A thesis submitted for the degree of Doctor of Philosophy

# Genomic Analyses of the Immune System

Jia-Yuan Zhang

St Cross College, University of Oxford

Trinity term 2023

# Contents

# Abstract

### Project 1. Genetic variation in key immune system components

Genes underpinning the diversity and plasticity of the human adaptive immune system, such as the HLA and immunoglobulins, are known for their complex structures and polymorphism. The emergence of long-read sequencing technologies has revolutionised genomics research, in particular the characterisation of segmental duplications and structural variation. Here, using long-read sequencing and additional genomics data from a healthy donor identified as HV31, I built two iterations of *de novo* personal genome assemblies for HV31 as a foundation to study the genetic variation of the immune system. I analysed complex structural variants found in genomic regions encoding key immune system components, and validated them against sequencing data. I also evaluated long-read sequencing accuracy and developed a tool for genomic data visualisation. Collectively, these efforts demonstrate the applications of personal genome assemblies in studying the immune system.

### Project 2. Effects of low-dose IL-2 immunotherapy in T and NK cells

Low-dose interleukin-2 (IL-2) immunotherapy is a promising treatment for type 1 diabetes (T1D). IL-2 supresses autoimmune reactions by increasing the number of regulatory T cells (Tregs). To better understand the mechanism of action of low-dose IL-2 immunotherapy, I analysed single-cell multiomics data of T and NK cells collected from T1D patients before and after low-dose IL-2 treatment. I confirmed that low-dose IL-2 selectively expanded thymic-derived FOXP3$^+$ HELIOS$^+$ regulatory T cells and CD56$^{br}$ NK cells, and showed that the treatment reduced the frequency of IL-21-producing CD4$^+$ T cells. In addition, I identified a long-lived gene expression signature induced by IL-2, which featured the upregulation of *CISH* and downregulation of *AREG*. Notably, I found that the signature remained

detectable one month after the treatment. Further analyses of publicly available COVID-19 cohort data revealed that SARS-CoV-2 infection induced opposite changes that persisted for several months after recovery. These findings suggested potential mechanisms of long COVID and longer-term benefits of IL-2 immunotherapy.

# Acknowledgements

First and foremost, I would like to thank my supervisors Professor John Todd, Professor Linda Wicker, Dr Gavin Band, Dr Ricardo Ferreira, and Dr Dan Crouch. I was inspired every day by the expertise and passion they demonstrated in their research. It was a great privilege to work with them.

I would like to thank everyone who helped along the way. In particular, Dr Chun-Xiao Song introduced Oxford to me during my summer internship in 2018. Professor Jim Hughes, Professor James Daves, and Dr Corey Watson provided helpful feedback to my work. Professor Andrew Pollard offered general advice and discussions on my DPhil study. I appreciate the opportunity to collaborate with Dr Tony Cutler, Dr Fiona Hamey, Dr Jamie Inshaw, Dr Jason Hendry, Dr Justin Whalley, Dr Andrew Brown, and Professor Julian Knight in my research. Discussions with Dr André Python and Dr Azim Ansari were always inspiring.

My DPhil study was supported by the China Scholarship Council, the Wellcome Trust, and JDRF.

Finally, I want to thank those who brighten my life. My parents Zhi-Wu Zhang and Gui-Lin Liu who are always wise and supportive. For me, to grow up is to constantly rediscover how great they are. My partner Yong Zhu understands me as who I am. With her, life becomes an adventure full of colours.

# Associated Publications

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

*PLOS Computational Biology* (2021)

Jia-Yuan Zhang, Hannah Roberts, David S. C. Flores, Antony J. Cutler, Andrew C. Brown, Justin P. Whalley, Olga Mielczarek, David Buck, Helen Lockstone, Barbara Xella, Karen Oliver, Craig Corton, Emma Betteridge, Rachael Bashford-Rogers, Julian C. Knight, John A. Todd, Gavin Band

### Low-dose IL-2 reduces IL-21$^{+}$ T cell frequency and induces anti-inflammatory gene expression in type 1 diabetes

*Nature Communications* (2022)

Jia-Yuan Zhang, Fiona Hamey, Dominik Trzupek, Marius Mickunas, Mercede Lee, Leila Godfrey, Jennie H. M. Yang, Marcin L. Pekalski, Jane Kennet, Frank Waldron-Lynch, Mark L. Evans, Timothy I. M. Tree, Linda S. Wicker, John A. Todd, Ricardo C. Ferreira

# Declarations

I declare that, unless otherwise stated, all work presented in this thesis is my own. Several aspects of the study relied upon collaboration where part of the work was conducted with or by others, as summarised below and detailed in corresponding sections.

## Project 1. Genetic Variation in Key Immune System Components

Genomic and functional data involved in the HV31 project were generated in multiple phases by collaborators from both within and outside the University of Oxford, and pre-processed by Hannah Roberts and Gavin Band. From September 2019, I have been responsible for the bioinformatics analyses of the HV31 genomic and functional data, supervised by Gavin Band, John Todd, and Julian Knight. David Flores, David Smith, and Qijing Shen also contributed significantly to several other aspects of the analyses. Some contents related to this project were adapted from previously published works[1,2] contributed by me and co-authors.

## Project 2. Effects of Low-dose IL-2 Immunotherapy in T and NK Cells

Clinical trial and sample collection for the DILfrequency study was conducted by collaborators from both within and outside the University of Oxford. Flow cytometry experiments and analyses were performed by Ricardo Ferreira and included in a previous publication[3]. Single-cell sequencing data was generated by Ricardo Ferreira and pre-processed and annotated by Dominik Trzupek, Fiona Hamey and Ricardo Ferreira. From February 2021, I have been responsible for the DILfrequency single-cell sequencing data, supervised by Ricardo Ferreira, Linda Wicker and John Todd. Fiona Hamey and Justin Whalley also contributed significantly to several other aspects of the analyses. Some contents related to this project were adapted from previously published works[4,5] contributed by me and co-authors.

# List of Figures

# List of Tables

# Project 1. Genetic Variation in Key Immune System Components

## 1.1. Introduction

### 1.1.1 Genetics of the human adaptive immune system

The adaptive immune system is critical for the body's defence against myriad environmental pathogens. The adaptive immune system found in human and other mammals is referred to as the BCR-TCR-MHC-based adaptive immune system[6], emphasising the central roles of three key components: the major histocompatibility complex (MHC), B cell receptors (BCR) and T cell receptors (TCR), each with specialised functions that work together to provide a coordinated immune response.

#### The human leukocyte antigen (HLA)

The human MHC, commonly known as the human leukocyte antigen (HLA), is a group of cell surface proteins that play a critical role in the recognition and presentation of foreign antigens to T cells[7]. HLA molecules, encoded by the HLA locus on chromosome 6 (p22.1), consist of two main classes: class I and class II. HLA class I molecules, encoded by *HLA-A*, *HLA-B*, and *HLA-C* genes, are expressed on the surface of almost all nucleated cells and are responsible for presenting intracellular antigens to CD8$^+$ T cells. These antigens are typically derived from viral or tumour proteins that are synthesised within the infected or abnormal cell. HLA class II molecules, encoded by *HLA-DP*, *HLA-DQ*, and *HLA-DR* genes, are expressed on the surface of specialised antigen-presenting cells (APCs) such as dendritic cells, macrophages, and B cells, and are responsible for presenting extracellular antigens to CD4$^+$ T cells. These antigens are typically derived from the surface or internal compartments of pathogens that have been engulfed and degraded by APCs.

HLA genes are highly polymorphic, with multiple alleles encoding different HLA molecules within the same individual and across the population, enabling the immune system to recognise a vast array of foreign antigens and respond accordingly. Genome-wide association studies (GWAS) have shown that the HLA locus is a major genetic risk factor for autoimmune diseases, such as type 1 diabetes[8], rheumatoid arthritis[9], and multiple sclerosis[10].

## B cell receptors and immunoglobulins

BCRs are the membrane-bound form of immunoglobulins (IGs), commonly known as antibodies, which recognise and bind to specific antigens, leading to their neutralization or elimination. BCRs and IGs are encoded by three loci in humans: the immunoglobulin heavy chain (IGH) locus, the immunoglobulin κ (IGK) locus, and the immunoglobulin λ (IGL) locus, each of which encodes a different peptide chain of the antibody. The IGH locus, located on chromosome 14 (q32.33), contains the highest number of gene segments and encodes the heavy chain of immunoglobulins. The IGK locus on chromosome 2 (p11.2) and the IGL locus on chromosome 22 (q11.22) encode the κ chain and the λ chain, which form the immunoglobulin light chain in humans.

In most cells of the human body, the three IG loci exist in their germline configurations in the genome, which include a number of variable (V) and joining (J) gene segments. The IGH locus additionally contains diverse (D) segments. Gene segments of the same type share similar sequences. In B cells, before immunoglobulins can be produced, a genetic process called the V(D)J recombination[11] removes certain V, D, and J segments, leaving the remaining segments randomly recombined. The outcome of this random process is the generation of a unique and diverse repertoire of immunoglobulins within each individual's B cells. In addition to the V(D)J recombination, the IGH locus also undergoes somatic hypermutation[12], which further expands the diversity of the immunoglobulin repertoire.

### T cell receptors

TCRs are heterodimeric proteins expressed on the surface of T cells and, analogous to BCRs, function as receptors for specific antigens. In humans, there are two types of TCRs: αβ TCRs, which are composed of α and β chains, and γδ TCRs, which are composed of γ and δ chains. The α and δ chains are encoded by the overlapping TRA and the TRG loci on chromosome 14 (q11.2), respectively, while the β and γ chains are encoded by the TRB locus on chromosome 14, and the TRG locus on chromosome 7 (q34). Similar to IG loci in B cells, TCR loci undergo V(D)J recombination in T cells, yielding an individual-specific diverse repertoire of TCRs.

### Killer-cell immunoglobulin-like receptors (KIRs)

Killer-cell immunoglobulin-like receptors (KIRs) are a group of transmembrane glycoproteins that are expressed on the surface of natural killer (NK) cells and a subset of $CD8^+$ T cells. By binding to specific HLA molecules, KIRs are involved in the recognition of self and non-self, and play a crucial role in the regulation of immune responses. Although KIRs are part of the innate immune system rather than the adaptive immune system, the genetics of KIRs is complex and diverse, resembling that of the HLA in terms of a high degree of polymorphism and variability between individuals. This likely originated from the coevolution of KIR and HLA, as the product of each KIR gene specifically recognises a distinct subset of HLA allotypes[13].

The KIR locus is located on chromosome 19 (q13.42) and are organised into two clusters, KIR A and KIR B[14]. The KIR A cluster contains genes encoding inhibitory receptors, while the KIR B cluster contains genes encoding both inhibitory and activating receptors.

### 1.1.2. DNA sequencing technologies

#### Sanger sequencing

Over the past 50 years, the development of DNA sequencing technologies has revolutionized genomics research. Sanger sequencing, named after its inventor Frederick Sanger, contributed significantly to the success of the Human Genome Project, and remains widely used today[15]. Also known as chain termination sequencing, Sanger sequencing was based on the selective incorporation of chain-terminating dideoxynucleotides (ddNTPs) during DNA synthesis. The original process[16] involves four separate DNA sequencing reactions, each containing a small amount of one of the four ddNTPs (ddATP, ddCTP, ddGTP, and ddTTP) mixed with ordinary deoxyribonucleotide triphosphates (dNTPs). In a given sequencing reaction, when the DNA polymerase incorporates the ddNTP instead of an ordinary dNTP, DNA synthesis is prematurely terminated at the current location due to the lack of the 3' hydroxyl group in ddNTP necessary for the ligation of the next nucleotide. This random process produces a series of fragments with varying lengths, corresponding to the locations of the given base in the template DNA. These fragments are subsequently separated using gel electrophoresis to reveal the template DNA sequence. Modern versions of the Sanger sequencing reaction typically include two major improvements over the original process: (i) each of the four ddNTPs is labelled with a unique fluorophore and mixed together with normal dNTPs in one sequencing reaction, eliminating the need for four separate reactions; (ii) capillary electrophoresis combined with multi-channel fluorescence detectors replaced gel electrophoresis, enabling better resolution for longer DNA fragments and automation of the whole sequencing process[17,18].

### Short-read high-throughput sequencing

High-throughput sequencing (HTS), also known as massive parallel sequencing (MPS) or next-generation sequencing (NGS), inherited the chain termination chemistry from Sanger sequencing and upgraded the technology to allow the parallel sequencing of millions of DNA fragments[17]. Various implementations of HTS platforms have been successfully commercialized by manufacturers such Illumina, Roche and MGI, all of which are based on a workflow termed cyclic reversible termination[17], which significantly improves throughput over Sanger sequencing by monitoring the fluorescence signal generated during DNA synthesis in parallel *in* situ, rather than sequentially after size separation. To achieve this, DNA fragments are typically first tethered to a surface and then amplified using localized PCR, producing a large number of colonies each containing identical copies of a fragment. During the sequencing reaction, fluorophore-labelled reversible DNA synthesis terminators are used in place of the mixture of ddNTPs and dNTPs, so that DNA synthesis is synchronised to terminate after the incorporation of each new base and resume after recording the fluorescence signal of each colony, which reveals the identity of the incorporated base.

HTS significantly facilitated genomics research by enabling the routine sequencing of whole animal or plant genomes. However, applications of HTS are limited by the length of sequence that can be reliably read from a given DNA molecule, known as the read length, which is typically 100 to 250 base pairs (bp) from both ends of the DNA molecule. Genomic regions that contain repeat sequences longer than the typical read length, which are common in animal and plant genomes[19,20], are challenging to analyse, mainly due to the difficulty in mapping each read to the correct copy of the repeat. For example, 3804 protein coding genes in the human genome, including key immune system genes such as *HLA-DRB5*, have been identified as containing regions challenging for high-throughput short-read sequencing technologies[21].

Technically, the read lengths of current short-read HTS platforms are primarily limited by the reliance upon the cyclic reversible termination workflow, whereas the sequencing chemistry does not yet allow rapid and synchronized termination and reversal of all local DNA synthesis reactions[22]. Within each cycle, a small fraction of molecules in each colony falls out of phase, meaning that either zero or more than one nucleotide is incorporated, which will produce noise in the fluorescent signal in future cycles. As the sequencing proceeds and the reads grow longer, this issue accumulates and leads to signal deterioration[23]. In addition, the termination and reversal reactions place a cap on the speed of sequencing reactions, limiting the number of bases that can be sequenced within a given time limit[24]. The cyclic reversible termination workflow, despite its limitations, is essential to short-read HTS platforms due to the technical difficulty in monitoring the DNA polymerase reaction at the single-molecule resolution[25]. Therefore, it is necessary to generate colonies of identical DNA fragments and ensure the synchronization of DNA synthesis within each colony using reversible terminators.

### Long-read sequencing

Long-read sequencing, sometimes known as third-generation sequencing (TGS), first made it possible to sequence single DNA molecules in real time and therefore circumvents the issues associated with the cyclic reversible termination workflow that limited the read length. Two distinct solutions of the single-molecule sequencing problem are implemented by the two major manufacturers of long-read sequencing platforms, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), both of which enabled the generation of read lengths on the order of 10-100 kb. The PacBio solution, termed single-molecule real-time (SMRT) sequencing, is based on a nanophotonic structure called the zero-mode waveguide (ZMW), which are essentially nanometre-scale chambers in which single-molecule DNA polymerase reactions can be monitored continuously[26]. The ONT solution, on the other hand, does not rely on the sequencing-by-synthesis paradigm established by Sanger sequencing and HTS. Instead of harnessing the natural DNA replication process mediated by DNA polymerases, ONT sequencing is based on ionic current changes detected near a protein nanopore when a single DNA molecule passes through, which fluctuates as different nucleotides interact with the nanopore, depending on the chemical structure of the base[27,28].

Both PacBio and ONT sequencing initially suffered from higher error rates (~10%) compared to short-read HTS (< 1%), which improved as the technologies matured[29]. Most notably, PacBio HiFi sequencing, which is based on the circular consensus sequencing (CCS)[30] chemistry, where each DNA molecule is circularised and sequenced multiple times before a consensus sequence is generated computationally, provided considerable accuracy improvements over the raw reads yielded from ZMWs[31], the latter also referred to as continuous long reads (CLR). ONT, on the other hand, achieved progressively higher accuracy though iterated optimisation of the nanopore structure and the base calling algorithm that decode ionic current changes into DNA sequences[32].

### Linked-read sequencing and optical mapping

In addition to PacBio and ONT long-read sequencing, other technologies also achieved success in obtaining long-range genomic information useful for challenging applications such as long-range phasing, structural variation detection and *de novo* assembly[33–35]. 10X Genomics linked-read sequencing[36] and MGI single-tube long-fragment read (stLFR) sequencing[33] enhance standard short-read HTS by labelling each DNA molecule with a unique barcode prior to fragmentation. When the DNA molecule is sequenced, multiple reads are yielded which all share the same barcode information, which can be computationally linked together to provide information at a range longer than individual reads. Meanwhile, Bionano optical mapping[37] directly analyses the physical structure of long DNA molecules using fluorescence imaging. Unlike sequencing-based methods, optical mapping works by first labelling specific sequence motifs in DNA molecules with fluorescent dyes, and then image the labelled molecule using fluorescence microscopy. The resulting high-resolution images are computationally analysed to provide approximate information about the relative positions of sequence motifs, revealing any large structural variants.

### 1.1.3. Applications of long-read sequencing

#### Structural variation detection

Structural variation (SV) is a class of genetic variants that involve large-scale changes in the structure of DNA sequence, typically categorised into insertions, deletions, duplications, inversions, translocations, copy number variants (CNVs), and complex SVs that involve a combination of different types of changes. In practice, SVs are commonly defined DNA sequence alterations larger than 50 bp[38]. The genomic difference between two randomly selected humans are expected to be predominantly (> 90%) attributed to SVs rather than single-nucleotide variants (SNVs)[39]. Specific SVs have been shown to play important roles in human diseases. For example, the Huntington's disease is a genetic disorder caused by the expansion of CAG repeats in the *HTT* gene[40]. In addition, *HER2*$^+$ breast cancer, characterized by overexpression or amplification of the human epidermal growth factor receptor 2 (*HER2*) gene, contributed by various types of somatic SVs including *HER2* gene amplification and chromosomal rearrangements[41].

Despite their contribution to genetic diversity and disease susceptibility, SVs remain less studied compared to SNVs, due to the technical challenges involved in SV detection using short-read sequencing, as SVs are often larger than the read length, involve complex sequence alternations, or exist in repetitive regions[38]. Long-read sequencing provides better opportunities to span SV sequences and resolve long repeats, which greatly facilitated SV detection. Various tools for long-read SV calling are available, many of which have been tailored for PacBio or ONT platforms, such as pbsv[42], Sniffles[43] and NanoSV[44].

### *De novo* assembly

*De novo* assembly refers to the computational process to fully or partially construct a genome without relying on a reference sequence, typically by piecing together a large number of sequencing reads each representing a small fragment of the target genome[45]. For species without existing reference genomes, *de novo* assembly is performed to build the first genomic sequences which can be used as reference genomes for future analyses. The Human Genome Project is a well-known *de novo* assembly effort, yielding the first human reference genome in 2003. The GRCh38 reference genome[46], the latest successor of the original sequences published by the Human Genome Project, was released in 2013 and remains widely used today.

For species with existing high-quality reference genomes available, in particular humans, model organisms, and domesticated animals and plants, *de novo* assembly is also applied to study individual genomes while eliminating reference bias, which refers to the problem that genomic regions that diverge from the reference genome are systematically harder to analyse compared to those identical to the reference genome[47]. As whole-genome sequencing continues to become cheaper and more accessible, personal genomes build from *de novo* assembly are gaining importance in precision medicine[48] and promoting deeper understanding of human genetic diversity[49].

Long-read sequencing provides the ability to generate contiguous and comprehensive assemblies of complex genomes such as the human genome, by allowing each read to represent a significantly larger fraction of the genome, which help bridge repetitive regions and resolve structural variants[50]. Various computational tools for long-read *de novo* assembly have been developed to address the relatively higher error rate and to accommodate the computational challenges associated with increased read lengths. Similar to SV callers, many long-read *de novo* assemblers are tailored for specific sequencing chemistries and sometimes support hybrid assembly from multiple platforms[51–53]. Widely used long-read *de novo* assemblers include Canu[54–56], Flye[57,58], wtdbg2[59], Falcon[31,60,61], Shasta[62], hifiasm[63], and Verkko[64], among others, many of which have shown success in generating personal assemblies with quality similar to, or even better than, existing human reference genomes. In particular, the T2T CHM13 assembly[65], which was based on a custom approach using both PacBio and ONT sequencing, among other technologies, and published in 2022 by the Telomere-to-Telomere (T2T) consortium, became the first complete human genome sequence without unresolved gaps.

The human genome is diploid, with two sets of 23 chromosomes, each of which is inherited from one of the parents. Ideally, a human *de novo* assembly should faithfully represent both alleles of any given locus, while preserving information about which alleles belong to the same haplotype. However, due to the technical difficulties in differentiating highly similar heterozygous alleles, especially in repetitive regions, early de novo assembly efforts mostly focused on producing mixed-haplotype assemblies, also referred to as consensus assemblies or haploid assemblies, in which only one arbitrarily selected allele is assembled for a given locus[1,56,57,62]. The T2T CHM13 assembly, on the other hand, circumvented this problem by sequencing a hydatidiform mole which is essentially haploid[65]. As the sequencing technology and assembly algorithms matured, it became feasible to produce high-quality haplotype-resolved assemblies, also referred to as phased assemblies or diploid assemblies, which contain both alleles of a locus, often with the help of phasing information from parental sequencing data[54,66], Hi-C[67] or Strand-seq[68].

## DNA modifications

DNA modifications are chemical modifications that yield various modified bases, such as 5-methylcytosine (5-mC), 5-hydroxymethylcytosine (5-hmC), and *N*$^6$-methyladenine (6-mA), in place of the corresponding unmodified bases. As an important focus of epigenetics research, DNA modifications have been shown to play a crucial role in regulating gene expression, genome stability, development, and various biological processes[69]. In short-read HTS, information about DNA modification is erased by the localised PCR amplification process, and by the inability of known DNA polymerases to synthesise modified DNA. Therefore, specialised chemical or enzymatic methods, such as bisulfite sequencing[70] and TAB-Seq[71], have been developed to specifically label modified bases so that their locations can be revealed through HTS.

In contrast, PacBio and ONT sequencing platforms are able to sequence single DNA molecules without amplification, thus preserving epigenetic modifications in the input DNA. In PacBio sequencing chemistry, the fluorescence signal is characterised not only by the emission spectra (i.e., colours), but also by the kinetics (i.e., timing) of fluorescence pulses. This kinetics information proved useful for inferring methylated bases including 6-mA, 5-mC and 5-hmC, as methylated bases are processed at different rate by the DNA polymerase, compared to their unmodified counterparts[72]. As such differences are subtle, methylation detection currently relies on the aggregation of kinetics signals using the CCS chemistry and is not available for the CLR chemistry[72,73].

As previously described, in ONT sequencing chemistry, the raw electrical current signal generated by each nanopore depends on the chemical structure of the base passing through, which allows each modified base to have its own signature that can be decoded using hidden Markov models[74]. The accuracy of ONT methylation calling was historically burdened by the low accuracy of base calling, but has significantly improved since the release of the ONT R10 chemistry[75,76].

### Full-length transcriptome

The median length of human gene transcripts is estimated to be between 2000 and 4000 nucleotides[77]. Unlike short-read sequencing, of which each read typically only captures a fraction of an exon, long-read sequencing allows for full-length transcript sequencing, capturing complete RNA molecules as individual reads. This enables more accurate identification of alternative splicing events, isoforms, and gene fusions with higher accuracy, providing a deeper understanding of gene regulation, transcript diversity, and functional implications[78].

PacBio Iso-Seq, which is HiFi sequencing applied to full-length cDNA, has shown success in identifying novel isoforms[79,80]. Taking advantage of the fact that HiFi reads are typically several times longer than mRNA transcripts, PacBio MAS-Seq optimises throughput for single-cell RNA sequencing by concatenating multiple transcripts from the same cell during library preparation, each which are then retrieved computationally by breaking the reads into corresponding segments[81,82].

On the other hand, thanks to its DNA polymerase-independent chemistry, ONT is uniquely capable for direct sequencing of native RNA strands, eliminating sequencing biases associated with reverse transcription and PCR amplification, while also unlocking the possibility to detect RNA modifications such as $N^6$-methyladenosine (m$^6$A) and 5-methylcytosine (5-mC)[83,84].

### 1.1.4. Research objectives

The overarching objective of this project is to explore possibilities of personalised genomic and functional analyses, enabled by the recent emergence of revolutionising sequencing technologies. We started from the blood sample of one healthy consenting participant, whom we refer to as HV31, and generated a broad set of genomic (long-read sequencing, short-read sequencing and genomic mapping) and functional (ATAC-Seq, ChIP-Seq, RNA-Seq) data. From there, we aimed at building a high-quality personal *de novo* assembly, which would not only reveal genetic variants that are challenging for previous methods, but also serve as the reference genome for joint interrogation of DNA modifications, chromatin accessibility and histone modifications, and mRNA transcription.

## 1.2. Results

### 1.2.1. Overview of genomic and functional data collected for HV31

Over the course the project, we generated genomic and functional data using multiple platforms using blood sample from a single individual identified as HV31 (Table 1). HV31 was recruited as a healthy female volunteer and identified as having European ancestry. Based on the time of collection and type of platform, the datasets available for HV31 can be categorised into three groups: 2019 genomics data, 2019 functional data, and 2022-23 genomics data. A large portion of analyses work in the HV31 project were conducted in the period between September 2019 and August 2021, before the 2022-23 genomics data became available.

| Name | Platform | Cell type | Coverage depth | Read length |
|---|---|---|---|---|
| HiFi-2019 | PacBio Sequel II Circular Consensus Sequencing (CCS) | CD14$^+$ | 12.3× | N50 = 12.7 kb |
| CLR-2019 | PacBio Sequel II Continuous Long Reads (CLR) | CD14$^+$ | 35× | N50 = 25.9 kb |
| ONT-2019 | Oxford Nanopore PromethION R9.4.1 chemistry | CD14$^+$ | 63× | N50 = 10.9kb |
| Bionano | Bionano Saphyr Direct Label and Stain (DLS) optical mapping | PBMC | 152.7× | N50 = 216.4 kb |
| MGI-standard | MGI DNA nanoball sequencing (DNBSEQ), StandardMPS chemistry, standard WGS library | PBMC | 56.8× | 100 bp paired-end |
| MGI-CoolMPS | MGI DNA nanoball sequencing (DNBSEQ), CoolMPS chemistry, standard WGS library | PBMC | 56.9× | 100 bp paired-end |

| | | | | |
|---|---|---|---|---|
| MGI-stLFR | MGI DNA nanoball sequencing (DNBSEQ), StandardMPS chemistry, single-tube long-fragment read (stLFR) WGS library | CD14$^+$ | 51.3× | 100 bp paired-end |
| 10X | 10X Linked-Read sequencing | CD14$^+$ | 40.2× | 150 bp paired-end |
| Illumina | Illumina PCR-free short-read sequencing on NovaSeq 6000 | PBMC | 44.2× | 151 bp paired-end |
| RNA-Seq | Illumina NovaSeq 6000 | CD4$^+$, CD8$^+$, CD14$^+$, CD19$^+$ | - | 75 bp paired-end |
| ATAC-Seq | Illumina NovaSeq 6000 | CD4$^+$, CD8$^+$, CD14$^+$, CD19$^+$ | - | 75 bp paired-end |
| ChIP-H3K4me3 | Illumina NovaSeq 6000 | CD4$^+$, CD8$^+$, CD14$^+$, CD19$^+$ | - | 75 bp paired-end |
| ChIP-H3K27ac | Illumina NovaSeq 6000 | CD4$^+$, CD8$^+$, CD14$^+$, CD19$^+$ | - | 75 bp paired-end |
| HiFi-2022 | PacBio Sequel IIe CCS | PBMC | 12.6× | N50 = 17.0 kb |
| HiFi-2023 | PacBio Revio CCS | PBMC | 58.8× | N50 = 15.9 kb |
| ONT-2022 | Oxford Nanopore PromethION R10.4.1 chemistry | PBMC | 68× simplex<br>3.7× duplex | N50 = 27.7 kb |

Table 1. Overview of genomic and functional data collected for HV31.

## The 2019 genomic data

The 2019 genomic data were collected between 2018 and 2019, shortly after the recruitment of the HV31 individual. Sample processing and sequencing was primarily the work of Antony Cutler and Andrew Brown, supervised by Prof John Todd and Prof Julian Knight. Hannah Roberts and Gavin Band contributed to the pre-processing of sequencing data. By the time I joined the project in September 2019, the 2019 genomic data had been fully collected and ready for analyses.

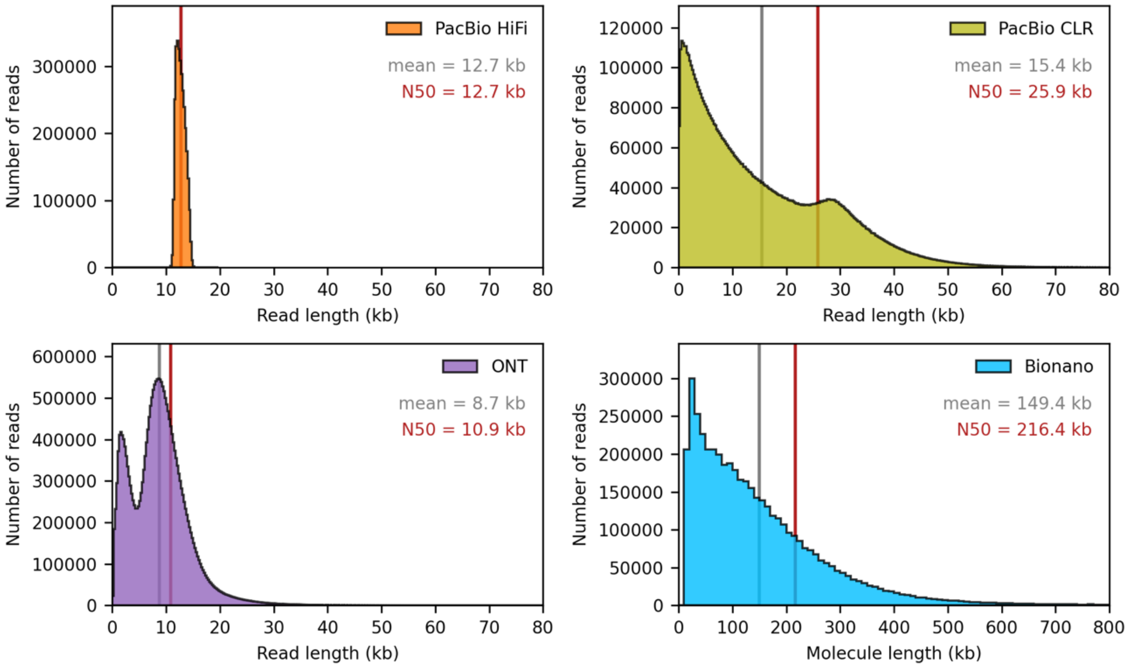


**Figure 1. Read/molecule length distribution of PacBio HiFi, PacBio CLR, ONT and Bionano optical mapping datasets in 2019 genomic data.**

Red and grey vertical lines denote the N50 and mean read/molecule length for each dataset, respectively. N50 was defined as 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.

The 2019 genomic data were generated to cover a broad range of long-read and short-range sequencing platforms, which were considered complementary to each other in terms of read length, accuracy and coverage depth (Table 1 and Figure 1). Specifically, 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 Stain (DLS) optical mapping (152.7× coverage by imaged molecules) data were collected. 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. To minimize the impact of cell-specific events including V(D)J recombination and somatic hypermutation and enable accurate assembly of the germline genome in immunoglobulin and T cell receptor regions, all long-read and linked-read datasets were collected from CD14$^+$ monocytes isolated from peripheral blood mononuclear cells (PBMCs) with antibody-conjugated beads. Bionano optical mapping data, as well as several short-read datasets, were generated directly from PBMCs.

The 2019 genomic data have been deposited to the European Genome-Phenome Archive (EGA), under the accession number EGAS00001005046.

### The 2019 functional data

Apart from long-read and short-read genomic data, which were useful for detecting genetic variants, the HV31 project aimed at investigating the functional impact of genetic variants, with a focus on immune cells. Therefore, transcriptome (RNA-Seq), chromatin accessibility (ATAC-Seq), and histone modification (ChIP-Seq; H3K4me3 and H3K27ac) data were also collected for CD4$^+$, CD8$^+$, CD14$^+$, and CD19$^+$ cells from PBMCs (Table 1). These functional datasets were generated around the same time when the 2019 functional data were collected, primarily by Antony Cutler and Andrew Brown, and preprocessed by Hannah Roberts and Gavin Band.

### The 2022/23 genomic data

Between 2019 and 2023, the two leading companies in long-read sequencing, PacBio and Oxford Nanopore, developed multiple generations of their respective long-read sequencing platforms, and achieved considerable improvements in terms of cost, throughput and accuracy through innovations in chemistry and bioinformatics. In particular, the PacBio Sequel IIe[85] and Revio[86] platforms improved throughput and cost efficiency compared to the previous Sequel II platform, while the ONT R10 chemistry were reported to yield more accurate reads compared to the previous R9 chemistry[32,87,88]. The 2022/23 genomic data were generated in collaboration with PacBio and Oxford Nanopore, separately, to evaluate these technical improvements and explore opportunities to build more complete and accurate personal genomes that harness these new platforms. Genomic sequencing data based on PacBio Sequel IIe, PacBio Revio, and Oxford Nanopore R10.4.1 chemistry platforms were collected (Table 1). All datasets were generated from frozen PBMCs.

## 1.2.2. A mixed-haplotype assembly of immune system regions

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

Eight genomic regions were selected as the focus of this assembly, 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 2). Immunoglobulin and T cell receptor regions were defined based on NCBI RefSeq locus definitions[89]. HLA and KIR regions were defined based on previously published gene ranges[14,90]. All regions were specified using GRCh38 coordinates, with an additional 1 Mb flanking sequence added to both sides (detailed on page 123). In the IGK region, I 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. The least reference sequence complexity was observed in the T cell receptor α, δ and γ regions (which contain < 2% repeat sequence and no listed alternate haplotypes). Meanwhile, the regions encoding immunoglobulin subunits contain the highest levels of duplication. Previous analyses[91,92] have also demonstrated significant structural diversity among known haplotypes in these regions.

| **Name** | **Coordinates and length** | **# core gene** | **% repeat** | **% SD** | **# gap** | **# alt. hap.** | **% novel** | **# patch** |
|---|---|---|---|---|---|---|---|---|
| Immuno-globulin heavy chain (IGH) | chr14 104,586,437 - 107,043,718 (2.46 Mb) | 164 | 6.8 | 31.1 (0.5) | 0 | 2 | 44.7 | 0 |
| Immuno-globulin κ (IGK) | chr2 87,857,361 – 91,902,511 (4.05 Mb) | 84 | 22.7 | 44.8 (22.7) | 3 | 2 | 31.5 | 1 |
| Immuno-globulin λ (IGL) | chr22 21,026,076 - 23,922,913 (2.90 Mb) | 89 | 7.4 | 34.0 (15.3) | 0 | 3 | 47.4 | 0 |
| Human leukocyte antigen (HLA) | chr6 28,602,238 - 34,409,896 (5.81 Mb) | 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 (2.93 Mb) | 115 | 1.9 | 3.5 (0) | 0 | 0 | 37.6 | 0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| T cell receptor β (TRB) | chr7<br>141,299,011 -<br>143,813,287<br>(2.51 Mb) | 78 | 5.3 | 19.5<br>(9.0) | 1 | 2 | 34.2 | 1 |
| T cell receptor γ (TRG) | chr7<br>37,240,024 -<br>39,368,055<br>(2.13 Mb) | 22 | 1.3 | 3.1<br>(0) | 0 | 0 | 0.2 | 0 |
| Killer-cell immune-globulin-like receptors (KIR) | chr19<br>53,724,447 -<br>55,867,209<br>(2.14 Mb) | 10 | 4.7 | 12.9<br>(0) | 0 | 50 | 47.4 | 0 |

**Table 2. Overview of eight selected immune system loci in GRCh38.**

% repeat, the proportion of repetitive DNA calculated as the proportion of 31-mers that are repeated at least once. % SD, the percentage of the region that is annotated as lying in a segmental duplication or (in brackets) a highly identical (≥ 95%) segmental duplication. # gap, the number of gaps (sequences of 'N' bases) in GRCh38. % novel, the percentage length of contigs that are new to GRCh38 rather than carried forward from GRCh37. # patch, the number of fix patches intersecting the locus in GRCh38 patch release 13. The GRCh38 segmental duplication, alternative haplotypes and fix patches annotations were downloaded from the UCSC Table Brower[93] based on the `genomicSuperDups`, `altSeqLiftOverPsl` and `fixSeqLiftOverPsl` datasets, respectively.

To confirm the effect of V(D)J recombination in PBMC-derived sequencing data, I aligned each sequencing dataset to the GRCh38 reference genome, and inspected the coverage depth patterns in the eight selected regions. I found that the coverage depth by PBMC-derived reads drops significantly around T cell receptor genes (Figure 2), consistent with an effect of V(D)J recombination in T cells which are the predominant cell type in PBMC, accounting for 70–85% to total cells[94]. Decrease of sequencing depths is not identified in immunoglobulin regions, presumably due to the relatively low fraction (5–10%) of B cells in PBMC[7]. Despite γδ T cells being rarer than αβ T cells, the coverage drop around T cell receptor γ and δ genes can be explained by the fact that the TRG and TRD loci undergo rearrangement in most αβ T cells in addition to γδ T cells[95,96].

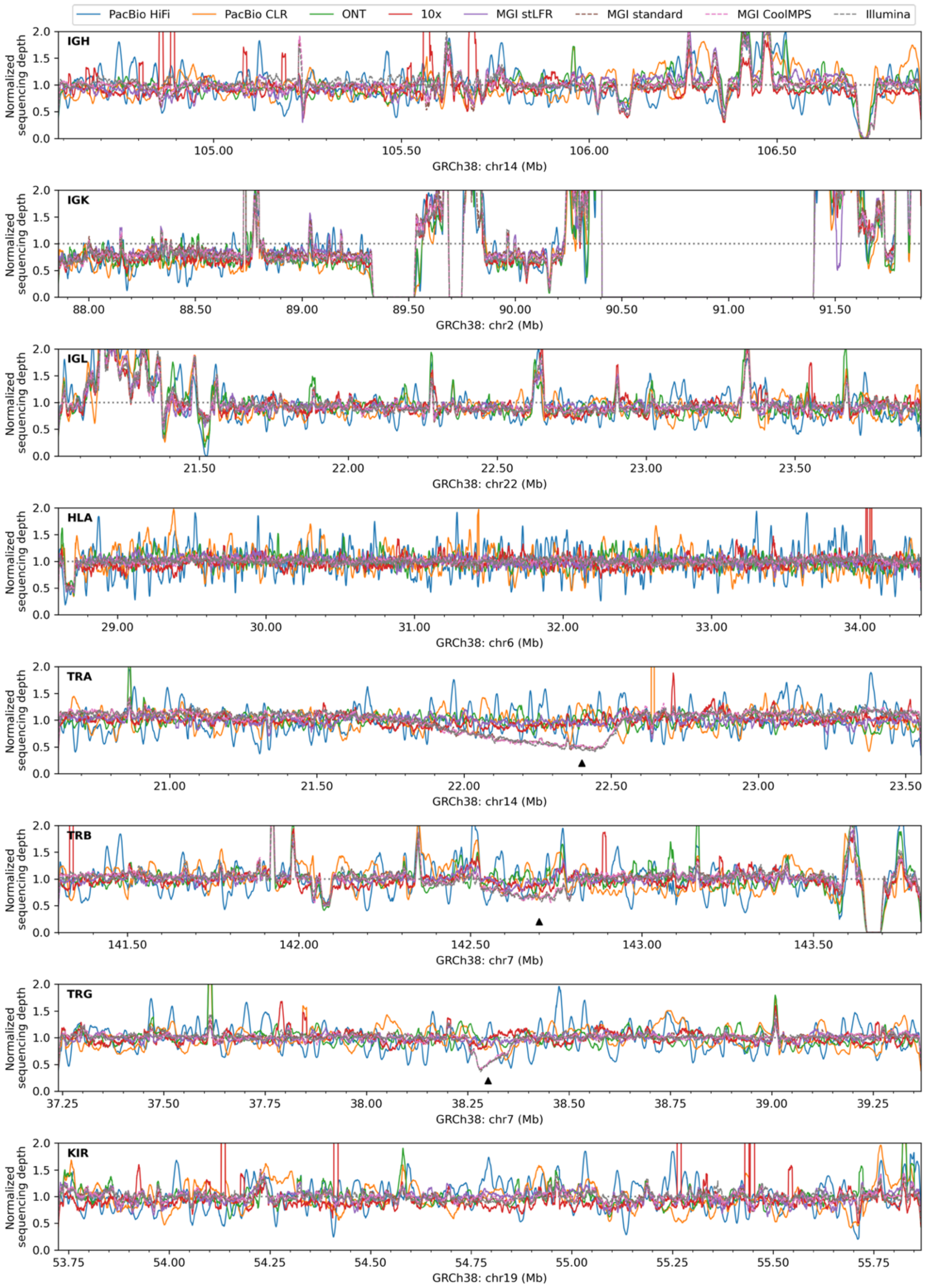


Figure 2. Depth of coverage of sequencing reads across 2019 genomics datasets, in the eight select regions.

Coverage depths of various sequencing datasets (colours) aligned to the GRCh38 reference genome for the eight selected regions (Table 2). Depths were normalized by the average depth across each region for each dataset. Areas with reduced coverage depths are highlighted with black triangles. Datasets generated using DNA from $CD14^{+}$ monocyte and PBMC samples are denoted with solid and dashed lines, respectively. Locations with zero coverage depths reflect gaps in the GRCh38 reference genome.

## Design of the *de novo* assembly pipeline

Our initial attempt to build a personal genome assembly for HV31 was conducted between September 2019 and July 2020, at a time when long-read sequencing, in particular PacBio HiFi sequencing (known then as circular consensus sequencing, CCS)[31] began to show its promise in resolving complex regions of the genome. Following the approach previously developed by Hannah Roberts, I started by testing several popular long-read *de novo* assembly tools available at that time, including wtdbg2[59], Flye[57], and Canu[56], all of which had been designed to generate mixed-haplotype assemblies, in which only one of the two inherited haplotypes, selected arbitrarily, is represented for any given locus. I generated local *de novo* assembly in the selected regions using HiFi-2019 data, and found that Canu yielded the most continuous assembly, with lowest fraction of assembly errors such as collapsed duplications or missing sequences (Figure 3). Meanwhile, Falcon[31,60,61], one of the few tools available then that supported the generation of haplotype-resolved assemblies, tended to generate highly fragmented assemblies for HV31, likely due to the relative low sequencing depth of the HiFi-2019 data, which was equivalent to around 6× coverage per haplotype.

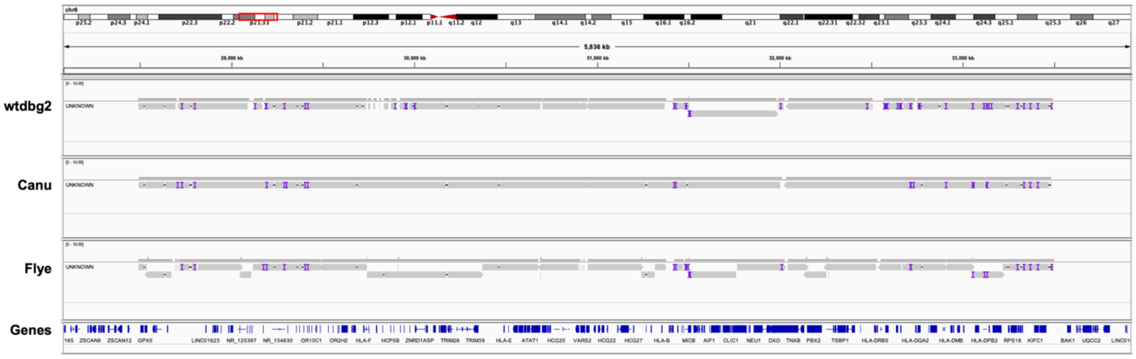


**Figure 3. wtdbg2, Canu and Flye assembles in the HLA region.**

Integrative Genome Viewer (IGV) screenshot showing wtdbg2, Canu and Flye local assembles in the HLA region aligned to the GRCh38 reference genome. Each assembly was generated using the same subset of HiFi-2019 reads that were extracted by alignment to the GRCh38 reference genome in the HLA region.

Considering the challenges involved in generating a haplotype-resolved assembly, in which both inherited haplotypes are represented and phased correctly, I took a pragmatic approach to generate an accurate representation of the eight regions in the HV31 genome, by developing a mixed-haplotype assembly for each region and listing heterozygous structural variants (SVs) that were not included in the assembly. The mixed-haplotype assembly, hereby referred to as the HV31-V1 assembly, and the SV list were used to jointly describe the HV31 genome. The overall pipeline consisted of four stages: initial assembly, scaffolding, gap closure, and polishing, as described below.

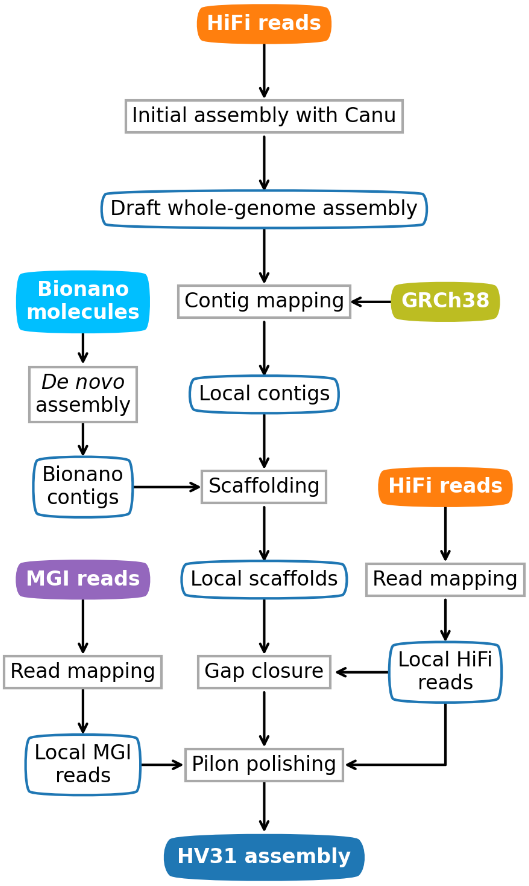


Figure 4. Workflow diagram for the HV31-V1 assembly.

Schematic representation of the HV31-V1 assembly workflow. Boxes with rounded corners represent input and output data and intermediate results. Boxes with square corners represent computational processes.

## Initial assembly

I used the Canu assembler, which had yielded the best results in previous evaluations, to produce a draft whole-genome assembly based on HiFi-2019 data. I 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 (Figure 5 and Figure 6), reflecting the unusual genomic complexity in these regions. The assembly also contained multiple shorter contigs aligning to the same location as longer contigs in some regions, which either represent assembly errors or genuine differences between haplotypes (Figure 6). These shorter contigs are hereby referred to as haplotigs.

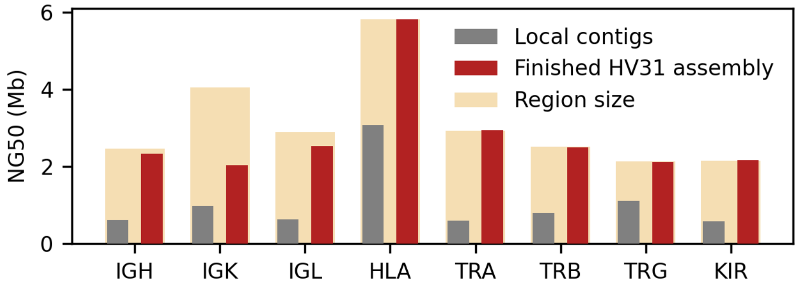


**Figure 5. Continuity of HV31-V1 assembly.**

Contig/scaffold continuity (NG50, y axis) for local contigs (grey) and finished HV31-V1 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.

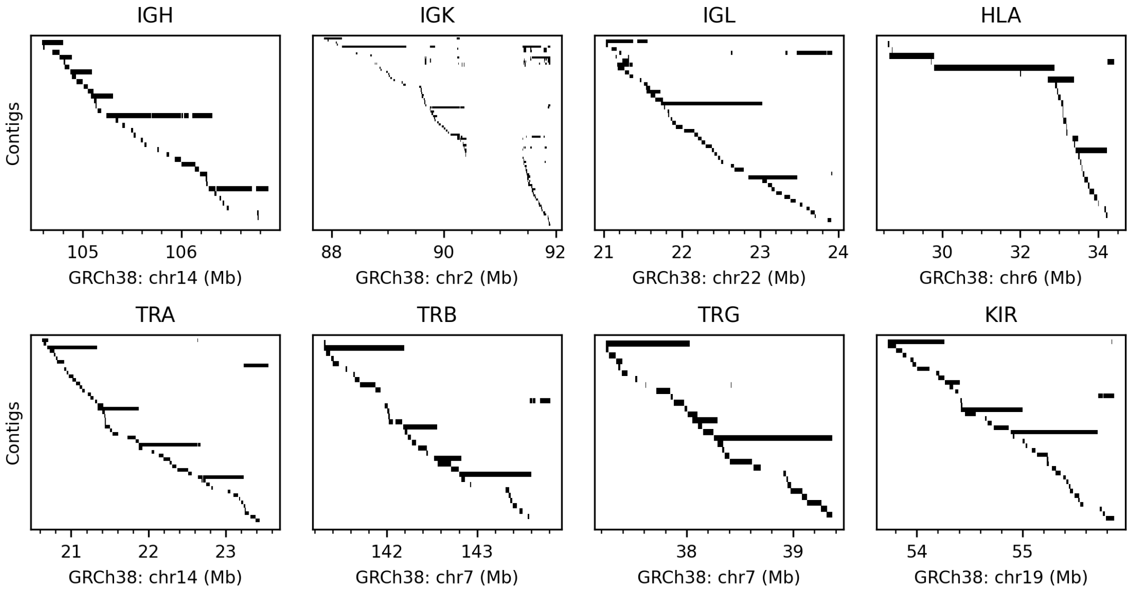


**Figure 6. Canu whole-genome assembly contigs aligned to GRCh38.**

Alignment of local contigs extracted from the draft HV31 whole-genome assembly produced by Canu for each of the eight selected regions (Table 2).

### Scaffolding

I 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 (Figure 1). I assembled these molecules using the proprietary Bionano Access software. As expected, the resulting Bionano contigs tended to be substantially longer than those in the draft whole-genome assembly (Figure 7), which allowed us to utilise the long-range information in Bionano contigs to order and orient the local contigs. I used the Bionano Solve algorithm to align the local contigs to the Bionano-assembled contigs and implemented a modified version of the BiSCoT algorithm[97] to generate local scaffolds (detailed on page 124). This process also removes or merges in haplotigs that can be effectively aligned to the local scaffolds. Finally, I confirmed that the remaining haplotigs represented substantial duplication of scaffolded contigs using a *k*-mer based method (detailed on page 124), 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 (Figure 8).

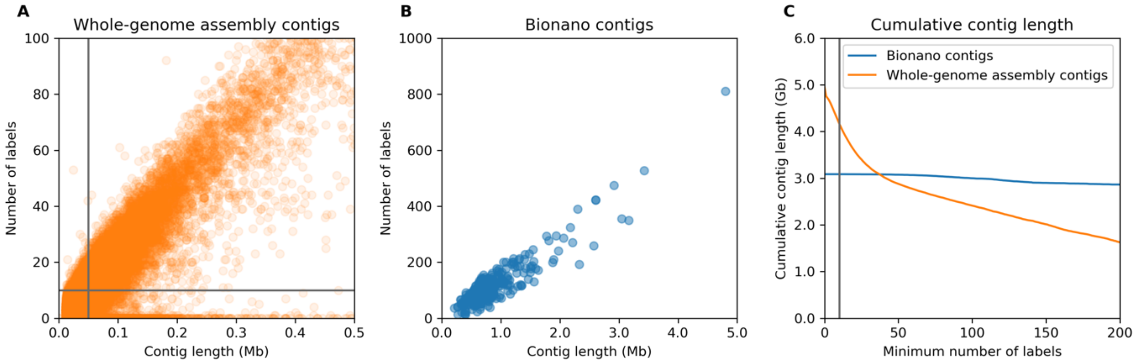


Figure 7. Comparison of contig lengths and counts of Bionano DLE-1 markers.

**(A-B)** Number of DLE-1 labels (y axis) plotted against contig length (x axis) for draft whole-genome assembly contigs **(A)** and Bionano contigs **(B)**. For reference, grey 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 grey vertical line denotes 10 DLE-1 labels.

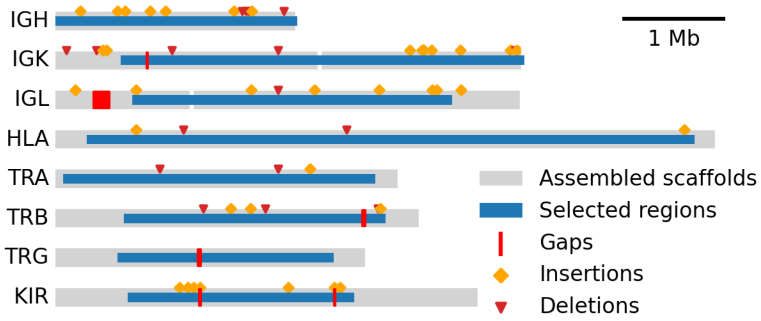


**Figure 8. Overview of assembled scaffolds in eight selected regions.**

Heterozygous SVs on the unassembled haplotype that are larger than 1 kb in size are shown as orange diamonds or red triangles. The assembled scaffolds (grey) were often larger than the predefined immune system regions (blue) in Table 2.

## Gap filling and polishing

The assembly quality was further improved by carrying out a gap-closing step (which fills in nucleotide information for missing bases between adjacent contigs in a scaffold) using TGS-GapCloser[98] applied to local HiFi reads, resulting in the closing of seven gaps. I also implemented a polishing step using Pilon[99] 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, the relevant reads were selected 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. This process left six gaps in the HV31 scaffolds (Figure 8), which lie outside regions aligning to core immune system genes but could potentially be improved with additional processing. The resulting assembly sequences that covered the eight selected genomic regions are referred to as the HV31-V1 assembly hereafter.

### Structural variant calling

The HiFi-2019, CLR-2019 and ONT-2019 datasets were used to call heterozygous SVs using the HV31-V1 assembly as reference. In brief, SVs were called separately from locally aligned HiFi-2019, CLR-2019 and ONT-2019 long reads using PBSV[42] (for HiFi and CLR) and Sniffles[43] (for HiFi, CLR and ONT). A computational approach based on unique *k*-mers[100] was used to refine read alignment before variant calling (detailed on page 126). 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 (Figure 9), as reported by SVanalyzer[101], which analyses the sequence information of each variant and identifies groups of compatible variants. As these numbers indicated, considerable discrepancy was observed between the individual SV calling approaches (Figure 10), reflecting the difficulty of aligning reads and calling SVs in paralogous regions.

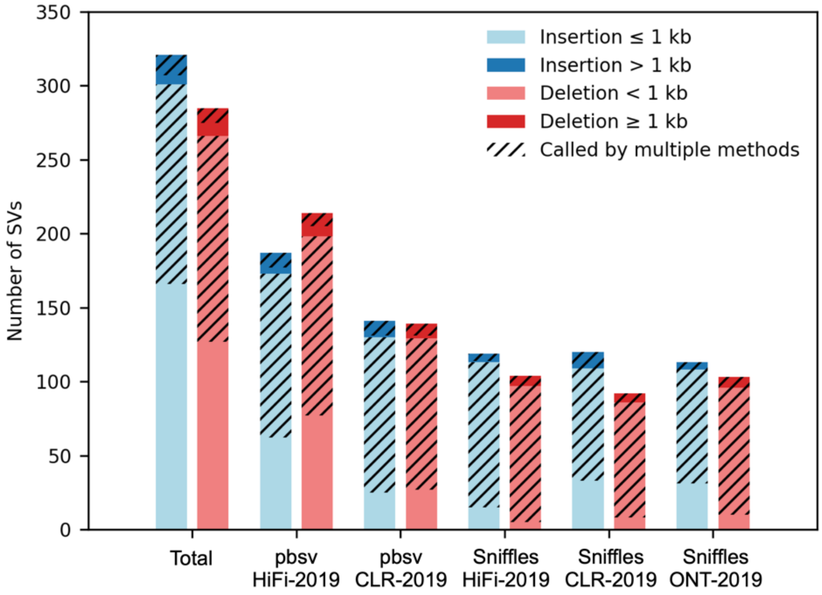


**Figure 9. Comparison of insertions and deletions identified by alignment-based structural variant calling methods.**

The number of insertions (blue) and number of deletions (red) identified by each combination of method and sequencing data (x axis) based on read alignment using the HV31-V1 assembly as the reference. SVs were broadly classified as insertions or deletions according to whether the alternative haplotype was longer or shorter than the HV31 haplotype. The number of SVs called by multiple methods, as identified by SVanalyzer[101], is indicated by shading.

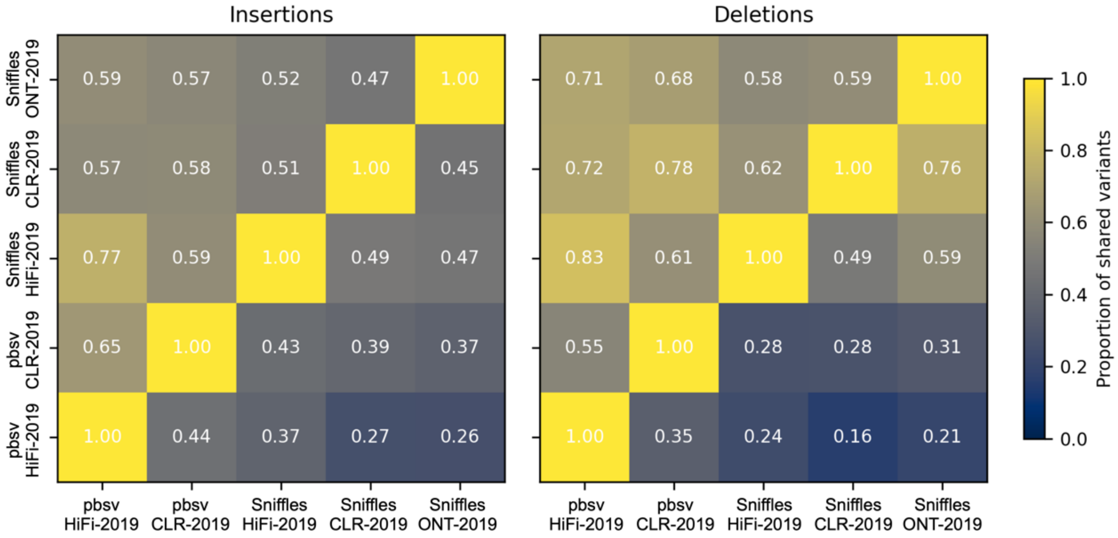


**Figure 10. Concordance of structural variants called by various methods.**

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[101].

## 1.2.3. Validation of the HV31-V1 assembly

### Comparing the HV31-V1 assembly with GRCh38

Below I use *k*-mer sharing dot plots[102] (Figure 11) to visually compare the HV31-V1 assembly with reference sequences. Each point in a *k*-mer sharing dot plot represents a short sequence of length *k* that is shared by both the reference sequence and the HV31-V1 assembly, with *k* = 50 for most such plots in this thesis. Patterns formed by these points provide a visualization of similarities and differences between the two sequences.

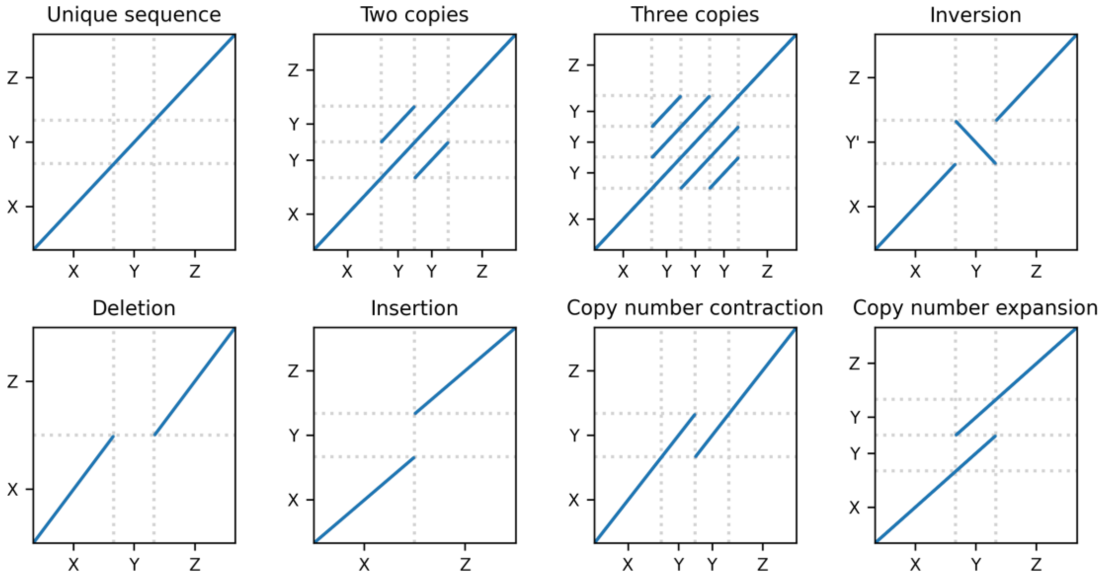


**Figure 11. Sequence duplications and structural variants demonstrated with *k*-mer sharing dot plots.**

Schematic examples of *k*-mer sharing dot plots demonstrating various sequence patterns. Each panel compares a reference sequence (x axis) with an alternate sequence (y axis). X, Y, and Z denote three sequence fragments that are mutually different. Y’ denotes the reverse complement of Y.

Comparing the HV31-V1 assembly with GRCh38 confirmed that the HV31-V1 assembly was highly accurate and complete for the eight regions, without apparent chimera sequences or missing sequences, apart from the six gaps described above (Figure 8). the HV31-V1 assembly contained two scaffold breaks at the IGK and IGL loci, both of which were located near long (≥ 100 kb) SDs that are highly identical (≥ 99%). In contrast, genomic loci with higher proportions of shorter, low-similarity SDs, such as the HLA and KIR, were fully resolved in the HV31-V1 assembly.

Close inspection of these plots revealed many large (≥ 1 kb) SVs that differ between GRCh38 and the HV31-V1 assembly. To systematically characterize these SVs, we aligned the assembly to GRCh38 and applied Assemblytics[103]. Assemblytics reported 145 SVs, 55 of which were ≥ 1 kb in size (Figure 8). 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 harboured the highest number (29) of SVs, followed by IGH (28) and IGK (24) regions.

