2015

Characterisation of M protein from Group A streptococcal isolates associated with invasive disease

David Manuel Pereira De Oliveira
University of Wollongong

Characterisation of M protein from Group A streptococcal isolates associated with invasive disease

A thesis submitted in fulfilment of the requirements for the award of the degree

Doctor of Philosophy (PhD)

From the

University of Wollongong

By

David Manuel Pereira De Oliveira

Bachelor of Science (Hons)

Illawarra Health and Medical Research Institute (IHMRI)

Certification

I, David M. P. De Oliveira, declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Department of Biological Sciences, University of Wollongong, is wholly my own work unless otherwise referenced and acknowledged. The document has not been submitted for qualifications at any other academic institution.

_____

David M. P. De Oliveira

July, 2015

Acknowledgments

Firstly I would like to thank my primary supervisor, Martina Sanderson-Smith. Throughout my time under her supervision, she has been an invaluable source of knowledge, support and guidance. Martina is an exemplary model of a supervisor who drives her students to excel while providing them with the stepping stones to succeed in their scientific careers. I regard her as a true friend.

I would also like to thank my co-supervisors Mark Walker and Jason McArthur. Mark – thank you for providing me with the opportunity to join such a fantastic research group, and Jason – thank you not only for your vast array of knowledge, but also for your company at the bar every Friday afternoon.

To be part of the team which developed the newly implemented GAS classification system, I would like to thank Pierre Smeesters. His dedication to this field of research is nothing short of an inspiration. I would also like to thank the research group at The Institute for Glycomics for allowing me to work in their lab and making me feel so welcome. Thankyou Lauren for giving me a place to stay and introducing me to your beautiful family.

A special mention to every past and present member of the McArthur/Sanderson-Smith lab. Specifically Simon, James, Diane, Patrick, Aleta, and Catherine. I thank you all for your friendship and kindness over the past few years. You guys not only helped shape my life, but also provided me with a network of support needed to complete this challenge. I'm sure our friendships will last for many more years to come.

Finally, I would like to thank my family. Without your support, patience and love this would have not been possible.

Abstract

Group A streptococcus (Group A streptococci; GAS; *Streptococcus pyogenes*) is a Gram positive bacterium which is responsible for a wide range of non-invasive and invasive diseases. Reports over the last three decades have indicated resurgences in GAS infection worldwide. The cell surface M protein is the most extensively studied GAS virulence factor with over 220 variants of this protein being described. Coded by the *emm* gene, the M protein mediates invasive GAS infection by binding to a diverse range of host proteins across its entire length. The diverse function of this protein promotes GAS colonisation while also facilitating the bacterium to overcome the host innate immune response. To date, M protein is the most commonly used substrate for epidemiological typing and is a commonly utilised vaccine antigen in GAS vaccine development. Studies investigating the global distribution of M-types have found that while the M-type distribution of GAS in high income settings is restricted to a few predominant M-types, the distribution in low-income settings is much greater. GAS typing is dependent on a small portion (10-15%) of the M protein. Prior studies analysing the sequence diversity of M-types in a low-income settings have found that they are highly similar in sequence, subsequently raising questions of type-specific immunity. In this study 1086 GAS isolates collected from 31 countries representing 175 M-types were phylogenetically assessed on their full length mature M protein sequence, with the aim of developing a new cluster-based classification system which would support future vaccine development and functional studies. Phylogenetic analysis identified that 175 M-types could be grouped in to 2 clades, 2 sub-clades and 48 clusters, 16 of which encompassed 82% of all analysed M-types. To evaluate the utility of this novel classification system as a tool for identification of M protein function, 26 M proteins representing 24 distinct M-types encompassing all major cluster groups were recombinantly expressed and screened for

binding against a select range of host proteins. Surface plasmon resonance (SPR) binding experiments identified that both fibrinogen and plasminogen binding function was restricted to clade-Y. Plasminogen binding was further restricted to cluster D4 whereby all M protein contained the highly predictive plasminogen binding motif. High affinity IgA binding was observed only by M proteins associated with *emm*-clusters E1 and E6 whereas IgG binding was observed by *emm*-clusters E1, E3, E4, E6, A-C3 and non-clustered M proteins M57 and M14. Complement inhibitor C4-binding protein was shown to only bind clade-X M protein representatives, specifically *emm*-clusters E1, E3, E4 and E6. Although all M proteins in study were shown to express the previously characterised human serum albumin binding domain, *emm*-cluster E4 and D4 M protein representatives M53 and M98 were shown not to bind human serum albumin. The ability of GAS vaccine candidate 'Streptococcal vaccine 1' (SV1) to effectively opsonise the C-repeat targeted domains of the M proteins in this study was assessed in the presence of IgG, IgA, plasminogen, fibrinogen and albumin. Competitive SPR binding analysis identified that α-SV1 IgG binding did not compete for binding with the respective host proteins. Only IgA binding by E1 M proteins was shown to both inhibit and outcompete α-SV1 IgG binding interactions. These results clearly demonstrate the utility of this GAS classification system based on full length M protein sequence with applications of facilitating future M protein functional studies, epidemiological surveillance and vaccine development.

Interactions between M protein and host proteins are central to multiple stages of GAS disease. Plasminogen acquisition mediated by the M protein PAM, and activation by SK is central to GAS pathogenesis. Two predominant glycoforms of plasminogen circulate in the host. Glycoform I plasminogen (GI-plasminogen) contains glycosylation sites at Asn_{289} and Thr_{346}, whereas glycoform II plasminogen (GII-plasminogen) is exclusively glycosylated at Thr_{346}. SPR experiments demonstrated that plasminogen binding group A streptococcal M

protein (PAM; M53; *emm*-cluster D4) exhibits relatively equal affinity for GI- and GII-plasminogen in the 'closed' conformation (GII-plasminogen, K_D = 27.4 nM; GI-plasminogen, K_D = 37.0 nM). When plasminogen was in the 'open' conformation, PAM exhibited an 11 fold increase in affinity for GII-plasminogen (K_D = 2.8 nM) compared with GI- plasminogen (K_D = 33.2 nM). The interaction of PAM with plasminogen is believed to be mediated by lysine binding sites within kringle (KR) 2 of plasminogen. PAM and GI-plasminogen interactions were fully inhibited with 100 mM of lysine analogue ε-aminocaproic acid (εACA), whereas PAM-GII-plasminogen interactions were shown to be reduced but not inhibited in the presence of 400 mM εACA. In contrast, binding to the KR1-3 domains of GII- plasminogen (angiostatin) by PAM were completely inhibited in the presence 5 mM εACA. Along with PAM, *emm* pattern D GAS isolates express a phenotypically distinct SK variant (type-2b SK) which requires plasminogen ligands such as PAM to activate plasminogen. Type-2b SK was able to generate an active site and activate GII-plasminogen at a significantly higher rate than GI-plasminogen when bound to PAM. Taken together, these data suggest that GAS selectively recruits and activates GII- plasminogen. Furthermore, it is proposed that the interaction between PAM and plasminogen may be partially mediated by a secondary binding site outside of kringle 2, affected by glycosylation at Asn_{289}.

Aside from protein-protein interactions, carbohydrates (glycans) have also been shown to play an important role in host-pathogen interactions for select bacterial species. As the dominant GAS cell surface protein, M protein is a strong candidate for mediating GAS-glycan interactions, however, to date, M protein-glycan interactions have not been comprehensively explored. Through the use of newly developed glycan microarray technology, the glycan binding profile of each *emm*-cluster was characterised against 319 distinct glycan structures. M protein representatives from each *emm*-cluster were shown to have distinct glycan binding profiles, with several *emm*-cluster groups demonstrating broad

specificity for multiple terminal galactose, glucose, fucose, mannose, sialic acid and glycosaminoglycan containing structures. To evaluate the physiological relevance of these M protein-glycan interactions in GAS pathogenesis, globally disseminated M1T1 GAS strain, 5448, and isogenic mutant 5448ΔM1 were selected for further characterisation. Whole GAS cell glycan microarray screening identified that 5448 could bind 19 different glycan structures in an M1 protein dependant manner. Due to the expression of lacto-*N*-tetraose and blood group H antigen type I structures on human tissue and in mucosal fluid, interactions of these structures with M1 were further investigated. SPR analysis identified that binding of lacto-*N*-tetraose (K_D = 127.3 ± 26.2 μM) and blood group H antigen type I (K_D = 307.2 ± 168.9 nM) were both localised to the B-repeat domains of M1 whereby Galβ1-4Gal was the minimum carbohydrate residue required for recognition. Following pre-incubation with lacto-*N*-tetraose, 5448 was shown to significantly associate more with Detroit-562 pharyngeal cells in a concentration dependant and tissue specific manner ($P < 0.01$). No significant differences in association of 5448 to Detroit-562 cells were observed in the presence of blood group H antigen type I. To assess the potential role of blood group antigen structures in upper respiratory tract colonisation by M1 serotypes, buccal epithelial cell and saliva samples from 20 individuals of unknown blood type were collected. Blood group status and expression were assessed via LC-ESI MS analysis of donor saliva and differences in association of 5448eGFP and 5448ΔM1eGFP to buccal cells from each donor were examined via flow cytometry. Blood group antigen expression analysis identified that 5448eGFP was able to associate with buccal epithelial cells expressing H antigen structures at a significantly higher level than cells expressing A ($P < 0.05$) and AB ($P < 0.01$) antigen structures. No significant differences in association were observed between 5448ΔM1eGFP and buccal epithelial cells expressing various blood group antigen structures. As only 1 subject was shown to express the B antigen, no statistical inferences were able to be made. Although preliminary, this data

suggests blood group antigen related structures play a role in GAS colonisation, further highlighting the complex nature of GAS disease.

This study clearly demonstrates the development and utility of the newly implemented GAS classification system. Furthermore, we have identified novel M protein functions which may contribute to the different stages during GAS infection. These results illustrate the homologous genetic and functional properties of distinct M-types in regions where the epidemiology of infection is diverse. A better understanding of M protein interaction with the host will broaden our knowledge of GAS disease progression, facilitating the development of future therapeutics.

Table of Contents

Publications arising from this thesis

This thesis includes chapters that have been published in the following journal articles:

Martina L. Sanderson-Smith, **David M. P. De Oliveira**, Julien Guglielmini, David D. J McMillan, Therese Vu, Jessica K. Holien, Anna Henningham, Andrew C. Steer, Debra E. Bessen, James B. Dale, Nigel Curtis, Bernard W Beall, Mark J. Walker, Michael W. Parker, Jonathan R. Carapetis, Laurence Van Melderen, Sri K. Sriprakash, Pierre R. Smeesters; M protein study group. (2014). A systematic and functional classification of Streptococcus pyogenes that serves as a new tool for molecular typing and vaccine development. *The Journal of Infectious Diseases*. **210**: 1325-38. **Incorporated in part as chapter 3**

Contributor	**Extent of contribution**
M. L. Sanderson-Smith	Directed and supervised research (35%) Cloning, expression and purification (10%) Preparation of figures (30%) Wrote the paper (40%)
D. M. P. De Oliveira	Cloning, expression and purification (70%) M protein secondary structure analysis (100%) SPR binding analysis (100%) Preparation of figures (20%) Wrote the paper (10%) Edited the paper (5%)
J. Guglielmini	Bioinformatic analysis (70%) Wrote the paper (10%) Edited the paper (5%)
D. D. J McMillan	Directed and supervised research (5%) Edited the paper (15%)
T. Vu	Cloning, expression and purification (15%)
J. K. Holien	Modelling analysis (50%)
A. Henningham	Cloning, expression and purification (5%)
A. C. Steer	Edited the paper (15%)
D. E. Bessen	Bioinformatic analysis (10%) Edited the paper (15%)
J. B. Dale	Directed and supervised research (5%) Edited the paper (5%)
N. Curtis	Edited the paper (5%)
B. W Beall	Directed and supervised research Edited the paper (5%)
M. J. Walker	Directed and supervised research (5%) Edited the paper (10%)

M. W. Parker	Modelling analysis (50%)
J. R. Carapetis	Edited the paper (5%)
L. Van Melderen	Directed and supervised research (5%) Edited the paper (5%)
S. K. Sriprakash	Directed and supervised research (5%) Edited the paper (10%)
P. R. Smeesters	Directed and supervised research (40%) Bioinformatic analysis (20%) Preparation of figures (50%) Wrote the paper (40%)

David M. P. De Oliveira, Ruby H. P. Law, Diane Ly, Simon M. Cook, Adam J. Quek, Jason D. McArthur, James C. Whisstock, and Martina L. Sanderson-Smith. (2015). Preferential acquisition and activation of plasminogen glycoform II by PAM positive Group A streptococcal isolates. *Biochemistry*. *In press*.

Contributor	**Extent of contribution**
D. M. P. De Oliveira	Protein expression and purification (100%) SPR binding experiments (100%) Western blot analysis (50%) Plasminogen activation (50%) Plasminogen structural analysis (100%) Preparation of figures (100%) Wrote the paper (100%)
R. H. P Law	Purified and prepared plasminogen samples (20%) Edited the paper (5%)
D. Ly	GAS cell surface plasmin acquisition (100%) Western blot analysis (50%) Edited the paper (5%)
S. M. Cook	Active site generation (100%) Plasminogen activation (50%) Edited the paper (5%)
A. J. Quek	Purified and prepared plasminogen samples (80%)
J. D. McArthur	Edited the paper (30%)
J. C. Whisstock	Edited the paper (5%)
M. L. Sanderson-Smith	Directed and supervised research (100%) Cloning (100%) Edited the paper (50%)

Other publications

Martina L. Sanderson-Smith, **David M. P. De Oliveira**, Marie Ranson, and Jason McArthur. (2012). Bacterial Plasminogen Receptors: Mediators of a Multifaceted Relationship. *Journal of Biomedicine and Biotechnology*. **2012:** 272148.

Conference proceedings

David M. P. De Oliveira, Simon M. Cook, Ruby H. P. Law, James C. Whisstock, Jason D. McArthur, Martina L. Sanderson-Smith. Human plasminogen glycoform variants display differences in binding affinity to the Group A streptococcus cell surface receptor Plasminogen Binding M-like Protein (PAM). Plasminogen Activation System in Pathology Workshop. Wollongong, New South Wales, Australia, September 23-25. *Poster presentation.*

David M. P. De Oliveira, Lauren H. Tassell, Christopher J. Day, Mark J. Walker, Michael P. Jennings, Martina L. Sanderson-Smith. Identification of Novel Glycan Interactions Associated With M1 Protein of Invasive Group A Streptococcal Isolate 5448. BacPath 12: Molecular Analysis of Bacterial Pathogens, Moreton Island, Queensland, Australia, September 29 – October 2 2013. *Poster Presentation*.

David M. P. De Oliveira, Ruby H. P. Law, Diane Ly, Simon M. Cook, Jason D. McArthur, James C. Whisstock, Martina L. Sanderson-Smith. Preferential Acquisition and Activation of Plasminogen Glycoform II by PAM Positive Group A Streptococcal Isolates. Lancefield International Symposium on Streptococci and streptococcal Diseases. Buenos Aires, Argentina. November 9 - 12, 2014. *Poster Presentation.*

List of figures

List of tables

Abbreviations

Ω	**Ohms**
[Θ]	**Molar elipticity**
α	**Anti**
ARF	**Acute rheumatic fever**
APSGN	**acute post-streptococcal glomerulonephritis**
BCA	**Bicinchoninic acid**
BSA	**Bovine serum albumin**
˚C	**Degrees Celsius**
CFU	**Colony forming units**
C4BP	**C4b-binding protein**
CD	**Circular dichroism**
Da	**Dalton**
DPBS	**Dulbecco's PBS**
DNA	**Deoxyribonucleic acid**
EDTA	**Ethylenediamine tetra-acetic acid**
εACA	**ε-aminocaprioic acid**
ESI	**Electrospray ionisation**
Fab region	**Fragment antigen-binding region**
FBS	**Fetal bovine serum**
Fc region	**Fragment crystallisable region**
g	**Gram**
g	**Gravitational force**
GAGs	**Glycosaminoglycans**
GAPDH	**Glyceraldehyde-3-phosphate dehydrogenase**
GAS	**Group A streptococcus**
h	**Hours**
HCl	**Hydrochloric acid**
Ig	**Immunoglobulin**
IPTG	**Isopropyl-β-D-thiogalactopyranoside**
k	**Killo**
K_D	**Binding affinity constant**
KOH	**Potassium hydroxide**
KR	**Kringle domain**
L	**Litre**
LB media	**Luria-Bertani media**
LNT	**Lacto-*N*-tetraose**
LnT	**Lacto-neo-tetraose**
LNFP I	**Lacto-*N*-fucopentaose I**
M	**Molar**
m (prefix)	**Milli**
m (suffix)	**Meters**
MFI	**Mean fluorescent intensity**
MHC	**Major histocompatibility complex**
min	**Minutes**
MS	**Mass spectrometry**

m/v	**Mass per volume**
m/z	**Mass to charge ratio**
n	**Nano**
Ni-NTA	**Nickle-nitrilotriacetic acid**
NT	**Northern Territory**
OD	**Optical density**
p	**Pico**
Pa	**Pascal**
PAGE	**Polyacrylamide gel electrophoresis**
PAM	**Plasminogen binding Group A streptococcal M protein**
PBS	**Phosphate buffer saline**
PBST	**Phosphate buffered saline with 0.05% Tween-20**
PCR	**Polymerase chain reaction**
PGC	**Porous graphitised carbon liquid chromatography**
Prp	**PAM related M protein**
PVDF	**Polyvinylidene fluoride**
RU	**Response units**
RPM	**Revolutions per minute**
s	**Seconds**
ScpA	**C5a peptidase**
SDS	**Sodium dodecyl sulphate**
SEN	**Surface enolase**
SF	**Sub-family**
Sfb1	**Fibronectin binding protein I**
SK	**Streptokinase**
SPR	**Surface plasmon resonance**
SpeB	**Streptococcal pyrogenic exotoxin B**
STSS	**Streptococcal toxic shock syndrome**
SV1	**Streptococcal vaccine 1**
TAE	**Tris-acetate EDTA**
THY	**Todd-Hewitt broth 1% yeast**
µ	**Micro**
UV	**Ultra violet**
V	**Volts**
v/v	**Volume per volume**

1. Introduction

1.1 Overview

Group A streptococcus (Group A streptococci; *Streptococcus pyogenes*; GAS) is a human pathogen which can cause a diverse range of both invasive and non-invasive infections. The ability of this Gram positive bacterium to colonise both deep and superficial sites has resulted in the World Health Organisation categorising GAS within the 'top 10' most lethal pathogens, exceeded only by pathogens causing HIV, tuberculosis, malaria and pneumonia (WHO 2005). The Australian Indigenous community living in remote and rural settings such as the Northern Territory (NT), experience significantly higher incidence and mortality rates due to invasive GAS infections than non-indigenous Australians. Reports indicate that GAS infection rates among Aboriginal children in these regions are as high as 70% (Carapetis *et al.* 1999).

M protein is coded by the *emm* gene, and is the major surface protein and virulence factor of GAS. Currently, M protein is used for the classification of GAS, and is a major therapeutic target. GAS mediate binding to a wide range of proteins found within the host's extracellular matrix (ECM) via the M protein, including fibrinogen, plasminogen, C4-binding protein (C4BP), IgG and IgA (Ringdahl *et al.* 2000). To date, over 220 M protein serotypes have been defined, exhibiting sequence variation primarily in the first 10-15% portion of the M protein. As a result of sequence diversity between M protein serotypes the specificity for host proteins is varied. A wide range of M serotypes are associated with invasive GAS infection in the NT of Australia. Study of the sequence identity and binding characteristics of M proteins expressed by invasive clinical GAS isolates from this region may lead to a better understanding of the mechanisms underlying invasive infection.

1.2 Classification of Group A Streptococcus

Streptococcal *emm* and *emm*-like genes which encode cell bound M and M-like proteins are clustered together at the *vir* (also known as *mga*) locus of the GAS chromosome and are divided into the *mrp*, *emm,* and *enn* gene groups on the basis of differences in conserved 5' and 3' regions (Beall *et al.* 1996). M protein, a major surface protein and virulence factor of GAS, is anchored to the cell wall peptidoglycan through the LPxTG motif (Fischetti *et al.* 1990; Barnett *et al.* 2002). Structural characterisation has revealed that the M protein is a fibrillary coiled-coil dimer that extends beyond the GAS cell surface as hairlike projections, typically consisting of 4 repeat regions (designated A to D) that vary in size and amino acid composition (Phillips *et al.* 1981; Smeesters *et al.* 2008). Differences in *emm* nucleotide and translated amino acid sequence, along with *emm* and *emm*-like gene patterns, have given rise to a range of GAS classification techniques which have functioned to categorise over 220 GAS *emm*-types to date (Walker *et al.* 2014).

1.2.1 M-serotyping classification system

M proteins universally exhibit the general structure of an elongated coil coiled alpha-helix containing a conserved C-terminus, with a variable N-terminus. The N-terminus is comprised of a signal sequence that targets protein expression to the cell surface and a series of hypervariable repeat domains (Hollingshead *et al.* 1986). Immunogenic bacterial surface determinants such as the M protein have been historically used to type GAS via serotype-specific antiserum raised against the M protein. The antigenic heterogeneity displayed in the N-terminus of M proteins provides the basis for serological typing via M protein antisera. Even though this form of classification has been used to identify a wide range of serotypes, it is not appropriate to evaluate the genetic diversity of a full length M protein based on this method. Furthermore, M-serotyping is unable to classify a number of isolates owing to their

lack of M-protein expression (Fischetti 1989). To combat this problem, molecular techniques such as vir-typing and *emm*-typing have been employed to provide further discrimination between GAS serotypes.

1.2.2 T-typing classification system

The T protein, encoded by the *tee* gene is a GAS surface protein expressed and presented alongside the M protein. The T protein, named after its observed resistance to trypsin digestion has been shown to be stable at the GAS cell surface throughout the multiple stages of GAS disease (Lancefield *et al.* 1946; Cunningham 2000). Prior studies analysing the sequence homology between *tee* genes found that they exhibit significantly greater sequence diversity compared to *emm* genes (Schneewind *et al.* 1990). Because of these factors, T-proteins have proved to be epidemiologically valuable for sub-classifying strains of GAS. T-typing classification is performed as an agglutination test using T-specific antisera. Initially, GAS are treated with trypsin, enabling the digestion of surface expressed trypsin-sensitive proteins, including the M protein, leaving the T protein exposed for interaction with T-specific antibody. Most (over 95%) GAS strains have well defined T-type antigens which are associated with specific M protein serotypes (Beall *et al.* 1998). When M-type is not identifiable, T-typing serves as an important classification tool, especially in the in the identification of GAS strains associated with disease outbreak (Cunningham 2000).

1.2.3 Vir/Mga-typing classification system

The virulence regulatory network found within the Mga of GAS houses and activates the transcription of several genes, including those for M protein (*emm*) and M-like proteins (*mrp*, *enn*, and *fcR*) (Podbielski *et al.* 1996). Mga expression is regulated by positive feedback, and regulation is controlled via the *mga* gene *(alternatively mry*/*virR).* As a result of intragenic and intergenic recombination events, Mga varies from serotype to serotype. This in turn,

allows for sequence divergence to occur among gene species, and ultimately facilitates the expression of different gene patterns to be housed within an Mga (Whatmore *et al.* 1995). Vir/Mga-typing is a powerful technique which utilises long polymerase chain reactions (PCR) and restriction enzyme digestion to characterise the 5-7 kb *mga* regulon found within the GAS genome (Gardiner *et al.* 1998). This highly discriminatory procedure is able to differentiate between all M-non-type-able strains of GAS. The ability of this technique to be highly effective and reproducible has allowed previously uncharacterised strains to be classified, such as those found within the NT (Carapetis *et al.* 1999).

1.2.4 *emm*-typing classification system

emm-typing is based on the nucleotide sequence of the *emm* gene which encodes the M protein. *emm*-typing is able to differentiate between GAS strains due to insertions, deletions and frameshift mutations within different *emm* genes. Mutational events within the 5' region of *emm* contributes variability within the M protein, and thus provide subsequent grounds for discrimination between GAS strains (Facklam *et al.* 1999). The degree of sequence conservation between *emm* genes, particularly at the 3' end, is thought to be responsible for the cross-reactivity between specific GAS strains and reference antisera. Although the *emm* typing system is a useful and reliable epidemiological tool for classification of GAS, this technique does not evaluate the genetic diversity of a full length *emm* gene/M protein. Previous attempts to link *emm*-type with the phenotypic characteristics of GAS have failed to account for sequence differences downstream of the *emm*-type region. Prior classification systems have failed to take full length M protein sequence into account and thus novel classification techniques based on full length sequence must be implemented.

1.2.5 *emm*-Pattern classification

GAS strains can encode multiple *emm* or *emm*-family genes within the Mga regulon. The number of *emm* and *emm*-like genes, their sub-family content and respective positioning on the chromosome forms the basis of 5 different *emm* patterns designated A through to E (Fig 1.1) (Hollingshead *et al.* 1993). *emm*-patterns A, B and C are typically associated with pharyngeal infections, where D strains are predominately associated with cutaneous skin infections. Pattern E strains have been found to be associated with both pharyngeal and cutaneous skin infections, and as such are termed generalists (Bessen *et al.* 1997; Smeesters *et al.* 2006).

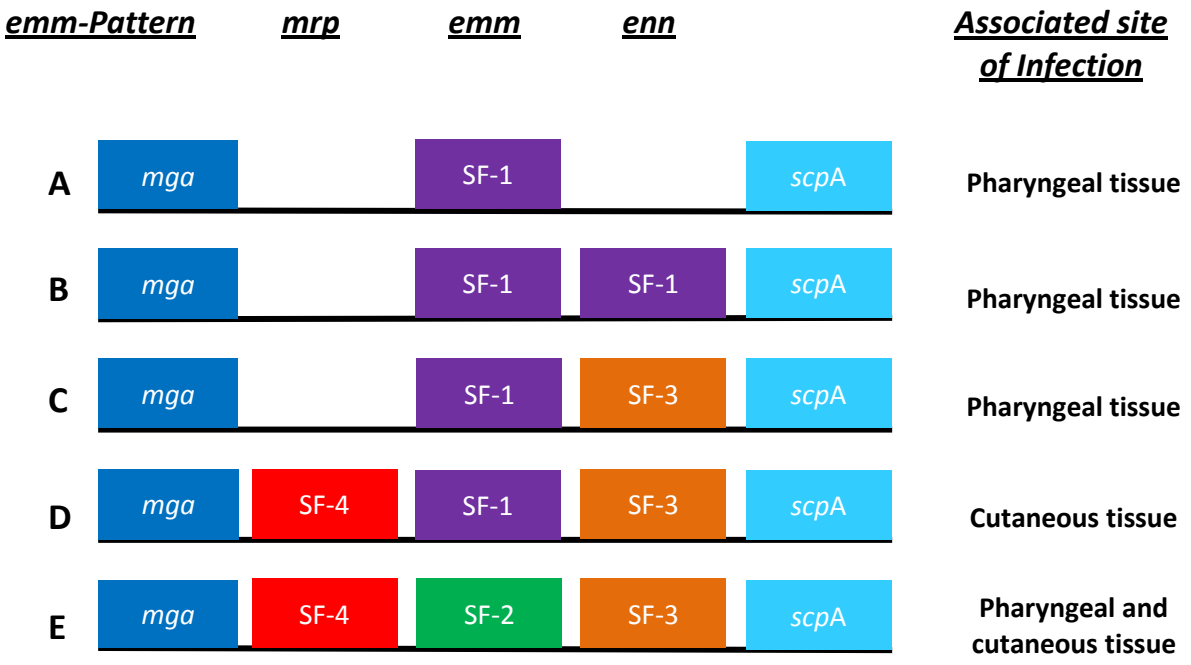


Figure 1.1: Assembly of *emm* and *emm*-like gene patterns, A-E. *emm* pattern groups are classified based on gene amplification using *emm*, *mrp*, or *enn* subfamily (SF) specific primers. *emm* and *emm*-family genes are expressed between *mga* and *scp*A (C5a peptidase) genes. Tissue tropism associated with each *emm*-pattern is indicated on the side. Figure was appended from Hollingshead *et al.* (1993) and Cunningham (2000).

1.3 GAS diseases

1.3.1 Non-invasive GAS diseases

Non-invasive GAS diseases are localised uncomplicated infections of pharyngeal or skin tissue. Common non-invasive GAS infections include pharyngitis, impetigo, and scarlet fever.

Pharyngitis commonly affects mucosal membrane sites such as the pharynx, larynx, and tonsils. Symptoms usually include inflammation of the infected area, headaches, nausea, abdominal pain and vomiting. Rapid antigen detection testing or a throat culture can be undertaken to clarify the diagnosis where by antibiotics such as amoxicillin or erythromycin will be prescribed for treatment purposes (Cunningham 2000). Impetigo usually occurs on areas such as the face and other points of contact where superficial skin layers may be broken. Lesions first begin as vesicles preceding the formation of pustules. Upon rupture, pustules form a yellow crust on the surface of the lesion (Gardiner *et al.* 1998). Scarlet fever often occurs in correlation with pharyngitis. Symptoms of scarlet fever include a characteristic red tongue, sand-paper like rash on the neck, chest, elbows and inner surface of the thighs (Gidaris *et al.* 2008). In 2011, an outbreak of scarlet fever hit Hong Kong where by 600 cases of scarlet fever were reported, with two resulting in death (Hsieha *et al.* 2011). Diseases proceeding scarlet fever such as kidney damage and rheumatic fever are considered much more severe than the original disease itself and although the disease is rarely observed in the Western world, it is still a serious communicable disease with the potential to cause severe epidemics (Gidaris *et al.* 2008).

1.3.2 Invasive GAS diseases

Invasive GAS diseases are typically restricted to sterile areas such as the blood and deep tissue. Although no single GAS serotype has been associated with a specific invasive disease, prior research has identified a general association between serotypes M1 and M3 with invasive infections (Johnson *et al.* 2002).

Cellulitis and bacteremia are the two most common invasive infections caused by GAS. Cellulitis is an infection which affects subcutaneous tissues and is characterised by inflammation of the skin with associated pain and swelling (Bernard 2008; Walker *et al.* 2014). Bacteraemia is defined by the presence of bacteria within the bloodstream which in turn can elicit a vigorous immune response (Bisno *et al.* 1996; Sriskandan *et al.* 2008). The presence of GAS within the blood stream facilitates spread of the bacteria away from the original site of infection, resulting in high fever, blood coagulation and organ failure (Jensen *et al.* 2010). Erysipelas is an acute bacterial infection commonly caused by GAS which affects dermal and hypodermal layers of the skin. Erysipelas infection can be easily defined by raised inflammation around the infected site. Recent studies have identified that between 7-29% of patients suffering from erysipelas experience repeated episodes of the disease (Leclerc *et al.* 2006). Chronic recurrence of erysipelas is considered to be the diseases most frequent complication as each additional episode raises the risk of further deterioration of underlying systemic conditions (e.g. cardiovascular and respiratory conditions), as well as local abnormalities affecting the lymphatic system (Leclerc *et al.* 2006).

Streptococcal toxic shock syndrome (STSS) is characterised by severe hypotension, multiple organ failure, systemic toxicity, and severe local pain making it one of the most severe GAS diseases. Initiation of infection occurs upon entry of GAS via mild trauma to the skin. Disease progression stems from a series of super-antigenic toxins secreted by GAS which

mediate binding of major histocompatibility complex (MHC) class II molecules with T cell receptors. This results in polyclonal T cell activation. The rapid amplification in T cell numbers induces high level release of cytokines within the body resulting in various bodily systems being affected (Holub *et al.* 2004). Necrotizing fasciitis which is normally accompanied by STSS, is a rare disease characterised by such symptoms as rapid necrosis of subcutaneous skin, gangrene and inflammation. Predisposing factors include varicella, penetrating injuries, surgical procedures and mild superficial trauma to the skin (Bisno *et al.* 1996).

1.3.3 Post infection sequelae

Infection with GAS can give rise to serious nonsuppurative sequelae such as acute rheumatic fever (ARF) and acute post-streptococcal glomerulonephritis (APSGN), whereby the highest morbidity and mortality rates are present within the Indigenous population of Australia (Carapetis *et al.* 1996; Jackson *et al.* 2011).

While ARF can be associated with GAS infection of the skin, the disease typically develops following GAS pharyngeal infection, with related symptoms affecting several tissues within the body. Common symptoms of ARF include migratory polyarthritis (inflammation of one or more joints), carditis (inflammation of heart tissue) and Sydenham's chorea (inflammation of the central nervous system). Although migratory polyarthritis is the most common symptom experienced by people suffering from ARF (approximately 75%), carditis is the most severe. Carditis, which forms a symptom of rheumatic heart disease (RHD) significantly increases the risk of endocarditis developing in patients. In rare cases, carditis can lead to heart failure, stroke, and eventually death (Cabell *et al.* 2003).

Molecular mimicry between GAS and host antigens has been a suggested mechanism for the pathology and manifestation of ARF (Johnson *et al.* 1992; Quinn *et al.* 2001). GAS induced

RHD is characterised by the presence of high levels of anti-group A carbohydrate epitope, and N-acetyl-beta-D-glucosamine (GlcNAc) antibodies which are cross reactive with both cardiac myosin in the myocardium and valve endothelium (Galvin *et al.* 2000). Autoimmune B-cell induced antibody responses against the Group A antigen have been correlated to the terminal *O*-linked GlcNAc residue which shares structural similarity to multiple host glycoconjugates (Fung *et al.* 1982). M protein has also been well documented to be involved in initiating RHD due to its highly homologous structure to select host glycoproteins (Ellis *et al.* 2005; Fae *et al.* 2006). T-cells have been shown to cross-react with M protein, cardiac myosin and laminin epitopes allowing T-cell infiltration into valve endothelium, subsequently triggering downstream inflammatory responses (Cunningham 2012). Although the rate of ARF in developed countries is low, rates among the Indigenous population of Australia remain high. Between 2005 and 2010, 98% of all recorded ARF cases in the NT of Australia were for Aboriginal of Torres Strait Islander people. The prevalence of infection during this time period for Indigenous Australians from the NT was 26 times higher than the rate for non-Indigenous people (AIHW 2013).

Acute post-streptococcal glomerulonephritis (APSGN) is commonly associated with GAS throat or skin infections occurring primarily in young children and adults with males affected twice as much as females. The latency period after primary infection may range from 1-4 weeks (Cunningham 2000). Common symptoms of APSGN include the sudden appearance of edema, hematuria, hypertension, urinary abnormalities and decreased serum complement levels (Rodriguez-Iturbe *et al.* 2007). The direct mechanism by which APSGN occurs is unknown, but GAS receptors glyceraldehyde-3-phosphate-dehydrogenase (GAPDH) and streptococcal pyogenic toxin (SpeB) have been implicated in inflammatory and complement deposition processes resulting in immune complex-mediated glomerulonephritis (Batsford *et al.* 2005). Only a sub-set of M-types have been identified as causative agents of

glomerulonephritis. M serotypes commonly associated with skin or throat infections such as 1, 2, 4, 12, 18, 25, 49, 55, 57, and 60 have been found to be highly nephritogenic (Nissenson *et al.* 1979; Bisno *et al.* 1996; Rodriguez-Iturbe *et al.* 2007). The early stages of APSGN are reversible and easily treated. Latter stages of disease progression often require renal replacement accounting for morbidity rates of approximately 1 % in all diagnosed cases (Hricik *et al.* 1998; WHO 2005).

1.4 Global burden of Group A Streptococcal diseases

GAS causes a substantial burden of disease on a global scale, effecting predominantly children and young adults in both developed and non-developed countries. Over 620 million people suffer from non-invasive GAS infections such as pharyngitis and pyoderma each year. Complicated diseases encompassing invasive GAS infections such as necrotizing fasciitis and STSS affect over 660 000 people each year, contributing to approximately 163 000 deaths. It is currently estimated that 15.6 million cases of post-sequelae rheumatic heart disease exist resulting in 233 000 deaths per year (Carapetis *et al.* 2005; WHO 2005). In total, invasive GAS diseases and diseases related to post infection sequelae contribute to over 500,000 deaths per year worldwide (Table 1.1).

The financial impacts of GAS diseases are quite significant in non-invasive diseases such as pyoderma and pharyngitis which eclipse invasive diseases in relation to total cases per year. Costs associated with antibiotic use, symptomatic medications and economic costs arising from absenteeism, place a significant burden on both families and healthcare systems. In the United States, the cost attributed to GAS pharyngitis alone is estimated to be up to $539 million per year (Pfoh *et al.* 2008).

1.5 Disease history and epidemiology

The turn of the twentieth-century coincided with a dramatic decline in the prevalence of serious infection caused by GAS in the Western world, with the majority of GAS infections resulting in only minor non-invasive disease (Carapetis *et al.* 1999). Since the mid-1980's an unexplained resurgence in severe GAS infection has been observed. One of the initial indications of this resurgence was an outbreak of the post infectious sequelae, rheumatic fever, which effected approximately 200 children in the USA over a five year period (Veasy *et al.* 1987). Towards the late 1980's, diseases such as STSS, bacteraemia, and severe invasive GAS skin and soft tissue infections were reported throughout USA and Europe (Pandey *et al.* 2009). In addition, the NT of Australia currently has a 70% incidence of skin infections such as impetigo in children below the age of five, along with one of the highest rates of rheumatic fever in the world (Cook *et al.* 2007; Middleton *et al.* 2014). Even though many M types have been responsible for the varying outbreaks, serotypes M1 and M3 have been found to be the most dominant in Western populations (Aziz *et al.* 2008). The fact that these strains are most frequently isolated from infected patients may reflect an epidemiological shift towards newly sequestered virulence factors within the bacterial species. The mechanisms underlying this epidemiological shift are yet to be fully determined. However, recent research indicates that the acquisition of bacteriophage encoded virulence factors may play a significant role (Walker *et al.* 2007; Davies *et al.* 2014).

Table 1.1: The global burden of Group A streptococcal diseases. Numbers relating to existing cases, new cases each year and morbidity rates were defined from the following regions: Sub-Saharan Africa, Indian Subcontinent, People's Republic of China, other countries in Asia (excluding Japan), Latin America, Middle East and North Africa, Eastern Europe, Pacific Island countries and indigenous Australian and New Zealand, and Established Market Economies. Table was appended from WHO/FCH/CAH/05.07 (WHO 2005).

Disease	Number of existing cases	Number of new cases each year	Number of deaths each year
Rheumatic heart disease	15.6 million	282,000	233,000
History of acute rheumatic fever without carditis, requiring secondary prophylaxis	1.88 million	188,000	-
RHD-related infective endocarditis	-	34,000	8000
RHD-related stroke	640 000	144,000	108 000
Acute post-streptococcal glomerulonephritis	-	472,000	5000
Invasive group A streptococcal diseases	-	663,000	163 000
Total severe cases	**18.1 million**	**1.78 million**	**517 000**
Pyoderma	111 million	-	-
Pharyngitis	-	616 million	-

In low socio-economic regions such as Brazil as well as in the NT, Australia, the epidemiology of GAS infection is less well defined. Infection in these regions is often linked to a variety of GAS serotypes, and neither M1 nor M3 appear to dominate (Stollerman 1997; Smeesters *et al.* 2006). Furthermore, in cases of non-invasive infection, GAS are highly susceptible to antibiotic treatment. Treatment options for invasive disease are limited to extremely high doses of intravenous penicillin and surgical debridement (Walker *et al.* 2014). Despite intensive research efforts, GAS vaccines remain elusive, however, M protein is still the most pursued vaccine target. It is for these reasons that there is a need to characterise the function of diverse M proteins associated with GAS from regions where the burden of infection is the highest.

1.6 M protein

1.6.1 M protein structure

M protein molecules appear as hair-like projections extending from the cell wall through the GAS capsule into the ECM (Fig 1.2). The protein structure itself is known to have an α-helical coiled coil amino acid arrangement whereby the first and fourth amino acids within the chain are hydrophobic making up the protein core, thus allowing for protein stability (McNamara *et al.* 2008). Additional biochemical support for the coiled coil structure of M protein became apparent when partial sequences of M5, M6, and M24 proteins were compared with tropomyosin. The sequences within these three M protein molecules contain a repeating seven residue periodicity of non-polar amino acids, a general characteristic of α-helical coiled coil proteins such as tropomyosin (Manjula *et al.* 1980).

Further studies analysing M protein structure through the use of X-ray diffraction indicated that M proteins were of similar width to tropomyosin and that a coiled coil structure was

present within the M protein (Phillips *et al.* 1981). Irregularities and instabilities throughout the α-helical coiled coil conformation have been found to be required for function, especially in the hypervariable regions of M1, where irregularities in conformation are required for fibrinogen binding, and subsequent downstream pro-inflammatory effects (McNamara *et al.* 2008). Similar structural irregularities occur in myosin and tropomyosin, and may partially explain the patterns of cross-reactivity seen in auto immune sequelae of GAS infection (Ellis *et al.* 2005; Fae *et al.* 2006).

Analysis of sequence data from various M protein serotypes has led to the development of a general model of M protein structure: an extended central coiled coil rod domain flanked by functional end domains. Within the central helical coiled coil region, sequence variances have been observed in M protein molecules. Based on these variances, the central rod region is divided into A, B, C, and D repeat blocks with a pepsin cleavage region between repeat B and repeat C (Fig 1.2) (Fischetti *et al.* 1988). The A and B repeat regions reside immediately downstream of the amino terminus which is exposed to the ECM, while the D repeat domain exists adjacent to the anchor region just above the peptidoglycan membrane of the bacterial cell (Fischetti 1989). Both the A and B repeat domains are highly variable among M protein serotypes. The size and serotype variation between M proteins is based on the number and sequence of repeating units in the A and B repeat regions, with over 200 described to date (ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/tsemm/). The A repeat domain of different M proteins have been found to bind plasminogen, IgA, IgG, C4BP and factor H, whilst the B repeat domain has been found to mediate binding to fibrinogen and fibronectin (Ringdahl *et al.* 2000; Persson *et al.* 2006).

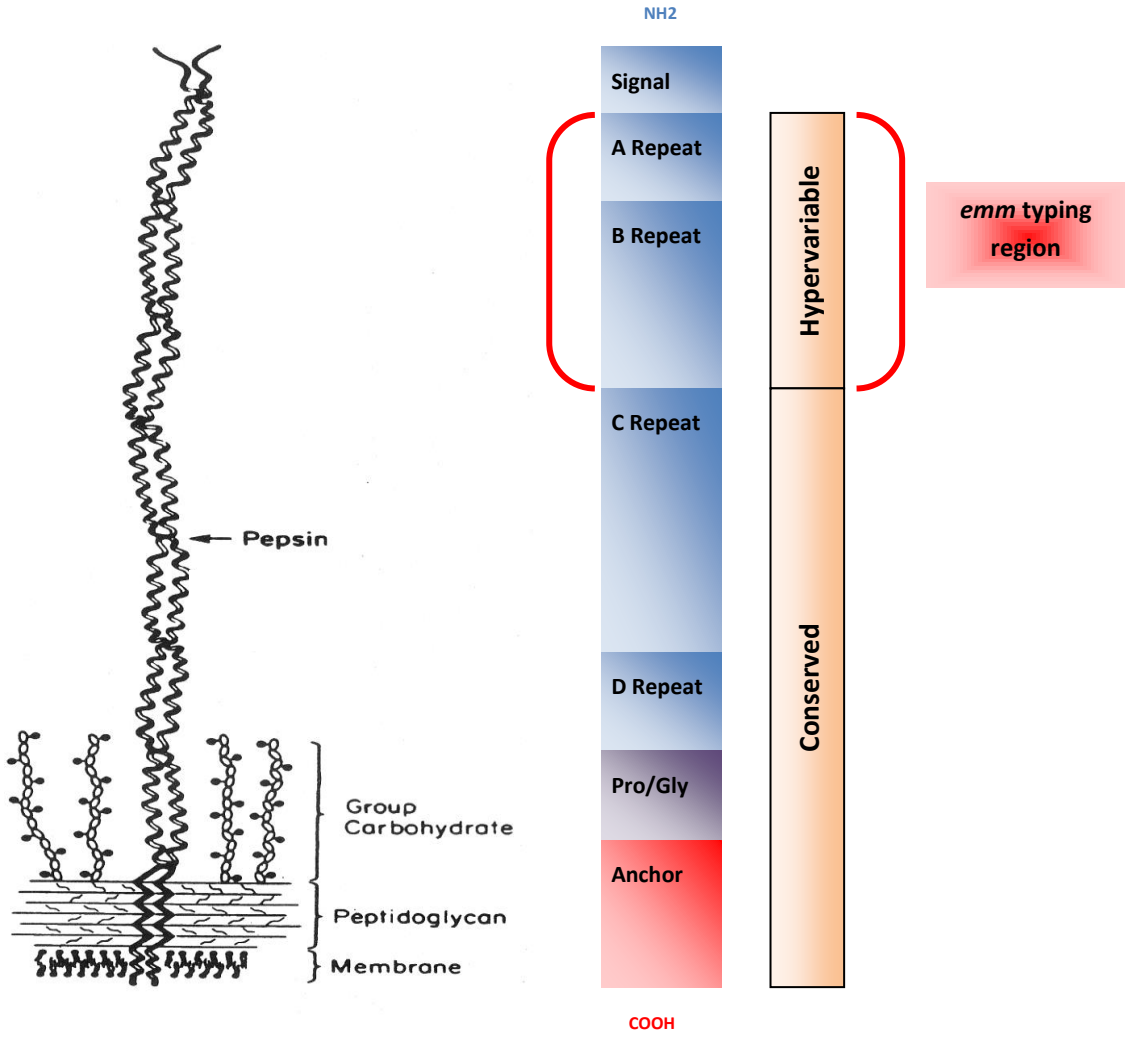


Figure 1.2: Proposed model of GAS M protein. The M protein structure is similar to tropomyosin whereby there is a repeating seven residue periodicity of non-polar amino acids, a general characteristic of α-helical coiled coil proteins. The A and B repeats are located at the N-terminal and are hypervariable among M protein serotypes, providing the basis for *emm*-typing. The C, D and anchor domains situated closer to the C-terminal are conserved across M protein serotypes (Fischetti 1989).

Incorporation of the protein into the peptidoglycan layer of the cell wall is mediated by the C-terminal anchor region. This region is highly conserved among M protein serotypes. It is postulated that the C-terminal cell-wall anchor may be necessary to keep the M protein molecule attached in the right orientation until the peptidoglycan is able to be cross linked around the proline-glycine region of the M protein (Fischetti 1989). This conserved C-terminal region extends throughout one-third of the molecule and aids in binding ECM proteins such as factor H, as well as human serum albumin (Ringdahl *et al.* 2000). The C-terminal homology is shared among most M protein serotypes, suggesting the functional importance of this region (Hollingshead *et al.* 1987).

1.7 Interactions of M protein with host ECM and plasma proteins

GAS pathogenesis involves multiple stages including adherence to, and colonisation of the skin, pharynx. Immediately upon colonisation, the pathogen must adapt to the new environment, necessitating the need to evade the host immune response. For some strains of GAS, colonisation is directly followed by tissue invasion, involving penetration of host mucosal layers. As a major virulence component of GAS, M protein is able to interact with a multitude of host factors involved in each stage of disease, highlighting the importance of M protein in GAS pathogenesis (Table 1.2).

Table 1.2: M proteins and known ligands. Table was adapted from Smeesters *et al.* 2006.

emm-pattern	M-type	Binding location	Ligand	References
A-C	M1	N-terminal	Fibronectin	(Akesson *et al.* 1994; Ben Nasr *et al.* 1995; Cue *et al.* 2001)
		Sub N-terminal	Kininogen	
		Sub N-terminal	Fibrinogen	
		Sub N-terminal	IgG3	
		C-repeat	Albumin	
	M3	N-terminal	Fibronectin	(Reichardt *et al.* 1995; Reichardt *et al.* 1997; Cue *et al.* 2001)
		Sub N-terminal	Fibrinogen	
		C-repeat	Albumin	
	M5	B-repeat	Fibrinogen	(Kotarsky *et al.* 2001; Sandin *et al.* 2006; Waldemarsson *et al.* 2009)
		C-repeat	Factor H	
		C-repeat	Albumin	
	M6	N-terminal	Kininogen	(Horstmann *et al.* 1992; Ben Nasr *et al.* 1995; Fischetti *et al.* 1995; Cue *et al.* 2001)
		N-terminal	Fibronectin	
		Sub N-terminal	Fibrinogen	
		C-repeat	Factor H	
	M12	Sub N-terminal	IgG3	(Retnoningrum *et al.* 1994)
		C-repeat	Albumin	
	M24	Sub N-terminal	Fibrinogen	(Whitnack *et al.* 1985)
	M46	N-terminal	Kininogen	(Ben Nasr *et al.* 1995)
	M55	Unknown	Albumin	(Boyle *et al.* 1995)
		Unknown	Fibrinogen	
		Unknown	IgG1,2,3,4	
D	M33	Sub N-terminal	Plasminogen	(Wistedt *et al.* 1995)
	M41	Sub N-terminal	Plasminogen	(Wistedt *et al.* 1995)
	M52	Sub N-terminal	Plasminogen	(Wistedt *et al.* 1995)
	M53	Sub N-terminal	Plasminogen	(Wistedt *et al.* 1995)
	M56	Sub N-terminal	Plasminogen	(Wistedt *et al.* 1995)
	M98	Sub N-terminal	Plasminogen	(Wistedt *et al.* 1995; Sanderson-Smith *et al.* 2007)
E	M2	N-terminal	C4BP	(Boyle *et al.* 1995; Morfeldt *et al.* 2001)
		Unknown	IgG1,2,3,4	
	M4	N-terminal	C4BP	(Stenberg *et al.* 1992; Johnsson *et al.* 1996)
		N-terminal	IgA	
		Unknown	IgG3	
	M22	N-terminal	C4BP	(Stenberg *et al.* 1992; Berggard *et al.* 2001)
		Sub N-terminal	IgA	
		Unknown	IgG1,2,3,4	
	M28	Unknown	IgA	(Stenberg *et al.* 1992)
		Unknown	IgG1,2,3,4	
	M49	C-repeat	Albumin	(Frick *et al.* 1994)
		Unknown	IgG	
	M50	Unknown	IgG	(Yung *et al.* 1996)
	M60	N-terminal	C4BP	(Stenberg *et al.* 1992; Johnsson *et al.* 1996)
		Unknown	IgA	
		Unknown	IgG3	

1.7.1 The role of M protein in GAS adhesion

Adhesion to the nasopharynx and skin is an important step in GAS virulence, impeding bacterial clearance via cilia and epithelial shedding. M protein mediated binding to epithelial cells has been well documented using human epithelial HEp-2 cells (Wang *et al.* 1994; Courtney *et al.* 1997). Courtney *et al.* have suggested a two-step model of GAS adherence whereby a low affinity primary adhesin, lipoteichoic acid, mediates a weak initial attachment, followed by a stronger more specific M protein or fibronectin binding protein (Sfb1) mediated interaction. Although fibronectin binding proteins are heavily implicated in adherence, clinically important GAS serotypes, M1 and M3, which are associated with invasive infection, do not express these surface proteins (Natanson *et al.* 1995).

In GAS serotypes deficient in Sfb1 such as M1 and M3, respective M proteins have been found to bind fibronectin through hypervariable B repeat domains. The ability of M1 to bind the glycoprotein, fibronectin, facilitates adhesion and internalisation into a range of epithelial cells. M1-fibronectin interactions act as a molecular bridge to bind integrin α5β1 present at the epithelial cell surface, ultimately stimulating internalisation into the host cell. *In vitro* experiments showed that M1 protein can associate with integrin α5β1, but only if the integrin is complexed with fibronectin (Fig 1.3). It is theorised that M1 protein binding changes the structural conformation of fibronectin, in turn affecting the interaction between fibronectin and host cell surface integrins (Cue *et al.* 2001).

Carbohydrates (glycans) have also been shown to play a vital role in cell recognition events, especially those involved in bacterial adherence to host cells (Crocker *et al.* 2007). Mammalian cells are distinguished by a dense diversity of cell surface glycans. Specific surface glycosylation patterns are indicative features of certain cell types such as pharyngal, kidney, and ABO blood group antigens (Arndt *et al.* 2011). The vast structural diversity

coalesced with accessibility at the cell surface makes glycan structures an effective medium for bacterial-host interactions. Protein-glycan interactions have been identified as adherence factors for a variety of commensal and pathogenic bacteria, including *Pseudomonas aeruginosa*, *Helicobacter pylori*, and *Escherichia coli* (Blixt *et al.* 2004; Walz *et al.* 2005; Wu *et al.* 2006).

A specific sub-set of glycans, glycosaminoglycans (GAGs), have been shown to interact with a large cohort of M protein serotypes including clinically relevant M-types M1, M3 and M12. Frick *et al.* (2003) were able to demonstrate that dermatan sulphate, highly sulphated fractions of heparin sulphate, and heparin, are able to bind to a range of M protein serotypes. Strains deficient in M protein surface expression failed to interact with this assembly of GAGs, and significantly reduced adhesion of GAS to GAG expressing human epithelial cells and skin fibroblasts. To date, M protein-glycan interactions and their potential role in GAS virulence have not been completely explored.

1.7.2 M Protein and the host immune response

The propensity of GAS to cause systemic disease reflects the pathogens diverse ability to evade phagocytic defences of the innate immune response. M protein contributes to this via interaction with a range of host factors such as Ig, factor H, fibrinogen and C4BP, which are able to both regulate and conceal GAS from host innate immunity. The role of each of these interactions in GAS infection will be discussed below.

Ig's are glycoprotein molecules which are produced by plasma cells in response to an immunogen, aiding in the neutralisation of foreign objects and pathogens. Although Ig structures vary, they typically consist of two large heavy chains and two small light chains. The base of the two heavy chains is termed the fragment crystallisation (Fc) region. This region interacts with Fc cell surface receptors along with some proteins of the complement

system, subsequently stimulating immune mechanisms such as opsonisation and degranulation of mast cells. There are five different Ig isotype variants (IgA, IgD, IgE, IgG, and IgM) each performing different roles in directing the appropriate immune response against pathogens (Munthe *et al.* 1972).

GAS can be divided into two major groups based on immunoreactivity with a pair of monoclonal antibodies (MAbs) directed at a specific epitope within the M protein. Class I isolates are able to bind one or both MAbs whereas Class II are unable to bind to either. In class I GAS, IgG binding activity is a means to distinguish between nasopharyngeal and impetigo isolates whereas IgA binding activity is a specific property of class II isolates (Bessen *et al.* 1992). Both IgA and IgG bind to the hypervariable N-terminus-repeat regions of the M protein via variable Fc regions. Although the M protein N-terminal is hypervariable, it is still able to bind Ig, suggesting that the tertiary structure of the M protein remains the same despite sequence heterogeneity (Akesson *et al.* 1994). Furthermore, M1 has been found to bind IgG through its S-region (Akesson *et al.* 1994) suggesting that different M protein serotypes may have evolved different IgG binding domains.

M protein-Ig binding interactions are thought to block the recognition of bacterial cells by the immune system, subsequently hiding them from host defences, facilitating the evasion of phagocytic uptake (O'Toole *et al.* 1992; Cunningham 2000). Although GAS Ig binding contributes to bacterial persistence in the host, studies have shown that antibodies raised against Ig binding M proteins are effective against the anti-phagocytic properties of the bacteria (Bessen *et al.* 1988). α-M6-IgA bound to an M6 GAS strain was shown to significantly decrease the adherence of this strain to pharyngeal cells where as α-M6-IgG was able to reduce bacterial internalisation following binding to the M6 protein (Fluckiger *et al.* 1998). The orientation in which Ig, specifically IgG, interacts with M protein has been shown

to be critical in evading phagocytosis. M1 binding the fragment crystallisable (Fc) region of IgG has been shown to allow the bacterium to persist in the host, blocking interaction with phagocytic cells. Alternatively, M1 interaction with the fragment antigen-binding (Fab) domain of IgG, a common interaction in blood, significantly increases the susceptibility of GAS to phagocytic uptake by neutrophils (Nordenfelt *et al.* 2012).

The capacity of certain M proteins to bind fibrinogen and factor H is thought to be an important mode of action in resisting phagocytosis. Factor H is a regulatory component of the alternative complement pathway. Factor H bound to the GAS cell wall inhibits activation of the alternative complement pathway by restricting the deposition of C3b on the cell surface (Cunningham 2000). Abrogation of factor H binding by M protein following deletion of the C1 and C2 domains suggest that factor H binds to the C repeat region of the M protein (Perez-Casal *et al.* 1995). Although factor H binding plays an important antiphagocytic role, it alone cannot prevent phagocytosis. Jack-Weis *et al.* (1982) quantified the amount of C3b deposition on the cell surface of M+ and M- GAS strains and found that even though M- strains exhibited greater levels of C3b deposition on the cell surface, the amount of C3b deposited on M+ strains was still sufficient for phagocytosis. In conjunction with factor H, the binding of fibrinogen has also been shown to assist in restricting phagocytosis (Fig 1.3). Fibrinogen, a well characterised plasma glycoprotein, is composed of three pairs of disulfide-bonded polypeptides with a combined molecular mass of 340 kDa. This soluble plasma protein contains a series of functional domains which interact with a variety of substrates such as hyaluronic acid, thrombin, and fibronectin. Several GAS proteins have also been found to bind fibrinogen, including M proteins and M related proteins (Ringdahl *et al.* 2000). Similar to factor H, it is suggested that fibrinogen is able to block the interaction of GAS with polymorphonuclear leukocytes via disrupting the alternative complement pathway (Whitnack *et al.* 1984). Although not completely understood it is thought that fibrinogen can interfere

with the opsonisation process in several ways. The first system proposes that fibrinogen is able to mask the conserved cell wall components of GAS which are either targets for antibodies or regions capable of fixing microbial opsonin, C3b. An alternative explanation may be that fibrinogen along with C3b competitively binds receptors such as CD11b/CD18 which are major receptors for activated C3 thus preventing output signals required for phagocytosis (Ringdahl *et al.* 2000). Studies analysing M5 and M6 serotypes show that fibrinogen binds to both the type-specific and cross-reactive regions of the M protein. To explain the common anti-phagocytic property of M proteins, it was suggested that the fibrinogen binding site may be conserved among different serotypes (Cunningham 2000).

A key regulator of the classical complement pathway is C4BP. C4BP which is selectively acquired by M protein at the cell surface interferes with membrane-bound C3 convertase (Gigi *et al.* 1979). Interactions between M protein and C4BP are thought to occur at the hypervariable region of M protein. Berggard *et al.* (2001) investigated the capacity of M22 to bind C4BP, and contribute to phagocytosis resistance. Through deletion of the hypervariable region, the ability of M22 to bind C4BP was lost, and phagocytosis resistance was significantly reduced. Furthermore, M protein-C4BP interactions were shown to completely block the effect of opsonising hypervariable-targeted antibodies. The lack of sequence consensus among hypervariable M protein-C4BP binding motifs infers that interactions are dependent on protein structure rather than amino acid sequence identity.

The host zymogen, plasminogen, has also been shown to mediate interactions which facilitate the evasion of the innate immune response. Plasminogen is a single chain glycoprotein zymogen of the broad spectrum serine protease plasmin. Glycosylation of plasminogen gives rise to two major species of the zymogen; glycoform I plasminogen (GI-plasminogen) and glycoform II plasminogen (GII-plasminogen). GI-plasminogen possesses carbohydrate chains

N-linked to Asn_{289} and *O*-linked to Thr_{346}, while GII-plasminogen contains a sole *O*-linked carbohydrate at Thr_{346}. GI-plasminogen and GII-plasminogen exist in humans at an approximate ratio of 2:3, respectively (Law *et al.* 2012). Plasminogen is found within plasma and extracellular fluids at concentrations of ~2 μM (Ponting *et al.* 1992; Andreasen *et al.* 1997). Cleavage of the peptide bond between Arg_{560} and Val_{561} of plasminogen by plasminogen activators results in the formation of active plasmin, a key regulator of GAS invasion (see below: section 1.7.3). Prior studies have demonstrated that plasmin at the GAS cell surface results in the degradation of C3b. Furthermore, plasminogen acquisition at the bacterial surface was shown to prevent C3b opsonisation and phagocytic uptake of GAS by neutrophils (Ly *et al.* 2014).

1.7.3 The role of M protein in invasion and the inflammatory response

To instigate invasive disease, GAS must invade and penetrate host tissues, facilitating the dissemination of GAS away from the initial site of infection. The ability of M protein to interact with multiple host proteins involved in fibrinolysis and inflammation, such as plasmin(ogen) and fibrinogen, is central to invasive disease initiation. Plasmin plays a key role in the fibrinolytic system as it is able to degrade fibrin clots, fibronectin, laminin and proteoglycans. Plasminogen activation is mediated by the plasminogen activators tissue type plasminogen activator (tPA) and urokinase plasminogen activator (uPA) which cleave the 92-kDa protein into a two-chain active plasmin molecule (Ponting *et al.* 1992). Plasminogen has five kringle domains with lysine binding sites located in kringles 1, 2, 4 and 5 which are responsible for the binding of plasminogen to plasminogen receptors (Ponting *et al.* 1992).

The acquisition of plasminogen by GAS can be achieved via two main pathways: direct and indirect binding (Fig 1.3). Over the last 12 years studies have shown that the direct binding of plasminogen is mediated by three known receptors; plasminogen binding M proteins (PAM

and Prp), surface enolase (SEN), and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (Ringdahl *et al.* 2000). Plasminogen binding by M proteins is mediated by two characteristic tandem repeats within the M protein. These 13-amino-acid repeats are designated A1/A2 and have been found to interact with plasminogens kringle 2 domain (Ringdahl *et al.* 2000; Cnudde *et al.* 2006). Plasminogen binding by PAM is mediated by lysine, arginine, and histidine residues within this repeat domain (Sanderson-Smith *et al.* 2006). Similar plasminogen binding motifs have also been identified in other M proteins such as the PAM-related protein (Prp) found in GAS strain NS88.2 (McKay *et al.* 2004). Once bound on the GAS cell surface, plasminogen can be activated by the host activators uPA and tPA, and the secreted GAS plasminogen activator, streptokinase (Sanderson-Smith *et al.* 2012).

The indirect pathway of plasminogen binding utilises a tri-molecular complex consisting of the polymorphic bacterial plasminogen activator streptokinase, plasminogen and fibrinogen (Cook *et al.* 2012). Although streptokinase is secreted by GAS, streptokinase itself has no protease activity, but instead forms a 1:1 complex with plasminogen (Wang *et al.* 1995). The complex is then able to adhere to fibrinogen molecules which have bound to fibrinogen binding receptors on the GAS cell surface. Indirect binding is dependent on the streptokinase-plasminogen activator complex. Once bound to the cell surface, plasminogen can then be converted to the protease plasmin via streptokinase (Walker *et al.* 2005).

Inflammation and systemic reactions that accompany ECM degradation are a hallmark of severe and invasive infection, and are frequently associated with M protein-fibrinogen interactions. Fibrinogen binding by M proteins were first described over 50 years ago when Kantor (1965) found that fibrinogen could be co-precipitated with M protein by acid precipitation. Further studies analysing the capacity of M1 and M5 to bind fibrinogen identified that binding was dependant on the B1 and B2 repeats within the M1 and M5 protein even though they were shown to not share homologous B1 and B2 sequences

(Ringdahl *et al.* 2000). These findings suggest that GAS may have evolved different fibrinogen-binding sequences within their M proteins, while retaining the same function.

M protein-interaction with fibrinogen is associated with inflammatory responses due to the activation of neutrophils. Upon release from the bacterial cell, M proteins are able to form complexes with fibrinogen, binding to β_2 integrins of polymorphonuclear neutrophils (PMN's) subsequently activating them. As a result, PMN's release heparin binding protein (HBP), a mediator of pulmonary vascular leakage (Herwald *et al.* 2004). This in turn helps to explain symptoms associated with GAS infections such as STSS and necrotizing fasciitis, whereby high levels of M protein-fibrinogen complexes are commonly detected within biopsy samples. M proteins also play a dual role as a potent activator of monocytes, inducing the release of pro-inflammatory cytokines such as IL-1β, IL-6, and TNFα through interactions with TLR2 receptors (Påhlman *et al.* 2006). Patients with severe GAS soft tissue infections can experience high concentrations of cytokines (>10 μg/ml) which is enhanced in the presence of HBP (Holub *et al.* 2004). Thus, binding of M protein to fibrinogen initiates a twofold reaction in the inflammatory response (Fig 1.3).

1.8 M protein as a vaccine candidate

M protein is the major GAS cell wall protein, and is expressed by a significant portion of GAS strains. Therefore, a major focus of research in the GAS field is on M protein derived prophylactic vaccines. Whole cell vaccines have been deemed impractical as T cells and antibodies targeting GAS cross-react with human tissue leading to autoimmune diseases such as rheumatic fever, rheumatic heart disease, and glomerulonephritis (Carapetis *et al.* 2005; Smeesters *et al.* 2010).

GAS vaccines centred on the hypervariable N-terminal portion of M protein are currently a major focus in GAS vaccine development. A 26-valent GAS vaccine encompassing the

hypervariable domains of 26 distinct M-types circulating North America has recently passed phase I and II trials (Hu *et al.* 2002; McNeil *et al.* 2005). The geographic distribution and subsequent isolation of these M-types has proven a major limitation in providing cross-protection against a wider range of clinically relevant M-types. Through further epidemiological surveillance of North America and Europe, this GAS vaccine candidate was redefined into a 30-valent vaccine which more adequately represented GAS serotypes associated with pharyngitis and invasive disease (Dale *et al.* 2011). Besides evoking an immunogenic against all 30 vaccine GAS serotypes, the 30-valent vaccine antisera was also shown to contain bactericidal antibodies against 24 of 40 non-vaccine GAS serotypes. One of the major implications of this 30-valent GAS vaccine is the selective pressure is places on non-detectable GAS serotypes to emerge, especially in regions of high *emm*-type diversity.

In industrialised countries M-types 1, 3, 12 and 28 account for approximately 40% of all GAS infections (Steer *et al.* 2009). In non-developed regions such as the NT of Australia, the epidemiology of GAS infection is less defined, whereby *emm* types are constantly expanding and subsiding, dependant on host immune selective pressures (Walker *et al.* 2014). A GAS vaccine based on the hypervariable region of the M protein is well suited for producing protection against specific homologous serotypes in particular industrialised geographic locations, but not for the development of a successful global vaccine encompassing all serotypes. Thus, vaccine strategies targeting the C-terminal domain have tried to overcome these limitations. Recent research has utilised a 29mer-peptide (named J-14) GAS vaccine candidate containing 14 amino acids from the conserved M protein C region enclosed in non-streptococcal peptide sequences designed to retain the native alpha-helix conformation (Hayman *et al.* 2002).

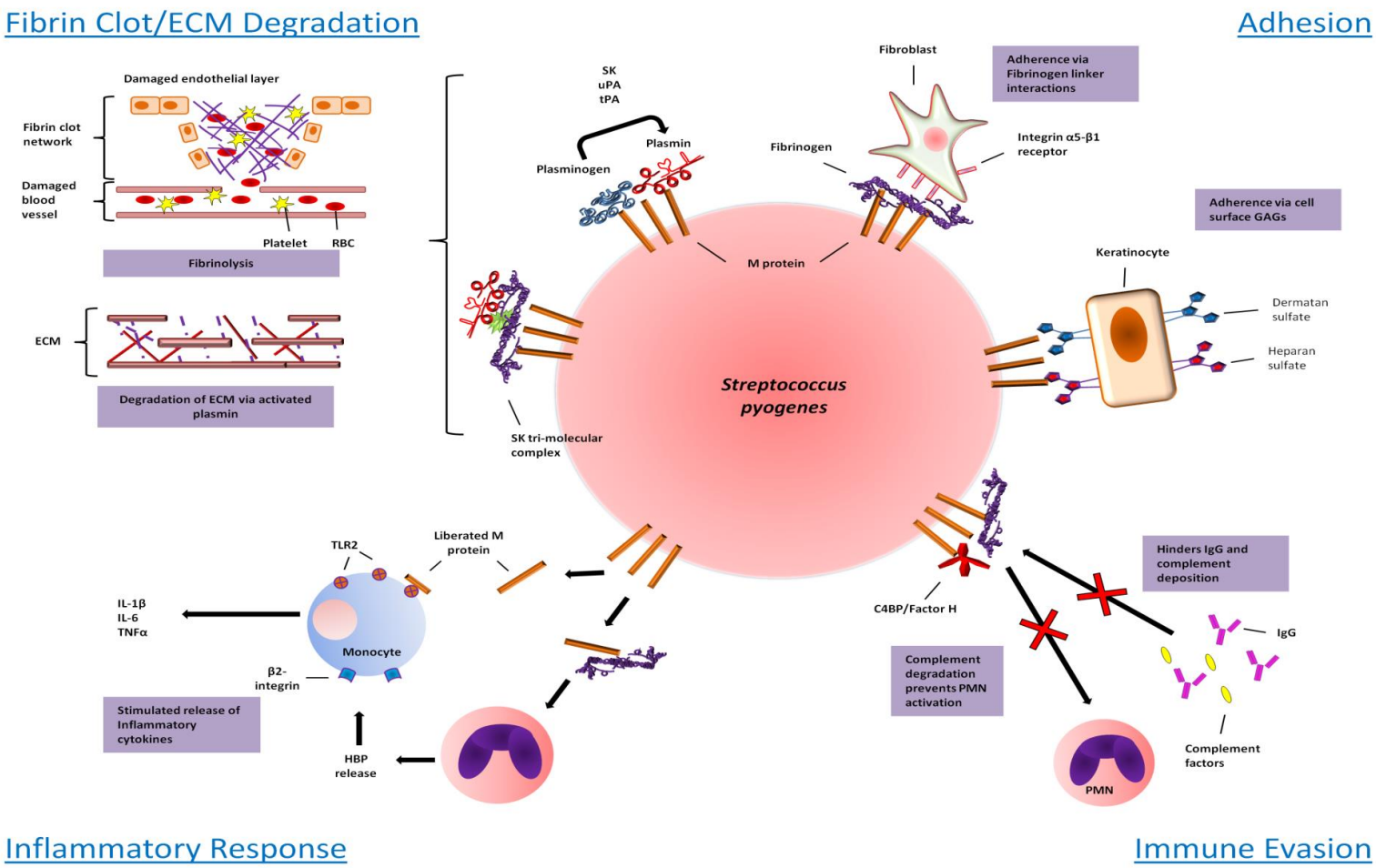


Figure 1.3: Role of M protein in GAS virulence. M protein interactions localised at the cell surface are involved in four main processes; 1) Fibrin clot and ECM degradation; 2) Adherence to host cells via direct and non-direct interactions with host cell surface receptors; 3) Inflammatory response through neutrophil and monocyte activation; 4) Immune evasion via hindering and degrading complement deposition. C4BP: C4 binding protein; ECM: Extra cellular matrix; HBP: Heparin binding protein; IgG: Immunoglobulin G; IL-1β: Interleukin-1β; IL-6: Interleukin-6; PMN: Polymorphic neutrophil; SK: Streptokinase; TLR2: Toll-like receptor-2; TMN: Tumour necrosis factorα; uPA: urokinase plasminogen activator; tPA: tissue-type plasminogen activator. Adapted from Sanderson-Smith *et al.* (2012).

When conjugated to the cholera toxin, the chimeric peptide induces murine antibody responses that kill GAS and protect mice from intraperitoneal and intranasal challenge (Dunn *et al.* 2002). The J-14 epitope, is derived from the C3 repeat domain of M protein. In highly mucoid strains, it is not known whether the C3 repeat domain is able to project far enough beyond the capsule for recognition by antibodies. This is further supported by recent studies which have identified that antibodies raised against the J-14 peptide alone, lack cross-reactivity with M protein antigens across a wide range of GAS serotypes (Penfound *et al.* 2010). To overcome these restrictions, variant J-14-like sequences have been found in multiple M-types, expressed within C1 and C2 repeats domains. A hybrid construct has therefore been developed utilising five common J-14 variant sequences found in 77 different M-types, termed SV1 (Bauer *et al.* 2012).

It is well established that different M-types have been shown to exhibit a diverse range of antigenic and functional properties (Stenberg *et al.* 1992; Johnsson *et al.* 1996; Sanderson-Smith *et al.* 2006). The ability of human plasma proteins to block opsonisation is a critical factor which should be carefully considered in the development of an M protein directed GAS vaccine. Prior research analysing the domains of M5 revealed that binding of M5 B-repeat and C-repeat specific antisera was abrogated by human fibrinogen and human serum albumin respectively (Akesson *et al.* 1994; Retnoningrum *et al.* 1994; Sandin *et al.* 2006). It is also important to note, that different host proteins may be present at different stages of GAS infection. To date, prior work has focused on plasma proteins which may be present in inflamed mucosal and skin infection sites, but it remains to be seen whether host proteins present during early stage GAS infection may interfere with host opsoninic processes (Retnoningrum *et al.* 1994; McArthur *et al.* 2006; Sandin *et al.* 2006). Although there is a great deal of focus on SV1 in relation to its viability in producing an effective immunogenic response, it is unknown whether SV1 can compete out M protein-host interactions required

for GAS virulence. Furthermore, the propensity of SV1 to bind M protein in the presence of other host factors is unknown, which consequently could provide a potential limitation to the candidature of this novel vaccine.

1.9 Aims and Objectives

The vast array of M protein interactions with host proteins has been shown to play a significant role in causing invasive GAS infection. The resurgence of GAS infections over the last two decades has demonstrated that current prevention and treatment techniques are inadequate. Although a large amount of research has been undertaken to understand the interactions between M protein and plasma/ECM proteins, M protein-glycan interactions remain under investigated. In general, M protein studies have focused on the small number of M serotypes responsible for infections in Western countries. The NT of Australia has one of the highest rates of GAS infection in the world and these rates of infection can be attributed to a diverse range of M serotypes. Although epidemiological studies have found positive correlations between host-M protein interactions and invasive disease, the large diversity of M types from this region and their interactions with host proteins remain largely unexplored. Furthermore, no prior study has made comparisons between related M proteins of different M-type and their ability to interact with common host factors, thus identifying if specific M-types are indicative of function and ultimately a specific state of disease. Given that M protein is the major focus of current prophylactic vaccine development for GAS infection, there is a need to better understand the role of diverse M proteins in virulence. As such, the aims of this study were to investigate sequence and binding properties of the M proteins from 26 clinical GAS isolates originating from the NT and other low-income setting regions which are associated with invasive infection. Specifically, the aims of this project are:

I. Conduct sequence analysis of 26 M proteins in order to phylogenetically classify each M protein and identify potential novel binding motifs.
II. Functionally characterise the biochemical interactions between the 26 M proteins and key host proteins
III. Examine the interaction of M protein and host proteins in the presence of SV1
IV. Examine how the different glycosylation profiles of human plasminogen affect the acquisition and activation by select GAS receptors.
V. Conduct preliminary analysis to appraise the role of M proteins in GAS-glycan binding.

The accumulation of sequence and binding data pertaining to these virulent strains may provide insight into the mechanisms that facilitate disease within the host, in turn, allowing potential therapies to be developed which can be used to treat GAS infection and reduce mortality on a global scale.

2. General Materials and Methods

2.1 General Materials

Details of general buffer compositions are listed in Appendix A.

2.2 Bacterial strains and culture methods

2.2.1 Escherichia coli

Escherichia coli (*E.coli*) containing plasmid DNA for DNA sequence analysis or protein expression were grown on Luria-Bertani (LB)-agar plates containing either 100 µg/mL ampicillin (Ap100) or 50 µg/mL kanamycin (Km50). LB-Agar plates were incubated at 37°C for 16 h. Liquid *E. coli* cultures were grown using a single colony of each strain in LB media containing either Ap100 or Km50 at 37°C overnight with shaking at 200 rpm. *E.coli* strains and plasmids used in this study are outlined in table 2.1.

