

RESEARCH ARTICLE

Using *de novo* assembly to identify structural variation of eight complex immune system gene regions

Jia-Yuan Zhang^{1*}, Hannah Roberts¹, David S. C. Flores¹, Antony J. Cutler¹, Andrew C. Brown¹, Justin P. Whalley¹, Olga Mielczarek¹, David Buck¹, Helen Lockstone¹, Barbara Xella², Karen Oliver³, Craig Corton³, Emma Betteridge³, Rachael Bashford-Rogers¹, Julian C. Knight^{1‡}, John A. Todd^{1‡}, Gavin Band^{1‡*}

1 Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **2** MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, United Kingdom, **3** Wellcome Sanger Institute, Hinxton, Cambridge, United Kingdom

‡ These authors are joint senior authors on this work.

* jzhang@well.ox.ac.uk (JYZ); gavin.band@well.ox.ac.uk (GB)



OPEN ACCESS

Citation: Zhang J-Y, Roberts H, Flores DSC, Cutler AJ, Brown AC, Whalley JP, et al. (2021) Using *de novo* assembly to identify structural variation of eight complex immune system gene regions. PLoS Comput Biol 17(8): e1009254. <https://doi.org/10.1371/journal.pcbi.1009254>

Editor: Aakrosh Ratan, University of Virginia, UNITED STATES

Received: February 4, 2021

Accepted: July 6, 2021

Published: August 3, 2021

Copyright: © 2021 Zhang et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: A complete list of data generated in this study are available through the European Genome-phenome Archive (EGAS00001005046).

Funding: The work of JYZ and JAT was supported by the Juvenile Diabetes Research Fund [5-SRA-2015-130-A-N], [4-SRA-2017-473-A-N]; the Wellcome [107212/Z/15/Z]; [203141/Z/16/Z]. Computation used the Oxford Biomedical Research Computing (BMRC) facility and was enabled by the Wellcome Trust Core Award Grant Number

Abstract

Driven by the necessity to survive environmental pathogens, the human immune system has evolved exceptional diversity and plasticity, to which several factors contribute including inheritable structural polymorphism of the underlying genes. Characterizing this variation is challenging due to the complexity of these loci, which contain extensive regions of paralogy, segmental duplication and high copy-number repeats, but recent progress in long-read sequencing and optical mapping techniques suggests this problem may now be tractable. Here we assess this by using long-read sequencing platforms from PacBio and Oxford Nanopore, supplemented with short-read sequencing and Bionano optical mapping, to sequence DNA extracted from CD14⁺ monocytes and peripheral blood mononuclear cells from a single European individual identified as HV31. We use this data to build a *de novo* assembly of eight genomic regions encoding four key components of the immune system, namely the human leukocyte antigen, immunoglobulins, T cell receptors, and killer-cell immunoglobulin-like receptors. Validation of our assembly using k-mer based and alignment approaches suggests that it has high accuracy, with estimated base-level error rates below 1 in 10 kb, although we identify a small number of remaining structural errors. We use the assembly to identify heterozygous and homozygous structural variation in comparison to GRCh38. Despite analyzing only a single individual, we find multiple large structural variants affecting core genes at all three immunoglobulin regions and at two of the three T cell receptor regions. Several of these variants are not accurately callable using current algorithms, implying that further methodological improvements are needed. Our results demonstrate that assessing haplotype variation in these regions is possible given sufficiently accurate long-read and associated data. Continued reductions in the cost of these technologies will enable application of these methods to larger samples and provide a broader catalogue of germline structural variation at these loci, an important step toward making these regions accessible to large-scale genetic association studies.

203141/Z/16/Z. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. JYZ was supported by the China Scholarship Council-University of Oxford Scholarship. GB is a member of the MalariaGEN resource centre, supported by Wellcome [204911/Z/16/Z]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: JAT is a member of the GSK Human Genetics Advisory Board.

Author summary

The human immune system is incredibly versatile underlying its capacity to defend the body against thousands of pathogens. At a molecular level, it recognizes pathogens using large libraries of antibodies and related protein receptors. These molecules are encoded by gene families that are particularly difficult to analyze due to their unusually complex patterns of similarities and differences between genes and individuals. To overcome this, we applied several sequencing methods to DNA from a single individual and developed methods to reconstruct the underlying sequence at eight of the immune-associated regions. Importantly, we used DNA extracted from monocytes to avoid capturing the further rearrangements that occur in active immune cells. We generated accurate assemblies by integrating multiple complementary data types, although we noted a small subset of locations that remain challenging. Moreover, we found that this individual contains multiple structural differences between the two inherited chromosomes and compared to previously analyzed genomes, affecting the copy number of immune system genes. Application of these methods in larger numbers of individuals will clearly uncover much more variation than is currently known, and might lead to new understanding of the effect of genetic variation on the broad range of human diseases determined by the immune response.

Introduction

The capability of the human immune system to respond to environmental pathogens results from its substantial diversity and variability, both among individuals within a population and among cells within a single host. Key components of the innate and adaptive immune system, including the human leukocyte antigen (HLA), immunoglobulins (IG), T cell receptors (TCR) and killer-cell immunoglobulin-like receptors (KIR), have evolved exceptional complexity in their genomic loci, featuring numerous highly similar genes interspersed with pseudogenes and repetitive elements. Variation in genes encoding some of these components have well-established associations with infectious, immune-mediated, and other disease traits. The major histocompatibility complex (MHC) encoding HLA is so far the best-studied example, with hundreds of associations now known across multiple classes of disease [1,2] including infections [3,4]. In some cases the underlying functional mechanisms have also been identified [5,6]. However, despite the clearly important role of immunoglobulins (IG), TCR and KIR [7–9], the underlying complexity of these genomic regions has so far prevented a full analysis of their contribution to human disease.

Three challenges must be overcome to make these regions accessible to future studies. First, key aspects of adaptive immunity are driven by somatic recombination and hypermutation of TCR and IG genes in immune cells. Consequently, DNA from non-recombining cell populations is needed to access germline genetic variation in these regions; these are not targeted in current surveys of haplotype variation based on lymphoblastoid cell lines or whole blood [10–12]. Second, extensive paralogy makes these regions intractable to short-read sequencing approaches [13], although analyses based on known immunogenetic sequences can be achieved [14]. Approaches using more costly long-read sequencing must therefore be employed [15,16], though even these methods are not always sufficient to solve the most complex regions [17]. Third, even if these technical challenges can be dealt with, the high diversity observed at these regions presents further difficulties for methods that identify, catalogue, and

genotype the underlying variation. Solving these challenges would in principle enable the development of large haplotype variation reference panels at these loci, complementing existing immunogenetic variation databases [18] and opening them to analysis in large disease association studies.

Motivated by these challenges, here we utilize genomic data from a single individual (HV31) to assemble eight regions that encode key components of the human adaptive and innate immune response. To achieve this, we use DNA extracted from CD14⁺ monocytes, which do not undergo systematic somatic recombination. We develop a pipeline that exploits PacBio HiFi long-read sequencing, Bionano optical mapping, and short-read sequencing data to produce high-quality *de novo* assemblies of these regions. We then use additional long-read and short-read datasets to assess assembly accuracy and to call heterozygous variations using computational approaches based on read alignment and the copy number distribution of short k-mers (i.e. short DNA fragments of fixed length *k*). We find that HV31 carries substantial structural differences between haplotypes and in comparison to the GRCh38 reference sequence, including multiple large variants that affect core immune system genes but are not accurately called by current methods, and we investigate several of these in detail. Lastly, we analyze four gaps in the GRCh38 reference sequence at the immunoglobulin κ and T cell receptor γ regions, that are fully or partially filled in our assembly.

Results

Immune system loci display a spectrum of complexity in the human reference sequence

We focused on eight genome regions that encode components of the human immune system, namely those encoding the HLA, immunoglobulins (IGH, IGL, IGK), T cell receptors (TRA, TRB, TRD, TRG), and the killer-cell immunoglobulin-like receptors (KIR) (Table 1). Regions were defined based on NCBI RefSeq locus definitions [19] (except HLA and KIR which were based on previously published gene ranges [20,21]), plus an additional 1Mb flanking sequence added to both sides (see Methods). In the IGK region we additionally expanded the range to include a ~1 Mb heterochromatin gap present in GRCh38. The expanded regions range from 2–6 Mb in length and vary considerably in terms of repetitive structure and haplotype diversity (Table 1). We noted the least reference sequence complexity in the T cell receptor α , δ and γ regions (which contain < 2% repeat sequence and no listed alternate haplotypes), but greater complexity in other regions. In particular, the regions encoding immunoglobulin subunits contain the highest levels of duplication; previous analyses [22,23] have demonstrated significant structural diversity among known haplotypes in these regions. GRCh38 also contains dozens of alternative haplotype sequences at the HLA and KIR regions, and four gaps in the IGK and TRB regions (Table 1). Comparison to the earlier GRCh37 assembly and the presence of fix patches highlights that these regions are likely to be challenging to assemble.

Assembling immune system regions with long-read, short-read and optical mapping data

We assessed whether the eight selected regions can be accurately assembled *de novo* using data from a single individual identified here as HV31. HV31 was recruited as a healthy volunteer and identified as having European ancestry. To facilitate accurate assembling of these complex regions, we generated data from multiple complementary platforms. Specifically, we performed PacBio Sequel II circular consensus sequencing (obtaining 12.3× genome coverage by ~12 kb HiFi reads), MGI short-read sequencing (56.8×) and Bionano Saphyr Direct Label and

Table 1. Overview of eight selected immune system loci in GRCh38.

Name	Acronym	Coordinates and length	# core genes	% repetitive ^a	% SD ^b	# Gaps ^c	# Alternate haplotypes	% Novel to GRCh38 ^d	# Fix patches ^e
Immunoglobulin heavy chain	IGH	chr14 104,586,437–107,043,718 (2.46Mb)	164	6.8	31.1 (0.5)	0	2	44.7	0
Immunoglobulin κ	IGK	chr2 87,857,361–91,902,511 (4.05Mb)	84	22.7	44.8 (22.7)	3	2	31.5	1
Immunoglobulin λ	IGL	chr22 21,026,076–23,922,913 (2.90Mb)	89	7.4	34.0 (15.3)	0	3	47.4	0
Human leukocyte antigen	HLA	chr6 28,602,238–34,409,896 (5.81Mb)	39	2.7	6.5 (1.1)	0	8	0	0
T cell receptor α and δ	TRA	chr14 20,621,904–23,552,132 (2.93Mb)	115	1.9	3.5 (0)	0	0	37.6	0
T cell receptor β	TRB	chr7 141,299,011–143,813,287 (2.51Mb)	78	5.3	19.5 (9.0)	1	2	34.2	1
T cell receptor γ	TRG	chr7 37,240,024–39,368,055 (2.13Mb)	22	1.3	3.1 (0)	0	0	0.2	0
Killer cell immunoglobulin-like receptors	KIR	chr19 53,724,447–55,867,209 (2.14Mb)	10	4.7	12.9 (0)	0	50	47.4	0

^a The proportion of repetitive DNA calculated as the proportion of 31-mers that are repeated at least once.

^b The percentage of the region that is annotated as lying in a segmental duplication or (in brackets) a highly identical ($\geq 95\%$) segmental duplication.

^c The number of gaps (sequences of 'N' bases) in GRCh38.

^d The percentage length of contigs that are new to GRCh38, i.e. not carried forward from GRCh37.

^e The number of fix patches intersecting the locus in GRCh38 patch release 13.

<https://doi.org/10.1371/journal.pcbi.1009254.t001>

Stain (DLS) optical mapping (152.7× coverage by imaged molecules). In addition, long-read and short-read sequencing data from PacBio continuous long read (CLR; 35×), Oxford Nanopore Technologies (ONT) PromethION (63×), 10x Genomics linked-reads (40.2×), Illumina Novaseq PCR-free (44.2×), MGI single-tube long fragment read (stLFR) (51.3×) and MGI CoolMPS (56.9×) platforms were also generated from the same blood sample (S1 Table). To minimize the impact of cell-specific events including V(D)J recombination and somatic hypermutation and enable accurate assembly of the germline genome, the long-read data used for our assembly below were collected from CD14⁺ monocytes isolated from peripheral blood mononuclear cells (PBMCs) with antibody-conjugated beads (see Methods). Bionano imaging data, as well as several of our short-read datasets, were generated directly from PBMCs (S1 Table). We note that coverage by PBMC-derived reads drops significantly around T cell receptor genes (S1 Fig), consistent with an effect of V(D)J recombination in T cells which are the predominant cell type in PBMC. Datasets generated in this study are further detailed in S1 Table and have been deposited with the European Genome-Phenome Archive (Data Availability).

