Title: A systematic and functional classification of *Streptococcus pyogenes* that serves as a new tool for molecular typing and vaccine development.

Martina Sanderson-Smith [a], David M.P. De Oliveira [a,1], Julien Guglielmini [b,c,1], David J. McMillan [d,e], Therese Vu [d,f], Jessica K. Holien [g], Anna Henningham [h], Andrew C. Steer [i,j,k], Debra E. Bessen [l], James B. Dale [m,n,o], Nigel Curtis [i,p,q], Bernard W. Beall [r], Mark J. Walker [h], Michael W. Parker [g,s], Jonathan R. Carapetis [t], Laurence Van Melderen [f], Kadaba S. Sriprakash [d] and Pierre R. Smeesters [f,i,*]; the M Protein Study Group.

a. Illawarra Health and Medical Research Institute and School of Biological Sciences, University of Wollongong, Wollongong, Australia.
b. Microbial Evolutionary Genomics, Département Génomes et Génétique, Institut Pasteur, Paris, France.
c. CNRS, UMR3525, F-75015, Paris, France.
d. Bacterial Pathogenesis Laboratory, Queensland Institute of Medical Research, Brisbane, Australia.
e. Inflamation and Healing Research and School of Health and Sports Sciences, University of the Sunshine Coast, Sippy Downs, Australia
f. Laboratoire de Génétique et Physiologie Bactérienne, Institut de Biologie et de Médecine Moléculaires, Faculté des Sciences, Université Libre de Bruxelles, Gosselies, Belgium.
g. Biota Structural Biology Laboratory, ACRF Rational Drug Discovery Centre. St. Vincent's Institute of Medical Research, Melbourne, Australia.
h. School of Chemistry and Molecular Biosciences and Australian Infectious Diseases Research Centre, University of Queensland, Brisbane, Australia.
i. Murdoch Children Research Institute, Melbourne, Australia.
j. Centre for International Child Health The University of Melbourne, Melbourne, Australia.
k. Department of General Medicine, Royal Children's Hospital Melbourne, Melbourne, Australia.
l. Department of Microbiology and Immunology, New York Medical College, Valhalla, United States of America.
m. Department of Medicine, The University of Tennessee Health Science Center, Memphis, United States of America.
n. Department of Veterans Affairs Medical Center, Memphis, United States of America.
o. Department of Microbiology, Immunology and Biochemistry, The University of Tennessee Health Science Center, Memphis, United States of America.
p. Infectious Diseases Unit, Royal Children's Hospital Melbourne, Melbourne, Australia.
q. Department of Paediatrics, The University of Melbourne, Melbourne, Australia
r. Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, United States of America.
s. Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Melbourne, Australia.
t. Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, Perth, Australia.

[1]: Contributed equally to this work.

Running title: S*treptococcus pyogenes* classification

Abstract word count: 146

Manuscript word count: 3497

Conflict of interest

J.B.D. is the inventor of certain technologies related to the development of GAS vaccines. The University of Tennessee Research Corporation has licensed the technology to Vaxent, LLC. J.B.D. serves as the Chief Scientific Officer of Vaxent. The other authors have no conflict of interest to declare.

Funding

This project has been funded by the European Society for Clinical Microbiology and Infectious Diseases, European Society for Paediatric Infectious Diseases, Fonds National de la Recherche Scientifique (Belgium), Fonds Brachet and Fondation Van Buuren (Belgium), Australian National Health and Medical Research Council (Australia) and the National Institutes of Health (USA). Funding was also obtained from the Victorian Government Operational Infrastructure Support Scheme to St Vincent's Institute and Murdoch Childrens Research Institute. MS-S is an NHMRC Career Development Fellow. MWP is an NHMRC Senior Principal Research Fellow. JKH is a joint Cure Cancer/ Leukaemia Foundation Postdoctoral Fellow. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Meetings

Preliminary data have been presented at the Australian Society for Microbiology, Annual Scientific Meeting in Brisbane, Australia, in 1-4 July 2012 and at the 52nd International Conference on Antimicrobial Agents and Chemotherapy (ICAAC), 9-12 September 2012, San Francisco.

Corresponding author

Pierre Smeesters
Laboratoire de Génétique et Physiologie Bactérienne, IBMM
Université Libre de Bruxelles
12 rue des professeurs Jeener et Brachet, 6041 Gosselies, Belgium
Tel : 32 2 650 97 76
psmeeste@ulb.ac.be

Abstract

Streptococcus pyogenes ranks amongst the main causes of mortality from bacterial infections worldwide. Currently there is no vaccine to prevent diseases such as rheumatic heart disease and invasive streptococcal infection. The streptococcal M protein that is used as the substrate for epidemiological typing is both a virulence factor and a vaccine antigen. Over 220 variants of this protein have been described, making comparisons between proteins difficult, and hindering M protein-based vaccine development. A functional classification based on 48 *emm*-clusters containing closely related M proteins that share binding and structural properties is proposed. The need for a paradigm shift from type-specific immunity against *Streptococcus pyogenes* to *emm*-cluster based immunity for this bacterium should be further investigated. Implementation of this *emm*-cluster-based system as a standard typing scheme for *Streptococcus pyogenes* will facilitate the design of future studies of M protein function, streptococcal virulence, epidemiological surveillance and vaccine development.

Keywords: *Streptococcus pyogenes*, vaccine, M protein, fibrinogen, plasminogen, IgA, IgG, molecular typing, epidemiology.

Introduction

Streptococcus pyogenes, (Group A streptococcus, GAS) infections result in over 500,000 deaths per year [1]. The greatest burden is due to rheumatic heart disease in low-income settings, affecting 12 million individuals and resulting in 350,000 deaths each year [1]. Invasive infections are also of significant concern, with a mortality rate from 15 to 30% and an incidence exceeding that of meningococcal disease in the pre-vaccine era [2]. Aside from rheumatic fever, there are no proven public health control strategies for GAS disease. Prevention strategies for rheumatic fever in low-income countries are difficult to implement. A safe and effective vaccine is therefore needed but remains commercially unavailable despite numerous initiatives [3].

The M protein is a surface protein, vaccine antigen and virulence factor of GAS [4, 5]. The M protein inhibits phagocytosis in the absence of opsonizing antibodies, promotes adherence to human epithelial cells and helps the bacterium overcome innate immunity. The multifunctional nature of this protein is further evidenced by its interaction with numerous host proteins occurring along its entire length [4]. The N-terminus consists of a highly variable amino acid sequence resulting in antigenic diversity, and is the basis for the nucleotide-based *emm*-typing scheme [6-8]. To date, 223 different *emm*-types have been reported [9] but only a small proportion of them have been properly characterized for their cross-reactive properties (the so called serotypes (M-types)) mentioned in earlier studies [10, 11].

Systematic reviews have highlighted differences in the *emm*-type distribution of GAS, especially between high-income countries and resource-poor regions [12, 13]. While only a relatively small number of predominant *emm*-types circulate in high-income countries, the diversity of strains associated with disease in low-income settings is much greater. This diversity has made epidemiologic comparisons complex to analyze, has hindered the development of M protein vaccines, and has made comprehensive microbiologic characterization of the global repertoire of GAS strains challenging. Most often, typing GAS relies on a small portion (10-15%) of the M protein. Preliminary analysis of the complete sequence of 51 M proteins suggested that the many *emm*-types circulating in low-income countries [14] are highly similar in sequence [15, 16], raising questions about the type-specificity of the immune response induced by such highly homologous M proteins [16, 17]. Pioneering work in the 1950s established the basis for “type-specific immunity” [10, 11, 18, 19], showing that M-type specific antibodies are responsible for immunity against the homologous M-type, with no effect on infection by heterologous M-types. However, this broadly accepted paradigm has only been tested with a limited number of *emm*-types and its applicability to the many *emm*-types circulating in low-income countries has not been investigated.

We described a worldwide comprehensive study of 1086 GAS isolates collected from 31 countries representing 175 *emm*-types [9] and investigate the feasibility and value of a new *emm*-cluster typing system. This *emm*-cluster system has strong phylogenetic support, serves as a functional classification scheme for GAS M proteins and can support vaccine design and evaluation.

Materials and Methods

Nucleotide and protein sequence analysis

PCR amplification and sequencing of *emm* genes was performed as previously described [9, 15]. The predicted amino acid sequences of M proteins were trimmed from the first amino acid of the predicted mature protein to the first amino acid of the D repeat near the sortase LPxTG motif [9, 15]. The absence of significant recombination events in this dataset has been demonstrated prior to phylogenetic analysis (See supplementary data).

Phylogenetic analysis

Multiple protein sequences alignments were obtained using MUSCLE [20] with default parameters as implemented in SeaView [21]. Informative sites were extracted from these alignments using default criteria from BMGE [22] (See supplementary data). Phylogenetic inferences were made using PhyML [23] with gamma parameter of 0.46 under the LG+Γ model of substitution from an optimized BioNJ starting tree. The definition of the *emm*-clusters was based on four bioinformatic criteria: 1) monophyletic or paraphyletic nature, 2) supported by an approximate Likelihood-Ratio Test (aLRT) higher than 80%, 3) demonstrating a minimal average pairwise identity of 70% between all M proteins included and 4) demonstrating a minimum pairwise identity of 60% between pair of M proteins (C repeat size variation was excluded from identity calculation). The selective pressure analysis is described in supplementary data.

Cloning, expression and purification of recombinant M proteins

A subset of 26 M proteins, representing 24 M types, was selected for binding studies; the M proteins chosen provide coverage of the major *emm*-cluster groups within the phylogenetic tree, and include positive and negative control proteins, based on previously published studies. Recombinant M proteins were produced essentially as previously described [24] (See supplementary data).

Binding assays

Host proteins were selected to provide analysis of interactions across the full length of the M protein (N-terminus, Central domain and C-terminus), and also based on the proposed contribution of these proteins to GAS virulence. Purified histidine-tagged recombinant M protein was analyzed for binding affinity to human glu-plasminogen (Haemotologic Technologies Inc., Essex Junction, USA), human fibrinogen and albumin (Sigma-Aldrich, Sydney, Australia), IgG (Life Technologies, Melbourne, Australia), IgA (Abcam, Sydney, Australia) and C4BP (Athens Research and Technology, Athens, USA) via single cycle kinetics, using a Biacore T200 (GE Healthcare, Sweden) at 20 °C. Detailed protocols are provided in the supplementary data.

Results

The emm-cluster system

Near complete *emm* sequences from 1086 isolates collected from 31 countries and belonging to 175 *emm*-types were used [9] to establish the *emm*-cluster system. As the *emm*-type is predictive of the whole M protein sequence [9], a single representative sequence for each of the 175 *emm*-types was selected for phylogenetic analysis (Table S1). Apart from 6 outlier proteins, two well-supported clades (Fig. 1; X and Y; 85 and 84 proteins respectively) were defined based on the general organization of the tree (Fig. 1). Clade Y was divided into two major sub-clades (Y1 and Y2). Clade X, sub-clades Y1 and Y2 were further subdivided into 48 *emm*-clusters. Thirty-two *emm*-clusters contained a single M protein (Fig.1 and Table 1). Notably, the number of *emm*-clusters comprising a single protein was higher in clade Y (n = 22) than in clade X (n = 4). The remaining 16 *emm*-clusters possessed multiple M proteins accounting for an additional 143 M proteins. The number of proteins per *emm*-cluster ranged from 2 to 32. Together, the six largest *emm*-clusters (E2-6 and D4) accounted for 101 M proteins indicating that many M proteins are highly related in sequence.

