Immunogenetics (2008) 60:669–676
DOI 10.1007/s00251-008-0325-z

ORIGINAL PAPER

The reported germline repertoire of human immunoglobulin kappa chain genes is relatively complete and accurate

Andrew M. Collins · Yan Wang · Viveka Singh · Phillip Yu · Katherine J. Jackson · William A. Sewell

Received: 19 May 2008 / Accepted: 16 July 2008 / Published online: 20 August 2008

Abstract We describe a bioinformatic analysis of germline and rearranged immunoglobulin kappa chain (IGK) gene sequences, performed in order to assess the completeness and reliability of the reported IGK repertoire. In contrast to the reported heavy-chain gene repertoire, which includes many dubious sequences, only five IGK variable gene (IGKV) alleles appear to have been reported in error. There was, however, insufficient evidence to justify removing these IGKV genes from the germline repertoire. Bioinformatic analysis of apparent mismatches between reported germline genes and 1,863 expressed IGK sequences suggested the existence of two unreported IGKV polymorphisms. Genomic screening of 12 individuals led to the confirmation of both of these polymorphisms, IGKV1-16*02 and IGKV2-30*02. We also show that in contrast to the heavy chain, the IGK repertoire is dominated by sequences that use just a handful of kappa variable (IGKV) and junction (IGKJ) gene pairs. There is also little modification of IGKV and IGKJ genes by the processes of exonuclease removal and N nucleotide addition. The expressed IGK repertoire therefore lacks diversity and the junction region is particularly constrained. Remarkably, the analysis of a dataset of 435 relatively unmutated rearranged kappa genes showed that ten amino acid sequences account for almost 10% of the rearrangements, with identical sequences being derived from as many as seven independent sources. Such dominant sequences are likely to have important roles in the operation of the humoral immune response.

Keywords Immunoglobulin · Light chain · IGKV · Allelic variants · Sequencing errors · Somatic point mutation

A. M. Collins · Y. Wang · V. Singh · P. Yu · K. J. Jackson
School of Biotechnology and Biomolecular Sciences,
University of New South Wales,
Kensington, Australia

W. A. Sewell
Institute of Laboratory Medicine, St Vincent's Hospital Sydney,
Darlinghurst, NSW, Australia

W. A. Sewell
St Vincent's Clinical School, University of New South Wales,
Sydney, Australia

A. M. Collins (✉)
School of Biotechnology and Biomolecular Sciences,
University of New South Wales,
Sydney, NSW 2052, Australia
e-mail: a.collins@unsw.edu.au

Introduction

Immunoglobulin heavy- and light-chain genes rearrange early in B-cell ontogeny. The rearranged light chain is formed by the recombination of genes selected from either the variable and joining genes of the kappa locus (IGKV and IGKJ), or the variable and joining genes of the lambda locus (IGLV and IGLJ). The genes and allelic variants of these loci were a focus of early immunoglobulin sequence studies in the mid-1980s. Major studies in the early 1990s led to the conclusion that the description of the human kappa locus was essentially complete (Cox et al. 1994; Zachau 1993), and that IGKV genes show relatively little polymorphism. Nevertheless, a number of later studies reported new alleles (e.g., Atkinson et al. 1996). A study that specifically set out to identify new alleles of the IGKV2-29 and IGKV2D-29 genes reported one new IGKV2-29 allele and two new IGKV2D-29 alleles, leading the authors to suggest that many more unreported alleles

probably remained to be detected (Atkinson et al. 1996; Feeney et al. 1996).

In a major initiative in the late 1990s, all reported immunoglobulin genes and allelic variants were incorporated into the ImMunoGeneTics (IMGT) databases of germline sequences and a systematic nomenclature was developed. The IMGT database of germline human kappa chain immunoglobulin gene sequences includes 103 IGKV genes and alleles, including 37 pseudogenes, 55 functional genes and allelic variants and 11 Open Reading Frames (ORFs) (Barbie and Lefranc 1998). Functional genes are defined as such by IMGT if the coding region has an open reading frame without stop codons, and where there are no apparent defects in the splicing sites, recombination signals, or regulatory elements of the genes.

