Factor VIII cross-matches to the human proteome reduce the predicted inhibitor risk in missense mutation hemophilia A

by Daniel P. Hart, Nazmiye Uzun, Stuart Skelton, Alison Kakoschke, Jacob Househam, David S. Moss, and Adrian J. Shepherd

Haematologica 2018 [Epub ahead of print]

Citation: Daniel P. Hart, Nazmiye Uzun, Stuart Skelton, Alison Kakoschke, Jacob Househam, David S. Moss, and Adrian J. Shepherd. Factor VIII cross-matches to the human proteome reduce the predicted inhibitor risk in missense mutation hemophilia A. Haematologica. 2018; 103:xxx
doi:10.3324/haematol.2018.195669

Publisher's Disclaimer.
E-publishing ahead of print is increasingly important for the rapid dissemination of science. Haematologica is, therefore, E-publishing PDF files of an early version of manuscripts that have completed a regular peer review and have been accepted for publication. E-publishing of this PDF file has been approved by the authors. After having E-published Ahead of Print, manuscripts will then undergo technical and English editing, typesetting, proof correction and be presented for the authors' final approval; the final version of the manuscript will then appear in print on a regular issue of the journal. All legal disclaimers that apply to the journal also pertain to this production process.
Factor VIII cross-matches to the human proteome reduce the predicted inhibitor risk in missense mutation hemophilia A
Abstract

Single missense mutations in the \( F8 \) gene encoding the coagulation protein factor VIII (FVIII) give rise predominantly to non-severe hemophilia A. Despite only a single amino acid sequence difference between the replacement, therapeutic FVIII (tFVIII) and the patient’s endogenous FVIII, tFVIII may still be perceived as foreign by the recipient’s immune system and trigger an immune response (inhibitor). Inhibitor formation is a lifelong risk for non-severe hemophilia A patients treated with tFVIII, but remains difficult to predict. The aim of this study was to understand whether fortuitous, primary sequence cross-matches between tFVIII and proteins in the human proteome are the reason why certain \( F8 \) mutations are not associated with inhibitor formation. We predicted which tFVIII differences are potentially perceived as foreign by helper T cells – a necessary precursor to inhibitor development – and then scanned potentially immunogenic peptides against more than 100,000 proteins in the proteome. As there are hundreds of disease-causing \( F8 \) missense mutations and the Human Leucocyte Antigen gene complex governing peptide presentation to helper T cells is highly polymorphic, these calculations pose a huge combinatorial challenge that we addressed computationally. We identify that cross-matches between tFVIII and the human proteome are commonplace and have a profound impact on the predicted risk of inhibitor development. Our results emphasize the importance of knowing both the \( F8 \) missense mutation and the Human Leucocyte Antigen alleles of a patient with missense mutation hemophilia A if his underlying risk of inhibitor development is to be estimated.
Introduction

All severities of hemophilia A (HA) are at risk of an alloimmune response (inhibitor formation) against infused, therapeutic FVIII (tFVIII) concentrate. It is well recognized that the more disruptive the F8 mutation, the more severe the hemophilia and the more likely it is that inhibitors will arise. Consequently, severe HA has been the priority for inhibitor-related research, surveillance and intervention over the past decades. However, it is also clear that only a single amino acid difference between an endogenous F8 genotype and the wild-type tFVIII sequence is sufficient to induce an immune response that results in clinically relevant inhibitors and that this risk is life-long in the context of non-severe HA.

HA caused by a missense mutation is typically associated with a less severe bleeding phenotype than HA caused by incomplete F8 transcripts. In contrast to severe HA, boys and men living with non-severe HA are more likely to remain hospital dependent for on-demand tFVIII administration throughout their lives in the event of injury or surgery. The treatment burden for this group is surprisingly high, with 44% of a large London cohort being reported to have received some hemostatic treatment in a 2-year observation window, 79% of whom received tFVIII concentrate. Consequently, inhibitor surveillance in non-severe HA requires adult treaters to be ever vigilant. In contrast to the systematic inhibitor screening in early tFVIII exposures for severe HA, inhibitor screening in the non-severe HA setting is currently more reactive and sporadic, but recognized to be of increasing importance given the ageing population of those living with non-severe HA. Inhibitor occurrence in non-severe hemophilia A can be devastating, with neutralization of infused FVIII concentrate and potential cross-reactivity with endogenous FVIII. This cross-reactivity occurs in at least 50% of identified cases and results in loss of a patient’s previous non-severe FVIII activity baseline level (FVIII:C) resulting in a worsening bleeding phenotype, often in latter decades of life. This results in increased bleed rates and an increased risk of premature mortality. In this context, the early detection of inhibitor occurrence – or, better still, the ability to reliably predict an individual’s risk of developing inhibitors before any have formed – has the potential to influence subsequent clinical decisions in ways that substantially improve patient outcomes.

The T-cell dependency of inhibitor generation is well described, with confirmed tFVIII-specific CD4+ T-cell responses and immunoglobulin class switching. CD4+ T-cell activation depends on their interaction with “foreign” peptides – in this case, tFVIII-derived peptides spanning the location of the endogenous F8 missense mutation – presented by Major Histocompatibility Complex (MHC) class II molecules. However, not all such foreign peptides are perceived to be immunologically different from self and, if the difference is undetected, there is presumed negligible risk of an immune response. There are two key mechanisms at work here. Firstly, not all peptides are capable of binding to an individual’s repertoire of MHC molecules and are therefore never presented to T cells. Secondly, not all binding peptides are distinguishable from self-peptides bound to the same MHC molecules; in such cases, T cells that are capable of binding to these
MHC:peptide complexes are expected to have been removed from the T-cell repertoire by self-tolerance mechanisms.\textsuperscript{17}

What is unclear, however, is whether an understanding of MHC presentation and self-tolerance can enable us to make useful predictions about the inhibitor risk of patients with missense mutation HA – for example, by accurately predicting whether individual patients have a negligible risk of developing inhibitors. The aim of this study is to directly address that point. Given that MHC molecules are encoded by genes that are among the most polymorphic in the human genome, and there are several hundred disease-causing $F^8$ missense mutations, the first aim of this study was to predict inhibitor risk based on an analysis of tFVIII peptide presentation by MHC molecules. Such an investigation poses a huge combinatorial challenge – one that is arguably impractical to address using purely \textit{in vitro} techniques. Building on the approach developed in an earlier study that we undertook using a much smaller dataset,\textsuperscript{18} we analyzed MHC:peptide complexes associated with 25 common HLA class II alleles and 956 distinct $F^8$ missense mutations, requiring over 4 million peptide-HLA isoform combinations to be evaluated.

