Germline V_H/V_L Pairing in Antibodies

Narayan Jayaram, Pallab Bhowmick and Andrew C.R. Martin*

Institute of Structural and Molecular Biology,
Division of Biosciences,
University College London,
Darwin Building,
Gower Street,
London WC1E 6BT.

27th April, 2012 – Updated 11th June 2012

Abstract

Antibodies are key molecules of the adaptive immune response and are now a major class of biopharmaceuticals. Pairing of heavy and light chains is one of the ways of generating antibody diversity and, while little is known about mechanisms governing V_H/V_L pairing, previous studies have suggested that the germline source from which chains are paired is random. By selecting paired antibody protein sequences from human and mouse antibodies from the KabatMan database and mapping them onto their corresponding germline sequences, we find that pairing preferences do exist in the germline, but only for a small proportion of germline gene segments; others are much more promiscuous showing no preferences. The closest equivalent human and mouse gene families were identified and pairing preferences compared. This work may impact on the ability to generate more stable antibodies for use as biopharmaceuticals.

Keywords: immunoglobulins; immunology; preferred pairing; antibody stability; humanization

* Corresponding author. EMail: andrew@bioinf.org.uk. Tel.: 0207 679 7034

1 Introduction

Antibodies are amongst the most important classes of proteins involved in the adaptive immune system. Together with T-cell receptors, they provide a robust system of defense against infection caused by foreign bodies (Berek and Milstein, 1987). Antibodies are capable of binding to a virtually infinite set of antigens, generally with high specificity and affinity. Recently there has been a huge resurgence of interest in using antibodies in the treatment of human disease. 206 antibodies underwent clinical trials between 1980 and 2005 (Reichert and Valge-Archer, 2007) and it is estimated that more than 400 monoclonal antibodies are currently in clinical trials, with almost a third of all drugs in development being monoclonal antibodies (Abhinandan and Martin, 2008; Reichert and Valge-Archer, 2007).

In order for the immune system to recognize and act against the enormous variety of pathogenic organisms, antibodies must be capable of recognizing a virtually infinite array of antigens. Antibody diversity is achieved by (i) the V region genes undergoing V(D)J recombination; (ii) the variable portions of the heavy and light chains pairing to form a domain dimer (the variable fragment, or Fv region); and (iii) somatic hypermutation occurring in order for the expressed antibody to be optimized towards a particular antigen (Maizels, 2005). The contribution of residues in the framework regions to interactions with the antigen remains poorly understood. It has been demonstrated that modification of residues distant from the antigen combining site of the antibody can have a significant effect on the binding affinity for the antigen (Chatellier *et al.*, 1996; Roguska *et al*., 1996; Adair *et al.*, 1999). For example, Adair and co-workers have demonstrated that modification of residue H23 could significantly affect binding (Adair *et al.*, 1999).

Therapeutic antibodies are used for a variety of conditions such as cancer (Larson *et al.*, 2002), transplant rejection (Berard *et al.*, 1999), rheumatoid arthritis (Maini *et al.*, 1999), antiviral prophylaxis (Saez-Llorens *et al.*, 1998) and Crohn's disease (Sandborn and Hanauer, 1999). A successful therapeutic antibody must be non-immunogenic, available in a high enough yield for the desired response to occur and must have a good binding affinity for the target antigen. In addition, it must be stable to avoid denaturation and aggregation, not only for long shelf-life and persistent bio-availability, but also because these factors can increase immunogenicity (Wang *et al.*, 2007; Mackay *et al.*, 2000). Thus, stability has a large influence on the efficacy of biotherapeutics (Brekke and Sandlie, 2003).

Interactions between the light and the heavy chain contribute significantly to the stability of the Fv. The V_H/V_L interface between the light chain and heavy chain has been shown to affect the binding kinetics of a peptide (Chatellier *et al.*, 1996) suggesting preference for particular pairings. Packing of the V_H and V_L domains was analyzed in detail by Chothia *et al.* (1985) and more recently by Abhinandan and Martin (2010) and by Narayanan *et al.* (2009).

Unfortunately, several difficulties exist in obtaining human monoclonal antibodies through traditional hybridoma technology. These included unstable human hybridomas, the low production of monoclonal antibodies, and the ethical and practical difficulties associated with using humans who have been immunised against a target antigen (Green, 1999; Winter and Milstein, 1991) as well as the fact that many therapeutic targets are human proteins that will not lead to the production of antibodies. Murine monoclonal antibodies, derived using the mouse hybridoma method (Kohler and Milstein, 1975), that are used therapeutically in humans, are likely to result in an immune response (the Human Anti Mouse Antibody or HAMA response) (Schroff *et al.*, 1985). However murine (or other non-human) antibodies can be engineered to make them appear 'more human'. Chimeric

antibodies are comprised of human constant regions and mouse variable regions (Morrison *et al.*, 1984). Humanization further reduces the immunogenicity by using human constant and variable regions into which mouse CDRs are inserted (Jones *et al.*, 1986; Verhoeyen *et al.*, 1988). However, in order to restore the binding affinity, some framework residues need to be converted to the equivalent mouse residue (Riechmann *et al.*, 1988). The Adair patent (Adair *et al.*, 1999) includes V_H/V_L interface residues as one of the classes of residues which may need to match their murine counterparts suggesting the importance of the paring of light and heavy chains in defining antibody affinity and stability. Finally, an alternative method of humanization called 'resurfacing' has been proposed by Roguska *et al.* (1994). Starting with a chimeric antibody, solvent accessible residues in the framework regions are replaced with human residues in an attempt to remove B-cell epitopes. A more recent development is the production of fully human antibodies from phage display libraries (McCafferty *et al.*, 1990; Burton, *et al*., 1991; Marks, *et al.*, 1991) or transgenic mice (Green, 1999).

