Mosaic deletion patterns of the human antibody heavy chain gene locus as revealed by Bayesian haplotyping

Moriah Gidoni[1], Omri Snir[2], Ayelet Peres[1], Pazit Polak[1], Ida Lindeman[2], Ivana Mikocziova[2], Vikas Kumar Sarna[2], Knut E. A. Lundin[2], Christopher Clouser[3], Francois Vigneault[3], Andrew M. Collins[4], Ludvig M. Sollid[2], and Gur Yaari[1]

[1]Faculty of Engineering, Bar Ilan University, Ramat Gan 5290002, Israel

[2]KG Jebsen Centre for Coeliac Disease Research and Department of Immunology, University of Oslo and Oslo University Hospital, 0372 Oslo, Norway

[3]AbVitro, Inc., Boston, MA, USA

[4]School of Biotechnology and Biomolecular Sciences, University of NSW, Kensington, Sydney, NSW 2052 Australia

Abstract

Analysis of antibody repertoires by high-throughput sequencing is of major importance in understanding adaptive immune responses. Our knowledge of variations in the genomic loci encoding antibody genes is incomplete, mostly due to technical difficulties in aligning short reads to these highly repetitive loci. The partial knowledge results in conflicting *V-D-J* gene assignments between different algorithms, and biased genotype and haplotype inference. Previous studies have shown that haplotypes can be inferred by taking advantage of *IGHJ6* heterozygosity, observed in approximately one third of the population. Here, we propose a robust novel method for determining *V-D-J* haplotypes by adapting a Bayesian framework. Our method extends haplotype inference to *IGHD*- and *IGHV*-based analysis, thereby enabling inference of complex genetic events like deletions and copy number variations in the entire population. We generated the largest multi individual data set, to date, of naïve B-cell repertoires, and tested our method on it. We present evidence for allele usage bias, as well as a mosaic, tiled pattern of deleted and present *IGHD* and *IGHV* nearby genes, across the population. The inferred haplotypes and deletion patterns may have clinical implications for genetic predispositions to diseases. Our findings greatly expand the knowledge that can be extracted from antibody repertoire sequencing data.

Introduction

The success of the immune system in fighting evolving threats depends on its ability to diversify and adapt. In each individual, a repertoire of extremely diverse antigen receptors is carried by T cells and B cells. In B cells, the antigen receptor is a membrane bound immunoglobulin. In effector B cells, i.e. plasma cells, the immunoglobulins are secreted as antibodies to survey the extracellular environment. Antibodies are symmetric molecules with a constant and a variable region. They are built from two identical heavy chains and two identical light chains. The heavy chains are assembled by a complex process involving somatic recombination of a large number of germline-encoded *IGHV, IGHD*, and *IGHJ* genes (for simplicity we will refer to them as *V, D,* and *J* from now onwards), along with junctional diversity that is added at the boundaries where these genes are joined together[1]. Pathogenic antigens are first recognized by lymphocytes carrying these relatively low affinity receptors. Following initial recognition, B cells undergo affinity maturation, which includes cycles of somatic hypermutation and affinity-dependent selection[2].

Thus, the antibody repertoire of an individual stores information about current and past threats that the body has encountered. Studying this diverse repertoire can teach us about fundamental processes underlying the immune system in healthy individuals[3], as well as reveal dysregulation in autoimmune diseases[4–6], infectious diseases[7–9], allergy[10], cancer[11,12], and aging[13].

Dramatic improvements in high-throughput sequencing (HTS) technologies now enable large-scale characterization of adaptive immune receptor repertoires (AIRR-seq)[14,15]. Extracting valuable information from these sequencing data is challenging, and requires tailored computational and statistical tools which are being constantly developed[16]. Much is being invested, especially by the AIRR community[17], in the collection and standardization of data preprocessing and analysis.

Correct assignment of antibody sequences to specific germline *V, D,* and *J* genes is a critical step in AIRR-seq analysis. For example, it is the basis for identifying somatic hypermutation, pairing biases, N additions and exonuclease removals, determination of gene usage distribution, and studying the link between AIRR-seq data and clinical conditions. Only very few complete or partial sequences of these loci in the human genome have been published thus far[18–22]. The reason for this insufficiency is that these are extremely long ($\sim 1.2Mb$) complex regions with many duplications, which impedes usage of traditional methods for sequencing and data interpretations. Because of the difficulty in performing physical sequencing of these loci, several computational tools have been developed for personal genotype inference from AIRR-seq data[3,23–25].

Although germline genotyping by itself is extremely helpful, deeper insight can be gained by going one step further and inferring chromosomal phasing (haplotyping). Since each antibody chain is generated from a single chromosome, it is important to know not only the presence of genes, but also their combination on the chromosomes. For example, inference of haplotype can provide much more accurate information regarding gene deletions and other copy number variations. These appear to be highly common, as shown by Watson et al.[18] by one complete and nine partial haplotype sequencing of the genomic region encoding the antibody heavy chain locus, using BACs and fosmids.

Haplotyping can be computationally inferred from antibody repertoire sequencing data, using a heterozygous *V/D/J* gene as an "anchor" to define the chromosomes. So far, a statistical framework for haplotyping has been developed for *J6*[26,27], which is heterozygous in $\sim 30\%$ of people (alleles *J6*02* and **03*). Here, we show that reliable haplotyping can also be performed using *D* or *V* genes as anchors (figure 1). Haplotype inference is performed using a Bayesian approach, and follows an initial deletion identification step based on a binomial test applied to gene usage. Using *D* or *V* genes as anchors also enables the *J* distribution to be examined, and expands the percentage of the population for which it is possible to infer haplotype. We present evidence for allele usage bias, as well as interesting mosaic-like deletion patterns that are common in many individuals and involving multiple genes. The dataset that we produced is the first of its kind in size and accuracy, and our findings greatly expand the knowledge that can be extracted from antibody repertoire sequencing data.

Results

Relative gene usage can indicate gene deletions

Naïve B cells from 100 individuals were sorted and their antibody heavy chain variable regions were sequenced using a unique molecular identifier protocol. This data allows us to infer the genetic variability of the antibody heavy chain locus across the largest cohort to date. We exploited the fact that only naïve cells were sequenced to infer and study the characteristics of their germline IGH locus. After filtering out six samples with low coverage (< 2000 sequences), personal genotypes of the IGH regions were constructed using a Bayesian genotype approach[28]. To eliminate further potential biases genotype construction was based on unique sequences with at most three mutations in their *V* region and no mutations in the *D* region. Furthermore, only sequences with single assignments for the *V, D,* and *J* genes were used, since sequences with multiple assignments may introduce biases (Supplementary table 1). In agreement with previous studies[23], genotyping resulted in a five-fold reduction in multiple assignments of a sequence for *V* genes, and two-fold reduction for *D* genes. This reduction was observed by genotyping sequences that were aligned using three different tools: IgBLAST[29], IMGT HighV-QUEST[30], and partis[25] (Supplementary figure 1A). $\sim 2\%$ of sequences were initially assigned to genes that were removed during genotyping. They were thus reassigned to genes present in the subject (Supplementary figure 1B).

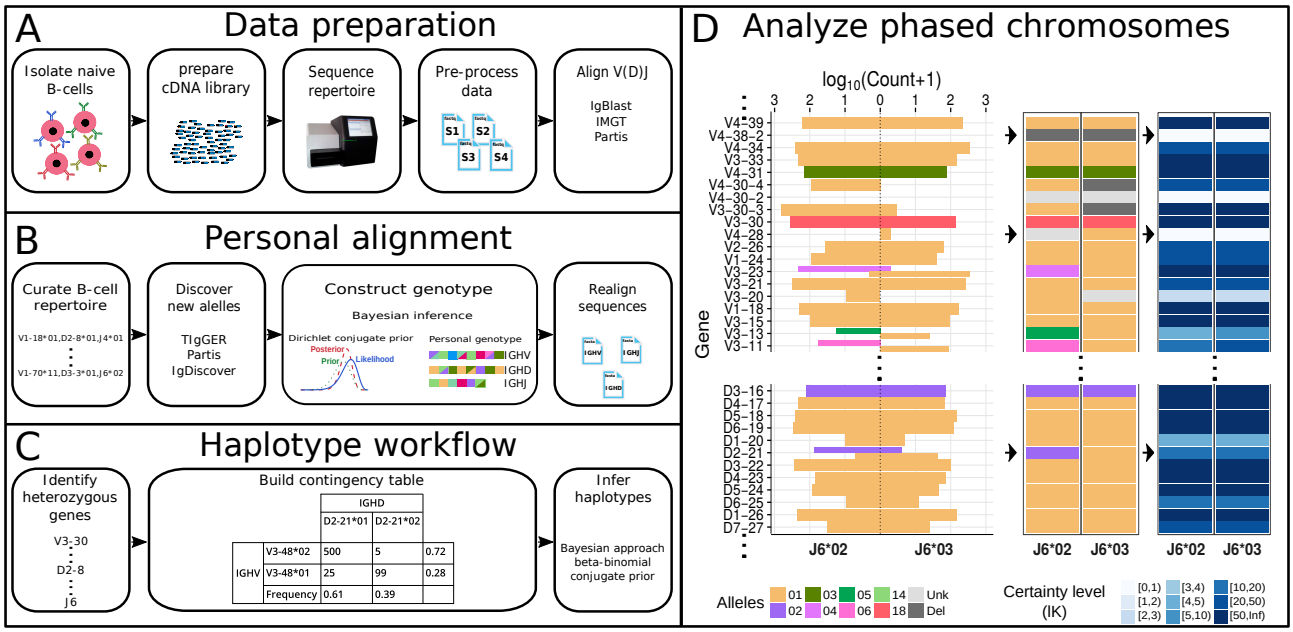


Figure 1: **Schematic view of the haplotype inference process** (A) Naïve B-cells are first isolated, followed by RNA purification. cDNA libraries are prepared, sequenced, the resulting data are pre-processed, and initial *VDJ* alignment is performed. (B) Novel alleles are discovered, and the genotype is constructed. An additional *VDJ* alignment is performed using the constructed genotype. (C) Contingency tables are constructed for *J-V/J-D/V-D* heterozygous gene pairs, and the haplotype is inferred using a Bayesian approach (see methods). (D) The output can be presented as a phased chromosomes graph. The left panel shows the count of each gene (Y axis) that is associated with its paired anchor gene (X axis). In this example the anchor gene is *J6*, and the inferred haplotypes are the *V/D* genes. Colors correspond to the different alleles. The thickness of the count bar is inversely proportional to the number of alleles found on the chromosome. The middle panel shows the called *V/D* haplotype, and the right panel shows the certainty level (lK) for each haplotype decision. The full haplotype output for the individual in this example is shown in figure 7A.

Deletion patterns of neighboring genes

Next, we wished to compare the relative usage of different antibody genes across the population. Applying a binomial test (see methods), we identified deletions in many individuals and multiple genes (figure 2A and 2B). Genes with extremely low expression across all samples were considered indeterminable (NA). In particular, *V1-45, V4-28,* and *D6-25* have very low expression across the vast majority of individuals. It could be that these genes occur only in a very small fraction of individuals with extremely low prevalence. Another possibility is that these are non-functional genes. Looking at the deletions of each sample by itself, several interesting patterns for groups of neighboring genes that are deleted together are observed along the locus (figure 2C and 2D). The most prominent examples are: i) In 46 of the 47 individuals that lack *V2-70D*, the adjacent gene *V1-69-2*, is also deleted. ii) In 16 of the 17 individuals that lack *V4-30-2*, the adjacent genes *V4-30-4* and *V3-30-3* are also deleted. *V3-30-5* is located between *V4-30-4* and *V3-30-3*, we could not infer its deletion, since *V3-30-5* alleles cannot be differentiated from those of *V3-30*. iii) Out of 57 individuals that lack *V3-43D*, 56 lack also *V4-38-2*. The sample that lacks only *V3-43D* had a low relative usage of *V4-38-2*, which was very close to the deletion threshold, but due to a small sample size, could not be inferred as deleted. iv) Two pairs of genes, *V3-9* and *V1-8*, and *V5-10-1* and *V3-64D*, are deleted in a mutually exclusive manner. This pattern has previously been observed for single haplotypes[18,31,32]. Here we show the prevalence of this pattern among a large cohort.

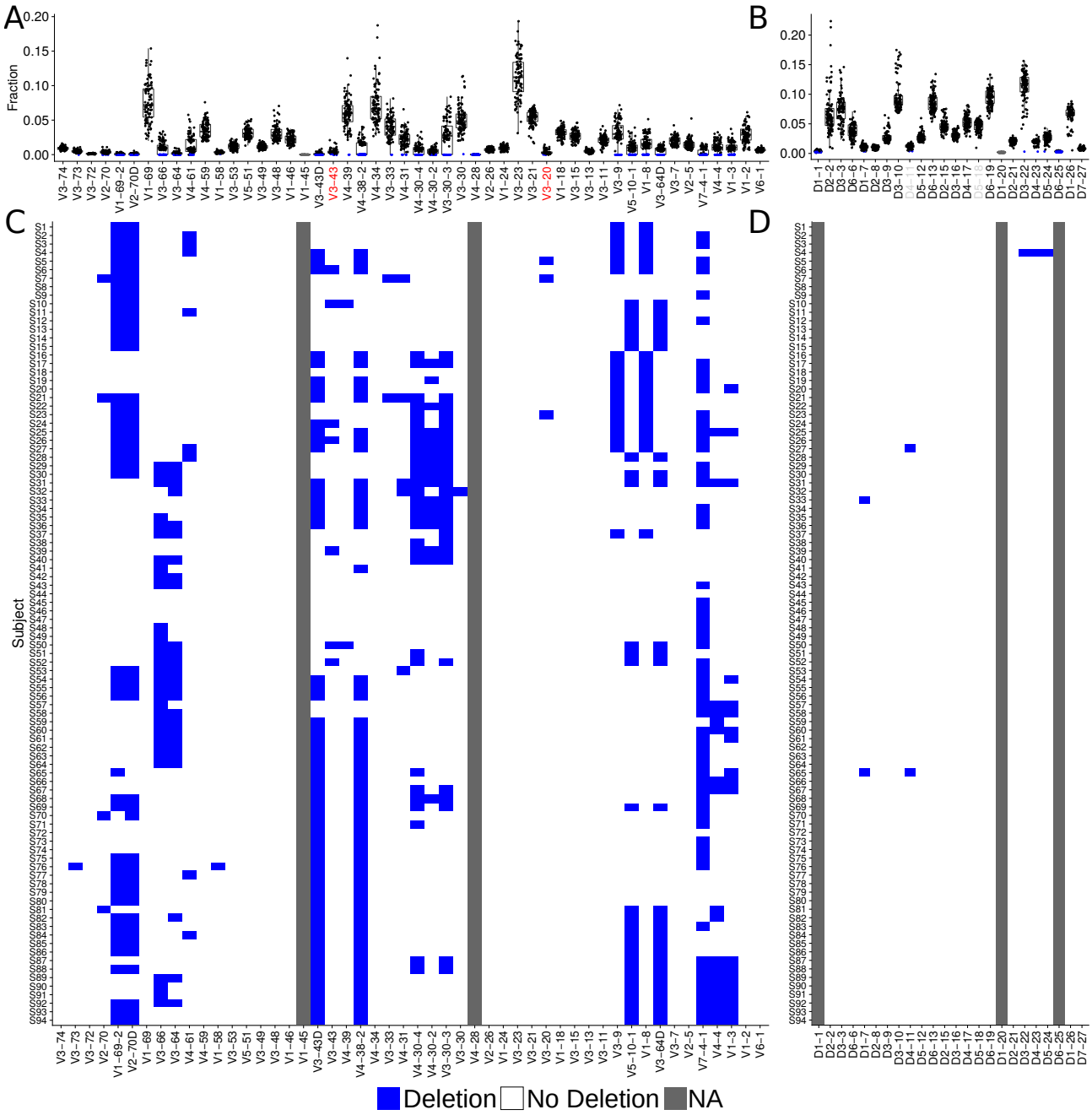


Figure 2: **Gene deletion inference by relative gene usage.** (A, B) Box plots of relative gene usage, where each dot represents a single individual. Blue represents deleted genes according to the binomial test (see methods). The order of genes here and in the following figures is based on their chromosomal location[18]. *V* gene labels marked in red represent low expressed genes for which deletions are inferred with low certainty. *D* gene labels marked in light gray represent indistinguishable genes due to high sequence similarity, therefore alignment call is less reliable. (C,D) Each row corresponds to an individual. Blue represents a deletion on both chromosomes. Gray represents a gene that was not expressed in more than 90% of individuals (marked as NA). Order of rows was determined based on a hierarchical clustering analysis performed in R using the heatmap function (see methods). Box plots elements here and onwards are: center line, median; box limits, upper and lower quartiles; whiskers, $1.5x$ interquartile range; points, all samples.

Ig heavy chain gene heterozygosity landscape

Inference of a personal genotype allows us to estimate the heterozygosity of genes in the population. We considered genes for which more than one allele is carried by an individual as heterozygous. Up to four distinct alleles in an individual's genotype were allowed, where four alleles would correspond to a mis-named gene duplication with both genes being heterozygous and without sharing between the genes (figure 3). It has been previously shown that approximately one third of the population is heterozygous for *J6*[26,27]. Our cohort agrees with this observation with $32/94$ heterozygous samples for the 02 and 03 alleles in this gene, and one individual carries alleles 03 and 04, to combine to a total of 33 heterozygous samples. In addition, we identified a large number of heterozygous *V* genes. Five out of the *V* genes (*V1-69, V3-53, V3-48, V3-49*, and *V3-11*) were heterozygous in more than 50% of the individuals with a defined genotype, and 19 in more than 20%. Three *D* genes, *D2-2, D2-8*, and *D2-21* were determined as heterozygous in $2\% - 36\%$ of the population (2, 16, and 28 individuals respectively after imposing the 30% threshold as described in the methods). In the region between *V1-69* and *V1-46* ($\sim 200K$ base pairs) the fraction of heterozygous individuals is dramatically higher than the surrounding regions (figure 3A). One possible explanation is that this region is a genomic hotspot for germline evolution giving rise to diverse alleles. Within this region, the three genes, *V3-66, V3-64*, and *V4-61* appear as mostly homozygous, i.e. the same allele is present on both chromosomes. However, there are many single chromosome deletions in *V3-66, V3-64* as shown in the following sections (Supplementary figure 2). In the case of *V4-61*, the allele number is less reliable since IMGT may have mis-classified several allele sequences in the *V4-4/V4-59/V4-61* complex.

