Germline polymorphism in the immunoglobulin kappa and lambda loci explain variation in the expressed light chain antibody repertoire

Eric Engelbrecht[1], Oscar L. Rodriguez[1,2], William Lees[3], Zach Vanwinkle[1], Kaitlyn Shields[1], Steven Schultze[1], William S. Gibson[1], David R. Smith[1], Uddalok Jana[1], Swati Saha[1], Ayelet Peres[4], Gur Yaari[4], Melissa L. Smith[1,†],and Corey T. Watson[1,†]

[1] Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY, USA; [2] Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [3] Clareo Biosciences, Louisville, Kentucky, USA; [4] Department of Pathology, Yale School of Medicine, New Haven, CT, USA

Abstract

Variation in antibody (Ab) responses contributes to disease outcomes and therapeutic responsiveness, the determinants of which remain incompletely understood. This study demonstrates that polymorphisms in immunoglobulin kappa (IGK) and lambda (IGL) loci influence baseline differences in the Ab repertoire that preclude antigen-driven responses. We pair IGK/IGL genomic sequencing with Ab repertoire profiling to comprehensively resolve coding and non-coding single nucleotide and structural variants that explain gene usage variation at >70% of light chain genes, with direct impacts on the amino acid composition of complementarity determining regions. The identification of both intergenic and coding repertoire-associated variants indicates genetic modulation of gene usage via diverse mechanisms. Relative to IGL, IGK is characterized by extensive linkage disequilibrium and genetic co-regulation of gene usage, highlighting differential regulatory and evolutionary features. These results establish the contribution of IG polymorphism in the formation of the Ab repertoire, shedding light on our understanding of Ab-mediated immunity.

Teaser

This study demonstrates that germline polymorphisms in the immunoglobulin light chain loci modulate inter-individual variation in the repertoire of expressed antibody light chains, directly influencing the composition of the antibody antigen-binding landscape through impacts on light chain gene usage.

Introduction

Antibodies (Abs) are critical components of the adaptive immune system and are one of the most diverse protein families in the human body. The circulating Ab repertoire comprises hundreds of millions of unique antibodies (*1*, *2*), and its composition varies significantly between individuals (*1–3*). The variability likely contributes to the diverse Ab responses observed across clinical settings, including infection (*4–8*), autoimmunity (*9*), and cancer (*10*). Identifying the factors that contribute to variation in B cell-mediated immunity will inform disease diagnosis and treatment.

Human Abs are composed of two pairs of identical 'heavy' chains and 'light' kappa or lambda chains, encoded by genes located at three primary loci in the genome: the immunoglobulin (IG) heavy chain locus (IGH; 14q32.33), and the IG lambda (IGL; 22q11.2) and kappa (IGK; 2p11.2) loci (*11*). Across the IG loci, there are >240 phylogenetically related functional/open reading frame variable (V), diversity (D) (specific to IGH), and joining (J) genes (*12–15*) Selection of individual IG heavy and light chain genes during V(D)J recombination is a foundational step in the process of Ab generation. Increasing evidence indicates that genetic variation within the IG loci modulates the generation of the formation of the human Ab repertoire, contributing to the observed receptor diversity seen between individuals. This was initially supported in twin studies, which demonstrated that both naïve and antigen-stimulated Ab repertoires possess heritable characteristics (*16–18*). Additionally, non-coding and coding IG heavy chain (*12*, *19–26*) and light chain (*20*, *27–29*) germline variants have been shown to affect Ab gene usage and antigen specificity.

Utilizing matched adaptive immune receptor repertoire (AIRR)-seq and comprehensive long-read sequencing genotyping of IGH in a cohort of 154 individuals, we recently demonstrated that approximately half of common germline variants in IGH were associated with variation in usage frequencies of the majority of IGHV, IGHD, and IGHJ genes within the IgM (naïve-enriched) repertoire (*12*). Subsequently we showed that these genetic variants contribute to repertoire variation in early B cell developmental stages in the bone marrow, indicating direct impacts on V(D)J recombination (*30*). This has raised the prospect that variants within the IGK and IGL loci exert similar effects on the formation of the light chain repertoire, and will associate with inter-individual Ab variation in the periphery. The heavy and light chains of an Ab must be paired and compatible to achieve specificity and functionality, and both heavy and light chains contribute to antigen binding. The identification and characterization of antigen-specific and disease-associated Abs requires comprehensive models of naïve and antigen-experienced repertoires, for which a detailed understanding of genetic variation in all three loci is essential. To this end, it is critical to recognize that IG loci are enriched with structural variants (SVs), including segmental duplications and inversions, limiting the utility of short-read sequencing to characterize genetic variation (*12–15*, *31–34*). We previously demonstrated that long-read sequencing of diverse IGH (*12*, *13*, *33*), IGL (*14*), and IGK (*15*) haplotypes identifies genetic variation not recorded in reference databases, including catalogues of single-nucleotide variants and gene alleles.

Here, we pair long-read genomic sequencing of IGK and IGL with AIRR-seq at population-scale to identify *cis*-acting variants that explain inter-individual variation in light chain Ab repertoire features. We find that genetic variants in IGK and IGL associate with gene usage frequency for the majority of light chain V and J genes. These associations between germline polymorphism and gene usage persisted even in antigen-experienced Ab repertoires. Analysis of lead variants revealed mechanisms by which genetic sequence can impact gene usage frequencies, including missense and nonsense substitutions, as well as substitutions in regulatory elements, such as

recombination signal sequences. We find distinct structures of linkage disequilibrium (LD) in IGK and IGL, with relatively high LD in IGK associating with coordinated usage of multi-gene clusters. Finally, we demonstrate that genetic effects on gene usage contribute to amino acid variation in V genes, as well as physicochemical properties of the CDR3, linking germline variations with Ab features that contribute to antigen binding.

Results

Long-read genomic sequencing and genotyping of IGK and IGL loci and expressed light chain antibody repertoire sequencing

We combined targeted long-read sequencing of IGK and IGL loci in 177 healthy individuals with newly and previously (*12*) generated AIRR-seq for nearly all donors in the cohort (IGK, n=173, IGL, n=171). Donors ranged in age from 18 to 57 years (mean: 32.3), representing both biological sexes (male, n=90; female, n=87), and diverse genetic ancestry groups (**Supplementary Table S1**). Using our previously published method (*13*), we performed targeted long-read single molecule real-time (SMRT) sequencing of the IGK proximal and distal regions (*15*), and the IGL locus (*14*) (see **Supplementary Material**). From these data, we generated sample-level variant callsets (SNVs and SVs), and IGK/L gene and allele germline sets, representing the most comprehensive collection of such data types collected to date (see **Supplementary Material**). Importantly, this dataset allowed for the identification of >300 novel germline IG alleles, as well as novel SNVs and SVs (**Supplementary Figure 2**).

To profile expressed IGK and IGL transcripts, AIRR-seq data was generated using 5' rapid amplification of complementary DNA ends (5' RACE) on total RNA isolated from PBMCs collected from 173 individuals. With germline IGKV, IGKJ, IGLV, and IGLJ alleles for each individual, we limited V and J germline allele calls to those present in the germline on a per-individual basis. To enrich for antigen-naïve BCR sequences, we selected those containing V and J segments that matched germline allele sequences with 100% identity, unlikely to have undergone somatic hypermutation (SHM) (i.e. unmutated). The opposite approach was used to enrich for antigen-experienced BCR sequences, for which either the J or V (or both) segment varied from the germline allele sequence (i.e., were mutated). Importantly, personalized germline sets allowed us to account for the presence of previously undocumented alleles, and thus more accurately infer SHM. The usage frequencies of V and J genes among all unmutated or mutated unique BCR sequences were calculated for each individual. Together, these datasets allowed us to resolve comprehensive variant callsets to perform genetic association analyses with gene usage variation observed in the expressed light chain Ab repertoires.

Light chain gene usage is strongly associated with common IGK and IGL genetic variants in both antigen naive and experienced repertoires

Throughout the genome, genetic variation has been associated with molecular traits such as gene expression and splicing (*35–38*). We previously demonstrated that genetic variants in the IGH locus mediate the composition of peripheral IgM and IgG repertoires through effects on IGHV, IGHD, and IGHJ gene usage (*12*). Here, we followed this same quantitative trait locus (QTL) framework to test if light gene usage was associated with IGK and IGL variant genotypes in *cis*. Allele assignments to AIRR-seq reads were derived from a personalized germline allele set for each individual. This permitted disambiguation of IGK gene paralogs for individuals wherein each allele of a proximal paralog was distinct from each allele of the distal paralog, including IGKV1-12 and IGKV1D-12, IGKV1-13 and IGKV1D-13, and IGKV6-21 and IGKV6D-21

(**Supplementary Table S5**). Paralog pairs for which at least 160 individuals could not be disambiguated included IGKV1-33/1D-33, IGKV1-37/1D-37, IGKV1-39/1D-39, IGKV2-28/2D-28, and IGKV2-40/2D-40, and are referred to as ambiguous or "ambi" (e.g. IGKV1-39/1D-39 is IGKV1-39ambi).

We performed genetic association tests on unmutated ("antigen naive") and mutated ("antigen experienced") sets separately to identify *cis* effects in each of the two repertoire sets. In total, our analysis included 41 IGKV, 5 IGKJ, 33 IGLV, and 4 IGLJ genes; 2792 SNVs and 2 SVs in IGK; and 5198 SNVs and 1 SV in IGL. In the unmutated IGK repertoire, after Bonferroni multiple-testing correction (P < 5.0e-05), we identified 2,413 unique variants (2,411 SNVs, 2 SVs) that were statistically associated with gene usage changes in 25 (60%) IGKV and 3 (60%) IGKJ and genes (**Figure 1, Supplementary Table S6**). In the unmutated IGL repertoire, a set of 918 unique variants (917 SNVs, 1 SV) were associated with gene usage changes in 22 (67%) IGLV genes and 3 (75%) IGLJ (**Figure 1**, **Supplementary Table S6**). Notably, a large fraction of the genes identified in both the IGK (n=14 genes) and IGL (n=19 genes) unmutated repertoires also had significant gene usage QTLs (guQTLs) in the mutated repertoires (**Supplementary Figures 7-9**, **Supplementary Table S7**). However, guQTLs in the unmutated repertoires tended to have lower *P* values and explain more variance (R2) in gene usage (**Supplementary Figure 7**). We also noted stronger genetic effects in IGK relative to IGL; this included the observation that overall genetic similarity among subjects associated with more highly correlated IGK gene usage, a signal that was blunted in IGL (see **Supplemental Material, Supplementary Figure 10**). Together, these results show that usage of IG light chain genes is broadly impacted by germline genetic variants, and while these genetic effects are more prominent in the antigen naive repertoire, for many genes, those effects persist even following antigen exposure.

Genomic locations of guQTLs implicate genetic roles in coding and non-coding driven processes underlying antibody repertoire formation

Features of IG J genes include a single exon and RSS sequence, whereas V genes include an RSS sequence, first exon, intron, and second exon that encodes the variable region (V-region) (**Figure 1A**). While the majority of lead guQTLs across the unmutated repertoires were intergenic (n=22 (79%), IGK; n=22 (85%), IGL), coding lead variants were identified for 7 IGK genes (IGKJ3, IGKV2-29, IGKV1-5, IGKV1-13, IGKV2D-29, IGKV1D-13, and IGKV1-39ambi) (**Figure 1B**, **Supplementary Figure 8**) and 3 IGL genes (IGLV8-61, IGLV5-48, IGLV3-21) (**Figure 2C, Supplementary Figure 7**). In addition, 1 IGK lead guQTL and 3 IGL guQTLs fell within RSSs; these were lead variants for IGKV1-6, IGLV5-37, IGLV3-16, and IGLV5-48 (**Figure 1C, Supplementary Figure 7**).

Examples of lead guQTLs in coding, RSS, and intergenic regions are shown in **Figure 2**, including IGKV2-29, IGKV1-5, IGLV3-16, and IGLV9-49. The SNV-driven guQTL in this dataset with the lowest P value was for IGKV2-29 (*P* value = 2.1e-58, **Figure 2A**). This variant introduced a stop codon in V-region amino acid position 93 (**Figure 2B**), resulting in decreased usage of IGKV2-29 (**Figure 2C**). We also identified a lead guQTL associated with missense variants. In the case of IGKV1-5, two linked lead guQTLs (r^2 = 1) within codon 50 associated with a lysine to aspartic acid (AAG→GAT, K50D) change, resulting in an alteration of residue charge (**Figure 2D-E**). Individuals homozygous for K50 alleles, which represented six different IGKV1-5 coding alleles in this cohort, had lower gene usage (**Figure 2F-G**). As an example of a guQTL in the RSS, two lead variants in perfect LD were identified at positions 8 and 23 of the spacer for IGLV3-16 (**Figure 2H-J**). The reference haplotype had a C at position 8, which was

represented among consensus bases (C and T) at this position (**Supplementary Figure 11**), whereas the alternate haplotype had a G (**Figure 2J**). Among the C/C and G/G lead guQTL genotype groups, IGLV3-16 usage varied 3.8-fold on average (**Figure 2K**). As noted above, the majority of lead guQTLs in this dataset were in non-coding regions. For example, the lead guQTL for IGLV9-49 was 86 bp upstream of the first exon, and guQTLs were not identified in coding sequence or the RSS (**Figure 2L**). Mean IGLV9-49 usage varied by 3.5-fold between homozygous-reference and homozygous-alternate individuals at this lead guQTL (**Figure 2M**). Consistent with previous observations in IGH (*12*), many guQTLs within IGL overlapped curated transcription factor binding sites, representing an enrichment over background SNVs (see **Supplementary Material, Supplementary Figures 12-13**), suggesting likely roles for non-coding variants in the regulation of V(D)J recombination. Additionally, in IGK, we noted that both coding and non-coding regulatory variants altered proximal and distal gene usage biases (see **Supplementary Material, Supplementary Figures 14-15**).

Finally, in addition to SNV guQTLs, SVs resulting in gene copy number changes also made significant impacts on gene usage. Specifically, SVs were lead guQTLs for the genes IGKV1-NL1, IGKV1D-8, and IGLV5-39. In all cases, differential usage between genotypes followed an additive model in which gene usage increased with every additional haploid gene copy (**Figure 2N**). The lead variant associated with IGKV1D-8 usage was the SV deletion of the entire IGKV distal region (see **Supplementary Figure 4**). We noted that the number of diploid IGLJ2-3 cassette copies associated with the usage of IGLJ1 and IGLJ2-3ambi (**Supplementary Figure 16**); however, this CNV was not the lead QTL for these genes. The complexity of this SV will likely require analysis in larger cohorts and more detailed assessment of potential haplotype-specific effects.

