Mosaic deletion patterns of the human antibody heavy chain gene locus

Moriah Gidoni[1], Omri Snir[2], Ayelet Peres[1], Pazit Polak[1], Ida Lindeman[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, which is 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. In addition to *IGHJ*-based inference, our method extends haplotype inference also to *IGHD*- and *IGHV*-based, in heterozygous individuals. We can thus infer complex genetic events like deletions and copy number variations in a much larger fraction of the population. We tested our method on the largest data set, to date, of naïve B-cell repertoires. We present evidence for allele usage bias, as well as a mosaic deletion pattern of *IGHD* and *IGHV* 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 preformed repertoire of antigen receptors is carried by an extremely large number of T cells and B cells, each of which is unique. 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 (Murphy, 2011). 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 (Hodgkin et al., 2007). 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 (Boyd et al., 2010), as well as reveal dysregulation in autoimmune diseases (Stern et al., 2014; Palanichamy et al., 2014; Snir et al., 2015), infectious diseases (Laserson et al., 2014; Tsioris et al., 2015; Sok et al., 2013), allergy (Wu et al., 2014), cancer (Fridman et al., 2012; Yahalom et al., 2013), and aging (Wu et al., 2012).

Dramatic improvements in high-throughput sequencing (HTS) technologies now enable large-scale characterization of adaptive immune receptor repertoires (AIRR-seq) (Benichou et al., 2012; Yaari and Kleinstein, 2015). Extracting valuable information from these sequencing data is challenging, and requires tailored computational and statistical tools which are being constantly developed (Wardemann and Busse, 2017). Much is being invested, especially by the AIRR community (Breden et al., 2017), 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 and partial sequences of these loci in the human genome have been published thus far (Watson et al., 2013; Matsuda et al., 1998; Corbett et al., 1997; Mattila et al., 1995; Ravetch et al., 1981). 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 (Boyd et al., 2010; Gadala-Maria et al., 2015; Corcoran et al., 2016; Ralph and Matsen IV, 2016).

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 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. (2013) by one complete and nine partial haplotype sequencing of the genomic region encoding the antibody heavy chain locus, using BACs and fosmids.

Because of the difficulty in performing physical sequencing of these loci, several computational tools have been developed for haplotype inference from AIRR-seq data (Kidd et al., 2012; Kirik et al., 2017). 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, statistical framework for haplotyping has been developed for *J6* (Kidd et al., 2012; Kirik et al., 2017), 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. This enables looking also at *J* distribution, and expanding the percentage of the population for which it is possible to infer haplotype. We present indications for allele usage bias, as well as interesting mosaic-like deletion patterns that are common in many individuals. Our findings greatly expand the knowledge that can be extracted from antibody repertoire sequencing data.

Methods

Library preparation and sequencing

100 individuals were enrolled in this study; 48 healthy subjects and 52 patients with celiac disease. Subjects with celiac disease were included to represent genetic variation that might be present among such patients, and not to perform association analysis. Naive B cells (defined as $CD19^{+}$, $CD27^{-}$, IgD^{+}, IgA^{-}, and IgG^{-}) were sorted on a FACSAria flow cytometer (BD) from all 100 individuals. The cells were immediately spun and cell pellets were kept at -80C until RNA extraction (using RNeasy Midi kit, Qiagen). Participants gave written informed consent. The research is covered by the approval of the Regy ional 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 manufacturers recommendations.

Data preprocessing and genotyping

pRESTO (Vander Heiden, Jason A and Yaari, Gur and Uduman, Mohamed and Stern, Joel NH and OConnor, Kevin C and Hafler, David A and Vigneault, Francois and Kleinstein, Steven H., 2014) version $0.5.4.0$ was applied to produce a high-fidelity repertoire, as previously described (Vander Heiden et al., 2017). Sequences were then aligned to the *V,D,* and *J* genes using IgBLAST (Ye et al., 2013). The reference germline was downloaded from IMGT website in December 2017.

Novel alleles were detected by applying TIgGER (Gadala-Maria et al., 2015) 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 (Gadala-Maria et al., 2018). 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. 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 (Li et al., 2013) or partis (Ralph and Matsen IV, 2016) (see figure S1). Five samples with low sequencing depth after filtration (< 2000 reads) were discarded from the analysis.

All sequencing data are available in EGA (accession numbers ERS2445766-ERS2445865).

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 different 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. 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-candidate samples with relative frequencies larger than 0.001. For a given gene, candidate samples with an adjusted p 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.

Haplotype inference

The process is illustrated in figure S2. 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 that is present on them. 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 the number of sequences that this *V* allele appeared with the two *J* alleles respectively. 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} 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 $lK = \log_{10}(K)$.

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.

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.

Results

Relative gene usage can indicate gene deletions

The data set analyzed here is first of a kind in its size and accuracy. 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 five samples with low coverage (< 2000 sequences), personal genotypes of the IGH regions were constructed using a Bayesian genotype approach (Gadala-Maria et al., 2018). 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 (table S1). In agreement with previous studies (Gadala-Maria et al., 2015), 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 (Ye et al., 2013), IMGT HighV-QUEST (Lefranc et al., 2009), and partis (Ralph and Matsen IV, 2016) (figure S1A). With the genotype step $\sim 2\%$ of the sequences that had gene assignments that were not included in the genotype were reassigned to genes that are included in it (figure S1B).

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 1A and 1B). Genes with extremely low expression across all samples were considered indeterminable (N/A). In particular, *V1-45, V4-28,* and *D6-25* have very low expression across the vast majority of individuals and thus are suspected to be non-functional. Looking at the deletions of each sample by itself several interesting patterns are observed along the locus (figure 1C and D). Specifically: A) In 44 of the 46 individuals that lack *V2-70D*, the adjacent gene *V1-69-2*, is also deleted. The two samples that lack only *V2-70D*, were borderline in terms of significance. In fact, these samples have only one assignment to this gene, but the adjusted p value is larger than the threshold due to small sample size. B) In 16 of the 17 individuals that lack *V4-30-2*, the adjacent genes: *V4-30-4* and *V3-30-3* are also deleted. Although *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*. C) Out of 56 individuals that lack *V3-43D*, 55 lack *V4-38-2*. The sample that lacks only *V4-38-2* had an adjusted p value larger than the threshold due to small sample size. D) There is a mosaic-like structure of two sets of genes whose deletion is mutually exclusive: *V3-9* and *V1-8*, and *V5-10-1* and *V3-64D*. A), B), and C) were also observed for a single haplotype shown in (Watson et al., 2013). Here we show the prevalence of these patterns among a large cohort. Further exploration of possible clinical implications caused by these deletion patterns is intriguing and remains an open challenge for future studies.