2.2.2 Streptococcus pyogenes

GAS strains used in this study were grown on horse blood agar (HBA) plates (Oxoid, UK) or Todd-Hewitt agar supplemented with 1% (w/v) yeast (THYA) (Difco, Australia). Static GAS cultures were grown overnight in Todd-Hewitt broth supplemented with 1 % (w/v) yeast (THY) at 37°C. GAS containing pDCerm were cultured in THY containing 2 µg/mL erythromycin. GAS strains and plasmids used in this study are outlined in table 2.1.

2.3 DNA manipulations

2.3.1 Plasmid extraction from *E.coli.*

Plasmid DNA was extracted and purified using the Wizard® *Plus* SV Minipreps DNA purification systems kit (Promega, USA). Bacterial cells were grown overnight in a total volume of 10 mL as stated above. Cells were harvested via centrifugation (5000 × *g*, 5 min)

and resuspended in 250 µl cell resuspension solution. Following resuspension, 250 µl of cell lysis solution was added to bacterial samples and mixed by inversion four times. Once cell lysis was achieved, 10 µl of alkaline protease solution was added to the cell lysate and mixed by inverting the microfuge tube 4 times. Following a 5 min incubation at room temperature, 350 µl of neutralisation solution was added to each sample and mixed by inversion. Bacterial lysates were then centrifuged (16 000 × *g*, 10 min) and supernatants transferred to a Wizard spin column. Spin columns were then washed with 750 µl Column wash solution via centrifugation (16 000 × *g*, 1 min). This was then repeated using 250 µl Column wash solution. Spin columns were then transferred to a sterile 1.5 mL Eppendorf microfuge tube and DNA was eluted in 60 µl dH_2O via centrifugation (16 000 × *g*, 1 min).

2.3.2 Chromosomal extraction from *S. pyogenes*

Chromosomal DNA from all GAS strains was prepared using the DNeasy Tissue Extraction Kit (Qiagen, Germany). GAS cells were pelleted from stationary phase cultures by centrifugation (5000 × *g*, 10 min), resuspended in 180 µl of enzymatic lysis buffer and incubated for 30 min at 37°C. Proteinase k (25 µl) and buffer AL (200 µl) were added and the cell suspension incubated at 70°C for 30 min before adding 200 µl of 100% ethanol and mixing thoroughly by vortexing. Mixtures were then transferred to a DNeasy mini column within a 2 ml collection tube and centrifuged (14000 × *g*, 1 min), with the flowthrough discarded. 500 µl of AW1 was then added to the DNeasy mini column and centrifuged (16000 × *g* for 1 min), with flow through discarded. 500 µl of AW2 was then added to the DNeasy mini column and centrifuged (16000 × *g*, 3 min), with flow through discarded. The DNeasy mini column was then transferred into a sterile 1.5 ml microfuge tube and 100 µl of dH_2O was added directly to the column membrane. Following 1 min incubation at room temperature, DNA was eluted by re-centrifugation as above.

Table 2.1: Bacterial strains and plasmid constructs utilised in this study. Where specified, symbols denote plasmid constructs generously provided by Prof. Mark Walker (University of Queensland, Australia) ‡ and Dr. David McMillan (University of the Sunshine Coast, Australia) # or constructed commercially (Sigma-Aldrich, USA)¶.

Bacterial strain or plasmid	Characteristics	References/ Origin
Escherichia coli		
Top10	F-mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 ΔlacX74 recA1 araD139 Δ(ara-leu)7697 galU galK rps L (StrR) endA1 nupG	Invitrogen
BL21(DE3)	*fhuA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5*	Lab stock
Streptococcus pyogenes		
NS696	M1 serotype. *emm*-pattern A. Isolated from uncomplicated infection: Throat (pharyngitis)	(McKay *et al.* 2004; Sanderson-Smith *et al.* 2007)
5448 (M1T1)	M1 serotype. *emm*-pattern A-C. Isolated from invasive infection: Necrotizing fasciitis and toxic shock	(Aziz *et al.* 2004; Anderson *et al.* 2014)
NS13	M53 serotype. *emm*-pattern D. Isolated from invasive infection: Blood (bacteremia)	(McKay *et al.* 2004; Sanderson-Smith *et al.* 2006; Sanderson-Smith *et al.* 2007)
NS88.2	M98 serotype. *emm*-pattern D. Isolated from invasive infection: Blood (bacteremia).	(McKay *et al.* 2004; Sanderson-Smith *et al.* 2007)
NS226	M4 serotype. *emm*-pattern E. Isolated from invasive infection: Wound (cellulitis).	NT of Australia isolate
NS730	M90 serotype. *emm*-pattern E. Isolated from invasive infection: Pus from left hip (necrotizing fasciitis)	(McKay *et al.* 2004)
NS192	M106 serotype. *emm*-pattern E. Isolated from Invasive infection: Blood (bacteremia)	(McKay *et al.* 2004)
NS179	M9 serotype. *emm*-pattern E. Isolated from invasive infection: Pustules on foot (bacteremia)	(McKay *et al.* 2004)
NS8	M85 serotype. *emm*-pattern D. Invasive status and clinical origin unknown.	NT of Australia isolate
NS414	M11 serotype. *emm*-pattern D. Isolated from invasive infection: Wound (cellulitis).	(McKay *et al.* 2004)
NS931	M69 serotype. *emm*-pattern D. Isolated from invasive infection: Blood (necrotizing fasciitis).	(McKay *et al.* 2004)
NS1140	M57 serotype. *emm*-pattern E. Isolated from invasive infection: patient disease status unknown.	NT of Australia isolate
NS178	M54 serotype. *emm*-pattern E. Invasive status and clinical origin unknown.	NT of Australia isolate
NS501	M14 serotype. *emm*-pattern C. Isolated from invasive infection: Blood (bacteremia)	(McKay *et al.* 2004)
NS80	M70 serotype. *emm*-pattern E. Invasive status and clinical origin unknown.	NT of Australia isolate
PRS66	M102 serotype. *emm*-pattern E. Invasive status and clinical origin unknown.	Lab stock

PRS18	M58 serotype. *emm*-pattern E. Invasive status and clinical origin unknown.	Lab stock
PRS2	M2 serotype. *emm*-pattern E. Invasive status and clinical origin unknown.	Lab stock
ALAB49	M53 serotype. *emm*-pattern D. Isolated from impetigo lesion.	(Svensson *et al.* 2002)
ALAB49ΔPAM	ALAB49 derivative with *Pam* replaced with erythromycin resistance gene.	(Svensson *et al.* 2002)
5448 (M1T1) (eGFP)	5448 (M1T1) derivative expressing eGFP via pDCerm	Lab stock
5448ΔM1 (M1T1) (eGFP)	5448 (M1T1) derivative with *emm*1 replaced with chloramphenicol resistance gene. Expresses eGFP via pDCerm	Lab stock
Plasmids		
pGEX2T	Ap^r, ColE1 origin, P *tac* promoter, GST tag, *lacl*q repressor.	GE Healthcare, UK
pET41	Km^r f1 pBR322 origin, T7 promoter, GST tag, his-tag, *lac*l repressor	Novagen, Australia
pET-28b(+)	Km^r, pBR322 origin, T7promoter, his-tag, *lac*l repressor	Novagen, Australia
pQE-30	Ap^r, ColE1 origin, his-tag, T5 promoter, *t*0 and T1 transcriptional terminators	Qiagen,USA
pDCerm	Erm^r, *phoZ*, gram-positive origin of replication	(Jeng *et al.* 2003)
pGEX2T-*emm* 1	pGEX2T derivative containing *emm* coding sequence from GAS strain NS696	(Sanderson-Smith *et al.* 2007)
pGEX2T-*emm* 2	pGEX2T derivative containing *emm* coding sequence from GAS strain PRS2	This study
pGEX2T-*emm 3*	pGEX2T derivative containing *emm* coding sequence from GAS strain SSI-1	This study ǂ
pGEX2T-*emm* 4.2	pGEX2T derivative containing *emm* coding sequence from GAS strain NS226	This study
pGEX2T-*emm* 9	pGEX2T derivative containing *emm* coding sequence from GAS strain NS179	This study
pET41-*emm* 9.1	pET41 derivative containing *emm* coding sequence from GAS strain PRS55	This study #
pGEX2T-*emm* 11	pGEX2T derivative containing *emm* coding sequence from GAS strain NS414	This study
pET41-*emm* 12	pET41 derivative containing *emm* coding sequence from GAS strain PRS8	This study #
pGEX2T-*emm* 14	pGEX2T derivative containing *emm* coding sequence from GAS strain NS501	This study
pET41-*emm* 19	pET41 derivative containing *emm* coding sequence from GAS strain PRS9	This study #
pET41-*emm* 19	pET41 derivative containing *emm* coding sequence from GAS strain PRS9	This study #
pET41-*emm* 48	pET41 derivative containing *emm* coding sequence from GAS strain PRS15	This study #
pGEX2T-*emm* 53	pGEX2T derivative containing *emm* coding sequence from GAS strain NS13	(Sanderson-Smith *et al.* 2006)
pGEX2T-*emm* 54	pGEX2T derivative containing *emm* coding sequence from GAS strain NS178	This study
pET41-*emm* 54	pET41 derivative containing *emm* coding sequence from GAS strain TVU5	This study #
pGEX2T-*emm* 57	pGEX2T derivative containing *emm* coding sequence from GAS strain NS1140	This study
pGEX2T-*emm* 58	pGEX2T derivative containing *emm* coding sequence from GAS strain PRS18	This study

pET41-*emm* 60	pET41 derivative containing *emm* coding sequence from GAS strain PRS20	This study [#]
pGEX2T-*emm* 69	pGEX2T derivative containing *emm* coding sequence from GAS strain NS931	This study
pGEX2T-*emm* 70	pGEX2T derivative containing *emm* coding sequence from GAS strain NS80	This study
pET41-*emm* 83	pET41 derivative containing *emm* coding sequence from GAS strain PRS30	This study [#]
pGEX2T-*emm* 85	pGEX2T derivative containing *emm* coding sequence from GAS strain NS8	This study
pGEX2T-*emm* 90	pGEX2T derivative containing *emm* coding sequence from GAS strain NS730	This study
pET41-*emm* 97	pET41 derivative containing *emm* coding sequence from GAS strain 88/30	This study [#]
pGEX2T-*emm* 98.1	pGEX2T derivative containing *emm* coding sequence from GAS strain NS88.2	(Sanderson-Smith *et al.* 2007)
pGEX2T-*emm* 102	pGEX2T derivative containing *emm* coding sequence from GAS strain PRS66	This study
pGEX2T-*emm* 106	pGEX2T derivative containing *emm* coding sequence from GAS strain NS192	This study
pET-28b(+)-*emm* 1-AB	pET-28b(+) derivative containing A and B repeat *emm* coding sequence from GAS strain NS696	This study
pET-28b(+)-*emm* 1-AC1	pET-28b(+) derivative containing A, B and C1 repeat *emm* coding sequence from GAS strain NS696	This study
pET-28b(+)-*emm* 1-A92C1	pET-28b(+) derivative containing *emm* coding sequence beginning from the first 92 nucleotides of the A repeat up to and including the C1 repeat from GAS strain NS696	This study
pET-28b(+)-*emm* 1-B	pET-28b(+) derivative containing B repeat *emm* coding sequence from GAS strain NS696	This study
pET-28b(+)-*emm* 1-BC1	pET-28b(+) derivative containing B repeat *emm* coding sequence from GAS strain NS696	This study
pET-28b(+)-*emm* 1-C	pET-28b(+) derivative containing C repeat *emm* coding sequence from GAS strain NS696	This study [¶]
pQE-30-SK_{ALAB49}	pQE-30 derivative containing *ska* coding sequence from GAS strain ALAB49	(Cook *et al.* 2012)

2.3.4 Polymerase chain reaction conditions

Polymerase chain reaction (PCR) amplification of all *emm* genes was performed using an Eppendorf Mastercycler thermal cycler (Eppendorf, Germany). PCR reaction mixtures consisted of 1 × PfuUltra™ II reaction buffer (Stratagene, USA**),** 1 U of PfuUltra™ II Fusion HS DNA polymerase (Stratagene, USA), 0.25 mM deoxynucleoside triphosphates (dNTPs), 50 pmol each of forward and reverse primer, and 200 µg of GAS chromosomal DNA, made up to a final volume of 50 µl with sterile dH_2O. The thermocycling conditions of PCR consisted of 35 cycles. Each cycle consisted of a 1 min denaturation step performed at 95°C,

a 15 sec annealing step performed at 55°C, and a 1 min extension step performed at 72°C. A final extension cycle of 4 min at 72°C was also performed. PCR screening of *E. coli* transformants for cloning was conducted with reaction mixtures consisting of 2 U of Mango *Taq* polymerase(Bioline, Australia), 1 × Mango *Taq* reaction buffer (Bioline, Australia), 0.25 mM of each dNTP, 50 pmol of each forward and reverse primer, and 2 mM $MgCl_2$ made up to a final volume of 50 µl with sterile dH_2O. A single bacterial colony was used as the DNA template. PCR thermocycling conditions were carried out as above. Details of primer sequences used for all PCR's are provided in appendix B.

2.3.5 Agarose gel electrophoresis

DNA amplicons were resolved by electrophoresis on a 1% agarose gel for 1 h at 90 V in 1× TAE electrophoresis buffer in a Bio-Rad Minisub™ (Bio-Rad, USA). DNA bands were visualised by ethidium bromide staining and recorded using the Bioimaging UVP E3C system (Biostrategy, USA). Hyperladder I molecular size marker (Fermentas, Canada) was used to determine the approximate size of DNA bands.

2.3.6 DNA extraction from agarose gels and PCR's

DNA amplified by PCR or subjected to agarose gel electrophoresis was extracted and purified using the Wizard® SV Gel and PCR clean-Up system (Promega, USA). Following electrophoresis, DNA bands of interest were excised and placed in a 1.5 mL microfuge tube. An equal weight/volume of membrane binding solution was added and allowed to incubate at 50-65˚C until the excised gel slice was completely dissolved. Following PCR, equal volumes of membrane binding solution were added to PCR mixtures and incubated at room temperature for 1 min. Dissolved gel and PCR mixtures were then transferred to an SV Minicolumn assembly and allowed to incubate for 1 min at room temperature. Following incubation, samples were centrifuged (16 000 × *g*, 1 min) and the flow through discarded. SV

Minicolumns were then washed with 700 µl Membrane wash solution and the flow through discarded. This was then repeated with 500 µl Membrane wash solution, and centrifuged as above. The SV Minicolumn was then transferred to a sterile 1.5 mL microfuge tube. To elute the DNA, 50 µl dH_2O was added to the SV Minicolumn and incubated at room temperature for 1 min. DNA was eluted from the SV Minicolumn via centrifugation (16 000 × *g*, 1 min).

2.3.7 DNA sequence analysis

emm amplicons were directly sequenced from products of PCR amplification. Each sequencing reaction consisted of 1 pmol of a single strain/plasmid specific primer (appendix B), 200 µg of purified PCR product or plasmid, 1 U of BigDye v 3.1 ready reaction mix (ABI, Australia), 1 U of sequencing buffer (ABI, Australia) made up to a final volume of 10 µl with sterile dH_2O**.** 25 cycles of PCR were performed with an Eppendorf Master Cycler (Eppendorf, Germany), consisting of a 1 min denaturation step at 96°C, a 15 sec annealing step at 55°C, and a 1 min extension step at 60°C. Resulting PCR products were purified using the Wizard® PCR clean-up kit (Promega, USA) as described above. Purified PCR products were sequenced by Margret Philips (University of Wollongong, Australia) using a 3130*xl* Genetic Analyser (Applied Biosystems, USA) as per manufacturer's specifications.

2.3.8 Nucleotide and protein sequence identity analysis

In order determine the level of sequence identity between *emm* genes, and between M proteins, nucleotide and protein sequence analysis was undertaken using MUSCLE with default parameters applied by Geneious *version 6.0* (Biomatters, USA). Prior to sequence analysis, *emm* gene sequences and translated M protein amino acid sequences were trimmed from the first codon/amino acid of the mature protein to the first codon/amino acid of the D repeat domain adjacent the LPXTG motif. Phylogenetic analysis between mature M proteins were inferred under a Blossum80 matrices using a Jukes-Cantor genetic distance model and a

neighbour joining tree method via Geneious *version 6.0* (Biomatters, USA). Strain collection data along with GenBank locus description of each *emm* gene in study are provided in appendix C.

2.4 *emm* gene cloning

2.4.1 Restriction enzyme digestion

Plasmid DNA and PCR amplicons were routinely digested to facilitate cloning and confirm the presence of insert DNA in to designated vectors: pGEX2T and pET-28b(+). Restriction enzyme digestion mixtures for cloning in to pGEX2T consisted of 2 × of Tango buffer ™ (Fermentas, Canada), 1 U of *Bam*HI (Fermentas, Canada), 0.5 U of *Eco*RI (Fermentas, Canada), and 300 ng of DNA made up to a total volume of 20 µl with dH_2O. Restriction enzyme digestion mixtures for cloning in to pET-28b(+) were made up of 2 × Tango buffer ™ (Fermentas, Canada), 1 U of *Nco*I (Fermentas, Canada), 1 U of *Xho*I (Fermentas, Canada), and 300 ng of DNA made up to a total volume of 20 µl with dH_2O. Digestions were carried out at 37°C for 2 h. Following digestion, mixtures were incubated at 80°C for 15 min to inactivate restriction enzymes.

2.4.2 Ligation

The amount of vector to insert used in each ligation was kept at a ratio of 3:1. DNA quantities were calculated using the following formula:

Amount of insert (ng) = 3/1 × [size of insert (kb)/size of vector (kb)] × 100 ng vector

To facilitate ligation, digested vector and insert mixtures were made to a total volume of 30 µl with dH_2O. Ligation mixtures were incubated at 60°C for 5 min, 37°C for 10 min, room temperature for 10 min, and left on ice for 5 min. 3 × ligase buffer (Fermentas, Canada) and 1

U T_4 DNA ligase (Fermentas, Canada) was then added to the ligation mixture and incubated over night at 10°C.

2.4.3 Preparation of electro-competent *E.coli*

Single colonies of *E.coli* were cultured in 10 mL LB media and grown overnight at 37°C with shaking at 200 rpm. Overnight 10 mL cultures were inoculated into 100 mL cultures, and were allowed to grow at 37°C with shaking until an optical density (OD_{600nm}) 1.0 was reached. Bacterial cells were harvested via centrifugation (5000 × *g*, 10 min) and washed twice in 80 mL of cold sterile dH_2O. Cell resuspensions were then pelleted as above (5000 × *g*, 10 min) and washed once more in 80 mL of cold sterile 10% glycerol. After washing, cell pellets were resuspended in 2 mL of cold sterile 10% glycerol, remaining on ice until transformation.

2.4.4 Transformation of electro-competent *E. coli*

Electro-competent *E.coli* were transformed using a Bio-Rad Gene Pulser (Bio-Rad, USA) according to the manufactures instructions. Transformation mixtures consisted of 3.3 µl of ligation mixture, 80 µl of electro-competent *E. coli* (Top10 or BL21(DE3)cells) and 8 µl of dH_2O in an ice cold cuvette. The Gene Pulser ^{TM} electroporator (Bio-Rad, USA) was set to 2.5 kV, 25 µ FD, and 200 Ω. After electroporation, cells were then transferred to preheated (37°C) LB broth and incubated at 37°C for 1 h. Aliquots of 50 µl, 150 µl, and 300 µl of transformation mixture were plated out on LB agar plates containing Ap100 or Km50 and grown overnight at 37°C. Single colonies were then screened via PCR, using cloning primers and DNA sequence analysis of the resulting amplicons as previously described.

2.5 M protein expression

2.5.1 Full length M protein expression

The fundamental aspects of M protein purification were carried out as previously described (Smith *et al.* 1988) with minor modifications. Initial 100 mL starter cultures of LB media containing Ap100 were inoculated with a single colony of *emm*-pGEX2T *E. coli* transformants and grown overnight at 37°C with shaking (200 rpm). LB supplemented with Ap100 made to a total volume of 900 ml was then inoculated with 100 mL stationary phase *E. coli* and incubated at 37°C with shaking until OD_{600nm} reached 0.6. At OD_{600nm} 0.6, phenylmethylsulfonyl fluoride (PMSF) and isopropyl β-D-1-thiogalactopyranoside (IPTG) were added to final concentrations of 1 mM and 0.1 mM respectively. PMSF was again added 2 h after induction. After a total of 4 hours post induction, bacterial cells were pelleted by centrifugation (4000 × *g*, 20 min) and resuspended in 50 ml of ice cold cell lysis buffer. Samples were then incubated on ice for 10 min and lysed by applying the bacterial cell suspension 3 times to an Avestin EmulsiFlex-C5-Homogeniser (ATA Scientific, Australia) at 120 000 kPa. The resulting lysate was then rotated end over end for 30 min at 4˚C. The lysate was then centrifuged (12 000 × *g*, 10 min) to remove cellular debris and the supernatant filtered using a 0.45 μm filter (Millipore, USA). Filtered lysate was then applied to a glutathione-agarose column (Sigma-Aldrich, USA) with a bed volume of 2 ml. The column was then washed with 50 ml of PBS at 4°C and 100 U of thrombin (Sigma-Aldrich, USA) was then run into the column and incubated overnight at 4°C. Cleaved protein was eluted from the column by washing the bed with 3 volumes of PBS. A 2 ml Ni-NTA column was then equilibrated in PBS, and each sample containing cleaved protein was re-applied to the column a total of 4 times. The Ni-NTA column was then washed with 20 ml of PBS and recombinant protein was eluted using 4 column volumes of native elution buffer. Samples of

the recombinant protein purification process were analysed via 12% SDS-PAGE analysis (described below: section 2.5.3).

2.5.2 Fragmented M-protein expression

M1 protein fragments were expressed and purified as previously described (Macheboeuf *et al.* 2011). Initial 100 mL starter cultures of LB media containing Km50 were inoculated with a single colony of *emm1AB; A92C1; AC1; B; BC1; C*-pET28b(+) *E. coli* transformants and grown, cultured and lysed as described above. The filtered lysate was then incubated with 2 mL of Ni-NTA resin equilibrated in PBS for 1 h at 4˚C with end-over-end rotation. Following incubation, lysate containing Ni-NTA resin was applied to a 15 mL column and washed with 20 mL PBS. Recombinant M1-fragments were eluted using 4 column volumes of native elution buffer. Samples of the recombinant protein purification process were analysed via 12% SDS-PAGE analysis (described below: section 2.5.3).

2.5.3 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS- PAGE)

SDS-PAGE analysis was routinely used to analyse/confirm protein expression and purification. Acrylamide gel ranging from 7-12% was poured into a MINI PROTEAN® 2 Cell system (Bio-Rad, USA) and overlayed with 100% ethanol. The gel was allowed to set for 30 min. Following rinsing with dH_2O, gels were overlayed with 4% stacking gel, and a 0.75 mm wide comb was inserted to enable well formation. The gel was allowed to set for 30 min. Protein samples were then mixed with an equal volume of reducing cracking buffer, and boiled for 10 min. Samples were loaded into wells and the gel was run at 160 V for 1 h in 1 × SDS-PAGE running buffer. Visualisation of protein bands was conducted after staining with Coomassie rapid stain solution overnight with slight agitation. Protein bands were resolved upon addition of rapid de-stain solution. PageRuler Prestained Protein Ladder (Fermentas,

USA) and Broad Range Protein Marker (Bio-Rad, USA) were used as molecular weight markers to assess protein size and concentration.

2.5.4 Protein quantification

For proteins such as the M protein which typically lack aromatic amino acid residues, protein concentration was determined using BCA protein assay (Sigma-Aldrich, USA). A 25 µl sample of each protein dialysed in PBS was added, in triplicate, to a 96-well microtitre plate. BSA standards of known concentration (0 – 1 mg/ml) dialysed in PBS, were also added to the microtitre plate in triplicate. Colorimetric determination of protein concentration was initiated upon addition of 196 µl of reagent A, and 4 µl of reagent B to each sample. Microtitre plates were then incubated at 37°C for 30 min to allow the reaction to proceed. Individual samples were then read at 560 nm in a Spectramax 250 plate reader (Molecular Devices, USA). Protein concentrations were extrapolated from a BSA standard curve using Softpromax software (Molecular Devices, USA). Quantification of all proteins other than M protein was undertaken using a NanoDrop 2000c (Thermofisher, Australia). Following automated calibration, 2 µl of each protein sample was loaded on to the pedestal and absorbance was read at 280 nm.

2.6 Western transfer analysis

Using a Mini Trans-Blot apparatus (Bio-Rad, USA), Western transfer analysis was routinely conducted to confirm the presence of M protein and human plasminogen following SDS-PAGE analysis. Cassettes were set up so that protein would travel from the negatively charged anode to the positively charged cathode. Once loaded in to the gel tank, and filled with Western transfer buffer, proteins were transferred at 100 V for 1 h at 4˚C with constant stirring to ensure even buffer temperature and ion distribution. Post-transfer, polyvinylidene

fluoride (PVDF) membranes were blocked in PBST with 10% (w/v) skim milk powder (Difco, Australia) overnight at 4˚C. PVDF membranes were then washed twice for 5 min in PBST and probed using either rabbit α-human plasminogen antibody (Calbiochem, USA) or rabbit polyclonal α-PAM_{NS13} antisera diluted 1:3000 and 1:30000 in PBST with 1% skim milk powder (Difco, Australia) respectively. Previous studies have shown that rabbit polyclonal α-PAM_{NS13} antisera antibody can recognise multiple M proteins, rather than PAM specifically, and does not detect proteins other than M proteins (Sanderson-Smith *et al.* 2008). After incubation for 1.5 h with agitation at room temperature, membranes were then washed three times in PBST. Both M protein and human plasminogen recognising primary antibodies were detected using goat α-rabbit IgG horseradish peroxidase conjugate (Invitrogen, USA) diluted 1:3000 in PBST containing 1% skim milk powder (Difco, Australia) for 1 h at room temperature with agitation. Unbound secondary antibody was washed from the membrane with three 5min washes of PBST, followed by one 5 min wash in PBS. Enhanced chemiluminescence detection (Pierce, Biotechnology, Ill, USA) was used to visualise protein bands of interest. Exposed X-ray films were scanned using a GS-800 calibrated densitometer (Bio-Rad, USA).

2.7 Circular dichroism

To identify potential secondary structure variation in all M proteins and M protein fragments, far UV circular dichroism (CD) spectroscopy was carried out. CD spectra were acquired using a Jasco J-810 Spectropolarimeter (Jasco, Canada). Analysis was undertaken at room temperature from 180 - 250 nm in a 0.1 cm pathlength cell containing 300 µl of protein solution in 10 mM sodium phosphate buffer (pH 7.4). Scanning mode was set to continuous, with a response time of 2 sec and a band width of 1 nm. Data pitch was set to 1, with

recorded data representing the average of 6 scans, corrected for buffer baseline. Molar residue ellipticity ([θ]) was calculated using the following formula:

[θ] = (θ × mean residue weight) / (pathlength in millimetres × concentration in mg/mL) (Greenfield 2006).

The percentage of α-helix for each protein was estimated from ellipticity at 222 nm using the following formula:

% α-helix = - ($\theta_{222\,nm}$ – 4,800) / 45,400 (Phillips *et al.* 1981).

2.8 Functional characterisation of recombinant M protein

Functional analysis of clustered M proteins was undertaken to assess binding to key host proteins which have been previously shown to interact with multiple distinctive M proteins. Binding interactions were analysed using surface plasmon resonance on a BIAcore T200 (GE Healthcare, Sweden).

2.8.1 Surface plasmon resonance – Plasminogen

Purified histidine-tagged M protein was analysed for binding affinity to human glu-plasminogen (Haemotologic Technologies Inc., USA) via single cycle kinetics, using a Biacore T200 (GE Healthcare, Sweden) at 20°C. α-histidine monoclonal antibody (Abcam, Australia) was immobilised to a series S CM5 sensor chip (Biacore AB) using an amine coupling kit as per the manufactures instructions (Biacore AB). Briefly, the chip was activated with a 1:1 mixture of 0.2 M *N*-ethyl-*N'*-(3-dimethylaminopropyl) carbodiimide and 0.05 M *N*-hydroxysuccimide. To capture M protein at the surface, α-histidine monoclonal antibody was coated onto the chip at 100 µg/ml in 10 mM sodium acetate (pH 4) to a level of 10000 response units (RU). Unoccupied binding sites were blocked using 1 M ethanolamine

(pH 8.5). Histidine-tagged M protein was captured at the surface of flow cells 2, 3 and 4 until a total of 80-100 RU was reached. Flow cell 1 was left blank to serve as a control. Analytes were diluted into running buffer (PBS, 0.05% Tween-20, pH 7.4), and kinetic assays were performed using human glu-plasminogen at varying concentrations (0-120 nM), over a series of five 60 s injections at a flow rate of 30 µl/min with a 900 s dissociation period. Regeneration of the surface was achieved with 10 mM glycine-HCl (pH 1.5) for 30 s at 30 µl/min. M protein-glu-plasminogen interactions were analysed by non-linear fitting of the single cycle kinetic sensograms according to a 1:1 Langmuir binding model using Biacore T200 evaluation software (Biacore AB).

2.8.2 Surface plasmon resonance – Fibrinogen, IgG, IgA, albumin and C4BP

Purified histidine-tagged M protein was analysed for binding affinity to human fibrinogen (Sigma-Aldrich, Australia) IgG (Life Technologies, Australia), IgA (Abcam, Australia), albumin (Sigma-Aldrich, Australia) and C4BP (Athens Research and Technology, USA**)** via single cycle kinetics, on a series S Ni-NTA chip (BIAcore AB), using a Biacore T200 at 20 °C. All four flow cells were activated with 0.5 mM $NiCl_2$ for 60 s at 5 µl/min and washed with 3 mM EDTA for 60 s at 5 µl/min. M protein was captured at the surface of flow cells 2, 3 and 4 until 100-200 RU was reached. Flow cell 1 remained as a blank control. Analytes were diluted into running buffer (PBS, 0.05% Tween-20, 50 µM EDTA, pH 7.4), and kinetic assays were performed using analyte at varying concentrations (0-200 nM) over a series of five 60 s injections at a flow rate of 30 µl/min with a 900 s dissociation period. Regeneration of the flow cell surface was achieved with two separate injections consisting of 50 mM NaOH and 350 mM EDTA (pH 8.3) for 50 s at 30 µL/min. M protein-host interactions were analysed as described above.

2.9 α-SV1-M protein binding interactions

The ability of rabbit-α-SV1 IgG (α-SV1 IgG) to bind phylogenetically distinct M proteins in the absence and presence of host protein was analysed using surface plasmon resonance on a BIAcore T200 (GE Healthcare, Sweden). α-SV1 IgG was produced and graciously provided by Dr. David McMillan (University of the Sunshine Coast, Australia).

2.9.1 Competitive binding analysis

Binding affinity interactions between individual M proteins and α-SV1 IgG was carried out as stated above. Competitive binding analysis between M protein and α-SV1 IgG was assessed using host proteins plasminogen, fibrinogen, albumin, IgG and IgA. Respective host proteins were selected to enable coverage of all potential binding motifs across each repeat domain within selected M proteins. Analysis was carried out via single cycle kinetics, on a series S Ni-NTA chip (BIAcore AB), using a Biacore T200 at 20 °C. All four flow cells were activated with 0.5 mM $NiCl_2$ for 60 s at 5 µl/min and washed with 3 mM EDTA for 60 s at 5 µl/min. M protein was captured at the surface of flow cells 2, 3 and 4 until 100-200 RU was reached. Flow cell 1 remained as a blank control. For competitive binding analysis, either host proteins or α-SV1 IgG was captured on flow cells 2, 3 and 4 until saturation of M protein occurred. Host proteins and α-SV1 were diluted into running buffer (PBS, 0.05% Tween-20, 50 µM EDTA, pH 7.4), and competitive kinetic assays were performed using analytes at varying concentrations (0-1000 nM) over a series of five 60 s injections at a flow rate of 30 µl/min with a 900 s dissociation period. Regeneration of the flow cell surface was achieved as previously described. Affinity interactions were analysed as described above.

2.10 Plasminogen glycoform interactions with PAM expressing GAS

Examination of how the varying glycosylation profiles of human plasminogen affect acquisition and activation by respective GAS plasminogen receptors, PAM and type-2b SK was analysed using host proteins, GI-plasminogen, GII-plasminogen and GII-angiostatin which were kindly provided by Prof. James Whisstock and Dr. Ruby Law (Monash University, Australia).

2.10.1 Surface plasmon resonance - PAM binding analysis

Recombinant histidine-tagged PAM was analysed for binding affinity to purified human GI-plasminogen, GII-plasminogen and GII-angiostatin via single cycle kinetics, on a series S Ni-NTA chip (GE Healthcare, Sweden), using a Biacore T200 (GE Healthcare, Sweden) at 20˚C as previously described. For competitive binding analysis, GI-plasminogen or GII-plasminogen was captured on flow cells 2, 3 and 4 until saturation of immobilised PAM occurred. Analytes were diluted in running buffer (PBS, 0.05% Tween-20, 50 µM EDTA, pH 7.4) or in running buffer containing varying concentrations (0-100 mM) of the lysine analogue εACA and/or benzamidine (Sigma-Aldrich, Australia). Kinetic assays were performed using varying concentrations of analyte (GI-plasminogen and GII-plasminogen, 0-200 nM; GII-angiostatin, 0-100 nM) over a series of five 120 s injections at a flow rate of 30 µL/min with a 600 s dissociation period. Regeneration of the flow cell surface was achieved as previously described. Affinity interactions were analysed as described above. Differences in binding affinity were determined by a one-way analysis of variance with Tukey's multiple comparison test.

2.10.2 Detecting conformational change in plasminogen glycoforms I and II

Plasminogen glycoforms were diluted to 1.5 µM in HEPES buffer (10 mM HEPES, 150 mM NaCl, pH 7.4). Methods were appended from Biggar *et al* (2012). Briefly, 5 µL of each plasminogen glycoform at a concentration of 1.5 µM was added to a thin walled 96-well unskirted polymerase chain reaction (PCR) microplate containing 2.5 µL of 40 × SYPRO® orange (Invitrogen, Carlsbad, CA, USA) and 12.5 µL of HEPES buffer with varying concentrations (0-100 mM) of εACA. Plates were read on a LightCycler® 480 II Real Time PCR system (Hoffman-La Roche, Basel, Switzerland) with excitation and emission set at 425 nm and 625 nm respectively. Fluorescent measurements were expressed as change in the initial fluorescence $(F_{abs}-F_o)/F_o=\Delta F/F_o$. As εACA was found to induce an immediate conformational change in plasminogen, intervals represent the end points of conformational change in relation to εACA concentration.

2.10.3 Cell surface plasminogen glycoform acquisition

To establish whether GAS preferentially captures GI-plasminogen or GII-plasminogen at the bacterial cell surface, PAM expressing GAS isolate NS13 was harvested at mid-log phase (OD_{600nm} 0.5), washed in PBS and incubated in citrated normal human plasma as a source of plasminogen for 1 h at 37 ˚C. Following two washes in PBS, bound plasminogen was eluted from the GAS cell surface using 100 mM glycine-HCl (pH 2.0) for 15 min at 25˚C. Eluted plasminogen was then probed using rabbit α-human plasminogen antibody (Calbiochem, USA), followed by goat α-rabbit IgG horseradish peroxidase conjugate (Invitrogen, USA). Enhanced chemiluminescence detection (Pierce, Biotechnology, USA) was used to visualise plasminogen protein bands. Purified GI-plasminogen and GII-plasminogen samples were used as positive controls to enable identification of bound plasminogen glycoforms from plasma by GAS based on electrophoretic mobility.

2.10.4 Cell surface plasminogen activation time-course

To assess the ability of GAS to accumulate plasmin at the bacterial cell surface in the presence of each glycoform variant, a time-course assay was performed. GAS strains NS13 and ALAB49 were harvested at mid-log phase (OD_{600nm} 0.5) and pre-incubated with 100 nM of either GI-plasminogen or GII-plasminogen at 37°C for either 5 min or 10 min. Following two washes in PBS, bacterial cells were resuspended in PBS containing 500 μM of chromogenic substrate S-2251 and incubated overnight. GAS cell surface plasmin activity was measured at A405 using a SpectraMax Plus 384 spectrophotometer (Molecular Devices, Sunnyvale, CA, USA). Differences in plasmin acquisition were determined by a one-way analysis of variance with Tukey's multiple comparison test.

2.10.5 Non-proteolytic active site generation in plasminogen

Non-proteolytic active site generation in GI-plasminogen and GII-plasminogen by type-2b SK was examined using the fluorescent active site titrant 4-methylumbelliferyl *p*-guanidinobenzoate (MUGB) (Marker Gene Technologies Inc., USA) in a POLARstar Omega fluorescence spectrophotometer (BMG LABTECH, Germany). plasminogen glycoforms (200 nM) were added to a black 96-well microtitre plate containing 1 μM MUGB in assay buffer (50 mM Tris-HCl, 100 mM NaCl, pH 7.4) and pre-incubated with PAM (200 nM) for 10 mins at 37°C. To initiate the reaction, SK was added to a final concentration of 400 nM in a total volume of 100 μL and the development of fluorescence was monitored continuously with excitation at 355 nm and emission at 460 nm. Data was normalised by subtracting a control reaction of both plasminogen glycoforms without the addition of SK and 1 μM MUGB. Fluorescence measurements were expressed as the fractional change in the initial fluorescence $(F_{abs}-F_o)/F_o=\Delta F/F_o$. Values for rates of active site generation were calculated using the initial linear portion of the reactions as $(\Delta F/F_o^{-1}).min^{-1}$. Initial rates of non-

proteolytic activation of GI-plasminogen and GII-plasminogen were compared using a two-tailed unpaired students t-test. Data sets were considered statistically significant at $p < 0.05$. All analysis including data transformation and linear regression was performed using GraphPad Prism version 5 (GraphPad Software, La Jolla, CA, USA).

2.10.6 GI-plasminogen and GII-plasminogen activation by type-2b SK variant plasminogen complexes

The ability of type-2b SK variant plasminogen complexes to activate substrate plasminogen was studied by the addition of type 2b-SK (final concentration 5 nM) to assay buffer (10 mM HEPES, 150 mM NaCl, 0.01% Tween-20, pH 7.4) containing an excess of GI-plasminogen or GII-plasminogen (500 nM) and chromogenic substrate S-2251 (500 μM) in a total volume of 100 μL. As type-2b SK does not possess the ability to activate plasminogen without the presence of PAM (Cook *et al.* 2013), recombinant PAM was also added to each experiment at 500 nM. The parabolic generation of plasmin was monitored by change in absorbance at A405 nm and measured over 30 min at 37°C as described above. For quantitative comparison of plasminogen variant effect, change in A405 nm, which is a function of S-2251 substrate cleavage by plasmin generated during the activation of plasminogen by SK, was plotted against t^2. The velocities of these reactions were then calculated from the gradient of A405 nm vs t^2. Initial rates of proteolytic activation of GI-plasminogen and GII-plasminogen were compared using a two-tailed unpaired students t-test. Data sets were considered statistically significant at $p < 0.05$. All analysis including data transformation and linear regression was performed using GraphPad Prism version 5 (GraphPad Software, La Jolla, CA, USA).

2.10.7 GI-plasminogen and GII-plasminogen activation at the GAS cell surface

Differences in GAS cell surface activation of GI-plasminogen and GII-plasminogen was carried out using type-2b expressing GAS strains ALAB49 and isogenic mutant ALAB49ΔPAM (Cook *et al.* 2013). GAS strains were cultured to mid-log phase (OD_{600nm} = 0.5) and collected by centrifugation (5000 × *g*, 10 min) and washed twice in sterile PBS. Following washing, 1×10^8 CFU of each GAS strain was resuspended in assay buffer (10 mM HEPES, 150 mM NaCl, 0.01 % Tween-20, pH 7.4) containing either GI-plasminogen or GII-plasminogen (500 nM). Protein mixtures were incubated for 10 min at 37˚C in a 96-well plate to allow binding of each plasminogen glycoform to the GAS cell surface. To initiate activation, SK_{ALAB49} and S-2251 were added to the wells at 5 nM and 500 µM respectively, in a final volume of 100 µl. The exponential generation of plasmin, monitored by change in absorbance at 405 nm was measured over 60 min at 37˚C using a SpectraMax Plus 384 spectrophotometer (Molecular Devices, USA). Initial rates of proteolytic activation of GI-plasminogen and GII-plasminogen at the cell surface were compared using a two-tailed unpaired students t-test. Data sets were considered statistically significant at $p < 0.05$. All analysis including data transformation and linear regression was performed using GraphPad Prism version 5 (GraphPad Software, La Jolla, CA, USA).

2.10.8 Ethics statement

Collection of blood was performed with the approval of the University of Wollongong Human Ethics Committee (HE08/250). Volunteers provided informed consent before donating blood samples.

2.11.0 Glycan interactions with *S. pyogenes*

2.11.1 Glycan microarray analysis – M protein

Glycan microarray analysis was undertaken at the Institute for Glycomics, Griffith University, Australia. Pre-printed glycan microarray slides were produced and generously donated by Dr. Lauren Hartley-Tassell and Prof. Michael Jennings (Institute for Glycomics, Griffith University, Australia). Glycan microarray techniques employed to identify novel M protein-glycan interactions were adapted from Day *et al* (2009). Purified recombinant M protein (0.5-1µg) was pre-incubated with mouse-α-histidine antibody (Abcam, Australia), rabbit-α-mouse Alexafluor488 antibody conjugate (Abcam, Australia) and goat α-rabbit Alexafluor488 conjugate, at a ratio of 4:2:1. For competitive binding analysis, M protein-antibody complexes were incubated with glycans of interest (60 pM-60 µM) purchased from Dextra Laboratories (UK). Protein mixtures were made up to a final volume of 75 µL in Dullbecos PBS (DPBS) and incubated at room temperature for 10 min to allow protein complex formation. Glycan microarray slides were blocked in DBPS with 1% (w/v) bovine serum albumin (Sigma-Aldrich, Australia) for 10 min at room temperature and dried by centrifugation (200 × *g*, 5 min). Following centrifugation, gene frames (Abgene, Australia) were bonded on to each glycan microarray slide, and 75 µL of pre-complexed protein mixture was applied. Protein mixtures were secured with the addition of a plastic cover, and samples were allowed to hybridise in the dark for 10 min at room temperature. Proceeding incubation, glycan microarray slides were washed in DPBS for 5 min with agitation, followed by a secondary wash in DPBS, 0.01% Tween-20 for 2 min with agitation. After washing, glycan microarray slides were dried by centrifugation (200 × *g*, 5 min). Fluorescence intensities of array spots were measured using the ProScanArray Microarray 4-Laser Scanner (Perkin Elmer, USA) using the Blue argon 488 nm excitation laser set to the FITC setting (494 nm

excitation and 518 nm emission), and gain set to 40-60 %. Image analysis was carried out using the ProScanArray imaging software ScanArray Express (Perkin Elmer, USA). Statistical analysis of the data was carried out using an independent sample T-test.

2.11.2 Glycan microarray analysis – *S. pyogenes*

Whole cell GAS glycan array analysis was undertaken using M1 serotype 5448 and isogenic mutant 5448ΔM1. GAS strains were harvested at $OD_{600\ nm}$ = 0.5 by centrifugation (5000 × *g*, 10 min) and 1×10^8 CFU were resuspended in DPBS containing 1 mM Cell Trace BODIPY TR methyl ester (Thermo-Fisher Scientific, Australia) to a final volume of 250 µL. Staining of bacteria was allowed to proceed in the dark at room temperature for 15 min. Following staining, bacterial cells were applied to a 5 µM Millex-SV Syringe filter unit (Merk-Millipore, Australia) to remove clumps of aggregated bacteria. Glycan microarray slides which were blocked and prepared as per above, were exposed to the stained filtered bacterial suspension for 10 min in the dark at room temperature. Glycan microarray slides were then washed, processed and analysed as previously described.

2.11.3 Surface plasmon resonance – glycan

Affinity interactions of M1 protein and M1-protein fragments with specific glycan subsets was undertaken via single cycle kinetics on a series S Ni-NTA chip (BIAcore AB) using a BIAcore T200 (GE Healthcare, Sweden) at 20˚C. Recombinant proteins were captured on flow cells 2, 3 and 4 until saturation occurred. Glycan analytes were diluted in PBS and kinetic assays were performed using varying concentrations of analyte (0-100 µM) over a series of five 120 s injections at a flow rate of 30 µL/min. Regeneration of the flow cell surface was achieved as previously described. M1/M1 fragment protein-glycan interactions were analysed via a steady-state affinity model using Biacore T200 evaluation software (Biacore AB).

It must be noted that increase in the RU of captured his-tagged lectin at the sensor surface increases the rate of dissociation of the lectin from the surface (Nieba *et al.* 1997). Due to unavoidable differences in lectin capture and dissociation across each of the three different flow cells relative to the negative control flow cell, binding affinity constants of each glycan binding interaction were shown to deviate. Furthermore, as the stoichiometry of each binding interaction was unknown, only steady state affinity binding analysis was able to be undertaken. Binding affinity constants were shown to have varying degrees of standard deviation, thus the binding data presented in this study provides a qualitative representation of the binding interaction between M1 protein/ M1 protein fragments and each respective glycan.

2.11.4 Preparation of human Detroit-562 and HaCaT cell lines

Human pharyngeal (Detroit-562) and keratinocyte (HaCaT) cells were cultured in DMEM-F12 (Invitrogen, Australia), supplements with 2 mM L-glutamine and 10 % v/v heat inactivated foetal bovine serum (FBS) (Bovogen Biologicals, Australia). For association assays with GAS, adherent eukaryotic cells were washed once with sterile PBS and treated with 200 µL of 0.25% Trypsin, 0.02% EDTA (Sigma-Aldrich, Australia). Cells were pelleted via centrifugation (500 × *g*, 5 min) and resuspended in pre-warmed DMEM-F12 (Invitrogen, Australia), supplements with 2 mM L-glutamine and 10 % v/v FBS. Following washing, 24-well plates were seeded with 1×10^5 cells and were cultured 37˚C with 5% CO_2 until 80-85% confluency (approximately 2×10^5 cells).

2.11.5 Pharyngeal cell association assays

Assays measuring total association of GAS (5448 eGFP) to human pharyngeal (Detroit-562) and keratinocyte (HaCaT) cells was performed as previously described (Ryan *et al.* 2001; Hollands *et al.* 2010). To determine M1 protein specificity, GAS strains 5448ΔM1 and

NS730 were used as negative controls. Mid-logarithmic GAS (2×10^8 CFU) were pre-incubated with either lacto-*N*-tetraose (Dextra Laboratories, UK) or lacto-*N*-fucopentose type I (Dextra Laboratories, UK) in PBS made up to a volume of 200 µl for 30 min at room temperature. Following incubation, GAS were pelleted by centrifugation (5000 × *g*, 10 min), washed in sterile PBS and resuspended in 2 mL of DMEM F12 media (Invitrogen, Australia) supplemented with 2 mM L-glutamine. Bacteria (2×10^6 CFU) were then incubated with respective eukaryotic cells (2×10^5) in a final volume of 200 µL for 2 h at 37˚C with 5% CO_2. After the subsequent incubation, GAS were washed 6 times in sterile PBS, and adherent eukaryotic cells were released upon the addition of 200 µL of 0.25% Trypsin, 0.02% EDTA (Sigma-Aldrich, Australia). To enable lysing of eukaryotic cells, 800 µL of dH_2O containing 0.025% Triton-X 100 (Sigma-Aldrich, Australia) was added to each sample. Released bacteria were serially diluted on THYA for enumeration. Bacterial association was calculated as a percentage of the original inoculum. Statistical significance was determined using a one-way ANOVA with the Tukey post-hoc test.

2.11.6 Collection of saliva and human buccal epithelial cells

Human buccal epithelial cells and human saliva were collected from 20 healthy individual donors of unknown blood type and secretor status. Buccal epithelial cells were collected via exfoliation of the buccal mucosa using flocked swabs (Micro Rheologics, Italy). Harvested cells were resuspended in sterile PBS, and sorted using a 70 µM cell strainer (Pacific Laboratories, Australia). Buccal cells were then pelleted via centrifugation (500 × *g*, 5 min) and resuspended in ice cold PBS. Along with buccal cells, 2 mL of human unstimulated saliva was also collected. After collection, saliva was immediately frozen down and stored at – 80˚C.

2.11.7 *N*- and *O*-linked glycan release for mass spectrometry

Release of *N*- and *O*-linked glycans was undertaken on salivary membrane glycoproteins according to Wilson *et al* (2002). Proteins were immobilised by dot-blotting on to a methanol activated PVDF membrane (Millipore, Australia). Once dry, PVDF membrane sections containing immobilised protein were excised and transferred to 1.5 mL Eppendorf tubes. *N*-linked glycans were released from immobilised proteins by incubation with 3 U of PNGase F (*Flavobacterium meningosepticum*, Roche Diagnostics, Australia) overnight at 37°C. Released *N*-linked glycans were further treated with 15 mM NH_4COOH, for 60 min at room temperature, and subsequently dried in a vacuum centrifuge (Eppendorf, Germany). Samples were then reduced with 20 µL of 1 M $NaBH_4$ in 50 mM KOH for 3 h at 50°C. Reduction reactions were quenched with 1 µL glacial acetic acid. *O*-linked glycans were released from the same excised PVDF membrane spots via β-elimination via overnight incubation with 20 µL of 0.5 M $NaBH_4$ in 50 mM KOH at 50°C. Reduction was quenched with 1 µL glacial acetic acid. Both released *N*- and *O*-linked glycans were desalted using custom made exchange columns containing 30 µL AG50W-X8 cation-exchange resin (BioRad, Australia) packed in to µC18 ZipTips. Glycan samples were washed three times with 200 µL methanol to remove residual borate and dried via vacuum centrifugation. Purified glycans were resuspended in 10 µL dH_2O and subjected to porous graphitised carbon liquid chromatography electrospray ionisation mass spectrometry (PGC-LC-ESI MS/MS).

2.11.8 Human buccal epithelial association assays

To measure the association of GAS to HBE cells, 2×10^5 HBE cells were incubated with 2×10^6 eGFP expressing GAS (5448 and 5448ΔM1) in 250 µL DPBS for 2 h at 37°C with 5% CO_2. Post-incubation, samples were transferred to FACS tubes, resuspended in 750 µL DPBS, pelleted (500 × *g*, 10 min) and resuspended in a further 500 µL DPBS. Association of

GAS to HBE cells was measured using a LSR II (BD Bioscience, USA) flow cytometer. Gated HBE cells were analysed with a 575/26 nm filter with data being recorded for 10000 events. Fluorescent intensity of uninfected HBE cells was also measured to allow for compensation of auto-fluorescence. The relative quantity of GAS associated with HBE cells was estimated by the mean fluorescence intensity (MFI) of eGFP positive HBE cells. Data was analysed and processed using Flowjo (Treestar, USA).

2.11.10 Mass spectrometry

ESI-MS of *N*- and *O*-glycans was carried out as previously described (Everest-Dass *et al.* 2012). Released *N*- and *O*-glycans were separated by PGC (5 µm Hypercarb, 180 µm ID × 100 mm, Thermo Scientific, USA). Separation of *N*-glycans was performed over an 85 min gradient of 0-45% (w/v) CH_3CN in 10 mM NH_4HCO_3. The flow rate set for *N*- and *O*-glycans was set to 3 µL/min, using a Agilent 1100 HPLC system (Agilent Technologies, USA) which was connected directly to an Aglient 6330 ESI source. Capillary voltage was set to 3 kV, and the dry gas maintained at 300°C. MS spectra was obtained in the negative-ion mode with scan range between *m/z* 200 and *m/z* 2200. MS data analysis was carried out in ESI-Compass 1.3 (Bruker Daltonics, Germany).

2.11.11 Ethics for human tissue collection

Collection of HBE cells and saliva was performed with the approval of the University of Wollongong Human Ethics Committee (HE08/250). Volunteers provided informed consent before donating tissue samples.

3. Phylogenetic and functional characterisation of M protein from invasive GAS isolates

A section of this work has been published in The Journal of Infectious Diseases.

Reference:

Martina L. Sanderson-Smith, David M. P. De Oliveira, Julien Guglielmini, David D. J McMillan, Therese Vu, Jessica K. Holien, Anna Henningham, Andrew C. Steer, Debra E. Bessen, James B. Dale, Nigel Curtis, Bernard W Beall, Mark J. Walker, Michael W. Parker, Jonathan R. Carapetis, Laurence Van Melderen, Sri K. Sriprakash, Pierre R. Smeesters; M protein study group. (2014). A systematic and functional classification of Streptococcus pyogenes that serves as a new tool for molecular typing and vaccine development. *The Journal of Infectious Diseases*. **210**: 1325-38.

3.1 Introduction

S. pyogenes is a major cause of global morbidity and mortality, accounting for approximately 19.9 million cases of severe infection and over 500 000 deaths per year (WHO 2005). The global burden of GAS, exemplified by current rates of rheumatic fever and rheumatic heart disease, has prompted the urgent development of a global vaccine (Steer *et al.* 2013). The GAS surface protein, the M protein, has been the most pursued GAS vaccine target to date. Despite numerous initiatives, a global GAS vaccine still remains commercially unavailable, principally due to the antigenic diversity displayed between M protein serotypes.

Antigenic variation between M proteins forms the basis of nucleotide based M-typing, a classification system utilising 10-15% of the mature protein. The representation of GAS isolates in developed settings is restricted to several predominant M-types, most notably M-types 1, 3, 12 and 28 which are frequently isolated from asymptomatic carriers (Ekelund *et al.* 2005). However, to date more than 220 M-types have been reported, with the majority circulating in low-income settings (McMillan *et al.* 2013). Within areas such as the NT of Australia, the diversity of M-types is significantly greater. Selective pressures influencing

these genetic disparities remain unclear. Potential factors such as climate, socio-economic status, population immunity and genetic susceptibilities have been proposed to play a role in explaining the high number of circulating M-types (Smeesters *et al.* 2008).

Few studies provide a detailed characterisation of M-types in poor-income regions, although preliminary analysis of M proteins from a subset of 51 Brazilian and Belgian GAS serotypes suggests that the majority of circulating M-types in low-socioeconomic countries and territories share a high level of sequence identity (Smeesters *et al.* 2006). Defined on the basis of whole surface-exposed M protein sequence, prior work analysing the distribution and diversity of Belgian and Brazilian isolates discovered that M-types could be categorised into 9 and 13 clusters respectively (Smeesters *et al.* 2008). These findings have questioned the concept of type-specific immunity and whether highly homologous M proteins are susceptible to cross-reactive immunity. A framework which enables the development of an immunologic cross-protective vaccine may also provide insight into the function of diverse M proteins in specific disease states.

A major function of the M protein is to initiate interactions between GAS and the host. As such, M proteins interact with a diverse range of host proteins. A small but diverse range of these interactions have been characterised, particularly binding of select host proteins which serve to mediate bacterial virulence and evade host immunity (Walker *et al.* 2014). A novel M protein based classification system based on full length amino acid sequence, encompassing both sequence diversity and function may help to explain anomalies observed in M-type cross-protection, while further enhancing our understanding of antibody mediated recognition and bacterial opsonisation. By using the sequences of 175 M-types from 1086 invasive GAS isolates, a phylogenetic classification system was developed. To evaluate the utility of this novel functional classification system, 26 M proteins from 24 distinct M-types, representative of each of the major *emm*-cluster groups were systematically characterised.

3.2 Results

3.2.1 DNA sequence analysis

PCR amplification of *emm* genes from 13 of the 26 GAS isolates used in this study was undertaken using serotype specific primers. The remaining *emm* gene sequences were kindly provided by Dr. Pierre Smeesters. PCR amplification of the 13 *emm* genes resulted in the production of amplicons ranging from 0.8-1.5 kb (Fig 3.1).

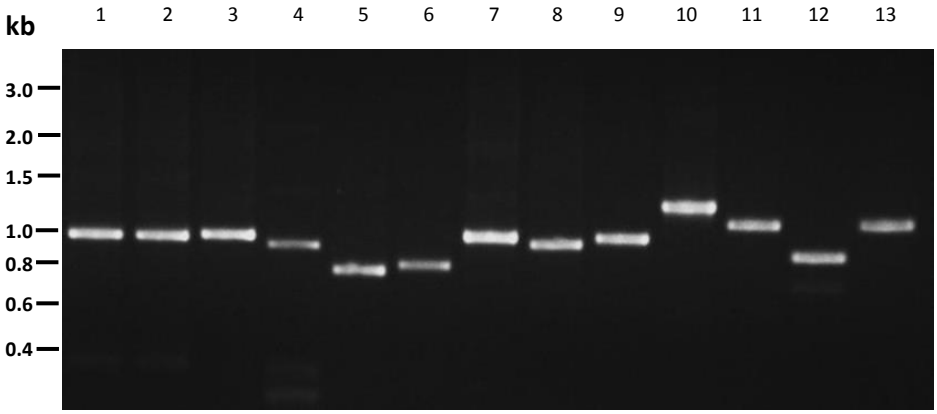


Figure 3.1: PCR amplification of *emm* genes from 13 GAS strains. Lane 1, NS226 *emm*4; lane 2, NS730 *emm*90; lane 3, NS192, *emm*106; lane 4, NS179, *emm*9; lane 5, PRS66, *emm*102; lane 6, PRS2, *emm*2; lane 7, NS8, *emm*85; lane 8, NS414, *emm*11; lane 9, NS931, *emm*69; lane 10, NS1140, *emm*57; lane 11, NS178, *emm*54; lane 12, NS501, *emm*14; lane 13, NS80, *emm*70. PCR amplicons were electrophoresed on a 1% agarose gel. Molecular weight markers are shown on the left in kilo base pairs.

In order to determine the extent of homology between M proteins, DNA sequence analysis was undertaken on *emm* genes from 26 GAS isolates and aligned via a Pairwise MUSCLE alignment using default settings. All *emm* sequences were trimmed from the first codon of the translated mature protein to the last codon expressing the first amino acid of the D-repeat domain. At the nucleotide level *emm* genes were found to be highly variable and demonstrated low sequence identity at the 5' regions which code A and B repeat domains. Alignment of N-terminal *emm* sequences starting from the first nucleotide codon of the mature protein up to the first nucleotide codon corresponding to the start of the C-repeat domain identified that *emm* genes had an average sequence identity of 29.7 % (Fig 3.2A). Unlike the N-terminal region, the 3' end which codes the C-repeat domains was found to be highly conserved among *emm* genes (Fig 3.2B). MUSCLE alignment starting from the first codon on the C-repeat domain up to the first codon of the D-repeat domain showed an 80.0% average in sequence identity among *emm* genes at the C-terminal region.

3.2.2 M protein sequence analysis

At the amino acid level, translated *emm* sequences from 26 GAS isolates were aligned using MUSCLE pairwise alignment as previously mentioned (Appendix E). Sequence identity between full length M proteins was found to be approximately 44%. N-terminal regions that encode A and B repeat domains and form the basis of M-serotyping were found to be highly variable. With the exception of M9, M9.1, M11 and M85 which displayed highly conserved sequences, sequence identities were found to range from 6% to 55% (Fig 3.3A). Sequence identity between M proteins at the C-terminal end which encodes C and D repeat domains was considerably more conserved with amino acid sequence identity ranging from 44% to 100%.

Using a Jukes-Cantor distance model, phylogenetic analysis identified that the 26 full length M proteins could be categorised into two major clades; clade X and clade Y. Clade X could be further sub-divided into 6 cluster groups whereas clade Y could be separated in to 3 minor cluster groups with the exception of M57, which was deemed an outlier (Fig 3.4). Within clade X, cluster groups E1-4 was exclusively comprised of M-serotypes originating from *emm*-pattern E GAS isolates. Cluster E5 consists of M proteins belonging to *emm*-pattern D and E whereas E6 comprised solely of A-C pattern isolates. Within clade Y, cluster D1 consisted of D-pattern isolates whereas clusters AC-1, AC-2 and outlier M57 were all made up of sequences originating from A-C pattern isolates.

The subset of 26 M proteins representing 24 M-types were included in a phylogenetic and functional classification study analysing 174 M-types derived from 1086 GAS isolates (Sanderson-Smith *et al.* 2014). Phylogentic analysis of all 174 M-types identified that M proteins could be arranged into two major clades, clade X and clade Y. Clade X was composed of 5 main M-pattern E clusters (E1 to E4 and E6), whereas clade Y was divided into two sub-clusters (Y1 and Y2) which contained 10 main M protein clusters (Y1: D1 to D5; Y2: A-C1 to A-C5). Major cluster groups within clade X and clade Y accounted for 145 of the 175 M-types in study (Fig 3.5).

A

	emm97	emm57	emm14	emm19	emm54	emm1	emm70	emm53	emm83	emm98	emm65	emm11	emm85	emm48	emm4.2	emm60	emm3	emm12	emm58	emm9	emm9.1	emm2	emm102	emm90	emm106
emm97		27.7%	23.0%	28.5%	23.9%	22.8%	26.6%	27.5%	27.5%	26.6%	20.0%	21.9%	22.4%	21.0%	22.2%	24.7%	32.1%	29.8%	26.3%	22.6%	22.8%	29.6%	27.0%	27.8%	26.3%
emm57	27.7%		23.0%	31.1%	25.5%	27.0%	21.0%	25.8%	24.1%	22.6%	20.0%	20.8%	21.4%	22.1%	21.8%	19.5%	30.4%	29.7%	27.1%	22.7%	22.7%	25.5%	24.4%	26.2%	25.7%
emm14	23.0%	23.0%		30.8%	30.8%	23.6%	24.3%	24.6%	25.3%	23.6%	20.4%	20.6%	20.8%	20.0%	21.8%	22.4%	26.7%	26.0%	26.2%	21.9%	22.0%	23.8%	23.2%	23.6%	27.4%
emm19	28.5%	31.1%	30.8%		36.7%	28.8%	23.3%	22.5%	27.2%	20.2%	18.6%	19.6%	20.5%	20.0%	23.4%	22.4%	30.5%	29.4%	25.0%	22.8%	22.8%	22.8%	23.5%	25.5%	27.3%
emm54	23.9%	25.5%	30.8%	36.7%		29.9%	34.0%	28.3%	23.4%	26.9%	27.1%	25.5%	26.7%	26.4%	32.8%	32.8%	26.7%	24.4%	36.5%	31.4%	31.1%	32.1%	28.7%	32.1%	36.4%
emm1	22.8%	27.0%	23.6%	28.8%	29.9%		26.2%	28.8%	28.5%	25.9%	21.9%	22.4%	23.5%	25.2%	28.6%	26.6%	24.9%	25.3%	31.4%	30.4%	30.4%	27.5%	25.6%	27.7%	28.2%
emm70	26.6%	21.0%	24.3%	23.3%	34.0%	26.2%		55.6%	39.6%	59.1%	31.0%	33.0%	33.4%	32.7%	28.3%	32.0%	27.9%	25.2%	35.7%	28.9%	28.9%	27.3%	25.7%	30.4%	32.3%
emm53	27.5%	25.8%	24.6%	22.5%	28.3%	28.8%	55.6%		56.0%	59.0%	31.4%	30.9%	31.8%	30.1%	31.9%	25.7%	26.4%	26.0%	29.0%	30.0%	30.0%	28.6%	26.7%	33.0%	31.9%
emm83	27.5%	24.1%	25.3%	27.2%	23.4%	28.5%	39.6%	56.0%		50.6%	25.7%	25.5%	25.8%	23.7%	26.1%	23.4%	24.7%	25.5%	24.8%	25.6%	25.8%	25.6%	24.3%	27.7%	27.7%
emm98	26.6%	22.6%	23.6%	20.2%	26.9%	25.9%	59.1%	59.0%	50.6%		34.0%	35.6%	35.7%	34.2%	28.5%	26.4%	26.5%	26.4%	28.7%	31.0%	31.0%	37.2%	35.7%	31.3%	30.9%
emm65	20.0%	20.0%	20.4%	18.6%	27.1%	21.9%	31.0%	31.4%	25.7%	34.0%		68.5%	71.7%	45.4%	33.9%	25.1%	19.1%	19.7%	31.8%	39.0%	39.2%	31.5%	29.9%	34.2%	35.9%
emm11	21.9%	20.8%	20.6%	19.6%	25.5%	22.4%	33.0%	30.9%	25.5%	35.6%	68.5%		87.8%	44.0%	34.6%	28.8%	20.2%	20.7%	30.8%	34.6%	34.9%	32.1%	29.4%	32.8%	34.5%
emm85	22.4%	21.4%	20.8%	20.5%	26.7%	23.5%	33.4%	31.8%	25.8%	35.7%	71.7%	87.8%		48.9%	33.9%	28.9%	20.5%	21.1%	32.3%	36.0%	36.0%	34.1%	30.3%	34.2%	36.0%
emm48	21.0%	22.1%	20.0%	20.0%	26.4%	25.2%	32.7%	30.1%	23.7%	34.2%	45.4%	44.0%	48.9%		30.1%	25.4%	21.3%	22.1%	31.6%	31.9%	31.9%	31.7%	28.9%	32.1%	33.7%
emm4.2	22.2%	21.8%	21.8%	23.4%	32.8%	28.6%	28.3%	31.9%	26.1%	28.5%	33.9%	34.6%	33.9%	30.1%		65.2%	25.4%	25.7%	38.8%	46.6%	46.6%	37.0%	35.1%	39.3%	36.0%
emm60	24.7%	19.5%	22.4%	22.4%	32.8%	26.6%	32.0%	25.7%	23.4%	26.4%	25.1%	28.8%	28.9%	25.4%	65.2%		27.7%	28.0%	40.7%	37.0%	37.0%	34.2%	33.1%	33.3%	35.1%
emm3	32.1%	30.4%	26.7%	30.5%	26.7%	24.9%	27.9%	26.4%	24.7%	26.5%	19.1%	20.2%	20.5%	21.3%	25.4%	27.7%		57.4%	30.9%	25.8%	25.8%	26.9%	26.2%	29.6%	29.3%
emm12	29.8%	29.7%	26.0%	29.4%	24.4%	25.3%	25.2%	26.0%	25.5%	26.4%	19.7%	20.7%	21.1%	22.1%	25.7%	28.0%	57.4%		28.3%	25.3%	25.3%	25.6%	25.9%	30.1%	29.6%
emm58	26.3%	27.1%	26.2%	25.0%	36.5%	31.4%	35.7%	29.0%	24.8%	28.7%	31.8%	30.8%	32.3%	31.6%	38.8%	40.7%	30.9%	28.3%		45.0%	45.0%	46.3%	43.1%	49.0%	51.7%
emm9	22.6%	22.7%	21.9%	22.8%	31.4%	30.4%	28.9%	30.0%	25.6%	31.0%	39.0%	34.6%	36.0%	31.9%	46.6%	37.0%	25.8%	25.3%	45.0%		99.7%	51.1%	50.8%	49.2%	47.6%
emm9.1	22.8%	22.7%	22.0%	22.8%	31.1%	30.4%	28.9%	30.0%	25.8%	31.0%	39.2%	34.9%	36.0%	31.9%	46.6%	37.0%	25.8%	25.3%	45.0%	99.7%		51.1%	50.8%	49.0%	47.4%
emm2	29.6%	25.5%	23.8%	22.8%	32.1%	27.5%	27.3%	28.6%	25.6%	37.2%	31.5%	32.1%	34.1%	31.7%	37.0%	34.2%	26.9%	25.6%	46.3%	51.1%	51.1%		74.1%	50.8%	48.3%
emm102	27.0%	24.4%	23.2%	23.5%	28.7%	25.6%	25.7%	26.7%	24.3%	35.7%	29.9%	29.4%	30.3%	28.9%	35.1%	33.1%	26.2%	25.9%	43.1%	50.8%	50.8%	74.1%		47.9%	46.4%
emm90	27.8%	26.2%	23.6%	25.5%	32.1%	27.7%	30.4%	33.0%	27.7%	31.3%	34.2%	32.8%	34.2%	32.1%	39.3%	33.3%	29.6%	30.1%	49.0%	49.2%	49.0%	50.8%	47.9%		58.4%
emm106	26.3%	25.7%	27.4%	27.3%	36.4%	28.2%	32.3%	31.9%	27.7%	30.9%	35.9%	34.5%	36.0%	33.7%	36.0%	35.1%	29.3%	29.6%	51.7%	47.6%	47.4%	48.3%	46.4%	58.4%	

B

	emm70	emm4.2	emm60	emm106	emm58	emm9	emm9.1	emm2	emm90	emm102	emm85	emm65	emm11	emm48	emm57	emm14	emm3	emm12	emm1	emm83	emm97	emm53	emm98	emm19	emm54
emm70		87.9%	88.2%	85.5%	84.0%	84.6%	84.6%	61.6%	85.5%	85.5%	88.4%	88.2%	89.9%	89.9%	69.9%	72.1%	73.2%	73.2%	86.8%	87.1%	86.4%	88.6%	88.4%	76.8%	76.8%
emm4.2	87.9%		99.8%	91.9%	90.4%	91.4%	91.4%	65.1%	91.9%	91.4%	95.0%	94.7%	95.6%	95.6%	65.8%	71.3%	71.7%	72.1%	75.9%	76.5%	76.3%	77.2%	77.4%	73.2%	73.5%
emm60	88.2%	99.8%		92.1%	90.6%	91.7%	91.7%	65.1%	92.1%	91.7%	95.2%	95.0%	95.8%	95.8%	65.8%	71.3%	71.7%	72.1%	76.1%	76.8%	76.5%	77.4%	77.6%	73.2%	73.5%
emm106	85.5%	91.9%	92.1%		97.9%	97.2%	97.2%	69.7%	98.2%	97.7%	91.2%	90.8%	91.2%	91.2%	63.0%	73.8%	74.3%	75.1%	74.6%	73.7%	73.5%	74.3%	74.6%	77.2%	76.6%
emm58	84.0%	90.4%	90.6%	97.9%		98.4%	98.4%	69.7%	97.9%	98.4%	89.7%	89.5%	89.7%	89.7%	61.9%	73.8%	74.0%	74.6%	73.5%	72.1%	73.2%	72.8%	73.2%	75.6%	74.9%
emm9	84.6%	91.4%	91.7%	97.2%	98.4%		100%	69.7%	98.2%	98.6%	89.9%	90.1%	89.9%	89.9%	63.0%	73.3%	74.7%	75.3%	73.7%	73.2%	74.4%	73.9%	74.3%	75.9%	75.6%
emm9.1	84.6%	91.4%	91.7%	97.2%	98.4%	100%		69.7%	98.2%	98.6%	89.9%	90.1%	89.9%	89.9%	63.0%	73.3%	74.7%	75.3%	73.7%	73.2%	74.4%	73.9%	74.3%	75.9%	75.6%
emm2	61.6%	65.1%	65.1%	69.7%	69.7%	69.7%	69.7%		70.3%	69.9%	64.5%	64.9%	64.0%	64.0%	44.7%	58.0%	54.3%	53.3%	53.5%	53.3%	51.9%	54.2%	53.9%	54.7%	54.7%
emm90	85.5%	91.9%	92.1%	98.2%	97.9%	98.2%	98.2%	70.3%		99.1%	91.2%	91.0%	91.2%	91.2%	63.7%	74.0%	74.5%	75.8%	75.0%	74.1%	74.6%	74.8%	75.2%	77.2%	77.0%
emm102	85.5%	91.4%	91.7%	97.7%	98.4%	98.6%	98.6%	69.9%	99.1%		91.2%	91.0%	91.2%	91.2%	63.2%	74.3%	74.5%	75.6%	75.0%	73.7%	74.9%	74.3%	74.8%	76.8%	76.6%
emm85	88.4%	95.0%	95.2%	91.2%	89.7%	89.9%	89.9%	64.5%	91.2%	91.2%		97.4%	96.7%	96.7%	66.9%	70.2%	70.8%	71.1%	77.6%	76.1%	75.8%	77.2%	77.4%	74.1%	73.5%
emm65	88.2%	94.7%	95.0%	90.8%	89.5%	90.1%	90.1%	64.9%	91.0%	91.0%	97.4%		96.7%	96.7%	66.7%	70.6%	71.7%	71.8%	77.9%	76.8%	76.8%	77.9%	77.6%	73.7%	73.2%
emm11	89.9%	95.6%	95.8%	91.2%	89.7%	89.9%	89.9%	64.0%	91.2%	91.2%	96.7%	96.7%		100%	66.9%	71.1%	72.1%	72.5%	78.7%	78.1%	76.5%	79.2%	78.9%	74.8%	74.3%
emm48	89.9%	95.6%	95.8%	91.2%	89.7%	89.9%	89.9%	64.0%	91.2%	91.2%	96.7%	96.7%	100%		66.9%	71.1%	72.1%	72.5%	78.7%	78.1%	76.5%	79.2%	78.9%	74.8%	74.3%
emm57	69.9%	65.8%	65.8%	63.0%	61.9%	63.0%	63.0%	44.7%	63.7%	63.2%	66.9%	66.7%	66.9%	66.9%		67.1%	73.3%	76.8%	79.1%	79.6%	80.9%	80.4%	79.8%	75.7%	76.3%
emm14	72.1%	71.3%	71.3%	73.8%	73.8%	73.3%	73.3%	58.0%	74.0%	74.3%	70.2%	70.6%	71.1%	71.1%	67.1%		77.5%	79.0%	77.4%	76.3%	77.5%	77.0%	77.0%	84.1%	85.5%
emm3	73.2%	71.7%	71.7%	74.3%	74.0%	74.7%	74.7%	54.3%	74.5%	74.5%	70.8%	71.7%	72.1%	72.1%	73.3%	77.5%		96.5%	82.9%	83.1%	83.1%	83.3%	83.6%	88.0%	86.7%
emm12	73.2%	72.1%	72.1%	75.1%	74.6%	75.3%	75.3%	53.3%	75.8%	75.6%	71.1%	71.8%	72.5%	72.5%	76.8%	79.0%	96.5%		83.3%	84.3%	83.3%	84.3%	84.0%	90.4%	89.6%
emm1	86.8%	75.9%	76.1%	74.6%	73.5%	73.7%	73.7%	53.5%	75.0%	75.0%	77.6%	77.9%	78.7%	78.7%	79.1%	77.4%	82.9%	83.3%		96.1%	96.0%	96.9%	96.1%	86.6%	87.1%
emm83	87.1%	76.5%	76.8%	73.7%	72.1%	73.2%	73.2%	53.3%	74.1%	73.7%	76.1%	76.8%	78.1%	78.1%	79.6%	76.3%	83.1%	84.3%	96.1%		96.7%	97.4%	97.4%	86.0%	86.4%
emm97	86.4%	76.3%	76.5%	73.5%	73.2%	74.4%	74.4%	51.9%	74.6%	74.9%	75.8%	76.8%	76.5%	76.5%	80.9%	77.5%	83.1%	83.3%	96.0%	96.7%		96.5%	97.4%	86.2%	85.9%
emm53	88.6%	77.2%	77.4%	74.3%	72.8%	73.9%	73.9%	54.2%	74.8%	74.3%	77.2%	77.9%	79.2%	79.2%	80.4%	77.0%	83.3%	84.3%	96.9%	97.4%	96.5%		98.2%	86.2%	86.6%
emm98	88.4%	77.4%	77.6%	74.6%	73.2%	74.3%	74.3%	53.9%	75.2%	74.8%	77.4%	77.6%	78.9%	78.9%	79.8%	77.0%	83.6%	84.0%	96.1%	97.4%	97.4%	98.2%		86.2%	85.7%
emm19	76.8%	73.2%	73.2%	77.2%	75.6%	75.9%	75.9%	54.7%	77.2%	76.8%	74.1%	73.7%	74.8%	74.8%	75.7%	84.1%	88.0%	90.4%	86.6%	86.0%	86.2%	86.2%	86.2%		95.9%
emm54	76.8%	73.5%	73.5%	76.6%	74.9%	75.6%	75.6%	54.7%	77.0%	76.6%	73.5%	73.2%	74.3%	74.3%	76.3%	85.5%	86.7%	89.6%	87.1%	86.4%	85.9%	86.6%	85.7%	95.9%	

Figure 3.2: Pairwise identity analysis of *emm* N-terminal (A) and C-terminal (B) gene sequences. Sequence alignments were carried out via the pairwise MUSCLE alignment tool using default settings on Geneious version 6.0, Biomatters. Sequence identity is presented as a heat map, whereby darker shades of grey correspond to an increase in percentage of identical bases between distinct *emm* gene sequences.