To generate an accurate representation of the eight regions in the HV31 genome, we developed an assembly pipeline consisting of four stages (Fig 1A). This approach deals with heterozygosity by producing a consensus assembly of each region, and a list of heterozygous

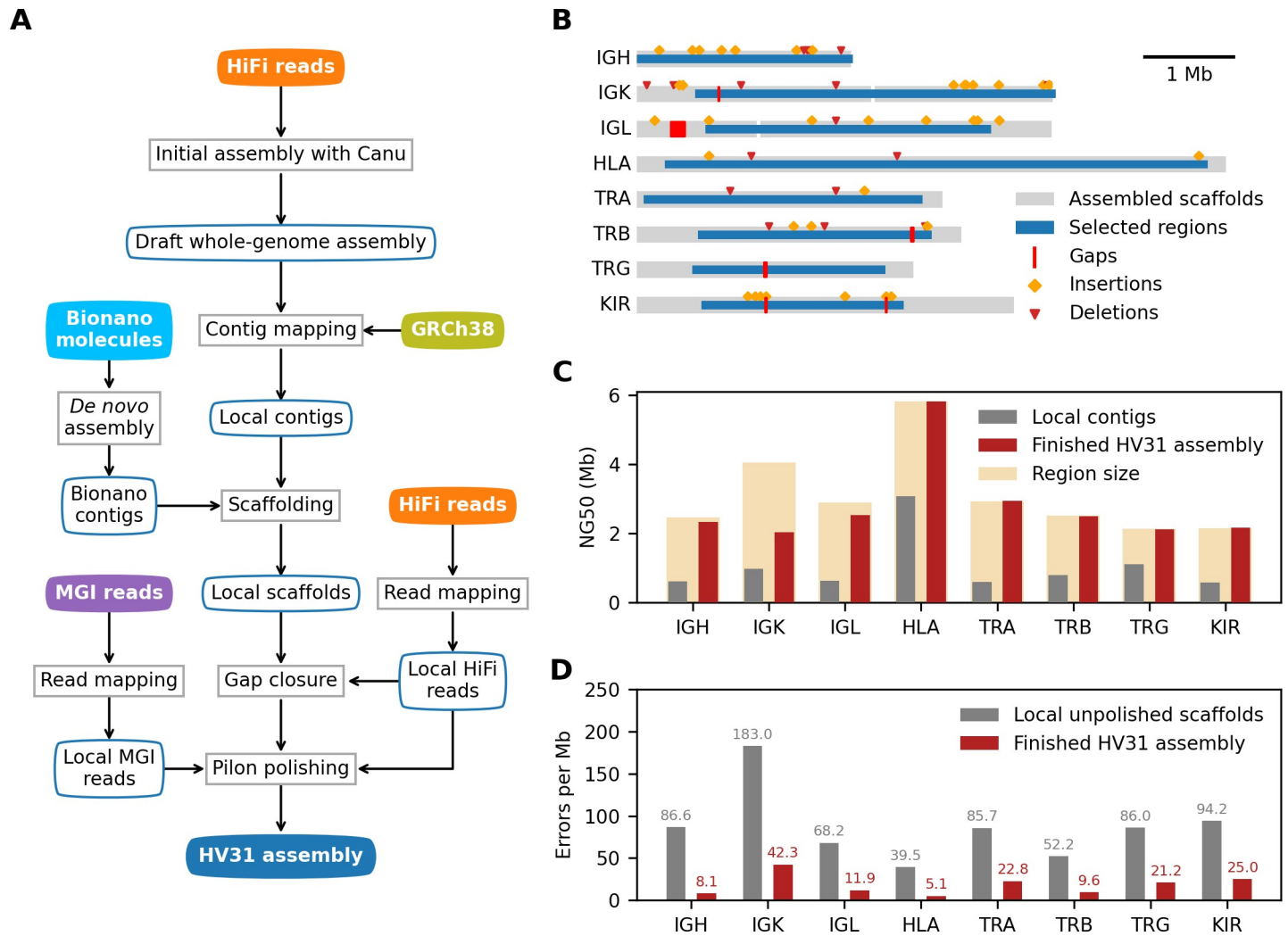


Fig 1. Evaluation of HV31 *de novo* assemblies. (A) Diagram of *de novo* assembly workflow. Processes and datasets are represented by blocks with square and rounded corners, respectively. (B) Overview of assembled scaffolds in 8 selected regions. Heterozygous SVs on the unassembled haplotype that are larger than 1 kb in size are shown as orange diamonds or red triangles. Note that the assembled scaffolds (gray) are often larger than the originally selected immune system regions (blue) defined in Table 1. (C) Contig/scaffold continuity (NG50, y axis) for local contigs (gray) and finished HV31 assembly scaffolds (red) in each region (x axis). NG50 is defined as the length of the longest contig/scaffold that, along with longer contigs/scaffolds, covers 50% percent of each locus, as determined by alignment to GRCh38. The size of the selected region on the GRCh38 reference is also shown. To ensure comparable results, for each contig/scaffold, only the length within region boundaries is taken into NG50 calculation. (D) The estimated number of errors per megabase in each region, before and after assembly polishing. Error rates are estimated using a modified version of the Merqury algorithm [24] as described in Methods.

<https://doi.org/10.1371/journal.pcbi.1009254.g001>

structural variants (SVs; Fig 1B and S1 Dataset) that jointly describe the HV31 genome. We describe the steps of this assembly process and a comparison to other approaches below.

Initial assembly. We used the Canu assembler [25] applied to HiFi reads to produce a draft whole-genome assembly. We aligned the resulting contigs to GRCh38 and extracted all contigs that overlap with the predefined regions of interest, hereafter referred to as local contigs, for further processing. Local contigs were highly fragmented (Figs 1C and S2), reflecting the unusual genomic complexity in these regions. The assembly also contained multiple shorter contigs (referred to as “haplotigs” below) aligning to the same location as longer contigs in some regions, which either represent assembly errors or genuine differences between haplotypes (S2 Fig).

Scaffolding. We next used the local contigs with Bionano optical imaging data to produce longer continuous scaffolds. Imaged DNA molecules had an observed mean length of 149 kb, substantially longer than reads from other datasets involved in this study (S3 Fig). We assembled these molecules using the proprietary Bionano Access software. As expected, the resulting contigs tended to be substantially longer than those in the draft whole-genome assembly (S4 Fig). We used the Bionano Solve algorithm to align the local contigs to the Bionano-assembled contigs and implemented a modified version of the BiSCoT algorithm [26] (Methods) to order and orient the local contigs accordingly. This process also removes or merges in haplotigs that can be effectively aligned to the hybrid scaffolds. Finally, we confirmed that the remaining haplotigs represented substantial duplication of scaffolded contigs using a k-mer based method (Methods), and removed these from downstream analysis. The scaffolds generated by this process fully covered six of the eight regions with a single scaffold, while the IGL and IGK regions were assembled with two scaffolds each (Fig 1B and 1C).

Gap filling and polishing. We further improved the assembly quality by carrying out a gap-closing step (which fills in nucleotide information for missing bases between adjacent contigs in a scaffold) using TGS-GapCloser [27] applied to local HiFi reads, resulting in the closing of seven gaps. We also implemented a polishing step using Pilon [28] applied to local HiFi and MGI reads, correcting erroneous bases in the assembly that likely originate from sequencing errors. To avoid bias due to read selection, for both processes we selected relevant reads using a double-alignment process that first aligns all reads to the initial whole-genome assembly, and then realigns the subset of reads mapping to local contigs to the fully scaffolded assembly (referred to as locally aligned reads below; see Methods). This process left six gaps (i.e. sequences of 'N' bases) in the HV31 scaffolds (Fig 1B); these lie outside regions aligning to core immune system genes but could potentially be improved with additional processing.

Structural variant calling. We used the available long-read data to call heterozygous SVs using the HV31 assembly as reference (Fig 1B). In brief, SVs were called separately from locally aligned HiFi, CLR and ONT long reads using PBSV (for HiFi and CLR) and Sniffles (for HiFi, CLR and ONT). A computational approach based on unique k-mers [29] was used to refine read alignment before variant calling (see Methods). Across the eight regions, 1,366 SVs were reported by PBSV or Sniffles, 491 of which were jointly supported by two or more dataset-software combinations (S5 Fig and S1 Dataset; including 179 >100 bp and 23 >1 kb in length), as reported by SVanalyzer [30], which analyzes the sequence information of each variant and identifies groups of compatible variants. As these numbers indicate, we observed considerable discrepancy between the individual SV calling approaches (S6 Fig), reflecting the difficulty of aligning reads and calling SVs in paralogous regions. As a comparison point, we also created a dataset of SV calls based on 10x Genomics linked-read sequencing data using the Long Ranger pipeline (S2 Dataset). We compare these structural variant calling results with the HV31 assembly further below. We also note that six gaps remain in the HV31 assembly (Fig 1B), which lie outside regions aligning to core immune system genes but could potentially be improved with additional processing.

We refer to the polished assembly scaffolds and SV dataset generated by these steps as "the HV31 assembly" hereafter; the assembly is summarized in Fig 1 and compared to the GRCh38 reference in Fig 2 and to other published assemblies in S7 Fig. Benchmarking information of the assembly pipeline is available as S3 Dataset. As we describe below, the per-base error rate of these assembled regions is on the order of 5–50 errors per Mb (Fig 1D), which is of a similar magnitude to recently published whole-genome assemblies based on HiFi data [17,31], although some structural errors do remain. The HV31 assembly is also comparable in coverage and continuity to other assemblies (S7 Fig), including to recently published high-quality assemblies of the homologous CHM13 cell line [29,32].

Structural variation revealed by comparison to GRCh38

We used k-mer sharing plots (i.e. “dot plots” [33]; S8 Fig) to compare HV31 to the GRCh38 reference sequence (Fig 2). Each point in these plots represents a short sequence of length k (here $k = 50$) that is shared by both the reference sequence and the HV31 assembly; the observed pattern of points therefore provides a visualization of similarities and differences between the two assemblies. This comparison further suggests that the HV31 assembly is relatively complete for the eight regions, without apparent missing sequence (apart from the six gaps mentioned above) or chimera sequences. HV31 contains two scaffold breaks at the IGK and IGL loci; both are located near long (≥ 100 kb) SDs that are highly identical ($\geq 99\%$) and indicate that this type of SD remains challenging for current assembly methods. In contrast, genomic loci with higher proportions of shorter, low-similarity SDs such as the HLA and KIR were completely resolved in the HV31 assembly.

Close inspection of these plots (S4 Dataset) reveals many large (≥ 1 kb) SVs that differ between GRCh38 and the HV31 primary assembly. To systematically characterize these SVs, we aligned the assembly to GRCh38 and applied Assemblytics [34]. Assemblytics reported 145 SVs, 55 of which were ≥ 1 kb in size (Fig 1B and S5 Dataset). The majority (65.5%) of the reported SVs involved expansions or contractions of repeat elements, while the rest were insertions or deletions of unique sequences. The KIR region harbors the highest number (29) of SVs, followed by IGH (28) and IGK (24) regions.

Validation of assembly accuracy using unassembled sequencing reads

Given that HV31 differs structurally from GRCh38, an important question is how the structure of our assembly of these regions can be confirmed (or conversely how any remaining errors can be identified) without reliance on a reference sequence. Motivated by previous work [24,35], we adopted an approach based on computing the multiplicity of each assembly k-mer in a validation dataset, which we here take as the set of sequence reads including PacBio HiFi reads and all short reads from MGI, 10x and Illumina platforms (S1 Table), which were chosen due to their relatively low error rates (S9 Fig). This dataset has over 150 \times coverage of k-mers appearing in both copies of the genome (Fig 3A), and is sufficiently high-coverage that heterozygous k-mers, and k-mers in higher repeat numbers, can be separated from homozygous k-mers (Fig 3A and 3B). Under the assumption that sequence reads are approximately uniform across the genome, the multiplicity of each k-mer in the validation dataset should be proportional to its copy number in the HV31 assembly [36]; any discrepancies therefore indicate heterozygous variation or assembly errors.