To better understand the phylogeny presented in Figure 1, the sequence from each protein was divided into three sections (See supplementary data). The tree based on the highly-conserved C-terminus regions (73% average pairwise identity, 11% of the sites identical in the multiple alignment) confirmed the general organization of 2 major clades (data not shown). The central regions, the length of which varied from 68 to 215 residues, were much more divergent (19% average pairwise identity), but strongly supported most of the previously defined *emm*-clusters (data not shown). As expected [15], the tree based on the amino-terminus region was not well supported due to low levels of sequence identity (10% average pairwise identity, no identical sites); however, it revealed several *emm*-types having closely related sequences, most of which were in the same *emm*-cluster group (data not shown).

To assess adaptive evolution, individual codons of M protein were analyzed for positive selection. Data show that the amino-terminal portion is largely under diversifying selection whereas the carboxy-terminal region is highly constrained (Fig. 1 and table S2). Importantly, different patterns of selective pressure were noted for different *emm*-clusters. The proportion of the mature M protein under diversifying selection varied from only 15-20% (the first 50 amino terminal residues) for some *emm*-clusters, to > 60% of the protein (the amino terminus plus central region) for other *emm*-clusters (Fig. 1 and Table S2). Only some *emm*-clusters had codons under diversifying selection within the carboxy-terminal region. Lastly, a unique pattern of neutral evolution was observed for *emm*-cluster A-C3, containing the clinically important M1 protein [2], indicating a higher degree of sequence flexibility across the complete sequence.

In summary, phylogenetic analysis confirmed that some M proteins are highly divergent from all others (32 single protein *emm*-clusters) while the majority (143 *emm*-types) are closely related and can be grouped into 16 homogeneous and well-supported *emm*-clusters whose evolution was driven by distinct selective pressures.

A functional classification

A diverse array of M-protein functions has been described, many of which involve binding to host proteins, which subsequently mediate bacterial virulence and/or provide protection against innate immune responses [4]. Functional analysis of representative M proteins from each of the dominant *emm*-clusters was undertaken to assess binding to key host proteins known to interact with M proteins (Table S3) [4]. M proteins belonging to clades X versus Y displayed distinct functional profiles, with immunoglobulin and C4BP-binding restricted largely to clade X and plasminogen- and fibrinogen-binding restricted to clade Y. Plasminogen-binding was further restricted to *emm*-cluster D4, indicating that these M proteins are highly specialised in function. Comparison of the *emm*-cluster D4 protein sequences with the previously published M protein plasminogen-binding motif [24, 25] and crystal structure data [26] revealed the presence of a highly-conserved plasminogen motif found exclusively in all *emm*-cluster D4 M proteins, and in the M140 protein, positioned just outside *emm*-cluster D4 (Fig. 1 and 2). This motif can therefore be considered predictive of plasminogen-binding M proteins.

High affinity IgA-binding was exhibited by M proteins associated with *emm*-clusters E1 and E6, with affinity constants ranging from 0.66-5.36 nM (Table S3). Of the four proteins functionally assessed from *emm*-cluster E6, all except M65 bind IgA. The previously described IgA-binding motif [27] has been refined based on these data (Fig. 3C). The refined IgA motif was present in *emm*-cluster E1 and E6 M proteins, and in sub-*emm*-cluster E4.1 (Fig. S1) and 4 M protein types outside these *emm*-clusters (Fig. 1). Many of the proteins included in sub-*emm*-cluster E4.1, such as M22, have been reported to bind IgA [28].

IgG-binding was observed for M proteins in *emm*-clusters E1-E4, E6 and A-C3 and in single *emm*-cluster M57 and M14 proteins (Fig. 1 and 3). *Emm*-cluster A-C3 M proteins contain the 'S' domain, reported to be responsible for IgG-binding in M1 [29]. A refined IgG-binding motif for M protein has been defined (Fig. 3F) and is present in most M proteins from clade X and *emm*-cluster A-C3 (Fig. 1). The motif matches a portion of the previously described EQ-rich region reported for IgG3-binding by M2 protein [30]. This IgG motif is however absent from both M14 and M57 proteins (sub-clade Y1), suggesting the existence of additional sites for IgG-binding.

Fibrinogen-binding was primarily restricted to *emm*-clusters D1, AC3-5 and a few M proteins from sub-clade Y1 (M57, M54, M19, M14). Fibrinogen binding to M5 has been localized to the B repeat domain [31]. For M1, fibrinogen binding was suggested to be dependent on irregularities within the coil-coil structure of the B repeats, specifically as a result of alanine residues at positions 'a' and 'd' within the heptad [32]. Although this region of the M protein has limited sequence similarity among the fibrinogen-binders [33], binding data suggests a more refined fibrinogen-binding motif can be described (Fig. 4).

All *emm*-clusters examined, with the exception of E4, contained representative proteins that bound human serum albumin (HSA) which is in accordance with previous data [34]. Binding of HSA by M proteins has been localized to the C repeat domain [29, 35, 36], and a putative HSA-binding motif proposed (RDLXXSRXAKKXXE) [35]. This motif was present in nearly all sequences from this study, including those that did not bind HSA. Interestingly, studies with the M23 (sub-clade Y1) [36] and M1 (A-C3 *emm*-cluster) [37] proteins suggested that regions adjacent to the C repeat domains are required to stabilize the coiled-coiled conformation essential for interaction with HSA. These data clearly highlight the utility of a whole M protein sequence-based approach for studying interactions between different M protein regions, and the impact of these interactions on the biology and virulence of the organism.

Apart from *emm*-cluster E2, C4BP-binding was exhibited with very high affinity (ranging from 4.7-119.93pM) by M proteins associated with *emm*-clusters belonging to the clade X, while no binding could be demonstrated in clade Y (Table S3 and Fig.1). In *emm*-cluster E4, we however observed that M2 bound C4BP while M102 did not. Binding of C4BP by M proteins has been previously localized to the hypervariable N-terminal region of the M protein, which may explain why a defined binding motif has yet to be identified [38].

Taken together, the *emm*-cluster classification correlates the function of 26 representative M proteins to 6 of the most important host ligands. The classification system is also concordant with refined binding motifs for an additional 119 M proteins. *Emm*-cluster classification is therefore likely to be of biological relevance and may provide insights into clinically relevant aspects of M protein function.

A vaccine development tool

The broadly accepted paradigm states that immunity to GAS infection is M-type specific [10, 11, 18, 19]. The M proteins tested in the seminal publications proposing type-specific immunity for GAS [10, 11] are highly divergent across their entire sequence. Most of these proteins are either in a single protein *emm*-cluster (M6, M5, M14, M26, M24) or representative of a unique member of a larger *emm*-cluster (M1, M2, M3, M12, M13, M15, M41) (Fig. 1). M proteins from different *emm*-clusters have very low sequence identity (average of 35% pairwise identity among the 48 *emm*-clusters) and possess different binding capacities. In striking contrast, M proteins included in the same *emm*-cluster demonstrate, by definition, an average pairwise identity >70% and share similar binding properties. Therefore, the *emm*-cluster system provides a working hypothesis for the recently discovered, but unexplained, cross protection between different *emm*-types [39, 40]. Serum from rabbits immunized with a multivalent vaccine containing amino-terminal peptides from 30 different *emm*-types was tested against 49 *emm*-types not included in the vaccine; unexpectedly, cross-opsonisation and killing was demonstrated for 39 of 49 of the *emm*-types tested [39, 40] (Fig. 1). For 12 *emm*-types, cross-opsonization may be due to sequence identity that resides in the amino-terminus [40]. For the remaining 27 *emm*-types, high sequence identity across the full-length of the M proteins within the same *emm*-cluster, together with similar binding properties, may explain the cross-protection observed. Although the sequence of the vaccine antigen region is different across these proteins, their sequences outside this region are nearly identical (Fig. 5). Most of the M proteins (27/39) demonstrating cross protection in rabbits belong to *emm*-clusters that possess at least one representative included in the vaccine (Fig. 1). M proteins belonging to the D4 *emm*-cluster do not demonstrate a high proportion of cross-protection (4/9 *emm*-types tested). This might be related to the large size of this *emm*-cluster and the single antigen included in the 30-valent vaccine. Outside *emm*-cluster D4, the only exception to the *emm*-cluster-based immunity hypothesis is M124 protein (*emm*-cluster E4) that would be predicted to be cross-opsonized by the 30-valent vaccine.

In some experimental models, antibodies directed to the conserved C-repeat region elicit protective immunity [41]. To assess the impact of this *emm*-cluster system on such vaccine strategies [42-45], the distribution of so-called 'J8' alleles was assessed. The J8 peptide is a leading vaccine candidate that has recently entered into clinical trials. Twenty-two J8 alleles are present amongst the 175 *emm*-types, whereby most J8 alleles differ by a single amino acid residue (Data not shown). *Emm*-clusters are largely predictive of a specific pattern of J8 alleles (Fig. 6). The selective pressure analysis implicated some C-repeat region residues (clade Y, *emm*-cluster E6 of clade X) as being under diversifying selection (Fig. 1, Table S2 and data not shown). This result was repeatedly observed within the various subsets of the tree used in this analysis. The potential impact of such diversifying selection pressure on immune escape is currently unknown but data presented here suggest that a deeper understanding of the relationship between C-repeat allele diversity and vaccine efficacy is required.

A reference-typing tool

The *emm*-clusters can be directly inferred from *emm*-typing results (Table 1). They predict both the C-repeat allelic content (such as the J8 alleles) and the *emm* pattern-typing scheme (Figure 1). *emm* pattern-typing distinguishes three distinct groupings (patterns A-C, D and E) based on the presence and arrangement of *emm* and *emm*-like genes within the GAS genome [46]. Specific *emm*-type share the same *emm* pattern grouping [9, 47] and *emm* pattern correlates well with tissue tropism (impetigo for pattern D, pharyngitis for pattern A-C and both for pattern E) [46]. Patterns A-C and D correspond to the previously called classI/*sof*$^{-}$ M proteins while pattern E correspond to the classII/*sof*$^{+}$ [4]. Our data show that patterns E and A-C M proteins are largely restricted to clade X and Y, respectively. In contrast, pattern D *emm*-types are found in three different portions of the tree. The first pattern D group is the highly specialised plasminogen-binding *emm*-cluster D4. *Emm*-cluster E5 and E6 (clade X) form the second group that equally include pattern D and E M proteins. The third group, although not as cohesive, is represented by the pattern D *emm*-types interspersed with pattern A-C in sub-clade Y1 and Y2. A phylogenetic analysis of the 67 pattern D proteins confirmed this differentiation into three lineages (data not shown). It also confirmed that *emm*-clusters E5-E6 and sub-*emm*-cluster D4.1 share some evolutionary history as previously suggested by the presence of J8.1 allele in sub-*emm*-cluster D4.1 (Fig. 6). Thus, pattern D M proteins form 3 discrete structural groups, implying that there may be multiple mechanisms for skin pathogenesis.

In conclusion, in comparison with the previous typing methods such as *emm* pattern and class I/II, the *emm*-cluster typing system provides complementary information in terms of sequence homology, characterisation of binding capacities to 6 different host ligands, prediction of the J8 vaccine candidate allele content and as a framework for investigating the cross-protection hypothesis.

