MAJOR ARTICLE

A Systematic and Functional Classification of *Streptococcus pyogenes* That Serves as a New Tool for Molecular Typing and Vaccine Development

Martina Sanderson-Smith,[1] David M. P. De Oliveira,[1,a] Julien Guglielmini,[2,3,a] David J. McMillan,[4,5] Therese Vu,[4,6] Jessica K. Holien,[7] Anna Henningham,[8] Andrew C. Steer,[9,10,11] Debra E. Bessen,[12] James B. Dale,[13,14,15] Nigel Curtis,[9,16,17] Bernard W. Beall,[18] Mark J. Walker,[8] Michael W. Parker,[7,19] Jonathan R. Carapetis,[20] Laurence Van Melderen,[6] Kadaba S. Sriprakash,[4] and Pierre R. Smeesters,[6,9] The M Protein Study Group

[1]Illawarra Health and Medical Research Institute and School of Biological Sciences, University of Wollongong, Australia; [2]Microbial Evolutionary Genomics, Département Génomes et Génétique, Institut Pasteur, and [3]CNRS, UMR3525, Paris, France; [4]Bacterial Pathogenesis Laboratory, QIMR Berghofer Medical Research Institute, Brisbane, and [5]Inflammation and Healing Research Cluster, School of Health and Sports Sciences, University of the Sunshine Coast, Sippy Downs, Australia; [6]Laboratoire de Génétique et Physiologie Bactérienne, Institut de Biologie et de Médecine Moléculaires, Faculté des Sciences, Université Libre de Bruxelles, Gosselies, Belgium; [7]Biota Structural Biology Laboratory, ACRF Rational Drug Discovery Centre, St. Vincent's Institute of Medical Research, Melbourne, [8]School of Chemistry and Molecular Biosciences and Australian Infectious Diseases Research Centre, University of Queensland, Brisbane, [9]Murdoch Children Research Institute, [10]Centre for International Child Health, The University of Melbourne, and [11]Department of General Medicine, Royal Children's Hospital Melbourne, Australia; [12]Department of Microbiology and Immunology, New York Medical College, Valhalla; [13]Department of Medicine, The University of Tennessee Health Science Center, [14]Department of Veterans Affairs Medical Center, and [15]Department of Microbiology, Immunology and Biochemistry, The University of Tennessee Health Science Center, Memphis; [16]Infectious Diseases Unit, Royal Children's Hospital Melbourne, and [17]Department of Paediatrics, The University of Melbourne, Australia; [18]Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, Georgia; [19]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, and [20]Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, Perth

Streptococcus pyogenes **ranks among 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 *S. 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 *S. 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.

Received 14 January 2014; accepted 25 April 2014; electronically published 5 May 2014.

[a]D. M. P. D. O. and J. G. Contributed equally to this work.

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

Correspondence: Pierre Smeesters, MD, PhD, Laboratoire de Génétique et Physiologie Bactérienne, IBMM, Université Libre de Bruxelles, 12 rue des professeurs Jeener et Brachet, 6041 Gosselies, Belgium (psmeeste@ulb.ac.be).

The Journal of Infectious Diseases 2014;210:1325–38

DOI: 10.1093/infdis/jiu260

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 prevaccine 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]. Although 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

Polymerase chain reaction (PCR) amplification and sequencing of *emm* genes was performed as described elsewhere [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 LP × TG motif [9, 15]. The absence of significant recombination events in this data set 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 4 bioinformatic criteria: (1) monophyletic or paraphyletic nature, (2) supported by an approximate likelihood-ratio test (aLRT) >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 described elsewhere [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, US), human fibrinogen and albumin (Sigma-Aldrich, Sydney, Australia), immunoglobulin G (IgG; Life Technologies, Melbourne, Australia), immunoglobulin A (IgA; Abcam, Sydney, Australia), and C4BP (Athens Research and Technology, Athens, US) 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

