frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 31 January 2023
DOI 10.3389/fimmu.2023.1073414

OPEN ACCESS

EDITED BY
Eric Pinaud,
UMR7276 Contrôle des réponses immunes B et des lymphoprolifération (CRIBL), France

REVIEWED BY
Scott A. Jenks,
Emory University, United States
Philippe Auguste Robert,
University of Oslo, Norway

*CORRESPONDENCE
Gunilla B. Karlsson Hedestam
gunilla.karlsson.hedestam@ki.se

SPECIALTY SECTION
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

RECEIVED 18 October 2022
ACCEPTED 09 January 2023
PUBLISHED 31 January 2023

CITATION
Hardt U, Corcoran MM, Narang S, Malmström V, Padyukov L and Karlsson Hedestam GB (2023) Analysis of IGH allele content in a sample group of rheumatoid arthritis patients demonstrates unrevealed population heterogeneity. *Front. Immunol.* 14:1073414. doi: 10.3389/fimmu.2023.1073414

Analysis of IGH allele content in a sample group of rheumatoid arthritis patients demonstrates unrevealed population heterogeneity

Uta Hardt[1,2], Martin M. Corcoran[2], Sanjana Narang[2], Vivianne Malmström[1], Leonid Padyukov[1] and Gunilla B. Karlsson Hedestam[2*]

[1]Division of Rheumatology, Department of Medicine Solna, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden and Karolinska University Hospital, Stockholm, Sweden, [2]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

Immunoglobulin heavy chain (IGH) germline gene variations influence the B cell receptor repertoire, with resulting biological consequences such as shaping our response to infections and altering disease susceptibilities. However, the lack of information on polymorphism frequencies in the IGH loci at the population level makes association studies challenging. Here, we genotyped a pilot group of 30 individuals with rheumatoid arthritis (RA) to examine IGH allele content and frequencies in this group. Eight novel IGHV alleles and one novel IGHJ allele were identified in the study. 15 cases were haplotypable using heterozygous IGHJ6 or IGHD anchors. One variant, IGHV4-34*01_S0742, was found in three out of 30 cases and included a single nucleotide change resulting in a non-canonical recombination signal sequence (RSS) heptamer. This variant allele, shown by haplotype analysis to be non-expressed, was also found in three out of 30 healthy controls and matched a single nucleotide polymorphism (SNP) described in the 1000 Genomes Project (1KGP) collection with frequencies that varied between population groups. Our finding of previously unreported alleles in a relatively small group of individuals with RA illustrates the need for baseline information about IG allelic frequencies in targeted study groups in preparation for future analysis of these genes in disease association studies.

KEYWORDS

immunoglobulin heavy chain, germline gene variation, haplotyping, recombination signal sequence, polymorphism, rheumatoid arthritis, population genetics

Introduction

The human immune system recognizes an indefinite number of antigenic determinants through the use of antigen receptors on naïve T and B cells (TCRs and BCRs). BCRs are transmembrane-bound immunoglobulin (IG) molecules composed of heavy and light chains, encoded by rearranged variable (V), diversity (D) and joining (J) genes, and a constant gene segment (1). The vast sequence diversity of naive IG heavy chain (IGH) repertoires is mediated by combinatorial recombination of V, D and J genes at recombination signal sequences (RSS). These are characterized by a conserved heptamer at the 5' end, a 12/23 bp spacer, and a conserved nonamer at the 3' end. The IG diversity is further increased by non-templated nucleotide insertions and/or trimming of nucleotides at the V-D and D-J junctions (2) during the recombination process, and subsequent somatic hypermutation (SHM) and isotype switching that increases antibody-antigen affinity and function.

The IGH locus in humans contains frequent copy number variations and an abundance of pseudogenes interspersed between highly similar functional genes, resulting in a challenge for genomic sequencing (3–5). Traditional short read sequencing approaches, such as those utilized for the 1000 Genomes Project (1KGP) (6, 7), result in ambiguous assemblies of the IGH region, limiting the ability to accurately identify IG gene variations. Furthermore, for complex genomic regions such as the IGH locus, high coverage genomic sequencing is low throughput. Scaling up to large numbers of individuals, such as those in disease cohorts, remains a major challenge. To date, high coverage sequencing of the IGH locus has been reported for a limited number of individuals (8, 9), but there are ongoing studies to extend this analysis for over 100 individuals (10).

In the recent years, the development of next generation sequencing (NGS) approaches that allow sufficient sequencing length and depth has enabled opportunities to infer germline alleles from full-length V(D)J sequences using tools such as IgDiscover (11), TIgGER (12), Partis (13, 14) or IMPre (15) to determine germline IGHV and IGHJ sequences at an individual level. NGS-based immune repertoire sequencing is high throughput offering possibilities to define allele frequencies in larger groups of individuals and enables the application of inferred haplotype analysis to reveal gene duplication and structural variation (16).

The Epidemiological Investigation of Rheumatoid Arthritis (EIRA) study is a population-based case-control study based on incident cases of rheumatoid arthritis in Sweden. EIRA comprises adult individuals in areas from southern and central Sweden from May 1996 and onwards. Cases were recruited through rheumatology clinics in the study area (17). Controls were randomly selected from the population registry shortly after case identification and were matched on age, sex and residential area. 96% of participating cases and 60% of participating controls provided blood samples. Cases and controls were invited to answer an extensive questionnaire. To date, the study population consists of several thousand cases and controls. Here, to examine IGH allelic variation in a pilot group comprising 30 individuals belonging to the EIRA study, in preparation for larger association studies, we generated IgM libraries and applied the germline inference tool IgDiscover to each case.

Within the 30 case dataset, we classified the IGH germline allelic genotype of each case, thereby enabling the identification of novel allelic variants and biased allelic expression in the IgM repertoire by inferred haplotype analysis. Variants identified were validated using haplotyping and Sanger sequencing. Most notably, we discovered a novel IGHV4-34 gene variant, which was present in 10% of the cases and that could be validated by restriction fragment length polymorphism analysis. Of critical importance, a set of 30 control cases was therefore included in the restriction fragment length polymorphism analysis to delineate the frequency of this allele in healthy individuals of the same population group. The results obtained here expand our knowledge of IGH gene diversity and highlight the importance of extended and population specific profiling of the baseline content and frequencies of IG alleles prior to performing disease association studies. Any such population based allelic frequency bias has the potential to confound association studies and the EIRA study set of patient samples and matched controls provides an ideal opportunity to extend our findings.

