HLA-Resolve: High-Resolution HLA Haplotyping Using Long-Read Hybrid Capture

Matthew R. Glasenapp[1], Muh-Ching Yee[1], Alexander E. Symons[1], Obed A. Garcia[1,2], Omar E. Cornejo[1]

[1]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz
[2]Department of Anthropology, University of Kansas

Abstract

Accurate HLA typing is critical for transplantation, pharmacogenomics, and disease risk prediction, yet short-read approaches cannot resolve the HLA region's extreme polymorphism. Long-read sequencing improves resolution, but its adoption has been limited by higher cost, reduced base accuracy, limited throughput, and reliance on long-range PCR. To overcome these limitations, we present a multiplexed long-read hybrid capture workflow for PacBio and Oxford Nanopore sequencing that enriches all classical HLA loci and the complete HLA Class III region. A single-step enzymatic fragmentation and barcoding strategy enables automated library prep. We also introduce HLA-Resolve, an HLA typing program optimized for HiFi reads, and validate workflow performance against the Genome in a Bottle, Human Pangenome Reference Consortium, and International Histocompatibility Working Group benchmarks using 32 geographically diverse samples. These advances offer a cost-effective approach for high-resolution HLA typing with clinical applicability and enable investigation of the role of HLA Class III variation in disease.

Main

The human leukocyte antigen (HLA) locus (6p21.3) encodes the molecules responsible for immune self-recognition and antigen presentation, thereby shaping host-pathogen interactions, susceptibility to infectious and autoimmune diseases, and donor compatibility in transplantation[1]. It is among the most polymorphic regions of the genome[2–4] and has the most disease associations of any locus[5]. HLA typing, the process of identifying the specific HLA haplotypes present at the antigen-presenting HLA genes, is routine in immunogenetics research[6,7] and donor-recipient matching in organ and stem cell transplantation[8–11]. Clinical HLA typing methods range from traditional antigen-level serological assays to more recently developed higher-resolution next-generation sequencing (NGS) tests, yet most HLA NGS approaches still struggle to resolve the extraordinary genetic variation present within these genes.

HLA typing methods relying on short-read gDNA sequencing often produce incorrect or ambiguous allele calls because short reads struggle to resolve the complexity of the HLA locus[12]. The extreme polymorphism of the HLA locus exacerbates reference bias, as short reads originating from HLA alleles divergent from the human reference genome may either fail to map or align with reduced quality, biasing downstream genotyping[13]. Reference genome alignment is further compromised by inter-gene sequence homology (e.g., HLA-B and HLA-C), large structural variants, and numerous paralogous loci with high sequence similarity (e.g., HLA-DRB)[12]. These challenges have prompted the development of increasingly complex HLA typing algorithms[14], including those that align reads directly to known HLA haplotype sequences[15–19], employ graph-based alignment strategies[20–23], or use local de novo assembly[24–27]. However, these algorithms often still yield ambiguous results, necessitating expert review and complex donor-selection algorithms for transplant patients[28,29]. While long-read HLA typing tools have begun to address these limitations[30–32], there remains a need for scalable, targeted long-read sequencing assays and complementary bioinformatics strategies to further improve the reliability of HLA typing and haplotype reconstruction[33].

The HLA Class III locus shares many of the complexities of the antigen-presenting HLA loci (Class I, Class II) and would also benefit from long-read approaches. HLA Class III is among the most gene-dense regions of the human genome[7], and encodes complement factors (e.g., C4A, C4B, CFB), cytokines (e.g., TNF, LTA, LTB), and numerous other genes involved in fundamental cellular processes[6]. Many well-known HLA disease associations originate from HLA Class III genes rather than the classical antigen-presenting HLA loci[5], including associations with C4 (systemic lupus erythematosus), TNXB (Ehlers-Danlos syndrome), and CYP21A2 (congenital adrenal hyperplasia)[34]. Short-read approaches struggle in this region as reads do not reliably map to HLA Class III genes with highly similar pseudogene paralogs (e.g., CYP21A2 vs. CYP21A1P) and cannot adequately characterize the complex structural variation of the RCCX module. This eight-gene cassette, which includes C4 and TNXB, exhibits extensive copy-number variation. These challenges underscore the importance of developing HLA assays

that encompass not only the classical Class I and Class II antigen-presenting loci but also the clinically relevant Class III region.

Long-read sequencing offers a powerful solution to many of the challenges inherent to HLA characterization, and is increasingly recognized as the future of clinically relevant variant detection[35–38]. Long reads enhance genome assembly, reference mapping, phasing accuracy, and structural variant detection, enabling more reliable resolution of full-gene HLA haplotypes[39–44]. HLA alleles are defined by four colon-separated fields that specify increasing levels of resolution, from the amino acid sequence (fields 1–2) to synonymous coding variation (field 3) and noncoding variation (field 4), with more than 43,000 alleles catalogued in the IPD-IMGT/HLA database[45,46]. Conventional methods restricted to the antigen-recognition domain or coding exons typically resolve alleles only at two- or three-field resolution. By sequencing through introns and untranslated regions, newer long-read approaches permit four-field, ultrahigh-resolution (UHR) HLA typing, revealing clinically meaningful noncoding variation while also improving accuracy at lower fields of resolution, as reads spanning intronic sequence bridge distant exons and enable read-based phasing across the gene. In a retrospective study of hematopoietic stem cell transplants, UHR HLA typing with PacBio sequencing reclassified the matching status of 29% of donor-recipient pairs[47], and five-year overall survival was higher for fully UHR-matched patients than for pairs with previously undetected HLA mismatches or pairs matched at coding sequences but mismatched in noncoding regions[47] (but see refs. [48,49]). Few studies have examined the influence of noncoding mismatches on transplant outcomes, but noncoding variation is known to affect HLA gene regulation and splicing[50–52], underscoring the clinical importance of characterizing the entire HLA gene sequence[53].

Despite the advantages of long-read sequencing for HLA typing, its adoption has been limited by higher costs, lower throughput, higher base error rates, and procedural challenges, such as mechanical shearing. However, recent advances in PacBio's HiFi sequencing chemistry and variant calling have made PacBio's small variant detection accuracy comparable to Illumina[36], and emerging target enrichment approaches have helped reduce cost and increase throughput[54–56]. Still, most existing targeted long-read HLA typing protocols rely on long-range PCR, which can compromise accuracy through amplification bias, allele dropout, chimerism from template switching, and replication slippage in low-complexity regions[14,30,53–55,57–59]. HLA typing would therefore benefit from a scalable, targeted long-read sequencing approach that avoids the use of gene-specific PCR primers. While hybrid capture has been successfully applied to long-read sequencing of other immunogenetic loci[60], existing HLA hybrid capture protocols remain mostly limited to short-read platforms[61–64] or to a subset of the HLA genes bundled with dozens of other non-HLA genes (e.g., Twist Alliance Long-Read PGx Panel).

Here, we present a cost-effective, high-throughput target-capture approach for multiplexed long-read HLA sequencing on the PacBio and Oxford Nanopore platforms. The workflow features one-step enzymatic fragmentation and barcoding, followed by hybrid capture with a custom biotinylated probe set (Twist Bioscience) and PacBio/ONT library. This workflow is readily automatable, avoids the need for expensive mechanical shearing, and is adaptable to

other challenging genomic regions. We also introduce a lightweight HLA typing algorithm tailored to high-coverage HiFi reads that returns star allele calls at four-field resolution. Our end-to-end workflow enables accurate reconstruction of full-length haplotypes for the classical antigen-presenting HLA genes and all HLA Class III protein-coding genes. We rigorously validated our approach with 32 geographically diverse samples from the Human Pangenome Reference Consortium (HPRC) and International Histocompatibility Working Group (IHWG) catalogs. These datasets allowed us to validate (i) genotyping accuracy against the Genome in a Bottle benchmark, (ii) haplotype sequence identity against the phased HPRC assemblies, and (iii) HLA star allele concordance against the IHWG benchmark. Our results demonstrate robust HLA typing performance across diverse global populations, and the resulting dataset provides a valuable resource for future method development and benchmarking.

Results

Hybrid capture sequencing consistently captures all 81 targeted protein-coding HLA genes

We developed a custom HLA hybrid capture assay for multiplexed long-read sequencing. We applied it to 33 geographically diverse human samples from the Human Pangenome Reference Consortium (HPRC) and the International Histocompatibility Working Group (IHWG) (Fig. 1, Supplementary Table 1). The samples were tagmented individually, yielding an average peak fragment size of 8,234 bases (Supplementary Fig. 1a). The tagmented samples were then pooled by cohort (HPRC, IHWG) for hybrid capture using a custom probe panel developed with Twist Bioscience, yielding an average peak fragment size of 5,637 bases (Supplementary Fig. 1b). The custom probe panel was created by combining Twist's existing probes for twelve HLA Class I (HLA-A, B, C) and II (HLA-DPA1, DPB1, DQA1, DQA2, DQB1, DRB1, DRB3, DRB4, DRB5) genes with a novel probe set targeting the 69 HLA Class III protein-coding genes annotated in the human reference genome (GRCh38). Probes were tiled at 1kb intervals across each gene to ensure uniform enrichment, including through intronic and untranslated regions. The two hybrid capture pools were then pooled for library preparation and sequenced on both the PacBio Revio (SMRTbell prep kit 3.0, 25M SMRT Cell) and the ONT PromethION (R10.4.1). After demultiplexing, the PromethION run yielded 68,573,197 reads (163.5 Gb) and the Revio run yielded 11,337,707 reads (50.8 Gb). The PromethION generally offers higher raw throughput than the Revio but requires more input DNA. Therefore, the PromethION flow cell was loaded with approximately twice the volume of DNA used for the Revio SMRT Cell. When stratified by hybrid capture cohort, raw read counts per sample were highly correlated between PacBio and ONT (HPRC: Pearson r = 0.89; IHWG: Pearson r = 0.96), indicating consistent sample-level performance across platforms. One sample from the HPRC cohort, HG01891, did not tagment and failed to produce sufficient sequencing reads (PacBio:

476, ONT: 1106) and was therefore excluded from downstream analyses. Across all samples and platforms, greater than 99.7% of the raw reads mapped to the human reference genome. The number of unique (de-duplicated) mapped reads ranged from 21,397 - 448,211 for PacBio, and 183,480 - 1,475,328 for ONT (Fig. 2a). Mean and median read lengths for target-overlapped reads were longer for PacBio (mean=4,093, median=3,910, N50=4,331) than ONT (mean=2,805, median=2,693, N50=3,306) (Fig. 2b, Supplementary Fig. 2). We observed high on-target mean coverage depth across samples for both PacBio (148X) and ONT (286X) (Fig. 2c). Target specificity was highly consistent across hybrid capture pools and across samples within a hybrid capture pool (Supplementary Fig. 3, Supplementary Tables 2 and 3). Variation in coverage depth among samples within cohorts was primarily driven by sequencing effort (i.e., reads per sample) rather than by variation in target specificity (Fig. 2a, Supplementary Fig. 3). Variation in coverage depth between hybrid capture cohorts is detailed in Supplementary Note 1. Fold enrichment, defined as the ratio of mean on-target to off-target depth, was high (505–921X) and consistent across sequencing platforms and among samples within each cohort, though it varied slightly between hybrid-capture cohorts (Fig. 2d).

Sequencing coverage profiles from the PacBio Revio and ONT PromethION strongly mirrored each other (Fig. 2e,f, Fig. 3). There was a near-perfect correlation between the mean coverage depth of captured protein-coding genes between PacBio Revio and ONT PromethION, when stratified by hybrid capture cohort (Fig. 2e,f). We consistently captured all 81 targeted protein-coding genes, including the 69 HLA Class III genes (Fig. 3, Supplementary Figs. 4 and 5). We also captured an additional four untargeted protein-coding genes, HLA-E, HLA-F, HLA-G, and HLA-DQB2, and numerous additional pseudogenes, lncRNAs, snoRNAs, and miRNAs annotated in the human reference genome, including HLA-H, HLA-J, and HLA-K. See Supplementary Tables 4–11 for per-sample mean coverage depth and proportion of bases covered by 20 reads, reported by gene and exon for both PacBio and ONT.

Although the read lengths generated from the hybrid capture protocol are shorter than those of standard PacBio and ONT whole-genome sequencing, the reads were sufficiently long to distinguish between highly paralogous genes, such as protein-coding gene CYP21A2 and its pseudogene paralog CYP21A1P (Fig. 4a). Our reads were also able to distinguish C4A from C4B, and differentiate between long and short isoforms with or without the long intron 9 HERV insertion (Fig. 4a). Across the HLA region, we were able to detect deletions up to 6,367 bp and insertions up to 3,655 bp (Fig. 4b, Supplementary Table 12) using pbsv with the PacBio reads (https://github.com/PacificBiosciences/pbsv). To confirm read-level support for detected structural variants, we manually reviewed all variants >500 bp in IGV and removed those lacking supporting reads. The genes harboring the most structural variants (insertions/deletions >= 50 bp) were the HLA Class II genes HLA-DRB1, HLA-DRB5, HLA-DRB6, HLA-DPB2, HLA-DQA1, and HLA-DQB1, and the Class III genes TSBP1, C4A, C4B, and C2, consistent with the known structural complexity of the Class II loci and the RCCX module of Class III (Fig. 4c). Structural variants were distributed relatively evenly across samples, ranging from 6 to 49

per sample (median = 33), with no individual samples disproportionately contributing to the total call set (Fig. 4d).

Clinical-level genotyping accuracy achieved with HiFi reads

To evaluate small-variant genotyping accuracy, we called variants with DeepVariant and Clair3 and compared the results to the Genome in a Bottle (GIAB) high-confidence benchmarks (v4.2.1). For the ONT reads, both DeepVariant and Clair3 were evaluated, and Clair3 performed better. We therefore present ONT results from Clair3 only in downstream analyses. Genotype concordance was assessed for samples HG002–HG007 using hap.py following the GA4GH/GIAB best practices[65]. We restricted analyses to GIAB's confident regions and loci targeted by our hybrid capture probes, resulting in 2,113 (HG003) - 3,146 (HG007) truth SNVs and 177 (HG002) – 250 (HG007) truth indels. We observed strong genotype concordance from both platforms, with PacBio + DeepVariant achieving clinical accuracy (mean SNV F1 = 99.8%, mean INDEL F1 = 98.4%) and outperforming ONT + Clair3 (SNV F1 = 98.4%, INDEL F1 = 94.1%; Fig. 5a, Supplementary Table 13). Discordant calls for both platforms were concentrated in C4A, C4B, and the HLA-DQ genes, with HLA-DQ accounting for the majority of false positives and negative genotypes. It is important to note that the distribution of SNVs in the GIAB benchmark is not uniform across our targeted regions, as SNPs are not evenly distributed among HLA genes. For example, more than 75% of the HG002 truth set variants fall within HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQB1, and TSBP1 (Supplementary Fig. 6). Furthermore, HLA-DRB1 is not included in GIAB's confident regions, so we were not able to evaluate genotype concordance within this gene.

To determine the minimum sequencing depth required for accurate genotyping, we iteratively downsampled the HG002 PacBio reads and recalculated coverage depth and genotype concordance with the GIAB benchmark. Because coverage depth varied across HLA classes, we downsampled and evaluated concordance separately for each HLA class. To account for sampling variance, ten independent downsampling replicates were generated at each coverage level, and mean recall, precision, and F1 score were estimated (Supplementary Figs. 7 and 8, Supplementary Table 14). As coverage depth decreased, recall declined sharply. In contrast, precision remained consistently high, reflecting accurate genotyping at the subset of variant sites that were still detected and a low rate of false positive calls. The coverage depth required to sustain recall thresholds of 98% and 95% differed by HLA class, with HLA Class III maintaining higher recall at lower depths than HLA Class I and II. For HLA Classes I and II, SNP recall varied substantially across downsampling replicates, exceeding 98% in all ten replicates only at 60X depth for Class I and at 153X for Class II, though Class II recall remained variable with individual replicates occasionally falling below 98% at higher coverage levels (Supplementary Fig. 8). In contrast, HLA Class III maintained SNP recall above 98% for all replicates at approximately 12x.

High-Accuracy haplotyping enables high-resolution HLA typing with HLA-Resolve

We developed HLA-Resolve, a Python command-line tool that performs four-field HLA star-allele calling directly from raw PacBio reads (FASTQ/uBAM). HLA-Resolve follows a standard variant-discovery workflow, including adapter trimming, PCR duplicate removal, reference-genome alignment, variant calling, structural variant calling, tandem repeat genotyping, and genotype phasing. The phased genotypes from HiPhase are applied to the human reference genome with vcf2fasta[66] to reconstruct coding and full-gene haplotype sequences, which are then queried against the IPD-IMGT/HLA database to identify the best-matching allele by edit distance. In addition to the reconstructed FASTA haplotypes and star allele assignments, HLA-Resolve outputs haplotagged BAMS, phased VCFs, and coverage depth metrics from mosdepth[67].

We also phased genotypes for the Class III genes, in addition to the classical HLA genes, resolving complete haplotypes (including introns and UTRs) for an average of 75 of 83 captured protein-coding genes per sample from the PacBio reads (excluding HLA-DRB3 and -DRB4 due to their absence from GRCh38 primary assembly). HLA class II genes were generally more challenging to phase, particularly HLA-DRB1, which contains a long and structurally variable first intron that frequently includes large insertions relative to the GRCh38 reference. The GRCh38 reference genome features one of the shortest HLA-DRB1 alleles (DRB1*15:01:01:01, see https://www.ebi.ac.uk/ipd/imgt/hla/help/genomics.html). Across alleles cataloged in the IPD-IMGT/HLA database, the length of HLA-DRB1 intron 1 varies by several kilobases. The shortest HLA-DRB1 intron 1 is 5,262 bases (DRB1*15:01:01:07), while the longest reported is 10,315 bases (DRB1*09:01:02:05) (IPD-IMGT/HLA v3.62.0). This substantial intronic length variability complicates read alignment, variant calling, and phasing continuity, particularly when large insertions relative to the reference genome are not completely spanned by many sequencing reads. A phasing break was frequently observed at the location of the large HLA-DRB1 intron 1 insertion relative to GRCh38, likely because the hybrid capture probes failed to capture DNA from this insertion.