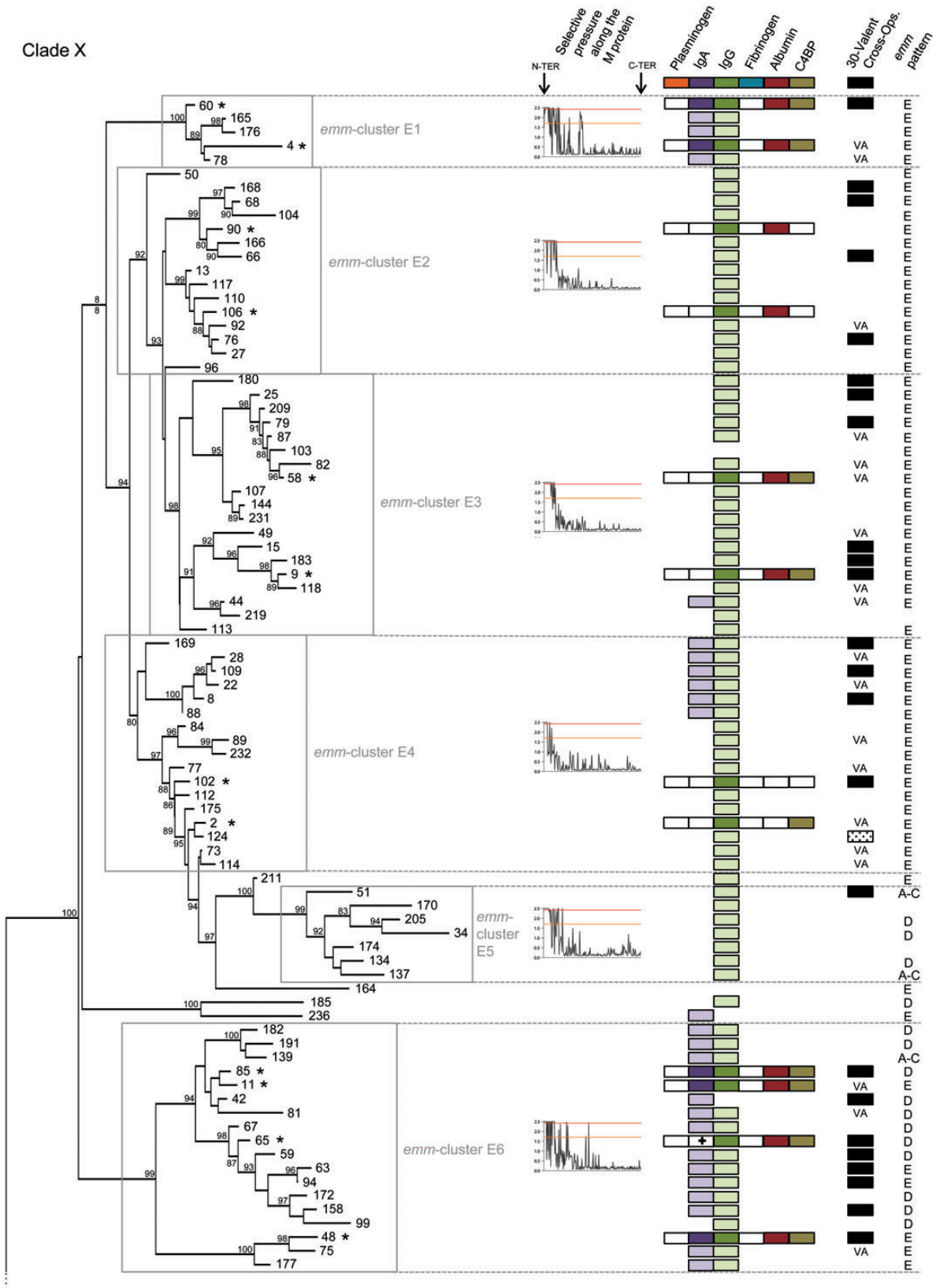

Figure 1. Phylogeny of M proteins and the *emm*-cluster classification system. Phylogenetic inferences of M protein sequences from 175 *emm*-types drawn by PhyML. The tree is drawn to scale, with branch lengths in the same units (number of amino acid substitutions per site) as those of the evolutionary distances used for the phylogenetic tree. Approximate likelihood-ratio test values >80% are indicated at the nodes. The tree has 2 main clades: Clade X is composed of 6 main *emm*-clusters (E1–E6), whereas clade Y is divided into 2 subclades (Y1 and Y2) that are then subdivided 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

