Immunogenetics (2012) 64:3–14
DOI 10.1007/s00251-011-0559-z

ORIGINAL PAPER

Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells

Katherine Jean Louise Jackson · Yan Wang · Bruno A. Gaeta · William Pomat · Peter Siba · Janet Rimmer · William A. Sewell · Andrew M. Collins

Received: 20 May 2011 / Accepted: 12 July 2011 / Published online: 26 July 2011

Abstract We have analysed the transcribed immunoglobulin kappa (IGK) repertoire of peripheral blood B cells from four individuals from two genetically distinct populations, Papua New Guinean and Australian, using high-throughput DNA sequencing. The depth of sequencing data for each individual averaged 5,548 high-quality IGK reads, and permitted genotyping of the inferred IGKV and IGKJ germline gene segments for each individual. All individuals were homozygous at each IGKJ locus and had highly similar inferred IGKV genotypes. Preferential gene usage was seen at both the IGKV and IGKJ loci, but only IGKV segment usage varied significantly between individuals. Despite the differences in IGKV gene utilisation, the rearranged IGK repertoires showed extensive identity at the amino acid level. Public rearrangements (those shared by two or more individuals) made up 60.2% of the total sequenced IGK rearrangements. The total diversity of IGK rearrangements of each individual was estimated to range from just 340 to 549 unique amino acid sequences. Thus, the repertoire of unique expressed IGK rearrangements is dramatically less than previous theoretical estimates of IGK diversity, and the majority of expressed IGK rearrangements are likely to be extensively shared in individual human beings.

Keywords IGK · Immunoglobulin · Repertoire · Public rearrangements · Stereotypical rearrangements

Introduction

Vertebrate evolution has given rise to adaptive immune receptors with enormous potential diversity, presumably to enable responses to the diverse antigens encountered throughout a lifetime. The primary immunoglobulin repertoire results from genomic rearrangement events that bring together a number of short genes to generate each final immunoglobulin gene rearrangement (Tonegawa 1983). In humans, each immunoglobulin consists of two identical heavy chains and two identical light chains joined by disulfide bonds. The light chains are encoded by rearranged variable (V) and joining (J) genes from the kappa locus (immunoglobulin kappa, IGK) or from the lambda locus

Electronic supplementary material The online version of this article (doi:10.1007/s00251-011-0559-z) contains supplementary material, which is available to authorized users.

K. J. L. Jackson · Y. Wang · A. M. Collins
School of Biotechnology and Biomolecular Sciences,
The University of New South Wales,
Sydney, NSW, Australia

K. J. L. Jackson (✉) · B. A. Gaeta
School of Computer Science and Engineering,
The University of New South Wales,
Sydney, NSW, Australia
e-mail: jacksonk@stanford.edu

W. Pomat · P. Siba
Papua New Guinea Institute of Medical Research,
Goroka, Eastern Highlands Province, Papua New Guinea

J. Rimmer
St Vincent's Clinic,
Darlinghurst, NSW, Australia

W. A. Sewell
The Garvan Institute of Medical Research and St Vincent's
Clinical School, University of New South Wales,
Sydney, NSW, Australia

Present Address:
K. J. L. Jackson
Department of Pathology, Stanford University,
300 Pasteur Drive, R214 MC 5324,
Stanford, CA 94305, USA

(immunoglobulin lambda, IGL). The human IGK locus, located on chromosome 2, is made up of two domains spaced approximately 800 kb apart. The IGK constant region genes and IGKJ genes are located in the IGK proximal domain, along with 40 IGKV genes, pseudogenes and open reading frames (ORFs). The distal domain is essentially a duplication of the IGKV region of the proximal domain and contains an additional 36 IGKV genes, pseudogenes and ORFs (Zachau 1993; Zachau 1989).

Diversity in the primary IGK repertoire is generated by the combinatorial selection and joining of single IGKV and IGKJ genes, as well as the imprecise nature of the rearrangement event, with the addition and loss of nucleotides accompanying the rearrangement process (Alt and Baltimore 1982; Lafaille 1989). A B cell with a functional immunoglobulin receptor that is not auto-reactive may later be selected into the secondary repertoire if it encounters an antigen bound by its receptor. The secondary repertoire is further diversified through the introduction of targeted point mutations by somatic hypermutation of immunoglobulin genes, as part of the process of affinity maturation (Kim et al. 1981).

The choice of individual IGKV and IGKJ genes in IGK rearrangements is not entirely random, for there is preferential use of some gene segments. The high frequency of IGKV3-20*01 in expressed kappa light chains has been noted in both adults and neonates (Cox et al. 1994; Klein et al. 1993; Weber et al. 1994). Single-cell PCR of IGK rearrangements from human B cells has additionally revealed preferential usage of IGKV3-15*01, IGKV3-11*01, IGKV1-5*03, IGKV2-30*01 and IGKV1-30*01/IGKV1D-39*01 (Foster et al. 1997). The over-representation of these genes in rearrangements was also apparent in 1,863 IGK rearrangements that formed part of a bioinformatic analysis of the completeness of the reported IGK germline repertoire (Collins et al. 2008). Preferential gene usage also extends to the IGKJ genes, with more frequent usage of IGKJ1 and IGKJ2 (Collins et al. 2008; Foster et al. 1997; Weber et al. 1994) and underutilisation of IGKJ3 and IGKJ5(Collins et al. 2008; Foster et al. 1997).

The IGK locus appears to show little polymorphism (Cox et al. 1994; Zachau 1993) relative to the human immunoglobulin heavy (IGH) chain (Li et al. 2002). Targeted investigations of the IGKV2-29/IGKV2D-29 genes revealed a number of new alleles, suggesting that additional unreported alleles may exist (Atkinson et al. 1996; Feeney et al. 1996). More recent analysis of IGK rearrangements collected from public sequence databases however concluded that the IGK germline repertoire was relatively complete (Collins et al. 2008). The same study also noted that a number of translated products of IGK variable regions reported from independent sources were identical. These rearrangements were termed ‘stereotypical’ rearrangements and their occurrence among independent reports is unexpected given that the IGK repertoire has been previously estimated to include up to 10^{11} unique sequences (Janeway 2001). Other studies, however, have failed to observe contributions from stereotyped rearrangements, with no identical sequences reported among 700 IGK rearrangements sequenced from individual human B cells (Foster et al. 1997).