To leverage this, we plotted validation multiplicity in comparison to scaffold multiplicity across all regions (S10 Fig; illustrated for part of the IGH region in Fig 3C and 3D). The assembly and validation dataset are in generally good agreement at most informative k-mers, (i.e. relatively few large regions show validation k-mer multiplicity systematically incompatible with assembly multiplicity), including across several repetitive regions with assembly copy number systematically greater than 1. We used a similar approach to estimate base-level errors, which we identified as clusters of k-mers with validation multiplicity < 5 (using $k = 22$ as recommended previously [24]; this approach is similar to but more stringent in practice than the previously published Merqury method [24]; S2 Table and Methods). These errors are relatively sparse; on average across regions the error rate was 18.1 per 1 Mb sequence (based on 3.16×10^7 k-mers, of which 8786 (0.0278%) were deemed erroneous; Fig 1D; improved from 85.3 per Mb prior to polishing). However, a number of locations show larger discrepancies between assembly and validation data (numbered regions in Panel B in S10 Fig; S3 Table). We examined these in detail and found that many reflect heterozygous structural variants

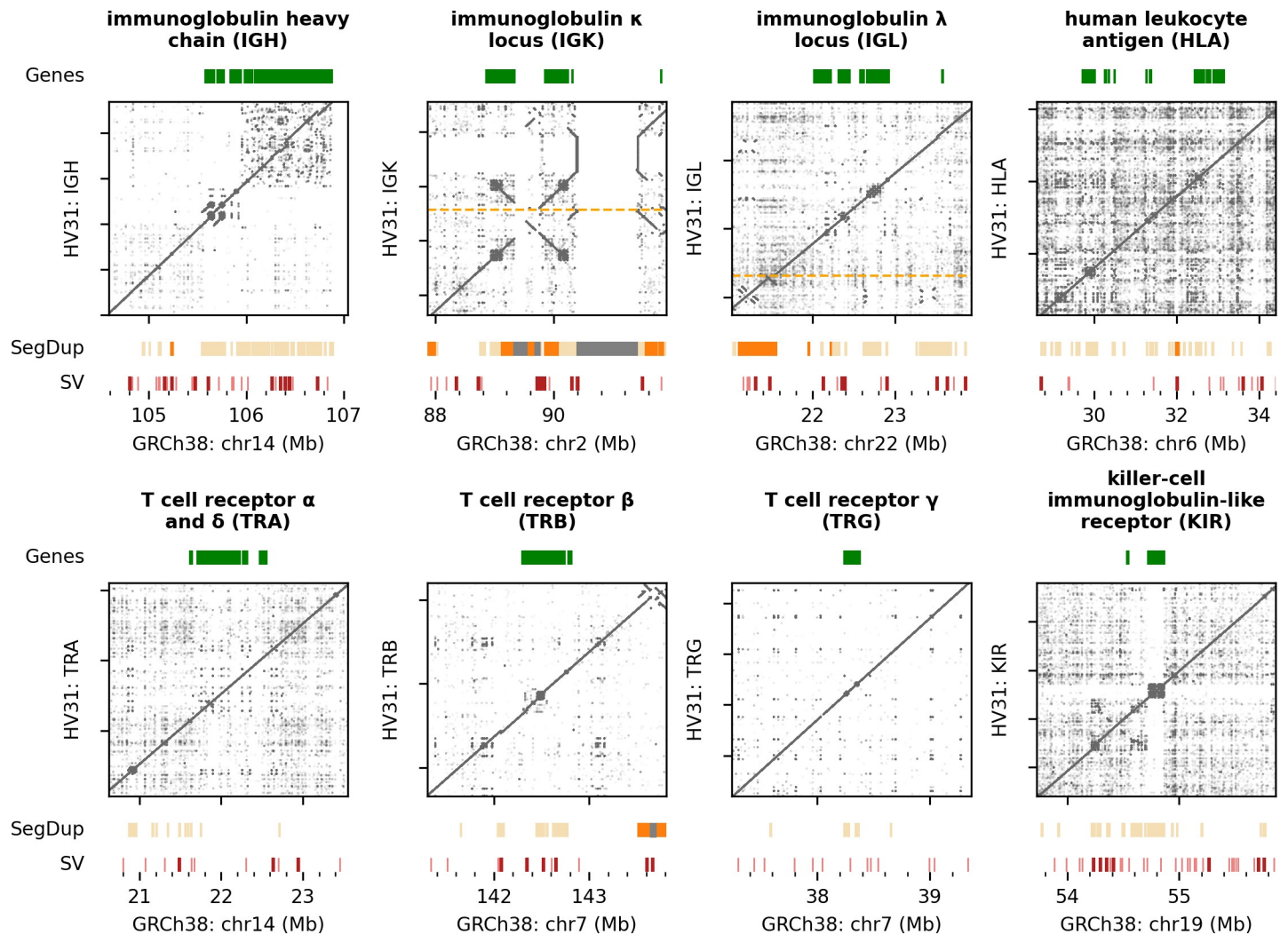


Fig 2. Comparing the HV31 assembly with GRCh38. Main panels show dot plots indicating locations of shared k-mers ($k = 50$) on the GRCh38 reference (x axis) and on the HV31 assembly (y axis) for each locus studied. Multiple scaffolds are separated with orange horizontal dashed lines. Plots are annotated as follows: Gene, core genes of each locus. SegDup, segmental duplications defined as sequence fragments that are ≥ 1 kb in length and $\geq 90\%$ identical to another fragment (segmental duplications with identity $\geq 99\%$ are highlighted in orange; reference gaps are shown in gray). SV, structural variants detected in the HV31 assembly relative to GRCh38. Structural variants larger than 1 kb are highlighted in dark red.

<https://doi.org/10.1371/journal.pcbi.1009254.g002>

(illustrated in Panel A in [S11 Fig](#) for a heterozygous deletion, and in [Fig 3D](#) for a complex heterozygous duplication that we discuss further below) as well as the aforementioned assembly gaps. However, a small subset of these locations indicate possible structural errors in our assembly. These include a ~ 30 kb duplication in the HLA region that we confirmed is incorrectly assembled in a single copy (i.e. “collapsed”) in HV31 (Panel B in [S11 Fig](#)), as well as three relatively extensive stretches of elevated multiplicity in the IGK and IGL regions where we were unable to fully confirm the assembly structure using the k-mer approach ([S10 Fig](#)). We also implemented a comparison to contigs *de novo* assembled from optical mapping data, which suggested that several of these regions were correctly assembled (Methods). In general, this analysis indicates that our assembly of HV31 is substantially accurate apart from three repeat-rich segments in the IGK, IGL and HLA regions that may still contain errors.

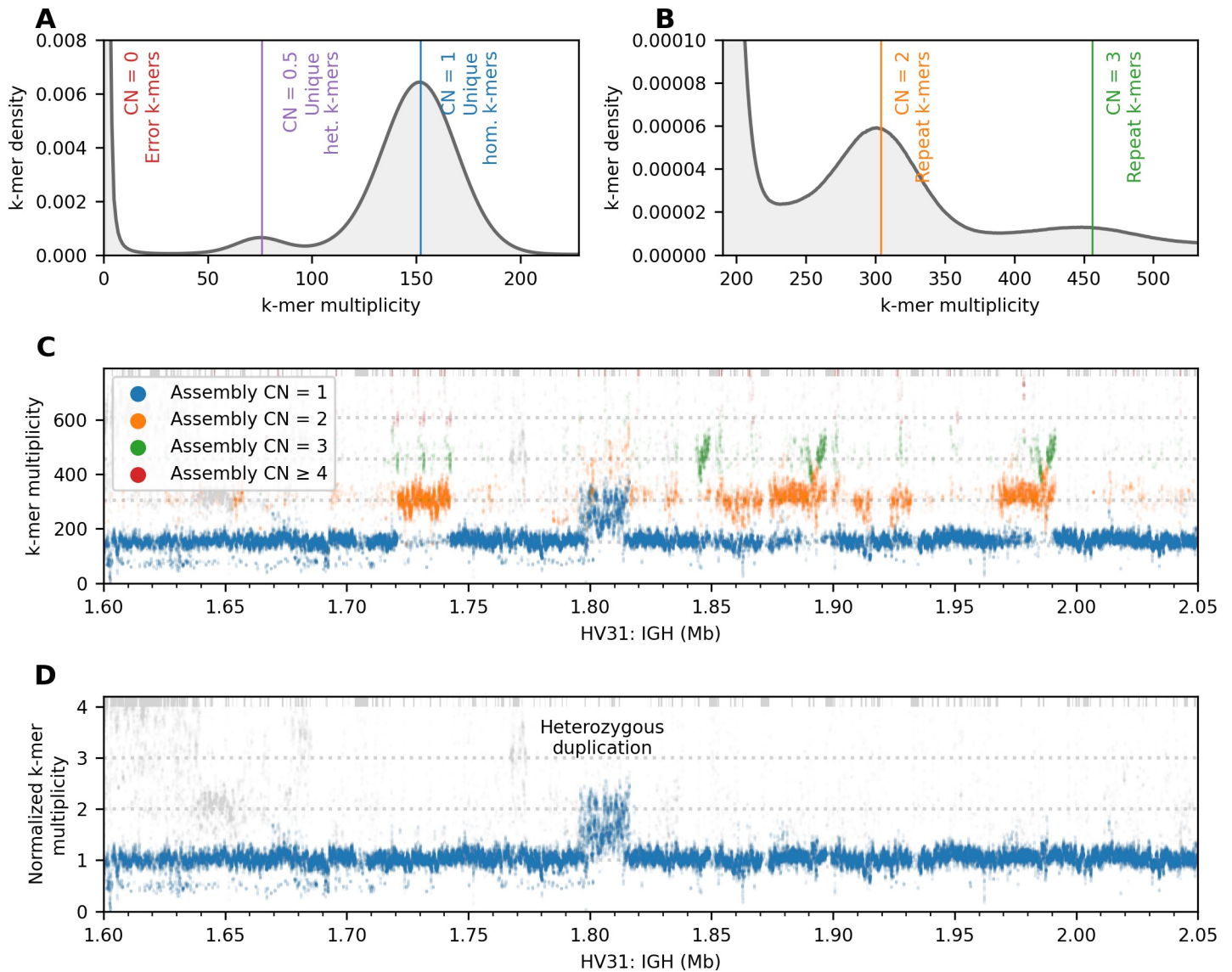


Fig 3. Reference-free assembly validation based on k-mer multiplicity. (A, B) Histogram of k-mer multiplicity ($k = 31$) in unassembled reads in the validation dataset. Vertical lines show locations of distribution peaks; text indicates interpretation of k-mers near these peaks (het., heterozygous; hom., homozygous; CN denotes assumed true copy number of the k-mer in the diploid HV31 genome). (C) k-mer multiplicity (y axis) plotted against k-mer position for a repeat-rich sequence fragment in the IGH region of the HV31 assembly. Green, blue and purple colors denote the k-mer copy number in the HV31 assembly scaffolds. Non-specific k-mers that are also found outside the IGH region are colored gray. (D) validation k-mer multiplicity normalized against assembly copy number for the same sequence fragment as in (C). To make the y axis center on 1, values are further normalized by dividing by the peak multiplicity of unique homozygous k-mers as shown in panel A. The region of discrepancy indicates a complex heterozygous duplication around IGHV3-30 that we discuss further below. In (C) and (D), k-mers with multiplicity beyond the axis limits are stacked at the top of the plots. Panel A and Panel B in S10 Fig show (C) and (D) extended to all eight regions, respectively.

<https://doi.org/10.1371/journal.pcbi.1009254.g003>

We note two issues that impede validation by short k-mers. First, accurate measurement of k-mer multiplicity in highly repetitive regions is challenging; in our assembly this is particularly relevant to three relatively extensive stretches of elevated multiplicity in the IGH and IGL regions where we were unable to fully confirm the assembly structure using the k-mer approach (S10 Fig). Secondly, there is a trend towards a drop in coverage of the validation k-mers in regions encoding T cell receptors (S10 Fig); as discussed above this is due to the use of DNA from PBMCs for some platforms (S1 Fig). An alternative approach based on coverage of

locally aligned nanopore reads (Panel B in [S10 Fig](#)) does not show this drop and confirms the structure of these regions.

HV31 contains diverse complex immune system structural variants

As detailed above, a nontrivial number of large structural variants exist between GRCh38 and HV31 as well as between the two haplotypes of HV31. To assess the impact of these SVs on core genes, we used an alignment process to identify the best-matching allelic variant of each immunoglobulin and T cell receptor variable gene segment, and each HLA and KIR gene within the relevant IMGT or IPD database ([Fig 4](#)). Relative to GRCh38, the HV31 scaffolds contain both insertions and deletions of gene sequence in the IGH, IGK, IGL and TRB regions. It also contains allelic variation in all regions except TRG. We also noted a small number of genes that differ from the best matching IMGT allele, and may represent novel sequence. We note that HV31 gene content also differs substantially from the GRCh37 assembly [[22](#)].

In interpreting these results, some care must be taken because of the consensus nature of the HV31 scaffolds, which do not necessarily represent a single haplotype at each locus. To elucidate underlying genetic variation, we investigated the genetic basis of the observed copy number changes in detail, focusing on the IGH and TRB regions and described in the following sections.

A tandem repeat within a 45kb CNV involving *IGHV1-69* and *IGHV2-70*

Variation in the copy number of *IGHV1-69* and *IGHV2-70* genes has previously been reported [[22](#)]. Both genes are present in two copies in GRCh38. In the HV31 scaffold, we found only one copy of *IGHV1-69* and *IGHV2-70* remaining, as the result of a 45 kb copy number contraction relative to GRCh38 (Figs [4](#) and [5A](#); variant IGH_b_29 in [S1 Dataset](#)). The earlier GRCh37 reference genome shared a similar haplotype in the IGH region, with only one copy of *IGHV1-69* and *IGHV2-70* genes. This haplotype has been suggested to be more common worldwide than the GRCh38 haplotype [[14,22](#)] and comparison to validation k-mers indicates it is homozygous in HV31 ([S10 Fig](#)). We noted that this CNV appears to be effectively callable by aligning reads to GRCh38, e. g. manifesting as a coverage gap in aligned PacBio HiFi reads ([Fig 5B](#)); a co-located deletion was also called by the 10x pipeline, though the endpoints and length appeared inaccurate ([S2 Dataset](#)).