However, this preliminary analysis did not take into account the possibility that fortuitous cross-matches between tFVIII-derived peptides and peptides at other locations – both within the FVIII protein sequence itself, and more generally to other proteins in the human proteome – may play a protective role by ensuring that T cells capable of triggering an immune response have been removed from the repertoire by self-tolerance mechanisms. This includes, but is far from limited to, the well-described homology between FVIII and factor V.\textsuperscript{19} Such human proteome cross-matching is the main focus of this study. Here we demonstrate that cross-matches between tFVIII and other parts of the proteome are commonplace and have a profound impact on the predicted inhibitor risk for individuals living with non-severe hemophilia A.

\textbf{Methods}

\textit{Novel peptide-MHC surfaces}

In previous work, we developed a methodology for predicting which patients with non-severe HA are at risk of developing antibodies against tFVIII.\textsuperscript{18} Specifically, we predicted which $F^8$ missense mutation/HLA isoform combinations would present novel peptide-MHC (pMHC) surfaces to CD4\textsuperscript{+} T cells, taking into account the reasonable assumption that T cells capable of binding to pMHC surfaces that are formed by endogenous FVIII peptides and presented by the same MHC molecules would have been removed from the T-cell repertoire by central tolerance mechanisms. Novelty arises when a tFVIII-derived peptide is an MHC-binder (most are not) \textit{and} either i) the equivalent endogenous FVIII-derived peptide is a non-binder (this may occur if the missense mutation is at an MHC-facing, peptide-anchoring position, as residues at such positions anchor the peptide to the MHC molecule) or ii) the relevant amino-acid difference is at a TCR-facing position (Figure 1A).
Our earlier work focused exclusively on the location of the hemophilia-causing \textit{F8} missense mutation and HLA-DR presentation. Here we extend our analysis to include HLA-DP and -DQ presentation and, crucially, to take into account the possibility of fortuitous peptide cross-matches to other locations – both within \textit{FVIII} itself, and more generally within other proteins in the human proteome (Figure 1B).

\textit{Peptide-MHC binding prediction}

We used the \textit{in silico} tool NetMHCII\textsuperscript{20} to predict the strength with which a given peptide binds to an MHC molecule – specifically, the portable version of NetMHCII 2.2 (Technical University of Denmark, http://www.cbs.dtu.dk/services/NetMHCII/). If a peptide is predicted to bind with a 50\% inhibitory concentration (IC\textsubscript{50}) of <1000 nmol/L – the conventional threshold used to indicate that peptide-MHC class II binding is of biological significance\textsuperscript{21} – it is a candidate for forming a novel pMHC surface. Deciding whether a given combination of tFVIII peptide and MHC molecule forms a novel pMHC surface in comparison to the corresponding endogenous \textit{FVIII}-derived peptide requires the considerations outlined in Figure 1 to be made, given: a) knowledge of the positions of the MHC anchoring pockets for the chosen HLA isoform (generally these are at positions 1, 4, 6 and 9); and b) the binding register of the peptide (predicted by NetMHCII), which specifies the stretch of 9 consecutive amino-acid residues that form the preferred binding core within the MHC groove. Predictions were made for 25 common HLA-DR, -DP and -DQ isoforms with estimated worldwide population coverages of >70\%, >90\% and >80\% respectively\textsuperscript{22} using UniProt\textsuperscript{23} \textit{FVIII} sequence P00451.

For the full set of 25 HLA class II alleles considered for this research, over 4 million peptide-HLA isoforms combinations were evaluated in order to identify the combinations that are predicted to form a novel pMHC surface – plus a similar order of additional evaluations necessary to identify potential cross-matches to the human proteome. A detailed breakdown of the calculations performed is given in \textit{Online Supplementary Table S1}.

\textit{Scanning the human proteome for cross-matches}

Our preliminary risk assessment focused exclusively on the location of the disease-causing \textit{F8} missense mutation – that is, we assessed whether one or more peptides would be perceived as foreign, given the binding properties and side-chain orientations of both tFVIII and endogenous \textit{FVIII} peptides spanning the location of the missense mutation (Figure 1A). The innovative hypothesis explored here is that some pMHC surfaces that were identified as risk-associated by this preliminary approach may not be novel within the broader context of the whole human proteome.

For this research, we compared pMHC surfaces formed by tFVIII peptides spanning the location of the \textit{F8} missense mutation with those formed by peptides from the human proteome. Following research showing that the “maximal representation of the ‘immunological self’”\textsuperscript{17} is made available for tolerance induction in the thymus, we used
the complete human proteome from Ensembl containing over 100,000 proteins, including alternative isoforms that have an associated protein product.

As previously, we confined our analysis to 15-mers, this being the most common peptide length chosen for MHC class II binding experiments. The canonical proteome was subdivided into all possible 9-mers. The resultant dataset consists of nearly 38 million 9-mers of which more than 11 million are non-identical.

A summary of the computational pipeline used to identify novel pMHC surfaces is shown in Figure 2.