Recent developments in sequencing technology now allow rapid sequencing of many thousands of IGK rearrangements from a single individual (Boyd et al. 2009). Here, we report the use of high-throughput sequencing technology to explore the rearranged IGK repertoire of four individuals from two genetically distinct populations; Papua New Guinean and Australian. Following sequencing of each individual's transcribed IGK repertoire, genotypes of the IGKV and IGKJ loci were inferred and examined for the presence of unreported polymorphisms, and the similarities and differences between individuals' expressed IGK gene rearrangements were quantitated. Surprisingly, even in this relatively small number of subjects, only a minority of IGK rearrangements were unique to each individual, with over 60% of rearrangements being seen in more than one individual. IGK repertoire members were shared with similar frequencies between all subjects in the study, indicating that the shared IGK repertoire is common to relatively divergent human population groups.

Methods

Sample preparation

After informed consent, and with the approval of both the UNSW Human Research Ethics Committee and the Papua New Guinea Medical Advisory Council, peripheral blood was collected from four individuals; two residents of Sydney, Australia (denoted as K01 and K02) and two residents of Masilakaiufa Village, Eastern Highlands Province, Papua New Guinea (K03 and K04). Mononuclear cells were prepared from peripheral blood samples by density gradient centrifugation and frozen at −70°C for later processing.

Amplification and sequencing

Total cellular RNA was extracted from mononuclear cells, previously prepared from peripheral blood, using TRIzol(R) Reagent (Invitrogen, Carlsbad, CA). Synthesis of cDNA was performed using Superscript(R) III Reverse Transcriptase (Invitrogen) according to the manufacturer's protocol. IGK rearrangements were then amplified using PCR. Specific primers for the three largest IGKV gene families (IGKV1—CATCCRGWTGACCCAGTCTCC, IGKV2—TTGTGATGACYCAGWCTCCACTCTC, IGKV3—TGACR

CAGTCTCCASSCACCCT) were used as forward primers and a primer specific for the IGK constant region was used as reverse primer (IGKC—GAAGATGAAGACAGATGGTGCAG). GS FLX Titanium Primer A (CGTATCGCCTCCCTCGCGCCATCAG) was added to the 5′ end of the forward IGKV family-specific primers and Primer B (CTATGCGCCTTGCCAGCCCGCTCAG) was added to the 5′ end of the IGKC reverse primers. A multiplex identifier sequence (ATCAGACACG) was also added between the GS FLX Titanium Primers and the template-specific primers, to distinguish the sequences reported here from sequences that were amplified as part of a separate study.

PCR amplification was performed using the FastStart High Fidelity PCR System (Roche, Mannheim, Germany) with additional 10 mM PCR grade Nucleotide Mix (Roche). Master mixes were made of 50 ng cDNA, 0.4 μM of each primer (International DNA Technologies, Coralville, IA), 200 μM of each dNTP (Roche), 1.25 units FastStart High Fidelity Enzyme Blend (Roche) and a buffer supplied by the manufacturer. The specific primers for the IGKV1, IGVK2 and IGKV3 gene subgroups were used in separate reactions. The PCR conditions were 95°C for 3 min, followed by 30 cycles of 95°C for 30 s, 60°C for 30 s, 72°C for 35 s, and then a final extension at 72°C for 5 min. The PCR products were then cleaned by gel extraction using QIAquick Gel Extraction Kits (QIAgen). The cleaned samples were then sent to the Ramaciotti Centre for Gene Function Analysis, University of New South Wales for sequencing using an FLX Titanium Genome Sequencer (454 Life Sciences).

Processing of IGK reads

The sequence datasets for the four individuals were separately analysed to identify IGK rearrangements. Reads were filtered to discard from analysis those that were shorter than 350 nucleotides, did not contain at least 150 nucleotides corresponding to an IGKV1, IGKV2 or IGKV3 gene or that included nucleotide insertions or deletions. IGKV genes were identified by BLAST alignment (Altschul et al. 1990) against an IGKV repertoire comprised of all genes plus in-frame pseudogenes and orphons obtained from the Immunogenetics Group (Lefranc et al. 2009). All remaining reads were converted to coding strand and the primer region and upstream sequence were removed. Reads for each of the four samples were then partitioned using the iHMMune-align alignment utility (Gaëta et al. 2007).

Genotyping and identification of unreported polymorphisms

The genotype for each of the four individuals was determined at the IGKV and IGKJ loci, as previously described (Boyd et al. 2010). The reads generated for each individual were also analysed for evidence of the rearrangement of as yet unidentified variants of IGKV and IGKJ genes. Analysis was undertaken as previously described (Boyd et al. 2010).

Analysis of IGK rearrangements

The copy number of unique rearrangements within each sample was determined allowing for shorter 100% identity matches to longer sequences within the same sample. The longest read was selected as the representative for each set of identical sequences and only rearrangements that were observed as at least two copies with a sense read accompanied by an anti-sense read were retained for further analysis. IGKV and IGKJ gene usage and IGKV to IGKJ pairing frequencies for each sample K01 through K04 was determined based on this non-redundant set of rearrangements.

One of the shortcomings of pyrosequencing is the introduction of nucleotide insertions or deletions (indels) through the misreading of homopolymer tracts (Margulies et al. 2005). In this study, any read containing an indel relative to the germline gene was therefore excluded from analysis. In addition, only rearrangements present as both sense and anti-sense reads were retained. The effectiveness of the filtering was assessed by clustering of the samples based on sequence homology, to identify sequence variants resulting from either somatic hypermutation or point mutation sequencing errors (Niu et al. 2010).

For evaluation of 'stereotypical' rearrangements within the samples, each unique rearrangement was reverted to the germline composition based on the IGK genes assigned by iHMMune-align. Finally, the reverted nucleotide sequences were translated to their amino acid sequences.

Diversity of IGK rearrangements

Shannon entropy measures the diversity of categorical datasets, in this case, IGK rearrangements categorised based on their amino acid translations. The alpha component (Whittaker 1972), a measure of IGK diversity for each individual reflecting the number of unique IGK rearrangements that comprise an individual's transcribed repertoire, was determined as the Shannon entropy within each sample. The gamma diversity (Whittaker 1972), or the 'global' diversity of IGK rearrangements across the four individual's studied, was calculated as the Shannon entropy of the combined samples. Finally, the beta diversity component (Whittaker 1972) for each sample, representing the relationship between the set of IGK rearrangements of an individual's IGK repertoire and those rearrangements that define the 'global' diversity of all samples, was derived from the alpha and gamma measures using the method of Jost (Jost 2007). As the alpha diversity of each of the individuals varied, all Shannon entropies were expressed as their number equivalents (Jost 2007).

