



King's Research Portal

DOI: 10.3389/fimmu.2016.00546

Document Version Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):

Martin, V. G., Wu, Y. C. B., Townsend, C. L., Lu, G. H. C., O'Hare, J. S., Mozeika, A., Coolen, A. C. C., Kipling, D., Fraternali, F., & Dunn-Walters, D. K. (2016). Transitional B cells in early human B cell development - Time to revisit the paradigm? *Frontiers in Immunology*, *7*(DEC), Article 546. https://doi.org/10.3389/fimmu.2016.00546

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research. •You may not further distribute the material or use it for any profit-making activity or commercial gain •You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Transitional B cells in early human B cell development - time to revisit the paradigm?

Victoria G. Martin³, Yu-Chang B. Wu², Catherine L. Townsend³, Grace H. Lu², Joselli S. O'Hare¹, Alexander Mozeika⁴, Anthonius C. Coolen⁴, David Kipling⁵, Franca Fraternali^{2, 4}, Deborah K. Dunn-Walters^{1, 3*}

¹Faculty of Health & Medical Sciences, University of Surrey, United Kingdom, ²Faculty of Life Sciences & Medicine, King's College London, United Kingdom, ³Faculty of Life Sciences & Medicine, King's College London, United Kingdom, ⁴Faculty of Life Sciences & Medicine, King's College London, United Kingdom, ⁵School of Medicine, Cardiff University, United Kingdom

Submitted to Journal: Frontiers in Immunology

Specialty Section: B Cell Biology

ISSN: 1664-3224

Article type: Original Research Article

Received on: 17 Aug 2016

Accepted on: 16 Nov 2016

Provisional PDF published on: 16 Nov 2016

Frontiers website link: www.frontiersin.org

Citation:

Martin VG, Wu YB, Townsend CL, Lu GH, O_hare JS, Mozeika A, Coolen AC, Kipling D, Fraternali F and Dunn-walters DK(2016) Transitional B cells in early human B cell development - time to revisit the paradigm?. *Front. Immunol.* 7:546. doi:10.3389/fimmu.2016.00546

Copyright statement:

© 2016 Martin, Wu, Townsend, Lu, O_hare, Mozeika, Coolen, Kipling, Fraternali and Dunn-walters. This is an open-access article distributed under the terms of the <u>Creative Commons Attribution</u> <u>License (CC BY)</u>. The use, distribution and reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms. This Provisional PDF corresponds to the article as it appeared upon acceptance, after peer-review. Fully formatted PDF and full text (HTML) versions will be made available soon.

Frontiers in Immunology | www.frontiersin.org



1 Title: Transitional B cells in early human B cell development – time to revisit the paradigm?

- Martin, V.1*, Wu, Y.-C2.*, Townsend, C.1, Lu, H-C.2, Silva O'Hare3, J., Mozeika, A.4, Coolen, A.4, 3 Kipling, D.⁵, Fraternali, F.^{2,4}, Dunn-Walters, D.K.^{1,3} 4
- 5 Sort title: Early human B cell development
- 6 Key words
- 7 Bone marrow, human, B cell development, transitional, regulatory B cells
- 8 ¹ Division of Infection, Immunity and Inflammatory Disease, Faculty of Life Sciences & Medicine,
- King's College London, London, United Kingdom 9
- 10 ²Randall Division of Cell and Molecular Biophysics, Faculty of Life Sciences & Medicine, King's College London, London, United Kingdom 11
- 12 ³ School of Biosciences and Medicine, Faculty of Health and Medical Sciences, University of Surrey, 13 Guildford, Surrey, United Kingdom.
- 14 ⁴ Institute for Mathematical and Molecular Biomedicine, Faculty of Life Sciences & Medicine, King's
- College London, London, United Kingdom 15
- 16 ⁵ Institute of Cancer & Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom ovi 1510100
- 17 *These authors contributed equally to this work.

18 Abstract

19 The B cell repertoire is generated in the adult bone marrow by an ordered series of gene rearrangement 20 processes that result in massive diversity of immunoglobulin (Ig) genes, and consequently an equally large number of potential specificities for antigen. As the process is essentially random, then cells 21 22 exhibiting excess reactivity with self-antigens are generated and need to be removed from the repertoire 23 before the cells are fully mature. Some of the cells are deleted, and some will undergo receptor editing 24 to see if changing the light chain can rescue an autoreactive antibody. As a consequence, the binding 25 properties of the B cell receptor are changed as development progresses through pre-26 B>>immature>>transitional>>naïve phenotypes. Using long-read, high-throughput, sequencing we 27 have produced a unique set of sequences from these four cell types in human bone marrow and matched 28 peripheral blood and our results describe the effects of tolerance selection on the B cell repertoire at the 29 Ig gene level. Most strong effects of selection are seen within the heavy chain repertoire, and can be 30 seen both in gene usage and in CDR-H3 characteristics. Age-related changes are small and only the 31 size of the CDR-H3 shows constant and significant change in these data. The paucity of significant changes in either kappa or lambda light chain repertoires implies that either the heavy chain has more 32 33 influence over autoreactivity than light chain and/or that switching between kappa and lambda light 34 chains, as opposed to switching within the light chain loci, may effect a more successful autoreactive 35 rescue by receptor editing. Our results show that the transitional cell population contains cells other 36 than those that are part of the pre-B>>immature>>transitional>>naïve development pathway, since the 37 population often shows a repertoire that is outside the trajectory of gene loss/gain between pre-B and 38 naïve stages.

39 Introduction

B cells development starts in the bone marrow, from a hematopoietic stem cell precursor, and 40 41 undergoes an ordered series of differentiation steps to ultimately generate mature naïve B cells in the 42 peripheral blood (1). As development progresses the B cell receptor (BCR) is generated and adjusted to ensure that cells are not overly autoreactive. First, at the initial pro-B cell stage heavy chain gene 43 44 recombination occurs, such that the random selection and joining of IGHV, IGHD and IHGJ genes 45 produces a complete heavy chain. As cells develop into pre-B cells the heavy chain is then presented 46 on the surface of the cell, in conjunction with a surrogate light chain, so that selection of productive 47 heavy chains can take place. Cells without a productive heavy chain gene rearrangement are removed 48 from the repertoire, whilst cells containing productive heavy chains undergo a few rounds of 49 proliferation and are designated 'large' pre-B cells (2). After this point light chain recombination of IGK or IGL genes occurs within each cell in order to produce cells with rearranged heavy (IgM) and 50 51 light chain genes (3-5). Expression of the complete antibody on the surface on these immature B cells 52 enables the first tolerance checkpoint such that some cells carrying receptors with too high an affinity