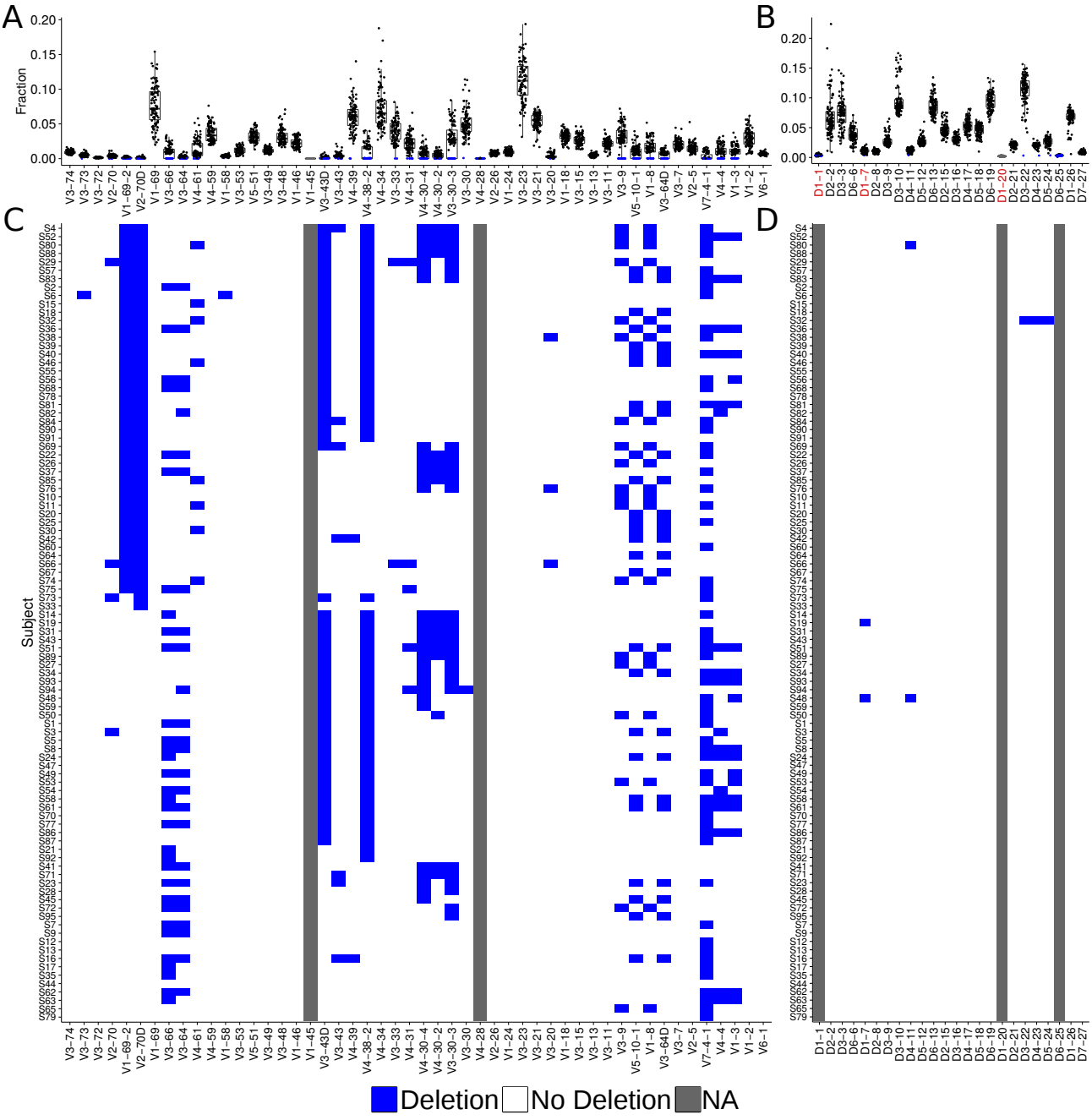

Figure 1: **Gene deletion inference by relative gene usage.** (A, B) Box plots of relative gene usage, where each dot represents a single individual. Blue represents deleted genes according to the binomial test (see methods). The order of genes here and in the following figures is based on their chromosomal location (Watson et al., 2013). *D* gene labels marked in red 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 with more than 90% deletions. Order of rows was determined by sorting the gene deletions first by *V2-70D*, then by *V3-43D*, and finally by *V4-30-4*.

Ig heavy chain gene heterozygosity landscape

Inference of personal genotype allows us to estimate the heterozygosity of these 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 individuals 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 2). It has been previously shown that approximately one third of the population is heterozygous for *J6* (Kidd et al., 2012; Kirik et al., 2017). Our cohort agrees with this observation with $31/95$ 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 32 heterozygous samples. In addition, we identified a large number of heterozygous *V* genes. Six out of the *V* genes (*V1-69, V3-53, V3-48, V3-49, V4-28*, and *V3-11*) were heterozygous in more than 50% of the individuals with a defined genotype, and 19 in more than 20%. Four *D* genes, *D2-2, D2-8, D2-21*, and *D3-16* were determined as heterozygous in $2\% - 36\%$ of the population (20, 30, 34, and 2 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 2A). This suggests a genomic hot region for germline recombination. Within this region, the three genes, *V3-66, V3-64*, and *V4-61* appear as mostly homozygous. This is not the case, however, because there are many single chromosome deletions in these genes as shown in the following sections.

