

Characterizing genetic variation across sequencing platforms: Applications to intracellular pathogens and immune receptor genotyping

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics
and Natural Sciences at the
Heinrich Heine University
Düsseldorf

presented by

Yomna Gohar
from Cairo

Düsseldorf, Mai 2026

from the Institute of Medical
Microbiology and Hospital Hygiene at the
Heinrich Heine University Düsseldorf

Published by permission of the Faculty of
Mathematics and Natural Sciences at the Heinrich
Heine University Düsseldorf

Supervisor: Herr Prof. Dr. Alexander T. Dilthey

Co-Supervisor: Herr Prof. Dr. Tobias Marschall

Date of the oral examination: 07.09.2026

Statement

I declare under oath that I have produced my thesis independently and without any undue as-sistance by third parties under consideration of the “Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf”.

Düsseldorf, Mai 2026

Yomna Gohar

List of publications

1. Yomna Gohar, Marie Neumann, Lisanna Hülse, Daniel Wind, Julia Mock, Karin Buchholz, Marcel Helle, Ursula R Sorg, Daniel Degrandi, Klaus Pfeffer, Alexander Dilthey
   Intra-Strain Genetic Heterogeneity in *Toxoplasma gondii* ME49: Oxford Nanopore Long-Read Sequencing Reveals Copy Number Variation in the ROP8-ROP2A Locus
   Published in: BMC Genomics 26, Article number: 1094 (2025)

2. Yomna Gohar, Alberto Descalzo Garcia, Katherine M. Kichula, Paul J. Norman, Alexander Dilthey
   KIR*BLOOM: Accurate KIR genotyping using a new copy number-aware integrated genotype likelihood framework
   Submitted to biorxiv.

Abstract

Genetic variation is enriched in structurally complex regions of the genome—tandem repeats, segmental duplications and highly polymorphic gene families—which evolve rapidly through mutation, recombination and copy-number change. Reliable detection of this variation is constrained by sequencing technology, reference representation, and the model used to infer genotypes. This dissertation addresses these constraints in two systems where conventional short-read analysis fails: genomic divergence among long-term laboratory isolates of the parasite *Toxoplasma gondii*, and natural diversity at the human killer-cell immunoglobulin-like receptor (KIR) locus. First, we compared three independently maintained *T. gondii* ME49 isolates using Oxford Nanopore long-read sequencing. Long reads enabled an accurate de novo assembly with substantially improved contiguity (NG50 6.68 Mb versus 1.2 Mb for the current RefSeq) and near-complete chromosome resolution. Pangenome graph-based genotyping against this reference revealed clear intra-strain heterogeneity, including SNVs, indels and structural variants. The most striking was a 13 kb tandem expansion at the virulence-associated ROP8-ROP2A locus that doubled copy number from three to six in two isolates. These results show that laboratory-maintained strains diverge measurably over time and argue for routine genomic monitoring to protect experimental reproducibility. Second, we developed KIR*BLOOM, a copy-number-aware genotyping method for the KIR locus from short-read data, where current tools struggle with near-identical paralogs, extensive copy-number variation and incomplete reference representation. KIR*BLOOM infers gene copy number from read depth, selects alleles from a curated database under that constraint while treating ambiguous mappings as informative, then refines coding sequences by variant calling. Developed on 52 samples from the 1000 Genomes Project with matched haplotype-resolved assemblies and evaluated on a further 45, it achieved 99.85% precision and 99.92% recall for copy-number inference, and 92.73% precision and 92.69% recall at five-digit allele resolution, outperforming T1K, GraphKIR and Geny. Together, both studies show that detecting complex variation depends critically on the choice of reference and algorithm. Long reads matched to an appropriate reference revealed structural differences between laboratory parasite strains, while a copy-number-aware model improved KIR typing from existing short reads—in both cases, the right combination of reference and method, rather than the sequencing data alone, enabled more reliable detection than standard pipelines.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Alexander Dilthey, for his outstanding support throughout this journey. Coming to bioinformatics from a different field, I am deeply grateful that he gave me the opportunity to take on this work, and for the time and patience he offered while I found my footing. His guidance shaped not only this dissertation but also the way I approach problems. Among the many lessons he taught me, one has stayed with me above all: "do the simple thing first." It is advice I will carry with me long after this work is complete.

I am also grateful to the many Professors, colleagues and students whose discussions helped shape this research, in particular Professor Tobias Marschall, Professor Klaus Pfeffer, Professor Johannes Hegemann, and Dr. Daniel Degrandi. My thanks also go to my master's students, Christian and Marie, for their dedication and for all I learned from working alongside them.

I gratefully acknowledge the support of the Jürgen Manchot Foundation, which funded this research.

Finally, my deepest thanks go to my husband, who was my greatest source of support throughout this journey.

Contents

Statement **3**

List of publications **5**

Abstract **7**

Acknowledgements **8**

Contents **9**

1 Introduction 10

2 Background 12

2.1 Common Foundation **12**

2.2 *Toxoplasma gondii* (*T. gondii*) **34**

2.3 Killer-cell Immunoglobulin-like Receptors (KIR) **37**

3 Publications 45

3.1 Publication 1: "Intra-strain genetic heterogeneity in Toxoplasma gondii ME49: Oxford Nanopore long-read sequencing reveals copy number variation in the ROP8-ROP2A locus" **45**

3.2 Supplementary material for Publication 1 ..... **64**

3.3 Publication 2: "KIR*BLOOM: Accurate KIR genotyping using a new copy number-aware integrated genotype likelihood framework" **79**

3.4 Supplementary material for Publication 2 ..... **107**

4 Discussion 115

5 References 118

1 Introduction

Since the discovery of the DNA double helix and its role as the carrier of genetic information, efforts to determine the sequence of DNA have driven major advances in molecular biology. Early sequencing methods, such as the chain-termination approach developed by Frederick Sanger in 1977 [1], were labor-intensive and limited to short stretches of DNA. Over the past decades, sequencing technologies have evolved dramatically, transitioning from low-throughput approaches to high-throughput platforms capable of generating millions to billions of bases in a single run [2, 3]. Today, short-read sequencing technologies, most prominently developed by Illumina, enable rapid and cost-effective genome analysis, while long-read platforms such as Pacific Biosciences [4] and Oxford Nanopore Technologies (ONT) [5] provide the ability to resolve complex genomic regions through reads spanning thousands to tens of thousands of bases.

These technological advances have fundamentally transformed biological research. DNA sequencing now underpins diverse fields, including evolutionary biology, cancer research, infectious disease biology, and immunology. It enabled large-scale analysis of genetic variation across individuals and populations, forming the basis of modern population genetics. In immunogenomics, sequencing provides the means to dissect immune receptor diversity and identify genetic determinants of immune responses. In infectious disease biology, it enables the study of pathogen evolution, transmission dynamics, and the genetic basis of virulence and antimicrobial resistance. As a result, biology has shifted from targeted, low-throughput investigations to a comprehensive, genome-wide and data-driven discipline.

Despite this progress, different sequencing technologies present distinct trade-offs. Short-read sequencing offers high accuracy and scalability. Moreover, the vast majority of existing datasets remain available only as short reads, making it the dominant approach for population-level studies [6,7], but it collapses in regions of high homology, segmental duplication, or extreme polymorphism [8]. Long-read sequencing resolves those structures but has historically carried higher error rates and higher cost per sample. The central challenge in modern genomics is therefore not simply generating more data, but choosing and adapting the right technological framework to accurately detect genetic variation and link it to phenotype, particularly for complex variation that is invisible to standard pipelines.

x

This challenge appears in very different biological systems, but the underlying problem is the same. The work presented in this dissertation addresses it in two complementary contexts that together span pathogen genomics and immune receptor genomics.

First, in parasite genomics, *Toxoplasma gondii* (*T. gondii*) serves as a model organism for studying intracellular parasitism [9, 10]. Although clonal lineages are widely used in the laboratory, phenotypic differences have been reported after prolonged culture or passage, including changes in growth rate, plaque size, differentiation capacity, and the loss of oocyst production in cats [11,12]. The genomic extent of variation underlying these phenotypic changes remains largely unknown, because the earlier studies that documented these phenotypes predated high-throughput sequencing, and later work has relied mainly on short reads that cannot reliably span repetitive regions or resolve large structural variants [13]. A further challenge is comparing isolates in a comparative genomics framework when data come from different Nanopore chemistries, especially older data with higher error rates. Technical differences between chemistries can mimic biological variation and make it difficult to distinguish true mutations from sequencing artifacts, with small Indels in homopolymeric tracts being particularly error-prone [14]. We therefore sequenced three independently maintained ME49 isolates with a shared clonal origin using ONT, and used de novo assembly with graph-based genotyping to integrate variation across chemistries and test for genomic changes among the three isolates.

Second, in human immune receptor genomics, the killer-cell immunoglobulin-like receptor (KIR) region exemplifies the opposite constraint [15, 16]. KIR is one of the most polymorphic and structurally complex loci in the human genome, with extensive gene-content variation, high sequence similarity, and direct relevance for natural killer cell mediated immune responses against infected and transformed cells [17]. Because most clinical and population cohorts have been sequenced with short reads, there is a pressing need for methods that can extract accurate KIR genotypes from existing short-read data, rather than requiring new long-read sequencing for every sample. Current methods remain insufficient, with additional challenges arising from the incomplete representation of KIR alleles in existing databases. In this dissertation, we developed a new method for accurate KIR genotyping directly from available short-read Illumina sequencing data.

Aims of this thesis:

1. To identify and characterize acquired small and large genomic variation in

laboratory strains of *T. gondii* using Oxford Nanopore long-read sequencing.

2. To develop a novel computational method that leverages widely available short-read data for accurate genotyping of gene content and allelic variation in the human KIR region.

Thesis structure:

This cumulative thesis is based on two first-author publications that form the scientific core.

1. Chapter 2: It includes the background necessary to understand the work in the papers.
2. Chapter 3.1: presents the study on *T. gondii*. It describes the long-read sequencing strategy, the detection of small and large structural variants arising during in vitro and in vivo propagation, and the interpretation of this heterogeneity in the context of parasite biology.
3. Chapter 3.3: presents the computational study on the KIR region. It introduces the algorithm, and its validation against state-of-the-art typing methods.
4. Chapter 4: provides a discussion that integrates both projects. It compares the technological trade-offs, discusses the implications for variations in pathogens and for immunogenetic screening in humans, and outlines future directions.

Together, these studies demonstrate that accurate characterization of complex variation requires matching the biological question to the appropriate sequencing framework, and that both experimental long-read approaches and improved short-read inference can close critical gaps left by standard genome analysis.

2 Background

This chapter sets out the biological background and bioinformatic principles that underpin the projects in this dissertation. Section 2.1 provides a common foundation relevant to both topics. Mentions of *Toxoplasma gondii (T. gondii)* and killer-cell immunoglobulin-like receptor (KIR), occur occasionally in this common foundation. Readers seeking a brief overview of both are recommended to begin with Sections 2.2.1 and 2.3.1, which present the topic-specific background for *T. gondii* and KIR respectively.

2.1 Common Foundation

2.1.1 DNA sequencing

DNA sequencing is the determination of the precise order of nucleotides in a DNA

molecule. Sanger sequencing dominated the field for approximately three decades due to its high accuracy (~99.9%) and relatively long read lengths (~400–900 bp) [1,18]. Its limitation was the low throughput and high cost per base [3]. The second-generation sequencing emerged between 2005 and 2007 with massively parallel short-read technologies from 454 Life Sciences/Roche, Solexa/Illumina, and SOLiD. These methods replaced electrophoresis-based approaches with highly parallelized sequencing chemistries, including sequencing-by-synthesis, pyrosequencing, and sequencing-by-ligation, enabling millions of simultaneous reactions [3]. This transition led to a dramatic reduction in the cost of sequencing a human genome, decreasing from millions of dollars per genome to only thousands of dollars by around 2014 [19]. Third-generation sequencing became commercially established with Pacific Biosciences' single-molecule real-time sequencing platform in 2011 and Oxford Nanopore's MinION nanopore platform in 2014, moving to single-molecule sequencing chemistries that initially traded lower raw per-base accuracy for much longer reads, ranging from tens to hundreds of kilobases [3].

2.1.2 Short read sequencing

Due to their low cost and high accuracy, short-read sequencing is still a fundamental aspect of large-scale human genetic studies, because it is commonly used in population-based analyses, and some of the largest available human datasets, such as UK Biobank, were generated using Illumina NovaSeq short–read sequence platforms [6, 7]. In this method, genomic DNA is typically fragmented to ~200–600 base pairs, ligated to sequencing adapters, and clonally amplified on a flow cell via bridge amplification. Fluorescently labeled nucleotides are incorporated one base at a time, imaged, and then cleaved, enabling cyclic base calling [20, 21] (Figure 1). This process generates massive parallel output, with high-throughput platforms such as the Illumina NovaSeq capable of producing up to 20 billion reads per run [2]. Typical outputs are paired-end reads of 150 or 300 base pairs, with a large proportion of bases achieving Q30 quality scores (corresponding to ~99.9% accuracy) and an error profile dominated by substitutions with a very low Indel rate [2, 22].

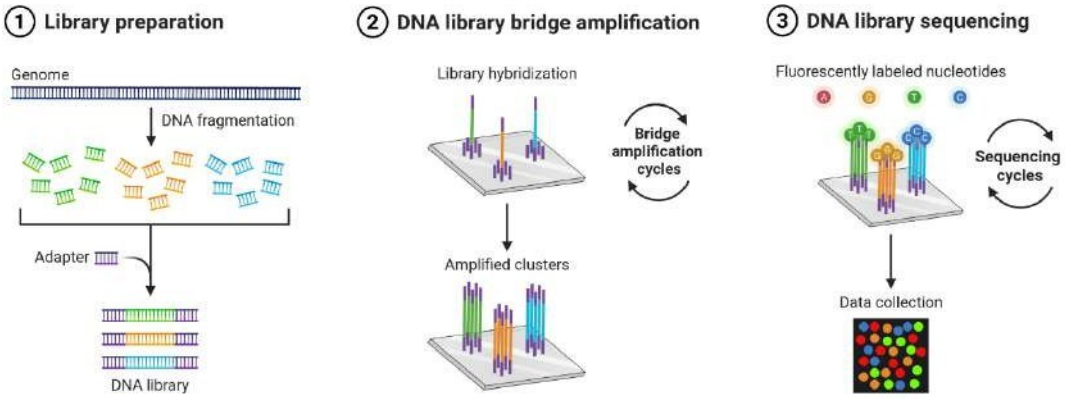


Figure 1. Overview of Illumina short-read sequencing workflow. Genomic DNA is first fragmented and ligated to sequencing adapters to generate a DNA library. Library molecules are then immobilized on a flow cell and clonally amplified by bridge amplification to form clusters. During sequencing-by-synthesis, fluorescently labelled nucleotides are incorporated in repeated cycles, allowing the sequence of each cluster to be recorded (Reprinted from BioRender, based on a BioRender template).

The main limitations of short-read methods arise from their read length. Structural variants larger than the insert size, including deletions, duplications, inversions and translocations, are inferred indirectly from discordant read pairs (pairs mapping in an unexpected way) or split-read alignments (different parts of the same read map to different positions) rather than observed directly as continuous molecules [11–13] (Figure 2, middle panel). Structural variations are discussed in more detail in Section 2.1.6.3. Complex haplotypes cannot be resolved over distances greater than the fragment length [25]. Additionally, because individual reads are shorter than most repetitive elements, they also struggle to span tandem repeats, segmental duplications, homopolymer tracts and other low-complexity regions, leading to ambiguous mapping, and collapsed or misassigned paralogs in these intervals [12, 15]. These constraints limit short-read sequencing to precise characterization of small variants, while leaving large-scale genome architecture incompletely resolved.