We ran HLA-Resolve on all HPRC and IHWG PacBio samples, evaluating haplotype accuracy against the HPRC haploid assemblies and star allele concordance against the IHWG reference typings. We excluded one IHWG sample (IHW09117) due to contamination signatures (see Supplementary Note 2). Comparison of our PacBio full-gene sequences to the HPRC assemblies showed high haplotype resolution accuracy (Fig. 5b). Across the HLA Class I genes (HLA-A, -B, -C), 45/47 (96%) of our queries had zero edit distance to the corresponding assembly, compared with 51% for HLA Class II, driven largely by HLA-DQA1 (2/13) and HLA-DRB1 (1/13) (Supplementary Table 15). Of our reconstructed coding haplotypes, 94% (467/496) had an identical match in the IPD-IMGT/HLA database, with five of eight typed genes having all perfect matches (HLA-A, -B, -C, -DPA1, -DPB1). Across 15 IHWG samples, we observed 97% star allele concordance at two-field (215/222) and three-field resolution (193/198), with perfect

two- and three-field concordance for HLA-A, -B, -C, and -DPA1, and 81% concordance at four-field resolution (Fig. 5c). We observed similar concordance against the HPRC reference typings established by Lai et al.[63] with 97% concordance at two-field (187/192) and three-field (185/190) resolution, and 81% at four-field (150/186). HLA-A, -B, -C, -DPA1, and -DPB1 achieved perfect concordance through three-field resolution. However, these truth labels derived from the HPRC assemblies have not been independently validated (see Discussion). For the PacBio IHWG samples, HLA-Resolve outperformed StarPhase[31] and SpecImmune[32], at all fields of resolution, including identifying ten additional correct star alleles at four-field resolution (HLA-Resolve: 131/162; StarPhase, SpecImmune: 121/162) (Fig. 6, Supplementary Table 16). HLA-Resolve is currently PacBio-specific, so we benchmarked StarPhase and SpecImmune on ONT reads, where both maintained performance similar to their PacBio results, with mean concordance 1–4 percentage points lower at two- and three-field resolution and near-identical four-field concordance (~73% on both platforms). We also evaluated HLA*LA[22] and observed strong performance and compatibility with our reads, but did not quantify star-allele concordance, as HLA*LA returns calls at G-group resolution.

HLA-Resolve is a lightweight tool that operates directly on raw FASTQ files with minimal computational requirements and no additional bioinformatics steps or manual intervention. In developing HLA-Resolve, we are ensuring our hybrid capture sequencing strategy has a validated, user-friendly informatics platform integration. It implements validated, open-source tools for haplotype reconstruction and a transparent, biologically intuitive algorithm for querying the IPD-IMGT/HLA allele database. HLA-Resolve is compatible with whole-genome or whole-exome sequencing data, but is more efficient with targeted data, as the rate-limiting step is reference genome alignment. We benchmarked runtime and memory usage for our hybrid capture data on a dual-socket AMD EPYC 7663 node running AlmaLinux 9.7. Each sample was processed using 12 CPU threads and allocated 25GB of RAM. Across PacBio samples, HLA-Resolve had an average wall time of 14 minutes and a peak memory usage of 17.7 GB. For ONT samples, HLA-Resolve was slightly slower due to the higher number of raw reads, with an average wall time of 42 minutes and a peak memory usage of 18.4 GB.

Discussion

We present an end-to-end long-read HLA typing solution that overcomes longstanding limitations in accuracy, cost, and workflow complexity. Our one-step tagmentation and barcoding method consistently generates high-complexity DNA fragments suitable for hybrid capture and multiplexed, platform-agnostic long-read sequencing. We also demonstrate consistent target enrichment across two independent hybrid capture experiments. Our protocol is fully compatible with automated liquid handlers and obviates the need for traditional mechanical shearing approaches requiring manual intervention (e.g., Diagenode Megaruptor or Covaris g-TUBE), thereby reducing costs, streamlining the workflow, improving reproducibility, and enabling comprehensive sample tracking to minimize human error. Although our captured

fragments were shorter than typical PacBio and ONT WGS libraries, they were sufficient for high-accuracy HLA genotyping, haplotyping, and star allele calling. We benchmarked three open-source long-read typers, SpecImmune[32], StarPhase[31], and HLA*LA[22] on both PacBio and ONT reads, all of which showed strong concordance on both platforms. For PacBio reads, our tool HLA-Resolve outperformed all three and provides a fully automated bioinformatic workflow compatible with WGS, WES, amplicon, and hybrid-capture sequencing. HLA-Resolve successfully reconstructed phased full-gene haplotypes with a standard variant-discovery workflow and accurately matched them to the IPD-IMGT/HLA database using a biologically interpretable edit distance metric. The high-coverage HiFi reads produced by our sequencing protocol support a conventional linear reference-based implementation that is lightweight, fast, and intuitive, avoiding the need for complex read binning or alignment heuristics. By reconstructing haplotypes before database matching, the inferred gene sequences are not influenced by the composition of the IPD-IMGT/HLA reference set, although database completeness influences the matching process and final star allele assignments. While we developed and tested the protocol on the HLA locus, the underlying targeted sequencing protocol is generalizable to any genomic region by modifying the hybrid capture probe panel.

Cost-effective, high-resolution HLA haplotyping remains a major unmet need in both research and clinical practice[53]. High-resolution typing is rarely implemented despite emerging evidence that increased typing resolution may improve transplant outcomes[47,49]. Our workflow offers a practical path toward high-resolution clinical typing at approximately $100/sample, with room for further reduction through increased multiplexing and continued improvements in PacBio sequencing chemistry (e.g., SPRQ-Nx). The recent introduction of PacBio's Vega benchtop sequencer further lowers the barrier to adoption for clinical and smaller research laboratories. Beyond clinical applications, long-read HLA haplotyping addresses a fundamental limitation in immunogenetics research. The classical HLA genes harbor numerous disease associations, including type 1 diabetes[68], multiple sclerosis[69], ankylosing spondylitis[70], celiac disease[71], among others[7], yet resolving haplotypes with short reads remains difficult, often requiring trio sequencing, statistical phasing, or imputation. Trio sequencing is costly and not always feasible, and imputation accuracy drops substantially in non-European populations due to limited representation in available reference panels[72]. By generating phased haplotypes directly from high-coverage long reads, our method provides a scalable approach to building diverse reference panels that could substantially improve imputation accuracy across underrepresented populations.

In addition to HLA typing, our method provides near-complete coverage of the HLA Class III region, which contains many clinically important genes that are poorly served by existing approaches. The most structurally complex of these is the RCCX module, a highly variable multi-gene cassette with structural variation that underlies several Mendelian and complex diseases. For example, CYP21A1P-CYP21A2 fusion deletion mutations cause congenital adrenal hyperplasia[34], low C4 copy-number is associated with systemic lupus erythematosus[7,73], and TNXB deficiency causes classical-like Ehlers-Danlos syndrome[74,75].

These variants are difficult to detect with short-read sequencing due to the high sequence identity between paralogous genes and pseudogenes. The long reads generated by our approach can be reliably assigned to CYP21A2 or the CYP21A1P pseudogene and are compatible with specialized tools such as Paraphase[76] for resolving pathogenic variants. Reference alignment at C4A and C4B remains challenging, but our long reads anchor to the sites that distinguish the two genes in GRCh38 and may be well suited for local de novo assembly. Together, the sequencing strategy and available analytical tools provide a cost-effective avenue for studying RCCX-associated disease in parallel with HLA typing.

While our protocol successfully captures all targeted genes, there are several areas where future iterations could improve. Across all samples, the mean coverage of HLA Class I and II genes was approximately threefold higher than that of Class III, reflecting an unequal mixture of the two biotinylated probe panels, which we have adjusted for future iterations of the assay. To expand the scope of our method, a second version of our probe panel is also under development, incorporating an expanded set of MHC and killer cell immunoglobulin-like receptor (KIR) gene targets (e.g., MICA, MICB, KIR2DL1). For HLA typing, our reference-based approach has limitations. One is reconciling overlapping variant calls from multiple callers (e.g., DeepVariant, pbsv), which may genotype the same variants but annotate them differently. HLA-Resolve handles this by screening for redundant genotypes before haplotype reconstruction (see Methods). A second limitation is that genes with high structural variation, such as HLA-DRB1, pose challenges for genotyping and phasing, as kilobase-scale insertions relative to GRCh38 are common in intron 1. These complex variants likely contributed to our reduced 4-field star allele concordance, as they may fail genotype quality filters or resist read-based phasing, introducing phasing breaks that prevent full-gene haplotype reconstruction, obscuring variation that distinguishes alleles at the fourth field. Improving coverage across complex variants, such as the HLA-DRB1 intron 1 insertion, is a priority. However, their high repeat content[77] may limit our ability to target them with hybrid capture. Local assembly is an alternative, but it remains unreliable for HLA genes even with reads far exceeding the lengths generated by our tagmentation-based library prep. In the HPRC v1 long-read assemblies, a notable proportion of samples had HLA genes with missing annotations, likely due to assembly gaps, a problem also observed in the high-coverage 1KGP ONT assemblies from Gustafson et al.[78].

The commercial HLA typing landscape is rapidly evolving toward targeted long-read sequencing. Most current assays use either Illumina-based hybrid capture (e.g., Thermo Fisher AlloSeq Tx 17, CareDx One Lambda HybridType NGS Assay, Omixon Holotype HLA) or ONT-based amplicon sequencing (e.g., Thermo Fisher's AllType Rapid Assay, Omixon's NanoType, and GenDx's NGS Pronto), but new long-read hybrid capture options are emerging, including the Qiagen QIAseq xHYB HLA panel for PacBio sequencing and GenDx's ONT-based NGS-Capto. Despite this progress, transparent benchmarking remains limited across the field. Several products have received regulatory approval, but detailed independent validation is sparse, and, to our knowledge, no study has applied multiple commercial solutions to the same IHWG samples for direct comparison. Vendors routinely use the IHWG benchmark during

development and participate in the International HLA and Immunogenetics Workshop (IHIWS), but rarely publish detailed performance metrics on haplotype accuracy or star allele concordance. Academic approaches face a similar problem. Custom sequencing protocols are commonly paired with commercial analysis software such as GenDx NGSengine or Omixon HLA Explore, but these hybrid strategies are rarely validated against gold-standard benchmarking datasets. Performance is more often reported as rates of ambiguous calls or the proportion of allele assignments at different resolution levels, with reproducibility emphasized over concordance against independent truth sets. This lack of standardized benchmarking across both commercial and academic settings limits our ability to identify the major unresolved challenges in high-resolution HLA typing and to direct efforts toward the most pressing problems. The proprietary nature and cost of commercial assays further limit their accessibility in academic research, motivating the development of open, cost-effective alternatives.

Benchmarking NGS-based HLA typing methodologies is inherently challenging because accuracy depends on the entire workflow, from raw sequencing read generation to the algorithms that produce star allele calls. The cause of incorrect star allele calls is not always immediately apparent, as errors may arise from data quality, the typing algorithm, or both. A meaningful comparison of HLA typing tools must therefore account for the sequencing platform, its error profile, and the library strategy, as WGS, WES, amplicon, and hybrid capture may yield fundamentally different coverage distributions across HLA genes. HLA typing algorithms also vary in their robustness to coverage depth and in their accuracy across loci, leading to situations in which the highest-performing tool depends on both the gene and the depth of coverage[79]. This variability is expected given the substantial differences among HLA genes in size, polymorphism, and repeat content. For these reasons, the IHWG star allele concordance values presented here should be interpreted as specific to the long-read target-capture data generated with our protocol. In this context, HLA-Resolve outperformed all freely available tools on our PacBio sequencing reads.

The benchmarking datasets used to develop and validate open-source HLA typing tools have important limitations. The most commonly used benchmark, including for long-read typers SpecImmune[32] and StarPhase[31], is the 1KGP HLA allele labels[80], which were generated from short-read whole-exome sequencing and the PolyPheMe HLA typing tool (Xegen, France) at two-field resolution. These labels have not been independently confirmed with newer sequencing and HLA typing technologies, and coverage depth varies considerably across the Phase 3 samples used to generate them[79]. Although ONT data are now freely available for many of the 1KGP samples, the 1KGP HLA truth labels have, to our knowledge, not been re-evaluated. More recently, Lai et al.[63] generated truth labels for the HPRC v1 cohort by comparing the assembled HLA haplotypes with alleles in the IPD-IMGT/HLA database. However, this approach has not been independently validated and assumes assembly accuracy, a nontrivial assumption given that a single miscalled nonsynonymous variant can change the allele group. For example, Chin et al.[81] developed a diploid MHC assembly-based benchmark for HG002 using sequencing reads

from multiple platforms, but likely misclassified HG002 as homozygous for DQB1*05:01:01G (label generated by HLA*ASM). Our analysis indicates heterozygosity for DQB1*05:01 and DQB1*03:02, consistent with Lai et al.[63] and Behera et al.[82]. Any individual approach to establishing ground truth will have limitations, underscoring the need for consortium-oriented benchmarking datasets representing consensus across independent groups using different wet-lab and computational methods. Comprehensive, high-coverage long-read sequencing of the IHWG catalog would provide a strong foundation for this effort.

There is a rapidly growing interest in Oxford Nanopore Technologies (ONT)-based workflows owing to the substantial advantages of long-read sequencing and the competitive pricing of the ONT MinION and GridION instruments. These workflows represent a considerable improvement over older PCR-SSP and PCR-SSO methods, perform well at lower levels of resolution, and are beginning to enter clinical use. In our data, ONT small-variant genotype concordance in the HLA region was notably lower than PacBio, despite sequencing on the PromethION platform with R10 chemistry at higher coverage depth. However, the reduction in HLA typing performance was modest: 1–2 percentage points for SpecImmune and 2–4 percentage points for StarPhase at two- and three-field resolution, with the gap narrowing to near zero at 4-field resolution, though ONT had ~3x higher mean coverage across the classical HLA genes. Commercial typing software such as GenDx NGSEngine or Omixon NanoTYPER may compensate for ONT's higher per-read error rate, but to our knowledge, no study has directly compared PacBio and ONT HLA typing on matched samples. Furthermore, even if commercial algorithms recover the correct star allele label, it remains unclear how accurately they reconstruct full-length haplotype sequences, an important consideration for evolutionary genetics and immunogenetics research. One confound in our comparison is that PacBio read lengths were approximately 1.5x longer than ONT, which may have affected phasing and haplotype reconstruction, though shorter reads alone do not explain the reduced SNP concordance given ONT's higher coverage depth.

In summary, we developed an open, end-to-end long-read HLA typing framework that combines a streamlined, automatable sequencing protocol with a fully integrated bioinformatics platform. Our results demonstrate that this approach delivers high-accuracy, high-resolution HLA allele calls across the classical HLA loci and yields informative coverage of clinically relevant genes in the HLA Class III region. By making both the sequencing protocol and HLA-Resolve freely available, we provide a practical and cost-effective alternative to proprietary commercial assays for clinical HLA typing and academic immunogenetics research.

Methods

Sample selection and cohort design

Thirty-three human cell line samples were selected from the Coriell Institute/Human Pangenome Reference Consortium (HPRC) and International Histocompatibility Working Group (IHWG) catalogs (Supplementary Table 1) to enable three complementary benchmarking analyses: (i) genotype concordance with the Genome in a Bottle (GIAB) benchmark, (ii) HLA haplotype reconstruction accuracy relative to the phased HPRC genome assemblies, and (iii) HLA typing concordance with the IHWG reference typings. The HPRC cohort comprised 17 gDNA samples obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research, thirteen of which were included in the first HPRC pangenome reference release[43]. To support trio-based evaluation of genotype concordance and phasing accuracy, we additionally included the parents of HG002 (NA24149, NA24143; NIST IDs HG003, HG004) and HG005 (NA24694, NA24695; NIST IDs HG006, HG007), forming the two GIAB benchmark trios. The second cohort consisted of gDNA from 16 B-lymphoblastoid cell lines from the IHWG Cell and DNA Bank, each with previously curated HLA typings, enabling independent benchmarking of HLA allele typing performance.

Tagmentation and Hybrid Capture

Long-read target-capture libraries were generated using a custom barcoded tagmentation strategy, followed by hybrid capture enrichment of HLA protein-coding genes. Custom barcoded transposase adapter oligonucleotides were designed for DNA tagmentation and adapter insertion. All oligonucleotides were ordered from IDT and resuspended at 100 µM in IDTE (Supplementary Table 1). Barcoded oligonucleotides have a 5' universal PCR adapter sequence (5'-cgaacatgtagctgactcaggtcac-3') followed by a 16bp barcode and end with the transposase mosaic end binding sequence (5'-AGATGTGTATAAGAGACAG-3'). Partially double-stranded adapters were formed by annealing each barcoded oligonucleotide to a mosaic end reverse oligonucleotide ([5'-PHO]CTGTCTCTTATACACATCT) following the Diagenode Transposome Assembly protocol. Briefly, 9 µL each of 100 µM barcoded and mosaic end reverse oligonucleotide were mixed with 2 µL 0.5M NaCl in a 20 µL final volume. The reaction was incubated at 95°C for 5 minutes, cooled to 65°C at -0.1°C per second, incubated at 65°C for 5 minutes, and then cooled to 4°C at -0.1°C per second. Annealed adapters were stored at -20°C.

Barcoded transposomes were assembled by binding 2 µL unloaded tagmentase (C01070010-20, Diagenode) with 2 µL partially double-stranded adapter at 23°C for 30 minutes. A single barcoded adapter was used per tagmentation reaction, ensuring that both fragment ends carry the same barcode and enabling amplification with a universal PCR primer. Assembled transposomes were diluted in Tagmentase Dilution Buffer (Diagenode, C01070011) for storage and titrated to working concentrations to yield desired fragment sizes.

Genomic DNA was tagmented using the assembled barcoded transposomes. Tagmentation reactions containing 0.5 µg of genomic DNA, 17 µL of 2X tagmentation buffer (C01019043-1000, Diagenode), and 2 µL of diluted barcoded transposome, in a final volume of 34 µL, were incubated at 55°C for 15 minutes. 8.5 µL of 0.2% SDS was added to stop the reaction, and the reaction was incubated at 68°C for 5 minutes to inactivate the tagmentase. Low-molecular-weight DNA and unincorporated adapters were removed using 21 µL (0.5X) of AxyPrep MAG PCR cleanup beads (Axygen MAGPCRCL500). After elution from beads, barcoded samples were pooled by sample cohort (Coriell/HPRC, IHWG) for gap repair. 50 µL of pooled tagmented DNA was mixed with 50 µL of 2X Kapa HiFi polymerase and incubated at 72°C for 10 minutes, followed by another bead cleanup. Fragment size distributions were assessed using the Agilent Genomic ScreenTape and reagents (Agilent 5067-5365, 5067-5366).