Results

Personalized genotyping of IGHV alleles

We generated IgM libraries from 30 female individuals without genetic indication of non-European ethnicity from the EIRA study using whole blood RNA as the input material. The libraries were generated by reverse transcription PCR using an IgM gene specific primer, followed by multiplex PCR using IGHV leader-specific primers, and subsequent index PCR to add the Illumina adapters, as previously described (18). After sequencing using the Illumina MiSeq platform, we used the IgDiscover germline inference tool to infer individualized genotypes of each case (Figure 1A). We inferred eight novel IGHV alleles among the 30 individuals as indicated in the allelic heatmap with an underscore and suffix S number. Four novel alleles, IGHV1-69*04_S4205, IGHV3-30*02_S4989, IGHV3-30*03_S8990 and IGHV3-43D*04_S5432, were present in one individual, two novel alleles, IGHV3-66*03_S2497 and IGHV4-39*01_S2720, were present in two individuals. IGHV3-13*01_S3164 was present in four individuals and IGHV3-66*03_S1480 was present in five individuals. The inferred expressed IGHV alleles for all 30 individuals are shown as a heatmap in Figure 1B.

Haplotyping by IGHJ6 revealing heterozygosity status

As previously shown, approximately 25-30 percent of humans are heterozygous for the IGHJ6 gene (19). Consistent with this, nine of the 30 individuals included in this study were found to be heterozygous with the presence of both IGHJ6*02 and IGHJ6*03. Since VDJ recombination occurs locally along a single chromosomal strand, IGHJ6 heterozygosity can be used to anchor IGHV alleles to a specific haplotype. In this manner IGHV alleles can be revealed as homozygous, heterozygous, duplicated or deleted (16, 20, 21). In examining haplotype plots for the haplotypable cases studied here, we observed an unexpected hemizygosity for IGHV4-34 in two

10.3389/fimmu.2023.1073414

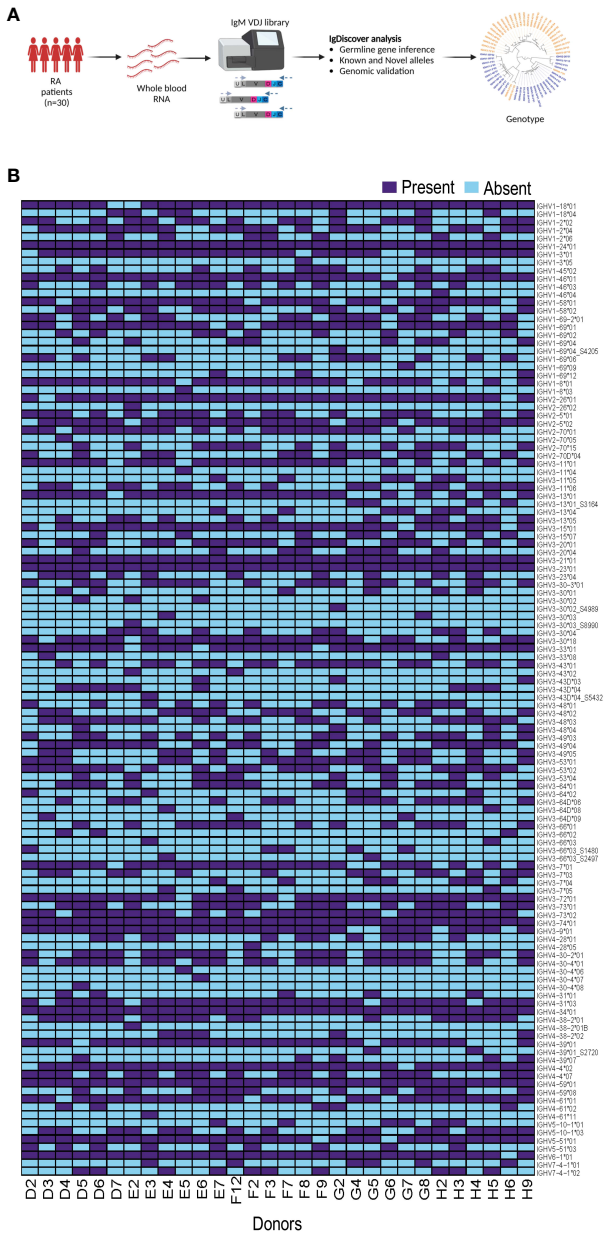

FIGURE 1
Production of immunoglobulin heavy chain genotype information from 30 individuals. **(A)** The flowchart shows the process from whole blood RNA isolation, IgM library preparation and subsequent IgDiscover analysis. **(B)** Heatmap showing expressed IGHV alleles for all 30 individuals.

individuals, H9 and F8. In both cases, we could map one allele, IGHV4-34*01, to one chromosome while expression of IGHV4-34 was absent from the other chromosome (Figures 2A, B). In the reference haplotype from individual E2, there is evident IGHV4-34*01 homozygosity (Figure 2C).

RSS SNP variant associating with decreased IGHV4-34 expression

IGHV4-34*01 is a very common allele of IGHV4-34, a gene that shows low levels of allelic variation in previous inference studies (19). Absence of an expressed allele on one chromosome can be explained by either a genomic deletion of IGHV4-34 on that chromosome or by mechanisms that interfere with either the recombination, expression or stability of the variant allele.

Genomic amplification and Sanger sequence analysis of IGHV4-34 in both cases with monoallelic expression (H9 and F8) resulted in the identification of IGHV4-34 heterozygosity, with two allelic variants found in both individuals. In each case, these variants included the common IGHV4-34*01 allele, and a novel variant, IGHV4-34*01_S0742, that shared 100% sequence identity to IGHV4-34*01 across the entire V sequence, but contained a single nucleotide polymorphism (SNP) within the second position of the seven bp RSS heptamer sequence, resulting in the non-canonical heptamer sequence, CTCAGTG.

Since 21 of the 30 cases could not be J6 haplotyped we could not investigate if there were additional cases that were heterozygous for the IGHV4-34 RSS variant allele using the haplotyping approach. However, the IGHV4-34 RSS variation results in the introduction of a restriction site for the enzyme DdeI that recognizes the target sequence CTNAG that is absent from the IGHV4-34*01 RSS. We therefore PCR amplified an 84 bp segment spanning the polymorphism using genomic DNA from all 30 cases. These were analysed by restriction fragment length polymorphism analysis (RFLP), which allowed us to identify heterozygous cases containing this allelic variant. DdeI digestion of an 84bp amplicon produced three diagnostic bands of 84, 53 and 31bp in heterozygous cases containing the variant (Figure 3A), while individuals without the variant yield only the undigested 84 bp product. We found that the two haplotyped individuals, F8 and H9, as well as one additional individual, G7, produced the diagnostic bands (Figure 3B; Supplemental Figure 1 and Supplemental Tables 1, 2), demonstrating a prevalence of the RSS variant allele, IGHV4-34*01_S0742, in 10 percent of cases in our study. Updated IGHV genotypes of H9, F8 and G7 including the non-functional IGHV4-34*01_S0742 allele are shown in Supplemental Figure 2. The non-canonical CTCAGTG heptamer is consistent with an interference in effective VDJ recombination, which explains the IGHV4-34 hemizygosity in these individuals.