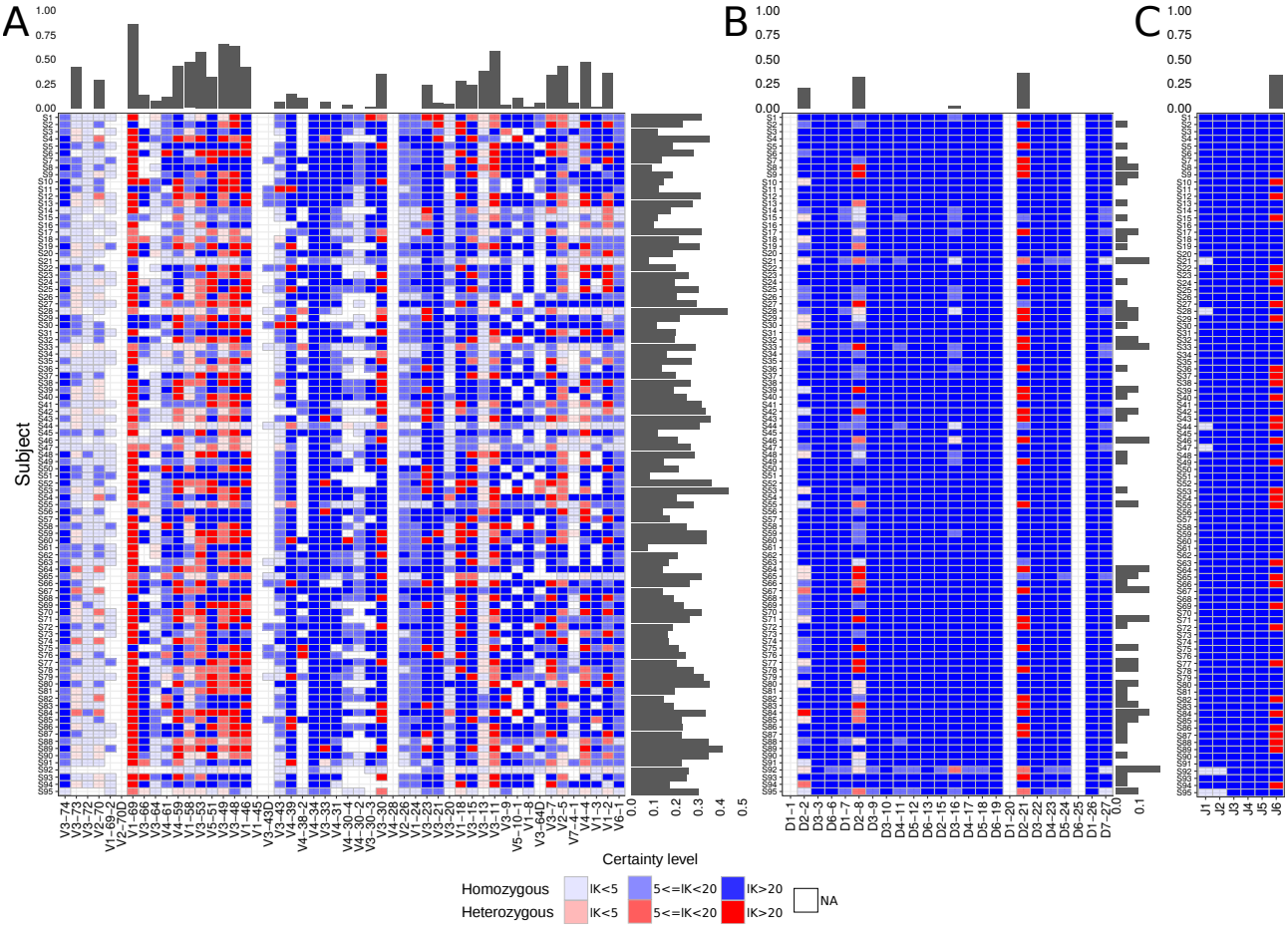

Figure 2: **Heterozygosity of the IGH genes.** Each row represents an individual, and each column represents *V* gene (A), *D* gene (B), and *J* gene (C). Red shades represent heterozygous genes, and blue shades homozygous genes. Transparency corresponds to the certainty level of genotype inference. White represents a gene with too low usage (fewer than 10 sequences) to enable clear genotype inference. Bars on top of each panel represent the fraction of heterozygous genes out of 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 3A). To statistically address whether there is a biased usage between pairs of alleles that are present in the same individual, a 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. Out of 42 allele pairs (23 genes) that were tested, in 17 allele pairs significant differences were found (14 genes, see figure 3B). In 10 allele pairs, the preferred allele was significantly more expressed than its partner in all individuals.