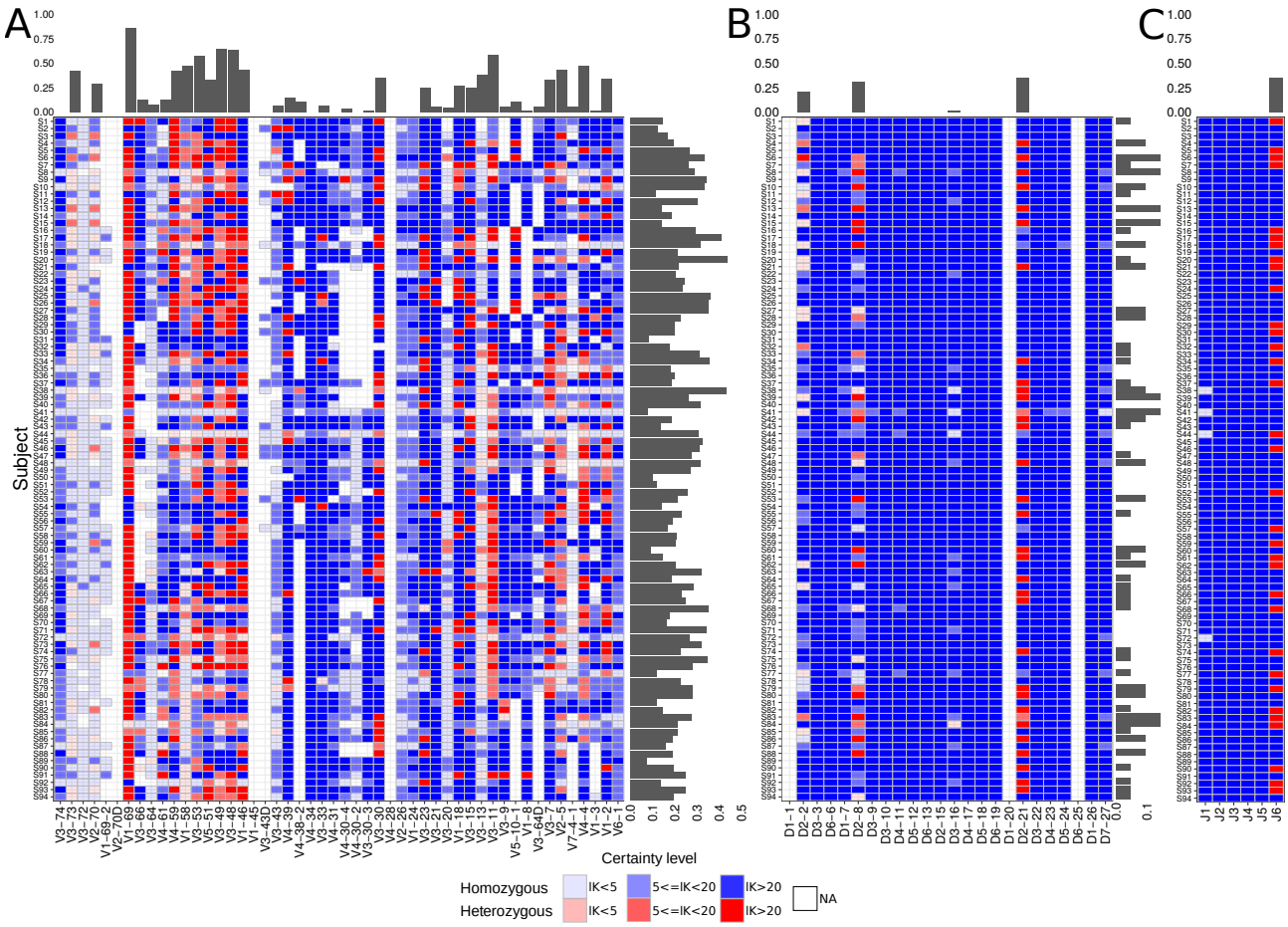


Figure 3: **Heterozygosity of the IGH genes.** Each row represents an individual, and each column represents *V* gene (A), *D* gene (B), and *J* gene (C). Red shades represent heterozygous genes, and blue shades homozygous genes. Transparency corresponds to the certainty level of genotype inference. White represents a gene with too low usage (fewer than 10 sequences, marked as NA) to enable clear genotype inference. Bars on top of each panel represent the ratio between the number of individuals with heterozygous genes and all individuals with a defined genotype for this gene. Bars on the side of each panel represent the fraction of heterozygous genes for each individual out of all genes with a defined genotype.

We next tested whether in heterozygous individuals, expression of both alleles is similar, or biased towards one of them. For each heterozygous gene, the relative usage of each allele was calculated for each individual (figure 4). For the test we considered only samples with a high sequence coverage ($> 10K$), and genes that appeared at a frequency higher than 1%. To statistically address whether there is a biased usage between pairs of alleles that are present in the same individual, a single sample sign test was applied. This test was formulated to consider binary outcomes across the population. For each individual, we asked whether the fraction of the first of the allele pair is larger or smaller than 0.5. Then we noted in how many individuals this fraction is larger than 0.5, and asked how likely this result is to occur by chance. P values were adjusted using the Benjamini-Hochberg method and are referred to as q values. Out of 33 allele pairs (23 genes) that were tested, significant differences were found in 13 allele pairs (11 genes, see figure 4). In 10 allele pairs, the preferred allele was significantly more expressed than its partner in all individuals. The range of allelic preferences observed between different individuals is large, most likely due to factors related to their heterogeneous genetic background.

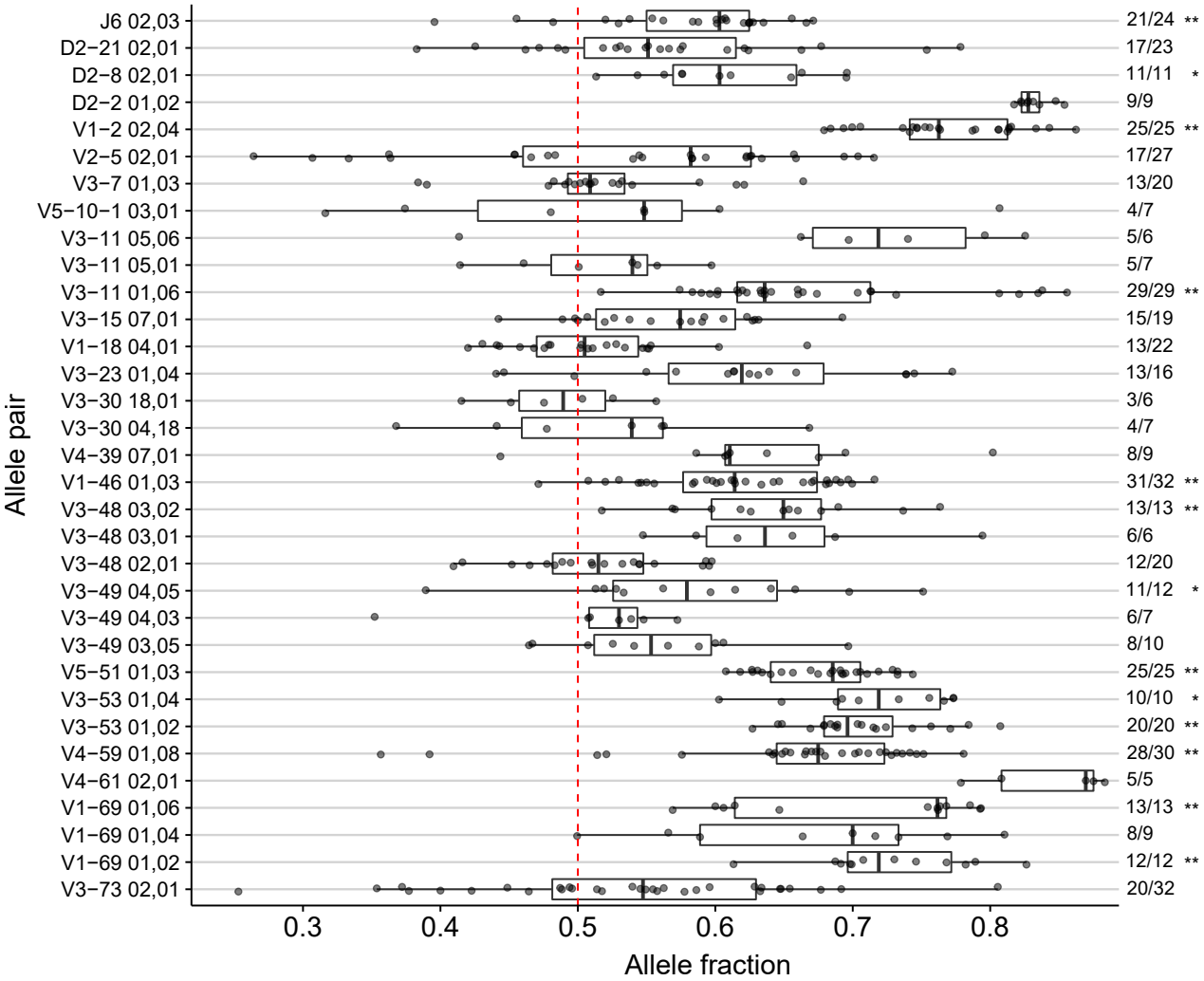


Figure 4: **Biased allele usage.** Box plot of allele pairs relative usage of heterozygous individuals. Only allele pairs which were observed in more than 5 individuals are shown. Each point represents an individual. The allele fraction (X axis) corresponds to the allele that is written first in each row, and is the dominant allele in most individuals. Asterisks indicate allele pairs with a statistically significant difference in the number of individuals with the same dominant allele. Statistical significance was determined with a binomial single sample sign test (see methods, $*$ indicates p value < 0.05, $**$ indicates p value < 0.01).

The single chromosome gene deletion pattern of the antibody loci is mosaic-like

To obtain new insights into the *V* and *D* gene chromosomal distribution in the population, we inferred the haplotypes of the 33 individuals in our cohort that are heterozygous for *J6*. We applied a Bayesian approach described in the methods section, and adapted a threshold on the level of confidence to call a deletion ($lK > 3$). Figure 5A shows the distribution of *V* and *D* deletions along both chromosomes in these individuals. The deletion likelihood is non uniform as there are regions along the chromosomes that are more prone to deletions in both chromosomes, and regions that are less.

To further investigate the patterns of deletion, we generated a heatmap of *V* and *D* deletions (and suspected deletions) for each individual (figure 5B). *V1-45* and *V4-28* are very rare and therefore their single chromosome deletions are hard to call. The heatmap depicts several interesting observations. First, individual S1 has a long deletion stretch in the chromosome carrying *J6*02*, spanning from *V4-28* until *V3-64D*. This region includes 15 *V* genes and over $230K$ base pairs, including the very frequently used *V3-23, V3-21,* and *V3-15*. It will be interesting to research any clinical implications this deletion might have on the people carrying it, and if such deletion in a homozygous setting can exist. This pattern might be not so rare, as another individual has been previously shown to have a similar deletion stretch[27].

Second, similar to the pattern observed in both chromosomes (figure 2), *V3-9* and *V1-8* deletion is mutually exclusive with *V5-10-1* and *V3-64D* deletion, in each of the chromosomes. Almost all individuals have one of these pairs deleted in each of the chromosomes. These genes are located sequentially on the DNA. In fact, in 47 of the 94 individuals who passed the sequencing quality filtering criteria, a deletion in both chromosomes of one of these gene pairs was detected using the binomial test (figure 2). This is consistent with the assumption that all individuals (not only the J6 heterozygous ones) have one of these deletions in each chromosome.

Third, nine individuals have deletions in the adjacent genes *D3-3* and *D6-6*. In fact, this deletion stretch might span also *D1-7* and *D2-8*, but we lack the statistical power to say it with confidence. *D4-4* and *D5-5* have the same sequences as *D4-11* and *D5-18* respectively, and therefore are not presented here (see methods). These genes are located within the above deletion stretch. Such a deletion stretch was shown in a previous study[3]. Out of these nine individuals, eight have also a *V3-9* and *V1-8* deletion, and one individual only has a *V5-10-1* and *V3-64D* deletion (q value of 0.01 by a binomial test). It will be interesting to research the structure of this region in the DNA, and also to find out whether there are any phenotypic differences between these groups.

Fourth, deletions in *D3-22* together with *D1-26* were observed in the *J6*03* chromosome in eight and seven individuals, respectively, and were not observed at all in the *J6*02* chromosome.

Since single chromosome deletion inferred by haplotype analysis, reflects a deletion polymorphism, we integrated these polymorphisms with other types of heterozygousity to better estimate heterozygosity levels throughout the population (Supplementary figure 2).

To verify that the above suspected deletions indeed appear at the genomic level, five selected genes were amplified from genomic DNA (gDNA) of T cells and monocytes using custom-designed gene-specific primers, and PCR products were analyzed by gel electrophoresis. Six individuals from whom gDNA was available were chosen for testing. Primers for amplification and other technical parameters related to the analysis are described in Supplementary table 2. The target amplicons of *V1-8, V3-9, V3-64D, V5-10-1,* and *V4-38-2* were predicted to be 428, 493, 470, 460, and 500 bp long, respectively. All PCR products were verified by Sanger sequencing and the identity of each sequence was subsequently confirmed by IMGT V-Quest[33]. The findings confirm our inferred observations from the cDNA sequencing data analysis (Supplementary figure 3). Individuals S49 and S42 expressed all of the tested genes, and were therefore used as positive controls. Individuals S16 and S4 lack genes V1-8, V3-9 and V4-38-2, but carry V3-64D and V5-10-1. In contrast, individuals S85 and S30 lack genes V3-64D and V5-10-1, but carry V1-8 and V3-9. Moreover, individual S30 also lacks the gene V4-38-2. Two of the subjects, S42 and S49, were found to have one synonymous mutation in the heavy gene V3-64D, which can potentially be a novel allele (Supplementary table 3).

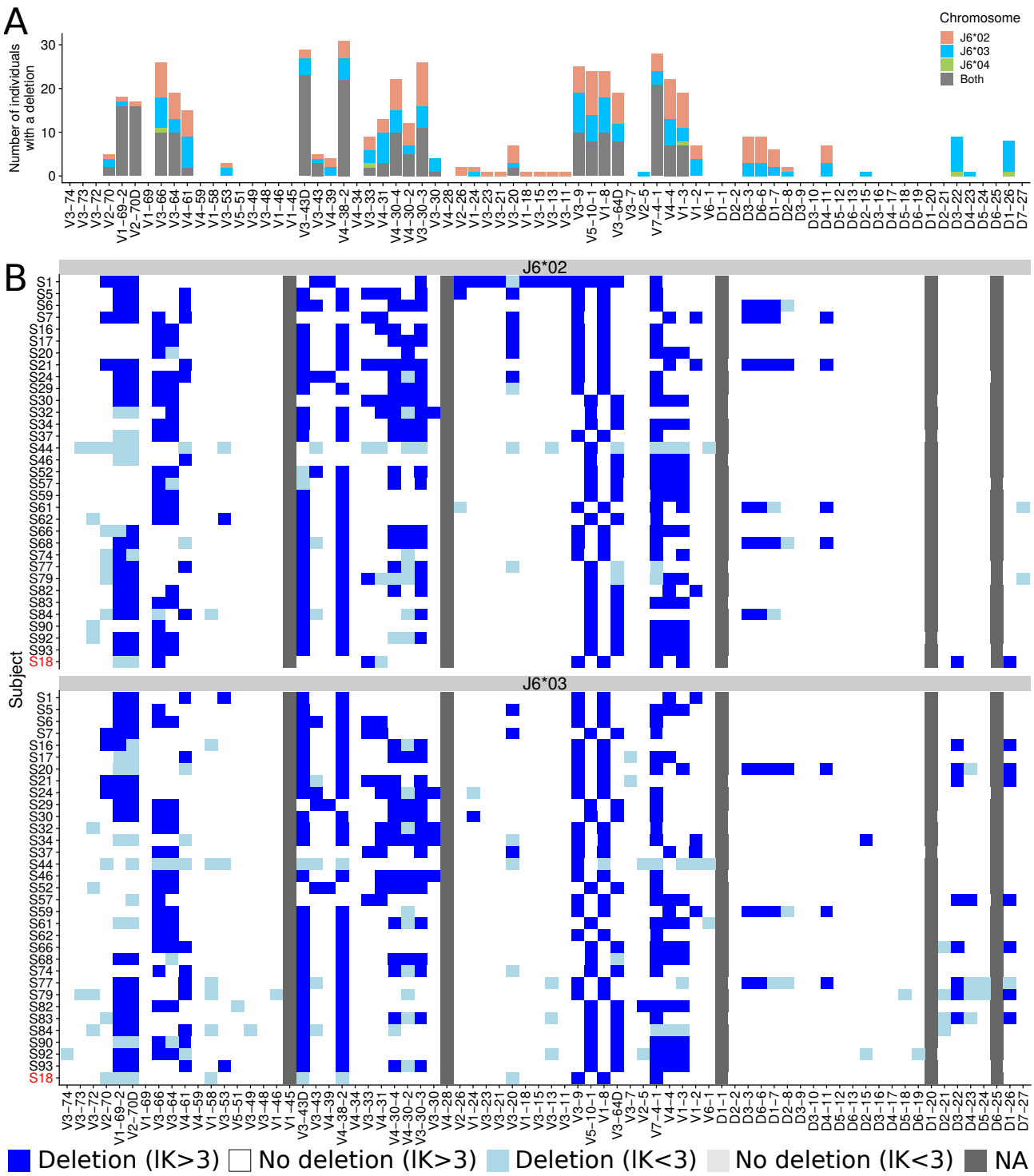


Figure 5: **Gene deletion inference along each chromosome.** (A) The distribution of *V* and *D* gene deletions along each chromosome in 33 individuals that are heterozygous for *J6*, as inferred by haplotype (light red, blue, and green) and by the binomial test (gray) (B) A heatmap of *V* and *D* gene deletions and suspected deletions for each of the 33 heterozygous individuals in *J6*. Each row represents an individual, and each column represents *V* or *D* gene. Blue represents a deletion ($lK > 3$), light blue represents a suspected deletion ($lK < 3$), and light gray represents no deletion on both chromosomes with low certainty ($lK < 3$). Dark gray represents a gene with an extremely low usage across all samples. The top panel represents the chromosome on which *J6*02* is present, and the bottom panel represents the chromosome on which *J6*03* is present. Sample S18, marked in red is heterozygous for *J6*03* and *J6*04*. For this individual, *J6*04* was added to the *J6*02* panel.

