Title: The human immunoglobulin heavy chain constant gene locus is enriched for large complex structural variants and coding polymorphisms that vary in frequency among human populations

Authors: Uddalok Jana[1], Oscar L. Rodriguez[1], William Lees[1,5], Eric Engelbrecht[1], Zach Vanwinkle[1], Ayelet Peres[5], William S. Gibson[1,6], Kaitlyn Shields[1], Steven Schultze[1], Abdullah Dorgham[1], Matthew Emery[2], Gintaras Deikus[2], Robert Sebra[2], Evan E. Eichler[3,4], Gur Yaari[5], 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 Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [3] Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA; [4]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA; [5] Bioengineering Program, Faculty of Engineering, Bar-Ilan University, Ramat Gan, 5290002, Israel; [6] Vaccine Research Center, National Institute of Allergy and Infectious Disease, National Institute of Health, Bethesda, MD

*To whom correspondence should be addressed: corey.watson@louisville.edu, ml.smith@louisville.edu

Abstract

The immunoglobulin heavy chain constant (IGHC) domain of antibodies (Ab) is responsible for effector functions critical to Ab mediated immunity. In humans, this domain is encoded by genes within the IGHC locus, where descriptions of genomic diversity remain incomplete. To address this, we utilized long-read genomic datasets to build a high-quality IGHC haplotype/variant catalog from 105 individuals of diverse ancestry, and developed a high-throughput approach for targeted long-read IGHC locus sequencing and assembly. From locally phased assemblies, we discovered previously uncharacterized single nucleotide variants (SNV) and complex structural variants (SVs, n=7), as well as novel genes and alleles. Of the 262 identified IGHC coding alleles, 235 (89.6%) were undocumented. SNV, SV, and gene allele/genotype frequencies revealed significant population differentiation, including; (i) hundreds of SNVs in African and East Asian populations exceeding fixation index (F_{ST}) of 0.3, (ii) and an IGHG4 haplotype carrying specific coding variants uniquely enriched in East and South Asian populations. Our results illuminate missing signatures of haplotype diversity in the IGHC locus, including evidence of natural selection, and establish a new foundation for investigating IGHC germline variation and its role in Ab function and disease.

Introduction

Antibodies (Abs) are critical components of the immune system (1), playing pivotal roles in adaptive immunity, with involvement in many disease states and other clinical outcomes (2). These proteins are produced by B lymphocytes and are either expressed at the cell surface as B cell receptors (BCRs) or are secreted as Abs. Abs and BCRs are composed of two pairs of identical 'heavy' and 'light' (kappa or lambda) chains, encoded by immunoglobulin (IG) genes located in three primary regions of the human genome: the IG heavy chain (IGH) locus, (chr14q32.33), and the light chain lambda (IGL; chr22q11.2) and kappa (IGK; chr2p11.2) loci. The expression of a given Ab depends on the somatic rearrangement of variable (V), diversity (D; heavy chain only), and joining (J) gene segments at IGH and one of the two light chain loci. The recombined V(D)J DNA template is then transcribed and spliced to a specific constant gene to comprise a full transcript. Genes within the IGH constant (IGHC) gene region encode and determine the functional Ab isotype (IgM, IgG, IgA, IgE, and IgD). These isotypes are responsible for mediating downstream immune processes (3).

The IGHC genes reside within a ~350 Kbp region adjacent to the IGHJ gene cluster (4). Within the T2TCHM13v2.0 genome reference, for example, nine IGHC genes are present (IGHM, IGHD, IGHG3, IGHG1, IGHA1, IGHG2, IGHG4, IGHE, and IGHA2; listed in 5'-3' order in Figure 1) each encoding a specific heavy chain isotype or subisotype. All IGHG genes share a common genomic structure consisting of CH1, CH2, CH3 domain exons and a hinge region between the CH1 and CH2 domain (5). The hinge exons of IGHG3 can vary in copy number, ranging from 2-4 (6). IGHD genes have a structure similar to IGHG3, including two hinge region exons (7). In contrast, IGHA, IGHE, and IGHM genes lack hinge region exons, and both IGHE and IGHM harbor an added CH4 domain exon downstream of CH3. IGHC genes encode a multi-domain protein structure that contains binding sites for complement and Fc receptors (8)(9), which modulate Ab effector functions and determine Ab transport and half-life (10).

The IGHC locus has been shaped by segmental duplication, and consists of multiple homologous sequence blocks that harbor functional and pseudogenized paralogs of IGHG, IGHA, and IGHE genes (11, 12). The region is also diverse at the population-level, both with respect to structural variation that can result in changes in IGHC gene copy number (4, 13–16), as well as SNVs within IGHC coding sequences that result in extensive allelic variation (14, 17, 18). Specifically, duplications and deletions involving IGHA1, IGHA2, IGHE, IGHG2 and IGHG4 have all been reported (4), but the majority of these are yet to be characterized with nucleotide-level resolution. With respect to allelic variation, there are currently 100 alleles cataloged in the IMmunoGeneTics Information System database (date: November 24, 2022; IMGT; imgt.org;) (19). However, recent reports indicate that there are likely to be many more alleles. A recent study of IGHG genes in a Brazilian cohort of 357 individuals from multiple population groups identified 28 novel alleles (17). This was consistent with our recent study, in which we identified a total of 24 IGHG and 4 IGHM novel alleles in only 10 healthy individuals (14). Furthermore, significant differences in IGHC allele frequency variation have been demonstrated between human populations (18, 20).

Functional consequences of IGHC polymorphism are known, but our understanding of the full extent of these impacts remains limited (21). These functional consequences can include impacts on inter-individual variation in isotype distributions within the expressed Ab repertoire, as well as serum Ab levels of various isotypes (15, 22–25). Additionally, coding allelic variation has been shown to alter protein function, such as induced effects on neonatal Fc receptor (FcRn) affinity (10), and influence on Ab dependent cellular cytotoxicity (ADCC) (26). These functional differences are expected to link IGHC polymorphism to complex disease phenotypes. Indeed, IGHG variants have been associated with disease outcomes in multiple

contexts, including but not limited to multiple sclerosis (MS) (27), human cytomegalovirus (HCMV) infection (28), Myasthenia Gravis (29), Myelin oligodendrocyte glycoprotein (MOG) antibody-associated disease (MOGAD) (30), and COVID severity (31).

Despite previous studies highlighting extensive allelic diversity and structural variation within IGHC, our knowledge of locus-wide haplotype diversity in this region at the genome-level remains markedly insufficient. To date, there are only two fully curated and complete locus representations of IGHC. These include the T2T-CHM13v2.0 and GRCh38 reference assemblies, both of which are derived from individuals of western European ancestry (32, 33). Historically, technical barriers have impeded the use of standard next-generation short-read sequencing to resolve complex loci (34, 35), specifically the IG loci (36–38). The emergence of advanced long-read sequencing techniques, such as single molecule real-time (SMRT) sequencing (Pacific Biosciences), with enhanced accuracies have shown significant improvements in comprehensive and precise detection of SNVs, SVs, repeat expansions, and complex genomic regions (35, 39, 40), thereby providing a deeper and more accurate understanding of genomic architecture and variation. These technologies have the potential to enable higher-resolution accurate characterization of the IGHC locus, including of segmental duplications and deletions. Considering the anticipated extent of genetic diversity across human populations, targeted approaches will be necessary to allow for high-throughput IGHC sequencing at population-scale.

In this study, we employ targeted long-read SMRT sequencing (Pacific Biosciences) of genomic DNA and large-insert fosmid clones from donors of the 1000 Genomes Project (1KGP) to generate high-quality haplotype resolved assemblies from five super populations (AFR, AMR, EUR, EAS, SAS). We first leverage a combination of fosmid clones and targeted long-read assemblies from four individuals of diverse genetic ancestries to generate manually curated and benchmarked high-quality haplotype-specific assemblies spanning the IGHC region. To facilitate additional benchmarking of our targeted long-read sequencing approach, we construct haplotype-phased assemblies for four trio probands, allowing for careful accuracy assessment utilizing parental datasets. We use these curated assemblies as the foundation for developing and evaluating a high-throughput computational pipeline for IGHC long-read assembly and variant calling. With this approach, we expand our survey of IGHC haplotype diversity to include an additional 97 individuals, leading to the largest curated set to date (n=105 donors). From this cohort, we identify extensive diversity at high resolution, including descriptions of novel SVs, SNVs, and short insertion-deletions (indels), as well as previously uncharacterized IGHC genes and alleles. This advancement enhances our understanding of IGHC gene diversity at the genome and population levels, highlighting the vast and underappreciated polymorphism in this region. Our analyses illuminate the need to expand genetic surveys in these critical loci if we are to fully elucidate the importance of germline variation in Ab effector functions in disease.

Materials and methods

Sample information

Samples used in this study were previously collected as a part of the 1000 Genomes Project (1KGP) (41). Genomic DNA was procured from the Coriell Institute for Medical Research (https://www.coriell.org/; Camden, NJ). Fosmid libraries were originally generated from lymphoblastoid cell lines (LCL) (42). Sample, population, sub-population, sex and relatedness information are reported in Table S1. While EBV (Epstein-Barr virus) transformed LCLs are valuable for studying human genetic diversity, the IGHC loci were carefully curated to account for class switch recombination, ensuring DNA integrity and haplotype preservation.

Constructing a reference assembly haplotype for the human IGHC locus

To build a reference assembly for the mapping and characterization of genetic variants in samples sequenced in our study, we utilized the IGHC locus from the T2T CHM13 v2.0 build (chr14:99730306-100132055) as our IGHC reference for all analyses conducted below. We chose to use this assembly rather than GRCh38/hg38 because it harbored a ~19.5 Kbp insertion, including a duplication of the gene IGHG4.

Fosmid isolation, sequencing and assembly

Fosmid clones were selected based on clone-end sequence mapping to the IGHC locus (GRCh38; chr14:105,510-105,860,000). For each individual, five non-overlapping pools of fosmids were selected (average n=16 per pool). Fosmids were grown in 96-well plates and DNA was extracted following previously published methods (42). On a per fosmid clone basis, DNA was sheared using g-tubes (Covaris, Woburn, MA, USA) to ~15 Kbp, followed by size selection of >10 Kbp fragments using the BluePippin system (Sage Science, Beverly, MA, USA). Following shearing and size-selection, DNA from non-overlapping fosmids was pooled in equimolar ratios, with numbers of samples per pool dependent on the SMRT sequencing platform used. Pooled DNA was then directly prepared for sequencing using SMRTbell Template Preparation Kit 1.0 (for RSII-sequenced samples) or 2.0 (for Sequel IIe-sequenced samples) (Pacific Biosciences, Menlo Park, CA, USA) following the manufacturer's standard protocol. SMRTbells were sequenced on either the RSII sequencing system (Pacific Biosciences) using P6/C4 chemistry and 6 hour movies, or the Sequel IIe system (Pacific Biosciences) using 2.0 chemistry and 30h movies. Fosmid-level metadata are provided in Table S2.

For each fosmid pool, sequencing data from individual fosmids were used to generate fosmid-specific assemblies by first removing the fosmid backbone sequence from the reads and mapping them to the GRCh38 reference. Then, for each fosmid, only reads that mapped within the fosmid end sequence coordinates were de novo assembled using Canu (42, 43). Clones with ends that did not map well to the reference genome (~40 Kbp chunks) were marked discordant; those were barcoded and sequenced. These fosmids were assembled *de novo* without mapping.

Targeted capture sequencing and assembly

For targeted long-read sequencing of the IGHC locus, we designed a tailored HyperCap DNA probe panel (Roche). This panel included target sequences from the GRCh38 IGHC locus (chr14:105,510-105,860,000). Genomic DNA (gDNA) samples were prepared and sequenced according to previously published methods (36, 38, 44, 45). Briefly, gDNA was sheared to approximately 8 Kbp using g-tubes (Covaris, Woburn, MA, USA) and size-selected with the Blue Pippin system (Sage Science, Beverly, MA, USA), collecting fragments in size ranging from 5–9 Kbp. Fragmented gDNA was ligated to universal adapters and amplified. Amplified individual samples were pooled in groups of 6 prior to target-enrichment using the custom HyperCap DNA

probes. After probe capture, targeted fragments were subjected to a second round of amplification and processed for SMRT sequencing using the SMRTbell Template Preparation Kit 2.0 as per the manufacturer's protocol. The resulting SMRTbell libraries were multiplexed across pools as 2-plex (12 samples) and sequenced using one SMRT cell 8M on the Sequel IIe system with 2.0 chemistry and 30-hour movies.

Following data acquisition, HiFi (high-fidelity; > Q30 or 99.9%) reads were generated and aligned to the IGHC reference for coverage analysis to first assess sequencing depth over the IGHC region; non-specific reads mapping to non-IGHC regions were discarded and the mean per-base read coverage for each sample is provided in Table S1. HiFi reads were used to generate phased de-novo assemblies using Hifiasm with default parameters (46). Assemblies were aligned to the IGHC reference using minimap2 (version 2.26) (47) with the '-x asm20' allowing maximum divergence in alignments and using custom gap penalty adjustments to accommodate better alignments over structural variants. HiFi reads were also used to generate orthogonal assemblies using the IGenotyper (38) pipeline with default parameters to generate reference-guided locally phased assemblies. The Hifiasm-based assemblies, which offer greater contiguity, were aligned to the reference and utilized for SV detection. Subsequent analysis involved SV annotation and allele curation across SV regions using haplotype-specific Hifiasm contigs. Assemblies generated by IGenotyper were guided by the reference genome and phased using contiguous read support in conjunction with the presence of heterozygous single SNPs. Phasing blocks were defined as the reference span between the two farthest SNPs connected by contiguous read support, and these blocks were utilized to generate the most accurate haplotype-specific contigs. These contigs were then aligned back to the reference genome to obtain contig alignment over the IGHC locus, excluding the IGHG4 SV region. The alignments were then examined for areas with missing contigs using the bedtools genomecov command, confirming the presence of two heterozygous contigs (hap=1, hap=2) or one homozygous contig (hap=0) in every alignment block. To fill in these alignments, curated missing region boundaries were supplemented using de novo assemblies generated by Canu (43). These assemblies were created based on available HiFi reads aligned specifically to the missing region boundary, resulting in the production of missing haplotypes or alternate haplotypes to complete the final assemblies. These contigs were also aligned against the IGHC reference using minimap2 (version 2.26) with the '-x asm20' for allowing maximum divergence in the alignment and using custom gap penalty adjustments as mentioned before.