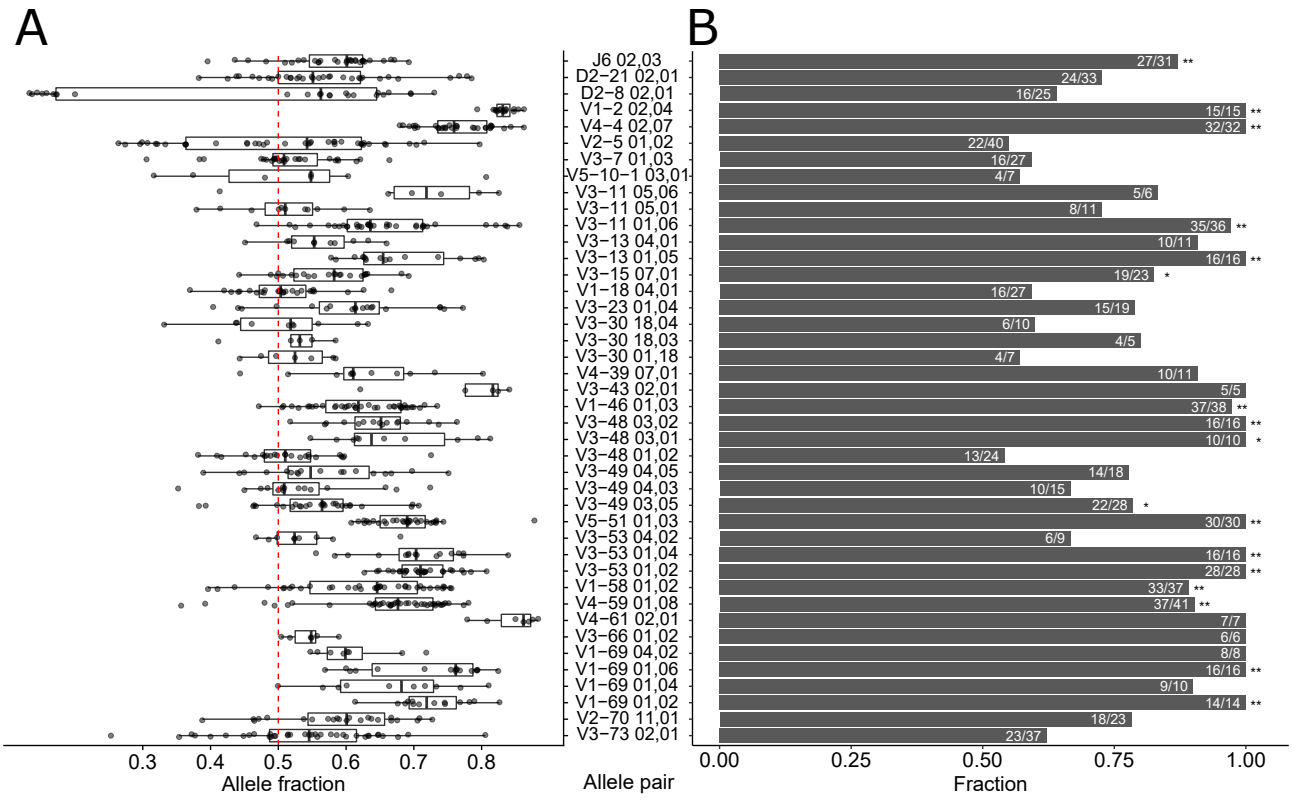

Figure 3: **Biased allele usage.** (A) Box plot of allele pairs relative usage of heterozygous individuals. Only allele pairs which were observed in more than 5 individuals are shown. Each point represents an individual. The allele fraction (X axis) corresponds to the allele that is written first in each row, and is the dominant allele in most individuals. (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 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 32 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 4A 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 4B). *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 S10 (second lowest individual) 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.

Second, similar to the pattern observed in both chromosomes (figure 1), *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 46 of the 95 individuals a deletion in both chromosomes of one of these gene pairs was detected using the binomial test (figure 1). This is consistent with the assumption that all individuals (not only the J6 heterozygous ones) have one of these deletions in each chromosome.

Nine individuals have deletions in the adjacent genes *D3-3* and *D6-6*. In fact, this deletion stretch might spread 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 (Boyd et al., 2010). 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 (p 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.