Within this 45 kb CNV, we also noticed a 2.66 kb cluster of tandem repeats with a 59-mer motif (Figs [5A](#) and [S12](#)) that was not correctly assembled in either GRCh37 or GRCh38 (see GenBank: AC245369.4). Similar repeat clusters have also been reported for CHM1 (from which the GRCh38 sequence for IGH region was derived) and NA19240 samples, though the copy numbers of the 59-mer motif varied [[15](#)].

A compound heterozygous CNV involving *IGHV3-30*

A second prominent feature of the IGH region is the loss of one copy of the *IGHV3-30* gene. GRCh38 carries two copies of the *IGHV3-30* gene, namely *IGHV3-30* and *IGHV3-33* [[14](#)]. We find that *IGHV3-33* is removed in the HV31 assembly, together with *IGHV4-31* ([Fig 5C](#)). However, inspection of Bionano contigs covering this region revealed a further unusual feature ([Fig 5D](#)): one of the two contigs covering this region contains a corresponding three-fold copy number expansion. To confirm this, we inspected validation k-mer multiplicity in the surrounding area, and observed elevated multiplicity compatible with a 3-fold expansion on the unassembled haplotype ([Fig 3D](#)). As further confirmation we also searched for reads spanning this region and observed a CLR read consistent with the expanded haplotype ([S13 Fig](#)). These results therefore indicate that the unassembled haplotype carries three copies of the region

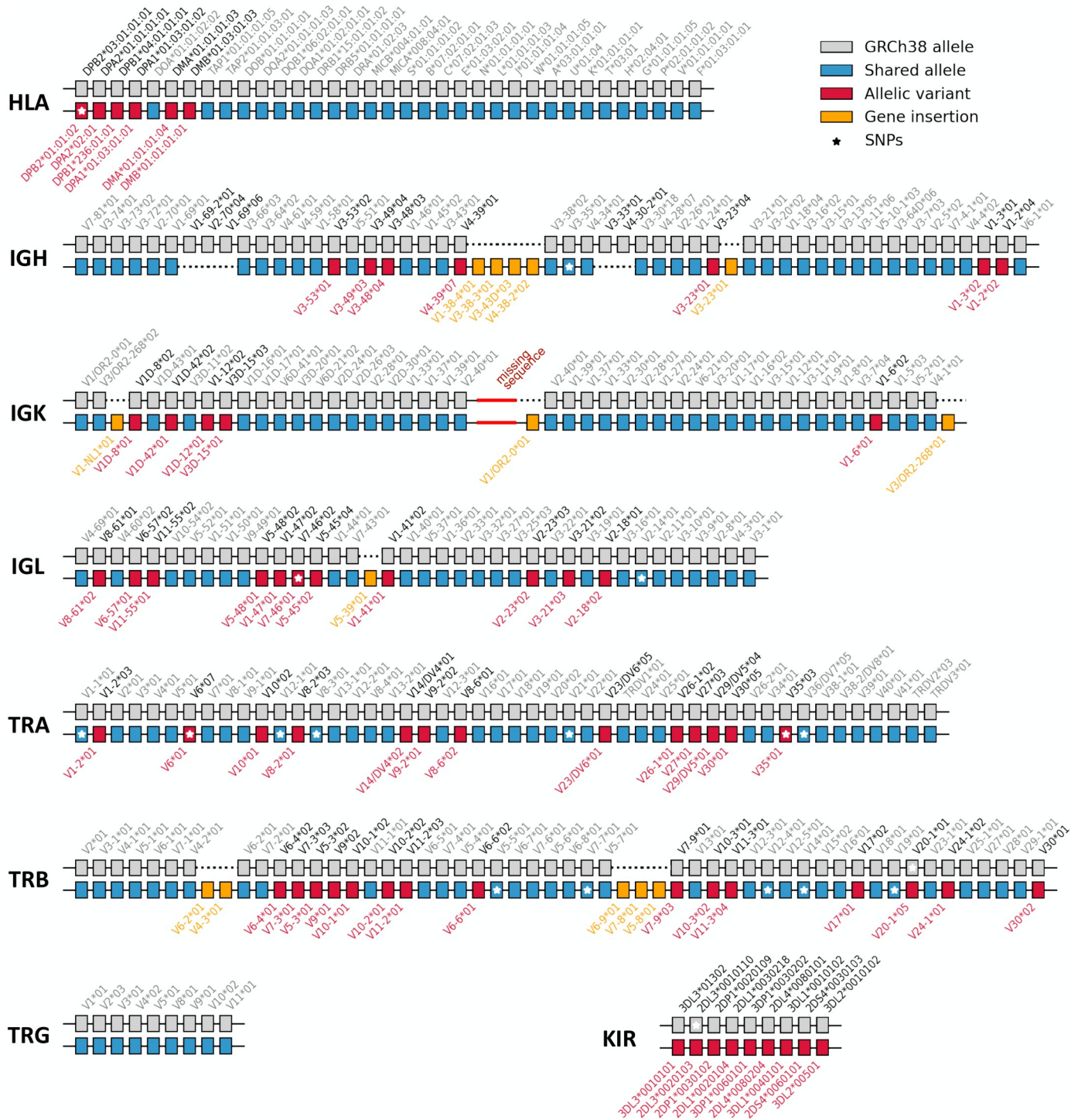


Fig 4. HV31 assembly content of immunoglobulin and T cell receptor variable (V) genes compared with GRCh38. Pseudogenes are not shown. V genes in each region are arranged according to their relative order on the positive-sense strand. Allelic variants refer to genes where the best-matching HV31 allele differs from the GRCh38 allele. Insertions refer to genes in the HV31 assembly that cannot be matched to a GRCh38 gene. Alleles with identical sequences, such as *TRBV6-2*01* and *TRBV6-3*01*, are not distinguished. Alleles that carry additional SNPs compared to the best-matching reference allele are marked with stars. The sequence fragment between IGK proximal and distal clusters that remains not fully resolved is denoted as a red line.

<https://doi.org/10.1371/journal.pcbi.1009254.g004>

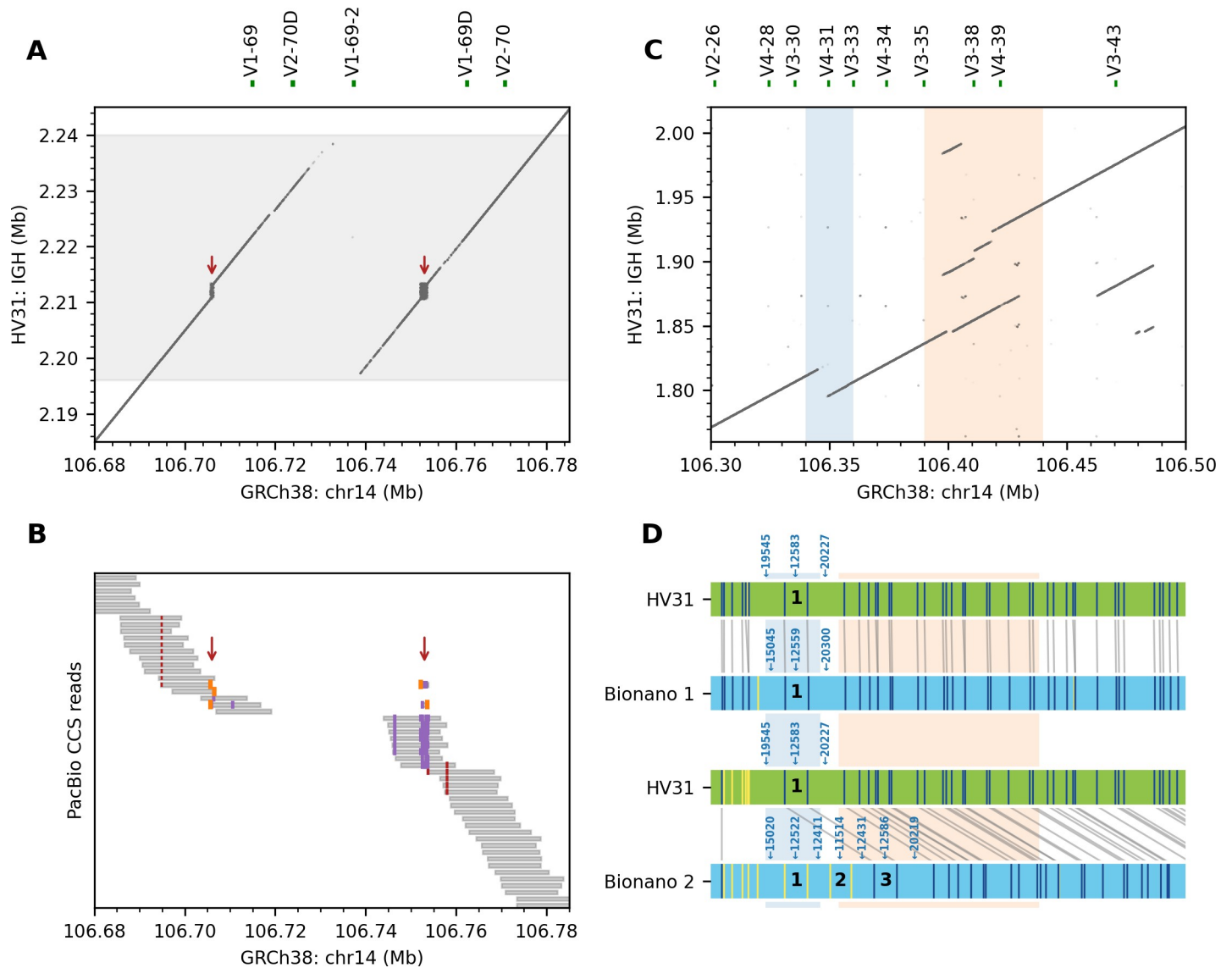


Fig 5. IGHV genes feature highly variable copy numbers. (A) Collapsed repeats (i.e. repeats incorrectly assembled in a single copy; red arrows) in GRCh38 were resolved in the HV31 assembly inside a 45 kb copy number contraction (i.e. reduction in copy number of a repeat unit relative to GRCh38, shaded in grey). (B) Pileup of HiFi reads aligned to GRCh38. The 45 kb CNV was evident from the uncovered central region. The resolved repeats were shown as alignment breakpoints (orange) or insertions (purple) within the aligned segments (gray). Nearby short deletions (red) were also evident. (C) k-mer sharing plot (k = 50) showing complex structural variations found between *IGHV3-21* and *IGHV3-43*, including a 25 kb copy number contraction (i.e. reduction in copy number relative to GRCh38 of a repeat unit; blue) and an 80 kb complex duplication event (orange). (D) Alignment patterns of HV31 assembly and two Bionano contigs covering the region shown in (C). The four rows show the HV31 assembly (green horizontal bars) and two Bionano contigs (blue horizontal bars). Vertical/diagonal grey lines between HV31 and the Bionano contigs show alignments of DLE-1 recognition markers; alignments are computed by the Bionano Solve algorithm based on inter-marker distances. The relevant marker positions are indicated by vertical lines inside the contigs (colored dark blue for aligned markers and yellow for markers that were not aligned). The alignment patterns indicate that Bionano contig 2 contains an expansion consistent with the triplication of the region highlighted in blue. Black numbers indicate the approximate repeat units and relevant inter-marker distances are annotated above the contigs.

<https://doi.org/10.1371/journal.pcbi.1009254.g005>

surrounding *IGHV3-30*, such that HV31 carries both a contraction and expansion of this region, with an overall copy number change that would not be evident from coverage analysis of reads aligned to the reference sequence. This observation is also compatible with previous work [22] which reports this region as a hotspot for SVs, with the diploid copy number of *IGHV3-30* and related genes ranging from zero to six. This CNV was not called accurately by any of the SV calling methods we employed (S1, S2 and S5 Datasets).

An 80 kb complex duplication involving multiple IGHV genes

HV31 carries additional copies of *IGHV1-38*, *IGHV3-43*, *IGHV4-38* and *IGHV3-38* genes compared to GRCh38, that are contained in a ~80 kb duplication with complex structure (Fig 4 and Fig 5C). Inspection of k-mer multiplicity data implies this duplication is homozygous (S10 Fig). However, this duplication was not called by any of the methods we used to call SVs (S1, S2 and S5 Datasets). We interpret this as resulting from difficulty in aligning the two sequences; consistent with this, we observed that the HiFi reads in this region displayed suspicious alignment patterns when mapped to GRCh38, which were improved when mapped to the HV31 assembly (S14 Fig).