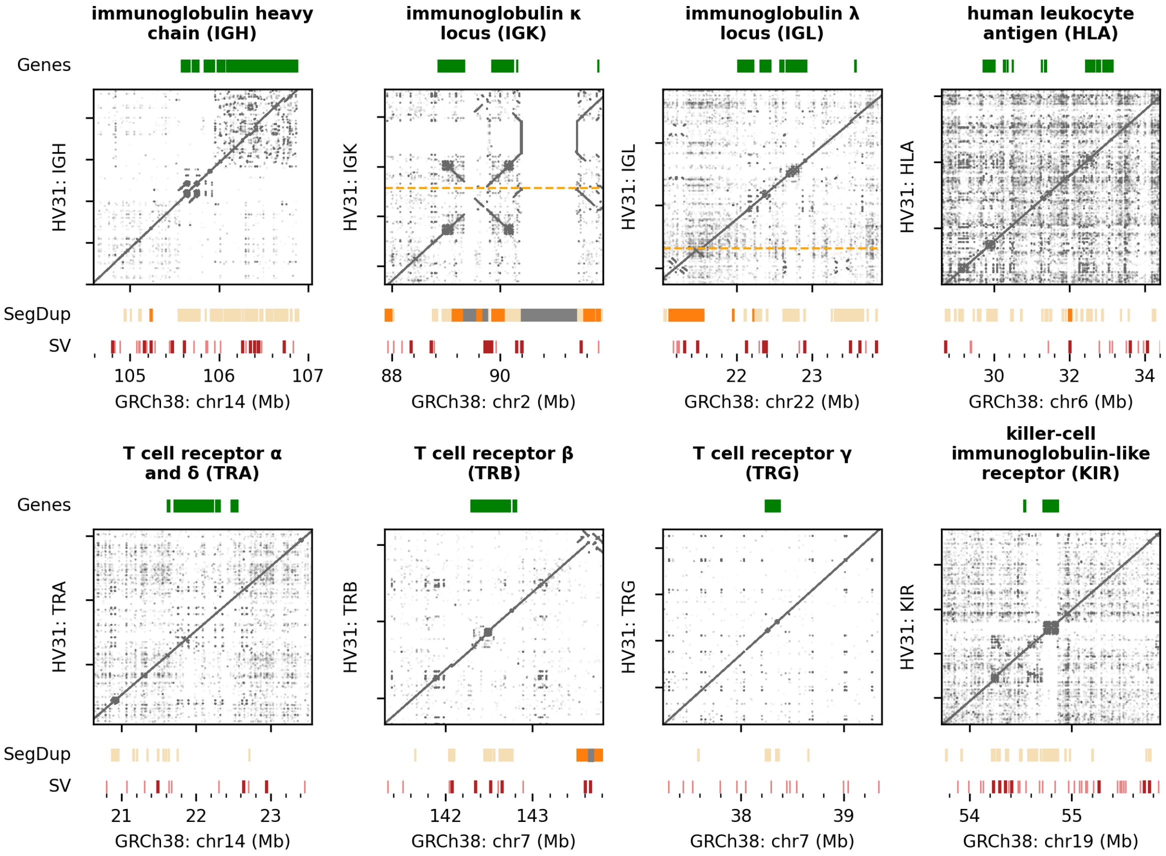


Figure 12. Comparing the HV31-V1 assembly with GRCh38.

*k*-mer sharing dot plots comparing the HV31-V1 assembly (y axis) with the GRCh38 reference (x axis) for the eight selected regions. Multiple scaffolds in the HV31-V1 assembly are separated with orange horizontal dashed lines. Gene, core genes of each locus. SegDup, segmental duplications defined as sequence fragments that are ≥ 1 kb in length and ≥ 90% identical to another fragment. Segmental duplications with identity ≥ 99% are highlighted in orange. Reference gaps are shown in grey. SV, structural variants detected in the HV31-V1 assembly relative to GRCh38. Structural variants larger than 1 kb are highlighted in dark red.

### Reference-free evaluation of structural errors

Given that the HV31-V1 assembly differed structurally from GRCh38, it was important to validate the structure of the HV31-V1 assembly, and identify assembly errors, without reliance on a reference sequence as the truth. Motivated by previous works[104,105], I considered raw sequencing reads for HV31, in particular those with high accuracy, a bias-free source of information for assembly validation. I utilised the correspondence between the number of occurrences of each $k$-mer in the genome (typically referred to as the copy number of the $k$-mer), and the number of occurrences of that $k$-mer in the sequencing reads (typically referred to as the multiplicity, depth, or frequency of the $k$-mer). Specifically, given a sequencing dataset of average coverage depth $D$, a $k$-mer with copy number $N$ is expected to have depth $x$ that equals $ND/2$, which allowed the inference of $N$ based on $x$, or *vice versa*[105,106].

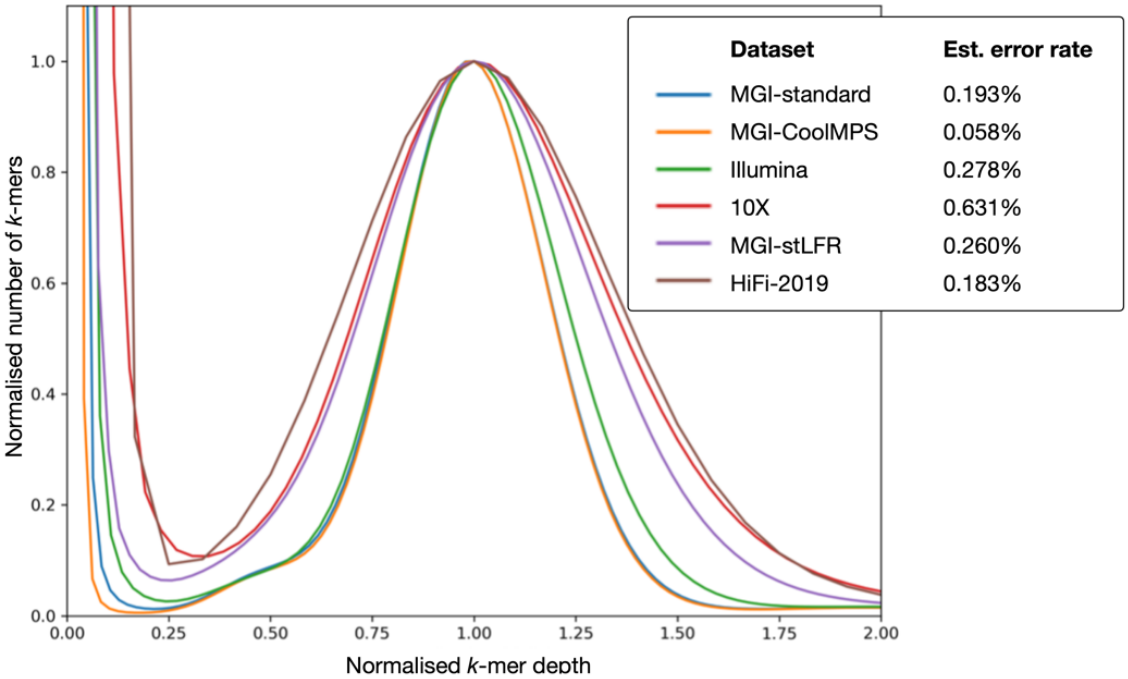


**Figure 13. Error rate estimation of PacBio HiFi and various short read datasets in 2019 genomic data using GenomeScope.**

The distribution of *k*-mer depths in each dataset after scaling depth values (x axis) and k-mer numbers (y axis) so that the peak of unique homozygous k-mers in each dataset overlap. *k* = 22. Est. error rate, the estimated per-base error rate of each dataset as estimated using GenomeScope[105].

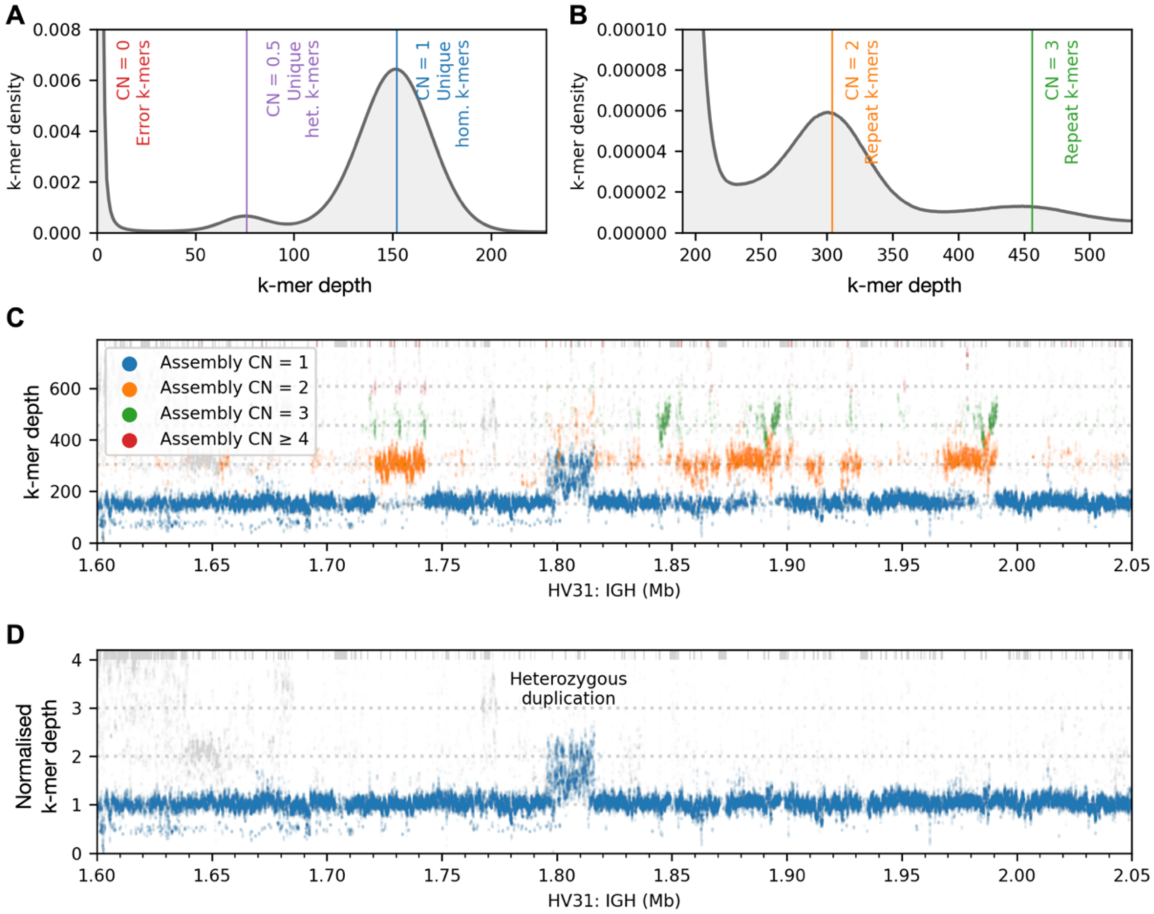


**Figure 14. Reference-free assembly validation based on *k*-mer depths.**

**(A-B)** Distribution of *k*-mer depths in the validation dataset. Vertical lines indicate local peaks. Het., heterozygous. Hom., homozygous; CN, the diploid copy number of a *k*-mer in the diploid HV31 genome. *k* = 31. The observed peak depth of one-copy *k*-mers (152×) are lower than the overall coverage depth of the validation dataset (262×) due to sequencing errors and the underrepresentation of both ends of short reads in 31-mers.

**(C)** Validation *k*-mer depths (y axis) plotted against *k*-mer position for a repeat-rich sequence fragment in the IGH region of the HV31-V1 assembly. Colours denote the copy number of a given *k*-mer in the HV31-V1 assembly.

**(D)** Normalised *k*-mer depths for the same sequence fragment as in **(C)**. Depth values were normalized by dividing by the peak depths of unique homozygous *k*-mers as shown in **(A)**. The location of a complex heterozygous duplication around *IGHV3-30* is labelled.

In **(C)** and **(D)**, *k*-mers with depths beyond the axis limits are stacked at the top of the plots. Non-specific *k*-mers, which exist both inside and outside the IGH region, are shown in grey.

Based on the principles described above, I used GenomeScope[105] to estimate the error rates of each short-read dataset in the 2019 genomics data, as well as the HiFi-2019 dataset, which confirmed that all of the evaluated datasets had error rates below 1% (Figure 13). These datasets were selected for the validation of the assembly. Collectively, the validation datasets had 262× genomic coverage depth, which was sufficient for inferring the copy number of each *k*-mer based on its depth (Figure 14, A-B). I assumed that any systematic discrepancies between the inferred copy number of each *k*-mer in the validation dataset, and the actual copy number of that *k*-mer in the HV31-V1 assembly, indicated either heterozygous variation or assembly errors. To leverage this, I plotted validation depths in comparison to scaffold depths (Figure 14, C-D). I found that most large regions showed validation *k*-mer depths compatible with assembly copy number, including several repetitive regions with higher *k*-mer copy numbers, indicating a good agreement between the HV31-V1 assembly and the validation dataset.

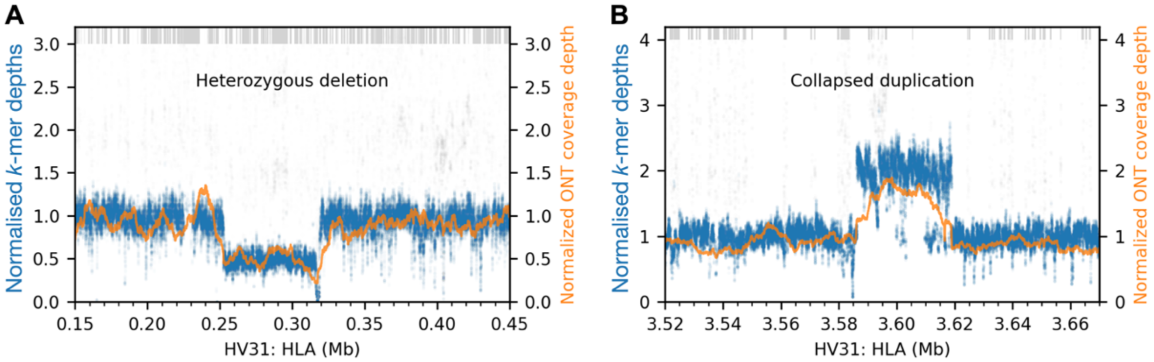


**Figure 15. Detecting heterozygous SVs and assembly errors from *k*-mer depths and coverage depth patterns in the HLA region.**

**(A)** A 63.9 kb heterozygous deletion in the HLA locus is revealed by the reduction in the normalised coverage depths of ONT-2019 reads (orange) and validation *k*-mers (blue).

**(B)** A collapsed duplication in the HLA locus is revealed by the elevation in the normalised coverage depths of ONT-2019 reads (orange) and validation *k*-mers (blue).

In **(A)** and **(B)**, *k*-mers with depths beyond the axis limits are stacked at the top of the plots. Non-specific *k*-mers, which exist both inside and outside the HLA region, are shown in grey.

Nevertheless, several locations showed larger discrepancies between assembly and validation data (Appendix Figure 1). I examined these locations in detail and found that many of the discrepancies reflected heterozygous structural variants. For example, I found a heterozygous deletion in the HLA region, manifesting as the *k*-mer depths in the validation data being approximately half of expected *k*-mer depth assuming homozygosity (Figure 15A), suggesting the existence of another, unassembled haplotype that did not have the corresponding sequence. Meanwhile, a small subset of these locations indicated potential structural errors in the HV31-V1 assembly, including a ~30 kb SD region in the HLA region, of which two copies should exist in each haplotype according to the *k*-mer depth data, while only one copy was found in the HV31-V1 assembly (Figure 15B), suggesting that the HV31-V1 likely failed to represent any one haplotype correctly for this SD. In addition, I found several stretches of elevated depths in the IGK region, but failed to fully confirm the assembly structure due to the lack of long-range information for resolving the long SDs in the region (Appendix Table 1). Overall, this analysis indicated that the HV31-V1 assembly was generally accurate even in complex regions, although several repeat-rich segments in the IGK and HLA regions likely still contained errors.

### Estimation of per-base error rates

In addition to the structural errors described above, small errors that involved the substitution, insertion or deletion of several bases, were analysed using a *k*-mer-based method similar to Merqury[104] (detailed on page 126). I estimated the presence 18.1 error bases per 1 Mb sequence in the HV31-V1 assembly averaged across the eight selected regions, which was comparable to other *de novo* assemblies based on PacBio HiFi data published at the time[55,66], and showed a significant improvement over the intermediate contigs generated before the polishing step, which had 85.3 error bases per 1 Mb sequence (Figure 16).

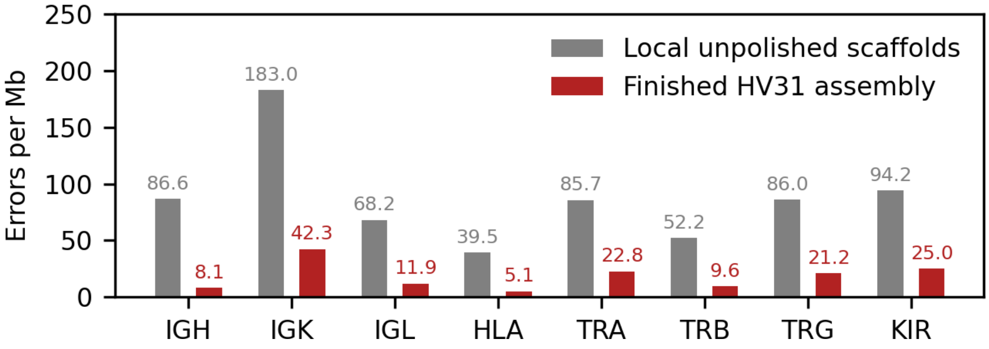


**Figure 16. Estimated per-base error rates in the HV31-V1 assembly.**

The number of errors per megabase in each region, before and after assembly polishing, estimated using a modified version of the Merqury algorithm[104] (detailed on page 126).

### Comparing the HV31-V1 assembly with other *de novo* assemblies

I compared the HV31-V1 assembly with several published *de novo* assemblies[54,55,57,61,62,65,65,66,100,107] based on long-read sequencing available at the time, by visualising contig alignment to GRCh38. I found that, in terms of completeness and continuity in the eight selected regions, the HV31-V1 assembly was comparable to exiting assemblies generated using various combinations of sequencing platforms and assembly tools, including the Telomere-to-Telomere (T2T) consortium assembly of the homologous CHM13 cell line[65,65], the first complete human genome assembly.

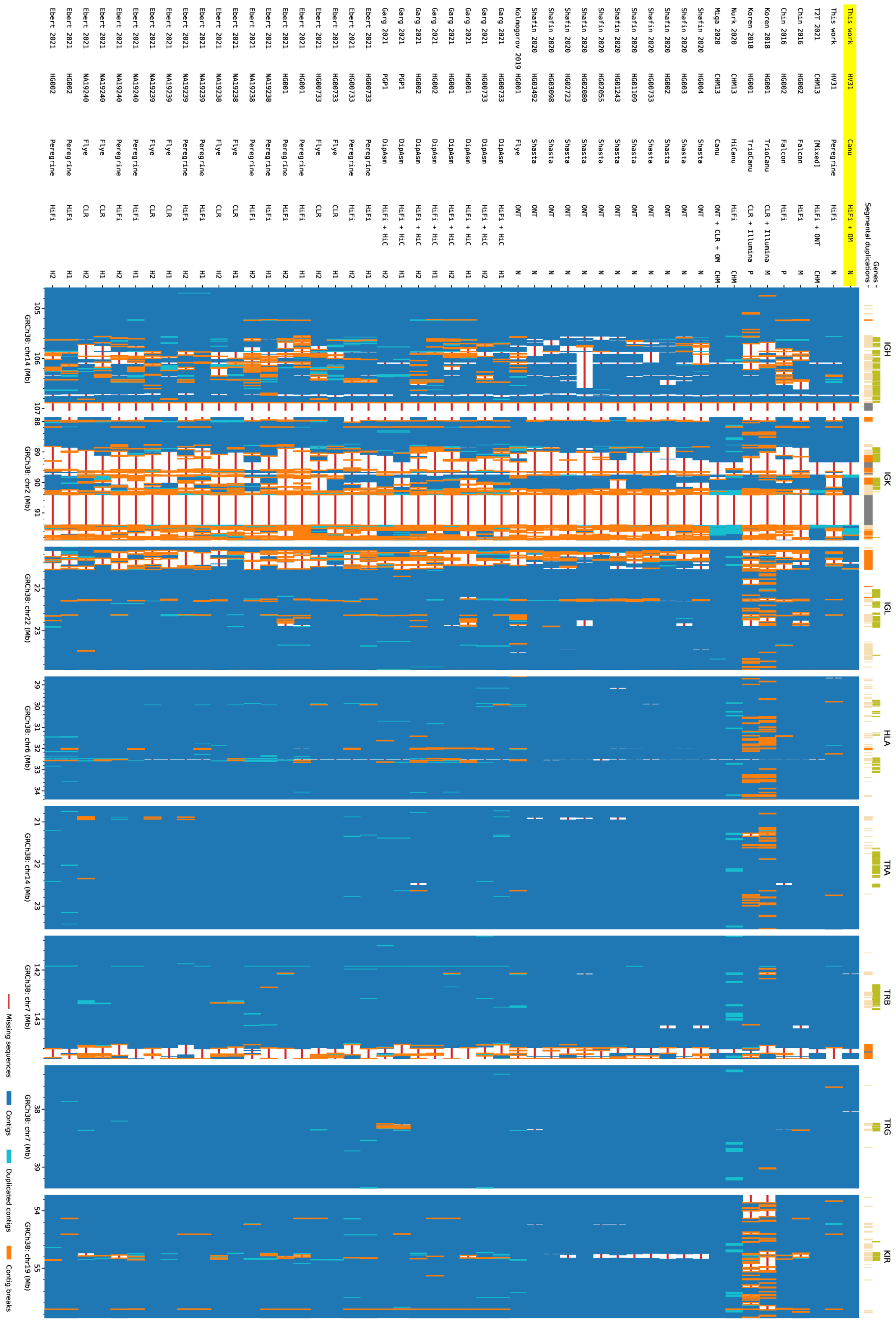


Figure 17. Comparison of published assemblies and alternative assembly methods in the eight selected regions.

Alignment of published assembly sequences to the GRCh38 reference genome. Duplicate contigs (cyan) were defined as shorter contigs that align within the span of a longer contig. Contig breaks (orange) were defined as endpoints of non-duplicate contigs. Regions of the GRCh38 reference that were not covered by the aligned assembly contigs denoted by red lines. Each assembly is labelled in the following order: publication, sample, key algorithms, key technology, and haplotype. OM, optical mapping. Hi-C, the Hi-C chromosome conformation capture approach. 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[108], Chin 2016[61], Koren 2018[54], Nurk 2020[55], Miga 2020[100], Shafin 2020[62], Kolmogorov 2019[57], Garg 2021[107], Ebert 2021[66]. Genes and segmental duplications are annotated above as in Figure 12.

## 1.2.4. Allelic and structural variants in the immune system

### A map of key immune gene allotypes

As detailed above, a 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, I used an alignment process to identify the best-matching allotype of each IG and TCR variable gene segment, and each HLA and KIR gene within the relevant IMGT or IPD database (Figure 18). The identified allotypes in the HLA region were compatible with HV31 HLA genotyping results (Table 3 and Table 4). Relative to GRCh38, the HV31-V1 assembly contain both insertions and deletions of gene sequence in the IGH, IGK, IGL and TRB regions, as well as allelic variation in all regions except TRG. A small number of genes that differed from the best matching IMGT allele, which may represent novel sequences. I note that HV31 gene allotypes also differ substantially from the GRCh37 assembly in the IGH region based on published results[91].

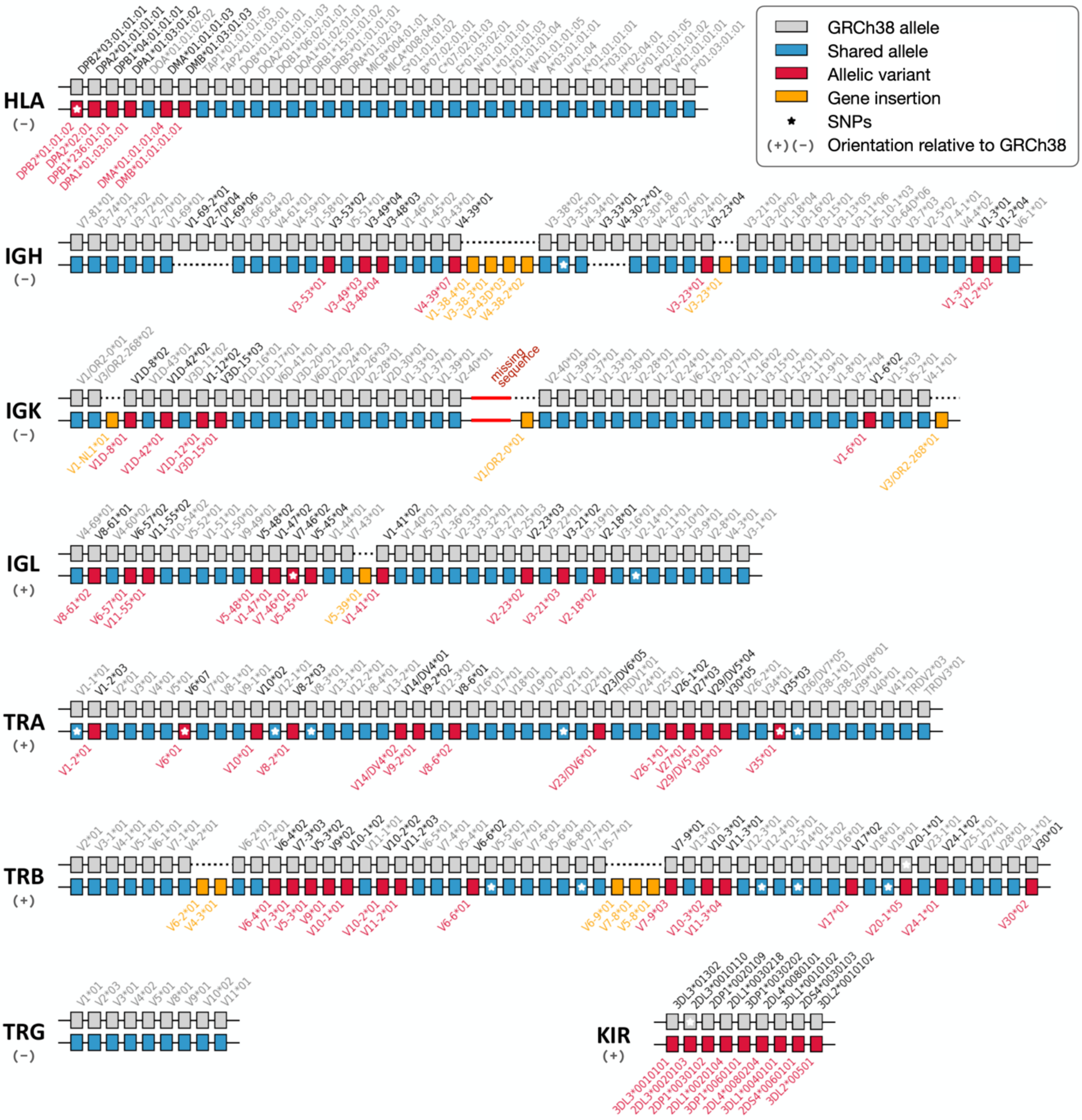


Figure 18. Allotypes of immune genes in the HV31-V1 assembly compared with GRCh38.

In immunoglobulin and T cell receptor regions, only V genes are shown. Genes in each region are arranged, from left to right, according to their relative order on the positive-sense strand, which may be either the plus (+) strand (with 5' terminus at the p arm) or the minus (-) strand (with 5' terminus at the q arm) of GRCh38. Pseudogenes are not shown. Allelic variants refer to genes where the best-matching HV31-V1 allele differs from the GRCh38 allele. Insertions refer to genes in the HV31-V1 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.

| Gene | Haplotype 1 | Haplotype 2 |
|---|---|---|
| HLA-A | 03:01 | 03:01 |
| HLA-C | 07:02 | 07:02 |
| HLA-B | 07:02 | 07:02 |
| HLA-DRB1 | 15:01 | 15:01 |
| HLA-DQA1 | 01:02 | 01:02 |
| HLA-DQB1 | 06:02 | 06:02 |

Table 3. HLA class I and class II genotyping results for HV31.

| Gene | Haplotype 1 | Haplotype 2 |
|---|---|---|
| HLA-DPA1 | 01:03 | 01:03 |
| HLA-DPB1 | 02:01 | 04:01 |

Table 4. HLA-DP genotyping results for HV31.

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 45 kb CNV involving *IGHV1-69* and *IGHV2-70*

Variation in the copy number of *IGHV1-69* and *IGHV2-70* has previously been reported[91]. Both genes are present in two copies in GRCh38. In the HV31-V1 assembly, I found only one copy of *IGHV1-69* and *IGHV2-70* remaining, as the result of a 45 kb copy number contraction relative to GRCh38 (Figure 19A). The earlier GRCh37 reference genome shared a similar haplotype in the IGH region, with only one copy of *IGHV1-69* and *IGHV2-70*. This haplotype has been suggested to be more common worldwide than the GRCh38 haplotype[91,109] and comparison to validation *k*-mers indicates it is homozygous in HV31 (Figure 14). This CNV manifested as a coverage gap in aligned HiFi-2019 reads aligned to GRCh38 (Figure 19B), suggesting the feasibility of accurate detection using alignment-based methods.

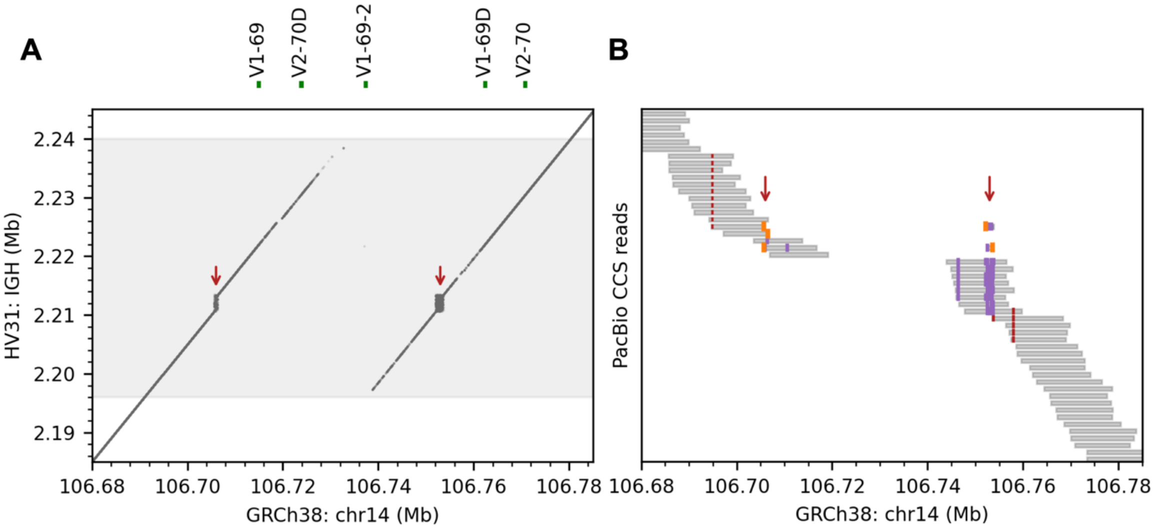


Figure 19. A tandem repeat within a 45 kb CNV involving *IGHV1-69* and *IGHV2-70.*

**(A)** k-mer sharing dot plot comparing the HV31-V1 assembly with GRCh38 in the IGH region, showing a 45 kb CNV (grey). Tandem repeats that were incorrectly assembled in a single copy are highlighted with red arrows.

**(B)** HiFi-2019 reads aligned to GRCh38 in the same region as in **(A)**. The coverage gap corresponded to the 45 kb CNV in **(A)**. The tandem repeats in **(A)** are highlighted with red arrows. Insertions, deletions, and alignment breakpoints are shown in purple, red, and orange, respectively.

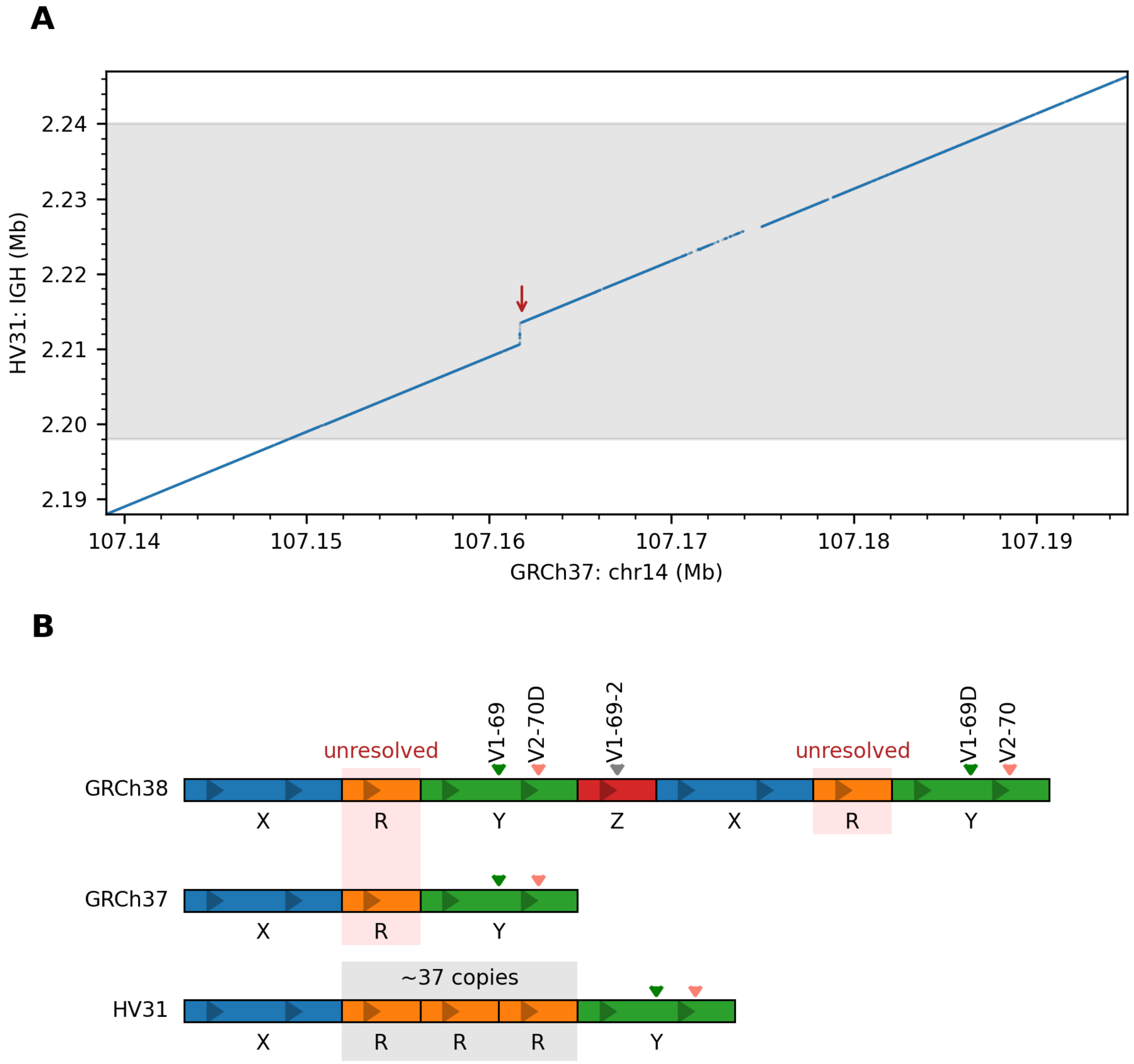


Figure 20. GRCh38 and GRCh37 represent different haplotypes for IGHV genes.

**(A)** *k*-mer sharing dot plot comparing the HV31-V1 assembly (y axis) with GRCh37 (x axis) in the IGH region. Similar to HV31, GRCh37 has only one copy of *IGHV1-69* and *IGHV2-70* genes. Locations corresponding to the 45 kb CNV and the tandem repeats in Figure 19 are highlighted with grey shade and a red arrow, respectively.

**(B)** Schematic representation of GRCh38, GRCh37 and the HV31-V1 assembly near the *IGHV1-69* and *IGHV2-70* genes. Fragment R denotes the tandem repeats in Figure 19, which was fully assembled in HV31.

Within this 45 kb CNV, I also noticed a 2.66 kb cluster of tandem repeats with a 59-mer motif (Figure 19 and Figure 20) that was not correctly assembled in either GRCh37 or GRCh38 (see GenBank sequence 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[110].

### A compound heterozygous CNV involving *IGHV3-30*

A second prominent feature of the IGH region is the loss of one copy of *IGHV3-30*. GRCh38 carries two copies of *IGHV3-30*, which are named *IGHV3-30* and *IGHV3-33*[109]. I found that *IGHV3-33* was removed in the HV31-V1 assembly, together with *IGHV4-31* (Figure 21A). However, inspection of Bionano contigs covering this region revealed a further unusual feature (Figure 21B): one of the two Bionano contigs covering this region contains a corresponding three-fold copy number expansion. To confirm this, I inspected validation *k*-mer depths in the surrounding area, and observed elevated depths compatible with a 3-fold expansion on the unassembled haplotype (Figure 14D). To confirmation this further, I fetched raw sequencing reads spanning this region, and observed a CLR-2019 read consistent with the expanded haplotype (Figure 22). These results indicate that the unassembled haplotype carries three copies of the region surrounding *IGHV3-30*, such that HV31 carries both a contraction and expansion of this region relative to GRCh38. This CNV was not called accurately by any of the SV calling methods I employed. I speculated that this type of copy number variation was particularly challenging for variant calling methods that depend on read alignment, considering the lack of overall diploid copy number changes of *IGHV3-30* genes.

The observation of this compound heterozygous CNV was also compatible with previous work[91] which reported this region as a hotspot for SVs, with the diploid copy number of *IGHV3-30* and related genes ranging from zero to six.

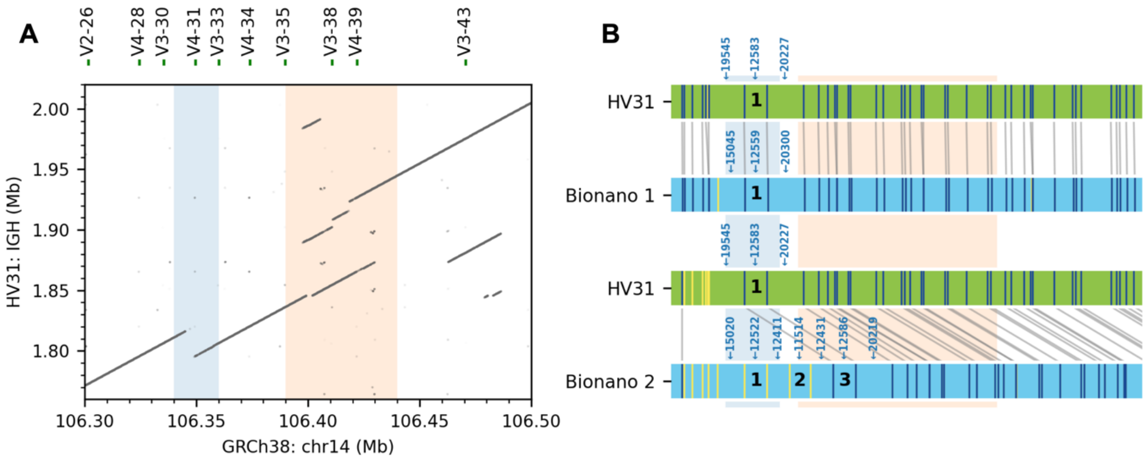


**Figure 21. Complex structural variation involving IGHV genes.**

**(A)** *k*-mer sharing dot plot showing complex structural variations found between *IGHV3-21* and *IGHV3-43*, including a 25 kb CNV (blue) and an 80 kb complex duplication event (orange).

**(B)** Alignment patterns of HV31-V1 assembly and two Bionano contigs covering the region shown in **(A)**. The four rows represent the HV31-V1 assembly (green horizontal bars) and two Bionano contigs (blue horizontal bars). Aligned and unaligned DLE-1 recognition markers are shown as dark blue and yellow vertical lines, respectively. Aligned markers are connected by grey lines. The approximate regions corresponding to the 25 kb CNV and the 80 kb complex duplication in **(A)** are highlighted in blue and orange, respectively. Locations of the repeat units of the 25 kb CNV are labelled with black numbers. Distances between adjacent DLE-1 markers are labelled with blue numbers for selected locations.

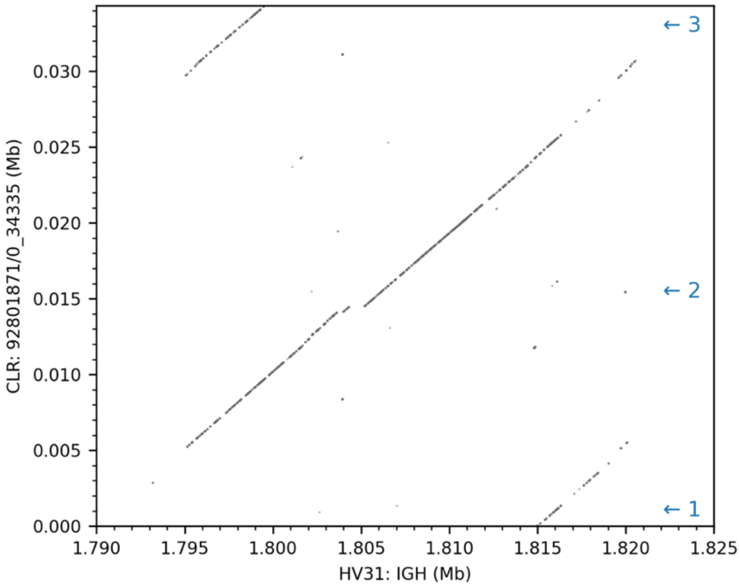


**Figure 22. A compound heterozygous CNV involving *IGHV3-30* confirmed by a CLR read.**

*k*-mer sharing dot plot comparing the CLR-2019 read with ID 92801871/034335 (y axis) with the HV31-V1 assembly (x axis). The read is consistent with the presence of a three-copy unassembled haplotype (Figure 21). Each copy of the repeat unit is annotated with a number and an arrow for clarity. *k* = 20.

### An 80 kb complex duplication involving multiple IGHV genes

The HV31-V1 assembly carried additional copies of *IGHV1-38*, *IGHV3-43*, *IGHV4-38* and *IGHV3-38* compared to GRCh38 (Figure 18), which were contained in an 80 kb duplication with a complex structure (Figure 21). Inspection of *k*-mer depth data implies this duplication is homozygous (Appendix Figure 1). This duplication was not called by any of the methods used to call SVs, which likely resulted from the difficulty in correctly aligning HV31 reads to GRCh38, given that the HV31 genome was considerably different from the GRCh38 sequence at this location. Consistent with this, the HiFi-2019 reads from this location displayed suspicious alignment patterns when mapped to GRCh38, while using the HV31-V1 assembly as the reference enabled the same set of reads to be aligned correctly (Figure 23).

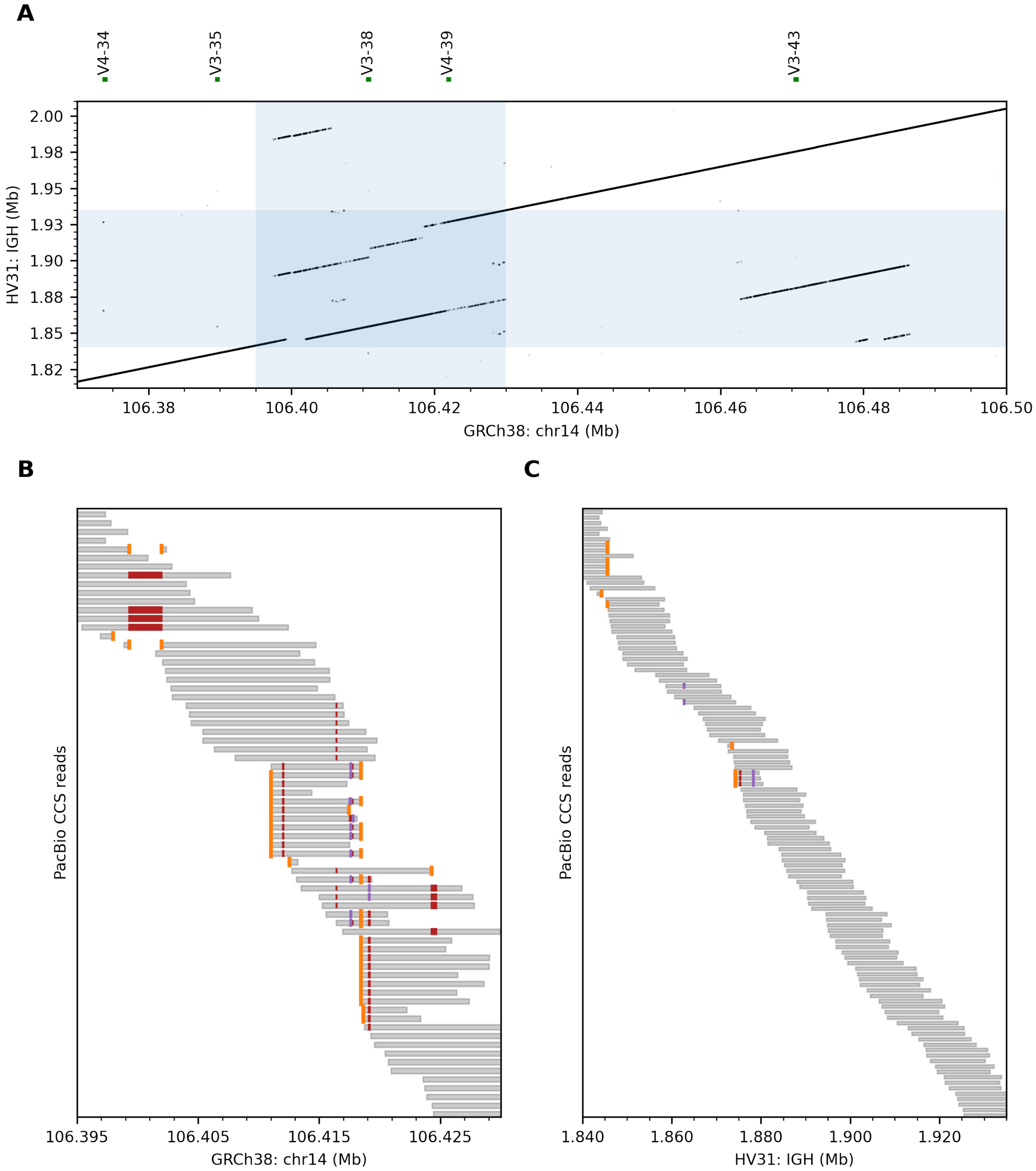


**Figure 23. Misalignment resulting from large structural rearrangements in the IGH locus.**

**(A)** *k*-mer sharing dot plot comparing the HV31-V1 assembly (y axis) with GRCh38 (x axis), showing the 80 kb insertion between *IGHV3-37* and *IGHV7-40* (highlighted in orange in Figure 21). The region further inspected in **(B)** and **(C)** is highlighted in blue.

**(B)** Alignment HiFi-2019 reads to GRCh38 in the region highlighted in **(A)**. Each row represents a read. Insertions, deletions, and alignment breakpoints are shown in purple, red, and orange, respectively.

**(C)** Same as **(B)**, with reads aligned to the HV31-V1 assembly.

### Large insertions incorporating novel TRBV genes

In the TRB region, I identified a ~11 kb homozygous insertion near *TRBV6-2* and another ~19 kb insertion near *TRBV5-7* (Figure 24A). Both insertions are supported by Bionano contigs (Figure 24B) and incorporated sequence fragments that are not found in GRCh38, with limited homology to adjacent sequences (Figure 24A). Comparison to *k*-mer validation implies both insertions are homozygous (Appendix Figure 1). Assemblytics[103] identified duplications at both locations but with inaccurate length and sequence content. The HV31 scaffold was consistent with an alternative reference sequence for the TRB locus (NCBI Reference Sequence: NG_001333.2) which was part of GRCh38 alternative haplotypes (Figure 25). By comparing NG_001333.2 with the GRCh38 primary sequence, I confirmed that the 11 kb insertion introduced two new genes (*TRBV4-3* and *TRBV6-2*) and a pseudogene (*TRBV3-2*), while the 19 kb insertion introduced three new genes (*TRBV6-9*, *TRBV7-8* and *TRBV5-8*; Figure 25).

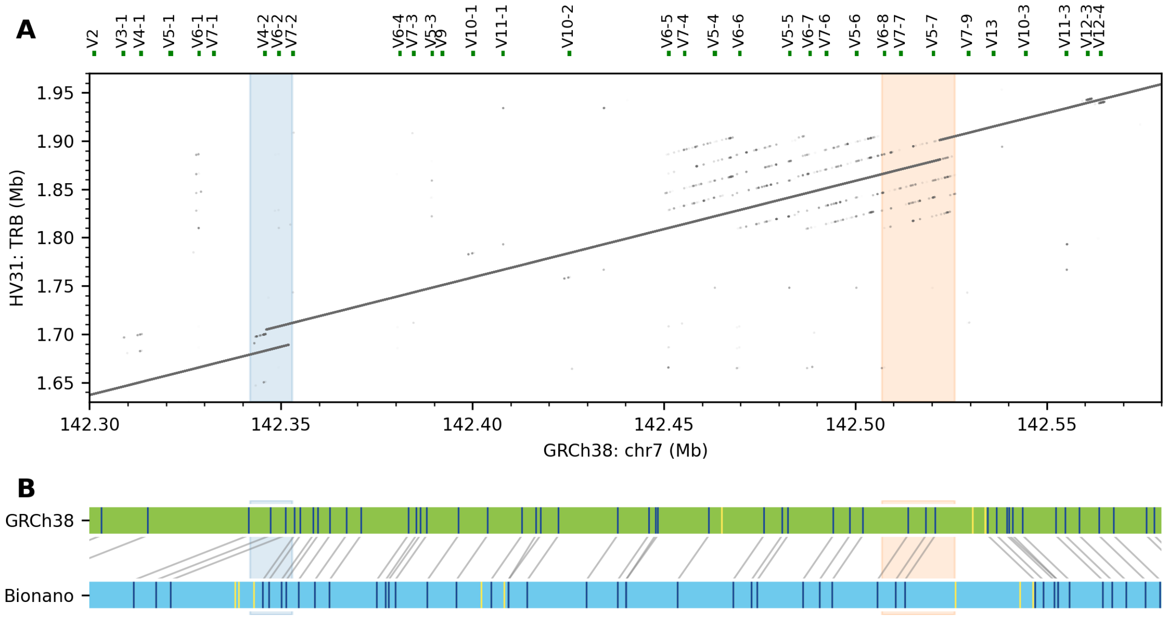


**Figure 24. Large insertions in the TRB region.**

**(A)** *k*-mer sharing dot plot comparing the HV31-V1 assembly (y axis) with GRCh38 (x axis), showing a 11 kb insertion (blue) and another 19 kb insertion (green).

**(B)** A Bionano contig (blue) aligned to GRCh38 (green), confirming the two insertions in **(A)**.

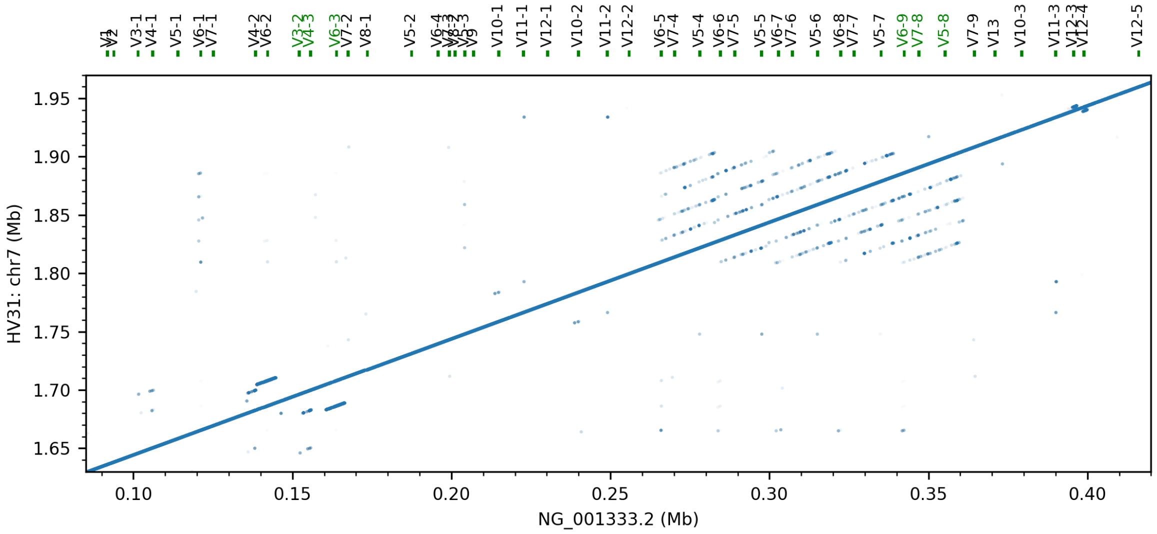


**Figure 25. The HV31-V1 assembly is consistent with NCBI RefSeq NG_001333.2.**

*k*-mer sharing dot plot comparing the HV31-V1 assembly (y axis) with the NG_001333.2 contig from NCBI RefSeq (x axis). TRBV genes not included in GRCh38 are highlighted in green.

### Reference gaps amidst complex segmental duplications resolved in the HV31-V1 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, such as heterochromatin regions which often consist of megabase-scale tandem microsatellite repeats, whose functional significance remains largely unexplored[111]. Eleven heterochromatin gaps exist in GRCh38, with estimated sizes ranging from 20 kb to 30 Mb, all of which, except the largest heterochromatin gap in chromosome X[100], remained unresolved in de novo assemblies, until the publication of the T2T CHM13 assembly[65]. This part of the analyses, in which I analysed the structure of a 600 kb heterochromatin sequence assembled near the IGK locus, was conducted at around same time when the first draft of the T2T CHM13 assembly[112] was released.

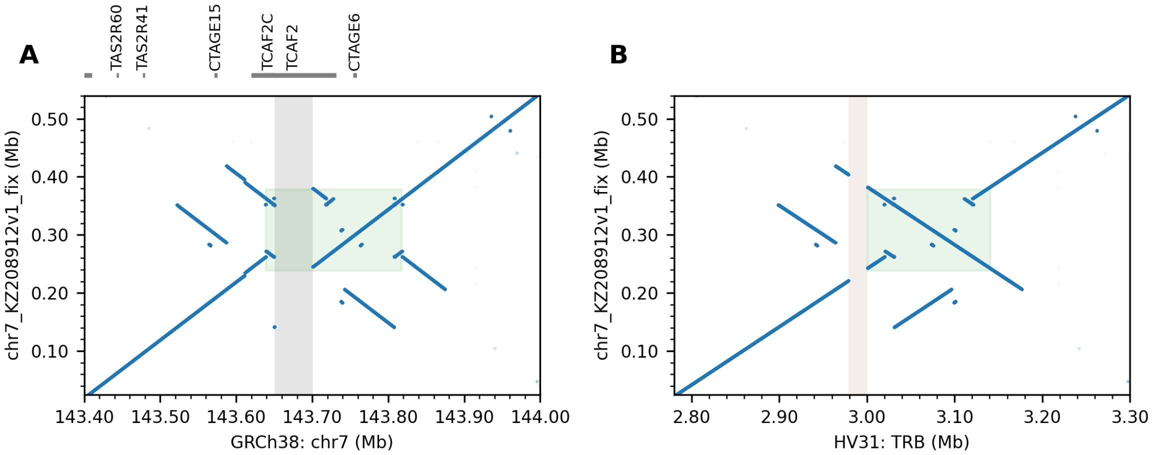


Figure 26. A 50 kb GRCh38 gap flanked by segmental duplications closed in the HV31-V1 assembly.

**(A)** *k*-mer sharing dot plot 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-V1 assembly (green region) and a 50 kb gap in GRCh38 (grey region) which is closed in the HV31-V1 assembly.

**(B)** The HV31-V1 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).

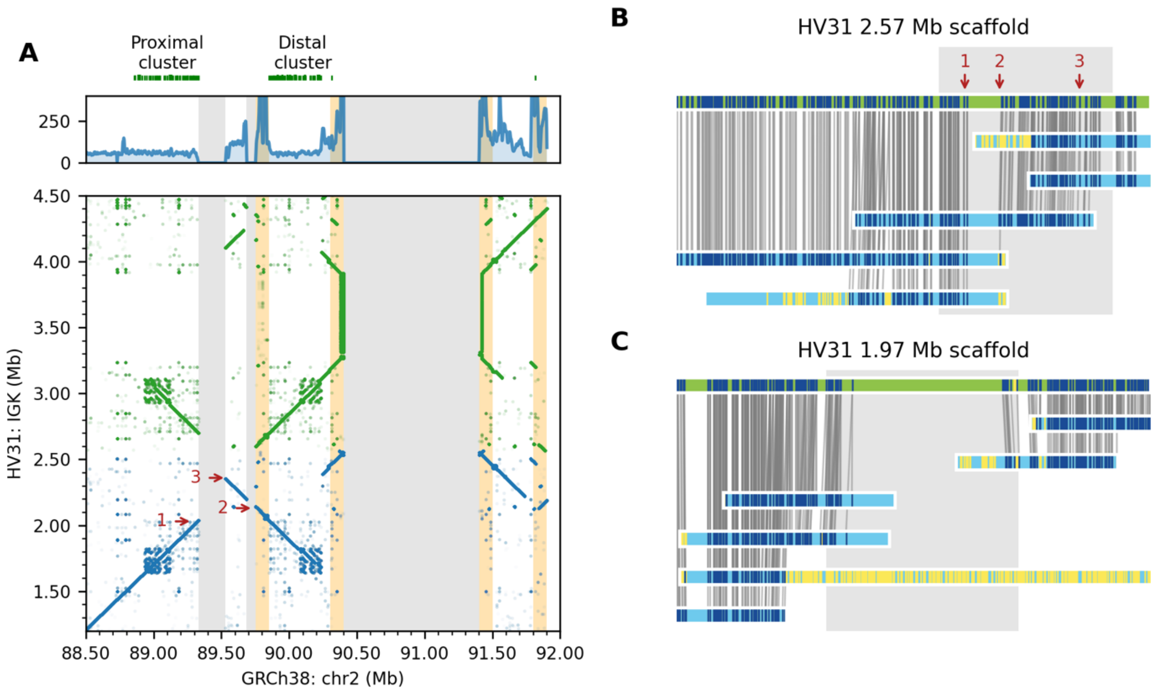


**Figure 27. Three gaps flanked by high-identity repeats were filled in the HV31-V1 assembly.**

**(A)** *k*-mer sharing dot plot comparing GRCh38 (x axis) with the HV31-V1 assembly (y axis). The 2.56 Mb scaffold and the 1.97 Mb scaffold in the HV31-V1 assembly are shown in blue and green, respectively. Coverage depths of ONT-2019 reads aligned to GRCh38, and locations of the proximal and distal IGK gene clusters, are shown above the dot plot. Gaps in GRCh38 are shaded in grey. Novel sequence junctions in the HV31-V1 assembly are annotated with red arrows. Sequence fragments of which extra copies were introduced in the HV31-V1 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 copy numbers of these fragments.

**(B)** Alignment of Bionano contigs (blue) to the 2.56 Mb scaffold in the HV31-V1 assembly (green). DLE-1 labels and their alignments are denoted by coloured lines within and between sequences. The approximate sequence region that maps to the GRCh38 gaps between IGK proximal and distal gene clusters is shaded in grey. For clarity, corresponding positions in the HV31-V1 assembly in **(A)** and **(B)** are labelled with red arrows.

**(C)** Alignment of Bionano contigs (blue) to the 1.97 Mb scaffold in the HV31-V1 assembly (green). Approximate sequence region that maps to the GRCh38 heterochromatin gap is shaded in grey.

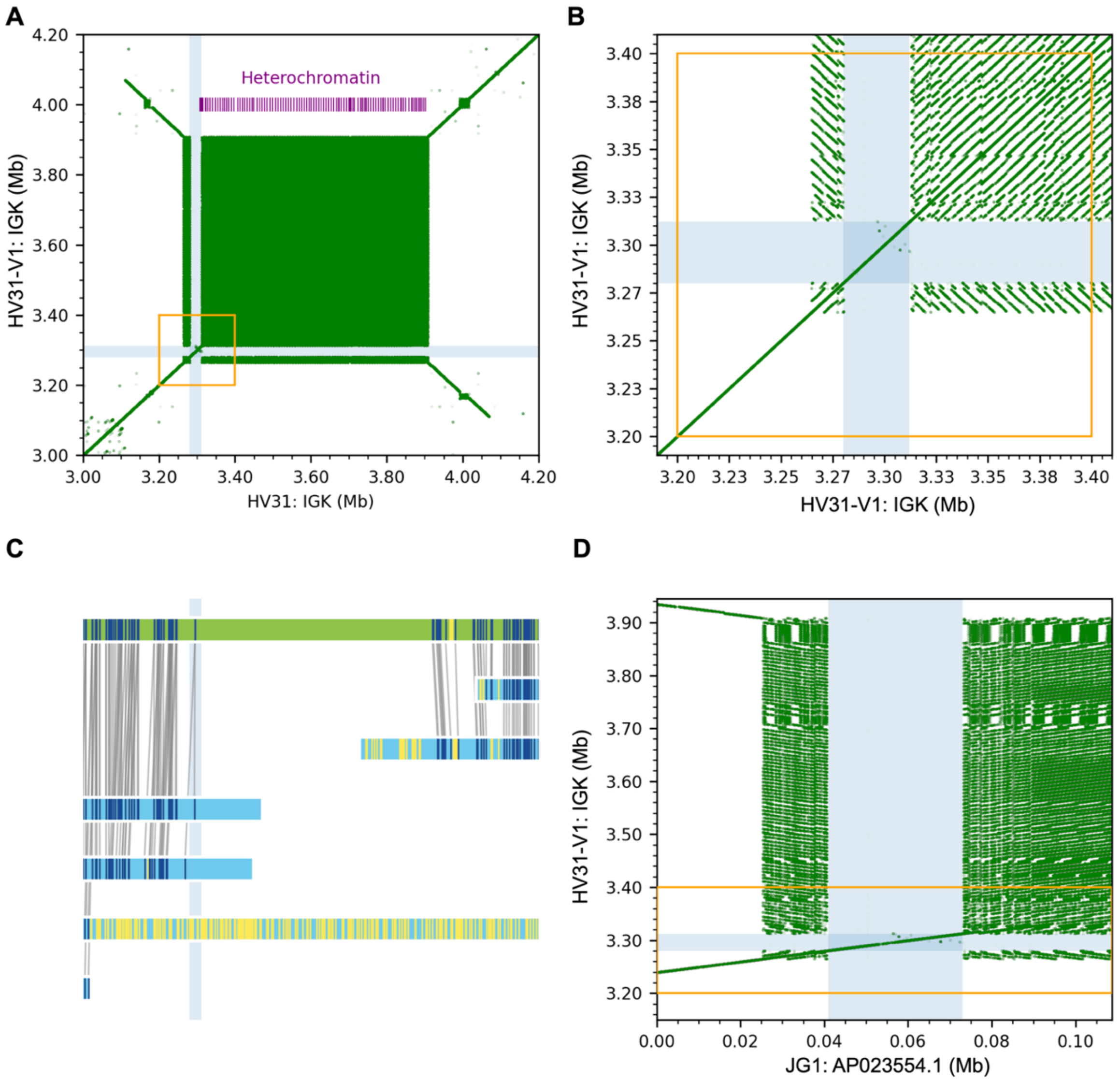


**Figure 28. The heterochromatin gap in the IGK locus was filled with 650 kb complex repeat sequence.**

**(A)** *k*-mer sharing dot plot comparing the HV31-V1 assembly with itself in the IGK heterochromatin region. Purple lines show the occurrence of a 22 bp HSat2B repeat signature sequence (`TTCGATTCCATTTGATGATTCCAT`). A 32 kb unique sequence fragment is highlighted in blue.

**(B)** Details of *k*-mer sharing dot plot in **(A)**, zoomed to reveal details of the unique sequence fragment and repeat structure.

**(C)** Comparison of HV31 contigs and Bionano contigs as in Figure 27C, zoomed in to show that the 32 kb unique fragment (blue shaded region) contained a DLE-1 recognition label that was confirmed by Bionano contigs.

**(D)** *k*-mer sharing dot plot comparing the HV31-V1 assembly (y axis) with the GenBank AP023554.1 contig in the JG1 assembly[51] (x axis).

In **(A)**, **(B)**, and **(D)**, the orange boxes denote approximately the same region.

The HV31-V1 assembly closed three large gaps in the GRCh38 reference sequence, and partially closed a fourth. One of the gaps was located within a 400 kb region of high-identity segmental duplications[113] ~1 Mb downstream of TRB genes (Figure 26), while the other three of these gaps were located in the IGK region (Figure 27 and Figure 28). The largest of them, a ~1 Mb gap annotated as heterochromatin in GRCh38, was located between the distal cluster of IGK genes and the centromere of chromosome 2 (Figure 18 and Figure 28). Examination of this gap revealed a ~650 kb sequence assembled as an array of approximately 115 imperfect tandem copies of 6 kb repeat units (Figure 28). Most of the repeat units contain a 22-bp signature sequence (`TTCGATTCCATTTGATGATTCCAT`), indicating that the heterochromatin sequence belongs to the human satellite HSat2B family[111].