Relative gene usage may indicate gene deletions on a single chromosome

Gene deletion identification is of major importance and might have critical clinical implications. In the first section of the results, we proposed to use a binomial test to detect deletions from both chromosomes. Haplotype inference offers an additional approach to detect deletions from one of the chromosomes only. We wished to learn the relative gene usage pattern in *J6* heterozygous individuals with single chromosome deletions. Most *V* and *D* genes showed lower usage when one of the genes was identified as deleted from one of the chromosomes according to haplotype inference (Supplementary figure 4A and 4B).

To assess statistically if relative gene usage can indicate gene deletions on a single chromosome, we took the following approach. For each gene, we divided the individuals into two groups, those with no deletions and those with a single chromosome deletion. For each group, we estimated gene usage with a normal distribution. From these estimations, a threshold was derived to call single chromosome deletions (figure 6A). By considering different thresholds we created receiver operating characteristics (ROC) curve for each gene (figure 6B), from which we extracted the threshold that yields a false positive rate (FPR) of α (Supplementary table 4). The obtained sensitivity distributions for $\alpha = 0.01$ and $\alpha = 0.05$ are shown in figure 6C. To support our assumption, we estimated the ratios between the mean gene usage of individuals with no gene deletion and individuals with a single chromosome gene deletion (figure 6D). The average fold change between the group means is 1.74 ± 0.18 (95% confidence interval), which agrees with the hypothesis that genes that are deleted in a single chromosome are expressed at half the frequency. Figure 6E shows gene usage distributions for the two groups for genes that have at least two individuals for each group with a certainty level of $lK >= 10$, along with the derived thresholds. An interesting exception is *V4-61*, in which the relative usage in individuals with a single chromosome deletion was sometimes higher than in individuals with no deletions. This could be the result of a mis-classification of sequences belonging to the *VH4* family. Over 120 sequences in the IMGT reference directory are unmapped. This is true of a majority of *VH4* sequences, which have been given their gene names by sequence alignment, despite the similarities of the *V4-4*, *V4-59* and *V4-61* genes in particular. The problem with this strategy was recently highlighted with the demonstration that the *V4-59*08* sequence maps to both the *V4-59* and *V4-61* loci[34]. This suggests that an individual with an apparent *V4-61* deletion could carry a misnamed sequence at the *V4-61* locus, and it is certainly possible that the high apparent expression of *V4-61* in an individual with a deletion polymorphism is actually high expression of a misnamed sequence.

When *D3-3* is deleted in one chromosome (in our cohort, this gene was not deleted from both chromosomes in any individual, see figure 2D), it appears to be compensated by higher *D3-10* usage (figure 6F, as suggested in[26]). A cutoff of 0.11 on *D3-10* usage correctly classifies all nine individuals with *D3-3* single chromosome deletions. Applying the same cutoff to *J6* homozygous individuals can thus be extrapolated for identifying *D3-3* single chromosome deletions. As shown above, *D3-3* deletion is accompanied by deletions in *D6-6, D1-7*, and *D2-8* which are harder to detect due to their low usage. Thus, *D3-10* usage higher than 0.11 implies the above *D* gene deletion stretch.

In the previous section we showed that in *J6* heterozygous individuals, the two *D* genes, *D3-22* and *D1-26* were deleted only in the chromosome carrying *J6*03*. Figure 6G shows the relative usage of these genes for all individuals. All *J6*02* homozygous individuals (black) have higher usage than the usage of the individuals carrying *J6*03* with a single chromosome deletion. In addition, a single individual (S16), with the lowest usage frequency in both, *D3-22* and *D1-26* genes, is *J6*03* homozygous and has been determined with *D3-22* to *D6-25* gene deletion according to the binomial test. For this sample, *D1-26* usage is just above the binomial test cutoff (0.0056) for being called as deleted which may imply its deletion if the threshold were determined for each gene independently. Thus, in this cohort, there were no cases in which *D3-22* and *D1-26* were deleted from the chromosome carrying *J6*02*.

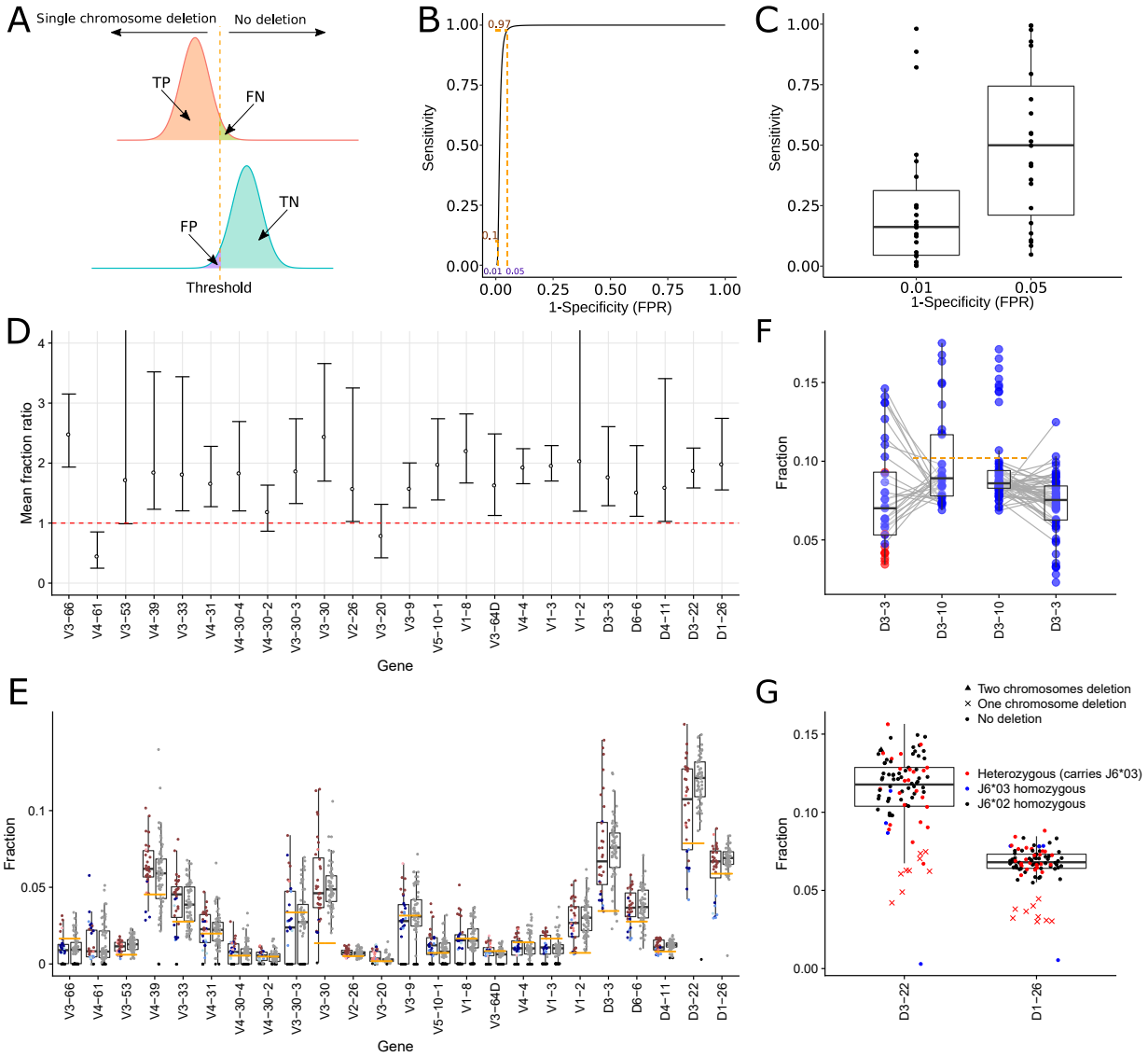


Figure 6: **Inferring single gene deletions by their relative usage.** (A) An example calculation of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) for two t distributions that estimate the gene usage of individuals with zero or one chromosome gene deletion for *V3-66*. (B) An example ROC curve for *V3-66*. (C) Box plots of sensitivity distributions for FPR of $\alpha = 0.01$ (left) and $\alpha = 0.05$ (right) (D) Point estimator and 95% confidence interval for the ratio between the mean gene usage of individuals with no gene deletion and individuals with a single chromosome gene deletion. Genes shown have at least two individuals for each group with a certainty level of $lK >= 10$. (E) Box plots of *V* and *D* gene usage for the genes appearing in D. Each gene distribution appears once for the *J6* heterozygous individuals (left) and once for the *J6* homozygous individuals (right). Each dot represents an individual. For the *J6* heterozygous individuals, color corresponds to gene deletion from both chromosomes (black), single chromosome (blue), or no deletion (red). Shades correspond to the certainty level of deletion inference. The orange cutoffs separate individuals with a single chromosome deletion from individuals with no deletions in that gene, with FPR of 0.05. (F) Box plots of *D3-3* and *D3-10* usage for *J6* heterozygous samples (left) and *J6* homozygous samples (right). Gray lines connect between *D3-3* and *D3-10* relative usage of the same individual. Orange cutoffs separate individuals with high and low *D3-10* usage that correspond to single chromosome deletions in *D3-3*. Blue points represent individuals with no *D3-3* deletion, and red points represent individuals with a single chromosome deletion. (G) Box plots of the usage of *D3-22* and *D1-26* for all individuals. Blue and black points represent homozygous individual in *J6*03* and *J6*02* alleles respectively, red points represent heterozygous individuals carrying *J6*03* allele. The shape of the point represents each individual's gene deletion state.

Finding useful D genes for haplotyping

Compared to *V* and *J* assignments, assigning *D* genes and alleles is challenging and error prone. This is due to the relatively short length of the *D* genes. As noted above, multiple possible assignments are partially resolved by genotyping, especially for *V* and *J* (Supplementary figure 1A). The *D* gene assignment, however, still suffers from significantly lower credibility. We calculated the allele bias present for the three candidate *D* genes that can be used for haplotyping (i.e., are heterozygous in a fraction of the population), and observed a distinct set of individuals with highly biased usage ($\sim 80\%$, see Supplementary figure 5A and 5B). Although we saw similar patterns in other genes (figure 3), for the purpose of *D*-based haplotyping we wanted to be conservative, and exclude individuals who present highly biased usage between the two chromosomes based on their *D* assignments. For this purpose, we built *V* gene haplotypes for a subset of individuals who are heterozygous for *J6* and either *D2-2, D2-21,* or *D2-8*. We have plotted the Jaccard distance between the haplotypes of these individuals as a function of allele bias (Supplementary figure 5C). Based on this analysis we set up a threshold of 30%, above which the Jaccard distance between the haplotypes is expected to be smaller (p value $< 2 \cdot 10^{-4}$ by Wilcoxon test). All of the samples which were initially determined as heterozygous for *D2-2* were set as homozygous after applying the 30% cutoff. Haplotype can be inferred only in individuals who carry heterozygous genes, therefore *D2-21* and *D2-8* emerge here as good candidate anchor genes for haplotyping, due to their relatively high rate of heterozygosity in the population. In our cohort the number of heterozygous individuals increased from 33 to 52 of 94 (Supplementary figure 6A). To test the *D*-based haplotype, we first inferred the haplotype of *D* by *J6* (Supplementary figure 6B). Next we inferred the *V* haplotypes from both anchor genes *J6* and *D2-21* (figure 7A and figure 7B respectively). We then calculated the distances between all resulting haplotypes (Jaccard distance < 0.1 for the same individual, Supplementary figure 5C). These distances were used to generate a hierarchical clustering dendogram (figure 7C). The dendogram shows high similarity between the *J6*- and *D2-21*-based haplotypes inferred for the same individual, compared to haplotypes of other individuals. A similar picture is obtained for the *D2-8*-based haplotypes (Supplementary figure 6C), indicating that these *D* genes can be used for reliable haplotype inference.

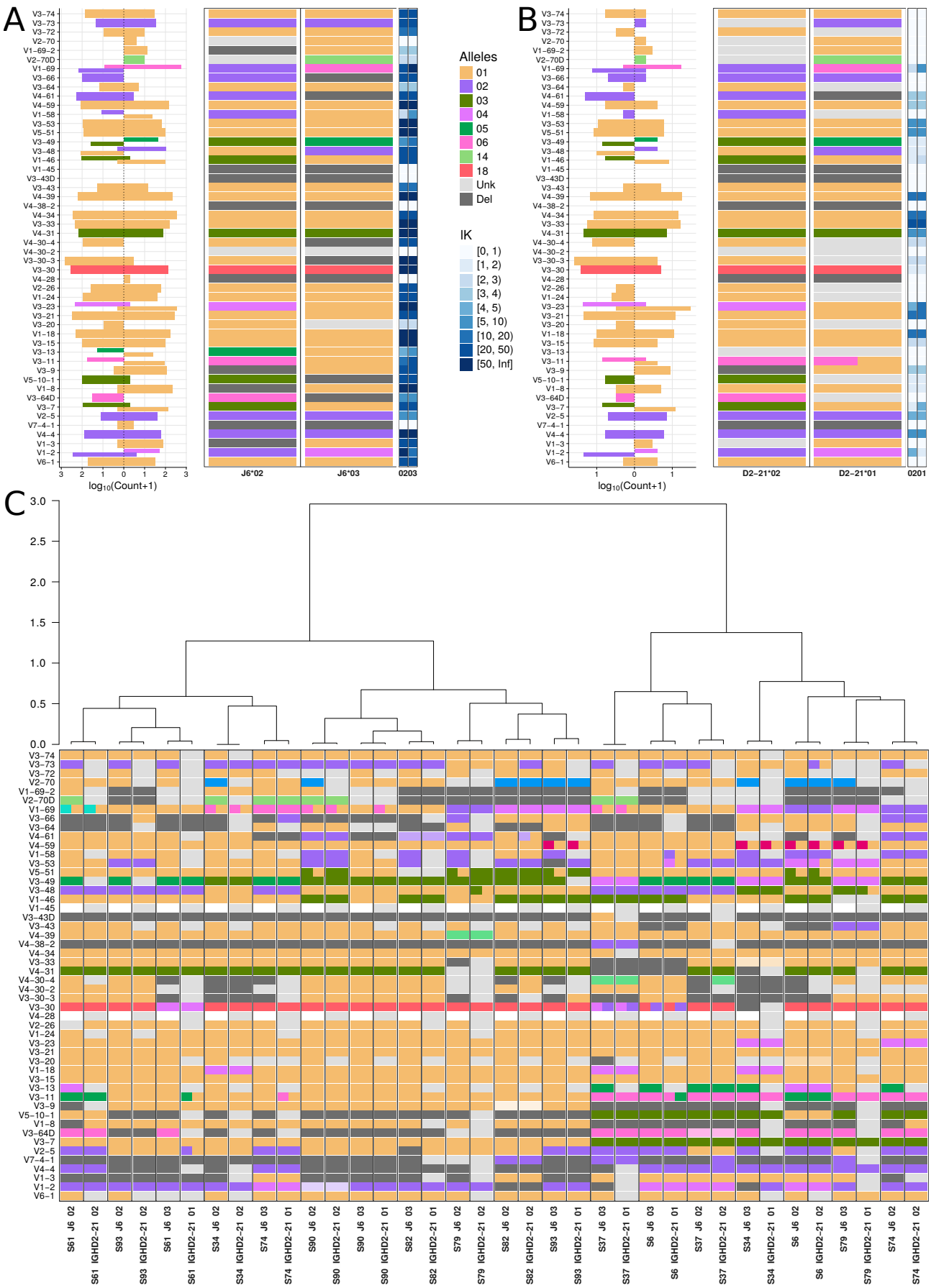


Figure 7: ***D*-based haplotype inference.** (A) Example for *V* haplotype by *J6*. The cutoff between an unknown and a deleted gene on a certain chromosome was determined by $lK >= 3$ (see methods). (B) Example for *V* haplotype by *D2-21*, for the same individual as in (A). (C) V haplotype map where each column corresponds to a *J6*- or *D2-21*-based haplotype of an individual that is heterozygous for both these genes. The order of columns was determined by a hierarchical dendogram based on the distances between individual haplotypes.

D deletion can be detected using *V* haplotype inference

In previous sections we showed how *D* gene deletions can be inferred either from both chromosomes using a binomial test or from a single chromosome by anchor *J6* gene haplotype. As indicated above, *J6* heterozygosity prevalence is approximately one third, leaving most of the population without the possibility to infer single chromosome *D* gene deletions. Since *V* gene heterozygosity is extremely common (figure 3), we pursued the option of inferring a haplotype based on *V* anchor genes. In our cohort, all individuals are heterozygous in at least two *V* genes. Thus, using *V* genes as anchors for haplotype inference could dramatically increase the number of people for which *D* haplotype can be inferred. However, reliable haplotype inference using *V* genes as anchors requires a much greater sequencing depth than haplotype inference using *J6* gene as an anchor. Since there are far more *V* genes than *J* genes, the relative frequencies of the *V* genes are much lower, making a single anchor *V* gene haplotype inference more challenging.

To overcome the low number of sequences that connect a given *V*-*D* allele pair, we applied an aggregation approach, in which information from several *V* heterozygous genes was combined to infer *D* gene deletions. The Bayesian approach utilizing a binomial likelihood and a conjugate beta prior, allows us to use the posterior output of one *V*-based inference as the prior to the next *V*-based inference. We do not know in advance on which chromosome each *V* allele is located. Therefore we attribute all dominant D alleles to the same chromosome. Hence, this V_{pooled} approach is exposed to allele mix, where contradicting chromosomal alleles assignments are inferred by different *V* genes.

To assess the power of the V_{pooled} approach, we compared the resulting *D* gene deletion patterns from V_{pooled} with *J6*. We compared *D* genes with minimum mean relative usage of 1.5% in the 32 *J6* heterozygous individuals (figure 8A left panel, red line). Due to the potential allele mix of the V_{pooled} approach we compared sensitivity and precision for a range of lK cutoffs (figure 8B). We identified an lK value ($lK = 12$) which optimized the precision rate ($\sim 94\%$ for $lK(J) = 2$ and $\sim 84\%$ for $lK(J) = 7$) with an acceptable price in sensitivity ($\sim 46\%$ for $lK(J) = 2$ and $\sim 59\%$ for $lK(J) = 7$). The relatively low levels of sensitivity result from an overall reduction in the number of identified deletions (figure 8C). Using the V_{pooled} anchor approach we were able to correctly identify most of the *D3-3, D6-6, D3-22*, and *D1-26* chromosome deletions (figure 8D). Applying the same approach to the entire cohort, we identified single chromosome *D* gene deletions also in *J6* homozygous individuals (Supplementary figure 7). *V* anchor gene haplotyping provides an important opportunity to identify *D* gene chromosome deletions in a much larger proportion of the population than solely by *J6*. Pooling together several heterozygous *V* genes as in the suggested V_{pooled} anchor approach, increases the power of *D* gene deletion identification for moderate sequencing depths.