A

	M57	M48	M65	M11	M85	M1	M4.2	M60	M9.1	M9	M58	M2	M102	M90	M106	M3	M12	M14	M19	M54	M97	M53	M70	M83	M98
M57		7.0%	11.1%	9.5%	11.1%	10.1%	7.7%	8.2%	11.4%	11.4%	7.3%	9.4%	9.4%	9.3%	7.0%	5.9%	6.0%	10.5%	7.3%	8.5%	11.3%	8.7%	9.0%	8.1%	9.9%
M48	7.0%		33.6%	31.8%	31.8%	8.6%	14.3%	12.3%	9.6%	8.7%	10.5%	11.0%	8.5%	10.2%	10.4%	9.3%	8.8%	9.2%	15.9%	13.3%	7.8%	10.6%	14.0%	8.0%	12.8%
M65	11.1%	33.6%		54.7%	54.7%	10.4%	15.7%	15.2%	17.0%	17.0%	13.2%	12.3%	12.3%	10.2%	12.6%	9.0%	5.5%	9.2%	13.5%	13.6%	10.7%	12.6%	15.8%	10.5%	16.1%
M11	9.5%	31.8%	54.7%		85.8%	11.7%	17.6%	19.1%	16.1%	16.1%	13.7%	11.1%	7.7%	14.0%	14.0%	8.4%	6.4%	11.0%	14.7%	12.4%	11.2%	13.8%	16.3%	7.9%	11.2%
M85	11.1%	31.8%	54.7%	85.8%		12.8%	21.3%	17.9%	17.0%	17.0%	11.4%	10.5%	8.8%	12.7%	15.3%	8.0%	5.5%	10.5%	15.1%	12.7%	12.2%	15.0%	19.2%	9.3%	13.3%
M1	10.1%	8.6%	10.4%	11.7%	12.8%		13.5%	11.6%	16.7%	16.7%	21.6%	15.4%	16.0%	16.9%	16.8%	11.0%	10.2%	16.3%	15.2%	14.4%	19.2%	18.6%	15.1%	14.1%	20.1%
M4.2	7.7%	14.3%	15.7%	17.6%	21.3%	13.5%		60.0%	23.9%	23.9%	24.1%	23.0%	20.4%	22.9%	27.7%	11.1%	11.5%	10.7%	15.9%	17.0%	11.5%	11.7%	12.5%	10.8%	13.9%
M60	8.2%	12.3%	15.2%	19.1%	17.9%	11.6%	60.0%		26.5%	26.5%	20.1%	19.2%	20.0%	20.7%	19.4%	7.8%	9.8%	9.1%	9.7%	14.6%	10.2%	9.2%	10.9%	9.0%	10.9%
M9.1	11.4%	9.6%	17.0%	16.1%	17.0%	16.7%	23.9%	26.5%		99.1%	41.7%	34.8%	33.9%	38.7%	36.3%	10.7%	7.1%	17.2%	16.8%	22.0%	15.2%	13.9%	11.5%	7.7%	14.7%
M9	11.4%	8.7%	17.0%	16.1%	17.0%	16.7%	23.9%	26.5%	99.1%		41.7%	34.8%	33.9%	38.7%	36.3%	10.7%	7.1%	17.2%	16.8%	22.0%	15.2%	13.9%	11.5%	7.7%	14.7%
M58	7.3%	10.5%	13.2%	13.7%	11.4%	21.6%	24.1%	20.1%	41.7%	41.7%		33.1%	31.5%	40.0%	39.7%	9.0%	8.9%	14.4%	14.1%	17.8%	10.1%	12.9%	10.4%	9.8%	14.2%
M2	9.4%	11.0%	12.3%	11.1%	10.5%	15.4%	23.0%	19.2%	34.8%	34.8%	33.1%		61.3%	36.5%	44.8%	9.7%	6.2%	10.4%	11.1%	14.3%	15.1%	14.8%	13.6%	12.5%	14.4%
M102	9.4%	8.5%	12.3%	7.7%	8.8%	16.0%	20.4%	20.0%	33.9%	33.9%	31.5%	61.3%		35.7%	41.6%	6.0%	6.6%	9.9%	9.0%	15.0%	13.2%	14.1%	14.4%	10.5%	14.4%
M90	9.3%	10.2%	10.2%	14.0%	12.7%	16.9%	22.9%	20.7%	38.7%	38.7%	40.0%	36.5%	35.7%		49.6%	8.2%	9.7%	15.8%	14.8%	16.9%	9.4%	16.7%	19.5%	9.3%	14.6%
M106	7.0%	10.4%	12.6%	14.0%	15.3%	16.8%	27.7%	19.4%	36.3%	36.3%	39.7%	44.8%	41.6%	49.6%		8.9%	7.6%	10.8%	10.4%	16.3%	11.7%	17.6%	15.9%	12.0%	15.4%
M3	5.9%	9.3%	9.0%	8.4%	8.0%	11.0%	11.1%	7.8%	10.7%	10.7%	9.0%	9.7%	6.0%	8.2%	8.9%		39.9%	9.6%	14.3%	11.1%	10.2%	12.6%	8.7%	17.7%	10.6%
M12	6.0%	8.8%	5.5%	6.4%	5.5%	10.2%	11.5%	9.8%	7.1%	7.1%	8.9%	6.2%	6.6%	9.7%	7.6%	39.9%		10.9%	14.6%	13.6%	8.8%	9.2%	7.7%	12.9%	8.4%
M14	10.5%	9.2%	9.2%	11.0%	10.5%	16.3%	10.7%	9.1%	17.2%	17.2%	14.4%	10.4%	9.9%	15.8%	10.8%	9.6%	10.9%		22.4%	15.5%	10.8%	12.3%	12.6%	11.0%	13.6%
M19	7.3%	15.9%	13.5%	14.7%	15.1%	15.2%	15.9%	9.7%	16.8%	16.8%	14.1%	11.1%	9.0%	14.8%	10.4%	14.3%	14.6%	22.4%		29.8%	10.3%	13.3%	11.2%	11.2%	12.7%
M54	8.5%	13.3%	13.6%	12.4%	12.7%	14.4%	17.0%	14.6%	22.0%	22.0%	17.8%	14.3%	15.0%	16.9%	16.3%	11.1%	13.6%	15.5%	29.8%		11.9%	18.4%	18.4%	14.5%	23.0%
M97	11.3%	7.8%	10.7%	11.2%	12.2%	19.2%	11.5%	10.2%	15.2%	15.2%	10.1%	15.1%	13.2%	9.4%	11.7%	10.2%	8.8%	10.8%	10.3%	11.9%		18.2%	17.6%	13.9%	16.6%
M53	8.7%	10.6%	12.6%	13.8%	15.0%	18.6%	11.7%	9.2%	13.9%	13.9%	12.9%	14.8%	14.1%	16.7%	17.6%	12.6%	9.2%	12.3%	13.3%	18.4%	18.2%		46.6%	43.9%	52.3%
M70	9.0%	14.0%	15.8%	16.3%	19.2%	15.1%	12.5%	10.9%	11.5%	11.5%	10.4%	13.6%	14.4%	19.5%	15.9%	8.7%	7.7%	12.6%	11.2%	18.4%	17.6%	46.6%		34.3%	49.6%
M83	8.1%	8.0%	10.5%	7.9%	9.3%	14.1%	10.8%	9.0%	7.7%	7.7%	9.8%	12.5%	10.5%	9.3%	12.0%	17.7%	12.9%	11.0%	11.2%	14.5%	13.9%	43.9%	34.3%		51.4%
M98	9.9%	12.8%	16.1%	11.2%	13.3%	20.1%	13.9%	10.9%	14.7%	14.7%	14.2%	14.4%	14.4%	14.6%	15.4%	10.6%	8.4%	13.6%	12.7%	23.0%	16.6%	52.3%	49.6%	51.4%	

B

	emm57	emm3	emm12	emm11	emm48	emm85	emm69	emm4.2	emm60	emm 106	emm9	emm9.1	emm58	emm102	emm2	emm90	emm14	emm19	emm54	emm70	emm97	emm83	emm53	emm98	emm1
emm57		66.1%	67.0%	51.3%	51.3%	50.4%	50.4%	52.1%	52.1%	54.5%	55.4%	55.4%	53.6%	44.3%	44.3%	56.3%	53.6%	65.2%	66.1%	53.8%	61.3%	60.5%	62.2%	62.2%	61.3%
emm3	66.1%		98.2%	63.9%	63.9%	62.2%	62.2%	62.2%	62.2%	66.1%	64.3%	64.3%	64.3%	57.1%	57.1%	65.2%	67.0%	82.1%	81.3%	66.4%	71.4%	73.1%	74.8%	74.8%	75.6%
emm12	67.0%	98.2%		63.9%	63.9%	62.2%	62.2%	62.2%	62.2%	66.1%	64.3%	64.3%	64.3%	57.1%	57.1%	65.2%	67.9%	83.0%	82.1%	66.4%	71.4%	73.1%	74.8%	74.8%	75.6%
emm11	51.3%	63.9%	63.9%		100%	95.8%	96.6%	94.1%	94.1%	86.6%	85.7%	85.7%	84.0%	79.2%	79.2%	86.6%	61.3%	68.9%	70.6%	84.0%	67.2%	73.9%	73.9%	73.9%	74.8%
emm48	51.3%	63.9%	63.9%	100%		95.8%	96.6%	94.1%	94.1%	86.6%	85.7%	85.7%	84.0%	79.2%	79.2%	86.6%	61.3%	68.9%	70.6%	84.0%	67.2%	73.9%	73.9%	73.9%	74.8%
emm85	50.4%	62.2%	62.2%	95.8%	95.8%		97.5%	93.3%	93.3%	86.6%	85.7%	85.7%	84.0%	79.2%	79.2%	86.6%	58.0%	67.2%	67.2%	80.7%	65.5%	70.6%	70.6%	70.6%	73.1%
emm69	50.4%	62.2%	62.2%	96.6%	96.6%	97.5%		94.1%	94.1%	88.2%	86.6%	86.6%	84.9%	80.5%	81.8%	88.2%	58.8%	67.2%	67.2%	80.7%	66.4%	70.6%	70.6%	70.6%	73.1%
emm4.2	52.1%	62.2%	62.2%	94.1%	94.1%	93.3%	94.1%		100%	88.2%	86.6%	86.6%	84.9%	80.5%	81.8%	88.2%	60.5%	66.4%	68.1%	83.2%	66.4%	73.1%	73.1%	73.1%	72.3%
emm60	52.1%	62.2%	62.2%	94.1%	94.1%	93.3%	94.1%	100%		88.2%	86.6%	86.6%	84.9%	80.5%	81.8%	88.2%	60.5%	66.4%	68.1%	83.2%	66.4%	73.1%	73.1%	73.1%	72.3%
emm 106	54.5%	66.1%	66.1%	86.6%	86.6%	86.6%	88.2%	88.2%	88.2%		96.4%	96.4%	96.4%	92.9%	94.3%	96.4%	62.5%	71.4%	69.6%	79.0%	65.5%	68.9%	68.9%	68.9%	70.6%
emm9	55.4%	64.3%	64.3%	85.7%	85.7%	85.7%	86.6%	86.6%	86.6%	96.4%		100%	98.2%	98.6%	97.1%	98.2%	62.5%	69.6%	69.6%	79.0%	66.4%	68.9%	68.9%	68.9%	70.6%
emm9.1	55.4%	64.3%	64.3%	85.7%	85.7%	85.7%	86.6%	86.6%	86.6%	96.4%	100%		98.2%	98.6%	97.1%	98.2%	62.5%	69.6%	69.6%	79.0%	66.4%	68.9%	68.9%	68.9%	70.6%
emm58	53.6%	64.3%	64.3%	84.0%	84.0%	84.0%	84.9%	84.9%	84.9%	96.4%	98.2%	98.2%		97.1%	95.7%	96.4%	63.4%	69.6%	67.9%	77.3%	64.7%	67.2%	67.2%	67.2%	68.9%
emm102	44.3%	57.1%	57.1%	79.2%	79.2%	79.2%	80.5%	80.5%	80.5%	92.9%	98.6%	98.6%	97.1%		98.6%	98.6%	61.4%	67.1%	67.1%	68.8%	61.0%	64.9%	64.9%	64.9%	67.5%
emm2	44.3%	57.1%	57.1%	79.2%	79.2%	79.2%	81.8%	81.8%	81.8%	94.3%	97.1%	97.1%	95.7%	98.6%		100%	61.4%	67.1%	67.1%	68.8%	61.0%	64.9%	64.9%	64.9%	67.5%
emm90	56.3%	65.2%	65.2%	86.6%	86.6%	86.6%	88.2%	88.2%	88.2%	96.4%	98.2%	98.2%	96.4%	98.6%	100%		63.4%	70.5%	70.5%	79.8%	67.2%	69.7%	69.7%	69.7%	71.4%
emm14	53.6%	67.0%	67.9%	61.3%	61.3%	58.0%	58.8%	60.5%	60.5%	62.5%	62.5%	62.5%	63.4%	61.4%	61.4%	63.4%		79.5%	78.6%	62.2%	61.3%	62.2%	64.7%	64.7%	64.7%
emm19	65.2%	82.1%	83.0%	68.9%	68.9%	67.2%	67.2%	66.4%	66.4%	71.4%	69.6%	69.6%	69.6%	67.1%	67.1%	70.5%	79.5%		94.6%	69.7%	73.1%	76.5%	77.3%	77.3%	79.0%
emm54	66.1%	81.3%	82.1%	70.6%	70.6%	67.2%	67.2%	68.1%	68.1%	69.6%	69.6%	69.6%	67.9%	67.1%	67.1%	70.5%	78.6%	94.6%		70.6%	75.6%	79.0%	79.0%	79.0%	80.7%
emm70	53.8%	66.4%	66.4%	84.0%	84.0%	80.7%	80.7%	83.2%	83.2%	79.0%	79.0%	79.0%	77.3%	68.8%	68.8%	79.8%	62.2%	69.7%	70.6%		80.7%	89.9%	89.9%	89.9%	87.4%
emm97	61.3%	71.4%	71.4%	67.2%	67.2%	65.5%	66.4%	66.4%	66.4%	65.5%	66.4%	66.4%	64.7%	61.0%	61.0%	67.2%	61.3%	73.1%	75.6%	80.7%		87.4%	90.8%	90.8%	91.6%
emm83	60.5%	73.1%	73.1%	73.9%	73.9%	70.6%	70.6%	73.1%	73.1%	68.9%	68.9%	68.9%	67.2%	64.9%	64.9%	69.7%	62.2%	76.5%	79.0%	89.9%	87.4%		96.6%	96.6%	94.1%
emm53	62.2%	74.8%	74.8%	73.9%	73.9%	70.6%	70.6%	73.1%	73.1%	68.9%	68.9%	68.9%	67.2%	64.9%	64.9%	69.7%	64.7%	77.3%	79.0%	89.9%	90.8%	96.6%		100%	95.8%
emm98	62.2%	74.8%	74.8%	73.9%	73.9%	70.6%	70.6%	73.1%	73.1%	68.9%	68.9%	68.9%	67.2%	64.9%	64.9%	69.7%	64.7%	77.3%	79.0%	89.9%	90.8%	96.6%	100%		95.8%
emm1	61.3%	75.6%	75.6%	74.8%	74.8%	73.1%	73.1%	72.3%	72.3%	70.6%	70.6%	70.6%	68.9%	67.5%	67.5%	71.4%	64.7%	79.0%	80.7%	87.4%	91.6%	94.1%	95.8%	95.8%	

Figure 3.3: Pairwise identity analysis of M protein N-terminal (A) and C-terminal (B) gene sequences. Sequence alignments were carried out via the pairwise MUSCLE alignment tool using default settings on *Geneious* version 6.0, Biomatters. Sequence identity is presented as a heat map, whereby darker shades of grey correspond to an increase in percentage of identical amino acids between distinct M protein sequences.

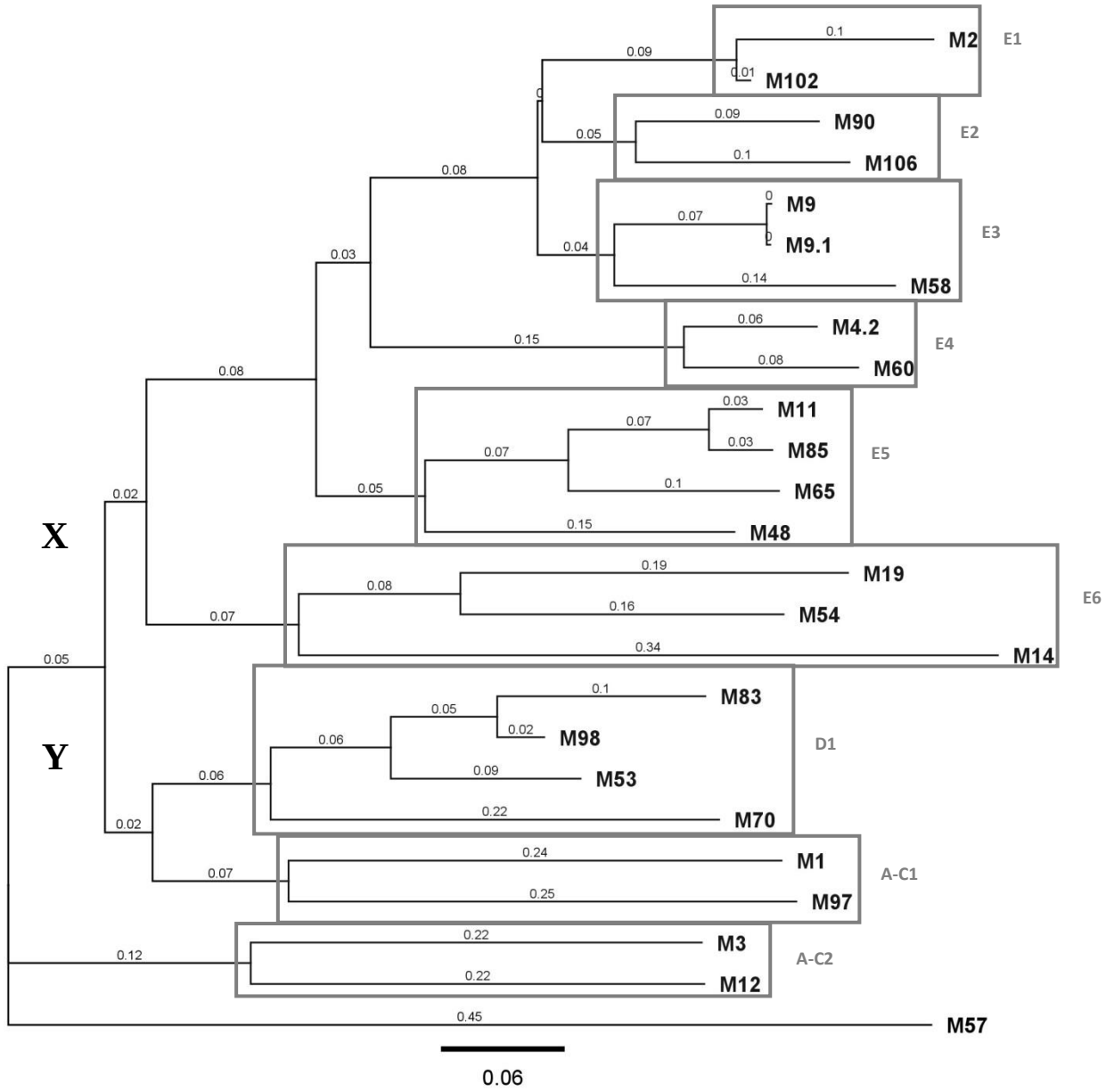
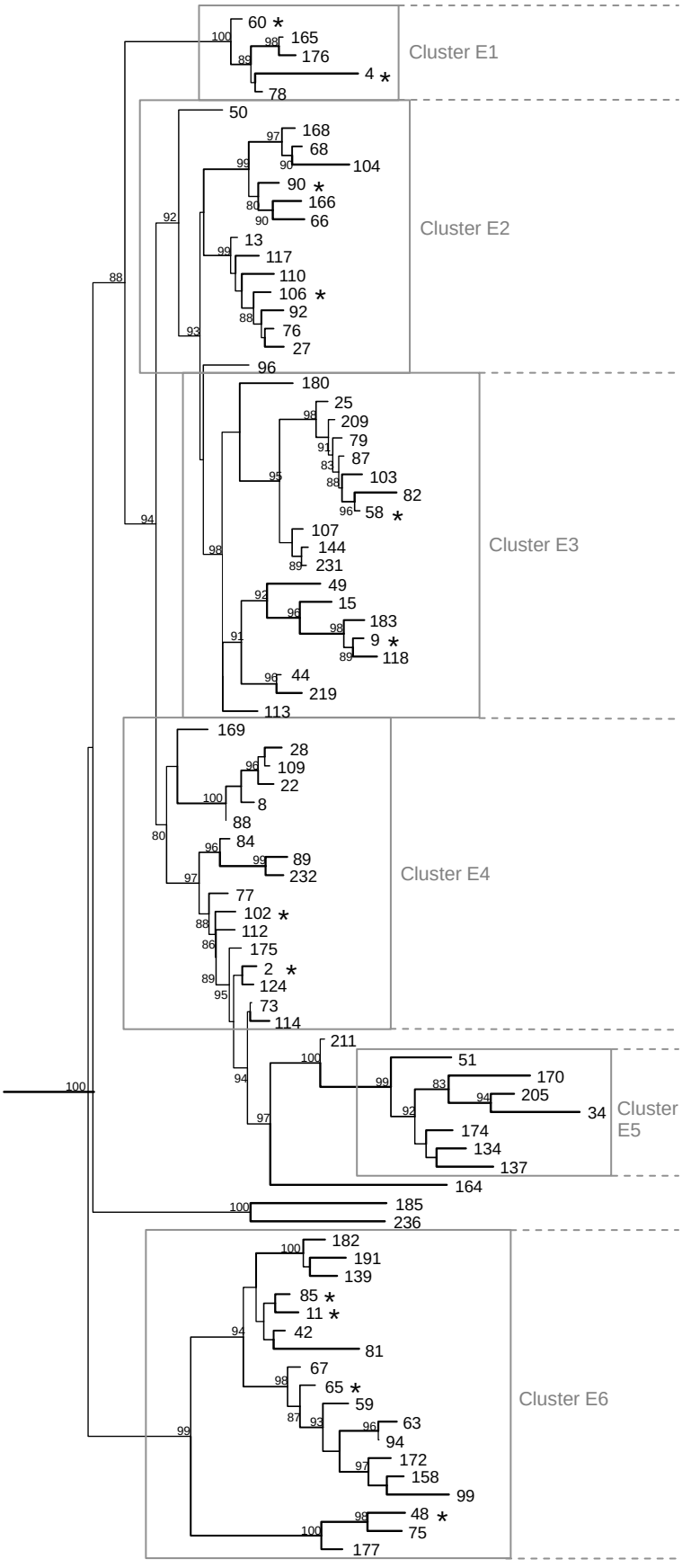


Figure 3.4: Phylogentic assembly based on M protein amino acid sequences from 25 M-types. The tree is drawn to scale, with branch length in the same units (number of amino acid substitutions per site) to those of the evolutionary distances used for the phylogenetic tree. For phylogenetic analysis M protein sequences were trimmed from the first amino acid of the mature protein to the first amino acid preceding the anchor LPxTG motif. The phylogenetic tree was constructed using Geneious *version 6.0* (Biomatters, USA) with default settings. M54 is represented twice through two different GAS strains.


Cluster E1
Cluster E2
Cluster E3
Cluster E4
Cluster E5
Cluster E6
60 *
165
176
4 *
78
50
168
68
104
90 *
166
66
13
117
110
106 *
92
76
27
96
180
25
209
79
87
103
82
58 *
107
144
231
49
15
183
9 *
118
44
219
113
169
28
109
22
8
88
84
89
232
77
102 *
112
175
2 *
124
73
114
211
51
170
205
34
174
134
137
164
185
236
182
191
139
85 *
11 *
42
81
67
65 *
59
63
94
172
158
99
48 *
75
177

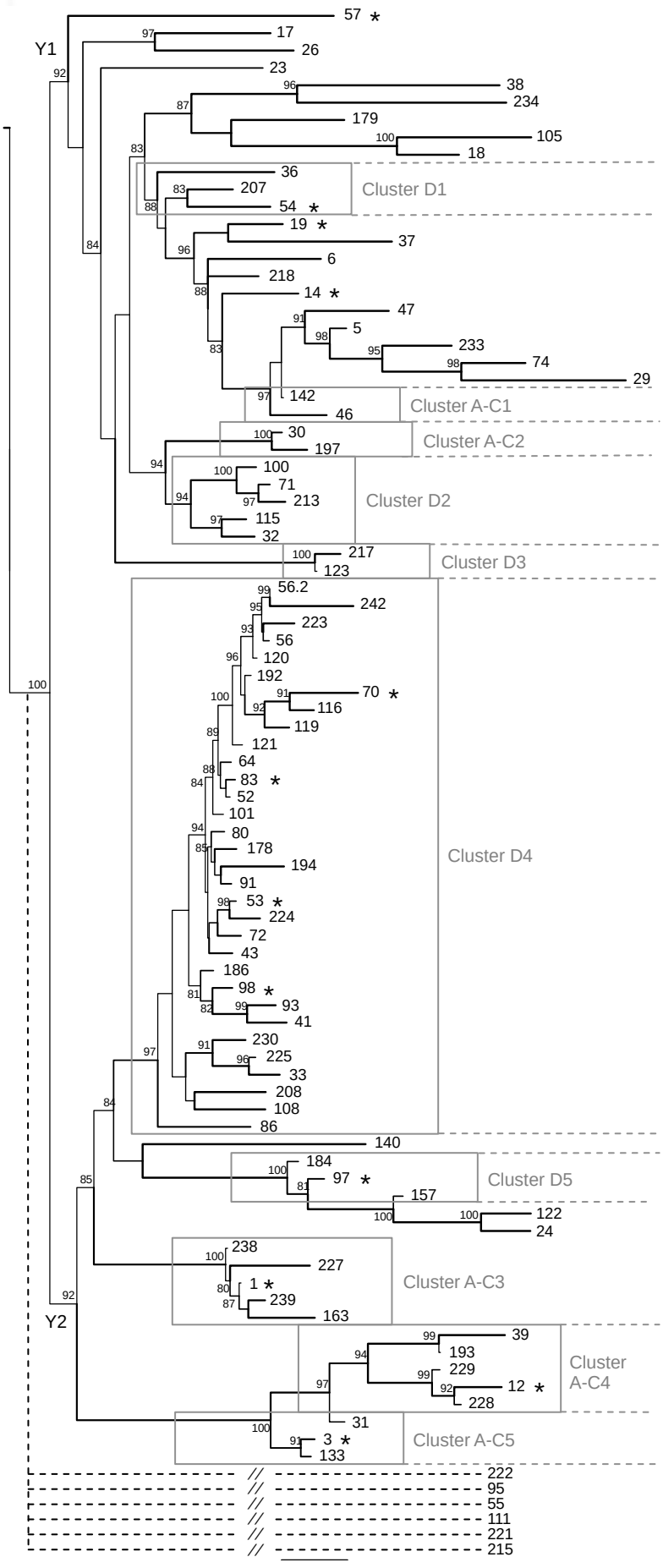

Y1
Y2
Cluster D1
Cluster A-C1
Cluster A-C2
Cluster D2
Cluster D3
Cluster D4
Cluster D5
Cluster A-C3
Cluster A-C4
Cluster A-C5
0.5

Figure 3.5: Systematic cluster classification via phylogenetic analysis of M proteins from 175 M-types. The tree is drawn to scale, with branch length in the same units (number of amino acid substitutions per site) to those of the evolutionary distances used for the phylogenetic tree. Phylogenetic inferences of M protein sequences were drawn by PhyML. For phylogenetic analysis, M protein sequences were trimmed from the first amino acid of the mature protein to the first amino acid preceding the anchor LPxTG motif. Test values >80% are indicated at the tree nodes. The tree has two major clades: Clade X is comprised of six main clades whereas clade Y is divided into two sub-clades that are then divided into 10 main clusters. Six main outliers are indicated via dashed lines. The (*) mark indicates recombinant M protein representatives expressed in *E. coli*.

3.2.3 Functional M protein classification

Functional analysis of 26 M proteins derived from each of the major cluster groups was undertaken to assess binding to key host proteins known to interact with M proteins. Following expression in *E. coli*, purified recombinant M proteins were found to range from 36-58 kDa (Fig 3.6). Some M proteins appeared as doublet bands following SDS-PAGE analysis, which is characteristic of certain M proteins (Cunningham 2000). To further confirm that purified proteins represented M protein, proteins were subjected to Western blot analysis using rabbit polyclonal α-M53$_{NS13}$ antisera. This antisera has been shown to recognise multiple M proteins due to cross-reactivity with the C repeat domain (Sanderson-Smith *et al.* 2006). All 26 recombinant M proteins were detected using this method at the same molecular weight as observed in SDS-PAGE analysis. High and low molecular weight proteins visible in some samples may be indicative of protein dimerisation or protein breakdown.

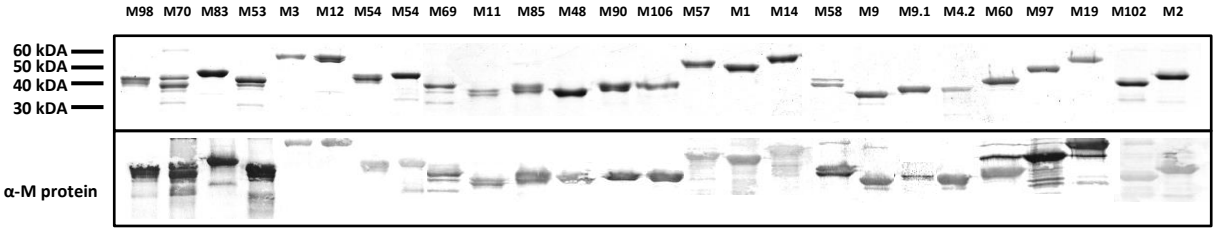


Figure 3.6: SDS-PAGE and Western blot analyses of purified recombinant M proteins. Proteins were electrophoresed under non-reducing conditions on a 12% acrylamide gel. Molecular weight markers are shown on the left and are given in kilodaltons (kDA) (above). All M proteins were shown to react with rabbit polyclonal α-M53$_{NS13}$ antisera (below).

The secondary structure characteristics of all 26 recombinant M proteins were analysed using far UV CD spectroscopy. All M protein were found to have a CD emission spectrum characteristic of an α-helical coiled-coil protein (Freifelder 1982), displaying two distinctive minima at 210 nm and 220 nm, and a maximum peak at 190 nm (Fig. 3.7A-L). These emission spectra were similar to that of previously characterised M proteins (Nilson *et al.* 1995). Percent α-helicity was found to range from 77% to 99%. *emm*-cluster D1 M protein representatives and M19 situated just outside of *emm*-cluster D1 demonstrated the lowest level of α-helicity ranging from 76-81%. No significant differences in percent α-helicity were observed between remaining *emm*-cluster groups.

To assess whether the phylogenetic classification system serves as a synergistic functional classification system, M protein cluster representatives were examined for binding to select host proteins; specifically, IgG, IgA, C4BP, plasminogen, fibrinogen and albumin. M proteins were expressed with C-terminal $(His)_6$ tags. The positioning of the C-terminal tag ensured that for SPR experiments, M proteins were captured and presented in an orientation which allowed the N-terminal portion of the protein to extend away from the BIAcore sensor surface. By capturing the C-terminal region of each M protein, we were able to closely model the presentation of M protein at the GAS cell surface, thus allowing for a more physiologically relevant comparison of each binding interaction. Sequence alignments with annotation and binding profiles of all truncated M protein can be found in Appendix E.

The binding of IgG was found to be restricted primarily to clade X with affinity constants ranging from 2.06 nM to 82.69 nM. IgG binding was specifically observed for clusters E1-E6 as well as clade Y representatives, M57 (non-clustered A-C) and M1 (A-C3) (Fig 3.8A & C). As previously identified in M2 (E4 representative), IgG3 binding is dependent on the presence of conserved glutamic acid (E) and glutamine (Q) rich amino acid sequences, a common property of IgG binding proteins (Pack *et al.* 1996). The 35 amino acid EQ-rich

sequence characterised in M2 was found to be highly conserved in clusters E1-4 with sequence identity ranging from 69% to 100% (Fig 3.8B). M1 has been previously shown to bind IgG, via the 38 amino acid S-domain (Akesson *et al.* 1994). M48 (E6 representative), did not express an EQ rich motif, but was shown to share a small level of sequence conservation (41%) with the S-region of M1.

No other M protein sequence in this study exhibited sequence conservation consistent with the previously characterised M1 S-domain (Fig 3.8D). With the exception of M48, sequence conservation between the EQ rich IgG binding motif present in clusters E1-E4 and the S-domain of M1 were absent in E6 clustered M proteins as well as in M57 and M14, suggesting the existence of an additional IgG binding site (Fig 3.8A-D).

As with IgG, IgA binding was limited to clade X. The IgA binding motif 'ALxGENxDLR' has been previously characterised in M4 protein (Bessen 1994). MUSCLE pairwise alignment analysis of M protein sequences with this IgA binding motif revealed that M protein representatives from cluster E4 demonstrated 50% sequence identity whereas M proteins from clusters E1, and E6 shared 64-100% sequence identity with the previously characterised IgA binding motif. High affinity IgA binding was demonstrated by cluster E1 and E6 M protein representatives with affinity constants ranging from 0.66 nM to 5.36 nM. Although M proteins from both cluster E1 and E6 contained the previously characterised IgA binding motif, all except M65 bound IgA (Fig 3.8E-F).

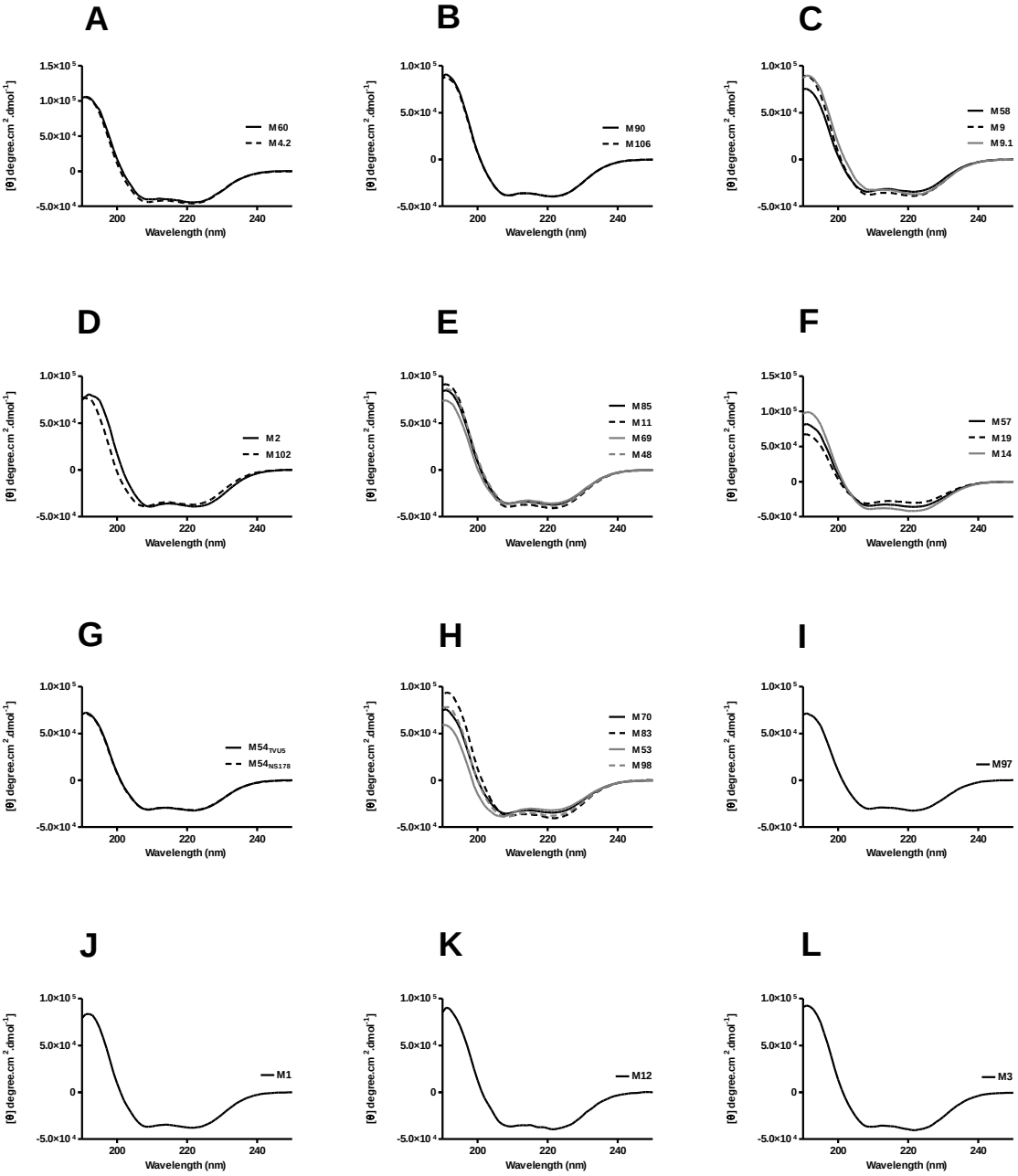


Figure 3.7: Circular dichroism spectra of phylogenetically clustered recombinant M protein. A, E1; B, E2; C, E3; D, E4; E, E6; F, non-clustered A-C M proteins; G, D1; H, D4; I, D5; J, A-C3; K, A-C4; L, A-C5. All M protein cluster representatives exhibit CD emission spectra characteristic of α-helical coiled-coil proteins, displaying two minima at approximately 210 nm and 220 nm and a maximum peak at 190 nm.

C4BP, a ligand that confers phagocytosis resistance, is known to bind the N-terminal, hypervariable region of M22 (Persson *et al.* 2006). C4BP was observed to bind exclusively to M protein from clade X. Clusters E1, E3, E4 and E6 bound C4BP with extremely tight affinity (K_D = 5.10 pM – 119.93 pM) (Fig 3.9A). M102 from cluster E4 demonstrated no binding to C4BP. No significant conservation in sequence identity was observed between the previously published N-terminal domain of M22 and the M proteins in this study. Since binding of C4BP has been previously localised to the hypervariable N-terminal section of the M protein, it may explain why a defined binding motif has yet to be identified.

The capacity to bind fibrinogen was only demonstrated by clade Y M protein representatives. M protein-fibrinogen binding interactions were restricted to cluster D1, AC3-5 and M proteins from subclade Y1 (M57, M54, M19 and M14) with affinity constants of 0.09 nM - 0.64 nM (Fig 3.9B). To date, a conserved fibrinogen binding motif has not been identified. Previous studies analysing the capacity of M1 and M5 to bind fibrinogen identified that binding was dependant on the B1 and B2 repeats within M1 and M5 protein even though they were shown to not share homologous B1 and B2 sequences (Ringdahl *et al.* 2000; Waldemarsson *et al.* 2009). Furthermore, for M1, fibrinogen binding was suggested to be dependent on irregularities within the coil-coiled structure of B1 and B2 domains (McNamara *et al.* 2008). Sequence alignment of M proteins from each cluster group identified limited sequence homology with the B repeat domains of M1 supporting the suggestion of a structural non-linear binding motif.

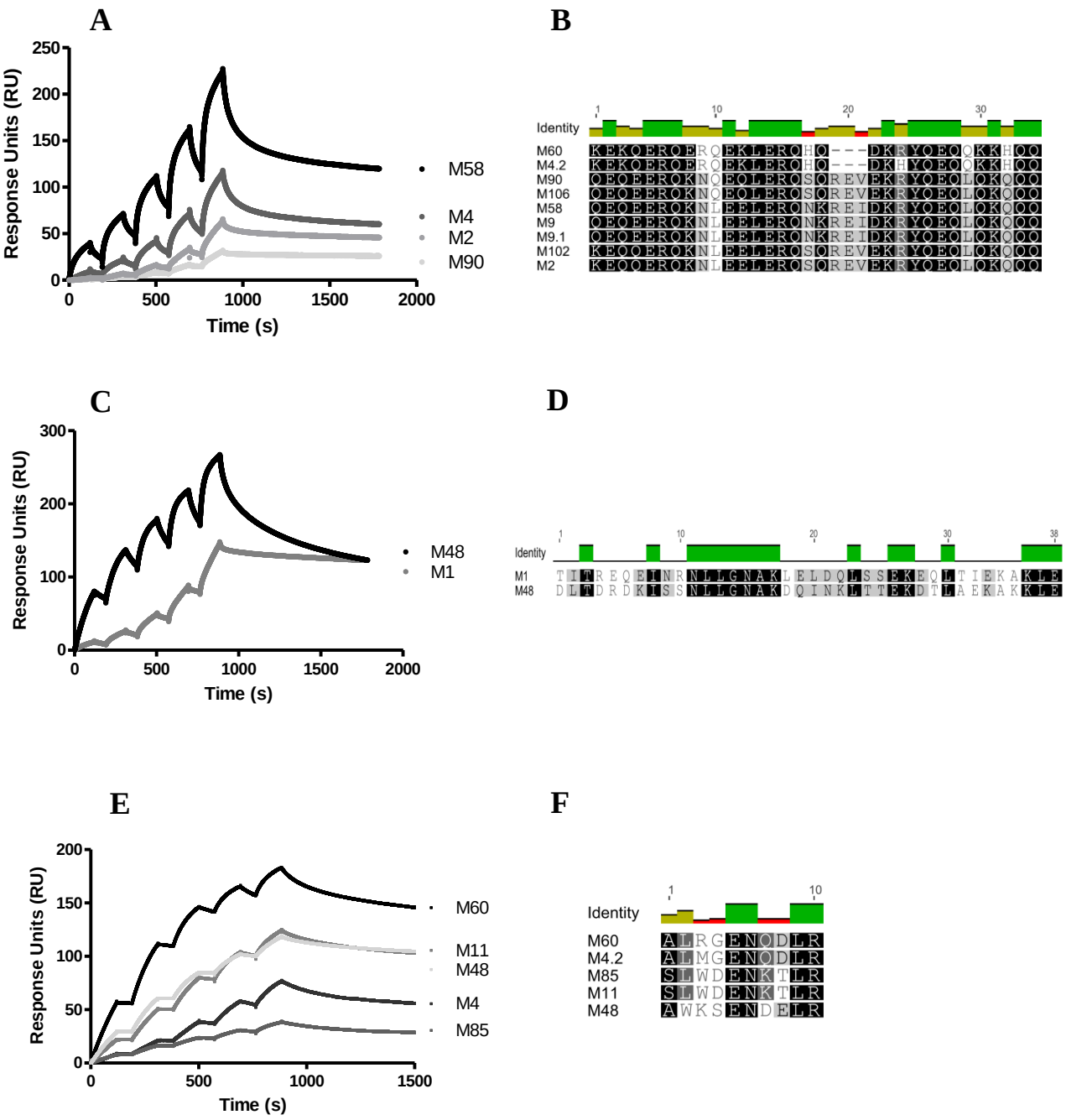


Figure 3.8: Binding of IgG and IgA by clustered M protein representatives. All M proteins from clusters E1-4, E6 and AC-3 bound IgG (A & C). Unclustered M57 and M14 were also shown to bind IgG (Data not shown. See appendix E). Sequence alignments identified that clusters E1-4 share EQ-rich IgG binding motifs where as M48 was shown to have considerable sequence conservation with the S-domain of M1 (B & D). IgA binding was restricted to cluster E1 and E6 M proteins which contained the previously characteriesed IgA binding motif (E & F). All kinetic assays were performed using analytes at varying concentrations (0-200 nM) over a series of five 60 s injections at a flow rate of 30 µl/min with a 900 s dissociation period. Affinity interactions were determined via surface plasmon resonance and were analysed by non-linear fitting of the single cycle kinetic sensograms according to a 1:1 Langmuir binding model.

As with fibrinogen, plasminogen binding was specific to clade Y. Conservation of lysine, arginine, and histidine residues in the N-terminal A1 and A2 repeats of M53 and M98 have been shown to be critical in the binding of plasminogen (Ringdahl *et al.* 2000; Sanderson-Smith *et al.* 2006; Sanderson-Smith *et al.* 2007). SPR analysis identified that plasminogen binding was a restricted to cluster D4 with affinity constants ranging from 1.33 nM to 3.06 nM (Fig 3.10A). The highly-conserved plasminogen binding motif (EAELERLKSERHD) was present only in D4 M proteins, as well as in M140 protein, situated directly adjacent to the D4 cluster with sequence conservation ranging from 62-100% (Fig 3.10B). This motif can therefore be considered predictive of plasminogen binding M proteins.

All cluster groups examined with the exception of E4, contained representative M proteins that bound albumin (Fig 3.10C), which is in agreement with prior studies (Sandin *et al.* 2006). Within cluster D4 only M53 and M98 were shown not to bind albumin. M protein-albumin interactions exhibited affinity constants ranging from 2.17 nM to 18.14 nM. The binding of albumin by M proteins has been attributed to a 14 amino acid motif (RDLDASREAKKQVE) localised in the C-repeat domain (Retnoningrum *et al.* 1994). This motif was present in all M proteins, including those which did not bind albumin whereby sequence identity ranged from 79-100 % (Fig 3.10D).

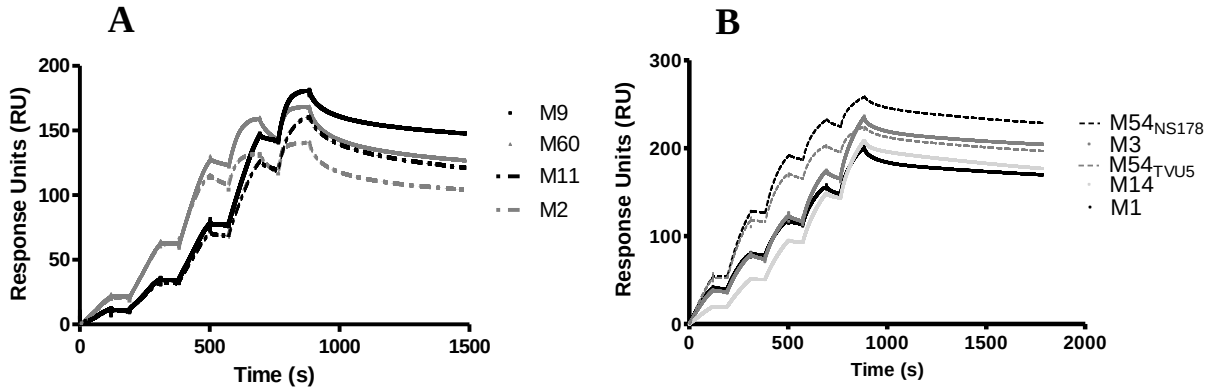


Figure 3.9: Binding of C4BP and fibrinogen by clustered M protein representatives. C4BP binding was restricted to clade-X M proteins from clusters E1, E3, E4 and E6 (A). Fibrinogen binding was only observed by M proteins from clade Y, specifically D1, D5 and A-C3-5 clusters. M proteins from subclade Y1 (M57, M54, M19 and M14) were also shown to bind fibrinogen (M57 and M19 data not shown. See appendix E) (B). Representative single cycle kinetic SPR sensograms are shown for M proteins originating from each major cluster group. All kinetic assays were performed using analytes at varying concentrations (0-200 nM) over a series of five 60 s injections at a flow rate of 30 µl/min with a 900 s dissociation period. Affinity interactions were analysed by non-linear fitting of the single cycle kinetic sensograms according to a 1:1 Langmuir binding model.

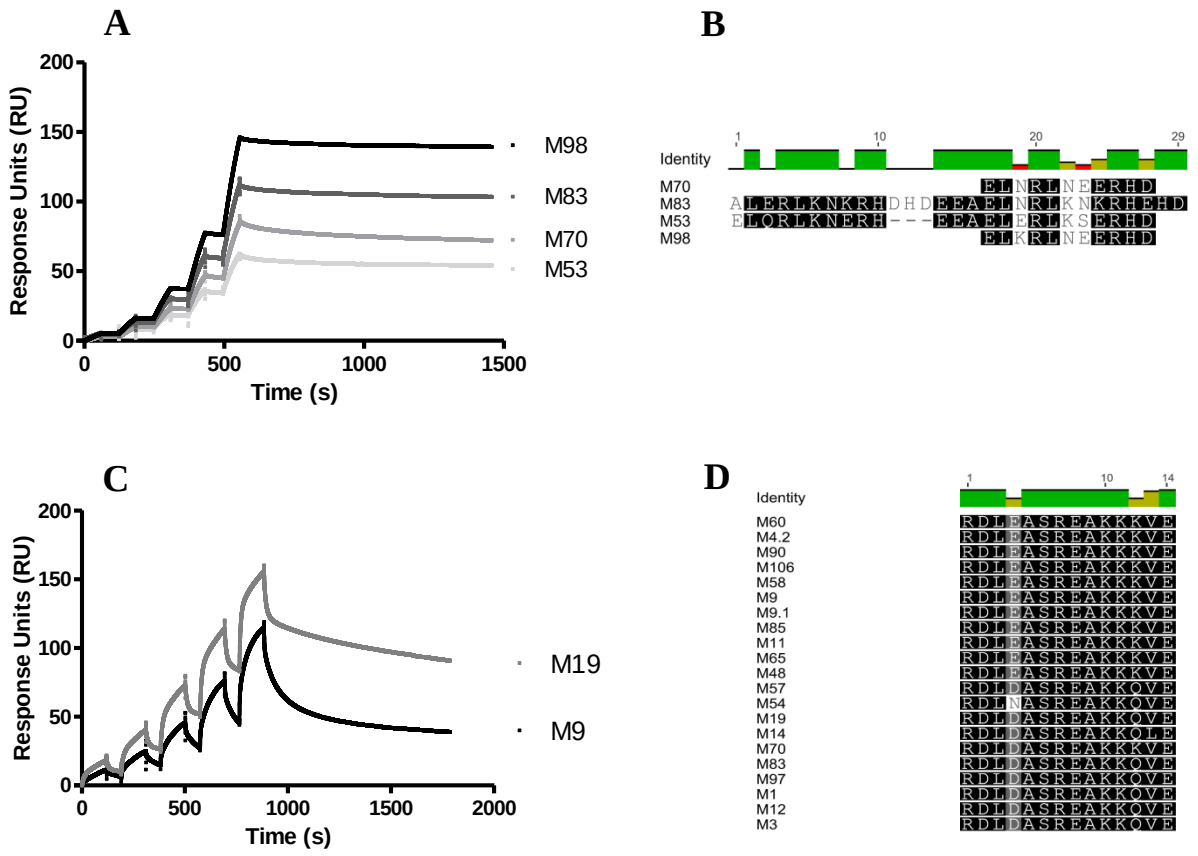


Figure 3.10: Binding of plasminogen and albumin by clustered M protein representatives. Plasminogen binding was restricted to cluster D4 M proteins (A) which contained the highly conserved plasminogen binding motif (B). Albumin was shown to bind all M proteins except protein representative from cluster E4, and M53 and M98 from cluster D4 (Only M19 and M9 shown. See appendix E for remaining data) (C). All M proteins, including those which did not bind albumin, possesed the previously characterised albumin binding motif (D). All kinetic assays were performed using analytes at varying concentrations (0-200 nM) over a series of five 60 s injections at a flow rate of 30 µl/min with a 900 s dissociation period. Affinity interactions were determined via surface plasmon resonance and were analysed by non-linear fitting of the single cycle kinetic sensograms according to a 1:1 Langmuir binding model.

3.2.4 SV1 competitive binding analysis

Recently, a J-14 variant encompassing sequences from 77 different M-types, termed SV1, was described (Bauer *et al.* 2012). Despite emphasis on its ability to elicit an effective immunogenic response, the ability of SV1 to effectively opsonise GAS via M protein directed antigens in the presence of host proteins has not been examined. This is of interest as blocking of host proteins could constrain the candidature of this novel GAS vaccine candidate. Alternatively, it has been hypothesised that the generation of antibodies which block M-protein – host protein interactions may further contribute to a reduction in GAS virulence following vaccination (Sandin *et al.* 2006).

To assess the recognition of α-SV1 IgG against M protein, and examine the associated downstream effects on M protein-host protein binding interactions, competitive binding analysis was performed using M protein and α-SV1 IgG in the presence of various host proteins (IgG, IgA, plasminogen, fibrinogen and albumin) (Appendix E). Host proteins were selected based on their known capacity to interact with different repeat domains of different M protein serotypes, ensuring complete coverage of ligand binding sites along the M protein, which may affect α-SV1 IgG-M protein interactions.

All M proteins in this study displayed high affinity for α-SV1 IgG, with affinity constants ranging from 6.82 nM to 108.75 nM. Known IgG binding motifs such as the EQ rich domain and the S-domain of M1 are situated within the central region of the M protein, approximately 20-25 residues upstream of the α-SV1 IgG target motif. Competitive binding assays indicated that the presence of IgG binding motif did not significantly reduce M protein recognition by SV1. When saturated with either IgG or α-SV1, IgG binding M protein representatives from clusters E1, E3, E4, E6, A-C3 along with M proteins M14 and M57 were able to bind α-SV1 IgG and IgG with slightly reduced affinity (IgG affinity: K_D = 6.29 nM – 102.63 nM; α-SV1 IgG affinity: K_D = 10.29 nM – 71.18 nM) (Fig 3.11A-B). The IgA

binding motif has been previously localised to the N-terminal portion of M protein, distal from the α-SV1 IgG binding motif (Johnsson *et al.* 1999). IgA binding by E1 cluster M proteins was shown to outcompete interactions with α-SV1 IgG, whereas M proteins from cluster E6 were able to bind α-SV1 IgG in the presence of IgA (K_D = 18.65 nM to 48.76 nM). M proteins from both cluster E1 and E6 displayed high affinity for IgA in the presence of α-SV1 IgG (K_D = 0.35 nM to 2.22 nM) (Fig 3.11C-D).

Plasminogen binding function has been attributed to the N-terminal A repeat domains of cluster D4 M proteins (Berge *et al.* 1993; Sanderson-Smith *et al.* 2006). As expected due to the proximity of both binding domains, saturation of D4 M protein with either plasminogen or α-SV1 IgG was unable to block the binding by respective ligands (plasminogen affinity: K_D = 11.10 nM – 39.29 nM; α-SV1 IgG affinity: K_D = 31.32 nM -106.95 nM) (Fig 3.12A-B). M protein-fibrinogen interactions have been shown to be dependent on a combination of sequence conservation and tertiary structure irregularities in the coiled-coil assembly of the B-repeat domain. Fibrinogen binding has been shown to be restricted to clusters D1, A-C3-5, and M proteins M57, M19 and M14. Neither α-SV1 IgG nor fibrinogen appeared to compete for binding to respective M protein (α-SV1 IgG affinity: K_D = 64.84 nM - 129.15 nM; fibrinogen affinity: K_D = 0.10 nM – 0.64 nM) (Fig 3.12C-D). Albumin binding by M proteins has been localised to the C-repeat domain which is highly conserved among M protein (Retnoningrum *et al.* 1994). The J-14 variant sequences which form the basis of SV1 are expressed within M protein C1 and C2 repeat domains which subsequently contains the previously characterised albumin binding motif (Bauer *et al.* 2012). Interestingly, M proteins from all cluster groups (except E4, M53 and M98) were able to bind α-SV1 and albumin in the presence of respective ligands (α-SV1 affinity: K_D = 9.10 nM – 158.0 nM; albumin affinity: K_D = 2.17 nM – 18.14 nM) (Fig 3.12E-F).

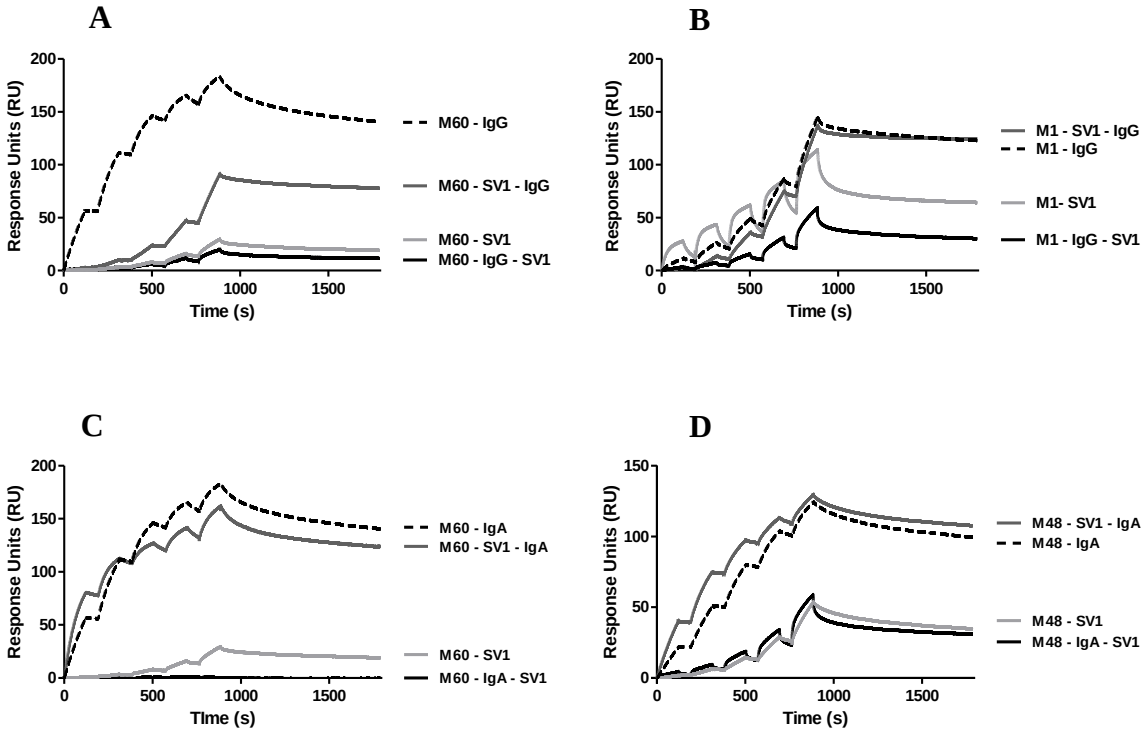


Figure 3.11: Saturation binding analysis of IgG, IgA and α-SV1-IgG to immobilised recombinant M protein. Recombinant poly-histidine tagged M protein was immobilised and saturated with IgG (A & B), IgA (C & D) and α-SV1 IgG (A-D). Competitive binding analysis between M proteins, IgG and IgA against α-SV1 IgG was measured and analysed by non-linear fitting of the single cycle sensograms according to a 1:1 Langmuir binding model using Biacore T200 evaluation software (Biacore AB). Competitive binding data of the remaining M proteins can be found in Appendix E.

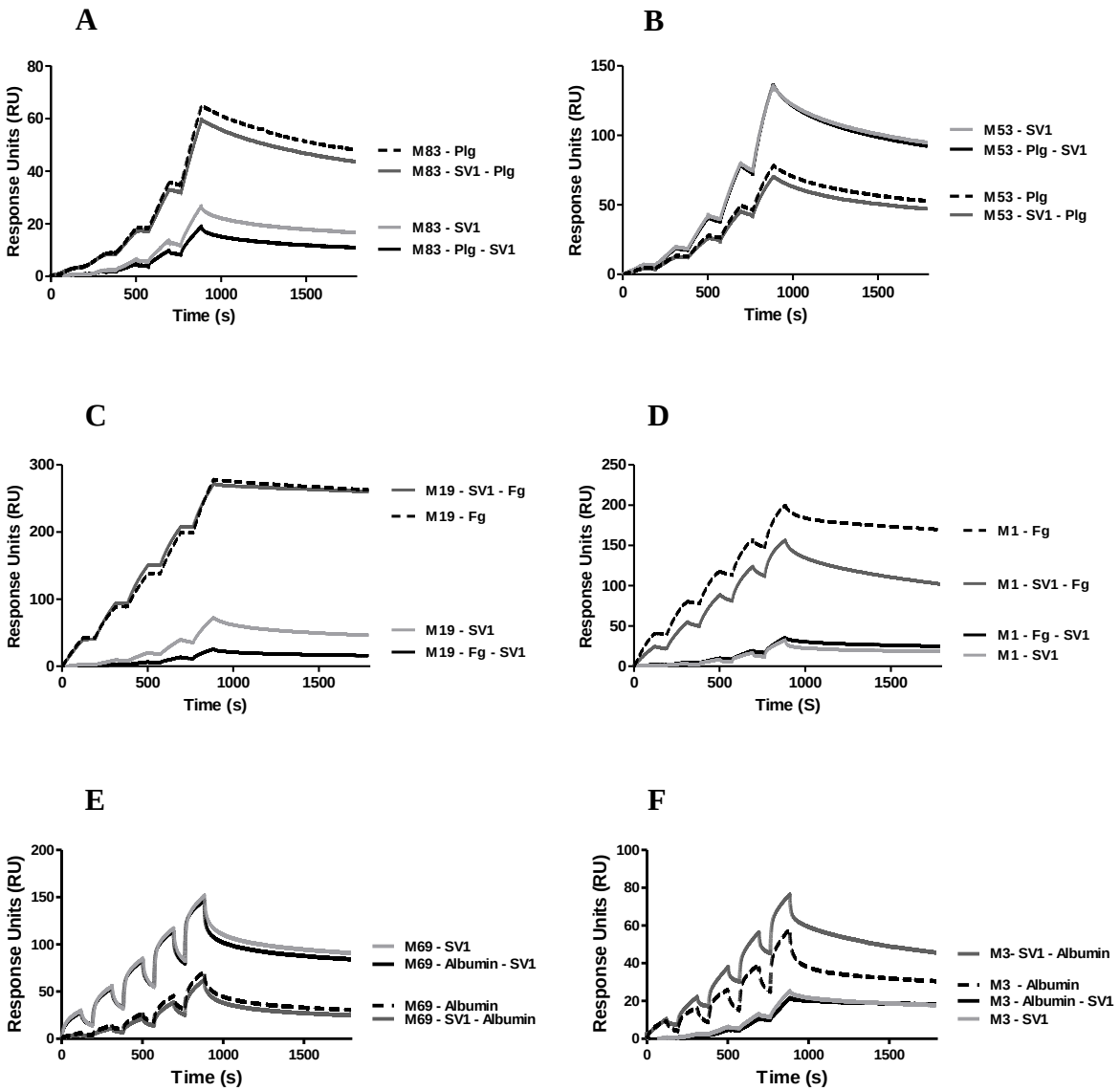


Figure 3.12: Saturation binding analysis of plasminogen, fibrinogen, albumin and α-SV1IgG to immobilised recombinant M protein. Recombinant poly-histidine tagged M protein was immobilised and saturated with plasminogen (A-B), fibrinogen (C-D), albumin (E-F) and α-SV1IgG (A-E). Competitive binding analysis between M proteins, plasminogen, fibrinogen and albumin against α-SV1-IgG was measured and analysed by non-linear fitting of the single cycle sensograms according to a 1:1 Langmuir binding model using Biacore T200 evaluation software (Biacore AB). Competitive binding data of the remaining M proteins can be found in Appendix E.

3.3 Discussion

Until recently, GAS classification has been based on sequence variation within the initial N-terminal section (10-15%). This has fostered concerns regarding type-specific immunity, and whether different M-types similar in sequence can exhibit cross-protection via M protein targeted GAS vaccines (Lancefield 1962; Bauer *et al.* 2012; McMillan *et al.* 2013). In this study, development of a novel GAS classification system that correlates full length M protein sequence with M protein function was undertaken using 1086 GAS isolates collected from 31 countries representing 175 M-types. This study confirmed the utility of a system which encompasses phylogenetic cluster based classification of full length M protein sequence in conjunction with M protein function.

At the amino acid level, phylogenetic analysis demonstrated that M proteins could be grouped into 2 clades, 2 sub-clades and 48 clusters, 16 of which encompass 82% of the M-types in this study. Both at the nucleotide and amino acid level, the N-terminal portions of respective *emm*-genes and translated M protein sequences demonstrated high levels of sequence divergence. Through both sequence and phylogenetic analysis, it is evident that different M protein cluster groups are under different patterns of diversifying selection. The observed diversification, occurring predominately at A and B repeat domains has given rise to 16 homogenous cluster groups. If *emm* genes/M proteins within a cluster group are experiencing similar rates of non-synonymous mutation, than it could be assumed that different M-types alike in sequence are undergoing similar levels of diversification, brought about through common environmental and host-related immune factors. Highly variable A and B domains possess opsonic epitopes and generally determine the function of individual M-types, facilitating the binding of certain host proteins (Ringdahl *et al.* 2000; Sanderson-Smith *et al.* 2006). It has been suggested that the functional domains expressed in A and B repeat domains may constrain the amount of sequence divergence exhibited by an M-type

(McNamara *et al.* 2008). Segregation of 32 single M protein cluster groups, divergent from the 16 main cluster groups, may have resulted through allelic variation within an M-type, giving rise to mutant populations with altered function and host immunity.

The plasminogen and IgA binding regions of M proteins are examples of functional domains which restrict diversifying selective pressures in the N-terminal domains of select subsets of clustered M proteins. The restricted binding function of plasminogen and IgG binding M protein has been shown to be integral in eliciting pathogenesis and facilitating immune evasion (O'Toole *et al.* 1992; Sanderson-Smith *et al.* 2008; Ly *et al.* 2014). The highly predictive plasminogen binding motif was restricted to cluster D4 M protein, which displayed high affinity for the respective host protein. Previous studies analysing the plasminogen binding characteristics of M53 identified that plasminogen binding is mediated via conserved arginine, histidine, and lysine residues in the A1 and A2 repeat domains (Wistedt *et al.* 1995; Sanderson-Smith *et al.* 2006). Together, lysine, arginine, and histidine make up a highly hydrophilic domain in the A1/A2 repeat domains of D4 M protein representatives. Prior analysis of the plasminogen binding receptor, urokinase plasminogen activator identified that a hydrophilic ligand binding domain is essential for the recruitment of plasminogen (Behrendt *et al.* 1991). IgA binding has been observed to correlate with infection of the nasopharyngeal mucosa (Bessen *et al.* 1990). M protein representatives from clusters E1 and E6 were shown to exclusively express the previously characterised N-terminal IgA binding motif (Bessen 1994). Interestingly, of the proteins functionally assessed from cluster E6, M65 was unable to bind IgA even though it expresses the IgA binding motif. If the IgA binding motif in M protein requires a particular conformation in order to bind IgA, then flanking sequences may be critical for maintaining this conformation.

Comparison of M protein central regions revealed a high level of sequence variation between M proteins. Perhaps the most well characterised function of the central region is the ability of

the B repeat domain of M1 and M5 to mediate fibrinogen binding (Waldemarsson *et al.* 2009; Macheboeuf *et al.* 2011). The ability of M proteins to bind fibrinogen is thought to prevent opsonisation, and ultimately phagocytosis of GAS. Fibrinogen binding was found to be restricted to clusters D1, A-C3-5 and M proteins M57, M54, M19 and M14 which are situated in close proximity to cluster D1. Ringdahl *et al* (2000), suggested that different GAS strains may have evolved different fibrinogen binding sequence motifs within their M proteins while retaining the same function. This suggestion is supported by the finding that no sequence identity was observed between the B repeat domains of the fibrinogen binding proteins in this study. IgG binding has also been shown to be a function of the central domains of M protein. IgG binding by M protein has been associated with the prevention of complement activation at the GAS cell surface (Stenberg *et al.* 1992; Carlsson *et al.* 2003). IgG binding clusters E1-E4 were observed to contain the previously described EQ rich region reported for IgG3 binding by M2 protein (Pack *et al.* 1996). This motif was not shared among other IgG binding M protein in study such as cluster E6 M protein, M1 (A-C3) and M57 (non-clustered A-C). Prior studies analysing M1 protein have suggested IgG-Fc regions bind M1 via the S-domain, directly upstream of the C repeats (Akesson *et al.* 1994). The lack of sequence conservation between the IgG binding motif of cluster E1-E4 M protein, the M1 S-domain, and the M protein from E6, M57 and M14 suggest the existence of additional uncharacterised IgG binding sites. Furthermore, the binding interactions between M protein from cluster E1-E4 and IgG were found to be weaker in affinity than those observed by other IgG binding M protein in study. This could imply that the different clusters of M proteins function to bind different subclasses of IgG (Jefferis *et al.* 1990).

Previous research has highlighted that the C-repeat domains of M protein are highly conserved (Fischetti 1989; Smeesters *et al.* 2008). Sequence analysis of C-repeat domains from the M proteins in this study identified that these regions share a high level of sequence

identity. The binding of albumin has been shown to be localised to the C-repeats domains. Although in this study albumin was not observed to block α-SV1 IgG binding, prior work has shown that albumin-M protein interaction is capable of blocking the recognition of C-repeat targeted antibodies (Akesson *et al.* 1994; Retnoningrum *et al.* 1994). The previously characterised albumin binding motif was present in nearly all sequences from this study, including those which did not bind albumin (M102, M2, M53 and M98). Prior work examining M23 (sub-clade Y) and M1 (A-C3 cluster) proteins suggested that regions adjacent to the C-repeat domains are involved in stabilising the coiled-coil conformation, essential for interaction with albumin (Gubbe *et al.* 1997; Hong 2007). These data clearly highlight the importance of a whole M protein sequence-based approach in examining interactions between different M protein regions, and the impact of these interactions on GAS pathogenesis.

To further examine the effects of M protein binding interactions on adjacent binding motifs, functional binding analysis was undertaken using α-SV1 IgG in the presence and absence of select host proteins. Recent studies have shown the J-14 variant M protein epitope is accessible to α-SV1 IgG at the GAS cell surface (Bauer *et al.* 2012). The ability of α-SV1 IgG to recognise a wider range of M protein targeted antigens was assessed in the presence of host proteins; IgG, IgA, plasminogen, fibrinogen and albumin. M protein representatives from every cluster group were observed to bind α-SV1 IgG with tight to moderate affinity. Variations in binding affinity between M proteins for α-SV1 IgG may be attributed to conformational differences in the C-repeat domain, resultant from upstream flanking residues (Hong 2007). Furthermore, IgG binding M protein representatives were generally observed to bind α-SV1 IgG with tighter affinity than non-IgG binding M proteins. This suggests that two binding sites such as the EQ rich domain of E1-E4 M protein or the S-domain of M1 may be co-operatively involved with the C-repeat α-SV1 IgG binding motif. The degree of

contribution each M protein IgG binding motif plays in binding α-SV1 is yet to be evaluated. The α-SV1 M protein epitope was found to be distinct from the M protein IgG binding domain in all characterised IgG binding M proteins. Thus, M protein-IgG interaction was not expected to block the α-SV1 binding site. Conversely, α-SV1 was hypothesised to bind both the IgG binding site and the α-SV1 epitope within the M protein. Based on the similar K_D values each M protein displays for α-SV1 IgG in the presence and absence of IgG, it can assumed that α-SV1 IgG interaction with M protein is predominately restricted to the α-SV1 epitope, and not the respective M protein IgG binding motifs.

Only M-types M4 (E1) M11 and M48 (E6) were shown to have significantly reduced affinity for both α-SV1 IgG and IgG when in the presence of each respective protein, supporting the presence of a more heterogeneous multi-domain α-SV1 IgG binding motif. Differences in the binding affinity of IgG binding M protein for IgG and α-SV1 IgG could be the consequence of discrete binding sites on IgG. Previous studies have identified that in IgG deficient areas such as saliva, binding of IgG by M1 is mediated via the Fc region of IgG, where as in blood, binding occurs via the Fab domain (Nordenfelt *et al.* 2012). Observed differences in affinity observed between IgG binding M proteins for IgG and α-SV1 IgG may be due to differences in affinity for Fab and Fc domains of IgG.

Functional analysis identified that M proteins from clusters E1 and E6 were able to bind IgA. IgA binding by E1 M proteins was shown to both inhibit and outcompete α-SV1 IgG binding interactions. Interestingly, IgA binding by E6 M proteins was not affected by or able to outcompete α-SV1 IgG binding. IgA binding motifs present in cluster E1 M proteins are situated closer in proximity to C-repeat domains than in cluster E6 M proteins. To explain the functional differences displayed by these two IgA cluster groups, two theories could be put forward. IgA is 320 kDA in size, thus steric hindrance may influence and mediate inhibition of E1 M protein function. Secondly, structural changes induced by IgA or α-SV1 IgG binding

may be more prominent in cluster E1 than cluster E6 M protein, which may subsequently alter the remaining protein binding sites.

Competitive α-SV1 IgG binding studies analysing the effects of plasminogen and fibrinogen binding identified that, generally, binding of either host protein did not sterically obstruct α-SV1 IgG C-repeat domain recognition. Both plasminogen and fibrinogen have been shown to bind to the N-terminal hypervariable region of M protein (Berge *et al.* 1993; Ringdahl *et al.* 2000; Macheboeuf *et al.* 2011). It could be inferred that these respective binding motifs are positioned in a manner, distal enough to the C-repeat domain which does not allow for steric hindrance or induce conformational changes, inhibiting further M protein interactions. Previous studies analysing the effects of C-repeat directed antibodies against M5 protein, identified that binding was not inhibited in the presence of fibrinogen (Sandin *et al.* 2006). Two exceptions were observed within this study. α-SV1 IgG interaction with M1 protein was shown to obstruct but not inhibit binding of fibrinogen. The M1 exclusive S-domain is responsible for binding IgG, and is likely to be a viable low affinity secondary binding site for α-SV1 IgG. The structure of the B-repeat domain which resides just upstream of the S-domain has been shown to be critical in the proteins ability to bind fibrinogen (McNamara *et al.* 2008). Therefore it is possible that binding of α-SV1 IgG to the S-domain may affect the irregular coiled-coil arrangement of the B-repeat domain structure which is required for high affinity fibrinogen binding. The second exception was observed in M19 interaction with α-SV1 IgG which was slightly inhibited in the presence of excess fibrinogen. Although this result was unexpected, it could be inferred that potential conformational rearrangement as a result of fibrinogen binding, in conjunction with a relatively weak affinity for α-SV1 IgG could account for the slight quantitative decrease in α-SV1 IgG binding.

M protein C-repeat domains are responsible for binding albumin. The M protein albumin binding motif is situated in close proximity to the SV1 domain, with both regions sharing a

high level of sequence identity (Akesson *et al.* 1994; Retnoningrum *et al.* 1994; Bauer *et al.* 2012). Binding of albumin did not prevent α-SV1 IgG-binding to M proteins in this study. Additionally α-SV1 IgG was unable to affect albumin-M protein interactions. SV1 development is centred on J14 sequence variants found in all three C-repeat domains of M protein (Smeesters *et al.* 2008; Bauer *et al.* 2012). It was originally postulated that the high affinity interactions between albumin and M protein would inhibit α-SV1 IgG binding activity, especially since both proteins share analogous binding motifs which are in close proximity. Using GAS serotype M5, it has been previously shown that albumin is capable of blocking C-repeat directed antibody opsonisation (Sandin *et al.* 2006). Unlike N-terminal domains, C-terminal domains of M protein which are covalently linked to the GAS cell surface have been shown to be ridged in α-helical structure (Fischetti *et al.* 1988; McNamara *et al.* 2008). A possible explanation for the contradicting competitive binding data may lie in the fluidic structural nature of the M protein molecule where unlike M protein covalently bound at the GAS cell surface, recombinantly expressed M protein may be able to accommodate both ligands due to the reduction in conformational restriction.

In this study, we have developed a unique M protein based classification system categorising a large subset of circulating M-types in to 16 major cluster groups. This classification system highlights the potential for cross-reactivity in GAS vaccine development but also serves as a tool for the identification of potential binding motifs and M protein function (Sanderson-Smith *et al.* 2014). The results presented here clearly demonstrate the utility of a GAS classification system based on full length M protein sequence with the goal of facilitating future M protein functional studies, epidemiological surveillance and vaccine development.