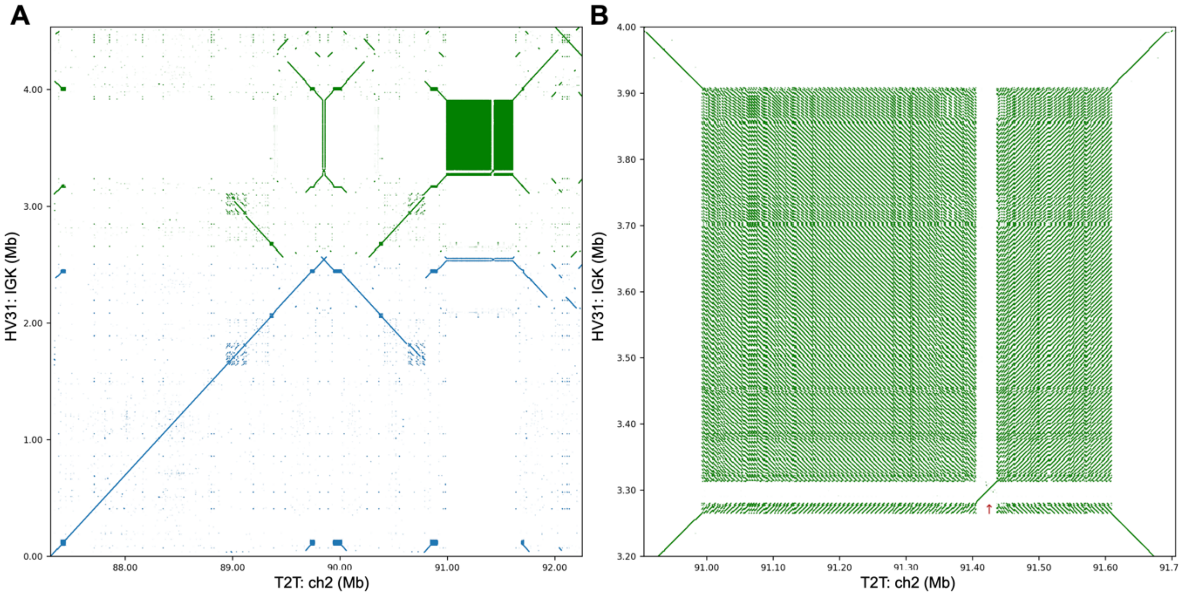


**Figure 29. Comparison of the HV31 and the T2T CHM13 assemblies in the IGK region.**

**(A)** *k*-mer sharing dot plot comparing the HV31-V1 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-V1 assembly are shown in blue and green, respectively.

**(B)** *k*-mer sharing dot plot as in **(A)**, zoomed in to show details of the heterochromatin region. The 32 kb unique sequence fragment is highlighted with a red arrow.

Notably, the assembled heterochromatin sequence also contained a nonrepetitive sequence fragment within the assembled heterochromatin sequence (Figure 28). 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 optical mapping, and I was therefore unable to directly confirm the arrangement using Bionano contigs, though a marker corresponding to the 32 kb unique sequence could be identified (Figure 28C).

I compared the HV31-V1 assembly with the recently reported T2T CHM13 assembly[65] (GenBank sequence: GCA_009914755.2), where the IGK region is fully reconstructed in one contig. The corresponding heterochromatin sequence in the

T2T CHM13 assembly is consistent with the HV31-V1 assembly in terms of total length and repeat unit sequences, though the specific order and orientation of these repeat units differ (Figure 29). 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 (Figure 29B). In addition, I found this 32 kb unique fragment, along with 76.8 kb flanking sequences, was highly consistent (Figure 28D) with a 108.8 kb unplaced sequence (GenBank AP023554.1) in the JG1 assembly[51], which was built from individuals of Japanese ethnicity. Additional analyses of this 32 kb unique fragment are described on page 101. Similar islands of unique sequence amid heterochromatin regions have previously been suggested for chromosome Y[114] and chromosome 21[115].

### 1.2.5. A haplotype-resolved personal genome for HV31

The first phase of the HV31 project, conducted between September 2019 and July 2021 and centred around a mixed-haplotype assembly of immune system regions for HV31, was summarized and published in August 2021[1]. Since then, I continued to work on the second phase of the HV31 project, with the two objectives as initially envisioned: (i) build a haplotype-resolved assembly for HV31, and (ii) investigate applications of the 2019 functional data. The two objectives were related in the sense that inter-haplotype comparisons were deemed useful for associating gene expression profiles in the functional data with specific genetic variants.

In late 2022, I began to develop a workflow for building a high-quality haplotype-resolved genome assembly for HV31, motivated by the arrival of the 2022/23 genomics data (Table 1) from collaborations with PacBio and ONT, and recent reports of advanced *de novo* assembly algorithms[63,64,116]. The following sections describe my current progress from this on-going effort.

## Overview of the haplotype-resolved assembly workflow

A major obstacle for building a haplotype-resolved *de novo* assembly in the first phase of the HV31 project was the lack of high-quality sequencing data with deep genomic coverage[1]: the high-accuracy HiFi-2019 dataset had only 12.3× coverage depth, while the CLR-2019 and ONT-2019 datasets, despite their relatively higher coverage depth, were based on error-prone PacBio CLR and ONT R9 chemistries (Table 1), with reported error rates of ~10%[76,117,118]. As a result, it was difficult to obtain sufficient information from sequencing reads for differentiating subtle differences between the two haplotypes, and among individual repeat units of segmental duplications[119,120].

This issue was largely alleviated by the arrival of 2022/23 genomic data, namely the HiFi-2022, HiFi-2023, and ONT-2022 datasets (Table 1). The new PacBio HiFi data, in particular the HiFi-2023 dataset, which was generated from the PacBio Revio platform[86] that featured throughput improvements over previous PacBio platforms, significantly increased the amount of high-accuracy sequencing data available for HV31, with a cumulative PacBio HiFi coverage reaching 83.7×, exceeding the coverage depth of PacBio HiFi reads (32.4×) used for initial assembly graph construction in the T2T CHM13 assembly[65]. The HiFi-2022 and HiFi-2023 datasets also had slightly better read lengths compared to the HiFi-2019 dataset (Table 1).

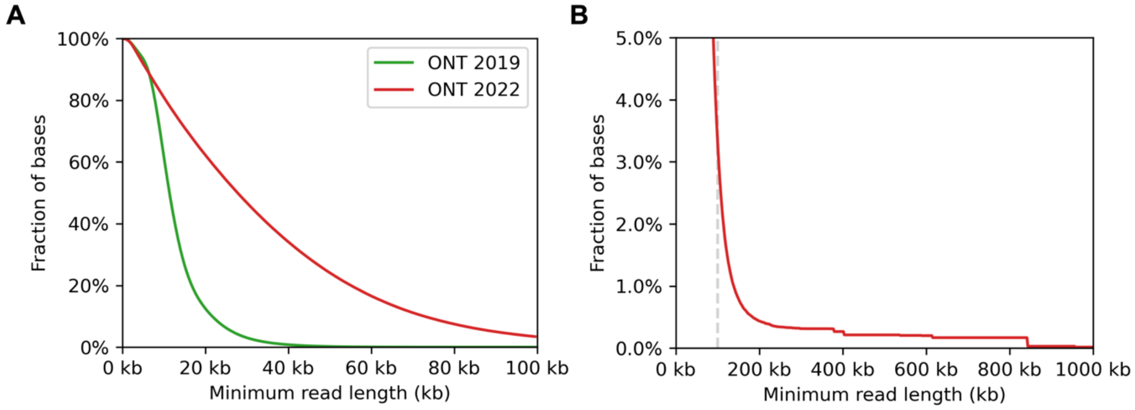


Figure 30. ONT-2022 had longer reads compared to ONT-2019.

**(A)** Read length distribution of ONT-2019 and ONT-2022 datasets, represented as the fraction of total bases (y axis) that exist in reads longer than a given threshold (x axis). No filtering was applied to either dataset.

**(B)** Same as **(A)**, showing only the ONT-2022 dataset, zoomed out to highlight reads longer than 100 kb. The grey dashed vertical line denotes the 100 kb threshold.

The ONT-2022 dataset was based on the ONT R10.4.1 chemistry, which had been reported to feature significant accuracy improvements[32,87,88]. Evaluations of the ONT-2019 and ONT-2022 datasets, described in a separate section below (page 105), confirmed that the ONT-2022 dataset had an overall error rate of ~2%, which could be further improved to ~1% after applying a basic filtering strategy (Figure 51). In addition, the ONT-2022 dataset also contained longer reads compared to the ONT-2019 dataset, with a small fraction of reads (~3% of total bases) longer than 100 kb (Figure 30), which was commonly considered the threshold for "ultra-long" reads that proved to be useful in resolving repeats and producing highly continuous assemblies[64,65,121].

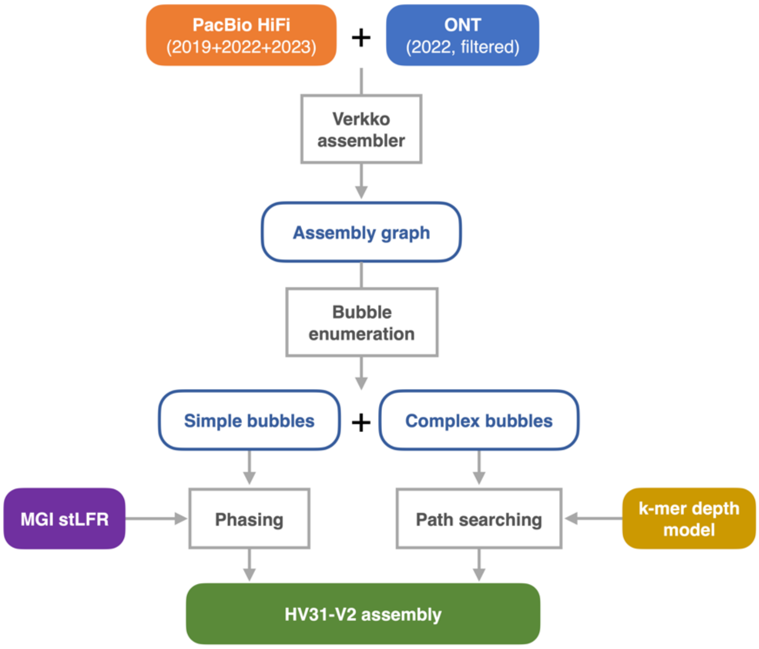


Figure 31. Overview of the HV31-V2 assembly workflow.

Schematic representation of the HV31-V2 assembly workflow. Boxes with rounded corners represent input and output data and intermediate results. Boxes with square corners represent computational processes.

Taking advantage of this rich category of sequencing data for HV31, I constructed a haplotype-resolved assembly for HV31, hereby referred to as the HV31-V2 assembly. The HV31-V2 assembly was build using a hybrid workflow that was primarily derived from the Verkko assembler[64] and modified to incorporate available sequencing data from PacBio HiFi, ONT R10 chemistry, MGI stLFR and Bionano optical mapping platforms, along with *k*-mer depth information generated from long-read and short-read sequencing (Figure 31).

### Assembly graph construction using Verkko

Assembly graphs are data structures widely used to represent sequencing reads and their overlaps. In an assembly graph, nodes typically represent sequences and edges their overlaps (Figure 32), such that a traversal of the graph, represented by contigs, reconstructs the underlying genome from which the sequencing data are generated. Assembly graphs broadly consist of two classes: de Bruijn graphs (DBGs) and string graphs. Depending on specific implementations, DBGs usually represent exact overlaps between fixed-length *k*-mers, while string graphs typically represent inexact overlaps between variable-length substrings of sequencing reads[122]. Given the large amount of input sequencing data, dentification of overlaps and graph building are often the most computationally intensive step in *de novo* assembly[116,122,123].

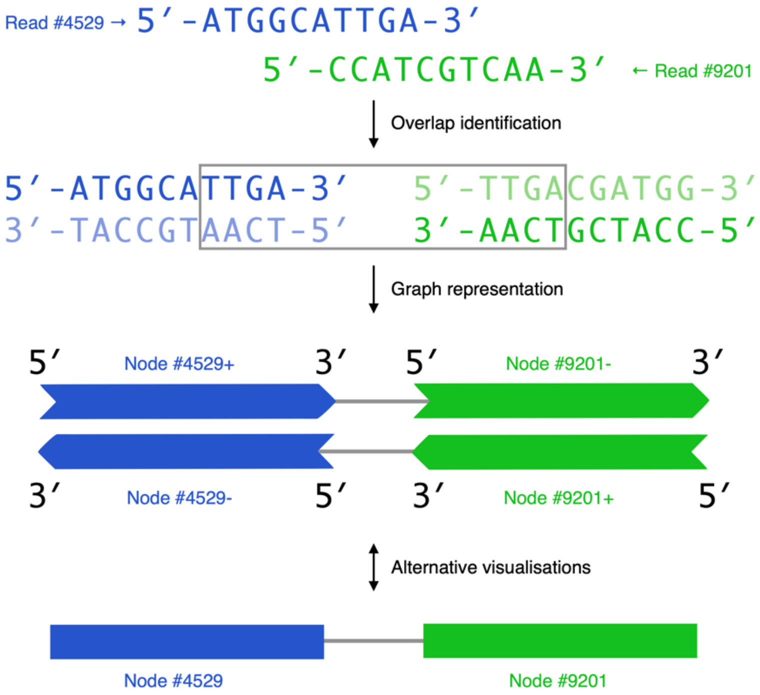


**Figure 32. Assembly graphs represent sequencing reads and their overlaps.**

Schematic illustration of assembly graph construction. Green and blue colours represent two hypothetical sequencing reads that overlap with each other. In the assembly graph, both reads are represented as nodes and their overlap is represented as an edge between the two nodes. In practice, the nodes in an assembly graph do not necessarily represent entire sequencing reads, but also fixed-length *k*-mers (as in de Bruijn graphs) or variable-length substrings (as in string graphs) in sequencing reads, and edges may represent inexact overlaps. The graph structure can be equivalently visualised in the "double" style (top) or the "single" style (bottom)[124]. All graph visualisations below use the "single" style.

The Verkko assembler[64] was applied for initial assembly graph construction in the HV31-V2 workflow. Verkko is a modular multiple-step *de novo* assembly pipeline that first builds a de Bruijn graph (DBG) from PacBio HiFi reads using MBG, which is then simplified by incorporating long-range information from ultra-long ONT reads[64]. Here, the collection of available PacBio HiFi reads, namely HiFi-2019, HiFi-2022, and HiFi-2023 (Table 1), were used for graph building, while the ONT-2022 dataset, filtered and trimmed to improve accuracy (detailed on page 106), was used for graph simplification (Figure 31). The simplified assembly graph, often referred to as the unitig graph, contains nodes that represent long sequences, or unitigs, concatenated from stretches of non-branching nodes, and edges that represent overlaps between them[64]. For the HV31 genome, the unitig graph generated by Verkko displayed linear topology for most chromosomes with very high contiguity (Figure 33), which was comparable to the reported HG002 Verkko assembly[64] and the T2T CHM13 assembly[65]. Out of the 23 chromosomes in the HV31 genome, 17 were assembled telomere-to-telomere for both haplotypes. Chromosome 3 was assembled in two parts, while the five acrocentric chromosomes (13, 14, 15, 21, and 22) formed a single connected component due to the similarity among their short arms, which contain ribosome DNA clusters (Figure 33). Similar graph structures were also observed in the HG002 Verkko assembly[64] and the T2T CHM13 assembly[65].

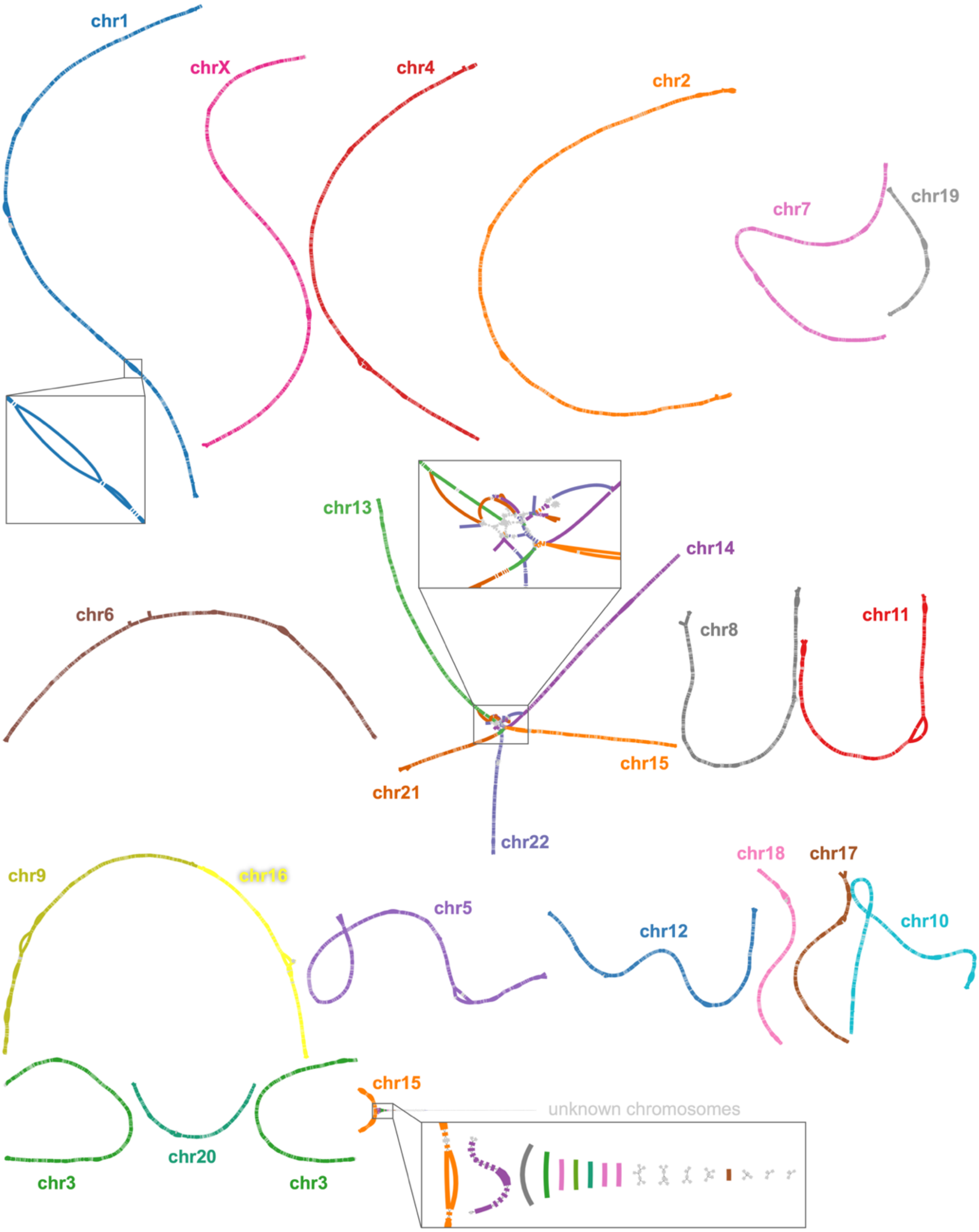


Figure 33. Verkko produced telomere-to-telomere graphs of most chromosomes.

Unitig graph of the HV31 genome generated using Verkko based on the HiFi-2019, HiFi-2022, HiFi-2023 and ONT-2022 datasets. Chromosomes were labelled based on alignment to the T2T reference sequence. Inlets show enlarged versions of selected regions. Unitigs with only low-quality alignments (length < 10 kb or mapping quality < 30) are coloured light grey. Visualisation was made using Bandage-NG[124].

With a high-quality unitig graph, the next step was to generate long, linear, haplotype-resolved sequences (contigs) by traversing each connected component in the graph. By design, following assembly graph building and simplification, Verkko produces haplotype-resolved contigs by resolving paths from the assembly graph using Rukki, which extracts long-range phasing information from specific types of data including trio sequencing, Hi-C, or Strand-Seq[64]. Considering the unavailability of such data for HV31, I developed methods to build contig paths across the assembly graph using information from MGI stLFR reads and *k*-mer depths, as described below, and engineered the Verkko pipeline to generate contigs directly from these externally built paths, skipping the Rukki step (Figure 31).

### Enumerating bubbles in the assembly graph

In a *de novo* assembly unitig graph, most nodes, except those containing complex repeats, typically exist in specific structures termed bubbles and bubble chains[125] (Figure 34). Here a bubble is loosely defined as a subgraph that contains a source node, a sink node, and one or more internal nodes between the source and the sink, and a bubble chain is a linear sequence of bubbles. Bubbles that contain only two internal nodes are called simple bubbles, while bubbles that contain more than two internal nodes, but no cycles (paths that visit the same node more than once), are called super bubbles (Figure 34). The definitions above are adapted from Onodera *et al*.[126], modified slightly to improve naming clarity. In addition, bubbles that contain cycles are referred to as cyclic bubbles here (Figure

35). Formal definitions of bubble classes and bubble chains are detailed on page 130.

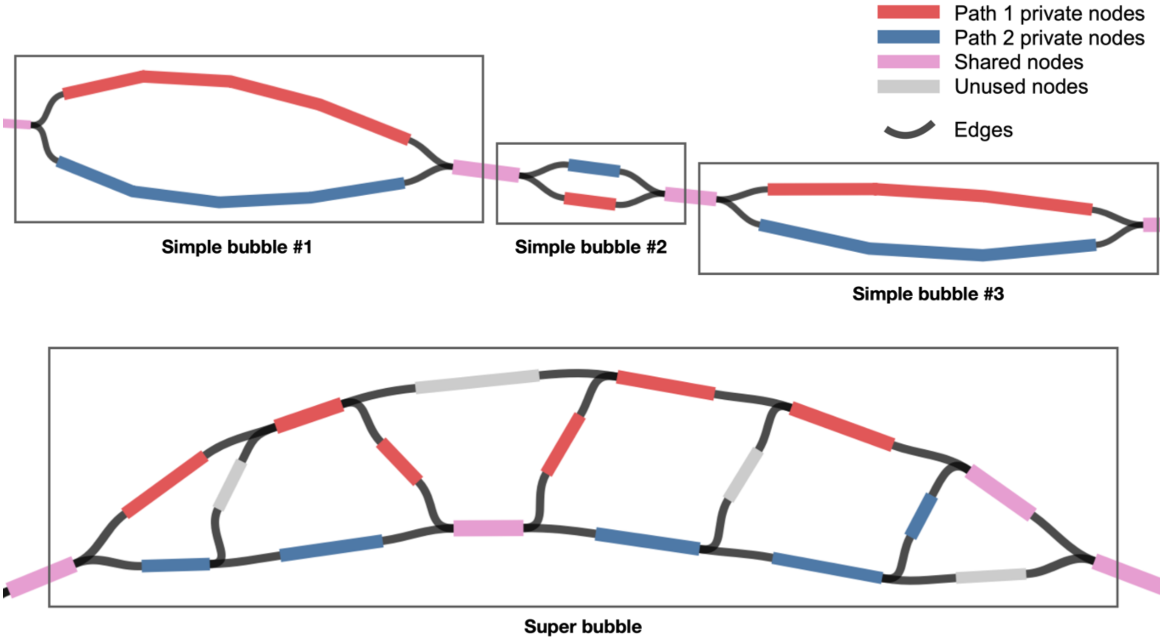


**Figure 34. Examples of bubbles and bubble chains.**

Two bubble chains are shown. The top chain contains three simple bubbles, and the bottom chain contains one super bubble. Node colours denote two possible paths that traverse the bubble chain, with path 1 private nodes, path 2 private nodes, and shared nodes, are shown in red, blue, and pink, respectively. Nodes not included in either path are shown in grey. In a diploid genome, the two paths that traverse a bubble chain likely represent the two haplotypes of the locus, in which case nodes not included in these two paths likely originate from sequencing errors, or similar sequences located in different parts of the genome. Formal definitions of bubbles and bubble chains are described on page 130.

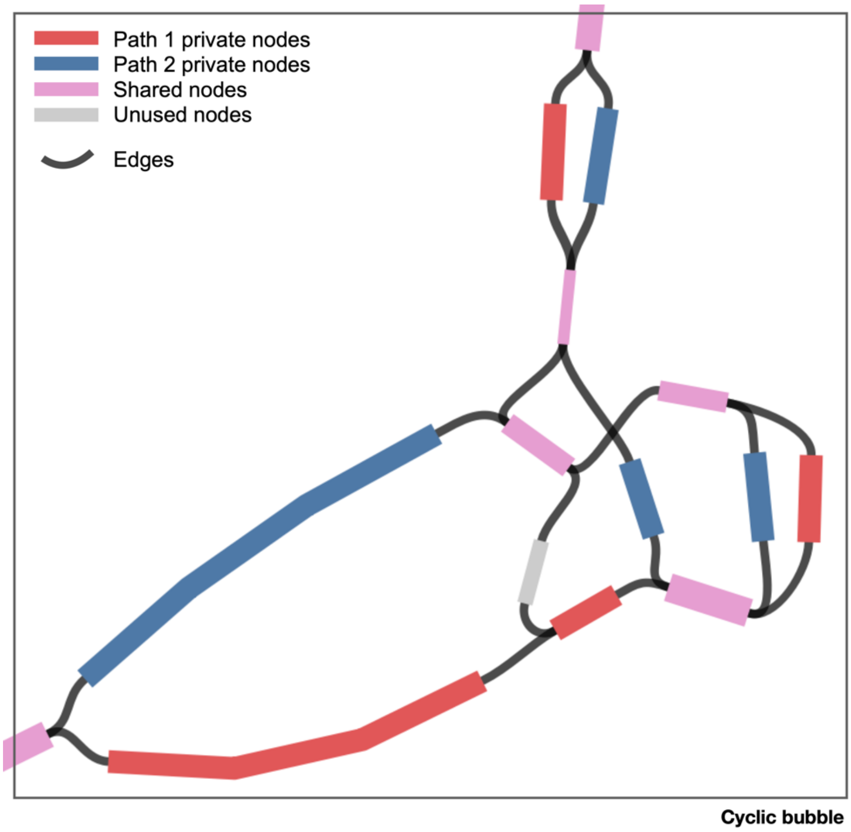


Figure 35. An example of a cyclic bubble.

A cyclic bubble in shown, which contains valid paths that visit the same node more than once. For example, it is possible to leave from the node at the bottom-left corner and return to the same node via multiple valid paths. Node colours denote two possible paths that traverse the bubble chain, with path 1 private nodes, path 2 private nodes, shared nodes, and nodes not included in either path coloured red, blue, pink, and grey, respectively.

Bubble chains are key structures in assembly graphs, as they can be traversed bubble by bubble, through source and sink nodes[125], while non-bubble structures likely involve complex branching patterns that sometimes require manual resolution[65]. Here I used BubbleGun[125], which implements the linear-time algorithm proposed by Onodera *et al*.[126], to enumerate all the simple and super bubbles from the unitig graph built using Verkko. Cyclic bubbles were identified up to a depth limit using a custom algorithm (detailed on page 130). Given that the human genome is diploid, and that internal nodes in bubbles typically arise from heterozygosity or sequencing errors[125], the *de novo* assembly problem here can be framed as finding two paths across each bubble chain that are collectively most consistent with the available data, among all valid paths. The identification of such traversal paths is described below.

### Resolving complex bubbles by *k*-mer depth modelling

Complex bubbles, here referring to the union of super and cyclic bubbles, typically represent segmental duplications, and are sometimes further complicated by sequencing errors. The interior nodes of complex bubbles cumulatively account for 10.1% of total sequence length in the Verkko unitig graph (Figure 36). In principle, the worst-case number of valid traversal paths within a complex bubble increases exponentially with the distance between the source and the sink nodes, measured as the number of branching internal nodes. Nevertheless, I observed that most complex bubbles in the Verkko unitig graph contained less than 15 internal nodes (Figure 36), making it possible to enumerate all traversal paths connecting the source node and the sink node of a complex bubble using a brute-force search algorithm. This reduces the problem of optimal path finding to the problem of path ranking, that is, deriving a metric to indicate the relative plausibility of each pair of paths given the available sequencing data.

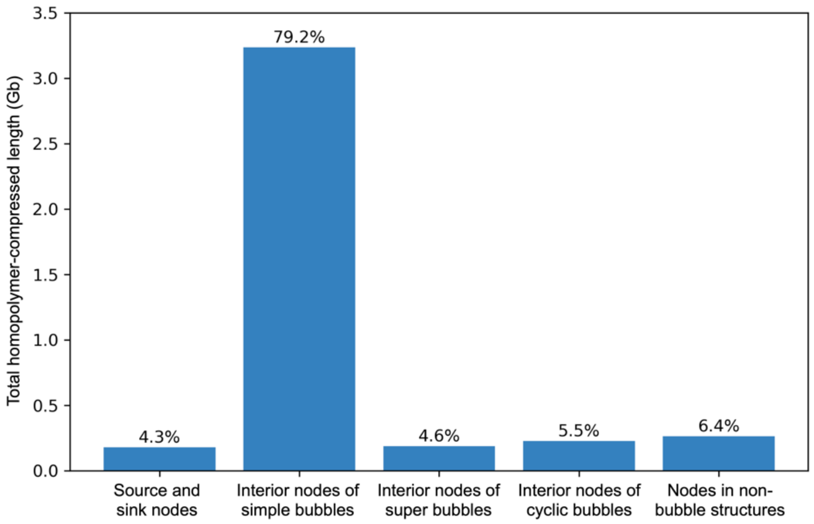


**Figure 36. Composition of node types in the Verkko unitig graph.**

The total homopolymer-compressed lengths (y axis) of each type of nodes in the Verkko unitig graph grouped by bubble topology (x axis).

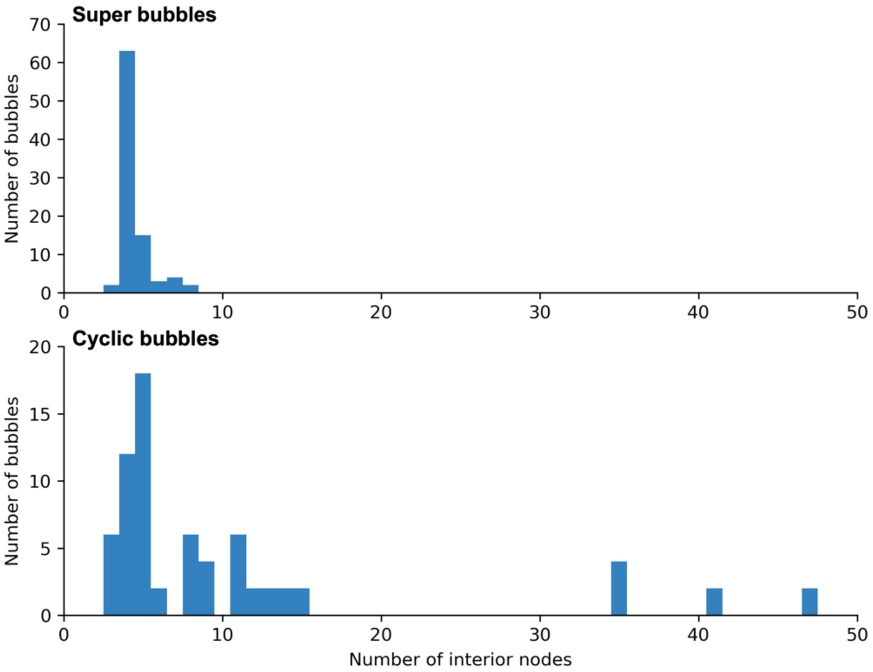


**Figure 37. Most complex bubbles had fewer than 15 internal nodes.**

The distribution of internal node counts for super bubbles (top) and cyclic bubbles (bottom) in the Verkko unitig graph.

Within a super bubble, valid paths differ only in terms of inclusion or exclusion of certain nodes (Figure 34). Within a cyclic bubble, valid paths may additionally differ in terms of node revisits, orders, and orientations (Figure 35). In both super bubbles and cyclic bubbles, including or revisiting a node manifests in the resulting assembly as differences in the copy number of the node sequence, which can be zero, one, two, or more. As described previously (Figure 14), the copy number of a sequence in the genome can be inferred from *k*-mer coverage depths in the sequencing data. Thus, *k*-mer coverage depths should be informative of the relative plausibility of path pairs. In other words, when selecting a pair of haplotype paths, nodes with higher *k*-mer depths should be included, or even revisited, while nodes with lower *k*-mer depths likely originate from sequencing errors and should be ignored.

Inferring the most plausible copy numbers of a given node based on the depths of its *k*-mers requires a quantitative model that describes the relationship between *k*-mer depths and *k*-mer copy numbers. Designed for this purpose, GenomeScope[105,127] is a mixture model that decomposes the *k*-mer profile, also known as the *k*-mer spectrum, which is the distribution of *k*-mer depths in a sequencing dataset, as the sum of several evenly spaced peaks. Each peak represents *k*-mers of a given copy number, and is modelled as the probability density function of a negative binomial distribution[105]. However, GenomeScope only considers peaks with copy numbers of one to four, while peaks with copy numbers of zero, which represent sequencing errors, and those with copy numbers greater than four, which represent multi-copy repeats, also need to be modelled for complex bubble resolution. Therefore, I extended the original GenomeScope implementation to allow for explicit modelling of sequencing errors and an arbitrary number of peaks (detailed on page 131).

For the HV31 genome, *k*-mer depth modelling was based on available accurate sequencing data, which included the HV31-V1 validation dataset (Figure 14), and additionally the HiFi-2022 and HiFi-2023 datasets, with an overall coverage depth of 333× (Table 1). The high coverage depth of accurate sequencing data enabled the *k*-mer profile to be decomposed into 13 peaks by fitting the extended model that I developed, including an error peak and 12 peaks corresponding to *k*-mers with diploid copy numbers from one to twelve (Figure 38).

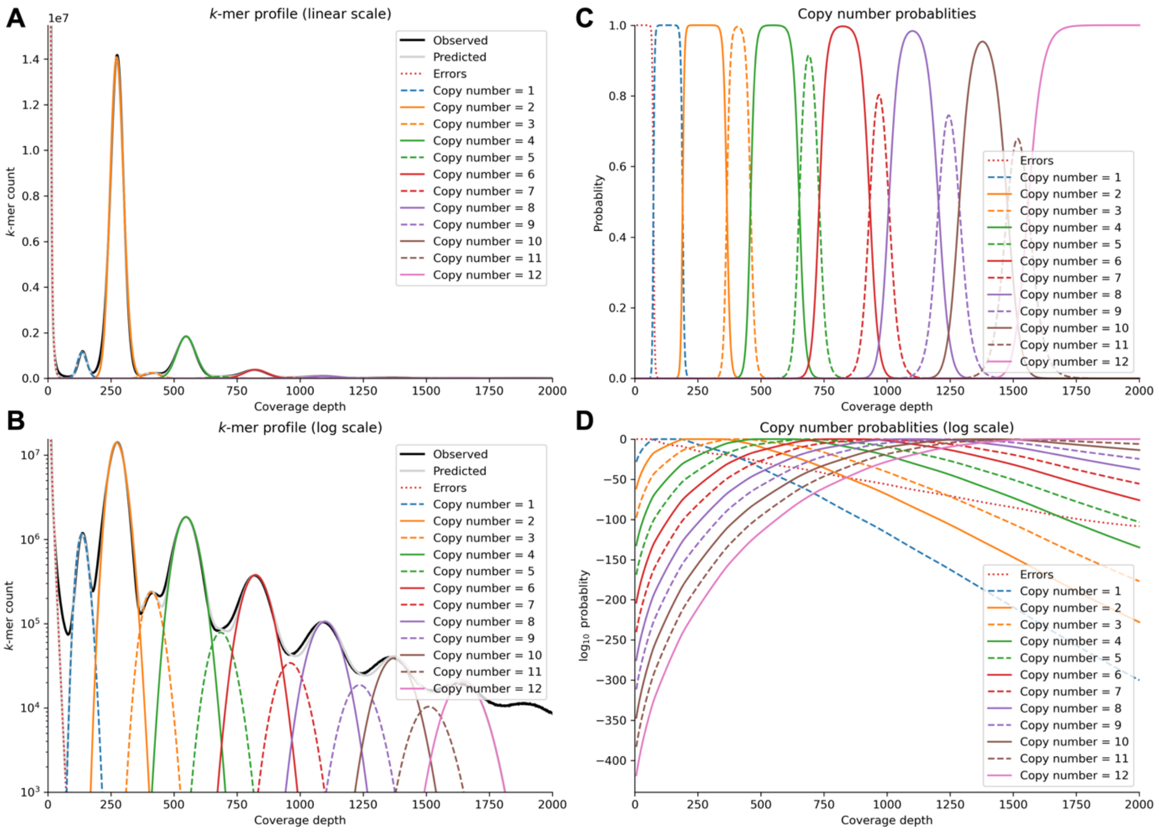


**Figure 38. Decomposition of the *k*-mer profile from sequencing data.**

**(A)** Observed (black line) and predicted (grey line) *k*-mer counts (y axis) for each *k*-mer coverage depth (x axis). Predicted *k*-mer counts were calculated using a modified version of the GenomeScope model fitted to selected HV31 sequencing datasets (detailed on page 131). Peaks corresponding to each diploid copy number are shown in coloured lines. *k* = 22. Homopolymer-compressed sequences were used for *k*-mer analyses.

**(B)** Same as **(A)**, with y axis shown in log scale.

**(C)** Inferred probabilities of each copy number (y axis) conditioned on the coverage depth of a given *k*-mer (x axis), based on the *k*-mer profile model in **(A)**.

**(D)** Same as **(C)**, with y axis shown in log scale.

This figure is related to Figure 14.

Based on the fitted *k*-mer profile model described above, the valid path pairs from each complex bubble were enumerated and ranked (detailed on page 133). For each node, private *k*-mers were identified as *k*-mers that appear only in this node, but not elsewhere in the assembly graph. The best path pair was defined as the pair of paths whose joint copy numbers of nodes had the highest likelihood given the private *k*-mer coverage depths of the corresponding nodes, assuming each *k*-mer was independent, and selected for subsequent contig construction (Figure 39).

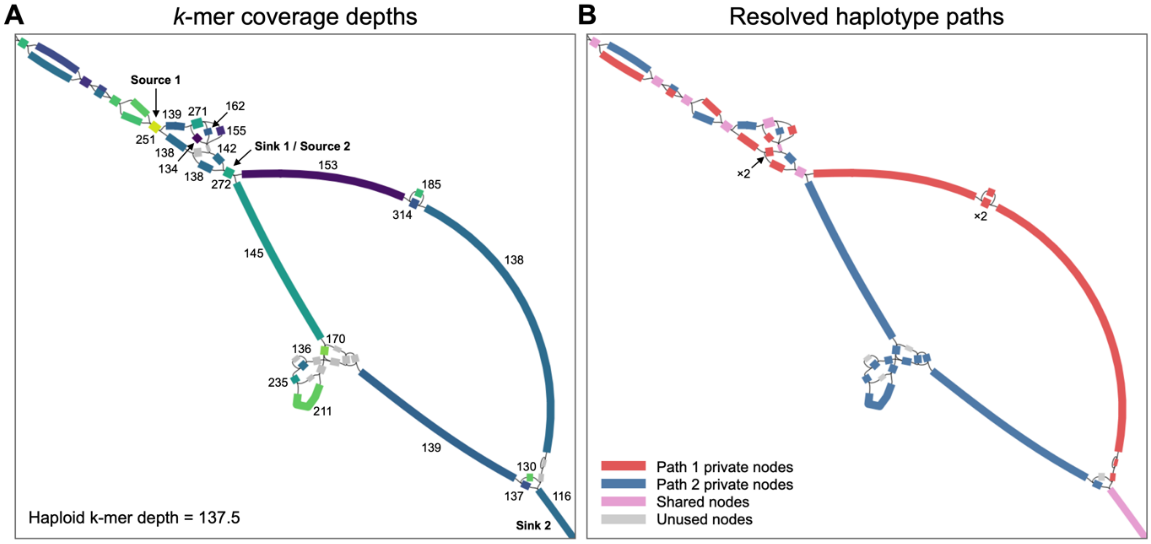


**Figure 39. An example of complex bubble resolution.**

**(A)** Resolution of two cyclic bubbles found in chromosome 20 in the HV31 genome. The source and sink nodes for each cyclic bubble are labelled. Each node in the cyclic bubbles is labelled by the median coverage depths of its private *k*-mers. Nodes without unique *k*-mers are coloured grey.

**(B)** Resolved haplotype paths of the two cyclic bubbles in **(A)**. Nodes visited twice in a single path are labelled “×2”.

The current approach for complex bubble resolution had several limitations. First, brute-force path enumeration was computationally intractable for large bubbles with complex internal structures, and requires a sufficiently simplified assembly graph to start with. Second, this method did not consider node order and orientation differences, though these differences were empirically rare in the Verkko assembly graph for HV31. Third, the lengths and number of private *k*-mers of a give node, which likely contains valuable information for path selection, was not specifically considered. Finally, the coverage depths of private *k*-mers within a node were assumed to be independent, which might not be true for *k*-mers in close proximity. Further development of this method will likely lead to more general, robust, and efficient approaches for complex bubble resolution.

### Phasing simple bubbles using MGI stLFR data

Unlike super and cyclic bubbles, which may contain one or more branching internal nodes, simple bubbles allow only two traversal paths, each though one of the two internal nodes (Figure 34), which often represent the two haplotypes. Nevertheless, to phase the simple bubbles relative to each other in a bubble chain, that is, to determine which internal nodes belong to the same haplotype, usually requires additional information, such as parental sequencing data, Strand-Seq, or Hi-C[54,64,128]. Here I took advantage of the MGI single-tube long-fragment read (stLFR) data available for HV31 to facilitate bubble phasing.

stLFR is a sequencing technology developed by MGI to produce barcoded short reads, such that reads sharing the same barcode usually originate from the same input DNA fragment[33], thereby retaining long-range information for the sequenced short reads. In the HV31 stLFR data, I observed the stLFR fragments were significantly longer than PacBio or ONT sequencing reads, with nearly 20% of total bases found in fragments longer than 100 kb (Figure 40).

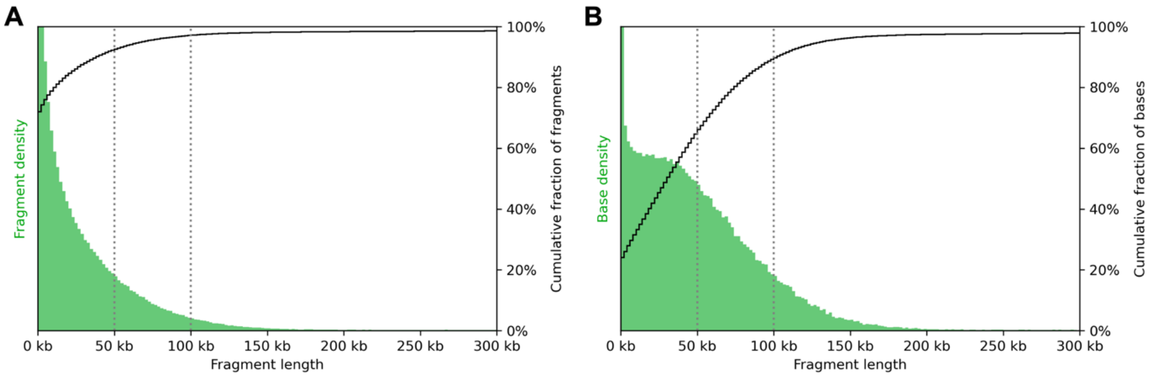


**Figure 40. MGI stLFR fragment lengths.**

**(A)** Empirical probability density function (green) and cumulative density function (black) of MGI stLFR fragment lengths.

**(B)** Empirical probability density function (green) and cumulative density function (black) of MGI stLFR fragment lengths weighted by the total number of bases sequenced in each fragment.

In **(A-B)**, MGI stLFR fragment lengths were estimated by aligning stLFR reads to the T2T reference genome. A fragment was defined as one or more reads sharing the same barcode, aligned to the same chromosome within a 1 Mb region. The fragment length was calculated as the distance between the start of the first read and the end of the last read.

Each stLFR read contains multiple tandem barcodes, which has been reported to improve barcoding uniqueness compared to 10x Linked-Reads sequencing, which uses a single barcode[33]. In the 2019 genomics data, the MGI-stLFR dataset contained a 30-nt barcode for each read pair, in the form of three tandem 10-nt segments, while the 10x dataset contains a 16-nt barcode for each read pair. Previous results also suggested that the MGI-stLFR dataset had significantly better accuracy compared to the 10x dataset (Figure 13).

Each 10-nt segment in a 30-nt stLFR barcode is randomly selected from a library of 1536 unique segments[33], which allows the 10-nt segment to encode up to 10.6 bit of information, assuming a uniform distribution and the absence of sequencing errors. Therefore, despite the fact that 30 bases are devoted to barcodes, less than 31.8 bit of barcoding information is available for each read pair. As a result, it remains possible for reads from multiple unrelated DNA fragments to share the same stLFR barcodes. Empirically, I found that over 35% of all stLFR fragments, or over 25% of fragments longer than 2 kb, were labelled with unique barcodes not shared by other fragments (Figure 41). Despite the random noise generated by fragments that happen to share the same barcode, stLFR reads have been successfully used for variant phasing in the human genome, yielding phase blocks with N50 up to 34 Mb[33]. Therefore, I developed a *k*-mer-based method that utilised the accurate long-range phasing information provided by stLFR data, which was complementary to PacBio HiFi and ONT long-read sequencing, to phase simplex bubble paths in a bubble chain relative to each other, as described below.

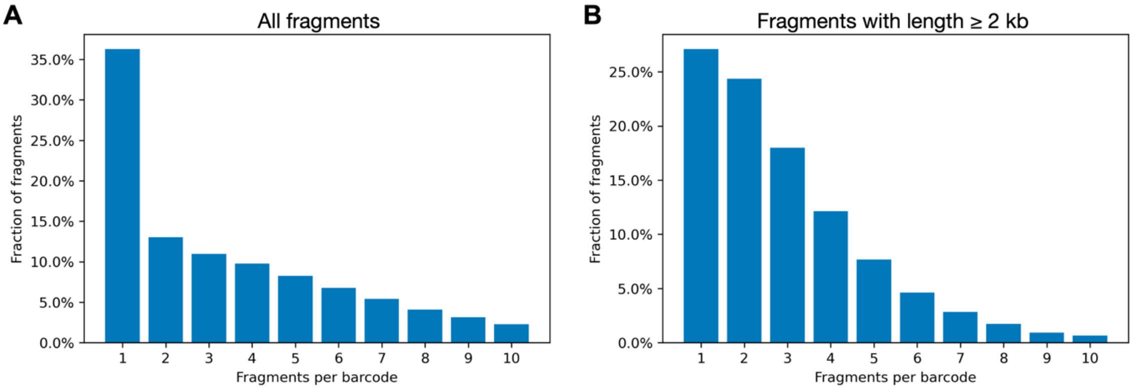


**Figure 41. stLFR fragments feature largely unique barcodes.**

**(A-B)** Fractions of fragments grouped by the number of fragments per barcode, for all fragments **(A)** or for fragments with length ≥ 2 kb **(B)**.

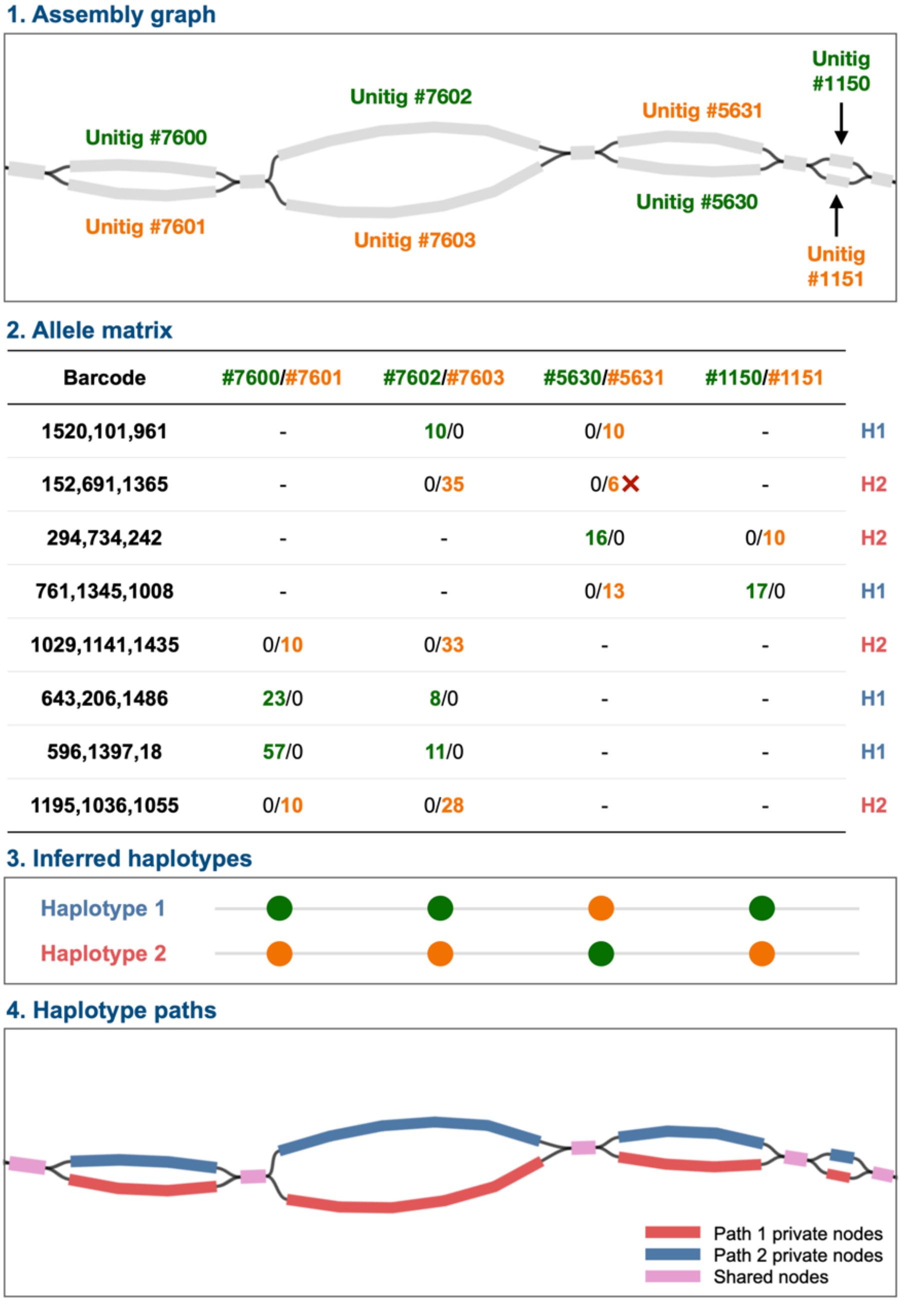
A common approach for diploid phasing is to identify heterozygous SNPs from sequencing reads aligned to a reference genome, and representing them as an allele matrix. The correct haplotype of these SNPs can then be inferred by solving a minimum error correction (MEC) problem[129–131]. The MEC problem provides a general, noise-tolerant framework for haplotype phasing of diploid genomes. Various algorithms have been developed to solve the MEC problem efficiently in the context of variant phasing, such as WhatsHap[132,133], HapCUT[134], and HapCUT2[135].

However, the current approach based on variant calling has several limitations, especially in diverse, repeat-rich regions of the genome, such as those encoding key immune system components (Table 2). In particular, aligning reads to a reference genome is likely to introduce reference bias or alignment errors in the presence of repeats or large structural variants (Figure 23). Moreover, structural variants, which are clearly informative for phasing, are challenging to detect using alignment-based methods (Figure 21).

As almost all heterozygous SNPs can be captured by heterozygous *k*-mers without dependence on reference genomes[136,137], I used heterozygous *k*-mers (diploid copy number = 1) identified from the *k*-mer profile model described above (Figure 38) instead of heterozygous SNPs to alleviate the aforementioned issues of reference bias and alignment errors, without losing compatibility to existing MEC solvers. I expected that most of the heterozygous *k*-mers overlapped with heterozygous SNPs, while others were *k*-mers that existed only on one haplotype due to structural variation. An example of a structural variant capture by heterozygous k-mers is shown in Figure 15A.

The heterozygous *k*-mers were used to associate stLFR fragments with specific internal nodes of simple bubbles (detailed on page 89). The resulting allele matrix, in which every row represented a stLFR fragment and every column represented a simple bubble, was solved as a MEC problem using the HapCUT2 heuristic algorithm[135], producing two haplotype paths for simple bubbles in each bubble chain (Figure 42),.

The stLFR-based bubble phasing practice developed here has several limitations. First, the method assumes that the two internal nodes of each simple bubble represent the two haplotypes, while in practise one of the nodes may actually originate from sequencing errors. Second, accurate identification of heterozygous *k*-mers depends on deep sequencing coverage and is affected by random local fluctuations in the sequencing depth, which may lead to more false positives and false negatives in heterozygous variant detection compared to alignment-based methods. Third, due to the lack of alternative sources of phasing information for the HV31 genome, I was not able to systematically evaluate the accuracy of stLFR-based bubble phasing. Finally, it is possible, in principle, to phase resolved paths in complex bubbles using a similar manner, which is not implemented currently.



| Barcode | #7600/#7601 | #7602/#7603 | #5630/#5631 | #1150/#1151 | |
|---|---|---|---|---|---|
| 1520,101,961 | - | 10/0 | 0/10 | - | H1 |
| 152,691,1365 | - | 0/35 | 0/6 ✗ | - | H2 |
| 294,734,242 | - | - | 16/0 | 0/10 | H2 |
| 761,1345,1008 | - | - | 0/13 | 17/0 | H1 |
| 1029,1141,1435 | 0/10 | 0/33 | - | - | H2 |
| 643,206,1486 | 23/0 | 8/0 | - | - | H1 |
| 596,1397,18 | 57/0 | 11/0 | - | - | H1 |
| 1195,1036,1055 | 0/10 | 0/28 | - | - | H2 |

Figure 42. Phasing bubbles using MGI stLFR linked reads.

Procedure of phasing bubbles in a bubbles chain using long-range information from MGI stLFR linked reads. Starting from an assembly graph, the objective was to identify a pair of haplotypes for each bubble chain that are most consistent with stLFR data, among all possible haplotype combinations. Heterozygous *k*-mers, i.e., *k*-mers with copy numbers of one in a diploid genome, were used to associate stLFR barcoded fragments with specific bubble nodes, represented as an allele matrix for the minimum error correction (MEC) problem[130,131], which was resolved using HapCUT2[135] to obtain the target haplotypes (detailed on page 133). In the allele matrix, each cell represents the numbers of heterozygous *k*-mer labels that a given stLFR fragment (rows) contains for a given pair of nodes (columns). For example, 8/0 denotes a fragment contains 8 labels for the first node in a pair, and 0 labels for the second node in the same pair. 0/0 is represented by dashes (-) for clarity. In this hypothetical example, the MEC solution identified one error in the allele matrix to be corrected, which is marked by a red cross.

After the optimal haplotype paths of each simple or complex bubble in the assembly graph were identified as described above, these paths were finally processed by Verkko, to produce contigs from the consensus of HiFi reads along each haplotype path[64] (Figure 31). In addition, unresolved unitigs which were not part of any simple or complex bubble were each included as a separate contig. The resulting collection of contigs is hereby referred to as the HV31-V2 assembly.

### Evaluation of the HV31-V2 assembly

The HV31-V2 assembly contained roughly 2.9 Gb resolved sequence for each of the two haplotypes, in addition to 264 Mb unresolved sequence for which haplotype information was missing. Aligning the HV31-V2 assembly to the T2T reference genome revealed that several chromosomes, including chromosomes 1, 7, 10, 12, 18, and X, were assembled with near telomere-to-telomere contiguity, while several other chromosomes reached telomere-to-centromere contiguity (Figure 43).

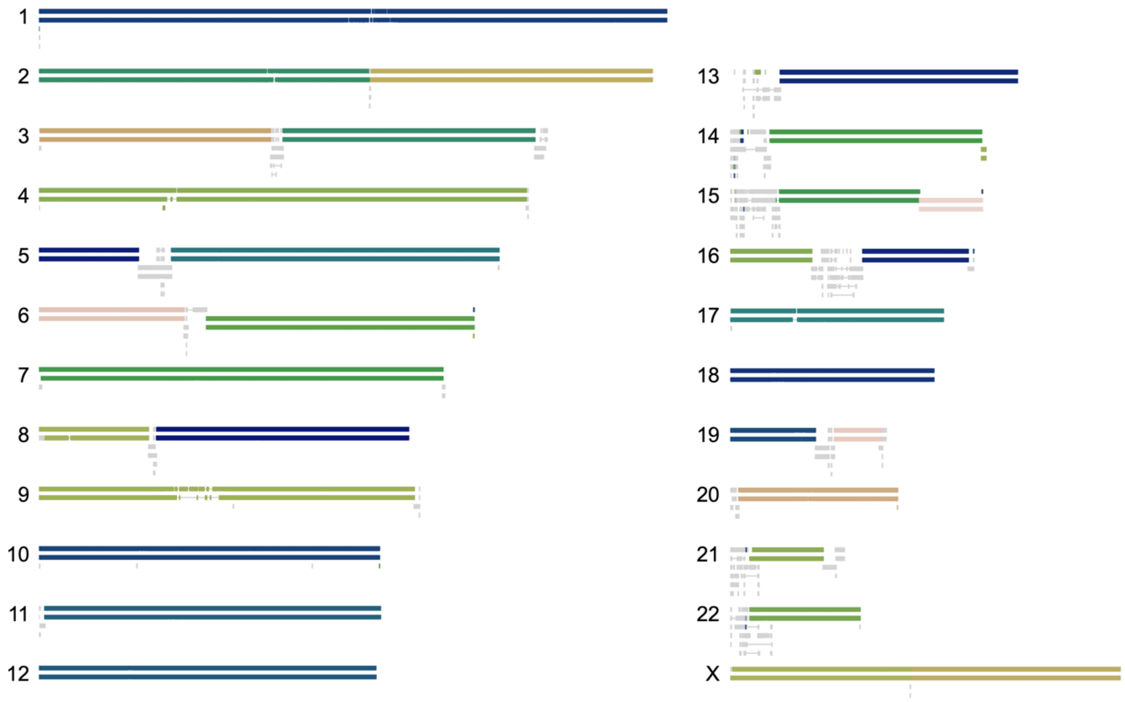


**Figure 43. The HV31-V2 assembly contains entire chromosome arms.**

Alignment of the HV31-V2 assembly to the T2T reference genome. Chromosome names are labelled next to corresponding HV31-V1 contigs. Each horizontal bar represents a contig. Each pair of haplotype contigs is assigned a random colour. Contigs that originated from unresolved unitigs are shown in grey.

More than half of the genome in each haplotype were assembled in contigs longer than 100 Mb (Figure 44), which was higher than GRCh38 and several published high-quality haplotype-resolved assemblies[49,128]. Although all simple and complex bubbles were resolved using methods described above, the Verkko assembly graph also contained complex non-bubble structures, such as the highly-connected structure representing the short arms of the five acrocentric chromosomes (Figure 33), which I was not able to resolve algorithmically. As a result, the HV31-V2 assembly contigs (Figure 43) were less contiguous than the corresponding assembly graph (Figure 33).

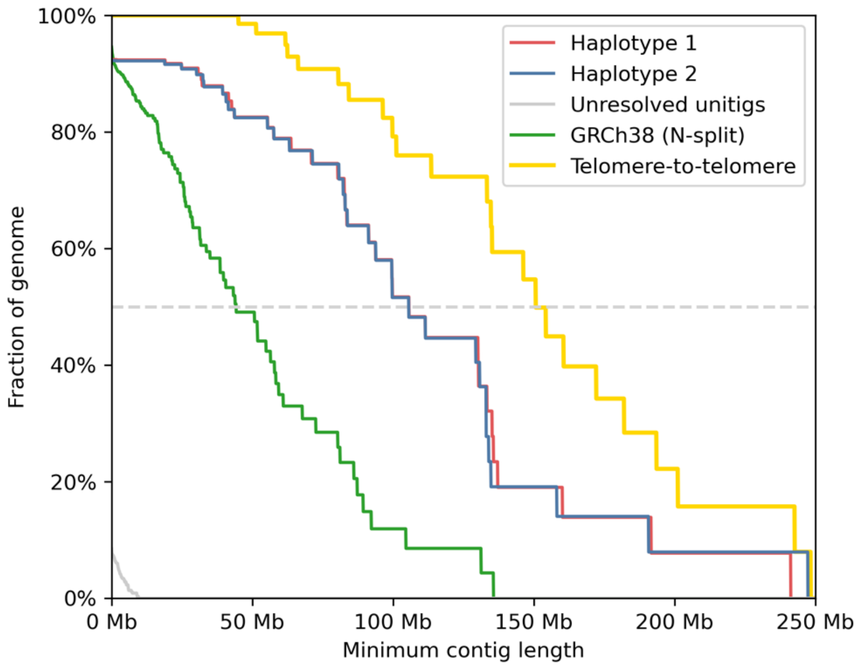


**Figure 44. The HV31-V2 assembly is highly contiguous.**

Distribution of HV31-V2 assembly contig lengths, presented as the fraction of genome assembled (y axis) in contigs longer than a given minimum length cut-off (x axis), for each haplotype (red and blue) and for unresolved contigs originating from complex non-bubble structures (grey). Contig lengths of the GRCh38 reference sequence, after splitting at unknown (N) bases, are shown in green. Chromosome lengths in the T2T reference sequence, which represent the upper limit of assembly contiguity, are shown in yellow. The dashed grey horizontal line represents the threshold where 50% of the genome is assembled, corresponding to NG50 values. Genome fractions were calculated based on the estimated female human genome of 3.05 Gb[65].

As mentioned previously, the coverage depth of a given *k*-mer is expected to be proportional to its copy number (CN) in the genome (Figure 38). This linear correlation between *k*-mer coverage depth and *k*-mer copy number, especially for low-copy (CN ≤ 4) *k*-mers, is commonly used to evaluate the error rate and completeness of a haplotype-resolved *de novo* assembly[104]. In the HV31-V2 assembly, I observed a highly consistent correlation between assembly *k*-mer copy numbers and corresponding *k*-mer coverage depths from the validation data

(Figure 45). In particular, almost all assembled heterozygous (CN = 1) and homozygous *k*-mers (CN = 2) had expected coverage depths, suggesting a reasonably complete representation of both haplotypes. Mismatches between the assembly copy number and the k-mer depths, manifesting as additional peaks in the *k*-mer depth distribution not corresponding to the expected copy numbers, likely resulted from various artefacts, such as incomplete or erroneous graph construction. In addition, for some local structures, the assumptions based on which optimal paths were computed might be broken. For example, in a simple bubble, it was likely that one of the internal nodes represented sequencing errors, as mentioned above, in which case both paths should go through the other node, instead of using both nodes. Further investigation will be required to develop more sophisticated algorithms for path resolution.

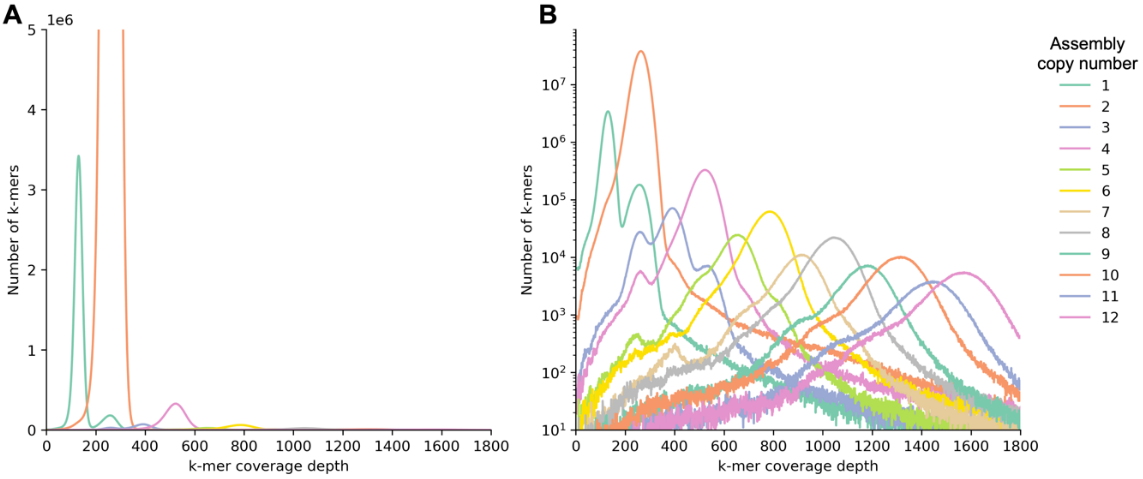


**Figure 45. HV31-V2 assembly *k*-mer copy numbers are consistent with sequencing data.**

**(A)** Distribution of *k*-mer coverage depths (x axis) in selected sequencing datasets, grouped by the corresponding *k*-mer copy numbers in the HV31-V2 assembly (colours). *k* = 31.