At the depth of sequencing used in this study, and considering the number of sequences that were observed only once in an individual's sample, it is likely that many rare IGK rearrangements in an individual's total repertoire were not detected. To account for the presence of unobserved species, all Shannon diversities were calculated as described by Chao and Shen (2002) as this provides for a non-parametric estimation of Shannon's index where there are undetected species. Chao and Shen's approach utilises the concept of sample coverage to adjust the Shannon index (H) for rearrangements that escaped sampling. The sample coverage is estimated from the proportion of unique IGK rearrangements that are observed just once within a dataset. The sample coverage adjusted Shannon entropy ($\hat{H}$) for the alpha ($\hat{H}_\alpha$) and gamma ($\hat{H}_\gamma$) components were then transformed to their numbers equivalent by taking their exponentials (Jost 2007). Finally, the beta component ($\hat{H}_\beta$) was calculated based on the relationship established in Jost (2007):

$$\widehat{\mathrm{H}}_\beta = \widehat{\mathrm{H}}\gamma / \widehat{\mathrm{H}}\alpha$$

Pair-wise similarities between the repertoire of IGK rearrangements from each individual were calculated using the Sorensen measure adjusted for the estimated number of unseen rearrangements within each sample as described by Chao and colleagues (Chao et al. 2004). Standard errors for pair-wise similarities were estimated from 200 bootstrap replicates.

The abundance data of IGK rearrangements for each individual was used to estimate the total number of unique IGK rearrangements within each sample. Estimations were made based on the contribution of abundant rearrangements plus an adjusted estimate for rare rearrangements. Rare rearrangements were defined as those that were observed just once in an individual. The estimate of the total number of unique IGK rearrangements was based on the sample coverage of the rare rearrangements and was calculated as described by Chao and Shen (2002).

IGK repertoire sampling

The completeness of the IGK rearrangement sampling within each individual was investigated by Monte Carlo re-sampling (Sepúlveda et al. 2010). Ten thousand simulations were performed for each of the four samples. For each simulation, as randomly selected sequences accumulated, the effect on diversity measured by numbers equivalent of Shannon's entropy was noted, along with whether or not the new addition represented a previously unseen IGK rearrangement. The mean diversity and the proportion of new IGK rearrangements based on the sequential addition were then calculated across all simulations.

Results

High-throughput sequencing of IGK rearrangements

To explore IGK rearrangements between individuals, cDNA was prepared from four individuals (K01–K04) and sequenced by GS-FLX sequencing. A total of 63,132 reads were generated: 16,652 reads from K01; 10,745 from K02; 18,285 from K03 and 17,450 from K04. Sequence read data for each individual is available at http://www.ihmmune.unsw.edu.au/IGKdiversity/. After filtering to remove low-quality sequencing data, the datasets included 5,215 rearrangements from K01, 3,580 from K02, 6,310 from K03 and 7,088 from K04. Cluster analysis indicated that filtering had suitably reduced potential pyrosequencing artefacts. Filtering reduced the overall extent of clustering and the remaining clusters were explained by the inherent similarity among the IGK rearrangements (data not shown).

IGK genotypes and unreported allelic variants

Genotypes encompassing the three largest IGKV families (IGKV1, IGKV2 and IGKV3) were inferred for each of the four individuals. These partial IGKV genotypes are presented as Table 1. Each person was found to have a distinct set of expressed IGKV1-IGKV3 genes. One individual (K03) was apparently homozygous for all observed IGKV. Two of the four individuals (K02, K04) appeared homozygous for all observed IGKV genes except IGKV2-30, for which they appeared to be heterozygous for the *01 and *02 alleles. The remaining individual (K01) carried two alleles of the IGKV1-5 gene (*01 and *03) in addition to the two IGKV2-30 alleles. At the IGKJ locus all four individuals were apparently homozygous for IGKJ1*01, IGKJ2*01, IGKJ3*01, IGKJ4*01 and IGKJ5*01. For less commonly expressed IGKV segments, there were too few rearrangements for confident assessment of allelic heterozygosity.

The individual datasets were examined for the presence of previously unreported polymorphisms. When an individual carries an unreported polymorphism, IGK rearrangements bearing the polymorphism will be assigned the most similar allele from the germline gene database used for partitioning. Within the dataset of rearrangements from an individual, an unreported polymorphism is therefore typically identified by the absence of apparently unmutated rearrangements of an allele, and by the presence of many examples of rearrange-

Table 1 IGKV genotypes for four individuals inferred from analysis of IGK rearrangements

IGKV	K01			K02			K03			K04		
IGKV1-5	***01**	*02	***03**	*01	*02	***03**	*01	*02	***03**	*01	*02	***03**
IGKV1-6	***01**			***01**			***01**			***01**		
IGKV1-8	***01**			***01**			***01**			***01**		
IGKV1-9	***01**			***01**			***01**			***01**		
IGKV3-11	***01**	*02		***01**	*02		***01**	*02		***01**	*02	
IGKV1-12	***01**	*02		***01**	*02		***01**	*02		***01**	*02	
IGKV1-13	*01 (P)	*02		*01 (P)	***02**		*01 (P)	*02		*01 (P)	*02	
IGKV3-15	***01**			***01**			***01**			***01**		
IGKV1-16	*01	***02**		*01	***02**		*01	***02**		*01	***02**	
IGKV1-17	***01**	*02		***01**	*02		***01**	*02		***01**	*02	
IGKV3-20	***01**	*02		***01**	*02		***01**	*02		***01**	*02	
IGKV2-24	***01**			***01**			***01**			***01**		
IGKV1-27	***01**			***01**			***01**			***01**		
IGKV2-28/IGKV2D-28	***01**			***01**			***01**			***01**		
IGKV2-29	*01P	*02	*03	*01P	*02	*03	*01P	*02	*03	*01P	*02	*03
IGKV2-30*01	***01**	***02**		***01**	***02**		***01**	*02		***01**	***02**	
IGKV1-33/IGKV1D-33	***01**			***01**			***01**			***01**		
IGKV1-37/IGKV1D-37	***01 (ORF)**			*01 (ORF)			*01 (ORF)			***01 (ORF)**		
IGKV1-39/IGKV1D-39	***01**			***01**			***01**			***01**		
IGKV2-40/IGKV2D-40	***01**			***01**			*01			***01**		
IGKV2D-30	***01**			***01**			***01**			***01**		
IGKV2D-29	***01**	*02		***01**	*02		***01**	*02		***01**	*02	
IGKV2D-26	*01	*02		*01	*02		*01	*02		*01	*02	
IGKV2D-24	***01 (ORF)**			***01 (ORF)**			*01 (ORF)			***01 (ORF)**		
IGKV3D-20	*01			*01			*01			*01		
IGKV1D-17	*01	*02		*01	*02		*01	*02		*01	*02	
IGKV1D-16	***01**	*02		***01**	*02		***01**	*02		***01**	*02	
IGKV3D-15	***01**	*02 (P)		*01	*02 (P)		*01	*02 (P)		*01	*02 (P)	
IGKV1D-13	*01			*01			*01			*01		
IGKV1D-12	***01**	*02		***01**	*02		***01**	*02		***01**	*02	
IGKV3D-11	*01			*01			*01			*01		
IGKV1D-43	***01**			***01**			*01			*01		
IGKV1D-8	*01			*01			*01			*01		
IGKV3D-7	*01			*01			*01			*01		

