The humanness of macaque antibody sequences

Philippe Thullier[a], Oliver Huish[b], Thibaut Pelat[a], Andrew C. R. Martin[*,b]

[a]*Centre de Recherche du Service de Santé des Armées, DBAT/Biotechnologies des Anticorps, 24, Avenue des Maquis du Grésivaudan, B.P. 87 38702 La Tronche Cedex - France*

[b]*Institute of Structural and Molecular Biology, Darwin Building, University College London, Gower Street, London WC1E 6BT, United Kingdom*

Abstract

Chimeric, humanized and human antibodies have successively been exploited as therapeutics because their increasing human ('self') character is expected to correspond with decreased immunogenicity, critical for their clinical development. Thus humanness has been inferred to predict antibody immunogenicity. Humanness of antibody variable (V) regions has recently been studied using a parameter (here referred to as the H-score) which evaluates similarity to expressed human sequences. Macaque (*Macaca fascicularis*) antibody sequences are of particular interest because they have been suggested to have extremely human-like character and recently macaque scFvs of very high affinity against various antigens have been isolated. In this study, the H-scores of all macaque antibody V-regions available in sequence databanks were compared with those of their human counterparts using statistical tests. The results were found to be influenced by the relative size of the human families to which the macaque V-regions are related. As the relevance of families to immunogenicity is suspected but unproven, a new parameter (the 'G-score') was derived from the H-score to avoid this influence and macaque V-regions sequences were re-analyzed using the G-score. Both parameters show that these regions cannot be regarded as human when they derive from heavy chains, but the humanness of light chains is variable. It was shown that 'germline humanization' of a macaque V-region favourably influenced its humanness as evaluated by both H- and G-scores. In addition, humanness of macaque sequences presented in patents has been analyzed. H- and G-scores define objectively the humanness of antibody V-regions and their use is exemplified here.

Key words: immunogenicity, humanization, patent, HAMA response, anti-antibody response

1. Introduction

Recombinant antibodies are a class of therapeutic molecules of ever-increasing applicability. However, the question of their immunogenicity is particularly important, because the frequency of anti-antibody response (AAR) varies by a wide margin (0 to 34% for humanized antibodies)[1] and may correspond to severe side effects (serum sickness or anaphylaxis). Murine antibodies are generally unsuitable for therapeutic purposes as they induce a strong AAR (also known as a HAMA – human anti-mouse antibody) response[1]. Historically, this has been reduced by replacement of murine constant domains (the Fc fragment together with C_H1 and C_L domains) with their human counterparts ('chimerization'[2]), and further reduced by the replacement of both their constant domains and variable frameworks (FRs) in a second approach known as 'humanization'[3,4]. These replacements of murine by human regions, because of the presumed lack of immunogenicity in humans of human 'self' antibodies, effectively made the use of antibodies as therapeutics possible. The trend to diminish the size of murine regions has culminated with the recent clinical availability of a 'fully human' antibody (Humira®), expected to be less immunogenic than previous antibodies[5]. According to this trend, assessing the 'humanness' of an antibody may help in prediction of immunogenicity[6,7] and this notion is directly supported by Tables 1–3 of Hwang and Foote's paper[1] which show percentages of human recipients exhibiting AAR. Averaging the AAR in these tables shows that the response decreases from approximately 60% when using mouse antibodies to 25 and 5% when using chimeric and humanized antibodies respectively. In other words, there is a clear decrease in immunogenicity as the human-like character of the antibodies increases. A numerical assessment of 'humanness' may also be of more formal use because international nomenclature attributes the suffix -mumab to 'human' antibodies, but with no objective or independent evaluation at present, even though the engineering of antibodies may increase ('humanization') or decrease (for example, during *in vitro* affinity maturation) the 'human character' of an antibody.

The 'raw humanness' (μ_i) of an antibody variable domain was recently defined[8] as the mean sequence identity of a variable (V) region sequence (i), scored against all distinct human variable regions of the same isotype (V_H, $V\kappa$ or $V\lambda$) present in the last public release of the Kabat database[9]. The 'raw humanness' parameter was centered to allow each sequence to be assigned a Z-score (a simple statistical measure also known as the 'standard score', $Z = (x - \mu)/\sigma$) which was defined as the 'humanness' for the considered antibody sequence. A sequence with a 'humanness' score of zero has average similarity with the panel of expressed human sequences seen in Kabat. Scores above zero have a higher than average similarity with

*Corresponding author

expressed sequences and are therefore more representative of the expressed repertoire in Kabat. Conversely sequences with a negative score have less than average similarity to the expressed sequences. Clearly, since this humanness score is based on observed expressed sequences, it is influenced by the relative use of different germline families in each of the three isotypes (V_H, Vκ or Vλ) as stated in the initial study[8].

In the present study, to avoid any influence of germline usage, we introduce a new Z-score normalized measure of humanness evaluated at the level of germline-derived families. To avoid any confusion, we will refer to these two different measures based on Z-scores as the 'H-score' (Humanness score) for the previously described human expressed repertoire-based Z-score, and the 'G-score' (Germline-derived score) for the human germline-derived family-based Z-score, respectively.

The present study begins by establishing a statistical test to examine whether or not any group of V-regions has a mean H-score that is statistically significantly different from the mean human value. If no significant difference is observed, these V-regions may be regarded as having the same humanness as their human counterparts – in short, they may be regarded as indistinguishable from human sequences. This test was used to analyze the humanness of the V-regions originating from cynomolgus macaques ('crab-eating macaques' or *Macaca fascicularis*) and available in sequence databanks. As expected, the results were found to be influenced by isotypes and families; V-regions belonging to the most highly used human germline families are rated as more human by H-scores. Families may have an important role in immunogenicity and one may suspect that the most utilized families are less immunogenic than less frequently used families; however, this is, as yet, unproven[8].

While a 'germinality index' has previously been presented to measure the similarity of a V-region sequence to a human germline in order to predict its immunogenicity[10], the quality of this prediction was not assessed and a benchmark characterizing antibodies in terms of expressed sequences can be seen as more straightforward and relevant and thus may be preferred. Thus a score similar to the H-score, but avoiding the effect of relative family use by examining each germline-derived family separately may be a better way of evaluating the human-like character of antibodies and thus may be utilized as part of this important, but difficult, immunogenicity prediction. We therefore introduce a new parameter, the G-score, which is calculated in the same way as the H-score, but avoids influence of family usage on evaluation of humanness by comparing against sequences from the same family (i.e. derived from the same human germline gene) separately. This new parameter was utilized to re-analyze V-regions from *M. fascicularis* and the results of both H- and G-score-based analyses are presented and compared. Both parameters were then used to assess if the 'germline humanization' of a macaque antibody, recently presented[10], can be predicted to reduce its immunogenicity. In addition, the sequences which are at the core of patents (parallel patents number 1 266 965 B1 in Europe, and 5,693,780 in the U.S.) restricting the therapeutic use of V-regions from macaques ('non-human primates of the New World') have been analyzed with H- and G-scores.

To supplement the H-score server at `http://www.bioinf.org.uk/abs/shab/` we have provided a G-score server at `http://www.bioinf.org.uk/abs/gscore/`, enabling any scientist to use both parameters freely to aid in prediction of the immunogenicity of antibody sequences. Of course it should be stressed that many other factors (such as aggregation) are important in immunogenicity and only studying similarity with human sequences is a simplification. Nonetheless, it may be anticipated that similarity to expressed human sequences (and thus the use of the H-score or the G-score) will have a major impact on immunogenicity and thus the clinical development of antibodies.