**(B)** Same as **(A)**, with the y axis shown in log scale.

## 1.2.6. Genetic variation analyses facilitated by a personal genome

The HV31-V2 assembly represents both haplotypes of the HV31 genome and provides a more complete view of genetic variation compared to the mixed-haplotype HV31-V1 assembly. To demonstrate the potential applications of the HV31-V2 assembly, I revisited the eight genomic regions encoding key immune system components investigated previously, and analysed how a haplotype-resolved assembly could contribute to a more complete understanding of these complex regions, as described below.

### A haplotype-resolved map of key immune gene allotypes

A major limitation of the HV31-V1 assembly, as mentioned above, was that only one of the two inherited alleles of each gene, arbitrarily selected by the assembler, was represented in the assembly. Despite my previous effort to capture information about the unassembled alleles by detecting heterozygous structural variants from read alignment, it remained unclear which subset of the assembled alleles belonged to the same haplotype. Such information, once validated, could help investigate the functional impacts of allelic and structural variants, and contribute to the existing population panels of known haplotypes in complex regions[49,109,138].

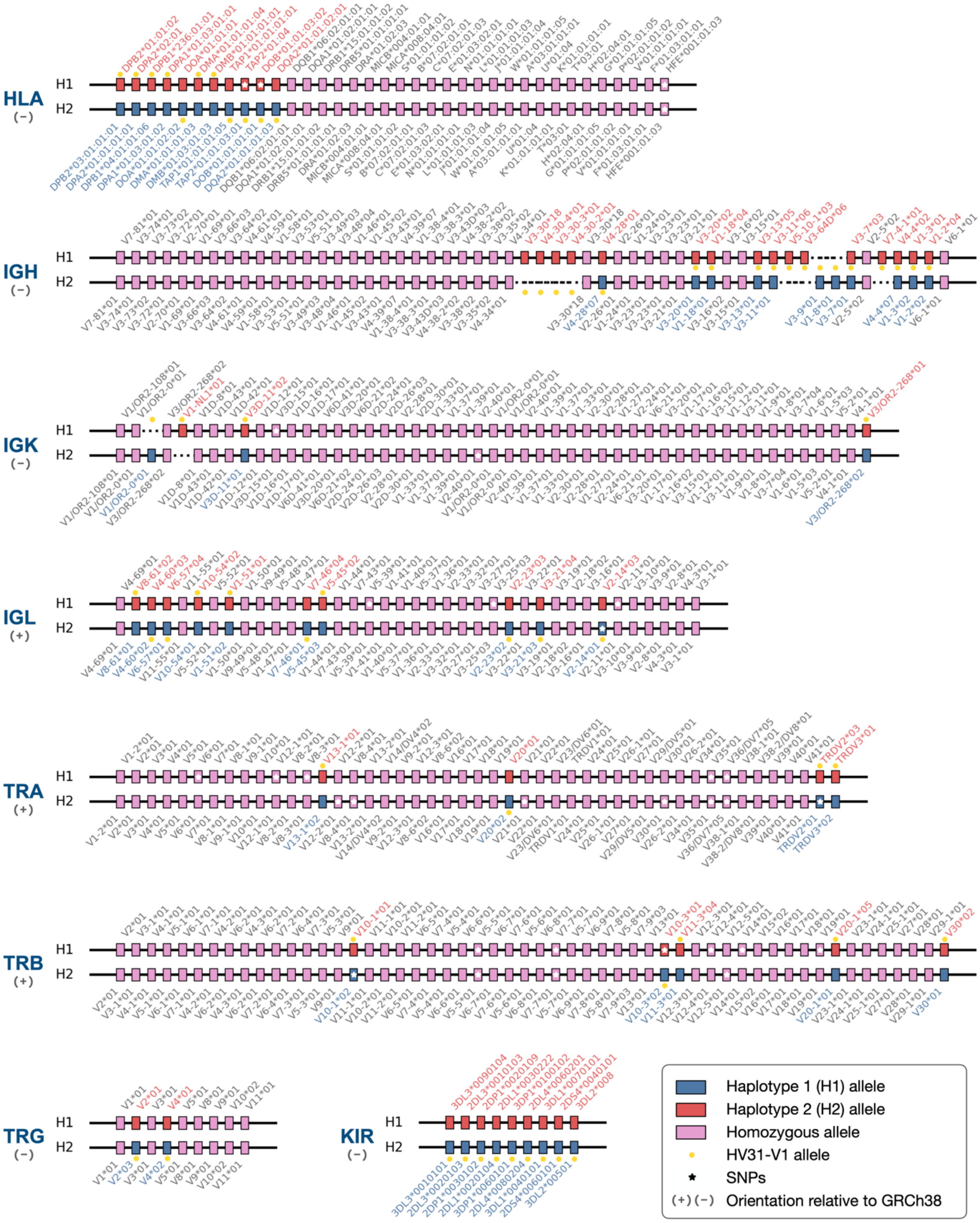


Figure 46. Allotypes of immune genes in the HV31-V2 assembly.

Heterozygous alleles in haplotype 1 (H1) and haplotype 2 (H2) are shown in red and blue, respectively. Homozygous alleles are shown in pink. Alleles represented in the HV31-V1 assembly are labelled with yellow dots. In immunoglobulin and T cell receptor regions, only V genes are shown. Genes in each region are arranged, from left to right, according to their relative order on the positive-sense strand, which may be either the plus (+) strand (with 5’ terminus at the p arm) or the minus (-) strand (with 5’ terminus at the q arm) of GRCh38. Pseudogenes are not shown. Alleles with identical sequences, such as *TRBV6-2**01 and *TRBV6-3**01, are not distinguished. This figure is related to Figure 18.

Here, I compared the allotypes of immune genes in both haplotypes revealed by the HV31-V2 assembly (Figure 46), which were identified by search the relevant IMGT or IPD databases (detailed on page 128). As expected, genomic loci of the greatest population diversity, namely the HLA, KIR, IGH and TRB loci, also featured the highest levels of heterozygosity. The HV31-V2 assembled confirmed all allotypes represented in the HV31-V1 assembly (Figure 18), and clearly showed how the HV31-V1 assembly wiggled through both haplotypes in most loci (Figure 46).

### Revisiting complex structure variants in the immune system

The complex structural variants in the IGH region revealed by the HV31-V1 assembly (Figure 19 and Figure 21) were confirmed by the HV31-V2 assembly (Figure 47). The 45 kb CNV involving *IGHV1-69* and *IGHV2-70* and the 80 kb complex duplication near *IGHV3-38* were homozygous in the HV31-V2 assembly (Figure 47), consistent with previous results (Figure 19 and Figure 21). In addition, the compound heterozygous CNV involving *IGHV3-30* was directly constructed in the HV31-V2 assembly, showing three copies of *IGHV3-30* genes on haplotype 1 and one copy on haplotype 2 (Figure 47).

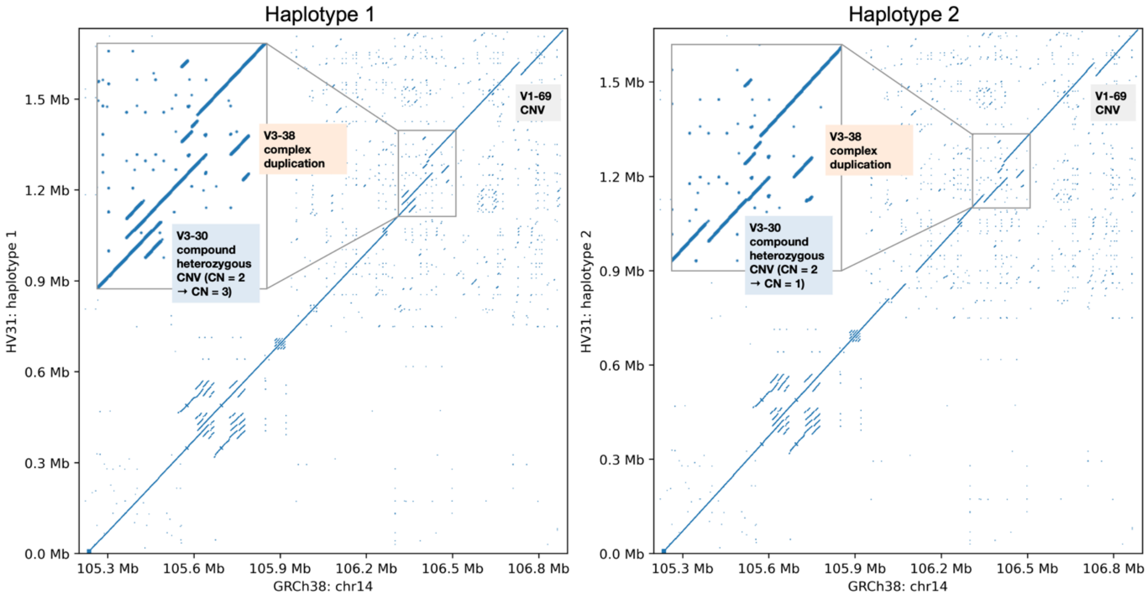


**Figure 47. The HV31-V2 assembly confirmed complex structural variants in the IGH region.**

*k*-mer sharing dot plots comparing the HV31-V2 assembly (y axis) with the GRCh38 reference genome (x axis) in the IGH region, highlighting complex structural variants in haplotype 1 (left) and haplotype 2 (right). Enlarged views of structural variants near *IGHV3-30* and *IGHV3-38* are shown (inset). This figure is related to Figure 19 and Figure 21.

The IGK region was fully assembled in one contig for each haplotype in the HV31-V2 assembly (Figure 48), an improvement over the HV31-V1 assembly, which assembled the region in two contigs, with missing sequences between them (Figure 27). The heterochromatin sequence near IGK genes described above (Figure 28) was constructed for both haplotypes, confirming the existence and location of the 32 kb unique fragment, and suggested that the heterochromatin had different sizes in the two haplotypes. Due to the highly repetitive nature of the heterochromatin sequence, it remains to be confirmed whether this size difference reflected heterozygosity or an assembly error. The locations of the 32 kb unique fragment within the heterochromatin sequence in the HV31 assemblies, including the HV31-V1 assembly and both haplotypes of the HV31-V2 assembly, were consistent with each other and with the JG1 assembly built from Japanese individuals[51] (Figure 28), but differed from the T2T reference genome, as previously described. Interrogation of individual ONT long reads from available CHM13 and HV31 sequencing data supported both locations of this 32 kb unique fragment in the corresponding genome (Appendix Figure 2 and Appendix Figure 3), suggesting that this difference was due to a structural variant rather than an assembly error.

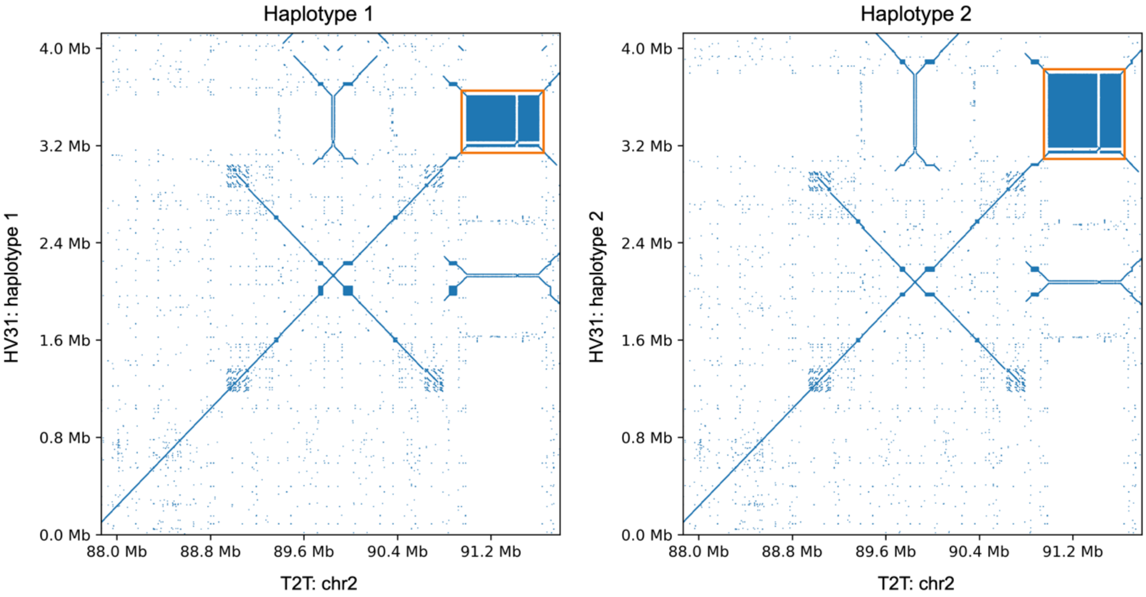


**Figure 48. The HV31-V2 assembly correctly reconstructed the IGK region.**

*k*-mer sharing dot plots comparing the HV31-V2 assembly (y axis) with the T2T reference genome (x axis) in the IGK region. Assembled heterochromatin sequences are highlighted in orange boxes. The assembly sequences were reversed and clipped for visual clarity. This figure is related to Figure 27, Figure 28 and Figure 29.

## Allele-specific expression of *HLA-DPB1*

Allele-specific expression (ASE) is the quantification of relative expression levels of the two alleles of a given gene in an individual, which helps to identify DNA polymorphism that regulates gene expression[139–141]. HLA has been a focus of ASE studies because of its disease associations as well as its high level of haplotype diversity, and by extension, high rate of heterozygosity in individuals[142–144]. ASE is typically identified by aligning RNA-Seq data to the reference genome and attributing reads to the two haplotypes based on SNP markers[140,145,146]. Here, the HV31-V2 assembly enabled personal-genome-based identification of allele-specific expression by aligning functional data to both haplotypes and count the number or coverage depths of reads on each haplotype (detailed on page 135). ASE analysis of the *HLA-DPB* gene estimated that *DPB1**236:01:01 had roughly 40% more copies of mRNA transcripts compared to the other allele, *DPB1**04:01:01:06 (Figure 49). This allelic imbalance was also associated with higher chromatin accessibility for *DPB1**236:01:01 compared to *DPB1**04:01:01:06, as measured by ATAC-Seq, and a similar, yet less pronounced, difference in histone modifications, as measured by ChIP-Seq (Figure 49). As an illustrative example, these preliminary results highlight the potential of joint analyses of genomic and functional data for the HV31 individual, though more sophisticated methods will be needed to evaluate ASE robustly across the whole genome.

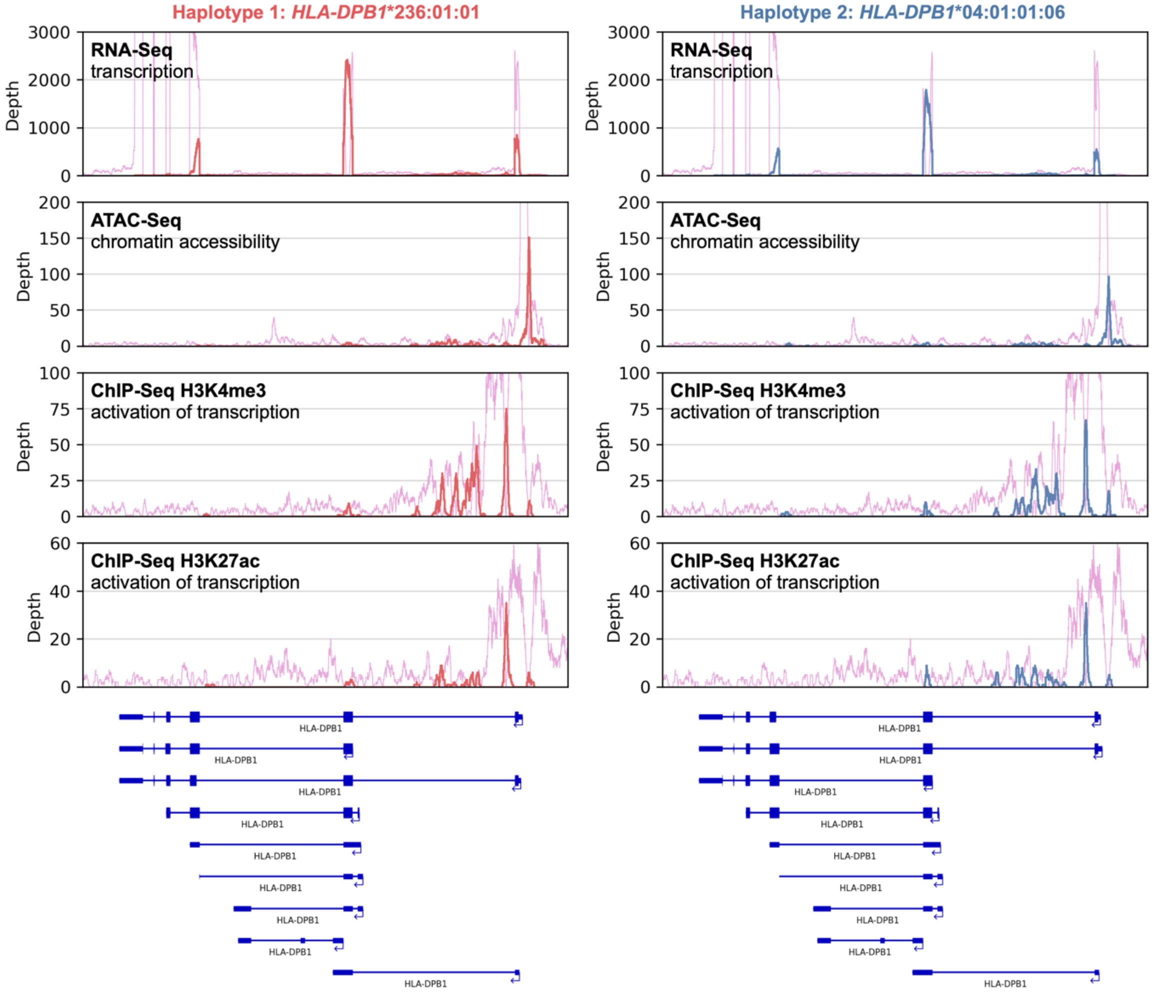


Figure 49. Allele-specific expression of *HLA-DPB1*

Coverage depths of RNA-Seq, ATAC-Seq and ChIP-Seq data aligned to the two haplotypes of the HV31-V2 assembly. Sequencing reads were assigned to haplotypes based on the edit distance of alignment. Reads assigned to haplotype 1 and haplotype 2 are represented in red and blue lines, respectively. Reads shared by the two haplotypes are represented in pink lines. Transcript annotations of the HV31-V2 assembly are shown at the bottom.

## 1.2.7. The error profile of the Oxford Nanopore R10.4.1 chemistry

Although long-read sequencing has greatly enhanced the accuracy and completeness of genome assemblies and enabled the investigation of challenging genomic features, such as structural variants and haplotype phasing, applications of long-read sequencing technologies have previously been limited by their relatively low accuracy compared to short-read sequencing[147]. Recently, Oxford Nanopore Technologies (ONT) released its R10 chemistry, a major platform upgrade which the company claimed to be significantly more accurate compared to its previous R9 chemistry[148]. In collaboration with ONT, we obtained genomic sequencing data for HV31 based on the latest available R10.4.1 chemistry in 2022 (Table 1). Here, I systematically evaluated sequencing accuracy improvements of the ONT-2022 data, compared to the ONT-2019 data which was based on the R9.4.1 chemistry. Results from these analyses later guided applications of ONT sequencing data in the design of the HV31-V2 assembly pipeline.

### Overall error profiles

One of the major challenges for accurately estimating sequencing errors in the HV31 genomic data was the lack of an established gold standard reference for HV31. The HV31-V2 assembly described above was partly based on the ONT data evaluated here, constructed at a later timepoint, and not fully validated. Hence, I was not able to use the HV31-V2 assembly as a gold standard for benchmarking ONT datasets. Considering this practical limitation, I evaluated the sequencing accuracy of ONT data against the T2T CHM13 reference genome, after excluding genomic regions of known repeats, to minimise alignment errors, and genetic variation between HV31 and CHM13, to avoid mistreating true variants as errors (Figure 50). This approach was similar those applied in previous analyses of sequencing accuracy in microbial genomes[32,76], and those in well-characterised human cell lines[31,121], such as NA12878, for which extensive sequencing data had been available[149]. I selected T2T genomic regions with no known microsatellites, segmental duplications, transposable elements or other types of repeats, and no SNPs, indels or SVs identified by independent short-read or long-read datasets, as high-confidence regions (detailed on page 137). Within the high-confidence regions, I defined the sequencing error rate in a given dataset as the total number of substituted, inserted and deleted bases, divided by the total number of sequenced bases (detailed on page 138).

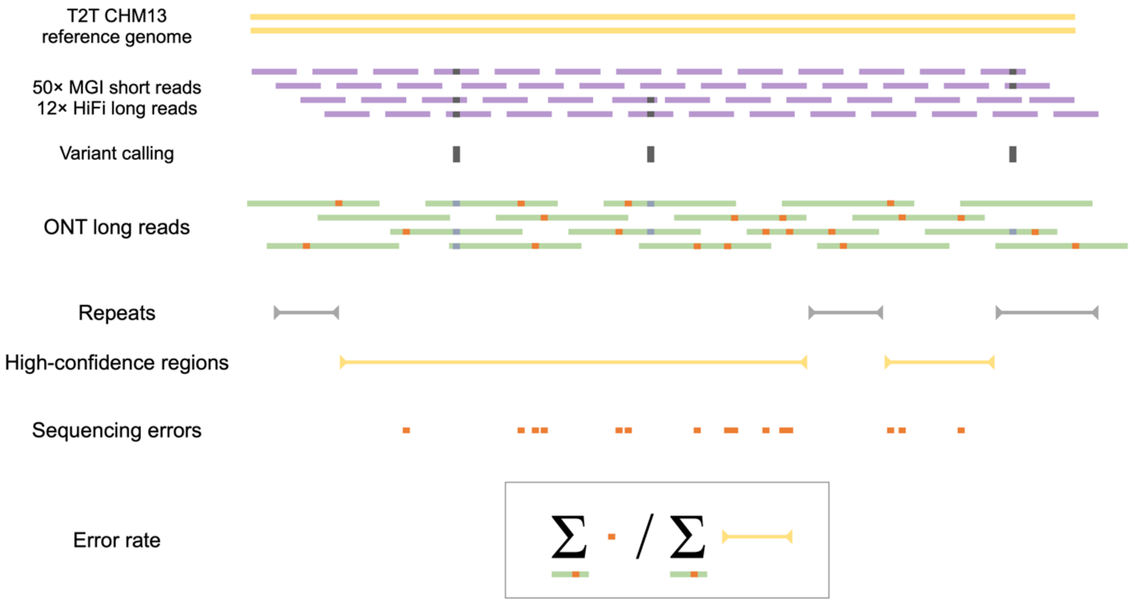


Figure 50. Schematic overview of the sequencing accuracy evaluation pipeline.

Variant calling was performed separately for MGI short reads using DeepVariant[150], and for HiFi long reads using pbsv[42].

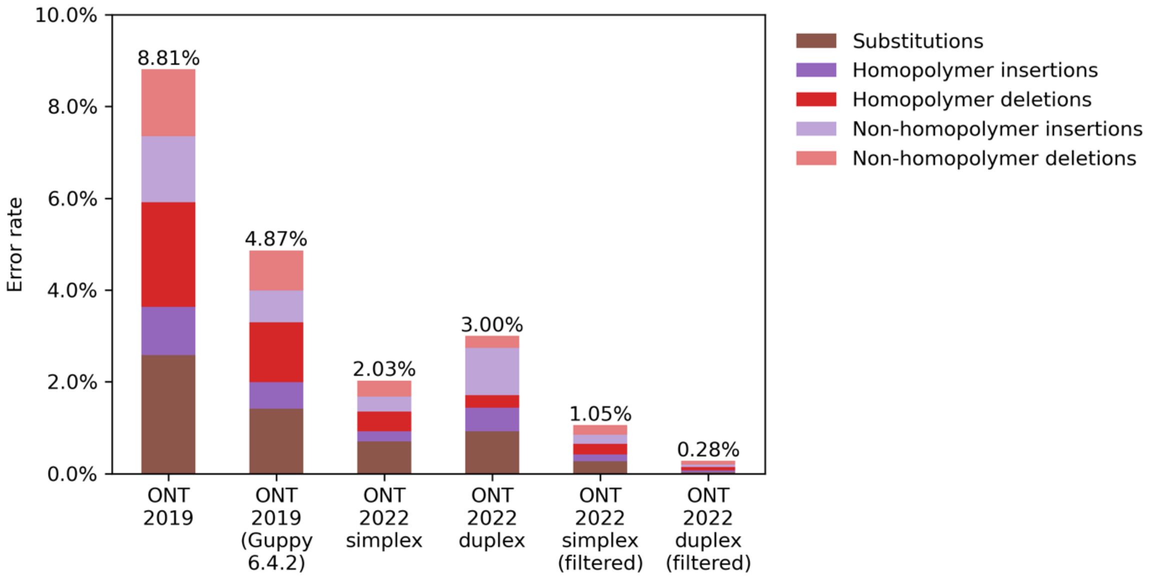


Figure 51. Overall error profiles of evaluated ONT datasets.

The contribution of each type of errors (colours) to the overall error rate (y axis) in each ONT dataset (x axis). Error rates were defined as the number of error bases divided by the number of total sequenced bases.

These analyses confirmed the sequencing accuracy of R10.4.1 chemistry was considerably improved over the previous R9.4.1 chemistry. The overall error rates were 2.03% for ONT-2022-simplex data and 3.00% for ONT-2022-duplex data, both of which were generated from the same R10.4.1 sequencing run, compared to 8.81% for the ONT-2019 data based on the R9.4.1 chemistry (Figure 51). Considering that simplex reads represented 94.6% of total data yielded from the R10.4.1 sequencing run (Table 1), these estimates suggested the R10.4.1 chemistry had an over four-fold accuracy improvement compared to the R9.4.1 chemistry.

Basecalling, which is the conversion from raw data generated from the sequencing chemistry (usually in the form of optical or electrical signals) to nucleic acid sequences, is a critical step for generating accurate long reads[29]. The ONT-2019 data was generated using Guppy v3.1.5, a proprietary basecalling software developed by ONT for which multiple iterations of optimisation had taken place ever since. Therefore, it was likely that part of the accuracy improvement observed for the ONT-2022 data was caused by improvements of the basecalling software, rather than the underlying sequencing chemistry. This distinction was important because basecalling only accounts for a small fraction of the overall cost of sequencing, and could be rerun at a low marginal cost when new basecalling software became available. To test this hypothesis, I re-basecalled the ONT-2019 data with Guppy v6.4.2, the latest version available then. I found that the resulting data had an overall error rate of 4.87%, which was considerably better than the original ONT-2019 data but still less accurate compared to the ONT-2022-simplex data (Figure 51). This confirmed the contribution of the basecalling method in the improved accuracy, and suggested a cost-effective way to obtain more accurate information from existing ONT sequencing data.

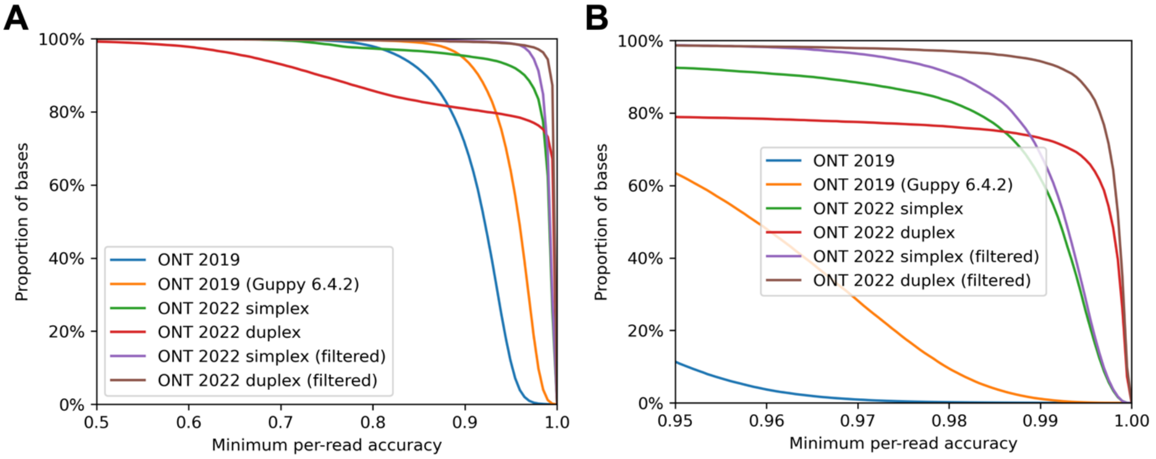


**Figure 52. Per-read accuracy distribution of evaluated ONT datasets.**

**(A)** The maximum proportional of total bases retained after filtering (y axis) plotted against the minimum required per-read accuracy (x axis) for each evaluated dataset (colours).

**(B)** Same as **(A)**, zoomed in for clarity.

Despite the overall accuracy improvement of the R10.4.1 chemistry described above, the duplex reads appeared less accurate compared to the simplex reads (Figure 51), inconsistent with the claimed higher accuracy of duplex reads[148]. To investigate this, I calculated the accuracy for each individual read in the evaluated datasets, and found that the minimum per-read accuracy of ONT-2022-duplex reads could be improved to 99% after filtering out less than 30% of total bases, while almost no read in the ONT-2022-simplex data could fulfil this criterion (Figure 52). This observation suggested that the apparent lower accuracy of duplex reads was due to the presence of a small number of low-quality reads in addition to a majority of high-quality reads, the latter having better accuracy than simplex reads.

Meanwhile, Gavin Band found that (i) part, but not all, of low-quality reads in the ONT-2022-simplex and ONT-2022-duplex data were explained by around 0.5%

of total reads being contamination; and (ii) both ends of each read in all evaluated ONT datasets were enriched of sequencing errors, compared to other parts of the read. Qijing Shen confirmed that the base quality scores of evaluated ONT datasets were predictive, to varying degrees depending on the specific dataset, of sequencing errors. In response to these findings, Gavin Band produced filtered subsets of ONT-2022-simplex and ONT-2022 data, by removing reads with low mean base qualities (12% of simplex reads and 23% of duplex reads) and trimming 100 bases from each end of remaining reads. The resulting filtered datasets displayed significantly higher accuracies compared to their unfiltered counterparts (Figure 51). In particular, filtered ONT-2022-duplex data had an error rate of only 0.28%, which was higher than the filtered ONT-2022-simplex data, and was better than the estimated error rate of 0.42% for HiFi-2022 data (data not shown). In light of this finding, I subsequently based the HV31-V2 assembly on the filtered ONT-2022 simplex and duplex reads.

### Substitution errors

Substitution errors, defined as a base being mistakenly sequenced as a single, different base, are analogous to SNPs and may cause false positives in SNP calling[151]. ONT R9.4.1 chemistry has been shown to produce a disproportionally large amount of A-to-G and G-to-A substitution errors[76], likely due to the similar chemical structure of adenosine and guanosine bases. For the same reason, A-to-G and G-to-A transitions are also the most common types of human SNPs, accounting for about 32% of all SNPs, followed by C-to-T and T-to-C transitions, accounting for about 29% of all SNPs[152].

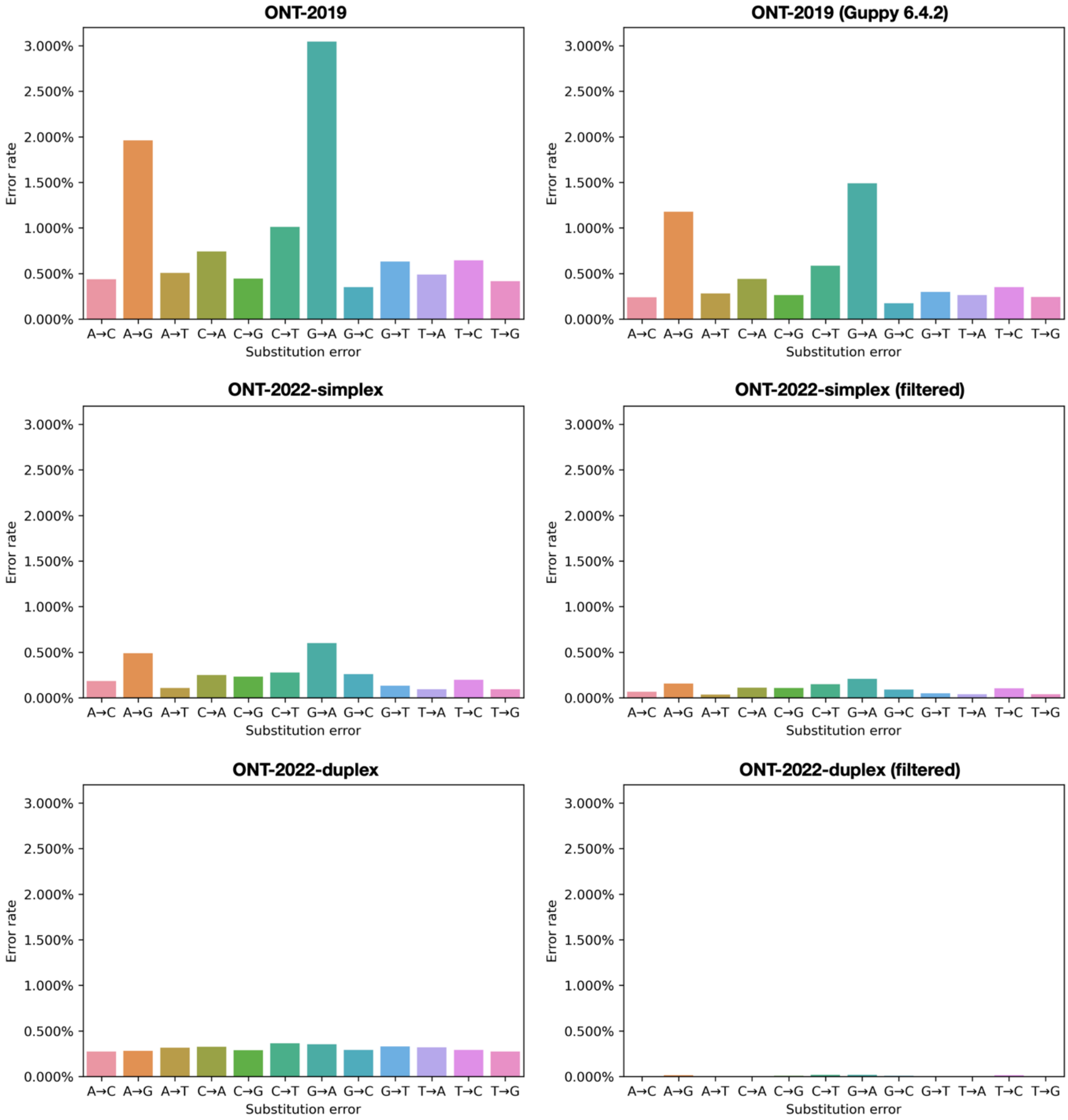


**Figure 53. Substitution error rates of evaluated sequencing datasets.**

For each dataset, the x axis shows possible types of substitution errors, represented as the true base, an arrow, and the read base. For example, A→G denotes an adenosine base mistakenly sequenced as a guanosine base. The y axis shows the corresponding substitution error rate, conditioned upon the true base. For example, if the A→G substitution has an error rate of 1%, then 1% of all adenosine bases will be mistakenly sequenced as guanosine bases.

Here I analysed substitution error rates grouped by the nucleotides involved, and confirmed the enrichment of A-to-G and G-to-A errors in the ONT-2019 data from R9.4.1 chemistry (Figure 53). I also observed similar enrichment in the ONT-2022-simplex data from R10.4.1 chemistry, despite improved overall accuracy, but not ONT-2022-duplex data from the same chemistry. The filtering strategy significantly reduced the substitution error rates, while the relative contribution of each type of substitution remained largely the same.

### Homopolymer errors

Homopolymers are repeats of identical bases, such as GGGG. Homopolymer errors, loosely defined as expansions or contractions of homopolymer lengths relative to true lengths, have been known to contribute to a large proportion of total errors in PacBio[55] and ONT[76] sequencing data. This phenomenon is unique to long-read sequencing, presumably due to the fact that long-read sequencing happen in real time by design, without chemically interrupted cycles present in Sanger or Illumina sequencing[29,31], preventing homopolymer lengths to be reliably inferred from the number of cycles.

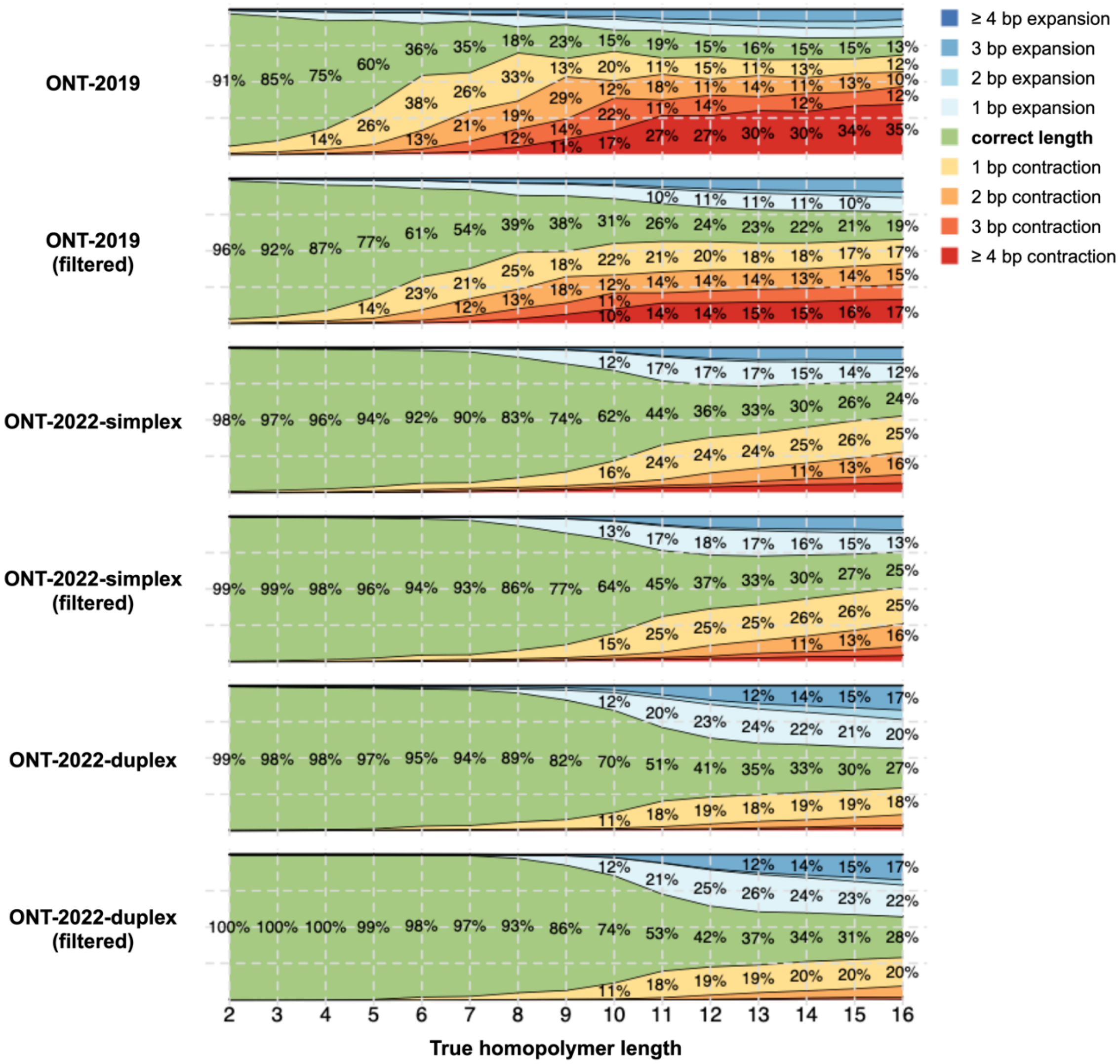


**Figure 54. Homopolymer accuracy of evaluated sequencing datasets.**

For each dataset, the length distribution of sequenced homopolymers (y axis) relative to the true homopolymer lengths (x axis) is visualised. Text labels indicate the proportions of each group. Visualisation made by Gavin Band.

Here, I analysed the distributions of homopolymer errors grouped by true homopolymer lengths (Figure 54). I found that homopolymers sequenced with the R9.4.1 chemistry, especially those longer than five nucleotides, were significantly biased towards shorter lengths. This bias was likely a compromise by design to achieve higher overall homopolymer accuracy, given that longer homopolymers are proportionally rarer in the human genome. In comparison, simplex reads from the R10.4.1 chemistry showed less bias in homopolymer lengths, while duplex reads showed almost no bias, enabling the accurate estimation of true homopolymer lengths from multiple overlapping reads. I also noted that homopolymer error rates for R10.4.1 chemistry only started to rise sharply above 10 nucleotides, presumably because of the improved pore design[29].

### Short tandem repeats

In addition to homopolymers, short tandem repeats (STRs) were shown to be difficult targets for ONT sequencing[29]. Certain STRs are strongly associated with genetic disorders, such as the Huntington's disease and the fragile X syndrome[153]. Here, I visualised the distribution of repeat lengths in ONT reads that covered the *HTT* tandem repeats which is linked to the Huntington's disease, along with PacBio HiFi reads for comparison (Figure 55). The *HTT* tandem repeats consists of two repeat units, `CAG` and `CCG`, and excessive (36 or more) copies of the `CAG` repeat predisposes the carrier to the Huntington's disease in a autosomal dominant manner[154]. The repeat length distributions of suggested that HV31 carried 13 copies of CAG (39 nucleotides) on one haplotype, and 20 copies (60 nucleotides) on the other haplotype (Figure 55). I found that ONT-2022 reads from the R10.4.1 chemistry displayed a tight distribution in repeat length, similar to that of HiFi-2022 data, in which most reads reported the expected repeat length, while ONT-2019 reads from the R9.4.1 chemistry had a broader repeat length distribution, which could cause ambiguities in determining the exact repeat lengths on each haplotype.

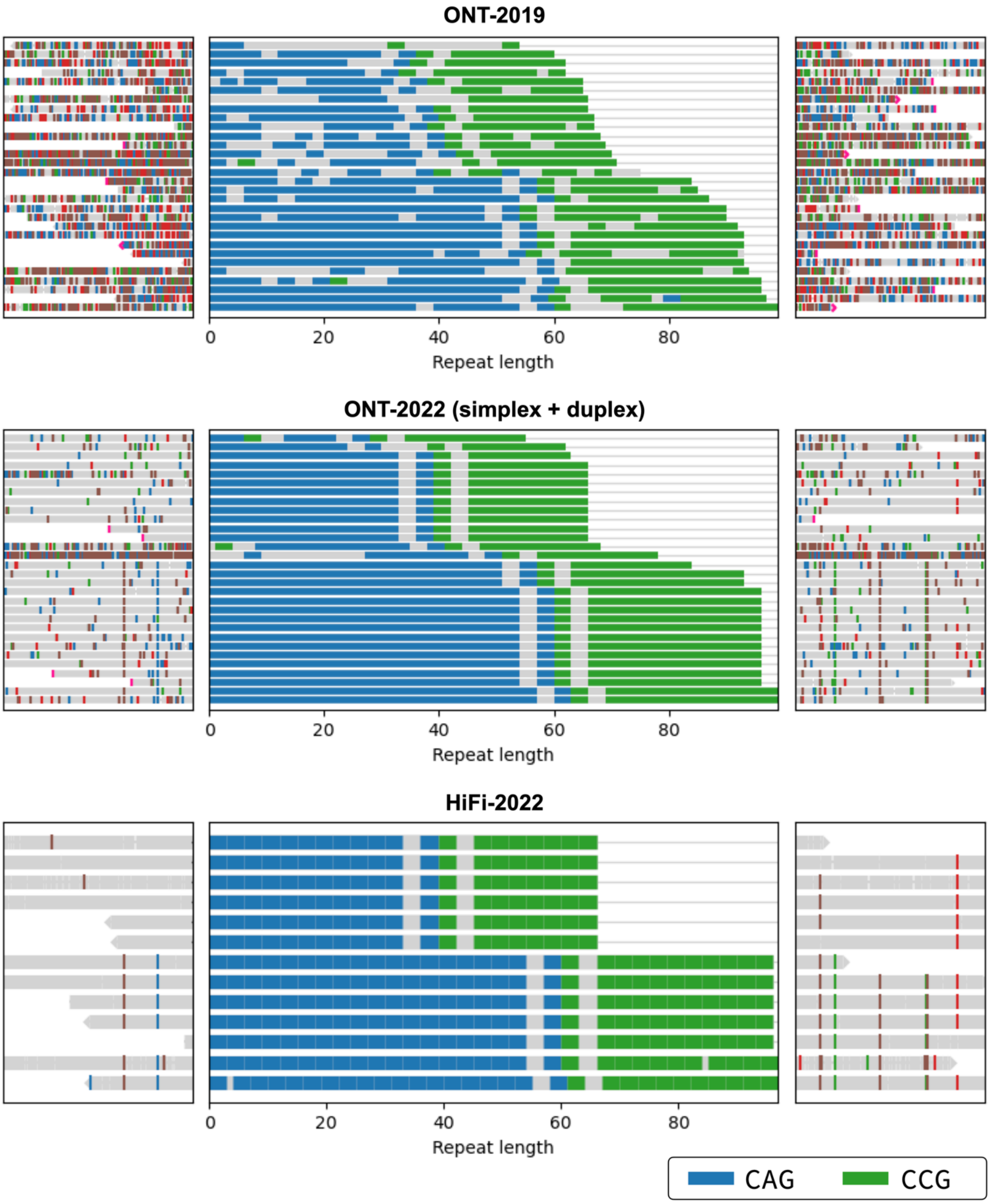


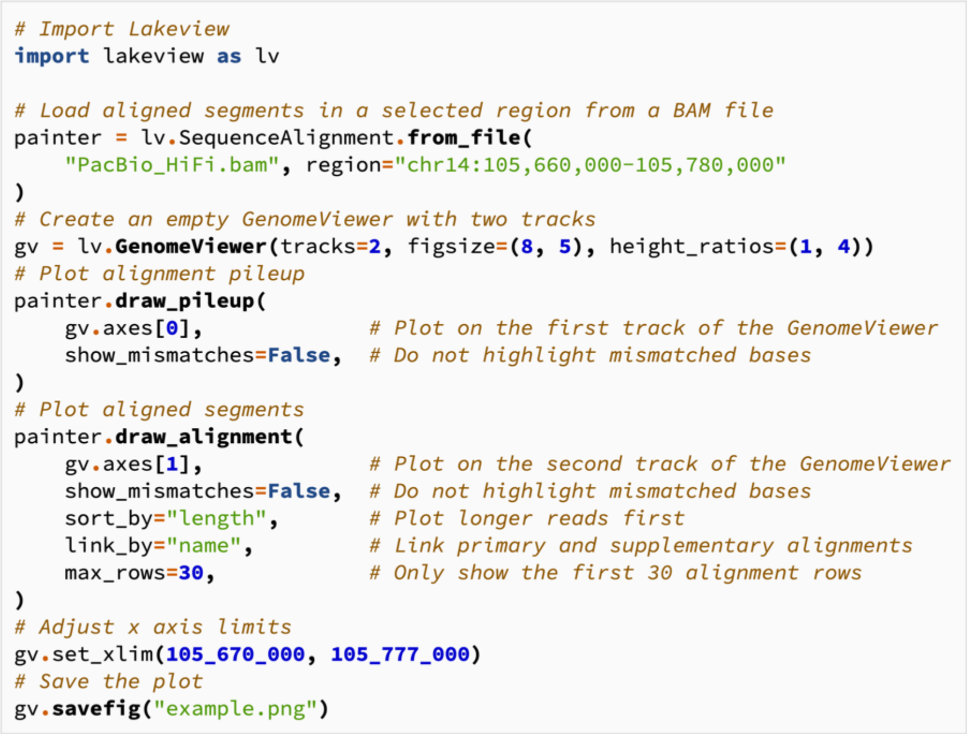
Figure 55. Short tandem repeats associated with the Huntington's disease.

Sequencing reads covering short tandem repeats associated with the Huntington's disease. Each row represents a single read, shown in three parts: the tandem repeat sequence (middle) and 5 kb flanking sequence upstream (left) and downstream (right) to the tandem repeats. CAG and CGG repeat units within each repeat sequence are shown in blue and green, respectively. Single-nucleotide mismatches in each flanking sequence that are different from the T2T reference sequence are shown as vertical line segments coloured according to the sequenced base. ONT reads were downsampled to 30× coverage depth for clarity. Visualisation inspired by TRGT[155].

### 1.2.8. Lakeview: a modular framework for genomic data visualisation

Data visualisation is an integral part of genomics analyses[156]. Various tools have been development for genomic data visualisation, most notably the Integrative Genome Viewer (IGV)[157], which provides a versatile and user-friendly interface for visualising common bioinformatics data formats. Designed for interactive exploration, the IGV provides only limited support for programmatic access, which restricts its application in certain use cases. For example, it remains challenging to use IGV for plotting bulk or clinically sensitive data stored on a remote server without a display device. In addition, although data-dependent customisation of plots, such as highlighting sequencing reads that satisfy certain predefined rules, is possible in IGV, such customisation often leads to difficulties in reproducing the plot once the underlying data changes. Apart from functionality issues, the visual design of IGV is also oriented towards interactive use, which necessitates post-processing steps when using IGV output in publications, such as removing control elements and increasing text label sizes (Figure 3). Various alternatives for genomic data visualisation have been developed[158–160], each with their own design focus, but these tools typically lack the visual clarity of IGV that many researchers have already been familiar with.

Here, motivated by the specific needs arising from the HV31 project, I developed Lakeview, a modular framework for genomic data visualisation, available as an open-source Python library (https://pypi.org/project/lakeview/). Built on top of Matplotlib[161], Lakeview is a Python 3 library for creating publication-quality genomic visualizations. Lakeview inherits the familiar and intuitive visual style of IGV, with a clear layout designed for publication and presentation (Figure 56). For remote data, visualisations can be programmatically created on the server storing the data files, and transmitted to the user for inspection. By avoiding the direct transfer of the genomic data to the user, this workflow saves local storage, computation and network resources, and may be desirable for bulk data or data that contains clinically sensitive information. Various figures included in this thesis were created using Lakeview following this workflow.



```python
# Import Lakeview
import lakeview as lv

# Load aligned segments in a selected region from a BAM file
painter = lv.SequenceAlignment.from_file(
    "PacBio_HiFi.bam", region="chr14:105,660,000-105,780,000"
)
# Create an empty GenomeViewer with two tracks
gv = lv.GenomeViewer(tracks=2, figsize=(8, 5), height_ratios=(1, 4))
# Plot alignment pileup
painter.draw_pileup(
    gv.axes[0],             # Plot on the first track of the GenomeViewer
    show_mismatches=False,  # Do not highlight mismatched bases
)
# Plot aligned segments
painter.draw_alignment(
    gv.axes[1],             # Plot on the second track of the GenomeViewer
    show_mismatches=False,  # Do not highlight mismatched bases
    sort_by="length",       # Plot longer reads first
    link_by="name",         # Link primary and supplementary alignments
    max_rows=30,            # Only show the first 30 alignment rows
)
# Adjust x axis limits
gv.set_xlim(105_670_000, 105_777_000)
# Save the plot
gv.savefig("example.png")
```

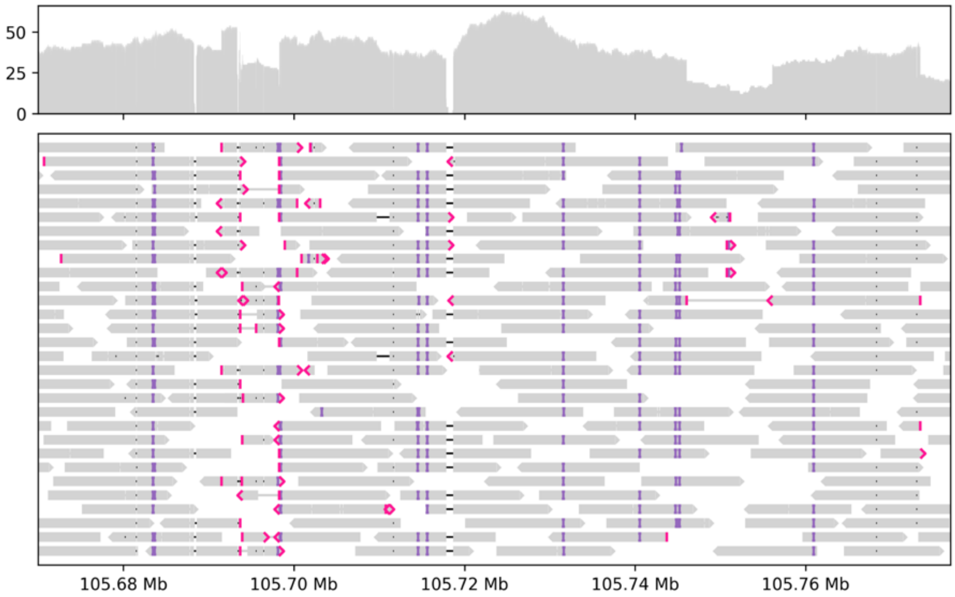


**Figure 56. Lakeview enables clear and reproducible visualisation of genomic data.**

A demonstrational code example (top) for plotting the coverage depths and alignment patterns in a given genomic region from a BAM file using Lakeview, and the corresponding output.

Designed with a focus on reproducibility, the output figures produced by Lakeview, including any customisations or annotations, are uniquely defined by the input data and the specific code used for creating the visualisation (Figure 56). This makes Lakeview suitable for making visualisations that need to be reproduced for multiple samples or genomic regions of interest, at a later time when the data changes, or by a different user working in a different environment.

Lakeview is a collection of several modular components, each responsible for parsing and visualising a certain type of genomic data. Currently, Lakeview supports the visualisation of sequence alignment (Figure 23, Figure 43, Figure 55, Appendix Figure 2, and Appendix Figure 3), coverage depth (Figure 27A and Figure 49), gene annotation (Figure 21, Figure 24 and Figure 49), and sequence comparison (Figure 12, Figure 27A, Figure 47, Appendix Figure 2, and Appendix Figure 3). Support for additional data types can be added in the future as additional modules that follow a consistent interface design. Lakeview was published under the GNU General Public License v3.0, with comprehensive documentation and automated testing available for prospective users and developers.

## 1.3. Discussion

The human adaptive immune system is encoded by a set of complex and highly diverse genes which protects the body from myriad environmental pathogens both at the individual level and for the long-term survival of the species. Genetic variation in these genes, such as those encoding the HLA and immunoglobulins, despite the technical difficulties in their characterisation, have been associated with various infection and autoimmune diseases[162–165]. Here, taking advantage of recent advances of high-throughput, long-range genomics technologies, in particular long-read sequencing from PacBio and ONT, I explored the frontier of human *de novo* genome assemblies based on the rich collection of genomic data available for the HV31 individual, and investigated complex structural variation found inside the HV31 personal genome.

In the first phase of the HV31 project, I constructed the HV31-V1 assembly, focusing on the eight selected genomic loci encoding key components of the immune system. Despite being a mixed-haplotype assembly due to technical limitations, the HV31-V1 assembly, after Bionano scaffolding and short-read polishing steps, was reasonably complete, accurate and contiguous in the selected regions. This revealed several large structural variants involving key immune genes which were otherwise challenging to characterise, many of which were inspected in depth and validated using a combination of methods based on *k*-mer coverage depths, read alignment, and optical mapping.

Further collaboration with PacBio and ONT, and the rapid development of bioinformatics algorithms tailored for long-read sequencing, enabled the second phase of the HV31 project aimed at building a complete personal genome. Assembly graph construction based on PacBio and ONT long reads, followed by path resolution based on MGI stLFR linked reads and *k*-mer profile modelling, yielded the HV31-V2 assembly, a haplotype-resolved assembly reaching telomere-to-telomere contiguity for multiple chromosomes. Revisiting the complex structural variants identified previously, I showed that the HV31-V2 assembly provided the missing pieces required for mapping the HV31 personal genome.

Nevertheless, several aspects of the current HV31-V2 workflow warrant further improvement. For example, the resolved paths of complex bubbles should optimally be phased using stLFR data in a way similar to simple bubble paths. In addition, given that MEC is an NP-hard problem[129], tailoring the HapCUT2 heuristic algorithm[135] based on unique features of bubble phasing and stLFR fragments may yield improved phasing results.

The analyses of ONT sequencing accuracy were an effort to understand the technical characteristics of ONT sequencing, the result of which guided the read filtering strategy applied in the HV31-V2 workflow. A recurring challenge of the HV31 project is the lack of a known ground truth for the HV31 genome, which was alleviated, in the case of error rate analyses, by comparing the sequencing data with the T2T reference genome after excluding regions of identified genetic variants. The considerable accuracy improvements achieved by the ONT R10.4.1 chemistry, as confirmed here, showed promise towards a future of affordable high-quality personal genomes.

Lakeview is a standalone Python library for genomic data visualisation, which evolved from specific use cases in the HV31 project, and underpinned many figures included here. With continued iterations, the library will likely become a modular toolbox for versatile visualisation of various genomic data formats, from which more sophisticated and user-friendly applications might be built.

I envision that the progress made in HV31 project will eventually contribute to the scientific community from three aspects: (i) the data generated from the HV31 individual will serve as a rich resource, similar to the Genome in a Bottle project, that fuels technical understanding of the underlying platforms and methodology development in bioinformatics; (ii) the bioinformatic approaches I developed for sequencing quality analyses, *de novo* assembly, assembly validation and functional data analyses may be reused by future researchers interested in personal genomes; (ii) specific findings from the HV31 project, such as structural variants in the IGH region, the heterochromatin structure in the IGK region, or allelic imbalance of HLA genes, may shed some light on the genetic complexity of the human adaptive immune system. With a long-standing focus on the application of personal genomes in the immune system, the HV31 project will move from current analyses which primarily investigate DNA sequences, towards a more comprehensive exploration of genetic variants in the immune system in terms of proteins, pathways, cells and diseases.

## 1.4. Methods

### 1.4.1. 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.

### 1.4.2. Definition of regions of interest

Eight genomic regions encoding key components of the human immune system, including HLA, IG, TCR and KIR were selected for investigation (Table 2). 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. For IG and TCR regions, the core range were selected based on the respective reference sequences in the NCBI RefSeq database[89]. For the HLA region, the core range was defined as the genomic range from *GABBR1* to *KIFC1*[90]. For the KIR region, the core range was defined as the genomic range from *KIR3DL3* to *KIR3DL2*[14]. 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, I expanded the centromeric flanking sequence in the IGK region by 0.67 Mb to bridge a 1 Mb heterochromatin gap present in GRCh38. A similar approach was applied to retrieve the corresponding coordinates of each region of interest in the T2T reference genome.

### 1.4.3. Construction of the HV31-V1 mixed-haplotype assembly

### Canu whole-genome de *novo* assembly

Canu v1.9[56] was used to perform whole-genome *de novo* assembly for HV31 based on HiFi-2019 reads, with the following command:

```
canu \
  -pacbio-hifi FASTQ_PATH \
  genomeSize=3235000000 \
  -minInputCoverage=1 \
  -stopOnLowCoverage=1
```

The resulting contigs were mapped to GRCh38 using Minimap2[166] with the following parameters: `-ax asm5 --secondary=no`. Contigs that mapped to the 8 loci of interest were extracted as local contigs.

### Peregrine whole-genome assembly

For comparison purposes, Peregrine[167] was used to generate a whole-genome *de novo* assembly for HV31 based on HiFi-2019 reads, with the following command:

```
python PEREGRINE_SCRIPT \
  asm FASTQ_LIST \
  16 16 16 16 16 16 16 16 16 \
  --with-consensus
```

### Hybrid scaffolding

Hybrid scaffolding was performed using Bionano Solve, a proprietary software provided by Bionano Genomics (https://bionanogenomics.com/), with default parameters. I used a custom script based on BiSCoT[168] to improve the contiguity and quality of the resulting scaffolds. Specifically, I 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 (a sequence of “N” bases) between them, the size of which was estimated based on the distance of nearest labelling sites. In addition, I 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, I removed duplicated contigs or scaffolds that presumably represent alternative haplotypes (haplotigs) using a custom *k*-mer based method. In brief, I 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, I first mapped the reads from each sequencing dataset using Minimap2, and then extracted reads that mapped to contigs that represent each locus of interest[133]. 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[100] 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, I 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 ≤ depth ≤ 231) and the reference sequence (copy number = 1; no occurrence outside the locus) were selected as anchoring *k*-mers. Then, I 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 was 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[98] v1.0.1 with HiFi-2019 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[99] with HiFi-2019 reads and MGI-standard 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 visualisation steps.

## 1.4.4. Evaluation of the HV31-V1 assembly

### Per-base error rate estimation

Jellyfish[169] was used to count the depth of each *k*-mer (k = 22 or 31) from a pooled FASTQ dataset consisting of the following datasets: HiFi-2019, MGI-standard, MGI-CoolMPS, MGI-stLFR, 10X and Illumina (Table 1), with the following parameters: `jellyfish count -m K -s 30G —min-qual-char "?" -C`. The cumulative sequencing depth of the pooled FASTQ dataset was 262×. In each read, *k*-mers that include bases with base quality < 20 were excluded. For error rate estimation, I applied a modified version of the Merqury[104] method. I estimated the error rates based on clusters of kmers with low validation coverage, which produced slightly higher error rate estimates empirically, compared to the original Merqury method, which counts each error *k*-mer individually[104] (Table 5). Specifically, *k*-mers (*k* = 22) in the HV31-V1 assembly with depth < 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-V1 assembly.

| | **Merqury** | | **This work** | |
|---|---|---|---|---|
| | Before polishing | After polishing | Before polishing | After polishing |
| IGH | 88.5 | 6.7 | 86.6 | 8.1 |
| IGK | 223.2 | 24.2 | 183.0 | 42.3 |
| IGL | 78.8 | 8.0 | 68.2 | 11.9 |
| HLA | 32.4 | 3.3 | 39.5 | 5.1 |
| TRA | 71.7 | 18.9 | 85.7 | 22.8 |
| TRB | 62.6 | 7.4 | 52.2 | 9.6 |
| TRG | 76.6 | 18.1 | 86.0 | 21.2 |
| KIR | 78.1 | 18.1 | 94.2 | 25.0 |

**Table 5. Comparison of original and modified Merqury error rate estimation results.**

Per-base error rate estimates for local scaffolds and the finished HV31-V1 assembly (Figure 16) using the Merqury algorithm, and the modified algorithm applied here.

### Identification of potential structural errors

For reference-free identification of potential structural errors, I define the normalized depth ($d$) of each $k$-mer ($k$ = 31) in the HV31-V1 assembly as $d = D / (C \times M)$, in which $D$ is the depth of that $k$-mer in the validation dataset, $C$ is the copy number of that $k$-mer in the HV31-V1 assembly, and $M$ is the mode depth of unique homozygous $k$-mers in the validation dataset, as estimated from the $k$-mer depth histogram (Figure 14). The normalized $k$-mer coverage was visualized against the position of the $k$-mer, along with the normalized coverage of ONT-

2019 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 (Appendix Figure 1).

## 1.4.5. Characterisation of allelic and structural variation

### Allelic variant detection

Reference variant sequences of IGHV, IGKV, IGLV, TRAV, TRDV, TRBG and TRGV genes were downloaded from the IMGT reference directory[170]. Reference variant sequences of HLA genes were downloaded from the IPD-IMGT/HLA database[171]. Reference variant sequences of KIR genes were downloaded from the IPD-KIR database[172]. The reference gene variant sequences were mapped to GRCh38, the HV31-V1 assembly, or the HV31-V2 assembly using Minimap2 with the following parameters: `-a -w1 -f1e-9`. I 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[173] (for IGHV, IGKV, IGLV, TRAV, TRDV, TRBV and TRGV genes) or NCBI BLAST+[174] (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.

### Structural variant calling

PBSV[42], a subprogram of SMRT tools was used to call heterozygous SVs from HiFi-2019 and CLR-2019 reads with default parameters. Sniffles[43] was used to call heterozygous SVs from HiFi-2019, CLR-2019 and ONT-2019 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[101] was used to cluster and merge SV records from output VCF files of PBSV and Sniffles, with default parameters.