In summary, these results indicate that many forms of genetic variation are associated with gene usage variation in the IGK and IGL repertoire. The variable localization of guQTLs in intergenic, RSS, and coding regions implicates causative roles for these genetic variants in plausibly regulating V(D)J recombination, transcription, and translation, as well as contributing to differential heavy-light chain pairing dynamics and antigen selection.

guQTLs within large linkage disequilibrium blocks in IGK create expansive networks of genes with correlated usage

In our previous study of IGH guQTLs, we observed that many SNVs were associated with the usage of individual genes. This included instances in which genes and associated guQTLs extended 10's to 100's of Kb; notably, these genes exhibited correlated usage patterns (*12*), suggestive of coordinated gene regulation. We sought to investigate whether similar features were present in the IGK and IGL loci.

First, we observed a difference between the number of guQTLs associated with IGK versus IGL gene usage in the unmutated repertoires. In IGK, the maximum number of associations with a single gene was 1,674, with 7 genes having over 400 (mean = 407) guQTL variants. By contrast, the maximum number of guQTL variants associated with a single IGL gene was 208 (mean = 47) (**Figure 1B-C**). This was not simply explained by the number of SNVs genotyped in the two loci, as we identified twice as many common variants in IGL relative to IGK. Thus, among all common SNVs in each locus, 84.5% in IGK and 17.7% in IGL were significantly associated with usage of at least one gene (**Figure 3A**). For IGK, we found that a large fraction of these guQTL variants were shared between at least two genes (n=2,049, 83.9%). This was in contrast to IGL, in which we only observed 254 (22.8%) IGL guQTL variants associated with >1 gene (**Figure 3B**). Likewise, at gene-level, 16 of 23 IGK guQTL genes shared at least

one significant variant with >5 other genes. In contrast, 0/26 IGL genes shared significant guQTL variants with >5 genes. The majority of IGL guQTL genes (24 of 26) shared variants with fewer than 3 other genes (**Figure 3C**).

To visualize these relationships between genes and guQTLs, we constructed networks in which nodes represented genes and edges represented connections between genes sharing at least one guQTL SNV. From these networks, we identified multi-member cliques, in which 2 or more genes were connected by at least one shared guQTL. For IGK, 21 of the 23 guQTL genes formed a single super clique, with embedded subcliques in which all genes were connected to one another through guQTL variants (**Figure 3D**). Demonstrative of interconnected gene usage, the largest subclique was composed of 9 guQTL genes associated with a single guQTL SNV (**Figure 3E**). In contrast to IGK, 13 of the 26 guQTL IGL genes were represented by 4 distinct cliques, all of which were smaller than the large clique observed in IGK and disconnected from one another (gene membership range = 2-5; **Figure 3D**).

The stark difference in IGK and IGL clique sizes (**Figure 3D**) suggested likely differences in the genetic haplotype structure between the two loci. To explore this, we estimated pairwise linkage disequilibrium (LD) between all common SNVs (MAF >= 5%) and determined blocks of LD (*39*, *40*) (**Supplementary Table S9**, see Methods). LD was more extensive in IGK (**Figure 4A**) relative to IGL (**Figure 4B**), with LD blocks >20 Kbp comprising 53.5% and 12.9% of the IGK and IGL loci, respectively, (**Figure 4C, Supplementary Table S10**). In contrast, 24.4% of IGL was within LD blocks <5 Kbp, compared to only 8.6% of the IGK locus (**Figure 4C**). The three largest LD blocks in IGK were 122 Kbp, 110 Kbp, and 76 Kbp, compared to the three largest LD blocks in IGL that were 34 Kbp, 26 Kbp and 24 Kbp (**Figure 4D**). As expected, the number of SNVs per block was positively correlated in both loci. Additionally, larger sets of genes in IGK fell within large LD blocks. We found that 27 of 47 IGKV genes were within LD blocks >20 Kbp, whereas 5 of 39 IGLV genes were within LD blocks >20 Kbp (**Figure 4F**). We also noted that 26 of 47 IGKV genes were within one of 6 LD blocks containing >1 gene, in contrast to IGL (**Supplementary Figure 17**), in which this was true for only 7 of 33 IGLV genes (**Supplementary Figure 18**). IGK guQTL SNVs were also more frequently in large LD blocks as 1,181 (49%) were in blocks >20 Kbp, whereas 233 (17%) IGL guQTL SNVs were in blocks >20 Kbp (**Figure 4G**).

These data demonstrate that a larger proportion of IGK sequence, genes, and guQTLs are contained within large LD blocks as compared to IGL. The overlap of LD blocks with guQTL and gene cliques suggests that extended haplotype structures within both loci likely contribute to coordinated gene regulation.

guQTLs are linked to missense variation in coding regions and physicochemical CDR3 properties in the IGK and IGL repertoires

Together, the data presented so far demonstrate that genetic variants within IGK and IGL associate with shifts in gene usage in the light chain repertoire. While the majority of lead guQTLs in both loci occurred in intergenic space, we wanted to see whether genetically driven usage shifts also associated with: 1) germline changes in V gene coding sequence spanning complementarity determining and framework regions (CDR1, CDR2, FWR1, FWR2, and FWR3); and 2) amino acid properties of CDR3 sequences spanning germline codons and junctions of recombined V and J genes. We reasoned that such associations would link changes in gene usage to BCR features likely to impact preferential pairing of available heavy and light chains and antigen binding. This has direct relevance to germline variants contributing to Abs associated with disease and vaccination (*20–22*, *24–26*, *41*).

First, we found that many lead guQTLs were associated with shifts in coding allele usage within the repertoire, representing LD between coding and non-coding SNVs. This was consistent with our previous investigation of IGH guQTLs (*12*). Specifically, for 18/32 (56%) tested IGK genes, individuals within different guQTL genotypes exhibited differential coding allele frequencies (Fisher's exact test, Bonferroni; $P < 0.002$). Likewise, in IGL we noted such associations for 9/24 (38%) guQTL genes (Fisher's exact test, Bonferroni; $P < 0.002$) (**Figure 5A-B**, **Supplementary Table S11**). Among these genes, the 12/18 (66%) in IGK, and 7/9 (77%) in IGL involved alleles carrying amino acid changes (**Figure 5C-D**). For the remaining genes in each locus, genes either exhibited allelic variation, but did not associate with guQTL genotype, or lacked appreciable allelic variation (major allele frequency >95%; **Figure 5C-D**). Examples of genes with coding allele variation linked to lead non-coding guQTL variants include IGKV2-30 and IGLV7-46, for which gene alleles were distributed differently among the non-coding lead guQTL genotypes (**Figure 5E-F**). In the case of IGKV2-30, the *02 allele, which harbored a missense variant in CDR1, was carried by 95.9% of individuals with genotype A/A at the lead guQTL variant, compared to only 7.9% of individuals with genotype G/G (**Figure 5E**). Likewise, in the case of IGLV7-46, the *02 allele, which harbored an amino acid change in FWR3, was carried by 100% of individuals with guQTL genotype G/G and by 0.7% of individuals with genotype C/C (**Figure 5F**).

We next asked whether shifts in gene usage also resulted in shifts in CDR3 physicochemical properties. To do this, we conducted an unbiased test for associations between variation in nine CDR3 properties and genotypes at all variants across IGK and IGL (**Supplementary Table S12**). In IGK, we identified SNVs associated (Bonferroni, $P < 3.9e\text{-}05$) with CDR3 properties of aromaticity, aliphaticity, acidity, polarity, bulk, and GRAVY index (**Figure 6A; Supplementary Figure 19**). Likewise, in IGL, SNVs were associated (Bonferroni, $P < 1.9e\text{-}05$) with CDR3 aromaticity, aliphaticity, GRAVY index, bulk, basicity, polarity, charge, and length (**Figure 6B; Supplementary Figure 20**). Lead variants in both loci overlapped guQTLs, linking CDR3 properties with gene usage variation. For example, the lead variant associated with IGK CDR3 aromaticity was also a guQTL for seven IGKV genes (**Figure 6A, 6C-E**) that are part of a previously described network clique (**Figure 3D**). Among these genes, we focused on those with usage patterns that were negatively correlated, and thus considered to differentially represent the guQTL genotypes. Specifically, IGKV1-13/1D-13 usage was highest in individuals with high CDR3 aromaticity. In contrast, the usage of IGKV2-40ambi, IGKV1-39ambi, and IGKV1D-12 were highest in individuals with low CDR3 aromaticity (**Figure 6D-E**). To determine whether these genes explain CDR3 aromaticity variation between genotype groups, we computed CDR3 aromaticity for BCRs utilizing only IGKV1-13/1D-13, IGKV2-40ambi, IGKV1-39ambi, or IGKV1D-12 (**Figure 6F**). This demonstrated IGKV1-13/1D-13-encoded BCRs had higher aromaticity than those encoded by all of the other three genes. This was regardless of the J gene contribution (**Supplementary Figure 20**), indicating the genetic effect on CDR3 aromaticity is through influence on V gene usage.

In IGL, CDR3 aliphaticity and aromaticity shared the same lead variant; this variant was also the lead guQTL for both IGLJ1 and IGLJ2-3ambi, linking genetic regulation of IGLJ gene usage with CDR3 properties (**Supplementary Figure 21**). At this variant, A/A individuals had relatively higher CDR3 aliphaticity and IGLJ2-3ambi usage, whereas G/G individuals had relatively higher CDR3 aromaticity and IGLJ1 usage. Analysis of BCRs using one or the other of these IGLJ genes revealed that sequences containing IGLJ2-3ambi have relatively high CDR3 aliphaticity, whereas sequences containing IGLJ1 have relatively high CDR3 aromaticity (**Supplementary**

Figure 21). Together, these data link genetic effects on IGLJ gene usage with IGL CDR3 properties.

In summary, 41% of IGLV and 63% of IGKV genes showed significant variation in coding alleles among lead guQTL genotype groups, indicating LD between non-coding variants and gene alleles. Additionally, we show that variation in gene usage also contributes to biases in CDR3 properties, at least in part explained by contributions of germline encoded amino acids at the 3' of V genes and 5" end of J genes. This is noteworthy, as it indicates that guQTLs not only impact general variation in gene usage, but also have the potential to modulate the availability of germline encoded residues in the baseline unmutated repertoire.

Discussion

Akin to other hypervariable immune loci, the IG loci exhibit extensive haplotype diversity at the population level, and are among the most structurally complex regions in the human genome (*12*, *14*, *15*, *33*, *42*). This complexity has limited our ability to accurately characterize inter-individual IG haplotype diversity and delineate its role in shaping the composition of the Ab repertoire (*41*). Here, by leveraging the strengths of long-read sequencing, we were able to overcome this barrier. We characterized complete genotype callsets for SNVs and SVs across the IGL and IGK loci, and combined these with matched light chain repertoires. First, this comprehensive dataset allowed for the discovery of tremendous population-level genetic variation within the IGK and IGL loci, including descriptions of novel SNVs, SVs, and coding alleles. These discoveries alone will now facilitate significant improvements in existing germline database resources for the IG loci (Peres et al., *in prep*). Second, for the first time, we were able to directly utilize personalized germline reference sets for each individual to increase the accuracy of V, D, and J gene/allele assignments, including resolution of duplicated paralogs between the proximal and distal duplicated regions of the IGK locus. Third, with these data in hand, we expanded our previous work in IGH (*12*) to demonstrate that IGK and IGL polymorphisms also contribute to variation in light chain repertoire gene usage and associated changes in the composition and availability of V, J, and CDR3 encoded amino acids among expressed Ab transcripts. Together, these findings solidify the pervasive impact of IG genetics on the adaptive immune system.

Across the IGK and IGL loci, genetic variants were statistically associated with usage variation in the majority of V and J genes. For a subset of IGKV genes, for example, we found that even single guQTLs could explain over 75% of the usage variation among donors. Notably, however, although we found that most guQTLs and associated genes were common between unmutated and mutated repertoires, the extent of variance in gene usage explained by guQTLs in the unmutated repertoire was on average higher for both IGK and IGL; V genes also had stronger genetic associations relative to J genes in both loci. The blunted genetic effects in the mutated repertoire may reflect shifts in usage in the memory repertoire driven by interactions with antigen. However, it is notable that even in antigen-experienced repertoires, variation can still be explained by genetic factors, indicating some degree of genetic constraint, consistent with observations in IGH (*12*, *16*). We also noted that on average, R2 values were lower in IGL; however, because this analysis only included an assessment of lead guQTLs, we have not accounted for additional variants that may make additional genetic contributions. We previously showed that secondary guQTLs in IGH were able to increase the variance in gene usage explained by *cis* genetic factors (*12*). As cohorts increase in size, we expect it will be possible to characterize additional guQTLs in IGK and IGL.

An assessment of guQTL positions within each locus provided initial evidence of the mechanisms by which they may exert their effects on the repertoire. Early studies in IGK were the first to show that IG polymorphism can directly impact the usage of particular genes in the repertoire. These specifically linked variation within the RSS of IGKV2D-29 and its usage frequency (*28*, *43*, *44*), demonstrating that RSS polymorphisms have the potential to influence the binding of RAG1/2 and the selection of genes by V(D)J recombination. Here, we found additional direct evidence for effects of RSS variants on several IGK and IGL genes. However, consistent with previous observations in IGH (*12*), the majority of light chain guQTLs were intergenic SNVs. The factors underlying the effects of non-coding guQTLs require further study, but are expected to influence various regulatory mechanisms (e.g., enhancer and promoter and function; formation of topologically associating domains) in the IG loci during V(D)J recombination (*45–49*). It is notable that in IGL we observed overlap between a subset of intergenic guQTLs and known TFBSs (**Supplemental Figure 12**); this included enrichments in binding sites for CTCF, a transcription factor known to be critical to the chromatin landscape within IG loci (*50*). We also found that SVs were the lead guQTLs for three genes (IGKV1D-8, IGKV1-NL1, and IGLV5-39). In all cases these SVs altered the number of diploid copies (range = 0-2) and thus the chance they could be selected by V(D)J recombination. We identified multiple less common SVs (MAF <5%) that will require study in larger sample sizes to more fully assess their contribution to gene usage. Overall, however, we noted that the number of genes impacted by SVs in each of the light chain loci were comparatively fewer than reported in IGH (*12*), which is likely a reflection of the fact that the IGH locus overall has a greater number of SVs.

We also observed examples in which guQTLs were localized to intronic and V gene coding sequences, the latter of which included examples resulting in the introduction of premature stop codons and amino acid changes. These examples indicate that, rather than effects on V(D)J recombination, some guQTLs potentially influence the light chain repertoire composition by impacting transcription and translation, with implications for light chain selection during B cell development. For example, we would expect that B cells expressing non-functional alleles would undergo receptor editing and/or negative selection in the bone marrow, leading to their absence in the periphery (*51–53*). Likewise, it is plausible that some light chain coding alleles within an individual may serve as less optimal partners for rearranged heavy chains (*54*), leading to a decrease in their frequency within the mature naive repertoire. This would be somewhat analogous to shifts in light chain gene distributions noted in the memory repertoire, which have been attributed to light chain coherence (*55*). However, fully delineating the roles of genetics in heavy and light pairing and in the context of different antigen-driven effects, will require careful investigation of guQTLs across developmental time points, and will need to consider combined effects of polymorphisms across the three loci.