**Statistical analysis**

We evaluated the accuracy of our method using $F8$ mutation data downloaded from the large source of $F8$ mutation data in the public domain (The European Association for Haemophilia and Allied Disorders. The Factor VIII Gene ($F8$) Variant Database. http://www.factorviii-db.org. Accessed November 26 2016). The dataset contained 956 distinct $F8$ missense mutations at 605 different loci from 3,243 individuals. Ninety of the missense mutations were associated with at least one reported case of inhibitor formation, with a total of 160 individuals (prevalence 4.9%) listed as having inhibitors.

We tested the null hypothesis that our predicted rate and the database reported rate of inhibitor formation are independent. In the absence of patient HLA-typing information, we predicted whether a patient with a given missense mutation has a risk of inhibitor formation based on the combined predictions for our chosen set of 25 common HLA class II isoforms. We evaluated different IC$_{50}$ binding thresholds for tFVIII peptides, on the grounds that binding strength is likely to be an important factor in inhibitor risk. However, when considering peptide-MHC binding in the context of self-tolerance prediction (e.g. the binding potential of proteome cross-matches), we maintained a binding threshold of 1000 nM. For a given tFVIII-peptide binding threshold, a patient’s predicted risk of inhibitor formation was deemed to fall within one of the following three categories: “predicted low/negligible risk”, “predicted at risk”, or “unknown risk”. A patient was deemed to be at low/negligible risk if no novel pMHC surfaces were predicted to be formed with any of the 25 HLA isoforms. A patient was deemed to be “at risk” if novel pMHC surfaces were predicted to be formed by all, or all but one, of the HLA isoforms encoded by at least one HLA gene complex – HLA-DRB1, HLA-DRB3/4/5, HLA-DP, and/or HLA-DQ isoforms. (Here the “all, or all but one” condition was designed to rule out the possibility that the patient was a heterozygous individual with a risk-free combination of common HLA isoforms, i.e. two or more low/negligible risk isoforms per gene complex.) The inhibitor risk for patients in neither of the preceding categories was considered “unknown”. Patients with unknown inhibitor risk were omitted from the statistical test.

The statistical test was undertaken using the two-tailed Fisher’s exact test implemented in the R statistical programming language. The Fisher’s exact test is necessary because the sample size and background inhibitor rate are both relatively low; in such circumstances,
the calculation of exact $P$ values is important. Following standard convention, a $P$ value of <0.05 was used to define statistical significance.

**Results**

*An example of proteome cross-matching*

Arg593Cys (R593C) (R612C using Human Genome Variation Society numbering) is a relatively common $F_8$ missense mutation that has been identified as being associated with an “increased” risk of inhibitor formation – for example, 12/106 (11.3%) of R593C individuals within the INSIGHT cohort have been reported as having an inhibitor.\(^8\) Taking the common HLA-DR allele HLA-DRB1*01:01 as an example, our analysis of predicted inhibitor risk proceeded as follows.

We began by using NetMHCII\(^20\) to predict whether any of the tFVIII 15-mers spanning position 593 are binders. Several such 15-mers were predicted to bind to the MHC molecule associated with HLA-DRB1*01:01, and some had binding cores that span position R593. There were two such cores – IQRFLPNPA and YLTENIQRF – both of which were associated with multiple binding peptides, as shown in Figure 3A.

The next step assessed whether either of these cores was associated with predicted pMHC surface novelty compared to their respective endogenous counterparts. Both of the endogenous cores – IQCFLPNPA and YLTENIQCF respectively (with a Cys (C) in place of the Arg (R) of the equivalent tFVIII peptides) – were associated with predicted binding 15-mers. Hence the question of predicted novelty hinged on the position of R593 within the 9-mer: whether at a position involved in MHC-binding (and hence invisible to a TCR), or at a TCR-facing position? As the binding pockets in the MHC groove for HLA-DRB1*01:01 are at positions 1, 4, 6 and 9, both IQRFLPNPA (with R593 at TCR-facing position 3) and YLTENIQRF (with R593 at TCR-facing position 8) are associated with the formation of pMHC surfaces that are novel in comparison to those formed by endogenous FVIII, as shown in Figure 3B.

Based on the preceding analysis alone, we would predict that a patient with the R593C mutation and HLA-DR allele HLA-DRB1*01:01 would be at risk of developing inhibitors. However, a tFVIII-derived pMHC surface that is novel with respect to an individual’s endogenous FVIII may not be novel in the wider context of his proteome. To evaluate this possibility of a proteome cross-match that reduces the risk of inhibitor development, we searched for a pattern for each of the 9-mer cores matching at T-cell facing positions 2, 3, 5, 7 and 8. For IQRFLPNPA and YLTENIQRF these patterns are XQRXLXNPX and XLTXNXQRX respectively, where the letter X matches any amino-acid type. These patterns were scanned against a library containing all 11,272,502 unique 9-mers from the human proteome. In this case, the pattern XQRXLXNPX matched the 9-mer FQRELNNPL in human tubulin polyglutamylase (UniProt\(^25\) Q6ZT98), and pattern XLTXNXQRX matched both the 9-mer GLTENSQRD in dystrobrevin binding protein 1
HUMAN PROTEOME AND HEMOPHILIA A INHIBITOR RISK

(dysbindin) (UniProt D6RJC6) and the 9-mer ELTKNAQRA in the uncharacterized human protein C2orf48 (UniProt Q96LS8), as shown in Figure 4A.

The final step was to check whether a given cross-matching, proteome-derived 9-mer occurred as a binding core for HLA-DRB1*01:01, as only then would we hypothesize tolerance. In this case, NetMHCII predicted that both FQRELNNPL in tubulin polyglutamylase and ELTKNAQRA in C2orf48 form cores within 15-mers with a predicted IC50 <1000 nmol/L, as shown in Figure 4B. Hence, we ultimately predicted that the F8 missense mutation/HLA allele combination R593C/HLA-DRB1*01:01 confers no, or negligible, risk of inhibitor formation owing to fortuitous cross-matches to peptides in the human proteome.