Large insertions incorporating novel TRBV genes

In the TRB region, we detected a ~11 kb homozygous insertion near *TRBV6-2* and another ~19 kb insertion near *TRBV5-7* (Fig 6A). Both are supported by Bionano contigs (Fig 6B), and both insertions incorporated sequence fragments that are not found in GRCh38, with limited homology to adjacent sequences (Fig 6A). Comparison to k-mer validation again implies both insertions are homozygous. Assemblytics identified duplications at both locations but with inaccurate length and sequence content (S5 Dataset). The HV31 scaffold was consistent with an alternative contig for the TRB locus included in GRCh38 (RefSeq NG001333.2; S15 Fig). By comparing NG001333.2 with GRCh38, we confirmed that the 11 kb insertion introduced *TRBV4-3* and *TRBV6-2* genes and a *TRBV3-2* pseudogene, while the 19 kb insertion introduced the *TRBV6-9*, *TRBV7-8* and *TRBV5-8* genes (Figs 6 and S15).

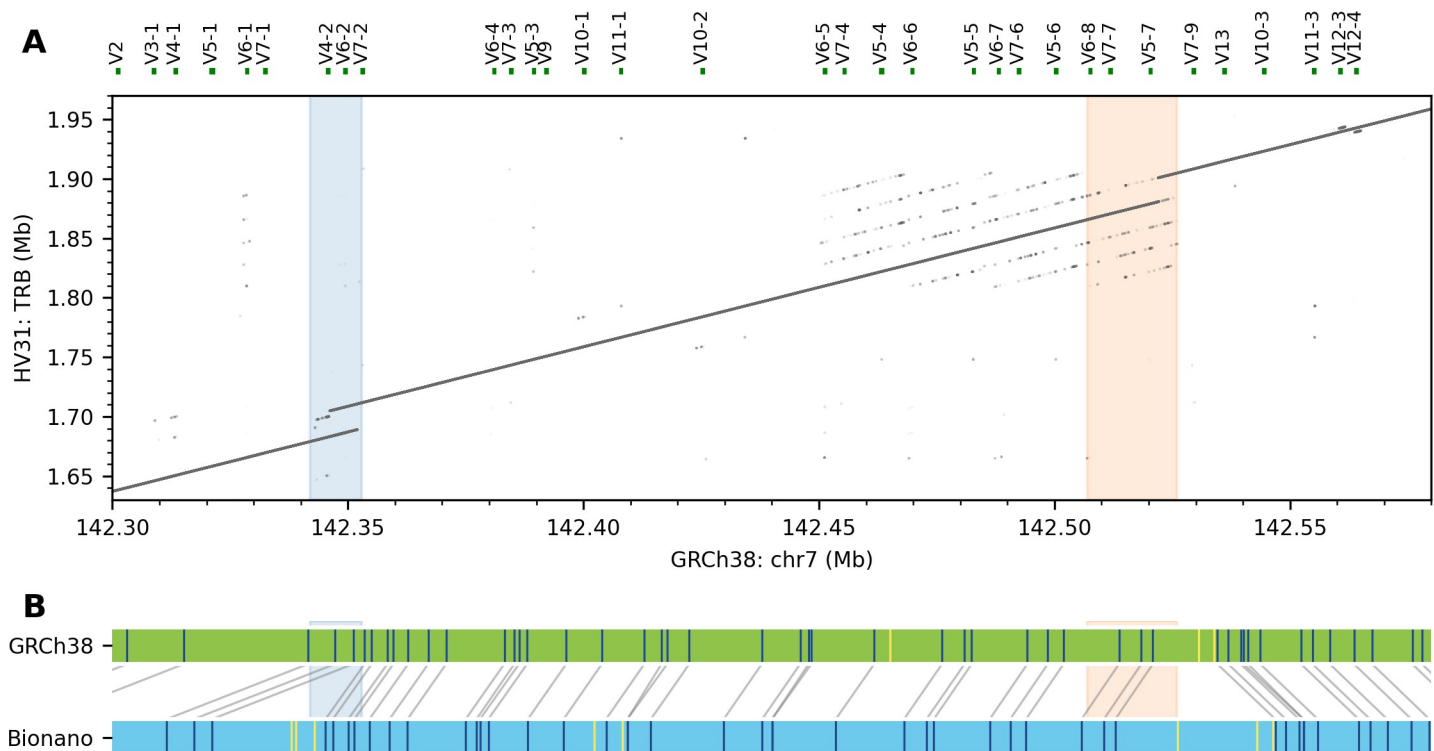


Fig 6. Large insertions in the TRB region. (A) HV31 harbors a 11 kb insertion (blue) and another 19 kb insertion (green) relative to GRCh38. (B) A Bionano contig (blue) aligned to GRCh38 (green), confirming the two insertions in (A).

<https://doi.org/10.1371/journal.pcbi.1009254.g006>

Reference gaps amidst complex segmental duplications resolved in the HV31 assembly

In addition to the complexities engendered by structural variation, gaps in the reference genome constitute another potential impediment to the analysis of genetic variation. Large gaps typically arise due to highly repetitive sequence that is challenging to assemble (e.g. heterochromatin regions that often consist of megabase-scale tandem satellite repeats), and their functional significance remains largely unexplored [37].

We were therefore interested to note that the HV31 assembly closes three large gaps in the GRCh38 assembly, and partly closes a fourth. Three of these gaps lie in the IGH region (S16 Fig), while the fourth lies within a 400 kb region of high-identity segmental duplications [38] ~1Mb downstream of TRB genes (S17 Fig). We here focus on the largest such gap, a ~1 Mb gap annotated as heterochromatin in GRCh38, located between the distal cluster of IGH genes and the centromere of chromosome 2 (S18 Fig). Examination of this gap revealed a ~650 kb sequence assembled as an array of approximately 115 imperfect tandem copies of 6 kb repeat units (Panel A in S19 Fig). Most of the repeat units contain a 22-bp signature sequence (TTCGATTCATTTGATGATTCAT), indicating that the heterochromatin sequence belongs to the human satellite HSat2B family [37].

Notably, the assembled heterochromatin sequence also contains a nonrepetitive sequence fragment within the assembled heterochromatin sequence (Panel B in S18 Fig). This 32 kb fragment appears unique to the region, sharing no significant homology with either the rest of the heterochromatin region nor any part of GRCh38. The heterochromatin sequence does not contain the recognition motif of DLE-1 (CTTAAG) used in generating optical maps, and we were therefore unable to directly confirm the arrangement using Bionano contigs, though a marker corresponding to the 32 kb unique sequence could be identified (Panel C in S18 Fig).

We compared the HV31 assembly with the recently reported T2T CHM13 assembly [32] (GenBank: GCA_009914755.2), where the IGH region is fully reconstructed in one contig. The corresponding heterochromatin sequence in the T2T CHM13 assembly is consistent with the HV31 assembly in terms of total length and repeat unit sequences, though the specific order and orientation of these repeat units differ (S19 Fig). This is of interest because it potentially reflects the structural variability in this heterochromatin region. A similar 32 kb unique sequence is also found in the T2T CHM13 assembly, though at a different location (Panel B in S19 Fig). In addition, we found this 32 kb fragment, along with 76.8 kb flanking sequences, was over 90% identical to a 108.8 kb unplaced sequence (GenBank: AP023554.1; Panel D in S18 Fig) assembled recently from individuals of Japanese ethnicity [39]. Similar 'islands' of unique sequence amid heterochromatin regions have previously been suggested for chromosome Y [40] and chromosome 21 [41].

Eleven heterochromatin gaps remain in GRCh38, with estimated sizes ranging from 20 kb to 30 Mb, and only the largest heterochromatin gap in chromosome X has been resolved in the form of a *de novo* assembly so far [29]. The resolved heterochromatin sequence for chromosome 2 in the HV31 assembly may provide insight into other problematic heterochromatin regions.

Discussion

Genetic regions encoding the human immune response are among the most important for medical science, harboring determinants of both infectious and immune-mediated diseases. Despite this, the complex structure and diverse nature of many of these loci has hindered assessment of their contribution to health in large-scale genetic studies. Resolving this will likely require the creation of reference haplotype variation datasets linked to genome function.

The greatest progress in this direction has been made for inference of HLA alleles [42–45] and to a lesser extent KIR alleles [46] based on large databases of known gene sequences. More recent approaches have used targeted long-read sequencing to characterize immunoglobulin variation [15]. However, much functional genetic variation affecting core immune system genes is still to be discovered and made accessible to larger studies. The decreasing cost and improving performance of long-read sequencing technologies suggests a possible route to this through the *de novo* assembly of representative individual genomes, although computational, cost, and analytic challenges remain in practice.

In this study, we contribute to this program by assembling eight of the most complex immune system regions in a single healthy European individual identified as HV31, who was recruited as part of a larger study of genomic and transcriptomic variation in immune cell types. To do this, we based our assembly on DNA extracted from CD14⁺ monocytes, allowing us to assess the germline haplotype configurations in the immunoglobulin and T cell receptor regions. We exploited accurate PacBio HiFi data for the initial assembly, and used optical mapping data to scaffold the assembled contigs, followed by additional gap closing and polishing steps (Fig 1). Alongside this, we also developed a set of methods to validate assembly correctness using available datasets. Although some structural errors do remain—and might be resolvable with future work based on the breadth of data we have generated here—our results suggest we have produced an essentially correct representation of the regions reported here, with some caveats that we have noted. Our assembly thus adds to growing numbers of reference-quality assemblies that can be utilized in these regions [15,23,47,48] and to the catalogue of known alleles at these loci. The sequencing data and finished assemblies generated in this study have also been deposited at the European Genome-Phenome Archive (Data Availability) for future use.

An ideal description of a diploid sample would involve two fully assembled contigs for every somatic chromosome, each representing one of the two inherited haplotypes. This approach is sought by several emerging experimental and computational approaches [48–51]. Generating a phased assembly inevitably involves a trade-off among cost, phasing accuracy and assembly continuity, which is further complicated by the presence of large high-identity duplications [51]. Given the limited depth of HiFi reads in our data (S1 Table), we instead chose to represent the HV31 genome as a set of consensus (i.e. mixed-haplotype [52]) scaffolds supplemented with a list of heterozygous variants (S1 Dataset). This is a pragmatic approach which also simplifies downstream analysis as the assembly can be directly used in place of a reference genome without additional preparation.

Fully assembling regions containing long, complex repeat structures from shotgun sequencing remains a challenging problem; any approach must somehow distinguish reads coming from different repeat copies but identify those that truly overlap—while allowing for sequencing errors. Exemplifying this challenge, our assembly of the IGK region fills a heterochromatin gap in GRCh38 that largely consists of high copy-number repeats but also contains unique sequence (S18 Fig). Direct confirmation of the assembled structure of IGK is difficult using our data and may require further experimental methods to achieve, but our assembly of this region is compatible with other recently reported assemblies. The fact that most of the repetitive regions within the loci studied here are correctly assembled is of interest in itself, as it implies that the repeats are old enough (or divergent enough) to be effectively distinguished.

Given the high diversity observed in these regions it is unsurprising that HV31 carries structural variants which make it differ from current genome builds and from other previously reported samples. Some of the variation we have identified in HV31 has previously been associated with phenotypes—for example, copy number polymorphism of *IGHV1-69* is known to strongly correlate with the prevalence of this gene in expressed antibody repertoires, which is

preferentially used in antibodies against certain influenza strains and the HIV-1 virus [7]. However, we have also highlighted extensive variant haplotypes that have not previously been reported. The degree of variation observed in this single sample indicates that much haplotype variation of immunoglobulins and T cell receptor regions remains to be discovered. It is also notable that many of the more complex variants we have identified are not accurately called by the variant calling methods we employed (S1, S2 and S5 Datasets), but this is also an area of active development [17,30,53]. The costs associated with current long-read platforms remain high, but if this can be overcome (e.g. through further improvements in throughput or targeted sequencing approaches [15]) then these methods will become applicable to larger samples, enabling a full catalogue of haplotype variation in these important regions to be generated.

Methods

Ethics statement

HV31 was recruited as a healthy volunteer under approval by the Oxfordshire Research Ethics Committee (COREC reference 06/Q1605/55). The donor provided written informed consent for the use of their blood in research.

Software

Bioinformatics software and algorithms used in this study are summarized in S4 Table.

DNA extraction, sequencing and optical mapping

Blood sample of a healthy female donor of European ancestry identified as HV31 was used in this study. For PacBio HiFi and CLR sequencing, Oxford Nanopore sequencing, 10x linked-read sequencing and MGI stLFR linked-read sequencing, the DNA was extracted from CD14⁺ monocytes isolated from PBMC using CD14 antibody-conjugated beads (Miltenyi Biotec). DNA extraction was performed with QIAGEN MagAttract HMW DNA kit following manufacturer's instructions, with slight adaptations. In brief, 220 μ l Buffer ATL and 20 μ l Proteinase K were added to a suspension of 1×10^6 CD14⁺ monocytes. The mixture was incubated overnight at 56°C, shaking at 900 rpm to lyse the cells. The lysate was then processed according to manufacturer's instructions for the purification of high-molecular-weight genomic DNA from tissue.