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

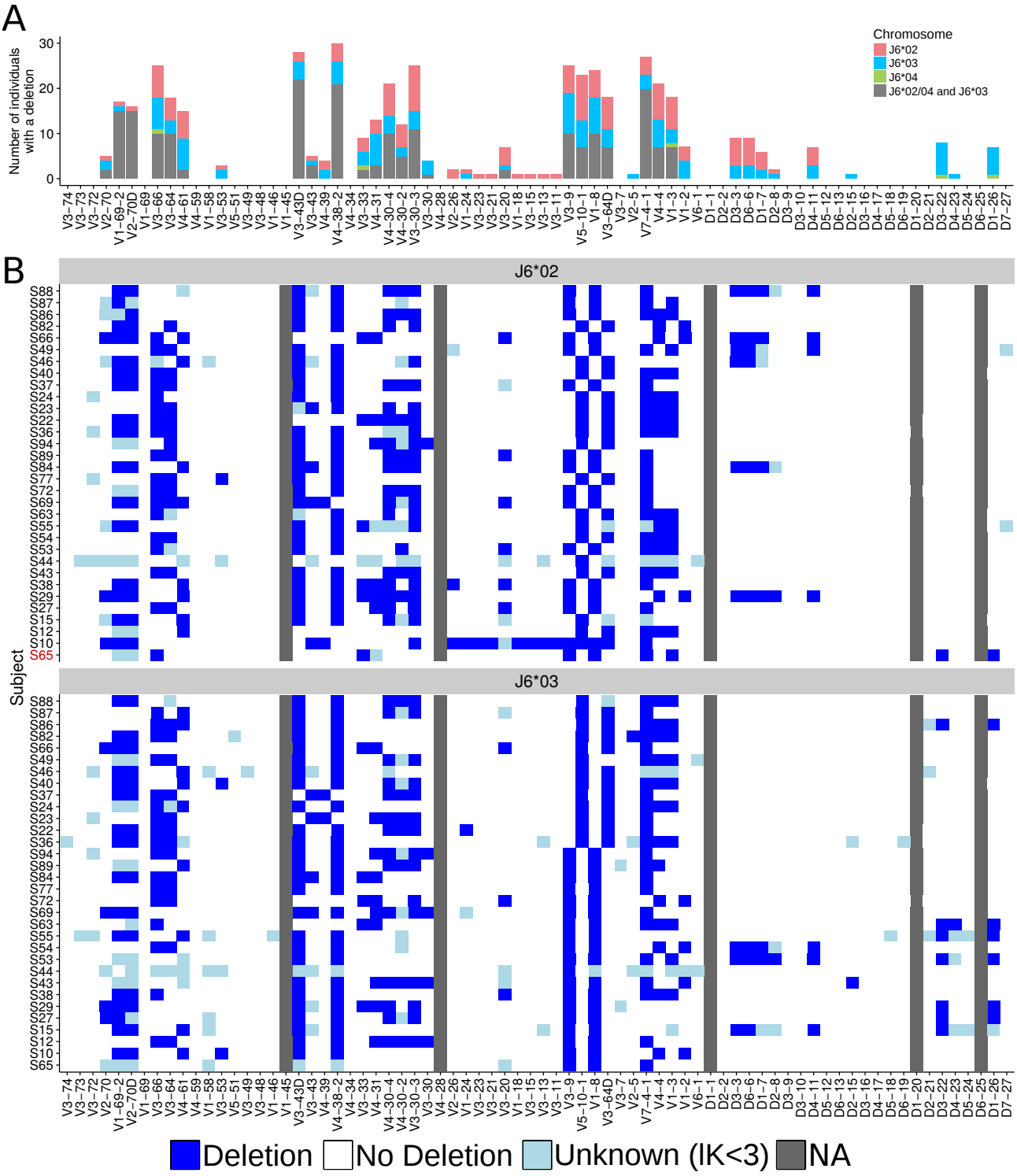

Figure 4: **Gene deletion inference along each chromosome.** (A) The distribution of *V* and *D* gene deletions along each chromosome in 32 individuals that are heterozygous for *J6*, as inferred by haplotype (light red, blue, and green) and by the binomial test (gray) (B) A heatmap of *V* and *D* gene deletions and suspected deletions for each of the 32 heterozygous individuals in *J6*. Each row represents an individual, and each column represents *V* or *D* gene. Blue represents a deletion ($lK > 3$), and light blue represents a suspected deletion ($lK < 3$). 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*02* is present. Sample S65, 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 (figure 5A and 5B). Five *V* genes, *V3-66, V4-39, V4-38-2, V1-8*, and *V1-3*, demonstrated, in most cases, a clear usage cutoff between samples with one chromosome deletion and samples with no deletion (figure 5C). 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 because IMGT has mis-classified several allele sequences in the *V4-4/V4-59/V4-61* complex. It may therefore be that individuals with apparently two *V4-61* alleles actually have a *V4-4* or *V4-59* allele with an erroneous *V4-61* name.

When *D3-3* is deleted in one chromosome (in our cohort, this gene was not deleted from both chromosomes in any individual, see figure 1D), it appears to be compensated by higher *D3-10* usage (figure 5D, as suggested in (Kidd et al., 2012)). 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 5E 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(S32), 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*.

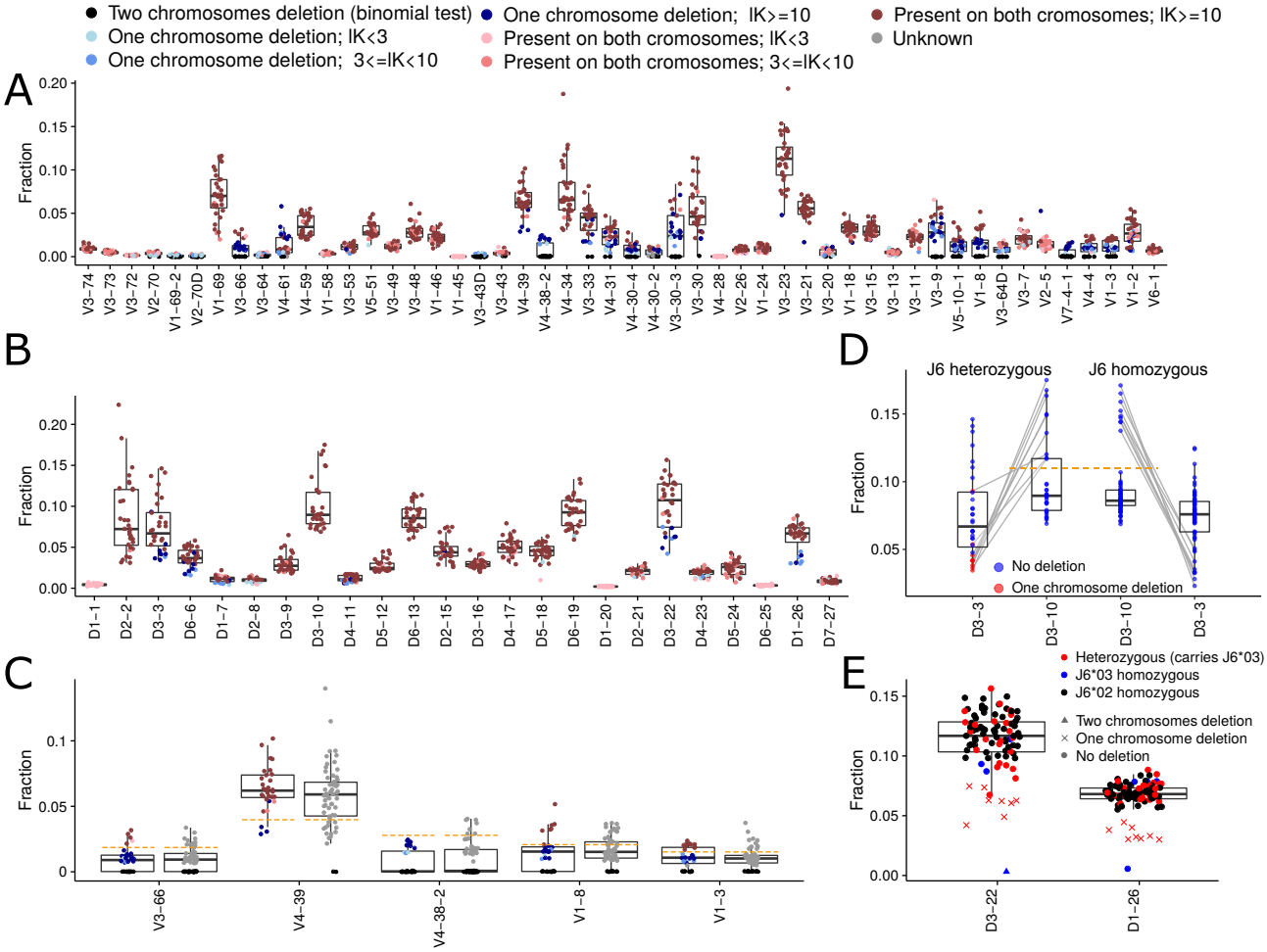

Figure 5: **Inferring single gene deletions by their relative usage.** (A,B) Relative usage of *V* and *D* genes from *J6* heterozygous individuals. Each dot represents an individual. Color corresponds to gene deletion from both chromosomes (black), single chromosome (blue), or no deletion (red). Shades correspond to the certainty level of deletion inference. (C) Box plots of the usage of five *V* genes. Each gene distribution appears once for the *J6* heterozygous individuals (left) and once for the *J6* homozygous individuals (right). The orange dashed cutoffs were placed to separate individuals with a single chromosome deletion from individuals with no deletions in that gene. (D) 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 dashed cutoffs were placed to separate individuals with high *D3-10* usage and low *D3-3* usage. Blue points represent individuals with no *D3-3* deletion, and red points represent individuals with a single chromosome deletion. (E) 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.