Proteome cross-matches and inhibitor risk stratification

In terms of individual combinations of F8 missense mutation and HLA-DR/DP/DQ isoforms, the impact of proteome cross-matches on predicted risk is shown in a comprehensive heat map (Online Supplementary Figure S1), with a subset of combinations shown in Figure 5. Each individual F8 missense mutation/HLA isoform combination is shown as a single square. An analysis of the full set of data indicates that the percentage of F8 missense mutation/HLA isoform combinations associated with predicted inhibitor risk falls appreciably when proteome cross-matches are taken into account: from 49% to 31% with a binding threshold of 1000 nM; and from 37% to 21% with a binding threshold of 500 nM.

These predictions strongly suggest that the risk of inhibitor formation is largely HLA-dependent and, in most cases, cannot be reliably predicted from knowledge of the F8 genotype alone. The proportion of reported F8 missense mutations that we predict to be risk-associated varies considerably between different HLA isoforms. For example, with a binding threshold of 1000 nM, it ranges between 21% (HLA-DQA1*03:01-DQB1*03:02 and HLA-DQA1*04:01-DQB1*04:02) to 86% (HLA-DRB1*01:01) without proteome cross-matches, and between 13% (HLA-DQA1*04:01-DQB1*04:02) and 48% (HLA-DRB1*01:01) with proteome cross-matches. Values for all 25 HLA isoforms and multiple binding thresholds are given in Online Supplementary Table S2. The number of F8 missense mutations associated with no, or negligible, predicted risk (black columns in the heatmap) is only 69 out of 956 (7%).

We consider any (human) protein that, in effect, contributes to a reduction in predicted inhibitor risk for one or more F8 missense mutation/HLA isoform combinations to be “protective”. Of the 20,300 proteins in the human proteome, 4,605 are protective. The protein that affords the most protection is factor V, which is recognized as having high homology with FVIII.19 The most protective proteins identified in our analysis are listed in Table 1. It is notable that FVIII and all the top entries in Table 1 (coagulation factor V, hephaestin-like protein 1, ceruloplasmin, and hephaestin) have copper-binding sites.

Evaluation of risk prediction accuracy
Given the paucity of published data specifying the HLA profiles of patients with missense mutation HA, we have based the evaluation of how accurate our approach is at predicting potential risk on data in the Factor VIII Gene (F8) Variant Database, focusing on F8 missense mutations that we predict to have low or negligible risk with any common HLA isoform.

Arguably the most important performance indicator for our method is the number of false negatives: are there individuals with missense mutation hemophilia A and with data in the Factor VIII Gene Variant Database that we predict to have a zero, or negligible, inhibitor risk but have, in fact, developed inhibitors? The results in Table 2 for our chosen set of 25 HLA alleles show that the number of false negatives (column 3) is very low with conservative cut-offs for peptide-MHC binding affinity. The number of patients predicted to have a zero, or negligible risk of inhibitor development is considerably higher when cross-matches to the proteome are taken into account, but the false negative rate remains comparatively low; these factors taken together contribute to the higher prevalence of statistically significant $P$ values with proteome cross-matching.

Discussion

In this paper, we highlight the potential value of an in silico predictive model of HLA class II antigen presentation as the basis for identifying patients with missense mutation hemophilia A who are at risk of developing inhibitors against tFVIII. Our pipeline incorporates a novel strategy that we term proteome scanning – the identification of fortuitous cross-matches between potentially antigenic tFVIII peptides and peptides arising elsewhere within the human proteome. Such cross-matching peptides are the basis for protection against inhibitor development because of presumed T-cell tolerance mechanisms. Here we focused on the surfaces formed by tFVIII peptides that span the locations of known disease-causing F8 missense mutations and are predicted to bind to the MHC molecules for 25 common HLA class II alleles. Given a conservative binding threshold of 1000 nM, the number of $F8$ missense mutation/HLA isoform combinations associated with a risk of developing inhibitors was predicted to fall by more than a third – from 49% to 31% – when cross-matches to the proteome are taken into account. These results were shown to be statistically significant with a dataset of missense mutation HA patients derived from the Factor VIII Gene Variant Database.

Although our proteome scanning approach reduces the number of patients predicted to be at risk of developing inhibitors, that number remains higher than, albeit closer to, the number of patients that have – at least to date – developed inhibitors. This is inevitable for a model based entirely on a consideration of MHC/TCR interactions, as a range of downstream factors may militate against inhibitor development. These include: the absence of sufficient T cells capable of binding to a given pMHC surface for reasons other than self-tolerance; the lack of co-stimulatory signalling; or the level of exposure to tFVIII being below the threshold necessary for inhibitor formation (only cursory
information about a patient’s degree of exposure to tFVIII is available in the Factor VIII Gene Variant Database database).

There are a number of ways in which this analysis could be refined. Firstly, we took no account of potential F8 genotype mismatches between tFVIII products (derived from common F8 genotypes H1 and H2 that differ only in the B domain) and rare genotypes H3-8, such as the M2238V found in approximately 23% of black people. Nor did we consider the antigenic impact of different linkers used in B-domain modified tFVIII products. Secondly, proteome scanning was performed against a single reference proteome. It is likely that additional cross-matches will be found if allelic variants are taken into account, adding further to the potential advantages of personalized risk assessment. Scanning against an individual’s own proteome would be the optimal predictive strategy. The impact of proteome variability will be assessed in future work using data from IGSR: The International Genome Sample Resource.

There are several more challenging issues. Our model of peptide-MHC binding is imperfect, for example: we do not take into account the impact of cathepsin cleavage on the availability of FVIII peptides for MHC class II binding (there are no established computational methods for predicting cleavage by cathepsins, and different sets of cathepsins occur in different professional antigen presenting cells; peptide differences at anchoring positions, or outside the binding core, are known to affect the formation of pMHC surface in specific cases (but the prevalence of such effects is poorly understood); and a given TCR may not be in contact with all TCR-facing residues (but the binding angle and register of individual TCRs is currently unpredictable).