Discussion

This study represents the first systematic analysis of the numerous GAS M protein variants and proposes a novel functional classification that correlates with sequence analysis. Our results demonstrate that 175 *emm*-types can be grouped into 2 clades, 2 sub-clades and 48 *emm*-clusters, 16 of which encompass 82% of the *emm*-types. The *emm*-clusters represent functionally distinct groups of M proteins, as shown by characterization of host protein binding of 24 representative *emm*-types. The *emm*-cluster system, combined with the structural information on specific binding motifs (data not shown), predicted function for an additional 119 *emm*-types. To date, many of the most thoroughly characterized M proteins belong to either small and divergent *emm*-clusters (eg. M1, M3, M12) or single protein *emm*-clusters (eg. M5, M6). Whilst the study of these *emm*-types is justified based on the ability to cause serious clinical manifestations, our current study suggests caution should be taken when attempting to generalize results to the many other M proteins belonging to the other *emm*-clusters. On the contrary, this classification enabled for the first time a model whereby functional attributes could potentially be ascribed to proteins from the same *emm*-cluster.

An effective GAS vaccine remains elusive. Recent studies show that immunization with a 30-valent vaccine generates an antibody response that cross-opsonizes non-vaccine *emm*-types [39, 40]. This represents a significant paradigm shift in the understanding of GAS immunology, but remains until now largely unexplained. If the cross-protection hypothesis is definitively not solved yet, the *emm*-cluster system provides a necessary framework to investigate this in more detail. Apart from the hypothesis that *emm*-types in the same *emm*-cluster are cross-reactive in nature, alternative hypothesis could be either that exposure to 30 diverse M peptide antigens generates broadly cross-reactive antibodies or that some of the most recently discovered *emm*-types generate in fact cross-reactive antibodies to many *emm*-types, including those inside and outside of the same *emm*-cluster. The fact that *emm*-clusters also correlate with single residue substitutions in the C-repeat region enhances the classification system utility as a vaccine development tool. Experience from vaccines targeting other bacteria such as *Streptococcus pneumoniae* show that the introduction of a vaccine may induce serotype replacement and strain emergence [48]. The *emm*-cluster classification provides a tool to predict this risk and to monitor epidemiological changes that might occur after the introduction of any vaccine.

Emm-clusters were defined based on bioinformatic criteria that allows for simple updating when new sequences are added into the dataset. However, three limitations should be acknowledged: rare outliers were observed; some characteristics, such as fibrinogen-binding capacity, seem to be linked to a higher phylogenetic hierarchy (sub-clades) rather than *emm*-clusters; and some findings (eg., the presence of the IgA-binding motif in sub-*emm*-cluster E4.1) correlate with entities smaller than *emm*-clusters.

The *emm*-cluster typing does not, and is not intended to, replace *emm*-typing but rather constitute a new complementary tool that adds meaningful information and may be widely used to analyze GAS molecular epidemiology. Future experiments aimed at characterising the cross-protection hypothesis might potentially refine the current *emm*-cluster system to provide immediate threshold for determining antigenic novelty. This functional classification and its further improvement will be hosted on the website from the streptococcal reference laboratory at the Centers for Disease Control and Prevention (CDC), Atlanta, USA.

Acknowledgments

The contributing members of the M protein study group (in addition to the authors of this paper) include Michael Batzloff and Rebecca Towers from Australia; Herman Goossens and Surhbi Malhotra-Kumar from Belgium; Luiza Guilherme and Rosangela Torres from Brazil; Donald Low and Allison Mc Geer from Canada; Paula Krizova from Czech Republic; Sawsan El Tayeb from Egypt; Joe Kado from Fiji; Mark van der Linden from Germany; Guliz Erdem from Hawai; Alon Moses and Ran Nir-Paz from Israel; Tadayoshi Ikebe and Haruo Watanabe from Japan; Samba Sow and Boubou Tamboura from Mali; Bard Kittang from Norway; José Melo-Cristino and Mario Ramirez from Portugal; Monica Straut from Romania; Alexander Suvorov and Artem Totolian from Russia; Mark Engel, Bongani Mayosi and Andrew Whitelaw from South Africa; Jessica Darenberg and Birgitta Henriques Normark from Sweden; Chuan Chiang Ni and Jiunn-Jong Wu from Taiwan; Aruni De Zoysa and Androulla Efstratiou from UK; Stanford Shulman and Robert Tanz from USA. We thank Dwight Johnson, and Velusamy Srinivasan for their work in expanding and maintaining the *emm* type database at http://www.cdc.gov/ncidod/biotech/strep/strepblast.htm. We are grateful to Scott Sammons, who leads CDC's Bioinformatics team within CDC's Division of Scientific Resources in Atlanta, GA, and his colleagues for supporting the global *emm* type database for all these years. We thank Tracy Nero for her advice in aspects of the structural modelling and Roy Robins-Browne for his overall support in this project.

References

1. Carapetis JR, Steer AC, Mulholland EK and Weber M. The global burden of group A streptococcal diseases. Lancet Infect Dis 2005;5:685-94
2. Steer AC, Lamagni T, Curtis N and Carapetis JR. Invasive group a streptococcal disease: epidemiology, pathogenesis and management. Drugs 2012;72:1213-27
3. Dale JB, Fischetti VA, Carapetis JR, et al. Group A streptococcal vaccines: Paving a path for accelerated development. Vaccine 2013;31 Suppl 2:B216-22
4. Smeesters PR, McMillan DJ and Sriprakash KS. The streptococcal M protein: a highly versatile molecule. Trends Microbiol 2010;18:275-282
5. Fischetti VA. Streptococcal M protein: molecular design and biological behavior. Clin Microbiol Rev 1989;2:285-314
6. Whatmore AM, Kapur V, Sullivan DJ, Musser JM and Kehoe MA. Non-congruent relationships between variation in emm gene sequences and the population genetic structure of group A streptococci. Mol Microbiol 1994;14:619-31
7. Beall B, Facklam R and Thompson T. Sequencing emm-specific PCR products for routine and accurate typing of group A streptococci. J Clin Microbiol 1996;34:953-8
8. Facklam RF, Martin DR, Lovgren M, et al. Extension of the Lancefield classification for group A streptococci by addition of 22 new M protein gene sequence types from clinical isolates: emm103 to emm124. Clin Infect Dis 2002;34:28-38
9. McMillan DJ, Dreze PA, Vu T, et al. 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 2013;19:E222-9
10. Denny FW, Jr., Perry WD and Wannamaker LW. Type-specific streptococcal antibody. The Journal of clinical investigation 1957;36:1092-100
11. Lancefield RC. Persistence of type-specific antibodies in man following infection with group A streptococci. The Journal of experimental medicine 1959;110:271-92
12. Smeesters PR, McMillan DJ, Sriprakash KS and Georgousakis MM. Differences among group A streptococcus epidemiological landscapes: consequences for M protein-based vaccines? Expert Rev Vaccines 2009;8:1705-20
13. Steer AC, Law I, Matatolu L, Beall BW and Carapetis JR. Global emm type distribution of group A streptococci: systematic review and implications for vaccine development. Lancet Infect Dis 2009;9:611-6
14. Smeesters PR, Vergison A, Campos D, de Aguiar E, Miendje Deyi VY and Van Melderen L. Differences between Belgian and Brazilian group A Streptococcus epidemiologic landscape. PLoS ONE 2006;1:e10
15. Smeesters PR, Mardulyn P, Vergison A, Leplae R and Van Melderen L. Genetic diversity of Group A Streptococcus M protein: implications for typing and vaccine development. Vaccine 2008;26:5835-42
16. Smeesters PR, Dramaix M and Van Melderen L. The emm-type diversity does not always reflect the M protein genetic diversity--is there a case for designer vaccine against GAS. Vaccine 2010;28:883-5
17. Smeesters PR. Immunity and vaccine development against Streptococcus pyogenes: is emm-typing enough? Proceedings of the Belgian Royal Academies of Medicine in press

18. Wannamaker LW, Denny FW, Perry WD, Siegel AC and Rammelkamp CH, Jr. Studies on immunity to streptococcal infections in man. A.M.A. American journal of diseases of children 1953;86:347-8
19. Watson RF, Rothbard S and Swift HF. Type-specific protection and immunity following intranasal inoculation of monkeys with group A hemolytic streptococci. The Journal of experimental medicine 1946;84:127-42
20. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792-7
21. Gouy M, Guindon S and Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular biology and evolution 2010;27:221-4
22. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC evolutionary biology 2010;10:210
23. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W and Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology 2010;59:307-21
24. Sanderson-Smith ML, Walker MJ and Ranson M. 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. J Biol Chem 2006;281:25965-71
25. Wistedt AC, Ringdahl U, Muller-Esterl W and Sjobring U. Identification of a plasminogen-binding motif in PAM, a bacterial surface protein. Mol Microbiol 1995;18:569-78
26. Rios-Steiner JL, Schenone M, Mochalkin I, Tulinsky A and Castellino FJ. Structure and binding determinants of the recombinant kringle-2 domain of human plasminogen to an internal peptide from a group A Streptococcal surface protein. J Mol Biol 2001;308:705-19
27. Bessen DE. Localization of immunoglobulin A-binding sites within M or M-like proteins of group A streptococci. Infect Immun 1994;62:1968-74
28. Johnsson E, Areschoug T, Mestecky J and Lindahl G. An IgA-binding peptide derived from a streptococcal surface protein. J Biol Chem 1999;274:14521-4
29. Akesson P, Schmidt KH, Cooney J and Bjorck L. M1 protein and protein H: IgGFc- and albumin-binding streptococcal surface proteins encoded by adjacent genes. Biochem J 1994;300 (Pt 3):877-86
30. Pack TD, Podbielski A and Boyle MD. Identification of an amino acid signature sequence predictive of protein G-inhibitable IgG3-binding activity in group-A streptococcal IgG-binding proteins. Gene 1996;171:65-70
31. Waldemarsson J, Stalhammar-Carlemalm M, Sandin C, Castellino FJ and Lindahl G. Functional dissection of Streptococcus pyogenes M5 protein: the hypervariable region is essential for virulence. PLoS ONE 2009;4:e7279
32. McNamara C, Zinkernagel AS, Macheboeuf P, Cunningham MW, Nizet V and Ghosh P. Coiled-coil irregularities and instabilities in group A Streptococcus M1 are required for virulence. Science 2008;319:1405-8
33. Ringdahl U, Sjobring U. Analysis of plasminogen-binding M proteins of Streptococcus pyogenes. Methods 2000;21:143-50
34. Sandin C, Carlsson F and Lindahl G. Binding of human plasma proteins to Streptococcus pyogenes M protein determines the location of opsonic and non-opsonic epitopes. Mol Microbiol 2006;59:20-30

