.

2.2 Flow cytometry

2.2.1 Cell surface staining for flow cytometry

Flow cytometry experiments were carried out in biological triplicate, and repeated three of more times, unless otherwise stated. Cells were removed from incubator to class II tissue culture hood. The culture medium was removed and discarded. Cells were washed once with sterile PBS. For cells other than 293T cells, 5mM EDTA in PBSA solution at 37 °C was used to detach cells from plate. For 293T cells, PBSA alone was used to detach cells. Cells were then transferred to a 96-well plate (maximum volume 200 µl per well). The cells were centrifuged at 1500rpm for 5 minutes, and the supernatant was discarded. Primary staining was carried out with 0.2 µg of primary antibody or isotype control in 100 µl cold PBSA with 1% BSA per well. The cells were incubated in the primary stain for 20-30 minutes at 4 °C. The cells were centrifuged again at 1500rpm for 5 minutes, before being washed with 150 µl cold PBSA per well. Secondary staining was carried out with 0.2 - 1 µg conjugated secondary antibody in 100 µl cold PBSA per well, at 4 °C for 20-30 minutes. The cells were then washed again with 150 µl cold PBSA, before being resuspended in 100 µl cold PBSA for analysis by flow cytometry.

2.2.2 Dead cell staining

An aliquot of cells were detached and washed with cold PBSA as described above. A propidium iodide (PI) staining buffer comprising 1 µg ml^{-1} of PI in PBSA was prepared. Cells were resuspended in 100 µl of PI staining buffer and stored at 4 °C in the dark for ∼15 minutes before flow cytometry. Where live and dead cells were distinguishable by FSc/SSc (forward scatter/side scatter) characteristics, PI staining was not routinely carried out to minimise fluorochrome interference, and limit the need for compensation.

Table 2.11: Total cellular staining for flow cytometry.

Round	Surface only	Total cellular
1	Surface stain (P)	Surface stain (P)
2	Fixation (P)	Fixation (P)
3	PBSA (P)	0.1% saponin (S)
4	Surface stain (P)	Intracellular stain (S)
5	Secondary stain (P)	Secondary stain (S)
6	Resuspension (S)	Resuspension (S)

(P) - PBSA-based buffer; (S) - Saponin-based buffer

2.2.3 Total cellular staining for flow cytometry

Combined cell surface and intracellular staining was carried out as outlined in **Table 2.11**. In each case of total cellular staining, an aliquot of the same cells was also stained at the surface only as a control. In addition, isotype controls were used for all samples to determine the background level of staining. Cells were harvested as described above, and washed with PBSA. Primary surface staining was carried out using 0.2 µg of primary antibody or isotype control for both surface and total cell stains in 100 µl PBSA/0.1% BSA per well of a 96-well plate. The cells were washed with 150 µl PBSA, and resuspended in 100 µl of 1% paraformaldehyde (PFA) in PBSA to fix for 30 minutes in the dark at room temperature. The cells were again washed with 150 µl PBSA to remove residual PFA.

To permeabilise the cells for total cell staining, cells were resuspended in 100 µl of 0.1% saponin in PBSA, while surface only cells were resuspended in PBSA only in the dark at 4 celsius. From this point on, cells for total cell staining were only resuspended in 0.1% saponin containing buffers. A second round of primary antibody/isotype control staining was carried out for both surface only and total cell staining samples, using 0.2 µg of antibody, and 0.1% BSA. Cells were washed with 150 µl PBSA or PBSA/0.1% saponin, before staining with a conjugated secondary antibody (0.2 - 1 µg per well) in PBSA/0.1% BSA, with or without 0.1% saponin. Cells were washed again, and resuspended in PBSA with 0.1% saponin and analysed by flow cytometry.

2.2.4 Staining for cell cycle analysis

Cells were trypsinized before being washed twice with PBSA and transferred to 15 ml centrifuge tubes. The cell pellets were then resuspended in 1 ml of 1% paraformaldehyde (PFA) and then placed on ice for 60 min. Cells were washed with 5 ml of ice cold PBSA, and resuspended in 0.5 ml of PBSA. While vortexing, 4.5 ml of ice cold 70% ethanol was added drop-wise to each sample. The samples were then incubated at 4 °C for two hours. The cells were washed twice in PBSA, and resuspended in 0.5 ml of PI staining solution containing 40 µg ml^{-1} PI, 100 µg ml^{-1} RNase A in PBS. The cells were incubated at 37 °C for 30 minutes before being analysed on a FACSCanto flow cytometer, measuring PI fluorescence emission in the linear mode. Doublets were excluded by initial PI width-area gating. PI voltage was adjusted to centre the G1 population on the 50k voltage area level. Cell cycle parameter modelling was carried out using FlowJo software.

2.2.5 CFSE proliferation assay

5(6)-Carboxyfluorescein diacetate N-succinimidyl ester (CFDA SE) freely diffuses across the mammalian cell membrane. Once inside the cell, the diacetate group rapidly undergoes hydrolysis, generating carboxyfluorescein N-succinimidyl ester (CFSE), which cannot cross the cell membrane. Relative changes in CFSE fluorescence over time occur through dilution of CFSE by cell division. For these experiments, CFDA SE was purchased from Sigma (21888). A 10 mM stock solution of CFDA SE in DMSO was prepared and stored at −20 °C. To label cells with CFSE at the beginning of an experiment, a working 5 µM solution of CFDA SE in warm (37 °C) PBS. The culture medium was removed from target cells, and cells were washed in situ with sterile PBS. The warm 5 µM CFDA SE solution was added to the adherent target cells to completely cover the cell monolayer in each well (typically 0.5 ml in a 12-well plate, and 1 ml in a 6-well plate). The cells were incubated with the CFDA SE solution for 10 minutes at 37 °C. The CFDA SE solution was then removed, the cells were

washed with sterile PBS, and fresh culture medium was added to each well, along with and indicated treatments. The cell culture plate was covered in foil and cultured in a standard incubator at 37 °C for the time interval indicated. CFSE fluorescence was measured by flow cytometry.

2.2.6 Flow cytometry data acquisition and analysis

Flow cytometry was carried out on a BD FACSCanto, using BD FACSDiva software for data collection. For typical data collection, unstained cells were initially acquired to adjust forward scatter (FSc), side scatter (SSc) and fluorescence voltages. For most cell types used, the height of the FSc voltage trace, and the area of the SSc voltage trace were used. **(Fig 2.9)**. Having adjusted FSc-H and SSc-A voltages to fit events into a detectable window, live (FSc/SSc healthy) cells were gated. Fluorescence voltages were adjusted, using the unstained cells, to set baseline voltage reading to negative. Where appropriate, propidium iodide staining was used to confirm that cells in the FSc/SSc healthy gate population had intact cell membranes.

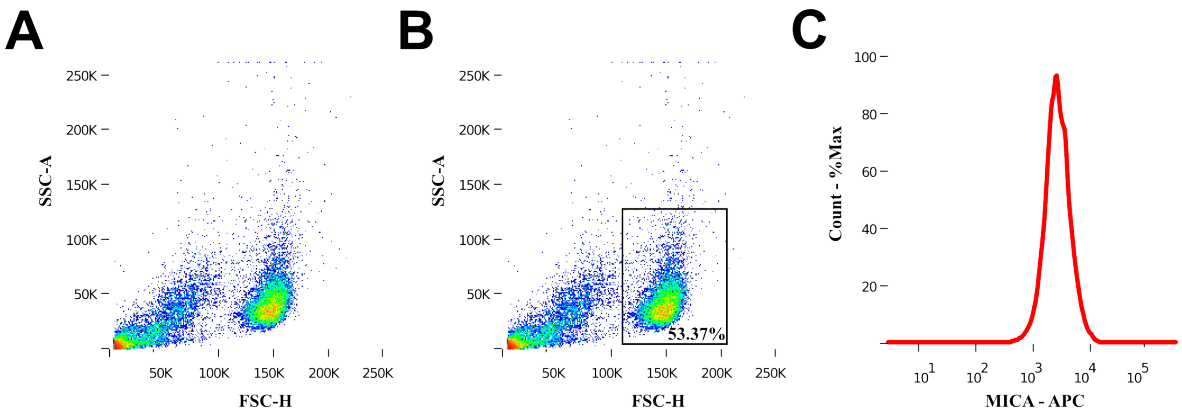

Figure 2.9: Acquisition and analysis of flow cytometry data. (A) Typically, forward scatter height voltage, and side scatter area voltage were adjusted to contain events within the observation window. (B) Based on forward scatter and side scatter characteristics, we gated on a live cell population. The cells within this window were consistently propidium iodide (PI) negative. (C) After adjusting baseline fluorescence voltages using unstained cells from the gate set in (B), fluorescence was recorded for 10,000 events in this gate for each treatment condition.

When two fluorochromes were used simultaneously, where the possibility of spectral overlap existed, compensation controls were used. Unstained cells and cells stained with each fluorochrome individually were run separately, and compensation was applied at the point

of data acquisition. For standard experiments, 10,000 gated (live cell) events were recorded for each sample. All samples were measured in parallel biological triplicate, unless otherwise state.

Following data acquisition, the complete dataset was exported to FlowJo for further analysis. Typically, initial gating was carried out as described above **(Fig 2.9)**. Having set the initial gate to include a maximum number of FSc/SSc healthy cells, a frequency distribution histogram of specific fluorescence within this population was created. The mean fluorescence intensity was recorded for each sample. Biological triplicate values were used to calculate the 95% confidence interval for the estimate of mean fluorescence intensity. Analysis of these figures was carried out using Microsoft Excel or Apple Numbers, and Graphpad Prism.

2.3 Cell-based techniques

2.3.1 Preparation of cell culture media

Standard DMEM and RPMI-1640 were purchased from Sigma or PAA laboratories. Fetal calf serum (FCS) was purchased from Sigma and First Link UK. To prepare DMEM and RPMI-1640 for routine cell culture, the FCS was heat inactivated at 55 °C for 15 minutes, and 50 ml (10%) of heat inactivated FCS was added to each 500 ml of culture medium. The medium was then supplemented with final concentrations of 2 mM glutamine, 50 units ml^{-1} penicillin, and 50 µg ml^{-1}streptomycin.

Standard DMEM contains 4.5 g l^{-1} or 24.98 mM glucose. Standard RPMI-1640 contains 2 g l^{-1} or 11.10 mM glucose. For experiments where alteration of the glucose concentration in the culture medium was required, glucose free DMEM was purchased from Invitrogen (#11966025). Each 500 ml bottle of glucose-free medium was supplemented with 50 ml of heat inactivated FCS, 1 mM sodium pyruvate, 50 units ml^{-1} penicillin, and 50 µg ml^{-1}streptomycin. A stock 2M glucose solution was made in autoclaved dH_2O, and filter sterilised at 0.2 µm and

stored at 4 °C in a parafilm sealed 50 ml tube. For each individual experiment, a "master mix" of supplemented glucose-free DMEM and glucose at the required concentration was made up in a class II hood.

2.3.2 Standard cell culture

The primary human fibroblasts used in the experiments presented were adult human dermal primary untransformed fibroblasts, purchased from Invitrogen (#C-013-5C). Stock cell cultures were maintained in culture in T75 flasks, and passaged every 2-4 days, aiming to maintain a maximum confluence of ∼90% at any time. Passage of cells was carried out in a class II hood. For passage of cells, the culture medium was removed from the flask and discarded. The cells were washed with 1 ml trypsin-EDTA solution (Sigma, T3924), which was then removed and discarded. A volume of 2 ml trypsin-EDTA was added to the culture flask, which was placed in a standard incubator at 37 °C for 5 - 10 minutes. The trypsin-EDTA was neutralised by adding 8 ml fresh culture medium at room temperature. The resulting cell suspension was transferred to a 15 ml sterile centrifuge tube, and centrifuged at 830 g for 5 minutes. The supernatant was discarded, and the cells were resuspended in fresh culture medium. The cells were typically plated again, at a 1:10 to 1:20 dilution. Cell culture flasks were discarded and replaced at least every 14 days.

2.3.3 Cell storage in liquid nitrogen

For long-term storage, cells were placed in liquid nitrogen. Fresh culture medium was added to the cells on the day before initial freezing. On the day of initial freezing, a cryostore solution of 10% DMSO, 30%FCS in culture medium was made. The culture medium was removed from the cells. The cells were washed with 1 ml of trypsin-EDTA solution, and then incubated with 2 ml of trypsin-EDTA solution at 37 °C for 5 minutes. The trypsinization was stopped by adding 8 ml of fresh culture medium to the flask. The resulting cell suspension was centrifuged at 830 g for 5 minutes to collect the cells in a pellet. The supernatant was

discarded, and the cells were resuspended in the cryostore solution. This cell suspension was distributed in 1.5 ml aliquots to sterile cryovials (Fisher, cry-960-090S). Cryovials containing the cells were placed in a cryopot, filled with isopropanol, and stored overnight at −80 °C. After 24 hours at −80 °C, cryovials were transferred on dry ice to liquid nitrogen.

To re-culture cells stored in liquid nitrogen, cryovials were rapidly thawed by warming in a 37 °C water bath. A Pasteur pipette was used to quickly transfer the thawed cells from the cryovial to a 15 ml centrifuge tube containing 10 ml of fresh culture medium at room temperature. This cel suspension was centrifuged at 830 g for 5 minutes to collect the cells in a pellet. The supernatant was discarded, and the cell pellet was resuspended in 10 ml of fresh culture medium. This cell suspension was transferred to a new, sterile culture flask and placed in the incubator at 37 °C. The culture medium was replaced after 24 hours.

2.3.4 Chromium-release cytotoxicity assay

1×10^6 target cells for each test condition were transferred to a sterile 15 ml tube. The cells were washed in RPMI-1640/10% FCS, and centrifuged at 524 g for 5 minutes to pellet. The supernatant was discarded, leaving 100 µl covering the pellet. These target cell pellets were resuspended in 50 µl chromium-51 (^{51}Cr) (0.05 mCi or 1.85 MBq) per sample and incubated for 1 hour at 37 °C.

After 1 hour incubation with ^{51}Cr, target cells were washed twice with warm RPMI, and resuspended in 10 ml RPMI. The target cells were plated in quadruplicate for each E:T ratio, including two blank lanes and two maximum lysis lanes, in a round bottomed 96-well plate. Effector NK92 cells were resuspended at a concentration of 1×10^6 ml^{-1} in NK92 culture medium (RPMI, 10% FCS, 2 mM glutamine, penicillin/streptomycin solution, 1 mM pyruvate and 100 units ml^{-1} interleukin-2 (IL2)) and serial dilutions were prepared for the different E:T ratios.

The NK92 cells were added to the target cells in the 96-well plate, and mixed by gentle pipetting. NK92 culture medium was added at the same volume to blank lanes to estimate

spontaneous ^{51}Cr release, and 5% triton solution was added to maximum lysis lanes to estimate maximum possible ^{51}Cr release. Effector and target cells were incubated at 37 °C for 4 hours. After 4 hours co-incubation of effector and target cells, the 96-well plates were centrifuged at 524 g for 5 minutes to pellet intact cells and cellular debris. Supernatant from each well (15-25 µl) was mixed with 150 µl OptiPhase Supermix scintillation fluid (Perkin Elmer), before reading on a microbeta TriLux liquid scintillation counter.

For NKG2D blocking experiments, effector NK92 cells were initially resuspended in 500 µl NK92 culture medium, and an anti-NKG2D antibody (BD biosciences; 552866) was added at a concentration of 15 µg ml^{-1}. The NK92 cells were incubated at 4 °C for 30 minutes, before being resuspended in RPMI, and added to target cells as described previously.

2.3.5 CMV propagation

Typically, propagation of CMV was carried out using 40 × T175 (175 cm^2) culture flasks, containing fibroblasts at ∼80% confluence with 25 ml supplemented RPMI-1640. Propagation began with infection of a single T75 culture flask of 80% confluent fibroblasts with cytomegalovirus (CMV) at a multiplicity of infection (MOI) of 0.1. When diffuse cytopathic effect was observed, the supernatant was used to infect 2 × T175 culture flasks of fibroblasts at ∼80% confluence. When ≥ 90% of fibroblasts in the T175 flasks demonstrated a cytopathic effect, the supernatant was harvested, and used to infect 40 × T175 flasks. When these cells demonstrated ≥ 90% cytopathic effect, the supernatant (containing CMV) was removed and stored at −80 °C, and 25 ml of fresh culture medium was added to the flasks. This was repeated twice, to give a total of three sets of CMV-containing supernatant.

To concentrate the virus and minimise the amount of cytokines transferred with the virus, the CMV was then concentrated. The supernatants were defrosted, and centrifuged at 830 g to remove any cellular debris. The cleared supernatant, containing the virus, was transferred to sterile 250 ml ultracentrifuge tubes (maximum 200 ml supernatant per tube), and centrifuged at 22000 g for 120 minutes at room temperature. The supernatant was then

aspirated from the viral pellet, and discarded in Virkon. The viral pellets were resuspended in RPMI, aliquoted to cryovials, and stored at −80 °C.

2.3.6 CMV titration

The viral titre was estimated by looking for cytopathic effect in plated fibroblasts infected with serial dilutions of the concentrated virus. Typically, Fibroblasts were cultured in 3 × 24-well plates to ∼80% confluence. For the serial viral dilutions, 100 µl of concentrated virus was added to 9.9 ml of supplemented RPMI (1:100 dilution). 100 µl of the 1:100 dilution was added to 9.9 ml of supplemented RPMI to make a 1:10000 dilution of the concentrated supernatant. This was repeated to make a $1{:}10^{-6}$ dilution. Similarly, $1{:}10^{-7}$, $1{:}10^{-8}$, $1{:}10^{-9}$, and $1{:}10^{-10}$ dilutions of the concentrated supernatant were made. Ten wells each were infected with each of the supernatant dilutions, from $1{:}10^{-4}$ to $1{:}10^{-10}$, leaving 12 uninfected control wells. The plates were observed for cytopathic effect every 24 hours, and the titre was estimated when no new wells became infected.

2.3.7 Lentivirus production

Lentivirus particles were made in HEK293T cells. HEK293T cells were cultured to ∼70% confluence in a T175 flask, and fresh standard DMEM culture medium (25 ml) was added 24 hours before transfection. On the day of transfection, 15 µg of pMD2.G lentiviral packaging plasmid, 30 µg of psPAX2 packaging plasmid, and 45 µg of the pTripZ construct encoding the cDNA of interest were added to a 15 ml polypropylene tube, and topped up a volume of 2.185 ml with dH_2O. A second polypropylene tube as filled with 2.5 ml of 2x HEPES buffered saline (Sigma, #51558). 315 µl of 2 M $CaCl_2$ was added drop wise to the tube containing the plasmid DNA. This solution was then added drop wise to the 2x HEPES buffered saline solution, while aerating the tube with a second pipette to ensure homogeneity. The total solution was added drop wise to the T175 flask containing the HEK293T cells. At 6 - 9 hours following transfection, the culture medium was removed, and replaced with 20 ml of fresh

Table 2.12: DNaseI-seq: high salt nuclear lysis buffer.

Compound	Concentration
Potassium chloride (KCl)	50 mM
Magnesium sulphate ($MgSO_4$)	10 mM
Dithiothreitol (DTT)	3 mM
HEPES	5 mM
IGEPAL(R) CA-630 (I8896)	0.05%

DMEM.

At 48 hours, the culture medium now containing lentivirus particles was removed from the flask, and centrifuged at 1500 RPM for 5 minutes before filtering through a 0.22 µm filter to remove and HEK293T cell debris or other contaminants. The supernatant was then stored at −80 °C prior to use. Fresh culture medium was added to the flask, and the harvesting was repeated after a further 24 hours.

To concentrate the lentivirus particles from the HEK293T cell supernatant, the supernatants were ultracentrifuged at 19500 RPM for 140 minutes at 4 °C without a break. The resulting viral pellets were resuspended in 100 µl DMEM per T175 flask, and stored at −80 °C.

2.4 DNaseI-seq

The DNaseI-seq protocol described below was adapted and used with some modifications from a protocol originally developed by Marco De Gobbi and Jim Hughes of the Computational Biology Research Group at the Wellcome Institute for Molecular Medicine, Oxford.

2.4.1 Nuclear isolation and DNaseI digestion

Between 1 and 5 × 10^7 cells were collected by trypsinization, and centrifuged at 830 g for 5 minutes at 4 °C. Cells were washed twice in ice cold PBS, and resuspended for 2 minutes on ice in 947 µl of high salt nuclear lysis buffer **(Table 2.12)** with 40 µl Complete protease inhibitor cocktail (Roche, #11836145001), 13 µl of 100 mM PMSF.

The cells were then centrifuged at 930 g for 5 minutes at 4 °C, before being resuspended

Table 2.13: DNaseI-seq: RSB buffer.

Compound	Concentration
Sodium chloride (NaCl)	10 mM
Magnesium chloride ($MgCl_2.6H_2O$)	3 mM
Tris-HCl (pH7.4)	10 mM

Table 2.14: DNaseI-seq: lysis buffer.

Compound	Concentration
Sodium chloride (NaCl)	10 mM
EDTA	10 mM
Tris-HCl (pH7.4)	10 mM
SDS	0.05%

in 2ml RSB buffer **(Table 2.13)**. Cells were washed twice in RSB buffer and ultimately resuspended in 10 ml RSB buffer, which was aliquoted equally into 10 × 1.5 ml centrifuge tubes on ice. 0.5, 1, 2, 4, 8, 16, 32 or 64 units of DNaseI (47167280, Roche) were added to the tubes containing cell nuclei in RSB buffer, with two DNaseI-free tubes. The samples were moved to a heating block at 37 °C for ∼10 minutes. To judge when to stop digestion, 10 µl aliquots of the digesting samples were mixed with 10 µl of lysis buffer **(Table 2.14)** to release the genomic DNA. Poorly digested samples were very viscous, and became progressively water-like in consistency as digestion proceeded. When the sample containing 16 units of DNaseI became water-like in consistency, the reaction was stopped by transferring all tubes to ice. The contents of each tube was transferred to a separate 12 ml round-bottom centrifuge tube containing 3 ml of lysis buffer and 21 µl of proteinase K. The samples were incubated overnight at 37 °C.

2.4.2 Phenol-chloroform extraction and ethanol precipitation

A phenol-chloroform extraction was carried out on all samples on the day following DNaseI digestion. A volume of 4 ml of phenol chloroform (Sigma, P2069; pH8) was added to each sample. The samples were mixed by inversion and centrifuged at 2400 g for 30 minutes at room temperature. The upper aqueous phase was transferred to a new 12 ml round-bottom

centrifuge tube using a Pasteur pipette, mixed by inversion, and centrifuged again at 2400 g for 30 minutes. Again, the aqueous phase was transferred to a fresh tube using a Pasteur pipette. To precipitate the DNA fragments, 400 µl of a 4 M NaCl solution and 8 ml of 100% ethanol were added to the sample. The sample was then mixed by inversion and stored at −20 °C overnight. The next day, the samples were centrifuged at 2400 g for 30 minutes at 4 °C to pellet the DNA. The ethanol was removed and the sample resuspended in 1.5 ml of 70% ethanol, and transferred to a 1.5 ml tube. The samples were then centrifuged at 16000 g for 30 minutes at 4 °C. The excess ethanol was removed completely and allowed to air dry for 5 minutes. The samples were finally resuspended in 150 µl dH_2O, and allowed to dissolve at room temperature for 20 minutes. The DNA was finally stored at −20 °C.

2.4.3 Pulsed-field gel electrophoresis

Pulsed field gel electrophoresis was carried out in Ben Davies' lab at the Wellcome Trust Centre for Human Genetics, with kind assistance from group members. A 1% agarose gel was used, in 1× Tris/borate/EDTA (TBE) running buffer. A yeast chromosome mid range marker ladder, PFG I, was used on each gel (N3551, NEB). For DNaseI-digested and phenol-chloroform extracted samples, 10 µl of DNA was added to 2 µl of 6× loading dye. The gel was run for 18 hours, at 16 °C, with a switch time of 20-60 seconds, at a voltage of 6 V cm^{-1}. The gel was stained with ethidium bromide, and imaged in a standard gel imaging dock under ultraviolet light.

2.4.4 Assessment of digestion by real-time PCR

To relatively quantify the amount of digestion at each DNaseI concentration, we performed qPCR to amplify two DNaseI insensitive sites, and one DNaseI hypersensitive site **(Fig 6.5)**. The primer sequences for the DNaseI insensitive site in the 3' untranslated region of the rhodopsin gene on chromosome 3 were obtained from previous publications (213, 214). The primer sequences for the haemoglobin beta insensitive site, and the hypersensitive site 5' of

Table 2.15: DNaseI-seq: qPCR sites for assessing DNaseI digestion.

Gene	Chr	Size	Status	Forward	Reverse
rhodopsin	3	234	closed	ggacatccaccaagacctactg	atatgttctcccttcccattca
haemoglobin beta	11	270	closed	catagactcaccctgaagttct	ctctctctgcctattggtcta
c-Myc	8	271	open	cgccctcgttgacatccag	ttgcagctcagcgttcaagtg

the c-Myc gene were chosen on the basis of consistent DNaseI sensitivities across several cell lines and primary cells from ENCODE DNaseI-seq data **(Fig 6.5)**. The primer sequences are shown in **Table 2.15**.

To measure the digestion, the concentration of all DNaseI-digested samples was measured using a NanoDrop ND-1000 spectrophotometer. Based on this measured concentration, 1 µg of DNaseI-digested DNA was added to 900 µl dH_2O, to give a DNA concentration of 10 ng per 9 µl of dH_2O. For all three sites, and all DNaseI-digested samples, a qPCR reaction consisting of 9 µl (10 ng) DNA, 0.5 µl of forward and reverse primers (10 µM), and 10 µl of Fast SYBR Green Master Mix (Applied Biosystems, #4385612), was carried out. The qPCR samples were run in using standard relative quantification settings on a StepOnePlus thermocycler (Applied Biosystems). Digestion at the c-myc hypersensitive site was measured as the fold change in the c-myc target level, relative to either insensitive site, between DNaseI-digested samples. Correction for amplification efficiency was calculated by the Pfaffl method (209).

2.4.5 Library production

On the basis of relative quantification of DNaseI digestion by qPCR, 1.5 µg of DNA was taken from each four consecutive DNaseI digested samples (e.g. from 32 unit, 16 unit and 8 unit and 4 unit digestions). This DNA (total 6 µg) was pooled, and diluted to a total volume of 20 µl in dH_2O. To fill in any overlaps at the DNA fragment ends, a T4 DNA polymerase reaction was carried out at 12 °C for 1 hour to generate blunt-ended DNA fragments **(Table 2.16)**.

The sample was cleaned up using the Qiagen PCR purification kit, according to the manufacturers instructions, and eluted in 50 µl dH_2O. dATP overhangs were added to the

Table 2.16: T4 DNA polymerase blunt-end reaction.

Component	Amount
Pooled DNA	20 µl (6 µg)
T4 DNA polymerase	2.4 µl
NEB 2 buffer (10×)	20 µl
dNTP (10 mM)	5 µl
BSA (100×)	2 µl
dH_2O	150.6 µl
Total volume	200 µl

T4 DNA polymerase - NEB, M0203S.

Table 2.17: Klenow dATP overhang reaction.

Component	Amount
Blunt-ended DNA	25 µl
Klenow Exo minus	1 µl
Klenow buffer (10×)	5 µl
dATP (1 mM)	10 µl
dH_2O	9 µl
Total volume	50 µl

Klenow Exo Minus - Promega, M218A.

blunt ends in a Klenow reaction for 30 minutes at 37 °C **(Table 2.17)**.

The sample was then cleaned up again with the Qiagen PCR purification kit, according to the manufacturers protocol, and eluted to 30 µl of dH_2O. The Illumina sequencing adapters were ligated to dA tailed fragments in an overnight reaction at 16 °C **(Table 2.18)**, and cleaned up the next day using the Qiagen PCR purification kit, eluting to 30 µl of dH_2O.

Next, the library was PCR amplified in a 50 µl reaction **(Table 2.19)** for a total of 10

Table 2.18: Illumina adapter ligation reaction.

Component	Amount
dA-tailed DNA	15 µl
T4 ligase	1 µl
T4 ligase buffer (10×)	10 µl
Illumina PE adaptor mix (100 µM)	0.5 µl
dH_2O	22.5 µl
Total volume	50 µl

T4 ligase - NEB, M0202S.

Table 2.19: Library amplification reaction.

Component	Amount
Adapter-ligated library	10 µl
Phusion HotStart Taq	0.5 µl
Phusion buffer (5×)	10 µl
Illumina PE1 primer (25 µM)	1 µl
Illumina PE2 primer (25 µM)	1 µl
dNTP (10 µM)	1 µl
dH_2O	26.5 µl
Total volume	50 µl

Phusion HotStart Taq - NEB, F-540S.

Table 2.20: Library amplification cycling conditions.

Stage	Temperature	Duration
Initial denaturation:	98 °C	30 s
10 cycles of:		
Denaturation:	98 °C	10 s
Annealing:	65 °C	30 s
Elongation:	72 °C	15 s
Final elongation:	72 °C	5 min
Storage:	4 °C	∞

cycles **(Table 2.20)**. Finally, the library was size selected and purified by loading the entire library onto a 2% low range ultra agarose gel. DNA fragments in the size range of ∼200-500bp were excised from the gel, and gel purified using the Qiagen Gel Extraction kit, according to the manufacturers protocol. The amount of DNA in the purified, size selected library was quantified using the fluorescence based PicoGreen kit (Invitrogen, P7589), and concentrated to a volume of 10 µl by SpeedVac. Between 10 and 30 ng of the DNA library sent to the high throughput sequencing genomics facility at the Wellcome Trust Centre for Human Genetics, University of Oxford, for sequencing on a Illumina GAII high throughput sequencer.

2.5 Computational methods

2.5.1 General

The computational work was carried locally on a HP Optiplex 990, running Ubuntu 12.04 LTS, and a MacBook Pro, Mid-2009 running the latest OSX version. The sequencing data

was stored locally on two internal hard drives, and backed up to an external hard drive and also stored on the University's Hierarchical File Server (HFS), a backup service provided by the Oxford University IT Services (formerly Oxford University Computing Services). Several text editors were used, including Sublime Text 2, Fraise, and the built-in Ubuntu text editor, gedit.

2.5.2 Programs used

The thesis was written and compiled in TeXShop Version 2.47. Numbers for Mac was used to analyse data, and images were created using several programs: Bar charts were made with Graphpad Prism v6, and Open Office (Peakhunter chapter). Figures were principally made with Inkscape v0.48, and some also with Pixelmator v2.2 and Microsoft Word 2011. Flow cytometry data was analysed with FlowJo. The chemical structure images were from SigmaAldrich in **Fig 3.14**, and drawn with MarvinSketch v6.0.0 in **Fig 4.27**. The programs were written using Perl version 5.12 on OSX and Ubuntu. The basic functions of completed programs were checked for compatibility on a PC running ActivePerl version 5.12. No additional Perl modules were used or are required to run our programs. The current version of PeakHunter is available for download (free of charge) from www.ccmp.ox.ac.uk/peakhunter. Other programs used for analysis of DNaseI-seq data are shown on **Table 2.21**.

Table 2.21: Programs used for data analysis.

Title	Version	Usage	URL
PeakHunter	v1.0	Manipulation of mapped read files, peak calling.	http://www.ccmp.ox.ac.uk/peakhunter/
Stampy	v1.0.20	Short read mapping.	http://www.well.ox.ac.uk/project-stampy/
BWA	v0.6.2	Short read mapping.	http://bio-bwa.sourceforge.net/
HOMER	v4.2	Peak calling.	http://biowhat.ucsd.edu/homer/
MACS	v1.4.2	Peak calling.	http://liulab.dfci.harvard.edu/MACS/
FSEQ	v1.84	Peak calling.	http://fureylab.web.unc.edu/software/fseq/
SAMtools	v1.4	Manipulation of SAM/BAM format files.	http://samtools.sourceforge.net/
UCSC utilities	v.Current	Manipulation of browser format files.	http://genome.ucsc.edu/util.html
NCBI SRA Toolkit	v2.2.0	Manipulation of SRA format files.	http://www.ncbi.nlm.nih.gov/Traces/sra/

3

The effect of glucose on NKG2D ligand expression

Contents

3.1 Abstract

The expression of NKG2D ligands on infected, malignant and proliferating cells increases cellular susceptibility to cytotoxicity mediated by NKG2D receptor expressing immune cells, including NK cells, $\gamma\delta$ T cells, NKT cells and $CD8^+$ $\alpha\beta$ T cells. It is unclear how cellular expression of NKG2D ligands is controlled at the molecular level. Increasingly evidence

suggests that settings in which NKG2D ligand expression is observed are also associated with distinctive metabolic changes, including increased glucose uptake, aerobic glycolysis and increased lactate production, known as the 'Warburg effect'. To examine how Warburg metabolism might influence NKG2D ligand regulation, we have established a simple HEK293T cell model of Warburg metabolism. We carry out a preliminary characterisation of this model, and demonstrate the role of glucose in NKG2D ligand expression. We show that in order for glucose to influence NKG2D ligand expression, glucose must be taken into the cell and metabolised, and that the changes in NKG2D ligand expression induced by glucose are functionally significant.

3.2 Introduction

3.2.1 The case for exploring the role of glucose

The NKG2D ligands are a group of eight immune proteins whose expression on the surface of a cell confers susceptibility to the effector functions of NKG2D receptor-bearing immune cells, including NK cells, $\gamma\delta$ T cells, NKT cells and $CD8^+$ $\alpha\beta$ T cells. These ligands are not typically expressed on healthy cells, but are expressed in a range of settings where cellular metabolism changes to a high glucose consumption state, referred to as the Warburg effect. These settings include infection, malignant transformation, cellular proliferation and toll-like receptor (TLR) activation (for references, see **Table 1.3**).

Despite the powerful nature of the immune response controlled by NKG2D ligand expression, the molecular mechanisms that regulate NKG2D ligand expression in these settings are unknown. Based on the striking co-occurrence of NKG2D ligand cell surface expression and the highly-glycolytic metabolic phenotype, we hypothesised that the mechanism regulating NKG2D ligand expression was dependent on these metabolic changes. To explore this hypothesis, we developed a HEK293T cell model of Warburg metabolism.

3.2.2 Glucose is known to affect gene regulation

Glucose is known to have widespread influences on gene expression. Several studies have examined differential microarray profiles of cells cultured in graded glucose concentrations (215, 216, 217). Each of these studies has found changes, both increases and decreases, in expression levels of hundreds of genes. However, for many of these genes, the mechanism by which glucose increases or decreases its expression is unclear. An example of this is the regulation of the insulin gene expression by glucose, a subject of intense study for several decades. Despite this research, significant uncertainty still surrounds the mechanism of its regulation (reviewed by MacDonald *et al* (218)).

3.2.3 Four potential categories of glucose activity

As a framework for identifying the basic mode of action of glucose in controlling NKG2D ligand expression, we considered four possible categories of activity, based on previously described regulatory mechanisms: that glucose might act from outside the cell, by affecting osmolality, or binding cell surface receptors, or that it was acting from inside the cell either as a structural ligand (the glucose molecule itself alters gene expression), or generating metabolites of glucose which in turn alter gene expression independently of the glucose molecule itself **(Fig 3.1)**.

Changes in extracellular glucose concentrations result in changes in osmolality, a measure of the density of solute particles in a solution. Osmolar stress is known to result in changes in gene expression, the most prominently studied example of this being the aquaporin water channel genes (*AQP1, AQP2*) expressed in renal medullary epithelial cells (219, 220). Cell surface receptors for glucose have not been identified in mammals, but have been found, and well characterised in yeast (reviewed by Özcan *et al* (221) and Zaman *et al* (222)). Signalling from these receptors helps to coordinate the changes in gene expression necessary to alter the metabolic patterns in yeast when glucose becomes available in the environment.

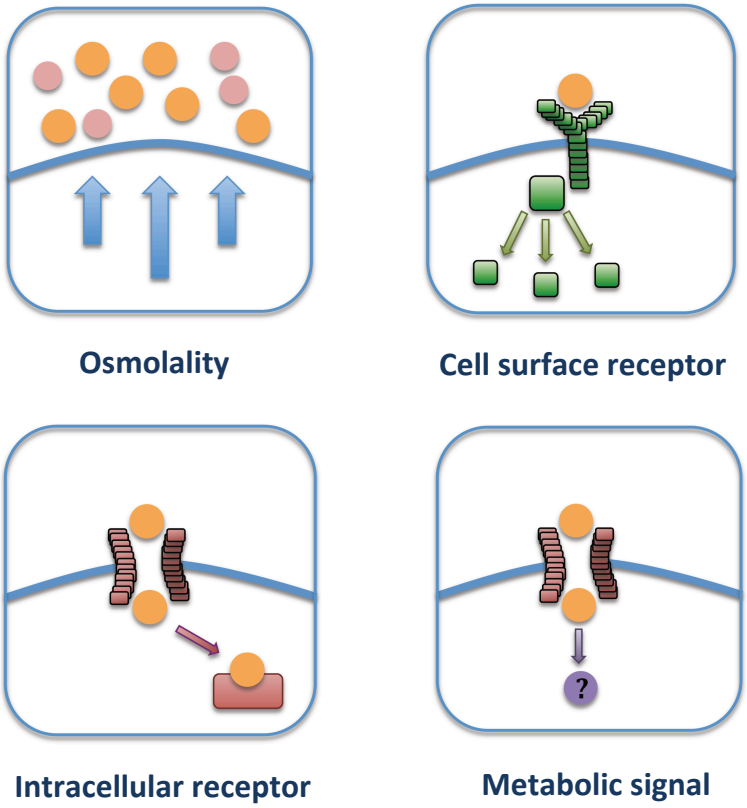

Figure 3.1: Potential mechanisms of gene regulation by glucose. To aid the exploration of potential glucose-mediated effects on NKG2D ligand regulation, we have divided known mechanisms of glucose mediated gene regulation into four categories. Changes in extracellular osmolality, which arise in the setting of changing extracellular glucose concentrations, are known to alter the expression of genes such as the aquaporin genes of renal medullary epithelium. Cell surface glucose receptors with intracellular signalling machinery are well characterised in yeast, but not known to exist in humans. Intracellular glucose or glucose 6-phosphate is known to act as a ligand itself, influencing gene transcription through direct interaction with transcription factors such as the carbohydrate response element binding protein (ChREBP). Finally genes such as insulin are thought to be regulated through mechanisms involving the detection of changes in metabolites of glycolysis.

A third potential mechanism is the transport of glucose into the cell, and direct binding of regulatory proteins by glucose, in turn leading to altered gene expression. An example of this is glucose 6-phosphate binding to the carbohydrate response element binding protein (Li *et al* (223) and reviewed by Filhoulaud *et al* (224)). Finally, we considered the possibility that glucose is taken into the cell and metabolised, and NKG2D ligand regulation may be influenced by down-stream products of glycolysis, as is thought to be the case with insulin (reviewed by MacDonald *et al* (218)).

3.2.4 Aims and objectives

To establish the wider relevance of glucose availability to NKG2D ligand expression, we first assess the impact of glucose supply on cell surface NKG2D ligand expression in a range of cell lines, and determine the functional significance of observed changes by measuring changes in NKG2D dependent cellular susceptibility to NK cell cytotoxicity. To explore the hypothesis that glucose or its metabolites directly influences the regulation of NKG2D ligand cell surface expression, we develop a simple cellular model of Warburg metabolism. We aim to characterise this cellular model, and use this to explore observed regulatory effects in terms of known mechanisms of glucose mediated-gene regulation outlined above.

3.3 Results

3.3.1 Glucose induces cell surface expression of MICA in cell lines

A key feature of Warburg metabolism is increased expression of cell surface glucose transporters, and increased glucose transport into the cell. To see if glucose uptake in Warburg metabolism affected NKG2D ligand expression, we cultured three different cell lines, including the non-malignant human embryonic kidney (HEK) cell line HEK293T, the cervical carcinoma cell line HeLa, and the breast cancer cell line MCF7, in a range of glucose concentrations for 48 hours, and measured cell surface MICA expression by flow cytometry. Because

commonly used *in vitro* mammalian cell culture media, such as DMEM and RPMI, contain 4.5g/L (25 mM) and 2.0g/L (11.1 mM) glucose respectively, to generate the culture media with different glucose concentrations, we used glucose-free DMEM with supplemental glucose. We found that in each cell line, cell surface MICA expression was proportional to the glucose concentration in the culture medium, with highest cell surface MICA expression occurring at 25mM glucose **(Fig 3.2)**.

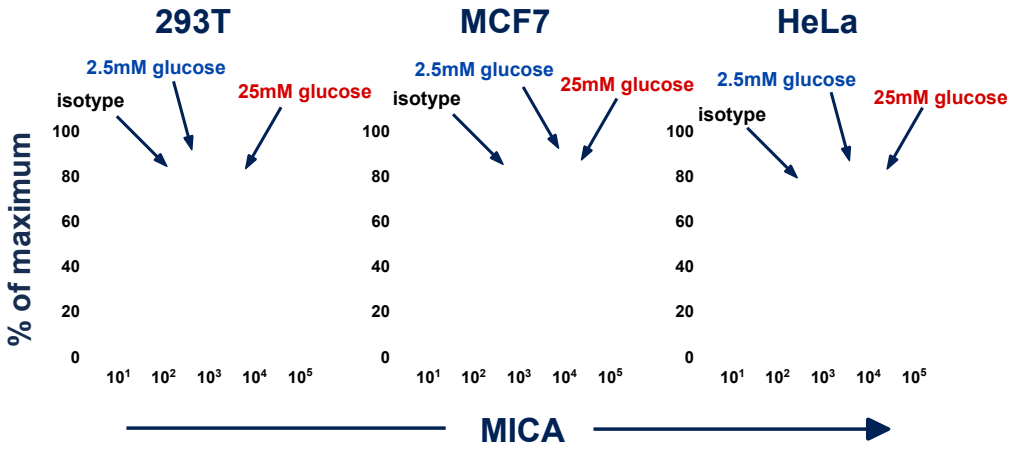

Figure 3.2: Glucose availability determines level of MICA cell surface expression. MCF7 cells, HeLa cells and 293T cells were cultured in 5mM glucose for 48 hours, and then transferred to wells containing DMEM with a range of glucose concentrations from 2.5mM glucose to 25mM glucose, and cultured for a further 48 hours. The cells were stained with anti-MICA or isotype control antibodies and measured by flow cytometry. For each cell line, cell surface MICA expression increased with increasing concentrations of glucose. This data is a representative experiment of three biological replicates.

3.3.2 Glucose increases NKG2D-dependent cell line immunogenicity

Key NKG2D effector functions, such as NK cell cytotoxicity, are dependent on a broad balance of activating and inhibitory receptor expression on the target cell surface. While we had observed glucose-induced MICA expression on three different cell lines, we wanted to know whether the net effect of glucose on these target cells resulted in significant changes in NKG2D-dependent effector functions. To determine if this was the case, we cultured each cell line for 48 hours in a range of different glucose concentrations (as described above), and measured glucose-dependent changes in susceptibility to NK cell meditated cytotoxicity, in

a series of ^{51}Cr release cytotoxicity assays. We found that cells cultured in high glucose concentrations, expressing higher levels of MICA, were more susceptible to NK cell mediated cytotoxicity **(Fig 3.3, *left*)**.

While it was clear that glucose increases the NK cell immunogenicity of these cell lines, we wanted to see if any of this killing was dependent on NKG2D signalling. Cells were cultured in either low (5mM) of high (25mM) glucose for 48 hours as above, and prior to carrying out the chromium release assay, the NKG2D receptor on the effector NK cells was blocked by incubation with an anti-NKG2D antibody. A greater amount of killing was prevented at high glucose concentrations compared to low glucose concentrations, showing that glucose increased the amount of NKG2D-dependent killing across these cell lines **(Fig 3.3, *right*)**.

3.3.3 Characterisation of the effect of glucose on HEK293T cells

Having demonstrated glucose-induced, immunogenicity altering cell surface NKG2D ligand expression in three different cell lines, we wanted to explore this relationship in greater detail, to try to understand how glucose was causing this effect. Because HEK293T cells were the only non-malignant cell line, and demonstrated the most pronounced changes in MICA expression level, and susceptibility to glucose-dependent NK cell cytotoxicity, we selected HEK293T cells for more detailed study.

3.3.3.1 Glucose affects HEK293T cell proliferation

We anticipated that altering glucose supply to cell lines may also alter the rate of proliferation of these cells. To characterise the baseline influence of glucose on relative cell proliferation in HEK293T cells, we carried out an epithelial cell CFSE proliferation assay **(Fig 3.4)**. Unlike traditional CFSE proliferation assays in cell populations with discrete and coordinate generations of proliferation, such as activated lymphocytes, in continuously dividing epithelial cells, proliferation leads to a progressive and continuous reduction in CFSE fluorescence in proliferating populations. No undivided CFSE^{high} cell populations remain.

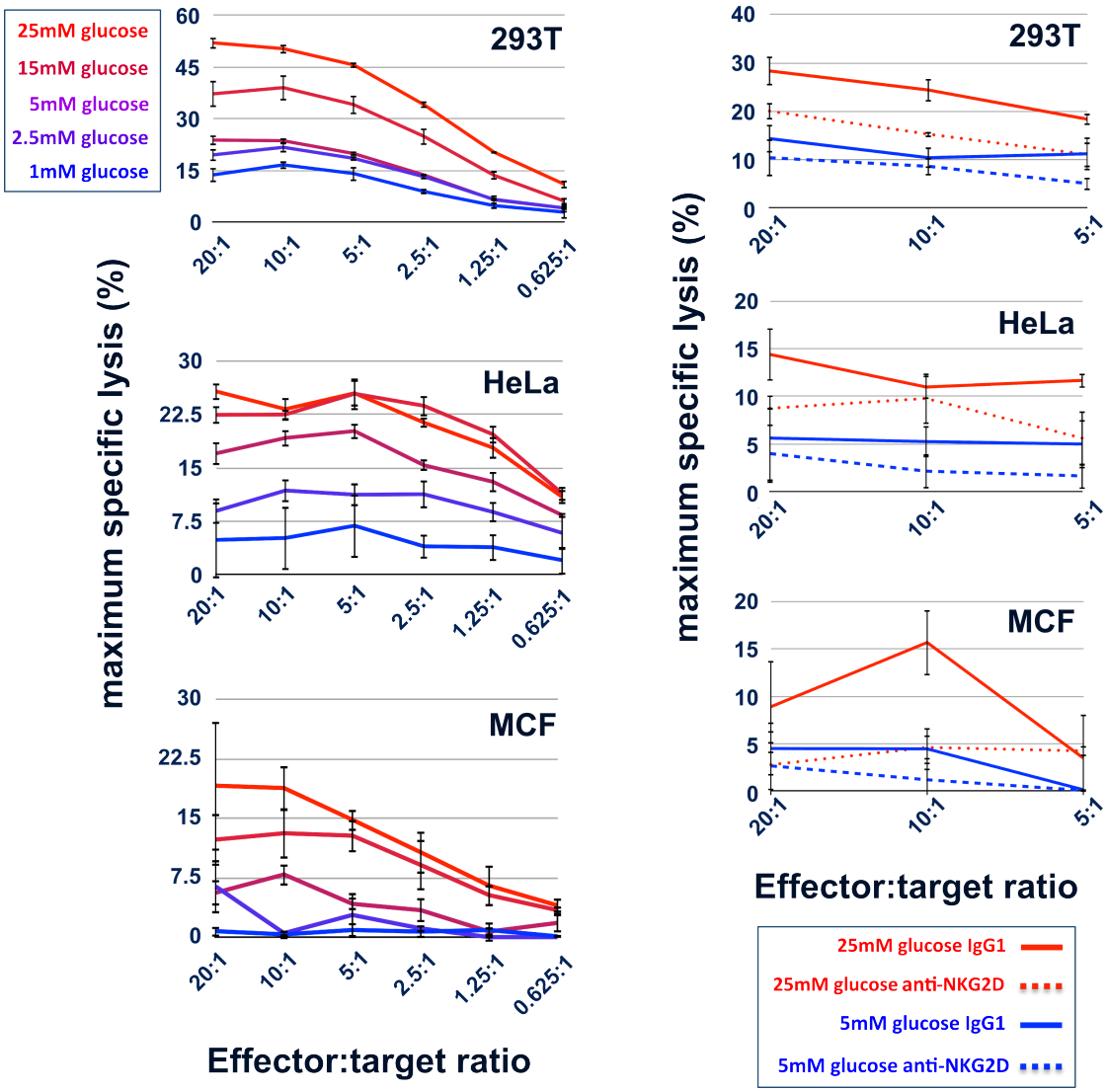

Figure 3.3: Glucose changes cellular immunogenicity in an NKG2D-dependent manner. ***Left:***Chromium release cytotoxicity assays were carried out using target cells cultured in a range of glucose concentrations from 1mM glucose to 25mM glucose. Following culture for 48 hours in different glucose concentrations, the target cells were incubated for 1 hour with chromium-51 (^{51}Cr). Excess ^{51}Cr that had not entered the cells was washed off, and these ^{51}Cr labelled cells were incubated with NK effector cells for four hours. The amount of chromium released into the supernatant after the incubation period is measured. To calculate the amount of effector cell-mediated ^{51}Cr release, background measurements (target cells incubated with no effector cells) are subtracted from effector cell-incubated measurements. The maximum possible ^{51}Cr release is found by incubating target cells with 5% triton solution. The percentage of maximum specific lysis (as displayed) is the amount of effector specific lysis as a percentage of maximum possible lysis. Here the percentage of maximum specific lysis was proportional to the glucose concentration for each cell line. ***Right:*** NKG2D specific lysis is calculated by pre-incubating effector NK cells with an anti-NKG2D antibody, blocking NKG2D dependent signalling. The amount of killing blocked by preventing NKG2D signalling was greater in high glucose, compared to that in low glucose, suggesting that glucose increased the NKG2D-dependent killing of target cells by NK effector cells.

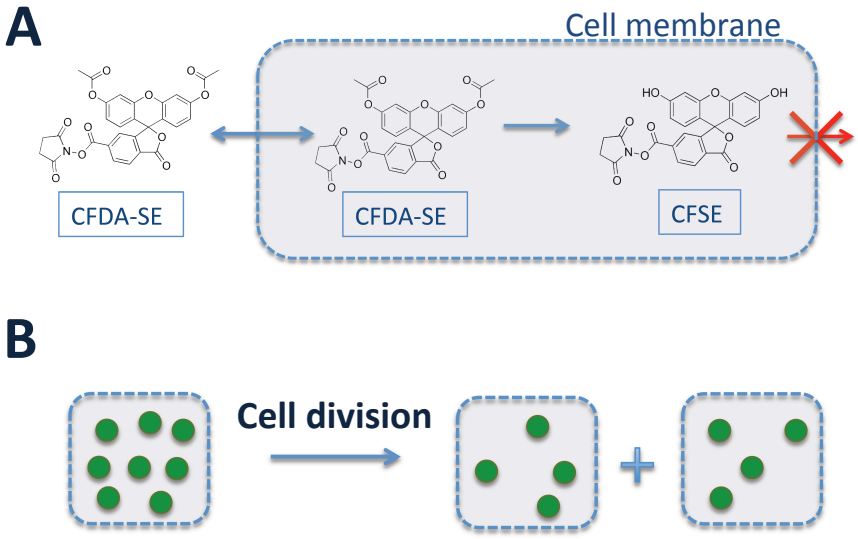

Figure 3.4: The CFSE proliferation assay. (A) CFDA-SE is a cell permeable fluorescent molecule. Upon incubation with mammalian cells, CFDA-SE passes freely into the cell, where the diacetate group is rapidly hydrolysed, generating CFSE. CFSE is much less permeable, trapping it inside the cell. (B) When CFSE is trapped within a cell, the only effective means of reducing intracellular CFSE concentration is through dilution by cell division. Because epithelial cell lines divide asynchronously and continuously, as a population, dilution of CFSE through cell division is represented by continuous and progressive shift of CFSE proliferation towards zero, with no discreet $CFSE^{high}$ populations remaining.

By measuring CFSE fluorescence 48 hours after staining HEK293T cells with CFSE and culturing the CFSE-stained cells in medium containing different glucose concentrations, we found that glucose altered the rate of HEK293T cell proliferation **(Fig 3.5a)**. The maximum rate of proliferation was seen at intermediate (12.5 mM) glucose concentrations. The reliability of the CFSE assay for estimating proliferation was verified by preparing HEK293T cells under identical conditions, and counting cells manually (using a hemocytometer) at 48hrs **(Fig 3.5a)**. The manual cell counts corresponded well to the CFSE fluorescence measurements.

3.3.3.2 Glucose does not affect the HEK293T cell cycle distribution

Given that changing glucose concentrations altered the rate of proliferation, and that disruption of major cellular processes such as transcription (225) can affect the cell cycle and vice versa (reviewed by Bregman *et al* (226)), we wanted to see if our stimulus disrupted cell

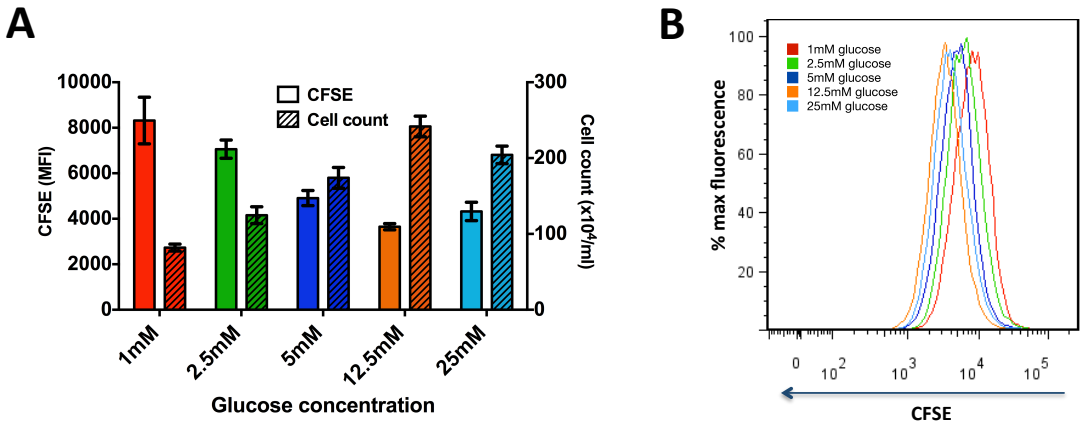

Figure 3.5: HEK293T CFSE proliferation assay: glucose affects proliferation. (A) To measure relative glucose-dependent changes in HEK293T cell proliferation, cells were cultured in 5 mM glucose for 48 hours, stained *in situ* with 5 µM CFDA-SE, and cultured for 48 hours in DMEM containing the glucose concentrations indicated. After 48 hours, cells were analysed by flow cytometry for CFSE fluorescence. The mean and 95% confidence interval of three CFSE mean fluorescence intensities is indicated. For cell counts, cells were prepared identically, and counted manually by hemocytometer after 48 hours. (B) Representative CFSE frequency distributions are shown from the data used to calculate CFSE mean fluorescence intensities in (A). With increasing rates of cell proliferation, the CFSE intensity frequency distribution shifts continuously to the left.

cycle progression in HEK293T cells.

When cells were cultured in DMEM containing low (5 mM) or high (25 mM) glucose concentrations for 48 hours, and then fixed, permeabilised and stained with propidium iodide, analysis of cell cycle phase distribution was carried out by flow cytometry, using FlowJo modelling software. We demonstrated that glucose availability does not alter cell cycle distribution in HEK293T cells **(Fig 3.6)**, despite the effect observed on proliferation.

3.3.3.3 Glucose alters expression of several NKG2D ligands

It was clear that glucose influenced cell surface MICA expression. We wanted to see if other ligands for NKG2D were similarly affected by glucose. We cultured HEK293T cells in 5 mM glucose for 48 hours, removed the culture medium, and replaced it with medium containing either 5 mM or 25 mM glucose, and cultured the cells for a further 48 hours. Measuring surface protein expression by flow cytometry, we found that like MICA, ULBP1, ULBP2 and ULBP3 expression was increased in medium containing higher glucose concentrations, while

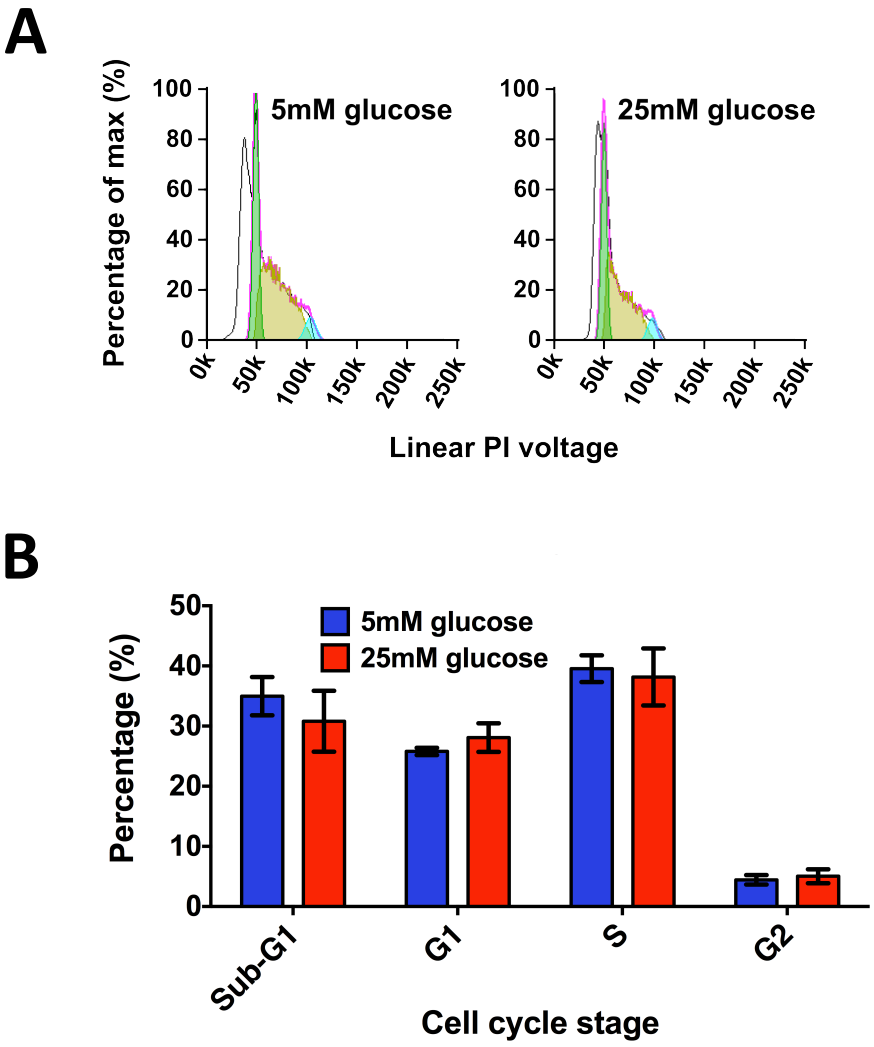

Figure 3.6: Cell cycle analysis of HEK293T cells in glucose. (A) To see if glucose availability affected the cell cycle distribution in cell lines, we cultured cells in 5 mM glucose for 48 hours, and then in 5 mM or 25 mM glucose for a further 48 hours. Propidium iodide based cell cycle analysis was carried out by fixing cells with paraformaldehyde, and permeabilising cells with 70% ethanol. RNA was digested with RNase, and DNA was stained with propidium iodide. Propidium iodide fluorescence was measured in linear signal collection mode by flow cytometry, and cell cycle phase modelling was carried out using FlowJo. A single profile for each glucose concentration is shown, representative of three biological triplicate samples analysed. (B) Statistical analysis of cell cycle phase percentages, carried out in biological triplicate, showed no significant differences in cell cycle phase distribution.

ULBP5 expression was consistently lower in higher glucose concentrations **(Fig 3.7)**.

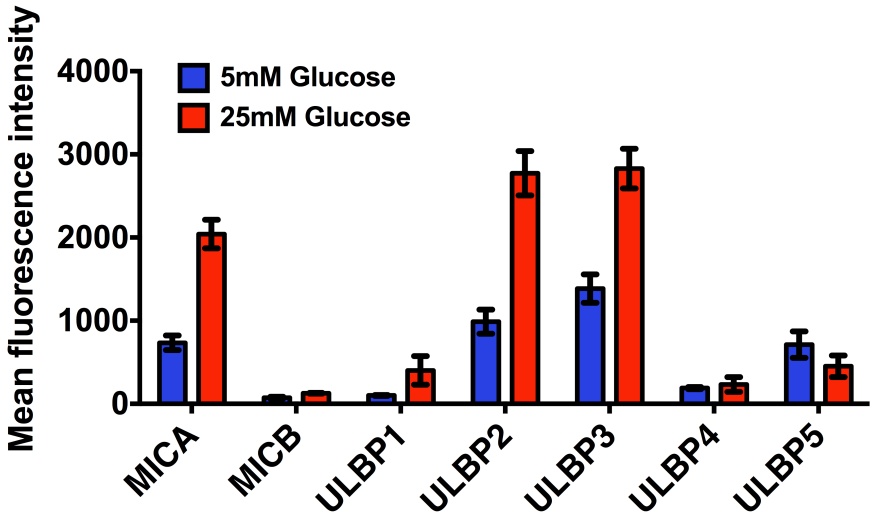

Figure 3.7: Several NKG2D ligands are affected by glucose. We measured the cell surface expression of other NKG2D ligands by flow cytometry in 293T cells cultured in either low (5 mM) or high (25 mM) glucose. In addition to driving increased MICA expression, glucose also induced cell surface expression of ULBP1, ULBP2 and ULBP3.

3.3.3.4 Glucose does not affect expression of all cell surface proteins

While we expected glucose to have diffuse effects on gene expression, we were somewhat concerned that glucose was having a generic effect on cell surface protein levels. To ascertain whether the increase in expression applied to a wider range of cell surface proteins, we cultured HEK293T cells in a finer gradient of glucose concentrations, and measured cell surface MICA, HLA class I (A, B and C), podoplanin (an unrelated cell surface signalling molecule) and ULBP5. While HLA class I and podoplanin cell surface expression also increased with glucose, ULBP5 expression decreased progressively and significantly with increasing glucose concentration **(Fig 3.8)**. The effect of glucose on ULBP5 confirms that glucose is not acting generically on all cell surface molecules.

3.3.3.5 Glucose does not affect expression of all cellular proteins

A second concern about the glucose-NKG2D cellular model, was that increasing glucose availability may act to alter gene transcription through generic mechanisms such as increasing

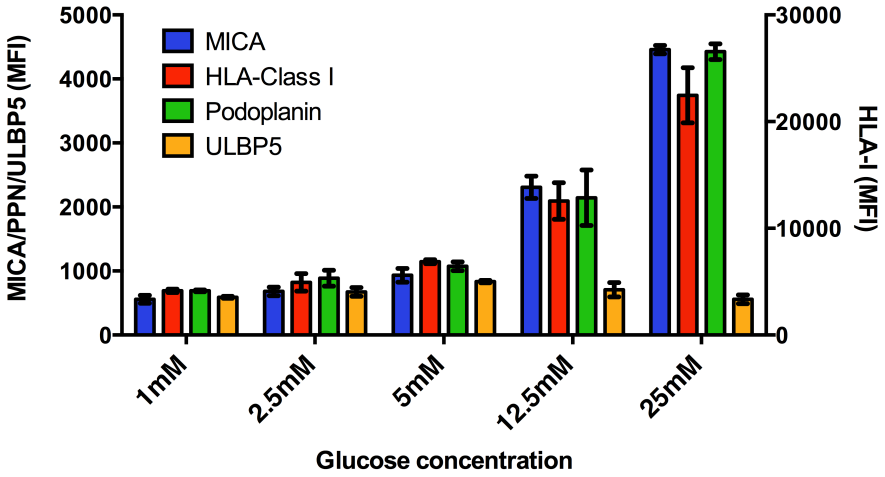

Figure 3.8: Not all cell surface proteins are increased by glucose. We measured the cell surface expression of MICA, HLA class I (A, B and C), immune signalling molecule podoplanin, and ULBP5 on the surface of HEK293T cells cultured in a glucose gradient by flow cytometry, While the levels of MICA, HLA class I and podoplanin increased proportionately with glucose, high levels of glucose suppressed ULBP5 cell surface expression.

the availability of nucleotides for the transcription process. To see if this were the case, we wanted to examine the affect of glucose on the expression of a gene over which we could exert precise transcriptional control. To achieve this, we made a doxycycline inducible lentiviral vector encoding eGFP. We infected HEK293T cells, and selected them with puromycin for ten days. The selected cells were then transferred to puromycin free DMEM containing 5mM glucose for 48 hours. The cells were split into low or high glucose, with or without $2\,\mu g\,ml^{-1}$ doxycycline, and cultured for a further 48 hours. The cells were then stained with anti-MICA antibody and analysed by flow cytometry. While MICA cell surface expression changed as expected with changing glucose concentration both with and without doxycycline, the expression of eGFP was not influenced by glucose concentration **(Fig 3.9)**.

3.3.4 Glucose increases total cellular MICA protein

Measured increases in protein at the cell surface doesn't necessarily imply an increase in total cellular protein. The amount of protein detectable at the cell surface can change as a result of impaired protein trafficking to the cell surface, as has been reported for MICA (148). To see if total cellular MICA changed with increasing glucose availability, we compared MICA

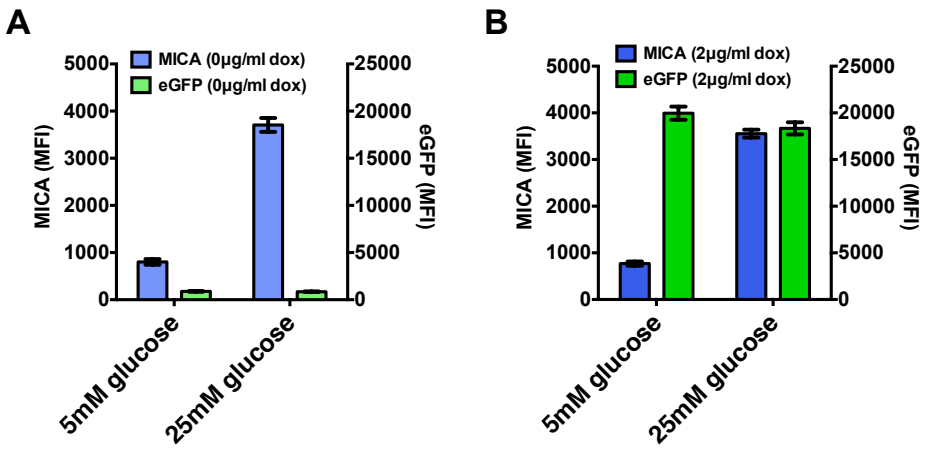

Figure 3.9: Glucose does not affect expression of all cellular proteins. To assess whether glucose globally influences protein expression of all transcribed genes, we infected HEK293T cells with a doxycycline inducible lentivirus encoding eGFP, and cultured infected cells in low (5 mM) or high (25 mM) glucose, with (A) or without (B) a set concentration of 2 µg ml^{-1} doxycycline. While glucose induced cell surface expression of MICA, it did not effect expression of eGFP under doxycycline control.

fluorescence on permeabilised and non-permeabilised cells across a range of glucose concentrations. We found that MICA levels increased in proportion to glucose concentration in both conditions **(Fig 3.10a)**. To ensure that cells had been permeabilised fully, we also stained permeabilised and non-permeabilised cells for proliferating cell nuclear antigen (PCNA), a nuclear protein. PCNA was not detectable on non-permeabilised cells, but was strongly detected in permeabilised cells, demonstrating successful permeabilisation to the nucleus **(Fig 3.10b)**.

3.3.5 Varying glucose supply mimics the 'Warburg' metabolic phenotype

Warburg metabolism is characterised by increased glucose uptake, rapid glycolysis, and increased lactate production. We wanted to see if our cellular HEK293T cell model reproduced these metabolic traits. Aliquots of DMEM containing 5 mM or 25 mM glucose were made. HEK293T cells were cultured in 5 mM glucose for 48 hours. The culture medium was removed, and 5 mM or 25 mM glucose-containing medium was added to these plates, and to empty plates, which were incubated together for a further 48 hours. Samples of the culture medium were taken at 48 hours, and sent to the Department of Biochemistry at the John

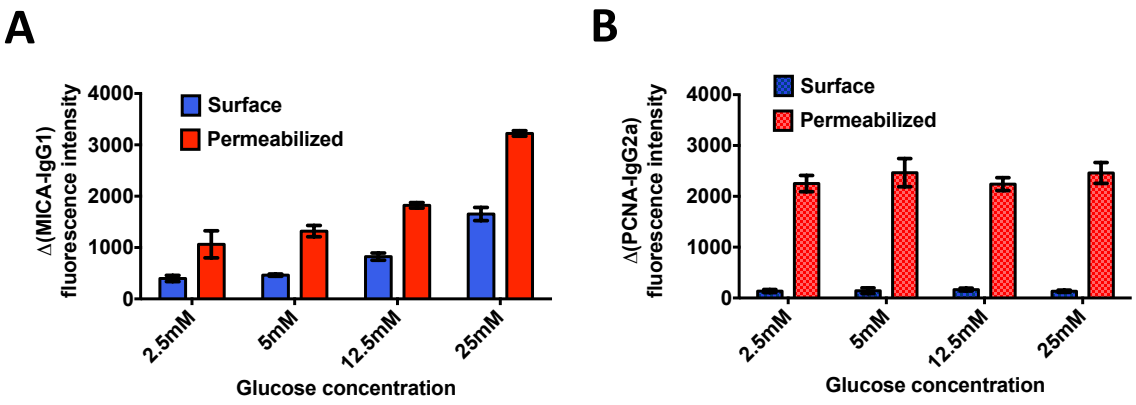

Figure 3.10: Glucose increases total cellular MICA protein. We compared the effect of glucose on total cellular MICA expression with cell surface MICA expression using cell-permeabilised flow cytometry. HEK293T cells were cultured in 5 mM glucose for 48 hours, before being transferred to a glucose gradient as indicated for a further 48 hours of culture. The cells were harvested, and cells from each glucose concentration were split into two portions, and stained with primary antibodies either before or after permeabilisation with 0.1% saponin. Antibody specific fluorescence was calculated by subtracting isotype fluorescence from primary antibody fluorescence. (A) Staining for MICA showed that both cell surface MICA and total cellular MICA levels increased with increasing glucose concentration. (B) As a positive control for permeabilisation, the same populations were stained with an antibody specific for the nuclear antigen, PCNA (proliferating cell nuclear antigen). While this demonstrates that cells were permeabilised to the nucleus, it also suggests that not all proteins are affected by glucose concentration.

Radcliffe Hospital, Oxford for glucose and lactate measurements. This showed that cells cultured in high glucose consumed more glucose, and produced more lactate during the culture period than cells cultured in low glucose, a pattern consistent with Warburg metabolism **(Fig 3.11)**.

In summary, these experiments demonstrate that glucose-induced NKG2D ligand expression in HEK293T cells represents a useful, metabolically accurate cellular model of 'Warburg-induced' NKG2D ligand expression. The induction of NKG2D ligand expression in this model results in increased NKG2D-dependent NK cell-mediated cellular immunogenicity. While the rate of cellular proliferation is altered by glucose, the distribution of cells through the cell cycle is unaffected. The effect of glucose on protein expression is not restricted to NKG2D ligands, but changing glucose concentration does not have a global effect on cell surface or cellular protein production. Additionally, the effect of glucose in increasing MICA protein expression is absolute at the whole cell level, and not limited to effects on cellular distribution of MICA.