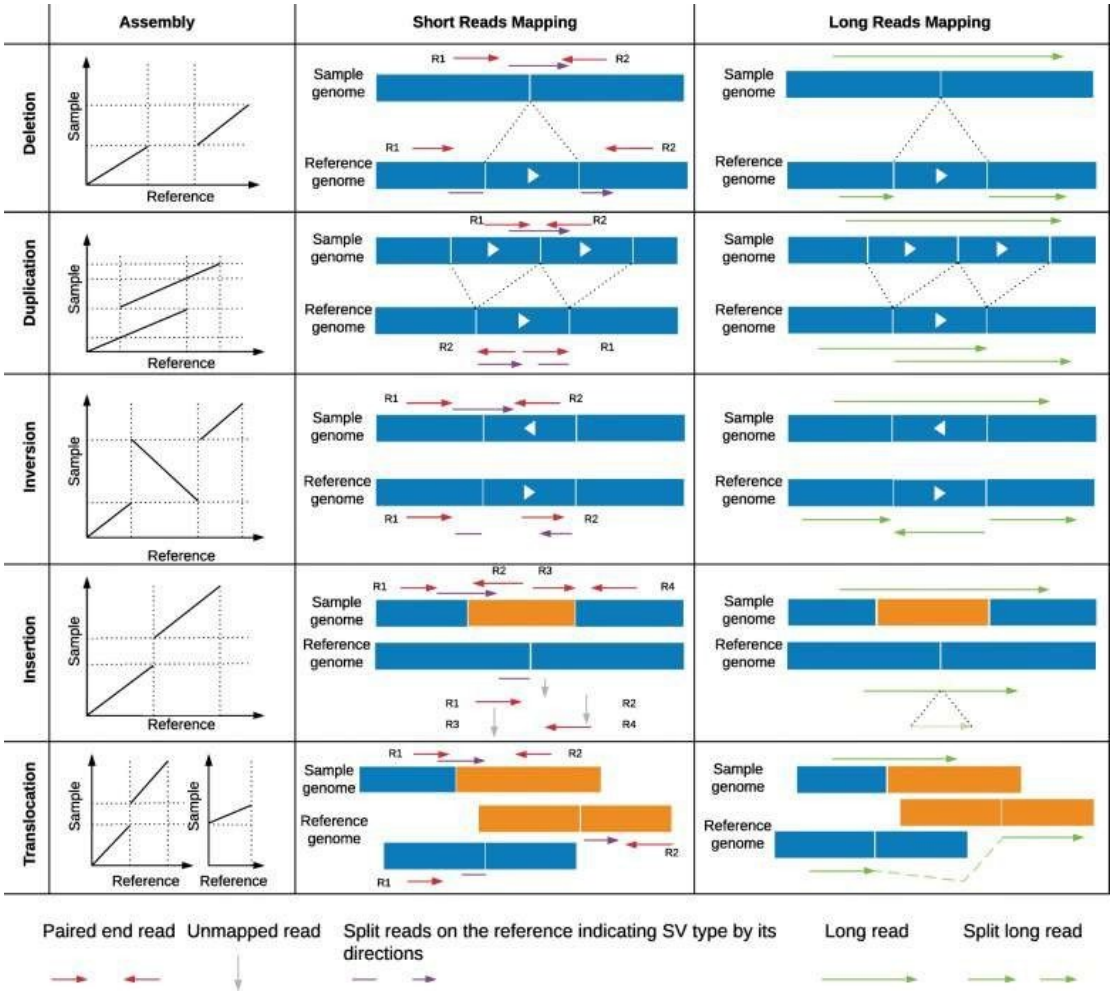


Figure 2. Overview of approaches for detecting structural variants. De novo assembly, short-read mapping, and long-read mapping provide different signals for identifying structural variants. In assembly-based approaches, structural variants can be inferred from changes in the relative order, orientation, or spacing of assembled genomic segments. In short-read mapping, paired-end reads, split reads, and read-depth changes are used to identify the type, size, and genomic position of variants, although repetitive regions can remain difficult to resolve. In long-read mapping, individual reads can span larger variants and repetitive sequences, allowing structural variants to be detected from long-read alignment patterns.
Reproduced from Mahmoud et al., Structural variant calling: the long and the short of it, Genome Biology 20, 246 (2019) [8], under the Creative Commons Attribution 4.0 International License. No changes were made.

2.1.3 Long read sequencing

Long-read sequencing has been advanced to overcome limitations of short reads in structurally complex regions. For example, Pacific Biosciences (PacBio) HiFi sequencing generates reads of approximately 15–25 kilobases with consensus accuracy exceeding 99.8%, achieved through multiple passes of the same molecule, resulting in largely random errors that can be effectively corrected [27] (Figure 3). Oxford Nanopore Technologies (ONT) sequences native DNA by passing single strands through a biological nanopore and inferring sequence from changes in ionic current in

real time [5] (Figure 3). This approach can produce read-length distributions with N50 values, defined as the read length L such that reads of length ≥ L contain at least 50% of all sequenced bases, in the tens to hundreds of kilobases depending on DNA extraction and library preparation; ultra-long ONT protocols have reported N50, values above 100 kb and individual reads approaching megabase scale [28, 29].

Both ONT and PacBio enable direct detection of DNA base modifications without chemical conversion, and they differ in the modification types they capture, such as 5-methylcytosine and 6-methyladenine [30–32]. Because long reads span repeats, structural variants including deletions, duplications, insertions, inversions, and translocations are observed as continuous alignments across breakpoints (Figure 2, left panel) [8,33]. Long reads also resolve tandem repeat arrays, segmental duplications, and other low-complexity regions inaccessible to short reads, allow haplotype phasing over tens of kilobases, and support de novo assembly into highly contiguous contigs without reliance on a reference genome [28, 29].

PacBio HiFi generally provides the highest single-read consensus accuracy, whereas ONT offers very long reads, real-time sequencing, portability, lower entry-cost instruments, and flexible throughput [5, 34]. Accordingly, ONT sequencing was employed in this dissertation to construct a genome assembly for *T. gondii* and enable simultaneous analysis of small and large sequence variation in different isolates (Chapter 3)

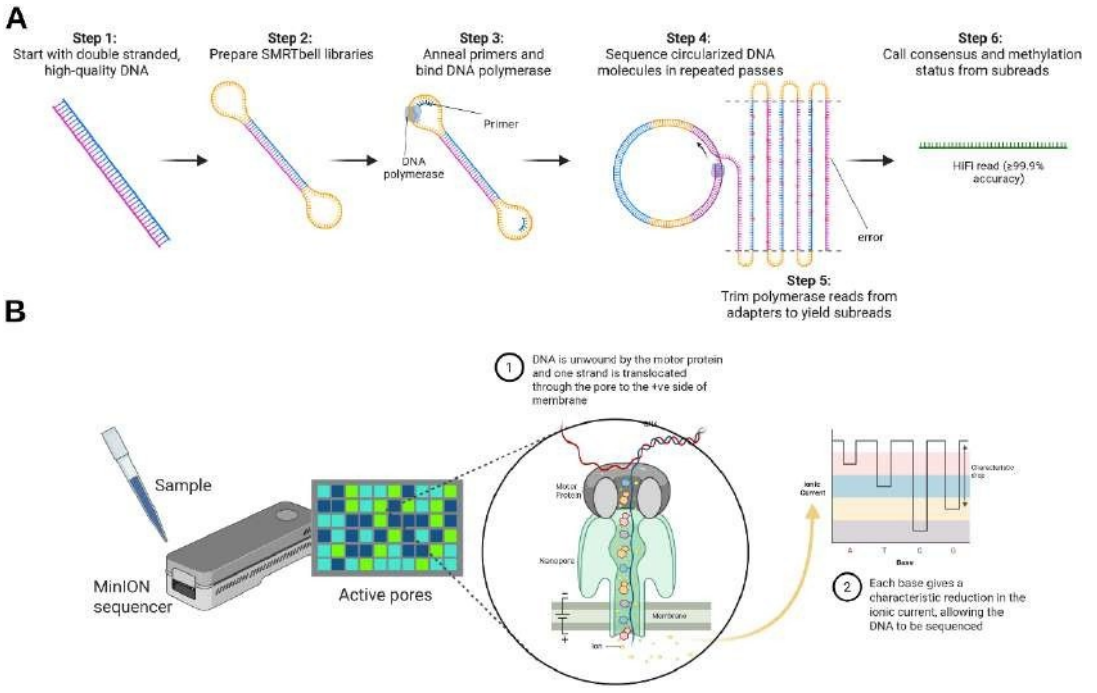


Figure 3. Overview of long-read sequencing technologies.

(A) PacBio HiFi sequencing begins with high-quality double-stranded DNA, which is converted into circular SMRTbell libraries. DNA polymerase repeatedly sequences the circular molecule, generating subreads that are combined into a high-accuracy consensus read. (B) Oxford Nanopore sequencing measures changes in ionic current as DNA is translocated through nanopores. Each nucleotide produces a characteristic signal pattern, allowing the DNA sequence to be inferred directly from the electrical signal. (Adapted from BioRender, based on a BioRender template.
)

2.1.4 Genome assembly

An assembled genome is not a directly observed chromosome, but a constructed model. Formally, assembly is the computational reconstruction of a genome from the millions of short or long fragments generated by a sequencer. Because DNA cannot yet be read continuously from telomere to telomere in a single pass, these fragments must be compared, connected, and merged into contiguous sequences (contigs) and then ordered and oriented into chromosome-scale scaffolds. This process, and the one used in this dissertation, is called de novo assembly, reconstruction without guidance from a pre-existing reference, in which genome contiguity, base accuracy, representation of repeats and haplotype resolution are determined by the sequencing data and the assembly algorithms themselves.

Currently, de novo genome assembly is primarily based on long-read sequencing data. Reads sequenced using PacBio HiFi or ONT platforms are assembled into contigs through graph-based algorithms implemented in assemblers such as hifiasm, Flye, and Verkko [35–37]. Gaps in centromeres, telomeres, and other highly repetitive regions can be closed when accurate long reads are combined with ultra-long reads that traverse large repeat arrays, as demonstrated by recent telomere-to-telomere assemblies [38]. Chromosome-scale scaffolding orders and orients constructed contigs, typically using proximity-ligation data such as Hi-C, to produce chromosome-level assemblies [39, 40]. This is particularly useful for de novo assemblies without a high-quality reference genome, for example in uncharacterized organisms.

Assemblies are evaluated with reference-independent metrics, including NG50 for contiguity, defined as the assembled sequence length L such that sequences of length ≥ L together add up to at least 50% of the expected genome size; consensus QV for base-level accuracy, where QV50 corresponds to approximately one error per 100,000 bases, or 99.999% accuracy; and BUSCO scores for gene-space completeness based on conserved single-copy orthologs [41, 42]. In this dissertation, de novo assembly was essential for the work on *T. gondii,* for which long-read Oxford Nanopore data were generated and assembled with Flye [36]. Flye is a long-read assembler designed for

PacBio and Oxford Nanopore reads. It uses a repeat-graph approach, but instead of constructing the graph directly from all read overlaps, it first assembles error-prone long reads into rough sequences called disjointigs. These disjointigs represent approximate paths through the genome and are then used to construct a repeat graph, in which unique regions and repeated sequences are represented as graph edges and nodes. Flye subsequently maps the original reads back to the graph to resolve repeats where possible, particularly when reads span repeat boundaries. Finally, Flye performs consensus polishing using the input reads to improve base-level accuracy and produce the final contig sequences (Figure 4).

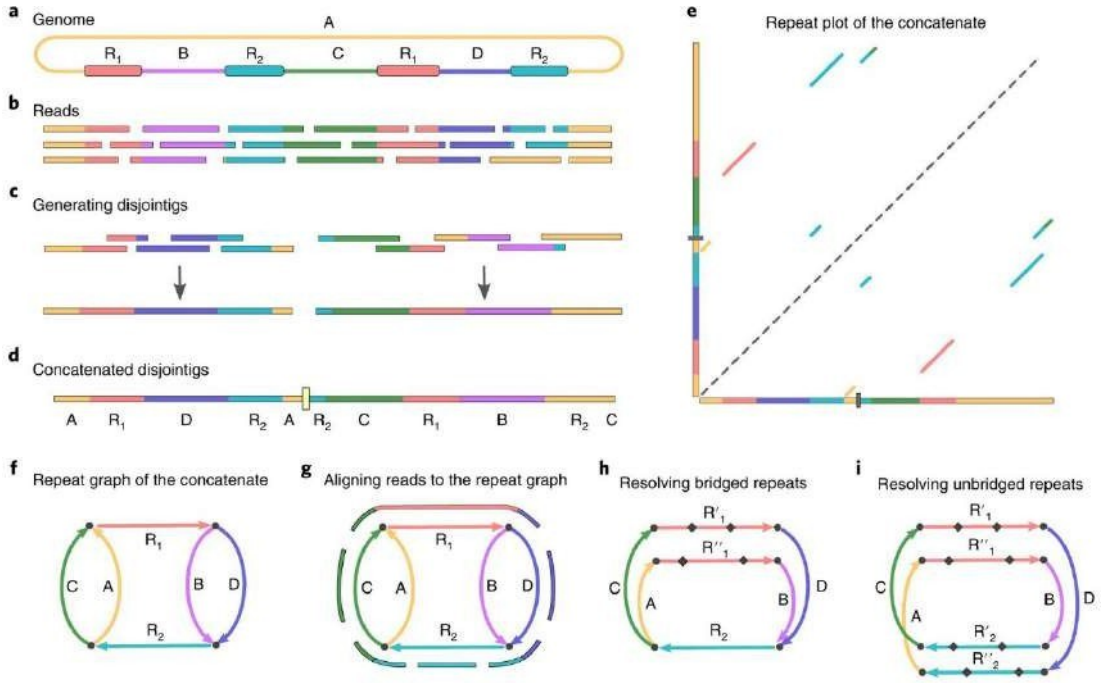


Figure 4. Repeat graph construction in Flye. The schematic illustrates a genome containing two pairs of highly similar repeats, the reads sampled from it, the construction of disjointigs, the generation of a repeat plot, and the conversion of this information into a repeat graph that can be used to resolve bridged and unbridged repeats. Reproduced from Kolmogorov et al., “Assembly of long, error-prone reads using repeat graphs,” Nature Biotechnology, 37, 540–546, 2019 [36], with permission from Springer Nature.

2.1.5 Reference genome

A reference genome provides a single, linear, haploid representation of a species' DNA sequence against which new sequencing data are aligned, compared, and interpreted. It serves as a coordinate system for variant calling, gene annotation, and functional studies, and therefore defines what is considered "reference" versus "alternative" in downstream analyses [43]. For human, the initial draft of the human genome was

produced by the International Human Genome Sequencing Consortium and published in 2001 [44]. It was assembled as a mosaic from DNA derived from multiple anonymous donors and contained more than 150,000 gaps [45]. The reference was later maintained and improved by the Genome Reference Consortium (GRC), an international collaboration founded in 2007 to improve the human, mouse, and zebrafish reference assemblies by closing gaps, correcting errors, and representing complex variation [46]. The current major human reference build, GRCh38, was released by the Genome Reference Consortium in 2013 and remains the standard GRC human reference assembly [47]. Although GRCh38 greatly improved earlier human reference builds, it is a mosaic linear reference derived from multiple individuals, introducing reference bias that can influence variant discovery and downstream analyses and some highly repetitive regions remained incomplete or difficult to represent accurately. For example, the GRC notes that complete and biologically accurate representation of some rDNA repeat clusters was still not possible in GRCh38 with the available resources [48]. A gapless telomere-to-telomere assembly was reported in 2022 for the CHM13 genome by the Telomere-to-Telomere Consortium. This genome is effectively homozygous due to its derivation from a hydatidiform mole, which simplified assembly [38]. However, the T2T-CHM13 reference has limitations: it represents a single genome lacking a Y chromosome and therefore does not capture normal human heterozygosity or population diversity. While the telomere-to-telomere CHM13 assembly addresses many gaps in the reference genome, GRCh38 remains the predominant coordinate system in clinical genomics due to its widespread adoption. The T2T-CHM13 assembly is primarily used as a complementary reference in research settings. In this dissertation, both references were used for read filtering and to assist in KIR genotyping, as described in Chapter 4.

For *T. gondii*, early references were constructed from Sanger sequencing, and not all circulating strains have a de novo assembly. The community has therefore relied heavily on lineage-representative references and ToxoDB resources [49], with ME49 commonly representing Type II strains, GT1 representing Type I, and VEG representing Type III [40,41]. More recently, long-read sequencing has been used to improve *T. gondii* genome assemblies and revise genome structure [42, 43]. Re-evaluation of apicomplexan genomes with long reads has uncovered previously hidden assembly errors, and karyotype differences, demonstrating that assembly quality directly changes biological interpretation. Xia et al., for example, used Oxford

Nanopore sequencing to generate near-complete assemblies for RH, ME49, and CTG [52]. Their work consolidated the *T. gondii* karyotype to 13 chromosomes instead of 14 and closed many repeat-associated gaps. Together, these studies show that reference quality is not only a technical issue, but also a determinant of how genome organization, structural variation, and strain-specific biology are interpreted in *T. gondii.*

2.1.6 Genetic variation

Genetic variation refers to differences in DNA sequences among individuals, strains, or populations, including single-nucleotide differences, insertions, deletions, duplications, copy-number changes, and rearrangements. Such variation arises from mutations and provides the substrate for evolution and for investigating biological differences among individuals or populations. Mutations can arise spontaneously during DNA replication or through DNA damage and repair processes. [53–55] Mutations are commonly classified according to how they arise, the cell lineage in which they occur, their genomic context, and their molecular size. By origin, variants are classified as inherited, having been present in the parental germline and transmitted to offspring, or as de novo, having arisen anew during gametogenesis or early development and absent from the parents [56,57]. By cellular context, mutations are designated germline when they occur in the lineage that gives rise to gametes and can therefore be passed to subsequent generations, or somatic when they arise in other tissues and are typically confined to the individual. Because a de novo event may occur in either compartment, the two classifications overlap: all inherited variants are by definition germline-derived, whereas de novo variants may be germline and potentially heritable or somatic and non-heritable in organisms with an early-segregated germline. In organisms lacking a distinct germline-soma architecture, this distinction is absent and newly arising mutations can be transmitted directly to descendant cells. Independent of origin and cellular-context, genomic location distinguishes non-coding changes in introns, untranslated regions, or intergenic DNA, which can influence regulation and splicing, from coding changes within exons that may alter the amino acid sequence [58]. Size provides a further classification that is driven primarily by detection methodology, and defines the principal variant classes encountered in genome analyses: single-nucleotide variants, small insertions and deletions, and large-scale structural variants such as large insertions, deletions, inversions, duplications, and translocations.