35. Retnoningrum DS, Cleary PP. M12 protein from Streptococcus pyogenes is a receptor for immunoglobulin G3 and human albumin. Infect Immun 1994;62:2387-94
36. Hong K. Characterization of group a streptococcal M23 protein and comparison of the M3 and M23 protein's ligand-binding domains. Curr Microbiol 2007;55:427-34
37. Gubbe K, Misselwitz R, Welfle K, Reichardt W, Schmidt KH and Welfle H. C repeats of the streptococcal M1 protein achieve the human serum albumin binding ability by flanking regions which stabilize the coiled-coil conformation. Biochemistry 1997;36:8107-13
38. Persson J, Beall B, Linse S and Lindahl G. Extreme sequence divergence but conserved ligand-binding specificity in Streptococcus pyogenes M protein. PLoS Pathog 2006;2:e47
39. Dale JB, Penfound TA, Chiang EY and Walton WJ. New 30-valent M protein-based vaccine evokes cross-opsonic antibodies against non-vaccine serotypes of group A streptococci. Vaccine 2011;29:8175-8
40. Dale JB, Penfound TA, Tamboura B, et al. Potential coverage of a multivalent M protein-based group A streptococcal vaccine. Vaccine 2013
41. Bessen D, Fischetti VA. Influence of intranasal immunization with synthetic peptides corresponding to conserved epitopes of M protein on mucosal colonization by group A streptococci. Infect Immun 1988;56:2666-72
42. Pandey M, Wykes MN, Hartas J, Good MF and Batzloff MR. Long-Term Antibody Memory Induced by Synthetic Peptide Vaccination Is Protective against Streptococcus pyogenes Infection and Is Independent of Memory T Cell Help. Journal of immunology 2013
43. Bauer MJ, Georgousakis MM, Vu T, et al. Evaluation of novel Streptococcus pyogenes vaccine candidates incorporating multiple conserved sequences from the C-repeat region of the M-protein. Vaccine 2012;30:2197-205
44. Guerino MT, Postol E, Demarchi LM, et al. HLA class II transgenic mice develop a safe and long lasting immune response against StreptInCor, an anti-group A streptococcus vaccine candidate. Vaccine 2011;29:8250-6
45. Batzloff MR, Hayman WA, Davies MR, et al. Protection against group A streptococcus by immunization with J8-diphtheria toxoid: contribution of J8- and diphtheria toxoid-specific antibodies to protection. J Infect Dis 2003;187:1598-608
46. Bessen DE, Lizano S. Tissue tropisms in group A streptococcal infections. Future Microbiol 2010;5:623-38
47. McGregor KF, Spratt BG, Kalia A, et al. Multilocus sequence typing of Streptococcus pyogenes representing most known emm types and distinctions among subpopulation genetic structures. J Bacteriol 2004;186:4285-94
48. Hausdorff WP, Van Dyke MK and Van Effelterre T. Serotype replacement after pneumococcal vaccination. Lancet 2012;379:1387-8; author reply 1388-9
49. Sanderson-Smith ML, Dowton M, Ranson M and Walker MJ. The plasminogen-binding group A streptococcal M protein-related protein Prp binds plasminogen via arginine and histidine residues. J Bacteriol 2007;189:1435-40
50. Sanderson-Smith ML, Dinkla K, Cole JN, et al. M protein-mediated plasminogen binding is essential for the virulence of an invasive Streptococcus pyogenes isolate. FASEB J 2008;22:2715-22

Table 1: Distribution of *emm*-types per *emm*-cluster

***emm*-types**	***Emm*-cluster**
4, 60, 78, 165 (st11014), 176 (st213)	E1
13, 27, 50 (50/62), 66, 68, 76, 90, 92, 96, 104, 106, 110, 117, 166 (st1207), 168 (st1389)	E2
9, 15, 25, 44 (44/61), 49, 58, 79, 82, 87, 103, 107, 113, 118, 144 (stknb1), 180 (st2460), 183 (st2904), 209 (st6735), 219 (st9505), 231 (stNS292)	E3
2, 8, 22, 28, 73, 77, 84, 88, 89, 102, 109, 112, 114, 124, 169 (st1731), 175 (st212), 232 (stNS554)	E4
34, 51, 134 (st2105), 137 (st465), 170 (st1815), 174 (st211), 205 (st5282)	E5
11, 42, 48, 59, 63, 65 (65/69), 67, 75, 81, 85, 94, 99, 139 (st7323), 158 (stxh1), 172 (st2037), 177 (st2147), 182 (st2861UK), 191 (st369)	E6
164 (st106M), 185 (st2917), 211 (st7406), 236 (sts104)	Single protein *emm*-cluster cl
36, 54, 207 (st6030)	D1
32, 71, 100, 115, 213 (st7700)	D2
123, 217 (st809)	D3
33, 41, 43, 52, 53, 56, 56.2 (st3850), 64, 70, 72, 80, 83, 86, 91, 93, 98, 101, 108, 116, 119, 120, 121, 178 (st22), 186 (st2940), 192 (st3757), 194 (st38), 208 (st62), 223 (stD432), 224 (stD631), 225 (stD633), 230 (stNS1033), 242 (st2926)	D4
97, 157 (stn165), 184 (st2911)	D5
46, 142 (st818)	A-C1
30, 197 (st4119)	A-C2
1, 163 (st412), 227 (stil103), 238 (1-2), 239 (1-4)	A-C3
12, 39, 193 (st3765), 228 (stil62), 229 (stmd216)	A-C4
3, 31, 133 (st1692)	A-C5
5, 6, 14, 17, 18, 19, 23, 24, 26, 29, 37, 38 (38/40), 47, 57, 74, 105, 122, 140 (st7395), 179 (st221), 218 (st854), 233 (stNS90), 234 (stpa57)	Single protein *emm*-cluster clade Y
55, 95, 111, 215 (st804), 221 (stCK249), 222 (stCK401)	Single protein *emm*-cluster outlier

emm-type nomenclature has recently been revised to a simplified system that includes the *emm*-types M1 to M242. A correspondence table between the old and new nomenclature is accessible at the CDC website (http://www.cdc.gov/ncidod/biotech/strep/strepblast.htm).

Figure Legends

Fig. 1. Phylogeny of M proteins and the *emm*-cluster classification system. Phylogenetic inferences of M protein sequences from 175 *emm*-types drawn by PhyML. The tree is drawn to scale, with branch lengths in the same units (number of amino acid substitutions per site) as those of the evolutionary distances used for the phylogenetic tree. Approximate Likelihood-Ratio Test values >80% are indicated at the nodes. The tree has two main clades: Clade X is comprised of 6 main *emm*-clusters (E1-E6) whereas clade Y is divided into two sub-clades (Y1 and Y2) that are then sub-divided into 10 main *emm*-clusters (D1 to D5 and A-C1 to A-C5). Six outlier *emm*-types are indicated by dashed lines (See supplementary data). Selective pressures analyses of M protein sequences are shown for the different *emm*-clusters and/or clades of the tree. The sites above the red and orange lines are positively selected (probability >0.95 and 0.5 respectively). M protein binding data to six human proteins are shown: dark-shaded color boxes indicate experimentally confirmed binding by M protein, white boxes indicate no binding, and light-shaded boxes represent predicted binding based on the presence of consensus binding motifs (plasminogen, IgA, IgG and fibrinogen). Hash marks (#) indicate proteins that bind by experimental testing but lack the predicted binding motif. The cross (+) indicates the presence of the IgA binding motif in the absence of experimental binding. Findings on cross-opsonisation elicited by the 30-valent vaccine [39, 40]: VA stands for vaccine antigen, black boxes indicate the presence of cross-opsonising antibodies in rabbit, and shaded boxes indicate a lack of cross-opsonisation. The *emm* pattern (pattern E, D and A-C) is indicated for each *emm*-type [9]. The asterisks (*) mark the representative M proteins expressed in *E. coli*.

Fig. 2. Binding of plasminogen by M proteins. Single cycle kinetic SPR sensorgrams for the interaction of M proteins with plasminogen are shown (**A**). Human glu-plasminogen was injected over immobilized M protein (concentrations of 7.5, 15, 30, 60, and 120 nM). Binding data was calculated 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). Only the four proteins from *emm*-cluster D4 bound plasminogen. Based on the protein sequence alignment of the 4 plasminogen-binding M proteins (**B**), the targeted mutagenesis data available in the literature [49, 50], and analysis of our protein dataset, a refined motif for M protein plasminogen-binding was defined (**C**). The search for this motif amongst the 175 *emm*-types yielded positive results for all M proteins of *emm*-cluster D4 and the closely related M140 protein (Figure 1); all other M proteins were negative for this motif. Plasminogen binding has not been described for any M protein outside these 33 proteins. Seventeen and 16 of the 33 proteins contained duplicate or single binding motifs respectively. The result of the multiple alignment of the 50 sequences containing a plasminogen binding motif is shown as a Sequence Logo representation (**B**).

Fig. 3. Binding of IgA and IgG by M proteins. Five of six proteins from *emm*-clusters E1 and E6 bound IgA (**A**). Based on the protein sequence alignment of the 5 IgA-binders (**B**) and the data available in the literature [27], a refined motif for binding of IgA by M protein is defined (**C**). Motif searching gave positive results for 28 *emm*-types in three main (sub-)*emm*-clusters (E1, E6 and E4.1). M proteins of four other *emm*-types were positive for this motif: M236 (close to E6), M44 (E3), M242 (D4) and M215 (Outlier Fig. 1). Findings from a multiple alignment of the 35 IgA-binding

sequences (3 *emm*-types contain a duplicate motif) are shown as a Sequence Logo representation (**B**). All 13 recombinant M proteins from *emm*-cluster E1-4, E6 and A-C3 bound IgG (Figure 1), as determined by surface plasmon resonance (SPR). Single cycle kinetic sensorgrams are shown for 4 representative M proteins (**D**). The protein sequence alignment of 4 representative IgG-binders (**E**) led to the definition of a motif for binding of IgG by M protein (**F**). Findings from a multiple alignment of the 101 IgG-binding sequences (15 *emm*-types contains duplicate motif) are shown as a Sequence Logo representation (**E**).

Figure 4. Binding of fibrinogen by M proteins. Eight recombinant M proteins from clade Y bound fibrinogen (Figure 1) and representative single cycle kinetic SPR sensorgrams are shown for 4 *emm*-types (**A**). Based on the fibrinogen-binding motif sequence previously described for M5 [31] and the alignment of fibrinogen-binders (**B**) a refined fibrinogen-binding motif is proposed (**C**). This motif was present in 25 M proteins from clade Y, but absent from M57. Findings from the multiple alignment of the 42 fibrinogen-binding sequences (9 and 4 proteins contain duplicate and triplicate motifs respectively) are shown as a Sequence Logo representation (**B**).