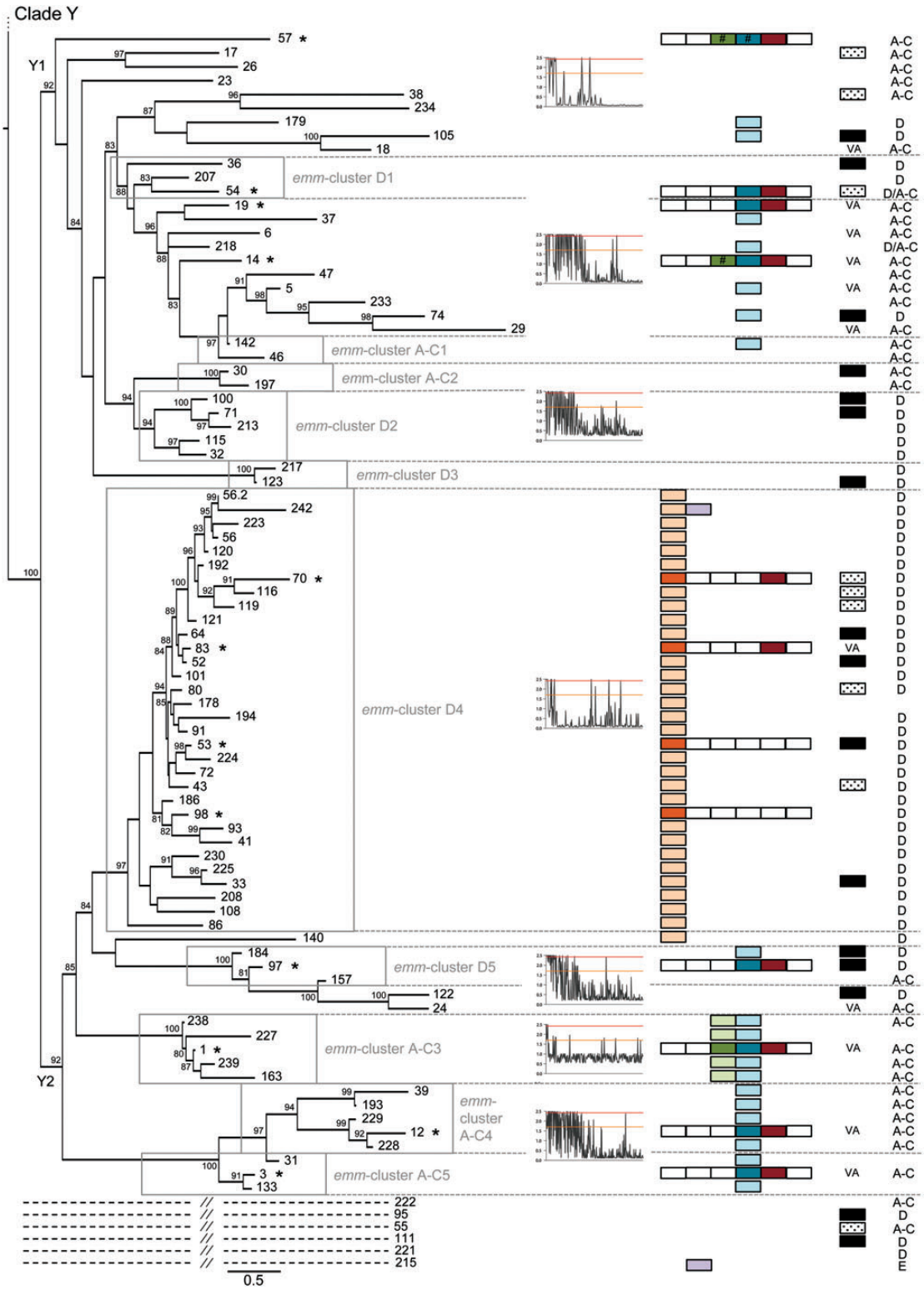

Figure 1 continued. sequences are shown for the different *emm*-clusters and/or clades of the tree. The sites above the red and orange lines are positively selected (probability >0.95 and 0.5, respectively). M protein binding data to 6 human proteins are shown: dark-shaded color boxes indicate experimentally 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-opsonization elicited by the 30-valent vaccine [39, 40]: VA stands for vaccine antigen, black boxes indicate the presence of cross- opsonizing antibodies in rabbit, and shaded boxes indicate a lack of cross-opsonization. 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*. Abbreviations: IgA, immunoglobulin A; IgG, immunoglobulin G.

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 clade X
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).