Over the last decade, the number of reported rearranged IGK genes has increased substantially, making it possible to reassess the functionality of reported alleles, and to assess the completeness of the reported repertoire. We have recently used this approach to review the reported human heavy chain gene repertoire (Lee et al. 2006, 2007; Wang et al. 2008). These studies identified both sequences that have been reported in error, and putative polymorphisms that remain unreported. In this study, we reconsider the reported IGKV and IGKJ gene repertoires by analysing features of the reported germline alleles, and by an analysis of the apparent usage of the different germline genes in a database of 1863 rearranged gene sequences.

In order to gain a better understanding of the overall diversity of kappa chains, we also analysed the processes that contribute to junctional diversity of the kappa chain repertoire during the gene rearrangement process. This diversity is introduced by deletions and additions of nucleotides at the joining IGKV and IGKJ gene ends. Analysis is presented showing relatively little addition or loss of nucleotides from the kappa chain gene ends. Analysis of the resulting junctional amino acids shows that the kappa chain repertoire includes a number of dominant sequences that have been reported from multiple independent studies.

Materials and methods

Sequence accumulation

Expressed human immunoglobulin kappa chain sequences were obtained from the European Molecular Biology Laboratory (EMBL) nucleotide sequence database (http://www.ebi.ac.uk/embl/) (Kanz et al. 2005). All sequences were derived from complementary DNA (cDNA) and had a minimum sequence length of 270 nucleotides. Not all sequences included identifiable IGKJ genes.

A database of IGKV germline genes was compiled from the IMGT IGKV gene database (http://imgt.cines.fr/) (Barbie and Lefranc 1998). The publications that reported each germline gene were identified from IMGT annotations, and each publication was reviewed to identify the polymerase chain reaction (PCR) primers that were used. The number of genomic sequences that have been reported for each reported germline gene was also determined.

The germline genes were analysed by multiple sequence alignment, using ClustalW (www.ebi.ac.uk/Tools/clustalw) (Thompson et al. 1994). Germline sequences that had only one or two nucleotide differences to other sequences were noted, as apparent allelic variants can be generated as a result of PCR errors.

Expressed kappa chain sequences were aligned against the IMGT IGKV germline repertoire using the Smith–Waterman algorithm (Smith and Waterman 1981) with the Gotoh optimisation (Gotoh 1982). Clonally elated sequences were identified by shared IGKV genes and shared somatic point mutations. No attempt was made to identify clonally related sequences amongst relatively unmutated sequences.

Partitioning

The 3′ ends of most IGKV genes are C-rich, with most genes ending with the nucleotides CCTCC. The neighbouring N-REGIONs, at the junction of the IGKV and IGKJ genes, are also often likely to be C-rich, because of the bias of N nucleotide addition towards additions of guanine and cytidine, and a tendency for the formation of homopolymer tracts (Jackson et al. 2007). There is therefore a risk that an N-REGION could be mistaken for the end of an IGKV gene, and this risk is particularly high if mutations occur near the junction. In order to accurately characterise the processing of IGKV and IGKJ gene ends, as well as N-REGIONS, kappa sequences were therefore chosen that had little likelihood of including mutations near the ends of the IGKV and IGKJ genes. The LowVMut database was compiled from 448 sequences, which had no more than three mutations in their IGKV genes, between codons 10 and 104.

These sequences were re-aligned, to determine IGKJ genes, using IMGT/VQUEST (Giudicelli et al. 2004), and after the removal of sequences that did not include rearranged IGKJ, 435 sequences remained. Where exonuclease removals meant that a sequence aligned equally well to two alleles, the sequence was ultimately assigned to the allele that was seen more often in the database of expressed kappa chain sequences.

The ends of the IGKV genes were determined by accepting just a single mutation within the nucleotides that were aligned against the final 15 nucleotides of the IGKV

germline sequence. Consecutive mismatches at the 3′ end of the IGKV genes were deemed to be N nucleotides, or part of an IGKJ sequence, and where the ends of an IGKV gene included a series of mismatches and matches, no runs of more than two matching nucleotides were deemed to have been removed from the end of any sequence. The start of the IGKJ genes was determined in a similar fashion, and the nucleotides between the ends of the IGKV and the start of the IGKJ genes were defined as the N-REGIONS.