Hybrid capture enrichment was performed for the HLA genes. A custom probe set was designed in collaboration with Twist Bioscience, combining their existing probe panel targeting twelve HLA Class I (HLA-A, B, C) and II (HLA-DPA1, -DPB1, -DQA1, -DQA2, -DQB1, -DRB1, -DRB3, -DRB4, and -DRB5) genes with a novel probe set targeting the 69 HLA Class III protein-coding genes annotated in the human reference genome. The probes were tiled at 1kb intervals across each gene to promote uniform enrichment. Hybrid capture was performed separately for the HPRC and IHWG pools. Tagmented, gap-repaired DNA was quantified with a Qubit DNA High Sensitivity kit (Thermo Fisher Q32851), and 3.5–4 µg of tagmented DNA was used directly as input for hybrid capture following Twist's Long Read Library Preparation and Standard Hyb v2 Enrichment beginning at Step 4 (Preparing Libraries for Hybridization), except that a 100 µM Universal PCR Primer was substituted for Twist's Universal Blockers. The Twist HLA Class I/II probe set (Twist_LRHLA_v3_T123plus; Twist Design ID: LRTE-97930942) and the custom HLA Class III probe set (MHC_Class_III_LRTE-96945019_hg38; Twist Design ID: LRTE-96945019) were combined at a 1:1 ratio for hybridization. Capture and wash steps were performed using the Twist Standard Hyb and Wash Kit v2 (Twist, 104446) with Dynabeads M-270 Streptavidin (Invitrogen, 65305), and post-capture amplification was carried out using KOD Xtreme Hot Start DNA polymerase (MilliporeSigma, 71975-3).

Sequencing and read processing

The PacBio sequencing library was prepared following PacBio's Preparing Target Enrichment Libraries for PacBio Sequencing using the SMRTbell Prep Kit 3.0 (PacBio, 102-141-700), with 150 ng of each post-capture amplification product as input. The library was sequenced on a Revio instrument at PacBio (Menlo Park, CA), with a movie length of 1,440 minutes (base caller v5.0). The ONT sequencing library was prepared using ONT's Ligation Sequencing Kit V14 (SQK-LSK114) with 250–275 ng of each post-capture amplification product as input and sequenced on a Promethion R10.4 flowcell at the UCSC Sequencing Technology Center.

For the PacBio data, circular consensus sequences (HiFi reads) were generated using ccs (v8.0.1), quality-trimmed with trim (v1.1.0), and demultiplexed and adapter-trimmed with lima

(v2.10.0). PCR duplicates were removed before reference genome alignment using pbmarkdup. For ONT, base calling, demultiplexing, and adapter trimming were performed with Dorado (v0.7.2+9ac85c6). For both platforms, the transposase mosaic end binding sequences from tagmentation were removed with cutadapt[83]. Reads were aligned to the GRCh38 reference genome (GCA_000001405.15; no-alt analysis set) using minimap2[84]. ONT duplicates were marked after alignment using Picard MarkDuplicates.

Sequencing coverage was assessed with mosdepth v0.3.3[67], including only primary alignments with mapping quality of 20 or higher. For each sample, we computed mean coverage depth (Supplementary Tables 4–11) and the proportion of each feature covered by 20 or more reads (Supplementary Tables 8–11) for all genes, ncRNA genes, pseudogenes, and exons annotated in the HLA region of the GRCh38 GFF3 file. Fold enrichment was calculated as the ratio of the mean on-target to the mean off-target depth. On-target depth was defined as the total number of target-aligned bases divided by the target region length (982,574 bp per the Twist Bioscience expected coverage BED file). Off-target depth was defined as the total number of non-target-aligned bases divided by the total off-target genome length, using a genome size of 3,099,922,541 bp (GRCh38 no-alt analysis set).

Variant calling, phasing, and genotype concordance

Small variants were called with DeepVariant v1.6.1[85] for PacBio and Clair3 v2.0.0 (model r1041_e82_400bps_sup_v500)[86] for ONT. Both DeepVariant and Clair3 were evaluated on ONT reads, and Clair3 performed better. For the PacBio reads, earlier DeepVariant models (v1.6) performed better than newer versions (v1.8, v1.9) on our target capture data. Structural variants were called with pbsv (https://github.com/PacificBiosciences/pbsv) for PacBio and Sniffles2[87] for ONT. Tandem repeats were genotyped from PacBio reads using TRGT[88]. Phasing was performed with HiPhase[44] for PacBio and LongPhase[89] for ONT. HiPhase outputs haplotype block coordinates directly, which were used to assess phasing performance in downstream analyses. For LongPhase, haplotype block coordinates were extracted from the phase set (PS) tags using WhatsHap stats[90].

Genotype concordance

Genotyping accuracy was assessed following the GA4GH best practices for benchmarking germline small-variant calls[65]. Genotype concordance was calculated against the Genome in a Bottle (GIAB) high-confidence benchmark (v4.2.1) for samples HG002–HG007 across both sequencing platforms. The concordance analysis was restricted to genomic intervals with observed coverage to avoid misclassifying untargeted regions as false negatives. Specifically, we restricted the analysis to the 83 protein-coding HLA genes consistently captured by our probe set, using coordinates from the GRCh38 GFF3 annotation. Although targeted by our probe set, HLA-DRB3 and HLA-DRB4 were excluded as they are not part of the GRCh38 primary assembly. We adjusted the coordinates for HLA-B, HLA-F, and HLA-DQA1 to remove

several small regions not targeted by the Twist probe panel, corresponding to exons beyond the canonical transcript boundaries (Supplementary Figs. 9–11). We also excluded intron 9 of C4A and C4B, which contains a human endogenous retrovirus (HERV) insertion identical in sequence between the two paralogs that impedes alignment to the reference genome and results in coverage gaps (Supplementary Figs. 12 and 13). In GRCh38, C4A and C4B differ at only 18 nucleotide sites, clustered in the region spanning exons 20–31. Reads originating from intron 9 rarely extended into unique flanking sequence and therefore mapped ambiguously to both paralogs, resulting in a mapQ value of 0[84]. These restrictions ensure that genotype concordance metrics reflect variant calling performance within confidently sequenced intervals. Genotype concordance was assessed with hap.py (https://github.com/Illumina/hap.py), using rtgtools v3.12.1 vcfeval[91] as the comparison engine.

We assessed the relationship between genotype concordance and coverage depth by iteratively downsampling PacBio reads from HG002. Due to differences in baseline coverage depth across the three HLA classes, each class was downsampled independently using Picard DownsampleSam. For each downsample level, we computed the mean coverage depth of protein-coding genes, called variants with DeepVariant, and evaluated genotype concordance using hap.py. To account for stochastic variation introduced by random downsampling, we performed ten replicates per level and computed the mean and standard error for recall, precision, and F1 score.

HLA-Resolve

We developed HLA-Resolve, an HLA typing algorithm optimized for high-coverage PacBio HiFi reads. The pipeline takes raw sequencing reads in FASTQ or unmapped BAM format, reconstructs haplotypes for each HLA gene through a conventional variant-discovery workflow, and queries the IPD-IMGT/HLA allele database to return best-guess star allele calls for each classical HLA gene. The pipeline begins with adapter trimming with fastplong[92] or cutadapt[83], followed by alignment to a slightly modified version of the GRCh38 no-alt analysis set reference genome (GCA_000001405.15; see Supplementary Note 3 for details) with minimap2[84]. Coverage depth and breadth are evaluated for each target gene (HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1) by mosdepth[67]. Genotyping is performed separately for each variant class: SNVs with bcftools[93], indels with DeepVariant[85], structural variants with pbsv (https://github.com/PacificBiosciences/pbsv), and tandem repeats with TRGT[88]. Separate genotypers for SNVs and indels were chosen because DeepVariant commonly assigns RefCall (0/0 or ./.) genotypes when one HLA haplotype is reference-like and the other is highly divergent, a pattern that mimics segmental duplication absent from the reference. In these cases, pileup-based genotypers proved sufficient given our high coverage depth. However, bcftools did not reliably genotype indels in our reads, so DeepVariant is used for indel calling, and HLA-Resolve implements a rescue function that promotes indel records from FILTER=RefCall to FILTER=PASS when genotypes show read depth (DP) >= 30 and allele depth (AD) >= 10. Rescued variants are classified as heterozygous if

the variant allele frequency (VAF) falls between 0.3 and 0.7, or homozygous ALT if VAF exceeds 0.7. SNV genotypes from bcftools are filtered for DP >= 2, GQ >= 20, and QUAL >= 10.

The resulting genotypes are jointly phased with HiPhase[44] and merged with bcftools concat. HLA-Resolve then filters redundant variant calls before haplotype reconstruction. Because DeepVariant and pbsv may both genotype the same indels, and both may call variants at repeat loci better handled by TRGT, duplicate records must be removed before haplotype reconstruction. HLA-Resolve's filter first collects all pass-filter, phased structural variant genotypes from pbsv and records their genomic coordinates and affected haplotype. Each non-SV indel is then checked for overlap with a collected SV on the same haplotype and suppressed if overlap is detected. An additional proximity check suppresses large indels (≥50 bp) within 10 bp of an SV breakpoint, which may represent redundant calls due to left-alignment annotation differences between callers. The SV overlap filter never suppresses SNPs, as they may represent true variants near SV breakpoints rather than redundant calls. A parallel filter is applied for TRGT, where non-TRGT variants falling entirely within annotated TRGT tandem repeat genotypes are suppressed, as the TRGT genotype already encodes the correct allele for the entire repeat span.

HLA-Resolve then inspects the haplotype blocks returned by HiPhase. Because haplotype block coordinates are defined by the first and last heterozygous genotype in a phase set, they exclude flanking homozygous bases. Following Holt et al.[44], HLA-Resolve extends each block overlapping an HLA gene through adjacent homozygous regions up to, but not including, the next heterozygous genotype in either direction, then verifies that each gene is completely spanned by a single haplotype block before proceeding. Genes containing one or fewer heterozygous genotypes often do not fall within a defined phase set or haplotype block but are nonetheless effectively phased. When a single heterozygous genotype is present, allele assignment to each haplotype is arbitrary, and fully homozygous genes are identical on both haplotypes. Accordingly, genes with at most one heterozygous genotype are considered fully phased. Genes with internal phasing breaks follow a separate pipeline to extract the largest haplotype block possible for database querying (detailed below).

For genes meeting coverage depth and phasing requirements, HLA-Resolve reconstructs full-length haploid nucleotide sequences by applying phased, pass-filter genotypes to the GRCh38 GFF3 gene models using vcf2fasta[66]. The default GRCh38 models (MANE Select) were used for all HLA genes, with minor modifications to HLA-C and HLA-DQB1. HLA-DQB1 is segregating for a splicing variant that skips exon 5, and the GRCh38 MANE Select transcript for HLA-DQB1 omits this exon, so we restored it, as HLA-Resolve performs germline HLA typing rather than transcript characterization. For HLA-C, most IPD-IMGT/HLA reference sequences include substantially more UTR than the GRCh38 HLA-C GFF3 annotation, so we added an extra kilobase of padding to each end of the gene model for full-gene haplotype reconstruction.

If a phasing break occurs within a gene, HLA-Resolve searches for the longest haplotype block that fully spans the antigen recognition sequence (ARS). When such a haplotype block exists, its coordinates are used to extract the corresponding region from the raw vcf2fasta output (FASTA). If no haplotype block fully spans the ARS, HLA-Resolve enters a rescue mode to extract as much information as possible (Supplementary Note 4). Even within defined haplotype blocks, a small fraction of genotypes may remain unphased, typically reflecting complex structural variants or low-confidence SNV genotypes. These unphased genotypes cannot be applied in a haplotype-aware manner and are therefore ignored, as are complex structural variant types such as breakend (BND), inversion (INV), and duplication (DUP), which are not currently compatible with tools like vcf2fasta and bcftools consensus. HLA-Resolve reports any such variants and any unphased genotypes that could not be applied.

Following haplotype reconstruction, HLA-Resolve compares reconstructed haplotypes to alleles in the IPD-IMGT/HLA database using an iterative, hierarchical classification algorithm. Query and reference sequences are compared using the edlib library[94], which computes Levenshtein edit distance and returns the corresponding optimal alignment path as a CIGAR string. Edit distance is calculated in infix (Hybrid Wunsch) mode, in which leading and trailing gaps are not penalized, accommodating partial or incomplete reference sequences in the database as well as variable start and stop coordinates arising from different PCR primer designs. HLA-Resolve derives two additional classification metrics from the alignment path: mismatch identity and match length. Mismatch identity is the proportion of matching bases at 1:1-aligned positions. Match length is defined as the total number of identical aligned nucleotide positions.

HLA-Resolve first attempts to assign each reconstructed HLA haplotype to a G group based on the antigen recognition sequence (ARS) of the protein (exons 2 and 3 for HLA Class I; exon 2 for HLA Class II). G-group reference sequences are derived from the allele exon annotations in the IPD-IMGT/HLA hla.xml database and compared against the query ARS. G-group assignment requires an exact (zero-edit-distance) match between the query and a G-group ARS sequence. If a perfect match is identified, the corresponding G group is assigned, and a second search is performed using the full concatenated exon sequence, restricted to alleles within that G group. Candidate alleles are ranked by edit distance to provide the three-field allele assignment. Ties are broken by match length, then alphabetically. If no exact G-group match is found, a full unrestricted search compares the concatenated exon query sequence against all alleles of the corresponding gene, using the same ranking and tiebreaking chain.

If the three-field allele assignment includes alleles with defined fourth-field values, a third search is performed using the full-gene reconstructed haplotype (including introns and UTRs). Candidates are ranked by mismatch identity rather than edit distance to avoid penalizing gaps from unreliable intronic reconstruction. Ties are broken first by match length, favoring alleles with more aligned evidence, then by the lowest fourth-field value. This last tiebreaker is effectively an allele-frequency prior, because 4th-field designations are assigned chronologically and earlier-annotated alleles tend to be the most common. HLA-Resolve reports a single best-match allele by default, as accommodating ambiguities complicates most research applications.

We verified that star allele concordance was not inflated by the 4th-field tiebreaking chain (See Supplementary Note 5).

The IHWG reference typings were used to benchmark HLA-Resolve (IPD-IMGT/HLA v3.63.0) against three long-read-friendly HLA typing algorithms: StarPhase v2.0.1 (IPD-IMGT/HLA v3.62.0-alpha)[31], HLA*LA v1.0.4[22], and SpecImmune v1.0.0 (release v0.0.2, IPD-IMGT/HLA v3.62.0)[32]. All three require a mapped BAM as input. For StarPhase and HLA*LA, reads were mapped to the GRCh38 no alts analysis set reference genome. For SpecImmune, reads were mapped to the GRCh38 full analysis set (GCA_000001405.15_GRCh38_full_plus_hs38d1_analysis_set) as recommended in its documentation. Reference HLA typings were obtained from the Fred Hutch Cancer Center (https://www.fredhutch.org/en/research/institutes-networks-ircs/international-histocompatibility-working-group.html) and were, in most cases, identical to those available in the IPD-IMGT/HLA database. A small number of conflicts were manually resolved (Supplementary Note 6). These edits did not affect concordance between the tested HLA typing algorithms because they occurred in cases where all algorithms returned identical calls. The final truth HLA star allele labels used in benchmarking are available in Supplementary Table 17.

To verify the accuracy of the reconstructed haplotypes, we calculated the edit distance between the HLA-Resolve output and the corresponding gene regions in the HPRC haploid genome assemblies. The HPRC v1 assemblies and their GFF3 gene annotations were downloaded from Ensembl (https://projects.ensembl.org/hprc/). Full-gene FASTA haplotypes for the classical HLA genes were extracted from each assembly using bedtools getfasta[95] with start and stop coordinates from the corresponding GFF3 annotation. Because HLA-Resolve's phased haplotypes are unlabeled with respect to parental origin, we tested both possible pairings against the HPRC maternal and paternal assemblies and selected the one with the lower total gap-compressed edit distance. The HLA haplotypes produced by HLA-Resolve were compared to their corresponding haplotypes in the HPRC assembly using edlib and parasail[96]. For each haplotype comparison, edlib was first run in infix (HW) mode to locate where the shorter sequence aligns within the longer, and the aligned region was then globally aligned with parasail using affine gap penalties (open=5, extend=1). Edlib computes minimum edit distance with uniform penalties for substitutions, insertions, and deletions, which produces reasonable alignments when the two sequences are highly similar, as was the case for HLA-A, -B, -C, -DPA1, -DPB1. However, for HLA-DQA1 -DQB1, and -DRB1, complex structural variants in introns and UTRs that are difficult for HLA-Resolve to genotype and phase may be absent from the reconstructed sequences, and edlib's uniform penalties yielded biologically unreasonable alignment paths. Parasail's affine gap model, which penalizes gap opening more heavily than gap extension, produced more realistic alignments for these regions. We therefore used edlib to determine alignment bounds and parasail for final alignment and scoring, though in practice, parasail was only necessary for the HLA-DQ and -DR genes. For each parasail alignment, we computed a gap-compressed edit distance by parsing the CIGAR string, treating each mismatch individually, and treating contiguous insertion or deletion characters as a single event.

Data Availability

Raw sequencing reads for the Coriell/HPRC samples are available through the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1417783 (SRR37075979-SRR37076010).

Code Availability

HLA-Resolve is available at https://github.com/matthewglasenapp/hla_resolve. The code supporting the findings of this study is available on Dryad (DOI: 10.5061/dryad.sf7m0cgn0).

Acknowledgments

We thank the following individuals for their contributions to this work. Arthur Chung, Casey Riegler, and Rebecca Barnard of Twist Bioscience developed the custom HLA Class III hybrid capture probe panel. PacBio sequencing was supported by Noelle Bittner, Ian McLaughlin, and Sarah Kingan, with additional bioinformatics support provided by Matthew Seetin and Xiao Chen. Brandy McNulty and Ivo Violich of the UC Santa Cruz Sequencing Technology Center assisted with nanopore sequencing and basecalling, and Jonathan Foos, Nicholas Scales, and Chris Still of Oxford Nanopore Technologies provided flow cells, sequencing kits, and technical support.