Comparison between celiac patients and healthy individuals

Approximately half of the subjects enrolled in the study have celiac disease (52 out of 100), and these subjects were included to represent genetic variation that might be present among patients with this disease. The study was not powered to perform association analysis. Yet looking at differences between the two groups, the most prominent difference was single chromosome deletions of the *D3-22* and *D1-26* genes; *D3-22* was deleted in 8 subjects, 7 of whom are celiac disease patients, and *D1-26* was deleted in 7 subjects, 6 of whom are celiac disease patients (figures 8 and 9). These differences are not statistically significant after multiple hypothesis correction.

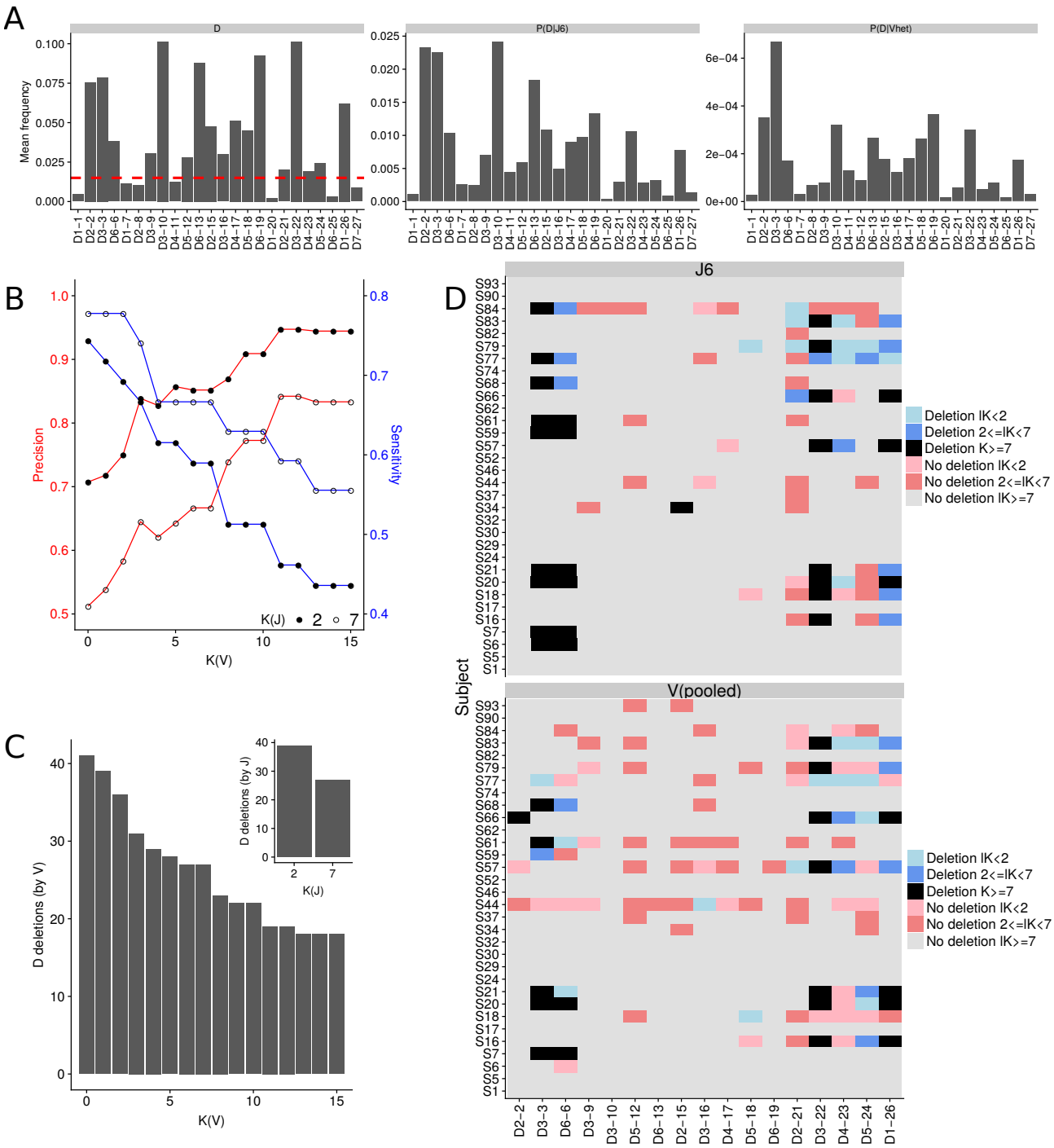


Figure 8: **Gene deletion inference along each chromosome by multiple *V* genes.** A comparison between *D* haplotype inference using a pool of *V* genes vs. *J6* as anchors, in 33 *J6* heterozygous individuals. (A) The mean relative *D* gene usage. Left: mean *D* gene usage. Dashed red line corresponds to the 1.5% threshold which was used to filter out lowly expressed genes for the rest of the analysis presented here. Middle: mean *D* gene usage in sequences containing *J6*. Right: *D* gene usage in sequences containing any heterozygous *V* gene. (B) Precision and sensitivity are described for *D* gene deletions. They are calculated to compare the *D* gene deletions by *J6* as anchor vs. by a pool of *V* genes. Different certainty levels are presented for *V* (X axis). Full circles correspond to $lK(J) > 2$, and empty circles correspond to $lK(J) > 7$. Precision is shown by the red curves and the left Y axis, and sensitivity is shown by the blue curves and the right Y axis. (C) The number of *D* gene deletions inferred by pooled *V* (main graph) and by *J6* (subgraph) as a function of the log of the Bayes factor (lK). (D) *D* gene deletions inferred by *J6* (upper panel) and by pooled *V* (lower panel). Each row represents an individual, and each column represents a *D* gene. Colors correspond to lKs as indicated in the caption. For the presented V_{pooled} approach only heterozygous *V* genes with minor allele fraction larger than 30% were included.

Discussion

Studying the genetic factors that determine the variable regions of B cell and T cell receptors is critical to our understanding of genetic predispositions to diseases. Despite their tremendous importance for the ability of our immune system to fight all sorts of diseases, these regions are understudied and rarely investigated as part of routine disease-association studies. The reason behind this discrimination is technical. The repetitive patterns present in these regions, combined with relatively short reads commonly used in HTS, make it challenging to map them, at both the genotype and the haplotype levels. On the other hand, the technology to produce reliable AIRR-seq data is advancing rapidly, and AIRR-seq studies are gaining popularity. From early days of AIRR-seq studies, ideas about how to connect these data to genotypes and haplotypes were proposed[3,26,24,25,23,35,25]. Here, we implemented similar ideas in a Bayesian framework that allowed us to: 1. Attribute a certainty level to each result, and 2. Infer haplotype based on *V, D,* or *J* genes. We generated the largest multiple individual dataset of naïve B cells, to date, and applied our method to it. Our study revealed many interesting patterns that are present in the antibody heavy chain locus, and should be investigated further in different populations, various clinical conditions, and using different sequencing technologies.

It had been previously demonstrated that there is a strong bias towards usage of particular genes[36] and between *D* and *J* gene recombinations[37]. In this study we have demonstrated an allele usage bias for various *V, D,* and *J* genes. Several hypotheses could explain such biases. The first, and most likely one, is differences in the recombination signal sequence (RSS) associated with alleles of the same gene[26,38,19]. Another possibility may be connected to the physical structure of the chromosomes - for example methylation patterns or other epigenetic modifications. Yet another hypothesis is that these biases result from a negative selection process against self-reactive antibodies. It is plausible that certain allele combinations result in self-reacting antibodies, and hence are excluded from the mature B cell repertoire. Note that the latter explanation is not relevant in all cases, since in three allele pairs (*V1-46*03,01, V4-59*01,08, V5-51*01,03*) the differentiating mutations are silent, i.e. the amino acid sequence is exactly the same.

We showed how gene deletion events on one or both chromosomes can be identified by applying a binomial test to genes with low usage. For the binomial test, we suggested one uniform cutoff for deletion candidates for *V* genes, and another cutoff for *D* genes. This uniform cutoff, however, may not be suitable for all genes and has to be adjusted according to additional parameters. For example, for the *D1-26* gene the cutoff threshold was a bit lower than the usage frequency needed to call it as deleted in individual S24, even though it should have been determined as a deletion by comparing its usage to other individuals (figure 6E). For single chromosome deletion detection, the cutoff is even harder to determine. Relying on deletions detected by haplotype, we observed that genes with one chromosome deletion mostly display a lower usage frequency than the same genes in individuals without a deletion. These cutoffs may be influenced by many factors, and as more data become available factors such as ethnic diversity should be taken into account in setting them. We showed that *V3-9* and *V1-8* deletion is mutually exclusive with *V5-10-1* and *V3-64D* (figure 5). This pattern can be utilized also as an anchor for haplotyping.

It is important to note that when we use the "deletion" terminology, we actually mean deletion from the repertoire. This does not necessarily imply that these genes were deleted from the germline DNA. It can be that there were mutations in the non coding region of the allele, the RSS. Such mutations can cause the specific "deleted" alleles not to appear in the repertoire. Hence, to validate inferred deletions and duplication events, sequencing of the genomic region encoding the antibody heavy chain locus is needed. Other major factors that have a strong influence on our approach are the number and type of cells sequenced, and sequencing depth. When sequencing PBMCs for example, a large fraction of the sequenced repertoire will belong to cells that were clonally expanded and have many mutations. This can influence the analysis by creating biases in gene usage estimation due to clonal expansion and allele mis-assignment due to somatic mutations. Increasing sequencing depth can help by effectively increasing the number of non-mutated cells. Our method was designed to confront the limitations of current sequencing technologies so that it can be used on typical depth datasets. Although small samples might result in undiscovered deletions, the deletions that are discovered in the small samples have a high reliability. The Bayes factor reflects the certainty level in cases of low numbers of sequences, and different thresholds can be imposed on it to obtain different levels of reliability.

The deletion patterns we discovered include several observations that will be interesting to follow up.

First, we observed several genes that are consistently deleted in groups, for example *V1-69-2* and *V2-70D* as well as *V4-30-4, V4-30-2* and *V3-30-3* (see figure 2C and figure 5B). These groups usually consist of adjacent genes, but not always, as in the case of *V3-43D* and *V4-38-2*, where they are separated by two pseudo genes. Second, there is a deletion in almost all chromosomes of either the gene pair *V3-9* and *V1-8* or *V5-10-1* and *V3-64D*, supporting the hypothesis that these gene pairs are persisting haplotype blocks.

The model parameters used here are based on simplifying assumptions such as a constant probability to miscall an allele (ϵ). This can be further developed in future studies to be based on empirical data, as outlined recently[39]. In particular, identifying single chromosome gene deletions from relative gene usage, relies on thresholds that we estimated here. These thresholds can be updated as more data accumulates, leading to more accurate estimations.

Methods

Library preparation and sequencing

One hundred subjects from Norway were enrolled in the study; 48 healthy controls (out of which 28 blood bank donors and 20 healthy subjects), and 52 patients with celiac disease. Naive B cells (defined as $CD19^+$, $CD27^-$, IgD^+, IgA^-, and IgG^-) were sorted on a FACSAria flow cytometer (BD) from all 100 individuals (Supplementary figure 10). The cells were immediately spun and cell pellets were kept at -80°C until RNA extraction (using RNeasy Midi kit, Qiagen). Participants gave written informed consent. The research is covered by the approval of the Regional Ethical Committee (projects REK 2010/2720 and REK 2011/2472, project leader Knut E. A. Lundin). RNA was reverse-transcribed using an oligo dT primer. An adaptor sequence was added to the 3' end, which contains a universal priming site and a 17-nucleotide unique molecular identifier. Products were purified, followed by PCR using primers targeting the IgD, IgM regions, and the universal adaptor. PCR products were then purified using AMPure XP beads. A second PCR was performed to add the Illumina P5 adaptor to the constant region end, and a sample-indexed P7 adaptor to the universal adaptor. Final products were purified, quantified with a TapeStation (Agilent Genomics), and pooled in equimolar proportions, followed by 2x300 paired-end sequencing with a 20% PhiX spike on the Illumina MiSeq platform according to the manufacturer's recommendations.

Verification of selected deletions from genomic DNA

Amplification of selected genes from gDNA was performed to verify a subset of the predicted deletions. Six patients (S16, S4, S85, S30, S42, and S49) were selected for testing, based on the predicted deletions and availability of gDNA samples (Supplementary table 2A). Due to high homology between the variable heavy chain genes, gene specificity of primers is often compromised. Therefore, only genes for which specific primers could be designed were selected, namely *V1-8, V3-9, V3-64D, V4-38-2*, and *V5-10-1*. gDNA was isolated from T cells and monocytes (QiaAmp DNA Mini kit, Qiagen) that had been sorted on a FACSAria flow cytometer (BD) from cryopreserved PBMCs (CD19-, CD3+/CD14+). Gene specific primers were designed using PrimerBLAST[40] (default settings) and reference genomic sequences from IMGT were used as search templates. The primers were ordered from biomers.net GmbH and Eurogentec (Supplementary table 2B). Target genes were amplified with Q5 Hot Start High-Fidelity DNA Polymerase (NEB) using touch-down PCR with annealing temperature decreasing from 70°C to 60°C, and 25 cycles in total. PCR products were analyzed by gel electrophoresis and the bands were visualized by staining with Midori Green Advance (Nippon Genetics, Techtum). Then, to validate that the correct sequences were amplified, PCR products were cleaned using the Monarch® PCR & DNA Cleanup Kit ($5\mu g$) (NEB), and Sanger sequencing of the products was performed by Eurofins Scientific/GATC. The identity of each sequence was verified by IMGT V-Quest[33].

Data preprocessing and genotyping

pRESTO[41] version $0.5.4.0$ was applied to produce a high-fidelity repertoire, as previously described[42]. Sequences were then aligned to the *V,D,* and *J* genes using IgBLAST[29]. The reference germline was downloaded from IMGT website in December 2017.

Novel alleles were detected by applying TIgGER[23,28] to the set of functional sequences. The *V/D/J* gene of a sequence with higher similarity to a novel allele than to the reference gene was reassigned to the novel allele. For each sample a genotype was constructed from sequences with a single assignment (only one best match), using TIgGER adapted for Bayesian approach[28]. Overall, 25 novel *V* alleles were identified and set as part of individuals' genotypes. Next, sequences were realigned according to the inferred personal genotype by IgBLAST, novel alleles that were part of the genotype, were included in the personal reference. Sequences with more than three mutations in the *V* locus and with at least one mutation in the *D* locus were filtered out leaving on average 86% of the sequences for each sample (range $58 - 91\%$). For additional analysis, genotypes were similarly inferred using IMGT/HighV-QUEST[43] version 1.5.7.1 or partis[25] version 0.13.0 (see Supplementary figure 1). Five samples with low sequencing depth after filtration (< 2000 reads) were discarded from the analysis. Sample names were given after ordering the samples in figure 1, using a hierarchical clustering analysis in R with the heatmap function.

Sequence data has been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001003337. Further information about EGA can

be found on https://ega-archive.org "The European Genome-phenome Archive of human data consented for biomedical research".

Binomial test for identifying gene deletions

The *V, D,* and *J* gene usage varies across genes and individuals. However, in some of the samples, the relative usage of some genes is much lower than in most of the population. To assess if the frequency is low enough to proclaim a certain gene as deleted in an individual, a binomial test was applied. In a given sample, *V* genes with relative frequency below 0.001 were set as candidates for deletion. This threshold was chosen based on the size of the smallest sequence set in the study (~ 2000). For such a sample, genes that included one or two reads were candidates for deletion by the binomial test. The binomial test has three parameters: number of trials (N), number of successes (x), and probability of success (p). Here, for a given individual, x was set to the number of sequences mapped to the *V* gene, N to the total number of sequences, and p to the lowest relative frequency of this gene among all non-deletion-candidate samples with relative frequencies larger than 0.001. For a given gene, candidate samples with an adjusted q value (Benjamini-Hochberg) below 0.01 were marked as deleted. *D* deletion detection was conducted in a similar way, but with a different candidate frequency threshold of 0.005.The larger threshold here was chosen to reflect the reduced reliability of *D* gene assignments. To evaluate the method's robustness for deletion inference in extremely lowly expressed genes, two gamma distributions were estimated for samples with gene usage below and above 0.001 threshold, respectively. In addition, empirical cumulative distribution function curves of gene usage were generated for all samples for a specific gene (Supplementary figure 11). Deletions in genes for which bimodal behavior could not be observed were considered less reliable.

Haplotype inference

The process is illustrated in figure 1. A Bayesian framework based on a binomial likelihood with a conjugate beta prior was adapted to haplotype inference. Using this framework, two biological models were compared. In one model, the considered allele is present on one of the chromosomes only, while in the other model it is present on both chromosomes. For the rest of this paragraph, we assume that we would like to infer the chromosome(s) on which a *V* allele resides, where the chromosomes are identified by the *J* allele they contain. For simplification, we also assume that each sequence represents a unique recombination event, and hence adds one to the number of *V*-*J* allele pair events. If the considered *V* allele appears with both *J* alleles, inference is expected to tell us that it is present on both chromosomes. If it almost always appears with one of the *J* alleles, we will infer that it is present on one of the chromosomes only. The posterior probability for each *V* allele usage is given by

$$P(\overrightarrow{\theta}|\overrightarrow{X})_{\beta} = \frac{P(\overrightarrow{X}|\overrightarrow{\theta})_{binomial} \cdot P(\overrightarrow{\theta})_{\beta}}{P(\overrightarrow{X})},$$

where $\overrightarrow{\theta}$ is the *V* allele probability distribution, and $\overrightarrow{X}$ is a two dimensional vector with components given by this *V* allele's number of occurrences in association with each of the two *J* alleles. Priors were fitted empirically for each individual based on their overall *V* allele usage. The two models are represented by two values of $\overrightarrow{\theta}$. For the one chromosome model, we expect all sequences with a given *V* allele to appear together with a specific *J* allele. Hence $\overrightarrow{\theta_1} = \frac{(1+\epsilon,\epsilon)}{1+2\epsilon}$, where ϵ accounts for the probability of allele mis-assignment. In the two chromosomes scenario, we expect the *V* allele to appear with both *J* alleles in similar proportions to the *J* allele usage, and hence $\overrightarrow{\theta_2} = \frac{(p+\epsilon,1-p+\epsilon)}{1+2\epsilon}$, where p is the fraction of the dominant *J* allele. The level of confidence in the most probable model is calculated using a Bayes factor, $K = \frac{P(H_{1st}|\theta)}{P(H_{2nd}|\theta)}$, where H_{1st}Âŋ and H_{2nd} correspond, to the posteriors of the most and second-most likely models, respectively. The larger the K, the greater the certainty in the model. If the evidence is not strong enough, haplotype inference is set to "unknown". Gene deletion events were called on a specific chromosome when for an "unknown" allele the Bayes factor was larger than 1000. For convenience we define $lK = \log_{10}(K)$. All relevant code used in this manuscript is available upon request.