Identification of unreported polymorphisms

The number of mutations in each rearranged sequence was noted from the output of the program, after removal of likely primer-mediated mismatches. The frequencies with which different levels of mutation were seen in alignments to each germline gene were then compared to the overall distribution of mutations in the dataset by Chi-squared test. Where the number of sequences that aligned to a particular gene/allele included an unexpectedly low number of unmutated sequences, additional analysis was performed using multiple sequence alignment. Where shared mismatches were commonly seen at a particular position within a sequence, unreported polymorphisms were inferred.

DNA isolation and amplification

In order to confirm the existence of putative alleles identified in the bioinformatic studies, genomic screening was undertaken. Buccal smears were collected from 12 volunteers, with the approval of the University of New South Wales Human Research Ethics Committee. DNA was extracted and IGKV sequences were amplified as previously described (Dahlke et al. 2006), using primers that were designed using the reported IGKV1-16*01 and IGKV2-30*01 sequences, as follows: 5′-tcccaggtaaggatggagaa-3′ (IGKV1–16 forward primer), 5′-gatggaggcatcaggagaag-3′ (IGKV1–16 reverse primer), 5′-ttgaaatatgaacaaatacatata tagcctg-3′ (IGKV2–30 forward primer) and 5′-ttcagggctg taccactgtg-3′ (IGKV2–30 reverse primer). PCR products were cloned and sequenced as previously described (Dahlke et al. 2006). Sequences were then aligned against the reported IGKV germline repertoire as described above.

Results

A database of germline IGKV and IGKJ genes was accessed from the IMGT website, and after removal of pseudogenes, 66 reportedly functional IGKV genes and ORFs remained, as well as nine reportedly functional IGKJ genes. The expression of each of the IGKV genes was then investigated by analysis of 1,863 rearranged kappa gene sequences, and the frequency of expression of the different IGKV genes are presented as Table 1. No alignments were seen for 22 reported IGKV genes. Nine of these IGKV genes have been reported from at least two independent studies, and therefore while the functionality of the genes may be questioned, the genes almost certainly exist. Five of the remaining genes are described as ORFs, and a lack of rearrangements could therefore be expected. Five sequences differ from other reported genes by at least three nucleotides, and therefore PCR errors or other PCR artefacts are unlikely to be responsible for their generation. The existence of the final 3 genes (IGKV1-5*02, IGKV3-NL2, IGKV3-NL4) may be queried, but no additional evidence could be found to suggest that these germline sequences were originally reported in error.

A further 13 germline genes were seen in fewer than five alignments. As rearranged sequences may be misaligned because of the similarity of different IGKV genes, the presence of a handful of alignments in a database of 1,863 rearranged sequences cannot be taken as unequivocal proof that reported genes are real and functional. But the existence of eight of the 13 genes has been confirmed by multiple reports of the germline sequences, while the low level of apparent mutations seen in the rearranged gene alignments gives credence to the existence of three of the remaining five genes. Only one of the final two genes (IGKV2-29*03) is sufficiently similar to another gene to suspect that it may have resulted from PCR error; however, there is insufficient evidence to be certain that this was the case.

A small number of alignments were seen to two sequences (IGKV1D-13*01 and IGKV1-37*01), which have previously been reported as ORFs (Barbie and Lefranc 1998), suggesting that, in some individuals, these may be functional genes.

The apparent absence of unmutated sequences utilising IGKV1-16*01, and the relative absence of unmutated sequences utilising IGKV2-30*01 led us to infer the likely existence of unreported alleles for these two genes. The putative IGKV1-16*02 allele includes AAG rather than AGG as codon 75. All 29 alignments to IGKV1-16*01 included this single nucleotide mismatch in codon 75, suggesting that the germline gene was originally reported in error. The putative IGKV2-30*02 allele includes CAC rather than TAC as codon 31. Of 68 alignments to IGKV2-30*01, 27 sequences included the C/T mismatch in codon 31. This suggests that the IGKV2-30*01 allele was reported accurately, but that an unreported polymorphism exists that may be present in many individuals. We carried out investigations of the putative alleles by genomic screening of buccal swabs from 12 individuals. This confirmed the existence of both IGKV1-16*02 (accession no. FM164406) and IGKV2–30*02 (accession no. FM164408).