For MGI standard short-read, MGI coolMPS and Illumina PCR-free sequencing, the DNA was extracted from PBMC with NEB Monarch Genomic DNA Purification Kit (T3010) following manufacturer's instructions. Library preparation, sequencing and optical mapping were performed following the instructions of the respective platform providers.

Region definition

Eight genomic regions encoding key components of the human immune system, including HLA, IG, TCR and KIR were selected for investigation (Table 1). Each region was defined as a core range in GRCh38 that contained genes related to immune system components, with additional flanking sequences added to both sides. The core range were typically selected based on the respective reference sequences in the NCBI RefSeq database [19]. As exceptions, the core range of the HLA region was defined as the genomic range from *GABBR1* to *KIFC1* genes [20], and the KIR region was defined as the genomic range from *KIR3DL3* to *KIR3DL2* genes [21]. The flanking sequence was typically 1 Mb on either side. As exceptions, the telomeric flanking sequence in the IGH region was limited to 164 kb by the length of chromosome 14. In

addition, we expanded the centromeric flanking sequence in the IGK region by 0.67 Mb to bridge a 1 Mb heterochromatin gap present in GRCh38.

GRCh38 annotations

The GRCh38 segmental duplication, alternative haplotypes and fix patches annotations were downloaded from the UCSC Table Browser [54] based on the genomicSuperDups, altSeqLiftOverPsl and fixSeqLiftOverPsl datasets, respectively.

Whole-genome *de novo* assembly

Canu v1.9 [25] was used to perform whole-genome *de novo* assembly for HV31 based on PacBio HiFi reads, with the following parameters: -pacbio-hifi <FASTQ> genome-Size = 3235000000 -minInputCoverage = 1 -stopOnLowCoverage = 1. The resulting contigs were mapped to GRCh38 using minimap2 [55] with the following parameters: -ax asm5—secondary = no. Contigs that mapped to the 8 loci of interest were extracted as local contigs. For comparison, Peregrine [56] was also used to generate a whole-genome *de novo* assembly for HV31 based on PacBio HiFi reads, with the following parameters: python <peregrine_script> asm <FASTQ_list_file> 16 16 16 16 16 16 16 16—with-consensus.

Hybrid scaffolding and haplotig removal

Hybrid scaffolding was performed using Bionano Solve, a proprietary software provided by Bionano Genomics (<https://bionanogenomics.com/>), with default parameters. We used a custom script based on BiSCoT [26] to improve the contiguity and quality of the resulting scaffolds. Specifically, we merged adjacent contigs in a scaffold if they overlap with each other, as inferred from shared enzymatic labelling sites or sequence alignment. If the two adjacent contigs were expected to be non-overlapping, they were joined with a gap (i.e. a sequence of “N” bases) between them, the size of which was estimated based on the distance of nearest labelling sites. In addition, we incorporated shorter contigs into longer ones if the shorter contig represented a subsequence of the longer contig, and aligned better with the Bionano genome maps.

After scaffolding, we removed duplicated contigs or scaffolds that presumably represent alternative haplotypes (‘haplotigs’) using a custom k-mer based method. In brief, we listed all unique 22-mers for each contig or scaffold and compare these sets of 22-mers in a pairwise manner. If a shorter contig had more than 80% of unique 22-mers shared with a longer contig, then the former was considered as a haplotig and removed from the assembly.

Read mapping

Sequencing reads from each locus of interest were required for various purposes including gap closing, polishing, error rate estimation and assembly validation based on alignment coverage and patterns. In order minimize reference bias, we first mapped the reads from each sequencing dataset using minimap2, and then extracted reads that mapped to contigs that represent each locus of interest [48]. The extracted reads were again mapped with minimap2 to the scaffolded or finalized assembly as appropriate for specific applications.

A unique k-mer anchoring method [29] was used to improve the mapping of long reads in repetitive regions. In brief, given a set of locus-specific reads and a corresponding reference sequence, we first defined a set of anchoring k-mers for each locus of interest. Only k-mers that appeared to be unique in both short read sequencing datasets ($31 \leq \text{multiplicity} \leq 231$) and the reference sequence (copy number = 1; no occurrence outside the locus) were selected as anchoring k-mers. Then, we mapped the reads to the reference with minimap2 using

parameters `-n 50 -r 10000`, which enabled the output of up to 50 alignments for each read, with gap sizes up to 10 kb in each alignment. An optimal alignment for each read were then selected based on the number of bases shared with the reference that were part of an anchoring *k*-mer. These selected alignments were pooled into a new BAM file, after filtering out alignments that were shorter than 5 kb. The resulting BAM file were used for polishing and reference-free alignment validation.

Gap closing and polishing

Gap closing was performed using TGS-GapCloser v1.0.1 [27] with PacBio HiFi reads. Sequencing reads were first mapped to the whole genome assembly produced by Canu, which enabled locus-specific read extraction. The extracted reads were used as input for TGS-GapCloser, which was executed using the following parameters: `-ne—tgstype pb—g_check`. Polishing was performed using Pilon [28] with HiFi reads and MGI paired-end short reads extracted in a similar manner. The default parameters were used. For clarity, the finalized scaffolds were displayed and coordinated based on the relative order and orientations of the corresponding sequence in GRCh38 in visualization steps.

Error rate estimation and reference-free assembly validation

Jellyfish [57] was used to count the multiplicity of each *k*-mer ($k = 22$ or 31) from a pooled FASTQ dataset of PacBio HiFi, MGI standard short-read, MGI CoolMPS, MGI stLFR linked read, 10x Linked-Read and Illumina PCR-free sequencing platforms (S1 Table), with the following parameters: `jellyfish count -m <k> -s 30G —min-qual-char "?" -C`. The accumulated sequencing depth of the pooled FASTQ dataset was $262\times$. In each read, *k*-mers that include bases with base quality < 20 were excluded. For error rate estimation, *k*-mers ($k = 22$) in the HV31 assembly with multiplicity < 5 were classified as erroneous *k*-mers, and clustered by their positions in the assembly, allowing a maximum of $k-1$ correct *k*-mers between two adjacent erroneous *k*-mers in each cluster. The number of erroneous *k*-mer clusters per Mb assembled sequence was used as an indicator of the error rate of the HV31 assembly.

For reference-free assembly validation, we define the normalized multiplicity (*N*) of each *k*-mer ($k = 31$) in the HV31 assembly as $N = M / (C \times D)$, where *M* is the multiplicity of that *k*-mer in the validation dataset, *C* is the copy number of that *k*-mer in the HV31 assembly, and *D* is the mode multiplicity of unique homozygous *k*-mers in the validation dataset, as estimated from the *k*-mer multiplicity histogram (Fig 3A). The normalized *k*-mer coverage was visualized against the position of the *k*-mer, along with the normalized coverage of ONT reads aligned to the assembly using the *k*-mer anchoring method. Regions where the normalized *k*-mer coverage or normalized ONT coverage deviated from 1 were labelled and inspected for potential assembly errors (S3 Table).

Variant calling

PBSV (<https://github.com/PacificBiosciences/pbsv>), a subprogram of SMRT tools was used to call heterozygous SVs from HiFi and CLR reads with default parameters. Sniffles [58] was used to call heterozygous SVs from HiFi and CLR and ONT reads with the following parameters: `-s 3 -q 20—ccs_reads—min_het_af 0.2` (HiFi), `-s 8 -q 20—min_het_af 0.2` (CLR), or `-s 15 -q 20—min_het_af 0.2` (ONT). Unique *k*-mer anchoring was applied prior to SV calling. SVmerge, a subprogram of SVanalyzer [30] was used to cluster and merge SV records from output VCF files of PBSV and Sniffles, with default parameters. SV calling from 10x data was performed using Long Ranger (<https://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger>) and GATK [59], with the following parameters: `longranger wgs—`

id <sample_id>—fastqs <FASTQ_directory>—reference <GRCh38_path>—localcores = 24
—localmem = 383 -vcmode <GATK_path>—disable-ui.

Allelic variant detection

Reference variant sequences of IGHV, IGKV, IGLV, TRAV, TRDV, TRBG and TRGV genes were downloaded from the IMGT reference directory [60] (http://www.imgt.org/download/V-QUEST/IMGTV-QUESTreference_directory.zip). Reference variant sequences of HLA genes were downloaded from the IPD-IMGT/HLA database [61] (ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/fasta/hla_gen.fasta). Reference variant sequences of KIR genes were downloaded from the IPD-KIR database [62] (ftp://ftp.ebi.ac.uk/pub/databases/ipd/kir/fasta/KIR_gen.fasta). The reference gene variant sequences were mapped to GRCh38 or the HV31 assembly using minimap2 with the following parameters: -a -w1 -f1e-9. We extracted subsequences in regions where at least one reference gene was mapped, with 20 bp flanking sequence at either side. These query sequence fragments were submitted to NCBI IgBLAST [63] (for IGHV, IGKV, IGLV, TRAV, TRDV, TRBV and TRGV genes) or NCBI BLAST+ [64] (for HLA and KIR genes) to search for matching sequences in the relevant databases, with default parameters. The top hit variant with the highest match score returned by NCBI IgBLAST or NCBI BLAST+ were assigned to each query fragment. Query fragments shorter than the top hit variant were considered to represent partial alignment and discarded.

Supporting information

S1 Fig. Depth of coverage of sequencing reads across platforms and regions. For each sequencing platform (colored lines, with platforms as in [S1 Table](#)), the plot shows the depth of coverage of reads aligned to GRCh38 across the eight selected regions ([Table 1](#)). Depths are normalized by the average depth across each region for each dataset. Areas with apparent systematic lower depth in the TRA, TRB and TRG regions are highlighted with black triangles. Datasets generated with DNA from CD14+ monocyte and PMBC are denoted with solid and dashed lines, respectively. Decrease of sequencing depths is not identified in immunoglobulin regions, presumably due to the relatively low fraction (5–15%) of B cells in PBMC. Despite $\gamma\delta$ T cells being rarer than α/β T cells, the coverage drop around T cell receptor γ genes can be explained by the fact that the γ locus is known to undergo rearrangement in most α/β T cells [65], and the drop around T cell receptor δ genes can be explained by the fact that any rearrangement at the α locus leads to the loss of the δ locus.[66]
(PNG)

S2 Fig. Whole-genome assembly contigs aligned to GRCh38 in each of the 8 selected loci. Each row represents one local contig extracted from the draft whole-genome assembly ([Fig 1A](#)) by alignment to GRCh38 using minimap2 (Methods). The x axis reflects GRCh38 coordinates across each of the selected regions ([Table 1](#)).
(PNG)

S3 Fig. Read/molecule length distribution of PacBio HiFi, PacBio CLR, ONT and Bionano optical mapping datasets involved in this study. Red and grey vertical lines denote the N50 (i.e. the maximal length such that reads/molecules longer than this length cumulatively account for at least 50% of the total length of reads/molecules in the dataset) and mean read/molecule length for each dataset, respectively.
(PNG)

S4 Fig. Comparison of contig lengths and counts of Bionano DLE-1 markers. (A,B) Number of DLE-1 labels (y axis) plotted against contig length (x axis) for draft whole-genome

assembly contigs (panel A) and Bionano contigs (panel B). For reference, gray vertical and horizontal lines in panel A denote 50 kb length and 10 DLE-1 labels, respectively. (C) Cumulative length of contigs (y axis) containing at least the given number of DLE-1 labels (x axis) is shown for whole-genome assembly contigs (orange) and Bionano contigs (blue). For reference, the gray vertical line denotes 10 DLE-1 labels.

(PNG)

S5 Fig. Comparison of insertions and deletions identified by alignment-based structural variant calling methods. Bars show the number of insertions (blue) and number of deletions (red) identified by each combination of method and sequencing data (x axis) after aligning reads to the HV31 assembly as described in main text and Methods. SVs are classified as insertions or deletions according to whether the alternative haplotype is longer or shorter than the HV31 haplotype. For comparison, the number of SVs called by multiple methods, as identified by SVanalyzer, is indicated by shading.

(PNG)

S6 Fig. Concordance of structural variants called by various methods. Each row shows the fraction of variants called by the corresponding method (y axis), that are also called by the method in the relevant column (x axis). Concordance of SVs is as determined by SVanalyzer.