53 for self-antigens undergo receptor editing to change the light chains (6). Lack of a functional 54 surrogate light chain somehow interferes with this tolerance checkpoint (7). It has been shown that 55 55.2% (n= 29) of early immature B cells carried polyreactive immunoglobulin genes, and this was 56 reduced by receptor editing, or deletion from the repertoire, so that only 7.4% (n= 72) of transitional 57 cells exiting the bone marrow carried polyreactive antibodies (8). The term "Transitional cells" was 58 originally coined to categorise the group of early emigrant cells from the bone marrow. These cells express IgD and CD10 alongside the IgM BCR so can be identified as IgD⁺ CD27⁻CD10^{hi/+} (9). Co-59 60 expression of high levels of CD24 and CD38 have also frequently been used to identify them, and it is important that CD27 be included if this is the case since the CD38^{hi}CD24^{hi} population can contain 61 62 CD27⁺ cells that may be more akin to the IgM memory populations (10). Heterogeneity has been 63 seen within transitional cells such that T1 (CD38+++CD24hiCD10++IgD^{lo/-}), T2 64 (CD38⁺⁺CD24^{hi}CD10⁺IgD⁺) and T3 (CD38⁺CD24⁺IgD⁺ABCB1⁻) subpopulations have been identified 65 (9, 11, 12). T1 cells have been shown to be highly prone to spontaneous apoptosis and are hard to rescue even with BCR or T cell stimulation (13, 14), thereby providing another opportunity for 66 67 negative selection during tolerance and removal of autoimmunity (8, 15). T2 cells were thought to be 68 less responsive to negative selection and more responsive to antigen stimulation allowing for positive selection to occur (13, 14, 16, 17). The functional classification of CD38^{hi}CD24^{hi}cells as transitional 69 70 cell intermediates between bone marrow and peripheral naïve B cells in development has also been complicated by the discovery of human regulatory B cells (Bregs), which are also CD38^{hi}CD24^{hi} (18) 71 72 In humans, the gradual loss of CD10, CD5 and IgM and the upregulation of CD22, CD44, CD21 and 73 CD23 as cells progress from immature to transitional (TI to T2 to T3) to mature naïve cells, along with 74 the generation of naïve cells from stimulated transitional cells (9, 19), lead to the current paradigm: 75 That B cells develop from pre-B cells through immature cells in the bone marrow to transitional cells 76 in the periphery and then to peripheral naïve cells in a linear pathway (20). 77 Positive and negative selection events that occur in B cell development are expected to shape the

repertoire of B cell populations in terms of in terms of V, D, J gene usage and CDRH3 properties. We

79 have previously shown that different stages of memory B cell development have distinct repertoire

80 characteristics (21-23). Notably an increase in IGHV3 family at the expense of IGHV1 family in IgM 81 memory cells, but not switched memory cells (21) has been seen, and a decrease in the overall CDR3 82 length which is partially (but not wholly) caused by an increase of *IGHJ4* family usage at the expense 83 of IGHJ6 family usage is observed in memory cells in general (21-25). The selection events that 84 occur during central and peripheral tolerance will shape the immunoglobulin repertoire due to the 85 removal of unwanted autoreactive cells. Comparison between passenger out-of-frame 86 immunoglobulin genes and in-frame immunoglobulin genes in human naïve cells indicates that B cell 87 selection has already occurred before exogenous antigen activation (26). Cloning of up to 131 Ig 88 genes from pre-B, immature and mature B cell subsets indicates there may be differences in CDRH3 89 characteristics due to negative selection processes (27). However, little information is available on the 90 expressed immunoglobulin repertoire as a whole in the early stages of development in the human 91 bone marrow. Here we have used high throughput sequencing to define the heavy and light chain B 92 cell repertoire in pre-B and immature cells from human bone marrow, alongside donor-matched 93 transitional and naïve B cells from the peripheral blood, to provide an overall picture of the 94 consequences of early selection events on human B cell repertoire.

96 Methods

97 Sample collection

Bone marrow and peripheral blood was obtained from 19 healthy adult donors (aged 24-86) with no
known disease affecting the immune system and undergoing total hip replacement surgery at Guy's
Hospital, London, UK. The samples were collected with informed consent under the REC number
11/LO/1266.

102

103 **B cell isolation and sorting**

104 The B cells were isolated and sorted as previously published (28). Briefly, bone marrow (BM)

105 material was removed from the head of the femur and filtered into RPMI-1640 (Sigma Aldrich). BM

106 mononuclear cells (BMMCs) and peripheral blood mononuclear cells (PBMCs) were isolated using

107 Ficoll-Paque PLUS (GE Healthcare Life Sciences) according to the manufacturer's instructions. For

the BMMCs, CD19⁺ B cells were then enriched to >98% using CD19 microbead magnetic separation
(Miltenyi).

110 BMMCs were stained using PE anti-human Ig light chain lambda (MHL-38, BioLegend), APC anti-

111 human Ig light chain kappa (MHK-49, BioLegend), PE/Cy7 anti-human CD38 (HIT2,

112 BioLegend), PerCP/Cy5.5 anti-human IgD (IA6-2, BioLegend), Pacific Blue anti-human IgM

113 (MHM-88, BioLegend), APC/Cy7 anti-human CD10 (HI10a, BioLegend) and FITC CD27 (M-

114 T271, Miltenyi Biotec). PBMCs were stained using CD19 APC (HIB19, BD BioScience), IgD

115 PerCP/Cy5.5 (IA6-2, BioLegend), CD27 FITC (M-T271, Miltenyi Biotec) and CD10 APC/Cy7

- 116 (HI10a, BioLegend).
- 117 B cells were sorted into Sort Lysis Reverse Transcription (SLyRT)(21) buffer using the FACS Aria
- 118 (BD BioSciences). B cells were sorted into four cell types: large pre-B (IgK-IgL-CD38+IgM+),
- 119 immature (IgK⁺ or IgL⁺CD27⁻IgM⁺IgD⁻CD10⁺), transitional (IgD⁺CD27⁻CD10⁺) and naïve
- 120 (IgD⁺CD27⁻CD10⁻) as shown in Figure 1. Due to the lytic (RNA stabilising) nature of the sort buffer

and the rarity of some of the cell populations we were unable to check post-sorting purity. We set thecollection gates well away from the FMO control gates as a precautionary measure (Figure 1b,c).