Table 1 IGKV gene usage in a database of 1863 rearranged human kappa chain genes

	Unmutated alignments	Total alignments	Single nt difference[a]	Published reports of germline gene	Doubtful allele
IGKV1-5*01	3	8	Y (1–5*02)	1[d]	
IGKV1-5*02	–	0	Y (1–5*01)	1	Yes
IGKV1-5*03	10	110	N	1	
IGKV1-6*01	2	29	N	2	
IGKV1-8*01	2	23	N	1	
IGKV1D-8*01	–	1	N	1	
IGKV1-9*01	3	36	N	2	
IGKV1-12*01	6	44	Y (1–12*02)	2	
IGKV1D-12*01	2	13	N	2	
IGKV1-12*02/IGKV1D-12*02	–	2	Y (1–12*01)	1	
IGKV1-13*01			Y (1D–13*01)	2	
IGKV1D-13*02	2	7	Y (1D–13*01)	1	
IGKV1D-13*01 (ORF)	2	4	Y (1–13*01, 1D–13*02)	1	
IGKV1-16*01	–	29 (0)[c]	Y (1–16*02)	1	Yes
IGKV1-16*02[b]	(1)[c]	(29)[c]	Y (1–16*01)		
IGKV1D-16*01	–	6	Y (1D–16*02)	2	
IGKV1D-16*02	–	0	Y (1D–16*01)	2	
IGKV1-17*01	5	32	Y (1–17*02)	3	
IGKV1-17*02	–	1	Y (1–17*01)	1	
IGKV1D-17*01	–	1	N	2	
IGKV1-27*01	5	37	N	2	
IGKV1-33*01	15	89	N	2	
IGKV1D-33*01		0	N	2	
IGKV1-37*01 (ORF)	1	1	N	3	
IGKV1D-37*01 (ORF)		0	N	3	
IGKV1-39*01	57	298	N	2	
IGKV1D-39*01		0	N	2	
IGKV1D-42*01 (ORF)		0	N	1	
IGKV1D-43*01		0	N	1	
IGKV1-NL1	–	15	N	1	
IGKV2-24*01	4	23	N	2	
IGKV2D-24*01 (ORF)		0	N	2	
IGKV2–28*01	23	103	N	2	
IGKV2D-28*01		0	N	3	
IGKV2-29*02	4	10	Y (2–29*03)	1	
IGKV2-29*03	–	1	Y (2–29*02)	1	Yes
IGKV2D-29*01	3	25	Y (2D–29*02)	2	
IGKV2D-29*02	5	19	Y (2D–29*01)	1	
IGKV2-30*01	8	68 (41)[c]	Y (2–30*02)	2	
IGKV2-30*02[b]	(6)[c]	(27)[c]	Y (2–30*01)		
IGKV2D-30*01	2	4	N	2	
IGKV2-40*01	1	8	N	2	
IGKV2-40*02		0	N	1	
IGKV2D-40*01		0	N	2	
IGKV3-7*01 (ORF)		0	Y (3–7*03)	1	
IGKV3-7*02 (ORF)		0	N	1	
IGKV3-7*03 (ORF)		0	Y (3–7*01)	1	
IGKV3D-7*01	–	1	N	2	
IGKV3-11*01	14	119	Y (3–11*02)	3	
IGKV3-11*02	–	1	Y (3–11*01)	1	
IGKV3D-11*01	–	2	Y (3–NL2)	2	
IGKV3-15*01	15	100	N	2	
IGKV3D-15*01	1	9	N	3	
IGKV3-20*01	64	415	N	6	
IGKV3-20*02		0	N	1	

Table 1 (continued)

	Unmutated alignments	Total alignments	Single nt difference[a]	Published reports of germline gene	Doubtful allele
IGKV3D-20*01	1	4	N	3	
IGKV3-NL1		0	N	1	
IGKV3-NL2		0	Y (3D–11*01)	1	Yes
IGKV3-NL3		0	N	1	
IGKV3-NL4		0	N	1	Yes
IGKV3-NL5		0	N	1	
IGKV4-1*01	28	157	N	5	
IGKV5-2*01	–	6	N	1	
IGKV6-21*01 (ORF)	–	2	N	2	
IGKV6D-21*01 (ORF)		0	N	2	
IGKV6D-41*01 (ORF)		0	N	2	
Total	288	1863	N		

[a] The coding sequence is identical to another allele at all but a single nucleotide. The highly similar allele name is shown in brackets.
[b] Allele reported for the first time in this study.
[c] Figures in brackets show the number of sequences seen after re-alignment using a modified repertoire that included the newly identified alleles.
[d] Where more than one sequence has been reported from a single study, this is recorded as one report.