### 1.4.6. Construction of the HV31-V2 haplotype-resolved assembly

#### Trimming and filtering low-quality sequences in the ONT-2022 dataset

Filtered versions of the ONT-2022-simplex and ONT-2022-duplex datasets were generated based on base quality scores, excluding reads that had less than 80% (for simplex reads) or less than 96% (for duplex reads) bases that had quality scores of at least 20. 100 base pairs were subsequently trimmed off both ends of each read.

The trimming and filtering strategy was developed by Gavin Band based on empirical evidence gathered by Gavin Band, Qijing Shen, and from my analyses of ONT sequencing accuracy.

#### Verkko *de novo* assembly

Verkko v1.3.1[64] was run twice during the HV31-V2 assembly workflow, first to generate the assembly graph, and then to construct contigs from read consensus along resolved paths. The first run was executed using pooled PacBio HiFi reads from HiFi-2019, HiFi-2022 and HiFi-2023 datasets, and trimmed filtered ONT-2019 reads, as inputs. The following command was used:

```
verkko \
  -d OUTPUT_FOLDER \
  --hifi INPUT_HIFI_FASTQ \
  --nano INPUT_ONT_FASTQ
```

After the first run, haplotype paths were identified from the assembly graph by resolving complex bubbles and phasing simple bubbles, as described below. The resulting haplotype paths were then used to replace the original paths generated by Verkko during the first run, which, in the absence of parental sequencing data or other forms of long-range phasing information natively supported by Verkko, represented each unitig as a path. Finally, Verkko was run for the second time using the same command, producing the HV31-V2 assembly.

### Enumeration of bubbles and bubble chains

Simple and supper bubbles were defined according to Onodera *et al*.[126]. For naming consistency, here “simple bubble” refers to “bubble” in the work of Onodera *et al*., and “super bubble” refers to “superbubble”. Cyclic bubbles were defined as subgraphs that contain cycles but otherwise satisfy the definition of “superbubbles” in the work of Onodera *et al*.[126] Formal definitions are described below:

Let $G = (V, E)$ be a directed assembly graph, where $V$ is the set of nodes and $E$ is the set of edges. If an ordered pair of distinct nodes $(s, t)$ in $G$ satisfies the following conditions:

(i) $t$ is reachable from $s$;

(ii) the set of nodes reachable from $s$ without passing through $t$ is equal to the set of nodes reachable from $t$ without passing through $s$;

(iii) no node in $U$ other than $t$ forms a pair with $s$ that satisfies conditions (i) and (ii), where $U$ the subgraph induced by the set of nodes in condition (ii),

then the subgraph $U$ defined in condition (iii) is said to be a *bubble*, for which $s$ and $t$ are referred to as the source and sink nodes, respectively. Other nodes in $U$ are referred to as internal nodes. If $U$ satisfies the following conditions:

(iv) no cyclic paths, defined as paths that visit the same node more than once, exist in $U$;

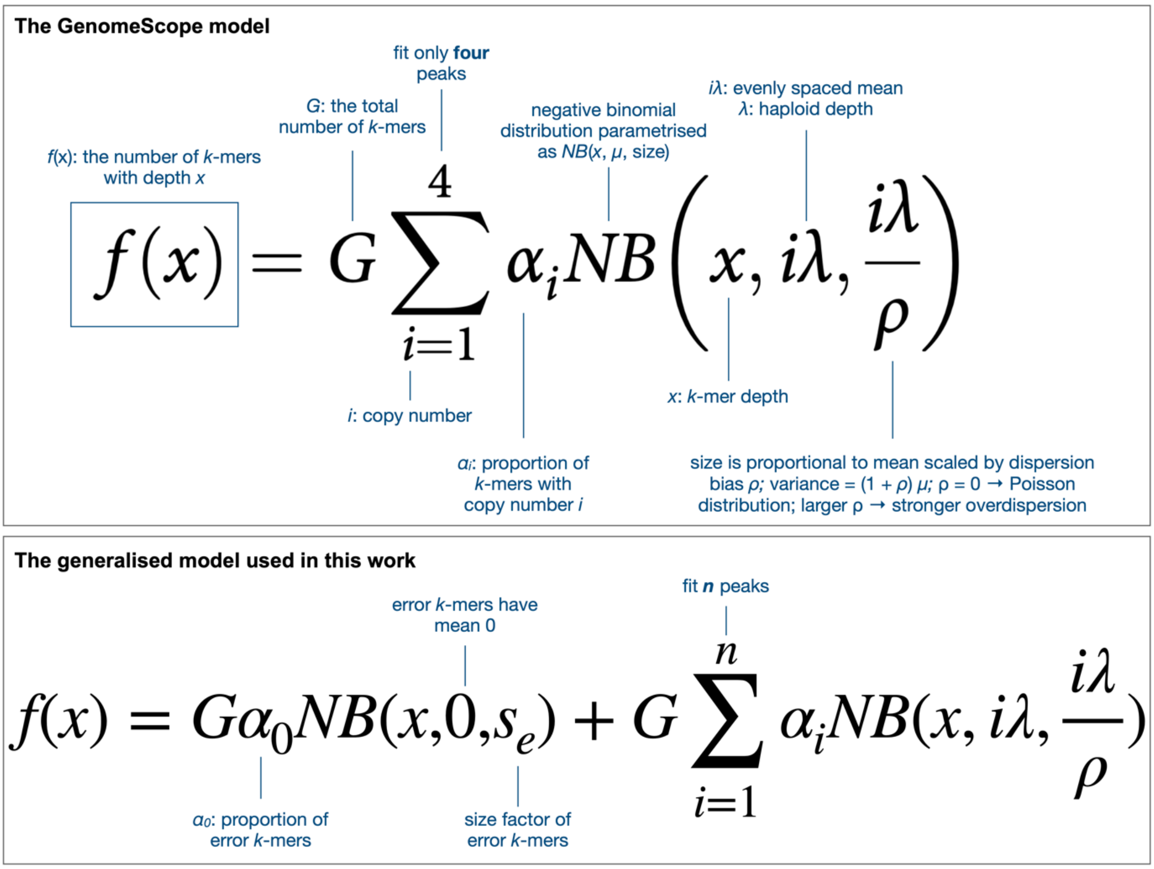
(v) only two internal nodes exist in $U$, both of which are directly connected to $s$, then $U$ is said to be a *simple bubble*, for which the two internal nodes in condition (v) are referred to as arms. If $U$ satisfies condition (iv) but does not satisfy condition (v), then $U$ is said to be a *super bubble*. If $U$ does not satisfy condition (iv), then $U$ is said to be a *cyclic bubble*. A *bubble chain* is defined as an ordered sequence of bubbles $U_1, U_2, \ldots, U_n$ such that the sink node of bubble $U_i$ is the source node of bubble $U_{i+1}$ for any $i \in \{1, 2, \ldots, n-1\}$.

Simple and supper bubbles, and bubble chains consisting of simple and supper bubbles, were enumerated from the assembly graph generated by Verkko using BubbleGun[125]. Cyclic bubbles were identified using a breath-first search algorithm implemented in a custom script, using entrance and exit nodes of known bubble chains as candidate source and sink nodes, implemented in a custom script. For performance reasons, cyclic bubbles for which the minimum numbers of connecting edges between the source and sink nodes were greater than 20 were not included.

### Decomposition of the *k*-mer profile from sequencing data

GenomeScope[105,127] is a tool widely used for modelling the k-mer profile, which uses a mixture of four evenly spaced binomial probability density distribution functions, corresponding to k-mers with copy number of one to four, respectively, to model the empirical distribution of *k*-mer depths (Figure 57). The GenomeScope model does not model *k*-mers resulting from sequencing errors.

Here, considering the need to explicitly model sequencing errors, and the possibility to model k-mers with higher copy numbers given the high coverage depth of HV31 sequencing data, I generalised the GenomeScope model to include additional peaks for error k-mers and k-mers with copy numbers up to a user-defined limit *n* (Figure 57). Similar to the GenomeScope model, the generalised model was fitted to the empirical distribution of *k*-mer depths using a least-squares optimiser implemented in the Scipy library[175]. The likelihood of each copy number given each depth value was estimated based on the relative contribution of each peak for that depth value (Figure 38). The custom code used for implementing and fitting the generalised model is available from GitHub (https://github.com/jzhang-dev/kmer-profile-decomposer).


The GenomeScope model
fit only **four** peaks
*G*: the total number of *k*-mers
negative binomial distribution parametrised as *NB*(*x*, *μ*, size)
*iλ*: evenly spaced mean
*λ*: haploid depth
*f*(x): the number of *k*-mers with depth *x*
*i*: copy number
*x*: *k*-mer depth
*αᵢ*: proportion of *k*-mers with copy number *i*
size is proportional to mean scaled by dispersion bias *ρ*; variance = (1 + *ρ*) *μ*; ρ = 0 → Poisson distribution; larger ρ → stronger overdispersion
The generalised model used in this work
error *k*-mers have mean 0
fit ***n*** peaks
*α₀*: proportion of error *k*-mers
size factor of error *k*-mers

$$\boxed{f(x)} = G\sum_{i=1}^{4}\alpha_i NB\left(x, i\lambda, \frac{i\lambda}{\rho}\right)$$

$$f(x) = G\alpha_0 NB(x,0,s_e) + G\sum_{i=1}^{n}\alpha_i NB(x, i\lambda, \frac{i\lambda}{\rho})$$

**Figure 57. The generalised *k*-mer profile model.**

The *k*-mer profile model used by GenomeScope[105] and the generalised model used in this work.

Eight HV31 datasets were used for *k*-mer profile modelling, including 10X, Illumina, MGI-CoolMPS, MGI-standard, MGI-stLFR, HiFi-2019, HiFi-2022, HiFi-2023 (Table 1). Meryl[104] was used for *k*-mer counting. For complex bubble resolution (Figure 38 and Figure 39) and simple bubble phasing (Figure 42), 22-mers from homopolymer-compressed sequences were analysed. For assembly validation (Figure 45), 31-mers from unmodified sequences were analysed.

### Resolving complex bubbles using *k*-mer coverage depths

For each complex bubble, all paths connecting the source node to the sink node were enumerated using a breadth-first search algorithm implemented in a custom script. For performance reasons, paths longer than 25, as measured by the number of path nodes, were not included. For each pair $j$ of identified paths with $n$ nodes in total, a score $S$ was calculated as $S_j = \sum_{i=1}^{n} \log_{10} P(c_i|d_i)$, where $P(c_i|d_i)$ was the likelihood of the copy number of the a node $i$ in the path pair $j$, given the median depth $d_i$of the private *k*-mers of node $i$, as estimated from the fitted *k*-mer profile model (Figure 38). For the complex bubble, the path pair with the highest score was selected and randomly assigned to the two haplotypes.

### Phasing simple bubbles using MGI stLFR data

Heterozygous *k*-mers, i.e., k-mers with inferred diploid copy numbers of one, were identified using the fitted *k*-mer profile model (Figure 38). For a given node, *k*-mers that appeared only once in that node and not found in any other nodes were defined as the unique private *k*-mers of that node. For each simple bubble, unique private *k*-mers of each of the two internal nodes that were also identified as heterozygous *k*-mers were used as heterozygosity markers. For each set of stLFR reads sharing the same barcode, the numbers of heterozygosity markers in the reads for each internal node was calculated. A barcode was assigned to an internal node of a simple bubble if the barcode had at least ten more markers for the node compared to the other node in that simple bubble. Otherwise, the barcode was treated as being not informative for that bubble. The resulting barcode assignment data, represented as an allele matrix, a sparse matrix implemented using the Scipy library[175], with each row representing a barcode, each column representing a bubble, and each value representing an optional assignment to one of the two internal nodes, was solved as an MEC problem using HapCUT2[135] (Figure 42). stLFR reads whose barcodes were not on the list of known barcodes were excluded. The custom code used for solving a general allele matrix using HapCUT2 is available from GitHub (https://github.com/jzhang-dev/hapcut2-mec-solver). The haplotypes produced by HapCUT2 were used to phase simple bubbles relative to each other in building haplotype paths from each bubble chain (Figure 42).

## 1.4.7. Functional data analyses based on the HV31-V2 assembly

### Gene annotation of the HV31-V2 assembly

Gene annotation of the HV31-V2 assembly was performed for each haplotype separately using LiftOff[176] v1.6.3, based on GENCODE v38 annotations[177], using default parameters.

### RNA-Seq alignment

RNA-Seq data was aligned to each haplotype of the HV31-V2 assembly separately using STAR[178] v 2.7.10b, based on LiftOff annotations. First, reference indices were built for each haplotype using the following command:

```
STAR \
  --runThreadN 8 \
  --runMode genomeGenerate \
  --genomeDir OUTPUT_FOLDER \
  --genomeFastaFiles ASSEMBLY_FASTA \
  --sjdbGTFfile ANNOTATION_GFF3 \
  --sjdbGTFtagExonParentTranscript Parent \
  --sjdbOverhang 74 \
  --outTmpDir OUTPUT_TEMP_FOLDER
```

Next, alignment was performed using the following command:

```
STAR \
  --runThreadN 4 \
  --runMode alignReads \
  --genomeDir REFERENCE_INDEX_FOLDER \
  --readFilesIn RNASEQ_FASTQ \
    --readFilesCommand zcat \
    --genomeSAindexNbases 14 \
    --outSAMattributes All \
    --outSAMtype BAM SortedByCoordinate \
    --outFileNamePrefix OUTPUT_PREFIX \
    --outTmpDir OUTPUT_TEMP_FOLDER
```

### ATAC-Seq and ChIP-Seq alignment and peak calling

Duplicates in ATAC-Seq and ChIP-Seq data were removed using Picard Tools[179] v2.18.7, using default parameters. Deduplicated ATAC-Seq and ChIP-Seq data were aligned to each haplotype of the HV31-V2 assembly separately using Bowtie2[180] v2.5.1. First, Bowtie2 reference indices were built using default parameters. Next, the following command was used for alignment:

```
  bowtie2 \
    -k 4 --minins 38 --maxins 2000 \
    --seed 489534229 \
    -x REFERENCE_INDEX_PREFIX \
    -1 READ1_FASTQ -2 READ2_FASTQ \
    -S /dev/stdout \
  | samtools sort -m 6G -T TEMP_FOLDER -@ 4 -o OUTPUT_BAM
```

### Allele-specific expression

For each RNA-Seq, ATAC-Seq and ChIP-Seq read, the total edit distance between the aligned segments of the read and the reference haplotype of the HV31-V2 assembly, as recorded in the `NM` tag, was calculated. Secondary or unmapped segments, segments failing quality control, and segments there were not properly paired, were excluded. Each read was assigned to the haplotype with lower edit distance. In case of a draw, the read was assigned to both haplotypes. Coverage depths were calculated, for the overall alignment, and for haplotype specific reads, using bedtools[181] with the following command:

```
  bedtools genomecov \
    -ibam INPUT_BAM -bg -split \
  | LC_COLLATE=C sort -k1,1 -k2,2n > OUTPUT_BEDGRAPH
```

For ATAC-Seq and ChIP-Seq data, the `-split` parameter was omitted.

## 1.4.8. Long-read sequencing error profiling

### ONT base-calling

Guppy v6.4.2, a proprietary software developed by Oxford Nanopore Technologies, was used to perform basecalling from the FAST5 files generated in the 2019 sequencing run, using the associated configuration file `dna_r9.4.1_450bps_modbases_5mc_cg_sup_prom.cfg`, which specified basecalling in the super accuracy (SUP) mode with 5-methylated cytosine detection.

### Strand-specific read alignment

It has been reported that heuristic algorithms implemented in read alignment tools may introduce alignment bias by discriminating reads that originate from different strands of the genomic DNA[76]. To eliminate such potential bias and to simplify the analyses workflow, I aligned each read separately to the forward strand and to the reverse strand of the T2T reference genome, while restricting reads to be only aligned in the forward orientation, and assign each read to the strand it aligns better, as suggested previously[76]. Specifically, I first generated a FASTA file containing the reverse complement of each chromosome sequence in the T2T CHM13 v2.0 reference genome, and then used Minimap2 v2.24 to align reads (downsampled to ~5× coverage depth) separately to the original reference genome and to its reverse complement, using the following command:

```
minimap2 \
  -ax map-ont REFERENCE QUERY \
  -secondary=no --MD --eqx --cs=short --for-only \
  --sam-hit-only -Y -I 10G -t 8
```

For downstream analyses, I retained only the longest aligned segment for each read that satisfied the following criteria: (i) aligned segment length / read length ≥ 0.5; and (ii) phred-scale mapping quality ≥ 30. Reads that failed to align were excluded.

### Short-read SNP and indel calling

MGI-standard short reads were aligned to the T2T reference genome using BWA-MEM[182] with default parameters. Variant calling was performed using DeepVariant[150] with default parameters for the WGS model.

### Long-read structural variant calling

HiFi-2019 reads were aligned to the T2T reference genome using pbmm2[183] with the following command:

```
pbmm2 \
  align REFERENCE QUERY OUTPUT_BAM \
  --preset CCS --sort -j 8 -J 8 --log-level INFO
```

Structural variant calling was performed using pbsv[42] with default parameters for HiFi reads.

### Definition of high-confidence regions

The high-confidence regions for sequencing error analyses were defined as T2T genome regions that appeared in none of following four sets of masked regions: (i) microsatellites, segmental duplications, transposable elements and other types of repeats annotated in a previous report[184]; (ii) regions with one base pair distance from homozygous or heterozygous SNPs and indels identified from MGI-standard reads using DeepVariant as described above; (iii) regions with one base pair distance from homozygous or heterozygous SVs identified from HiFi-2019 reads using pbsv as described above.

### Sequencing error rate estimation

Sequencing errors were defined as sequences that aligned to and differed from the high-confidence regions of the T2T reference genome, including substitutions, insertions and deletions. The error rate of a given dataset was defined as the total number of error bases divided by the total number of sequenced bases from the high-confidence regions.

To implement the definition above, I inspected the CIGAR string of each read alignment, and counted the occurrence of each base that was substituted, inserted or deleted, as well as the length of the overlap between the alignment and the high-confidence regions, using a custom Python script based on pysam[185], a wrapper around htslib[186]. The results were summarised across the high-confidence regions of the reference genome.

### Homopolymer error analyses

Homopolymer length errors were analysed using an approach similar to counterr[187]. Gavin Band implemented a custom C++ program, `find-homopolymers`, which was applied to efficiently enumerate all homopolymer locations in the T2T reference genome. For each homopolymer in the T2T reference genome, I analysed the read sequences covering the homopolymer, and recorded the reference and read homopolymer lengths. Read homopolymer lengths were calculated as the corresponding reference length, plus bases inserted next to and consistent with the homopolymer, minus any deleted bases in the homopolymer. The results were summarised across the high-confidence regions of the reference genome.

Homopolymer error analyses were conducted by Gavin Band and me in close collaboration. Gavin Band implemented a custom C++ programme `find-homopolymers` and visualised the final output. I analysed the length of each homopolymer in the aligned reads based on `find-homopolymers` output, and generated a summary file for visualisation.

# Project 2. Effects of Low-dose IL-2 Immunotherapy in T and NK Cells

## 2.1. Introduction

### 2.1.1. Type 1 diabetes

Type 1 diabetes (T1D) is a chronic autoimmune disorder that results in the destruction of pancreatic beta cells, leading to a deficiency in the production of insulin, a key hormone responsible for regulating glucose metabolism. As a result, individuals with T1D suffer from hyperglycaemia, which can lead to a range of acute and chronic complications affecting various organs and systems in the body. T1D is typically diagnosed in childhood or adolescence, with higher incidence rates observed Europe[188]. T1D prevalence in children is estimated to be doubling approximately every 20 years in Europe[189]. In 2021, about 8.4 million individuals lived with T1D worldwide, a number that has been projected to increase to 13.5-17.4 million in 2040 (ref[190]).

Currently, there is no cure for T1D, and disease management requires lifelong insulin therapy and blood glucose monitoring, which only delays disease progression. Despite significant advances in diabetes care, individuals with T1D still face significant challenges in achieving optimal glycaemic control and maintaining a good quality of life[191], and suffer from the loss of 11-13 years of life expectancy[192]. Ongoing research efforts are focused on developing new therapies to treat this complex and debilitating disease, such as beta cell replacement therapy, which involves transplanting functional beta cells into individuals with T1D to restore insulin production[193]. In addition, immunotherapies that aim to modulate the immune response that leads to beta cell destruction in T1D have shown promise. For example, teplizumab, an anti-CD3 monoclonal antibody, has been shown to delay the onset of T1D in individuals at high risk of developing the disease[194]. Low-dose IL-2 immunotherapy, which constrains autoimmunity by boosting the number of regulatory T cells, is another promising approach for T1D treatment, as detailed below.

### 2.1.2. Regulatory T cells

Regulatory T cells (Tregs) are a specialized subset of T lymphocytes that plays a crucial role in maintaining immune tolerance and preventing excessive immune responses against self-antigens. Tregs are heterogeneous and can be broadly classified into two main subtypes: thymically-derived Tregs (tTregs) and peripherally induced Tregs (pTregs)[195]. tTregs develop in the thymus during T cell maturation, where they acquire self-tolerance and constitutively express the transcription factor FOXP3, which is the master controller for their suppressive function[196]. The majority of tTregs also express the transcription factor HELIOS[197]. On the other hand, pTregs are generated in the periphery from naïve conventional T cells (Tconvs) through induction by specific environmental factors, such as exposure to antigens in the presence of the anti-inflammatory cytokine TGF-β. pTregs only transiently express FOXP3 and are therefore less functionally stable compared to tTregs. Under certain conditions, pTregs may differentiate into effector cells that are pathogenic in autoimmune diseases[198].

Tregs can inhibit the activation and effector functions of T cells through various mechanisms[199,200], such as: (i) constitutive expression of CTLA-4, which outcompetes CD28 for the binding of CD80/CD86 on APCs, thus depriving the co-stimulation signal required for TCR activation; (ii) the production of inhibitory cytokines such as IL-10 and TGF-β; (iii) granzyme- and perforin-dependent killing of effector T cells; (iv) metabolic disruption mediated by IL-2-deprivation, cAMP, or adenosine receptor 2A; (v) inhibition of dendric cell (DC) maturation and function, causing the release of immunosuppressive molecule 2,3-dioxygenase.

Tregs play a crucial role in controlling the autoimmune attack on pancreatic beta cells in T1D, and Treg dysfunction has been suggested as a cause for T1D[201,202]. Therefore, Tregs have been established as an important therapeutic target for T1D. Adoptive transfer of Tregs has been shown to protect insulin production capacity in animal models[203] and newly diagnosed T1D patients[204]. *In vivo* induction of Tregs via low-dose IL-2 immunotherapy is another research focus for T1D treatment, as detailed below.

### 2.1.3. Low-dose IL-2 immunotherapy

Low-dose IL-2 immunotherapy is an emerging treatment approach for T1D that aims to restore immune tolerance and preserve beta cell function by boosting Treg numbers[205]. IL-2 is a 15-kDa cytokine with pleiotropic effects on the immune system and plays a critical role in the regulation of the immune response[206]. At higher doses, IL-2 promotes T and NK cell activation, expansion, and cytokine production, which is mediated through the low-affinity dimeric IL-2 receptor consisting of CD122 and CD132 expressed on effector-type lymphocytes[206], IL-2 promotes T and NK cell activation, expansion, and cytokine production, which is mediated through the low-affinity dimeric IL-2 receptor consisting of CD122 and CD132 expressed on effector-type lymphocytes[206]. High-dose IL-2 immunotherapy has been approved for the treatment of several types of cancer, including metastatic melanoma and renal cell carcinoma, by enhancing the immune response against cancer cells[207]. In contrast, low-dose IL-2 has been shown to selectively expand IL-2-dependent Tregs, which constitutively express the high-affinity trimeric IL-2 receptor consisting of CD25, CD122, and CD132, and play a critical role in maintaining immune tolerance. Low-dose IL-2 immunotherapy has shown safety and efficacy in several inflammatory and autoimmune conditions[208], such as graft-versus-host disease[209], systemic lupus erythematosus[210], and rheumatic diseases[211].

Genetic studies of T1D have also highlighted the roles of IL-2 and related genes in T1D. Multiple genes in the IL-2 pathway, including *IL2RA* (encoding CD25), *IL-2*, *IL-21*, *BACH2*, *PTPN2* and *IL-10* have been linked to T1D susceptibility in genome-wide association studies[212,213]. In particular, lower expression of the CD25 subunit of the IL-2 trimeric receptor was found to be correlated with increased risk of T1D[214].

Motivated by these findings, various clinical trials have been conducted to evaluate the potential of low-dose IL-2 immunotherapy in boosting Treg function and restoring immune tolerance in T1D patients. One clinical trial confirmed the safety of low-dose IL-2 immunotherapy in adult T1D patients [215]. The DILT1D trial[216] led by Prof John Todd's group found that Tregs were desensitised for at least 24 hours after IL-2 injection, while single doses over $0.38 \times 10^6$ IU/m$^2$ activated conventional T cells in addition to Tregs. In addition, sustained activation of memory effector T (Teff) cells was observed for doses over $1.0 \times 10^6$ IU/m$^2$. These findings signified the importance of appropriate IL-2 dosing regimen. Subsequently, the DILfrequency study[3] led by Prof John Todd's group confirmed an optimal three-day IL-2 dosing interval, with doses ranging from $0.20–0.47\times10^6$ IU/m$^2$. This part of my thesis analyses samples collected from the DILfrequency study, and focuses on understanding the effects of interval administration of low-dose IL-2 in T and NK cells using a single-cell multiomics approach, as detailed below.

### 2.1.4. Single-cell sequencing

Single-cell sequencing has revolutionised our understanding of cellular composition, heterogeneity and development, and has emerged as a crucial tool in various research areas such as cancer, neuroscience and immunology[217].

Unlike bulk sequencing approaches, which analyse a population of cells together and provide an averaged representation of their gene expression profiles, single-cell sequencing operates on isolated cells and allows the identification of cell-to-cell variability with unprecedented resolution.

In a typical single-cell sequencing experiment, a single-cell suspension is first prepared from the biological sample of interest, using various methods such as enzymatic digestion, mechanical dissociation, or fluorescence-activated cell sorting (FACS), while preserving cell integrity and viability. Individual cells are then partitioned into droplets containing molecular barcodes and lysis reagents using microfluidics devices. Next, library preparation is conducted according to the specific type and platform of sequencing. For example, in single-cell RNA sequencing using the 10x Genomics Chromium platform, mRNA molecules in each cell are captured and reverse-transcribed into complementary DNA (cDNA), which then undergoes fragmentation, end repair, adapter ligation, and PCR amplification processes[218]. Finally, the sequencing library is loaded onto a sequencer, which yields barcoded reads that can be traced back to individual cells using bioinformatics approaches[219,220].

In immunology research, single-cell sequencing has been instrumental in understanding immune responses, immune cell development, and immune-related diseases[221]. In particular, single-cell sequencing approaches have been extensively used in identifying and characterising immune cell types and subtypes[222,223], revealing heterogeneity within cell types[224], tracing immune cell development and lineages[225–227], and profiling the T and B cell receptor repertoires[228,229].

### 2.1.5. Research objectives

This project aims to explore in depth the effects of interval low-dose IL-2 immunotherapy, based on blood samples collected from participants of the

DILfrequency clinical trial[3]. I analysed single-cell multiomics data from isolated T and NK cells using a targeted approach based on the BD Rhapsody platform, which enabled simultaneous measurement of selected mRNA and surface protein markers at single-cell resolution. Specific research objectives include: (i) perform a detailed immunophenotypic characterisation of both T and NK cell compartments, (ii) confirm that low-dose IL-2 immunotherapy selectively expanded functional Treg subsets, without activating effector T cells, (iii) identify short-term and long-term changes in cellular composition and gene expression profiles induced by low-dose IL-2 immunotherapy, and (iv) provide guidance for future clinical studies T1D treatment using low-dose IL-2 immunotherapy.

## 2.2. Results

### 2.2.1. Targeted single-cell multiomics

The DILfrequency study was a nonrandomized, open-label, response-adaptive clinical trial  aimed at identifying the optimal IL-2 dosing regimen in patients with T1D[3]. During the study, dosing regimens with varying dose and frequency were tested on 38 adult participants with T1D, leading to the conclusion that a three-day IL-2 dosing interval with doses ranging from 0.20–0.47×$10^6$ IU/$m^2$ (Figure 58) was able to elicit and maintain the desired steady-state increase in Tregs without Tconv expansion[3].

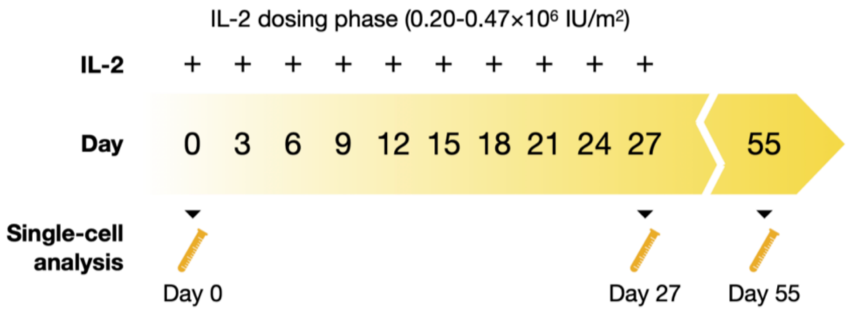


Figure 58. IL-2 dosing regimen and sampling timepoints.

Schematic representation of the optimal IL-2 dosing regimen and sampling timepoints used in the current study[3]. 13 participants were selected for single-cell analyses. For each participant, three longitudinal blood samples were used, which were taken on Day 0 (before the first IL-2 injection), Day 27 (before the last IL-2 injection), and Day 55.

To further investigate the effects of interval low-dose IL-2 (iLD-IL-2) immunotherapy, here I profiled T and NK cells from 13 selected DILfrequency participants treated with the optimal dosing regimen (0.20-0.47×$10^6$ IU/$m^2$ injection once every three days) using a targeted single-cell multiomics approach based on the BD Rhapsody platform, which enabled the parallel quantification of preselected mRNA and cell-surface proteins in each cell (Figure 59).

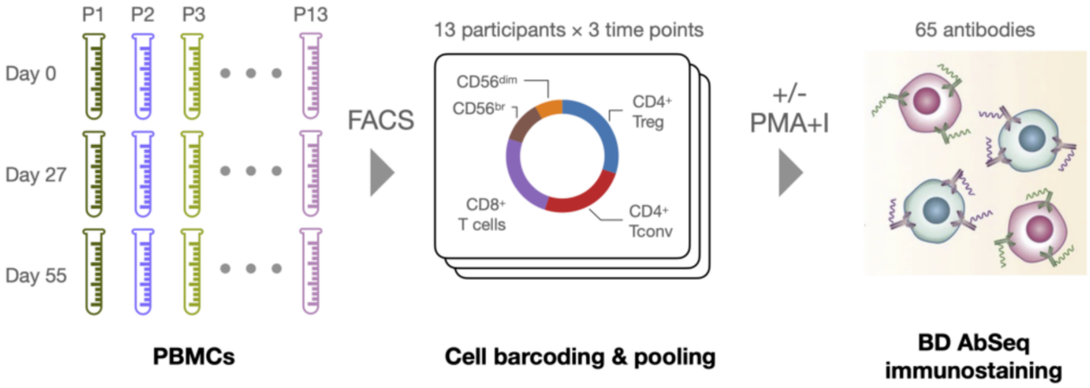


Figure 59. Single-cell sequencing experiment design.

Schematic representation of the single-cell sequencing experiments conducted to study the effects in low-dose IL-2 in T and NK cells. 39 PBMC samples the 13 selected participants taken at three timepoints were sorted using flow cytometry to isolate and enrich five major cell populations: CD4$^+$ Tregs (30%), CD4$^+$ Tconvs (25%), CD8$^+$ T cells (25%), CD56$^{br}$ NK cells (12%) and CD56$^{dim}$ NK cells (8%), where the numbers in the brackets represent the approximate proportion of each cell type included in the single-cell sequencing experiments.  Roughly half of the cells were stimulated *in vitro* using phorbol myristate acetate and ionomycin (PMA+I). Both stimulated and unstimulated cells were then tagged using BD AbSeq oligo-conjugated antibodies, and sequenced on the BD Rhapsody platform. All flow cytometry and sequencing experiments were conducted by Ricardo Ferreira.

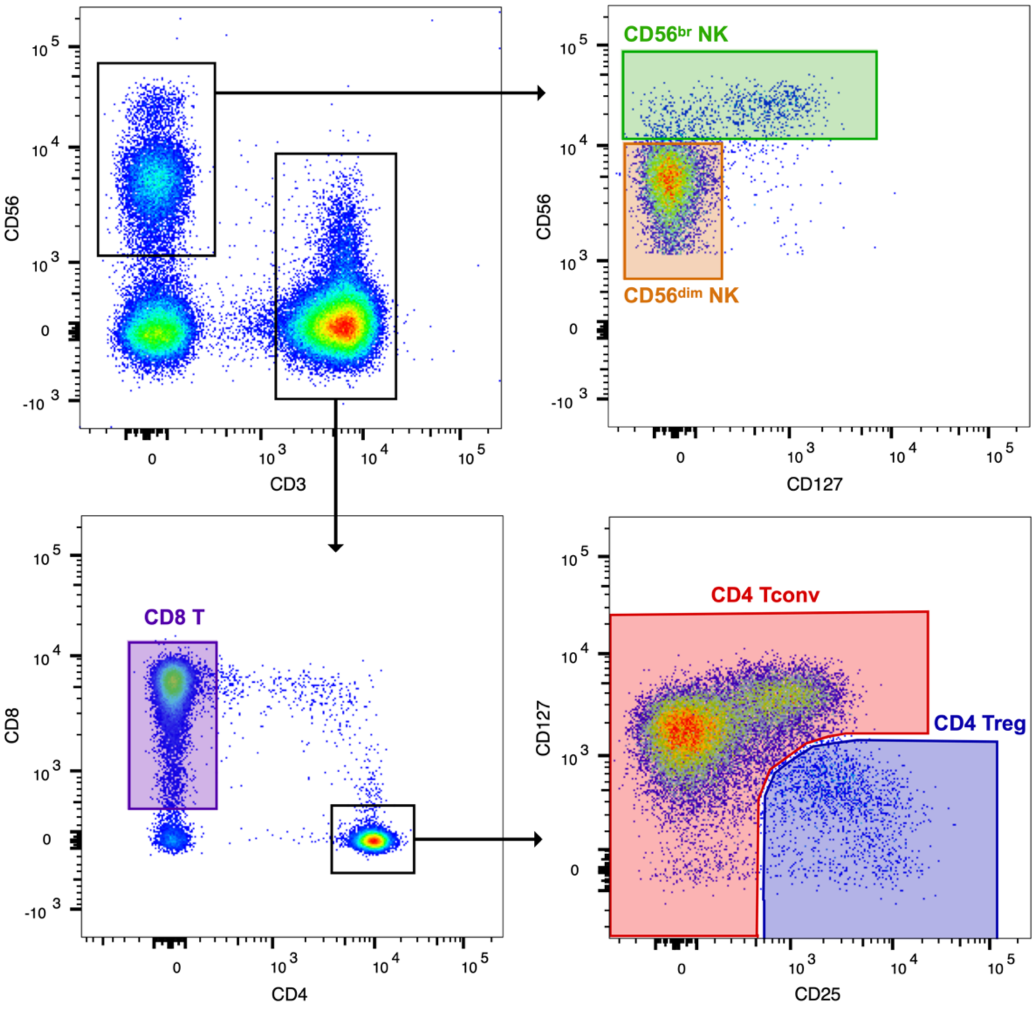


**Figure 60. FACS gating strategy for the delineation of T and NK populations.**

Representative illustration of the FACS gating strategy applied to isolate the five T and NK cell populations analysed using single-cell sequencing. All flow cytometry experiments were conducted by Ricardo Ferreira. Visualisation made by Ricardo Ferreira.

From each participant, three longitudinal blood samples were selected (Figure 58): Day 0 (baseline), Day 27 (immediately before the last IL-2 injection) and Day 55 (four weeks after the last IL-2 dose). From each cryopreserved PBMC sample, five immune cell populations were isolated using fluorescence-activated cell sorting FACS (Figure 60), including CD4$^+$ Treg (defined as CD3$^+$ CD4$^+$ CD127$^{low}$ CD25$^{hi}$), CD4$^+$ Tconv (defined as CD3$^+$ CD4$^+$ CD127$^{hi}$), CD8$^+$ T (defined as CD3$^+$ CD8$^+$), CD56$^{br}$ NK (defined as CD56$^{hi}$), and CD56$^{dim}$ NK (defined as CD56$^{low}$ CD127$^{low}$).

Among the five isolated populations, only CD4$^{+}$ Treg and CD56$^{br}$ NK were shown to expand significant after IL-2 treatment[3,5]. However, the two populations collectively account for less than 5% of PBMCs. Therefore, to maximize the statistical power to detect small changes elicited by IL-2 in CD4$^{+}$ Treg and CD56$^{br}$ NK, a cell enrichment strategy was adopted to increase the numbers cells from these two populations profiled in single-cell sequencing experiments (Figure 59). Specifically, the five isolated populations were labelled with oligo-conjugated sample multiplexing antibodies, which enabled the separation of sequenced cells by their original FACS gate, and mixed according to the following proportions: 30% CD4$^{+}$ Treg, 25% CD4$^{+}$ Tconv, 25% CD8$^{+}$ T, 12% CD56$^{br}$ NK, and 8% CD56$^{dim}$ NK (Figure 59). As a result of this enrichment strategy, the composition of each cell population in the sequencing data did not reflect the corresponding composition in PBMC samples. Therefore, most downstream analyses were conducted within specific cell populations.

The cell mixtures were split into two batches, one of which were sequenced directly, while the other were stimulated *in vitro* prior to sequencing. Single-cell sequencing was conducted using a custom multiomics panel previously developed by Ricardo Ferreira specifically for the investigation of T and NK cell populations[230], which consisted of 565 mRNA probes and 65 oligo-labelled antibodies that target cell-surface proteins (Figure 59). Compared to whole-transcriptome single-cell sequencing, this targeted approach based on the BD Rhapsody platform enabled deeper sequencing of key differentiation markers for T and NK cells at a lower cost, which was essential for mapping the heterogeneity within these immune cell subsets[222,230].

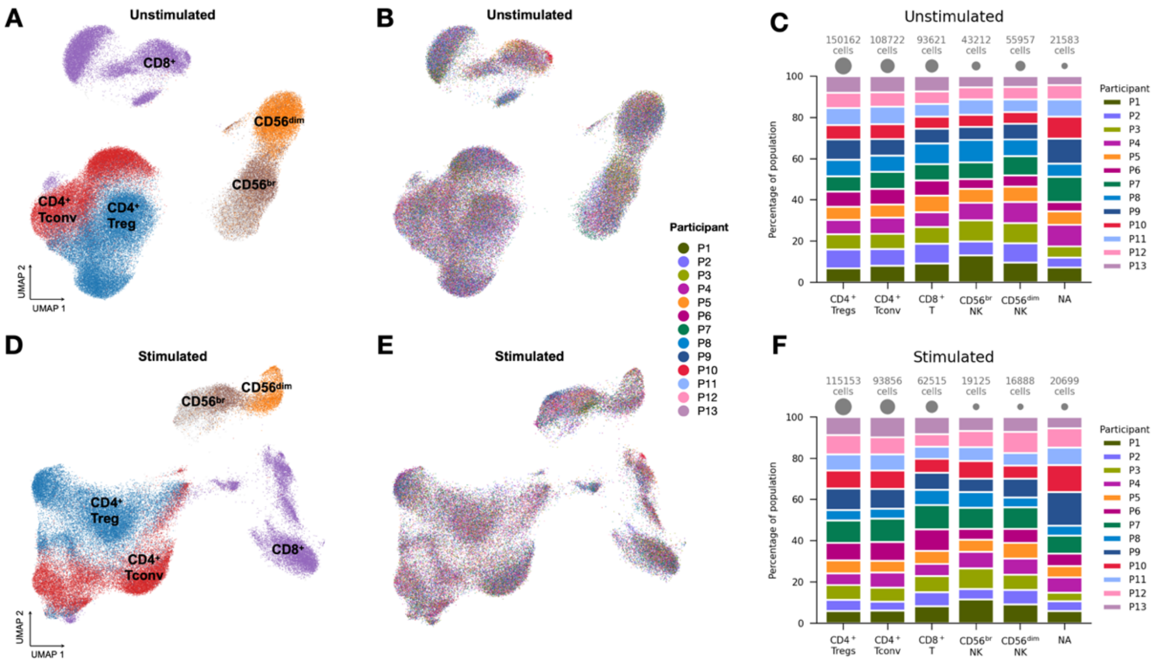


**Figure 61. Integration of the single-cell multiomics data.**

**(A, D)** Uniform Manifold Approximation and Projection (UMAP) embedding of unstimulated **(A)** and stimulated **(D)** cells in T and NK populations. Cells were coloured according to the FACS-isolated T and NK populations. Sparse grey dots in the CD56$^{br}$ clusters correspond to a small fraction of untagged CD56$^{br}$ NK cells, whose sample barcode information are missing due to suboptimal sample tagging efficiency.

**(B, E)** Same as **(A)** and **(D)**, respectively, with cells coloured according to the DILfrequency participants.

**(C, F)** Relative proportion of unstimulated **(C)** or stimulated **(F)** cells from each participant in each of the isolated populations. Areas of grey circles represent the number of cells sequenced in each population.

Batch effects represent an important technical challenge in single-cell sequencing[231]. To minimise the impact of batch effects in the identification of IL-2-induced changes, I applied a combination of canonical correlation analysis and identification of mutual nearest neighbours, implemented in Seurat[232,233], to correct for batch effects among samples from different participants (detailed on page 189). Following the batch-effect correction, I observed good overlap among cells from different participants, all of whom were well represented in the dataset (Figure 61).

### 2.2.2. Selective expansion of thymically derived Tregs during iLD-IL-2 treatment

Based on specific cell types identified using unsupervised clustering methods, I first investigated whether the iLD-IL-2 regimen applied here altered the relative composition of cells in the PBMC. Because of the cell subset enrichment strategy adopted in designing the single-cell sequencing experiments (Figure 59), the proportion of each major cell population did not reflect the corresponding physiological proportion in PBMC. Therefore, I conducted the differential abundance analyses separately for each of the five major cell types defined by the corresponding FACS sorting gates (detailed on page 191).

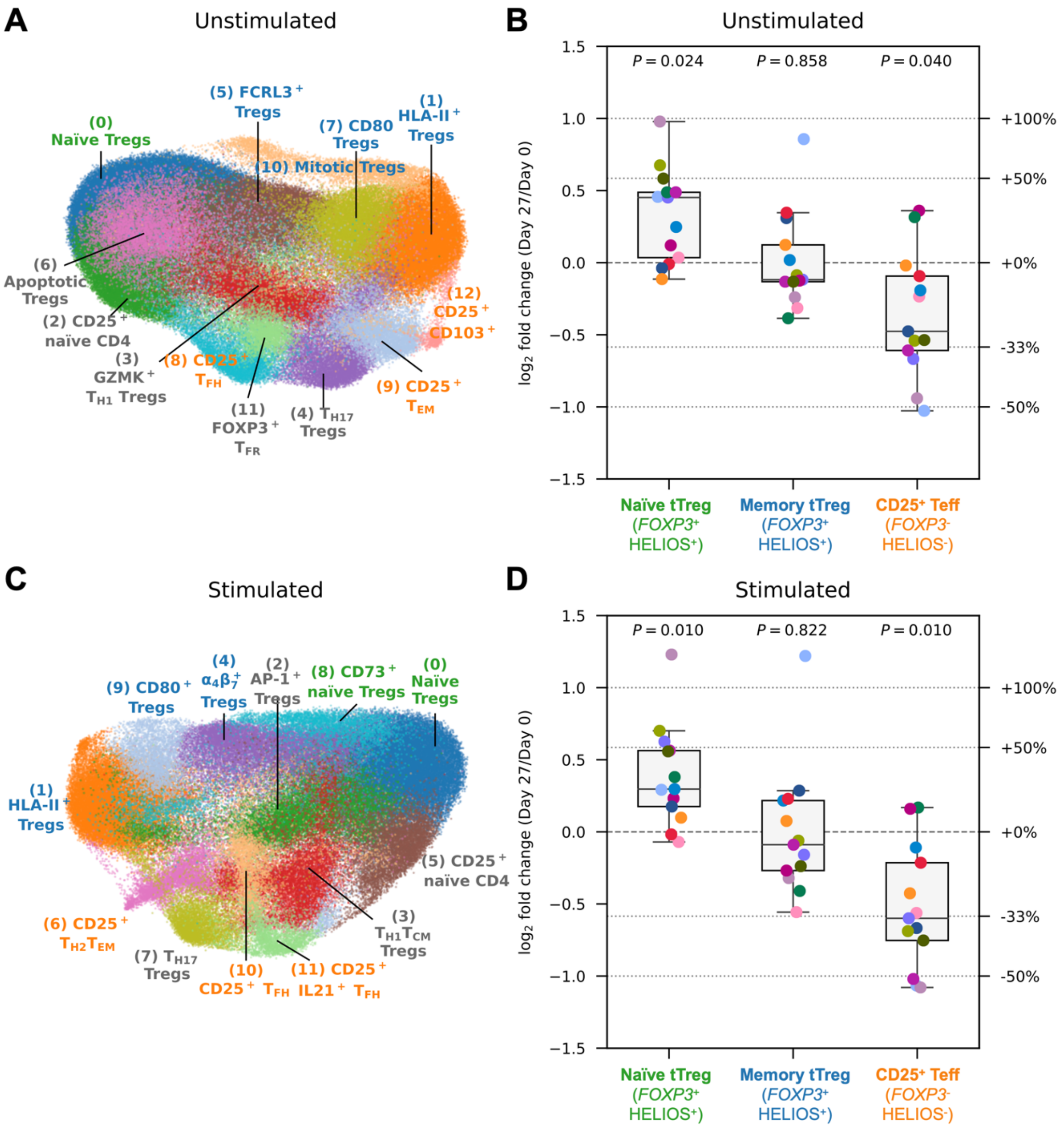


**Figure 62. iLD-IL-2 selectively expanded naïve tTreg subsets.**

**(A, C)** UMAP embedding of unstimulated **(A)** and stimulated **(C)** cells from the CD4$^{+}$ Treg population, coloured by identified clusters. Clusters were manually annotated based on the expression of key mRNA and protein markers. Clusters classified as naïve tTregs (FOXP3$^{+}$ HELIOS$^{+}$ CD45RA$^{+}$), memory tTregs (FOXP3$^{+}$ HELIOS$^{+}$ CD45RA$^{-}$), and CD25$^{+}$ Teffs (FOXP3$^{-}$ HELIOS$^{-}$ CD45RA$^{-}$) are annotated in green, blue and orange, respectively. Other miscellaneous clusters are annotated in black. Selected markers of each cluster are shown in Appendix Figure 4.

**(B, D)** Abundance changes of naïve tTreg, memory tTreg, and CD25$^+$ Teff subsets on Day 27 compared to Day 0, for unstimulated **(B)** and stimulated **(D)** cells. Each dot represents cells from a single participant. Dots with the same colour represent the same participant. Calculations of fold changes and P values are described on page 191.

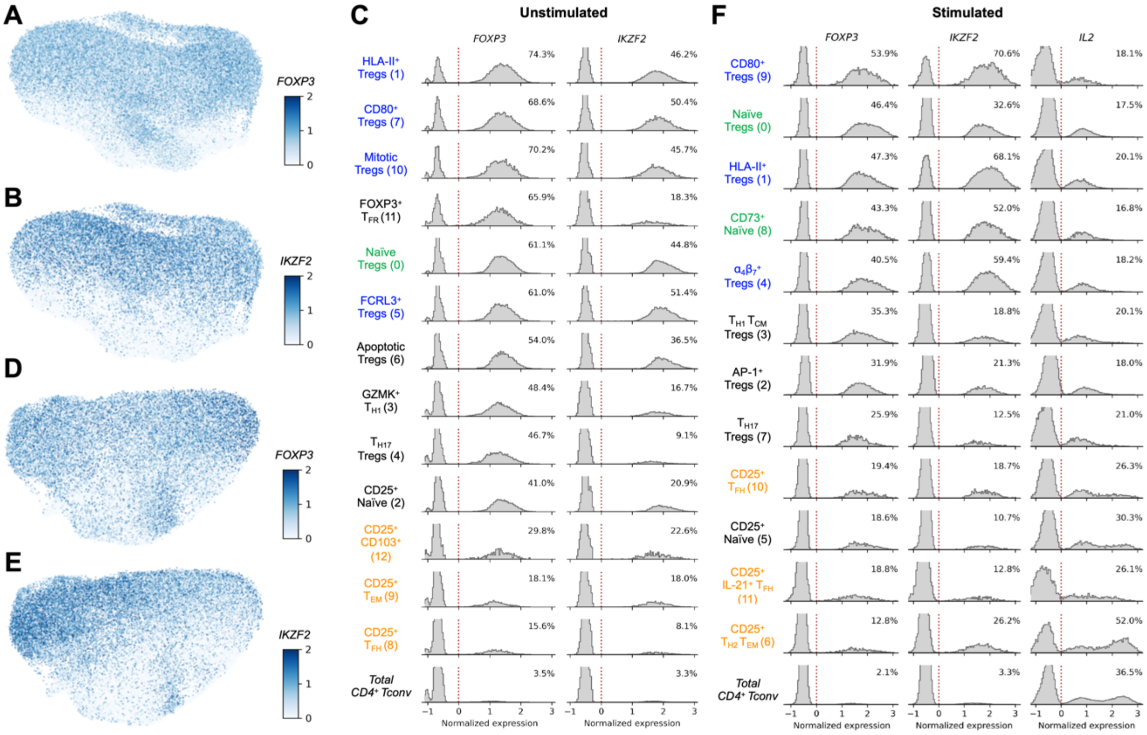


**Figure 63. Expression of *FOXP3*, *IKZF2* and *IL2* differentiates functional CD4$^+$ Treg and Tconv subsets.**

**(A-B)** UMAP embedding of unstimulated cells in the CD4+ Treg population coloured by the expression levels of *FOXP3* **(A)** and *IKZF2* **(B)**.

**(C)** Distribution of *FOXP3* and *IKZF2* expression levels in each cluster in the unstimulated CD4$^+$ Treg population.

**(D-E)** Same as **(A)** and **(B)**, respectively, for stimulated cells.

**(F)** Distribution of *FOXP3*, *IKZF2*, and IL2 expression levels in each cluster in the stimulated CD4$^+$ Treg population.

In **(C)** and **(F)**, clusters classified as naïve tTregs, memory tTregs, and CD25$^+$ Teffs are annotated in green, blue and orange, respectively. Other miscellaneous clusters are annotated in black.

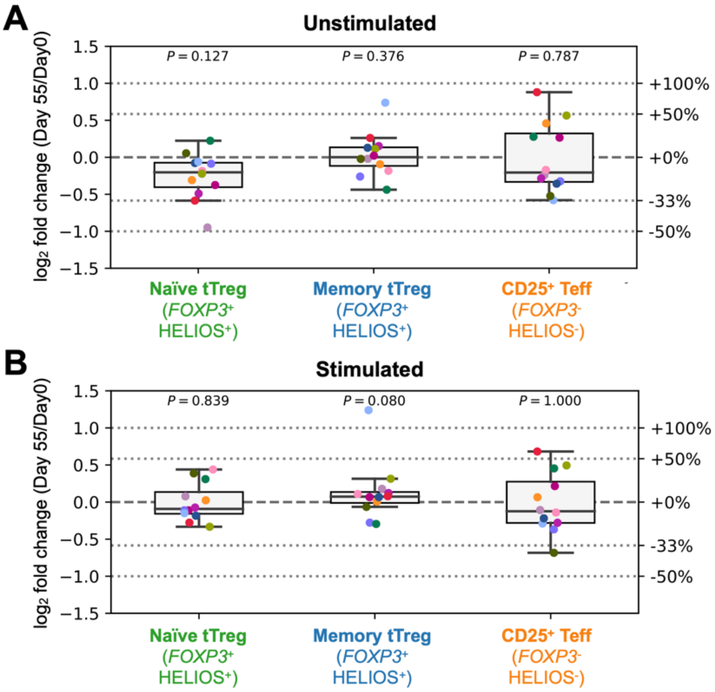


**Figure 64. Compositional changes in the CD4$^+$ Treg population on Day 55.**

**(A-B)** Abundance changes of naïve tTreg, memory tTreg, and CD25$^+$ Teff subsets on Day 55 compared to Day 0, for unstimulated **(A)** and stimulated **(B)** cells.

Expression of FOXP3 and HELIOS are considered canonical markers for stable Tregs[198,234], usually tTregs. However, the quantification of these two transcriptional factors required intracellular staining, which was incompatible with downstream single-cell sequencing. Therefore, the CD4$^+$ Treg population isolated using FACS was defined by the low expression of CD127 and high expression of CD25, as mentioned above (Figure 60). The Treg population defined this way was known to be heterogeneous, containing not only tTregs, which express FOXP3 and HELIOS, but also a fraction activated Teffs that express high levels of CD25 and low levels of CD127[198], which may be pathogenic in T1D patients.

13 clusters were identified and annotated in the CD4$^+$ Treg population using single-cell sequencing (Figure 62 and Appendix Figure 4). Analysis of the mRNA levels of *FOXP3* and *IKZF2* (encoding HELIOS) and *IL2* (encoding cytokine IL-2) in the CD4$^+$ Treg population allowed the stratification of naïve tTregs (*FOXP3*$^+$ HELIOS$^+$ CD45RA$^+$), memory tTregs (*FOXP3*$^+$ HELIOS$^+$ CD45RA$^-$), and CD25$^+$ Teffs (*FOXP3*$^-$ HELIOS$^-$ CD45RA$^-$), among other subsets (Figure 62 and Figure 63). Differential abundance analysis revealed an increase in the abundance of naïve tTregs and a concomitant reduction of CD25$^+$ Teffs on Day 27, after the conclusion of the four-week IL-2 dosing phase, which was replicated in stimulated cells (Figure 62). This compositional change of the CD4$^+$ Treg population reverted to baseline on Day 55 (Figure 64). Abundance changes of specific CD4$^+$ Treg clusters can be found in Appendix Figure 5.

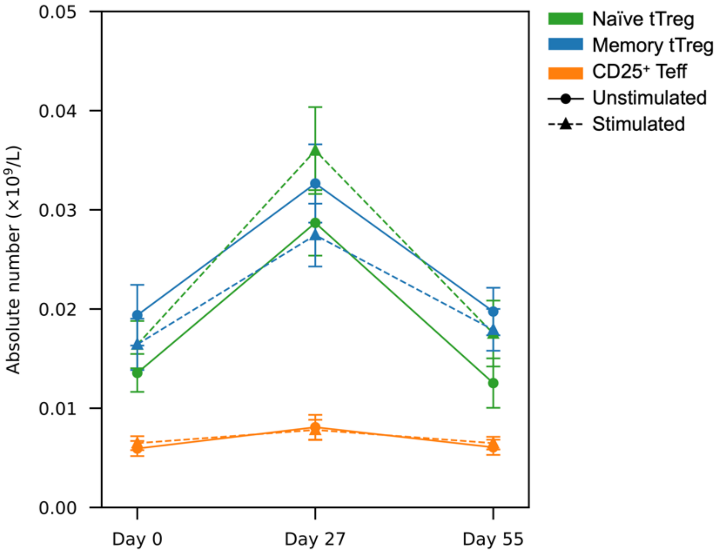


Figure 65. Estimated absolute Treg cell numbers in circulation

Absolute cell numbers of naïve tTregs, memory tTregs, and CD25$^+$ Teffs on Day 0, Day 27, and Day 55, estimated from unstimulated and stimulated cells in the CD4$^+$ Treg population. Absolute cell numbers of the Treg subsets were calculated by multiplying the absolute number of cells in the CD4$^+$ Treg population, which was measured by FACS and reported previously[3], and the relative abundance of the respective subset, which was measured by single-cell sequencing. Error bars denote the standard error of mean (SEM). Ricardo Ferreira contributed to this analysis.

Previous analyses of flow cytometry data from the same DILfrequency samples had shown a ~50% overall expansion of cell counts in the CD4$^+$ Treg population, which sustained during the IL-2 treatment and returned to the baseline on Day 55[3,5]. Combining this result with the compositional changes within the CD4$^+$ Treg population described above (Figure 62), I calculated the absolute numbers of each Treg subset, and found that the IL-2-induced expansion on Day 27 was restricted to the naïve and memory tTreg subsets (Figure 65). I observed very good concordance between unstimulated and stimulated conditions, with naïve tTregs showing a two-fold increase and memory tTregs displaying a ~70% increase in cell numbers compared to pre-treatment levels. In contrast, the number of CD25$^+$ Teffs within this population, despite their expression of the high affinity trimeric IL-2 receptor, showed no signs of increase (Figure 65).

### 2.2.3. Reduction of IL-21-producing T cells induced by IL-2

IL-21 signalling play a critical role in the development of T1D[235,236]. In particular, IL-21 production by T follicular helper ($T_{FH}$) cells is increased in T1D patients[237,238], and anti-interleukin-21 antibody, combined with liraglutide, is able to preserve β-cell function in recently diagnosed T1D patients[239]. In addition to cytokine IL-21, canonical markers for $T_{FH}$ cells include cell surface receptors CXCR5, ICOS, and PD-1, and transcription factors *MAF* and *BCL6*[240].

Within the unstimulated CD4$^+$ Treg and Tconv populations, two clusters were identified with transcriptional profiles resembling T$_{FH}$, including a cluster in the CD25$^+$ Teff subset, which expressed CXCR5, PD-1, ICOS, and MAF, and was annotated as CD25$^+$ T$_{FH}$ (Appendix Figure 4). Another cluster in the CD4$^+$ Tconv population that expressed CXCR5 and showed a profile consistent with central memory (T$_{CM}$) cells, a subset known to be enriched with circulating precursors of T$_{FH}$ cells[241], was annotated as *CXCR5*$^+$ T$_{CM}$ (Appendix Figure 6). After IL-2 treatment, CD25$^+$ T$_{FH}$ showed a slight decrease in abundance on Day 27, which returned to baseline on Day 55 (Appendix Figure 5), while *CXCR5*$^+$ T$_{CM}$ abundance displayed no significant changes.

*In vitro* stimulation was able to elicit the expression of cytokine IL-21 in T cells, allowing better discrimination of the T$_{FH}$ clusters. This led to the identification of four clusters with IL-21 expression and transcriptional profiles resembling T$_{FH}$ from stimulated CD4$^+$ Treg and Tconv populations, including two clusters in the CD4$^+$ Tconv population annotated as CXCR5$^+$ T$_{CM}$ and CXCR5$^{low}$ IL-21$^+$, and another two clusters in the CD4$^+$ Treg population annotated as CD25$^+$ T$_{FH}$ and CD25$^+$ IL-21$^+$ T$_{FH}$ (Appendix Figure 4, Appendix Figure 7, and Figure 66). After IL-2 treatment, CXCR5$^+$ T$_{CM}$ showed no abundance changes, while CXCR5$^{low}$ IL-21$^+$, CD25$^+$ T$_{FH}$, and CD25$^+$ IL-21$^+$ T$_{FH}$ displayed various degrees of decrease in abundance on Day 27, which returned to baseline on Day 55 (Appendix Figure 5 and Appendix Figure 7).

Meanwhile, a cluster in the unstimulated CD4$^+$ Treg population that with distinct Treg features, such as the expression of *POU2AF1* and *CCR9*, in addition to a T$_{FH}$ transcription profile, was annotated as *FOXP3*$^+$ T follicular regulatory (T$_{FR}$) cells (Appendix Figure 4). Unlike other T$_{FH}$ clusters, *FOXP3*$^+$ did not show a significant reduction in abundance on Day 27 (Appendix Figure 5).

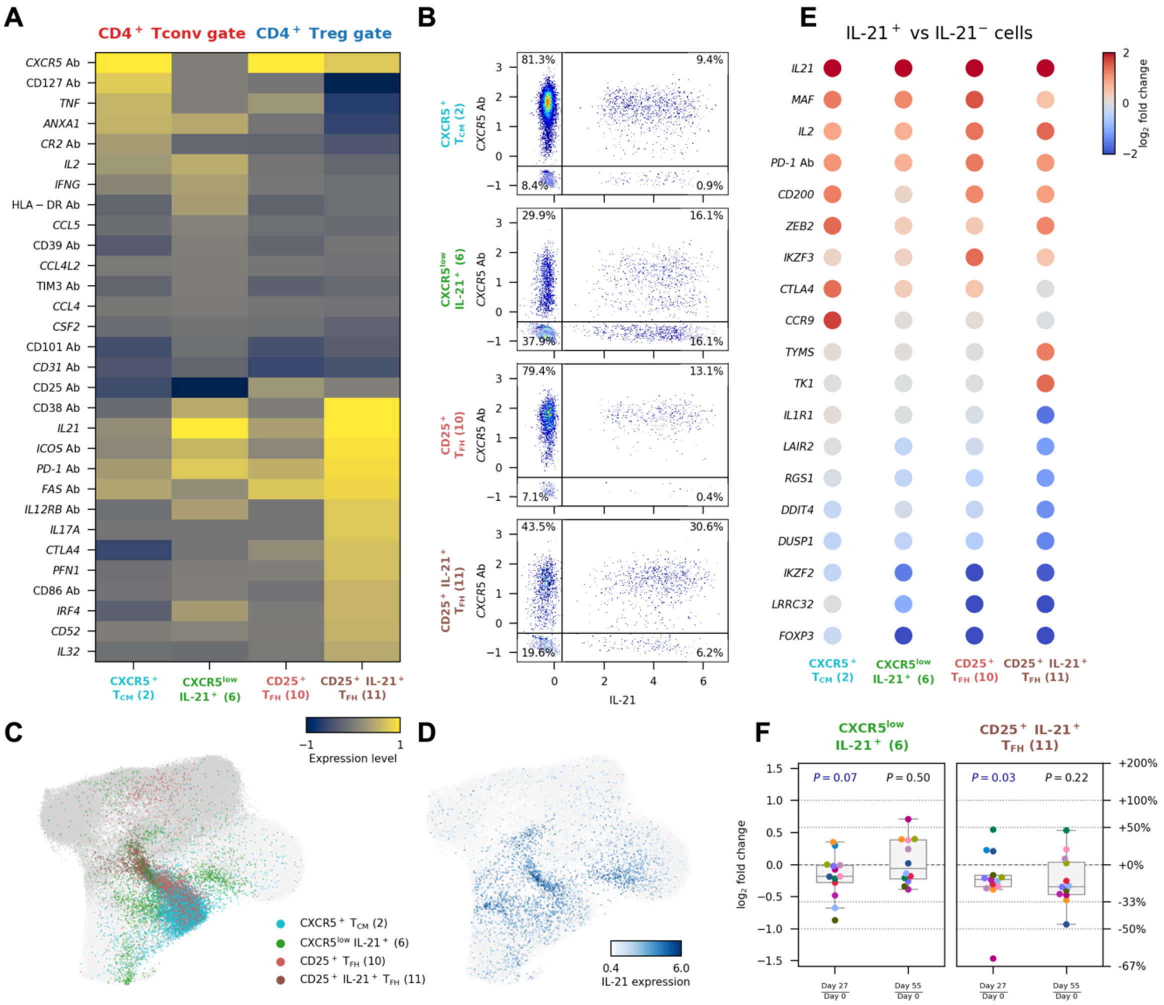


**Figure 66. IL-2 inhibited the differentiation of IL-21-producing CD4⁺ T cells**

**(A)** Expression levels of 30 differentially expressed mRNA and protein markers in four IL-21-producing subsets from stimulated cells in the CD4⁺ Treg and Tconv populations.

**(B)** Distribution of expression levels of CXCR5 protein and *IL21* mRNA in the four subsets described in **(A)**.

**(C)** UMAP embedding of stimulated cells in the CD4⁺ Treg and Tconv populations. The four subsets described in **(A)** are highlighted cyan, green, red, and brown. Other cells from the CD4⁺ Treg and Tconv populations were shown in dark and light grey, respectively.

**(D)** Same as **(C)**, with cells coloured by IL-21 expression levels.

**(E)** Gene expression changes between IL-21$^+$ and IL-21$^-$ cells within each of the four subsets described in **(A)**. Genes with absolute $\log_2$ fold change ≥ 1.2 and FDR-adjusted $P < 0.01$ in at least one subset are shown. Calculations of fold changes and *P* values are described on page 193.

**(F)** Abundance changes of IL-21$^+$ cells within the CXCR5$^{low}$ IL-21$^+$ and CD25$^+$ IL-21$^+$ T$_{FH}$ subsets, comparing Day 27 or Day 55 with Day 0.