Population frequency of IGHV4-34*01_S0742

The IGHV4-34*01_S0742 RSS variant nucleotide was consistent with SNP rs148342179 (A/T) (Figure 3C). This SNP was found to be present in 0.5 percent of all samples of the 1000 Genomes Project (ALL) and in 1.8 percent of the European samples (EUR). It was not found in either the African (AFR) or East Asian samples (EAS) and was only present at low frequency in the other populations such as the American (AMR: 0.6 percent) and the South Asian (SAS: 0.1 percent). However, rs148342179 is present at much higher frequency among Finnish individuals (FIN: 4.5 percent) within the European population (Figure 3D). To determine the frequency of this allele in a matched control population, we performed genomic PCR and DdeI restriction digestion on a set of 30 control samples. We found the variant was present in three control samples, I6, J5 and K2 and we validated the presence of the IGHV4-34*01_S0742 allele in all three

10.3389/fimmu.2023.1073414

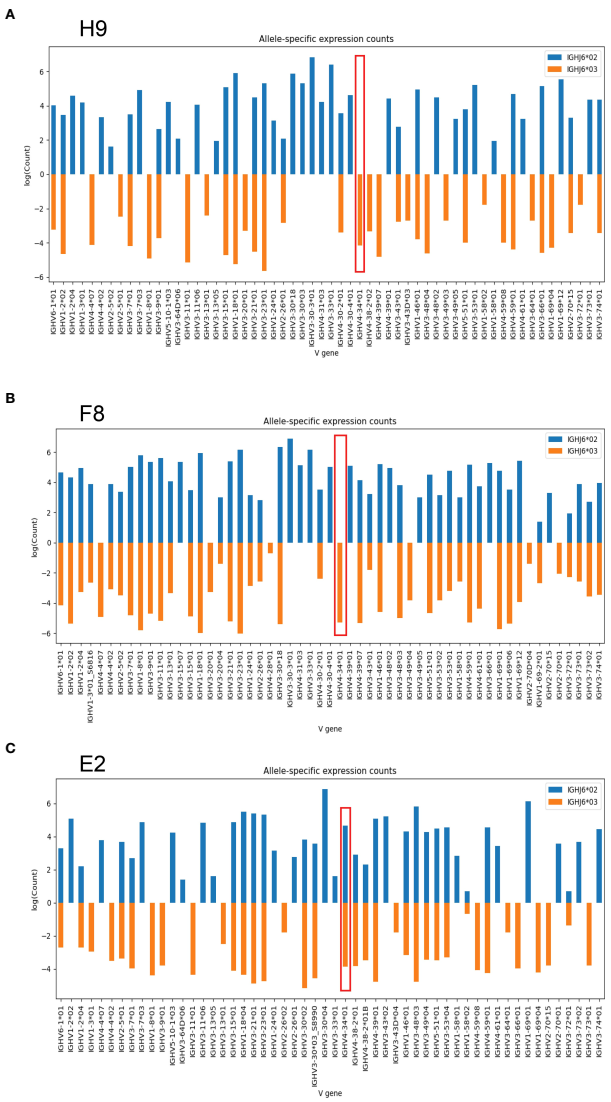

FIGURE 2
Haplotype analysis of two individuals based on IGHJ6 heterozygosity. **(A)** J6-anchored haplotype analysis of individual H9 reveals IGHV4-34 hemizygosity (red box). **(B)** J6-anchored haplotype analysis of individual F8 reveals IGHV4-34 hemizygosity (red box). **(C)** Reference haplotype from individual E2 reveals IGHV4-34 homozygosity.

individuals by targeted genomic PCR and Sanger sequencing (Figure 3E and Supplemental Tables 1, 2).

Identifying a novel IGHJ6 allele as an anchor for haplotyping

In addition to the novel IGHV4-34 allele, we found a novel IGHJ6 allele, IGHJ6*05_S6029, in individual D5. This variant J allele was characterized by the deletion of a triplet base GGT, which we validated using targeted genomic PCR and Sanger sequencing. This deletion is consistent with a SNP variant, rs74454466 and results in a deletion of a single glycine at the protein level (Figure 4A and Supplemental Table 2). The presence of this variant IGHJ6 allele provided an additional heterozygous J6 anchor, enabling IGHV haplotype analysis of D5 (Figure 4B), thereby allowing J6-anchored haplotype analysis of a total of 10 of the 30 individuals. In addition to the anchoring by J6 gene heterozygosity, it is possible to use IGHD3-10 or IGHD2-21 heterozygosity to infer V gene haplotypes. In the current study, this allowed haplotyping of five additional individuals of the 30 cases. In case D5, we identified a novel IGHD3-10 allele, IGHD3-10*03_S2198, providing two heterozygous anchors in this individual. This allowed a direct comparison of using heterozygous J and D genes for haplotyping, illustrating that the use of J anchors is preferred as this gives higher sequence counts for each V allele (Figures 4B, C). Finally, to validate the presence of the novel IGHJ6*05_S6029 allele, we used a heterozygous V gene, IGHV4-30-4*01/IGHV4-30-4*08, to haplotype J genes in D5, which clearly demonstrated the presence of IGHJ6*02 and IGHJ6*05_S6029 on the two separate chromosomes (Figure 4D).

Common structural variation

At the gene level, several common structural variations were identified in the 30 individuals. Duplication of the IGHV3-30 gene was apparent in three individuals and of the IGHV1-69 gene in five individuals. The complete set of IGHV3-30, IGHV3-30-3, IGHV4-31, IGHV3-33, IGHV4-30-2 and IGHV4-30-4 genes was present in 14/30 individuals and the genes for IGHV2-70D and IGHV1-69-2 were present in 13/30 individuals. The IGHV4-38-2 gene was absent in 10 individuals, while the IGHV3-43D gene was absent in 18 individuals. In addition, the genes for IGHV1-8 and IGHV3-9 were present in a homozygous state in 25/30 individuals, while the genes for IGHV3-64D and IGHV5-10-1 were present in homozygous state only in 15/30 individuals (19). (Figure 5). Overall, our analyses demonstrate an extensive variation in the IGH locus, at both structural and allelic levels.