2.1.6.1 Single nucleotide polymorphisms (SNP)

SNPs are substitutions affecting a single base and represent the most abundant form of

variation in genomes [59]. Spontaneous base-substitution mutations can arise during DNA replication when DNA polymerases misincorporate a nucleotide, forming a non-canonical base pair that escapes proofreading. Misincorporation is promoted by transient mispairing, most notably by the rare tautomeric forms of bases first proposed by Watson and Crick, who suggested that a shift from the common keto form to the rarer imino or enol tautomer (Figure 5) allows a mismatched base to adopt a Watson-Crick-like (WC-like) geometry and therefore be accommodated in the polymerase active site [53, 60–62]. Later studies also showed that mismatches can also adopt non-Watson–Crick conformations, such as G·T wobble base pairs, in which the bases are displaced relative to canonical Watson–Crick pairs but remain accommodated within the DNA helix (Figure 5).

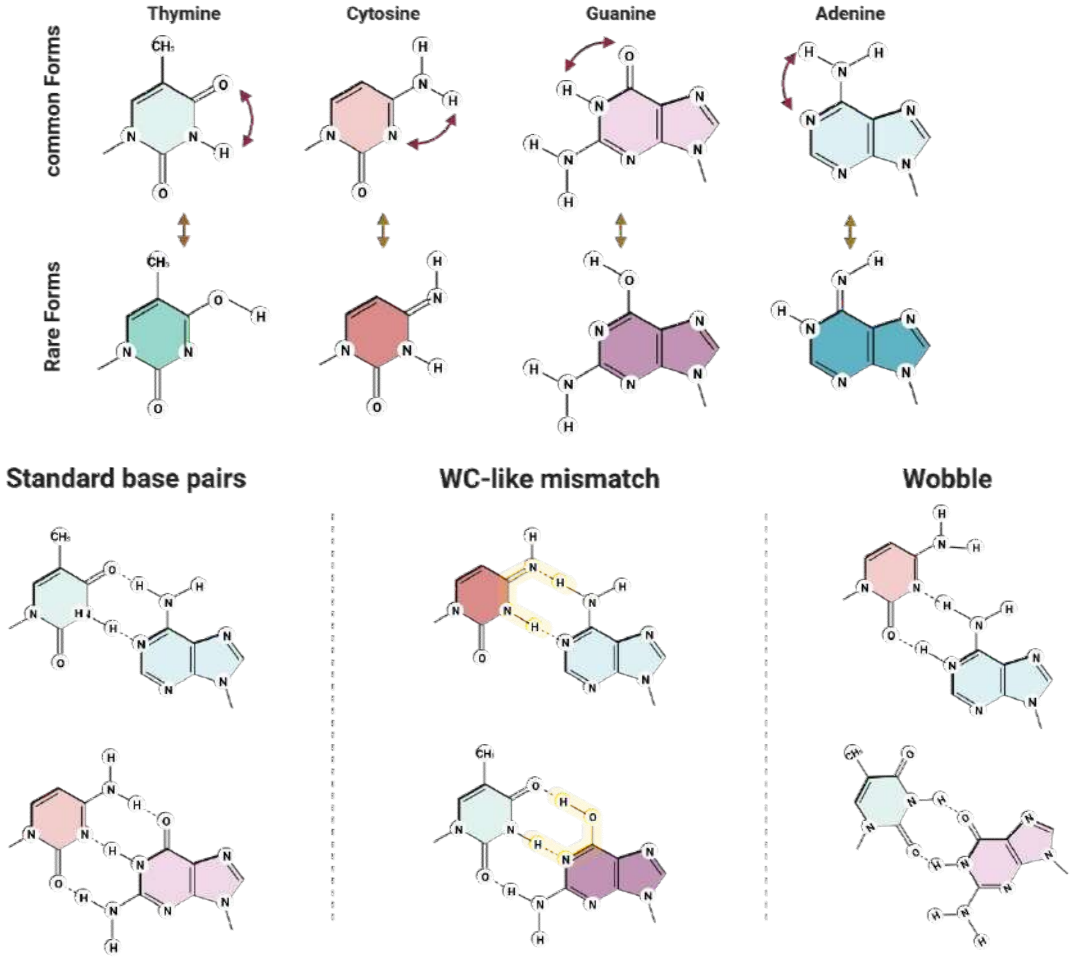


Figure 5. Rare base forms and non-canonical DNA base pairing.
DNA bases predominantly exist in common tautomeric forms but can transiently shift to rare tautomeric states through proton rearrangements. These rare tautomers alter hydrogen-bond donor and acceptor patterns, allowing mismatched bases to form Watson–Crick-like geometries, such as G–T or A–C mismatches. In contrast, wobble base pairs do not necessarily require tautomeric shifts; they can arise from alternative base-pair conformations or changes in protonation state

influenced by the local chemical environment, such as acidity. Such non-canonical base-pairing events can contribute to replication errors if they escape repair. (Created in BioRender)

When SNPs occur in coding regions, the changes are categorized by their translational impact into synonymous, missense and nonsense. Synonymous or silent variants do not change the encoded amino acid. Missense variants replace one amino acid with another, potentially altering protein structure or function. Nonsense variants create a premature stop codon, resulting in premature termination of translation and production of a truncated protein [63].

2.1.6.2 Small insertions and deletions (Indel)

The second most common class of small-scale variation, in humans, are small insertions and deletions, collectively termed Indels [64]. They are defined as insertions or deletions of fewer than 50 base pairs relative to the reference sequence, with insertions adding nucleotides absent from the reference and deletions removing nucleotides present in it. Indels can arise through a process known as strand slippage [53], which occurs frequently in homopolymeric or repetitive regions when either the template strand or the newly synthesized strand transiently loops out during DNA replication as illustrated in Figure 6.

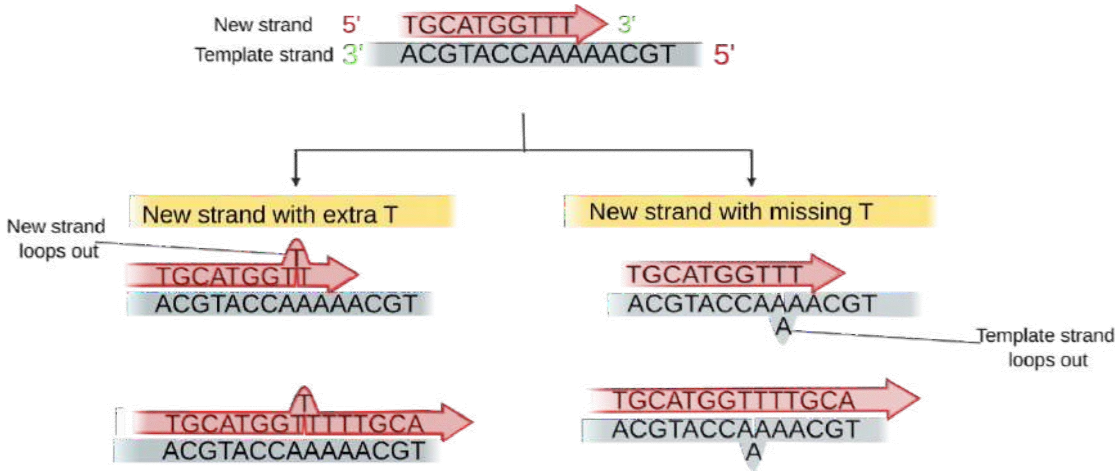


Figure 6. Replication slippage in a homopolymer tract.
During DNA replication, repetitive sequences can cause transient misalignment between the template and newly synthesized strand. If the new strand loops out, DNA polymerase may insert an extra base, resulting in an insertion. If the template strand loops out, the polymerase may skip a base, resulting in a deletion. These insertion–deletion errors are common in homopolymer and other repetitive regions.

In coding regions, the consequences of indels depend on size. Insertions or deletions that are not multiples of three nucleotides cause a frameshift, which alters the downstream reading frame and often result in premature termination or loss of protein

function. In-frame indels, which occur in multiples of three, maintain the reading frame while adding or removing amino acids.

2.1.6.3 Structural variations

Large-scale differences are classified as structural variations and are generally defined as genomic alterations of at least 50 base pairs [13, 56]. The five main classes are deletions, insertions, duplications, inversions, and translocations. Deletions can range from a few bases to entire genes or chromosomal segments. Insertions may involve novel sequence or the integration of mobile genetic elements. Duplications occur when a segment of DNA is present in more than one copy within the genome. They can occur as tandem duplications, where the copies are located adjacent to each other, or as interspersed duplications, where the duplicated segment is inserted at a different genomic location. Inversions arise when a chromosomal segment breaks at two points and is rejoined in the opposite orientation, with sizes ranging from kilobases to megabases, and are frequently located in repetitive regions that complicate detection. Translocations describe the exchange of genetic material between non-homologous chromosomes or between distant loci on the same chromosome. A depiction of different types of SV is presented in Figure 7.

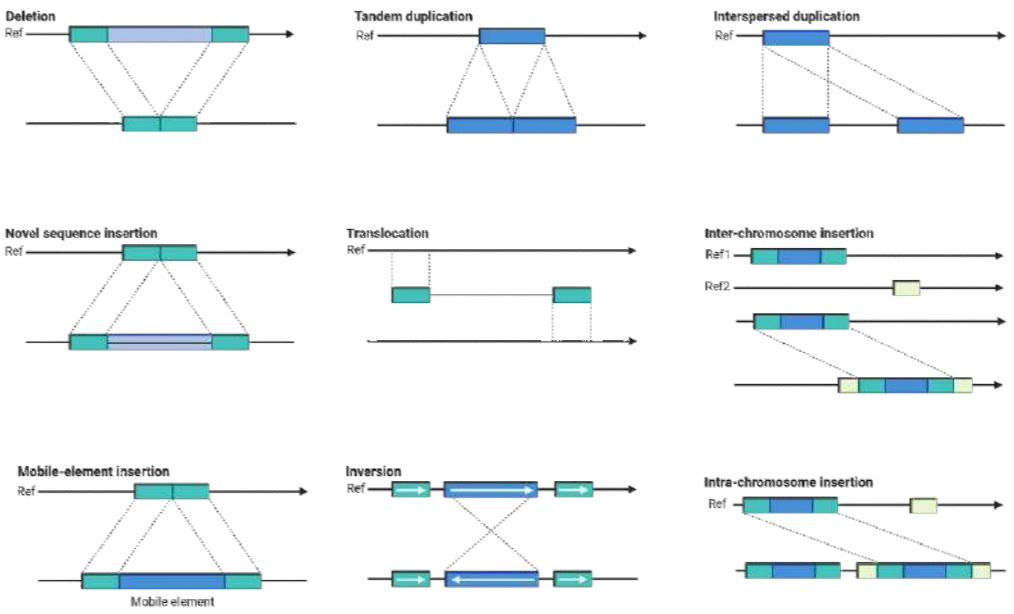


Figure 7. Depiction of major classes of structural variation. (Reprinted from BioRender, based on a BioRender template.)

Structural variants arise through several broad categories of mechanisms, among the

best characterized of which are: replication-based mechanisms, including Fork Stalling and Template Switching (FoSTeS), Microhomology-Mediated Break-Induced Replication (MMBIR), and Break-Induced Replication (BIR); repair-based mechanisms, including Non-Homologous End Joining (NHEJ) and Microhomology-Mediated End Joining (MMEJ); and recombination-based mechanisms, including Non-Allelic Homologous Recombination (NAHR).

Replication-based SVs stem from problems during DNA replication, notably fork stalling and collapse. During replication, the replisome unwinds DNA and polymerases synthesize the second copy. The fork stalls when it meets obstacles such as secondary structures from repeats, DNA lesions, transcription-replication conflicts, or nucleotide depletion [66]. One rescue pathway is Fork Stalling and Template Switching, FoSTeS. When the fork stalls, the lagging strand can disengage and, via microhomology, switch to a nearby template on the same chromosome or sister chromatid and resume synthesis. This template switching can occur multiple times in series, with the lagging strand repeatedly disengaging and switching to successive nearby templates. The final synthesized product is therefore a mosaic of sequences copied from different templates, stitched together at microhomology junctions. Depending on the relative positions of the templates involved, this process can generate deletions, duplications, or complex rearrangements [65, 67] (Figure 8, left panel).

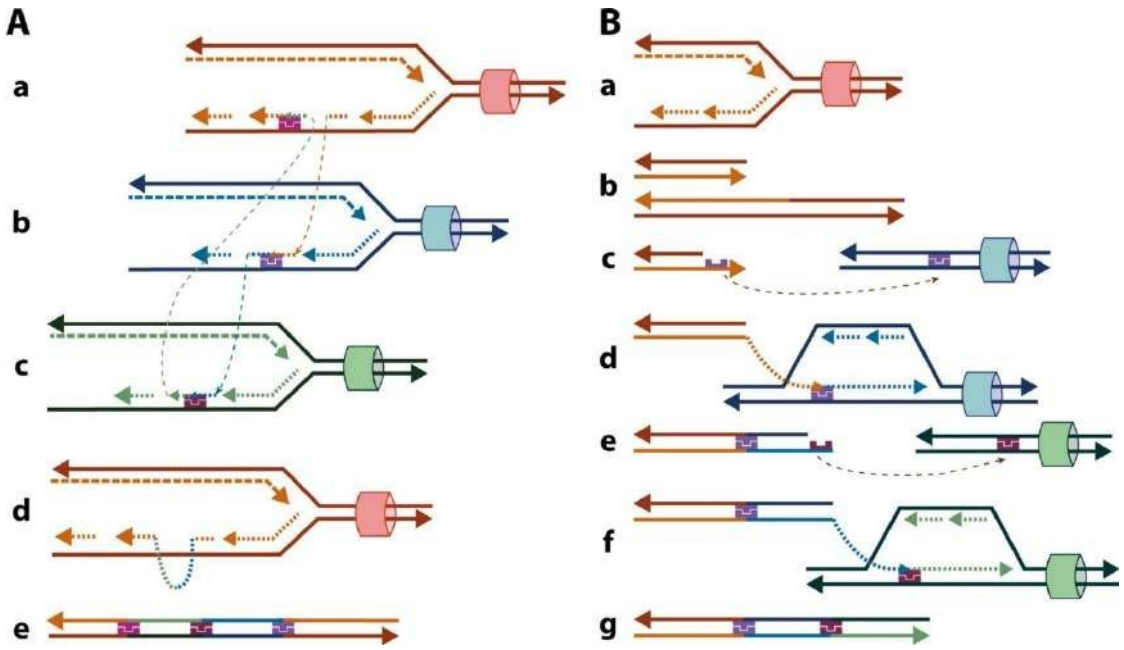


Figure 8. Replication-based mechanisms of complex structural rearrangement.
(A) In fork stalling and template switching (FoSTeS), replication fork stalling allows the nascent lagging strand to disengage from its original template. Through short regions of microhomology, the nascent strand can anneal to a different template at another active replication fork and restart

DNA synthesis. Repeated strand disengagement and template invasion can join segments from different genomic regions into a single rearranged product. (B) In microhomology-mediated break-induced replication (MMBIR), replication fork collapse generates a single-ended double-strand break. After 5′–3′ resection, the exposed 3′ overhang can anneal to microhomologous sequence at another genomic location and initiate DNA synthesis. Repeated template switching can generate complex rearrangements containing segments copied from multiple regions. Figure reproduced from Burssed et al., 2022, [65] under the Creative Commons Attribution 4.0 International License. No changes were made.