2. Results and Discussion

2.1. Humanness of macaque variable domains retrieved from databanks, according to Z scores

Analysis of the 131 macaque (*M. fascicularis*) sequences was performed for each chain type, according to the H-score definition. With a mean H-score equal to -0.449, the 74 macaque heavy-chain sequences were found to be significantly different from human with a very high probability ($p < 0.00001$). On the contrary, the variable light domains were not significantly different from human sequences, with a mean H-score of -0.044 for the 47 κ-related macaque sequences, and a mean H-score of -0.33 for the 10 macaque λ-related sequences.

Because humanness results, as evaluated by the H-score, differed depending on the chain type, it was decided to expand on the V-region nomenclature and divide macaque sequences into families. The results presented in Table 1 show that all groups of macaque sequences related to human V_H families have mean H-scores much lower than the smallest individual human H-score in the family. Clearly these mean scores could not have been obtained from sampling human sequences. This result fully supports the conclusion above that macaque heavy-chain sequences have a lower humanness than human sequences and are thus not human-like in character. However, it should also be noted that, contrary to all other families, the H-score of the macaque sequences related to the human V_H-III family is positive. While this may reflect a higher degree of similarity of these sequences with human sequences, it may also reflect the very frequent use of the V_H-III family in the human repertoire.

In the light chains, the mean H-scores of the groups of macaque sequences related to Vκ-I and Vκ-II families are higher than the smallest human H-score in the corresponding family. Contrary to what was observed above for heavy chains, this indicates that macaque κ chains from these macaque families have a higher humanness than certain human sequences of the equivalent family and they are thus more human-like in character than the heavy chains. However, this was not the case for the Vκ-III and Vκ-IV-related macaque sequences, whose mean H-score was lower than the smallest human H-score in that family. The λ-related macaque sequences showed a similar trend with Vλ-III-related macaque sequences having a relatively high mean H-score indicating a human-like character, in contrast to the Vλ-I-related macaque sequences (Table 1). These data illustrate the importance of studying humanness at the family level.

Table 1: Summary of H-scores for macaque antibody sequences.

Heavy chain sequences	V_H-I	V_H-III	V_H-IV	V_H-V
Number of sequences	9	32	31	2
Mean H-score	-1.097	0.243	-0.938	-1.032
Lowest human H-score in the family	-0.84	0.89	-0.01	0.18
κ-related sequences	Vκ-I	Vκ-II	Vκ-III	Vκ-IV
Number of sequences	41	3	2	1
Mean H-score	0.031	-1.081	0.092	-0.261
Lowest human H-score in the family	-0.19	-2.37	0.87	0.07
λ-related sequences	Vλ-I	Vλ-III		
Number of sequences	7	3		
Mean H-score	-0.458	-0.038		
Lowest human H-score in the family	0.4	-0.17		

Macaque sequences were assigned to the most similar human family based on the best G-scores as described in Table 3

It should be noted however, that no statistical tests were performed on the data presented above and obtained at the family level, owing to the absence of a mean H-score for each family; trends should therefore only be regarded as indicative. The introduction of the G-score was aimed precisely at improving this analysis.

2.2. Introduction of the G-score for humanness

The fact that the influence of family usage on immunogenicity has not been experimentally validated, the discrepancies between the results evaluated family-by-family using H-scores, and the inability to perform statistical tests on those results, led us to derive a new parameter, the G-score. Whereas the H-score (which relates humanness to the complete expressed human repertoire) divides the available sequences into heavy, λ and κ, the G-score divides the available human sequences into families based on the germline genes from which they are derived. Each human expressed sequence in the Kabat database had to be assigned to a germline gene from which it was most likely to be derived, thus defining families on which new calculations of mean and standard deviation sequence identity were performed. Macaque sequences were then tested against each of these germline-derived families in turn (see Materials and Methods). Thus the G-score gives a measure of humanness that is related to how typical a sequence is of a family and is not influenced by the relative usage of different germline genes.

2.3. Assignment of expressed human sequences to germline families

A summary of the human germline sequence data collated by IMGT[11, 12] and available through the NCBI web site is shown in Table 2. Assignment of expressed human sequences from Kabat to germline families was performed using `tblastn` (see Materials and Methods). Of the 147 possible germline sequences used, the expressed human sequences were assigned to 127, but of these, 25 had only two human sequences leaving 102 families that could be used for G-score analysis. Within each of these 102 families, the mean and standard deviation 'representativeness' (see Materials and Methods) were calculated and used for evaluation of G-scores.

Table 2: Summary of germline sequence data extracted via the NCBI website.

	Number of alleles	Number of unique germline genes
V_H	229	53
Vλ	77	39
Vκ	73	55
Total	379	147

The 'number of alleles' reflects the total number of sequences extracted while the 'unique germline genes' reflects the grouping of allelic variations of the same gene (see Materials and Methods).

2.4. G-score humanness of macaque variable domain sequences retrieved from databanks

Each of the 131 macaque sequences was compared against each of the 102 human families using ssearch35. Average sequence similarity ('raw representativeness') was calculated against each family and converted into a 'G-score' using the family-specific mean and standard deviation. The results for the 102 families were sorted and the highest score was reported as the final G-score for the sequence. Results are presented in Table 3, which also shows a *p*-value, obtained from the 'error function' (the integral of the Gaussian distribution). This is the probability of seeing this G-score or above by chance when examining human sequences.

2.5. Evaluation of the humanness of macaque V-region families retrieved from sequence databanks using G-scores

The relatedness of each macaque sequence to a human family was defined based on the highest G-score obtained and macaque

