Published in final edited form as:

Immunogenetics. 2012 May ; 64(5): 337–350. doi:10.1007/s00251-011-0595-8.

Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations

Ponraj Prabakaran,

Protein Interactions Group, Center for Cancer Research, Nanobiology Program, National Cancer Institute (NCI)-Frederick, National Institutes of Health (NIH), Bldg 469, Rm 150B, Frederick, MD 21702, USA

Basic Research Program, Science Applications International, Corporation-Frederick, Inc., NCI-Frederick, Frederick, MD 21702, USA

Weizao Chen,

Protein Interactions Group, Center for Cancer Research, Nanobiology Program, National Cancer Institute (NCI)-Frederick, National Institutes of Health (NIH), Bldg 469, Rm 150B, Frederick, MD 21702, USA

Maria G. Singarayan,

1901 Harpers Court, Frederick, MD 21702, USA

Claudia C. Stewart,

The Laboratory of Molecular Technology, Science Applications, International Corporation-Frederick, Inc., NCI-Frederick, Frederick, MD 21702, USA

Emily Streaker

Protein Interactions Group, Center for Cancer Research, Nanobiology Program, National Cancer Institute (NCI)-Frederick, National Institutes of Health (NIH), Bldg 469, Rm 150B, Frederick, MD 21702, USA

Basic Research Program, Science Applications International, Corporation-Frederick, Inc., NCI-Frederick, Frederick, MD 21702, USA

Yang Feng, Dimiter S. Dimitrov

Protein Interactions Group, Center for Cancer Research, Nanobiology Program, National Cancer Institute (NCI)-Frederick, National Institutes of Health (NIH), Bldg 469, Rm 150B, Frederick, MD 21702, USA

Abstract

Human cord blood cell-derived IgM antibodies are important for the neonate immune responses and construction of germline-based immunoglobulin libraries. Several previous studies of a relatively small number of sequences found that they exhibit restrictions in the usage of germline

---

dimitr.dimitrov@nih.gov.

Electronic supplementary material The online version of this article (doi:10.1007/s00251-011-0595-8) contains supplementary material, which is available to authorized users.

genes and in the diversity of the variable heavy chain complementarity determining region 3 compared to adults. To further characterize such restrictions on a larger scale and to compare the early B-cell diversity to adult IgM repertoires, we performed 454 sequencing and IMGT/HighV-QUEST analysis of cord blood IG libraries from two babies and determined germline gene usage, V-D-J rearrangement, VHCDR3 diversity, and somatic mutations to characterize human neonate repertoire. Most of the germline subgroups were identified with frequencies comparable to those present in the adult IgM repertoire except for the IGHV1–2 gene that was preferentially expressed in the cord blood cells. The gene usage diversity contributed to 1,430 unique IGH V-D-J rearrangement patterns while the exonuclease trimming and N region addition at the V-D-J junctions along with gene diversity created a wide range of VHCDR3 with different lengths and sequence variability. We observed a lower degree of somatic mutations in the CDR and framework regions of antibodies from cord blood cells compared to adults. These results provide insights into the characteristics of human cord blood antibody repertoires, which have gene usage diversity and VHCDR3 lengths similar to that of the adult IgM repertoire but differ significantly in some of the gene usages, V-D-J rearrangements, junctional diversity, and somatic mutations.

Keywords

454 Sequencing; IMGT/HighV-QUEST; Antibodyome; Human cord blood; Antibody repertoire; IgM; Immunoglobulin; Antibody library

Introduction

Characterization of large antibody libraries and potentially the immunoglobulin (IG) repertoire generated at different times by the immune system are now possible with the advent of high-throughput next-generation sequencing (NGS) technologies (Dimitrov 2010; Fischer 2011; Persson 2009). Recently, large numbers of antibodies from zebra fish (Jiang et al. 2011; Weinstein et al. 2009) and humans (Boyd et al. 2009, 2010; Collins et al. 2011; Fischer et al. 2010; Glanville et al. 2009; Wu et al. 2011) have been sequenced and analyzed using the NGS methods. The large-scale sequencing studies have provided valuable insights on the breadth of antibody repertoires expressed and immune B cell diversity—germline gene usage, junctional diversity, and somatic mutations. As a part of our studies of components of the human IG repertoire and to further understand initial responses to immunogens, we have been generating IgM antibody libraries and characterizing IgM monoclonal antibodies (Dimitrov 2010). Particularly, we are interested in the umbilical cord blood libraries containing B lymphocytes that presumably have not been exposed to exogenous antigens and have been used as a source of naturally occurring unmutated or minimally mutated pre-immune antibodies (Casali and Schettino 1996; Ridings et al. 1997). These germline-encoded natural antibodies could possibly play a vital role in the development and physiology of the human B cell repertoire as well as have the ability to bind a variety of exogenous antigens (Bobrzynski et al. 2005; Cooke et al. 1993; Messmer et al. 1999; Rodman et al. 2001). Furthermore, the human neonatal IgM antibody repertoire was found to be invariant and directed toward a limited set of self-antigens that could represent the targets for unmutated antibody specificities (Mouthon et al. 1995). It was already noted that human fetal, cord blood, and adult sources exhibit comparable IG

repertoires with respect to IGHV and IGHJ gene usage with only modest differences in the variable heavy chain complementarity determining region 3 (VH CDR3; Kolar et al. 2004). However, several studies of human and mouse antibody repertoires expressed during fetal life and at birth showed restrictions in variable (V) gene usage, somatic mutations, and VH CDR3 length (Bauer et al. 2001; Mortari et al. 1992; Ridings et al. 1997; Schroeder et al. 1987, 1999; Williams et al. 2009). Specifically, previous studies on human cord blood antibody repertoires contributed to the understanding of some of the characteristics of early B cell repertoires (Feeney et al. 1997; Mortari et al. 1993; Richl et al. 2008). However, all previous studies on the cord blood-derived antibodies were limited by sampling statistics due to the unavailability of massive parallel sequencing at that time. Here, we used high-throughput 454 sequencing and IMGT/HighV-QUEST (Alamyar et al. 2010, 2011) analysis of human cord blood antibody libraries from two babies to assess the extent of germline gene usage, distinct patterns of V-D-J rearrangements, exonuclease trimming and N region addition in the creation of VH CDR3 and somatic mutations existing in the cord blood repertoire. We found that gene usage is similar to that of adult IgM but with a significant shift in the IGHV1–2 gene usage in the neonate repertoire, and that VH CDR3 lengths are not restricted in the cord blood-derived heavy chains; however, amino acid (AA) compositions could not be the same between cord blood and adult IgM repertoires due to differences in V-D-J rearrangements and in junctional diversity due to exonuclease trimming and N addition. The large datasets of cord blood IgM antibody sequences analyzed in this study will be helpful to understand human IG repertoires and provide insights into the functional paratope diversity available for human neonates.