A closely related mechanism is Microhomology-Mediated Break-Induced Replication (MMBIR), which shares similarities with FoSTeS but differs fundamentally in its initiation. While FoSTeS is initiated by a transiently stalled fork, MMBIR begins when a replication fork stalls and subsequently collapses. Fork collapse occurs when the replication machinery encounters a lesion it cannot bypass, causing the fork structure to break down and endonuclease cleavage to generate a one-ended double-strand break. This one-ended break is then resected by nucleases including MRE11, EXO1, and DNA2, producing a 3' single-stranded overhang. This 3' overhang subsequently anneals to a nearby region of microhomology on the same chromosome or sister chromatid and primes new DNA synthesis. As in FoSTeS, this process can involve multiple template switches in series, with the invading strand repeatedly disengaging and annealing to successive microhomologous regions, ultimately generating complex genomic rearrangements including deletions, and duplications at breakpoints characterized by short microhomology sequences [65,67,68] (Figure 8, right panel).

The third replication-associated mechanism is Break-Induced Replication (BIR), which like MMBIR is initiated by a one-ended double-strand break arising from replication fork collapse. Following resection of the broken end to generate a 3' single-stranded overhang, BIR diverges fundamentally from MMBIR in how it proceeds. Rather than annealing to a nearby template via short microhomology sequences and switching templates repeatedly, the 3' overhang in BIR invades a double-stranded template, typically the sister chromatid, using extensive sequence homology of hundreds of base pairs. This invasion displaces one strand of the duplex to form a stable D-loop structure, from which replication continues all the way to the end of the chromosome without disengaging from the template. Because synthesis proceeds entirely on the sister chromatid template from the point of invasion to the chromosome end, the original sequence information in that region is replaced, generating loss of heterozygosity over large genomic intervals [69].

Non-Homologous End Joining (NHEJ) is a repair-based mechanism that operates on

two-ended double-strand breaks, which arise when both strands of the DNA duplex are severed simultaneously by exogenous sources of damage such as ionizing radiation, chemotherapeutic agents, or reactive oxygen species. Unlike the replication-based mechanisms discussed above, NHEJ does not use sequence homology to guide repair. Instead, the two broken ends are recognized and processed by a coordinated set of polymerase, nuclease, and ligase complexes that can act on the broken ends in any order, independently at each end, with multiple rounds of nucleotide resection and addition possible before final ligation joins the ends together. Because this end processing is flexible and iterative rather than precise, NHEJ can generate a spectrum of outcomes ranging from small insertions or deletions at the junction to large-scale structural variants including inversions, deletions, and translocations, the latter arising when two spatially proximate double-strand breaks occur simultaneously and their ends are erroneously joined together across different genomic locations [61, 62] (Figure 9).

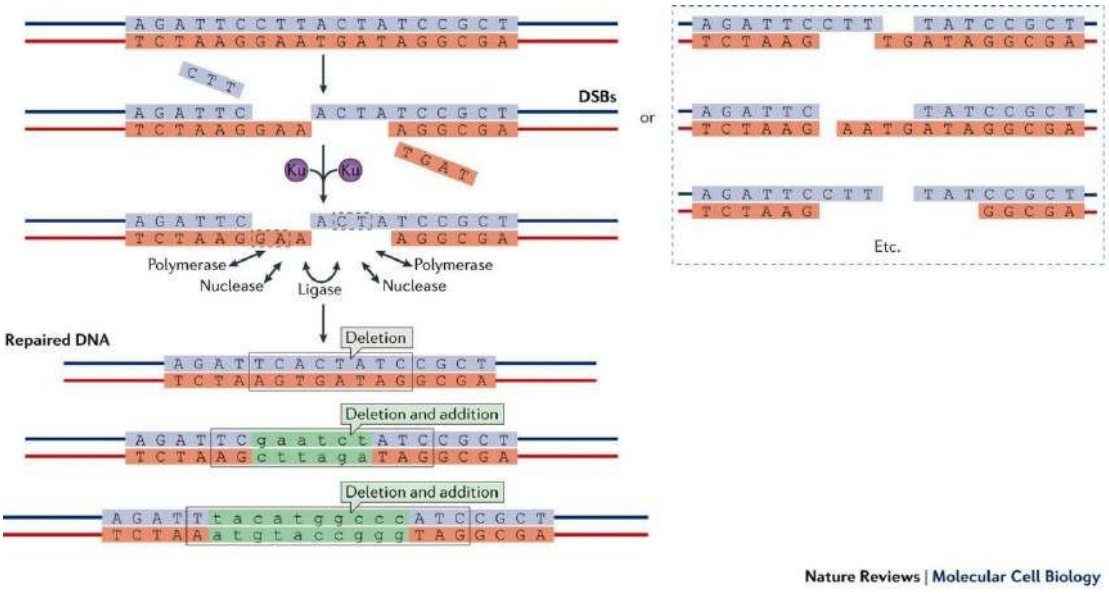


Figure 9. Non-homologous end joining repair of DNA double-strand breaks.
DNA double-strand breaks are recognized by Ku, which recruits NHEJ-associated polymerase, nuclease, and ligase activities. During repair, DNA ends can undergo nuclease-mediated resection and polymerase-mediated nucleotide addition before ligation. Short regions of microhomology may guide end joining, either when already present at the broken ends or generated by nucleotide addition. Because end processing can occur in multiple ways, NHEJ is error-prone and can produce different repair junctions, including deletions and insertions. Added nucleotides are shown in green lowercase. Reproduced from Chang et al., 2017 [71]; licence obtained from Springer Nature. No changes were made

A related pathway is Microhomology-Mediated End Joining (MMEJ), which also repairs two-ended double-strand breaks but differs from NHEJ in that the broken ends undergo more extensive resection to expose single-stranded overhangs that anneal via short regions of microhomology, typically 2 to 25 base pairs, before ligation. This

produces a characteristic signature of deletions flanked by microhomology at the breakpoint junctions, generally larger than those produced by NHEJ [63, 64] (Figure 10)

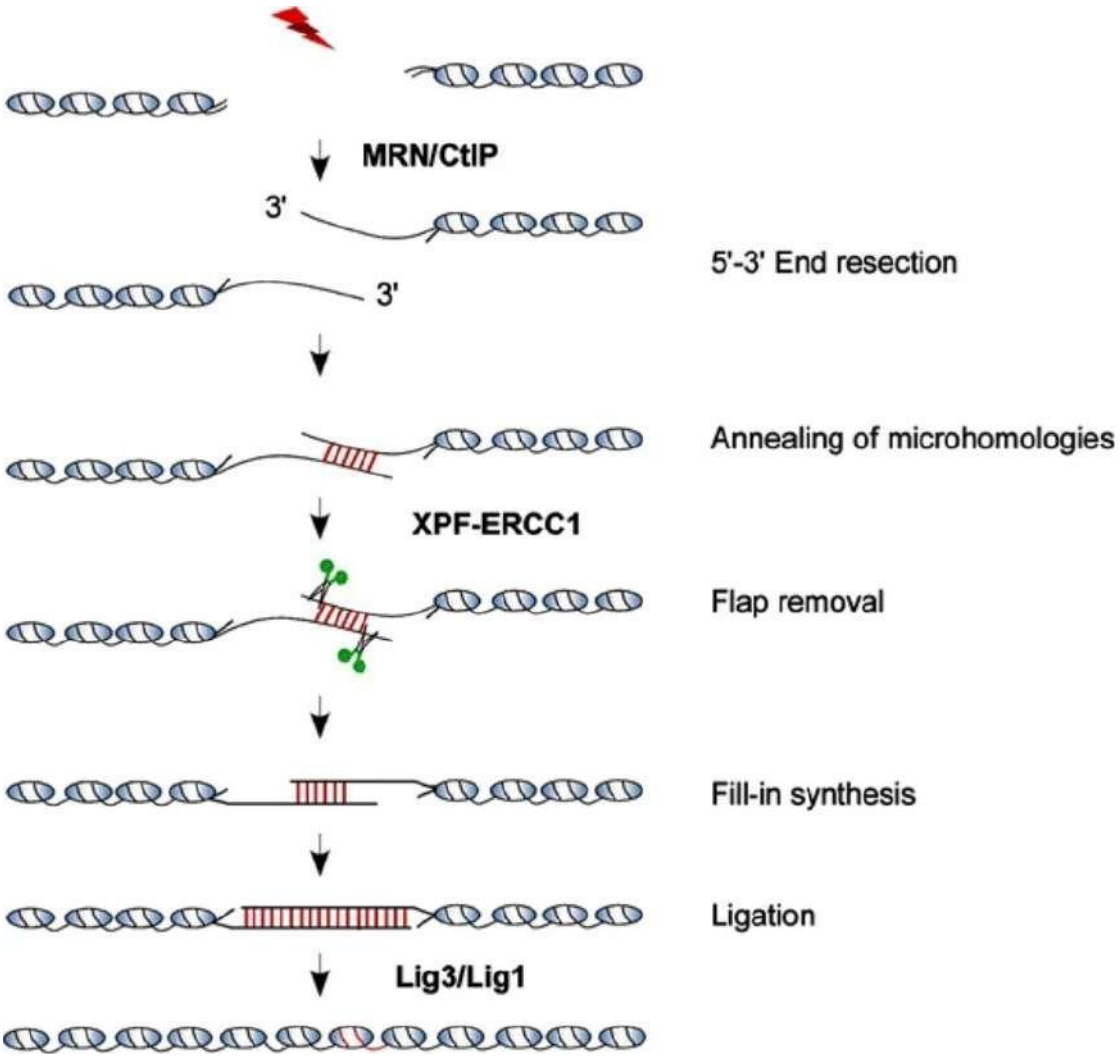


Figure 10. Model of microhomology-mediated end joining repair of DNA double-strand breaks. Following a double-strand break, 5′–3′ end resection exposes short single-stranded microhomologous regions. These microhomologies anneal to each other, forming an intermediate with 3′ flaps and gaps. The non-homologous flaps are removed, the gaps are filled by DNA synthesis, and the repaired DNA is sealed by ligation. Because sequence between the annealed microhomologies is lost, MMEJ is an error-prone repair pathway. Reproduced from Wang and Xu, Microhomology-mediated end joining: new players join the team, Cell & Bioscience 7, 6 (2017) [73], under the Creative Commons Attribution 4.0 International License. No changes were made

Non-Allelic Homologous Recombination (NAHR) is a recombination-based mechanism that generates structural variants when the recombination machinery misaligns paralogous sequences during meiosis. Normal homologous recombination relies on extensive sequence homology to pair allelic positions on homologous chromosomes and facilitate crossing over. However, in regions containing highly

similar duplicated sequences arranged in tandem, such as the KIR locus, the recombination machinery can mistake one paralog for another and align non-allelic positions. Once misaligned, crossing over proceeds normally, but the resulting products are unequal: one chromosome gains extra sequence, producing a duplication, while the other loses sequence, producing a deletion. If the crossover point falls within a gene rather than between genes, the recombination joins the beginning of one gene to the end of a different gene, generating a fusion gene [56, 65] (Figure 11).

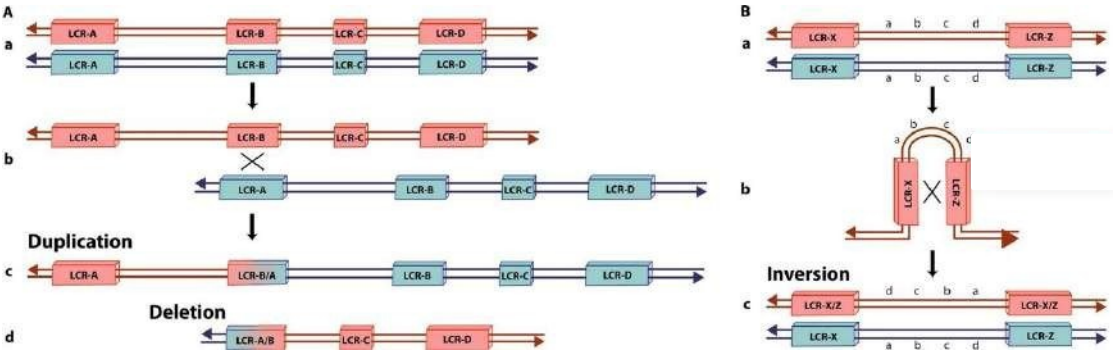


Figure 11. Non-Allelic Homologous Recombination (NAHR) (A) NAHR leading to duplication and deletion. (a) Normal chromosome pairing and alignment of low-copy repeats (LCRs) in the same orientation. (b) Because LCRs share high sequence similarity, misalignment can occur during recombination, leading to unequal crossing-over. This can generate (c) a chromosome carrying a duplication and (d) a chromosome carrying the corresponding deletion. (B) NAHR leading to inversion. (a) Normal chromosome pairing and alignment of LCRs. LCR-X and LCR-Z share similar DNA sequences but are present in opposite orientations. (b) Misalignment and recombination between these oppositely oriented LCRs can generate (c) an inversion of the intervening genomic segment. Reproduced from Burssed et al., 2022 [65], under the Creative Commons Attribution 4.0 International License. No changes were made.

2.1.7 Variant calling

Variant calling is the process of identifying differences between sequenced reads and a reference genome. After reads are aligned, algorithms scan the column-by-column view of how many reads support each base at each genomic position, known as the alignment pileup, to detect places where the sample sequence deviates from the reference. These deviations are classified as single-nucleotide variants (SNVs), small insertions and deletions (Indels), or larger structural variants (SVs). Different algorithms exist for variant calling depending on the sequencing technology and the structure of the reference, with methods developed for both linear and graph-based references (explained in section 2.1.8) and for both short- and long-read data. Algorithms are also typically developed separately for small variants and for structural variants, because the two require fundamentally different detection strategies: small variants are identified by analyzing base-level disagreements within the pileup, whereas structural variants are

inferred from larger-scale signals such as split reads, clipped alignments, read-pair orientation, and changes in read depth across genomic regions. Broadly, variant callers fall into two algorithmic categories. Classical statistical callers, such as GATK, and FreeBayes, apply explicit probabilistic models to the alignment pileup, calculating the likelihood that the observed reads at a given position reflect a true variant rather than a sequencing error, based on parameters such as read depth, base quality scores, and expected error rates [75, 76]. Machine learning-based callers, such as DeepVariant, and Medaka, instead learn the patterns that distinguish true variants from sequencing errors directly from training data, typically using neural networks that take the pileup as input and output a prediction of the most likely variation at each position [77–79]. Machine learning approaches have become particularly important for long-read data, where the error profiles of platforms such as Oxford Nanopore are too complex and context-dependent to be fully captured by simple statistical models.

In this dissertation, which integrates two studies, variant calling is implemented differently in each biological context. First, in the *T. gondii* ME49 intra-strain heterogeneity study, a long-read pangenome approach was employed using three complementary tools: Medaka [79], Sniffles2 [80], and vg [81]. Medaka was used for ONT-based small-variant calling. Medaka uses a neural network specifically trained on the error patterns produced by particular combinations of nanopore chemistry and basecaller versions. Reads are first aligned to the reference with minimap2 [82]; Medaka then converts the alignment into a strand-separated pileup where each genomic position is encoded as a count matrix of 10 features (A, C, G, T and deletion, each for forward and reverse strand). Counts are normalized to frequencies to remove coverage bias, and a bidirectional recurrent neural network evaluates the local sequence context to predict the most probable true base [83]. Differences to the reference are emitted as SNPs and Indels in VCF file format.

Sniffles2 was used for structural variant discovery from the same long-read alignments. Unlike short-read SV callers that rely on discordant pairs, Sniffles2 exploits the length of nanopore reads to detect SV signatures directly within individual alignments. It scans the BAM for split-reads, large gaps, clipped ends, and depth changes, clusters supporting reads by SV type and breakpoint, and applies a repeat-aware genotyping model to distinguish heterozygous from homozygous and mosaic events. This enables sensitive detection of deletions, insertions, inversions, duplications and translocations in the 50 bp to megabase range, which are typically missed by small-variant callers like

Medaka.

Variant calling with vg was carried out within the genotyping workflow and is described in section 2.1.9

Second, in the KIR genotyping study, variant calling was implemented as a simple pileup-based consensus approach applied at the level of coding sequences. The most likely CDS genotype was first inferred from the read data, and reads were then re-examined position by position against this initial genotype. At each position, if the proportion of reads supporting an alternative base exceeded a predefined threshold of 75% to 80%, the CDS sequence was corrected accordingly. This refinement step served two purposes: it improved the accuracy of the called allele where the initial genotype was imperfect, and it enabled the detection of novel alleles that were not represented in the existing reference database.

