

Supplementary Information for
Germline polymorphisms in the immunoglobulin kappa and lambda loci underpinning
antibody light chain repertoire variability

Eric Engelbrecht¹, Oscar L. Rodriguez^{1,2}, William Lees³, Zach Vanwinkle¹, Kaitlyn Shields¹, Steven Schultze¹, William S. Gibson¹, David R. Smith¹, Uddalok Jana¹, Swati Saha¹, Ayelet Peres⁴, Gur Yaari⁴, 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

For correspondence: corey.watson@louisville.edu

- This file includes:**
- Supplementary Figures
 - Supplementary Notes
 - Supplementary References

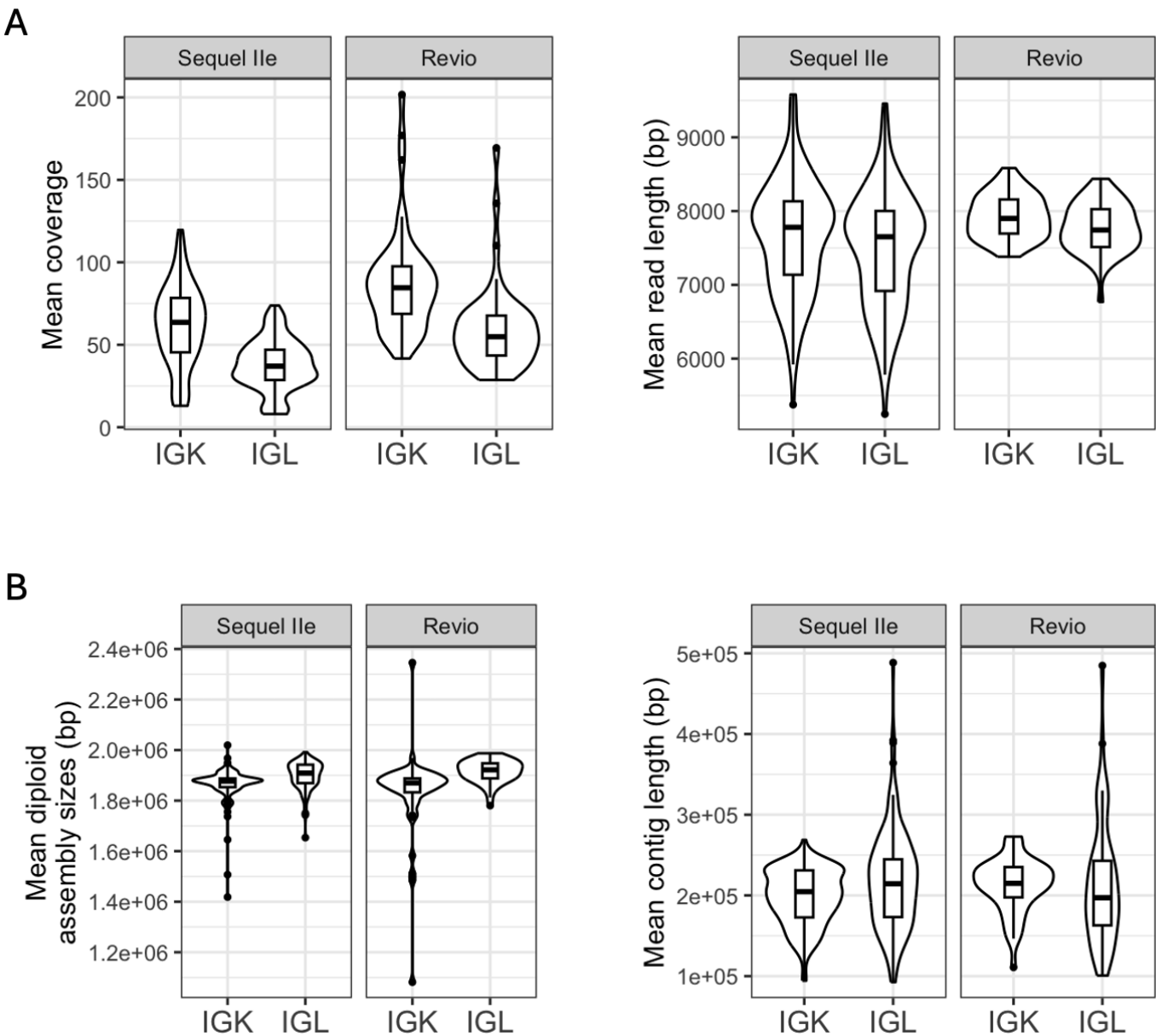


Figure S1. PacBio sequencing and assembly statistics.

For both IGK and IGL, the statistics described include (A) locus coverage (left), read length (right), (B) assembly lengths (left), and contig lengths (right). Each statistic is separated by PacBio platforms. The number of sequencing runs on the Sequel IIe and Revio platforms is 134 and 43, respectively. Violin plots with overlaid 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.

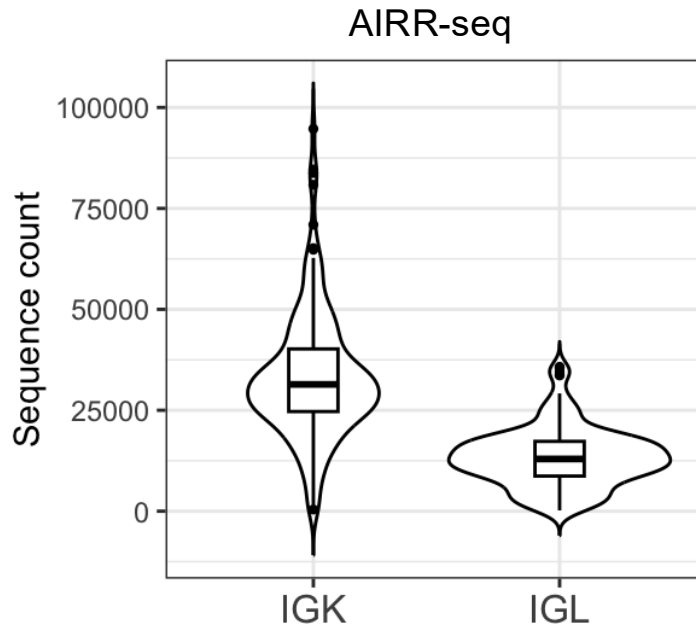
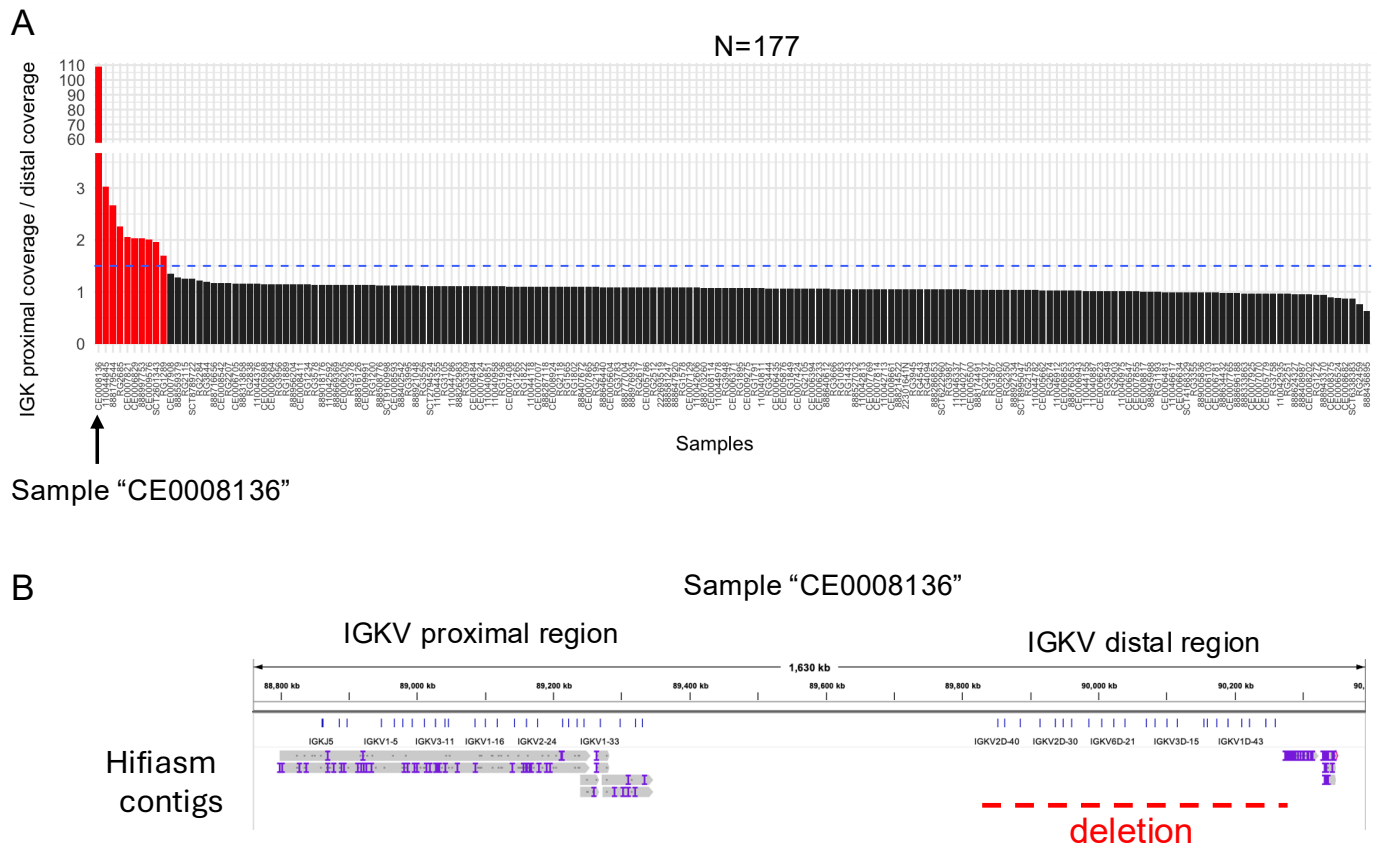


Figure S3. Number of merged reads for each AIRR-seq dataset after processing.

Number of unique V-J sequences per-sample after filtering duplicate reads for IGK (n=164) and IGL (n=168) AIRR-seq. Violin plots with overlaid 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.

134



135

136

137

138

139

140

141

142

143

144

145

Figure S4. 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-generated contigs for sample CE0008136 aligned to the IGK locus; no contigs map to the IGKV distal region.