Construction of hybrid assemblies

Fosmid contigs were aligned to the IGHC reference for each individual, and phased into haplotype groups using overlapping SNP profiles detected through manual inspection of contigs in the Integrative Genomics viewer (IGV) (48), (49). Haplotype-specific fosmid contigs were merged to form larger contigs using a custom python script, requiring a minimum 2 Kbp overlap, with a maximum 5 bp mismatches. Gaps in the resultant merged assemblies were filled using assemblies generated from long-read targeted sequencing from the same donors. To validate bases in hybrid assemblies, HiFi reads from IGHC locus-wide targeted capture sequencing from the same donor were realigned to each diploid assembly. Base discrepancies between reads and assembly bases were assessed using the WhatsHap (50) find_snv_candidates command to detect SNPs potentially representing erroneous bases in these diploid reference assemblies.

Construction of targeted capture phased assemblies in trio probands

We selected 4 human trios for which we generated targeted long-read sequencing, and for which the proband dataset had a mean 30X read depth over IGHC locus bases. The parental HiFi reads were processed using Yak to generate a count of k-mers present in each parent, utilizing the parameters "-k31 -b37". Subsequently, the proband was assembled using Hifiasm in trio binning mode (46), employing the parental k-mer count.

Additionally, the proband HiFi reads were assembled using IGenotyper (38). The proband contigs obtained from the Hifiasm and IGenotyper pipelines were inspected in IGV alongside their respective parental Hifiasm contigs to ensure that phased contigs in the proband were supported by those generated in the parents. Any phase switch errors in Hifiasm contigs were corrected using IGenotyper contigs to produce final gapless contigs for each haplotype of the proband. Proband contigs derived from trio-based analyses underwent orthogonal validation, employing a method similar to that used for hybrid assemblies. This involved realigning capture reads to their personalized diploid references and assessing the presence of erroneous bases within the diploid reference using SNP calls obtained through the WhatsHap find_snv_candidates command.

Detection and genotyping of IGHC single nucleotide variants in phased hybrid and proband assemblies

Samples with hybrid assemblies and trio-based proband assemblies were processed using ‘ bcftools mpileup ’ (51), (52) command with ‘ -Ou -f -B ’ to call variants and ‘ -r ’ was used to specify the region of targeted capture alignment, for the HiFi reads based on capture probes, then ‘ bcftools call ’ command was used with ‘ -m -Ov -V indels ’ to get all bases reported alongside variants, skipping the indels. The ‘bcftools view’ command was used with “ -i ‘QUAL >= 0’ ” to filter positions with zero quality scores, then the ‘bedtools intersect’ (53) command was used with ‘-b masking_regions.bed’, which consists of common IGHC SV coordinates to skip any entries in the vcf file over SV regions. These vcf files were used as ground truth to assess targeted capture-only diploid assemblies processed through our custom IGHC genotyping pipeline.

To analyze population statistics, variant data from targeted capture-based assemblies with a MAF > 5% were utilized. The R package SNPrelate (54) was utilized to perform principal component analysis (PCA), enabling the assessment of population relatedness using identity-by-descent (IBD) measures. Pairwise F_{ST} analysis was conducted between a specific population group and the rest of the individuals using the same variant dataset to evaluate genetic differentiation. Weir and Cockerham’s F_{ST} (55) values were calculated for all IGHC SNPs having a minor allele frequency >5% using VCFtools v0.1.16 (56).

Benchmarking the accuracy of high-throughput IGHC targeted long-read sequencing in diploid samples

To create test SNV genotype call sets for benchmarking, targeted capture assemblies for the same individuals (fosmid donors and trio probands, as described above) were processed through both the Hifiasm and IGenotyper pipelines. The contigs generated by each computational pipeline were aligned to the IGHC reference to produce assembly alignments (bam files). These bam files were subsequently processed to call SNVs, after masking common SV regions, such as the IGHG4 and IGHG2-IGHA1 inversion regions, to avoid confounding the results. For the highly contiguous Hifiasm assemblies, SNVs were curated using ‘bcftools mpileup’ (52). In the case of IGenotyper assemblies, variant calls were first made directly from the assembly contigs aligned against the reference genome using ‘bcftools mpileup’, followed by filtering of this call set to only include SNVs supported by HiFi read pileups. To do this, HiFi reads were processed using two approaches: SNP calling via ‘bcftools mpileup’ and ‘whatshap find_snv_candidates’ with the ‘--minrel 0.05’ option (50), which includes all variant positions while masking SV regions. The HiFi-based vcf file was constructed by combining SNV positions identified by whatshap and entries from bcftools mpileup, with overlapping variants resolved by prioritizing whatshap calls to create the final HiFi-reads-based vcf. The final HiFi-based vcf and capture-assembly-based vcf were then intersected to produce the most accurate set of variant calls for each individual. Finally, genotype calls from the capture-only assemblies were compared against ground truth sets to evaluate the performance of the high-throughput pipeline.

Identification and genotyping of structural variants

For structural variants in the IGHC region, assembly bam files were manually inspected in IGV. In this study, SV genotyping was mainly focused on the most commonly found SVs in the IGHC region resolved by the assembly process. Alignment files from IGenotyper and Hifiasm assemblies were manually inspected in IGV to identify large deletions (IGHG4D), inversions (between IGHG2 and IGHA1) and hinge deletions (IGHG3), which were then genotyped accordingly. Additional large SVs identified in only a small number of individuals were identified and curated through manual assessment of the assembly contigs (Figure S2). From our targeted capture assemblies we initially observed the presence of a third haplotype-like contig over the IGHG2-IGHA1 region. To be certain with the observation we parallelly looked at the publicly available long-read whole genome sequencing (WGS) data (Pacific Biosciences) of these individuals from Human Pangenome Reference Consortium (HPRC) and confirmed similar instances from soft clipped/ supplementary alignments. From the alignment on the current reference assembly over the IGHE-pseudogene (IGHEP1) region manual inspection of the plausible third haplotype sequence for putative INDELs (2-3 Kbp) and review of the softclipped unaligned portion using the blat toolkit in the UCSC genome browser revealed alignment hits over IGHE and/or IGHG4, hinting at the large segmental duplication (segdup) structure previously described by (16). We further used a 150k-sized k-mer based read simulation from the WGS contigs to confirm the plausible breakpoint using the GRCh38 IGHC reference as the guiding frame (Figure S2). Once confirmed we separated the contigs and used the haplotype-specific contigs from both WGS and targeted capture to merge them based on terminal overlap and sequence identity, producing the complete contiguous haplotype for the duplication and alternate haplotype, respectively. Further we used these haplotypes as a personalized reference for realignment of the HiFi reads for orthogonal validation of the SV.

To generate genotypes for the complex IGHG4 region, we extracted sequences from each haplotype-resolved Hifiasm assembly spanning the IGHG4 region with flanking sequence length of +/- 5 Kbp. This was followed by a length binning approach based on tandem repeat size to ensure comparison among similar copy number variation (CNV) haplotypes, and then multiple sequence alignment was performed for all haplotypes in our cohort. From this alignment, a pairwise distance-based matrix was formulated, followed by hierarchical clustering to identify IGHG4 region spanning haplotype clusters. The distance threshold (0.1) determined the number of clusters and were assigned to haplotypes. Individuals were then genotyped based on the identified haplotype clusters present in each donor.

Identification of IGHC alleles

Allele sequences for the IGHC genes were extracted from haplotype grouped IGenotyper and Hifiasm assembly alignments using a custom script (*https://github.com/williamdlees/VDJbase-Genomics*) to generate an annotation file for all the genes in each sample. IGHC alleles were assigned IMGT allele names for identical sequence matches. Alleles with $<100\%$ identity alignments were assigned to the closest IMGT allele, followed by designation of polymorphic positions following the VDJbase nomenclature schema (*https://vdjbase.org/*). For IGHC genes that were determined to be outside of an SV in a given sample, allele sequences were extracted using the IGHC_gene_coords.bed based on the alignment of a given assembly to the IGHC reference, requiring that a given extracted allele had to include all exons within a single contig. The reference database used for the assignment of “known” and “novel” alleles was downloaded from IMGT (57) in November 2022.

Support from HiFi reads for curated alleles from each assembly contig was assessed using a custom read-support script (*https://github.com/Watson-IG/wasp/tree/master/annotation/read-support*). Specifically, donor-level HiFi reads were realigned to the respective contig assemblies generated for that donor. For each allele, the number of fully-spanning HiFi reads was enumerated, counting the number of HiFi reads with 100%

base agreement with the curated allele across all exons in the personalized alignment. In addition, read support was calculated for each base position of the curated allele from all reads aligned to the personalized reference, regardless of whether that read fully spanned the gene. The percentage of reads supporting the base at that position in the assembly was recorded (using samtools mpileup; (52)). We set our read support threshold to require that at least 10 reads mapped to all exonic positions, and that at least 80% of those reads supported the base represented by the assembly and curated allele.

Results

Building an initial IGHC haplotype reference set from orthogonal long-read datasets and parent-offspring trios

Given the paucity of available highly-vetted haplotype assemblies for the complex IGHC region, we initially focused our effort on constructing high-quality haplotype-resolved assemblies from a selected number of individuals representing diverse genetic ancestry using multiple, orthogonal long-read datasets. To do this, we followed our previous approach (44), using long-read SMRT sequencing data from target-enrichment capture libraries, and overlapping fosmid clones isolated from the same donors (Figure S3). For each donor, an average of ~32 (range, 20-46) fosmid clones were selected and sequenced on the RSII system. Each fosmid was individually assembled and, through manual curation, overlapping clones were joined to create haplotype-resolved contigs. In addition, for each individual, we generated long-read IG-capture libraries using our published protocol (38, 44, 45), which were sequenced on the Sequel IIe system with average read lengths of 5.4 Kbp. Capture reads were used to fill gaps in haplotype-specific fosmid-based assemblies and error correct bases (Figure S3). Together, these data allowed us to construct 8 fully resolved gapless haplotype assemblies from NA18507 (AFR), NA18555 (EAS), NA19129 (AFR), and NA12156 (EUR) spanning the entire IGHC locus.

To construct additional phased reference assemblies, we generated long-read targeted capture data for 4 mother-father-child trios from the 1KGP: HG00671, HG00672, and HG00673 (EAS); HG01926, HG01927, and HG01928 (AMR); HG01950, HG01951, and HG01952 (AMR); and HG02715, HG02716, and HG02717 (AFR). We generated IGHC assemblies using IGenotyper and Hifiasm from HiFi reads for all 12 individuals. Proband contigs were manually inspected alongside the parental assemblies in IGV to select the haplotype-specific contigs based on parent-of-origin (Figure S4). Selected contigs were merged to create phased, gapless assemblies for each proband, resulting in 8 gapless IGHC haplotypes. To assess support from haplotypes generated in the parental assemblies, phased proband haplotype contigs were compared to the assemblies of each parent. For 3 of the 4 probands, we found that parent assemblies spanned >98% of their respective proband haplotypes. However, the maternally inherited haplotype of HG02717 was only partially covered due to haplotype drop-out in the assembly from the mother (HG02716), likely due to a class-switch recombination event (Figure S5).

We validated bases at each position in the hybrid and trio-based proband assemblies by alignment of HiFi reads for each sample back to their personalized IGHC reference. On average, we observed a mean of 41.5X coverage for each capture library relative to the respective haplotype-resolved assemblies. Across samples, we found that a median of 97.5% of haploid assembly bases had at least 10X coverage. For these positions, we calculated the number of mapped reads supporting each base in the contig assembly, relative to the total number of reads mapping to that position. Requiring that at least 80% of mapped reads matched the base found in the assembly, HiFi reads supported >=99.9% of assembly bases (those with >9X coverage) across the 8 diploid assemblies (Table S3).

Curating genetic diversity from high-quality IGHC haplotypes from diverse genetic backgrounds

From these 16 highly vetted haplotypes (Figure 2A), we next curated variant call sets, including SVs and SNVs, as well as IGHC genes and alleles. We detected 3 SVs (ranging in size from 188 bp to 19.5 Kbp) across the IGHC locus (Figure 2A). While earlier studies had noted the occurrence of IGHC SVs (58–60), these represented the first curated descriptions of SVs with nucleotide resolution. The SVs identified among these 8 donors included: (1) an insertion-deletion (19.5 Kbp) variant region spanning the IGHG4 gene region, which was associated with variable IGHG4 copy number; (2) an intergenic inversion (4.5 Kbp) between IGHG2 and

IGHGA1; and (3) an insertion-deletion (indel, 188 bp) variant that resulted in variable copies (3-4) of the IGHG3 hinge exons (Figure 2A). Focusing first on SNVs in gene regions, we next curated IGHC alleles from the 16 haplotypes. Consistent with recent reports (14, 17, 18), we uncovered an extensive number of IGHC alleles, including those that were either incomplete or missing from existing databases (i.e., IMGT). In total, across the 9 annotated IGHC genes, we identified 70 alleles, 77% of which were not fully curated in IMGT (Figure 2B). Considering both allelic variation and SVs, we observed tremendous diversity across these 16 haplotypes (Figure 2A). Remarkably, no two haplotypes in this set of donors shared the same composition of SV and IGHC alleles.

In addition, outside SV regions, we identified 3462 total SNVs across the 8 diploid assemblies relative to our reference assembly. The number of SNVs varied between samples and were composed of both private SNVs (unique to a given sample) and those also observed in at least one other sample (Figure 2C). In general, we observed a greater overlap of SNVs between samples of the same genetic ancestry (Figure 2D). The majority of SNVs resided within intergenic regions, followed by those in IGHC exons and introns (Figure 2C). Interestingly, across the 8 individuals, 7.25% (n=251) and 78% (n=2671) of SNVs were not found in either dbSNP or the 1KGP call sets, respectively (Figure 2E, F). This is a likely reflection of the difficulty in using short reads to detect variants within the IGHC locus. This point was further demonstrated when we directly compared our SNV callsets to those generated by the 1KGP Phase 3 sets at the individual level. In this comparison we detected SNVs that were unique to either our long-read assemblies (~76.4%) or the 1KGP (1.3%) callsets, indicating the likely detection of both false-negative and -positive genotypes called by the 1KGP using short reads (Figure 2G). This is consistent with results we have reported previously for variant calling in the IG loci (38, 44, 45).

