Germline polymorphisms 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 variable disease outcomes and therapeutic responsiveness, the determinants of which are incompletely understood. This study demonstrates that polymorphisms in immunoglobulin (IG) light chain loci dictate the composition of the Ab repertoire, establishing fundamental baseline differences that preclude functional Ab-mediated responses. Using long-read genomic sequencing of the IG kappa (IGK) and IG lambda (IGL) loci, we comprehensively resolved genetic variation, including novel structural variants, single nucleotide variants, and gene alleles. By integrating these genetic data with Ab repertoire profiling, we found that all forms of IG germline variation contributed to inter-individual gene usage differences for >70% of light chain genes in the repertoire, directly impacting the amino acids of expressed light chain transcripts, including complementarity determining region domains. The genomic locations of usage-associated variants in both intergenic and coding regions indicated that IG polymorphisms modulate gene usage via diverse mechanisms, likely including the modulation of V(D)J recombination, heavy and light chain pairing biases, and transcription/translation. Finally, relative to IGL, IGK was characterized by more extensive linkage disequilibrium and genetic co-regulation of gene usage, illuminating differential regulatory and evolutionary features between the two light chain loci. These results firmly establish the critical contribution of IG light chain polymorphism in Ab repertoire diversity, with important implications for investigating Ab responses in health and disease.

42 **Main**

43 Antibodies (Abs) are critical components of the adaptive immune system and are one of the most
44 diverse protein families in the human body. The circulating Ab repertoire comprises hundreds of millions of
45 unique antibodies [1,2], and its composition varies significantly between individuals [1–3]. The variability likely
46 contributes to the diverse Ab responses observed across clinical settings, including infection [4–8], autoimmunity
47 [9], and cancer [10]. Identifying the factors that contribute to variation in B cell-mediated immunity will inform
48 disease diagnosis and treatment.

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

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

76 Here, we pair long-read genomic sequencing of IGK and IGL with AIRR-seq at population-scale to
77 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 Fig. 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.

114

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

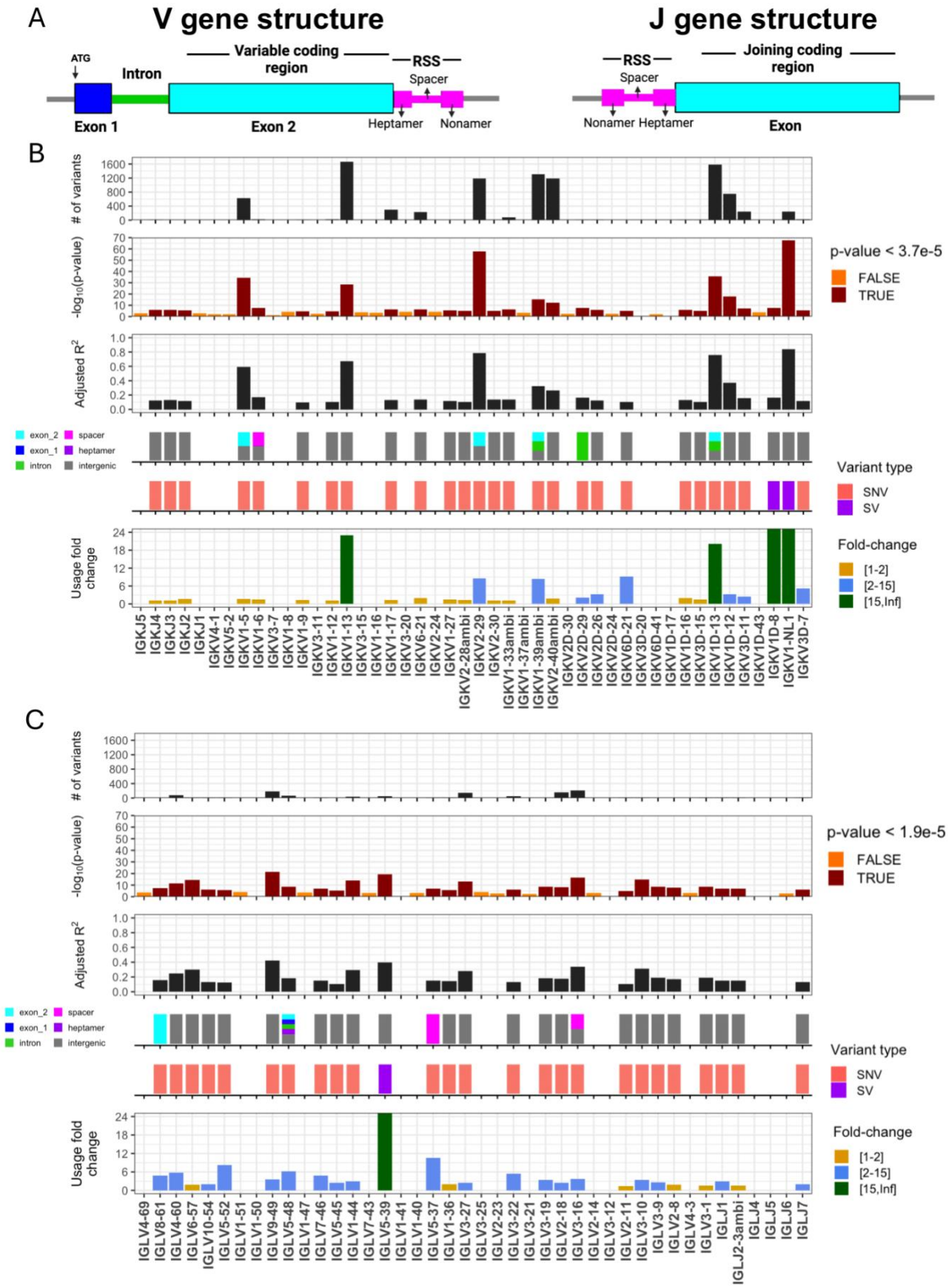
117 Throughout the genome, genetic variation has been associated with molecular traits such as gene
118 expression and splicing [35–38]. We previously demonstrated that genetic variants in the IGH locus mediate the
119 composition of peripheral IgM and IgG repertoires through effects on IGHV, IGHD, and IGHJ gene usage [12].
120 Here, we followed this same quantitative trait locus (QTL) framework to test if light gene usage was associated
121 with IGK and IGL variant genotypes in *cis*. Allele assignments to AIRR-seq reads were derived from a
122 personalized germline allele set for each individual. This permitted disambiguation of IGK gene paralogs for
123 individuals wherein each allele of a proximal paralog was distinct from each allele of the distal paralog,
124 including IGKV1-12 and IGKV1D-12, IGKV1-13 and IGKV1D-13, and IGKV6-21 and IGKV6D-21
125 (**Supplementary Table S5**). Paralog pairs for which at least 160 individuals could not be disambiguated
126 included IGKV1-33/1D-33, IGKV1-37/1D-37, IGKV1-39/1D-39, IGKV2-28/2D-28, and IGKV2-40/2D-40, and are
127 referred to as ambiguous or "ambi" (e.g. IGKV1-39/1D-39 is IGKV1-39ambi).

128 We performed genetic association tests on unmutated ("antigen naive") and mutated ("antigen
129 experienced") sets separately to identify *cis* effects in each of the two repertoire sets. In total, our analysis
130 included 41 IGKV, 5 IGKJ, 33 IGLV, and 4 IGLJ genes; 2792 SNVs and 2 SVs in IGK; and 5198 SNVs and 1
131 SV in IGL. In the unmutated IGK repertoire, after Bonferroni multiple-testing correction ($P < 5.0e-05$), we
132 identified 2,413 unique variants (2,411 SNVs, 2 SVs) that were statistically associated with gene usage
133 changes in 25 (60%) IGKV and 3 (60%) IGKJ and genes (**Fig. 1, Supplementary Table S6**). In the unmutated
134 IGL repertoire, a set of 918 unique variants (917 SNVs, 1 SV) were associated with gene usage changes in 22
135 (67%) IGLV genes and 3 (75%) IGLJ (**Fig. 1**, **Supplementary Table S6**). Notably, a large fraction of the
136 genes identified in both the IGK (n=14 genes) and IGL (n=19 genes) unmutated repertoires also had significant
137 gene usage QTLs (guQTLs) in the mutated repertoires (**Supplementary Figures 7-9**, **Supplementary Table**
138 **S7**). However, guQTLs in the unmutated repertoires tended to have lower *P* values and explain more variance
139 (R^2) in gene usage (**Supplementary Fig. 7**). We also noted stronger genetic effects in IGK relative to IGL; this
140 included the observation that overall genetic similarity among subjects associated with more highly correlated
141 IGK gene usage, a signal that was blunted in IGL (see **Supplemental Material, Supplementary Fig. 10**).
142 Together, these results show that usage of IG light chain genes is broadly impacted by germline genetic
143 variants, and while these genetic effects are more prominent in the antigen naive repertoire, for many genes,
144 those effects persist even following antigen exposure.