Discussion

Here, we defined the IGH genotype of 30 individuals from the EIRA study. We found a novel IGHV4-34*01_S0742 allele, which was present in 10 percent of the individuals. This allele was characterized by a SNP, rs148342179, in the RSS region, which interfered with successful recombination, resulting in hemizygous expression of IGHV4-34. The location of the variant T nucleotide at position 2 of the RSS heptamer deviates from the canonical CA sequence described in previous analyses of RSS functionality and may therefore have a major inhibitory effect on recombination (22), as reflected by the absence of expression of the allele in cases H9 and F8. In addition to the inferred haplotype analysis, which shows a hemizygous loss of a functional IGHV4-34 in heterozygous cases, it is important to note that an A to T change in the RSS heptamer has not been described for any functional RSS heptamer in humans or other species. The starting CA dinucleotide is believed to be critical for the recombination process as shown by Kim et al. (23), with CA facilitating reduced base-stacking and enabling bending of the DNA helix during the recombination process. Consistently, Hu et al. (24) report 107 cryptic RSS sites, all of which require a CAC triplet at the beginning of the heptamer RSS sequence. Finally, Hoolehan et al. (25) recently showed that the CAC triplet is crucial feature in functional RSS heptamers,

10.3389/fimmu.2023.1073414

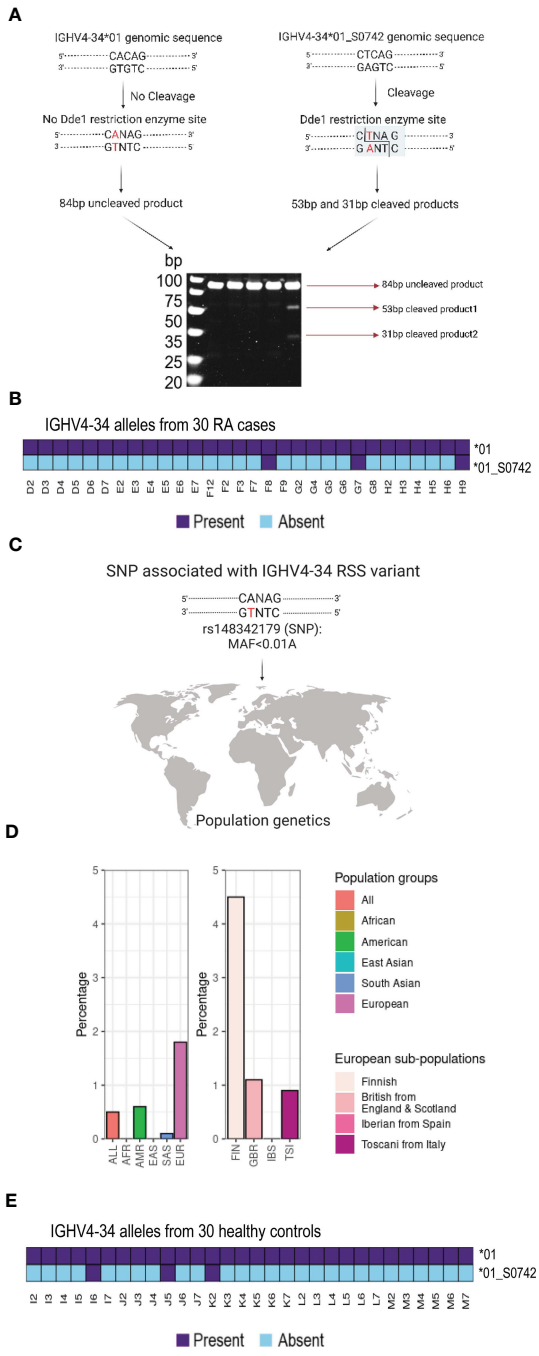

FIGURE 3
RSS variant analysis in the study group and in populations **(A)** Scheme of the restriction fragment length polymorphism is shown. **(B)** Three RA individuals contain the RSS variant of the IGHV4-34 gene (IGHV4-34*01_S0742). **(C)** Schematic overview of the SNP number associated with the polymorphism **(D)** Data from the 1000 Genomes Project reveals that the coding minus strand has a T/A variant in SNP rs148342179 that is especially frequent in the Finnish (FIN) population. **(E)** Three healthy controls contain the RSS variant of the IGHV4-34 gene (IGHV4-34*01_S0742).

with all functional non-canonical RSS heptamer sequences showing nucleotide variation solely in the final four nucleotides.

Although the population origin was not recorded within the EIRA study, we know that the participants were collected in the middle and southern parts of Sweden. Data from the 1000 Genomes Project show that SNP rs148342179 found in our cases was present at 4.5 percent in the Finnish population. The population frequency of IGHV4-34*01_S0742 identified in our study can be expected to result in around one percent of the Finnish/Scandinavian population being homozygous for this RSS variant, thereby resulting in the full absence of IGHV4-34 expression in the immunoglobulin repertoires in these individuals. Similarly, homozygous deletions of the IGHV3-30/IGHV4-31 region (hv3005) were found to be enriched in RA patients (26) and SLE patients compared to ethnically matched healthy individuals of Korean (27) or Caucasian (28) ethnicity. This deletion has also been associated with susceptibility to chronic idiopathic thrombocytopenic purpura in Caucasians (29).

The IGHV4-34 gene has previously been shown to be associated with autoimmunity (30), but the specificities responsible for this remain unclear. IGHV4-34 is highly used in both the IgM and IgG repertoire (31), and has been identified in studies of autoreactivity towards type I blood antigens (32). In particular, autoreactive antibodies encoded by the IGHV4-34 gene have been shown to be raised in patients with systemic lupus erythematosus (SLE) using the rat monoclonal antibody 9G4 (33) that has been claimed to recognize human IGHV4-34. In a repertoire analysis study, Bashford-Rogers et al. (34) found increased usage of IGHV4-34 in 10 patients with SLE of mixed Caucasian and Asian ethnicity, 11 patients with eosinophilic granulomatosis with polyangiitis of mainly northern European ancestry and 23 patients with Crohn's disease of mainly northern European ancestry compared to healthy individuals of mainly northern European ancestry. In SLE, there is a plethora of autoreactivities described, however the over-representation of IGHV4-34 among SLE clonal expansions (35) is insufficient to conclusively prove specific autoreactivity of that germline gene, particularly in the context of other diseases. The IGHV4-34*01 allele contains a germline-encoded Asn-X-Ser/Thr motif in its CDR2 region, which allows N-linked glycosylation at this site (36). High levels of SHM-introduced variable domain glycans have been associated with autoantibodies in rheumatoid arthritis (37). At the same time, it has been shown, that an antibody produced by a self-reactive B cell had reduced capacity of autoantigen binding, when N-linked glycosylation was introduced (38). However, we note that in this particular RA group analyzed, the frequency of the IGHV4-34*01_S0742 variant exactly matches that of the control group. While the allele may have functional significance within the population at large, particularly in the case of homozygosity, we did not find a clear signal that it was relevant to the RA group in the current study.