Materials and methods

DNA isolation, amplification, and 454 sequencing

Two cord blood samples from an African-American female baby and a Caucasian male baby with normal medical histories, designated as CB1 and CB2, respectively, were obtained from the National Disease Research Interchange, Philadelphia, PA, USA. Care was taken not to contaminate the cord blood samples with maternal blood. As a source for amplification of antibody sequences, cDNA was prepared from cord blood of two babies according to the reported protocol (Chen et al. 2009). The complete set of primers used in the PCR amplification of IgM-derived heavy and light chains were described in detail elsewhere (Zhu and Dimitrov 2009). Briefly, PCR amplifications were performed with a mixture of primers in which the 3'-ends corresponded to the first seven codons of IGHV1–7, IGKV1–6, and IGLV1–9 families according to the IMGT numbering for coding regions (Lefranc 2011b; Lefranc et al. 2003). PCR amplifications of constant domains were achieved by a sense primer specific for CH1 domain of IGHM spanning first eight codons and anti-sense primers specific for IGKC and IGLC domains spanning last eight codons at the 3' corresponding to the numbering schemes of IGKC and IGLC1 regions annotated in the IMGT (Lefranc 2011b; Lefranc et al. 2005). For 454 sequencing, primer combinations used to amplify cDNA in separate reactions included the Roche A and B adaptor sequences along with target amplification sequence for heavy and light chain variable domains. For heavy chain, rHF5: 5'-CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG ATC AGA CAC GTG GGG CGG ATG CAC TCCC-3' (sense, for cord blood 1), rHF6: 5'-CCA TCT CAT CCC TGC

GTG TCT CCG ACT CAG ATA TCG CGA GTG GGG CGG ATG CAC TCCC-3' (sense, for cord blood 2), and rHR1: 5'-CCT ATC CCC TGT GTG CCT TGG CAG TCT CAG GCT GCC CAA CCA GCC ATG GCC-3' (antisense, for both cord blood). For light chain, rLF5: 5'-CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG ATC AGA CAC GTG CAG CCA CAG TTCG-3' (sense, for kappa light chains of cord blood 1), rLF5-1: 5'-CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG ATC AGA CAC GCG AKG GGG YRG CCT TGGG-3' (sense, for lambda light chains of cord blood 1), rLF6: 5'-CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG ATA TCG CGA GTG CAG CCA CAG TTCG-3' (sense, for kappa light chains of cord blood 2), rLF6-1: 5'-CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG ATA TCG CGA GCG AKG GGG YRG CCT TGGG-3' (sense, for lambda light chains of cord blood 2), and rLR1: 5'-CCT ATC CCC TGT GTG CCT TGG CAG TCT CAG GGG CCC AGG CGG CC-3' (antisense, for kappa and lambda light chains of both cord blood). The antisense primers for amplification of heavy (VH) and light (VL) chain variable domains target the very beginning of the 5' ends of CH1 and CL gene fragments, respectively. The sense primers for VH and VL match the 3' ends of the PelB and OmpA leader peptide gene sequences, respectively, which are embedded in the phagemid pComb3x and direct the translational processing of immunoglobulin sequences. The attachment of Roche A and B Multiplex Identifier (MID) adaptor sequences, MID5 "ATCAGACACG" and MID6 "ATATCGCGAG", to the sense primers allow the sequencing to start from 3' ends and terminates at the 5' ends of VH and VL. The gene fragments were amplified in 20 cycles of PCR using the High Fidelity PCR Master from Roche. The standard Roche 454 GS Titanium shotgun library protocol was adapted as found in the Roche sequencing technical bulletin. The Multiplex Identifier, MID-containing adaptor sequences were used for pooling, parallel sequencing of libraries, and retrieving of sequences from each sample.

IMGT/HighV-QUEST sequence data analysis

The 454 sequence data were trimmed for quality and only sequences typically of length more than 300 nucleotides, covering the antibody variable domains consisting of the three complementarity determining regions (CDR) along with framework regions (FR), were retained for the sequence analysis. A total of 91,160 unique sequences of heavy and light chains from the two cord blood samples were analyzed using IMGT/HighV-QUEST (Alamyar et al. 2010, 2011). IMGT/HighV-QUEST is the high-throughput version of IMGT/V-QUEST for deep sequencing NGS data analysis resource. Details of the algorithm and method used in the tools for determining antibody functionality and identifying insertion/deletion errors, germline genes and alleles, V region mutations, CDR and junctional modification were described previously (Brochet et al. 2008; Giudicelli et al. 2011). Both tools integrate IMGT/JunctionAnalysis for a detailed and accurate analysis of the junctions (Giudicelli and Lefranc 2011; Yousfi Monod et al. 2004). Briefly, V, D, and J genes were identified by performing the global pairwise alignment against different subsets of the IMGT reference directory followed by a similarity evaluation based on alignment score and identity percentage. The identities of D and J gene alleles were determined as consecutive matches of at least more than four nucleotides occurring within the junction, and the best matches among the possible alleles were chosen. The IMGT Reference directory is from IMGT Repertoire (Lefranc and Lefranc 2001) and IMGT/GENE-DB (Giudicelli et al.

2005). The following versions were used: “IMGT/HighV-QUEST” (version 1.0.3), “IMGT/V-QUEST” (Version 3.2.17) and “IMGT/V-QUEST reference directory release” (201048–1) used at the time of the analysis, as indicated on the IMGT/HighV-QUEST portal web site. Total numbers of productive sequences of VH and light kappa (V-KAPPA) and lambda (V-LAMBDA) domains, and unique sequences of V regions and CDR3 at the amino acid level, along with the numbers of respective unproductive sequences, are given in Table 1. Results from the IMGT/HighV-QUEST analysis in CSV files were imported into PostgreSQL, an open source object-relational database system, and Structured Query Language (SQL) was used to retrieve the data for statistical analysis. Statistical calculations were carried out using SAS JMP9® statistical software (SAS Institute, Cary, NC, USA).

Results

Germline gene usage in human cord blood cell antibody repertoires

We investigated the germline gene usage in a large antibody repertoire of 28,169 VH and 24,530 VL domains (18,928 V-KAPPA (77%) and 5,602 V-LAMBDA (23%)) derived from human cord blood samples of two babies. The rearranged sequences selected for analysis were typically more than 300 nt in length and unique at the amino acid level with productive functionality as determined by IMGT/V-QUEST (Giudicelli et al. 2011). Figure 1 shows the frequencies of the IGHV subgroups, IGHD sets and IGHJ genes in the VH domains, IGKV subgroups and IGKJ genes in the V-KAPPA, and IGLV subgroups and IGLJ genes in the V-LAMBDA for the two cord blood samples, CB1 and CB2, observed in productive and unproductive sequences. For the productive VH repertoires, all IGHV gene subgroups, IGHD sets and IGHJ genes were represented though certain preferential gene usages were noted (Fig. 1a–c). Among the IGHV subgroups, IGHV1, IGHV3, and IGHV4 together accounted for 91.8% of the total IGHV while IGHV2, IGHV6, and IGHV7 subgroups were underrepresented with each less than 1%. No significant preferential usage among the IGHD sets and IGHJ genes existed, except for the IGHJ1 and IGHJ4 genes that were under and overrepresented, respectively. For the productive VL repertoires, all IGKV and IGLV gene subgroups were represented (IGKV1 to IGKV7, IGLV1, to IGLV10) (Fig. 1d). However, IGKV1, IGKV2, IGKV3, and IGKV4 subgroups, and IGLV1, IGLV2, and IGLV3 subgroups were mainly used. These subgroups contributed to 76% and 22%, respectively, of the total IGKV and IGLV.