References

1. Kennedy, A. E., Ozbek, U. & Dorak, M. T. What has GWAS done for HLA and disease associations? *Int. J. Immunogenet.* **44**, 195–211 (2017).
2. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
3. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
4. Long, X. & Xue, H. Genetic-variant hotspots and hotspot clusters in the human genome facilitating adaptation while increasing instability. *Hum. Genomics* **15**, 19 (2021).
5. Lenz, T. L., Spirin, V., Jordan, D. M. & Sunyaev, S. R. Excess of deleterious mutations around HLA genes reveals evolutionary cost of balancing selection. *Mol. Biol. Evol.* **33**, 2555–2564 (2016).
6. Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.* **54**, 15–39 (2009).
7. Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annu. Rev. Genomics Hum. Genet.* **14**, 301–323 (2013).
8. Lee, S. J. *et al.* High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. *Blood* **110**, 4576–4583 (2007).
9. Spellman, S. R. *et al.* A perspective on the selection of unrelated donors and cord blood units for transplantation. *Blood* **120**, 259–265 (2012).
10. Montgomery, R. A., Tatapudi, V. S., Leffell, M. S. & Zachary, A. A. HLA in transplantation. *Nat. Rev. Nephrol.* **14**, 558–570 (2018).
11. Kok, G. *et al.* The effect of genetic HLA matching on liver transplantation outcome: A systematic review and meta-analysis. *Ann. Surg. Open* **4**, e334 (2023).
12. Dilthey, A. T. State-of-the-art genome inference in the human MHC. *Int. J. Biochem. Cell Biol.* **131**, 105882 (2021).
13. Brandt, D. Y. C. *et al.* Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project Phase I data. *G3 Genes Genomes Genet.* **5**, 931–941 (2015).
14. Klasberg, S., Surendranath, V., Lange, V. & Schöfl, G. Bioinformatics strategies, challenges, and opportunities for next generation sequencing-based HLA genotyping. *Transfus. Med. Hemotherapy* **46**, 312–325 (2019).
15. Bai, Y., Ni, M., Cooper, B., Wei, Y. & Fury, W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics* **15**, 325 (2014).
16. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).
17. Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
18. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.* **38**, 788–797 (2017).
19. Song, L., Bai, G., Liu, X. S., Li, B. & Li, H. Efficient and accurate KIR and HLA genotyping with massively parallel sequencing data. *Genome Res.* **33**, 923–931 (2023).
20. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R. & McVean, G. Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47**, 682–688 (2015).
21. Dilthey, A. T. *et al.* High-accuracy HLA type inference from whole-genome sequencing data using population reference graphs. *PLOS Comput. Biol.* **12**, e1005151 (2016).
22. Dilthey, A. T. *et al.* HLA*LA-HLA typing from linearly projected graph alignments. *Bioinformatics* **35**, 4394–4396 (2019).
23. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
24. Liu, C. *et al.* ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res.* **41**, e142 (2013).
25. Huang, Y. *et al.* HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome Med.* **7**, 25 (2015).

26. Lee, H. & Kingsford, C. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome Biol.* **19**, 16 (2018).
27. Wang, S. *et al.* SpecHLA enables full-resolution HLA typing from sequencing data. *Cell Rep. Methods* **3**, 100589 (2023).
28. Bochtler, W. *et al.* A comparative reference study for the validation of HLA-matching algorithms in the search for allogeneic hematopoietic stem cell donors and cord blood units. *HLA* **87**, 439–448 (2016).
29. Maiers, M. *et al.* GRIMM: Graph imputation and matching for HLA genotypes. *Bioinformatics* **35**, 3520–3523 (2019).
30. Liu, C. *et al.* Accurate typing of human leukocyte antigen class I genes by Oxford Nanopore sequencing. *J. Mol. Diagn.* **20**, 428–435 (2018).
31. Holt, J. *et al.* StarPhase: comprehensive phase-aware pharmacogenomic diplotyper for long-read sequencing data. 2024.12.10.627527 Preprint at https://doi.org/10.1101/2024.12.10.627527 (2024).
32. Wang, S. *et al.* A scalable framework for comprehensive typing of polymorphic immune genes from long-read data. *Adv. Sci.* e21531 (2025) doi:10.1002/advs.202521531.
33. Kulski, J. K., Suzuki, S. & Shiina, T. Human leukocyte antigen super-locus: nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes. *Hum. Genome Var.* **9**, 1–15 (2022).
34. White, P. C. & Speiser, P. W. Congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Endocr. Rev.* **21**, 245–291 (2000).
35. Liu, C. A long road/read to rapid high-resolution HLA typing: The nanopore perspective. *Hum. Immunol.* **82**, 488–495 (2021).
36. Hammond, N. *et al.* Analytical validation of germline small variant detection using long-read HiFi genome sequencing. *Genome Res.* **35**, 1391–1399 (2025).
37. Höps, W. *et al.* HiFi long-read genomes for difficult-to-detect, clinically relevant variants. *Am. J. Hum. Genet.* **112**, 450–456 (2025).
38. Negi, S. *et al.* Advancing long-read nanopore genome assembly and accurate variant calling for rare disease detection. *Am. J. Hum. Genet.* **112**, 428–449 (2025).
39. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
40. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
41. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
42. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
43. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
44. Holt, J. *et al.* HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* **40**, btae042 (2024).
45. Barker, D. J. *et al.* The IPD-IMGT/HLA Database. *Nucleic Acids Res.* **51**, D1053–D1060 (2023).
46. Marsh, S. G. E. *et al.* Nomenclature for factors of the HLA system, 2026. *HLA* **107**, e70595 (2026).
47. Mayor, N. P. *et al.* Recipients receiving better HLA-matched hematopoietic cell transplantation grafts, uncovered by a novel HLA typing method, have superior survival: A retrospective study. *Biol. Blood Marrow Transplant.* **25**, 443–450 (2019).
48. Hurley, C. K. *et al.* Regarding "recipients receiving better hla-matched hematopoietic cell transplantation grafts, uncovered by a novel hla typing method, have superior survival: a retrospective study". *Biol. Blood Marrow Transplant.* **25**, e268–e269 (2019).
49. Mayor, N. P. *et al.* Impact of previously unrecognized HLA mismatches using ultrahigh resolution typing in unrelated donor hematopoietic cell transplantation. *J. Clin. Oncol.* **39**, 2397–2409 (2021).
50. Handunnetthi, L., Ramagopalan, S. V., Ebers, G. C. & Knight, J. C. Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. *Genes Immun.* **11**, 99–112 (2010).
51. Baxter-Lowe, L. A. The changing landscape of HLA typing: understanding how and when HLA typing data can be used with confidence from bench to bedside. *Hum. Immunol.* **82**, 466–477 (2021).
52. Arumugam, T., Adimulam, T., Gokul, A. & Ramsuran, V. Variation within the non-coding genome influences genetic and epigenetic regulation of the human leukocyte antigen genes. *Front. Immunol.* **15**, 1422834 (2024).
53. Mayor, N. P. *et al.* HLA typing for the next generation. *PLOS ONE* **10**, e0127153 (2015).
54. Albrecht, V. *et al.* Dual redundant sequencing strategy: full-length gene characterisation of 1056 novel and confirmatory HLA alleles. *HLA* **90**, 79–87 (2017).

55. Liu, C. *et al.* High-resolution HLA typing by long reads from the R10.3 Oxford nanopore flow cells. *Hum. Immunol.* **82**, 288–295 (2021).
56. Centurião, N. de F., Maciel, G. C., Goldenstein, M. F. L., Alonso, E. O. & Torres, M. A. The novel HLA-C*03:678 allele identified using nanopore sequencing. *HLA* **105**, e70038 (2025).
57. Laver, T. W. *et al.* Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci. Rep.* **6**, 21746 (2016).
58. Yin, Y. *et al.* Application of high-throughput next-generation sequencing for HLA typing on buccal extracted DNA: results from over 10,000 donor recruitment samples. *PLOS ONE* **11**, e0165810 (2016).
59. Turner, T. R. *et al.* Single molecule real-time DNA sequencing of HLA genes at ultra-high resolution from 126 International HLA and Immunogenetics Workshop cell lines. *HLA* **91**, 88–101 (2018).
60. Steiert, T. A. *et al.* High-throughput method for the hybridisation-based targeted enrichment of long genomic fragments for PacBio third-generation sequencing. *NAR Genomics Bioinforma.* **4**, lqac051 (2022).
61. Wittig, M. *et al.* Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res.* **43**, e70 (2015).
62. Norman, P. J. *et al.* Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. *Genome Res.* **27**, 813–823 (2017).
63. Lai, S.-K. *et al.* A novel framework for human leukocyte antigen (HLA) genotyping using probe capture-based targeted next-generation sequencing and computational analysis. *Comput. Struct. Biotechnol. J.* **23**, 1562–1571 (2024).
64. Wade, K. J. *et al.* MHConstructor: a high-throughput, haplotype-informed solution to the MHC assembly challenge. *Genome Biol.* **25**, 274 (2024).
65. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
66. Sanchez-Ramirez, S. vcf2fasta. (2017).
67. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
68. Noble, J. A. & Erlich, H. A. Genetics of Type 1 Diabetes. *Cold Spring Harb. Perspect. Med.* **2**, a007732 (2012).
69. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
70. Brown, M. A., Kenna, T. & Wordsworth, B. P. Genetics of ankylosing spondylitis—insights into pathogenesis. *Nat. Rev. Rheumatol.* **12**, 81–91 (2016).
71. Sollid, L. M. & Thorsby, E. HLA susceptibility genes in celiac disease: Genetic mapping and role in pathogenesis. *Gastroenterology* **105**, 910–922 (1993).
72. Pappas, D. J. *et al.* Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. *Pharmacogenomics J.* **18**, 367–376 (2018).
73. Yang, Y. *et al.* The intricate role of complement component C4 in human systemic lupus erythematosus. in *Complement in Autoimmunity* 98–132 (Karger, 2003). doi:10.1159/000075689.
74. Burch, G. H. *et al.* Tenascin–X deficiency is associated with Ehlers–Danlos syndrome. *Nat. Genet.* **17**, 104–108 (1997).
75. Schalkwijk, J. *et al.* A recessive form of the ehlers–danlos syndrome caused by tenascin-X deficiency. *N. Engl. J. Med.* **345**, 1167–1175 (2001).
76. Chen, X. *et al.* Comprehensive *SMN1* and *SMN2* profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. *Am. J. Hum. Genet.* **110**, 240–250 (2023).
77. Satta, Y., Mayer, W. E. & Klein, J. HLA-DRB intron 1 sequences: implications for the evolution of HLA-DRB genes and haplotypes. *Hum. Immunol.* **51**, 1–12 (1996).
78. Gustafson, J. A. *et al.* High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res.* **34**, 2061–2073 (2024).
79. Thuesen, N. H., Klausen, M. S., Gopalakrishnan, S., Trolle, T. & Renaud, G. Benchmarking freely available HLA typing algorithms across varying genes, coverages and typing resolutions. *Front. Immunol.* **13**, 987655 (2022).
80. Abi-Rached, L. *et al.* Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLOS ONE* **13**, e0206512 (2018).
81. Chin, C.-S. *et al.* A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat. Commun.* **11**, 4794 (2020).

82. Behera, S. *et al.* Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat. Biotechnol.* **43**, 1177–1191 (2025).
83. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.j.* **17**, 10–12 (2011).
84. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
85. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
86. Zheng, Z. *et al.* Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci.* **2**, 797–803 (2022).
87. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
88. Dolzhenko, E. *et al.* Characterization and visualization of tandem repeats at genome scale. *Nat. Biotechnol.* **42**, 1606–1614 (2024).
89. Lin, J.-H., Chen, L.-C., Yu, S.-C. & Huang, Y.-T. LongPhase: an ultra-fast chromosome-scale phasing algorithm for small and large variants. *Bioinformatics* **38**, 1816–1822 (2022).
90. Patterson, M. *et al.* WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **22**, 498–509 (2015).
91. Cleary, J. G. *et al.* Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. 023754 Preprint at https://doi.org/10.1101/023754 (2015).
92. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, e107 (2023).
93. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
94. Šošić, M. & Šikić, M. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics* **33**, 1394–1395 (2017).
95. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
96. Daily, J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinformatics* **17**, 81 (2016).
97. Alexandrov, N., Wang, T., Blair, L., Nadon, B. & Sayer, D. HLA-OLI: a new MHC class I pseudogene and HLA-Y are located on a 60 kb indel in the human MHC between HLA-W and HLA-J. *HLA* **102**, 599–606 (2023).

Tables and Figures

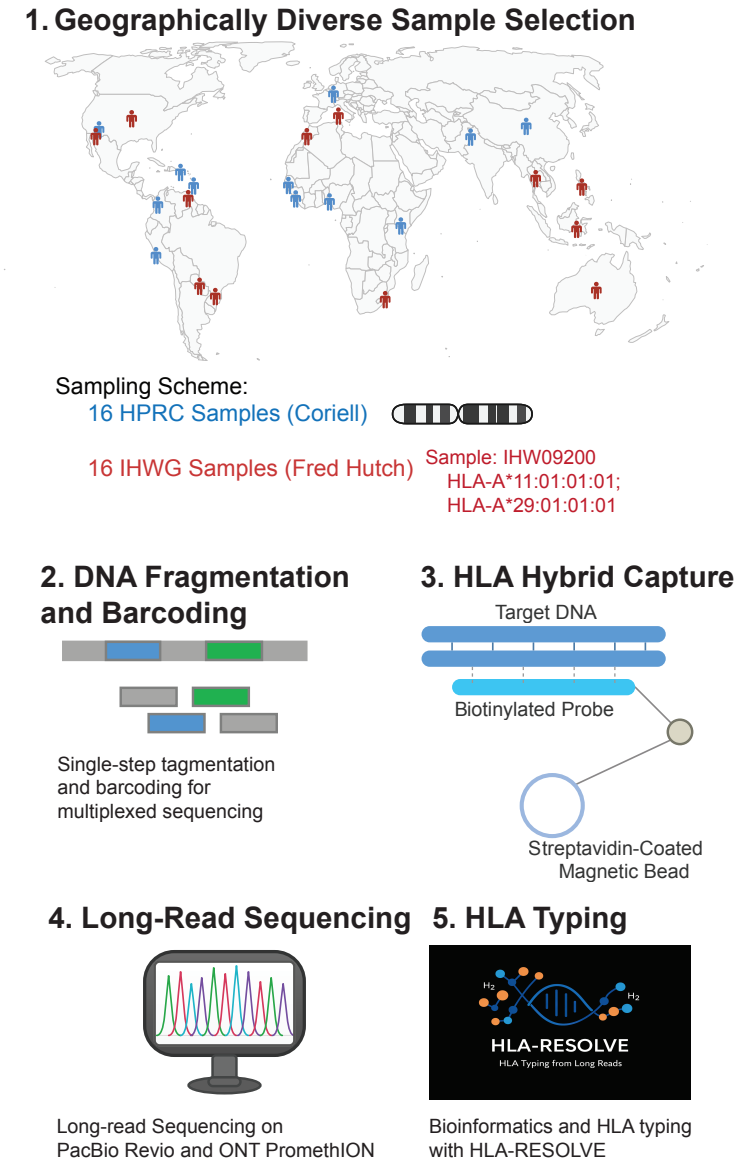


Fig. 1: Overview of the targeted long-read sequencing workflow. (1) Thirty-three geographically diverse samples were selected from two cohorts, 17 from the Human Pangenome Reference Consortium (HPRC, Coriell) and 16 from the International Histocompatibility Working Group (IHWG, Fred Hutch). Some samples share a common ancestral origin and are represented by a single map location. (2) Genomic DNA from each sample was tagmented and barcoded for multiplexed sequencing. (3) Barcoded libraries were pooled by cohort for hybrid capture using a custom probe set from Twist Bioscience targeting protein-coding genes in the HLA region. (4) Post-capture libraries from both cohorts were sequenced on PacBio Revio and Oxford Nanopore PromethION platforms. (5) HLA typing was performed with HLA-Resolve.

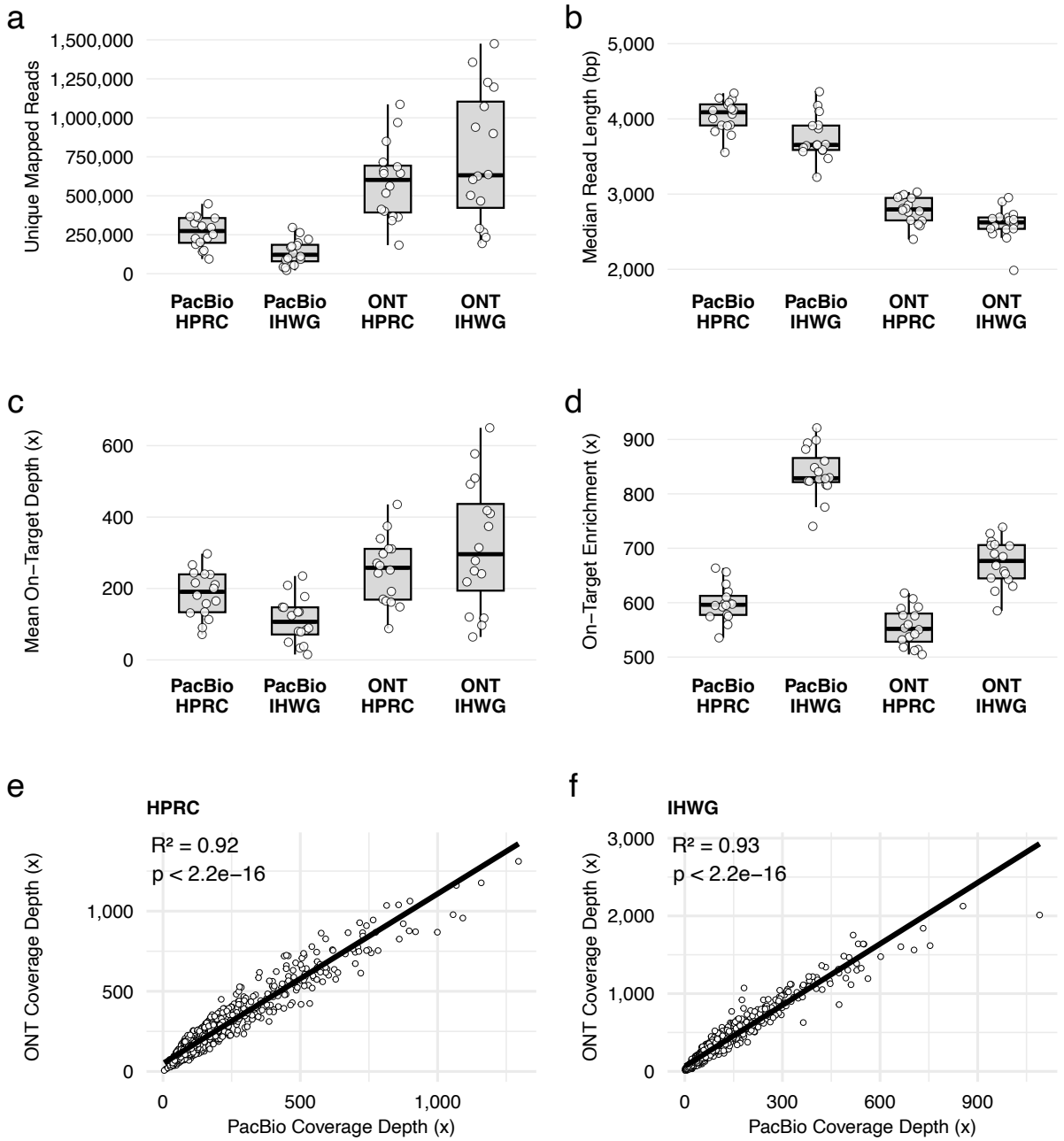


Fig. 2: Summary of targeted long-read sequencing performance. a, Number of mapped sequencing reads per sample, with PCR duplicates removed. **b,** Median read length of target-overlapped primary alignments. **c,** Mean on-target coverage depth (x). **d,** On-target enrichment, defined as the ratio of mean on-target to off-target coverage depth. For panels a–d, results are stratified by sequencing platform (PacBio or ONT) and hybrid capture cohort (HPRC or IHWG), and points represent individual samples. **e–f**, Comparison of ONT versus PacBio coverage depth across 83 consistently captured protein-coding genes for matched samples from the HPRC (e) and IHWG (f) cohorts. Each point represents a gene-sample pair. The solid line shows the linear regression fit, with R^2 and p-values annotated.