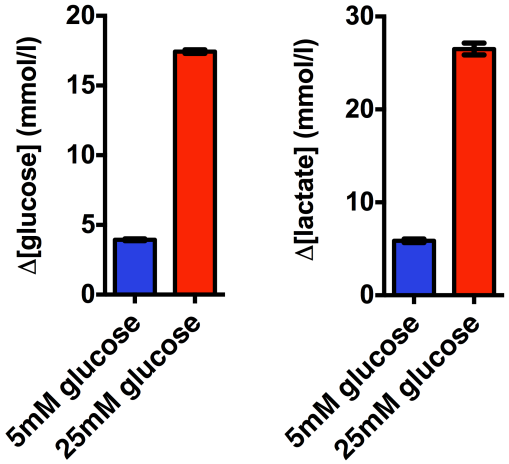

Figure 3.11: Culture of HEK293T cells in a glucose gradient mimics a Warburg metabolic phenotype. HEK293T cells were cultured in DMEM containing low (5 mM) or high (25 mM) glucose for 48 hours, and the change in glucose and lactate concentrations in the culture media was measured. Cells cultured in high glucose display increased glucose utilisation, and increased lactate production, similar to that observed in 'Warburg metabolism'.

3.3.6 Osmolality does not affect MICA expression

Having characterised the HEK293T cell model of NKG2D ligand induction, we next wanted to address some of the basic questions surrounding the mechanism of action of glucose in driving NKG2D ligand expression. Namely, whether glucose was acting from outside or inside the cell and whether its intracellular metabolism was necessary to mediate these effects **(Fig 3.1)**. We began by addressing the simple question of osmolality.

Osmolality is a measurement of the density of solute particles in a solution. When two solutions of different osmolality are placed on either side of a semipermeable membrane, water will move from the solution with lower osmolality to that with higher osmolality. The normal osmolality of human extracellular fluid is 275 – 295 milliosmole per kilogram of fluid. We expected each 1 mM of glucose added to DMEM to increase the osmolality of the culture medium by 1 mOsm kg^{-1}. To adjust the osmolality of culture medium by similar amounts without using glucose, we added 1 mM mannitol or 0.5 mM NaCl for each 1 mOsm kg^{-1} adjustment in osmolality required. We sent samples of our prepared osmolality-adjusted culture media to the Department of Biochemistry at the John Radcliffe Hospital, Oxford, to verify our calculated osmolalities. Osmolalities were measured using the method of freezing

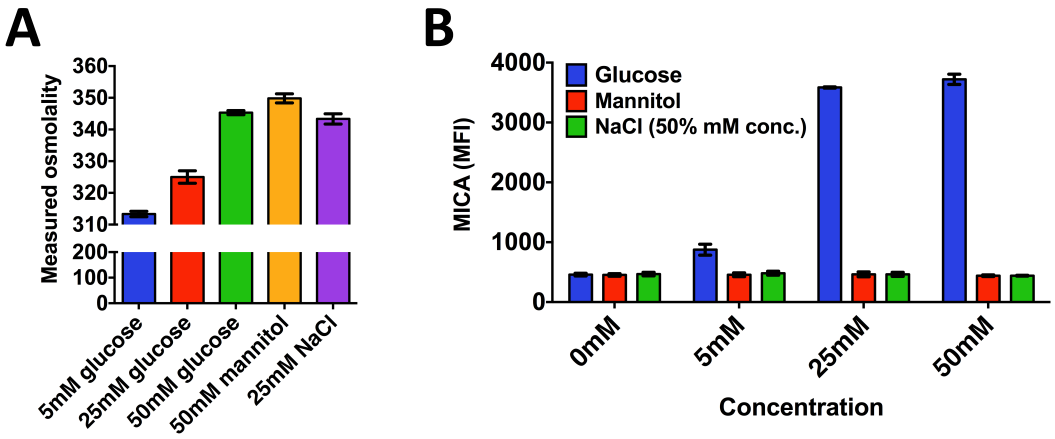

Figure 3.12: Osmolality does not affect NKG2D ligand expression. (A) To verify the actual osmolality of our culture media solutions, we sent samples to the Department of Biochemistry at the John Radcliffe Hospital for measurement by freezing point depression. The measurements agreed with the calculated osmolalities that we had expected. (B) HEK293T cell sere placed in medium containing increasing concentrations of glucose, and cultured along side cells grown in medium containing either mannitol or sodium chloride at equivalent osmolalities. Only glucose altered MICA expression, suggesting that osmolality was not responsible for the changes observed.

point of water depression. We found a good correlation between our calculated osmolality, and measured osmolality **(Fig 3.12a)**. We next determined if the change in osmolality caused by adding glucose to DMEM was responsible for the change in MICA expression, by culturing HEK293T cells in DMEM with increasing concentrations of glucose, or iso-osmolal amounts of mannitol or NaCl. Only glucose increased MICA expression, suggesting that changes in osmolality were not responsible for this outcome **(Fig 3.12b)**.

3.3.7 Uptake and metabolism of glucose is necessary for NKG2D ligand expression

3.3.7.1 Inhibition of GLUT transporters limits the effect of glucose

In order to determine if glucose was acting inside the cell rather than from outside the cell, we first assessed the affect of inhibiting glucose transporters on the expression of MICA in response to glucose. Glucose transport in mammalian cells is complex and still incompletely understood. There are two groups of glucose transporters, the facilitative trans-

porters (GLUT proteins; *SLC2A* gene family), and the sodium co-transporters (SGLT proteins; *SLC5A* gene family). The co-transporters are thought to be responsible for glucose absorption/re-absorption in intestinal epithelial cells and renal tubule epithelial cells. The facilitative transporters are a growing family, and are thought to be responsible for the uptake of 'metabolisable' glucose into cells throughout the body (for a review of the facilitative transporters and their inhibitors, see Augustin *et al* (227)). We used two well characterised chemical inhibitors of facilitative glucose transporters, phloretin and cytochalasin B to assess the role of glucose transport in MICA expression **(Fig 3.13a)**.

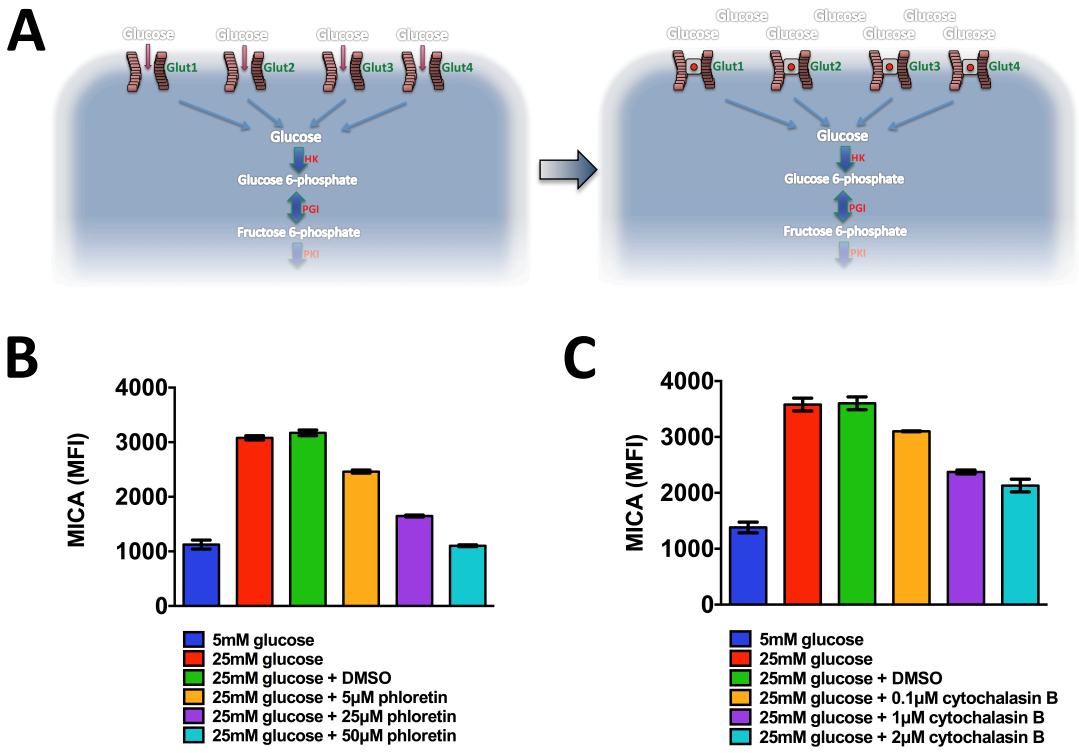

Figure 3.13: Inhibitors of glucose transport limit the effect of glucose on MICA expression. To address the question of whether glucose transport into the cell affects it's ability to alter NKG2D ligand expression, we treated HDK293T cells with two recognised inhibitors of GLUT transport, phloretin, and cytochalasin B. (A) Both phloretin and cytochalasin B act by physically blocking transport through GLUT1-4 facilitative transporters. Inhibition of MICA cell surface expression by phloretin (B) and cytochalasin B (C) occurs in a dose dependent manner. This suggests that glucose transport into the cell is important for MICA expression to occur.

HEK293T cells were cultured in low (5 mM) glucose for 48 hours. Fresh culture medium was then added to the cells containing high (25 mM) glucose, and increasing concentrations of the GLUT inhibitors, phloretin **(Fig 3.13b)** or cytochalasin B **(Fig 3.13c)**. Both compounds

limited the effect of glucose on MICA induction. This suggests that glucose transport into the cell is important for its effect on NKG2D ligand cell surface expression.

3.3.7.2 2-Deoxyglucose inhibits MICA expression

2-Deoxyglucose is a chemical analogue of glucose that lacks a hydroxyl group on the 2nd carbon of the hexose ring **(Fig 3.14)**. It has been used as a tool in metabolic research since the 1950s (228), and as such, it is well characterised. 2-Deoxyglucose does not act as a substrate for glycolysis. It is transported into the cell through glucose transporters and transformed to 2-deoxyglucose 6-phosphate through the action of hexokinase. 2-Deoxyglucose 6-phosphate inhibits hexokinase and phosphoglucose isomerase, but can influence gene transcription in situations where glucose 6-phosphate itself acts as a ligand in signal transmission, such as ChREBP binding (223).

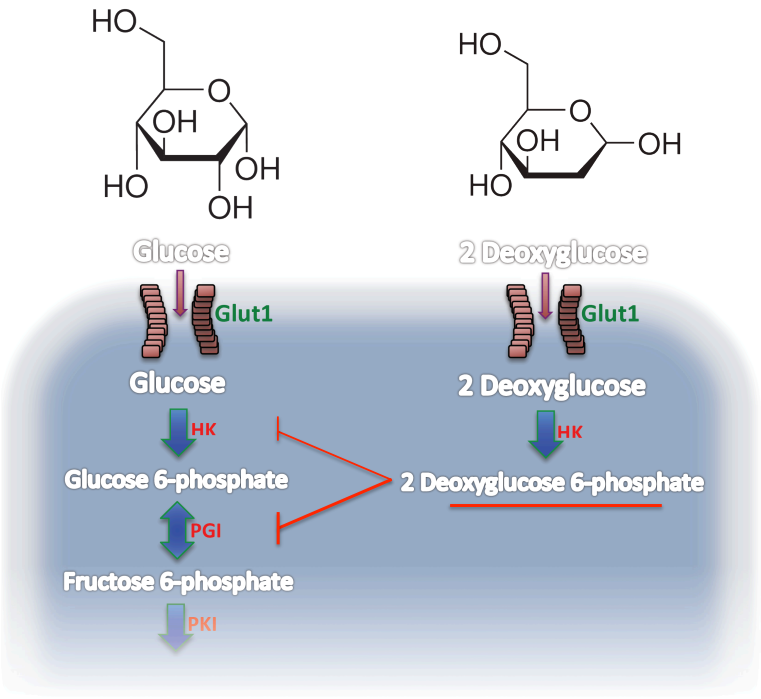

Figure 3.14: 2-deoxyglucose inhibits hexokinase and phosphoglucose isomerase. 2-deoxyglucose is structurally identical to glucose, apart from the substitution of a hydrogen for a hydroxyl group on the 2nd carbon. It is a well characterised compound. 2-deoxyglucose is efficiently transported into the cell through regular facilitative GLUT glucose transporters. Inside the cell it is readily phosphorylated by hexokinase to form 2-deoxyglucose 6-phosphate. In the process, hexokinase and phosphoglucose isomerase are inhibited. While 2-deoxyglucose 6-phosphate is not a substrate for further glycolysis, it has been shown to activate some glucose 6-phosphate driven transcriptional events, such as some ChREBP targets.

We first wanted to see whether 2-deoxyglucose 6-phosphate could re-capitulate the effect of glucose on NKG2D ligand expression, which would suggest a ChREBP-like activity. However, when HEK293T cells were cultured with equi-molar amounts of glucose or 2-deoxyglucose, cell surface MICA expression was only induced by glucose, and not by 2-deoxyglucose **(Fig 3.15a)**. Next, we wanted to see if the documented ability of 2-deoxyglucose to inhibit hexokinase and phosphoglucose isomerase, essential enzymes of early glycolytic metabolism, would limit the ability of glucose to induce cell surface MICA expression. As expected, the addition of 2-deoxyglucose to HEK293T cells cultured in 25 mM glucose limited the up-regulation of MICA **(Fig 3.15b)**. This suggests that for glucose to drive expression of NKG2D ligands, it must be metabolised beyond glucose 6-phosphate.

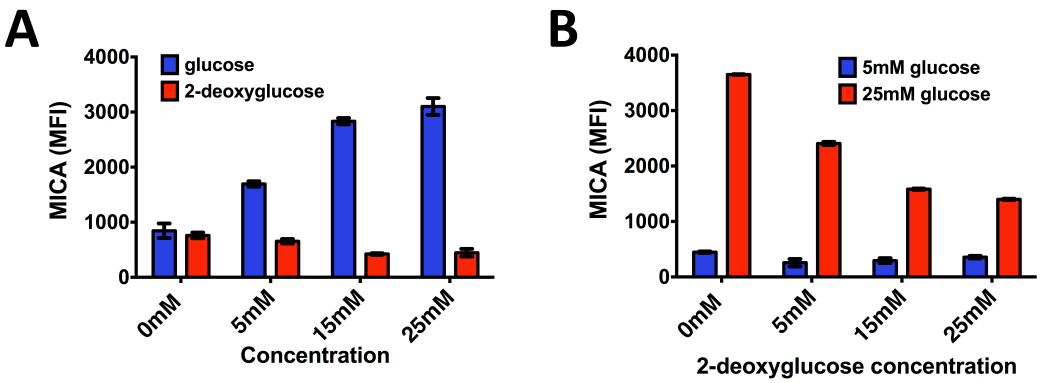

Figure 3.15: 2-Deoxyglucose does not induce MICA expression, and inhibits the effect of glucose. (A) Despite not acting as a glycolytic substrate, 2-deoxyglucose 6-phosphate can activate some glucose 6-phosphate signalling mechanisms, such as ChREBP. HEK293T cells were cultured in 5 mM glucose for 48 hours, and then changed to a gradient of either glucose, or 2-deoxyglucose for 48 hours. Only glucose induced MICA expression, suggesting that 2-deoxyglucose is incapable of controlling NKG2D ligand expression. (B) 2-Deoxyglucose is also known to inhibit the metabolism of glucose, acting on hexokinase and phosphoglucose isomerase. HEK293T cells were cultured in 5mM or 25mM glucose for 48 hours, before the medium was refreshed, and 2-deoxyglucose was added at the indicated concentrations. 2-deoxyglucose inhibited the effect of glucose on MICA expression, suggesting that glucose needs to be metabolised beyond this point in glycolysis to effect changes in cell surface MICA.

3.3.7.3 Mannose equally induces MICA expression

Looking at the need for glucose uptake and metabolism from an alternative perspective, we identified two alternative six carbon sugars capable of acting as glycolytic substrates, namely mannose and fructose **(Fig 3.16)**.

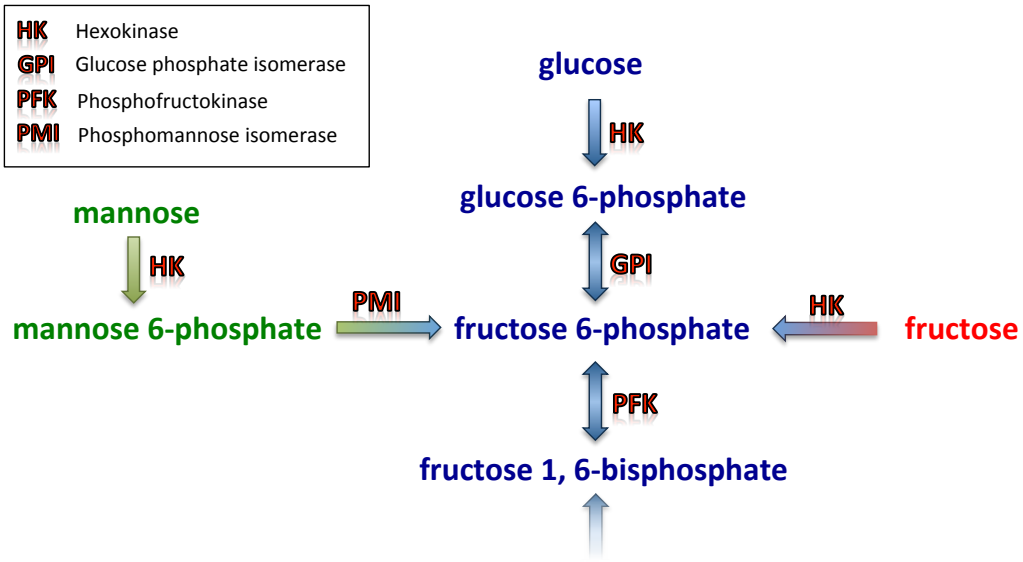

Figure 3.16: Glycolytic metabolism of six carbon sugars. Glucose is the principal substrate for glycolysis in mammals. However, mannose can be converted directly to fructose 6-phosphate, through the actions of hexokinase, and phosphomannose isomerase. Similarly, fructose can be converted directly to fructose 6-phosphate through the action of hexokinase. The phosphorylation of fructose by hexokinase occurs at a slower rate than the phosphorylation of glucose.

When we cultured HEK293T cells in media containing increasing concentrations of glucose, mannose or fructose, we found that glucose and mannose induced MICA expression equally well, while fructose had a limited and fixed impact on MICA expression **(Fig 3.17)**. The effect of mannose on inducing MICA cell surface expression suggested that a glycolytic substrate was needed to drive MICA expression, as opposed to glucose per sé.

3.3.7.4 Fructose induces MICA expression, following transient Glut5 transfection

All mammalian cells have glucose transporters. While studies show that separate mannose transporters exist, there is strong evidence that standard glucose transporters possess mannose-carrying capacity (229, 230, 231). In contrast, the transport of fructose into cells is more restricted. GLUT5 (*SLC2A5*) is a dedicated fructose transporter with limited tissue distribution (232). While other less well characterised transporters in the *SLC2A* gene family may have fructose transporting capacity (231), the details of this are less certain. We hypothesised that increasing fructose transport capacity would restore the ability of this glycolytic

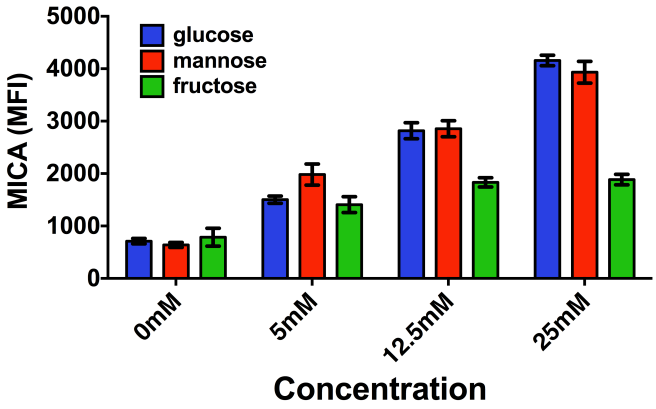

Figure 3.17: Mannose induces MICA expression independently of glucose. HEK293T cells were cultured in 5 mM glucose for 48 hours, before being transferred to DMEM containing the stated concentrations of glucose, mannose or fructose. Mannose induced MICA expression as efficiently as glucose, while fructose has a limited, fixed effect

substrate to induce NKG2D ligand expression **(Fig 3.18)**.

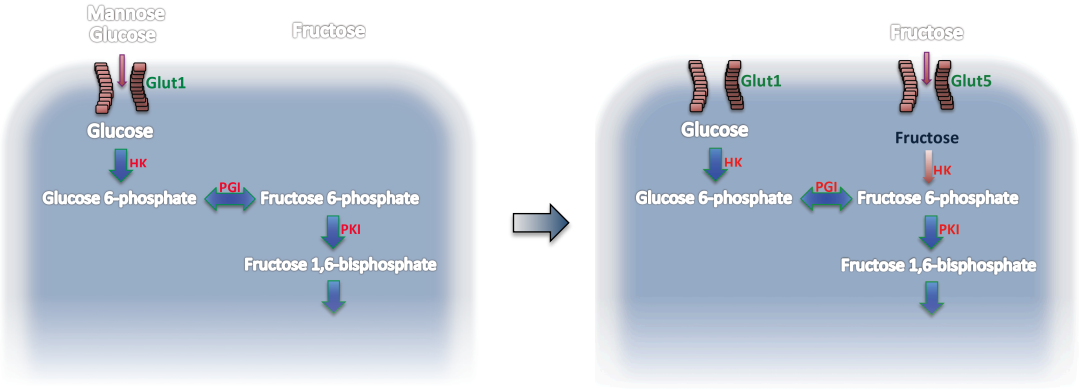

Figure 3.18: Enhancing fructose transport capacity through GLUT5 transfection. The cellular distribution of the fructose transporter, GLUT5, is limited. We hypothesised that cells lacking GLUT5 (*left*) would have limited fructose transport capacity, but that transfection of cells with a plasmid encoding the GLUT5 transporter (*right*) would restore the cellular capacity to take up fructose from the environment, and metabolise fructose through glycolysis.

To test this, we cloned full length *GLUT5* from MCF7 cell cDNA into a pCDNA3.1 expression vector. HEK293T cells were transiently transfected with *GLUT5* cDNA. The expression of *GLUT5* mRNA in these cells was confirmed by extracting total cellular RNA, digesting DNA (including plasmid DNA) with DNaseI, and carrying out RTPCR on the resulting cDNA. This showed that untransfected HEK293T cDNA did not contain GLUT5 mRNA, while cDNA from transfected HEK293T cells contained GLUT5 mRNA **(Fig 3.19a)**.

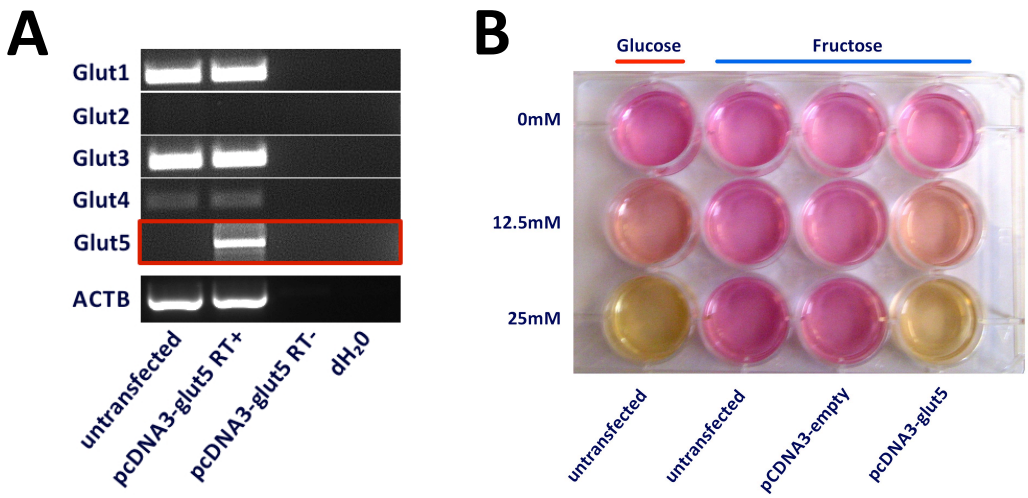

Figure 3.19: GLUT5 transfection of HEK293T cells enhances fructose transport capacity. (A) HEK293T cells were transiently transfected with an expression vector encoding the GLUT5 fructose transporter. GLUT5 expression was confirmed in transfected cells by RTPCR using DNase digested mRNA for cDNA synthesis. (B) GLUT5 transfected cells cultured in the presence of fructose acidified the culture medium (through lactate production) as efficiently as cells cultured with the same concentrations of glucose. Untransfected cells were not as metabolically active.

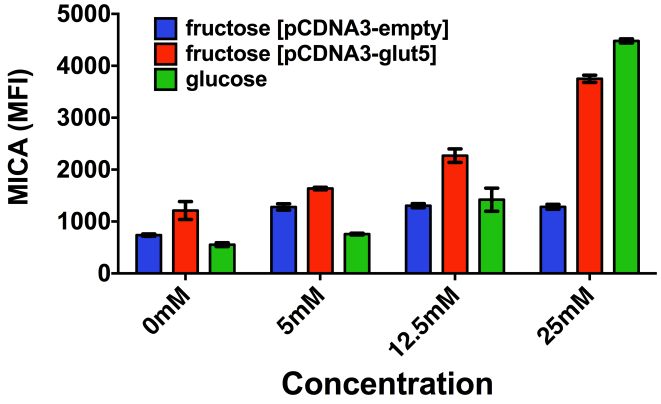

Figure 3.20: Fructose drives MICA expression in cells with fructose transport capacity. In HEK293T cells transfected with an empty pcDNA3.1 vector, fructose had limited impact on MICA expression, as previously observed **(Fig 3.18)**. In cells transfected with GLUT5 (confirmed by RT-PCR; **(Fig 3.19a)**, fructose induced MICA expression as strongly as that seen with glucose.

To see if conferring HEK293T cells with enhanced fructose transport capacity altered cellular ability to use fructose as a glycolytic substrate, we cultured untransfected, empty vector-transfected or *GLUT5*-transfected HEK293T cells in increasing concentrations of fructose. Cells cultured with high fructose concentrations acidified the culture medium at 48 hours only if they were transfected with *GLUT5*. This suggests that *GLUT5*-transfected HEK293T cells were able to utilise fructose as a glycolytic substrate, producing lactate, like cells cultured in high glucose concentrations.

Finally, we wanted to see if these cells, with enhanced capacity to utilise fructose as a glycolytic substrate, would display fructose-dependent NKG2D ligand cell surface expression. *GLUT5*-transfected or empty transfected cells were cultured in a fructose gradient. Cells transfected with an empty vector displayed a limited increase in MICA expression at the cell surface, while cells transfected with *GLUT5* displayed a 'glucose-like' MICA cell surface induction **(Fig 3.20)** .

3.4 Discussion

Our central hypothesis, that glucose metabolism is a key molecular determinant of cell surface NKG2D ligand expression, was formulated on the basis of the observation that NKG2D ligand expression is not known to occur in the absence of changes in glucose metabolism, across a diverse range of physiological settings (for references, see **Table 1.3**). In this chapter we have presented experimental data that directly supports this hypothesis, and begins to describe how this regulatory role of glucose is mediated by intracellular metabolites of glucose.

3.4.1 Glucose plays a key role in determining NKG2D ligand expression

A key characteristic of the 'Warburg effect' is the increased uptake and metabolism of glucose. Evidence for this comes from decades of *in vitro* experiments (reviewed by Cairns *et al* (160)) and is practically attested to by the success of ^{18}FDG-PET in identifying cancer cells *in vivo*

(183). Here we have shown that the amount of glucose metabolised by three diverse cell lines, HEK293T cells, MCF7 cells and HeLa cells influences the ability of cells to express the NKG2D ligand, MICA at the cell surface **(Fig 3.2)**. Furthermore, the amount of glucose metabolised in these cell lines is directly proportional to the immunogenicity of these cells *in vitro*, and leads to a significant change in NKG2D-dependent cytotoxicity **(Fig 3.3)**. The fact that this occurs across three genetically and histologically heterogenous cell lines provides compelling evidence as to the importance of glucose metabolism to the NKG2D ligand response on a wider scale. On the basis of the data presented, it remains a possibility that immunogenic factors other than NKG2D ligand expression are also controlled by glucose, such as other NK cell activating molecules, and this may merit separate investigation in its own right. Nevertheless, taken together, the results of these experiments are consistent with our hypothesis, that glucose or its metabolism has a causal role in NKG2D ligand expression, and demonstrate why this phenomenon warrants more finely detailed exploration.

3.4.2 The HEK293T model of Warburg metabolism

In order to understand the role of glucose metabolism in NKG2D ligand regulation in more detail, we developed a HEK293T cell model of the process of Warburg metabolism. This model recapitulated key features of Warburg metabolism, and allowed us to gather further evidence in support of our glucose-NKG2D hypothesis.

Glucose concentrations. In considering potential models of Warburg metabolism, a key requirement was the ability to alter the amount of glucose available to cells for metabolism. In physiological settings, this occurs principally through changes in glucose transport capacity (reviewed by Szablewski (233)). Because replicating this *in vitro*, for example with inducible glucose transporter expression constructs, may itself influence NKG2D ligand expression, we achieved variable glucose delivery by adjusting culture medium glucose concentration. Physiological plasma glucose concentrations in healthy human subjects range between 4 mM and 10 mM, when monitored continuously over 24 hours (234). Modern *in vitro* cell culture me-

dia were optimised through the 1950s and 1960s on the basis of experiments such as those carried out by Munyon *et al* (154). These experiments identified the optimal composition of culture media for mammalian cells grown in the *in vitro* environment, including the optimal concentrations of glucose. As a result, common culture media in use today, such as RPMI1640 and DMEM contain 11.1 mM and 25 mM glucose respectively. For this reason, for central experiments, we chose a 'low' glucose concentration of 5 mM, and a 'high' glucose concentration of 25 mM.

Biological and metabolic characterisation. This method of controlling HEK293T cell access to glucose served well in reproducing the central biological and metabolic characteristics of Warburg metabolism in our cells. Critically, we verified that controlling cell access to glucose in this way reliably changed glucose consumption and lactate production **(Fig 3.11)**. At the key glucose concentrations of 5 mM and 25 mM, despite these large and significant differences in metabolic output, the HEK293T cells remained functionally intact in several core biological categories. The ability to proliferate is retained at both glucose concentrations (with relatively small differences in rate; **Fig 3.4**), the cell cycle distribution was unchanged **(Fig 3.6)** and cells retain the ability to express genes, even in culture medium with 5 mM glucose **(Fig 3.9)**. Collectively, these observations support the assertion that the control of glucose availability to HEK293T cells represents a strong model of Warburg metabolism, and that 5 mM and 25 mM concentrations of glucose allow a clear metabolic distinction to be made, without abolishing core cellular functions.

The HEK293T - NKG2D ligand response. Because of the large and significant change in NKG2D-dependent cellular immunogenicity across a range of glucose concentrations, having established a robust model of Warburg metabolism, we next examined the breadth of the impact that glucose supply has on NKG2D ligand expression in HEK293T cells. No suitable monoclonal antibodies are available as yet for ULBP6. Examination of the impact of glucose on expression of other NKG2D ligands revealed significant changes in surface expression of MICA, ULBP1, ULBP2, and ULBP3 **(Fig 3.7)**. Glucose had the

reproducible and significant effect of decreasing cell surface ULBP5 expression **(Fig 3.7 & 3.8)**. This is an important observation, as it excludes the possibility of generic influences on cell surface protein expression by glucose, such as changes in cell surface area, or protein turnover. Interestingly, the *ULBP5* and *ULBP2* gene promoters are located side-by-side on chromosome 6, with their gene bodies encoded in opposite orientations, raising the possibility of a transcription-level reciprocal regulation, which would be of interest in its own right as a regulatory mechanism. The fact we find an increase in total cellular MICA **(Fig 3.10)**, along with the observed increase in cell surface MICA, adds further weight to the argument that this cellular model demonstrates glucose-mediated absolute increases in NKG2D ligand protein expression, as opposed to distributional, transport or relative changes in NKG2D ligand levels.

3.4.3 Intracellular metabolites of glucose control NKG2D ligand expression

Having established the HEK293T cell model of Warburg metabolism, and characterised the glucose-mediated changes in NKG2D ligand expression in this model, we next turned to the question of mechanism, and considered four broad and inclusive possibilities for the mode of action of glucose in this model: acting from outside the cell through effects on osmolality, or cell surface receptor binding, or from within the cell, either as glucose itself, or as a metabolite of glucose. Collectively, our results demonstrate that an intracellular metabolite of glucose is responsible for these effects. This is a key finding, as it significantly reduces the number of possible mechanistic explanations for the action of glucose in this capacity.

The conclusion that an intracellular metabolite of glucose is responsible for NKG2D ligand regulation rests on two main observations. First, that inhibition of glucose transport into cells using chemical inhibitors lessens the effect of glucose on increasing NKG2D ligand expression **(Fig 3.13)**. If glucose were acting from outside the cell, we would expect an equal effect on NKG2D ligand expression, regardless of glucose transport capacity. Second, the ob-

servation that an alternative substrate for glycolysis, fructose, can control MICA expression as effectively as glucose, and this effect is dependent on the expression of the cell surface fructose transporter, GLUT5 **(Fig 3.20)**.

Th idea that glucose can alter the expression of genes as a result of its metabolism is not without precedent. The regulation of expression of the insulin gene by glucose is thought to depend on signalling transmitted as a result of changes in either NAD/NADH oxidation status, or possibly through changes in the phosphorylation levels of intracellular adenosine nucleosides (reviewed by MacDonald (218) and Melloul *et al* (235)). The data presented here suggests that that NKG2D ligand expression is regulated in a similar manner.

3.4.4 Summary

In this chapter we demonstrated that glucose has a functionally significant effect on NKG2D ligand expression. We developed and characterised a HEK293T cell model of Warburg metabolism, and used this to show the significant impact of glucose on expression of a range of NKG2D ligands. We demonstrated that glucose is transported into the cell and metabolised before mediating these effects on NKG2D ligand expression. Together these findings are consistent with our hypothesis that the metabolism of glucose is deterministically important for functional cell surface NKG2D ligand expression.

4

Nucleotide metabolism and NKG2D ligand expression

Contents

4.1 Abstract

Warburg metabolism comprises a state of cellular metabolic activation seen in malignancy, cellular proliferation and infection, with increased glucose uptake, aerobic glycolysis, and lactate production. Warburg metabolism commonly occurs in the setting of NKG2D ligand expression. To explore the relationship between cellular glucose metabolism and cell surface

NKG2D ligand expression, we developed a HEK293T cell model of Warburg metabolism, and demonstrated that glucose metabolism is important for NKG2D ligand expression. Here, we show that purine nucleotide synthesis, dependent on the glucose metabolites, ribose 5-phosphate and PRPP, is central to the up-regulation of NKG2D ligands by glucose. Furthermore, we demonstrate that purine nucleosides are themselves sufficient to induce strong NKG2D ligand expression, independently of glucose. We define a set of purine molecules with the capacity to drive NKG2D ligand expression, and discuss some potential mechanisms through which this might occur.

4.2 Introduction

4.2.1 Summary of previous work

There is increasing evidence to suggest that the cellular contexts in which cell surface NKG2D ligand expression is observed, are also settings in which the Warburg metabolic phenotype is established **(Table 1.3)**. We have hypothesised that the process of changing from quiescent metabolism to active Warburg metabolism plays an important role in the molecular control of cell surface NKG2D ligand expression. To explore this hypothesis, we developed a simple cellular model of Warburg metabolism, by controlling glucose availability to HEK293T cells cultured *in vitro*. This model recapitulates the metabolic features seen in Warburg metabolism, and demonstrates glucose-driven cell surface NKG2D ligand expression that alters cellular NKG2D-dependent susceptibility to NK cell cytotoxicity. Further, we have presented evidence to suggest that for glucose to mediate cell surface NKG2D ligand expression in this setting, glucose must be transported into the cell, and metabolised further than glucose 6-phosphate.

4.2.2 Metabolites of glucose support nucleotide synthesis

Purine nucleotides are synthesised (both *de novo* and salvage pathways) from the glucose metabolite, phosphoribosyl pyrophosphate (PRPP; nucleobase nomenclature is described in **Fig 4.1**; see **Fig 4.2** for the basic organisation of the purine biosynthetic pathway; a more detailed view of purine synthesis and other metabolic pathways is available from the KEGG database at www.genome.jp/kegg/kegg2.html (236)). PRPP is produced from glucose via the pentose phosphate pathway (PPP), a group of interwoven metabolic pathways that branch from several points in the upper portion of the glycolytic pathway, eventually generating ribose 5-phosphate. Ribose 5-phosphate is then metabolised to PRPP by the enzyme PRPP synthetase. PRPP forms the pentose sugar component of the nucleotide.

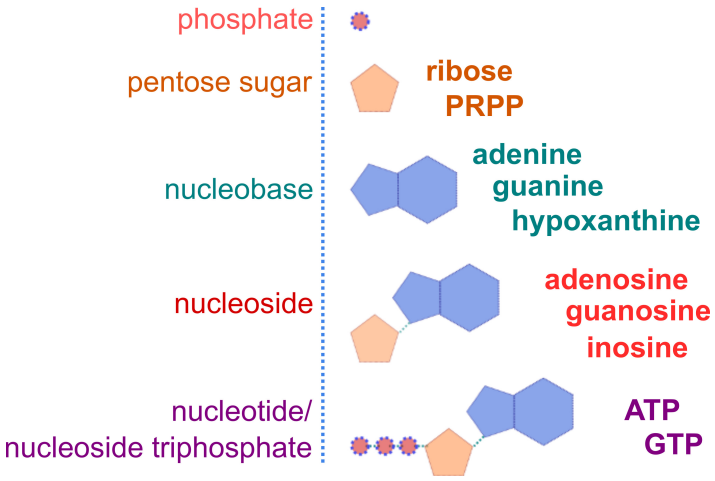

Figure 4.1: Nucleobase nomenclature. The bases that make up DNA and RNA are comprised of three components: a nucleobase, a pentose sugar, and a triphosphate group. The term 'nucleobase' refers to the base without a pentose sugar. The key purine nucleases are adenine, guanine and hypoxanthine. While pyrimidine bases can be synthesised *de novo* without the pentose sugar, PRPP, in humans, purine nucleobases can only be built upon the sugar backbone, PRPP. The term 'nucleoside' refers to a nucleobase attached to a pentose sugar. The central purine nucleases are adenosine, guanosine and inosine. The terms nucleotide and nucleoside triphosphate are interchangeable, and refer to a phosphorylated nucleoside, for example ATP and GTP.

For salvage purine nucleotide synthesis **(Fig 4.2; *green text*)**, a purine base is combined with PRPP to generate the corresponding nucleoside monophosphate. This is a more energy efficient process than *de novo* synthesis. *De novo* purine synthesis involves the step-wise construction of the purine base on top of PRPP. The enzyme amidophosphoribosyl transferase

(APRT) commits PRPP to the *de novo* purine synthesis pathway **(Fig 4.2; *red text*)**. Unlike purine bases, which cannot be synthesised *de novo* without the PRPP founder molecule, the *de novo* synthesis of the pyrimidine base proceeds independently of PRPP, which is added to the base at a later point. Following many decades of research into purine metabolism, the regulation of these pathways in humans remains unclear.

Despite this lack of clarity, some studies have tried to describe the relationship between glucose concentration, PRPP synthesis, and purine synthesis. Recent metabolomic mass spectrometer-based studies have estimated that in HEK293 cells, ∼15% of consumed glucose is channeled through the pentose phosphate pathway (237). It is not yet clear from metabolomic studies how this flux changes with increased glucose uptake, or with quiescent-Warburg transitions. Instead, our best understanding of the relationship between glucose, ribose 5-phosphate, and purine nucleotide synthesis comes from the biochemical and chromatographical methods of ∼20 years ago. Boer *et al* addressed this question using chromatographical methods to measure metabolic intermediates in the HepG2 hepatocellular carcinoma cell line. One of his key findings was that increasing the concentration of glucose in cell culture medium lead to a proportional increases in both ribose 5-phosphate and PRPP (238). Further, the rate of *de novo* purine synthesis was proportional to the level of available PRPP. This relationship between PRPP availability and *de novo* purine synthesis has been confirmed by several authors (239, 240).

Taken together, this evidence suggests, at least in the context of cell lines, increasing glucose availability leads to increased PRPP, and ultimately increased purine nucleotide synthesis.

4.2.3 Aims and objectives

In this chapter, we use our HEK293T cell model of glucose-induced NKG2D ligand expression to examine our hypothesis that ribose 5-phosphate, in its capacity to drive purine nucleotide

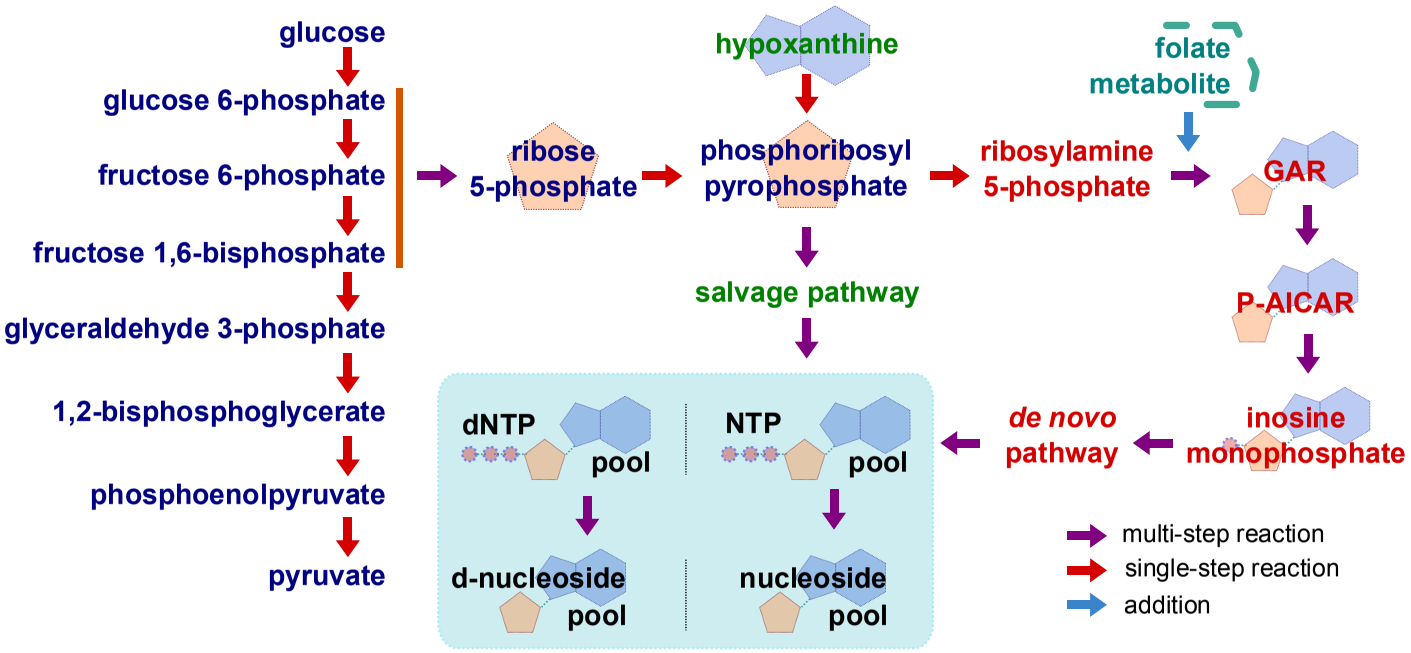

Figure 4.2: *De novo* and salvage purine synthesis pathways. Glycolysis refers to the metabolism of glucose to pyruvate (*left hand side*). Purine biosynthesis is dependent on the pentose phosphate pathway (PPP), which consists of a collection of inter-meshed metabolic reactions that can arise from the upper glycolytic intermediates, indicated by the orange vertical bar. For purine biosynthesis, a key product of the PPP, and therefore of glucose, is the five carbon sugar, ribose 5-phosphate. The enzyme phosphoribosyl pyrophosphate (PRPP) synthetase converts ribose 5-phosphate to PRPP, in what is considered the first committed step of purine biosynthesis. PRPP is a central molecule, essential for both salvage and *de novo* purine synthesis pathways, forming the sugar component of nucleosides. Salvage purine synthesis (*green text*) refers to the direct combination of a PRPP molecule with a recycled nucleobase, such as hypoxanthine, adenine, or guanine, to form the corresponding nucleoside monophosphate. The salvage pathway is energetically favourable, and cells use the salvage pathway in preference to *de novo* synthesis when recycled nucleobases are available (241, 242). The enzyme amidophosphoribosyl transferase (APRT) catalyses the first committed step of the *de novo* pathway (*red text*). *De novo* biosynthesis of purines simply refers to the step wise construction of a new nucleobase onto the PRPP foundation. While synthesis of the pyrimidine ring can occur without PRPP, this is not the case for purines. Building of the purine ring onto PRPP, which also requires folate intermediates, ultimately leads to the synthesis of inosine monophosphate (IMP). Here, purine synthesis branches, and IMP can form either adenosine nucleotides or guanosine nucleotides. This is shown in greater detail in **Fig 4.9**.

synthesis, is one of the key metabolites of glucose 6-phosphate causing NKG2D ligand cell surface expression. We probe purine biosynthetic pathways using chemical inhibitors of central purine pathway enzymes to describe the relationship between purine nucleotides and NKG2D ligand expression. We will explore some prospective signalling mechanisms that may explain the impact of purine nucleotides on NKG2D ligand expression. Finally we discuss potential future approaches to advancing this line of inquiry.

4.3 Results

4.3.1 Purine synthesis is necessary for glucose-driven MICA expression

4.3.1.1 The role of PRPP in *de novo* purine synthesis is essential for the induction of MICA by glucose

Because ribose 5-phosphate and phosphoribosyl pyrophosphate (PRPP) are metabolites of glucose, and are essential for purine nucleotide synthesis, we first wanted to see if the purine synthesis pathway was necessary for the effect of glucose on MICA expression in our HEK293T cell model. To test this, we used two chemical inhibitors of the enzyme amidophosphoribosyl transferase (APRT), which is the first enzyme of the *de novo* purine synthesis pathway **(Fig 4.3c)**. These compounds, azaserine and 6-diazo oxonorleucine (DON), through inhibition of APRT, prevent *de novo* purine synthesis (243, 244). Because there are no salvage pathway substrates in culture medium (apart from a biologically insignificant amount of hypoxanthine, estimated to have a concentration of 10-15 µM in fetal calf serum (241)) both azaserine and DON effectively inhibit purine synthesis.

HEK293T cells were cultured in DMEM for 48 hours. The culture medium was discarded, and fresh culture medium containing either 5 mM or 25 mM glucose was added to the cells, with or without azaserine **(Fig 4.3a)** or DON **(Fig 4.3b)**. Following 48 hours of cell culture in these conditions, cell surface MICA expression was measured by flow cytometry. Both of these chemical inhibitors prevented glucose from inducing MICA cell surface

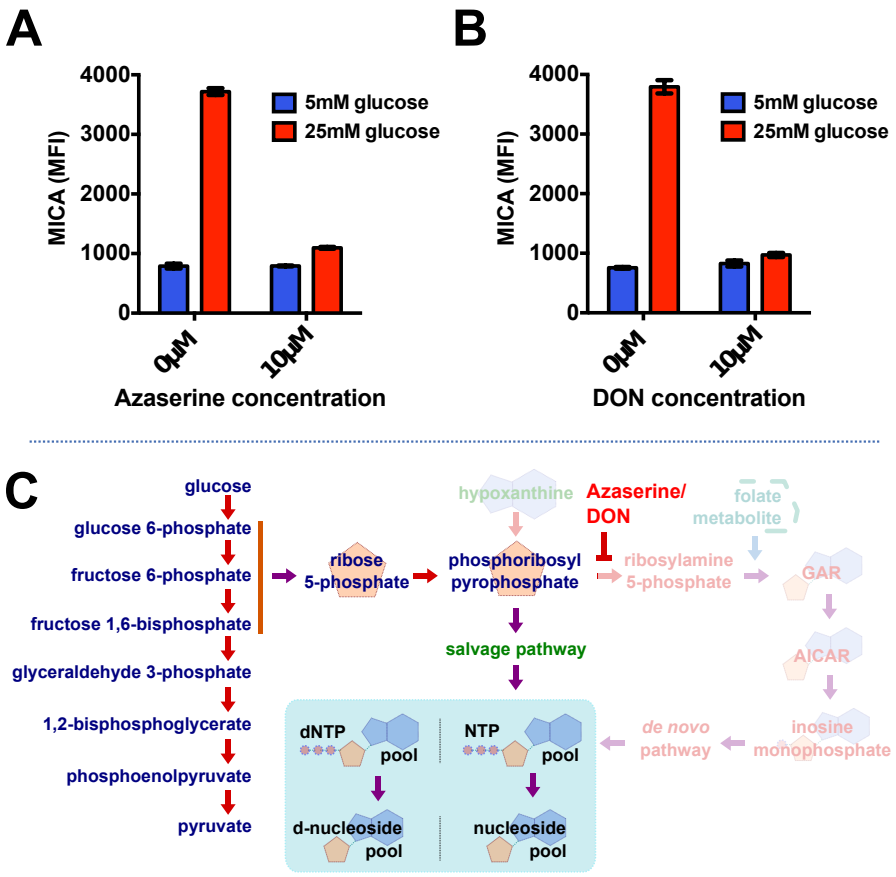

Figure 4.3: Inhibition of *de novo* purine synthesis blocks glucose induced MICA expression. Ribose 5-phosphate and PRPP are metabolites of glucose. Evidence suggests that the cellular level of ribose 5-phosphate and PRPP is proportional to the amount of available glucose, and this influences the rate of purine nucleotide synthesis (238). We blocked the contribution of PRPP to *de novo* purine synthesis using APRT inhibitors, azaserine and DON (C). HEK293T cells were treated in 5 mM glucose for 48 hours, and then transferred to 5 mM or 25 mM glucose with or without azaserine (A) or DON (B) for a further 48 hours. Cell surface MICA expression was measured by flow cytometry. Both compounds prevented glucose-induced MICA expression, suggesting that *de novo* purine synthesis is necessary for glucose-induced MICA expression. The histograms represent the mean +/– 95% confidence interval of mean fluorescence measurements from three biological replicate samples.

expression, suggesting that the purine synthesis pathway is necessary for glucose-mediated MICA expression.

4.3.1.2 Azaserine specifically inhibits amidophosphoribosyl transferase to block MICA expression

One difficulty of working with chemical inhibitors is the occurrence of off-target effects. Because this was a key observation, we wanted to see if both APRT inhibitors, azaserine and DON, were working on-target (against APRT) to bring about the effect of blocking MICA's induction by glucose. To do this, we used two populations of HEK293T cells: a regular population, and a hypoxanthine, aminopterin and thymidine (HAT)-selected population. The metabolic process of HAT selection is shown in **Fig 4.4**.

HAT selection was most famously employed by Köhler and Milstein in 1975 to generate antibody hybridomas (245). Köhler and Milstein used $HGPRT^{-/-}$ immortal myeloma cell lines, and fused these cells with primary $HGPRT^{+/+}$ antibody-secreting plasma cells. HGPRT is the enzyme that combines hypoxanthine with PPRP to generate IMP, which in turn can generate adenosine nucleotides or guanosine nucleotides. This process is also known as the salvage pathway of purine synthesis. Cells without HGPRT are fully dependent on the *de novo* pathway of purine synthesis to survive. The aminopterin in HAT selection medium **(Fig 4.4b)** inhibits dihydrofolate reductase. *De novo* purine synthesis cannot continue without the folate metabolites derived from dihydrofolate reductase activity. The hypoxanthine provides a substrate for the salvage pathway of purine synthesis. The net result of culture in HAT selection medium, is that cells use only salvage pathway purine synthesis. In the case of Köhler and Milstein, if the $HGPRT^{+/+}$ antibody secreting primary plasma cells successfully fused with the immortal $HGPRT^{-/-}$ myeloma cells, the resulting hybridoma would survive culture in HAT medium. Unfused, $HGPRT^{-/-}$ myeloma cells would die, through an inability to make purines.

HEK293T cells cultured in regular DMEM are dependent on *de novo* purine synthesis,

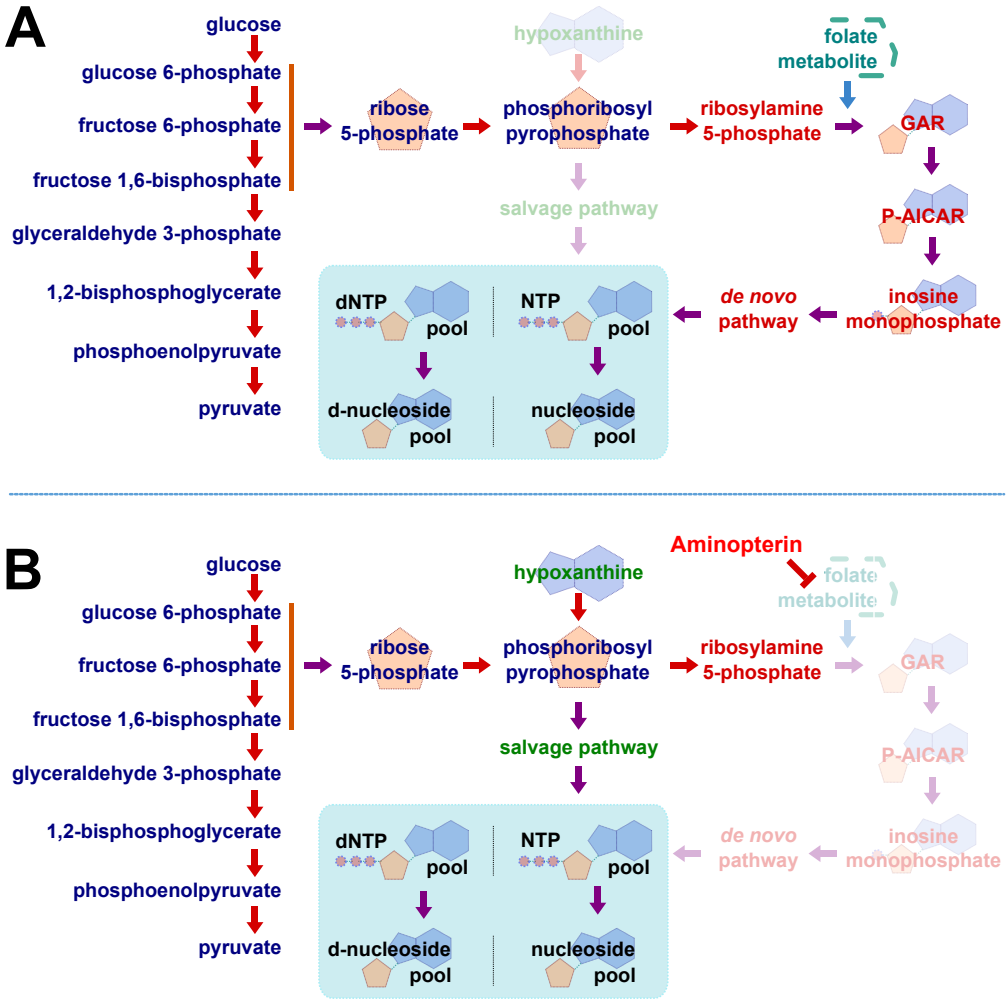

Figure 4.4: Metabolic principle of HAT culture medium selection. (A) Cells proliferating in regular culture medium are dependent on the *de novo* purine synthesis pathway (red text) for purine production. While the salvage pathway is active and energetically preferable to *de novo* synthesis, a sufficient supply of salvage pathway substrate (e.g. hypoxanthine) is not available from regular culture medium (or plasma *in vivo*). (B) Culturing cells in HAT selection medium forces cells to use the salvage pathway only for purine synthesis. The aminopterin present in HAT inhibits dihydrofolate reductase. This prevents the synthesis of the folate metabolites necessary for *de novo* purine synthesis, and therefore inhibits *de novo* synthesis. The hypoxanthine present in HAT provides a salvage pathway substrate to allow purines to be supplied solely from salvage procedures.

because they lack a substrate for salvage pathway synthesis **(Fig 4.4a)**. HEK293T cells cultured in HAT selection DMEM are completely dependent on salvage pathway purine synthesis **(Fig 4.4b)**. If the effect of azaserine and DON in preventing glucose-induced MICA expression was solely through inhibition of the first step of *de novo* purine synthesis, we hypothesised that azaserine and DON would not prevent glucose-mediated MICA expression in HAT-selected cells, as the *de novo* pathway is not active in these cells.

HEK293T cells or HAT-selected HEK293T cells (HEK293T^{HAT}) were cultured in DMEM or DMEM-HAT with 5 mM glucose for 48 hours, and then transferred to DMEM **(Fig 4.5a)** or DMEM-HAT **(Fig 4.5b)** with 5 mM or 25 mM glucose. Each group was treated with azaserine or DON to inhibit *de novo* purine synthesis, and cultured for a further 48 hours before measuring cell surface MICA expression by flow cytometry. Cells in each group were also treated with phloretin, a glucose transport inhibitor, as a 'negative control', to show how MICA expression is affected in HEK293T^{HAT} cells by an inhibitor that is known *not* to act on purine synthesis pathways. While all three compounds (azaserine, DON and phloretin) blocked the effect of glucose on MICA expression in HEK293T cells as expected **(Fig 4.5a)**, azaserine had no effect on glucose induced MICA expression in HEK293T^{HAT} cells **(Fig 4.5b)**, suggesting that azaserine is indeed acting on the *de novo* purine synthesis pathway. MICA expression on HEK293T^{HAT} cells cultured in high glucose was partly reduced by DON, suggesting that DON also acts through a separate unknown mechanism to limit MICA expression. HAT selection had no effect on the ability of phloretin to reduce MICA expression, as expected.

These experiments suggest that of the APRT inhibitors, azaserine is acting on-target, so this compound was used in later experiments.

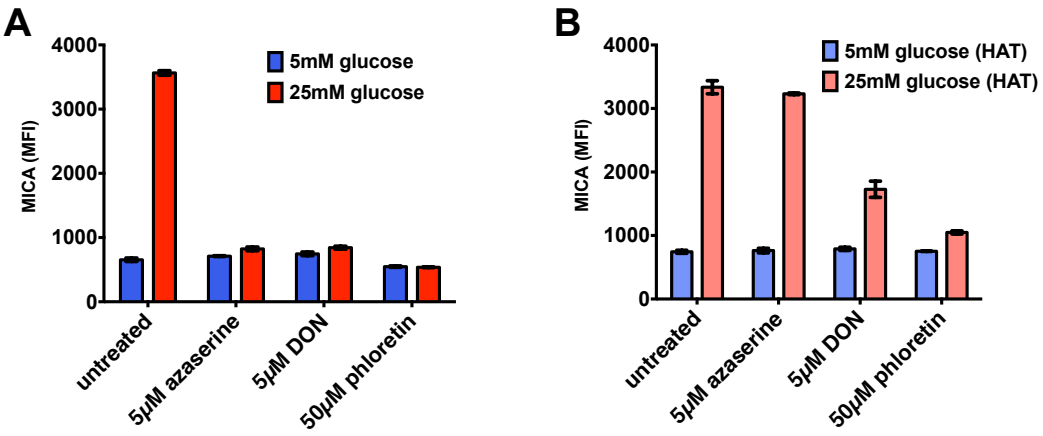

Figure 4.5: Azaserine inhibits MICA expression by inhibiting *de novo* purine synthesis. (A) HEK293T cells were cultured in 5 mM glucose for 48 hours, then transferred to 5 mM or 25 mM glucose, and treated with azaserine, DON or phloretin (cell surface glucose transporter inhibitor) for a further 48 hours. These cells are dependent on the *de novo* pathway for the synthesis of new purines **(Fig 4.4a)**. As expected, azaserine, DON and phloretin each prevented glucose-mediated NKG2D ligand expression. (B) HEK293T^{HAT} cells were prepared identically, with culture medium containing HAT. HEK293T^{HAT} cells use only the salvage pathway for purine synthesis. Azaserine had no effect on glucose-mediated MICA expression in these cells, suggesting that the action of azaserine to block glucose-mediated MICA expression is localised to the *de novo* purine synthesis pathway. DON still reduced MICA expression, suggesting that it also has MICA limiting activity in alternative pathways. Phloretin, by limiting glucose uptake into the cell, and therefore inhibiting both salvage and *de novo* purine synthesis, reduced MICA expression under all conditions. MICA cell surface expression was analysed by flow cytometry.

4.3.1.3 Activation of salvage pathway purine synthesis rescues MICA expression in the presence of azaserine

To confirm that the utilisation of PRPP (a metabolite of glucose) in purine nucleotide synthesis was necessary for the effect of glucose on induction of MICA expression, we cultured HEK293T cells in 5mM glucose for 48 hours, and then transferred the cells to DMEM containing 25 mM glucose and azaserine, to prevent the glucose being used for *de novo* purine synthesis **(Fig 4.6b)**. To allow the glucose-derived PRPP to be utilised in purine nucleotide synthesis, we added increasing doses of hypoxanthine, a salvage pathway substrate, that can be combined with PRPP by the enzyme HGPRT to make inosine monophosphate (IMP), restoring purine nucleotide pools through the salvage pathway (refer also to **Fig 4.9**). Cells were also stained with CFSE to track proliferation (decreasing CFSE fluorescence corresponding to increased rate of proliferation; see **Fig 3.4**), and cultured for 48 hours before cell surface MICA and CFSE fluorescence were measured by flow cytometry.

The left histogram **(i)** in **Fig 4.6a** shows the values for MICA expression and CFSE fluorescence in HEK293T cells cultured in DMEM with 5 mM or 25 mM glucose. The first data point on the right histogram **(ii)** shows that in cells cultured in high glucose, azaserine prevents MICA expression and relatively reduces the rate of proliferation (high CFSE fluorescence). However, as salvage pathway synthesis increases, with increasing doses of hypoxanthine (from left to right), MICA expression and proliferation are restored.

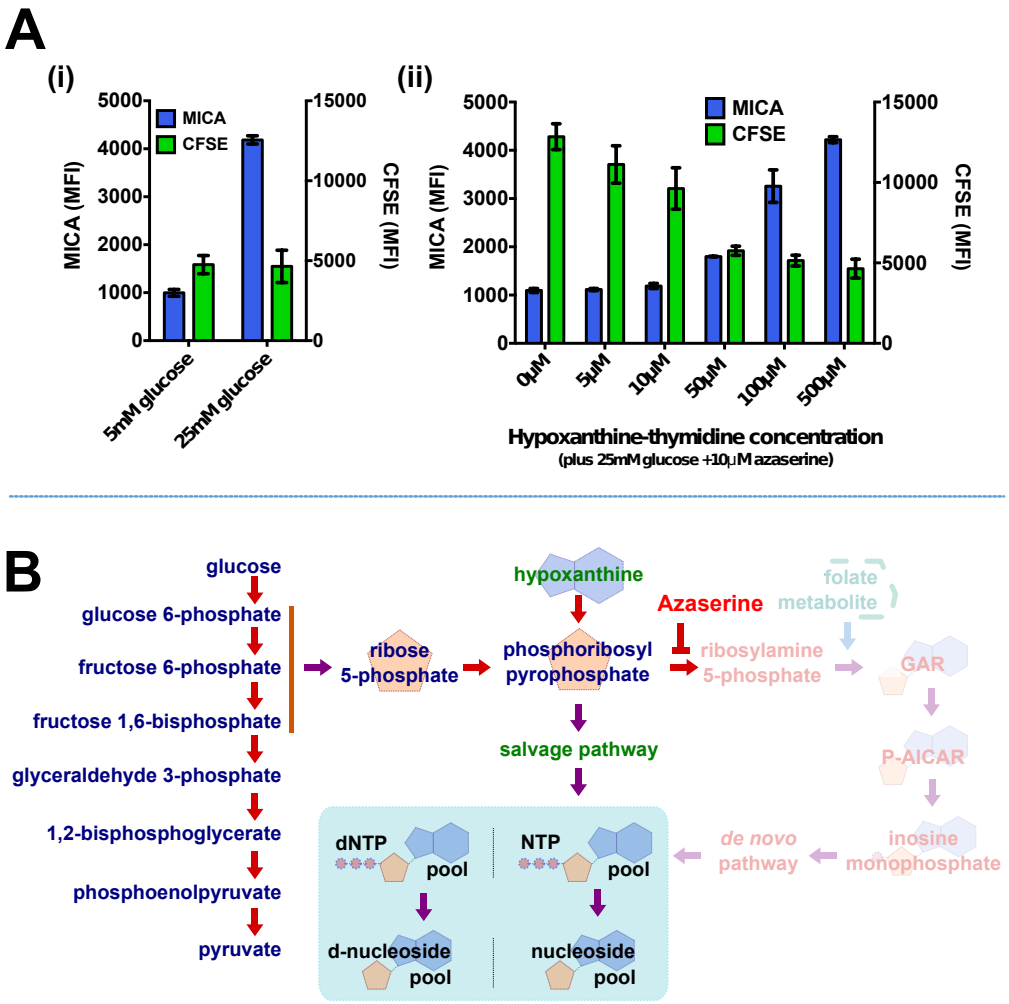

Figure 4.6: MICA expression in cells inhibited with azaserine is rescued by hypoxanthine-thymidine. (A) HEK293T cells were cultured in 5 mM glucose for 48 hours, before being stained with CFSE, and then cultured in the treatment conditions indicated for 48 hours. Cell surface MICA and CFSE was measured by flow cytometry at 48 hours. Control HEK293T cells (left histogram) were cultured in 5 mM or 25 mM glucose only, showing glucose-induced MICA expression in 25 mM glucose. In the right histogram, all cells were cultured in high glucose, with azaserine to inhibit the *de novo* purine synthesis pathway. The first group of cells show that MICA expression is reduced (by inhibition of *de novo* purine synthesis) and proliferation is also limited. The addition of a salvage pathway substrate (hypoxanthine) to these cells shows dose dependent recovery of MICA expression, and proliferation. This suggests that glucose-induced MICA expression is dependent on purine synthesis. (B) Schematic representation of the culture conditions for cells depicted in the right histogram in (A)

This shows that the utilisation of PRPP (and therefore glucose) for purine nucleotide synthesis is necessary for the impact of glucose on MICA expression.

The converse experiment, where the hypoxanthine concentration and azaserine concentration are constant and the glucose concentration changes progressively also supports this assertion **Fig 4.7**. Here, glucose is being predominantly utilised through the *de novo* pathway (*blue*), neither pathway (*red*; inhibited *de novo* synthesis, and no salvage pathway substrates), or the salvage pathway (*green*). Cell surface MICA expression is induced when glucose can be used in purine synthesis through either salvage or *de novo* pathways, but not when both of these pathways are blocked.

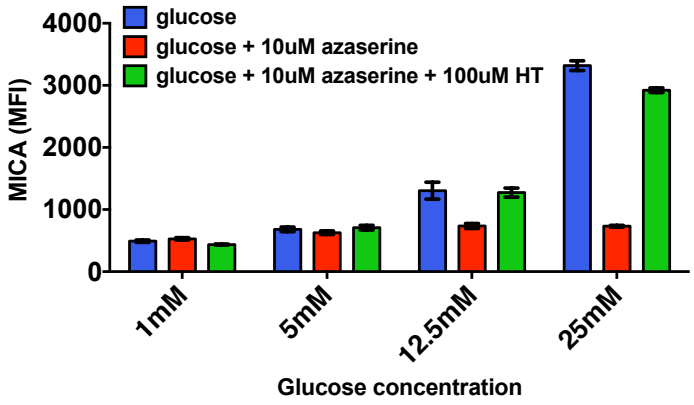

Figure 4.7: Glucose induces MICA expression through salvage or *de novo* purine synthesis pathways. To compare glucose-mediated cell surface MICA expression via salvage or *de novo* purine synthesis pathways, HEK293T cells were cultured in 5 mM glucose for 48 hours, and then transferred to DMEM in increasing glucose concentrations from 1 mM to 25 mM. Cells were also cultured in the presence of additional azaserine, or azaserine and hypoxanthine. Cell surface MICA expression was measured by flow cytometry. MICA expression was induced by glucose when either the salvage pathway, or *de novo* pathways were active, but MICA expression was not induced when neither pathway was active. HT - Hypoxanthine-thymidine.

4.3.1.4 Inhibition of *de novo* nucleotide synthesis does not inhibit global gene expression

We were concerned by the possibility that azaserine, in preventing purine nucleotide synthesis, would prevent the expression of all genes by depleting nucleotide pools for mRNA synthesis. To see if gene expression machinery remained intact in the presence of azaserine, we infected HEK293T cells with a lentivirus encoding doxycycline-inducible eGFP. The cells were selected

with puromycin for 10 days, and transferred to puromycin-free DMEM with 5 mM glucose for 48 hours. The cells were then transferred to the conditions shown, either without **(Fig 4.8a)** or with **(Fig 4.8b)** doxycycline. All cells were cultured in these conditions for 48 hours, when cell surface MICA expression and eGFP fluorescence was measured by flow cytometry. Azaserine prevented glucose-mediated MICA expression as previously demonstrated, but did not prevent doxycycline-mediated expression of eGFP **(Fig 4.8b)**, illustrating that over the 48 hour time period, azaserine does not preclude gene expression.

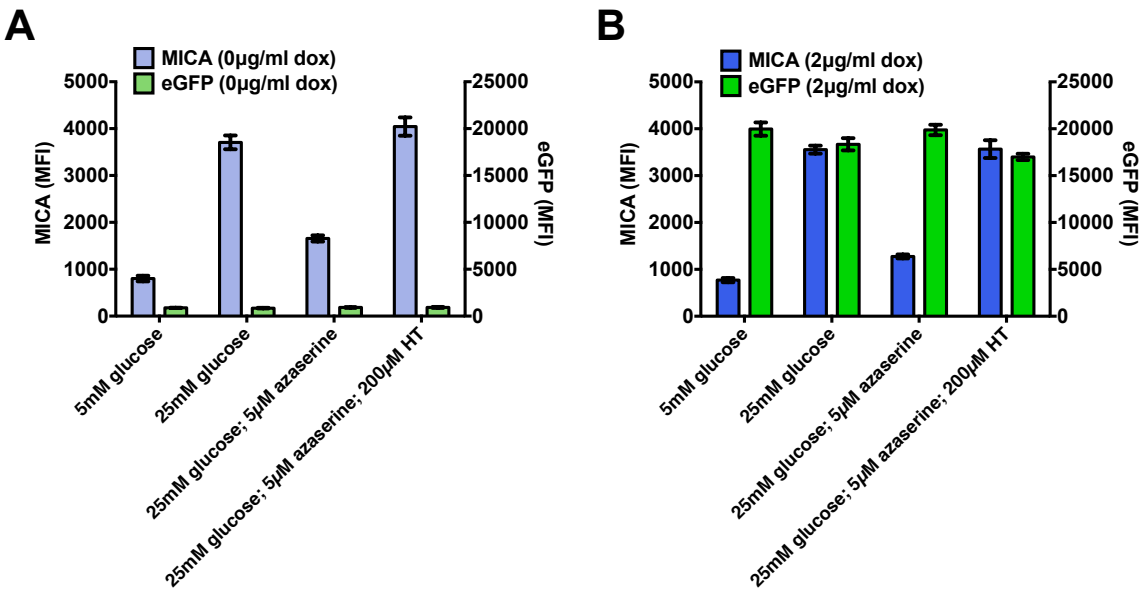

Figure 4.8: Azaserine does not prevent gene expression. HEK293T cells were infected with a doxycycline-inducible eGFP-encoding lentivirus. We compared glucose-induced MICA expression with doxycycline-induced eGFP expression, and the effect of azaserine on both. Cells were not treated (A) or treated (B) with $2\,\mu g\,ml^{-1}$ of doxycycline. Azaserine prevented glucose-induced MICA expression, but not doxycycline induced eGFP expression (B). Similarly, the expression of eGFP was not affected by glucose, demonstrating that in both low glucose and high glucose with azaserine, the gene expression machinery is intact.