Table 3: Best G-scores for each macaque sequence

Macaque ID	Human Germline Family ID	G-score	$\bar{\mu}$	σ	N	p-value
A29590	IGKV1-6	-0.278	79.175	3.631	14	0.610
A29591	IGKV1D-13	0.175	84.887	3.703	6	0.431
A29592	IGKV1-33	-0.128	78.269	4.590	54	0.551
A29594	IGKV1-6	0.334	79.175	3.631	14	0.369
A29595	IGKV1-17	-1.290	83.630	7.784	20	0.902
AJ619770	IGHV4-30-4	-1.199	74.874	4.869	36	0.885
AJ619771	IGKV1-6	0.562	79.175	3.631	14	0.287
AJ810486	IGHV4-30-4	-0.989	74.874	4.869	36	0.839
AJ810487	IGKV1-17	-0.883	83.630	7.784	20	0.811
AM406799*H2LF	IGHV3-66	-1.166	69.776	4.219	12	0.878
AM406800*L2LF	IGKV1-NL1	-0.517	80.437	4.820	11	0.697
BD218853	IGHV3-11	-1.504	75.311	6.734	33	0.934
BD218854	IGHV4-30-4	-2.212	74.874	4.869	36	0.987
BD218855	IGHV5-51	-1.457	82.503	6.412	148	0.927
BD218856	IGHV4-30-4	-1.091	74.874	4.869	36	0.862
BD218857	IGHV4-34	-3.703	79.591	6.494	135	1.000
BD218858	IGHV3-11	-0.831	75.311	6.734	33	0.797
BD218859	IGHV4-30-4	-2.392	74.874	4.869	36	0.992
BD218860	IGHV4-30-4	-0.767	74.874	4.869	36	0.779
BD218861	IGKV1-33	-0.517	78.269	4.590	54	0.697
BD218862	IGKV4-1	-0.294	87.543	5.415	69	0.616
BD218863	IGKV2-30	-1.249	85.258	4.022	26	0.894
BD218864	IGKV1-NL1	-0.706	80.437	4.820	11	0.760
BD218865	IGKV1-6	0.629	79.175	3.631	14	0.265
BD218866	IGKV1-17	-1.945	83.630	7.784	20	0.974
DQ065576	IGHV3-11	-1.216	75.311	6.734	33	0.888
DQ065577	IGHV4-30-4	-1.577	74.874	4.869	36	0.943
DQ065578	IGHV1-46	-1.651	70.308	6.313	41	0.951
DQ065579	IGHV3-15	-1.323	80.983	5.261	77	0.907
DQ065580	IGHV3-15	-1.323	80.983	5.261	77	0.907
DQ065581	IGHV4-30-4	-1.476	74.874	4.869	36	0.930
DQ065582	IGHV1-46	-1.547	70.308	6.313	41	0.939
DQ065583	IGHV4-34	-2.627	79.591	6.494	135	0.996
DQ065584	IGHV4-34	-0.987	79.591	6.494	135	0.838
DQ065585	IGHV3-11	-1.996	75.311	6.734	33	0.977
DQ065586	IGHV4-30-4	-1.290	74.874	4.869	36	0.901
DQ065587	IGHV4-30-4	-2.730	74.874	4.869	36	0.997
DQ065588	IGHV4-30-4	-1.131	74.874	4.869	36	0.871
DQ065589	IGHV4-b	-2.067	80.056	5.687	17	0.981
DQ065590	IGHV3-11	-2.139	75.311	6.734	33	0.984
DQ065591	IGHV4-30-4	-1.662	74.874	4.869	36	0.952
DQ065592	IGHV4-30-4	-1.437	74.874	4.869	36	0.925
DQ065593	IGHV3-11	-2.131	75.311	6.734	33	0.983
DQ065594	IGHV3-15	-1.343	80.983	5.261	77	0.910
DQ065595	IGHV4-30-4	-1.351	74.874	4.869	36	0.912
DQ065596	IGHV3-11	-2.813	75.311	6.734	33	0.998
DQ065597	IGHV5-51	-2.320	82.503	6.412	148	0.990
DQ065598	IGHV4-30-4	-1.366	74.874	4.869	36	0.914
DQ065599	IGHV4-30-4	-0.807	74.874	4.869	36	0.790
DQ065600	IGHV1-46	-2.278	70.308	6.313	41	0.989
DQ065601	IGHV4-59	-2.545	79.702	5.543	60	0.995
DQ065602	IGHV1-46	-1.248	70.308	6.313	41	0.894
DQ065603	IGHV1-46	-1.771	70.308	6.313	41	0.962
DQ065604	IGHV3-49	-2.042	79.442	4.493	12	0.979
DQ065605	IGHV4-30-4	-2.536	74.874	4.869	36	0.994
EU346094	IGKV1-NL1	-0.706	80.437	4.820	11	0.760
EU346095	IGHV3-66	-0.379	69.776	4.219	12	0.648
EU346096	IGKV1-NL1	-0.706	80.437	4.820	11	0.760
EU346097	IGHV3-66	-0.421	69.776	4.219	12	0.663
EU346098	IGKV1-NL1	-0.706	80.437	4.820	11	0.760
EU346099	IGHV3-66	-0.383	69.776	4.219	12	0.649
EU346100	IGKV1-6	0.398	79.175	3.631	14	0.345
EU346101	IGHV3-66	-1.466	69.776	4.219	12	0.929
EU346102	IGKV1-17	-1.019	83.630	7.784	20	0.846
EU346103	IGHV3-66	-1.264	69.776	4.219	12	0.897
EU346104	IGKV1-NL1	-0.706	80.437	4.820	11	0.760
EU346105	IGHV3-66	-0.405	69.776	4.219	12	0.657
EU346106	IGKV1-33	-1.353	78.269	4.590	54	0.912
EU346107	IGKV1-33	-1.302	78.269	4.590	54	0.904
EU346108	IGKV1-NL1	-0.706	80.437	4.820	11	0.760
EU346109	IGHV3-66	-0.405	69.776	4.219	12	0.657
EU346110	IGKV1-NL1	-0.481	80.437	4.820	11	0.685
EU346111	IGHV3-66	-0.379	69.776	4.219	12	0.648
EU346112	IGKV1-NL1	-0.706	80.437	4.820	11	0.760
EU346113	IGHV3-66	-0.565	69.776	4.219	12	0.714
EU346114	IGKV1-NL1	-0.928	80.437	4.820	11	0.823
EU346115	IGHV3-43	-0.136	75.136	2.801	9	0.554
EU346116	IGKV1-NL1	-0.706	80.437	4.820	11	0.760
EU346117	IGHV3-66	-0.379	69.776	4.219	12	0.648
EU346118	IGKV1-17	-4.584	83.630	7.784	20	1.000
EU346119	IGHV3-66	-0.650	69.776	4.219	12	0.742
EU346120	IGKV1-NL1	-0.706	80.437	4.820	11	0.760
EU346121	IGHV3-66	-0.725	69.776	4.219	12	0.766
EU346122	IGKV1-33	-0.687	78.269	4.590	54	0.754
EU346123	IGHV3-11	-1.849	75.311	6.734	33	0.968
EU346124	IGKV1-NL1	-0.706	80.437	4.820	11	0.760
EU346125	IGHV3-66	-0.405	69.776	4.219	12	0.657
EU346128	IGKV1-NL1	-0.711	80.437	4.820	11	0.762
EU346129	IGHV3-66	-1.452	69.776	4.219	12	0.927
EU346130	IGKV1-33	-1.074	78.269	4.590	54	0.859
EU346131	IGHV3-11	-1.209	75.311	6.734	33	0.887
EU346132	IGKV1-33	-0.801	78.269	4.590	54	0.788
EU346133	IGHV3-66	-0.443	69.776	4.219	12	0.671
EU346134	IGHV3-66	-0.650	69.776	4.219	12	0.742
EU346135	IGKV1-17	-0.841	83.630	7.784	20	0.800
EU346136	IGHV3-66	-0.942	69.776	4.219	12	0.827
EU346137	IGKV1-NL1	-1.317	80.437	4.820	11	0.906
EU346138	IGHV3-66	-0.650	69.776	4.219	12	0.742
EU359720	IGHV4-30-4	-1.019	74.874	4.869	36	0.846
EU359721	IGLV1-36	-1.597	76.481	4.240	35	0.945
EU359722	IGHV4-30-4	-0.436	74.874	4.869	36	0.669
EU359723	IGKV1-9	-1.323	84.176	6.455	13	0.907
EU359724	IGHV3-11	-0.586	75.311	6.734	33	0.721
EU359725	IGLV1-36	-2.096	76.481	4.240	35	0.982
EU359726	IGHV3-11	-1.140	75.311	6.734	33	0.873
EU359727	IGKV1-6	-0.469	79.175	3.631	14	0.681
EU359728	IGHV4-30-4	-0.558	74.874	4.869	36	0.712
EU359729	IGLV3-21	-1.762	82.725	4.726	99	0.961
EU359730	IGKV3-NL5	-0.072	80.675	3.633	8	0.529
EU359731	IGHV4-30-4	0.064	74.874	4.869	36	0.474
EU359732	IGHV1-46	-0.774	70.308	6.313	41	0.780
EU359733	IGKV1-6	-0.813	79.175	3.631	14	0.792
EU359734	IGHV1-f	1.023	68.771	2.277	8	0.153
EU359735	IGKV1-6	-0.813	79.175	3.631	14	0.792
EU359736	IGHV1-46	-0.972	70.308	6.313	41	0.835
EU359737	IGKV3-NL5	-1.321	80.675	3.633	8	0.907
EU359738	IGHV4-30-4	-0.674	74.874	4.869	36	0.750
EU359739	IGKV1-6	0.062	79.175	3.631	14	0.475
EU359740	IGLV1-36	-2.289	76.481	4.240	35	0.989
EU359741	IGHV4-30-4	-1.001	74.874	4.869	36	0.841
EU359742	IGKV1-6	-0.715	79.175	3.631	14	0.763
EU359743	IGHV3-11	-0.586	75.311	6.734	33	0.721
EU359744	IGLV1-36	-2.096	76.481	4.240	35	0.982
EU359745	IGHV1-46	-1.095	70.308	6.313	41	0.863
EU359746	IGHV4-30-4	-1.019	74.874	4.869	36	0.846
EU359747	IGLV1-36	-1.213	76.481	4.240	35	0.887
EU359748	IGHV3-66	-0.352	69.776	4.219	12	0.637
EU359749	IGKV1-6	-0.546	79.175	3.631	14	0.707
EU359750	IGHV4-30-4	-1.058	74.874	4.869	36	0.855
EU359751	IGLV3-21	-1.370	82.725	4.726	99	0.915
EU359752	IGHV4-30-4	-1.019	74.874	4.869	36	0.846
EU359753	IGLV1-36	-1.213	76.481	4.240	35	0.887
EU359754	IGHV4-30-4	-1.019	74.874	4.869	36	0.846
EU359755	IGLV1-36	-1.213	76.481	4.240	35	0.887
EU359756	IGLV1-40	-1.873	88.430	4.943	82	0.969
EU359757	IGHV4-30-4	-1.019	74.874	4.869	36	0.846
EU359758	IGLV1-36	-1.863	76.481	4.240	35	0.969
EU359759	IGLV1-36	-0.716	76.481	4.240	35	0.763
EU359760	IGHV3-66	-0.873	69.776	4.219	12	0.809
EU359761	IGKV1-6	-0.536	79.175	3.631	14	0.704
EU359762	IGHV4-30-4	-1.013	74.874	4.869	36	0.845
EU359763	IGLV3-21	-1.370	82.725	4.726	99	0.915
EU359764	IGHV4-30-4	-0.436	74.874	4.869	36	0.669
EU359765	IGLV3-21	-0.862	82.725	4.726	99	0.806
EU359766	IGHV4-30-4	-1.058	74.874	4.869	36	0.855
EU359767	IGKV1-6	-1.120	79.175	3.631	14	0.869
L13307	IGHV4-30-4	-0.016	74.874	4.869	36	0.506
L13308	IGKV1-6	-0.329	79.175	3.631	14	0.629
L13309	IGKV1D-13	0.193	84.887	3.703	6	0.424
L13310	IGKV1-33	-0.268	78.269	4.590	54	0.606
L13311	IGKV2-30	-1.030	85.258	4.022	26	0.848
L13312	IGKV1-6	0.688	79.175	3.631	14	0.246
L13313	IGKV1-17	-1.161	83.630	7.784	20	0.877
L13314	IGKV2-30	-0.793	85.258	4.022	26	0.786
L13315	IGKV1-9	-0.268	84.176	6.455	13	0.606
L13316	IGKV1D-13	0.445	84.887	3.703	6	0.328
L13317	IGKV1-6	0.284	79.175	3.631	14	0.388