To date, differences in the genetic architecture of the IGK and IGL loci have been underexplored. Previous comparisons of a small number of IGK and IGL haplotypes indicated that SNV densities were lower in IGK compared to IGL (*31*). Our data confirmed this pattern at the population level, revealing that the number of common SNVs was almost 2-fold higher in IGL. Additionally, we found that IGL was also characterized by less extensive LD. These stark differences in genetic architecture were reflected in the interconnectedness of gene usage profiles in the repertoire, specifically that a greater proportion of genes in the IGK repertoire shared overlapping guQTLs. This suggests that we could expect the regulatory landscapes to also be different. To date, our knowledge of V(D)J regulation in the light chain loci come from studies in mice; however, it is unlikely that we can extrapolate much detail from these studies, as these

loci show little structural resemblance to those in humans. Given the ordered engagement of IGK and IGL genes in the formation of functional BCRs during B cell development, it's plausible that the differences in genetic structure have been shaped by their differing functional roles. Recent work looking at IG locus genetic features across a range of vertebrate species suggest co-evolution of the light chain loci (*56*).

These data highlight the impacts of germline variation on gene usage variation in the repertoire, and their underlying mechanisms. We argue that the influence of genetics should be considered when seeking to understand how inter-individual differences in repertoire composition directly contribute to antigen-driven responses (*19–22*, *24*, *57–62*). Given coding differences between genes, usage variation by default alters the landscape of available germline encoded residues among expressed BCRs, which can also bias SHM patterns (*63*). Here, we demonstrate that guQTLs are also directly linked to amino acid differences between alleles of individual genes; thus, while some coding variants may be present within the genome of an individual, their frequency within the repertoire is dependent on guQTL genotype. Taking this a step further, we showed that gene usage also correlated with variation in CDR3 properties, driven by direct contributions of 3' V and 5' J gene germline-encoded bases to junction amino acids. This is consistent with observations of CDR3 comparisons between monozygotic twins and unrelated individuals (*17*). The link between guQTLs and amino acid features among expressed light chain transcripts elevates the likelihood that guQTLs significantly impact the antigen-binding landscape of expressed Abs.

Together, these findings advance our basic understanding of repertoire development, illuminating regions of IGK and IGL that not only regulate gene usage but establish biases in the amino acid diversity observed among expressed Abs. In combination with our previous work in the IGH locus, our data lay a foundation for integrating genetic contributions from all three IG loci to establish more complete and precise models of sequence diversity in the expressed repertoire. This will be critical for refining our understanding of Ab repertoire dynamics in health and disease.

Materials and Methods

Sample information

PBMCs (n=177) were procured from STEMCELL Technologies (Vancouver, Canada). Sample-level demographic information, including age, biological sex, and ancestry informative marker (AIM)-determined ancestry are reported in **Supplementary Table S1**.

Single-molecule real-time (SMRT) long-read library preparation and sequencing

DNA was extracted from ~3-5 million PBMCs per donor using the DNA/RNA co-extraction AllPrep kit (Qiagen, Germantown, MD, USA), and genomic DNA was processed using our published "IG-capture" targeted long-read sequencing protocol (*12–15*). Briefly, high molecular weight DNA (~2.5 μg) was sheared to ~15 Kbp using g-tubes (Covaris, Woburn, MA, USA) and size-selected using Pippin systems (Sage Science, Beverly, MA, USA) using the "high pass" protocol to select fragments greater than 5 Kbp. Size-selected DNA was ligated to universal barcoded adapters and amplified, and small fragments and excess reagents were removed using 0.7X KAPA Pure beads (Roche, Indianapolis, IN, USA). Individual samples were pooled in groups of six prior to IGK and IGL enrichment using custom Roche HyperCap DNA probes described previously (*14*, *15*). Targeted fragments were amplified after capture to increase total mass for sequencing library construction.

Enriched IGK and IGL libraries were prepared for sequencing using the SMRTBell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA) and SMRTBell Enzyme Cleanup Kit 1.0 (Pacific Biosciences), according to the manufacturer's protocol. Resulting SMRTbell libraries were multiplexed in pools of 12 and sequenced using one SMRT cell 8□M on the Sequel IIe system (n=134) (Pacific Biosciences) using 2.0 chemistry and 30-hour movies. For Revio sequencing (n=43), SMRTbell libraries were pooled in 36-plexes and sequenced using one SMRT cell 25M on the Revio system (Pacific Biosciences) using Revio Polymerase Kit v1.0 (PacBio; 102-739-100) and 30 hour movies. High Fidelity ("HiFi") intramolecular circular consensus reads with accuracies >99.9% (Q20) were generated on instrument and used for all downstream analyses.

IGK and IGL AIRR-seq

AIRR-seq libraries were prepared and sequenced for 173 individuals (IGK) 171 (IGL). RNA was extracted from ~3-5 million PBMCs per donor using the AllPrep DNA/RNA Kit (Qiagen). AIRR-seq libraries were generated using a 5' Rapid Amplification of cDNA Ends (RACE) approach. For IGK and IGL 5'RACE AIRR-seq, libraries were produced using the SMARTer Human BCR Profiling Kit (Takara Bio, San Jose, CA, USA), according to the manufacturer's instructions. Quality and quantity of individually indexed IGK libraries were determined using the 2100 Bioanalyzer High Sensitivity DNA Assay Kit (Agilent, Santa Clara, CA, USA) and IGL libraries with the Qubit 3.0 Fluorometer dsDNA High Sensitivity Assay Kit (Life Technologies, Carlsbad, CA, USA). Libraries were pooled at 10□nM and sequenced on the Illumina NextSeq system using 300□bp paired-end reads with the 600-cycle NextSeq P1 Reagent Kit (n=161) or with the 600-cycle MiSeq Reagent Kit v3 (n=11) (Illumina, San Diego, CA, USA).

Construction of a custom linear reference assembly

We used our previously described custom linear reference (*15*), which includes modifications to the IGH (*12*, *13*) and IGK (*15*) loci to include sequences not present in the GRCh38 assembly. To include IGL sequences not present in GRCh38, chromosome 22 was removed and replaced with the T2T (CHM13v2.0) chromosome 22, including the IGLV5-39 structural variant (SV, insertion) sequence. In addition, 46,423 bp of sequence (chr22:23315600-23362023) is from a Human Pangenome Reference Consortium (HPRC) haplotype from sample HG00621, which includes a 16,093 bp insertion relative to the CHM13v2.0 reference, reflecting 3 additional copies of the IGLJ-C3 cassette relative to the CHM13v2.0 reference. This reference is publicly available at https://github.com/Watson-IG/immune_receptor_genomics/tree/main.

Phased assembly of IGK and IGL

Phased assemblies were generated as described previously (*15*). HiFi reads were used to generate haplotype-phased *de novo* (i.e. reference-agnostic) assemblies using hifiasm (*64*) (v0.18.2-r467) with default parameters. For each sample, hifiasm contigs were concatenated into a FASTA file, then redundant contigs were filtered out using the seqkit toolkit (*65*) command 'seqkit rmdup --by-seq <hifiasm_contigs.fasta>' (seqkit v2.4.0). Hifiasm contigs were mapped to the custom reference assembly using minimap2 (v2.26) with the '-x asm20' option. HiFi reads were also processed using IGenotyper (*13*); the programs "phase" and "assemble" were run with default parameters to generate phased contigs and HiFi read alignments to the custom reference.

For each sample, aligned HiFi reads as well as aligned hifiasm-generated contigs and IGenotyper-generated contigs were viewed in the Integrative Genomics Viewer

(IGV) application (*66*) for manual selection of phased contigs. Contigs were evaluated for read support from mapped HiFi reads, and contigs harboring one or more SNVs that lacked read support were not selected during manual curation. Where a hifiasm and an IGenotyper contig were identical throughout a phased block, the hifiasm contig was selected, as described previously (*15*). Curated, phased assemblies were aligned to the custom reference using minimap2 (*67*) with the '-x asm20' option.

To assess accuracy of manually curated assemblies, personalized references were first generated by N-masking the IGK (chr2:88837160-90280100) and IGL (chr22:22378775-23423320) loci of the custom reference and, for each sample, appending the reference FASTA with IGK and IGL curated contigs. All HiFi reads from each individual were aligned to the corresponding personalized reference using minimap2 with the '-x map-hifi' preset; coverage and read length metrics were extracted from these alignments. Positions in assemblies with > 25% of aligned HiFi reads mismatching the assembly were identified by parsing the output of samtools 'mpileup' using custom scripts; assembly accuracy was determined using the formula [total (diploid) bases without a mismatch / total (diploid) assembly length (bp)] * 100 = % accuracy (**Supplementary Table S1**).

Identification of IGK and IGL gene alleles

Sequences corresponding to V, J, and C exons were obtained from assembly BAM files and aligned to all known human immunoglobulin alleles from the IMGT database (version downloaded on 2023-09-26) using custom Python scripts and functions provided by the receptor_utils library (https://pypi.org/project/receptor-utils/). Metrics of HiFi read support for each allele were calculated using custom scripts written in Python and bash. Briefly, for each sample, curated IGK and IGL assembly contigs were appended to our custom reference with the IGK and IGL loci N-masked to generate a personalized reference FASTA, as described above. All HiFi reads from IG-capture for a given individual were mapped to the personalized reference using minimap2 with the '-ax map-hifi' preset, then the resulting BAM file was input to samtools 'mpileup' with iteration over allele coordinates in the personalized reference. The outputs of this script are in **Supplementary Table S3** (IGK) and **Supplementary Table S4** (IGL), and include a column for total number of HiFi reads spanning the novel allele ('Fully_Spanning_Reads') and a column for the number of HiFi reads spanning the novel allele with 100% sequence identity ('Fully_Spanning_Reads_100%_Match'). A complete description of HiFi read support metrics for alleles is available at https://vdjbase.org/.

Genetic ancestry

IGenotyper (*13*) was used to call SNVs at ancestry-informative markers (AIMs) by aligning, phasing, and locally assembling reads at AIM regions, then directly identifying SNVs from the assembled sequences. Genetic ancestry was determined using these AIMs and the STRUCTURE program (*68*). SNV VCFs were processed to extract AIM-specific data from IG-Capture libraries using custom scripts and VCFtools. Coverage of AIMs was assessed using BAM files and the pysam library, ensuring a minimum read depth threshold for inclusion. Genotypes were converted into haplotypes by separating phased alleles, and samples were coded alongside reference populations from the 1000 Genomes Project. STRUCTURE (v2.3.4) was run with K = 5, representing five global ancestry groups (European, African, East Asian, South Asian, and American), using default admixture and allele frequency models. For each sample, the two highest ancestry proportions were identified. If the difference between these two proportions was less than or equal to 5%, the sample was classified as "Mixed." Otherwise, the sample was assigned to the ancestry category with the highest proportion.

Processing AIRR-seq data and calculating gene usage

Paired-end sequences ("R1" and "R2") were processed using the pRESTO/Change-O toolkit (*69*, *70*). All R1 and R2 reads were trimmed to Q□=□20 using the function "FilterSeq.py trimqual". Constant region (IGKC and IGLC) primers were identified with an error rate of 0.3 and corresponding chains were recorded in the fastq headers using "MaskPrimers align." The 12 base UMI, located directly after the constant region primer, was extracted using "MaskPrimers extract." Annotations between mate pairs, including UMI barcodes and constant region calls, were synchronized using PairSeq.py to sort reads into mate pairs and remove unpaired reads.

UMI groups sharing the same barcode were processed to generate consensus sequences using the BuildConsensus.py function. The following criteria were applied: a minimum UMI group size of one, a maximum mismatch error rate of 10%, and at least 60% agreement on the constant region call within the group. Reads with lower-quality consensus sequences (Q < 30) were masked using FilterSeq.py maskqual. Duplicate counts ("Dupcounts") were recorded, and duplicate sequences were collapsed using CollapseSeq.py to retain one representative sequence per cell, with the total number of sequences contributing to each consensus recorded as "Conscount." Collapsed consensus sequences supported by fewer than two contributing reads (Conscount < 2) were discarded using SplitSeq.py. Samples containing fewer than 100 unique sequences were excluded from downstream analysis; after filtering, 170 individuals were included for IGK analyses and 170 individuals were included for IGL analyses (173 individuals total; **Supplementary Table S1**). After processing, the repertoires contained a mean of 33,678 unique BCR sequences (IGK) and 13,281 (IGL) (**Supplementary Table S1, Supplementary Figure 3**).

Germline allele designations were assigned to sequences using a personalized allele database during the IgBLAST step. For each individual, IGK and IGL germline allele databases were generated from the set of alleles identified in genomic assemblies derived from long-read sequencing. Separate BLAST databases were created for V and J segments using makeblastdb. The resulting databases were used as input for igblastn, and IgBLAST output files ("Change-O" tables) of unique BCRs were generated for IGK and IGL separately. This process permitted theoretical disambiguation of IGK gene paralogs for individuals wherein the sequence of each allele of a proximal paralog was distinct from the sequence of each allele of the distal paralog (results in **Supplementary Table S5**). In the case of IGKV1-13 and IGKV1D-13, a subset of the cohort (n=114) met this disambiguation criteria and was carried forward to identify guQTLs for these genes.

Due to sequence identity between IGKV paralog allele sequences, genes collapsed into a single ambiguous ("ambi") entity in Change-O tables included:

1) IGKV1-37 and IGKV1D-37 replaced with IGKV1-37ambi
2) IGKV1-39 and IGKV1D-39 with IGKV1-39ambi
3) IGKV2-40 and IGKV2D-40 with IGKV2-40ambi
4) IGKV1-33 and IGKV1D-33 with IGKV1-33ambi
5) IGKV2-28 and IGKV2D-28 with IGKV2-28ambi

In addition, IGLJ gene calls that corresponded to a IGLJ2 or IGLJ3 cassette (IGLJ2, IGLJ3-1, IGLJ3-2, IGLJ3-3, IGLJ3-4) were collapsed to "IGLJ2-3ambi".

Change-O tables of unique BCRs were analyzed using the alakazam package (*70*). To enrich for antigen-naïve BCRs, only unmutated light chain sequences with 100% identity to the assigned germline V and J alleles were included in downstream analyses. For analysis of mutated sequences (i.e. the fraction of sequences enriched with those containing SHM), sequences with less than 100% identity to either the germline V or J allele were included. Gene usage was quantified for IGK and IGL light

chains using the countGenes function with the parameters gene = "v_call" and "j_call", groups = "sample_id", mode = "gene", and genes with a sequence count (seq_count) of at least 10 in at least one sample were retained. A m□×□n usage matrix C was created, where m are the genes and n are the samples. Each value in C represented usage frequency among all unique (unmutated or mutated) BCR sequences for a given gene in a given sample.

CDR3 physicochemical properties were computed using the aminoAcidProperties function (47) with options seq = "junction", trim = TRUE, label = "cdr3"; resulting values were then averaged across all sequences within each sample to obtain sample-level means.