The numbers of unique sequences identified for each V gene subgroup, D gene set, and J gene in the total pool of heavy and light chain sequences including both those functionally productive and unproductive are given in Table S1. Detailed classification of identified genes for each productive VH, V-KAPPA, or V-LAMBDA domain is given in Table 2. IGHD genes were identified in a total of 27,854 rearranged sequences. All seven IGHD sets and 25 different IGHD genes were identified (Table 3). Among the 52 IGHV genes found to be rearranged in VH domains identified as productive by IMGT/HighV-QUEST (no stop codon in the V region and in-frame junction), we observed that the pseudogene IGHV3-h (no initiation codon) occurred 16 times and that four genes IGHV1-c, IGHV3–22, IGHV3–35, and IGHV3-d appeared only once. We noted that IGHV3–22, a pseudogene gene carrying a stop codon at position 67, in which a point mutation “g” to “c” changed it to tyrosine. For

the VL domains, 45 IGKV and 34 IGLV genes were identified. The distribution of the IGKJ and IGLJ genes were analyzed (Fig. 1e and f). All five IGKJ genes were utilized in the V-KAPPA but only three IGLJ genes (IGLJ1, IGLJ2, and IGLJ3) were mostly utilized in the V-LAMBDA. The IGLJ7 gene was found only five times, and IGLJ5 and IGLJ6 were found twice for each in the rearranged V-LAMBDA. The IGLJ6 may be expressed if associated to the functional IGLC6 allele (Dariavach et al. 1987), whereas IGLJ5 is associated to IGLC5 which is always a pseudogene. For the unproductive VH and VL domains, we observed similar gene usages although their functionalities were affected by stop codons and out-of-frame rearrangements. We found four IGHV genes that were not seen in the productive VH domains, namely, IGHV4–55 appearing five times and other three, IGHV3–71, IGHV7–40, and IGHV7–81, were found once in the unproductive VH repertoire. It was already noted that IGHV4–55 is a pseudogene carrying a stop codon at position 39 but found to be rearranged (Brochet et al. 2008).

Length distributions of the CDR1 and CDR2 of the VH and VL

The complementarity determining regions CDR1 and CDR2 of the VH and VL domains are encoded by the germline genes. Their lengths are defined using the IMGT unique numbering (Lefranc et al. 2003; Lefranc 2011a) and the IMGT Collier de Perles graphical representations (Lefranc 2011b; Ruiz and Lefranc 2002). The length distributions of CDR1 and CDR2 of the VH, V-KAPPA, and V-LAMBDA (Fig. 2) reflect the usage of the different germline genes. Although this was expected, these results provide an important quality control on the sequences themselves and show how the analysis of IMGT/HighV-QUEST can be useful for sequence validation. For example, those sequences that had VH CDR1 and V-KAPPA CDR2 lengths of seven and two amino acids, respectively (bottom panel of Fig. 2a, b) do not correspond to correct germline CDR lengths and revealed the sequencing errors associated.

Somatic hypermutations in the V regions of the VH and VL

The numbers of unique, mutated V regions of productive VH, V-KAPPA, and V-LAMBDA were 13,888, 12,164, and 3,948, respectively, representing 49% of VH and 66% of VL. These results unambiguously show that somatic hypermutations are present in the cord blood-derived IgM antibodies. The extensive documentation on IG germline V regions at IMGT® (Lefranc et al. 2009) allowed us to quantify the exact number of AA changes from the closest germline genes in the cord blood IgM antibody repertoire. We compared the number of V region AA changes present in the VH and VL domains (Fig. 3). We did not include the mutations at the CDR3 and FR4 due to their involvement in the rearrangement process. The total number of AA changes in the V regions of the VH and VL domains ranged from 0 to 12 (Fig. 3a). There were V regions without any AA change for 5%, 3%, and 10% of the total sequences of VH, V-KAPPA, and V-LAMBDA, respectively. Overall, about 70% of VH and VL possessed between one and five AA changes in their V region. However, the average numbers of AA changes at CDR1 and CDR2 were remarkably lower as a majority of V domains (63% of VH, 70% of V-KAPPA, and 75% of V-LAMBDA) used germline-encoded CDR1 and CDR2 and has no AA change in these regions (Fig. 3b). In contrast, there were only about 7% of VH, 4% of V-KAPPA, and 14% of V-LAMBDA

whose V region FR (FR1, FR2, and FR3) was without AA change. The average numbers of AA changes along the V-region are shown in Fig. 3c.

V-D-J junction diversity due to trimming and P and N addition

We analyzed 28,169 productive and unique VH domains, from which IGHD genes were identified for 27,845 sequences by IMGT/HighV-QUEST. The V-D-J junction diversity results from the addition of palindromic “P” nucleotides (P region), exonuclease trimming and addition of “N” nucleotides (N region) at the V→D (N1) and D→J (N2) junctions (Fig. 4). In the V-D-J junction, P addition was observed at the 3′V, 5′D, and 5′J ends (no P region was found at the 3′D), whereas exonuclease trimming was observed at the 3′V, 5′D, 3′D, and 5′J ends. Exonuclease trimming and N regions were observed in 97% and 93%, respectively, of the VH while P regions were found in 26% of VH. Average numbers of nucleotides gained due to N regions (N1+N2) and P regions were 8.98 (± 0.04) and 1.71 (± 0.01), respectively. Average number of nucleotides lost due to exonuclease trimming was 15.00 (± 0.05). To compare these productive junctional properties with that of unproductive sequence data, we analyzed 13,803 unique unproductive VH domains for which 13,662 an IGHD gene was identifiable. We found that more than 95% of sequences exhibited N region addition and exonucleolytic loss with the average numbers of 10.98 (± 0.06) and 16.44 (± 0.07), respectively, while 27% of sequences had P region addition with an average value of 1.72 (± 0.02) that was similar to the corresponding productive P region.

CDR3 diversity in lengths and amino acid compositions

The rearranged CDR3 (positions 105–117 according to the IMGT unique numbering) is resulting of the V-(D)-J rearrangements and largely determines the diversity of expressed VH and VL repertoires. Out of 18,928 and 5,602 unique V-J sequences of V-KAPPA and V-LAMBDA domains, respectively, 2,927 and 1,193 sequences were only found with unique CDR3 regions correspondingly. The CDR3 length distributions of V-KAPPA and V-LAMBDA ranged up to 14 amino acids but highly concentrated in between 8 and 12 (Fig. 5a, b). In contrast, the VH repertoire contained 24,188 unique CDR3 regions out of 28,169 that were analyzed for their length diversity and AA composition. The CDR3 lengths ranged from 4 to 32 amino acids following a Poisson distribution with mean 14.03 (± 0.02 ; Fig. 5c). The most populated bin contained 12% of the total sequences having the CDR3 length of 13 amino acids. We examined the AA usage by calculating AA composition of all VH CDR3 sequences whose lengths ranged from 4 to 32. The average distribution of individual amino acids as percent of the total amino acids in the VH CDR3 showed the preferred AA usage for tyrosine, serine, glycine, and alanine along with charged amino acids arginine, and aspartic acid (Fig. 5d). To identify VH CDR3 AA distribution without any length-dependent bias, position-specific AA frequencies were calculated for the most common group of VH CDR3 sequences with the length of 13 amino acids positions 105–117 in the IMGT numbering (Lefranc 2011b; Lefranc et al. 2003; Fig. 5e). At the N-terminus of VH CDR3, 97% of sequences contained an alanine (Ala A) at position 105 and mainly positively charged amino acids at position 106 with 74% arginine (Arg R) and 8% lysine (Lys K; “AR” being encoded by most germline IGHV regions). At the C-terminus of VH CDR3, phenylalanine (Phe F), aspartic acid (Asp D), and tyrosine (Tyr Y) were found in 73%, 92%, and 56% of sequences, respectively, at the three consecutive positions 115, 116, and 117

(encoded by the J region). Other hydrophobic amino acids such as leucine and isoleucine were also observed in 10% and 20% of sequences at the position 117 as compared to the previous statistical analysis of IG VH amino acid properties (Pommié et al. 2004). Glycine and serine were spread across the CDR where AA composition varied greatly due to alternate IGHD reading frame usage and junctional diversity due to P regions, exonuclease trimming and N regions. Cysteine were observed about 1% each at positions 107, 108, 113, and 114, which could indicate possible disulfide bridges in some VH CDR3 sequences. For the unproductive rearrangements, there were 9,937 VH CDR3 sequences ranging from 4 to 30 amino acid lengths and the most common length was 14 amino acids representing 14% of the total populations.