Continued…

…Continued

The macaque sequence identifiers are shown together with the human germline-derived family from which they obtained the best G-score. Also shown are the mean ($\bar{\mu}$), standard deviation (σ), number of family members (N) and a simple p-value for the sequence being human derived directly from the G-score.

Table 4: G-scores for macaque antibody sequences.

Human Family[a]	NSeq[b]	Mean G-score	*p*-value
IGHV1-46	8	-1.417	< 0.0001
IGHV1-f	1	1.023	Not tested
IGHV3-11	11	-1.583	< 0.00001
IGHV3-15	3	-1.330	Not tested
IGHV3-43	1	-0.136	Not tested
IGHV3-49	1	-2.042	Not tested
IGHV3-66	16	-0.759	< 0.01
IGHV4-30-4	26	-1.220	< 0.0000001
IGHV4-34	3	-2.439	Not tested
IGHV4-59	1	-2.545	Not tested
IGHV4-b	1	-2.067	Not tested
IGHV5-51	2	-1.889	Not tested
IGKV1-17	6	-1.191	Not tested
IGKV1-33	8	-0.766	< 0.05
IGKV1-6	15	-0.123	NS[d]
IGKV1-9	2	-0.795	Not tested
IGKV1D-13	3	0.271	Not tested
IGKV1-NL1	7	-0.766	NS
IGKV2-30	3	-1.024	Not tested
IGKV3-NL5	2	-0.700	Not tested
IGKV4-1	1	-0.294	Not tested
IGLV1-36	6	-1.629	Not tested
IGLV1-40	1	-1.873	Not tested
IGLV3-21	3	-1.332	Not tested

The *p*-value indicates the probability that the family is not related to a human sequence. [a]Family names are assigned based on the human germline-derived family showing the highest G-score for the macaque sequences — germline sequences and names come from IMGT; [b]NSeq = Number of macaque sequences related to the family; [c]NS = Not significant

sequences were then grouped based on these families. The G-score was used for statistical analysis by averaging across the macaque sequences related to each family. This analysis is shown in Table 4 where the number of macaque sequences, their mean G-score and results of the statistical test for humanness are indicated. Significance can only be tested where there are 7 or more macaque sequences in a family, which applied to 4 heavy-chain families and 3 κ-chain families. All four heavy-chain families show that macaque V-regions are significantly different from their human counterparts (p from < 0.01 to < 0.0000001). The three families related to human light chains show variation in agreement with the previous tests: macaque sequences related to the IGKV1-33 germline gene are significantly different from human ($p < 0.05$) while those related to IGKV1-6 and IGKV1-NL1 are not.

In summary, the statistical tests based on the G-score, which avoids the influence of family use in the human repertoire, show that V_H regions from cynomolgus macaques are not similar to their human counterparts while some macaque V_L variable domains are human-like, and others are not.