[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 (Supplementary Table 1). Apart from 6 outlier proteins, 2 well-supported clades (Figure 1; X and Y; 85 and 84 proteins, respectively) were defined based on the general organization of the tree (Figure 1). Clade Y was divided into 2 major subclades (Y1 and Y2). Clade X, subclades Y1 and Y2 were further subdivided into 48 *emm*-clusters. Thirty-two *emm*-clusters contained a single M protein (Figure 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 6 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 3 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 (Figure 1 and Supplementary Table 2). 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 (Figure 1 and Supplementary Table 2). Only some *emm*-clusters

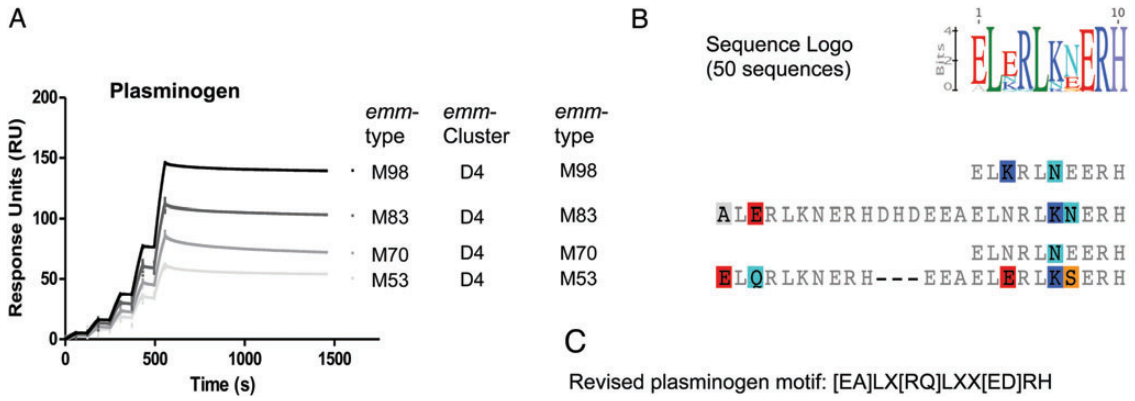

Figure 2. Binding of plasminogen by M proteins. Single cycle kinetic SPR sensorgrams for the interaction of M proteins with plasminogen are shown (*A*). Human glu-plasminogen was injected over immobilized M protein (concentrations of 7.5, 15, 30, 60, and 120 nM). Binding data were calculated by nonlinear fitting of the single cycle kinetic sensograms according to a 1:1 Langmuir binding model using Biacore T200 evaluation software (Biacore AB). Only the 4 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 data set, a refined motif for M protein plasminogen-binding was defined (*C*). The search for this motif among 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. In sum, 17 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*). Abbreviation: SPR, surface plasmon resonance.

had codons under diversifying selection within the carboxy-terminal region. Finally, 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), whereas 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 (Supplementary Table 3) [4]. M proteins belonging to clades X vs 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 specialized 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 (Figures 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 to 5.36 nM (Supplementary Table 3). Of the 4 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 (Figure 3*C*). The refined IgA motif was present in *emm*-cluster E1 and E6 M proteins, and in sub-*emm*-cluster E4.1 (Supplementary Figure 1) and 4 M protein types outside these *emm*-clusters (Figure 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 (Figures 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 (Figure 3*F*) and is present in most M proteins from clade X and *emm*-cluster A-C3 (Figure 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 (subclade 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 subclade Y1 (M57, M54, M19, M14). Fibrinogen binding to M5 has been localized to the B repeat domain [31]. For M1, fibrinogen binding was suggested