The LowVMut database, a smaller dataset of 435 relatively unmutated sequences, was used to investigate IGKJ gene/allele usage, and is available upon request. Analysis was restricted to these sequences because accurate partitioning is essential for investigation of the usage of reported IGKJ2 and IGKJ4 alleles, which only differ at the 5′ ends of the sequences. Database analysis showed that seven of the nine reported IGKJ genes could be identified in rearranged sequences, though their frequency of expression varies from less than 0.5% of sequences (IGKJ2*02) to greater than 25% of sequences (IGKJ1*01, IGKJ4*01) (Table 2). Neither the IGKJ2*03 or IGKJ4*02 alleles were seen in the VLowMut Database, though IGKJ4*02 is well established as a rare allele (Atkinson et al. 1996).

Table 2 IGKJ gene usage in a database of 435 relatively unmutated rearranged human kappa chain genes

	Unmutated alignments	Total alignments	Published reports	Doubtful allele
IGKJ1*01	29	119	1	
IGKJ2*01	14	101	1	
IGKJ2*02	2	2[a]	1[b]	
IGKJ2*03	–	0[a]	1[b]	Yes
IGKJ2*04	2	8	2[b]	
IGKJ3*01	26	49	1	
IGKJ4*01	39	121	1	
IGKJ4*02	–	0	2	
IGKJ5*01	16	35	1	
Total	128	435		

[a] A small number of sequences that had four or more 5′ nucleotides removed by exonuclease activity aligned equally well to this allele as to another more common allele.
[b] Only reported from cDNA sequences.

IGKJ2*03 has only been reported from cDNA, and its existence must be queried.

Summaries of exonuclease removals from the IGKV and IGKJ genes are provided as Fig. 1a and b, respectively. An average of 2.4 nucleotides were apparently removed from IGKV, and 1.8 nucleotides were removed from IGKJ. These are likely to be slight underestimates, for in some cases nucleotide removal will not be identified because of addition of the same nucleotide by the process of N nucleotide addition. The detectable levels of N addition are presented as Fig. 1c. On average, just 1.5 nucleotides were added, and in 46% of cases, there was no addition at all. The lack of both addition and removal of nucleotides means that the overall length of IGK genes shows little variability, as shown in Fig. 1d.

The small amount of processing of the VJ junction, as well as the strong biases in the use of certain IGKV and IGKJ genes, led to additional investigations of IGKV/IGKJ pairings in the expressed repertoire. The ten most common IGKV/IGKJ gene combinations collectively accounted for 40.9% of all alignments (Table 3). These combinations reflect the biases in expression of the IGKV and IGKJ genes, and there were no additional biases in the combinations themselves.

Amino acid sequences of the VJ junction were then analysed, and many examples of identical junctions were noted (Table 3). Not all such identical sequences can be assumed to have arisen independently, for identical sequences were seen that were derived from the same study. Most sequences, however, were independently derived. The eight sequences encoding the dominant junction CQQYGSSPRTF of the IGKV3-20*01/IGKJ4*01 pairing came from numerous sources, including two sequences from patients with

