**Rationale for choosing heavy chains**

To be sure that my data analysis was not being influenced too much by the presence of unproductive sequences, I did my data analysis again, having removed all unproductive sequences. I got similar results to when the unproductive sequences are left in.

Data removed from the heavy chain repertoire to produce the following results:

* CDR3s >35 amino acids long (61 in total)
* CDR3s of length "NA" (5 in total)
* CDR3s containing an "X" in their AA sequence (Edited\_AA\_junction column) (149 in total)
* CDR3s containing an "\*" in their AA sequence (Edited\_AA\_junction column - determined using LEN() and comparing with the Num\_AAs column) (2896 in total)
* All unproductive entries (had a # in their CDR3 region) were then removed (7115 were removed - this is probably not the total number of unproductives as the above had all been removed before unproductives were removed)
* Removed remaining 464 unproductives
* Removed the 1 patient "NA" (also had Ig class G)

Number of entries in original database: 39638. Number of entries after above entries removed: 29017 (some of the entries fall in to more than one category so the number removed will not sum to the number in each category)

**VDJ usage:**



**Figure 1:** VDJ family usage was analysed using two-way repeated measures ANOVA followed by Turkey’s multiple comparisons post-test. The bars show the mean frequency of usage +/- 1SD. **A)** VH3 is significantly less prevalent in the naïve repertoire. **B)** DH2 is significantly less prevalent in the periphery than in the bone marrow. **C)** JH6 is significantly less common in the naïve repertoire compared to the immature repertoire.



**Figure 2: A)** Histogram of the frequency of each CDRH3 length per cell type. This shows that the distribution of CDRH3 lengths of naïve cells is skewed towards shorter CDRH3s than the other three cell types. **B)** Mean CDRH3 length per cell type, +/- one standard deviation. This graph shows that the mean CDRH3 length of the naïve repertoire is significantly shorter than the other three repertoires. Significance was calculated using one-way repeated measures ANOVA followed by Holm-Sidak’s multiple comparisons post-test.



**Figure 3: A)** Mean GRAVY index +/- one standard deviation per cell type. The mean GRAVY index significantly decreases after central tolerance. **B)** Mean Boman index +/- one standard deviation per cell type. The mean Boman index is significantly higher in the naïve repertoire. Significance was calculated using one-way repeated measures ANOVA followed by Holm-Sidak’s multiple comparisons post-test.

**Note:** each data point shape represents a different patient (same shape per patient for all three graphs). Filled-in data points are young patients, hollow data points are old patients.

**Parameters by which I filtered the dataset to leave predicted autoreactive heavy chains:**

1. IGHV3
2. IGHD2
3. IGHJ6
4. Num\_AA. greater-than 22AA (i.e. 23 or more) (mean + 2SD)
5. GRAVY\_index greater than 0.163 (mean + 1SD)
6. Boman less than 0.712 (mean - 1SD)