V-D-J rearrangement patterns in human cord blood heavy chain repertoire

We observed that 52 IGHV, 7 IGHD, and 6 IGHJ genes in productive VH repertoire containing 27,854 sequences contributed to a total of 1,430 unique V-D-J rearrangement patterns determined by IMGT/HighV-QUEST, which was 65% of the total number of possible V-D-J ($52 \text{ V} \times 7 \text{ D} \times 6 \text{ J} = 2,184 \text{ V-D-J}$). Each V-D-J pattern could refer to a distinct set of antibodies resulting from different germline-encoded alleles found with different junctions and somatic hypermutations, but sharing the same IGHV, IGHD, and IGHJ genes. We noted that the V-D-J rearrangement patterns occurred at different frequencies (Fig. 6). Up to 61% of V-D-J rearrangement patterns had lower frequencies, occurring between one and nine times, while the remaining 39% were found at higher frequencies ranging in between 10 and 670 times. Specifically, there were 54 V-D-J patterns that occurred more than 100 times utilizing 10 different IGHV genes, and all IGHD and IGHJ genes, except IGHJ1 (Table S3). These frequently occurring V-D-J rearrangements contributed to the formation of 9,928 unique VH, i.e., 36% of the total VH domains found productive in cord blood repertoire. Particularly, the IGHV1–2-IGHD6-IGHJ4 rearrangement had the highest frequency which indicated a large number of individual antibodies existing with different alleles, junctions, and somatic mutations to possess this specific V-D-J rearrangement pattern. In the unproductive VH repertoire of 13,662 unique sequences, we found 54 IGHV, 7 IGHD, and 6 IGHJ genes contributing to a total of 1,338 unique V-D-J rearrangements following a similar pattern as in the productive repertoire. However, there were 13 high frequency V-D-J patterns utilizing six IGHV genes (IGHV1–2, IGHV1–69, IGHV3–23, IGHV4–34, IGHV4–59, and IGHV5–51), four IGHD genes (IGHD1, IGHD2, IGHD3, and IGHD6), and mostly one IGHJ gene (IGHJ6 in 12 patterns out of 13) which all combined to produce 12% of the total unproductive VH domains.

Discussion

We performed 454 sequencing and IMGT/HighV-QUEST analysis of the human cord blood antibody repertoire to gain insights into germline diversity and other early B cell characteristics related to VH CDR3 diversity, V-D-J rearrangement patterns, and somatic hypermutation of naturally occurring IgM, as related to somatic generation of antibody diversity (Tonegawa 1983). Heavy and light chain sequence data sets were separately derived from the cord blood samples of two babies and then combined for the analysis (Table 1). IMGT/HighV-QUEST analysis revealed the extended usage of all genes in heavy

and light chain domains (VH, V-KAPPA, and V-LAMBDA) for both productive and unproductive transcripts (Fig. 1). However, we noted significant preferential usage of several subgroups, IGHV1, IGHV3, and IGHV4 for VH; IGKV1, IGKV2, and IGKV3 for V-KAPPA; and IGLV1, IGLV2, and IGLV3 for V-LAMBDA. These IGHV, IGKV, and IGLV subgroup frequencies were comparable to those previously observed in 18,158 rolling circle amplified shotgun reads derived from 654 human donor IgM repertoires (Glanville et al. 2009). One notable exception was in regards to the most predominant IGHV subgroup usage between IgM repertoires; 46% of the total IGHV genes used in the cord blood repertoire were found to belong to the IGHV1 subgroup, whereas the other IgM repertoires of adult and mixed populations retain the IGHV3 subgroup as the most populated subgroup (~35% of the total IGHV genes; Glanville et al. 2009). This drastic shift in the IGHV subgroup usage might be due different molecular mechanisms in control of allelic exclusion or receptor editing during the B cell development (Chen et al. 1995; Koralov et al. 2006; Loffert et al. 1996; Nemazee 2006). Nonetheless, the IGHV1 subgroup appears as one of the most dominant with germline genes being used in cord blood as well as adult IgM repertoires (Boyd et al. 2010; Glanville et al. 2009). At the gene level, 52 IGHV genes, 45 IGKV, and 34 IGLV genes for VH, V-KAPPA, and V-LAMBDA, respectively, were identified which all contribute to the productive, expressed cord blood antibody repertoire (Table 2). A similar gene usage was also identified in the unproductive cord blood antibody repertoire containing 54 IGHV genes, 44 IGKV, and 34 IGLV genes for VH, V-KAPPA, and V-LAMBDA, respectively. It was found that the ratio of productive to unproductive transcripts was approximately 1:2 in the cord blood repertoire except for IGHV1 and IGLJ3 genes, which had the lowest and highest ratios, respectively (Fig. 1a and f). Accordingly, the most common IGHV gene in the repertoire was IGHV1–2 which accounted for 20% of the total heavy chains. The IGKV3–20 gene was used in 24% of the total kappa chains while the IGLV2–14 gene accounted for 14% of the total lambda chains along with other genes IGLV1–40, IGLV3–1, and IGLV3–21, each contributing >10% of the total V-LAMBDA repertoire.

In the IGHD gene repertoire, IGHD1–26 and IGHD6–13 together formed about 20% of the total IGHD genes (Table 3). The shortest and longest IGHD genes, IGHD7–27 and IGHD3–16, contributed to 7% and 3%, respectively. The IGHD repertoire was much larger and included some of the rare allelic variants such as IGHD3–3*02, IGHD3–10*02, and IGHD3–16*01, each appearing hundreds of times, although their existence was questioned as not seen previously (Boyd et al. 2010). Similarly, all six IGHJ genes were expressed with multiple allelic variants, with some restriction of IGHJ1. In the case of V-KAPPA and V-LAMBDA, all five IGKJ genes were utilized in kappa chains but only three IGLJ genes were mostly utilized for lambda chains. Overall, the gene usage in the present study was addressed to an extent that had not been previously recognized and which in the earlier studies of the human cord blood repertoire was limited by sampling statistics.