123

124 High throughput sequencing and data clean up

High throughput sequencing was carried out as previously described (21, 29). Briefly, reverse
transcription was performed directly on the sample immediately after sorting and then a semi-nested
PCR was performed, adding multiplex identifiers (MIDs) to distinguish patients (29). High

128 throughput sequencing was carried out using the Roche 454 GS FLX system (LGC Genomics) and

129 data clean-up was performed as before Wu, Kipling (29). In addition, for analysis of the CDR3

130 peptide sequence character, the data was cleaned to remove sequences where the CDR3 was likely

131 inaccurate as a result of sequencing error, i.e. CDR3 regions outside the normal distribution of CDR3

132 lengths (1 to 35 amino acids for heavy chain and 1 to 20 amino acids for light chain) and/or sequences

133 identified by IMGT as unproductive.

134 V(D)J gene assignment was carried out using IMGT/HighV-QUEST (30, 31). The physicochemical

properties of the CDR3 amino acid sequences were calculated using the R package Peptides (32, 33)

136 and clustering analysis of the Ig gene sequences was carried out using levenstein distance on the

137 CDR3 regions using R scripts available on our website (34).

138 As all of the repertoires were antigen-naïve, then true clonal expansions would be negligible.

139 Therefore, in order to remove biases caused by PCR amplification, only unique gene rearrangements

140 were used for this analysis. Where the clustering identified more than one related sequence a modal

sequence was used to represent the gene rearrangement. The data was stored in CSV files and data

analysis was performed using Microsoft Excel, GraphPad Prism and R.

143

144 Analysis and Statistics

145 Frequency of gene usage in the repertoire

The frequency of each gene (both at the individual gene and at the gene family level) observed in the data was calculated for each cell subset from each donor. The frequency (in percentage) of each VDJ family combination (heavy chain), or VJ family combination (light chain) was also calculated for each cell subset from each donor. The mean values of gene combination frequencies were calculated for each cell subset and 3D bubble plots were created using the R package *plot3D* (35). Statistical analysis (Mann-Whitney Wilcoxon test and ANOVA, with post-test analysis where appropriate) was performed using R or GraphPad Prism.

153

Physicochemical properties of CDR3 regions

154 The physicochemical properties of CDR3 regions at heavy and light chains were compared between 155 different cell types. These properties consisted of length, hydrophobicity indicated by GRAVY index 156 (36), Boman index (37), molecular weight (Mr), Isoelectric point (pI) (38), aliphatic index (39), 157 frequency of amino acid classes in the CDR3 region, and Kidera factors (40). The R package lem4 158 (41) was used for fitting and analysing the mixed model of our data, describing the fixed-effect (cell 159 types) and the random-effect (patients) in a linear predictor expression. The Likelihood Ratio Test 160 was calculated with the statistical method ANOVA to estimate the statistical significance between 161 populations, i.e. a pair of cell subsets.

162 163

Clustering and Principal Component Analysis

Principal Component Analysis (PCA) and clustering, using Minkowski distance, were applied to the 164 165 kidera factors and gene usage frequencies from the CDR3 data as follows. Firstly, the mean values of 166 the kidera factors and gene usage frequencies were computed for each donor. Secondly, the mean 167 values and frequencies of all donors were grouped and then analysed by PCA and clustering. 168 PCA was performed using the prcomp function in R. The Minkowski distances (with power of 4) 169 were calculated using *dist(method="minkowski")* function in R based on all CDR3 properties. The 170 distances were then plotted with dendrograms (trees) using the *dendrapply* function in R. 171 Randomised datasets were generated by randomly shuffling the sequences across four cell subpopulations. PCA analysis was then performed to be compared with the original dataset in order to 172 173 show that our observations of differences between cell sub-populations were not random events.

174

175 Mass Cytometry

176 PBMCs were stained with FITC anti-human CD14 and APC anti-human CD3 (clone M5E2 and 177 UCHT1 respectively), and a population of enriched B cells (CD3⁻CD14⁻) was collected into 50% FCS 178 (Biosera) and 50% RPMI-1640 (Gibco). The CD3⁻CD14⁻ enriched B cells were labelled with a 179 rhodium intercalator (Rh103, DVS Sciences) followed by intracellular and extracellular staining with 180 a panel of 30 different metal-tagged antibodies (DVS Sciences, BD BioSciences and BioLegend). 181 Cells were fixed, iridium stained (Ir193, DVS Sciences), and normalization beads (DVS Sciences) 182 were added before analysis on the mass cytometry system (DVS Sciences). Between 1 and 5×10^5 183 stained cells were analysed per sample. 184 Data were normalized and files were concatenated and cleaned up to remove debris (by gating on cell 185 length and DNA⁺ cells), to exclude normalization beads (Ce140⁻ cells), to positively select intact cells 186 (Ir191⁺Ir193⁺), to positively select live cells (Rh103⁻Ir193⁺), and to identify CD19⁺ and/or CD20⁺ B cells. CD38hiCD24hi B cells were identified and exported as a new group prior to performing 187 SPADE (Spanning-tree Progression Analysis of Density-normalised Event) analysis (42). SPADE 188 189 analysis groups cells into 'nodes' based on the expression of all 30 markers to produce a two 190 dimensional tree. Using a colour coded expression scale, the nodes in the tree were manually grouped 191 into larger 'bubbles' to collect together nodes, and therefore cells, that had similar expression, i.e. all 192 those with high IgM expression were grouped together in one bubble.

193 **Results**

194 Heavy chain gene family usage distinguishes cell types

195 Pre-B (large pre-B) and immature B cells, from BM samples, and matched transitional and naïve B

196 cells, from PB samples, were sorted (Figure 1b and c) prior to high throughput sequencing using an

197 IgM specific constant region primer. Both the heavy and light chain (kappa and lambda) Ig genes

198 were amplified with a total of 96,593 heavy and 49,101 light chain sequences generated after initial 199 data clean up. These B cell populations are all thought to be exogenous antigen-naïve and therefore 200 will not have been activated to undergo somatic hypermutation and expansion. We do not see 201 evidence of somatic hypermutation in the gene sequences (data not shown) and therefore we have 202 assumed that any sequences with the same CDR3 region arise from PCR duplication. Therefore only 203 one example sequence of any unique gene rearrangement was used in this analysis, resulting in 39,577 204 heavy chain and 42,542 light chain sequences grouped by donor and cell type. Sequencing error does 205 not substantially affect the assignment of germline Ig genes to the sequences, however, for the CDR3 peptide analysis we further removed sequences where the peptide sequence may be inaccurate due to 206 207 sequencing error. This resulted in 29,074 heavy chain and 29,128 light chain sequences. Sequences 208 can be accessed on the National Center for Biotechnology Information's Sequence Read Archive in 209 raw format (BioProject: PRJNA39946; Sequence Read Archive accession: SRP081849) or in 210 processed format with metadata at www.bcell.org.uk.