Assessing the performance of high-throughput IGHC targeted long-read sequencing in diploid samples

Using the the allele annotations and variant callsets from these highly-vetted hybrid and trio-based assemblies, we next assessed the use of targeted long-read sequencing data alone for generating accurate SNV/SV genotypes and gene/allele calls from haploid-resolved assemblies.

Initially, we evaluated the capture read coverage across the 8 diploid assemblies. For each haplotype, a mean of 95% of the bases were represented by >4X HiFi read coverage (Figure 3A). We next benchmarked assemblies in these same donors using only targeted long-read data, processing the data through two pipelines, IGenotyper (38) and Hifiasm (46). IGenotyper adopts a reference-guided assembly strategy, using reads spanning phased blocks demarcated by haplotype-specific SNVs to generate haplotype-specific local assemblies from locally phased reads with Canu (43). Hifiasm, in contrast, generates *de novo* diploid assemblies. On average, across samples we found that IGenotyper yielded 27 IGHC assembly contigs per donor on average, with a mean length of 31 Kbp (Figure 3B). In contrast, Hifiasm generated 9 contigs per donor on average, with a mean contig length of 114.8 Kbp (Figure 3B). These results indicated that Hifiasm assemblies resulted in longer contigs across the locus overall. However, despite the greater length of Hifiasm contigs, a modest number of phase-switch errors (in which a single contig has both paternally and maternally inherited bases) were observed in 4/8 individuals (Figure 3B). We also observed instances in which Hifiasm contigs included assembly errors caused by reads representing class switch recombination (CSR) events (Figure S6); however, it is worth noting that this issue would not be expected to arise in non-LCL samples.

We next sought to assess the accuracy of SV and SNV calls generated by the IGenotyper and Hifiasm pipelines compared to the ground truth hybrid/trio assembly-based call sets. For the 3 SVs, we observed overall greater concordance between the Hifiasm SV calls and ground truth sets, relative to those produced by IGenotyper (Figure 3C). For example, IGenotyper did not fully and accurately resolve IGHG4 SV genotypes in any sample, whereas Hifiasm assemblies did in all 8 donors. This was due to the fragmented assemblies

generated by IGenotyper over the IGHG4 region. However, the manual joining of IGenotyper assemblies did allow for the curation of IGHG4 single- and 2-copy gene haplotypes in all samples. Likewise, while both Hifiasm and IGenotyper had perfect concordance with ground truth genotype calls for the IGHG3 hinge exon deletion, only Hifiasm accurately called the inversion SV genotype in all samples. Again, fragmented IGenotyper assemblies spanning the inversion required manual curation to resolve SV genotypes correctly in a subset of samples.

To assess SNV call concordance, we considered genotypes at every base in each diploid assembly, whether per base genotypes were homozygous reference, heterozygous, or homozygous for the alternate non-reference allele. This allowed us to assess concordance between ground truth assemblies with IGenotyper and Hifiasm pipeline genotypes at each IGHC region base in the reference assembly. After excluding indels and SV-associated bases, the number of IGHC reference positions for which genotypes were called in the ground-truth datasets for each sample ranged from 288,503 to 288,985 base pairs. For each pipeline, genotypes were only called at a given position if spanned by a diploid assembly and a minimum of >3 HiFi reads. Positions not meeting both of these criteria were classified as “missing”. For positions meeting these criteria, SNVs were called both from the assemblies directly and from mapped reads; in instances in which assembly-based SNV calls and read-based SNV calls were discordant, the pipelines defaulted to using the read-based genotype. After compiling genotype call sets from each pipeline, they were compared to those generated using the ground-truth assemblies. For IGenotyper call sets, the number of “missing” SNVs ranged from 3 to 143 (0.001% to 0.049%)(Figure 3D); for Hifiasm, the range was from 2 to 6761 (0.0006% to 2.34%)(Figure 3D). Notably, the sample with the high number of missing SNVs (n=6761) was caused by contig assembly dropout in this region, despite the occurrence of reads (Figure S7). Of the remaining positions for which SNVs were called in both the ground-truth and pipeline sets, concordance ranged from 99.87% to 99.99% for IGenotyper, and 97.6% to 99.99% for Hifiasm (Figure 3E).

Finally, we assessed the ability of both pipelines to provide accurate IGHC allele calls. To do this, we developed a novel allele curation pipeline for extracting and annotating alleles for each IGHC gene present in either IGenotyper or Hifiasm assemblies, including an assessment of each annotation with respect to both assembly contig and HiFi read support (see Materials and Methods). Using IGenotyper assemblies, the automated pipeline returned correct gene allele calls for all 8 donors over IGHA2, IGHE, IGHG2, IGHA1, IGHG1, and IGHG3 (Figure 3F). Accuracy was lower for the genes IGHG4, IGHD, IGHM, as correct allele genotypes were assigned in only 6 or 7 (depending on the gene) out of 8 donors. For the IGHG4 miscalls, a split alignment over the SV region (IGHG4) inhibited the automated extraction of genes from the haplotype-resolved contigs in a single individual; manual curation of these assemblies revealed the correct alleles. With respect to the IGHM and IGHD miscalls, we observed lower sequencing coverage over these genes in those donors, resulting in either the generation of partial contigs, or loss of contigs representing alternate haplotypes. However, in all cases for these three genes, manual assembly-curation using the long-reads specific to each haplotype revealed that the correct alleles were present in all samples. Similarly, allele curation using Hifiasm assemblies achieved accurate donor assignments for all individuals across IGHA2, IGHE, IGHG2, IGHG1, IGHG3, and IGHD genes (Figure 3F), while correctly identifying 6 or 7 (depending on the gene) individuals for IGHG4, IGHA1, and IGHM. We suspect that the lower read coverage observed in these select samples reflected DNA loss associated with class-switch recombination represented in the LCLs.

Assembling the IGHC locus using long reads in 97 individuals across 5 super-populations

To further characterize IGHC diversity we used IG-capture to conduct targeted long-read sequencing in 97 additional individuals from the 1KGP, increasing our total sample set to 105 individuals (AFR:n=16, AMR:n=28,

EAS:n=22, EUR:n=19, SAS:n=20), encompassing 210 haplotypes. The mean read coverage for the IGHC reference locus across samples was 79.89X, ranging from 33X-250X. We processed all individuals through the IGenotyper and Hifiasm pipelines, to produce both reference guided and *de novo* assemblies. The IGenotyper pipeline produced a mean of 27 contigs with a mean length of 29.76 Kbp. Consistent with our benchmarking analysis (Figure 3), Hifiasm assemblies yielded an average of 12 contigs per sample, with an average length of 80 Kbp; in these samples, the IGHG4 segmental duplication/deletion region was fully resolved within a single contig for 208/210 haplotypes.

Population-level genotyping allows for comprehensive characterization of additional structural variants, SNVs, and IGHC alleles

We first curated large SVs in our expanded cohort. Given the greater accuracy of Hifiasm assembly-based SV genotyping (Figure 3C), we chose to use these assemblies for this population-level assessment. As expected, we identified all three polymorphic SVs identified from the hybrid and trio-proband assemblies (Figure 2). In addition, we discovered the presence of 4 additional SVs, all of which were relatively rare in this cohort, occuring in < 3 individuals. In 2 AFR individuals (HG03516, HG01891), we found evidence for a larger insertion variant, which was initially represented by the apparent presence of multiple distinct Hifiasm contigs overlapping the same region of the reference genome (Figure S2). However, manual curation of these contigs allowed for the reconstruction of a large insertion (~120 Kbp), harboring an additional copy each of IGHE, IGHG4, IGHG2, and IGHA1 (Figure 4A). This SV was also recently discovered in an additional unrelated cohort using Oxford Nanopore Technologies (ONT)-based adaptive sampling sequencing (61). Further, we identified a smaller insertion haplotype in 1 EUR individual (HG00096) that contained an additional copy of IGHG4, resulting in a tandem triplication of IGHG4 genes. In addition to these insertions, we also characterized multiple gene deletion events in EAS individuals. This included a deletion that resulted in the complete loss of both copies of the IGHG4 tandem duplication, which was observed in 2 EAS individuals (NA18543, NA19062); and a second larger deletion in a single EAS individual (NA18956) resulting in a putative fusion of the IGHA2-IGHA1 genes, and the loss of IGHE, IGHG4, and IGHG2 genes (Figure 4A).

After accounting for SVs, we next characterized SNVs across the IGHC locus. We identified a total of 5,889 variant sites; 92 of these sites were multi-allelic. Among these, 2,192 SNVs were observed at a minor allele frequency (MAF) ≥ 5%. Consistent with observations made in our benchmarking dataset, 790 (36%) of these common SNV sites were not represented in dbSNP (Figure 4C, Table S6). Additionally, sample-level comparison of SNV positions identified in our study and those derived from short-read data (1KGP) also revealed that, on average, 73% of SNV positions were uniquely detected in our curated assemblies (Figure 4B, Table S7). This observation is consistent with the well-known limitations of short reads for accurately genotyping genomic regions enriched with duplicated sequences (62–64).

IGHC alleles were annotated for all IGHC ORF genes present in the haplotypes from the 105 individuals. Because the duplicate genes identified in the insertion haplotypes do not yet have official gene names, alleles were aggregated across all paralogs (i.e., IGHE, IGHG4, IGHG2, and IGHA1 duplicates) for simplicity. In total, we identified 262 alleles, only 27 of which (10%) were documented in the IMGT database (Figure 5A). Of the putative novel alleles, coding bases annotated from assemblies for 186 alleles were supported by at least 8 HiFi reads (see Materials and Methods) in at least one individual (Figure S8); additionally, 94 of these alleles were also observed in >1 individual (Figure S8). Putative novel alleles were identified for all 9 IGHC genes, and these novel alleles outnumbered previously curated alleles for each gene. Notably, all IGHD alleles (n=21) identified were novel, not having been previously described or documented. This means that none of the 5 currently documented alleles for IGHD were observed in this cohort, potentially calling their validity into question.

All 105 individuals in our dataset carried at least 1 novel allele, with individuals of African descent harboring ~11 novel alleles on average (range: 8-14); this was in contrast to individuals of European descent, which were found to carry on average ~5 novel alleles (range: 3-8)(Figure 5B; Figure S15). Well-supported undocumented alleles included both non-synonymous and synonymous nucleotide differences, as indicated in the example shown for IGHG3 alleles (Figure 5C; see also Figure S9). The distributions of known and previously undocumented alleles varied widely among super populations, including a number of alleles observed in only a single population (Figure 5C and 5D). Not surprisingly, the extensive number of alleles identified for each gene resulted in high heterozygosity rates for IGHC genes across all populations (Figure 5E), with the majority of individuals carrying >1 allele per IGHC genes (Figure S15). A notable exception was the IGHE gene in both AMR and EUR individuals, which showed reduced heterozygosity relative to other populations and genes (Figure 5E).

SVs and IGHC allele frequencies vary between human populations

Mounting evidence indicates that signatures of haplotype diversity in the IG loci, including examples of SVs and both coding and non-coding SNVs, can differ by genetic ancestry (17, 18, 20, 65–67). To investigate this, we leveraged the comprehensive IGHC SV, SNV, and allele genotype call sets generated here to identify differences between human populations represented in our cohort.

First, for all three common SVs, we observed significant variation in copy number and/or allele frequency variation between populations (Figure 6). For example, we observed that IGHG4 copy number was lower in EAS donors than all other populations (Figure 6A). Similarly, we observed that the inversion SV was invariant among donors of AFR populations, in contrast to all other populations in which both inversion alleles were found (Figure 6B). We also determined the presence of the hinge deletion SV in IGHG3 (Figure 6C). In our study cohort, we identified 14 haplotypes (~7%) lacking a hinge exon in IGHG3, consistent with that previously reported (18). These deletion haplotypes were observed in 12 individuals, two homozygotes (HG02011-AFR and HG03491-SAS) and 10 heterozygotes.

In addition to analysis of copy number variation, population-level analysis of the IGHG4 gene duplication region also revealed additional complexity and suggested the potential of multiple non-recurrent deletions and duplications. Specifically, based on the mapping of single-copy IGHG4 spanning contigs to the reference assembly we identified several distinct putative breakpoints (Figure S10). This signature suggested the potential occurrence of more than one deletion and/or duplication event in this region, as well as the potential for sequence divergence following an SV event. However, given this complexity and the multitude of potential deletion events represented, we concluded that the delineation of specific breakpoints relative to a single reference genome was fraught, and likely to lead to false inferences of breakpoints, SNV genotypes, and mis-assignment of single copy IGHG4 genes and alleles. Thus to genotype this region across the population, we employed a haplotype clustering approach, based on the pairwise distances of alignments of all IGHG4 spanning haplotypes (see **Materials and Methods**). We identified 23 haplotype clusters, including 1 cluster representing 2-copy IGHG4 haplotypes (i.e., sharing the same structure as the reference assembly haplotype), 1 unique 3-copy haplotype, 2 unique haplotypes with a complete IGHG4 deletion (0-copies), and 19 single-copy IGHG4 haplotype-associated clusters. Cluster 1 (a single-copy cluster) was the most frequent across the cohort (n=74 haplotypes), followed by 2-copy IGHG4 haplotypes (n=30) and cluster 2 (n=22 haplotypes). Analysis of haplotype-associated alleles revealed that the most frequent clusters (1 and 2) alongside nine other clusters (5, 6, 9, 10, 11, 13, 15, 17, and 20) primarily harbored *01 and similar alleles (by sequence identity) alleles (novel alleles with a closest match of *01), whereas clusters 7, 8, 12, 16, and 18 carried *04 and related alleles, and clusters 3, 4, 5, and 21 carried *09 and related alleles (Figure 6F).

Notably, when analyzed at the resolution of haplotype clusters, we found that the frequency and diversity of observed haplotypes varied considerably across populations; this was evident for both allele and genotype frequencies **(**Figure 6E**)**. Of single-copy haplotypes, clusters 7 and 11 were predominant among AFR individuals, whereas cluster 1 was the most common among EUR, AMR and SAS groups alongside the reference like 2-copy IGHG4 haplotype (cluster 22). In addition, cluster 4 was observed in 17 out of 42 haplotypes in the EAS population. Interestingly, a greater number of clusters were observed in AFR and AMR individuals (9 and 11, respectively) whereas EUR had the least variation with 6 clusters.