Selecting common variants for gene usage QTL analysis

SNVs were genotyped from curated assemblies as described previously (*15*). Variants were called from assemblies using 'bcftools mpileup' (bcftools v1.15.1) with options '-f -B -a QS', then 'bcftools call' with options '-m --ploidy 2'. A mutli-sample VCF file was generated using 'bcftools merge' with the '-m both' option. Multiallelic SNVs were split into biallelic records using the 'bcftools norm' command with options '-a -m-'. The VCF file was annotated for V-exon, introns, L-Part1, RSS sequences (heptamer, nonamer, spacer), were added using vcfanno and BED files corresponding to our reference (available at https://github.com/Watson-IG/immune_receptor_genomics/). Biallelic SNVs with MAF ≥ 5% were selected using bcftools view with options "-m2 -M2 -v snps -i 'INFO/MAF >= 0.05'" and used for guQTL analysis.

All SVs in IGKV and IGLV gene regions were genotyped by manual inspection using IGV (*66*). SVs with a MAF less than 0.05 were not included in the guQTL analysis.

Gene usage QTL analysis

Genotypes at common SNVs and SVs were tested for association with usage using linear regression to determine significance and additional metrics (e.g., beta coefficients and R2 values). To adjust for multiple comparisons, a Bonferroni correction was applied on a per-gene basis. Pairwise r^2 (LD) values were computed using vcftools '--geno-r2' (*71*). Variants in complete linkage disequilibrium (LD r^2=1) were considered as a single variant during correction, representing only one association test.

Haplotype block analysis

LD blocks were computed and visualized using "LDBlockShow" (*39*) using the multi-sample VCF of common SNVs (MAF >= 5%) as input, with options '-SeleVar 2 -BlockType 1' to use normalized linkage disequilibrium coefficients (D') as described by described by Gabriel et al. (2002) (*40*) to determine haplotype blocks. The program was run using '-Region chr2:88837160-90280100' for IGK and '-Region chr22:22378775-23423320' for IGL. LD block boundaries, sizes, and SNVs within blocks are included in **Supplementary Table S9**. Genes overlapping LD blocks were identified using bedtools (*72*) 'intersect' (bedtools v2.30.0).

Network analysis

Variants significantly associated with IGK and IGL gene usage (after Bonferroni correction) were compiled to create a set of guQTL variants for each gene. Pairwise comparisons between genes were performed to calculate Jaccard similarity indices, defined as the ratio of shared variants to the total unique variants across the two genes being compared. Only gene pairs with nonzero similarity (i.e., at least one shared

variant) were included in the analysis. Separate IGK and IGL network graphs were constructed and visualized using igraph (*73*) and ggraph (*74*) (IGL),. In these graphs, nodes represented individual genes and were positioned using the Fruchterman-Reingold force-directed algorithm, edges connected pairs of genes that shared guQTL variants, and edges were labeled to reflect the number of shared guQTL variants.

Regulatory analysis

ENCODE transcription factor binding site data were obtained from the UCSC Genome Browser under the group "Regulation," track "TF Clusters," and table "encRegTfbsClustered." SNVs associated with gene usage were overlapped with this track and enrichment over all SNVs overlapping each track was calculated using a one-sided Fisher Exact Test (**Supplementary Table S8**).

Figure Legends and Supplementary Figure Legends

Figure Legends

Figure 1. IGK and IGL variants impact gene usage in the naïve Ab repertoire

(A) General structure of V and J genes in the IGK and IGL loci, including location of the recombination signal sequences (RSS). (B-C) Per gene (x axis, all panels) statistics from linear regression guQTL analysis for the repertoire of unmutated IGK (B) and IGL (C) light chains, including: (i) the number of associated variants after Bonferroni correction (IGK; P < 3.7e-5, IGL; P < 1.9e-5), (ii) −log10(P value) of the lead guQTL, (iii) adjusted R2 for variance in gene usage explained by the lead guQTL, (iv) the location and (v) type of variant for the lead guQTL and (vi) the fold change in gene usage between genotypes at the lead guQTL. Summary statistics are provided in Supplementary Table S6.

Figure 2. Examples of coding and non-coding lead guQTLs

(A) Manhattan plot showing the −log10(P value) for all SNVs in IGK tested for association with usage of IGKV2-29, with SNVs colored according to LD (r2) with the lead variant (marked with an X). (B) Sequence alignment of the germline IGKV2-29 alleles in this cohort from codons 90 to 95, with the lead variant indicated. Alleles encoding C93 and X93 (STOP codon) alleles are indicated. (C) Boxplot of IGKV2-29 usage in lead guQTL genotype groups. (D) Manhattan plot of associations (−log10(P value)) between all IGK SNVs and usage of IGKV1-5, with SNVs colored according to LD (r2) with the lead variant. (E) Sequence alignment of the reference and alternate haplotypes at the lead guQTL, with two missense variants in perfect LD in codon 50 indicated, resulting in K50D in the alternate haplotype. (F) Boxplot of IGKV1-5 usage in lead guQTL (shown in (E)) genotype groups. (G) Alignment of translated germline IGKV1-5 alleles with codon 50 boxed. (H) Manhattan plot of associations (−log10(P value)) between all IGL SNVs and usage of IGLV3-16, with SNVs colored according to LD (r2) with the lead variant. Two lead variants in perfect LD are in the RSS spacer. (I) Sequence of the RSS spacer in reference and alternate lead guQTL haplotypes. (J) Boxplot of IGLV3-16 usage in lead guQTL genotype groups. (K) (Top) Manhattan plot of associations (−log10(P value)) between all IGL SNVs and usage of IGLV9-49, with SNVs colored according to LD (r2) with the lead variant. (Bottom) Zoom-in on an 8 Kbp window centered on IGLV9-49 with the lead non-coding variant indicated. (L) Boxplot of IGLV9-49 usage in lead guQTL genotype groups. (M) Gene usage boxplots of genes for

which the lead variant was a deletion ("DEL") SV, including IGKV1-NL1, IGKV1D-8, and IGLV5-39.

Figure 3. Genetic coordination of IG light chain gene usage is more prevalent in IGK relative to IGL

(A) Stacked bar plot showing the proportion of total IGK and IGL common SNVs that are a guQTL. (B) Bar plot showing the number of IGK and IGL SNVs (guQTLs) significantly associated with varying numbers of genes (n = 1–9). For IGK, this includes a large number of SNVs (n=2,049) that were associated with >1 gene. (C) For each gene, the number of genes sharing at least one guQTL variant is plotted for indicated IGK (left) and IGL (right) genes (x-axis). (D-E) Network analysis identified a large clique of genes and guQTLs in IGK (D) and 4 cliques for IGL (E), demarcating groups of genes associated with overlapping sets of guQTLs. For each clique, genes are shown as nodes, connected by edges color coded according to the number of shared guQTL variants.

Figure 4. IGK has larger LD blocks and lower density of SNVs relative to IGL

(A-B) LD heatmaps of the IGK (A) and IGL (B) loci. LD blocks are illustrated as triangles. (C) Stacked bar plot of the percent of each locus (IGK, IGL) that is within LD blocks of various lengths (colors). (D) Plots of LD blocks in IGK and IGL depicting the length of each block (y-axis) and number of SNVs in each blot (x-axis). (E) Bar plot of the overall SNV density in the IGK and IGL loci. (F-G) Barplots of the counts of IGK (F) or IGL (G) genes in LD blocks with lengths indicated along y-axes.

Figure 5. Linkage between IGKV and IGLV coding region alleles and lead guQTL genotypes

(A-B) Variation in the proportion of different coding gene alleles among lead guQTL genotype groups was determined by Fisher's exact test for guQTL genes in IGK (A) and IGL (B). Barplots shows -log10(P value). (C-D) For each gene, the frequency of coding alleles in the cohort is shown, with unique alleles color coded. Genes that lack appreciable allelic variation (major allele frequency >95%) are indicated with an asterisk. Circles above each gene indicate whether coding allele variation is linked to the lead guQTL. guQTLs linked to coding allele variation are associated with missense or nonsense variants. (E-F) Stacked bar plots showing the distributions of the respective coding allele genotypes across individuals partitioned by guQTL genotype for IGKV2-30 (E) and IGLV7-46 (F).

Figure 6. IGK and IGL variants impact CDR3 physicochemical properties in the naïve Ab repertoire

(A-B) For each CDR3 physicochemical property (x-axis), mean values were computed for each individual and tested for association (linear regression) with all common variants in IGK (A) and IGL (B). Barplots show (i) the number of QTL variants (Bonferroni-corrected) for each property, (ii) the -log10(P value) for lead variants, and (iii) the number of guQTL genes identified for the lead CDR3 property QTL variant. Summary statistics are provided in Supplementary Table S12. (C) Manhattan plot shows the −log10(P value) for all SNVs in the IGK locus tested for association with CDR3 aromaticity, with QTLs colored dark red and the lead QTL labelled. (D) Boxplot of the mean IGK CDR3 aromaticity with individuals separated by genotype at the lead QTL. (E) Boxplots of usages for seven IGK genes that are guQTLs at the lead CDR3 aromaticity variant. (F) BCR sequences that used the indicated V genes were selected from the Ab

repertoire, then mean CDR3 aromaticity of each repertoire subset was computed and plotted with individuals separated by genotype at the lead CDR3 aromaticity QTL.

Supplementary Figure Legends

Supplementary Figure 1. PacBio sequencing and assembly statistics.
For both IGK and IGL, the statistics described include (a) locus coverage, (b) read length, (c) assembly lengths, and (d) contig lengths. Each statistic is separated by PacBio platforms. The number of sequencing runs on the Sequel IIe and Revio platforms is 134 and 43, respectively. Boxplots display the median,
25th percentile, 75th percentile, and whiskers that extend up to 1.5 times the inter-quartile range
(IQR) from the respective percentiles. Any data points outside the whiskers are also plotted.

Supplementary Figure 2. IGK and IGL genetic variation identified by long-read sequencing in a cohort of 172 individuals.
(A) Diagram of genes and structural polymorphisms in the human IGK (A) and IGL (B) loci. SVs include deletions, an inversion, a CNV region in IGL (Gibson et al., 2023), as well as a region of high SNV density that likely resulted from a gene conversion event in which at least ~16 Kbp of sequence containing IGKV1-12 and IGKV1-13 replaced paralogous sequence in the distal region, described previously (Watson et al., 2015; Engelbrecht et al., 2024). (C) Barplot of the number of common SNVs (MAF ≥ 5%) identified in this cohort. (D-E) Pie charts indicate the proportion of common SNVs present and absent in dbSNP. (F-G) Stacked bar plots of the number of unique novel and previously curated (in IMGT) alleles for IGKV and IGLV genes. (H-I) Number of samples (y-axis) wherein each novel allele (x-axis, colored by substitution type) was identified. Novel alleles were first filtered to include only those with >10 supporting HiFi reads. (NS: non-synonymous, S: synonymous, STOP: premature stop codon (nonsense)).

Supplementary Figure 3. Number of merged reads for each AIRR-seq dataset after processing.
Number of unique V-J sequences per-sample after filtering duplicate reads for IGK (n=173) and IGL (n=171) AIRR-seq. Boxplots display the median, 25th percentile, 75th percentile, and whiskers that extend up to 1.5 times the inter-quartile range (IQR) from the respective percentiles. Any data points outside the whiskers are also plotted.

Supplementary Figure 4. Identification of putative IGKV distal region deletion by coverage analysis
(A) Ratio of proximal:distal region read coverage is shown. If IGKV distal region gDNA was absent from the sample due to deletion, we anticipate this would be reflected in the computed ratio as ~2-fold more reads mapping to the proximal region as compared to the distal region. The dashed black line is at y=1.5. (B) IGV screenshot of hifiasm contigs for sample CE0008136 aligned to the IGK locus; no contigs map to the IGKV distal region.

Supplementary Figure 5. Common SNVs in IGK and IGL in genic and RSS regions.
Number of common SNVs in IGK and IGL partitioned according to location in genic and RSS (heptamer, spacer, nonamer) regions.

Supplementary Figure 6. Many IGK and IGL gene alleles are not catalogued in IMGT.
Pie charts indicate the proportion of IGK and IGL gene alleles identified in our cohort that are novel (not cataloged in IMGT).

Supplementary Figure 7. guQTLs in naïve (unmutated) and antigen-experienced (mutated) Ab repertoires.
(A-B) Barplots show the strength of associations for lead guQTLs in the unmutated (naïve) and mutated (antigen-experienced) IGK (A) and IGL (B) repertoires. Plots include only guQTL genes. (C) Venn diagrams indicating the number of unique and overlapping guQTL genes in unmutated and mutated IGK and IGL Ab repertoires (see also Supplementary Table S7). (D) The variance explained (adjusted R2) for guQTL genes identified in both unmutated and mutated Ab repertoires for IGK (14 genes) and IGL (19 genes). Gray lines connect the variance explained in the unmutated and mutated repertoires.

Supplementary Figure 8. guQTLs in naïve (unmutated) and antigen-experienced (mutated) Ab repertoires.
(A-B) Per gene (x axis, all panels) statistics from linear regression guQTL analysis for the repertoire of mutated IGK (A) and IGL (B) light chains, including: (i) the number of associated variants after Bonferroni correction (IGK; $P < 3.7e-5$, IGL; $P < 1.9e-5$), (ii) $-\log10$(P value) of the lead guQTL, (iii) adjusted R2 for variance in gene usage explained by the lead guQTL, (iv) the location and (v) type of variant for the lead guQTL and (vi) the fold change in gene usage between genotypes at the lead guQTL. Summary statistics are provided in Supplementary Table S6.

Supplementary Figure 9. Usage of IGK and IGL guQTL and non-guQTL genes in unmutated and mutated Ab repertoires.
Each panel shows usage of V or J genes in IGK or IGL. guQTL genes in unmutated and mutated repertoires are annotated “U” and “M”, respectively, with “U,M” indicating guQTL genes in both unmutated and mutated repertoires.

Supplementary Figure 10. Individuals sharing a greater number of guQTL genotypes have more correlated repertoire-wide light chain gene usage profiles.
Repertoire-wide gene usage correlations between all individuals in the cohort (pairwise) were calculated using the Pearson’s Correlation coefficient (y-axis). Pairs of samples are separated according to allele sharing distance (ASD), calculated using all guQTL variants in the respective locus. Boxplots plots show pairwise IGK and IGL repertoire-wide gene usage correlations partitioned by ASD. Boxplots display the median, 25th percentile, 75th percentile, and whiskers that extend up to 1.5 times the inter-quartile range (IQR) from the respective percentiles. Data points outside the whiskers are also plotted.

Supplementary Figure 11. IGLV RSS consensus.
IGLV RSS consensus sequence logo computed using all unique IGLV RSSs in our cohort.