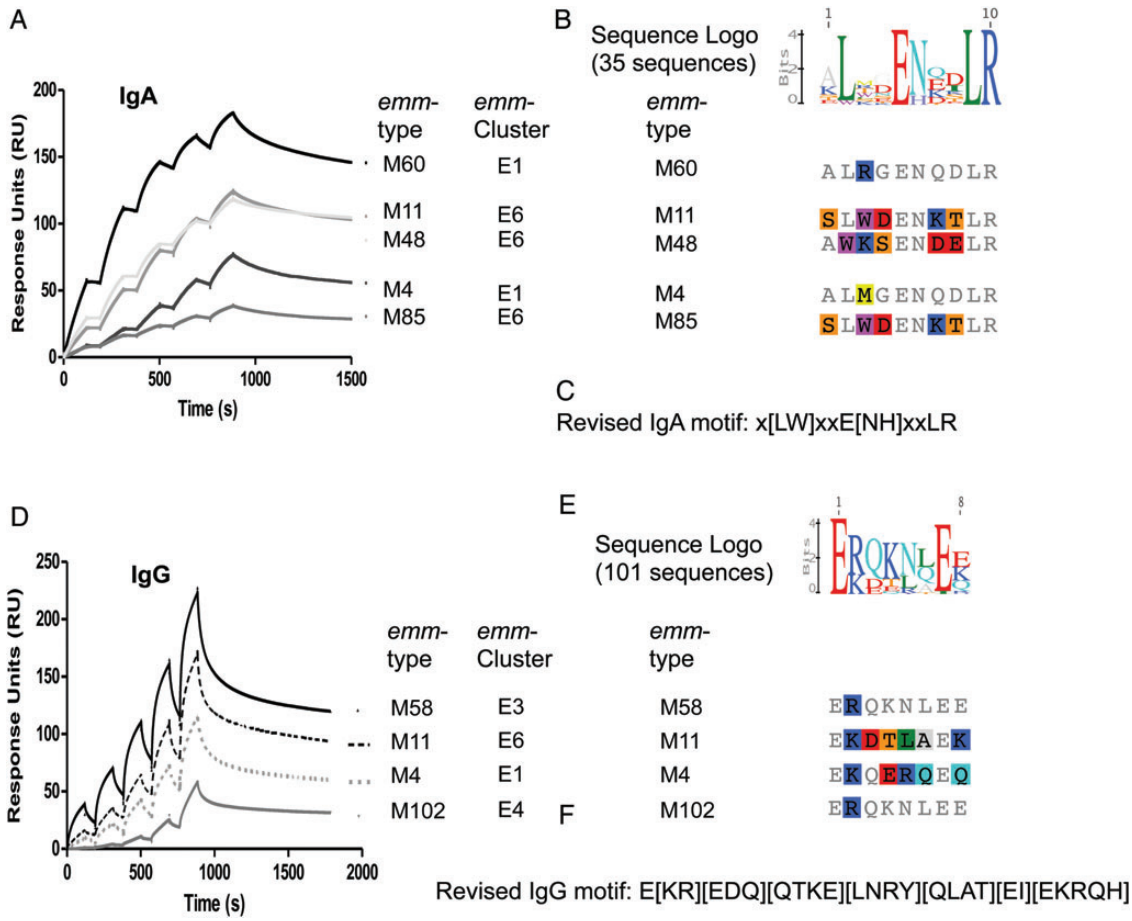

Figure 3. Binding of IgA and IgG by M proteins. In sum, 5 of 6 proteins from *emm*-clusters E1 and E6 bound IgA (*A*). Based on the protein sequence alignment of the 5 IgA-binders (*B*) and the data available in the literature [27], a refined motif for binding of IgA by M protein is defined (*C*). Motif searching gave positive results for 28 *emm*-types in three main (sub-)*emm*-clusters (E1, E6, and E4.1). M proteins of 4 other *emm*-types were positive for this motif: M236 (close to E6), M44 (E3), M242 (D4), and M215 (Outlier Figure 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*). Abbreviations: IgA, immunoglobulin A; IgG, immunoglobulin G.

to be dependent on irregularities within the coil-coil structure of the B repeats, specifically as a result of alanine and other destabilizing 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 suggest a more refined fibrinogen-binding motif can be described (Figure 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 (subclade 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 to 119.93pM) by M proteins associated with *emm*-clusters belonging to clade X, whereas no binding could be demonstrated in clade Y (Supplementary Table 3 and Figure 1). In *emm*-cluster E4, however we observed that M2 bound C4BP while M102 did not. Binding of