Coding and non-coding SNVs highlight regions of IGHC exhibiting intra-population differences

A strength of our dataset is our accounting for the presence of SVs in our assemblies for improved SNV calling and genotyping. Thus, in addition to assessing population-level variation for SVs, we asked whether SNVs outside of SV regions also exhibited signatures of population divergence. We first assessed this by conducting PCA using common SNVs in non-SV regions; individuals with large SVs (n=2, AFR; n=1, EAS) were excluded from this analysis. Assessment of the first three principal components (PCs) indeed revealed separation by super-population (Figure 7A & B); in particular, EAS and AFR samples were the most differentiated along PC2 and PC1, respectively. We next tested SNV-level allele frequency differentiation using pairwise F_{ST} (Weir and Cockeram (55)), comparing each population group against all remaining groups collectively. Consistent with results from the PCA, drastically higher F_{ST} values were observed for AFR (F_{ST} as high as 0.95) and EAS (F_{ST} as high as 0.86). For each of the five population comparisons, 742 (AFR), 0 (AMR), 280 (EAS), 74 (EUR), and 0 (SAS) SNVs exceeded an F_{ST} threshold of 0.3. Interestingly, high-F_{ST} (>0.3) sites were observed in IGHG1 and IGHG3 for three of the population comparisons (AFR, EAS and EUR). However, strong differentiation of AFR was also uniquely found in IGHM, IGHA1, and IGHG2, whereas no high-F_{ST} SNVs in these genes were observed in the other population comparisons (Figure S11).

Discussion

IGHC genes encode critical domains of expressed Abs, facilitating downstream effector immune function(s). Despite the biological importance of IGHC genes, their roles in health and disease, and even their relevance to therapeutic development, our understanding of IGHC genetic diversity has remained extremely limited. Here, extending our previously published methods, we demonstrated the utility of a high-throughput molecular and informatics protocol that leverages highly accurate, long-read sequencing for reconstructing haplotype-resolved assemblies across the IGHC locus. From benchmarking using an initial set of vetted assemblies, we first showed that our pipeline provides accurate, locally phased assembly contigs representing all IGHC genes across the locus using a single molecular assay. Further, with our automated analysis workflow we were able to generate reliable SNV and SV callsets and curate alleles (both novel and known) from accurately mapped IGHC genes. Second, by applying this approach at scale in the largest surveyed cohort to date, we characterized an extensive range of genomic variations, including undocumented SNVs and SVs, as well as novel IGHC genes and coding alleles, each with substantial inter-population variation. For the first time, these data provide a snapshot of the tremendous IGHC diversity circulating in the human population, including the most broadly representative collection of genetic ancestries.

The IGHC locus as represented in current linear reference genomes (i.e., GRCh38 and T2T) is composed of many large segmental duplications, including duplications of functional and pseudo IGHC genes. It is now understood that these references are missing genes present across populations. This complex genomic architecture and “missing” data has thwarted previous efforts to accurately reconstruct IGHC haplotypes with nucleotide-level resolution using short-read sequencing. As a result, our knowledge of SVs in the IGHC region almost entirely draws from those inferred from RFLP, targeted PCRs, and expressed Ab repertoire data (4, 15, 16, 59, 68). Thus, even though large polymorphic duplications and deletions in the IGHC locus have been indicated since the 1990s (4, 13, 15, 59, 60), the majority of these have remained undescribed and uncurated. We recently described complete haplotype-resolved IGHC assemblies using a combination of Oxford Nanopore and PacBio HiFi data from multiple donors, allowing for the first characterization of several large duplications, each harboring additional copies of IGHC genes, which had never been described (61). The largest of these included a duplication of ~120 Kbp, spanning 4 functional genes. Critically, these duplicated genes, which represented both IgG and IgA isotypes and subisotypes were shown to be expressed, highlighting deficits in our basic knowledge of IG isotype gene regulation and fundamental Ab biology. Here, we drastically expand our prior discovery and survey of IGHC SVs, identifying 7 SVs in total across 210 haplotypes and five superpopulations. Importantly, our assembly resolution allowed us to delineate the presence of multiple overlapping SVs across the locus, indicating that the IGHC locus as a whole, is a SV hotspot, and a site of recurrent events. With respect to the large multi-gene SV noted above, we characterized IGHC haplotypes from 2 individuals (HG01891, HG03516) carrying a previously described large segmental duplication and additionally resolved a multigene deletion haplotype in one EAS individual (NA18956) that resulted in the fusion of IGHA2 and IGHA1; resolution of these complex haplotypes was confirmed using a combination of targeted long-read assembly, publicly available WGS, fosmids and/or supporting parental assemblies with manual inspection over each contig. In a second subregion of IGHC, spanning IGHG4 and its close paralogs, we identified 23 distinct haplotype clusters, which were defined by variation in IGHG4 gene copy number (0-3 haploid copies), differences in nucleotide/SNV profiles, coding alleles, as well as different putative SV breakpoints. Notably, IGHG4 was previously reported to represent a site of local population-specific adaptation and neanderthal introgression (20). Consistent with this, in our cohort we observed that the frequency of IGHG4 SV and gene allele haplotypes varied considerably between human populations, with the AFR population exhibiting the greatest diversity in terms of the number of unique clusters and their frequency distributions.

Our resolution of SVs also allowed us to improve the accuracy of SNV detection and genotyping. Specifically, haplotype-level characterization of SVs allowed us to improve the accuracy of SNV detection and genotyping, mitigating errors associated with the presence of segmental duplications that conflate paralogous sequence variants (PSVs) with SNVs. Demonstrating this, we compared the variant call set derived from our long-read sequencing data to that obtained by the 1KGP for these same samples. Interestingly, we found that a median of 74.3% of SNVs were only detected in our long-read-based variant call set. This was also true for comparisons to the dbSNP-common set, for which we found that 36% of common (MAF ≥ 5%) SNVs in our dataset were not present in dbSNP. These observations emphasize the substantial contribution of long-read sequencing to uncovering genetic variation that may have been missed or underrepresented in previous datasets. Notably, even at the individual level, we observed that on average, 73% of SNVs genotyped in our long-read data were not called by the 1KGP. These data indicate the additional work needed to better integrate missing variation in genetic association studies.

The same pitfalls that obfuscate SNV genotyping in complex loci can also impede accurate curation of genes and alleles. Our data clearly demonstrate this to be true for the IGHC genes. First, among the most remarkable findings from this cohort was the extraordinary number of undocumented IGHC alleles identified. Across the 9 core functional genes (and their newly curated paralogs) we identified a total 262 alleles, 235 (89.6%) of which were not documented in IMGT. This includes 21 IGHD alleles, none of which had been previously documented. The IMGT version we used for this paper (downloaded November 24, 2022) contained only 100 IGHC alleles. Thus, the collection of newly discovered alleles here represents a 235% increase. Importantly, we noted that a large fraction of these undocumented alleles came from donors of non-European ancestries, stressing the ongoing need to generate improved genomic resources for the IG loci that are more representative of global populations. This is consistent with observations more recently made in the V, D, and J genes for IGH, IGK and IGL as well (14, 37, 44, 45, 65, 69, 70). To contribute to this growing effort, data from this study and from our growing collection of long-read based assemblies, will be made public in VDJbase (69) to enhance the representation of population diversity in existing IGHC germline allele databases. From VDJbase, users can browse and download annotated assemblies, allele sequences and usage statistics. Previously undocumented allele sequences are in the process of being submitted to the Nomenclature Review Committee of the International Union of Immunological Societies for ratification and naming.

In addition to cataloging a larger set of alleles, our long-read haplotype data also afforded us the ability to map these alleles to specific gene loci. To date, the majority of curated alleles have been identified and named lacking knowledge and context of gene duplications, raising uncertainty in the gene assignments of existing alleles for genes with uncharacterized paralogs. A clear example is IGHG4 and related paralogs. Of the 13 alleles currently assigned to IGHG4 and the IGHG4 duplicate in IMGT, 10 of these alleles are assigned to IGHG4; however, the majority of these alleles were curated from targeted gene sequencing, rather than full-locus IGHC sequencing and assembly. Thus, whether these 10 alleles have been correctly assigned to their proper IGHG4 locus will need to be reassessed, deprecated and renamed when necessary. This is further complicated by the fact that, in addition to 3-copy IGHG4 haplotypes (which to date have not been discussed in the literature), we observed significant sequence variation among single-copy IGHG4 haplotypes including distinct deletion events, and carrying different IGHG4 paralogs. Thus, careful phylogenetic analysis will be required that accounts for this complex duplication structure, in order to more fully delineate relationships between IGHG4 genes and alleles. Together, these observations illuminate significant flaws in our current nomenclature system.

An added strength of our comprehensive haplotype dataset for the curation of variants is our sampling of donors from multiple global populations, allowing us to assess differences in IGHC haplotype variation between genetic ancestries. Previous targeted IGHC gene sequencing studies have shown that allele frequencies vary significantly between human populations, and that coding variants exhibit high pairwise F_{ST} values between certain populations (18). In line with these observations, a genome-wide survey also found that a small 33 bp SV within IGHG4 was observed at a significantly higher frequency in southeast Asian populations (20). Here, we showed that these signatures extend to large SVs and SNVs in both coding and non-coding regions. Every common SV we assessed showed significant variation between populations. SV allele frequencies in AFR donors tended to deviate most from the other populations; for example, inversion SV genotypes were variable in all populations except AFR, in which only a single allele was observed in all samples. Similarly, AFR donors only carried single-copy IGHG4 haplotypes, whereas deletion haplotypes and 3-copy haplotypes were observed uniquely in EAS and EUR populations, respectively. Additionally, a specific single-copy IGHG4 haplotype (cluster 4) was only observed in EAS and SAS, and particularly enriched in EAS; this haplotype was primarily associated with alleles most similar to IGHG4*09, and also confirmed to represent haplotypes previously reported to be sites of local adaptation and introgression (20) (Figure S12). Expanding population comparisons to locus-wide SNVs also demonstrated that both coding and non-coding variants contributed to population stratification. This was demonstrated through analysis of F_{ST} as well as through PCA, revealing that while the majority of IGHC intergenic and genic regions were associated with high F_{ST} estimates when comparing AFR to all other populations combined, smaller subregions of the locus surrounding only a subset of IGHC genes were associated with differentiating signatures for EAS and EUR as well. Interestingly, across all population comparisons, F_{ST} values were more modest in IGHD and IGHM regions, suggesting that these regions are potentially more homogeneous across genetic ancestries. These patterns may point to historical demographic events, geographic isolation, co-evolution with localized pathogenic environments, and/or variation in selective pressures. However, considering the relatively small subpopulation sample sizes collected here, further analysis in larger cohorts will be required to investigate these scenarios.

Together these data provide evidence that IGHC diversity has been vastly underestimated and is deserving of more focused study. In addition to further clarifying population differentiation in the IGHC locus, future work should seek to understand how this genetic diversity influences the composition of the expressed Ab repertoire and contributes to variation in Ab effector function(s). Allelic diversity has been demonstrated to influence functional outcomes by altering FcR and IG interactions, thereby affecting Ab half-life (25). Additionally, SVs have been linked to differential ADCC and enhanced complement activation, including IGHG3 hinge copy number (71, 72). IGHG4, which we have shown boasts extensive copy number variation differences across different ancestries, is generally associated with low inflammatory activity and weak complement activation. However, coding variants of IGHG4 have been reported to be associated with elevated expression levels, as well as variation in Fab arm exchange and patterns of glycosylation (73, 74). In addition to coding variation, it will also be worth exploring whether non-coding variants influence biases in CSR frequencies, leading to shifts in the distribution of particular IGHC genes and linked IGHV, D, and J transcripts. Our recent studies of near-full length IgM and IgG transcripts (14, 61) have revealed that multi-copy IGHC genes (such as IGHG4 and IGHA and their close paralogs) are detected in the Ab repertoire. Using these more recently developed long-read based approaches will be important for identifying interindividual variation in the Ab repertoire, accounting for full BCR transcripts, and identifying potential disease-associated signatures.

In conclusion, our population-level survey has revealed that the human IGHC locus harbors extensive genetic diversity, including variants that differentiate human populations, with the likely potential of contributing to Ab function and disease. In addition, we have shown that our work in this locus is only just beginning, and continued efforts to characterize haplotype variation at scale will be needed to delineate functional variants of biomedical importance. We show here that the use of long-read sequencing offers a tractable solution to this challenge. However, future work should focus on expanding the capacity of existing molecular and analysis tools to allow for more automated assessments of long-read based assemblies, with respect to their contiguity and phasing accuracy, as well as ensuring precision in SV and SNV discovery and genotyping. Finally, the discovered haplotype complexity also demands greater attention to the development of pangenome reference resources that can better represent haplotype variation (63, 75–77). Moving forward, these will allow for the more ordered cataloguing of IGHC genetic variants of all types, as well as the curation of genes and alleles with new and improved nomenclatures that more aptly reflect the population and evolutionary history of these loci.

Acknowledgements

We sincerely thank Eric Rouchka, Juw Won Park for their valuable insights and feedback on the computational analysis. We also appreciate the contributions of Easton Ford, Kamille Rasche, Chandrima Bharadwaj for their assistance with manuscript preparation and a special thanks to Andy Lauer and the staff of the UofL Sequencing Technology Center for assistance in the sequencing.

Funding

This work was supported by an award made to C.T.W. and M.L.S. by the National Institute of Allergy and Infectious Disease (R24 AI138963) and in part, to E.E.E by US National Institutes of Health (NIH) grant R01HG010169. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Conflicts of interest statement

C.T.W., M.L.S., and W.L. are founders and shareholders of Clareo Biosciences, Inc. and serve on its Executive Board. M.L.S. and C.T.W. are listed inventors of patent filing PCT/US2024/044692. E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc. None of the other authors declare any competing interests.

References

1. Murphy,K., Travers,P., Walport,M. and Walter,P. (2010) Janeway's Immunobiology Garland Science.

2. Lu,R.-M., Hwang,Y.-C., Liu,I.-J., Lee,C.-C., Tsai,H.-Z., Li,H.-J. and Wu,H.-C. (2020) Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.*, **27**, 1–30.

3. Lefranc,M.-P. and Lefranc,G. (2020) Immunoglobulins or Antibodies: IMGT® Bridging Genes, Structures and Functions. *Biomedicines*, **8**, 319.

4. Brusco,A., Cariota,U., Bottaro,A., Boccazzi,C., Plebani,A., Ugazio,A.G., Galanello,R., Guerra,M.G. and Carbonara,A.O. (1995) Variability of the immunoglobulin heavy chain constant region locus: a population study. *Hum. Genet.*, **95**, 319–326.

