Supporting Information

Gadala-Maria et al. 10.1073/pnas.1417683112

SI Text

Polymorphism Signal Strength in the Case of Heterozygote Alleles. Fig. 3 shows that the perceived mutation frequency of polymorphic positions decreases very rapidly as a function of the IGHV mutation count when the gene is heterozygous for the novel allele. This behavior results from the interaction of two properties of the Ig sequence data: (*i*) when considering total B cells, the frequency of sequences carrying different numbers of mutations is skewed, with a peak at 0 and a rapidly decreasing frequency as mutation counts increase (Fig. S2, solid black lines); and (*ii*) for novel alleles, this distribution will be shifted to the right because polymorphic positions will be perceived as additional mutations (Fig. S2, solid gray lines). In the heterozygous case where a known and novel allele are close enough to be assigned the same IGHV segment allele, the observed frequency of sequences carrying the polymorphism will be a combination of these two distributions. As a result of the skew and shift, a disproportionate number of observed sequences are contributed by the novel allele at low mutation counts (assuming similar use of the two alleles) (Fig. S2, dashed lines). This process causes the perceived mutation frequency of polymorphic positions to decrease much more rapidly than observed in the homozygous case, where the decrease is driven only by mutations at the polymorphic positions (shown in Fig. 3).

Whereas the results shown in Fig. 3 are based on equal allele frequencies (i.e., 50% known and 50% novel), the predicted patterns will be qualitatively conserved for a wide range of novel allele frequencies. However, the pattern will be shifted down along the *y* axis as the novel allele is used less frequently. This behavior means it will be harder to discover novel alleles that are used at low frequencies (because the *y*-intercept may fall below the required threshold). However, simulations suggest that novel alleles should still be detectable at a frequency of one in eight, the minimum frequency required by the genotype inference stage.

Putative IGHV Alleles Excluded from Table 2. Six alleles identified by TIgGER were excluded from Table 2, as several lines of evidence indicated that these are artifacts. Five of these putative alleles differed from known germlines at the same position (position 56) with the same mutation (a T-to-C transition), even though the nearest known germlines were different in each case (*IGHV3-21*01*, *IGHV3-23*01*, *IGHV3-48*02*, *IGHV3-66*02*, and *IGHV3-74*01*). All of these false positives were found in 454 data from subject 420IV. Two of these alleles were also detected in 454 data from subject PGP1, but none were found in the MiSeq data from subject PGP1 (MiSeq data were not obtained from subject 420IV). The sixth allele was observed in subject M2, differing from *IGHV4-39*01* at position 278 by a C-to-T transition. In addition, the *y*-intercepts for these putative polymorphisms (the result of a linear fit to a plot of mutation frequency at the polymorphic position versus mutation count in the entire V, as described in the main text) were low (range of 0.14–0.28). Furthermore, these putative germline alleles constituted less than 4% of unmutated alleles aligning to their respective genes. Thus, all of these putative alleles are excluded from the final genotype determined by TIgGER, which uses a 12.5% cut-off for genotype inclusion.

Manual inspection of the raw sequences carrying the putative polymorphism at position 56 revealed that in nearly every case IMGT/HighV-QUEST had attempted to remove an apparent sequencing artifact (this was usually an indel for sequences derived using the 454 platform). The microsequence flanking position 56 (underlined) in the IMGT reference alleles was TCCCTGAGA, whereas the observed sequence submitted to IMGT was TCCCGTCGAGA. IMGT/High-VQUEST interpreted the G and T nucleotides as insertions, and corrected the sequence to TCCCCGAGA, thus producing the T-to-C transition "mutation." If, instead, the G and T were identified as the insertions, then the sequence would perfectly match the germline allele in the database. Raw sequences from all of the other polymorphic sequences identified by TIgGER were manually analyzed to confirm the absence of nearby indels.

In contrast to position 56, manual inspection of the sequences carrying the putative polymorphism at position 278 did not reveal the possibility of a sequencing artifact. However, the extremely small number of sequences that perfectly matched the putative germline (39 sequences) compared with those that perfectly matched the database version of the allele (1,156) supports the classification of this position as a false-positive result, likely because of random fluctuations. In any case, this putative novel allele, representing <4% of the sequence assigned to the gene, is excluded at the genotype inference step.