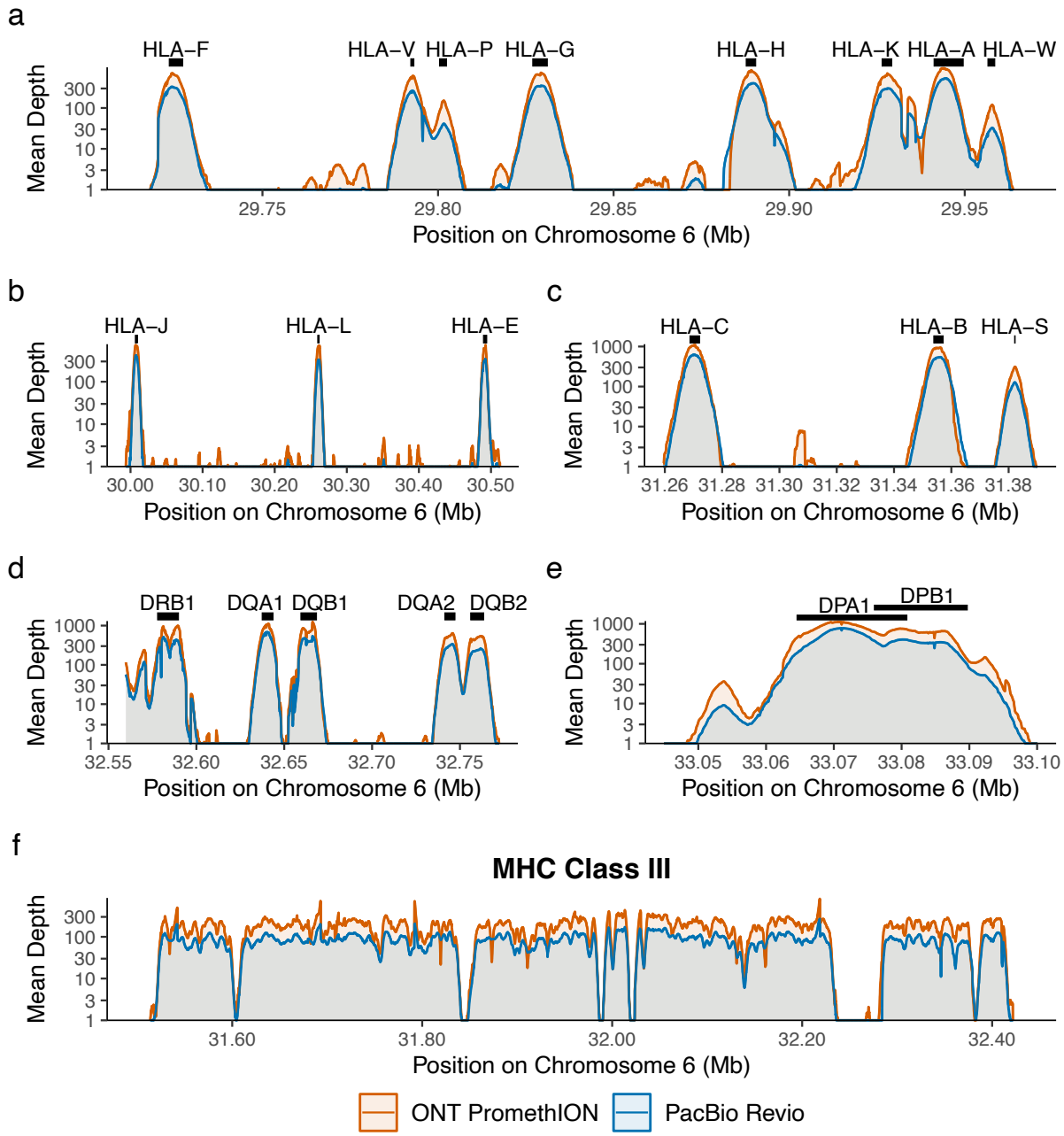


Fig. 3: Average coverage depth across selected HLA regions. Mean coverage depth (log scale) across the HLA region is shown for PacBio Revio (blue) and ONT PromethION (orange), using per-base coverage averaged across samples and aggregated into non-overlapping 100 bp windows. **a-c,** The HLA Class I region, including the classical antigen-presenting genes, HLA-A, HLA-B, HLA-C, and selected non-classical HLA genes and pseudogenes. **d-e,** The HLA Class II region, including the Class II antigen-presenting genes. **f,** The entire HLA Class III region, containing 69 protein-coding genes (not labeled). Coverage gaps reflect intergenic regions not targeted by the hybrid capture probes, except for the two small gaps flanking 32.00 Mb, corresponding to the C4A and C4B intron 9 HERV insertion.

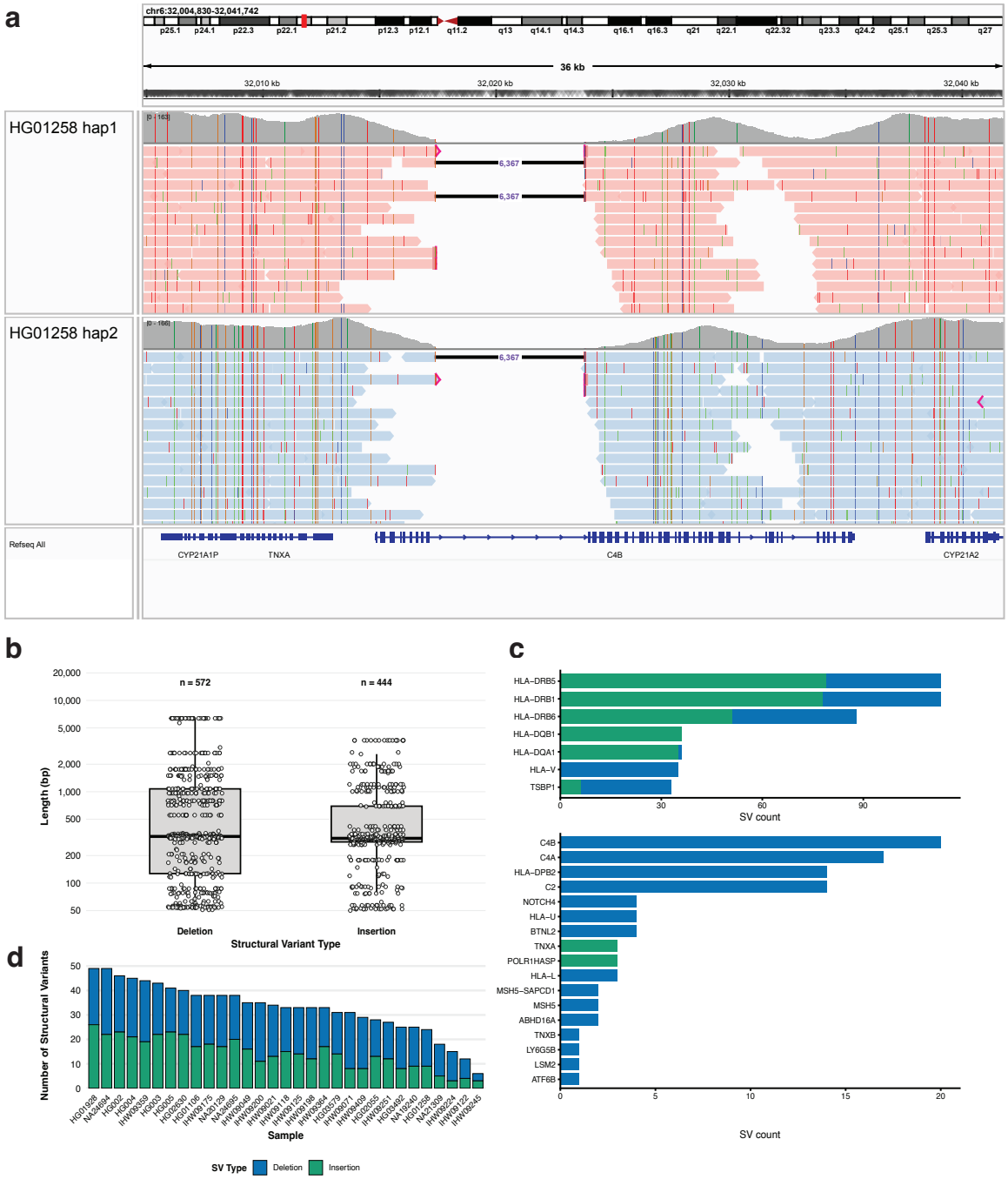
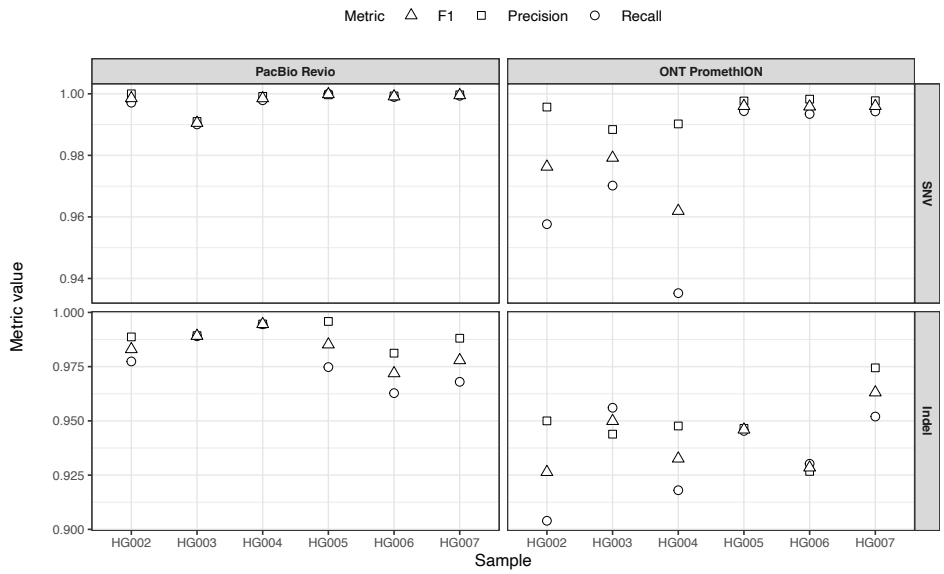
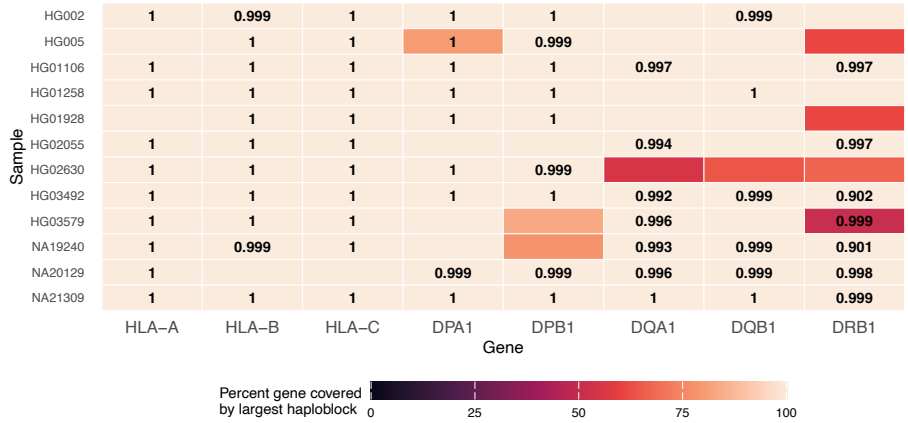


Fig. 4: Characterization of CYP21 read mapping and HLA-wide structural variation. a, An IGV plot showing primarily alignments to CYP21A2 and CYP21A1P. Also shown is a large 6,367 bp deletion in C4B, representing the short C4B isoform lacking the intron 9 HERV insertion. **b**, Distribution of structural variant (SV) lengths by type. Each point represents one SV in an individual sample. The y-axis is log10-scaled. Total SV counts across all samples are shown above each column. **c**, Number of SVs overlapping annotated genes, stratified by SV type. The plot is split into two panels with different x-axis scales: x > 30 (top) and x < 30 (bottom). **d**, Total SV count per sample, stratified by SV type. Samples are ordered by decreasing total SV count. For **b-d**, only SVs greater than 49 base pairs in length were included.

a



b



c

Resolution	HLA–A	HLA–B	HLA–C	HLA–DPA1	HLA–DPB1	HLA–DQA1	HLA–DQB1	HLA–DRB1
1–field	100% (30/30)	100% (30/30)	100% (30/30)	100% (26/26)	96% (27/28)	96% (27/28)	96% (27/28)	90% (27/30)
2–field	100% (28/28)	100% (28/28)	100% (28/28)	100% (26/26)	96% (27/28)	96% (27/28)	96% (27/28)	86% (24/28)
3–field	100% (28/28)	100% (26/26)	100% (27/27)	100% (25/25)	100% (21/21)	96% (23/24)	96% (25/26)	86% (18/21)
4–field	100% (24/24)	100% (19/19)	83% (20/24)	78% (18/23)	94% (15/16)	55% (11/20)	70% (14/20)	62% (10/16)

Fig. 5: HLA genotype concordance and HLA haplotype accuracy: a, Single-nucleotide variant (top) and indel (bottom) genotype concordance with the Genome in a Bottle (GIAB) benchmark for HG002–HG007, stratified by sequencing platform (PacBio Revio, left; ONT PromethION, right). Precision, recall, and F1 score are shown for each sample, computed across captured protein-coding genes within the HLA region. **b,** Heat map showing, for each PacBio-sequenced sample and HLA gene, the percentage of the gene sequence spanned by the largest contiguous haplotype block generated by HiPhase (color scale). Text annotations within each cell report haplotype sequence identity to the corresponding Human Pangenome Reference Consortium (HPRC) phased assembly sequences, computed as the mean of per-haplotype sequence identities. Identity values are truncated to three decimal places for display. A value of 1 is shown only when both haplotypes exhibit perfect sequence identity to the HPRC. Cells with missing values reflect missing annotations in the HPRC GFF3 files. **c,** Star allele concordance of HLA-Resolve calls with IHWG reference typings, stratified by gene and resolution level.

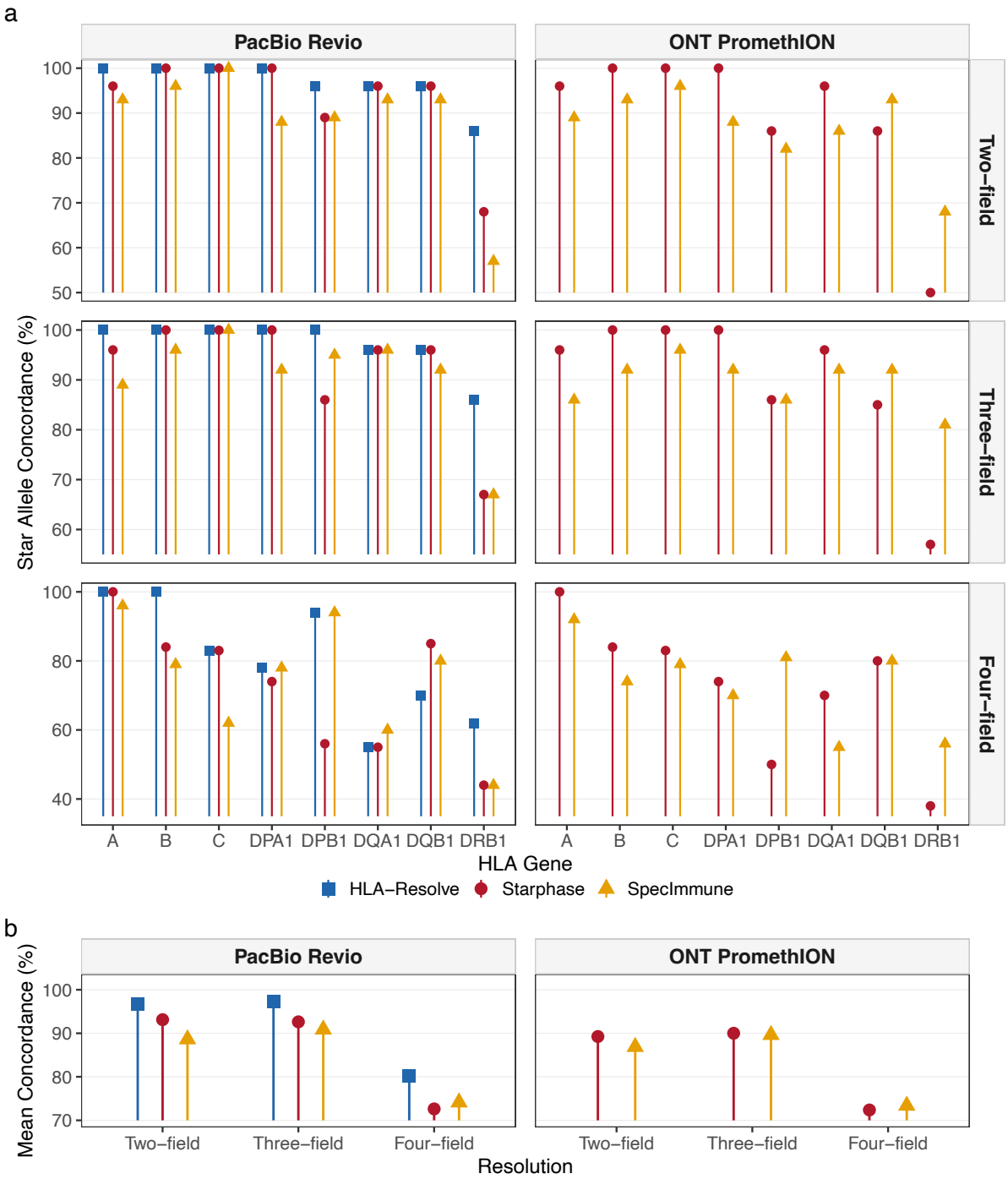


Fig. 6: Star allele concordance with the IHWG reference typings for the 15 IHWG samples. a, Star allele concordance (%) with IHWG reference typings for HLA-Resolve (squares), StarPhase (circles), and SpecImmune (triangles), shown for each HLA gene and stratified by typing resolution (two-, three-, or four-field) and sequencing platform (PacBio Revio, ONT PromethION). Each point represents the concordance for a given gene across the 15 IHWG samples. HLA-Resolve results are shown for PacBio Revio only. **b**, Mean concordance across all eight HLA genes, summarized by resolution and platform.