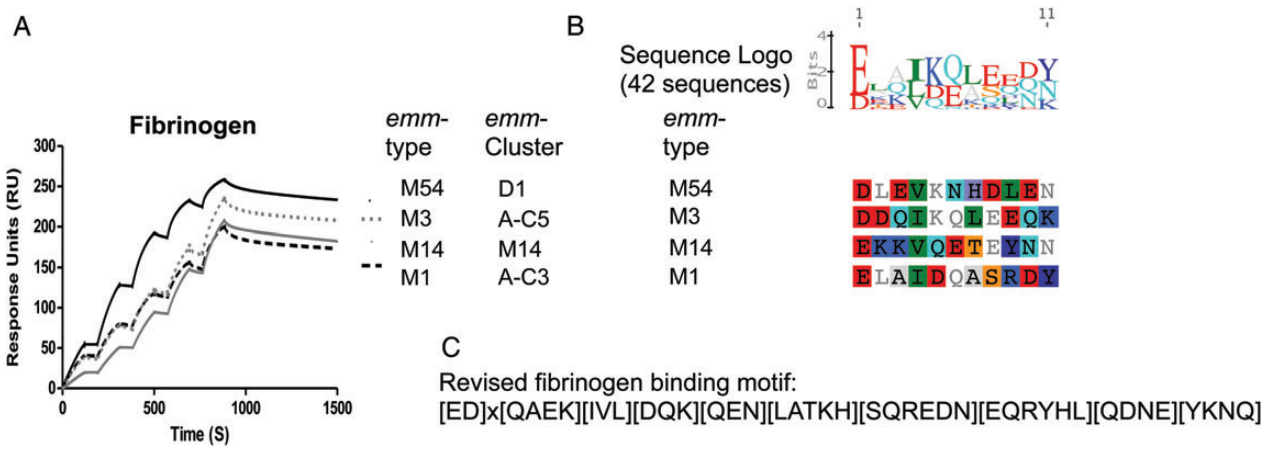

Figure 4. Binding of fibrinogen by M proteins. Eight recombinant M proteins from clade Y bound fibrinogen (Figure 1) and representative single cycle kinetic SPR sensorgrams are shown for 4 *emm*-types (*A*). Based on the fibrinogen-binding motif sequence previously described for M5 [31] and the alignment 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*). Abbreviation: SPR, surface plasmon resonance.

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; Figure 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- opsonization and killing was demonstrated for 39 of 49 of the *emm*-types tested [39, 40] (Figure 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 (Figure 5). Most of the M proteins (27/39) demonstrating cross protection in rabbits belong to *emm*-clusters that possess at least 1 representative included in the vaccine (Figure 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 among 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 (Figure 6). The selective pressure analysis implicated some C-repeat region residues (clade Y, *emm*-cluster E6 of clade X) as being under diversifying selection (Figure 1, Supplementary Table 2 and data not shown). This result was