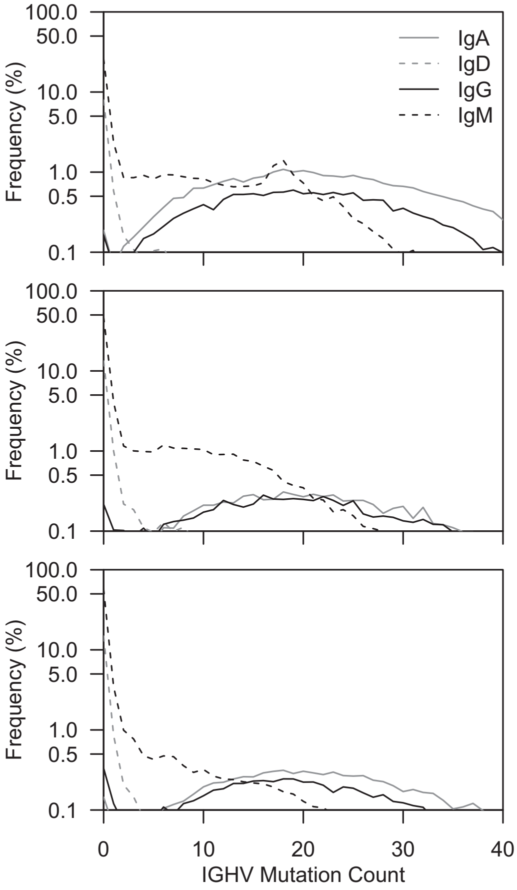

Fig. S1. Distribution of IGHV mutation counts. Sequences generated by 454 were compared with their germline IGHV alleles reported by IMGT/High-VQUEST to identify mutations in 454 data from subjects hu420143 (*Top*), PGP1 (*Middle*), and 420IV (*Bottom*).

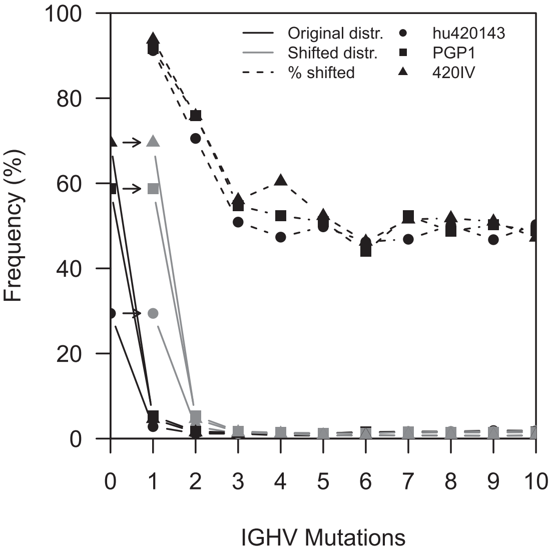

Fig. S2. Expected polymorphism signal strength for heterozygous alleles. Sequences generated by 454 were compared with their germline IGHV alleles reported by IMGT/High-VQUEST, and the frequency of sequences carrying different numbers of IGHV mutations was determined (solid black lines). The frequency of a novel allele with a single nucleotide polymorphism was estimated by shifting this distribution right by one additional IGHV mutation (solid gray lines). Finally, the perceived frequency of the polymorphic position (dashed lines) was estimated as the portion of sequences that would be attributed to the shifted distribution (solid gray lines) if it were combined with the original distribution (solid black lines).