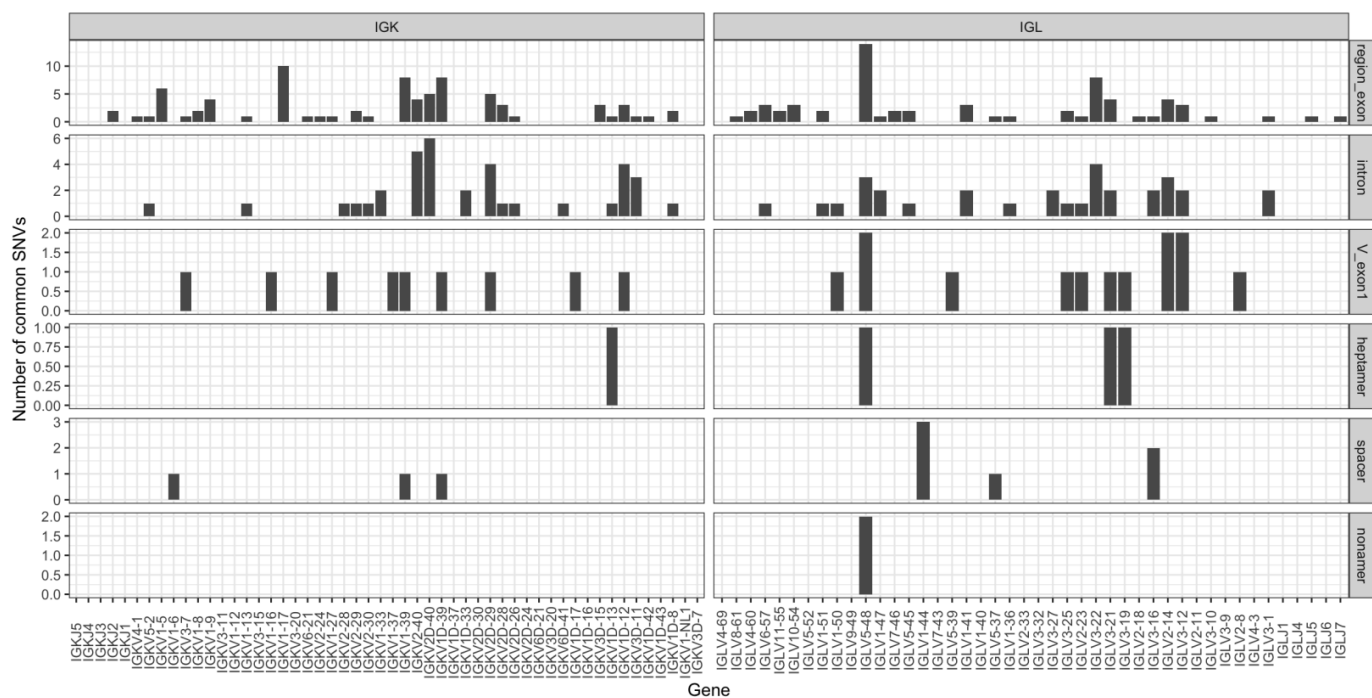
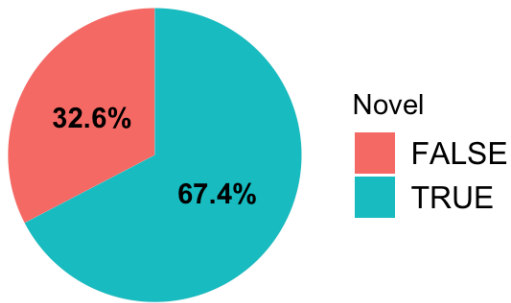


Figure S5. Common SNVs in IGK and IGL in genic and RSS regions.

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

178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218

All unique IGK alleles



All unique IGL alleles

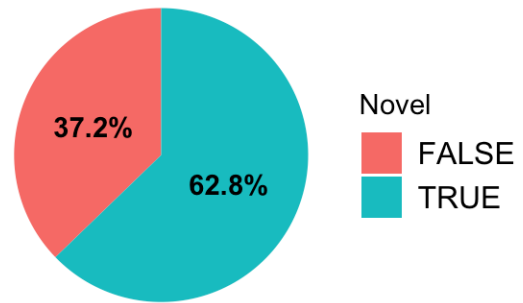


Figure S6. Many IGK and IGL gene alleles are not catalogued in IMGT.

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

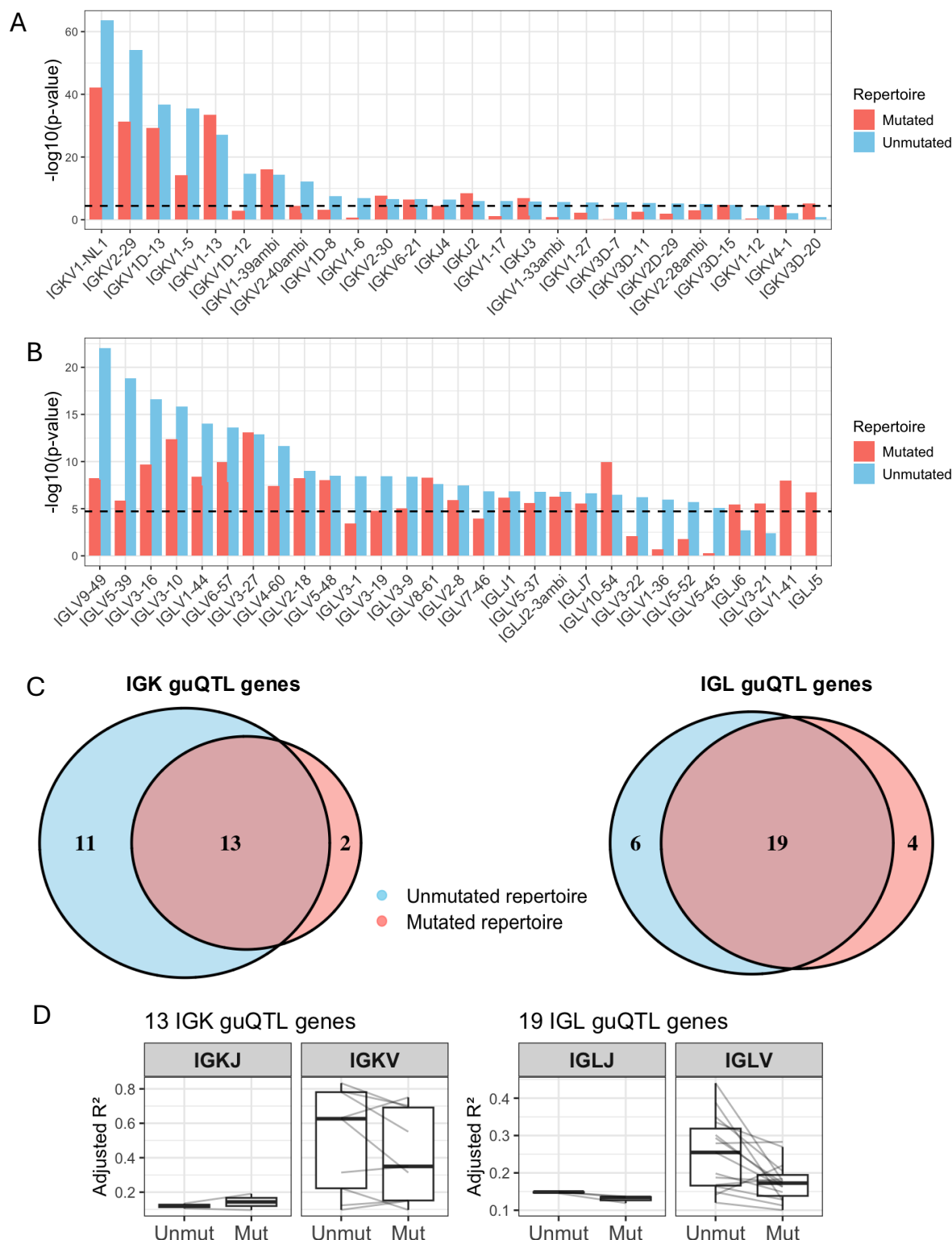


Figure S7. guQTLs in naïve (unmutated) and antigen-experienced (mutated) Ab repertoires. (A-B) Barplots show the strength of associations for lead guQTLs in the unmutated (naïve) and mutated (antigen-experienced) IGK (A) and IGL (B) repertoires. Plots include only guQTL genes. (C) Venn diagrams indicating the number of unique and overlapping guQTL genes in unmutated and mutated IGK and IGL Ab repertoires (see also **Supplementary Data 7**). (D) The variance explained (adjusted R^2) for guQTL genes identified in both unmutated and mutated Ab repertoires for IGK (13 genes) and IGL (19 genes). Gray lines connect the variance explained in the unmutated and mutated repertoires.

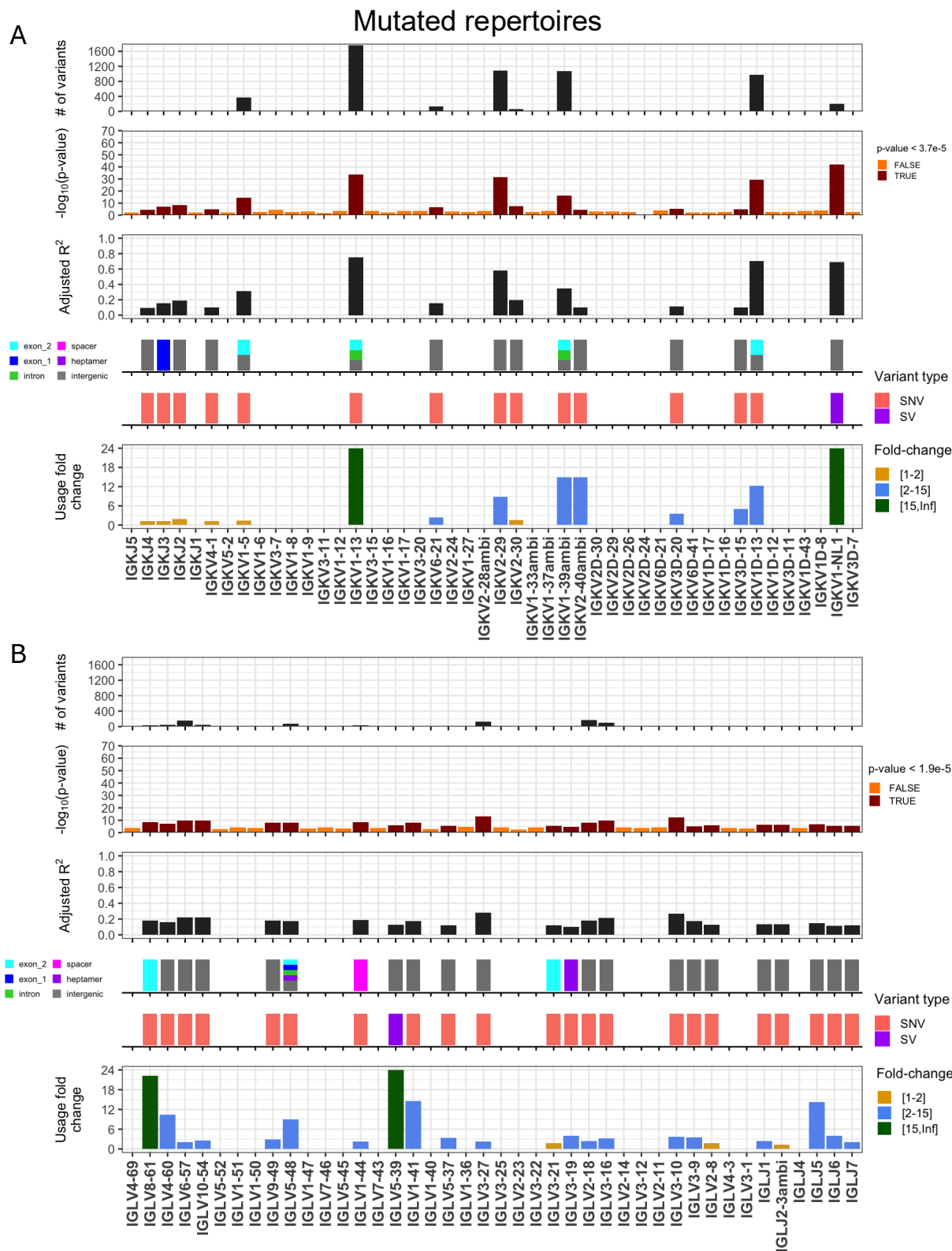


Figure S8. guQTLs in antigen-experienced (mutated) Ab repertoires.