Supplementary Information

Supplementary Note 1: Variation in sequencing depth and duplication rates between hybrid capture cohorts

In the ONT PromethION sequencing run, the IHWG samples received approximately 5 times as many reads as the HPRC samples, likely due to unequal pooling of the two capture reactions before ONT library preparation (Supplementary Table 3). This imbalance was not present in the PacBio data, where the IHWG-to-HPRC read ratio was approximately 1:1 (Supplementary Table 2). PCR duplication rates were consistent within each cohort but differed substantially between cohorts due to overamplification of the IHWG pool during post-capture PCR. The average duplication rate for the HPRC samples was 19% for PacBio and 18% for ONT, compared to 61% (PacBio) and 83% (ONT) for the IHWG samples (Supplementary Tables 2 and 3). For PacBio, the higher duplication rate in the IHWG cohort led to slightly lower on-target depth than in the HPRC cohort. This difference was not observed in the ONT data, where the disproportionate share of sequencing output allocated to the IHWG cohort offset its elevated duplication rate. Additionally, the equal mixture of the two Twist probe panels resulted in 3-4X greater coverage depth for Class I/II genes than for Class III genes on both platforms.

Supplementary Note 2: Contamination in sample IHW09117

One sample, IHW09117, showed signatures of contamination and was removed from the analysis. This sample showed an abundance of tri-allelic sites, a pattern absent from all other samples. Manual inspection of read alignments in IGV revealed >2 distinct haplotypes with strong support at multiple HLA genes. We were unable to determine whether the contamination occurred at the Fred Hutchinson Cancer Center or during our procedure.

Supplementary Note 3: Reference genome modifications for HLA-Resolve

Mismapping of reads from paralogous loci introduced spurious genotype calls, with HLA-Y reads mapping to HLA-A and HLA-DRB3 and -DRB4 reads mapping to HLA-DRB1. HLA-Y and HLA-OLI are newly identified pseudogenes located within a polymorphic 60kb insertion that is absent from the GRCh38 no alt analysis set[97]. HLA-Y shows the highest sequence similarity to HLA-A of the HLA Class I paralogs/pseudogenes (see Fig. 3 from ref. [97]), and when present in a sample, reads originating from HLA-Y may map to HLA-A[14]. We verified the identity of HLA-Y reads by comparing them to the HLA-Y alleles in the latest IPD-IMGT/HLA database, and preventing them from mapping to HLA-A by appending the 60kb insertion sequence containing HLA-Y and HLA-OLI from GenBank sample HG00735

(GenBank accession JAHBCG010000038.1) to the GRCh38 reference genome as an additional scaffold.

We devised a similar approach to handle reads originating from different HLA-DRB paralogs that mismap to HLA-DRB1. The GRCh38 primary assembly contains two functional HLA-DRB paralogs, HLA-DRB1 and HLA-DRB5. The Twist probes also captured two additional functional paralogs, HLA-DRB3 and HLA-DRB4, whose reads can mismap to HLA-DRB1 and interfere with allele calling. To filter out potentially mismapped reads, we constructed a multi-allele reference containing several full-length genomic sequences of HLA-DRB alleles from the IPD-IMGT/HLA database: thirteen HLA-DRB1 alleles (one representative per allele group), three HLA-DRB3 alleles, and one HLA-DRB4 allele. An early step in the HLA-Resolve workflow maps all reads against this reference to identify and remove those with primary alignments to HLA-DRB3 or HLA-DRB4.

Supplementary Note 4: HLA-Resolve haplotype block recovery logic

Low variant density can interfere with read-based phasing and haplotype block assignment. When no single haplotype block fully spans the antigen-recognition sequence (ARS) of an HLA gene, HLA-Resolve attempts to recover as much information as possible by evaluating the distribution of heterozygous genotypes along the gene model. These edge cases are hypothetical and did not occur in our data, but could arise when (a) there are overlapping haploblocks, but none completely span the ARS, or (b) there are no overlapping haplotype blocks and more than 1 unphased heterozygous genotype. For example, if one of a sample's HLA genes contained unphased heterozygous genotypes in introns/UTRs only, with the exons/CDS fully homozygous, accurate coding sequences could still be reconstructed. Furthermore, if unphased heterozygous sites occurred only in noncoding or non-ARS exons, an accurate ARS could still be reconstructed and used for HLA typing at G group resolution. To implement this rescue, heterozygous genotypes are counted separately across all CDS exons and across the subset that falls within the ARS. Depending on the distribution of heterozygous genotypes, HLA-Resolve does one of the following (documented in the HLA-Resolve GitHub under hla_resolve/docs/technical_reference.md):

a. If there are zero CDS heterozygous genotypes, the raw concatenated exon vcf2fasta output can be used, and the largest ARS-overlapping interval containing <= 1 heterozygous genotype is extracted from the vcf2fasta full-gene output.
b. If exactly one heterozygous genotype is present across the entire CDS, the raw concatenated exon vcf2fasta output can be used, and a new interval is defined relative to the location of the heterozygous CDS genotype as (prev_het + 1) to (next_het -1). If this interval includes the full ARS, it is extracted from the vcf2fasta full-gene output and used in matching. If the interval does not fully span the full ARS, allele matching stops after

coding sequence matching (three-field resolution), and the haplotype does not receive a 4th field assignment.

c. If more than one CDS heterozygous genotype exists, but at most one falls within the ARS CDS exons, only the ARS CDS subsequence is extracted, and the genomic sequence is extended outward from the ARS boundaries to the nearest flanking heterozygous genotypes. For HLA class I genes under this condition, an additional check is performed for heterozygous genotypes in intron 2 (the intron separating the two ARS exons), and the genomic sequence is not reported if any are found. If more than one heterozygous genotype falls within the ARS CDS exons, the gene is excluded from typing.

Supplementary Note 5: Performance of HLA-Resolve's tiebreaking chain

To confirm that four-field concordance was not inflated by the alphabetical tiebreaker defaulting to the :01 allele, we examined how allele assignments were distributed across the tiebreaking chain and stratified concordance by fourth-field value. Across all assignments (including both HPRC and IHWG samples), 65% were resolved by mismatch identity, 8% by match length, and 27% by alphabetical order. Among the 162 IHWG truth alleles at four-field resolution used in star allele benchmarking, only 21% (34/162) were resolved alphabetically. Concordance remained high even when the truth allele's 4th field was not :01 (43/60, 72%), compared to 88/102 (86%) when it was :01. Among these truth alleles, :01 was the fourth field 63% of the time (102/162), which is expected given that the most common alleles tend to be discovered and named first.

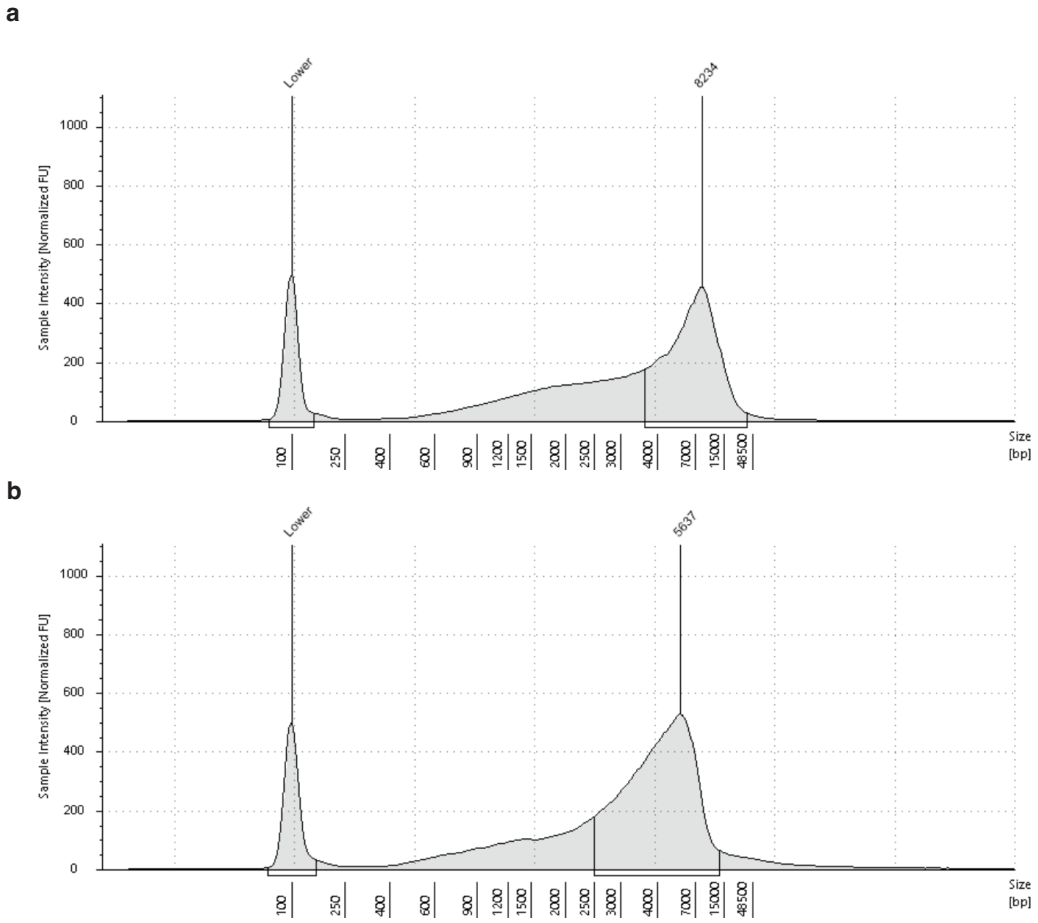
Supplementary Note 6: Resolution of Fred Hutch - IPD-IMGT/HLA conflicts

- IHW09122 HLA-C: IPD-IMGT/HLA shows C*07:02, C*02:02:02, while Fred Hutch as C*07:02:01:03, C*15:02:01:01. We chose the Hutch typings because they are at higher resolution and supported by all the typing tools used in the benchmarking analysis.
- IHW09125 HLA-C: Both the IPD-IMGT/HLA and Fred Hutch show C*01. However, all typing algorithms tested, including our own, showed a heterozygous HLA type C*01/C*08 in both the PacBio and ONT data. Therefore, we edited the truth labels to C*01/C*08, but left the call at one-field resolution.
- IHW09359 HLA-DRB1: Fred Hutch has missing DRB1 HLA types, while IPD-IMGT/HLA shows DRB1*04 (homozygous). All HLA typing algorithms tested supported this call on both sequencing platforms, so we updated our truth set to include DRB1*04, DRB1*04.

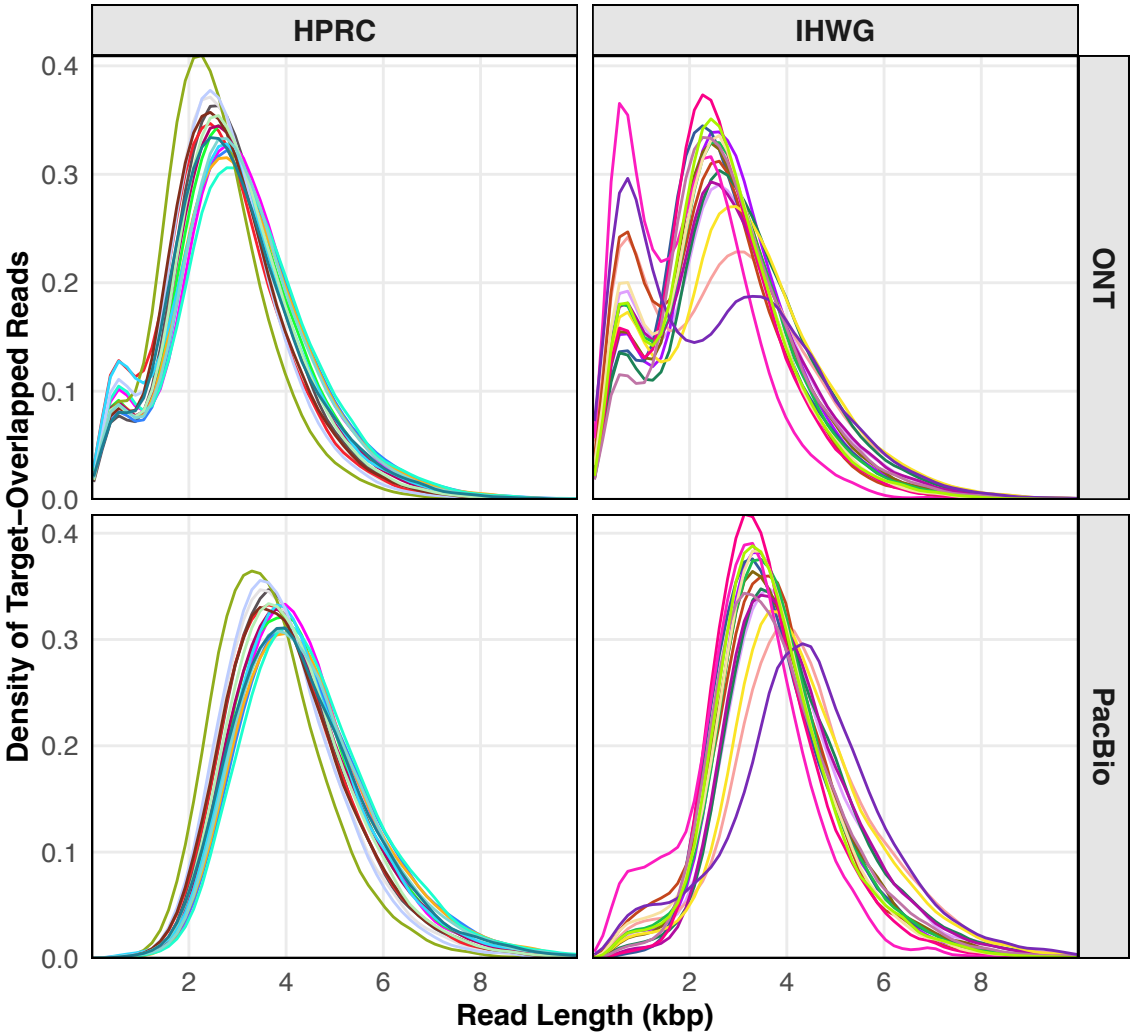
For the remaining IPD-Hutch conflicts, we chose the star allele call made more recently or at higher resolution, which was likely to be typed more recently.

- IHW09198 HLA-DQB1: IPD shows DQB1*03:01:01, DQB1*05:02:01 while the Hutch shows DQB1*03:01:01:01/DQB1*03:01:01:03_NewIntronicVariant_hybridx1, DQB1*05:02:01
- IHW09224 HLA-DQA1: IPD shows DQA1*01:01:01:07, DQA1*01:02:01:05, while the Hutch shows DQA1*01:01:01:01/DQA1*01:01:01:02/DQA1*01:01:01:03, DQA1*01:02:01:01/DQA1*01:02:01:03/DQA1*01:02:01:05. Here, we chose the IPD typings, as they appeared updated more recently.
- IHW09224 HLA-DQB1: IPD tools shows DQB1*05:01:24:01, DQB1*05:02:01:01, while the Hutch shows DQB1*05:01:01:03, DQB1*05:02:01. Here, we chose the IPD typings, as they appeared updated more recently.
- IHW09251 HLA-DPB1: IPD shows DPB1*02:01:02, DPB1*04:02 while the Hutch shows DPB1*02:01:02|DPB1*416:01, DPB1*04:02:01:02|DPB1*105:01.