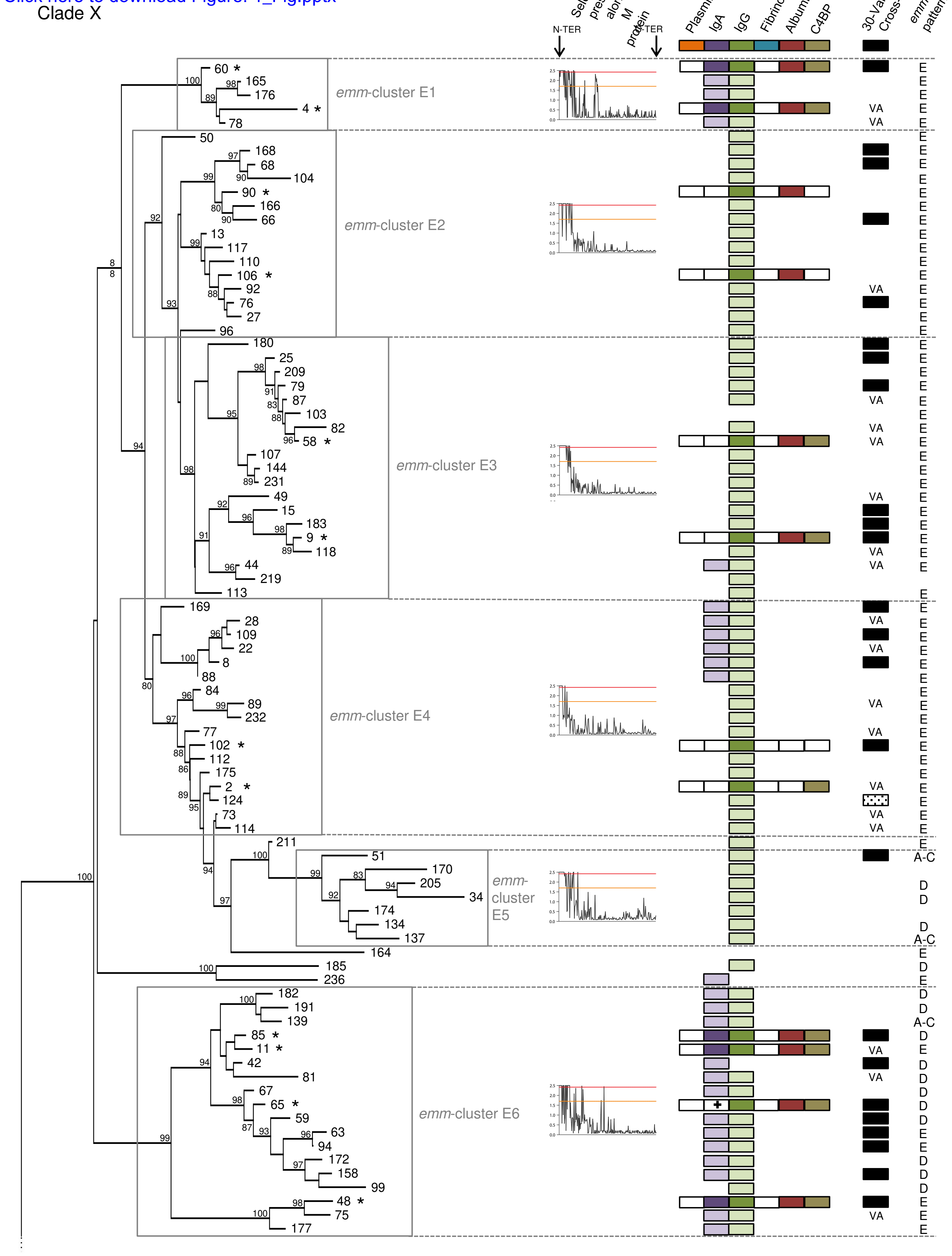
Figure 5. Correlation between immunological cross-protection and M protein sequence *emm*-clusters. M proteins sharing the same *emm*-cluster have different amino-terminal regions but possess nearly identical sequences for the rest of the protein (Figure 1); *emm*-cluster E6 is shown as an example (**A**). VA stands for vaccine antigen and indicates the M proteins of *emm*-cluster E6 that are included in the 30-valent vaccine [39]. The black squares show the M proteins that demonstrate cross-opsonization in rabbits following vaccination with the 30-valent vaccine [39,

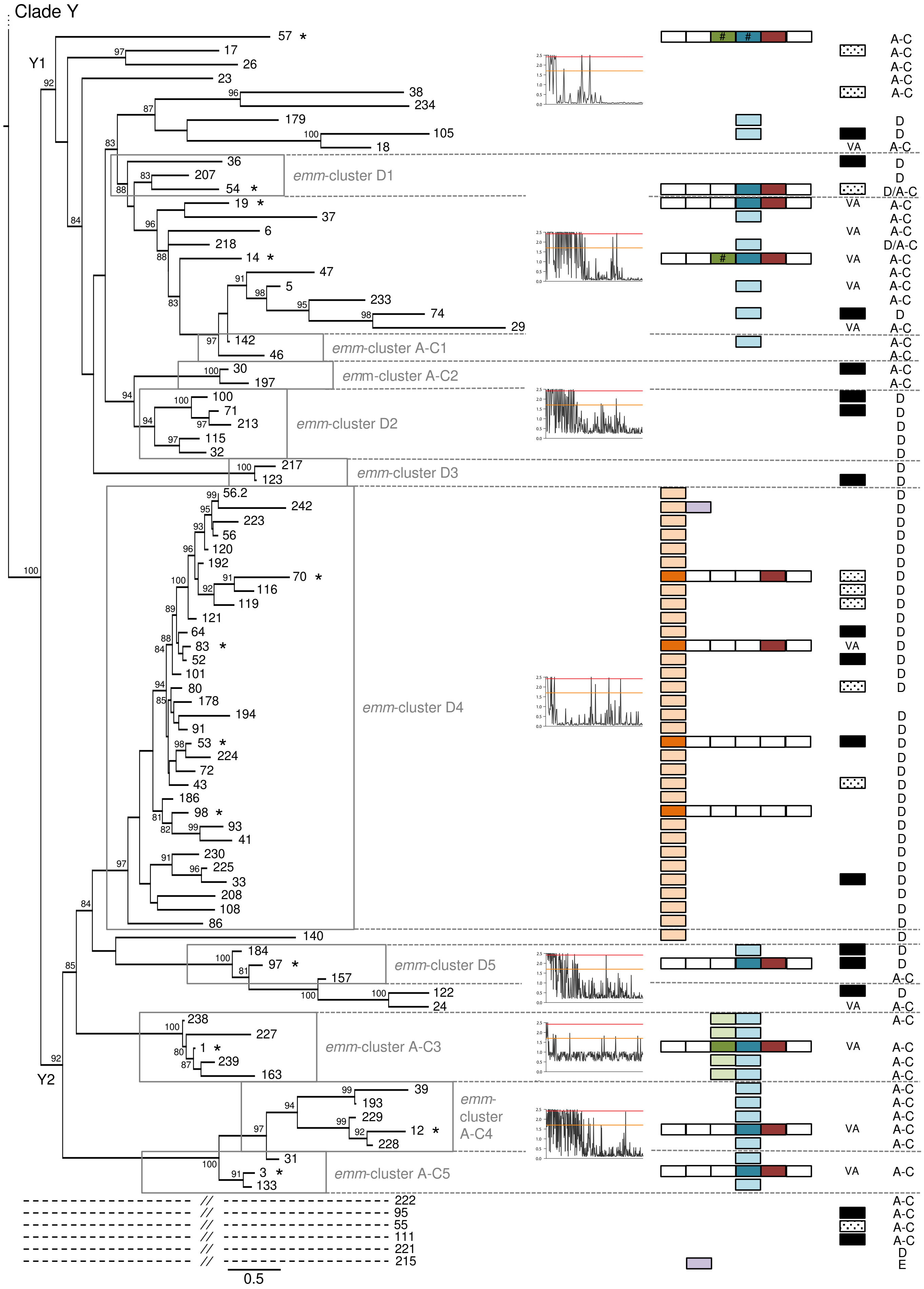
40]. The pairwise identity values of the whole M protein sequences within a *emm*-cluster is by definition >70% (average pairwise identity of 77.8%) (**B**). Multiple sequence alignments are shown for the whole M protein (**C**) and for the 50 amino-terminal residues only (**D**). Amino acid differences are highlighted by color shading and identity is represented in grey. Red boxes highlight vaccine antigens (the 50 amino-terminal residues). Pairwise identity values for the first 50 residues (average pairwise identity of 33.3%) is shown (**E**).

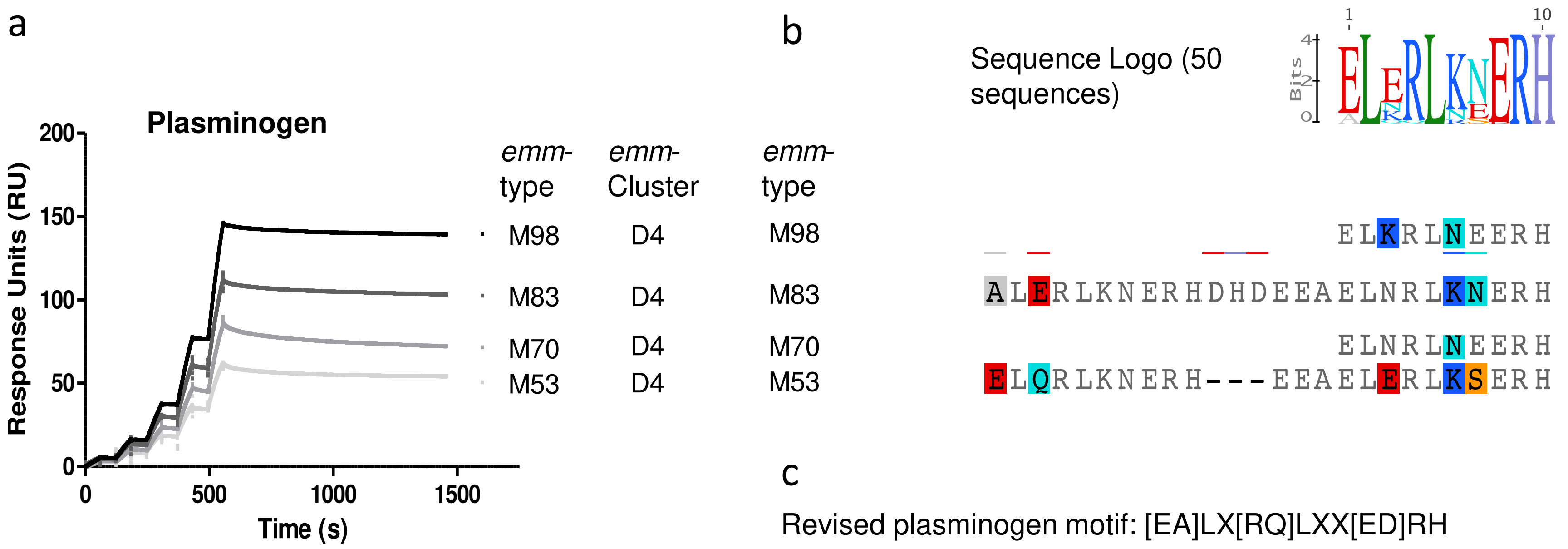
Figure 6. The *emm*-cluster typing system predicts the presence of J8 alleles. The presence of 11 alleles of the J8 vaccine antigen is presented for each *emm*-type. 22 different alleles of the J8 vaccine antigen were found in our dataset. The 11 alleles present in at least 5 *emm*-types were represented in this figure. A correlation between clades, sub-clades and *emm*-clusters with the presence of specific J8 alleles is evident. J8, the vaccine candidate, is present in all but 13 *emm*-types from clade Y while absent from clade X. In contrast, J8.1 is present in 5 of the 6 *emm*-clusters constituting clade X. 173 of the 175 *emm*-types included in this study contains either J8 or J8.1 (M93, M122 and M224 do not). J8.29 and J8.8 are exclusively present in *emm*-cluster E2, E3 and E4. They are never present together in an *emm*-type and only differ by a single amino acid. J8.36 is exclusively present in *emm*-cluster E6, while a combination of J8.1-J8.12 and J8.12-J8.40 are specific for *emm*-cluster E1 and E5 respectively. The whole clade Y1 is characterized by a combination of J8, J8.2 and J8.4. In contrast, J8.4 is rarely found in clade Y2. J8.84 is specific of *emm*-clusters A-C4 and A-C5. Interestingly, *emm*-cluster D4 seems divided by the presence of either J8.1 or J8.57.