The length distributions of CDR observed for the V domains of heavy and light chains (Fig. 2) reflect the extensive paratope diversity which is comparable to that found using all antibody sequences in the NCBI GenBank, IMGT/LIGM-DB, Protein Data Bank, and IMGT/3Dstructure-DB (North et al. 2011; Zhao and Lu 2011). Further, to investigate how much of the V germline genes were diversified by somatic mutations, we estimated that 49%

of heavy and 66% of light chains had somatic hypermutations in the cord blood-derived IgM antibodies. We found that about 70% of heavy and light chains displayed between one and five AA changes in their V regions. Most of somatic AA changes were confined to the framework regions as the majority of CDR1 and CDR2 displayed no AA changes (Fig. 3). The average number of AA changes within the FR1, FR2, and FR3 (0–2%) and CDR1 and CDR2 (<1%), indicates that mutations leading to AA changes are rare in the cord blood IgM repertoire as previously reported (Mortari et al. 1993; Ridings et al. 1997; Williams et al. 2009). It should be noted that the average lengths of FR are significantly higher than the CDR and that could be taken into account when calculating the mutations rates. Also, one or two amino acid residues of FR1 were not normally counted as mutations if they were located at the N terminus and might have been introduced through the degenerate primers used in the PCR cloning process as observed in a previous study (Zhu et al. 2011). The somatic mutations observed to a lesser extent in the cord blood antibodies could have resulted from the exposure to self-antigens or placentally transferred antigens.

Although the IGHV gene usage contributes to the diversity of the repertoire, the major source of diversity arises from the VH CDR3 as it involves both combinatorial and junctional diversification. During the V-D-J rearrangement leading to VH CDR3, the junctional diversity is created by palindromic P-REGION, exonuclease trimming at the 3'V, 5'D, 3'D, and 5'J ends, and non-germline encoded at random nucleotide addition (N region; Fig. 4). The average number of N nucleotide additions in productive VH CDR3 from human cord blood is 8.98 ± 0.04 which is close to that observed in CDR3 from human fetus (8.8 ± 0.6), but less than in CDR3 from human adults (15.2 ± 0.8); the average number of nucleotides lost due to exonuclease trimming was 15.00 ± 0.05 in productive cord blood VH which is far less than found in human fetus and adult, approximately 25 nucleotides for each, as calculated using a total of 183 sequences in a previous study (Link et al. 2005). Therefore, smaller degree of exonuclease trimming in cord blood when compared to the adult having that in larger might compensate the decreased usage of IGHJ6 and reduced N addition. This could possibly explain why the VH CDR3 length distribution in the IgM cord blood repertoire was similar to that of adult repertoire. Besides, the diversified cord blood heavy chain repertoire consisting of 24,188 unique sequences analyzed in this study showed moderate junctional diversity preserving most of the IGHD gene germline identities (>98%).

The junctional diversification observed through V-D-J rearrangement, P/N addition, and end trimming creates the VH CDR3 diversity with different lengths and AA compositions (Fig. 5). A distribution of VH CDR3 lengths was found ranging from 4 to 32 amino acids which was similar to the VH CDR3 length diversity seen in the human adult IgM repertoire (Glanville et al. 2009). We also found a similar VH CDR3 length distribution among productive and unproductive cord blood HV repertoires though the latter had relatively increased average numbers of N nucleotide addition (10.98 ± 0.06) and nucleotide lost (16.44 ± 0.07). We noted that use of the longest IGHD3–16 and IGHJ6 genes was not restricted in the repertoire, and even the combination of these two longest IGHD and IGHJ genes accounted for 0.4% of the total D-J rearrangements (Table S2). The overall and position-specific AA frequencies along the VH CDR3 showed the characteristic amino acids typically observed in human antibodies including tyrosine, serine, glycine, alanine, and charged residues, arginine, and aspartic acid (Ehrenmann et al. 2010; Kirkham et al. 2003;

Link et al. 2005). Previously, structure-based conformation analyses of VH CDR3 loops were helpful in identifying the roles of amino acids based on their positions and in predicting the structural features from the VH CDR3 sequence (Lesk et al. 1998; Nakamura et al. 1999; Shirai et al. 1996; Tramontano et al. 1997). Position-specific AA frequencies of VH CDR3 with lengths of 13 amino acids the most represented length in this analysis and interestingly, also the basic CDR3 length in IMGT Collier de Perles (Lefranc et al. 2003, 2011a; Ruiz and Lefranc 2002) showed arginine and aspartic acid predominantly at positions 106 and 116 that could form a salt bridge at the base of the CDR3 loop; the carboxy termini of the VH CDR3 frequently contained phenylalanine, aspartic acid, and tyrosine at positions 115, 116, and 117, which might be due to the predominant use of the IGHJ4 in the repertoire. The other common amino acids such as tyrosine, serine, glycine, and alanine occurred in the CDR3 with varied composition. From this study, VH CDR3 lengths appear not to be restricted in neonates but AA composition may not be similar due to differences in V-D-J rearrangements, P regions, end trimming, and N regions when compared to the adult repertoire.

Out of 1,430 V-D-J rearrangement patterns observed in the cord blood repertoire, 54 dominant V-D-J patterns appeared having high sequence abundance, each appearing between 101 and 670 instances (Table S3). In a recent study of zebrafish antibody maturation, a similar observation was found in which the antibody repertoire was stereotyped with a few dominant V-D-J rearrangements in early stages which decreased dramatically when fish matured (Jiang et al. 2011). However, most of the V-D-J patterns in the human cord blood had lower abundance indicating the diversity of V-D-J rearrangements. One of the most abundant V-D-J rearrangement pattern is IGHV1–2-IGHD6-IGHJ4. The overall contribution of IGHV1–2 gene was 20% in the cord blood IgM repertoire which decreased to 1.7% (Boyd et al. 2010) and 8% (Glanville et al. 2009) in adult IgM repertoires. Interestingly, the VH of one of the most broadly neutralizing anti-HIV antibodies, VRC01, was found to origin from the IGHV1–2 subgroup (Wu et al. 2011). It was recently discovered that the IGHV1–2 gene rearrangements produce polyreactive antibodies that react against self-antigens (Warsame et al. 2011) which might be the reason for the loss of IGHV1–2 genes in the adult repertoires developed through the maturation process. The other predominantly appearing IGHV genes included IGHV1–69, IGHV5–51, IGHV1–8, and IGHV4–59, each accounting for 5–10% of the total VH repertoire. All those IGHV genes appeared in the respective unproductive cord blood repertoire with similar V-D-J rearrangement patterns and frequencies although some restricted gene usages found in IGHD and IGHJ genes. The frequently occurring unproductive V-D-J rearrangements patterns exclusively utilized the longest IGHJ6 in combination with IGHD1, IGHD2, IGHD3, and IGHD6.

In conclusion, using 454 sequencing and IMGT/HighV-QUEST analysis of human cord blood repertoire, we found that most of the germline genes expressed in human cord blood cell repertoires are also found in the adult ones but with differences in the extent of gene utilization, somatic mutation, and VH CDR3 composition. The findings suggest that gene usage and VH CDR3 length may not be restricted, but V-D-J rearrangements, junctional diversity due to P region, exonuclease trimming and N addition, and the level of somatic hypermutation have definitive characteristics for the cord blood cell repertoires.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Laboratory of Molecular Technology of SAIC-Frederick Inc. for providing Roche 454 sequencing service. We are grateful to Eltaf Alamyar and to the IMGT® team for providing access to IMGT/HighV-QUEST. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research, and by Federal funds from the NIH, National Cancer Institute, under contract no. NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the US Government.