2.1.8 Pangenomes and genome graphs

A linear reference as explained in section 2.1.5 and its assembly in section 2.1.4 is a pragmatic necessity, but it also imposes important limitations. A linear reference represents only one possible path through the space of population genomes, usually corresponding to a single haplotype or consensus sequence derived from one or a small number of individuals. Consequently, it collapses heterozygous variation, excludes sequences absent from the reference genome, and describes all other genomes in terms of deviations from that single path [84]. Structural variants such as large insertions, deletions, inversions, duplications, and highly divergent haplotypes can be difficult to align to a linear reference and may be misrepresented, incompletely characterized, or missed entirely [85–87]. As population-scale sequencing continues to uncover extensive sequence and structural diversity, reliance on a single linear coordinate system introduces reference bias and limits the accuracy of read mapping, variant discovery, and genotyping. These limitations motivate the development of pangenome references, in which genomic diversity across individuals is incorporated directly into the reference. Formally, a pangenomic reference is a data structure that represents the genomic sequences of a population, metagenome, clade, species, or strains. Pangenomic models can take many forms, but the most widely used representation is the sequence graph, which is a data structure consisting of nodes and edges used to compress redundant input sequences into a smaller structure that remains representative of the full set [84]. In the node-labeled formulation, nodes carry DNA sequences and edges

indicate observed adjacencies between them; walks through the graph reconstruct the input genomes. Sequence graphs can be implemented as directed or bidirected graphs. A directed graph stores each DNA strand separately: the forward sequence and its reverse complement require distinct nodes, which doubles the graph and complicates the representation of strand-specific events. Because DNA is double-stranded, pangenome graphs are typically implemented as bidirected graphs. In bidirected graphs each node stores a single DNA sequence but is understood to have two ends: a left (5') and a right (3') end. Edges connect specific ends rather than whole nodes, thereby encoding orientation. Traversing a node from left to right yields the forward sequence; traversing from right to left yields its reverse complement. This allows both strands to be represented without duplication, and it enables inversions to be encoded naturally as walks that enter a node through the opposite end, rather than as separate subgraphs.

When the sequence graph is built to represent whole-genome relationships, these structures are called genome graphs [84]; walks correspond to the included haplotypes, and regions where multiple paths diverge and reconverge (bubbles) encode variation. A further refinement is the variation graph, in which the linear haplotypes of the pangenome are embedded as named paths through the sequence graph. Because positions are defined along paths rather than on graph topology alone, the coordinate system remains stable regardless of how the graph was constructed, enabling coordination of positions, annotations, and alignments with linear references [88].

In this dissertation, a variant graph was constructed using vg toolkit [81], which implements the variation graph model as a bidirected sequence graph with embedded haplotype paths. Construction proceeds in two steps: (i) the linear reference genome is converted into an initial graph backbone, and (ii) variants called across multiple samples (SNPs, indels, and structural variants in VCF format) are incrementally added. Each variant creates an alternative path through a local bubble, preserving the reference path while embedding sample-specific alleles as additional traversals. The resulting graph therefore compresses shared sequence into common nodes and represents population diversity explicitly as alternative walks, rather than as deviations from a single coordinate. Once constructed, the graph serves as the reference for sequence alignment. Unlike linear aligners, which align reads to a single reference path, graph aligners can align reads to alternative paths through the graph topology, allowing reads carrying non-reference alleles to map without being forced onto the linear reference sequence. Here, reads from each sample are aligned to the variation graph using GraphAligner

[89]. Alignment output, such as GAM or GAF, records how each read traverses the graph, including the nodes and edges supporting its placement. These graph-based alignments can then be used by downstream genotyping methods to estimate read support for reference and alternative alleles embedded in the graph.

2.1.9 Genotyping

Genotyping is the inference of the allelic state at a locus by choosing the best-supported option from a set of alternatives. In practice, no single tool covers all forms because workflows differ in how that set is built and how evidence is evaluated. In standard variant-calling pipelines, the set is built during analysis; candidate sites are identified typically after sequencing read alignment to a reference, and genotyping assigns the diploid state at each site, for example 0/0, 0/1, or 1/1 for a biallelic A/G variant. Inference is therefore site-level and tightly coupled to discovery. In catalogue-based genotyping, the set is supplied beforehand. The catalogue may contain full-length sequences, as in HLA or KIR typing with IPD-IMGT/HLA [90] and IPD-KIR [91] (typed with tools like HLA*LA [92] or Locityper [93]), or a curated panel of individual sites, as in SNP arrays or pharmacogenomic panels. The task is to select the allele, allele pair, or site-level combination best supported by the data, using evidence from read alignments or local assemblies. Inference can be site-level or whole-locus depending on the catalogue. Similar to variant calling, method choice is shaped by the sequencing data type, short versus long reads with their different error profiles and phasing capacity, the class of variation targeted from SNVs and Indels to structural variants and complex haplotypes, and the reference representation, linear reference genome (as used by most conventional genotypers like GATK and FreeBayes) versus pangenome graph representations [94] (as in vg and PanGenie [95]).

At an abstract level, a genotyping workflow proceeds by first establishing a set of candidate alleles, either discovered from the reads or assembly or taken from a reference panel, variation graph, or curated allele database. Evidence for each candidate is then extracted from the data, for example through read alignments, k-mer counts, assembly alignments, or read-depth patterns. These observations are evaluated within a statistical or probabilistic model that produces genotype scores, expressed as likelihoods, posterior probabilities, HMM state probabilities, or coverage-based support scores. Where available, priors or cohort-level information such as allele frequencies, haplotype-panel constraints, copy-number expectations, or locus-specific databases are

incorporated to constrain or refine these scores. The workflow concludes by reporting the best-supported genotype, which may take the form of a site-level call, a phased haplotype pair, a structural variant genotype, a repeat size, or a copy-number state [96]. While all variant callers comprise an implicit genotyping step that assigns the most likely genotype to a detected variant, the focus here is on the catalog-based genotyping approach. This catalog approach was particularly relevant for two parts of this dissertation. For *T. gondii* structural variant analysis, a key challenge arises when comparing variants across samples: the same biological event is often reported with slightly different breakpoint coordinates, or missed entirely in one sample due to a false negative, because of differences in read quality, sequencing depth, alignment uncertainty, or chemistry. This creates a problem for comparative analysis, because two calls may represent the same underlying variant even if their reported coordinates are not identical. Graph-based genotyping provides a way to address this issue by representing known variants in a shared structure rather than evaluating each sample only against a single linear reference.

In this dissertation, vg was used for graph-based genotyping within the variation graph framework. The first step is to construct a variation graph in which all candidate alleles, derived from a catalog of structural variants identified across samples, are embedded as alternative paths through the graph alongside the reference sequence. Variable regions of the graph are identified as snarls, graph structures containing alternative traversals between a common entry and exit point, and each snarl represents a site at which one of several candidate alleles must be selected. Sequencing reads from a given sample are then aligned to the entire graph rather than to a single linear reference, allowing reads that match an alternative path to be aligned correctly even when they would have mapped poorly to the linear reference. For each snarl, the alignments are summarized as read support across the nodes and edges of the candidate paths, and each candidate traversal is scored according to how well it is supported by the data. The genotype is then assigned from the best-supported path or pair of paths, as illustrated in Figure 12.

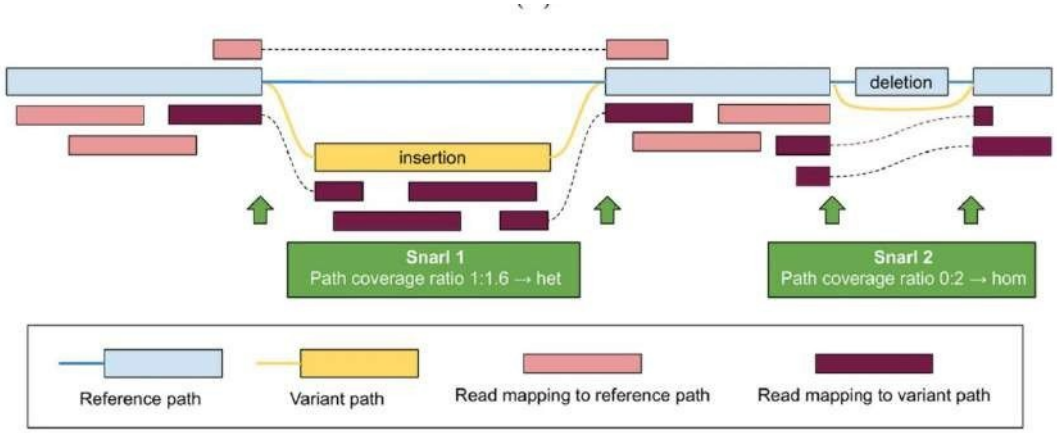


Figure 12**.** Schematic of structural variant genotyping in a pangenome graph. Reads are mapped to a graph containing both reference paths (blue) and variant paths (yellow), and genotypes are inferred from the relative read coverage of each path through a site (a "snarl"). At Snarl 1, an insertion is supported by reads mapping to both the reference and the variant path (coverage ratio 1:1.6), yielding a heterozygous call (het). At Snarl 2, a deletion is supported only by reads traversing the variant path (coverage ratio 0:2), yielding a homozygous call (hom). Reproduced from Hickey et al., "Genotyping structural variants in pangenome graphs using the vg toolkit" (*Genome Biology*, 2020) [81], under the Creative Commons Attribution 4.0 International License. No changes were made.

For the KIR locus, a different catalog-based approach was developed because the highly homologous paralogous gene structure of KIR makes general-purpose graph genotyping insufficient. A custom genotyping model was implemented that uses the IPD-KIR allele database as its reference catalog, selecting the most likely combination of catalog alleles based on read support across the coding sequences and applying a refinement step to detect novel alleles not yet represented in the catalog. The details of this workflow are described in Chapter 4.

2.2 *Toxoplasma gondii* (T. gondii)

2.2.1 Life cycle and transmission

T. gondii is an obligate intracellular parasite of the phylum Apicomplexa. It is the causative agent of toxoplasmosis, a highly prevalent disease estimated to infect approximately one-third of the global population [97]. Infection is usually asymptomatic but poses significant risk to vulnerable groups, particularly causing congenital infection when acquired during pregnancy and life-threatening disease in immunocompromised individuals, such as those with HIV/AIDS [98]. The parasite was first described in 1908 from the North African rodent *Ctenodactylus gondi,* it infects almost all warm-blooded vertebrates and can invade nearly any nucleated cell type [99]. It has a two-host life cycle: sexual reproduction occurs exclusively in felids, the definitive hosts, while asexual replication occurs in a broad range of intermediate hosts,

including humans, other mammals and birds. The life cycle involves three infectious forms: sporozoites within oocysts, tachyzoites, and bradyzoites within tissue cysts. Cats acquire *T. gondii* by ingesting tissue cysts in prey or sporulated oocysts from the environment. In the feline intestine, released bradyzoites or sporozoites invade enterocytes, replicate asexually as schizonts to produce merozoites, and then differentiate into microgametes and macrogametes. Fertilization yields oocysts that are shed in feces and sporulate to form two sporocysts, each containing four sporozoites. Intermediate hosts become infected by ingesting oocysts or tissue cysts; the liberated parasites invade nucleated cells and replicate as tachyzoites within a parasitophorous vacuole. In response to environmental stress, tachyzoites subsequently convert into bradyzoites, which persist in tissue cysts predominantly in muscle and neurons and can reactivate to tachyzoites during immunosuppression [9]. The life cycle is illustrated in Figure 13.

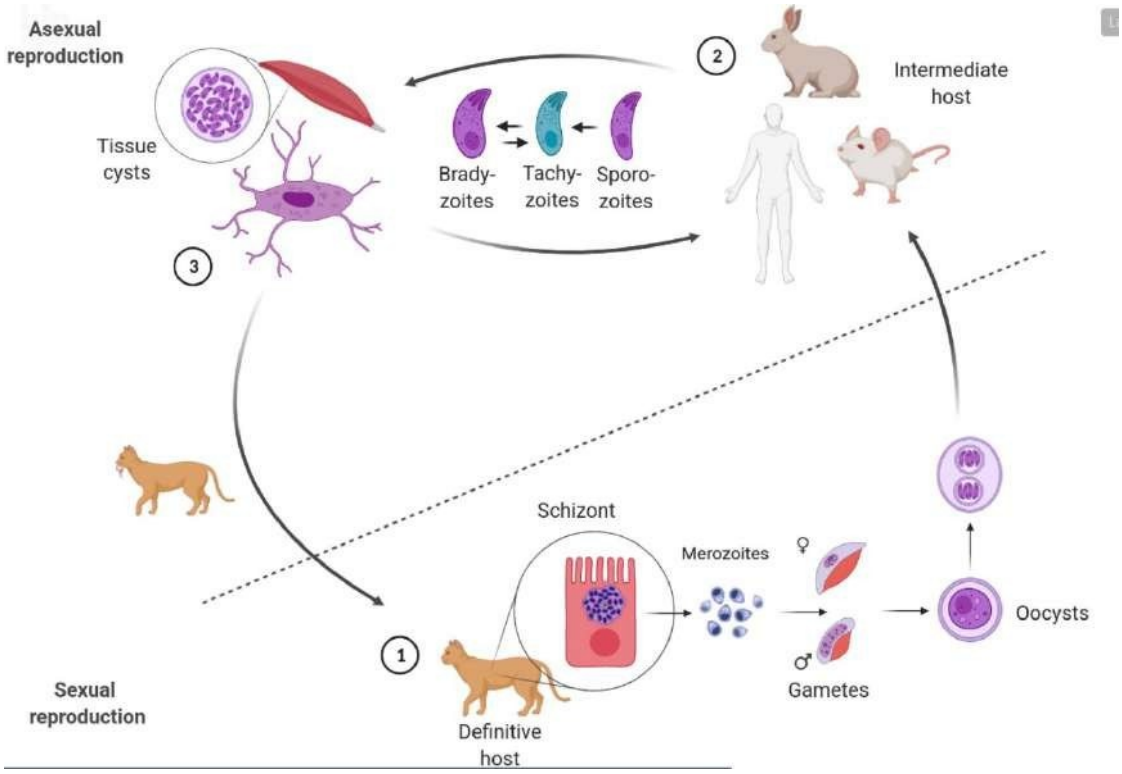


Figure 13. Life cycle of *Toxoplasma gondii*.
T. gondii undergoes sexual reproduction only in felids, the definitive hosts, where intestinal stages produce gametes that fuse to form oocysts. Oocysts are shed into the environment and can infect intermediate hosts, including humans, rodents, and other warm-blooded animals. In intermediate hosts, parasites undergo asexual replication as rapidly dividing tachyzoites and can differentiate into slow-growing bradyzoites within tissue cysts, particularly in muscle and neural tissues. Infection can occur through ingestion of sporulated oocysts from the environment or tissue cysts from infected meat. (Created with BioRender).

2.2.2 Genetic diversity of *T. gondii*

Almost all life-cycle stages of *T. gondii* are haploid; diploidy is restricted to the transient zygote formed in felids by fusion of a macrogamete and a microgamete, and is maintained within the immature oocyst until meiosis during sporulation restores haploidy [9].
The first population-level picture came from low-resolution PCR-RFLP analyses, which surveyed only a small number of loci in isolates drawn predominantly from North America and Europe and largely derived from humans and domestic animals [100]. This limited marker set and biased sampling identified three predominant allele combinations, designated types I, II, and III, leading to the view that *T. gondii* was largely clonal in these regions. These archetypal lineages are closely related, showing very low within-lineage diversity and approximately 1–3% divergence between lineages, consistent with recent expansion from a common ancestor.
Recognition of this geographic bias prompted a shift to intermediate-resolution multilocus approaches applied to globally diverse collections, particularly isolates from South America and wild hosts. These studies employed multilocus microsatellite typing and sequencing of neutral introns from housekeeping genes. The increased number of independent markers resolved substantially greater genetic diversity and grouped strains into 15 major haplogroups [101]. In this framework, the original RFLP types were retained as haplogroups 1, 2, and 3, while genotypes previously classified as “atypical” were distributed across haplogroups 4 through 15. A fourth largely clonal lineage prevalent in North American wildlife, not distinguishable using the original RFLP panel, was subsequently designated haplogroup 12. High-resolution whole-genome sequencing has since added further depth to this haplogroup framework by revealing the underlying genome architecture of *T. gondii*. Genome-wide analyses of globally distributed isolates demonstrated that *T. gondii* chromosomes are composed of mosaic haploblocks with distinct ancestral origins, and that specific haploblocks enriched in secretory pathogenesis determinants are inherited in a clade-specific manner. This mosaic architecture provides a finer-grained view of population structure that complements haplogroup classification, helping to explain how recombination between ancestral lineages has shaped contemporary strain diversity [13].