A similar observation to the IGHV4-34*01_S0742 allele, where an RSS polymorphism affected the V gene usage, was reported for a variant kappa V gene, IGKV2-29D, enriched in a Native American population. In that study the variant was associated with *Haemophilus influenzae* type b susceptibility (39, 40). Of critical importance to this study was the observation that the frequency of the heterozygous IGHV4-34 variant in the patient group was identical to that found in the matched control group. The frequencies of IGHV4-34*01_S0742 in both the Swedish EIRA patient samples and in the healthy controls studied here are consistent with 1000 Genomes population data showing SNP rs148342179 is found at highest frequencies in the Finnish population set, a region geographically and historically closely linked to Sweden. It is interesting to note that the haplotypable cases H9 and F8 share an identical string of 14 alleles,

10.3389/fimmu.2023.1073414

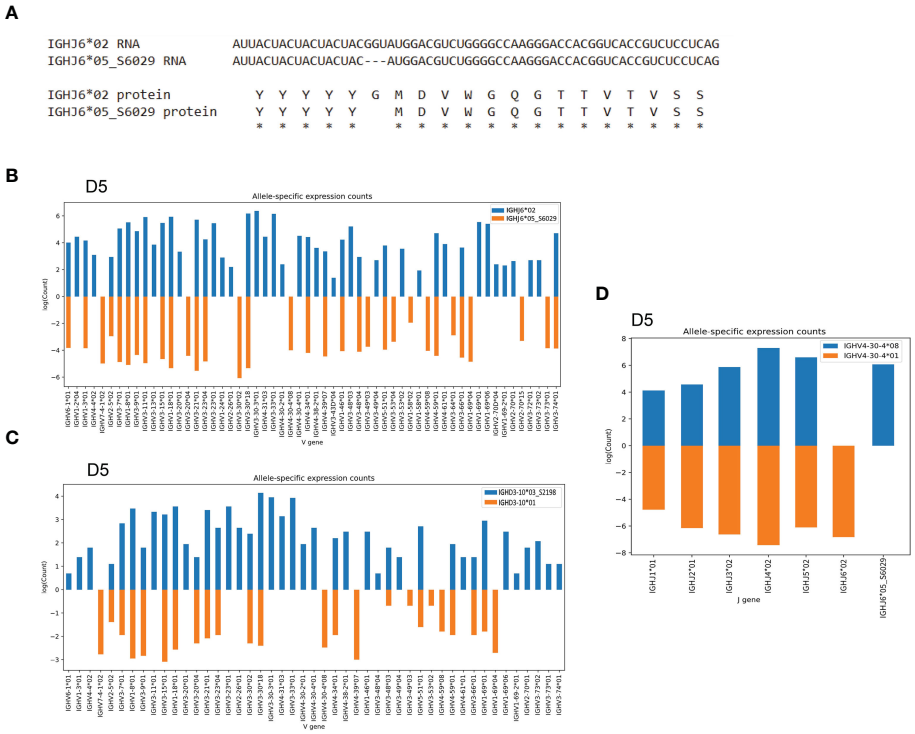

FIGURE 4
Novel IGHJ6 allele as anchor for haplotyping **(A)** The alignment shows the deletion of a base triplet in the novel IGHJ6*05_S6029 results in a deletion of a glycine residue at protein level. **(B)** The novel IGHJ6*05_S6029 variant in combination with the regular IGHJ6*02 allele can be used for IGHV gene haplotyping in the D5 individual. **(C)** Case D5 can be haplotyped based on IGHD3-10 gene heterozygosity. **(D)** The J6 alleles are haplotypable using heterozygous IGHV4-30-4 alleles as anchors, showing the clear separation of the IGHJ6*05_S6029 allele to a different chromosome compared to the IGHJ6*02 allele.

from IGHV3-30-3*01 to IGHV5-51*01, that encompasses the non-functional IGHV4-34*01_S0742 allele, indicating the possibility that the mutation may be historically recent. Further studies may reveal how common this shared segment is within this and other populations.

SNP arrays and whole genome sequencing enabling association studies have revolutionized the field of complex disease genetics. Genome wide association studies have implicated causal genes and mechanisms, drug targets, disease biomarkers and risk prediction. There are more than 150 risk loci reported in genome wide association studies of rheumatoid arthritis patients from different populations (41). The class II human leukocyte antigen (HLA) loci show a very strong association with rheumatoid arthritis. Together with efficacy of anti-CD20 therapy, rituximab (42), these data indicate a role for T and B cells in rheumatoid arthritis pathology. Genotyping the complex IGHV locus with the aim of identifying disease associated variation is challenging. The process necessitates the ability to produce accurate and comprehensive genotypes encompassing genes that are utilized at both low and high frequency in the B cell repertoire. Structural variations due to gene deletion or duplication events, in addition novel alleles that may be present in the population (18) should also be identifiable. High variation in the IG loci may introduce confounding problems in association studies if the case and control groups are not population-stratified (43) since the goal is to identify disease related polymorphisms that are present at different frequencies in the disease group compared to the control group. The observation that IG genes display great population variation has been discussed with the respect to limited sampling groups currently represented by databases (44).

Notably, genome wide association studies currently lack high quality information for the highly complex immunoglobulin heavy and light chain loci. As immunoglobulins play an important role as B cell effector molecules in rheumatoid arthritis, these loci may equally be a source of genetic variance within the general population that contributes to disease susceptibility and are therefore an important target for disease association studies. An example of such variation was identified in the indigenous populations of the South Pacific, who expose a high burden of rheumatic heart disease. The allele IGHV4-61*02 is associated with an increased risk of rheumatic heart disease (45) in that population set. Identification of the functional allelic variation in the IGHV locus is challenging, but feasible using specialized next generation sequencing methods and computational analysis, as applied in our studies.

Previous studies on antibody repertoires in rheumatoid arthritis are to our knowledge mostly limited to studies of heavy chain CDR3 sequences derived from comparatively small libraries constructed with a limited number of IGHV primers (46). In our studies, we use a 5' MTPX primer set that was previously shown to capture a comprehensive set of IGHV genes and we sequence full-length VDJ transcripts and determine IG genotypes in each study participant (18). Individualized genotyping of individuals provides the basis for analysing IGHV germline gene variation and expressed repertoires in

10.3389/fimmu.2023.1073414

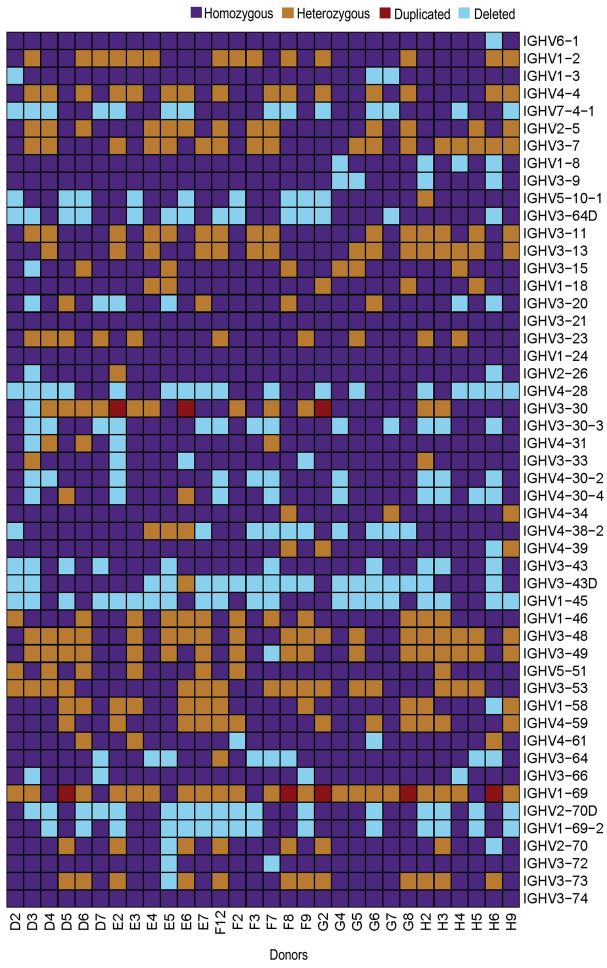

FIGURE 5
Heatmap showing IGHV genes for all 30 individuals. IGHV genes are represented on y-axis in chromosomal order with color-key on the top-right.

any disease, including rheumatoid arthritis. However, interpreting IGHV allelic variation can be challenging since there may be population-based variations in allelic frequencies. Identification of immunoglobulin gene polymorphisms in a disease study group, even at high frequency, should not be assumed to be disease related if that frequency matches that of the population from which the study group is drawn.