211 Gene family repertoire can distinguish early human B cell subsets

212 Heavy chain V, D and J family usage did not show any significant differences in repertoire between 213 pre-B and immature cells from the bone marrow. There were, however, significant differences 214 between these BM cells and the peripheral transitional and naïve cells (Figure 2). IGHV3 family 215 genes are the most predominant genes in the human peripheral repertoire. It was interesting that in 216 the bone marrow this was particularly the case, with *IGHV3* cells actually being removed from the 217 repertoire during B cell maturation: There is a highly significant >13% decrease in the use of IGHV3 218 family genes in naïve cells with small increases in all other families to compensate (Figure 2a). Naïve 219 cells also showed a significantly decreased use of *IGHJ6* and, together with transitional cells, a > 6%220 reduction in use of IGHD2 family genes.

221 Since we had expected that peripheral transitional cells would fall between immature bone marrow

cells and peripheral naïve cells in the development pathway, and that any changes in repertoire we

saw would reflect this, we were surprised to see that this was not always the case. There was a

significant 5% increased frequency of *IGHD3* family usage in transitional cells compared to all other

cell types. Furthermore, there was a significant >9% increase in *IGHJ6* usage, compensated for by decreases in *IGHJ3*,4 and 5 usage, in transitional cells compared to all other cell types. This is reflected in the different size of bubble V3D3J6 in the bubble plots (Figure 2B). The different repertoire of transitional and naïve cells compared to the bone marrow cells (p<0.05, Wilcoxon), and compared to each other (p<0.001, Wilcoxon) is illustrated by a PCA analysis of gene family usage (Figure 2c).

231 Light chain repertoire is less variable.

232 In contrast to the heavy chain repertoire, the light chain gene family repertoire does not distinguish 233 between cell types. There are no significant changes in kappa family usage (Figure 3a). Some differences were seen in lambda families (Figure 3b). The IGLV2 family usage is significantly 234 235 increased by 10-15%, at the expense of all other families, and IGLJ1 family usage is significantly 236 increased by 2-5%, at the expense of IGLJ3. As a result of this an ANOVA analysis of the 237 combinatorial lambda family repertoire showed a significant difference between the immature and the transitional and naïve stages of development (p < 0.001). However, clustering by PCA showed that 238 239 any differences in light chain V-J gene usage were not strong enough to be able to distinguish 240 between the different cell types (Figure 3e). Nor were there any obvious differences between the 241 different cell types in lambda CDR3 amino acid sequence, since PCA of the kidera factors to assess 242 the physicochemical character of the CDR3 did not distinguish between the groups (Figure 3f).

243 Selection of individual *IGH* genes in early development

As the above analysis of gene family repertoire indicated that there were repertoire changes between

cell types, we analysed all the genes individually to check if we had missed any significant gene

selection due to the averaging effect of looking at the family level (Figure 4). Not all the *IGHV3*

- 247 family genes are decreased in naïve cells compared to bone marrow cells. While there are significant
- decreases in IGHV3-15, IGHV3-30 and IGHV3-33 in particular, IGHV3-9 is actually increased
- 249 (Figure 4a). Other notable increases occur in the two main *IGHV1* family genes: *IGHV1-18* and
- 250 IGHV1-69, and in the IGHV6 gene. The IGHD2 family decreases are contributed by IGHD2-15 and
- 251 IGHD2-2, and while the compensatory increase in other IGHD genes seemed unremarkable across the

board, *IGHD1-7* and *IGHD4-17* did show significant differences (Figure 4b). In spite of the

- significant change in *IGHD3* family use in transitional cells, this did not show up at the individual
- gene level, implying that the increase occurs throughout the *IGHD* gene family. Despite the lack of
- significant changes in *IGK* family repertoire there was a small (~3.8%) but significant increase in
- 256 *IGKV3-11* gene use in naïve cells compared to immature cells. This appeared to be at the expense of
- small (<3%) decreases in *IGKV3-20* and *IGKV4-1* genes. The increase in *IGLV2* family during
- development seemed to be mainly due to significant increases of 12.8% and 7% in IGL2-14 and
- 259 *IGL2-23* respectively (Figure 4c).

There is a certain amount of inter-individual variation that occurs in these analyses but the trends for selection of these genes in the repertoire are consistent, as illustrated in Figure 4, where the individual donors are shown separately for genes that are removed from the repertoire (Figure 4d) or that are increased in the repertoire (Figure 4e) during early development.

264 Heavy chain CDR3 properties are also strongly selected.

265 Although much of the CDR3 region is comprised of contributions from the individual IGHV, IGHD 266 and IGHJ genes, reflecting some of the repertoire selection effects that are captured in the analysis above, the actual amino acid sequences encoded by CDR3 varies tremendously even within the same 267 268 VDJ combinations. In addition to the direct effects of endonuclease action on the genes, and N region insertion by terminal deoxynucleotidyl transferase, the reading frame of the IGHD region can also 269 270 vary. Since the CDR3 region encodes a crucial part of the antibody binding site, and key functional 271 aspects of its structure are dependent on the primary sequence (43), we also analysed the biophysical 272 characteristics of the CDR3 amino acid sequence. Initially we used Kidera factors, which are a set of 273 10 orthogonal factors that encapsulate information from ~140 different measurable biophysical 274 characteristics of peptides. The data from PCA analysis of the CDR3 Kidera factors is in accordance 275 with that for the VDJ gene analysis, showing that the characteristics of pre-B and immature cells are 276 found in overlapping clusters (Figure 5a). Naïve cells and transitional cells, however, form separate 277 yet non-overlapping clusters. The data from heavy chain CDR3 Kidera analysis separates the groups 278 of cells better than the gene usage data, with 30% of the data contributing to PC1. To elucidate which

characteristics were mainly responsible for the differences we analysed some of the most common ones individually. The numbers of charged, basic and aromatic amino acids in each sequence, and the sequence Boman index, were significantly increased in naïve cells compared to pre-B cells (Figure 5b). Conversely, the number of small amino acids per sequence, the hydrophobicity (GRAVY index), aliphatic index and overall length of sequence were all disfavoured characteristics that were removed from the repertoire during development (Figure 5c). Interestingly the selection on the size of CDR3 region did not seem as strong in the older donors as it did in the younger ones (Figure 5d).