2.6. Germline humanization of a macaque Fab, according to H- and G-scores

The humanization of a macaque Fab ($35PA_{83}$, derived from an animal immunized with a sub-unit of the anthrax lethal toxin) was recently performed[10] using human germline antibody sequences as the template (or 'acceptor') sequence. This approach was called 'germline humanization', or 'germlinization' for short. The goal of this approach was to increase the degree of similarity of the macaque V-region to human germline-encoded V-regions. The rationale behind germline humanization is the fact that human, germline-encoded V-regions should be encountered as part of IgMs by all humans and thus might be expected not to be immunogenic. The germlinization process was evaluated with a parameter (the 'Germinality Index', GI) which measures the percentage similarity between parental V-regions (or their germline-humanized variants) and their most closely related, human, germline-encoded sequences. In the course of that study, the GI showed a significant increase, from a value of 0.876 to a value of 0.977 (the maximum theoretical GI value is 1.0). The GI parameter is different from H- or G-scores in that it measures similarity with germline encoded sequences (which may be expressed as IgM), but not with sequences expressed as IgG – the most common circulating antibodies – which may be more relevant when evaluating immunogenicity. Consequently, to test the robustness of this germline humanization, the H- and G-scores of the parental, macaque V-regions and of their germline humanized variants were studied. This individual example did not allow statistical tests to be performed, but the values of H- and G-scores were studied in comparison with the distribution of human sequences, using the *p*-value (from the integral of the Gaussian distribution) as presented in Table 3.

The V_H of $35PA_{83}$ belongs to the V_H-IV family and had an H-score of -0.241: this value corresponded to a humanness better than the humanness of 41% of expressed human V_H sequences. The germlinized V_H had an H-score of -0.082, corresponding to

a humanness better than 47% of human V_H-regions, thus only showing a small improvement. However, when evaluated using G-scores, the result of the V_H germlinization was much more striking. The G-score of $35PA_{83}$ V_H was -0.541, (higher than 29% of human sequences) and, after germline humanization, increased to 0.191 (higher than 57% of human sequences). The discrepancies between the H- and G-scores of the parental V_H are a result of the relatively frequent use of V_H-IV and this example clearly shows how the H-score is influenced by family usage, in contrast to the G-score. However, both analyses converge to suggest that the germlinized V_H sequence has a humanness that is near the average value for human sequences.

The macaque Vκ, belonging to the Vκ-I family, had an H-score of -0.215, corresponding to a humanness above that of 42% of human sequences, but the Vκ-I family is used frequently which could account for this relatively high H-score. After germlinization, the H-score increased to 0.684 (better than 76% of human sequences), but again may be influenced by the high usage of Vκ-I germline genes. The G-score for the macaque sequence was much lower (-0.930, higher than only 17% of human sequences) and, after germline humanization, changed to 0.339 (higher than than 63% of human sequences). Thus, quite remarkably, the germlinized $35PA_{83}$ Vκ can be regarded as having a better humanness than average human Vκ, in contrast to the parental macaque sequence. For this example at least, the H- and G-score results for germline humanization of $35PA_{83}$ V-regions converge with those of the 'germlinality index' and this triple convergence tends to show the robustness of the germlinization approach intended to lower the immunogenicity of macaque V-regions.

2.7. Humanness of macaque variable domains described in European patent 1 266 965 B1, using H- and G-scores

The distribution of the 29 macaque sequences (15 V_H sequences plus 8 variants (see below), 5 Vκ, 1 Vλ; H-scores presented in Table 5) indicated in European patent 1 266 965 B1 is such that no family is represented by more than 6 of these sequences (see Materials and Methods). As a result, no statistical study could be done at the family level using G-scores, and only the 15 V_H sequences could be statistically analyzed with H-scores. As described in the Materials and Methods, 4 macaque sequences related to V_H-III showed ambiguities (clones 9, 34, 36 and 40); in these cases, the variants having the highest H-scores (9G, 34G, 36G, 40G) were selected for further analysis. Despite this precaution, the result of the analysis is that the V_H sequences cannot be regarded as similar to human sequences ($p < 0.02$). This is in perfect accordance with the other results on macaque antibody sequences presented in this study.

Understanding the immunogenicity of antibodies is an important question, influenced by many factors including the quality of the preparation itself. However, based on the non-immunogenicity of human 'self' proteins, it is assumed that the more human-like an antibody is, the less immunogenic it will be. This reasoning is coherent with the observed decreased immunogenicity of chimerized antibodies compared with murine antibodies, and the further decreased immunogenicity of humanized antibodies. As a consequence, a major aspect of the

Table 5: H- and G-scores for sequences in the EPO 1 266 965 B1 and US 5,693,780 patents

Name	Family[a]	H-score	G-score
Clone 1-2	IGHV1-46	-0.788	-1.322
Clone 1-14	IGHV1-46	-0.353	-0.229
Clone 2-10	IGHV2-70	-1.681	-1.776
Clone 2-13	IGHV2-70	-1.754	-2.260
Clone 2-3	IGHV2-70	-1.545	-0.990
Clone 34A	IGHV3-11	0.691	-0.803
Clone 36A	IGHV3-11	-0.011	-1.921
Clone 40A	IGHV3-11	0.691	-0.607
Clone 40G	IGHV3-11	0.757	-0.622
Clone 40S	IGHV3-11	0.697	-0.512
Clone 34G	IGHV3-66	0.757	-0.790
Clone 34S	IGHV3-66	0.697	-0.654
Clone 36G	IGHV3-66	0.037	-1.859
Clone 36S	IGHV3-66	0.003	-1.728
Clone 9A	IGHV3-66	0.480	-0.725
Clone 9G	IGHV3-66	0.549	-0.725
Clone 9S	IGHV3-66	0.484	-0.541
Clone 4-13	IGHV4-30-4	-0.625	-1.029
Clone 4-14	IGHV4-30-4	-0.768	-0.746
Clone 4-16	IGHV4-30-4	-0.998	-0.672
Clone ANTI-CD4 CHIM	IGHV4-30-4	-0.949	-0.828
Clone SC CHIM (H)	IGHV4-30-4	-0.628	-0.673
Clone 5-11	IGHV5-51	-1.106	-2.309
Clone k1-3	IGKV1-17	-0.077	-0.930
Clone k1-14	IGKV1-6	-0.557	-1.122
Clone k1-7	IGKV1-6	0.631	-0.261
Clone K2-8	IGKV2-30	-1.736	-3.158
Clone SC CHIM (L)	IGKV3-NL4	-2.653	-3.893
Clone ANTI-CD4	IGLV3-21	-0.343	-1.966

[a]The 'Family' refers to the closest human germline-derived family assigned on the basis of the highest G-score for the macaque sequence.

question of the immunogenicity of any antibody – in particular examining its suitability for clinical development – may be viewed as evaluating whether the antibody will be regarded by the immune system as part of the human 'self'. In other words, it is a case of evaluating its 'humanness'. Human IgGs are however all unique proteins, owing to mutations introduced in the V-region during affinity maturation; this corresponds to the virtually infinite range of antigens that antibodies may have to bind with high affinity and specificity. Thus the 'humanness' of a V-region is not as easily determined as the 'humanness' of any other protein. Consequently, 'raw humanness' was previously defined as the mean percentage identity of the antibody V_H or V_L protein sequence with all human sequences of the same isotype (V_H, Vκ or Vλ), present in the Kabat database. From this, the H-score parameter was mathematically defined as a Z-score which centers the 'raw humanness' (i.e. its mean value is zero). When a V-region belongs to a frequently-used family, sequences of this family are over-represented in the Kabat database used to define the H-scores. Consequently, the 'raw humanness' and H-score of such a V-region increases. It has been postulated that this frequent use may reflect a lower immunogenicity, on the basis that *variations* of these sequences are seen more frequently by the immune system, and any high immunogenicity would have been detected. However, there is no direct experimental evidence to support this hypothesis. On the other hand, based on the physiology of the immune system, it may be assumed that the more frequently-used families are selected because they are more likely to bind a large variety of antigens than less-frequently used families and not because of their reduced immunogenicity. The G-score was thus defined in a way equivalent to the H-score, but with the means and standard deviations calculated at the family level, thus avoiding the influence of family usage on the results.