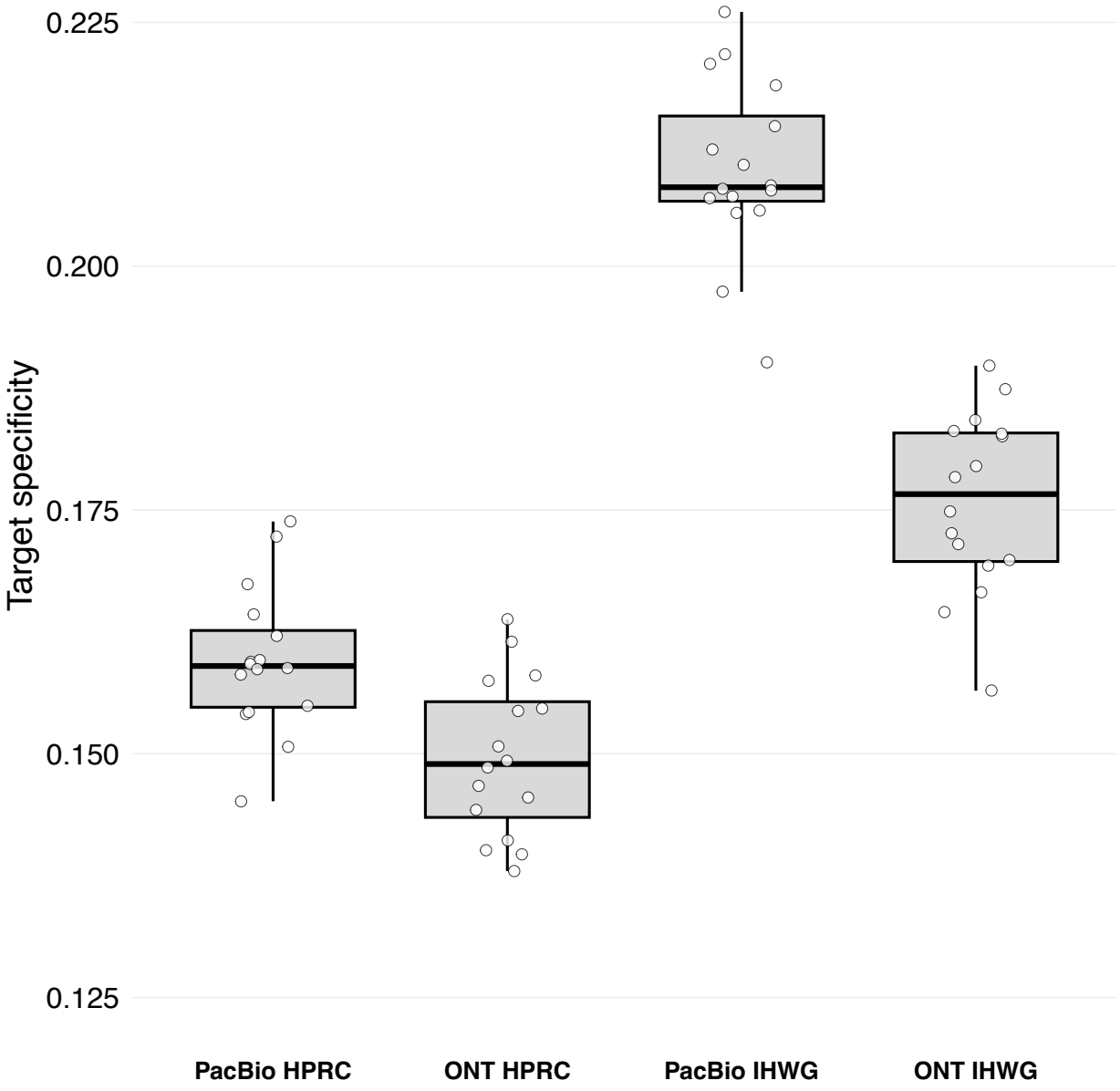
Supplementary Figures



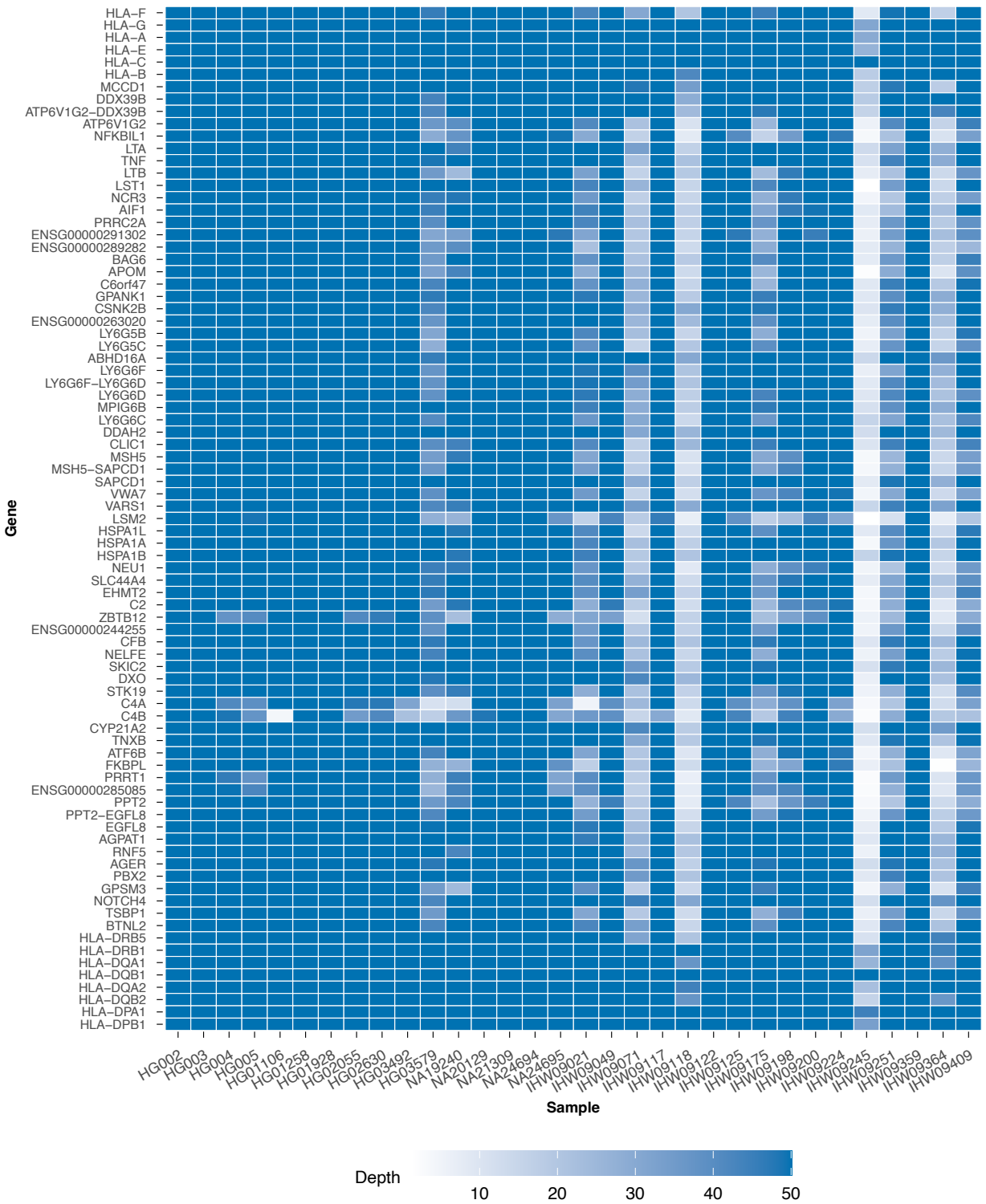
Supplementary Figure 1: **Distribution of DNA fragment lengths**. Genomic DNA ScreenTape after tagmentation, before hybrid capture (**a**) and after hybrid capture and PCR amplification (**b**).



Supplementary Figure 2: **Distribution of sequencing read lengths for target-overlapped reads**. Read-length distributions (in kilobases) for target-overlapped reads, stratified by sample cohort (HPRC, IHWG) and sequencing platform (ONT, PacBio). Each curve represents an individual sample.

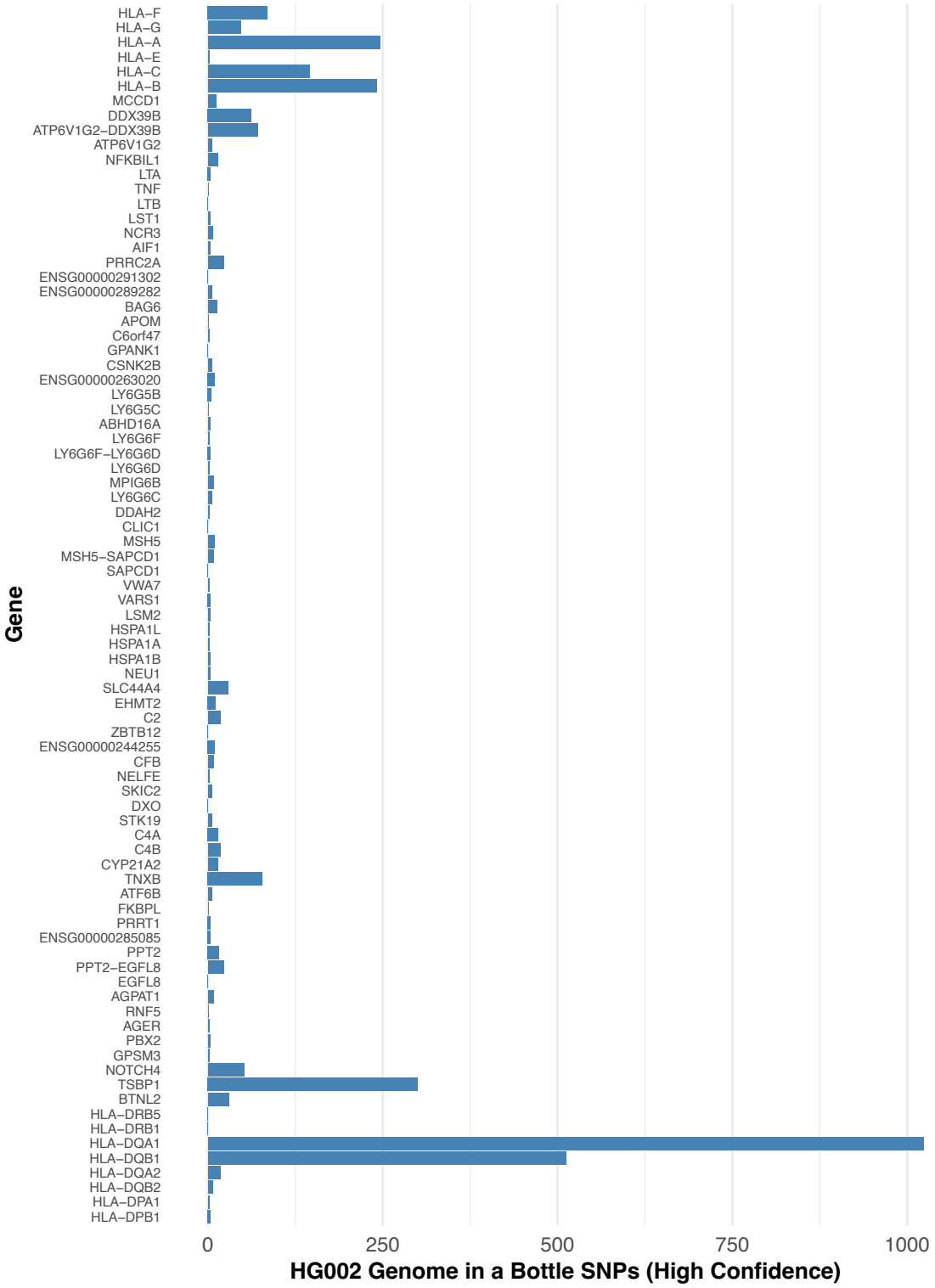


Supplementary Figure 3: Target specificity by hybrid capture experiment and sequencing platform. Target specificity was calculated as the proportion of on-target aligned bases. Points represent individual samples.

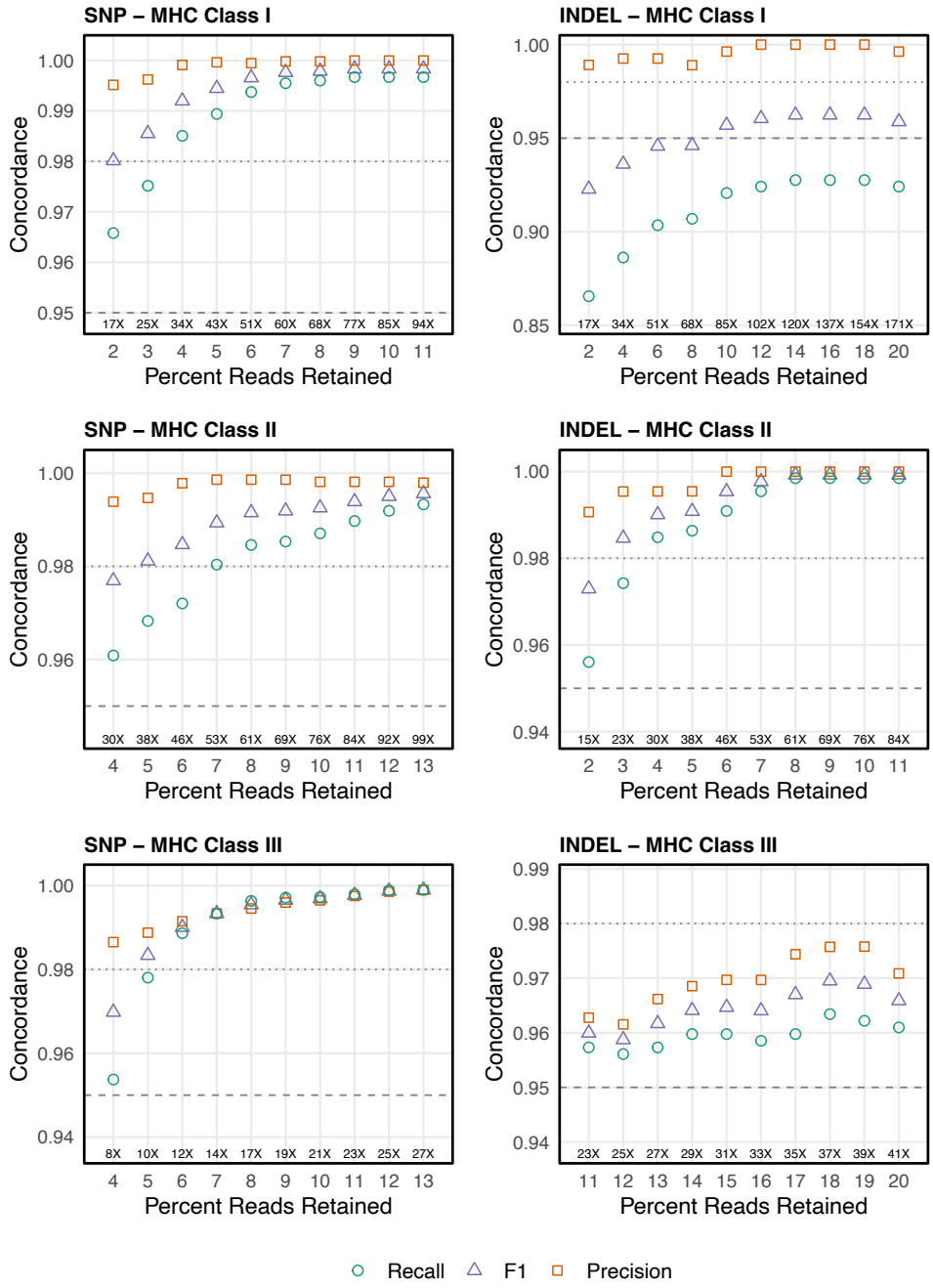


Supplementary Figure 4: Gene-level PacBio Revio coverage across the captured HLA region. Heatmap showing mean sequencing depth for each captured protein-coding gene in the HLA region and each sample. Genes are ordered by genomic position along chromosome 6.

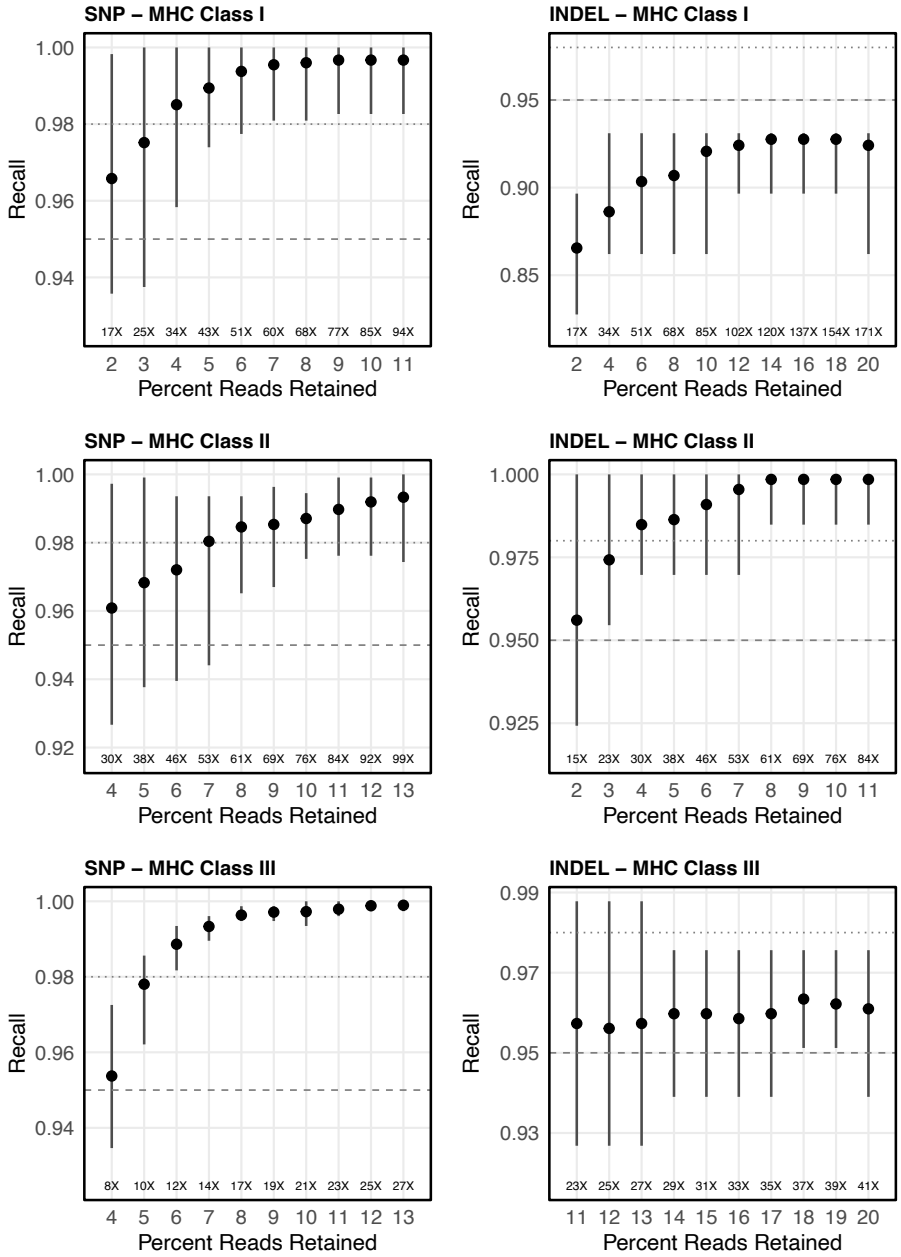
Supplementary Figure 5: Gene-level ONT PromethION coverage across the captured HLA region. Heatmap showing mean sequencing depth for each captured protein-coding gene in the HLA region and each sample. Genes are ordered by genomic position along chromosome 6.



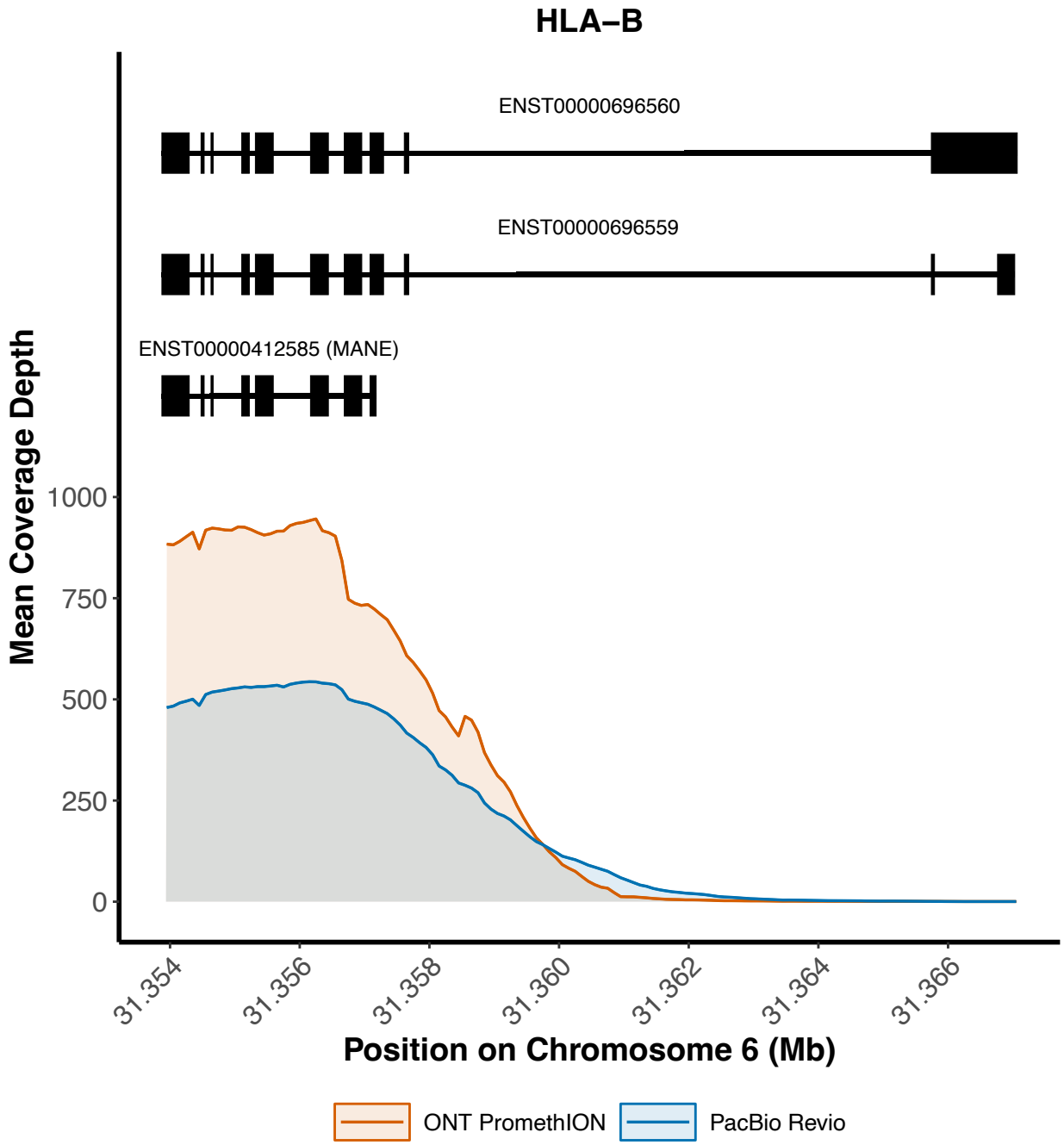
Supplementary Figure 6: Distribution of HG002 Genome in a Bottle SNPs by HLA gene.