The current study was not designed or powered for association analysis; however, this can be performed with the larger EIRA study with matching control samples already available (47). Association studies for immunoglobulin alleles are so far limited to studies that are reviewed elsewhere (48, 49); thus the impact of this variation on disease risks is insufficiently investigated (50, 51). The results of our pilot investigation demonstrates that population-based genetic variance of IG alleles is likely to be common (45, 52–54). Without adequate information of the expected frequencies of immunoglobulin alleles in the population, erroneous associations may be identified. Likewise, real associations may become apparent only when accurate information of allelic frequencies in the target population is well established.

Materials and methods

Experimental design

We collected whole blood samples in PAXgene tubes (Qiagen) from 30 rheumatoid arthritis (RA) patients and genomic DNA samples from 30 healthy controls. The inclusion criteria were recruitment within the EIRA (Epidemiological Investigation of Rheumatoid Arthritis) study (55), no indication from the Immunochip Array genotyping (Illumina) for non-European ancestry. Whole blood RNA (extracted with PAXgene Blood miRNA Kit, PreAnalytiX, Qiagen) was used to prepare cDNA for subsequent construction of IgM libraries (18) to infer Immunoglobulin heavy chain genotypes. Sampled DNA was used to validate inferred novel variants by RFLP and Sanger sequencing.

Patients

All RA patients and healthy controls were recruited as part of the EIRA study under ethics permits #1023-96 and #2006/476-31/4 obtained from Regionala Etikprövningsnämnden, Stockholm. This study comprises cases and control subjects from the middle and southern parts of Sweden. All samples were taken in hospital-based or privately run rheumatology units in the study area in accordance with the Helsinki Declaration and written informed consent was given by each patient before entering the study. In the current study, we included 30 female RA patients comprising 10 shared epitope-negative (SE negative) anti-citrullinated protein antibody-negative (ACPA negative) individuals with a mean age of 62.5 years, eight SE positive ACPA negative individuals with a mean age of 57.4 years and 12 SE positive ACPA positive individuals with a mean age of 58.7 years.

Library preparation

IgM libraries were prepared according a previously published protocol (18). In brief, 200 ng of whole blood mRNA was reverse transcribed using the Sensiscript Reverse Transcription kit (Qiagen) and reverse gene specific primer with a unique molecular identifier (UMI) and a universal reverse amplification sequence. 2 µl of purified (Qiagen MinElute PCR purification kit) cDNA was amplified using the universal reverse primer and the chain-specific 5' forward leader primer mix, using the KAPA HiFi Hotstart Ready Mix (Roche). The product of around 480bp was gel purified (Qiagen MinElute Gel Extraction kit). 5 to 10ng of the gel-purified product were used for the indexing PCR, as detailed previously (18). The forward indexing primer P5_R1 and the reverse indexing primer P7_R2_I1-27 were added in 10 cycle PCR reaction using the KAPA HiFi Hotstart Ready Mix (Roche). The final libraries were purified and quantified according to Illumina's manufacturer's instructions. The Illumina Version 3 (2x300bp) sequencing kit was employed for sequencing the libraries with the addition of 13% PhiX174 DNA (12pM) as positive control.

Computational analysis

Library analysis was performed using the IgDiscover version 0.12.4 with default settings. IgDiscover pre-processed the libraries for quality control and subsequently performed expression analysis and generated individualized databases. The IMGT reference database (May 2019) was used (56), with the addition of some recently described new alleles (57). The databases were aligned using CLUSTAL W (58) and the trees were plotted using FigTree (version 1.4.4). Haplotypes were generated using the plotalleles module of IgDiscover with a chromosomal filter of 25%.

Software

Heatmaps and 1000 Genomes Project data were plotted with R (version 3.6.3) using R studio (version 1.2.1335). In particular, we used tidyverse (version 1.3.0), cowplot (version1.1.1) and gplots (version 3.1.3) packages.

Restriction fragment length polymorphism

84bp long genomic DNA around the polymorphism was amplified using specific primers designed using BLAT (59) (UCSC genome browser, Supplemental Table 3). Amplicons were digested by DdeI (Thermo Fisher Scientific) for 4h at 37°C. Samples were run on a TBE 20% polyacrylamide gel (Invitrogen) at 100V for 3.5h and stained with SYBR Green I nucleic acid gel stain (Thermo Fisher Scientific) in TE buffer (10mM Tris, 1mM EDTA, pH8) for 40 min while shaking. Images were taken on a Biorad Geldoc instrument.

Genomic validation

Primers encompassing the IGHV4-34 gene and the IGHJ6 gene were designed using BLAT (UCSC genome browser) to validate the presence of novel alleles (Supplemental Table 3). 10ng of genomic DNA template were amplified using KAPA Hifi Hotstart Ready Mix (Roche). The product was gel purified (Qiagen MinElute Gel Extractration kit) and ligated to CloneJet pJET 1.2 vector (Thermo Scientific). 1 to 2µl of ligation product were transformed in XL10-Gold Ultracompetent Cells (Agilent) following manufacturer's instructions. Transformed Cells were grown overnight on a 100 mg/ml ampicillin LB agarose plate before colony screening. IGHV4-34 gene transformed colonies were screened using the RFLP primers, IGHJ6 gene transformed colonies were screened for inserts. Positive colonies were grown in LB medium overnight. Bacterial cultures were purified with GeneJET Plasmid Miniprep Kit (Thermo Scientific) and Sanger sequenced (Genewiz). Sanger sequences of validated alleles have been deposited in the GenBank under accession numbers OL807662 (IGHV4-34*01_S0742) and OL807663 (IGHJ6*05_S6029).

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://figshare.com/, 10.17044/scilifelab.21316677 https://www.ncbi.nlm.nih.gov/genbank/, OL807662, OL807663.

Ethics statement

The studies involving human participants were reviewed and approved by Regionala Etikprövningsnämnden, Stockholm. The patients/participants provided their written informed consent to participate in this study.