At the beginning of this study we confirmed that the distribution of Z-scores derived from a normal distribution of data have a standard deviation equal to one (Appendix A), such that any Z-score is a 'normalized' parameter, in addition to being 'centered'. Consequently, since H- and G-scores are Z-scores, both are centered and normalized which were necessary properties for the statistical tests. In particular, these tests assess whether or not the mean H- or G-score of V-regions originating from macaques is equivalent to the mean scores of their human counterparts. That assessment requires that a sample of sequences may be seen as representative of the macaque sequences – in other words, according to standard statistical rules, a sample of macaque V_H, Vκ or Vλ regions must contain more than 6 sequences. If the H-scores of macaque sequences have no statistical difference when compared with their human counterparts, it means that the macaque sequences have the same humanness as the human sequences and are therefore indistinguishable from them, leading to the prediction that the immunogenicity for macaque V-regions will be the same as the immunogenicity of their human counterparts (all other factors being equal). Indeed, these tests allow prediction of the immunogenicity of V-regions of any origin and are of wide interest. High affinity (nanomolar[13, 14, 15, 16]) and very high affinity (picomolar[17]) antibody fragments of macaque origin have recently been isolated for therapeutic purposes explaining our particular interest in antibodies from this source.

The tests based on the H-score were applied to all V-domains of macaque (*M. fascicularis*) origin, published and available in sequence databanks (74 V_H, 47 Vκ and 10 Vλ). These were regarded here as being representative of macaque V-regions in general. Publication of the sequences was regarded as testifying for the quality of the study and, in particular, of the sequencing process, as we wished to avoid any effect of sequencing errors that might have lowered the apparent humanness. Previous studies presenting V-regions of macaque origin[13, 14, 15, 16, 17] have shown a high degree of similarity of the corresponding V-regions with their human, germline-encoded counterparts. However, in the absence of a benchmark for 'human antibodies', the question of whether or not these antibody fragments may be regarded as human had not been fully answered, although it was described as central to the validity of the approach[18]. Here, use of the H-score-based statistical test allowed us to show that macaque V_L (Vκ or Vλ) may be regarded as human, while V_H may not. The discrepancy between V_L and V_H results, and the fact that the degree to which different families are used is known to influence H-score analysis, prompted us to perform a family-by-family analysis. This confirmed the influence of families on the results as, for instance, macaque sequences related to V_H-III (the most highly represented human V_H family) scored best, but the exact significance of this result – are macaque V_H-III sequences more human-like and thus less immunogenic than other macaque V_H sequences, or is this result only representative of the more frequent usage of the V_H-III family – was unknown.

To overcome this difficulty, a new parameter was designed to assess humanness independently of family membership. This parameter, which we call the G-score, is also calculated as a Z-score, but on each set of expressed sequences of the same family present in the Kabat database, instead of grouping all V_H, Vκ or Vλ sequences together. With this simple modification to our previous scoring, the frequency of use of each family does not influence G-score results, and prediction of immunogenicity using the G-score is purely based on assessment of similarity with human sequences that are members of the same family, thus derived from the same human germline gene. When applied to the macaque sequences, G-scores showed that none of the four family-related sets of macaque V_H, including two sets related to V_H-III, could be considered as 'equivalent to' human. These results are coherent with the H-score results for V_H, and leave no ambiguity regarding the highly used V_H-III family. Considering V_L, the results were less clear, as one set (related to IGKV1-33) was evaluated as different from human at the limit of significance ($p < 0.05$) while two others (IGKV1-6 and IGKV1-NL1) were not. With regard to results at the limit of significance, it might be that the number of macaque sequences available in the database at present is sometimes still too small to give unambiguous results. Alternatively certain families of macaque V_L are not significantly different from human V_L while others are. This is in contrast to macaque V_H where all families were rated as significantly different from their human counterparts.

3. Conclusions

These results are of great interest, justifying *a posteriori* the 'germline humanization' of macaque V-regions, of which an example ($35PA_{83}$) was previously presented[10]. Recent preliminary results show that germline humanization of macaque sequences reduces the number of potential human T-cell epitopes (Thullier and Carr, unpublished). This process increases the similarity of V-regions to their human-germline-encoded counterparts so it was interesting to evaluate whether the similarity to *expressed* human IgGs (derived from the same germline) also improved for $35PA_{83}$, as prediction of tolerance is more frequently based on similarity to expressed sequences[6, 7]. The V-regions of $35PA_{83}$ belonged to families V_H-IV and Vκ-I, both of which are rather frequently used. Consequently, the macaque H-scores were higher than the G-scores and the effect of germline humanization was more striking with G-scores (averaged G-score increase of $35PA_{83}$ V-regions: +1.005) than with H-scores (averaged H-score increase: +0.529). However both results converged to show the efficacy of the approach, showing in particular that the germline-humanized variant had a greater humanness than average human sequences (H- and G-scores averaged on germline humanized V-regions: 0.602 and 0.530 respectively). These high H- and G-scores for the germline humanized variant of $35PA_{83}$, predicting low immunogenicity, were obtained starting with parental V-regions whose negative H- and G-score values showed a humanness lower than the humanness of average human V-regions (H- and G-scores averaged on V_H and V_L: -0.228 and -0.735 respectively), in accordance with our earlier results regarding the non-human character of macaque V-regions. More broadly, these studies show that macaque V-regions may not be regarded as human, but that they may be engineered to increase their humanness and probably lower their immunogenicity to the level of human antibodies.

While this approach of isolating, then germline-humanizing, macaque V-regions may be of wide interest, it might be limited by parallel patents in Europe and the U.S.A. These patents state that macaque antibodies are 'indistinguishable' from their human counterparts, based on unquantified comparisons of 20 macaque V-regions with human 'consensus' sequences (although no precise definition of these terms is given in the patent). Not only is this statement contradicted by our statistical analysis (using H- and G-scores) of macaque V-regions whose sequences are deposited in public databanks, but it is also contradicted by our H-score analysis of the sequences presented in the patents themselves.

In conclusion, isolating and engineering V-regions of macaque origin is still a rarely-used path to obtain therapeutic antibodies, but data are accumulating to show that it should be regarded as having the same efficiency, at least, as other strategies, while potentially leading to less immunogenicity and perhaps being less constrained by valid commercial or legal aspects.

4. Materials and Methods

4.1. Retrieval of macaque sequences in databanks and European patent 1 266 965 B1

Sequences of cynomolgus macaque (*Macaca fascicularis*) V-regions were retrieved from the LIGM database, on the IMGT website (`http://imgt.cines.fr/cgi-bin/IMGTlect.jv`), utilizing the query parameter 'crab-eating macaque' as the 'English name of the species', 'RNA' or 'cDNA' as 'nucleic acid type', 'rearranged' as 'configuration' and leaving all other parameters at their default values. Only the sequences corresponding to previous publications were retained, as a quality control. Of note, two of the co-authors of the present study are the largest contributors of such sequences, with 71 of the 131 non-redundant retrieved sequences corresponding either to anti-tetanus toxin, anti-protective antigen (*B. anthracis*), anti-lethal factor (*B. anthracis*) or anti-ricin V-regions isolated from libraries constructed from immunized *M. fascicularis*.