Figure 1

Click here to download Figure: 1_Fig.pptx



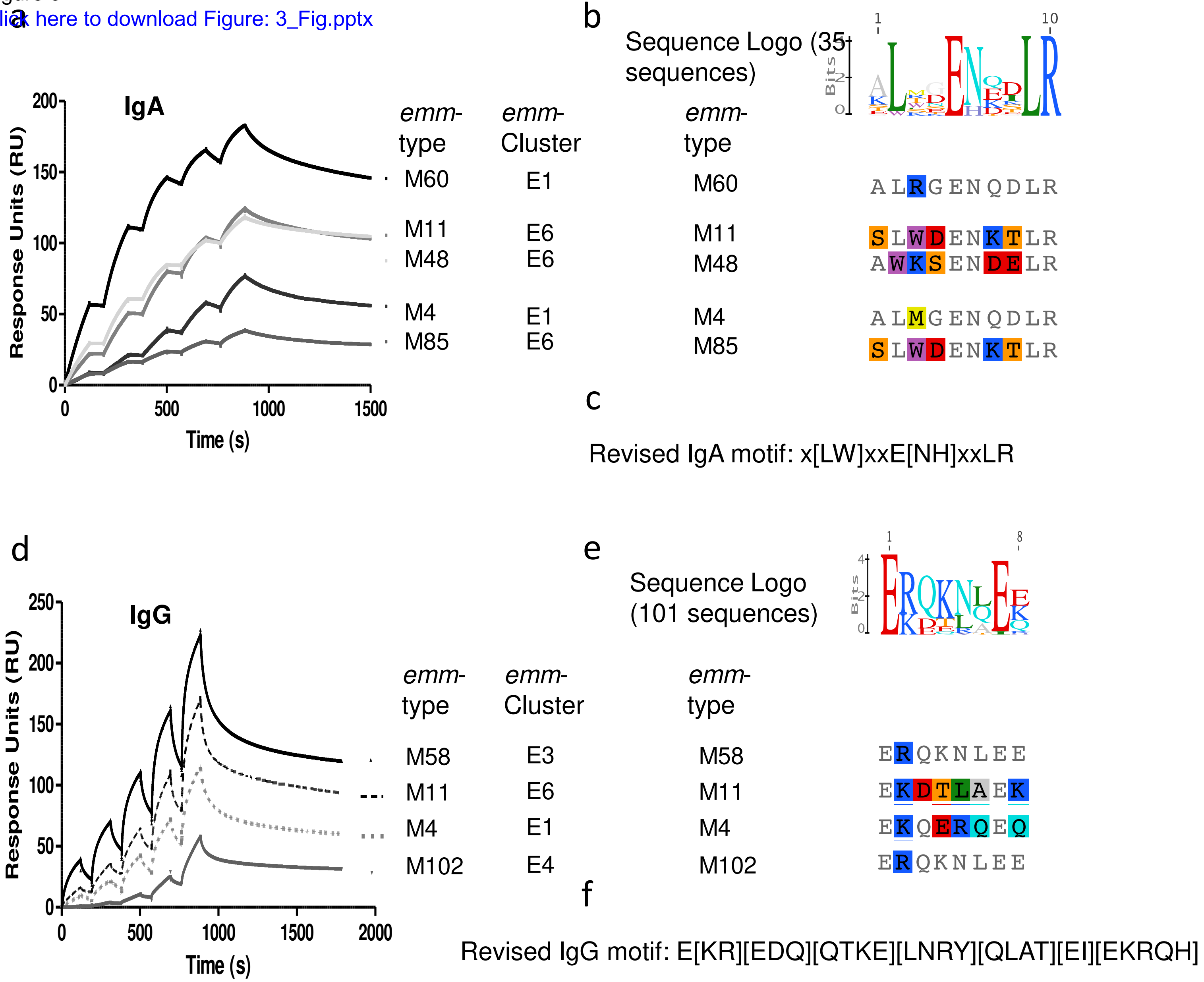

Clade Y
Y1
57 *
17
26
23
38
234
179
105
18
36
207
54 *
emm-cluster D1
19 *
37
6
218
14 *
47
5
233
74
29
142
46
emm-cluster A-C1
30
197
emm-cluster A-C2
100
71
213
115
32
emm-cluster D2
217
123
emm-cluster D3
56.2
242
223
56
120
192
70 *
116
119
121
64
83 *
52
101
80
178
194
91
53 *
224
72
43
186
98 *
93
41
230
225
33
208
108
86
emm-cluster D4
140
184
97 *
157
122
24
emm-cluster D5
238
227
1 *
239
163
emm-cluster A-C3
39
193
229
12 *
228
31
emm-cluster A-C4
3 *
133
emm-cluster A-C5
Y2
222
95
55
111
221
215
0.5
#
#
VA
A-C
D
D/A-C
E



Revised plasminogen motif: [EA]LX[RQ]LXX[ED]RH

Figure 3

Click here to download Figure: 3_Fig.pptx



emm-type	*emm*-Cluster
M60	E1
M11	E6
M48	E6
M4	E1
M85	E6

emm-type	
M60	ALRGENQDLR
M11	SLWDENKTLR
M48	AWKSENDELR
M4	ALMGENQDLR
M85	SLWDENKTLR

Revised IgA motif: x[LW]xxE[NH]xxLR

emm-type	*emm*-Cluster
M58	E3
M11	E6
M4	E1
M102	E4

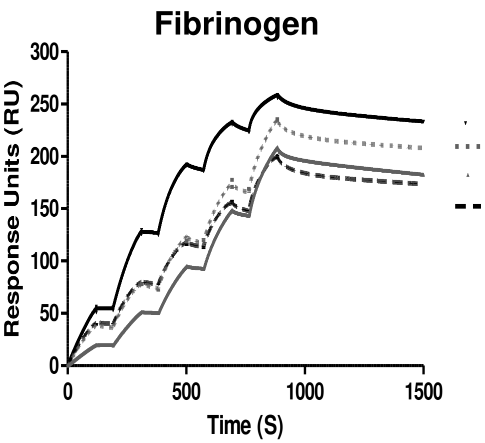
emm-type	
M58	ERQKNLEE
M11	EKDTLAEK
M4	EKQERQEQ
M102	ERQKNLEE

Revised IgG motif: E[KR][EDQ][QTKE][LNRY][QLAT][EI][EKRQH]

Figure 4

Click here to download Figure: 4_Fig.pptx

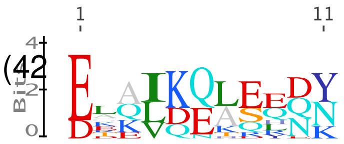
a



emm-type	*emm*-Cluster	*emm*-type	
M54	D1	M54	DLEVKNHDLEN
M3	A-C5	M3	DDQIKQLEEQK
M14	M14	M14	EKKVQETEYNN
M1	A-C3	M1	ELAIDQASRDY

b

Sequence Logo (42 sequences)

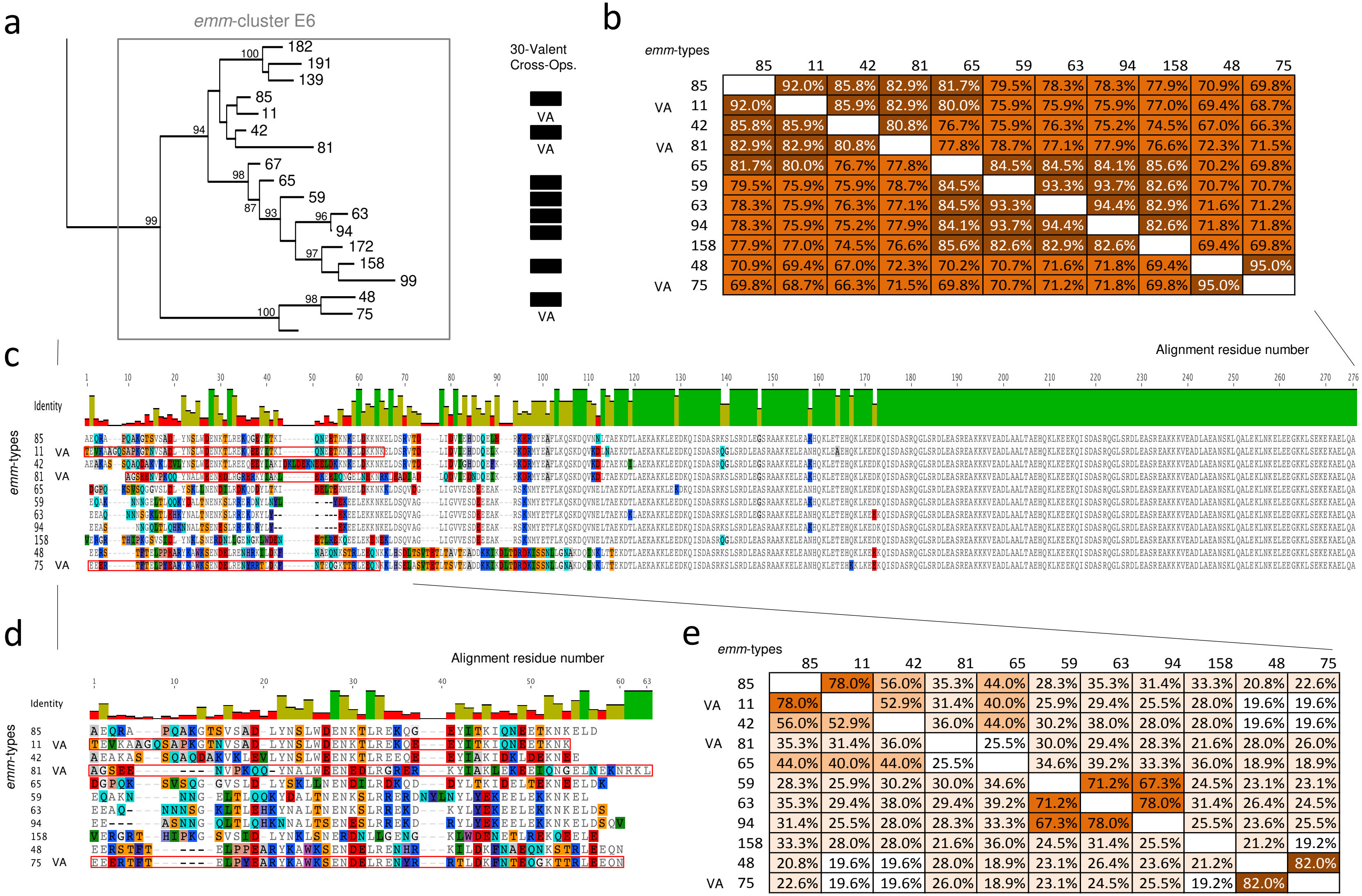


c

Revised fibrinogen binding motif:
[ED]x[QAEK][IVL][DQK][QEN][LATKH][SQREDN][EQRYHL][QDNE][YKNQ]