4.3.1.5 Nucleobases and nucleosides can equally salvage MICA expression

Having established that glucose metabolite PRPP was capable of controlling cell surface MICA expression on HEK293T cells through either salvage or *de novo* purine synthesis pathways, we were interested in discovering whether other nucleobases or nucleosides possessed the cell permeability and metabolic capacity to rescue MICA expression through the salvage

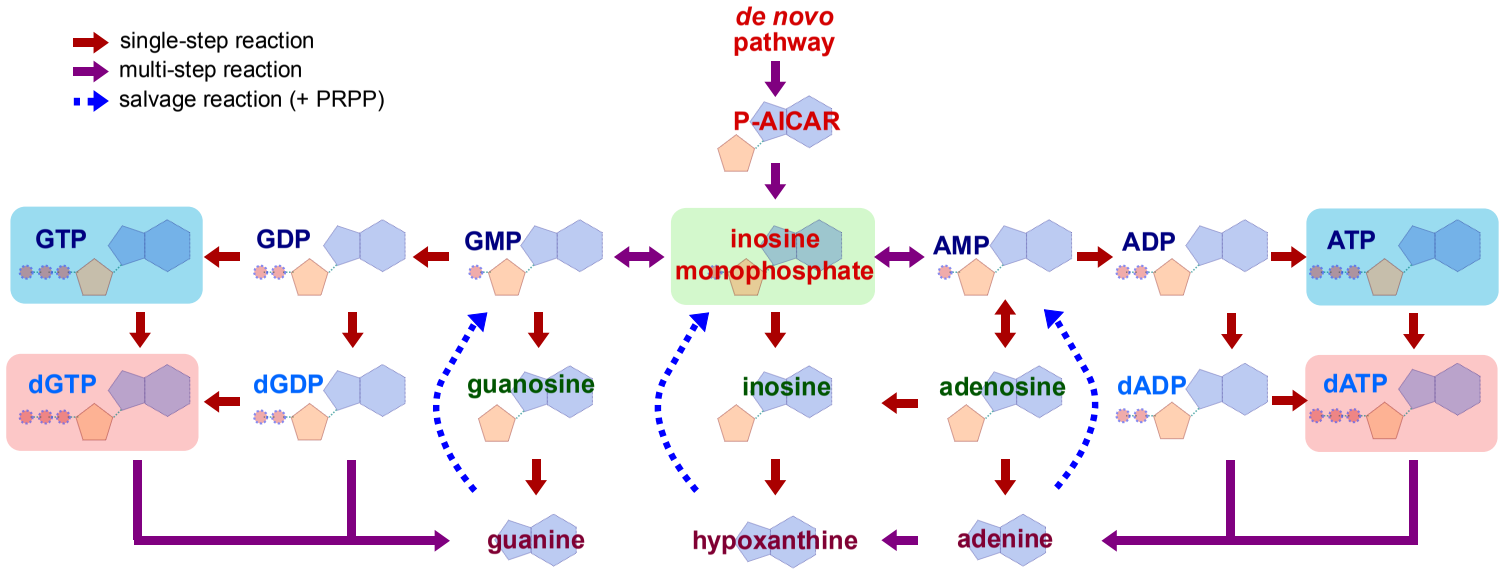

Figure 4.9: Purine nucleotide metabolism. Following purine metabolism downstream from the pathway depicted in **Fig 4.2**, the common pathway of purine synthesis centres on inosine monophosphate (IMP). From IMP, both adenosine and guanosine based nucleotides and deoxynucleotides can be synthesised. Salvage pathway purine synthesis (*blue dashed arrows*) is simpler and more energy efficient. The salvage reaction demonstrated experimentally in **Fig 4.6** is the conversion of hypoxanthine to inosine monophosphate, mediated by the enzyme hypoxanthine guanine phosphoribosyl transferase (HGPRT), and requiring a molecule of PRPP (a metabolite of glucose). As the name suggests, HGPRT also catalyses a similar reaction with guanine. With the addition of PRPP, guanine is converted to GMP, which can in turn be converted to IMP. Adenine phosphoribosyl transferase (APRT) is catalyses the equivalent reaction for adenine and PRPP. In addition to salvage from nucleobases (guanine, hypoxanthine and adenine), the nucleosides can be converted to nucleobases and fuel purine salvage synthesis similarly. Unique to the adenosine arm of purine salvage, humans possess an enzyme called adenosine kinase, which can salvage from adenosine by directly phosphorylating it to form AMP. Notably, adenosine kinase also catalyses the formation of P-AICAR (ZMP) from the exogenous unphosphorylated *de novo* pathway intermediate, AICAR.

pathway following inhibition by azaserine, or whether the ability to rescue MICA expression was unique to hypoxanthine. While it is probable that HEK293T cells have nucleobase and nucleoside transport capacity (246), this is not entirely clear. The non-hypoxanthine routes to purine salvage are illustrated in **Fig 4.9**. The nucleobases guanine, hypoxanthine and adenine (*dark red*) can be salvaged by HGPRT, HGPRT and adenine phosphoribosyl transferase (AdPRT) respectively, with the addition of one molecule of PRPP. The nucleosides (*green*) can act as salvage pathway substrates by being metabolised to their corresponding nucleobases. Adenosine can also be salvaged directly by conversion to AMP by the enzyme adenosine kinase.

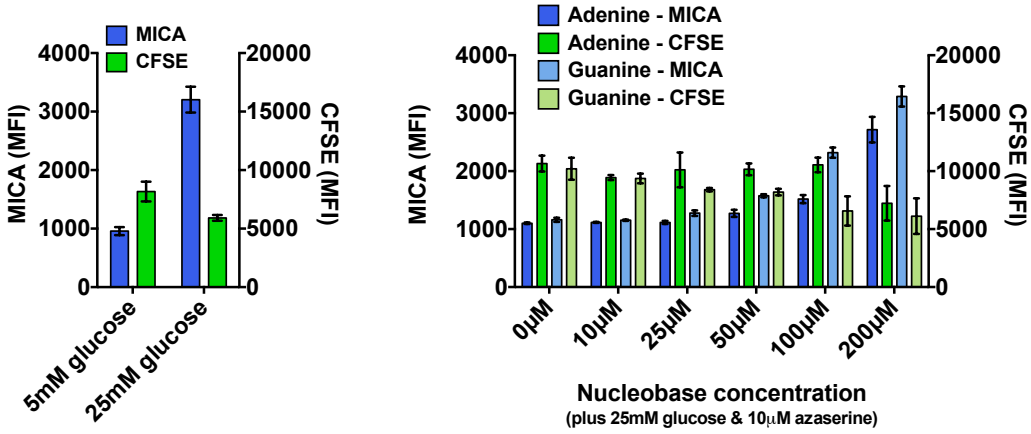

Figure 4.10: Adenine and guanine salvage MICA expression and proliferation. HEK293T cells were cultured in DMEM with 5 mM glucose for 48 hours. Cells were then transferred to either 5 mM and 25 mM glucose (*left*) as a control for MICA expression, or to DMEM containing 25 mM glucose and 10 µM azaserine to inhibit *de novo* purine synthesis (*right*). This inhibited MICA expression and proliferation, but both MICA expression and proliferation were rescued by adding increasing concentrations of either of the nucleobases, adenine or guanine. The pathway to purine nucleotide salvage from adenine and guanine can be seen in **Fig 4.9**. MICA expression and CFSE fluorescence was measured by flow cytometry.

We first tested the ability of other *nucleobases*, namely adenine and guanine, to rescue MICA expression via the purine salvage pathway **(Fig 4.10)**. In the presence of high glucose concentrations, and azaserine to inhibit *de novo* purine synthesis, both adenine and guanine rescued MICA expression, and cellular proliferation. Similarly, nucleosides, deoxynucleosides **(Fig 4.11a)** and the nucleoside, inosine **(Fig 4.11b)** had the same effect. The observation that a wide range of purine salvage pathway substrates can rescue MICA expression in the

presence of *de novo* purine synthesis pathway inhibition, further supports the conclusion that glucose-driven cell surface MICA expression is dependent on purine synthesis.

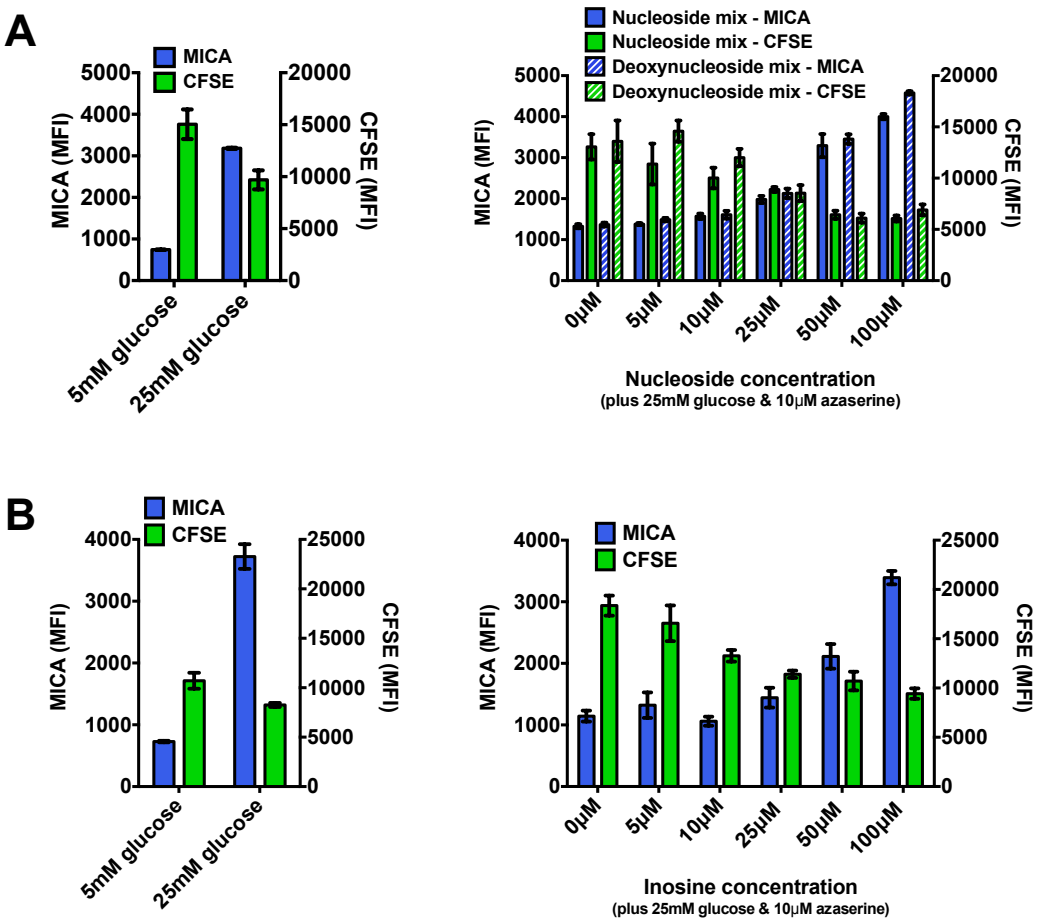

Figure 4.11: Purine nucleosides salvage MICA expression in the presence of glucose and azaserine. (A) HEK293T cells were cultured in DMEM with 5 mM glucose for 48 hours, and then transferred to the indicated conditions. Cells were cultured for a further 48 hours, before MICA expression was measured by flow cytometry. As expected, cells cultured in 25 mM glucose had higher levels of cell surface MICA expression (*left*). Azaserine inhibited both MICA expression and proliferation (*right*). MICA expression and proliferation were restored by increasing concentrations of nucleosides (adenosine, guanosine, cytidine and thymidine) and deoxynucleosides. (B) Cells were prepared in an identical fashion to (A). A separate nucleoside, inosine, was added as a purine salvage pathway substrate, and inosine also rescued MICA expression and cell proliferation.

4.3.2 Purine nucleosides induce MICA expression independently of glucose

Having ascertained that the synthesis of purine nucleotides from PRPP was *necessary* for glucose-driven MICA expression, we wanted to find out if purine nucleotides were *sufficient* for the induction of cell surface MICA expression. Given the evidence for the existence and function of human nucleoside/nucleobase transporters (reviewed by Young *et al* (246)),

and the fact that nucleosides and nucleobases were capable of rescuing MICA expression and cellular proliferation in the context of facilitating the purine salvage pathway, we were confident that cell permeability for these compounds was evident.

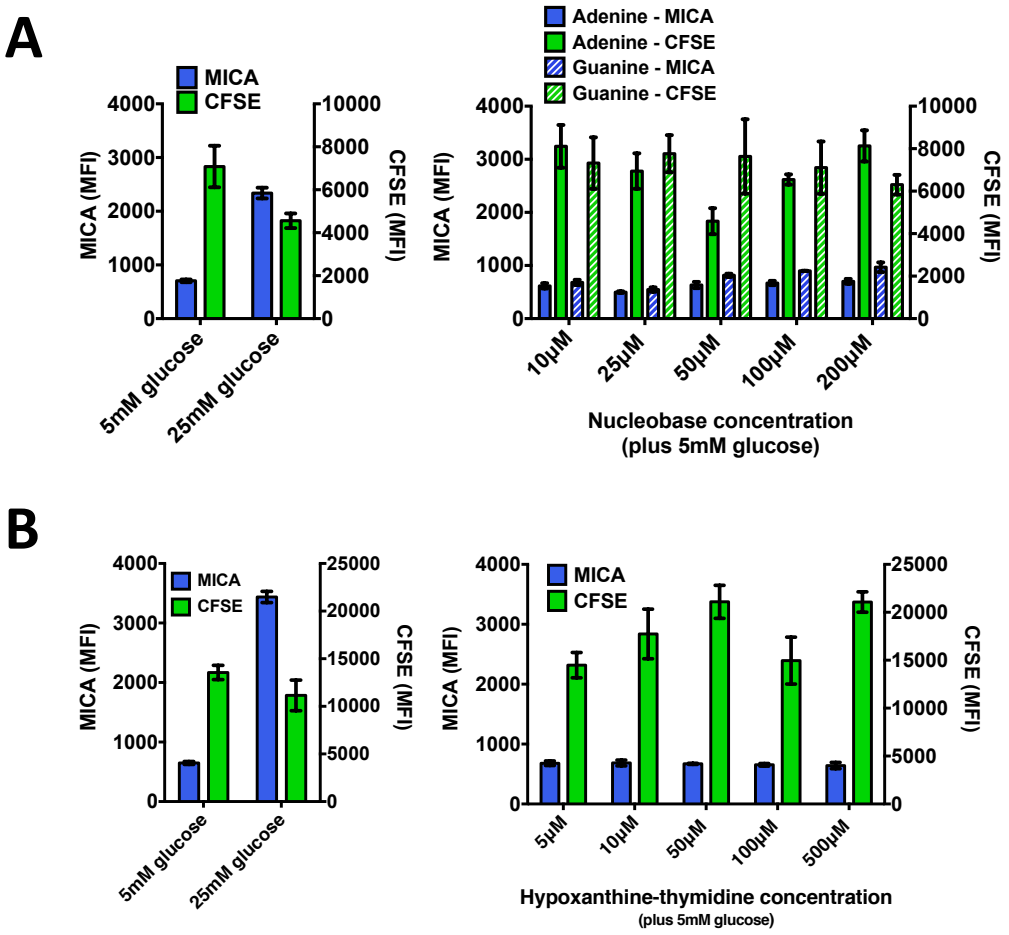

Figure 4.12: Nucleobases are not sufficient for MICA expression. To test if the nucleobases, adenine and guanine (A) or hypoxanthine (B) were sufficient to induce MICA expression as well as being necessary, we cultured HEK293T cells in 5 mM glucose for 48 hours, before removing the culture medium, staining cells with CFSE (to track proliferation) and adding culture medium containing 5 mM glucose and increasing concentrations of nucleobase, up to concentrations found to be capable of salvaging MICA expression in the presence of high glucose and azaserine. Cells were cultured for 48 hours, and MICA expression and CFSE fluorescence were measured by flow cytometry. Adenine, guanine and hypoxanthine did not induce MICA expression at these concentrations.

4.3.2.1 At salvage concentrations, salvage pathway substrates are not sufficient to induce MICA expression

Based on our previous experiments, we began by culturing HEK293T cells in DMEM containing 5 mM glucose for 48 hours, and then transferring cells to DMEM containing 5 mM glucose (instead of 25 mM glucose and azaserine) with increasing doses of nucleobases, up to the same concentrations of nucleobase that salvaged MICA expression from azaserine inhibition. Unexpectedly, the nucleobases, adenine, guanine **(Fig 4.12a)** and hypoxanthine **(Fig 4.12b)** failed to induce MICA expression, or significantly alter cellular proliferation. Similarly, an equivalent experiment with nucleosides or deoxynucleosides **(Fig 4.13a)** and inosine **(Fig 4.13b)** failed to induce cell surface MICA expression.

4.3.2.2 Purine nucleosides (but not pyrimidine nucleosides) induce MICA expression at high concentrations

Metabolomic studies have suggested that ∼15% of glucose transported into a cell is channelled to the pentose phosphate pathway (237). In this case, cells cultured in 25 mM glucose would have ∼2-2.5 mM PRPP to support *de novo* purine synthesis and salvage pathways, or in the case of cells cultured with azaserine, salvage pathways alone. In light of this, given that salvage concentrations of nucleosides and nucleobases failed to induce MICA expression, we cultured CFSE-stained 293T cells for 48 hours in DMEM containing 5 mM glucose, and milli molar concentrations of purine or pyrimidine nucleosides, and then measured MICA expression and CFSE fluorescence by flow cytometry **(Fig 4.14)**. We found that milli molar concentrations of purine nucleosides, but not pyrimidine nucleosides, did induce cell surface MICA expression **(Fig 4.14a)**. Surprisingly, and in contrast with our previous observations, proliferation in these cells was inversely proportional to the level of purine-induced MICA expression, i.e. cells with higher purine concentrations, had higher MICA expression, and higher CFSE levels, meaning relatively lower rates of proliferation **(Fig 4.14b)**.

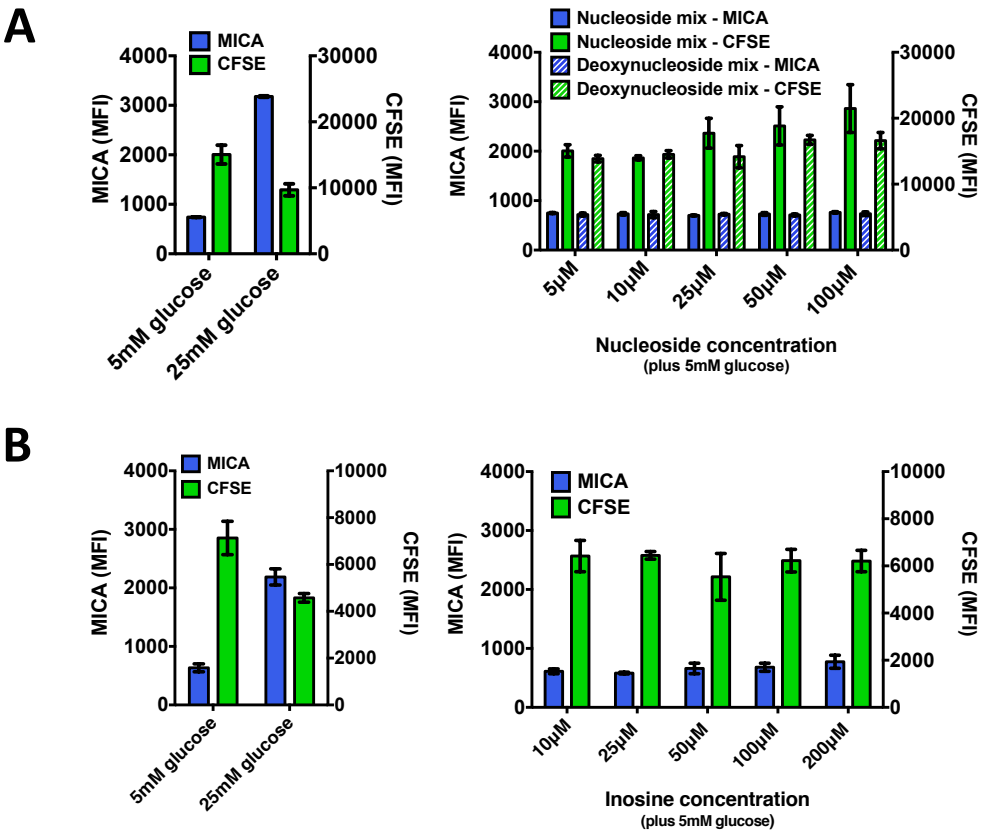

Figure 4.13: Nucleosides are not sufficient for MICA expression. HEK293T cells were stained with CFSE and cultured in medium containing 5 mM glucose and increasing doses of a mixture of nucleosides or deoxynucleosides (adenosine, guanosine, cytidine, thymidine), up to concentrations capable of salvaging purine synthesis (A), or an equivalent range of concentrations for the nucleoside, inosine (B). After 48 hours of culture, flow cytometry was used to measure MICA and CFSE fluorescences. No significant change in the level of proliferation or MICA expression was found.

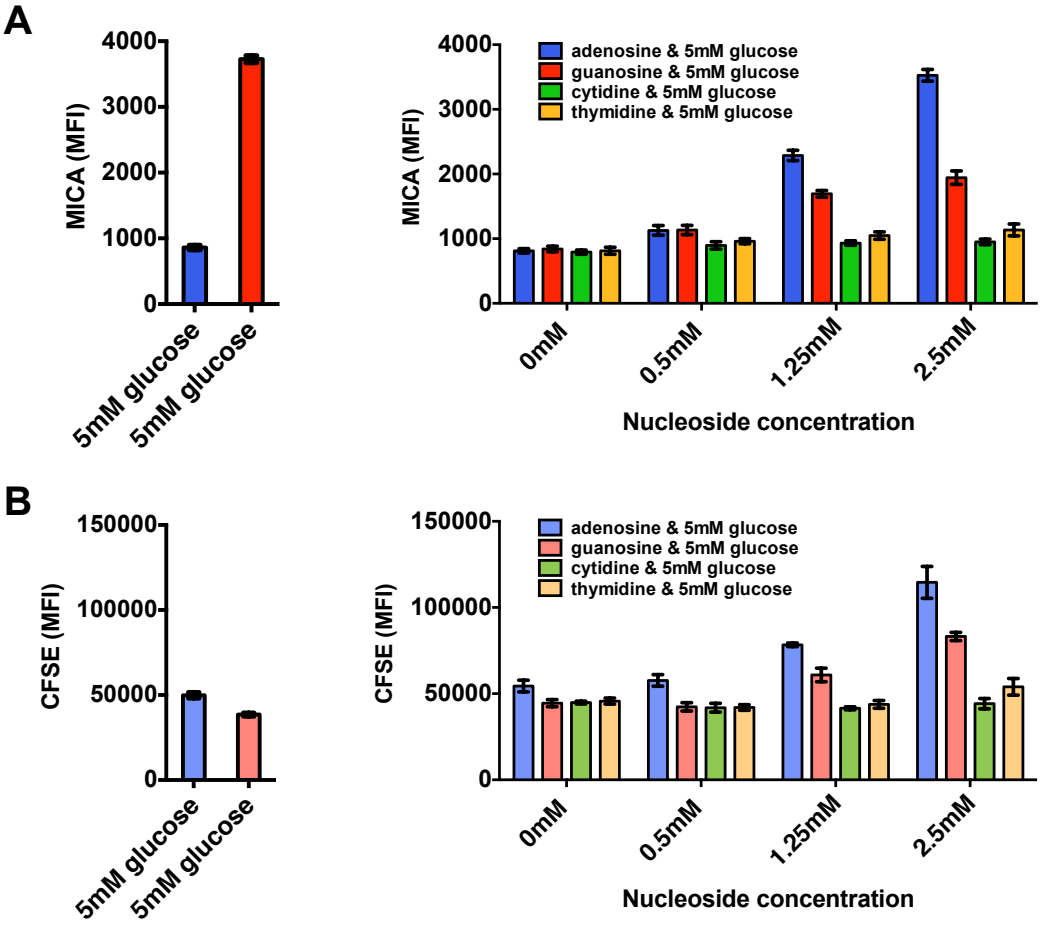

Figure 4.14: High dose purine nucleosides induce MICA expression. In order for glucose to affect cell surface NKG2D regulation, HEK293T cells must have a functional purine synthesis pathway **(Fig 4.7)**. To see if purine nucleosides were *sufficient* to induce cell surface MICA expression, as well as being necessary, we cultured HEK293T cells in 5 mM glucose for 48 hours, before staining the cells with CFSE, and replacing the culture medium with DMEM containing 5 mM glucose, and increasing concentrations of pyrimidine or purine nucleosides. After 48 hours of culture in these conditions, we measured cell surface MICA expression (A), and CFSE fluorescence (B) by flow cytometry. We found that the purine nucleosides, adenosine and guanosine both induced cell surface MICA expression, while pyrimidine nucleosides had no effect on MICA. Further, opposite to the effect of glucose on proliferation, purine nucleosides reduced cellular proliferation (higher CFSE fluorescence) in these cells.

4.3.2.3 Purine nucleosides (but not purine nucleobases) induce MICA expression at high concentrations

Further evidence for the importance of the contribution of the ribose component of purine nucleosides at high concentrations to NKG2D ligand induction was found when comparing nucleosides with their corresponding nucleobases. HEK293T cells were cultured in 5 mM glucose for 48 hours, and then cultured in 5 mM glucose with increasing doses of nucleosides or nucleobases for 48 hours, before MICA expression was measured by flow cytometry. We found that nucleosides (adenosine and inosine; pentose sugar-containing), but not their corresponding nucleobases (adenine and hypoxanthine; no pentose sugar), are capable of inducing MICA expression at high concentrations **(Fig 4.15)**.

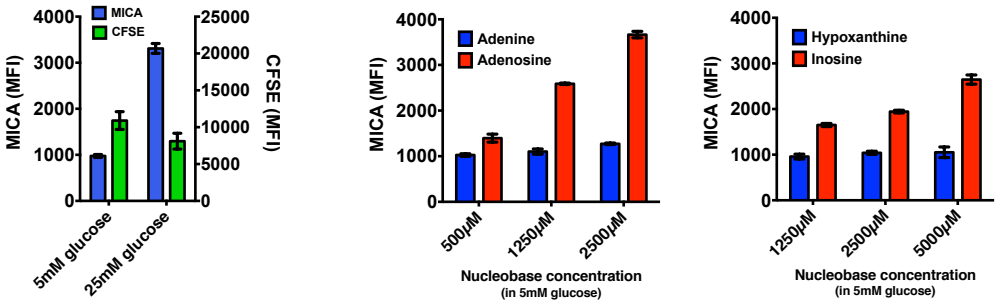

Figure 4.15: Nucleosides, but not nucleobases, induce MICA expression. To see if the ribose moiety of purine nucleosides was important for their ability to induce cell surface MICA expression in low glucose culture medium, we cultured 293T cells in 5 mM glucose for 48 hours, and then in fresh medium, with the conditions indicated for a further 48 hours, before measuring MICA expression by flow cytometry. The left histogram shows the magnitude of cell surface MICA induction for comparison. Both adenosine (middle graph, red) and inosine (right graph, red) induced cell surface MICA expression, independently of the glucose concentration. However the corresponding nucleobases, adenine and hypoxanthine respectively had no impact on MICA expression. This suggests that the ribose moiety is important for the effect of purines on MICA expression.

4.3.3 AICAR, an unphosphorylated *de novo* pathway intermediate, induces MICA expression

5-Aminoimidazole-4-carboxamide 1-β-D-ribofuranoside (AICAR) is the nucleoside (unphosphorylated) counterpart of the nucleotide P-AICAR (also referred to in the literature as ZMP), an intermediate in the common pathway of *de novo* purine synthesis, formed after

PRPP, and two enzymatic reactions before IMP, the guanosine-adenosine branch-point of the *de novo* pathway (see **Figs 4.2, 4.9 & 4.27**). Exogenous AICAR can be transformed to endogenous P-AICAR through the action of the enzyme, adenosine kinase (247). In this form, in addition to being a metabolic intermediate in the *de novo* purine synthesis pathway, as a structural analogue of AMP, P-AICAR is also recognised as a chemical activator of the 'energy sensing' signalling protein, AMP-activated protein kinase (AMPK). AMPK plays a role in conserving cellular energy in times of 'low-energy stress', normally activated when cellular AMP levels are high (reviewed by Vazquez-Martin *et al* (248)). Given that we can infer from the multitude of published AICAR-AMPK signalling papers that the mammalian cell membrane is permeable to AICAR, and that AICAR is also a purine nucleoside, we explored its impact on cell surface MICA expression in our HEK293T cell model.

4.3.3.1 Salvage, and low dose AICAR-induced MICA induction

We first replicated previous experiments to see if AICAR could act as a salvage pathway substrate, and rescue MICA expression in cells cultured in medium containing 25 mM glucose, with inhibition of the *de novo* purine synthesis pathway by azaserine. We found that, in common with previously tested nucleosides and nucleobases, AICAR at low doses was able to salvage MICA expression and cellular proliferation under these conditions **(Fig 4.16)**.

Similarly, when we treated HEK293T cells cultured in medium containing 5 mM glucose with the same concentrations of AICAR that were sufficient to rescue MICA expression from azaserine inhibition for 48 hours, we found that AICAR at these concentrations was not sufficient to induce cell surface MICA expression (measured by flow cytometry; **Fig 4.17)**.

4.3.3.2 AICAR at higher doses strongly induces MICA expression and inhibits cell proliferation

In our earlier experiments, we observed that higher doses of purine nucleosides, but not pyrimidine nucleosides or purine nucleobases, were able to induce cell surface MICA ex-

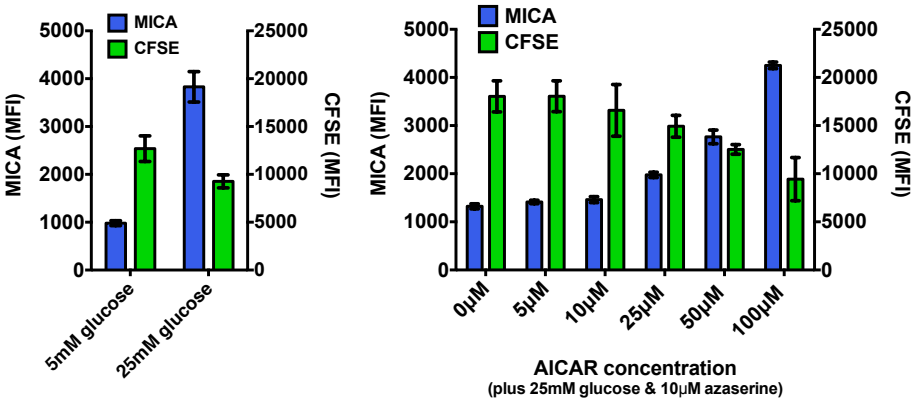

Figure 4.16: AICAR can salvage MICA expression. AICAR is a *de novo* purine synthesis pathway intermediate, that is cell permeable and located two enzymatic steps from the central common nucleotide, IMP. To see if AICAR was capable of acting as a salvage pathway substrate, we cultured HEK293T cells in 5 mM glucose-containing DMEM for 48 hours, and then replaced this medium with DMEM containing 25 mM glucose and azaserine to inhibit the *de novo* pathway. The magnitude of the effect of glucose on MICA expression is shown on the left. In the graph on the right, as expected, azaserine inhibited both MICA expression and cellular proliferation in the presence of 25 mM glucose. Increasing doses of AICAR over came this inhibition, and restored MICA expression, suggesting that AICAR can also act as a salvage pathway substrate to induce cell surface MICA expression in these conditions.

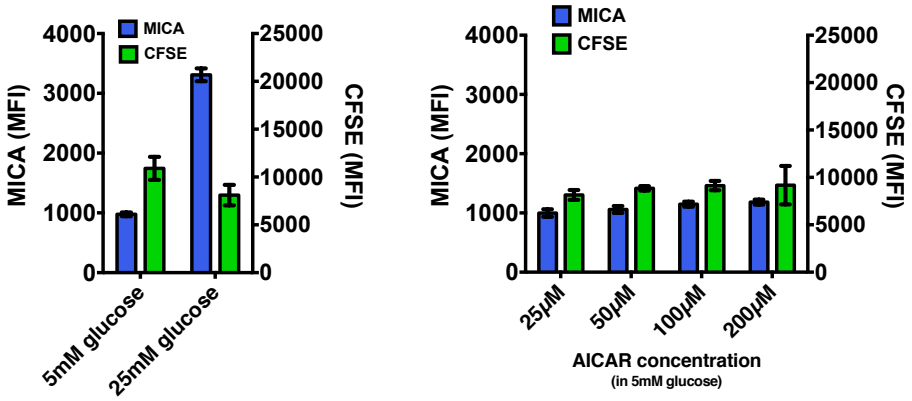

Figure 4.17: AICAR at low doses does not induce MICA expression. Given the location of AICAR in the *de novo* purine synthesis pathway, we expected that AICAR at doses required to salvage purine synthesis may also be able to induce MICA expression at the cell surface. We cultured HEK293T cells in medium containing 5 mM glucose for 48 hours, and then replaced this with fresh DMEM, with 5mM glucose, and the AICAR concentrations shown. The magnitude of the response of MICA to glucose is shown on the left. AICAR alone at low doses failed to induce MICA expression.

pression independently of the glucose concentration. Because AICAR is a *de novo* purine synthesis pathway intermediate, we anticipated that high doses of AICAR might also induce MICA expression independently of glucose. When we tested this, by culturing HEK293T cells in DMEM with 5 mM glucose and AICAR, we found that like adenosine and guanosine, AICAR induced cell surface MICA expression, while inhibiting cell proliferation **(Fig 4.18)**.

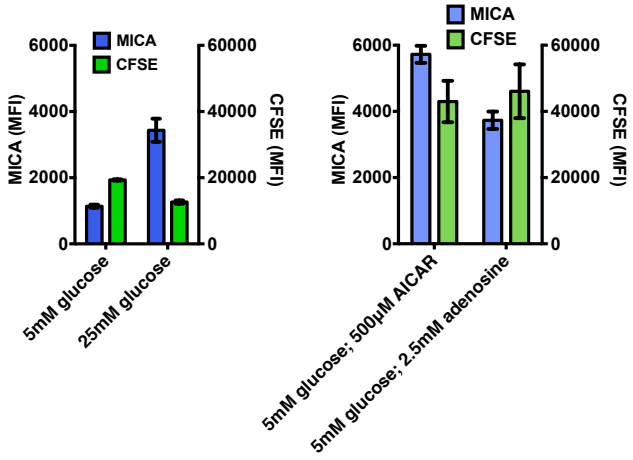

Figure 4.18: AICAR at high doses induces MICA expression. As a nucleoside with close metabolic relationship with adenosine and guanosine, we expected that higher doses of AICAR may induce cell surface MICA expression. To test this HEK293T cells were cultured in DMEM with 5 mM glucose for 48 hours, at which point the cells were stained with CFSE to track proliferation, and the culture medium was then replaced with fresh DMEM containing 5 mM glucose and AICAR or adenosine at the concentrations shown (*right*). The magnitude of cell surface MICA induction and proliferation is shown on the left. Similar to adenosine, AICAR strongly induce cell surface MICA expression, and inhibited cellular proliferation.

4.3.3.3 AICAR-induced MICA expression is dependent on adenosine kinase activity

Once inside the cell, AICAR can be phosphorylated to P-AICAR (ZMP) and metabolised to IMP and other downstream purine nucleotides, or in excess can lead to the formation of the other Z-nucleotides, ZDP and ZTP (249). To see if P-AICAR/ZMP formation was important for the effect of AICAR on MICA induction, we used two inhibitors of adenosine kinase, adenosine kinase inhibitor (a proprietary compound from Merck; **Fig 4.19a**), and iodotubercidin **(Fig 4.19b)**. Both of these compounds successfully prevented AICAR-induced MICA expression on HEK293T cells. While the impact of glucose on MICA expression was not lim-

ited by the adenosine kinase inhibitors, and the effects of adenosine and inosine were slightly reduced, it should be noted that glucose, adenosine and inosine each have metabolic routes to AMP (ZMP structural analogue) formation that do not require the action of adenosine kinase (via the normal purine *de novo* or AdPRT/HGPRT salvage pathways). Taken together, these results suggest that P-AICAR/ZMP formation is largely responsible for AICAR-induced MICA expression on HEK293T cells, and that adenosine salvage by adenosine kinase may also play a role in adenosine-induced MICA expression.

4.3.4 Glucose and purine nucleosides: a common mechanism?

It is difficult to prove that glucose, AICAR, and the purine nucleosides, adenosine, inosine and guanosine are inducing cell surface NKG2D ligand expression in our HEK293T cell model by acting through a common mechanism. To develop some evidence that would support this position, we hypothesised that if these molecules shared a common mechanism, they would induce expression of the same NKG2D ligands (i.e. each stimulus would induce the expression of MICA, ULBP1, ULBP2, ULBP3 and HLA class I, similarly to glucose; **Fig 3.7**), and also that azaserine, which reproducibly and conclusively prevents the effect of glucose on cell surface MICA expression through inhibition of *de novo* purine synthesis, would have no impact on the ability of AICAR and purine nucleosides to induce MICA expression, as these molecules are distal to APRT in the purine synthesis pathway.

4.3.4.1 AICAR, purine nucleosides, and glucose similarly induce NKG2D ligand expression

We had previously observed that glucose induces the expression of MICA, ULBP1, ULBP2 and ULBP3 **(Fig 3.7)**, as well as HLA class I **(Fig 3.8)**. If AICAR and purine nucleosides were acting through a common mechanism with glucose to induce cell surface MICA expression, we expected AICAR and purine nucleosides to have similar effects on these other NKG2D ligands, and HLA class I. To test this, we cultured CFSE-stained HEK293T cells

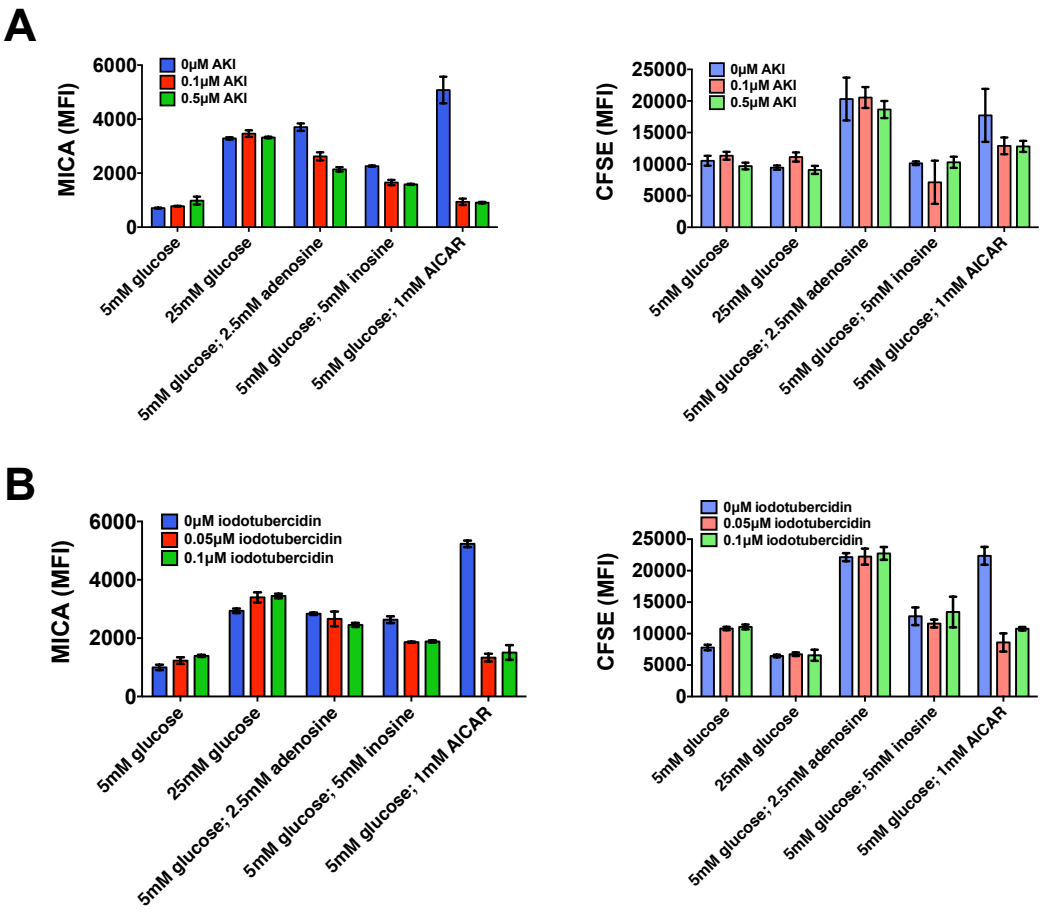

Figure 4.19: Adenosine kinase is necessary for AICAR-induced MICA expression. To see if the action of adenosine kinase in phosphorylating AICAR to form P-AICAR/ZMP was important for AICAR-induced cell surface MICA expression, we cultured 293T cells in 5 mM glucose for 48 hours, stained the cells with CFSE, and then added fresh DMEM containing 5 mM glucose and the additional treatments shown. Cells were cultured in these conditions for a further 48 hours, before surface MICA expression and CFSE fluorescence were measured by flow cytometry. The passage of glucose, via ribose 5-phosphate and PRPP to finished purine nucleotides via the *de novo* pathway occurs without the involvement of adenosine kinase. While adenosine can be salvaged to AMP by adenosine kinase, it can also be salvaged via salvage from nucleobases, without requiring adenosine kinase. We found that both the proprietary adenosine kinase inhibitor (A), and iodotubercidin (B) almost completely blocked the ability of AICAR to induce MICA expression. Both compounds had modest effects on the ability of adenosine and inosine to induce MICA expression, and no effect on the ability of glucose to induce MICA expression. The adenosine kinase inhibitors also reduced AICAR's ability to inhibit cellular proliferation, while the impact of glucose, adenosine, and inosine on cell proliferation was unaffected.

in 5 different conditions: DMEM with 5 mM glucose, 25 mM glucose, 5 mM glucose and 2.5 mM adenosine, 5 mM glucose and 5mM inosine, and 5mM glucose with 750 µM AICAR. After 48 hours, we stained cells with antibodies for each NKG2D ligand, and HLA class I, and measured cell surface expression and CFSE fluorescence by flow cytometry **(Fig 4.20)**.

As expected, we found that AICAR and purine nucleosides, in addition to inducing MICA expression, also induced ULBP2, ULBP3 and HLA class I with marginal effects on ULBP1. This observation is consistent with a common mechanism of action. We also measured the CFSE fluorescence in each sample, and found that NKG2D ligand induction and proliferation were not related, in agreement with previous observations **(Fig 4.21)**.

4.3.4.2 Purine-driven MICA expression is independent of azaserine inhibition

If AICAR and purines, like glucose, were acting through purine *de novo* or salvage pathways to induce cell surface NKG2D ligand expression, we would also expect that unlike glucose, AICAR and nucleoside-mediated MICA expression would not be inhibited by azaserine, which acts on the *de novo* purine synthesis pathway. As expected, azaserine inhibited the effect of glucose on cell surface MICA expression, but had no effect on MICA induction by adenosine, inosine, or AICAR **(Fig 4.22)**.

4.3.5 Purine receptors, NLRC5 activity, and ATM/ATR signalling do not account for the effect of glucose and purines

It remains unclear how glucose, AICAR and purine nucleosides influence cell surface NKG2D ligand expression. We know that the total amount of cellular protein changes **(Fig 3.10)**, which suggests that the mechanism is not an effect on cellular NKG2D ligand distribution or transport. Below we describe some mechanisms that are unlikely to account for the observed effects.

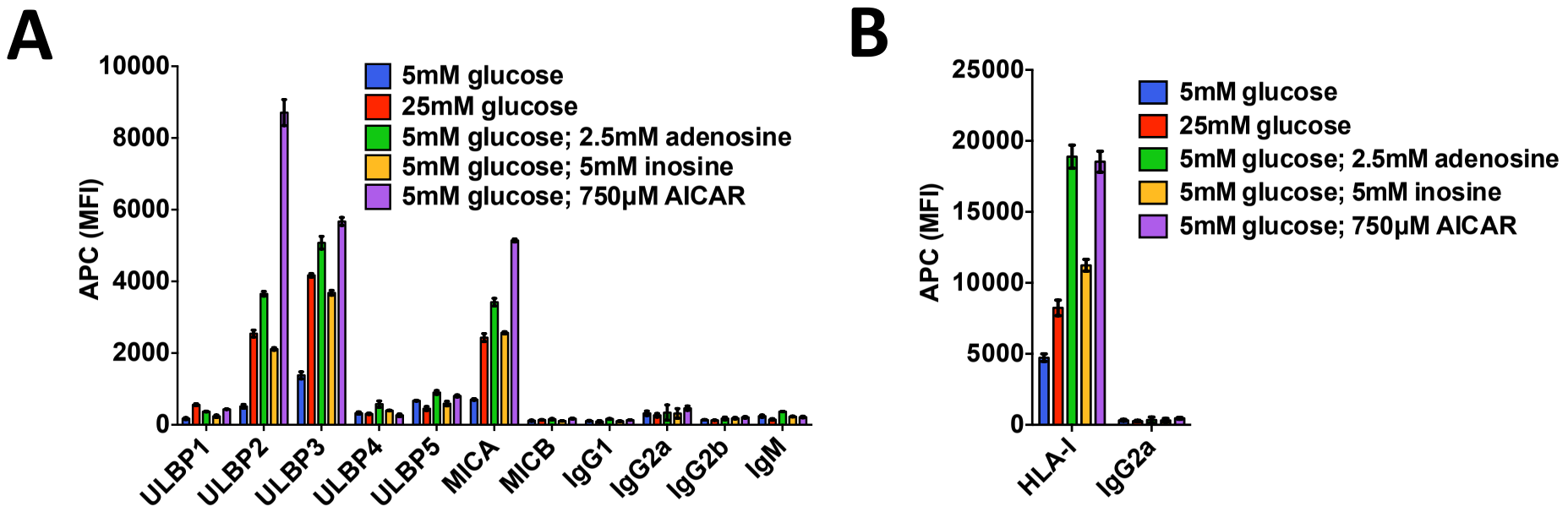

Figure 4.20: Purines and glucose share NKG2D ligand expression patterns. To see if AICAR and purine nucleosides induce cell surface expression of the same NKG2D ligands as glucose, we cultured HEK293T cells in 5 mM glucose for 48 hours, stained the cells with CFSE, and then cultured cells for a further 48 hours in DMEM containing the glucose and purine concentrations indicated. Examination of NKG2D ligand expression by flow cytometry showed that glucose, AICAR and purine nucleosides each induced cell surface expression of HLA class I, MICA, ULBP2, and 3, with marginal effects on ULBP2. This is consistent with action through a common mechanism.

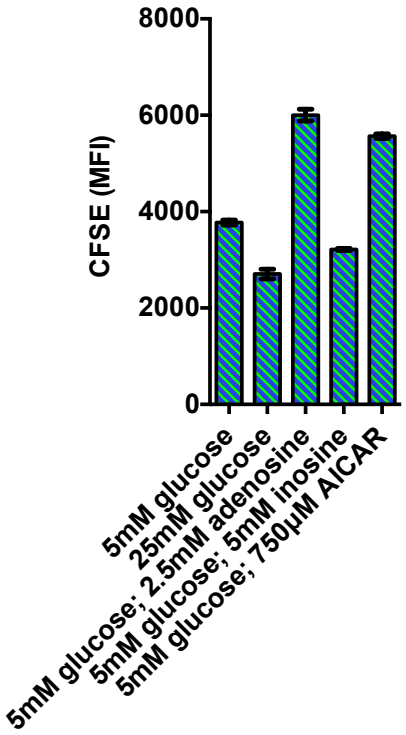

Figure 4.21: Purine and glucose stimuli have differing effects on cellular proliferation. HEK293T cells from **(Fig 4.20)** were stained with CFSE immediately prior to culture in the test conditions for 48 hours. Flow cytometry was used to measure CFSE fluorescence. While glucose, AICAR and purine nucleosides induce expression of the same NKG2D ligands, they each have diverse effects on cellular proliferation, and this does not correlate with the direction of NKG2D ligand expression.

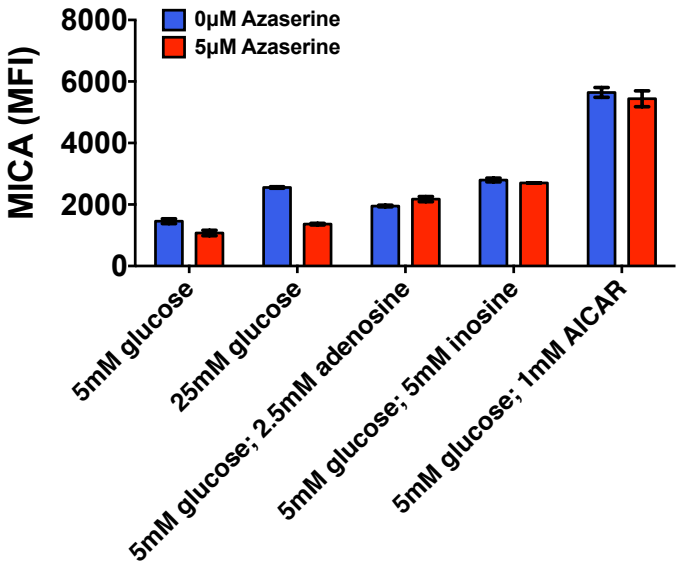

Figure 4.22: Purine-mediated MICA expression is not susceptible to azaserine inhibition. To see if adenosine, inosine and AICAR could induce cell surface MICA expression in the presence of *de novo* pathway inhibitor, azaserine, we cultured HEK293T cells in 5 mM glucose for 48 hours, and then replaced the medium with fresh DMEM in the conditions shown. Cells were cultured for a further 48 hours, before cell surface MICA expression was measured by flow cytometry. We found that azaserine inhibited glucose induced MICA expression, but did not affect nucleoside- or AICAR-induced MICA expression, suggesting that AICAR and purine nucleosides act downstream of APRT, the enzyme inhibited by azaserine.

4.3.5.1 Purine receptor antagonists don't block MICA up-regulation

Purine receptors are a large and growing family of cellular receptors, divided into two subgroups. The P1 purine receptors are receptors for adenosine (nucleoside), and the P2 receptors are nucleotide receptors. Purine receptors are primarily localised to the cell membrane, but some reports have described intracellular localisation (250). They are thought to play roles in a wide range of cellular processes from the control of glycolysis (251), to cell proliferation (252, 253), and cellular immunity (reviewed by Haskó *et al* (254)). The purine receptors and their agonists/antagonists are reviewed in detail by Burnstock (255).

We wanted to explore the possibility that glucose- or purine nucleoside-induced MICA expression in our HEK293T cell model is dependent on purine receptor signalling. To do this, we used two general inhibitors of P1 adenosine receptors, 3-isobutyl 1-methylxanthine (IBMX) and theophylline, and two general inhibitors of P2 nucleotide receptors, suramin and PPADS. To see if we could block glucose-mediated NKG2D ligand induction with these

purine receptor inhibitors, we cultured HEK293T cells in 5 mM glucose for 48 hours, and then replaced this medium with DMEM containing 5 mM or 25 mM glucose, with increasing concentrations of each inhibitor centring on published effective sub-toxic concentrations **(Fig 4.23)**. The cells were cultured in these conditions for 48 hours, before cell surface MICA expression was measured by flow cytometry. While suramin slightly limited glucose-induced MICA expression at its higher concentrations, there was some cytotoxicity at these high doses. Taken together, it is less probable that glucose induces NKG2D ligand expression through purine receptor signalling.

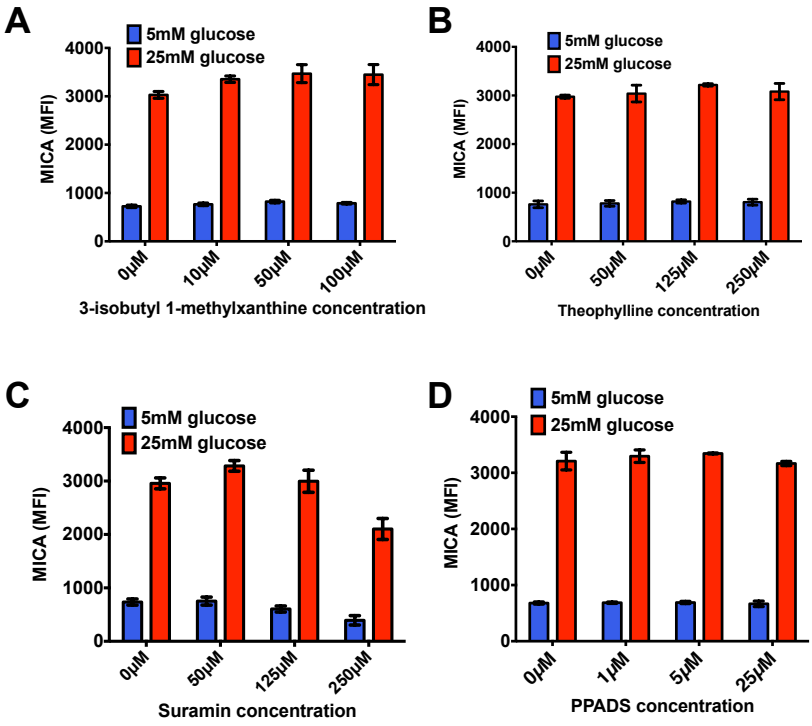

Figure 4.23: Purine receptor inhibitors fail to block the effect of glucose on MICA expression. To see if purine receptor signalling was likely to be responsible for glucose- and purine nucleoside-induced NKG2D ligand expression, we treated HEK293T cells cultured in low and high glucose concentrations with two inhibitors of P1 purine receptors, IBMX (A) and theophylline (B), and two inhibitors of P2 receptors, suramin (C) and PPADS (D). Suramin marginally limited glucose-induced MICA expression at high concentrations. However, some cytotoxicity was noted at this dose, and this dose is above the expected therapeutic dose to inhibit purine receptors. On balance, it is less likely that glucose-induced NKG2D ligand expression is dependent on purine receptor signalling.

In a similar experiment, the four purine receptor inhibitors failed to prevent adenosine-mediated MICA expression, for HEK293T cells cultured in 5 mM glucose **(Fig 4.24)**. Again, these results make a role for purine receptor signalling less probable in this setting.

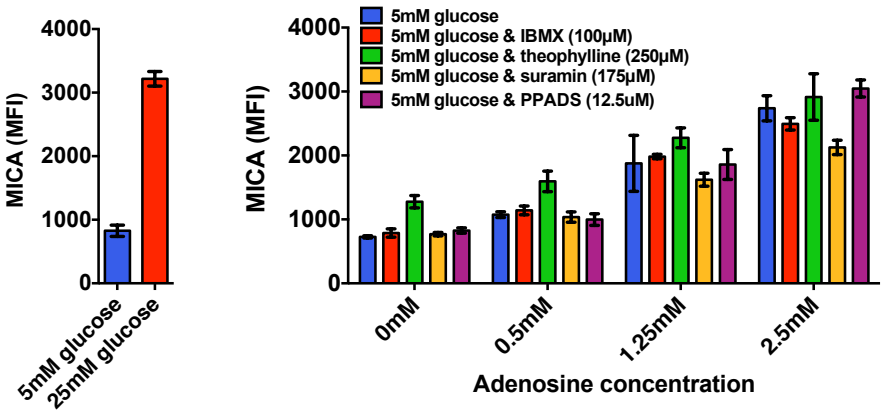

Figure 4.24: Purine receptor inhibitors fail to block the effect of adenosine on MICA expression. HEK293T cells were cultured in 5 mM glucose for 48 hours, which was then replaced with fresh DMEM containing 5 mM glucose, and increasing concentrations of adenosine. all four purine receptor inhibitors failed to prevent adenosine-induced up-regulation of MICA, suggesting that purine receptor signalling is unlikely to be involved in purine nucleoside-mediated NKG2D ligand regulation.

4.3.5.2 NLRC5 does not control MICA expression

Nucleotide-binding oligomerization domain-like receptor C5 (NLRC5) is an intracellular nucleotide binding receptor that has been shown to control cell surface levels of HLA class I molecules (256). MICA and MICB are encoded on chromosome 6, adjacent to the HLA class I locus. While the dose-response relationship between nucleotide binding and activity of NLRC5 has not been clearly established, we thought it a plausible hypothesis that increased purine nucleotide production would lead to increased NLRC5 nucleotide binding, and increased expression of HLA class I and NKG2D ligands, as we have observed in our model.

To test this, we replicated NLRC5 over-expression experiments carried out by Meissner *et al* (256). We purchased a 5.6 kb myc-tagged NLRC5 construct in the pCDNA3 mammalian expression vector from Addgene (deposited by Neerincx *et al* (257)). We cultured HEK293T cells in 5 mM glucose for 48 hours, before replacing the medium with either 5 mM or 25 mM glucose. Cells were either transfected with a low (0.5 µg) or high (2 µg) amount of the NLRC5 expression plasmid. A pCDNA3.1-glut5 expression plasmid and untransfected cells

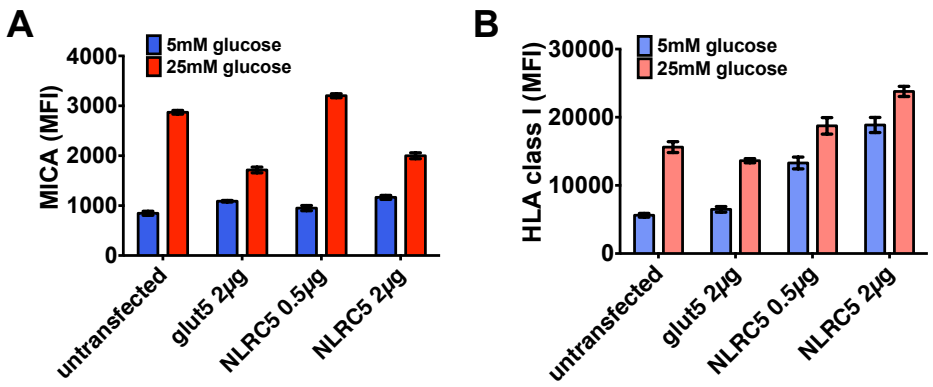

Figure 4.25: NLRC5 does not control MICA expression in 293T cells. We measured cell surface MICA (A) or HLA class I (B) expression on HEK293T cells transfected with low or high doses of an expression plasmid encoding NLRC5. As a negative control, we also stained untransfected cells, and cells transfected with an expression plasmid encoding and unrelated protein (pCDNA3.1-glut5). The NLRC5 expression vector induced cell surface HLA class I expression in a dose-dependent manner as anticipated, however, it had no effect to increase MICA cell surface expression at low or high glucose. This suggests that it is improbable that NLRC5 plays a role in glucose-induced NKG2D ligand expression.

were used as negative controls **(Fig 4.25a)**. As a positive control, we separately measured HLA class I expression on the same cells, to confirm the effectiveness of transfected NLRC5 in inducing HLA class I expression **(Fig 4.25b)**. While HLA class I expression was efficiently induced by transfected NLRC5 at high and low glucose concentrations, it had no effect on MICA expression under the same conditions. This result suggests that NLRC5 signalling is unlikely to be involved in glucose- or purine nucleoside-mediated NKG2D ligand expression.

4.3.5.3 ATM/ATR signalling (DNA damage) is unlikely to explain the effect of purines

Finally, several reports have implicated ATM/ATR signalling, acting as part of the DNA damage response, in the regulation of NKG2D ligand expression (71, 100). This is most commonly demonstrated in publications by the ability of caffeine to inhibit NKG2D ligand expression in response to a given stimulus. If ATM/ATR signalling was the final common mechanism for NKG2D ligand induction by glucose, AICAR and purine nucleosides, we should expect up-regulation of MICA in response to these compounds to be blocked effectively by caffeine. We cultured HEK293T cells in 5 mM glucose for 48 hours, and then replaced

the medium with fresh DMEM and the treatment conditions shown in **Fig 4.26**. Cells were cultured for a further 48 hours before cell surface MICA expression was measured by flow cytometry. Interestingly, caffeine blocked the up-regulation of cell surface MICA expression by glucose, and AICAR, but not by adenosine or inosine. Caffeine is a purine derivative, and work in the late 1970s demonstrated that caffeine does inhibit *de novo* purine synthesis itself (258). In light of this, as both glucose and AICAR are 'further up' the purine synthesis pathway than the already completed purine bases, adenosine and inosine, this suggests that caffeine is acting on the *de novo* purine synthesis pathway, and not ATM/ATR signalling.

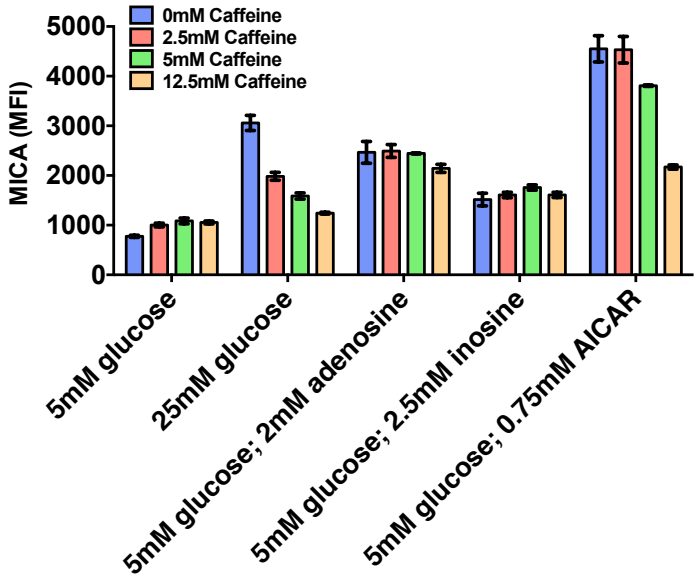

Figure 4.26: ATM/ATR does not control MICA induction by purines in 293T cells. HEK293T cells were treated with NKG2D ligand inducers: glucose, adenosine, inosine and AICAR. Caffeine, itself a purine metabolite, and inhibitor of ATM/ATR signalling, limited the up-regulation of MICA expression by glucose, and by the incomplete purine nucleoside, AICAR. Given that caffeine failed to affect MICA up-regulation by adenosine and inosine, it is unlikely that ATM/ATR signalling represents a final common mechanism for MICA induction by these compounds. Caffeine is also known to inhibit *de novo* purine synthesis, and this is a more probable explanation for the results, considering only the complete purine nucleosides, inosine and adenosine, were unaffected by caffeine.

4.4 Discussion

Having established in **Chapter 3** that an intracellular metabolite of glucose was responsible for glucose-mediated NKG2D ligand expression in our HEK293T cell model, we hypothesised that PRPP, a molecule central to both *de novo* and salvage purine synthesis pathways, was the

glucose metabolite responsible for this activity. Throughout this chapter, we have presented evidence to support this hypothesis, and to show that not only is purine synthesis *necessary* for NKG2D ligand expression in our HEK293T cell model, but purines nucleosides themselves are also *sufficient* to mediate the same effect.

4.4.1 Glucose-induced NKG2D ligand expression requires purine synthesis

Research on metabolic pathways is experimentally challenging, not least owing to limitations imposed by cell membrane permeability and solute transport. This is attested to by the uncertainties that still surround heavily researched metabolic topics, such as glucose-sensing in pancreatic β islet cells (reviewed by MacDonald *et al* (218) and Melloul *et al* (235)). We have capitalised on the existence in mammalian cells of two distinct purine nucleotide synthesis pathways, linked by their mutual dependence on the glucose metabolite, PRPP, to demonstrate that this process is central to NKG2D ligand expression. The data presented in **Figs 4.3 & 4.5** show that azaserine, acting specifically on the *de novo* purine synthesis pathway after PRPP prevents glucose-mediated increases in MICA expression. Importantly, we show that azaserine blocks MICA expression, while leaving the gene expression machinery intact **(Fig 4.8)**. The ability of glucose to control MICA expression through either salvage or *de novo* purine pathways **(Fig 4.7)** underlines the central role that purine nucleosides play in this process, and further supports our original hypothesis that an intracellular metabolite of glucose determines the level of NKG2D expression in these cells.

4.4.2 Purine nucleosides are *sufficient* for NKG2D ligand induction

The data we have presented suggests that over the 48 hour time course of the experiment, a 100 - 200 µM concentration of nucleobase, in the context of abundant PRPP, is sufficient to rescue MICA expression and cell proliferation from *de novo* pathway inhibition by azaserine **(Figs 4.6, 4.10, 4.11 & 4.16)**. At these concentrations, any individual purine nucleobase or nucleoside, including AICAR, overcomes azaserine with equal efficacy. The failure of

both nucleobases and nucleosides at the same concentrations to induce MICA expression in the presence of 5 mM glucose (depleted PRPP; **Figs 4.12, 4.13 & 4.17**) suggests that a metabolite of glucose (our earlier results point to PRPP) is key to rescue in this setting.

This conclusion, that both an adequate ribose source and nucleobase supply together are sufficient to mediate this effect is also supported by the results of our high-dose purine experiments. We noticed that purine nucleosides (purine nucleobase & ribose sugar) but not pyrimidine nucleosides (pyrimidine nucleobase & ribose sugar; **Fig 4.14**), or purine nucleobases (no ribose sugar; **Fig 4.15**) are sufficient to induce NKG2D ligand expression, even in low glucose concentrations. In other words, the role of the ribose sugar of the nucleoside molecule appears as important as the nucleobase component. There is evidence from the literature to suggest that the ribose moiety of nucleosides is, like the nucleobase, salvageable by the purine synthesis pathway, when free nucleobases are present (259, 260). This helps to understand why high dose, but not low dose purines are sufficient to induce MICA expression. If ribose salvage is indeed occurring in our high dose purine nucleoside experiments (e.g. 2.5 mM adenosine), the molar amount of ribose generated would be insufficient to reconstitute glycolysis to the same extent as 25 mM glucose, but the induction of MICA expression by 25 mM glucose or 2.5 mM adenosine is equally strong. This makes it probable that a complete nucleoside (or downstream nucleotide) is truly sufficient to induce NKG2D ligand expression in our HEK293T model, even if some of these nucleosides are broken down and salvaged.

4.4.3 The importance of nucleoside phosphates?

The experiments discussed above provide evidence that purine nucleosides, the synthesis of which is dependent on the glucose metabolite PRPP, are important for NKG2D ligand induction, but do not give an indication whether nucleosides themselves, or their phosphorylated counterparts, nucleotides (see **Fig 4.1** for the distinction) are the active metabolites in mediating this effect. Two results suggest that nucleotides might be more important. First, as discussed above, the ribose moiety of nucleosides appears to be as important as the

nucleobase component in the sufficiency of nucleosides for NKG2D ligand induction. The phosphate group of IMP (the final common molecule of the *de novo* purine synthesis pathway), and the phosphate group of GMP, IMP and AMP when guanine, inosine, and adenosine respectively are salvaged by the HGPRT/AdPRT mediated addition of PPRP, in both cases originates on the PRPP molecule (and before that on the ribose 5-phosphate molecule). Our data in these cases would be best explained if upon entering the cell, nucleosides were split to their constituent parts, ribose and nucleobase, the ribose moiety is phosphorylated to PRPP, and the free nucleobases are subsequently salvaged by the 'recharged' PRPP molecule. This would explain both why low dose nucleosides, and nucleobases alone even in high dose, do not induce MICA expression.

Second, it is well recognised that the purine nucleoside AICAR can be directly phosphorylated by the enzyme adenosine kinase to form the *de novo* pathway intermediate P-AICAR/ZMP. P-AICAR/ZMP, a structural analogue of AMP, is a purine nucleotide with an 'incomplete' purine ring, and can be further phosphorylated to ZDP, and ZTP (249). AICAR, even in low glucose, induces cell surface expression of the same NKG2D ligands and HLA class I as glucose and other purine nucleosides, suggesting a common mechanism of action **(Fig 4.20)**. By blocking the phosphorylation of AICAR to form ZMP, the effect of AICAR on the induction of NKG2D ligands is blocked **(Fig 4.19)**, suggesting that P-AICAR/ZMP (or ZDP/ZTP), and not AICAR itself, is responsible for the effect on MICA expression. It is widely recognised that P-AICAR/ZMP acts as a structural analogue of AMP, and in this capacity can induce signalling through AMP binding signalling molecules such as AMP activated protein kinase (reviewed by Towler and Hardie (261)). While little data exists regarding the downstream fate or biological functionality of ZDP and ZTP, it would be tempting to speculate, in light of our data, that these Z-nucleotides are unlikely to support RNA or DNA synthesis per sé, and that the action of AICAR, purine nucleosides and glucose (via purine nucleosides) on NKG2D ligand expression is based on the action of nucleotides (or a specific nucleotide) as a ligand for an intracellular signalling molecule, analogous to the activation of

AMPK by ZMP. In fact, a wide variety of nucleotide binding signalling proteins are known to exist, not least the nod-like (nucleotide binding oligomerization domain) receptors which are already recognised as important signalling components of the innate immune response (reviewed by Rubino *et al* (262)). The HLA class I-regulating protein, NLRC5, does not however control the glucose mediated NKG2D ligand regulatory process **(Fig 4.25)**.

4.4.4 The role of ATM/ATR signalling

The induction of NKG2D ligand signalling through activation of the DNA damage response pathway and ATM/ATR signalling, discussed in the main introduction, is one of the more attractive published hypotheses describing a molecular mechanism which controls physiological NKG2D ligand induction. It is consistent with reason that DNA damage response pathways would be activated in cancer cells, virus-infected cells, and proliferating cells, where NKG2D ligand induction is also known to occur. The majority of the evidence supporting this hypothesis in several publications, is that caffeine, an inhibitor of ATM/ATR signalling, prevents the induction of NKG2D ligand expression by aphidicolin in mouse fibroblasts (100), HIV in $CD4^+$ T cells (71), and TLR4 agonists in human macrophages (78). If this was a common final mechanism of action of glucose, purines and AICAR, we would expect caffeine to prevent NKG2D ligand induction by these stimuli. Instead, we have found that caffeine limits NKG2D ligand induction by glucose and AICAR, but doesn't influence NKG2D ligand induction by adenosine and inosine, both downstream of AICAR and glucose in the purine synthesis pathway. This raises the possibility that in spite of Gasser *et al*'s ATR silencing experiments, which clearly prevent NKG2D ligand expression in aphidicolin-treated mouse fibroblasts, the impact of caffeine, itself a purine metabolite, is principally an effect on inhibition of purine synthesis, an effect that has been documented previously (258).