The sequences of macaque V-regions presented in European patent 1 266 965 B1, which is the parallel of U.S. patent 5,693,780, are grouped by families: 3 sequences for the V_H-I family, 3 for V_H-II, 4 for V_H-III, 5 for V_H-IV, 1 for V_H-V, 3 for Vκ-I, 2 for Vκ-II and 1 for Vλ-VIII. Amino-acids (one-letter code) are sometimes written above macaque sequences in the patents, but their significance is unclear, so they were not taken into account in the study. Four macaque sequences related to the V_H-III family (Clones 9, 34, 36 and 40) included an 'X' with three amino-acids mentioned above the X. The three variant sequences for each macaque V-region were tested, with the number indicating the name of the clone followed by a letter indicating which amino-acid was taken into account (Table 5). For each of the four V-regions, the most human-like variant of each of the three variants was retained for the rest of this study.

4.2. Z-score-based statistics for germline-derived human families (G-scores)

The derivation of H-scores has been described previously[8]. G-scores are calculated in an analogous way, but by dividing expressed sequences into families based on the closest germline from which they are most likely to be derived. The procedure is described below with all custom code being implemented in Perl.

Using KabatMan[19], expressed variable domain light and heavy chain amino acid sequences were extracted from the most recent publicly available release of the Kabat database[9] dated July 2000. Any identical sequences were removed from the set.

Sets of functional human nucleotide germline sequences were extracted for V_H, Vκ and Vλ using the NCBI server at `http://www.ncbi.nlm.nih.gov/igblast/showGermline.cgi`. These data, which come from IMGT[11, 12], were combined into a BLAST[20] database. Each of the expressed human amino acid sequences was then searched against this database using `tblastn` to assign the most likely parent germline variable fragment for each expressed sequence. The expressed antibody sequences were then grouped based on their parent

germline sequence. Germline sequences have names such as IGHV1-18*01 and IGHV1-18*02, with the numbers following the '*' (here '01' and '02') reflecting different alleles of the same gene[11]. For the purposes of parent assignment the part following the '*' was dropped and so expressed sequences matching either of these alleles would be assigned to the same gene, IGHV1-18.

The mean and standard deviation for sequence similarity was calculated for each germline-derived family of expressed human sequences as follows. Each human germline-derived family is treated separately. Within each family, every human variable domain sequence was compared with every other human variable domain sequence using `ssearch35`[21] to generate pairwise alignments and sequence identities. Thus for a family of N members, each sequence is associated with $N-1$ sequence identities. For each sequence, i, a mean sequence identity is calculated as:

$$\mu_i = \sum_{j=1, j\neq i}^{N} P_{ij}/(N-1) \quad (1)$$

where P_{ij} is the pairwise sequence identity between the *i'th* and the *j'th* sequence in the query and target dataset respectively (a sequence is not compared against itself). This value, μ_i, referred to in our previous paper as the 'raw humanness', gives a measure of how representative a sequence is of the sequences derived from the same germline. To avoid confusion with our previous paper we call it 'representativeness' here.

From this we can calculate a 'mean representativeness' ($\bar{\mu}$) for each dataset:

$$\bar{\mu} = \sum_{i=1}^{N} \mu_i/N \quad (2)$$

and the standard deviation, σ:

$$\sigma = \sqrt{\sum_{i=1}^{N} (\mu_i - \bar{\mu})^2/N} \quad (3)$$

These values, $\bar{\mu}$ and σ are then used in evaluation of individual sequences.

4.3. Z-scores

The Z-score for sequence i is calculated as:

$$Z_i = (\mu_i - \bar{\mu})/\sigma \quad (4)$$

Any germline-gene-derived family containing < 3 sequences is excluded from this analysis as $P_{ij} = P_{ji}$. Therefore $\sigma = 0$ resulting in an infinite Z-score.

The Z-score (which is simply the number of standard deviations a value is away from the mean) is a method of normalizing a normal distribution. The mean of the Z-scores will be zero and they will have a standard deviation of 1 (as proved in Appendix A). We refer to this germline-family-derived Z-score as the G-score.

4.4. Analysis of the macaque sequences

Given the germline-derived families of expressed human sequences, together with the $\bar{\mu}$ and σ values for that family, a novel sequence may be scanned against the family using `ssearch35` to obtain a set of sequence identities. From this, one can calculate μ_i and hence the G-score.

Each of the macaque sequences was compared against each of the families and a G-score with respect to each family was calculated. These were then sorted and the best G-score was reported.

4.5. Statistics when the samples of macaque and human sequences to be compared contain at least 30 sequences

These tests are aimed at evaluating whether the sequences of macaque V-regions, as retrieved from the IMGT database or from patents, may be regarded as equivalent to their human counterparts. The tests compare the mean H-score of a representative sample of macaque sequences with the mean H-score of a reference group of human sequences. By definition of the H-score, the group of human sequences used as the reference consists of all distinct sequences of the corresponding isotype (V_H, Vκ, or Vλ), retrieved from the Kabat database.

Generally, to compare the two means (m_1 and m_2), of two groups (groups '1' and '2' containing n_1 and n_2 elements respectively), whose standard deviations are σ_1 and σ_2 respectively, the formula to be used is:

$$E = \frac{(m_1 - m_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \quad (5)$$

Here, the macaque sequences to be tested are called group 1. Group 2 is thus the reference group of (V_H, Vκ, or Vλ) human sequences. Owing to our definition of H-scores, $m_2 = 0$ and $\sigma_2 = 1$. The size, n_2, of the set of sequences on which the H-scores are established (1857 for heavy sequences, 645 for κ sequences, 1003 for the λ sequences) is high when compared with σ_2. Thus, σ_2^2/n_2 is negligible and equation 5 may be simplified as:

$$E = m_1/(\sigma_1/\sqrt{n_1}) \quad (6)$$

The absolute value of E has to be compared to ξ, read from a table of the Gaussian law and which depends on the value of the α risk. If E is greater than ξ, the two groups are statistically different. i.e. the group of macaque sequences is *not* equivalent to the reference group of human sequences; if the two groups are not significantly different, the two groups are regarded as coming from the same source or, in other words, macaque sequences may be regarded as equivalent to human.

4.6. Statistics when one of the samples of macaque and human sequences to be compared contains fewer than 30 sequences

The observed sample of macaque sequences may have fewer than 30 sequences, as the use of macaque V-regions for therapeutic purposes has been scarcely exploited thus far and relatively few sequences are available. Owing to the higher number of families (102) utilized to define G-scores, as compared

with the three isotypes utilized to define H-scores, some human families also comprise < 30 reference sequences. When the observed sample size of macaque sequences (n_1), or the reference sample size of human sequences used to define the G-score distribution (n_2) is ≤ 29 (but n_1 and $n_2 \geq 7$, otherwise the sample is not representative) the formula to be used for the comparison of means is:

$$T = \frac{(m_1 - m_2)}{\sqrt{\sigma^2/n_1 + \sigma^2/n_2}} \tag{7}$$

where σ^2 is defined as:

$$\sigma^2 = \frac{((n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2)}{(n_1 + n_2 - 2)} \tag{8}$$

and where m_2 is the mean human H- or G-score, with σ_2 being the standard deviation about this mean. By the definition of H- and G-scores, m_2 and σ_2 are respectively equal to 0 and 1, such that:

$$\sigma^2 = \frac{(n_1 - 1)\sigma_1^2 + n_2 - 1}{(n_1 + n_2 - 2)} \tag{9}$$

and, from equation 7,

$$T = \frac{m_1}{\sqrt{\sigma^2/n_1 + \sigma^2/n_2}} \tag{10}$$

thus

$$T \approx m_1/(\sigma/\sqrt{n_1}) \tag{11}$$

The absolute value of T has to be compared to t (read from the table of Student's law) at (n_1+n_2-2) degrees of freedom. (If (n_1+n_2-2) is ≥ 30, Student's law is equivalent to the Gaussian law.) If $T > t$, the two groups are statistically different and thus the group of macaque sequences is not equivalent to the reference group of human sequences; otherwise the two groups are regarded as coming from the same source or, in other words, macaque sequences may be regarded as 'human'.