5. Vidarsson,G., Dekkers,G. and Rispens,T. (2014) IgG subclasses and allotypes: from structure to effector functions. *Front. Immunol.*, **5**, 520.

6. Dard,P., Lefranc,M.-P., Osipova,L. and Sanchez-Mazas,A. (2001) DNA sequence variability of IGHG3 alleles associated to the main G3m haplotypes in human populations. *Eur. J. Hum. Genet.*, **9**, 765–772.

7. White,M.B., Shen,A.L., Word,C.J., Tucker,P.W. and Blattner,F.R. (1985) Human Immunoglobulin D: Genomic Sequence of the Delta Heavy Chain. *Science*, 10.1126/science.3922054.

8. Bournazos,S. and Ravetch,J.V. (2017) Diversification of IgG effector functions. *Int. Immunol.*, **29**, 303–310.

9. Goldberg,B.S. and Ackerman,M.E. (2020) Antibody-mediated complement activation in pathology and protection. *Immunol. Cell Biol.*, **98**, 305–317.

10. Stapleton,N.M., Einarsdóttir,H.K., Stemerding,A.M. and Vidarsson,G. (2015) The multiple facets of FcRn in immunity. *Immunol. Rev.*, **268**, 253–268.

11. Garzón-Ospina,D. and Buitrago,S.P. (2022) Immunoglobulin heavy constant gamma gene evolution is modulated by both the divergent and birth-and-death evolutionary models. *Primates*, **63**, 611–625.

12. Migone,N., de Lange,G., Piazza,A. and Cavalli-Sforza,L.L. (1985) Genetic analysis of eight linked polymorphisms within the human immunoglobulin heavy-chain region. *Am. J. Hum. Genet.*, **37**, 1146–1163.

13. Wiebe,V., Helal,A., Lefranc,M.P. and Lefranc,G. (1994) Molecular analysis of the T17 immunoglobulin CH multigene deletion (del A1-GP-G2-G4-E). *Hum. Genet.*, **93**, 520–528.

14. Ford,E.E., Tieri,D., Rodriguez,O.L., Francoeur,N.J., Soto,J., Kos,J.T., Peres,A., Gibson,W.S., Silver,C.A., Deikus,G., *et al.* (2023) FLAIRR-Seq: A Method for Single-Molecule Resolution of Near Full-Length Antibody H Chain Repertoires. *J. Immunol.*, **210**, 1607–1619.

15. Rabbani,H., Kondo,N., Smith,C.I. and Hammarström,L. (1995) The influence of gene deletions and duplications within the IGHC locus on serum immunoglobulin subclass levels. *Clin. Immunol. Immunopathol.*, **76**, S214–8.

16. Bottaro,A., Cariota,U., DeMarchi,M. and Carbonara,A.O. (1991) Pulsed-field electrophoresis screening for immunoglobulin heavy-chain constant-region (IGHC) multigene deletions and duplications. *Am J Hum Genet*, **48**, 745–756.

17. Calonga-Solís,V., Malheiros,D., Beltrame,M.H., Vargas,L. de B., Dourado,R.M., Issler,H.C., Wassem,R., Petzl-Erler,M.L. and Augusto,D.G. (2019) Unveiling the Diversity of Immunoglobulin Heavy Constant Gamma () Gene Segments in Brazilian Populations Reveals 28 Novel Alleles and Evidence of Gene Conversion and Natural Selection. *Front. Immunol.*, **10**, 1161.

18. Bashirova,A.A., Zheng,W., Akdag,M., Augusto,D.G., Vince,N., Dong,K.L., O'hUigin,C. and Carrington,M. (2021) Population-specific diversity of the immunoglobulin constant heavy G chain (IGHG) genes. *Genes & Immunity*, **22**, 327–334.

19. Lefranc,M.-P., Giudicelli,V., Kaas,Q., Duprat,E., Jabado-Michaloud,J., Scaviner,D., Ginestoux,C., Clément,O., Chaume,D. and Lefranc,G. (2005) IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.*, **33**, D593–7.

20. Yan,S.M., Sherman,R.M., Taylor,D.J., Nair,D.R., Bortvin,A.N., Schatz,M.C. and McCoy,R.C. (2021) Local adaptation and archaic introgression shape global diversity at human structural variant loci. *Elife*, **10**.

21. Warrender,A.K. and Kelton,W. (2020) Beyond Allotypes: The Influence of Allelic Diversity in Antibody Constant Domains. *Front. Immunol.*, **11**, 2016.

22. Hassan,M.S., Islam,K.B., Hammarström,L. and Smith,C.I. (1992) Regulation of C gamma 3 expression. Role of switch in the allotype-associated variation of human serum IgG3 levels. *J. Immunol.*, **148**, 2555–2562.

23. Pandey,J.P. and French,M.A. (1996) GM phenotypes influence the concentrations of the four subclasses of immunoglobulin G in normal human serum. *Hum. Immunol.*, **51**, 99–102.

24. Lefranc,M.-P. and Lefranc,G. (2012) Human Gm, Km, and Am allotypes and their molecular characterization: a remarkable demonstration of polymorphism. *Methods Mol. Biol.*, **882**, 635–680.

25. Ternant,D., Arnoult,C., Pugnière,M., Dhommée,C., Drocourt,D., Perouzel,E., Passot,C., Baroukh,N., Mulleman,D., Tiraby,G., *et al.* (2016) IgG1 Allotypes Influence the Pharmacokinetics of Therapeutic Monoclonal Antibodies through FcRn Binding. *J. Immunol.*, **196**, 607–613.

26. de Taeye,S.W., Bentlage,A.E.H., Mebius,M.M., Meesters,J.I., Lissenberg-Thunnissen,S., Falck,D., Sénard,T., Salehi,N., Wuhrer,M., Schuurman,J., *et al.* (2020) FcγR Binding and ADCC Activity of Human IgG Allotypes. *Front. Immunol.*, **11**, 740.

27. Pandey,J.P., Kistner-Griffin,E., Radwan,F.F., Kaur,N., Namboodiri,A.M., Black,L., Butler,M.A., Carreón,T. and Ruder,A.M. (2014) Immunoglobulin genes influence the magnitude of humoral immunity to cytomegalovirus glycoprotein B. *J. Infect. Dis.*, **210**, 1823–1826.

28. Simon,B., Weseslindtner,L., Görzer,I., Pollak,K., Jaksch,P., Klepetko,W. and Puchhammer-Stöckl,E. (2016) Subclass-specific antibody responses to human cytomegalovirus in lung transplant recipients and their association with constant heavy immunoglobulin G chain polymorphism and virus replication. *J. Heart Lung Transplant.*, **35**, 370–377.

29. Obaid,A.H., Zografou,C., Vadysirisack,D.D., Munro-Sheldon,B., Fichtner,M.L., Roy,B., Philbrick,W.M., Bennett,J.L., Nowak,R.J. and O'Connor,K.C. (2022) Heterogeneity of Acetylcholine Receptor Autoantibody-Mediated Complement Activity in Patients With Myasthenia Gravis. *Neurol Neuroimmunol Neuroinflamm*, **9**.

30. Yandamuri,S.S., Filipek,B., Obaid,A.H., Lele,N., Thurman,J.M., Makhani,N., Nowak,R.J., Guo,Y., Lucchinetti,C.F., Flanagan,E.P., *et al.* (2023) MOGAD patient autoantibodies induce complement, phagocytosis, and cellular cytotoxicity. *JCI Insight*, **8**.

31. Bastard,P., Rosen,L.B., Zhang,Q., Michailidis,E., Hoffmann,H.-H., Zhang,Y., Dorgham,K., Philippot,Q., Rosain,J., Béziat,V., *et al.* (2020) Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science*, **370**.

32. Miga,K.H. and Wang,T. (2021) The Need for a Human Pangenome Reference Sequence. *Annu Rev Genomics Hum Genet*, **22**, 81–102.

33. Taylor,D.J., Eizenga,J.M., Li,Q., Das,A., Jenike,K.M., Kenny,E.E., Miga,K.H., Monlong,J., McCoy,R.C., Paten,B., *et al.* (2024) Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References. *Annu Rev Genomics Hum Genet*, **25**, 77–104.

34. Huddleston,J., Ranade,S., Malig,M., Antonacci,F., Chaisson,M., Hon,L., Sudmant,P.H., Graves,T.A., Alkan,C., Dennis,M.Y., *et al.* (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.*, **24**, 688–696.

35. Engelbrecht,E., Rodriguez,O.L. and Watson,C.T. (2024) Addressing technical pitfalls in pursuit of molecular factors that mediate immunoglobulin gene regulation. *bioRxiv*, 10.1101/2024.03.06.582860.

36. Rodriguez,O.L., Silver,C.A., Shields,K., Smith,M.L. and Watson,C.T. (2022) Targeted long-read sequencing facilitates phased diploid assembly and genotyping of the human T cell receptor alpha, delta, and beta loci. *Cell Genomics*, **2**.

37. Rodriguez,O.L., Safonova,Y., Silver,C.A., Shields,K., Gibson,W.S., Kos,J.T., Tieri,D., Ke,H., Jackson,K.J.L., Boyd,S.D., *et al.* (2023) Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat. Commun.*, **14**, 4419.

38. Rodriguez,O.L., Gibson,W.S., Parks,T., Emery,M., Powell,J., Strahl,M., Deikus,G., Auckland,K., Eichler,E.E., Marasco,W.A., *et al.* (2020) A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. *Front. Immunol.*, **11**, 2136.

39. Warburton,P.E. and Sebra,R.P. (2023) Long-Read DNA Sequencing: Recent Advances and Remaining Challenges. *Annu Rev Genomics Hum Genet*, **24**, 109–132.

40. Mahmoud,M., Huang,Y., Garimella,K., Audano,P.A., Wan,W., Prasad,N., Handsaker,R.E., Hall,S., Pionzio,A., Schatz,M.C., *et al.* (2024) Utility of long-read sequencing for All of Us. *Nat Commun*, **15**, 837.

41. 1000 Genomes Project Consortium, Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

42. Kidd,J.M., Cooper,G.M., Donahue,W.F., Hayden,H.S., Sampas,N., Graves,T., Hansen,N., Teague,B., Alkan,C., Antonacci,F., *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.

43. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.

44. Gibson,W.S., Rodriguez,O.L., Shields,K., Silver,C.A., Dorgham,A., Emery,M., Deikus,G., Sebra,R., Eichler,E.E., Bashir,A., *et al.* (2023) Characterization of the immunoglobulin lambda chain locus from diverse populations reveals extensive genetic variation. *Genes Immun.*, **24**, 21.

45. Engelbrecht,E., Rodriguez,O.L., Shields,K., Schultze,S., Tieri,D., Jana,U., Yaari,G., Lees,W.D., Smith,M.L. and Watson,C.T. (2024) Resolving haplotype variation and complex genetic architecture in the human immunoglobulin kappa chain locus in individuals of diverse ancestry. *Genes Immun.*, 10.1038/s41435-024-00279-2.

46. Cheng,H., Concepcion,G.T., Feng,X., Zhang,H. and Li,H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*, **18**, 170–175.

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

48. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

49. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.

50. Martin,M., Ebert,P. and Marschall,T. (2023) Read-Based Phasing and Analysis of Phased Variants with WhatsHap. *Methods Mol. Biol.*, **2590**, 127–138.

51. Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

52. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M., *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**.

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

54. Zheng,X., Levine,D., Shen,J., Gogarten,S.M., Laurie,C. and Weir,B.S. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.

55. Weir,B.S. and Cockerham,C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358.

56. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T., *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

57. Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.*, **33**, D256–D261.

58. Bottaro,A., Cariota,U., de Lange,G.G., DeMarchi,M., Gallina,R., Oliviero,S., Vlug,A. and Carbonara,A.O. (1990) Multiple levels of analysis of an IGHG4 gene deletion. *Hum. Genet.*, **86**, 191–197.

59. Brusco,A., Saviozzi,S., Cinque,F., Bottaro,A. and DeMarchi,M. (1999) A recurrent breakpoint in the most common deletion of the Ig heavy chain locus (del A1-GP-G2-G4-E). *J. Immunol.*, **163**, 4392–4398.

60. Lefranc,M.P. and Lefranc,G. (1987) Human immunoglobulin heavy-chain multigene deletions in healthy individuals. *FEBS Lett.*, **213**, 231–237.

61. Gornitzka,M.B., Røsjø,E., Jana,U., Ford,E.E., Tourancheau,A., Lees,W.D., Vanwinkle,Z., Smith,M.L., Watson,C.T. and Lossius,A. (2024) Ultra-long sequencing for contiguous haplotype resolution of the human immunoglobulin heavy chain locus. *bioRxivorg*, 10.1101/2024.12.14.628445.

62. Ebbert,M.T.W., Jensen,T.D., Jansen-West,K., Sens,J.P., Reddy,J.S., Ridge,P.G., Kauwe,J.S.K., Belzil,V., Pregent,L., Carrasquillo,M.M., *et al.* (2019) Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol.*, **20**, 97.

63. Liao,W.-W., Asri,M., Ebler,J., Doerr,D., Haukness,M., Hickey,G., Lu,S., Lucas,J.K., Monlong,J., Abel,H.J., *et al.* (2023) A draft human pangenome reference. *Nature*, **617**, 312–324.

64. Watson,C.T. and Breden,F. (2012) The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun.*, **13**, 363–373.

65. Watson,C.T., Steinberg,K.M., Huddleston,J., Warren,R.L., Malig,M., Schein,J., Willsey,A.J., Joy,J.B., Scott,J.K., Graves,T.A., *et al.* (2013) Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet*, **92**, 530–546.

66. Avnir,Y., Watson,C.T., Glanville,J., Peterson,E.C., Tallarico,A.S., Bennett,A.S., Qin,K., Fu,Y., Huang,C.-Y., Beigel,J.H., *et al.* (2016) Erratum: IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep*, **6**, 23876.

67. Luo,S., Yu,J.A., Li,H. and Song,Y.S. (2019) Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans. *Life Sci Alliance*, **2**.

68. Flanagan,J.G. and Rabbitts,T.H. (1982) Arrangement of human immunoglobulin heavy chain constant region genes implies evolutionary duplication of a segment containing gamma, epsilon and alpha genes. *Nature*, **300**, 709–713.