145

146

147


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
# of variants
-log10(p-value)
Adjusted R2
p-value < 1.9e-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]
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

148

149 **Figure 1. IGK and IGL variants impact gene usage in the naïve Ab repertoire**
150 (**A**) General structure of V and J genes in the IGK and IGL loci, including location of the recombination signal sequences
151 (RSS). (**B-C**) Per gene (x axis, all panels) statistics from linear regression guQTL analysis for the repertoire of unmutated
152 IGK (B) and IGL (C) light chains, including: (i) the number of associated variants after Bonferroni correction (IGK; P <
153 3.7e-5, IGL; P < 1.9e-5), (ii) −log10(P value) of the lead guQTL, (iii) adjusted R^2 for variance in gene usage explained by
154 the lead guQTL, (iv) the location and (v) type of variant for the lead guQTL and (vi) the fold change in gene usage
155 between genotypes at the lead guQTL. Summary statistics are provided in **Supplementary Table S6**.
156

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

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

167 Examples of lead guQTLs in coding, RSS, and intergenic regions are shown in **Fig. 2**, including IGKV2-
168 29, IGKV1-5, IGLV3-16, and IGLV9-49. The SNV-driven guQTL in this dataset with the lowest P value was for
169 IGKV2-29 (*P* value = 2.1e-58, **Fig. 2A**). This variant introduced a stop codon in V-region amino acid position 93
170 (**Fig. 2B**), resulting in decreased usage of IGKV2-29 (**Fig. 2C**). We also identified a lead guQTL associated
171 with missense variants. In the case of IGKV1-5, two linked lead guQTLs (r^2 = 1) within codon 50 associated
172 with a lysine to aspartic acid (AAG→GAT, K50D) change, resulting in an alteration of residue charge (**Fig. 2D-**
173 **E**). Individuals homozygous for K50 alleles, which represented six different IGKV1-5 coding alleles in this
174 cohort, had lower gene usage (**Fig. 2F-G**). As an example of a guQTL in the RSS, two lead variants in perfect
175 LD were identified at positions 8 and 23 of the spacer for IGLV3-16 (**Fig. 2H-J**). The reference haplotype had a
176 C at position 8, which was represented among consensus bases (C and T) at this position (**Supplementary**
177 **Fig. 11**), whereas the alternate haplotype had a G (**Fig. 2J**). Among the C/C and G/G lead guQTL genotype
178 groups, IGLV3-16 usage varied 3.8-fold on average (**Fig. 2K**). As noted above, the majority of lead guQTLs in
179 this dataset were in non-coding regions. For example, the lead guQTL for IGLV9-49 was 86 bp upstream of the
180 first exon, and guQTLs were not identified in coding sequence or the RSS (**Fig. 2L**). Mean IGLV9-49 usage
181 varied by 3.5-fold between homozygous-reference and homozygous-alternate individuals at this lead guQTL
182 (**Fig. 2M**). Consistent with previous observations in IGH [12], many guQTLs within IGL overlapped curated

183 transcription factor binding sites, representing an enrichment over background SNVs (see **Supplementary**
184 **Material, Supplementary Figures 12-13**), suggesting likely roles for non-coding variants in the regulation of
185 V(D)J recombination. Additionally, in IGK, we noted that both coding and non-coding regulatory variants
186 altered proximal and distal gene usage biases (see **Supplementary Material, Supplementary Figures 14-**
187 **15**).

188 Finally, in addition to SNV guQTLs, SVs resulting in gene copy number changes also made significant
189 impacts on gene usage. Specifically, SVs were lead guQTLs for the genes IGKV1-NL1, IGKV1D-8, and IGLV5-
190 39. In all cases, differential usage between genotypes followed an additive model in which gene usage
191 increased with every additional haploid gene copy (**Fig. 2N**). The lead variant associated with IGKV1D-8 usage
192 was the SV deletion of the entire IGKV distal region (see **Supplementary Fig. 4**). We noted that the number of
193 diploid IGLJ2-3 cassette copies associated with the usage of IGLJ1 and IGLJ2-3ambi (**Supplementary Fig.**
194 **16**); however, this CNV was not the lead QTL for these genes. The complexity of this SV will likely require
195 analysis in larger cohorts and more detailed assessment of potential haplotype-specific effects.

196 In summary, these results indicate that many forms of genetic variation are associated with gene usage
197 variation in the IGK and IGL repertoire. The variable localization of guQTLs in intergenic, RSS, and coding
198 regions implicates causative roles for these genetic variants in plausibly regulating V(D)J recombination,
199 transcription, and translation, as well as contributing to differential heavy-light chain pairing dynamics and
200 antigen selection.
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218

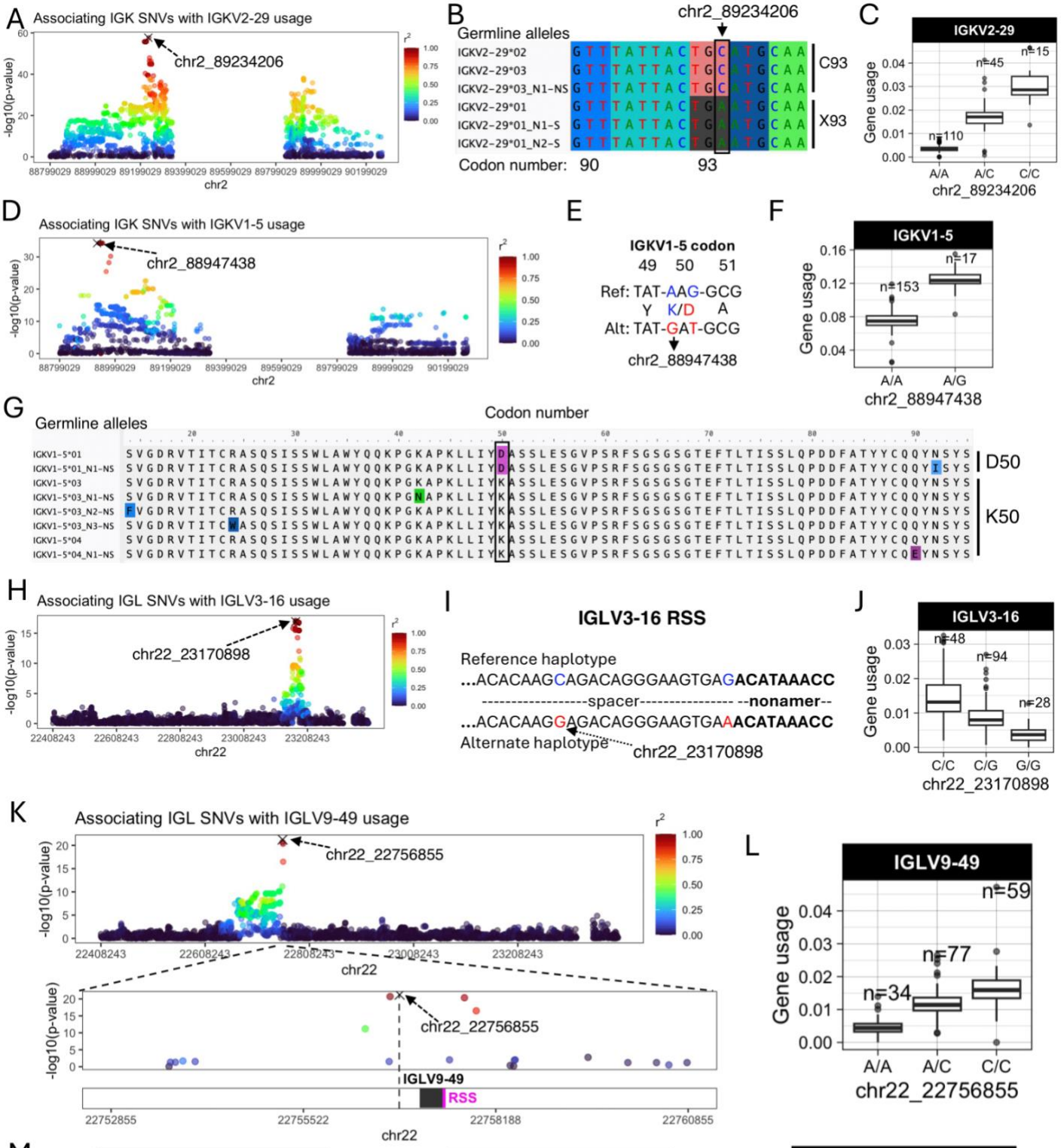

A
Associating IGK SNVs with IGKV2-29 usage
chr2_89234206
-log10(p-value)
chr2
B
chr2_89234206
Germline alleles
IGKV2-29*02
IGKV2-29*03
IGKV2-29*03_N1-NS
IGKV2-29*01
IGKV2-29*01_N1-S
IGKV2-29*01_N2-S
C93
X93
Codon number: 90 93
C
IGKV2-29
Gene usage
n=15
n=45
n=110
A/A A/C C/C
chr2_89234206
D
Associating 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
REF/REF REF/DEL DEL/DEL
Genotype
IGKV1D-8
n=161
n=8
n=1
REF/REF REF/DEL DEL/DEL
Genotype
IGLV5-39
n=41
n=76
n=53
REF/REF REF/DEL DEL/DEL
Genotype

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 (r^2) 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 (r^2) 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 (r^2) 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 (r^2) 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.

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) (**Fig. 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 (**Fig. 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 (**Fig. 3B**). Likewise, at gene-level, 16 of 23 IGK guQTL genes shared at least one significant variant with >5 other genes. In contrast,