Determining the heterozygosity cutoff for anchor D genes

To estimate the distance between two haplotypes inferred by different genes, a Jaccard distance was calculated in the following way: (i) For each gene, one minus the ratio between the number of shared alleles

over the number of unique alleles from both samples was calculated. For example, for two haplotyped allele assignments a and b the Jaccard distance was defined as $1 - \frac{a \cap b}{a \cup b}$. Genes that appeared in only one of the samples were excluded. (ii) The overall distance between two haplotypes was calculated as an average of all individual gene distances. Only individuals that had a minimum of 5 genes tested in the Jaccard comparison and each of the genes had a minimum of five linkages with the question allele were included. Wilcoxon test was used to assert the cutoff that differentiates between the two groups' means. A cutoff of 0.3 was set with a p-value $< 2e - 04$.

Determining thresholds for single chromosome deletions

Gene usage for haplotyped individuals was used to determine a threshold for a single chromosome deletion. After determining single chromosome deletions in haplotyped individuals for a specific gene, we divided the data for each gene into three groups: $0, 1,$ or 2 chromosome gene deletions. This division relies upon a minimum certainty level (lK), which was set to 10 to be conservative. We then estimated the mean and standard deviation for the groups with zero or one deletion. Since the standard deviation is unknown in advance, the probability distribution from which we can extract confidence intervals is the t distribution with $n - 1$ degrees of freedom; i.e. $x \sim \mu + \sigma \cdot t(\nu = n - 1)$ where μ and σ are the estimated mean and standard deviation respectively. Using these two distributions we calculated TP, TN, FP, and FN for various thresholds (figure 6A). From these, we extracted the thresholds for which the FPR is 0.05 and 0.01 (figure 6B-C). Confidence intervals for the ratio between gene usage of different groups (figure 6D) was calculated with the ttestratio function from the R package mratios[44].

Comparing *J6*- and *D2-8/D2-21*-based *V* haplotypes

V haplotypes of samples heterozygous for both J6 and D2-8 and for both J6 and D2-21 were used to assess the reliability of using D genes as anchors for haplotype inference. To compare the similarity between *V* inferred haplotypes, a Jaccard distance was calculated once for samples heterozygous for J6 and D2-8 and once for samples heterozygous for J6 and D2-21. The distance between two V haplotypes was calculated in the following way: (i) For each gene, one minus the ratio between the number of shared alleles over the number of unique alleles from both samples was calculated. For example, for two haplotyped allele assignments a and b the Jaccard distance was defined as $1 - \frac{a \cap b}{a \cup b}$. Genes that appeared in only one of the samples were excluded. (ii) The overall distance between two haplotypes was calculated as an average of all individual gene distances.

Gene filtration

For the haplotype inference only functional genes, according to IMGT and NCBI, were used. IMGT ORF and pseudo-genes were removed after genotype inference. *V1-69D* was also removed since for most alleles it is not possible to distinguish it from *V1-69*. *V4-30-1* was removed as well, as IMGT does not have the annotation sequence reference. Two *D* gene pairs have identical sequences: *D4-4/D4-11* and *D5-5/D5-18*. Therefore only *D4-11* and *D5-18* were used in the inference.

Single sample sign test

A special case of the binominal test was used to statistically compare the distribution of values below and above a 0.5 threshold. The p values obtained from the test were then corrected using the Benjamini-Hochberg method and referred to as q values.

Acknowledgments

This research was supported by grants from ISF (grant number 832/16) to G.Y., P.P., A.P, and M.G., and grants from the Research Council of Norway through its Centre of Excellence funding scheme (project number 179573/V40), the South-Eastern Norway Regional Health Authority (project 2016113) and Stiftelsen KG Jebsen (SKGMED-017) to L.M.S..

References

[1] Murphy, K. *Janeway's Immunobiology* (Garland Science, 2011), 8 edn.

[2] Hodgkin, P. D., Heath, W. R. & Baxter, A. G. The clonal selection theory: 50 years since the revolution. *Nature Immunology* **8**, 1019–1026 (2007). URL `http://www.nature.com/ni/journal/v8/n10/abs/ni1007-1019.html`.

[3] Boyd, S. D. *et al.* Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *Journal of immunology (Baltimore, Md.: 1950)* **184**, 6986–6992 (2010).

[4] Stern, J. N. *et al.* B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Science Translational Medicine* **6**, 248ra107–248ra107 (2014).

[5] Palanichamy, A. *et al.* Immunoglobulin class-switched B cells form an active immune axis between CNS and periphery in multiple sclerosis. *Sci. Transl. Med.* **6**, 248ra106–248ra106 (2014). URL `http://stm.sciencemag.org/cgi/doi/10.1126/scitranslmed.3008930`.

[6] Snir, O. *et al.* Analysis of celiac disease autoreactive gut plasma cells and their corresponding memory compartment in peripheral blood using high-throughput sequencing. *J Immunol* 1402611 (2015).

[7] Laserson, U. *et al.* High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4928–4933 (2014). URL `http://www.pnas.org/content/111/13/4928`.

[8] Tsioris, K. *et al.* Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. *Integrative Biology* (2015).

[9] Sok, D. *et al.* The Effects of Somatic Hypermutation on Neutralization and Binding in the PGT121 Family of Broadly Neutralizing HIV Antibodies. *PLoS Pathog* **9**, e1003754 (2013). URL `http://dx.doi.org/10.1371/journal.ppat.1003754`.

[10] Wu, Y.-C. B. *et al.* Influence of seasonal exposure to grass pollen on local and peripheral blood IgE repertoires in patients with allergic rhinitis. *Journal of Allergy and Clinical Immunology* **134**, 604–612 (2014).

[11] Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer* **12**, 298–306 (2012).

[12] Yahalom, G. *et al.* An antibody-based blood test utilizing a panel of biomarkers as a new method for improved breast cancer diagnosis. *Biomarkers in cancer* **5**, 71 (2013).

[13] Wu, Y.-C. B., Kipling, D. & Dunn-Walters, D. K. Age-related changes in human peripheral blood IGH repertoire following vaccination. *Frontiers in immunology* **3**, 193 (2012).

[14] Benichou, J., Ben-Hamo, R., Louzoun, Y. & Efroni, S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* **135**, 183–191 (2012).

[15] Yaari, G. & Kleinstein, S. H. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome medicine* **7**, 121 (2015).

[16] Wardemann, H. & Busse, C. E. Novel approaches to analyze immunoglobulin repertoires. *Trends in immunology* **38**, 471–482 (2017).

[17] Breden, F. *et al.* Reproducibility and reuse of Adaptive Immune Receptor Repertoire data. *Frontiers in immunology* **8** (2017).

[18] Watson, C. T. *et al.* Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *The American Journal of Human Genetics* **92**, 530–546 (2013).

[19] Matsuda, F. *et al.* The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *Journal of Experimental Medicine* **188**, 2151–2162 (1998).

[20] Corbett, S. J., Tomlinson, I. M., Sonnhammer, E. L., Buck, D. & Winter, G. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments,"minor" D segments or DD recombination1. *Journal of molecular biology* **270**, 587–597 (1997).

[21] Mattila, P. S., Schugk, J., Wu, H. & Mäkelä, O. Extensive allelic sequence variation in the J region of the human immunoglobulin heavy chain gene locus. *European journal of immunology* **25**, 2578–2582 (1995).

[22] Ravetch, J. V., Siebenlist, U., Korsmeyer, S., Waldmann, T. & Leder, P. Structure of the human immunoglobulin μ locus: characterization of embryonic and rearranged J and D genes. *Cell* **27**, 583–591 (1981).

[23] Gadala-Maria, D., Yaari, G., Uduman, M. & Kleinstein, S. H. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E862–E870 (2015).

[24] Corcoran, M. M. *et al.* Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nature communications* **7**, 13642 (2016).

[25] Ralph, D. K. & Matsen IV, F. A. Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS computational biology* **12**, e1004409 (2016).

[26] Kidd, M. J. *et al.* The inference of phased haplotypes for the immunoglobulin H chain v region gene loci by analysis of VDJ gene rearrangements. *The Journal of Immunology* **188**, 1333–1340 (2012).

[27] Kirik, U., Greiff, L., Levander, F. & Ohlin, M. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. *Molecular Immunology* **87**, 12–22 (2017).

[28] Gadala-Maria, D. *et al.* Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. *bioRxiv* (2018). URL `https://www.biorxiv.org/content/early/2018/08/31/405704`. `https://www.biorxiv.org/content/early/2018/08/31/405704.full.pdf`.

[29] Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research* gkt382 (2013). URL `http://nar.oxfordjournals.org/content/early/2013/05/11/nar.gkt382`.

[30] Lefranc, M.-P. *et al.* Imgt®, the international immunogenetics information system®. *Nucleic acids research* **37**, D1006–D1012 (2008).

[31] Watson, C. & Breden, F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes and immunity* **13**, 363 (2012).

[32] Kirik, U., Greiff, L., Levander, F. & Ohlin, M. Data on haplotype-supported immunoglobulin germline gene inference. *Data in brief* **13**, 620–640 (2017).

[33] Brochet, X., Lefranc, M.-P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for ig and tr standardized v-j and v-d-j sequence analysis. *Nucleic acids research* **36**, W503–W508 (2008).

[34] Parks, T. *et al.* Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in oceania. *Nature communications* **8**, 14946 (2017).

[35] Elhanati, Y. *et al.* Inferring processes underlying b-cell repertoire diversity. *Phil. Trans. R. Soc. B* **370**, 20140243 (2015).

[36] Schroeder Jr, H. W. The evolution and development of the antibody repertoire. *Frontiers in immunology* **6**, 33 (2015).

[37] Kidd, M. J., Jackson, K. J., Boyd, S. D. & Collins, A. M. DJ pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes. *The Journal of Immunology* **196**, 1158–1164 (2016).

[38] Oettinger, M. A., Schatz, D. G., Gorka, C. & Baltimore, D. RAG-1 and RAG-2, adjacent genes that synergistically activate V (D) J recombination. *Science* **248**, 1517–1523 (1990).

[39] Olson, B. J. & Matsen IV, F. A. The Bayesian optimist's guide to adaptive immune receptor repertoire analysis. *arXiv* **1804.10964 [q-bio.PE]** (2018).

[40] Ye, J. *et al.* Primer-blast: a tool to design target-specific primers for polymerase chain reaction. *BMC bioinformatics* **13**, 134 (2012).

[41] Vander Heiden, Jason A and Yaari, Gur and Uduman, Mohamed and Stern, Joel NH and O'Connor, Kevin C and Hafler, David A and Vigneault, Francois and Kleinstein, Steven H. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30**, 1930–1932 (2014). URL `+http://dx.doi.org/10.1093/bioinformatics/btu138. /oup/backfile/content_public/journal/bioinformatics/30/13/10.1093_bioinformatics_btu138/2/btu138.pdf`.

[42] Vander Heiden, J. A. *et al.* Dysregulation of B cell repertoire formation in myasthenia gravis patients revealed through deep sequencing. *The Journal of Immunology* **198**, 1460–1473 (2017).

[43] Li, S. *et al.* IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nature communications* **4**, 2333 (2013).

[44] Djira, G., Hasler, M., Gerhard, D. & Schaarschmidt, F. *mratios: Ratios of Coefficients in the General Linear Model* (2018). URL `https://CRAN.R-project.org/package=mratios`. R package version 1.4.0.

Mosaic deletion patterns of the human antibody heavy chain gene locus as revealed by Bayesian haplotyping

Moriah Gidoni[1], Omri Snir[2], Ayelet Peres[1], Pazit Polak[1], Ida Lindeman[2], Ivana Mikocziova[2], Vikas Kumar Sarna[2], Knut E. A. Lundin[2], Christopher Clouser[3], Francois Vigneault[3], Andrew M. Collins[4], Ludvig M. Sollid[2], and Gur Yaari[1]

[1]Faculty of Engineering, Bar Ilan University, Ramat Gan 5290002, Israel

[2]KG Jebsen Centre for Coeliac Disease Research and Department of Immunology, University of Oslo and Oslo University Hospital, 0372 Oslo, Norway

[3]AbVitro, Inc., Boston, MA, USA

[4]School of Biotechnology and Biomolecular Sciences, University of NSW, Kensington, Sydney, NSW 2052 Australia

Supplementary material

Sample	Number of naive B cells	RIN	Paired sequences	Unique sequences	Single assignment VDJ
S1	3.5e5	10	936044	34756	28977
S2	4.5e5	9.8	896721	31108	26849
S3	4.2e5	9.4	736291	38716	32937
S4	3.5e5	8.8	872281	26747	21282
S5	3.5e5	8.5	662652	37984	32223
S6	4e5	9.5	1016502	53829	43847
S7	3e5	9.4	717247	42842	36776
S8	3.5e5	9.3	821769	5469	4464
S9	7e5	8.5	747468	34650	29217
S10	2.5e5	7.6	953488	6656	5595
S11	3.5e5	9.1	847040	26612	22682
S12	3.5e5	9.9	788419	30277	25721
S13	3e5	9.3	917742	54470	45287
S14	3.5e5	10	625901	19042	16499
S15	2.4e5	8.5	1006142	54030	45021
S16	2.5e5	9.2	876395	39773	33357
S17	3e5	7	1192473	26036	21957
S18	8.5e5	8	1206532	7798	6218
S19	4e5	9.1	781842	46688	39487
S20	5e5	8.7	537152	27361	22607
S21	3.5e5	9.7	1133385	27712	22205
S22	3.5e5	8.5	815183	11902	10170
S23	2.8e5	9.3	967969	48795	42069
S24	4e5	8.9	700587	35643	29149
S25	4.5e5	8.8	749075	34025	28290
S26	2.5e5	9.5	1115341	55694	46951
S27	2.3e5	9.6	716100	33553	28048
S28	1e5	9.2	775067	28194	23614
S29	3.5e5	8.6	841336	37040	31401
S30	3.5e5	9.6	1044881	37573	32322
S31	6e5	8.8	786272	36442	30946
S32	2.5e5	9.7	653913	22147	18532

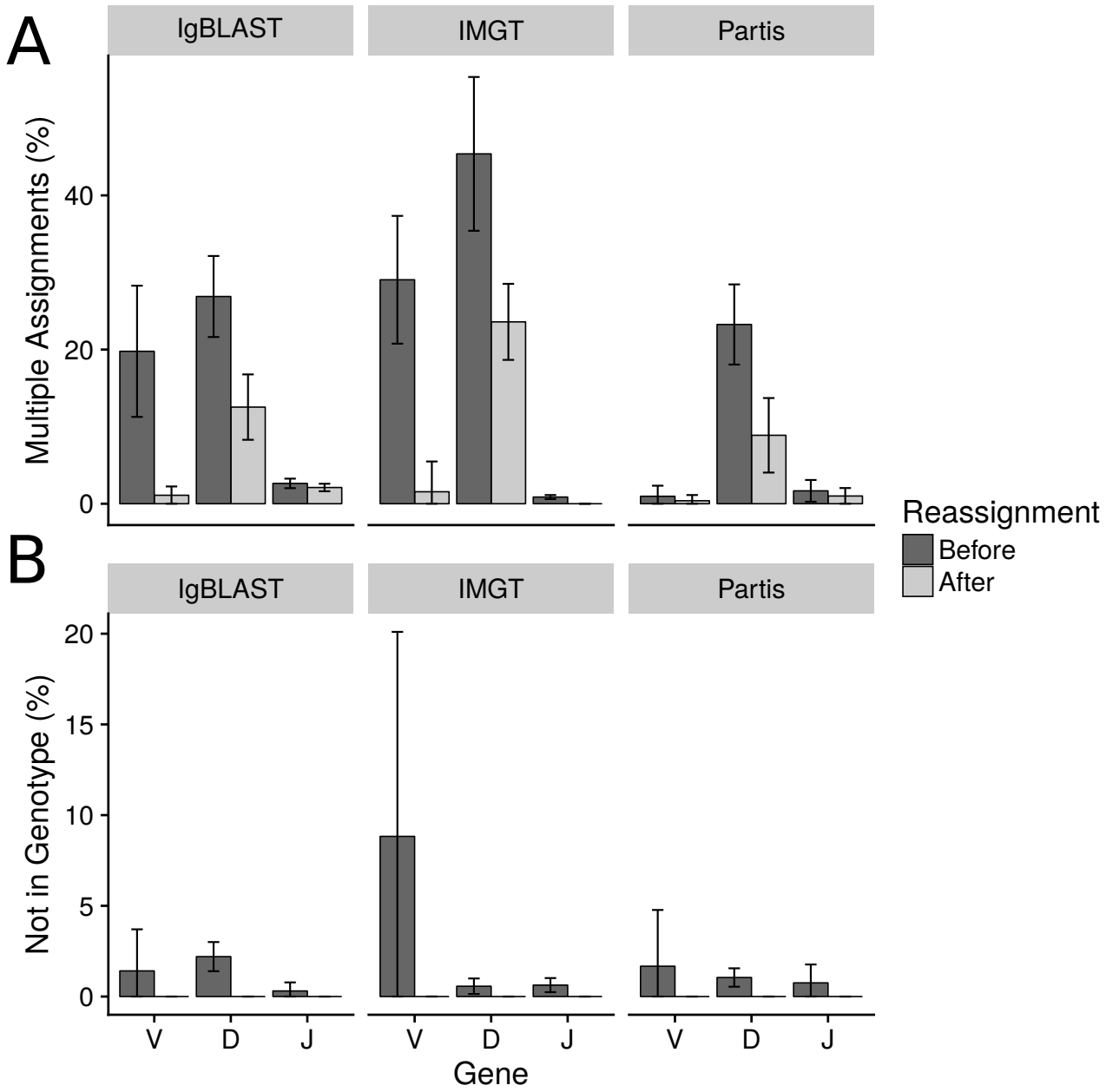
Supplementary Table 1: **Sequencing data summary.** The first column represents the sample name. The second column shows the number of sorted naïve B cells. The third column shows the sample RIN scores. The fourth column shows the number of assembled paired sequences for each sample. The fifth column shows the number of sequences after the initial *VDJ* alignment using IgBLAST, which were functional, had a read count above 2, a *V* gene mutation count below 3, and no mutations in the *D* gene. The sixth and last column shows the number of sequences with a single assignment for each of the *VDJ* genes for each sample.