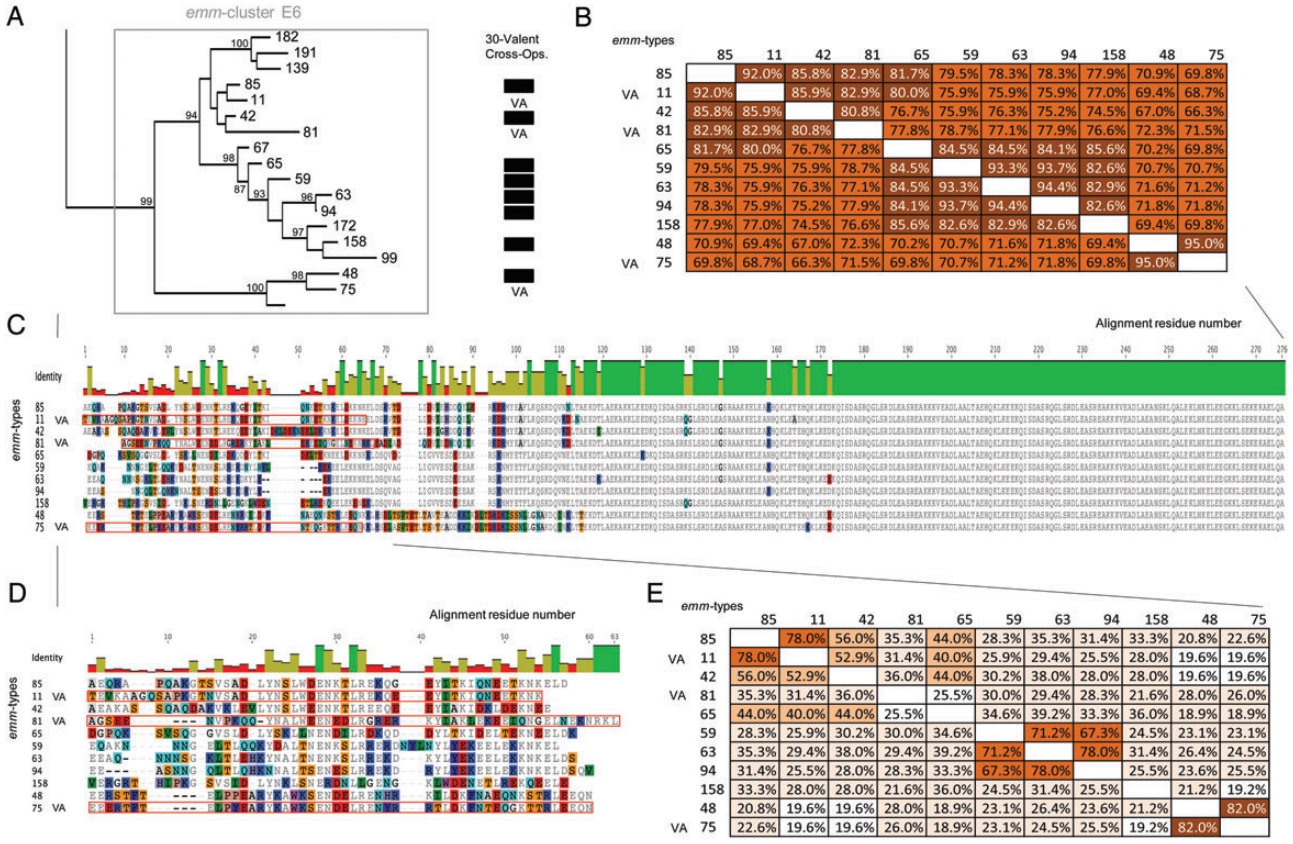

Figure 5. Correlation between immunological cross-protection and M protein sequence *emm*-clusters. M proteins sharing the same *emm*-cluster have different amino-terminal regions but possess nearly identical sequences for the rest of the protein (Figure 1); *emm*-cluster E6 is shown as an example (*A*). 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 average pairwise identity values of the whole M protein sequences within an *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 gray. 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*).

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). The *emm* pattern-typing distinguishes 3 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*-types 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 class I/*sof*$^-$ M proteins, whereas pattern E correspond to the class II/*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 3 different portions of the tree. The first pattern D group is the highly specialized 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 subclade Y1 and Y2. A phylogenetic analysis of the 67 pattern D proteins confirmed this differentiation into 3 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 (Figure 6). Thus, pattern D M proteins form 3 discrete structural groups,