4. Preferential acquisition and activation of plasminogen glycoform II by PAM positive Group A streptococcal isolates

A section of this work has been published in Biochemistry.

Reference:

David M. P. De Oliveira, Ruby H. P. Law, Diane Ly, Simon M. Cook, Adam J. Quek, Jason D. McArthur, James C. Whisstock, and Martina L. Sanderson-Smith. (2015). Preferential acquisition and activation of plasminogen glycoform II by PAM positive Group A streptococcal isolates. *Biochemistry*. *In press*.

4.1 Introduction

Multiple species of bacteria interact with the host protein plasminogen and increasing evidence suggests the ability to bind plasminogen is central to multiple stages of bacterial pathogenesis (Sanderson-Smith *et al.* 2008). Plasminogen is a single chain glycoprotein zymogen of the broad spectrum protease plasmin that circulates in plasma and extracellular fluids at an approximate concentration of 2 µM (Ponting *et al.* 1992; Andreasen *et al.* 1997). Regulation of the plasminogen activation system is essential for the maintenance of homeostatic function and the conversion of plasminogen into plasmin is a major regulatory target. The conformation of the plasminogen molecule greatly influences the regulation of plasminogen activation. Intramolecular binding between lysine residues and the lysine binding sites (LBS) of kringle (KR)-4, KR5 and the N-terminal Pan-apple (PAp) domain maintains circulating plasminogen in a compact and internally rigid 'closed' conformation that is highly resistant to activation (Ponting *et al.* 1992). Upon binding to specific cell surface receptors or ligands, such as fibrin, plasminogen adopts an 'open' conformation that becomes more susceptible to activation (Law *et al.* 2012). It is now well recognised that these interactions play a central role in modulating key steps in plasminogen activation by host and bacterial plasminogen activators (Law *et al.* 2012; Sanderson-Smith *et al.* 2012).

Plasminogen receptors and activators are expressed by a range of bacterial species including *Streptococcus pneumoniae, Staphylococcus aureus* and GAS (Bergmann *et al.* 2003; Sanderson-Smith *et al.* 2006; Sanderson-Smith *et al.* 2006; Kwiecinski *et al.* 2010; Sanderson-Smith *et al.* 2012). The interaction of GAS with plasminogen has been well characterised. GAS has evolved numerous highly specialised interactions with plasminogen/plasmin which are critical for virulence. Central to this are the plasminogen binding group A streptococcal M-proteins (PAM), and the secreted plasminogen activator, streptokinase (SK). A distinct 43 kDa PAM protein (M-type 53) belonging to *emm*-cluster D4 has been shown to bind plasminogen with high affinity ($K_D \sim 1$ nM) (Berge *et al.* 1993). Plasminogen is known to interact with ligands via LBS within specific KR structures. KR1, 4 and 5 demonstrate the highest affinity for lysine-containing ligands, with KR2 displaying the weakest affinity (Marti *et al.* 1997). The high affinity interaction between PAM and plasminogen is unique, in that PAM lacks a typical C-terminal lysine, a common feature of most but not all plasminogen receptors. Early reports suggest that Lys_{98} and Lys_{111} of PAM mediates binding to KR2 of plasminogen (Wistedt *et al.* 1995). More recent work has shown that binding of plasminogen to PAM is primarily mediated by positively charged Arg_{101}, Arg_{114}, His_{102} and His_{115} in the A1 and A2 repeat domains of PAM (Sanderson-Smith *et al.* 2006). In addition to the expression of PAM, GAS also secrete SK, a potent plasminogen activating protein. SK mediates plasminogen activation by the formation of a 1:1 (SK:plasminogen) complex, with substrate recognition occurring at KR5 of plasminogen (Tharp *et al.* 2009).

SK produced by GAS displays considerable genetic and phenotypic diversity. Phylogenetic studies of *ska* sequences from GAS isolates have revealed two main sequence clusters (cluster type-1 and -2) with cluster type-2 sequences being further subdivided (cluster type-2a and -2b) (Kalia *et al.* 2004; McArthur *et al.* 2008). Recent studies analysing these distinct

SK variants from specific GAS isolates suggest that the phenotypic differences displayed by SK may directly affect GAS pathogenic outcomes (Cook *et al.* 2012). Epidemiological studies have shown the type-2b *ska* lineage to be largely restricted to PAM positive GAS strains. In these isolates PAM and SK play a co-operative role in the acquisition and activation of plasminogen/plasmin. Specific ligand-binding induced conformational changes in plasminogen mediated by PAM and other host proteins such as fibrinogen are required for type-2b SK to form a functional activator complex with plasminogen. Furthermore the combination of PAM and fibrinogen is required for these type-2b SK variants to display resistance to inhibition by circulating α_2-antiplasmin, which is a hallmark of classical SK mediated activation process (Cook *et al.* 2013).

Glycosylation of plasminogen gives rise to two major species of the zymogen; GI-plasminogen and GII-plasminogen. GI-plasminogen possesses carbohydrate chains *N*-linked to Asn_{289} and *O*-linked to Thr_{346}, while GII-plasminogen contains a sole *O*-linked carbohydrate at Thr_{346}. GI-plasminogen and GII-plasminogen exist in humans at an approximate ratio of 2:3, respectively (Law *et al.* 2012). Differences in carbohydrate content between the two glycoforms of plasminogen lead to disparities in the kinetic and activation properties of the zymogen (Gonzalez-Grownow *et al.* 2002). The ability of highly specialised plasminogen-binding pathogens such as GAS to interact with different glycoforms of plasminogen and the downstream consequences of these interactions has not been investigated. Here, we examine the biochemical, structural and functional differences between the interaction of GI-plasminogen and GII-plasminogen with GAS proteins, PAM (M53 – cluster D4) and SK, with a view to resolve how binding and activation of these two different glycoforms may contribute to the interaction of GAS with the plasminogen activation system.

4.2 Results

4.2.1 Conformational change analysis between plasminogen GI and GII

To examine the effect of εACA on protein-domain unfolding in each plasminogen glycoform, GI-plasminogen and GII-plasminogen were monitored in the presence of εACA using the hydrophobic protein denaturation marker SYPRO® orange as previously described (Biggar *et al.* 2012). εACA is a synthetic lysine analogue which at high concentrations can inhibit the plasminogen activation system by blocking high affinity LBS (Griffin *et al.* 1978; Sun *et al.* 2002). At low concentrations, εACA has been shown to induce conformational changes in plasminogen, allowing the molecule to adopt an open, activation susceptible conformation (Law *et al.* 2012). Unfolding analysis identified that both GI-plasminogen and GII-plasminogen undergo structural change in the presence εACA. Dramatic conformational change was observed with both plasminogen glycoforms at a εACA concentration of 2.5 mM, indicating potential exposure of the typically compact KR2-5 and serine protease domains (Fig 4.1A). In the presence of εACA concentrations of 5 mM and above, hydrophobic exposure was consistently greater in GI-plasminogen than GII-plasminogen. Higher levels of hydrophobicity displayed by GI-plasminogen over GII-plasminogen may be the result of the mobile KR3 domain in GI-plasminogen which has been implicated in allowing the molecule to more readily adopt an 'open' conformation than its GII-plasminogen counterpart (Law *et al.* 2012). Subsequently, these data may also suggest that GII-plasminogen undergoes a smaller degree of εACA induced conformational change than GI-plasminogen (Fig. 4.1A).

4.2.2 Binding analysis of PAM for GI and GII plasminogen

In order to analyse the biochemical interactions of PAM with GI-plasminogen and GII-plasminogen, PAM_{NS13} was cloned and expressed as a truncated recombinant protein, lacking

the N-terminal signal sequence and C-terminal anchor domain (Sanderson-Smith *et al.* 2006). SPR analysis indicated that PAM has a comparative high affinity for closed forms of GI-plasminogen and GII-plasminogen with K_D of 27.4 ± 1.6 and 37.0 ± 6.6 nM respectively. Crystallisation studies of full length human plasminogen have revealed that only the LBS of KR1 is unprotected in closed plasminogen, however upon adopting an 'open' conformation KR5 moves away from the core of the plasminogen structure allowing it to be transiently exposed to interact with lysine-containing ligands (Law *et al.* 2012). To assess the impact of plasminogen conformation on the interaction with PAM, an open conformation was induced by the addition of lysine analogue εACA to the binding buffer. As shown in figure 4.1A, at εACA concentrations of 5 mM and above, plasminogen adopts a fully 'open' conformation, however large excesses of εACA will also block LBS and reduce interactions with ligands. In the presence of 5 mM εACA, both GI-plasminogen and GII-plasminogen maintained the ability to interact with PAM with a high affinity. However, binding of GI-plasminogen by PAM was completely inhibited in the presence of 100 mM εACA, whereas GII-plasminogen-PAM was still able to form in 400 mM εACA, although with reduced affinity (Fig 4.1B-C). Unlike GI-plasminogen, the affinity of PAM for GII-plasminogen was increased approximately 10-fold in the presence of 5 mM εACA (GI-plasminogen K_D = 33.2 ± 4.4 nM, $p > 0.05$; GII-Plasminogen K_D = 2.8 ± 0.6 nM, $p = 0.0001$), suggesting the conformational change to an 'open' conformation: 1) promotes a significantly higher affinity interaction between PAM and LBS of GII-plasminogen than occurs between LBS and εACA or 2) exposes a secondary PAM binding site outside of the LBS of GII-plasminogen which cannot be outcompeted with εACA.

To further investigate the higher affinity displayed by PAM for 'open' GII-plasminogen over GI-plasminogen, competitive binding analysis was undertaken via SPR, using both 'closed' and 'open' forms of each plasminogen glycoform. In the 'closed' conformation, neither

plasminogen glycoform was able to out-compete binding of the other to immobilised PAM. The low level binding displayed by each 'closed' plasminogen glycoform in the presence of immobilised PAM-GI-plasminogen or PAM-GII-plasminogen complex is likely to be attributed to the dissociation of respective PAM-plasminogen glycoforms, subsequently unmasking further binding sites. Interestingly, when in the 'open' conformation, induced by the addition of 5 mM εACA, GII-plasminogen was able to out-compete GI-plasminogen-PAM binding interactions. Furthermore, once 'open' conformation GII-plasminogen was bound to immobilised PAM, this interaction was unable to be disrupted by 'open' form GI-plasminogen (Fig. 1D). Collectively, these data suggest a higher affinity interaction between PAM and GII-plasminogen compared with GI-plasminogen when plasminogen is in the 'open' conformation.

4.2.3 Competitive binding analysis between PAM and GII-angiostatin in the presence of εACA

The KR2 LBS has been reported to mediate the interaction between PAM and plasminogen, however, our finding that 400 mM εACA is unable to fully disrupt the interaction between PAM and GII-plasminogen suggests the possibility of a secondary binding site within the plasminogen molecule. The first three KR domains of plasminogen (angiostatin) have been previously shown to interact with PAM via LBS (Berge *et al.* 1993; Cnudde *et al.* 2006). To determine if a novel binding domain exists outside of KR1-3, GII-angiostatin-PAM binding-kinetic analysis was undertaken with increasing concentrations of εACA (Fig. 2A). In contrast to our findings with full-length GII-plasminogen, the interaction of GII-angiostatin with immobilised PAM was completely inhibited in the presence of 5 mM εACA. This supports previous findings that LBS within KR1-3 are involved in the PAM-plasminogen interaction. However, as comparative concentrations of εACA were unable to inhibit

interactions of immobilised PAM with full length GI-plasminogen and GII-plasminogen, these data also suggests that a secondary site outside of KR1-3 may mediate the interaction of PAM with plasminogen. This interaction may represent a non-LBS mediated event, or alternatively a secondary higher affinity interaction within full-length plasminogen.

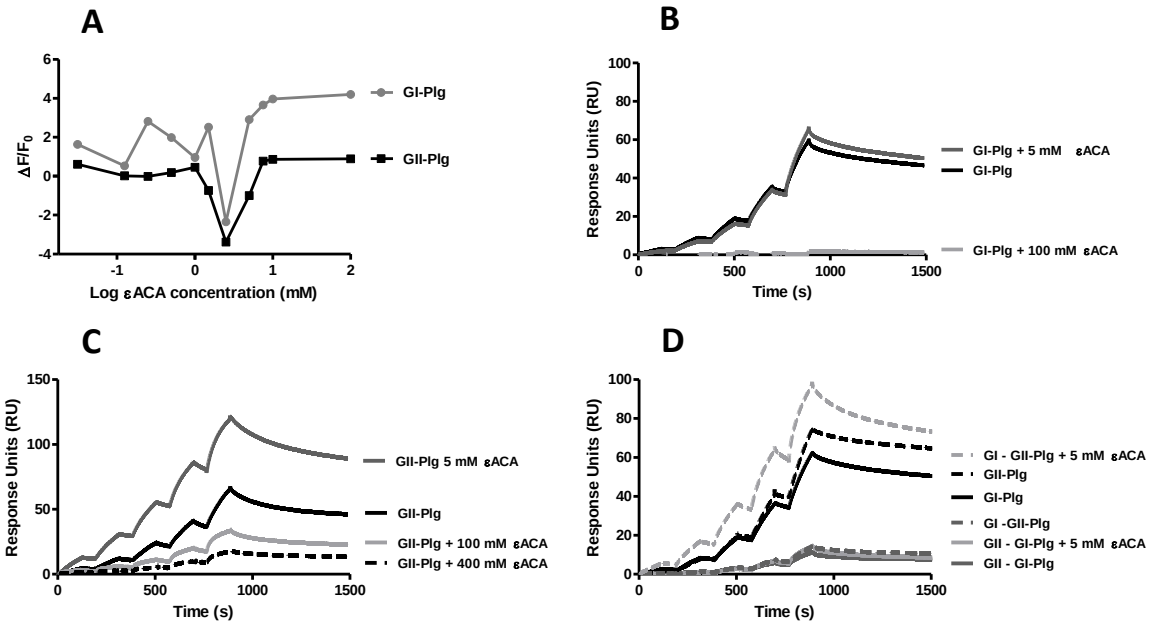


Figure 4.1: Effects of εACA on the structural conformation of GI/GII-plasminogen and associated affinity for PAM. (A) Conformational changes in GI-plasminogen and GII-plasminogen induced by the lysine analogue, εACA. Changes in structural conformation were undertaken using the conformational sensitive dye SYPRO orange. Analysis was undertaken at 37 ˚C using a LightCycler® 480 II Real Time PCR system, with excitation and emission set at 425 nm and 625 nm respectively. Fluorescent measurements were expressed as change in the initial fluorescence (F_{abs}-F_o)/F_o=$\Delta F/F_o$. Single cycle kinetic BIAcore sensograms analysing the affinity of immobilised PAM for plasminogen glycoform GI-plasminogen (B) and GII-plasminogen (C) in closed and open conformations. (D) Competitive binding analysis between GI-plasminogen and GII-plasminogen against immobilised PAM in the presence and absence of 5 mM εACA. Each run reflects five injections of plasminogen at 12.5, 25, 50, 100 and 200 nM. A dissociation period of 600 s followed the last injection. PAM-plasminogen interactions were analysed by non-linear fitting of the single cycle sensograms according to a 1:1 Langmuir binding model using Biacore T200 evaluation software (Biacore AB).

4.2.4 Competitive binding analysis between PAM-GI/GII-plasminogen in the presence of benzamadine

Plasminogen is known to adopt three different conformational states; the compact 'closed' conformation, the partially extended intermediate conformation and the fully extended 'open' conformation (Ponting *et al.* 1992; Marshall *et al.* 1994). The PAp domain serves as an important factor in the precise arrangement of 'closed' plasminogen by interacting with KR5, preventing the reversible conformational change to intermediate state plasminogen and/or extended 'open' form plasminogen (Ponting *et al.* 1992; Law *et al.* 2012). Prior studies analysing the interactions of PAM with the PAp-KR5 interface are limited (Wistedt *et al.* 1995; Cnudde *et al.* 2006). To determine whether PAM is able to bind outside of KR1-3 in the PAp domain, SPR analysis using Lys-plasminogen which lacks the PAp domain and exists in the intermediate conformation was undertaken both in the presence and absence of 10 mM εACA. εACA has been previously shown to result in Lys-plasminogen adopting the extended 'open' conformation (Marshall *et al.* 1994). In the absence of εACA, immobilised PAM was shown to bind Lys-plasminogen with high affinity (K_D = 0.44 ± 0.03 nM) indicating the PAp domain is not required for high affinity PAM-plasminogen interactions. In the presence of εACA, binding affinity decreased 9-fold (K_D = 3.83 ± 0.1 nM), but was still high affinity (Fig 4.2B). This decrease in affinity may be the result of blocking the PAM-K2 interaction, as seen for angiostatin. To further dismiss the PAp-KR5 interface as a factor in PAM-plasminogen recognition, benzamidine, a KR5 binding site competitor able to induce an intermediate conformation in plasminogen was employed (Varadi *et al.* 1981; Marshall *et al.* 1994). The addition of 5 mM benzamidine with and without equimolar concentrations of εACA was unable to inhibit binding in either plasminogen glycoform, but was shown to significantly increase the affinity ($p < 0.001$) of PAM-GII-plasminogen complexes which is similar to the effect seen in the presence of low concentrations of εACA. This is most likely

due to the higher affinity interaction between GII-plasminogen vs GI-plasminogen in the potential open conformation as a result of the secondary glycosylation site in GI-plasminogen (Fig 4.2C).

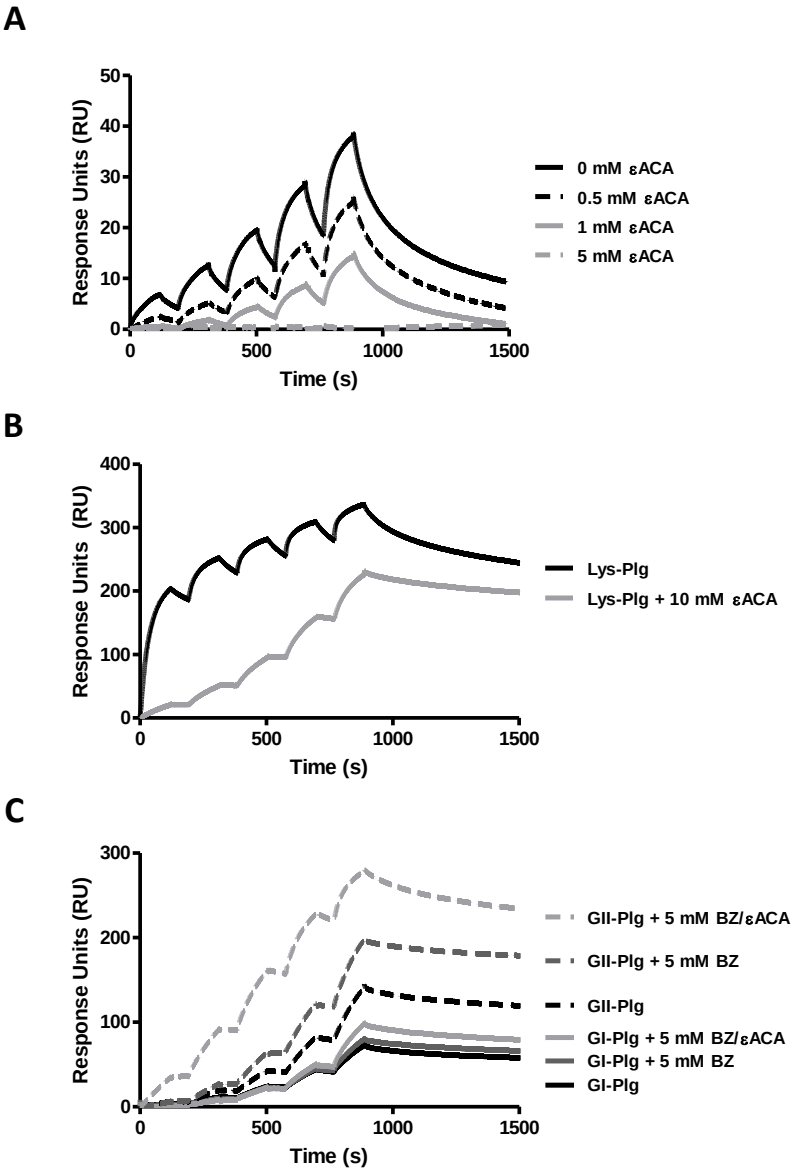


Figure 4.2: Isolating a potentially novel non-lysine dependant PAM binding site in GI/GII-plasminogen by SPR. (A) Affinity of immobilised PAM for GII-plasminogen angiostatin (KR1-3) in the presence of varying concentrations of εACA (0-5 mM). (B) Affinity of immobilised PAM for lys-plasminogen with and without 10 mM εACA. (C) Affinity of immobilised PAM for GI/GII-plasminogen in the presence of KR5 binding competitor, benzamidine (BZ), and εACA. Each run reflects five injections of analyte at 12.5, 25, 50, 100 and 200 nM. A dissociation period of 600 s followed the last injection. PAM-GI/GII-plasminogen interactions were analysed by non-linear fitting of the single cycle sensograms according to a 1:1 Langmuir binding model using Biacore T200 evaluation software (Biacore AB).

4.2.5 Selective plasminogen glycoform recruitment at the GAS cell surface

To determine if PAM preferentially sequesters a particular glycoform of plasminogen at the GAS cell surface, the PAM expressing GAS strain NS13 was incubated in human plasma. Eluted GAS cell surface bound protein (encompassing plasminogen) and positive control GI- and GII-plasminogen samples were then screened using α-human plasminogen Western blot analysis. The molecular weight of GI-plasminogen and GII-plasminogen correspond to 93 KDa and 89 KDa respectively. Plasminogen acquisition at the GAS cell surface was found to be GII-plasminogen specific (Fig. 4.3A), corresponding to an approximate molecular weight of 89 KDa. The presence of a 93 kDa band consistent with GI-plasminogen was absent from the elution sample.

4.2.6 Plasminogen activation at the GAS cell surface

For each glycoform, recruitment and activation at the GAS cell surface was qualitatively measured using the chromogenic substrate S-2251. S-2251 hydrolysis in this assay is given as a function of activated plasmin bound to the GAS cell surface. SK expression during the course of the assay and subsequent formation of SK-plasminogen/plasmin complexes at the cell surface is assumed to have a minimal contribution to pNA generation and is equal between isolates. This is due to the large molar excess of exogenous plasminogen added to the experiment versus endogenous SK and evidenced by subsequent saturation of plasmin activity between pre-incubation time points. PAM and type-2b SK expressing GAS strains NS13 and ALAB49 were incubated in the presence of equimolar levels of either GI-plasminogen or GII-plasminogen for 5 or 10 min periods. At both time points, plasminogen acquisition by GAS strains NS13 and ALAB49 was significantly higher in the presence of GII-plasminogen compared to GI-plasminogen ($p < 0.05$; Fig. 4.3B). This suggests that PAM

expressing GAS isolates preferentially recruit GII-plasminogen/plasmin over GI-plasminogen/plasmin at the cell surface.

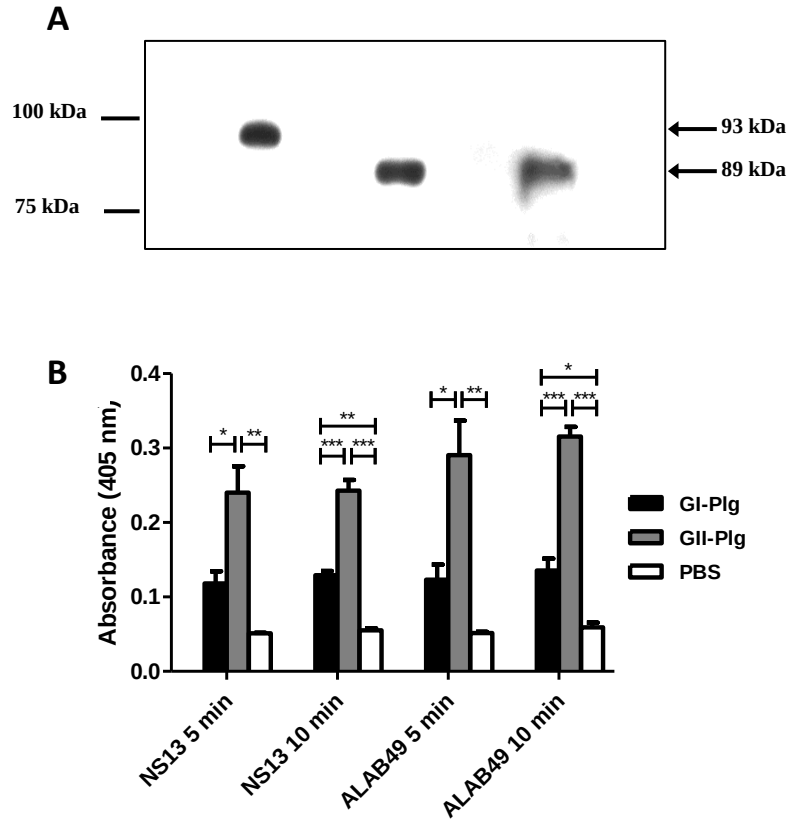


Figure 4.3: Selective plasminogen glycoform variant recruitment and plasmin glycoform acquisition at the GAS cell surface. (A) GAS isolate NS13 was incubated in human plasma and bound plasminogen was eluted using 100 mM Glycine-HCl (pH 2.0). Purified controls GI-plasminogen (lane 1) and, GII-plasminogen (lane 2) and eluted GAS cell surface plasminogen samples (lane 3) were visualized via an α-human plasminogen Western blot. (B) GAS isolates NS13 and ALAB49 were incubated with GI-plasminogen, GII-plasminogen (100 nM) or PBS for 5 or 10 min intervals at 37˚C. GAS cell surface plasmin acquisition was measured by the addition of S-2251 and change in absorbance was monitored at 405 nm at 37˚C.

4.2.7 SK mediated non-proteolytic active site generation and plasminogen activation

SK from the type-2b expressing GAS strain ALAB49 was examined for its ability to generate an active site in GI-plasminogen and GII-plasminogen. Non-proteolytic active site generation was examined using the fluorescent active site titrant MUGB as previously described (Cook *et al.* 2012). The rate of active site generation by type-2b SK was significantly higher for GII-plasminogen (21.0×10^{-3} $\Delta F.F_o^{-1}.min^{-1}$) compared to GI-plasminogen (5.4×10^{-3} $\Delta F.F_o^{-1}.min^{-1}$) ($p = 0.01$) (Fig. 4.4A). Type-2b SK mediated plasminogen activation in the presence of PAM was monitored using the chromogenic substrate, S-2251. Type-2b SK was able to generate plasmin activity in both GI-plasminogen/GII-plasminogen-PAM complexes. Plasmin activity generated by type-2b SK was observed to be significantly higher for GII-plasminogen (3.0 $mAbs_{405nm}.min^{-2}$) than GI-plasminogen containing samples (0.90 $mAbs_{405nm}.min^{-2}$) ($p < 0.001$) (Fig. 4.4B). Using GAS strain ALAB49, cell surface, type-2b SK dependant activation of plasminogen glycoform variants was shown to be significantly higher in GII-plasminogen (0.70 $mAbs_{405nm}.min^{-2}$) than GI-plasminogen (0.49 $mAbs_{405nm}.min^{-2}$) ($p < 0.01$). GAS strain ALABΔ*pam* was unable to activate either plasminogen glycoform confirming the importance of PAM in type-2b SK mediated plasminogen activation (Fig. 4.4C). Differences in the lag-time of plasmin generation displayed by both plasminogen glycoforms may be attributed to delay in active site generation, but could also be influenced by the differing affinities of PAM for each plasminogen glycoform.

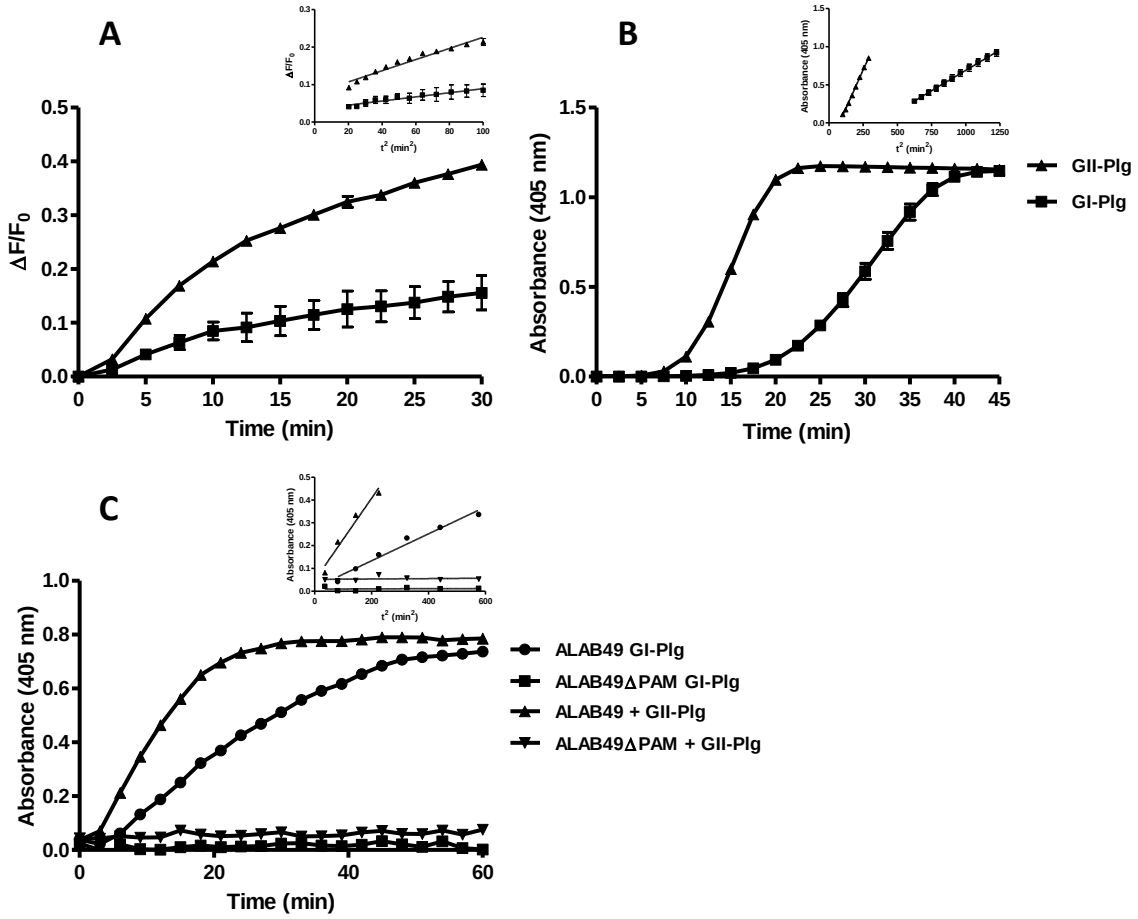


Figure 4.4: Non-proteolytic active site generation in plasminogen glycoforms by type-2b SK and the influence of variant plasminogen conformation on SK mediated plasminogen activation. (A) GI/GII-plasminogen (200 nM) and PAM (200 nM) were added to 1 µM MUGB in assay buffer (50 mM Tris-HCl, 100 mM NaCl, pH 7.4) and pre-incubated at 37 °C for 10 min. To initiate the reaction, SK was added to a final concentration of 400 nM in a total volume of 100 µL and the development of fluorescence was monitored continuously with excitation at 355 nm and emission at 460 nm. (B) SK (5 nM final concentration) was added to assay buffer (10 mM HEPES, 150 mM NaCl, 0.01% Tween-20, pH 7.4) containing variant GI/GII-plasminogen (500 nM), PAM (500 nM) and S-2251 (500 µM). (C) GAS cell surface plasminogen activation using GAS strains ALAB49 and ALABΔ*pam* containing GI/GII-plasminogen (200 nM) and exogenous SK (5 nM) carried out in assay buffer (10 mM HEPES, 150 mM NaCl, 0.01% Tween-20, pH 7.4). The generation of plasminogen activation activity was monitored at an absorbance of 405 nm at 37 °C. Insets show transformed plot of absorbance at 405 nm vs. t^2 with linear regression of the linearised region.

4.3 Discussion

The interaction of PAM and plasminogen combined with SK plays an important role in the pathogenicity of GAS. A key virulence mechanism of GAS is the bacterium's ability to sequester and activate plasminogen both extracellularly and at the GAS cell surface, aiding in the migration of the bacterium from cutaneous and mucosal surfaces to deeper tissue sites, resulting in severe invasive infection (Walker *et al.* 2005). Plasminogen has been previously shown to typically bind receptors via C-terminal lysine residues (Broder *et al.* 1991; Pancholi *et al.* 1998). In contrast, plasminogen binding to PAM has been attributed to a combination of internal lysine residues, and a series of positively charged arginine and histidine residues present within a1 and a2 repeat domains of PAM (Sanderson-Smith *et al.* 2006). In this study, we have shown that PAM preferentially binds GII-plasminogen over GI-plasminogen and propose that both plasminogen glycoforms contain secondary PAM binding sites outside of KR2 which are impacted by glycosylation at Asn_{289}. Furthermore, GII-plasminogen was also shown to have a higher rate of activation than GI-plasminogen by type-2b SK.

GI-plasminogen and GII-plasminogen circulate in plasma at an approximate ratio of 2:3. Biochemical and affinity analysis presented in this study demonstrated that GAS preferentially recruits and activates GII-plasminogen over GI-plasminogen. Analysis of the interaction between PAM and GI-plasminogen or GII-plasminogen, in combination with plasminogen recruitment at the cell surface, demonstrated that PAM preferentially binds GII-plasminogen, the most abundant glycoform. Earlier work investigating the dynamics of interaction between GI-plasminogen and GII-plasminogen at the site of injury in the thoracic aorta of rabbits found that GII-plasminogen accumulates approximately five times faster than GI-plasminogen (Hatton *et al.* 1999). This result is consistent with an earlier report identifying that GII-plasminogen is synthesized and released by the liver five times faster than GI-plasminogen in eukaryotes (Hatton *et al.* 1994). The differences in the abundance

between the two glycoforms of plasminogen in both serum and at the site of cutaneous injury may have served as a selective pressure for preferential recruitment of GII-plasminogen by PAM.

PAM is one of the few plasminogen receptors able to bind plasminogen via residues other than lysine (Sanderson-Smith *et al.* 2006). Crystallography studies of the first three KR domains of plasminogen (angiostatin) bound to VEK-30, a peptide derived from PAM, suggest that interactions between PAM and plasminogen occur exclusively at KR2 (Cnudde *et al.* 2006). Using GII-angiostatin, we have shown that binding of PAM to KR1-3 can be inhibited with the lysine analogue εACA, supporting the role of the KR2 LBS in the interaction between PAM and GII-plasminogen (Rios-Steiner *et al.* 2001). In contrast, whole molecule binding data using full length GI-plasminogen and GII-plasminogen demonstrated that interactions with PAM in α-helical coiled coil conformation could not be inhibited with comparative concentrations of εACA (Sanderson-Smith *et al.* 2006). Berge and Sjöbring (1993) previously identified that interaction between PAM and radiolabeled plasminogen could be outcompeted with 100 mM εACA. Although we cannot explain the discrepancies between these two data sets, observed differences may be a direct result of differences in binding methodology. Within this study, SPR experiments captured and presented PAM in an orientation which allows the N-terminal portion of the protein to extend away from the BIAcore sensor surface. By capturing the C-terminal of PAM, we were able to closely model the presentation of PAM at the GAS cell surface, thus allowing for a more physiologically relevant comparison of binding with each plasminogen glycoform.

The affinity between PAM and GII-plasminogen was significantly increased in the presence of εACA, an inducer of 'open' conformation plasminogen. Unlike GII-plasminogen, GI-plasminogen is glycosylated at Asn_{289} on KR3. Previous studies have shown that glycosylation at Asn_{289} results in reduced affinity of KR1 for fibrin and α_2-antiplasmin

(Lijnen *et al.* 1981; Thorsen *et al.* 1981). Conformational change to full extended 'open' form plasminogen has been suggested to be dependent on co-operative binding interactions between two ligands and plasminogen (Christensen *et al.* 1992). Christensen and Molgard proposed that co-operative binding involving two separate LBS where an initial low-affinity ligand interaction induces a rate-determining conformational change in plasminogen allowing subsequent interaction with a secondary ligand of significantly higher affinity (Christensen *et al.* 1992). The presence of this additional glycosylation site at the KR3/SP interface which also lies in the vicinity of KR2 in GI-plasminogen may act to sterically hinder but not inhibit a secondary PAM interaction, potentially limiting the degree of conformational change undertaken by GI-plasminogen. Although glycosylation at Asn_{289} lies in the vicinity of KR2, we cannot discern if KR2 function of GI-plasminogen is impaired by glycosylation at Asn_{289} in GI-plasminogen. PAM has been previously shown to bind KR1-3, but not truncated KR4 and KR5 fragments of plasminogen (Wistedt *et al.* 1998). Taken together with SPR data discounting the role of the PAp and KR5 LBS in PAM-plasminogen binding interactions, we suggest a secondary binding interaction may be occurring between PAM and the SP domain of plasminogen, or via a cryptic binding site which is more accessible when GII-plasminogen is in the open conformation. Recent superposition experiments have revealed that SK can bind to the SP domain of plasminogen in a way which avoids the KR domains (Law *et al.* 2012). Therefore, we cannot rule out that at the SP domain in both glycoforms of plasminogen, two independent binding sites may exist where both PAM and SK are able to co-operatively interact.

Epidemiological studies have associated type-2b SK with skin tropic, PAM positive GAS isolates (Kalia *et al.* 2004). Type-2b SK variants have been previously shown not to induce the formation of an active site in plasminogen, and have 25 fold less affinity for plasminogen compared to other SK polymorphic variants (Cook *et al.* 2012). Unlike GII-plasminogen, GI-

plasminogen contains a glycosylation site at Asn_{289} which does not allow for the formation of a hydrogen bond between KR3 and the Gly_{714} residue from the SP domain (Law *et al.* 2012). The absence of a hydrogen bond at the KR3-SP interface allows GI-plasminogen to more readily adopt an 'open' conformation (Molgaard *et al.* 1997; Law *et al.* 2012). Consequently, this has been shown to be responsible for higher rates of activation of GI-plasminogen by select host and bacterial plasminogen activators where by only in the presence of a lysine analogue, is the activation rate of GII-plasminogen enhanced to the same extent of GI-plasminogen (Takada *et al.* 1983). Only in the presence of either fibrinogen and/or PAM has type-2b SK been shown to create an active site in plasminogen (McArthur *et al.* 2008; Cook *et al.* 2013). In the presence of PAM, our results indicate that type-2b SK can form both an active site and mediate plasminogen activation significantly faster in GII-plasminogen than GI-plasminogen. This suggests that differences in glycosylation profile between the two glycoforms, particularly at Asn_{289} in GI-plasminogen, affect the potential of type-2b SK to generate active plasmin. It is well established that PAM and type-2b SK play a co-operative role in plasminogen activation, dependent on ligand-binding induced conformational change in plasminogen. As stipulated above, differences in binding interaction between each plasminogen glycoform and PAM may prompt significant disparities in conformational change between GI-plasminogen and GII-plasminogen, ultimately influencing activation in a manner which is unique to type-2b SK.

The acquisition of plasminogen/plasmin by PAM and SK is a fundamental step in GAS virulence. In this study we demonstrate for the first time that PAM expressing GAS selectively bind GII-plasminogen over GI-plasminogen with an affinity that is significantly higher than the affinity of GII-plasminogen for εACA. This suggests a role for either a non-LBS secondary interaction or the induction of a high affinity LBS dependent interaction in the open form of GII-plasminogen. Furthermore, we have identified significant differences in

type-2b SK mediated activation of plasminogen glycoforms. Due to similarities in M protein structure, and conservation of the highly predictive plasminogen binding motif, it can be suggested that *emm*-cluster D4 GAS isolates preferentially recruit GII-plasminogen. Type-2b SK is restricted to *emm*-cluster D4 GAS isolates and thus activation of plasminogen in other *emm*-cluster D4 GAS strains is expected to parallel the results shown in this study. Recognising the molecular mechanisms of how GAS sequesters plasminogen and how it is influenced by the distinct glycosylation profiles of plasminogen may significantly contribute to our understanding of GAS pathogenesis.

5. Characterisation of M protein mediated interactions with host glycan structures

5.1 Introduction

Glycans (or carbohydrates) are the most abundant and diverse biopolymers in nature, playing key roles in numerous cell recognition events, including those involved in the recognition of the immune system, and in the attachment of pathogenic organisms to host tissue (Sharon 1996; Crocker *et al.* 2007). The use of glycan array technology is becoming widely employed as a tool to efficiently identify carbohydrate binding proteins expressed both on human mammalian cells and select species of pathogenic bacteria (Tao *et al.* 2008; Day *et al.* 2009; Arndt *et al.* 2011). Several pathogenic bacterial species such as *Pseudomonas aeruginosa*, *Helicobacter pylori* (*H. pylori*), and *Campylobacter jejuni* have been shown to mediate adherence to host tissue via cell surface exposed glycan residues (Ilver *et al.* 1998; Mitchell *et al.* 2002; Day *et al.* 2009). To date, only a select subset of M protein-glycan interactions have been characterised. Glycosaminoglycans (GAG's) such as dermatan sulphate, highly sulphated fractions of heparin sulphate and heparin have been observed to facilitate M protein dependant binding of a diverse range of GAS serotypes to human epithelial cells (Frick *et al.* 2003). Specific surface glycosylation patterns are indicative features of certain cell types such as pharyngeal, skin and kidney as well as forming the basis of ABO blood group antigens (Arndt *et al.* 2011).

To date, M protein-glycan interactions and their potential role in GAS virulence have not been adequately explored. Glycan binding analysis of M protein representatives from the newly developed cluster based classification system outlined in chapter 3 may further reveal the importance of glycan interactions in different stages of GAS pathogenesis. Epidemiological studies have identified the existence of distinct tissue tropic GAS strains, which subsequently attach to cell types with markedly different glycosylation profile (Bessen *et al.* 2010; Arndt *et al.* 2011)**.** The, identification of novel M protein-glycan interactions may

provide insight in to the selective pressures which drive GAS tissue tropic pathogenesis. Our interest into better understanding the role of M protein-glycan interactions may help identify which glycans and/or glycosylated cell surface proteins are involved in contributing to the spectrum of diseases caused by GAS.

5.2 Results

5.2.1 Glycan microarray analysis

To further assess the application of the novel cluster based GAS classification system, M protein representatives from the 11 major cluster groups were screened for glycan binding using newly developed microarray technology (Day *et al.* 2009). From the 26 M proteins expressed in this study, 20 were chosen for glycan microarray analysis representing 20 different M-types from 11 major cluster groups. Positive binding was determined when binding fluorescence was above the mean background fluorescence (defined as greater than one fold increase of the mean background plus 3 standard deviations). Statistical analysis of the data was performed by a Student's t-test with a confidence level of 99.99% ($p \leq 0.0001$) and only glycans that met these criteria for three biologically independent samples (n = 12 glycans spot replicates) were interpreted as positive binding interactions. Complete M protein-glycan binding profiles are shown in Figure 5.1 and describe each respective glycan structure with designated glycan index number.

Carbohydrates carrying terminal galactose residues have been shown to be common adhesion receptors for an array of bacterial pathogens such as *Streptococcus suis*, *Shigella dysenteriae* and *Neisseria gonorrhoeae* (Lindberg *et al.* 1987; Calander *et al.* 1988; Haataja *et al.* 1993). Binding of terminating-galactose containing glycan structures was an observed characteristic of all *emm* clusters. Galactose covalently attached via α-linkage (1, Galα) was shown to bind

clusters E2-E6, A-C3, A-C4 and M proteins M57 and M19. Binding of mono-galactose moieties covalently attached via β-linkage (2, Galβ) were observed to exclusively bind clusters E2, E6, A-C4 and M protein M14 suggesting the importance of residue presentation in lectin recognition. Galactose structures Galβ1-3GlcNaAcβ (11) Galβ1-4Glcβ (16) were collectively bound by M protein representatives from clusters E1-E6, D1, A-C3 and non-clustered M proteins M14 and M57. Of particular note, binding to blood group antigen precursor lacto-*N*-tetraose (53, Galβ1-3GlcNAcβ1-3Galβ1-4Glc) was found to be an exclusive property of M proteins from clusters E3, A-C3, A-C4 and M proteins M57 and M19 suggesting a potential role in blood group antigen recognition. Cluster D4 was shown to exhibit little, to no conservation in binding of galactose-terminating structures between representative M proteins. From the glycans screened D4 M protein representatives were universally able to bind only GalNAcα1-3GalNAcβ (16), Galβ1-4GlcNAcβ1-6(Galβ1-3GlcNAcβ1-3)Galβ1-4Glc (56), GalNAcβ1-3Gal (65) and Galα1-4Galβ1-4GlcNAc (67) suggesting D4 M proteins are selectively restricted in function. M proteins which demonstrated the ability to bind a diverse range of galactose-terminating structures included non-clustered M proteins M57, M19 and A-C3 clustered M1.

Glucose terminating glycan structures are present in human secretions such as saliva and have been previously shown to be a fundamental source of carbon, involved in bacterial metabolism (Fuhrer *et al.* 2005; Wiesen *et al.* 2012). Recognition of glucose terminating glycan structures was shown to vary between M protein cluster groups. Unlike most clustered M protein, non-clustered M19 was observed to uniquely bind a wide range of terminal glucose glycan structures. M protein binding profiles of maltose (76, Glcα1-4Glcβ), maltotriose (82, $(Glc\alpha 1\text{-}4)_3\beta$), isomaltotriose (83, $(Glc\alpha 1\text{-}6)_3\beta$), maltotetraose (84, $(Glc\alpha 1\text{-}4)_4\beta$), isomaltotetraose (85, $(Glc\alpha 1\text{-}6)_4\beta$), isomaltopentaose (86, $(Glc\alpha 1\text{-}6)_5\beta$), and maltohexaose (87, $(Glc\alpha 1\text{-}6)_6\beta$) were shown to vary between cluster groups suggesting the

binding of maltose containing glycans by M protein is dependent on the number of repeat Glc residues and linkage.

*N´*acetyl glucosamine, a derivative of glucose, is an important biopolymer in bacterial cell wall structure, which when cross-linked with various oligopeptides, forms the bacterial peptidoglycan layer (Mesnage *et al.* 2014). Terminal *N´*acetyl glucosamine glycan binding patterns exhibited by clustered M protein were found to differ between M protein cluster groups. Of the glycans assessed, M protein cluster groups E4 and D1 were unable to bind any terminal *N´*acetyl glucosamine glycan structure. Conversely, M protein representative from clusters E6, A-C3, and non-clustered and M19 were observed to bind a wide range of *N´*acetyl glucosamine glycan structures, including common binding structures GlcNAcβ1-6(GlcNAcβ1-3)Galβ1-4GlcNAcβ (103), $(GlcNAc\beta1\text{-}4)_5\beta$ (104), and GlcNAcβ1-4MurNAc (110).

Mannose residues presented on the surface of bacteria are an important marker in host innate immunity, mediating non-opsonic phagocytic uptake by macrophages (Stahl *et al.* 1998). Prior studies analysing the pathogenic *S. pneumoniae* have suggested that association of the bacterium to Schwann cells is mediated via mannose specific lectins on the bacterial cell surface (Macedo-Ramos *et al.* 2014). With the exception of clade-X representative, cluster E6, binding of mannose containing glycan structures was over-represented by clade-Y M protein. Cluster groups E1-E4 were observed to bind little, to no mannose containing glycan structures. Clusters E6, D5, A-C3, A-C4 and M proteins M57 and M19 demonstrated an ability to bind a broad range of mannose containing structures including manβ (112), ManNAcβ (13) and Manα1-6(Manα1-3)Manα1-6(Manα1-3)Man (124) (124 did not bind D5). GlcNAcβ1-3Manβ was also observed to bind all M protein representatives belonging to clade-Y clusters D4, D5, A-C3, A-C4, A-C5 and M19 suggesting a conserved function among these M proteins.

Fucosylation of terminating galactose structures is an important step in blood group antigen synthesis (Yang *et al.* 1991). Prior studies have shown that fucose containing structures are utilised by a subset of bacterial species for host colonisation and metabolism (Stahl *et al.* 2011; Pacheco *et al.* 2012). Glycan array analysis of M proteins in this study identified binding to fucosylated structures by the majority of M protein clusters groups. Only M protein representatives from cluster E2 lacked conserved binding function to fucosylated glycans. Of particular note, the ability to bind lacto-*N*-fucopentaose I, (150, Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glc; LNFP I, H antigen I), the most commonly expressed blood group antigen was displayed by M protein clusters E3, D4, A-C3 and M proteins M57 and M19. With the exception of clusters D2, D4 and A-C4, binding of one or more complex multi-residue blood group related Lewis antigen containing structures (538-543) was observed between all major cluster groups and non-clustered M protein. For bacterial species such as *H. pylori*, phenotypic differences in host blood group antigen expression have been shown to be correlated with disease status (Jaff 2011). Here in this study, the varying patterns of A, B, O(H) and Lewis blood group antigen binding profiles displayed between all *emm*-cluster groups may provide avenues for future research investigating host blood group expression and GAS disease status.

N-acetylneuraminic acid (Neu5Ac) is the predominant sialic acid found in complex glycans attached to mammalian cell membranes. Prior studies analysing the ability of M6 protein to bind Neu5Ac identified that M protein interactions with Neu5Ac mediated association to host pharyngeal tissue (Ryan *et al.* 2001). M6 protein is situated in clade Y adjacent to non-clustered M protein and M19. Glycan array analysis identified that both M19 and M1 are able to recognise a wide range of Neu5Ac structures. Little to no binding of Neu5Ac containing structures was observed for clusters E1-E4, D4, and D5. Neu5Ac linkage has been shown to play a major role in M protein recognition. Neu5Ac α2-6 linked with galactose was suggested

to be a key sialylated linkage for M6 protein recognition (Ryan *et al.* 2001). Glycan array data identified that all Neu5Ac binding cluster groups were able to recognise both α2-3 and α2-6 linked Neu5Ac, demonstrating little to no linkage specificity. Furthermore, siaylation of galactose containing structures shown to bind various M proteins did not affect binding function. Lacto-*N*-tetraose and Lewis^{a} binders, M1 (A-C3) and M19 (non-clustered), demonstrated siaylated lacto-*N*-tetraose (215, Neu5Acα2-3Galβ1-3GlcNAcβ1-3Galβ1-4Glc) and siaylated Lewis^{a} (213, Neu5Acα2-3Galβ1-3(Fucα1-4)GlcNAc) binding function, highlighting the ability of select lectins to recognise conserved residues in distinct polysaccharides.

Sulphated oligosaccharides are a known adherence mediating factor for bacterial pathogen *Mycoplasma pneumoniae*, while also serving as a biomarker for the characterisation of cystic fibrosis and bronchitis (Krivan *et al.* 1989; Degroote *et al.* 2003; Xia *et al.* 2005). Glycan array analysis identified that clusters E2-E4 and D4 have little to no affinity for sulphated structures. M proteins representative of cluster E6 along with non-cluster M57 and M19 protein were shown to recognise a wide range of sulphated structures, irrespective of linkage or adjacent connecting oligosaccharides. Sulphation of galactose containing structures Galβ1-3GlcNaAcβ (237, 3-O-Su-Galβ1-3GalNAcα; 238, 6-O-Su-Galβ1-3GalNAcα) and Galβ1-4Glcβ (239, 3-O-Su-Galβ1-4Glcβ; 240, 6-O-Su-Galβ1-4Glcβ) was shown to inhibit binding to clusters E1-E4, D1, D5 and non-clustered M protein M14.

Carrageenans a family of sulphate containing carbohydrates commonly found in various types of marine seaweed have been shown to have antibacterial properties against *Enterobacter cloacae, S. aureus* and *Enterococcus faecalis* (Briones *et al.* 2014). Of the carrageenans screened, E1 clustered M protein and non-clustered M57 protein were observed to bind 5 of 6 carrageenan structures (273, 274, 276, 277 and 278) whereas M1 (A-C3) was able to bind all 6 (273-278). Remaining M protein cluster groups exhibited varying patterns

of carrageenan recognition whereby only cluster E2 demonstrated no conserved carrageenan binding function. Glycosaminoglycans (GAGs) belong to an assemblage of glycan structures which are expressed both on host cell surfaces and within the ECM of connective tissue. GAGs such as chondroitin sulphate, dermatan sulphate and heparin sulphate are typically comprised of repeating disaccharide residues, forming linear sulphated polysaccharides (Kjellen *et al.* 1991). Glycan array analysis revealed that binding of heparin (298) and heparin digested fragments (279-285) was a shared function among most M protein cluster groups. Only M proteins representative of clusters E2 and D5 did not exhibit conserved heparin and digested heparin fragment binding function. Sulphation of heparin (314) was observed to inhibit binding in cluster E1, D4, A-C4 and A-C5 M protein. Hyaluronic acid and related hyaluronan fragments are a common component of host epithelial and connective tissue. All M protein cluster groups were shown to bind hyaluronic acid (271), or various hyaluronan fragments (294-297; 302-313) whereby the number of repeating hyaluronan core units (GlcAβ1-3GlcNAcβ1-4) was observed to be a key factor in M protein recognition.

Terminal galactose glycan structures															
Index	**Glycan structure**	**E1**	**E2**	**E3**	**E4**	**E6**	**AC - M57**	**D1**	**AC - M14**	**AC - M19**	**D4**	**D5**	**A-C3**	**A-C4**	**A-C5**
1	Galα														
2	Galβ														
3	Galα1-2Galβ														
4	Galα1-3Galβ														
5	Galα1-3GalNAcβ														
6	Galα1-3GalNAcα														
7	Galα1-3GlcNAcβ														
8	Galα1-4GlcNAcβ														
9	Galα1-6Glcβ														
10	Galβ1-2Galβ														
11	Galβ1-3GlcNaAcβ														
12	Galβ1-3Galβ														
13	Galβ1-3GalNAcβ														
14	Galβ1-3GalNAcα														
15	Galβ1-4Glcβ														
16	GalNAcα1-3GalNAcβ														
17	GalNAcα1-3Galβ														
18	GalNAcα1-3GalNAcα														
19	GalNAcβ1-4GlcNAcβ														
20	Galβ1-2Galα1-4GlcNAcβ														
21	Galβ1-3Galβ1-4GlcNAcβ														
22	Galβ1-4GlcNAcβ1-3GalNAcα														
23	Galβ1-4GlcNAcβ1-6GalNAcα														
24	GalNAcβ1-4Galβ1-4Glcβ														

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
25	Galβ1-3GalNAcβ1-3Gal				■				■	■			■		
26	Galβ1-4Galβ1-4GlcNAc					■			■	■			■		
27	Galα1-3Galβ1-4GlcNAcβ1-3Galβ							■		■		■			■
28	Galα1-4GlcNAcβ1-3Galβ1-4GlcNAcβ				■	■	■	■	■	■		■	■		
29	Galβ1-3GlcNAcβ1-3Galβ1-3GlcNAcβ								■	■					
30	Galβ1-3GlcNAcα1-3Galβ1-4GlcNAcβ			■											
31	Galβ1-3GlcNAcβ1-3Galβ1-4GlcNAcβ		■	■			■			■			■		
32	Galβ1-3GlcNAcα1-6Galβ1-4GlcNAcβ			■						■			■		
33	Galβ1-3GlcNAcβ1-6Galβ1-4GlcNAcβ			■											■
34	Galβ1-3GalNAcβ1-4Galβ1-4Glcβ	■	■			■	■	■	■	■		■	■		
35	Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ		■		■	■	■	■	■	■		■	■		
36	Galβ1-4GlcNAcβ1-6Galβ1-4GlcNAcβ	■				■	■			■			■		■
37	Galβ1-4GlcNAcβ1-6(Galβ1-3)GalNAcα					■	■			■			■		
38	GalNacβ1-3GalαGalβ1-4Glcβ					■	■			■					
39	Galβ1-3GlcNAcβ1-3Galβ1-3GlcNAcβ						■			■		■			
40	Galα1-3Galβ1-4GlcNAcβ1-3Galβ1-4Glcβ						■			■			■		
41	Galβ1-4GlcNAcβ1-6(Galβ1-4GlcNAcβ1-3)GalNAcα					■	■			■					■
42	Galβ1-4GlcNAcβ1-3(GlcNAcβ1-6)Galβ1-4GlcNAcβ						■			■		■			
43	Galβ1-4GlcNAcβ1-6(GlcNAcβ1-3)Galβ1-4GlcNAcβ						■								
44	$(Galβ1\text{-}4GlcNAcβ1\text{-}3)_3$					■	■						■		
45	Galβ1-4GlcNAcβ1-6(Galβ1-4GlcNAcβ1-3)Galβ1-4GlcNAc					■				■			■		
46	Galβ1-3GalNAcβ1-3Galα1-4Galβ1-4Glcβ	■				■	■			■			■		

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
47	Galβ1-3GlcNAc														
48	Galβ1-4GlcNAc														
49	Galβ1-4Gal														
50	Galβ1-6GlcNAc														
51	Galβ1-3GlcNAc														
52	Galb1-3GalNAcβ1-4Galβ1-4Glc														
53	Galβ1-3GlcNAcβ1-3Galβ1-4Glc														
54	Galβ1-4GlcNAcβ1-3Galβ1-4Glc														
55	Galβ1-4GlcNAcβ1-6(Galβ1-4GlcNAcβ1-3)Galβ1-4Glc														
56	Galβ1-4GlcNAcβ1-6(Galβ1-3GlcNAcβ1-3)Galβ1-4Glc														
57	Galα1-4Galβ1-4Glc														
58	GalNAcα1-O-Ser														
59	Galβ1-3GalNAcα1-O-Ser														
60	Galα1-3Gal														
61	Galα1-3Galβ1-4GlcNAc														
62	Galα1-3Galβ1-4Glc														
63	Galα1-3Galβ1-4Galα1-3Gal														
64	Galβ1-6Gal														
65	GalNAcβ1-3Gal														
66	GalNAcβ1-4Gal														
67	Galα1-4Galβ1-4GlcNAc														
68	GalNAcα1-3Galβ1-4Glc														
69	Galβ1-3GlcNAcβ1-3Galβ1-4GlcNAcβ1-6(Galβ1-3GlcNAcβ1-3)Galβ1-4Glc														

Terminal glucose glycan structures															
Index	**Glycan structure**	**E1**	**E2**	**E3**	**E4**	**E6**	**AC - M57**	**D1**	**AC - M14**	**AC - M19**	**D4**	**D5**	**A-C3**	**A-C4**	**A-C5**
70	Glcα														
71	Glcβ														
72	GlcN(Gc)β														
73	$HOCH_2(HOCH)_4CH_2NH_2$														
74	GlcAα														
75	GlcAβ														
76	Glcα1-4Glcβ														
77	Glcβ1-4Glcβ														
78	Glcβ1-6Glcβ														
79	GlcAβ1-3GlcNAcβ														
80	GlcAβ1-3Galβ														
81	GlcAβ1-6Galβ														
82	$(Glc\alpha1\text{-}4)_3\beta$														
83	$(Glc\alpha1\text{-}6)_3\beta$														
84	$(Glc\alpha1\text{-}4)_4\beta$														
85	$(Glc\alpha1\text{-}6)_4\beta$														
86	$(Glc\alpha1\text{-}6)_5\beta$														
87	$(Glc\alpha1\text{-}6)_6\beta$														
Terminal *N´*acetyl glucosamine structures															
88	GalNAcα														
89	GlcNAcβ														
90	GlcNAcβ1-3GalNAcα														
91	GlcNAcβ1-4GlcNAcβ														
92	GlcNAcβ1-6GalNAcα														

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
93	GlcNAcβ1-4-[HOOC(CH_3)CH]-3-O-GlcNAcβ						■			■					
94	GlcNAcβ1-[HOOC(CH_3)CH]-3-O-GlcNAcβ-L-alanyl-D-i-glutaminyl-L-lysine	■				■			■	■					■
95	GlcNAcβ1-2Galβ1-3GalNAcα									■					
96	GlcNAcβ1-3Galβ1-3GalNAcα									■					
97	GlcNAcβ1-3Galβ1-4Glcβ					■				■		■			
98	GlcNAcβ1-3Galβ1-4GlcNAcβ					■				■					
99	GlcNAcβ1-4Galβ1-4GlcNAcβ														
100	GlcNAcβ1-6Galβ1-4GlcNAcβ									■					
101	GlcNAcβ1-6(Galβ1-3)GalNAcα						■		■	■			■		■
102	GlcNAcβ1-6(GlcNAcβ1-3)GalNAcα						■						■		
103	GlcNAcβ1-6(GlcNAcβ1-3)Galβ1-4GlcNAcβ			■		■	■		■	■			■		■
104	$(GlcNAcβ1-4)_5β$	■				■	■			■			■		
105	$(GlcNAcβ1-4)_6β$						■						■		
106	GlcNAcβ1-4GlcNAc	■					■			■			■		
107	GlcNAcβ1-4GlcNAcβ1-4GlcNAc									■			■		
108	GlcNAcβ1-4GlcNAcβ1-4GlcNAcβ1-4GlcNAc			■		■							■		
109	$(GlcNAcβ1-4GlcNAc)_3$ β1-4	■					■			■		■	■	■	■
110	GlcNAcβ1-4MurNAc					■			■	■	■		■	■	
Mannose containing structures															
111	Manα		■							■		■	■	■	
112	Manβ					■	■			■		■	■		■
113	ManNAcβ					■	■			■	■		■	■	
114	6-H_2PO_3Manα							■		■					

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
115	GlcNAcβ1-3Manβ														
116	Manβ1-4GlcNAcβ														
117	GlcNAcβ1-2Man														
118	GlcNAcβ1-2Manα1-6(GlcNAcβ1-2Manα1-3)Man														
119	Manα1-2Man														
120	Manα1-3Man														
121	Manα1-4Man														
122	Manα1-6Man														
123	Manα1-6(Manα1-3)Man														
124	Manα1-6(Manα1-3)Manα1-6(Manα1-3)Man														
Fucosylated structures															
125	Fucα														
126	Fucα1-3GlcNAcβ														
127	Fucα1-4GlcNAcβ														
128	Fucα1-2Galβ1-4GlcNAcβ														
129	Fucα1-2Galβ1-3GalNAcα														
130	Fucα1-2Galβ1-4Glcβ														
131	Galα1-3(Fucα1-2)Galβ														
132	Galα1-3(Fucα1-2)Galβ1-3GlcNAcβ														
133	Galα1-3(Fucα1-2)Galβ1-4GlcNAcβ														
134	Galα1-3(Fucα1-2)Galβ1-3GalNAcα														
135	Galα1-3(Fucα1-2)Galβ1-3GalNAcβ														
136	Galα1-3Galβ1-4(Fucα1-3)GlcNAcβ														
137	GalNAcα1-3(Fucα1-2)Galβ1-3GlcNAcβ														

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
138	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ														
139	GalNAcα1-6(Fucα1-2)Galβ1-3GalNAcα														
140	Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4GlcNAcβ														
141	Galα1-3(Galβ1-4)Fucα1-2(Fucα1-3)GlcNAcβ														
142	Fucα1-4(Fucα1-2Galβ1-3)GlcNAcβ1-3Galβ1-4Glcβ														
143	Fucα1-2Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4Glcβ														
144	Le^x1-6'(Le^c1-3')Lac														
145	LacNAc1-6'(Le^d1-3')Lac														
146	Le^x1-6'(6'SLN1-3')Lac														
147	Le^x1-6'(Le^d1-3')Lac														
148	Le^cLe^x1-6'(Le^c1-3')Lac														
149	Le^x1-6'(Le^b1-3')Lac														
150	Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glc														
151	Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4Glc														
152	Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4Glc														
153	Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4Glc														
154	Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4(Fucα1-3)Glc														
155	Fucα1-2Gal														
156	Fucα1-2Galβ1-4Glc														
157	Galβ1-4(Fucα1-3)Glc														
158	Galβ1-4(Fucα1-3)GlcNAc														
159	Galβ1-3(Fucα1-4)GlcNAc														
160	GalNAcα1-3(Fucα1-2)Gal														

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
161	Fucα1-2Galβ1-4(Fucα1-3)Glc														
162	Galβ1-3(Fucα1-2)Gal														
163	Fucα1-2Galβ1-4(Fucα1-3)GlcNAc														
164	Fucα1-2Galβ1-3GlcNAc														
165	Fucα1-2Galβ1-3(Fucα1-4)GlcNAc														
166	SO_3-3Galβ1-3(Fucα1-4)GlcNAc														
167	SO_3-3Galβ1-4(Fucα1-3)GlcNAc														
168	Galβ1-3GlcNAcβ1-3Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4Glc														
169	Galβ1-4(Fucα1-3)GlcNAcβ1-6(Galβ1-3GlcNAcβ1-3)Galβ1-4Glc														
170	Galβ1-4(Fucα1-3)GlcNAcβ1-6(Fucα1-2Galβ1-3GlcNAcβ1-3)Galβ1-4Glc														
171	Galβ1-4(Fucα1-3)GlcNAcβ1-6(Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ1-3)Galβ1-4Glc														
Siaylated structures															
172	Neu5Acα														
173	Neu5Acα														
174	Neu5Gcα														
175	9-NAc-Neu5Acα														
176	Neu5Acα2-3Galβ														
177	Neu5Acα2-6Galβ														
178	Neu5Acα2-3GalNAcα														
179	Neu5Acα2-6GalNAcα														
180	Neu5Gcα2-6GalNAcα														
181	Neu5Acα2-8Neu5Acα2														
182	Neu5Acα2-6GalNAcβ														

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
183	Neu5Gcα2-3Gal														
184	Neu5Acα2-6(Galα1-3)GalNAcα														
185	Neu5Acα2-6(Galβ1-3)GalNAcα														
186	Neu5Acα2-3Galβ1-3GalNAcα														
187	Neu5Acα2-3Galβ1-3GlcNAcβ														
188	Neu5Gcα2-3Galβ1-4GlcNAcβ														
189	Neu5Gcα2-6Galβ1-4GlcNAcβ														
190	9-NAc-Neu5Acα2-6Galβ1-4GlcNAcβ														
191	Neu5Acα2-3Galβ1-4-(6-O-Su)GlcNAcβ														
192	Neu5Acα2-3Galβ1-3-(6-O-Su)GalNAcβ														
193	Neu5Acα2-6Galβ1-4-(6-O-Su)GlcNAcβ														
194	Neu5Acα2-3-(6-O-Su)Galβ1-4GlcNAcβ														
195	$(Neu5Ac\alpha2\text{-}8)_3$-sp3														
196	Neu5Acα2-6Galβ1-3GlcNAc-sp3														
197	Neu5Acα2-6Galβ1-3(6-O-Su)GlcNAc														
198	Neu5Gcα2-3Galβ1-3GlcNAcβ-sp3														
199	GalNAcβ1-4(Neu5Acα2-3)Galβ1-4Glcβ														
200	Neu5Acα2-3Galβ1-4GlcNAcβ1-3Galβ														
201	Neu5Acα2-3Galβ1-4(Fucα1-3)6-O-Su-GlcNAcβ														
202	Neu5Acα2-3(6-O-Su)Galβ1-4(Fucα1-3)GlcNAcβ														
203	Neu5Acα2-6(Neu5Acα2-3Galβ1-3)GalNAcα														
204	Neu5Acα2-8Neu5Acα2-3Galβ1-4Glcβ														

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
205	Neu5Acα2-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ						■						■		
206	Neu5Acα2-3 Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ									■		■	■		
207	Neu5Acα2-6(Galβ1-3)GlcNAcβ1-3Galβ1-4Glcβ					■	■						■		
208	GalNAcβ1-4(Neu5Acα2-8Neu5Acα2-3)Galβ1-4Glc						■			■			■		
209	Neu5Acα2-8Neu5Acα2-8Neu5Acα2-3Galβ1-4Glc	■				■	■			■			■		
210	GalNAcβ1-4(Neu5Acα2-8)$_2$Neu5Acα2-3Galβ1-4Glc				■	■	■	■		■			■		
211	Neu5Acα2-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ					■	■	■		■					
212	Neu5Acα2-3Galβ1-4GlcNAcβ1-3Galβ1-4Glcβ-sp4	■					■			■			■		
213	Neu5Acα2-3Galβ1-3(Fucα1-4)GlcNAc	■				■	■	■	■	■			■	■	
214	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAc	■		■		■	■	■		■	■				■
215	Neu5Acα2-3Galβ1-3GlcNAcβ1-3Galβ1-4Glc	■			■	■	■	■		■			■	■	■
216	Galβ1-4(Fucα1-3)GlcNAcβ1-6(Neu5Acα2-6Galβ1-4GlcNAcβ1-3)Galb1-4Glc	■				■	■	■		■			■		
217	Neu5Acα2-3Galβ1-3(Neu5Acα2-6)GalNAc			■		■	■		■	■			■		
218	Neu5Acα2-6Galβ1-3GlcNAcβ1-3Galβ1-4(Fucα1-3)Glc						■					■	■	■	
219	Neu5Acα2-3Galβ1-4GlcNAc	■					■			■			■	■	■
220	Neu5Acα2-6Galβ1-4GlcNAc	■			■		■						■	■	■
221	Neu5Acα2-3Galβ1-3GlcNAcβ1-3Galβ1-4Glc	■		■	■	■	■	■		■			■	■	
222	Galβ1-3(Neu5Acα2-6)GlcNAcβ1-3Galβ1-4Glc	■					■	■	■	■		■	■		■
223	Neu5Acα2-6Galβ1-4GlcNAcβ1-3Galβ1-4Glc						■			■			■		■
224	Neu5Acα2-3Galβ1-3(Neu5Acα2-6)GlcNAcβ1-3Galβ1-4Glc	■		■	■	■	■	■	■	■			■		
225	Neu5Acα2-3Galβ1-4Glc	■		■		■	■			■			■		

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
226	Neu5Acα2-6Galβ1-4Glc														
227	(Neu5Acα2-8Neu5Ac)n (n<50)														
228	Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-6(Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-6)Manβ1-														
Sulphated structures															
229	3-O-Su-Galβ														
230	3-O-Su-GalNAcα														
231	6-O-Su-GlcNAcβ														
232	3-O-Su-GlcNAcβ														
233	Galβ1-3(6-O-Su)GlcNAcβ														
234	Galβ1-4(6-O-Su)Glcβ														
235	Galβ1-4(6-O-Su)GlcNAcβ														
236	GlcNAcβ1-4(6-O-Su)GlcNAcβ														
237	3-O-Su-Galβ1-3GalNAcα														
238	6-O-Su-Galβ1-3GalNAcα														
239	3-O-Su-Galβ1-4Glcβ														
240	6-O-Su-Galβ1-4Glcβ														
241	3-O-Su-Galβ1-3GlcNAcβ														
242	3-O-Su-Galβ1-4GlcNAcβ														
243	4-O-Su-Galβ1-4GlcNAcβ														
244	6-O-Su-Galβ1-3GlcNAcβ														
245	6-O-Su-Galβ1-4GlcNAcβ														
246	3-O-Su-Galβ1-4(6-O-Su)Glcβ														
247	3-O-Su-Galβ1-4(6-O-Su)GlcNAcβ														
248	6-O-Su-Galβ1-4(6-O-Su)Glcβ														

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
249	6-O-Su-Galβ1-3(6-O-Su)GlcNAcβ														
250	6-O-Su-Galβ1-4(6-O-Su)GlcNAcβ														
251	3,4-O-Su_2-Galβ1-4GlcNAcβ														
252	3,6-O-Su_2-Galβ1-4GlcNAcβ														
253	4,6-O-Su_2-Galβ1-4GlcNAcβ														
254	4,6-O-Su_2-Galβ1-4GlcNAcβ														
255	3,6-O-Su_2-Galβ1-4(6-O-Su)GlcNAcβ														
256	GalNAcβ1-4(6-O-Su)GlcNAcβ														
257	3-O-Su-GalNAcβ1-4GlcNAcβ														
258	6-O-Su-GalNAcβ1-4GlcNAcβ														
259	6-O-Su-GalNAcβ1-4-(3-O-Su)GlcNAcβ														
260	3-O-Su-GalNAcβ1-4(3-O-Su)-GlcNAcβ														
261	3,6-O-Su_2-GalNAcβ1-4GlcNAcβ														
262	4,6-O-Su_2-GalNAcβ1-4GlcNAcβ														
263	4,6-O-Su_2-GalNAcβ1-4-(3-O-Ac)GlcNAcβ														
264	4-O-Su-GalNAcβ1-4GlcNAcβ														
265	3,4-O-Su_2-Galβ1-4GlcNAcβ														
266	6-O-Su-GalNAcβ1-4(6-O-Su)GlcNAcβ														
267	Galβ1-4(6-O-Su)GlcNAcβ														
268	4-O-Su-GalNAcβ1-4GlcNAcβ														
269	3-O-SuGalβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ														
270	4-O-SuGalβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ														

Carrageen and glycosaminoglycan structures															
Index	**Glycan structure**	**E1**	**E2**	**E3**	**E4**	**E6**	**AC - M57**	**D1**	**AC - M14**	**AC - M19**	**D4**	**D5**	**A-C3**	**A-C4**	**A-C5**
271	(GlcAβ1-4GlcNAcβ1-3)$_8$-NH_2-ol														
272	(Sia2-6A-GN-M)$_2$-3,6-M-GN-GNβ-sp4														
273	Neocarratetraose-41, 3-di-*O*-sulphate (Na^+)														
274	Neocarratetraose-41-*O*-sulphate (Na^+)														
275	Neocarrahexaose-24,41, 3, 5-tetra-*O*-sulphate (Na^+)														
276	Neocarrahexaose-41, 3, 5-tri-*O*-sulphate (Na^+)														
277	Neocarraoctaose-41, 3, 5, 7-tetra-*O*-sulphate (Na^+)														
278	Neocarradecaose-41, 3, 5, 7, 9-penta-*O*-sulphate (Na^+)														
279	ΔUA-2S → GlcNS-6S Na_4 (I-S)														
280	ΔUA → GlucNS-6S Na_3 (II-S)														
281	ΔUA → 2S-GlcNS Na_3 (III-S)														
282	ΔUA → 2S-GlcNAc-6S Na_3 (I-A)														
283	ΔUA → GlcNAc-6S Na_2 (II-A)														
284	ΔUA → 2S-GlcNAc Na_2 (III-A)														
285	ΔUA → GlcNAc Na (IV-A)														
286	ΔUA → GalNAc-4S Na_2 (Δ Di-4S)														
287	ΔUA → GalNAc-6S Na2 (Δ Di-6S)														
288	ΔUA → GalNAc-4S,6S Na_3 (Δ Di-disE)														
289	ΔUA → 2S-GalNAc-4S Na_2 (Δ Di-disB)														
290	ΔUA → 2S-GalNAc-6S Na_3 (Δ Di-disD)														
291	ΔUA → 2S-GalNAc-4S-6S Na_4 (Δ Di-tisS)														
292	ΔUA → 2S-GalNAc-6S Na_2 (ΔDi-UA2S)														

Index	Glycan structure	E1	E2	E3	E4	E6	AC - M57	D1	AC - M14	AC - M19	D4	D5	A-C3	A-C4	A-C5
293	ΔUA → GlcNAc Na (Δ Di-HA)	■				■	■						■	■	
294	(GlcAβ1-3GlcNAcβ1-4)n (n=4)	■		■	■	■	■	■	■	■			■		■
295	(GlcAβ1-3GlcNAcβ1-4)n (n=8)	■		■	■	■	■	■	■	■			■	■	■
296	(GlcAβ1-3GlcNAcβ1-4)n (n=10)	■		■	■	■	■	■		■	■		■	■	■
297	(GlcAβ1-3GlcNAcβ1-4)n (n=12)	■	■	■	■	■	■	■	■	■		■	■		■
298	(GlcA/IdoAα/β1-4GlcNAcα1-4)n (n=200)	■				■	■	■		■	■		■	■	■
299	(GlcA/IdoAβ1-3(±4/6S)GalNAcβ1-4)n (n<250)	■		■	■	■	■	■					■	■	■
300	((±2S)GlcA/IdoAα/b1-3(±4S)GalNAcβ1-4)n (n<250)	■				■	■			■			■	■	
301	(GlcA/IdoAβ1-3(±6S)GalNAcβ1-4)n (n<250)	■		■		■	■		■	■	■		■	■	■
302	(GlcAβ1-3GlcNAcβ1-4)n (n=4)	■		■	■	■	■	■	■	■			■		■
303	(GlcAβ1-3GlcNAcβ1-4)n (n=6)	■		■	■	■	■	■	■	■		■	■		■
304	(GlcAβ1-3GlcNAcβ1-4)n (n=8)	■		■	■	■	■	■		■			■	■	■
305	(GlcAβ1-3GlcNAcβ1-4)n (n=10)	■				■	■	■	■	■			■		■
306	(GlcAβ1-3GlcNAcβ1-4)n (n=12)	■		■	■	■	■	■	■	■			■	■	■
307	(GlcAβ1-3GlcNAcβ1-4)n (n=14)	■	■	■	■	■	■	■	■	■			■		■
308	(GlcAβ1-3GlcNAcβ1-4)n (n=16)	■		■	■	■	■	■	■	■			■	■	■
309	(GlcAβ1-3Glc*N*Ac β1-4)n									■			■		
310	(GlcAβ1-3Glc*N*Ac β1-4)n		■				■			■					■
311	(GlcAβ1-3Glc*N*Ac β1-4)n	■	■		■	■	■	■		■			■		■
312	(GlcAβ1-3Glc*N*Ac β1-4)n	■	■				■			■					■
313	(GlcAβ1-3Glc*N*Ac β1-4)n	■	■			■	■	■		■			■		■
314	(GlcA/IdoAα/ IdoASα/β1-4Glc*N*Ac/GlcNS/Glc*N*Ac6Sα1-4)n				■	■	■	■		■			■		

Figure 5.1: Glycan binding profile of phylogenetically clustered M protein. M protein binding to a glycan was defined as a value $\geq$ 1 fold increase above mean background RFU. The mean background was calculated from the average RFU of all empty spots on the array plus three standard deviations. Glycan binding was examined using ProScanArray imaging software, ScanArray Express (PerkinElmer, USA) before data was exported to Microsoft Excel for further analysis. Statistical analysis of the data was performed by a Student's t-test with a confidence level of 99.99% ($p \leq 0.0001$) and only glycans that met these criteria for three biologically independent samples (n = 12 glycans spot replicates) were interpreted as positive binding interactions. Green blocks, indicative of glycan binding, are only presented if all M protein representatives within a cluster group were able to bind a respective glycan structures.