**This filtering left the following 62 entries, which I predict are likely to be autoreactive/polyreactive:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Seq\_ID** | **patient** | **cell.type** | **Subclass** | **Vgene** | **Jfamily** | **Dgene** | **Num\_AAs** | **pI** | **Gravy\_index** | **Boman** |
| preh119\_M\_UN\_212620 | 119 | preB | IgM | IGHV3-30-3 | IGHJ6 | IGHD2-8 | 23 | 4.36 | 0.378261 | 0.433913 |
| immh141\_M\_V3\_41894 | 141 | immature | IgM | IGHV3-11 | IGHJ6 | IGHD2-15 | 23 | 3.75 | 0.191304 | -0.35304 |
| trah141\_UN\_V3\_52939 | 141 | transitional | NA | IGHV3-9 | IGHJ6 | IGHD2-15 | 27 | 3.57 | 0.674074 | -0.05926 |
| immh119\_M\_V3\_113890 | 119 | immature | IgM | IGHV3-23 | IGHJ6 | IGHD2-2 | 29 | 4.37 | 0.193103 | 0.452069 |
| preh149\_M\_V3\_297170 | 149 | preB | IgM | IGHV3-23 | IGHJ6 | IGHD2-15 | 25 | 4.37 | 1 | 0.1152 |
| preh119\_M\_V3\_51502 | 119 | preB | IgM | IGHV3-30 | IGHJ6 | IGHD2-21 | 25 | 8.91 | 0.284 | 0.3652 |
| preh118\_M\_V3\_316900 | 118 | preB | IgM | IGHV3-23 | IGHJ6 | IGHD2-2 | 24 | 3.83 | 0.670833 | -0.325 |
| immh146\_M\_UN\_45274 | 146 | immature | IgM | IGHV3-9 | IGHJ6 | IGHD2-2 | 29 | 3.83 | 0.510345 | -0.20448 |
| preh119\_M\_V3\_37944 | 119 | preB | IgM | IGHV3-30 | IGHJ6 | IGHD2-15 | 23 | 4.11 | 0.33913 | 0.052609 |
| immh149\_M\_V3\_59726 | 149 | immature | IgM | IGHV3-30 | IGHJ6 | IGHD2-2 | 24 | 5.41 | 0.720833 | -0.45333 |
| trah119\_M\_V3\_46902 | 119 | transitional | IgM | IGHV3-23 | IGHJ6 | IGHD2-2 | 23 | 3.61 | 0.195652 | 0.426087 |
| preh149\_M\_UN\_598450 | 149 | preB | IgM | IGHV3-30 | IGHJ6 | IGHD2-2 | 24 | 9.03 | 0.229167 | 0.024167 |
| preh146\_M\_V3\_336300 | 146 | preB | IgM | IGHV3-9 | IGHJ6 | IGHD2-2 | 23 | 9.29 | 0.452174 | 0.396522 |
| immh146\_M\_UN\_34672 | 146 | immature | IgM | IGHV3-73 | IGHJ6 | IGHD2-2 | 25 | 9.1 | 0.492 | 0.55 |
| naih111\_M\_V3\_35305 | 111 | naïve | IgM | IGHV3-23 | IGHJ6 | IGHD2-2 | 28 | 6.07 | 0.278571 | 0.018571 |
| preh146\_M\_V3\_253300 | 146 | preB | IgM | IGHV3-30 | IGHJ6 | IGHD2-2 | 25 | 3.83 | 0.356 | 0.2444 |
| naih111\_M\_V3\_30355 | 111 | naïve | IgM | IGHV3-30 | IGHJ6 | IGHD2-2 | 25 | 4.11 | 0.312 | 0.0844 |
| preh122\_M\_V3\_380030 | 122 | preB | IgM | IGHV3-23 | IGHJ6 | IGHD2-2 | 30 | 6.34 | 0.506667 | -0.08233 |
| naih149\_M\_V3\_40880 | 149 | naïve | IgM | IGHV3-7 | IGHJ6 | IGHD2-2 | 25 | 10.01 | 0.464 | 0.3484 |
| preh146\_M\_V3\_367320 | 146 | preB | IgM | IGHV3-21 | IGHJ6 | IGHD2-15 | 28 | 4.19 | 0.253571 | 0.136071 |
| preh111\_M\_V3\_217890 | 111 | preB | IgM | IGHV3-21 | IGHJ6 | IGHD2-2 | 24 | 4.11 | 0.170833 | 0.58875 |
| preh140\_M\_V3\_140960 | 140 | preB | IgM | IGHV3-48 | IGHJ6 | IGHD2-8 | 24 | 6.23 | 0.645833 | -0.37667 |