254 0/26 IGL genes shared significant guQTL variants with >5 genes. The majority of IGL guQTL genes (24 of 26)
255 shared variants with fewer than 3 other genes (**Fig. 3C**).
256 To visualize these relationships between genes and guQTLs, we constructed networks in which nodes
257 represented genes and edges represented connections between genes sharing at least one guQTL SNV.
258 From these networks, we identified multi-member cliques, in which 2 or more genes were connected by at
259 least one shared guQTL. For IGK, 21 of the 23 guQTL genes formed a single super clique, with embedded
260 subcliques in which all genes were connected to one another through guQTL variants (**Fig. 3D**). Demonstrative
261 of interconnected gene usage, the largest subclique was composed of 9 guQTL genes associated with a single
262 guQTL SNV (**Fig. 3E**). In contrast to IGK, 13 of the 26 guQTL IGL genes were represented by 4 distinct
263 cliques, all of which were smaller than the large clique observed in IGK and disconnected from one another
264 (gene membership range = 2-5; **Fig. 3D**).
265 The stark difference in IGK and IGL clique sizes (**Fig. 3D**) suggested likely differences in the genetic
266 haplotype structure between the two loci. To explore this, we estimated pairwise linkage disequilibrium (LD)
267 between all common SNVs (MAF >= 5%) and determined blocks of LD [39,40] (**Supplementary Table S9**, see
268 Methods). LD was more extensive in IGK (**Fig. 4A**) relative to IGL (**Fig. 4B**), with LD blocks >20 Kbp
269 comprising 53.5% and 12.9% of the IGK and IGL loci, respectively, (**Fig. 4C, Supplementary Table S10**). In
270 contrast, 24.4% of IGL was within LD blocks <5 Kbp, compared to only 8.6% of the IGK locus (**Fig. 4C**). The
271 three largest LD blocks in IGK were 122 Kbp, 110 Kbp, and 76 Kbp, compared to the three largest LD blocks in
272 IGL that were 34 Kbp, 26 Kbp and 24 Kbp (**Fig. 4D**). As expected, the number of SNVs per block was
273 positively correlated in both loci. Additionally, larger sets of genes in IGK fell within large LD blocks. We found
274 that 27 of 47 IGKV genes were within LD blocks >20 Kbp, whereas 5 of 39 IGLV genes were within LD blocks
275 >20 Kbp (**Fig. 4F**). We also noted that 26 of 47 IGKV genes were within one of 6 LD blocks containing >1
276 gene, in contrast to IGL (**Supplementary Fig. 17**), in which this was true for only 7 of 33 IGLV genes
277 (**Supplementary Fig. 18**). IGK guQTL SNVs were also more frequently in large LD blocks as 1,181 (49%)
278 were in blocks >20 Kbp, whereas 233 (17%) IGL guQTL SNVs were in blocks >20 Kbp (**Fig. 4G**).
279 These data demonstrate that a larger proportion of IGK sequence, genes, and guQTLs are contained
280 within large LD blocks as compared to IGL. The overlap of LD blocks with guQTL and gene cliques suggests
281 that extended haplotype structures within both loci likely contribute to coordinated gene regulation.
282
283
284
285
286
287
288
289

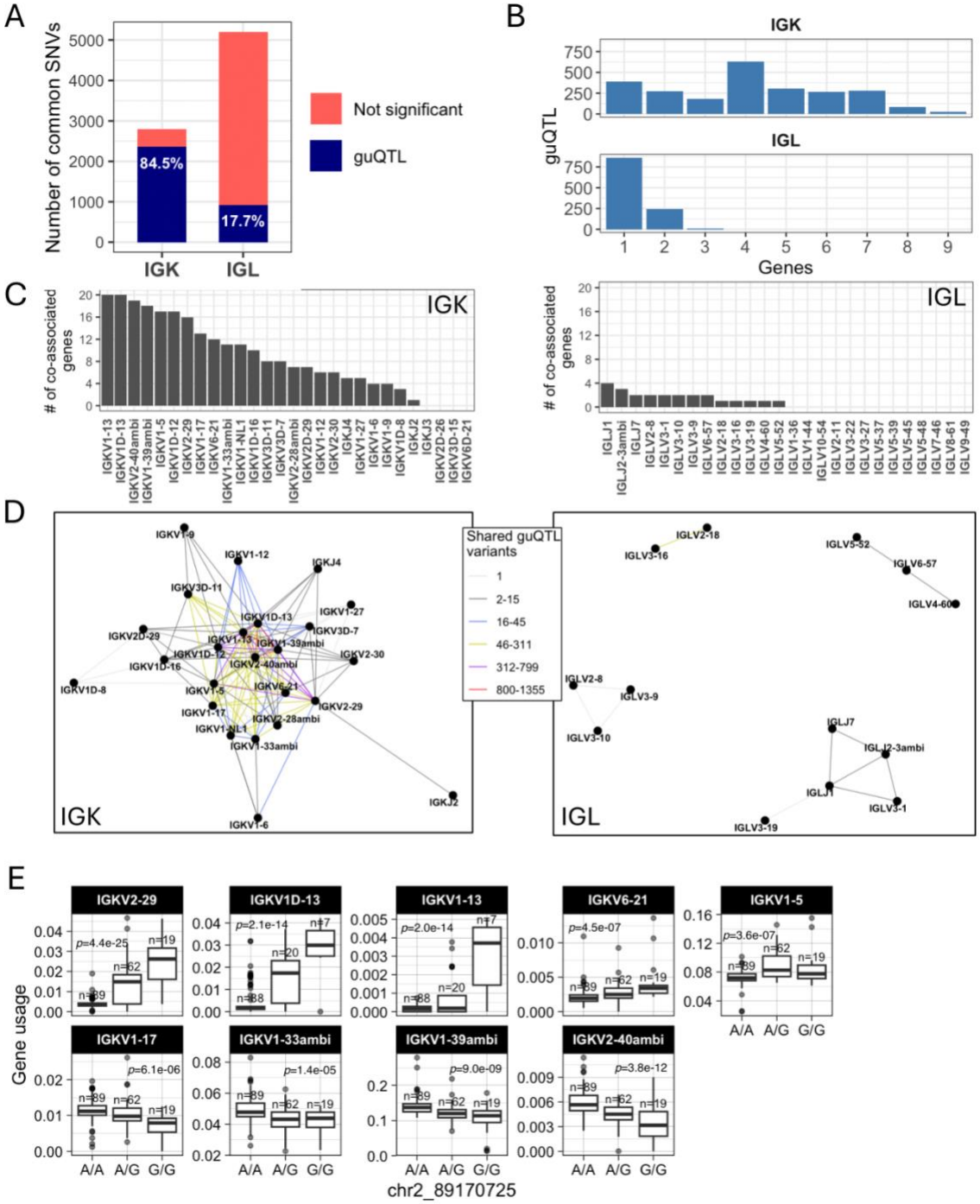
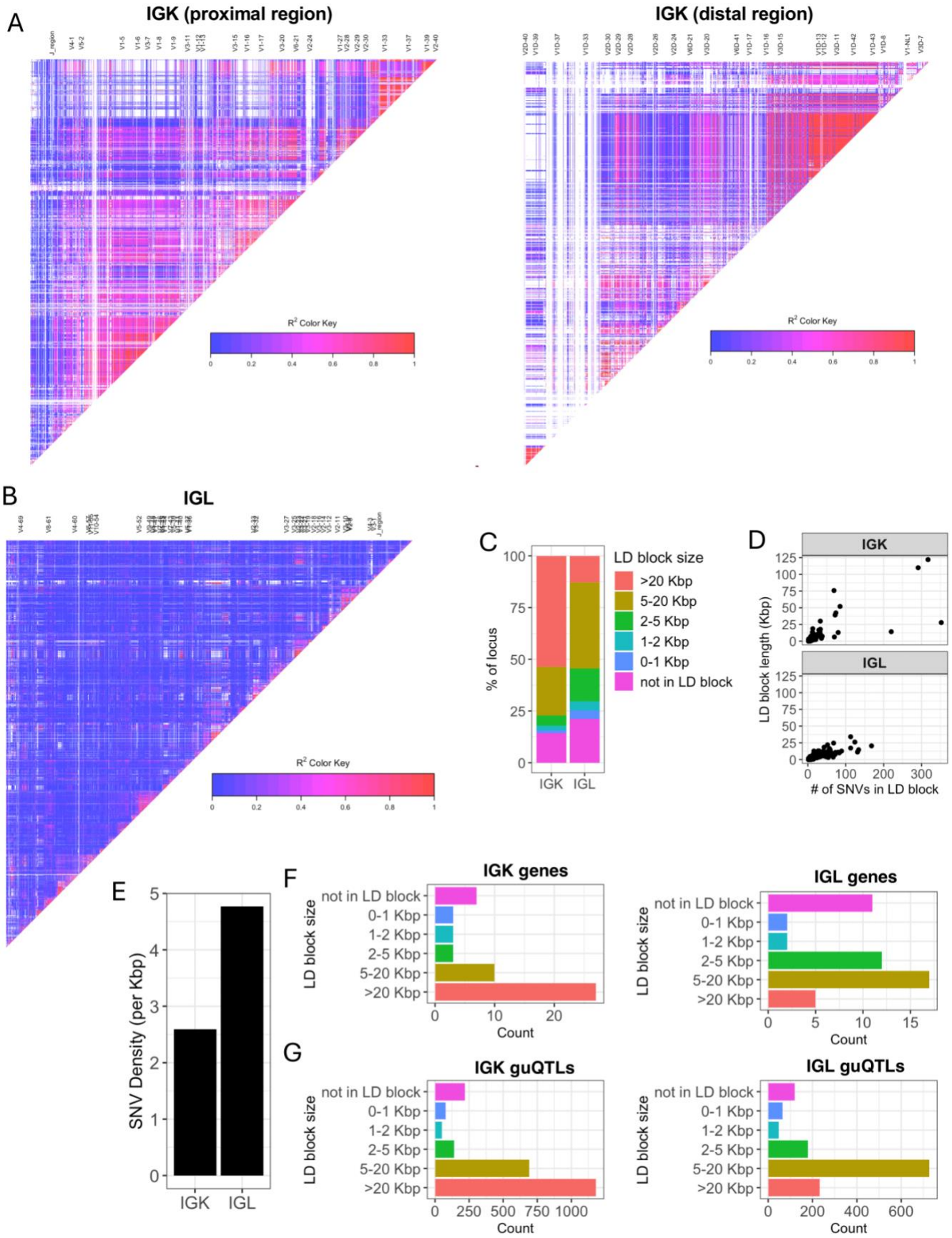


Figure 3. Genetic coordination of IG light chain gene usage is more prevalent in IGK relative to IGL

(**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.