References

- Alamyar E, Giudicelli V, Duroux P, Lefranc MP (2010) IMGT/HighV-QUEST: a high-throughput system and web portal for the analysis of rearranged nucleotide sequences of antigen receptors - High-throughput version of IMGT/V-QUEST 11èmes Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), 7–9 9 2010 Montpellier, France
- Alamyar E, Giudicelli V, Duroux P, Lefranc MP (2011) IMGT/HighV-QUEST 2011. 12èmes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM), 28 June–1 7 2011 Paris, France
- Bauer K, Zemlin M, Hummel M, Pfeiffer S, Devers S, Zemlin C, Stein H, Versmold HT (2001) The diversity of rearranged immunoglobulin heavy chain variable region genes in peripheral blood B cells of preterm infants is restricted by short third complementarity-determining regions but not by limited gene segment usage. *Blood* 97:1511–1513 [PubMed: 11222402]
- Bobrzynski T, Fux M, Vogel M, Stadler MB, Stadler BM, Miescher SM (2005) A high-affinity natural autoantibody from human cord blood defines a physiologically relevant epitope on the Fc epsilon RI alpha. *J Immunol* 175:6589–6596 [PubMed: 16272313]
- Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL, Fire AZ (2009) Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. *Sci Trans Med* 1 (12):12ra23
- Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL, Collins AM (2010) Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* 184:6986–6992 [PubMed: 20495067]
- Brochet X, Lefranc MP, Giudicelli V (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* 36: W503–W508 [PubMed: 18503082]
- Casali P, Schettino EW (1996) Structure and function of natural antibodies. *Immunol Silicones* 210:167–179
- Chen C, Nagy Z, Prak EL, Weigert M (1995) Immunoglobulin heavy chain gene replacement: a mechanism of receptor editing. *Immunity* 3:747–755 [PubMed: 8777720]
- Chen W, Zhu Z, Xiao X, Dimitrov DS (2009) Construction of a human antibody domain (VH) library. *Methods Mol Biol* 525:81–99, xiii [PubMed: 19252835]
- Collins AM, Wang Y, Jackson KJ, Gaeta B, Pomat W, Siba P, Sewell WA (2011) Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* 63:259–265 [PubMed: 21249354]
- Cooke WD, Orr AS, Wiseman BL, Rouse SB, Murray WC, Ranck SG (1993) Human cord-blood contains an Igm antibody to the 41kd flagellar antigen of *Borrelia burgdorferi*. *Scand J Immunol* 38:407–409 [PubMed: 8211003]
- Dariavach P, Lefranc G, Lefranc MP (1987) Human immunoglobulin C lambda 6 gene encodes the Kern+Oz-lambda chain and C lambda 4 and C lambda 5 pseudogenes. *Proc Natl Acad Sci USA* 84:9074–9078 [PubMed: 3122211]

- Dimitrov DS (2010) Therapeutic antibodies, vaccines and antibodyomes. *Mabs* 2:347–356 [PubMed: 20400863]
- Ehrenmann F, Kaas Q, Lefranc MP (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res* 38:D301–D307 [PubMed: 19900967]
- Feeney AJ, Lugo G, Escuro G (1997) Human cord blood kappa repertoire. *J Immunol* 158:3761–3768 [PubMed: 9103441]
- Fischer N (2011) Sequencing antibody repertoires: the next generation. *Mabs* 3:17–20 [PubMed: 21099370]
- Fischer N, Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, Magistrelli G, Farinelli L, Kosco-Vilbois MH (2010) By-passing in vitro screening-next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res* 38(21):e193 [PubMed: 20846958]
- Giudicelli V, Lefranc MP (2011) IMGT/junctionanalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* 2011:716–725 [PubMed: 21632777]
- Giudicelli V, Chaume D, Lefranc MP (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res* 33:D256–D261 [PubMed: 15608191]
- Giudicelli V, Brochet X, Lefranc MP (2011) IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* 2011:695–715 [PubMed: 21632778]
- Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GM, Cox D, Rajpal A, Pons J (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A* 106:20216–20221 [PubMed: 19875695]
- Jiang N, Weinstein JA, Penland L, White RA 3rd, Fisher DS, Quake SR (2011) Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc Natl Acad Sci U S A* 108:5348–5353 [PubMed: 21393572]
- Kirkham PM, Zemlin M, Klinger M, Link J, Zemlin C, Bauers K, Engler JA, Schroeder HW (2003) Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J Mol Biol* 334:733–749 [PubMed: 14636599]
- Kolar GR, Yokota T, Rossi MID, Nath SK, Capra JD (2004) Human fetal, cord blood, and adult lymphocyte progenitors have similar potential for generating B cells with a diverse immunoglobulin repertoire. *Blood* 104:2981–2987 [PubMed: 15256421]
- Koralov SB, Novobrantseva TI, Konigsmann J, Ehlich A, Rajewsky K (2006) Antibody repertoires generated by VH replacement and direct VH to JH joining. *Immunity* 25:43–53 [PubMed: 16860756]
- Lefranc MP (2011a) IMGT Collier de Perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* 2011:643–651 [PubMed: 21632788]
- Lefranc MP (2011b) IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* 2011:633–642 [PubMed: 21632789]
- Lefranc MP, Lefranc G (2001) *The immunoglobulin FactsBook*. Academic, London
- Lefranc MP, Pommie C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* 27:55–77 [PubMed: 12477501]
- Lefranc MP, Pommie C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, Foulquier E, Thouvenin V, Lefranc G (2005) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* 29:185–203 [PubMed: 15572068]

- Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, Regnier L, Ehrenmann F, Lefranc G, Duroux P (2009) IMGT, the International ImmunoGeneTics Information System. *Nucleic Acids Res* 37:D1006–D1012 [PubMed: 18978023]
- Lesk AM, Morea V, Tramontano A, Rustici M, Chothia C (1998) Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J Mol Biol* 275:269–294 [PubMed: 9466909]
- Link JM, Larson JE, Schroeder HW (2005) Despite extensive similarity in germline DH and JH sequence, the adult *Rhesus macaque* CDR-H3 repertoire differs from human. *Mol Immunol* 42:943–955 [PubMed: 15829286]
- Loffert D, Ehlich A, Muller W, Rajewsky K (1996) Surrogate light chain expression is required to establish immunoglobulin heavy chain allelic exclusion during early B cell development. *Immunity* 4:133–144 [PubMed: 8624804]
- Messmer BT, Sullivan JJ, Chiorazzi N, Rodman TC, Thaler DS (1999) Two human neonatal IgM antibodies encoded by different variable-region genes bind the same linear peptide: Evidence for a stereotyped repertoire of epitope recognition. *J Immunol* 162:2184–2192 [PubMed: 9973494]
- Mortari F, Newton JA, Wang JY, Schroeder HW Jr (1992) The human cord blood antibody repertoire. Frequent usage of the VH7 gene family. *Eur J Immunol* 22:241–245 [PubMed: 1730251]
- Mortari F, Wang JY, Schroeder HW (1993) Human cord blood antibody repertoire. Mixed population of Vh gene segments and Cdr3 distribution in the expressed C-alpha and C-gamma repertoires. *J Immunol* 150:1348–1357 [PubMed: 8381834]
- Mouthon L, Nobrega A, Nicolas N, Kaveri SV, Barreau C, Coutinho A, Kazatchkine MD (1995) Invariance and restriction toward a limited set of self-antigens characterize neonatal IgM antibody repertoires and prevail in autoreactive repertoires of healthy adults. *Proc Natl Acad Sci U S A* 92:3839–3843 [PubMed: 7731992]
- Nakamura N, Shirai H, Kidera A (1999) H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Lett* 455:188–197 [PubMed: 10428499]
- Nemazee D (2006) Receptor editing in lymphocyte development and central tolerance. *Nat Rev Immunol* 6:728–740 [PubMed: 16998507]
- North B, Lehmann A, Dunbrack RL Jr (2011) A new clustering of antibody CDR loop conformations. *J Mol Biol* 406:228–256 [PubMed: 21035459]
- Persson MA (2009) Twenty years of combinatorial antibody libraries, but how well do they mimic the immunoglobulin repertoire? *Proc Natl Acad Sci U S A* 106:20137–20138 [PubMed: 19934042]
- Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc MP (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* 17:17–32 [PubMed: 14872534]
- Richl P, Stern U, Lipsky PE, Girschick HJ (2008) The lambda gene immunoglobulin repertoire of human neonatal B cells. *Mol Immunol* 45:320–327 [PubMed: 17675156]
- Ridings J, Nicholson IC, Goldsworthy W, Haslam R, Robertson DM, Zola H (1997) Somatic hypermutation of immunoglobulin genes in human neonates. *Clin Exp Immunol* 108:366–374 [PubMed: 9158112]
- Rodman TC, Lutton JD, Jiang SL, Al-Kouatly HB, Winston R (2001) Circulating natural IgM antibodies and their corresponding human cord blood cell-derived Mabs specifically combat the Tat protein of HIV. *Exp Hematol* 29:1004–1009 [PubMed: 11495706]
- Ruiz M, Lefranc MP (2002) IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics* 53:857–883 [PubMed: 11862387]
- Schroeder HW Jr, Hillson JL, Perlmutter RM (1987) Early restriction of the human antibody repertoire. *Science* 238:791–793 [PubMed: 3118465]
- Schroeder HW, Shiokawa S, Mortari F, Lima JO, Nunez C, Bertrand FE, Kirkham PM, Zhu SG, Dasanayake AP (1999) IgM heavy chain complementarity-determining region 3 diversity is constrained by genetic and somatic mechanisms until two months after birth. *J Immunol* 162:6060–6070 [PubMed: 10229847]
- Shirai H, Kidera A, Nakamura H (1996) Structural classification of CDR-H3 in antibodies. *FEBS Lett* 399:1–8 [PubMed: 8980108]

- Tonegawa S (1983) Somatic generation of antibody diversity. *Nature* 302:575–581 [PubMed: 6300689]
- Tramontano A, Morea V, Rustici M, Chothia C, Lesk AM (1997) Antibody structure, prediction and redesign. *Biophys Chem* 68:9–16 [PubMed: 9468606]
- Warsame AA, Aasheim HC, Nustad K, Troen G, Tierens A, Wang V, Randen U, Dong HP, Heim S, Brech A, Delabie J (2011) Splenic marginal zone lymphoma with VH1–02 gene rearrangement expresses poly- and self-reactive antibodies with similar reactivity. *Blood* 118:3331–3339 [PubMed: 21725051]
- Weinstein JA, Jiang N, White RA 3rd, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324:807–810 [PubMed: 19423829]
- Williams JV, Weitkamp JH, Blum DL, LaFleur BJ, Crowe JE Jr (2009) The human neonatal B cell response to respiratory syncytial virus uses a biased antibody variable gene repertoire that lacks somatic mutations. *Mol Immunol* 47:407–414 [PubMed: 19804909]
- Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, Chen X, Longo NS, Louder M, McKee K, O'Dell S, Peretto S, Schmidt SD, Shi W, Wu L, Yang Y, Yang ZY, Yang Z, Zhang Z, Bonsignori M, Crump JA, Kapiga SH, Sam NE, Haynes BF, Simek M, Burton DR, Koff WC, Doria-Rose N, Connors M, Mullikin JC, Nabel GJ, Roederer M, Shapiro L, Kwong PD, Mascola JR (2011) Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science*. doi:10.1126/science.1207532
- Yousfi Monod M, Giudicelli V, Chaume D, Lefranc MP (2004) IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONS. *Bioinformatics* 20(Suppl 1):i379–i385 [PubMed: 15262823]
- Zhao S, Lu J (2011) A bioinformatics pipeline to build a knowledge database for in silico antibody engineering. *Mol Immunol* 48:1019–1026 [PubMed: 21310488]
- Zhu Z, Dimitrov DS (2009) Construction of a large naive human phage-displayed Fab library through one-step cloning. *Methods Mol Biol* 525:129–142, xv [PubMed: 19252833]
- Zhu Z, Qin HR, Chen W, Zhao Q, Shen X, Schutte R, Wang Y, Ofek G, Streaker E, Prabakaran P, Fouda GG, Liao HX, Owens J, Louder M, Yang Y, Klaric KA, Moody MA, Mascola JR, Scott JK, Kwong PD, Montefiori D, Haynes BF, Tomaras GD, Dimitrov DS (2011) Cross-reactive HIV-1-neutralizing human monoclonal antibodies identified from a patient with 2F5-like antibodies. *J Virol* 85:11401–11408 [PubMed: 21880764]

Fig. 1.

Gene usage frequencies observed for the VH and VL domains of IG or antibody cDNA from human cord blood samples of two babies (CB1 and CB2). Gene usage was calculated as the percentage of the total unique population of productive (P) and unproductive (UP) sequences. **a–c** VH domains: IGHV subgroup, IGHD set, and IGHJ gene utilization observed in productive and unproductive sequences. **d–f** VL domains: IGKV and IGLV subgroups as well as IGKJ and IGLJ gene utilization in productive and unproductive sequences

Fig. 2.

Length distributions of the complementarity determining regions CDR1 and CDR2 observed in the V regions of the VH, V-KAPPA, and V-LAMBDA domains. CDR lengths and number of sequences in percent are shown in the horizontal and vertical axes, respectively. **a** Length distributions of VH CDR1 and CDR2. **b** Length distributions of V-KAPPA CDR1 and CDR2. **c** Length distributions of V-LAMBDA CDR1 and CDR2

Fig. 3.

Number of amino acid (AA) changes from the closest germline counterparts found in the combined unique V region sequences of 13,888 VH, 12,164 V-KAPPA, and 3,948 V-LAMBDA, obtained from two human cord blood samples. **a** Total number of AA changes in the V regions of VH and VL domains. **b** Number of AA changes observed in the CDR1 and CDR2 of VH, V-KAPPA, and V-LAMBDA domains. **c** The average number of AA changes observed in FR1, CDR1, FR2, CDR2, and FR3 of VH, V-KAPPA, and V-LAMBDA

Fig. 4.