Supplementary Figure 12. IGL guQTLs are enriched with ENCODE3 TFBS.
(A) Bar plot showing the fraction of lead IGL guQTL SNVs that overlapped ENCODE3 TFBS, compared to the overlap observed for the non-guQTL set of variants used in the

IGL guQTL analysis. TFBS for which statistically significant enrichments were observed are indicated by asterisks: One-side Fisher's Exact Test; *P value < 0.05; **P value < 0.005; ***P value < 0.0005; ****P value < 0.00005 (see also Supplementary Table S8). (B) IGV screenshot with annotations for genes, guQTLs, and ENCODE3 TF binding regions. The lead guQTL for IGLV2-11 overlaps multiple TFBS. The lead guQTL for 3 IGLV genes (IGLV3-10, IGLV3-9, IGLV2-8) also overlaps multiple TFBS. (C) Boxplots of usages of IGLV3-10, IGLV3-9, IGLV2-8 with individuals separated by genotype at the lead guQTL variant (indicated in (B) for these three genes.

Supplementary Figure 13. The lead IGLV9-49 guQTL overlaps multiple TFBS.
(A) IGV screenshot with annotations for genes, guQTLs, and ENCODE3 TF binding regions. The lead guQTL for IGLV9-49 overlaps multiple TFBS.

Supplementary Figure 14. Analysis of IGKV gene paralog usage.
(A) Usage of indicated IGKV gene paralogs within individual (data points), with blue and orange dashed lines indicating higher usage of the proximal and distal paralog, respectively. Differences between proximal and distal gene usage were determined by paired t-tests. (B) Alignment of all RSSs for IGKV1-13 and IGKV1D-13 alleles in our cohort. (C) The frequency of each IGKV1-13 and IGKV1D-13 allele in our cohort, with alleles colored according to having a canonical or non-canonical RSS heptamer. (D) The frequency of IGKV1D-13 alleles in lead guQTL IGKV1D-13 genotype groups. (E) Boxplot of IGKV1D-13 usage with individuals separated according to genotype at the lead IGKV1D-13 guQTL. (F) Boxplot of IGKV2-29 (proximal) and IGKV2D-29 (distal) gene usage in individuals, with blue and orange dashed lines indicating higher usage of the proximal and distal paralog, respectively. Individuals are grouped (columns) according to genotype at the lead IGKV2-29 guQTL (discussed in Figure 2A-C).

Supplementary Figure 15. IGKV1-13 and IGKV1D-13 germline alleles.
(A) Alignment of translated V-regions of IGKV1-13 and IGKV1D-13 alleles. (B-C) Frequency of IGKV1-13 (B) and IGKV1D-13 (C) alleles among individuals in the cohort.

Supplementary Figure 16. Usage of IGLJ1 and IGLJ2-3ambi, with individuals grouped according to number of diploid copies of the IGLJ2-3 cassette. P values from indicated tests are shown.

Supplementary Figure 17. LD blocks in IGK.
(Top) Heat corresponds to pairwise correlations between SNVs; LD blocks (black triangles) were computed using LDBlockShow (Dong et al., 2021). (Bottom) LD blocks are plotted along the x-axis from largest to smallest (left to right), and the count of the number of genes within each block is shown. Arrows are drawn for the 3 largest LD blocks.

Supplementary Figure 18. LD blocks in IGK.
(Top) Heat corresponds to pairwise correlations between SNPs; LD blocks (black triangles) were computed using LDBlockShow (Dong et al., 2021). (Bottom) LD blocks are plotted along the x-axis from largest to smallest (left to right), and the count of the number of genes within each block is shown. Genes within the 3 largest LD blocks are indicated.

Supplementary Figure 19. Associations between CDR3 physicochemical properties and germline variants in IGK and IGL.

Manhattan plots show the −log10(P value) for all SNVs in IGK (top) or IGL (bottom) tested for association with indicated CDR3 physicochemical properties in naïve (unmutated) Ab repertoires. Dashed lines indicate Bonferroni-corrected significance (IGK; P < 3.7e-5, IGL; P < 1.9e-5).

Supplementary Figure 20. CDR3 aromaticity of BCR sequences composed of specific IGKV and IGKJ genes.
Boxplot of mean CDR3 aromaticity (per-individual) of unmutated IGK BCR sequences comprised of indicated IGKV and IGKJ gene pairs.

Supplementary Figure 21. Genetic effects on IGLJ gene usages are associated with CDR3 aliphaticity and aromaticity.
(A-B) Manhattan plots show the −log10(P value) for all SNVs in the IGL locus tested for association with CDR3 aliphaticity (A) or aromaticity (B), with QTLs colored dark red and the lead QTL labelled. These two CDR3 properties share a lead QTL variant (labelled). (C-D) Boxplots of the mean IGL CDR3 aliphaticity (C) and aromaticity (D) with individuals separated by genotype at the lead QTL. (E) Boxplot of IGLJ1 and IGLJ2-3ambi usages with individuals separated by genotype at the lead guQTL. (F-G) BCR sequences that used IGLJ1 or IGLJ2-ambi were selected from the Ab repertoire, then mean CDR3 aromaticity of each repertoire subset was computed and plotted with individuals separated by genotype at the lead variant.

Funding
This work was supported by grant R24AI138963 to C.T.W and M.L.S. from the National Institute of Allergy and Infectious Disease.

Author Contributions
E.E., O.L.R., M.L.S., and C.T.W. contributed to the conception of the work. M.L.S. and C.T.W. directed wet lab/bench and sequencing experiments. E.E., D.S., K.S., S.S. prepared sequencing libraries. E.E., U.J., O.L.R., W.L., W.S.G., A.P., G.Y. and Z.V. developed the computational pipeline to analyze the sequencing data. E.E., G.Y., C.T.W., M.L.S contributed to and supervised the analysis and data interpretation. C.T.W. and M.L.S. acquired funding. E.E. and C.T.W. drafted the manuscript and all authors contributed to revising the manuscript.

Competing interests
C.T.W., M.L.S., and W.L. are founders and shareholders of Clareo Biosciences, Inc. and serve on its Executive Board.

References

1. B. Briney, A. Inderbitzin, C. Joyce, D. R. Burton, Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).
2. C. Soto, R. G. Bombardi, A. Branchizio, N. Kose, P. Matta, A. M. Sevy, R. S. Sinkovits, P. Gilchuk, J. A. Finn, J. E. Crowe Jr, High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566**, 398–402 (2019).
3. S. D. Boyd, B. A. Gaëta, K. J. Jackson, A. Z. Fire, E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, K. C. Nadeau, M. Egholm, D. B. Miklos, J. L. Zehnder, A. M. Collins, Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* **184**, 6986–6992 (2010).
4. K. Röltgen, A. E. Powell, O. F. Wirz, B. A. Stevens, C. A. Hogan, J. Najeeb, M. Hunter, H. Wang, M. K. Sahoo, C. Huang, F. Yamamoto, M. Manohar, J. Manalac, A. R. Otrelo-Cardoso, T. D. Pham, A. Rustagi, A. J. Rogers, N. H. Shah, C. A. Blish, J. R. Cochran, T. S. Jardetzky, J. L. Zehnder, T. T. Wang, B. Narasimhan, S. Gombar, R. Tibshirani, K. C. Nadeau, P. S. Kim, B. A. Pinsky, S. D. Boyd, Defining the features and duration of antibody responses to SARS-CoV-2 infection associated with disease severity and outcome. *Sci. Immunol.* **5**, eabe0240 (2020).
5. W. M. P. B. Wahala, A. M. de Silva, The human antibody response to dengue virus infection. *Viruses* **3**, 2374–2395 (2011).
6. J. Overbaugh, L. Morris, The antibody response against HIV-1. *Cold Spring Harb. Perspect. Med.* **2**, a007039 (2012).
7. F. Krammer, The human antibody response to influenza A virus infection and

vaccination. *Nat. Rev. Immunol.* **19**, 383–397 (2019).
8. N. Muñoz-Durango, M. S. Pizarro-Ortega, E. Rey-Jurado, F. E. Díaz, S. M. Bueno, A. M. Kalergis, Patterns of antibody response during natural hRSV infection: insights for the development of new antibody-based therapies. *Expert Opin. Investig. Drugs* **27**, 721–731 (2018).
9. L. Del Pozo-Yauner, G. A. Herrera, J. I. Perez Carreon, E. A. Turbat-Herrera, F. J. Rodriguez-Alvarez, R. A. Ruiz Zamora, Role of the mechanisms for antibody repertoire diversification in monoclonal light chain deposition disorders: when a friend becomes foe. *Front. Immunol.* **14**, 1203425 (2023).
10. G. V. Sharonov, E. O. Serebrovskaya, D. V. Yuzhakova, O. V. Britanova, D. M. Chudakov, B cells, plasma cells and antibody repertoires in the tumour microenvironment. *Nat. Rev. Immunol.* **20**, 294–307 (2020).
11. K. Murphy, C. Weaver, *Janeway's Immunobiology* (Garland Science, 2016).
12. O. L. Rodriguez, Y. Safonova, C. A. Silver, K. Shields, W. S. Gibson, J. T. Kos, D. Tieri, H. Ke, K. J. L. Jackson, S. D. Boyd, M. L. Smith, W. A. Marasco, C. T. Watson, Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat. Commun.* **14**, 4419 (2023).
13. O. L. Rodriguez, W. S. Gibson, T. Parks, M. Emery, J. Powell, M. Strahl, G. Deikus, K. Auckland, E. E. Eichler, W. A. Marasco, R. Sebra, A. J. Sharp, M. L. Smith, A. Bashir, C. T. Watson, A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. *Front. Immunol.* **11**, 2136 (2020).
14. W. S. Gibson, O. L. Rodriguez, K. Shields, C. A. Silver, A. Dorgham, M. Emery, G. Deikus, R. Sebra, E. E. Eichler, A. Bashir, M. L. Smith, C. T. Watson, Characterization of the immunoglobulin lambda chain locus from diverse populations reveals extensive genetic variation. *Genes Immun.* **24**, 21–31 (2023).
15. E. Engelbrecht, O. L. Rodriguez, K. Shields, S. Schultze, D. Tieri, U. Jana, G. Yaari, W. D. Lees, M. L. Smith, C. T. Watson, Resolving haplotype variation and complex genetic architecture in the human immunoglobulin kappa chain locus in individuals of diverse ancestry. *Genes Immun.*, doi: 10.1038/s41435-024-00279-2 (2024).
16. J. Glanville, T. C. Kuo, H.-C. von Büdingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, G. R. Mehta, J. R. Oksenberg, S. L. Hauser, D. R. Cox, A. Rajpal, J. Pons, Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20066–20071 (2011).
17. F. Rubelt, C. R. Bolen, H. M. McGuire, J. A. Vander Heiden, D. Gadala-Maria, M. Levin, G. M. Euskirchen, M. R. Mamedov, G. E. Swan, C. L. Dekker, L. G. Cowell, S. H. Kleinstein, M. M. Davis, Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat. Commun.* **7**, 11112 (2016).
18. C. Wang, Y. Liu, M. M. Cavanagh, S. Le Saux, Q. Qi, K. M. Roskin, T. J. Looney, J.-Y. Lee, V. Dixit, C. L. Dekker, G. E. Swan, J. J. Goronzy, S. D. Boyd, B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 500–505 (2015).
19. T. Parks, M. M. Mirabel, J. Kado, K. Auckland, J. Nowak, A. Rautanen, A. J. Mentzer, E. Marijon, X. Jouven, M. L. Perman, T. Cua, J. K. Kauwe, J. B. Allen, H. Taylor, K. J. Robson, C. M. Deane, A. C. Steer, A. V. S. Hill, Pacific Islands Rheumatic Heart Disease Genetics Network, Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nat. Commun.* **8**, 14946 (2017).
20. P. Pushparaj, A. Nicoletto, D. J. Sheward, H. Das, X. Castro Dopico, L. Perez Vidakovics, L. Hanke, M. Chernyshev, S. Narang, S. Kim, J. Fischbach, S. Ekström, G. McInerney, B. M. Hällberg, B. Murrell, M. Corcoran, G. B. Karlsson Hedestam,

Immunoglobulin germline gene polymorphisms influence the function of SARS-CoV-2 neutralizing antibodies. *Immunity* **56**, 193–206.e7 (2023).

21. Y. Avnir, C. T. Watson, J. Glanville, E. C. Peterson, A. S. Tallarico, A. S. Bennett, K. Qin, Y. Fu, C.-Y. Huang, J. H. Beigel, F. Breden, Q. Zhu, W. A. Marasco, IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci. Rep.* **6**, 20842 (2016).
22. M. Sangesland, A. Torrents de la Peña, S. Boyoglu-Barnum, L. Ronsard, F. A. N. Mohamed, T. B. Moreno, R. M. Barnes, D. Rohrer, N. Lonberg, M. Ghebremichael, M. Kanekiyo, A. Ward, D. Lingwood, Allelic polymorphism controls autoreactivity and vaccine elicitation of human broadly neutralizing antibodies against influenza virus. *Immunity* **55**, 1693–1709.e8 (2022).
23. Y. Mashimo, K. Yamazaki, T. Kageyama, S. Tanaka, T. Taniguchi, K. Matsushita, H. Igari, H. Hanaoka, K. Yokote, H. Nakajima, Y. Onouchi, Germline variants of IGHV3-53 / V3-66 are determinants of antibody responses to the BNT162b2 mRNA COVID-19 vaccine. *J. Infect.* **85**, 702–769 (2022).
24. J. H. Lee, L. Toy, J. T. Kos, Y. Safonova, W. R. Schief, C. Havenar-Daughton, C. T. Watson, S. Crotty, Vaccine genetics of IGHV1-2 VRC01-class broadly neutralizing antibody precursor naïve human B cells. *NPJ Vaccines* **6**, 113 (2021).
25. K. Sonehara, Y. Uwamino, R. Saiki, M. Takeshita, S. Namba, S. Uno, T. Nakanishi, T. Nishimura, T. Naito, G. Sato, M. Kanai, A. Liu, S. Uchida, T. Kurafuji, A. Tanabe, T. Arai, A. Ohno, A. Shibata, S. Tanaka, M. Wakui, S. Kashimura, C. Tomi, A. Hara, S. Yoshikawa, K. Gotanda, K. Misawa, H. Tanaka, S. Azekawa, Q. S. Wang, R. Edahiro, Y. Shirai, K. Yamamoto, G. Nagao, T. Suzuki, M. Kiyoshi, A. Ishii-Watabe, S. Higashiue, S. Kobayashi, H. Yamaguchi, Y. Okazaki, N. Matsumoto, A. Masumoto, H. Koga, A. Kanai, Japan COVID-19 Task Force, Biobank Japan Project, Y. Oda, Y. Suzuki, K. Matsuda, Y. Kitagawa, R. Koike, A. Kimura, A. Kumanogoh, A. Yoshimura, S. Imoto, S. Miyano, T. Kanai, K. Fukunaga, N. Hasegawa, M. Murata, H. Matsushita, S. Ogawa, Y. Okada, H. Namkoong, Germline variants and mosaic chromosomal alterations affect COVID-19 vaccine immunogenicity. *Cell Genom.*, 100783 (2025).
26. A. C. deCamp, M. M. Corcoran, W. J. Fulp, J. R. Willis, C. A. Cottrell, D. L. V. Bader, O. Kalyuzhniy, D. J. Leggat, K. W. Cohen, O. Hyrien, S. Menis, G. Finak, L. Ballweber-Fleming, A. Srikanth, J. R. Plyler, F. Rahaman, A. Lombardo, V. Philiponis, R. E. Whaley, A. Seese, J. Brand, A. M. Ruppel, W. Hoyland, C. R. Mahoney, A. Cagigi, A. Taylor, D. M. Brown, D. R. Ambrozak, T. Sincomb, T.-M. Mullen, J. Maenza, O. Kolokythas, N. Khati, J. Bethony, M. Roederer, D. Diemert, R. A. Koup, D. S. Laufer, J. M. McElrath, A. B. McDermott, G. B. Karlsson Hedestam, W. R. Schief, Human immunoglobulin gene allelic variation impacts germline-targeting vaccine priming. *NPJ Vaccines* **9**, 58 (2024).
27. I. Mikocziova, A. Peres, M. Gidoni, V. Greiff, G. Yaari, L. M. Sollid, Germline polymorphisms and alternative splicing of human immunoglobulin light chain genes. *iScience* **24**, 103192 (2021).
28. A. J. Feeney, M. J. Atkinson, M. J. Cowan, G. Escuro, G. Lugo, A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to haemophilus influenzae type b disease. *J. Clin. Invest.* **97**, 2277–2282 (1996).
29. E. L. Shrock, R. T. Timms, T. Kula, E. L. Mena, A. P. West Jr, R. Guo, I.-H. Lee, A. A. Cohen, L. G. A. McKay, C. Bi, Y. Leng, E. Fujimura, F. Horns, M. Li, D. R. Wesemann, A. Griffiths, B. E. Gewurz, P. J. Bjorkman, S. J. Elledge, Germline-encoded amino acid-binding motifs drive immunodominant public antibody responses. *Science* **380**, eadc9498 (2023).
30. O. L. Rodriguez, X. Qiu, K. Shields, C. Dunn, A. Singh, M. Kaileh, C. T. Watson, R. Sen, Human genetic variation shapes the antibody repertoire across B cell development,

Genomics (2025). https://www.biorxiv.org/content/10.1101/2025.05.19.654982v1.