Validating the accuracy of inhibitor risk prediction for patients with non-severe HA is also particularly problematic. In practice, the current, clinical gold standard for inhibitor detection is a functional, clotting-based Bethesda assay. However, heat treatment modifications in the presence of residual FVIII:C (i.e. non-severe HA) are often omitted, resulting in reduced sensitivity of detection. More importantly, a purely “functional” clotting assay does not detect the totality of antibody responses against a protein therapeutic. The absence of a more “neutral” screening assay (e.g. ELISA based) to pick up any anti-tFVIII antibody response first, and subsequently for the functional assay (Bethesda) to determine its inhibitory potential and clinical relevance, compromises our knowledge of the totality of anti-tFVIII responses in our patient cohorts. It is also evident that in contrast to severe HA, screening practice for antibody responses in non-severe HA are often opportunistic and passive, further reducing the likelihood of detecting the totality of anti-tFVIII antibody responses by missing the optimal immunological windows for screening after tFVIII exposure. Given the life-long risk of inhibitor formation in non-severe HA, we have concerns that true negatives (i.e. patients confirmed to have a zero risk of inhibitor development) are impossible to identify in a non-severe HA clinical study, even when factors such as age and exposure are taken into account.

Notwithstanding these limitations, this study provides compelling evidence of the importance of HLA class II genotyping for analyzing the inhibitor risk of patients with missense mutation HA. Moreover, we have demonstrated that an innovative
computational pipeline incorporating proteome scanning predicts that a large proportion of \( F8 \) missense mutation/HLA isoform combinations afford a negligible risk of inhibitor development, with a low error rate when evaluated using the largest available dataset of HA patients with \( F8 \) missense mutations and conservative MHC binding thresholds. This represents an important step forward, as it closes part of the gap between predicted/potential inhibitor risk and observed inhibitor rates. These insights may ultimately contribute to the design of future clinical studies (with HLA typing of missense mutation HA patients) that are of direct translational relevance.

**Acknowledgements**

D.P.H. received funding from the British Society of Haematology.

**References**


### Table 1. Human proteins that afford the greatest proteome cross-match protection

<table>
<thead>
<tr>
<th>UniProt ID</th>
<th>Protein name</th>
<th>Protected peptide count</th>
</tr>
</thead>
<tbody>
<tr>
<td>P12259</td>
<td>Coagulation factor V</td>
<td>640</td>
</tr>
<tr>
<td>Q6MZM0</td>
<td>Hephaestin-like protein 1</td>
<td>457</td>
</tr>
<tr>
<td>P00450</td>
<td>Ceruloplasmin</td>
<td>437</td>
</tr>
<tr>
<td>Q9BQS7</td>
<td>Hephaestin</td>
<td>389</td>
</tr>
<tr>
<td>P00451</td>
<td>Coagulation factor VIII [match to different, but homologous, location within the protein]</td>
<td>251</td>
</tr>
<tr>
<td>O75445</td>
<td>Usherin</td>
<td>150</td>
</tr>
<tr>
<td>Q14585</td>
<td>Zinc finger protein 345</td>
<td>142</td>
</tr>
<tr>
<td>Q14587</td>
<td>Zinc finger protein 268</td>
<td>134</td>
</tr>
<tr>
<td>Q96M86</td>
<td>Dynein heavy chain domain-containing protein</td>
<td>83</td>
</tr>
<tr>
<td>O00154</td>
<td>Cytosolic acyl coenzyme A thioester hydrolase</td>
<td>76</td>
</tr>
<tr>
<td>Q7LBC6</td>
<td>Lysine-specific demethylase 3B</td>
<td>75</td>
</tr>
<tr>
<td>Q9UKF2</td>
<td>Disintegrin and metalloproteinase domain-containing protein 30</td>
<td>75</td>
</tr>
<tr>
<td>Q9Y2P0</td>
<td>Zinc finger protein 835</td>
<td>74</td>
</tr>
<tr>
<td>P10745</td>
<td>Retinol-binding protein 3</td>
<td>73</td>
</tr>
<tr>
<td>Q5T5N4</td>
<td>Uncharacterized protein C6orf118</td>
<td>67</td>
</tr>
</tbody>
</table>

The protected peptide count for a given combination of human protein $p$, $F8$ missense mutation $m$ and HLA isoform $h$ is incremented by 1 every time a peptide in $tFVIII$ that spans the location of $m$ is a) associated with a predicted risk of inhibitor development for $h$ prior to considering cross-matches to the human proteome, and b) cross-matches to a peptide in $p$ that is a predicted binder for $h$. Hence a peptide that cross-matches to multiple binding peptides at different locations within $p$ will be counted multiple times. The final count for $p$ is the aggregate of individual counts for all $F8$ missense mutation/HLA isoform combinations considered in this study.
Table 2. Evaluation of zero/negligible inhibitor risk prediction with and without proteome scanning