Author contributions

UH, SN and MC designed and performed the experiments. UH, MC, SN, VM, LP and GKH analysed the data. UH, MC and GKH wrote the paper. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by a European Research Council Advanced Grant (agreement number 78816) and a Distinguished Professor grant from the Swedish Research Council (agreement number 00968) to GKH.

Acknowledgments

We thank Barbro Larsson for help with the EIRA sample collection and the EIRA study group for sharing data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1073414/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Restriction digestion of a PCR product comprising the variant position identified in IGHV4-34*01_S0742 (rs148342179) shows the diagnostic 53 and 31bp bands in individuals F8 and G7, and H9.

SUPPLEMENTARY FIGURE 2
IGHV genotypes of H9, F8 and G7 including the IGHV4-34*01_S0742 allele.

References

1. Nielsen SCA, Boyd SD. Human adaptive immune receptor repertoire analysis–past, present, and future. *Immunol Rev* (2018) 284:9–23. doi: 10.1111/IMR.12667

2. Hozumi N, Tonegawa S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc Natl Acad Sci U.S.A.* (1976) 73:3628–32. doi: 10.1073/pnas.73.10.3628

3. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* (2010) 184:6986, LP – 6992. doi: 10.4049/jimmunol.1000445

4. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, Boyd SD, et al. The inference of phased haplotypes for the immunoglobulin h chain V region gene loci by analysis of VDJ gene rearrangements. *J Immunol* (2012) 188:1333–40. doi: 10.4049/jimmunol.1102097

5. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, et al. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *Front Immunol* (2020) 11:2136. doi: 10.3389/fimmu.2020.02136

6. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. *Nat 2015 5267571* (2015) 526:68–74. doi: 10.1038/nature15393

7. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nat 2015 5267571* (2015) 526:75–81. doi: 10.1038/nature15394

8. Matsuda F, Ishii K, Bourvagnet P, Kuma KI, Hayashida H, Miyata T, et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med* (1998) 188:2151–62. doi: 10.1084/JEM.188.11.2151

9. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet* (2013) 92:530–46. doi: 10.1016/J.AJHG.2013.03.004

10. Rodriguez OL, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT, et al. Antibody repertoire gene usage is explained by common genetic variants in the immunoglobulin heavy chain locus. *bioRxiv* (2022) 2022.07.04.498729. doi: 10.1101/2022.07.04.498729

11. Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) 7:13642. doi: 10.1038/ncomms13642

12. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. 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.* (2015) 112:E862–70. doi: 10.1073/pnas.1417683112

13. Ralph DK, Matsen FA4. Consistency of VDJ rearrangement and substitution parameters enables accurate b cell receptor sequence annotation. *PloS Comput Biol* (2016) 12:e1004409. doi: 10.1371/journal.pcbi.1004409

14. Ralph DK, Matsen FA. Per-sample immunoglobulin germline inference from b cell receptor deep sequencing data. *PloS Comput Biol* (2019) 15:e1007133. doi: 10.1371/JOURNAL.PCBI.1007133

15. Zhang W, Wang I-M, Wang C, Lin L, Chai X, Wu J, et al. IMPre: An accurate and efficient software for prediction of T- and b-cell receptor germline genes and alleles from rearranged repertoire data. *Front Immunol* (2016) 7:457. doi: 10.3389/fimmu.2016.00457

16. 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. *Mol Immunol* (2017) 87:12–22. doi: 10.1016/j.molimm.2017.03.012

17. Klareskog L, Nordmark B, Lindblad S. On the organization of an early arthritis clinic. *Best Pract Res Clin Rheumatol* (2001) 15:1–15. doi: 10.1053/berh.2000.0122

18. Vázquez Bernat N, Corcoran M, Hardt U, Kaduk M, Phad GE, Martin M, et al. High-quality library preparation for NGS-based immunoglobulin germline gene inference and repertoire expression analysis. *Front Immunol* (2019) 10:660. doi: 10.3389/fimmu.2019.00660

19. Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I, et al. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nat Commun* (2019) 10:628. doi: 10.1038/s41467-019-08489-3

20. Kirik U, Greiff L, Levander F, Ohlin M. Data on haplotype-supported immunoglobulin germline gene inference. *Data Br* (2017) 13:620–40. doi: 10.1016/J.DIB.2017.06.031

21. Vázquez Bernat N, Corcoran M, Nowak I, Kaduk M, Castro Dopico X, Narang S, et al. Rhesus and cynomolgus macaque immunoglobulin heavy-chain genotyping yields comprehensive databases of germline VDJ alleles. *Immunity* (2021) 54:355–366.e4. doi: 10.1016/j.immuni.2020.12.018

22. Ru H, Zhang P, Wu H. Structural gymnastics of RAG-mediated DNA cleavage in V(D)J recombination. *Curr Opin Struct Biol* (2018) 53:178–86. doi: 10.1016/J.SBI.2018.11.001

23. Kim MS, Chuenchor W, Chen X, Cui Y, Zhang X, Zhou ZH, et al. Cracking the DNA code for V(D)J recombination. *Mol Cell* (2018) 70:358–370.e4. doi: 10.1016/j.molcel.2018.03.008

24. Hu J, Zhang Y, Meng F-L, Schatz DG, Correspondence FWA, Zhao L, et al. Chromosomal loop domains direct the recombination of antigen receptor genes. *Cell* (2015) 163:947–59. doi: 10.1016/j.cell.2015.10.016

25. Hoolehan W, Harris JC, Byrum JN, Simpson DA, Rodgers KK. An updated definition of V(D)J recombination signal sequences revealed by high-throughput recombination assays. *Nucleic Acids Res* (2022) 50:11696–711. doi: 10.1093/nar/gkac1038

26. Olee T, Yang PM, Siminovitch KA, Olsen NJ, Hillson J, Wu J, et al. Molecular basis of an autoantibody-associated restriction fragment length polymorphism that confers susceptibility to autoimmune diseases. *J Clin Invest* (1991) 88:193–203. doi: 10.1172/JCI115277

27. Cho M-L, Chen PP, Seo Y-I, Hwang S-Y, Kim W-U, Min D-J, et al. Association of homozygous deletion of the Humhv3005 and the VH3-30.3 genes with renal involvement in systemic lupus erythematosus. *Lupus* (2003) 12:400–5. doi: 10.1191/0961203303lu385oa

28. Huang DF, Siminovitch KA, Liu XY, Olee T, Olsen NJ, Berry C, et al. Population and family studies of three disease-related polymorphic genes in systemic lupus erythematosus. *J Clin Invest* (1995) 95:1766–72. doi: 10.1172/JCI117854

29. Mo L, Leu SJ, Berry C, Liu F, Olee T, Yang YY, et al. The frequency of homozygous deletion of a developmentally regulated vh gene (Humhv3005) is increased in patients with chronic idiopathic thrombocytopenic purpura. *Autoimmunity* (1996) 24:257–63. doi: 10.3109/08916939608994718