A
IGK (proximal region)
IGK (distal region)
R² Color Key
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
G
IGK guQTLs
IGL guQTLs
LD block size
Count

302 **Figure 4. IGK has larger LD blocks and lower density of SNVs relative to IGL**
303 (**A-B**) LD heatmaps of the IGK (A) and IGL (B) loci. LD blocks are illustrated as triangles. (**C**) Stacked bar plot of the
304 percent of each locus (IGK, IGL) that is within LD blocks of various lengths (colors). (**D**) Plots of LD blocks in IGK and IGL
305 depicting the length of each block (y-axis) and number of SNVs in each blot (x-axis). (**E**) Bar plot of the overall SNV
306 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
307 indicated along y-axes.
308

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

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

320 First, we found that many lead guQTLs were associated with shifts in coding allele usage within the
321 repertoire, representing LD between coding and non-coding SNVs. This was consistent with our previous
322 investigation of IGH guQTLs [12]. Specifically, for 18/32 (56%) tested IGK genes, individuals within different
323 guQTL genotypes exhibited differential coding allele frequencies (Fisher's exact test, Bonferroni; $P < 0.002$).
324 Likewise, in IGL we noted such associations for 9/24 (38%) guQTL genes (Fisher's exact test, Bonferroni;
325 $P < 0.002$) (**Fig. 5A-B**, **Supplementary Table S11**). Among these genes, the 12/18 (66%) in IGK, and 7/9
326 (77%) in IGL involved alleles carrying amino acid changes (**Fig. 5C-D**). For the remaining genes in each locus,
327 genes either exhibited allelic variation, but did not associate with guQTL genotype, or lacked appreciable allelic
328 variation (major allele frequency >95%; **Fig. 5C-D**). Examples of genes with coding allele variation linked to
329 lead non-coding guQTL variants include IGKV2-30 and IGLV7-46, for which gene alleles were distributed
330 differently among the non-coding lead guQTL genotypes (**Fig. 5E-F**). In the case of IGKV2-30, the *02 allele,
331 which harbored a missense variant in CDR1, was carried by 95.9% of individuals with genotype A/A at the lead
332 guQTL variant, compared to only 7.9% of individuals with genotype G/G (**Fig. 5E**). Likewise, in the case of
333 IGLV7-46, the *02 allele, which harbored an amino acid change in FWR3, was carried by 100% of individuals
334 with guQTL genotype G/G and by 0.7% of individuals with genotype C/C (**Fig. 5F**).

335 We next asked whether shifts in gene usage also resulted in shifts in CDR3 physicochemical
336 properties. To do this, we conducted an unbiased test for associations between variation in nine CDR3
337 properties and genotypes at all variants across IGK and IGL (**Supplementary Table S12**). In IGK, we
338 identified SNVs associated (Bonferroni, $P < 3.9e-05$) with CDR3 properties of aromaticity, aliphaticity, acidity,

339 polarity, bulk, and GRAVY index (**Fig. 6A; Supplementary Fig. 19**). Likewise, in IGL, SNVs were associated
340 (Bonferroni, $P < 1.9e-05$) with CDR3 aromaticity, aliphaticity, GRAVY index, bulk, basicity, polarity, charge, and
341 length (**Fig. 6B; Supplementary Fig. 20**). Lead variants in both loci overlapped guQTLs, linking CDR3
342 properties with gene usage variation. For example, the lead variant associated with IGK CDR3 aromaticity was
343 also a guQTL for seven IGKV genes (**Fig. 6A, 6C-E**) that are part of a previously described network clique
344 (**Fig. 3D**). Among these genes, we focused on those with usage patterns that were negatively correlated, and
345 thus considered to differentially represent the guQTL genotypes. Specifically, IGKV1-13/1D-13 usage was
346 highest in individuals with high CDR3 aromaticity. In contrast, the usage of IGKV2-40ambi, IGKV1-39ambi, and
347 IGKV1D-12 were highest in individuals with low CDR3 aromaticity (**Fig. 6D-E**). To determine whether these
348 genes explain CDR3 aromaticity variation between genotype groups, we computed CDR3 aromaticity for BCRs
349 utilizing only IGKV1-13/1D-13, IGKV2-40ambi, IGKV1-39ambi, or IGKV1D-12 (**Fig. 6F**). This demonstrated
350 IGKV1-13/1D-13-encoded BCRs had higher aromaticity than those encoded by all of the other three genes.
351 This was regardless of the J gene contribution (**Supplementary Fig. 20**), indicating the genetic effect on CDR3
352 aromaticity is through influence on V gene usage.
353 In IGL, CDR3 aliphaticity and aromaticity shared the same lead variant; this variant was also the lead
354 guQTL for both IGLJ1 and IGLJ2-3ambi, linking genetic regulation of IGLJ gene usage with CDR3 properties
355 (**Supplementary Fig. 21**). At this variant, A/A individuals had relatively higher CDR3 aliphaticity and IGLJ2-
356 3ambi usage, whereas G/G individuals had relatively higher CDR3 aromaticity and IGLJ1 usage. Analysis of
357 BCRs using one or the other of these IGLJ genes revealed that sequences containing IGLJ2-3ambi have
358 relatively high CDR3 aliphaticity, whereas sequences containing IGLJ1 have relatively high CDR3 aromaticity
359 (**Supplementary Fig. 21**). Together, these data link genetic effects on IGLJ gene usage with IGL CDR3
360 properties.
361 In summary, 41% of IGLV and 63% of IGKV genes showed significant variation in coding alleles among
362 lead guQTL genotype groups, indicating LD between non-coding variants and gene alleles. Additionally, we
363 show that variation in gene usage also contributes to biases in CDR3 properties, at least in part explained by
364 contributions of germline encoded amino acids at the 3' of V genes and 5" end of J genes. This is noteworthy,
365 as it indicates that guQTLs not only impact general variation in gene usage, but also have the potential to
366 modulate the availability of germline encoded residues in the baseline unmutated repertoire.
367
368
369
370
371
372

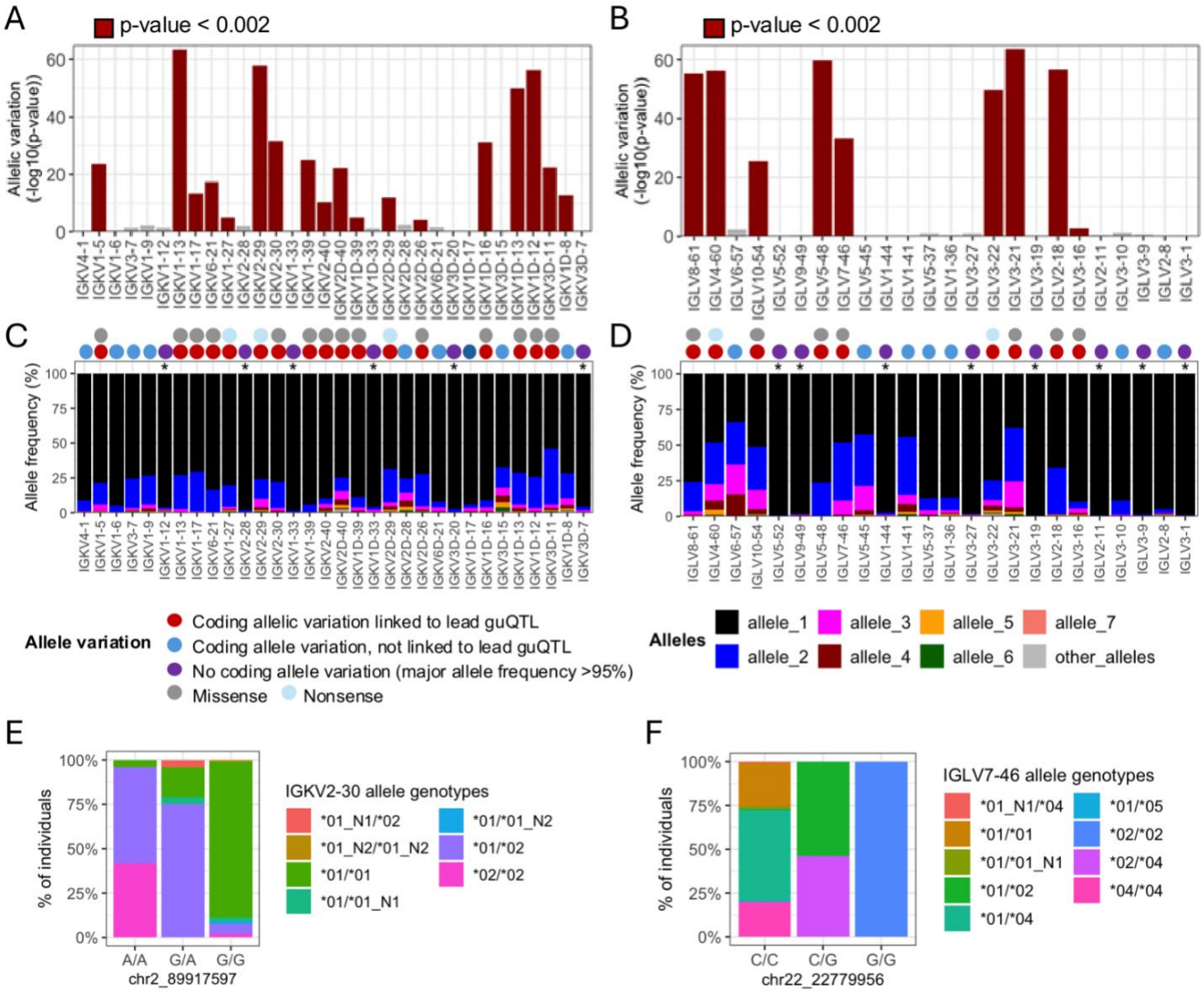
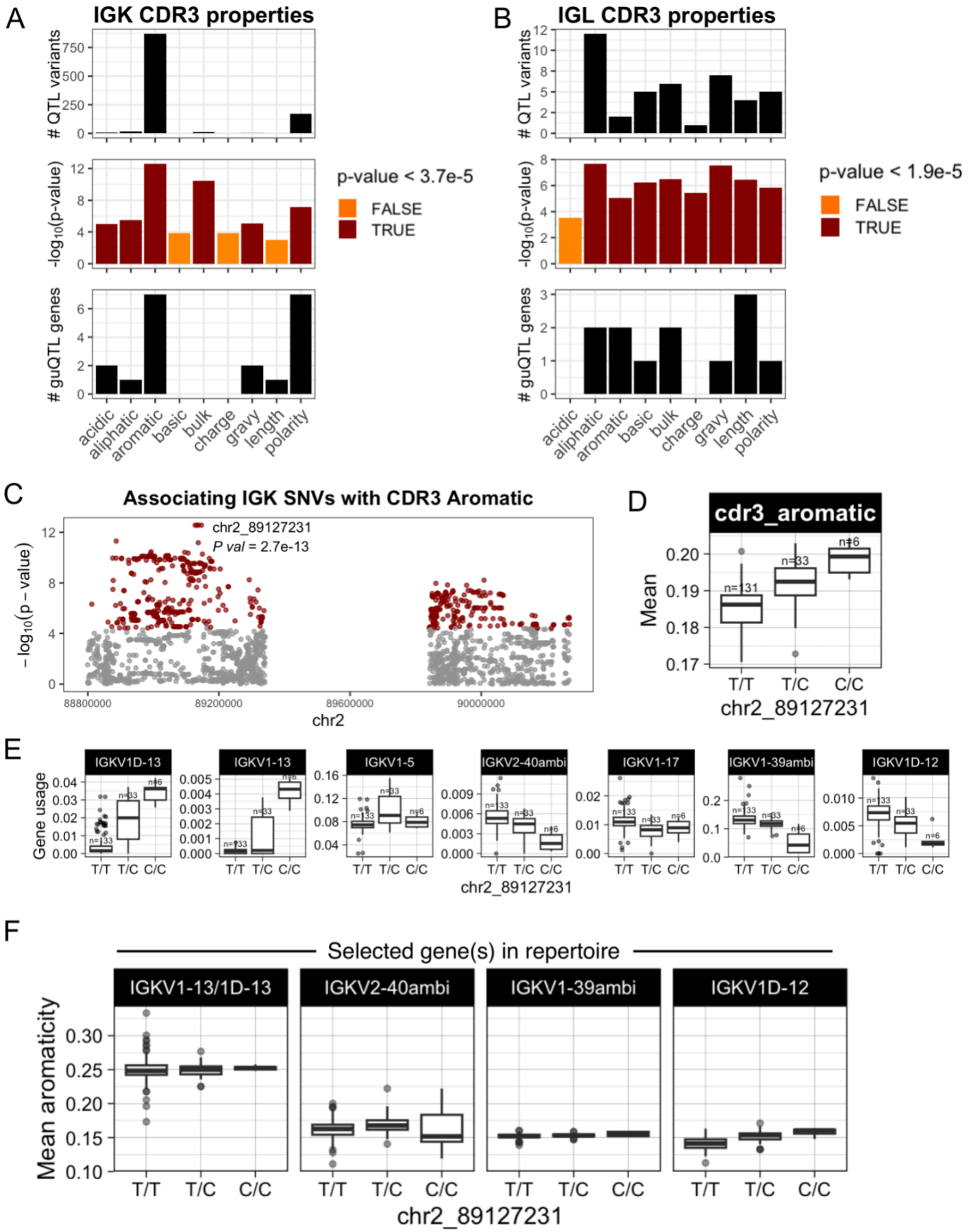