N, P regions and trimming at the variable heavy chain complementarity determining region V_H CDR3. The V-D-J junctional diversity created by P region, exonuclease trimming and N region addition significantly contributes to a large V_H CDR3 repertoire. The histograms show the average numbers of N and P nucleotides added at the V-D-J junctions, and the average numbers of nucleotides deleted at the 3'V, 5'D, 3'D and 5'J by exonuclease trimming

Fig. 5. CDR3 lengths in VH and VL domains, and amino acid usage observed in the cord blood-derived VH CDR3 regions. **a** Length distribution of 2,927 unique V-KAPPA CDR3 regions. **b** Length distribution of 1,193 unique V-LAMBDA CDR3 regions. **c** Length distribution of 24,188 unique VH CDR3 regions showing the highly diversified CDR3 repertoire with different lengths. **d** Average values of AA composition calculated for all unique VH CDR3 sequences in percent shows the preferred AA usage. The amino acids are arranged by relative hydrophobicity values as found in Kyte-Doolittle scale (Pommié et al. 2004). **e** The frequency of individual amino acids at specific positions for the most prevalent VH CDR3 of 2,963 sequences having the length of 13 amino acids (the basic length of CDR3-IMGT) is shown. CDR3 positions are shown according to the IMGT unique numbering (Lefranc et al. 2003)

Fig. 6. Frequency of V-D-J unique rearrangements derived from the same IGFCV, IGHD and IGHJ genes and expressed in the cord blood VH repertoire. The A-axis represents the IGFCV genes used in association with pertinent IGFCJ and IGHJ genes which are represented by the left and right y-axes, respectively. A total of 1,430 unique V-D-J rearrangement patterns that share the same IGFCV, IGHD and IGFCJ genes are shown in different shapes indicating the frequency count for each V-D-J rearrangement pattern observed in the cord blood VH repertoire

Table 1

The number of unique sequence data derived from the 454 sequencing and IMG/HighV-QUEST analysis of cord blood IgM libraries

IgM chains	Total V-D-J length>300 nt	Translated V-D-J AA sequences	V region AA sequences	CDR3 AA sequences
IGH	28,825 (13,890)	28,169 (13,803)	13,888 (11,379)	24,188 (12,951)
IGK	25,389 (9,355)	18,928 (8,674)	12,164 (6,731)	2,927 (3,149)
IgL	8,828 (4,873)	5,602 (4,258)	3,948 (3,523)	1,193 (1,220)
Total	63,042 (28,118)	52,699 (26,735)	30,000 (21,633)	28,308 (17,320)

Values within parentheses represent the number of unproductive sequences observed
m Nucleotide, *AA* amino acid, *IGH* heavy chain, *IGK* light kappa chain, *IgL* light lambda chain

Table 2

The frequency distribution of IGHV, IGKV and IGLV genes from VH, V-KAPPA and V-LAMBDA domains, as observed in human cord blood repertoire

IGHV genes	Percentage	IGKV genes	Percentage	IGLV genes	Percentage
IGHV1-2	20.19	IGKV3-20	23.73	IGLV2-14	13.96
IGHV1-69	9.20	IGKV3-11	9.92	IGLV1-40	12.21
IGHV5-51	6.85	IGKV2-28	7.32	IGLV3-1	11.94
IGHV1-8	6.52	IGKV4-1	6.41	IGLV3-21	11.09
IGHV4-59	6.26	IGKV3-15	6.09	IGLV3-19	9.14
IGHV4-34	5.21	IGKV1-39	5.39	IGLV1-44	8.66
IGHV1-18	5.09	IGKV3D-20	4.25	IGLV2-8	6.60
IGHV3-23	4.94	IGKV1-33	3.98	IGLV1-47	6.16
IGHV4-61	3.11	IGKV1-5	3.39	IGLV1-51	4.96
IGHV3-30	3.10	IGKV3-NL5	3.24	IGLV2-11	3.11
IGHV4-39	2.72	IGKV2-29	2.70	IGLV2-23	2.25
IGHV4-4	2.68	IGKV1-27	2.66	IGLV6-57	1.80
IGHV1-46	2.23	IGKV2-30	2.55	IGLV1-50	1.64
IGHV3-48	2.13	IGKV1-13	2.07	IGLV8-61	1.00
IGHV3-21	1.97	IGKV2D-29	1.84	IGLV3-25	0.96
IGHV3-33	1.96	IGKV2-24	1.55	IGLV7-46	0.89
IGHV3-7	1.82	IGKV1-9	1.53	IGLV2-18	0.57
IGHV3-11	1.35	IGKV3D-15	1.45	IGLV9-49	0.55
IGHV3-9	1.26	IGKV1-8	1.33	IGLV7-43	0.50
IGHV1-3	1.25	IGKV1-16	1.27	IGLV1-36	0.48
IGHV4-b	1.16	IGKV1-12	1.22	IGLV10-54	0.37
IGHV3-66	1.07	IGKV6-21	1.11	IGLV3-10	0.23
IGHV1-24	1.04	IGKV1-17	1.02	IGLV3-27	0.16
IGHV3-15	0.91	IGKV5-2	0.61	IGLV1-41	0.11
IGHV6-1	0.80	IGKV1D-17	0.45	IGLV3-16	0.11
IGHV4-31	0.70	IGKV1D-13	0.44	IGLV5-39	0.11
IGHV3-74	0.60	IGKV1D-16	0.39	IGLV2-5	0.09
IGHV4-30-2	0.54	IGKV2-40	0.34	IGLV5-37	0.09
IGHV3-53	0.48	IGKV1-6	0.30	IGLV3-9	0.07
IGHV3-73	0.36	IGKV3D-7	0.27	IGLV4-3	0.05
IGHV3-30-3	0.33	IGKV3-NL1	0.25	IGLV3-12	0.04
IGHV3-13	0.31	IGKV3-NL3	0.22	IGLV5-45	0.04
IGHV7-4-1	0.30	IGKV1-37	0.11	IGLV5-52	0.04
IGHV3-49	0.28	IGKV1D-43	0.08	IGLV3-22	0.02
IGHV3-64	0.26	IGKV1D-8	0.08		
IGHV1-58	0.24	IGKV2D-26	0.08		
IGHV3-43	0.14	IGKV2D-30	0.07		
IGHV3-72	0.14	IGKV3D-11	0.07		

IGHV genes	Percentage	IGKV genes	Percentage	IGLV genes	Percentage
IGHV5-a	0.13	IGKV2D-24	0.06		
IGHV4-30-4	0.11	IGKV6D-41	0.04		
IGHV3-20	0.07	IGKV7-3	0.03		
IGHV3-h	0.06	IGKV3-NL4	0.03		
IGHV2-5	0.05	IGKV1D-12	0.02		
IGHV1-45	0.02	IGKV2-18	0.02		
IGHV2-70	0.02	IGKV1-NL1	0.01		
IGHV2-26	0.02				
IGHV4-28	0.01				
IGHV1-f	0.01				
IGHV1-c	<0.01				
IGHV3-22	<0.01				
IGHV3-35	<0.01				
IGHV3-d	<0.01				

Table 3

Frequency distribution of IGHD genes observed in the cord blood VH repertoire

IGHD genes	Percentage
IGHD1-1	2.35
IGHD1-14	1.08
IGHD1-20	0.21
IGHD1-26	9.56
IGHD1-7	2.14
IGHD2-15	4.32
IGHD2-2	4.69
IGHD2-21	3.40
IGHD2-8	2.43
IGHD3-10	7.69
IGHD3-16	3.20
IGHD3-22	6.61
IGHD3-3	3.80
IGHD3-9	2.19
IGHD4-17	3.06
IGHD4-23	2.68
IGHD4-4	1.42
IGHD5-12	3.30
IGHD5-24	2.14
IGHD5-5	3.47
IGHD6-13	9.78
IGHD6-19	5.85
IGHD6-25	0.49
IGHD6-6	6.71
IGHD7-27	7.44