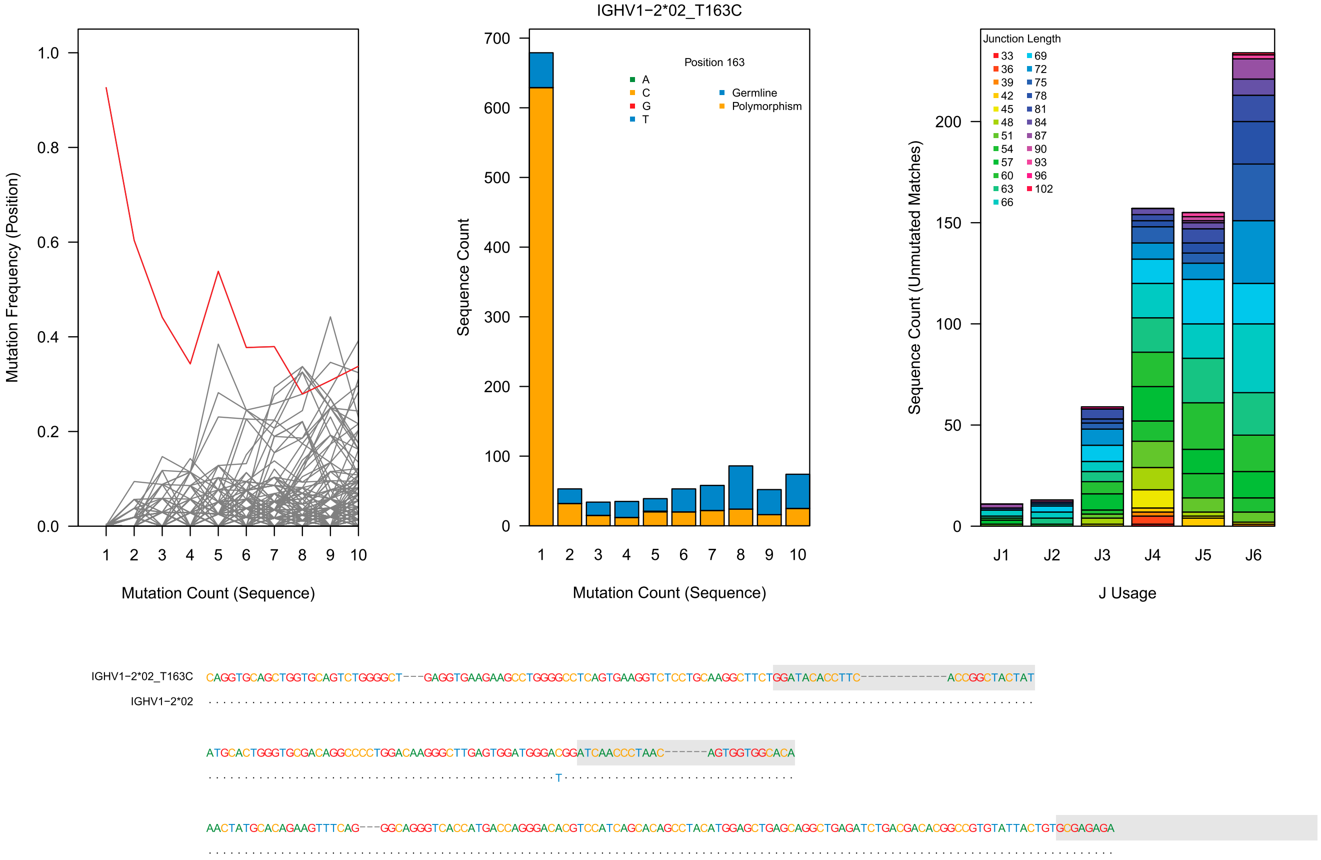

Fig. S3. Supporting evidence for IGHV polymorphism detection provided by TIgGER. TIgGER uses several criteria to identify novel IGHV alleles. For each prediction, a summary of the analysis consisting of four plots is provided in a single report generated by the software. First, the mutation frequency is plotted as a function of IGHV-wide mutation count (*Upper Left*, analogous to Fig. 3). The putative polymorphisms, based on the *y*-intercept of a linear model fit, are shown in red. Second, the distribution of sequences as a function of mutation count is shown, with each bar colored according to the nucleotide distribution at the hypothesized polymorphic position (*Upper Center*). Third, among sequences which perfectly match the hypothesized novel allele, the distribution of J use and junction lengths is shown (*Upper Right*). Finally, the nucleotide sequence of the hypothesized novel allele is shown in comparison with the nearest database allele (*Lower*).

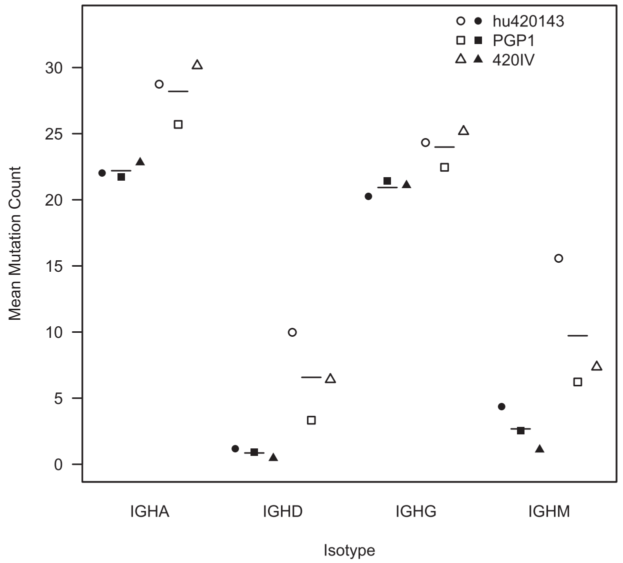

Fig. S4. Mutation counts are higher in sequences initially assigned to alleles that TIgGER excludes from the genotype. Sequences from the three subjects sequenced by 454 were divided into two groups depending on whether the IGHV allele initially assigned by IMGT/High-VQUEST was included (solid points) or excluded (open points) from the personalized germline database determined by TIgGER. Sequences with multiple allele assignments were included in the latter group if any of the alleles were excluded by TIgGER. Mutation counts were determined using the final V(D)J assignments provided by TIgGER.