Figure 5. Linkage between IGKV and IGLV coding region alleles and lead guQTL genotypes

(**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 -$\log_{10}$(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**).


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
F
Selected gene(s) in repertoire
IGKV1-13/1D-13
IGKV2-40ambi
IGKV1-39ambi
IGKV1D-12
Mean aromaticity

387

388 **Figure 6. IGK and IGL variants impact CDR3 physicochemical properties in the naïve Ab repertoire**
389 (**A-B**) For each CDR3 physicochemical property (x-axis), mean values were computed for each individual and tested for
390 association (linear regression) with all common variants in IGK (**A**) and IGL (**B**). Barplots show (i) the number of QTL
391 variants (Bonferroni-corrected) for each property, (ii) the -log10(P value) for lead variants, and (iii) the number of guQTL
392 genes identified for the lead CDR3 property QTL variant. Summary statistics are provided in **Supplementary Table S12**.
393 (**C**) Manhattan plot shows the −log10(P value) for all SNVs in the IGK locus tested for association with CDR3 aromaticity,
394 with QTLs colored dark red and the lead QTL labelled. (**D**) Boxplot of the mean IGK CDR3 aromaticity with individuals
395 separated by genotype at the lead QTL. (**E**) Boxplots of usages for seven IGK genes that are guQTLs at the lead CDR3
396 aromaticity variant. (**F**) BCR sequences that used the indicated V genes were selected from the Ab repertoire, then mean
397 CDR3 aromaticity of each repertoire subset was computed and plotted with individuals separated by genotype at the lead
398 CDR3 aromaticity QTL.
399

400 **Discussion**

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

417 Across the IGK and IGL loci, genetic variants were statistically associated with usage variation in the
418 majority of V and J genes. For a subset of IGKV genes, for example, we found that even single guQTLs could
419 explain over 75% of the usage variation among donors. Notably, however, although we found that most
420 guQTLs and associated genes were common between unmutated and mutated repertoires, the extent of
421 variance in gene usage explained by guQTLs in the unmutated repertoire was on average higher for both IGK
422 and IGL; V genes also had stronger genetic associations relative to J genes in both loci. The blunted genetic
423 effects in the mutated repertoire may reflect shifts in usage in the memory repertoire driven by interactions with

424 antigen. However, it is notable that even in antigen-experienced repertoires, variation can still be explained by
425 genetic factors, indicating some degree of genetic constraint, consistent with observations in IGH [12,16]. We also
426 noted that on average, R^2 values were lower in IGL; however, because this analysis only included an
427 assessment of lead guQTLs, we have not accounted for additional variants that may make additional genetic
428 contributions. We previously showed that secondary guQTLs in IGH were able to increase the variance in gene
429 usage explained by *cis* genetic factors [12]. As cohorts increase in size, we expect it will be possible to
430 characterize additional guQTLs in IGK and IGL.

431 An assessment of guQTL positions within each locus provided initial evidence of the mechanisms by
432 which they may exert their effects on the repertoire. Early studies in IGK were the first to show that IG
433 polymorphism can directly impact the usage of particular genes in the repertoire. These specifically linked
434 variation within the RSS of IGKV2D-29 and its usage frequency [28,43,44], demonstrating that RSS polymorphisms
435 have the potential to influence the binding of RAG1/2 and the selection of genes by V(D)J recombination. Here,
436 we found additional direct evidence for effects of RSS variants on several IGK and IGL genes. However,
437 consistent with previous observations in IGH [12], the majority of light chain guQTLs were intergenic SNVs. The
438 factors underlying the effects of non-coding guQTLs require further study, but are expected to influence various
439 regulatory mechanisms (e.g., enhancer and promoter and function; formation of topologically associating
440 domains) in the IG loci during V(D)J recombination [45–49]. It is notable that in IGL we observed overlap between
441 a subset of intergenic guQTLs and known TFBSs (**Supplemental Fig. 12**); this included enrichments in
442 binding sites for CTCF, a transcription factor known to be critical to the chromatin landscape within IG loci [50].
443 We also found that SVs were the lead guQTLs for three genes (IGKV1D-8, IGKV1-NL1, and IGLV5-39). In all
444 cases these SVs altered the number of diploid copies (range = 0-2) and thus the chance they could be
445 selected by V(D)J recombination. We identified multiple less common SVs (MAF <5%) that will require study in
446 larger sample sizes to more fully assess their contribution to gene usage. Overall, however, we noted that the
447 number of genes impacted by SVs in each of the light chain loci were comparatively fewer than reported in IGH
448 [12], which is likely a reflection of the fact that the IGH locus overall has a greater number of SVs.

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

460 will require careful investigation of guQTLs across developmental time points, and will need to consider
461 combined effects of polymorphisms across the three loci.

462 To date, differences in the genetic architecture of the IGK and IGL loci have been underexplored.
463 Previous comparisons of a small number of IGK and IGL haplotypes indicated that SNV densities were lower in
464 IGK compared to IGL [31]. Our data confirmed this pattern at the population level, revealing that the number of
465 common SNVs was almost 2-fold higher in IGL. Additionally, we found that IGL was also characterized by less
466 extensive LD. These stark differences in genetic architecture were reflected in the interconnectedness of gene
467 usage profiles in the repertoire, specifically that a greater proportion of genes in the IGK repertoire shared
468 overlapping guQTLs. This suggests that we could expect the regulatory landscapes to also be different. To
469 date, our knowledge of V(D)J regulation in the light chain loci come from studies in mice; however, it is unlikely
470 that we can extrapolate much detail from these studies, as these loci show little structural resemblance to
471 those in humans. Given the ordered engagement of IGK and IGL genes in the formation of functional BCRs
472 during B cell development, it's plausible that the differences in genetic structure have been shaped by their
473 differing functional roles. Recent work looking at IG locus genetic features across a range of vertebrate species
474 suggest co-evolution of the light chain loci [56].

475 These data highlight the impacts of germline variation on gene usage variation in the repertoire, and
476 their underlying mechanisms. We argue that the influence of genetics should be considered when seeking to
477 understand how inter-individual differences in repertoire composition directly contribute to antigen-driven
478 responses [19–22,24,57–62]. Given coding differences between genes, usage variation by default alters the
479 landscape of available germline encoded residues among expressed BCRs, which can also bias SHM patterns
480 [63]. Here, we demonstrate that guQTLs are also directly linked to amino acid differences between alleles of
481 individual genes; thus, while some coding variants may be present within the genome of an individual, their
482 frequency within the repertoire is dependent on guQTL genotype. Taking this a step further, we showed that
483 gene usage also correlated with variation in CDR3 properties, driven by direct contributions of 3' V and 5' J
484 gene germline-encoded bases to junction amino acids. This is consistent with observations of CDR3
485 comparisons between monozygotic twins and unrelated individuals [17]. The link between guQTLs and amino
486 acid features among expressed light chain transcripts elevates the likelihood that guQTLs significantly impact
487 the antigen-binding landscape of expressed Abs.