In a given cell, the V_H/V_L pairing is unique. A significant number of the available heavy and light chain germline gene segments are used in V_H/V_L pairing and a given V_H sequence can pair with many light chain sequences of both lambda and kappa light chain classes (Edwards *et al.*, 2003). Previous work failed to reveal any evidence for preferences in pairing of particular V_H and V_L gene families and it was concluded that pairing of heavy and light chains occurred at random: Brezinschek *et al.* (1998) looked at V_H and kappa chain sequences obtained from 144 individual human (CD19+/IgM+) B cells using single chain PCR, while de Wildt *et al.* (1999) used 365 human IgG+ B cells from peripheral blood using PCR amplification of cDNA. However these authors stated that their studies were limited by the amount of data available and they could not rule out the possibility that preferences in pairing could exist.

In this paper we re-examine this problem using larger datasets of paired light and heavy chains from the Kabat dataset (Johnson and Wu, 2001) using the KabatMan database (Martin, 1996) and find that, contrary to earlier work, pairing preferences do occur. In particular, using 545 human antibodies we found that the human heavy-1 (hHV1) family shows a strong preference for Kappa-3 (hKV3). Mouse sequences from 1456 antibodies show a larger number of strong preferences: heavy-2 (mHV2) for kappa-4 (mKV4) at the expense of kappa-3 (mKV3); heavy-5 (mHV5) for kappa-2 (hKV2); heavy-6 (mHV6) for kappa-11 (mKV11); heavy-7 (mHV7) for kappa-7 (mKV7) and kappa-8 (mKV8) at the expense of kappa-10 (mKV10); heavy-8 (mHV8) for kappa-13 (mKV13); and heavy-11 (mHV11) for kappa-14 (mKV14). Some of these preferences are driven by over-representation of certain antigens in the Kabat database, but others (mHV6/mKV11, mHV7/mKV8, mHV11/mKV14) are not influenced in this way.

2 Materials and Methods

Variable light and heavy chain protein sequences were obtained for 'complete' human and mouse antibodies (i.e. light and heavy chains are both present and sequence data are present for at least 75 residues eliminating short gene segments and cases where the data are incomplete) from the July 2000 release of the Kabat database (Johnson and Wu, 2001) (the latest publicly available release) using Abysis (http://www.abysis.org/) and KabatMan (http://www.bioinf.org.uk/abs/kabatman.html) (Martin, 1996). Any duplicated sequences (100% identity) were filtered out and the sequences were converted to FASTA format. The database contained 545 human and 1456 mouse distinct antibodies.

Functional human germline V-gene segment sequences for heavy, lambda and kappa chains were obtained from the NCBI IGBLAST server in FASTA format (http://www.ncbi.nlm.nih.gov/igblast/showGermline.cgi). These sequences originate from

the IMGT database (Lefranc *et al.*, 2005). Similarly, functional mouse germline V-gene segment sequences for heavy, lambda and kappa chains were obtained from the VBASE2 database (Retter *et al.*, 2005) in FASTA format (http://www.vbase2.org/vbdownload.php).

BLAST databases were produced from these germline sequences and TBLASTN (Altschul *et al.*, 1990) was then used to map each antibody protein sequence to its closest germline DNA sequence by selecting the hit with the best e-value.

IMGT germline identifiers are of the form IG*ltf-g*aa* (e.g. IGKV4-1*01) where *l* denotes the locus (H, K or L – K (kappa) in this case), *t* denotes the type of gene segment (V, D, J or C – V in this case), *f* corresponds to the gene family (4 in this case), *g* corresponds to the individual gene (1 in this case) and *aa* corresponds to the allele (01 in this case) (Barbie and Lefranc, 1998). The allele information was omitted such that sequences corresponding to different alleles of a given gene were all considered as corresponding to that gene. This approach was adapted from the method used by Thullier *et al.* (2010).

Having identified the most likely parent germline gene segment for each complete antibody, the frequencies with which each light and heavy chain germline gene segment is seen to be paired could be counted. The raw pairing data at the individual gene level are provided in Supplementary Data files S1 and S4[1]. The number of zero-cells means that statistical tests are not practical without grouping the data, the assumptions of the χ^2 test (Bland, 2000, Section 13.1) being that no more than 20% of the expected values are below five and that no single expected value is below one (Dytham, 2011).

Consequently another version of the dataset was prepared by omitting the gene number such that the data were considered at the gene family level. Pairing counts at the gene family level are provided in Tables I and II with expected counts shown in Supplementary Data files S2 and S5. Both human and mouse datasets still had >20% of expected values below 5 and also had several expected values less than one (indicated by parenthesis in Tables I and II). Further grouping was performed in order to perform an overall χ^2 test as detailed in the legends to Tables I and II and in Supplementary Data files S3 and S6.

The significance of individual pairings being either favoured or disfavoured was calculated by creating a 2x2 contingency table:

x, y	¬x, y
¬y, x	¬x, ¬y

For example, considering the pairing IGHV1/IGKV1, we would have 4 cells containing IGHV1/IGKV1, IGHV1/not-IGKV1, not-IGHV1/IGKV1, not-IGHV1/not-IGKV1. Significance for these was evaluated using a Fisher Exact test (Bland, 2000, Section 13.4). A Bonferroni correction for multiple testing (Bland, 2000, Section 9.10) was not performed as this is often considered over-conservative (Perneger, 1998). However, marginally significant results should be treated with caution.

The closest mouse equivalent to each human germline gene segment was identified using BLASTN searches (Supplementary Data file S7). The human and mouse germline sequences were separated into heavy, kappa and lambda sequences. Each human heavy chain sequence (respectively, kappa and lambda) was searched against the database of mouse heavy chain germline sequence (respectively, kappa and lambda).

Frequencies of occurrence of each human-to-mouse mapping were calculated. In cases where members of a particular human family mapped onto more than one mouse gene family, the highest frequency mapping was assumed to be correct. These steps were also performed in the opposite direction (i.e. each mouse sequence was searched against the

[1] Supplementary data files are available at http://www.bioinf.org.uk/chainpairing/

corresponding database of human sequences) to ensure all mouse germline families were mapped to a human family.