Sample	Number of naive B cells	RIN	Paired sequences	Unique sequences	Single assignment VDJ
S33	3.5e5	8.6	820166	27688	23960
S34	2.5e5	8.1	699674	18231	15428
S35	2.5e5	7.8	886829	6270	5439
S36	3.5e5	9.2	777346	36919	31690
S37	4e5	9	796176	39383	33301
S38	2.5e5	7.2	646166	4074	3471
S39	5e5	9.8	678202	19349	15486
S40	2.5e5	7.8	948734	12746	10942
S41	1.5e5	7.2	780665	2548	2109
S42	3.5e5	9.1	717585	30336	25666
S43	5e5	8.5	582384	36305	30872
S44	3e5	7.5	730609	3117	2708
S45	1.1e5	9.3	814191	17652	15363
S46	3.5e5	9.9	933332	39848	33827
S47	2.5e5	10	1014845	32405	27696
S48	3e5	6.4	1021241	5756	4870
S49	3.8e5	9.3	634033	14603	12470
S50	3e5	8.5	777167	15824	13442
S51	3e5	8.9	789479	27543	23722
S52	5e5	8.7	880247	31001	26093
S53	1.2e5	9.1	974061	22964	19558
S54	4e5	9	989353	50241	44485
S55	2.5e5	9.1	946647	39230	33253
S56	4.5e5	8.7	903587	41298	35776
S57	3.5e5	8.4	756901	13783	12083
S58	9e5	6.8	618095	35802	30762
S59	2.5e5	9.3	906100	36518	32093
S60	2.8e5	8.8	974636	45215	38689
S61	2.6e5	8.8	473494	10419	9098
S62	1.8e5	9	804185	35565	29070

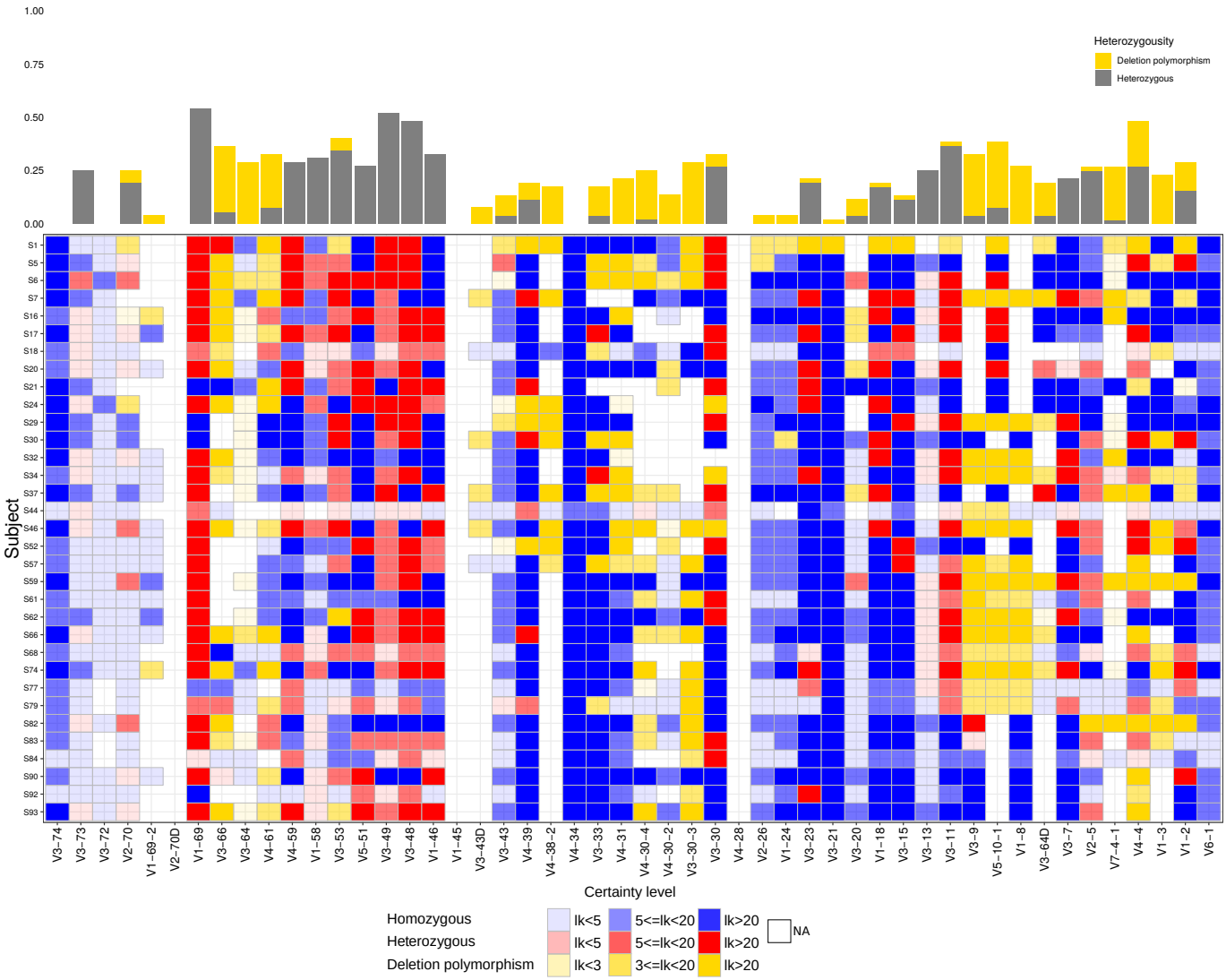
Supplementary Table 1: **Sequencing data summary (Cont.)**

Sample	Number of naive B cells	RIN	Paired sequences	Unique sequences	Single assignment VDJ
S63	2.5e5	9.4	892123	28823	23674
S64	3.5e5	8.8	1247072	50146	43362
S65	2.5e5	9.4	757705	37277	31002
S66	4e5	9.3	1089824	46583	40108
S67	2.8e5	9.7	816466	32359	27960
S68	3e5	9.1	841563	9949	8559
S69	4e5	9.1	952996	51314	44066
S70	4e5	7.9	711166	23258	20097
S71	4e5	8.6	979793	28455	24291
S72	3.1e5	8.4	741015	6594	5732
S73	4e5	8.2	679834	35316	30945
S74	4e5	9	737264	38276	32565
S75	2.8e5	9.3	977269	13592	11917
S76	3e5	8.9	829329	46567	39337
S77	3.5e5	9.8	1150471	6281	5117
S78	3e5	7.3	1002758	13714	12056
S79	3.5e5	9	917183	7880	6642
S80	1.1e5	9.8	576024	15714	13135
S81	1.2e5	9.4	795323	27966	23910
S82	4e5	9.2	677994	32794	28225
S83	3e5	9	927048	11647	8634
S84	3e5	9.2	569276	8066	6678
S85	3e5	8.8	900189	11158	9440
S86	2.5e5	8.4	745469	21820	18232
S87	3.5e5	9.6	1456927	7558	6530
S88	1.5e5	9.2	902418	36937	31043
S89	4e5	8.9	666271	34469	29569
S90	3.5e5	9.3	915147	28631	24805
S91	2.3e5	9.4	823833	26501	22922
S92	2.5e5	8.8	1386237	8624	7278
S93	2.5e5	7.7	730300	29955	25341
S94	4e5	8.7	493745	21285	18253

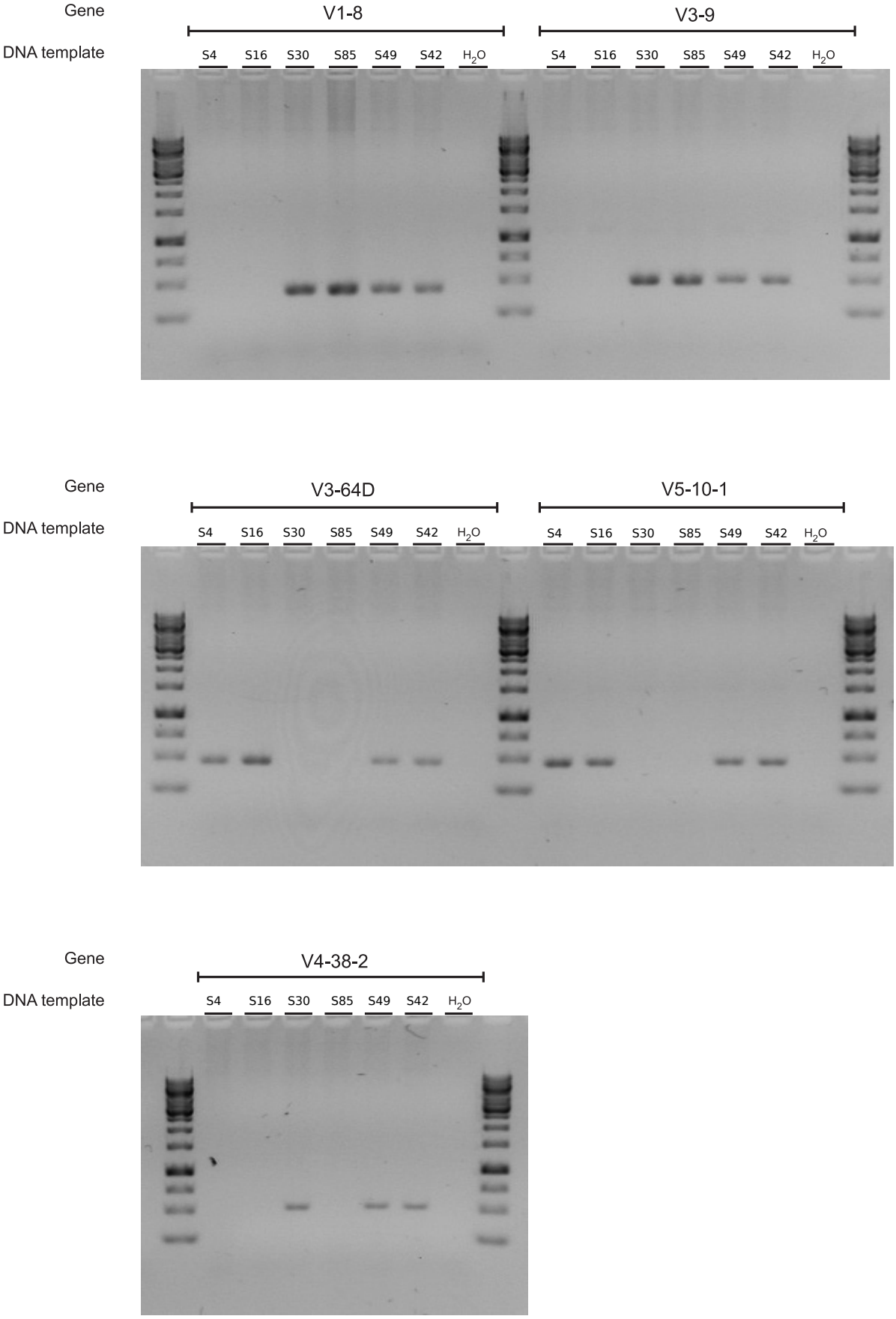
Supplementary Table 1: **Sequencing data summary (Cont.)**



Supplementary Figure 1: **Genotyping reduces multiple allele assignments** The mean percent of multiple assignments and the percentage of genes not in genotype for all individuals across antibody heavy chain genes, before and after genotyping. Alignment was performed using IgBLAST, partis or IMGT. (A) Multiple assignment is reduced by approximately half. (B) After the second assignment, genes originally not assigned within the personal genotype were eliminated. Alignment with partis was applied using version $0.13.0$[25]. In order to obtain the data for the pre-genotype process and the genotype annotation we provided the aligner with our own customized germlines. We used the following flags for partis: '- -skip-unproductive','- -dont-find-new-alleles', and '- -initial-germline-dir'. We ran the received data through an in-house parser to get the needed layout for TIgGER use, there we defined a multiple assignment when the difference in score between the first assignment and the others was less than 0.2. Alignment with IMGT/HighV-QUEST[43] was applied using version $1.5.7.1$ with a parameter modification to allow more assignment in *D* genes. Since IMGT does not accept customized germline, we had to rename the genes after the first alignment. To implement the second realignment step with the individual's own customized reference, we used the reassignAlleles function with Hamming distance from TIgGER for the *J* genes, and a weighted Levenshtein distance for *D* genes.



Supplementary Figure 2: **Heterozygosity of the IGH genes including deletion polymorphisms.** Each row represents a *J6* heterozygous individual, and each column represents a *V* gene. Red shades represent heterozygous genes, blue shades homozygous genes, and purpule shades genes with deletion polymorphisms. Transparency corresponds to the certainty level of genotype or haplotype inference. White represents a gene with too low usage (fewer than 10 sequences) to enable clear genotype inference. Bars on top represent the ratio between the number of individuals with heterozygous genes and all individuals with a defined genotype for this gene. Bar color correspond to the type of heterozygosity (deletion polymorphism, and heterozygosity due to two distinct alleles).



Supplementary Figure 3: **V-gene deletions on the genomic level.** Five genes were amplified by custom-designed gene-specific primers from gDNA and the products were analyzed by gel electrophoresis. GeneRuler 1kb (ThermoFisher) was used as a ladder. As a negative control, water was used instead of a gDNA template. Individuals S42 and S49, who were not predicted to have deletions of the tested genes, were used as positive controls.

A

	10^6 Cells	ng/µl
S4	3.5	46.4
S16	3.0	38.7
S30	4.3	57.1
S85	3.5	57.5
S49	3.5	46.8
S42	3.0	50.9

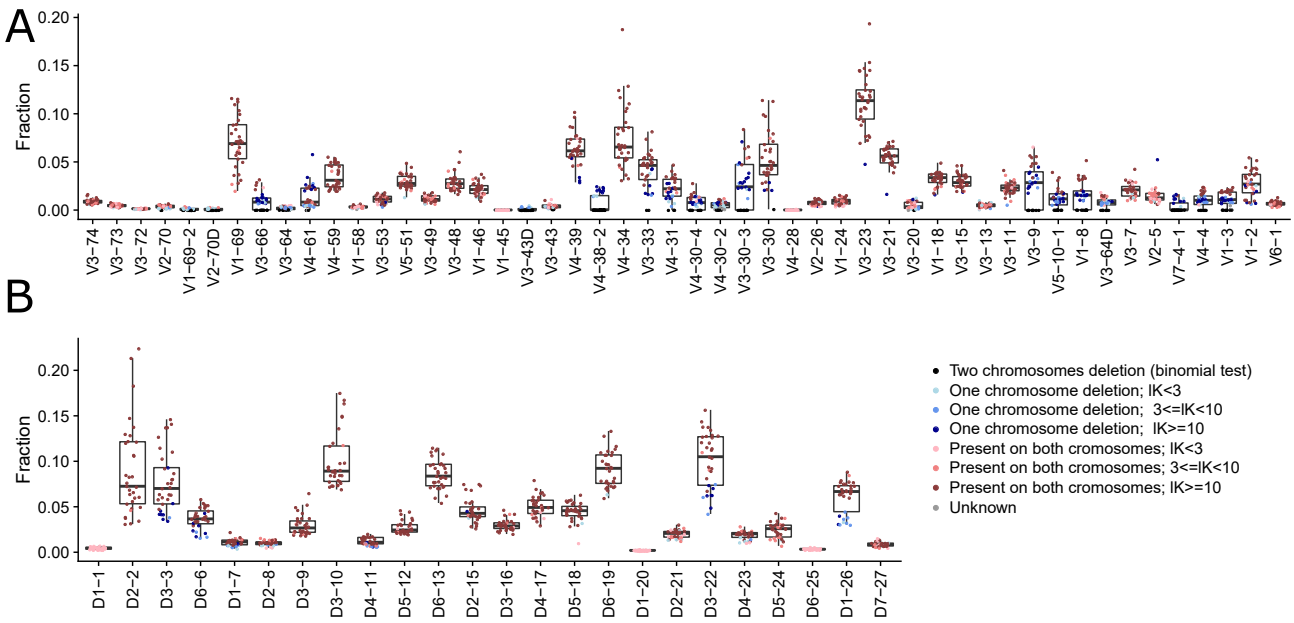
B

	Sequence (5'–3')
IGHV1–8_fwd	GTAAGGGGCTTCCTAGTCTCAAAG
IGHV1–8_rev	TCTGACTCTCTGAGGATGTGGTTT
IGHV3–9_fwd	AGGACTCACCATGGAGTTGG
IGHV3–9_rev	TTTTTGTCTGGGCTCTCGCT
IGHV3–64D_fwd	TTCATGGAGAACTAGAGATAGTGTG
IGHV3–64D_rev	GCTGTTTTTCTCCAGCGTTCC
IGHV4–38–2_fwd	AGGGATCCAGACGTGAAGATA
IGHV4–38–2_rev	GGCCTTGTATTCCGTGAGC
IGHV5–10–1_fwd	CATCCTTGGCCTCCTCCTG
IGHV5–10–1_rev	GGGTTTTAGACGGGCTCAGT

Supplementary Table 2: **Amplification of selected immunoglobulin variable heavy chain genes from gDNA.** (A) Cell number and gDNA concentration from T cells and monocytes together (non-B cells). (B) Custom-designed primers for amplification of selected immunoglobulin variable heavy chain genes from gDNA .