Figure 5

Click here to download Figure: 5_Fig.pptx



b

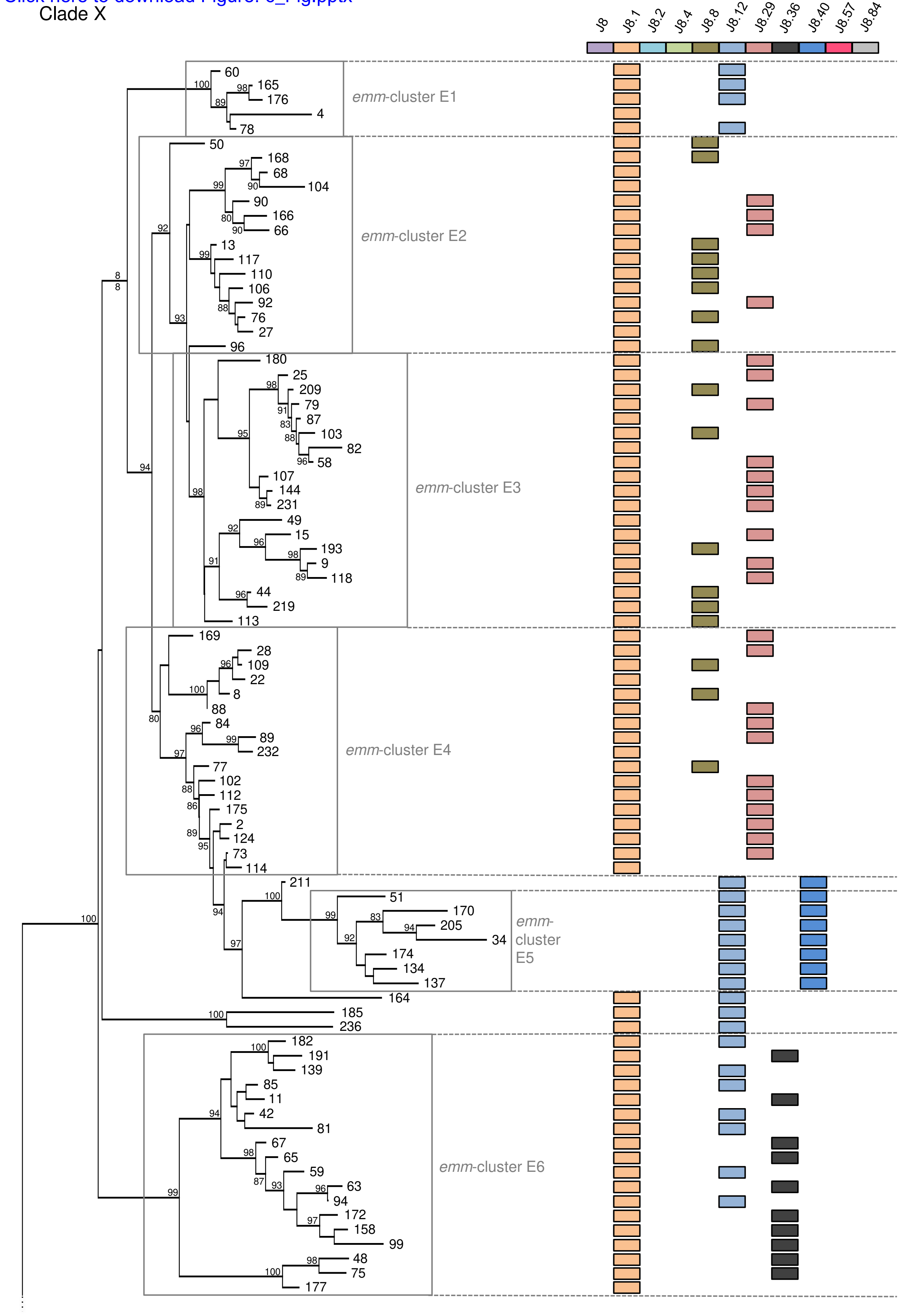
	emm-types	85	11	42	81	65	59	63	94	158	48	75
	85		92.0%	85.8%	82.9%	81.7%	79.5%	78.3%	78.3%	77.9%	70.9%	69.8%
VA	11	92.0%		85.9%	82.9%	80.0%	75.9%	75.9%	75.9%	77.0%	69.4%	68.7%
	42	85.8%	85.9%		80.8%	76.7%	75.9%	76.3%	75.2%	74.5%	67.0%	66.3%
VA	81	82.9%	82.9%	80.8%		77.8%	78.7%	77.1%	77.9%	76.6%	72.3%	71.5%
	65	81.7%	80.0%	76.7%	77.8%		84.5%	84.5%	84.1%	85.6%	70.2%	69.8%
	59	79.5%	75.9%	75.9%	78.7%	84.5%		93.3%	93.7%	82.6%	70.7%	70.7%
	63	78.3%	75.9%	76.3%	77.1%	84.5%	93.3%		94.4%	82.9%	71.6%	71.2%
	94	78.3%	75.9%	75.2%	77.9%	84.1%	93.7%	94.4%		82.6%	71.8%	71.8%
	158	77.9%	77.0%	74.5%	76.6%	85.6%	82.6%	82.9%	82.6%		69.4%	69.8%
	48	70.9%	69.4%	67.0%	72.3%	70.2%	70.7%	71.6%	71.8%	69.4%		95.0%
VA	75	69.8%	68.7%	66.3%	71.5%	69.8%	70.7%	71.2%	71.8%	69.8%	95.0%	

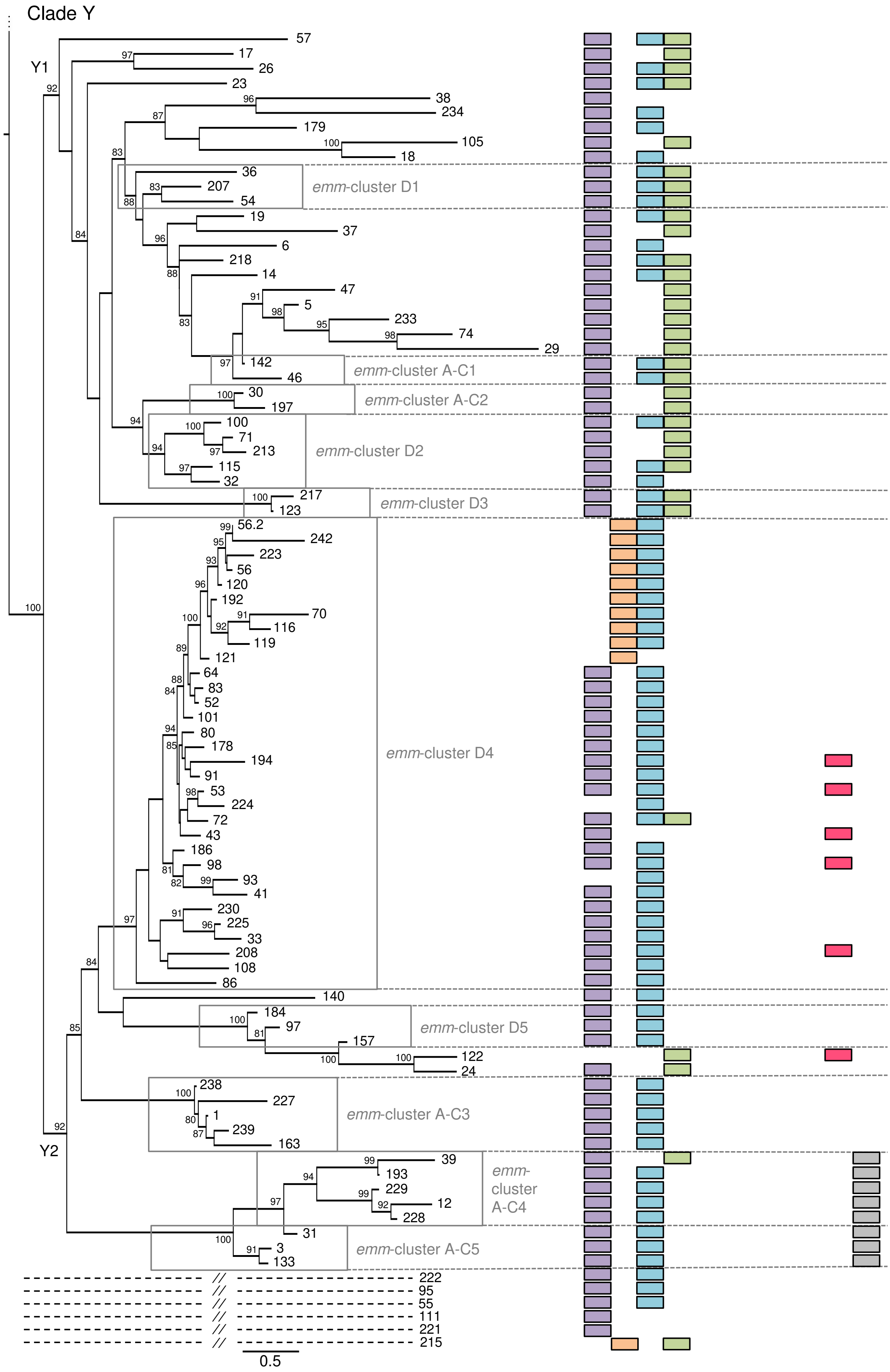
e

	emm-types	85	11	42	81	65	59	63	94	158	48	75
	85		78.0%	56.0%	35.3%	44.0%	28.3%	35.3%	31.4%	33.3%	20.8%	22.6%
VA	11	78.0%		52.9%	31.4%	40.0%	25.9%	29.4%	25.5%	28.0%	19.6%	19.6%
	42	56.0%	52.9%		36.0%	44.0%	30.2%	38.0%	28.0%	28.0%	19.6%	19.6%
VA	81	35.3%	31.4%	36.0%		25.5%	30.0%	29.4%	28.3%	21.6%	28.0%	26.0%
	65	44.0%	40.0%	44.0%	25.5%		34.6%	39.2%	33.3%	36.0%	18.9%	18.9%
	59	28.3%	25.9%	30.2%	30.0%	34.6%		71.2%	67.3%	24.5%	23.1%	23.1%
	63	35.3%	29.4%	38.0%	29.4%	39.2%	71.2%		78.0%	31.4%	26.4%	24.5%
	94	31.4%	25.5%	28.0%	28.3%	33.3%	67.3%	78.0%		25.5%	23.6%	25.5%
	158	33.3%	28.0%	28.0%	21.6%	36.0%	24.5%	31.4%	25.5%		21.2%	19.2%
	48	20.8%	19.6%	19.6%	28.0%	18.9%	23.1%	26.4%	23.6%	21.2%		82.0%
VA	75	22.6%	19.6%	19.6%	26.0%	18.9%	23.1%	24.5%	25.5%	19.2%	82.0%	

Figure 6

Click here to download Figure: 6_Fig.pptx




Clade Y
Y1
57
17
26
23
38
234
179
105
18
36
207
54
emm-cluster D1
19
37
6
218
14
47
5
233
74
29
142
46
emm-cluster A-C1
30
197
emm-cluster A-C2
100
71
213
115
32
emm-cluster D2
217
123
emm-cluster D3
56.2
242
223
56
120
192
70
116
119
121
64
83
52
101
80
178
194
91
53
224
72
43
186
98
93
41
230
225
33
208
108
86
emm-cluster D4
140
184
97
157
emm-cluster D5
122
24
238
227
1
239
163
emm-cluster A-C3
Y2
39
193
229
12
228
31
emm-cluster A-C4
3
133
emm-cluster A-C5
222
95
55
111
221
215
0.5

A systematic and functional classification of *Streptococcus pyogenes* that serves as a new tool for molecular typing and vaccine development

Supplementary material

Supplementary Data

Absence of significant recombination event in the dataset

Several lines of evidences demonstrate that recombination is not frequent among this dataset. Firstly, a large dataset of 1086 GAS isolates representing 175 different *emm*-types recovered from 31 countries on six continents has been used to search for recombination events. Sequences have been carefully annotated for the presence of repeated sequences and visually analysed using the Geneious® 6.1 software. M proteins assigned to the same *emm*-type are highly conserved across their surface exposed portions, despite differences in both geographical origins and clinical manifestations [1]. Single M-type paired with multiple, highly divergent regions, was not observed, suggestive of a relative lack of recombination events in this dataset. Secondly, we used the RDP software [2] to detect recombination among the 175 representative *emm*-types sequences included in the present study and were unable to detect significant recombination event. Thirdly, concordant evolutionary signals were detected when phylogenetic trees were reconstructed with either different regions of the M protein separately (see below) or by using different algorithm such as Neighbor-joining, Maximum Likelihood Ratio and UPGMA on the complete M protein sequence, suggesting again a lack of recombination events.