Next, I compared the transcriptional profile of the four IL-21-producing T$_{FH}$-like clusters identified in the stimulated CD4$^+$ Treg and Tconv populations. I found a continuum of T$_{FH}$ differentiation among these clusters, with CD25$^+$ IL-21$^+$ T$_{FH}$ representing the terminal stage of T$_{FH}$ maturation in blood, as illustrated by their classical T$_{FH}$ profile and high IL-21 production (Figure 66). Analysis of the co-expression of CXCR5 and IL-21 suggested that IL-21 was predominantly expressed in the CXCR5$^+$ cells rather than CXCR5$^-$ cells, with the exception that, in the CXCR5$^{low}$ IL-21$^+$ cluster, 50% of IL-21$^+$ cells express low levels of CXCR5 (Figure 66B). These IL-21$^+$ cells likely corresponded to circulating T peripheral helper (T$_{PH}$) cells, a cell type that was defined as CXCR5$^-$ PD-1$^{hi}$ and previously shown to be increased in T1D patients[242]. In addition, though originating from different isolated populations, CXCR5$^+$ T$_{CM}$ in CD4$^+$ Tconv and CD25$^+$ T$_{FH}$ in CD4$^+$ Treg shared similar transcriptional profiles (Figure 66A), with the latter likely representing an earlier stage of T$_{FH}$ differentiation. In agreement with this shared identity and developmental stage, CXCR5$^+$ T$_{CM}$ and CD25$^+$ T$_{FH}$ clusters had largely overlapping UMAP embeddings among all stimulated CD4$^+$ T cells (Figure 66C). In contrast to the other three T$_{FH}$-like subsets, the CXCR5$^{low}$ IL-21$^+$ cluster in the CD4$^+$ Teff population represented a much more heterogeneous subset with scattered UMAP embedding and high IL-21 production (Figure 66C-D).

To better discriminate the effect of IL-2 on IL-21 production, I stratified cells in these four T$_{FH}$-like subsets according to their expression of IL-21 (Figure 66E). This revealed a reduction in the relative proportion of IL-21$^+$ cells on Day 27 in CXCR5$^{low}$

IL-21$^+$ and CD25$^+$ IL21$^+$ T$_{FH}$ subsets (Figure 66F), both of which displayed high levels of IL-2 production (Figure 66B).

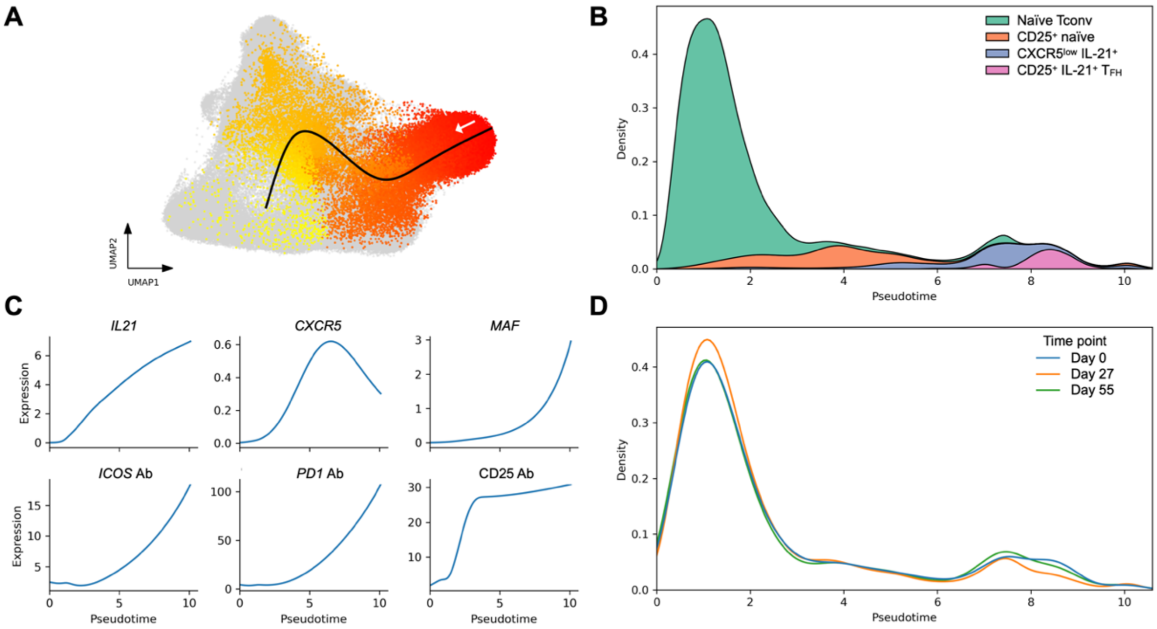


Figure 67. Pseudotime analysis of the T$_{FH}$ differentiation trajectory.

**(A)** UMAP embedding of stimulated cells in the CD4$^+$ Treg and Tconv populations, coloured by pseudotime. The black line denotes an identified pseudotime trajectory that was related to T$_{FH}$ differentiation. The white arrow denotes the predefined starting point of the pseudotime trajectory. Identification of pseudotime trajectories is described on page 192.

**(B)** Distribution of cells along the pseudotime trajectory described in **(A)**, stratified by cluster labels.

**(C)** Fitted expression curves of selected T$_{FH}$ marker genes along the pseudotime trajectory described in **(A)**.

**(D)** Same as **(B)**, with cells stratified by timepoints.

Finally, I used pseudotime analysis to confirm the differentiation trajectory of T$_{FH}$ cells. Pseudotime analysis is a computational approach for trajectory inference using single-cell sequencing data[243]. Conceptually speaking, pseudotime analysis works by ordering the cells linearly so that similar cells are close to each other. A cross-sectional sample of living humans, properly ordered, reveals the temporal trajectory of growth and aging. Analogously, a cross-sectional sample of single-cells often contains cells at various development stages that can be used to build pseudotime trajectories, which often resemble trajectories of cell differentiation[226,227,244].

Here, pseudotime analysis in stimulated CD4$^+$ Treg and Tconv cells revealed a trajectory of $T_{FH}$ cell differentiation characterised by the acquisition of a $T_{FH}$ phenotype and production of IL-21 (Figure 67A-C). The distribution of cells along this pseudotime differentiation trajectory confirmed that Day 27 samples had fewer differentiated $T_{FH}$ cells compared to Day 0 samples (Figure 67D), consistent with differential abundance results described above (Appendix Figure 5 and Appendix Figure 7). In addition, the $T_{FH}$ cells on Day 27 appeared to have a lower degree of differentiation compared to those on Day 0 (Figure 67D). Meanwhile, the pseudotime distribution on Day 55 was similar to that on Day 0. These results further supported a potential effect of iLD-IL-2 immunotherapy in inhibiting the differentiation of IL-21-producing cells, which had not been reported previously.

In addition IL-21-producing cells, I also found evidence for the reduction of another cytokine-producing subset in the stimulated CD4$^+$ Treg population annotated as CD25$^+$ $T_{H2}$ $T_{EM}$, which expressed various pro-inflammatory cytokines including GM-CSF, IL-21 and IL-2, IL-4 and IL-13 (Appendix Figure 4 and Appendix Figure 5B).

### 2.2.4. Effects of iLD-IL-2 on CD8$^+$ T and CD56$^+$ NK cells

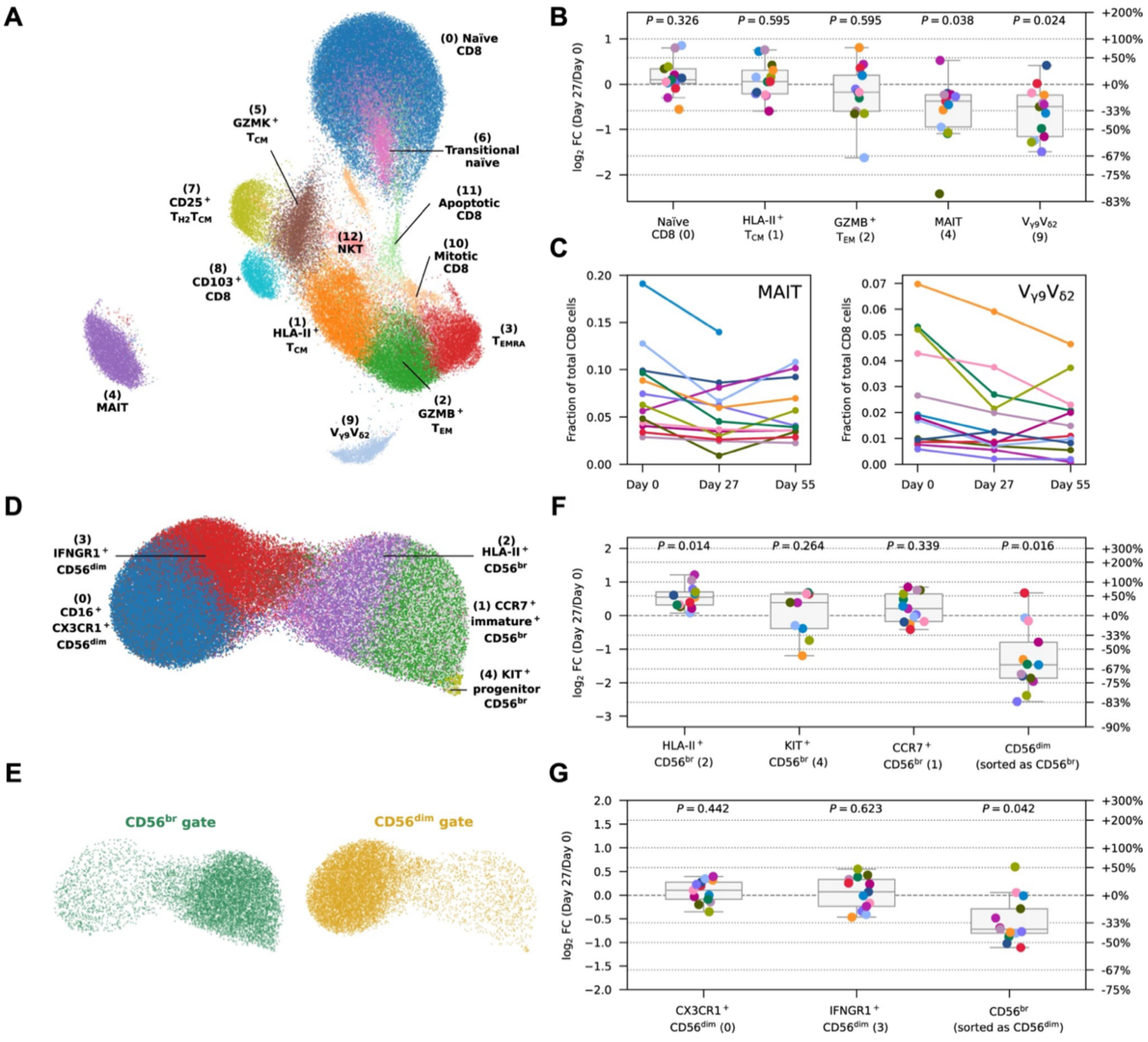


**Figure 68. IL-2 reduced MAIT and $V_{\gamma9}V_{\delta2}$ T cells and increased HLA-II$^+$ CD56$^{br}$ NK cells.**

**(A)** UMAP embedding of unstimulated cells in the CD8$^+$ T population, coloured by identified clusters. Selected markers of each cluster are shown in Appendix Figure 8A.

**(B)** Abundance changes of five selected clusters in **(A)**, comparing Day 27 with Day 0.

**(C)** Abundance of innate-like mucosal-associated invariant T cells (MAIT) and $V_{\gamma9}V_{\delta2}$ clusters within the CD8$^+$ T population on Day 0, Day 27, and Day 55. Each line represents cells from a participant. Lines with the same colour represent the same participant.

**(D)** UMAP embedding of unstimulated cells in the CD56$^{br}$ and CD56$^{dim}$ NK populations coloured by identified clusters. Selected markers of each cluster are shown in Appendix Figure 9A.

**(E)** Same as **(D)**, showing the CD56$^{br}$ and CD56$^{dim}$ NK populations separately.

**(F-G)** Abundance changes of five selected clusters in the CD56$^{br}$ **(F)** or CD56$^{dim}$ **(G)** NK populations in **(D)**, comparing Day 27 with Day 0.

In the CD8$^{+}$ T population, differential abundance analyses revealed the reduction of two clusters annotated as mucosal-associated invariant T cells (MAIT) and $V_{\gamma9}V_{\delta2}$, both of which were innate-like CD8$^{+}$ T cells (Figure 68A-B and Appendix Figure 8). Unlike compositional changes in the CD4$^{+}$ Treg and Tconv populations, the reduction of MAIT and $V_{\gamma9}V_{\delta2}$ cells sustained on Day 55, one month after the last dose of IL-2 (Figure 68C). Meanwhile, no discernible differences were found in the abundance of the other clusters, which represented conventional αβ CD8$^{+}$ T cell subsets.

CD56$^{br}$ and CD56$^{dim}$ NK populations had relative low heterogeneity compared to T cell populations (Figure 68D-E). Similar to CD4$^{+}$ Treg, the number of cells in the CD56$^{br}$ NK population increased significantly during IL-2 treatment in the DILfrequency participants, as previously reported[3,5]. Here, I found that this increase of CD56$^{br}$ NK was particularly driven by a subset annotated as HLA-II$^{+}$ CD56$^{br}$, which had an increased fraction on Day 27 (Figure 68F) but not Day 55 (Appendix Figure 9B-C), compared to Day 0. Meanwhile, I found no compositional changes in the CD56$^{dim}$ NK population (Figure 68G and Appendix Figure 9), except the reduction of a small fraction of contaminating cells in both CD56$^{br}$ and CD56$^{dim}$ NK populations on Day 27, including CD56$^{dim}$ sorted as CD56$^{br}$, and CD56$^{br}$ sorted as CD56$^{dim}$ (Figure 68F-G).

### 2.2.5. Reduction of cycling Treg and CD56$^{br}$ NK cells in blood

Cycling cells express a set of specific marker genes, such as *MCM4*, *TOP2A,* and *MKI67* (encoding Ki-67). Based on a proliferation score derived from 11 cell-cycle markers (detailed on page 194), I identified subsets of cycling cells in the unstimulated T and NK populations (Figure 69A). Cycling cells were not present in the stimulated populations, likely due to the high susceptibility of cycling cells to apoptosis following activation[245]. Comparing Day 27 with Day 0, I found a reduction in the fractions of proliferating cells within the CD56$^{br}$ NK and CD4$^{+}$ Treg populations (Figure 69B). This was inconsistent with previous reports that the fractions of Tregs and CD56$^{br}$ NK cells expressing Ki-67, a proliferation marker widely used in flow cytometry, increased after LD-IL-2 treatment[210,246,247]. A possible explanation of this discrepancy is that Ki-67 is not an optimal proliferation marker in this context. Ki-67 is a long-lived protein that is synthesised during the S and G2/M phases of the cell cycle, which remains detectable in the G1 phase after cell division[248]. Therefore, the increased levels of Ki-67$^{+}$ cells identified previously by FACS likely reflected protein accumulation in recently divided cells in the G1 phase, whereas the reduction of proliferating cells shown here (Figure 69) reflects a reduction of actively dividing cells in blood. This reduction was possibly due to an increased likelihood for cells to enter the cell cycle in tissues rather than blood, as IL-2 accumulates on the extracellular matrix during dosing[249–251].

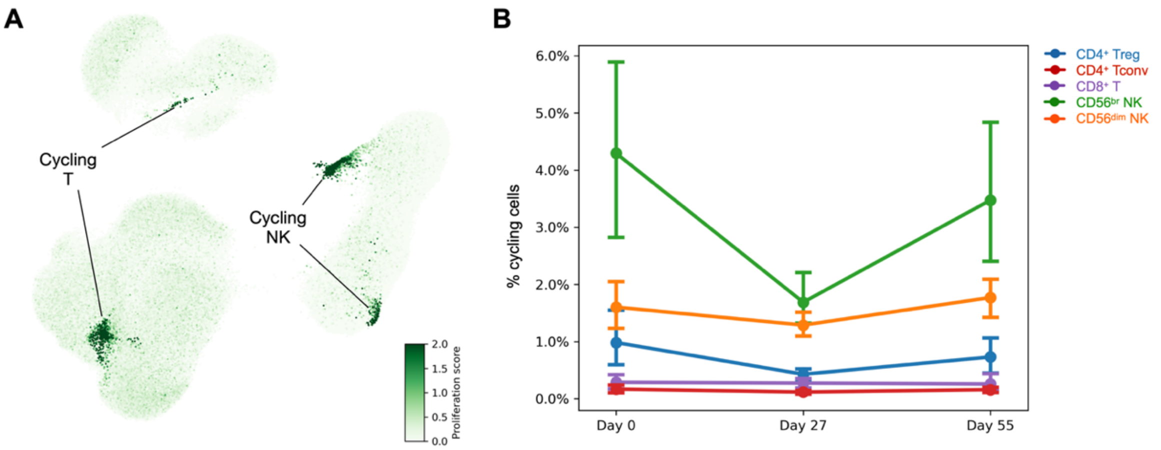


Figure 69. IL-2 reduced proliferating Treg and CD56$^{br}$ NK cells in blood.

**(A)** UMAP embedding of unstimulated cells in the five T and NK populations sequenced, coloured by a predefined proliferation score. Calculation of the proliferation score is described on page 194.

**(B)** Abundance of proliferating cells within each T or NK population sequenced on Day 0, Day 27, and Day 55. Proliferating cells were defined based on identified clusters marked by high proliferation scores. Dots denote mean values across 13 participants. Error bars denote 95% confidence intervals of mean values.

### 2.2.6. A long-lasting anti-inflammatory gene expression signature

In addition to differential expression analyses described above, I investigated whether IL-2, induced short-term or long-term changes in the gene expression profile of each T and NK population. On Day 27, I identified 40 genes that were differentially expressed in one or more populations of unstimulated and stimulated cells (Figure 70). Consistent with previous flow cytometry analyses[3], CD25 was strongly upregulated on Tregs on Day 27. The differentially expressed genes were mostly restricted to the CD4+ Treg and CD56br NK populations, which were known to be sensitive to IL-2 treatment[3,5]. These genes could be broadly categorised into CD4$^{+}$ Treg and CD56$^{br}$ NK signature genes, reflecting the increased abundance of naïve tTregs (Figure 64) after iLD-IL-2 treatment, as well as the decrease of contaminating CD56$^{dim}$ NK cells (Figure 68F). In addition, several genes related to cell cycle, including *TK1*, *TYMS*, *PCLAF* and *TOP2A*, were found downregulated in the CD56$^{br}$ NK population, and to a lesser degree in the CD4$^{+}$ Treg population (Figure 70), which was consistent with the decrease of cycling cells in the CD4$^{+}$ Treg and CD56$^{br}$ NK populations described above (Figure 69).

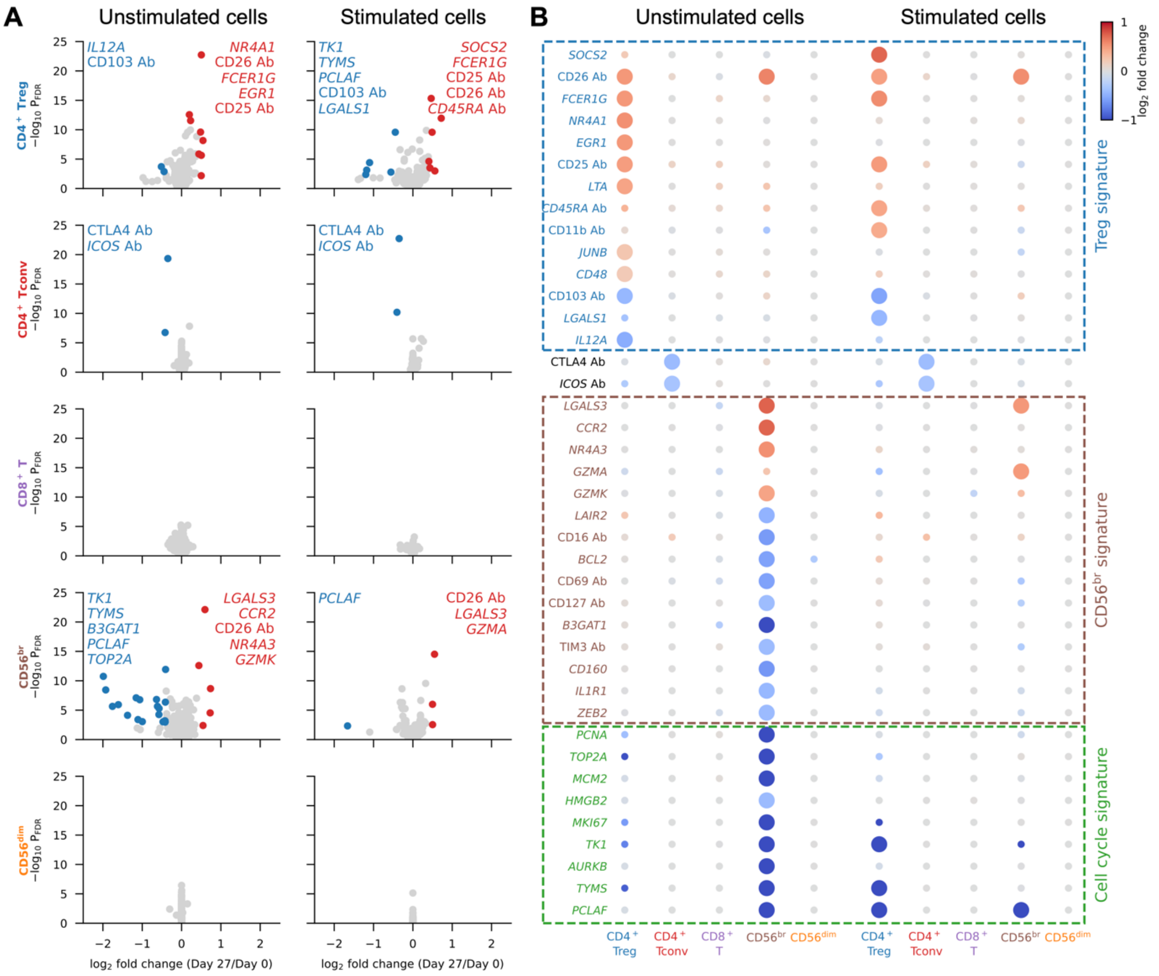


**Figure 70. iLD-IL2 selectively modulated gene expression of Tregs and CD56$^{br}$ NK cells on Day 27.**

**(A)** Volcano plots showing differential expression between Day 0 and Day 27 for the five T and NK populations sequenced. Significantly upregulated and downregulated genes are coloured in red and blue, respectively. Names of the top five upregulated and downregulated genes as defined by fold change values are labelled on each panel. Calculations of fold changes and *P* values and the definition of significantly differential expression are described on page 193.

**(B)** Differential expression of the 40 genes that are significantly differentially expressed in at least one population in **(A)**. Dot colours represent $\log_2$ fold changes between Day 0 and Day 27. Larger dots represent genes that are significantly differentially expressed in the respective cell population. Manually annotated gene subsets are labelled with dashed boxes.

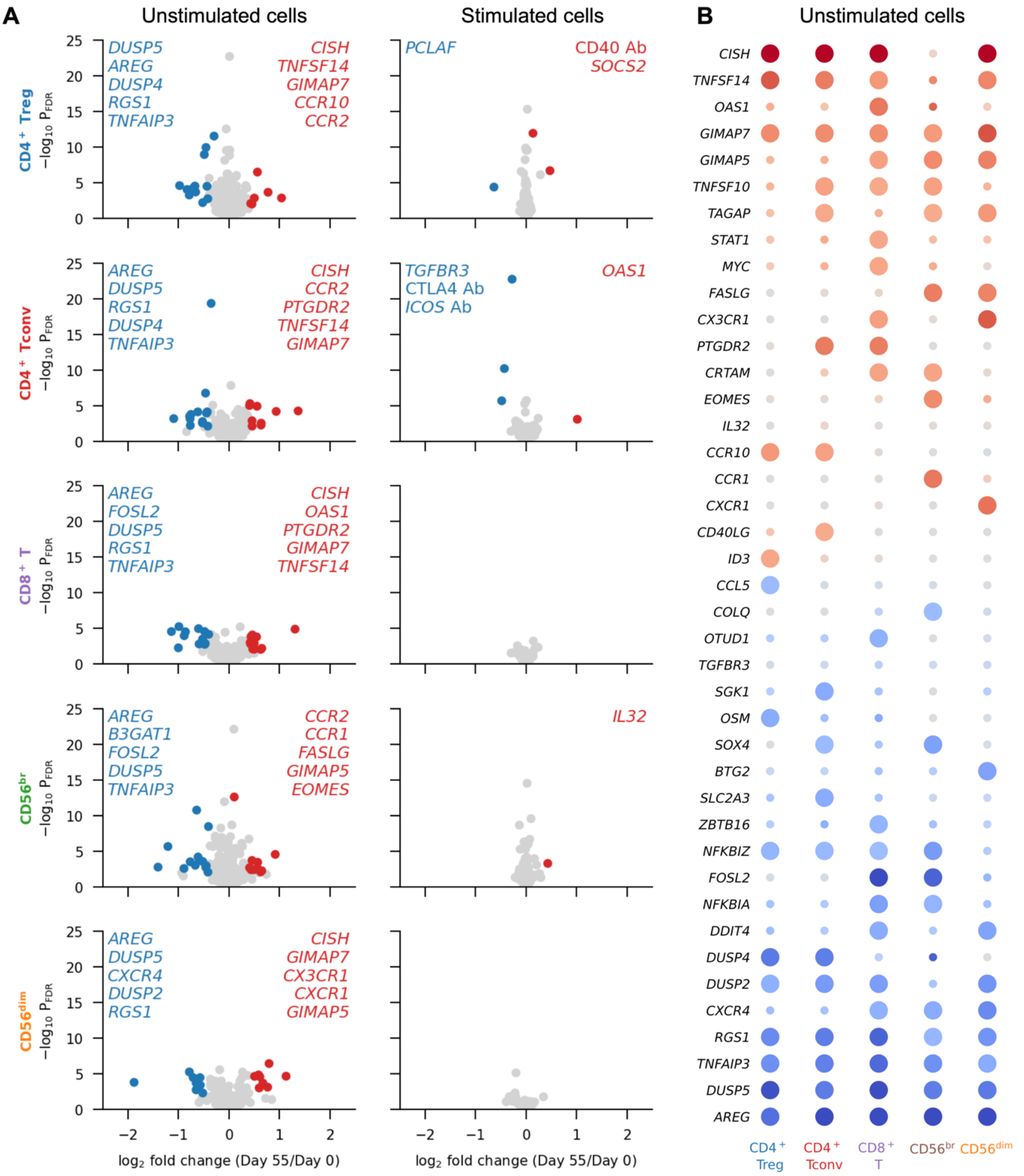


Figure 71. iLD-IL2 induced a long-lasting anti-inflammatory gene expression signature.

**(A)** Volcano plots showing differential expression between Day 0 and Day 55 for the five T and NK populations sequenced. Significantly upregulated and downregulated genes are coloured in red and blue, respectively. Names of the top five upregulated and downregulated genes as defined by fold change values are labelled on each panel.

**(B)** Differential expression of the 41 genes that are significantly differentially expressed in at least one population in **(A)**. Dot colours represent log2 fold changes between Day 0 and Day 55. Larger dots represent genes that are significantly differentially expressed in the respective cell population.

Comparing Day 55 with Day 0, I found a consistent differential expression pattern in unstimulated cells in all five T and NK populations (Figure 71). This shared gene expression signature induced by the iLD-IL-2 dosing regimen, which was sustained one month after the last IL-2 injection, featured most prominently the upregulation of Cytokine Inducible SH2 Containing Protein (*CISH*), a gene encoding a well-characterised negative regulator of cytokine signalling, and the downregulation of Amphiregulin (*AREG*), a secreted protein with pleiotropic roles in inflammation[252]. This Day 55 differential expression signature were enriched of genes associated with cytokine signalling, most notably in the TNF signalling pathway, which, in addition to *CISH* and *AREG*, also included *TNFSF14*, *TNFSF10*, *STAT1*, *SGK1*, *NFKBIZ*, *NFKBIA*, *DUSP2*, *DUSP4*, *DUSP5*, *RGS1* and *TNFAIP3* (Figure 71B). Furthermore, I observed a downregulation of several TNF-inducible genes that play a central role in curtailing TNF signalling, including the inhibitors of NFκB, *NFKBIZ* and *NFKBIA*, *RGS1*, *TNFAIP3* and *AREG*. Consistent with this observation, Oncostatin M (*OSM*), a pro-inflammatory cytokine shown to be increased in inflammatory bowel disease patients with poor response to anti-TNF therapy[253], was also downregulated after iLD-IL-2 treatment. Increased expression of *TNFSF14*, which encodes LIGHT, is notable as decreased expression is associated with higher susceptibility to multiple sclerosis (MS)[254]. In addition, in a MS mouse model, LIGHT expression in the brain limits disease severity[255].

Based on the function of involved genes, I refer to this gene expression signature as the IL-2-induced anti-inflammatory signature (IL2-AIS) hereafter, which includes 41 differentially expressed genes identified from Day 55 samples in one or more populations (Figure 71B).

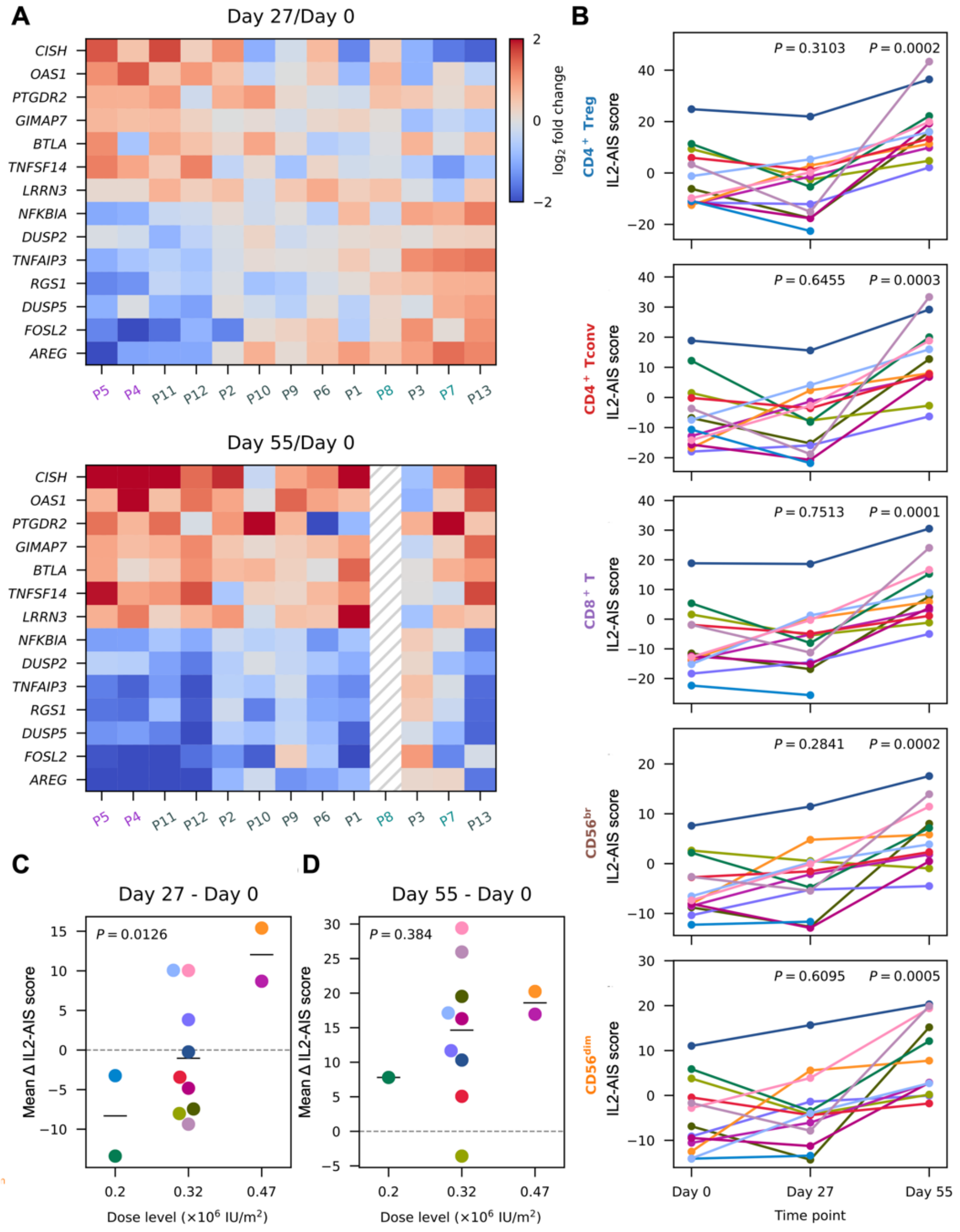


Figure 72. Induction of IL2-AIS on Day 27 is IL-2 dose-dependent.

**(A)** Participant-specific differential expression in CD8$^+$ T cells comparing Day 27 (top) or Day 55 (bottom) with Day 0. Top 5 upregulated and downregulated genes ranked by Day 55/Day 0 fold-changes in CD8$^+$ T cells are shown. Participant receiving IL-2 doses of 0.2 ×10$^6$ IU/m$^2$, 0.32 ×10$^6$ IU/m$^2$, and 0.47 ×10$^6$ IU/m$^2$ are labelled in cyan, grey, and purple, respectively. Day 55 data for participant P8 was excluded due to technical reasons detailed on page 189.

**(B)** IL2-AIS scores on Day 0, Day 27, and Day 55, stratified by participant (colours) and cell type (rows). Calculation of IL2-AIS scores is described on page 195. *P* values were calculated using a two-tailed paired *t* test comparing Day 27 or Day 55 with Day 0.

**(C-D)** Differences of the IL2-AIS scores between Day 27 **(C)** or Day 55 **(D)** and Day 0 averaged across the five T and NK populations stratified by IL-2 doses. Each dot represents a participant. *P* values were calculated using linear regression, with IL-2 doses treated as a continuous variable.

In contrast to the consistent observation of the IL2-AIS on Day 55 across different participants, the expression changes of IL2-AIS genes on Day 27 displayed considerable heterogeneity among participants (Figure 72A). Some participants exhibited concordant expression changes on Day 27 and Day 55, compared to Day 0, while others showed discordant expression profiles, with *CISH* being downregulated and *AREG* upregulated on Day 27, which likely explained why the IL2-AIS could not be identified by comparing Day 27 with Day 0.

Considering this correlation among different IL2-AIS genes, I derived a linear score to quantify IL2-AIS changes for each participant, referred to as the IL2-AIS score, with higher scores reflecting higher expression of signature genes upregulated in IL2-AIS, such as *CISH*, and lower expression of signature genes downregulated in IL2-AIS, such as *AREG*. As expected, the IL2-AIS score showed considerable inter-individual variation that was consistent across different cell types (Figure 72B). Correlation between the IL2-AIS scores and IL-2 doses suggested that participants receiving higher doses of IL-2 were more likely to have increased IL2-AIS scores on both Day 27 and Day 55 compared the baseline, while participants receiving lower doses tended to show such increases only on Day 55, but not Day 27 (Figure 72C-D). The mechanism of action of this dose-dependent induction of IL2-AIS on Day 27 remains unknown.

### 2.2.7. A pro-inflammatory gene expression signature related to IL2-AIS in COVID-19 patients

Since 2019, the COVID-19 pandemic has led to unprecedented collaborative efforts to elucidate the mechanisms of the immune response to SARS-CoV-2 infection. In particular, recent reports highlighted the differential expression of several key IL2-AIS genes in COVID-19 patients, including *CISH*, *AREG*, *DUSP2*, *NFKBIA* and *TNFAIP3*[229,256]. In light of this, I examined the dynamics of IL2-AIS in the context of the inflammatory responses in immune cell samples from two large-scale COVID-19 cohorts, the COMBAT cohort[229] and the INCOV cohort[256] (Figure 73).

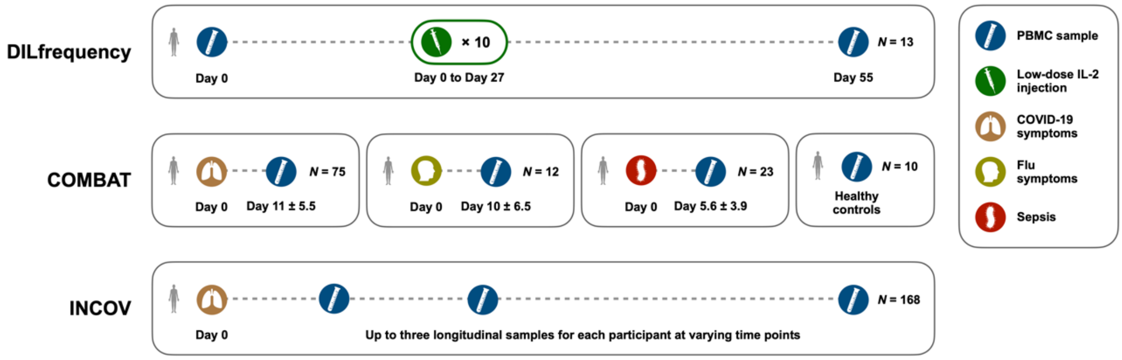


**Figure 73. Study design of the DILfrequency, COMBAT and INCOV cohorts.**

Schematic representation of the study design of the DILfrequency[3,5], COMBAT[229] and INCOV[256] cohorts. Only participants included in the current analyses are shown.

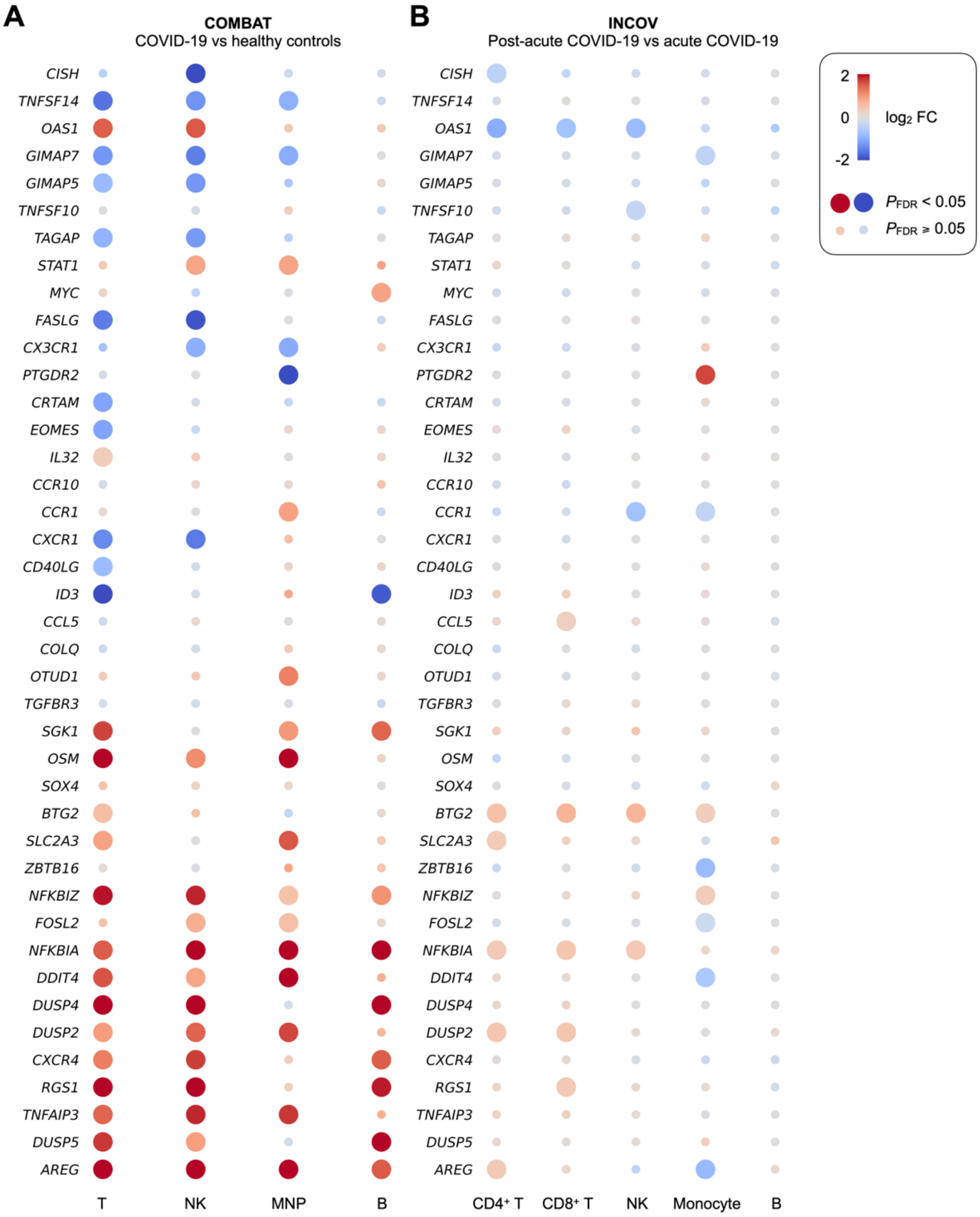


Figure 74. Differential expression of IL2-AIS genes in COVID-19 patients.

**(A-B)** Differential expression of the 41 IL2-AIS genes in the COMBAT **(A)** or INCOV **(B)** cohort. Differential expression was identified by comparing the expression between COVID-19 patients of varying disease severity with healthy controls in COMBAT **(A)**, and by comparing post-acute phase (29 to 84 days post symptoms) samples to acute phase (1-14 days post symptoms) samples of the same COVID-19 participants in INCOV **(B)**. Dot colours represent $\log_2$ fold change values. Larger dots represent genes with FDR-adjusted $P < 0.05$. Calculations of fold change and P values are described on page 193.

Differential expression analyses suggested that many IL2-AIS genes were regulated in the opposite direction in COVID-19 patients compared to healthy controls (Figure 71 and Figure 74A). For example, in direct contrast to the observed transcriptional changes induced by iLD-IL-2 in T1D patients, I observed the downregulation of *CISH* and the upregulation of *AREG* in COVID-19 patients during acute infection, across multiple immune cell types.

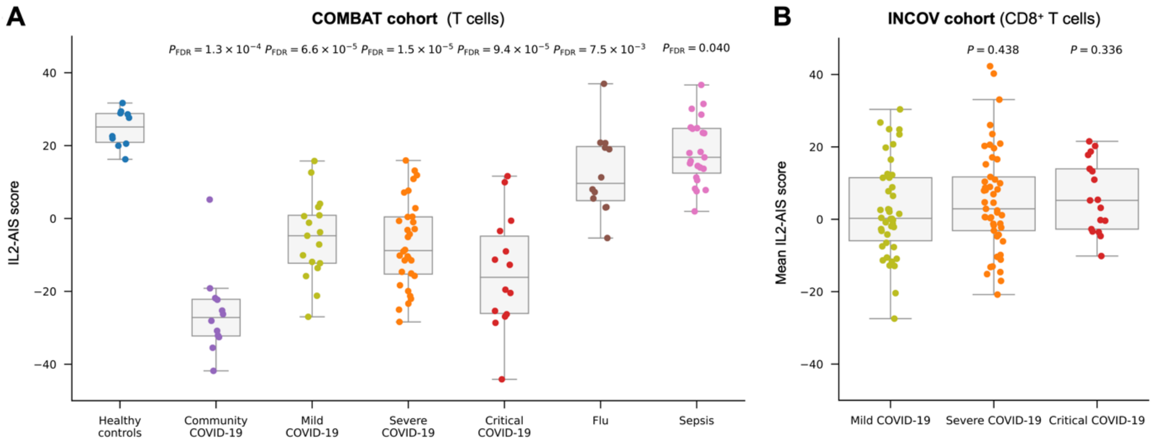


**Figure 75. IL2-AIS scores decreased in COVID-19 patients of varying severity.**

**(A)** IL2-AIS scores in T cells from each participant group in the COMBAT cohort. Data was stratified by disease group and COVID-19 disease severity. *P* values were calculated by comparing each patient group with healthy controls using two-sided Mann–Whitney *U* test followed by FDR adjustment based on the six comparisons.

**(B)** Mean IL2-AIS scores in CD8$^{+}$ T cells from each participant group in the INCOV cohort from samples collected during the first 14 days after symptoms onset. Data was stratified by COVID-19 disease severity. For participants with multiple longitudinal samples available, mean IL2-AIS scores across these samples are shown. *P* values are calculated by comparing the severe or critical COVID-19 group with the mild COVID-19 group using two-sided Mann–Whitney *U* test.

Evaluation of the IL2-AIS score in the COMBAT and INCOV cohorts revealed that this inverted modulation of the IL2-AIS was specifically observed in COVID-19 patients, but not in hospitalised patients in the COMBAT cohort who had other pro-inflammatory conditions, such as severe influenza or sepsis (Figure 75A). Furthermore, the COVID-19-specific decrease of the IL2-AIS scores was observed across all disease severity groups and did not appear to be correlated with COVID-19 severity (Figure 75).

### 2.2.8. Sustained transcriptional changes in COVID-19 patients

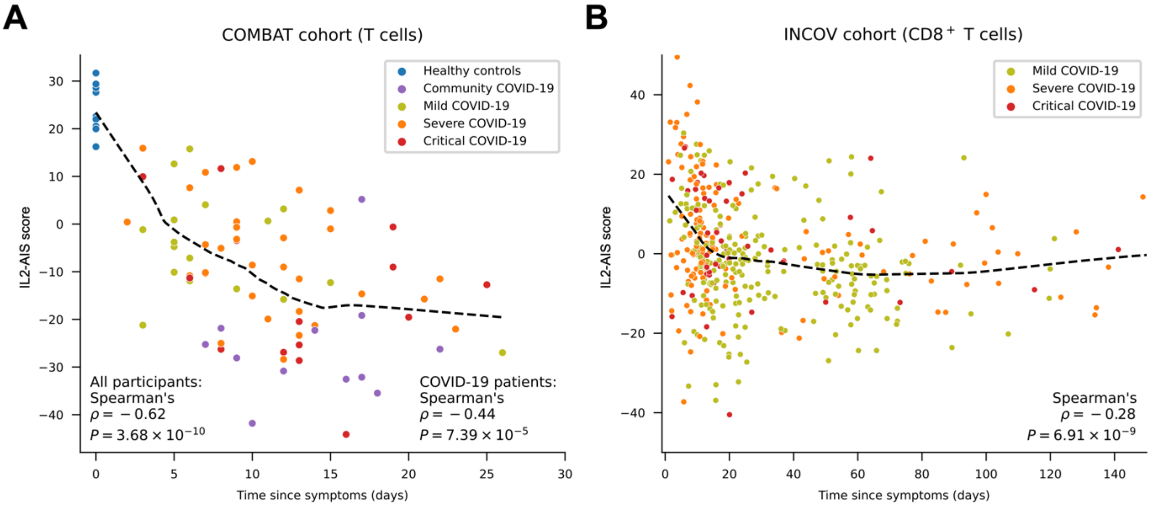


Figure 76. IL2-AIS scores progressively decreased after SARS-CoV-2 infection.

**(A-B)** Decrease of IL2-AIS scores after the onset of symptoms in COVID-19 patients from the COMBAT cohort **(A)** or INCOV **(B)** cohorts. Each dot represents a clinical sample, and colours depict the different COVID-19 disease severity groups. In the INCOV cohort, patients are grouped by their worst recorded COVID-19 severity. Dashed black lines represent locally weighted scatterplot smoothing (LOWESS) curves. The IL2-AIS score was calculated in T cells and CD8$^+$ T cells in the COMBAT and INCOV cohorts, respectively. *P* values were computed using Spearman's rank correlation. IL2-AIS scores for additional cell types are shown in Appendix Figure 10. IL2-AIS scores for specific cell subsets in the COMBAT cohort are shown in Appendix Figure 11.

Considering that the IL2-AIS was observed one month after the final IL-2 injection (Figure 72), and that the decrease of the IL2-AIS score was especially pronounced in a group of community COVID-19 patients in the COMBAT cohort (Figure 75) who were recruited from healthcare works and sampled relatively late (at least seven days post symptoms) compared to other patient groups[229], I speculated that IL2-AIS changes were progressively induced in both iLD-IL-2 and COVID-19 contexts. In support of this, analyses of IL2-AIS score dynamics in both COMBAT and INCOV cohorts revealed a progressive decrease during the first 2-3 weeks after SARS-CoV-2 infection (Figure 76). In addition, in the INCOV cohort, where longer-term follow-up samples were available, the decrease of IL2-AIS scores sustained for 2-3 months before showing slight signs of recovery towards the baseline (Figure 76B). The dynamics of the IL2-AIS transcriptional changes were largely consistent in all immune populations assessed in both COVID-19 cohorts, while the magnitude was largest in T and NK cells and smallest in B cells (Appendix Figure 10). Based on available cell-type annotations in the COMBAT dataset, I also investigated the dynamics of IL2-AIS in specific cell subsets, which confirmed that almost all evaluated cell subsets displayed IL2-AIS score changes (Appendix Figure 11) similar to that observed for the aggregated cell types (Figure 76A and Appendix Figure 10). This suggested that the IL2-AIS changes reflected differential expression within each cell subset, rather than relative compositional changes among different subsets.

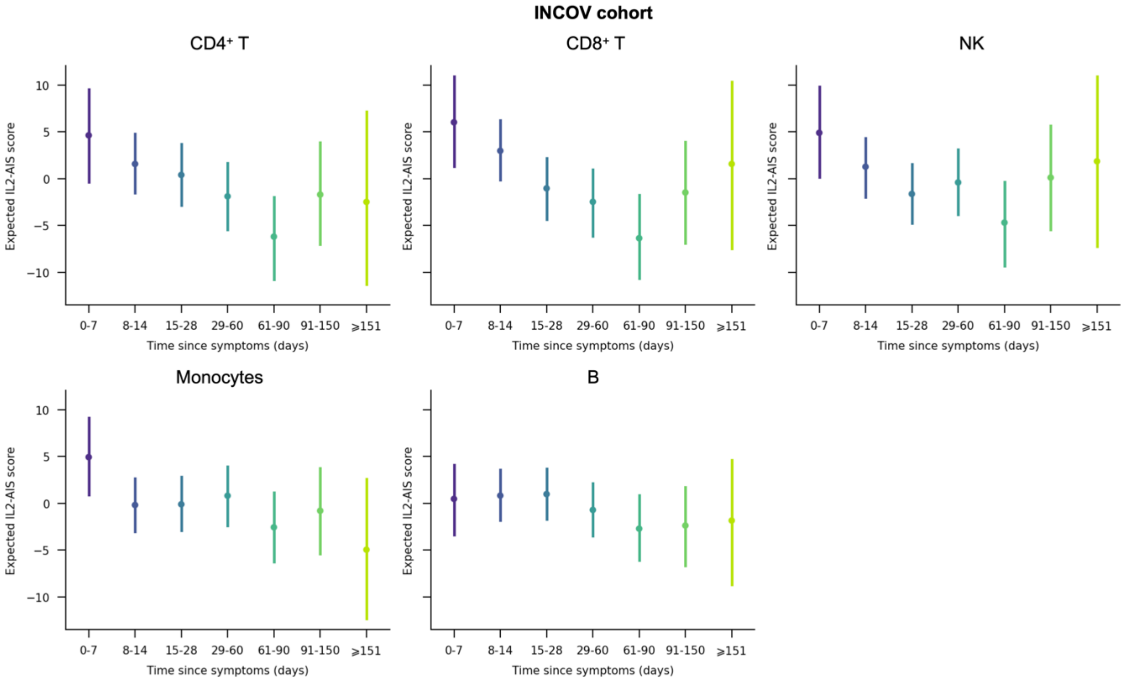


**Figure 77. Modelling the dynamics of IL2-AIS scores in the INCOV cohort.**

Posterior mean values (dots) and 95% confidence intervals (error bars) of the expected IL2-AIS scores (y axis) for each time range after COVID-19 symptoms onset (x axis). The expected IL2-AIS scores were estimated using a Bayesian linear model with regularising priors, detailed on page 195.

Considering that most participants in the INCOV cohort were sampled multiple times after symptom onset, I modelled the participant-specific effects and time-specific effects to the IL2-AIS scores in the INCOV cohort (detailed on page 195), which confirmed the progressive decline during the first 90 days post infection in CD4⁺ T, CD8⁺ T and NK cells, followed by a slow recovery (Figure 77). In monocytes and B cells, this pattern was less clear, though a general trend of decline was evident.

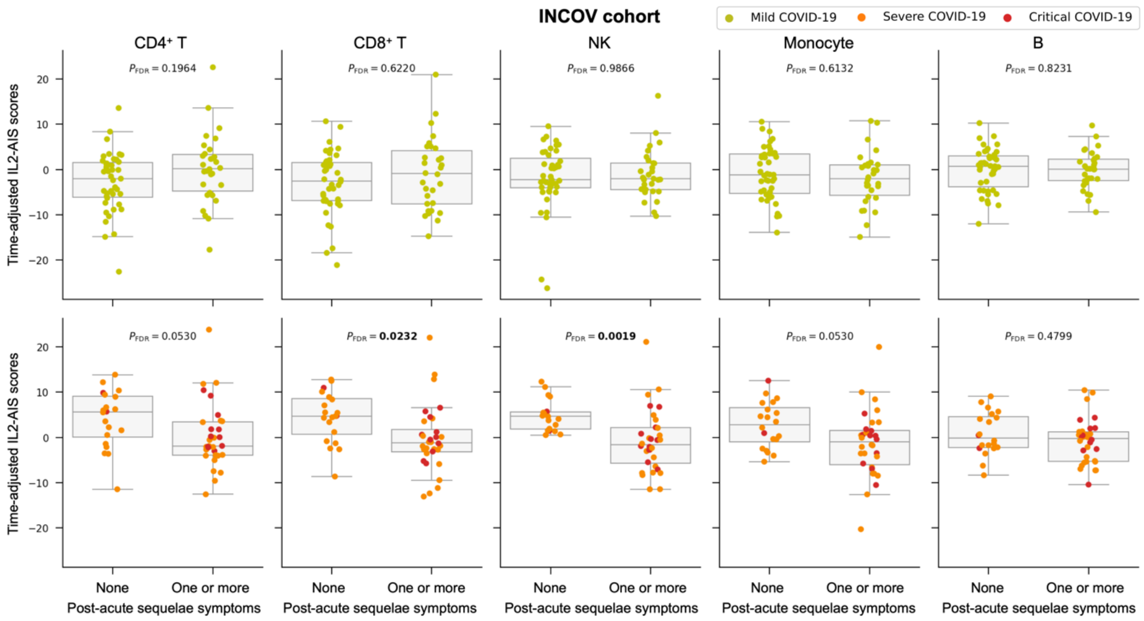


**Figure 78. Severe/critical patients reporting post-acute sequelae symptoms had lower IL2-AIS scores.**

Time-adjusted IL2-AIS scores for patients with or without post-acute sequelae of COVID-19 (PASC) symptoms, shown for the five cell types in the INCOV cohort. Each dot represents a patient. Patients are stratified by their maximum COVID-19 severity (colours), with mild patients shown on the top row, and severe/critical patients shown on the bottom row. The *P* values were calculated using two-sided Mann–Whitney U test and corrected for multiple testing using the Benjamini–Hochberg procedure.

The specificity of the IL2-AIS changes towards SARS-CoV-2 infection (Figure 75) and its persistence in samples taken months after the acute phase (Figure 76) implied the existence of potential long-term impairment of the immune response, which could underpin the manifestation of post-acute sequalae symptoms of COVID-19 (PASC). In agreement with this hypothesis, I found that severe and critical COVID-19 patients in the INCOV cohort who reported one or more PASC symptoms had lower IL2-AIS scores compared to those reporting no PASC symptoms (Figure 78). However, although this trend was consistently observed in all analysed immune subsets, the effect size was relatively small compared to intra-group variance, likely due to the heterogeneity of PASC symptoms[256].

In all three datasets analysed, I found the IL2-AIS score was not correlated with age or sex (Appendix Figure 12), although both factors were found to be associated with COVID-19 severity and PASC symptoms[257,258].

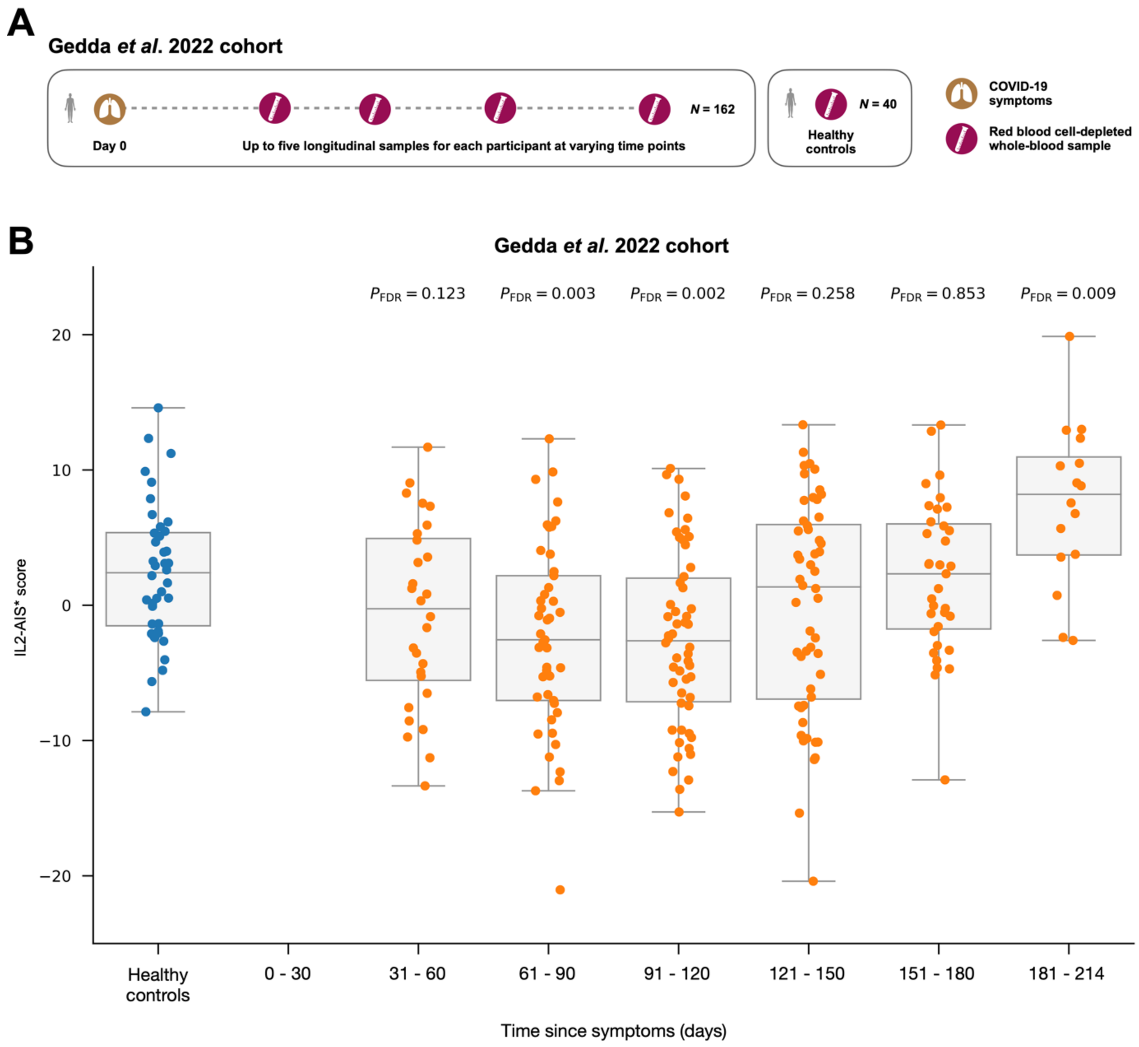


**Figure 79. Replicating a subset of IL2-AIS using the NanoString nCounter transcriptomics platform.**

**(A)** Overview of the Gedda *et al*. 2022 cohort[259].

**(B)** Distribution of the IL2-AIS* scores from 139 convalescent COVID-19 patients and 40 healthy control participants in the Gedda *et al*. 2022 cohort. The IL2-AIS* scores were calculated using a subset of 16 IL2-AIS constituent genes profiled using the NanoString nCounter Human Host Response panel: *MYC*, *CXCR1*, *OAS1*, *TNFSF10*, *FASLG*, *CCR10*, *STAT1*, *CX3CR1*, *CD40LG*, *IL32*, *EOMES*, *CCR1*, *OSM*, *SLC2A3*, *CXCR4*, and *CCL5*. Each dot represents a whole-blood sample from a participant visit. Each participant has up to five longitudinal samples. *P* values are calculated by comparing each group of COVID-19 patients with healthy controls using two-sided Mann–Whitney U test followed by FDR adjustment.

To obtain further evidence of these transcriptional changes, I assessed data from a recent study published by Gedda *et al*.[259], which used the NanoString nCounter platform to profile the transcriptional landscape of red blood cell-depleted whole-blood samples taken from 162 convalescent COVID-19 patients and 40 healthy controls (Figure 79A). The NanoString nCounter Human Host Response panel used in this study included 16 of the 41 IL2-AIS genes, allowing us to derive a signature score capturing a subset of the original IL2-AIS, referred to as the IL2-AIS* score. Consistent with previous observations, the IL2-AIS* score showed a progressive decrease during the first two months after infection, followed by a gradual recovery towards the baseline level measured in healthy control participants (Figure 79B). Despite the limited overlap between the IL2-AIS genes and the NanoString transcriptional panel, as well as differences in cell types, the consistency observed in the Gedda *et al*. cohort provides additional validation of the longevity of the transcriptional alterations induced by SARS-CoV-2 infection.

### 2.2.9. Network analysis of transcriptional changes in COVID-19 patients

As the IL2-AIS was initially identified in the DILfrequency study using a targeted panel of 565 transcripts designed for profiling T and NK cells, I sought to interpret the observed changes in the broader transcriptional landscape of immune cells in COVID-19 patients. In the COMBAT study, the authors employed a multi-parametric tensor decomposition analysis combining gene expression, surface protein expression (CITE-seq), plasma proteomics and cell subset abundance (flow cytometry) data, to identify 130 COVID-19-associated gene expression components categorized into 14 clusters, with each component containing a large number of weighted genes likely differentially expressed in COVID-19 patients[229]. I analysed whether IL2-AIS genes, compared to other genes assayed in the DILfrequency study, were enriched in specific components. I found three similar components in Cluster 3 (Component 211, Component 187 and Component 178), each showing strong enrichment (odds ratio > 8) of the 41 IL2-AIS genes (Figure 80A). From these three components in Cluster 3, Component 187 was highlighted by the authors as the most significant COVID-19-specific signature out of all the identified signatures[229]. Furthermore, I found that among the signatures associated with COVID-19 disease severity, Component 187 was unique in the longevity of the signature (Figure 81). In contrast, all other COVID-19-associated Components were only transiently modulated immediately after the onset of symptoms, indicating pathogenic alterations in cell composition and activation state during the acute phase of the disease, as illustrated for example by the classical monocyte-derived type 1 interferon (IFN) signature (Figure 81). Therefore, I focused on Component 187 for further analyses.