286 Human transitional cells are not just precursors to naïve cells.

287 The heavy chain gene and CDR3 PCA analysis (Figure 2c and Figure 5a) indicated that transitional 288 cells, in additional to being distinctive from pre-B cells and immature cells, also had a different 289 repertoire to naïve cells. We used cluster analysis (based on Minkowski distances) to investigate the 290 relationships further, which confirmed, by both VDJ usage (Figure 6a) and Kidera factors (Figure 6c), 291 that transitional cells have a different repertoire to the other cell types. Naïve cells formed a subbranch of the cluster containing pre-B and immature cells suggesting that the naïve repertoire is more 292 293 similar to the bone marrow cells than to the transitional cells. Clear examples of individual genes 294 where the usage in transitional cells differs from the rest of the cells can be seen in Figure 6b, and 295 biophysical characteristics showing the significantly different character of the heavy chain CDR3 in 296 transitional cells are shown in Figure 6d. Since this subset of cells has been reported to contain 297 regulatory B cells, as well as being the precursor to naïve B cells, we investigated the heterogeneity of 298 the population by mass cytometric analysis of surface markers. Although the population is small, it 299 does appear to contain a number of different potential subpopulations, as illustrated by the IgM 300 SPADE plot in Figure 6e.

301 Discussion

The lack of difference between the heavy chain repertoire in pre-B and immature B cells implies that there is very little selective pressure at this developmental stage, which is in agreement with current thinking on the tolerance checkpoints (44). As expected, we do see a major difference between 305 immature bone marrow B cells and the transitional and naïve mature peripheral B cells, where we 306 would expect the repertoire to reflect the changes incurred as a result of the post-immature selective 307 processes that can remove up to 50% of the repertoire (8). There is a wealth of literature on the heavy 308 chain gene usage in different conditions, and both negative and positive associations have been made 309 for various genes. For example, the common IGHV1 family genes IGHV1-18 and IGHV1-69 have 310 been associated with responses to viral infections as well as with stereotypical receptors in CLL. It is 311 interesting that these two genes increase, and a number of *IGHV3* family genes decrease, since this 312 recapitulates the change in repertoire between naïve and switched memory repertoire (21). Indeed the 313 relative use of IGHV1 and IGHV3 genes seems to be a marker that distinguished between a number of 314 different B cell types (25). Furthermore, the significant changes in CDRH3 are to be expected from a 315 selected population, since this forms the most important part of the antibody binding site in all except 316 the smallest CDRH3 regions. What was particularly striking from this data was that the selection in 317 CDRH3 appeared to change with age even at this early stage in development, particularly in the 318 length of the CDRH3 region. We, and others, have previously noted that shorter CDRH3 regions are 319 selected upon exogenous antigen selection (21, 28, 45), and that older people have longer CDRH3 320 regions than in the young when measured in peripheral blood IgM-expressing cells. This data shows that a longer CDRH3 exists in B cells even before exogenous antigen stimulation so is likely a result 321 322 of changes in bone marrow tolerance selection rather than any exogenous antigen selection of IgM 323 memory cells.

324 Receptor editing to rescue potentially autoreactive B cells can occur after the immature B cell stage 325 once the light chain has been co-expressed. The light chain loci continue rearrangement to form a 326 new gene. The kappa light chain locus rearranges before the lambda locus, and has the potential to 327 rearrange a number of times. However, at some point the kappa locus would run out of genes to 328 rearrange, or the kappa deleting element would be used, in which case then the lambda locus would 329 start rearrangement (3, 5). With this in mind, the paucity of differences in light chain repertoire 330 between immature, transitional and naïve cells is quite surprising. The kappa repertoire in particular 331 does not change much, possibly indicating that that the ability of different kappa genes to rescue a

332 potentially autoreactive heavy chain gene does not vary much. Only IGKV3-20 and IGKV4-1 show a 333 significant decrease in use (Figure 4c), implying a potential contribution to autoreactive BCR. 334 Indeed, *IGKV4-1* has previously been shown to be overrepresented in systemic lupus erythromatosus, 335 celiac disease and type 1 diabetes (46, 47), and we have also shown that its actual expression in the 336 peripheral repertoire is significantly lower than its frequency of rearrangement in the genomic DNA 337 (48). IGKV3-11 may possibly be a rescue gene, showing a significant increase in use, and our previous analysis also showed an increase in expression of this gene in the expressed repertoire 338 339 compared to its expected frequency of rearrangement (48). Two IGLV2 lambda genes were noted as being increased within the lambda repertoire, presumably in preference to the IGLV1 family genes 340 that showed a slight decrease. Not much is known about the potential significance of lambda light 341 342 chain genes, although it has been reported that POEMS syndrome of neuropathy is associated with 343 monoclonal expansions of *IGLV1* family plasma cells (49). It has been reported that lambda light 344 chains have a good potential for rescuing autoreactive B cells (50). Since the primer sets we used for 345 these experiments amplified the kappa and lambda light chains separately we cannot comment on any changes in kappa/lambda ration between immature and later B cells. Given the inability of the light 346 347 chain repertoire characteristics to distinguish between the different cell types, as shown by the PCA of 348 Figure 3e and 3f, it is possible that any light chain-mediated autoreactive rescue would be more likely 349 to be performed by a switch from kappa to lambda than by a switch within the loci. Alternatively, the 350 lack of cell type-distinguishing features in the light chain repertoire could mean that the central 351 selection events are mainly driven by heavy chain-encoded binding specificities. The selection in 352 heavy chain but not light chain also implies that the heavy-light chain pairing is mostly random, since 353 if the pairing had biases then the same selection effects would appear in both chains. This is in 354 agreement with previous data where a large number of paired heavy and light chain rearrangements 355 were sequenced (51, 52). It has been previously reported that a particular CDRH3 stereotype on a 356 IGHV1-69 background might be associated with a particular light chain gene, but this was on a small 357 sample size (n=66) of selected CLL sequences (53) and the data here represents a much larger 358 diversity in a normal unselected population of cells.