Over-represented antigens were identified as follows. The observed number of antibodies binding a given antigen (O_A) was calculated together with the total number of antibodies with any known antigen (N_a) and the number of distinct antigens (N_g). If evenly distributed, the expected number of antibodies for every antigen is $E=N_a/N_g$. For an antigen, A, we have an observed count (O_A) and expected count (E). We also have the observed count for all other antigens ($O_{\neg A} = N_a-O_A$) and the expected count for all other antigens ($E_{\neg A} = N_a-E$). We can now calculate a χ^2 value with 1 degree of freedom. To apply the Bonferroni correction, we divide the normal alpha value for significance (0.05) by the number of distinct antigens (N_g).

χ^2 and Fisher's Exact tests were implemented in C and all other custom code was written in Perl.

3 Results and Discussion

Tables I and II show the frequencies of the pairings of the human and mouse variable heavy chain and variable light chain germline families. The overall χ^2 value for the human data shown in Table I (after grouping) was 38.33 with 15 degrees of freedom, giving $p < 8.1\text{x}10^{-4}$ while, for the mouse data, χ^2 was 321.72 with 54 degrees of freedom ($p \approx 0.0$). This clearly shows that, contrary to previous analyses (Brezinschek *et al.*, 1998; de Wildt *et al.*, 1999), light and heavy chain pairing does not occur at random.

In order to find out whether this was a small effect spread across all pairings, or whether specific pairings were statistically significant, 2x2 contingency tables were constructed and Fisher Exact tests were performed as described in the Materials and Methods. The counts in Tables I and II are annotated with '+' signs or '-' signs to indicate significant over- or under-representation.

In the case of human germlines, the kappa locus is divided into normal and distal sub-loci. For example, the Kappa-1 (hKV1) gene also occurs in the distal locus (hKV1D). The analysis of individual gene family pairing was repeated grouping the normal and distal loci gene families as shown in Table III. The overall significance remains the same as for the data in Table I where normal and distal families were also grouped for the purpose of calculating the overall significance.

Tables I, II and III illustrate that only certain gene families show any pairing preference. For humans, 5 of 7 heavy chain families (4 of 7 if distal loci are treated with primary loci) and 5 of 15 light chain families (5 of 12 if distal loci are treated with primary loci) show some preference (up or down). For mice, 13 of 14 heavy chain families and 15 of 21 light chain families show some preference. If a Bonferroni correction is made (dividing α – normally 0.05 – by the number of tests made – i.e. the number of cells in the tables), then for human sequences only one heavy family (hHV1) and one light chain family (hKV3 – hKV3(D) if distal loci are treated with primary loci) remain significant.

In terms of individual pairings, 9.5% of human heavy/light chain family pairs show some significant preference (10.7% if distal loci are treated with primary loci) while 14.6% of mouse chain family pairs show some significant preference. If a Bonferroni correction is made, this falls to a single significantly preferred pairing (0.95%, or 1.2% of possible pairings if distal loci are treated with primary loci) in human antibodies: hHV1 with hKV3. In the case of mouse antibodies, eight (2.7%) of the pairings remain significant after a Bonferroni correction: mHV5/mKV2, mHV2/mKV4, mHV7/mKV7, mHV7/mKV8, mHV6/mKV11, mHV8/mKV13, mHV11/mKV14 and mHV3/mKV4.

All of the statistically valid preferences that remain after a Bonferroni correction in humans

are positive preferences while in mouse, all but one of the remaining preferences (mHV3/mKV4) is positive. This clearly suggests that pairing preferences are driven by positive selection (i.e. preferred pairs have some useful property such as enhanced stability) rather than their being any negative selection (i.e. particular pairs being somehow incompatible).

Where there was a statistical preference in the mouse pairings, the equivalent human pairing (Supplementary Data file S7) was examined (see Tables IV and V). In particular we were looking for cases where a particular mouse pairing was favoured, but the equivalent human pairing was disfavoured. Such cases would give concern if, for example, humanization protocols selected a disfavoured pair of human acceptor frameworks.

Both mHV1/mKV10 and mHV9/mKV10 were significantly favoured pairs in mice while the human equivalents (hHV1/hKV1D in both cases – hHV1/hKV1(D) if distal gene segments are treated with the normal versions) are significantly disfavoured. However, none of these pairings (in mouse or human) is significant if a Bonferroni correction is made. Conversely, mHV2/mKV10, mHV3/mKV10 and mHV3/mKV4 are all disfavoured in mouse, but the human equivalent (hHV4/hKV1D in all cases) is favoured in humans. The human equivalent if distal gene segments are treated with the normal version (hHV4/hKV1(D)) is not significantly over-represented and again none of the pairings is significant if a Bonferroni correction is made.

It is possible that some of the pairing preferences are the result of bias in the Kabat dataset. In particular, it would be expected that any bias in the antigens against which antibodies are present might lead to the selection of particular germlines since early-response IgM and IgD antibodies are close in sequence to germline. Consequently, we examined the distribution of antigens to identify those that are statistically over-represented, with and without a Bonferroni correction. Antigen names for 'complete' antibodies were extracted using KabatMan. (See Supplementary Data Files S8 and S9 for human and mouse results respectively). For all significantly over-represented antigens, the paired V-gene segments were identified (Supplementary Data Files S10, S11 and S12). All significant germline pairings from Tables I, II and III were then examined to find out whether the over-representation of each pair resulted from over-represented antigens.