31. C. T. Watson, K. M. Steinberg, T. A. Graves, R. L. Warren, M. Malig, J. Schein, R. K. Wilson, R. A. Holt, E. E. Eichler, F. Breden, Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. *Genes Immun.* **16**, 24–34 (2015).
32. C. T. Watson, K. M. Steinberg, J. Huddleston, R. L. Warren, M. Malig, J. Schein, A. J. Willsey, J. B. Joy, J. K. Scott, T. A. Graves, R. K. Wilson, R. A. Holt, E. E. Eichler, F. Breden, 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.* **92**, 530–546 (2013).
33. U. Jana, O. L. Rodriguez, W. Lees, E. Engelbrecht, Z. Vanwinkle, A. Peres, W. Gibson, K. Shields, S. Schultze, A. G. Dorgham, M. Emery, G. Deikus, R. Sebra, E. E. Eichler, G. Yaari, M. L. Smith, C. T. Watson, The human immunoglobulin heavy chain constant gene locus is enriched for large complex structural variants and coding polymorphisms that vary in frequency among human populations, *Genomics* (2025). https://www.biorxiv.org/content/10.1101/2025.02.12.634878v1.
34. E. T. Engelbrecht, O. L. Rodriguez, C. T. Watson, Addressing technical pitfalls in pursuit of molecular factors that mediate immunoglobulin gene regulation. *bioRxiv*, 2024–2003 (2024).
35. GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
36. Y. Ye, Z. Zhang, Y. Liu, L. Diao, L. Han, A multi-omics perspective of quantitative trait loci in precision medicine. *Trends Genet.* **36**, 318–336 (2020).
37. K. Yamaguchi, K. Ishigaki, A. Suzuki, Y. Tsuchida, H. Tsuchiya, S. Sumitomo, Y. Nagafuchi, F. Miya, T. Tsunoda, H. Shoda, K. Fujio, K. Yamamoto, Y. Kochi, Splicing QTL analysis focusing on coding sequences reveals mechanisms for disease susceptibility loci. *Nat. Commun.* **13**, 4659 (2022).
38. M. van der Wijst, D. H. de Vries, H. E. Groot, G. Trynka, C. C. Hon, M. J. Bonder, O. Stegle, M. C. Nawijn, Y. Idaghdour, P. van der Harst, C. J. Ye, J. Powell, F. J. Theis, A. Mahfouz, M. Heinig, L. Franke, The single-cell eQTLGen consortium. *Elife* **9** (2020).
39. S.-S. Dong, W.-M. He, J.-J. Ji, C. Zhang, Y. Guo, T.-L. Yang, LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Brief. Bioinform.* **22** (2021).
40. S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, D. Altshuler, The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
41. C. T. Watson, J. Glanville, W. A. Marasco, The Individual and Population Genetics of Antibody Immunity. *Trends Immunol.* **38**, 459–470 (2017).
42. M. Gidoni, O. Snir, A. Peres, P. Polak, I. Lindeman, I. Mikocziova, V. K. Sarna, K. E. A. Lundin, C. Clouser, F. Vigneault, A. M. Collins, L. M. Sollid, G. Yaari, Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nat. Commun.* **10**, 628 (2019).
43. M. J. Atkinson, M. J. Cowan, A. J. Feeney, New alleles of IGKV genes A2 and A18 suggest significant human IGKV locus polymorphism. *Immunogenetics* **44**, 115–120 (1996).
44. B. Nadel, A. Tang, G. Lugo, V. Love, G. Escuro, A. J. Feeney, 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 influenzae type b disease. *J. Immunol.* **161**, 6068–6073 (1998).

45. N. M. Choi, S. Loguercio, J. Verma-Gaur, S. C. Degner, A. Torkamani, A. I. Su, E. M. Oltz, M. Artyomov, A. J. Feeney, Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom V gene rearrangement frequencies. *J. Immunol.* **191**, 2393–2402 (2013).
46. D. J. Bolland, H. Koohy, A. L. Wood, L. S. Matheson, F. Krueger, M. J. T. Stubbington, A. Baizan-Edge, P. Chovanec, B. A. Stubbs, K. Tabbada, S. R. Andrews, M. Spivakov, A. E. Corcoran, Two mutually exclusive local chromatin states drive efficient V(D)J recombination. *Cell Rep.* **15**, 2475–2487 (2016).
47. K. H. Bhat, S. Priyadarshi, S. Naiyer, X. Qu, H. Farooq, E. Kleiman, J. Xu, X. Lei, J. F. Cantillo, R. Wuerffel, N. Baumgarth, J. Liang, A. J. Feeney, A. L. Kenter, An Igh distal enhancer modulates antigen receptor diversity by determining locus conformation. *Nat. Commun.* **14**, 1225 (2023).
48. E. M. Barajas-Mora, L. Lee, H. Lu, J. A. Valderrama, E. Bjanes, V. Nizet, A. J. Feeney, M. Hu, C. Murre, Enhancer-instructed epigenetic landscape and chromatin compartmentalization dictate a primary antibody repertoire protective against specific bacterial pathogens. *Nat. Immunol.* **24**, 320–336 (2023).
49. A. L. Kenter, C. T. Watson, J.-H. Spille, Igh locus polymorphism may dictate topological chromatin conformation and V gene usage in the ig repertoire. *Front. Immunol.* **12**, 682589 (2021).
50. N. M. Choi, A. J. Feeney, CTCF and ncRNA regulate the three-dimensional structure of antigen receptor loci to facilitate V(D)J recombination. *Front. Immunol.* **5**, 49 (2014).
51. D. Nemazee, Mechanisms of central tolerance for B cells. *Nat. Rev. Immunol.* **17**, 281–294 (2017).
52. F. Melchers, Checkpoints that control B cell development. *J. Clin. Invest.* **125**, 2203–2210 (2015).
53. S. Lee, Y. Ko, T. J. Kim, Homeostasis and regulation of autoreactive B cells. *Cell. Mol. Immunol.* **17**, 561–569 (2020).
54. P. Dudzic, D. Chomicz, W. Bielska, I. Jaszczyszyn, M. Zieliński, B. Janusz, S. Wróbel, M.-M. L. Pannérer, A. Philips, P. Ponraj, S. Kumar, K. Krawczyk, Conserved heavy/light contacts and germline preferences revealed by a large-scale analysis of natively paired human antibody sequences and structural data, *Immunology* (2024). https://www.biorxiv.org/content/10.1101/2024.12.20.629642v3.full.
55. D. B. Jaffe, P. Shahi, B. A. Adams, A. M. Chrisman, P. M. Finnegan, N. Raman, A. E. Royall, F. Tsai, T. Vollbrecht, D. S. Reyes, N. L. Hepler, W. J. McDonnell, Functional antibodies exhibit light chain coherence. *Nature* **611**, 352–357 (2022).
56. M. Pospelova, K. Voss, A. Zamyatin, C. T. Watson, K.-P. Koepfli, A. Bankevich, M. Pennell, Y. Safonova, Comparative analysis of mammalian adaptive immune loci revealed spectacular divergence and common genetic patterns, *Genomics* (2025). https://www.biorxiv.org/content/10.1101/2025.04.01.646651v2.
57. K. Peng, Y. Safonova, M. Shugay, A. B. Popejoy, O. L. Rodriguez, F. Breden, P. Brodin, A. M. Burkhardt, C. Bustamante, V.-M. Cao-Lormeau, M. M. Corcoran, D. Duffy, M. Fuentes-Guajardo, R. Fujita, V. Greiff, V. D. Jönsson, X. Liu, L. Quintana-Murci, M. Rossetti, J. Xie, G. Yaari, W. Zhang, M. S. Abedalthagafi, K. O. Adekoya, R. A. Ahmed, W.-C. Chang, C. Gray, Y. Nakamura, W. D. Lees, P. Khatri, H. Alachkar, C. Scheepers, C. T. Watson, G. B. Karlsson Hedestam, S. Mangul, Diversity in immunogenomics: the value and the challenge. *Nat. Methods* **18**, 588–591 (2021).
58. M. M. Corcoran, G. E. Phad, N. Vázquez Bernat, C. Stahl-Hennig, N. Sumida, M. A. A. Persson, M. Martin, G. B. Karlsson Hedestam, Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat. Commun.* **7**, 13642 (2016).
59. A. A. Marsden, M. Corcoran, G. Karlsson Hedestam, N. Garrett, S. S. A. Karim, P. L.

Moore, D. Kitchin, L. Morris, C. Scheepers, Novel polymorphic and copy number diversity in the antibody IGH locus of South African individuals. *Immunogenetics* **77**, 6 (2024).

60. X. Castro Dopico, M. Mandolesi, G. B. Karlsson Hedestam, Untangling associations between immunoglobulin genotypes, repertoires and function. *Immunol. Lett.* **259**, 24–29 (2023).
61. P. Pushparaj, A. Nicoletto, X. C. Dopico, D. J. Sheward, S. Kim, S. Ekström, B. Murrell, M. Corcoran, G. B. Karlsson Hedestam, Frequent use of IGHV3-30-3 in SARS-CoV-2 neutralizing antibody responses. *Front. Virol.* **3**, 1128253 (2023).
62. Y. A. Yeung, D. Foletti, X. Deng, Y. Abdiche, P. Strop, J. Glanville, S. Pitts, K. Lindquist, P. D. Sundar, M. Sirota, A. Hasa-Moreno, A. Pham, J. Melton Witt, I. Ni, J. Pons, D. Shelton, A. Rajpal, J. Chaparro-Riggers, Germline-encoded neutralization of a Staphylococcus aureus virulence factor by the human antibody repertoire. *Nat. Commun.* **7**, 13376 (2016).
63. U. Kirik, H. Persson, F. Levander, L. Greiff, M. Ohlin, Antibody heavy chain variable domains of different germline gene origins diversify through different paths. *Front. Immunol.* **8**, 1433 (2017).
64. H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
65. W. Shen, S. Le, Y. Li, F. Hu, SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).
66. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
67. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
68. R. Kosoy, R. Nassir, C. Tian, P. A. White, L. M. Butler, G. Silva, R. Kittles, M. E. Alarcon-Riquelme, P. K. Gregersen, J. W. Belmont, F. M. De La Vega, M. F. Seldin, Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutat.* **30**, 69–78 (2009).
69. J. A. Vander Heiden, G. Yaari, M. Uduman, J. N. H. Stern, K. C. O'Connor, D. A. Hafler, F. Vigneault, S. H. Kleinstein, pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30**, 1930–1932 (2014).
70. N. T. Gupta, J. A. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, S. H. Kleinstein, Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**, 3356–3358 (2015).
71. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
72. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
73. G. Csardi, T. Nepusz, *The Igraph Software Package for Complex Network Research* (2006).
74. T. Pedersen, *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks* (2024).