359 What we had not expected to see in these data was the large difference between transitional and naïve 360 B cells, which does not seem in accord with an immature-transitional-naïve pathway of development. 361 One assumes that processes in nature have evolved to require minimum energy or resource, and if this 362 is the case then any change in repertoire between creation (pre-B cells) and end point (naïve B cells) 363 would be in a single linear direction. The actual cell-cell differences may vary depending on which point the selection pressure were applied, but one would not expect to see a change in direction of 364 365 increase/decrease one way, followed by a change in direction back again, half way through a 366 development pathway, i.e. for a gene that was being removed from the repertoire through the 367 development pathway we would expect the percentage representation in the repertoire to be pre-368 B>immature>transitional>naïve. In actual fact, for some genes, we see varying patterns such as 369 transitional >[pre-B=immature]>naïve. For this reason, and in the light of results exemplified by use 370 of IGHV3-53 or use of non-polar CDR3 amino acids (Figure 6b and d), we assume that a large 371 proportion of the cells in our transitional subset are not intermediates between bone marrow immature and peripheral naïve B cells. We sorted our CD19⁺IgD⁺CD10^{hi}CD27⁻ cells, based on the previous 372 373 information that CD10, CD24, CD38 decrease as cells develop from immature to naïve. This 374 information had been obtained by studying the reconstitution of different phenotypic subsets after B 375 cell depletion (9). There are three subsets of mature non-memory B cells by the expression of CD10 (high, medium, and low) that have parallels in the differing strengths of CD24^{hi}CD38^{hi} expression in 376 377 humans. These distinctions were first described in mice as T1, T2 and T3 subsets and this 378 nomenclature has been carried over into human studies (54). The transitional cell subset in humans 379 has been shown to contain B cells with regulatory activity after stimulation in vitro (55), and have 380 also been shown to contain cells with different homing integrins (56). It is clear from our high 381 dimensional phenotyping in Figure 6 that the population can be quite heterogeneous. Since the FACS 382 gates that we used were quite stringent, we skewed our cells towards the equivalent of the mouse T1 383 population which may be less diverse and less representative of that portion of cells that are 384 precursors to naïve cells. In this context it is interesting that a prior comparison of human T1 and T2 385 cells also showed a difference in IGHJ6 usage (11). Without the immature B cell repertoire to give 386 this context this could be interpreted as IGHJ6 being removed gradually from the repertoire.

387 However, in the light of the fact that our transitional cells have higher *IGHJ6* than either immature

388 cells or naïve cells then this is unlikely. In reality, this CD10 very high population has a very

389 distinctive repertoire in many other respects also, and therefore likely has a completely different

390 function. Whether this would be the regulatory B cell subset or not would require further

391 investigation in the future.

392 In summary, we have shown that there are strong selective influences over the B cell repertoire in early B cell development, and we can identify genes and characteristics that are likely to be detrimental by 393 394 the fact that they disappear from the repertoire in development. The selection effects are mainly on the 395 heavy chain rather than the light chain genes. This is surprising considering the role that receptor editing 396 is thought to play in central tolerance, and may mean that either the heavy chain plays a dominant role 397 in receptor specificity, or that switching between kappa and lambda is the chief mode of receptor 398 editing. An unexpected finding was that the transitional subset of cells with the highest level of CD10 399 expression may not really be a transitional stage between immature and naïve B cells, and further work will be required to determine whether these represent the regulatory B cells. 400

401

402 Acknowledgements.

403 The authors are extremely grateful to all the staff and patients at the orthopaedic unit of Guy's

404 Hospital and to the funders: This work was supported by a joint programme from the MRC and

405 BBSRC (MR/L01257X/1) and also by funds from the Dunhill Medical Trust (R279/0213) and a 406 CASE award from the BBSRC in conjunction with MedImmune (BB/L015854/1).

References

410	1. Gathings WE, Lawton AR, Cooper MD. Immunofluorescent studies of the development of
411	pre-B cells, B lymphocytes and immunoglobulin isotype diversity in humans. Eur J Immunol.
412	1977;7(11):804-10.
413	2. Hardy RR, Carmack CE, Shinton SA, Kemp JD, Hayakawa K. Resolution and
414	characterization of pro-B and pre-pro-B cell stages in normal mouse bone marrow. J Exp Med.
415	1991;173(5):1213-25.
416	3. Hardy RR, Hayakawa K. B cell development pathways. Annu Rev Immunol. 2001;19:595-
417	621.
418	4. Schroeder HW, Jr. The evolution and development of the antibody repertoire. Front Immunol.
419	2015;6:33.
420	5. Santos P, Arumemi F, Park KS, Borghesi L, Milcarek C. Transcriptional and epigenetic
421	regulation of B cell development. Immunol Res. 2011;50(2-3):105-12.
422	6. Halverson R, Torres RM, Pelanda R. Receptor editing is the main mechanism of B cell
423	tolerance toward membrane antigens. Nat Immunol. 2004;5(6):645-50.
424	7. Keenan RA, De Riva A, Corleis B, Hepburn L, Licence S, Winkler TH, et al. Censoring of
425	autoreactive B cell development by the pre-B cell receptor. Science. 2008;321(5889):696-9.
426	8. Wardemann H, Yurasov S, Schaefer A, Young JW, Meffre E, Nussenzweig MC. Predominant
427	autoantibody production by early human B cell precursors. Science. 2003;301(5638):1374-7.
428	9. Palanichamy A, Barnard J, Zheng B, Owen T, Quach T, Wei C, et al. Novel human
429	transitional B cell populations revealed by B cell depletion therapy. J Immunol. 2009;182(10):5982-
430	93.
431	10. Boyd SD, Liu Y, Wang C, Martin V, Dunn-Walters DK. Human lymphocyte repertoires in
432	ageing. Curr Opin Immunol. 2013;25(4):511-5.
433	11. Sims GP, Ettinger R, Shirota Y, Yarboro CH, Illei GG, Lipsky PE. Identification and
434	characterization of circulating human transitional B cells. Blood. 2005;105(11):4390-8.
435	12. Simon Q, Pers JO, Cornec D, Le Pottier L, Mageed RA, Hillion S. In-depth characterization
436	of CD24(high)CD38(high) transitional human B cells reveals different regulatory profiles. J Allergy
437	Clin Immunol. 2016;13/(5):15/ $/$ -84 e10.
438	13. Petro JB, Gerstein RM, Lowe J, Carter RS, Sninners N, Knan WN. I ransitional type 1 and 2 D knowle such sate such sate and differentially represented states resented signaling. J Dial Chart
439	B lymphocyte subsets are differentially responsive to antigen receptor signaling. J Biol Chem.
440	2002;277(50):48009-19.
441	14. Chung JB, Sater KA, Fleids ML, Effkson J, Molifoe JG. CD25 defines two distinct subsets of immeture R calls which differ in their represented to T call help signals. Int Immunol. 2002;14(2):157
442	finitiature B cens which unter in their responses to 1 cen help signals. Int finitumor. 2002,14(2).157-
443 AAA	15 von Boehmer H. Melchers F. Checkpoints in lymphocyte development and autoimmune
445	disease Nat Immunol 2010:11(1):14-20
446	16 Su TT Rawlings DI Transitional B lymphocyte subsets operate as distinct checkpoints in
447	murine splenic B cell development [Immuno] 2002:168(5):2101-10
448	17 Roy V Chang NH Cai Y Bonventi G Wither I Aberrant IgM signaling promotes survival
449	of transitional T1 B cells and prevents tolerance induction in lupus-prone New Zealand black mice. I
450	Immunol. 2005:175(11):7363-71.
451	18. Blair PA, Norena LY, Flores-Boria F, Rawlings DJ, Isenberg DA, Ehrenstein MR, et al.
452	CD19(+)CD24(hi)CD38(hi) B cells exhibit regulatory capacity in healthy individuals but are
453	functionally impaired in systemic Lupus Erythematosus patients. Immunity. 2010;32(1):129-40.
454	19. Wehr C, Eibel H, Masilamani M, Illges H, Schlesier M, Peter HH, et al. A new CD21low B
455	cell population in the peripheral blood of patients with SLE. Clin Immunol. 2004;113(2):161-71.
456	20. Bemark M. Translating transitions - how to decipher peripheral human B cell development. J
457	Biomed Res. 2015;29(4):264-84.