488 Together, these findings advance our basic understanding of repertoire development, illuminating
489 regions of IGK and IGL that not only regulate gene usage but establish biases in the amino acid diversity
490 observed among expressed Abs. In combination with our previous work in the IGH locus, our data lay a
491 foundation for integrating genetic contributions from all three IG loci to establish more complete and precise
492 models of sequence diversity in the expressed repertoire. This will be critical for refining our understanding of
493 Ab repertoire dynamics in health and disease.
494
495

496 **Materials and Methods**

497

498 **Sample information**

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

502

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

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

513 Enriched IGK and IGL libraries were prepared for sequencing using the SMRTBell Express Template
514 Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA) and SMRTBell Enzyme Cleanup Kit 1.0 (Pacific
515 Biosciences), according to the manufacturer's protocol. Resulting SMRTbell libraries were multiplexed in pools
516 of 12 and sequenced using one SMRT cell 8 M on the Sequel IIe system (n=134) (Pacific Biosciences) using
517 2.0 chemistry and 30-hour movies. For Revio sequencing (n=43), SMRTbell libraries were pooled in 36-plexes
518 and sequenced using one SMRT cell 25M on the Revio system (Pacific Biosciences) using Revio Polymerase
519 Kit v1.0 (PacBio; 102-739-100) and 30 hour movies. High Fidelity ("HiFi") intramolecular circular consensus
520 reads with accuracies >99.9% (Q20) were generated on instrument and used for all downstream analyses.

521

522 **IGK and IGL AIRR-seq**

523 AIRR-seq libraries were prepared and sequenced for 173 individuals (IGK) 171 (IGL). RNA was
524 extracted from ~3-5 million PBMCs per donor using the AllPrep DNA/RNA Kit (Qiagen). AIRR-seq libraries
525 were generated using a 5' Rapid Amplification of cDNA Ends (RACE) approach. For IGK and IGL 5'RACE
526 AIRR-seq, libraries were produced using the SMARTer Human BCR Profiling Kit (Takara Bio, San Jose, CA,
527 USA), according to the manufacturer's instructions. Quality and quantity of individually indexed IGK libraries
528 were determined using the 2100 Bioanalyzer High Sensitivity DNA Assay Kit (Agilent, Santa Clara, CA, USA)
529 and IGL libraries with the Qubit 3.0 Fluorometer dsDNA High Sensitivity Assay Kit (Life Technologies,
530 Carlsbad, CA, USA). Libraries were pooled at 10 nM and sequenced on the Illumina NextSeq system using

531 300 bp paired-end reads with the 600-cycle NextSeq P1 Reagent Kit (n=161) or with the 600-cycle MiSeq
532 Reagent Kit v3 (n=11) (Illumina, San Diego, CA, USA).
533
534 **Construction of a custom linear reference assembly**
535 We used our previously described custom linear reference [15], which includes modifications to the IGH
536 [12,13] and IGK [15] loci to include sequences not present in the GRCh38 assembly. To include IGL sequences not
537 present in GRCh38, chromosome 22 was removed and replaced with the T2T (CHM13v2.0) chromosome 22,
538 including the IGLV5-39 structural variant (SV, insertion) sequence. In addition, 46,423 bp of sequence
539 (chr22:23315600-23362023) is from a Human Pangenome Reference Consortium (HPRC) haplotype from
540 sample HG00621, which includes a 16,093 bp insertion relative to the CHM13v2.0 reference, reflecting 3
541 additional copies of the IGLJ-C3 cassette relative to the CHM13v2.0 reference. This reference is publicly
542 available at https://github.com/Watson-IG/immune_receptor_genomics/tree/main.
543
544 **Phased assembly of IGK and IGL**
545 Phased assemblies were generated as described previously [15]. HiFi reads were used to generate
546 haplotype-phased *de novo* (i.e. reference-agnostic) assemblies using hifiasm [64] (v0.18.2-r467) with default
547 parameters. For each sample, hifiasm contigs were concatenated into a FASTA file, then redundant contigs
548 were filtered out using the seqkit toolkit [65] command 'seqkit rmdup --by-seq <hifiasm_contigs.fasta>' (seqkit
549 v2.4.0). Hifiasm contigs were mapped to the custom reference assembly using minimap2 (v2.26) with the '-x
550 asm20' option. HiFi reads were also processed using IGenotyper [13]; the programs "phase" and "assemble"
551 were run with default parameters to generate phased contigs and HiFi read alignments to the custom
552 reference.
553 For each sample, aligned HiFi reads as well as aligned hifiasm-generated contigs and IGenotyper-
554 generated contigs were viewed in the Integrative Genomics Viewer (IGV) application [66] for manual selection of
555 phased contigs. Contigs were evaluated for read support from mapped HiFi reads, and contigs harboring one
556 or more SNVs that lacked read support were not selected during manual curation. Where a hifiasm and an
557 IGenotyper contig were identical throughout a phased block, the hifiasm contig was selected, as described
558 previously [15]. Curated, phased assemblies were aligned to the custom reference using minimap2 [67] with the '-x
559 asm20' option.
560 To assess accuracy of manually curated assemblies, personalized references were first generated by
561 N-masking the IGK (chr2:88837160-90280100) and IGL (chr22:22378775-23423320) loci of the custom
562 reference and, for each sample, appending the reference FASTA with IGK and IGL curated contigs. All HiFi
563 reads from each individual were aligned to the corresponding personalized reference using minimap2 with the
564 '-x map-hifi' preset; coverage and read length metrics were extracted from these alignments. Positions in
565 assemblies with > 25% of aligned HiFi reads mismatching the assembly were identified by parsing the output
566 of samtools 'mpileup' using custom scripts; assembly accuracy was determined using the formula [total

567 (diploid) bases without a mismatch / total (diploid) assembly length (bp)] * 100 = % accuracy (**Supplementary**
568 **Table S1**).
569
570 **Identification of IGK and IGL gene alleles**
571 Sequences corresponding to V, J, and C exons were obtained from assembly BAM files and aligned to
572 all known human immunoglobulin alleles from the IMGT database (version downloaded on 2023-09-26) using
573 custom Python scripts and functions provided by the receptor_utils library (https://pypi.org/project/receptor-
574 utils/). Metrics of HiFi read support for each allele were calculated using custom scripts written in Python and
575 bash. Briefly, for each sample, curated IGK and IGL assembly contigs were appended to our custom reference
576 with the IGK and IGL loci N-masked to generate a personalized reference FASTA, as described above. All HiFi
577 reads from IG-capture for a given individual were mapped to the personalized reference using minimap2 with
578 the '-ax map-hifi' preset, then the resulting BAM file was input to samtools 'mpileup' with iteration over allele
579 coordinates in the personalized reference. The outputs of this script are in **Supplementary Table S3** (IGK) and
580 **Supplementary Table S4** (IGL), and include a column for total number of HiFi reads spanning the novel allele
581 ('Fully_Spanning_Reads') and a column for the number of HiFi reads spanning the novel allele with 100%
582 sequence identity ('Fully_Spanning_Reads_100%_Match'). A complete description of HiFi read support
583 metrics for alleles is available at https://vdjbase.org/.
584
585 **Genetic ancestry**
586 IGenotyper [13] was used to call SNVs at ancestry-informative markers (AIMs) by aligning, phasing, and
587 locally assembling reads at AIM regions, then directly identifying SNVs from the assembled sequences.
588 Genetic ancestry was determined using these AIMs and the STRUCTURE program [68]. SNV VCFs were
589 processed to extract AIM-specific data from IG-Capture libraries using custom scripts and VCFtools. Coverage
590 of AIMs was assessed using BAM files and the pysam library, ensuring a minimum read depth threshold for
591 inclusion. Genotypes were converted into haplotypes by separating phased alleles, and samples were coded
592 alongside reference populations from the 1000 Genomes Project. STRUCTURE (v2.3.4) was run with K = 5,
593 representing five global ancestry groups (European, African, East Asian, South Asian, and American), using
594 default admixture and allele frequency models. For each sample, the two highest ancestry proportions were
595 identified. If the difference between these two proportions was less than or equal to 5%, the sample was
596 classified as "Mixed." Otherwise, the sample was assigned to the ancestry category with the highest
597 proportion.
598
599 **Processing AIRR-seq data and calculating gene usage**
600 Paired-end sequences ("R1" and "R2") were processed using the pRESTO/Change-O toolkit [69,70]. All
601 R1 and R2 reads were trimmed to Q = 20 using the function "FilterSeq.py trimqual". Constant region (IGKC and
602 IGLC) primers were identified with an error rate of 0.3 and corresponding chains were recorded in the fastq