Informative sites for phylogenetic analysis and phylogenetic controls

Informative sites were extracted from these alignments using default criteria from BMGE [3]. BMGE trims multiple sequence alignments according to an entropy score which is calculated for each site. This score depends on a similarity matrix (BLOSUM or PAM) and the proportion of gaps at the site. The alignment produced by Muscle was 682 amino-acids long, and BMGE retained 249 of them.

For some analyses (including for searching for recombination event, see above), the sequence from each protein was divided into three regions, designated as the amino-terminus, central region and carboxy-terminus. The amino-terminus is defined as the first 50 amino terminal amino acids of the mature M protein. The central region starts at residue 51 and extends to the residue before the first C-repeat (the definition of the C-repeat can be found in reference [1]). The carboxy-terminal end of the M protein is defined as the region starting at the first residue from the first C-repeat to the first residue of the first D repeat as previously described [1].

The *emm*-clusters presented in figure 1 have been sub-divided into sub-*emm*-clusters using higher sequence identity thresholds based on three bioinformatics criteria: 1) monophyletic of paraphyletic nature 2) demonstrating a minimal average pairwise identity of 80% between all M proteins included and 3) demonstrating a minimal pairwise identity of 70% between each pairs of M proteins included (the sites including gaps in the C-repeat regions (e.g.: variation in the number of C-repeats) were excluded from sequence identity calculation). Most of the analyses presented in this paper do not support the use of the sub-*emm*-cluster level as a meaningful threshold for classification of M variants.

Outlier proteins

Six M proteins (M55, M95, M111, M215, M221 and M222) were initially excluded from this tree based on a significantly lower sequence similarity with the other 169 M sequences; their position on the tree (Fig.1) is not possible to define reason why they are indicated with dashed lines in Fig 1. Based on the phylogenetic analysis of the 3

M protein regions (data not shown), their *emm*-pattern and the presence of specific binding motifs and J8 alleles (Fig. 1 and Fig. 6), one could however argue that they are more closely related to the Y clade than the X clade.

Selective pressure analysis

In order to estimate the ratio of non-synonymous to synonymous substitution rates, multiple codon alignments were generated from the corresponding aligned protein sequences using PAL2NAL [4]. The codeml program from the PAML package [5] was implemented, using "site models". The graphs were generated from the output of the Bayes empirical Bayes. All analyses have a Likelihood-Ratio Test (LRT) value >99%. For the data output, values >1 were normalized so that they range from 1 to 2.5. Sites for which this probability is highly significant (i.e. probability > 0.95) have a normalized omega value >2.4. The sites indicated as positively selected, with a probability > 0.5 have a normalized omega value >1.7. *emm*-types M49 and M219 (*emm*-cluster E3) were excluded from selective pressure analyses because significantly shorter than other *emm*-types.

Cloning, expression and purification of recombinant M proteins

Recombinant M proteins were produced essentially as previously described [6], using the expression vector pGEX-2T. The nucleotide sequences of the cloned genes were confirmed. The purity and secondary structure of recombinant proteins was confirmed using Western blot analysis and circular dichroism spectroscopy as described previously [6] (data not shown).

Binding assays

Recombinant M proteins were analyzed for their binding affinity to human glu-plasminogen (Haemotologic Technologies Inc., Essex Junction, USA) via single cycle kinetics, using a Biacore T200 (GE Healthcare, Sweden) at 20 °C. Anti-histidine monoclonal antibody was immobilized to a series S CM5 sensor chip (Biacore AB) using an amine coupling kit as per the manufacturer's instructions. 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, anti-histidine monoclonal antibody was coated onto the chip at 100 µg ml^{-1} 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 response units was reached. Flow cell 1 was left blank to serve as a control. Host proteins 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^{-1} 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^{-1}. Interactions were analyzed by non-linear fitting of the single cycle kinetic sensograms according to a 1:1 Langmuir binding model using Biacore T200 evaluation software.

Purified histidine-tagged recombinant M protein was analyzed for binding affinity to human fibrinogen (Sigma-Aldrich, Sydney, Australia) IgG (Life Technologies, Melbourne, Australia), IgA (Abcam, Sydney, Australia), C4BP (Athens Research and Technology, Athens, USA) and albumin (Sigma-Aldrich, Sydney, Australia) 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^{-1} and washed with 3 mM EDTA for 60 s at 5 µl min^{-1}. 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^{-1} with a 900 s dissociation period. Regeneration of the surface was achieved with 5 M urea, 300 mM EDTA (pH 8.3) for 50 s at 30 µl min^{-1}. M protein-host interactions were analyzed as described above.

Supplementary table S1: Accession numbers and strain collection

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.

Supplementary table S2: Selective pressure along the M-protein sequences.

Clade, sub-clades and *emm*-cluster (Nbr *emm*-types)	Nbr codon incl. / Nbr codon align. (%)	N-term	Central section	C-term
		Nbr codon + /codon incl. (%)	Nbr codon + /codon incl. (%)	Nbr codon + /codon incl. (%)
Clade X (85)	98/392 (25)	2/4 (50)	1/25 (4)	0/69 (0)
emm-cluster E1 (5)	221/276 (80)	19/43 (44)	0/68 (0)	0/110 (0)
emm-cluster E2* (32)	174/329 (53)	16/18 (89)	3/60 (5)	0/96 (0)
emm-cluster E3 (17)	184/319 (58)	13/13 (100)	3/75 (4)	0/68 (0)
emm-cluster E4* (26)	153/315 (49)	7/21 (33)	0/36 (0)	0/96 (0)
emm-cluster E5 (8)	170/275 (62)	16/41 (39)	0/33 (0)	0/96 (0)
emm-cluster E6 (18)	241/281 (86)	21/30 (70)	1/58 (2)	1/153 (1)
Clade Y (84)	117/641 (18)		7/13 (54)	2/104 (2)
Sub-clade Y1 (33)	109/564 (19)		8/12 (67)	2/97 (2)
emm-cluster D1* (15)	223/442 (50)	27/42 (64)	22/43 (51)	1/138 (1)
emm-cluster A-C1* (7)	303/428 (71)	16/34 (47)	67/131 (51)	1/138 (1)
emm-cluster A-C2* (7)	251/469 (54)	16/43 (37)	8/70 (11)	0/138 (0)
emm-cluster D2 (5)	251/318 (79)	14/43 (33)	9/70 (13)	0/138 (0)
Sub-clade Y2 (51)	195/568 (34)		11/57 (19)	0/138 (0)
emm-cluster D4 (32)	209/335 (62)	5/10 (50)	3/54 (6)	2/145 (1)
emm-cluster D5* (5)	283/413 (69)	20/42 (48)	4/103 (4)	0/138 (0)
emm-cluster A-C3 (5)	246/346 (71)	2/6 (33)	0/88 (0)	0/152 (0)
emm-cluster A-C4 (5)	357/441 (81)	13/24 (54)	10/188 (5)	1/145 (1)
emm-cluster A-C5* (8)	345/447 (77)	4/12 (33)	28/188 (15)	0/145 (0)

Only the codons under positive selection that demonstrate an omega value higher than 2.4 (95% significant) are included in this table. * indicates that the clade analyzed encompasses, but is not restricted, to the sole *emm*-cluster mentioned (as only monophyletic clades can be included selective pressure analysis).

Supplementary Table S3: Binding raw data

GAS strain	*emm* type	*emm* pattern	Major Clade	*emm*-cluster	Plg. (KD)	IgA (KD)	IgG (KD)	Fg. (KD)	Alb. (KD)	C4BP
PRS20	60	E	X	E1	NB	0.84 ±0.04nM	20.07 ±9.70nM	NB	6.62 ±0.14nM	7.09 ±3.75pM
NS226	4	E	X	E1	NB	5.36 ±0.20nM	12.95 ±0.45nM	NB	4.18 ±0.12nM	9.81 ±4.76pM
NS730	90	E	X	E2	NB	NB	10.96 ±1.29nM	NB	7.54 ±0.45nM	NB
NS192	106	E	X	E2	NB	NB	18.54 ±0.53nM	NB	9.33 ±0.85nM	NB
PRS18	58	E	X	E3	NB	NB	5.75 ±0.22nM	NB	8.76 ±0.59nM	5.93 ±2.40pM
PRS55	9	E	X	E3	NB	NB	4.17 ±0.66nM	NB	5.93 ±0.21nM	4.70 ±1.59pM
NS179	9	E	X	E3	NB	NB	6.45 ±0.95nM	NB	6.45 ±0.25nM	5.18 ±1.10pM
PRS66	102	E	X	E4	NB	NB	82.69 ±13.87nM	NB	NB	NB
PRS2	2	E	X	E4	NB	NB	29.76 ±7.33nM	NB	NB	45.42 ±8.62pM
NS8	85	D	X	E6	NB	1.73 ±0.78nM	2.06 ±0.11nM	NB	3.89 ±0.16nM	6.99 ±1.22pM
NS414	11	E	X	E6	NB	1.77 ±0.09nM	10.44 ±0.78nM	NB	18.14 ±0.51nM	119.93 ±23.13pM
NS931	65	D	X	E6	NB	NB	4.51 ±0.18nM	NB	6.38 ±0.01nM	5.10 ±1.28pM
PRS15	48	E	X	E6	NB	0.66 ±0.02nM	3.11 ±0.08nM	NB	11.67 ±0.22nM	7.21 ±1.70pM
NS1140	57	A-C	Y1	M57	NB	NB	7.18 ±0.17nM	0.10 ±0.01nM	4.68 ±0.23nM	NB
NS178	54	A-C and D	Y1	D1	NB	NB	NB	0.11 ±0.02nM	2.24 ±0.16nM	NB
TVU5	54	A-C and D	Y1	D1	NB	NB	NB	0.09 ±0.01nM	4.74 ±0.14nM	NB
PRS9	19	A-C	Y1	M19	NB	NB	NB	0.64 ±0.04nM	3.02 ±0.15nM	NB
NS501	14	A-C	Y1	M14	NB	NB	18.64 ±0.69nM	0.45 ±0.07nM	2.17 ±0.03nM	NB
NS80	70	D	Y2	D4	3.06 ±0.37nM	NB	NB	NB	4.43 ±1.08nM	NB
PRS30	83	D	Y2	D4	1.66 ±0.31nM	NB	NB	NB	5.94 ±0.15 nM	NB
NS13	53	D	Y2	D4	2.19 ±0.73nM	NB	NB	NB	NB	NB
NS88.2	98	D	Y2	D4	1.33 ±0.32nM	NB	NB	NB	NB	NB
88/30	97	D	Y2	D5	NB	NB	NB	0.45 ±0.06nM	4.82 ±0.06nM	NB
NS696	1	A-C	Y2	AC3	NB	NB	5.53 ±0.13nM	0.15 ±0.03nM	4.36 ±0.43nM	NB
PRS8	12	A-C	Y2	AC4	NB	NB	NB	0.20 ±0.01nM	2.77 ±0.09nM	NB
M3	3	A-C	Y2	AC5	NB	NB	NB	0.20 ±0.08nM	2.86 ±0.05nM	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).

References

1. McMillan DJ, Dreze PA, Vu T, et al. 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 2013;19:E222-9
2. Martin D, Rybicki E. RDP: detection of recombination amongst aligned sequences. Bioinformatics 2000;16:562-3
3. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC evolutionary biology 2010;10:210
4. Suyama M, Torrents D and Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic acids research 2006;34:W609-12
5. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution 2007;24:1586-91
6. Sanderson-Smith ML, Walker MJ and Ranson M. 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. J Biol Chem 2006;281:25965-71