5.3 Discussion

Glycan array technology represents a powerful tool for the rapid identification and characterisation of novel lectin-glycoconjugate interactions. Microbial lectins, typically presented as elongated surface structures have been shown to interact with a wide range of glycan residues, subsequently defining the susceptibility of specific host tissue to bacterial pathogenesis. To date, the characterisation of M protein-glycan interactions have been restricted to large glycosaminoglycans such as heparin, heparin sulphate, and dermatan sulphate (Winters *et al.* 1993; Frick *et al.* 2003). Prior to the use of glycan array technology, characterisation of GAS-glycan interactions has served as a challenge. Here, we have systematically categorised conserved glycan binding function across 11 major *emm*-cluster groups, further evaluating the functionality of a full length M protein sequence based classification system. Coalesced with the *emm*-cluster system, glycan array technology serves as a predictive functional tool, with potential to aid in the development of future therapeutics for GAS disease.

Bacterial pathogens express a range of galactose specific lectins which allows for the interaction with various host tissues (Calander *et al.* 1988; Sharon 1996; Day *et al.* 2009). One of the most well characterised bacterial lectins capable of binding galactose is the *E. coli* surface expressed receptor, PapG the tip adhesin of Pap pilus, which recognises Galα1-4Gal residues present in the core of glycosphingolipids (Sung *et al.* 2001). Other bacterial lectins have also been shown to possess broad substrate specificity for terminal galactose structures. *C. jejuni*, a common pathogen of the digestive system expresses both galactose specific receptors as well as broad range galactose receptors, capable of binding both Galα1-3/4Gal and Galβ- linked glycans (Day *et al.* 2009). M protein representatives from all 11 *emm*-cluster groups displayed contrasting specificity for a range of terminal galactose glycan

structures, independent of linkage and sub-terminal carbohydrate class. Although the nature of each M protein galactose binding interaction awaits elucidation, binding data suggests that all M proteins in this study recognise galactose structures which form distinct sub-units within specific A, B, O and Lewis blood group antigen structures. It is well established that blood group antigen related structures are abundant on the surface of cells and within extracellular matrices. (Arndt *et al.* 2011; Everest-Dass *et al.* 2012). It is conceivable that blood group antigen recognition by GAS may contribute to M protein mediated GAS adhesion.

Glucose containing structures are a primary nutritional source for bacteria persisting within the host, specifically within secretions such as saliva and those pertaining to the human digestive system (Fuhrer *et al.* 2005; Eilam *et al.* 2014). With the exception of *emm*-cluster D4, all remaining clusters exhibited binding to several terminal glucose glycan structures. Derivatives of maltose (maltose, maltotriose, isomaltotriose, maltotetraose, isomaltotetraose, isomaltopentaose and maltohexaose) were observed to bind every M protein representative, except those from cluster E6. The varying patterns of recognition displayed by each M protein for maltose containing structures was shown to be influenced by the number of repeating glucose residues and corresponding linkage. The observed differences in specificity for various analogues of maltose may be indicative of certain selective pressures brought about through host tissue localisation and specificity. Glucose derivative, *N´*acetyl glucosamine, is a common constituent of the bacterial peptidoglycan layer (Mesnage *et al.* 2014). Binding recognition of *N´*acetyl glucosamine related structures was shown to vary between each *emm*-cluster whereby clusters E4 and D1 exhibited no conserved binding function. Apart from being an important biopolymer in the bacterial peptidoglycan layer, *N´*acetyl glucosamine has also been a target receptor for select bacterial *N´*acetyl glucosamine specific lectins. Gram negative *Eikenella corrodens* has been previously shown to mediate

adhesion to human oral epithelial cells via surface expressed *N´*acetyl glucosamine structures (Yamada *et al.* 2002). The capacity of each M protein representative within each *emm*-cluster group to bind various *N´*acetyl glucosamine structures may confer a broad function in mediating GAS adherence.

With the exception of clade-X representative, *emm*-cluster E6, binding of mannose related structures was highly conserved within clade-Y. Mannose binding lectin (MBL), a soluble host serum protein, is a well-established constituent of the innate immune response, functioning to recognise select mannose structures on the bacterial cell surface. MBL has been previously shown to exhibit high levels of binding to GAS regardless of clinical origin (Neth *et al.* 2000). The expansive recognition of mannose structures displayed primarily by clade-Y M protein may act to enhance MBL binding function, facilitating macrophage mediated phagocytic uptake subsequently inducing a pro-inflammatory cascade. Bacterial mannose binding lectins have also been shown to play a role in adherence to host tissue. The vast majority of bacterial mannose lectins are able to recognise a wide range of mannose structures and have been suggested to be critical in the initial adherence to host tissue (Thomas *et al.* 2002; Thomas *et al.* 2004; Day *et al.* 2009). The broad range mannose binding specificity displayed by the different M-types in this study may be reflective of the general mannose-specificity observed in other bacterial species, ultimately facilitating the initial adherence of bacteria to host tissue.

The acquisition of sialic acid based structures has been previously shown to be a fundamental element in bacterial metabolism and resistance of the host innate immune response (Severi *et al.* 2007). Bacterial species *Haemophilus influenzae* and *E. coli* K1 have been previously shown to utilise sialic acid as a carbon and nitrogen source via the enzymatic action of *N*-acetylneuraminate aldolase (Vimr *et al.* 2000; Severi *et al.* 2005). Furthermore, the ability to sequester and/or express siaylated glycan structures at the bacterial cell surface has been

hypothesised to allow multiple bacterial pathogens to circumvent and counteract the host immune response through molecular mimicry (Ram *et al.* 1998; Harvey *et al.* 2001; Vimr *et al.* 2002). Glycan array binding data revealed that only *emm*-cluster E2 M protein was unable to bind to any siaylated glycan structures. Interestingly, clade-X representatives E1, E3 and E4 only demonstrated specificity for siaylated Lewis antigens and galactose-containing blood group antigen precursor structures. The specificity shown by clade-X representatives for these respective structures may infer that binding recognition is independent of sialylation, and instead, a function of galactose or fucose residue expression. The binding of sialic acid-containing structures to clusters E6, D5, A-C3, A-C5 and non-clustered M protein, M57, M14 and M19 was found to be independent of galactose or fucose sub-terminal sugars, suggesting these M protein are capable of binding directly to sialic acid residues. Although respective siaylated binding functions are yet to be determined, sialyation of select galactose containing glycans has been previously shown to block galactose-specific lectin recognition (Day *et al.* 2009). Binding of lacto-*N*-biose (Galβ1-3GlcNAc) was found to be highly conserved across all clade-X *emm*-cluster groups. Terminal sialyation of lacto-*N*-biose (Neu5Acα2-3Galβ1-3GlcNAc) was shown to abrogate binding to all clade-X M protein representatives. These results may suggest that GAS serotypes belonging to clade-X preferentially interact with niches within the host which express reduced levels of siaylated glycoconjugates.

To date, investigation of M protein interaction with sulphated glycans have predominately focused on large sulphated glycosaminoglycans such as heparin sulphate and dermatan sulphate (Frick *et al.* 2003). Glycan array analysis of sulphated galactose and terminal *N*' acetyl glucosamine structures demonstrated that sulphation could alter the binding properties of select *emm*-clusters for specific glycans. Sulphation of lacto-*N*-biose and lactose was shown to inhibit binding by M protein belonging to *emm*-clusters E1-E4, D1, D5 and non-clustered M protein M14. Although the glycan binding domains of each M protein have not

yet been elucidated, the subsequent net negative change in charge upon addition of a sulphate group may prevent binding of lacto-*N*-biose or lactose to the respective binding domains of each M protein. Further analysis of sulphated glycans revealed that *emm*-cluster E6 along and non-cluster M57 and M19 protein can recognise a broad range of sulphate structures, irrespective of linkage and adjacent connecting oligosaccharides. The expansive specificity exhibited by these M proteins for sulphated structures suggest a defined function for sulphate-M protein binding interactions in GAS host-interaction. Such function has been demonstrated for *M. pneumoniae*, for which two distinct receptors have been reported to mediate binding of the bacterium to host cells: glycolipids containing sulphated galactose structures and glycoproteins containing sulphated forms of terminal *N*' acetyl glucosamine (Krivan *et al.* 1989). The broad range binding specificity exhibited by M proteins may be akin to binding interactions in *M. pneumoniae* which are fundamental to the process of colonisation.

GAS has been previously shown to mediate adhesion to cell surface and ECM connective tissue by M protein mediated interaction with glycosaminoglycans; heparin sulphate, dermatan sulphate and heparin (Bergey *et al.* 1988; Frick *et al.* 2003). Glycan array analysis identified that binding of heparin digested fragments, heparin sulphate, hyaluronic acid and hyaluronan fragments was a conserved function across multiple *emm*-cluster groups. Binding interactions between GAGs and M proteins, M1 and M5 have been previously attributed to electrostatic interactions between negatively charged sulphate group chains and positively charged N- and C-terminal regions of the M protein. (Frick *et al.* 2003). Since the C repeat domains are highly conserved among M protein, it is expected that binding of sulphated GAGs is a characteristic binding property of most if not all M protein. Furthermore, as GAG binding has also been partially attributed to the N-terminal domains of M1 and M5 protein, differences in binding affinities may be the result of amino acid sequence disparities in the A

and B repeat domains of respective *emm*-clustered M protein. Apart from being expressed on cell surfaces and within ECM connective tissue, soluble GAGs are abundant in wounds where dermatan sulphate is highly abundant (Penc *et al.* 1998). Therefore in such microenvironments, it could be assumed that GAS pathogenesis could be heightened via interaction with these GAG's. Moreover, GAS is also known to express the broad range cysteine protease, SpeB, which can cleave fragments of M protein from the bacterial surface (Anderson *et al.* 2014). Once released from the GAS cell surface, M protein could potentially modulate interaction with GAGs known to be involved in the binding of select cytokines and mediators of inflammation (Iozzo *et al.* 1996).

Fucosylated glycans and glycoproteins are frequently found in human saliva and are abundantly expressed on the surface of several types of cell epithelia (Arndt *et al.* 2011; Everest-Dass *et al.* 2012). Fucosylated glycans are also associated with A, B, H and Lewis blood group antigens which are commonly derived from the fucosylation of specific galactose containing precursors (Galβ1-3GlcNAcβ1-3Galβ1-4Glc; Galβ1-4GlcNAcβ1-3Galβ1-4Glc). Interestingly, *emm*-clusters which demonstrated conserved binding to various blood group antigen precursor or precursor related structures, also exhibited binding to a wide range of A, B, H and Lewis antigen structures. Thus, it is likely that galactose residues play a central role in in blood group antigen recognition. The ability of bacteria to bind fucosylated glycan structures has been previously associated with pathogenesis, with *P. aeruginosa* representing a well characterised example. The fucose-specific lectin of *P. aeruginosa* (PA-IIL) plays a central role in *Pseudomonas* infection by facilitating adherence and colonisation to heavily fucosylated mucosal layers (Rhim *et al.* 2001). M protein-mediated GAS interactions with fucosylated glycan structures may mediate two different outcomes in GAS colonisation, particularly in A-C pattern throat isolates. One potential function of binding fucosylated glycans is the promotion of GAS adherence to cell epithelia, possibly

contributing to non-invasive disease such as pharyngitis. Alternatively, binding interactions may serve as host protective measures against GAS adherence and colonisation. Since fucosylated glycoproteins are amply expressed in saliva, binding of fucosylated residues may result in the occlusion of other M domains utilised by GAS to adhere to host cells. In certain A-C pattern serotypes like M1 and M3 which do not express the major adherence protein Sfb1, the M protein serves as one of the primary adherence factors which may subsequently employ binding of fucose structures as a major means of adherence (Natanson *et al.* 1995). The finding that *emm*-clusters display unique patterns of binding to different blood group antigen structures may represent a selective pressure brought about through varying patterns of geographic serotype restriction. Different blood group antigens predominate in ethnically distinct populations which may subsequently reflect the divergent phenotypic binding capacity of M proteins from each *emm*-cluster (Farhud *et al.* 2013). Furthermore, these findings may also provide insight into niche adaptation between GAS isolates which typically colonise skin, versus those which colonise the throat, a site high in blood group antigen expression.

To date, glycan interactions and their roles in GAS pathogenesis have not been comprehensively explored. The results presented here demonstrate the diverse binding profiles of each *emm*-cluster group for an array of diverse glycans. Furthermore, the newly implemented M protein classification system based on full length M protein sequence has provided a means by which conserved binding function can be identified across different M-types sharing moderate levels of sequence identity. Although glycan array technology can be utilised to effectively identify novel bacterial glycan interactions, limitations exist in deciphering the role of each interaction. To determine the physiological relevance of these interactions, additional techniques must be employed.

6. Characterisation of M1 protein mediated interactions with blood group antigen related structures

6.1 Introduction

Human blood group antigen expression has long been shown to play a key role in modulating bacterial colonisation for the host, while also inhibiting yeast colonisation of the oral cavity (Everest-Dass *et al.* 2012; Makivuokko *et al.* 2012). The typical human blood group antigen classification system is based on three glycan structures, A, B and O (H) with precursor lacto-*N*-tetraose (LNT; Gal-β-1,3-GlcNAc-β-1,3-Gal-β-1,4-Glc) forming the backbone of all three antigens (Nishihara *et al.* 1999). Host gene expression of α-1,2-fuco-transferase enables synthesis of H antigen, which can be further processed by N-acetylgalactosaminyl-transferase or D-galactosyl-transferase to produce A and B antigens respectively (Yang *et al.* 1991; Persson *et al.* 2007) (Fig. 6.1). In addition to blood, A, B, H and related precursor structures are also expressed in other tissues and fluids such as saliva, tears, breast milk and oral epithelial cells (Slomiany *et al.* 1978; De Leoz *et al.* 2012; Everest-Dass *et al.* 2012). Expression of A, B, and H antigens on protein surfaces and secretion into extracellular fluid is dependent on expression of the *Se*(FUT2) gene encoding fucosyl-transferase-2 which enables secretion of the corresponding blood group antigens by mucin secreting goblet cells (Slomiany *et al.* 1978). Frequency population studies have shown that on average, 76% of people are blood group antigen secretors and 24% are non-secretors. Furthermore antigen H expressing individuals are found to have a significantly higher frequency of secretor status than non-antigen H expressing individuals (Jaff 2010).

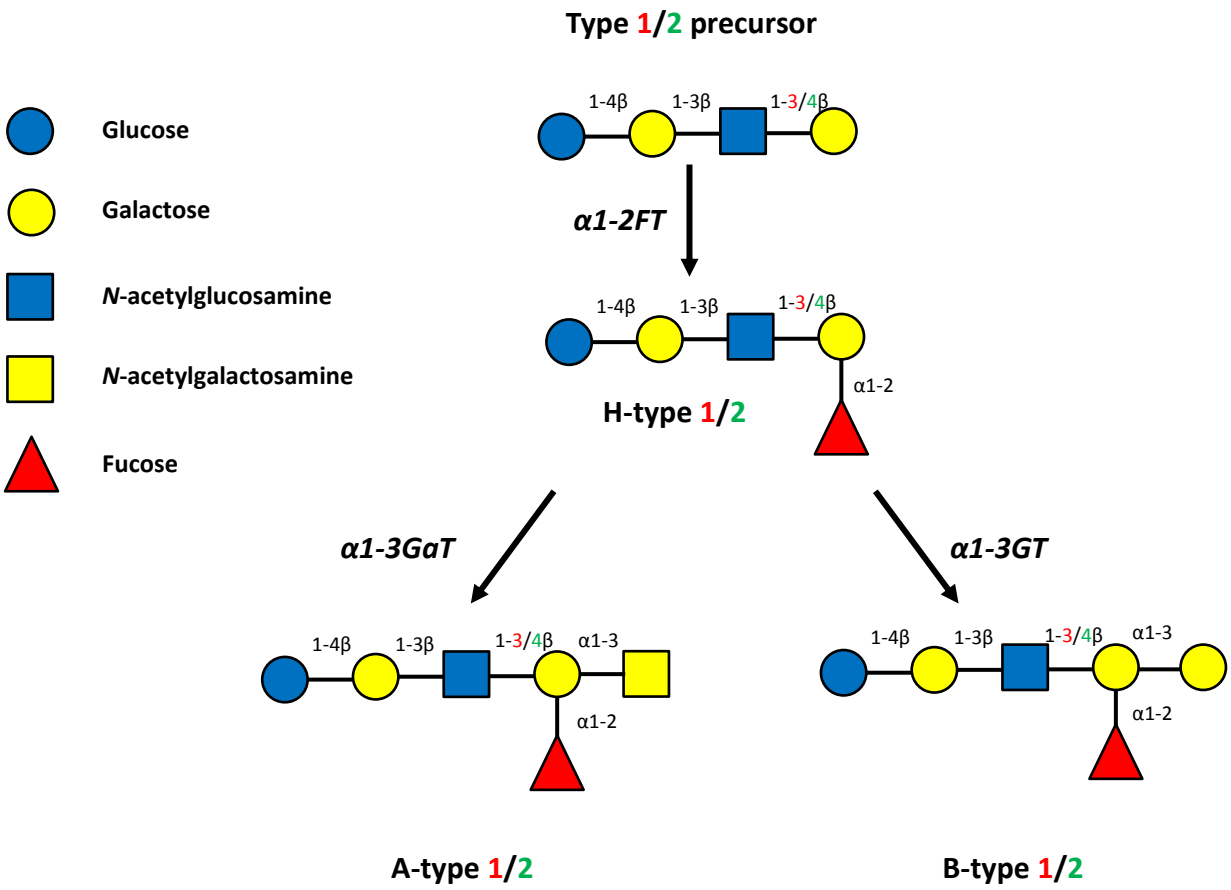


Figure 6.1: Synthesis of A, B and O(H) blood group determinants in humans. Only the two main precursor oligosaccharides are depicted (type 1, lacto-*N*-tetraose; type 2: lacto-*N*-neotetraose). Enzymes required for the synthesis of each blood group antigen are presented in italics. *α1-2FT*, α1-2 fucosyltransferase; *α1-3GaT*, α1-3 *N*-acetylgalactosyltransferase; *α1-3GT*, α1-3 galactosyltransferase.

Since the 1980's, the M1T1 GAS clone has disseminated globally accounting for a significant proportion of clinical isolates in the developed world. Although the M1T1 clone is able to mediate invasive disease, it is classically a throat tropic isolate frequently associated with uncomplicated infection such as pharyngitis (Stevens 1992; Johnson *et al.* 2002). Previous studies analysing the ability of M1T1 clones to colonise glycan rich sites such as pharyngeal mucosa have been limited to the study of M-protein-host protein interactions, with little focus on the potential role of expressed glycan residues. Blood group antigens and their precursors are heavily represented on oral epithelia and in mucosal fluid, suggesting a potential role in pharyngeal associated infection. As such, the interaction between M1 and these glycan structures was explored in more detail.

6.2 Results

6.2.1 5448 whole cell glycan microarray

The ability of M1 to bind glycan structures at the cell surface was assessed using M1T1 GAS strain 5448 and the previously described isogenic mutant, 5448ΔM1 (Walker *et al.* 2007). 5448 was found to bind 19 different glycan structures including blood group antigen precursor lacto-*N*-tetraose (53) and blood group antigen H type I (150) in an M1 dependant fashion. With deletion of M1 at the cell surface, 5448 was no longer able to recognise these structures on the array (Fig. 6.2). Furthermore, 5448 and 5448ΔM1 were also shown to bind multiple blood group A (137-139) and B (131, 368 and 392) antigen species, suggesting the presence of additional blood group antigen/fucose receptors at the GAS cell surface (appendix G). Due to the expression of lacto-*N*-tetraose and blood group H antigen related structures on human tissue and in mucosal fluid, interaction of these structures with M1 were further investigated (Slomiany *et al.* 1978; Everest-Dass *et al.* 2012).

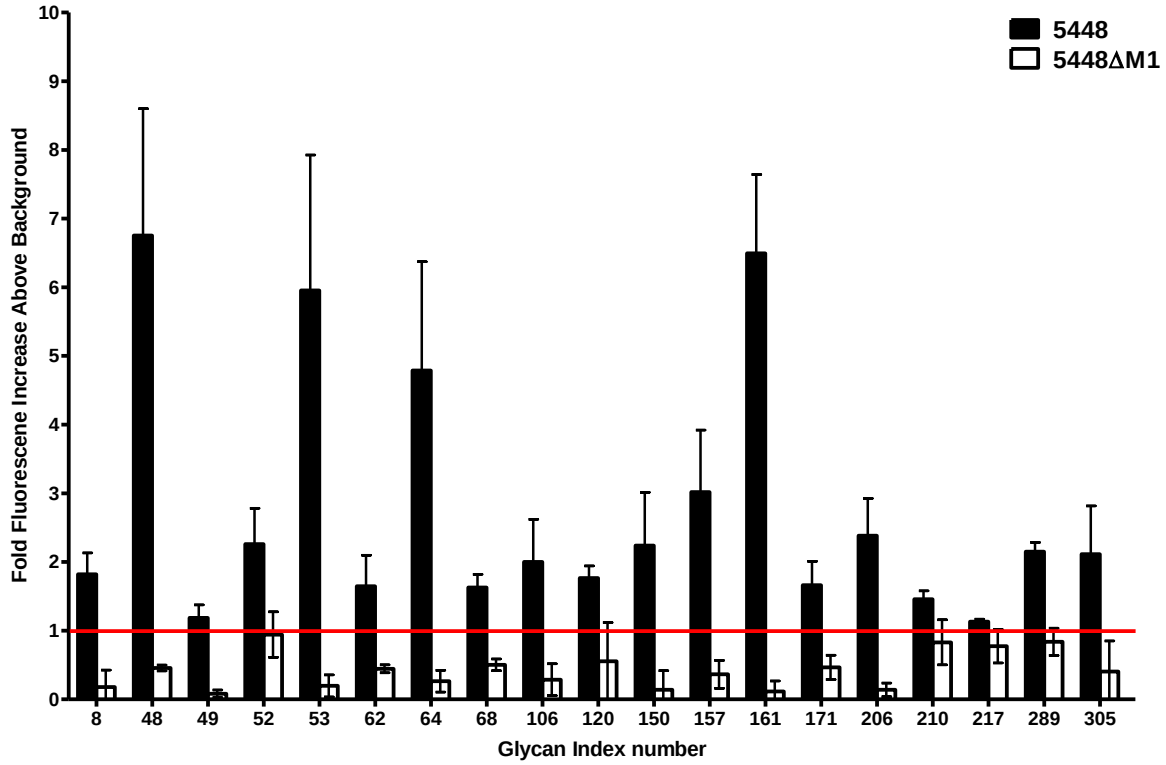


Figure 6.2: Glycan binding profile of M1T1 clone GAS strains 5448 and 5448ΔM1. Glycan binding was analysed using ProScanArray imaging software, ScanArray Express (PerkinElmer, USA) before data was exported to Microsoft Excel for further analysis. Bacterial binding to a glycan was defined as a value ≥ 1 fold increase above mean background RFU. The mean background was calculated from the average RFU of all empty spots on the array plus three standard deviations. Furthermore, statistical analysis of the data was performed by a Student's *t*-test with a confidence level of 99.99% ($p \leq 0.0001$) and only glycans that met these criteria for three biologically independent samples (n = 12 glycans spot replicates) were interpreted as positive binding interactions. Glycan index: 8, Galα1-4GlcNAcβ; 48, Galβ1-4GlcNAc; 49, Galβ1-4Gal; 52, Galb1-3GalNAcb1-4Galb1-4Glc; 53, Galβ1-3GlcNAcβ1-3Galβ1-4Glc; 62, Galα1-3Galβ1-4Glc; 64, Galβ1-6Gal; 68, Galα1-4Galβ1-4GlcNAc; 106, GlcNAcβ1-4GlcNAc; 120, Manα1-3Man; 150, Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glc; 157, Galβ1-4(Fucα1-3)Glc; 161, Fucα1-2Galβ1-4(Fucα1-3)Glc; 171, Galβ1-4(Fucα1-3)GlcNAcβ1-6(Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ1-3)Galβ1-4Glc; 206, Neu5Acα2-3 Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ; 210, GalNAcβ1-4(Neu5Acα2-8)$_2$Neu5Acα2-3Galβ1-4Glc; 217, Neu5Acα2-3Galβ1-3(Neu5Acα2-6)GalNAc; 289, ΔUA ® 2S-GalNAc-4S Na_2 (Δ Di-disB); 305, (GlcAβ1-3GlcNAcβ1-4)n (n=10).

6.2.2 M1-glycan affinity interactions via SPR

To further characterise the interaction of M1 with lacto-*N*-tetraose (index number, 53) and lacto-*N*-fucopentaose type I (index number, 150; blood group antigen H type I), single cycle kinetic SPR analysis was undertaken using a BIAcore T200 (GE Healthcare, Sweden). Data were fitted with a steady state affinity model. SPR binding analysis revealed that M1 was able to bind lacto-*N*-tetraose (Galβ1-3GlcNAcβ1-3Galβ1-4Glc) with an affinity of 127.3 ± 26.2 μM (Fig 6.3A). The binding of M1 to lacto-*N*-fucopentaose type I (Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glc) was observed to be > 400 fold higher (307.2 ± 168.9 nM) than exhibited by M1-lacto-*N*-tetraose suggesting a role of fucose residues in high affinity M1 glycan interactions (Fig. 6.3B).

6.2.3 Characterisation of glycan binding sites in M1

In order to assess whether a common glycan binding site exists within M1, competitive glycan array analysis was undertaken using both structurally related and structurally distinct glycan species known to interact with GAS strain 5448 in an M1 dependant manner. Lacto-*N*-tetraose, lacto-*N*-fucopentaose type I and β1-4galactosyl-galactose (Galβ1-4Gal) were selected due to Galβ1-4Gal residue conservation among these three glycan structures. α1-3-mannobiose (Manα1-3Man) was also included as it is structurally distinct from the aforementioned glycan structures. Competition binding analysis was undertaken by pre-incubating M1 protein with varying concentrations of either lacto-*N*-tetraose or lacto-*N*-fucopentaose type I (60 – 0.6 μM) and binding was assessed against the aforementioned set of glycans immobilised at molar amounts of 5 mmol and 500 nmol. Pre-incubation of M1 with either lacto-*N*-tetraose or lacto-*N*-fucopentaose type I was shown to out-compete M1 interaction with all 4 immobilised glycan structures (Fig. 6.4 A-D). Furthermore, when immobilised at a molar amount of 5 mmol, lacto-*N*-fucopentaose I binding to M1 was unable

to be inhibited by pre-incubation with lacto-*N*-tetraose confirming M1 has a higher affinity for lacto-*N*-fucopentaose type I compared with lacto-*N*-tetraose.

To better resolve which repeat domains of M1 are involved in the binding of lacto-*N*-tetraose and related blood group antigen structures, truncated M1 protein fragments with overlapping repeat domains were constructed as previously described (McNamara *et al.* 2008; Macheboeuf *et al.* 2011). The observed molecular weight of respective M1-fragments were shown to range from 27 kDa to 9 kDa and exhibited a far UV CD spectra characteristic of α-helical coiled coil protein (Fig. 6.5A & B). The ability of each respective glycan to bind fragmented M1 protein constructs was measured using SPR and modelled using steady state affinity fitting of the single cycle kinetic sensograms. Binding analysis revealed that with the exception of fragment M1-C, all M1 protein fragments were able to bind lacto-*N*-tetraose (K_D = 35.6 – 11.0 μM) and lacto-*N*-fucopentaose type I (K_D = 238.3 – 609.6 nM) (Table 6.1). These data in conjunction with aforementioned competition glycan array analysis suggests the existence of a multifunctional glycan binding motif, localised to the B repeat domains of M1.

To ascertain whether individual B repeat domains of M1 are capable of binding blood group related glycan structures, SPR binding analysis was undertaken using peptide amino acid sequences corresponding to the B1 and B2 repeat domains of M1. Binding analysis revealed that both B1 and B2 peptides were able to bind lacto-*N*-tetraose (M1-B1: K_D = 989 ± 153 nM; M1-B2: K_D = 1.49 ± 0.16 μM) and lacto-*N*-fucopentaose type I (M1-B1: K_D = 30.3 ± 3.6 nM; M1-B2: K_D = 50.0 ± 1.7 nM) with varying affinity (Fig 6.6A-B).

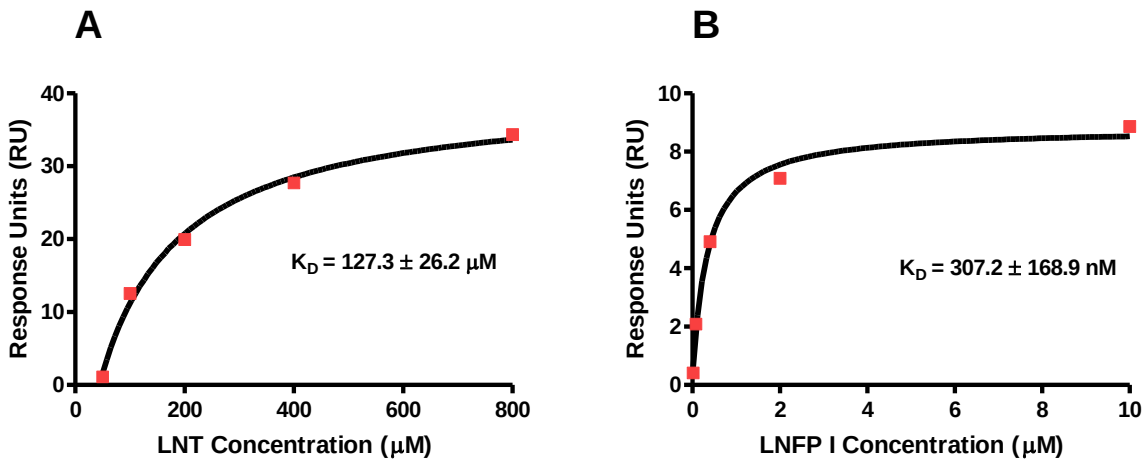


Figure 6.3: M1 binding of blood group antigen related structures. M1 binding to (A) lacto-*N*-tetraose (LNT; Galβ1-3GlcNAcβ1-3Galβ1-4Glc) and (B) lacto-N-fucopentaose type I (LNFP I; Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glc) was determined via surface plasmon resonance and were modelled using steady state affinity fitting of the single cycle kinetic sensograms.

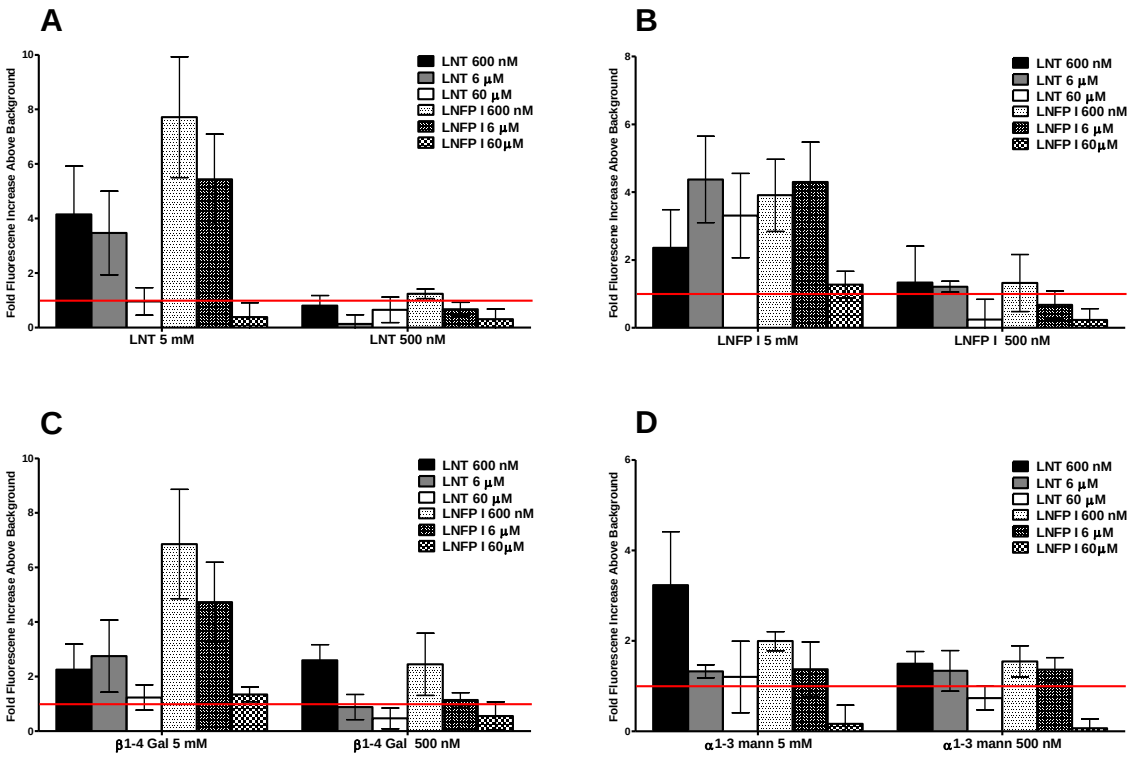


Figure 6.4: Competitive binding glycan array analysis. M1 binding to immobilised (A) lacto-*N*-tetraose (LNT; Galβ1-3GlcNAcβ1-3Galβ1-4Glc), (B) lacto-*N*-fucopentaose type I (LNFP I; Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glc), (C) β1-4galactosyl-galactose (Galβ1-4Gal), and (D) α1-3-mannobiose (Manα1-3Man) was assessed in the presence of varying competing concentrations of LNT and LNFP I. Glycan binding was analysed using ProScanArray imaging software, ScanArray Express (PerkinElmer, USA) before data was exported to Microsoft Excel for further analysis. M1 protein binding to a glycan was defined as a value $\geq$ 1 fold increase above mean background RFU. The mean background was calculated from the average RFU of all empty spots on the array plus three standard deviations. Furthermore, statistical analysis of the data was performed by a Student's t-test with a confidence level of 99.99% ($p \leq 0.0001$) and only glycans that met these criteria for three biologically independent samples (n = 12 glycans spot replicates) were interpreted as positive binding interactions.

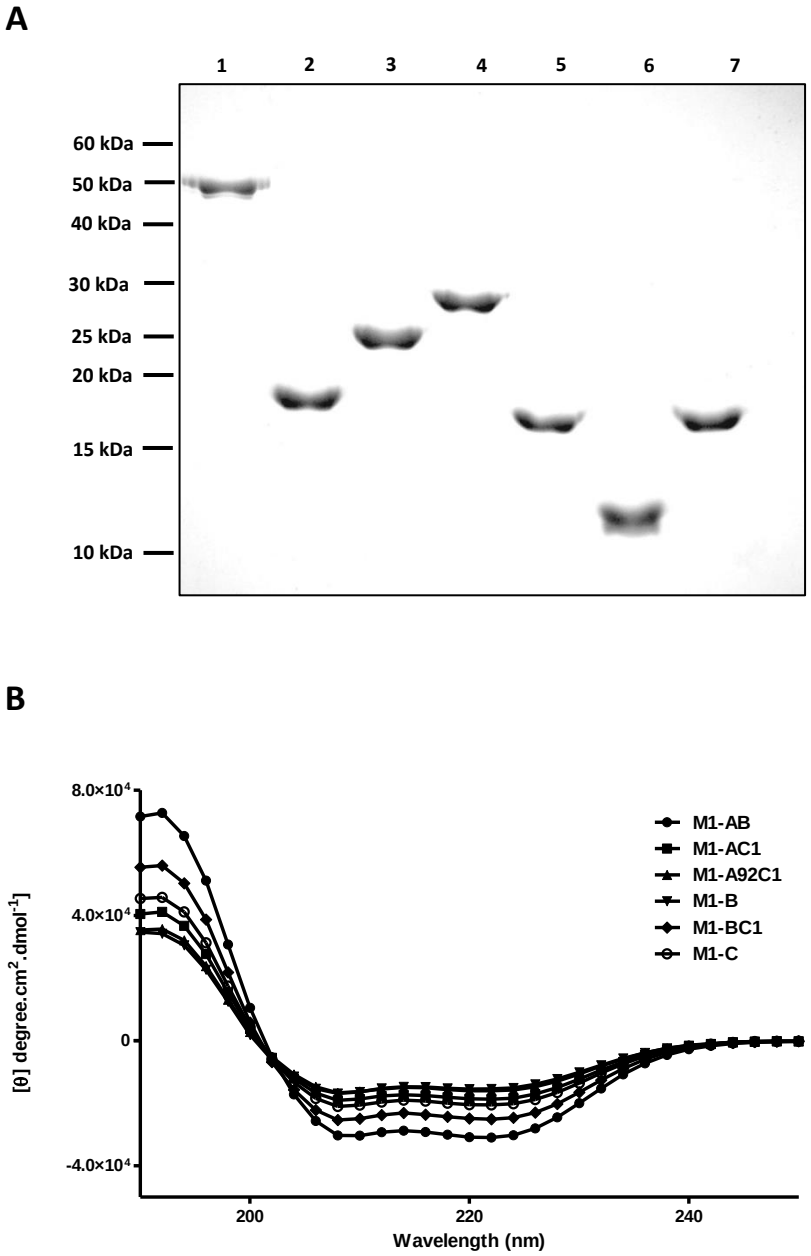


Figure 6.5: Recombinant M1 fragment protein analysis. (A) SDS-PAGE analysis of M1 (lane 1), M1-AB (lane 2), M1-A92C1 (lane 3), M1-AC1 (lane 4), M1-BC1 (lane 5), M1-B (lane 6) and M1-C (lane 7). Proteins were electrophoresed under non-reducing conditions on a 12% SDS-PAGE gel. Molecular weight markers are shown on the left and are given in kilodaltons (kDA). (B) CD spectra of M1 fragment proteins. All M1 fragment proteins exhibited CD emission spectra characteristic of α-helical coiled-coil proteins, displaying two minima at approximately 210 nm and 220 nm and a maximum peak at 190 nm.

Table 6.1: Summary glycan binding profile of recombinant M1 fragments. Interactions were analysed via SPR using single cycle kinetics fitted with a binding model of steady state affinity. *Nb, no binding.

	Lacto-N-tetraose	**Lacto-*N*-fucopentaose type I**
M1	127.3 ± 26.2 µM	307.2 ± 168.9 nM
M1-AB	21.3 ± 8.0 µM	238.3 ± 203.0 nM
M1-A92C1	35.6 ± 9.7 µM	484.0 ± 127.4 nM
M1-AC1	33.7 ± 22.9 µM	311.0 ± 223.1 nM
M1-B	11.0 ± 4.4 µM	609.6 ± 176.2 nM
M1-BC1	15.6 ± 11.5 µM	511.5 ± 84.2 nM
M1-C	Nb*	Nb*

To potentially refine a glycan binding motif within the B repeat domains of M1, the first 12 amino acids of the B1 repeat domain (M1-B1 motif: LEKELEEKKEAL$_{HHHHHHHHHH}$), sharing 92% sequence identity with B2 was constructed. Binding analysis via SPR demonstrated that the M1-B1 motif peptide exhibited binding to lacto-*N*-tetraose (K_D = 3.8 ± 0.72 µM) and lacto-*N*-fucopentaose type I (K_D = 81.5 ± 46.3 nM). To further validate the glycan binding function of the M1-B1 motif sequence, a negative control scrambled adaptation of the M1-B1 motif sequence (M1-B1 motif*; KLLKEKEAELEE$_{HHHHHHHHHH}$) was constructed. Binding analysis demonstrated that the M1-B1 motif* was not able to recognise lacto-*N*-tetraose or lacto-*N*-fucopentaose type I. These data suggest the importance of this 12 amino acid sequence in the recognition of specific blood group related antigen structures (Fig. 6.6A-B). Secondary structure analysis via far UV CD showed that M1-B1, M1-B2, M1-B1 motif and M1-B1 motif* peptides exhibited secondary structures consistent with a monomeric random coil (Ogawa *et al.* 2006) (Fig. 6.6C). Thus, binding of lacto-*N*-tetraose and lacto-*N*-fucopentaose type I may not be dependent on the secondary coiled coil structure of M1, but instead dependant on the primary amino acid sequence of the B-repeat domain. Furthermore, the observed increase in binding affinity displayed by the M1-B repeat peptides for lacto-*N*-

tetraose and lacto-*N*-fucopentaose type I may suggest that typical 'core' forming amino acids are able to interact with these respective glycan structures once exposed.

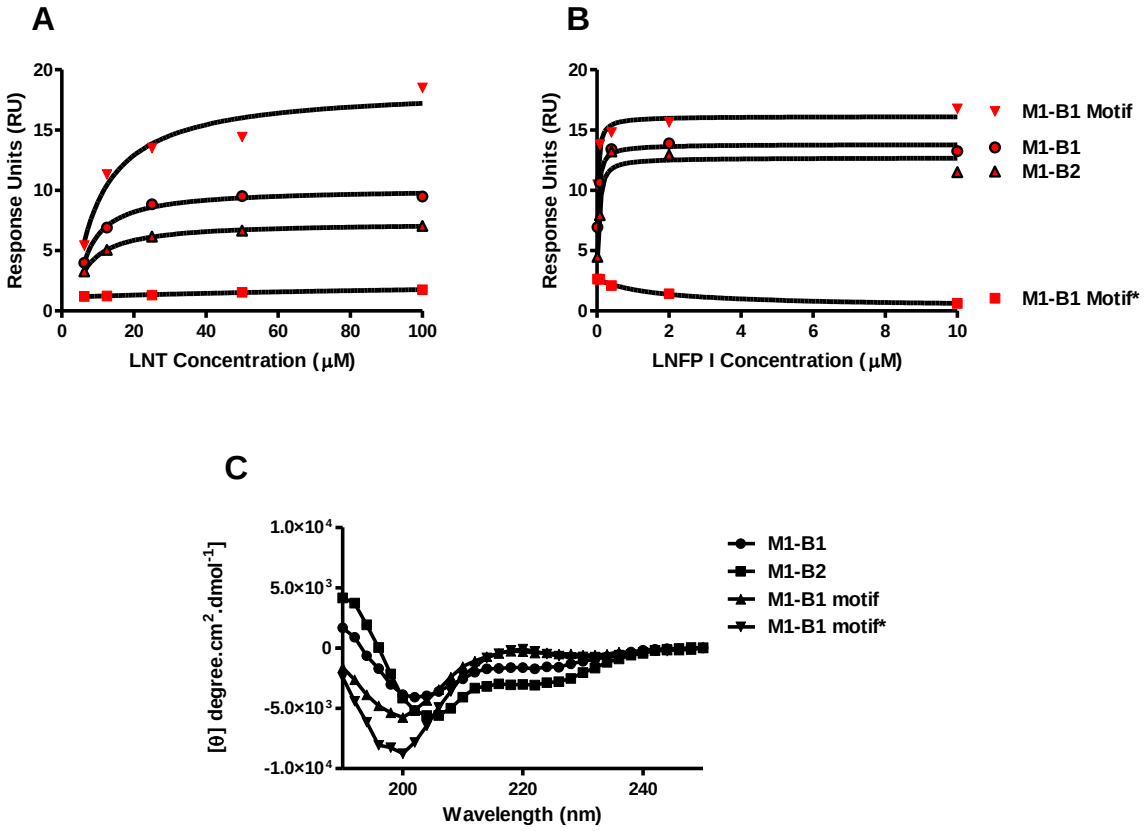


Figure 6.6: SPR analysis of M1 peptide glycan interactions. Binding analysis of M1 derived peptides M1-B1, M1-B2, M1-B1 motif, and M1-B1 motif scrambled were screened for binding against (A) lacto-*N*-tetraose (LNT; Galβ1-3GlcNAcβ1-3Galβ1-4Glc) and (B) lacto-*N*-fucopentaose type I (LNFP I; Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glc). Binding was determined via surface plasmon resonance and were modelled using steady state affinity fitting of the single cycle kinetic sensograms. (C) CD spectra of M1 derived peptides. Peptides, M1-B1, M1-B2 and M1-B1 motif exhibited CD emission spectra consistent with that of a monomeric random coil (Ogawa *et al.* 2006).

6.2.3 Functional characterisation of M1 mediated glycan interactions

Pharyngitis is frequently attributed to a subset of M types including A-C3-5 clustered serotypes, M12, M3 and M1 (Walker *et al.* 2014). The oral mucosa and epithelium of pharyngeal tissue is a glycan rich environment expressing a vast multitude of *O*- and *N*-linked

carbohydrate structures (Ryan *et al.* 2001; Everest-Dass *et al.* 2012). To examine whether specific M1-dependant glycan interactions facilitate the association of M1 serotype 5448 to pharyngeal tissue, association assays were undertaken using Detroit 562 pharyngeal cells in the presence of exogenous lacto-*N*-tetraose and lacto-*N*-fucopentaose type I.

Following pre-incubation with lacto-*N*-tetraose (900 µM), 5448 WT was shown to associate significantly more with Detroit-562 cells compared to the PBS control ($P < 0.01$). Association of 5448 WT with Detroit-562 cells was significantly attenuated in the absence of lacto-*N*-tetraose. No significant differences in pharyngeal epithelial cell association were observed by 5448ΔM1 when pre-incubated with either lacto-*N*-tetraose or PBS (Fig 6.7A). Furthermore, lacto-*N*-tetraose was shown to promote association of 5448 WT in a concentration dependant manner (Fig. 6.7B). M90, expressed on the surface of GAS strain NS730 was shown not to bind lacto-*N*-tetraose via glycan array analysis (see chapter 5). The ability of lacto-*N*-tetraose to mediate increased NS730 association to Detroit-562 cells was therefore assessed to confirm the specificity of the observed increase in M1 binding. No significant difference was observed in the association levels of NS730 to Detroit-562 in the presence and absence of lacto-*N*-tetraose (Fig 6.7C). Numerous studies support the existence of throat-tropic GAS isolates, restricted primarily to *emm* pattern A-C isolates (Bessen *et al.* 2010; Walker *et al.* 2014). To investigate whether the interaction of M1 and lacto-*N*-tetraose plays a role in the association of GAS to other forms of host tissue, association assays examining the binding of 5448 to HaCaT keratinocyte cells were employed in the presence of lacto-*N*-tetraose. Pre-incubation with lacto-*N*-tetraose did not alter the binding of 5448 WT or 5448ΔM1 to HaCat cells (Fig. 6.7D). Due to the high affinity of M1 for lacto-*N*-fucopentaose I, and the physiological significance of the blood group antigen, association of 5448 WT to Detroit-562 cells was examined in the presence and absence of exogenous lacto-*N*-fucopentaose type I. Upon pre-incubation of 5448 WT with 900 µM lacto-*N*-fucopentaose

type I, 5448 WT demonstrated a slight but non-significant increase in association to Detroit-562 cells. Association of 5448ΔM1 exhibited no increase in the presence of lacto-*N*-fucopentaose type I (Fig. 6.8).

6.2.4 Glycosylation profiles of human buccal epithelial cells and their potential role in GAS colonisation

Terminal glycan epitope A, B, and H antigen structures have previously been shown to be expressed in high abundance at the surface of buccal epithelial cells. These cells are part of the oral mucosal surface, and represent a physiologically relevant cell type for investigating GAS-host cell interaction (Everest-Dass *et al.* 2012). The current study demonstrates that GAS strain 5448 is able to bind a variety of blood group antigen structures (See above 6.2.1). Binding of M1 protein to A (type I and II; GalNAcα1-3(Fucα1-2)) and B (type I-IV; Galα1-3(Fucα1-2)Gal) antigen trisaccharide was assessed using SPR via single cycle kinetics. Recombinant M1 was shown to bind both A and B trisaccharide structures with binding dissociation constants of 2.41 ± 1.14 μM and 527.5 ± 252.7 nM respectively (Fig. 6.9A & B). To assess the potential role of blood group antigen structures in upper respiratory tract colonisation by GAS, buccal epithelial cell and saliva samples from 20 individuals of unknown blood type and secretory status were collected. Blood group status and expression were assessed via LC-ESI MS analysis of donor saliva and differences in association of 5448eGFP WT and 5448ΔM1eGFP to buccal cells from each donor were examined via flow cytometry.

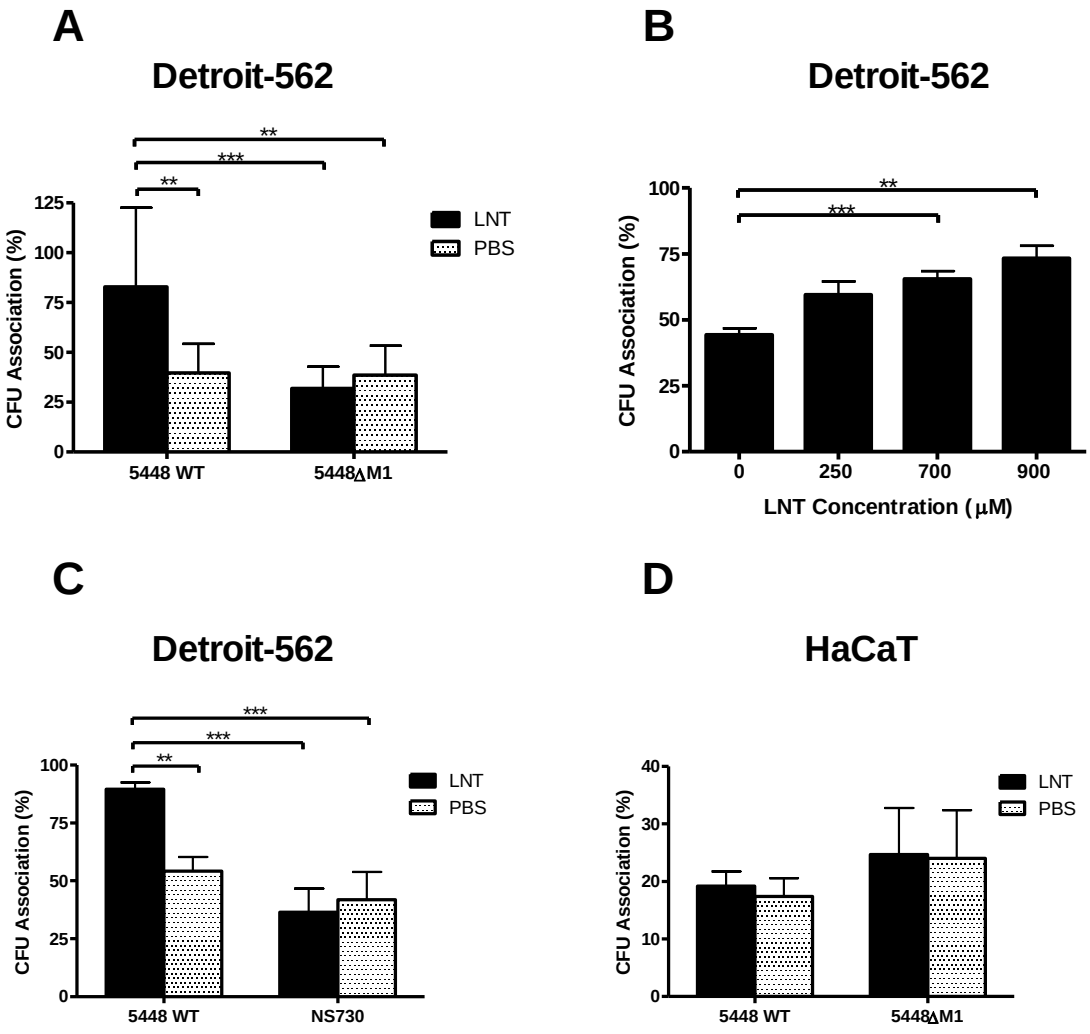


Figure 6.7: Lacto-*N*-tetraose mediated GAS association to host tissue. (A) Association of 5448 WT and 5448ΔM1 to Detroit-562 pharyngeal cells in the presence and absence of lacto-*N*-tetraose. (B) Association of 5448 to Detroit-562 cells in the presence of varying concentration of lacto-*N*-tetraose. (C) Comparable association of 5448 WT and NS730 (M90) to Detroit-562 cells in the presence and absence of lacto-*N*-tetraose. (D) Association of 5448 WT and 5448ΔM1 to HaCaT keratinocyte cells in the presence and absence of lacto-*N*-tetraose. GAS strains 5448 WT, 5448ΔM1 and NS730 were pre-incubated with 900 µM lacto-*N*-tetraose (unless stated otherwise) for 30 min at room temp, washed and incubated with Detroit-562 or HaCaT cells for 2 h at 37˚C, 5 % CO_2. Associated CFU was calculated against the initial inoculum.

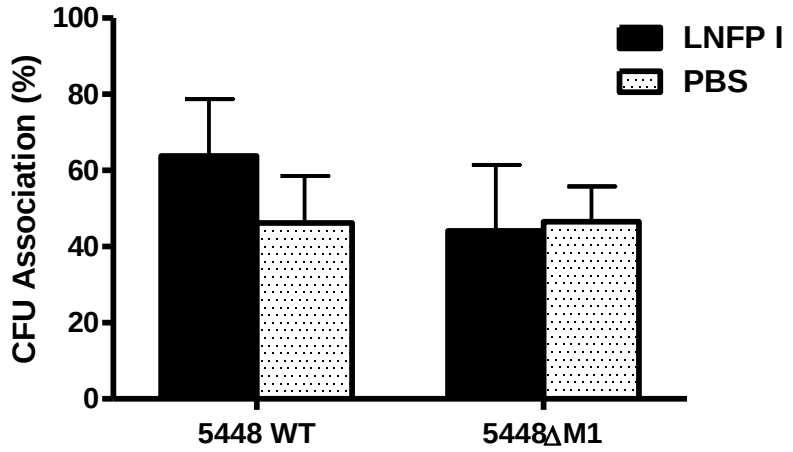


Figure 6.8: Lacto-N-fucopentaose type I mediated GAS association to host tissue. Association of 5448 WT and 5448ΔM1 to Detroit-562 pharyngeal cells was measured in the presence and absence of lacto-*N*-fucopentaose type I. GAS strains 5448 WT and 5448ΔM1 were pre-incubated with 900 µM lacto-*N*-fucopentaose type I for 30 min at room temp, washed and incubated with Detroit-562 cells for 2 h at 37˚C, 5 % CO_2. Associated CFU was calculated against the initial inoculum.

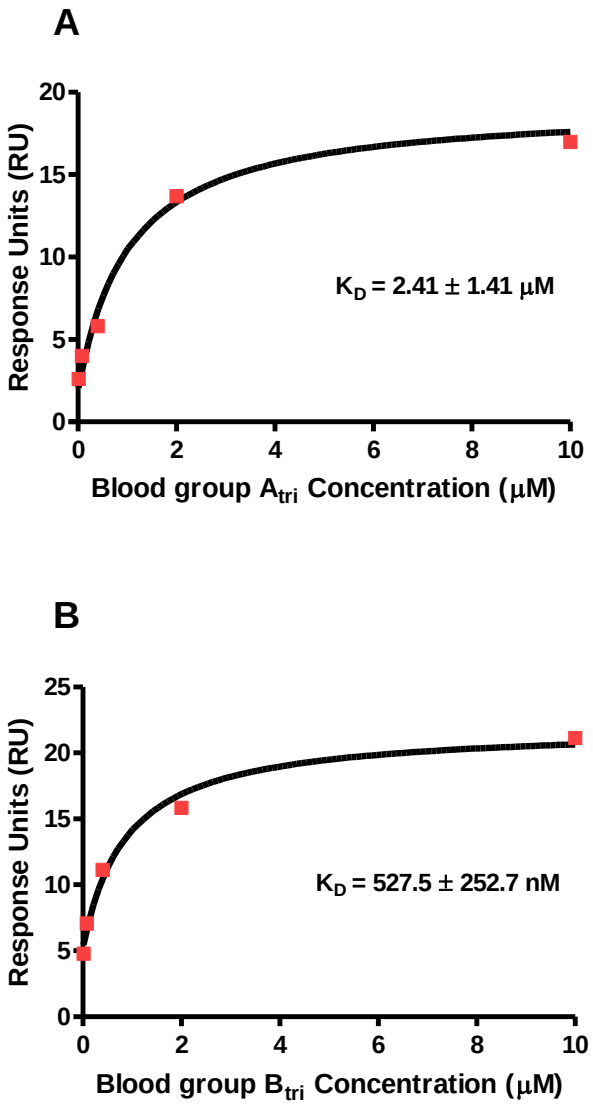


Figure 6.9: M1 binding of blood group A and B antigen related structures. M1 binding to (A) blood group A trisaccharide (A_{tri}; GalNAcα1-3(Fucα1-2)) and (B) blood group B trisaccharide (B_{tri}; Galα1-3(Fucα1-2)Gal) was determined via SPR and were modelled using steady state affinity fitting of the single cycle kinetic sensorgrams.

LC-ESI MS analysis of released *N*- and *O*-linked glycans from salivary glycoproteins revealed that of the 20 subjects, only 2 were blood group antigen non-secretors. The remaining 18 donors were shown to express a range of A, B and H antigen structures. From the 18 samples which tested positive for blood group antigen expression, 8 expressed the A antigen, 5 expressed A and B antigens, 4 expressed lone H antigen structures and only 1 was shown to express the B antigen moiety. Fragmented ions generated in the negative-ion MS^2 fragmentation spectra were used for detailed structural analysis of several of the many blood group antigen determining structures. Signals at *m/z* 350, 553 and 2002.7 indicate the presence of core fucosylation in reduced glycan structures, typical of blood group antigen related glycans (Everest-Dass *et al.* 2012) (Appendix G).

To explore the potential relationship between blood group antigen expression and GAS association to epithelial cells of the oral mucosa, human buccal epithelial cells were examined in the presence of GFP expressing 5448 WT and 5448ΔM1 via flow cytometry. Human buccal epithelial cells were identified by their characteristic forward and side scatter profile and a single gate was used to exclude debris and non-associated GAS (Everest-Dass *et al.* 2012). Observed human buccal epithelial cell auto-fluorescence was accounted for in the binding exhibited by 5448 WT and 5448ΔM1 (Fig 6.10A-C). Flow cytometry analysis demonstrated that 5448 WT was able to associate with buccal epithelial cells expressing H antigen structures at a significantly higher level than cells expressing A ($P < 0.05$) and AB ($P < 0.01$) antigen structures. As only 1 subject was shown to express the B antigen, no statistical inferences were able to be made (Fig. 6.11A). No statistical differences in 5448ΔM1 association were observed between buccal epithelial cells expressing A, A and B, B or H antigens (Fig. 6.11B). 5448ΔM1 was able to associate with buccal cells at a significantly higher level that 5448 WT ($P < 0.001$) (Fig 6.11A-B). This is supports previous

findings that SpeB mediated cleavage of M1 protein from the GAS cell surface enhances binding of the bacterium to epithelial cells of the oral tract (Anderson *et al.* 2014).

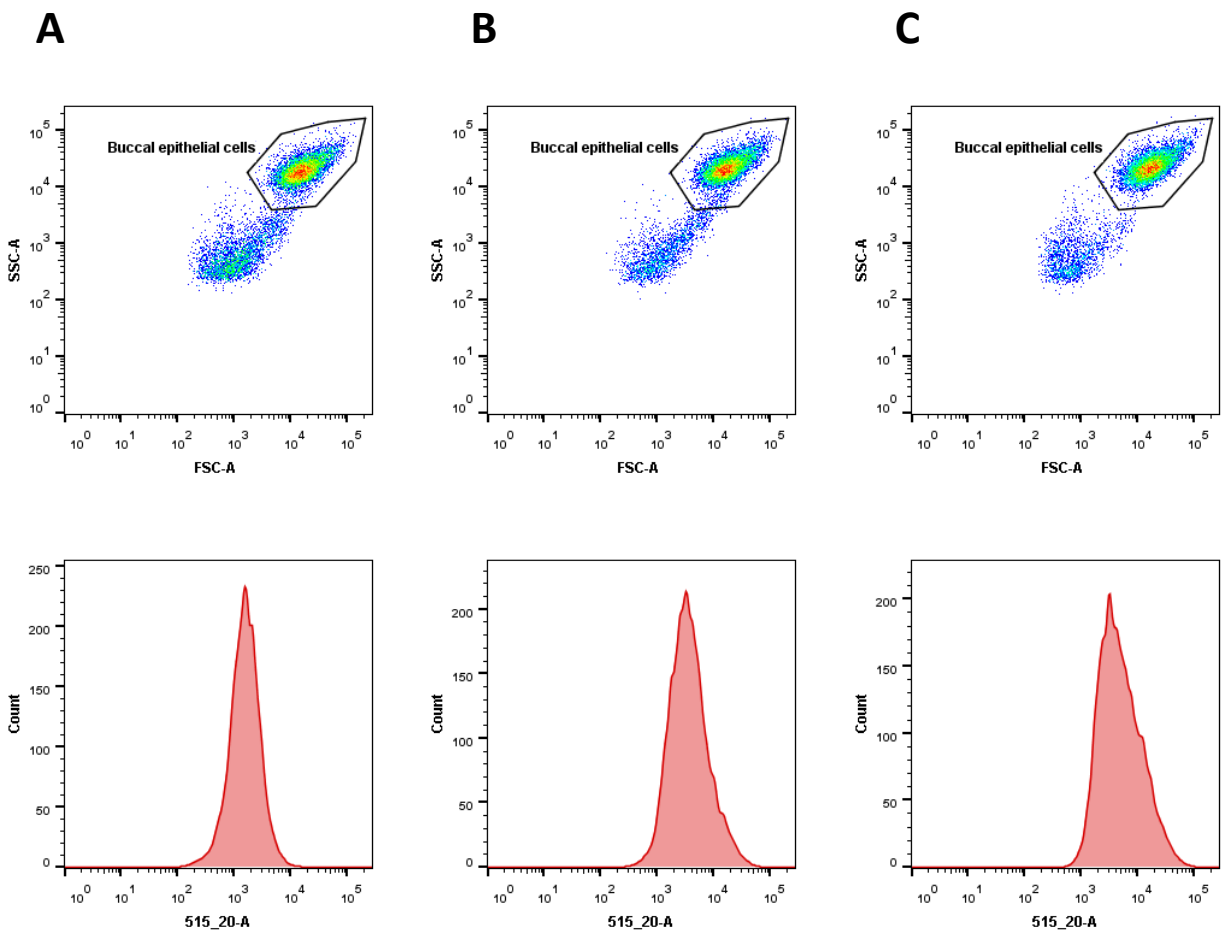


Figure 6.10: Flow cytometry assay of 5448 and 5448ΔM1 binding to human buccal epithelial cells. Dot plots at the top show the forward scatter (FSC-A) and the side scatter (SSC-A) distribution while the corresponding histograms of buccal epithelial cell count vs. fluorescence intensity at 515-520 nm is shown below. (A) Auto-fluorescence of buccal epithelial cells alone. (B) 5448eGFP WT incubated with buccal epithelial cells. (C) 5448ΔM1eGFP incubated with buccal epithelial cells.

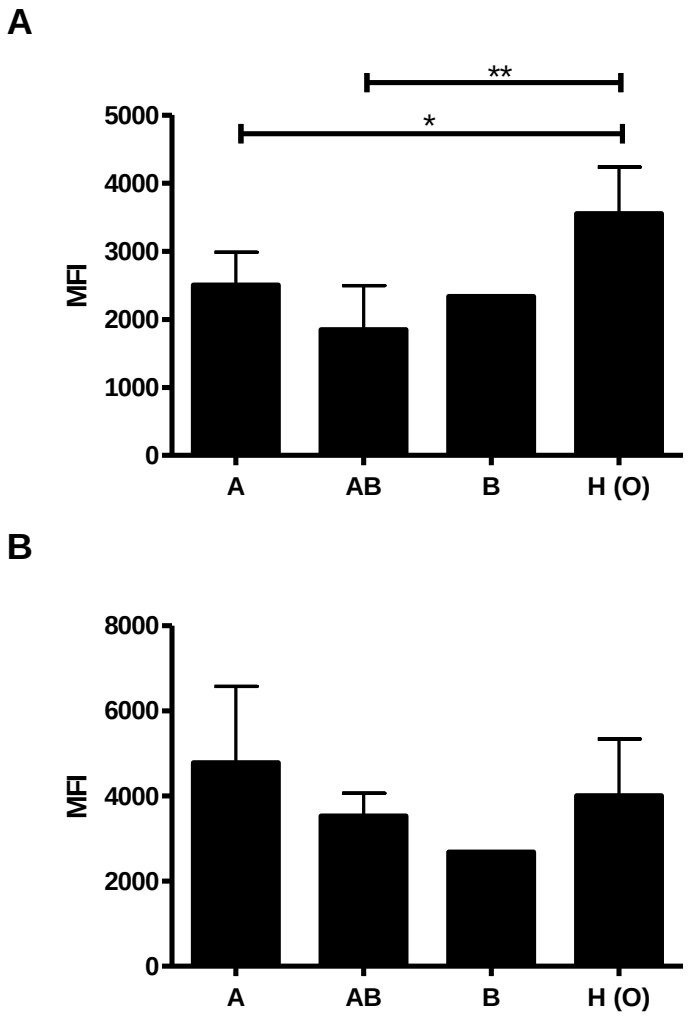


Figure 6.11: 5448 WT and 5448ΔM1 association to human buccal epithelial cells expressing different blood group antigen structures. Differences in association of 5448 WT (A) and 5448ΔM1 (B) to human buccal epithelial cells expressing either A, B, A and B, or H blood group antigen structures.

6.3 Discussion

Since the 1980's, the increase in GAS disease has been paralleled by the emergence of the globally disseminated M1T1 clone (Johnson *et al.* 2002). Epidemiological studies have identified that M1 serotypes are frequently isolated from throat cultures, representing a common cause of pharyngitis (Kaplan *et al.* 2001). Previous research examining the role of M1 protein-glycan interactions in GAS disease have focused on a select range of ECM-expressed GAG's. In this study, we have characterised novel M1-dependant glycan interactions which may be involved in M1 serotype colonisation of the host upper-respiratory tract.

Whole GAS cell glycan microarray analysis identified that GAS M1 serotype 5448 was able to bind 19 different glycan structures in an M1 dependant fashion. Differences in M1-glycan binding were noted between glycan microarray analysis using recombinant M1 protein and whole cell 5448. These variances in binding could be influenced by a series of factors including glycan presentation at the microarray epoxy surface along with M1 binding motif proximity to the presented glycans. M1 has been previously shown to be cleaved from the GAS cell surface via GAS secreted cysteine protease speB (Anderson *et al.* 2014). Cleavage of M1 from the cell surface may act to overcome restrictive factors encompassing glycan residue presentation required for binding.

Blood group antigen structures have been shown to be expressed on a range of host cell and protein surfaces (De Leoz *et al.* 2012; Everest-Dass *et al.* 2012). Interestingly, M1 dependent 5448 binding partners included blood group antigen precursor lacto-*N*-tetraose and lacto-*N*-fucopentaose I (H type I antigen). Blood group A and B antigen structures were found to bind recombinant M1, 5448 and 5448ΔM1, suggesting that along with M1, another lectin with blood group antigen specificity may be present at the GAS cell surface. Competitive glycan microarray analysis identified that structurally related lacto-*N*-tetraose, blood group H-

antigen type I and β1-4galactosyl-galactose share a homogenous binding domain whereby Galβ1-4Gal was the minimum carbohydrate residue required for recognition. Furthermore, binding by structurally distinct α1-3-mannobiose was also observed to be outcompeted by excess lacto-*N*-tetraose and blood group H-antigen type I, suggesting the existence of a broad range carbohydrate binding domain within M1. In the current study, lacto-*N*-tetraose and lacto-*N*-fucopentaose type I interactions were restricted to a 12 amino acid sequence (LEKELEEKKE/KAL) present at the start of both the B1 and B2 repeat domains of M1. The function of M1 is largely attributed to the ability of the protein to exchange between monomer and dimer states, a result of irregularities in B-repeat alpha-helicies (McNamara *et al.* 2008). This structurally distinct binding domain may represent a flexible region within M1 which is able to recognise multiple carbohydrates that are distinct in structure.

Fucosylation of lacto-*N*-tetraose, which results in blood group H-antigen type I was shown to significantly increase the affinity of M1 for lacto-*N*-tetraose. SPR analysis of blood group A and B trisaccharides which both contain α2-linked fucose residues demonstrated a significantly higher affinity for M1 than observed for precursor lacto-*N*-tetraose. Studies analysing the blood group recognition properties of rotavirus show that fucose moieties do not play a direct role in blood group antigen binding (Bohm *et al.* 2015). Alternatively, it has been hypothesised that fucose residues may play an indirect role by stabilising the blood group antigen via non-conventional hydrogen bonding, a reported binding property of Lewis^{x} antigen by norovirus (Zierke *et al.* 2013). Although the lectin binding mechanisms of M1 are yet to be elucidated, we have established that fucose is not essential for recognition of blood group related structures, but instead mediates a higher affinity interaction with the B-repeats of M1 which may parallel the observed lectin properties of both rotavirus and norovirus (Zierke *et al.* 2013; Bohm *et al.* 2015).

M1 serotypes are typically isolated from the upper respiratory tract during the course of a pharyngeal infection (Kaplan *et al.* 2001). Examination of the M1-lacto-*N*-tetraose binding identified that exogenous lacto-*N*-tetraose and not lacto-*N*-fucopentaose type I, significantly promoted the association of M1 GAS serotype 5448 to Detroit-562 pharyngeal cells in a tissue specific manner. If already present at the Detroit-562 cell surface, it was expected that exogenous lacto-*N*-tetraose would abrogate binding to the eukaryotic cell surface. The observed differences in association may suggest that lacto-*N*-tetraose can act as a bridging molecule between M1 and a specific cell surface receptor. Alternatively, M1-bound lacto-*N*-tetraose may be capable of homologously binding to various structurally related carbohydrates at the Detroit-562 cell surface. Homologous glycan-glycan interactions have been previously reported in marine eukaryotes, facilitating cell recognition and adhesion events (Bucior *et al.* 2004). As saliva is a source of lacto-*N*-tetraose related structures, the ability of lacto-*N*-tetraose to either act as a bridging molecule or mediate carbohydrate self-recognition events both represent conceivable means of GAS interaction with pharyngeal tissue.

To further examine the potential role of blood group antigen structures in GAS colonisation, buccal epithelial cells from 20 individuals of different blood group status were measured for binding against M1 serotype 5448 and 5448ΔM1. To our knowledge, this is the first study to investigate the potential link between GAS disease and host blood group status. The ability of 5448 to associate at a significantly higher level to buccal epithelial cells expressing H antigen structures may underlie differences in individual susceptibility to GAS disease, particularly pharyngitis. Correlations between blood group status and disease susceptibility have been shown for *H. pylori* with blood group H expressing individuals more predisposed to infection and symptomatic gastrointestinal complication (Jaff 2011). Although differences in association of GAS to buccal epithelial cell donors was correlated with blood group antigen

status, potential contributing factors such as donor age, ethnicity, and gender must not be overlooked in prospective studies. Secretor status of individuals has been previously associated with disease susceptibility (Jaff 2010). Due to sample size constraints, no definitive trend or correlation was able to be determined between secretor and non-secretor donor samples in the current study. Future studies examining differences in secretor status in relation to GAS colonisation may aid in future therapeutics centred on anti-adhesives that act to mimic host glycan receptors.

Globally disseminated M1T1 is a significant cause of GAS disease. Here, we have identified numerous novel M1 protein dependent glycan interactions while suggesting a role for lacto-*N*-tetraose in GAS colonisation of the upper respiratory tract. Furthermore, we have for the first time, provided evidence to support a potential link between host blood group antigen expression and GAS disease status. This study exemplifies the diverse nature of lectin-glycan interactions as well as the complexity of GAS disease and highlights the need for further studies focused on GAS-glycan interaction.

7. Conclusions and Future Research

The capacity of *S. pyogenes* to cause a wide range of non-invasive and invasive diseases makes it responsible for some of the highest morbidity and mortality rates worldwide. It is currently estimated that invasive GAS disease and disease related to post infection sequelae account for over 500 000 deaths per year (Carapetis *et al.* 2005; WHO 2005). In urbanised Western populations, M-types associated with outbreaks are typically over represented by M1 and M3 serotypes (Aziz *et al.* 2008). However, in regions such as the NT, Australia, the epidemiology of GAS infection is less defined, and is often linked to a variety of GAS serotypes (Stollerman 1997; McMillan *et al.* 2013). As M protein is still one of the most pursued GAS vaccine targets, there has been an increasing need to characterise diverse M proteins associated with GAS strains from regions which display the highest burden of infection. Mechanisms by which GAS initiate pathogenesis are yet to be fully understood, but it is thought that the interaction with host factors by M protein plays a pivotal role in triggering invasive disease. As such this study aimed to systematically characterise the function of a diverse sub-set of M proteins associated with invasive disease

The diversity of circulating *emm*-types has made epidemiological and functional comparisons extremely complex to analyse, and has hindered the development of a global GAS vaccine. Recent studies have challenged the concept of M-type specific immunity, which underlies the development of current M protein based vaccine candidates (Smeesters *et al.* 2008; Smeesters *et al.* 2010; Baroux *et al.* 2014). Here, the development of a cluster classification system which correlates full length *emm*/M protein sequence with function was undertaken, providing a strong foundation for prospective GAS vaccine development. Our results demonstrate that the cluster system can serve as a functional classification system which can facilitate future M protein functional studies, epidemiological surveillance and vaccine development. The recent application of this *emm*-cluster classification system against a sub-set of New Caledonia GAS isolates further exemplifies the value of the *emm*-cluster system

in analysing epidemiological trends and evaluating subsequent cross-protective vaccine strategies (Baroux *et al.* 2014).