(A-B) Per gene (x axis, all panels) statistics from linear regression guQTL analysis for the repertoire of mutated IGHK (A) and IGL (B) light chains, including: (i) the number of associated variants after Bonferroni correction (IGHK; $P < 3.7e-5$, IGL; $P < 1.9e-5$), (ii) $-\log_{10}(P \text{ value})$ of the lead guQTL, (iii) adjusted R^2 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 Data 6**.

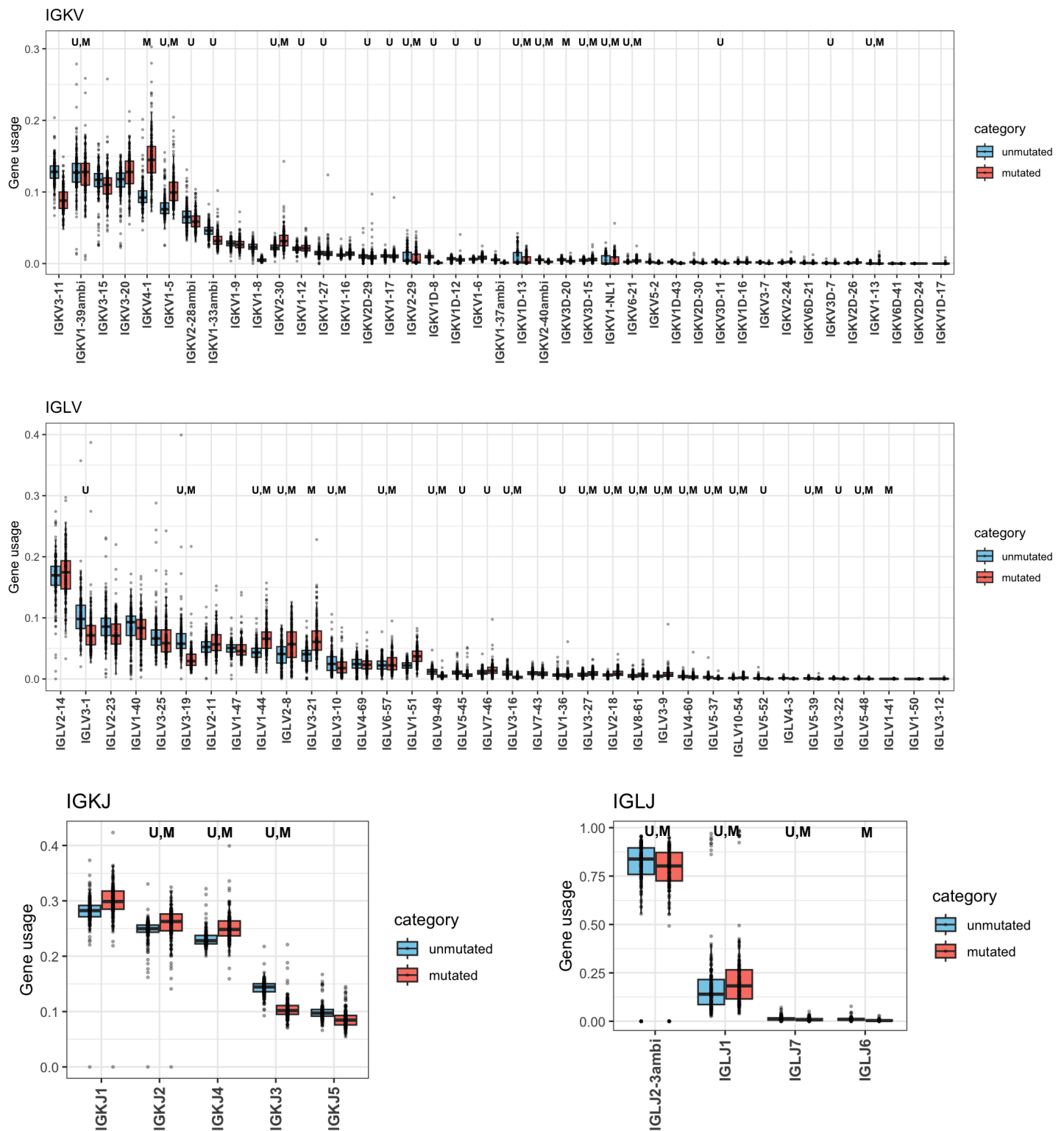
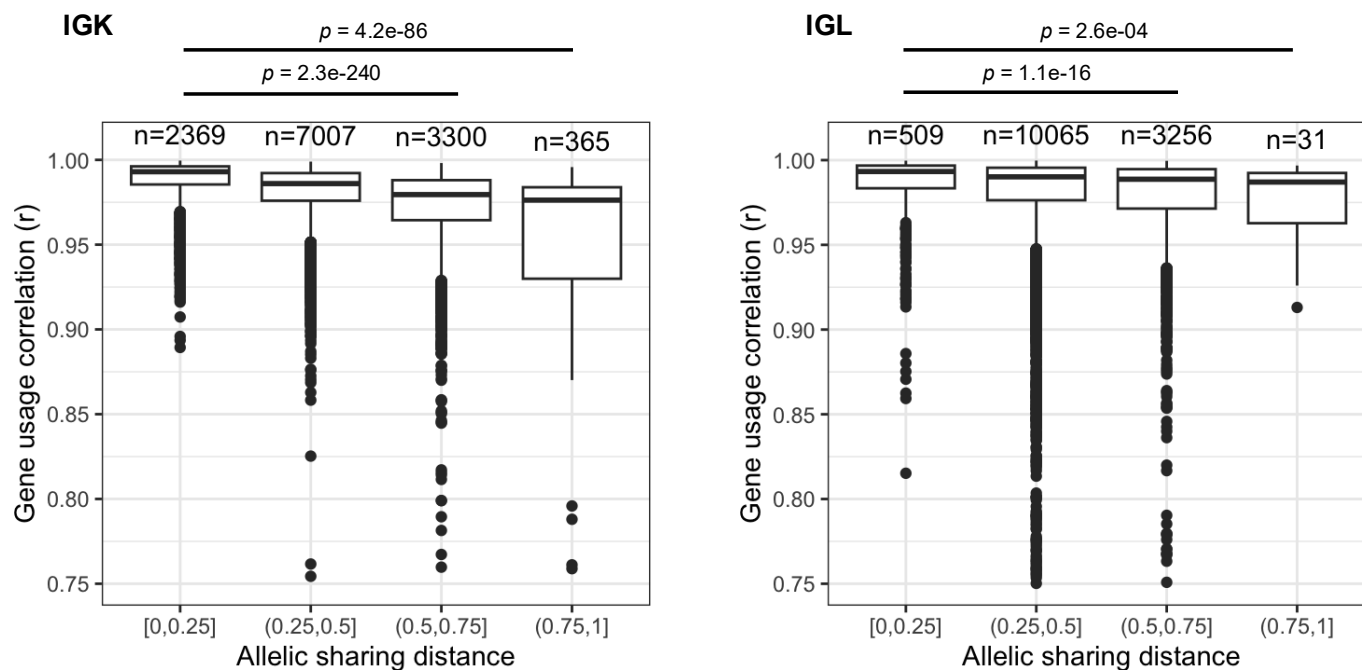


Figure S9. Usage of IGK and IGL guQTL and non-guQTL genes in unmutated and mutated Ab repertoires. Each panel shows usage of V or J genes in IGK or IGL. guQTL genes in unmutated and mutated repertoires are annotated “U” and “M”, respectively, with “U,M” indicating guQTL genes in both unmutated and mutated repertoires.

344
345
346
347
348



349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367

Figure S10. 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.

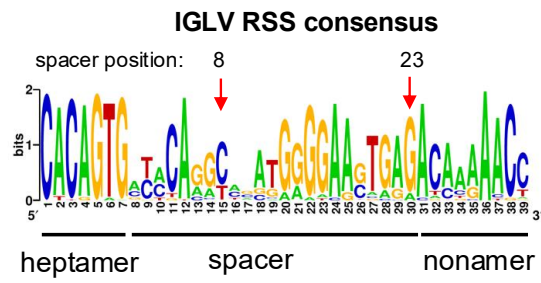


Figure S11. IGLV RSS consensus.