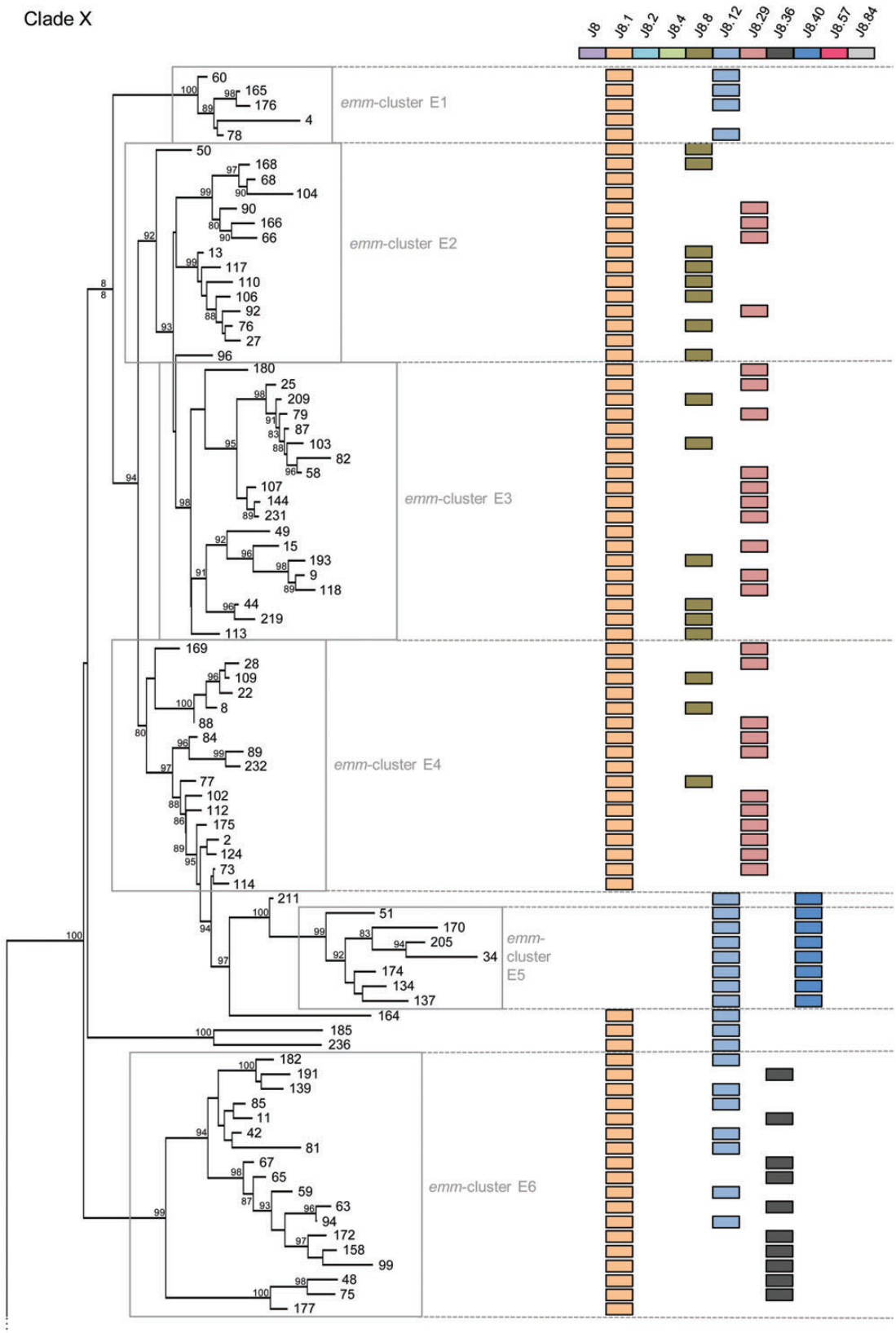

Figure 6. The *emm*-cluster typing system predicts the presence of J8 alleles. The presence of 11 alleles of the J8 vaccine antigen is presented for each *emm*-type. In total, 22 different alleles of the J8 vaccine antigen were found in our data set. The 11 alleles present in at least 5 *emm*-types were represented in this figure. A correlation between clades, subclades, 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

Figure 6 continued. *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, whereas 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.

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, characterization 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

To our knowledge, 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). Although 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 nonvaccine *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 hypotheses 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 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 data set. However, 3 limitations should be acknowledged: rare outliers were observed; some characteristics, such as fibrinogen-binding capacity, seem to be linked to a higher phylogenetic hierarchy (subclades) 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 constitutes a new complementary tool that adds meaningful information and may be widely used to analyze GAS molecular epidemiology. Future experiments aimed at characterizing 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, Georgia.

Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online (http://jid.oxfordjournals.org/). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Acknowledgments. The contributing members of the M protein study group (in addition to the authors of this article) 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.

Financial support. 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. M. S.-S. is an NHMRC Career Development Fellow. M. W. P. is an NHMRC Senior Principal Research Fellow. J. K. H. 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.

Potential 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. All other authors report no potential conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Carapetis JR, Steer AC, Mulholland EK, Weber M. The global burden of group A streptococcal diseases. Lancet Infect Dis **2005**; 5:685–94.
2. Steer AC, Lamagni T, Curtis N, 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, Sriprakash KS. The streptococcal M protein: a highly versatile molecule. Trends Microbiol **2010**; 18:275–82.
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, 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, 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. Clin Microbiol Infect **2013**; 19:E222–9.
10. Denny FW Jr, Perry WD, Wannamaker LW. Type-specific streptococcal antibody. J Clin Invest **1957**; 36:1092–100.
11. Lancefield RC. Persistence of type-specific antibodies in man following infection with group A streptococci. J Exp Med **1959**; 110:271–92.
12. Smeesters PR, McMillan DJ, Sriprakash KS, 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, 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, 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, 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, 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? In: Proceedings of the Belgian Royal Academies of Medicine **2014**; 3:89–98.
18. Wannamaker LW, Denny FW, Perry WD, Siegel AC, Rammelkamp CH Jr. Studies on immunity to streptococcal infections in man. Am J Dis Child **1953**; 86:347–8.
19. Watson RF, Rothbard S, Swift HF. Type-specific protection and immunity following intranasal inoculation of monkeys with group A hemolytic streptococci. J Exp Med **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, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol **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 Evol Biol **2010**; 10:210.
23. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **2010**; 59:307–21.
24. Sanderson-Smith ML, Walker MJ, 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, 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, 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, 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, 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, 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, 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, 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, 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, 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, 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, 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**; 31:1576–81.
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, 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. J Immunol **2013**; 190:2692–701.
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, 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, 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.