As previously stated, one of the most pursued antigens for a global GAS vaccine is the M protein. The binding of host plasma proteins plays an important role in M protein vaccine development. These interactions can greatly influence the efficacy of selected opsonic epitopes. In M-types such as M6, these binding interactions have been shown to inhibit binding of antibodies directed against the B- and C-repeat domains of the protein. Furthermore in the case of IgG binding M protein, M protein-directed antibodies may also recognise directed and non-directed epitopes. The ability of α-SV1 IgG to recognise a wide range of M-types was reduced in the presence of various host plasma proteins. However, none of the plasma proteins examined in this study were able to completely inhibit α-SV1 IgG interaction with M protein. Similarly, the ability of M proteins to bind host plasma proteins was not fully blocked in the presence of α-SV1 IgG. The single exception to this finding was IgA binding by *emm*-cluster E1 M proteins, whereby IgA was found to inhibit α-SV1 IgG binding. Although the data presented here highlights the biochemical capabilities of α-SV1 IgG in recognising a broad range of M-types, it is not an absolute reflection of the interactions which may occur *in vivo*. Expression of M protein at the GAS cell surface has been shown to vary in response to disease progression and host selective pressures (Anderson *et al.* 2014). Future studies which examine α-SV1 IgG directed GAS recognition in human plasma, taking into account the aforementioned factors will significantly strengthen the utility of SV1 as a global GAS vaccine candidate.

The ability to bind host zymogen, plasminogen, was shown to be a restricted function of *emm*-cluster D4 M protein. The capacity of GAS to bind plasminogen has been previously shown to be central to multiple stages of pathogenesis (Sanderson-Smith *et al.* 2008; Ly *et al.* 2014; Walker *et al.* 2014). Differences in the glycosylation profile between GI-plasminogen

and GII-plasminogen are known to contribute to conformational differences between both glycoforms, subsequently affecting the kinetic and activational properties of certain plasminogen receptors (Takada *et al.* 1983; Gonzalez-Grownow *et al.* 2002; Law *et al.* 2012). In the current study, D4 M protein representative, PAM was shown display significantly higher affinity for GII-plasminogen over GI-plasminogen when plasminogen was in the 'open' conformation suggesting glycosylation at Asn_{289} in GI-plasminogen hinders interaction with PAM. The ability of PAM to bind plasminogen has been attributed to lysine dependant interactions with KR2 of plasminogen (Cnudde *et al.* 2006). Analysis of PAM binding interaction with GII-angiostatin in the presence of εACA confirmed the role of KR2 LBS in PAM-GII-Plg binding interactions. In contrast, whole molecule binding analysis revealed that comparative concentrations of εACA were unable to inhibit the interaction of PAM with full length GI-plasminogen and GII-plasminogen. The findings presented here suggest either: 1) the presence of a non-LBS dependent PAM binding site outside of KR2 which is more accessible in GII- plasminogen, or alternatively 2) a secondary higher-affinity LBS interaction in full-length plasminogen. Along with PAM, *emm* D-pattern GAS also express phenotypically distinct type-2b SK, which requires plasminogen ligands such as PAM to activate plasminogen (Cook *et al.* 2014). In the presence of PAM, type-2b SK was shown to generate an active site and activate GII-plasminogen at a significantly higher rate than GI-plasminogen. The accumulations of these findings clearly demonstrate the preferential acquisition and activation of GII-plasminogen over GI-plasminogen by PAM expressing GAS isolates. Additionally, these data highlight the impact of glycosylation on protein structure and function, representing a model for future research analysing the impact of glycan expression on host proteins known to be utilised in GAS pathogenesis.

To date, the characterisation of direct glycan-GAS interaction has focused on M protein interaction with a small selection of host GAG's. The development of the *emm*-cluster

classification system in conjunction with novel glycan array technology has enabled qualitative high throughput identification of novel M protein-glycan interactions, a large degree of which were shown to be *emm*-cluster specific. Glycan array analysis identified unique patters of binding function across each designated *emm*-cluster, further highlighting the utility of this novel cluster classification system based on full length M protein sequence. Although *emm*-clusters for the most part, demonstrated unique patters of binding, conservation in binding was observed across multiple *emm*-cluster groups, with multiple *emm*-clusters recognising similar terminal galactose, sialic acid, and GAG containing structures. The capacity of these *emm*-cluster groups to demonstrate conserved binding function suggests that these interactions may be important in mediating GAS disease. Future studies of the biochemical and physiological roles of these interactions may further contribute to our understanding of GAS pathogenesis which may subsequently aid in future therapeutic development.

Recent findings suggest that glycan interaction is an important step in initiating bacterial pathogenesis (Day *et al.* 2009; Everest-Dass *et al.* 2012). Results presented here indicate the M1T1 clone 5448 interacts with 19 glycan structures in an M1 protein dependant manner. Binding site analysis identified the presence of a homogenous glycan binding domain localised to the flexible disordered α-helical B-repeats of M1 (McNamara *et al.* 2008). Globally disseminated M1 is commonly associated with infection of the upper respiratory tract, a region which expresses high levels of terminal galactose and related blood group antigen structures (Johnson *et al.* 1992; Kaplan *et al.* 2001; Everest-Dass *et al.* 2012). Interaction with blood group antigen precursor, lacto-*N*-tetraose was shown to significantly promote adherence to Detroit-562 pharyngeal cells in an M1 dependent manner. Furthermore expression of blood group H antigen at the surface of human buccal epithelial cells was shown to significantly increase association of GAS strain 5448 over cells expressing blood

group antigens A, A and B or B. Although preliminary, these findings suggest that blood group antigens and related structures may be involved in GAS colonisation. Prior population studies have identified that pharyngitis is highly prevalent in children aged 5-15 (Carapetis *et al.* 2005). To date, no investigation has analysed GAS infection in relation to phenotypic differences in blood group antigen expression. The results presented in this study provide a scope for future research into the quantitative assessment of blood antigen expression in adolescents with respect to GAS disease status. Prospective studies such as these may further aid our understanding pharyngitis with potential for therapeutic development.

The results presented here clearly highlight the utility of the newly implemented GAS classification system. Furthermore, this study has identified novel M protein functions which may contribute to the different stages of GAS infection. Understanding the role of genetically diverse M protein in GAS pathogenesis has widespread implications for both the study and treatment of GAS disease.

References

AIHW (2013). Rheumatic heart disease and acute rheumatic fever in Australia: 1996–2012. Cardiovasular Disease. Canberra, Australian Institute of Health and Welfare. **Cat. no. CVD 60**.

Akesson, P., K. H. Schmidt, et al. (1994). "M1 protein and Protein H: IgGFc- and albumin-binding streptococcal surface proteins encoded by adjacent genes." Journal of Biochemistry **300**: 877-886.

Anderson, E. L., J. N. Cole, et al. (2014). "The fibrinogen-binding M1 protein reduces pharyngeal cell adherence and colonization phenotypes of M1T1 group A Streptococcus." The Journal of Biological Chemistry **289**(6): 3539-3546.

Andreasen, P. A., L. Kjøller, et al. (1997). "The urokinase-type plasminogen activator system in cancer metastasis: A review." International Journal of Cancer **72**: 1-22.

Arndt, N. X., J. Tiralongo, et al. (2011). "Differential Carbohydrate Binding and Cell Surface Glycosylation of Human Cancer Cell Lines." Journal of Cellular Biochemistry **112**: 2230-2240.

Aziz, R. K. and M. Kotb (2008). "Rise and Persistence of Global M1T1 Clone of *Streptococcus pyogenes*." Emerging Infectious Diseases **14**: 1511-1517.

Aziz, R. K., M. J. Pabst, et al. (2004). "Invasive M1T1 group A Streptococcus undergoes a phase-shift in vivo to prevent proteolytic degradation of multiple virulence factors by SpeB." Molecular Microbiology **51**(1): 123-134.

Barnett, T. C. and J. R. Scott (2002). "Differential recognition of surface proteins in Streptococcus pyogenes by two sortase gene homologs." Journal of Bacteriology **184**(8): 2181-2191.

Baroux, N., E. D'Ortenzio, et al. (2014). "The emm-cluster typing system for Group A Streptococcus identifies epidemiologic similarities across the pacific region." Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **59**(7): 84-92.

Batsford, S. R., S. Mezzano, et al. (2005). "Is the nephritogenic antigen in post-streptococcal glomerulonephritis pyrogenic exotoxin B (SPE B) or GAPDH?" Kidney international **68**(3): 1120-1129.

Bauer, M. J., M. M. Georgousakis, et al. (2012). "Evaluation of novel *Streptococcus pyogenes* vaccine candidated incorporating multiple conserved sequences from the c-repeat region of M-protein." Vaccine **30**: 2197-2205.

Beall, B., R. Facklam, et al. (1996). "Sequencing emm-Specific PCR Products for Routine and Accurate Typing of Group A Streptococci." Journal of Clinical Microbiology **34**(953-958).

Beall, B., R. R. Facklam, et al. (1998). "Streptococcal emm types associated with T-agglutination types and the use of conserved emm gene restriction fragment patterns for subtyping group A streptococci." Journal of medical microbiology **47**(10): 893-898.

Behrendt, N., M. Ploug, et al. (1991). "The ligand-binding domain of the cell surface receptor for urokinase-type plasminogen activator." The Journal of Biological Chemistry **266**(12): 7842-7847.

Ben Nasr, A. B., H. Herwald, et al. (1995). "Human kininogens interact with M protein, a bacterial surface protein and virulence determinant." The Biochemical journal **305 (Pt 1)**: 173-180.

Berge, A. and U. Sjöbring (1993). "PAM, a novel plasminogen-binding protein from Streptococcus pyogenes." J Biol Chem **268**: 25417-25424.

Bergey, E. J. and M. W. Stinson (1988). "Heparin-inhibitable basement membrane-binding protein of Streptococcus pyogenes." Infection and Immunity **56**(7): 1715-1721.

Berggard, K., E. Johnsson, et al. (2001). "Binding of human C4BP to the hypervariable region of M protein: a molecular mechanism of phagocytosis resistance in *Streptococcus pyogenes*." Molecular Microbiology **42**: 539-551.

Bergmann, S., D. Wild, et al. (2003). "Identification of a novel plasmin(ogen)-binding motif in surface displayed α-enolase of Streptococcus pneumoniae." Molecular Microbiology **49**: 411-423.

Bernard, P. (2008). "Management of common bacterial infections of the skin." Current opinion in infectious diseases **21**(2): 122-128.
Bessen, D. and V. A. Fischetti (1988). "Passive acquired mucosal immunity to group A streptococci by secretory Immunoglobulin A." Journal of Experimental Medicine **167**: 1945-1950.
Bessen, D. and V. A. Fischetti (1990). "A human IgG receptor of group A streptococci is associated with tissue site of infection and streptococcal class." The Journal of infectious diseases **161**(4): 747-754.
Bessen, D. E. (1994). "Localization of immunoglobulin A-binding sites within M or M-like proteins of group A streptococci." Infection and Immunity **62**(5): 1968-1974.
Bessen, D. E., T. R. Fiorentino, et al. (1997). "Molecular markers for throat and skin isolates of group A streptococci." Advances in experimental medicine and biology **418**: 537-543.
Bessen, D. E. and V. A. Fischetti (1992). "Nucleotide sequences of two adjacent M or M-like protein gene of group A streptococci: Different RNA transcript levels and identification of a unique immunoglobulin A-binding protein." Infection and Immunity **60**: 124-135.
Bessen, D. E. and S. Lizano (2010). "Tissue tropisms in group A streptococcal infections." Future microbiology **5**(4): 623-638.
Biggar, K. K., N. J. Dawson, et al. (2012). "Real-time protein unfolding: a method for determining the kinetics of native protein denaturation using a quantitative real-time thermocycler." BioTechniques **53**(4): 231-238.
Bisno, A. L. and D. L. Stevens (1996). "Streptococcal Infections of Skin and Soft Tissues." The New England Journal of Medicine **344**: 240-245.
Blixt, O., S. Head, et al. (2004). "Printed covalent glycan array for ligand profiling of diverse glycan binding proteins." Proceedings of the National Academy of Sciences of the United States of America **101**: 17033-17038.
Bohm, R., F. E. Fleming, et al. (2015). "Revisiting the role of histo-blood group antigens in rotavirus host-cell invasion." Nature communications **6**: 5907.
Boyle, M. D., J. Weber-Heynemann, et al. (1995). "Characterization of a gene coding for a type IIo bacterial IgG-binding protein." Molecular immunology **32**(9): 669-678.
Briones, A. V., T. Sato, et al. (2014). "Antibacterial activity of polyethylenimine/carrageenan multilayer against pathogenic bacteria." Advances in Chemical Engineering and Science **4**(2): 233-241.
Broder, C. C., R. Lottenberg, et al. (1991). "Isolation of a prokaryotic plasmin receptor. Relationship to a plasminogen activator produced by the same micro-organism." The Journal of Biological Chemistry **266**(8): 4922-4928.
Bucior, I., S. Scheuring, et al. (2004). "Carbohydrate-carbohydrate interaction provides adhesion force and specificity for cellular recognition." The Journal of cell biology **165**(4): 529-537.
Cabell, C. H., E. Abrutyn, et al. (2003). "Cardiology patient page. Bacterial endocarditis: the disease, treatment, and prevention." Circulation **107**(20): e185-187.
Calander, N., K. A. Karlsson, et al. (1988). "On the dissection of binding epitopes on carbohydrate receptors for microbes using molecular modelling." Biochimie **70**(11): 1673-1682.
Carapetis, J. R. and B. J. Currie (1996). "Group A streptococcus, pyoderma, and rheumatic fever." Lancet **347**(9010): 1271-1272.
Carapetis, J. R., A. C. Steer, et al. (2005). "The global burden of group A streptococcal diseases." The Lancet. Infectious diseases **5**(11): 685-694.
Carapetis, J. R., A. C. Steer, et al. (2005). "The global burden of group A streptococcal diseases." Lancet Infectious Diseases **5**(11): 685-694.
Carapetis, J. R., A. C. Steer, et al. (2005). "The global burden of group A streptococcal diseases." The Lancet Infectious Diseases **5**: 685-694.
Carapetis, R., A. M. Walker, et al. (1999). "Clinical and Epidemiological Features of Group A Streptococcal Bacteraemia in a Region with Hyperendemic Superficial Streptococcal Infection." Epidemiology and Infection **122**: 59-65

Carlsson, F., K. Berggard, et al. (2003). "Evasion of phagocytosis through cooperation between two ligand-binding regions in Streptococcus pyogenes M protein." The Journal of experimental medicine **198**(7): 1057-1068.

Christensen, U. and L. Molgaard (1992). "Positive co-operative binding at two weak lysine-binding sites governs the Glu-plasminogen conformational change." Biochem J **285 (Pt 2)**: 419-425.

Cnudde, S. E., M. Prorok, et al. (2006). "X-ray crystallographic structure of the angiogenesis inhibitor, angiostatin, bound to a peptide from the group A streptococcal surface protein PAM." Biochemistry **45**: 11052-11060.

Cook, S., I. Ellis, et al. (2007). "Headlice: a precursor to Group A Streptococcal infection in remote Indigenous children." Primary Intention: The Australian Journal of Wound Management **15**: 181-184.

Cook, S. M., A. Skora, et al. (2012). "Streptokinase variants from Streptococcus pyogenes isolates display altered plasminogen activation characteristics - implications for pathogenesis." Molecular Microbiology **86**: 1052-1062.

Cook, S. M., A. Skora, et al. (2013). "Site Restricted Plasminogen Activation Mediated by Group A Streptococcal Streptokinase Variants." The Biochemical journal.

Cook, S. M., A. Skora, et al. (2014). "Site-restricted plasminogen activation mediated by group A streptococcal streptokinase variants." The Biochemical journal **458**(1): 23-31.

Courtney, H. S., I. Ofek, et al. (1997). "M protein mediated adhesion of M type 24 *Streptococcus pyogenes* stimulates release of interleukin-6 by HEp-2 tissue culture cells." FEMS Microbiology Letters **151**: 65-70.

Crocker, P. R., J. C. Paulson, et al. (2007). "Siglecs and their roles in the immune system." Nature Reviews Immunology **7**: 255-266.

Cue, D., H. Lam, et al. (2001). "Genetic dissection of the Streptococcus pyogenes M1 protein: regions involved in fibronectin binding and intracellular invasion." Microbial Pathogenesis **31**: 231-242.

Cunningham, M. W. (2000). "Pathogenesis of Group A Streptococcal Infections." American Society of Microbiology **13**: 470-511.

Cunningham, M. W. (2012). "Streptococcus and rheumatic fever." Current opinion in rheumatology **24**(4): 408-416.

Dale, J. B., T. A. Penfound, et al. (2011). "New 30-valent M protein-based vaccine evokes cross-opsonic antibodies against non-vaccine serotypes of group A streptococci." Vaccine **29**(46): 8175-8178.

Davies, M. R., M. T. Holden, et al. (2014). "Emergence of scarlet fever Streptococcus pyogenes emm12 clones in Hong Kong is associated with toxin acquisition and multidrug resistance." Nature genetics.

Day, C. J., J. Tiralongo, et al. (2009). "Differential carbohydrate recognition by Campylobacter jejuni strain 11168: influences of temperature and growth conditions." PLoS ONE **4**(3): e4927.

De Leoz, M. L., S. C. Gaerlan, et al. (2012). "Lacto-N-tetraose, fucosylation, and secretor status are highly variable in human milk oligosaccharides from women delivering preterm." Journal of proteome research **11**(9): 4662-4672.

Degroote, S., E. Maes, et al. (2003). "Sulfated oligosaccharides isolated from the respiratory mucins of a secretor patient suffering from chronic bronchitis." Biochimie **85**(3-4): 369-379.

Dunn, L. A., D. J. McMillan, et al. (2002). "Parenteral and mucosal delivery of a novel multi-epitope M protein-based group A streptococcal vaccine construct: investigation of immunogenicity in mice." Vaccine **20**: 2635-2640.

Eilam, O., R. Zarecki, et al. (2014). "Glycan degradation (GlyDeR) analysis predicts mammalian gut microbiota abundance and host diet-specific adaptations." mBio **5**(4).

Ekelund, K., J. Darenberg, et al. (2005). "Variations in emm type among group A streptococcal isolates causing invasive or noninvasive infections in a nationwide study." Journal of Clinical Microbiology **43**(7): 3101-3109.
Ellis, N. M., Y. Li, et al. (2005). "T cell mimicry and epitope specificity of cross-reactive T cell clones from rheumatic heart disease." Journal of Immunology **175**(8): 5448-5456.
Everest-Dass, A. V., D. Jin, et al. (2012). "Comparative structural analysis of the glycosylation of salivary and buccal cell proteins: innate protection against infection by Candida albicans." Glycobiology **22**(11): 1465-1479.
Facklam, R., B. Beall, et al. (1999). "emm typing and validation of provisional M types for group A streptococci." Emerging Infectious Diseases **5**: 247-253.
Fae, K. C., D. D. da Silva, et al. (2006). "Mimicry in recognition of cardiac myosin peptides by heart-intralesional T cell clones from rheumatic heart disease." Journal of Immunology **176**(9): 5662-5670.
Farhud, D. D. and M. Zarif Yeganeh (2013). "A brief history of human blood groups." Iranian journal of public health **42**(1): 1-6.
Fischetti, V. A. (1989). "Streptococcal M Protein: Molecular Design and Biological Behaviour." Clinical Microbiology Reviews **2**: 285-314.
Fischetti, V. A., R. D. Horstmann, et al. (1995). "Location of the complement factor H binding site on streptococcal M6 protein." Infection and Immunity **63**(1): 149-153.
Fischetti, V. A., V. Pancholi, et al. (1990). "Conservation of a hexapeptide sequence in the anchor region of surface proteins from gram-positive cocci." Molecular Microbiology **4**(9): 1603-1605.
Fischetti, V. A., D. A. Parry, et al. (1988). "Conformational characteristics of the complete sequence of group A streptococcal M6 protein." Proteins **3**: 60-69.
Fluckiger, U., K. F. Jones, et al. (1998). "Immunoglobulins to group A streptococcal surface molecules decrease adherence to and invasion of human pharyngeal cells." Infection and Immunity **66**: 974-979.
Freifelder, D. (1982). Physical biochemistry. Applications to biochemistry and molecular biology. New York, W.H Freeman and Company.
Frick, I.-M., A. Schmidtchen, et al. (2003). "Interactions between M proteins of *Streptococcus pyogenes* and glycosaminoglycans promote bacterial adhesion to cells." European Journal of Biochemistry **270**: 2303-2311.
Frick, I. M., P. Akesson, et al. (1994). "Protein H--a surface protein of Streptococcus pyogenes with separate binding sites for IgG and albumin." Molecular Microbiology **12**(1): 143-151.
Fuhrer, T., E. Fischer, et al. (2005). "Experimental identification and quantification of glucose metabolism in seven bacterial species." Journal of Bacteriology **187**(5): 1581-1590.
Fung, J. C., K. Wicher, et al. (1982). "Immunochemical analysis of streptococcal group A, B, and C carbohydrates, with emphasis on group A." Infection and Immunity **37**(1): 209-215.
Galvin, J. E., M. E. Hemric, et al. (2000). "Cytotoxic mAb from rheumatic carditis recognizes heart valves and laminin." The Journal of clinical investigation **106**(2): 217-224.
Gardiner, D. L., A. M. Goodfellow, et al. (1998). "Group A Streptococcal Vir Types Are M-Protein Gene (emm) Sequence Type Specific." Journal of Clinical Microbiology **36**: 902-907.
Gidaris, D., D. Zafeiriou, et al. (2008). "Scarlet Fever and hepatitis: a case report." Hippokratia **12**: 186-187.
Gigi, I., T. Fujita, et al. (1979). "Modulation of the classical pathway C3 convertase by plasma proteins C4 binding protein and C3b inactivator." Proceedings of the National Academy of Sciences of the United States of America **76**: 6596-6600.
Gonzalez-Grownow, M., G. Gawid, et al. (2002). "Tissue Factor is the receptor for plasminogen type 1 on 1-LN human prostate cancer cells." Blood **99**: 4562-4567.
Greenfield, N. J. (2006). "Using circular dichroism spectra to estimate protein secondary structure." Nature protocols **1**(6): 2876-2890.

Griffin, J. D. and L. Ellman (1978). "Epsilon-aminocaproic acid (EACA)." Seminars in thrombosis and hemostasis **5**(1): 27-40.
Gubbe, K., R. Misselwitz, et al. (1997). "C repeats of the streptococcal M1 protein achieve the human serum albumin binding ability by flanking regions which stabilize the coiled-coil conformation." Biochemistry **36**(26): 8107-8113.
Haataja, S., K. Tikkanen, et al. (1993). "Characterization of a novel bacterial adhesion specificity of Streptococcus suis recognizing blood group P receptor oligosaccharides." The Journal of Biological Chemistry **268**(6): 4311-4317.
Harvey, H. A., W. E. Swords, et al. (2001). "The mimicry of human glycolipids and glycosphingolipids by the lipooligosaccharides of pathogenic neisseria and haemophilus." Journal of autoimmunity **16**(3): 257-262.
Hatton, M. W., S. Day, et al. (1999). "Plasminogen II accumulates five times faster than plasminogen I at the site of a balloon de-endothelializing injury in vivo to the rabbit aorta: comparison with other hemostatic proteins." Journal of Laboratory and Clinical Medicine **134**(3): 260-266.
Hatton, M. W. C., S. Southward, et al. (1994). "Catabolism of Plasminogen Glycoforms I and II in Rabbits: Relationship to Plasminogen Synthesis by the Rabbit Liver In Vitro." Metabolism **43**: 1430-1437.
Hayman, W. A., I. Toth, et al. (2002). "Enhancing the immunogenicity and modulating the fine epitope recognition of antisera to a helical group A streptococcal peptide vaccine candidate from the M protein using lipid-core peptide technology." Immunology & Cell Biology **80**: 178-187.
Herwald, H., H. Cramer, et al. (2004). "M Protein, a classical bacterial virulence determinant, forms complexes with fibrinogen that induce vascular leakage." Cell **116**: 367-379.
Hollands, A., M. A. Pence, et al. (2010). "Genetic switch to hypervirulence reduces colonization phenotypes of the globally disseminated group A streptococcus M1T1 clone." The Journal of infectious diseases **202**(1): 11-19.
Hollingshead, S., T. Readdy, et al. (1993). "Structural heterogeneity of the emm gene cluster in group A streptococci." Molecular Microbiology **8**: 707-717.
Hollingshead, S. K., V. A. Fischetti, et al. (1986). "Complete nucleotide sequence of type 6 M protein of the group A Streptococcus. Repetitive structure and membrane anchor." The Journal of Biological Chemistry **261**(4): 1677-1686.
Hollingshead, S. K., V. A. Fischetti, et al. (1987). "A highly conserved region present in transcripts encoding heterologous M protein of group A streptococci." Infection and Immunology **55**: 3237-3239.
Holub, M., M. Helec, et al. (2004). "Kinetics of immune parameters in a patient with sepsis caused by Streptococcus pyogenes treated with activated protein C." Scandinavian Journal of Infectious Diseases **36**: 485-488.
Hong, K. (2007). "Characterization of group a streptococcal M23 protein and comparison of the M3 and M23 protein's ligand-binding domains." Current microbiology **55**(5): 427-434.
Horstmann, R. D., H. J. Sievertsen, et al. (1992). "Role of fibrinogen in complement inhibition by streptococcal M protein." Infection and Immunity **60**(12): 5036-5041.
Hricik, D. E., M. Chung-Park, et al. (1998). "Glomerulonephritis." The New England Journal of Medicine **339**(13): 888-899.
Hsieha, Y.-C. and Y.-C. Huanga (2011). "Scarlet fever outbreak in Hong Kong, 2011." Journal of Microbiology, Immunology and Infection **44**: 409-411.
Hu, M. C., M. A. Walls, et al. (2002). "Immunogenicity of a 26-valent group A streptococcal vaccine." Infection and Immunity **70**(4): 2171-2177.
Ilver, D., A. Arnqvist, et al. (1998). "Helicobacter pylori adhesin binding fucosylated histo-blood group antigens revealed by retagging." Science **279**(5349): 373-377.
Iozzo, R. V. and A. D. Murdoch (1996). "Proteoglycans of the extracellular environment: clues from the gene and protein side offer novel perspectives in molecular diversity and function."

FASEB journal : official publication of the Federation of American Societies for Experimental Biology **10**(5): 598-614.
Jack-Weis, J., Y. Kim, et al. (1982). "Restricted Deposition of C3 on M+ group A streptococci: correlation with resistance to phagocytosis." Journal of Immunology **128**: 1897-1902.
Jackson, S. J., A. C. Steer, et al. (2011). "Systematic Review: Estimation of global burden of non-suppurative sequelae of upper respiratory tract infection: rheumatic fever and post-streptococcal glomerulonephritis." Tropical medicine & international health : TM & IH **16**(1): 2-11.
Jaff, M. S. (2010). "Higher frequency of secretor phenotype in O blood group - its benefits in prevention and/or treatment of some diseases." International journal of nanomedicine **5**: 901-905.
Jaff, M. S. (2011). "Relation between ABO blood groups and Helicobacter pylori infection in symptomatic patients." Clinical and experimental gastroenterology **4**: 221-226.
Jefferis, R. and D. S. Kumararatne (1990). "Selective IgG subclass deficiency: quantification and clinical relevance." Clinical and Experimental Immunology **81**(3): 357-367.
Jeng, A., V. Sakota, et al. (2003). "Molecular genetic analysis of a group A Streptococcus operon encoding serum opacity factor and a novel fibronectin-binding protein, SfbX." Journal of Bacteriology **185**(4): 1208-1217.
Jensen, U. S., J. D. Knudsen, et al. (2010). "Recurrent bacteraemia: A 10-year regional population-based study of clinical and microbiological risk factors." journal of Infection **60**: 191-199.
Johnson, D., J. T. Wotton, et al. (2002). "A comparison of group a streptococci from invasive and uncomplicated infections: Are virulent clones responsible for serious streptococcal infections?" Journal of Infectious Diseases **185**: 1586-1597.
Johnson, D. R., D. L. Stevens, et al. (1992). "Epidemiologic analysis of group A streptococcal serotypes associated with severe systemic infections, rheumatic fever, or uncomplicated pharyngitis." The Journal of infectious diseases **166**(2): 374-382.
Johnson, D. R., J. T. Wotton, et al. (2002). "A comparison of group A streptococci from invasive and uncomplicated infections: are virulent clones responsible for serious streptococcal infections?" The Journal of infectious diseases **185**(11): 1586-1595.
Johnsson, E., T. Areschoug, et al. (1999). "An IgA-binding peptide derived from a streptococcal surface protein." The Journal of Biological Chemistry **274**(21): 14521-14524.
Johnsson, E., A. Thern, et al. (1996). "A highly variable region in members of the streptococcal M protein family binds the human complement regulator C4BP." Journal of Immunology **157**(7): 3021-3029.
Kalia, A. and D. E. Bessen (2004). "Natural selection and evolution of streptococcal virulence genes involved in tissue-specific adaptations." Journal of Bacteriology **186**: 110-121.
Kantor, F. S. (1965). "Fibrinogen precipitation by streptococcal M protein." Journal of Experimental Medicine **121**: 849-859.
Kaplan, E. L., J. T. Wotton, et al. (2001). "Dynamic epidemiology of group A streptococcal serotypes associated with pharyngitis." Lancet **358**(9290): 1334-1337.
Kjellen, L. and U. Lindahl (1991). "Proteoglycans: structures and interactions." Annual review of biochemistry **60**: 443-475.
Kotarsky, H., M. Gustafsson, et al. (2001). "Group A streptococcal phagocytosis resistance is independent of complement factor H and factor H-like protein 1 binding." Molecular Microbiology **41**(4): 817-826.
Krivan, H. C., L. D. Olson, et al. (1989). "Adhesion of Mycoplasma pneumoniae to sulfated glycolipids and inhibition by dextran sulfate." The Journal of Biological Chemistry **264**(16): 9283-9288.
Kwiecinski, J., E. Josefsson, et al. (2010). "Activation of plasminogen by staphylokinase reduces the severity of Staphylococcus aureus systemic infection." The Journal of Infectious Diseases **202**: 1041-1049.

Lancefield, R. C. (1962). "Current knowledge of type-specific M antigens of group A streptococci." Journal of Immunology **89**: 307-313.
Lancefield, R. C. and V. P. Dole (1946). "The Properties of T Antigens Extracted from Group a Hemolytic Streptococci." The Journal of experimental medicine **84**(5): 449-471.
Law, R. H. P., T. Caradoc-Davies, et al. (2012). "The X-ray Crystal Structure of Full-Length Human Plasminogen." Cell Reports **1**: 185-190.
Leclerc, S., A. Teixeira, et al. (2006). "Recurrent Erysipelas: 47 Cases." Clinical and Laboratory Investigations **214**: 52-57.
Lijnen, H. R., B. Van Hoef, et al. (1981). "On the role of the carbohydrate side chains of human plasminogen in its interaction with alpha 2-antiplasmin and fibrin." European journal of biochemistry / FEBS **120**(1): 149-154.
Lindberg, A. A., J. E. Brown, et al. (1987). "Identification of the carbohydrate receptor for Shiga toxin produced by Shigella dysenteriae type 1." The Journal of Biological Chemistry **262**(4): 1779-1785.
Ly, D., J. M. Taylor, et al. (2014). "Plasmin(ogen) acquisition by group A Streptococcus protects against C3b-mediated neutrophil killing." Journal of innate immunity **6**(2): 240-250.
Macedo-Ramos, H., A. F. Batista, et al. (2014). "Evidence of involvement of the mannose receptor in the internalization of Streptococcus pneumoniae by Schwann cells." BMC microbiology **14**: 211.
Macheboeuf, P., C. Buffalo, et al. (2011). "Streptococcal M1 protein constructs a pathological host fibrinogen network." Nature **472**(7341): 64-68.
Makivuokko, H., S. J. Lahtinen, et al. (2012). "Association between the ABO blood group and the human intestinal microbiota composition." BMC microbiology **12**: 94.
Manjula, B. N. and V. A. Fischetti (1980). "Tropomyosin-like seven residue periodicity in three immunologically distinct streptococcal M proteins and its implication for the antiphagocytic property of the molecule." Journal of Experimental Medicine **151**: 695-708.
Marshall, J. M., A. J. Brown, et al. (1994). "Conformational studies of human plasminogen and plasminogen fragments: evidence for a novel third conformation of plasminogen." Biochemistry **33**(12): 3599-3606.
Marti, D. N., C. K. Hu, et al. (1997). "Ligand preferences of kringle 2 and homologous domains of human plasminogen: canvassing weak, intermediate, and high-affinity binding sites by 1H-NMR." Biochemistry **39**: 11591-11604.
McArthur, J. D., F. C. McKay, et al. (2008). "Allelic variants of streptokinase from *Streptococcus pyogenes* display functional differences in plasminogen activation." FASEB **22**: 3146-3153.
McArthur, J. D. and M. J. Walker (2006). "Domains of group A streptococcal M protein that confer resistance to phagocytosis, opsonization and protection: implications for vaccine development." Molecular Microbiology **59**(1): 1-4.
McKay, F. C., J. D. McArthur, et al. (2004). "Plasminogen Binding by Group A Streptococcal Isolates from a Region of Hyperendemicity for Streptococcal Skin Infection and a High Incidence of Invasive Infection." Infection and Immunology **72**: 364-370.
McMillan, D. J., P. A. Dreze, et al. (2013). "Updated model of group A Streptococcus M proteins based on a comprehensive worldwide study." Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **19**(5): E222-229.
McNamara, C., A. S. Zinkernagel, et al. (2008). "Coiled-Coil Irregularities and Instabilities in Group A Streptococcus M1 Are Required for Virulence." Science **319**: 1405-1408.
McNeil, S. A., S. A. Halperin, et al. (2005). "Safety and immunogenicity of 26-valent group a streptococcus vaccine in healthy adult volunteers." Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **41**(8): 1114-1122.
Mesnage, S., M. Dellarole, et al. (2014). "Molecular basis for bacterial peptidoglycan recognition by LysM domains." Nature communications **5**: 4269.

Middleton, B., P. Morris, et al. (2014). "Invasive group A streptococcal infection in the Northern Territory, Australia: Case report and review of the literature." Journal of paediatrics and child health **50**(11): 869-873.
Mitchell, E., C. Houles, et al. (2002). "Structural basis for oligosaccharide-mediated adhesion of Pseudomonas aeruginosa in the lungs of cystic fibrosis patients." Nature structural biology **9**(12): 918-921.
Molgaard, L., C. P. Ponting, et al. (1997). "Glycosylation of Asn-289 facilitates the ligand-induced conformation changes of human Glu-plasminogen." FEBS Letters **405**: 363-368.
Morfeldt, E., K. Berggard, et al. (2001). "Isolated hypervariable regions derived from streptococcal M proteins specifically bind human C4b-binding protein: implications for antigenic variation." Journal of Immunology **167**(7): 3870-3877.
Munthe, E. and J. B. Natvig (1972). "Immunoglobulin classes, subclasses and complexes of IgG rheumatoid factor in rheumatoid plasma cells." Clinical and Experimental Immunology **12**: 55-70.
Natanson, S., S. Sela, et al. (1995). "Distribution of fibronectin-binding proteins among group A streptococci of different M types." Journal of Infectious Diseases **171**: 871-878.
Neth, O., D. L. Jack, et al. (2000). "Mannose-binding lectin binds to a range of clinically relevant microorganisms and promotes complement deposition." Infection and Immunity **68**(2): 688-693.
Nieba, L., S. E. Nieba-Axmann, et al. (1997). "BIACORE Analysis of Histidine-Tagged Proteins Using a Chelating NTA Sensor Chip " Analytical biochemistry **252**: 217-228.
Nilson, B. H., I. M. Frick, et al. (1995). "Structure and stability of protein H and the M1 protein from Streptococcus pyogenes. Implications for other surface proteins of gram-positive bacteria." Biochemistry **34**(41): 13688-13698.
Nishihara, S., T. Hiraga, et al. (1999). "Molecular mechanisms of expression of Lewis b antigen and other type I Lewis antigens in human colorectal cancer." Glycobiology **9**(6): 607-616.
Nissenson, A. R., L. J. Baraff, et al. (1979). "Poststreptococcal acute glomerulonephritis: fact and controversy." Annals of internal medicine **91**(1): 76-86.
Nordenfelt, P., S. Waldemarson, et al. (2012). "Antibody orientation at bacterial surfaces is related to invasive infection." The Journal of experimental medicine **209**(13): 2367-2381.
O'Toole, P., L. Stenberg, et al. (1992). "Two major classes in the M protein family in group A streptococci." Proceedings of the National Academy of Sciences of the United States of America **89**: 8661-8665.
Ogawa, M. Y., J. Fan, et al. (2006). "Electron-Transfer Functionality of Synthetic Coiled-Coil Metalloproteins." Journal of the Brazilian Chemical Society **17**(8): 1516-1521.
Pacheco, A. R., M. M. Curtis, et al. (2012). "Fucose sensing regulates bacterial intestinal colonization." Nature **492**(7427): 113-117.
Pack, T. D., A. Podbielski, et al. (1996). "Identification of an amino acid signature sequence predictive of protein G-inhibitable IgG3-binding activity in group-A streptococcal IgG-binding proteins." Gene **171**(1): 65-70.
Påhlman, L. I., M. Mörgelin, et al. (2006). "Streptococcal M Protein: A Multipotent and Powerful Inducer of Inflammation." Journal of Immunology **117**: 1221-1228.
Pancholi, V. and V. A. Fischetti (1998). "alpha-enolase, a novel strong plasmin(ogen) binding protein on the surface of pathogenic streptococci." The Journal of Biological Chemistry **273**(23): 14503-14515.
Pandey, M., S. Sekuloski, et al. (2009). "Novel strategies for controlling Streptococcus pyogenes infection and associated diseases: from potential peptide vaccines to antibody immunotherapy." Immunology and Cell Biology **87**: 391-399.
Penc, S. F., B. Pomahac, et al. (1998). "Dermatan sulfate released after injury is a potent promoter of fibroblast growth factor-2 function." The Journal of Biological Chemistry **273**(43): 28116-28121.

Penfound, T. A., E. Y. Chiang, et al. (2010). "Protective efficacy of group A streptococcal vaccines containing type-specific and conserved M protein epitopes." Vaccine **28**(31): 5017-5022.
Perez-Casal, J., N. Okada, et al. (1995). "Role of the conserved C repeat region of the M protein of streptococcus pyogenes." Molecular Microbiology **15**: 907-916.
Persson, J., B. Beall, et al. (2006). "Extreme sequence divergence but conserved ligand-binding specificity in Streptococcus pyogenes M protein." PLoS pathogens **2**(5): e47.
Persson, M., J. A. Letts, et al. (2007). "Structural effects of naturally occurring human blood group B galactosyltransferase mutations adjacent to the DXD motif." The Journal of Biological Chemistry **282**(13): 9564-9570.
Pfoh, E., M. R. Wessels, et al. (2008). "Burden and economic cost of group A streptococcal pharyngitis." Pediatrics **121**(2): 229-234.
Phillips, G. N., P. F. Flicker, et al. (1981). "Streptococcal M protein: alpha-helical coiled-coil structure and arrangement on the cell surface." Proceedings of the National Academy of Sciences of the United States of America **78**: 4698-4693.
Phillips, G. N., Jr., P. F. Flicker, et al. (1981). "Streptococcal M protein: alpha-helical coiled-coil structure and arrangement on the cell surface." Proceedings of the National Academy of Sciences of the United States of America **78**(8): 4689-4693.
Podbielski, A., M. Woischnik, et al. (1996). "What is the size of the group A streptococcal vir regulon? The Mga regulator affects expression of secreted and surface virulence factors." Medical Microbiology and Immunology **185**: 171-181.
Ponting, C. P., S. K. Holland, et al. (1992). "The compact domain conformation of human Glu-plasminogen in solution." Biochimica et biophysica acta **1159**(2): 155-161.
Ponting, C. P., J. M. Marshall, et al. (1992). "Plasminogen: a structural review." Blood Coagul Fibrinolysis **3**: 605-614.
Quinn, A., S. Kosanke, et al. (2001). "Induction of autoimmune valvular heart disease by recombinant streptococcal m protein." Infection and Immunity **69**(6): 4072-4078.
Ram, S., A. K. Sharma, et al. (1998). "A novel sialic acid binding site on factor H mediates serum resistance of sialylated Neisseria gonorrhoeae." The Journal of experimental medicine **187**(5): 743-752.
Reichardt, W., K. Gubbe, et al. (1995). "M3-protein with close sequence homology to M12 protein binds fibrinogen, albumin, fibronectin, but not to any subclass of IgG-localization of binding regions." Developments in biological standardization **85**: 179-182.
Reichardt, W., K. H. Schmidt, et al. (1997). "Mapping of binding sites for human serum albumin and fibrinogen on the M3-protein. Molecular model and function in the pathogenic mechanism." Advances in experimental medicine and biology **418**: 577-579.
Retnoningrum, D. S. and P. P. Cleary (1994). "M12 protein from Streptococcus pyogenes is a receptor for immunoglobulin G3 and human albumin." Infection and Immunity **62**(6): 2387-2394.
Rhim, A. D., L. Stoykova, et al. (2001). "Terminal glycosylation in cystic fibrosis (CF): a review emphasizing the airway epithelial cell." Glycoconjugate journal **18**(9): 649-659.
Ringdahl, U. and U. Sjobring (2000). "Analysis of Plasminogen-Binding M Proteins of Streptococcus pyogenes." Methods **21**: 143-150.
Ringdahl, U., H. G. Svensson, et al. (2000). "A role for the fibrinogen-binding regions of streptococcal M proteins in phagocytosis resistance." Molecular Microbiology **37**: 1318-1326.
Rios-Steiner, J. L., M. Schenone, et al. (2001). "Structure and binding determinants of the recombinant kringle-2 domain of human plasminogen to an internal peptide from a group A Streptococcal surface protein." Journal of Molecular Biology **308**(4): 705-719.
Rodriguez-Iturbe, B. and S. Batsford (2007). "Pathogenesis of poststreptococcal glomerulonephritis a century after Clemens von Pirquet." Kidney international **71**(11): 1094-1104.
Ryan, P. A., V. Pancholi, et al. (2001). "Group A streptococci bind to mucin and human pharyngeal cells through sialic acid-containing receptors." Infection and Immunity **69**(12): 7402-7412.

Sanderson-Smith, M., M. Batzloff, et al. (2006). "Divergence in the plasminogen-binding group A streptococcal M protein family - Functional conservation of binding site and potential role for immune selection of variants." Journal of Biological Chemistry **281**(6): 3217-3226.
Sanderson-Smith, M., M. Batzloff, et al. (2006). "Divergence in the plasminogen-binding group a streptococcal M protein family: functional conservation of binding site and potential role for immune selection of variants." The Journal of Biological Chemistry **281**(6): 3217-3226.
Sanderson-Smith, M., D. M. De Oliveira, et al. (2014). "A Systematic and Functional Classification of Streptococcus pyogenes That Serves as a New Tool for Molecular Typing and Vaccine Development." The Journal of infectious diseases **210**(8): 1325-1338.
Sanderson-Smith, M. L., D. M. De Oliveira, et al. (2012). "Bacterial plasminogen receptors: mediators of a multifaceted relationship." Journal of biomedicine & biotechnology **2012**: 272148.
Sanderson-Smith, M. L., K. Dinkla, et al. (2008). "M protein-mediated plasminogen binding is essential for the virulence of an invasive Streptococcus pyogenes isolate." FASEB **22**(8): 2715-2722.
Sanderson-Smith, M. L., M. Dowton, et al. (2007). "The plasminogen-binding group A streptococcal M protein-related protein Prp binds plasminogen via arginine and histidine residues." Journal of Bacteriology **189**(4): 1435-1440.
Sanderson-Smith, M. L., M. J. Walker, et al. (2006). "The Maintenance of High Affinity Plasminogen Binding by Group A Streptococcal Plasminogen-binding M-like protein Is Mediated by Arginine and Histidine Residues Within the a1 and a2 Repeat Domains." The Journal of Biological Chemistry **281**: 25965-25971.
Sanderson-Smith, M. L., M. J. Walker, et al. (2006). "The maintenance of high affinity plasminogen binding by group A streptococcal plasminogen-binding M-like protein is mediated by arginine and histidine residues within the a1 and a2 repeat domains." Journal of Biological Chemistry **281**(36): 25965-25971.
Sandin, C., F. Carlsson, et al. (2006). "Binding of human plasma proteins to Streptococcus pyogenes M protein determines the location of opsonic and non-opsonic epitopes." Molecular Microbiology **59**(1): 20-30.
Schneewind, O., K. F. Jones, et al. (1990). "Sequence and structural characteristics of the trypsin-resistant T6 surface protein of group A streptococci." Journal of Bacteriology **172**(6): 3310-3317.
Severi, E., D. W. Hood, et al. (2007). "Sialic acid utilization by bacterial pathogens." Microbiology **153**(Pt 9): 2817-2822.
Severi, E., G. Randle, et al. (2005). "Sialic acid transport in Haemophilus influenzae is essential for lipopolysaccharide sialylation and serum resistance and is dependent on a novel tripartite ATP-independent periplasmic transporter." Molecular Microbiology **58**(4): 1173-1185.
Sharon, N. (1996). "Carbohydrate-lectin interactions in infectious disease." Advances in experimental medicine and biology **408**: 1-8.
Slomiany, B. L. and A. Slomiany (1978). "ABH-blood-group antigens and glycolipids of human saliva." European journal of biochemistry / FEBS **85**(1): 249-254.
Smeesters, P. R., M. Dramaix, et al. (2010). "The emm-type diversity does not always reflect the M protein genetic diversity--is there a case for designer vaccine against GAS." Vaccine **28**(4): 883-885.
Smeesters, P. R., P.-A. Drèze, et al. (2010). "Group A Streptococcus virulence and host factors in two toddlers with rheumatic fever following toxic shock syndrome." International Journal of Infectious Disease **14**: 403-409.
Smeesters, P. R., P. Mardulyn, et al. (2008). "Genetic diversity of Group A Streptococcus M protein: implications for typing and vaccine development." Vaccine **26**(46): 5835-5842.
Smeesters, P. R., A. Vergison, et al. (2006). "Differences between Belgian and Brazilian Group A Streptococcus Epidemiologic Landscape." PLoS ONE **1**: e10.

Smith, D. B. and K. S. Johnson (1988). "Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase." Gene **67**(1): 31-40.
Sriskandan, S. and D. M. Altmann (2008). "The immunology of sepsis." The Journal of pathology **214**(2): 211-223.
Stahl, M., L. M. Friis, et al. (2011). "L-fucose utilization provides Campylobacter jejuni with a competitive advantage." Proceedings of the National Academy of Sciences of the United States of America **108**(17): 7194-7199.
Stahl, P. D. and R. A. Ezekowitz (1998). "The mannose receptor is a pattern recognition receptor involved in host defense." Current opinion in immunology **10**(1): 50-55.
Steer, A. C., J. B. Dale, et al. (2013). "Progress toward a global group a streptococcal vaccine." The Pediatric infectious disease journal **32**(2): 180-182.
Steer, A. C., I. Law, et al. (2009). "Global emm type distribution of group A streptococci: systematic review and implications for vaccine development." The Lancet. Infectious diseases **9**(10): 611-616.
Stenberg, L., P. O'Toole, et al. (1992). "Many group A streptococcal strains express two different immunoglobulin-binding proteins, encoded by closely linked genes: characterization of the proteins expressed by four strains of different M-type." Molecular Microbiology **6**(9): 1185-1194.
Stevens, D. L. (1992). "Invasive group A streptococcus infections." Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **14**(1): 2-11.
Stollerman, G. H. (1997). "Rheumatic fever." Lancet **349**: 935-942.
Sun, Z., Y. H. Chen, et al. (2002). "The blockage of the high-affinity lysine binding sites of plasminogen by EACA significantly inhibits prourokinase-induced plasminogen activation." Biochimica et biophysica acta **1596**(2): 182-192.
Sung, M. A., H. A. Chen, et al. (2001). "Sequential assignment and secondary structure of the triple-labelled carbohydrate-binding domain of papG from uropathogenic E. coli." Journal of biomolecular NMR **19**(2): 197-198.
Svensson, M. D., U. Sjobring, et al. (2002). "Roles of the plasminogen activator streptokinase and the plasminogen-associated M protein in an experimental model for streptococcal impetigo." Microbiology **148**(Pt 12): 3933-3945.
Takada, A. and Y. Takada (1983). "The activation of two isozymes of glu-plasminogen (I and II) by urokinase and streptokinase." Thrombosis research **30**(6): 633-642.
Tao, S. C., Y. Li, et al. (2008). "Lectin microarrays identify cell-specific and functionally significant cell surface glycan markers." Glycobiology **18**(10): 761-769.
Tharp, A. C., M. L. M, et al. (2009). "Plasminogen substrate recognition by the streptokinase-plasminogen catalytic complex is facilitated by Arg253, Lys256, and Lys257 in the streptokinase beta-domain and kringle 5 of the substrate." The Journal of Biological Chemistry **284**: 19511-19521.
Thomas, W. E., L. M. Nilsson, et al. (2004). "Shear-dependent 'stick-and-roll' adhesion of type 1 fimbriated Escherichia coli." Molecular Microbiology **53**(5): 1545-1557.
Thomas, W. E., E. Trintchina, et al. (2002). "Bacterial adhesion to target cells enhanced by shear force." Cell **109**(7): 913-923.
Thorsen, S., I. Clemmensen, et al. (1981). "Adsorption to fibrin of native fragments of known primary structure from human plasminogen." Biochimica et biophysica acta **668**(3): 377-387.
Varadi, A. and L. Patthy (1981). "Kringle 5 of human plasminogen carries a benzamidine-binding site." Biochemical and biophysical research communications **103**(1): 97-102.
Veasy, L., S. Wiedmeier, et al. (1987). "Resurgence of acute rheumatic fever in the intermountain area of the United States." The New England Journal of Medicine **316**: 421-427.
Vimr, E. and C. Lichtensteiger (2002). "To sialylate, or not to sialylate: that is the question." Trends in Microbiology **10**(6): 254-257.

Vimr, E., C. Lichtensteiger, et al. (2000). "Sialic acid metabolism's dual function in Haemophilus influenzae." Molecular Microbiology **36**(5): 1113-1123.
Waldemarsson, J., M. Stalhammar-Carlemalm, et al. (2009). "Functional dissection of Streptococcus pyogenes M5 protein: the hypervariable region is essential for virulence." PLoS ONE **4**(10): e7279.
Walker, M. J., T. C. Barnett, et al. (2014). "Disease manifestations and pathogenic mechanisms of group a Streptococcus." Clinical Microbiology Reviews **27**(2): 264-301.
Walker, M. J., H. Hollands, et al. (2007). "Selection pressure by the innate immune system switches Streptococcus pyogenes M1T1 into a hyperinvasive pathogen." Nature Medicine **13**(981-985).
Walker, M. J., J. D. McArthur, et al. (2005). "Is plasminogen deployed as a Streptococcus pyogenes virulence factor?" Trends in Microbiology **13**(7): 308-313.
Walker, M. J., J. D. McArthur, et al. (2005). "Is plasminogen deployed as a Streptococcus pyogenes virulence factor." Trends in Microbiology **13**: 308-313.
Walz, A., S. Odenbreit, et al. (2005). "Identification and characterization of binding properties of Helicobacter pylori by glycoconjugate arrays." Glycobiology **15**(700-708).
Wang, H., R. Lottenberg, et al. (1995). "Analysis of the interaction of group A streptococci with fibrinogen, streptokinase and plasminogen." Microbial Pathogenesis **18**: 153-166.
Wang, J.-R. and M. W. Stinson (1994). "M Protein Mediates Streptococcal Adhesion to HEp-2 Cells." Infection and Immunity **62**: 442-448.
Whatmore, A. M., V. Kapur, et al. (1995). "Molecular population genetic analysis of the enn subdivision of group A streptococcal emm-like genes: horizontal gene transfer and restricted variation among enn genes." Molecular Microbiology **15**: 1039-1048.
Whitnack, E. and E. H. Beachey (1985). "Biochemical and biological properties of the binding of human fibrinogen to M protein in group A streptococci." Journal of Bacteriology **164**(1): 350-358.
Whitnack, E., J. B. Dale, et al. (1984). "Common protective antigens of group A streptococcal M proteins masked by fibrinogen." Journal of Experimental Medicine **159**(1201-1212).
WHO (2005). The Current Evidence for the Burden of Group A Streptococcal Diseases WHO/FCH/CAH/.
Wiesen, M. H., F. Farowski, et al. (2012). "Liquid chromatography-tandem mass spectrometry method for the quantification of mycophenolic acid and its phenolic glucuronide in saliva and plasma using a standardized saliva collection device." Journal of chromatography. A **1241**: 52-59.
Wilson, N. L., B. L. Schulz, et al. (2002). "Sequential analysis of N- and O-linked glycosylation of 2D-PAGE separated glycoproteins." Journal of proteome research **1**(6): 521-529.
Winters, B. D., N. Ramasubbu, et al. (1993). "Isolation and characterization of a Streptococcus pyogenes protein that binds to basal laminae of human cardiac muscle." Infection and Immunity **61**(8): 3259-3264.
Wistedt, A. C., H. Kotarsky, et al. (1998). "Kringle 2 mediates high affinity binding of plasminogen to an internal sequence in streptococcal surface protein PAM." The Journal of Biological Chemistry **273**(38): 24420-24424.
Wistedt, A. C., U. Ringdahl, et al. (1995). "Identification of a plasminogen-binding motif in PAM, a bacterial surface protein." Molecular Microbiology **18**: 569-578.
Wu, A. M., J. H. Wu, et al. (2006). "Interactions of the fucose-specific Pseudomonas aeruginosa lectin, PA-IIL, with mammalian glycoconjugates bearing polyvalent Lewis(a) and ABH blood group glycotopes." Biochimie **88**: 1479-1492.
Xia, B., J. A. Royall, et al. (2005). "Altered O-glycosylation and sulfation of airway mucins associated with cystic fibrosis." Glycobiology **15**(8): 747-775.

Yamada, M., H. Nakae, et al. (2002). "N-acetyl-D-galactosamine specific lectin of Eikenella corrodens induces intercellular adhesion molecule-1 (ICAM-1) production by human oral epithelial cells." Journal of medical microbiology **51**(12): 1080-1089.

Yang, N. and B. Boettcher (1991). "Conversion of the human blood group H antigen to A antigen in vitro." Immunology and Cell Biology **69 (Pt 2)**: 111-118.

Yung, D. L. and S. K. Hollingshead (1996). "DNA sequencing and gene expression of the emm gene cluster in an M50 group A streptococcus strain virulent for mice." Infection and Immunity **64**(6): 2193-2200.

Zierke, M., M. Smiesko, et al. (2013). "Stabilization of branched oligosaccharides: Lewis(x) benefits from a nonconventional C-H...O hydrogen bond." Journal of the American Chemical Society **135**(36): 13464-13472.

Appendix A: Media and general buffer compositions

Unless stated otherwise, all media and buffers were made up to volume with distilled H_2O

Media composition

Luria-Bertani (LB) broth

Tryptone	10 g/L
Yeast extract	5 g/L
NaCl	10 g/L

LB agar

LB broth	
Agar	15 g/L

Todd-Hewitt broth (THY)

Todd-Hewitt	30 g/L
Yeast extract	10 g/L

THY agar (THYA)

THY	
Agar	15g/L

Buffer composition

1 × Phosphate buffer saline (PBS)

NaCl	8 g/L
KCl	0.2 g/L
Na_2HPO_4	1.44 g/L
KH_2PO_4	0.24 g/L

Make up to 800 mL volume with dH_2O, adjust to pH 7.4 and expand to 1L with dH_2O.

1 × Dulbecco's PBS

NaCl	8 g/L
KCl	0.2 g/L
Na_2HPO_4	1.44 g/L
KH_2PO_4	0.24 g/L
$CaCl_2$	99.88 mg/L
$MgCl_2$	46.65 mg/L

1 × Tris-acetate-EDTA (TAE) buffer

Tris base	4.84 g/L

Glacial acetic acid	1.14 ml/L
EDTA	0.292 g/L

TAE agarose gel 1%

1 × TAE buffer	100 mL
Agarose	1 g

DNA loading buffer

Bromophenol blue	5 mg/10 mL
Glycerol	7.5 mL/10 mL

Make up to 10 mL with TAE buffer

Ethidium bromide staining solution

Ethidium bromide	1 µg/100 mL

5 × protein cracking buffer

Tris-HCl (pH 6.8)	4.5 mL/ 10 mL
Glycerol	5 mL/ 10 mL
SDS	0.5 g/ 10 mL
Bromophenol blue	5 mg/ 10 mL

1 × SDS-PAGE running buffer

Tris base	3.03 g/L
Glycine	14.33 g/L
SDS	1 g/L

Coomassie blue rapid stain

Coomassie blue R-250	2 g/L
Methanol	400 mL
Glacial acetic acid	100 mL

Make up to 1 L with dH_2O.

Rapid destain

Glacial acetic acid	100 mL
Methanol	400 mL

Make up to 1 L with dH_2O.

Western transfer buffer

Glycine	14.4 g/L
Tris base	3.02 g/L
Methanol	200 m/L

Make up to 1 L with dH_2O.

Native cell lysis buffer

Lysozyme	50 mg/50 ml
DNase I	25 μg/50 ml
PMSF	8.725 mg/ 50 ml
$MgCl_2$	476 μg/50 ml
$CaCl_2$	1.4 μg/50 ml
Triton X-100	50 μl/50 ml
NaCl	1.055 g/50 ml
NaH_2PO_4	0.3 g/50 ml
Imidazole	85 μg/50 ml

Native wash buffer – nickel affinity chromatography

PBS	1 L
Imidazole	1.36 g/L

Native elution buffer – nickel affinity chromatography

NaCl	17.5 g/L
NaH_2PO_4	6.9 g/L
Imidazole	17 g/L

Circular dichroism phosphate buffer

Na_2HPO_4	0.336 g/L
NaH_2PO_4	1.02 g/L

BIAcore running buffer

1× PBS	1 L
EDTA	29.22 mg/L

BIAcore regeneration solution 1

Glycine-HCl (pH 1.5)	1.11 g/L

BIAcore regeneration solution 2

NaOH	2 g/L

BIAcore regeneration solution 3

HEPES	2.38 g/L
EDTA	102.28 g/L
NaCl	8.77 g/L

Appendix B: Primers used in this study

Table B.1: List of primers used in this study

Primer name	Sequence(5'-3')
Sequencing	
M1noRE	AGAAAATTAAAAACAGGTACGGCAT
M2noRE	TTACCATCAACAGGGGGAAACA
Cloning/sequencing	
M4.2 pGEX2T F	GGGGGATCCGCGGAGATTAAAAAGCCTCAG
M4.2 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCGTTCTCTTTTGTTGCTTCAT
M90 pGEX2T F	GGGGGATCCGCGGAAGCGGTTAGTCGATCT
M90 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCGTTCTCTTTTGTTGTGTCAT
M106 pGEX2T F	GGGGGATCCACTACGGGCTCCCAGAAACAG
M106 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCGTTCTCTTTTGTTGCGTCAT
M58 pGEX2T F	GGGGGATCCGATTCTTCCAGAGAAGTAACC
M58 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCGTTCTCTTTTGTTGTGTCAT
M9.1 pGEX2T F	GGGGGATCCGAAGGGGTTAAGAAGGCGGAA
M9.1 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCGTTCTCTTTTGTTGCGTCAT
M102 pGEX2T F	GGGGGATCCGACAATCCGAGCTCTGTCCCT
M102 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCGTTCTCTTTTGTTGTGTCAT
M2 pGEX2T F	GGGGGATCCAACAGTAAGAACCCTGTCCCT
M2 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGTGTTCTCTTTTGTTGCGTCAT
M85 pGEX2T F	GGGGGATCCACTGAAGTTAAGGCTGCGGGG
M85 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCTGTCTCTTAGTTTCCTTCAT
M11 pGEX2T F	GGGGGATCCACTGAAGTTAAGGCTGCGGGG
M11 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCGTTCTCTTTTGTTGCGTCAT
M65 pGEX2T F	GGGGGATCCGATGGCCCCCAGAAAAGCGTT
M65 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCTGTCTCTTAGTTTCCTTCAT
M57 pGEX2T F	GGGGGATCCAATGACGATATTACTTCGATG
M57 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCTGTCTCTTAGTTTCCTTCAT
M54 pGEX2T F	GGGGGATCCGAAGTATTGACTAGGCGTCAG
M54 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCTGTCTCTTAGTTTCCTTCAT
M14 pGEX2T F	GGGGGATCCAGAGTTAGTAGGTCTATGTCA
M14 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCTGTCTCTTAGTTTCCTTCAT
M70 pGEX2T F	GGGGGATCCGAAGAGCATGAGAGCGTAACA
M70 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCTGTCTCTTAGTTTCCTTCAT
M53 pGEX2T F	GGGGGATCCAATAGAGCAGACGACGCTAGA
M53 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCTGTCTCTTAGTTTCCTTCAT
M98 pGEX2T F	GGGGGATCCGATAGATATACCGATGCTCAC
M98 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCTGTCTCTTAGTTTCCTTCAT
M1 pGEX2T F	GGGGGATCCAACGGTGATGGTAATCCTAGG
M1 pGEX2T R	GGGGAATTCTCAGTGATGGTGATGGTGATGCTGTCTCTTAGTTTCCTTCAT
M1 A-B pET-28b(+) F	GGGCCATGGTTAACGGTGATGGTAATCCTAGG
M1 A-B pET-28b(+) R	GGGCTCGAGTTCTAACTCTTTTTCTAAGAC
M1 A-C1 pET-28b(+) F	GGGCCATGGTTAACGGTGATGGTAATCCTAGG
M1 A-C1 pET-28b(+) R	GGGCTCGAGTAAATCTTTTTCAACCTGTTT
M1 A92-C1 pET-28b(+) F	GGGCCATGGTTGAAGAACTTGAAAAAGCAAAA
M1 A92-C1 pET-28b(+) R	GGGCTCGAGTAAATCTTTTTCAACCTGTTT
M1 B pET-28b(+) F	GGGCCATGGTTTGGGATAGACAAAGACTTGAA
M1 B pET-28b(+) R	GGGCTCGAGTAAATCTTTTTCAACCTGTTT
M1 B-C1 pET-28b(+) F	GGGCCATGGTTTGGGATAGACAAAGACTTGA
M1 B-C1 pET-28b(+) R	GGGCTCGAGTTCTAACTCTTTTTCTAAGAC

Appendix C: *emm* gene accession numbers and classification

Table C.1: *emm* genes used for phylogenetic analysis in this study along with respective accession numbers.

M-type	M-type (old nomenclature)	*emm* pattern	*emm*-cluster	Locus Accession number
1	1	A-C	A-C3	JX028599
2	2	E	E4	KC978826
3	3	A-C	A-C5	KC978816
4	4	E	E1	KC978806
5	5	A-C	Single protein *emm*-cluster clade Y	KC978827
6	6	A-C	Single protein *emm*-cluster clade Y	KC978835
8	8	E	E4	KC978796
9	9	E	E3	KC978828
11	11	E	E6	KC978833
12	12	A-C	A-C4	KC978829
13	13	E	E2	JX028611
14	14	A-C	Single protein *emm*-cluster clade Y	JX028612
15	15	E	E3	KC978775
17	17	A-C	Single protein *emm*-cluster clade Y	JX028614
18	18	A-C	Single protein *emm*-cluster clade Y	KC978771
19	19	A-C	Single protein *emm*-cluster clade Y	KC978837
22	22	E	E4	KC978795
23	23	A-C	Single protein *emm*-cluster clade Y	JX028618
24	24	A-C	Single protein *emm*-cluster clade Y	JX028619
25	25	E	E3	JX028620
26	26	A-C	Single protein *emm*-cluster clade Y	JX028621
27	27	E	E2	JX028622
28	28	E	E4	KC978790
29	29	A-C	Single protein *emm*-cluster clade Y	KC978834
30	30	A-C	A-C2	KC978842
31	31	nd	A-C5	KC978840
32	32	D	D2	JX028627
33	33	D	D4	JX028628
34	34	D	E5	JX472406
36	36	D	D1	JX028629
37	37	A-C	Single protein *emm*-cluster clade Y	JX028630
38	38/40	A-C	Single protein *emm*-cluster clade Y	JX028631
39	39	A-C	A-C4	JX028632
41	41	D	D4	KC978805
42	42	D	E6	KC978792

43	43	D	D4	KC978807
44	44/61	E	E3	KC978823
46	46	A-C	A-C1	JX028637
47	47	A-C	Single protein *emm*-cluster clade Y	JX028638
48	48	E	E6	KC978808
49	49	E	E3	KC978809
50	50/62	E	E2	JX028641
51	51	A-C	E5	JX028642
52	52	D	D4	JX028643
53	53	D	D4	KC978810
54	54	D	D1	JX028645
55	55	A-C	Single protein *emm*-cluster clade Y	KC978839
56	56	D	D4	JX028647
56.2	st3850	D	D4	JX028745
57	57	A-C	Single protein *emm*-cluster clade Y	JX028648
58	58	E	E3	KC978785
59	59	D	E6	KC978836
60	60	E	E1	KC978811
63	63	E	E6	KC978812
64	64	D	D4	KC978830
65	65/69	D	E6	KC978788
66	66	E	E2	KC978813
67	67	D	E6	KC978803
68	68	E	E2	KC978841
70	70	D	D4	JX028658
71	71	D	D2	KC978780
72	72	D	D4	JX028660
73	73	E	E4	KC978814
74	74	D	Single protein *emm*-cluster clade Y	KC978815
75	75	E	E6	KC978786
76	76	E	E2	KC978772
77	77	E	E4	KC978787
78	78	E	E1	KC978838
79	79	E	E3	JX028667
80	80	D	D4	JX028668
81	81	D	E6	KC978783
82	82	E	E3	KC978794
83	83	D	D4	KC978817
84	84	E	E4	JX028672
85	85	D	E6	JX028673
86	86	D	D4	JX028674
87	87	E	E3	KC978818
88	88	E	E4	JX028676

89	89	E	E4	KC978831
90	90	E	E2	JX028678
91	91	D	D4	JX028679
92	92	E	E2	KC978819
93	93	D	D4	KC978804
94	94	E	E6	KC978832
95	95	D	Single protein *emm*-cluster clade Y	KC978820
96	96	E	E2	JX028684
97	97	D	D5	KC978797
98	98	D	D4	KC978821
99	99	D	E6	JX028687
100	100	D	D2	JX028688
101	101	D	D4	KC978798
102	102	E	E4	KC978781
103	103	E	E3	KC978799
104	104	E	E2	JX028692
105	105	D	Single protein *emm*-cluster clade Y	KC978800
106	106	E	E2	KC978801
107	107	E	E3	JX028695
108	108	D	D4	KC978793
109	109	E	E4	JX028697
110	110	E	E2	KC978779
111	111	D	Single protein *emm*-cluster clade Y	JX028699
112	112	E	E4	KC978773
113	113	E	E3	JX028701
114	114	E	E4	KC978791
115	115	D	D2	JX028703
116	116	D	D4	KC978774
117	117	E	E2	JX028705
118	118	E	E3	KC978822
119	119	D	D4	JX028707
120	120	D	D4	JX028708
121	121	D	D4	JX028709
122	122	D	Single protein *emm*-cluster clade Y	KC978784
123	123	D	D3	KC978777
124	124	E	E4	JX028712
133	st1692	nd	A-C5	JX028730
134	st2105	D	E5	JX028734
137	st465	A-C	E5	JX028723
139	st7323	A-C	E6	JX028750
140	st7395	D	Single protein *emm*-cluster clade Y	JX028751
142	st818	A-C	A-C1	JX028726
144	stknb1	E	E3	JX028763

157	stn165	A-C	D5	JX028765
158	stxh1	D	E6	JX028772
163	st412	A-C	A-C3	JX028722
164	st106M	E	Single protein *emm*-cluster clade X	JX028716
165	st11014	E	E1	KC978789
166	st1207	E	E2	JX028728
168	st1389	E	E2	JX028729
169	st1731	E	E4	JX028731
170	st1815	REA	E5	KC978824
172	st2037	D	E6	JX028733
174	st211	REA	E5	JX028717
175	st212	E	E4	JX028718
176	st213	E	E1	JX028719
177	st2147	E	E6	JX028735
178	st22	nd	D4	JX028713
179	st221	D	Single protein *emm*-cluster clade Y	JX028720
180	st2460	E	E3	KC978778
182	st2861UK	D	E6	JX028737
183	st2904	E	E3	KC978825
184	st2911	D	D5	JX028739
185	st2917	D	Single protein *emm*-cluster clade X	JX028740
186	st2940	D	D4	KC978782
191	st369	D	E6	JX028721
192	st3757	D	D4	JX028743
193	st3765	A-C	A-C4	JX028744
194	st38	D	D4	JX028714
197	st4119	A-C	A-C2	JX028746
205	st5282	D	E5	JX028747
207	st6030	D	D1	KC978776
208	st62	D	D4	JX028715
209	st6735	E	E3	JX028749
211	st7406	E	Single protein *emm*-cluster clade X	JX028752
213	st7700	D	D2	JX028753
215	st804	E	Single protein *emm*-cluster clade Y	JX028724
217	st809	D	D3	JX028725
218	st854	D	Single protein *emm*-cluster clade Y	JX028727
219	st9505	nd	E3	JX028754
221	stck249	D	Single protein *emm*-cluster clade Y	JX028756
222	stck401	A-C	Single protein *emm*-cluster clade Y	JX028757
223	std432	D	D4	JX028758
224	std631	D	D4	JX028759
225	std633	D	D4	KC978802
227	stil103	nd	A-C3	JX028762

228	stil62	A-C	A-C4	JX028761
229	stmd216	A-C	A-C4	JX028764
230	stns1033	D	D4	JX028769
231	stns292	E	E3	JX028767
232	stns554	E	E4	JX028768
233	stns90	A-C	Single protein *emm*-cluster clade Y	JX028766
234	stpa57	nd	Single protein *emm*-cluster clade Y	JX028770
236	sts104	E	Single protein *emm*-cluster clade X	JX028771
238	1.2	A-C	A-C3	JX028600
239	1.4	A-C	A-C3	JX028601
242	st2926	D	D4	JX028741

REA, rearranged *emm* pattern (atypical amplification patterns). ND, not determined.

Appendix D: Vector construction and protein purification

Primer pairs used for all PCR reactions listed here are given in parentheses, and sequences are given in Appendix B.

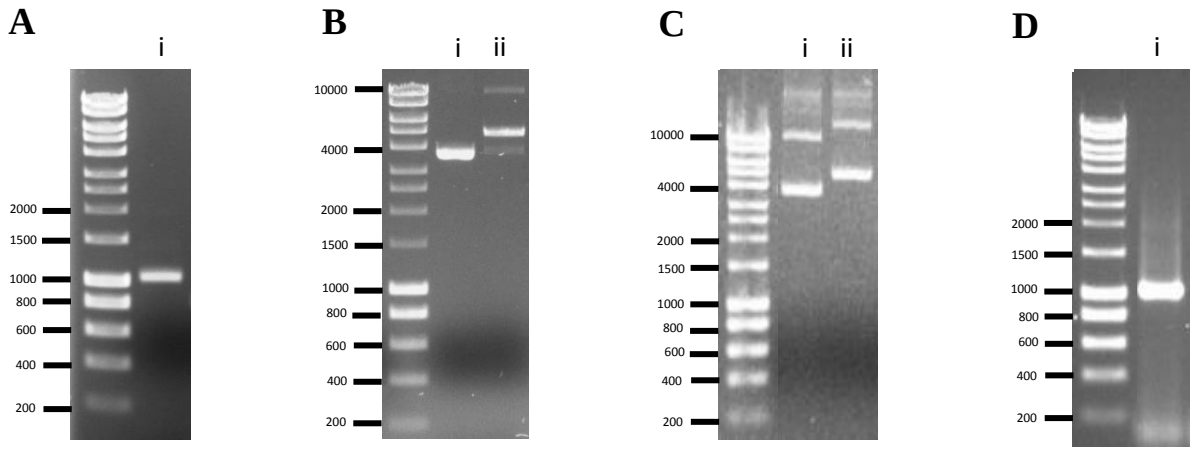


Figure D.1: Exemplar construction of *emm*-pGEX2T via ligation independent cloning. (A) PCR amplification of *emm* 85 gene (i) using M85 pGEX2T F and M85 pGEX2T R primers. (B) Agarose gel showing *EcoRI*/*BamHI* digested (i) and non-digested (ii) pGEX2T. (C) Representative agarose gel showing plasmid extraction of pGEX2T vector (i) and *emm*85 positive tranformant (ii). PCR amplification of *emm*85 gene from *emm*85 pGEX2T positive tranformant using M85 pGEX2T F and M85 pGEX2T R primers. Molecular size markers are from Hyperladder I (Bioline, Australia), with sizes indicated in base pairs.

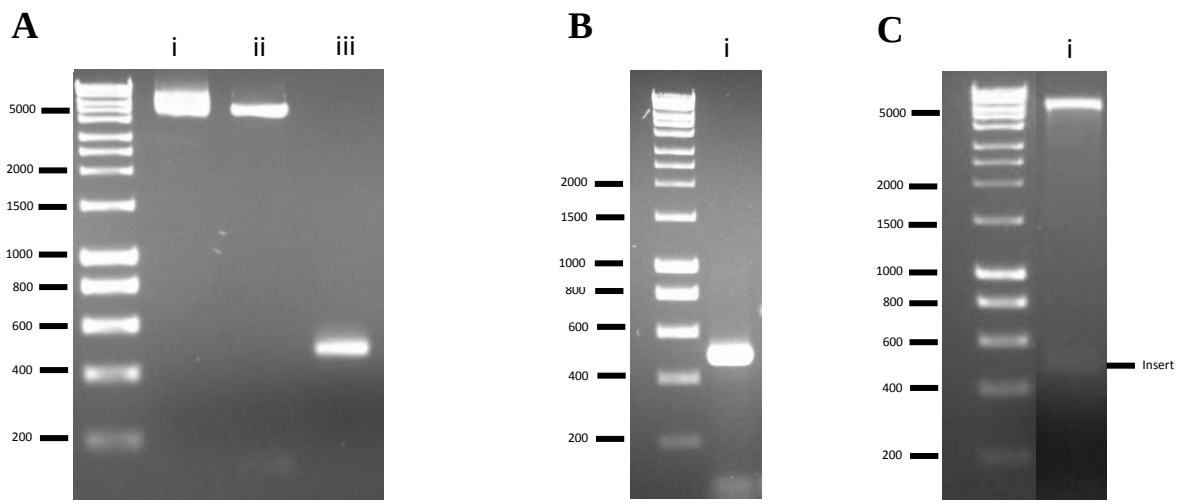


Figure D.2: Exemplar construction of *emm*1 fragment-pET-28b(+) via ligation independent cloning. (A) Agarose gel showing non-digested (i) and *Xho*I/*Nco*I enzyme digested (ii) pET-28b(+) along with PCR amplified A and B repeat coding sequence from *emm*1. (B) PCR amplification of *emm*1 A-B fragment gene from *emm*1 AB pET-28b(+) positive transformant. (C) Due to the low molecular weight of *emm*1 A-B, restricition enzyme digestion of *emm*1 AB pET-28b(+) was undertaken via *Xho*I/*Nco*I restriction enzyme digestion. PCR amplification was achived using M1 A-B pET-28b(+) F and M1 A-B pET-28b(+) R primers. Molecular size markers are from Hyperladder I (Bioline, Australia), with sizes indicated in base pairs.