For each IGKV gene the presence of an allele in a given individual is indicated by grey shading. Where distal and proximal genes had identical coding regions no attempt was made to delineate between the two. Functional genes outside of IGKV1, IGKV2 and IGKV3 have been excluded. Gene names are ordered according to their position in the IGKV loci from closest to farthermost proximity to the IGKJ locus and alleles that are either open reading frames (ORFs) or pseudogenes (P) are noted

ments that include apparently shared mismatches. While such novel alleles appear frequently in high-throughput sequencing of immunoglobulin heavy chain rearrangements (Boyd et al. 2009), no evidence of new IGKV or IGKJ allelic variants was seen among the four individuals studied.

IGK gene usage

Each filtered dataset was re-analysed using only those IGK genes confirmed to be part of each individual's genotype, to ensure assignment only of genes present in the genotype. To avoid biases from variation in copy number of identical rearrangements, a dataset of unique rearrangements was created for each individual. Unique rearrangements were observed with on average 5.2 copies, up to a maximum of 100 copies. The datasets containing just a single copy of each unique rearrangement included 445 IGK rearrangements from K01, 337 from K02, 554 from K03 and 547 from K04.

Individuals expressed rearrangements involving between 22 (K03) and 25 (K01 and K04) of the 42 reported, functional IGKV1, IGKV2 and IGKV3 genes. Across the four individuals, 28 of the functional IGKV1, IGKV2 and IGKV3 genes were observed. Clear preferences for IGKV utilisation were observed (Table 2). In each of the four samples, the three most frequently rearranged IGKV genes accounted for at least 35% of total rearrangements, ranging from 35.3% (K01) to 55.2% (K02) of rearrangements. The highly utilised genes differed significantly (p<0.000001, Pearson's chi-square) for each of the individuals (Table 2). For example, IGHV1-5*03 was present only in rearrangements from K01 where it was observed in 7.2% of rearrangements (p<0.001, Pearson's chi-square). The utilisation of IGKV3-20*01 was also striking. The gene was present in expressed rearrangements by all individuals and accounted for 32.1% and 30.7% of all rearrangements in K02 and K03, respectively. The other individuals studied showed high utilisation of IGKV3-20*01, but at a frequency approximately half that of K02 and K03 (K01, 17.1%; K04, 16.1%; p<0.000001, Pearson's chi-square). Individuals also displayed preferential utilisation of the proximal domain genes, with 69.1% of rearrangements including genes whose coding sequences are unique to the proximal domain.

The inclusion of IGKJ genes in expressed rearrangements was also biassed. On average 30.1% of the 1,883 unique rearrangements utilised IGKJ1*01, 26.9% utilised IGKJ2*01 and 24.1% included IGKJ4*01. IGKJ3 and IGKJ5 were both observed at a frequency of 9.4%. Despite these biases, the frequency of utilisation of the five IGKJ genes did not differ between the four individuals (p=0.14, Pearson's chi-square).

The pairing of IGKV and IGKJ genes was investigated in each individual to determine if particular IGKV were observed to pair more, or less, frequently with the three most frequently utilised IGKJ genes. The frequency of rearrangements incorporating given IGKV and IGKJ1, IGKJ2 and IGKJ4 genes displayed (Supplementary Table 1)

Table 2 IGKV gene usage in rearrangement from four individuals

IGKV gene	K01	K02	K03	K04	Significance
IGKV1-5*01	7.2				***
IGKV1-5*03	4.3	4.5	8.1	6.2	
IGKV1-6*01	0.7	0.3	1.3	0.6	
IGKV1-8*01		0.3		0.2	
IGKV1-9*01	2.5	2.4	6.0	2.0	
IGKV3-11*01	9.2	5.0	5.2	5.9	
IGKV1-12*01	2.3	3.0	1.4	2.2	
IGKV1-13*02		1.8			**
IGKV3-15*01	8.8	14.2	7.9	8.6	
IGKV1-16*02	1.1	1.2	1.4	1.3	
IGKV1-17*01	4.5	3	5.1	2.6	
IGKV3-20*01	17.1	32.0	31.0	16.1	***
IGKV2-24*01	3.2	4.5	0.4	3.7	*
IGKV1-27*01	0.9	1.5	1.8	1.7	
IGKV2-28*01/IGKV2D-28*01	9.0	8.9	5.6	15.5	***
IGKV2-30*01	4.3	6.8	0.9	9.9	***
IGKV2-30*02	3.4	2.1		4.8	***
IGKV1-33*01/IGKV1D-33*01	5.8	3.9	5.1	5.7	
IGKV1-39*01/IGKV1D-39*01	7.6	3.3	15.3	7.7	***
IGKV2-40*01/IGKV2D-40*01	1.1	0.6		0.6	
IGKV2D-30*01			0.2	1.1	
IGKV2D-29*01	2.7	0.9	1.4	1.7	
IGKV1D-16*01	1.4		0.7	0.7	
IGKV3D-15*01	2.0				***
IGKV1D-12*01	0.9		1.4	1.7	
IGKV1D-43*01	0.2				

The frequency of IGKV gene usage across the four individuals is expressed as percentage of total rearrangements for an individual. Blank cells indicate that a gene was absent from rearrangements for that individual

* p<0.001, ** p<0.0001 and *** p<0.00001; significant differences in the use of a gene across the four individuals

biases suggestive of preferential gene pairing in three of the four individuals (contingency tables: K01, $p<0.001$; K03, $p<0.001$; K04, $p<.001$). Rearrangements from the repertoire of K02 did not display such biases (K02, $p=0.007$), however, this was the smallest of the sample datasets and the sparseness of the data may impact that analysis.

Contribution of public sequences to IGK repertoires of individuals

Each of the 1,833 unique IGK rearrangements was translated to its amino acid sequence following the reversion of any mutations to the corresponding nucleotide from either the germline IGKV or IGKJ. If an IGK rearrangement's amino acid translation was observed in the repertoire of two or more individuals then the IGK rearrangement was classified as shared or public. The proportion of public rearrangements within each individual's repertoire ranged from 57.5% (K01) to 61.5% (K02) and these accounted for 60.2% of all rearrangements studied (Fig. 1). Thirty unique amino acid translations were observed in the repertoires of all four individuals representing 500 of the 1,883 IGK rearrangements (Fig. 2). An additional 46 amino acid translations were observed in three of the four samples and a further 93 amino acid translations appeared in two of the four individuals. A summary of the public IGK rearrangements from all individuals is presented as Supplementary Table 2 and details of the private rearrangements unique to each of the four individuals is presented as Supplementary Table 3.