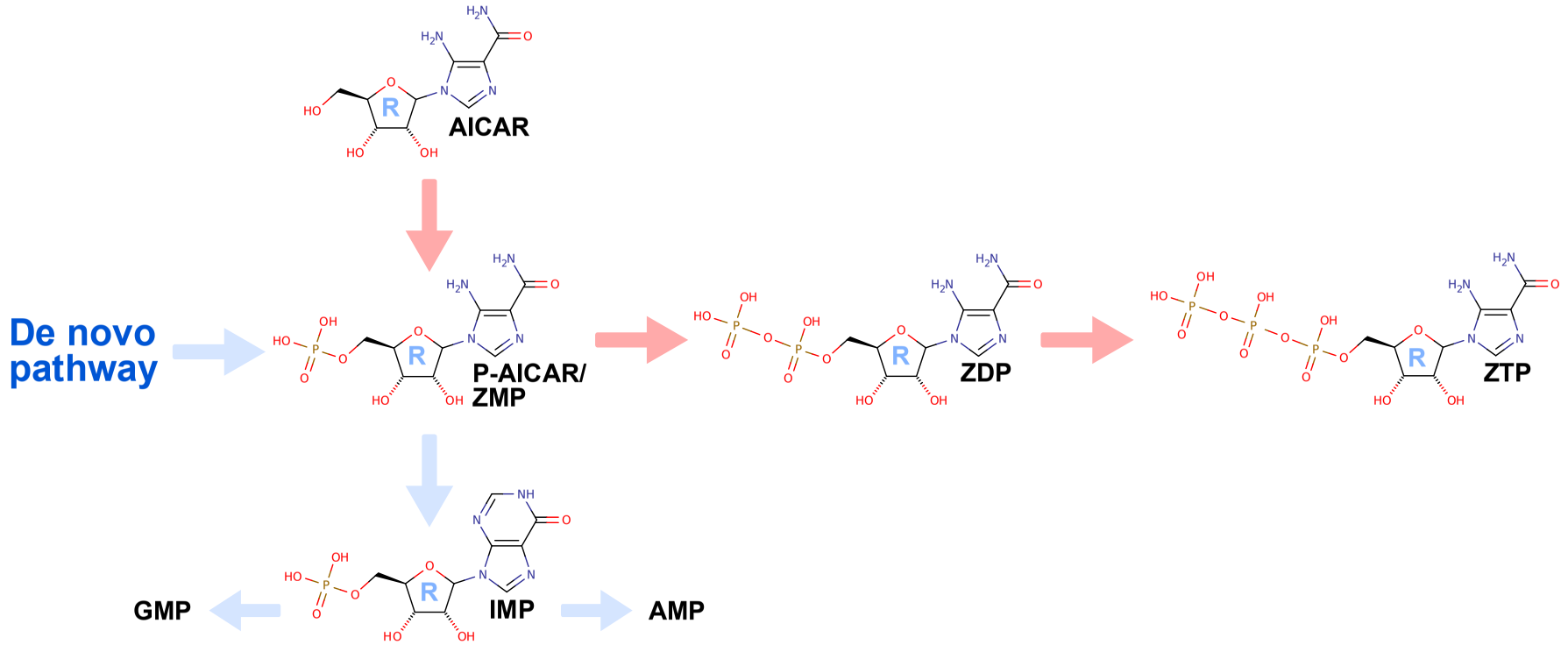

Figure 4.27: The role of exogenous AICAR in purine metabolism. *De novo* purine synthesis from PRPP (represented by *blue* arrows) involves the stepwise construction of a purine nucleobase around the PRPP molecule. In this way, each purine nucleoside intermediate already contains within its structure the phosphoribosyl group that contains the 'phosphate' of inosine monophosphate (IMP). Exogenous AICAR is an unphosphorylated P-AICAR molecule, that exists *in vivo* only in its phosphorylated form (P-AICAR/ZMP) . Exogenous AICAR can be phosphorylated intracellularly by adenosine kinase, to form P-AICAR/ZMP. When exogenous AICAR is present in excess, the multi-phosphorylated AICAR nucleosides ZDP and ZTP can form (*red* arrows). Adenosine kinase inhibitors prevent this reaction. In this case, the ribose moiety of AICAR is remains available for salvage. The chemical structures were drawn with MarvinSketch v6.0.0. R - ribose.

4.4.5 Summary

In summary, in this chapter we have found that the impact of glucose on NKG2D ligand expression in our HEK293T cell model is dependent on glucose contributing to purine synthesis, and that purine nucleosides are sufficient to induce NKG2D ligand expression in their own right, independently of glucose. This evidence remains consistent with our hypothesis that glucose metabolism plays a causal role in the NKG2D immune response, and that purine nucleotides are central to this process.

5

Development of a CMV model of NKG2D ligand expression

Contents

5.1 Abstract

Examination of cell line models suggests that glucose plays an important role in cell surface NKG2D ligand expression, and that the impact of glucose on NKG2D ligand expression is dependent on its role as a substrate for purine synthesis. We sought to identify and develop

a physiological cellular model of NKG2D ligand induction in which to test our hypothesis, supported by this cell line model evidence, that glucose plays a key role in NKG2D ligand induction in physiological settings. Consistent with previous reports, we find that infection of primary human fibroblasts with CMV leads to a shift to Warburg metabolism and cell surface induction of NKG2D ligand expression. We find that NKG2D ligand induction is dependent on *de novo* mRNA synthesis. Additionally, glucose availability determines the level of NKG2D ligand expression achieved, and the role of glucose in this physiological model is dependent on purine synthesis. These results suggest that our findings in the cell line model are relevant to physiological NKG2D ligand induction.

5.2 Introduction

5.2.1 Summary of previous relevant work

The NKG2D ligands are cell surface proteins, the expression of which can induce cytotoxic or co-stimulatory responses from NKG2D receptor expressing immune cells. How NKG2D ligand expression is controlled at the molecular level is unknown. Settings in which cellular NKG2D ligand expression is induced commonly display altered cellular metabolism, known as the Warburg effect. We hypothesised that the metabolic changes that occur in Warburg metabolism are important for the induction of NKG2D ligand expression. To test this, we developed a HEK293T cell model of Warburg metabolism to explore how these metabolic processes might be relevant. We have demonstrated that glucose availability is important for NKG2D ligand expression, and glucose-driven cell surface NKG2D ligand expression can lead to a significant increase in NKG2D-dependent NK cell-mediated cytotoxicity. Furthermore, purine synthesis, a process fuelled by the glucose metabolites ribose 5-phosphate and PRPP, is a central to this effect of glucose. Indeed, purine nucleosides themselves (but not purine nucleobases, or pyrimidines) can induce NKG2D ligand expression independently of glucose availability. Having made these observations in a non-malignant cell line model, HEK293T

cells, we sought to assess the relevance of these observations in a physiological model of NKG2D ligand induction.

5.2.2 CMV as a physiological model of NKG2D ligand expression

The choice of a cytomegalovirus (CMV)-primary fibroblast model for these experiments was based on its fulfilment of several key criteria. It is a human primary cell model, which is genetically and metabolically 'normal'. The expression of NKG2D ligands is inducible, with a 'healthy', metabolically quiescent, NKG2D ligand-negative, uninfected state, and an NKG2D ligand-expressing, metabolically active infected state. This stimulus (infection with CMV) can be induced at a precise time, and is reproducible. It is a scalable system, which can potentially be used for larger experiments, and is a good example of the physiological induction of NKG2D ligand expression. There are several lines of evidence that suggest that this is physiologically valid model.

CMV has wide cellular tropism in humans, with the capacity to infect most human cell types (reviewed by Sinzger *et al* (263)). Specifically, there is clear evidence that fibroblasts are an *in vivo* target of CMV (264). This documentary evidence of fibroblast infection by CMV derived from *ex vivo* tissue blocks is useful, because of all the cell types which CMV can productively infect, evidence of NKG2D ligand induction by CMV only exists for fibroblasts. Groh *et al* were first to notice NKG2D ligand expression following CMV infection in 2001 (60). They reasoned that MICA and MICB may play a role in the immune response to CMV, by way of immune compensation for CMV-induced down-regulation of HLA class I molecules. When they examined this possibility by infecting primary fibroblasts with the AD169 CMV strain, they found strong induction of MICA/B at the cell surface. Cell surface and mRNA-level NKG2D ligand induction in fibroblasts by CMV strains has since been confirmed by several authors (65, 67).

The physiological relevance of NKG2D ligand-mediated signalling in the *in vivo* immune response to CMV can be inferred from the genes encoding NKG2D evasion mechanisms that

have evolved in the CMV genome to limit NKG2D ligand cell surface expression in infected cells (reviewed by Mocarski *et al* (265)). CMV's UL16 protein, from which the ULBPs get their name, binds to and reduces cell surface expression of MICB, ULBP1 and ULBP2 (23, 67, 266). UL142 reduces cell surface expression of MICA and ULBP3 (267, 268, 269). Stern-Ginossar *et al* also report that the CMV genome encodes microRNAs that target MICB (141).

There is strong evidence that cellular infection with CMV induces Warburg metabolism in infected cells (155, 156, 270).

5.2.3 The CMV AD169ΔUL16 laboratory strain

AD169 is a laboratory CMV strain, originally isolated in 1956 from a 7 year old girl with adenoid tonsillitis (271). AD169 is known to contain several mutations in comparison to wild type CMV, including a deletion of the UL133-UL150 genes (272) and a frameshift mutation in the UL131A gene (273). UL131A is important for endothelial and epithelial cell tropism, and its disruption in AD169 limits efficient productive infection to fibroblasts (274). In addition, Mocarski described a deletion disrupting UL42 and UL43 genes in a UK based AD169 strain, not present in the ATCC AD169 reference strain (275). Here we use an engineered UL16-deficient strain of AD169, AD169ΔUL16. Because AD169ΔUL16 lacks both UL16 and UL142, measurement of cell surface NKG2D ligand expression by flow cytometry is possible.

5.2.4 Aims and objectives

In the work presented here, we aim to confirm the induction of NKG2D ligand expression in primary human fibroblasts by infecting cells with the CMV laboratory strain, AD169ΔUL16, and measuring cell surface NKG2D ligand expression by flow cytometry, and new transcription by RTPCR targeted at unspliced cDNA products. We aimed to quantify the metabolic changes induced by cellular infection by measuring changes in glucose and lactate concen-

trations in culture medium from infected and uninfected cells. We explore the possibility that the available glucose concentration is an important determinant of NKG2D ligand surface expression in infected cells, and that purine synthesis is necessary for glucose-mediated changes in expression. Finally, we assess the influence of glucose and its purine derivatives on NKG2D ligand expression in uninfected cells.

5.3 Results

5.3.1 The effect of CMV AD169ΔUL16 on NKG2D ligands in primary fibroblasts

While there is clear evidence that CMV infection of fibroblasts induces NKG2D ligand expression, the pattern of CMV-induced NKG2D ligand expression reported has been inconsistent. Before assessing the impact of glucose on NKG2D ligand expression, we first wanted to characterise the response of primary human fibroblasts to infection with AD169ΔUL16.

5.3.1.1 Infection with CMV induces cell surface NKG2D ligand expression

Primary human fibroblasts were cultured to confluence in T175 flasks in RPMI1640 culture medium (11.1mM glucose). We added concentrated AD169ΔUL16 viral particles to the confluent fibroblasts at a multiplicity of infection (MOI) of 0.1. Both infected and uninfected flasks were then cultured until ∼90% cytopathic effect had been reached, assessed by light microscopy, in the AD169ΔUL16 infected samples, which occurred at 7 days post-infection. The cells were stained with antibodies to MICA, MICB, ULBP1-5 or HLA class I, or with the appropriate isotype control antibodies, and surface expression was measured by flow cytometry **(Fig 5.1)**. As expected, down-regulation of cell surface HLA class I expression was observed in the CMV-infected cells, confirming active CMV infection. Of the NKG2D ligands, cell surface expression of MICA, MICB and ULBP1 was induced from a baseline of no expression. While uninfected confluent fibroblasts expressed small quantities of ULBP2,

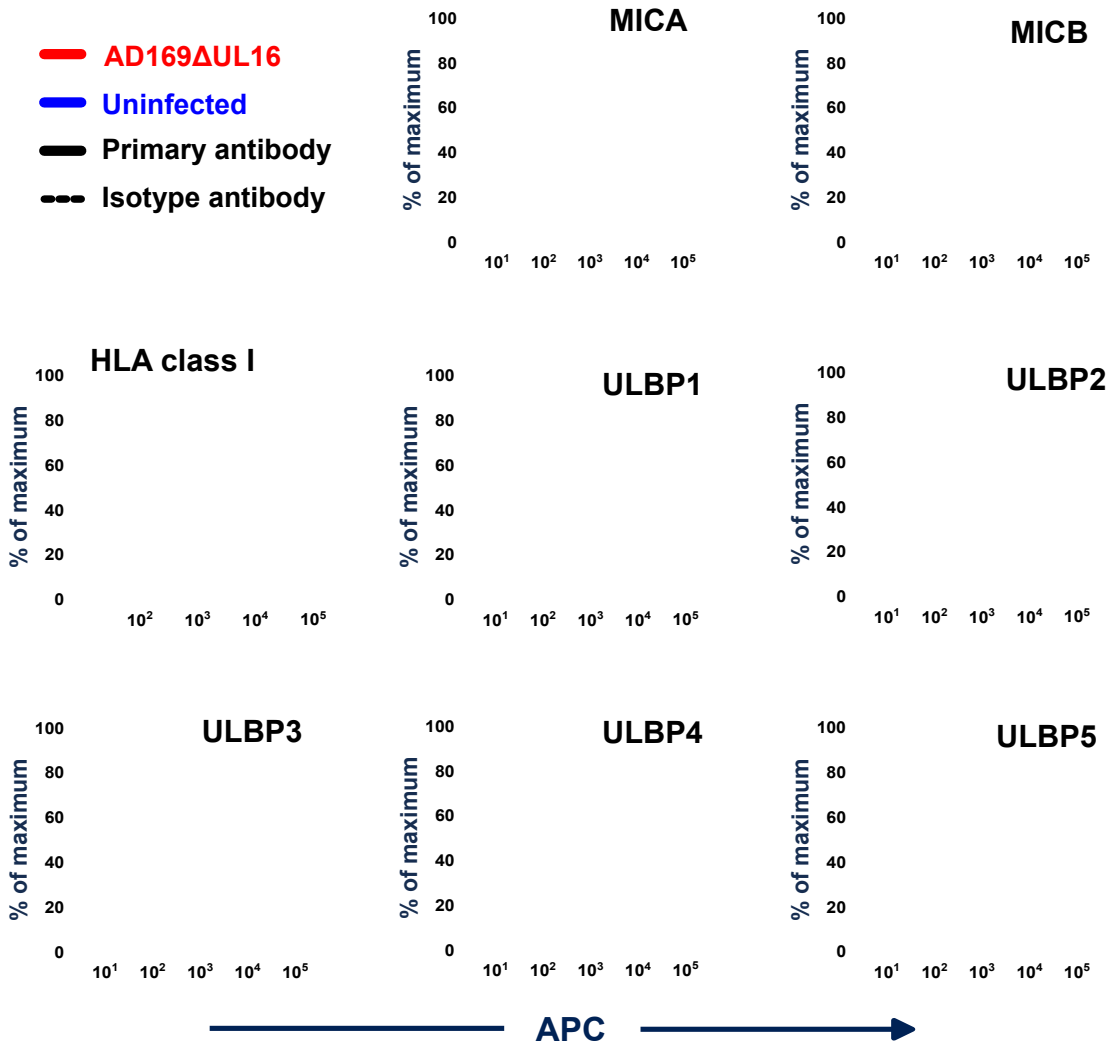

Figure 5.1: Induction of NKG2D ligand expression by AD169ΔUL16. Primary human fibroblasts were cultured to confluence in RPMI1640. Concentrated AD169ΔUL16 viral particles at a multiplicity of infection (MOI) of 0.1 or virus-free RPMI1640 was added to the fibroblasts, and cells were cultured for a further 7 days until ∼90% cytopathic effect was achieved in the CMV-infected flasks. Surface expression of HLA class I and the NKG2D ligands was measured by flow cytometry. HLA class I expression was decreased in CMV infected flasks as expected. Levels of MICA, MICB, ULBP1 and ULBP2 were increased, while ULBP3-5 were unaffected by CMV.

CMV infection increased ULBP2 expression further. CMV did not affect the expression of ULBP3-5.

5.3.1.2 Confirming CMV infection

We have used three indicators to confirm CMV infection of fibroblasts at the population level. HLA class I down-regulation is a recognised CMV immune evasion mechanism (276). By measuring HLA class I expression across a population of fibroblasts, functional CMV infection may be inferred **(Fig 5.1)**. Second, CMV is a cytopathic virus. The AD169 strain was isolated because of these cytopathic properties (271). This cytopathic effect can be observed by light microscopy **(Fig 5.2a)**, and the percentage of cytopathic cells in the population can be estimated. Finally, we have measured the expression of CMV immediate early antigen 1 (*IE1*; *UL123*) mRNA expression by reverse transcription PCR **(Fig 5.2b)**. Because the *UL123* gene does not contain introns, we carried out a DNaseI digestion of extracted RNA to eliminate the *UL123* genomic sequence prior to cDNA synthesis.

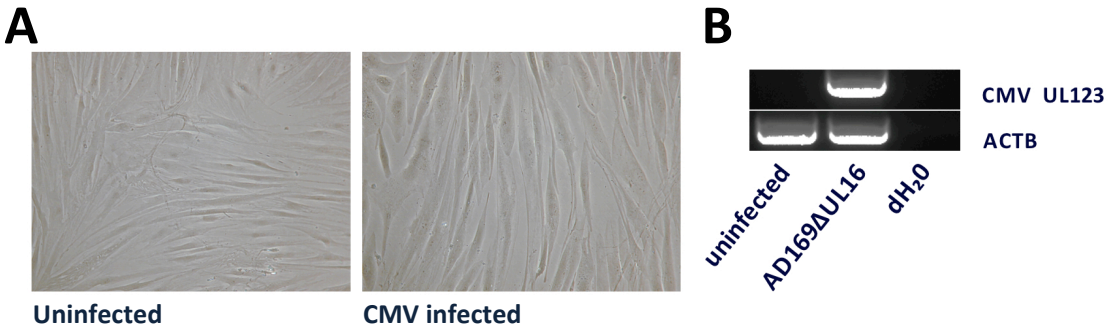

Figure 5.2: Signs of infection with CMV. In addition to measuring cell surface HLA class I expression **(Fig 5.1)**, we confirmed CMV infection in our primary fibroblast populations by estimating the prevalence of cytopathic cells in the population by light microscopy (A), and measuring the expression of CMV genes, such as *UL123*, by RTPCR (B). To avoid detection of genomic UL123, extracted RNA was digested with DNaseI on a column before cDNA synthesis.

5.3.1.3 CMV induces *de novo* transcription of NKG2D ligands

To establish whether CMV infection had a transcriptional effect on NKG2D ligand induction, we carried out a time course to measure unspliced NKG2D ligand transcripts in CMV-infected fibroblasts. Unlike spliced mRNA transcript levels, which are mechanistically susceptible to

alterations in stability and decay rate, as well as alterations in transcription rate, increasing unspliced RNA transcript levels require new transcription instances, and therefore reflects the rate of transcription itself. Measurement of unspliced RNA transcripts in total cell RNA involves two key steps. First, the primer pair must be designed with one primer inside an exon, and the other in an adjacent intron (see ULBP1 primer pair, **Fig 5.3**). Secondly, immediately following RNA extraction, the RNA is digested with DNaseI to remove any contaminating genomic DNA from the preparation, which would contain the same sequence as the unspliced RNA transcript. Any remaining genomic DNA is detected by PCR from a duplicate cDNA synthesis reaction, carried out without adding reverse transcriptase.

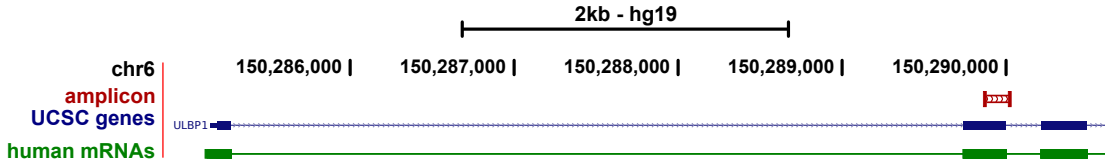

Figure 5.3: Design of *de novo* transcription primers. To semi-quantitatively estimate changes in the rate of transcription of NKG2D ligands, we measured changes in the amount of unspliced transcript for each ligand over time. Primer pairs were designed to flank an exon-intron boundary (*red*). The resulting primer pair would not detect spliced mRNAs, the amount of which depends on decay rate/stability in addition to transcription. Each PCR product was sequenced by Sanger sequencing to determine the specificity of the primers.

The CMV infection time-course demonstrated that CMV increases the rate of transcription of ULBP1, ULBP2 and ULBP3, while ULBP4-6 are unaffected **(Fig 5.4)**. The rate of MICA transcription is also increased, however this assay cannot distinguish between transcription of the functional MICA transcript, and the non-functional non-protein coding alternative MICA transcript **(Fig 5.4)**.

5.3.1.4 Infection with CMV induces Warburg metabolic phenotype

While there is strong evidence for the induction of Warburg metabolism by CMV (155, 156, 270), we wanted to confirm this in our model. Primary human fibroblasts were cultured to confluence in RPMI1640, at which point the culture medium was replaced with DMEM containing 5 mM glucose, 25 mM glucose, 25 mM glucose with azaserine (inhibitor of *de*

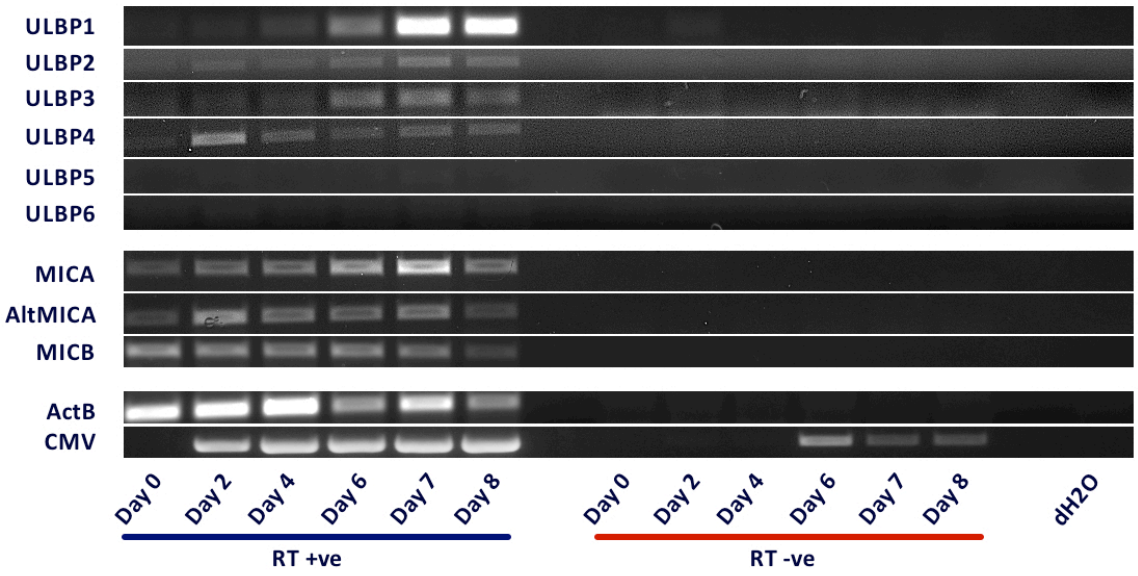

Figure 5.4: CMV induces *de novo* transcription of NKG2D ligand mRNA. CMV-induced change in the rate of new transcript synthesis was estimated by measuring changes in the level of unspliced NKG2D ligand transcripts over a 7 day time course. Primary human fibroblasts were cultured to confluence before infection with CMV AD169ΔUL16 at a MOI of 0.1. Following RNA extraction, each RNA sample was treated with DNaseI to remove any contaminating genomic DNA. Duplicate cDNA synthesis reactions were carried out for each sample, with or without reverse transcriptase. An increase in the amount of unspliced RNA transcripts over time was detected for ULBP1-3, with no change in ULBP4-6. Two transcripts of MICA are known to occur, a functional protein coding transcript (labelled MICA), and a non-functional non-coding transcript (labelled AltMICA). Because the complete sequence of the functional MICA transcript is contained within the unspliced non-coding transcript, changes in the rate of coding MICA transcription can not be directly estimated by this method. However, the increase in transcription observed in the coding MICA transcript, without a corresponding increase in the non-coding transcript suggests that CMV also induces transcription of coding MICA. Equivalent PCR reactions using reverse transcriptase negative cDNA demonstrates the absence of genomic DNA contamination. RT- Reverse transcriptase; ActB - Beta actin.

novo purine synthesis), or 25 mM glucose and azaserine with the purine salvage pathway substrate, hypoxanthine. To measure the amount of glucose utilisation and lactate production attributable to the cells, the same 5 mM or 25 mM glucose containing DMEM was added to cell free culture plates. The cells were infected with AD169ΔUL16 at a MOI of 0.1, and all plates were incubated for 7 days. The culture medium was removed, filtered at 0.2 µm, and stored at −80 °C before being sent to the Department of Biochemistry at the John Radcliffe Hospital, Oxford, UK for measurement of glucose and lactate concentrations.

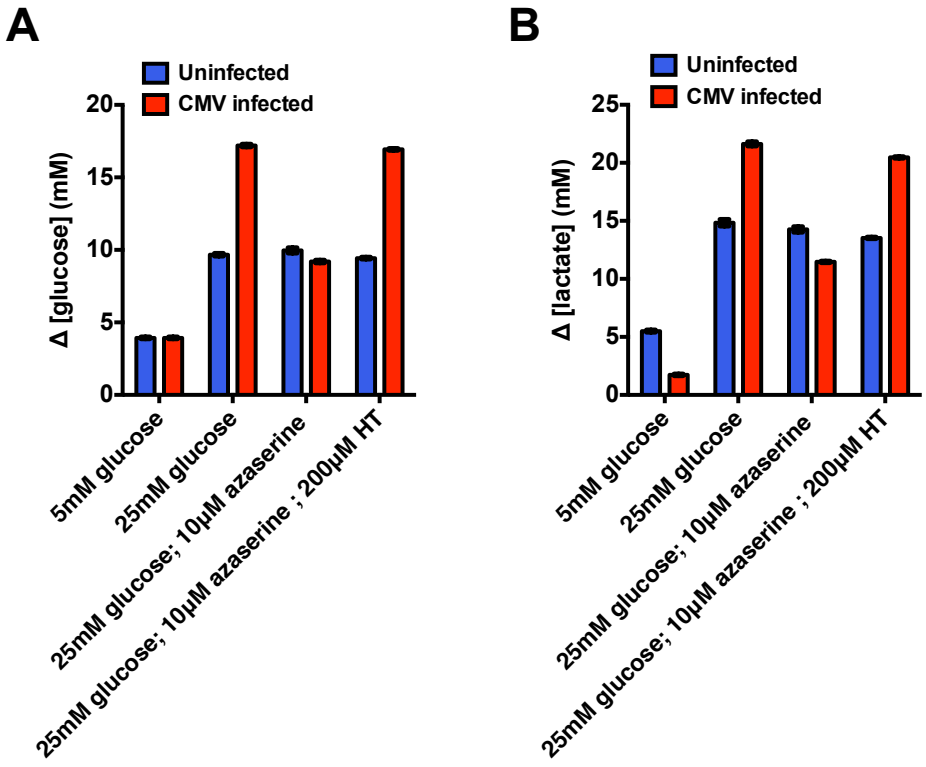

Figure 5.5: Culture media measured glucose and lactate concentrations. We compared the glucose usage and lactate production of confluent uninfected or CMV-infected primary human fibroblasts by measuring glucose and lactate concentrations in culture medium after 7 days of incubation in the stated conditions. CMV-infected cells used more glucose and produced more lactate than uninfected cells. This pattern was blocked by inhibiting *de novo* purine synthesis, and restored by providing a salvage pathway substrate.

We found that CMV-infected fibroblasts used greater amounts of glucose, and produced more lactate than uninfected fibroblasts. This is consistent with 'Warburg metabolism' in the infected cells **(Fig 5.5)**. Further, cells cultured with 25 mM glucose and azaserine showed the same glucose utilisation and lactate production as uninfected cells. The Warburg metabolic phenotype was restored to azaserine-treated cells by the addition of the salvage

pathway substrate, hypoxanthine. This data is consistent with the hypothesis that CMV infection induces Warburg metabolism, and that the purine synthesis pathway is necessary to maintain this process.

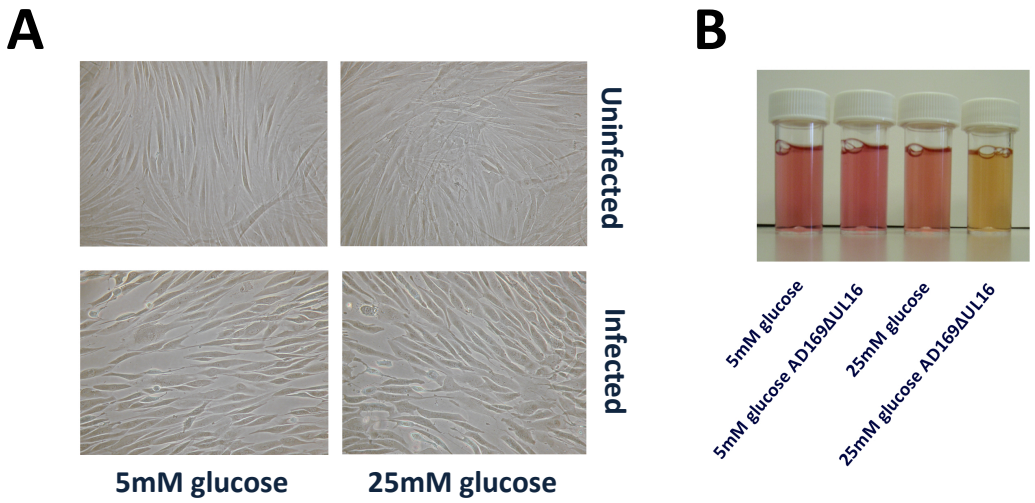

Figure 5.6: Signs of CMV infection in different glucose concentrations. (A) Primary human fibroblasts infected with CMV AD169ΔUL16 demonstrated a cytopathic effect whether cultured in 5 mM or 25 mM glucose, suggesting that successful infection by CMV occurs in both cases. (B) DMEM contains the pH indicator phenol red, which is red at high basic pH and yellow at low acidic pH. The culture medium at day 7 from cells cultured in 25 mM glucose DMEM was more acidic when fibroblasts were infected with CMV, compared to the medium from uninfected fibroblasts. This suggests that more lactate was produced by infected cells over the incubation period.

5.3.2 The effect of glucose on NKG2D ligand induction in the CMV-primary fibroblast model

Having established that CMV infection induces cell surface expression of MICA, MICB, ULBP1 and ULBP2 proteins on primary fibroblasts, we wanted to see if the amount of glucose available to these cells would influence the extent of NKG2D ligand induction. Primary fibroblasts were cultured to confluence in T175 flasks, and the culture medium was replaced with DMEM containing 5mM or 25 mM glucose. Half of the cells were infected with CMV AD169ΔUL16 at a MOI of 0.1, and all flasks were then cultured for 7 days. Successful infection was confirmed by observing a cytopathic effect by light microscopy **(Fig 5.6a)**, HLA class I down-regulation by flow cytometry **(Fig 5.7a)**, and CMV *UL123* gene expression by

RTPCR **(Fig 5.7b)**. Infected cells cultured in DMEM containing 25 mM glucose produced more lactate than uninfected cells, illustrated by pH changes in the culture medium, and consistent with Warburg metabolism **(Fig 5.6b)**. Measurement of NKG2D ligand expression by flow cytometry demonstrated higher cell surface expression of MICA, MICB, ULBP1 and ULBP2 in cells cultured in DMEM with 25 mM glucose compared to cells cultured in DMEM with 5 mM glucose **(Fig 5.7c)**. This observation confirms the findings in our HEK293T cell model, that glucose availability is important for NKG2D ligand induction.

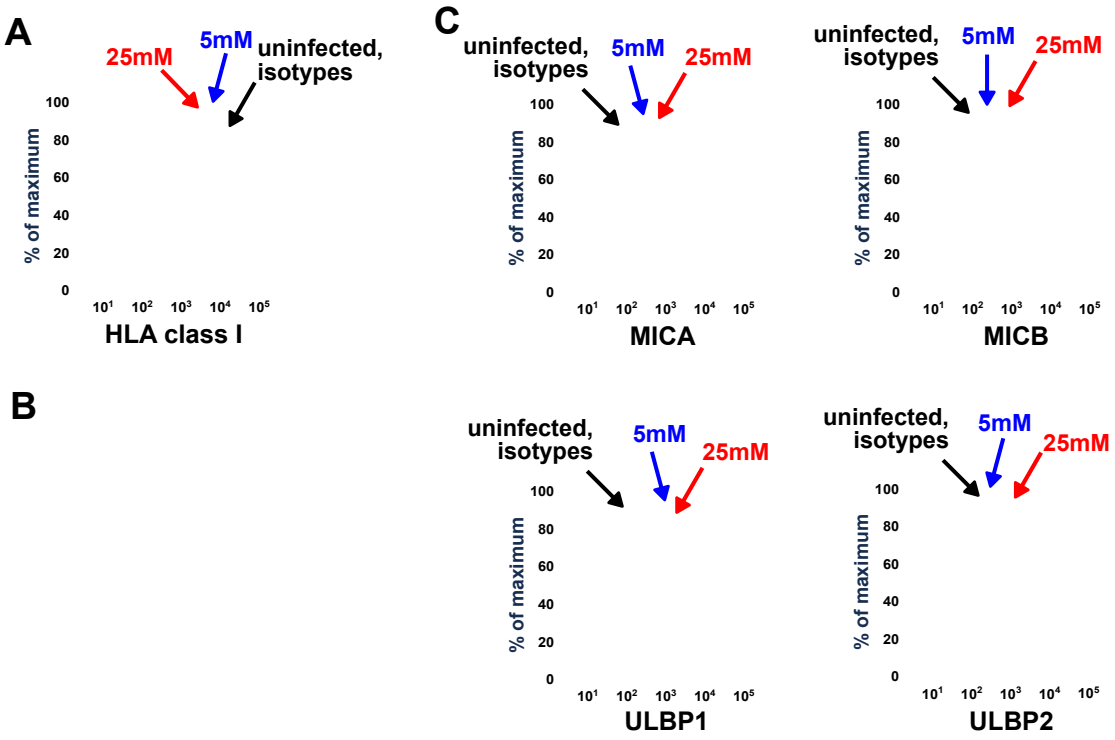

Figure 5.7: Induction of NKG2D ligands by CMV depends on glucose availability. Primary human fibroblasts were cultured to confluence, infected with CMV AD169ΔUL16 at a MOI of 0.1, and cultured for 7 days in DMEM containing 5 mM or 25 mM glucose. (A) CMV caused down-regulation of HLA class I in fibroblasts cultured in 5 mM glucose and 25 mM glucose. (B) The expression of CMV gene *UL123* was evident in infected fibroblasts grown in 5 mM or 25 mM glucose. (C) Cells cultured in the presence of 25 mM glucose demonstrated higher levels of MICA, MICB, ULBP1 and ULBP2 expression compared to cells grown in medium containing 5 mM glucose.

5.3.3 Purine synthesis is important for the impact of glucose on NKG2D ligand expression

In our HEK293T cell model, we established that purine synthesis was necessary for glucose to influence NKG2D ligand expression, by inhibiting *de novo* purine synthesis with azaserine, and restoring purine synthesis via the salvage pathway with hypoxanthine **(Fig 4.6)**. To see if purine synthesis plays a similar role in glucose-mediated NKG2D ligand expression in primary human fibroblasts infected with CMV, we infected confluent fibroblasts with CMV AD169ΔUL16 at a MOI of 0.1 for 7 days, in four different culture conditions: DMEM with 5 mM glucose, 25 mM glucose, 25 mM glucose with azaserine, and 25 mM glucose with azaserine and hypoxanthine. The culture medium of cells grown in DMEM with 25 mM glucose only, or 25 mM glucose, azaserine and hypoxanthine (the purine-competent samples) was more acidic at 7 days, suggesting greater lactate production, in keeping with previous observations **(Fig 5.8a)**. Light microscopy revealed a cytopathic effect in all culture conditions **(Fig 5.8b)**. Cell surface examination by flow cytometry confirmed down-regulation of HLA class I expression in all four culture conditions, consistent with active CMV infection **(Fig 5.9)**. Infected fibroblasts cultured in 5 mM glucose, or 25 mM glucose and azaserine displayed limited increases in cell surface expression of MICA, MICB, ULBP1 and ULBP2, while purine-competent fibrobasts, cultured in 25 mM glucose, or 25mM glucose, azaserine and hypoxanthine had higher levels of cell surface expression of these ligands **(Fig 5.9)**. These results suggest that purine synthesis plays an important role in glucose-mediated NKG2D ligand expression in the physiological context.

5.3.4 NKG2D ligand expression in uninfected cells is not glucose-dependent

Increased glucose transport capacity through up-regulation of GLUT transporters is a documented consequence of CMV infection in primary fibroblasts (270). Cellular uptake of purine nucleosides is also transporter dependent (reviewed by Young *et al* (246)). To see if glucose

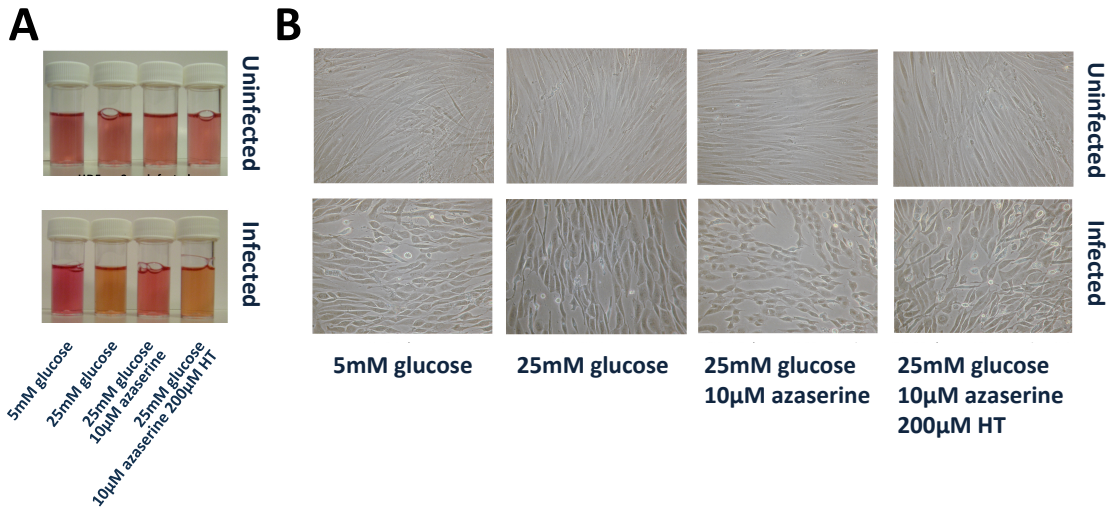

Figure 5.8: CMV infection of fibroblasts with inhibition of purine synthesis. (A) The culture medium from CMV-infected cells cultured in purine-competent conditions (25 mM glucose, or 25 mM glucose, azaserine and hypoxanthine) was more acidic than culture medium from uninfected cells, cells cultured in low glucose medium, or cells cultured with 25 mM glucose and azaserine, indicating greater lactate production in purine-competent samples. (B) Light microscopy demonstrates a cytopathic effect in all culture conditions, demonstrating the competence of CMV infection, regardless of purine synthesis capacity.

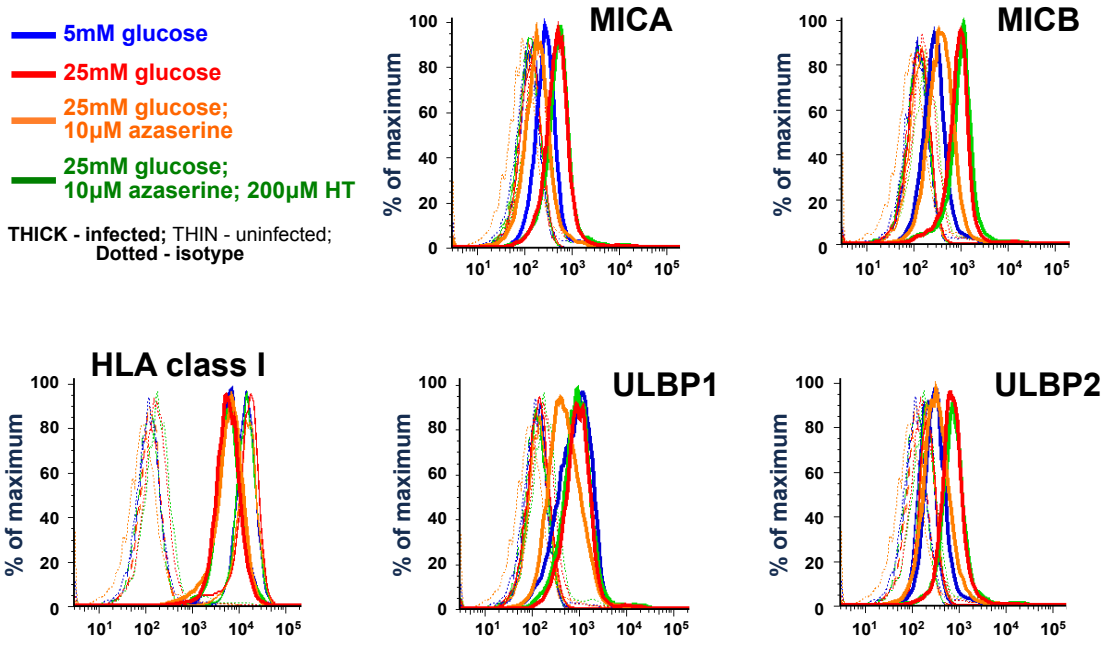

Figure 5.9: CMV-induced NKG2D ligand expression depends on purine synthesis. Analysis of cell surface protein expression by flow cytometry demonstrated down-regulation of cell surface HLA class I in CMV infected cells, regardless of purine synthesis capacity. The induction of cell surface MICA, MICB, ULBP1 and ULBP2 by CMV infection was limited by glucose availability and azaserine-mediated inhibition of purine synthesis. Higher levels of cell surface expression of these NKG2D ligands was observed in purine-competent culture conditions.

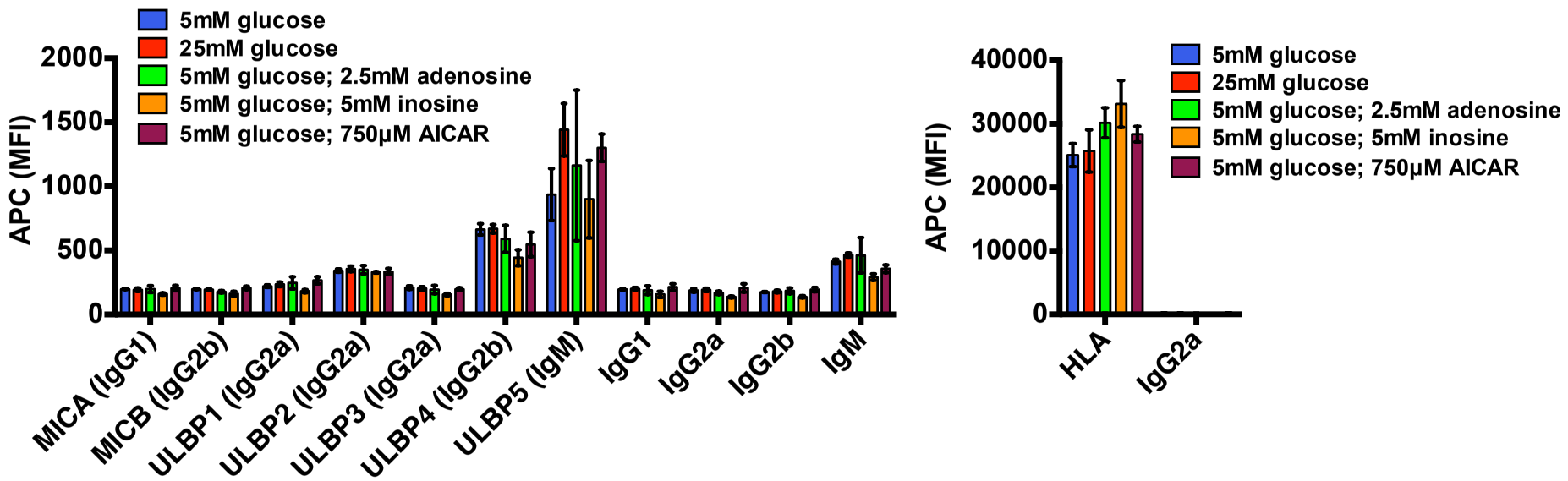

Figure 5.10: Purines do not induce NKG2D ligand expression on quiescent fibroblasts. While glucose and purine synthesis are *necessary* for NKG2D ligand cell surface expression in the physiological setting, we wanted to see if these stimuli were *sufficient* to induce NKG2D ligand expression. Primary human fibroblasts were cultured to confluence, and then cultured in DMEM with the treatments as indicated for 72 hours. Examination of cell surface protein expression by flow cytometry at 72 hours revealed that neither glucose nor purine nucleosides were sufficient for cell surface NKG2D ligand induction. Histograms represent the mean +/– 95% confidence interval of mean fluorescence intensities from three biological replicate measurements.

or purine nucleosides could induce cell surface NKG2D ligand expression on uninfected primary human fibroblasts, we cultured fibroblasts to confluence, and replaced the medium with DMEM containing 5 mM or 25 mM glucose, or 5 mM glucose and either 2.5 mM adenosine, 5 mM inosine or 750 µM AICAR. The fibroblasts were cultured for a further 72 hours, and cell surface expression of NKG2D ligands or HLA class I was measured by flow cytometry. No effect on NKG2D ligand expression was observed **(Fig 5.10)**.

5.4 Discussion

In **Chapters 3 & 4**, we provided strong evidence to support the hypothesis that induction of the Warburg effect (i.e. the increased uptake and metabolism of glucose) played a causative role in NKG2D ligand induction in our HEK293T cell model. Further, we demonstrated that the transport of glucose and its metabolism within the cell was responsible for this effect, and that the effect of glucose was dependent on purine nucleotide synthesis. In this chapter we sought to investigate whether these metabolic effects, central to NKG2D ligand expression in HEK293T cells, were important in a physiological model of NKG2D ligand induction.

5.4.1 CMV infection of primary fibroblasts as a physiological model

To test our hypothesis that the induction of Warburg metabolism was important of NKG2D ligand induction in a physiological setting, we used a CMV-primary fibroblast infection model. We confirmed that CMV AD169ΔUL16 induces cell surface expression of MICA, MICB, ULBP1 and ULBP2 in confluent primary human fibroblasts **(Fig 5.1)**, and that this expression is associated with the induction of Warburg metabolism **(Fig 5.5)**. The observations that glucose availability **(Fig 5.7)** and purine synthesis **(Fig 5.9)** are important for NKG2D ligand induction in this setting are novel findings, and directly support our hypothesis that these events are important in NKG2D ligand regulation, in a range of settings. While previous studies have reported increases in steady-state spliced mRNA transcript levels, we show

that CMV infection induces *de novo* synthesis of ULBP1, 2 and 3 RNA transcripts. This suggests that the regulation of NKG2D ligand expression is either at transcriptional or pre-transcriptional stages. This information is useful, as a physiological, inducible transcriptional model of NKG2D ligand induction will be helpful for identifying inducible regulatory DNA elements using methods such as DNaseI-seq. It is also interesting that CMV AD169ΔUL16 infection induces *de novo* transcription of ULBP3, while no cell surface expression of ULBP3 was observed. This raises the possibility that in addition to known CMV NKG2D-evasion mechanisms, further mechanisms may exist to prevent the translation or cell surface expression of ULBP3 in CMV-infected cells.

5.4.2 Confirmation of CMV infection

The documentation of CMV infection is important to understand whether our interventions act to influence the cellular response to infection, or the infection itself (discussed below). The use of three independent markers of CMV infection, namely RTPCR measurement of CMV gene expression (*UL123*; intermediate early antigen 1 (*IE1*)), and two functional outcomes of CMV infection: morphological changes by light microscopy representative of the CMV-induced cytopathic effect, and CMV-induced HLA class I down-regulation (reviewed by Hengel *et al* (277)), strengthens our ability to make judgements about the success of infection under different treatment conditions.

It is difficult to prove that the CMV life cycle is entirely unaffected by our interventions in these experiments. Two observations suggest that our intervention is affecting the cellular response (NKG2D ligand induction), rather than CMV infection itself. First, in all experiments, there is clear evidence that functional CMV infection occurred in infected cells, according to the criteria outlined above. Second, Venkataraman *et al* (136), examining MICA and MICB promotor function demonstrated that transfection of primary fibroblasts with CMV intermediate early genes 1 and 2 alone was sufficient to induce transcriptional activity from these promoters. We have provided evidence that in all our test conditions,

including low (5 mM) glucose, and with azaserine, CMV infection has proceeded to this stage. The demonstration by RTPCR that CMV *IE1* (*UL123*) gene expression is taking place to an equal extent with each treatment suggests that the effect of these interventions on cell surface NKG2D ligand induction is at a later point in this process. This conclusion would also be consistent with our HEK293T cell data.

5.4.3 Glucose and purine synthesis are important for NKG2D induction in a physiological setting

A key conclusion from the experiments presented in this chapter is that glucose and purine synthesis are important for NKG2D ligand induction in the physiological setting of CMV-primary fibroblast infection. Culturing primary human fibroblasts at in culture medium with 5 mM glucose or 25 mM glucose with azaserine limited or prevented the induction of all expressed cell surface NKG2D ligands in CMV AD169ΔUL16 infected cells **(Figs 5.7 & 5.9)**. The low and high glucose concentrations, as well as the azaserine and hypoxanthine concentrations were chosen on the basis of our previous experiments in HEK293T cells from earlier chapters. These concentrations may be relatively generous, considering that the number of HEK293T cells is ∼6-8 fold greater than the number of primary human fibroblasts per unit surface area in confluent flasks (e.g. a confluent T75 flask of primary human fibroblasts typically contains $\sim 1 \times 10^6$ cells, while for HEK293T cells this figure is closer to $\sim 6–8 \times 10^6$ cells). This relative excess of glucose compared to the HEK293T cell model may have been a limiting factor on the magnitude of the effects observed.

5.4.4 Uninfected fibroblasts, glucose and purine nucleosides

We were somewhat surprised to note that neither glucose, nor purine nucleosides (including AICAR), had any effect on NKG2D ligand expression in quiescent primary fibroblasts, despite the strong induction by these stimuli in HEK293T cells. There are many possible explanations for this. It is certainly the case that in our model, CMV infection is necessary to permit

the influence of glucose. One possible explanation is that CMV affects glucose/nucleoside transport capacity. We noted in earlier experiments **(Fig 3.20)** that a limited capacity for fructose transport limited the impact of fructose on MICA expression in HEK293T cells, and that transfection of HEK293T cells with the fructose transporter, GLUT5, allowed fructose to induce MICA expression as effectively as glucose. It is also clear that upon CMV infection of primary fibroblasts, cell surface expression of the glucose transporter, GLUT4, is increased (157). It is possible that inducing transporter expression in primary fibroblasts would restore the ability of glucose and nucleosides to influence NKG2D ligand expression. We plan to test this hypothesis in future experiments.

5.4.5 Summary

In this chapter we have demonstrated that glucose availability is an important determinant of NKG2D ligand expression in an independent physiological model of NKG2D ligand induction. Further, purine synthesis is key to this process. The observations that glucose and purine synthesis are important in two typical settings where NKG2D ligand expression is observed (cellular proliferation - **Chapters 3 & 4**, and viral infection - **Chapter 5**) adds further weight to this argument. The results presented in this chapter support our underlying hypothesis, that the induction of Warburg metabolism is a key deterministic event in controlling NKG2D ligand induction.

6

DNaseI hypersensitivity and NKG2D ligand expression.

Contents

6.1 Abstract

Earlier experiments have demonstrated that the induction of cell surface NKG2D ligand expression in quiescent primary human fibroblasts by cytomegalovirus is dependent on changes in new gene transcription of these ligands. To identify the regulatory DNA elements that control this transcriptional NKG2D ligand response, we have carried out DNaseI-seq analysis to identify the open chromatin sites that are located at the NKG2D ligand gene loci in the

breast cancer cell line, MCF7. In the course of these experiments, we adapt and optimise a DNaseI-seq protocol for use in human cells and develop a data analysis 'pipeline' to aid in open chromatin site identification. In total, we have identified 10 open chromatin sites at the MICA/MICB locus within the MHC, and 17 open chromatin sites at the ULBP locus on chromosome 6. We anticipate that the biological and computational methods developed here will be directly useful in future experiments.

6.2 Introduction

6.2.1 Summary of previous work

In attempting to identify elements which control the induction of cell surface NKG2D ligand expression, we have demonstrated that glucose and purine synthesis are important components of the induction process, both in a HEK293T cell line model, and in a physiological CMV-primary fibroblast model. In examining the CMV-primary fibroblast model, we have found that the induction of NKG2D ligand expression by CMV involves an increase in *de novo* NKG2D ligand transcription **(Fig 5.4)**. Given that these transcriptional changes occur, we wanted to identify areas of open chromatin at the NKG2D ligand loci that may be important in controlling these transcriptional effects.

6.2.2 The identification of regulatory genomic DNA elements

In vivo, genomic DNA exists in the form of chromatin, a term referring to DNA wrapped around histone proteins. The basic structural unit of chromatin is the nucleosome, which consists of ~200 base pairs coiled around a single histone protein octamer core (reviewed by Bell *et al* (278)). This tight association between DNA and histone proteins limits the ability of the enzyme DNaseI to access genomic DNA.

The exploration of these structural features of chromatin began in 1976, when Weintraub and Groudine noticed that genes that were being actively transcribed were more sensitive

to digestion with DNaseI than untranscribed genes (279). In 1980, working with drosophila genes, Wu noticed that discrete sites of hypersensitivity to DNaseI digestion existed upstream of the transcription start site (280). This was the first DNaseI hypersensitivity assay.

Genomic open chromatin sites develop where histone displacement occurs in the presence of DNA-binding factors (281). This gives rise to three chromatin sensitivity states **(Fig 6.1)**. Nucleosomal DNA wrapped tightly around histone proteins is resistant to DNaseI activity, while increased accessibility to DNaseI resulting from nucleosome displacement confers hypersensitivity ((282); also reviewed by Weake and Workman (283)). Within hypersensitive sites, proteins (such as transcription factors) binding non-covalently to DNA can provide protection from DNaseI digestion, relative to the accessibility at the surrounding hypersensitive sites. This relative protection forms the basis of the related but distinct method of DNaseI footprinting, which aims to identify the specific base sequences to which a transcription factor is bound (284).

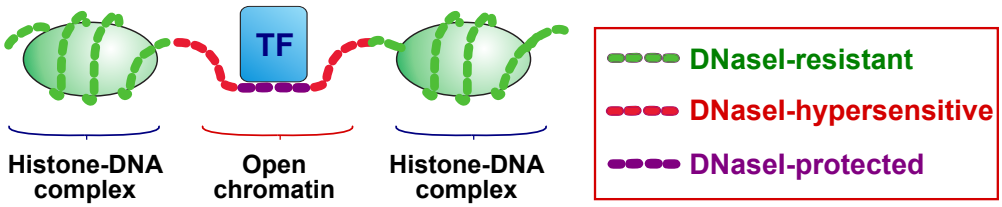

Figure 6.1: Chromatin and DNaseI sensitivity. Variability in sensitivity of genomic DNA to digestion by the enzyme DNaseI forms the basis of the DNaseI-seq assay. Nucleosomal DNA (*green*) is tightly wrapped around histone protein complexes. DNaseI has poor access to this DNA which is therefore DNaseI resistant. Displacement of histones at sites of DNA-trans-acting factor interaction leads to enhanced DNaseI enzyme accessibility, which confers DNaseI hypersensitivity to these sites (*red*). Within these sites, non-covalent interaction between DNA and protein (e.g. a transcription factor) may relatively protect DNA from digestion at the point of binding. This forms the basis of the DNaseI footprinting assay. TF - Transcription factor.

Methods for high throughput identification of DNaseI-hypersensitive sites have been in development over the past 10 years. This began with Sanger sequencing of cloned DNaseI-hypersensitive DNA fragments (285, 286), followed by a microarray-based approach (287, 288), and ultimately, ligation-mediated high-throughput sequencing of these fragments (289, 290). One of the most widely used methods involves a "two-hit" approach **(Fig 6.2)**. Here, isolated cellular nuclei are exposed to DNaseI, total DNA is isolated (undigested genomic

DNA strands and DNaseI released fragments), and adapters are ligated. The short DNaseI-released fragments are isolated by a combination of agarose gel extraction-size selection and short course PCR amplification. These tagged DNA fragments are sequenced in parallel, generating tens of millions of short DNA sequences, each originating in a DNaseI-hypersensitive site. DNA sequences can be mapped back to the reference human genome to localise the DNaseI signal.

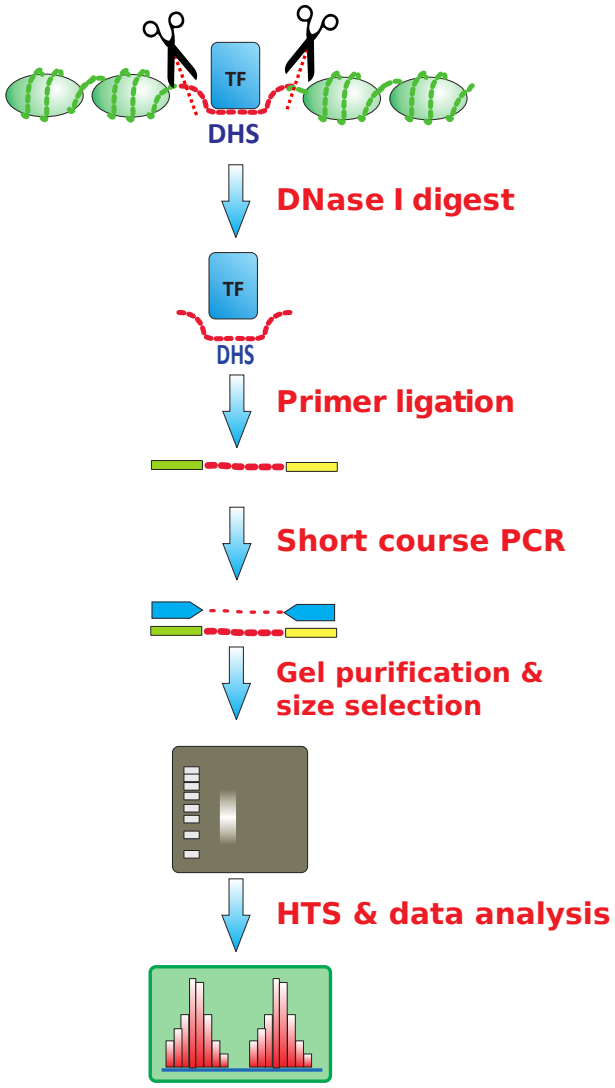

Figure 6.2: The "two-hit" approach to DNaseI-seq. Permeabilized cellular nuclei are isolated and incubated with the enzyme, DNaseI. Where DNaseI cuts the genomic DNA twice in a single open chromatin site, small DNA fragments are released. Primer adapters are ligated to any cut DNA ends, and short fragments are size selected by gel extraction and short course PCR. The resulting 'DNaseI library' is sequenced using high throughput methods, and mapped back to the reference human genome sequence. DHS - DNaseI hypersensitive site; TF - transcription factor; HTS - High throughput sequencing.

6.2.3 Aims and objectives

In this chapter, we aim to adapt, optimise and validate a DNaseI-seq protocol originally designed for use with mouse cell lines (291). We also aim to develop a qPCR assay to quantitatively estimate digestion at DNaseI hypersensitive sites in DNaseI-treated DNA samples. Finally, we plan to set up a data analysis 'pipeline' for the processing of sequencing data, to enable the identification of discrete open chromatin sites at the NKG2D ligand gene loci.

6.3 Results

6.3.1 Optimisation of DNaseI-seq method in the MCF7 cell line

The DNaseI-seq protocol was originally designed by Dr. Marco De Gobbi (MRC Molecular Haematology Unit) (291), and modifications to the protocol were made following helpful discussions with Dr. Jim Hughes (MRC Molecular Haematology Unit).

6.3.1.1 Assessment of nuclear isolation by confocal microscopy

In order for the DNaseI enzyme to gain access to the cell nucleus, it is necessary to disrupt the cell membrane. This is achieved by resuspending cells in an NP40-based lysis buffer until cell membrane disruption is achieved. In the original protocol, the quality and extent of cell membrane disruption is assessed by light microscopy. Because we were adapting a mouse cell protocol for use with a human cell line, we wanted to identify a more robust means of making this assessment.

MCF7 cells were cultured in T175 flasks with DMEM containing 25 mM glucose until a total of 7×10^7 cells were obtained. After harvesting and washing the cells, they were incubated with Hoechst 33258 (light blue), a cell membrane permeable DNA-binding dye, for 10 minutes. The cells were then divided into seven 1×10^7 aliquots, and resuspended in lysis buffer for the amount of time specified. The cells were washed twice in PBS to remove

Figure 6.3: Confocal microscopy assessment of nuclear isolation. To assess the efficiency of cell membrane disruption with the lysis buffer, and identify the optimal duration of digestion, we pre-stained MCF7 cell nuclei with the blue DNA binding dye, Hoechst33258, and resuspended aliquots of the stained cells in lysis buffer for various time periods. The lysis buffer was removed by washing cells with PBS. Prior to microscopy, the samples were incubated with the lipid-bilayer dye, CellMask Orange, and the cell membrane impermeable DNA binding dye, Sytox green. Unlysed cells did not take up Sytox green. ~80% of cells were Sytox green positive at 30 seconds, and this proportion increased to >95% by 2.5 minutes of lysis.

residual lysis buffer and transferred on ice to the cellular imaging core. Immediately prior to examination by confocal microscopy, Cellmask Orange, a lipid bilayer binding dye, and Sytox green, a cell membrane impermeable DNA-binding dye, were added to the samples. The samples were mixed by flicking and inversion, and incubated on ice for 5 minutes. Confocal microscopy showed no uptake of Sytox green in unlysed cells. After just 30 seconds of incubation with the lysis buffer, ∼80% of cells were positive for Sytox green. At 2.5 minutes of lysis, >95% of cells were Sytox green positive, and CellMask Orange was staining internal cell membranes **(Fig 6.3)**. Based on these observations, we chose a 2.5 minute lysis time for further experiments.

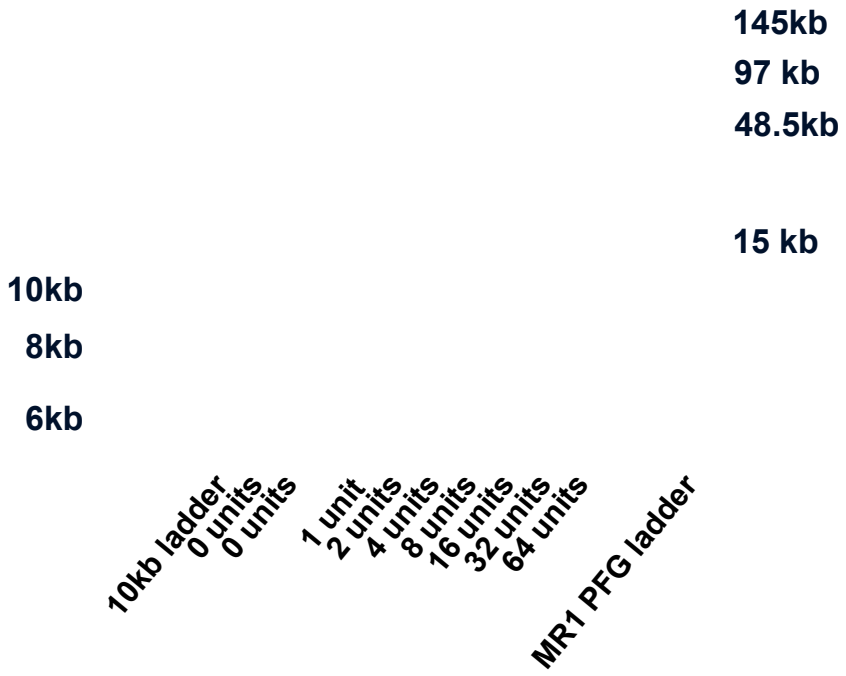

Figure 6.4: Assessment of DNaseI digestion by pulsed-field gel electrophoresis. Pulsed-field gel electrophoresis was carried out to assess the DNaseI-mediated digestion of genomic DNA following DNA extraction and precipitation. The gel was run overnight, and stained with ethidium bromide to visualise the DNA under UV light. This confirmed progressive digestion of genomic DNA with increasing doses of DNaseI.

6.3.1.2 Assessment of DNaseI digestion by pulsed-field gel electrophoresis

A DNaseI-seq library was prepared by culturing 7×10^7 MCF7 cells in T175 flasks with DMEM containing 25 mM glucose. The cells were harvested and lysed for 2.5 minutes, as described above. Following cell membrane disruption, the isolated nuclei were aliquoted, and increasing concentrations of DNaseI were mixed with the cell aliquots to release DNA fragments from sites of open chromatin. Once digested, the DNA was released from the nuclei by resuspension of the sample in a nuclear lysis buffer, followed by phenol-chloroform extraction and ethanol precipitation. To quantify the amount of digestion achieved in the DNaseI-digested samples, we initially carried out pulsed-field gel electrophoresis **(Fig 6.4)**. This confirmed DNaseI concentration-dependent digestion across each aliquot.

6.3.1.3 Assessment of DNaseI digestion by real-time PCR

We wanted to develop a qPCR-based assay to quantify the amount of DNaseI digestion more precisely, so as to improve the likelihood of being able to match samples with the same level of digestion in future experiments. For DNaseI-insensitive control sites, we chose two loci, one in the rhodopsin (*RHO*) gene and one in the haemoglobin beta (*HBB*) gene. The rhodopsin site has been demonstrated to be DNaseI insensitive by Dorschner *et al* (214), and we used their primers for this site **(Fig 6.5a)**. For the second insensitive site in the *HBB* gene, we looked at available ENCODE DNaseI hypersensitivity data on the UCSC browser and made primers for an amplicon away from known hypersensitive sites **(Fig 6.5b)**. Finally, for our control DNaseI *hypersensitive* site, we made primers for a site upstream of the *MYC* gene that is consistently hypersensitive in all available ENCODE datasets on the UCSC browser **(Fig 6.5c)**.

Having identified and made primers for each of these sites, we measured the extent of digestion across our DNaseI-digested MCF7 samples by qPCR using all three primer pairs. In keeping with the pulsed-field gel analysis, we found DNaseI dose-dependent digestion at the *MYC* site relative to either *RHO* or *HBB* sites **(Fig 6.6)**. Based on this, we pooled

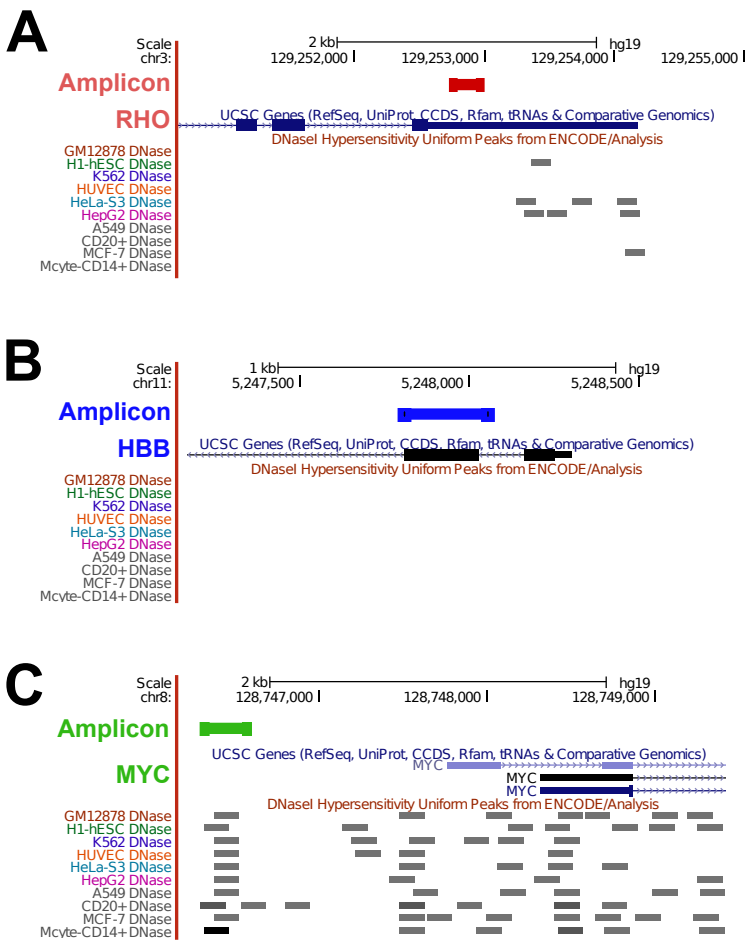

Figure 6.5: Assessment of DNaseI digestion: amplicons for qPCR sites. To semi-quantify DNaseI digestion between DNA samples exposed to different amounts of DNaseI, we performed real time qPCR. We measured the C_t of target in a hypersensitive site 5' of the c-Myc gene, compared to the C_t at two DNaseI insensitive sites at the rhodopsin and hemoglobin beta genes. Primers for the rhodopsin insensitive site (A) were obtained from previous publications (213, 214). The haemoglobin beta site (B), and c-Myc site (C) were chosen on the basis of ENCODE DNaseI-seq data across multiple primary cells and cell lines suggesting consistent DNaseI insensitivity and hypersensitivity respectively. Each of the figures shows the results of an *in silico* PCR reaction, using the respective amplicon primer sequences, on the UCSC genome browser (292). ENCODE hypersensitive sites are indicated as grey bars below the UCSC gene tracks (293). RHO – rhodopsin; HBB – hemoglobin beta; MYC – c-Myc.

DNA from the 8-, 16-, and 32-DNaseI unit samples to generate the DNaseI-seq library for sequencing. The sequencing was carried out on an Illumina GAII high throughput sequencing machine, on a single end, 50 base pair sequencing run.