- 458 21. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High459 throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and
 460 switched memory B-cell populations. Blood. 2010;116(7):1070-8.
- 461 22. Wu YC, Kipling D, Dunn-Walters DK. The relationship between CD27 negative and positive
 462 B cell populations in human peripheral blood. Front Immunol. 2011;2:81.
- 463 23. Martin V, Bryan Wu YC, Kipling D, Dunn-Walters D. Ageing of the B-cell repertoire. Philos
 464 Trans R Soc Lond B Biol Sci. 2015;370(1676).
- 465 24. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, et al. Large-scale
 466 sequence and structural comparisons of human naive and antigen-experienced antibody repertoires.
 467 Proc Natl Acad Sci U S A. 2016;113(19):E2636-45.
- 468 25. Martin V, Wu YC, Kipling D, Dunn-Walters DK. Age-related aspects of human IgM(+) B
 469 cell heterogeneity. Ann N Y Acad Sci. 2015;1362:153-63.
- 470 26. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH
 471 repertoires revealed by deep sequencing. J Immunol. 2012;189(6):3221-30.
- 472 27. Meffre E, Davis E, Schiff C, Cunningham-Rundles C, Ivashkiv LB, Staudt LM, et al.
 473 Circulating human B cells that express surrogate light chains and edited receptors. Nat Immunol.
 474 2000;1(3):207-13.
- Townsend C. L., Laffy J. MJ, Wu YC, Silva O'Hare J, Martin V., David K., et al. Significant
 differences in physicochemical properties of human immunoglobulin kappa and lambda CDR3
- 477 regions. Frontiers Immunology. 2016;Manuscript submitted.
- 478 29. Wu YC, Kipling D, Dunn-Walters D. Assessment of B Cell Repertoire in Humans. Methods
 479 in molecular biology (Clifton, NJ). 2015;1343:199-218.
- 480 30. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc MP. IMGT/HIGHV-QUEST: The IMGT®
 481 web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high
 482 throughput and deep sequencing. Immunome research. 2012;8(1):2.
- 483 31. Lefranc MP. Antibody Informatics: IMGT, the International ImMunoGeneTics Information
 484 System. Microbiol Spectr. 2014;2(2).
- 485 32. Osorio D, Rondon-Villarreal P, Torres R. Peptides: Calculate Indices and Theoretical
 486 Properties of Protein Sequences. R package version 1.1.1. 2015.
- 487 33. Team RC. R: A Language and Environment for Statistical Computing. Vienna, Austria: R
 488 Foundation for Statistical Computing; 2015.
- 489 34. Dunn-Walters D. Dunn-Walters' Lab. <u>http://wwwbcellorguk</u>.
- 490 35. Soetaert K. plot3D : Tools for plotting 3-D and 2-D data. R package version 10-2. 2014.
- 491 36. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein.
 492 Journal of Molecular Biology. 1982;157(1):105-32.
- 493 37. Boman HG. Antibacterial peptides: basic facts and emerging concepts. Journal of Internal
 494 Medicine. 2003;254(3):197-215.
- 495 38. Rice P, Longden I, Bleasby A. EMBOSS: The European molecular biology open software
 496 suite. Trends in Genetics. 2000;16(6):276-7.
- 497 39. Ikai A. Thermostability and aliphatic index of globular proteins. Journal of Biochemistry.
 498 1980;88(6):1895-8.
- 499 40. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical
- 500 properties of the 20 naturally occurring amino acids. Journal of Protein Chemistry. 1985;4(1):23-55.
- 501 41. Douglas Bates MM, Ben Bolker, Steve Walker. Fitting Linear Mixed-Effects Models using
 502 lme4. Journal of Statistical Software. 2015;67(1):1-48.
- 503 42. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Jr., Bruggner RV, Linderman MD, et al.
- 504 Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol.505 2011;29(10):886-91.
- 506 43. Casali P, Schettino EW. Structure and function of natural antibodies. Curr Top Microbiol
 507 Immunol. 1996;210:167-79.
- 508 44. Tussiwand R, Bosco N, Ceredig R, Rolink AG. Tolerance checkpoints in B-cell development:
 509 Johnny B good. Eur J Immunol. 2009;39(9):2317-24.
- 510 45. Rosner K, Winter DB, Tarone RE, Skovgaard GL, Bohr VA, Gearhart PJ. Third
- 511 complementarity-determining region of mutated VH immunoglobulin genes contains shorter V, D, J,
- 512 P, and N components than non-mutated genes. Immunology. 2001;103(2):179-87.