Pair-wise similarities for the four IGK repertoires were calculated using Sorensen similarity measure and are presented as Table 3. The four IGK repertoires were found to be highly similar with each other. The beta diversity component of each individual dataset was calculated following the determination of the Shannon entropy alpha and gamma diversities (Table 4). Shannon diversities were expressed as their number equivalents and therefore describe the number of equally likely elements needed to produce the given diversity index (Jost 2007). The beta diversity was on average 1.5, suggesting that each individual's set of IGK rearrangements is composed of two distinct components that would appear unequal in their contributions to an individual's repertoire.

The completeness of the sampling of each individual's IGK repertoire was examined through Monte Carlo resampling simulations (Fig. 3a, b). The samples were dominated by rare IGK rearrangements that were observed just one or two times (Fig. 3c). The repertoire sampling was substantial but not complete as the diversity measure does not achieve complete plateau and the proportion of new rearrangements fails to reach zero. An estimate of the lower bounds of the number of unique IGK rearrangements for each sample was made and is presented as Table 4.

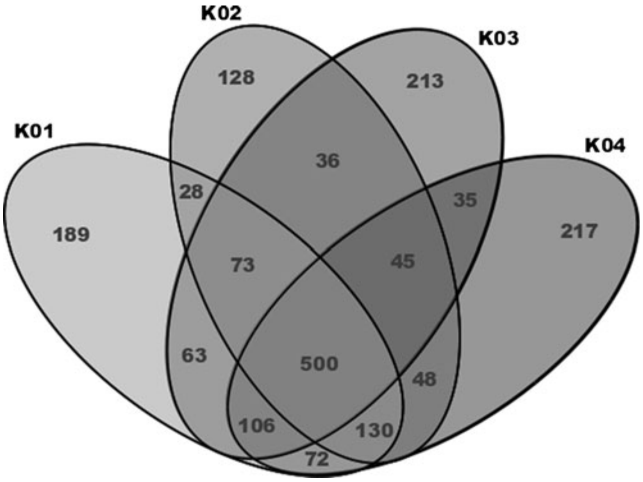

Fig. 2 The distribution of public and private sequences in the IGK repertoire of four individuals. Private sequences were observed only in the repertoire of single individuals. Public sequences were shared by between two and four individuals. Counts reflect the number of unique nucleotide sequences in each sample that once reverted to the germline composition shared the same amino acid translation

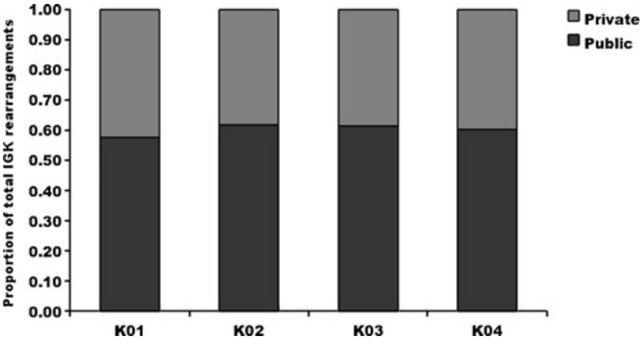

Fig. 1 The contribution of public and private sequences to the IGK repertoire of four individuals. Public sequences were present in two or more individuals, while private sequences were present in only a single individual. The IGK repertoire for each individual was examined following the reversion of mutated rearrangements to their germline composition

Table 3 Pair-wise similarities between the repertoires of IGK rearrangements were calculated for K01 to K04

	K01	K02	K03
K02	0.62 (0.078)		
K03	0.60 (0.059)	0.53 (0.070)	
K04	0.58 (0.055)	0.49 (0.062)	0.46 (0.062)

Similarity was measured by Sorensen index adjusted for contribution from undetected rearrangements. Standard errors were estimated from 200 bootstrap replicates and are shown in brackets

Table 4 Diversity of IGK repertoires of four individuals was assessed by Shannon entropy using a non parametric approach

	Alpha diversity	Beta diversity	Estimated unique rearrangements (amino acid translations)
K01	404.37	1.00	548.1
K02	279.19	1.45	350.7
K03	204.66	1.98	339.7
K04	282.94	1.43	475.8
Overall	414.80[a]	–	1,016.4

Shannon entropies are presented as their numbers equivalents. Sample coverage, based on the contribution to each sample from rearrangements that were observed just once, was used to estimate the lower bounds of the number of IGK rearrangements within each sample. IGK repertoires were analysed at the amino acid level, following reversion of any somatic point mutations to the germline nucleotide

[a] The alpha diversity of combined set of rearrangements is equivalent to the gamma diversity of the four samples

Rearrangement of public and private sequences

We evaluated the IGK sequence sets to determine if there were any significant differences in the V and J gene segments that were used in public rearrangements compared to private rearrangements in the four individuals studied. Only one individual (K03) differed significantly in IGKV ($p<$ 0.000001) and IGKJ ($p<0.000001$) gene usage between public and private rearrangements. The source of difference for IGKV genes was an increased tendency for rearrangements using the most frequently rearranged gene segment, IGKV3-20*01 to be categorised as public (23.9% private, 34.9% public). A similar trend was observed across all samples, but only achieved significance in K03. K03 also utilised IGKJ3*01 in fewer public rearrangements than expected, with just 4.7% of public compared to 21.1% of private rearrangements using this gene. Conversely, IGKJ4*01 was more prominent in the public rearrangements, where it comprised 31.7% of rearrangements compared to 12.7% of private rearrangements.

Junctional diversification mechanisms appeared to result in similar junction lengths in IGK sequences in public and private rearrangements, but via different mechanisms: public rearrangements lost on average 3.1 nucleotides during joining, while private rearrangements lost on average 2.8 nucleotides ($p=0.0015$, Student's t test). The public rearrangements however had on average 0.4 N-nucleotide additions, compared to 2.5 nucleotides for the private rearrangements ($p<0.001$, Student's t test). Similar levels of net processing were therefore achieved as increased additions to the private rearrangements were offset by increased nucleotide losses from both the IGKV ($p<0.001$, Student's t test: 2.3 nucleotides public, 3.2 nucleotides private) and IGKJ ends ($p<0.001$, Student's t test:1.3 nucleotides public, 2.1 nucleotides private).