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

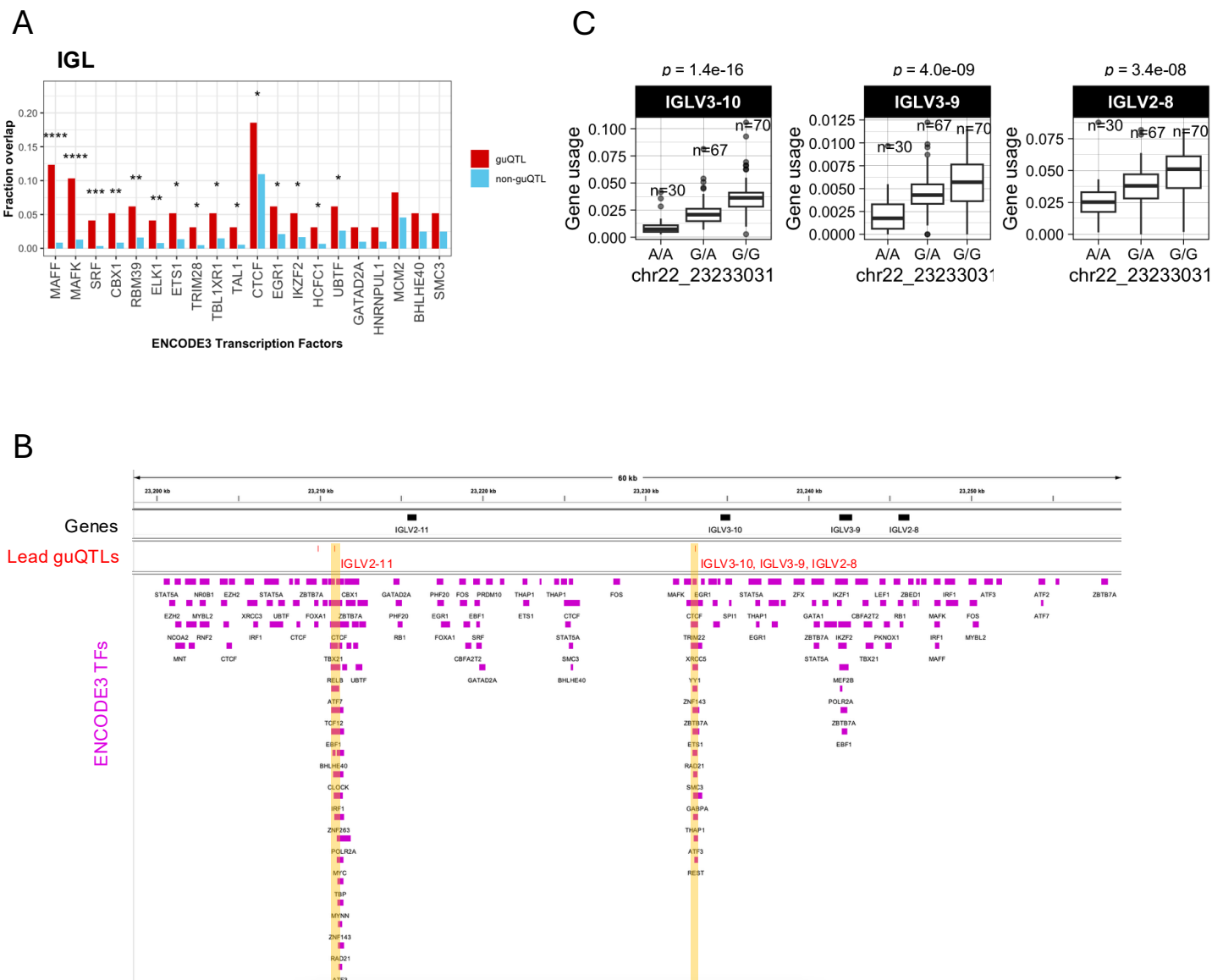
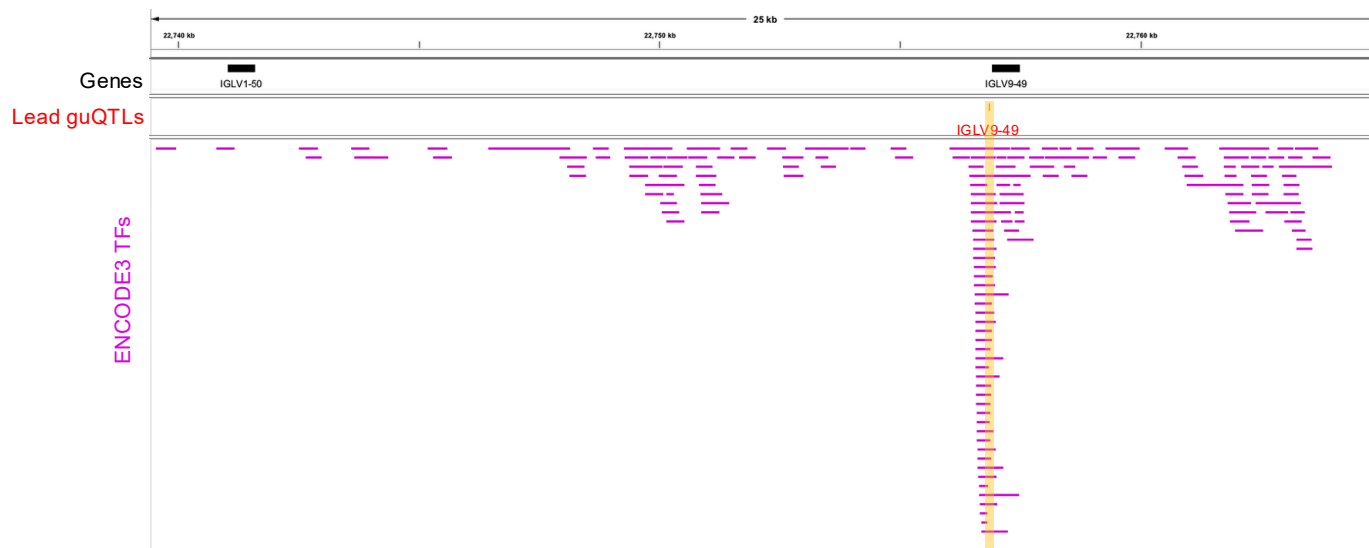


Figure S12. IGL guQTLs are enriched with ENCODE3 TFBS.

(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-sided Fisher's Exact Test; *P value < 0.05; **P value < 0.005; ***P value < 0.0005; ****P value < 0.00005 (see also **Supplementary Data 8**). (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*, and *IGLV2-8* with individuals separated by genotype at the lead guQTL variant (indicated in (B)) for these three genes. 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.

419



420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

Figure S13. 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.

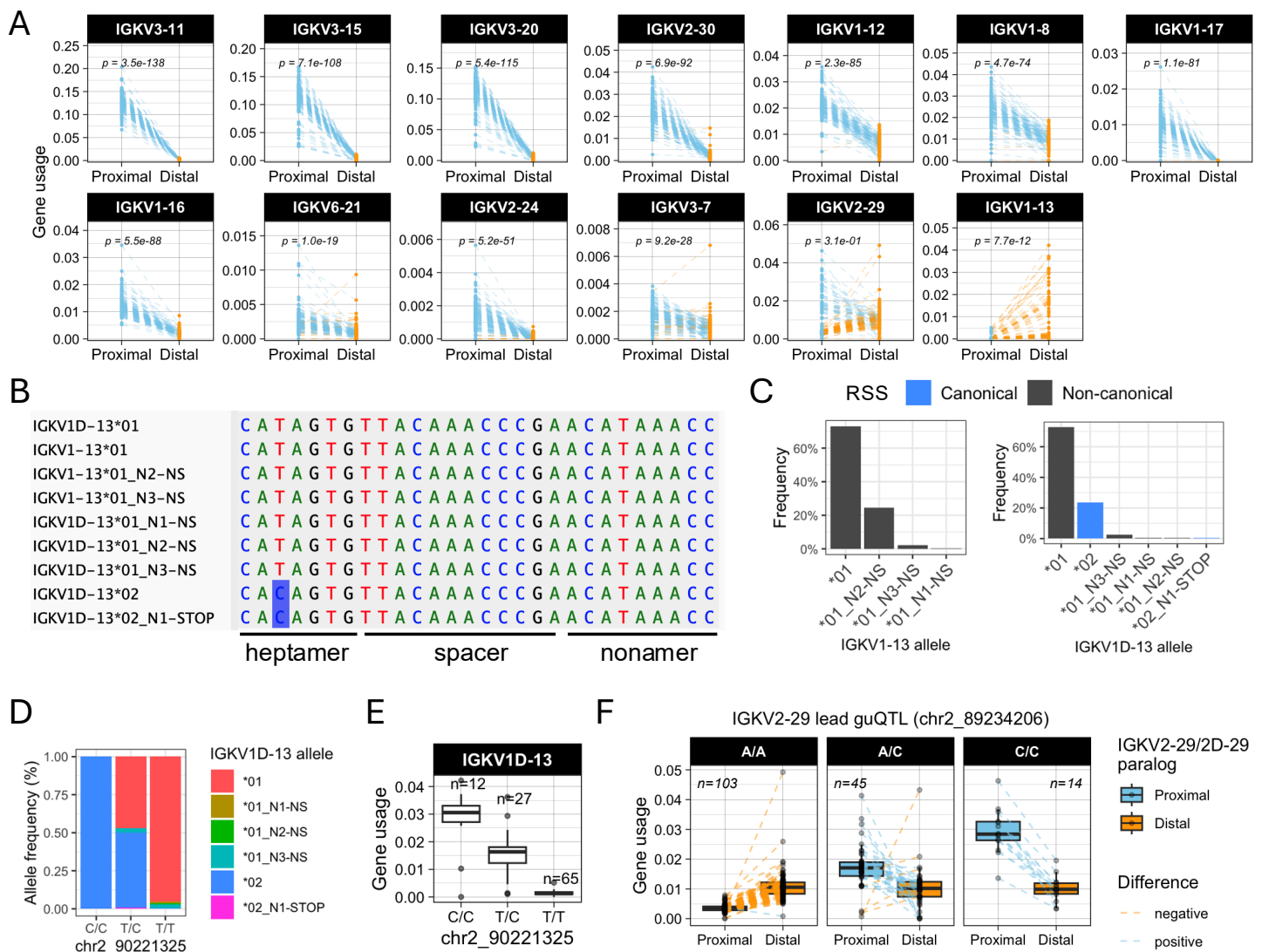


Figure S14. Analysis of IGKV gene paralog usages.

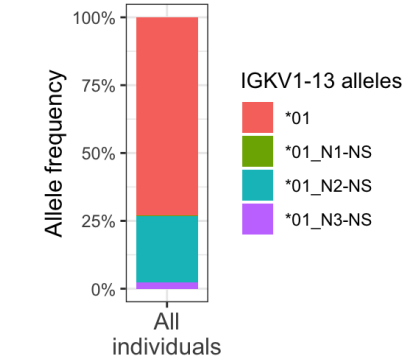
(A) Usage of indicated IGKV gene paralogs within individual (data points), with blue and orange dashed lines indicating higher usage of the proximal and distal paralog, respectively. Differences between proximal and distal gene usage were determined by paired t-tests. (B) Alignment of all RSSs for *IGKV1-13* 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 usages 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**). 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.