603 headers using “MaskPrimers align.” The 12 base UMI, located directly after the constant region primer, was
604 extracted using “MaskPrimers extract.” Annotations between mate pairs, including UMI barcodes and constant
605 region calls, were synchronized using PairSeq.py to sort reads into mate pairs and remove unpaired reads.
606 UMI groups sharing the same barcode were processed to generate consensus sequences using the
607 BuildConsensus.py function. The following criteria were applied: a minimum UMI group size of one, a
608 maximum mismatch error rate of 10%, and at least 60% agreement on the constant region call within the
609 group. Reads with lower-quality consensus sequences (Q < 30) were masked using FilterSeq.py maskqual.
610 Duplicate counts (“Dupcounts”) were recorded, and duplicate sequences were collapsed using CollapseSeq.py
611 to retain one representative sequence per cell, with the total number of sequences contributing to each
612 consensus recorded as “Conscount.” Collapsed consensus sequences supported by fewer than two
613 contributing reads (Conscount < 2) were discarded using SplitSeq.py. Samples containing fewer than 100
614 unique sequences were excluded from downstream analysis; after filtering, 170 individuals were included for
615 IGK analyses and 170 individuals were included for IGL analyses (173 individuals total; **Supplementary Table**
616 **S1**). After processing, the repertoires contained a mean of 33,678 unique BCR sequences (IGK) and 13,281
617 (IGL) (**Supplementary Table S1, Supplementary Fig. 3**).
618 Germline allele designations were assigned to sequences using a personalized allele database during
619 the IgBLAST step. For each individual, IGK and IGL germline allele databases were generated from the set of
620 alleles identified in genomic assemblies derived from long-read sequencing. Separate BLAST databases were
621 created for V and J segments using makeblastdb. The resulting databases were used as input for igblastn, and
622 IgBLAST output files (“Change-O” tables) of unique BCRs were generated for IGK and IGL separately. This
623 process permitted theoretical disambiguation of IGK gene paralogs for individuals wherein the sequence of
624 each allele of a proximal paralog was distinct from the sequence of each allele of the distal paralog (results in
625 **Supplementary Table S5**). In the case of IGKV1-13 and IGKV1D-13, a subset of the cohort (n=114) met this
626 disambiguation criteria and was carried forward to identify guQTLs for these genes.
627 Due to sequence identity between IGKV paralog allele sequences, genes collapsed into a single
628 ambiguous (“ambi”) entity in Change-O tables included:
629 1) IGKV1-37 and IGKV1D-37 replaced with IGKV1-37ambi
630 2) IGKV1-39 and IGKV1D-39 with IGKV1-39ambi
631 3) IGKV2-40 and IGKV2D-40 with IGKV2-40ambi
632 4) IGKV1-33 and IGKV1D-33 with IGKV1-33ambi
633 5) IGKV2-28 and IGKV2D-28 with IGKV2-28ambi
634 In addition, IGLJ gene calls that corresponded to a IGLJ2 or IGLJ3 cassette (IGLJ2, IGLJ3-1, IGLJ3-2, IGLJ3-
635 3, IGLJ3-4) were collapsed to “IGLJ2-3ambi”.
636 Change-O tables of unique BCRs were analyzed using the alakazam package [70]. To enrich for antigen-
637 naïve BCRs, only unmutated light chain sequences with 100% identity to the assigned germline V and J alleles
638 were included in downstream analyses. For analysis of mutated sequences (i.e. the fraction of sequences

639 enriched with those containing SHM), sequences with less than 100% identity to either the germline V or J
640 allele were included. Gene usage was quantified for IGK and IGL light chains using the countGenes function
641 with the parameters gene = "v_call" and "j_call", groups = "sample_id", mode = "gene", and genes with a
642 sequence count (seq_count) of at least 10 in at least one sample were retained. A m × n usage matrix C was
643 created, where m are the genes and n are the samples. Each value in C represented usage frequency among
644 all unique (unmutated or mutated) BCR sequences for a given gene in a given sample.
645 CDR3 physicochemical properties were computed using the aminoAcidProperties function (47) with
646 options seq = "junction", trim = TRUE, label = "cdr3"; resulting values were then averaged across all
647 sequences within each sample to obtain sample-level means.
648

649 **Selecting common variants for gene usage QTL analysis**

650 SNVs were genotyped from curated assemblies as described previously [15]. Variants were called from
651 assemblies using 'bcftools mpileup' (bcftools v1.15.1) with options '-f -B -a QS', then 'bcftools call' with options
652 '-m --ploidy 2'. A mutli-sample VCF file was generated using 'bcftools merge' with the '-m both' option.
653 Multiallelic SNVs were split into biallelic records using the 'bcftools norm' command with options '-a -m-'. The
654 VCF file was annotated for V-exon, introns, L-Part1, RSS sequences (heptamer, nonamer, spacer), were
655 added using vcfanno and BED files corresponding to our reference (available at https://github.com/Watson-
656 IG/immune_receptor_genomics/). Biallelic SNVs with MAF ≥ 5% were selected using bcftools view with options
657 "-m2 -M2 -v snps -i 'INFO/MAF >= 0.05'" and used for guQTL analysis.
658 All SVs in IGKV and IGLV gene regions were genotyped by manual inspection using IGV [66]. SVs with a
659 MAF less than 0.05 were not included in the guQTL analysis.
660

661 **Gene usage QTL analysis**

662 Genotypes at common SNVs and SVs were tested for association with usage using linear regression to
663 determine significance and additional metrics (e.g., beta coefficients and R^2 values). To adjust for multiple
664 comparisons, a Bonferroni correction was applied on a per-gene basis. Pairwise r^2 (LD) values were computed
665 using vcftools '--geno-r2' [71]. Variants in complete linkage disequilibrium (LD r^2=1) were considered as a single
666 variant during correction, representing only one association test.
667

668 **Haplotype block analysis**

669 LD blocks were computed and visualized using "LDBlockShow" [39] using the multi-sample VCF of
670 common SNVs (MAF >= 5%) as input, with options '-SeleVar 2 -BlockType 1' to use normalized linkage
671 disequilibrium coefficients (D') as described by described by Gabriel et al. (2002) [40] to determine haplotype
672 blocks. The program was run using '-Region chr2:88837160-90280100' for IGK and '-Region chr22:22378775-

673 23423320' for IGL. LD block boundaries, sizes, and SNVs within blocks are included in **Supplementary Table**
674 **S9**. Genes overlapping LD blocks were identified using bedtools [72] 'intersect' (bedtools v2.30.0).
675
676 **Network analysis**
677 Variants significantly associated with IGK and IGL gene usage (after Bonferroni correction) were
678 compiled to create a set of guQTL variants for each gene. Pairwise comparisons between genes were
679 performed to calculate Jaccard similarity indices, defined as the ratio of shared variants to the total unique
680 variants across the two genes being compared. Only gene pairs with nonzero similarity (i.e., at least one
681 shared variant) were included in the analysis. Separate IGK and IGL network graphs were constructed and
682 visualized using igraph [73] and ggraph [74] (IGL),. In these graphs, nodes represented individual genes and were
683 positioned using the Fruchterman-Reingold force-directed algorithm, edges connected pairs of genes that
684 shared guQTL variants, and edges were labeled to reflect the number of shared guQTL variants.
685
686 **Regulatory analysis**
687 ENCODE transcription factor binding site data were obtained from the UCSC Genome Browser under the
688 group "Regulation," track "TF Clusters," and table "encRegTfbsClustered." SNVs associated with gene usage
689 were overlapped with this track and enrichment over all SNVs overlapping each track was calculated using a
690 one-sided Fisher Exact Test (**Supplementary Table S8**).
691
692 **Data availability**
693 Long-read sequencing data and AIRR-seq datasets generated in this study have been deposited in the
694 BioProject repository PRJNA1274485 . Previously published AIRR-seq datasets are available in the BioProject
695 repository PRJNA555323. Metadata and sequencing summary statistics for this study are provided in
696 Supplementary Table S1.
697
698 **Funding**
699 This work was supported by grant R24AI138963 to C.T.W and M.L.S. from the National Institute of Allergy and
700 Infectious Disease.
701
702 **Author Contributions**
703 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
704 lab/bench and sequencing experiments. E.E., D.S., K.S., S.S. prepared sequencing libraries. E.E., U.J.,
705 O.L.R., W.L., W.S.G., A.P., G.Y. and Z.V. developed the computational pipeline to analyze the sequencing
706 data. E.E., G.Y., C.T.W., M.L.S contributed to and supervised the analysis and data interpretation. C.T.W. and
707 M.L.S. acquired funding. E.E. and C.T.W. drafted the manuscript and all authors contributed to revising the
708 manuscript.

709

710 **Competing interests**

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

713

714 **References**

715 1. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the
716 baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).

717 2. Soto, C. *et al.* High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566**, 398–
718 402 (2019).

719 3. Boyd, S. D. *et al.* Individual variation in the germline Ig gene repertoire inferred from variable region gene
720 rearrangements. *J. Immunol.* **184**, 6986–6992 (2010).

721 4. Röltgen, K. *et al.* Defining the features and duration of antibody responses to SARS-CoV-2 infection
722 associated with disease severity and outcome. *Sci. Immunol.* **5**, eabe0240 (2020).

723 5. Wahala, W. M. P. B. & Silva, A. M. de. The human antibody response to dengue virus infection. *Viruses* **3**,
724 2374–2395 (2011).

725 6. Overbaugh, J. & Morris, L. The antibody response against HIV-1. *Cold Spring Harb. Perspect. Med.* **2**,
726 a007039 (2012).

727 7. Krammer, F. The human antibody response to influenza A virus infection and vaccination. *Nat. Rev.*
728 *Immunol.* **19**, 383–397 (2019).

729 8. Muñoz-Durango, N. *et al.* Patterns of antibody response during natural hRSV infection: insights for the
730 development of new antibody-based therapies. *Expert Opin. Investig. Drugs* **27**, 721–731 (2018).

731 9. Del Pozo-Yauner, L. *et al.* Role of the mechanisms for antibody repertoire diversification in monoclonal
732 light chain deposition disorders: when a friend becomes foe. *Front. Immunol.* **14**, 1203425 (2023).

733 10. Sharonov, G. V., Serebrovskaya, E. O., Yuzhakova, D. V., Britanova, O. V. & Chudakov, D. M. B cells,
734 plasma cells and antibody repertoires in the tumour microenvironment. *Nat. Rev. Immunol.* **20**, 294–307
735 (2020).

736 11. Murphy, K. & Weaver, C. *Janeway's Immunobiology*. (Garland Science, 2016).