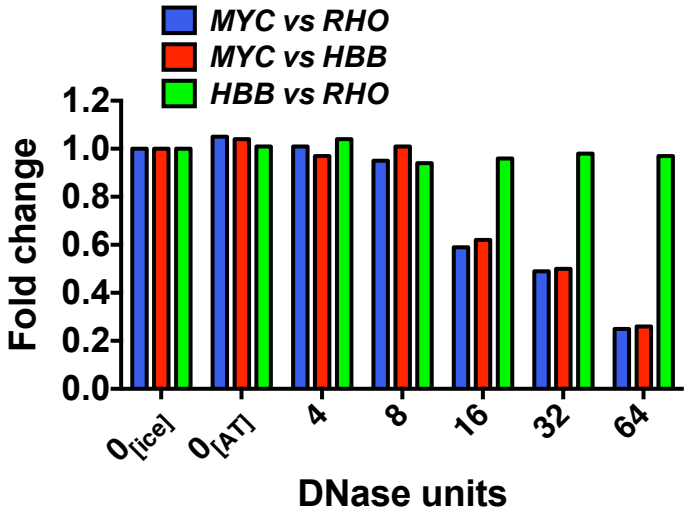

Figure 6.6: Assessment of DNaseI digestion by real time qPCR. To obtain a quantitative estimate of the extent of hypersensitive site digestion in DNaseI treated samples, we carried out qPCR at one DNaseI hypersensitive site *MYC* and two DNaseI insensitive sites (rhodopsin, *RHO* and haemoglobin beta, *HBB*) We found progressive digestion at the *MYC* site relative to the *HBB* and *RHO* sites, with a ∼4 fold decrease in signal strength at the hypersensitive site in maximally digested samples.

6.3.2 DNaseI-seq data analysis

The strategy for DNaseI-seq data analysis was chosen following helpful discussions with Dr. Linda Hughes (Bioinformatics Department, WTCHG) and Dr. Jim Hughes (MRC Molecular Haematology Unit).

6.3.2.1 Mapping MCF7 DNaseI-seq data

In total, 38,478,799 reads were obtained from the sequencing run. The raw sequence file data from the MCF7 DNaseI-seq experiment was mapped to the reference human genome sequence (GRCh37/hg19) using the hash-based short read mapping program, Stampy (294). Stampy identifies the genomic locus corresponding to the best possible sequence match for each read. Using sequencing quality scores, and information about base mismatches at the mapped locus, Stampy generates its own overall quality score (MAPQ score) for each mapped

read. This allows the removal of poor quality mapped reads. When a read maps equally well to two distinct genomic sites, the read is given a default quality score of zero (poor quality).

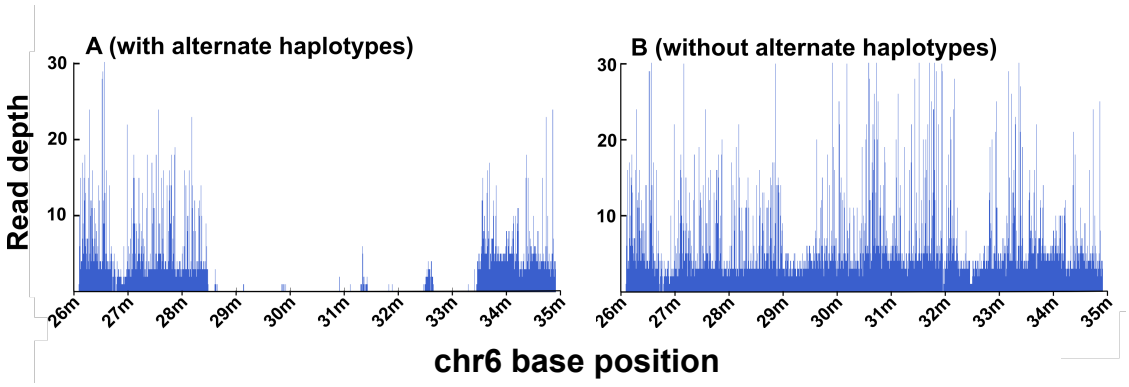

Figure 6.7: Removing alternate haplotypes from the reference genome improves mapping to alternate haplotype regions. (A) When we initially mapped reads to the full human genome reference sequence, and quality filtered the resulting mapped reads, we found little or no mapping across large portions of the MHC region of the genome on chromosome 6. Inspection of the unfiltered data files showed that reads were mapped normally across the MHC, but were assigned a mapping quality (MAPQ) score of zero by the mapping program. (B) When we removed alternate haplotype chromosomes from the reference genome sequence and re-mapped the data, reads mapped to the MHC region with good quality scores, despite the polymorphism in the region.

We initially mapped the DNaseI-seq data to the standard complete version of the hg19 reference sequence, downloaded from UCSC, using Stampy, with the commands:

./stampy.py -G hg19 /chromsomeDirectory/*.fa.gz

./stampy.py -g hg19 -h hg19

to build the reference genome. To map data with Stampy, we used the command:

./stampy.py -g hg19 -h hg19 -M reads.fastq >mappedReads.sam

We filtered out any mapped reads with a quality score (MAPQ score) of lower than 15, with advice from the authors of Stampy. When we did this, we found that most of MHC region, which contains two genes of direct interest, *MICA* and *MICB*, had no mapped reads **(Fig 6.7a)**. Closer inspection of the full, unfiltered dataset showed good read mapping across the region, but the majority of reads had a quality score of zero. The standard refer-

Table 6.1: Alternate haplotype chromosomes in hg19

Alternate haplotype	Chromosome	hg19 coordinates
chr6_apd_hap1	chr6	28696604-33335493
chr6_cox_hap2	chr6	28477797-33351542
chr6_dbb_hap3	chr6	28696604-33329076
chr6_mann_hap4	chr6	28696604-33225977
chr6_mcf_hap5	chr6	28696604-33359642
chr6_qbl_hap6	chr6	28696604-33379750
chr6_ssto_hap7	chr6	28659143-33448354
chr4_ctg9_hap1	chr4	69170077-69878175
chr17_ctg5_hap1	chr17	43384864-44913631

ence genome contains nine alternate haplotype chromosomes, seven of which cover the MHC region **(Table 6.1)**. These haplotypes represent sequences of duplicate regions from different individuals of a genomic locus already present in the corresponding standard reference sequence chromosome, but are presented along with the reference sequence as distinct chromosomes. We hypothesised that most reads mapping to these loci were assigned MAPQ scores of zero by Stampy, because they mapped equally well to two or more alternate haplotype chromosomes, despite polymorphism in the region. To test this, we built a new reference genome sequence containing all hg19 chromosomes, except the alternate haplotype chromosomes, and remapped our MCF7 data. With this strategy, reads mapping to the MHC locus were assigned high quality MAPQ scores by Stampy, despite the existing polymorphism **(Fig 6.7b)**. A total of 27,161,980 (70.6%) reads mapped successfully and passed the quality filters.

6.3.2.2 Open chromatin sites at NKG2D ligand loci

Having mapped the sequence data, we next wanted to identify clusters of reads in the mapped data that might represent areas of open chromatin. Following helpful discussions about available analytical tools with Dr. Jim Hughes (MRC Molecular Haematology Unit), we decided to write our own analytical software, a program called PeakHunter (see **Chapter 7**). PeakHunter identifies areas of the genome that contain more mapped reads above the background level than expected by chance. To identify the threshold read density for peak calling, PeakHunter uses a kernel density estimator, to generate a probability distribution

for read density scores at randomly sampled sites throughout the genome. With random sampling of the background read densities to generate a probability distribution, PeakHunter takes into account the signal-noise ratio variation that exists between independent DNaseI-seq datasets in setting the threshold read density for peak calling.

In total, we identified 10 open chromatin sites over a ∼200 kb segment of the MICA and MICB locus **(Fig 6.8a)**, and 17 open chromatin sites over a ∼200 kb segment of the ULBP locus **(Fig 6.8b)**. For comparison, we have also shown the open chromatin sites identified in the same dataset with three other peak calling programs, HOMER (295), MACS (296) and FSEQ (297).

6.3.3 Validation of MICA-associated open chromatin sites

To validate the MCF7 DNaseI-seq data, we picked two hypersensitive sites identified by DNaseI-seq surrounding the MICA promotor, and designed qPCR primer pairs at several loci spread through the region. Primers were designed using the Primer3Plus web tool, and checked with NCBI's primer blast, and UCSC's *in silico* PCR tools. We measured the change in amount of the target amplicon relative to the *HBB* DNaseI-insensitive site in undigested or 64 unit DNaseI-digested genomic DNA, and displayed the results on the UCSC genome browser, aligned with the MCF7 DNaseI-seq data and the peaks called by PeakHunter **(Fig 6.9)**. As expected, we observed a reduction in the amount of target amplicon at the hypersensitive sites, and little change at DNaseI-insensitive sites.

6.4 Discussion

Our specific aim in this chapter was to map the open chromatin sites at NKG2D ligand loci, to develop and optimise the laboratory methods, including a means to quantitatively measure DNaseI digestion at the pre-library stage. We also aimed to develop the analytical 'pipeline' necessary to identify useful features in this data. We have achieved each of these aims.

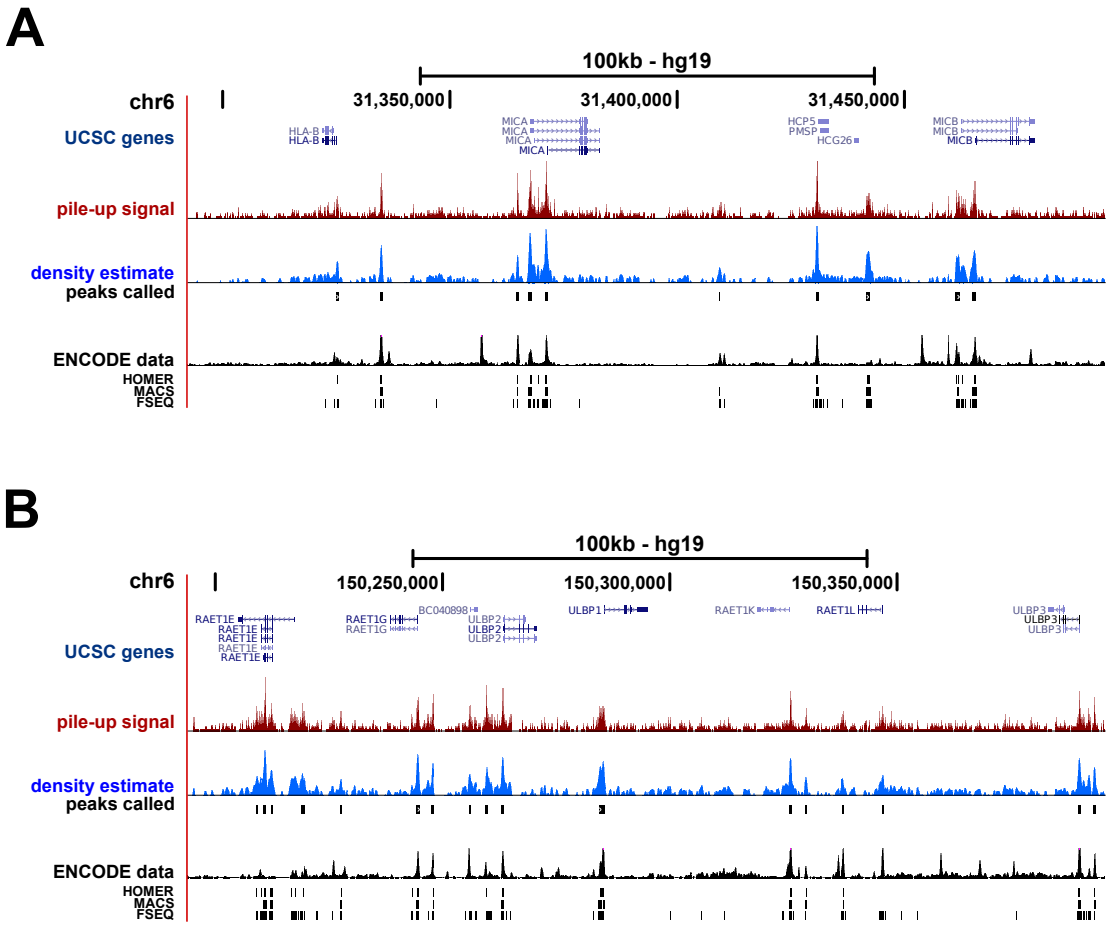

Figure 6.8: Open chromatin sites near NKG2D ligands in MCF7 breast cancer cells. To display the MCF7 DNaseI-seq data results on the UCSC genome browser (298), we loaded the data to local servers, and accessed these files via the UCSC browser following the instructions provided by UCSC (http://genome-preview.ucsc.edu/goldenPath/help/customTrack.html). (A) A 200 kb segment of the MHC including the *MICA* and *MICB* NKG2D ligand genes is depicted. The raw pile-up signal (*red*), the smoothed probability density signal (*blue*), and the sites identified as open chromatin by PeakHunter (*peaks called*) are shown. After completing our DNaseI-seq experiment, the ENCODE group also released a DNaseI-seq dataset in MCF7 cells, and this is included below. Finally, for comparison with PeakHunter, we called peaks in our MCF7 dataset with three other currently available peak callers, HOMER, MACS and FSEQ, and the areas of open chromatin identified by these programs is shown below. (B) An analysis identical to that described in section (A) is presented for the ULBP encoding locus on chr6.

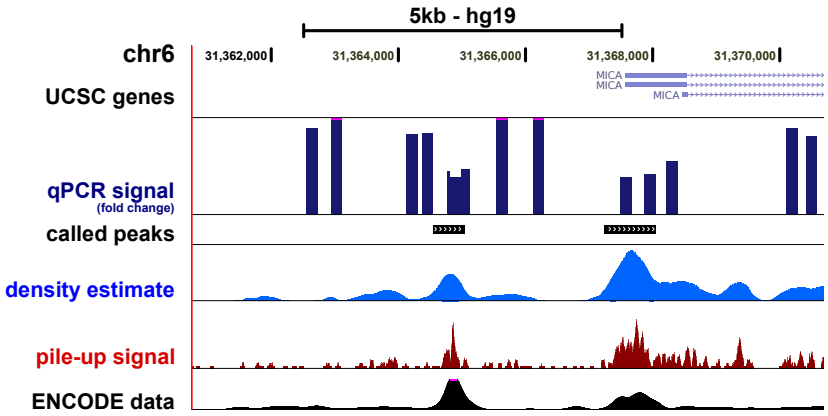

Figure 6.9: Validation of two MICA hypersensitive sites by qPCR. We sought to validate our MCF7 DNaseI-seq dataset results by using DNaseI-qPCR. We designed primers for 14 sites spread through a 10 kb region upstream of the MICA gene. We measured changes in the amount of each target amplicon in relation to a DNaseI-insensitive site (*HBB*) in the same DNaseI-digested or undigested samples. Amplicons at sites identified as hypersensitive by DNaseI-seq were reduced, while amplicons at insensitive sites were unchanged.This confirms hypersensitivity at the DNaseI-seq identified sites.

6.4.1 Measuring digestion

The DNaseI hypersensitivity assay, involving the stepwise, semi-quantitative identification of DNaseI hypersensitive sites by southern blotting and radioactive probing of DNaseI digested material, was once the gold standard assay in this field. The development of DNaseI-seq, which allows the genome-wide identification of open chromatin sites, has transformed the field in recent years. DNaseI-seq has its own challenges. A key challenge that we have observed during this work is the variability in signal-noise ratio between data sets, which is critically dependent on the quantity of DNaseI digestion. We have demonstrated a qPCR based method to quantify DNaseI digestion in this setting, prior to construction of the DNaseI library for sequencing **(Fig 6.6)**. The results of this qPCR assay correlated well with the amount of digestion visualised by pulsed-field gel analysis **(Fig 6.4)**, and it offers two advantages over pulsed-field gel electrophoresis. First it offers a means of precise numerical quantification of the digestion due to DNaseI activity that has taken place. A second pragmatic benefit is that the equipment necessary to carry out qPCR is generally widely available, while pulsed-field gel apparatus is more is likely to have lower availability.

There are some potential limitations with this qPCR method. Following DNaseI digestion of the isolated nuclei, total DNA is purified by phenol-chloroform extraction, and ethanol precipitation. Particularly in slightly digested samples, the presence of long strands of genomic DNA makes the material viscous, which leads to difficulty separating the aqueous upper DNA-containing phase from the protein rich interphase. In contrast, heavily digested samples are much less viscous, and separation is less technically challenging. If extracted DNA samples contain contaminants from this process, particularly digestion dependent contamination with phenol or chloroform for example, it may affect the accuracy and reliability of qPCR as a measurement of digestion. To overcome this, it is important to carry out amplification curves from serial dilutions of digested DNA samples, which will demonstrate the presence of such contaminants by skewing the calculated amplification efficiency.

6.4.2 Signal-noise ratio

The signal-noise ratio in our data set was low (for comparison with other publicly available DNaseI-seq data set signal-noise ratios, see **Fig 7.1**). We expect that this is largely due to the inclusion in the library preparation of over-digested DNA, most likely from the 32-unit sample. Because our principal interest is in hypersensitivity, we expect that relatively light digestion would expose these sites better, and reduce the amount of non-hypersensitive DNaseI cleavage sites. On the opposite end of the scale, under digested samples with very high signal-noise ratios may miss some small hypersensitive sites of interest. In future experiments we hope to quantify this, and improve our ability to identify optimally digested DNaseI samples before sequencing.

6.4.3 DNaseI hypersensitive sites

We have identified the DNaseI hypersensitive sites at the NKG2D ligand loci, in a cell line where most of the NKG2D ligands are expressed at the cell surface **(Fig 6.8)**. This is a useful baseline map of the open chromatin in the area. It is likely that open chromatin sites

at the beginning of each annotated gene represent the gene promoter. From this data, it is not possible to presume that open chromatin sites outside the promoters represent relevant regulatory DNA elements that are important for gene regulation, or even gene expression. To achieve this, we plan to continue this approach by looking for inducible DNaseI hypersensitive sites in the region, and to correlate inducibility with local gene expression. This will help to build a clearer picture of the DNA elements necessary to the process of NKG2D ligand expression.

6.4.4 Summary

In summary, we achieved our principle objectives of identifying the open chromatin sites that occur at the NKG2D loci in MCF7 cells cultured in high glucose. We identified several challenges in the DNaseI-seq methodology, and have made several steps forward in finding solutions to these problems. We have developed a bioinformatic 'pipeline' to enable swift, efficient and accurate data analysis for DNaseI-seq assays. These experiments provide a good platform for our next experiment in a physiological model, geared to identifying dynamic changes in open chromatin sites at the NKG2D ligand loci.

7

PeakHunter: a kernel density estimator based program for DNaseI-seq data analysis

Contents

7.1 Abstract

The development of high throughput sequencing in the past decade has lead to sweeping changes in the field of gene regulation. In particular, high throughput sequencing of DNaseI

hypersensitive genomic DNA (DNaseI-seq) permits the rapid identification of open chromatin sites throughout the genome. The identification of open chromatin sites requires mapping of sequenced short reads to a reference genome, and subsequent identification of genomic sites associated with significant clustering of mapped reads. While several software applications exist for short read mapping, the identification of sites with significant clustering of mapped DNaseI-seq reads ('peaks') is challenging. Variation in DNaseI enzyme activity between experiments leads to inherent inter-sample signal-noise ratio differences, which is reflected in the resulting high-throughput sequencing data. Existing peak callers demonstrate wide variation in the number of peaks identified between datasets, and there are wide quantitive and qualitative differences in peaks identified in any given dataset with different peak callers. These differences are largely dependent on inherent signal-noise discrepancy. The PeakHunter algorithm employs kernel density estimation of the 'background' DNaseI signal (noise), and uses the resulting probability distribution to identify an appropriate read density threshold, above which read density is significantly different from the background signal. Combined with a range of other mapped read manipulation tools, PeakHunter should make a useful addition to the available peak calling software repertoire.

7.2 Introduction

7.2.1 The DNaseI-seq assay

DNaseI-seq is a high throughput approach to the genome-wide identification of open chromatin sites (described in greater detail in **Chapter 6**). The method is based on the principle that open chromatin (genomic DNA from which histone proteins have been displaced; see **Fig 6.1**) is more susceptible to digestion by the enzyme DNaseI than is DNA wrapped around histones. There is increasing evidence that such sites of open chromatin are functional in gene regulation, or have greater functional potential than closed chromatin (289, 290). To identify these sites in a population of cells, intact nuclei are exposed to the enzyme DNaseI,

which cuts out DNA fragments from accessible loci. These fragments of open chromatin DNA are isolated and purified, standard sequencing DNA oligomers are ligated, and the millions of resulting DNA fragments are sequenced in parallel **(Fig 6.2)**. The sequences are bioinformatically mapped back to the reference genome, and open chromatin sites are identified as clusters of mapped reads.

7.2.2 Inherent problems with the DNaseI-seq approach

A key challenge with this approach is the variability in DNaseI activity between experiments. Variation in the amount of DNaseI digestion leads to variation in DNaseI-signal intensity **(Fig 7.1a)**. Because there are no independent, precise markers of the quality of DNaseI-digestion during library preparation, variation in signal intensity and therefore signal-noise ratio is common in these datasets **(Fig 7.1b)**. As this is an intrinsic characteristic of the DNaseI enzyme, it is difficult to eliminate experimentally.

7.2.3 Current DNaseI-seq data analysis practice

DNaseI-seq data is analysed in two stages. The first is short read mapping, and a wide range of software applications exist to achieve this, including Stampy (294), bowtie (299), BWA (300) and MAQ (301). Identifying the DNaseI-seq signal in the mapped data is less straightforward. While simple empirical peak calling is an accepted practice in current literature (291), a range of software applications also exist for identifying clusters of mapped reads using various approaches. However, most of these applications are designed for ChIP-seq data, and require an input control data set that is typically carried out with these methods, but does not exist for DNaseI-seq.

When analysing our mapped DNaseI-seq data, we initially used three separate peak calling programs: FSEQ (297), MACS (296) and HOMER (295). We found that there was wide variation in the number of peaks identified in individual DNaseI-seq datasets using different peak callers, and between different datasets with individual peak callers **(Fig 7.2)**. FSEQ,

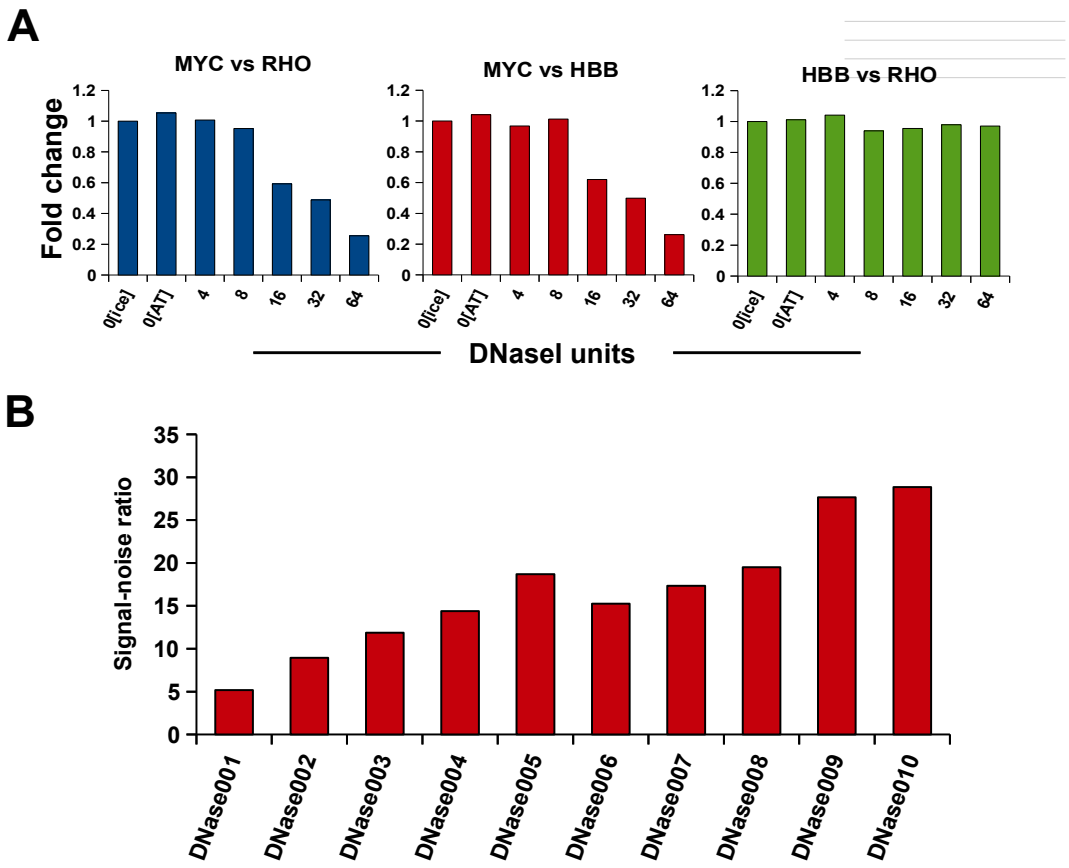

Figure 7.1: DNaseI signal intensity varies with extent of digestion. (A)We used real time qPCR to quantify the amount of DNaseI digestion that had occurred at a hypersensitive site, when isolated nuclei were digested with increasing concentrations of the enzyme DNaseI. Two DNaseI insensitive sites at the rhodopsin (RHO) and haemoglobin beta (HBB) genes were measured, along with one hypersensitive site, upstream of the c-myc gene. These sites were chosen on the basis of DNaseI-seq results available on the UCSC browser. The optimal level of digestion to demonstrate hypersensitivity in a DNaseI-seq library is unclear, and hypersensitivity at any given positive control site is liable to change between cell types and environmental conditions. (B) To estimate signal-noise ratio, we used the open chromatin sites discovered by the ENCODE consortium in 125 DNaseI-seq datasets. For our 'signal' estimate, we measured the mean read density in a 300 bp bin at the centre of ∼50,000 randomly selected sites that have been identified as open chromatin in 60 or more ENCODE datasets. For our estimate of random 'noise', we made the same measurement at sites found by ENCODE to be closed chromatin sites. The score shown for each data set is a ratio of the two values **(table 6.1)**. This demonstrates wide variation in signal-noise ratios across DNaseI-seq data sets.

which uses a local threshold model for peak calling tended to call more peaks in datasets with low signal-noise ratio, while HOMER and MACS, which use a local enrichment over background model, identified fewer peaks in these datasets.

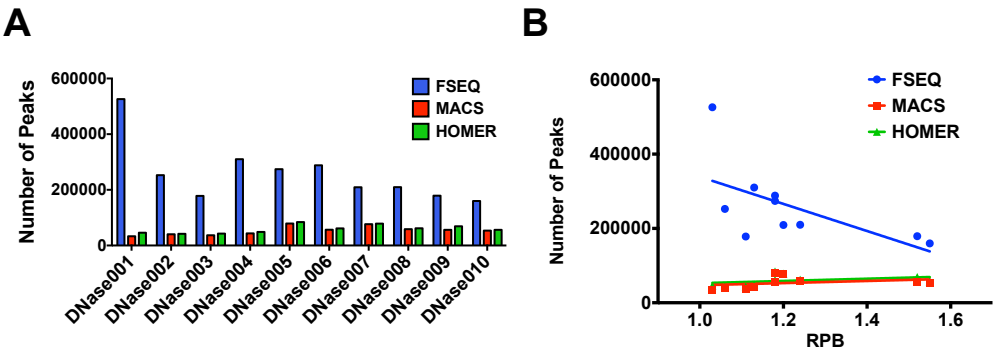

Figure 7.2: Variability in peak numbers with different available peak callers. (A) As a measure of signal spread in DNaseI-seq data sets, we calculated the number of reads in each data set, and the number of bases over which these reads were spread to calculate the number of reads per base (RPB) in each data set. To see how different available peak callers performed across a range of DNaseI-seq datasets with variable RPB scores, we called peaks in 10 sample DNaseI-seq datasets **(Table 7.2)**. We found wide variation in the numbers of peaks identified with different peak callers, and between different datasets with the same peak callers. (B) Plotting the number of peaks identified in each dataset against the RPB score of the associated sam file suggests a correlation between the two.

7.2.4 An alternative approach to DNaseI-seq data analysis

In order to overcome some of the limitations we experienced in analysing our DNaseI-seq data, we sought to develop a set of analytical computational tools to identify and validate 'peaks' of hypersensitivity in our DNaseI-seq data. Our objective was to develop a set of programs that would run on all major computing platforms and that was easy to install and use. From a theoretical point of view, we wanted a program that would identify a simple global density threshold, based on a model of background read density distribution, accounting for signal-noise ratio variation between DNaseI-seq datasets. Finally, wanted the set of tools capable of carrying out all of the necessary analytical steps from the mapped short read file through to a file containing a list of peaks. PeakHunter is written in Perl and is available for download at http://www.ccmp.ox.ac.uk/peakhunter.

7.3 PeakHunter methods and design

7.3.1 An overview of the data-analysis process

PeakHunter is designed to analyse mapped short-read sequence files in the standard SAM format (302). An overview of the analysis 'pipeline' is shown below **(Fig 7.3)**. PeakHunter has two key outputs: a whole-genome read density track in fixedStep wig format at single base resolution ('density analyser'; see http://genome.ucsc.edu for file format definitions), and a bed file listing peaks identified from the mapped read data ('peak caller'). It also contains sorting and filtering utilities as described in **Table 7.1**.

PeakHunter is designed to work with any genome, using a genome specific, user-provided chromosome size specification file. This is an ordinary text file containing a list of chromosome names, and chromosome sizes, which can be used to control chromosome inclusion and order of analysis, and also to ensure that peak boundaries do no extend beyond chromosome size limits.

7.3.2 Preparation of data for density analyser

For each of the datasets analysed, the raw sequence data was downloaded from the NCBI SRA (short read archive) data repository, and mapped to an alternative haplotype-free reference genome using Stampy. To build the reference genome for Stampy, we used the commands:

./stampy.py -G hg19 /chromsomeDirectory/*.fa.gz

./stampy.py -g hg19 -h hg19

and to map data with stampy, we used the command:

./stampy.py -g hg19 -h hg19 -M reads.fastq >mappedReads.sam

The mapped sequence files were filtered to exclude poorly mapped or sequenced reads (in this case a MAPQ score of less than 15 or UQ score greater than 150). To reduce the

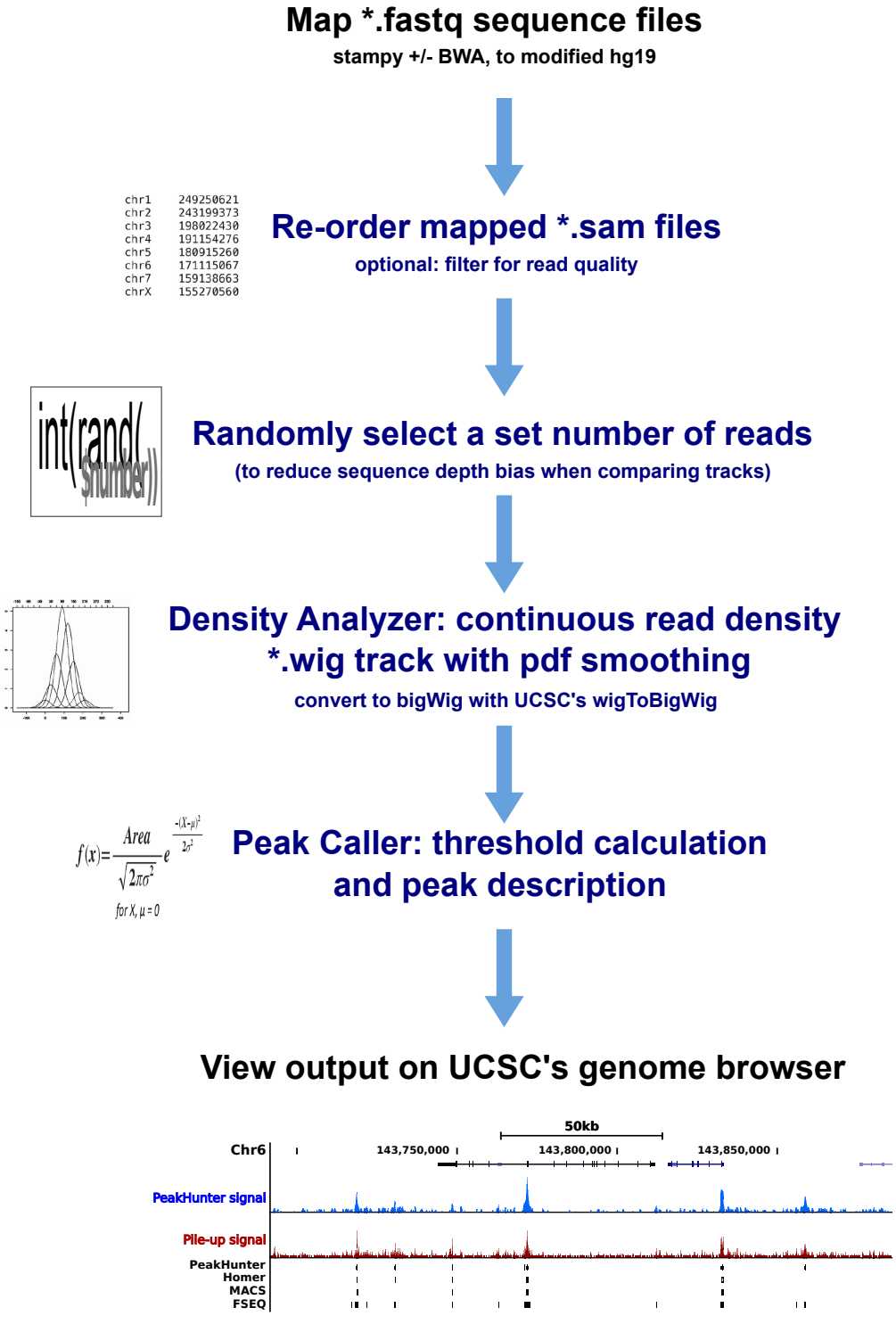

Figure 7.3: DNaseI-seq data analysis pipeline. Raw sequencer output files (fastq) can be mapped to the reference human genome sequence using a range of available programs (we have used Stampy (294)). Mapped short read data, can be re-ordered and quality filtered as required. To reduce the impact of variation in depth of sequencing, we then randomly select a set number of reads (17,000,000 in our analysis) from the filtered good quality reads. The density analyser component of PeakHunter creates the density track, while the peak caller function identifies significant read clusters as peaks in the data. All steps shown in blue can be carried out using tools provided by PeakHunter.

Table 7.1: PeakHunter functions and parameters.

Function	Options	Description
Sam Filter (-F)	-q [mapq threshold] -u [uq threshold] -i [include header] -g [chromosome size file]	Filters out reads with a mapq score less than (-q) and mismatch score greater than (-uq). The output only includes chromosomes listed in the chromosome size file, in the order in which they are listed.
Density Analyzer (-D)	-t [minimum bin threshold] -m [maximum bin threshold] -n [one-sided length of bin] -s [Gaussian step size] -d [Gaussian sigma value] -o [track offset] -g [chromosome size file]	Analyzes an ordered mapped reads file in sam format. Assigns a unitless probability density score to each base in the genome, representing its probability of being a hypersensitive site. Only chromosomes given in the chromosome size file are scored.
Peak Caller (-P)	-g [chromosome size file] -b [blue print file from ENCODE data] -bin [target bin size] -background [target background size] -npBack [number of bins for background estimation] -sig [threshold significance level]	Identifies peaks in mapped sam files. If a blueprint file is specified (-b), background densities are sampled from ENCODE closed chromatin sites, otherwise the background signal is estimated from randomly selected sites.
Random Read Selector (-R)	-nr [number of reads]	Randomly selects a target (-nr) number of reads from a given sam file.
Signal Intensity (-S)	-dc [distance from centre] -np [number of peaks] -t [target sam file]	Selects [-np] peaks from given peak bed file, and calculates the mean reads per base score in target sam file (-t) for [-dc] bases either side of the peak centre.
Peak Selector (-T)	-n [number of peaks]	Selects the top (-n) scoring peaks in given bed peak file.
Numerical Sam Sort (-NS)		Sorts all reads in given sam file into ascending numerical order by base position.

impact of sequencing depth on our analysis, we randomly selected 17,000,000 reads from each dataset for analysis. 17,000,000 is an arbitrary figure, corresponding to the number of reads in the smallest sequencing file used in our analysis.

7.3.3 Density analyser (probability density function)

The density analyser component of PeakHunter is designed to improve the visual localisation of the DNaseI signal. It achieves this by smoothing the raw signal using a probability density function. In practical terms, we consider that the starting base of each mapped read represents a location in the genome that the DNaseI enzyme has cut the genomic DNA. We measure the density of these DNaseI cuts, by dividing the genome into overlapping bins, and counting the number of DNaseI cuts in each bin **(Fig 7.4a)**. The total number of cuts in any given bin represents the read density score of that bin. This density score is redistributed through the bases making up that bin according to a gaussian distribution (probability density function), where σ is the standard deviation of the distribution, and x is the distance in bases from the centre of the distribution:

$$G(x) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

Because the bins overlap, this generates a continuous probability score through the genome, representing the probability that DNaseI has cut the genome at any particular base. The probability scores given to each base by overlapping bins are added together to give a total probability score for that base. The degree of smoothing can be adjusted by changing the standard deviation (σ) of the probability density function **(Fig 7.4b)**.

The resulting density scores are recorded in a text file in the standard wig fixedStep format at a single base resolution. To visualise these density tracks, we converted them to the bigWig file format and loaded them to the UCSC genome browser as described on the UCSC website. An example of the density track is shown in **Fig 7.4c.**

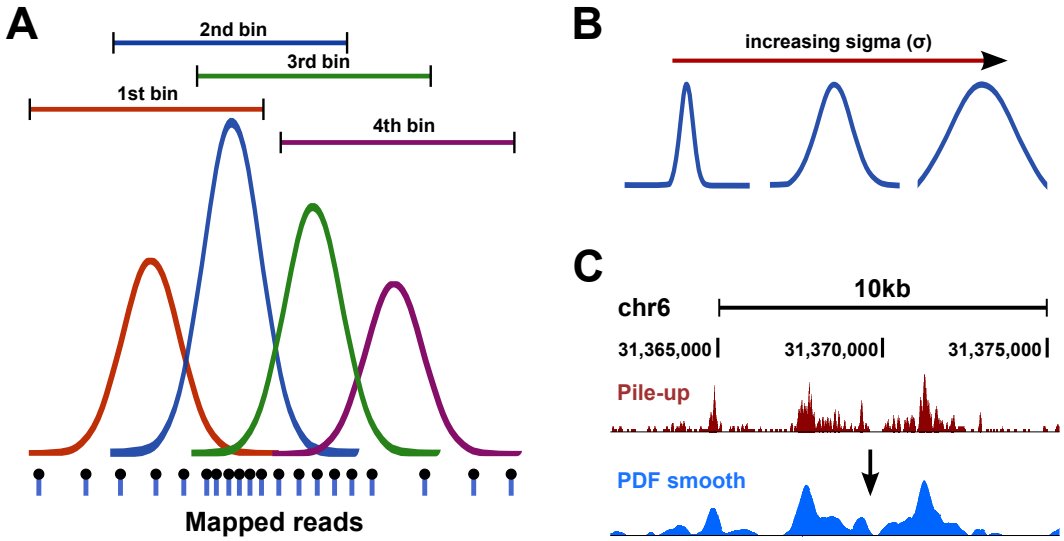

Figure 7.4: Using a probability density function to smooth read density scores. (A) To improve the visual localisation of the signal, we measured read density, as opposed to raw read counts (pile-up), and smoothed these density scores, using a probability density function (PDF). The PDF describes the coordinates of a Gaussian distribution. The pdf moves through the genome in consecutive, overlapping bins. In each bin, the number of reads is counted, and this read score is distributed normally throughout that bin according to the PDF values. Scores from overlapping bins are added together to give a relative probability density score to each individual base. (B) By adjusting the standard deviation (σ) of the probability density function, we can control the degree of smoothing with minimal effect on computational performance. (C) An example of the smoothing in real data. PDF - Probability density function.

7.3.4 Threshold selection and peak calling

7.3.4.1 The PeakHunter sampling bin

The central aim of the PeakHunter peak calling program is to identify genomic loci where clustering of mapped reads is significantly above background. To achieve this aim, measurements of read clustering are made using a simple sampling bin **(Fig 7.5)**. This bin is used to calculate the read density above local background for any given site. To calculate the read density for any given bin, the number of DNaseI-seq read start sites falling within the central bin are counted. The total number of reads in a wider local background bin are also counted. The number of reads expected to occur in the central bin, given the number of reads in the local background is calculated to the nearest whole number, and subtracted from the number of reads in the central bin, to give the density score. The sizes of both central and background bins is user-adjustable. For this analysis we have set the central bin size to 300 bp, the expected average size of an open chromatin site, and the background bin size to 3000 bp. When we refer to density sampling, or density measurement at a later point, the sampling and measurement is carried out using this bin.

7.3.4.2 The ENCODE open chromatin sites

The gold standard assay for confirming hypersensitivity is the DNaseI hypersensitivity assay first used by Wu in 1980 (280). However, this method is restricted to short segments of genomic DNA, and quantification of hypersensitivity is not precise. The only method capable of genome-wide DNaseI hypersensitivity measurement is DNaseI-seq itself, which presents some challenges in validating the open chromatin sites detected. In 2012, the ENCODE consortium published DNaseI-seq data, along with peak calling analysis across 125 different cell lines (303). This data is currently the best available definition of open chromatin sites across the human genome. We made use of this data in the design of PeakHunter, and the validation of PeakHunter's output. From the peaks identified by ENCODE in all 125 DNaseI-

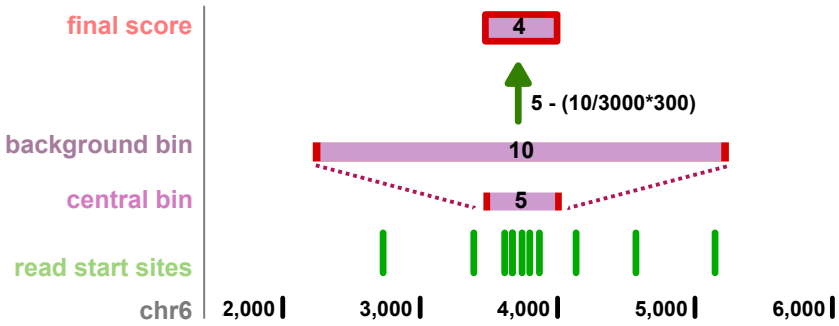

Figure 7.5: Sampling bin used by PeakHunter. A simple sampling bin is used to measure read density by PeakHunter. This consists of a central bin (by default set to 300bp; this value is user-adjustable), and a larger background bin (by default set to 3000bp; this value is user-adjustable). To calculate the read density above local background for any given site, the number of DNaseI-seq read start sites lying within the central bin is counted.The expected read density given the local background is calculated, by counting the number of reads in the background bin, and calculating the expected number of reads that should occur in a bin of central bin size (in this case, 10 reads in a background bin of 3000bp, equates to an expected 1 read per 300bp bin). This value is subtracted from the central bin read count, giving the read density above local background.

seq data sets, we generated an open chromatin 'master track' **(Fig 7.6)**. This compresses the open chromatin sites from all 125 data sets, into one file. The result is a continuous genome-wide list of elements, annotated with the number of ENCODE data sets that have an open chromatin site within that element. From this file we can identify genomic loci that exist as 'known' open chromatin and sites that exist as closed chromatin in all ENCODE-tested cell lines.

7.3.4.3 Estimation of signal-noise ratio

Given the technical challenges of precisely controlling DNaseI-seq activity and measuring this digestion, we expected to see variation in signal-noise ratio between different DNaseI-seq datasets. To estimate signal-noise ratio in the sequencing data, we used the ENCODE 'master track'. For the estimation of signal strength, we randomly selected genomic loci that were frequently (present in 60 or more data sets) designated as open chromatin across the 125 data sets. Using our sampling bin, we calculated the mean read density above background at the centre of each selected site. Similarly, to estimate noise, we carried out the same procedure,

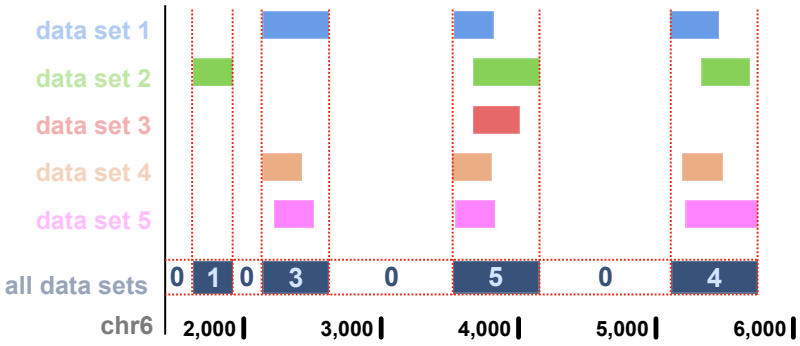

Figure 7.6: ENCODE open chromatin sites. The ENCODE consortium have recently published the coordinates of open chromatin sites, measured by DNaseI-seq, across 125 separate cell lines. This represents the best available map of open chromatin elements in the human genome that has been published to date. We have capitalised on this by amalgamating all 125 data sets into one 'master track' of open chromatin, covering the entire genome. This track is annotated with the number of cell lines (out of 125) that have open chromatin at each site. We have used this data for calculation of signal-noise ratios, estimation of background signal frequency distribution, and the validation of PeakHunter output (below).

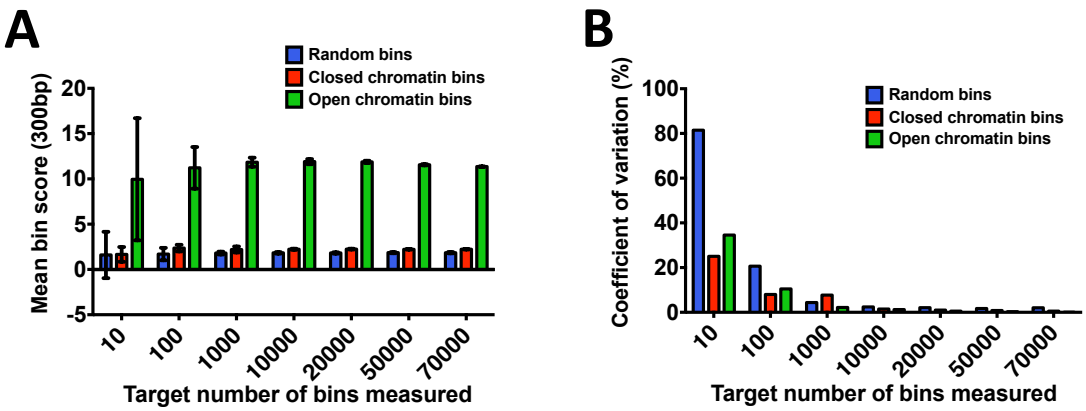

Figure 7.7: Estimation of signal-noise ratio in a low signal-noise data set. To identify the number of randomly selected sites required to achieve a reliable estimate of mean read density at either ENCODE open chromatin (open in 60 or more of 125 analysed data sets), ENCODE closed chromatin, or randomly selected genomic sites in a low signal-noise ratio data set, we randomly selected increasing numbers of sites from each category, and measured the mean read density score +/– 95% confidence interval (A) and calculated the coefficient of variation of the density scores (B). In all categories, the coefficient of variation dropped below 1% when ≥50,000 sites were sampled.

randomly sampling sites designated as closed chromatin in all 125 analysed datasets.

To establish the number of randomly sampled sites required to achieve a reliable estimate of signal or noise, we sampled an increasing number of randomly selected sites in each category, and calculated the coefficient of variation at each level. We found that the coefficient of variation fell below 1% for all estimates at 50,000 sites in both low signal-noise ratio data sets **(Fig 7.7)**, and high signal-noise ratio data sets **(Fig 7.8)**. For these calculations, the displayed number of sites were randomly selected, and any overlapping sites were discarded.

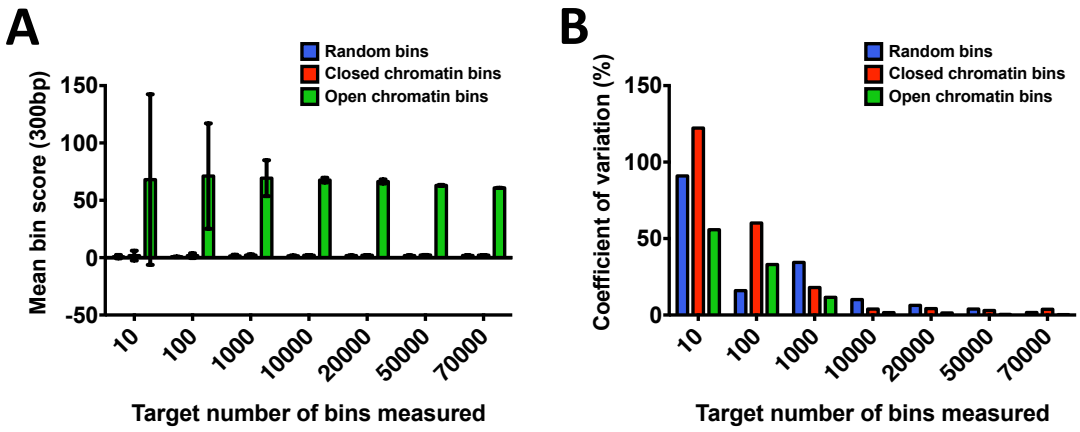

Figure 7.8: Estimation of signal-noise ratio in a high signal-noise data set. We repeated the calculations performed in **Fig 7.7** to identify the number of sites necessary to accurately estimate mean read densities in a high signal-noise ratio dataset, at either ENCODE open chromatin (open in 60 or more of 125 analysed data sets), ENCODE closed chromatin, or randomly selected genomic sites (A). Similarly, we found that the coefficient of variation fell below 1% in all categories when ≥50,000 sites were sampled.

7.3.4.4 Estimation of 'background' signal and threshold selection

A key challenge in statistically distinguishing signal from noise is measuring the noise itself. We approached this problem by employing a Gaussian kernel density estimator, a non-parametric approach to calculating a probability distribution from empirical data **(Fig 7.9)**. To generate a probability distribution for background noise using this method, a set number of genomic loci are chosen at random to represent the background signal. The read density above local background at each site is scored using the sampling bin. Each measured density score is given its own Gaussian probability distribution 'kernel', of area 1 unit, centred on the

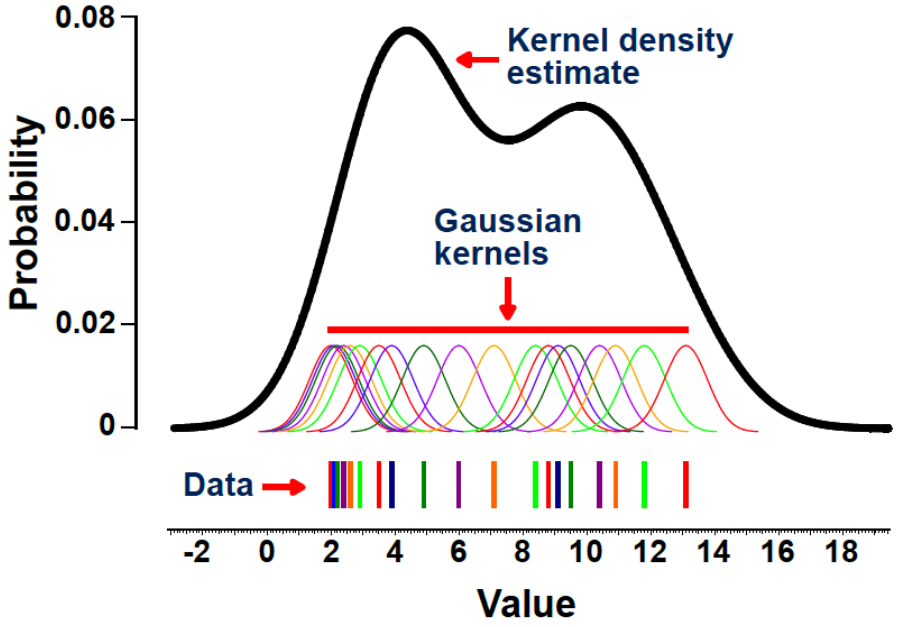

Figure 7.9: Gaussian kernel density estimator. Kernel density estimation is a non-parametric method to generate a probability distribution from empirical data. The sample 'data' illustrated above is not normally distributed. Therefore, the usually summary statistics that are useful for normally distributed data, such as the sample mean and standard deviation, are less useful. Instead, we assign each value a probability kernel, with a surface area of 1 unit, and centred at its value on the X axis. Many different kernel shapes have been described, for PeakHunter, we have chosen Gaussian kernels. To calculate the probability at any given value of X, we sum all the probabilities in overlapping kernels at that point, the total sum being divided by the total number of data points used in the calculation. Calculating the sum of all probability scores from each kernel at all values of X gives the probability distribution shown (*black line*). By sampling read densities from the background signal (ENCODE closed chromatin, or randomly selected genomic sites), PeakHunter similarly constructs the probability distribution of the 'noise' present in the data set. To calculate the threshold density at which a read density score is significantly different form the noise, PeakHunter calculates the value of P for increasing values of X. The threshold applied is the value of X at which the value of P first falls below the desired significance level.

read density score itself. The probability (P) that any given read density score (X) belongs to the population of 'background' read density scores can be calculated by summing together all the probabilities contributed by each individual Gaussian kernel to that value of X.

The probability distributions are calculated with:

$$f(x) \quad = \quad \frac{1}{nh} \sum_{i=1}^{n} K_u \left(\frac{x - x_i}{h} \right)$$

where *K(u)* for a Gaussian kernel shape is calculated with:

$$K_u \quad = \quad \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}$$

To illustrate these distributions, we have calculated the distribution values in a low signal-noise ratio data set **(Fig 7.10)**, and a high signal-noise ratio data set **(Fig 7.11)**, and extracted from this the calculated kernel probability estimates for side-by-side comparison **(Fig 7.12)**. To calculate the threshold read density for peak calling, PeakHunter uses the probability distributions calculated from randomly selected ENCODE closed chromatin, or from randomly selected genomic sites. For increasing values of X (read density), PeakHunter calculates the probability (P) that the tested read density score lies within the distribution of background signal read density scores. When the value of P reaches a defined significance level, the corresponding value of X is used as the peak threshold. For the data presented here, we have used a probability cut-off of $P \leq 0.001$. This significance level can be set by the user.

To identify the number of ENCODE closed chromatin or randomly selected sites required to obtain a reliable estimate of the threshold read density, we calculated the threshold and coefficient of variation of threshold estimates using increasing numbers of sites in either category, for both a low signal-noise ratio data set **(Fig 7.13)**, and a high signal-noise ratio

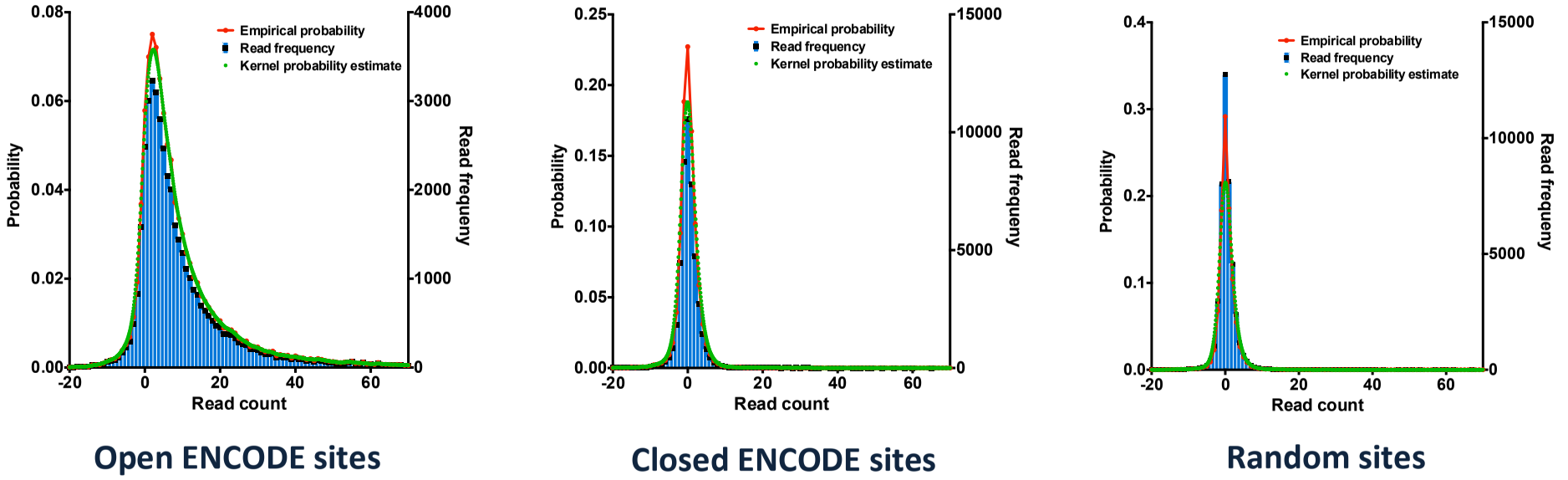

Figure 7.10: Kernel density estimation of probability distribution in a low signal-noise ratio data sets. To derive a probability distribution using kernel density estimation, we first calculated the read density frequencies at ~50,000 randomly selected sites either from ENCODE open chromatin (open chromatin present at that site in 60 or 125 analysed DNaseI-seq data sets), ENCODE closed chromatin, or randomly selected genomic sites. The blue columns depict the read density frequencies in the selected sites. A simple empirical probability can be derived by calculating the number of read densities as a fraction of the total number of read densities measured for each value of read density directly from the frequency distribution (*red*). The kernel probability estimates *green* use this empirical frequency distribution to generate a continuous probability distribution that closely tracks the empirical probability where values are available. The probabilities calculated from randomly selected sites closely match those calculated from ENCODE closed chromatin sites.

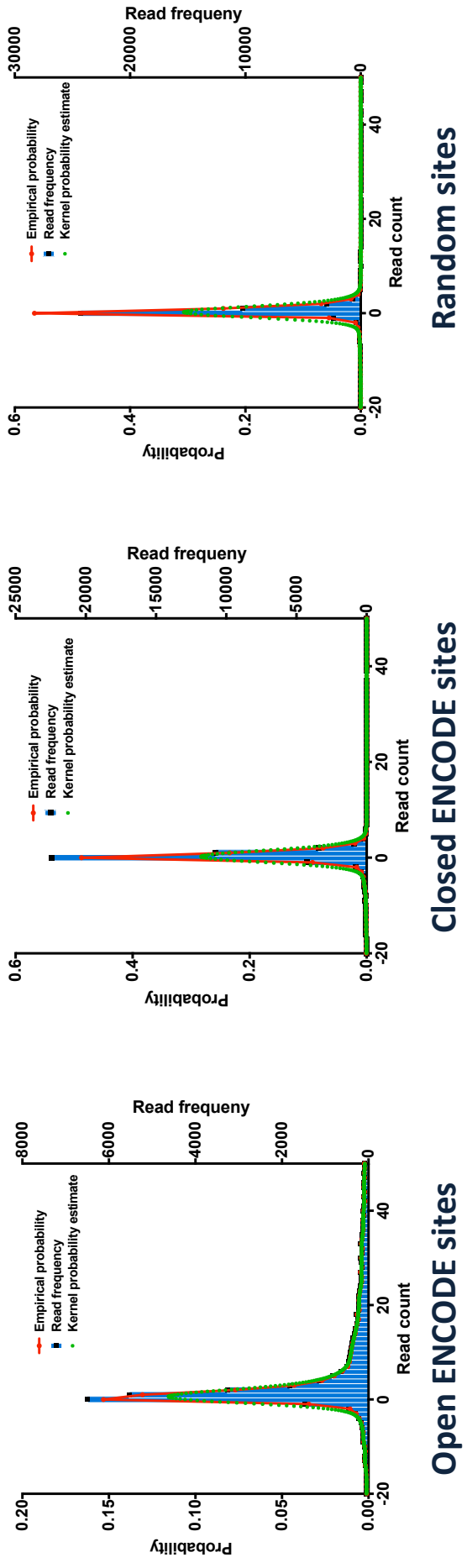

Figure 7.11: Kernel density estimation of probability distribution in a high signal-noise ratio data sets. We repeated the calculations described in **Fig 7.10** on a dataset with high signal-noise ratio.

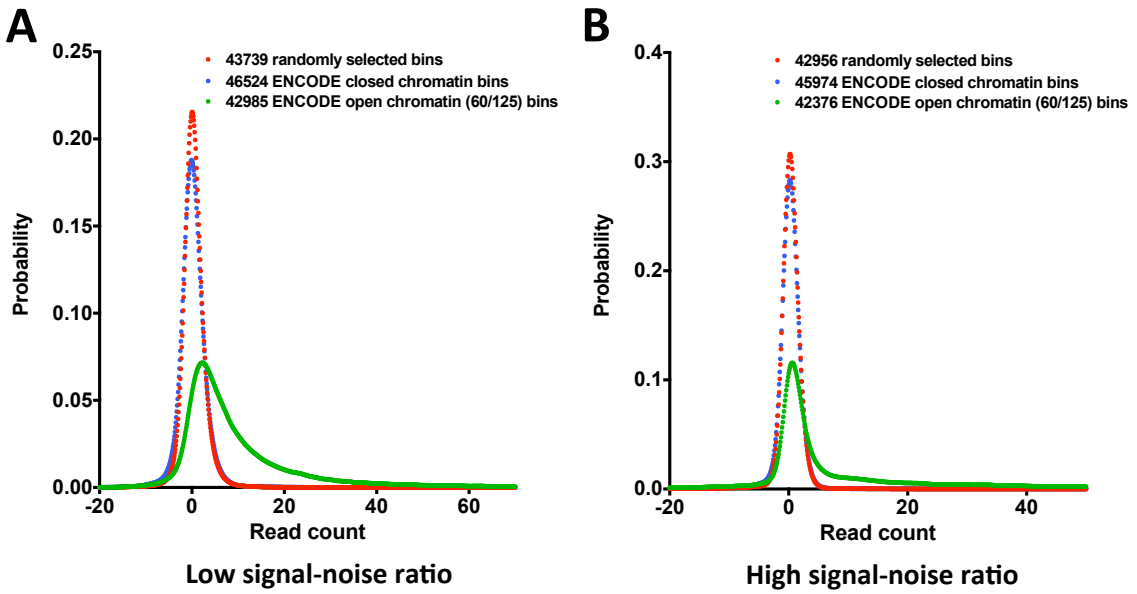

Figure 7.12: Kernel density estimation of probability distribution in high and low signal-noise ratio data sets. The kernel probability estimates depicted in **Fig 7.10** (A) and **Fig 7.10** (B) are shown. For selection of a read density threshold for peak calling, PeakHunter uses the ENCODE closed chromatin (*blue*) when this data is available. However, when the bins for sampling are randomly selected from the genome (*red*), the probability distribution closely matches that calculated from ENCODE closed chromatin

data set **(Fig 7.14)**. The coefficient of variation dropped below 1% when $\geq$ 50,000 sites were used in the calculation.

Peaks identified by the peak calling function of PeakHunter are recorded in bed format. Each peak description contains start and end genomic coordinates and the maximum density of the peak. These bed files can be viewed directly in most genome browsers.

7.3.5 Peak-sorting tools

PeakHunter also contains a simple tool to sort identified peaks by maximum density scores. The peak selector function will find the highest scoring n peaks in the bed file, and list them in order of maximum density.

The signal intensity tool, also contained within PeakHunter, allows the calculation of the mean signal intensity in a mapped reads sam file for a randomly selected number of peak centres in a PeakHunter generated peaks bed file.

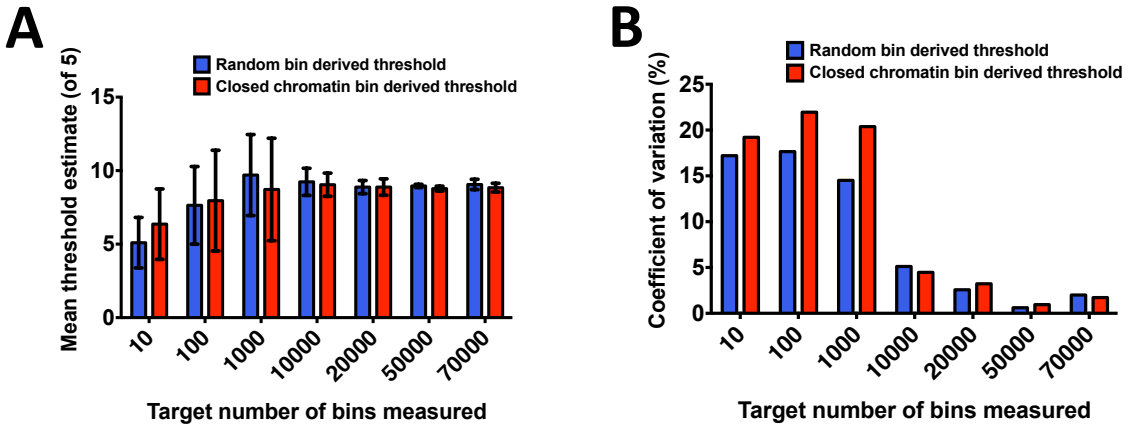

Figure 7.13: Threshold estimation in a low signal-noise ratio data set. To establish how many sites should be sampled from either ENCODE closed chromatin or random genomic sites to gain a reliable estimate of the threshold for peak calling, we calculated the threshold five times using increasing numbers of sampled sites and displayed this data as the mean threshold +/- 95% confidence interval (A). Variation in the calculated thresholds was measured by assessing the coefficient of variation (CV) (B). The CV fell below 1% when ≥50,000 sites were used for threshold calculation .

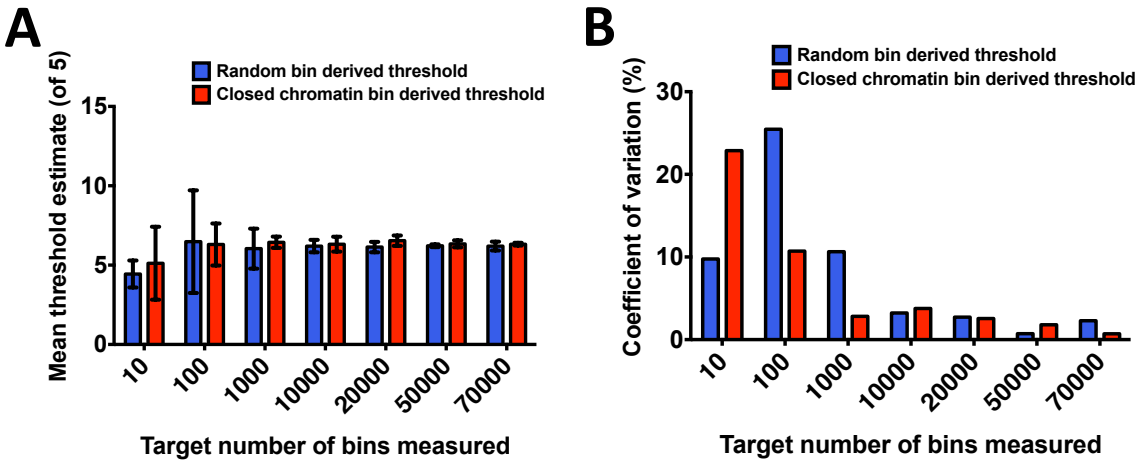

Figure 7.14: Threshold estimation in a high signal-noise ratio data set. We repeated the calculations described in **Fig 7.13** to identify the number of peaks needed to reliable estimate the read density threshold for peak calling in a high signal-noise ratio data set. We found that the coefficient of variation dropped below 1% when ≥50,000 sites were used for calculation of the threshold.

Table 7.2: Sample datasets used for PeakHunter validation

Label	NCBI SRA ref	Cell type of origin
DNase001	WTCHGMCF7	MCF7 (breast cancer)
DNase002	SRR171529	CD56+ primary cells
DNase003	SRR352544-46	CD4+ primary cells
DNase004	SRR171543	CD3+ primary cells
DNase005	SRR066352	CD34+ cord blood
DNase006	SRR352516-17	CD8+ primary cells
DNase007	SRR171576	CD14+ primary cells
DNase008	SRR171561	CD3+ primary cells
DNase009	SRR171574	CD4+ primary cells
DNase010	SRR171575	CD8+ primary cells

7.4 Results

To demonstrate PeakHunter, we analysed ten publicly available DNaseI-seq data sets obtained from NCBI's short reads archive (www.ncbi.nlm.nih.gov/sra; **Table 7.2**). Each data set was analysed as described above. Briefly, the raw sequence files were mapped to hg19 (without alternate haplotype chromosomes) by Stampy. Mapped reads with quality scores (MAPQ score) less than 15 were discarded. From the remaining good quality reads, 17,000,000 reads were randomly selected and used for peak finding by PeakHunter. For calculation of the signal-noise ratio, and background probability distributions in each data set, ∼50,000 sites were randomly selected from ENCODE open or closed chromatin sites as appropriate. The significance level used for read density threshold selection was $P \leq 0.001$.

The results of analysis by PeakHunter of all 10 data sets are summarised on **Table 7.3**. Empirical false discovery rates were calculated by generating simulated random DNaseI-seq data sets with 17,000,000 reads, and calculating the number of peaks at equivalent read density thresholds.

7.4.1 Quantitative comparison with other peak callers

To quantitatively assess PeakHunter's performance in comparison to other available peak calling programs, we called peaks in the 10 sample data sets with FSEQ, MACS and HOMER **(Fig 7.15a)**. To measure the variation in number of peaks called across the 10 sample

Table 7.3: PeakHunter analysis summary

code	name	mean signal	mean noise	SNR	threshold	no. of peaks	empirical FDR (%)
DNase001	WTCHGMCF7	11.53	2.23	5.17	9.1	37809	0.01
DNase002	SRR171529	18.15	2.03	8.94	7.6	51582	0.77
DNase003	SRR352544-46	25.61	2.16	11.86	7.7	51625	0.77
DNase004	SRR171543	30.63	2.13	14.38	8	56213	0.71
DNase005	SRR066352	34.94	1.87	18.68	8.9	87055	0.06
DNase006	SRR352516-17	30.35	1.99	15.25	7.7	69284	0.57
DNase007	SRR171576	33.45	1.93	17.33	7.4	85144	0.47
DNase008	SRR171561	39	2	19.5	7.1	64596	0.61
DNase009	SRR171574	60.85	2.2	27.66	6.7	69038	4.24
DNase010	SRR171575	63.79	2.21	28.86	6.3	63238	4.63

50,000 target sampled sites; cut off $p \leq 0.001$; FDR - False discovery rate; SNR - Signal-noise ratio.

datasets by each individual peak caller, we calculated the mean number of peaks per peak caller (+/– 95% confidence interval; **Fig 7.15b**), and the coefficient of variation for each peak caller over the 10 data sets **(Fig 7.15c)**. This demonstrated that of the four peak callers, PeakHunter had the lowest variation in the number of peaks identified across 10 data sets. To visualise a sample of this data, we converted PeakHunters density track output to the bigWig format, and loaded this track, along with bed files generated by each of the four peak callers, to the UCSC genome browser **(Fig 7.15d)**.