A

Germline alleles

	10	20	30	40	50	60	70	80	90	
IGKV1-13*01	AIQLTQSPSSLSASVGD	RVITITCRASQGISS	ALAYQQKPGKAPKLL	YDASSLES	GVPSRFS	SGSGSGT	DFTLT	ISSLP	EDFAT	YCYQQFNNYP
IGKV1-13*01_N1-NS	AIQLTQSPSSLSASVGD	RVITITCRASQGISS	ALAYQQKPGKAPKLL	YDASSLES	GVPSRFS	SGSGSGT	DF	KAVEH	VAD	SGHHRDTHRAVEHY
IGKV1-13*01_N2-NS	AIQLTQSPSSLSASVGD	RVITITCRASQGISS	ALAYQQKPGKAPKLL	YDASSLES	GVPSRFS	SGSGSGT	DFTLT	ISSLP	EDFAT	YCYQQFNNYP
IGKV1-13*01_N3-NS	AIQLTQSPSSLSASVGD	RVITITCRASQGISS	ALAYQQKPGKAPKLL	YDASSLES	GVPSRFS	SGSGSGT	DFTLT	ISSLP	EDFAT	YCYQQFNNYP
IGKV1D-13*01	AIQLTQSPSSLSASVGD	RVITITCRASQGISS	ALAYQQKPGKAPKLL	YDASSLES	GVPSRFS	SGSGSGT	DFTLT	ISSLP	EDFAT	YCYQQFNNYP
IGKV1D-13*01_N1-NS	AIQLTQSPSSLSASVGD	RVITITCRASQGISS	ALAYQQKPGKAPKLL	YDASSLES	GVPSRFS	SGSGSGT	DFTLT	ISSLP	EDFAT	YCYQQFNNYP
IGKV1D-13*01_N2-NS	AIQLTQSPSSLSASVGD	RVITITCRASQGISS	ALAYQQKPGKAPKLL	YDASSLES	GVPSRFS	SGSGSGT	DFTLT	ISSLP	EDFAT	YCYQQFNNYP
IGKV1D-13*01_N3-NS	AIQLTQSPSSLSASVGD	RVITITCRASQGISS	ALAYQQKPGKAPKLL	YDASSLES	GVPSRFS	SGSGSGT	DFTLT	ISSLP	EDFAT	YCYQQFNNYP
IGKV1D-13*02	AIQLTQSPSSLSASVGD	RVITITCRASQGISS	ALAYQQKPGKAPKLL	YDASSLES	GVPSRFS	SGSGSGT	DFTLT	ISSLP	EDFAT	YCYQQFNNYP
IGKV1D-13*02_N1-STOP	AIQLTQSPSSLSASVGD	RVITITCRASQGISS	ALAYQQKPGKAPKLL	YDASSLES	GVPSRFS	SGSGSGT	DFTLT	ISSLP	EDFAT	YCYQQFNNYP

B



C

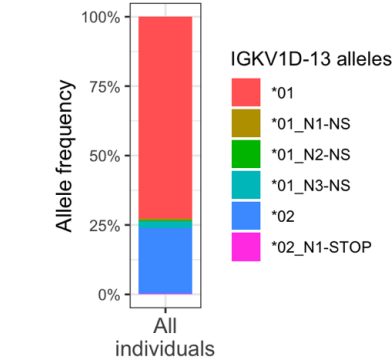


Figure S15. IGKV1-13 and IGKV1D-13 germline alleles.

(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.

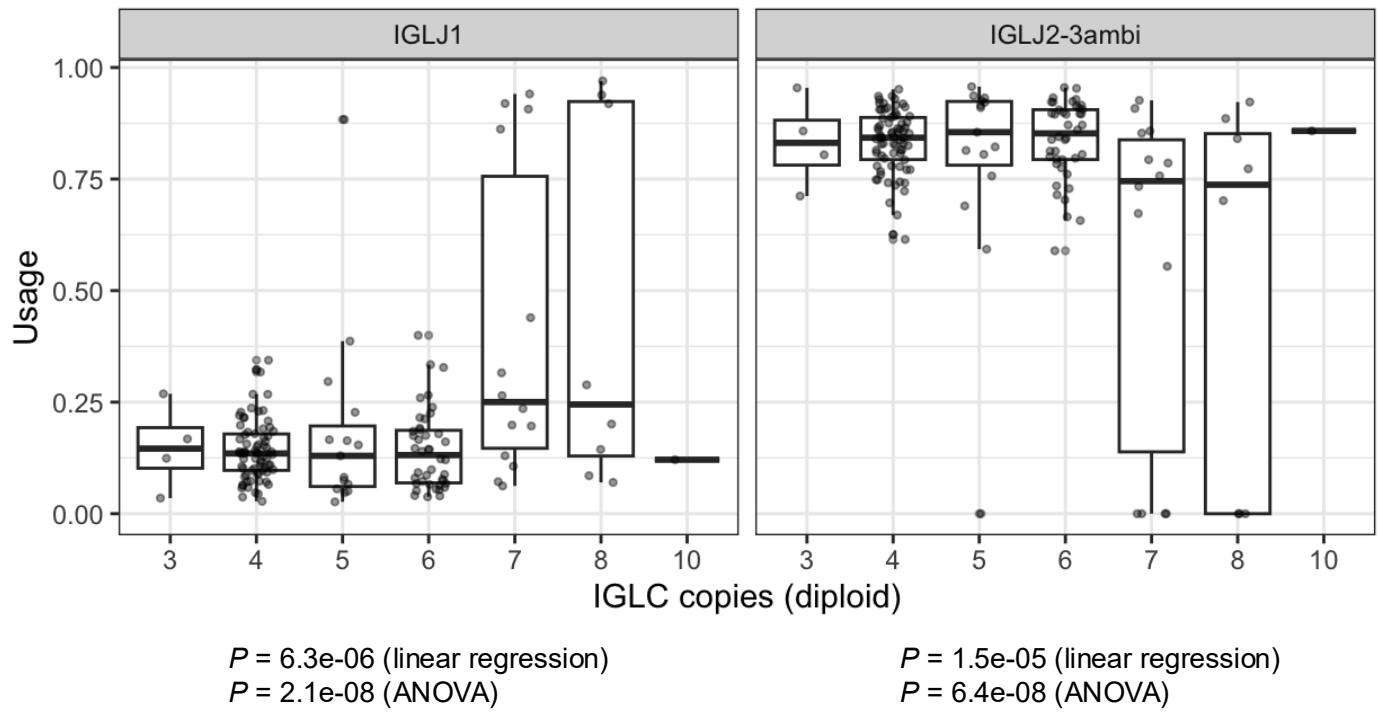


Figure S16. Usage of *IGLJ1* and *IGLJ2-3ambi*, with individuals grouped according to the number of diploid copies of the *IGLJ2-3* cassette. P values from indicated tests are shown.

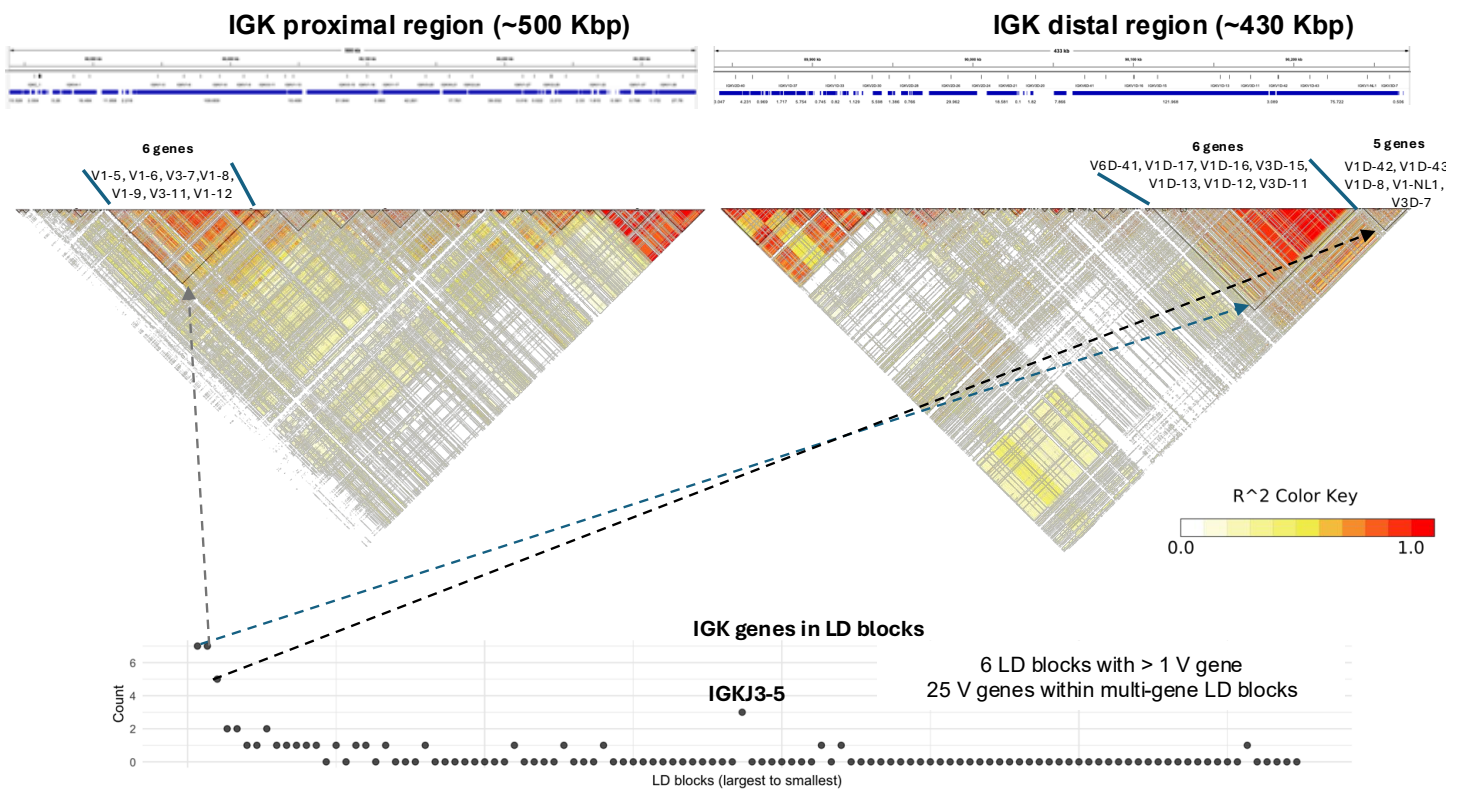


Figure S17. LD blocks in IGK.

(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.