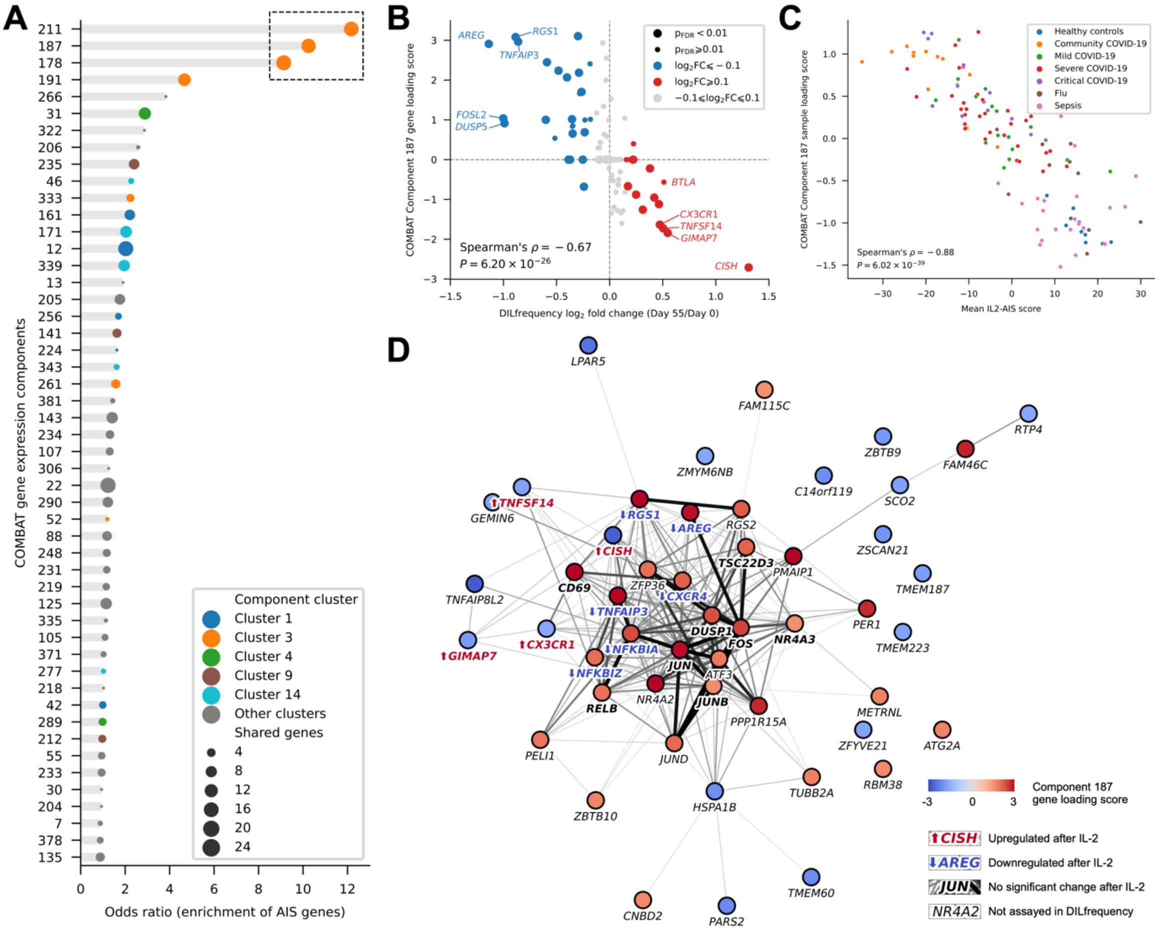


**Figure 80. NF-kB is central to the transcriptional alterations in COVID-19 patients.**

**(A)** Enrichment of IL2-AIS constituent genes in each gene expression component reported in the COMBAT study. Each dot represents a disease-associated gene expression component identified in the COMBAT study. Dot sizes represent the number of shared genes between the respective component and the 41 IL2-AIS constituent genes. The top 50 components are shown, ranked by their odds ratios of enrichment (x axis). Colours depict cluster membership as reported in the COMBAT study. Components in the same cluster are associated with diseases in a similar way. Cluster 3 is associated with all severity groups of COVID-19. The three components with odds ratio ⩾ 5 are highlighted with a dashed blacked box.

**(B)** Component 187 loading scores were negatively correlated with fold change values of differential expression after iLD-IL2 treatment. Of the 1,419 genes included in Component 187, 69 were also present in the target transcriptional panel applied to the DILfrequency dataset. The effects of iLD-IL2 immunotherapy and SARS-CoV-2 infection on the induction of these genes are compared by correlating component gene loading score (y axis) with $\log_2$ fold change (x axis) in the CD8 population. The positive and negative values in Component 187 loading scores represent the upregulation and downregulation of the corresponding gene, respectively, after SARS-CoV-2 infection. Each dot represents a gene. Larger dots represent genes with FDR-adjusted *P* values < 0.01 on Day 55 after iLD-IL2 treatment.

**(C)** Correlation between the IL2-AIS score and the Component 187 sample loading score. Each dot represents a participant in the COMBAT cohort and colours depict the different participant groups.

**(D)** STRING[260] protein interaction network of top 50 genes in Component 187. Each node represents a gene. Node colours represent Component 187 gene loading scores. IL2-AIS genes are labelled in bold text with directions of IL-2-induced differential expression shown in arrows. Each edge represents experimental or inferred protein-protein interaction between two genes. Edge widths and colours represent interaction scores, with thicker lines and darker colours representing higher scores.

Of the 1,419 genes contributing to Component 187, 77 were present in the transcriptional panel used in the DILfrequency study. I found that virtually all genes upregulated in IL2-AIS had negative loading scores in Component 187, and vice versa (Figure 80B), indicating a strong inverse correlation between the differential gene expression induced by IL-2 treatment in T1D patients and the response to infection in COVID-19 patients. Correspondingly, I observed a strong negative correlation between the mean Component 187 sample loading score and the mean IL2-AIS score among COMBAT participants (Figure 80C), suggesting that Component 187 and the IL2-AIS likely shared common underlying mechanisms.

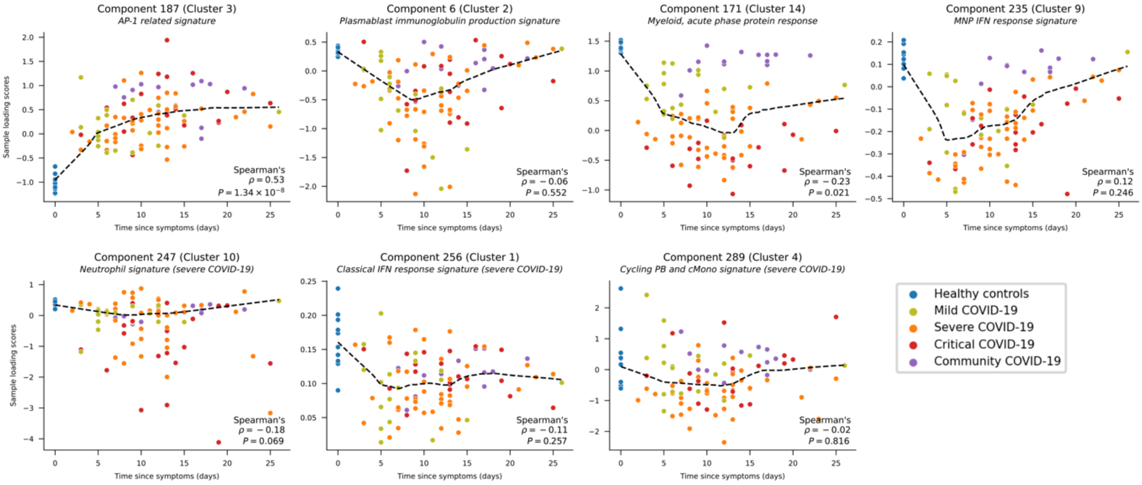


**Figure 81. Temporal patterns of selected gene expression components in COMBAT.**

The sample loading scores of seven gene expression components identified in the COMBAT dataset as being correlated with COVID-19 infection. Descriptions of each component are shown in italics. For each component, the direction of changes carries no predefined meaning and needs to be interpreted with respect to specific genes. Each dot represents a participant. Dashed black lines represent LOWESS curves.

Given the strong correlation between these two transcriptional signatures, I next performed a network analysis on the top 50 genes of Component 187 to gain some insight into the putative biological mechanism underpinning the identified IL2-AIS. This analysis highlighted a central role of the transcription factor NF-κB on the regulation of this transcriptional programme, as evidenced by a number of target genes previously shown[261] to be modulated by NF-κB in SARS-CoV-2 infected CD8$^+$ T cells, including *NFKBIA*, *NFKBIZ*, *TNFAIP3* and *CXCR4* (Figure 80D). Furthermore, I also observed a significant enrichment of known NF-kB target genes[262] within the top 50 genes of Component 187, including *RELB*, *NR4A2*, *DUSP1, CD69* and the AP-1 transcription factor complex genes (*FOS*, *FOSB*, *JUN*, *JUNB* and *JUND*). However, I note that these core NF-kB target genes, including the AP-1 genes, were not identified in the IL2-AIS (Figure 81D), suggesting that they were not significantly modulated by low-dose IL-2 therapy and are specifically modulated by SARS-CoV-2 infection. Notably, all identified NF-kB target genes show positive Component 187 loading scores (Figure 81D), suggesting that increased expression of early response factors such as NF-kB and the transcription factor complex AP-1 is involved in sustaining the pro-inflammatory gene expression profile in circulating immune cells during the post-acute phase of COVID-19.

## 2.3. Discussion

The findings reported here extended previous studies on the clinical application of low-dose IL-2 immunotherapy in T1D patients[215,263,264], and provided further insight into the mechanism of action of LD-IL-2, as detailed below.

First, I showed that low-dose IL-2 administered every three days for a period of one month was able to selectively increase the number of tTregs, which express FOXP3 and HELIOS, markers for stable Treg function, while showing no detectable impact on the expansion or cytotoxic gene expression of conventional effector T or CD56$^{dim}$ NK cells, demonstrating the high sensitivity of these immune subsets to LD-IL-2 treatment. I observed a particularly pronounced increase in the frequency of naïve FOXP3$^{+}$ HELIOS$^{+}$ Tregs after iLD-IL-2 therapy, suggesting that the Treg increases are caused by an increased representation of thymic-derived Tregs, and not the proliferation of peripherally-induced Tregs. This finding was consistent with the increased clonal diversity in T cells following an escalating LD-IL-2 regimen reported previously in graft-versus-host disease patients[265,266]. In addition to Tregs, the CD56$^{br}$ NK cell population is probably also contributing to the overall anti-inflammatory effects of iLD-IL-2, in an immunoregulatory role, as previously described[267,268].

Second, I found that iLD-IL-2 was associated with a reduction in IL-21-producing T cells, consistent with previous reports[247,269,270] that LD-IL-2 decreased the frequency of $T_{FH}$ cells in systemic lupus erythematosus patients. The effect was likely mediated through the inhibition of $T_{FH}$ differentiation, as IL-2 signalling has been shown to inhibit the differentiation of $T_{FH}$ cells *in vivo* in mice[271], and in an *in vitro* model in humans[272]. In humans, IL-21 has been implicated in the pathogenesis of several autoimmune diseases[273,274] including T1D[237,238,242]. In particular, an anti-IL-21 monoclonal antibody in combination with liraglutide showed promising results in preserving pancreatic beta cell function[239]. The observations here provided further support for the use of iLD-IL-2 in T1D, as well as other diseases associated with increased IL-21 production, such as systemic lupus erythematosus, rheumatoid arthritis, and psoriasis.

Third, I observed a decrease in circulating innate-like CD8$^+$ MAIT and $V_{\gamma9}V_{\delta2}$ T cell subsets during treatment and one month after. Since both subsets function in anti-viral and anti-bacterial defence[275,276], these findings suggest a role of IL-2 immunotherapy in the recruitment of MAIT and $V_{\gamma9}V_{\delta2}$ T cells to tissues following treatment, thereby increasing defence against viral and bacterial infections[277]. This could provide a mechanism for the previously observed decreased incidence of viral infections in SLE patients undergoing LD-IL-2 immunotherapy[278].

Finally, I identified a long-lived gene expression signature (IL2-AIS) in all assessed immune populations, which suggest a novel immunoregulatory mechanism of iLD-IL-2. IL2-AIS was marked by the upregulation of *CISH*, a well-characterised negative regulator of cytokine signalling[279], and the downregulation of key TNF-inducible genes such as *TNFAIP3*, *RGS1* and *AREG*, suggesting that iLD-IL-2 can decrease the homoeostatic levels of TNF for at least one month after the cessation of dosing. The mechanism of the dose-dependent establishment of the IL2-AIS observed on Day 27 remained unknown.

In light of the unprecedented breath of single-cell transcriptome datasets generated in COVID-19 patient cohorts, in particular the COMBAT cohort[229] and the INCOV cohort[256], which associated several key IL2-AIS genes with SARS-CoV-2 infection, I investigated IL2-AIS dynamics in the context of COVID-19. I found that SARS-CoV-2 infection leads to a long-lived alteration of the transcriptional profile of immune cells in blood for over three months after the onset of the clinical symptoms, which resembled the opposite of IL2-AIS, manifesting as the progressive and sustained decrease of IL2-AIS scores.

The longevity of transcriptional changes represented by the IL2-AIS in both iLD-IL-2 and COVID-19 context pointed to cytokines bound to the extracellular matrix as one possible explanation for the increased period of biological activity. Both

TNF and IL-2 are among the cytokines known to bind to the extracellular matrix, leading to their retention in within tissues extending far beyond their initial release[249–251,251]. The balance between pro- and anti-inflammatory cytokines bound to the extracellular matrix likely provides a regulatory mechanism to control immune responses. A perturbation of this balance via a sustained regulatory (e.g. iLD-IL-2 immunotherapy) or inflammatory (e.g. COVID-19) environment is therefore likely to remodel the transcriptional profile of immune cells and an alter their threshold for further stimulation, leading to long-term effects such as the PASC.

Analyses of top constituent genes of the Component 187 identified in the COMBAT study[229], which likely provides a more complete picture of the transcriptional changes related to the IL2-AIS in the COVID-19 context, revealed a central role of the NF-κB signalling pathway. This was supported by the upregulation of the NF-κB inhibitor genes *NFKBIA* and *NFKBIZ*, the AP-1 transcription factor complex genes *FOS*, *JUN*, *FOSB* and *JUNB*, as well as other NF-κB target genes such as *TNFAIP3*, *RELB*, *NR4A2*, *DUSP1* and *CD69*. The significance of the NF-κB pathway in COVID-19 patients was confirmed in a recent study and contrasted against influenza patients, who displayed a stronger type I interferon response[280]. Cell-line models infected with SARS-CoV-2 also highlighted the upregulation of NF-κB pathway genes including *NFKBIA*, *FOS*, *JUN* and *TNFAIP3*[281,282].

An important limitation of this study is the lack of mechanistic insights underpinning the induction of transcriptional changes related to IL2-AIS in either iLD-IL2 immunotherapy or COVID-19 contexts. Regarding the application of low-dose IL2 immunotherapy in T1D patients, which was the initial motivation of this study, it remains unclear whether the specific dosing regimen adopted in the DILfrequency study is necessary or sufficient to induce the IL2-AIS, how long does the transcriptional changes last beyond one month, and whether these changes can be translated into clinical benefits. Further investigations on the IL2-AIS, and more broadly the clinical effects of iLD-IL2 immunotherapy, have been planned by Prof John Todd's lab based on samples from the Interleukin-2 Therapy of Autoimmunity in Diabetes (ITAD) trial[283] and follow-up clinical studies, which would help answer these questions and eventually contribute to the possible future use of iLD-IL-2 treatment in the prevention of T1D diagnosis.

# 2.4. Methods

## 2.4.1. Study design and ethics statement

The study was performed in accordance with the guidelines for good clinical practice and the Declaration of Helsinki. Study participants included T1D patients enrolled in the adaptive study of IL-2 dose frequency on regulatory T cells in type 1 diabetes (DILfrequency), a response-adaptive trial of repeated doses of IL-2 administered using an interval dosing approach[3]. All 13 study participants, who were all adult patients diagnosed with T1D over 60 months before recruitment into the study, were selected from the 3-day interval dosing group and were treated with IL-2 with doses ranging from 0.2 to $0.47\times10^6$ IU/$m^2$.

Approval was obtained from the Health Research Authority, National Research Ethics Service (14/EE/1057), London, United Kingdom. The trials were registered at the International Standard Randomised Controlled Trial Number Register (ISRCTN40319192) and ClinicalTrials.gov (NCT02265809). The study protocols were published in advance of the completion and final analysis of the trials[284]. All participants provided written informed consent prior to their participation in the studies.

## 2.4.2. Normalisation, integration, dimensionality reduction and clustering

Preprocessing of single-cell sequencing data, including read filtering, read alignment, error correction, doublet removal, and additional quality control were conducted by Dominik Trzupek and Fiona Hamey.

After preprocessing, data normalisation was performed using Seurat[232]. The RNA count matrices were normalised by log normalisation. Specifically, the RNA counts for each cell were (i) divided by the total counts for that cell, (ii) multiplied by a scale factor of 10,000, and (iii) natural-log transformed after adding one pseudo count. The AbSeq counts for each cell were considered compositional and normalised by a centred log ratio transformation[285]. Normalised RNA and AbSeq expression matrices were then centred and scaled. Specifically, each feature was linearly regressed against selected latent variables including the number of RNA and AbSeq features, and the donor of origin, where appliable. The resulting residuals were mean-centred and divided by their standard deviations.

Integration of the scaled AbSeq expression matrices from different participants was performed using a combination of canonical correlation analysis and identification of mutual nearest neighbours, implemented in Seurat[232]. After integration was performed, the updated AbSeq expression matrices produced by the integration algorithm were used in subsequent dimensionality reduction and clustering.

Dimensionality reduction and clustering were performed using Seurat[232]. Specifically, principal components (PCs) were calculated separately for each assay using the top variable features of the scaled or integrated expression matrices, and a weighted nearest neighbour (WNN) graph as well as a weighted shared nearest neighbour (WSNN) graph were generated on the ten top PCs from each assay, unless otherwise specified. I used $k$=30 nearest neighbours for each cell. Uniform Manifold Approximation and Projection (UMAP)[286] embeddings were then calculated based on the WNN graph. The Louvain algorithm[287] was applied to identify clusters of cells based on a WSNN graph, at the selected resolution level. Clusters with fewer than ten cells were removed. The resulting clusters were manually annotated by Ricardo Ferreira based on marker genes and sorting gate information to elucidate cell types and label suspicious clusters.

During preprocessing, a subset of cells, most of which inferred to be CD56$^{br}$ NK cells, were identified to lack barcoding information that were used to delimitate FACS-isolated cell populations and the time of sampling. Although the reason for this barcode loss was unclear, the data quality for these cells was acceptable. Therefore, CD56$^{br}$ NK cells missing barcode information were preserved in dimensionality reduction and clustering (Figure 61), but excluded in analyses for which timepoint information was relevant, such as differential abundance and differential expression analyses. Other cells missing barcode information were excluded.

I observed that the PBMC aliquot corresponding to the Day 55 sample of Participant 8 yielded very low cell numbers (603 unstimulated and 67 stimulated cells) and displayed a compromised FACS staining profile. These cells were preserved in dimensionality reduction and clustering procedures but excluded from downstream analyses. All other 38 samples corresponding to 38 visits yielded the expected cell frequencies and numbers.

### 2.4.3. Differential abundance analysis

To identify specific cell abundance changes across different sampling timepoints, I calculated the abundance of each cell subset in the five T and NK populations. A cell subset was defined as a group of cells from the same cluster, or several clusters deemed to be similar. The abundance of each cell subset for each participant and each timepoint was calculated as the proportion of that cell subset within all cells from the same participant, timepoint and FACS-isolated population. For each cell subset, a two-sided Wilcoxon signed-rank test was performed to compare the abundance values on Day 0 with the abundance values on Day 27 or Day 55 in the 13 participants selected for single-cell analysis. Benjamini-Hochberg FDR correction was applied after pooling all resulting *P* values. For each cell type and each participant, the $\log_2$ fold change values between corresponding abundance values were calculated and visualised. Cells were compared only if they were from the same FACS-isolated population, as cells from different populations were mixed in predefined proportions that are not biologically relevant.

### 2.4.4. Pseudotime trajectory analysis

Pseudotime trajectory analysis was performed using Slingshot[226] and tradeSeq[288]. First, 10-dimensional UMAP embeddings were generated for stimulated cells in the CD4$^+$ Treg and Tconv populations based on RNA and AbSeq data, as described above. Second, the Slingshot algorithm was applied to UMAP embedding data and manually annotated cluster labels, with the naïve Tconv subset selected as the starting point, and other parameters set to the default. Finally, the resulting pseudotime data were used as the input for the tradeSeq algorithm to model the expression levels of each gene along each trajectory.

### 2.4.5. Generation and normalisation of pseudo-bulk expression data

Pseudo-bulk expression were calculated for three datasets: DILfrequency (this study), COMBAT[229] and INCOV[256]. For each dataset, pseudo-bulk expression matrices were generated by aggregating mRNA counts from cells from the same donor, timepoint (if applicable) and cell population. For the DILfrequency dataset, which was based on a custom mRNA panel of 585 transcripts, pseudo-bulk samples with less that 100 total mRNA counts were removed. For the COMBAT and INCOV datasets, which were based on whole-transcriptome sequencing, pseudo-bulk samples with less than 2,000 total mRNA counts or less than 500 IL2-AIS mRNA counts were removed. The pseudo-bulk expression matrices for each dataset were normalised by dividing raw counts with sample-specific scale factors calculated using the median-of-ratios method previously described[289]. The total mRNA counts of pseudo-bulk samples in the COMBAT and INCOV datasets are summarised in Table 6.

| Dataset | Cell type | Median count | IQR | Excluded |
|---|---|---|---|---|
| COMBAT | T | 10892548.5 | 9297326.5 | 2 |
| | B | 875124 | 862786.25 | 16 |
| | NK | 1274245.5 | 1499071.25 | 3 |
| | MNP | 6964687.5 | 5177158.25 | 4 |
| INCOV | CD4 | 1374058 | 1678171.5 | 0 |
| | CD8 | 703045 | 843708.5 | 0 |
| | B | 151089 | 203788 | 0 |
| | Monocytes | 1040401 | 1642817.5 | 0 |
| | NK | 357648 | 495790.5 | 0 |

**Table 6. Summary statistics of total mRNA counts of pseudo-bulk samples.**

Median and interquartile range (IQR) of total mRNA counts in each group of pseudo-bulk samples in the COMBAT and INCOV cohorts, and the number of excluded samples in each group. MNP, mononuclear phagocytes.

### 2.4.6. Differential expression analysis

Differential expression analyses were performed separately for each dataset based on the pseudo-bulk expression matrix using DESeq2[290]. For the DILfrequency dataset, the likelihood ratio test was used, with a full model including timepoints and the participants as independent variables, and a reduced model including only participants as the independent variable. For the COMBAT dataset, the Wald test was used, with patient groups (COVID-19 or healthy control) as the independent variable. For the INCOV dataset, considering the samples were taken during a wide range of time post COVID-19 symptoms, for each participant, I assigned the earliest sample taken 0-14 days post COVID-19 symptoms as the acute phase sample, and the earliest sample taken 29-84 days post COVID-19 symptoms as the post-acute phase sample. Only participants that have both acute and post-acute phase samples available were included in the differential expression analysis. The cut-off timepoints for acute and post-acute phase samples were selected to maximise the number of available participants, while minimizing the heterogeneity within each group. The likelihood ratio test was used, with a full model including sampling timepoints (acute or post-acute phase) and participants as independent variables, and a reduced model including only participants as the independent variable. For all datasets, the apeglm method[291] was applied to shrink the resulting fold change values, and the Benjamini-Hochberg FDR correction was applied after pooling all resulting *P* values.

## 2.4.7. Deriving the cell proliferation score

Single-cell proliferation scores were calculated using a previously described approach[292] based on the average normalised expression levels of a pre-selected set of 11 mRNA features related to cell cycle (*AURKB*, *HMGB2*, *HMMR*, *MCM4*, *MKI67*, *PCLAF*, *PCNA*, *TK1*, *TOP2A*, *TYMS*, and *UBE2C*), subtracted by the aggregated expression of a control set of 50 randomly selected mRNA features. The proliferation scores were used to identify two cycling clusters from the 15 functional T and NK cell subsets identified in blood from the initial clustering (Figure 69A). I noted that the distinct co-expression of these cell-cycle genes present in the targeted transcriptional panel superseded more subtle functional differences and aggregated all cycling T and NK cells into two respective clusters, regardless of their original FACS-isolated population. Given this lack of functional differentiation and to avoid the potential overwhelming effect of these cell-cycle genes on the resulting clustering visualisation, I opted to separate these cycling T and NK cells from the rest of the cells included in subsequent clustering steps to assess IL-2-induced changes in their relative abundance in blood.

### 2.4.8. Deriving the IL2-AIS score

For the DILfrequency dataset, the IL2-AIS scores were calculated based on the normalised pseudo-bulk expression levels of the 20 upregulated signature genes (CISH, TNFSF14, OAS1, *GIMAP7*, *GIMAP5*, *TNFSF10*, *TAGAP*, *STAT1*, *MYC*, *FASLG*, *CX3CR1*, *PTGDR2*, *CRTAM*, *EOMES*, *IL32*, *CCR10*, *CCR1*, *CXCR1*, *CD40LG*, and *ID3*) and the 21 downregulated signature genes (*AREG*, *DUSP5*, *TNFAIP3*, *RGS1*, *CXCR4*, *DUSP2*, *DUSP4*, *DDIT4*, *NFKBIA*, *FOSL2*, *NFKBIZ*, *ZBTB16*, *SLC2A3*, *BTG2*, *SOX4*, *OSM*, *SGK1*, *TGFBR3*, *OTUD1*, *COLQ*, and *CCL5*). For the COMBAT and INCOV datasets, a similar approach was applied to calculate the IL2-AIS scores for each pseudo-bulk sample. Specifically, z-scores of normalised expression levels of the 41 signature genes were first calculated for each pseudo-bulk sample within each major cell type. The IL2-AIS scores were then derived as the sum of z-scores of upregulated signature genes, subtracted by that of downregulated signature genes. As the z-scores were calculated from samples within a dataset, one important limitation of this definition was that comparisons of IL2-AIS scores were only allowed within the same dataset, but not across different datasets.

### 2.4.9. Modelling the dynamics of the IL2-AIS score

For the INCOV cohort, where multiple longitudinal samples are available for most individuals, I modelled the IL2-AIS scores using a Bayesian linear model to account for the inter-individual variation:

$$S_{i,t} \sim \mathrm{Normal}\left(\mu_{i,t}, \sigma\right)$$

$$\mu_{i,t} = \alpha_i + \beta_t$$

$$\alpha_i \sim \mathrm{Normal}(0, 15) \;\; \text{for } i = 1..168$$

$$\beta_t \sim \mathrm{Normal}(0, 10) \;\; \text{for } t = 1..7$$

$$\sigma \sim \mathrm{LogNormal}(0, 5)$$

where $S$ is the observed IL2-AIS scores, $i$ is the index of the individual, and $t$ is the index of sampling time represented as a categorical variable with seven levels: 0-7 days, 8-14 days, 15-28 days, 29-60 days, 61-90 days, 91-150 days, and ≥ 151 days. $\alpha_i$ and $\beta_t$ represent the individual-specific effect and the effect of sampling time, respectively. As the IL2-AIS score was formulated as the sum of z-scores of the 41 IL2-AIS genes, $\mathrm{E}(S) = 0$. Therefore, an intercept term for $\mu_{i,t}$ was not included. I interpreted $\alpha_i$ as the time-adjusted IL2-AIS score of individual $i$, and $\beta_t$ as the expected IL2-AIS score given sampling time $t$. Regularising priors were used for $\alpha_i$ and $\beta_t$. Posterior mean values and 95% confidence intervals of parameters were estimated using a Markov Chain Monte Carlo approach implemented in Turing.jl[293].

## 2.4.10. NanoString transcriptomic data analysis

The processed NanoString bulk transcriptomic data for the Gedda 2022 cohort[259] was accessed from Gene Expression Omnibus using accession code GSE211378. Among the 162 convalescent COVID-19 participants, 23 were excluded due to the lack of precise date of onset of COVID-19 symptoms. All 40 healthy control participants were included. Among the 785 genes profiled in the dataset using the NanoString nCounter Human Host Response panel, 16 IL2-AIS genes were present, including 12 upregulated genes (*MYC*, *CXCR1*, *OAS1*, *TNFSF10*, *FASLG*, *CCR10*, *STAT1*, *CX3CR1*, *CD40LG*, *IL32*, *EOMES*, and *CCR1*) and four downregulated genes (*OSM*, *SLC2A3*, *CXCR4*, and *CCL5*). The IL2-AIS* scores were defined for each sample as the sum of z-scores of normalised transcription levels of upregulated IL2-AIS genes, subtracted by that of downregulated IL2-AIS genes.

## 2.4.11. STRING network analysis

From the 1,419 constituent genes of COMBAT Component 187, the top 50 genes with the highest loading scores were selected for STRING protein interaction network analysis[260]. Given the very strong correlation between the relative contribution of the same core set of genes to both the IL2-AIS and Component 187, the selection of the top 50 genes of Component 187 not only facilitated the visualisation of the gene network, but also allowed to focus on the main biological pathways contributing specifically to the IL2-AIS identified in this study. A gene network and pathway analysis on the full 1,419 constituent genes of Component 187 is provided in the COMBAT study[229]. Physical and functional interactions identified from text mining, experimental evidence, annotated databases, and co-expression were used. The minimum required interaction score was set to 0.15.

# General discussions

This thesis includes two projects that are connected scientifically by their implications to autoimmune diseases and technically by the application of high-throughput sequencing. Underpinning these connections is the complexity of the human immune system in terms of the diversity of immune genes and the heterogeneity of immune cells.

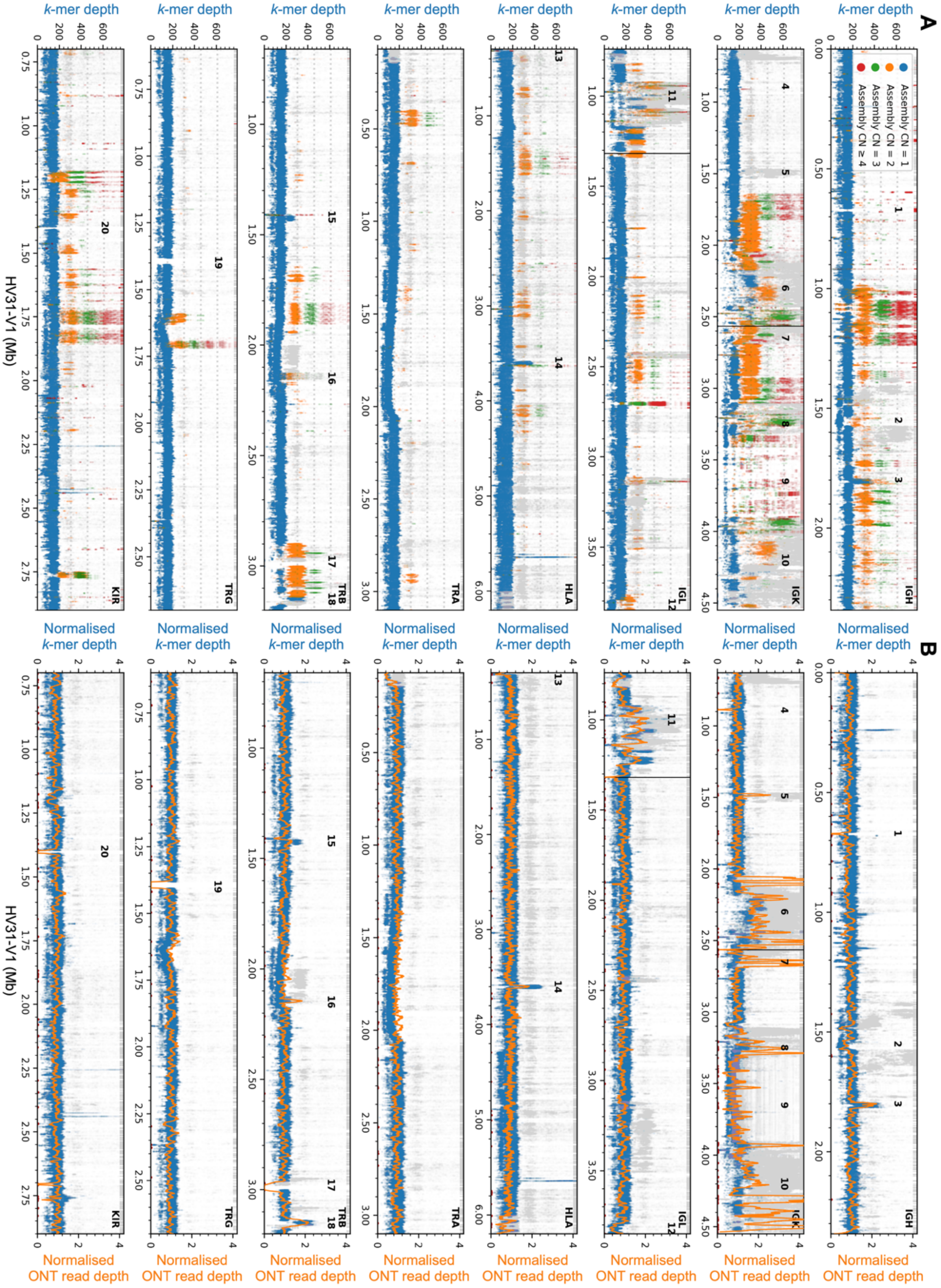
The achievements made in the *de novo* assembly of the human genome during the past few years were largely driven by the emergence of better sequencing platforms, new bioinformatics algorithms, and growing computational infrastructure. With the T2T reference genome[65], a manually curated, gap-free, haploid assembly, and the draft human pan-genome references[49], a collection of algorithmically built, high-quality, diploid assemblies, it is easy to envision an era when complete personal diploid assemblies, along with multi-modal, high-resolution functional data, become widely accessible. What that will reveal about the immune system, human genetics, and life sciences in general, remains to be seen.

The development of low-dose IL-2 immunotherapy as a treatment for T1D and other autoimmune diseases, if clinically proved, may become a textbook example of interdisciplinary drug discovery that pieced together key information from a broad spectrum of research, such as the high affinity of the trimeric IL-2 receptor encoded by *IL2RA*[294], the genetic associations between *IL2RA* and T1D[213,295], the heterogeneity of Tregs[296,297], recombination IL-2 and its application in cancer immunotherapy[207], and the optimal dosing region of IL-2 in T1D patients[3,216]. The key question of whether low-dose IL-2 immunotherapy provides clinical benefits to T1D patients, is currently under active investigation. With the development of other treatments targeting early-stage T1D, such as anti-CD3[194], anti-IL-21[239] and

Treg cell therapies[298], the prospect of identifying high-risk individuals genetically and delay or even prevent T1D onset clinically is becoming increasingly realistic.

# Appendix

## Appendix figures

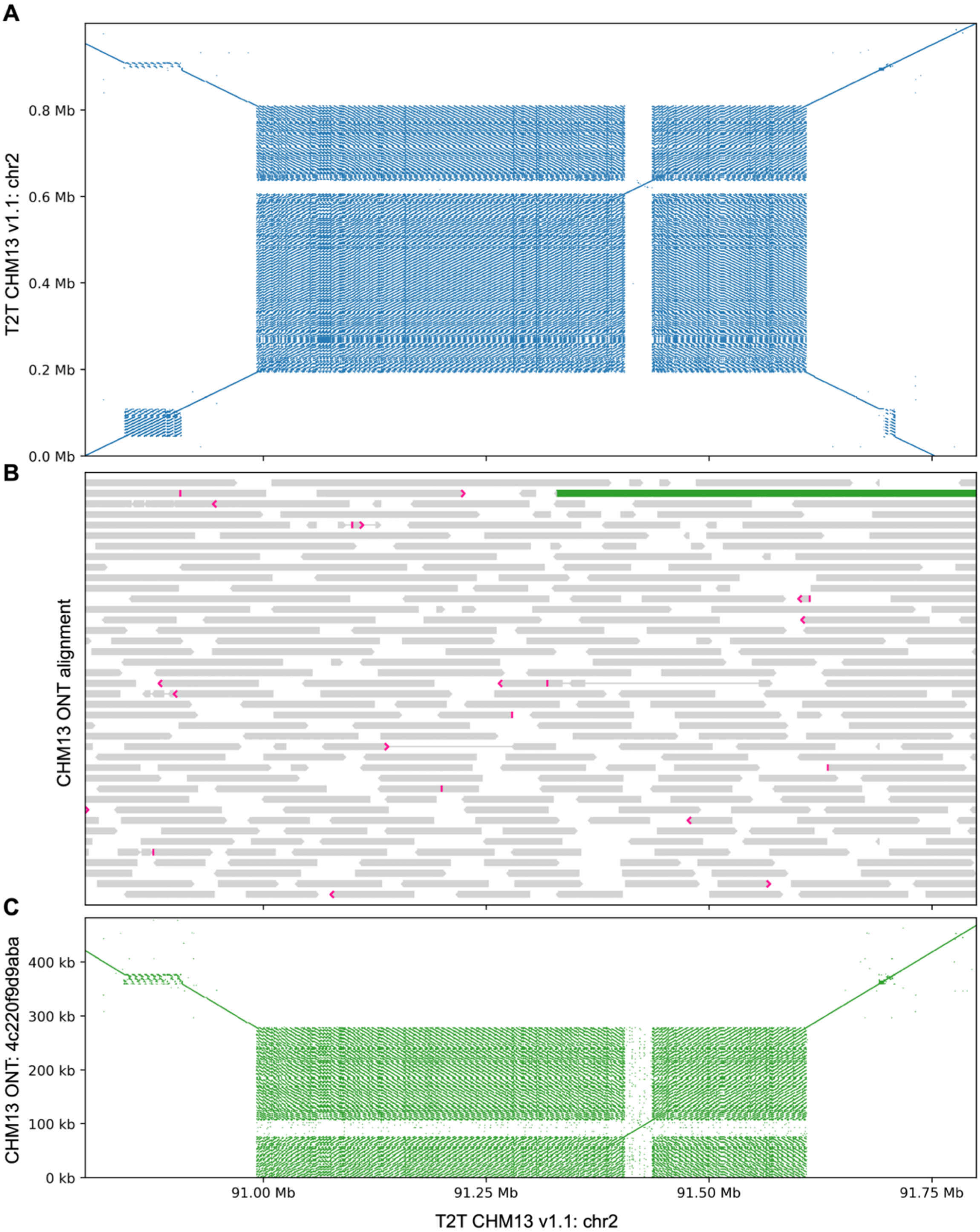


## Appendix Figure 1. Assembly validation based on *k*-mer depths.

**(A)** Validation *k*-mer depths for each region in the HV31-V1 assembly. Colours denote the copy number of a given *k*-mer in the HV31-V1 assembly. *k* = 31.

**(B)** Normalised *k*-mer depths for each region in the HV31-V1 assembly. Depth values were normalized by dividing by the peak depths of unique homozygous *k*-mers as shown in Figure 14A. Orange lines show ONT-2019 read coverage depth normalized to the genome-wide average coverage depth (63×).

In **(A-B)**, *k*-mers that found both inside and outside the given regions are shown in grey. Additional information about each numbered location is detailed in Appendix Table 1.

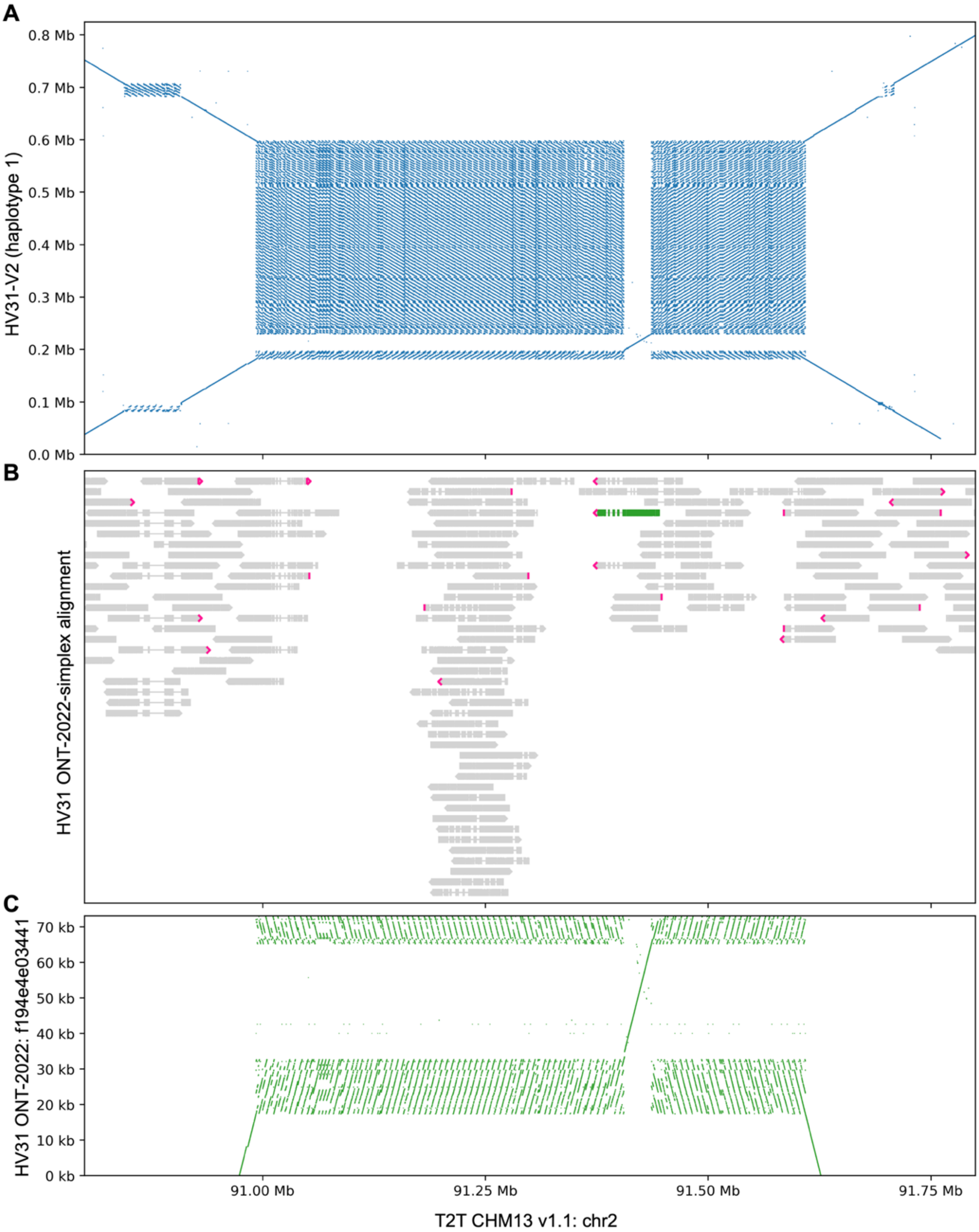


**Appendix Figure 2. Validation of IGK heterochromatin sequence in the T2T CHM13 assembly.**

**(A)** *k*-mer sharing dot plot comparing the heterochromatin sequence near the IGK locus in the T2T CHM13 v1.1 assembly to itself.

**(B)** Alignment of publicly available ONT sequencing data[299] to the T2T CHM13 v1.1 assembly. ONT read 4c220f9d9aba covering the 32 kb unique sequence is highlighted in green.

**(C)** *k*-mer sharing dot plot comparing ONT read 4c220f9d9aba to the T2T CHM13 v1.1 assembly.

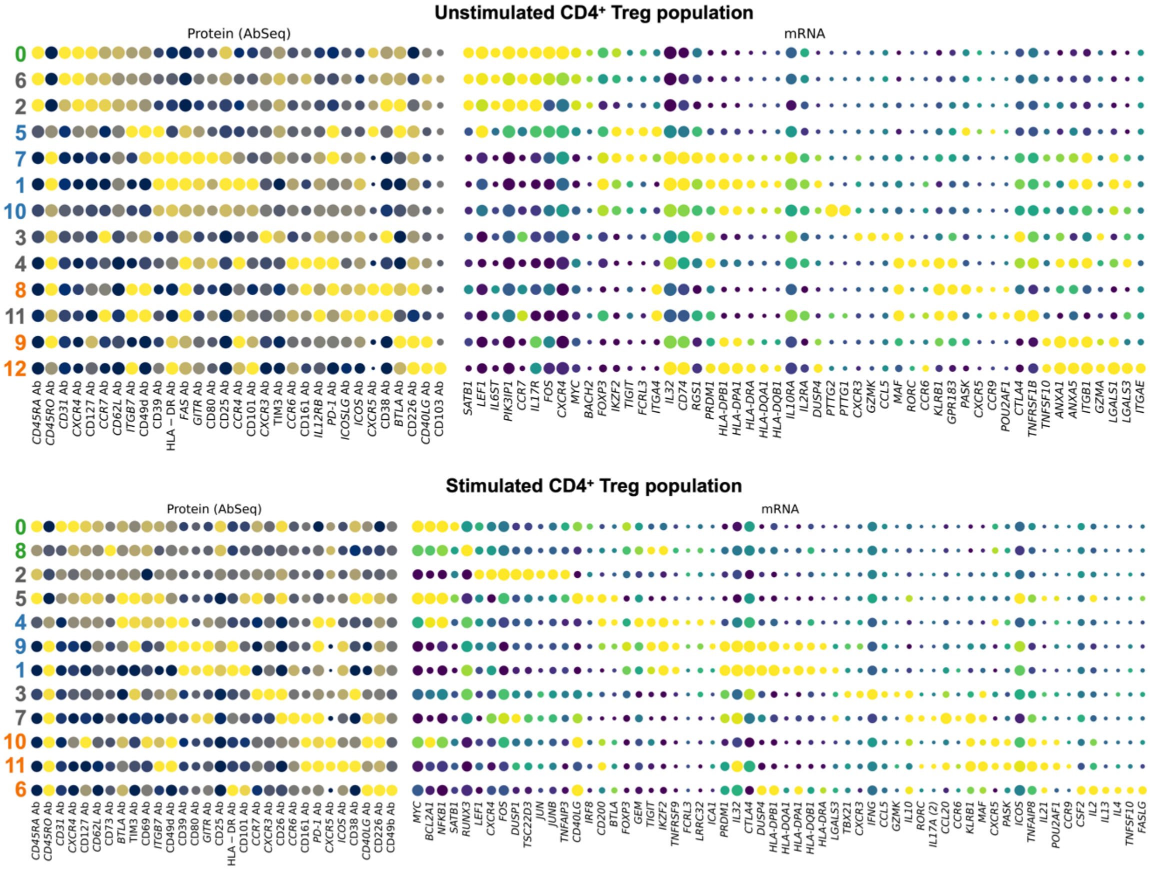


**Appendix Figure 3. Validation of IGK heterochromatin sequence in the HV31-V2 assembly.**

**(A)** *k*-mer sharing dot plot comparing the heterochromatin sequence near the IGK locus in the HV32-V1 assembly haplotype 1 (y axis) with the same region in the T2T CHM13 v1.1 assembly (x axis).

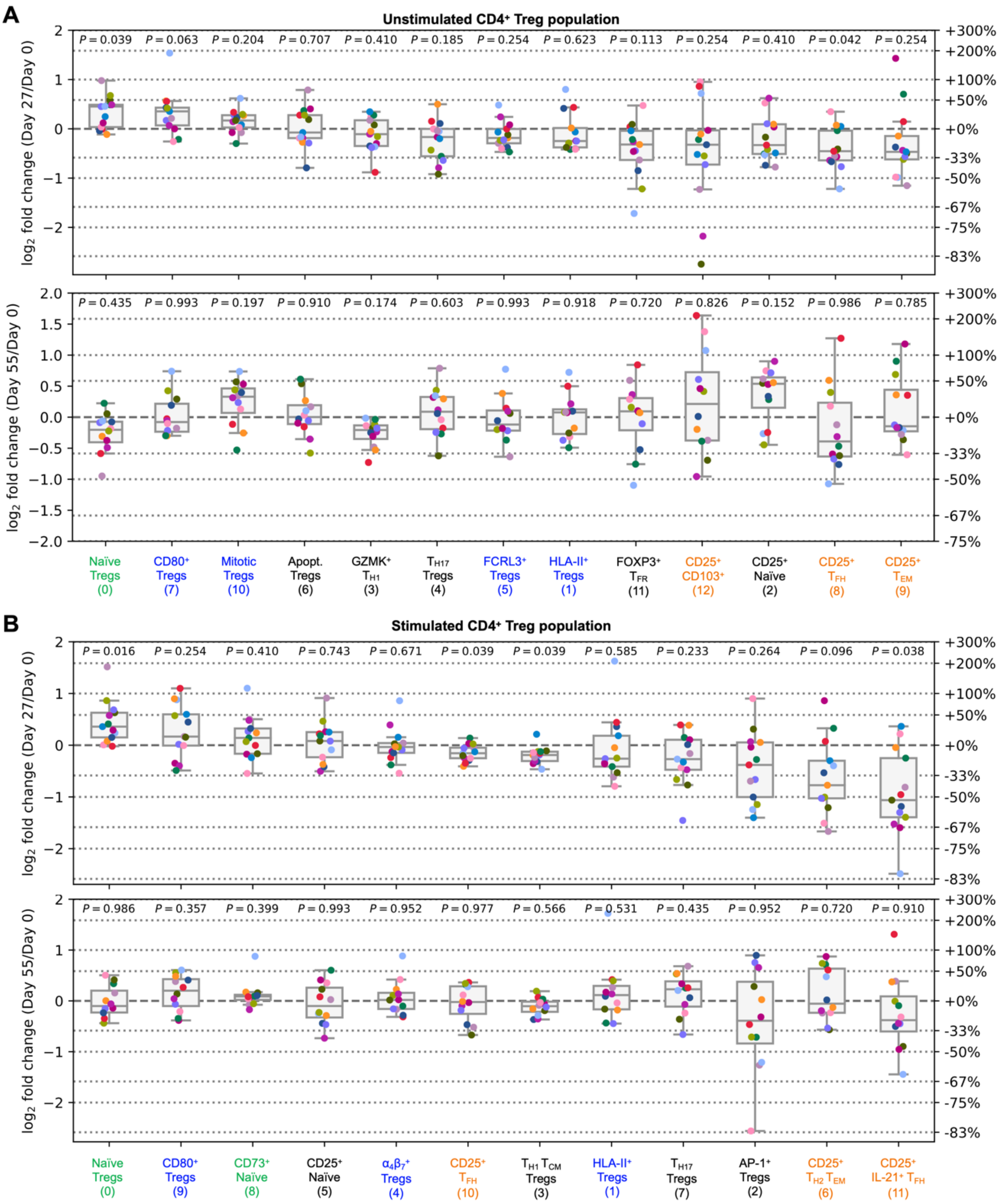
**(B)** Alignment of HV31 ONT-2022 simplex sequencing to the T2T CHM13 v1.1 assembly. ONT read f194e4e03441 covering the 32 kb unique sequence is highlighted in green.

**(C)** *k*-mer sharing dot plot comparing ONT read f194e4e03441 to the T2T CHM13 v1.1 assembly.



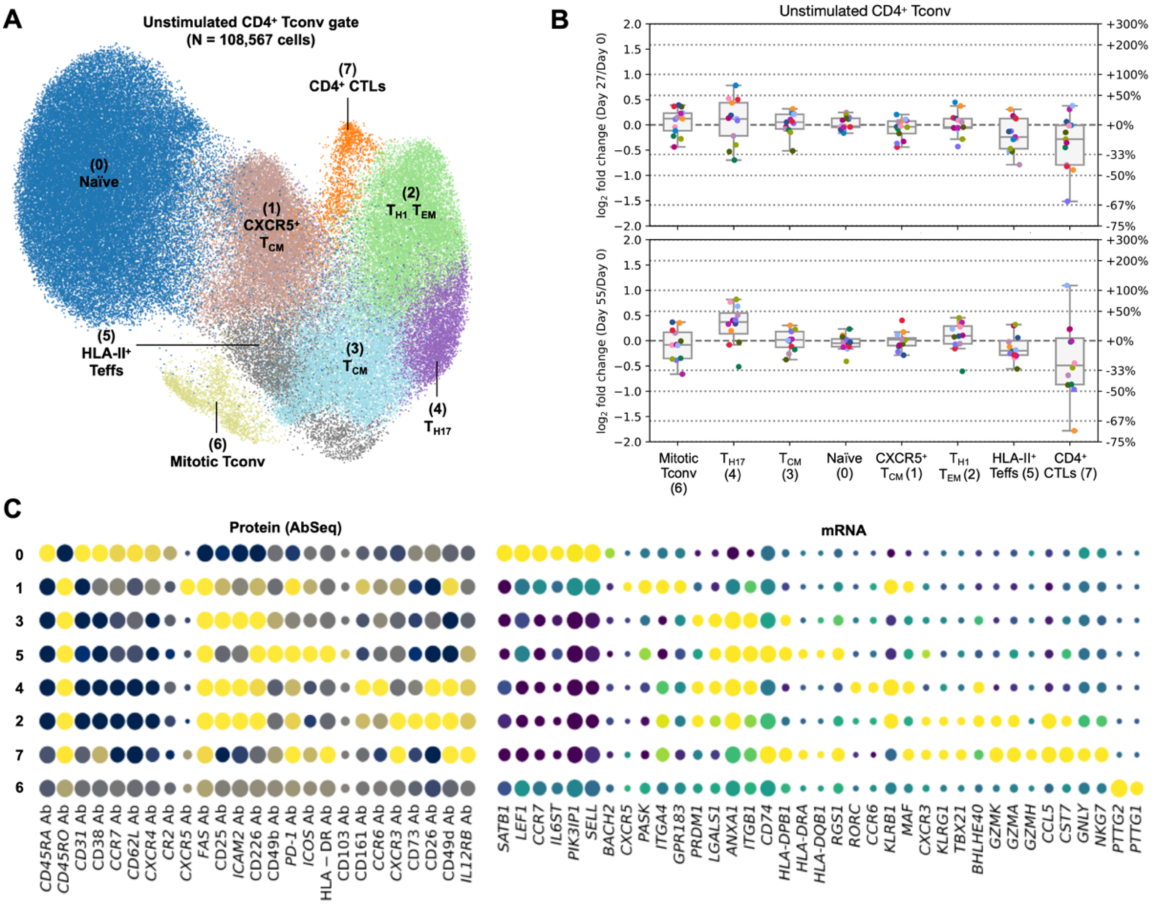
**Appendix Figure 4. Functional annotation of CD4$^+$ Treg subsets.**

Expression levels of selected mRNA and protein markers in each identified cluster in the CD4$^+$ Treg population. Cluster annotations are shown in Figure 62. Larger dots represent higher fraction of cells express the marker, while brighter colours represent higher mean expression levels. Clusters classified as naïve tTregs, memory tTregs, and CD25$^+$ Teffs are annotated in green, blue and orange, respectively. Other miscellaneous clusters are annotated in black.



**Appendix Figure 5. Abundance changes of CD4$^{+}$ Treg subsets.**

**(A-B)** Abundance changes of identified clusters in unstimulated **(A)** or stimulated **(B)** cells in the CD4$^{+}$ Treg population comparing Day 27 (top) or Day 55 (bottom) with Day 0. Clusters classified as naïve tTregs, memory tTregs, and CD25$^{+}$ Teffs are annotated in green, blue and orange, respectively. Other miscellaneous clusters are annotated in black.

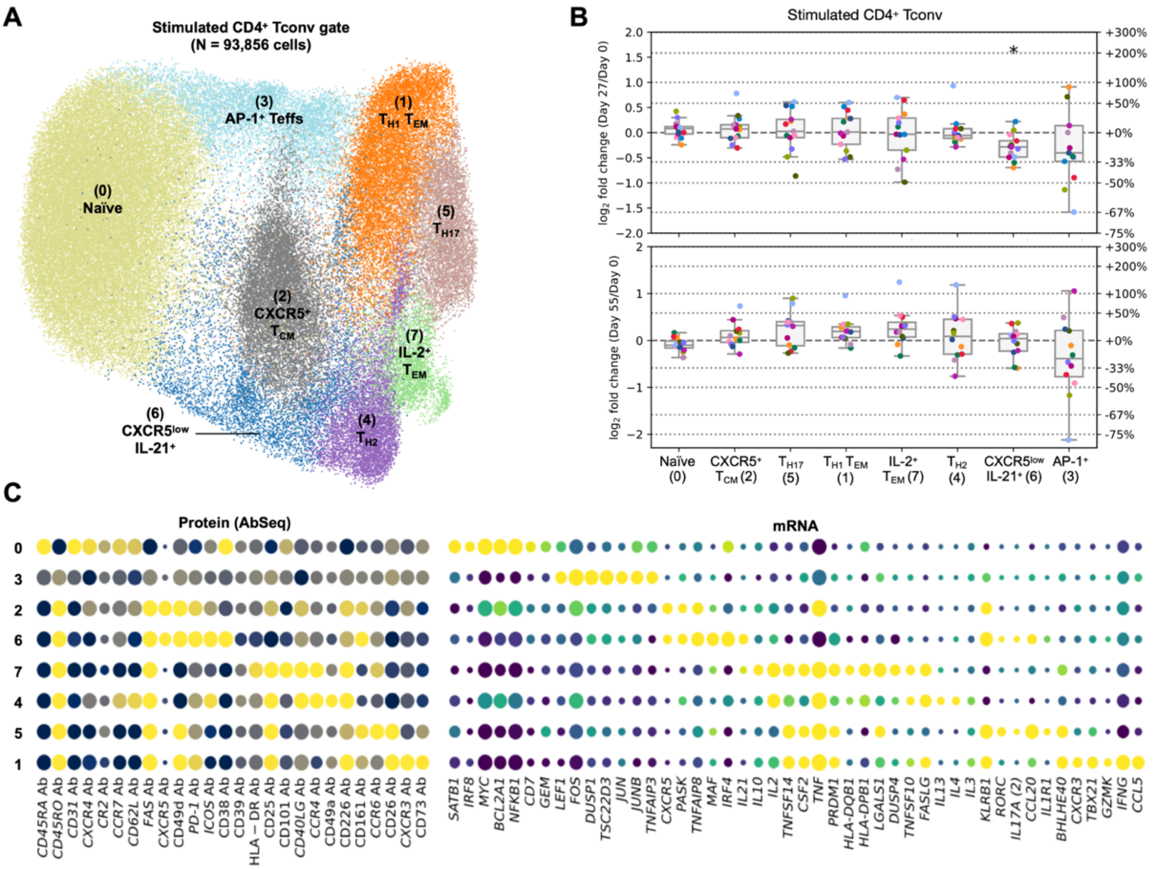


**Appendix Figure 6. Functional annotation and abundance changes of unstimulated CD4$^+$ Tconv subsets.**

**(A)** UMAP embedding unstimulated cells in the CD4$^+$ Tconv population, coloured by identified clusters.

**(B)** Abundance changes of each cluster in **(A)** comparing Day 27 (top) or Day 55 (bottom) with Day 0.

**(C)** Expression levels of selected mRNA and protein markers in each cluster in **(A)**. Larger dots represent higher fraction of cells express the marker, while brighter colours represent higher mean expression levels.

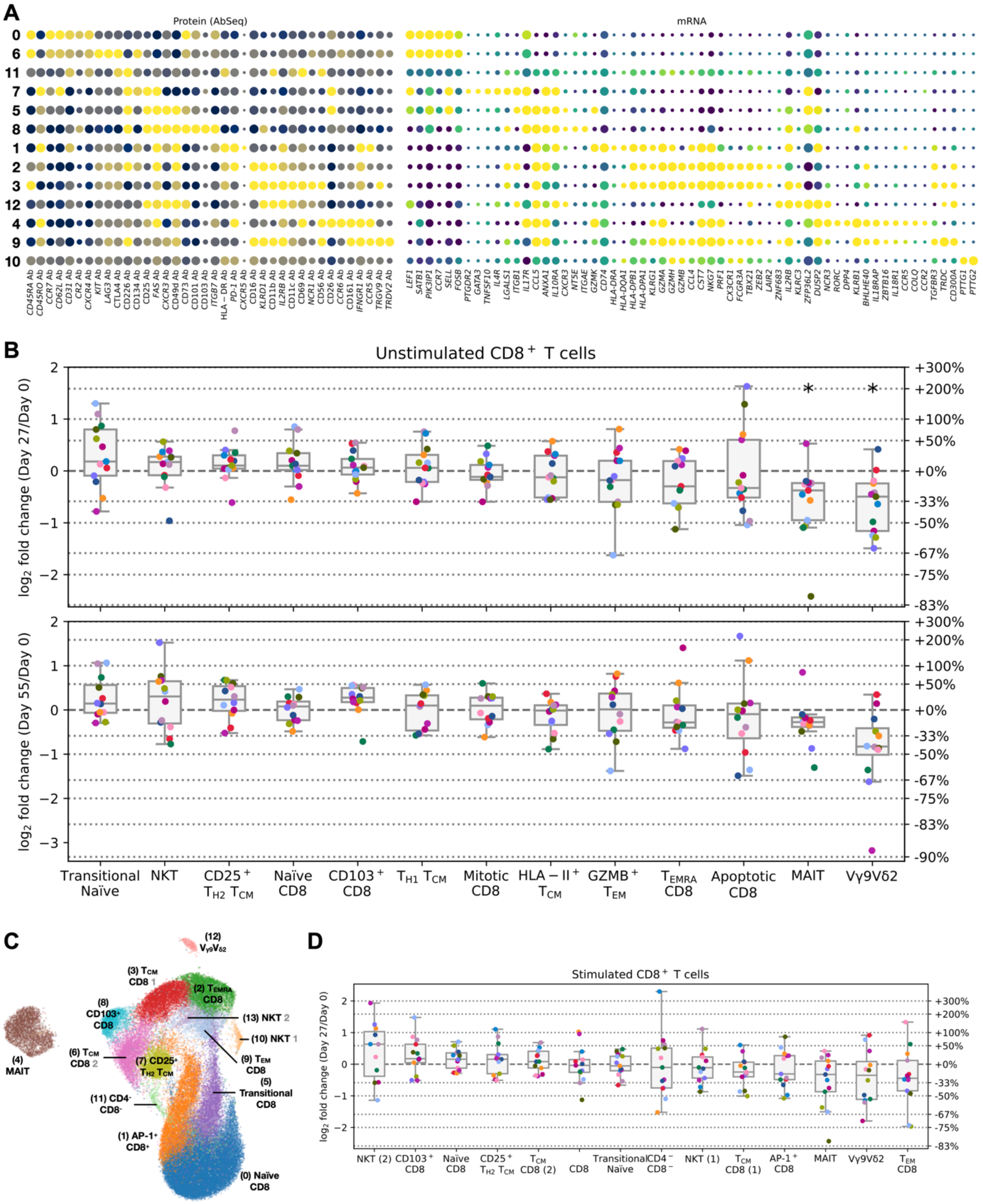


Appendix Figure 7. Functional annotation and abundance changes of stimulated CD4⁺ Tconv subsets.

**(A)** UMAP embedding stimulated cells in the CD4⁺ Tconv population, coloured by identified clusters.

**(B)** Abundance changes of each cluster in **(A)** comparing Day 27 (top) or Day 55 (bottom) with Day 0.

**(C)** Expression levels of selected mRNA and protein markers in each cluster in **(A)**. Larger dots represent higher fraction of cells express the marker, while brighter colours represent higher mean expression levels.



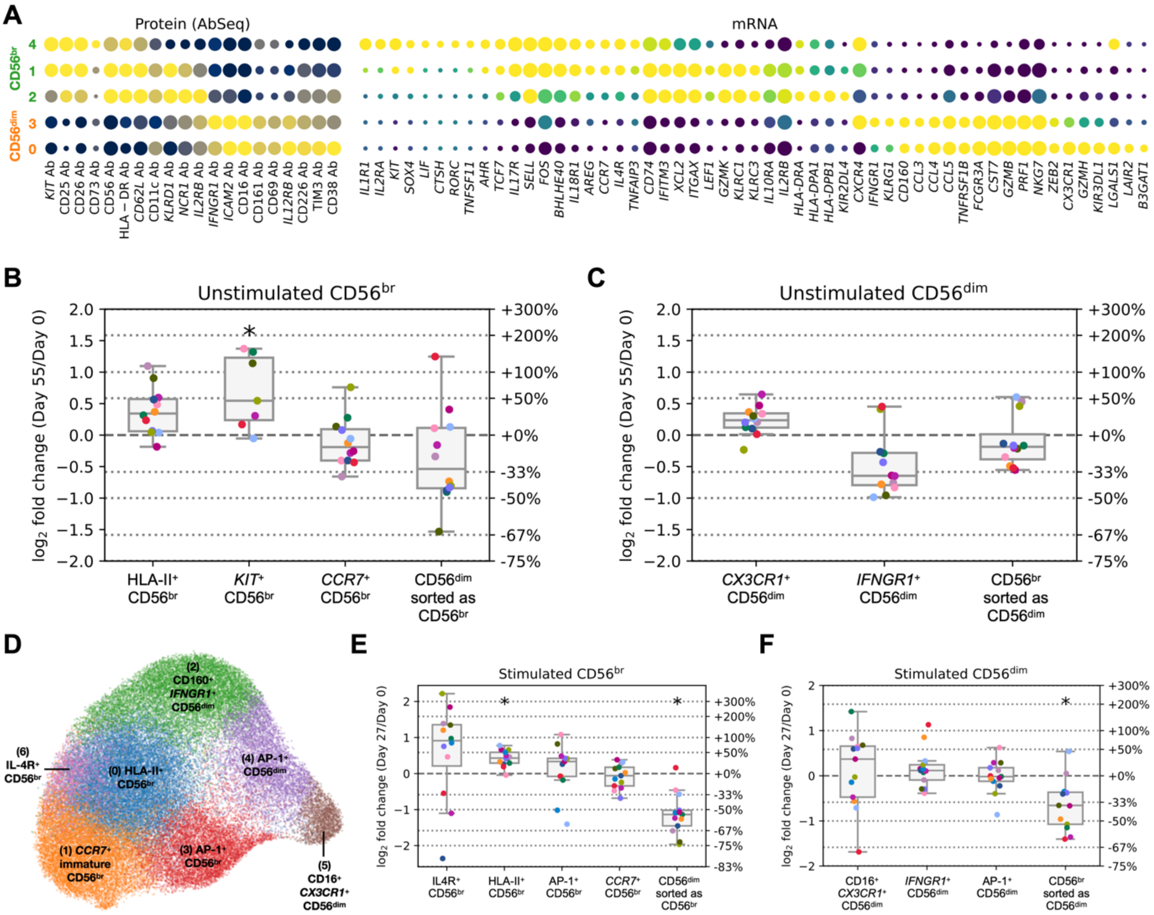
**Appendix Figure 8. Functional annotation and abundance changes of CD8$^+$ T subsets.**

**(A)** Expression levels of selected mRNA and protein markers in each cluster of unstimulated cells in the CD8$^+$ population. Cluster annotations are shown in Figure 68A. Larger dots represent higher fraction of cells express the marker, while brighter colours represent higher mean expression levels.