Haplotype can be inferred using the D2-8 and D2-21 genes

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* (figure S1A). 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 figure S3A and S3B). 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 based on the anchor *J6* gene and on the anchors *D2-2, D2-21, and D2-8* genes for a subset of heterozygous individuals for these genes. We have plotted the Jaccard distance between the haplotypes of these individuals as a function of allele bias (figure S3C). 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). Only samples with at least 5 *V* genes that can be compared were taken into account. This resulted in a reduction in *D* heterozygous samples to 31% for *D2-21* and 17% for *D2-8* (figure S4A). 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 32 to 51 of 95 (figure S4A).

To test the *D*-based haplotype, we first inferred the haplotype of *D* by *J6*. This resulted in a chromosomal linkage map between the alleles of these two genes (see example for one individual in figure S4B). Then, we compared the *V* haplotype inferred based on *J6* with the ones inferred by the new candidate *D* genes (see example for one individual in figure S4C and S4D). The comparison showed excellent resemblance between the haplotype inferred by *J6* and by *D2-8* and *D2-21* (Jaccard distance < 0.1, figure S3C), indicating that these *D* genes can be used for reliable haplotype inference.

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 2), 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. Since we do not know in advance the *V* gene haplotype, we cannot determine the connection between the major allele in the haplotype resulting from a given *V* gene and the haplotype resulting from the next one. Hence, this V_{pooled} approach is exposed to contradicting assignments of alleles 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 6A 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 6B). We identified an lK value ($lK = 12$) which optimized the precision rate ($\sim 90\%$ for $lK(J) = 2$ and $\sim 70\%$ for $lK(J) = 7$) with an acceptable price in sensitivity ($\sim 43\%$ for $lK(J) = 2$ and $\sim 56\%$ for $lK(J) = 7$). The relatively low levels of sensitivity result from an overall reduction in the number of identified deletions (figure 6C). 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 6D). Applying the same approach to the entire cohort, we identified single chromosome *D* gene deletions also in *J6* homozygous individuals (figure S5). *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.