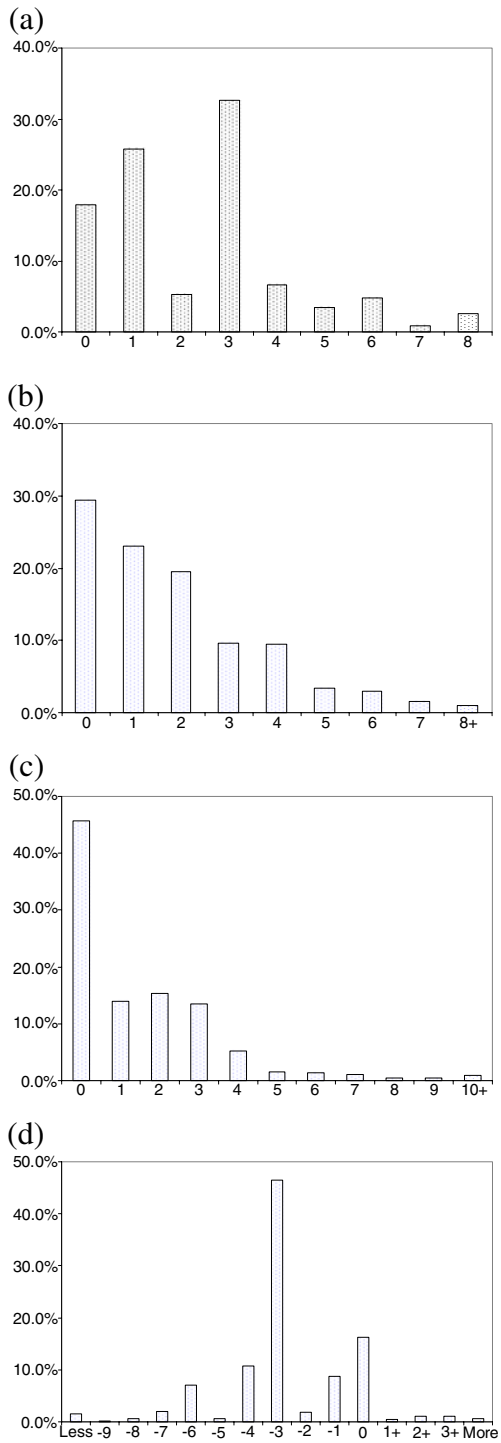

Fig. 1 Analysis of a dataset of 435 relatively unmutated rearranged human kappa chain genes, showing (**a**) exonuclease removals from the IGKV genes, (**b**) exonuclease removals from the IGKJ genes, (**c**) N nucleotide addition at the VJ junction, and (**d**) overall nucleotide gains and losses to the VJ rearrangements as a result of N addition and exonuclease removals

chronic lymphocytic leukemia (Z46310, DQ101165), one with systemic lupus erythematosus (AY422527), one with MALT lymphoma (AY281340), one with follicular lymphoma (AM040568), as well as two identical sequences (AB064107, AB064108) from an unpublished study involving an antibody gene library and one sequence (DQ187538) that is said to be derived from pneumococcal polysaccharide-specific B cells. On the other hand, the dominance of the most frequently seen junction in the study (CQQYGSSPLTF) is likely to be exaggerated by the presence of 11 identical sequences derived from a single study (Meijer et al. 2006).

Identical junctions were often produced in different ways. For example, the CQQYGSSPLTF junctions of the IGKV3-20*01/IGKJ4*01 sequences were the result of one, three or six nucleotide removals from the IGKV end, zero, one or three N additions, and zero, one or two nucleotide removals from the IGKJ end. Four different processing pathways led to the same junction sequence. Not surprisingly, very similar junctions were also seen at high frequency. For example, in addition to the seven independent instances of the junction CQQYGSSPRTF amongst IGKV3-20*01/IGKJ4*01 sequences, there were two independent instances each of CQQYGSSP[E/G/Q/W]TF.

Discussion

We have recently shown that 104 of the 226 reported heavy chain IGHV genes almost certainly contain errors and should be removed from the available repertoire (Wang et al. 2008). Many additional IGHV genes may also contain sequencing errors, but there was insufficient evidence to conclude this with confidence. This present study casts some doubt on five of the 55 functional IGKV genes, but there was insufficient evidence to firmly conclude that any of the sequences contained errors.