Discussion

Compared to the immunoglobulin heavy chain, the light-chain gene loci have a decreased capacity for sequence diversity, due to the lack of D segments, and fewer non-templated bases inserted at the V–J junction, but empirical measurement of the true diversity of expressed IGK or IGL chains, and the degree of similarity of the expressed light chain repertoire between different individuals has not been reported. High-throughput sequencing data of expressed IGK rearrangements permits the germline gene segments in the subjects studied to be inferred, and any ‘public’ or shared rearrangements between individuals to be identified. We have identified a surprisingly high proportion of such public rearrangements.

Inference of the IGKJ genotype for each individual in our study showed them all to be homozygous for the *01 alleles of each of the IGKJ genes. Two additional alleles, one each of IGKJ2 and IGKJ4, have been previously reported in humans and confirmed by genomic sequencing of unrearranged IGKJ genes (Feeney 2000). The absence of these IGKJ alleles in the four individuals included in this study is consistent with their reported very low frequencies in the general population (Feeney 2000). A further two IGKJ2 alleles have been proposed (Barbié and Lefranc 1998) but remain unconfirmed by genomic sequencing (Collins et al. 2008). At the IGKV locus, each individual possessed a highly similar set of expressed IGKV gene alleles. This high degree of similarity did not, however, translate to highly similar frequencies of IGKV gene usage by the individuals. The average frequency of IGKV gene usage by the four individuals studied was however consistent with earlier reports (Collins et al. 2008; Cox et al. 1994; Foster et al. 1997; Weber et al. 1994). At the IGKJ locus, where all individuals shared the same genotype, they also each displayed the same biassed IGKJ gene usage. The biases were consistent with previous reports of frequent utilisation of IGKJ1, IGKJ2 (Foster et al. 1997; Weber et al. 1994) and IGKJ4, and much less use of IGKJ3 and IGKJ5 (Foster et al. 1997). Frequent IGKJ4 usage has also been reported from analysis of a collection of 435 sequences representing relatively unmutated IGK rearrangements deposited into public sequence databases (Collins et al. 2008). Given that the IGKJ genes are thought to contribute very few residues to the antigen binding site of immunoglobulins, the IGKJ biases observed may be more likely to arise from different recombination frequencies such as differences in the recombination signal sequences that flank each functional IGK gene (van Gent et al. 1996) or from differences in the accessibility of different IGK genes on the chromosome (Cobb et al. 2006).

The differences in observed IGKV gene usage, despite highly similar genotypes, could be explained either by

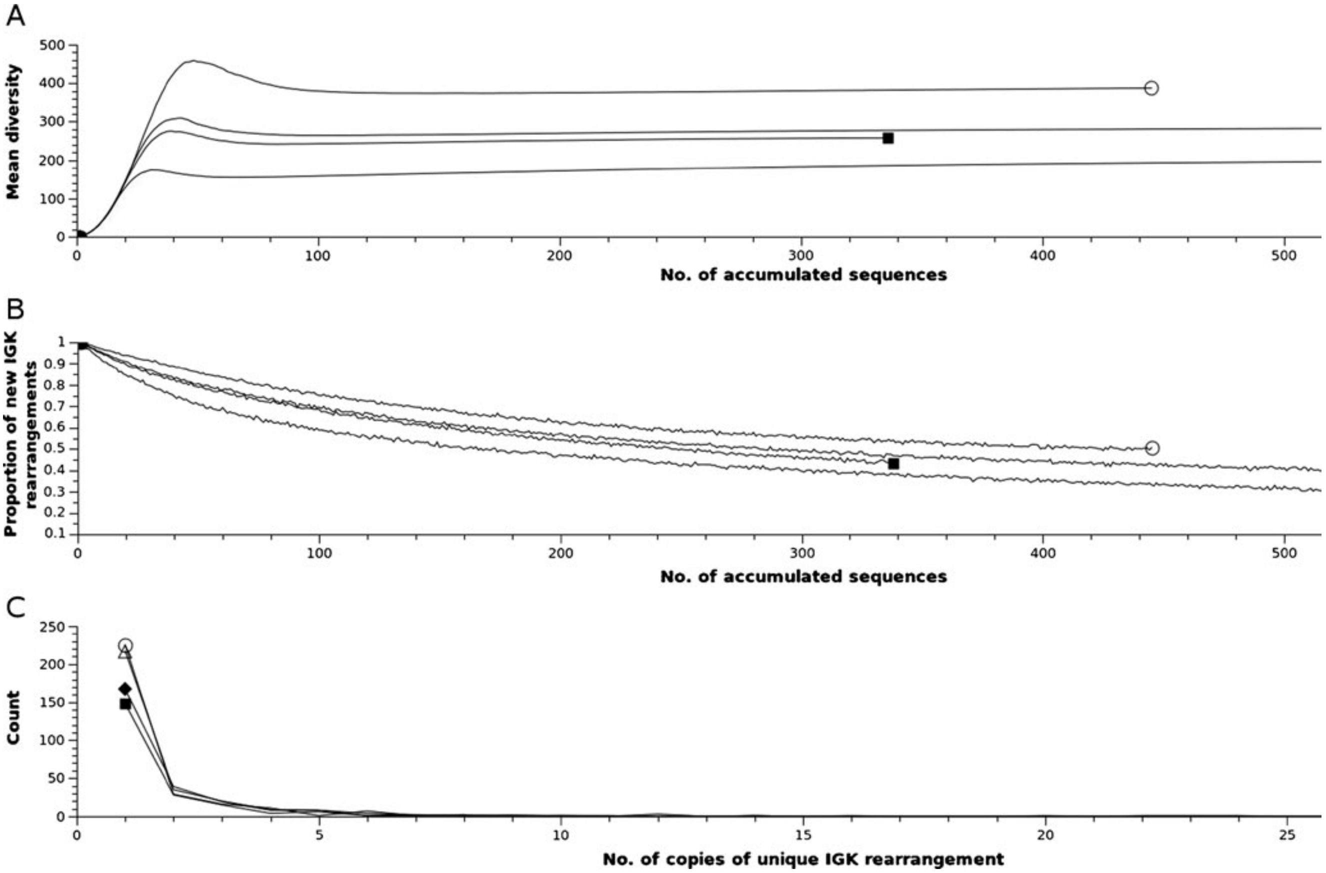

Fig. 3 The sampling of the IGK rearrangement sampling from K01 (*ellipse*), K02 (*rectangle*), K03 (*diamond*) and K04 (*triangle*) was assessed by Monte Carlo simulations. Ten thousand re-sampling simulations were performed and the impact on sample diversity (**a**) and accumulation of previously unseen rearrangements (**b**) was assessed. The shape of the IGK rearrangement repertoire was also assessed based on abundance of unique IGK rearrangements (**c**)