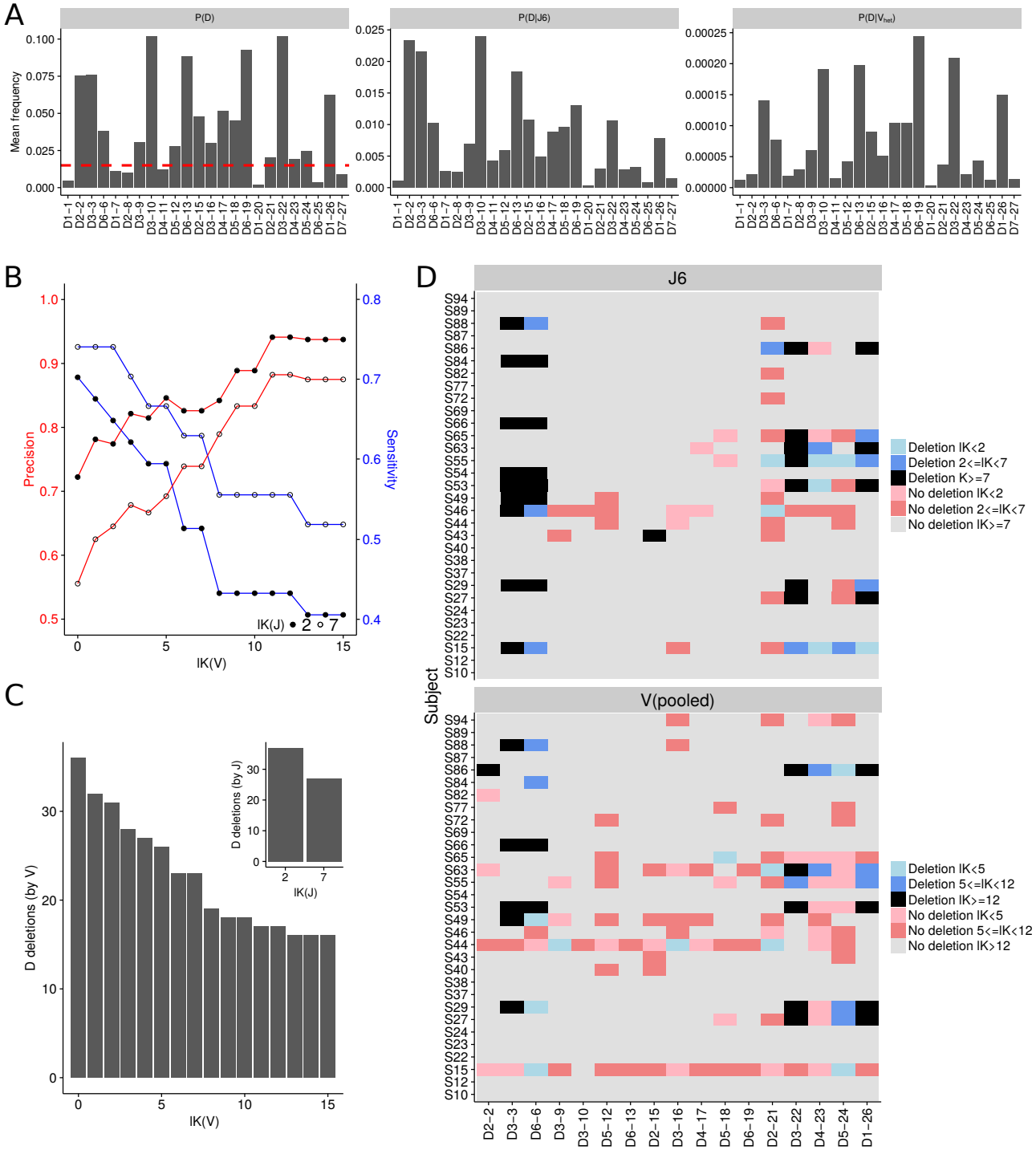

Figure 6: **Gene deletion inference along each chromosome by multiple V genes.** A comparison between *D* haplotype inference using a pool of *V* genes vs. *J6* as anchors, in 32 *J6* heterozygous individuals. (A) The mean relative *D* gene usage. Left: mean *D* gene usage. Dashed red line corresponds to the 1.5% threshold which was used to filter out low 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 genetics 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 (Boyd et al., 2010; Kidd et al., 2012; Gadala-Maria et al., 2015; Ralph and Matsen IV, 2016; Corcoran et al., 2016). Here, we implemented similar ideas in a Bayesian framework that allowed us to: 1. Attribute a confidence level to each result, and 2. Infer haplotype based on *V, D,* or *J* genes. We applied our method to the largest dataset, to date, of naïve B cells. 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 (Schroeder Jr, 2015) and between *D* and *J* gene recombinations (Kidd et al., 2016). 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 (Kidd et al., 2012; Oettinger et al., 1990; Matsuda et al., 1998). 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 of 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 its usage frequency to call it as deleted in individual S32, even though it should have been determined as a deletion by comparing its usage to other individuals (figure 5E). 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. Nonetheless, only for 5 *V* genes, we could suggest a usage frequency cutoff implying deletion on a single chromosome in samples without inferred haplotype (figure 5D). We showed that *V3-9* and *V1-8* deletion is mutually exclusive with *V5-10-1* and *V3-64D* (figure 4). 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 coding region of the allele, the RSS. Such mutations can cause the specific "deleted" alleles not to appear in the repertoire. Hence, in order to validate inferred deletions and duplication events, sequencing of the genomic region encoding the antibody heavy chain locus is needed. Other major factors that influence the strength of our approach are the 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.

Acknowledgments

This research was supported by grants from ISF (grant number 832/16) to G.Y., P.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 (SKGHMED-017) to L.M.S..

References

Benichou, Jennifer, Rotem Ben-Hamo, Yoram Louzoun, and Sol Efroni, 2012, Rep-Seq: uncovering the immunological repertoire through next-generation sequencing, *Immunology* 135, 183–191.

Boyd, Scott D, Bruno A Gäeta, Katherine J Jackson, Andrew Z Fire, Eleanor L Marshall, Jason D Merker, Jay M Maniar, Lyndon N Zhang, Bita Sahaf, Carol D Jones, Birgitte B Simen, Bozena Hanczaruk, Khoa D Nguyen, Kari C Nadeau, Michael Egholm, David B Miklos, James L Zehnder, and Andrew M Collins, 2010, Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements, *Journal of immunology (Baltimore, Md.: 1950)* 184, 6986–6992.

Breden, Felix, Eline T Luning Prak, Bjoern Peters, Florian Rubelt, Chaim A Schramm, Christian E Busse, Jason A Vander Heiden, Scott Christley, Syed Ahmad Chan Bukhari, Adrian Thorogood, et al., 2017, Reproducibility and reuse of Adaptive Immune Receptor Repertoire data, *Frontiers in immunology* 8.

Corbett, Simon J, Ian M Tomlinson, Erik LL Sonnhammer, David Buck, and Greg Winter, 1997, 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.

Corcoran, Martin M, Ganesh E Phad, Néstor Vázquez Bernat, Christiane Stahl-Hennig, Noriyuki Sumida, Mats AA Persson, Marcel Martin, and Gunilla B Karlsson Hedestam, 2016, Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity, *Nature communications* 7, 13642.

Fridman, Wolf Herman, Franck Pagès, Catherine Sautès-Fridman, and Jérôme Galon, 2012, The immune contexture in human tumours: impact on clinical outcome, *Nature Reviews Cancer* 12, 298–306.

Gadala-Maria, Daniel, Moriah Gidoni, Gur Yaari, and Steven H. Kleinstein, 2018, Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data, *Frontiers in Immunology* SUBMITTED.

Gadala-Maria, Daniel, Gur Yaari, Mohamed Uduman, and Steven H Kleinstein, 2015, 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.

Hodgkin, Philip D., William R. Heath, and Alan G. Baxter, 2007, The clonal selection theory: 50 years since the revolution, *Nature Immunology* 8, 1019–1026.

Kidd, Marie J, Zhiliang Chen, Yan Wang, Katherine J Jackson, Lyndon Zhang, Scott D Boyd, Andrew Z Fire, Mark M Tanaka, Bruno A Gaëta, and Andrew M Collins, 2012, 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.

Kidd, Marie J, Katherine JL Jackson, Scott D Boyd, and Andrew M Collins, 2016, DJ pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes, *The Journal of Immunology* 196, 1158–1164.

Kirik, Ufuk, Lennart Greiff, Fredrik Levander, and Mats Ohlin, 2017, Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery, *Molecular Immunology* 87, 12–22.

Laserson, Uri, Francois Vigneault, Gur Yaari, Daniel Gadala-Maria, Mohamed Uduman, Jason A. Vander Heiden, William Kelton, Sang Taek Jung, Yi Liu, Jonathan Laserson, Raj Chari, Je-Hyuk Lee, Ido Bachelet, Brendan Hickey, Erez Lieberman-Aiden, Bozena Hanczaruk, Birgitte B. Simen, Michael Egholm, Daphne Koller, George Georgiou, Steven H. Kleinstein, and George M. Church, 2014, High-resolution antibody dynamics of vaccine-induced immune responses, *Proc. Natl. Acad. Sci. U. S. A.* 111, 4928–4933.

Lefranc, M.-P., V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, Y. Wu, E. Gemrot, X. Brochet, J. Lane, L. Regnier, F. Ehrenmann, G. Lefranc, and P. Duroux, 2009, IMGT, the international ImMunoGeneTics information systemR), *Nucleic Acids Research* 37, D1006–D1012.