(PNG)

S7 Fig. Comparison of published assemblies and alternative assembly methods in the eight selected regions. Each row visualizes the alignment pattern of the corresponding assembly with GRCh38 as the reference. Duplicate contigs (i.e. shorter contigs that align within the span of a longer contig) and contig breaks (identified as endpoints of non-duplicate contigs) are shown in cyan and orange respectively, with regions of the GRCh38 reference that are not covered by the aligned assembly contigs denoted by red lines, according to the legend. Each assembly is labeled in the following order: publication, sample, key algorithms, key technology (with OM denoting optical mapping and Hi-C denoting the Hi-C chromosome conformation capture approach) and haplotype (N, not haplotype-resolved; M, maternal haplotype; P, paternal haplotype; H1, haplotype 1; H2, haplotype 2; CHM, complete hydatidiform mole). Relevant publications are: T2T 2021 [32], Chin 2016 [49], Koren 2018 [67], Nurk 2020 [31], Miga 2020 [29], Shafin 2020 [68], Kolmogorov [69], Garg 2021 [16], Ebert 2021 [17]. Genes and segmental duplications are annotated above as in Fig 2.

(PNG)

S8 Fig. Schematic examples of sequence duplications and structural variations demonstrated with k-mer sharing plots. Each panel shows a schematic of the expected pattern visualized on the k-mer sharing plots, given the pattern of sequence duplication or structural variant indicated by the panel label. For clarity, specific sequences are labelled as follows: X, Y and Z denote sequence fragments that are different from each other, and Y' denotes the reverse complement of Y. In each panel, the reference sequence is depicted on the x axis and the alternate sequence is depicted on the Y axis. The size of each structural variant can be estimated from the distance between relevant breakpoints on the plot.

(PNG)

S9 Fig. Error rate estimation of PacBio HiFi and various short read datasets involved in this study using GenomeScope. The histogram of k-mer ($k = 22$) multiplicity in each dataset is shown, after scaling multiplicity values (x axis) and k-mer numbers (y axis) so that the peak of unique homozygous k-mers in each dataset overlap at $x = 1, y = 1$. Numbers in brackets

show the estimated per-base error rate of each dataset as estimated using GenomeScope [35].
(PNG)

S10 Fig. Assembly validation based on k-mer multiplicity. (A) For each k-mer ($k = 31$) that appears in the HV31 assembly, the multiplicity of that k-mer in the validation dataset (y axis) is plotted against the position of that k-mer (x axis), colored by the copy number of that k-mer in the HV31 assembly as shown in the legend. (B) the normalized k-mer multiplicity (y axis), defined as ratio of validation k-mer multiplicity to assembly k-mer multiplicity computed using the values shown in (A), plotted against the position of that k-mer (x axis). Values are further normalized by dividing by the peak multiplicity of unique homozygous k-mers as shown in Fig 3A, such that these kmers are expected to lie near $y = 1$. k-mers that found both inside and outside the given regions are considered noninformative and are shown in gray. Orange lines show ONT read coverage depth normalized to the genome-wide average coverage depth ($63\times$).
(PNG)

S11 Fig. Detecting heterozygous SVs and assembly errors from k-mer multiplicity and coverage depth patterns. (A) A 63.9 kb heterozygous deletion in the HLA locus is revealed by reduced ONT coverage depth (orange) and validation k-mer multiplicity ($k = 31$; blue) appearing at normalized multiplicity close to 0.5. (B) A collapsed duplication in the HLA locus is revealed by elevated ONT coverage depth (orange) and validation k-mer multiplicity (blue) appearing at normalized multiplicity close to 2. k-mers that found both inside and outside the IGH region are considered noninformative and are shown in gray. In (A) and (B), k-mers with multiplicity beyond the axis limits are stacked at the top of the plots.
(PNG)

S12 Fig. GRCh38 and GRCh37 represent different IGH haplotypes. (A) k-mer sharing plot ($k = 50$) comparing GRCh37 and GRCh38 in the IGH region. Similar to HV31, GRCh37 has only one copy of IGHV1-69 and IGHV2-70 genes. The unresolved repeats are highlighted with red arrows. Gray shade marks the position of the 45 kb CNV in HV31 relative to GRCh38 (see Fig 3A). (B) Schematic representation of GRCh38, GRCh37 and HV31 near IGHV1-69 and IGHV2-70 genes. Fragment R denotes the unresolved duplication which was assembled in HV31.
(PNG)

S13 Fig. A compound heterozygous CNV involving IGHV3-30. k-mer sharing plot ($k = 20$) comparing the CLR read with ID 92801871/034335 (y axis) with the HV31 assembly (x axis). The read is consistent with the presence of a three-copy unassembled haplotype as described in main text and Fig 5D. Each copy of the repeat unit is annotated with a number and an arrow for clarity.
(PNG)

S14 Fig. Misalignment resulting from large structural rearrangements in the IGH locus. (A) k-mer sharing plot ($k = 50$) comparing the HV31 assembly (y axis) with GRCh38 (x axis), detailing the 80 kb insertion between IGHV3-37 and IGHV7-40 highlighted in orange in Fig 5C. The insertion introduces extra copies of several gene fragments as annotated above. The region further inspected in panels (B) and (C) is highlighted in blue. (B) Alignment patterns of HiFi reads (rows) to GRCh38 (x axis) in the region highlighted in panel A. Grey bars denote aligned segments, with deletions and insertions denoted in red and purple respectively. Orange vertical lines indicate alignment breakpoints (i.e. alignments are clipped or split at these points). (C) Alignments of HiFi reads (rows) to HV31 (x axis) in the same region. The complex

pattern of insertions, deletions and read clipping in panel (B) arise from between-copy misalignments that are largely resolved when aligning to HV31.

(PNG)

S15 Fig. The HV31 assembly is consistent with NCBI RefSeq NG_001333.2. k-mer sharing plot ($k = 50$) comparing the HV31 assembly (y axis) with the NG_001333.2 contig from NCBI RefSeq (x axis). TRBV genes not included in GRCh38 are highlighted in green.

(PNG)

S16 Fig. Three gaps flanked by high-identity repeats were filled in the HV31 assembly. (A) k-mer sharing plot ($k = 50$) comparing GRCh38 (x axis) with the HV31 assembly (y axis). The 2.56 Mb scaffold and the 1.97 Mb scaffold in the HV31 assembly are shown in blue and green, respectively. Coverage of ONT reads aligned to GRCh38 is displayed above, and the proximal and distal clusters are annotated. Gaps in GRCh38 are shaded in gray. Novel sequence junctions in the HV31 assembly are annotated with red arrows. Sequence fragments of which extra copies were introduced in the HV31 assembly to fill in the gaps between IGK proximal and distal gene clusters in GRCh38 are highlighted in yellow; corresponding read coverage peaks confirm increased genome multiplicity of these fragments. (B) Alignment of Bionano contigs (blue) to the 2.56 Mb scaffold in the HV31 assembly (green). DLE-1 labels and their alignments are denoted by colored lines within and between scaffolds as described in Fig 5D legend; note that all gray alignment lines connect the HV31 scaffold to each of the Bionano contigs (no between-Bionano alignments are shown). The approximate sequence region that maps to the GRCh38 gaps between IGK proximal and distal gene clusters is shaded in gray. For clarity, corresponding positions in the HV31 assembly in panels (A) and (B) are labelled with red arrows. (C) Alignment of BioNano contigs (blue) to the 1.97 Mb scaffold in the HV31 assembly (green). Approximate sequence region that maps to the GRCh38 heterochromatin gap is shaded in gray.

(PNG)

S17 Fig. A 50 kb GRCh38 gap flanked by segmental duplications closed in the HV31 assembly. (A) k-mer sharing plot ($k = 50$) comparing the chr7_KZ208912v1_fix patch sequence (y axis) with GRCh38 (x axis), highlighting the genomic position corresponding to the 140 kb inversion in the HV31 assembly (green region) and a 50 kb gap in GRCh38 (gray region) which is closed in the HV31 assembly. (B) The HV31 assembly (x axis) is consistent with chr7_KZ208912v1_fix sequence (y axis) except for the 21.9 kb gap (brown) and the 140 kb inversion (green).

(PNG)

S18 Fig. The heterochromatin gap in the IGK locus was filled with 650 kb complex repeat sequence. (A) k-mer sharing plot ($k = 50$) comparing the HV31 assembly with itself in the IGK heterochromatin region. Purple lines show the occurrence of a 22 bp HSat2B repeat signature sequence (TTCGATTCATTTGATGATTCCAT). A 32 kb unique sequence fragment is highlighted in blue. (B) Details of k-mer sharing plot in panel (A), zoomed to reveal details of the unique sequence fragment and repeat structure. (C) Comparison of HV31 contigs and Bionano contigs as in panel C in S16 Fig, zoomed in to show that the 32 kb unique fragment (blue shaded region) contains a DLE-1 recognition label that was confirmed by Bionano contigs. (D) k-mer sharing plot ($k = 50$) comparing the HV31 assembly (y axis) with the GenBank AP023554.1 contig (x axis). For reference, the orange box in panels A, B and D denote approximately the same region.

(PNG)

S19 Fig. Comparison of the HV31 and the T2T CHM13 assemblies in the IGK region. (A) k-mer sharing plot ($k = 50$) comparing the HV31 assembly (y axis) with the T2T CHM13 assembly (x axis) in the IGK region. The 2.56 Mb scaffold and the 1.97 Mb scaffold in the HV31 assembly are shown in blue and green, respectively. (B) k-mer sharing plot as in panel (A), zoomed in to show details of the heterochromatin region. The assemblies contain similar sequence in the heterochromatin region, but with some differences including in the position of the unique sequence (highlighted with a red arrow) as noted in main text. (PNG)

S1 Table. Summary of sequencing and genome mapping datasets. Table shows the name, cell type, genome coverage estimated by alignments to GRCh38, and read length statistics for each dataset generated in this study. (PDF)

S2 Table. Comparison of original and modified Merqury error rate estimation results. Table shows per-base error rate estimates for local scaffolds and the finished HV31 assembly (Fig 1A) using the Merqury algorithm, and the modified algorithm described in Methods and Fig 1D legend. As described in Methods, the key difference is that we estimate based on clusters of kmers with low validation coverage and this produces slightly higher estimates in practice than the Merqury method. (PDF)

S3 Table. Potentially problematic regions in the HV31 assembly as identified from k-mer multiplicity. Table shows details of locations where validation k-mers show discrepancy from expectation (as numbered in S10 Fig) and our conclusion about the region. For each region, the table lists the relevant assembled region (Table 1), number as shown in S10 Fig, type of evidence we have inspected to provide additional evidence, and our conclusion about the cause of discrepancy. (PDF)

S4 Table. Detail of bioinformatics tools used in this study. (PDF)

S1 Dataset. List of structural variants reported by SVanalyzer. Table shows structural variants detected by applying pbsv and sniffles to the long-read sequence reads aligned to the HV31 assembly, as output by SVanalyzer and described in main text and methods. Columns reflect the region, scaffold, position and identifier of the SV, the assembly and alternate alleles, and the genotypes assigned by each combination of long-read dataset and SV calling method. (TSV)

S2 Dataset. Structural variants detected by the 10x Genomics Long Ranger pipeline. Table reflects SVs identified by 10x Long Range pipelines applied to 10x reads aligned to the GRCh38 reference sequence in the eight selected regions (Table 1). Columns show the region, GRCh38 chromosome, position and identifier of the SV, followed by the GRCh38 allele and the alternate allele. The last column reflects detailed information including SV type (deletion or insertion) and length, as output by the Long Ranger pipeline. (TSV)

S3 Dataset. Benchmarking information of the assembly pipeline. Table provides a summary of computational resources used for each step in our assembly pipeline (Fig 1A). Columns reflect an identifier for the pipeline step, bioinformatics tool used, the total number of processes run, the average time taken per job, the number of cores allocated to each job and the

average memory used per job (computed as the maximum over the lifetime of the job).
(XLSX)

S4 Dataset. k-mer sharing plots (k = 50) comparing the HV31 assembly with GRCh38 in each of the eight regions of interest. Plots provide further detail of regional k-mer sharing plots shown in Fig 2, with details as described in Fig 2 legend and main text. In the IGK and IGL regions, colors reflect the distinct HV31 scaffolds.
(ZIP)

S5 Dataset. Structural variants between GRCh38 and the HV31 assembly detected by Assemblytics. Table shows the output of Assemblytics applied to GRCh38 and HV31 in each region as described in main text and Methods. Columns show the region, chromosome and position in GRCh38, SV identifier, the identified type of the SV, details of the SV length and the matching coordinates in HV31.
(TSV)

Author Contributions

Conceptualization: Antony J. Cutler, Rachael Bashford-Rogers, Julian C. Knight, John A. Todd, Gavin Band.

Data curation: Jia-Yuan Zhang, Hannah Roberts, David S. C. Flores, Justin P. Whalley, Gavin Band.

Formal analysis: Jia-Yuan Zhang, Hannah Roberts, David S. C. Flores.

Funding acquisition: Julian C. Knight, John A. Todd.

Investigation: Jia-Yuan Zhang, Hannah Roberts, David S. C. Flores, Antony J. Cutler, Andrew C. Brown, Justin P. Whalley, Olga Mielczarek, Barbara Xella, Karen Oliver, Craig Corton, Emma Betteridge, Gavin Band.

Methodology: Jia-Yuan Zhang, Antony J. Cutler, Andrew C. Brown, Gavin Band.