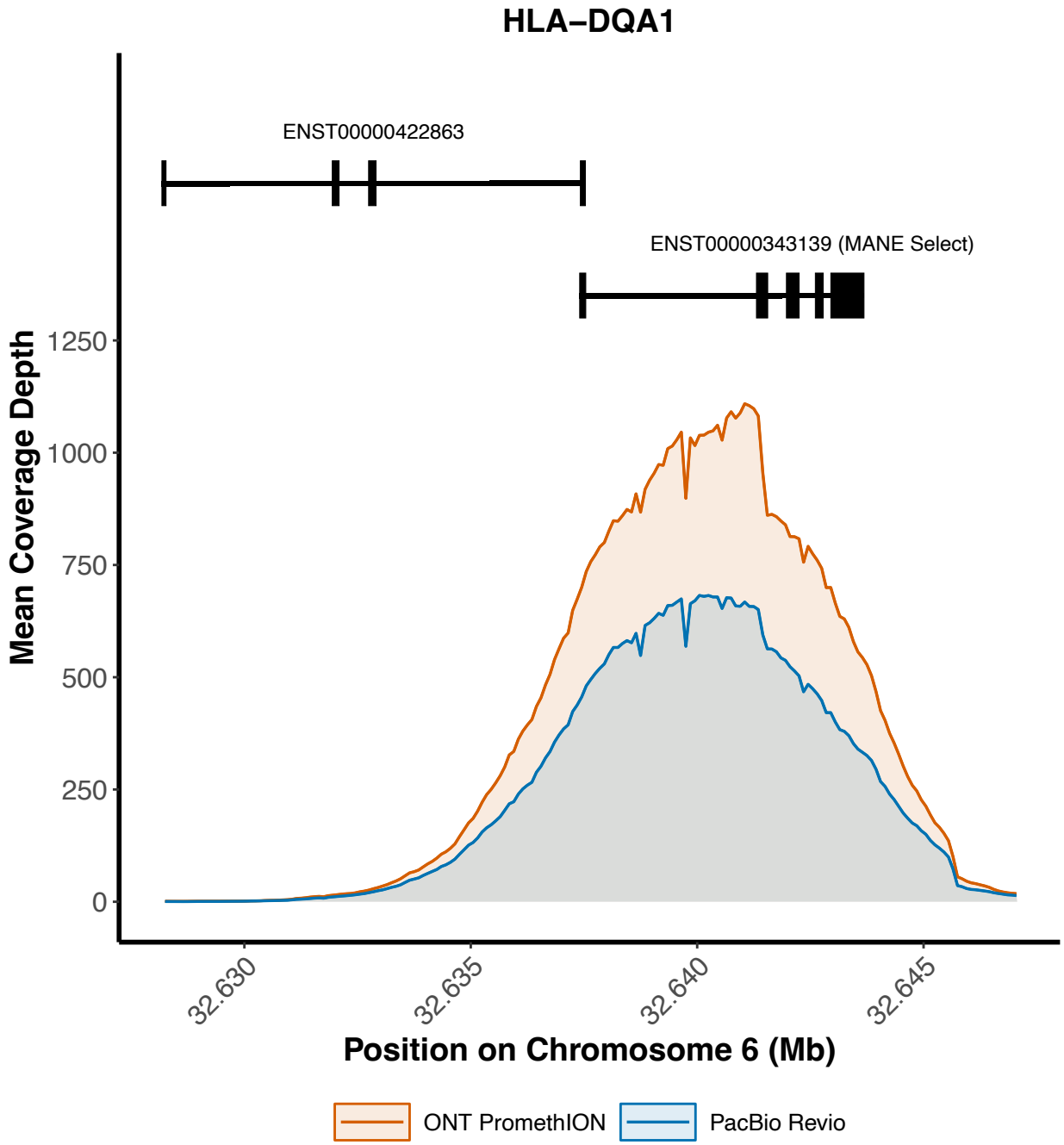
Supplementary Figure 7: Effect of downsampling on genotype concordance. Recall, precision, and F1 for single-nucleotide variants (left) and indels (right) are shown for the HLA Class I (top), Class II (middle), and Class III (bottom) regions across progressive levels of read downsampling. Metrics were computed from PacBio data after random read subsampling and genotyping with DeepVariant v1.6.1. The x-axis shows the resulting mean on-target coverage depth. Genotype concordance with the Genome in a Bottle benchmark was assessed for variants within captured protein-coding genes using ten independent downsampling replicates. Points indicate the mean metric value across replicates.



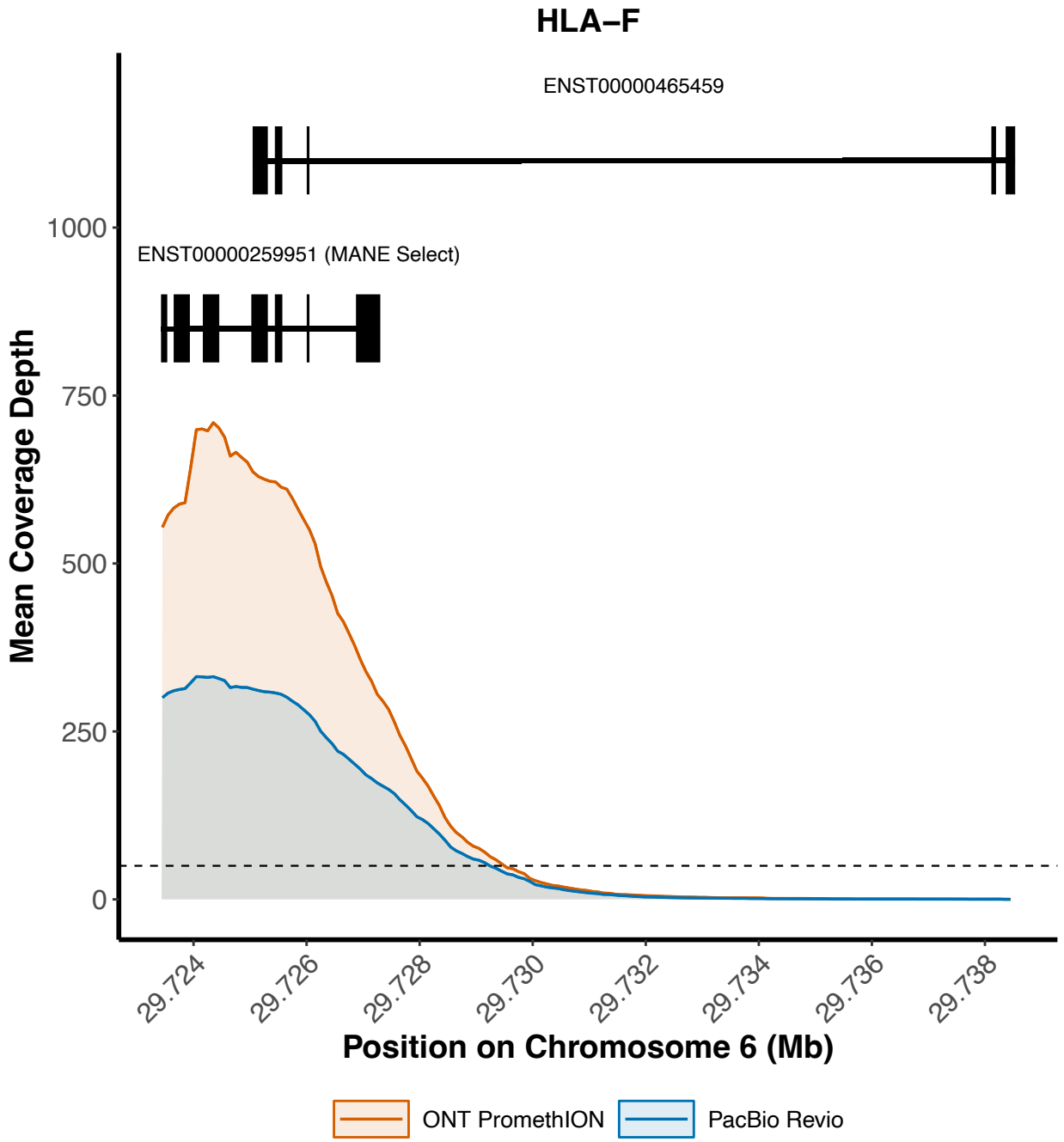
Supplementary Figure 8: Effect of downsampling on genotype recall. Recall for single-nucleotide variants (left) and indels (right) is shown for the HLA Class I (top), Class II (middle), and Class III (bottom) regions across increasing levels of read downsampling. Metrics were computed from PacBio data following random read subsampling and genotyping with DeepVariant v1.6.1. The x-axis shows the resulting mean on-target coverage depth (×). Genotype concordance with the Genome in a Bottle benchmark was evaluated for genotypes within captured protein-coding genes across ten independent downsampling replicates. Points denote the mean recall across replicates, and vertical bars indicate the observed minimum-maximum range.



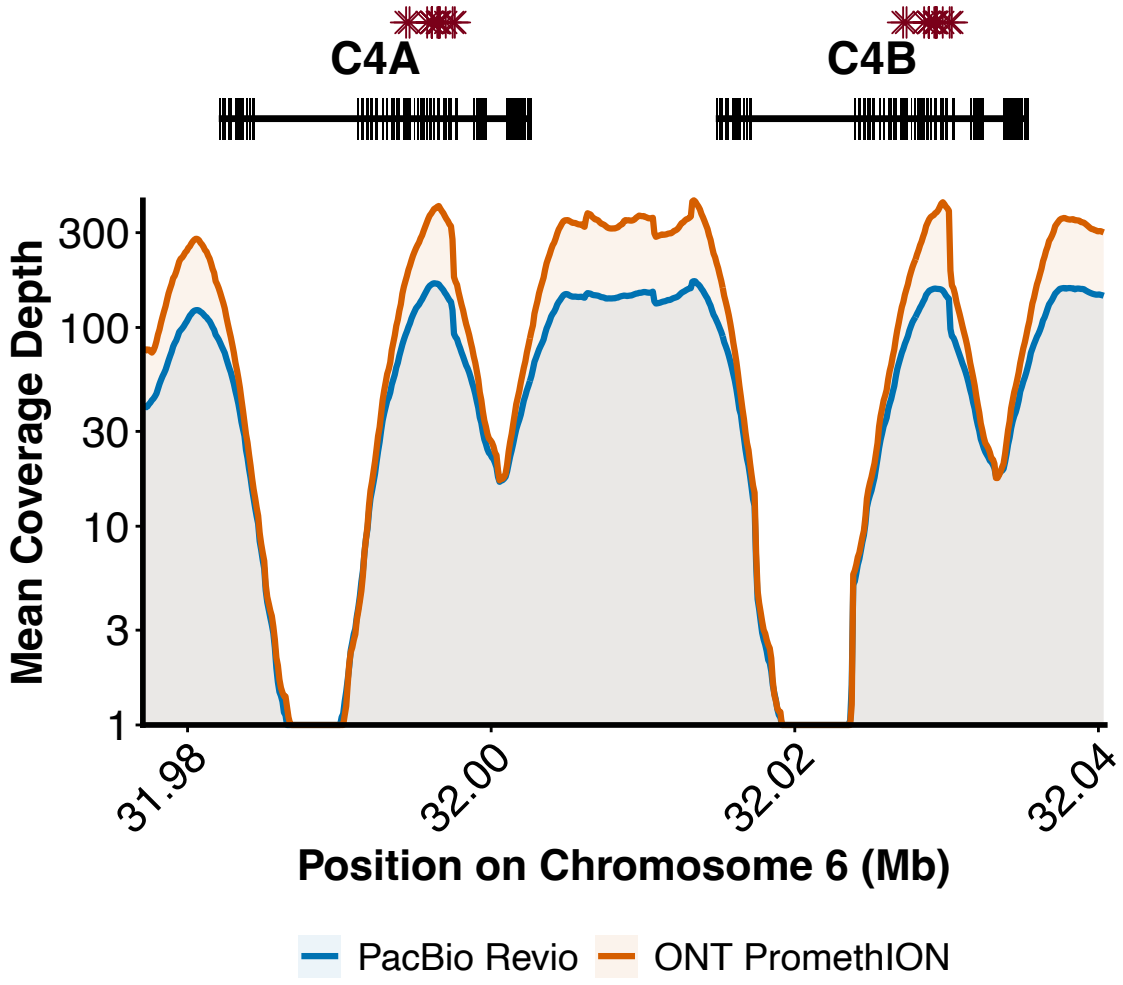
Supplementary Figure 9: Average coverage depth across HLA-B. Mean coverage depth for HLA-B is shown for PacBio Revio (blue) and ONT PromethION (orange), using per-base coverage averaged across samples and aggregated into non-overlapping 100 bp windows. Three annotated HLA-B isoforms are shown, with horizontal line segments indicating transcript boundaries and rectangles denoting exons.



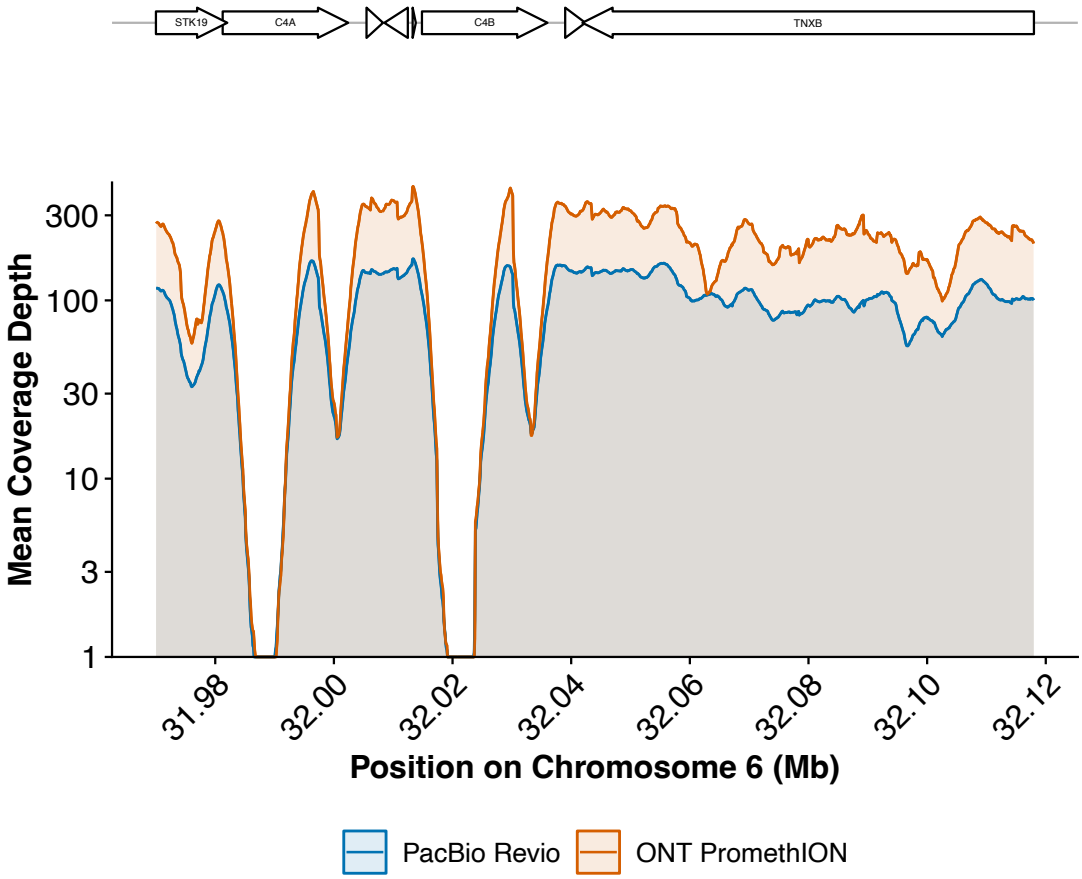
Supplementary Figure 10: **Average coverage depth across HLA-DQA1.** Mean coverage depth for HLA-DQA1 is shown for PacBio Revio (blue) and ONT PromethION (orange), using per-base coverage averaged across samples and aggregated into non-overlapping 100 bp windows. Two annotated HLA-DQA1 isoforms are shown, with horizontal line segments indicating transcript boundaries and rectangles denoting exons.



Supplementary Figure 11: Average coverage depth across HLA-F. Mean coverage depth for HLA-F is shown for PacBio Revio (blue) and ONT PromethION (orange), using per-base coverage averaged across samples and aggregated into non-overlapping 100 bp windows. The horizontal dashed line is placed at 50X coverage depth. Two annotated HLA-F isoforms are shown, with horizontal line segments indicating transcript boundaries and rectangles denoting exons.



Supplementary Figure 12: Average coverage depth across C4A and C4B. Mean coverage depth (log scale) across C4A and C4B is shown for PacBio Revio (blue) and ONT PromethION (orange), using per-base coverage averaged across samples and aggregated into non-overlapping 100 bp windows. Nucleotide positions that differ between C4A and C4B in the GRCh38 human reference genome are highlighted with red asterisks.



Supplementary Figure 13: Average coverage depth across the RCCX locus. Mean coverage depth (log scale) across the GRCh38 RCCX locus is shown for PacBio Revio (blue) and ONT PromethION (orange), using per-base coverage averaged across samples and aggregated into non-overlapping 100 bp windows.