- Dorner T, Foster SJ, Farner NL, Lipsky PE. Immunoglobulin kappa chain receptor editing in 513 46. systemic lupus erythematosus. J Clin Invest. 1998;102(4):688-94. 514 515 Woodward EJ, Thomas JW. Multiple germline kappa light chains generate anti-insulin B cells 47. 516 in nonobese diabetic mice. J Immunol. 2005;175(2):1073-9. 517 Hehle V, Fraser LD, Tahir R, Kipling D, Wu YC, Lutalo PM, et al. Immunoglobulin kappa 48. variable region gene selection during early human B cell development in health and systemic lupus 518 519 erythematosus. Mol Immunol. 2015;65(2):215-23. 520 Abe D, Nakaseko C, Takeuchi M, Tanaka H, Ohwada C, Sakaida E, et al. Restrictive usage of 49. 521 monoclonal immunoglobulin lambda light chain germline in POEMS syndrome. Blood. 522 2008;112(3):836-9. Prak EL, Trounstine M, Huszar D, Weigert M. Light chain editing in kappa-deficient animals: 523 50. 524 a potential mechanism of B cell tolerance. J Exp Med. 1994;180(5):1805-15. Brezinschek HP, Foster SJ, Dorner T, Brezinschek RI, Lipsky PE. Pairing of variable heavy 525 51. 526 and variable kappa chains in individual naive and memory B cells. J Immunol. 1998;160(10):4762-7. Jayaram N, Bhowmick P, Martin AC. Germline VH/VL pairing in antibodies. Protein Eng 527 52. 528 Des Sel. 2012;25(10):523-9. 53. Widhopf GF, 2nd, Goldberg CJ, Toy TL, Rassenti LZ, Wierda WG, Byrd JC, et al. 529 530 Nonstochastic pairing of immunoglobulin heavy and light chains expressed by chronic lymphocytic 531 leukemia B cells is predicated on the heavy chain CDR3. Blood. 2008;111(6):3137-44. 532 54. Agrawal S, Smith SA, Tangye SG, Sewell WA. Transitional B cell subsets in human bone 533 marrow. Clin Exp Immunol. 2013;174(1):53-9. 534 55. Mauri C, Menon M. The expanding family of regulatory B cells. Int Immunol. 535 2015;27(10):479-86. 536 Vossenkamper A, Blair PA, Safinia N, Fraser LD, Das L, Sanders TJ, et al. A role for gut-56. 537 associated lymphoid tissue in shaping the human B cell repertoire. J Exp Med. 2013;210(9):1665-74.
- 538

539 Figure Legends

- 540 *Figure 1.* Isolation of B cells early in development. *a)* B cell development pathway with phenotype
- 541 used to distinguish each cell type. Starting from a CD19+ population: B) Example showing the
- 542 sorting strategy used to isolate preB (red:IgK⁻IgL⁻CD38⁺IgM⁺) and immature (orange: IgK⁺ or
- 543 IgL⁺CD27⁻IgM⁺IgD⁻CD10⁺) B cells from bone marrow mononuclear cells (BMMCs). C) Sorting
- 544 strategy used to isolate transitional (green: $IgD^+CD27^-CD10^+$) and naïve (blue: $IgD^+CD27^-CD10^-$)
- 545 cells from matched peripheral blood mononuclear cells (PBMC)s. Dotted lines on the plots represent
- 546 the gates based on FMO controls and the solid lined boxes represent the gating used to collect the
- 547 *different subsets.*

548 Figure 2. Heavy chain VDJ gene family usage distinguishes cell types *a*) Mean frequency

- 549 histograms of individual V, D and J family usage for the heavy chain gene families of PreB (red),
- 550 *immature (yellow), transitional (green) and naïve (blue) cells (* p<0.05 by 2 way ANOVA with*
- 551 multiple analysis correction. Error bars are SEM). b) VDJ family combination usage in the different

- 552 *cell types. The size of a bubble represents the mean frequency of that VDJ combination. c)*
- 553 Transitional and Naïve cells show difference in VDJ family usage by principle component analysis
- 554 (PCA) (left) compared to a randomised data set (right).

555 Figure 3: Light chain gene usage and CDR3 properties cannot distinguish between cell types

- 556 *a-b)* V and J family usage for kappa (a) and lambda (b) light chain gene families between immature
- 557 (yellow), transitional (green) and naïve cell types (* p<0.05 by 2 way ANOVA with multiple analysis
- 558 correction. Error bars are SEM). c-d) light chain VJ usage for kappa (c) and lambda (d) light chains
- 559 in immature (yellow), transitional (green) and naïve (blue) B cells. The size of a circle indicates the
- 560 relative mean frequency of the VJ combination. e-f) Principle component analysis (PCA) of VJ usage
- 561 (e) and kidera factors (f) in three different cell types for kappa (top) and lambda (bottom).

562 Figure 4: Individual genes can be favoured or disfavoured as B cells mature. a-c) Frequency of

- 563 IGHV (a) and IGHD (b) gene usage in heavy chain and IGKV and IGLV usage in light chains (c) of
- 564 *different cell types are compared (* p<0.05 by 2 way ANOVA with multiple analysis correction.*
- 565 Error bars are SEM). d-e) The frequency for each cell type in each individual donor is shown for
- 566 genes that are decreased during selection (d) and those that are increased (e).
- 567 Figure 5. Heavy chain CDR3 characteristics distinguish between cell types. (a) Distinction
- 568 between the different cell types by kidera factors as illustrated by Principle component analysis
- 569 (PCA). Distribution of CDRH3 physicochemical properties that have an increased trend from preB
- 570 (P) immature (I), transitional (T) to naïve (N) cells (b), and a decrease in naïve cells compared to
- 571 preB cells (c). (* p<0.05 ANOVA). (d) The heavy chain CDR3 length in all cells types in young and
- 572 old donors. (young donors: 18 to 50; old donors: over 65). (* p<0.05 ANOVA). Values on the y axis of
- 573 *b)* to *d*) are as per the individual graph titles.

574 Figure 6: Transitional cells have a unique heavy chain immunoglobulin repertoire. (*a*, *c*)

- 575 Minkowski distance clustering analysis of heavy chain VDJ family usage (a) and CDRH3 kidera
- 576 *factors for preB (P) immature (I), transitional (T) and naïve (N) cells in each donor (c). (b) The*
- 577 *frequency of gene use (%) for different cell types in each individual donor for genes that have a*

- 578 distinctive distribution in transitional cells. (d) CDRH3 physicochemical properties in different cell
- 579 types for properties that have a distinctive distributions in transitional cells. (* p<0.05 ANOVA).
- 580 Values on the y axis of are as per the individual graph titles. (e) High dimensional clustering of
- 581 CD24hiCD38hi transitional B cells indicates heterogeneity within the transitional population with
- 582 respect to IgM expression, illustrated as a SPADE plot. Populations numbered 1 to 13 have been
- 583 grouped according to expression of IgM, IgD, CD21, CD23 as shown in supplementary figure 1.
- 584 Figure S1. High High dimensional clustering of CD24hiCD38hi transitional B cells indicates
- 585 *heterogeneity within the transitional population with respect to IgD, CD21, CD23 expression,*
- 586 illustrated as a SPADE plot. Populations numbered 1 to 13 have been grouped according to
- 587 *expression of IgM, IgD, CD21, CD23, see figure 6e for a tabulated summary.*