effects of selection on B cells, or potentially by IGKV copy number variation between individuals. All four individuals in this study carry IGKV1-5*03 and there is no significant variation in the utilisation of this gene across the four individuals ($p>0.05$). A single individual, however, carries the additional IGKV1-5*01 allele. Surprisingly, this is not associated with a lower frequency of utilisation of the IGKV1-5*03 allele. It is possible that the two IGKV1-5 genes are not the result of a heterozygosity at a single IGKV1-5 gene loci, but rather that one IGKV haplotype includes a duplication of the IGKV1-5 gene. It is also possible that the three individuals whose IGK rearrangements include only the *03 allele carry a haplotype that includes a deletion polymorphism of this gene. Possible copy number variations are also evident for gene loci where all individuals appear apparently homozygous. The most striking of these is IGKV3-20*01, where two individuals utilise the gene at approximately twice the frequency of the other two subjects. Haplotypes that include gene deletions have previously been identified by restriction fragment length polymorphism studies of IGK genes (Pargent et al. 1991; Schaible et al. 1993). One of the proposed haplotypes lacks the complete distal copy of the IGK locus and 0.7% of the population were predicted to be homozygous for the deletion (Pargent et al. 1991). The low frequency of distal gene rearrangement in the current study could in part be explained by heterozygosity of distal deletion haplotypes.

The four individuals in this study were examined for evidence of previously unreported polymorphism. No evidence of such polymorphisms was found for either the IGKV or IGKJ genes. This is consistent with previous observations that the described IGK germline repertoire is essentially complete (Collins et al. 2008), but is in stark contrast to the IGH locus. In a separate pyrosequencing study of the IGHV locus, that included one individual from this study (K04), 25% of the germline IGHV genes that were seen represented new allelic variants (Wang et al. 2011). The lack of polymorphism in the IGK locus could be consistent with selection pressure to maintain only a small number of allelic variants of IGK genes, while evolution has given rise to many polyallelic genes in the IGH locus.

Examination of the Shannon entropy of the IGK repertoires of the four individuals showed that they were comprised of two unique diversity components relative to the overall diversity of the IGK rearrangements across the samples. The two distinct components of each individual's repertoire can be thought of as the public and private rearrangements. Public rearrangements are those that are shared by two or more individuals, while private rearrangements are observed only within a single individual. The IGK repertoire of each person is highly similar due to substantial contributions from public rearrangements. Surprisingly, approximately 60% of the IGK rearrangements in any given person are public and this explains the reporting of the same rearrangements by a number of independent studies (Collins et al. 2008).

The high degree of sharing among individuals is unexpected given previous estimates of the potential size of the primary human immunoglobulin repertoire. Such estimates vary greatly, ranging from 10^{11} unique immunoglobulins (Janeway 2001) to 3.27×10^{74} unique sequences (Saada et al. 2007). The processing of the VJ junction provided just a 4.5-fold increase in IGK diversity in the individuals studied. This is much less than the expectation from models and theoretical calculations (Janeway 2001; Saada et al. 2007). The biassed gene usage across all individuals additionally constrains the primary IGK repertoires. The cumulative effects of the biassed gene usage, the highly similar genotypes and lack of junctional diversity lead to a pool of unique IGK rearrangements likely to be measured in the thousands, rather than millions.

The IGK repertoire is also biassed towards public rearrangements. Those rearrangements that are subject to higher than average junctional processing go on to form the private IGK rearrangements. Both the public and private IGK rearrangements, however, maintain the same average net processing, suggesting that the maintenance of junction length is structurally important for functional light chains. Structurally or functionally preferred conformations created by the pairing of particular IGKV and IGKJ genes could potentially explain the overall restricted diversity of the transcribed IGK repertoire. Such preferences are hinted at by the biases in IGKV to IGKJ pairings noted in three of the four individuals under study. Particular IGK rearrangements could be excluded from the transcribed repertoire due to the existence of a limited number of rearrangements compatible with IGH chains or due to the counter selection of auto-reactive CDR3 specificities. Similarly, rearrangements that include stop codons or culminate in frame-shifts of the IGKJ gene will also bias the primary repertoire. The implications of such factors on the primary IGK repertoire warrants further investigation particularly among the non-functional IGK rearrangements sequenced from genomic DNA templates.

Any constraints on the primary repertoire however only partially explain the dominance of public rearrangements. The IGK rearrangements studied here were collected from the periphery, and 87.0% of sequences carried five or more somatic point mutations (Supplementary Tables 2 and 3). Given the mutation levels, these public rearrangements must be considered to have been selected into the secondary repertoire. Selection into the secondary repertoire requires interaction with an antigen for which the B cell's immunoglobulin has some specificity. One possibility for such selection of public rearrangements across a number of individuals is IGK chain involvement in superantigen-like responses (Zouali 1995). The B-cell superantigen response is proposed to be analogous to that which occurs in the T-cell compartment, where antigen interacts directly with the variable region at sites outside of usual antigen binding sites. This often allows interaction with a number of members of the same V family due to sequence conservation among family members (Zouali 1995). Public sequences have been observed in the T-cell receptor (TcR) repertoire where the same TcRs may dominate responses to the same antigenic epitope in multiple individuals (Venturi et al. 2008).

IGKV gene usage has been studied in a number of clinical conditions including chronic lymphocytic leukaemia (CLL; Hadzidimitriou et al. 2008; Stamatopoulos et al. 2005). The reported stereotypical nature of CLL rearrangements refers to a different type of conservation of IGK rearrangements than that of the public IGK repertoire. In the context of CLL, the presence of highly similar rearrangements in a number of individuals may point to the potential involvement of antigen in the development of B cells crucial to these conditions. The stereotyped IGK CLL sequences previously reported (Hadzidimitriou et al. 2008) stand in stark contrast to the both the public and private IGK repertoires reported here. The frequency at which individual IGKV genes were utilised in IGK rearrangements sequenced from CLL B cells did not differ to those reported here (Ghiotto et al. 2006; Hadzidimitriou et al. 2008). The CLL rearrangements rather differed in their significant contributions, up to 11 amino acids, from N-nucleotide additions (Hadzidimitriou et al. 2008). The more extensively processed private IGK repertoire averaged just 2.5 N-nucleotide additions, the CLL IGKV rearrangements may therefore represent selection of rare and unusual rearrangements.