737 12. Rodriguez, O. L. *et al.* Genetic variation in the immunoglobulin heavy chain locus shapes the human
738 antibody repertoire. *Nat. Commun.* **14**, 4419 (2023).
739 13. Rodriguez, O. L. *et al.* A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human
740 Immunoglobulin Heavy Chain Locus. *Front. Immunol.* **11**, 2136 (2020).
741 14. Gibson, W. S. *et al.* Characterization of the immunoglobulin lambda chain locus from diverse populations
742 reveals extensive genetic variation. *Genes Immun.* **24**, 21–31 (2023).
743 15. Engelbrecht, E. *et al.* Resolving haplotype variation and complex genetic architecture in the human
744 immunoglobulin kappa chain locus in individuals of diverse ancestry. *Genes Immun.* (2024)
745 doi:10.1038/s41435-024-00279-2.
746 16. Glanville, J. *et al.* Naive antibody gene-segment frequencies are heritable and unaltered by chronic
747 lymphocyte ablation. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20066–20071 (2011).
748 17. Rubelt, F. *et al.* Individual heritable differences result in unique cell lymphocyte receptor repertoires of
749 naïve and antigen-experienced cells. *Nat. Commun.* **7**, 11112 (2016).
750 18. Wang, C. *et al.* B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc.*
751 *Natl. Acad. Sci. U. S. A.* **112**, 500–505 (2015).
752 19. Parks, T. *et al.* Association between a common immunoglobulin heavy chain allele and rheumatic heart
753 disease risk in Oceania. *Nat. Commun.* **8**, 14946 (2017).
754 20. Pushparaj, P. *et al.* Immunoglobulin germline gene polymorphisms influence the function of SARS-CoV-2
755 neutralizing antibodies. *Immunity* **56**, 193–206.e7 (2023).
756 21. Avnir, Y. *et al.* IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with
757 IGHV utilization shifts and varies by ethnicity. *Sci. Rep.* **6**, 20842 (2016).
758 22. Sangesland, M. *et al.* Allelic polymorphism controls autoreactivity and vaccine elicitation of human broadly
759 neutralizing antibodies against influenza virus. *Immunity* **55**, 1693–1709.e8 (2022).
760 23. Mashimo, Y. *et al.* Germline variants of IGHV3-53 / V3-66 are determinants of antibody responses to the
761 BNT162b2 mRNA COVID-19 vaccine. *J. Infect.* **85**, 702–769 (2022).
762 24. Lee, J. H. *et al.* Vaccine genetics of IGHV1-2 VRC01-class broadly neutralizing antibody precursor naïve
763 human B cells. *NPJ Vaccines* **6**, 113 (2021).

764 25. Sonehara, K. *et al.* Germline variants and mosaic chromosomal alterations affect COVID-19 vaccine
765 immunogenicity. *Cell Genom.* 100783 (2025).
766 26. deCamp, A. C. *et al.* Human immunoglobulin gene allelic variation impacts germline-targeting vaccine
767 priming. *NPJ Vaccines* **9**, 58 (2024).
768 27. Mikocziova, I. *et al.* Germline polymorphisms and alternative splicing of human immunoglobulin light chain
769 genes. *iScience* **24**, 103192 (2021).
770 28. Feeney, A. J., Atkinson, M. J., Cowan, M. J., Escuro, G. & Lugo, G. A defective Vkappa A2 allele in
771 Navajos which may play a role in increased susceptibility to haemophilus influenzae type b disease. *J.*
772 *Clin. Invest.* **97**, 2277–2282 (1996).
773 29. Shrock, E. L. *et al.* Germline-encoded amino acid-binding motifs drive immunodominant public antibody
774 responses. *Science* **380**, eadc9498 (2023).
775 30. Rodriguez, O. L. *et al.* Human genetic variation shapes the antibody repertoire across B cell development.
776 *Genomics* (2025).
777 31. Watson, C. T. *et al.* Sequencing of the human IG light chain loci from a hydatidiform mole BAC library
778 reveals locus-specific signatures of genetic diversity. *Genes Immun.* **16**, 24–34 (2015).
779 32. Watson, C. T. *et al.* Complete haplotype sequence of the human immunoglobulin heavy-chain variable,
780 diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.*
781 **92**, 530–546 (2013).
782 33. Jana, U. *et al.* The human immunoglobulin heavy chain constant gene locus is enriched for large complex
783 structural variants and coding polymorphisms that vary in frequency among human populations. *Genomics*
784 (2025).
785 34. Engelbrecht, E. T., Rodriguez, O. L. & Watson, C. T. Addressing technical pitfalls in pursuit of molecular
786 factors that mediate immunoglobulin gene regulation. *bioRxiv* 2024–2003 (2024).
787 35. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues.
788 *Science* **369**, 1318–1330 (2020).
789 36. Ye, Y., Zhang, Z., Liu, Y., Diao, L. & Han, L. A multi-omics perspective of quantitative trait loci in precision
790 medicine. *Trends Genet.* **36**, 318–336 (2020).

791 37. Yamaguchi, K. *et al.* Splicing QTL analysis focusing on coding sequences reveals mechanisms for
792 disease susceptibility loci. *Nat. Commun.* **13**, 4659 (2022).
793 38. van der Wijst, M. *et al.* The single-cell eQTLGen consortium. *Elife* **9**, (2020).
794 39. Dong, S.-S. *et al.* LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and
795 haplotype blocks based on variant call format files. *Brief. Bioinform.* **22**, (2021).
796 40. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229
797 (2002).
798 41. Watson, C. T., Glanville, J. & Marasco, W. A. The Individual and Population Genetics of Antibody
799 Immunity. *Trends Immunol.* **38**, 459–470 (2017).
800 42. Gidoni, M. *et al.* Mosaic deletion patterns of the human antibody heavy chain gene locus shown by
801 Bayesian haplotyping. *Nat. Commun.* **10**, 628 (2019).
802 43. Atkinson, M. J., Cowan, M. J. & Feeney, A. J. New alleles of IGKV genes A2 and A18 suggest significant
803 human IGKV locus polymorphism. *Immunogenetics* **44**, 115–120 (1996).
804 44. Nadel, B. *et al.* Decreased frequency of rearrangement due to the synergistic effect of nucleotide changes
805 in the heptamer and nonamer of the recombination signal sequence of the V kappa gene A2b, which is
806 associated with increased susceptibility of Navajos to Haemophilus influenzae type b disease. *J. Immunol.*
807 **161**, 6068–6073 (1998).
808 45. Choi, N. M. *et al.* Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom
809 V gene rearrangement frequencies. *J. Immunol.* **191**, 2393–2402 (2013).
810 46. Bolland, D. J. *et al.* Two mutually exclusive local chromatin states drive efficient V(D)J recombination. *Cell*
811 *Rep.* **15**, 2475–2487 (2016).
812 47. Bhat, K. H. *et al.* An Igh distal enhancer modulates antigen receptor diversity by determining locus
813 conformation. *Nat. Commun.* **14**, 1225 (2023).
814 48. Barajas-Mora, E. M. *et al.* Enhancer-instructed epigenetic landscape and chromatin compartmentalization
815 dictate a primary antibody repertoire protective against specific bacterial pathogens. *Nat. Immunol.* **24**,
816 320–336 (2023).
817 49. Kenter, A. L., Watson, C. T. & Spille, J.-H. Igh locus polymorphism may dictate topological chromatin

818 conformation and V gene usage in the ig repertoire. *Front. Immunol.* **12**, 682589 (2021).
819 50. Choi, N. M. & Feeney, A. J. CTCF and ncRNA regulate the three-dimensional structure of antigen receptor
820 loci to facilitate V(D)J recombination. *Front. Immunol.* **5**, 49 (2014).
821 51. Nemazee, D. Mechanisms of central tolerance for B cells. *Nat. Rev. Immunol.* **17**, 281–294 (2017).
822 52. Melchers, F. Checkpoints that control B cell development. *J. Clin. Invest.* **125**, 2203–2210 (2015).
823 53. Lee, S., Ko, Y. & Kim, T. J. Homeostasis and regulation of autoreactive B cells. *Cell. Mol. Immunol.* **17**,
824 561–569 (2020).
825 54. Dudzic, P. *et al.* Conserved heavy/light contacts and germline preferences revealed by a large-scale
826 analysis of natively paired human antibody sequences and structural data. *Immunology* (2024).
827 55. Jaffe, D. B. *et al.* Functional antibodies exhibit light chain coherence. *Nature* **611**, 352–357 (2022).
828 56. Pospelova, M. *et al.* Comparative analysis of mammalian adaptive immune loci revealed spectacular
829 divergence and common genetic patterns. *Genomics* (2025).
830 57. Peng, K. *et al.* Diversity in immunogenomics: the value and the challenge. *Nat. Methods* **18**, 588–591
831 (2021).
832 58. Corcoran, M. M. *et al.* Production of individualized V gene databases reveals high levels of
833 immunoglobulin genetic diversity. *Nat. Commun.* **7**, 13642 (2016).
834 59. Marsden, A. A. *et al.* Novel polymorphic and copy number diversity in the antibody IGH locus of South
835 African individuals. *Immunogenetics* **77**, 6 (2024).
836 60. Castro Dopico, X., Mandolesi, M. & Karlsson Hedestam, G. B. Untangling associations between
837 immunoglobulin genotypes, repertoires and function. *Immunol. Lett.* **259**, 24–29 (2023).
838 61. Pushparaj, P. *et al.* Frequent use of IGHV3-30-3 in SARS-CoV-2 neutralizing antibody responses. *Front.*
839 *Virol.* **3**, 1128253 (2023).
840 62. Yeung, Y. A. *et al.* Germline-encoded neutralization of a Staphylococcus aureus virulence factor by the
841 human antibody repertoire. *Nat. Commun.* **7**, 13376 (2016).
842 63. Kirik, U., Persson, H., Levander, F., Greiff, L. & Ohlin, M. Antibody heavy chain variable domains of
843 different germline gene origins diversify through different paths. *Front. Immunol.* **8**, 1433 (2017).
844 64. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using

845 phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

846 65. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File
847 Manipulation. *PLoS One* **11**, e0163962 (2016).

848 66. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

849 67. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

850 68. Kosoy, R. *et al.* Ancestry informative marker sets for determining continental origin and admixture
851 proportions in common populations in America. *Hum. Mutat.* **30**, 69–78 (2009).

852 69. Vander Heiden, J. A. *et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of
853 lymphocyte receptor repertoires. *Bioinformatics* **30**, 1930–1932 (2014).

854 70. Gupta, N. T. *et al.* Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire
855 sequencing data. *Bioinformatics* **31**, 3356–3358 (2015).

856 71. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

857 72. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.
858 *Bioinformatics* **26**, 841–842 (2010).

859 73. Csardi, G. & Nepusz, T. *The Igraph Software Package for Complex Network Research*. (2006).

860 74. Pedersen, T. *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. (2024).

861
862
863
864
865
866
867
868