Most of the errors in reported IGHV genes appear to be the result of Taq polymerase-mediated PCR errors, while some errors came from the use of coding region primers and from the amplification of chimeric sequences (Wang et al. 2008). These errors led to the impression that IGHV genes are more highly polymorphic than is the case. The IGHV3-30 gene, for example, has been reported to have 19 allelic variants (Pallares et al. 1999), with 16 of these variants being identified in a single study (Olee et al. 1991). It seems likely that 14 IGHV3-30 alleles have been reported in error, but the gene is certainly polymorphic, and over half of all IGHV genes are polymorphic (Wang et al. 2008). In contrast, only 11 of the 51 IGKV genes considered here have apparent allelic variants, with no more than three variants being reported for any gene. Functional polymorphisms could initially only be confirmed in this study for IGKV1-5 and IGKV2D-29. In each case, just two functional alleles were identified. With the sequencing of the IGKV2-30*02 allele, we have identified a third gene with functional allelic variants. The locus nevertheless

Table 3 Commonly seen IGKV/IGKJ pairings in a database of 435 relatively unmutated rearranged human kappa chain genes, and the most commonly seen amino acid junction sequences that were seen in such pairings

	No.	Percent	Dominant junctions	No.	Unique instances[a]
IGKV3–20*01 IGKJ4*01	32	7.4	CQQYGSSPLTF	17	5
IGKV3–20*01 IGKJ1*01	25	5.7	CQQYGDSPRTF	8	7
IGKV1–39*01 IGKJ4*01	23	5.3	CQQSYSTPLTF	8	6
IGKV1–39*01 IGKJ1*01	20	4.6	CQQSYSTPWTF	7	5
IGKV1–39*01 IGKJ2*01	17	3.9	CQQSYSTPYTF	6	3
IGKV3–20*01 IGKJ5*01	14	3.2	CQQYGSSPITF	5	4
IGKV4–1*01 IGKJ1*01	14	3.2	CQQYYSTPWTF	3	3
IGKV2–28*01 IGKJ1*01	12	2.8	CMQALQTPRTF	4	4
IGKV4–1*01 IGKJ2*01	11	2.5	CQQYYSTPYTF	4	2
IGKV3–20*01 IGKJ2*01	10	2.3	CQQYGSSPPYTF	2	2
TOTAL	178	40.9		64	41

[a] Unique sequences were derived from independent studies, or were identical sequences reported from a single study where the sequences were subject to different exonuclease removals and N nucleotide addition.

shows little polymorphism, and it is likely that few if any further common IGKV alleles remain to be identified.

Not only is there little diversity in the germline genes of the kappa locus, but little junctional diversity is generated during the process of gene rearrangement. The small amount of processing of rearranged kappa genes has been previously noted (Tomlinson et al. 1995), and parallels reports of lambda gene rearrangements, where the overwhelming majority of sequences are subject to little exonuclease activity and where very few sequences have more than a handful of nucleotides added by TdT activity (Farner et al. 1999). This means that the length of the lambda light chain CDR3 regions is highly constrained (Farner et al. 1999), as was also observed here for the kappa chain CDR3 regions.

We extended these observations by studying the amino acid sequences of the VJ junction, and noted a number of sequences that have been repeatedly observed in independent studies. Highly similar heavy chain sequences, termed stereotypical sequences have also been reported (Stamatopoulos et al. 2007); however, these 'stereotypical' heavy and light chain sequences may arise for different reasons. The formation of the heavy chain junction region generates far more diversity than is the case for the light chain, and stereotypical heavy chain rearrangements are truly improbable events. In contrast, 'stereotypical' light chains may simply arise because of the overuse of certain IGKV and IGKJ genes, combined with the general lack of exonuclease activity and N nucleotide addition. Although biases towards pairings of downstream IGKV genes with upstream IGKJ genes, and vice versa, have been suggested by modelling studies (Mehr et al. 1999), such biases were not seen here. In fact the prominent IGKV3-20 gene was frequently associated with four of the five IGKJ genes.

Whatever the theoretical diversity of the light chain, the present study suggests that the expressed repertoire is dominated by a relatively small number of amino acid sequences. Many different sequences are each seen at a frequency of more than 1% of randomly selected sequences. The immunoglobulin gene repertoire can therefore be characterised as having extraordinary diversity in the heavy chain, and limited diversity in the light chain. In order to understand the operation of the humoral immune response, it is likely that an understanding of the role of such commonly expressed light chains will be as important as an understanding of the legendary but perhaps overstated diversity of the theoretical repertoire.