<table>
<thead>
<tr>
<th>IC₅₀ binding threshold (nmol/L)</th>
<th>Patients predicted to have zero/negligible risk</th>
<th>Patients predicted to have an inhibitor risk</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No inhibitors</td>
<td>Inhibitors</td>
<td>Inhibitors</td>
</tr>
<tr>
<td>Without proteome scanning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>28</td>
<td>1</td>
<td>116</td>
</tr>
<tr>
<td>500</td>
<td>49</td>
<td>3</td>
<td>92</td>
</tr>
<tr>
<td>300</td>
<td>122</td>
<td>3</td>
<td>84</td>
</tr>
<tr>
<td>200</td>
<td>179</td>
<td>9</td>
<td>76</td>
</tr>
<tr>
<td>100</td>
<td>362</td>
<td>20</td>
<td>37</td>
</tr>
<tr>
<td>50</td>
<td>593</td>
<td>36</td>
<td>31</td>
</tr>
<tr>
<td>With proteome scanning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>103</td>
<td>4</td>
<td>80</td>
</tr>
<tr>
<td>500</td>
<td>157</td>
<td>7</td>
<td>65</td>
</tr>
<tr>
<td>300</td>
<td>322</td>
<td>14</td>
<td>57</td>
</tr>
<tr>
<td>200</td>
<td>465</td>
<td>26</td>
<td>53</td>
</tr>
<tr>
<td>100</td>
<td>777</td>
<td>42</td>
<td>23</td>
</tr>
<tr>
<td>50</td>
<td>1,114</td>
<td>66</td>
<td>22</td>
</tr>
</tbody>
</table>

P values were calculated by applying Fisher’s exact test to patients from the Factor VIII Gene Variant Database falling into the following categories: predicted to have zero/negligible risk, observed to have no inhibitors (column 2); predicted to have zero/negligible risk, observed to have inhibitors (column 3); predicted to have inhibitor risk, observed to have inhibitors (column 4); and predicted to have inhibitor risk, observed to have no inhibitors (column 5). P values <0.05 are deemed statistically significant and are shown in italics.
Figure Legends

Figure 1. Schematic diagram showing how side-chain differences may, or may not, lead to novel peptide-MHC surfaces. A) If a tFVIII peptide spanning the location of the F8 missense mutation is a non-binder, it poses no risk of forming a novel pMHC surface capable of inducing an immune response. Otherwise, one needs to consider the position of the F8 missense mutation within the MHC groove of the tFVIII peptide and the corresponding endogenous peptide. For most HLA class II isoforms, positions 1, 4, 6 and 9 are MHC-facing, and positions 2, 3, 5, 7 and 8 are TCR-facing. Where the F8 missense mutation is at a downward, MHC-facing position (top row of A, denoted by a diamond), there are 2 scenarios: both tFVIII and endogenous peptides are binders, implying no risk; or the tFVIII peptide is a binder and the endogenous peptide is a non-binder, implying a potential risk. Where the F8 missense mutation is at an upward, TCR-facing position (bottom row of A, denoted by a diamond) and both peptides are binders, there is a potential risk. B) Where a tFVIII peptide is associated with a potential risk according to the preceding assessment, a peptide from elsewhere in the human proteome that a) has the same TCR-facing residues and b) is a binder, will militate against this risk, as no novel pMHC surface will be formed.

Fig 2. Flowchart for the assessment of human proteome cross-matching and missense mutation HA inhibitor risk. Shown is the process by which the presence, or otherwise, of a novel peptide-MHC surface is determined given a specific combination of endogenous F8 missense mutation and HLA isoform.

Fig 3. A peptide-MHC binding and novel surface example: Arg593Cys for HLA-DRB1*0101. (A) NetMHCII predicts there are binding tFVIII 15-mers that have two cores – IQRFLPNPA and YLTENIQRF – containing Arg593 (R593). (B) These two cores form pMHC surfaces that are novel compared to the equivalent surfaces for endogenous FVIII.

Fig 4. A proteome cross-matching example: Arg593Cys for HLA-DRB1*0101. This example concerns the two tFVIII binding cores from Figure 3, IQRFLPNPA and YLTENIQRF. (A) Both cores cross-match to 9-mers from proteins in the human proteome at TCR-facing positions 2, 3, 5, 7 and 9. (B) Both of the matched 9-mers (one each from tubulin polyglutamylase and C2orf48, but none from dysbindin) are predicted by NetMHCII to form binding cores within 15-mers derived from these proteins. Hence, we conclude that the F8 missense mutation/HLA isoform combination Arg593Cys/HLA-DRB1*0101 is associated with negligible risk of inhibitor formation.

Figure 5. MHC-binding strengths of F8 peptides predicted to form novel pMHC surfaces with and without proteome scanning. Heatmap showing the predicted occurrence of novel pMHC surfaces and binding strengths for 25 HLA-DR/DP/DQ isoforms (y axis) covering the first 50 missense mutations in the Factor VIII Gene (F8) Variant Database (x axis). Black and grey squares indicate F8 missense mutation/HLA isoform combinations that are not predicted to form a novel pMHC surface. Otherwise the temperature color scale indicates the predicted binding strength of the strongest binding peptide with a novel pMHC surface for each remaining F8 missense mutation/HLA isoform combination. The full heatmap for all missense F8 mutations is given in Online Supplementary Figure S1. (A) MHC-binding strengths of F8 peptides predicted to form novel pMHC surfaces (colored squares), or not (black squares), without proteome scanning. (B) MHC-binding strengths of F8 peptides predicted to form novel pMHC surfaces with proteome scanning. Grey squares indicate F8 missense mutation/HLA isoform combinations that are no longer predicted to form a novel pMHC surface after cross-matches to the proteome are taken into account.
A

- tFVIII peptide
  - TCR
  - MHC
  - different MHC-facing side-chain
  - 2 3 5 7 8
  - 1 4 6 9

- endogenous FVIII peptide
  - TCR
  - MHC
  - 2 3 5 7 8
  - 1 4 6 9

B

- tFVIII peptide
  - TCR
  - MHC
  - matching TCR-facing side-chain
  - 2 3 5 7 8
  - 1 4 6 9

- matching proteome peptide
  - TCR
  - MHC
  - 2 3 5 7 8
  - 1 4 6 9
F8 MUTATION

Calculate spanning and binding tFVIII 15-mers

Remaining tFVIII 15-mers?

yes

Select strongest binding tFVIII 15-mer

Is there an endogenous FVIII 15-mer in same register?

no

Is it a binder?

no

Scan proteome: find 9-mers that match binding core of tFVIII 15-mer at TCR-facing positions

Remaining proteome 9-mers that match tFVIII binding core?

no

Does the 9-mer form the core of a binding proteome 15-mer?

yes

Proteome protection

no

Is the mutation at a TCR-facing position?

yes

NO NOVEL SURFACE

no
A

The image shows a diagram of the A2 domain with a sequence spanning 593 and its corresponding MHC II binding cores and affinities. The sequence is FDEBRSWYLTIQRFLPNPAGVQLEDPE.