Results are given in Table VI. In summary, for human germline sequences, none of the statistically preferred germline pairings can be explained by over-representation of antigens, except possibly the hHV1/hKV3(D) pairing (when distal and proximal gene families are grouped). Indeed, the hHV3/hKV3 pairing which is strongly favoured by the over-represented anti-HIV-1-GP120 antibodies (11 of 31 antibodies use this pairing) is statistically *under*-represented overall. For mouse antibodies, the story is somewhat different. Some of the statistically preferred germline pairings are dominated by certain over-represented antigens (mHV2/mKV4 by 2-phenyl-oxazolone, mHV3/mKV14 by Musk-odorant [traseolide-(6-acetyl-1-isopropyl-2,3,3,5-tetramethyl-indane)], mHV4/mKV4 by β-1,6-D-galactan, mHV5/mKV2 by influenza virus hemagglutinin, mHV7/mKV7 by phosphorylcholine, mHV8/mKV13 by human interferon-γ receptor, and mHV12/mKV4 by phosphatidyl-choline). Of particular note is mHV8/mKV13, where all seven of the observed pairings come from anti-human-interferon-γ-receptor antibodies. In the case of mHV1/mKV3 and mHV1/mKV5, three and two (respectively) completely different over-represented antigens contribute to this preference (see Table VI), so it is difficult to say that it is the over-representation of the antigens that is contributing to this preference, rather than this preference being observed in several over-represented antigens. Similarly, while most of the observed mHV10/mKV1 pairings (6 of 8) are a result of anti-DNA antibodies, this only represents 6 of the 117 anti-DNA antibodies, so it is hard to suggest that it is the over-representation of anti-DNA antibodies that is responsible for this germline pairing preference. It is also interesting to note that mHV1/mKV5 is seen in 6 of 21 anti-dsDNA

antibodies and in 6 of 10 anti-cardiolipin antibodies; it is known that antibodies for these antigens can cross-react (Collis *et al.*, 2003). None of the other statistically preferred pairings is unduly influenced by the over-represented antigens.

In addition, we looked for any correlation between the germline pairing and the V_H/V_L packing angle (Abhinandan and Martin, 2010), but found no such correlation (data not shown).

In summary, χ^2 tests performed across the full sets of data show very clearly that there are significant pairing preferences. Analysis of individual pairings using Fisher Exact tests shows that a number of individual pairings are significant. When the very strict Bonferroni correction is made for multiple testing, a number of pairings still show significant preferences, but all of the human and all but one of the mouse pairings (mHV3/mKV4) are positive preferences rather than disfavoured pairings. However, in no case in our analysis is a favoured pairing in mouse (after Bonferroni correction) disfavoured in the equivalent human pairing.

These significant preferences are in disagreement with the conclusions of previous studies by Brezinschek *et al.* (1998) and de Wildt *et al.* (1999) which concluded that pairing occurs at random. Brezinschek *et al.* (1998) point out that pairing preferences could place a restriction on the diverse array of antibodies that are expressed by the mature B cells, but our results show only one significantly disfavoured pairing after Bonferroni correction (in mice) suggesting that the full range of pairings is possible. Thus it appears that pairing preferences have a negligible effect on antibody diversity.

The reason for these pairing preferences is currently unclear. We have shown that some of the preferences in mouse antibodies are the result of bias in the antigens to which the antibodies in the Kabat database bind, but this is not the case for many of the mouse pairings, or for any of the human pairings. Thus, it seems more likely that preferences – almost all of which are positive – result from some innate properties of these antibodies such as high stability of the V_H/V_L interface. Anecdotal evidence does suggest that certain over-represented pairings do result in more stable antibodies. Therefore the results of this analysis could be used as part of a pipeline to select antibodies that are likely to be stable. Such antibodies are more likely to be effective in the clinic, having a longer shelf life and being less susceptible to denaturation and aggregation, thus lowering immunogenicity and helping to cross one of the hurdles to regulatory approval (Reichert and Valge-Archer, 2007). In those instances where over-represented antigens do show a preference for particular germline pairs, antibody engineers may find it useful to exploit these preferences when generating antibodies against related antigens.

References

Abhinandan, K. R. and Martin, A. C. R. (2008). *Mol. Immunol*., **45**, 3832-3839.

Abhinandan, K. R. and Martin, A. C. R. (2010). *Protein Eng. Des. Sel*., **23**, 689-697.

Adair, J. R., Athwal, D. S. and Emtage, J. S., (1999). *Humanised antibodies*. US patent 5,859,205.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). *J. Mol. Biol.,* **215**, 403-410.

Barbie, V. and Lefranc, M. P. (1998). *Exp. and Clin. Immunogenet.,* **15**, 171-183.

Berard, J. L., Velez, R. L., Freeman, R. B. and Tsunoda, S. M. (1999). *Pharmacotherapy*, **19**, 1127-1137.

Berek, C. and Milstein, C. (1987). *Immunol. Rev.,* **96**, 23-41.

Brekke, O. H. and Sandlie, I. (2003). *Nat. Rev. Drug Discov.*, **2**, 52-62.

Brezinschek, H. P., Foster, S. J., Dorner, T., Brezinschek, R. I. and Lipsky, P. E. (1998). *J. Immunol.*, **160**, 4762-4767.

Bland, M. J. (2000). *An Introduction to Medical Statistics,* 3rd Edition. Oxford Medical Publications.

Burton, D. R., Barbas, C. F. 3rd, Persson, M. A., Koenig, S., Chanock, R. M. and Lerner, R. A. (1991) *Proc. Natl. Acad. Sci., U.S.A*., **88**, 10134-10137.

Chatellier, J., Van Regenmortel, M. H., Vernet, T. and Altschuh, D. (1996). *J. Mol. Biol.,* **264**, 1-6.

Chothia, C., Novotny, J., Bruccoleri, R. and Karplus, M. (1985). *J. Mol. Biol.*, **186**, 651-663.

Collis, A. V. J., Brouwer, A. R. and Martin, A. C. R. (2003). *J. Mol. Biol.*, **325**, 337-354.

de Wildt, R. M. T., Hoet, R. M. A., van Venrooij, W. J., Tomlinson, I. M. and Winter, G. (1999). *J. Mol. Biol.,* **285**, 895-901.

Dytham, C., (2011). *Choosing and using statistics: a biologist's guide*. Wiley-Blackwell.

Edwards, B. M., Barash, S. C., Main, S. H., Choi, G. H., Minter, R., Ullrich, S., Williams, E., Du Fou, L., Wilton, J., Albert, V. R., Ruben, S. M. and Vaughan, T. J. (2003). *J. Mol. Biol.,* **334**, 103-118.

Green, L. L. (1999). *J. Immunol. Meth.*, **231**, 11-23.

Johnson, G. and Wu, T. T. (2001). *Nucleic Acids Res.*, **29**, 205-206.

Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S. and Winter, G. (1986). *Nature (London),* **321**, 522-525.

Kohler, G. and Milstein, C. (1975). *Nature (London),* **256**, 495-497.

Larson, R. A., Boogaerts, M., Estey, E., Karanes, C., Stadtmauer, E. A., Sievers, E. L., Mineur, P., Bennett, J. M., Berger, M. S., Eten, C. B., Munteanu, M., Loken, M. R., van Dongen, J. J. M., Bernstein, I. D. and Appelbaum, F. R. (2002). *Leukemia*, **16**, 1627-1636.

Lefranc, M.-P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clement, O., Chaume, D. and Lefranc, G. (2005). *Nucleic Acids Res*., **33**, D593-D597.

Mackay, I. R., Rosen, F. S., Delves, P. J. and Roitt, I. M. (2000). *New Eng. J. Med.,* **343**, 37-49.

Maini, R., St Clair, E. W., Breedveld, F., Furst, D., Kalden, J., Weisman, M., Smolen, J., Emery, P., Harriman, G., Feldmann, M. and Lipsky, P. (1999). *Lancet*, **354**, 1932-1939.

Maizels, N. (2005). *Annu. Rev. Genet.*, **39**, 23-46.

Marks, J. D., Hoogenboom, H. R., Bonnert, T. P., McCafferty, J., Griffiths, A. D., and Winter, G. (1991) *J. Mol. Biol*., **222**, 581-597.

Martin, A. C. R. (1996). *Proteins: Struct., Funct., Genet.*, **25**, 130-133.

McCafferty, J., Griffiths, A. D., Winter, G. and Chiswell, D. J. (1990). *Nature (London),* **348**, 552-554.

Morrison, S. L., Johnson, M. J., Herzenberg, L. A. and Oi, V. T. (1984). *Proc. Natl. Acad. Sci. USA*, **81**, 6851-6855.

Narayanan, A., Sellers, B. D. and Jacobson, M. P. (2009). *J. Mol. Biol.*, **388**, 941-953.

Perneger, T. V. (1998). *Brit. Med. J.*, **316**, 1236-1238.

Reichert, J. M. and Valge-Archer, V. E. (2007). *Nat. Rev. Drug Discov.*, **6**, 349-356.

Retter, I., Althaus, H. H., Munch, R. and Muller, W. (2005). *Nucleic Acids Res.*, **33**, D671-D674.

Riechmann, L., Clark, M., Waldmann, H. and Winter, G. (1988). *Nature (London),* **332**, 323-327.

Roguska, M. A., Pedersen, J. T., Henry, A. H., Searle, S. J., Roja, C. M., Avery, B., Hoffee, M., Cook, S., Lambert, J. M., Blättler, W. A., Rees, A. R. and Guild, B. C. (1996). *Protein Eng*., **9**, 895-904.

Roguska, M. A., Pedersen, J. T., Keddy, C. A., Henry, A. H., Searle, S. J., Lambert, J. M., Goldmacher, V. S., Blättler, W. A., Rees, A. R. and Guild, B. C. (1994). *Proc. Natl. Acad. Sci. USA*, **91**, 969-973.

Sáez-Llorens, X., Castaño, E., Null, D., Steichen, J., Sanchez, P. J., Ramilo, O., Top, F. H. and Connor, E. (1998). *Pediat. Infect. Dis. J.*, **17**, 787-791.

Sandborn, W. J. and Hanauer, S. B. (1999). *Inflam. Bowel Dis.,* **5**, 119-133.

Schroff, R. W., Foon, K. A., Beatty, S. M., Oldham, R. K. and Morgan, A. C. (1985). *Cancer Res.*, **45**, 879-885.

Thullier, P., Huish, O., Pelat, T. and Martin, A. C. R. (2010). *J. Mol. Biol.*, **396**, 1439-1450.

Verhoeyen, M., Milstein, C. and Winter, G. (1988). *Science*, **239**, 1534-1536.

Wang, W., Singh, S., Zeng, D. L., King, K. and Nema, S. (2007). *J. Pharmaceut. Sci.*, **96**, 1-26.

Winter, G. and Milstein, C. (1991). *Nature (London)*, **349**, 293.

Table I: Pairing frequencies of human germline families

	hKV1	**hKV1D**	**hKV2**	**hKV2D**	**hKV3**	**hKV3D**	**hKV4**	**hLV1**	**hLV2**	**hLV3**	**hLV4**	**hLV5**	**hLV6**	**hLV7**	**hLV8**
hHV1	20	14-	4	4	55+++	3	13	7-	11	8-	0	(0)	(0)	(1)	(1)
hHV2	1	5+	(0)	(0)	1	(0)	(0)	(0)	(2)	1	(0)	(0)	(0)	(0)	(0)
hHV3	44	36	8	8	39-	16	15	26	26	27	4	(0)	(2)	(0)	2
hHV4	9	23+	0	0	12	4	2	14+	4	17+	(0)	(1)	(0)	(1)	(0)
hHV5	0	4	(0)	(0)	3	(2)	1	3	3	3	(0)	(0)	(0)	(0)	(0)
hHV6	1	3	(0)	(0)	1	(2)	(0)	(1)	(1)	(0)	(0)	(0)	(0)	(0)	(0)
hHV7	(1)	(1)	(2)+	(0)	0	(0)	(1)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)

Overall χ^2 (after grouping): 321.72 with 54 degrees of freedom ($p \approx 0.0$). To avoid expected counts <1 and to ensure <20% of expecteds were <5, the following grouping was necessary: IGKV1, IGKV1D, IGKV2 and IGKV2D were grouped; IGKV3 and IGKV3D were grouped; IGHV1 and IGHV2 were grouped; IGHV5, IGHV6 and IGHV7 were grouped; IGLV3, IGLV4, IGLV5, IGLV6, IGLV7 and IGLV8 were grouped. ( ) expected value below 1; Significantly up: + $p<0.05$, ++ $p<1x10^{-3}$, +++ $p<1x10^{-5}$; Significantly down: - $p<0.05$, -- $p<1x10^{-3}$, --- $p<1x10^{-5}$; Percentage of expecteds <5 = 77% Percentage of expecteds <5 (after grouping) = 16.7%