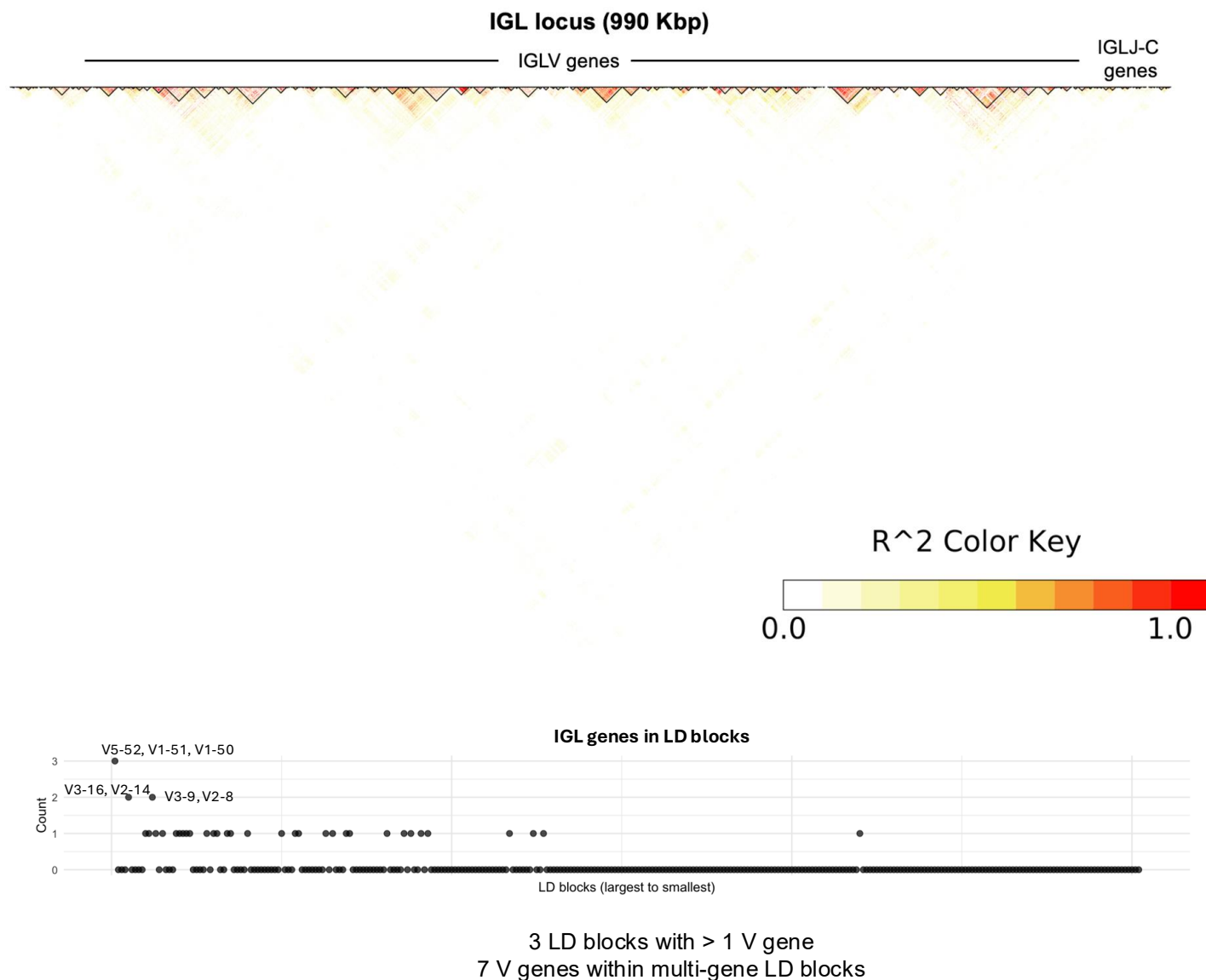
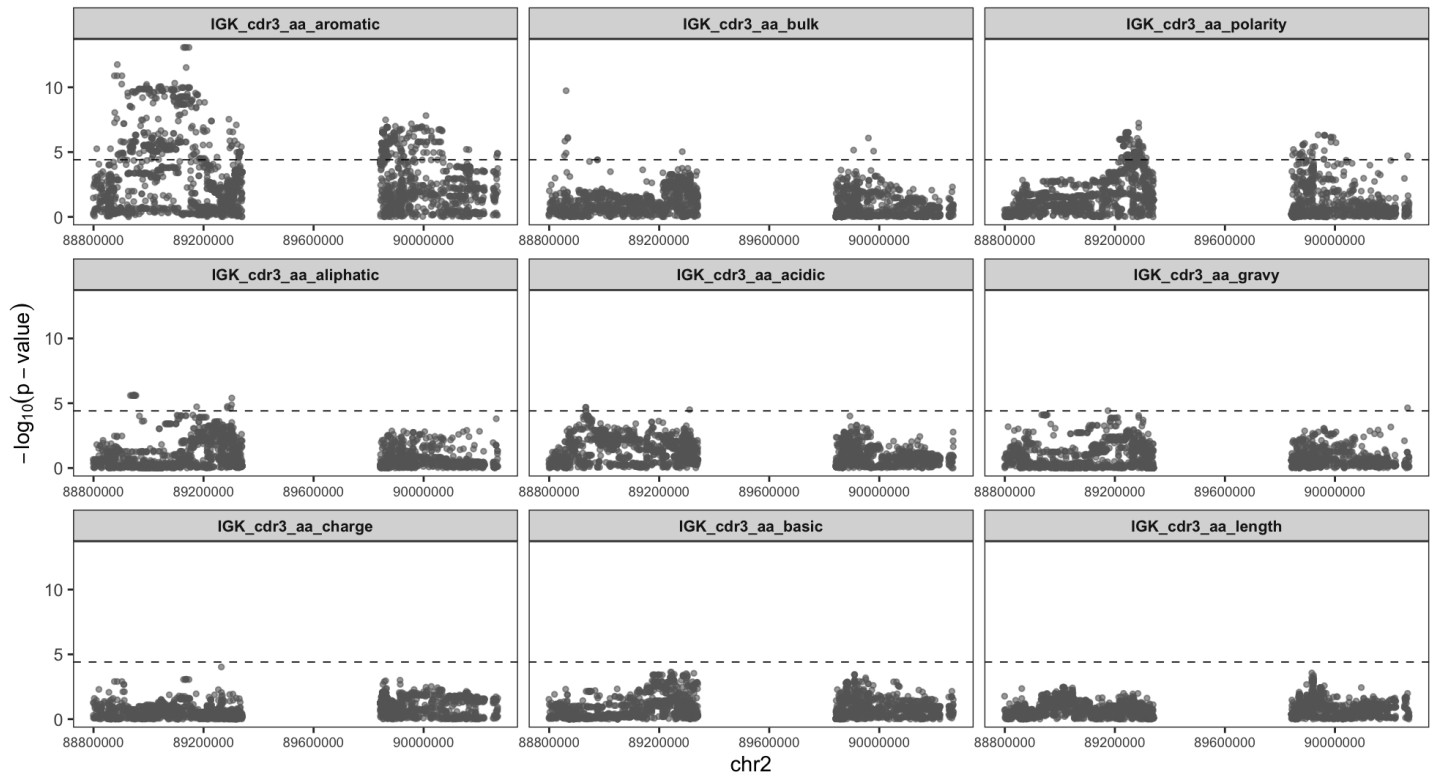


Figure S18. LD blocks in IGL.

(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.



Unmutated repertoire

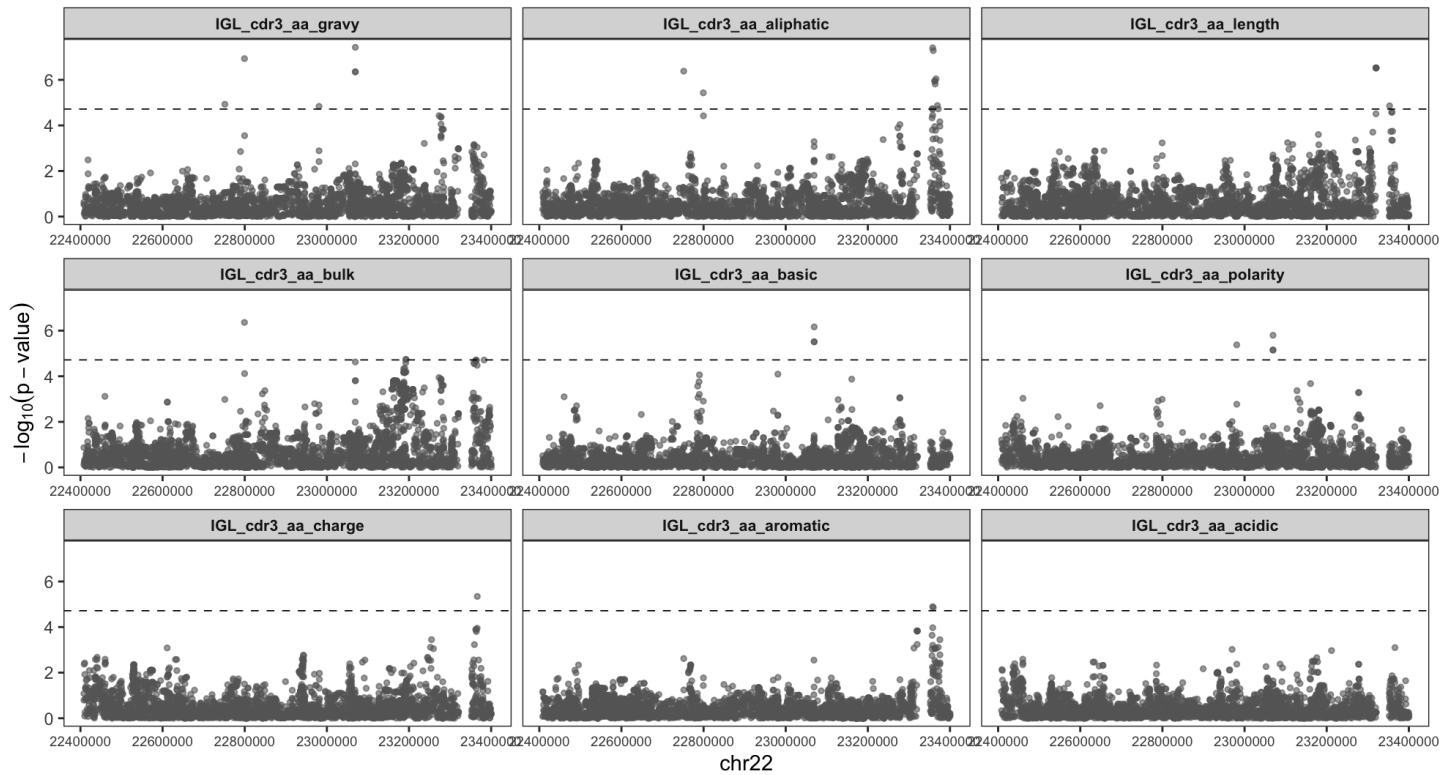
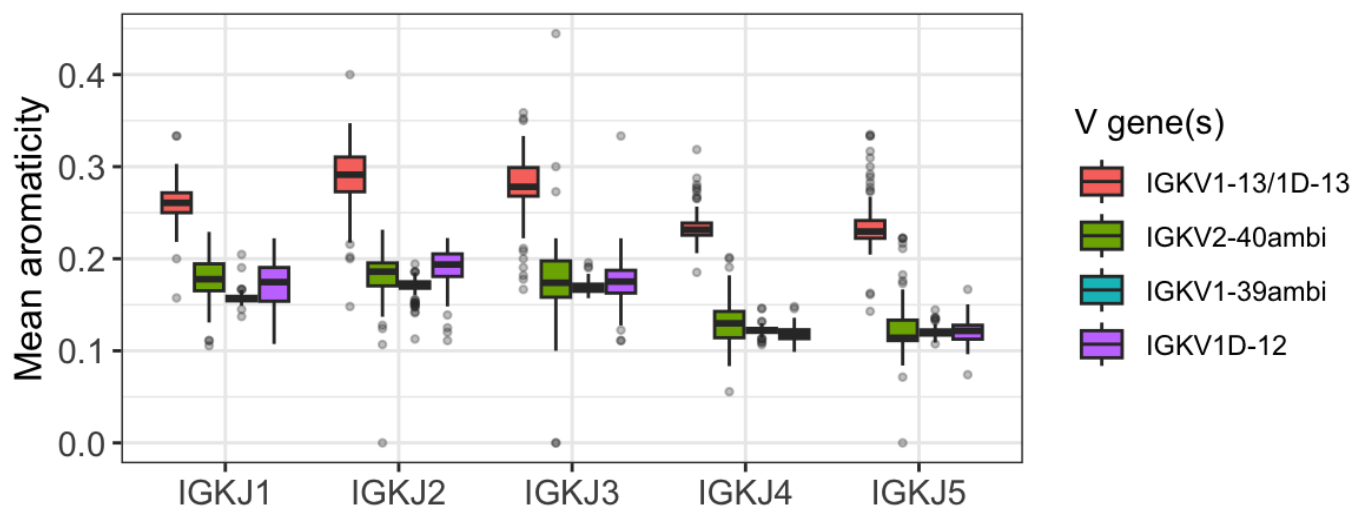