69. Omer,A., Shemesh,O., Peres,A., Polak,P., Shepherd,A.J., Watson,C.T., Boyd,S.D., Collins,A.M., Lees,W. and Yaari,G. (2020) VDJbase: an adaptive immune receptor genotype and haplotype database. *Nucleic Acids Res*, **48**, D1051–D1056.

70. Marsden,A.A., Corcoran,M., Hedestam,G.K., Garrett,N., Karim,S.S.A., Moore,P.L., Kitchin,D., Morris,L. and Scheepers,C. (2024) Novel polymorphic and copy number diversity in the antibody IGH locus of South African individuals. *Immunogenetics*, **77**, 6.

71. Michaelsen,T.E., Aase,A., Westby,C. and Sandlie,I. (1990) Enhancement of complement activation and cytolysis of human IgG3 by deletion of hinge exons. *Scand. J. Immunol.*, **32**, 517–528.

72. Giuntini,S., Granoff,D.M., Beernink,P.T., Ihle,O., Bratlie,D. and Michaelsen,T.E. (2016) Human IgG1, IgG3, and IgG3 Hinge-Truncated Mutants Show Different Protection Capabilities against Meningococci Depending on the Target Antigen and Epitope Specificity. *Clin. Vaccine Immunol.*, **23**, 698–706.

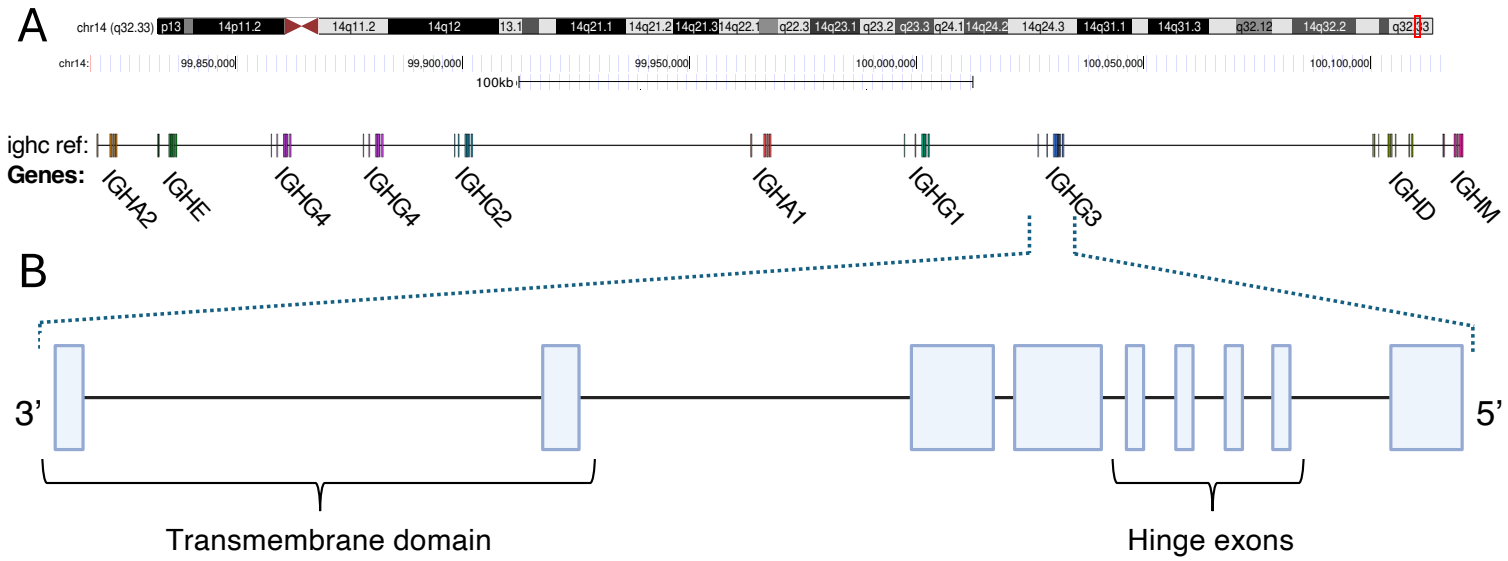
73. Damelang,T., Brinkhaus,M., van Osch,T.L.J., Schuurman,J., Labrijn,A.F., Rispens,T. and Vidarsson,G. (2023) Impact of structural modifications of IgG antibodies on effector functions. *Front Immunol*, **14**, 1304365.

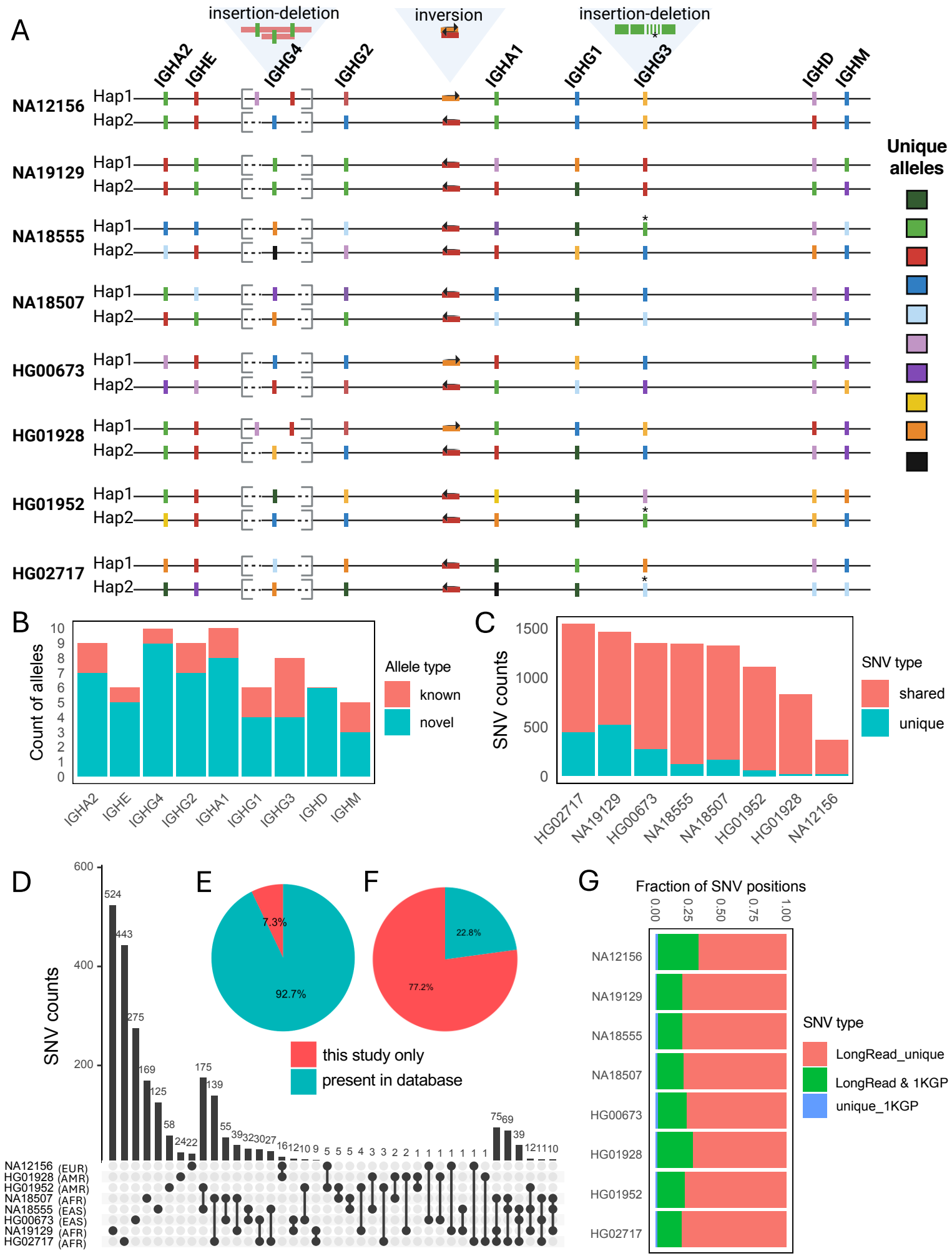
74. Oskam,N., Damelang,T., Streutker,M., Ooijevaar-de Heer,P., Nouta,J., Koeleman,C., Van Coillie,J., Wuhrer,M., Vidarsson,G. and Rispens,T. (2023) Factors affecting IgG4-mediated complement activation. *Front Immunol*, **14**, 1087532.

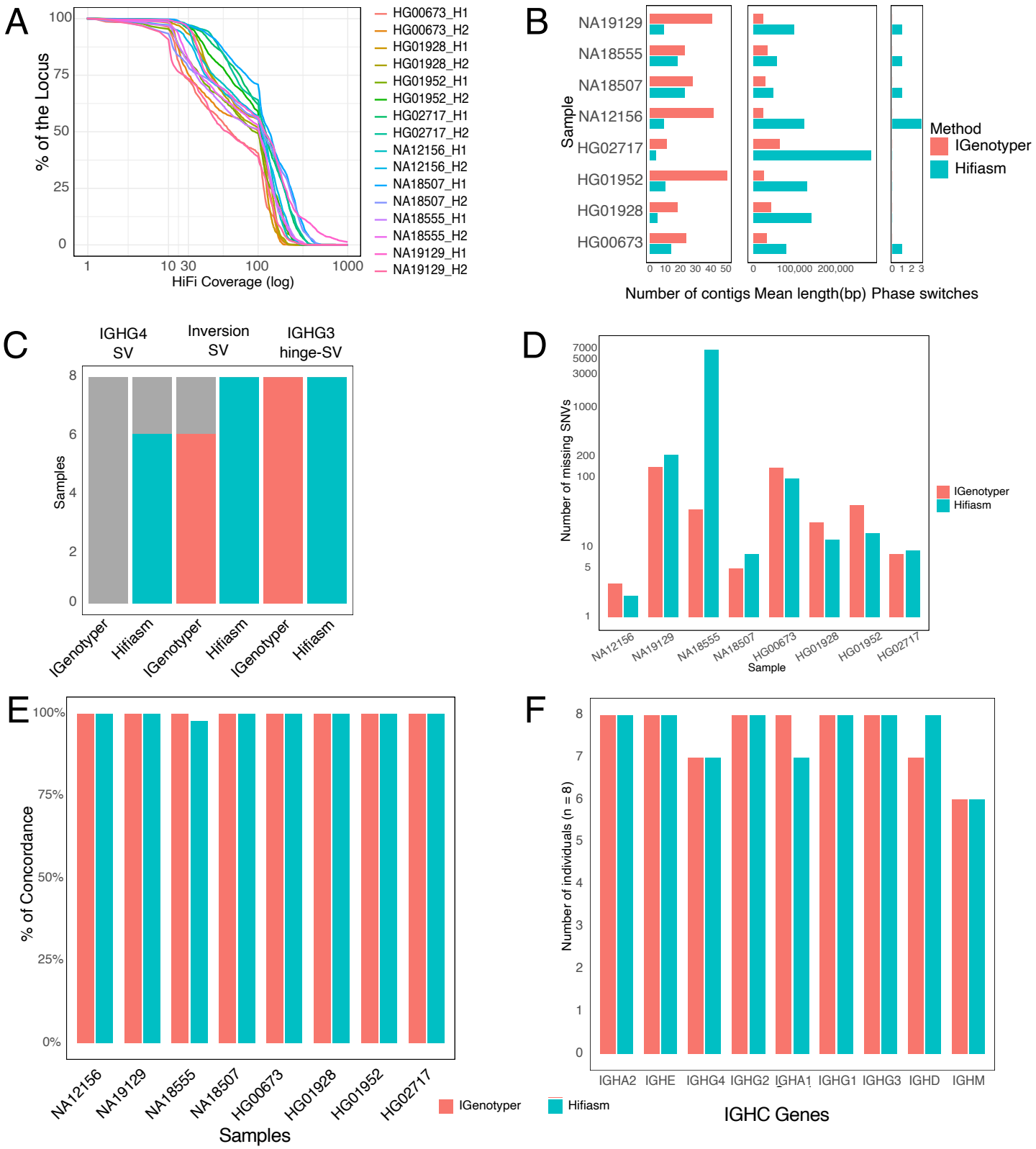
75. Ebler,J., Ebert,P., Clarke,W.E., Rausch,T., Audano,P.A., Houwaart,T., Mao,Y., Korbel,J.O., Eichler,E.E., Zody,M.C., *et al.* (2022) Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet*, **54**, 518–525.