- **15-mer**
  - FDEBRSWYLTIQR
  - DEBRSWYLTIQRF
  - EBRSWYLTIQRFL
  - BRSWYLTIQRFLP
  - RSWYLTIQRFLPN
  - SWYLTIQRFLPNP
  - WYLTIQRFLPNPA
  - YLTIQRFLPNPAG
  - LTIQRFLPNPAGV
  - TLIQRFLPNPAGVQ
  - ENIQRFLPNPAGVQL
  - NIQRFLPNPAGVQLE
  - IQRFLPNPAGVQLED
  - QRFLPNPAGVQLEDP
  - RFLPNPAGVQLEDPE

- **MHC II binding core**
  - FDEBRSWYL
  - YLTENIQR
  - YLTENIQRF
  - YLTENIQRFL
  - YLTENIQRFLP
  - YLTENIQRFLPN
  - YLTENIQRFLPNP
  - YLTENIQRFLPNPA
  - YLTENIQRFLPNPAG
  - YLTENIQRFLPNPAGV
  - YLTENIQRFLPNPAGVQ
  - YLTENIQRFLPNPAGVQLE
  - YLTENIQRFLPNPAGVQLED
  - YLTENIQRFLPNPAGVQLEDP
  - YLTENIQRFLPNPAGVQLEDPE

- **predicted binding affinity, IC50 (nmol/L)**
  - 601.9
  - 118.9
  - 33.3
  - 39.4
  - 28.4
  - 75.8
  - 103.9
  - 54.8
  - 21.3
  - 11.9
  - 9.0
  - 8.9
  - 9.6
  - 16.7
  - 33.8

B

The image compares the binding of tFVIII peptides to MHC II binding cores and endogenous FVIII peptides. The peptides and their corresponding MHC binding are depicted in the TCR and MHC compartments.
**A**

FVIII: IQRFPLNPAA
matching pattern: XQRXLXNPX
tubulin polyglutamylase: FQRELNNPL

FVIII: YLTENIQRF
dysbindin: GLTENSQRD
C2orf48: ELTKNAQRA

**B**

<table>
<thead>
<tr>
<th>15-mer</th>
<th>MHC II binding core</th>
<th>predicted binding affinity, IC$_{50}$ (nmol/l)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tubulin polyglutamylase matches to pattern XQRXLXNPX:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGRAASFQRELNNPL</td>
<td>FQRELNNPL</td>
<td>26.3</td>
</tr>
<tr>
<td>GRAASFQRELNNPLK</td>
<td>FQRELNNPL</td>
<td>13.8</td>
</tr>
<tr>
<td>RAASFQRELNNPLKRM</td>
<td>FQRELNNPL</td>
<td>7.9</td>
</tr>
<tr>
<td>AASFQRELNNPLKRMK</td>
<td>FQRELNNPL</td>
<td>5.3</td>
</tr>
<tr>
<td>ASFQRELNNPLKRMK</td>
<td>FQRELNNPL</td>
<td>6.4</td>
</tr>
<tr>
<td>SFQRELNNPLKRMKE</td>
<td>FQRELNNPL</td>
<td>8.2</td>
</tr>
<tr>
<td>FQRELNNPLKRMKEE</td>
<td>FQRELNNPL</td>
<td>11.3</td>
</tr>
<tr>
<td>dysbindin matches to pattern XLTNXQRX:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSSPGLTENSQRDPS</td>
<td>GLTENSQRD</td>
<td>8207.8</td>
</tr>
<tr>
<td>SSPGLTENSQRDPSSE</td>
<td>GLTENSQRD</td>
<td>7036.4</td>
</tr>
<tr>
<td>uncharacterized protein C2orf48 match to pattern XLTNXQRX:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGWELTKNAQRALGS</td>
<td>ELTKNAQRA</td>
<td>197.9</td>
</tr>
</tbody>
</table>
Supplemental Information

Supplemental Table S1. Breakdown of calculations required to identify novel peptide-MHC surfaces.

<table>
<thead>
<tr>
<th>Description</th>
<th>Count</th>
<th>Calculation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1: calculations required for a single residue</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15-mers spanning residue</td>
<td>15</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Positions 15-mer can bind to MHC</td>
<td>7</td>
<td>15 × 7</td>
<td>105</td>
</tr>
<tr>
<td>HLA alleles</td>
<td>25</td>
<td>105 × 25</td>
<td>2,625</td>
</tr>
<tr>
<td><strong>Step 2: calculations required for endogenous FVIII</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(using total from step 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reported F8 missense mutations</td>
<td>956</td>
<td>2,625 × 956</td>
<td>2,509,500</td>
</tr>
<tr>
<td><strong>Step 3: calculations required for tFVIII</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(using total from step 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Locations associated with F8 mutations</td>
<td>605</td>
<td>2,625 × 605</td>
<td>1,588,125</td>
</tr>
<tr>
<td><strong>Step 4: summation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum totals from step 2 and step 3</td>
<td></td>
<td>2,509,500 + 1,588,125</td>
<td>4,097,625</td>
</tr>
<tr>
<td><strong>Step 5: proteome scanning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two different types of calculation are required: a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>matching 9-mer cores of risk-associated peptides (from</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>step 4) to proteome 9-mers (see Figure 4A); and b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>calculating the MHC binding cores and binding strengths</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>of matched peptides (see Figure 4B). We do not keep</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>track of the exact number of calculations performed, so</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>here we estimate the lower bounds.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>a) Proteome matching:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk-associated peptides</td>
<td>≥12,189</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-identical proteome 9-mers</td>
<td>11,272,502</td>
<td>12,189 × 11,272,502</td>
<td>137,400,526,878</td>
</tr>
<tr>
<td><strong>b) MHC binding:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F8 mutation/HLA allele combinations with reduced risk</td>
<td>4,302</td>
<td></td>
<td></td>
</tr>
<tr>
<td>after scanning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rough estimate of ratio of non-binders to binders</td>
<td>10</td>
<td>4,302 × 10</td>
<td>43,020</td>
</tr>
</tbody>
</table>