Figure S19. Associations between CDR3 physicochemical properties and germline variants in IGK and IGL. Manhattan plots show the $-\log_{10}(P\text{ value})$ for all SNVs in IGK (top) or IGL (bottom) tested for association with indicated CDR3 physicochemical properties in naïve (unmutated) Ab repertoires. Dashed lines indicate Bonferroni-corrected significance (IGK; $P < 3.7 \times 10^{-5}$, IGL; $P < 1.9 \times 10^{-5}$).

524
525
526
527
528
529
530
531



532
533
534
535
536
537
538
539
540
541
542
543
544

Figure S20. 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.

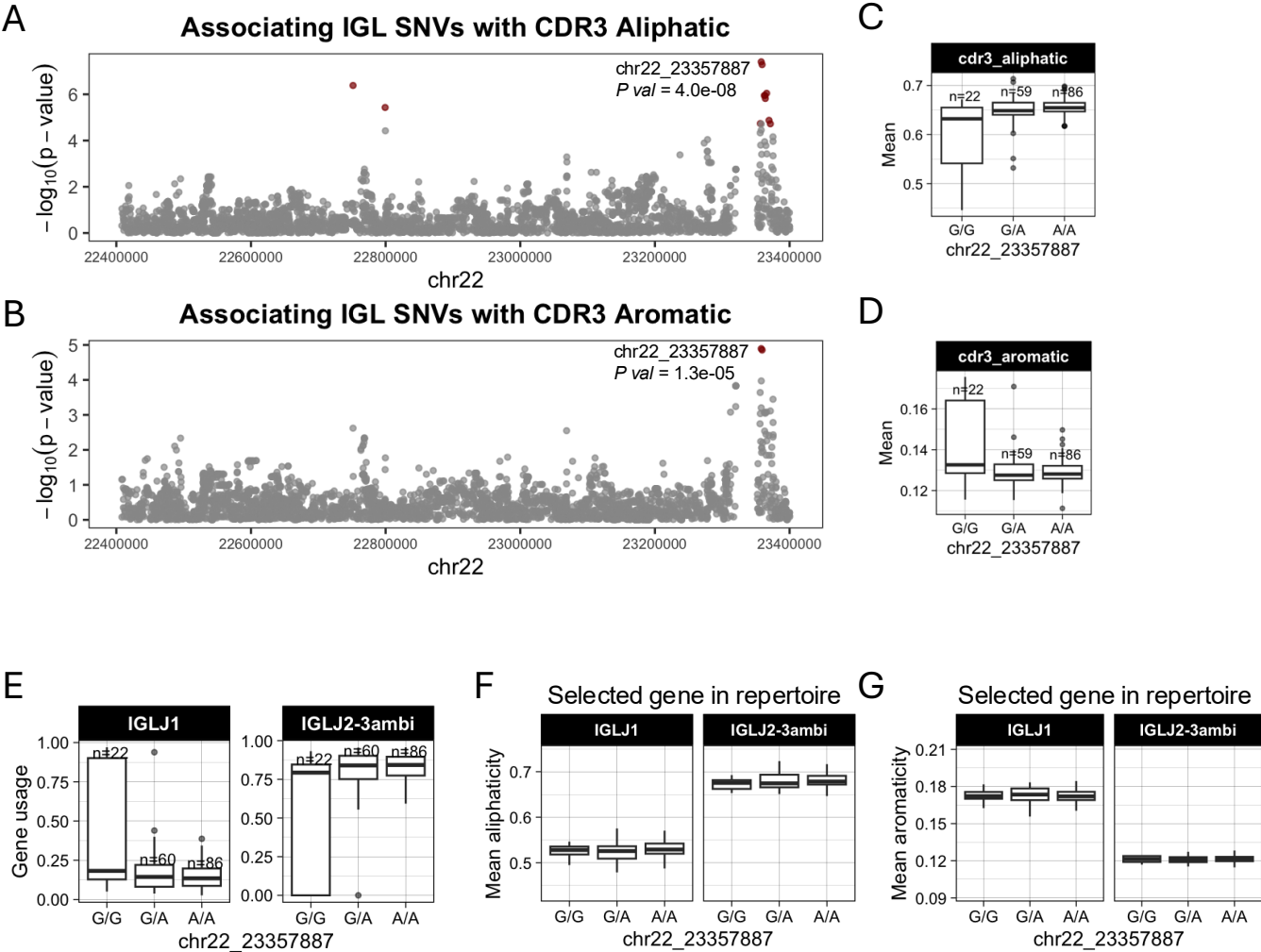


Figure S21. Genetic effects on IGLJ gene usages are associated with CDR3 aliphaticity and aromaticity. (A-B) Manhattan plots shows the $-\log_{10}(P \text{ value})$ for all SNVs in the IGL locus tested for association with CDR3 aliphaticity (A) or aromaticity (B), with QTLs colored dark red and the lead 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.

Supplementary Notes

Identification of SVs, SNVs, and gene alleles in IGK and IGL using long-read sequencing

To genotype variants in IGK and IGL, we used probe-based targeted capture long-read single molecule real-time (SMRT) sequencing¹ of the IGK proximal and distal regions², and the IGL locus³. IGK was sequenced to a coverage of 65.9X on average, with a mean of 1,858,094 diploid base pairs assembled per individual and mean assembly accuracy of 99.96% (**Supplementary Data 1, Supplementary Fig. 1A-B**). IGL was sequenced to a coverage of 41.3X on average, with a mean of 1,903,810 diploid base pairs assembled per individual and mean assembly accuracy of 99.93% (**Supplementary Data 1, Supplementary Fig. 1A-B**).

One of the primary objectives of this study was to develop a high-confidence set of genetic variants in IGK and IGL to enable downstream genetic association analysis. We previously described common SVs in IGK² and IGL³, including polymorphisms associated with gene conversion in IGK, a large inversion in IGK, a deletion involving *IGKV1-NL1*, a deletion involving *IGLV5-39*, and a deletion involving *IGLV3-16*, *IGLV2-18*, *IGLV3-19*, *IGLV3-21*, *IGLV3-22*, and *IGLV2-23* (termed IGLV2-18) (**Supplementary Fig. 2A-B**).

In addition to these SVs, we identified a haplotype wherein the entire IGKV distal region is deleted, which was previously reported^{4,5}, removing 23 functional IGKV genes. Nine individuals with this deletion haplotype exhibited ~2-3-fold higher sequencing coverage over the proximal relative to distal region (**Supplementary Fig. 4**), suggestive of hemizygosity. An additional individual lacked both HiFi reads and assemblies mapped to the IGK distal region, despite having 87X coverage and diploid assemblies for the proximal region (**Supplementary Fig. 4**), indicating homozygous absence of the IGKV distal region in this sample. SV allele frequencies for samples in this cohort are summarized in **Supplemental Table S2**.

SNVs within segmental duplications are difficult to characterize. We previously demonstrated that SNV callsets from haplotype-resolved IGK² and IGL³ assemblies derived from long-read SMRT sequencing outperform short-read derived variant calls in phase 3⁶ of the 1KGP. We identified variants for each individual and generated a callset of common SNVs, defined as those with a minor allele frequency (MAF) ≥ 0.05 , totaling 2,792 and 5,198 common SNVs in IGK and IGL, respectively (**Supplementary Fig. 2C**). Comparison of these SNVs with those in the dbSNP “common” (MAF ≥ 0.01) catalog revealed substantial differences; 59.3% and 38.8% of common SNVs in IGK and IGL, respectively, were absent from this dbSNP catalog (**Supplementary Fig. 2D-E**). These data indicate that dbSNP lacks accurate genotype information for about half of the common variants in IGK and about a third of the common variants in IGL.

While the majority of common SNVs were intergenic (IGK, 95%; IGL, 97%), SNVs were identified within features of V and J genes, including coding exons, V gene introns, and RSS heptamers, spacers, and nonamers (**Supplementary Fig. 5**). 34 of 47 IGKV genes (72%) and 29 of 39 IGLV genes (74%) harbored at least one common SNV within a gene feature.

Analysis of AIRR-seq data critically relies on assignment of AIRR-seq reads to specific IG gene alleles, which are typically identified from a germline allele database. For a given individual, accurate AIRR-seq analysis requires inclusion of all of the individual's alleles in the germline database to permit accurate assignment of reads to gene alleles; these assignments are used for analyzing a variety of Ab repertoire features, including gene usage and somatic hypermutation. We previously demonstrated that there is significant variation among the IGK² and IGL³ gene alleles in the human population, and many of these alleles are not documented in the commonly used germline alleles database called the ImMunoGeneTics Information System (IMGT; imgt.org). With haplotype-resolved IG assemblies in-hand, one can use a personalized germline allele database for AIRR-seq analysis of the individual.