Table II: Pairing frequencies of mouse germline families

	mKV1	mKV2	mKV3	mKV4	mKV5	mKV6	mKV7	mKV8	mKV9	mKV10	mKV11	mKV12	mKV13	mKV14	mKV15	mKV16	mKV17	mKV19	mLV1	mLV2	mLV3
mHV1	93-	15	70+	130	26+	34	0--	57	12	55+	4	22	0-	23	12	4	2	11	30	0	(0)
mHV2	25	0-	3--	65+++	1	8	0	10	1	3-	0	16+	(0)	4	3	1	(0)	2	6	(1)	(0)
mHV3	20	0	14	7--	7+	8	0	4	2	1-	(0)	12+	(0)	10+	0	(3)	(0)	0	6	(2)+	(0)
mHV4	4	4	2	20+	0	1	(0)	0-	(1)	9+	(0)	1	(0)	0	(0)	(0)	(0)	(0)	5+	(0)	(0)
mHV5	40	19+++	25	27-	5	16	0	22	3	14	0	5	1	5	0	0	(1)	3	7	(0)	(1)
mHV6	11+	(2)	0	3	(0)	1	(0)	1	(0)	0	(8)+++	0	(1)	0	(0)	(0)	(0)	(0)	1	(0)	(0)
mHV7	27	3	9	12-	0	3	15+++	30+++	0	0--	(0)	0-	(0)	0-	1	1	(0)	2	1	(1)	(0)
mHV8	3	0	0-	22++	0	4	(0)	1	(0)	0	(0)	2	(7)+++	2	(0)	(0)	(0)	(0)	1	(0)	(0)
mHV9	13	1	1	6	1	7+	(0)	0-	(1)	9+	(0)	2	(0)	0	(0)	(2)	(0)	(0)	0	(0)	(0)
mHV10	8+	(0)	1	1	(0)	(0)	(0)	3	(0)	2	(0)	(0)	(0)	(0)	(0)	(1)	(0)	(0)	(0)	(0)	(0)
mHV11	0	(0)	0	0-	(0)	0	(0)	0	(0)	0	(0)	(0)	(0)	(15)+++	(0)	(0)	(0)	(0)	(1)	(1)	(0)
mHV12	0	(0)	(0)	7++	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
mHV13	(0)	(0)	(1)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
mHV14	19+	0	6	11	2	4	(0)	2	(2)	0-	(0)	6	(0)	3	(2)	(2)	(0)	(2)	1	(0)	(0)

Overall χ^2 (after grouping): 38.33 with 15 degrees of freedom ($p < 8.1\text{x}10^{-4}$). To avoid expected counts <1 and to ensure <20% of expecteds were <5, the following grouping was necessary: IGKV5, IGKV6, IGKV7, IGKV8 and IGKV9 were grouped; IGKV2 and IGKV3 were grouped; IGKV11, IGKV12, IGKV13, IGKV14, IGKV15, IGKV16, IGKV17, IGKV18 and IGKV19 were grouped; IGLV1, IGLV2 and IGLV3 were grouped; IGHV10, IGHV11, IGHV12, IGHV13 and IGHV14 were grouped. ( ) expected value below 1; Significantly up: + $p<0.05$, ++ $p<1\text{x}10^{-3}$, +++ $p<1\text{x}10^{-5}$; Significantly down: - $p<0.05$, -- $p<1\text{x}10^{-3}$, --- $p<1\text{x}10^{-5}$; Percentage of expecteds <5 = 79.2% Percentage of expecteds <5 (after grouping) = 17.1%

Table III: Pairing frequencies of human germline families, grouping normal and distal variants

	hKV1(D)	hKV2(D)	hKV3(D)	hKV4	hLV1	hLV2	hLV3	hLV4	hLV5	hLV6	hLV7	hLV8
hHV1	34-	8	58+++	13	7-	11	8-	0	(0)	(0)	(1)	(1)
hHV2	6	(0)	1	(0)	(0)	(2)	1	(0)	(0)	(0)	(0)	(0)
hHV3	80	16	55-	15	26	26	27	4	(0)	(2)	(0)	2
hHV4	32	0-	16	2	14+	4	17+	(0)	(1)	(0)	(1)	(0)
hHV5	4	(0)	5	1	3	3	3	(0)	(0)	(0)	(0)	(0)
hHV6	4	(0)	3	(0)	(1)	(1)	(0)	(0)	(0)	(0)	(0)	(0)
hHV7	2	(2)+	0	(1)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)

Overall χ^2 (after grouping): 321.72 with 54 degrees of freedom ($p \approx 0.0$). To avoid expected counts <1 and to ensure <20% of expecteds were <5, the following grouping was necessary: IGKV1(D) and IGKV2(D) were grouped; IGHV1 and IGHV2 were grouped; IGHV5, IGHV6 and IGHV7 were grouped; IGLV3, IGLV4, IGLV5, IGLV6, IGLV7 and IGLV8 were grouped. ( ) expected value below 1; Significantly up: + $p<0.05$, ++ $p<1x10^{-3}$, +++ $p<1x10^{-5}$; Significantly down: - $p<0.05$, -- $p<1x10^{-3}$, --- $p<1x10^{-5}$; Percentage of expecteds <5 = 74.0% Percentage of expecteds <5 (after grouping) = 17.4%