7.4.2 Qualitative comparison with other peak callers

In addition to quantitative comparison, we also wanted to know whether peaks identified in the sample datasets by PeakHunter or other peak callers were likely to overlap known open chromatin sites. To estimate this, we returned to the ENCODE 'master track', consisting of open chromatin sites identified by ENCODE across 125 different cell lines by DNaseI-seq **(Fig 7.6)**. For each peak in each of the 125 data sets, we counted the number of other data sets with open chromatin at that site as a percentage of the total number of peaks in that dataset. **Fig 7.16a** shows the mean percentage of unique peaks for the ENCODE data sets (at x = 1), the mean percentage of total peaks shared with one other data set (at x = 2), two other datasets, etc (*red dots*). The blue lines represent 1.96 × the standard deviation at each value of x, to illustrate the range of variation within the 125 ENCODE data sets. From this data we can conclude that in the 125 ENCODE data sets, 3.61 +/– 8.20% of the total peaks in each data set are unique. For our 10 test datasets, the amount of unique peaks identified by PeakHunter in each dataset ranges between 2.83 and 6.95% of the total number of peaks identified. In contrast, for low signal-noise ratio data sets, such as our MCF7 DNaseI-seq data, other peak callers, despite identifying fewer peaks in total, have higher percentages of unique peaks: 7.38% of the total number of peaks identified by MACS, 31.40% of peaks

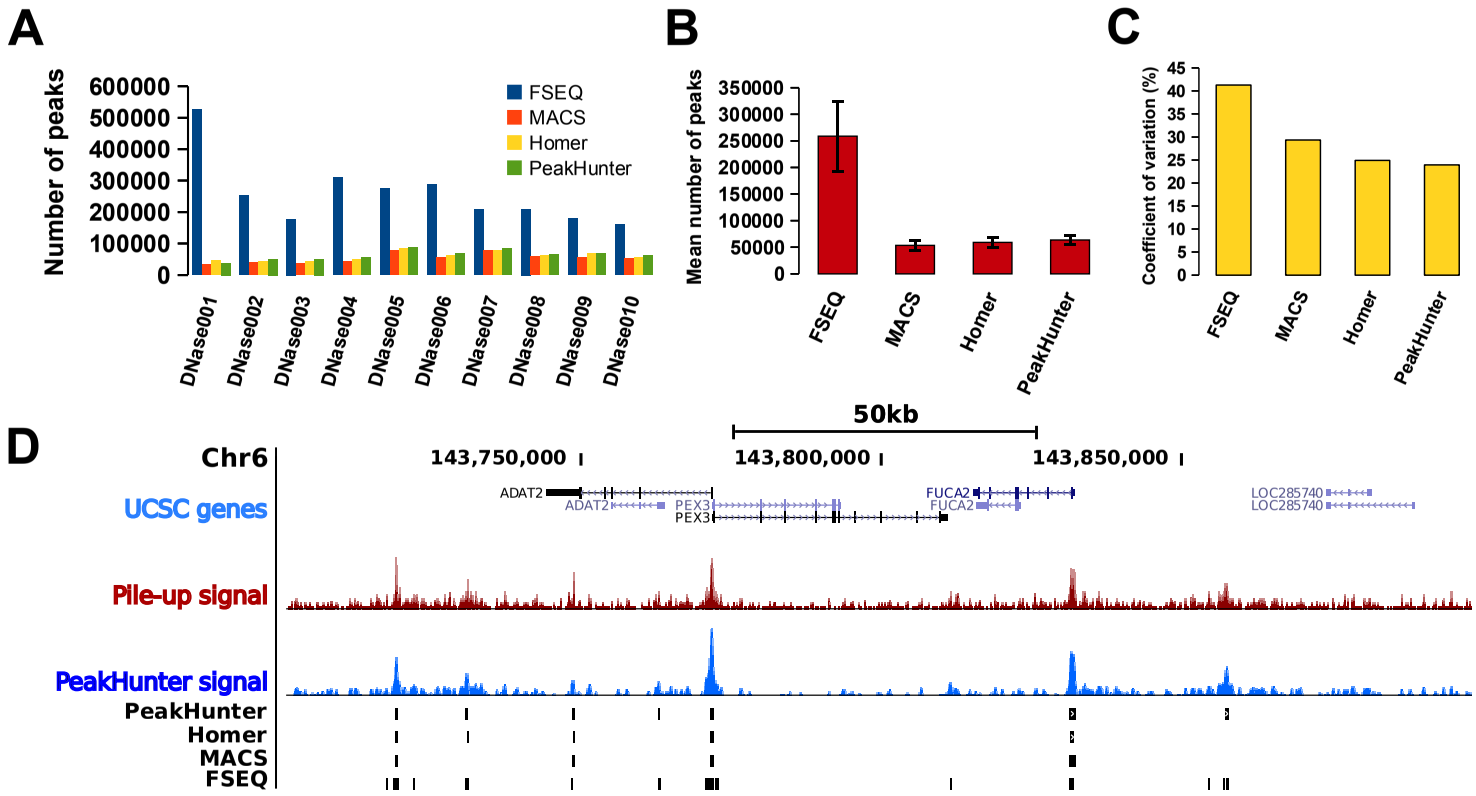

Figure 7.15: Quantitative benchmarking of PeakHunter. (A) We called peaks in 10 sample DNaseI-seq datasets with PeakHunter and compared this with the results obtained using three alternative PeakCalling programs **Fig 7.2**. (B) To quantify the variation in peak numbers, we calculated the mean number of peaks identified (+/– 95% confidence interval) across 10 datasets for each peak caller. (C) To quantify the variation shown in (B), we calculated the coefficient of variation for each peak caller. (D) The difference in performance between peak callers is illustrated by the depiction of a typical segment of the genome from the DNase001 data set, displayed on the UCSC genome browser.

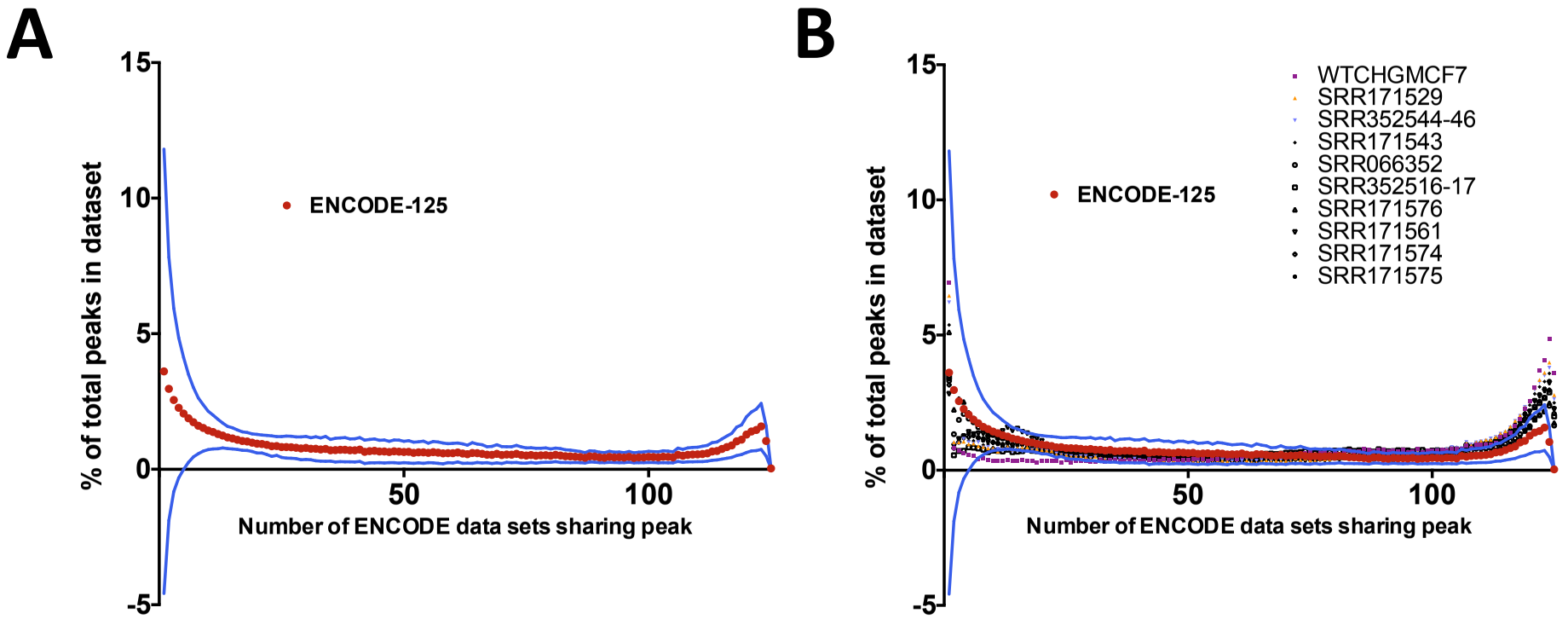

Figure 7.16: Qualitative validation of PeakHunter identified open chromatin sites. We wanted to determine whether the peaks identified by PeakHunter in DNaseI-seq data were likely to be localised to known open chromatin sites. (A) As a 'gold standard' for defining open chromatin, we used the 125 ENCODE data sets, and for each individual data set determined how many peaks were unique to that data set (across all 125 data sets), how many peaks in each data set were found in only one other data set of the 125, and so on, up to the number of peaks per data set that were common to all 125 data sets. The mean number of peaks in each category is shown, expressed as a percentage of total peaks in the individual data set (*red dots*), along with +/– 1.96 standard deviations (*blue lines*). For the most interesting data point, the number of unique peaks per DNaseI-seq data set, 3.61 +/– 8.20% of open chromatin sites in each data set were unique. (B) We used PeakHunter to identify peaks of open chromatin in our 10 test data sets. We found that the number of unique sites in each test data set (over all 125 ENCODE data sets) was within the expected range of variation. In the 'most unique' test data set, 93.05% of sites present had been designated as open chromatin in one or more of the 125 ENCODE data sets.

identified by FSEQ, and 9.64% of peaks identified by HOMER in the MCF7 data set were not found in the ENCODE data set **(Fig 7.17)**. This result suggests that the peaks identified by PeakHunter in DNaseI-seq datasets are more likely to be known open chromatin sites, than those identified by other available peak callers.

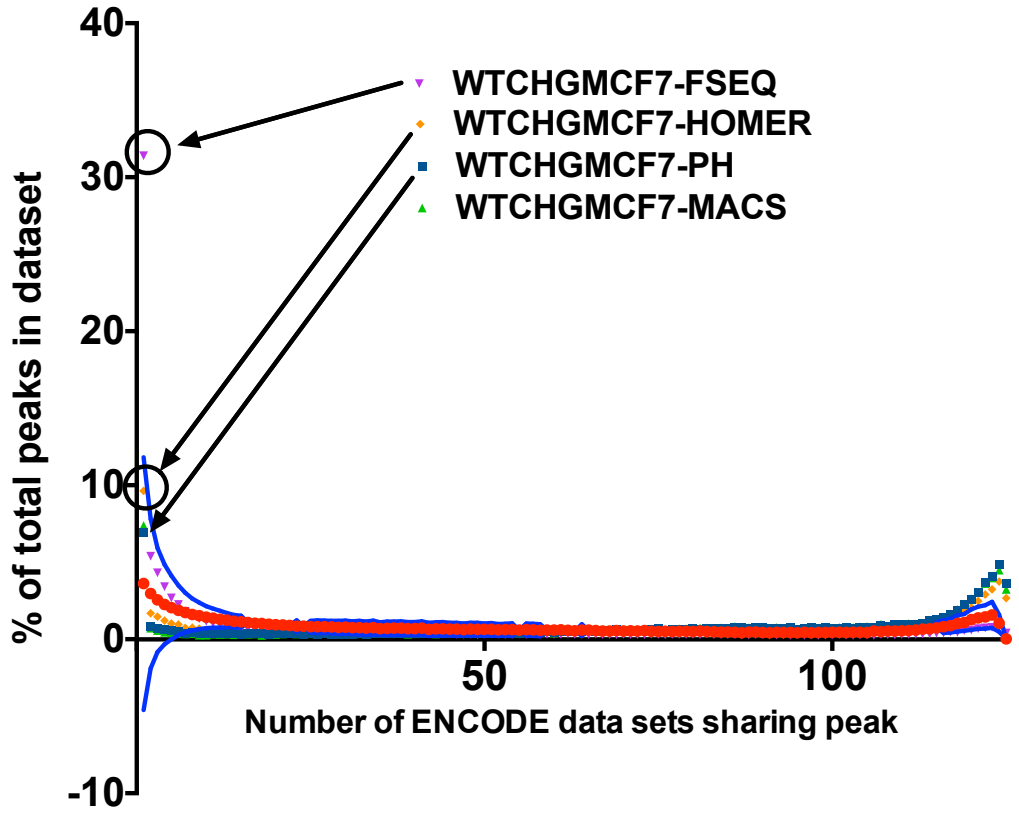

Figure 7.17: Peak calling of the MCF7 DNaseI-seq dataset by four peak callers. To compare the qualitative performance of PeakHunter with other available peak callers, we took our 'most unique' data set, and called peaks in that sample with three other PeakCallers, HOMER, MACS and FSEQ. The sites identified as open chromatin by PeakHunter were more likely to be known open chromatin sites (identified previously by ENCODE) than the sites identifies by HOMER, MACS, or FSEQ. See **Fig 7.16** for a description of the method used.

7.5 Discussion

7.5.1 Continuous sensitivity, signal, noise, and the gold standard

One of the major difficulties with analysis and interpretation of DNaseI-seq data is defining the cut-off between signal and noise. Noise in these experiments may come from a variety of sources, including random DNA sheering, PCR amplification bias, sequencing machine bias

or error, and read mismapping. The distribution of these different sources of noise throughout any given dataset might not be even, and is not easy to measure or model.

Of greater challenge however is genuine DNaseI sensitivity. DNaseI sensitivity is a continuous (as opposed to discrete) characteristic of genomic DNA, that is dependent on the total extent of exposure (duration and concentration) of genomic DNA to the DNaseI enzyme, as well as the 'openness' of the chromatin **(see Fig 7.1a)**. In library preparation, there is no universal, precise and reliable method of determining how much digestion has occurred, and what amount of digestion is optimal to detect DNaseI-hypersensitive sites with minimal background noise. Indeed there is no positive control, or gold standard of what "DNaseI hypersensitivity" is, and no universal functional assay for open chromatin.

Because of these factors, it is both difficult to accurately and reproducibly control the DNaseI digestion, and to measure it accurately prior to sequencing. This leads to analytical challenges, particularly when attempting to compare, in a quantitative sense, datasets with differing signal-noise ratios. Despite this difficulty, the comparison of DNaseI hypersensitivity in cells cultured in different conditions should be a key aim of DNaseI hypersensitivity testing, in order to gain information about the functionality of these sites.

We do not yet have an ideal solution to this problem. Positioning a given signal intensity on the scale of DNaseI sensitivity from hypersensitivity to sensitivity and simply noise, with enough reliability for quantitative comparison, is not straightforward. This is particularly the case considering that estimates from the ENCODE consortium made from large amounts of DNaseI-seq assay data suggest that there are on average ∼200,000 DNaseI hypersensitive sites in any given cell type under any given environmental conditions (304). An ideal solution should be two-fold: improvement of experimental methods to improve data quality, and improved analytical methods to quantify this signal. By finding an answer to the question: "How different from the background is the signal at any given point?", we feel that PeakHunter makes a good attempt at addressing the second question.

7.5.2 Estimates of background signal and threshold selection

A significant part of the problem of distinguishing signal from noise, is identifying the noise. PeakHunter approaches this by random sampling. We have demonstrated that whether sampling from known closed chromatin sites only (defined as closed by the ENCODE data set), or sampling random sites through the genome, very similar estimates of background signal distribution are reached. This is reassuring and useful. For genomes where data sets like that published by ENCODE are not available (i.e. non-human genomes), PeakHunter should still work well, by random background sampling.

The kernel density method also offers the possibility of a statistical, quantitative description of how different from background the signal is at any given read density. A potential problem with this method is posed by very large read densities, where, although it is still possible to calculate a p-value from the background distribution, these results may be less meaningful and accurate, existing so far away from the background distribution. Conversely, PeakHunter should be very powerful in threshold estimation, because good quality p-values will be available for read densities closer to the background distribution. We have demonstrated that even in data sets with low signal-noise ratio, where threshold selection should be more difficult, PeakHunter performs better than other available peak callers, with a greater percentage of identified peaks coming from known open chromatin sites.

7.5.3 Threshold selection and signal estimation: static or dynamic?

For our analysis, we have chosen a global cut-off threshold for peak identification, as opposed to a dynamic threshold. In practice, both approaches are used in published literature (291, 305). Rather than use a dynamic *threshold*, PeakHunter uses dynamic *signal estimation*, in the form of a sampling bin, where the signal in any given bin is adjusted for the local background signal. In the analysis presented here, we chose a bin size of 300 bp, based on the size of open chromatin sites that we might expect to find, and an arbitrarily set

background bin size of 3000 bp. All of these parameters are open to user-adjustment, but we find empirically, in the results presented, that these settings work well. As better quality reference DNaseI-seq data sets, and methods emerge, it may be possible to fine tune these settings further. In principle, smaller bins, and larger background sizes would give more precise results, however, identifying the optimal trade-off between computational speed and precision of results is an ever-present challenge.

7.5.4 Further potential improvements

We believe that PeakHunter achieves its main objectives, which are to provide a set of tools to identify 'peaks' in mapped DNaseI-seq data in a simple, easily accessible manner. While there is not a gold standard for DNaseI-seq data peak identification that would allow us to truly benchmark its performance, PeakHunter does perform well compared with other available peak callers from both quantitative and qualitative perspectives. The task of peak calling would be made more accurate if a reliable method of quantifying DNaseI-digestion accurately prior to sequencing were developed. Higher quality maps of open chromatin that should arise in time will improve the capacity to benchmark the performance of individual peak callers, and thereby improve their performance.

8

Discussion

Contents

8.1 Summary of key experimental findings

In the introduction to this thesis **(Section 1.1.2)**, two specific aims were outlined: to add to the current understanding of the molecular events that regulate NKG2D ligand expression, first, by addressing the hypothesis that Warburg metabolism was of central importance to NKG2D ligand regulation, and second, by pursuing the identification of open chromatin sites at NKG2D ligand gene loci.

In **Chapters 3 & 4**, we presented detailed metabolic experiments showing that glucose availability does determine NKG2D ligand expression, and cellular immunogenicity in an NKG2D-dependent manner, that this effect requires transport of glucose into the cell, metabolism of glucose, and the utilisation of glucose in purine synthetic pathways. Further, we demonstrated that purine nucleosides have the capacity to reproduce these effects on NKG2D ligand expression independently of glucose.

In **Chapter 5**, we provided evidence that these same effects are important in a separate physiological model of NKG2D ligand expression: the infection of primary human fibroblasts

with CMV. Collectively, these results provide evidence to strongly support our hypothesis that the Warburg metabolic transformation that is seen to accompany NKG2D ligand induction in *in vivo* or physiological settings plays a central role in the induction process. This is a novel, previously unreported finding.

In **Chapters 6 and 7**, we identified the open chromatin sites at the NKG2D gene loci in breast cancer cell line, MCF7, developing a quantitative qPCR assay to measure DNaseI digestion prior to DNaseI-seq library generation. We developed an improved analytical software tool, PeakHunter, to identify clustering ('peaks') of reads in mapped DNaseI-seq data. These outcomes, while providing useful outcomes in their own right, provide a solid platform for further exploration of this topic in the future.

8.2 Process and regulation in gene expression

The experimental study of gene regulation provides several challenges. Gene expression is a complex physical *process* **(Fig 1.13)**. All genes share certain definite physical properties, principally surrounding the process of transcription. For example, all genes have proximal and core promoters, transcription start sites, transcription rates and mRNA. Reasonably good methodology exists to identify and describe these common physical properties (306). Both before and after these central aspects gene transcription, the complexity of the gene expression process increases. Not all genes have enhancers, and enhancers can exist anywhere up to 300 kb from the promoter. Transcription factors (or combinations thereof) vary between genes. The number of cytoplasmic and nuclear proteins capable of 'signal transmission' is large. The methodology available for systematically identifying these components is less well defined. For any single gene, the number of potential regulatory factors is large, the likelihood of identifying the 'correct' hypothesis is small, and identifying the methodology to prove or disprove that hypothesis is challenging.

A second challenge relates to identifying the mechanism of *regulation*. When a regulatable

gene is functionally expressed as a protein, the physical processes of gene expression described above (and schematically represented in **Fig 1.13**) may be considered 'active' or 'intact'. A wide range of events may be subtracted from, or added to this process to physiologically switch it on or off. It is possible to understand the physical process of gene expression fully, without understanding how this expression is controlled. By studying only cells where gene expression is 'switched on', it is only possible to learn about the process of expression. To understand how regulation occurs, it is necessary to examine the differences between the physiological 'off' state, and the physiological 'on' state. Identifying useful physiological models is challenging.

8.3 Glucose metabolism and NKG2D ligand regulation

In terms of process and regulation, we developed our hypothesis by looking for potential regulatory factors common to NKG2D ligand expressing cells. It is interesting that NKG2D ligand expression occurs not only in pathological settings (e.g. cancer and viral infection), but also in apparently healthy cells, such as proliferating primary fibroblasts (85), activated proliferating T cells (83) and in gastrointestinal epithelium ((106); observed in *ex vivo* tissue samples) which is in a constant state of renewal through stem cell division. In any setting that cells divide *in vivo*, new nucleotides must be synthesised to replicate the ~3,000,000,000 deoxynucleotide base pairs of the genome, to double the RNA biomass, along with signalling and metabolic nucleotides such as cAMP (cyclic AMP) and NAD (nicotinamide adenine dinucleotide) respectively, and glucose is necessary for this nucleotide synthesis.

Is there any rationale for this process of increased nucleoside synthesis being auto-immunogenic (NKG2D ligand-inducing) in both physiological and pathological situations? Cerboni *et al* demonstrate that activated proliferating NKG2D ligand-expressing $CD8^+$ $\alpha\beta$ T cells and $CD4^+$ $\alpha\beta$ T cells are in fact susceptible to autologous NKG2D-dependent NK cell cytotoxicity, providing evidence from a physiological setting for the hypothesis that NKG2D signalling

may indeed be a self-directed immune response to high energy-biosynthetic 'self' cells where proliferation is excessive, unregulated or disadvantageous for the host ((83); reviewed by Zingoni *et al* (307)). It is a plausible hypothesis that this is the primary physiological role of NKG2D signalling, and that this physiological mechanism is activated when cancer cells acquire this 'high energy-biosynthetic' status through a combination of metabolic dysregulation and darwinian selection, or when viruses 'hi-jack' the cellular capacity to generate this 'high energy-biosynthetic' state to replicate the viral genome. In evolutionary terms, in my view, it is unlikely that an immune mechanism would have evolved to specifically protect against cancer, which is primarily an illness that occurs after reproductive age; it is plausible that such an immune response may have arisen primarily to protect against viral infection, but more probable that this immune response evolved to immunologically oversee the continuous and incessant transitions that occur between quiescent and active, 'high energy-biosynthetic' metabolism that occurs in perfectly healthy cells throughout the body. An example of this is shown by Cerboni *et al*, discussed above.

It is not likely to be the case that *all* cellular, biochemical and genetic features exhibited by cells in these settings would regulate NKG2D ligand expression. However, we have provided evidence suggesting that the availability of glucose to such cells for metabolism *is* central to the expression of NKG2D ligands in a range of cell lines **(Fig 3.2)** and importantly also in a physiological primary cell model **(Fig 5.7)**. It was interesting to observe that while the expression of non-immune proteins, such as PCNA does not change in response to glucose availability **(Fig 3.10)**, the cell surface expression of other key immune proteins such as HLA class I was also increased with higher glucose availability **(Fig 3.8)**. While we were unable to verify glucose-dependent HLA class I expression in the CMV model owing to the HLA class I suppressing immune evasion mechanisms of CMV (308), the results from HEK293T cells suggest that glucose may have wider effects on mammalian cell immunity than we have observed in this thesis. Given the central relationship that exists between cells and glucose, it would be unsurprising that such widespread effects on cellular immunity could have evolved.

It is also of interest that, as we have shown, NKG2D ligand expression relies on purine synthesis **(Figs 4.7 & 5.9)**, a process dependent on, and downstream of, glucose metabolism via the pentose phosphate pathway. While purine and pyrimidine nucleotides and deoxynucleotides play equally important roles in RNA and DNA respectively, in cellular metabolism, purines play a much more important part: cyclic nucleotides are key second messengers (reviewed by Zaccolo and Movsesian (309)), NAD and NADP control intracellular reduction-oxidation balance (310), ATP stores and transports energy through the cell, and other complex purine nucleotides are thought to have signal transmitting properties (311). Our data suggest that phosphorylated nucleosides, rather than nucleosides themselves, are the mediators of NKG2D ligand expression and that this activity occurs independently of contributions to DNA/RNA synthesis (discussed in **Section 4.4.3**). Nucleotide signalling through a ligand-protein mechanism is certainly not unprecedented, a well researched example being AMP activated protein kinase (AMPK) (261). While unpublished data from our laboratory suggests that AMPK is not involved in NKG2D ligand regulation, work on this molecule demonstrates clearly how nucleotides can control intracellular signalling networks.

In summary, the hypothesis that the metabolic transformation described by Warburg leads to the up-regulation of NKG2D ligand expression, through increased glucose transport, metabolism, and purine nucleotide synthesis, is good in explaining NKG2D ligand expression in the settings in which it is known to occur, and we have provided evidence, both in cell lines and a physiological model, that supports this hypothesis directly.

8.4 Open chromatin and NKG2D ligand gene regulation

In **Chapter 6**, we identified the open chromatin elements that are present surrounding the NKG2D ligand genes. In contrast to the work described above, this experimental approach aims to identify the regulatory DNA elements which form part of the physical *processes* necessary for NKG2D ligand expression. In these experiments, the aims were to adapt, optimise

and establish the methodology required to execute these experiments, and to map the open chromatin sites at the NKG2D ligand loci **(Fig 6.8)**, and we achieved both of these aims. MCF7 cells were initially chosen for this experiment on the basis that they represented a malignant cell line, one of the settings in which NKG2D ligand expression is characteristically induced. Additionally, MCF7 cells have been demonstrated in this laboratory to express a higher number of NKG2D ligands at the cell surface compared with other cell lines (unpublished). In this respect, the enhancers, proximal promoters and core promoters necessary for gene expression can be expected to be open and detectable.

While the signal-noise ratio was low in the MCF7 DNaseI-seq data set, a key outcome of this was the development of a qPCR assay for assessing the amount digestion in a given sample, which will allow us to avoid low signal-noise data sets in future. Fortunately, PeakHunter allowed us to overcome much of this limitation, by identifying 37,809 peaks (significant clustering of mapped reads) at an empirical false discovery rate of 0.01% **(Table 7.3)**. The accurate identification of open chromatin sites through DNaseI-seq in the future is likely to depend on improvements in library preparation, and in statistical data analysis, and here, we have made significant contributions to both parts of this strategy. Having our own software is an advantage, as it allows us to understand the analysis that we apply to our data, to fine-tune the programming as needed, and to have confidence in the output. We hope that other researchers also find this software useful in the future.

8.5 Context and existing research

Glucose and purine metabolism. In **Section 1.2.6**, we outlined current understanding of how NKG2D ligand expression is regulated. Of the pre-transcriptional mechanisms discussed **(Section 1.2.6.2)**, our data is consistent with a mechanism of NKG2D ligand regulation acting down-stream of TLR stimulation and 'growth factor/activation-driven' NKG2D ligand expression. In both settings, the induction of Warburg metabolism is known to occur **(Table**

1.3). The observation that histone deacetylase inhibitors induce NKG2D ligand expression is not inconsistent with our hypotheses or data, but the physiological settings in which histone deacetylase inhibiting activity is regulated in a directed manner are unclear. In my view, activation of DNA repair pathways is a good hypothesis to explain the regulation of NKG2D ligand expression in a range of physiological settings. However, the existing published data to support that hypothesis is not very strong, and our data **(Fig 4.26)** suggests that caffeine, an ATM/ATR inhibitor that forms the basis of much of the evidence for the DNA damage pathway signalling hypothesis, inhibits purine synthesis as well as ATM/ATR signalling, and doesn't block NKG2D ligand expression in all circumstances in our HEK293T cell model.

The available research describing peri-transcriptional regulatory events **(Section 1.2.6.3)** are all plausibly consistent with our hypothesis, irrespective of whether glucose and purines act at a predominantly pre- or post-transcriptional level. In terms of published post-transcriptional mechanisms of regulation **(Section 1.2.6.4)**, our data is inconsistent with the hypothesis that post transcriptional glycosylation of NKG2D ligand proteins controls their expression in physiological settings, and specifically in our model, as we have observed that total MICA protein expression increases with glucose relative to PCNA protein in permeabilised cells **(Fig 3.10)**. This makes physiological regulation of cell surface NKG2D ligand expression unlikely to be distributional. It is possible, in view of our results, that glucose and purine levels act at a post-transcriptional stage to regulate NKG2D ligand expression through control of miRNA expression or proteasomal activity. It is also plausible that all three mechanisms may influence NKG2D ligand expression independently.

On balance, our hypothesis that the induction of Warburg metabolism in these settings, leading to increased glucose uptake, and increased glucose-dependent purine nucleotide synthesis, plays a causal role in the regulation of NKG2D ligand induction explains NKG2D ligand expression in all the physiological situations in which NKG2D ligand expression is known to occur. The evidence we have presented supports this hypothesis, as is consistent with, or at least as strong as, the evidence available to support alternative hypotheses.

Open chromatin sites. The DNaseI-seq derived evidence for the open chromatin sites identified at the NKG2D ligand loci in MCF7 breast cancer cells is consistent with subsequently generated data sets published by the ENCODE consortium. ChIP-seq derived transcription factor binding data, also from the ENCODE consortium, independently supports the assertion that the sites we have identified are open chromatin, and transcription factor binding loci **(Figs 1.6, 1.7 & 1.8)**. We have shown that PeakHunter performs qualitatively and quantitatively (number of peaks identified) as well as other currently available peak calling programs in identifying open chromatin sites. PeakHunter outperforms other available peak callers in low-signal noise data sets, where a greater percentage of peaks identified by PeakHunter were at known open chromatin sites **(Fig 7.17)**. In this context, PeakHunter is a valuable addition to the available peak calling software repertoire.