Note that the number of calculations at step 2 is higher than the number at step 3 because a single location may be associated with more than one reported missense mutation, e.g. K48E, K48T and K48N.
Supplemental Table S2. Percentage of risk-associated F8 missense mutations for different HLA alleles, before and after proteome scanning

<table>
<thead>
<tr>
<th>HLA allele</th>
<th>risk (%) with 1000 nM threshold</th>
<th>risk (%) with 500 nM threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRB1*01:01</td>
<td>86 before 48 after</td>
<td>78 before 41 after</td>
</tr>
<tr>
<td>DRB1*03:01</td>
<td>34 before 25 after</td>
<td>24 before 16 after</td>
</tr>
<tr>
<td>DRB1*04:01</td>
<td>62 before 38 after</td>
<td>47 before 25 after</td>
</tr>
<tr>
<td>DRB1*04:04</td>
<td>66 before 38 after</td>
<td>53 before 26 after</td>
</tr>
<tr>
<td>DRB1*04:05</td>
<td>60 before 35 after</td>
<td>49 before 28 after</td>
</tr>
<tr>
<td>DRB1*07:01</td>
<td>70 before 46 after</td>
<td>60 before 39 after</td>
</tr>
<tr>
<td>DRB1*08:02</td>
<td>36 before 23 after</td>
<td>16 before 10 after</td>
</tr>
<tr>
<td>DRB1*09:01</td>
<td>66 before 41 after</td>
<td>50 before 25 after</td>
</tr>
<tr>
<td>DRB1*11:01</td>
<td>58 before 37 after</td>
<td>43 before 24 after</td>
</tr>
<tr>
<td>DRB1*13:02</td>
<td>34 before 25 after</td>
<td>26 before 16 after</td>
</tr>
<tr>
<td>DRB1*15:01</td>
<td>64 before 36 after</td>
<td>49 before 25 after</td>
</tr>
<tr>
<td>DRB3*01:01</td>
<td>35 before 25 after</td>
<td>24 before 17 after</td>
</tr>
<tr>
<td>DRB4*01:01</td>
<td>60 before 33 after</td>
<td>39 before 16 after</td>
</tr>
<tr>
<td>DRB5*01:01</td>
<td>58 before 40 after</td>
<td>48 before 31 after</td>
</tr>
<tr>
<td>DPA1<em>01-DPB1</em>04:01</td>
<td>41 before 27 after</td>
<td>33 before 23 after</td>
</tr>
<tr>
<td>DPA1<em>01:03-DPB1</em>02:01</td>
<td>46 before 29 after</td>
<td>36 before 21 after</td>
</tr>
<tr>
<td>DPA1<em>02:01-DPB1</em>01:01</td>
<td>59 before 34 after</td>
<td>47 before 25 after</td>
</tr>
<tr>
<td>DPA1<em>02:01-DPB1</em>05:01</td>
<td>30 before 18 after</td>
<td>17 before 8 after</td>
</tr>
<tr>
<td>DPA1<em>03:01-DPB1</em>04:02</td>
<td>52 before 33 after</td>
<td>41 before 22 after</td>
</tr>
<tr>
<td>DQA1<em>01:01-DQB1</em>05:01</td>
<td>27 before 21 after</td>
<td>19 before 15 after</td>
</tr>
<tr>
<td>DQA1<em>01:02-DQB1</em>06:02</td>
<td>54 before 34 after</td>
<td>36 before 20 after</td>
</tr>
<tr>
<td>DQA1<em>03:01-DQB1</em>03:02</td>
<td>21 before 15 after</td>
<td>10 before 6 after</td>
</tr>
<tr>
<td>DQA1<em>04:01-DQB1</em>04:02</td>
<td>21 before 13 after</td>
<td>11 before 6 after</td>
</tr>
<tr>
<td>DQA1<em>05:01-DQB1</em>02:01</td>
<td>36 before 18 after</td>
<td>24 before 12 after</td>
</tr>
<tr>
<td>DQA1<em>05:01-DQB1</em>03:01</td>
<td>58 before 32 after</td>
<td>40 before 20 after</td>
</tr>
</tbody>
</table>
Supplemental Figure S1. MHC-binding strengths of F8 peptides predicted to form novel pMHC surfaces with and without proteome scanning.

Heatmap showing the predicted occurrence of novel pMHC surfaces and binding strengths for 25 HLA-DR/DP/DQ alleles (y axis) covering the complete set of missense mutations in the Factor VIII Gene (F8) Variant Database (x axis). Black and grey squares indicate F8 missense mutation/HLA allele combinations that are not predicted to form a novel pMHC surface. Otherwise the temperature color scale indicates the predicted binding strength of the strongest binding peptide with a novel pMHC surface for each remaining F8 missense mutation/HLA allele combination. (A) MHC-binding strengths of F8 peptides predicted to form novel pMHC surfaces without proteome scanning. (B) MHC-binding strengths of F8 peptides predicted to form novel pMHC surfaces with proteome scanning. Grey squares indicate F8 missense mutation/HLA allele combinations that are no longer predicted to form a novel pMHC surface after cross-matches to the proteome are taken into account.