Table IV: Comparison of murine and human preferences

Significant murine pairing	p-value	Equivalent human pairing	p-value
mHV1/mKV1	-8.57×10^{-03}	hHV1/hKV2D	(+0.742)
mHV1/mKV10	$+1.07 \times 10^{-03}$	hHV1/hKV1D	-0.016
mHV1/mKV13	-0.013	hHV1/hKV1	(-1.000)
mHV1/mKV3	+0.013	hHV1/hKV3	$+6.95 \times 10^{-09}$ *
mHV1/mKV5	+0.011	hHV1/hKV6D	N/A
mHV1/mKV7	-2.81×10^{-04}	hHV1/hKV1D	-0.016
mHV10/mKV1	$+4.32 \times 10^{-03}$	hHV3/hKV2D	(+0.248)
mHV11/mKV14	$+9.98 \times 10^{-20}$ *	hHV3/hKV1D	(-0.197)
mHV11/mKV4	-0.019	hHV3/hKV1D	(-0.197)
mHV12/mKV4	$+2.44 \times 10^{-05}$ *	hHV4/hKV1D	+0.010
mHV14/mKV1	+0.019	hHV1/hKV2D	(+0.742)
mHV14/mKV10	-0.031	hHV1/hKV1D	-0.016
mHV2/mKV10	-0.014	hHV4/hKV1D	+0.010
mHV2/mKV12	$+1.39 \times 10^{-03}$	hHV4/hKV1	(-0.316)
mHV2/mKV2	-0.011	hHV4/hKV2D	(-0.232)
mHV2/mKV3	-2.95×10^{-04}	hHV4/hKV3	(-0.084)
mHV2/mKV4	$+4.77 \times 10^{-10}$ *	hHV4/hKV1D	+0.010
mHV3/mKV10	-0.017	hHV4/hKV1D	+0.010
mHV3/mKV12	$+1.15 \times 10^{-03}$	hHV4/hKV1	(-0.316)
mHV3/mKV14	$+7.45 \times 10^{-03}$	hHV4/hKV1D	+0.010
mHV3/mKV4	-1.05×10^{-04} *	hHV4/hKV1D	+0.010
mHV3/mKV5	+0.021	hHV4/hKV6D	N/A
mHV3/mLV2	+0.040	hHV4/hLV8	(-1.000)
mHV4/mKV10	$+2.73 \times 10^{-03}$	hHV3/hKV1D	(-0.197)
mHV4/mKV4	$+1.83 \times 10^{-03}$	hHV3/hKV1D	(-0.197)
mHV4/mKV8	-0.018	hHV3/hKV4	(-1.000)
mHV4/mLV1	+0.043	hHV3/hLV3	(-1.000)
mHV5/mKV2	$+1.17 \times 10^{-06}$ *	hHV3/hKV2D	(+0.248)
mHV5/mKV4	-2.76×10^{-03}	hHV3/hKV1D	(-0.197)
mHV6/mKV1	+0.011	hHV3/hKV2D	(+0.248)
mHV6/mKV11	$+3.91 \times 10^{-12}$ *	hHV3/hKV1D	(-0.197)
mHV7/mKV10	-8.54×10^{-04}	hHV3/hKV1D	(-0.197)
mHV7/mKV12	-0.013	hHV3/hKV1	(+0.082)
mHV7/mKV14	-0.012	hHV3/hKV1D	(-0.197)
mHV7/mKV4	-4.77×10^{-03}	hHV3/hKV1D	(-0.197)
mHV7/mKV7	$+4.68 \times 10^{-18}$ *	hHV3/hKV1D	(-0.197)
mHV7/mKV8	$+2.05 \times 10^{-09}$ *	hHV3/hKV4	(-1.000)
mHV8/mKV13	$+4.35 \times 10^{-10}$ *	hHV2/hKV1	(-1.000)
mHV8/mKV3	-0.028	hHV2/hKV3	(-0.697)
mHV8/mKV4	$+1.24 \times 10^{-05}$ *	hHV2/hKV1D	+0.014
mHV9/mKV10	$+1.41 \times 10^{-03}$	hHV1/hKV1D	-0.016
mHV9/mKV6	+0.013	hHV1/hKV4	(+0.098)
mHV9/mKV8	-0.028	hHV1/hKV4	(+0.098)

Insignificant p-values for human pairing are shown in parentheses. Pairings that remain significant after a Bonferroni correction are indicated with a '*'.

Table V: Comparison of murine and human preferences grouping distal and proximal human variants