| immh141\_M\_V3\_48900 | 141 | immature | IgM | IGHV3-30-3 | IGHJ6 | IGHD2-2 | 26 | 4.11 | 0.407692 | 0.089615 |
| preh118\_M\_V3\_142450 | 118 | preB | IgM | IGHV3-11 | IGHJ6 | IGHD2-2 | 25 | 3.7 | 0.492 | 0.41 |
| preh118\_M\_V3\_244030 | 118 | preB | IgM | IGHV3-21 | IGHJ6 | IGHD2-2 | 23 | 4.11 | 0.595652 | -0.04391 |
| immh159\_M\_UN\_24014 | 159 | immature | IgM | IGHV3-21 | IGHJ6 | IGHD2-2 | 24 | 4.31 | 0.275 | 0.58875 |
| preh119\_M\_V3\_94854 | 119 | preB | IgM | IGHV3-30-3 | IGHJ6 | IGHD2-2 | 29 | 9.05 | 0.331035 | 0.685517 |
| preh141\_M\_V3\_547430 | 141 | preB | IgM | IGHV3-48 | IGHJ6 | IGHD2-2 | 27 | 4.11 | 0.17037 | 0.444815 |
| preh149\_UN\_V3\_39620 | 149 | preB | NA | IGHV3-48 | IGHJ6 | IGHD2-2 | 25 | 4.31 | 0.664 | 0.5648 |
| preh146\_UN\_V3\_606250 | 146 | preB | NA | IGHV3-21 | IGHJ6 | IGHD2-2 | 31 | 7.53 | 0.625807 | 0.165161 |
| preh138\_M\_V3\_413660 | 138 | preB | IgM | IGHV3-13 | IGHJ6 | IGHD2-2 | 24 | 4.11 | 0.5375 | 0.197083 |
| preh118\_M\_V3\_855230 | 118 | preB | IgM | IGHV3-33 | IGHJ6 | IGHD2-2 | 27 | 3.83 | 0.340741 | 0.20037 |
| preh146\_M\_V3\_877550 | 146 | preB | IgM | IGHV3-21 | IGHJ6 | IGHD2-2 | 23 | 3.74 | 0.165217 | 0.546522 |
| immh119\_M\_V3\_58979 | 119 | immature | IgM | IGHV3-21 | IGHJ6 | IGHD2-2 | 24 | 3.88 | 0.383333 | 0.45875 |
| preh118\_M\_V3\_180330 | 118 | preB | IgM | IGHV3-74 | IGHJ6 | IGHD2-2 | 26 | 6.34 | 0.330769 | 0.575769 |
| preh119\_M\_V3\_323150 | 119 | preB | IgM | IGHV3-21 | IGHJ6 | IGHD2-2 | 25 | 4.19 | 0.396 | 0.1004 |
| preh146\_M\_V3\_757000 | 146 | preB | IgM | IGHV3-21 | IGHJ6 | IGHD2-15 | 26 | 7.7 | 0.692308 | -0.02846 |
| immh118\_M\_UN\_35947 | 118 | immature | IgM | IGHV3-74 | IGHJ6 | IGHD2-2 | 23 | 6.5 | 0.308696 | 0.438261 |
| immh118\_M\_UN\_25365 | 118 | immature | IgM | IGHV3-21 | IGHJ6 | IGHD2-2 | 25 | 4.11 | 0.364 | -0.0364 |
| preh146\_M\_V3\_476100 | 146 | preB | IgM | IGHV3-23 | IGHJ6 | IGHD2-2 | 28 | 9.55 | 0.560714 | 0.298929 |
| naih160\_M\_V3\_35691 | 160 | naïve | IgM | IGHV3-11 | IGHJ6 | IGHD2-2 | 24 | 8.99 | 0.75 | -0.11917 |
| preh138\_M\_UN\_389980 | 138 | preB | IgM | IGHV3-48 | IGHJ6 | IGHD2-2 | 24 | 6.34 | 0.420833 | 0.217083 |
| preh140\_M\_V3\_284890 | 140 | preB | IgM | IGHV3-21 | IGHJ6 | IGHD2-2 | 24 | 8.92 | 0.320833 | 0.11 |
| preh138\_M\_V3\_201820 | 138 | preB | IgM | IGHV3-30 | IGHJ6 | IGHD2-21 | 24 | 9.17 | 0.208333 | 0.537083 |
| immh146\_M\_V3\_16858 | 146 | immature | IgM | IGHV3-30-3 | IGHJ6 | IGHD2-2 | 24 | 9.07 | 0.4375 | 0.6675 |
| preh146\_M\_UN\_693120 | 146 | preB | IgM | IGHV3-23 | IGHJ6 | IGHD2-2 | 24 | 8.93 | 0.495833 | 0.408333 |
| immh146\_M\_UN\_37601 | 146 | immature | IgM | IGHV3-23 | IGHJ6 | IGHD2-2 | 24 | 6.23 | 0.6875 | -0.13958 |
| immh141\_M\_V3\_114388 | 141 | immature | IgM | IGHV3-11 | IGHJ6 | IGHD2-2 | 27 | 4.11 | 0.503704 | 0.205926 |