Supervision: Jia-Yuan Zhang, Antony J. Cutler, David Buck, Helen Lockstone, Rachael Bashford-Rogers, Julian C. Knight, John A. Todd, Gavin Band.

Visualization: Jia-Yuan Zhang.

Writing – original draft: Jia-Yuan Zhang, Gavin Band.

Writing – review & editing: Jia-Yuan Zhang, Antony J. Cutler, Andrew C. Brown, Justin P. Whalley, Rachael Bashford-Rogers, Julian C. Knight, John A. Todd, Gavin Band.

References

1. Venkataraman GR, Olivieri JE, DeBoever C, Tanigawa Y, Justesen JM, Dilthey A, et al. Pervasive additive and non-additive effects within the HLA region contribute to disease risk in the UK Biobank. *BioRxiv*. 2020.
2. de Bakker PI, Raychaudhuri S. Interrogating the major histocompatibility complex with high-throughput genomics. *Hum Mol Genet*. 2012; 21: R29–R36. <https://doi.org/10.1093/hmg/dds384> PMID: 22976473
3. Tian C, Hromatka BS, Kiefer AK, Eriksson N, Noble SM, Tung JY, et al. Genome-wide association and HLA ion fine-mapping studies identify susceptibility loci for multiple common infections. *Nat Commun*. 2017; 8: 1–13. <https://doi.org/10.1038/s41467-016-0009-6> PMID: 28232747
4. Fellay J, Ge D, Shianna KV, Colombo S, Ledergerber B, Cirulli ET, et al. Common genetic variation and the control of HIV-1 in humans. *PLoS Genet*. 2009; 5: e1000791. <https://doi.org/10.1371/journal.pgen.1000791> PMID: 20041166

5. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. *Nat Rev Immunol*. 2018; 18: 325–339. <https://doi.org/10.1038/nri.2017.143> PMID: 29292391
6. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. *Nature*. 2016; 530: 177–183. <https://doi.org/10.1038/nature16549> PMID: 26814963
7. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun*. 2012; 13: 363–373. <https://doi.org/10.1038/gene.2012.12> PMID: 22551722
8. Bashford-Rogers R, Bergamaschi L, McKinney E, Pombal D, Mescia F, Lee J, et al. Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature*. 2019; 574: 122–126. <https://doi.org/10.1038/s41586-019-1595-3> PMID: 31554970
9. Kusnierczyk PK. Killer cell immunoglobulin-like receptor gene associations with autoimmune and allergic diseases, recurrent spontaneous abortion, and neoplasms. *Front Immunol*. 2013; 4: 8. <https://doi.org/10.3389/fimmu.2013.00008> PMID: 23372569
10. The 1000 Genomes Project Consortium, Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, et al. A global reference for human genetic variation. *Nature*. 2015; 526: 68–74. <https://doi.org/10.1038/nature15393> PMID: 26432245
11. Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Collins RL, Brand H, Karczewski KJ, Zhao X, et al. A structural variation reference for medical and population genetics. *Nature*. 2020; 581: 444–451. <https://doi.org/10.1038/s41586-020-2287-8> PMID: 32461652
12. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *BioRxiv*. 2019; 563866.
13. Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol*. 2019; 20: 97. <https://doi.org/10.1186/s13059-019-1707-2> PMID: 31104630
14. Luo S, Jane AY, Li H, Song YS. Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans. *Life Sci Alliance*. 2019; 2. <https://doi.org/10.26508/lsa.201800221> PMID: 30808649
15. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, et al. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *Front Immunol*. 2020; 11. <https://doi.org/10.3389/fimmu.2020.02136> PMID: 33072076
16. Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol*. 2021; 39: 309–312. <https://doi.org/10.1038/s41587-020-0711-0> PMID: 33288905
17. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021; 372: eabf7117. <https://doi.org/10.1126/science.abf7117> PMID: 33632895
18. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT, the international ImMunoGeneTics information system 25 years on. *Nucleic Acids Res*. 2015; 43: D413–D422. <https://doi.org/10.1093/nar/gku1056> PMID: 25378316
19. O’Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016; 44: D733–D745. <https://doi.org/10.1093/nar/gkv1189> PMID: 26553804
20. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet*. 2009; 54: 15–39. <https://doi.org/10.1038/jhg.2008.5> PMID: 19158813
21. Carrington M, Norman P. The KIR gene cluster. *Natl Cent Biotechnol Inf US*. 2003.
22. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation. *Am J Hum Genet*. 2013; 92: 530–546. <https://doi.org/10.1016/j.ajhg.2013.03.004> PMID: 23541343
23. Watson CT, Steinberg KM, Graves TA, Warren RL, Malig M, Schein J, et al. Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. *Genes Immun*. 2015; 16: 24–34. <https://doi.org/10.1038/gene.2014.56> PMID: 25338678
24. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020; 21: 245. <https://doi.org/10.1186/s13059-020-02134-9> PMID: 32928274
25. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res*. 2017; 27: 722–736. <https://doi.org/10.1101/gr.215087.116> PMID: 28298431

26. Istace B, Belsler C, Aury J-M. BiSCoT: Improving large eukaryotic genome assemblies with optical maps. *Bioinformatics*; 2019 Jun. <https://doi.org/10.1101/674721>
27. Xu M, Guo L, Gu S, Wang O, Zhang R, Fan G, et al. TGS-GapCloser: fast and accurately passing through the Bermuda in large genome using error-prone third-generation long reads. *Bioinformatics*; 2019 Nov. <https://doi.org/10.1101/831248>
28. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Wang J, editor. *PLoS ONE*. 2014; 9: e112963. <https://doi.org/10.1371/journal.pone.0112963> PMID: 25409509
29. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*. 2020 [cited 5 Aug 2020]. <https://doi.org/10.1038/s41586-020-2547-7> PMID: 32663838
30. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, et al. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol*. 2020; 38: 1347–1355. <https://doi.org/10.1038/s41587-020-0538-8> PMID: 32541955
31. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res*. 2020; 30: 1291–1305. <https://doi.org/10.1101/gr.263566.120> PMID: 32801147
32. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *bioRxiv*. 2021; 2021.05.26.445798. <https://doi.org/10.1101/2021.05.26.445798>
33. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol*. 2004; 5: R12. <https://doi.org/10.1186/gb-2004-5-2-r12> PMID: 14759262
34. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*. 2016; 32: 3021–3023. <https://doi.org/10.1093/bioinformatics/btw369> PMID: 27318204
35. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 2017; 33: 2202–2204. <https://doi.org/10.1093/bioinformatics/btx153> PMID: 28369201
36. Myers EW. Toward Simplifying and Accurately Formulating Fragment Assembly. *J Comput Biol*. 1995; 2: 275–290. <https://doi.org/10.1089/cmb.1995.2.275> PMID: 7497129
37. Altemose N, Miga KH, Maggioni M, Willard HF. Genomic Characterization of Large Heterochromatic Gaps in the Human Genome Assembly. Ouzounis CA, editor. *PLoS Comput Biol*. 2014; 10: e1003628. <https://doi.org/10.1371/journal.pcbi.1003628> PMID: 24831296
38. Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, et al. The evolution and population diversity of human-specific segmental duplications. *Nat Ecol Evol*. 2017; 1: 0069. <https://doi.org/10.1038/s41559-016-0069> PMID: 28580430
39. Takayama J, Tadaka S, Yano K, Katsuo F, Gocho C, Funayama T, et al. Construction and integration of three de novo Japanese human genome assemblies toward a population-specific reference. *Nat Commun*. 2021; 12: 226. <https://doi.org/10.1038/s41467-020-20146-8> PMID: 33431880
40. Kirsch S, Weiß B, Miner TL, Waterston RH, Clark RA, Eichler EE, et al. Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome. *Genome Res*. 2005; 15: 195–204. <https://doi.org/10.1101/gr.3302705> PMID: 15653831
41. Lyle R, Prandini P, Osogawa K, Ten Hallers B, Humphray S, Zhu B, et al. Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. *Genome Res*. 2007; 17: 1690–1696. <https://doi.org/10.1101/gr.6675307> PMID: 17895424
42. Dilthey A, Cox C, Iqbal Z, Nelson MR, McVean G. Improved genome inference in the MHC using a population reference graph. *Nat Genet*. 2015; 47: 682–688. <https://doi.org/10.1038/ng.3257> PMID: 25915597
43. Dilthey AT, Gourraud P-A, Mentzer AJ, Cereb N, Iqbal Z, McVean G. High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. Franke A, editor. *PLoS Comput Biol*. 2016; 12: e1005151. <https://doi.org/10.1371/journal.pcbi.1005151> PMID: 27792722
44. Motyer A, Vukcevic D, Dilthey A, Donnelly P, McVean G, Leslie S. Practical use of methods for imputation of HLA alleles from SNP genotype data. *bioRxiv*. 2016; 091009.
45. Vince N, Douillard V, Geffard E, Meyer D, Castelli EC, Mack SJ, et al. SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics. *Genet Epidemiol*. 2020; 44: 733–740. <https://doi.org/10.1002/gepi.22334> PMID: 32681667

46. Vukcevic D, Traherne JA, Næss S, Ellinghaus E, Kamatani Y, Dilthey A, et al. Imputation of KIR types from SNP variation data. *Am J Hum Genet.* 2015; 97: 593–607. <https://doi.org/10.1016/j.ajhg.2015.09.005> PMID: 26430804
47. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* 2017; 27: 849–864. <https://doi.org/10.1101/gr.213611.116> PMID: 28396521
48. Chin C-S, Wagner J, Zeng Q, Garrison E, Garg S, Fungtammasan A, et al. A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat Commun.* 2020; 11: 4794. <https://doi.org/10.1038/s41467-020-18564-9> PMID: 32963235
49. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods.* 2016; 13: 1050–1054. <https://doi.org/10.1038/nmeth.4035> PMID: 27749838
50. Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, et al. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J Comput Biol.* 2015; 22: 498–509. <https://doi.org/10.1089/cmb.2014.0157> PMID: 25658651
51. Porubsky D, Garg S, Sanders AD, Korbel JO, Guryev V, Lansdorp PM, et al. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat Commun.* 2017; 8: 1293. <https://doi.org/10.1038/s41467-017-01389-4> PMID: 29101320
52. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol.* 2019; 37: 1155–1162. <https://doi.org/10.1038/s41587-019-0217-9> PMID: 31406327
53. Heller D, Vingron M, Church G, Li H, Garg S. SDip: A novel graph-based approach to haplotype-aware assembly based structural variant calling in targeted segmental duplications sequencing. *Bioinformatics*; 2020 Feb. <https://doi.org/10.1101/2020.02.25.964445>
54. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 2004; 32: D493–D496. <https://doi.org/10.1093/nar/gkh103> PMID: 14681465
55. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018; 34: 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191> PMID: 29750242
56. Chin C-S, Khalak A. Human Genome Assembly in 100 Minutes. *BioRxiv*; 2019 Jul. <https://doi.org/10.1101/705616>
57. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* 2011; 27: 764–770. <https://doi.org/10.1093/bioinformatics/btr011> PMID: 21217122
58. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods.* 2018; 15: 461–468. <https://doi.org/10.1038/s41592-018-0001-7> PMID: 29713083
59. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20: 1297–1303. <https://doi.org/10.1101/gr.107524.110> PMID: 20644199
60. Giudicelli V, Brochet X, Lefranc M-P. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc.* 2011; 2011: pdb-prot5633.
61. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SG. IPD-IMGT/HLA Database. *Nucleic Acids Res.* 2020; 48: D948–D955. <https://doi.org/10.1093/nar/gkz950> PMID: 31667505
62. Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SG. IPD—the immuno polymorphism database. *Nucleic Acids Res.* 2012; 41: D1234–D1240. <https://doi.org/10.1093/nar/gks1140> PMID: 23180793
63. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 2013; 41: W34–W40. <https://doi.org/10.1093/nar/gkt382> PMID: 23671333
64. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009; 10: 421. <https://doi.org/10.1186/1471-2105-10-421> PMID: 20003500
65. Alexandre D, Lefranc M-P. The human γ/δ^+ and α/β^+ T cells: a branched pathway of differentiation. *Mol Immunol.* 1992; 29: 447–451. [https://doi.org/10.1016/0161-5890\(92\)90001-e](https://doi.org/10.1016/0161-5890(92)90001-e) PMID: 1533010
66. Janeway CA Jr, Travers P, Walport M, Shlomchik MJ. T-cell receptor gene rearrangement. *Immunobiology: The Immune System in Health and Disease* 5th edition. Garland Science; 2001.
67. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol.* 2018; 36: 1174–1182. <https://doi.org/10.1038/nbt.4277> PMID: 30346939

68. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol.* 2020; 38: 1044–1053. <https://doi.org/10.1038/s41587-020-0503-6> PMID: 32686750
69. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol.* 2019; 37: 540–546. <https://doi.org/10.1038/s41587-019-0072-8> PMID: 30936562