9

Appendix: PeakHunter code

```
#!/usr/bin/perl
use strict;
if ((@ARGV==0) || ($ARGV[0] eq "-h"))
        {
                usage();
                die"\n"
        }

if ($ARGV[0] eq "-v")
        {
                die "\nPeakHunter v2.0; May 2013\n\n";
        }

my $genomeSizeFile = 0;
my $genomeBaseCount = 0;
my @lineArray = ();
my $lineCount = 0;
my $totalArea = 0;
my $count = 0;
my $currentStep = 0;
my @wordArray = ();
use POSIX 'acos';
use POSIX 'fmod';
#stamp package:
use Cwd;
my $dir = getcwd;
my $user = getlogin();
my @timeData = split(" ", localtime(time));
my $timeString = join(" ", @timeData);
my $progLoc = $0;
my $command = join(" ", @ARGV);
my $sys = $^O;
print STDERR "\nProgram: PeakHunter ($progLoc)\nVersion: v2.0\n\nstart time: $timeString\nuser: $user\ncwd: $dir\
    nparameters: $command\nsystem: $sys\n\n";
#end stamp package
##default values for density analyzer and peak caller, subject to change by user command line input
my $stepSize = 100;
my $peakBinSize = 300;
my $backgroundBinSize = 3000;
my $tail = $peakBinSize/2;
our $sigma = 50;
my $thresh = "DEFAULT";
my $offSet = 0;
my $maxThresh = 100000000;
my $bandWidth = $tail*2 +1;
my $binSize = $tail*2 +1;
my $thresholdSignificanceLevel = 0.001;
my $peaksForBackgroundCalc = 50000;
our %densityFrequencyHash;
my %pValueTable = ();
my $pcrDuplicates = "OFF";
my $closedAndOpen = "NO";
my $maxPeakPValue = "OFF";
my $mapqThresh = 0;
my $uqThresh = 10000;
my $peakStringency = 1;
my $powerDigit = 5;
my $numberDataSetsPerPeak = 60;

#end default values
my $peakThresholdValue = 0;
# check to ensure that an even number of items exist on the command line before converting @ARGV to a hash
die "ERROR: uneven option-value pairings.\n\n" if @ARGV%2 != 0;
my %arguments = @ARGV;
my @arguments = ();

## checks to make sure that only one function is selected
my $processCount = 0;
$processCount += 1 if exists $arguments{"-D"};
$processCount += 1 if exists $arguments{"-F"};
$processCount += 1 if exists $arguments{"-P"};
$processCount += 1 if exists $arguments{"-R"};
$processCount += 1 if exists $arguments{"-S"};
$processCount += 1 if exists $arguments{"-T"};
$processCount += 1 if exists $arguments{"-NS"};
```

```
$processCount += 1 if exists $arguments{"-RS"};
$processCount += 1 if exists $arguments{"-EFDR"};

die "too many processes selected.\n\n" if $processCount > 1;
die "no processes selected.\n\n" if $processCount < 1;
#################################################################

##### runs the random read sam selection option
my @randomTitleCheck = ();
my $samForRandom = 0;
my $randomNumber = 0;

if (exists $arguments{"-R"})
        {
                #check correct file extension
                $samForRandom = $arguments{"-R"}; delete($arguments{"-R"});
                die "Input file (set to \'$samForRandom\') not found.\n\n" unless -e $samForRandom;
                my @samTestResults = fileTest('SAM', $samForRandom);
                print STDERR "\n\tINPUT file format test report \'SAM\': $samForRandom\n\n\t\tSAM format ... "
                      $samTestResults[0]\n\t\tSAM read order ... $samTestResults[1]\n\t\tSAM chr order ... "
                      $samTestResults[3]\n\t\tSAM header ... $samTestResults[2]\n\n";
                die "sam file format for inputfile (\'$samForRandom\') not recognised.\n\n"      if $samTestResults
                      [0] eq 'BAD';
                print STDERR "warning: sam file reads for inputfile (\'$samForRandom\') appear disordered.\n\n" if
                      $samTestResults[1] eq 'BAD';
                print STDERR "warning: sam file chromosomes for inputfile (\'$samForRandom\') appear disordered.\n\
                      n"    if $samTestResults[3] eq 'BAD';
                #assign variables and delete from %arguments hash
                $randomNumber = $arguments{"-nr"} if exists $arguments{"-nr"}; delete($arguments{"-nr"});
                #checks for and kills if extra command line arguments
                @arguments = %arguments;
                die "unused command line arguments: @arguments\n\n" if keys %arguments > 0;
                #sets limits for assigned parameters
                die "random number of reads ($randomNumber; -nr) must be a positive integer. \n\n" unless ((
                      $randomNumber =~ /[0-9]/) && (fmod($randomNumber,1) == 0) && ($randomNumber > 0));
                #executes subroutine
                print STDERR "Random Read Selector v1.0\n\nwith parameters:\n\tinput file: $samForRandom\n\ttarget
                      number of reads: $randomNumber\n\n";
                randomReadSelection($samForRandom, $randomNumber);#<- need to write subroutine
        }

# runs the top peaks selection option
my @topPeaksTitleCheck = ();
my $bedForTop = 0;
my $topPeaksNumber = 'ALL';

if (exists $arguments{"-T"})
        {
                $bedForTop = $arguments{"-T"}; delete($arguments{"-T"});
                #check correct file extension
                die "Input file (set to \'$bedForTop\') not found.\n\n" unless -e $bedForTop;
                my @bedTestResults = fileTest('BED', $bedForTop);
                print STDERR "\n\tINPUT file format test report \'BED\': $bedForTop\n\n\t\tBED format ... "
                      $bedTestResults[0]\n\n";
                die "bed file format for inputfile (\'$bedForTop\') not recognised.\n\n"         if $bedTestResults
                      [0] eq 'BAD';
                #assign variables and delete from %arguments hash
                $topPeaksNumber = $arguments{"-n"} if exists $arguments{"-n"}; delete($arguments{"-n"});
                #checks for and kills if extra command line arguments
                @arguments = %arguments;
                die "unused command line arguments: @arguments\n\n" if keys %arguments > 0;
                #sets limits for assigned parameters()
                die "number of peaks (set to \'$$topPeaksNumber\') must be either 'ALL' or a positive integer. \n\n
                      " unless (((($topPeaksNumber =~ /[0-9]/) & (fmod($topPeaksNumber,1) == 0) && ($topPeaksNumber
                      >= 0)) ||  ($topPeaksNumber eq 'ALL'));

                #executes subroutine
                print STDERR "Top Peak Selector v1.0\n\nwith parameters:\n\tinput file: $bedForTop\n\ttarget number
                      of reads: $topPeaksNumber\n\n";
                topPeaks($bedForTop, $topPeaksNumber);
        }

##### runs the sam numerical sort option
my @samNSortTitleCheck = ();
my $samForNSort = 0;

if (exists $arguments{"-NS"})
        {
                $samForNSort = $arguments{"-NS"}; delete($arguments{"-NS"});
                die "Input file (set to \'$samForNSort\') not found.\n\n" unless -e $samForNSort;
                my @samTestResults = fileTest('SAM', $samForNSort);
                print STDERR "\n\tINPUT file format test report \'SAM\': $samForNSort\n\n\t\tSAM format ... "
                      $samTestResults[0]\n\t\tSAM read order ... $samTestResults[1]\n\t\tSAM chr order ... "
                      $samTestResults[3]\n\t\tSAM header ... $samTestResults[2]\n\n";
                die "sam file format for inputfile (\'$samForNSort\') not recognised.\n\n"       if $samTestResults
                      [0] eq 'BAD';
                #checks for and kills if extra command line arguments
                @arguments = %arguments;
                die "unused command line arguments: @arguments\n\n" if keys %arguments > 0;
                #executes subroutine
                print STDERR "Numerical Sam Sort v1.0\n\nwith parameters:\n\tinput file: $samForNSort\n\n";
                samNSort($samForNSort);
        }
##### runs the sam filter option
my @filterTitleCheck = ();
my @headerTitleCheck = ();
my $samForFilter = 0;
my $samForHeader = 0;
if (exists $arguments{"-F"})
        {
                $samForFilter = $arguments{"-F"}; delete($arguments{"-F"});
                die "Input file (set to \'$samForFilter\') not found.\n\n" unless -e $samForFilter;
```

```
                my @samTestResults = fileTest('SAM', $samForFilter);
                print STDERR "\n\tINPUT file format test report \'SAM\': $samForFilter\n\n\t\tSAM format .... 
                    $samTestResults[0]\n\t\tSAM read order .... $samTestResults[1]\n\t\tSAM chr order .... 
                    $samTestResults[3]\n\t\tSAM header .... $samTestResults[2]\n\n";
                die "sam file format for inputfile (\'$samForFilter\') not recognised.\n\n"      if $samTestResults
                    [0] eq 'BAD';
                die "sam file reads for inputfile (\'$samForFilter\') appear disordered.\n\n"    if $samTestResults
                    [1] eq 'BAD';
                print STDERR "warning: sam file chromosomes for inputfile (\'$samForFilter\') appear disordered.\n\
                    n"    if $samTestResults[3] eq 'BAD';

                die "genome size file (-g) must be set.\n\n" unless exists $arguments{"-g"};

                # checks that -g has been set (mandatory)
                $genomeSizeFile = $arguments{"-g"} if exists $arguments{"-g"}; delete($arguments{"-g"});
                die "chromosome size file (set to \'$genomeSizeFile\') required but not found.\n\n" unless ((
                    defined $genomeSizeFile) && (-e $genomeSizeFile));
                my @genomeTestResults = fileTest('CHROM', $genomeSizeFile);
                print STDERR "\n\tchromosome size file format test report \'CHROM\': $genomeSizeFile\n\n\t\tCHROM 
                    format .... $genomeTestResults[0]\n\n";
                die "chromosome size file format for inputfile (\'$genomeSizeFile\') not recognised.\n\n"       if
                    $genomeTestResults[0] eq 'BAD';

                #assign target file name
                $samForFilter = $arguments{"-F"}; delete($arguments{"-F"});
                $samForHeader = $samForFilter unless exists $arguments{"-i"};
                $samForHeader = $arguments{"-i"} if exists $arguments{"-i"};

                if (exists $arguments{"-i"})
                        {
                                die "$samForHeader (-i) not found.\n\n" unless -e $samForHeader;
                                my @samTestHeaderResults = fileTest('SAM', $samForHeader);
                                print STDERR "\n\tINPUT file format test report \'SAM\': $samForHeader\n\n\t\tSAM 
                                    format .... $samTestHeaderResults[0]\n\t\tSAM read order .... 
                                    $samTestHeaderResults[1]\n\t\tSAM chr order .... $samTestHeaderResults[3]\n\t\t
                                    SAM header .... $samTestHeaderResults[2]\n\n";
                                die "sam file header for inputfile (\'$samForHeader\') is not recognised.\n\n"  if
                                    $samTestHeaderResults[2] eq 'NONE';
                                delete($arguments{"-i"});
                        }
                #assign variables and delete from %arguments hash
                $mapqThresh = $arguments{"-q"} if exists $arguments{"-q"}; delete($arguments{"-q"});
                $uqThresh = $arguments{"-u"} if exists $arguments{"-u"}; delete($arguments{"-u"});
                $pcrDuplicates = $arguments{"-PCR"} if exists $arguments{"-PCR"}; delete($arguments{"-PCR"});

                #checks for and kills if extra command line arguments
                @arguments = %arguments;
                die "unused command line arguments: @arguments\n\n" if keys %arguments > 0;
                #sets limits for assigned parameters
                die "mapq threshold (set to \'$mapqThresh\') must be a non-negative integer.\n\n" unless ((
                    $mapqThresh =~ /[0-9]/) && (fmod($mapqThresh,1) == 0) && ($mapqThresh >= 0));
                die "uq threshold (set to \'$uqThresh\') must be a non-negative integer.\n\n" unless (($uqThresh =~
                     /[0-9]/) && (fmod($uqThresh,1) == 0) && ($uqThresh >= 0));
                #die "$samForHeader (-i) not found.\n\n" if !-e $samForHeader;
                die "PCR duplicates (set to \'$pcrDuplicates\') can only have the values ON (remove PCR duplicates)
                    or OFF (retain PCR duplicates).\n\n" if (($pcrDuplicates ne "OFF") && ($pcrDuplicates ne "ON"
                    ));
                #executes subroutine
                print STDERR "Sam Filter v1.0.\n\nwith parameters:\n\tinput file: $samForFilter\n\theader file: 
                    $samForHeader\n\tmapq threshold: $mapqThresh\n\tuq threshold: $uqThresh\n\tPCR duplicate 
                    removal: $pcrDuplicates\n\n";
                samFilter($samForHeader, $samForFilter, $mapqThresh, $uqThresh);
        }

##### runs the density reader option
my $samForDensity = 0;
my @samTitleCheck = ();

if (exists $arguments{"-D"})
        {
                $samForDensity = $arguments{"-D"}; delete($arguments{"-D"});
                die "Input file (set to \'$samForDensity\') not found.\n\n" unless -e $samForDensity;

                my @samTestResults = fileTest('SAM', $samForDensity);
                print STDERR "\n\tINPUT file format test report \'SAM\': $samForDensity\n\n\t\tSAM format .... 
                    $samTestResults[0]\n\t\tSAM read order .... $samTestResults[1]\n\t\tSAM chr order .... 
                    $samTestResults[3]\n\t\tSAM header .... $samTestResults[2]\n\n";
                die "sam file format for inputfile (\'$samForDensity\') not recognised.\n\n"     if $samTestResults
                    [0] eq 'BAD';
                die "sam file reads for inputfile (\'$samForDensity\') appear disordered.\n\n"  if $samTestResults
                    [1] eq 'BAD';
                die "sam file chromosomes for inputfile (\'$samForDensity\') appear disordered.\n\n"    if
                    $samTestResults[3] eq 'BAD';
                #check correct file extension
                # checks that -g has been set (mandatory)
                die "genome size file (-g) must be set.\n\n" unless exists $arguments{"-g"};
                $genomeSizeFile = $arguments{"-g"} if exists $arguments{"-g"}; delete($arguments{"-g"});
                die "chromosome size file (set to \'$genomeSizeFile\') required but not found.\n\n" unless ((
                    defined $genomeSizeFile) && (-e $genomeSizeFile));
                my @genomeTestResults = fileTest('CHROM', $genomeSizeFile);
                print STDERR "\n\tchromosome size file format test report \'CHROM\': $genomeSizeFile\n\n\t\tCHROM 
                    format .... $genomeTestResults[0]\n\n";
                die "chromosome size file format for inputfile (\'$genomeSizeFile\') not recognised.\n\n"       if
                    $genomeTestResults[0] eq 'BAD';

                #assign variables and delete from %arguments hash
                $thresh = $arguments{"-t"} if exists $arguments{"-t"}; delete($arguments{"-t"});
                $maxThresh = $arguments{"-m"} if exists $arguments{"-m"}; delete($arguments{"-m"});
                $tail = $arguments{"-n"} if exists $arguments{"-n"}; delete($arguments{"-n"});
                $sigma = $arguments{"-d"} if exists $arguments{"-d"}; delete($arguments{"-d"});
                $stepSize = $arguments{"-STEP"} if exists $arguments{"-STEP"}; delete($arguments{"-STEP"});
                $offSet = $arguments{"-o"} if exists $arguments{"-o"}; delete($arguments{"-o"});
```

```
		$genomeSizeFile = $arguments{"-g"} if exists $arguments{"-g"}; delete($arguments{"-g"});

		#checks for and kills if extra command line arguments
		@arguments = %arguments;
		die "unused command line arguments: @arguments\n\n" if keys %arguments > 0;

		die "minimum bin threshold (set to \'$thresh\') be 'DEFAULT' or a non-negative integer.\n\n" unless
		      ((($thresh =~ /[0-9]/) && (fmod($thresh,1) == 0) && ($thresh >= 0)) || ($thresh eq 'DEFAULT')
		      );
		die "maximum bin threshold (set to \'$maxThresh\') be a positive integer.\n\n" unless (($maxThresh
		      =~ /[0-9]/) && (fmod($maxThresh,1) == 0) && ($maxThresh > 0));
		die "tailSize must (set to \'$tail\') be a positive integer.\n\n" unless (($tail =~ /[0-9]/) && (
		      fmod($tail,1) == 0) && ($tail > 0));
		die "sigma must (set to \'$sigma\') be a positive integer.\n\n" unless (($sigma =~ /[0-9]/) && (
		      fmod($sigma,1) == 0) && ($sigma > 0));
		die "step size must (set to \'$stepSize\') be a positive integer.\n\n" unless (($stepSize =~
		      /[0-9]/) && (fmod($stepSize,1) == 0) && ($stepSize > 0));
		die "offset must (set to \'$offSet\') be an integer.\n\n" unless (($offSet =~ /[0-9]/) && (fmod(
		      $offSet,1) == 0));

		#executes subroutine
		print STDERR "Density Analyzer v1.0.\n\nwith parameters:\n\tinput file: $samForDensity\n\tstep
		      size: $stepSize\n\ttail length: $tail\n\tsigma: $sigma\n\tminimum bin threshold: $thresh\n\
		      tmaximum bin threshold: $maxThresh\n\ttrack offset: $offSet\n\tbin size: $binSize\n\n";
		densityAnalyzer($samForDensity, $stepSize, $tail, $sigma, $thresh, $offSet, $maxThresh);
	}

##### runs the peakCaller option
my $bluePrintFile = "NONE";
my $samForReport = 0;
my $samRPBScore = 0;
@samTitleCheck = ();
my @blueTitleCheck = ();
our $flatThreshold = "NONE";

if (exists $arguments{"-P"})
	{
		#### NB thresholdSignificanceLevel to build in
		#check correct file extension
		$samForReport = $arguments{"-P"}; delete($arguments{"-P"});
		die "Input file (set to \'$samForReport\') not found.\n\n" unless -e $samForReport;
		my @samTestResults = fileTest('SAM', $samForReport);
		print STDERR "\n\tINPUT file format test report \'SAM\': $samForReport\n\n\t\tSAM format .... 
		      $samTestResults[0]\n\t\tSAM read order .... $samTestResults[1]\n\t\tSAM chr order .... 
		      $samTestResults[3]\n\t\tSAM header .... $samTestResults[2]\n\n";
		die "sam file format for inputfile (\'$samForReport\') not recognised.\n\n"      if $samTestResults
		      [0] eq 'BAD';
		die "sam file reads for inputfile (\'$samForReport\') appear disordered.\n\n"    if $samTestResults
		      [1] eq 'BAD';
		die "sam file chromosomes for inputfile (\'$samForReport\') appear disordered.\n\n"      if
		      $samTestResults[3] eq 'BAD';
		$closedAndOpen = "YES" if exists $arguments{"-b"};
		$bluePrintFile = $arguments{"-b"} if exists $arguments{"-b"};

		if (exists $arguments{"-b"})
			{
				die "Blueprint file (set to \'$bluePrintFile\') not found.\n\n" unless -e
				      $bluePrintFile;
				my @bedTestResults = fileTest('BED', $bluePrintFile);
				print STDERR "\n\tBLUEPRINT file format test report \'BED\': $bluePrintFile\n\n\t\
				      tBED format .... $bedTestResults[0]\n\n";
				die "bed file format for inputfile (\'$bluePrintFile\') not recognised.\n\n"     if
				      $bedTestResults[0] eq 'BAD';
				delete($arguments{"-b"});
			}

		# checks that -g has been set (mandatory)
		die "genome size file (-g) must be set.\n\n" unless exists $arguments{"-g"};
		$genomeSizeFile = $arguments{"-g"} if exists $arguments{"-g"}; delete($arguments{"-g"});
		die "chromosome size file (set to \'$genomeSizeFile\') required but not found.\n\n" unless ((
		      defined $genomeSizeFile) && (-e $genomeSizeFile));
		my @genomeTestResults = fileTest('CHROM', $genomeSizeFile);
		print STDERR "\n\tchromosome size file format test report \'CHROM\': $genomeSizeFile\n\n\t\tCHROM 
		      format .... $genomeTestResults[0]\n\n";
		die "chromosome size file format for inputfile (\'$genomeSizeFile\') not recognised.\n\n"        if
		      $genomeTestResults[0] eq 'BAD';

		#assign variables and delete from %arguments hash
		#$bluePrintFile = $arguments{"-b"} if exists $arguments{"-b"}; delete($arguments{"-b"});
		$peakBinSize = $arguments{"-bin"} if exists $arguments{"-bin"}; delete($arguments{"-bin"});
		$backgroundBinSize = $arguments{"-back"} if exists $arguments{"-back"}; delete($arguments{"-back"})
		      ;
		$peaksForBackgroundCalc = $arguments{"-npBack"} if exists $arguments{"-npBack"}; delete($arguments{
		      "-npBack"});
		$thresholdSignificanceLevel = $arguments{"-sig"} if exists $arguments{"-sig"}; delete($arguments{"-
		      sig"});
		#$genomeSizeFile = $arguments{"-g"} if exists $arguments{"-g"}; delete($arguments{"-g"});
		#$samForReport = $arguments{"-s"} if exists $arguments{"-s"}; delete($arguments{"-s"});
		$maxPeakPValue = $arguments{"-PVAL"} if exists $arguments{"-PVAL"}; delete($arguments{"-PVAL"});
		$flatThreshold = $arguments{"-FLAT"} if exists $arguments{"-FLAT"}; delete($arguments{"-FLAT"});
		$stepSize = $arguments{"-STEP"} if exists $arguments{"-STEP"}; delete($arguments{"-STEP"});

		#checks for and kills if extra command line arguments
		@arguments = %arguments;
		die "unused command line arguments: @arguments\n\n" if keys %arguments > 0;

		#sets limits for assigned parameters
		die "peak bin size (set to \'$peakBinSize\') must be a positive integer.\n\n" unless (($peakBinSize
		      !~ /[^0-9\.]/) && (fmod($peakBinSize,1) == 0) && ($peakBinSize > 0));
		die "background bin size (set to \'$backgroundBinSize\') must be a positive integer.\n\n" unless ((
		      $backgroundBinSize !~ /[^0-9\.]/) && (fmod($backgroundBinSize,1) == 0) && ($backgroundBinSize
		      > 0));
```

```
                die "step size (set to \'$stepSize\') must be a positive integer.\n\n" unless (($stepSize !~
                    /[^0-9\.]/) && (fmod($stepSize,1) == 0) && ($stepSize > 0));
                die "number of peaks for background sampling (set to \'$peaksForBackgroundCalc\') must be a
                    positive integer.\n\n" unless (($peaksForBackgroundCalc !~ /[^0-9\.]/) && (fmod(
                    $peaksForBackgroundCalc,1) == 0) && ($peaksForBackgroundCalc > 0));
                die "flat threshold (set to \'$flatThreshold\') must be 'NONE' or a non-negative number.\n\n"
                    unless ((($flatThreshold !~ /[^0-9\.]/) && ($flatThreshold >= 0)) || ($flatThreshold eq 'NONE'
                    ));
                die "threshold significance level (set to \'$thresholdSignificanceLevel\') must be in the range [0
                    ..1].\n\n" unless (($thresholdSignificanceLevel >= 0) && ($thresholdSignificanceLevel <= 1));

                die "background bin size (set to \'$backgroundBinSize\') must be greater than peak bin size (set to
                    \'$peakBinSize\').\n\n" if $backgroundBinSize < $peakBinSize;
                die "peak score p values (-PVAL; set to \'$maxPeakPValue\') must be either ON (for p values) or OFF
                    (for maximum density score)\n\n" if (($maxPeakPValue ne "OFF") && ($maxPeakPValue ne "ON"));

                #executes subroutine
                print STDERR "Peak Caller 2.0.\n\nwith parameters:\n\tinput sam file: $samForReport\n\tblueprint
                    file: $bluePrintFile\n\tbin size/background size: $peakBinSize/$backgroundBinSize\n\tpeak
                    target for threshold/binscore assessment: $peaksForBackgroundCalc\n\tcalculated peak threshold
                    significance: $thresholdSignificanceLevel\n\tstep size: $stepSize\n\tflat threshold:
                    $flatThreshold\n\tpeak score p values: $maxPeakPValue\n\n";
                #$samRPBScore = readsPerBase($samForReport) unless (($samRPBScore > 0) || ($flatThreshold > 0));
                $peakThresholdValue = threshold($samForReport, $bluePrintFile, $peakBinSize, $backgroundBinSize,
                    $peaksForBackgroundCalc); #if $flatThreshold eq "NONE";### <- need to add powerDigit to
                    subroutine
                $peakThresholdValue = $flatThreshold if $flatThreshold ne "NONE"; ### <- need to test this
                peakCall($samForReport, $peakThresholdValue, $peakBinSize, $backgroundBinSize);
        }

##### runs the peak average signal option
my $bedForSignal = 0;
my @bedTitleCheck = ();
my $distanceFromCentre = 0;
my $numberOfRandomPeaks = 0;
my $samForSignal = 0;

if (exists $arguments{"-S"})
        {
                #check correct file extension
                @bedTitleCheck = split(/\./, $arguments{-S});
                @samTitleCheck  = split(/\./, $arguments{-t});
                die "Mean Signal Intensity (-S) requires *.bed file input\n\n" if $bedTitleCheck[-1] ne "bed";
                die "Mean Signal Intensity (-t) requires target *.sam file input\n\n" if $samTitleCheck[-1] ne "sam
                    ";
                #assign targe file name
                $bedForSignal = $arguments{"-S"}; delete($arguments{"-S"});
                die "$bedForSignal (-S) not found.\n\n" if !-e $bedForSignal;
                #assign variables and delete from %arguments hash
                $distanceFromCentre = $arguments{"-dc"} if exists $arguments{"-dc"}; delete($arguments{"-dc"});
                $numberOfRandomPeaks = $arguments{"-np"} if exists $arguments{"-np"}; delete($arguments{"-np"});
                $samForSignal = $arguments{"-t"} if exists $arguments{"-t"}; delete($arguments{"-t"});
                die "$samForSignal (-t) not found.\n\n" if !-e $samForSignal;
                #checks for and kills if extra command line arguments
                @arguments = %arguments;
                die "unused command line arguments: @arguments\n\n" if keys %arguments > 0;
                #sets limits for assigned parameters
                die "distance from peak centre (-dc) out of range.\n\n" if $distanceFromCentre == 0;
                die "number of random peaks (-np) out of range.\n\n" if $numberOfRandomPeaks < 50;
                #executes subroutine
                print STDERR "Mean Signal Intensity v1.0\nwith parameters:\n\tinput peak file: $bedForSignal\n\
                    ttarget sam file: $samForSignal\n\tdistance from peak centre: $distanceFromCentre\n\tnumber of
                    peaks included: ~$numberOfRandomPeaks\n\n";
                peakSignalIntensity($bedForSignal, $distanceFromCentre, $numberOfRandomPeaks, $samForSignal); ###
                    <- need to write subroutine
        }

my $RSname = 0;
my $RSnumber = 0;

############################## sub routines ################################

sub readCount
        {
                my $RCtarget = $_[0];
                my $RCcount = 0;
                my $buffer = 0;
                my @RCarray = ();
                my $RCheaderCount = 0;
                open (TARGET, "<", $RCtarget);
                while (<TARGET>)
                        {
                                @RCarray = split ("", $_);
                                $RCheaderCount += 1 if $RCarray[0] eq "@";
                                last if $RCarray[0] ne "@";
                        }
                close TARGET;
                open (my $targetHandle, "<", $RCtarget);
                while(sysread $targetHandle, $buffer, 4096)
                        {
                                $RCcount += ($buffer =~ tr/\n//);
                        }
                close $targetHandle;
                $RCcount = $RCcount - $RCheaderCount;
                return $RCcount;
        }

sub randomSam
        {
                my $outputfile = $_[0];
                $outputfile = "$outputfile" . ".sam";
                my $targetReadNumber = $_[1];
```

```
                my $chromosomeSizeFile = $_[2];
                my $index = 0;
                my $x = 0;
                my $range = 0;
                my @line = ();
                my $key = 0;
                my $readName = 0;
                my $randomReadNumber = 0;
                my @chromName = ();
                my @chromSize = ();
                my $chromCount = 0;
                my $end = 0;
                my $start = 0;
                my $total = 0;
                my $readStart = 0;
                my @chromStart = ();
                my @chromEnd = ();
                my $chrom = 0;
                my %indexHash = ();
                my $multipleCount = 0;
                open (CHROMS, "<", "$chromosomeSizeFile");
                while (<CHROMS>)
                        {
                                @line = split("\t", $_);
                                $end = $line[1] + $total;
                                $start =  1 + $total;
                                push (@chromName, $line[0]);
                                push (@chromStart, $start);
                                push (@chromEnd, $end);
                                $total = $total + $line[1];
                                $chromCount += 1;
                        }
                close CHROMS;
                $range = $total - 1;
                while ($randomReadNumber < $targetReadNumber)
                        {
                                $index = int(rand($range)) + 1;
                                $indexHash{$index} += 1;
                                $randomReadNumber+=1;
                        }

                print STDERR "\t$targetReadNumber\treads\tselected...\n";
                $randomReadNumber = 0;
                open (OUTPUT, ">", $outputfile);
                for ($x=0; $x<$total; $x +=1)
                        {
                                if (exists $indexHash{$x})
                                        {
                                while ($multipleCount < $indexHash{$x})
                                        {
                                                print OUTPUT "#rand\t0\t$chrom\t$readStart\t37\t37M\t*\t0\t0\
                                                    tRANDOM_Seq\tRANDOM_Qual\tUQ:i:0\tNM:i:0\n" if exists
                                                    $indexHash{$x};
                                                $multipleCount += 1;
                                        }
                                $multipleCount = 0;
                        }
                                print STDERR "$x\tbases\tcovered\n" if $x%100000000 == 0;
                        }
                close OUTPUT;
                print STDERR "\t$outputfile\tfile\tcreated.\n\n";
        }

sub threshold
        {
                #print STDERR "\n\n***in sub threshold***\n\n";
                my $samForAnalysis = $_[0];
                my $bluePrint = $_[1];
                my $halfRegionSize = ($_[2])/2;
                my $halfBackgroundSize = ($_[3])/2;
                my $numberOfPeaks = $_[4]; # the number of sites used to calculate p value: default 10000; probably
                    won't give option to change this.
                my @wholeBlue = ();
                my $index = 0;
                my $open = 0;
                my $closed = 0;
                my $x = 0;
                my $range = 0;
                my %closedChrom = ();
                my %openChrom = ();
                my @closedChromArray = ();
                my @openChromArray = ();
                my @line = ();
                my $siteLine = 0;
                my @globalSelection = ();
                my $key = 0;
                my $count = 0;
                my $subRange = 0;
                my $readName = 0;
                my $thresholdFromThreshold = 0;
                my $start = 0;
                my $end = 0;
                my @chromLine = ();
                my @chromName = ();
                my @chromStart = ();
                my @chromEnd = ();
                my $total = 0;
                my $chromCount = 0;
                my $randomPeakNumber = 0;
                my $randomPeakStart = 0;
                my $randomPeakEnd = 0;
                my $randomPeakString = 0;
```

```
my $chrom = 0;
my @randomChromArray = ();
my $randomRef = 0;
my %randomChromHash = ();
my $randomStringData = 0;
my @random = ();
my $randomReadsPerBin = ();
my $randomPeaksCompleted = ();
my $openRef = 0;
my $openStringData = 0;
my @open = ();
my $openReadsPerBin = ();
my $openPeaksAssessed = ();
my $closedRef = 0;
my $closedStringData = 0;
my @closed = ();
my $closedReadsPerBin = ();
my $closedPeaksCompleted = ();

# make arrays containg $numberOfPeaks bed lines from target categories
my $closedCount = 0;
my $openCount = 0;
my $sum = 0;
my @blueLine = ();
#need to write in: if blueprint exists then this: else random set skip to 70 lines.
if ($closedAndOpen eq "YES")
        {
                open (BLUE, "<", "$bluePrint");
                while (<BLUE>)
                        {
                                push (@wholeBlue, $_);
                                @blueLine = split(" ", $_);
                        }
                close BLUE;

                $range = @wholeBlue;

                print STDERR "randomly selecting ~$numberOfPeaks sites to sample in target sam file
                    .\n\n";
                while (($closed < $numberOfPeaks) || ($open < $numberOfPeaks))
                        {
                                $index = int(rand($range));
                                $siteLine = $wholeBlue[$index];
                                @line = split (" ", $siteLine);
                                next if $globalSelection[$index] > 0;
                                if (($line[4] == 0) && ($closed < $numberOfPeaks))
                                        {
                                                $closedChrom{$index} = $wholeBlue[$index];
                                                $closed += 1;
                                        }

                                if (($line[4] >= $numberDataSetsPerPeak) && ($open < $numberOfPeaks
                                    ))
                                        {
                                                $openChrom{$index} = $wholeBlue[$index];
                                                $open += 1;
                                        }
                                $globalSelection[$index] = 1;
                        }

                $x = 0;
                for $key ( sort {$a<=>$b} keys %closedChrom)
                        {
                           $closedChromArray[$x] = $closedChrom{$key};
                           $x += 1;
                        }
                $x = 0;
                for $key ( sort {$a<=>$b} keys %openChrom)
                        {
                           $openChromArray[$x] = $openChrom{$key};
                           $x += 1;
                        }

                $openRef = \@openChromArray;
                $closedRef = \@closedChromArray;
                # free some memory! 20130425
                %closedChrom = ();
                %openChrom = ();
                @wholeBlue = ();

                ##################
                # calculate the mean read counts in each site category in the target sam file
                ##################
                print STDERR "measuring signal intensities in $samForAnalysis\n\n";
                $openStringData = signalIntensity($openRef, 0, $samForAnalysis, $numberOfPeaks);
                @open = split(" ", $openStringData);
                $openReadsPerBin = $open[1];
                $openPeaksAssessed = $open[2];
                print STDERR "$openReadsPerBin reads per bin ($peakBinSize bp) at
                     $openPeaksAssessed randomly selected sites common to $numberDataSetsPerPeak of
                     125 open chromatin data sets\n";
                $closedStringData = signalIntensity($closedRef, 1, $samForAnalysis, $numberOfPeaks)
                    ;
                @closed = split(" ", $closedStringData);
                $thresholdFromThreshold = $closed[0];
                $closedReadsPerBin = $closed[1];
                $closedPeaksCompleted = $closed[2];

                print STDERR "\trBin reads per bin ($peakBinSize bp) at $closedPeaksCompleted
                     randomly selected ENCODE closed chromatin sites.\n";
                print STDERR "\tcalculated read density threshold for peak calling (p<
                     $thresholdSignificanceLevel): $thresholdFromThreshold\n" if $flatThreshold eq
```

```
                                        "NONE";;
                                print STDERR "\tusing flat threshold at $flatThreshold for peak calling.\n\n" if
                                        $flatThreshold ne "NONE";

                        }
                # need to put option here also for random data selection if no blueprint available.
                if ($closedAndOpen eq "NO")
                        {
                                open (CHROMS, "<", "$genomeSizeFile");
                                while (<CHROMS>)
                                        {
                                                @chromLine = split(" ", $_);
                                                $end = $chromLine[1] + $total;
                                                $start =  1 + $total;
                                                push (@chromName, $chromLine[0]);
                                                push (@chromStart, $start);
                                                push (@chromEnd, $end);
                                                $total = $total + $chromLine[1];
                                                $chromCount += 1;
                                        }
                                close CHROMS;
                                $range = $total - 1;
                                my $randomPeakNumber = 0;
                                while ($randomPeakNumber < $numberOfPeaks)
                                        {
                                                $index = int(rand($range)) + 1;
                                                for ($x = 0; $x < $chromCount; $x += 1)
                                                        {
                                                                if (($index >= $chromStart[$x]) && ($index <=
                                                                        $chromEnd[$x]))
                                                                   {
                                                                        $randomPeakNumber+= 1;
                                                                        $randomPeakStart = $index -
                                                                                $chromStart[$x];
                                                                        $randomPeakEnd = $randomPeakStart +
                                                                                10;
                                                                        $chrom = $chromName[$x];
                                                                        $readName = "#" ."$randomPeakNumber
                                                                                ";
                                                                        $randomPeakString = "$chrom\
                                                                                t$randomPeakStart\
                                                                                t$randomPeakEnd\t$readName\n";
                                                                        $randomChromHash{$index} =
                                                                                $randomPeakString;
                                                                   }
                                                        }
                                        }

                                #need to sort the randomchromarray
                                $x = 0;
                                for $key ( sort {$a<=>$b} keys %randomChromHash)
                                        {
                                           $randomChromArray[$x] = $randomChromHash{$key};
                                           $x += 1;
                                        }

                                $randomRef = \@randomChromArray;
                                $randomStringData = signalIntensity($randomRef, 1, $samForAnalysis, $numberOfPeaks)
                                        ;
                                @random = split(" ", $randomStringData);
                                $thresholdFromThreshold = $random[0];
                                $randomReadsPerBin = $random[1];
                                $randomPeaksCompleted = $random[2];
                                print STDERR "\t$randomReadsPerBin reads per bin ($peakBinSize bp) at
                                        $randomPeaksCompleted random sites.\n";
                                print STDERR "\tcalculated read density threshold for peak calling (p<
                                        $thresholdSignificanceLevel): $thresholdFromThreshold\n" if $flatThreshold eq
                                        "NONE";;
                                print STDERR "\tusing flat threshold at $flatThreshold for peak calling.\n\n" if
                                        $flatThreshold ne "NONE";
                        }

                return $thresholdFromThreshold;
        }
sub signalIntensity
        {
                #print STDERR "\n\n***in sub signalIntensity***\n\n";
                print STDERR "\tassessing signal intensity...\n\n";
                my @tedPeaksArray = @{$_[0]};
                my @bedPeaksArray = @tedPeaksArray;

                my $kernalOn = $_[1];

                my $numberOfSubPeaks = $_[3];
                my $targetSamFile = $_[2];
                my $sizeOfBin = $peakBinSize;
                my $sizeOfBackground = $backgroundBinSize;
                my $distanceFromPeakCentre = $sizeOfBin/2;
                my $backgroundFromPeakCentre = $sizeOfBackground/2;
                my $reCalc = 1;
                my $currentTargetPeak = 0;
                my $lowPeak = 0;
                my $highPeak = 0;
                my $peakCentre = 0;
                my $peakChrom = 0;
                my $peakName = 0;
                my $lowLimit = 0;
                my $highLimit = 0;
                my $actualPeakCount = 0;
                my $currentValue = 0;
                my $headerLineCheckerOff = 0;
                my $s = 0;
```

```
                my $flyCount = 0;
                my @globalBin = ();
                my @splitArray = ();
                my @headerCheck = ();
                my @splitSamArray = ();
                my @lineChromSizes = ();
                my %chromOrderScore = ();
                my $simpleCount = 1;
                my @scoreArray = ();
                my @probabilityArray = ();
                my $backgroundScore = 0;
                my $backgroundLowLimit = 0;
                my $backgroundHighLimit = 0;
                my $correctedFlyCount = 0;
                my $meanBackground = 0;
                my $totalFlyCount = 0;

# these are the varibales for returning
                my $returnString = 0;
                my $thresholdFromSignalIntensity = 0;
                my $peaksCompleted = 0;
                my $arrayChecker = 0;
                my $readsPerBin = 0;

                my @lowBack = ();
                my @highBack = ();
                my @lowBin = ();
                my @highBin = ();

                my @centrePoint = ();
                my $x = 0;
                my @lineSplitter = ();
                my @chromosomeNamesForIntensity = 0;

                #open target bed file, and randomly select $absoluteReadNumber reads, and put these into an array
                     new file

                # get the order of chromosomes to exclude skips;
                open (CHROM, "<", "$genomeSizeFile") or die "$genomeSizeFile not found\n\n";
                while (<CHROM>)
                        {
                                @lineChromSizes = split(" ", $_);
                                $chromOrderScore{"$lineChromSizes[0]"} = $simpleCount;
                                $simpleCount += 1;
                        }
                close CHROM;

                $actualPeakCount = @bedPeaksArray;

                print STDERR "\t$actualPeakCount sites were selected randomly.\n";

                for ($x = 0; $x<$actualPeakCount; $x += 1)
                        {
                                @lineSplitter = split (" ", $bedPeaksArray[$x]);
                                $centrePoint[$x] = int(($lineSplitter[1] + $lineSplitter[2])/2);
                                $chromosomeNamesForIntensity[$x] = $lineSplitter[0];
                        }
# make the outer bins, and exclude any overlapping bins
my $lowBackCount = 0;
my $highBackCount = 0;
my $checkingCounters = 0;
my $nextEx = 0;
my $loopCount = 0;
my $diffCount = 0;
my $spliced = 0;
my $totalRawCentralBinScore = 0;
my $totalCheckingCounters = 0;
while ($loopCount <= 10)
        {
                $actualPeakCount = @centrePoint;

                for ($x = 0; $x<$actualPeakCount; $x += 1)
                                        {
                                                $lowBack[$x] = $centrePoint[$x] - $backgroundFromPeakCentre;
                                                $highBack[$x] = $centrePoint[$x] + $backgroundFromPeakCentre;
                                        }

                $lowBackCount = @lowBack;
                $highBackCount = @highBack;

                for ($x = 0; $x<$actualPeakCount; $x += 1)
                                        {
                                                $nextEx = $x + 1;

                                                if (($highBack[$x] > $lowBack[$nextEx]) && (
                                                     $chromosomeNamesForIntensity[$x] eq
                                                     $chromosomeNamesForIntensity[$nextEx]))
                                                   {
                                                           $checkingCounters += 1;
                                                           $spliced = splice (@centrePoint, $nextEx, 1);
                                                           splice (@lowBack, $nextEx, 1);
                                                           splice (@highBack, $nextEx, 1);
                                                           splice (@chromosomeNamesForIntensity, $nextEx, 1);
                                                   }
                                        }
                $loopCount += 1;
                $totalCheckingCounters = $totalCheckingCounters + $checkingCounters;
                last if $checkingCounters == 0;
                $checkingCounters = 0;
        }

print STDERR "\t$totalCheckingCounters sites discarded for overlapping.\n\n";
```

```
# make the boundary coordinates of inner bins
$actualPeakCount = @centrePoint;
for ($x = 0; $x<$actualPeakCount; $x += 1)
                {
                        $lowBin[$x] = $centrePoint[$x] - $distanceFromPeakCentre;
                        $highBin[$x] = $centrePoint[$x] + $distanceFromPeakCentre;
                }

# find out how many reads in each bin:
my $chr = 0;
my $lowBackground = 0;
my $lowCentre = 0;
my $highCentre = 0;
my $highBackground = 0;
my $targetLine = 0;
my @samArray = ();
my $backGroundClock = 0;
my $binClock = 0;
my @centreScoresArray = ();
my @backgroundScoresArray = ();
my $chrNameCheck = 0;
my $loopChromNameCheck = 0;

open (SAM, "<", $targetSamFile);
while (<SAM>)
        {
                $targetLine = $_;
                @samArray = split(" ", $targetLine);
                @headerCheck = split("", $_);
                last if $headerCheck[0] ne "@";
                $headerLineCheckerOff = 1 if $headerCheck[0] ne "@";
        }

## check sites on each chromosome
my %chromSiteCount = ();
for ($x=0; $x<$actualPeakCount; $x+=1)
        {
                $chromSiteCount{"$chromosomeNamesForIntensity[$x]"} += 1;

        }

my @chromiumName = ();
my @lineChecker = ();

for ($x=0; $x<$actualPeakCount; $x+=1)
        {
                $chr = $chromosomeNamesForIntensity[$x];
                if ($chr ne $chrNameCheck)
                        {
                                print STDERR "\tsampling signal at $chromSiteCount{$chr} sites on $chr...\n";
                                $samArray[3] = 0;
                        }
                $lowBackground = $lowBack[$x];
                $lowCentre = $lowBin[$x];
                $highCentre = $highBin[$x];
                $highBackground = $highBack[$x];
                while ($samArray[3] <= $highBackground)
                        {
                                #print STDERR "FOR: $targetLine";
                                $backGroundClock += 1 if (($samArray[3] > $lowBackground) && ($samArray[3] <
                                     $highBackground));
                                $binClock += 1 if (($samArray[3] > $lowCentre) && ($samArray[3] < $highCentre));
                                $targetLine = <SAM> unless $loopChromNameCheck ne $samArray[2];
                                @samArray = split(" ", $targetLine);
                                if ($samArray[2] ne $chr)
                                        {
                                                $samArray[3] =0;
                                                last if $loopChromNameCheck ne $samArray[2];
                                                #print STDERR "stuck inthis loop\n";
                                        }
                                $loopChromNameCheck = $samArray[2];
                        }

                push(@centreScoresArray, $binClock);
                push(@backgroundScoresArray, $backGroundClock);

                $binClock = 0;
                $backGroundClock = 0;
                $chrNameCheck = $chr;
        }


my @standardArray = ();
my $correctedFlyCountPlus5000 = 0;

# take bin scores and make background corrected density scores
my $maximumBinScore = 0;
my $backgroundCorrectedDensity;
#%densityFrequencyHash is global to allow calculation of pvalueTable.

for ($x=0; $x<$actualPeakCount; $x+=1)
        {
                $meanBackground = ($backgroundScoresArray[$x]/$sizeOfBackground)*$sizeOfBin;
                $backgroundCorrectedDensity = $centreScoresArray[$x] - $meanBackground;
                $backgroundCorrectedDensity = sprintf ("%.2f", $backgroundCorrectedDensity);
                $densityFrequencyHash{$backgroundCorrectedDensity} += 1;
                $maximumBinScore = $backgroundCorrectedDensity if $backgroundCorrectedDensity > $maximumBinScore;
                $totalRawCentralBinScore = $totalRawCentralBinScore + $centreScoresArray[$x];
        }
# calculate some numbers
$peaksCompleted = @centreScoresArray;
```

```
$readsPerBin = $totalRawCentralBinScore/$peaksCompleted;
$readsPerBin = sprintf("%.2f", $readsPerBin);


my $standardArrayCount = @standardArray;


                # $threshold and @dataForPValue $readsPerBin and $peaksCompleted are global variables. $threshold
                      is used and must be set here prior to peak calling (peak calling sub will crash without it).
                      @dataForPValue is a list of bin scores in closed chromatin, and can be called on by the
                      peakPValue sub. $readsPerBin and $peaks completed are collected on the fly by threshold sub.
                $thresholdFromSignalIntensity = kernel(\%densityFrequencyHash, $thresholdSignificanceLevel) if
                      $kernalOn == 1;

                #make a hash for looking up peak P values
                my %pValueTable = ();
                if (($kernalOn == 1) && ($maxPeakPValue eq "ON"))
                        {
                                print STDERR "\n\tcreating peak density p value table...\n";

                                for ($x=0; $x<=50000; $x+=1)
                                        {
                                                $pValueTable{$x} = peakPValue($x);
                                                $x=50001 if $pValueTable{$x} == "0.0000000001";
                                        }
                        }

                @probabilityArray = ();
                @bedPeaksArray = ();
                my $pValueTableRef = \%pValueTable;
                # $meanTotalMeanSignalIndex = $totalMeanSignalIndex/$totalPositions;
                print STDERR "\n\t$peaksCompleted sites sampled.\n\n";

                $returnString = "$thresholdFromSignalIntensity " . "$readsPerBin " . "$peaksCompleted";
                open (PEAKNUMBFORTHRESH, ">>", "backpeaksForThresh.txt");
                print PEAKNUMBFORTHRESH "***this next two lines are not threshold lines:\n" if $kernalOn == 0;
                print PEAKNUMBFORTHRESH "***threshold lines:\n" if $kernalOn == 1;
                print PEAKNUMBFORTHRESH "$targetSamFile\t$numberOfSubPeaks\t$peaksCompleted\
                      t$thresholdFromSignalIntensity\n";
                print PEAKNUMBFORTHRESH "readsPerBin: $readsPerBin\n\n";
                close PEAKNUMBFORTHRESH;

                return $returnString;
        }

# sub kernel runs only once: it takes an array of bin scores from the closed chromatin sites, and returns a
      threshold corresponding to the globally set $thresholdSignificanceLevel
        }
sub kernel
        {
                #print STDERR "\n\n***in sub kernal***\n\n";
                my %hashForKernel = %{$_[0]};
                my $thresholdSignificanceLevel = $_[1];

                my $n;
                foreach my $key ( keys %hashForKernel )
                        {
                                $n += $hashForKernel{$key};
                        }

                my $bandWidth = 1;
                my $oneOverRootTwoPi = 0.39894228;
                my $e = 2.718281828;
                my $round = 0;
                my $probTotal = 0;
                my $eFactor = 0;
                my $eFigure = 0;
                my $gaussFigure = 0;
                my $rankNumber = 0;
                my $fracFactor = 0;
                my $interTotal = 0;
                my $actualTotal = 0;
                my @rankProbArray = ();
                my $posCount = 0;
                my $diff = 0;
                my $absDiff = 0;
                my $thresholdFromKernel = 0;



                for ($rankNumber = 0; $rankNumber <= 100; $rankNumber += 0.1)
                        {

                                foreach my $key ( keys %hashForKernel )
                                {
                                  for (my $x = 0; $x < $hashForKernel{$key}; $x += 1)
                                        {
                                                $diff = $rankNumber - $key;
                                                $absDiff = abs($diff);
                                                $eFactor = (($absDiff*$absDiff)/$bandWidth)/-2;
                                                $eFigure = $e**$eFactor;
                                                $gaussFigure = $oneOverRootTwoPi*$eFigure;
                                                $interTotal = $gaussFigure;
                                                $probTotal = $probTotal + $interTotal;
                                        }
                                }

                                $actualTotal = $probTotal/($n*$bandWidth);
                                #print STDOUT "KERNEL: **Rank $rankNumber\t Total: $actualTotal\n";
                                # stops the looping as soon as P is less than $thresh significance level
                                if ($actualTotal < $thresholdSignificanceLevel)
                                        {
```

```
                                                $thresholdFromKernel = sprintf("%.1f", $rankNumber);
                                                @rankProbArray = (); # frees memory;
                                                return $thresholdFromKernel;
                                                last;
                                        }

                                push (@rankProbArray, $actualTotal);
                                $probTotal = 0;
                                $posCount = 0;
                        }

        }


# calculates p values for individual peak maxima, using the global data array, @dataForPValue, which is filled by
      the signalIntensity sub.
sub peakPValue
        {
                my $maximaForP = $_[0];
                my $nP = 0;
                foreach my $key ( keys %densityFrequencyHash )
                        {
                                $nP += $densityFrequencyHash{$key};
                        }

                die "array with sample bin scores is empty!\n\n" if ($nP == 0);
                my $bandWidth = 1;
                my $oneOverRootTwoPi = 0.39894228;
                my $e = 2.718281828;
                my $round = 0;
                my $probTotal = 0;
                my $eFactor = 0;
                my $eFigure = 0;
                my $gaussFigure = 0;
                my $fracFactor = 0;
                my $interTotal = 0;
                my $actualTotal = 0;
                my $diff = 0;
                my $absDiff = 0;

                foreach my $key ( keys %densityFrequencyHash )
                        {
                          for (my $x = 0; $x < $densityFrequencyHash{$key}; $x += 1)
                                {

                                        $diff = $maximaForP - $key;
                                        $absDiff = abs($diff);
                                        $eFactor = (($absDiff*$absDiff)/$bandWidth)/-2;
                                        $eFigure = $e**$eFactor;
                                        $gaussFigure = $oneOverRootTwoPi*$eFigure;
                                        $fracFactor = $absDiff/$bandWidth;
                                        $interTotal = $fracFactor*$gaussFigure;
                                        $probTotal = $probTotal + $interTotal;
                                }
                        }
                $actualTotal = $probTotal/($nP*$bandWidth);
                $actualTotal = sprintf("%.10f", $actualTotal);
                $actualTotal = "0.0000000001" if $actualTotal < 0.0000000001;
                return $actualTotal;
        }

# sub peakCaller takes two arguments: the name of the file to call peaks on, and the threshold density to call
      peaks at
sub peakCall
        {
                #print STDERR "\n\n***in sub peakCall***\n\n";
                my $filename = $_[0];
                my $thresholdPeakCall = $_[1];
                my $centralBinSize = $_[2];
                my $binSize = $_[3];
                die "threshold calculation failed in peakCall sub.\n\n" if $thresholdPeakCall == 0;
                my %reads = ();
                my @line = ();
                my $token = 0;
                my $count = 0;
                my $binTotal = 0;
                my $backgroundTotal = 0;
                my @headerCheck = ();
                my $headerLineCheckerOff = 0;
                my @currentLineSplit = ();
                my $chromosome = 0;
                my $chromosomeSize = 0;
                my $currentBinStart = 0;
                my $currentBinEnd = 0;
                my $centralBinStart = 0;
                my $centralBinEnd = 0;
                my $centralBinScore = 0;
                my $binScore = 0;
                my $startString = 1;
                my @readStartPositionArray = ();
                my $arraySize = 0;
                my $peakOn = 0;
                my $backgroundScore = 0;
                my $segmentScore = 0;
                my $peakChr = 0;
                my $peakStart = 0;
                my $peakEnd = 0;
                my $peakLine = 0;
                my $peakMax = 0;
                my $peakNumber = 0;
                my @peakArray = ();
```

```
        my $i = 0;
        # create hash with chromkey and chromsize value
                        my %hashChromSizes = ();
                        open (CHROMSIZE, "<", "hg19.chrom.sizes"); #$genomeSizeFile");
                        while (<CHROMSIZE>)
                                {
                                        chomp;
                                        my ($keyChromName, $valChromSize) = split /\t/;
                                        $hashChromSizes{$keyChromName} = $valChromSize;
                                }
                        close CHROMSIZE;
                        ### print wig header in new wig file
                        print STDOUT "track type=wiggle_0 name=$trackName description=$trackDescription
                             viewLimits=0:500 autoScale=off gridDefault=on color=0,160,255 maxHeightPixels
                             =60:60:11 visibility=full windowingFunction=mean+whiskers\n";
                        ###start of density track making
open (ORIGINALSAM, "<$filename");
while (<ORIGINALSAM>)
        {
                if ($headerLineCheckerOff == 0)
                        {
                                @headerCheck = split("", $_);
                                $headerLineCheckerOff = 1 if $headerCheck[0] ne "@";
                                next if $headerLineCheckerOff == 0;
                        }
                @currentLineSplit = split(" ", $_);
                if ("$currentLineSplit[2]" ne "$chromosome")
                        {
                                $chromosome = "$currentLineSplit[2]";
                                $chromosomeSize = $hashChromSizes{$chromosome};
                                print STDERR "\tfinding peaks in $chromosome...\n";
                                $currentBinStart = $startString; #+ $offSetSubDR;
                                $currentBinEnd = $currentBinStart + $binSize;
                                $centralBinStart = ($currentBinEnd/2 - $centralBinSize/2) - 0.5;
                                $centralBinEnd = $centralBinStart + $centralBinSize;
                                @readStartPositionArray = ();
                        }
        push(@readStartPositionArray, $currentLineSplit[3]);
        while (("$currentLineSplit[3]" > "$currentBinEnd") && ($currentBinEnd < $chromosomeSize ))
                {
                        while ($readStartPositionArray[0]<$currentBinStart)
                                {
                                        shift(@readStartPositionArray);
                                }

                        $arraySize = @readStartPositionArray;
                        for ($i = 0; $i < "$arraySize"; $i += 1)
                                {
                if (($readStartPositionArray[$i] >= $currentBinStart) && ($readStartPositionArray[$i] <
                     $currentBinEnd) && ($currentBinStart < $currentBinEnd))
                                {
                                $binScore += 1;
                                }

                if (($readStartPositionArray[$i] >= $centralBinStart) && ($readStartPositionArray[$i] <
                     $centralBinEnd) && ($centralBinStart < $centralBinEnd))
                                {
                                        $centralBinScore += 1;
                                }
                }
$backgroundScore = int(($binScore/$binSize)*$centralBinSize);
$segmentScore = $centralBinScore - $backgroundScore;
if (($segmentScore >= $thresholdPeakCall) && ($peakOn ==0))
        {
                $peakOn = 1;
                $peakStart = $centralBinStart;
                $peakMax = $segmentScore;
                $peakEnd = $centralBinEnd;
                $peakChr = $chromosome;
        }
if (($segmentScore >= $thresholdPeakCall) && ($peakOn ==1))
        {
                $peakMax = $segmentScore if $segmentScore > $peakMax;
                $peakEnd = $centralBinEnd;
        }
if (($segmentScore < $thresholdPeakCall) && ($peakOn == 1))
        {
                $peakNumber += 1;
                if ($maxPeakPValue eq "ON")
                        {
                                $peakMax = "$pValueTable{$peakMax}";
                                $peakMax = "0.0000000001" if $peakMax < 0.0000000001;
                        }
                $peakLine = "$peakChr\t$peakStart\t$peakEnd\tpk$peakNumber\t$peakMax\t+\n";
                print STDOUT "$peakLine";
                $peakOn = 0;
        }
if (($segmentScore < $thresholdPeakCall) && ($peakOn == 0))
        {
        }
$currentBinStart += "$stepSize";
$currentBinEnd += "$stepSize";
$centralBinStart += "$stepSize";
$centralBinEnd += "$stepSize";
$binScore = 0;
$centralBinScore = 0;
}
}
close ORIGINALSAM;
print STDERR "\n\t$peakNumber peaks found in $filename.\n\n";
return $peakNumber;
}
```

```
sub randomReadSelection
        {
                die "no parameters sent to Random Read Selection subroutine: requires filename and number of reads
                     required\n\n" unless @_;
                my $file = $_[0];
                my $absoluteReadNumber = $_[1];
                my $counter = 0;
                my $numerator = 0;
                my $denominator = 0;
                my $percentage = 0;
                my $random = 0;
                my $check = 0;
                my $headerOff = 0;
                my @headerCheckerArray = ();
                print STDERR "counting reads...\n";
                open (FILE, "<$file");
                while (<FILE>)
                        {
                                if ($headerOff == 0)
                                        {
                                                @headerCheckerArray = split("", $_);
                                                $headerOff = 1 if $headerCheckerArray[0] ne '@';
                                        }

                                if ($headerOff ==1)
                                        {
                                                $counter += 1;
                                        }
                        }
                close FILE;
                $headerOff = 0;
                $percentage = ($absoluteReadNumber/$counter)*100;
                ### more zeros gives more precision in estimation of numbers of reads to select.
                $numerator = int($percentage*10000);
                $denominator = 1000000;
                my $roundedPercent = sprintf("%.1f", $percentage);
                print STDERR "selecting ~$roundedPercent% of $counter reads...\n";
                open (FILE, "<$file");
                while (<FILE>)
                        {
                                if ($headerOff == 0)
                                        {
                                                @headerCheckerArray = split("", $_);
                                                print STDOUT "$_";
                                                $headerOff = 1 if $headerCheckerArray[0] ne '@';
                                        }

                                if ($headerOff ==1)
                                        {
                                                $random = int(rand(1000000));
                                                if ($random <= $numerator)
                                                        {
                                                                print STDOUT "$_";
                                                                $check += 1;
                                                        }
                                        }
                        }
                close FILE;
                print STDERR "$check reads randomly selected.\n\n";
        }

sub samFilter
        {
                #       takes a mapped unsorted sam file that has been sorted by chromosome position with unix
                     command:
                #       sort -n --key=4,5 filename.sam > sortedFilename.sam

                my $headerFile = $_[0];
                my $sortedFile = $_[1];
                my $mapq = $_[2];
                my $uq = $_[3];
                my $currentTempFile = 0;
                my $headerOff = 0;
                my @headerCheckerArray = ();
                my @splitArray = ();
                my @currentUQArray = ();
                my $passedFilter = 0;
                my $totalReads = 0;
                my $i = 0;
                my $percentagePassedFilter = 0;
                my $numberOfChromosomes = 0;
                my $tempSwitch = 0;
                my $unmappedCount = 0;
                my $percentageUnmapped = 0;
                my $chromCount = 0;
                my @chromNames = ();
                my @chromSizes = ();
                my @line = ();
                my %chromReadCount = ();
                my $headOne = 0;
                my $samHeader = "FALSE";
                my $percentageDuplicates = 0;
                my $testString = 0;
                my $oldTestString = 0;
                my $duplicateCount = 0;
                print STDOUT "\@CO\ttime:$timeString\tprog:$0\tuser:$user\tcwd:$dir\tmapq:$mapq\tuq:$uq\n"; #
                     inserts new header comment line with current additional details
                open(HEADERFILE, "<$headerFile");
                while (<HEADERFILE>) # then prints remainder of original header and stops at with first non-header
                     line
                        {
```

```
                @splitArray = split("", $_);
                if ($splitArray[0] eq '@')
                        {
                                $samHeader = "TRUE";
                                print STDOUT "$_";
                        }
                else {last;}

        }
close HEADERFILE;
print STDERR "sam header detected, original header transferred to new file. \n" if $samHeader eq "
    TRUE"; #confirms status of header
print STDERR "warning: sam header not detected.\n" if $samHeader eq "FALSE";
# create new @SQ tags in the order of $genomeSizeFile. These tags are indicated by @SQ\tSNN:<chrom.
    name>:hg19ChromSize\tLN:<chrom.size>. The ultrafiltered sam file will be made in this order
print STDERR "creating chromosome order list from $genomeSizeFile...\n";
open (CHROMSIZES, "<$genomeSizeFile");
while (<CHROMSIZES>)
        {
                @line = split(" ", "$_");
                push (@chromNames, "$line[0]");
                push (@chromSizes,  "$line[1]");
        }

# initialise chromosome passed filter counts hash with zero value for each:
for ($i = 0; $i < @chromNames; $i += 1)
        {
                $chromReadCount{$chromNames["$i"]} = 0;
        }
$currentTempFile = "$headerFile" . ".temp";
# makes a new temporary file containing only reads that pass the mapq and uq filters
print STDERR "undertaking initial mapq and uq filtering...\n";
open (SORTEDINPUT, "<$sortedFile");
open (TEMP, ">$currentTempFile");
my %presentChroms = ();
my @chromsToPrint = ();
my @testChroms = ();
while (<SORTEDINPUT>)
        {
                if ($headerOff == 0)
                        {
                                @headerCheckerArray = split("", $_);
                                $headerOff = 1 if $headerCheckerArray[0] ne '@';
                        }
                if ($headerOff ==1)
                        {
                                @splitArray = split(" ", $_);
                                if ($pcrDuplicates eq "ON") # next seven lines added 20130507 to
                                    put in PCR duplicate checker
                                    {
                                            $oldTestString = $testString;
                                            $testString = "$splitArray[2]"."$splitArray[3]"."
                                                $splitArray[4]"."$splitArray[8]";
                                            $duplicateCount += 1 if $testString eq
                                                $oldTestString;
                                            next if $testString eq $oldTestString;
                                    }

                                $unmappedCount += 1 if $splitArray[3] == 0;
                                if ($splitArray[3] > 0)
                                        {
                                                @currentUQArray = split(":", $splitArray[11]);
                                                if (($splitArray[4] >= $mapq) && ($currentUQArray
                                                    [2] <= $uq))
                                                    {
                                                            $presentChroms{"$splitArray[2]"} =
                                                                0 unless exists $presentChroms
                                                                {"$splitArray[2]"};



                                                                                        print
                                                                TEMP "$_";
                                                            $passedFilter += 1;
                                                            $chromReadCount{$splitArray[2]} +=
                                                                1;
                                                    }
                                        }
                                $totalReads += 1;
                        }
        }
close SORTEDINPUT;
close TEMP;

@testChroms = %presentChroms; # makes an array of chromosomes that have been found in the file
for ($i = 0; $i < @chromNames; $i += 1) # adds to a new array chromosomes found in the file, in the
    order they appear in the chrom.sizes file
        {
                push (@chromsToPrint, "$chromNames[$i]") if exists $presentChroms{"$chromNames[$i]"
                    };
        }

# adding @SQ lines to sorted sam header that contain number of reads that have passed the filters
my $accuratePercent = $passedFilter/$totalReads*100;
$accuratePercent = sprintf("%.2f", $accuratePercent);

for ($chromCount = 0; $chromCount < @chromNames; $chromCount += 1)
        {
                print STDOUT "\@SQ\tSN:$chromNames[$chromCount]:readsPassedFilter:$chromReadCount{
                    $chromNames[$chromCount]}\tLN:$chromSizes[$chromCount]\n";
        }
```

```
                print STDOUT "\@CO\ttotalReads:$totalReads\treadsPassedFilter:$passedFilter\t percentPassed:
                    $accuratePercent%\tmapQpoint:$mapq\tUQpoint:$uq\n";
                print STDERR "new header completed.\nbeginning ordering of chromosomes...\n\n";

                ## this is the ordering section
                $numberOfChromosomes = @chromsToPrint;
                for ($i = 0; $i < $numberOfChromosomes; $i += 1)
                        {
                                print STDERR "sorting $chromsToPrint[$i]...\n";
                                open (TEMPA, "<$currentTempFile");
                                while (<TEMPA>)
                                        {
                                                print STDOUT "$_" if ($_ =~ m#\b$chromsToPrint[$i]\
                                                    b#);
                                        }
                                close TEMPA;
                        }
                unlink $currentTempFile;
                $percentagePassedFilter = ($passedFilter/$totalReads)*100;
                $percentageUnmapped = ($unmappedCount/$totalReads)*100;
                $percentagePassedFilter = sprintf("%.2f", $percentagePassedFilter);
                $percentageUnmapped = sprintf("%.2f", $percentageUnmapped);
                $percentageDuplicates = ($duplicateCount/$totalReads)*100;
                $percentageDuplicates = sprintf("%.2f", $percentageDuplicates);
                print STDERR "\n$percentageUnmapped% of total reads were unmapped.\n";
                print STDERR "$duplicateCount PCR duplicates ($percentageDuplicates% of total) were removed.\n" if
                    $pcrDuplicates eq "ON";
                print STDERR "$passedFilter ($percentagePassedFilter%) reads of total $totalReads successfully
                    passed filter.\n\n";
        }

sub densityAnalyzer
        {
                #removed requirement for HCS passed filter
                # definition of sub variables
                my $filename = $_[0];
                my $stepSizeSubDR = $_[1];
                my $tailSubDR = $_[2];
                our $sigmaSubDR = $_[3];
                my $threshSubDR = $_[4];
                my $offSetSubDR = $_[5];
                my $bandWidth = $tailSubDR*2 +1;
                my $binSize = $tailSubDR*2 +1;
                my $span = $stepSizeSubDR;
                my $gaussCount = 0;
                my $headerLineCheckerOff = 0;
                my $currentBinStart = 0;
                my $currentBinEnd = 0;
                my $positionCount = 0;
                my $i = 0;
                my $chromosome = 0;
                my $startString = 1;
                my $binScore = 0;
                my $arraySize = 0;
                my $chromosomeSize = 0;
                my $headerSwitch = 0;
                my $gaussOutput = 0;
                my $memSaverGauss = 0;
                my $sum = 0;
                my $roundedFigure = 0;
                my $gapCount = 0;
                my $stub = $filename;chomp $stub;for ($i = 0; $i < 4; $i += 1){chop $stub;}
                my $trackName = "$stub.step$stepSizeSubDR.tail$tailSubDR.sig$sigmaSubDR.thresh$threshSubDR.
                    off$offSetSubDR";
                my $trackDescription = "$stub" . "-$stepSizeSubDR.$tailSubDR.$sigmaSubDR.$threshSubDR.$offSetSubDR"
                    . "-";
                my @currentLineSplit = ();
                my @readStartPositionArray = ();
                my @headerCheck = ();
                my @splitArray = ();
                my @gaussArray = ();
                our @fixedGaussArray = ();
                my @gaussSubArray = ();

                # generates the gaussian values based on sigma and tailsize; eg for tail of 150, 301 values are
                    generated and placed in the fixedGauss reference array
                for ($i = -$tailSubDR; $i <= $tailSubDR; $i += 1)
                        {
                                $gaussOutput = $sigmaSubDR*gaussEngine($i);
                                push(@fixedGaussArray, $gaussOutput);
                        }

                # if thresh is not set by user, calculates default bin thresh
                if ($threshSubDR eq "DEFAULT")
                        {
                                my $threshCalcHeaderOff = 0; #<-removed == from here 20130703
                                my $threshCalcReadCount = 0;
                                my @threshCalcHeader = ();
                                print STDERR "calculating default minimum bin threshold...\n";
                                $genomeBaseCount = baseCountGenomeSizeFile($genomeSizeFile);

                                open (THRESHCALC, "<$filename");
                                while (<THRESHCALC>)
                                        {
                                                if ($threshCalcHeaderOff == 0)
                                                        {
                                                                @threshCalcHeader = split("", $_);
                                                                $threshCalcHeaderOff = 1 if $threshCalcHeader[0] ne
                                                                    "@";
                                                                next if $threshCalcHeaderOff == 0;
                                                        }
```

```
                                $thresh_calc_read_count += 1
                        }
                $thresh_sub_dr = 2*(($thresh_calc_read_count/$genome_base_count)*$bin_size);
                $thresh_sub_dr = sprintf("%.0f", $thresh_sub_dr);
                print STDERR "genome size file: $genome_size_file\ngenome total size:
                        $genome_base_count\ntotal number of reads: $thresh_calc_read_count\ndefault
                        calculated minimum bin threshold: $thresh_sub_dr\n\n";
        }

# create hash with chromkey and chromsize value
my %hash_chrom_sizes = ();
open (CHROMSIZE, "<$genome_size_file");
while (<CHROMSIZE>)
        {
                chomp;
                my ($key_chrom_name, $val_chrom_size) = split /\t/;
                $hash_chrom_sizes{$key_chrom_name} = $val_chrom_size;
        }
close CHROMSIZE;

### print wig header in new wig file
print STDOUT "track type=wiggle_0 name=$track_name description=$track_description viewLimits=0:500
      autoScale=off gridDefault=on color=0,160,255 maxHeightPixels=60:60:11 visibility=full
      windowingFunction=mean+whiskers\n";
###start of density track making
open (ORIGINALSAM, "<$filename");
while (<ORIGINALSAM>)
        {
                if ($header_line_checker_off == 0)
                        {
                                @header_check = split("", $_);
                                $header_line_checker_off = 1 if $header_check[0] ne "@";
                                next if $header_line_checker_off == 0;
                        }
                @current_line_split = split(" ", $_);
                if ("$current_line_split[2]" ne "$chromosome")
                        {
                                $chromosome = "$current_line_split[2]";
                                $chromosome_size = $hash_chrom_sizes{$chromosome};
                                print STDERR "making density track for $chromosome...\n";
                                $current_bin_start = $start_string + $off_set_sub_dr;
                                $current_bin_end = $current_bin_start + $bin_size;
                                @read_start_position_array = ();
                                @gauss_array = ();
                                $position_count = 0;
                                $gap_count = 1;
                        }
                        push(@read_start_position_array, $current_line_split[3]);
                        while (("$current_line_split[3]" > "$current_bin_end") && ($current_bin_end <
                              $chromosome_size ))
                                {
                                        while ($read_start_position_array[0]<$current_bin_start)
                                                {
                                                        shift(@read_start_position_array);
                                                }
                                        $array_size = @read_start_position_array;
                                        for ($i = 0; $i < "$array_size"; $i += 1)
                                                {
                                                        if (($read_start_position_array[$i] >=
                                                              $current_bin_start) && (
                                                              $read_start_position_array[$i] <
                                                              $current_bin_end) && ($current_bin_start <
                                                               $current_bin_end))
                                                                {
                                                                        $bin_score += 1;
                                                                }
                                                }
                                        if ($bin_score >= $thresh_sub_dr)
                                                {
                                                        for ($gauss_count = 0; $gauss_count <
                                                              $band_width; $gauss_count += 1)
                                                                {
                                                                        $gauss_array[$gauss_count] =
                                                                              $gauss_array[
                                                                              $gauss_count] +
                                                                              $bin_score*
                                                                              $fixed_gauss_array[
                                                                              $gauss_count];
                                                                }
                                                }
                                        @gauss_sub_array = splice(@gauss_array, 0, "$step_size_sub_dr");
                                        $sum += $_ for (@gauss_sub_array);
                                        $rounded_figure = int(10*$sum);
                                        if ($rounded_figure == 0)
                                                {
                                                        $gap_count = 1;
                                                        @gauss_array = ();
                                                }
                                        if (($rounded_figure >0) && ($gap_count == 0))
                                                {
                                                        for ($gauss_count = 0; $gauss_count <
                                                              $step_size_sub_dr; $gauss_count += 1)
                                                                {
                                                                        $mem_saver_gauss = int(10*
                                                                              $gauss_sub_array[
                                                                              $gauss_count]);
                                                                        print STDOUT "
                                                                              $mem_saver_gauss\n";
                                                                }
                                                }
                                        if (($rounded_figure > 0) && ($gap_count > 0))
                                                {
```

```
                                                                        #$newPosition = $startString +
                                                                            $positionCount*$stepSizeSubDR;
                                                                        print STDOUT "fixedStep chrom=$chromosome
                                                                            start=$currentBinStart step=1\n";

                                                                        for ($gaussCount = 0; $gaussCount <
                                                                            $stepSizeSubDR; $gaussCount += 1)
                                                                            {
                                                                                $memSaverGauss = int(10*
                                                                                    $gaussSubArray[
                                                                                    $gaussCount]);
                                                                                print STDOUT "
                                                                                    $memSaverGauss\n";
                                                                            }
                                                                        $gapCount = 0;
                                                                }

                                                            $sum = 0;
                                                            $positionCount += 1;
                                                            $currentBinStart += "$stepSizeSubDR";
                                                            $currentBinEnd += "$stepSizeSubDR";
                                                            $binScore = 0;
                                                    }
                                    }
                            close ORIGINALSAM;
                            print STDERR "\nDensity track created.\n\n";
                    }

sub peakSignalIntensity
                    {
                            my $file = $_[0];
                            my $distanceFromPeakCentre = $_[1];
                            my $numberOfPeaks = $_[2];
                            my $targetSamFile = $_[3];
                            my $counter = 0;
                            my $numerator = 0;
                            my $denominator = 0;
                            my $percentage = 0;
                            my $random = 0;
                            my $check = 0;
                            my $actualPeakCount = 0;
                            my $reCalc = 1;
                            my $currentTargetPeak = 0;
                            my $lowPeak = 0;
                            my $highPeak = 0;
                            my $peakCentre = 0;
                            my $peakChrom = 0;
                            my $peakName = 0;
                            my $lowLimit = 0;
                            my $highLimit = 0;
                            my $readValue = 0;
                            my $peaksCompleted = 0;
                            my $currentValue = 0;
                            my $currentXCoord = 0;
                            my $totalPositions = 0;
                            my $headerLineCheckerOff = 0;
                            my $s = 0;
                            my @meanGlobalBin = ();
                            my @globalBin = ();
                            my @splitArray = ();
                            my @headerCheck = ();
                            my @bedPeaksArray = ();
                            my @splitSamArray = ();
                            #open target bed file, and randomly select $absoluteReadNumber reads, and put these into an array
                                new file
                            open (FILE, "<$file");
                            while (<FILE>)
                                {
                                        $counter += 1;
                                }
                            close FILE;

                            $percentage = ($numberOfPeaks/$counter)*100;
                            $numerator = int($percentage*100000);
                            $denominator = 10000000;
                            my $roundedPercent = sprintf("%.2f", $percentage);
                            open (FILE, "<$file");
                            while (<FILE>)
                                {
                                        @splitArray = split(" ", $_);
                                                $random = int(rand(10000000));
                                                if ($random < $numerator)
                                                        {
                                                                chomp $_;
                                                                $bedPeaksArray[$actualPeakCount] = $_;
                                                                $actualPeakCount += 1;
                                                        }
                                }
                            close FILE;
                            print STDERR "\ttotal peaks: $counter; percentage for $numberOfPeaks peaks: $roundedPercent%\n\
                                t$actualPeakCount peaks were randomly selected.\n";
                            print STDERR "\tcompleted target peak survey.\n\tcalculating average signal intensity for
                                $actualPeakCount peaks...\n\n";
                            open (SAMFILE, "<$targetSamFile");
                            while (<SAMFILE>)
                                {
                                        if ($headerLineCheckerOff == 0)
                                                {
                                                        @headerCheck = split("", $_);
                                                        next if $headerCheck[0] eq "@";
                                                        $headerLineCheckerOff = 1 if $headerCheck[0] ne "@";
                                                }
```

```
                        if ($headerLineCheckerOff == 1)
                                {
                                        @splitSamArray = split("\t", $_);
                                        if ($reCalc == 1)
                                                {
                                                        $currentTargetPeak = $bedPeaksArray[$currentValue];
                                                        @splitArray = split("\t", $currentTargetPeak);
                                                        $lowPeak = $splitArray[6];
                                                        $highPeak = $splitArray[7];
                                                        $peakCentre = int(($lowPeak + $highPeak)/2);
                                                        $peakChrom =  $splitArray[0];
                                                        $peakName =  $splitArray[3];
                                                        $lowLimit = $peakCentre - $distanceFromPeakCentre;
                                                        $highLimit = $peakCentre + $distanceFromPeakCentre;
                                                        last if $lowPeak == 0;
                                                        $reCalc = 0;
                                                }

                                        if (($splitSamArray[2] eq "$peakChrom") && ($splitSamArray[3] >= "
                                            $lowLimit") && ($splitSamArray[3] <= "$highLimit"))
                                                {
                                                        $readValue = $splitSamArray[3] - $lowLimit;
                                                        $globalBin[$readValue] += 1;
                                                }

                                        if ((($splitSamArray[2] eq "$peakChrom") && ($splitSamArray[3] > "
                                            $highLimit")) || (($splitSamArray[2] ne "$peakChrom") && (
                                            $splitSamArray[3] < "$lowLimit")))
                                                {
                                                        $reCalc = 1;
                                                        $currentValue += 1;
                                                        $peaksCompleted += 1;
                                                        #print STDERR "$peaksCompleted peaks completed\n";
                                                }
                                }
                        }
                close SAMFILE;
                $currentXCoord = -1*$distanceFromPeakCentre;
                $totalPositions = ($distanceFromPeakCentre*2) + 1;
                for ($s = 0; $s < $totalPositions; $s += 1)
                        {
                                $meanGlobalBin[$s] = $globalBin[$s]/$peaksCompleted;
                                print STDOUT "$currentXCoord\t$meanGlobalBin[$s]\n";
                                $currentXCoord += 1;
                        }
                print STDERR "$peaksCompleted peaks completed.\n\n";
        }

sub topPeaks
        {
                my $file = $_[0];
                my $numberToPrint = $_[1];
                my @line = ();
                my %largeSORTHash = ();
                my $tempList = "$file" . ".temp";
                my $tempCounter = 0;
                print STDERR "sorting $numberToPrint peaks in $file...\n\n";
                open (HASHSORT, "<$file");
                while (<HASHSORT>)
                        {
                                @line = split(" ", $_);
                                push @{$largeSORTHash{$line[4]}}, $_;
                                #$largeSORTHash{$line[4]} = "$_";
                        }

                open (TEMPLIST, ">$tempList");
                foreach my $key (reverse sort {$a <=> $b} keys %largeSORTHash)
                        {
                                print TEMPLIST @{$largeSORTHash{"key"}};
                        }
                close TEMPLIST;
                open (TEMPLIST, "<$tempList");
                while (<TEMPLIST>)
                        {
                                print STDOUT "$_";
                                $tempCounter += 1;
                                unless ($numberToPrint eq "ALL")
                                        {
                                                last if $tempCounter == $numberToPrint;
                                        }
                        }
                close TEMPLIST;
                unlink $tempList;
        }

sub samNSort
        {
                my $file = $_[0];
                my @line = ();
                my %largeSORTHash = ();
                my $currentKey = 0;
                my $maxValue = 0;
                print STDERR "sorting $file...\n\n";
                open (HASHSORT, "<$file");
                while (<HASHSORT>)
                        {
                                @line = split(" ", $_);
                                push @{$largeSORTHash{$line[3]}}, $_;
                                $maxValue = $line[3] if $line[3] > $maxValue;
                        }
                my $number = keys %largeSORTHash;
                for ($currentKey = 0; $currentKey <= $maxValue; $currentKey += 1)
```

```
                        {
                                print STDOUT @{$largeSORTHash{"$currentKey"}} if defined $largeSORTHash{"
                                    $currentKey"};
                        }
        }

sub gaussEngine
        {
                my $in = "@_";
                my $denominator = 1/sqrt(2*3.14159265*$sigma**2);
                my $sigmaSubGEsquaredTwice = ($sigma*$sigma)*2;
                my $power = -($in**2)/$sigmaSubGEsquaredTwice;
                my $eToPower = 2.7182818284590451**$power;
                my $gaussed= $eToPower*$denominator;
                my $roundGaussed = sprintf("%.5f", $gaussed);
                return $roundGaussed;
        }

sub baseCountGenomeSizeFile
        {
                my $gSizeFile = $_[0];
                my $gTotalSize = 0;
                my @gline = ();
                open (GSIZE, "<$gSizeFile");
                while (<GSIZE>)
                        {
                                @gline = split(" ", $_);
                                $gTotalSize += $gline[1];
                        }
                return $gTotalSize;
        }

sub fileTest
        {
                my $format = $_[0];
                my $file = $_[1];
                my $headerOff = 0;
                my @line;
                my @sameLine;
                my $samFORMAT;
                my $samDisorder;
                my $samOrder = 0;
                my $header;
                my $samCount = 0;
                my $chrOrder;
                my $chrDisorder;
                my %samTags = ('@CO' => 1, '@RG' => 1, '@SQ' => 1, '@HD' => 1, '@PG' => 1);

                ### defining SAM format
if ($format eq 'SAM')
        {
                $header = 'NONE';
                $samFORMAT = 'OK';
                $samDisorder = 'OK';
                $chrDisorder = 'OK';

                if (-z $file)
                        {
                                $samFORMAT = 'BAD';
                                $samDisorder = 'BAD';
                                $header = 'NONE';
                                $chrDisorder = 'BAD';
                        }
                        open (my $samFile, "<", $file) or die "Unable to open $file.\n\n";
                        while (<$samFile>)
                                        {
                                                next unless $_ =~ /\w/;
                                                if (eof)
                                                        {
                                                                $samFORMAT = 'BAD';
                                                                $samDisorder = 'BAD';
                                                                $header = 'NONE';
                                                                $chrDisorder = 'BAD'
                                                        }
                                                if ($headerOff == 0)
                                                        {
                                                                @line = split(" ", $_);
                                                                $header = 'OK' if defined $samTags{$line[0]};
                                                                $headerOff = 1 unless exists $samTags{$line[0]};
                                                                #print STDERR "Header off = $headerOff\n";
                                                                #$headerOff = 1 if !defined $line[0];
                                                        }
                                                if ($headerOff == 1)
                                                        {
                                                                @sameLine = split(" ", $_);
                                                                $samFORMAT = 'BAD' unless defined $sameLine[7];
                                                                if (defined $sameLine[7])
                                                                        {
                                                                                if (
                                                                                        ($sameLine[0] =~ /[^!-?A
                                                                                            -~]/) ||
                                                                                        ($sameLine[1] =~ /\D/) ||
                                                                                        ($sameLine[2] !~ /chr/) ||
                                                                                        ($sameLine[3] =~ /\D/) ||
                                                                                        ($sameLine[4] =~ /\W/) ||
                                                                                        ($sameLine[5] =~ /\W/) ||
                                                                                        ($sameLine[6] =~ /[^!-?A
                                                                                            -~]/) ||
                                                                                        ($sameLine[7] =~ /[^!-?A
                                                                                            -~]/) ||
                                                                                        ($#sameLine < 8)
                                                                                )
```

```
                                                                                {
                                                                                        $samFORMAT = 'BAD';
                                                                                }
                                                                        }

                                                $samDisorder = 'BAD' unless defined $sameLine[2];
                                                $chrDisorder = 'BAD' unless defined $sameLine[3];
                                                if ((defined $sameLine[3]) && ($samCount >= 1))
                                                        {
                                                if ($sameLine[3] =~ /\d/)
                                                        {
                                                                unless ($sameLine[3] >= $samOrder)
                                                                        {
                                                                                $samDisorder = 'BAD' unless $sameLine[2] ne
                                                                                        $chrOrder;
                                                                        }
                                                        }
                                                else
                                                        {
                                                                $samDisorder = 'BAD'
                                                        }
                                                }

                                        if ((defined $sameLine[2]) && ($samCount >= 1))
                                                {
                                                        unless ($sameLine[2] eq $chrOrder)
                                                        {
                                                                $chrDisorder = 'BAD';
                                                        }
                                                }
                                        $samOrder = $sameLine[3] if defined $sameLine[3];
                                        $chrOrder = $sameLine[2] if defined $sameLine[2];
                                        $samCount += 1;
                                        last if $samCount == 20;
                                }
}
        return ($samFORMAT, $samDisorder, $header, $chrDisorder);
        }
                if ($format eq 'BED')
                        {
                                my $bedFORMAT = 'OK';
                                my $bedCount = 0;

                                if (-z $file)
                                        {
                                                $bedFORMAT = 'BAD';
                                        }
                                open (my $bedFile, "<", $file) or die "Unable to open $file\n\n";

                                while (<$bedFile>)
                                        {
                                                @line = split (" ", $_);

                                                if (eof)
                                                        {
                                                                $bedFORMAT = 'BAD';
                                                        }
                                                if (
                                                                ($line[0] !~ /chr/) ||
                                                                ($line[1] =~ /\D/) ||
                                                                ($line[2] =~ /\D/)
                                                        )
                                                                {
                                                                        $bedFORMAT = 'BAD';
                                                                }

                                                $bedCount += 1;
                                                last if $bedCount == 2;
                                        }
                                return ($bedFORMAT);
                        }

                if ($format eq 'CHROM')
                        {
                                my $chromFORMAT = 'OK';
                                my $chromCount = 0;
                                if (-z $file)
                                        {
                                                $chromFORMAT = 'BAD';
                                        }
                                open (my $chromFile, "<", $file) or die "Unable to open $file\n\n";

                                while (<$chromFile>)
                                        {
                                                @line = split (" ", $_);

                                                if (eof)
                                                        {
                                                                $chromFORMAT = 'BAD';
                                                        }
                                                if (
                                                                ($line[0] =~ /[^!-?A-~]/) ||
                                                                ($line[1] =~ /\D/) ||
                                                                ($#line != 1)
                                                        )
                                                                {
                                                                        $chromFORMAT = 'BAD';
                                                                }
                                                $chromCount += 1;
                                                last if $chromCount == 1;
                                        }
                                return ($chromFORMAT);
```

```
                        }
        }

sub usage
        {
                print STDOUT "
Program: PeakHunter v2.0
Version: 2.0 [July 2013]

Usage   PeakHunter.pl [Function] <targetfile> [parameter] <value>> [resultsFileName]

Functions and parameters

-NS (Numerical Sam Sorter) <mappedReads.sam>
\tSorts sam format reads by base start position, irrespective of chromosome.

-F (Sam Filter) <mappedReads.sam>
\tOrders sam format reads by chromosome, as indicated in chromosome size text file of format:

\t[chrA]\t[chrA size]
\t[chrB]\t[chrB size]
\t[....]\t[.........]

\t-g (chromosome size file; mandatory) <path>
\t\tpath to text file containing list of chromosome names/sizes.
\t-q (mapq score cut off) <value>
\t\tReads with mapq score less than this value will be excluded. Default value 0.
\t-u (uq mismatch score cut off) <value>
\t\tReads with uq base mismatch scores greater than this value will be excluded. Default value 10000.
\t-i (include header; optional) <headerFileName>
\t\tIncludes alternative sam file header in new file.
\t-PCR (exclude replicate reads; default OFF) <ON | OFF>
\t\tIncludes alternative sam file header in new file.

-R (Random Read Selector) <mappedReads.sam>
\tRandomly selects a given number of reads from a sam file.
\t-nr (number of reads to select; mandatory) <value>

-D (Density Analyzer) <orderedReads.sam>
\tcreates a fixedStep, step=1 wig density track from ordered sam file input

\t-g (chromosome size file; mandatory) <path>
\t\tPath to text file containing list of chromosome names/sizes.
\t-t (minimum bin read threshold) <value>
\t\tRead number threshold for exclusion from density calculation. Default calculated from file.
\t-m (maximum bin read threshold) <value>
\t\tExcludes bins containing greater than <value> reads from density calculation. Default value 100000.
\t-n (one-sided bin tail size) <value>
\t\tHalf the size of Gaussian bin. Default value 180.
\t-d (sd of PDF) <value>
\t\tDetermines the shape of the Gaussian distribution. Default value 50.
\t-STEP (bin step size; default 100) <value>
\t\tstep-size in numbers of bases that the sampling bin takes in moving along the genome.
\t-o (track offset) <value>
\t\tOffsets the entire wig track by <value> base positions

-P (Peak Caller) <orderedReads.sam>
\tCalculates the density threshold and calls peaks.

\t-g (chromosome size file; mandatory) <path>
\t\tPath to text file containing list of chromosome names/sizes.
\t-b (genome chromatin blue-print file) <masterTrack.bed>
\t\tA bed file listing known open chromatin sites, and the number of cell types known to have open chromatin at
      that site.
\t-bin (central density bin; default 300bp) <score>
\t\tManually adjusts the bin size.
\t-back (local background density bin; default 3000bp) <score>
\t\tManually adjusts the size of the local background bin for central bin score normalization.
\t-FLAT (flat threshold) <value>
\t\tSets a flat threshold for peak calling. Supersedes other threshold settings.
\t-npBack (number of sites do estimate background; default 50,000) <value>
\t\tBackground signal is estimated by random sampling of signal densities. -npBack sets the number of sites to be
      sampled in this estimate.
\t-sig (significance level; default p< 0.001) <value>
\t\tSets the p value from the KDE background density probability distribution, below which densities are considered
      significantly different from background densities.
\t-STEP (bin step size; default 100) <value>
\t\tstep-size in numbers of bases that the sampling bin takes in moving along the genome.
\t-PVAL (p value based peak scoring; default OFF) <ON | OFF>
\t\tBy default, called peaks are scored according to their maximum density. Setting -PVAL to ON replaces these
      scores with their equivalent p value from the background probability distribution.

-T (Top Peak Selector) <peakFile.bed>
\tOrders peaks in bed file from highest maximum density to lowest.

\t-n (number of peaks) <value>
\t\tSets the number of peaks to print from target file.

-S (Mean Signal Intensity) <peakFile.bed>
\tcalculates the mean reads per base score in sam file surrounding randomly selected peak summits.

\t-np (number of peaks) <value>
\t\tSets the approximate number of peaks to randomly select.
\t-dc (distance from centre) <value>
\t\tSets the distance (bases) either side of peak summit to calculate mean rpb score.
\t-t (target sam file) <mappedReads.sam>
\t\tSelects the sam file for mean rpb calculaton.\n\n"
        }
```

References

[1] Gary W Litman, John P Cannon, and Larry J Dishaw. **Reconstructing immune phylogeny: new perspectives.** *Nature reviews Immunology*, **5**(11):866–879, November 2005. 4

[2] Akie Sato, Werner E Mayer, Peter Overath, and Jan Klein. **Genes encoding putative natural killer cell C-type lectin receptors in teleostean fishes.** *Proceedings of the National Academy of Sciences of the United States of America*, **100**(13):7779–7784, June 2003. 4

[3] Benny P Shum, Laura R Flodin, David G Muir, Raja Rajalingam, Salim I Khakoo, Sophia Cleland, Lisbeth A Guethlein, Markus Uhrberg, and Peter Parham. **Conservation and variation in human and common chimpanzee CD94 and NKG2 genes.** *Journal of immunology (Baltimore, Md : 1950)*, **168**(1):240–252, January 2002. 4

[4] M L Labonte and N L Letvin. **Variable NKG2 expression in the peripheral blood lymphocytes of rhesus monkeys**. *Clinical and experimental immunology*, **138**(2):205–212, 2004. 4

[5] R Biassoni, M Fogli, C Cantoni, P Costa, R Conte, G Koopman, A Cafaro, B Ensoli, A Moretta, L Moretta, and A De Maria. **Molecular and functional characterization of NKG2D, NKp80, and NKG2C triggering NK cell receptors in rhesus and cynomolgus macaques: Monitoring of NK cell function during simian HIV infection**. *Journal of Immunology*, **174**(9):5695–5705, 2005. 4

[6] Anne Averdam, Heiner Kuhl, Mario Sontag, Tamara Becker, Austin L Hughes, Richard Reinhardt, and Lutz Walter. **Genomics and diversity of the common marmoset monkey NK complex.** *Journal of immunology (Baltimore, Md : 1950)*, **178**(11):7151–7161, June 2007. 4

[7] D Yim, H B Jie, J Sotiriadis, Y S Kim, K S Kim, M F Rothschild, L L Lanier, and Y B Kim. **Molecular cloning and characterization of pig immunoreceptor DAP10 and NKG2D.** *Immunogenetics*, **53**(3):243–249, April 2001. 4

[8] Y Fikri, J Nyabenda, J Content, and K Huygen. **Cloning, sequencing, and cell surface expression pattern of bovine immunoreceptor NKG2D and adaptor molecules DAP10 and DAP12**. *Immunogenetics*, **59**(8):653–659, 2007. 4

[9] D W Wolan, L Teyton, M G Rudolph, B Villmow, S Bauer, D H Busch, and I A Wilson. **Crystal structure of the murine NK cell-activating receptor NKG2D at 1.95 Å**. *Nature immunology*, **2**(3):248–254, 2001. 4

[10] Megan A Cooper, Julie M Elliott, Peter A Keyel, Liping Yang, Javier A Carrero, and Wayne M Yokoyama. **Cytokine-induced memory-like natural killer cells**. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(6):1915–1919, 2009. 4

[11] R Romee, S E Schneider, J W Leong, J M Chase, C R Keppel, R P Sullivan, M A Cooper, and T A Fehniger. **Cytokine activation induces human memory-like NK cells**. *Blood*, **120**(24):4751–4760, 2012. 4

[12] I C MacLennan and G Loewi. **Effect of specific antibody to target cells on their specific and non-specific interactions with lymphocytes.** *Nature*, **219**(5158):1069–1070, September 1968. 5

[13] A H GREENBERG, L HUDSON, L SHEN, and I M ROITT. **Antibody-dependent cell-mediated cytotoxicity due to a "null" lymphoid cell**. *Nature*, **242**(117):111–113, 1973. 5

[14] J P Houchins, T Yabe, C McSherry, N Miyokawa, and F H Bach. **Isolation and characterization of NK cell or NK/T cell-specific cDNA clones**. *Journal of Molecular and Cellular Immunology*, **4**(6):295–306, 1990. 5

[15] J P Houchins, T Yabe, C McSherry, and F H Bach. **DNA sequence analysis of NKG2, a family of related cDNA clones encoding type II integral membrane proteins on human natural killer cells.** *The Journal of experimental medicine*, **173**(4):1017–1020, April 1991. 5

[16] S Bahram, M Bresnahan, DE Geraghty, and T Spies. **A second lineage of mammalian major histocompatibility complex class I genes**. *Proceedings of the National Academy of Sciences of the United States of America*, **91**(14):6259, 1994. 5, 6

[17] V Groh, A Steinle, S Bauer, and T Spies. **Recognition of stress-induced MHC molecules by intestinal epithelial gammadelta T cells.** *Science (New York, NY)*, **279**(5357):1737–1740, March 1998. 5

[18] S Bauer, V Groh, A Steinle, J H Phillips, L L Lanier, and T Spies. **Activation of NK cells and T cells by NKG2D, a receptor for stress-inducible MICA**. *Science (New York, NY)*, **285**(5428):727–729, July 1999. 6, 8, 10

[19] D Cosman, J Müllberg, CL Sutherland, and W Chin. **ULBPs, novel MHC class I–related molecules, bind to CMV glycoprotein UL16 and stimulate NK cytotoxicity through the NKG2D receptor**. *Immunity*, 2001. 6, 10

[20] N J Chalupny, C.L. Sutherland, W A Lawrence, A. Rein-Weston, and D Cosman. **ULBP4 is a novel ligand for human NKG2D**. *Biochemical and Biophysical Research Communications*, **305**(1):129–135, 2003. 6

[21] Wei Cao, Xueyan Xi, Zhun Wang, Liling Dong, Zhiyong Hao, Lianxian Cui, Chi Ma, and Wei He. **Four novel ULBP splice variants are ligands for human NKG2D.** *International immunology*, **20**(8):981–991, August 2008. 6

[22] L. Bacon, R.A. Eagle, M. Meyer, N. Easom, N.T. Young, and J Trowsdale. **Two human ULBP/RAET1 molecules with transmembrane regions are ligands for NKG2D**. *Journal of Immunology*, **173**(2):1078–1084, 2004. 6

[23] R.A. Eagle, J.A. Traherne, J.R. Hair, I. Jafferji, and J Trowsdale. **ULBP6/RAET1L is an additional human NKG2D ligand**. *European Journal of Immunology*, **39**(11):3207–3216, 2009. 6, 164

[24] M Nomura, Z Zou, T Joh, Y Takihara, Y Matsuda, and K Shimada. **Genomic structures and characterization of Rae1 family members encoding GPI-anchored cell surface proteins and expressed predominantly in embryonic mouse brain**. *Journal of Biochemistry*, **120**(5):987–995, 1996. 6, 8

[25] A Cerwenka, A B H Bakker, T McClanahan, J Wagner, J. Wu, J H Phillips, and L L Lanier. **Retinoic acid early inducible genes define a ligand family for the activating NKG2D receptor in mice**. *Immunity*, **12**(6):721–727, 2000. 6, 8

[26] M Girardi, D E Oppenheim, C R Steele, J M Lewis, E Glusac, R Filler, P Hobby, B Sutton, R E Tigelaar, and A C Hayday. **Regulation of cutaneous malignancy by gammadelta T cells**. *Science (New York, NY)*, **294**(5542):605–609, October 2001. 6, 8

[27] Leonidas N Carayannopoulos, Olga V Naidenko, Daved H Fremont, and Wayne M Yokoyama. **Cutting edge: murine UL16-binding protein-like transcript 1: a newly described transcript encoding a high-affinity ligand for murine NKG2D.** *Journal of immunology (Baltimore, Md : 1950)*, **169**(8):4079–4083, October 2002. 6, 8

[28] A Takada, S Yoshida, M Kajikawa, Y Miyatake, U Tomaru, M Sakai, H Chiba, K Maenaka, D Kohda, and K Fugo. **Two novel NKG2D ligands of the mouse H60 family with differential expression patterns and binding affinities to NKG2D**. *The Journal of Immunology*, **180**(3):1678, 2008. 6, 9

[29] Mirjana Radosavljevic, Benoît Cuillerier, Michael J Wilson, Oliver Clément, Sophie Wicker, Susan Gilfillan, Stephan Beck, John Trowsdale, and Seiamak Bahram. **A cluster of ten novel MHC class I related genes on human chromosome 6q24.2-q25.3.** *Genomics*, **79**(1):114–123, January 2002. 6

[30] M Ohashi, R.A. Eagle, and J Trowsdale. **Post-translational modification of the NKG2D ligand RAET1G leads to cell surface expression of a glycosylphosphatidylinositol-linked isoform**. *Journal of Biological Chemistry*, **285**(22):16408–16415, 2010. 6

[31] M. Champsaur and L L Lanier. **Effect of NKG2D ligand expression on host immune responses**. *Immunological reviews*, **235**(1):267–285, 2010. 6, 23

[32] P Li, D L Morris, B E Willcox, A Steinle, T Spies, and R K Strong. **Complex structure of the activating immunoreceptor NKG2D and its MHC class I-like ligand MICA**. *Nature immunology*, **2**(5):443–451, May 2001. 7, 8

[33] J. Wu, Y Song, A B H Bakker, S Bauer, T Spies, L L Lanier, and J H Phillips. **An activating immunoreceptor complex formed by NKG2D and DAP10**. *Science (New York, NY)*, **285**(5428):730–732, 1999. 7

[34] K Jiang, B Zhong, D L Gilvary, B C Corliss, E Hong-Geller, S Wei, and J Y Djeu. **Pivotal role of phosphoinositide-3 kinase in regulation of cytotoxicity in natural killer cells.** *Nature immunology*, **1**(5):419–425, November 2000. 7

[35] J L Upshaw, L N Arneson, R A Schoon, C J Dick, D D Billadeau, and P J Leibson. **NKG2D-mediated signaling requires a DAP10-bound Grb2-Vav1 intermediate and phosphatidylinositol-3-kinase in human natural killer cells**. *Nature immunology*, **7**(5):524–532, 2006. 7

[36] Brian Rabinovich, Jennifer Li, Martin Wolfson, William Lawrence, Courtney Beers, Jan Chalupny, Rose Hurren, Brad Greenfield, Richard Miller, and David Cosman. **NKG2D splice variants: a reexamination of adaptor molecule associations.** *Immunogenetics*, **58**(2-3):81–88, April 2006. 7

[37] Lucie Novakova, Zuzana Nevoralova, and Jan Novak. **Innate-like behavior of human invariant natural killer T cells during herpes simplex virus infection.** *Cellular immunology*, **278**(1-2):16–20, July 2012. 8

[38] Andrea Sáez-Borderías, Mónica Gumá, Ana Angulo, Beatriz Bellosillo, Daniela Pende, and Miguel López-Botet. **Expression and function of NKG2D in CD4+ T cells specific for human cytomegalovirus.** *European Journal of Immunology*, **36**(12):3198–3206, December 2006. 8

[39] M Komatsu-Wakui, K Tokunaga, Y Ishikawa, C Leelayuwat, K Kashiwase, H Tanaka, S Moriyama, F Nakajima, M H Park, G J Jia, N O Chimge, E W Sideltseva, and T Juji. **Wide distribution of the MICA-MICB null haplotype in East Asians**. *Tissue antigens*, **57**(1):1–8, 2001. 8

[40] P Li, S T Willie, S Bauer, D L Morris, T Spies, and R K Strong. **Crystal structure of the MHC class I homolog MIC-A, a gammadelta T cell ligand.** *Immunity*, **10**(5):577–584, May 1999. 8

[41] MA Holmes, PW Li, EW Petersdorf, and RK Strong. **Structural studies of allelic diversity of the MHC class I homolog MIC-B, a stress-inducible ligand for the activating immunoreceptor NKG2D**. *Journal of immunology (Baltimore, Md : 1950)*, **169**(3):1395–1400, 2002. 8

[42] Steffen Müller, Georg Zocher, Alexander Steinle, and Thilo Stehle. **Structure of the HCMV UL16-MICB complex elucidates select binding of a viral immunoevasin to diverse NKG2D ligands.** *PLoS pathogens*, **6**(1):e1000723, January 2010. 8

[43] A V Romphruk, A Romphruk, T.K. Naruse, S Raroengjai, C Puapairoj, H Inoko, and C Leelayuwat. **Polymorphisms of NKG2D ligands: Diverse RAET1/ULBP genes in Northeastern Thais**. *Immunogenetics*, **61**(9):611–617, 2009. 8

[44] S Radaev, B Rostro, A G Brooks, M. Colonna, and P D Sun. **Conformational plasticity revealed by the cocrystal structure of NKG2D and its class i MHC-like ligand ULBP3**. *Immunity*, **15**(6):1039–1049, 2001. 8

[45] CA O'Callaghan, A Cerwenka, BE Willcox, LL Lanier, and PJ Bjorkman. **Molecular competition for NKG2D: H60 and RAE1 compete unequally for NKG2D with dominance of H60**. *Immunity*, **15**(2):201–211, 2001. 9

[46] LN Carayannopoulos, OV Naidenko, J Kinder, EL Ho, DH Fremont, and WM Yokoyama. **Ligands for murine NKG2D display heterogeneous binding behavior**. *European Journal of Immunology*, **32**(3):597–605, 2002. 9

[47] LL Lanier. **Up on the tightrope: natural killer cell activation and inhibition**. *Nature immunology*, **9**(5):495–502, 2008. 10

[48] D Pende, C Cantoni, P Rivera, M Vitale, R Castriconi, S Marcenaro, M Nanni, R Biassoni, C Bottino, A Moretta, and L Moretta. **Role of NKG2D in tumor cell lysis mediated by human NK cells: cooperation with natural cytotoxicity receptors and capability of recognizing tumors of nonepithelial origin.** *European Journal of Immunology*, **31**(4):1076–1086, April 2001. 10

[49] B Rincon-Orozco, V Kunzmann, P Wrobel, D Kabelitz, A Steinle, and T Herrmann. **Activation of V gamma 9V delta 2 T cells by NKG2D**. *The Journal of Immunology*, **175**(4):2144, 2005. 10, 12

[50] Yan Y Kong, Wei W Cao, Xueyan X Xi, Chi C Ma, Lianxian L Cui, and Wei W He. **The NKG2D ligand ULBP4 binds to TCRgamma9/delta2 and induces cytotoxicity to tumor cells through both TCRgammadelta and NKG2D.** *Blood*, **114**(2):310–317, July 2009. 10, 12, 13

[51] C Maccalli, D Pende, C Castelli, MC Mingari, PF Robbins, and G Parmiani. **NKG2D engagement of colorectal cancer-specific T cells strengthens TCR-mediated antigen stimulation and elicits TCR independent anti-tumor activity**. *European Journal of Immunology*, **33**(7):2033–2043, 2003. 10

[52] Michael R Verneris, Mobin Karami, Jeanette Baker, Anishka Jayaswal, and Robert S Negrin. **Role of NKG2D signaling in the cytotoxicity of activated and expanded CD8+ T cells.** *Blood*, **103**(8):3065–3072, April 2004. 10

[53] Matthieu Allez, Vannary Tieng, Atsushi Nakazawa, Xavier Treton, Vincent Pacault, Nicolas Dulphy, Sophie Caillat Zucman, Pascale Paul, Jean-Marc Gornet, Corinne Douay, Sophie Ravet, Ryad Tamouza, Dominique Charron, Marc Lémann, Lloyd Mayer, and Antoine Toubert. **CD4(+)NKG2D(+) T cells in Crohn's disease mediate inflammatory and cytotoxic responses through MICA interactions**. *Gastroenterology*, **132**(7):2346–2358, 2007. 10

[54] Zhenpeng Dai, Cameron J Turtle, Garrett C Booth, Stanley R Riddell, Theodore A Gooley, Anne M Stevens, Thomas Spies, and Veronika Groh. **Normally occurring NKG2D+CD4+ T cells are immunosuppressive and inversely correlated with disease activity in juvenile-onset lupus.** *The Journal of experimental medicine*, **206**(4):793–805, April 2009. 10

[55] R Alonso-Arias, M A Moro-García, A López-Vázquez, L Rodrigo, J Baltar, F M S García, J J S Jaurrieta, and C López-Larrea. **NKG2D expression in CD4+ T lymphocytes as a marker of senescence in the aged immune system**. *Age*, **33**(4):591–605, 2011. 10

[56] P André, R Castriconi, M Espéli, N Anfossi, T Juarez, S Hue, H Conway, F Romagné, A Dondero, M Nanni, S Caillat-Zucman, D H Raulet, C Bottino, E Vivier, A Moretta, and P Paul. **Comparative analysis of human NK cell activation induced by NKG2D and natural cytotoxicity receptors**. *European Journal of Immunology*, **34**(4):961–971, 2004. 12

[57] M Kubin, L Cassiano, J Chalupny, W Chin, D Cosman, W Fanslow, J Müllberg, A M Rousseau, D Ulrich, and R Armitage. **ULBP1, 2, 3: novel MHC class I-related molecules that bind to human cytomegalovirus glycoprotein UL16, activate NK cells.** *European Journal of Immunology*, **31**(5):1428–1437, May 2001. 12

[58] YT Bryceson, ME March, HG Ljunggren, and EO Long. **Synergy among receptors on resting NK cells for the activation of natural cytotoxicity and cytokine secretion**. *Blood*, **107**(1):159, 2006. 12

[59] C Fauriat, EO Long, HG Ljunggren, and YT Bryceson. **Regulation of human NK-cell cytokine and chemokine production by target cell recognition**. *Blood*, **115**(11):2167, 2010. 12

[60] V Groh, R Rhinehart, J Randolph-Habecker, M S Topp, S R Riddell, and T Spies. **Costimulation of CD8alphabeta T cells by NKG2D via engagement by MIC induced on virus-infected cells**. *Nature immunology*, **2**(3):255–260, March 2001. 12, 13, 36, 163

[61] AI Roberts, L Lee, E Schwarz, V Groh, T Spies, EC Ebert, and B Jabri. **Cutting edge: NKG2D receptors induced by IL-15 costimulate CD28-negative effector CTL in the tissue microenvironment**. *The Journal of Immunology*, **167**(10):5527, 2001. 12

[62] AI Marusina, JE Coligan, and F Borrego. **Cutting edge: NKG2D is a costimulatory receptor for human naive CD8+ T cells**. *The Journal of Immunology*, **174**(8):4480, 2005. 12

[63] L.I.R. Ehrlich, K Ogasawara, J.A. Hamerman, R Takaki, A Zingoni, J.P. Allison, and L L Lanier. **Engagement of NKG2D by cognate ligand or antibody alone is insufficient to mediate costimulation of human and mouse CD8+ T cells**. *Journal of Immunology*, **174**(4):1922–1931, 2005. 12

[64] K Rajasekaran, V Xiong, L Fong, J Gorski, and S Malarkannan. **Functional dichotomy between NKG2D and CD28-mediated co-stimulation in human CD8+ t cells**. *PloS one*, **5**(9):1–10, 2010. 12

[65] Stefan A Welte, Christian Sinzger, Stefan Z Lutz, Harpreet Singh-Jasuja, Kerstin Laib Sampaio, Ute Eknigk, Hans-Georg Rammensee, and Alexander Steinle. **Selective intracellular retention of virally induced NKG2D ligands by the human cytomegalovirus UL16 glycoprotein.** *European Journal of Immunology*, **33**(1):194–203, January 2003. 13, 163

[66] Robert A Eagle, James A Traherne, Omodele Ashiru, Mark R Wills, and John Trowsdale. **Regulation of NKG2D ligand gene expression**. *Human immunology*, **67**(3):159–169, 2006. 13, 36

[67] Alexander Rölle, Mehrdad Mousavi-Jazi, Mikael Eriksson, Jenny Odeberg, Cecilia Söderberg-Nauclér, David Cosman, Klas Kärre, and Cristina Cerboni. **Effects of human cytomegalovirus infection on ligands for the activating NKG2D receptor of NK cells: up-regulation of UL16-binding protein (ULBP)1 and ULBP2 is counteracted by the viral UL16 protein.** *Journal of immunology (Baltimore, Md : 1950)*, **171**(2):902–908, July 2003. 13, 36, 163, 164

[68] Mar Valés-Gómez, Helena Browne, and Hugh T Reyburn. **Expression of the UL16 glycoprotein of Human Cytomegalovirus protects the virus-infected cell from attack by natural killer cells**. *BMC immunology*, **4**(1):4, 2003. 13

[69] Isabel Y Pappworth, Eddie C Wang, and Martin Rowe. **The switch from latent to productive infection in epstein-barr virus-infected B cells is associated with sensitization to NK cell killing.** *Journal of Virology*, **81**(2):474–482, January 2007. 13

[70] Jeffrey Ward, Matthew Bonaparte, Jennifer Sacks, Jacqueline Guterman, Manuela Fogli, Domenico Mavilio, and Edward Barker. **HIV modulates the expression of ligands important in triggering natural killer cell cytotoxic responses on infected primary T-cell blasts.** *Blood*, **110**(4):1207–1214, August 2007. 13

[71] Jeffrey Ward, Zachary Davis, Jason DeHart, Erik Zimmerman, Alberto Bosque, Enrico Brunetta, Domenico Mavilio, Vicente Planelles, and Edward Barker. **HIV-1 Vpr triggers natural killer cell-mediated lysis of infected cells through activation of the ATR-mediated DNA damage response.** *PLoS pathogens*, **5**(10):e1000613, October 2009. 13, 26, 153, 158

[72] Jonathan Richard, Sardar Sindhu, Tram N Q Pham, Jean-Philippe Belzile, and Eric A Cohen. **HIV-1 Vpr up-regulates expression of ligands for the activating NKG2D receptor and promotes NK cell-mediated killing.** *Blood*, **115**(7):1354–1363, February 2010. 13

[73] Monia Draghi, Achal Pashine, Bharati Sanjanwala, Ketevan Gendzekhadze, Claudia Cantoni, David Cosman, Alessandro Moretta, Nicholas M. Valiante, and Peter Parham. **NKp46 and NKG2D recognition of infected dendritic cells is necessary for NK cell activation in the human response to influenza infection**. *Journal of immunology (Baltimore, Md : 1950)*, **178**(5):2688–2698, 2007. 13