To annotate germline IGK and IGL gene alleles in this cohort, we used haplotype-resolved assemblies and identified both documented and undocumented (novel) alleles, defined as alleles absent from IMGT (<https://www.imgt.org>). In total, we identified 160 and 145 high-confidence novel IGKV and IGLV alleles, respectively, defined as alleles with exact matches to ≥ 10 HiFi reads that mapped to the position of the allele sequence in the assembly (**Supplementary Fig. 2F-I; Supplementary Data 3, Supplementary Data 4**). Among all IGKV and IGLV alleles, 67.4% and 62.8% were not documented in IMGT (**Supplementary Fig. 6**). Among the novel alleles, 81 IGKV and 70 IGLV alleles were identified in more than 1 individual (**Supplementary Fig. 2H-I**). The majority of novel alleles resulted in non-synonymous substitutions (as compared to the closest-matching allele in IMGT), with 7 IGKV and 7 IGLV novel alleles encoding premature STOP codons (**Supplementary Fig. 2H-I**). We noted that 44 novel IGKV alleles were also identified in our previous survey of IGKV alleles using samples collected as part of the 1KGP². To access and explore curated genetic resources from this dataset further see Peres et al. (in prep).

IGK and IGL repertoire-wide gene usage profiles are more highly correlated in individuals carrying shared genotypes

Monozygotic twin studies have shown that gene usage frequencies in genetically identical individuals correlated to a greater degree than in unrelated individuals⁷⁻⁹. For IGH, we extended this observation at the population level and demonstrated that repertoire-wide gene usage profiles are more highly correlated in individuals carrying shared genotypes at guQTL SNVs¹⁰. To assess this in IGK and IGL, we used the same approach by estimating allele sharing distance (ASD)^{11,12} in our cohort across IGK or IGL, and comparing the gene usage correlations between groups of individuals with higher and lower ASDs. Repertoire-wide gene usage correlations between samples were calculated using the Pearson's Correlation coefficient. Using all guQTL variants for each gene, individuals with the most overlapping guQTL genotypes (low ASD) had a higher mean gene usage correlation than those in the group with the highest ASD scores for both IGK (0.988 vs. 0.957; KS test $p = 4.2e-86$) and IGL (0.931 vs. 0.710 ; KS test $p = 2.6e-04$) (**Supplementary Fig. 10**). These results indicated that genetic background makes a contribution to the overall gene usage composition of the

repertoire, and expand on observations from twin studies⁷⁻⁹ by demonstrating that heritable components of the light chain repertoire can be directly linked to germline variants in light chain loci.

Variants associated with IGL gene usage variation are enriched in regulatory regions

Large-scale studies utilizing expression, epigenomic, and variant datasets associated with diseases or traits have uncovered non-coding variants within regulatory elements linked to specific phenotypes¹³⁻¹⁶. In the context of V(D)J recombination, RSS are recognized by RAG1/RAG2 proteins to direct double-strand DNA breaks and initiate somatic recombination¹⁷. This process is regulated by cis-elements that interact with chromatin-binding proteins which collectively determine the probability of somatic recombination for a given gene¹⁸⁻²². Given that non-coding variants associated with IGH gene usage variation are enriched in regulatory regions, we hypothesized that non-coding variants might impact light-chain gene usage by affecting regulatory elements.

We tested for enrichment of candidate transcription factor binding sites (TFBS, ENCODE3 Transcription Factor ChIP-seq dataset) in IGL (302 regulatory elements) and IGK (129 regulatory elements) overlapping lead guQTL variants in IGL (97 lead SNVs) and IGK (208 lead SNVs) versus the remainder of common SNVs for each locus, which included 4,562 and 2,499 non-lead variants in IGL and IGK, respectively. We found that lead IGL guQTLs were enriched with TFBS for multiple factors, including SRF, CBX1, TRIM28, TAL1, ETS1, and EGR1 (**Supplementary Fig. 12**). The guQTL variants overlapping these TFBS were associated with usage of *IGLV3-10*, *IGLV2-8*, *IGLV3-9*, *IGLV3-27*, *IGLV2-11*, and *IGLV9-49* (**Supplementary Data 8**). In contrast with IGL, there was not a significant enrichment of lead IGK guQTL variants within TFBS.

The lead variant associated with the usage of *IGLV3-10*, *IGLV2-8*, and *IGLV3-9* (**Supplementary Fig. 12**) overlaps an EGR1 site in addition to >10 distinct annotated TFBS including SMC3, YY1, ETS1, ATF1, CTCF, and RAD21. Similarly, the lead variant associated with *IGLV9-49* usage overlaps CBX1 and TRIM28 sites in addition to sites for >10 additional TF's, among which were SMC3, CREB1, MYC, JUNB, STAT5A, CTCF, and RAD21 (**Supplementary Fig. 13**). We noted that the *IGLV9-49* lead variant is 50 bp upstream of the *IGLV9-49* 5' UTR.

These data suggest that a subset of IGL guQTLs are within cis-elements that may regulate V(D)J recombination and therefore gene usage in the peripheral antibody repertoire.

Genetic variants disrupt biases in differential usage of IGKV proximal and distal paralogs

IGKV proximal and distal gene paralogs vary substantially with respect to distance from the IGKJ region along the linear genome, with at least 1.0 Mbp separating distal V genes from J segments²³. To date, there has not been an empirical evaluation of IGKV proximal versus distal gene usage. Historically, usage of IGKV genes has been difficult to assess due to high sequence similarity between paralogs^{24,25}. With paired AIRR-seq and germline alleles available for the first time at population-scale, we determined the individual usage of 13 IGKV paralog pairs and observed higher usage of the proximal gene in 11 cases, with the exception of

IGKV2-29/2-29, for which usage was not significantly different, and *IGKV1-13/1D-13*, for which distal paralog usage was higher (paired t-tests, P value $< 1.0\text{e-}10$) (**Supplementary Fig. 14A**).

IGKV1-13 was used at a lower frequency than *IGKV1D-13* in all but one individual. Notably, all *IGKV1-13* alleles have a non-canonical heptamer (CATAGTG) (**Supplementary Fig. 14B**), and *IGKV1-13*01*, the most frequent allele (frequency 72.3%, **Supplementary Fig. 14C**) encodes a premature STOP codon (**Supplementary Fig. 15**). Therefore, *IGKV1-13* resembles a pseudogene, with both regulatory and coding loss-of-function features across haplotypes. *IGKV1D-13* sequences, in contrast, included an allele (*IGKV1D-13*02*) with a canonical heptamer (CACAGTG) identified at a frequency of 24.5% (**Supplementary Fig. 14C**). The lead *IGKV1D-13* guQTL tagged this heptamer variation, as all individuals in the higher-usage C/C genotype group were also *IGKV1D-13*02/IGKV1D-13*02*, whereas 64 of 65 individuals in the T/T genotype group did not carry the *02 allele (**Supplementary Fig. 14D-E**), implicating heptamer variation as a mechanism underlying usage variation.

IGKV2-29 was used at a lower frequency than *IGKV2D-29* in 106 out of 162 individuals. This inter-individual variation in *IGKV2-29* versus *IGKV2D-29* usage was explained by the lead *IGKV2-29* guQTL, with 102 out of 103 individuals homozygous for the reference allele having higher *IGKV2D-29* usage and all 14 individuals homozygous for the alternate allele having higher *IGKV2-29* usage (**Supplementary Fig. 14F**). Mechanistically, the lead *IGKV2-29* guQTL results in a premature STOP codon in the *01, *01_N1, and *01_N2 alleles (**Fig. 2B**), which are the only *IGKV2-29* alleles carried by the 103 individuals homozygous for the guQTL reference allele.

Supplementary references

1. Rodriguez, O. L. *et al.* A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. *Front. Immunol.* **11**, 2136 (2020).
2. Engelbrecht, E. *et al.* Resolving haplotype variation and complex genetic architecture in the human immunoglobulin kappa chain locus in individuals of diverse ancestry. *Genes Immun.* (2024) doi:10.1038/s41435-024-00279-2.
3. Gibson, W. S. *et al.* Characterization of the immunoglobulin lambda chain locus from diverse populations reveals extensive genetic variation. *Genes Immun.* **24**, 21–31 (2023).
4. Schaible, G., Rappold, G. A., Pargent, W. & Zachau, H. G. The immunoglobulin kappa locus: polymorphism and haplotypes of Caucasoid and non-Caucasoid individuals. *Hum. Genet.* **91**, 261–267 (1993).
5. Pargent, W., Schäble, K. F. & Zachau, H. G. Polymorphisms and haplotypes in the human immunoglobulin kappa locus. *Eur. J. Immunol.* **21**, 1829–1835 (1991).
6. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
7. Glanville, J. *et al.* Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20066–20071 (2011).
8. Rubelt, F. *et al.* Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat. Commun.* **7**, 11112 (2016).
9. Wang, C. *et al.* B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 500–505 (2015).
10. Rodriguez, O. L. *et al.* Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat. Commun.* **14**, 4419 (2023).
11. Gao, X. & Martin, E. R. Using allele sharing distance for detecting human population stratification. *Hum. Hered.* **68**, 182–191 (2009).
12. Gao, X. & Starmer, J. Human population structure detection via multilocus genotype clustering. *BMC Genet.* **8**, 34 (2007).
13. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
14. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
15. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature*.
16. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

- 738 17. Fugmann, S. D., Lee, A. I., Shockett, P. E., Villey, I. J. & Schatz, D. G. The RAG proteins and V(D)J
739 recombination: complexes, ends, and transposition. *Annu. Rev. Immunol.* **18**, 495–527 (2000).
- 740 18. Hill, L. *et al.* Igh and Igk loci use different folding principles for V gene recombination due to distinct
741 chromosomal architectures of pro-B and pre-B cells. *Nat. Commun.* **14**, 2316 (2023).
- 742 19. Bhat, K. H. *et al.* An Igh distal enhancer modulates antigen receptor diversity by determining locus
743 conformation. *Nat. Commun.* **14**, 1225 (2023).
- 744 20. Barajas-Mora, E. M. *et al.* Enhancer-instructed epigenetic landscape and chromatin compartmentalization
745 dictate a primary antibody repertoire protective against specific bacterial pathogens. *Nat. Immunol.* **24**,
746 320–336 (2023).
- 747 21. Bolland, D. J. *et al.* Two mutually exclusive local chromatin states drive efficient V(D)J recombination. *Cell*
748 *Rep.* **15**, 2475–2487 (2016).
- 749 22. Schatz, D. G. *et al.* The mechanism, regulation and evolution of V(D)J recombination. in *Molecular Biology*
750 *of B Cells* (eds. Honjo, T., Reth, M., Radbruch, A., Alt, F. & Martin, A.) 13–57 (Elsevier, 2024).
- 751 23. Kawasaki, K. *et al.* Evolutionary dynamics of the human immunoglobulin kappa locus and the germline
752 repertoire of the Vkappa genes. *Eur. J. Immunol.* **31**, 1017–1028 (2001).
- 753 24. Collins, A. M. *et al.* AIRR-C IG Reference Sets: curated sets of immunoglobulin heavy and light chain
754 germline genes. *Front. Immunol.* **14**, 1330153 (2023).
- 755 25. Mikocziova, I. *et al.* Germline polymorphisms and alternative splicing of human immunoglobulin light chain
756 genes. *iScience* **24**, 103192 (2021).