2.2.3 Genomic changes in lab strains of *T.* gondii

Previous studies on the RH strain, a type I lineage originally isolated from a 1939 case of toxoplasmic encephalitis, showed that laboratory strains accumulate tens to low hundreds of spontaneous single nucleotide variations (SNVs) over decades of culture, with a transition/transversion ratio of ~0.9 [102]. These SNVs are not distributed randomly: only 19.1% fall in exons, and most of those are synonymous, the pattern expected if replication errors occur genome-wide but purifying selection removes coding changes. Based on SNV analysis, there are no apparent hot- or cold-spots for

variations generated by either in vitro culture or chemical mutagenesis. Notably, the apicoplast genome shows ∼100-fold more spontaneous polymorphisms than the nucleus, which may reflect lower fidelity of plastid genome replication or a distinct deployment of DNA repair enzymes in that organelle [102]. *T. gondii*, like other eukaryotes, acquires this variation through two principal routes: spontaneous errors during DNA replication and during the repair of DNA lesions. In laboratory strains, the intrinsic source to acquire mutations is driven by the parasite's mode of division. *T. gondii* replicates within the parasitophorous vacuole (PV) by endodyogeny that typically completes in 6–9 h depending on strain [100]. This rapid replication imposes significant replicative stress, leading to fork stalling and collapse, which has been shown to trigger the homologous recombination repair response in *T. gondii*, mediated by RAD51 [103] Prolonged stalling leads to fork collapse into single-ended double strand break (DSB) [104], and activation of DNA repair, principally homologous recombination, which is not entirely error-free [105].

Beyond intrinsic sources of mutation, *T. gondii* experiences exogenous mutagenic pressure during infection, particularly from reactive oxygen species (ROS) generated by the host immune response [106]. Activated macrophages and other immune cells produce ROS as part of their antiparasitic defense, exposing intracellular *T. gondii* to oxidative damage that can cause base oxidation, single-strand breaks, and double-strand breaks in the parasite genome. *T. gondii* also possesses a functional non-homologous end joining (NHEJ) pathway, mediated by KU70 and KU80, which is in fact the preferred DSB repair mechanism in this parasite and contributes to genomic instability through its inherently error-prone joining of broken ends [107].

2.3 Killer-cell Immunoglobulin-like Receptors (KIR)

2.3.1 KIRs and the Regulation of NK Cell Recognition

KIRs are a family of innate immune receptors expressed primarily on natural killer (NK) cells and subsets of T cells. NK cells are innate lymphocytes that recognize and eliminate virus-infected and cancerous cells. Upon target recognition, NK cells kill through the directed release of cytotoxic granules containing perforin and granzymes: perforin forms pores in the target cell membrane through which granzymes enter and trigger apoptosis [17]. This killing response depends on the integration of signals from two opposing receptor classes: activating receptors, such as NKG2D and the natural cytotoxicity receptors (NCRs), recognize stress-associated ligands upregulated on

infected or malignant cells and promote killing, while inhibitory receptors, including KIRs, bind self MHC class I molecules on healthy cells and suppress it [108].

The best-characterized pathway governing this balance is missing-self recognition as illustrated in Figure 14. Inhibitory KIRs continuously survey target cells for self MHC class I; in healthy cells this interaction delivers a dominant inhibitory signal that prevents NK cell activation [109,110]. When a cell downregulates MHC class I, a common immune evasion strategy in viral infection and malignant transformation, the inhibitory signal is lost and the NK cell is released from restraint. However, this alone is often insufficient to trigger a full cytotoxic response, as efficient activation requires the concurrent engagement of at least one activating receptor by its cognate stress ligand. Missing self therefore lowers the threshold for activation, but an activating signal is still needed to start the response [111,112]. This two-signal requirement helps NK cells attack abnormal cells while sparing healthy tissue.

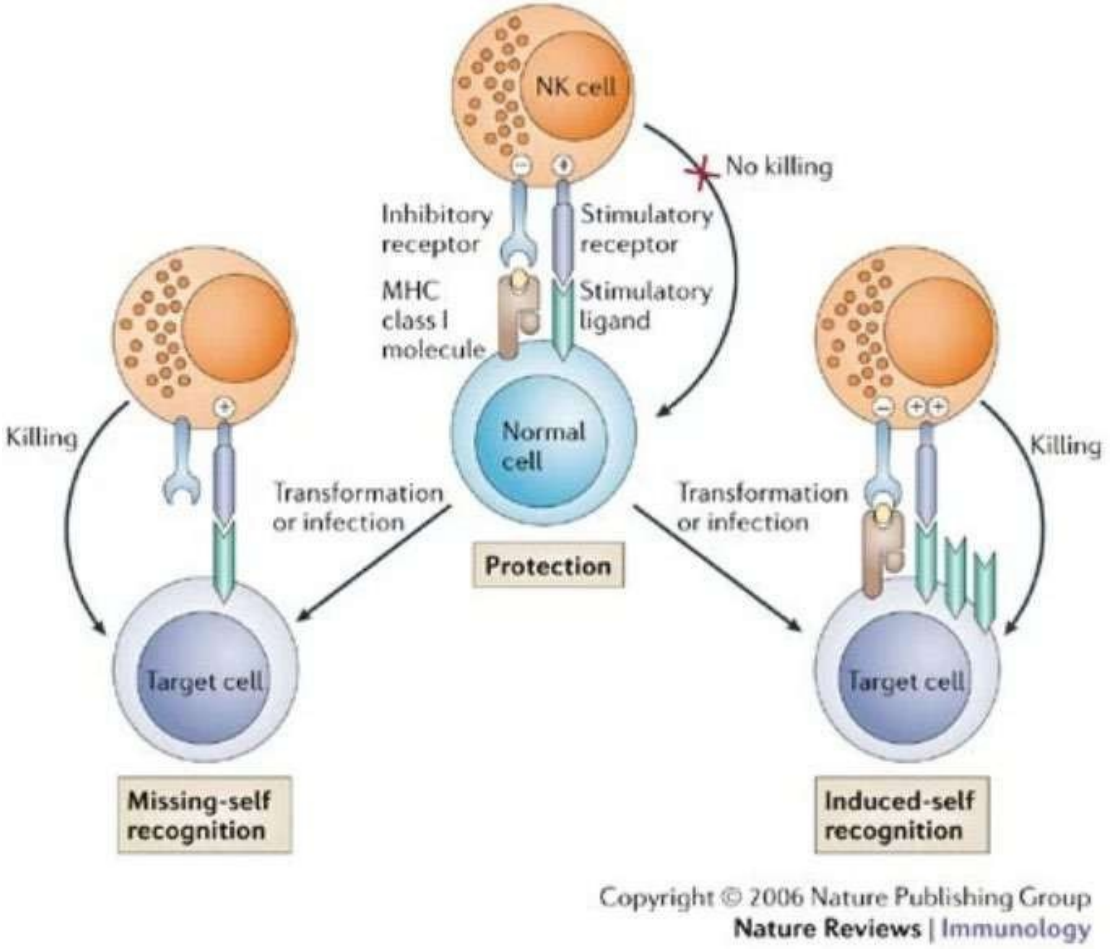


Figure 14. NK cell recognition of normal and altered self.

Natural killer (NK) cell activity is determined by the balance between inhibitory and activating signals. Normal cells expressing MHC class I engage inhibitory NK-cell receptors and are protected from killing, even when activating ligands are present. In transformed or infected cells, loss of MHC

class I can remove inhibitory signalling and trigger missing-self recognition, whereas increased expression of stress-induced activating ligands can trigger induced-self recognition. Both mechanisms shift the balance toward NK-cell activation and target-cell killing. Reproduced from Raulet et al., 2006 [111]; licence obtained from Springer Nature. No changes were made.

While missing-self recognition and the two-signal requirement describe how NK cells respond to the absence of inhibitory input, the discovery of activating KIRs introduced a more controversial dimension to NK cell biology. Rather than suppressing NK cell responses, activating KIRs contribute to the recognition of infected or malignant cells. However, the ligands for most activating KIRs remain poorly defined, and the finding that some activating KIRs recognize the same MHC class I ligands as their inhibitory counterparts is paradoxical, as it raises the question of how activating KIR engagement does not trigger killing of healthy cells that still express MHC class I, making the activating KIR family an active area of ongoing investigation [113,114].

2.3.2 The KIR gene family

The KIR gene family is clustered within a 100–200 kb region on chromosome 19q13.4. This region is part of a larger immune gene complex known as the Leukocyte Receptor Complex (LRC) [91]. It includes 8 inhibitory KIR genes, such as KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL5A, KIR2DL5B, KIR3DL1, KIR3DL2, and KIR3DL3; 6 activating KIR genes, including KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS4, KIR2DS5, and KIR3DS1; the atypical gene KIR2DL4, which can mediate both inhibitory and activating functions; and two pseudogenes, KIR2DP1 and KIR3DP1. In most genomic regions, a gene is defined by a fixed position on the chromosome, while alleles are sequence variants of that gene at the same position. The KIR region does not always follow this simple model because KIR genes can occur in different combinations and orders across haplotypes [16]. As shown in Figure 15, one haplotype may contain a certain arrangement of KIR genes, while another may have extra, missing, or repositioned genes. For example, KIR2DS3 can be found at two different positions within the KIR cluster. Therefore, KIR diversity includes both allelic variation and structural variation, making the distinction between a gene and an allele less clear than in a typical single-locus region.

Within the KIR cluster, KIR genes are arranged head-to-tail approximately 2.4 kb apart and are organized into two broad haplotype groups. Group A haplotypes have relatively fixed gene content and are dominated by inhibitory KIR genes, with KIR2DS4 as the only activating gene: KIR3DL3-KIR2DL3-KIR2DL1-KIR2DP1-KIR2DL4-KIR3DP1-

xl

KIR3DL1-KIR2DS4-KIR3DL2. In contrast, group B haplotypes are more variable and often contain additional activating KIR genes, including one or more of KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS5, and KIR3DS1, as well as variable inhibitory genes such as KIR2DL2 and KIR2DL5A/B.

Structural variation in the KIR region is mainly generated by recombination between highly similar KIR genes. Recombination at a central hotspot can exchange the centromeric and telomeric parts of the cluster, producing new haplotype combinations. In addition, unequal crossing-over between misaligned KIR genes can add, remove, or rearrange genes, leading to copy-number variation and gene-content diversity. If recombination occurs within a gene, it can produce fusion genes made from parts of different KIR genes. In addition to this structural variation, individual KIR genes also show extensive allelic diversity, with many sequence variants differing in their coding regions, regulatory regions, or both. As a result, accurate KIR genotyping is challenging because it requires distinguishing highly similar paralogous genes, identifying gene presence and copy number, resolving haplotype structure, detecting hybrid genes, and assigning alleles within genes that are themselves highly polymorphic.

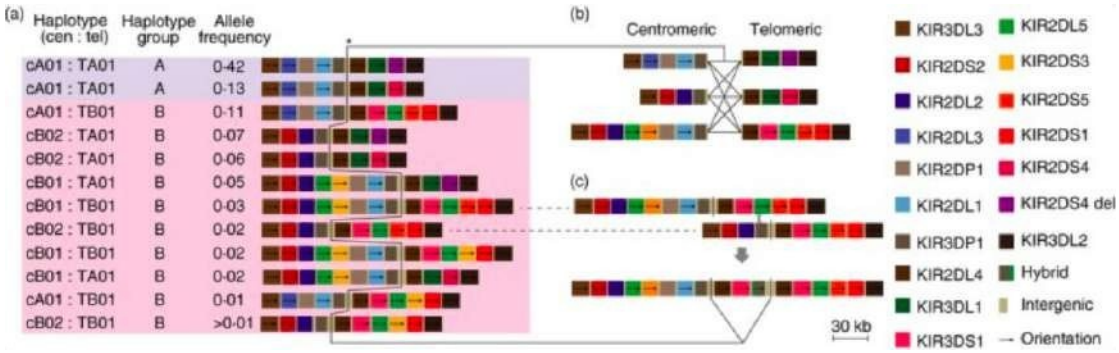


Figure 15. Structural haplotypes of the KIR gene cluster and mechanisms generating KIR diversity.

(A) Common KIR haplotypes differ in gene content, gene order, and allele frequency. Group A haplotypes have relatively conserved gene content and are dominated by inhibitory KIR genes, whereas group B haplotypes are more variable and often contain additional activating genes. KIR2DP1 and KIR3DP1 are pseudogenes. The figure also illustrates how the distinction between genes and alleles can be blurred in the KIR region, as some genes, such as KIR2DS3, can occur at different positions within the cluster. (B) Homologous recombination at a central recombination hotspot can shuffle the centromeric and telomeric parts of the KIR locus, producing new haplotype combinations. (C) Additional diversity is generated by unequal crossing-over, which can delete, duplicate, reposition, or create hybrid KIR genes. Reproduced from Béziat et al., 2016 [16] under a Creative Commons licence. No changes were made.

KIR alleles are named following an HLA-like nomenclature system coordinated by the IPD-KIR Database. The gene name first describes the structural features of the receptor.

In names such as KIR2DL1 or KIR3DS1, the number 2D or 3D indicates whether the receptor has two or three extracellular immunoglobulin-like domains, while L or S indicates the length of the cytoplasmic tail. L denotes a long cytoplasmic tail, usually associated with inhibitory signalling through ITIM motifs, whereas S denotes a short cytoplasmic tail, usually associated with activating signalling through adaptor proteins (Figure 16, upper panel). After the gene name and an asterisk, alleles receive a numerical designation: the first three digits distinguish alleles that differ in protein sequence; the next two digits distinguish alleles with synonymous differences within the coding sequence; and the final two digits distinguish alleles that differ only in introns, promoters, or other non-coding regions. Thus, a 3-digit name defines the protein sequence, a 5-digit name adds synonymous coding variation, and a 7-digit name represents the fully resolved genomic allele (Figure 16, lower panel). Seven-digit resolution is particularly challenging for KIR because introns are not under the same selective pressure as exons encoding the HLA-binding domains. Intronic regions accumulate neutral variation rapidly, leading to a large number of closely related alleles that differ only outside the coding sequence. This inflates allelic diversity without functional consequence, complicates phasing across the highly homologous and repetitive KIR cluster, and requires long-read sequencing spanning entire genes to resolve confidently.

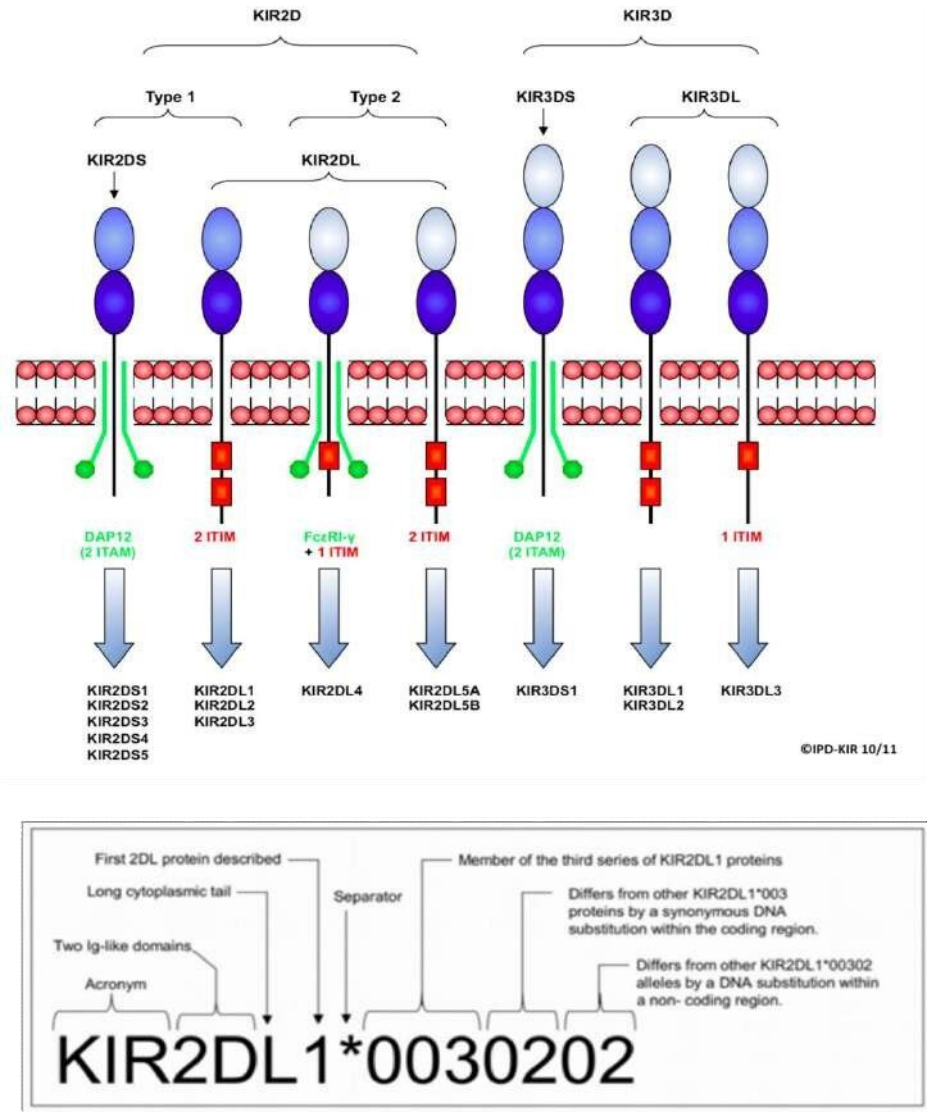


Figure 16. KIR receptor structure and allele nomenclature.
KIR gene names describe structural features of the receptor: 2D or 3D indicates the number of extracellular immunoglobulin-like domains, while L and S indicate long or short cytoplasmic tails, respectively. Long-tailed KIRs usually contain inhibitory ITIM motifs, whereas short-tailed KIRs signal through adaptor proteins such as DAP12 containing ITAM motifs. The lower panel illustrates the IPD-KIR allele nomenclature system, in which the gene name is followed by an asterisk and numerical fields that distinguish protein-changing, synonymous coding, and non-coding sequence variation. Adapted from the IPD-KIR Database.