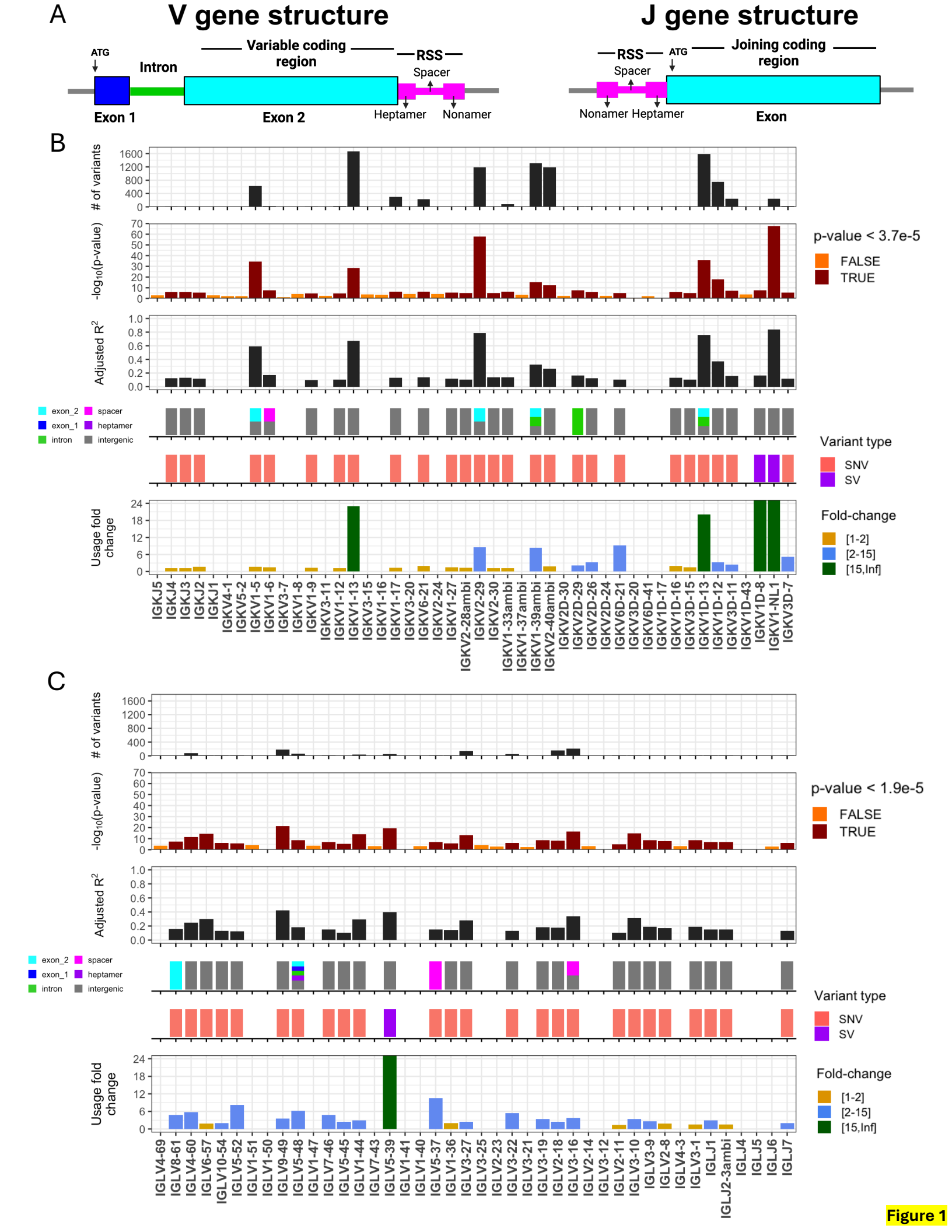

A
V gene structure
J gene structure
ATG
Intron
Variable coding region
RSS
Spacer
Exon 1
Exon 2
Heptamer
Nonamer
Joining coding region
Exon
B
# of variants
-log10(p-value)
Adjusted R2
p-value < 3.7e-5
FALSE
TRUE
exon_2
exon_1
intron
spacer
heptamer
intergenic
Variant type
SNV
SV
Usage fold change
Fold-change
[1-2]
[2-15]
[15,Inf]
IGKJ5 IGKJ4 IGKJ3 IGKJ2 IGKJ1 IGKV4-1 IGKV5-2 IGKV1-5 IGKV1-6 IGKV3-7 IGKV1-8 IGKV1-9 IGKV3-11 IGKV1-12 IGKV1-13 IGKV3-15 IGKV1-16 IGKV1-17 IGKV3-20 IGKV6-21 IGKV2-24 IGKV1-27 IGKV2-28ambi IGKV2-29 IGKV2-30 IGKV1-33ambi IGKV1-37ambi IGKV1-39ambi IGKV2-40ambi IGKV2D-30 IGKV2D-29 IGKV2D-26 IGKV2D-24 IGKV6D-21 IGKV3D-20 IGKV6D-41 IGKV1D-17 IGKV1D-16 IGKV3D-15 IGKV1D-13 IGKV1D-12 IGKV3D-11 IGKV1D-43 IGKV1D-8 IGKV1-NL1 IGKV3D-7
C
p-value < 1.9e-5
IGLV4-69 IGLV8-61 IGLV4-60 IGLV6-57 IGLV10-54 IGLV5-52 IGLV1-51 IGLV1-50 IGLV9-49 IGLV5-48 IGLV1-47 IGLV7-46 IGLV5-45 IGLV1-44 IGLV7-43 IGLV5-39 IGLV1-41 IGLV1-40 IGLV5-37 IGLV1-36 IGLV3-27 IGLV3-25 IGLV2-23 IGLV3-22 IGLV3-21 IGLV3-19 IGLV2-18 IGLV3-16 IGLV2-14 IGLV3-12 IGLV2-11 IGLV3-10 IGLV3-9 IGLV2-8 IGLV4-3 IGLV3-1 IGLJ1 IGLJ2-3ambi IGLJ4 IGLJ5 IGLJ6 IGLJ7

Figure 1

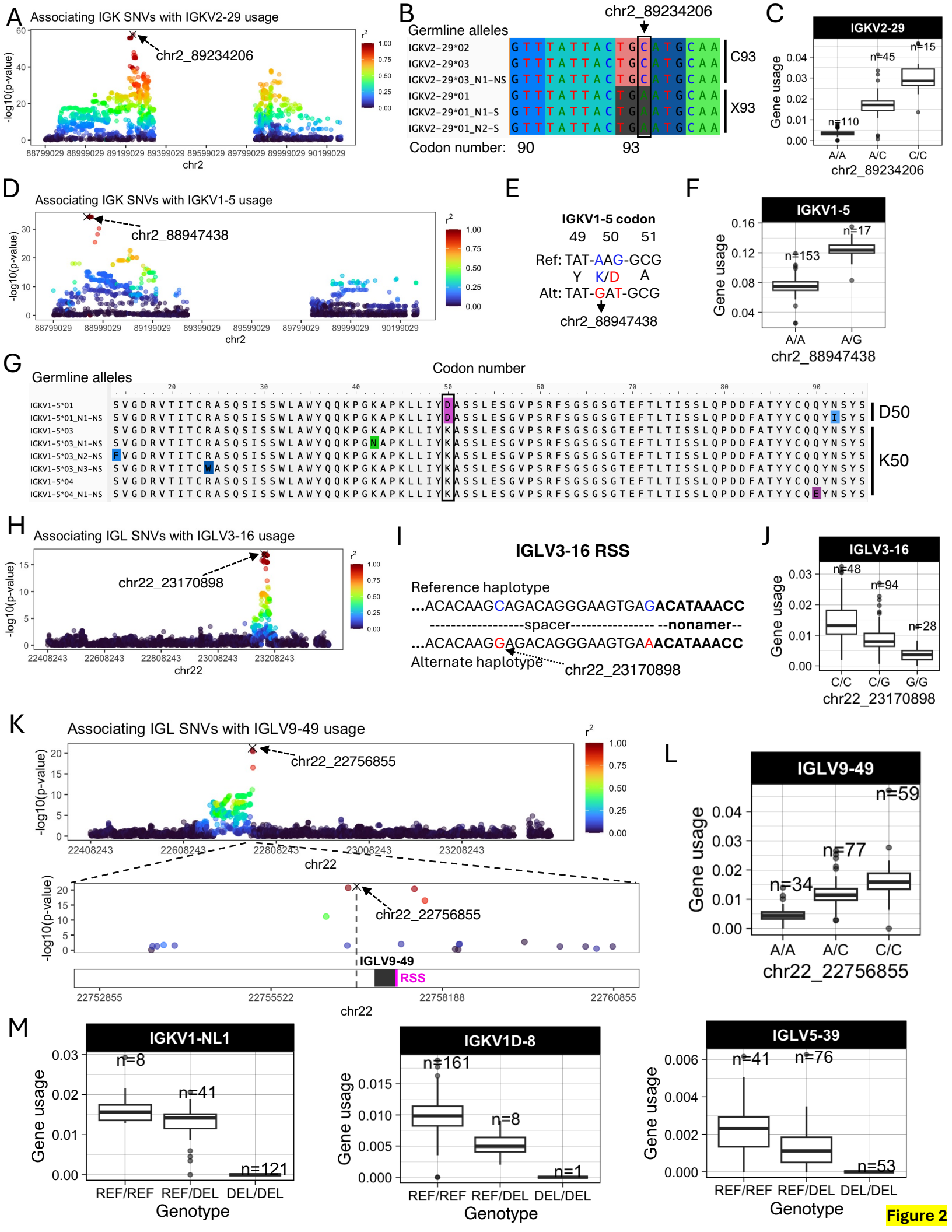
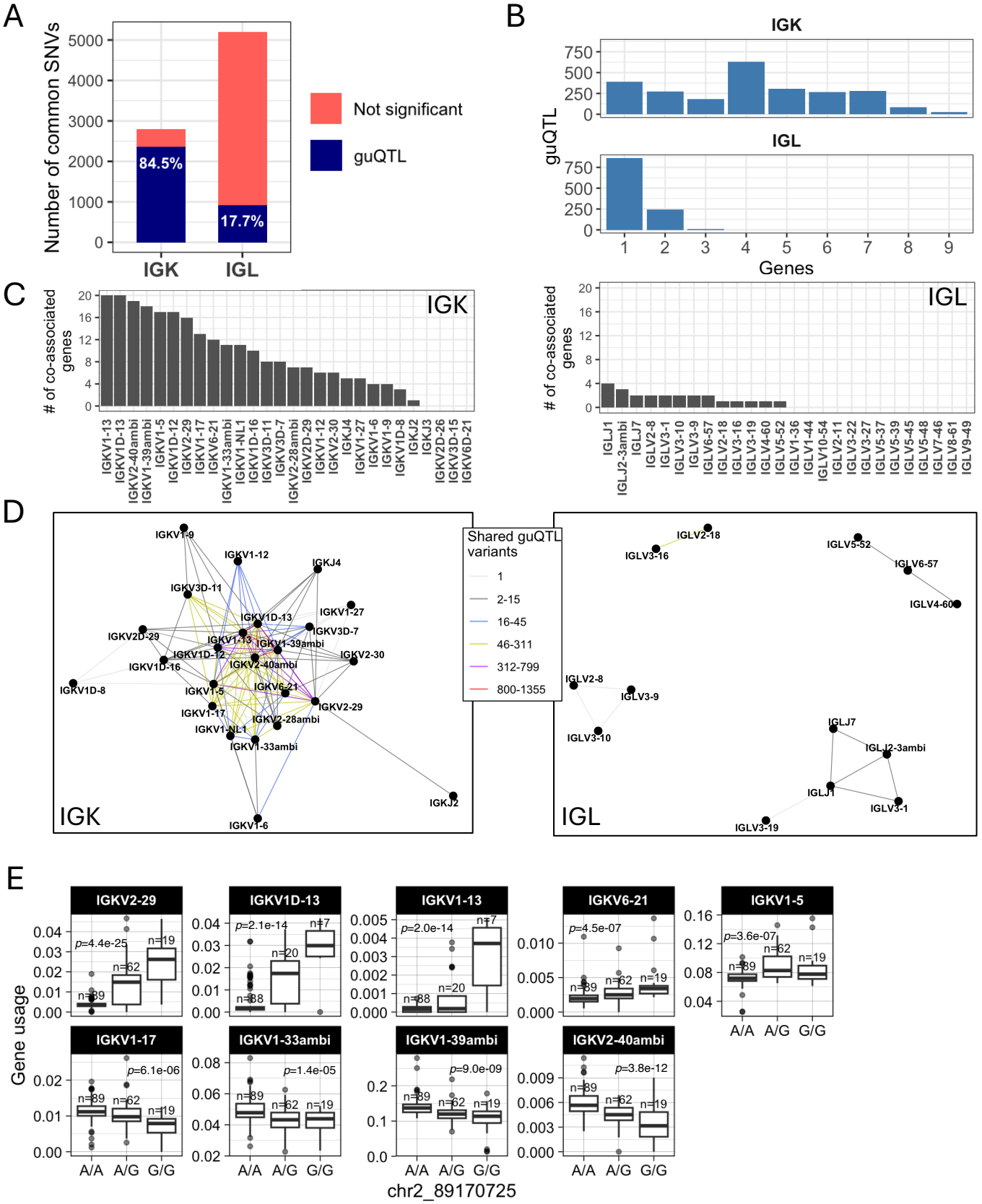

A
Associating IGK SNVs with IGKV2-29 usage
chr2_89234206
-log10(p-value)
chr2
B
Germline alleles
chr2_89234206
IGKV2-29*02
IGKV2-29*03
IGKV2-29*03_N1-NS
IGKV2-29*01
IGKV2-29*01_N1-S
IGKV2-29*01_N2-S
C93
X93
Codon number: 90 93
C
IGKV2-29
Gene usage
n=15
n=45
n=110
A/A A/C C/C
chr2_89234206
D
Associating IGK SNVs with IGKV1-5 usage
chr2_88947438
-log10(p-value)
chr2
E
IGKV1-5 codon
49 50 51
Ref: TAT-AAG-GCG
Y K/D A
Alt: TAT-GAT-GCG
chr2_88947438
F
IGKV1-5
Gene usage
n=17
n=153
A/A A/G
chr2_88947438
G
Germline alleles
Codon number
IGKV1-5*01
IGKV1-5*01_N1-NS
IGKV1-5*03
IGKV1-5*03_N1-NS
IGKV1-5*03_N2-NS
IGKV1-5*03_N3-NS
IGKV1-5*04
IGKV1-5*04_N1-NS
D50
K50
H
Associating IGL SNVs with IGLV3-16 usage
chr22_23170898
-log10(p-value)
chr22
I
IGLV3-16 RSS
Reference haplotype
...ACACAAGCAGACAGGGAAGTGAGACATAAACC
----------spacer---------- --nonamer--
...ACACAAGGAGACAGGGAAGTGAAACATAAACC
Alternate haplotype
chr22_23170898
J
IGLV3-16
Gene usage
n=48
n=94
n=28
C/C C/G G/G
chr22_23170898
K
Associating IGL SNVs with IGLV9-49 usage
chr22_22756855
-log10(p-value)
chr22
IGLV9-49
RSS
L
IGLV9-49
Gene usage
n=59
n=77
n=34
A/A A/C C/C
chr22_22756855
M
IGKV1-NL1
n=8
n=41
n=121
IGKV1D-8
n=161
n=8
n=1
IGLV5-39
n=41
n=76
n=53
Gene usage
REF/REF REF/DEL DEL/DEL
Genotype
Figure 2

Figure 3


A
Number of common SNVs
5000
4000
3000
2000
1000
0
84.5%
17.7%
IGK
IGL
Not significant
guQTL
B
IGK
IGL
guQTL
750
500
250
0
1
2
3
4
5
6
7
8
9
Genes
C
IGK
# of co-associated genes
20
16
12
8
4
0
IGKV1-13
IGKV1D-13
IGKV2-40ambi
IGKV1-39ambi
IGKV1-5
IGKV1D-12
IGKV2-29
IGKV1-17
IGKV6-21
IGKV1-33ambi
IGKV1-NL1
IGKV1D-16
IGKV3D-11
IGKV3D-7
IGKV2-28ambi
IGKV2D-29
IGKV1-12
IGKV2-30
IGKJ4
IGKV1-27
IGKV1-6
IGKV1-9
IGKV1D-8
IGKJ2
IGKJ3
IGKV2D-26
IGKV3D-15
IGKV6D-21
IGL
IGLJ1
IGLJ2-3ambi
IGLJ7
IGLV2-8
IGLV3-1
IGLV3-10
IGLV3-9
IGLV6-57
IGLV2-18
IGLV3-16
IGLV3-19
IGLV4-60
IGLV5-52
IGLV1-36
IGLV1-44
IGLV10-54
IGLV2-11
IGLV3-22
IGLV3-27
IGLV5-37
IGLV5-39
IGLV5-45
IGLV5-48
IGLV7-46
IGLV8-61
IGLV9-49
D
IGK
IGKV1-9
IGKV1-12
IGKJ4
IGKV3D-11
IGKV1-27
IGKV1D-13
IGKV3D-7
IGKV2D-29
IGKV1-13
IGKV1-39ambi
IGKV1D-12
IGKV2-30
IGKV1D-16
IGKV2-40ambi
IGKV1D-8
IGKV6-21
IGKV1-5
IGKV2-29
IGKV1-17
IGKV2-28ambi
IGKV1-NL1
IGKV1-33ambi
IGKJ2
IGKV1-6
Shared guQTL variants
1
2-15
16-45
46-311
312-799
800-1355
IGL
IGLV2-18
IGLV3-16
IGLV5-52
IGLV6-57
IGLV4-60
IGLV2-8
IGLV3-9
IGLV3-10
IGLJ7
IGLJ2-3ambi
IGLJ1
IGLV3-1
IGLV3-19
E
Gene usage
IGKV2-29
p=4.4e-25
n=89
n=62
n=19
IGKV1D-13
p=2.1e-14
n=88
n=20
n=7
IGKV1-13
p=2.0e-14
n=88
n=20
n=7
IGKV6-21
p=4.5e-07
n=89
n=62
n=19
IGKV1-5
p=3.6e-07
n=89
n=62
n=19
IGKV1-17
p=6.1e-06
n=89
n=62
n=19
IGKV1-33ambi
p=1.4e-05
n=89
n=62
n=19
IGKV1-39ambi
p=9.0e-09
n=89
n=62
n=19
IGKV2-40ambi
p=3.8e-12
n=89
n=62
n=19
A/A
A/G
G/G
chr2_89170725