**(B)** Abundance changes of each cluster of unstimulated cells in the CD8$^+$ population comparing Day 27 (top) or Day 55 (bottom) with Day 0.

**(C)** UMAP embedding of stimulated cells in the CD8$^+$ T population, coloured by identified clusters.

**(D)** Abundance changes of each cluster in **(C)** comparing Day 27 with Day 0.



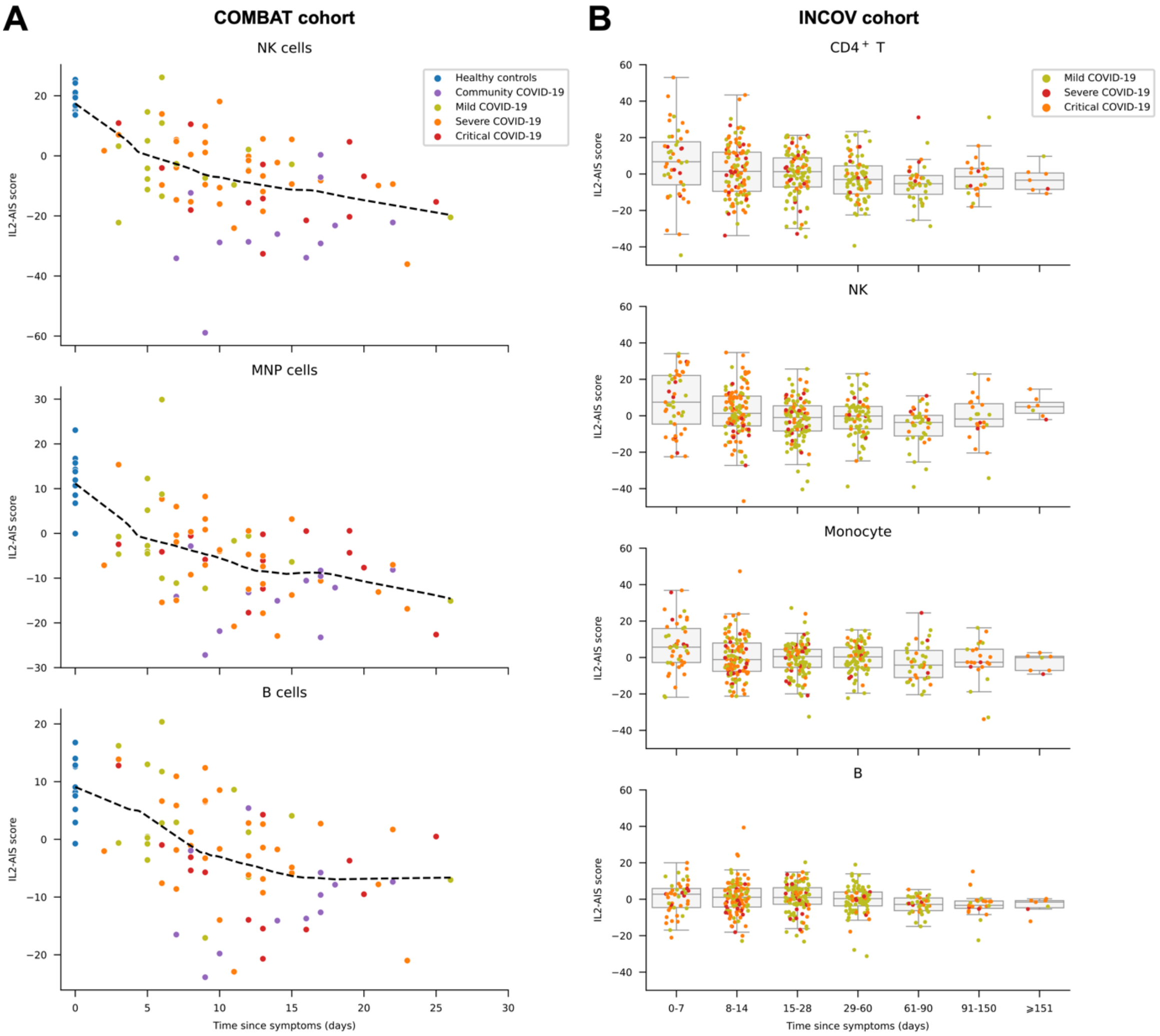
**Appendix Figure 9. Functional annotation and abundance changes of CD56$^{br}$ and CD56$^{dim}$ NK subsets.**

**(A)** Expression levels of selected mRNA and protein markers in each cluster of unstimulated cells in the CD56$^{br}$ and CD56$^{dim}$ NK populations. Cluster annotations are shown in Figure 68D. Larger dots represent higher fraction of cells express the marker, while brighter colours represent higher mean expression levels.

**(B-C)** Abundance changes of each cluster of unstimulated cells in the CD56$^{br}$ and CD56$^{dim}$ NK populations comparing Day 27 **(B)** or Day 55 **(C)** with Day 0.

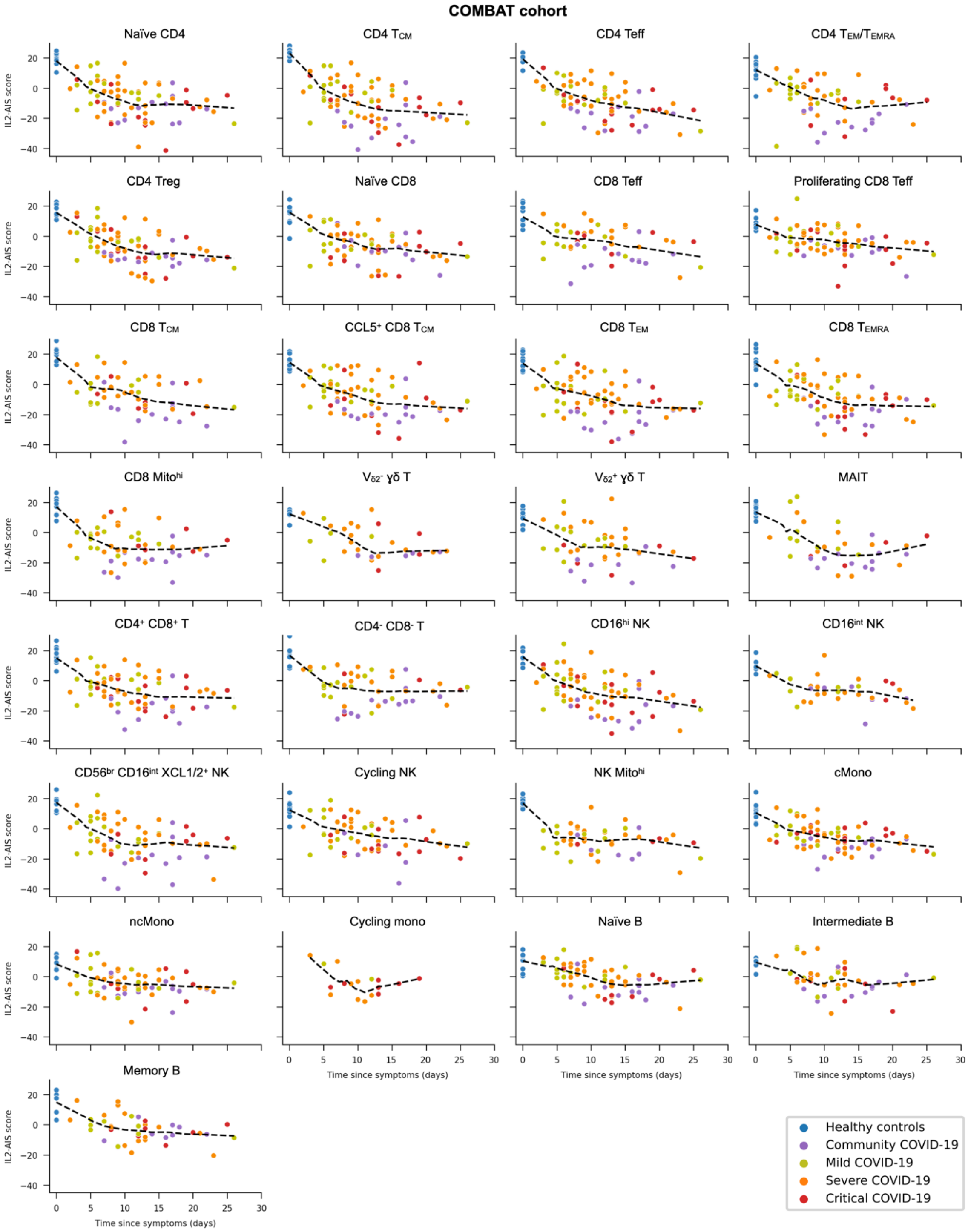
**(D)** UMAP embedding of stimulated cells from the CD56$^{br}$ and CD56$^{dim}$ NK populations, coloured by identified clusters.

**(E-F)** Abundance changes of each cluster in **(D)** comparing Day 27 **(E)** or Day 55 **(F)** with Day 0.



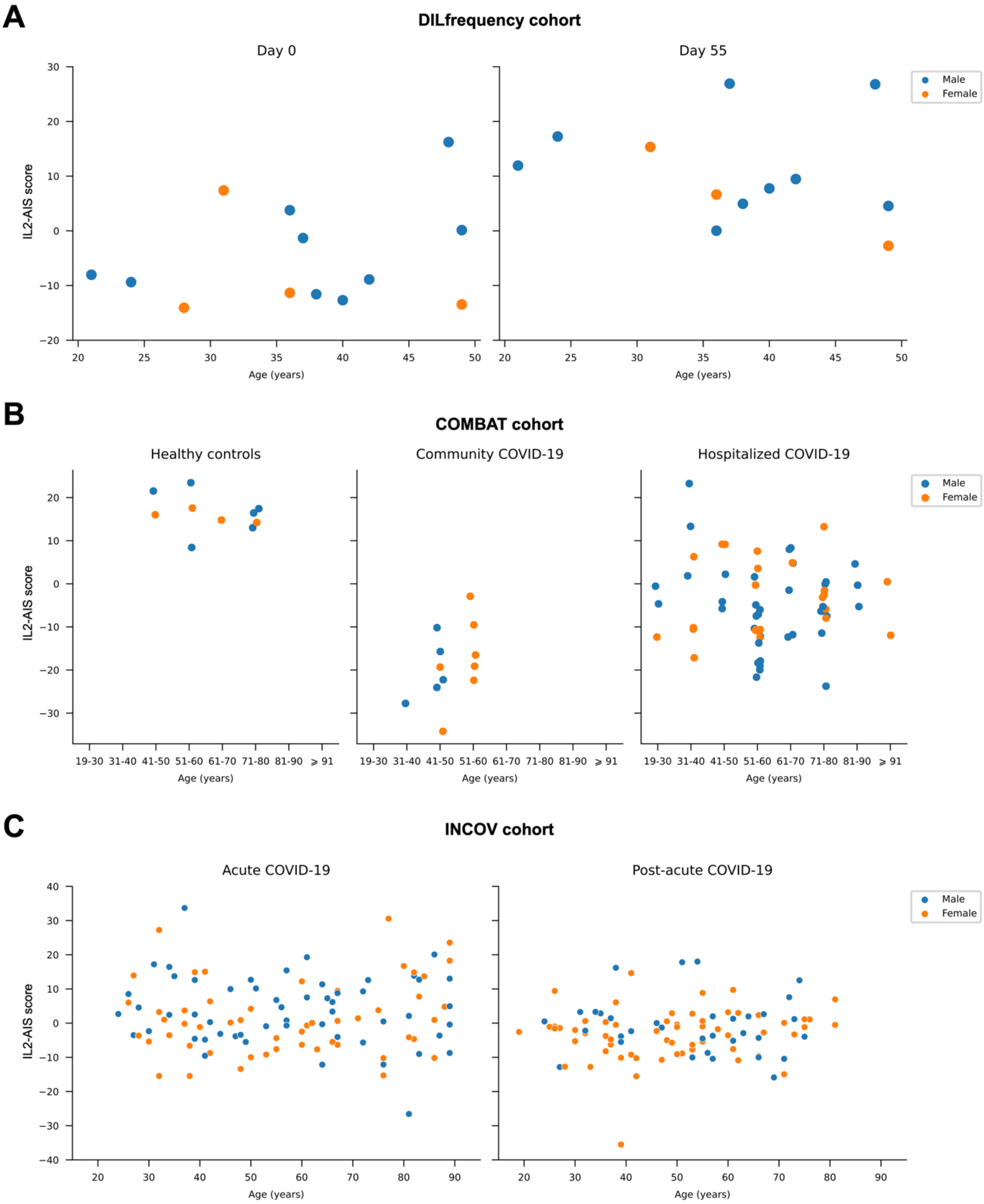
**Appendix Figure 10. Consistent decrease of IL2-AIS score after COVID-19 infection in multiple cell populations.**

**(A-B)** Decrease of IL2-AIS scores after the onset of symptoms in COVID-19 patients from the COMBAT cohort **(A)** or INCOV **(B)** cohorts. Data shown represents the variation of the IL2-AIS scores from the identified NK, mononuclear phagocytes (MNP) and B cell populations in COMBAT and from the CD4$^+$ T, NK, Monocyte and B cell populations in INCOV. Each dot represents a clinical sample, and colours depict the different COVID-19 disease severity groups. In the INCOV cohort, patients are grouped by their worst recorded COVID-19 severity. In **(A)**, dashed black lines represent LOWESS curves.



**Appendix Figure 11. IL2-AIS dynamics in specific cell subsets in the COMBAT cohort.**

Decrease of IL2-AIS scores after the onset of symptoms in COVID-19 patients from the COMBAT cohort, shown in 29 cell subsets for which at least 100 pseudo-bulk samples are available. Each dot represents a pseudo-bulk sample. Colours depict the different COVID-19 disease severity groups. Dashed black lines represent LOWESS curves.



Appendix Figure 12. The IL2-AIS score was not associated with age or sex.

**(A**-**C**) Correlation between IL2-AIS scores, age, and sex in the DILfrequency **(A)**, COMBAT cohort **(B)**, and INCOV **(C)** cohorts. Each dot represents a participant, with colours representing the sex. In the INCOV cohort, IL2-AIS scores were calculated separately from samples collected in the acute (defined as the earliest sample taken 0-14 days post COVID-19 symptoms) or post-acute (defined as the earliest sample taken 29-84 days post COVID-19 symptoms) phases of the disease.

# Appendix tables

| Region | Number | Evidence | Speculation |
|---|---|---|---|
| IGH | 1 | CLR reads | Misalignment. |
| IGH | 2 | CCS reads | Heterozygous deletion. |
| IGH | 3 | CLR reads | Heterozygous duplication; see S8 Fig. |
| IGK | 4 | HV31-V1 assembly | Assembly gap. |
| IGK | 5 | Bionano contigs and CCS reads | Misalignment. |
| IGK | 6 | - | Inconclusive. |
| IGK | 7 | - | Inconclusive. |
| IGK | 8 | - | Inconclusive. |
| IGK | 9 | ONT reads | Heterochromatin microsatellite array; see S12 Fig |
| IGK | 10 | - | Inconclusive. |
| IGL | 11 | Bionano contigs | Misalignment; possible assembly error outside the IGL region. |
| IGL | 12 | Bionano contigs | Heterozygous deletion. |
| HLA | 13 | CCS reads | Heterozygous deletion. |
| HLA | 14 | CCS reads | Assembly error (collapsed duplications). |
| TRB | 15 | Bionano contigs and CCS reads | Heterozygous duplication |
| TRB | 16 | CLR reads | Misalignment. |
| TRB | 17 | HV31-V1 assembly | Assembly gap. |

| | | | |
|---|---|---|---|
| TRB | 18 | Bionano contigs and HV31-V1 assembly | Misalignment due to assembly gap (see number 17). |
| TRG | 19 | HV31-V1 assembly | Assembly gap. |
| KIR | 20 | HV31-V1 assembly | Assembly gap. |

**Appendix Table 1. Potentially problematic regions in the HV31-V1 assembly identified from *k*-mer depths.**

Details of locations where validation *k*-mers show discrepancy from expectation and speculated reasons. Locations are numbered as shown in Appendix Figure 1. Locations 13 and 14 are also shown in Figure 15.

# Bibliography

1. Zhang JY, Roberts H, Flores DSC, et al. Using de novo assembly to identify structural variation of eight complex immune system gene regions. Ratan A, ed. *PLOS Comput Biol*. 2021;17(8):e1009254. doi:10.1371/journal.pcbi.1009254

2. Zhang JY, Roberts H, Flores DS, et al. Using de novo assembly to identify structural variation of complex immune system gene regions. *bioRxiv*. Published online 2021:2021-02.

3. Seelig E, Howlett J, Porter L, et al. The DILfrequency study is an adaptive trial to identify optimal IL-2 dosing in patients with type 1 diabetes. *JCI Insight*. 2018;3(19):e99306. doi:10.1172/jci.insight.99306

4. Zhang JY, Hamey F, Trzupek D, et al. Low-dose IL-2 reduces IL-21+ T cells and induces a long-lived anti-inflammatory gene expression signature inversely modulated in COVID-19 patients. *medRxiv*. Published online 2022:2022-04.

5. Zhang JY, Hamey F, Trzupek D, et al. Low-dose IL-2 reduces IL-21+ T cell frequency and induces anti-inflammatory gene expression in type 1 diabetes. *Nat Commun*. 2022;13(1):7324. doi:10.1038/s41467-022-34162-3

6. Flajnik MF, Kasahara M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat Rev Genet*. 2010;11(1):47-59. doi:10.1038/nrg2703

7. Cruz-Tapias P, Castiblanco J, Anaya JM. Major histocompatibility complex: antigen processing and presentation. In: *Autoimmunity: From Bench to Bedside [Internet]*. El Rosario University Press; 2013.

8. Sticht J, Álvaro-Benito M, Konigorski S. Type 1 diabetes and the HLA region: Genetic association besides classical HLA Class II genes. *Front Genet*. 2021;12:683946.

9. Mathebula EM, Sengupta D, Govind N, et al. A genome-wide association study for rheumatoid arthritis replicates previous HLA and non-HLA associations in a cohort from South Africa. *Hum Mol Genet*. 2022;31(24):4286-4294.

10. Hollenbach JA, Oksenberg JR. The immunogenetics of multiple sclerosis: A comprehensive review. *J Autoimmun*. 2015;64:13-25.

11. Roth DB. V(D)J Recombination: Mechanism, Errors, and Fidelity. *Microbiol Spectr*. 2014;2(6). doi:10.1128/microbiolspec.MDNA3-0041-2014

12. Papavasiliou FN, Schatz DG. Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. *Cell*. 2002;109(2):S35-S44.

13. Campbell KS, Purdy AK. Structure/function of human killer cell immunoglobulin-like receptors: lessons from polymorphisms, evolution, crystal structures and mutations. *Immunology*. 2011;132(3):315-325.

14. Carrington M, Norman P. The KIR gene cluster. *Natl Cent Biotechnol Inf US*. Published online 2003.

15. Shendure J, Balasubramanian S, Church GM, et al. DNA sequencing at 40: past, present and future. *Nature*. 2017;550(7676):345-353. doi:10.1038/nature24286

16. Sanger F, Air GM, Barrell BG, et al. Nucleotide sequence of bacteriophage ϕX174 DNA. *nature*. 1977;265(5596):687-695.

17. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26(10):1135-1145.

18. Karger BL, Guttman A. DNA sequencing by CE. *Electrophoresis*. 2009;30(S1):S196-S202.

19. Panchy N, Lehti-Shiu M, Shiu SH. Evolution of gene duplication in plants. *Plant Physiol*. 2016;171(4):2294-2316.

20. Sharp AJ, Locke DP, McGrath SD, et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet*. 2005;77(1):78-88.

21. Ebbert MTW, Jensen TD, Jansen-West K, et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol*. 2019;20(1):97. doi:10.1186/s13059-019-1707-2

22. Metzker ML. Sequencing technologies — the next generation. *Nat Rev Genet*. 2010;11(1):31-46. doi:10.1038/nrg2626

23. EC SEQ BIOINFORMATICS. Why does the per base sequence quality decrease over the read in Illumina? Accessed May 19, 2023. https://web.archive.org/web/20220621162301/https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina

24. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323(5910):133-138.

25. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. *science*. 2003;299(5607):682-686.

26. Eid et al. - 2009 - Real-Time DNA Sequencing from Single Polymerase Mo.pdf.

27. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*. 2021;39(11):1348-1365. doi:10.1038/s41587-021-01108-x

28. Chen P, Gu J, Brandin E, Kim YR, Wang Q, Branton D. Probing single DNA molecule transport using fabricated nanopores. *Nano Lett*. 2004;4(11):2293-2298.

29. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*. 2020;21(1):1-16.

30. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res*. 2010;38(15):e159-e159. doi:10.1093/nar/gkq543

31. Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;37(10):1155-1162. doi:10.1038/s41587-019-0217-9

32. Sereika M, Kirkegaard RH, Karst SM, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods*. 2022;19(7):823-826. doi:10.1038/s41592-022-01539-7

33. Wang O, Chin R, Cheng X, et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res*. 2019;29(5):798-808.

34. Zhang L, Zhou X, Weng Z, Sidow A. Assessment of human diploid genome assembly with 10x Linked-Reads data. *Gigascience*. 2019;8(11):giz141.

35. Yuan Y, Chung CYL, Chan TF. Advances in optical mapping for genomic research. *Comput Struct Biotechnol J*. 2020;18:2051-2062.

36. 10X Genomics. *An Introduction to Linked-Read Technology for a More Comprehensive Genome and Exome Analysis*.; 2017. Accessed May 19, 2023. https://support.10xgenomics.com/permalink/6pTp2FO3Ic42cuKwwKIUCY

37. Chan S, Lam E, Saghbini M, et al. Structural variation detection and analysis using Bionano optical mapping. *Copy Number Var Methods Protoc*. Published online 2018:193-203.

38. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. *Genome Biol*. 2019;20(1):246. doi:10.1186/s13059-019-1828-7

39. Pang AW, MacDonald JR, Pinto D, et al. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol*. 2010;11:1-14.

40. Caron NS, Wright GE, Hayden MR. Huntington disease. Published online 2020. https://www.ncbi.nlm.nih.gov/books/NBK1305/

41. Gutierrez C, Schiff R. HER2: biology, detection, and clinical implications. *Arch Pathol Lab Med*. 2011;135(1):55-62.

42. PacificBiosciences. pbsv: PacBio structural variant (SV) calling and analysis tools. Github. https://github.com/PacificBiosciences/pbsv

43. Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*. 2018;15(6):461-468. doi:10.1038/s41592-018-0001-7

44. Cretu Stancu M, Van Roosmalen MJ, Renkens I, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun*. 2017;8(1):1326.

45. Sohn J il, Nam JW. The present and future of de novo whole-genome assembly. *Brief Bioinform*. 2018;19(1):23-40.

46. Genome Research Consortium. Announcing GRCh38. GenomeRef. Published December 24, 2013. Accessed May 20, 2023. http://genomeref.blogspot.com/2013/12/announcing-grch38.html

47. Lischer HEL, Shimizu KK. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*. 2017;18(1):474. doi:10.1186/s12859-017-1911-6

48. Xiao W, Wu L, Yavas G, Simonyan V, Ning B, Hong H. Challenges, solutions, and quality metrics of personal genome assembly in advancing precision medicine. *Pharmaceutics*. 2016;8(2):15.

49. Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature*. 2023;617(7960):312-324.

50. De Coster W, Weissensteiner MH, Sedlazeck FJ. Towards population-scale long-read sequencing. *Nat Rev Genet*. 2021;22(9):572-587. doi:10.1038/s41576-021-00367-3

51. Takayama J, Tadaka S, Yano K, et al. Construction and integration of three de novo Japanese human genome assemblies toward a population-specific reference. *Nat Commun*. 2021;12(1):226.

52. Mostovoy Y, Levy-Sakin M, Lam J, et al. A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Methods*. 2016;13(7):587-590. doi:10.1038/nmeth.3865

53. Wang JR, Holt J, McMillan L, Jones CD. FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics*. 2018;19(1):50. doi:10.1186/s12859-018-2051-3

54. Koren S, Rhie A, Walenz BP, et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol*. 2018;36(12):1174-1182. doi:10.1038/nbt.4277

55. Nurk S, Walenz BP, Rhie A, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res*. 2020;30(9):1291-1305. doi:10.1101/gr.263566.120

56. 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(5):722-736. doi:10.1101/gr.215087.116

57. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 2019;37(5):540-546. doi:10.1038/s41587-019-0072-8

58. Bankevich A, Pevzner P. MosaicFlye: Resolving Long Mosaic Repeats Using Long Reads. In: Schwartz R, ed. *Research in Computational Molecular Biology*. Vol 12074. Lecture Notes in Computer Science. Springer International Publishing; 2020:226-228. doi:10.1007/978-3-030-45257-5_16

59. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*. 2020;17(2):155-158.

60. Gordon D, Huddleston J, Chaisson MJP, et al. Long-read sequence assembly of the gorilla genome. *Science*. 2016;352(6281):aae0344-aae0344. doi:10.1126/science.aae0344

61. Chin CS, Peluso P, Sedlazeck FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 2016;13(12):1050-1054. doi:10.1038/nmeth.4035

62. Shafin K, Pesout T, Lorig-Roach R, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol*. 2020;38(9):1044-1053. doi:10.1038/s41587-020-0503-6

63. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. Published online February 1, 2021. doi:10.1038/s41592-020-01056-5

64. Rautiainen M, Nurk S, Walenz BP, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol*. Published online 2023:1-9.

65. Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science*. 2022;376(6588):44-53.

66. Ebert P, Audano PA, Zhu Q, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021;372(6537):eabf7117. doi:10.1126/science.abf7117

67. Kronenberg ZN, Rhie A, Koren S, et al. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nat Commun*. 2021;12(1):1935.

68. Porubsky D. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol*. Published online 2020:13.

69. Kumar S, Chinnusamy V, Mohapatra T. Epigenetics of modified DNA bases: 5-methylcytosine and beyond. *Front Genet*. 2018;9:640.

70. Li Y, Tollefsbol TO. DNA methylation detection: bisulfite genomic sequencing analysis. *Epigenetics Protoc*. Published online 2011:11-21.

71. Yu M, Hon GC, Szulwach KE, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*. 2012;149(6):1368-1380.

72. Flusberg BA, Webster DR, Lee JH, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*. 2010;7(6):461-465. doi:10.1038/nmeth.1459

73. PacBio. *Detecting DNA Base Modifications Using Single Molecule, Real-Time Sequencing*.; 2015. Accessed May 22, 2023. https://www.pacb.com/wp-content/uploads/2015/09/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf

74. Simpson JT, Workman RE, Zuzarte P, David M, Dursi L, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*. 2017;14(4):407-410.

75. Liu Y, Rosikiewicz W, Pan Z, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol*. 2021;22(1):1-33.

76. Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. Andrés-León E, ed. *PLOS ONE*. 2021;16(10):e0257521. doi:10.1371/journal.pone.0257521

77. Lopes I, Altab G, Raina P, De Magalhães JP. Gene size matters: an analysis of gene length in the human genome. *Front Genet*. 2021;12:559998.

78. Glinos DA, Garborcauskas G, Hoffman P, et al. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature*. 2022;608(7922):353-359.

79. Feng S, Xu M, Liu F, Cui C, Zhou B. Reconstruction of the full-length transcriptome atlas using PacBio Iso-Seq provides insight into the alternative splicing in Gossypium australe. *BMC Plant Biol*. 2019;19(1):1-16.

80. Ali A, Thorgaard GH, Salem M. PacBio Iso-Seq improves the rainbow trout genome annotation and identifies alternative splicing associated with economically important phenotypes. *Front Genet*. 2021;12:683408.

81. Shi ZX, Chen ZC, Zhong JY, et al. High-throughput and high-accuracy single-cell RNA isoform analysis using PacBio circular consensus sequencing. *Nat Commun*. 2023;14(1):2631.

82. PacBio. *MAS-Seq for Single-Cell Isoform Sequencing*.; 2023. Accessed May 22, 2023. https://www.pacb.com/wp-content/uploads/Application-note-MAS-Seq-for-single-cell-isoform-sequencing.pdf

83. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat Commun*. 2019;10(1):3359.

84. Garalde DR, Snell EA, Jachimowicz D, et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods*. 2018;15(3):201-206.

85. PacBio. Meet the New Sequel IIe System: Delivering Fast & Affordable HiFi Sequencing. PacBio Blog. Published October 5, 2020. Accessed June 18, 2023. https://www.pacb.com/products-protocols/meet-the-new-sequel-iie-system/

86. PacBio. *Revio System: Reveal More with Accurate Long-Read Sequencing at Scale*.; 2023. https://www.pacb.com/wp-content/uploads/Revio-brochure.pdf

87. Oxford Nanopore. R10.3: the newest nanopore for high accuracy nanopore sequencing – now available in store. Published February 13, 2020. Accessed May 23, 2023. https://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store

88. Sanderson ND, Kapel N, Rodger G, et al. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. *Microb Genomics*. 2023;9(1). doi:10.1099/mgen.0.000910

89. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-D745.

90. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet*. 2009;54(1):15-39.

91. Watson CT, Steinberg KM, Huddleston 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(4):530-546. doi:10.1016/j.ajhg.2013.03.004

92. Watson CT, Steinberg KM, Graves TA, 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(1):24-34. doi:10.1038/gene.2014.56

93. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004;32(suppl_1):D493-D496.

94. Kleiveland CR. Peripheral blood mononuclear cells. *Impact Food Bioact Health Vitro Ex Vivo Models*. Published online 2015:161-167.

95. Alexandre D, Lefranc MP. The human γ/δ+ and α/β+ T cells: a branched pathway of differentiation. *Mol Immunol*. 1992;29(4):447-451.

96. Lefranc MP, Lefranc G. 3 - Synthesis of the T cell receptor chains. In: Lefranc MP, Lefranc G, eds. *The T Cell Receptor FactsBook*. Factsbook. Academic Press; 2001:25-46. doi:https://doi.org/10.1016/B978-012441352-8/50006-0

97. Istace B, Belser C, Aury JM. BiSCoT: improving large eukaryotic genome assemblies with optical maps. *PeerJ*. 2020;8:e10150.

98. Xu M, Guo L, Gu S, et al. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience*. 2020;9(9):giaa094.

99. Walker BJ, Abeel T, Shea T, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Wang J, ed. *PLoS ONE*. 2014;9(11):e112963. doi:10.1371/journal.pone.0112963

100. Miga KH, Koren S, Rhie A, et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*. Published online July 14, 2020. doi:10.1038/s41586-020-2547-7

101. Zook JM, Hansen NF, Olson ND, et al. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol*. 2020;38(11):1347-1355. doi:10.1038/s41587-020-0538-8

102. Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes. *Genome Biol*. 2004;5(2):R12.

103. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*. 2016;32(19):3021-3023. doi:10.1093/bioinformatics/btw369

104. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21(1):245. doi:10.1186/s13059-020-02134-9

105. Vurture GW, Sedlazeck FJ, Nattestad M, et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 2017;33(14):2202-2204.

106. Myers EW. Toward Simplifying and Accurately Formulating Fragment Assembly. *J Comput Biol*. 1995;2(2):275-290. doi:10.1089/cmb.1995.2.275

107. Garg S, Fungtammasan A, Carroll A, et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol*. 2021;39(3):309-312. doi:10.1038/s41587-020-0711-0

108. Nurk S, Rogaev EI, Eichler EE, Miga KH, Phillippy AM. The complete sequence of a human genome [preprint]. Published online 2021.

109. 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(2).

110. Rodriguez OL, Gibson WS, Parks T, et al. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *Front Immunol*. 2020;11.

111. Altemose N, Miga KH, Maggioni M, Willard HF. Genomic Characterization of Large Heterochromatic Gaps in the Human Genome Assembly. Ouzounis CA, ed. *PLoS Comput Biol*. 2014;10(5):e1003628. doi:10.1371/journal.pcbi.1003628

112. Adam Phillippy. The (near) complete sequence of a human genome. Genome Informatics Section. Published September 22, 2020. Accessed May 11, 2021. https://genomeinformatics.github.io/CHM13v1/

113. Dennis MY, Harshman L, Nelson BJ, et al. The evolution and population diversity of human-specific segmental duplications. *Nat Ecol Evol*. 2017;1(3):0069. doi:10.1038/s41559-016-0069

114. Kirsch S, Weiß B, Miner TL, et al. Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome. *Genome Res*. 2005;15(2):195-204.

115. Lyle R, Prandini P, Osoegawa K, et al. Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. *Genome Res*. 2007;17(11):1690-1696.

116. Bankevich A, Bzikadze AV, Kolmogorov M, Antipov D, Pevzner PA. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat Biotechnol*. 2022;40(7):1075-1081. doi:10.1038/s41587-022-01220-6

117. PacBio. *UNDERSTANDING ACCURACY IN DNA SEQUENCING*.; 2022. Accessed May 23, 2023. https://www.pacb.com/wp-content/uploads/Informational-Guide-Understanding-Accuracy-in-DNA-Sequencing.pdf

118. Zhang H, Jain C, Aluru S. A comprehensive evaluation of long read error correction methods. *BMC Genomics*. 2020;21(S6):889. doi:10.1186/s12864-020-07227-0

119. Vollger MR, Dishuck PC, Sorensen M, et al. Long-read sequence and assembly of segmental duplications. *Nat Methods*. 2019;16(1):88-94. doi:10.1038/s41592-018-0236-3

120. Bailey JA. Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly. *Genome Res*. 2001;11(6):1005-1017. doi:10.1101/gr.GR-1871R

121. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018;36(4):338-345. doi:10.1038/nbt.4060

122. Liao X, Li M, Zou Y, Wu FX, Yi-Pan, Wang J. Current challenges and solutions of de novo assembly. *Quant Biol*. 2019;7(2):90-109. doi:10.1007/s40484-019-0166-9

123. Rautiainen M, Marschall T. MBG: Minimizer-based sparse de Bruijn Graph construction. *Bioinformatics*. 2021;37(16):2476-2478.

124. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo* genome assemblies: Fig. 1. *Bioinformatics*. 2015;31(20):3350-3352. doi:10.1093/bioinformatics/btv383

125. Dabbaghie F, Ebler J, Marschall T. BubbleGun: enumerating bubbles and superbubbles in genome graphs. Kelso J, ed. *Bioinformatics*. 2022;38(17):4217-4219. doi:10.1093/bioinformatics/btac448

126. Onodera T, Sadakane K, Shibuya T. Detecting Superbubbles in Assembly Graphs. In: Darling A, Stoye J, eds. *Algorithms in Bioinformatics*. Vol 8126. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013:338-348. doi:10.1007/978-3-642-40453-5_26

127. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;11(1):1432. doi:10.1038/s41467-020-14998-3

128. Ebert P, Audano PA, Zhu Q, et al. De novo assembly of 64 haplotype-resolved human genomes of diverse ancestry and integrated analysis of structural variation. *bioRxiv*. Published online 2020.

129. Cilibrasi R, Van Iersel L, Kelk S, Tromp J. The complexity of the single individual SNP haplotyping problem. *Algorithmica*. 2007;49:13-36.

130. Bonizzoni P, Dondi R, Klau GW, Pirola Y, Pisanti N, Zaccaria S. On the Minimum Error Correction Problem for Haplotype Assembly in Diploid and Polyploid Genomes. *J Comput Biol*. 2016;23(9):718-736. doi:10.1089/cmb.2015.0220

131. Majidian S, Kahaei MH, de Ridder D. Minimum error correction-based haplotype assembly: Considerations for long read data. Singh TR, ed. *PLOS ONE*. 2020;15(6):e0234470. doi:10.1371/journal.pone.0234470

132. Patterson M, Marschall T, Pisanti N, et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J Comput Biol*. 2015;22(6):498-509.

133. Chin CS, Wagner J, Zeng Q, et al. A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat Commun*. 2020;11(1):4794. doi:10.1038/s41467-020-18564-9

134. Bansal V, Bafna V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*. 2008;24(16):i153-i159. doi:10.1093/bioinformatics/btn298

135. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res*. 2017;27(5):801-812. doi:10.1101/gr.213462.116

136. Li Y, Patel H, Lin Y. Kmer2SNP: Reference-Free Heterozygous SNP Calling Using k-mer Frequency Distributions. In: *Variant Calling: Methods and Protocols*. Springer; 2012:257-265.

137. Audano PA, Ravishankar S, Vannberg FO. Mapping-free variant calling using haplotype reconstruction from k-mer frequencies. *Bioinformatics*. 2018;34(10):1659-1665.

138. Rodriguez OL, Silver CA, Shields K, Smith ML, Watson CT. Targeted long-read sequencing facilitates phased diploid assembly and genotyping of the human T cell receptor alpha, delta, and beta loci. *Cell Genomics*. 2022;2(12):100228. doi:10.1016/j.xgen.2022.100228

139. Fan J, Hu J, Xue C, et al. ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genet*. 2020;16(5):e1008786.

140. Sánchez-Ramírez S, Cutter AD. *CompMap: An Allele-Specific Expression Read-Counter Based on Competitive Mapping*. Bioinformatics; 2021. doi:10.1101/2021.02.12.431019

141. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol*. 2015;16(1):195. doi:10.1186/s13059-015-0762-6

142. Johansson T, Yohannes DA, Koskela S, Partanen J, Saavalainen P. HLA RNA sequencing with unique molecular identifiers reveals high allele-specific variability in mRNA expression. *Front Immunol*. 2021;12:629059.

143. Johansson T, Partanen J, Saavalainen P. HLA allele-specific expression: Methods, disease associations, and relevance in hematopoietic stem cell transplantation. *Front Immunol*. 2022;13:1007425. doi:10.3389/fimmu.2022.1007425

144. Gutierrez-Arcelus M, Baglaenko Y, Arora J, et al. Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nat Genet*. 2020;52(3):247-253. doi:10.1038/s41588-020-0579-4

145. Krueger F, Andrews SR. SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Research*. 2016;5.

146. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods*. 2015;12(11):1061-1063. doi:10.1038/nmeth.3582

147. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*. Published online June 5, 2020. doi:10.1038/s41576-020-0236-x

148. Oxford Nanopore Technologies. The power of Q20+ chemistry. The power of Q20+ chemistry. Accessed May 2, 2023. https://nanoporetech.com/q20plus-chemistry

149. Wagner J, Olson ND, Harris L, et al. Benchmarking challenging small variants with linked and long reads. *Cell Genomics*. 2022;2(5):100128. doi:10.1016/j.xgen.2022.100128

150. Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983-987. doi:10.1038/nbt.4235

151. Olson ND, Lund SP, Colman RE, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet*. 2015;6:235.

152. The 1000 Genomes Project Consortium, Gibbs RA, Boerwinkle E, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

153. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet*. 2018;19(5):286-298. doi:10.1038/nrg.2017.115

154. Walker FO. Huntington's disease. *The Lancet*. 2007;369(9557):218-228.

155. PacificBiosciences. TRGT: Tandem Repeat Genotyper. Github. Accessed May 2, 2023. https://github.com/PacificBiosciences/trgt

156. Nusrat S, Harbig T, Gehlenborg N. Tasks, Techniques, and Tools for Genomic Data Visualization. *Comput Graph Forum*. 2019;38(3):781-805. doi:10.1111/cgf.13727

157. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26.

158. Xu W, Zhong Q, Lin D, et al. CoolBox: a flexible toolkit for visual analysis of genomics data. *BMC Bioinformatics*. 2021;22:1-9.

159. L'Yi S, Wang Q, Lekschas F, Gehlenborg N. Gosling: A Grammar-based Toolkit for Scalable and Interactive Genomics Data Visualization. Published online 2021. doi:10.31219/osf.io/6evmb

160. Carver T, Harris SR, Otto TD, Berriman M, Parkhill J, McQuillan JA. BamView: visualizing and interpretation of next-generation sequencing read alignments. *Brief Bioinform*. 2013;14(2):203-212.

161. Hunter JD. Matplotlib: A 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90-95. doi:10.1109/MCSE.2007.55

162. Cruz-Tapias P, Castiblanco J, Anaya JM. HLA association with autoimmune diseases. In: *Autoimmunity: From Bench to Bedside [Internet]*. El Rosario University Press; 2013.

163. Mikocziova I, Greiff V, Sollid LM. Immunoglobulin germline gene variation and its impact on human disease. *Genes Immun*. 2021;22(4):205-217. doi:10.1038/s41435-021-00145-5

164. Sanchez-Mazas A. A review of HLA allele and SNP associations with highly prevalent infectious diseases in human populations. *Swiss Med Wkly*. 2020;150(1516):w20214-w20214.

165. Tian C, Hromatka BS, Kiefer AK, et al. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat Commun*. 2017;8(1):599.

166. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094-3100.

167. Chin CS, Khalak A. *Human Genome Assembly in 100 Minutes*. BioRxiv; 2019. doi:10.1101/705616

168. Istace B, Belser C, Aury JM. *BiSCoT: Improving Large Eukaryotic Genome Assemblies with Optical Maps*. Bioinformatics; 2019. doi:10.1101/674721

169. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27(6):764-770.

170. Giudicelli V, Brochet X, Lefranc MP. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc*. 2011;2011(6):pdb-prot5633.

171. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SG. IPD-IMGT/HLA Database. *Nucleic Acids Res*. 2020;48(D1):D948-D955.

172. Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SG. IPD—the immuno polymorphism database. *Nucleic Acids Res*. 2012;41(D1):D1234-D1240.

173. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res*. 2013;41(W1):W34-W40.

174. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10(1):421.

175. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat Methods*. 2020;17:261-272. doi:10.1038/s41592-019-0686-2

176. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Valencia A, ed. *Bioinformatics*. 2021;37(12):1639-1643. doi:10.1093/bioinformatics/btaa1016

177. Frankish A, Diekhans M, Jungreis I, et al. GENCODE 2021. *Nucleic Acids Res*. 2021;49(D1):D916-D923.

178. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

179. Broad Institute. Piccard. Published 2022. Accessed June 5, 2023. https://broadinstitute.github.io/picard/

180. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):1-10.

181. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842.

182. Vasimuddin Md, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE; 2019:314-324. doi:10.1109/IPDPS.2019.00041

183. PacificBiosciences. pbmm2. Github. Accessed May 2, 2023. https://github.com/PacificBiosciences/pbmm2

184. Hoyt SJ, Storer JM, Hartley GA, et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science*. 2022;376(6588):eabk3112. doi:10.1126/science.abk3112

185. Heger A, Jacobs K. pysam: htslib interface for python. Read the Docs. Accessed May 2, 2023. https://pysam.readthedocs.io/en/latest/

186. Bonfield JK, Marshall J, Danecek P, et al. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience*. 2021;10(2):giab007.

187. Counterr. The power of Q20+ chemistry. Accessed May 2, 2023. https://github.com/dayzerodx/counterr

188. Green A, Hede SM, Patterson CC, et al. Type 1 diabetes in 2017: global estimates of incident and prevalent cases in children and adults. *Diabetologia*. 2021;64:2741-2750.

189. Patterson CC, Harjutsalo V, Rosenbauer J, et al. Trends and cyclical variation in the incidence of childhood type 1 diabetes in 26 European centres in the 25 year period 1989–2013: a multicentre prospective registration study. *Diabetologia*. 2019;62:408-417.

190. Gregory GA, Robinson TIG, Linklater SE, et al. Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study. *Lancet Diabetes Endocrinol*. 2022;10(10):741-760. doi:10.1016/S2213-8587(22)00218-2

191. Care D. Outcomes: Standards of Medical Care in Diabetesd2021. *Diabetes Care*. 2021;44:S53.

192. Livingstone SJ, Levin D, Looker HC, et al. Estimated life expectancy in a Scottish cohort with type 1 diabetes, 2008-2010. *Jama*. 2015;313(1):37-44.

193. Pagliuca FW, Millman JR, Gürtler M, et al. Generation of functional human pancreatic β cells in vitro. *Cell*. 2014;159(2):428-439.

194. Herold KC, Bundy BN, Long SA, et al. An anti-CD3 antibody, teplizumab, in relatives at risk for type 1 diabetes. *N Engl J Med*. 2019;381(7):603-613.

195. Abbas AK, Benoist C, Bluestone JA, et al. Regulatory T cells: recommendations to simplify the nomenclature. *Nat Immunol*. 2013;14(4):307-308.

196. Barbi J, Pardoll D, Pan F. Treg functional stability and its responsiveness to the microenvironment. *Immunol Rev*. 2014;259(1):115-139.

197. Himmel ME, MacDonald KG, Garcia RV, Steiner TS, Levings MK. Helios+ and Helios- cells coexist within the natural FOXP3+ T regulatory cell subset in humans. *J Immunol*. 2013;190(5):2001-2008.

198. Shevyrev D, Tereshchenko V. Treg heterogeneity, function, and homeostasis. *Front Immunol*. 2020;10:3100.

199. Shevach EM. Mechanisms of foxp3+ T regulatory cell-mediated suppression. *Immunity*. 2009;30(5):636-645.

200. Vignali DAA, Collison LW, Workman CJ. How regulatory T cells work. *Nat Rev Immunol*. 2008;8(7):523-532. doi:10.1038/nri2343

201. Bettini M, Bettini ML. Function, failure, and the future potential of Tregs in type 1 diabetes. *Diabetes*. 2021;70(6):1211-1219.

202. Buckner JH. Mechanisms of impaired regulation by CD4+ CD25+ FOXP3+ regulatory T cells in human autoimmune diseases. *Nat Rev Immunol*. 2010;10(12):849-859.

203. Zhou L, He X, Cai P, et al. Induced regulatory T cells suppress Tc1 cells through TGF-β signaling to ameliorate STZ-induced type 1 diabetes mellitus. *Cell Mol Immunol*. 2021;18(3):698-710.

204. Marek-Trzonkowska N, Myśliwiec M, Dobyszuk A, et al. Administration of CD4+ CD25highCD127- regulatory T cells preserves β-cell function in type 1 diabetes in children. *Diabetes Care*. 2012;35(9):1817-1820.

205. Dwyer CJ, Ward NC, Pugliese A, Malek TR. Promoting immune regulation in type 1 diabetes using low-dose interleukin-2. *Curr Diab Rep*. 2016;16:1-10.

206. Raeber ME, Sahin D, Karakus U, Boyman O. A systematic review of interleukin-2-based immunotherapies in clinical trials for cancer and autoimmune diseases. *eBioMedicine*. 2023;90:104539. doi:10.1016/j.ebiom.2023.104539

207. Jiang T, Zhou C, Ren S. Role of IL-2 in cancer immunotherapy. *OncoImmunology*. 2016;5(6):e1163462. doi:10.1080/2162402X.2016.1163462

208. Klatzmann D, Abbas AK. The promise of low-dose interleukin-2 therapy for autoimmune and inflammatory diseases. *Nat Rev Immunol*. 2015;15(5):283-294.

209. Koreth J, Matsuoka K ichi, Kim HT, et al. Interleukin-2 and regulatory T cells in graft-versus-host disease. *N Engl J Med*. 2011;365(22):2055-2066.

210. von Spee-Mayer C, Siegert E, Abdirama D, et al. Low-dose interleukin-2 selectively corrects regulatory T cell defects in patients with systemic lupus erythematosus. *Ann Rheum Dis*. 2016;75(7):1407-1415.

211. Kolios AG, Tsokos GC, Klatzmann D. Interleukin-2 and regulatory T cells in rheumatic diseases. *Nat Rev Rheumatol*. 2021;17(12):749-766.

212. Todd JA, Walker NM, Cooper JD, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet*. 2007;39(7):857-864.

213. Lowe CE, Cooper JD, Brusko T, et al. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat Genet*. 2007;39(9):1074-1082.

214. Garg G, Tyler JR, Yang JH, et al. Type 1 diabetes-associated IL2RA variation lowers IL-2 signaling and contributes to diminished CD4+ CD25+ regulatory T cell function. *J Immunol*. 2012;188(9):4644-4653.

215. Hartemann A, Bensimon G, Payan CA, et al. Low-dose interleukin 2 in patients with type 1 diabetes: a phase 1/2 randomised, double-blind, placebo-controlled trial. *Lancet Diabetes Endocrinol*. 2013;1(4):295-305.

216. Todd JA, Evangelou M, Cutler AJ, et al. Regulatory T Cell Responses in Participants with Type 1 Diabetes after a Single Dose of Interleukin-2: A Non-Randomised, Open Label, Adaptive Dose-Finding Trial. Huizinga TWJ, ed. *PLOS Med*. 2016;13(10):e1002139. doi:10.1371/journal.pmed.1002139

217. Tang X, Huang Y, Lei J, Luo H, Zhu X. The single-cell sequencing: new developments and medical applications. *Cell Biosci*. 2019;9:1-9.

218. Anaparthy N, Ho YJ, Martelotto L, Hammell M, Hicks J. Single-cell applications of next-generation sequencing. *Cold Spring Harb Perspect Med*. 2019;9(10):a026898.

219. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin Transl Med*. 2022;12(3):e694.

220. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. 2018;50(8):1-14.

221. Cao Y, Qiu Y, Tu G, Yang C. Single-cell RNA sequencing in immunology. *Curr Genomics*. 2020;21(8):564-575.

222. Trzupek D, Dunstan M, Cutler AJ, et al. Discovery of CD80 and CD86 as recent activation markers on regulatory T cells by protein-RNA single-cell analysis. *Genome Med*. 2020;12(1):55. doi:10.1186/s13073-020-00756-z

223. Seumois G, Vijayanand P. Single-cell analysis to understand the diversity of immune cell types that drive disease pathogenesis. *J Allergy Clin Immunol*. 2019;144(5):1150-1153.

224. Luo Y, Xu C, Wang B, et al. Single-cell transcriptomic analysis reveals disparate effector differentiation pathways in human Treg compartment. *Nat Commun*. 2021;12(1):3913.

225. Miragaia RJ, Gomes T, Chomka A, et al. Single-cell transcriptomics of regulatory T cells reveals trajectories of tissue adaptation. *Immunity*. 2019;50(2):493-504.

226. Street K, Risso D, Fletcher RB, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*. 2018;19(1):477. doi:10.1186/s12864-018-4772-0

227. Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*. 2019;566(7745):496-502. doi:10.1038/s41586-019-0969-x

228. Pai JA, Satpathy AT. High-throughput and single-cell T cell receptor sequencing technologies. *Nat Methods*. 2021;18(8):881-892.

229. Ahern DJ, Ai Z, Ainsworth M, et al. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell*. 2022;185(5):916-938.e58. doi:10.1016/j.cell.2022.01.012

230. Trzupek D, Lee M, Hamey F, Wicker LS, Todd JA, Ferreira RC. Single-cell multi-omics analysis reveals IFN-driven alterations in T lymphocytes and natural killer cells in systemic lupus erythematosus. *medRxiv*. Published online 2021:2021-04.

231. Tran HTN, Ang KS, Chevrier M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol*. 2020;21:1-32.

232. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031

233. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *bioRxiv*. Published online 2020.

234. Lam AJ, Uday P, Gillies JK, Levings MK. Helios is a marker, not a driver, of human Treg stability. *Eur J Immunol*. 2022;52(1):75-84.

235. Lu J, Liu J, Li L, Lan Y, Liang Y. Cytokines in type 1 diabetes: mechanisms of action and immunotherapeutic targets. *Clin Transl Immunol*. 2020;9(3):e1122.

236. Spolski R, Kashyap M, Robinson C, Yu Z, Leonard WJ. IL-21 signaling is critical for the development of type I diabetes in the NOD mouse. *Proc Natl Acad Sci*. 2008;105(37):14028-14033.

237. Kenefeck R, Wang CJ, Kapadi T, et al. Follicular helper T cell signature in type 1 diabetes. *J Clin Invest*. 2015;125(1):292-303.

238. Ferreira RC, Simons HZ, Thompson WS, et al. IL-21 production by CD4+ effector T cells and frequency of circulating follicular helper T cells are increased in type 1 diabetes patients. *Diabetologia*. 2015;58:781-790.

239. von Herrath M, Bain SC, Bode B, et al. Anti-interleukin-21 antibody and liraglutide for the preservation of β-cell function in adults with recent-onset type 1 diabetes: a randomised, double-blind, placebo-controlled, phase 2 trial. *Lancet Diabetes Endocrinol*. 2021;9(4):212-224.

240. Laurent C, Fazilleau N, Brousset P. A novel subset of T-helper cells: follicular T-helper cells and their markers. *Haematologica*. 2010;95(3):356.

241. Hale JS, Ahmed R. Memory T follicular helper CD4 T cells. *Front Immunol*. 2015;6:16.

242. Ekman I, Ihantola EL, Viisanen T, et al. Circulating CXCR5- PD-1 hi peripheral T helper cells are associated with progression to type 1 diabetes. *Diabetologia*. 2019;62:1681-1688.

243. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. *Nat Biotechnol*. 2019;37(5):547-554.

244. Reid JE, Wernisch L. Pseudotime estimation: deconfounding single cell time series. *Bioinformatics*. 2016;32(19):2973-2980.

245. Zhan Y, Carrington EM, Zhang Y, Heinzel S, Lew AM. Life and death of activated T cells: how are they different from naïve T cells? *Front Immunol*. 2017;8:1809.

246. Ito S, Bollard CM, Carlsten M, et al. Ultra-low dose interleukin-2 promotes immune-modulating function of regulatory T cells and natural killer cells in healthy volunteers. *Mol Ther*. 2014;22(7):1388-1395.

247. Humrich JY, von Spee-Mayer C, Siegert E, et al. Low-dose interleukin-2 therapy in refractory systemic lupus erythematosus: an investigator-initiated, single-centre phase 1 and 2a clinical trial. *Lancet Rheumatol*. 2019;1(1):e44-e54.

248. Di Rosa F, Cossarizza A, Hayday AC. To ki or not to ki: re-evaluating the use and potentials of Ki-67 for T cell analysis. *Front Immunol*. 2021;12:653974.

249. Miller JD, Clabaugh SE, Smith DR, Stevens RB, Wrenshall LE. Interleukin-2 is present in human blood vessels and released in biologically active form by heparanase. *Immunol Cell Biol*. 2012;90(2):159-167.

250. Wrenshall LE, Platt JL, Stevens ET, Wight TN, Miller JD. Propagation and control of T cell responses by heparan sulfate-bound IL-2. *J Immunol*. 2003;170(11):5470-5474.

251. Ariel A, Yavin EJ, Hershkoviz R, et al. IL-2 induces T cell adherence to extracellular matrix: inhibition of adherence and migration by IL-2 peptides generated by leukocyte elastase. *J Immunol*. 1998;161(5):2465-2472.

252. Singh SS, Chauhan SB, Kumar A, et al. Amphiregulin in cellular physiology, health, and disease: Potential use as a biomarker and therapeutic target. *J Cell Physiol*. 2022;237(2):1143-1156.

253. West NR, Hegazy AN, Owens BM, et al. Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor–neutralizing therapy in patients with inflammatory bowel disease. *Nat Med*. 2017;23(5):579-589.

254. Richard AC, Peters JE, Lee JC, et al. Targeted genomic analysis reveals widespread autoimmune disease association with regulatory variants in the TNF superfamily cytokine signalling network. *Genome Med*. 2016;8:1-15.

255. Maña P, Liñares D, Silva DG, et al. LIGHT (TNFSF14/CD258) is a decisive factor for recovery from experimental autoimmune encephalomyelitis. *J Immunol*. 2013;191(1):154-163.

256. Su Y, Yuan D, Chen DG, et al. Multiple early factors anticipate post-acute COVID-19 sequelae. *Cell*. 2022;185(5):881-895.e20. doi:10.1016/j.cell.2022.01.014

257. Bai F, Tomasoni D, Falcinella C, et al. Female gender is associated with long COVID syndrome: a prospective cohort study. *Clin Microbiol Infect*. 2022;28(4):611-e9.

258. Covid C, Team R, Bialek S, et al. Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. *Morb Mortal Wkly Rep*. 2020;69(12):343.

259. Gedda MR, Danaher P, Shao L, et al. Longitudinal transcriptional analysis of peripheral blood leukocytes in COVID-19 convalescent donors. *J Transl Med*. 2022;20(1):587. doi:10.1186/s12967-022-03751-7

260. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607-D613.

261. Kusnadi A, Ramírez-Suástegui C, Fajardo V, et al. Severely ill patients with COVID-19 display impaired exhaustion features in SARS-CoV-2–reactive CD8+ T cells. *Sci Immunol*. 2021;6(55):eabe4782.

262. Boston University Biology. NF-kB Target Genes. Accessed May 18, 2023. https://www.bu.edu/nf-kb/gene-resources/target-genes/

263. Rosenzwajg M, Churlaud G, Mallone R, et al. Low-dose interleukin-2 fosters a dose-dependent regulatory T cell tuned milieu in T1D patients. *J Autoimmun*. 2015;58:48-58. doi:10.1016/j.jaut.2015.01.001

264. Rosenzwajg M, Salet R, Lorenzon R, et al. Low-dose IL-2 in children with recently diagnosed type 1 diabetes: a Phase I/II randomised, double-blind, placebo-controlled, dose-finding study. *Diabetologia*. 2020;63:1808-1821.

265. Whangbo JS, Kim HT, Nikiforow S, et al. Functional analysis of clinical response to low-dose IL-2 in patients with refractory chronic graft-versus-host disease. *Blood Adv*. 2019;3(7):984-994.

266. Whangbo JS, Kim HT, Mirkovic N, et al. Dose-escalated interleukin-2 therapy for refractory chronic graft-versus-host disease in adults and children. *Blood Adv*. 2019;3(17):2550-2561.

267. Bielekova B, Catalfamo M, Reichert-Scrivner S, et al. Regulatory CD56bright natural killer cells mediate immunomodulatory effects of IL-2Rα-targeted therapy (daclizumab) in multiple sclerosis. *Proc Natl Acad Sci*. 2006;103(15):5941-5946.

268. Melsen JE, Lugthart G, Lankester AC, Schilham MW. Human circulating and tissue-resident CD56bright natural killer cell populations. *Front Immunol*. 2016;7:262.

269. Liang K, He J, Wei Y, et al. Sustained low-dose interleukin-2 therapy alleviates pathogenic humoral immunity via elevating the Tfr/Tfh ratio in lupus. *Clin Transl Immunol*. 2021;10(6):e1293.

270. He J, Zhang X, Wei Y, et al. Low-dose interleukin-2 treatment selectively modulates CD4+ T cell subsets in patients with systemic lupus erythematosus. *Nat Med*. 2016;22(9):991-993.

271. Ballesteros-Tato A, León B, Graf BA, et al. Interleukin-2 inhibits germinal center formation by limiting T follicular helper cell differentiation. *Immunity*. 2012;36(5):847-856.

272. Locci M, Wu JE, Arumemi F, et al. Activin A programs the differentiation of human TFH cells. *Nat Immunol*. 2016;17(8):976-984.

273. Ren HM, Lukacher AE, Rahman ZS, Olsen NJ. New developments implicating IL-21 in autoimmune disease. *J Autoimmun*. 2021;122:102689.

274. Crotty S. T follicular helper cell biology: a decade of discovery and diseases. *Immunity*. 2019;50(5):1132-1148.

275. Künkele KP, Wesch D, Oberg HH, Aichinger M, Supper V, Baumann C. Vγ9Vδ2 T cells: can we re-purpose a potent anti-infection mechanism for cancer therapy? *Cells*. 2020;9(4):829.

276. Provine NM, Klenerman P. MAIT cells in health and disease. *Annu Rev Immunol*. 2020;38:203-228.

277. Tao H, Pan Y, Chu S, et al. Differential controls of MAIT cell effector polarization by mTORC1/mTORC2 via integrating cytokine and costimulatory signals. *Nat Commun*. 2021;12(1):2029.

278. Zhou P, Chen J, He J, et al. Low-dose IL-2 therapy invigorates CD8+ T cells for viral control in systemic lupus erythematosus. *PLoS Pathog*. 2021;17(10):e1009858.

279. Sobah ML, Liongue C, Ward AC. SOCS proteins in immunity, inflammatory diseases, and immune-related cancer. *Front Med*. 2021;8:727987.

280. Lee JS, Park S, Jeong HW, et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci Immunol*. 2020;5(49):eabd1554. doi:10.1126/sciimmunol.abd1554

281. Finkel Y, Gluck A, Nachshon A, et al. SARS-CoV-2 uses a multipronged strategy to impede host protein synthesis. *Nature*. 2021;594(7862):240-245.

282. Ravindra NG, Alfajaro MM, Gasque V, et al. Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium identifies target cells, alterations in gene expression, and cell state changes. *PLoS Biol*. 2021;19(3):e3001143.

283. Marcovecchio ML, Wicker LS, Dunger DB, et al. Interleukin-2 Therapy of Autoimmunity in Diabetes (ITAD): a phase 2, multicentre, double-blind, randomized, placebo-controlled trial. *Wellcome Open Res*. 2020;5.

284. Truman LA, Pekalski ML, Kareclas P, et al. Protocol of the adaptive study of IL-2 dose frequency on regulatory T cells in type 1 diabetes (DILfrequency): a mechanistic, non-randomised, repeat dose, open-label, response-adaptive study. *BMJ Open*. 2015;5(12):e009799.

285. Aitchison J. The statistical analysis of compositional data. *J R Stat Soc Ser B Methodol*. 1982;44(2):139-160.

286. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2019;37(1):38-44. doi:10.1038/nbt.4314

287. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp*. 2008;2008(10):P10008.

288. Van den Berge K, Roux de Bézieux H, Street K, et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat Commun*. 2020;11(1):1201. doi:10.1038/s41467-020-14766-3

289. Anders S, Huber W. Differential expression analysis for sequence count data. Published online 2010:12.

290. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

291. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*. 2019;35(12):2084-2092.

292. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016;352(6282):189-196. doi:10.1126/science.aad0501

293. Ge H, Xu K, Ghahramani Z. Turing: A Language for Flexible Probabilistic Inference. In: Storkey A, Perez-Cruz F, eds. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Vol 84. Proceedings of Machine Learning Research. PMLR; 2018:1682-1690. https://proceedings.mlr.press/v84/ge18b.html

294. Malek TR, Castro I. Interleukin-2 receptor signaling: at the interface between tolerance and immunity. *Immunity*. 2010;33(2):153-165.

295. Vella A, Cooper JD, Lowe CE, et al. Localization of a Type 1 Diabetes Locus in the IL2RA/CD25 Region by Use of Tag Single-Nucleotide Polymorphisms. *Am J Hum Genet*. 2005;76(5):773-779. doi:10.1086/429843

296. Rainbow DB, Yang X, Burren O, et al. Epigenetic analysis of regulatory T cells using multiplex bisulfite sequencing. *Eur J Immunol*. 2015;45(11):3200.

297. Ferreira RC, Simons HZ, Thompson WS, et al. Cells with Treg-specific FOXP3 demethylation but low CD25 are prevalent in autoimmunity. *J Autoimmun*. 2017;84:75-86.

298. Dong S, Hiam-Galvez KJ, Mowery CT, et al. The effect of low-dose IL-2 and Treg adoptive cell therapy in patients with type 1 diabetes. *JCI Insight*. 2021;6(18):e147474. doi:10.1172/jci.insight.147474

299. CHM13 draft v1.1 ONT Guppy 3.6.0. https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/alignments/chm13.draft_v1.1.ont_guppy_3.6.0.wm_2.01.pri.bam