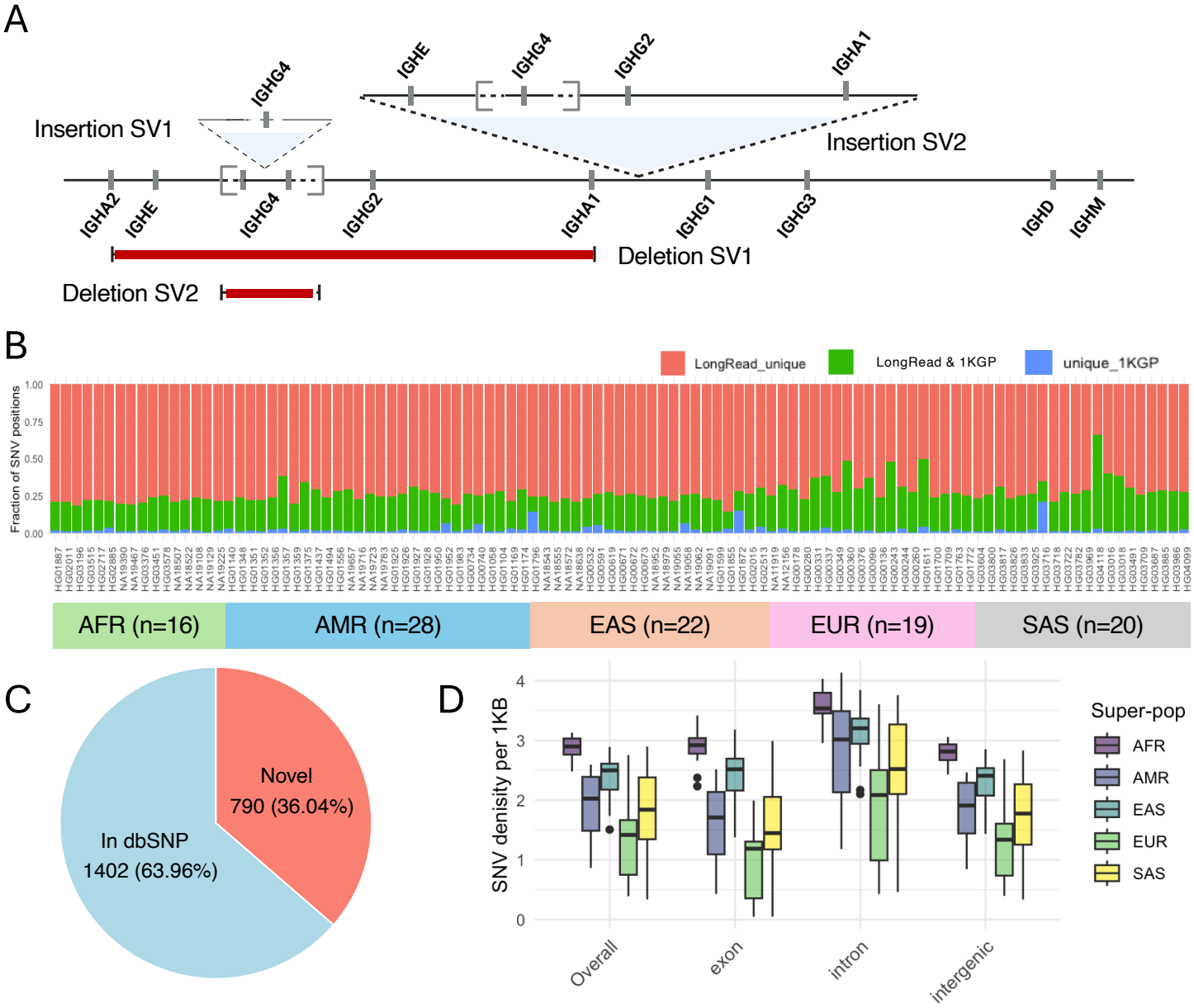
76. Hung,T.-K., Liu,W.-C., Lai,S.-K., Chuang,H.-W., Lee,Y.-C., Lin,H.-Y., Hsu,C.-L., Chen,C.-Y., Yang,Y.-C., Hsu,J.S., *et al.* (2024) Genetic complexity of killer-cell immunoglobulin-like receptor genes in human pangenome assemblies. *Genome Res*, **34**, 1211–1223.

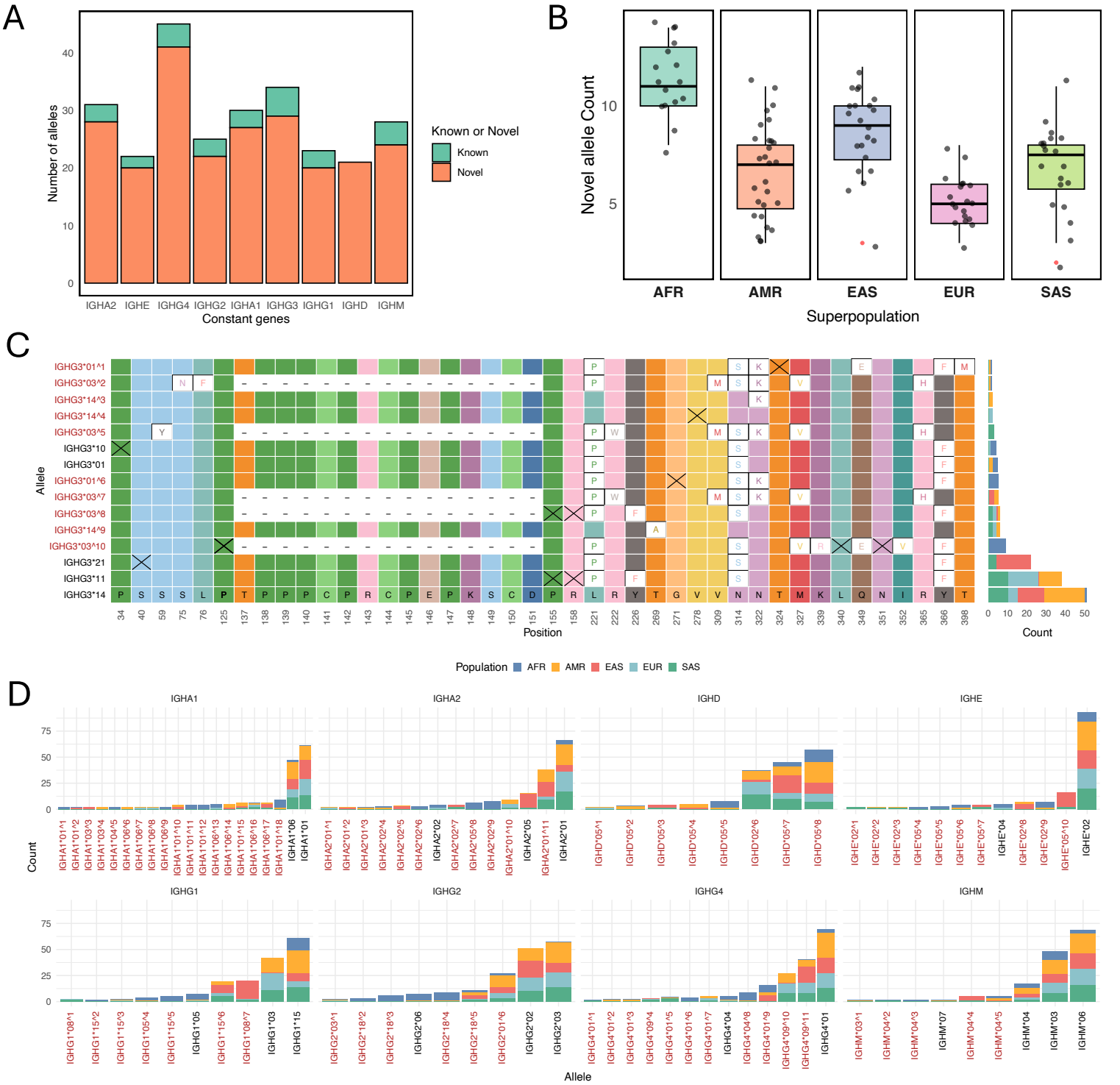
77. Chin,C.-S., Behera,S., Khalak,A., Sedlazeck,F.J., Sudmant,P.H., Wagner,J. and Zook,J.M. (2023) Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nat Methods*, **20**, 1213–1221.

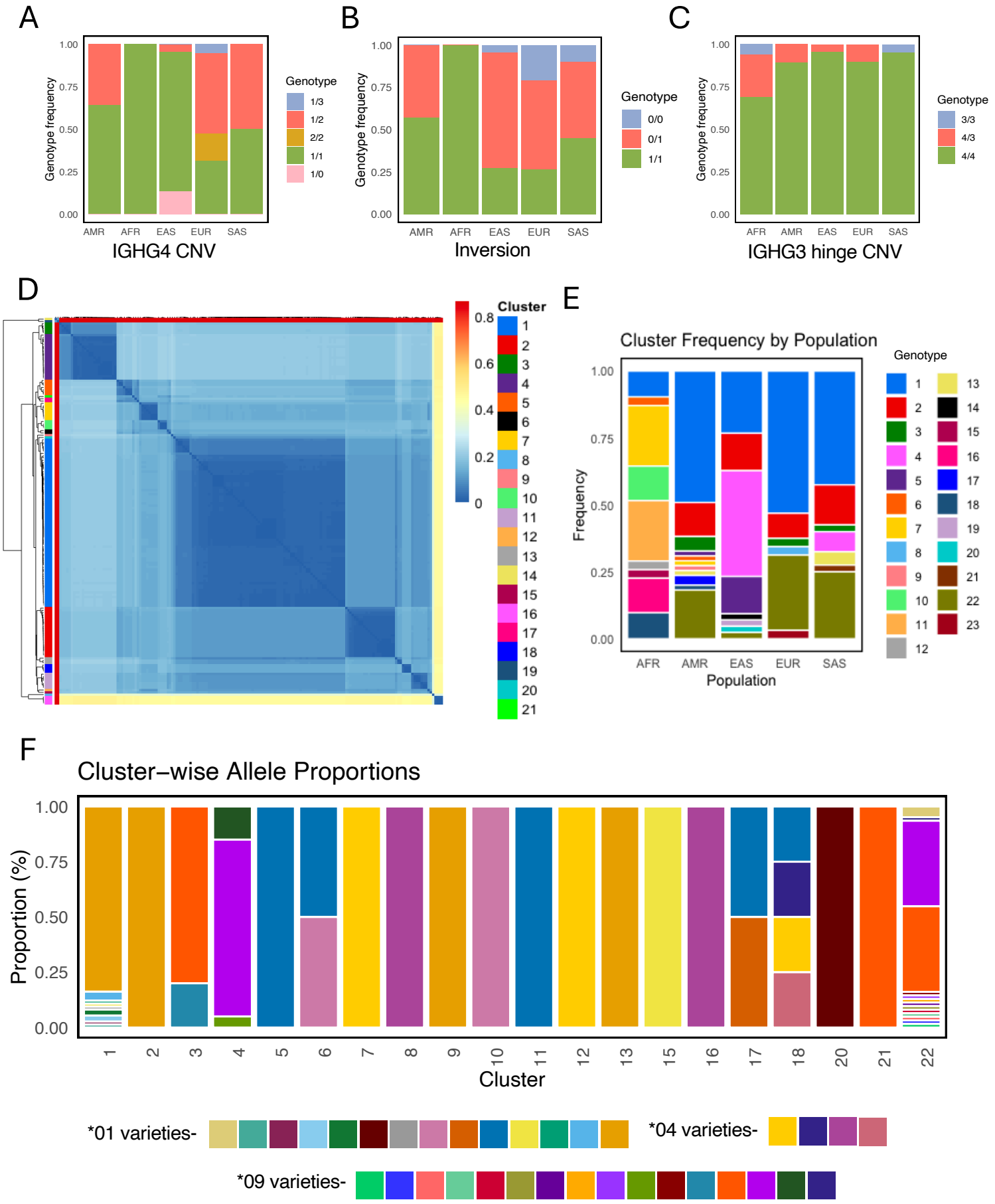

A
chr14 (q32.33)
p13
14p11.2
14q11.2
14q12
13.1
14q21.1
14q21.2
14q21.3
14q22.1
q22.3
14q23.1
q23.2
q23.3
q24.1
14q24.2
14q24.3
14q31.1
14q31.3
q32.12
14q32.2
q32.33
chr14:
99,850,000
99,900,000
99,950,000
100,000,000
100,050,000
100,100,000
100kb
ighc ref:
Genes:
IGHA2
IGHE
IGHG4
IGHG4
IGHG2
IGHA1
IGHG1
IGHG3
IGHD
IGHM
B
3'
5'
Transmembrane domain
Hinge exons


A
insertion-deletion
inversion
insertion-deletion
IGHA2
IGHE
IGHG4
IGHG2
IGHA1
IGHG1
IGHG3
IGHD
IGHM
NA12156
NA19129
NA18555
NA18507
HG00673
HG01928
HG01952
HG02717
Hap1
Hap2
Unique alleles
B
Count of alleles
Allele type
known
novel
C
SNV counts
SNV type
shared
unique
D
SNV counts
E
7.3%
92.7%
F
22.8%
77.2%
this study only
present in database
NA12156 (EUR)
HG01928 (AMR)
HG01952 (AMR)
NA18507 (AFR)
NA18555 (EAS)
HG00673 (EAS)
NA19129 (AFR)
HG02717 (AFR)
G
Fraction of SNV positions
SNV type
LongRead_unique
LongRead & 1KGP
unique_1KGP


A
% of the Locus
HiFi Coverage (log)
HG00673_H1
HG00673_H2
HG01928_H1
HG01928_H2
HG01952_H1
HG01952_H2
HG02717_H1
HG02717_H2
NA12156_H1
NA12156_H2
NA18507_H1
NA18507_H2
NA18555_H1
NA18555_H2
NA19129_H1
NA19129_H2
B
Sample
Number of contigs
Mean length(bp)
Phase switches
Method
IGenotyper
Hifiasm
C
IGHG4 SV
Inversion SV
IGHG3 hinge-SV
Samples
IGenotyper
Hifiasm
D
Number of missing SNVs
Sample
IGenotyper
Hifiasm
E
% of Concordance
Samples
IGenotyper
Hifiasm
F
Number of individuals (n = 8)
IGHC Genes
IGHA2
IGHE
IGHG4
IGHG2
IGHA1
IGHG1
IGHG3
IGHD
IGHM


A
IGHG4
IGHE
IGHG4
IGHG2
IGHA1
Insertion SV1
Insertion SV2
IGHA2
IGHE
IGHG4
IGHG2
IGHA1
IGHG1
IGHG3
IGHD
IGHM
Deletion SV1
Deletion SV2
B
LongRead_unique
LongRead & 1KGP
unique_1KGP
Fraction of SNV positions
1.00
0.75
0.50
0.25
0.00
AFR (n=16)
AMR (n=28)
EAS (n=22)
EUR (n=19)
SAS (n=20)
C
Novel
790 (36.04%)
In dbSNP
1402 (63.96%)
D
SNV denisity per 1KB
4
3
2
1
0
Overall
exon
intron
intergenic
Super-pop
AFR
AMR
EAS
EUR
SAS


A
Number of alleles
Constant genes
IGHA2 IGHE IGHG4 IGHG2 IGHA1 IGHG3 IGHG1 IGHD IGHM
Known or Novel
Known
Novel
B
Novel allele Count
AFR AMR EAS EUR SAS
Superpopulation
C
Allele
Position
Count
Population AFR AMR EAS EUR SAS
D
IGHA1
IGHA2
IGHD
IGHE
IGHG1
IGHG2
IGHG4
IGHM
Count
Allele
E
AFR AMR EAS EUR SAS
Proportion of individuals
type
Heterozygous
Homozygous
Gene


A
Genotype frequency
1.00
0.75
0.50
0.25
0.00
AMR AFR EAS EUR SAS
IGHG4 CNV
Genotype
1/3
1/2
2/2
1/1
1/0
B
Genotype frequency
AMR AFR EAS EUR SAS
Inversion
Genotype
0/0
0/1
1/1
C
Genotype frequency
AFR AMR EAS EUR SAS
IGHG3 hinge CNV
Genotype
3/3
4/3
4/4
D
Cluster
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
E
Cluster Frequency by Population
Frequency
1.00
0.75
0.50
0.25
0.00
AFR AMR EAS EUR SAS
Population
Genotype
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
F
Cluster−wise Allele Proportions
Proportion (%)
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 15 16 17 18 20 21 22
Cluster
*01 varieties-
*04 varieties-
*09 varieties-