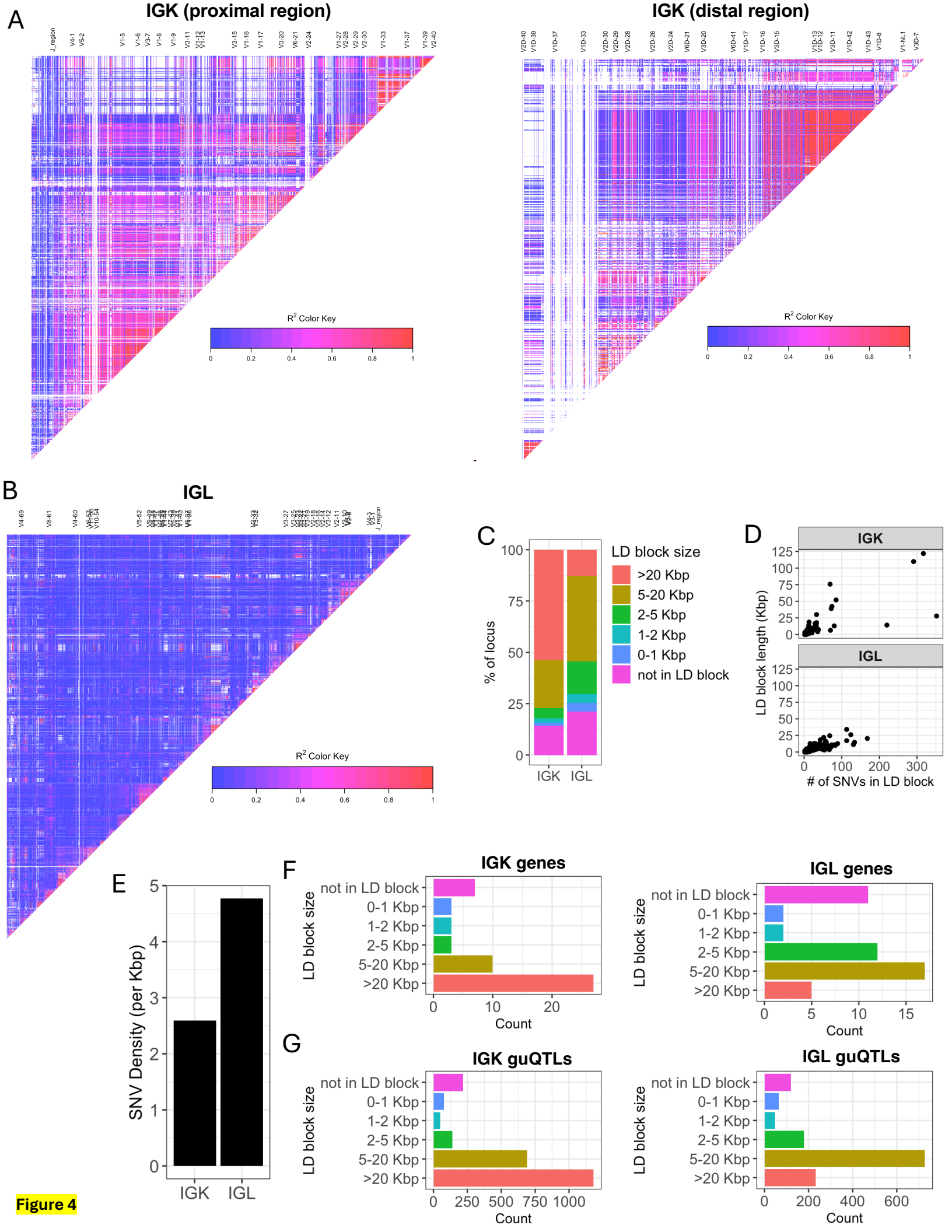

A
IGK (proximal region)
IGK (distal region)
R2 Color Key
0
0.2
0.4
0.6
0.8
1
B
IGL
C
LD block size
>20 Kbp
5-20 Kbp
2-5 Kbp
1-2 Kbp
0-1 Kbp
not in LD block
% of locus
IGK
IGL
D
LD block length (Kbp)
# of SNVs in LD block
E
SNV Density (per Kbp)
F
IGK genes
IGL genes
LD block size
Count
G
IGK guQTLs
IGL guQTLs

Figure 4

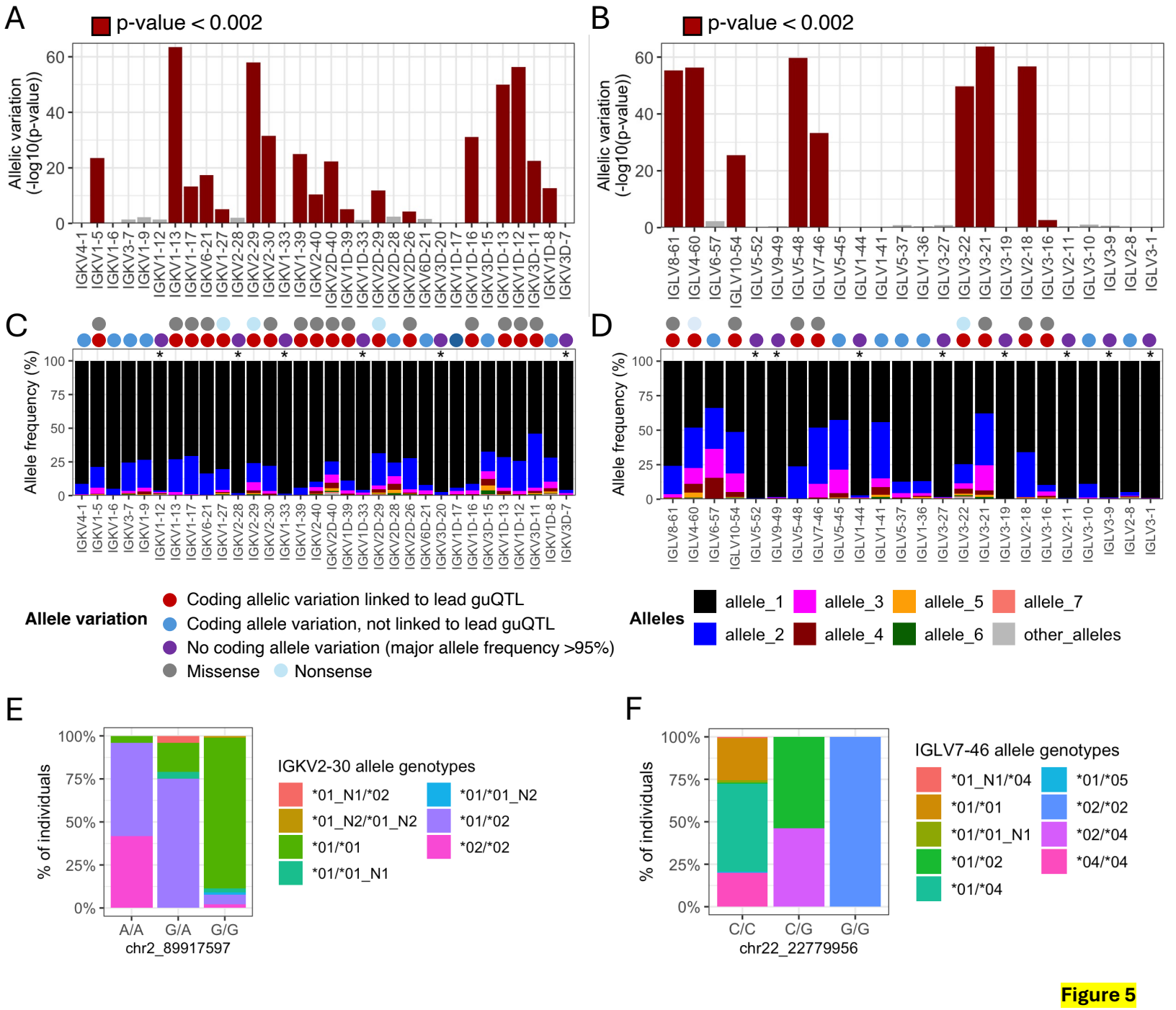

A
p-value < 0.002
Allelic variation (-log10(p-value))
B
p-value < 0.002
Allelic variation (-log10(p-value))
C
Allele frequency (%)
D
Allele frequency (%)
Allele variation
Coding allelic variation linked to lead guQTL
Coding allele variation, not linked to lead guQTL
No coding allele variation (major allele frequency >95%)
Missense
Nonsense
Alleles
allele_1
allele_2
allele_3
allele_4
allele_5
allele_6
allele_7
other_alleles
E
% of individuals
A/A
G/A
G/G
chr2_89917597
IGKV2-30 allele genotypes
*01_N1/*02
*01_N2/*01_N2
*01/*01
*01/*01_N1
*01/*01_N2
*01/*02
*02/*02
F
% of individuals
C/C
C/G
G/G
chr22_22779956
IGLV7-46 allele genotypes
*01_N1/*04
*01/*01
*01/*01_N1
*01/*02
*01/*04
*01/*05
*02/*02
*02/*04
*04/*04

Figure 5

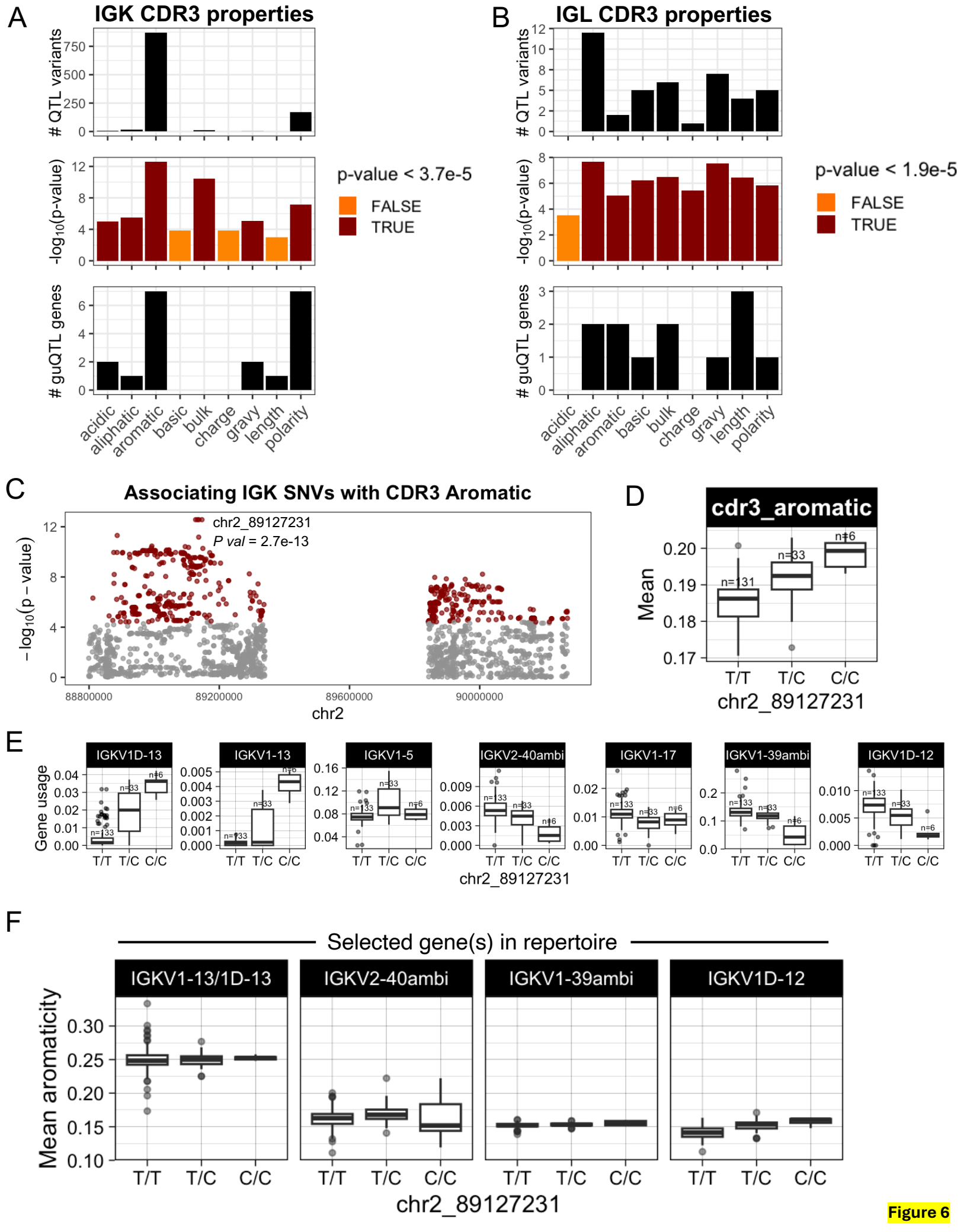

A
IGK CDR3 properties
QTL variants
-log10(p-value)
guQTL genes
p-value < 3.7e-5
FALSE
TRUE
acidic
aliphatic
aromatic
basic
bulk
charge
gravy
length
polarity
B
IGL CDR3 properties
p-value < 1.9e-5
C
Associating IGK SNVs with CDR3 Aromatic
chr2_89127231
P val = 2.7e-13
−log10(p−value)
chr2
D
cdr3_aromatic
Mean
T/T
T/C
C/C
chr2_89127231
E
IGKV1D-13
IGKV1-13
IGKV1-5
IGKV2-40ambi
IGKV1-17
IGKV1-39ambi
IGKV1D-12
Gene usage
chr2_89127231
F
Selected gene(s) in repertoire
IGKV1-13/1D-13
IGKV2-40ambi
IGKV1-39ambi
IGKV1D-12
Mean aromaticity
chr2_89127231

Figure 6