5. Acknowledgments

This work was financed in part by grant PEA 030802/09co301-1 awarded to PT by the Délégation générale de l'armement. We thank Dr. K.R. Abhinandan for his original work on humanness of antibody sequences, sponsored by the BBSRC and GlaxoSmithKline under the Dorothy Hodgkin Postgraduate Award scheme.

A. Standard deviation of Z-scores

For our statistics to work correctly we needed to confirm that the standard deviation (σ) of a set of Z-scores (and thus H- or G-scores) was 1. This is often stated to be the case, but we prove it formally thus:

The standard deviation is defined as

$$\sigma = \sqrt{\frac{\sum(\mu_i - \bar{\mu})^2}{M}} \tag{12}$$

The Z-score, z_i is defined as:

$$z_i = (\mu_i - \bar{\mu})/\sigma \tag{13}$$

Let the standard deviation of the Z-scores be SD. Thus:

$$SD = \sqrt{\frac{\sum(z_i - \bar{z})^2}{M}} \tag{14}$$

Substituting equation 13 in equation 14:

$$SD = \sqrt{\frac{\sum(\frac{\mu_i - \bar{\mu}}{\sigma} - \bar{z})^2}{M}} \tag{15}$$

By definition, the mean of the Z-score distribution is zero, so:

$$SD = \sqrt{\frac{\sum(\frac{\mu_i - \bar{\mu}}{\sigma})^2}{M}} \tag{16}$$

Since σ is a constant:

$$SD = \sqrt{\frac{\sum(\mu_i - \bar{\mu})^2}{\sigma^2 M}} \tag{17}$$

However from equation 12:

$$\sigma^2 = \sum(\mu_i - \bar{\mu})^2/M \tag{18}$$

Now, substituting equation 18 into equation 17:

$$SD = \sqrt{\frac{\sum(\mu_i - \bar{\mu})^2 M}{\sum(\mu_i - \bar{\mu})^2 M}} \tag{19}$$

and so

$$SD = 1 \tag{20}$$

References

[1] Hwang, W. Y. K. & Foote, J. (2005). Immunogenicity of engineered antibodies. *Methods* **36**, 3–10.

[2] Morrison, S. L., Johnson, M. J., Herzenberg, L. A. & Oi, V. T. (1984). Chimeric human antibody molecules: Mouse antigen-binding domains with human constant region domains. *Proc. Natl. Acad. Sci. USA* **81**, 6851–6855.

[3] Verhoeyen, M., Milstein, C. & Winter, G. (1988). Reshaping human antibodies: Grafting an antilysozyme activity. *Science* **239**, 1534–1536.

[4] Riechmann, L., Clark, M., Waldmann, H. & Winter, G. (1988). Reshaping human antibodies for therapy. *Nature (London)* **332**, 323–327.

[5] Lorenz, H. M. (2002). Technology evaluation: Adalimumab, Abbott laboratories. *Curr. Opin. Mol. Ther.* **4**, 185–190.

[6] Schellekens, H. (2002). Immunogenicity of therapeutic proteins: Clinical implications and future prospects. *Clin. Ther.* **24**, 1720–1740.

[7] De Groot, A. S. & Scott, D. W. (2007). Immunogenicity of protein therapeutics. *Trends Immunol.* **28**, 482–490.

[8] Abhinandan, K. R. & Martin, A. C. R. (2007). Analyzing the "degree of humanness" of antibody sequences. *J. Mol. Biol.* **369**, 852–862.

[9] Johnson, G. & Wu, T. T. (2001). Kabat Database and its applications: Future directions. *Nuc. Ac. Res.* **29**, 205–206.

[10] Pelat, T., Bedouelle, H., Rees, A. R., Crennell, S. J., Lefranc, M.-P. & Thullier, P. (2008). Germline humanization of a non-human primate antibody that neutralizes the anthrax toxin, by in vitro and in silico engineering. *J. Mol. Biol.* **384**, 1400–1407.

[11] Giudicelli, V., Chaume, D. & Lefranc, M.-P. (2005). IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nuc. Ac. Res.* **33**, D256–D261.

[12] Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G. & Duroux, P. (2009). IMGT, the international ImMunoGeneTics information system. *Nuc. Ac. Res.* **37**, D1006–D1012.

[13] Schütte, M., Thullier, P., Pelat, T., Wezler, X., Rosenstock, P., Hinz, D., Kirsch, M. I., Hasenberg, M., Frank, R., Schirrmann, T., Gunzer, M., Hust, M. & Dübel, S. (2009). Identification of a putative Crf splice variant and generation of recombinant antibodies for the specific detection of Aspergillus fumigatus. *PLoS One* **4**, e6625.

[14] Pelat, T., Hust, M., Laffly, E., Condemine, F., Bottex, C., Vidal, D., Lefranc, M.-P., Dübel, S. & Thullier, P. (2007). High-affinity, human antibody-like antibody fragment (single-chain variable fragment) neutralizing the lethal factor (LF) of Bacillus anthracis by inhibiting protective antigen-LF complex formation. *Antimicrob. Agents Chemother.* **51**, 2758–2764.

[15] Laffly, E., Danjou, L., Condemine, F., Vidal, D., Drouet, E., Lefranc, M.-P., Bottex, C. & Thullier, P. (2005). Selection of a macaque Fab with framework regions like those in humans, high affinity, and ability to neutralize the protective antigen (PA) of Bacillus anthracis by binding to the segment of PA between residues 686 and 694. *Antimicrob. Agents Chemother.* **49**, 3414–3420.

[16] Chassagne, S., Laffly, E., Drouet, E., Hérodin, F., Lefranc, M.-P. & Thullier, P. (2004). A high-affinity macaque antibody Fab with human-like framework regions obtained from a small phage display immune library. *Mol. Immunol.* **41**, 539–546.

[17] Pelat, T., Hust, M., Hale, M., Lefranc, M.-P., Dübel, S. & Thullier, P. (2009). Isolation of a human-like antibody fragment (scFv) that neutralizes ricin biological activity. *BMC Biotechnol.* **9**, 60.

[18] Pelat, T. & Thullier, P. (2009). Non-human primate immune libraries combined with germline humanization: an (almost) new, and powerful approach for the isolation of therapeutic antibodies. *mAbs* **4**, 377–381. http://www.landesbioscience.com/journals/mabs/article/8635.

[19] Martin, A. C. (1996). Accessing the Kabat antibody sequence database by computer. *Proteins: Struct., Funct., Genet.* **25**, 130–133.

[20] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

[21] Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.