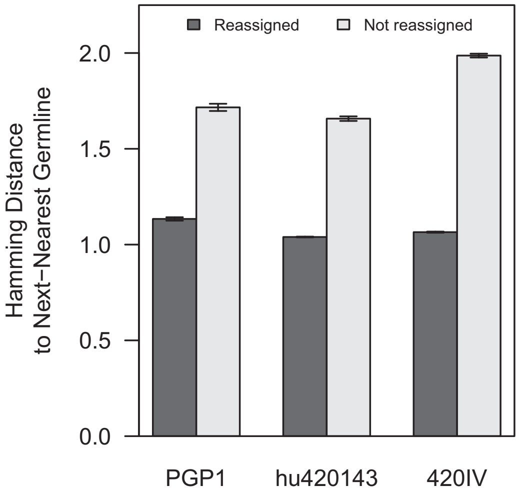

Fig. S5. Alleles initially assigned to reassigned sequences are more similar to other alleles. Sequences from the three subjects sequenced by 454 were divided into two groups, depending on whether the IGHV alleles initially assigned by IMGT/High-VQUEST were reassigned by TIgGER (dark gray bars) or not (light gray bars). For each group, the Hamming distance between the initially assigned germline allele and nearest germline allele was calculated (see Fig. 4 for methods). Error bars represent the 95% confidence interval.

Table S1. Personalized IGHV genotypes inferred by TIgGER

IGHV type	hu420143	PGP1	420IV	M2	M3	M4	M5
IGHV1-2	02, N	02, 04	02, 04	04	02, 04	02, N	N
IGHV1-3	01	01	01	01	01	01	01
IGHV1-8	01	01, N	01	01	01	01	01
IGHV1-18	01	01	01	01, N	01	01	01
IGHV1-24	01	01	01	01	01	01	01
IGHV1-46	01, 03	01	01	01, 03	01, 03	03	01
IGHV1-58	01, 02	01, 02	02	01, 02	01, 02	02	01
IGHV1-69	01, 06	01, 04, 06	01, 02, 06	01/12	01/12, N	02, 04/09	01/12, 06
IGHV2-5	01	01	01		01	01	
IGHV2-26	01	01	01	01	01	01	01
IGHV2-70	01, 04	01, 04, N	01, 04	01	01, 04	11	01, 04
IGHV3-7	01	01	01, 02	01, 03	01	01	01
IGHV3-9	01	01, N	01	01	01	01	01
IGHV3-11	01	01, 04	01, N	01	01	01	01
IGHV3-13	01	01, 04	01, 04	01	01	01	01
IGHV3-15	01	01	01	01	01	01	01
IGHV3-20	01	01, N	01		01		01
IGHV3-21	01	01	01	01/02	01/02	01/02	01/02
IGHV3-23	01	01	01	01/04	01/04	01/04	01/04
IGHV3-30	18	18	18	01, 18	18	04	18
IGHV3-30-3	01	01			01		01
IGHV3-33	01	01	01	01	01	01	01
IGHV3-43	01	01, 02	01, N				01
IGHV3-48	01, 03	02, 03	02, 03	02	01, 02	01	01, 02
IGHV3-49	03, 04	04, 05	03, 04	04, 05	03, 05	03	03, 05
IGHV3-53	01	01, 04	02, 04	01/02, 04	01/02		01/02
IGHV3-64	01	01	01	N		01	
IGHV3-66		01, 01/04	02	02	02	01, 01/04	
IGHV3-72	01		01	01	01	01	01
IGHV3-73	02	01, 02	01, 02		01/02	01/02	01/02
IGHV3-74	01	01	01	01, 01/02	01, 01/02	01, 01/02	01, 01/02
IGHV4-4	02, 07	02, 07	02, 07	02	02, 07	02, 07	02
IGHV4-30-2	01	01			01		01
IGHV4-30-4	01	01			01		01
IGHV4-31	03	03	03	01, 03	03	03	03
IGHV4-34	01	01	01	01/02	01/02	01/02	01/02
IGHV4-39	01	01	01	01			01
IGHV4-59	01, 08	01, 08	01, 08	01	01	01, 08	01
IGHV4-61	01	01, 02	02	01, 02	01, 02		01
IGHV5-51	03	01	01, 03	01	01, 03	03	01, 03
IGHV5-10-1			01	01/03, 03			
IGHV6-1	01	01	01	01	01	01	01
IGHV7-4-1	02			01		02	02

For each row representing an IGHV gene, the alleles of that gene carried by each individual in our study are indicated in the column. Commas separate heterozygous alleles, and slashes separate alleles that could not be differentiated because of the position of primers. Novel alleles identified by this study are indicated as N. Gaps in the table represent cases in which too few sequences aligned a to particular gene, and may indicate that the gene is not present in the indicated individual. Most of these excluded genes are known to be deleted in certain haplotypes (1, 2).

1. Watson CT, Breden F (2012) The immunoglobulin heavy chain locus: Genetic variation, missing data, and implications for human disease. *Genes Immun* 13(5):363–373.

2. Kidd MJ, et al. (2012) The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J Immunol* 188(3):1333–1340.