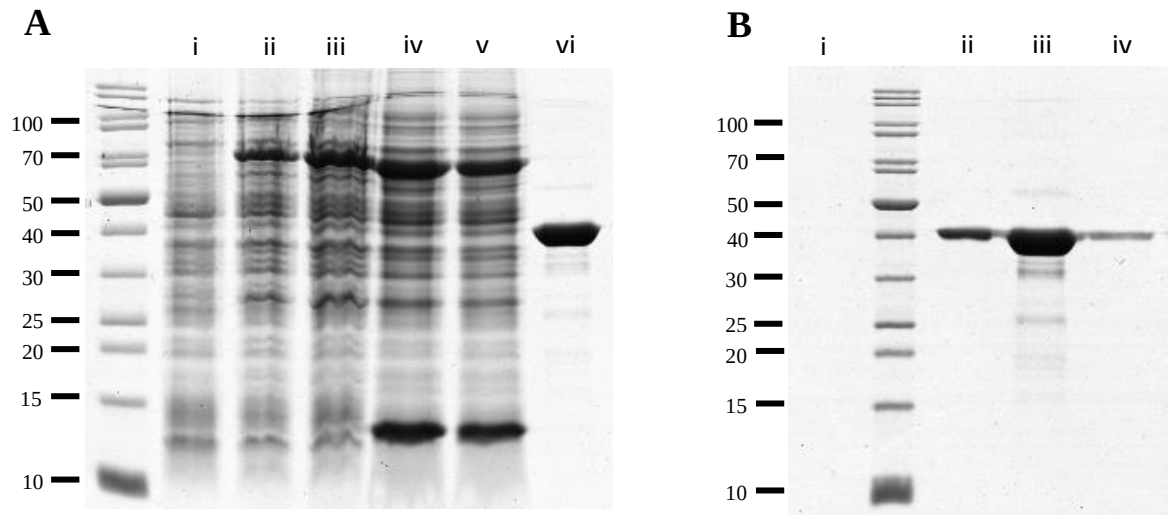


Figure D.3: Exemplar M protein expression and purification. (A) Glutathione sepharose chromatography showing representative expression and purification of the pGEX2T-*emm* 58 vector in Top10 *E. coli* with uninduced crude lysate (i) and culture induced with 0.1 mM IPTG for 2 h (ii) and 4 h (iii). Bacterial suspensions were pelleted and lysed in 50 mL lysis buffer and were incubated with glutathione sepharose resin for 30 min at 4˚C with end over end rotation, with pre-purification (iv) and post-purification (v) samples taken. M protein was released from the glutathione sepharose resin upon incubation with 100 U human thrombin and eluted with 1 × PBS (vi). (B) To facilitate the separation of human thrombin and M protein, samples were added in batch to 2 mL 50 % Nickel-NTA resin, incubated for 1 h at 4˚C with end over end rotation. After incubation, wash samples (i) and elution fractions (ii-iv) were collected. Results are shown from a 12% SDS-PAGE gel, Molecular weight markers are from Unstained Pageruler with weights indicated in kilo-Daltons.

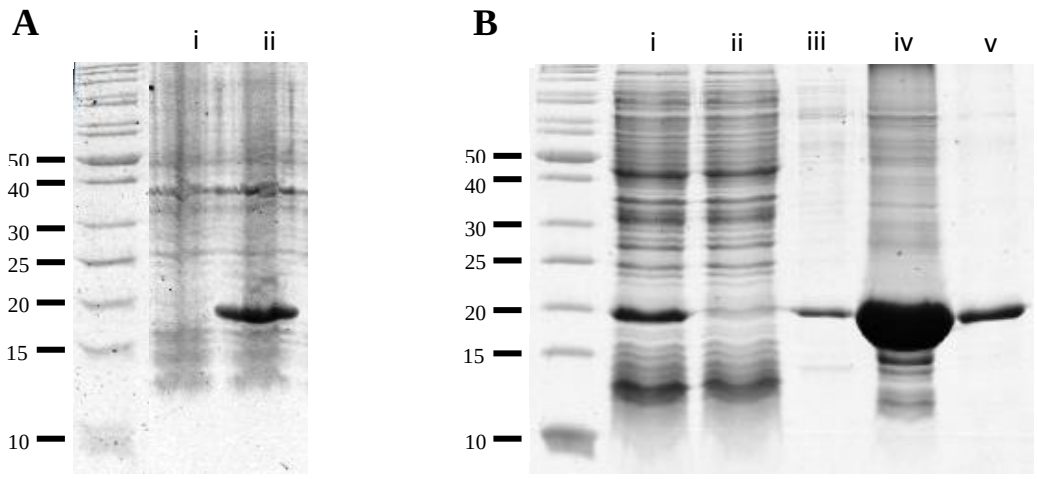
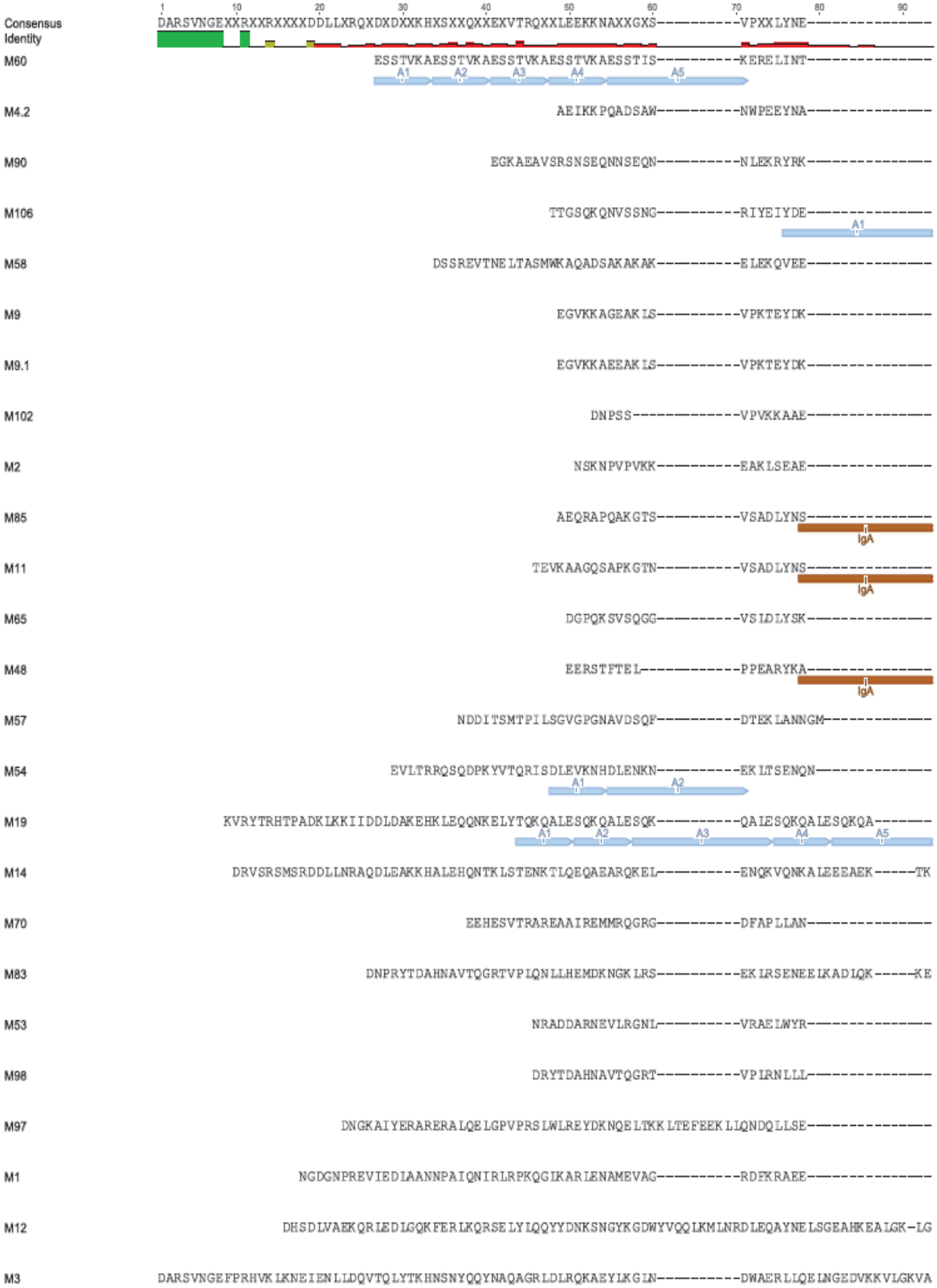
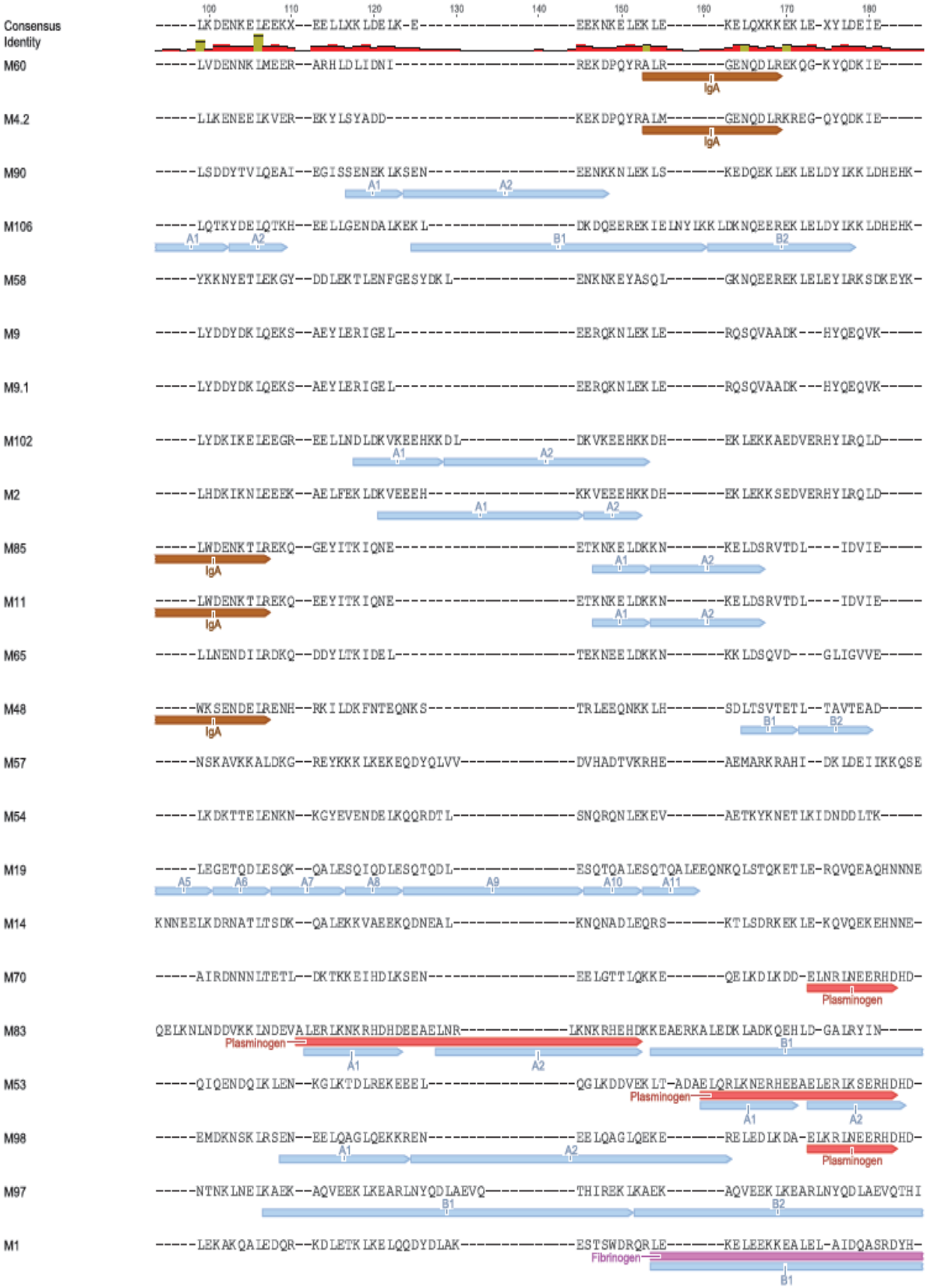
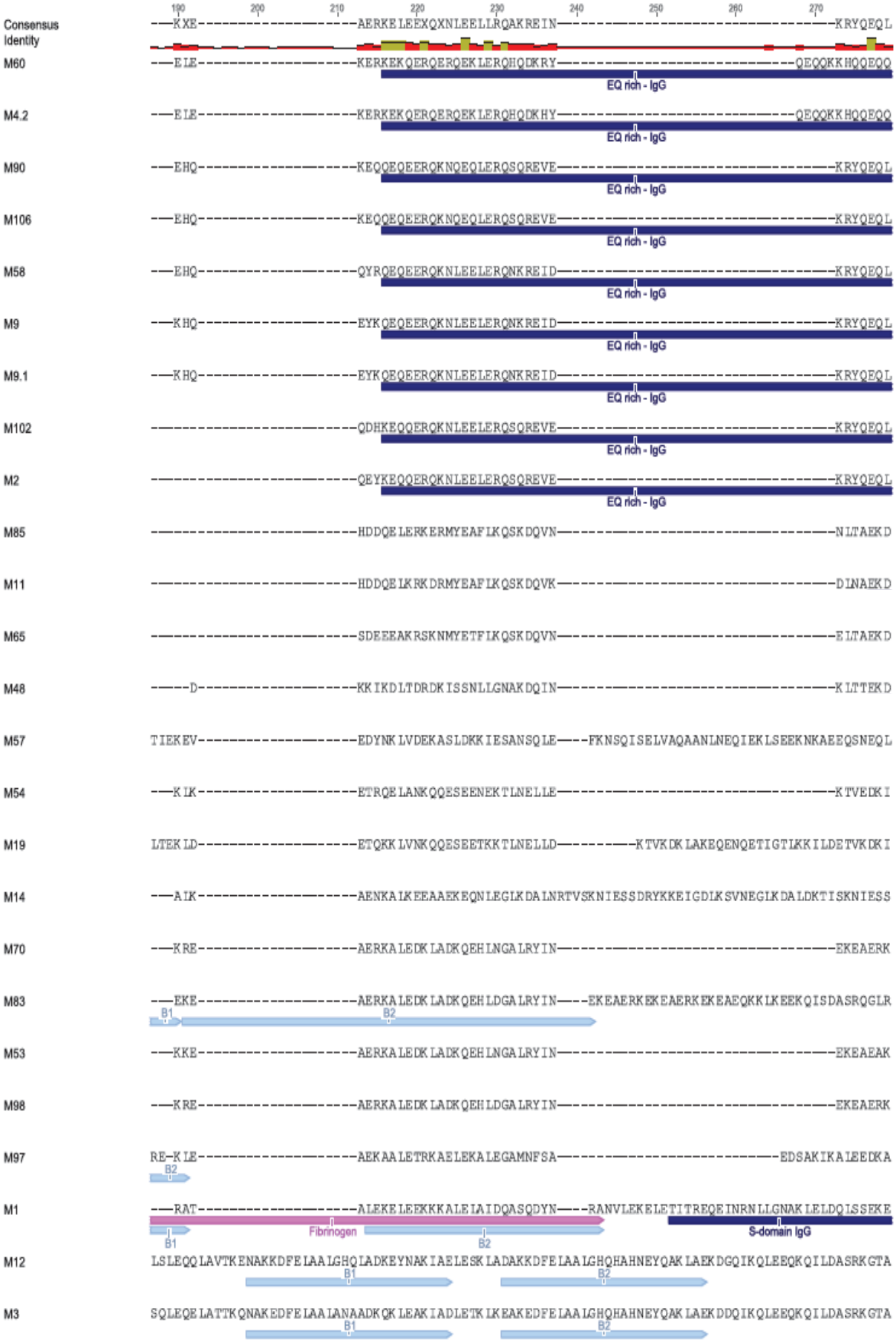


Figure D.4: Exemplar M1 fragment protein expression and purification. (A) Representative expression and purification of the pET-28b(+)-*emm*1 A-B vector in BL21(DE3) *E. coli* with uninduced crude lysate (i) and culture induced with 1 mM IPTG for 4 h (ii). (B) Bacterial suspensions were pelleted and lysed in 50 mL lysis buffer and batch incubated with 2 mL nickel-NTA resin for 1 h at 4˚C with end over end rotation with pre-purification (i) and post purification (ii) samples taken. Elutions of bound M1 A-B fragment were collected (iii-v). Results are shown from a 10% SDS-PAGE gel, Molecular weight markers are from Unstained Pageruler with weights indicated in kilo-Daltons.

Appendix E: M protein sequence analysis and binding affinity constants




Consensus
Identity
M60
M4.2
M90
M106
M58
M9
M9.1
M102
M2
M85
M11
M65
M48
M57
M54
M19
M14
M70
M83
M53
M98
M97
M1
M12
M3
IgA
A1
A2
B1
B2
Plasminogen
Fibrinogen


190
200
210
220
230
240
250
260
270
Consensus
Identity
M60
M4.2
M90
M106
M58
M9
M9.1
M102
M2
M85
M11
M65
M48
M57
M54
M19
M14
M70
M83
M53
M98
M97
M1
M12
M3
EQ rich - IgG
B1
B2
Fibrinogen
S-domain IgG


280
290
300
310
320
330
340
350
360
370
Consensus
Identity
M60
M4.2
M90
M106
M58
M9
M9.1
M102
M2
M85
M11
M65
M48
M57
M54
M19
M14
M70
M83
M53
M98
M97
M1
M12
M3
EQ rich - IgG
S-domain IgG
C1
Linker
C2
Albumin


380
390
400
410
420
430
440
450
462
Consensus
Identity
M60 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M4.2 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M90 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M106 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M58 LEASREAKKKVEADLAALNAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M9 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M9.1 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M102 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M2 ----------EHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M85 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M11 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M65 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M48 LEASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M57 LDASREAKKQVEKDLANLTAELDKVKEDKQISDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
M54 LNASREAKKQVEKDLANLTAELDKVKEEKQVSDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
M19 LDASREAKKQVEKDLANLTAELNKVKEEKQTSDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
M14 LDASREAKKQLE-------AEHQKLEEQNKISEASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
M70 LDASREAKKKVEADLAALTAEHQKIKEEKQISDASRQGLSRDLEASREAKKKVEADLAEANSKLQALEKINKELEEGKKLSEKEKAEIQA
M83 LDASREAKKQVEKDLANLTAELDKVKEEKQVSDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
M53 LDASREAKKQVEKGLANLTAELDKVKEEKQISDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
M98 LDASREAKKQVEKGLANLTAELDKVKEEKQISDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
M97 LDASREAKKQVEKGLANLTAELDKVKEEKQISDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
M1 LDASREAKKQVEKDLANLTAELDKVKEEKQISDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
M12 LDASREAKKQVEKDLANLTAELDKVKEEKQISDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
M3 LDASREAKKQVEKDLANLTAELDKVKEEKQISDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKINKELEESKKLTEKEKAEIQA
Albumin
Linker
C3
C2
C1

Figure E.1: Alignment 26 translated M protein amino acid sequences. M protein sequences were trimmed from the first codon/amino acid of the mature protein to the first codon/amino acid of the D repeat domain adjacent the LPXTG motif. Representational graphs highlighting the sequence identity between all M proteins are given at the beginning of each figure section. A, B and C repeat domains are indicated in blue. Characterised M protein IgA, IgG, plasminogen, fibrinogen, and albumin binding motifs are annotated within respective sequences. M54 amino acid sequences were only included once as M protein sequences from GAS strains NS178 and TVU5 are identical. Sequence alignment was undertaken with *Geneious* 6.0 using a MUSCLE alignment with default settings.

Table E.1: Binding raw data pertaining to the phylogenetic classification of full length M protein. Interactions were analysed by non-linear fitting of the single cycle kinetic sensograms according to a 1:1 Langmuir binding model using Biacore T200 evaluation software (Biacore AB).

GAS strain	*emm*-type	*emm* pattern	Major Clade	*emm*-cluster	Plg (K_D)	IgA (K_D)	IgG (K_D)	Fg. (K_D)	Alb. (K_D)	C4BP (K_D)
PRS20	60	E	X	E1	NB	0.84 ± 0.04 nM	20.07 ± 9.70 nM	NB	6.62 ± 0.14 nM	7.09 ± 3.75 pM
NS226	4.2	E	X	E1	NB	5.36 ± 0.20 nM	12.95 ± 0.45 nM	NB	4.18 ± 0.12 nM	9.81 ± 4.76 pM
NS730	90	E	X	E2	NB	NB	10.96 ± 1.29 nM	NB	7.54 ± 0.45 nM	NB
NS192	106	E	X	E2	NB	NB	18.54 ± 0.53 nM	NB	9.33 ± 0.85 nM	NB
PRS18	58	E	X	E3	NB	NB	5.75 ± 0.22 nM	NB	8.76 ± 0.59 nM	5.93 ± 2.40 pM
PRS55	9	E	X	E3	NB	NB	4.17 ± 0.66 nM	NB	5.93 ± 0.21nM	4.70 ± 1.59 pM
NS179	9.1	E	X	E3	NB	NB	6.45 ± 0.95 nM	NB	6.45 ± 0.25 nM	5.18 ± 1.10 pM
PRS66	102	E	X	E4	NB	NB	82.69 ± 13.87nM	NB	NB	NB
PRS2	2	E	X	E4	NB	NB	29.76 ± 7.33 nM	NB	NB	45.42 ± 8.62 pM
NS8	85	D	X	E6	NB	1.73 ± 0.78 nM	2.06 ± 0.11 nM	NB	3.89 ± 0.16 nM	6.99 ± 1.22 pM
NS414	11	E	X	E6	NB	1.77 ± 0.09 nM	10.44 ± 0.78 nM	NB	18.14 ± 0.51 nM	119.93 ± 23.13 pM
NS931	65	D	X	E6	NB	NB	4.51 ± 0.18 nM	NB	6.38 ± 0.01 nM	5.10 ± 1.28 pM
PRS15	48	E	X	E6	NB	0.66 ± 0.02 nM	3.11 ± 0.08 nM	NB	11.67 ± 0.22 nM	7.21 ± 1.70 pM
NS1140	57	A-C	Y1	M57	NB	NB	7.18 ± 0.17 nM	0.10 ± 0.01 nM	4.68 ± 0.23 nM	NB
NS178	54	A-C and D	Y1	D1	NB	NB	NB	0.11 ± 0.02 nM	2.24 ± 0.16 nM	NB
TVU5	54	A-C and D	Y1	D1	NB	NB	NB	0.09 ± 0.01 nM	4.74 ± 0.14 nM	NB
PRS9	19	A-C	Y1	M19	NB	NB	NB	0.64 ± 0.04 nM	3.02 ± 0.15 nM	NB
NS501	14	A-C	Y1	M14	NB	NB	18.64 ± 0.69 nM	0.45 ± 0.07 nM	2.17 ± 0.03 nM	NB
NS80	70	D	Y2	D4	3.06 ± 0.37 nM	NB	NB	NB	4.43 ± 1.08 nM	NB
PRS30	83	D	Y2	D4	1.66 ± 0.31 nM	NB	NB	NB	5.94 ±0.15 nM	NB
NS13	53	D	Y2	D4	2.19 ± 0.73 nM	NB	NB	NB	NB	NB
NS88.2	98	D	Y2	D4	1.33 ± 0.32 nM	NB	NB	NB	NB	NB
88/30	97	D	Y2	D5	NB	NB	NB	0.45 ± 0.06 nM	4.82 ± 0.06 nM	NB
NS696	1	A-C	Y2	AC3	NB	NB	5.53 ±0.13nM	0.15 ± 0.03 nM	4.36 ± 0.43 nM	NB
PRS8	12	A-C	Y2	AC4	NB	NB	NB	0.20 ± 0.01 nM	2.77 ± 0.09 nM	NB
M3	3	A-C	Y2	AC5	NB	NB	NB	0.20 ± 0.08 nM	2.86 ± 0.05 nM	NB

NB: Non-binder. Plg.; Fg.; and Alb.; stands for plasminogen, fibrinogen and albumin respectively. No difference in binding phenotype was observed between two different isolates from both M9 (PRS55 and NS179) and M54 (NS178 and TVU5).

Table E.2: α-SV1 competitive binding affinity data. Competitive binding analysis between M protein and α-SV1 IgG was assessed using host proteins plasminogen, fibrinogen, albumin, IgG and IgA. Interactions were analysed by non-linear fitting of the single cycle kinetic sensograms according to a 1:1 Langmuir binding model using Biacore T200 evaluation software (Biacore AB).

GAS strain	*emm-type*	*emm* cluster	α-SV1 (K_D - nM)	Host protein - α-SV1 (K_D - nM)	α-SV1 – Host protein (K_D - nM)	Host protein (K_D - nM)
IgA						
PRS20	M60	E1	142.3 ± 29.8	n.b. *	0.35 ± 0.05	0.84 ± 0.04
NS226	M4.2	E1	129.4 ± 2.80	n.b. *	0.90 ± 0.02	5.36 ± 0.20
NS8	M85	E6	36.27 ± 9.85	48.76 ± 4.44	2.22 ± 0.27	1.73 ± 0.78
NS414	M11	E6	45.74 ± 0.27	34.52 ± 10.73	1.41 ± 0.07	1.77 ± 0.09
PRS15	M48	E6	14.76 ± 1.20	18.65 ± 3.09	0.56 ± 0.03	0.66 ± 0.02
IgG						
PRS20	M60	E1	60.14 ± 6.55	50.86 ± 5.29	33.35 ± 9.29	20.07 ± 9.70
NS226	M4.2	E1	56.12 ± 5.99	56.60 ± 10.51	24.6 ± 1.32	12.95 ± 0.45
PRS18	M58	E3	8.62 ± 1.02	10.29 ± 0.95	6.64 ± 0.54	5.75 ± 0.22
PRS55	M9	E3	14.54 ± 2.04	26.79 ± 6.50	6.29 ± 0.43	4.17 ± 0.66
NS179	M9.1	E3	49.63 ± 4.25	31.15 ± 5.85	8.45 ± 1.02	6.45 ± 0.95
PRS66	M102	E4	34.62 ± 2.52	71.18 ± 1.31	69.35 ± 2.99	82.69 ± 13.87
PRS2	M2	E4	14.84 ± 0.49	29.52 ± 2.32	21.81 ± 4.64	29.67 ± 7.33
NS8	M85	E6	6.82 ± 0.51	51.24 ± 33.47	9.22 ± 2.66	2.06 ± 0.11
NS414	M11	E6	45.74 ± 0.27	36.88 ± 15.28	75.45 ± 16.40	10.44 ± 0.78
NS931	M65	E6	10.78 ± 0.85	22.62 ± 13.24	9.14 ± 0.44	4.51 ± 0.18
PRS15	M48	E6	14.76 ± 1.20	61.64 ± 13.10	102.63 ± 39.81	3.11 ± 0.08
NS1140	M57	M57	61.65 ± 1.35	64.81 ± 5.52	9.21 ± 1.24	7.18 ± 0.17
NS501	M14	M14	86.04 ± 7.27	59.85 ± 20.39	49.71 ± 14.32	18.64 ± 0.69
NS696	M1	A-C3	10.24 ± 1.07	39.62 ± 3.20	8.04 ± 1.60	5.53 ± 0.13
Plasminogen						
NS13	M53	D4	28.59 ± 3.91	31.23 ± 2.35	39.29 ± 5.61	26.43 ± 2.62
NS88.2	M98	D4	96.12 ± 5.88	83.18 ± 13.29	11.10 ± 3.72	15.61 ± 2.10
PRS30	M83	D4	102.22 ± 6.38	91.84 ± 6.32	24.10 ± 3.51	22.90 ± 3.14
NS80	M70	D4	108.75 ± 3.65	106.95 ± 3.15	17.00 ± 1.97	15.35 ± 1.84
Fibrinogen						
NS1140	M57	M57	61.65 ± 1.35	112.85 ± 2.65	0.06 ± 0.01	0.10 ± 0.01
NS178	M54	D1	87.13 ± 6.14	76.58 ± 3.95	0.25 ± 0.04	0.11 ± 0.02
TVU5	M54	D1	84.54 ± 7.42	88.86 ± 4.86	0.09 ± 0.01	0.12 ± 0.03
PRS9	M19	M19	74.86 ± 3.87	92.29 ± 6.43	0.64 ± 0.04	0.64 ± 0.04
NS501	M14	M14	86.04 ± 7.27	129.15 ± 3.60	0.28 ± 0.02	0.45 ± 0.07
88/30	M97	D5	61.20 ± 5.93	64.84 ± 3.15	0.13 ± 0.01	0.45 ± 0.06
NS696	M1	A-C3	10.24 ± 1.07	10.87 ± 0.50	0.61 ± 0.05	0.15 ± 0.03
PRS8	M12	A-C4	98.78 ± 4.13	105.23 ± 8.45	1.50 ± 0.18	0.20 ± 0.01
SSI-1	M3	A-C5	97.87 ± 3.45	116.35 ± 6.55	0.17 ± 0.03	0.20 ± 0.08
Albumin						
PRS20	M60	E1	60.14 ± 6.55	60.68 ± 6.81	3.97 ± 0.09	6.62 ± 0.14
NS226	M4.2	E1	56.12 ± 5.99	50.7 ± 2.05	10.79 ± 3.06	4.18 ± 0.12
NS730	M90	E2	45.94 ± 5.94	64.12 ± 5.38	8.36 ± 2.14	7.54 ± 0.45
NS192	M106	E2	57.81 ± 9.99	71.36 ± 7.66	17.98 ± 1.15	9.33 ± 0.85
PRS18	M58	E3	8.62 ± 1.02	9.10 ± 1.30	4.70 ± 0.26	8.76 ± 0.59
PRS55	M9	E3	14.54 ± 2.04	26.92 ± 1.60	7.31 ± 0.36	5.93 ± 0.21
NS179	M9.1	E3	13.46 ± 0.80	28.38 ± 1.66	9.15 ± 0.27	6.45 ± 0.25
NS8	M85	E6	6.82 ± 0.51	32.24 ± 1.97	70.25 ± 2.85	3.89 ± 0.16

NS414	M11	E6	45.74 ± 0.27	34.09 ± 2.81	13.99 ± 1.37	18.14 ± 0.51
NS931	M65	E6	10.78 ± 0.85	9.42 ± 0.72	4.78 ± 0.10	6.38 ± 0.01
PRS15	M48	E6	14.76 ± 1.20	12.18 ± 0.47	4.45 ± 0.17	11.67 ± 0.22
NS1140	M57	M57	61.65 ± 1.35	58.56 ± 4.86	4.85 ± 2.19	4.68 ± 0.23
NS178	M54	D1	87.13 ± 6.14	89.45 ± 10.31	2.63 ± 0.11	2.24 ± 0.16
TVU5	M54	D1	60.48 ± 3.90	58.68 ± 2.91	3.89 ± 0.32	4.74 ± 0.14
PRS9	M19	M19	74.86 ± 3.87	68.34 ± 5.78	1.32 ± 0.14	3.02 ± 0.15
NS501	M14	M14	86.04 ± 7.27	98.76 ± 10.09	1.37 ± 0.07	2.17 ± 0.03
NS80	M70	D4	108.75 ± 3.65	103.56 ± 5.02	4.25 ± 0.18	4.43 ± 1.08
PRS30	M83	D4	102.22 ± 6.38	158 ± 26.36	4.49 ± 0.16	5.94 ± 0.15
88/30	M97	D5	61.20 ± 5.93	73.34 ± 3.69	3.54 ± 0.14	4.82 ± 0.06
NS696	M1	A-C3	10.24 ± 1.07	13.69 ± 0.85	3.87 ± 1.21	4.36 ± 0.43
PRS8	M12	A-C4	98.78 ± 4.13	73.25 ± 7.20	1.69 ± 0.10	2.77 ± 0.09
SSI-1	M3	A-C5	97.87 ± 3.45	80.02 ± 4.70	2.82 ± 0.07	2.86 ± 0.05

NB: Non-binder. No difference in binding phenotype was observed between two different isolates from both M9 (PRS55 and NS179) and M54 (NS178 and TVU5).

Appendix F: PAM – GI- GII-plasminogen interactions

Table F.1: Summary of PAM – GI- GII-plasminogen binding affinities

Analyte/condition	K_D (nM)
GI-Plg	37.0 ± 6.6
GI-Plg+ 5 mM ε	33.2 ± 4.4
GI-Plg + 100 mM εACA	Nb*
GI-Plg + 5 mM BZ	20.0 ± 6.3
GI-Plg + 5 mM BZ/εACA	54.5 ± 12.2
GII-Plg	27.4 ± 1.6
GII-Plg + 5 mM εACA	2.8 ± 0.6
GII-Plg + 100 mM εACA	16.7 ± 7.0
GII-Plg + 400 mM εACA	94.1 ± 4.7
GII-Plg + 5 mM BZ	20.8 ± 1.6
GII-Plg + 5 mM BZ/εACA	2.2 ± 0.1
GI-Plg – GII-Plg	Nb*
GII-Plg – GI-Plg	Nb*
GI-Plg – GII-Plg + 5 mM εACA	5.1 ± 1.9
GII-Plg – GI-Plg + 5 mM εACA	Nb*
Lys-Plg	0.44 ± 0.1
Lys-Plg + 10 mM εACA	3.8 ± 0.1
GII-Plg KR1-3	14.6 ± 1.6
GII-Plg KR1-3 + 0.5 mM εACA	28.2 ± 2.1
GII-Plg KR1-3 + 1 mM εACA	101.8 ± 9.6
GII-Plg KR1-3 + 5 mM εACA	Nb*

NB: Non-binder; binding fit could not be determined.

Appendix G: M protein-glycan array analysis

Glycan structures present on the glycan array are given in table G.1

From the 26 M proteins expressed in study, 20 were chosen for glycan microarray analysis representing 20 different M-types from 11 major cluster groups. Positive binding was determined when binding fluorescence was above the mean background fluorescence (defined as greater than one fold increase of the mean background plus 3 standard deviations). Statistical analysis of the data was performed by a Student's t-test with a confidence level of 99.99% ($p \leq 0.0001$) and only glycans that met these criteria for three biologically independent samples (n = 12 glycans spot replicates) were interpreted as positive binding interactions (Fig. G1-20).

Table G.1: Glycan structures present on the array

Index	Name	Structure
	Terminal Galactose	
1	α-Galactose	Galα
2	β-Galactose	Galβ
3	Galα1-2Galβ	Galα1-2Galβ
4	B^{di}	Galα1-3Galβ
5	$T_{\alpha\beta}$	Galα1-3GalNAcβ
6	$T_{\alpha\alpha}$	Galα1-3GalNAcα
7	-	Galα1-3GlcNAcβ
8	α-LacNAc	Galα1-4GlcNAcβ
9	Melibiose	Galα1-6Glcβ
10	B^{di}	Galβ1-2Galβ
11	Le^c	Galβ1-3GlcNaAcβ
12	-	Galβ1-3Galβ
13	$T_{\beta\beta}$	Galβ1-3GalNAcβ
14	TF	Galβ1-3GalNAcα
15	Lac	Galβ1-4Glcβ
16	Fs-2	GalNAcα1-3GalNAcβ
17	A_{di}	GalNAcα1-3Galβ
18	core 5	GalNAcα1-3GalNAcα
19	LacdiNAc	GalNAcβ1-4GlcNAcβ
20	Ab2aLN	Galβ1-2Galα1-4GlcNAcβ
21	Ab3`LN	Galβ1-3Galβ1-4GlcNAcβ
22	LN3Tn	Galβ1-4GlcNAcβ1-3GalNAcα
23	LN6Tn	Galβ1-4GlcNAcβ1-6GalNAcα
24	GA_2, $GgOse_3$	GalNAcβ1-4Galβ1-4Glcβ
25	Tbb-A	Galβ1-3GalNAcβ1-3Gal

26	Ab4`LN	Galβ1-4Galβ1-4GlcNAc
27	Galili (tetra)	Galα1-3Galβ1-4GlcNAcβ1-3Galβ
28	-	Galα1-4GlcNAcβ1-3Galβ1-4GlcNAcβ
29	-	Galβ1-3GlcNAcβ1-3Galβ1-3GlcNAcβ
30	-	Galβ1-3GlcNAcα1-3Galβ1-4GlcNAcβ
31	-	Galβ1-3GlcNAcβ1-3Galβ1-4GlcNAcβ
32	-	Galβ1-3GlcNAcα1-6Galβ1-4GlcNAcβ
33	-	Galβ1-3GlcNAcβ1-6Galβ1-4GlcNAcβ
34	Asialo-GM1	Galβ1-3GalNAcβ1-4Galβ1-4Glcβ
35	LN2	Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ
36	-	Galβ1-4GlcNAcβ1-6Galβ1-4GlcNAcβ
37	H (type IV)	Galβ1-4GlcNAcβ1-6(Galβ1-3)GalNAcα
38	Gb4, P	GalNacβ1-3GalαGalβ1-4Glcβ
39	Le^c3Le^c	Galβ1-3GlcNAcβ1-3Galβ1-3GlcNAcβ
40	Galili (penta)	Galα1-3Galβ1-4GlcNAcβ1-3Galβ1-4Glcβ
41	-	Galβ1-4GlcNAcβ1-6(Galβ1-4GlcNAcβ1-3)GalNAcα
42	-	Galβ1-4GlcNAcβ1-3(GlcNAcβ1-6)Galβ1-4GlcNAcβ
43	-	Galβ1-4GlcNAcβ1-6(GlcNAcβ1-3)Galβ1-4GlcNAcβ
44	$(LN)_3$	$(Galβ1-4GlcNAcβ1-3)_3$
45	-	Galβ1-4GlcNAcβ1-6(Galβ1-4GlcNAcβ1-3)Galβ1-4GlcNAc
46	Gb5	Galβ1-3GalNAcβ1-3Galα1-4Galβ1-4Glcβ
47	Lacto-*N*-Biose I	Galβ1-3GlcNAc
48	N-Acetyllactosamine	Galβ1-4GlcNAc
49	β1-4galactosyl-galactose	Galβ1-4Gal
50	β1-6galactosyl-*N*-acetylglucosamine	Galβ1-6GlcNAc
51	β1-3galactosyl-*N*-acetylgalactosamine	Galβ1-3GlcNAc
52	Galb1-3GalNAcβ1-4Galβ1-4Glc	
53	Lacto-*N*-tetraose	Galβ1-3GlcNAcβ1-3Galβ1-4Glc
54	Lacto-*N*-neotetraose	Galβ1-4GlcNAcβ1-3Galβ1-4Glc

55	Lacto-*N*-neohexaose	Galβ1-4GlcNAcβ1-6(Galβ1-4GlcNAcβ1-3)Galβ1-4Glc
56	Lacto-N-hexaose	Galβ1-4GlcNAcβ1-6(Galβ1-3GlcNAcβ1-3)Galβ1-4Glc
57	Globotriose	Galα1-4Galβ1-4Glc
58	Tn Antigen	GalNAcα1-O-Ser
59	Galactosyl-Tn Antigen	Galβ1-3GalNAcα1-O-Ser
60	α1-3 Galactobiose	Galα1-3Gal
61	Linear B-2 Trisaccharide	Galα1-3Galβ1-4GlcNAc
62	Linear B-6 Trisaccharide	Galα1-3Galβ1-4Glc
63	α1-3, β1-4, α1-3 Galactotetraose	Galα1-3Galβ1-4Galα1-3Gal
64	β1-6Galactobiose	Galβ1-6Gal
65	Terminal disaccharide of globotriose	GalNAcβ1-3Gal
66	Receptor for pili of *P. aeruginosa*	GalNAcβ1-4Gal
67	P1 Antigen	Galα1-4Galβ1-4GlcNAc
68	α-D-*N*-acetylgalactosaminyl-1-3Gal-β1-4Glc	GalNAcα1-3Galβ1-4Glc
69	iso-Lacto-*N*-octaose	Galβ1-3GlcNAcβ1-3Galβ1-4GlcNAcβ1-6(Galβ1-3GlcNAcβ1-3)Galβ1-4Glc
	Terminal glucose structures	
70	α-Glc	Glcα
71	β-Glc	Glcβ
72	β-GlcN(Gc)	GlcN(Gc)β
73	aminoglucitol	$HOCH_2(HOCH)_4CH_2NH_2$
74	α-glucoronic acid	GlcAα
75	β-glucoronic acid	GlcAβ
76	maltose	Glcα1-4Glcβ
77	cellobiose	Glcβ1-4Glcβ
78	gentiobiose	Glcβ1-6Glcβ
79	GUb3GN	GlcAβ1-3GlcNAcβ
80	GUb3A	GlcAβ1-3Galβ
81	GUb6A	GlcAβ1-6Galβ

82	maltotriose	$(Glc\alpha 1\text{-}4)_3\beta$
83	isomaltotriose	$(Glc\alpha 1\text{-}6)_3\beta$
84	maltotetraose	$(Glc\alpha 1\text{-}4)_4\beta$
85	isomaltotetraose	$(Glc\alpha 1\text{-}6)_4\beta$
86	isomaltopentaose	$(Glc\alpha 1\text{-}6)_5\beta$
87	maltohexaose	$(Glc\alpha 1\text{-}6)_6\beta$
	Terminal *N*`Acetyl glucosamine	
88	N-Acetyl-*D*-glucosamine α	GalNAcα
89	N-Acetyl-*D*-glucosamine β	GlcNAcβ
90	core 3	GlcNAcβ1-3GalNAcα
91	chitobiose	GlcNAcβ1-4GlcNAcβ
92	core 6	GlcNAcβ1-6GalNAcα
93	GlcNAc-Mur	GlcNAcβ1-4-[HOOC(CH_3)CH]-3-O-GlcNAcβ
94	GMDP-Lys	GlcNAcβ1-[HOOC(CH_3)CH]-3-O-GlcNAcβ-L-alanyl-D-i-glutaminyl-L-lysine
95	GN2`TF	GlcNAcβ1-2Galβ1-3GalNAcα
96	GN3`TF	GlcNAcβ1-3Galβ1-3GalNAcα
97	GN3`Lac	GlcNAcβ1-3Galβ1-4Glcβ
98	GN3`LN	GlcNAcβ1-3Galβ1-4GlcNAcβ
99	GN4`LN	GlcNAcβ1-4Galβ1-4GlcNAcβ
100	GN6`LN	GlcNAcβ1-6Galβ1-4GlcNAcβ
101	core 2	GlcNAcβ1-6(Galβ1-3)GalNAcα
102	Core 4	GlcNAcβ1-6(GlcNAcβ1-3)GalNAcα
103	Tk	GlcNAcβ1-6(GlcNAcβ1-3)Galβ1-4GlcNAcβ
104	chitopentaose	$(GlcNAc\beta 1\text{-}4)_5\beta$
105	chitohexaose	$(GlcNAc\beta 1\text{-}4)_6\beta$
106	*N*,*N'*-Diacetyl chitobiose	GlcNAcβ1-4GlcNAc
107	*N*,*N'*,*N''*-Triacetyl chitotriose	GlcNAcβ1-4GlcNAcβ1-4GlcNAc
108	*N*,*N'*,*N''*,*N'''*-Tetraacetyl chitotetraose	GlcNAcβ1-4GlcNAcβ1-4GlcNAcβ1-4GlcNAc
109	*N*,*N'*,*N''*,*N'''*,*N''''*,*N'''''*-Hexaacetyl	GlcNAcβ1-4GlcNAcβ1-4GlcNAcβ1-4GlcNAcβ1-

	chitohexaose	4GlcNAcβ1-4GlcNAc
110	Bacterial cell wall muramyl disaccharide	GlcNAcβ1-4MurNAc
	Mannose containing structures	
111	Manα-sp3	Manα
112	Manβ-sp4	Manβ
113	ManNAcβ-sp4	ManNAcβ
114	α-Man6P	6-H_2PO_3Manα
115	GN3M	GlcNAcβ1-3Manβ
116	Mb4GN	Manβ1-4GlcNAcβ
117	β1-2-*N*-Acetylglucosamine-mannose	GlcNAcβ1-2Man
118	Bianntennary *N*-linked core pentasaccharide	GlcNAcβ1-2Manα1-6(GlcNAcβ1-2Manα1-3)Man
119	α1-2-Mannobiose	Manα1-2Man
120	α1-3-Mannobiose	Manα1-3Man
121	α1-4-Mannobiose	Manα1-4Man
122	α1-6-Mannobiose	Manα1-6Man
123	α1-3, a1-6-Mannobiose	Manα1-6(Manα1-3)Man
124	α1-3, α1-3, α1-6-Mannopentaose	Manα1-6(Manα1-3)Manα1-6(Manα1-3)Man
	Fucosylated structures	
125	L-α-Fuc	Fucα
126	Fa3GN	Fucα1-3GlcNAcβ
127	LeC	Fucα1-4GlcNAcβ
128	H (type 2)	Fucα1-2Galβ1-4GlcNAcβ
129	H (type 3)	Fucα1-2Galβ1-3GalNAcα
130	H (type 6)	Fucα1-2Galβ1-4Glcβ
131	B_{tri}	Galα1-3(Fucα1-2)Galβ
132	B (type I)	Galα1-3(Fucα1-2)Galβ1-3GlcNAcβ
133	B (type II)	Galα1-3(Fucα1-2)Galβ1-4GlcNAcβ
134	B (type III)	Galα1-3(Fucα1-2)Galβ1-3GalNAcα
135	B (type IV)	Galα1-3(Fucα1-2)Galβ1-3GalNAcβ

136	αGalLe^{x}	Galα1-3Galβ1-4(Fucα1-3)GlcNAcβ
137	A (type 1)	GalNAcα1-3(Fucα1-2)Galβ1-3GlcNAcβ
138	A (type 2)	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ
139	A (type 3)	GalNAcα1-6(Fucα1-2)Galβ1-3GalNAcα
140	H (type1) penta	Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4GlcNAcβ
141	BLe^{Y}	Galα1-3(Galβ1-4)Fucα1-2(Fucα1-3) GlcNAcβ
142	Le^{b}-Lac	Fucα1-4(Fucα1-2Galβ1-3)GlcNAcβ1-3Galβ1-4Glcβ
143	Le^{y}-Lac	Fucα1-2Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4Glcβ
144	MFLNH III	Le^{x}1-6'(Le^{c}1-3')Lac
145	MFLNH I	LacNAc1-6'(Le^{d}1-3')Lac
146	MSMFLNnH	Le^{x}1-6'(6'SLN1-3')Lac
147	DFLNH (a)	Le^{x}1-6'(Le^{d}1-3')Lac
148	MF(1-3)iLNO	Le^{c}Le^{x}1-6'(Le^{c}1-3')Lac
149	TFLNH	Le^{x}1-6'(Le^{b}1-3')Lac
150	Lacto-*N*-fucopentaose I	Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glc
151	Lacto-*N*-fucopentaose II	Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4Glc
152	Lacto-*N*-fucopentaose III	Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4Glc
153	Lacto-*N*-difucohexaose I	Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4Glc
154	Lacto-*N*-difucohexaose II	Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4(Fucα1-3)Glc
155	H-disaccharide	Fucα1-2Gal
156	2'-Fucosyllactose	Fucα1-2Galβ1-4Glc
157	3'-Fucosyllactose	Galβ1-4(Fucα1-3)Glc
158	Lewis^{x}	Galβ1-4(Fucα1-3)GlcNAc
159	Lewis^{a}	Galβ1-3(Fucα1-4)GlcNAc
160	Blood Group A-trisaccharide	GalNAcα1-3(Fucα1-2)Gal
161	Lactodifucotetraose	Fucα1-2Galβ1-4(Fucα1-3)Glc
162	Blood Group B-Trisaccharide	Galβ1-3(Fucα1-2)Gal
163	Lewis^{y}	Fucα1-2Galβ1-4(Fucα1-3)GlcNAc
164	Blood Group H Type II Trisaccharide	Fucα1-2Galβ1-3GlcNAc
165	Lewis^{b} tetrasaccharide	Fucα1-2Galβ1-3(Fucα1-4)GlcNAc

166	Sulpho Lewis[a]	SO_3-3Galβ1-3(Fucα1-4)GlcNAc
167	Sulpho Lewis[x]	SO_3-3Galβ1-4(Fucα1-3)GlcNAc
168	Monofucosyl-para-Lacto-*N*-hexaose IV	Galβ1-3GlcNAcβ1-3Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4Glc
169	Monofucosyllacto-*N*-hexaose III	Galβ1-4(Fucα1-3)GlcNAcβ1-6(Galβ1-3GlcNAcβ1-3)Galβ1-4Glc
170	Difucosyllacto-*N*-hexaose	Galβ1-4(Fucα1-3)GlcNAcβ1-6(Fucα1-2Galβ1-3GlcNAcβ1-3)Galβ1-4Glc
171	Trifucosyllacto-*N*-hexaose	Galβ1-4(Fucα1-3)GlcNAcβ1-6(Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ1-3)Galβ1-4Glc
	Neu5Ac containing structures	
172	Sia	Neu5Acα
173	Sia-Bn	Neu5Acα
174	aNeuGc	Neu5Gcα
175	9NAcSia	9-NAc-Neu5Acα
176	GM4	Neu5Acα2-3Galβ
177	Sia6A	Neu5Acα2-6Galβ
178	3-SiaTn	Neu5Acα2-3GalNAcα
179	SiaTn	Neu5Acα2-6GalNAcα
180	Neu5GCTn	Neu5Gcα2-6GalNAcα
181	$(Sia)_2$	Neu5Acα2-8Neu5Acα2
182	6SiaANb	Neu5Acα2-6GalNAcβ
183	Neu5Gc3A	Neu5Gcα2-3Gal
184	6-SiaTF	Neu5Acα2-6(Galα1-3)GalNAcα
185	A3a(Sia)Tn	Neu5Acα2-6(Galβ1-3)GalNAcα
186	3`-Sia-TF	Neu5Acα2-3Galβ1-3GalNAcα
187	3`-SiaLe[c]	Neu5Acα2-3Galβ1-3GlcNAcβ
188	3`SLN(Gc)	Neu5Gcα2-3Galβ1-4GlcNAcβ
189	6`SLN(Gc)	Neu5Gcα2-6Galβ1-4GlcNAcβ
190	9NAc-6`SLN	9-NAc-Neu5Acα2-6Galβ1-4GlcNAcβ
191	3`SLN6Su	Neu5Acα2-3Galβ1-4-(6-O-Su)GlcNAcβ

192	3`SiaTF6Su	Neu5Acα2-3Galβ1-3-(6-O-Su)GalNAcβ
193	6`SLN6Su	Neu5Acα2-6Galβ1-4-(6-O-Su)GlcNAcβ
194	3`SLN6`Su	Neu5Acα2-3-(6-O-Su)Galβ1-4GlcNAcβ
195	(Sia)3	$(Neu5Ac\alpha2\text{-}8)_3$
196	6`SiaLeC	Neu5Acα2-6Galβ1-3GlcNAc
197	6`SiaLeC6Su	Neu5Acα2-6Galβ1-3(6-O-Su)GlcNAc
198	3`SiaLeC(GC)	Neu5Gcα2-3Galβ1-3GlcNAcβ
199	GM2	GalNAcβ1-4(Neu5Acα2-3)Galβ1-4Glcβ
200	3`SLNb3A	Neu5Acα2-3Galβ1-4GlcNAcβ1-3Galβ
201	SiaLeX6Su	Neu5Acα2-3Galβ1-4(Fucα1-3)6-O-Su-GlcNAcβ
202	SiaLeX6`Su	Neu5Acα2-3(6-O-Su)Galβ1-4(Fucα1-3)GlcNAcβ
203	Sia2-3`,6TF	Neu5Acα2-6(Neu5Acα2-3Galβ1-3)GalNAcα
204	GC3	Neu5Acα2-8Neu5Acα2-3Galβ1-4Glcβ
205	3`SLN-LacNAc	Neu5Acα2-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ
206	$SiaLe^x$-3Gal	Neu5Acα2-3 Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ
207	LSTb	Neu5Acα2-6(Galβ1-3)GlcNAcβ1-3Galβ1-4Glcβ
208	GD2	GalNAcβ1-4(Neu5Acα2-8Neu5Acα2-3)Galβ1-4Glc
209	GT3	Neu5Acα2-8Neu5Acα2-8Neu5Acα2-3Galβ1-4Glc
210	GT2	$GalNAc\beta1\text{-}4(Neu5Ac\alpha2\text{-}8)_2Neu5Ac\alpha2\text{-}3Gal\beta1\text{-}4Glc$
211	6`SLN-LacNAc	Neu5Acα2-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ
212	LSTa	Neu5Acα2-3Galβ1-4GlcNAcβ1-3Galβ1-4Glcβ
213	Sialyl $Lewis^a$	Neu5Acα2-3Galβ1-3(Fucα1-4)GlcNAc
214	Sialyl $Lewis^x$	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAc
215	Sialyllacto-*N*-tetraose a	Neu5Acα2-3Galβ1-3GlcNAcβ1-3Galβ1-4Glc
216	Monosialyl, monofucosyllacto-*N*-neohexaose	Galβ1-4(Fucα1-3)GlcNAcβ1-6(Neu5Acα2-6Galβ1-4GlcNAcβ1-3)Galb1-4Glc
217	-	Neu5Acα2-3Galβ1-3(Neu5Acα2-6)GalNAc
218	-	Neu5Acα2-6Galβ1-3GlcNAcβ1-3Galβ1-4(Fucα1-3)Glc
219	2,3'-Sialyllactosamine	Neu5Acα2-3Galβ1-4GlcNAc
220	2,6'-Sialyllactosamine	Neu5Acα2-6Galβ1-4GlcNAc

221	LS-Tetrasaccharide a	Neu5Acα2-3Galβ1-3GlcNAcβ1-3Galβ1-4Glc
222	LS-Tetrasaccharide b	Galβ1-3(Neu5Acα2-6)GlcNAcβ1-3Galβ1-4Glc
223	LS-Tetrasaccharide c	Neu5Acα2-6Galβ1-4GlcNAcβ1-3Galβ1-4Glc
224	Disialyllacto-*N*-tetraose	Neu5Acα2-3Galβ1-3(Neu5Acα2-6)GlcNAcβ1-3Galβ1-4Glc
225	2,3'-Sialyllactose	Neu5Acα2-3Galβ1-4Glc
226	2,6'-Sialyllactose	Neu5Acα2-6Galβ1-4Glc
227	Colominic acid	(Neu5Acα2-8Neu5Ac)n (n<50)
228	Biantennary 2,6-sialylated-*N*-glycan-Asn	Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-6(Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAc-Asn
	Sulphate containing glycans	
229	3-O-Su-Galβ-sp3	3-O-Su-Galβ
230	3-O-Su-GalNAcα-sp3	3-O-Su-GalNAcα
231	6-O-Su-GlcNAcβ-sp3	6-O-Su-GlcNAcβ
232	GN3Su	3-O-Su-GlcNAcβ
233	LeC6Su	Galβ1-3(6-O-Su)GlcNAcβ
234	Lac6Su	Galβ1-4(6-O-Su)Glcβ
235	LN6Su	Galβ1-4(6-O-Su)GlcNAcβ
236	Ch2-6Su	GlcNAcβ1-4(6-O-Su)GlcNAcβ
237	TF3`Su	3-O-Su-Galβ1-3GalNAcα
238	TF6`Su	6-O-Su-Galβ1-3GalNAc
239	Lac3`Su	3-O-Su-Galβ1-4Glcβ
240	Lac6`Su	6-O-Su-Galβ1-4Glcβ
241	LeC3`Su	3-O-Su-Galβ1-3GlcNAcβ
242	LN3`Su	3-O-Su-Galβ1-4GlcNAcβ
243	LN4`Su	4-O-Su-Galβ1-4GlcNAcβ
244	LeC6`Su	6-O-Su-Galβ1-3GlcNAcβ
245	LN6`Su	6-O-Su-Galβ1-4GlcNAcβ
246	Lac3`,6Su2	3-O-Su-Galβ1-4(6-O-Su)Glcβ

247	LN3`6Su2	3-O-Su-Galβ1-4(6-O-Su)GlcNAcβ
248	Lac6,6`Su2	6-O-Su-Galβ1-4(6-O-Su)Glcβ
249	LeC6,6`Su2	6-O-Su-Galβ1-3(6-O-Su)GlcNAcβ
250	LN66`Su2	6-O-Su-Galβ1-4(6-O-Su)GlcNAcβ
251	LN3`4`Su2	3,4-O-Su_2-Galβ1-4GlcNAcβ
252	LN3`6`Su2	3,6-O-Su_2-Galβ1-4GlcNAcβ
253	LN4`6`Su2	4,6-O-Su_2-Galβ1-4GlcNAcβ
254	LN4`6`Su2-C3	4,6-O-Su_2-Galβ1-4GlcNAcβ
255	LN3`66`Su3	3,6-O-Su_2-Galβ1-4(6-O-Su)GlcNAcβ
256	LacdiNAc6Su	GalNAcβ1-4(6-O-Su)GlcNAcβ
257	LacdiNAc3`Su	3-O-Su-GalNAcβ1-4GlcNAcβ
258	LacdiNAc6`Su	6-O-Su-GalNAcβ1-4GlcNAcβ
259	3Ac-LacdiNAc6`Su	6-O-Su-GalNAcβ1-4-(3-O-Su)GlcNAcβ
260	LacdiNAc3,3`Su2	3-O-Su-GalNAcβ1-4(3-O-Su)-GlcNAcβ
261	LacdiNAc3`,6`Su2	3,6-O-Su_2-GalNAcβ1-4GlcNAcβ
262	LacdiNAc4`,6`Su2	4,6-O-Su_2-GalNAcβ1-4GlcNAcβ
263	3Ac-LacdiNAc4`,6`Su2	4,6-O-Su_2-GalNAcβ1-4-(3-O-Ac)GlcNAcβ
264	LacdiNAc4`Su	4-O-Su-GalNAcβ1-4GlcNAcβ
265	LacdiNAc3`,4`Su2	3,4-O-Su_2-Galβ1-4GlcNAcβ
266	LacdiNAc6,6`Su2	6-O-Su-GalNAcβ1-4(6-O-Su)GlcNAcβ
267	LN6Su	Galβ1-4(6-O-Su)GlcNAcβ
268	LacdiNAc4`Su-C2	4-O-Su-GalNAcβ1-4GlcNAcβ
269	(3`SuLN)3`LN	3-O-SuGalβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ
270	(4`SuLN)3`LN	4-O-SuGalβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ
	Carrageenan and Glycosaminoglycans	
271	hyaluronic acid	(GlcAβ1-4GlcNAcβ1-3)$_8$-NH_2-ol
272	11-OS, YDS	(Sia2-6A-GN-M)$_2$-3,6-M-GN-GNβ
273	Neocarratetraose-41, 3-di-*O*-sulphate (Na^+)	$C_{24}H_{36}O_{25}S_2Na_2$ (Mixed anomers. Tetrasaccharide of regular κ - carrageenan)
274	Neocarratetraose-41-*O*-sulphate (Na^+)	$C_{24}H_{37}O_{22}SNa$ (Mixed anomers. Derived from C1003

		by removal of the non-reducing terminal 4-sulphate)
275	Neocarrahexaose-24,41, 3, 5-tetra-*O*-sulphate (Na^+)	$C_{36}H_{52}O_{40}S_4Na_4$ (Mixed anomers. A hybrid sequence comprising carrageenan disaccharides in the order κ -i-κ, derived from the carrageenan from *Chondrus crispus*)
276	Neocarrahexaose-41, 3, 5-tri-*O*-sulphate (Na^+)	$C_{36}H_{53}O_{37}S_3Na_3$ (Mixed anomers. Hexasaccharide of regular κ -carrageenan)
277	Neocarraoctaose-41, 3, 5, 7-tetra-*O*-sulphate (Na^+)	$C_{48}H_{70}O_{49}S_4Na_4$ (Mixed anomers. Octasaccharide of regular κ -carrageenan)
278	Neocarradecaose-41, 3, 5, 7, 9-penta-*O*-sulphate (Na^+)	$C_{60}H_{87}O_{61}S_5Na_5$ (Mixed anomers. Decasaccharide of regular κ - carrageenan)
279	ΔUA-2S → GlcNS-6S Na_4 (I-S)	$C_{12}H_{15}NO_{19}S_3Na_4$ (Predominant disaccharide produced from heparin by heparinase I and II)
280	ΔUA → GlucNS-6S Na_3 (II-S)	$C_{12}H_{16}NO_{16}S_2Na_3$ (Produced from heparinase II digestion of heparin and heparin sulphate)
281	ΔUA → 2S-GlcNS Na_3 (III-S)	$C_{12}H_{16}NO_{16}S_2Na_3$ (Produced from heparin by digestion with heparinase I and II)
282	ΔUA → 2S-GlcNAc-6S Na_3 (I-A)	$C_{14}H_{18}NO_{17}S_2Na_3$ (Minor component produced from heparin by heparinase II)
283	ΔUA → GlcNAc-6S Na_2 (II-A)	$C_{14}H_{19}NO_{14}SNa_2$ (Product of the action of heparinases II and III on heparin and heparan sulphate)
284	ΔUA → 2S-GlcNAc Na_2 (III-A)	$C_{14}H_{19}NO_{14}SNa_2$ (Minor product of the action of heparinase II on heparin)
285	ΔUA → GlcNAc Na (IV-A)	$C_{14}H_{20}NO_{11}Na$ (Produced from heparin sulphate by digestion With heparinase III)
286	ΔUA → GalNAc-4S Na_2 (Δ Di-4S)	$C_{14}H_{19}NO_{14}SNa_2$ (Produced from various chondroitin sulphates By the action of chondroitinases ABC, B and AC-1)
287	ΔUA → GalNAc-6S Na2 (Δ Di-6S)	C14H19NO14SNa2 (Produced from various chondroitin sulphates By the action of chondroitinases ABC, AC-1 and C)
288	ΔUA → GalNAc-4S,6S Na_3 (Δ Di-disE)	$C_{14}H_{18}NO_{17}S_2Na_3$ (Produced from various chondroitin sulphates By the action of chondroitinases ABC, B and AC-1)
289	ΔUA → 2S-GalNAc-4S Na_2 (Δ Di-disB)	$C_{14}H_{18}NO_{17}S_2Na_3$ (Produced from various chondroitin sulphates by action of chondroitinase ABC and/or B. Most typically from chondroitin sulphate B (dermatan sulphate)
290	ΔUA → 2S-GalNAc-6S Na_3 (Δ Di-disD)	$C_{14}H_{18}NO_{17}S_2Na_3$ (Produced from various chondroitin sulphates by the action of chondroitinase ABC)

291	ΔUA → 2S-GalNAc-4S-6S Na_4 (Δ Di-tisS)	$C_{14}H_{17}NO_{20}S_3Na_4$ (Produced as a minor component by the action of chondroitinase ABC on various chondroitin sulphates, particularly B)
292	ΔUA → 2S-GalNAc-6S Na_2 (ΔDi-UA2S)	$C_{14}H_{19}NO_{14}SNa_2$ (Produced as a minor component from various chondroitin sulphates by the action of chondroitinase ABC)
293	ΔUA → GlcNAc Na (Δ Di-HA)	$C_{14}H_{20}NO_{11}Na$ (The only unsaturated disaccharide produced from hyaluronic acid by the action of chondroitinase ABC or AC-1)
294	Hyaluronan fragments (4mer)	(GlcAβ1-3GlcNAcβ1-4)n (n=4)
295	Hyaluronan fragments (8mer)	(GlcAβ1-3GlcNAcβ1-4)n (n=8)
296	Hyaluronan fragments (10mer)	(GlcAβ1-3GlcNAcβ1-4)n (n=10)
297	Hyaluronan fragments (12mer)	(GlcAβ1-3GlcNAcβ1-4)n (n=12)
298	Heparin	(GlcA/IdoAα/β1-4GlcNAcα1-4)n (n=200)
299	Chondroitin sulfate	(GlcA/IdoAβ1-3(±4/6S)GalNAcβ1-4)n (n<250)
300	Dermatan sulfate	((±2S)GlcA/IdoAα/b1-3(±4S)GalNAcβ1-4)n (n<250)
301	Chondroitin 6-Sulfate	(GlcA/IdoAβ1-3(±6S)GalNAcβ1-4)n (n<250)
302	HA-4 (10mM)	(GlcAβ1-3GlcNAcβ1-4)n (n=4)
303	HA-6 (10mM)	(GlcAβ1-3GlcNAcβ1-4)n (n=6)
304	HA-8 (9.7mM)	(GlcAβ1-3GlcNAcβ1-4)n (n=8)
305	HA-10 (7.83mM)	(GlcAβ1-3GlcNAcβ1-4)n (n=10)
306	HA-12 (6.5mM)	(GlcAβ1-3GlcNAcβ1-4)n (n=12)
307	HA-14 (5.6mM)	(GlcAβ1-3GlcNAcβ1-4)n (n=14)
308	HA-16 (4.9mM)	(GlcAβ1-3GlcNAcβ1-4)n (n=16)
309	HA-30,000da 2.5mg/ml	(GlcAβ1-3Glc*N*Ac β1-4)n
310	HA-107,000da 2.5mg/ml	(GlcAβ1-3Glc*N*Ac β1-4)n
311	HA-190,000da 2.5mg/ml	(GlcAβ1-3Glc*N*Ac β1-4)n
312	HA-220,000da 2.5mg/ml	(GlcAβ1-3Glc*N*Ac β1-4)n
313	HA-1,600,000da 2.5mg/ml	(GlcAβ1-3Glc*N*Ac β1-4)n
314	Heparin Sulfate 5mg/ml	(GlcA/IdoAα/ IdoASα/β1-4Glc*N*Ac/GlcNS/Glc*N*Ac6Sα1-4)n
315	Beta-1-3Glucan	(Glcβ1-3Glcβ1-3)n

Non-categorised glycans

316	Rhamnose	Rhaα
317	G6P	6-H_2PO_3Glcβ
318	9-OS	$(A\text{-}GN\text{-}M)_2$-3,6-M-GN-GNβ
319	7-OS	$(GN\text{-}M)_2$-3,6-M-GN-GNβ

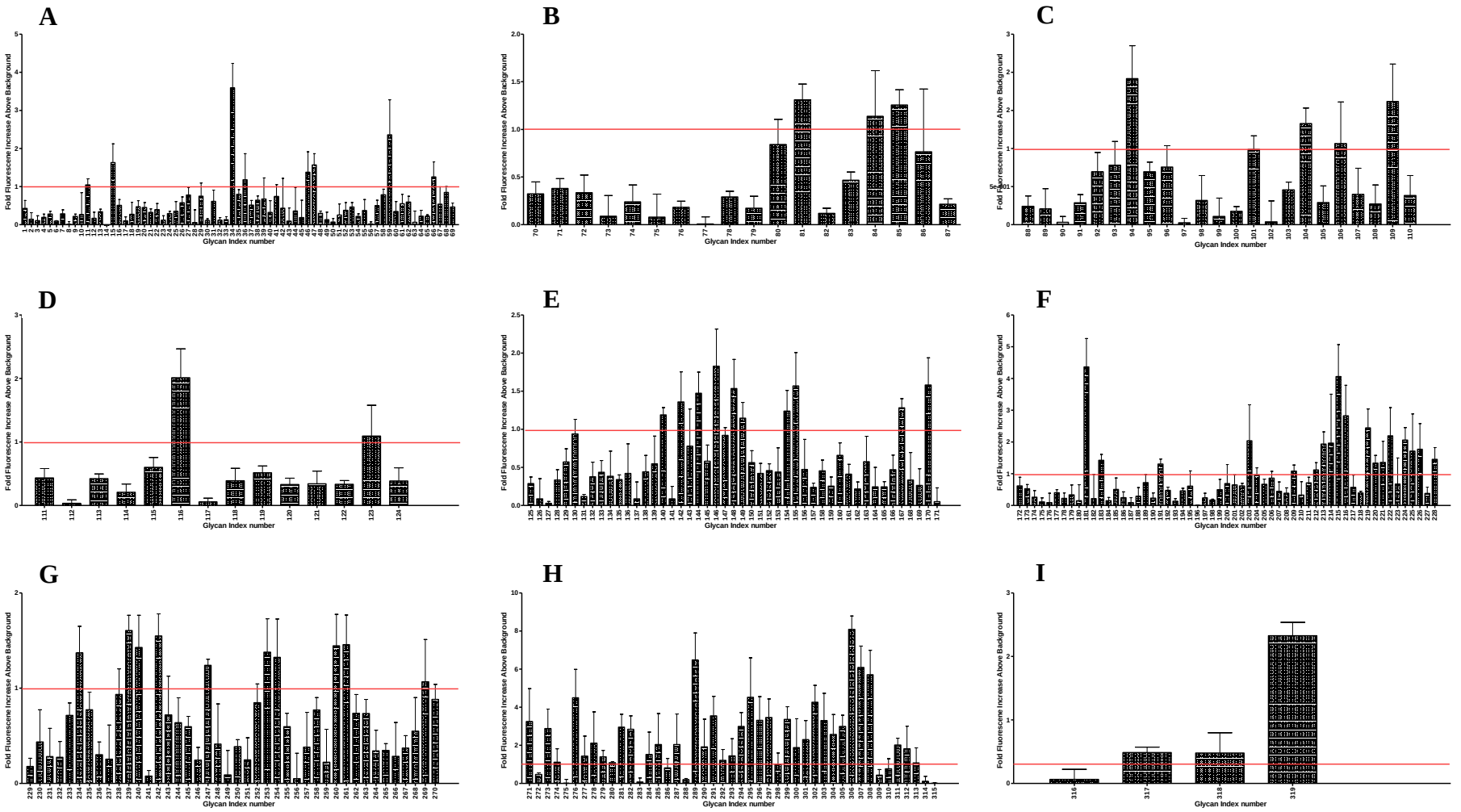


Figure G.1: M60 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures**.**

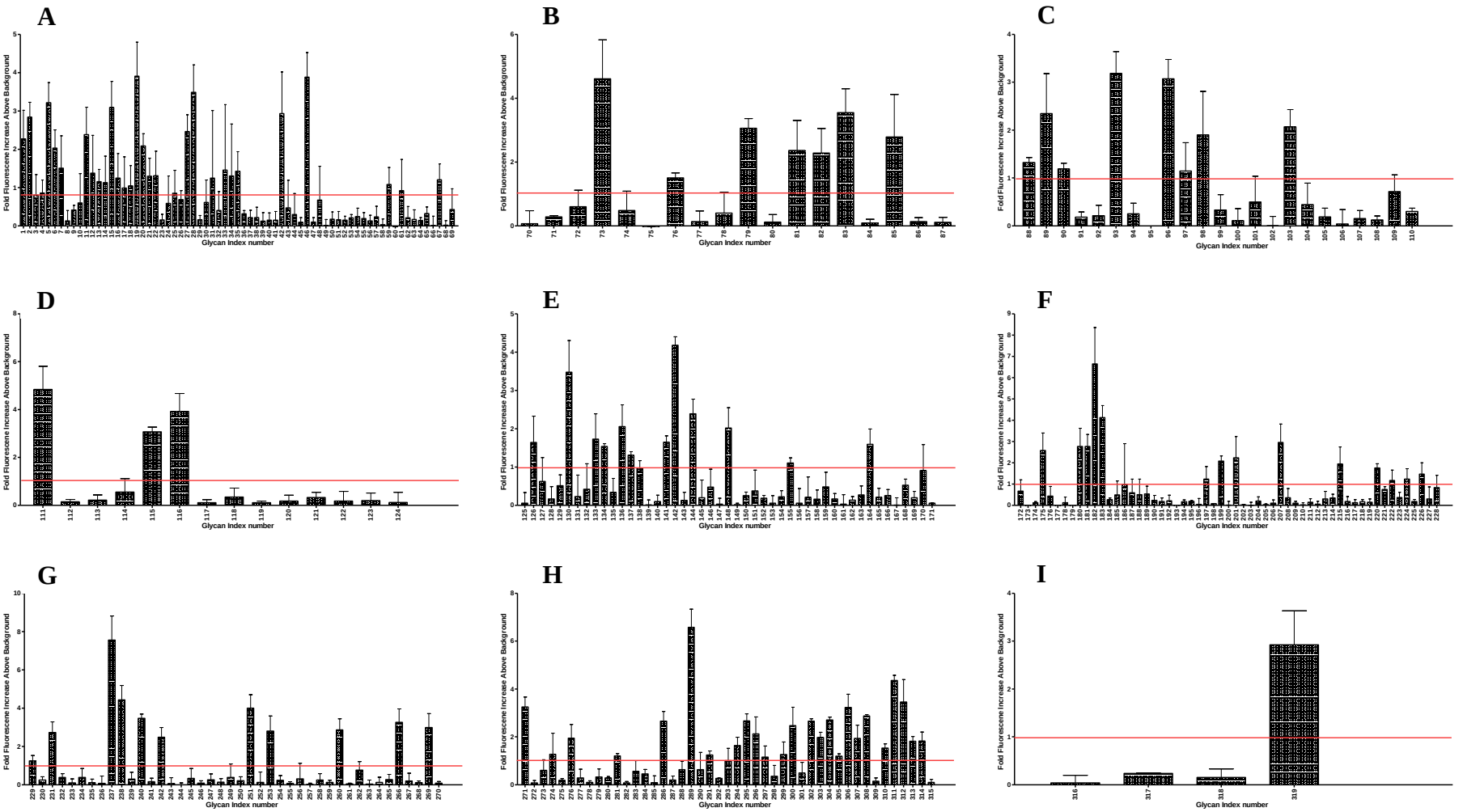


Figure G.2: M90 binding glycan profileof (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