Significant murine pairing	p-value	Equivalent human pairing	p-value
mHV1/mKV1	-8.57×10^{-03}	hHV1/hKV2(D)	(+0.653)
mHV1/mKV10	$+1.07 \times 10^{-03}$	hHV1/hKV1(D)	-0.043
mHV1/mKV13	-0.013	hHV1/hKV1(D)	-0.043
mHV1/mKV3	+0.013	hHV1/hKV3(D)	$+6.37 \times 10^{-06}$ *
mHV1/mKV5	+0.011	hHV1/hKV6(D)	N/A
mHV1/mKV7	-2.81×10^{-04}	hHV1/hKV1(D)	-0.043
mHV10/mKV1	$+4.32 \times 10^{-03}$	hHV3/hKV2(D)	(+0.227)
mHV11/mKV14	$+9.98 \times 10^{-20}$ *	hHV3/hKV1(D)	(+0.777)
mHV11/mKV4	-0.019	hHV3/hKV1(D)	(+0.777)
mHV12/mKV4	$+2.44 \times 10^{-05}$ *	hHV4/hKV1(D)	(+0.205)
mHV14/mKV1	+0.019	hHV1/hKV2(D)	(+0.653)
mHV14/mKV10	-0.031	hHV1/hKV1(D)	-0.043
mHV2/mKV10	-0.014	hHV4/hKV1(D)	(+0.205)
mHV2/mKV12	$+1.39 \times 10^{-03}$	hHV4/hKV1(D)	(+0.205)
mHV2/mKV2	-0.011	hHV4/hKV2(D)	-0.013
mHV2/mKV3	-2.95×10^{-04}	hHV4/hKV3(D)	(-0.083)
mHV2/mKV4	$+4.77 \times 10^{-10}$ *	hHV4/hKV1(D)	(+0.205)
mHV3/mKV10	-0.017	hHV4/hKV1(D)	(+0.205)
mHV3/mKV12	$+1.15 \times 10^{-03}$	hHV4/hKV1(D)	(+0.205)
mHV3/mKV14	$+7.45 \times 10^{-03}$	hHV4/hKV1(D)	(+0.205)
mHV3/mKV4	-1.05×10^{-04} *	hHV4/hKV1(D)	(+0.205)
mHV3/mKV5	+0.021	hHV4/hKV6(D)	N/A
mHV3/mLV2	+0.040	hHV4/hLV8	(-1.000)
mHV4/mKV10	$+2.73 \times 10^{-03}$	hHV3/hKV1(D)	(+0.777)
mHV4/mKV4	$+1.83 \times 10^{-03}$	hHV3/hKV1(D)	(+0.777)
mHV4/mKV8	-0.018	hHV3/hKV4	(-1.000)
mHV4/mLV1	+0.043	hHV3/hLV3	(-1.000)
mHV5/mKV2	$+1.17 \times 10^{-06}$ *	hHV3/hKV2(D)	(+0.227)
mHV5/mKV4	-2.76×10^{-03}	hHV3/hKV1(D)	(+0.777)
mHV6/mKV1	+0.011	hHV3/hKV2(D)	(+0.227)
mHV6/mKV11	$+3.91 \times 10^{-12}$ *	hHV3/hKV1(D)	(+0.777)
mHV7/mKV10	-8.54×10^{-04}	hHV3/hKV1(D)	(+0.777)
mHV7/mKV12	-0.013	hHV3/hKV1(D)	(+0.777)
mHV7/mKV14	-0.012	hHV3/hKV1(D)	(+0.777)
mHV7/mKV4	-4.77×10^{-03}	hHV3/hKV1(D)	(+0.777)
mHV7/mKV7	$+4.68 \times 10^{-18}$ *	hHV3/hKV1(D)	(+0.777)
mHV7/mKV8	$+2.05 \times 10^{-09}$ *	hHV3/hKV4	(-1.000)
mHV8/mKV13	$+4.35 \times 10^{-10}$ *	hHV2/hKV1(D)	(+0.076)
mHV8/mKV3	-0.028	hHV2/hKV3(D)	(-0.467)
mHV8/mKV4	$+1.24 \times 10^{-05}$ *	hHV2/hKV1(D)	(+0.076)
mHV9/mKV10	$+1.41 \times 10^{-03}$	hHV1/hKV1(D)	-0.043
mHV9/mKV6	+0.013	hHV1/hKV4	(+0.098)
mHV9/mKV8	-0.028	hHV1/hKV4	(+0.098)

Insignificant p-values for human pairing are shown in parentheses. Pairings that remain significant after a Bonferroni correction are indicated with a '*'.

Table VI: Statistically preferred pairings and the influence of over-represented antigens

Germline Pairing	Observed Count	Expected Count	Over-represented antigen(s) where this pairing is dominant (* may be responsible for this germline pairing preference)	Count of antibodies binding this antigen (with this germline pairing / total)
Human germline families				
hHV1/hKV3	55	29.9	CD19	(5/8)
hHV4/hKV1D	23	14.3	DNA	(3/8)
hHV4/hLV1	14	8.5	GLUTAMATE-DECARBOXYLASE-(MAJOR-ISLET-CELL-AUTOANTIGEN)	(1/7)
hHV4/hLV3	17	9.3	ERYTHROCYTE-RH(D)-ALLOANTIGEN	(4/10)
Human germline families (distal and proximal gene familes grouped)				
hHV1/hKV3(D)	58	37.2	* CD19 * HIV-1-GP120	(7/8) (6/31)
hHV4/hLV1	14	8.5	GLUTAMATE-DECARBOXYLASE-(MAJOR-ISLET-CELL-AUTOANTIGEN)	(1/7)
hHV4/hLV3	17	9.3	ERYTHROCYTE-RH(D)-ALLOANTIGEN	(4/10)
Mouse germline families				
mHV1/mKV3	70	56.3	* CD4 * HUMAN-AND-MOUSE-TYPE-II-COLLAGEN-C1-EPITOPE,-RESIDUES-316-TO-333 * MOUSE-TYPE-II-COLLAGEN * DNA-AUTOANTIBODY	(7/9) (9/11) (8/15) (3/4)
mHV1/mKV5	26	17.9	* CARDIOLIPIN * DS-DNA	(6/10) (6/21)
mHV1/mKV10	55	39.7	P-AZOPHENYLARSONATE-HYBRIDOMA	(7/8)
mHV2/mKV4	65	32.9	* 2-PHENYL-OXAZOLONE-HYBRIDOMA	(52/76)
mHV2/mKV12	16	7.0	None	
mHV3/mKV5	7	2.9	None	
mHV3/mKV12	12	4.5	None	
mHV3/mKV14	10	4.2	* MUSK-ODORANT-TRASEOLIDE-(6-ACETYL-1-ISOPROPYL-2,3,3,5-TETRAMETHYL-INDANE)	(5/7)
mHV4/mKV4	20	10.4	* BETA-1,6-D-GALACTAN-HYBRIDOMA	(8/8)
mHV4/mKV10	9	3.1	None	
mHV4/mLV1	5	2.0	None	
mHV5/mKV2	19	6.1	* INFLUENZA-VIRUS-HEMAGGLUTININ-HYBRIDOMA	(10/17)
mHV6/mKV1	11	5.2	None	
mHV6/mKV11	8	0.2	None	
mHV7/mKV7	15	1.1	* PHOSPHORYLCHOLINE-(S.PNEUMONIAE-STR.-R36A)-HYBRIDOMA	(11/13)
mHV7/mKV8	30	9.7	INFLUENZA-VIRUS-(A/PR/8/34)-HEMAGGLUTININ-SECONDARY-ANTIBODIES-(SB)-HYBRIDOMA	(3/7)
mHV8/mKV4	22	9.3	None	
mHV8/mKV13	7	0.3	* HUMAN-INTERFERON-GAMMA-RECEPTOR	(7/9)
mHV9/mKV6	7	2.6	None	
mHV9/mKV10	9	2.8	None	
mHV10/mKV1	8	3.0	* DNA	(6/117)
mHV11/mKV14	15	0.7	PHOSPHATIDYL-CHOLINE	(4/9)
mHV12/mKV4	7	0.3	PHOSPHATIDYL-CHOLINE	(5/9)