Diversity within the immunoglobulin response has been thought to be a major hallmark of a normally functioning humoral immune system, but the true diversity of light chains in expressed human immunoglobulins has not been studied in sufficient depth to assess the degree of uniqueness between individuals. In this study, IGK repertoires of four individuals demonstrated that each person's pool of rearranged IGK chains are highly similar, even between members of relatively divergent human population groups (Papua New Guinean, and Australian). This is unexpected with the given previous predictions of IGK repertoire size. The extensive public light chain rearrangements play could represent frequent rearrangements that do not interfere with antigen specificity primarily mediated by the immunoglobulin heavy chain, but could also be due to other mechanisms such as responses to superantigen-like molecules expressed by microbes. One practical consequence of limited light chain diversity is the prospect that future antibody expression studies could be facilitated by using a relatively small set of 'generic' light chains to pair with specific heavy chains, to create recombinant antibodies of suitable affinity for diagnostic or therapeutic uses.

Acknowledgements This was partially supported by the Chronic Lymphocytic Leukaemia Australian Research Consortium (CLLARC).

References

Alt FW, Baltimore D (1982) Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. Proc Natl Acad Sci USA 79:4118–4122

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi:10.1006/jmbi.1990.9999

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

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

Boyd SD, Gaëta BA, Jackson KJ et al (2010) Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol 184:6986–6992. doi:10.4049/jimmunol.1000445

Boyd SD, Marshall EL, Merker JD et al (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med 1:12ra23

Chao A, Chazdon RL, Colwell RK, Shen T-J (2004) A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol Lett 8:148–159. doi:10.1111/j.1461-0248.2004.00707.x

Chao A, Shen T-J (2002) Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environ Ecol Stat 10:429–443. doi:10.1023/A:1026096204727

Cobb RM, Oestreich KJ, Osipovich OA, Oltz EM (2006) Accessibility control of V(D)J recombination. Adv Immunol 91:45–109. doi:10.1016/S0065-2776(06)91002-5

Collins AM, Wang Y, Singh V et al (2008) The reported germline repertoire of human immunoglobulin kappa chain genes is relatively complete and accurate. Immunogenetics 60:669–676. doi:10.1007/s00251-008-0325-z

Cox JP, Tomlinson IM, Winter G (1994) A directory of human germ-line V kappa segments reveals a strong bias in their usage. Eur J Immunol 24:827–836. doi:10.1002/eji.1830240409

Feeney AJ (2000) New alleles of human immunoglobulin kappa J segments IGKJ2 and IGKJ4. Immunogenetics 51:487–488

Feeney AJ, Atkinson MJ, Cowan MJ et al (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. doi:10.1172/JCI118669

Foster SJ, Brezinschek HP, Brezinschek RI, Lipsky PE (1997) Molecular mechanisms and selective influences that shape the kappa gene repertoire of IgM+B cells. J Clin Invest 99:1614–1627. doi:10.1172/JCI119324

Gaëta BA, Malming HR, Jackson KJL et al (2007) iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. Bioinformatics 23:1580–1587. doi:10.1093/bioinformatics/btm147

van Gent DC, McBlane JF, Ramsden DA et al (1996) Initiation of V(D)J recombinations in a cell-free system by RAG1 and RAG2 proteins. Curr Top Microbiol Immunol 217:1–10

Ghiotto F, Fais F, Albesiano E et al (2006) Similarities and differences between the light and heavy chain Ig variable region gene repertoires in chronic lymphocytic leukemia. Mol Med 12:300–308. doi:10.2119/2006-00080.Ghiotto

Hadzidimitriou A, Darzentas N, Murray F et al (2008) Evidence for the significant role of immunoglobulin light chains in antigen recognition and selection in chronic lymphocytic leukemia. Blood 113:403–411. doi:10.1182/blood-2008-07-166868

Janeway C (2001) Immunobiology the immune system health & disease. Garland, New York

Jost L (2007) Partitioning diversity into independent alpha and beta components. Ecology 88:2427–2439

Kim S, Davis M, Sinn E et al (1981) Antibody diversity: somatic hypermutation of rearranged VH genes. Cell 27:573–581. doi:10.1016/0092-8674(81)90399-8

Klein R, Jaenichen R, Zachau HG (1993) Expressed human immunoglobulin kappa genes and their hypermutation. Eur J Immunol 23:3248–3262. doi:10.1002/eji.1830231231

Lafaille J (1989) Junctional sequences of T cell receptor gamma delta genes: implications for delta T cell lineages and for a novel intermediate of V-(D)-J joining. Cell 59:859–870. doi:10.1016/0092-8674(89)90609-0

Lefranc M-P, Giudicelli V, Ginestoux C et al (2009) IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 37:D1006–D1012. doi:10.1093/nar/gkn838

Li H, Cui X, Pramanik S, Chimge N-O (2002) Genetic diversity of the human immunoglobulin heavy chain VH region. Immunol Rev 190:53–68. doi:10.1034/j.1600-065X.2002.19005.x

Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380. doi:10.1038/nature03959

Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinforma 11:187. doi:10.1186/1471-2105-11-187

Pargent W, Schäble KF, Zachau HG (1991) Polymorphisms and haplotypes in the human immunoglobulin kappa locus. Eur J Immunol 21:1829–1835. doi:10.1002/eji.1830210808

Saada R, Weinberger M, Shahaf G, Mehr R (2007) Models for antigen receptor gene rearrangement: CDR3 length. Immunol Cell Biol 85:323–332. doi:10.1038/sj.icb.7100055

Schaible G, Rappold GA, Pargent W, Zachau HG (1993) The immunoglobulin kappa locus: polymorphism and haplotypes of Caucasoid and non-Caucasoid individuals. Hum Genet 91:261–267

Sepúlveda N, Paulino CD, Carneiro J (2010) Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. J Immunol Methods 353:124–137. doi:10.1016/j.jim.2009.11.009

Stamatopoulos K, Belessi C, Hadzidimitriou A et al (2005) Immunoglobulin light chain repertoire in chronic lymphocytic leukemia. Blood 106:3575–3583. doi:10.1182/blood-2005-04-1511

Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302:575–581

Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? Nat Rev Immunol 8:231–238. doi:10.1038/nri2260

Wang Y, Jackson K, Gaeta B et al (2011) Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. Immunogenetics. doi:10.1007/s00251-010-0510-8

Weber JC, Blaison G, Martin T et al (1994) Evidence that the V kappa III gene usage is nonstochastic in both adult and newborn peripheral B cells and that peripheral CD5+ adult B cells are oligoclonal. J Clin Invest 93:2093–2105. doi:10.1172/JCI117204

Whittaker RH (1972) Evolution and measurement of species diversity. Taxon 21:213. doi:10.2307/1218190

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

Zachau HG (1989) Immunoglobulin genes. Immunol Today 10:S9–S10

Zouali M (1995) B-cell superantigens: implications for selection of the human antibody repertoire. Immunol Today 16:399–405