| trah119\_M\_V3\_17885 | 119 | transitional | IgM | IGHV3-48 | IGHJ6 | IGHD2-15 | 27 | 3.75 | 0.27037 | -0.20481 |
| preh140\_M\_UN\_477980 | 140 | preB | IgM | IGHV3-30-3 | IGHJ6 | IGHD2-15 | 25 | 6.07 | 0.34 | 0.212 |
| trah119\_M\_V3\_11895 | 119 | transitional | IgM | IGHV3-23 | IGHJ6 | IGHD2-8 | 25 | 3.74 | 0.296 | 0.1816 |
| preh159\_M\_V3\_60711 | 159 | preB | IgM | IGHV3-15 | IGHJ6 | IGHD2-2 | 24 | 4.1 | 0.975 | -0.85833 |
| immh103\_M\_UN\_29876 | 103 | immature | IgM | IGHV3-15 | IGHJ6 | IGHD2-2 | 27 | 4.31 | 0.340741 | 0.566296 |
| immh149\_M\_V3\_13110 | 149 | immature | IgM | IGHV3-21 | IGHJ6 | IGHD2-2 | 23 | 6.23 | 0.847826 | -0.44217 |
| preh138\_M\_V3\_494120 | 138 | preB | IgM | IGHV3-15 | IGHJ6 | IGHD2-2 | 24 | 6.34 | 0.379167 | 0.625833 |
| preh149\_M\_UN\_713160 | 149 | preB | IgM | IGHV3-15 | IGHJ6 | IGHD2-2 | 23 | 6.34 | 0.491304 | 0.283043 |
| preh138\_UN\_V3\_212530 | 138 | preB | NA | IGHV3-15 | IGHJ6 | IGHD2-15 | 24 | 3.49 | 0.2 | 0.23125 |
| immh103\_M\_V3\_24476 | 103 | immature | IgM | IGHV3-15 | IGHJ6 | IGHD2-15 | 24 | 6.07 | 0.466667 | 0.046667 |
| preh119\_M\_V3\_110284 | 119 | preB | IgM | IGHV3-15 | IGHJ6 | IGHD2-2 | 23 | 6.5 | 0.395652 | -0.24478 |
| immh159\_M\_UN\_59280 | 159 | immature | IgM | IGHV3-15 | IGHJ6 | IGHD2-2 | 28 | 5.28 | 0.496429 | -0.47393 |
| immh103\_M\_V3\_22024 | 103 | immature | IgM | IGHV3-15 | IGHJ6 | IGHD2-2 | 23 | 4.11 | 0.721739 | 0.076087 |
| immh159\_M\_V3\_42289 | 159 | immature | IgM | IGHV3-23 | IGHJ6 | IGHD2-15 | 25 | 4.11 | 0.628 | -0.1492 |

Chi-square shows that there are significantly fewer post-central-tolerance cells in these 62 than would be expected by chance:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Pre-B | Immature | Transitional | Naïve | Total |
| Observed | 35 | 19 | 4 | 4 | 62 |
| Expected | 26.464 | 13.902 | 6.722 | 14.912 | 62 |

**Table 1:** Observed count of each cell type in the 62 entries remaining when the heavy chain database is filtered based on: IGHV3, IGHD2, IGHJ6, CDRH3>22AA, GRAVY index >0.163 and Boman index<0.712. The expected row shows the expected proportion of 62 entries to be made up of each cell type, if 62 entries were chosen from the unfiltered dataset at random. A chi-square test showed that the difference between the observed and expected counts is significantly different (p<0.01), indicating that there are significantly more pre-central tolerance cells and significantly fewer post-central tolerance cells than would be expected by chance, indicating that by filtering entries on these properties, I am choosing antibodies that are selected against at central tolerance and are therefore more likely to be autoreactive.