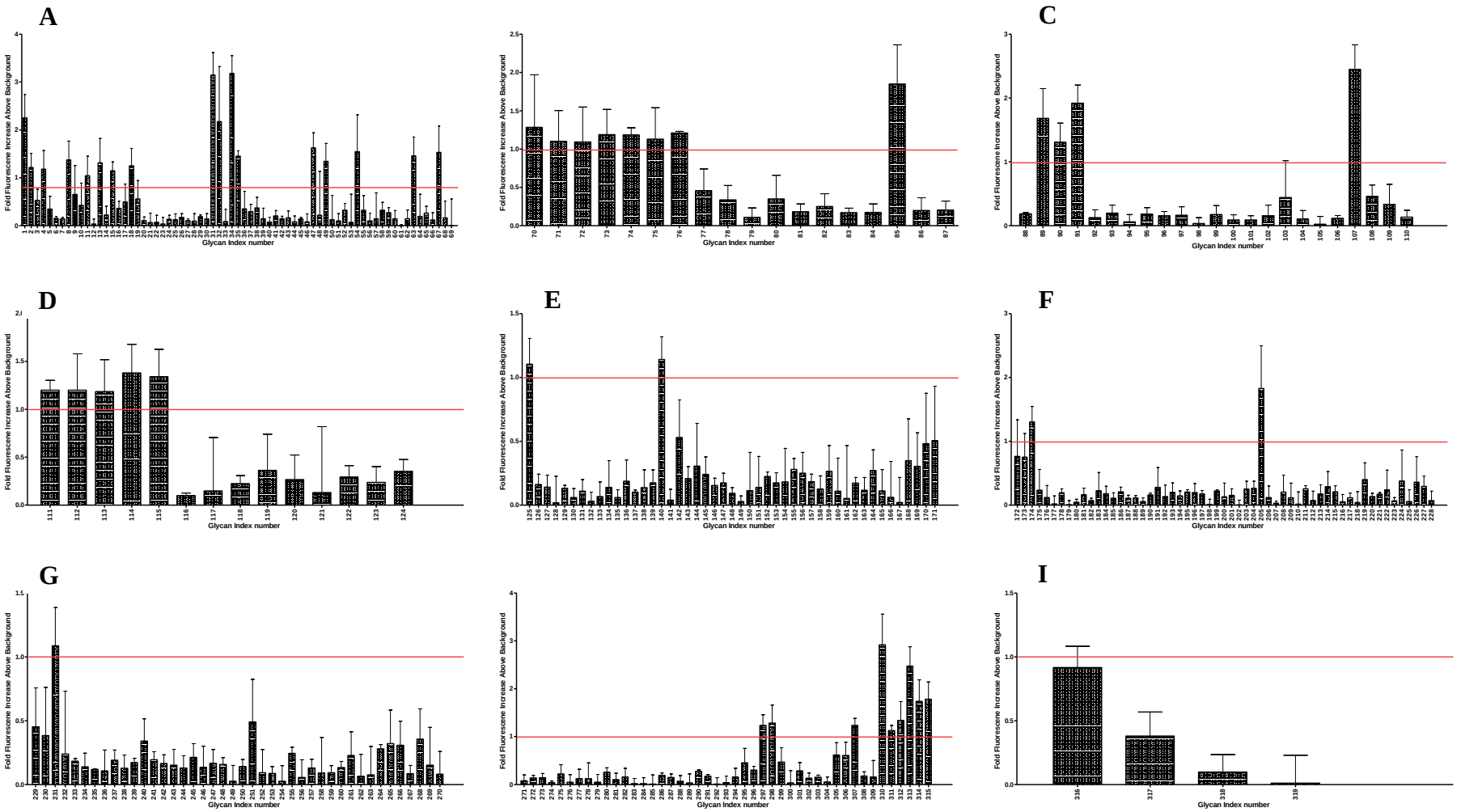


Figure G.3: M106 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

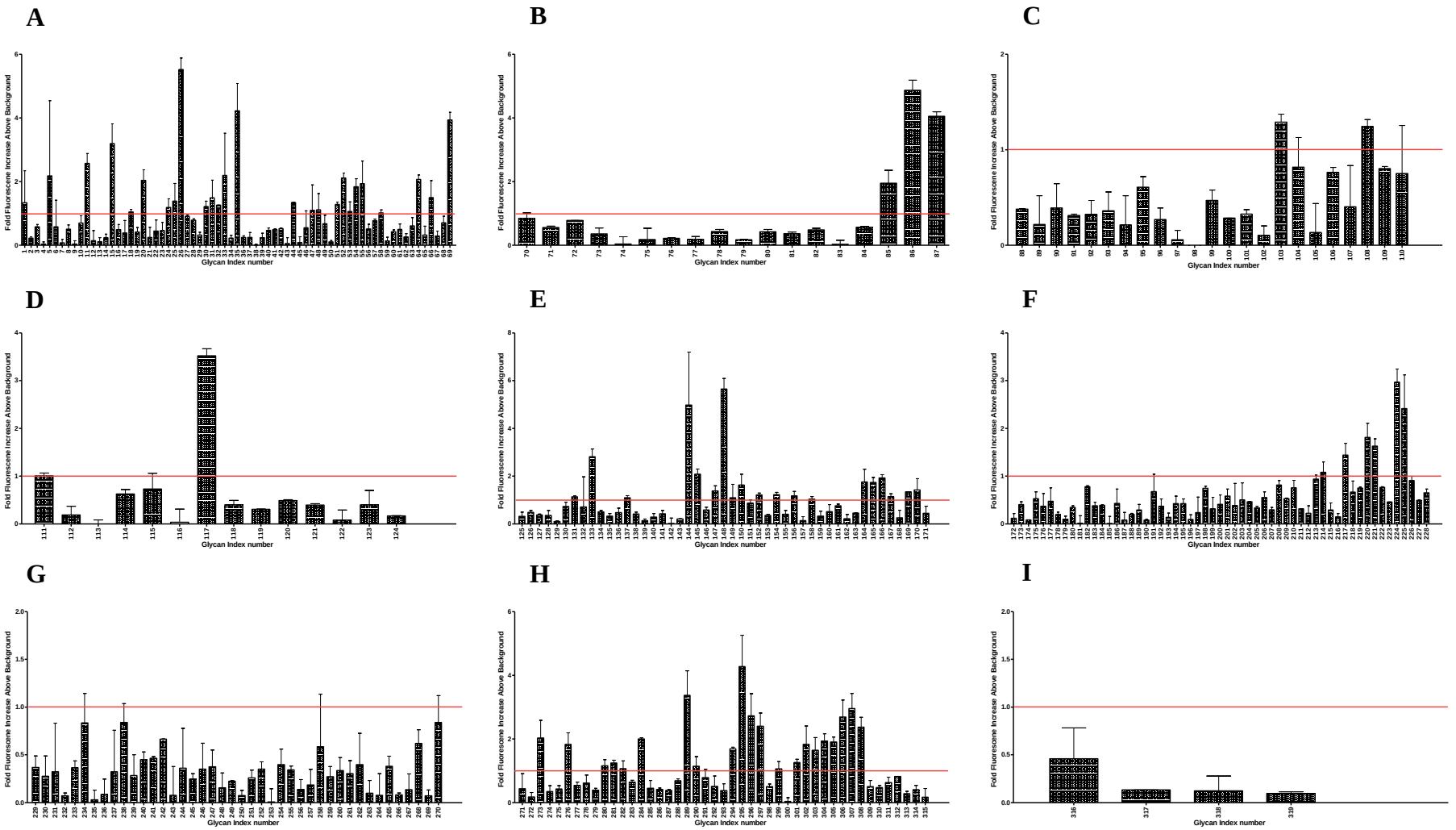


Figure G.4: M58 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

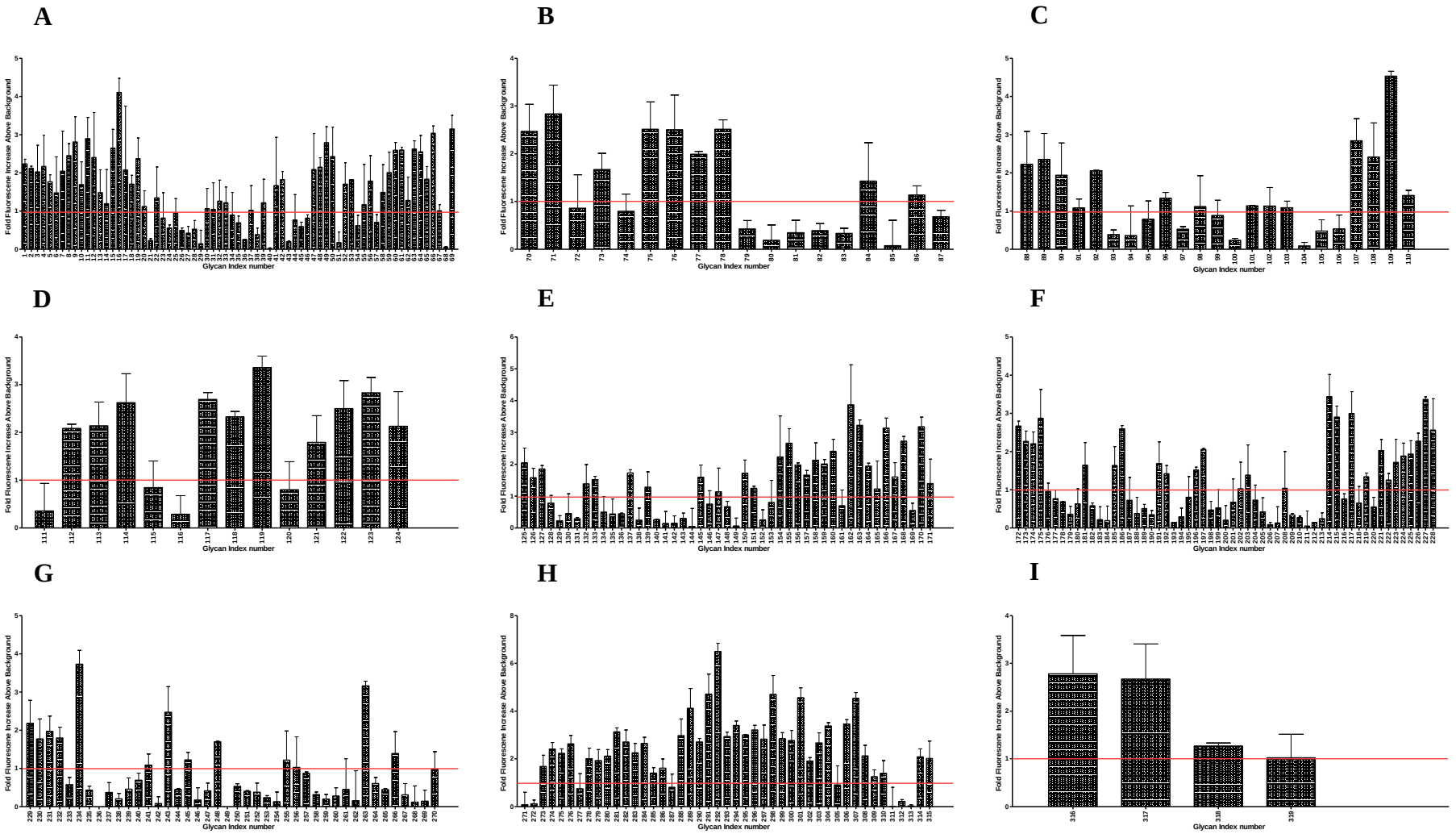


Figure G.5: M9 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

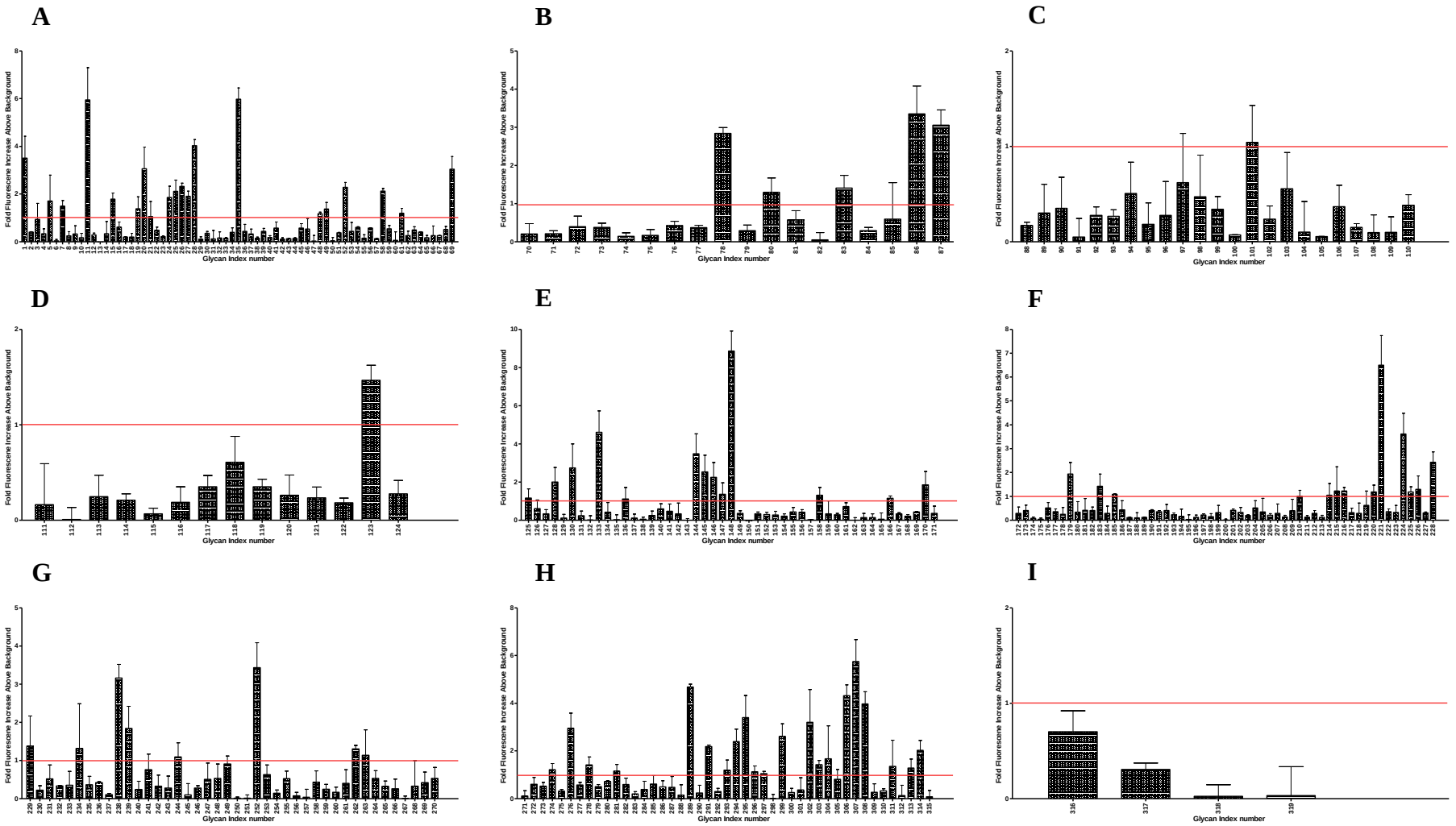


Figure G.6: M102 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

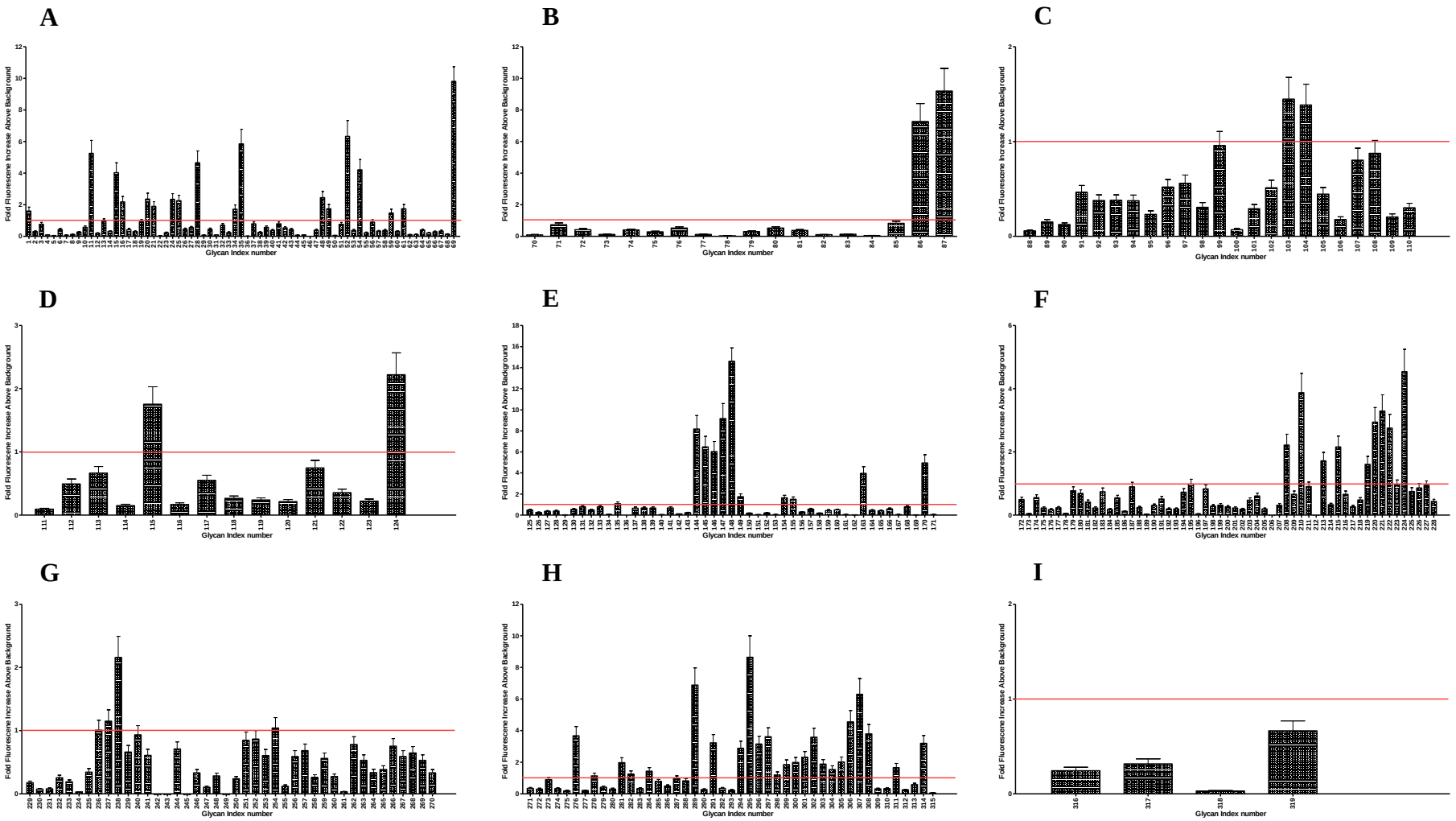


Figure G.7: M2 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

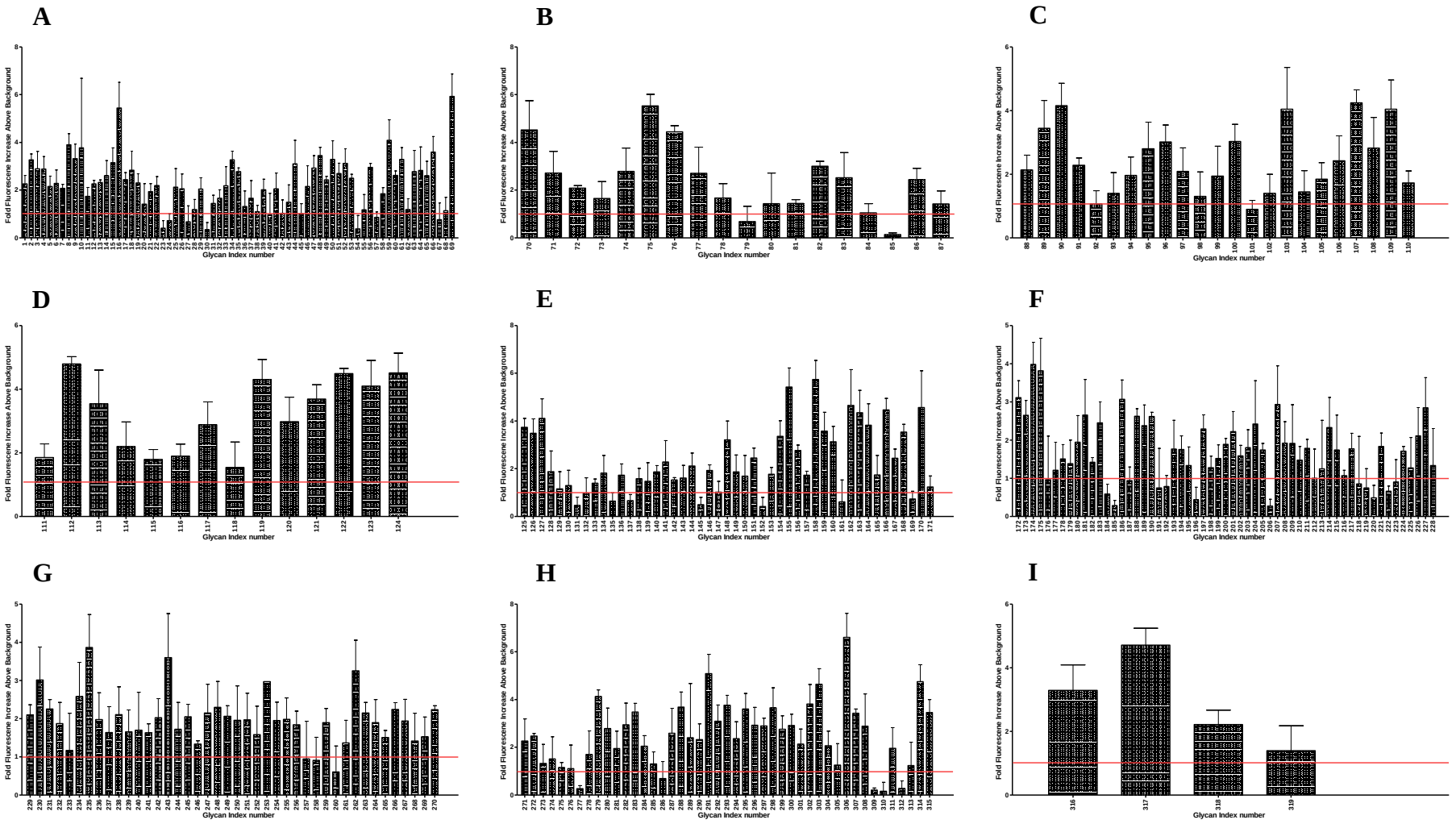


Figure G.8: M11 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

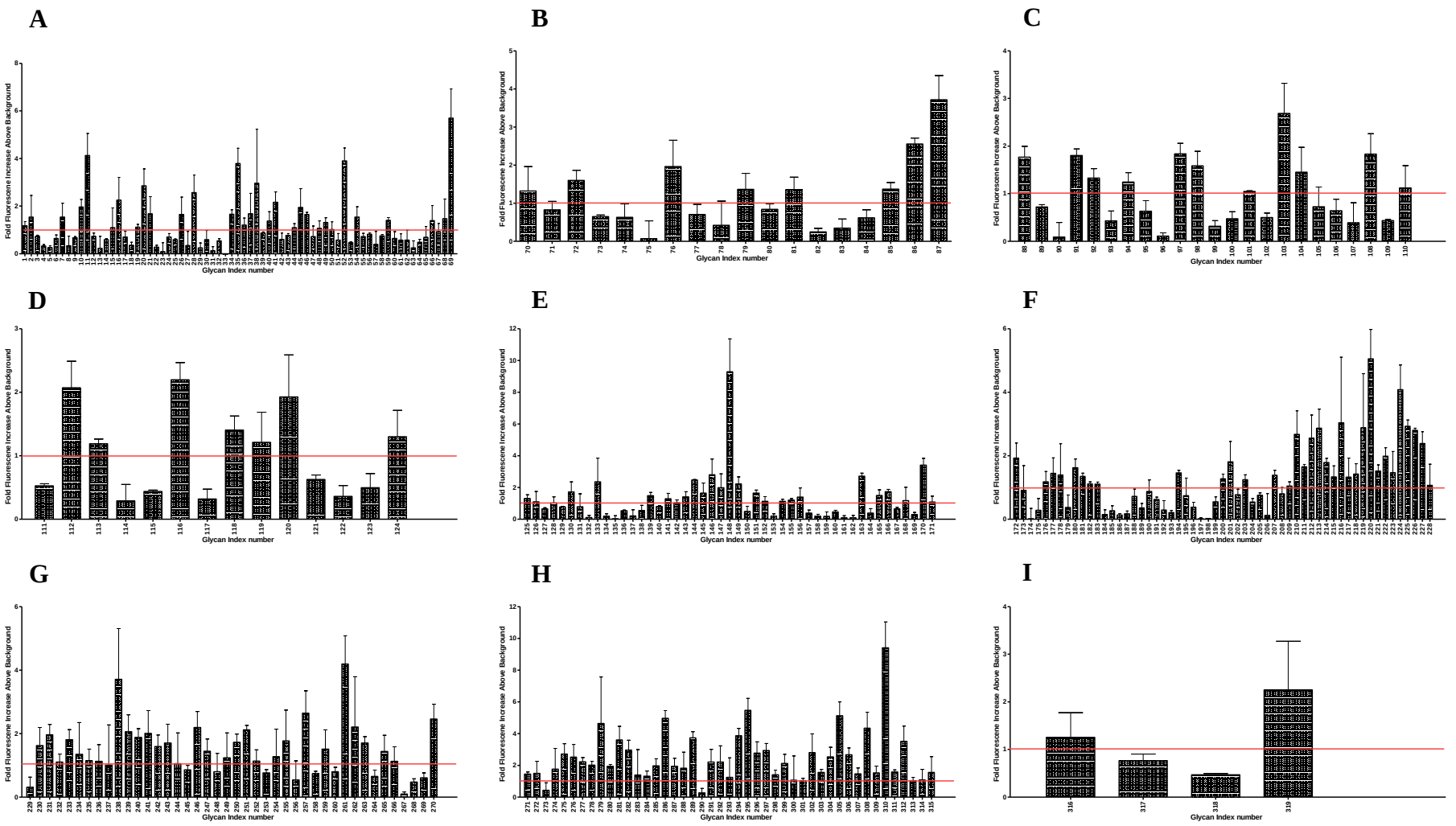


Figure G.9: M65 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

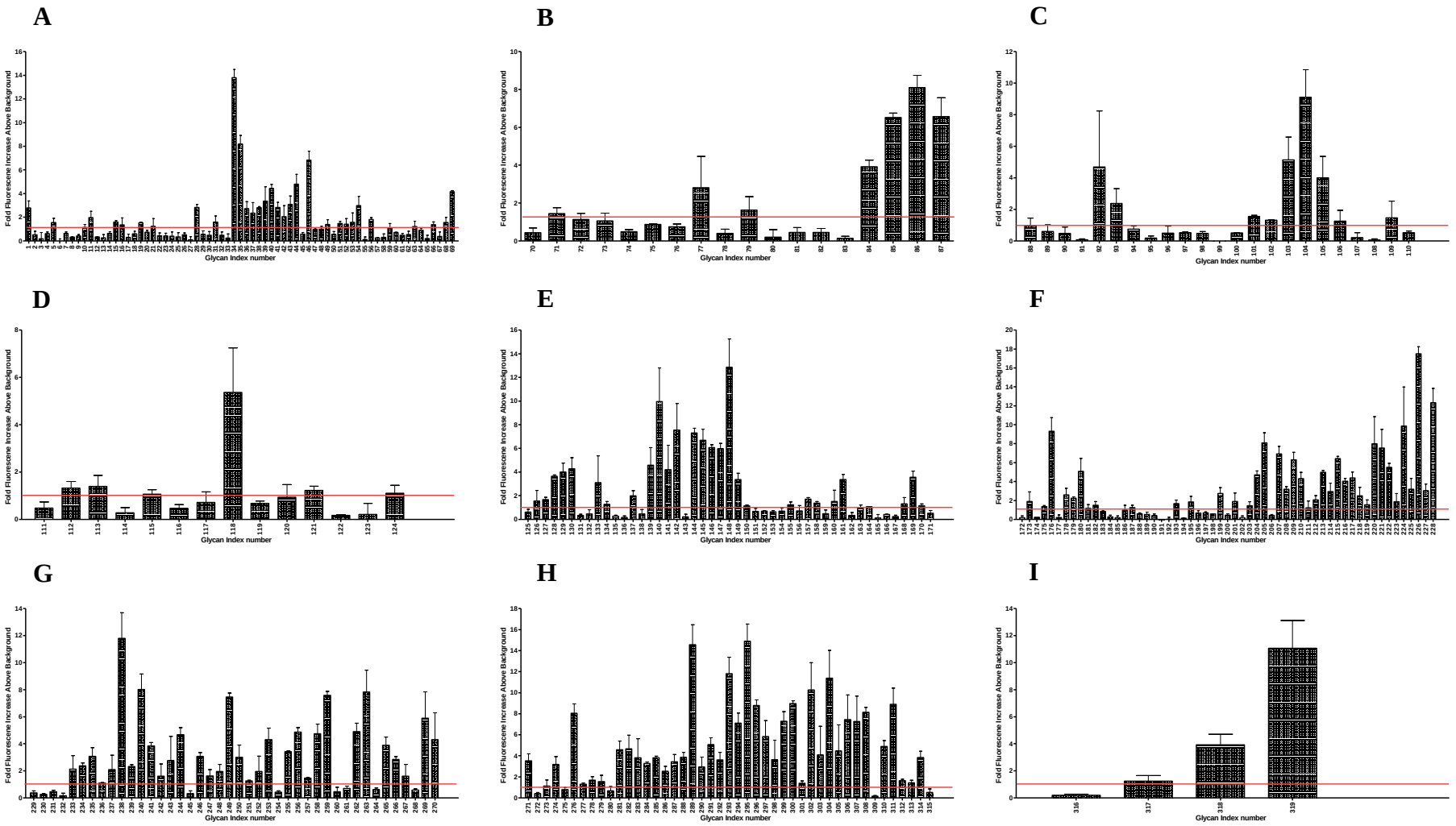


Figure G.10: M57 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

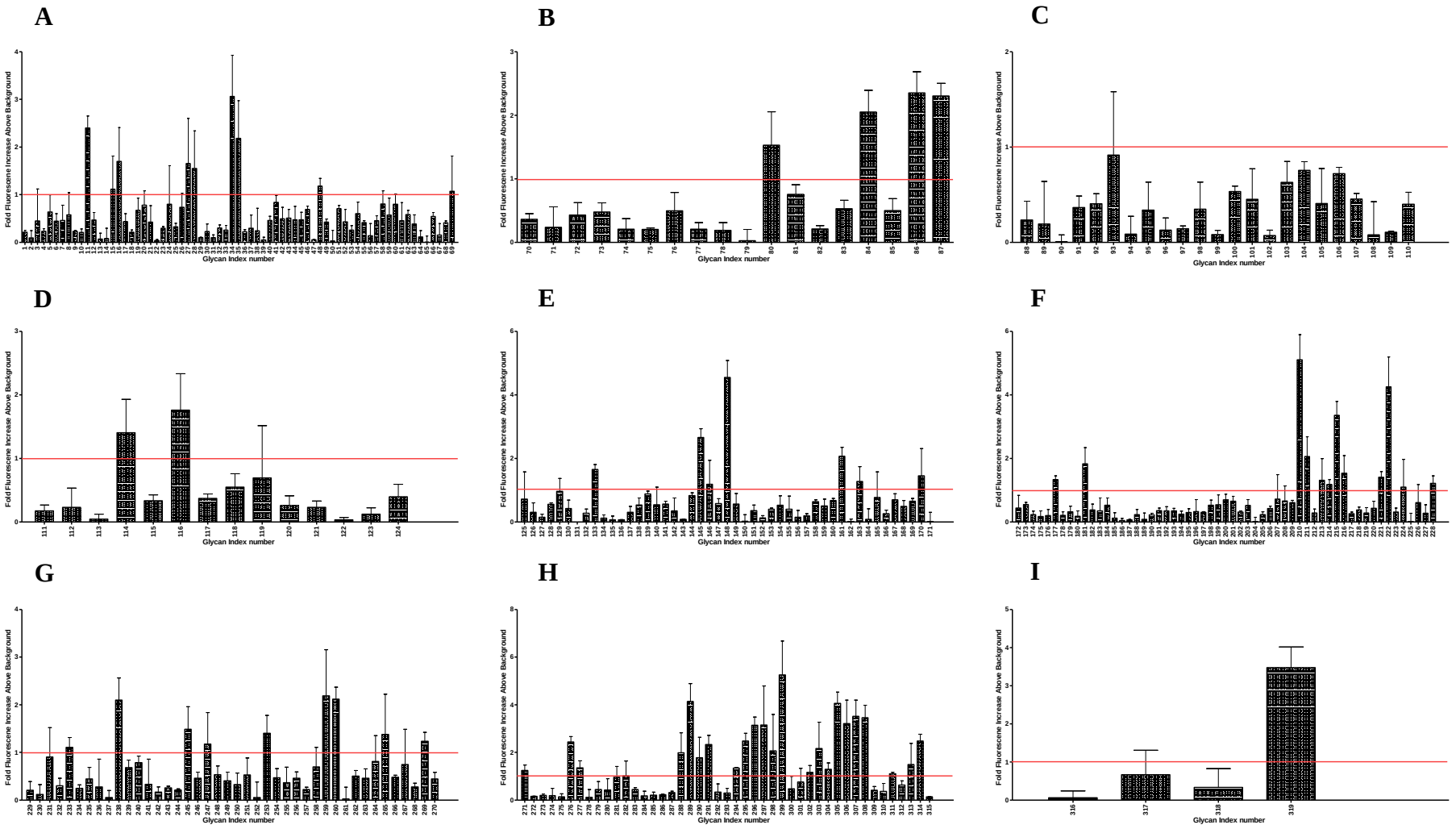


Figure G.11: M54 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

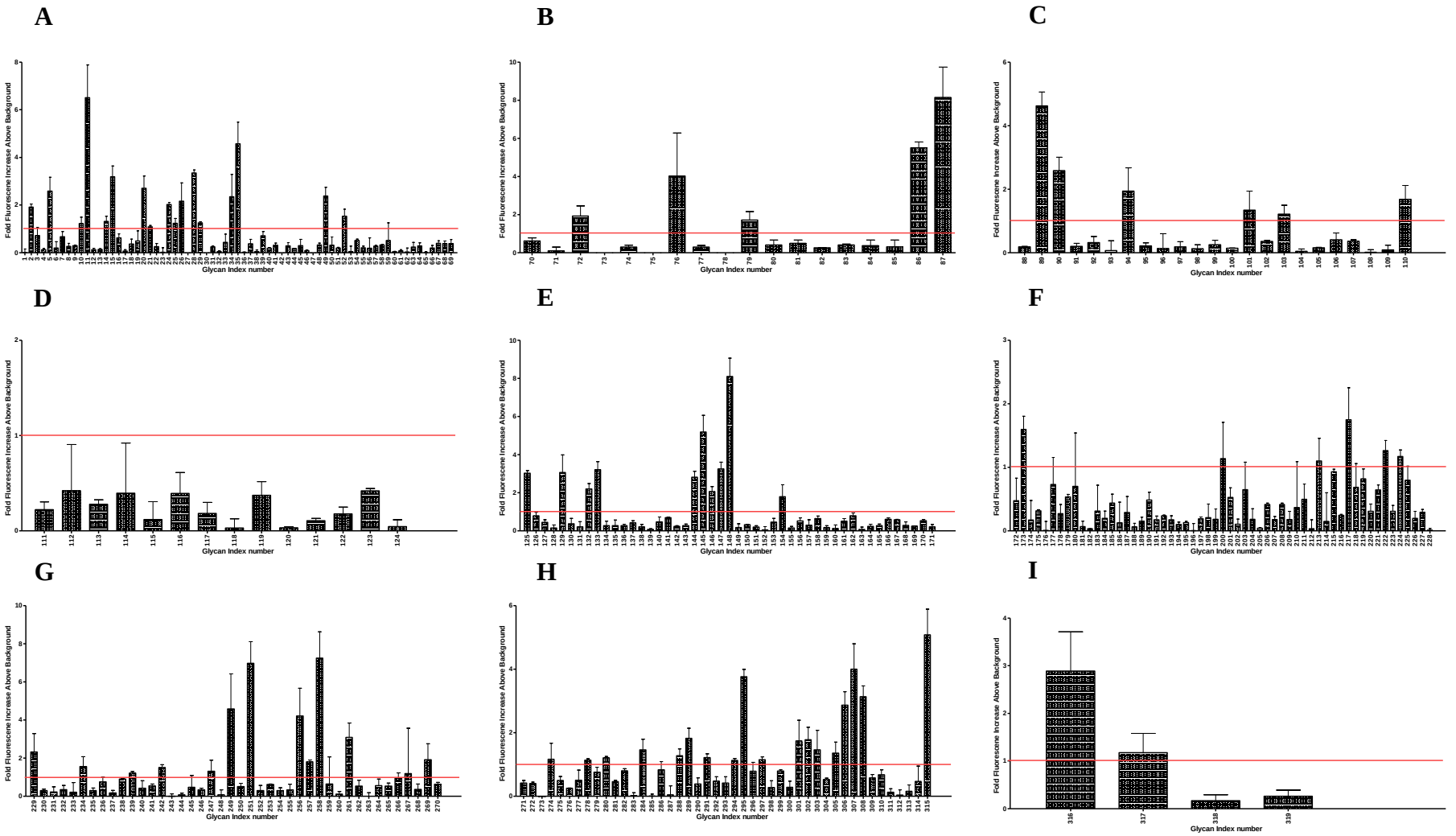


Figure G.12: M14 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

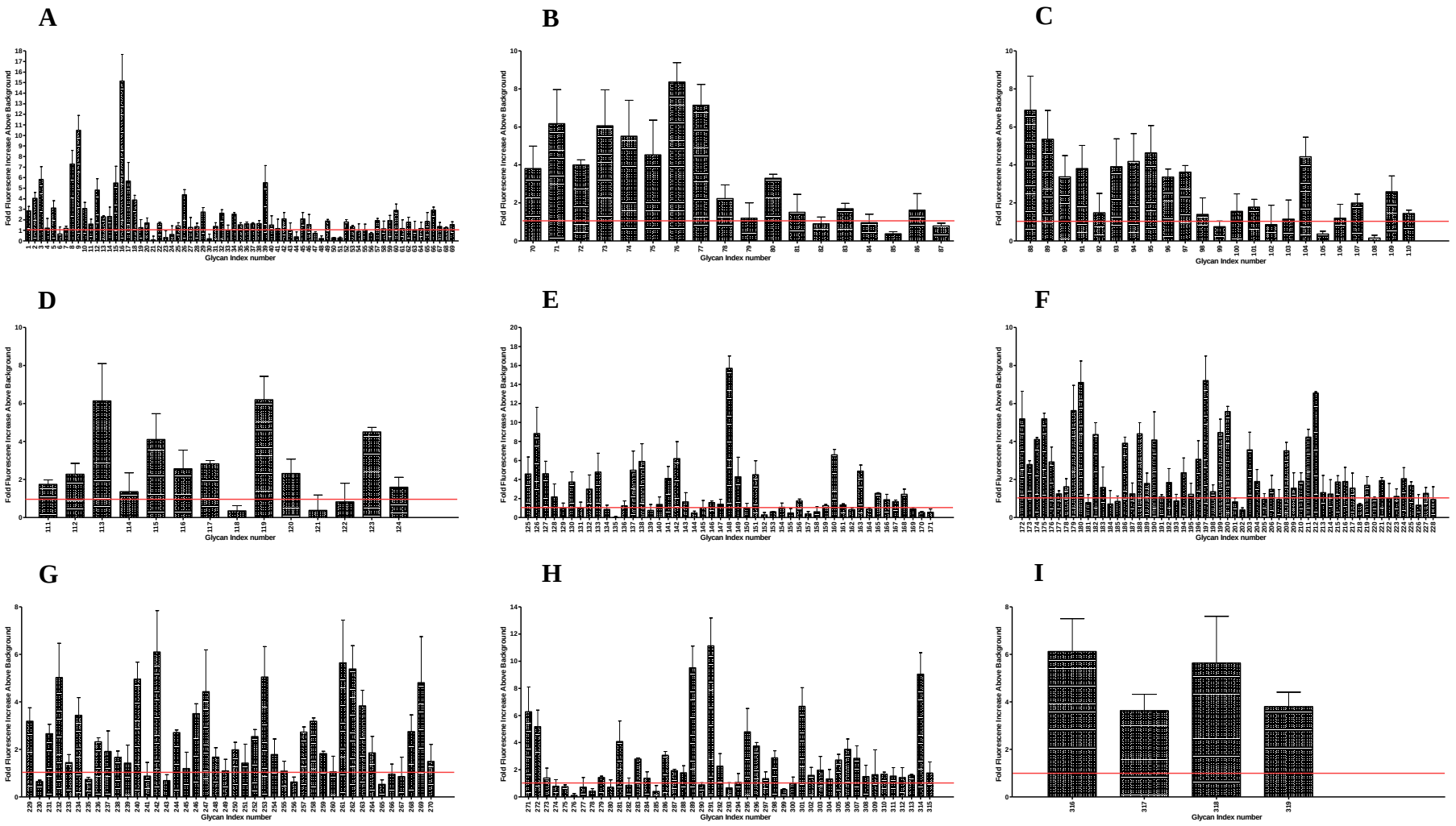


Figure G.13: M19 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

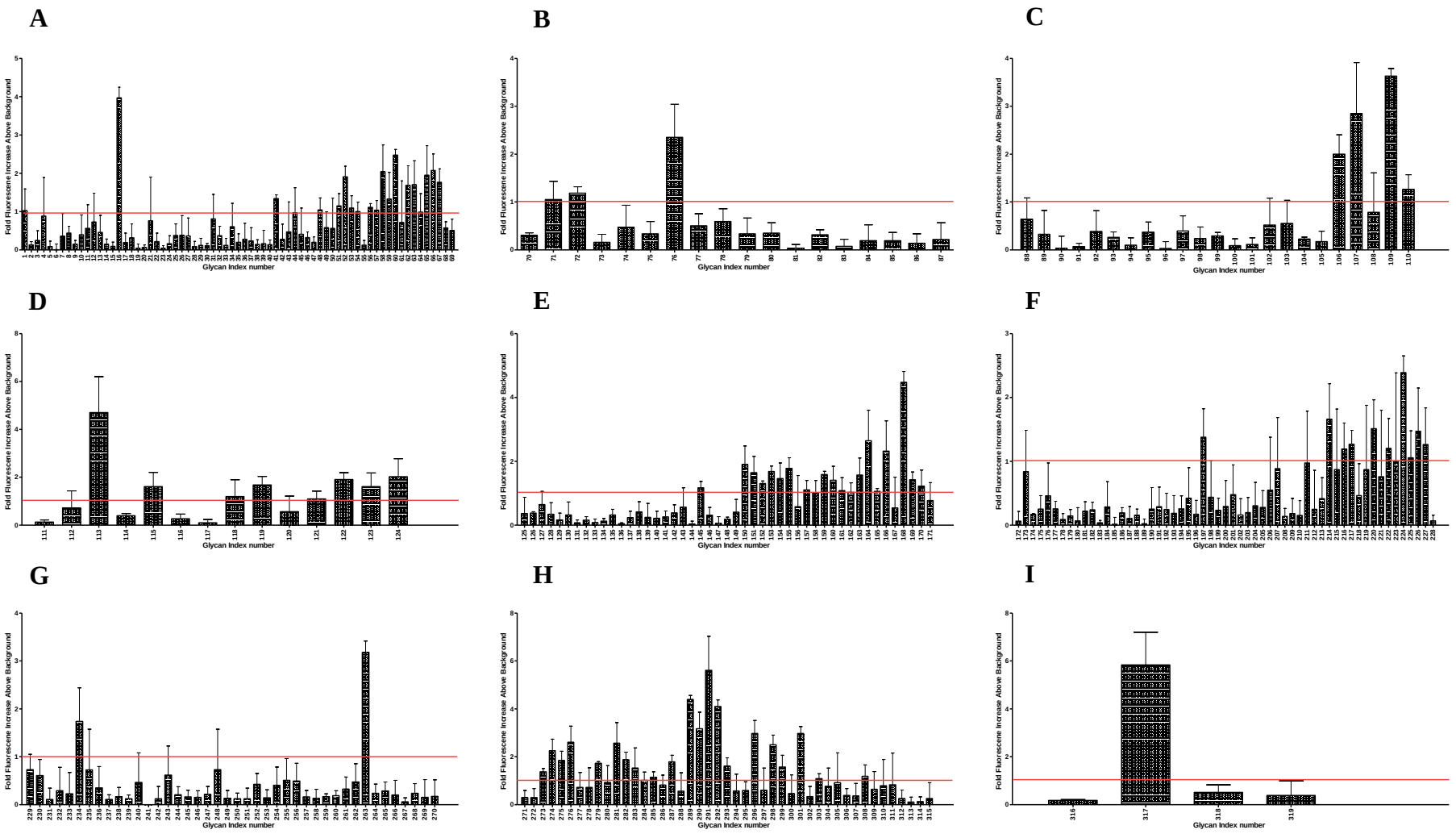


Figure G.14: M70 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

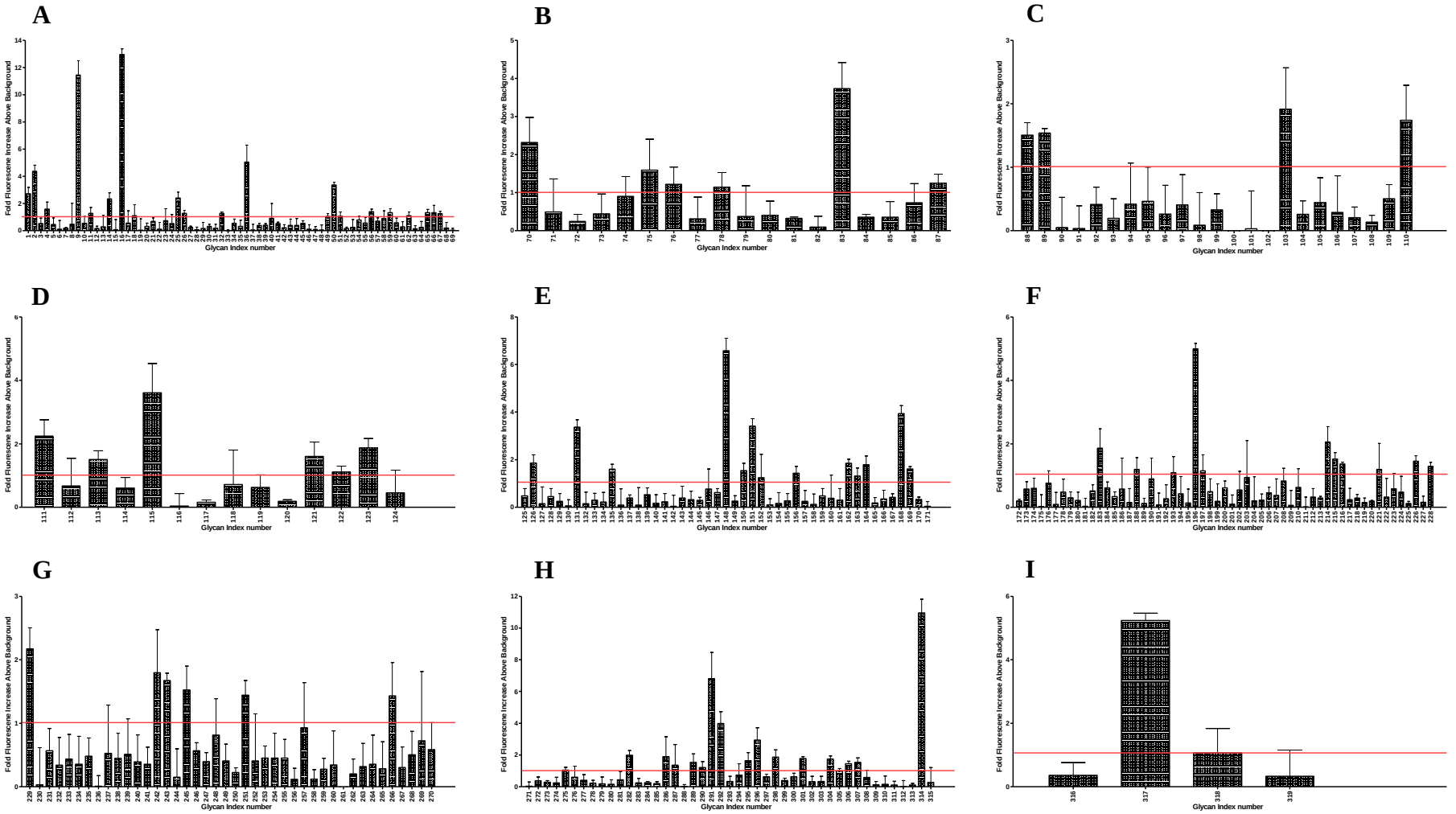


Figure G.15: M53 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

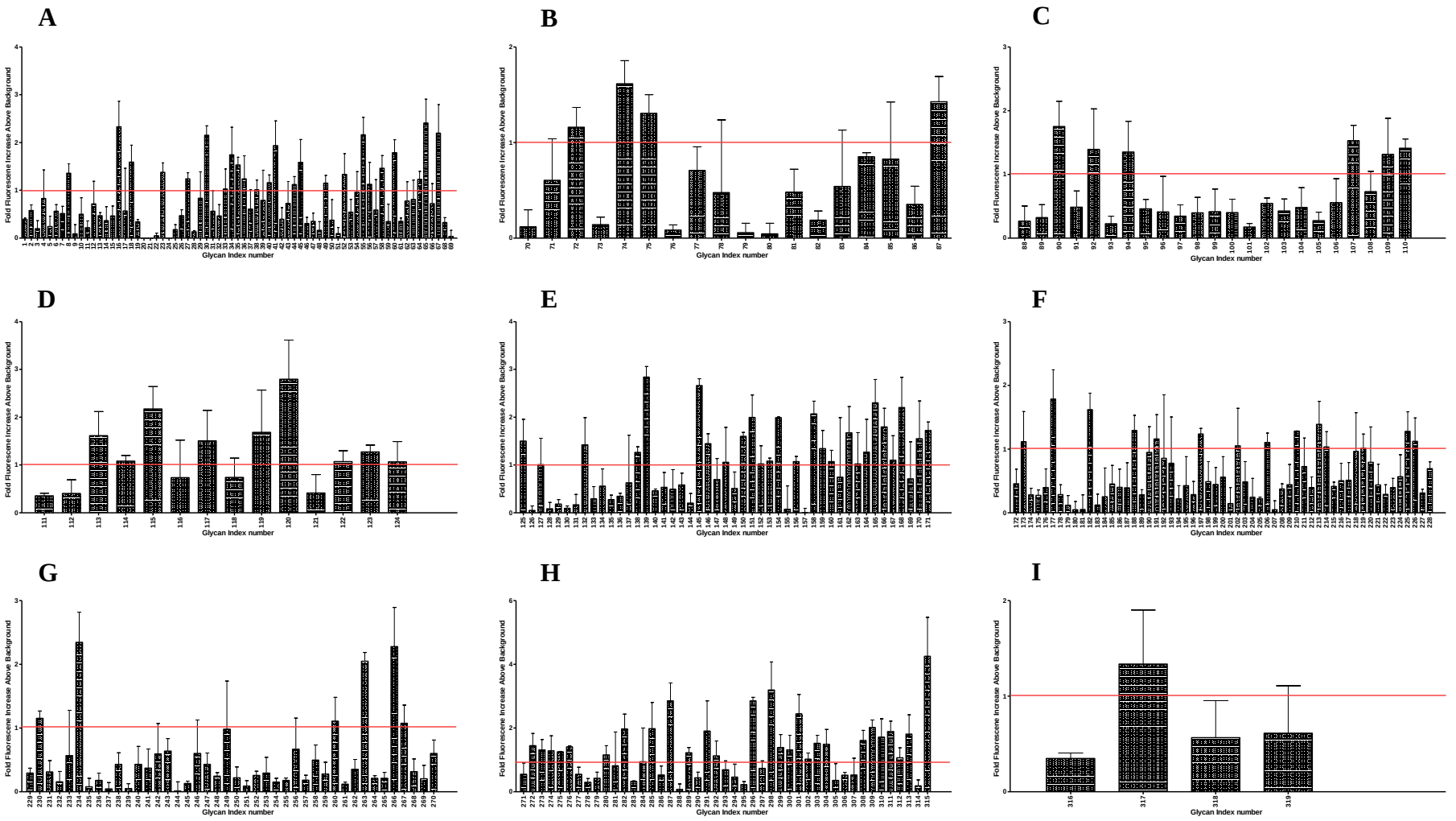


Figure G.16: M98 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

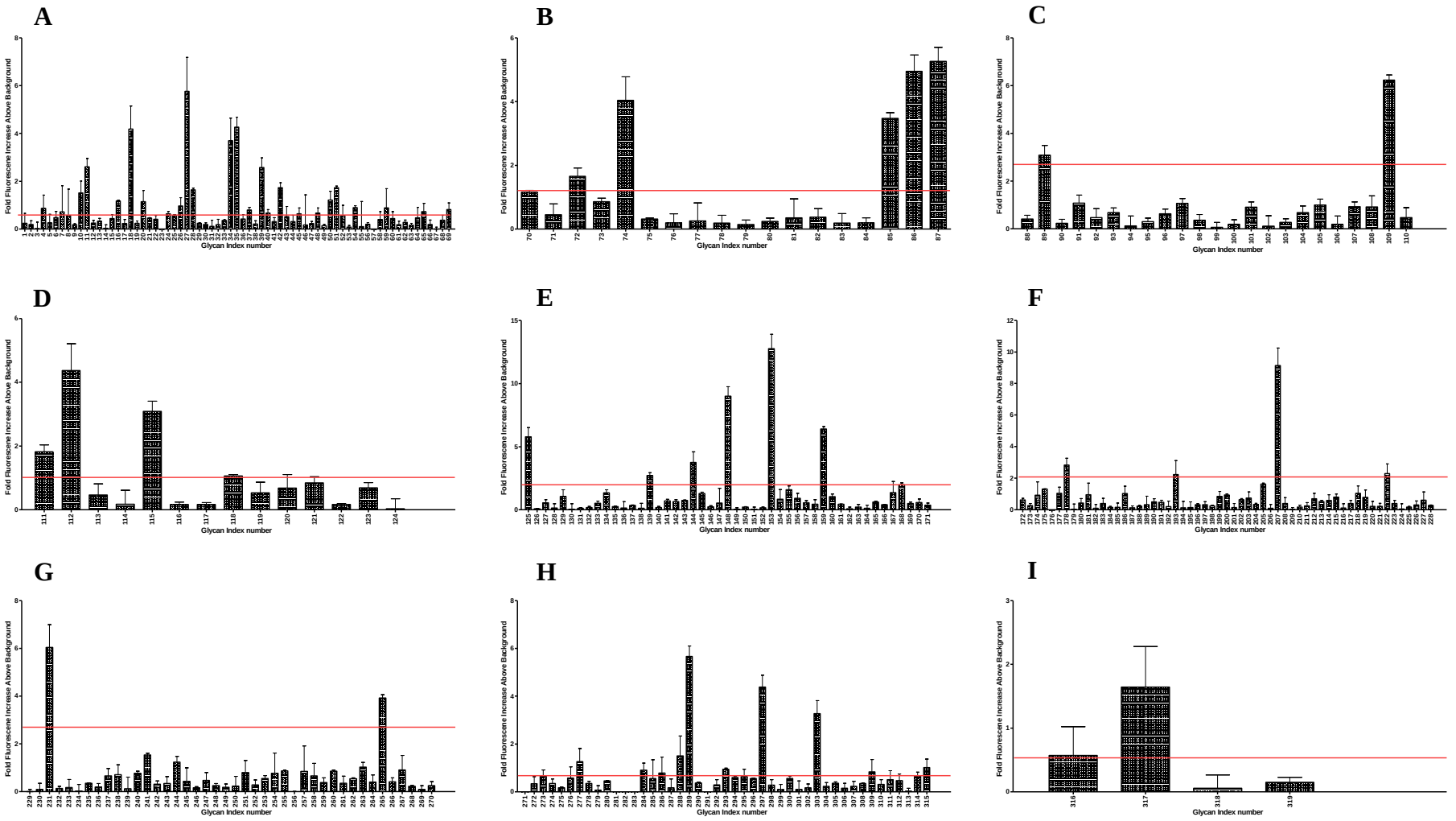


Figure G.17: M97 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

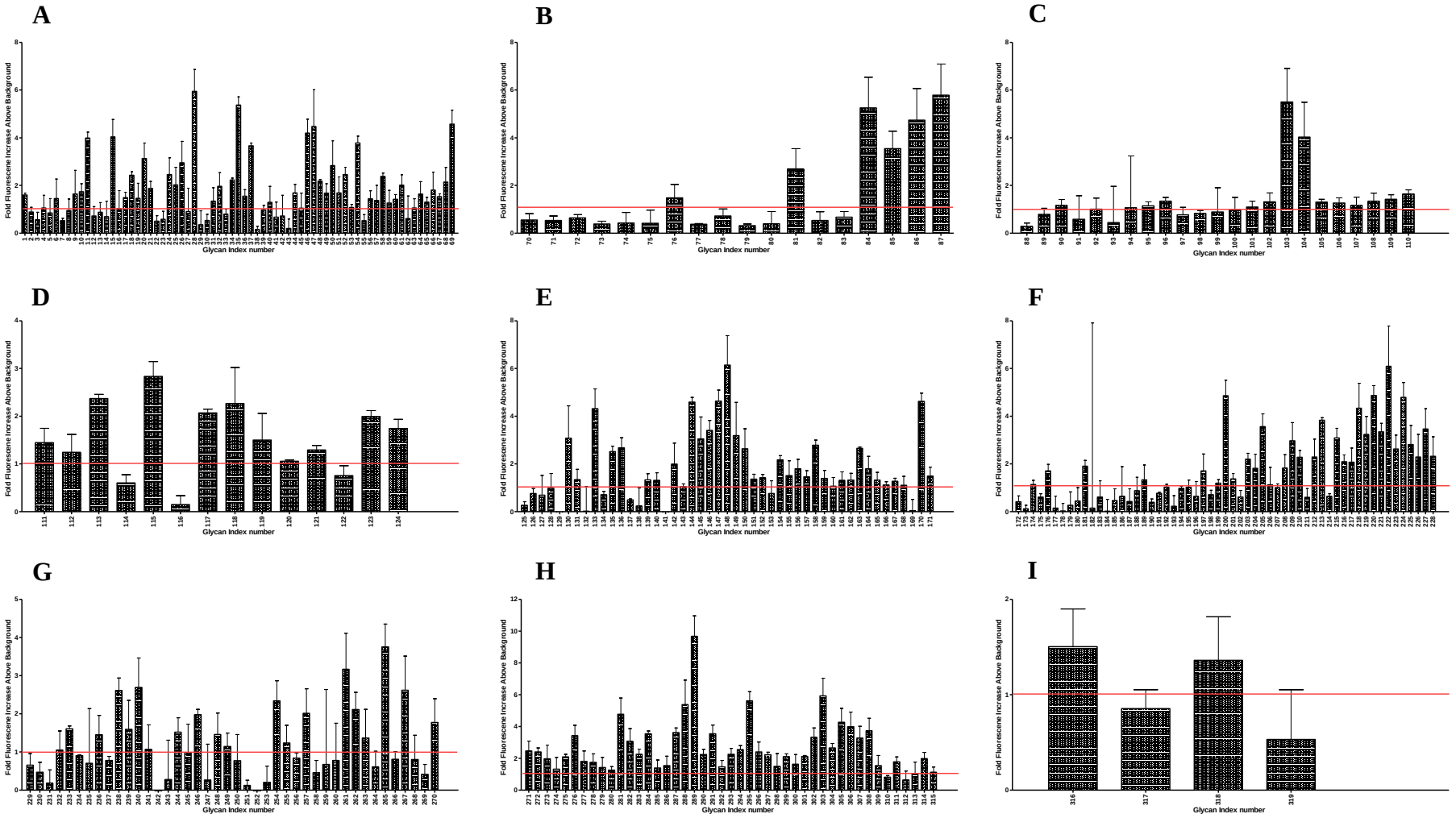


Figure G.18: M1 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

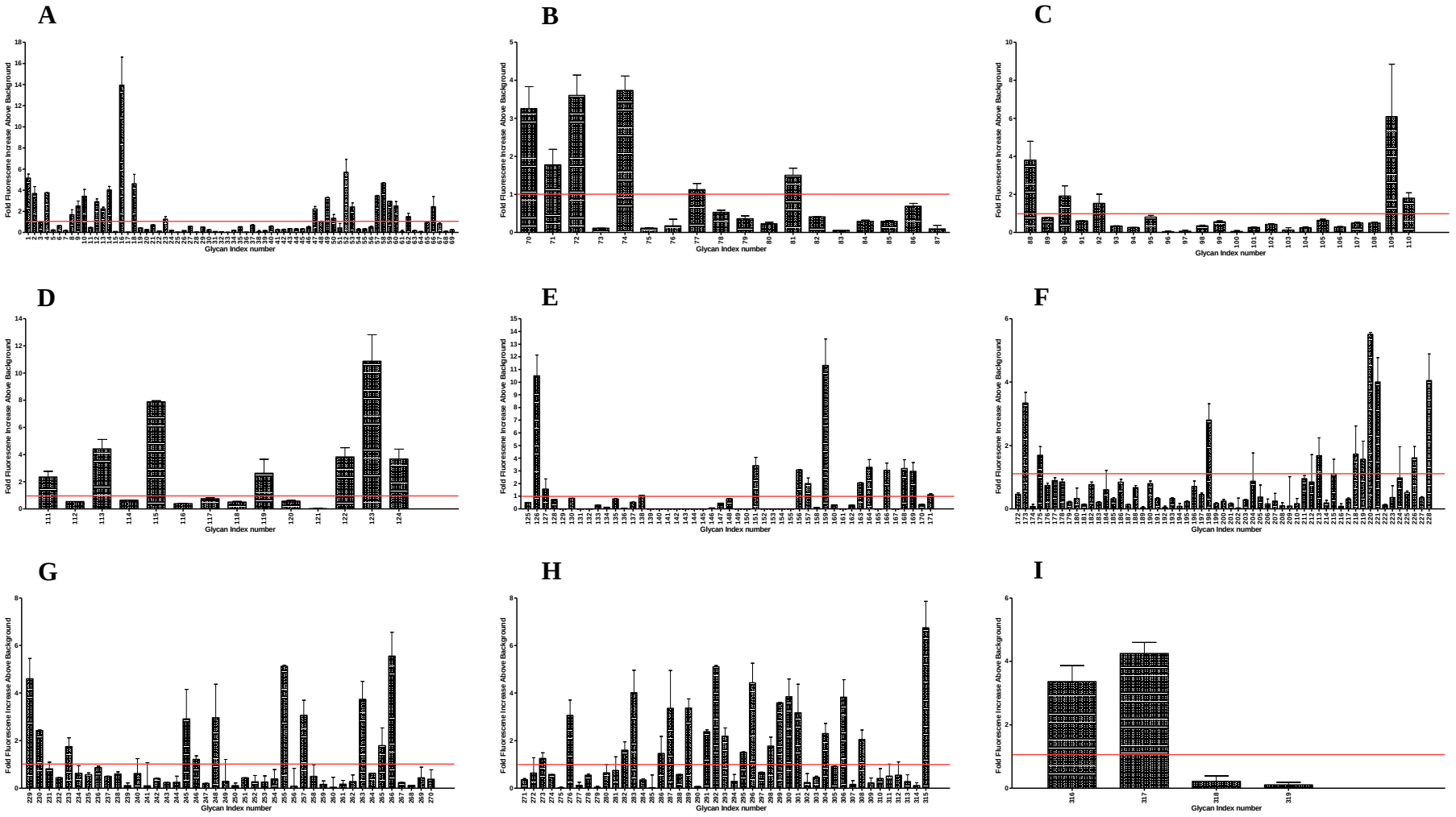


Figure G.19: M12 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

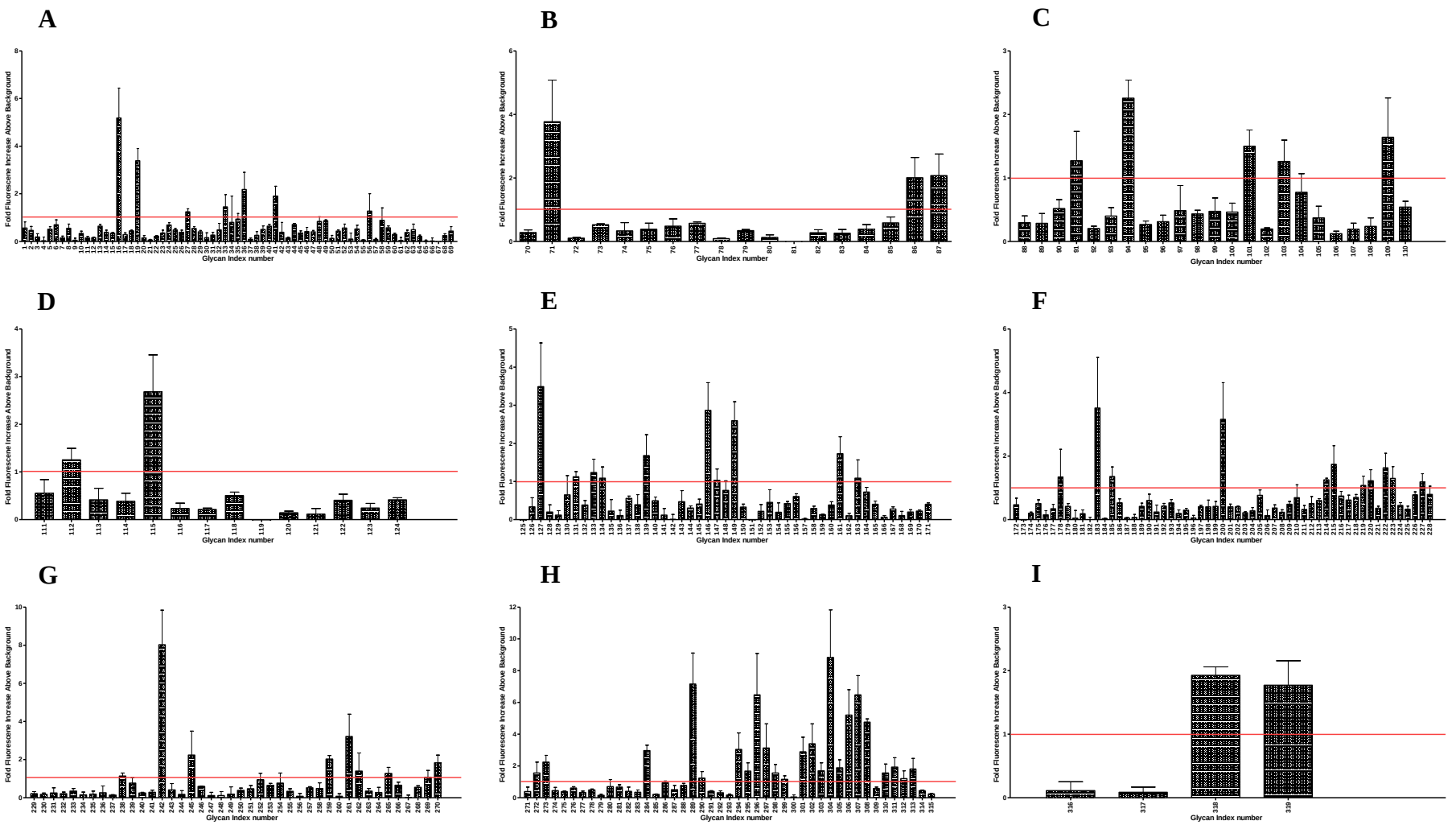


Figure G.20: M3 binding glycan profile of (A) terminal galactose, (B) terminal glucose, (C) terminal N acetyl glucosamine, (D) mannose, (E) fucosylated, (F) Sialic acid containing, (G) sulphated, (H) GAG's and (I) non-categorised structures.

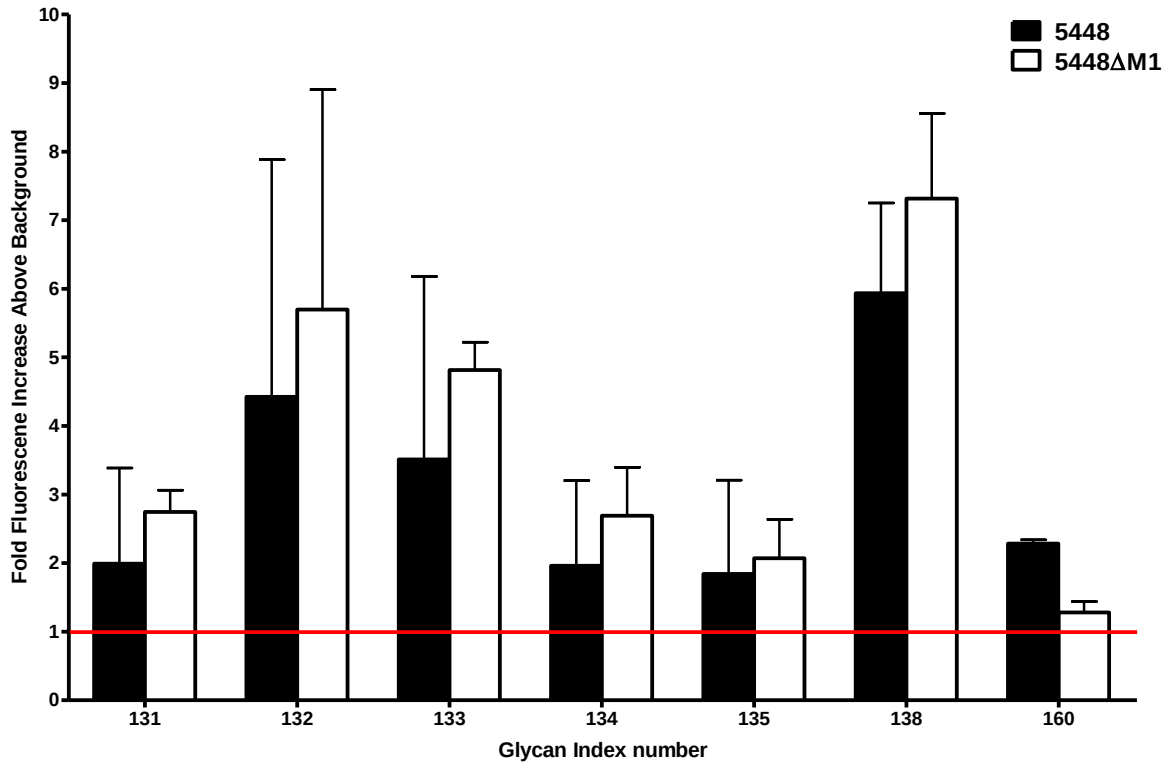


Figure G.21: Glycan binding profile of M1T1 clone GAS strains 5448 and 5448ΔM1 to blood group antigen A and B related structures. Glycan binding was analysed using ProScanArray imaging software, ScanArray Express (PerkinElmer, USA) before data was exported to Microsoft Excel for further analysis. Bacterial binding to a glycan was defined as a value $\geq$ 1 fold increase above mean background RFU. The mean background was calculated from the average RFU of all empty spots on the array plus three standard deviations. Furthermore, statistical analysis of the data was performed by a Student's *t*-test with a confidence level of 99.99% ($p \leq 0.0001$) and only glycans that met these criteria for three biologically independent samples (n = 12 glycans spot replicates) were interpreted as positive binding interactions. Glycan index: 131, Galα1-3(Fucα1-2)Galβ; 132, Galα1-3(Fucα1-2)Galβ1-3GlcNAcβ; 133, Galα1-3(Fucα1-2)Galβ1-4GlcNAcβ; 134, Galα1-3(Fucα1-2)Galβ1-3GalNAcα; 135, Galα1-3(Fucα1-2)Galβ1-3GalNAcβ; 138, GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ; 160, GalNAcα1-3(Fucα1-2)Gal.

Appendix H: Blood group antigen typing from human saliva

In this study, the blood group antigen status of each donor was typed via capillary carbon LC-ESI-IT MS/MS.

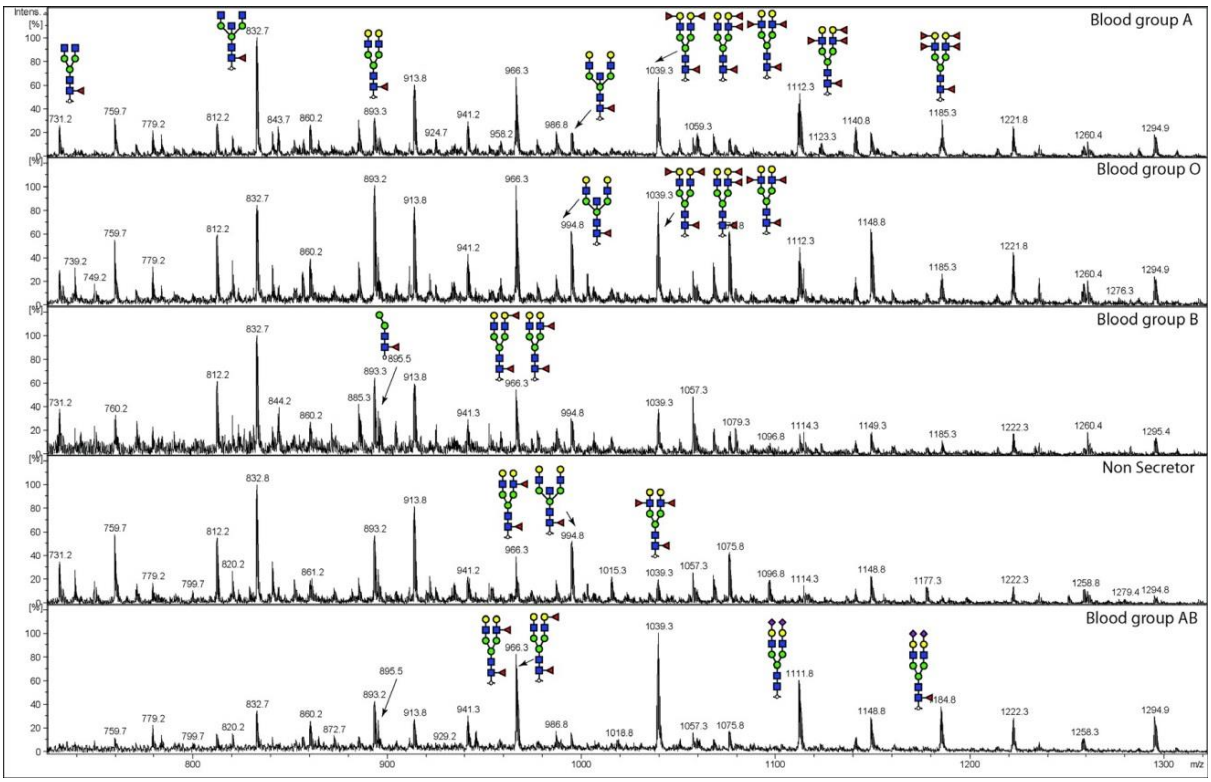


Figure H.1: Representative mass spectra of *N*-glycans released from salivary glycoproteins proteins which is subsequently indicative of donor blood group type and secretor status. Released *N*-linked glycans were analysed by capillary carbon LC-ESI-IT MS/MS.

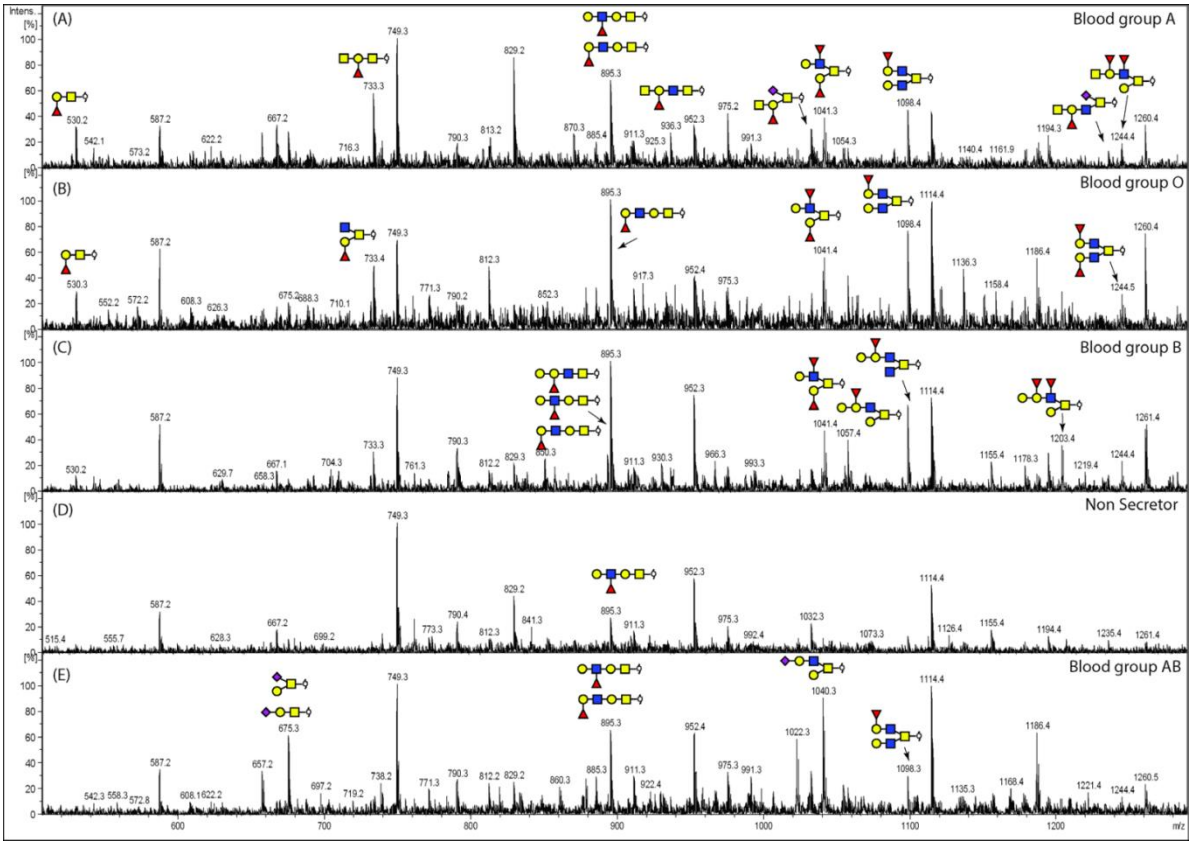


Figure H.2: Representative mass spectra of *O*-glycans released from salivary glycoproteins proteins which is subsequently indicative of donor blood group type and secretor status. Released *O*-linked glycans were analysed by capillary carbon LC-ESI-IT MS/MS.