30. Pascual V, Victor K, Spellerberg M, Hamblin TJ, Stevenson FK, Capra JD. VH restriction among human cold agglutinins. the VH4-21 gene segment is required to encode anti-I and anti-i specificities. *J Immunol* (1992) 149:2337–44. doi: 10.4049/jimmunol.149.7.2337

31. Phad GE, Pinto D, Foglierini M, Akhmedov M, Rossi RL, Malvicini E, et al. Clonal structure, stability and dynamics of human memory b cells and circulating plasmablasts. *Nat Immunol* (2022) 23:1–10. doi: 10.1038/s41590-022-01230-1

32. Parr TB, Johnson TA, Silberstein LE, Kipps TJ. Anti-b cell autoantibodies encoded by VH 4-21 genes in human fetal spleen do not require *in vivo* somatic selection. *Eur J Immunol* (1994) 24:2941–9. doi: 10.1002/eji.1830241204

33. Stevenson FK, Wrightham M, Glennie MJ, Jones DB, Cattan AR, Feizi T, et al. Antibodies to shared idiotypes as agents for analysis and therapy for human b cell tumors. *Blood* (1986) 68:430–6. doi: 10.1182/BLOOD.V68.2.430.430

34. Bashford-Rogers RJM, Bergamaschi L, Mckinney F, Pombal DC, Mescia F, Lee JC, et al. Analysis of the b cell receptor repertoire in six immune-mediated diseases. *Nature* (2019) 574:122–6. doi: 10.1038/s41586-019-1595-3

35. Tipton CM, Fucile CF, Darce J, Chida A, Ichikawa T, Gregoretti I, et al. Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat Immunol* (2015) 16:755–65. doi: 10.1038/ni.3175

36. van de Bovenkamp FS, Derksen NIL, Ooijevaar-de Heer P, van Schie KA, Kruithof S, Berkowska MA, et al. Adaptive antibody diversification through n-linked glycosylation of the immunoglobulin variable region. *Proc Natl Acad Sci U.S.A.* (2018) 115:1901–6. doi: 10.1073/pnas.1711720115

37. Youings A, Chang SC, Dwek RA, Scragg IG. Site-specific glycosylation of human immunoglobulin G is altered in four rheumatoid arthritis patients. *Biochem J* (1996) 314 (Pt 2):621–30. doi: 10.1042/bj3140621

38. Sabouri Z, Schofield P, Horikawa K, Spierings E, Kipling D, Randall KL, et al. Redemption of autoantibodies on anergic b cells by variable-region glycosylation and mutation away from self-reactivity. *Proc Natl Acad Sci U.S.A.* (2014) 111:E2567–75. doi: 10.1073/pnas.1406974111

39. Feeney AJ, Atkinson MJ, Cowan MJ, Escuro G, Lugo G. A defective vkappa A2 allele in navajos which may play a role in increased susceptibility to haemophilus influenzae type b disease. *J Clin Invest* (1996) 97:2277–82. doi: 10.1172/JCI118669

40. Nadel B, Tang A, Lugo G, Love V, Escuro G, Feeney AJ. Decreased frequency of rearrangement due to the synergistic effect of nucleotide changes in the heptamer and nonamer of the recombination signal sequence of the V kappa gene A2b, which is

associated with increased susceptibility of navajos to haemophilus. *J Immunol* (1998) 161:6068–73. doi: 10.4049/jimmunol.161.11.6068

41. Padyukov L. Genetics of rheumatoid arthritis. *Semin Immunopathol* (2022) 44:47–62. doi: 10.1007/S00281-022-00912-0

42. Edwards JCW, Szczepa´nski S, Szechi´nski S, Filipowicz-Sosnowska A, Emery P, Close DR, et al. Efficacy of B-Cell–targeted therapy with rituximab in patients with rheumatoid arthritis. *N Engl J Med* (2004) 350:2572–81. doi: 10.1056/NEJMOA032534

43. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* (2019) 20:467–84. doi: 10.1038/s41576-019-0127-1

44. Peng K, Safonova Y, Shugay M, Popejoy AB, Rodriguez OL, Breden F, et al. Diversity in immunogenomics: the value and the challenge. *Nat Methods 2021 186* (2021) 18:588–91. doi: 10.1038/s41592-021-01169-5

45. Parks T, Mirabel MM, Kado J, Auckland K, Nowak J, Rautanen A, et al. Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nat Commun* (2017) 8:14946. doi: 10.1038/ncomms14946

46. Doorenspleet ME, Klarenbeek PL, Hair MJH, Schaik BDC, REE E, Kampen AHC, et al. Rheumatoid arthritis synovial tissue harbours dominant b-cell and plasma-cell clones associated with autoreactivity. *Ann Rheum Dis* (2014) 73:756–62. doi: 10.1136/ANNRHEUMDIS-2012-202861

47. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet* (2012) 44:1336–40. doi: 10.1038/ng.2462

48. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun* (2012) 13:363–73. doi: 10.1038/gene.2012.12

49. Mikocziova I, Greiff V, Sollid LM. Immunoglobulin germline gene variation and its impact on human disease. *Genes Immun* (2021) 22:205–17. doi: 10.1038/s41435-021-00145-5

50. Watson CT, Glanville J, Marasco WA. The individual and population genetics of antibody immunity. *Trends Immunol* (2017) 38:459–70. doi: 10.1016/j.it.2017.04.003

51. Collins AM, Yaari G, Shepherd AJ, Lees W, Watson CT. Germline immunoglobulin genes: Disease susceptibility genes hidden in plain sight? *Curr Opin Syst Biol* (2020) 24:100–8. doi: 10.1016/j.coisb.2020.10.011

52. Ohlin M. Poorly expressed alleles of several human immunoglobulin heavy chain variable genes are common in the human population. *Front Immunol* (2021) 11:603980/FULL. doi: 10.3389/FIMMU.2020.603980/FULL

53. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep* (2016) 6:20842. doi: 10.1038/srep20842

54. Sasso EH, Buckner JH, Suzuki LA. Ethnic differences of polymorphism of an immunoglobulin VH3 gene. *J Clin Invest* (1995) 96:1591–600. doi: 10.1172/JCI118198

55. Klareskog L, Lorentzen J, Padyukov L, Alfredsson L. Genes and environment in arthritis: can RA be prevented? *Arthritis Res* (2002) 4:S31–6. doi: 10.1186/AR566/FIGURES/2

56. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res* (2015) 43:D413–22. doi: 10.1093/NAR/GKU1056

57. Mikocziova I, Gidoni M, Lindeman I, Peres A, Snir O, Yaari G, et al. Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. *Nucleic Acids Res* (2020) 48:5499–510. doi: 10.1093/nar/gkaa310

58. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* (1994) 22:4673–80. doi: 10.1093/NAR/22.22.4673

59. Kent WJ. BLAT–the BLAST-like alignment tool. *Genome Res* (2002) 12:656–64. doi: 10.1101/GR.229202