Li, Shuo, Marie-Paule Lefranc, John J Miles, Eltaf Alamyar, Véronique Giudicelli, Patrice Duroux, J Douglas Freeman, Vincent DA Corbin, Jean-Pierre Scheerlinck, Michael A Frohman, et al., 2013, IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling, *Nature communications* 4, 2333.

Matsuda, Fumihiko, Kazuo Ishii, Patrice Bourvagnet, Kei-ichi Kuma, Hidenori Hayashida, Takashi Miyata, and Tasuku Honjo, 1998, The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus, *Journal of Experimental Medicine* 188, 2151–2162.

Mattila, Petri S, Jan Schugk, Hongyan Wu, and Olli Mäkelä, 1995, Extensive allelic sequence variation in the J region of the human immunoglobulin heavy chain gene locus, *European journal of immunology* 25, 2578–2582.

Murphy, Kenneth, 2011, *Janeway's Immunobiology*, 8th edition (Garland Science).

Oettinger, Marjorie A, David G Schatz, Carolyn Gorka, and David Baltimore, 1990, RAG-1 and RAG-2, adjacent genes that synergistically activate V (D) J recombination, *Science* 248, 1517–1523.

Palanichamy, A., L. Apeltsin, T. C. Kuo, M. Sirota, S. Wang, S. J. Pitts, P. D. Sundar, D. Telman, L. Z. Zhao, M. Derstine, A. Abounasr, S. L. Hauser, and H.-C. von Budingen, 2014, Immunoglobulin class-switched B cells form an active immune axis between CNS and periphery in multiple sclerosis, *Sci. Transl. Med.* 6, 248ra106–248ra106.

Ralph, Duncan K, and Frederick A Matsen IV, 2016, Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation, *PLoS computational biology* 12, e1004409.

Ravetch, Jeffrey V, Ulrich Siebenlist, Stanley Korsmeyer, Thomas Waldmann, and Philip Leder, 1981, Structure of the human immunoglobulin μ locus: characterization of embryonic and rearranged J and D genes, *Cell* 27, 583–591.

Schroeder Jr, Harry W, 2015, The evolution and development of the antibody repertoire, *Frontiers in immunology* 6, 33.

Snir, Omri, Luka Mesin, Moriah Gidoni, Knut EA Lundin, Gur Yaari, and Ludvig M Sollid, 2015, Analysis of celiac disease autoreactive gut plasma cells and their corresponding memory compartment in peripheral blood using high-throughput sequencing, *J Immunol* 1402611.

Sok, Devin, Uri Laserson, Jonathan Laserson, Yi Liu, Francois Vigneault, Jean-Philippe Julien, Bryan Briney, Alejandra Ramos, Karen F. Saye, Khoa Le, Alison Mahan, Shenshen Wang, Mehran Kardar, Gur Yaari, Laura M. Walker, Birgitte B. Simen, Elizabeth P. St. John, Po-Ying Chan-Hui, Kristine Swiderek, Stephen H. Kleinstein, Galit Alter, Michael S. Seaman, Arup K. Chakraborty, Daphne Koller, Ian A. Wilson, George M. Church, Dennis R. Burton, and Pascal Poignard, 2013, The Effects of Somatic Hypermutation on Neutralization and Binding in the PGT121 Family of Broadly Neutralizing HIV Antibodies, *PLoS Pathog* 9, e1003754.

Stern, Joel NH, Gur Yaari, Jason A Vander Heiden, George Church, William F Donahue, Rogier Q Hintzen, Anita J Huttner, Jon D Laman, Rashed M Nagra, Alyssa Nylander, et al., 2014, B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes, *Science Translational Medicine* 6, 248ra107–248ra107.

Tsioris, Konstantinos, Namita T Gupta, Adebola O Ogunniyi, Ross M Zimnisky, Feng Qian, Yi Yao, Xiaomei Wang, Joel N Stern, Raj Chari, Adrian W Briggs, et al., 2015, Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing, *Integrative Biology* .

Vander Heiden, Jason A, Panos Stathopoulos, Julian Q Zhou, Luan Chen, Tamara J Gilbert, Christopher R Bolen, Richard J Barohn, Mazen M Dimachkie, Emma Ciafaloni, Teresa J Broering, et al., 2017, Dysregulation of B cell repertoire formation in myasthenia gravis patients revealed through deep sequencing, *The Journal of Immunology* 198, 1460–1473.

Vander Heiden, Jason A and Yaari, Gur and Uduman, Mohamed and Stern, Joel NH and OConnor, Kevin C and Hafler, David A and Vigneault, Francois and Kleinstein, Steven H., 2014, pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires, *Bioinformatics* 30, 1930–1932.

Wardemann, Hedda, and Christian E Busse, 2017, Novel approaches to analyze immunoglobulin repertoires, *Trends in immunology* 38, 471–482.

Watson, Corey T, Karyn M Steinberg, John Huddleston, Rene L Warren, Maika Malig, Jacqueline Schein, A Jeremy Willsey, Jeffrey B Joy, Jamie K Scott, Tina A Graves, et al., 2013, 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.

Wu, Yu-Chang B, Louisa K James, Jason A Vander Heiden, Mohamed Uduman, Stephen R Durham, Steven H Kleinstein, David Kipling, and Hannah J Gould, 2014, 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.

Wu, Yu-Chang Bryan, David Kipling, and Deborah K Dunn-Walters, 2012, Age-related changes in human peripheral blood IGH repertoire following vaccination, *Frontiers in immunology* 3, 193.

Yaari, Gur, and Steven H Kleinstein, 2015, Practical guidelines for B-cell receptor repertoire sequencing analysis, *Genome medicine* 7, 121.

Yahalom, Galit, Daria Weiss, Ilya Novikov, Therese B Bevers, Laszlo G Radvanyi, Mei Liu, Benjamin Piura, Stefano Iacobelli, Maria T Sandri, Enrico Cassano, et al., 2013, An antibody-based blood test utilizing a panel of biomarkers as a new method for improved breast cancer diagnosis, *Biomarkers in cancer* 5, 71.

Ye, Jian, Ning Ma, Thomas L. Madden, and James M. Ostell, 2013, IgBLAST: an immunoglobulin variable domain sequence analysis tool, *Nucleic Acids Research* gkt382.