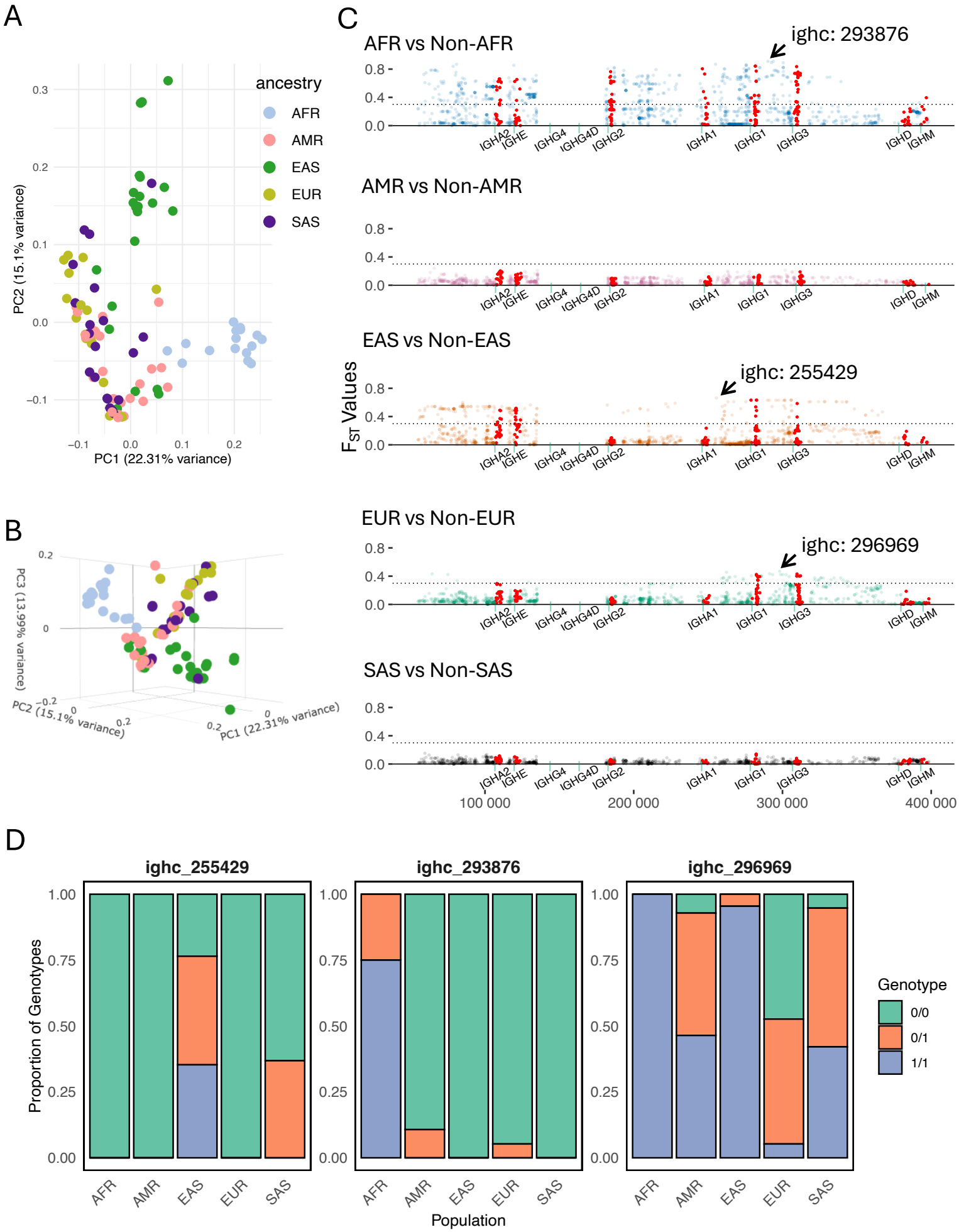

A
ancestry
AFR
AMR
EAS
EUR
SAS
PC2 (15.1% variance)
PC1 (22.31% variance)
B
PC3 (13.99% variance)
PC2 (15.1% variance)
PC1 (22.31% variance)
C
AFR vs Non-AFR
ighc: 293876
AMR vs Non-AMR
EAS vs Non-EAS
ighc: 255429
EUR vs Non-EUR
ighc: 296969
SAS vs Non-SAS
FST Values
IGHA2
IGHE
IGHG4
IGHG4D
IGHG2
IGHA1
IGHG1
IGHG3
IGHD
IGHM
100 000
200 000
300 000
400 000
D
ighc_255429
ighc_293876
ighc_296969
Proportion of Genotypes
AFR
AMR
EAS
EUR
SAS
Population
Genotype
0/0
0/1
1/1

Figure 1. Schematic map of the immunoglobulin heavy chain constant (IGHC) region.
(A) Genomic organization of the IGHC locus in the T2T-CHM13v2.0 reference (chr14:99730306-100132055). **(B)** The expanded illustration depicts the exon-intro gene structure of IGHG3, representing the structure common to IGHC genes. This includes three constant heavy (CH) domain exons CH1–CH2–CH3 (an additional CH4 domain is present in IGHM and IGHE), as well as hinge exons located between CH1 and CH2. The number of hinge exons varies within and among isotypes; no hinge exons are present in IGHA, IGHE, and IGHM, whereas 1-4 hinge exons are observed in IGHG and IGHD. Here the structure depicts 4 hinge exons as observed in IGHG3.

Figure 2. Resolving the IGHC locus in fosmid/trio-supported individuals from diverse ancestries.
(A) Haplotype-resolved maps of the IGHC locus across individuals from diverse ancestries (n=8), indicating the locations of polymorphic genes (allelic variation denoted by color) and SVs. **(B)** Frequency distribution of novel and known alleles (i.e., those present in IMGT) observed in the study cohort. **(C)** Counts of catalogued SNVs in each individual, partitioned by whether the SNVs are shared or private to each individual. **(D)** Upset plot showing the overlap of SNVs among individuals. **(E,F)** Comparison of observed SNVs to those curated in dbSNP v155 **(E)** and the 1KGP Phase 3 variant call set **(F)**. **(G)** Individual level comparison of genotyped SNVs using long-read sequencing (this study) to those detected using short-read data (1KGP Phase 3).

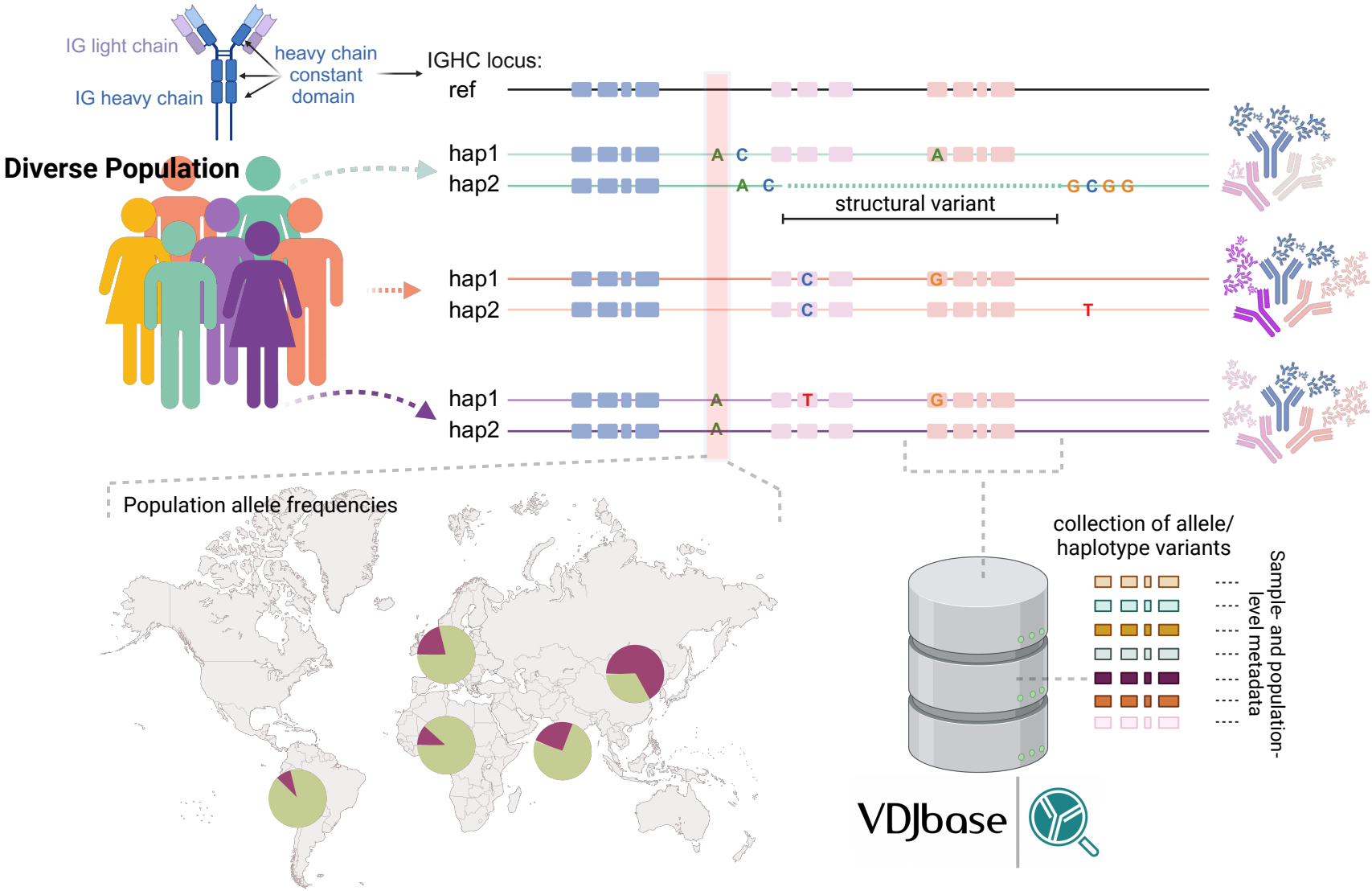
Figure 3. Benchmarking the performance of high-throughput IGHC targeted long-read sequencing against highly vetted ground-truth assemblies from individuals (n=8) of diverse ancestry.
(A) Per sample haplotype-specific HiFi read coverage, plotted as percent of haplotype-specific IGHC locus bases covered (Y axis) by haplotype-specific read depth (x axis). **(B)** Metrics (contig count, mean contig length, and number of phase-switch errors) for assemblies generated using IGenotyper and Hifiasm pipelines. **(C)** Reproducibility of ground-truth common SV annotations using automated IGenotyper and Hifiasm pipelines, indicating whether manual curation (marked in grey) was required. **(D)** Bar plots showing the number of variants in ground-truth assemblies missed by assemblies generated using IGenotyper and Hifiasm assemblies. **(E)** The base concordance (%) between IGenotyper or Hifiasm assemblies relative to covered ground-truth bases. **(F)** Accuracy of alleles curated from IGenotyper and Hifiasm assemblies compared to those curated from ground-truth assemblies.

Figure 4. Characterization of SVs and SNVs from population-level dataset.
(A) A map of additional characterized SVs in expanded population-level cohort: a tandem triplication of gene IGHG4 (Insertion SV1); ~120 Kb insertion including paralogs of IGHE, IGHG4, IGHG2, IGHA1 (Insertion SV2); large deletion resulting in a putative fusion allele of IGHA1 and IGHA2 genes, resulting in deletion IGHE, IGHG4, and IGHG2 genes (Deletion SV1); deletion of IGHG4 genes (Deletion SV2). **(B)** Per sample SNVs identified in our long-read datasets relative to those observed in the 1KGP Phase 3 SNV callsets. **(C)** Proportion of common SNVs (MAF >= 5%) in the expanded cohort also found in the dbSNP common SNV set. **(D)** Distribution of IGHC SNV density across samples in the expanded cohort, partitioned by genomic feature and super population.

Figure 5. Characterization of IGHC allele diversity from population-level dataset.
(A) Count of known and novel alleles identified for 9 IGHC genes; genes identified for close paralogs were consolidated into a single non-redundant set (e.g., alleles for IGHG4 and close duplicates). **(B)** Density plot showing distribution of donors in each super population carrying varying numbers of novel alleles. **(C)** Multiple sequence alignment of amino acid sequences representing IGHG3 alleles identified in our population; only variant positions are shown for alleles found in >1 individual, with "X" denoting synonymous variants. The number of individuals carrying each allele partitioned by population is shown in the stacked barplot at right. **(D)** Counts of individuals by super population carrying identified alleles for IGHC genes, including only alleles observed in >1 sample. **(E)** Stacked bar plot showing the percent heterozygosity/homozygosity for each IGHC gene, partitioned by super population.

Figure 6. Comprehensive characterization of SVs across diverse genetic ancestries.
(A-C) Stacked bar plot representing genotype frequencies by super population for three common SVs. **(D)** Heatmap representing pairwise nucleotide distances between IGHG4 region haplotypes; haplotypes are grouped as indicated by the dendrogram on left, representing hierarchical clustering. All sequences in heatmap are labeled by their assigned haplotype cluster, as determined by distance based clustering, using a distance cutoff of 0.1. **(E)** Stacked bar plot showing the frequency IGHG4 haplotype clusters among super populations **(F)** Stacked bar plot showing the distribution of IGHG4 alleles across defined haplotype clusters.

Figure 7. Coding and non-coding IGHC SNVs reveal inter-population genetic differences.
(A-B) Principal component analysis of IGHC locus-wide SNVs; points represent individuals in our cohort, colored by super population. **(C)** F_{st} values for IGHC locus-wide SNVs from pairwise tests of each individual population against all other populations. **(D)** Frequency of genotypes for three example SNVs with the highest observed F_{st} values for three populations.


IG light chain
IG heavy chain
heavy chain constant domain
IGHC locus:
ref
Diverse Population
hap1
hap2
A C
A
A C
G C G G
structural variant
hap1
hap2
C
G
C
T
hap1
hap2
A
T
G
A
Population allele frequencies
collection of allele/ haplotype variants
Sample- and population-level metadata
VDJbase