2.3.3 KIR genetic diversity and functional consequences

The extent of KIR genetic diversity, encompassing gene content variation, copy number variation, and allelic polymorphism, has direct functional consequences for NK cell biology. At the most fundamental level, the presence or absence of a KIR gene determines whether an NK cell can express that receptor at all, shaping the repertoire of receptors available to survey target cells. Importantly, gene presence does not equate to expression, as KIR expression is stochastic, allele-dependent, and epigenetically

regulated [115]. Pseudogenes are never expressed, several common alleles are transcriptionally silent or retained intracellularly, and each NK cell clone expresses only a subset of its inherited KIR genes. Beyond gene content, copy number variation influences the frequency of NK cells expressing a given KIR: individuals carrying two copies of KIR3DL1 have a higher proportion of KIR3DL1-positive NK cells than those with a single copy, suggesting that each gene copy is regulated independently and contributes additively to the size of the expressing NK cell subset [116]. This has functional implications because a larger pool of KIR3DL1-expressing NK cells provides greater surveillance capacity against targets that have downregulated HLA-Bw4, the cognate ligand of KIR3DL1, and higher KIR3DL1 copy number has been associated with improved control of HIV infection [117,118]. Allelic variation adds a further layer of complexity: even when a KIR gene is present, the functional output depends on which allele is carried. For example, KIR2DL1 allotypes differ in binding affinity for HLA-C2, such that alleles with weaker affinity produce less well-educated NK cells (explained later) with lower interferon-γ production upon target recognition [16,119,120].

The interplay between KIR and HLA class I variation further amplifies this diversity: because KIR and HLA genes are encoded on separate chromosomes, the number of possible KIR/HLA genotype combinations across individuals is enormous, tending toward individuality, with consequences for resistance to infection, susceptibility to autoimmune disease, reproductive outcomes, and transplantation. First, in establishing NK cell function is a developmental process called NK cell education, through which NK cells are calibrated to respond appropriately to abnormal cells [121,122]. During development, NK cells expressing inhibitory KIRs encounter healthy surrounding cells expressing self HLA class I molecules. When an inhibitory KIR engages its matching self HLA class I ligand, it delivers a chronic calibrating signal that progressively raises the NK cell's internal activation potential, a process described by the rheostat model [111,121].NK cells whose inhibitory KIRs find a matching self HLA ligand become fully licensed and highly responsive, while those lacking a matching KIR/HLA pair remain hyporesponsive even when they later encounter abnormal cells [122]. Education therefore depends critically on the match between the inhibitory KIR receptors expressed by an NK cell and the HLA class I ligands present in that individual. For example, KIR3DL1 recognizes HLA-Bw4. In individuals who carry KIR3DL1 but lack HLA-Bw4, NK cells expressing KIR3DL1 do not receive KIR3DL1-mediated licensing

and are generally less responsive through this pathway. However, these cells may still be educated through other self-HLA-specific inhibitory receptors that they co-express [118,120].

Additionally, KIR/HLA interactions shape reproductive biology through uterine NK cells: maternal KIR2DS1 binding to fetal HLA-C2 on trophoblast cells activates decidual NK cells to secrete growth factors including GM-CSF, promoting trophoblast migration and healthy placentation, while triggering of activating KIR2DS4 on uterine NK cells induces a distinct cytokine and chemokine profile including GM-CSF, XCL1, and CCL1 [123,124]. In the context of viral infection, distinct KIR/HLA compound genotypes have been shown to alter the kinetics of NK cell responses in influenza A infection models, suggesting that individual genetic variation in these receptor-ligand pairs shapes not only the magnitude but also the timing of innate immune responses [125]. Finally, in therapeutic settings, KIR/HLA ligand context influences the efficacy of antibody-dependent cellular cytotoxicity: in lymphoma patients treated with rituximab, low or absent KIR ligand expression on tumour cells favoured stronger NK cell ADCC responses, suggesting that KIR genotyping may have value in predicting or optimising responses to NK cell-engaging therapies [126,127]

Taken together, these findings strongly support the importance of comprehensively characterizing KIR genotypes to understand their functional consequences. As computational methods for resolving the highly homologous paralogs, copy number, and complex haplotype structure of the KIR locus continue to improve, increasingly refined genotype-phenotype correlations are expected to emerge, deepening our understanding of NK cell biology and its role in human health and disease.

3 Publications

3.1 Publication 1: “Intra-strain genetic heterogeneity in Toxoplasma gondii ME49: Oxford Nanopore long-read sequencing reveals copy number variation in the ROP8-ROP2A locus”

Manuscript published in: BMC Genomics

Impact factor: 3.7 (2024)

DOI: https://doi.org/10.1186/s12864-025-12372-5

Authorship: Shared first author

Contribution to the publication:

In this study, I performed the genome assembly and developed the overall analysis strategy. I designed and implemented the full analysis pipeline, including the variant analysis and graph-based genotyping framework, which was developed jointly with MN. I carried out the data analysis and led the interpretation of the results. The manuscript was written jointly with AD. Parasite samples and sequencing data were provided by LH, DW, JM, KB, and MH. US, DD, and KP contributed through scientific discussion and feedback. The project was supervised by AD. All authors contributed to the revision and approved the final manuscript.

Gohar *et al. BMC Genomics* (2025) 26:1094
https://doi.org/10.1186/s12864-025-12372-5

BMC Genomics

RESEARCH Open Access

Intra-strain genetic heterogeneity in *Toxoplasma gondii* ME49: Oxford Nanopore long-read sequencing reveals copy number variation in the ROP8-ROP2A locus

Yomna Gohar[1*†], Marie Neumann[1†], Lisanna Hülse[1], Daniel Wind[1], Julia Mock[1], Karin Buchholz[1], Marcel Helle[1], Ursula R. Sorg[1], Daniel Degrandi[1], Klaus Pfeffer[1] and Alexander Dilthey[1,2]

Abstract

Background *Toxoplasma gondii* is an important pathogen and model organism for studying mechanisms of immune evasion and defense. Within the same strain, model organisms are typically assumed to be isogenic; for *T. gondii*, within-strain genetic divergence has been detected based on phenotypic changes and older molecular techniques but not characterized at the genomic level. We therefore used Oxford Nanopore long-read sequencing to characterize three independently maintained *T. gondii* ME49 isolates: 2015T and 2020T (obtained from ATCC and propagated in cell culture), and 2000B (propagated in mice).

Results We *de novo* assembled a new *T. gondii* ME49 reference genome and, using state-of-the-art variant calling combined with pangenomic genotyping, detected variants between the sequenced isolates. Our new reference genome exceeded existing reference genomes in continuity (NG50 = 6.68 Mb versus 1.2 Mb in RefSeq) and structural accuracy, resolving all chromosomes except for a single break in the ribosomal DNA region. For isolates 2000B and 2020T, we identified 106 and 128 variants, respectively, across a final call set of 79 SNVs, 93 INDELs, and five structural variants; 18 small non-synonymous variants included genes associated with *T. gondii* life cycle (AP2X-8) and virulence in vivo (6-phosphogluconate dehydrogenase). A 13 kb expansion in the ROP8-ROP2A virulence locus increased the copy number of ROP2A-ROP8 genes in isolates 2000B and 2020T from three to six.

Conclusions We provide an improved *T. gondii* ME49 reference genome and demonstrate the potentially confounding effect of intra-strain genetic heterogeneity, highlighting the need for continuous genomic monitoring for long-term genetic identity.

Keywords Toxoplasma gondii, Within-strain variation, Oxford nanopore long-read sequencing, Structural variation, Genome assembly

[†]Yomna Gohar and Marie Neumann contributed equally to this work.

*Correspondence:
Yomna Gohar
Yomna.Gohar@med.uni-duesseldorf.de

[1]Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[2]Center for Digital Medicine, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Background

Model organisms are essential in biological research, providing controlled systems to study fundamental biological processes. A key assumption is that representatives of a given strain are largely isogenic. However, prolonged laboratory maintenance, for example after acquisition from a commercial vendor (e.g., ATCC) or a partner lab, can lead to genetic divergence, which—unless it has obvious phenotypic effects—often goes undetected [1–4]. Studies have demonstrated that genetic heterogeneity accumulates during laboratory maintenance and impacts research outcomes and phenotypic variation in key model organisms, including laboratory mice (*Mus musculus*) [5–8], *Drosophila melanogaster* [9], *Aedes aegypti* (a vector for the yellow fever virus) [10], bacteria (for example, *Escherichia coli* and *Myxococcus xanthus*) [11–13], *Caenorhabditis elegans* [14] and different human cell lines including human pluripotent stem cells (hPSCs) [15–18], which serve as in vitro model systems for studying human biology and disease. Genetic heterogeneity can encompass several types of variation, including single nucleotide polymorphisms (SNPs), small insertions and deletions (INDELs), and structural variants (SVs), the latter of which is often difficult to detect [8, 19, 20]. Mechanisms for the generation of genetic heterogeneity include replication errors, DNA repair mechanisms, and selection pressures [21–23]. In vivo passaging, in particular, exposes strains to host immune responses, accelerating genetic drift and selection [24, 25].

Among microbial pathogens, *Toxoplasma gondii* serves as a crucial model organism for studying host-pathogen interactions, immune evasion, and virulence mechanisms [26, 27]. Globally, the parasite *T. gondii* exhibits a broad spectrum of genotypes, with highly diverse non-archetypal lineages especially prevalent in South America, whereas the strains found in North America and Europe mainly belong to three clonal (archetypal) lineages—types I, II, and III [28]. Type II strains are responsible for most of the human infections studied in North America and Europe, and are likewise prevalent in livestock from these regions [28]. Previous studies on various *T. gondii* strains have shown that continuous passaging in mice or cell culture can lead to genetic and phenotypic changes. For example, research on Type I RH-derived clonal lineages, using restriction fragment length polymorphism (RFLP) analysis, has demonstrated that repeated passaging could induce genomic heterogeneity [29]. Some RH-derived lineages have also showed phenotypic divergence, such as variation in plaque size, growth rate, differentiation, and ability to survive outside host cells [30]. Additionally, prolonged laboratory maintenance of *T. gondii* through serial passaging in mice or cell culture has been linked to the loss of oocyst production in cats, as observed in multiple strains, including M-7741, GT-1, and RH, likely due to accumulated genetic changes over time [31–33].

While these studies have demonstrated that laboratory propagation of *T. gondii* can induce genetic heterogeneity associated with phenotypic changes, they were conducted before the emergence of high-throughput sequencing and did not include a characterization of genetic heterogeneity at the sequence level. What is more, comparative genomic studies in the *T. gondii* field in general have traditionally relied on short-read sequencing technologies [34, 35]. Short-read sequencing technologies have inherent limitations in detecting complex structural variations, as short reads cannot span large genomic rearrangements, repetitive regions, or highly polymorphic loci [36]. As a result, the importance of complex genetic variations to overall *T. gondii* genetic diversity remains incompletely characterized.

In contrast, long-read sequencing technologies, such as Oxford Nanopore sequencing, can overcome these challenges by producing long continuous reads, enabling more comprehensive detection of genomic variants. Recent advances, particularly in the Oxford Nanopore R10.4 platform, have further improved sequencing accuracy, allowing the generation of near-finished microbial genomes without the need for short reads or reference-based polishing [37].

In this study (Fig. 1), we leveraged Oxford Nanopore long-read sequencing to assess genomic divergence among three independently maintained *T. gondii* ME49 isolates—a widely used type II reference strain with moderate virulence [38, 39]. Employing de novo assembly and a multimethod variant calling approach, followed by graph genome-based genotyping (Supplementary Figure S1), our analysis revealed genetic heterogeneity among these isolates, characterized by both small variants and large structural variations that would have been inaccessible using traditional sequencing methods. Our findings contribute to a broader understanding of *T. gondii* genomic variations, with implications for strain characterization and experimental reproducibility.

Results

Generation of nanopore sequencing data for three *T. gondii* ME49 isolates

We selected three ME49 *T. gondii* isolates for characterization: 2015T, obtained from ATCC (catalog number 50611) thawed in 2015, expanded for two passages, refrozen and then continuously propagated in cell culture from 2022 onward; 2020 T, obtained in 2020 from ATCC (catalog number 50611) and propagated in cell culture; and 2000B, originally obtained in the early 1990s and propagated in CD1 mice (Fig. 2A; see Supplementary Note for additional details on the propagation history of these isolates). We generated single-molecule long-read

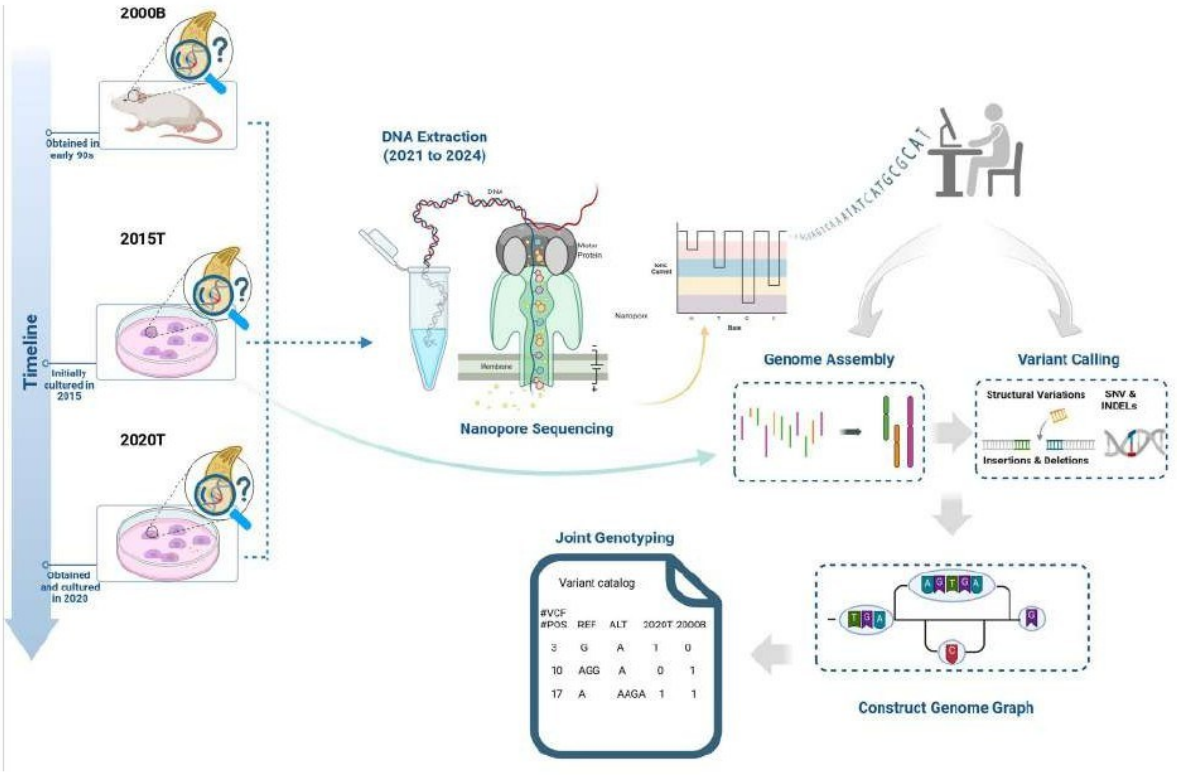
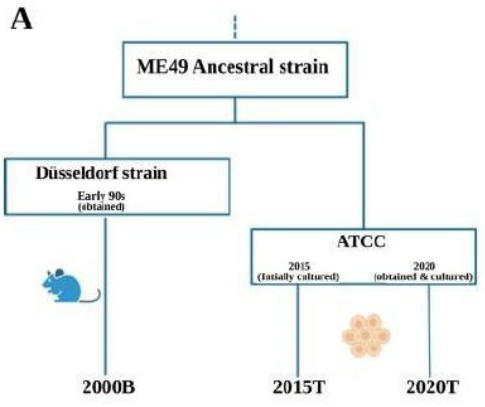


Fig. 1 Overview of the sequencing and variant analysis pipeline for T. gondii ME49 isolates. Created in BioRender. https://BioRender.com



B

	2015T [c]	2020T	2000B
Total sequencing data (Gb)	16	4	31.8
T. gondii sequencing data (Gb) [d]	13.35	1.31	1.55
Estimated genome coverage in X [a,b]	203	20	23.6
Number of reads [a]	716,210	468,681	672,080
Mean read length [a]	18,645	2,805	2,307
Read N50 [a]	37,228	5,061	12,898
Number of reads >=10kb [a]	362,791	20,913	28,001
Number of reads >=100kb [a]	4,542	2	228

[a] statistics are based on *T. gondii* reads in the second row
[b] calculated genome size as 65,500,000
[c] statistics are calculated after HERRO correction.
[d] calculated from non-human/mouse reads

Fig. 2 Origin and sequencing summary of *Toxoplasma gondii* ME49 isolates used in this study. (**A**) Schematic overview of the propagation and acquisition history of the ME49 isolates 2000B, 2015T, and 2020T. (**B**) Summary statistics for Oxford Nanopore long-read sequencing data generated for each isolate

Oxford Nanopore R10 sequencing data for the three isolates (Fig. 2B), yielding 203X genomic coverage at a read N50 of 37 kb for isolate 2015T; 20X genomic coverage at a read N50 of 5 kb for isolate 2020T; and 23X genomic coverage at a read N50 of 12.8 kb for isolate 2000B. For isolate 2015T, all sequencing data was generated with the R10.4.1 sequencing chemistry and base calling was carried out with Dorado [40] in "super high accuracy" mode, followed by HERRO (haplotype-aware error correction of ultra-long nanopore reads) [41] for error correction; for the other isolates, multiple sequencing chemistries in the R10 family were employed and base calling was carried out with Guppy [42] in "super high accuracy" mode (see Supplementary Table 1 for details on the generated sequencing data, proportion of host DNA, sequencing chemistries and base calling). Preliminary analyses of the generated data indicated that reads containing mitochondrial DNA fragments led to challenges during further

data analyses; such reads were therefore removed (Supplementary Notes).