SUBJECT−gene−primer	IMGT annotation	Identity	Mutation
S42−HV1−8−HV1−8_fwd	Homsap IGHV1−8*01 F	100.00%.(288/288 nt)	−
S42−HV3−9−HV3−9_fwd	Homsap IGHV3−9*01 F	100.00%.(288/288 nt)	−
S42−HV3−64D−HV3−64D_fwd	Homsap IGHV3−64D*06 F	99.65%.(287/288 nt)	g258>t (T86)
S42−HV4−38−2−HV4−38−2_fwd	Homsap IGHV4−38−2*01 F	100.00%.(288/288 nt)	−
S42−HV5−10−1−HV5−10−1_fwd	Homsap IGHV5−10−1*03 F	100.00%.(288/288 nt)	−
S49−HV1−8−HV1−8_fwd	Homsap IGHV1−8*01 F	100.00%.(288/288 nt)	−
S49−HV3−9−HV3−9_fwd	Homsap IGHV3−9*01 F	100.00%.(288/288 nt)	−
S49−HV3−64D−HV3−64D_fwd	Homsap IGHV3−64D*06 F	99.65%.(287/288 nt)	g258>t (T86)
S49−HV4−38−2−HV4−38−2_fwd	Homsap IGHV4−38−2*01 F	100.00%.(288/288 nt)	−
S49−HV5−10−1−HV5−10−1_fwd	Homsap IGHV5−10−1*03 F	100.00%.(288/288 nt)	−
S85−HV1−8−HV1−8_fwd	Homsap IGHV1−8*01 F	100.00%.(288/288 nt)	−
S85−HV3−9−HV3−9_fwd	Homsap IGHV3−9*01 F	100.00%.(288/288 nt)	−
S30−HV1−8−HV1−8_fwd	Homsap IGHV1−8*01 F	100.00%.(288/288 nt)	−
S30−HV3−9−HV3−9_fwd	Homsap IGHV3−9*01 F	100.00%.(288/288 nt)	−
S30−HV4−38−2−HV4−38−2_fwd	Homsap IGHV4−38−2*02 F	100.00%.(288/288 nt)	−
S16−HV3−64D−HV3−64D_fwd	Homsap IGHV3−64D*06 F	100.00%.(288/288 nt)	−
S16−HV5−10−1−HV5−10−1_fwd	Homsap IGHV5−10−1*01 F	100.00%.(288/288 nt)	−
S4−HV3−64D−HV3−64D_fwd	Homsap IGHV3−64D*06 F	100.00%.(288/288 nt)	−
S4−HV5−10−1−HV5−10−1_fwd	Homsap IGHV5−10−1*03 F	100.00%.(288/288 nt)	−

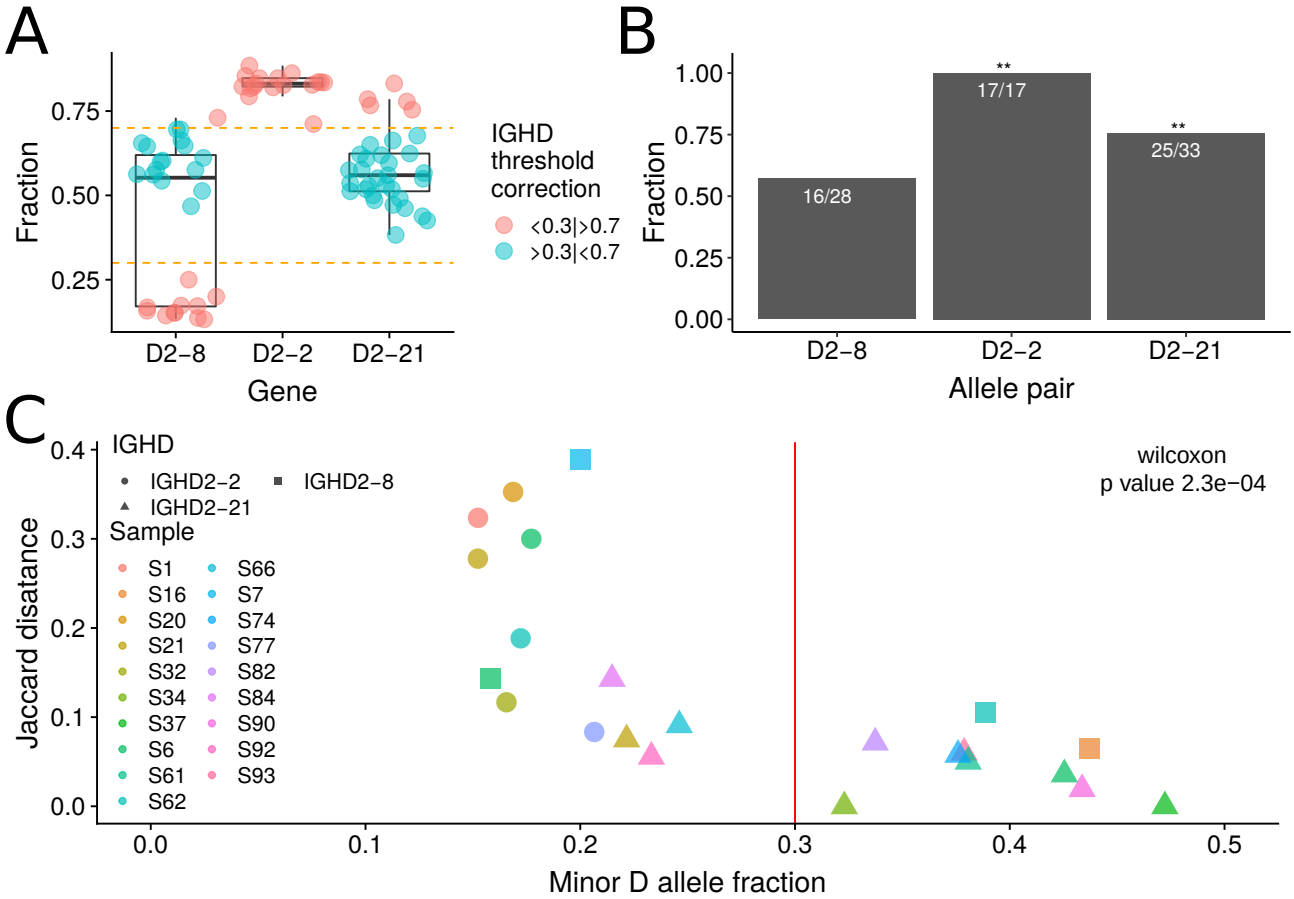
Supplementary Table 3: **Sanger sequencing of PCR products.** PCR products, obtained by amplification of genomic DNA of selected individuals using gene-specific primers, were sequenced to confirm the specificity of the primers and verify the identity of the targeted genes. Novel allele *IGHV3-64D*06_G258T* was also found by TIgGER as part of 25 novel alleles described in methods.



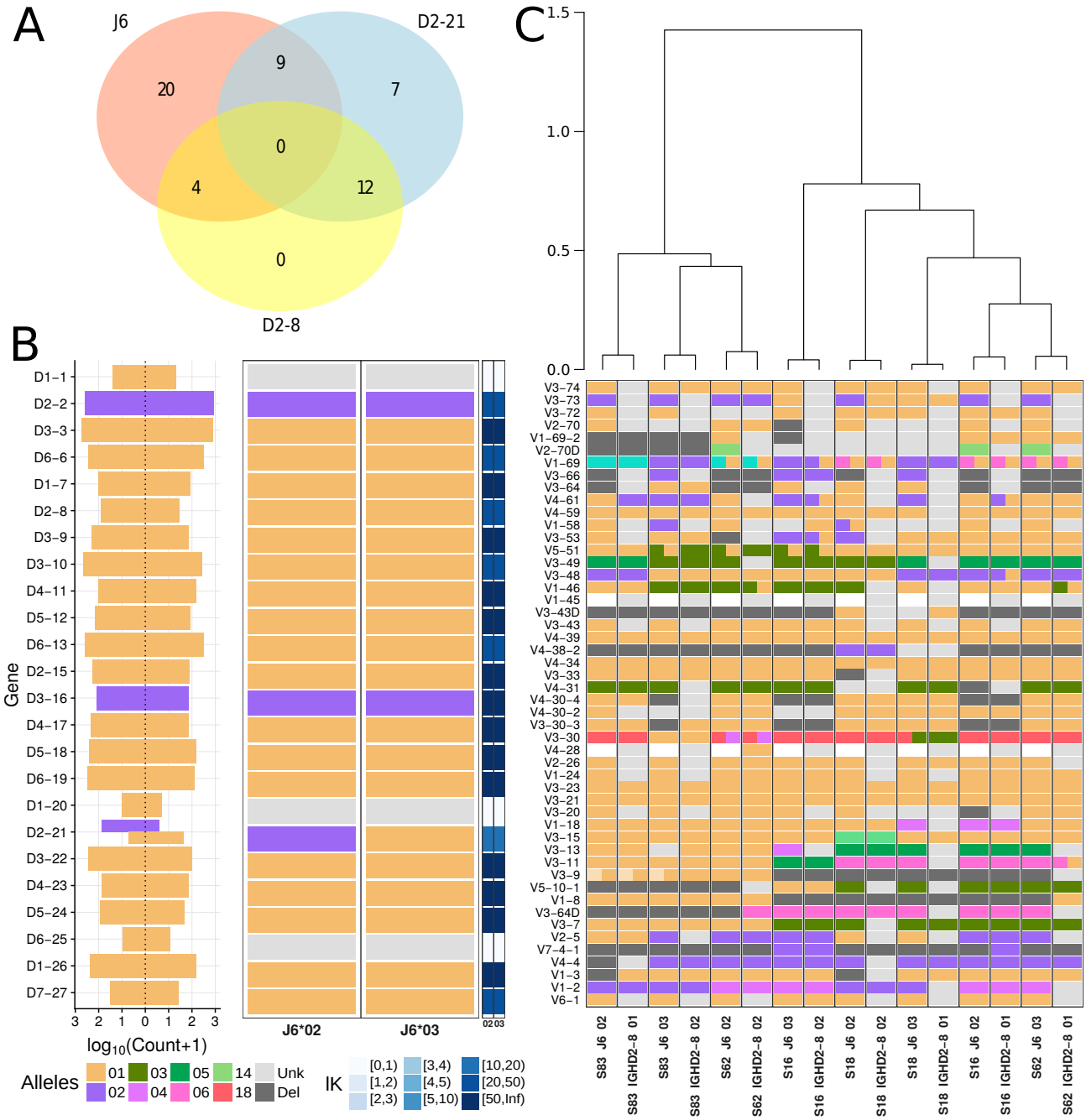
Supplementary Figure 4: **Relative gene usage of *J6* heterozygous individuals.** Relative usage of *V* (A) and *D* (B) genes from *J6* heterozygous individuals. Each dot represents an individual. Color corresponds to gene deletion from both chromosomes (black), single chromosome (blue), or no deletion (red). Shades correspond to the certainty level of deletion inference.

Gene	mu1	sd1	n1	mu2	sd2	n2	Sensitivity 0.01	Threshold 0.01	Sensitivity 0.05	Threshold 0.05
V3−66	0.0109	0.0026	14	0.0270	0.0043	4	0.1007	0.0072	0.9786	0.0165
V3−53	0.0071	0.0018	2	0.0123	0.0034	18	0.1469	0.0033	0.3633	0.0060
V4−39	0.0369	0.0116	4	0.0683	0.0134	25	0.4355	0.0348	0.7393	0.0453
V3−33	0.0271	0.0126	6	0.0492	0.0125	22	0.2443	0.0174	0.5175	0.0276
V4−31	0.0196	0.0063	8	0.0326	0.0070	13	0.1921	0.0135	0.5264	0.0198
V4−30−4	0.0093	0.0026	7	0.0171	0.0058	6	0.0062	0.0000	0.0989	0.0054
V4−30−2	0.0075	0.0014	4	0.0089	0.0019	5	0.0127	0.0018	0.0761	0.0048
V3−30−3	0.0339	0.0148	11	0.0633	0.0138	5	0.0824	0.0114	0.5009	0.0336
V3−30	0.0235	0.0046	3	0.0574	0.0256	27	0.0184	0.0000	0.0852	0.0135
V2−26	0.0055	0.0017	2	0.0087	0.0019	21	0.2511	0.0036	0.4709	0.0051
V3−20	0.0103	0.0015	2	0.0082	0.0026	4	0.0469	0.0000	0.0586	0.0018
V3−9	0.0305	0.0083	11	0.0482	0.0083	6	0.1253	0.0201	0.5486	0.0315
V5−10−1	0.0120	0.0036	10	0.0238	0.0089	9	0.0042	0.0000	0.1088	0.0072
V1−8	0.0156	0.0029	11	0.0344	0.0092	7	0.0035	0.0054	0.6331	0.0165
V3−64D	0.0084	0.0016	3	0.0138	0.0023	4	0.0462	0.0033	0.5042	0.0084
V4−4	0.0104	0.0013	10	0.0201	0.0030	7	0.6163	0.0105	0.9920	0.0141
V1−3	0.0102	0.0017	8	0.0201	0.0019	11	0.9831	0.0147	0.9960	0.0165
V1−2	0.0157	0.0079	5	0.0321	0.0144	20	0.0592	0.0000	0.1718	0.0072
D3−3	0.0486	0.0189	8	0.0860	0.0299	23	0.0432	0.0108	0.2420	0.0345
D6−6	0.0278	0.0092	6	0.0421	0.0083	23	0.2535	0.0210	0.4996	0.0276
D4−11	0.0086	0.0034	2	0.0137	0.0031	21	0.2901	0.0057	0.4828	0.0081
D3−22	0.0621	0.0088	5	0.1165	0.0219	22	0.4730	0.0613	0.9363	0.0787
D1−26	0.0355	0.0065	2	0.0706	0.0067	23	0.8908	0.0538	0.9139	0.0589

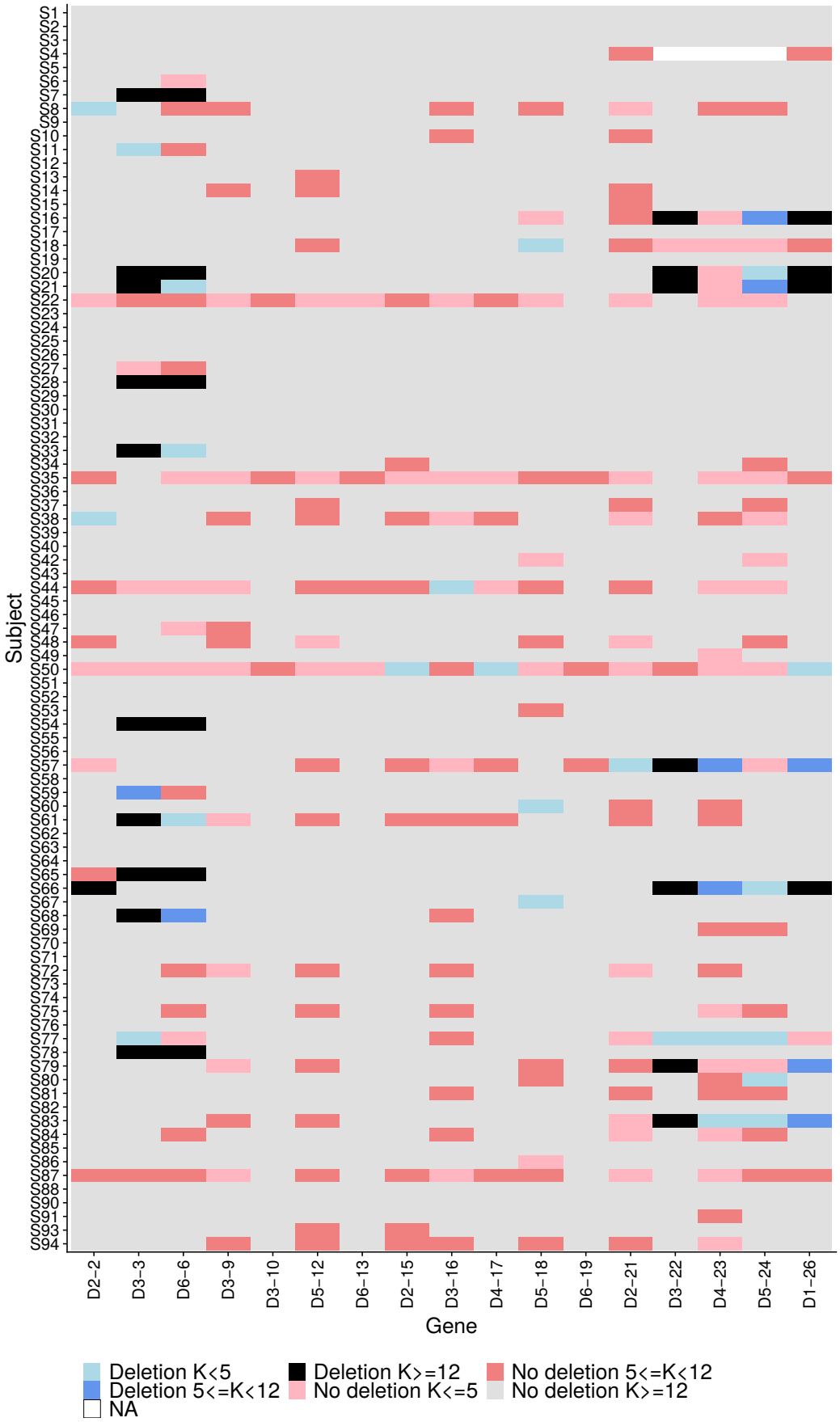
Supplementary Table 4: **Threshold statistics for figure 6** The annotation $mu1, sd1, n1$ represent the single chromosome deletion group, and $mu2, sd2, n2$ for no deletion.