[74] Brian P McSharry, Hans-Gerhard Burgert, Douglas P Owen, Richard J Stanton, Virginie Prod'homme, Martina Sester, Katja Koebernick, Veronika Groh, Thomas Spies, Steven Cox, Ann-Margaret Little, Eddie C Y Wang, Peter Tomasec, and Gavin W G Wilkinson. **Adenovirus E3/19K promotes evasion of NK cell recognition by intracellular sequestration of the NKG2D ligands major histocompatibility complex class I chain-related proteins A and B.** *Journal of Virology*, **82**(9):4585–4594, May 2008. 13

[75] K Takeda, T Kaisho, and S Akira. **Toll-like receptors**. *Annual review of immunology*, **21**:335–376, 2003. 14, 23

[76] R Medzhitov, P Preston-Hurlburt, and C A Janeway. **A human homologue of the Drosophila Toll protein signals activation of adaptive immunity.** *Nature*, **388**(6640):394–397, July 1997. 14

[77] Mercedes Kloss, Patrice Decker, Katrin M Baltz, Tina Baessler, Gundram Jung, Hans-Georg Rammensee, Alexander Steinle, Matthias Krusch, and Helmut R Salih. **Interaction of monocytes with NK cells upon Toll-like receptor-induced expression of the NKG2D ligand MICA.** *The Journal of Immunology*, **181**(10):6711–6719, November 2008. 14, 24

[78] Philipp Eissmann, J Henry Evans, Maryam Mehrabi, Emma L Rose, Shlomo Nedvetzki, and Daniel M Davis. **Multiple mechanisms downstream of TLR-4 stimulation allow expression of NKG2D ligands to facilitate macrophage/NK cell crosstalk.** *The Journal of Immunology*, **184**(12):6901–6909, June 2010. 14, 24, 26, 30, 158

[79] Satarupa Basu, Mikael Eriksson, Patricia A Pioli, Jose Conejo-Garcia, Teddy F Mselle, Satoshi Yamamoto, Charles R Wira, and Charles L Sentman. **Human uterine NK cells interact with uterine macrophages via NKG2D upon stimulation with PAMPs.** *American journal of reproductive immunology (New York, N.Y. : 1989)*, **61**(1):52–61, January 2009. 14

[80] Shlomo Nedvetzki, Stefanie Sowinski, Robert A Eagle, James Harris, Frédéric Vély, Daniela Pende, John Trowsdale, Eric Vivier, Siamon Gordon, and Daniel M Davis. **Reciprocal regulation of human natural killer cells and macrophages associated with distinct immune synapses.** *Blood*, **109**(9):3776–3785, May 2007. 14

[81] Luciana L Molinero, Mercedes B Fuertes, Gabriel A Rabinovich, Leonardo Fainboim, and Norberto W Zwirner. **Activation-induced expression of MICA on T lymphocytes involves engagement of CD3 and CD28**. *Journal of leukocyte biology*, **71**(5):791–797, May 2002. 14, 25

[82] B A Rabinovich, J Li, J Shannon, R Hurren, J Chalupny, D Cosman, and R G Miller. **Activated, but not resting, T cells can be recognized and killed by syngeneic NK cells**. *Journal of Immunology*, **170**(7):3572–3576, 2003. 14

[83] C Cerboni, A Zingoni, M Cippitelli, M Piccoli, L Frati, and A Santoni. **Antigen-activated human T lymphocytes express cell-surface NKG2D ligands via an ATM/ATR-dependent mechanism and become susceptible to autologous NK-cell lysis**. *Blood*, **110**(2):606, 2007. 14, 25, 27, 36, 228, 229

[84] Cristina Cerboni, Michele Ardolino, and Angela Santoni. **Detuning CD8+ T lymphocytes by down-regulation of the activating receptor NKG2D: role of NKG2D ligands released by activated T cells.** *Blood*, **113**(13):2955–2964, March 2009. 14

[85] YZ Zou, F Mirbaha, and P Stastny. **Contact inhibition causes strong downregulation of expression of MICA in human fibroblasts and decreased NK cell killing**. *Human immunology*, **67**(3):183–187, 2006. 15, 25, 36, 228

[86] R.W. McGilvray, R.A. Eagle, N.F.S. Watson, A. Al-Attar, G. Ball, I. Jafferji, J Trowsdale, and L.G. Durrant. **NKG2D ligand expression in human colorectal cancer reveals associations with prognosis and evidence for immunoediting**. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **15**(22):6993–7002, 2009. 15, 16, 21, 36

[87] Xiaohui Duan, Langmei Deng, Xiong Chen, Yebin Lu, Qi Zhang, Kejing Zhang, Yongjun Hu, Jie Zeng, and Weijia Sun. **Clinical significance of the immunostimulatory MHC class I chain-related molecule A and NKG2D receptor on NK cells in pancreatic cancer.** *Medical Oncology*, **28**(2):466–474, June 2011. 16, 36

[88] Masahisa Jinushi, Tetsuo Takehara, Tomohide Tatsumi, Tatsuya Kanto, Veronika Groh, Thomas Spies, Ritsuko Kimura, Takuya Miyagi, Kiyoshi Mochizuki, Yutaka Sasaki, and Norio Hayashi. **Expression and role of MICA and MICB in human hepatocellular carcinomas and their regulation by retinoic acid**. *International journal of cancer Journal international du cancer*, **104**(3):354–361, April 2003. 16

[89] Kui Li, Masaki Mandai, Junzo Hamanishi, Noriomi Matsumura, Ayako Suzuki, Haruhiko Yagi, Ken Yamaguchi, Tsukasa Baba, Shingo Fujii, and Ikuo Konishi. **Clinical significance of the NKG2D ligands, MICA/B and ULBP2 in ovarian cancer: high expression of ULBP2 is an indicator of poor prognosis**. *Cancer Immunology, Immunotherapy*, **58**(5):641–652, September 2008. 16, 36

[90] R.W. McGilvray, R.A. Eagle, P. Rolland, I. Jafferji, J Trowsdale, and L.G. Durrant. **ULBP2 and RAET1E NKG2D ligands are independent predictors of poor prognosis in ovarian cancer patients**. *International journal of cancer Journal international du cancer*, **127**(6):1412–1420, 2010. 16, 36

[91] X Xu, G. Rao, M.J. Gaffud, H.G. Ding, G. Maki, H.-G. Klingemann, V Groh, T Spies, S Caillat-Zucman, P. Gattuso, J. Plate, and R.A. Prinz. **Clinicopathological significance of major histocompatibility complex class I-related chain A and B expression in thyroid cancer**. *Journal of Clinical Endocrinology & Metabolism*, **91**(7):2704–2712, 2006. 16

[92] Claudia S Vetter, Veronika Groh, Per thor Straten, Thomas Spies, Eva-B Bröcker, and Jürgen C Becker. **Expression of stress-induced MHC class I related chain molecules on human melanoma.** *The Journal of investigative dermatology*, **118**(4):600–605, April 2002. 16, 36

[93] C S Vetter, W Lieb, E-B Bröcker, and J C Becker. **Loss of nonclassical MHC molecules MIC-A/B expression during progression of uveal melanoma.** *British Journal of Cancer*, **91**(8):1495–1499, October 2004. 16

[94] J D Wu, L.M. Higgins, A Steinle, D Cosman, K Haugk, and S.R. Plymate. **Prevalent expression of the immunostimulatory MHC class I chain-related molecule is counteracted by shadding in prostate cancer**. *Journal of Clinical Investigation*, **114**(4):560–568, 2004. 16, 36

[95] Masahisa Jinushi, Matthew Vanneman, Nikhil C Munshi, Yu-Tzu Tai, Rao H Prabhala, Jerome Ritz, Donna Neuberg, Kenneth C Anderson, Daniel Ruben Carrasco, and Glenn Dranoff. **MHC class I chain-related protein A antibodies and shedding are associated with the progression of multiple myeloma.** *Proceedings of the National Academy of Sciences*, **105**(4):1285–1290, January 2008. 16

[96] HR Salih and HG Rammensee. **Cutting edge: down-regulation of MICA on human tumors by proteolytic shedding**. *The Journal of . . .*, 2002. 16, 32

[97] Nicolas Boissel, Delphine Rea, Vannary Tieng, Nicolas Dulphy, Manuel Brun, Jean-Michel Cayuela, Philippe Rousselot, Ryad Tamouza, Philippe Le Bouteiller, François-Xavier Mahon, Alexander Steinle, Dominique Charron, Hervé Dombret, and Antoine Toubert. **BCR/ABL oncogene directly controls MHC class I chain-related molecule A expression in chronic myelogenous leukemia**. *Journal of immunology (Baltimore, Md : 1950)*, **176**(8):5108–5116, April 2006. 16, 26

[98] Manuel A Friese, Michael Platten, Stefan Z Lutz, Ulrike Naumann, Steffen Aulwurm, Felix Bischof, Hans-Jörg Bühring, Johannes Dichgans, Hans-Georg Rammensee, Alexander Steinle, and Michael Weller. **MICA/NKG2D-mediated immunogene therapy of experimental gliomas.** *Cancer Research*, **63**(24):8996–9006, December 2003. 16

[99] Joo-Young Kim, Young-Ok Son, Soon-Won Park, Jae-Ho Bae, Joo Seop Chung, Hyung Hoi Kim, Byung-Seon Chung, Sun-Hee Kim, and Chi-Dug Kang. **Increase of NKG2D ligands and sensitivity to NK cell-mediated cytotoxicity of tumor cells by heat shock and ionizing radiation.** *Experimental & molecular medicine*, **38**(5):474–484, October 2006. 16

[100] Stephan Gasser, Sandra Orsulic, Eric J Brown, and David H Raulet. **The DNA damage pathway regulates innate immune system ligands of the NKG2D receptor**. *Nature*, **436**(7054):1186–1190, July 2005. 16, 27, 153, 158

[101] J H Dobloug, O Førre, T K Kvien, T Egeland, and M Degré. **Natural killer (NK) cell activity of peripheral blood, synovial fluid, and synovial tissue lymphocytes from patients with rheumatoid arthritis and juvenile rheumatoid arthritis.** *Annals of the Rheumatic Diseases*, **41**(5):490–494, October 1982. 17

[102] Kalle Söderström, Emily Stein, Paula Colmenero, Ulrich Purath, Ulf Müller-Ladner, Cristina Teixeira de Matos, Ingo H Tarner, William H Robinson, and Edgar G Engleman. **Natural killer cells trigger osteoclastogenesis and bone destruction in arthritis.** *Proceedings of the National Academy of Sciences*, **107**(29):13028–13033, July 2010. 17

[103] Veronika Groh, Anja Bruhl, Hani El-Gabalawy, J Lee Nelson, and Thomas Spies. **Stimulation of T cell autoreactivity by anomalous expression of NKG2D and its MIC ligands in rheumatoid arthritis.** *Proceedings of the National Academy of Sciences of the United States of America*, **100**(16):9452–9457, August 2003. 17

[104] Philippe Saikali, Jack P Antel, Jia Newcombe, Zhihong Chen, Mark Freedman, Manon Blain, Romain Cayrol, Alexandre Prat, Jeffery A Hall, and Nathalie Arbour. **NKG2D-mediated cytotoxicity toward oligodendrocytes suggests a mechanism for tissue injury in multiple sclerosis.** *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **27**(5):1220–1228, January 2007. 18

[105] Juan Luís Fernández-Morera, Sandra Rodríguez-Rodero, Carlos Lahoz, Alberto Tuñon, Aurora Astudillo, Olivia Garcia-Suarez, Jesús Martínez-Borra, Antonio López-Vázquez, Luis Rodrigo, Segundo González, and Carlos López-Larrea. **Soluble MHC class I chain-related protein B serum levels correlate with disease activity in relapsing-remitting multiple sclerosis.** *HIM*, **69**(4-5):235–240, April 2008. 18

[106] V Groh, S Bahram, S Bauer, A Herman, M Beauchamp, and T Spies. **Cell stress-regulated human major histocompatibility complex class I gene expressed in gastrointestinal epithelium**. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(22):12445–12450, October 1996. 18, 228

[107] Raffaella La Scaleia, Antonella Stoppacciaro, Salvatore Oliva, Stefania Morrone, Giovanni Di Nardo, Angela Santoni, Salvatore Cucchiara, and Gabriella Palmieri. **NKG2D/Ligand dysregulation and functional alteration of innate immunity cell populations in pediatric IBD.** *Inflammatory bowel diseases*, **18**(10):1910–1922, October 2012. 19

[108] Stine Kjellev, Claus Haase, Dorthe Lundsgaard, Birgitte Ursø, Ditte Tornehave, and Helle Markholst. **Inhibition of NKG2D receptor function by antibody therapy attenuates transfer-induced colitis in SCID mice**. *European Journal of Immunology*, **37**(5):1397–1406, May 2007. 19

[109] S Makino, K Kunimoto, Y Muraoka, Y Mizushima, K Katagiri, and Y Tochino. **Breeding of a non-obese, diabetic strain of mice.** *Jikken dobutsu. Experimental animals*, **29**(1):1–13, January 1980. 19

[110] Y Kanazawa, K Komeda, S Sato, S Mori, K Akanuma, and F Takaku. **Non-obese-diabetic mice: immune mechanisms of pancreatic beta-cell destruction.** *Diabetologia*, **27 Suppl**(s1):113–115, June 1984. 20

[111] Kouetsu Ogasawara, Jessica A Hamerman, Lauren R Ehrlich, Helene Bour-Jordan, Pere Santamaria, Jeffrey A Bluestone, and Lewis L Lanier. **NKG2D blockade prevents autoimmune diabetes in NOD mice.** *Immunity*, **20**(6):757–767, June 2004. 20

[112] Melanie Rodacki, Britta Svoren, Vincent Butty, Whitney Besse, Lori Laffel, Christophe Benoist, and Diane Mathis. **Altered natural killer cells in type 1 diabetic patients.** *Diabetes*, **56**(1):177–185, January 2007. 20

[113] C Åkesson, K Uvebrant, C Oderup, K Lynch, R A Harris, Å Lernmark, C D Agardh, and C M Cilio. **Altered natural killer (NK) cell frequency and phenotype in latent autoimmune diabetes in adults (LADA) prior to insulin deficiency**. *Clinical and experimental immunology*, **161**(1):48–56, 2010. 20

[114] Kim G Hankey, Cinthia B Drachenberg, John C Papadimitriou, David K Klassen, Benjamin Philosophe, Steven T Bartlett, Veronika Groh, Thomas Spies, and Dean L Mann. **MIC expression in renal and pancreatic allografts.** *Transplantation*, **73**(2):304–306, January 2002. 21

[115] Jim Kim, Catherine K Chang, Tracy Hayden, Feng-Chun Liu, Jonathan Benjamin, Jessica A Hamerman, Lewis L Lanier, and Sang-Mo Kang. **The activating immunoreceptor NKG2D and its ligands are involved in allograft transplant rejection.** *Journal of immunology (Baltimore, Md : 1950)*, **179**(10):6416–6420, November 2007. 21

[116] Elena Sánchez-Zapardiel, María J Castro-Panete, Marcela Castillo-Rama, Pablo Morales, David Lora-Pablos, Diana Valero-Hervás, Raquel Ruiz-García, Jacqueline Apaza, Paloma Talayero, Amado Andrés, José M Morales, and Estela Paz-Artal. **Harmful Effect of Preformed Anti-MICA Antibodies on Renal Allograft Evolution in Early Posttransplantation Period.** *Transplantation*, April 2013. 21

[117] Yizhou Zou, Peter Stastny, Caner Süsal, Bernd Döhler, and Gerhard Opelz. **Antibodies against MICA antigens and kidney-transplant rejection.** *The New England journal of medicine*, **357**(13):1293–1300, September 2007. 21

[118] Steven T ST Cox, Henry A F HA Stephens, Raymond R Fernando, Aliyye A Karasu, Mark M Harber, Alexander J AJ Howie, Stephen S Powis, Yizhou Y Zou, Peter P Stastny, J Alejandro JA Madrigal, and Ann-Margaret AM Little. **Major histocompatibility complex class I-related chain A allele mismatching, antibodies, and rejection in renal transplantation.** *HIM*, **72**(10):827–834, September 2011. 21

[119] T O'Sullivan, G P Dunn, D Y Lacoursiere, R D Schreiber, and J D Bui. **Cancer immunoediting of the NK group 2D ligand H60a**. *Journal of Immunology*, **187**(7):3538–3545, 2011. 22

[120] A Cerwenka, J L Baron, and L L Lanier. **Ectopic expression of retinoic acid early inducible-1 gene (RAE-1) permits natural killer cell-mediated rejection of a MHC class I-bearing tumor in vivo**. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(20):11521–11526, September 2001. 22

[121] A Diefenbach, E R Jensen, A M Jamieson, and D H Raulet. **Rae1 and H60 ligands of the NKG2D receptor stimulate tumour immunity**. *Nature*, **413**(6852):165–171, September 2001. 22

[122] Yoshihiro Hayakawa, Janice M Kelly, Jennifer A Westwood, Phillip K Darcy, Andreas Diefenbach, David Raulet, and Mark J Smyth. **Cutting edge: tumor rejection mediated by NKG2D receptor-ligand interaction is dependent upon perforin**. *Journal of immunology (Baltimore, Md : 1950)*, **169**(10):5377–5381, November 2002. 22

[123] N Guerra, YX Tan, NT Joncker, A Choy, F Gallardo, N Xiong, S Knoblaugh, D Cado, NR Greenberg, and DH Raulet. **NKG2D-deficient mice are defective in tumor surveillance in models of spontaneous malignancy**. *Immunity*, **28**(4):571–580, 2008. 22

[124] Biljana Zafirova, Sanja Mandarić, Ronald Antulov, Astrid Krmpotić, Helena Jonsson, Wayne M Yokoyama, Stipan Jonjić, and Bojan Polić. **Altered NK cell development and enhanced NK cell-mediated resistance to mouse cytomegalovirus in NKG2D-deficient mice.** *Immunity*, **31**(2):270–282, August 2009. 22

[125] Paula J Kaplan-Lefko, Tsuey-Ming Chen, Michael M Ittmann, Roberto J Barrios, Gustavo E Ayala, Wendy J Huss, Lisette A Maddison, Barbara A Foster, and Norman M Greenberg. **Pathobiology of autochthonous prostate cancer in a pre-clinical transgenic mouse model.** *The Prostate*, **55**(3):219–237, May 2003. 22

[126] David H Raulet, Stephan Gasser, Benjamin G Gowen, Weiwen Deng, and Heiyoun Jung. **Regulation of ligands for the NKG2D activating receptor.** *Annual review of immunology*, **31**:413–441, 2013. 23

[127] A.R. Mistry and C.A. O'Callaghan. **Regulation of ligands for the activating receptor NKG2D**. *Immunology*, **121**(4):439–447, 2007. 23

[128] Gang Eric Chen, Huiling Wu, Jin Ma, Steven J Chadban, and Alexandra Sharland. **Toll-like receptor 4 engagement contributes to expression of NKG2D ligands by renal tubular epithelial cells.** *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association*, **26**(12):3873–3881, December 2011. 24

[129] Takashi Ebihara, Hisayo Masuda, Takashi Akazawa, Masashi Shingai, Hideaki Kikuta, Tadashi Ariga, Misako Matsumoto, and Tsukasa Seya. **Induction of NKG2D ligands on human dendritic cells by TLR ligand stimulation and RNA virus infection.** *International immunology*, **19**(10):1145–1155, October 2007. 24, 36

[130] Jessica E Bolden, Melissa J Peart, and Ricky W Johnstone. **Anticancer activities of histone deacetylase inhibitors.** *Nature reviews. Drug discovery*, **5**(9):769–784, September 2006. 25

[131] N Kato, J Tanaka, J Sugita, T Toubai, Y Miura, M Ibata, Y Syono, S Ota, T Kondo, M Asaka, and M Imamura. **Regulation of the expression of MHC class I-related chain A, B (MICA, MICB) via chromatin remodeling and its impact on the susceptibility of leukemic cells to the cytotoxicity of NKG2D-expressing cells**. *Leukemia*, **21**(10):2103–2108, 2007. 25

[132] Lars Andresen, Helle Jensen, Marianne T Pedersen, Karen A Hansen, and Søren Skov. **Molecular regulation of MHC class I chain-related protein A expression after HDAC-inhibitor treatment of Jurkat T cells.** *Journal of immunology (Baltimore, Md : 1950)*, **179**(12):8235–8242, December 2007. 25

[133] Cai Zhang, Yiping Wang, Zhixia Zhou, Jian Zhang, and Zhigang Tian. **Sodium butyrate upregulates expression of NKG2D ligand MICA/B in HeLa and HepG2 cell lines and increases their susceptibility to NK lysis.** *Cancer Immunology, Immunotherapy*, **58**(8):1275–1285, January 2009. 25

[134] Luciana L Molinero, Mercedes B Fuertes, Leonardo Fainboim, Gabriel A Rabinovich, and Norberto W Zwirner. **Up-regulated expression of MICA on activated T lymphocytes involves Lck and Fyn kinases and signaling through MEK1/ERK, p38 MAP kinase, and calcineurin.** *Journal of leukocyte biology*, **73**(6):815–822, June 2003. 25

[135] Riki Okita, Dimitrios Mougiakakos, Takashi Ando, Yumeng Mao, Dhifaf Sarhan, Erik Wennerberg, Barbara Seliger, Andreas Lundqvist, Kousaku Mimura, and Rolf Kiessling. **HER2/HER3 signaling regulates NK cell-mediated cytotoxicity via MHC class I chain-related molecule A and B expression in human breast cancer cell lines.** *The Journal of Immunology*, **188**(5):2136–2145, March 2012. 26

[136] G.M. Venkataraman, D. Suciu, V Groh, J.M. Boss, and T Spies. **Promoter region architecture and transcriptional regulation of the genes for the MHC class I-related chain A and B ligands of NKG2D**. *Journal of Immunology*, **178**(2):961–969, 2007. 27, 177

[137] Da Lin, Hayley Lavender, Elizabeth J Soilleux, and Christopher A O'Callaghan. **NF-B regulates MICA gene transcription in endothelial cell through a genetically inhibitable control site.** *The Journal of biological chemistry*, **287**(6):4299–4310, February 2012. 27

[138] Romain Bedel, Antoine Thiery-Vuillemin, Camille Grandclement, Jeremy Balland, Jean-Paul Remy-Martin, Bernadette Kantelip, Jean-René Pallandre, Xavier Pivot, Christophe Ferrand, Pierre Tiberghien, and Christophe Borg. **Novel role for STAT3 in transcriptional regulation of NK immune cell targeting receptor MICA on cancer cells.** *Cancer Research*, **71**(5):1615–1626, March 2011. 27

[139] Liang Wei, Jun Lu, Li Feng, Dan Long, Juan Shan, Shengfu Li, and Youping Li. **HIF-1alpha accumulation upregulates MICA and MICB expression on human cardiomyocytes and enhances NK cell cytotoxicity during hypoxia-reoxygenation.** *Life sciences*, **87**(3-4):111–119, July 2010. 27

[140] David P Bartel. **MicroRNAs: target recognition and regulatory functions.** *Cell*, **136**(2):215–233, January 2009. 30

[141] Noam Stern-Ginossar, Chamutal Gur, Moshe Biton, Elad Horwitz, Moran Elboim, Noa Stanietsky, Michal Mandelboim, and Ofer Mandelboim. **Human microRNAs regulate stress-induced immune responses mediated by the receptor NKG2D.** *Nature immunology*, **9**(9):1065–1073, September 2008. 30, 164

[142] Anja Heinemann, Fang Zhao, Sonali Pechlivanis, Jürgen Eberle, Alexander Steinle, Sven Diederichs, Dirk Schadendorf, and Annette Paschen. **Tumor suppressive microRNAs miR-34a/c control cancer cell expression of ULBP2, a stress-induced ligand of the natural killer cell receptor NKG2D.** *Cancer Research*, **72**(2):460–471, January 2012. 30

[143] Sorin Armeanu, Matthias Krusch, Katrin M Baltz, Thomas S Weiss, Irina Smirnow, Alexander Steinle, Ulrich M Lauer, Michael Bitzer, and Helmut R Salih. **Direct and natural killer cell-mediated antitumor effects of low-dose bortezomib in hepatocellular carcinoma.** *Clinical cancer research : an official journal of the American Association for Cancer Research*, **14**(11):3520–3528, June 2008. 32

[144] James E Butler, Mikel B Moore, Steven R Presnell, Huei-Wei Chan, N Jan Chalupny, and Charles T Lutz. **Proteasome regulation of ULBP1 transcription.** *The Journal of Immunology*, **182**(10):6600–6609, May 2009. 32

[145] G Liu, C L Atteridge, X Wang, A D Lundgren, and J D Wu. **Cutting Edge: The Membrane Type Matrix Metalloproteinase MMP14 Mediates Constitutive Shedding of MHC Class I Chain-Related Molecule A Independent of A Disintegrin and Metalloproteinases**. *Journal of immunology (Baltimore, Md : 1950)*, **184**(7):3346–3350, April 2010. 32

[146] Inja Waldhauer, Dennis Goehlsdorf, Friederike Gieseke, Toni Weinschenk, Mareike Wittenbrink, Andreas Ludwig, Stefan Stevanovic, Hans-Georg Rammensee, and Alexander Steinle. **Tumor-associated MICA is shed by ADAM proteases**. *Cancer Research*, **68**(15):6368–6376, August 2008. 32

[147] Stefan Holdenrieder, Petra Stieber, Andrea Peterfi, Dorothea Nagel, Alexander Steinle, and Helmut Rainer Salih. **Soluble MICA in malignant diseases**. *International journal of cancer Journal international du cancer*, **118**(3):684–687, August 2005. 32

[148] Lars Andresen, Sarah Line Skovbakke, Gry Persson, Michael Hagemann-Jensen, Karen Aagaard Hansen, Helle Jensen, and Søren Skov. **2-deoxy D-glucose prevents cell surface expression of NKG2D ligands through inhibition of N-linked glycosylation.** *The Journal of Immunology*, **188**(4):1847–1855, February 2012. 32, 104

[149] Zhangguo Chen, Lanfen Chen, Kristi Baker, Torsten Olszak, Sebastian Zeissig, Yu-Hwa Huang, Timothy T Kuo, Ofer Mandelboim, Nicole Beauchemin, Lewis L Lanier, and Richard S Blumberg. **CEACAM1 dampens antitumor immunity by down-regulating NKG2D ligand expression on tumor cells.** *The Journal of experimental medicine*, **208**(13):2633–2640, December 2011. 32

[150] Willem H Koppenol, Patricia L Bounds, and Chi V Dang. *Otto Warburg's contributions to current concepts of cancer metabolism.*, **11**. Nat Rev Cancer, May 2011. 33, 36

[151] O WARBURG. **Über den Stoffwechsel der Carcinomzelle**. *Die Naturwissenschaften*, **12**(50):1131–1137, 1924. 33, 36

[152] D Hanahan and R A Weinberg. **The hallmarks of cancer**. *Cell*, **100**(1):57–70, January 2000. 33

[153] Douglas Hanahan and Robert A Weinberg. **Hallmarks of cancer: the next generation.** *Cell*, **144**(5):646–674, March 2011. 33

[154] W H MUNYON and D J MERCHANT. **The relation between glucose utilization, lactic acid production and utilization and the growth cycle of L strain fibroblasts**. *Experimental cell research*, **17**(3):490–498, June 1959. 35, 36, 117

[155] M P Landini. **Early enhanced glucose uptake in human cytomegalovirus-infected cells.** *The Journal of general virology*, **65 (Pt 7)**(7):1229–1232, July 1984. 35, 36, 164, 168

[156] Yongjun Yu, Amy J Clippinger, and James C Alwine. **Viral effects on metabolism: changes in glucose and glutamine utilization during human cytomegalovirus infection.** *Trends in microbiology*, **19**(7):360–367, July 2011. 35, 36, 164, 168

[157] Yongjun Yu, Tobi G Maguire, and James C Alwine. **Human cytomegalovirus activates glucose transporter 4 expression to increase glucose uptake during infection.** *Journal of Virology*, **85**(4):1573–1580, February 2011. 35, 179

[158] Connie M Krawczyk, Thomas Holowka, Jie Sun, Julianna Blagih, Eyal Amiel, Ralph J DeBerardinis, Justin R Cross, Euihye Jung, Craig B Thompson, Russell G Jones, and Edward J Pearce. **Toll-like receptor-induced changes in glycolytic metabolism regulate dendritic cell activation**. *Blood*, **115**(23):4742–4749, 2010. 35, 36

[159] O WARBURG, F Wind, and E Negelein. **THE METABOLISM OF TUMORS IN THE BODY.** *The Journal of general physiology*, **8**(6):519–530, March 1927. 36, 37

[160] R.A. Cairns, I.S. Harris, and T.W. Mak. **Regulation of cancer cell metabolism**. *Nature reviews Cancer*, **11**(2):85–95, 2011. 36, 115

[161] T Wang, C Marquardt, and J Foker. **Aerobic glycolysis during lymphocyte proliferation.** *Nature*, **261**(5562):702–705, June 1976. 36, 47

[162] Carl F Cori and Gerty T Cori. **THE CARBOHYDRATE METABOLISM OF TUMORS II. CHANGES IN THE SUGAR, LACTIC ACID, AND CO2-COMBINING POWER OF BLOOD PASSING THROUGH A TUMOR**. *The Journal of biological chemistry*, **65**(2):397–405, 1925. 37

[163] O WARBURG. **On the origin of cancer cells.** *Science (New York, NY)*, **123**(3191):309–314, February 1956. 37, 47

[164] J B Murphy and J A Hawkins. **COMPARATIVE STUDIES ON THE METABOLISM OF NORMAL AND MALIGNANT CELLS.** *The Journal of general physiology*, **8**(2):115–130, September 1925. 37

[165] Frederic Parker Jr Eugene C Glover With the Assistance of Isabelle Newton Henry Jackson, Jr. **STUDIES OF DISEASES OF THE LYMPHOID AND MYELOID TISSUES. I : THE CHEMICAL METABOLISM OF NORMAL AND PATHOLOGICAL LYMPH NODES**. *The Journal of experimental medicine*, **52**(4):547, September 1930. 38

[166] H GOLDBLATT and G CAMERON. **Induced malignancy in cells from rat myocardium subjected to intermittent anaerobiosis during long propagation in vitro.** *The Journal of experimental medicine*, **97**(4):525–552, April 1953. 40

[167] C Edwards, D Heath, P Harris, Yolanda Castillo, H Krüger, and J Arias Stella. **The carotid body in animals at high altitude**. *The Journal of pathology*, **104**(4):231–238, 1971. 40

[168] J Arias Stella and J Valcarcel. **Chief cell hyperplasia in the human carotid body at high altitudes; physiologic and pathologic significance.** *Human Pathology*, **7**(4):361–373, July 1976. 40

[169] D Astuti, F Latif, A Dallol, P L Dahia, F Douglas, E George, F Sköldberg, E S Husebye, C Eng, and E R Maher. **Gene mutations in the succinate dehydrogenase subunit SDHB cause susceptibility to familial pheochromocytoma and to familial paraganglioma.** *American journal of human genetics*, **69**(1):49–54, July 2001. 40

[170] S Niemann and U Müller. **Mutations in SDHC cause autosomal dominant paraganglioma, type 3.** *Nature genetics*, **26**(3):268–270, November 2000. 40

[171] Ian P M Tomlinson, N Afrina Alam, Andrew J Rowan, Ella Barclay, Emma E M Jaeger, David Kelsell, Irene Leigh, Patricia Gorman, Hanan Lamlum, Shamima Rahman, Rebecca R Roylance, Simon Olpin, Stephen Bevan, Karen Barker, Nicholas Hearle, Richard S Houlston, Maija Kiuru, Rainer Lehtonen, Auli Karhu, Susa Vilkki, Päivi Laiho, Carita Eklund, Outi Vierimaa, Kristiina Aittomäki, Marja Hietala, Pertti Sistonen, Anders Paetau, Reijo Salovaara, Riitta Herva, Virpi Launonen, Lauri A Aaltonen, and Multiple Leiomyoma Consortium. **Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer.** *Nature genetics*, **30**(4):406–410, April 2002. 40

[172] Houman Ashrafian, Linda O'Flaherty, Julie Adam, Violetta Steeples, Yuen-Li Chung, Phil East, Sakari Vanharanta, Heli Lehtonen, Emma Nye, Emine Hatipoglu, Melroy Miranda, Kimberley Howarth, Deepa Shukla, Helen Troy, John Griffiths, Bradley Spencer-Dene, Mohammed Yusuf, Emanuela Volpi, Patrick H Maxwell, Gordon Stamp, Richard Poulsom, Christopher W Pugh, Barbara Costa, Chiara Bardella, Maria Flavia Di Renzo, Michael I Kotlikoff, Virpi Launonen, Lauri Aaltonen, Mona El-Bahrawy, Ian Tomlinson, and Patrick J Pollard. **Expression profiling in progressive stages of fumarate-hydratase deficiency: the contribution of metabolic changes to tumorigenesis.** *Cancer Research*, **70**(22):9153–9165, November 2010. 40, 41

[173] Youfeng Yang, Vladimir Valera, Carol Sourbier, Cathy D Vocke, Minghui Wei, Lisa Pike, Ying Huang, Maria A Merino, Gennady Bratslavsky, Min Wu, Christopher J Ricketts, and W Marston Linehan. **A novel fumarate hydratase-deficient HLRCC kidney cancer cell line, UOK268: a model of the Warburg effect in cancer**. *Cancer Genetics*, **205**(7-8):377–390, July 2012. 41

[174] Saroj P Mathupala, Young H Ko, and Peter L Pedersen. **Hexokinase-2 bound to mitochondria: cancer's stygian link to the "Warburg Effect" and a pivotal target for effective therapy**. *Seminars in cancer biology*, **19**(1):17–24, 2009. 42, 43

[175] E Bustamante and P L Pedersen. **High aerobic glycolysis of rat hepatoma cells in culture: Role of mitochondrial hexokinase**. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 3735–3739, 1977. 42, 43

[176] E Bustamante, H P Morris, and P L Pedersen. **Energy metabolism of tumor cells. Requirement for a form of hexokinase with a propensity for mitochondrial binding**. *Journal of Biological Chemistry*, **256**(16):8699–8704, 1981. 42

[177] S Heikkinen, S Suppola, M Malkki, S S Deeb, J Jänne, and M Laakso. **Mouse hexokinase II gene: Structure, cDNA, promoter analysis, and expression pattern**. *Mammalian Genome*, **11**(2):91–96, 2000. 43

[178] S Heikkinen, M Pietilä, M Halmekytö, S Suppola, E Pirinen, S S Deeb, J Jänne, and M Laakso. **Hexokinase II-deficient mice. Prenatal death of homozygotes without disturbances in glucose tolerance in heterozygotes.** *The Journal of biological chemistry*, **274**(32):22517–22523, August 1999. 43

[179] S S Deeb, M Malkki, and M Laakso. **Human hexokinase II: sequence and homology to other hexokinases.** *Biochemical and Biophysical Research Communications*, **197**(1):68–74, November 1993. 43

[180] Amparo Wolf, Sameer Agnihotri, Johann Micallef, Joydeep Mukherjee, Nesrin Sabha, Rob Cairns, Cynthia Hawkins, and Abhijit Guha. **Hexokinase 2 is a key mediator of aerobic glycolysis and promotes tumor growth in human glioblastoma multiforme.** *The Journal of experimental medicine*, **208**(2):313–326, February 2011. 43

[181] Peng Guo-Qing, Yang Yuan, Zhong Cai-Gao, Yin Hongling, Hu Gonghua, and Tian Yan. **A study of association between expression of hOGG1, VDAC1, HK-2 and cervical carcinoma.** *Journal of experimental & clinical cancer research : CR*, **29**(1):129, 2010. 43

[182] Minsuk Rho, Jin Kim, Chang Do Jee, You Mie Lee, Hee Eun Lee, Min A Kim, Hye Seung Lee, and Woo Ho Kim. **Expression of type 2 hexokinase and mitochondria-related genes in gastric carcinoma tissues and cell lines.** *Anticancer research*, **27**(1A):251–258, January 2007. 43

[183] T A D Smith. **The rate-limiting step for tumor [18F]fluoro-2-deoxy-D-glucose (FDG) incorporation**. *Nuclear Medicine and Biology*, **28**(1):1–4, 2001. 43, 116

[184] C I Pogson. **Two interconvertible forms of pyruvate kinase in adipose tissue.** *Biochemical and Biophysical Research Communications*, **30**(3):297–302, February 1968. 43

[185] K Imamura, K Taniuchi, and T Tanaka. **Multimolecular forms of pyruvate kinase: II. purification of m2-type pyruvate kinase from yoshida ascites hepatoma 130 cells and comparative studies on the enzymological and immunological properties of the three types of pyruvate kinases, l, m1, and m2**. *Journal of Biochemistry*, **72**(4):1001–1015, 1972. 44

[186] M A Steinmetz and W C Deal, Jr. **Metabolic control and structure of glycolytic enzymes. III. Dissociation and subunit structure of rabbit muscle pyruvate kinase**. *Biochemistry*, **5**(4):1399–1405, 1966. 44

[187] Shuichi Saheki, Kikuko Harada, Yoshikazu Sanno, and Takehiko TANAKA. **Hybrid isozymes of rat pyruvate kinase. Their subunit structure and developmental changes in the liver**. *Biochimica et Biophysica Acta (BBA) - Enzymology*, **526**(1):116–128, September 1978. 44

[188] J Marie, A Kahn, and P Boivin. **Pyruvate kinase isozymes in man. I. M type isozymes in adult and foetal tissues, electrofocusing and immunological studies.** *Human genetics*, **31**(1):35–45, January 1976. 44, 47

[189] C E SHONK, H P Morris, and G E BOXER. **PATTERNS OF GLYCOLYTIC ENZYMES IN RAT LIVER AND HEPATOMA.** *Cancer Research*, **25**:671–676, June 1965. 44

[190] C H Lo, F Farina, H P Morris, and S WEINHOUSE. **Glycolytic regulation in rat liver and hepatomas.** *Advances in enzyme regulation*, **6**(C):453–464, 1968. 44

[191] K Imamura and T Tanaka. **Multimolecular forms of pyruvate kinase from rat and other mammalian tissues. I. Electrophoretic studies.** *Journal of Biochemistry*, **71**(6):1043–1051, June 1972. 44

[192] Cheng-Fu Zhou, Xue-Bing Li, Heng Sun, Bo Zhang, Yu-Song Han, Yi Jiang, Qiu-Lin Zhuang, Jing Fang, and Guo-Hao Wu. **Pyruvate kinase type M2 is upregulated in colorectal cancer and promotes proliferation and migration of colon cancer cells.** *IUBMB life*, **64**(9):775–782, September 2012. 44

[193] C Zhan, Y Shi, C Lu, and Q Wang. **Pyruvate kinase M2 is highly correlated with the differentiation and the prognosis of esophageal squamous cell cancer.** *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus / I.S.D.E*, January 2013. 44

[194] J Mukherjee, J J Phillips, S Zheng, J Wiencke, S M Ronen, and R O Pieper. **Pyruvate Kinase M2 Expression, but Not Pyruvate Kinase Activity, Is Up-Regulated in a Grade-Specific Manner in Human Glioma**. *PloS one*, **8**(2), 2013. 44

[195] H Kato, T Fukuda, C Parkison, P McPhie, and S Y Cheng. **Cytosolic thyroid hormone-binding protein is a monomer of pyruvate kinase**. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 7861–7865. Laboratory of Molecular Biology, National Cancer Institute, Bethesda, MD 20892., 1989. 44

[196] K Ashizawa, M C Willingham, C M Liang, and S Y Cheng. **In vivo regulation of monomer-tetramer conversion of pyruvate kinase subtype M2 by glucose is mediated via fructose 1,6-bisphosphate.** *The Journal of biological chemistry*, **266**(25):16842–16846, September 1991. 44

[197] C Parkison, K Ashizawa, P McPhie, K H Lin, and S Y Cheng. **The monomer of pyruvate kinase, subtype M1, is both a kinase and a cytosolic thyroid hormone binding protein.** *Biochemical and Biophysical Research Communications*, **179**(1):668–674, August 1991. 45

[198] Jill D Dombrauckas, Bernard D Santarsiero, and Andrew D Mesecar. **Structural Basis for Tumor Pyruvate Kinase M2 Allosteric Regulation and Catalysis ,**. *Biochemistry*, **44**(27):9417–9429, July 2005. 45

[199] Y Ikeda, T Tanaka, and T Noguchi. **Conversion of non-allosteric pyruvate kinase isozyme into an allosteric enzyme by a single amino acid substitution**. *Journal of Biological Chemistry*, **272**(33):20495–20501, 1997. 45

[200] P Presek, M Reinacher, and E Eigenbrodt. **Pyruvate kinase type M2 is phosphorylated at tyrosine residues in cells transformed by Rous sarcoma virus.** *FEBS letters*, **242**(1):194–198, December 1988. 45

[201] Heather R Christofk, Matthew G vander Heiden, Ning Wu, John M Asara, and Lewis C Cantley. **Pyruvate kinase M2 is a phosphotyrosine-binding protein**. *Nature*, **452**(7184):181–186, March 2008. 45

[202] Dimitrios Anastasiou, Yimin Yu, William J Israelsen, Jian-Kang Jiang, Matthew B Boxer, Bum Soo Hong, Wolfram Tempel, Svetoslav Dimov, Min Shen, Abhishek Jha, Hua Yang, Katherine R Mattaini, Christian M Metallo, Brian P Fiske, Kevin D Courtney, Scott Malstrom, Tahsin M Khan, Charles Kung, Amanda P Skoumbourdis, Henrike Veith, Noel Southall, Martin J Walsh, Kyle R Brimacombe, William Leister, Sophia Y Lunt, Zachary R Johnson, Katharine E Yen, Kaiko Kunii, Shawn M Davidson, Heather R Christofk, Christopher P Austin, James Inglese, Marian H Harris, John M Asara, Gregory Stephanopoulos, Francesco G Salituro, Shengfang Jin, Lenny Dang, Douglas S Auld, Hee-Won Park, Lewis C Cantley, Craig J Thomas, and Matthew G vander Heiden. **Pyruvate kinase M2 activators promote tetramer formation and suppress tumorigenesis.** *Nature Chemical Biology*, **8**(10):839–847, October 2012. 45

[203] Katharina Bluemlein, Nana-Maria Grüning, René G Feichtinger, Hans Lehrach, Barbara Kofler, and Markus Ralser. **No evidence for a shift in pyruvate kinase PKM1 to PKM2 expression during tumorigenesis.** *Oncotarget*, **2**(5):393–400, 2011. 47

[204] Heather R Christofk, Matthew G vander Heiden, Marian H Harris, Arvind Ramanathan, Robert E Gerszten, Ru Wei, Mark D Fleming, Stuart L Schreiber, and Lewis C Cantley. **The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth**. *Nature*, **452**(7184):230–233, March 2008. 47

[205] Casey J Fox, Peter S Hammerman, and Craig B Thompson. **Fuel feeds function: energy metabolism and the T-cell response**. *Nature reviews Immunology*, **5**(11):844–852, November 2005. 47

[206] Avraham Mayevsky. **Mitochondrial function and energy metabolism in cancer cells: past overview and future perspectives.** *Mitochondrion*, **9**(3):165–179, June 2009. 48

[207] S WEINHOUSE, R H MILLINGTON, and C E WENNER. **Metabolism of neoplastic tissue. I. The oxidation of carbohydrate and fatty acids in transplanted tumors.** *Cancer Research*, **11**(11):845–850, November 1951. 48

[208] V. Gogvadze, B. Zhivotovsky, and S. Orrenius. **The Warburg effect and mitochondrial stability in cancer cells**. *Molecular Aspects of Medicine*, **31**(1):60–74, 2010. 48

[209] MW Pfaffl. **A new mathematical model for relative quantification in real-time RT-PCR**. *Nucleic acids research*, **29**(9):–, 2001. 59, 87

[210] K Reinicke, P Sotomayor, P Cisterna, C Delgado, F Nualart, and A Godoy. **Cellular distribution of glut-1 and glut-5 in benign and malignant human prostate tissue**. *Journal of Cellular Biochemistry*, **113**(2):553–562, 2012. 60

[211] Danna Hargett and Thomas E Shenk. **Experimental human cytomegalovirus latency in CD14+ monocytes.** *Proceedings of the National Academy of Sciences*, **107**(46):20039–20044, November 2010. 60

[212] Kim D Pruitt, Jennifer Harrow, Rachel A Harte, Craig Wallin, Mark Diekhans, Donna R Maglott, Steve Searle, Catherine M Farrell, Jane E Loveland, Barbara J Ruef, Elizabeth Hart, Marie-Marthe Suner, Melissa J Landrum, Bronwen Aken, Sarah Ayling, Robert Baertsch, Julio Fernandez-Banet, Joshua L Cherry, Val Curwen, Michael Dicuccio, Manolis Kellis, Jennifer Lee, Michael F Lin, Michael Schuster, Andrew Shkeda, Clara Amid, Garth Brown, Oksana Dukhanina, Adam Frankish, Jennifer Hart, Bonnie L Maidak, Jonathan Mudge, Michael R Murphy, Terence Murphy, Jeena Rajan, Bhanu Rajput, Lillian D Riddick, Catherine Snow,

Charles Steward, David Webb, Janet A Weber, Laurens Wilming, Wenyu Wu, Ewan Birney, David Haussler, Tim Hubbard, James Ostell, Richard Durbin, and David Lipman. **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.** *Genome research*, **19**(7):1316–1323, July 2009. 63

[213] J M Taylor, K Wicks, C Vandiedonck, and J C Knight. **Chromatin profiling across the human tumour necrosis factor gene locus reveals a complex, cell type-specific landscape with novel regulatory elements**. *Nucleic acids research*, **36**(15):4845–4862, July 2008. 86, 188

[214] M.O. Dorschner, M. Hawrylycz, R. Humbert, J.C. Wallace, A. Shafer, J. Kawamoto, J. Mack, R. Hall, J. Goldy, P.J. Sabo, A. Kohli, Q. Li, M McArthur, and J.A. Stamatoyannopoulos. **High-throughput localization of functional elements by quantitative chromatin profiling.** *Nature methods*, **1**(3):219–225, 2004. 86, 187, 188

[215] M Bensellam, L Van Lommel, L Overbergh, F C Schuit, and J C Jonas. **Cluster analysis of rat pancreatic islet gene mRNA levels after culture in low-, intermediate- and high-glucose concentrations.** *Diabetologia*, **52**(3):463–476, March 2009. 94

[216] Davis W Cheng, Yan Jiang, Anath Shalev, Renu Kowluru, Errol D Crook, and Lalit P Singh. **An analysis of high glucose and glucosamine-induced gene expression and oxidative stress in renal mesangial cells.** *Archives of physiology and biochemistry*, **112**(4-5):189–218, October 2006. 94

[217] G C Webb, M S Akbar, C Zhao, and D F Steiner. **Expression profiling of pancreatic cells: Glucose regulation of secretory and metabolic pathway genes**. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 5773–5778. Department of Biochemistry and Molecular Biology, University of Chicago, and Howard Hughes Medical Institute, 5841 South Maryland Avenue, Room N216, Chicago, IL 60637, USA., 2000. 94

[218] P.E. MacDonald, J.W. Joseph, and P. Rorsman. **Glucose-sensing mechanisms in pancreatic -cells**. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**(1464):2211–2225, 2005. 94, 96, 119, 155

[219] F Umenishi and R W Schrier. **Hypertonicity-induced aquaporin-1 (AQP1) expression is mediated by the activation of MAPK pathways and hypertonicity-responsive element in the AQP1 gene**. *Journal of Biological Chemistry*, **278**(18):15765–15770, 2003. 94

[220] U Hasler, P Nunes, R Bouley, H A J Lu, T Matsuzaki, and D Brown. **Acute hypertonicity alters aquaporin-2 trafficking and induces a MAPK-dependent accumulation at the plasma membrane of renal epithelial cells**. *Journal of Biological Chemistry*, **283**(39):26643–26661, 2008. 94

[221] S Özcan, J Dover, and M Johnston. **Glucose sensing and signaling by two glucose receptors in the yeast Saccharomyces cerevisiae**. *The EMBO journal*, **17**(9):2566–2573, 1998. 94

[222] Shadia Zaman, Soyeon Im Lippman, Xin Zhao, and James R Broach. **How Saccharomyces Responds to Nutrients**. *Annual review of genetics*, **42**:27–81, 2008. 94

[223] Ming V Li, Weiqin Chen, Romain N Harmancey, Alli M Nuotio-Antar, Minako Imamura, Pradip Saha, Heinrich Taegtmeyer, and Lawrence Chan. **Glucose-6-phosphate mediates activation of the carbohydrate responsive binding protein (ChREBP).** *Biochemical and Biophysical Research Communications*, **395**(3):395–400, May 2010. 96, 110

[224] Gaëlle Filhoulaud, Sandra Guilmeau, Renaud Dentin, Jean Girard, and Catherine Postic. **Novel insights into ChREBP regulation and function.** *Trends in endocrinology and metabolism: TEM*, **24**(5):257–268, May 2013. 96

[225] S Adolph, S Brusselbach, and R Muller. **Inhibition of transcription blocks cell cycle progression of NIH3T3 fibroblasts specifically in G1.** *Journal of cell science*, **105 (Pt 1)**(1):113–122, May 1993. 100

[226] D B Bregman, R G Pestell, and V J Kidd. **Cell cycle regulation and RNA polymerase II**. *Frontiers in bioscience : a journal and virtual library*, **5**:D244–57, February 2000. 100

[227] R. Augustin. **The protein family of glucose transport facilitators: It's not only about glucose after all**. *IUBMB life*, **62**(5):315–333, 2010. 109

[228] A N WICK, D R DRURY, and T N MORITA. **2-Deoxyglucose; a metabolic block for glucose.** In *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.)*, pages 579–582, 1955. 110

[229] Patricio Rodríguez, Coralia I Rivas, Alejandro Godoy, Marcela Villanueva, Jorge Fischbarg, Juan Carlos Vera, and Alejandro M Reyes. **Redefining the facilitated transport of mannose in human cells: absence of a glucose-insensitive, high-affinity facilitated mannose transport system.** *Biochemistry*, **44**(1):313–320, January 2005. 112

[230] K Panneerselvam and H H Freeze. **Mannose enters mammalian cells using a specific transporter that is insensitive to glucose**. *Journal of Biological Chemistry*, **271**(16):9417–9421, 1996. 112

[231] S Tazawa, T Yamato, H Fujikura, M Hiratochi, F Itoh, M Tomae, Y Takemura, H Maruyama, T Sugiyama, A Wakamatsu, T Isogai, and M Isaji. **SLC5A9/SGLT4, a new Na +-dependent glucose transporter, is an essential transporter for mannose, 1,5-anhydro-D-glucitol, and fructose**. *Life sciences*, **76**(9):1039–1050, 2005. 112

[232] C F Burant, J Takeda, E Brot-Laroche, G I Bell, and N O Davidson. **Fructose transporter in human spermatozoa and small intestine is GLUT5.** *The Journal of biological chemistry*, **267**(21):14523–14526, July 1992. 112

[233] L Szablewski. **Expression of glucose transporters in cancers**. *Biochimica et biophysica acta*, **1835**(2):164–169, 2013. 116

[234] E T SHAPIRO, H TILLIL, K S POLONSKY, V S FANG, A H RUBENSTEIN, and E V CAUTER. **Oscillations in Insulin Secretion During Constant Glucose Infusion in Normal Man: Relationship to Changes in Plasma Glucose**. *Journal of Clinical Endocrinology & Metabolism*, **67**(2):307–314, August 1988. 116

[235] D Melloul, S Marshak, and E Cerasi. **Regulation of insulin gene transcription.** *Diabetologia*, **45**(3):309–326, March 2002. 119, 155

[236] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic acids research*, **40**(D1):D109–D114, 2012. 122

[237] Olivier Henry, Mario Jolicoeur, and Amine Kamen. **Unraveling the metabolism of HEK-293 cells using lactate isotopomer analysis.** *Bioprocess and biosystems engineering*, **34**(3):263–273, March 2011. 123, 138

[238] P Boer and O Sperling. **Role of cellular ribose-5-phosphate content in the regulation of 5-phosphoribosyl-1-pyrophosphate and de novo purine synthesis in a human hepatoma cell line.** *Metabolism: clinical and experimental*, **44**(11):1469–1474, November 1995. 123, 126

[239] S Brosh, P Boer, B Kupfer, A de Vries, and O Sperling. **De novo synthesis of purine nucleotides in human peripheral blood leukocytes. Excessive activity of the pathway in hypoxanthine-guanine phosphoribosyltransferase deficiency.** *The Journal of clinical investigation*, **58**(2):289–297, August 1976. 123

[240] O Sperling, S Brosh, P Boer, B Kupfer, B Benjamin, A Weinberger, and J Pinkhas. **De novo synthesis of purine nucleotides and metabolic availability of phosphoribosylpyrophosphate in leukemic leukocytes.** *Biomedicine / [publiée pour l'A.A.I.C.I.G.]*, **31**(1):20–23, February 1979. 123

[241] R.B. Gordon, L. Thompson, L.A. Johnson, and B.T. Emmerson. **Regulation of purine de novo synthesis in cultured human fibroblasts: The role of P-Ribose-PP**. *Biochimica et biophysica acta*, **562**(1):162–176, 1979. 124, 125

[242] M E King, J M Honeysett, and S B Howell. **Regulation of de novo purine synthesis in human bone marrow mononuclear cells by hypoxanthine.** *The Journal of clinical investigation*, **72**(3):965–970, September 1983. 124

[243] E C Moore and G A LePage. **In vivo sensitivity of normal and neoplastic mouse tissues to azaserine.** *Cancer Research*, **17**(8):804–808, September 1957. 125

[244] R M Liskay and D Patterson. **A selective medium (GAMA) for the isolation of somatic cell hybrids from HPRT−and APRT−mutant cells**. *Cytogenetic and Genome Research*, **23**(1-2):61–69, 2008. 125

[245] G Kohler and C Milstein. **Continuous cultures of fused cells secreting antibody of predefined specificity.** *Nature*, **256**(5517):495–497, August 1975. 127

[246] James D Young, Sylvia Y M Yao, Jocelyn M Baldwin, Carol E Cass, and Stephen A Baldwin. **The human concentrative and equilibrative nucleoside transporter families, SLC28 and SLC29.** *Molecular Aspects of Medicine*, **34**(2-3):529–547, April 2013. 135, 136, 173

[247] M F Vincent, F Bontemps, and G Van den Berghe. **Substrate cycling between 5-amino-4-imidazolecarboxamide riboside and its monophosphate in isolated rat hepatocytes.** *Biochemical pharmacology*, **52**(7):999–1006, October 1996. 142

[248] Alejandro Vazquez-Martin, Cristina Oliveras-Ferraros, Eugeni Lopez-Bonet, and Javier A Menendez. **AMPK: Evidence for an energy-sensing cytokinetic tumor suppressor**. *Cell cycle (Georgetown, Tex)*, **8**(22):3679–3683, November 2009. 142

[249] Michael A Menze, Nilay Chakraborty, Matthew Clavenna, Mitali Banerjee, Xiang-Hong Liu, Mehmet Toner, and Steven C Hand. **Metabolic preconditioning of cells with AICAR-riboside: improved cryopreservation and cell-type specific impacts on energetics and proliferation.** *Cryobiology*, **61**(1):79–88, August 2010. 144, 157

[250] P Krzeminski, I Misiewicz, P Pomorski, T Kasprzycka-Guttman, and J Brańska. **Mitochondrial localization of P2Y1, P2Y2 and P2Y12 receptors in rat astrocytes and glioma C6 cells**. *Brain Research Bulletin*, **71**(6):587–592, 2007. 150

[251] F Amoroso, S Falzoni, E Adinolfi, D Ferrari, and F Di Virgilio. **The P2X7 receptor is a key modulator of aerobic glycolysis.** *Cell death & disease*, **3**:e370, 2012. 150

[252] J S Miller, T Cervenka, J Lund, I J Okazaki, and J Moss. **Purine metabolites suppress proliferation of human NK cells through a lineage-specific purine receptor.** *Journal of immunology (Baltimore, Md : 1950)*, **162**(12):7376–7382, June 1999. 150

[253] M T Tu, S F Luo, C C Wang, C S Chien, C T Chiu, C C Lin, and C M Yang. **P2Y(2) receptor-mediated proliferation of C(6) glioma cells via activation of Ras/Raf/MEK/MAPK pathway.** *British journal of pharmacology*, **129**(7):1481–1489, April 2000. 150

[254] G Haskó and B N Cronstein. **Adenosine: An endogenous regulator of innate immunity**. *Trends in immunology*, **25**(1):33–39, 2004. 150

[255] Geoffrey Burnstock. **Purinergic signalling: Its unpopular beginning, its acceptance and its exciting future.** *BioEssays*, **34**(3):218–225, March 2012. 150

[256] T B Meissner, A Li, A Biswas, K H Lee, Y J Liu, E Bayir, D Iliopoulos, P J Van Den Elsen, and K S Kobayashi. **NLR family member NLRC5 is a transcriptional regulator of MHC class I genes**. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 13794–13799. Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute, Boston, MA 02115, USA., 2010. 152

[257] Andreas Neerincx, Galaxia M Rodriguez, Viktor Steimle, and Thomas A Kufer. **NLRC5 controls basal MHC class I gene expression in an MHC enhanceosome-dependent manner.** *The Journal of Immunology*, **188**(10):4940–4950, May 2012. 152

[258] C A Waldren and D Patterson. **Effects of caffeine on purine metabolism and ultraviolet light-induced lethality in cultured mammalian cells.** *Cancer Research*, **39**(12):4975–4982, December 1979. 154, 158

[259] F Giorgelli, C Bottai, L Mascia, C Scolozzi, M Camici, and P L Ipata. **Recycling of alpha-D-ribose 1-phosphate for nucleoside interconversion.** *Biochimica et biophysica acta*, **1335**(1-2):6–22, April 1997. 156

[260] L Mascia, M Cappiello, S Cherri, and P L Ipata. **In vitro recycling of alpha-D-ribose 1-phosphate for the salvage of purine bases.** *Biochimica et biophysica acta*, **1474**(1):70–74, March 2000. 156

[261] Mhairi C Towler and D Grahame Hardie. **AMP-activated protein kinase in metabolic control and insulin signaling.** *Circulation research*, **100**(3):328–341, February 2007. 157, 230

[262] Stephen J Rubino, Thirumahal Selvanantham, Stephen E Girardin, and Dana J Philpott. **Nod-like receptors in the control of intestinal inflammation.** *Current opinion in immunology*, **24**(4):398–404, August 2012. 158

[263] C. Sinzger, M Digel, and G Jahn. **Cytomegalovirus cell tropism.** *Current topics in microbiology and immunology*, **325**:63–83, 2008. 163

[264] C. Sinzger, A Grefte, B Plachter, A S Gouw, T H The, and G Jahn. **Fibroblasts, epithelial cells, endothelial cells and smooth muscle cells are major targets of human cytomegalovirus infection in lung and gastrointestinal tissues.** *The Journal of general virology*, **76 (Pt 4)**:741–750, April 1995. 163

[265] Edward S Mocarski, Thomas Shenk, and Robert F Pass. **Cytomegalovirus**. In Knipe and Howley, editors, *Fields virology*, pages 2703–2772. LWW, December 2008. 164

[266] Claire Dunn, N Jan Chalupny, Claire L Sutherland, Stephanie Dosch, P.V. Sivakumar, David C Johnson, and David Cosman. **Human cytomegalovirus glycoprotein UL16 causes intracellular sequestration of NKG2D ligands, protecting against natural killer cell cytotoxicity.** *The Journal of experimental medicine*, **197**(11):1427–1439, June 2003. 164

[267] N J Chalupny, A. Rein-Weston, S. Dosch, and D Cosman. **Down-regulation of the NKG2D ligand MICA by the human cytomegalovirus glycoprotein UL142**. *Biochemical and Biophysical Research Communications*, **346**(1):175–181, 2006. 164

[268] N J Bennett, O Ashiru, F.J.E. Morgan, Y. Pang, G. Okecha, R.A. Eagle, J Trowsdale, J.G.P. Sissons, and M R Wills. **Intracellular sequestration of the NKG2D ligand ULBP3 by human cytomegalovirus**. *Journal of immunology (Baltimore, Md : 1950)*, **185**(2):1093–1102, 2010. 164

[269] Omodele Ashiru, Neil J Bennett, Louise H Boyle, Mair Thomas, John Trowsdale, and Mark R Wills. **NKG2D ligand MICA is retained in the cis-Golgi apparatus by human cytomegalovirus protein UL142.** *Journal of Virology*, **83**(23):12345–12354, December 2009. 164

[270] Yongjun Yu, Tobi G Maguire, and James C Alwine. **Human cytomegalovirus activates glucose transporter 4 expression to increase glucose uptake during infection.** *Journal of Virology*, **85**(4):1573–1580, February 2011. 164, 168, 173

[271] W P ROWE, J W HARTLEY, S WATERMAN, H C TURNER, and R J HUEBNER. **Cytopathogenic agent resembling human salivary gland virus recovered from tissue cultures of human adenoids.** *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.)*, **92**(2):418–424, June 1956. 164, 167

[272] T A Cha, E Tom, G W Kemble, G M Duke, E.S. Mocarski, and R R Spaete. **Human cytomegalovirus clinical isolates carry at least 19 genes not found in laboratory strains.** *Journal of Virology*, **70**(1):78–83, January 1996. 164

[273] AJ Davison, A Dolan, P Akter, C Addison, DJ Dargan, DJ Alcendor, DJ McGeoch, and GS Hayward. **The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome**. *The Journal of general virology*, **84**(1):17–28, 2003. 164

[274] Barbara Adler, Laura Scrivano, Zsolt Ruzcics, Brigitte Rupp, Christian Sinzger, and Ulrich Koszinowski. **Role of human cytomegalovirus UL131 A in cell type-specific virus entry and release**. *The Journal of general virology*, **87**(Pt 9):2451–2460, 2006. 164

[275] ES Mocarski, MN Prichard, CS Tan, and JM Brown. **Reassessing the organization of the UL42-UL43 region of the human cytomegalovirus strain AD169 genome**. *Virology*, **239**(1):169–175, 1997. 164

[276] Penelope D Barnes and Jane E Grundy. **Down-regulation of the class I HLA heterodimer and beta 2-microglobulin on the surface of cells infected with cytomegalovirus.** *The Journal of general virology*, **73 (Pt 9)**:2395–2403, September 1992. 167

[277] Hartmut Hengel, Wolfram Brune, and Ulrich H Koszinowski. **Immune evasion by cytomegalovirus—survival strategies of a highly adapted opportunist**. *Trends in microbiology*, **6**(5):190–197, 1998. 177

[278] Oliver Bell, Vijay K Tiwari, Nicolas H Thomä, and Dirk Schübeler. **Determinants and dynamics of genome accessibility.** *Nature reviews Genetics*, **12**(8):554–564, 2011. 181

[279] H Weintraub and M Groudine. **Chromosomal subunits in active genes have an altered conformation.** *Science (New York, NY)*, **193**(4256):848–856, September 1976. 182

[280] C. Wu. **The 5'ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I.** *Nature*, **286**(5776):854, 1980. 182, 207

[281] Lingyun Song, Zhancheng Zhang, Linda L Grasfeder, Alan P Boyle, Paul G Giresi, Bum-Kyu Lee, Nathan C Sheffield, Stefan Gräf, Mikael Huss, Damian Keefe, Zheng Liu, Darin London, Ryan M McDaniell, Yoichiro Shibata, Kimberly A Showers, Jeremy M Simon, Teresa Vales, Tianyuan Wang, Deborah Winter, Zhuzhu Zhang, Neil D Clarke, Ewan Birney, Vishy R Iyer, Gregory E Crawford, Jason D Lieb, and Terrence S Furey. **Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.** *Genome research*, July 2011. 182

[282] T. Owen-Hughes and J.L. Workman. **Remodeling the chromatin structure of a nucleosome array by transcription factor-targeted trans-displacement of histones**. *The EMBO journal*, **15**(17):4702–4712, 1996. 182

[283] Vikki M Weake and Jerry L Workman. **Inducible gene expression: diverse regulatory mechanisms.** *Nature reviews Genetics*, **11**(6):426–437, 2010. 182

[284] D J Galas and A Schmitz. **DNAase footprinting: a simple method for the detection of protein-DNA binding specificity**. *Nucleic acids research*, **5**(9):3157–3170, 1978. 182

[285] Gregory E Crawford, Ingeborg E Holt, James C Mullikin, Denise Tai, Robert Blakesley, Gerard Bouffard, Alice Young, Catherine Masiello, Eric D Green, Tyra G Wolfsberg, Francis S Collins, and National Institutes Of Health Intramural Sequencing Center. **Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites.** *Proceedings of the National Academy of Sciences of the United States of America*, **101**(4):992–997, January 2004. 182

[286] PJ Sabo, M. Hawrylycz, JC Wallace, R. Humbert, M Yu, A. Shafer, J. Kawamoto, R. Hall, J. Mack, MO Dorschner, M McArthur, and JA Stamatoyannopoulos. **Discovery of functional noncoding elements by digital analysis of chromatin structure**. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(48):16837–16842, 2004. 182

[287] G.E. Crawford, S. Davis, Scacheri, P.C., G. Renaud, M.J. Halawi, M.R. Erdos, R. Green, P.S. Meltzer, T.G. Wolfsberg, and F.S. Collins. **DNase-chip: A high-resolution method to identify DNase I hypersensitive sites using tiled microarrays**. *Nature methods*, **3**(7):503–509, 2006. 182

[288] Peter J Sabo, Michael S Kuehn, Robert Thurman, Brett E Johnson, Ericka M Johnson, Hua Cao, Man Yu, Elizabeth Rosenzweig, Jeff Goldy, Andrew Haydock, Molly Weaver, Anthony Shafer, Kristin Lee, Fidencio Neri, Richard Humbert, Michael A Singer, Todd A Richmond, Michael O Dorschner, Michael McArthur, Michael Hawrylycz, Roland D Green, Patrick A Navas, William S Noble, and John A Stamatoyannopoulos. **Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays.** *Nature methods*, **3**(7):511–518, July 2006. 182

[289] A.P. Boyle, S. Davis, H.P. Shulha, P. Meltzer, E.H. Margulies, Z Weng, T.S. Furey, and G.E. Crawford. **High-Resolution Mapping and Characterization of Open Chromatin across the Genome**. *Cell*, **132**(2):311–322, 2008. 182, 198

[290] Jay R Hesselberth, Xiaoyu Chen, Zhihong Zhang, Peter J Sabo, Richard Sandstrom, Alex P Reynolds, Robert E Thurman, Shane Neph, Michael S Kuehn, William S Noble, Stanley Fields, and John A Stamatoyannopoulos. **Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.** *Nature Chemical Biology*, **6**(4):283–289, April 2009. 182, 198

[291] M S Kowalczyk, J R Hughes, D Garrick, M D Lynch, J A Sharpe, J A Sloane-Stanley, S J McGowan, M De Gobbi, M Hosseini, D Vernimmen, J M Brown, N E Gray, L Collavin, R J Gibbons, J Flint, S Taylor, V J Buckle, T A Milne, W G Wood, and D R Higgs. **Intragenic Enhancers Act as Alternative Promoters**. *Molecular Cell*, **45**(4):447–458, 2012. 184, 199, 224

[292] Robert M Kuhn, David Haussler, and W James Kent. **The UCSC genome browser and associated tools.** *Briefings in bioinformatics*, October 2012. 188

[293] Kate R Rosenbloom, Timothy R Dreszer, Michael Pheasant, Galt P Barber, Laurence R Meyer, Andy Pohl, Brian J Raney, Ting Wang, Angie S Hinrichs, Ann S Zweig, Pauline A Fujita, Katrina Learned, Brooke Rhead, Kayla E Smith, Robert M Kuhn, Donna Karolchik, David Haussler, and W James Kent. **ENCODE whole-genome data in the UCSC Genome Browser.** *Nucleic acids research*, **38**(Database issue):D620–5, January 2010. 188

[294] G Lunter and M Goodson. **Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads**. *Genome research*, **21**(6):936–939, 2011. 189, 199, 203

[295] S Heinz, C Benner, N Spann, E Bertolino, Y C Lin, P Laslo, J X Cheng, C Murre, H Singh, and C.K. Glass. **Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities**. *Molecular Cell*, **38**(4):576–589, 2010. 192, 199

[296] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. **Model-based analysis of ChIP-Seq (MACS).** *Genome biology*, **9**(9):R137, 2008. 192, 199

[297] Alan P Boyle, Justin Guinney, Gregory E Crawford, and Terrence S Furey. **F-Seq: a feature density estimator for high-throughput sequence tags**. *Bioinformatics*, **24**(21):2537–2538, 2008. 192, 199

[298] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. **The human genome browser at UCSC.** *Genome research*, **12**(6):996–1006, June 2002. 193

[299] B Langmead, C Trapnell, M Pop, and S.L. Salzberg. **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome biology*, **10**(3):R25, 2009. 199

[300] H Li and R. Durbin. **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics*, **26**(5):589–595, 2010. 199

[301] H Li, J Ruan, and R. Durbin. **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome research*, **18**(11):1851–1858, 2008. 199

[302] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics (Oxford, England)*, **25**(16):2078–2079, August 2009. 202

[303] R E Thurman, E Rynes, R. Humbert, J Vierstra, M T Maurano, E Haugen, N C Sheffield, A B Stergachis, H Wang, and B Vernot. **The accessible chromatin landscape of the human genome**. *Nature*, **489**(7414):75–82, 2012. 207

[304] ENCODE Project Consortium, Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, and Michael Snyder. **An integrated encyclopedia of DNA elements in the human genome.** *Nature*, **489**(7414):57–74, September 2012. 223

[305] J Feng, T Liu, B Qin, Y Zhang, and X S Liu. **Identifying ChIP-seq enrichment using MACS**. *Nature protocols*, **7**(9):1728–1740, 2012. 224

[306] Michael F Carey, Craig L Peterson, and Stephen T Smale. *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniqes: Concepts, Strategies and Techniques*. Cold Spring Harbor Laboratory Press, 2nd edition, February 2009. 227

[307] Alessandra Zingoni, Michele Ardolino, Angela Santoni, and Cristina Cerboni. **NKG2D and DNAM-1 activating receptors and their ligands in NK-T cell interactions: role in the NK cell-mediated negative regulation of T cell responses**. *Frontiers in Immunology*, **3**, 2013. 229

[308] Sarah E Jackson, Gavin M Mason, and Mark R Wills. **Human cytomegalovirus immunity and immune evasion.** *Virus research*, **157**(2):151–160, May 2011. 229

[309] Manuela Zaccolo and Matthew A Movsesian. **cAMP and cGMP signaling cross-talk: role of phosphodiesterases and implications for cardiac pathophysiology.** *Circulation research*, **100**(11):1569–1578, June 2007. 230

[310] Rémi Dumollard, Zoe Ward, John Carroll, and Michael R Duchen. **Regulation of redox metabolism in the mouse oocyte and embryo.** *Development (Cambridge, England)*, **134**(3):455–465, February 2007. 230

[311] Irit Carmi-Levy, Nurit Yannay-Cohen, Gillian Kay, Ehud Razin, and Hovav Nechushtan. **Diadenosine tetraphosphate hydrolase is part of the transcriptional regulation network in immunologically activated mast cells.** *Molecular and cellular biology*, **28**(18):5777–5784, September 2008. 230