Acknowledgements This study was supported by a grant from the National Health and Medical Research Council.

References

Atkinson MJ, Cowan MJ, Feeney AJ (1996) New alleles of IGKV genes A2 and A18 suggest significant human IGKV locus polymorphism. Immunogenetics 44:115–120

Barbie V, Lefranc MP (1998) The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. Exp Clin Immunogenet 15:171–183

Cox JPL, Tomlinson IM, Winter G (1994) A directory of human germ-line Vk segments reveals a strong bias in their usage. Eur J Immunol 24:827–836

Dahlke I, Nott DJ, Ruhno J, Sewell WA, Collins AM (2006) Antigen selection in the IgE response of allergic and non-allergic individuals. J Allergy Clin Immunol 117:1477–1483

Farner NL, Dorner T, Lipsky PE (1999) Molecular mechanisms and selection influence the generation of the human V lambda J lambda repertoire. J Immunol 162:2137–2145

Feeney AJ, Atkinson MJ, Cowan MJ, Escuro G, Lugo G (1996) A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to haemophilus influenzae type b disease. J Clin Invest 97:2277–2282

Giudicelli V, Chaume D, Lefranc MP (2004) IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res 32:W435–W440

Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162:705–708

Jackson KJ, Gaeta B, Collins AM (2007) Identifying highly mutated IGHD genes in the junctions of rearranged human immunoglobulin heavy chain genes. J Immunol Methods 324:26–37

Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, Browne P, van den Broek A, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Gamble J, Diez FG, Harte N, Kulikova T, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Sobhany S, Stoehr P, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R (2005) The EMBL nucleotide sequence database. Nucleic Acids Res 33:1

Lee CEH, Gaëta B, Malming HR, Bain ME, Sewell WA, Collins AM (2006) Reconsidering the human immunoglobulin heavy chain locus. 1. An evaluation of the expressed human IGHD gene repertoire. Immunogenetics 57:917–925

Lee CEH, Jackson KJL, Sewell WA, Collins AM (2007) Use of IGHJ and IGHD gene mutations in analysis of immunoglobulin sequences for the prognosis of chronic lymphocytic leukemia. Leuk Res 31:1247–1252

Meijer P-J, Andersen PS, Haahr Hansen M, Steinaa L, Jensen A, Lantto J, Oleksiewicz MB, Tengbjerg K, Poulsen TR, Coljee VW, Bregenholt S, Haurum JS, Nielsen LS (2006) Isolation of human antibody repertoires with preservation of the natural heavy and light chain pairing. J Mol Biol 358:764–772

Mehr R, Shannon M, Litwin S (1999) Models for antigen receptor gene rearrangement. I. Biased receptor editing in B cells: implications for allelic exclusion. J Immunol 163:1793–1798

Olee T, Yang PM, Siminovitch KA, Olsen NJ, Hillson J, Wu J, Kozin F, Carson DA, Chen PP (1991) Molecular basis of an autoantibody-associated restriction fragment length polymorphism that confers susceptibility to autoimmune diseases. J Clin Invest 88:193–203

Pallares N, Lefebvre S, Contet V, Matsuda F, Lefranc MP (1999) The human immunoglobulin heavy variable genes. Exp Clin Immunogen 16:36–60

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Stamatopoulos K, Belessi C, Moreno C, Boudjograh M, Guida G, Smilevska T, Belhoul L, Stella S, Stavroyianni N, Crespo M, Hadzidimitriou A, Sutton L, Bosch F, Laoutaris N, Anagnostopoulos A, Montserrat E, Fassas A, Dighiero G, Caligaris-Cappio F, Merle-Beral H, Ghia P, Davi F (2007) Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: pathogenetic implications and clinical correlations. Blood 109:259–270

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tomlinson IM, Cox JP, Gherardi E, Lesk AM, Chothia C (1995) The structural repertoire of the human V kappa domain. EMBO J 14:4628–4638

Wang Y, Jackson KJL, Sewell WA, Collins AM (2008) Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. Immunol Cell Biol 86:111–115

Zachau HG (1993) The immunoglobulin kappa locus-or-what has been learned from looking closely at one-tenth of a percent of the human genome. Gene 135:167–173