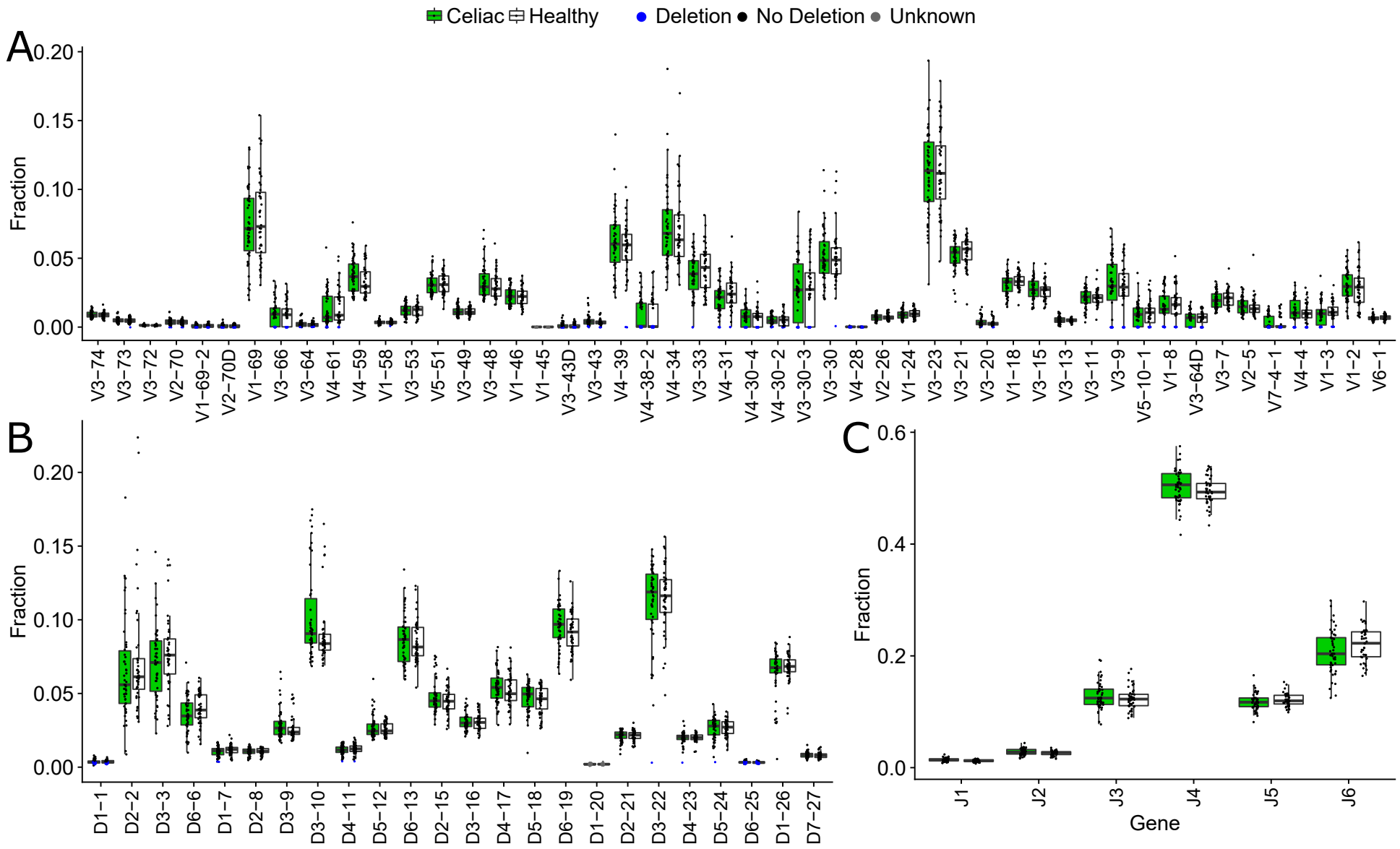
Supplementary Figure 5: **A *D* gene cutoff to be considered as heterozygous** Determining the heterozygosity cutoff for anchor *D* genes. (A) Box plot of anchor *D* gene allele pairs relative usage in heterozygous individuals. We considered allele pairs which were observed in more than 5 individuals, and each individual had a minimum of 5 *V-D-J* linkages for each of the alleles. Each point represents an individual. The allele fraction (Y axis) corresponds to the allele that is written first for each *D* gene, and is the dominant allele in most individuals. (B) The fraction of individuals with relative allele usage (of the first written allele, as in (A)) that is larger than 0.5. Asterisks indicate allele pairs with a statistically significant difference in the number of individuals with the same dominant allele. Statistical significance was determined with a binomial single sample sign test (see methods, $*$ indicates p value < 0.05, $**$ indicates p value < 0.01). (C) To estimate the distance between two haplotypes inferred by different genes, a Jaccard distance was calculated (see methods). The maximum Jaccard distance between the two anchor *D* gene alleles was plotted against the minor allele fraction, where each point represents an individual that had a minimum of 5 genes tested in the Jaccard comparison and each of the genes had a minimum of five linkages with the question allele. This resulted in a reduction in *D* heterozygous samples to 31% for *D2-21* and 17% for *D2-8* (Supplementary figure 6A). The three shapes of the dots: circle, triangle, and square correspond to the three anchor *D* genes *D2-2, D2-21,* and *D2-8*, respectively. Wilcoxon test was used to assert the cutoff that differentiates between the two groups' means. A cutoff of 0.3 was set with a p-value $< 2e-03$.



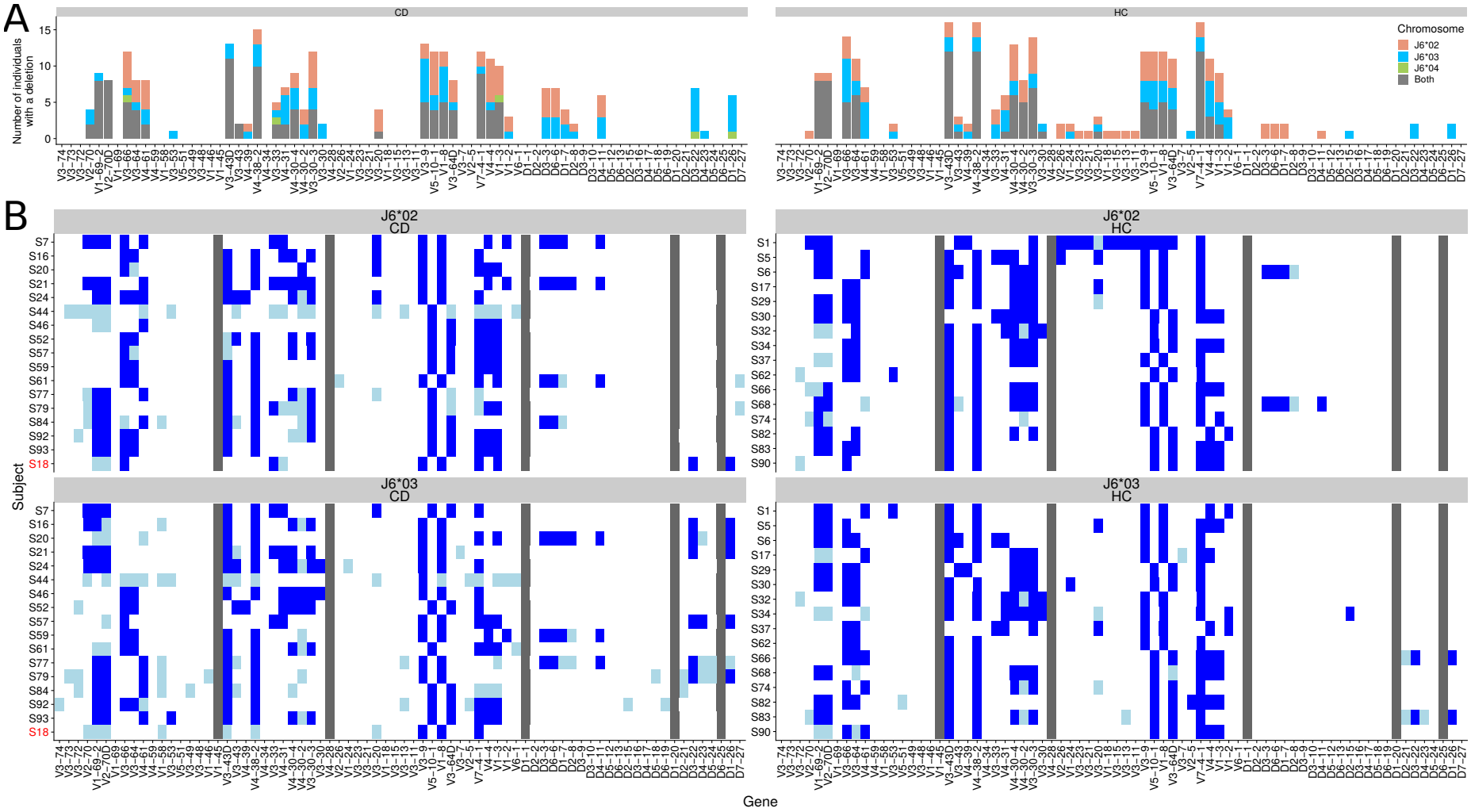
Supplementary Figure 6: ***D*-based haplotype inference.** (A) A Venn diagram of heterozygous individuals for *J6*, *D2-8* and *D2-21* genes. 54% of the individuals were identified as heterozygous in at least one of these anchor genes. (B) An example for *D* gene haplotype anchored by *J6* gene for a single individual. The left panel shows the count of each *D* gene (Y axis) that is associated with its paired anchor *J6* gene (X axis). The thickness of the count bar is inversely proportional to the number of alleles found on the chromosome. Colors correspond to the different alleles. The middle panel shows the called *D* haplotype, and the right panel shows the certainty level (lK) for each haplotype decision. (C) V haplotype map where each column corresponds to a *J6*- or *D2-8*-based haplotype of an individual that is heterozygous for both these genes. The order of columns was determined by a hierarchical dendogram based on the distances between individual haplotypes.



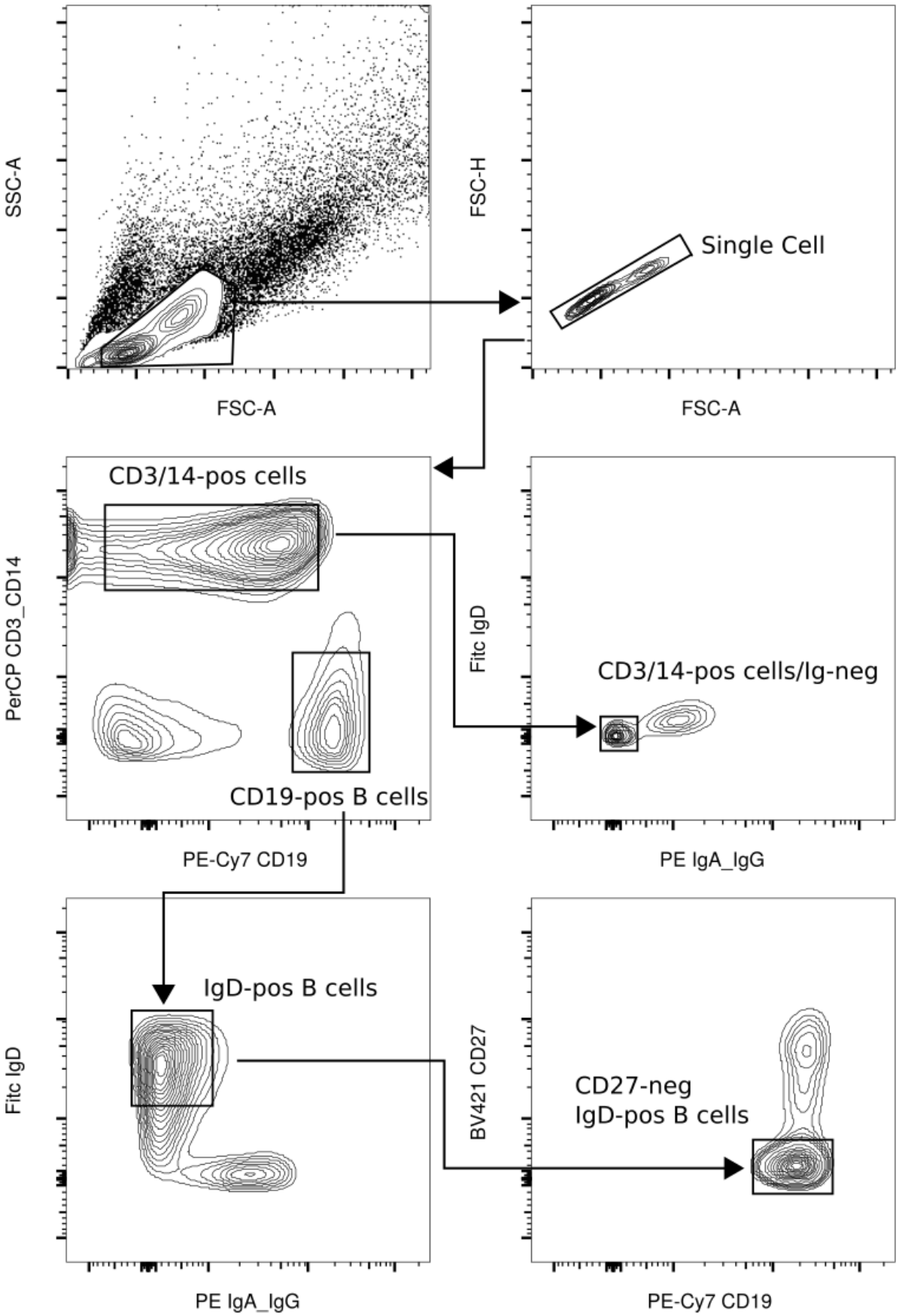
Supplementary Figure 7: **Chromosome deletion detection by pooled heterozygous *V* genes in the entire cohort.** For each individual (Y axis) and gene (X axis), each cell represents the deletion status of a gene (deletion/no deletion). The color indicates the certainty level of the decision. For the presented V_{pooled} approach only heterozygous *V* genes with minor allele fraction larger than 30% were included.



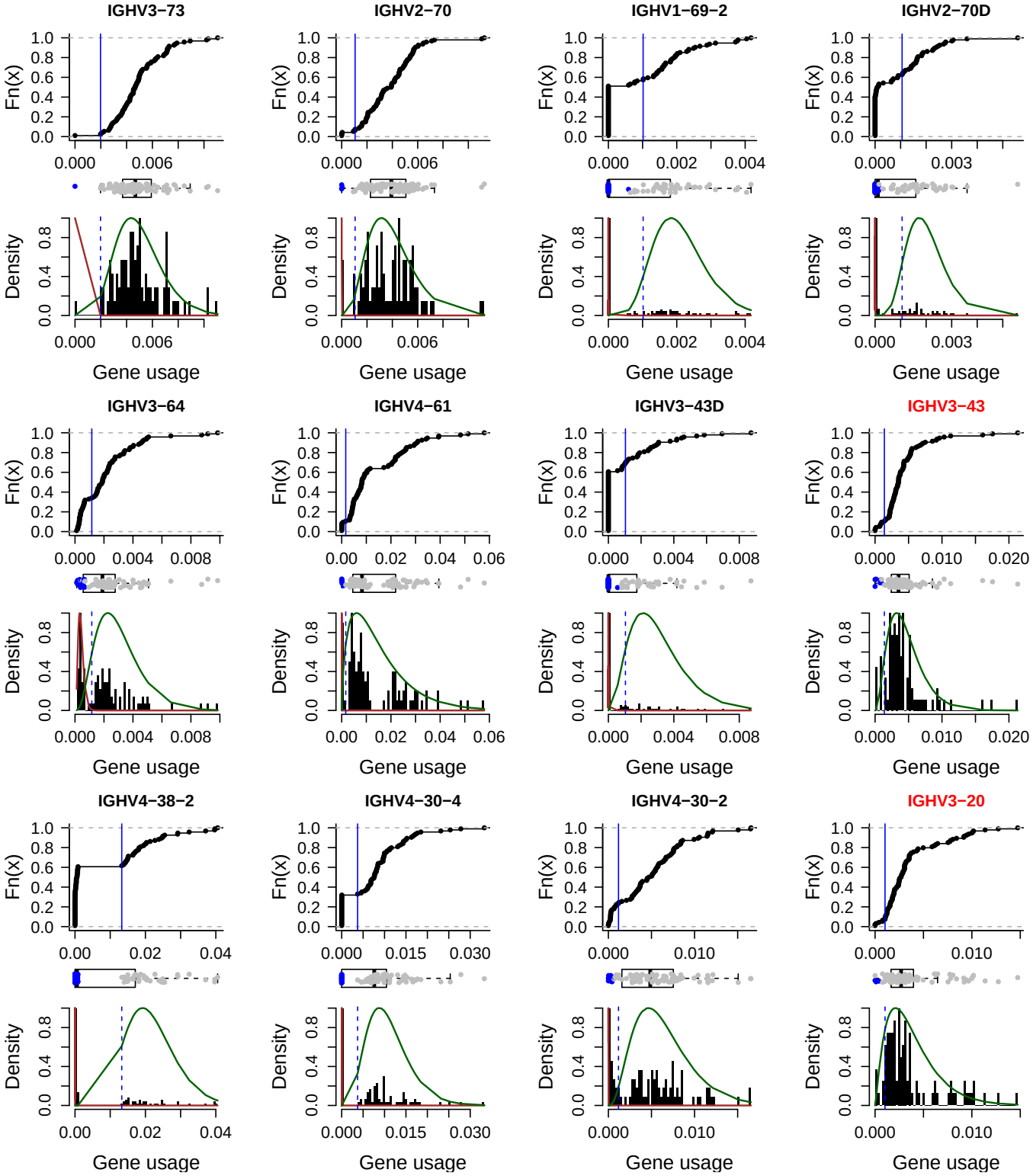
Supplementary Figure 8: **Gene deletion inference by relative gene usage with respect to clinical status.** Box plots of relative gene usage, where each dot represents a single individual, separated into celiac patients (green) and healthy individuals (white), for *V* (panel A), *D* (panel B), and *J* (panel C) genes. Blue represents deleted genes according to the binomial test (see methods).



Supplementary Figure 9: **Gene deletion inference along each chromosome according to clinical status.** (A) The distribution of *V* and *D* gene deletions along each chromosome in 32 individuals that are heterozygous for *J6*, separated into celiac disease patients (left panel) and healthy individuals (right panel), as inferred by haplotype (light red, blue, and green) and by the binomial test (gray) (B) A heatmap of *V* and *D* gene deletions and suspected deletions for each of the 32 heterozygous individuals in *J6*, separated into celiac patients (left panel) and healthy individuals (right panel). Each row represents an individual, and each column represents *V* or *D* gene. Blue represents a deletion ($lK > 3$), light blue represents a suspected deletion ($lK < 3$), and light gray represents no deletion on both chromosomes with low certainty ($lK < 3$). Dark gray represents a gene with an extremely low usage across all samples. The top panel represents the chromosome on which *J6*02* is present, and the bottom panel represents the chromosome on which *J6*03* is present. Sample S18, marked in red, is heterozygous for *J6*03* and *J6*04*. For this individual, *J6*04* was added to the *J6*02* panel.



Supplementary Figure 10: **Gating strategy for cell sorting of non B cells and naïve IgD+ B cells.** Gating strategy for sorting non-B cells (for gDNA extraction) and naïve IgD-expressing B cells (for RNA extraction). Non-B cells were sorted for being CD19-negative, CD3/CD14-positive and negative for any true/false signal from anti-IgD/IgA/IgG antibodies (mid right panel). Naïve B cells were defined as CD19-positive cells that were negative for CD3 and CD14, IgD-positive and IgA/IgG-negative that were also negative for the memory B marker CD27 (lower right panel). Arrows indicate sequential gating.



Supplementary Figure 11: **Bimodal distribution of low expressed genes.** For each gene with low usage, an empirical cumulative distribution function curve of the gene usage (upper panel for each gene), boxplot of the gene usage where the deletions are marked in blue (middle panel for each gene), and a histogram of the gene usage with the estimated gamma distributions(lower panel for each gene) are shown. Genes with labels marked in red do not follow a bimodal behavior, therefore deletion detection is considered less reliable for them.