High-quality genome assembly for isolate 2015T

We assembled the high-quality HERRO-corrected sequencing data for isolate 2015T using Flye [43] and obtained a genome assembly consisting of 22 contigs and a total size of 64.5 Mb at an NG50 of 6.7 Mb (Fig. 3A). We investigated the chromosomal completeness of our assembly and found that 12 of 13 *T. gondii* ME49 chromosomes were almost completely covered by single assembly contigs; the remaining chromosome IX was split into two contigs (Fig. 3B and C). The genome of the *T. gondii* ME49 apicoplast, which is approximately 35 kb in size and particularly challenging to assemble due to the presence of inverted repeats [44], was represented by two contigs of approximately 23 and 5 kb in size. We annotated our assembly with Companion [45] and found that it contained 9,137 genes and 166 pseudogenes.

A

	TGA4 Refseq	Xia et al. assembly	2015T assembly
Assembly length (Mb)	65.66	64.77	64.53
Number of contigs	2507	36	22
Contig N50 (Mb)	1.20	6.67	6.68
Number of chromosome-level contigs (99% of chromosome length)	3	12	12
Erroneous assembly positions (visual inspection)	-	55	2
Suspicious assembly positions (visual inspection)	-	0	2
number of mapped genes	8,597	8,768	8,770
Fraction of perfectly mapped genes	0.99	0.73	0.89
Apicoplast chromosome length	34,996	34,779	24,398 + 5,153

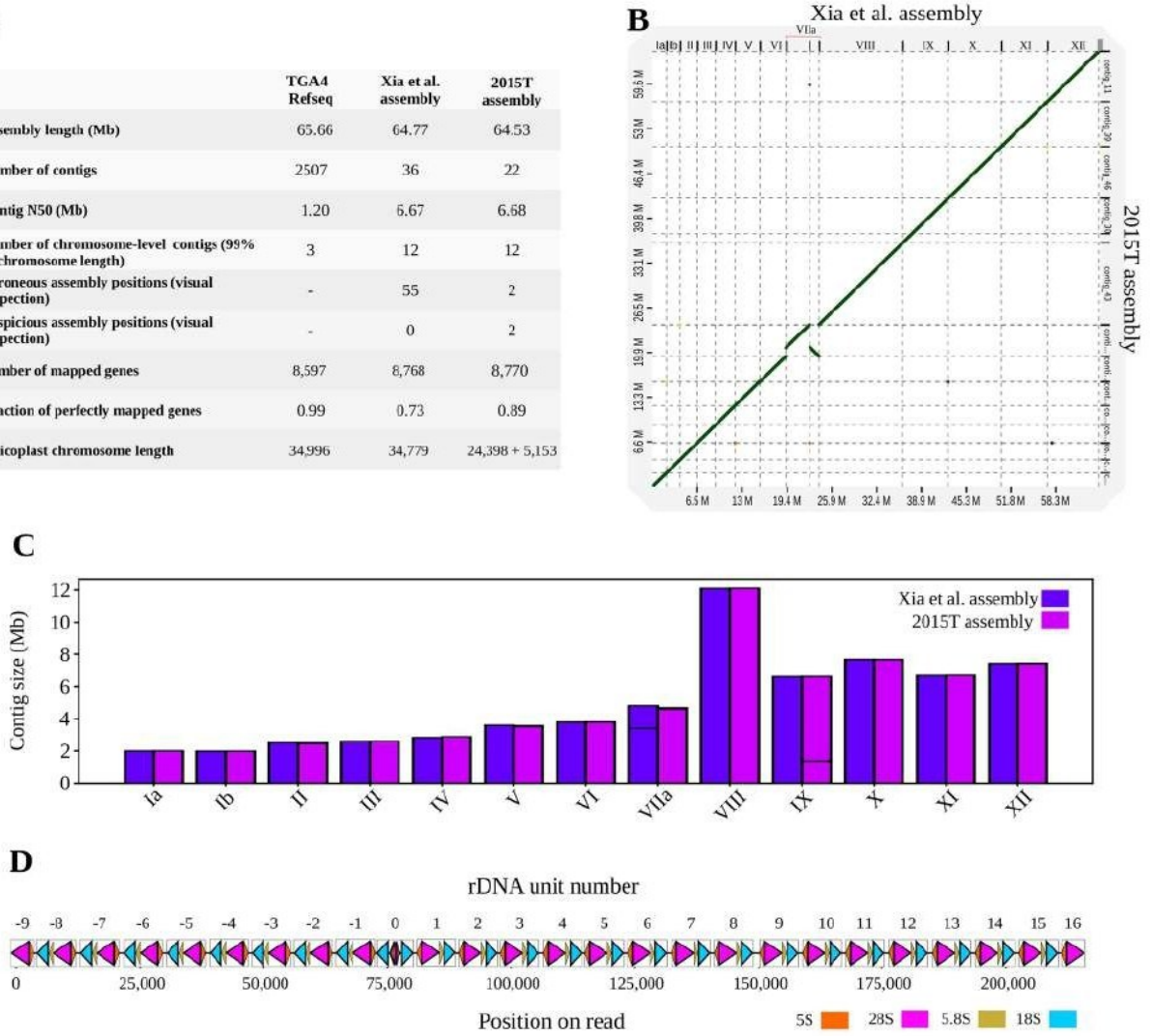


Fig. 3 Comparison of the 2015T assembly to existing *T. gondii* ME49 references and detailed view of the rDNA locus. (**A**) Comparison of the 2015T genome assembly to previously published *T. gondii* ME49 reference genomes. For TGA4, the number of contigs and the contig N50 values were derived from the NCBI Genome website (assembly accession GCF_000006565.2). (**B**) Dot plot showing structural alignment between the 2015T and Xia et al. Assemblies. (**C**) Contig size comparison, with horizontal lines indicating breaks between contigs. (**D**) Partial sequence structure of the ribosomal DNA (rDNA) locus in isolate 2015T, visualized using the longest read aligned to this region. Triangles mark the positions of 18 S, 28 S, 5.8 S, and 5 S rRNA genes along the read. The orientation of each triangle indicates the direction of rRNA gene alignment: left-facing triangles represent reverse-strand mappings, and right-facing triangles represent forward-strand mappings

Comparison of 2015T to other existing reference assemblies

To evaluate the quality of our assembly, we compared it to two existing *T. gondii* ME49 reference genomes (Fig. 3A), the TGA4 *T. gondii* reference genome from RefSeq and a more recent single-molecule long-read sequencing-based assembly [46], referred to as the "Xia et al. assembly" for the remainder of this paper. Compared to TGA4 (NCBI RefSeq assembly: GCF_000006565.2), our assembly was much more contiguous (NG50 of 6.68 Mb for our assembly compared to 1.20 Mb for TGA4); in addition, TGA4 was reported to contain unresolved duplications (e.g., in the ROP4/ROP7 locus) and an erroneous inversion on chromosome 4 [46]. The assembly from Xia's study exhibited comparable contiguity to our assembly (36 contigs at a NG50 of 6.67 Mb); we thus carried out an in-depth comparison. First, we carried out a dot plot analysis which revealed a high degree of collinearity between Xia et al. and our assembly, with few deviations observed for Chromosomes VIIa and IX (Fig. 3B). Chromosome VIIa was split in two fragments of approximately 3.4 Mb and 1.3 Mb in size in Xia et al., whereas our assembly contained a near-complete chromosome VIIa contig of 4.5 Mb in size. Chromosome IX was split into two contigs in our assembly, whereas in Xia et al. it was represented by a single contig. Otherwise, both assemblies exhibited almost identical contig sizes (Fig. 3C). Second, we investigated the size differences between the two assemblies. Using nucmer [47], we found that each assembly contained a stretch of sequence nearly identical in length, that could not be aligned to the other assembly (76 kb for Xia et al. compared to our assembly and 86 kb for our assembly compared to Xia et al.), suggesting that both assemblies covered comparable fractions of the *T. gondii* ME49 genome despite the slightly increased size of the Xia et al. assembly.

Third, the proportion of *T. gondii* gene sequences from ToxoDB [48] with perfect alignments was higher for our assembly than for Xia et al. (89% compared to 73%, see Fig. 3A and Figure S2 for examples of genes exhibiting more accurate mapping in our assembly), and annotation of the Xia et al. assembly using the same Companion approach we applied to our 2015T assembly revealed a higher fraction of pseudogenes (8891 genes and 557 pseudogenes). These results indicated a higher consensus quality of our assembly compared to the Xia et al. assembly, likely associated with the higher error rate of older Nanopore sequencing data used by Xia et al. (MinION R9.4.1).

Fourth, manual inspection of read alignments (based on the reads used by Xia et al. for assembly construction) in Integrative Genomic Viewer (IGV) [49] identified 55 regions in the Xia et al. assembly that were likely structurally incorrect (Supplementary Table 2, Supplementary Notes). We projected the coordinates of these onto our assembly and, based on visual assessment in IGV, found that all but 4 of the evaluated positions were correctly resolved in our assembly. One of the 4 unresolved positions coincided with the Chromosome IX breakpoint (see above); visual inspection in Xia et al. was indicative of a misassembled repetitive region with a significant read depth increase and numerous reads with a mapping quality of zero at that position (Figure S3). Analysis of reads from the region using BLAST [50] indicated the presence of ribosomal DNA (rDNA) sequences. We extracted the longest read (207,813 bp) from our 2015T dataset that mapped to the rDNA locus in the Xia et al. assembly and found that it contained at least 23 blocks of repetitive rDNA units (Fig. 3D). Notably, the read indicated a shift in relative rDNA repeat unit orientation in the middle of the rDNA region captured by the read, which may contribute to the assembly challenges at this locus. In contrast, the corresponding position in the Xia et al. assembly showed only 16 rDNA blocks (Figure S3). Together, the discrepancy in rDNA copy number and the visual evidence from IGV suggested that, despite the representation of Chromosome IX as a single contig in the Xia et al. assembly, the assembly of this region in Xia et al. was not accurate.

We thus concluded that our assembly was - with the possible exception of the apicoplast genome, which was represented as a single contig of approximately 35 kb in size in the Xia et al. assembly (Figure S4) - of higher quality than the Xia et al. assembly, likely reflecting the markedly higher coverage of the dataset used to construct our assembly (203X vs. 30X) and the improved quality of R10.4.1 sequencing data. A final visual inspection of the complete assembly confirmed the structural accuracy of our assembly, with the possible exception of telomeric regions, for which we did not carry out an in-depth assessment.

Variant discovery in isolates 2020T and 2000B

To detect candidate polymorphisms among the three isolates sequenced by in this study, we carried out variant calling for isolates 2020T and 2000B against the 2015T assembly, supplemented with the apicoplast sequence from Xia et al. to prevent read misalignment. For discovery of candidate small variants, i.e. of single nucleotide polymorphisms (SNPs) and small insertions/deletions (INDELs), we used medaka [51]. To minimize false-positive calls, we applied stringent filtering criteria (DP ≥ 10; GQ ≥ 10; removal of variants from structurally anomalous regions; see Materials and Methods), reducing the total set of small variants calls from 4,076 to 365 in sample 2020T and from 2,041 to 376 in 2000B (Supplementary Table 3). The post-filtering candidate variant set contained 67 SNPs and 298 INDELs for 2020T and 48

SNPs and 328 INDELs for 2000B (Supplementary Table 3). We observed a strong enrichment of candidate variant calls in repetitive and homopolymer regions (e.g., for both 2020 T and 2000B about 97% of INDELs were in tandem repeat or homopolymeric regions; see Supplementary Table 4), suggesting the presence of false-positive calls despite the application of strict filtering criteria. For discovery of SVs, we employed Sniffles2 [52]. We performed manual inspection of SV candidates to remove false-positive calls in IGV (Materials and Methods), which reduced the total size of the Sniffles2-based SV call set from 116 to 2 for 2020 T and from 36 to 3 in 2000B (Supplementary Tables 5 and Supplementary Table 6). Post-filtering, we identified 367 small and large variants (8.7% of the total variants called) in 2020 T and 379 variants (18% of the total variants called) in 2000B. Together, these variants comprised 693 unique variants across 669 unique loci, including SNPs, small indels, and structural variants.

Discovery of a large structural variant in the ROP8-ROP2A locus by manual curation

During manual inspection of SV calls for isolate 2000B, we observed an enrichment of SV calls with low reported allele frequencies in the ROP8-ROP2A region on Chromosome X; while these calls were individually rejected during manual inspection, the pattern of the aligned reads, with the presence of >4,000 bp of inserted sequence and exhibiting high frequencies of clipping and supplementary alignments, suggested the presence of a large structural variant that was not correctly resolved by Sniffles2 (Figure S5). Attempts at targeted re-assembly of the locus using standard assembly algorithms were not successful. However, we were able to manually assemble the locus by first determining its overall sequence structure based on overlaps between individual long reads ("read stitching"), followed by polishing (Supplementary Notes), revealing the presence of a 13 kb insertion. To confirm the existence of this 13 kb sequence in 2000B, we modified the 2015T assembly by incorporating the insertion using bcftools consensus. Alignment of 2000B reads to the modified reference showed a uniform coverage pattern without significant read clipping, showing that we had correctly resolved the structural variant (Figure S5). An initial inspection of the inserted sequence suggested that it contained additional copies of ROP8- and ROP2A-related genes (see below). Furthermore, visual inspection of the region in 2020 T also indicated the likely presence of a structural variant relative to the 2015T assembly (Figure S6); however, due to the shorter read lengths of 2020 T, we could not manually resolve this locus for 2020T. The 13 kb insertion (contig_46:7310767) from the stitched 2000B sequence was included in the SV variant set for subsequent joint genotyping.

To further characterize the sequence content of the inserted sequence in 2000B in comparison to the 2015T reference assembly, we carried out a fine-scale annotation effort, producing a set of consensus annotations based on manual curation of the output of three different tools (Augustus [53], Companion, BLAST; see Methods). We then compared the identified genes to the ME49 ROP2A and ROP8 sequences in ToxoDB and found that isolate 2000B contained three ROP2A-like genes and three ROP8-like genes, similar to the genome assembled by Xia et al., whereas 2015T had one ROP2A-like gene and two ROP8-like genes. Notably, the ME49 ROP2A sequence in ToxoDB is incomplete; gene classification in our assemblies was thus based on the shared sequence components of ROP2A and ROP8 reference genes. Accordingly, we refer to the ROP8-like sequences as ROP8-1, ROP8-2, and ROP8-3, and the ROP2A-like sequences as ROP2A-1, ROP2A-2 A, and ROP2A-2B (Fig. 4A).

Furthermore, ROP8-1 and ROP8-2 were identical between isolates 2015T and 2000B, as was ROP2A-2B. ROP2A-2 A was found only in 2000B, and its sequence was identical to ROP2A-2B (Fig. 4B). Given that ROP8 and ROP2A are members of the ROP2 superfamily in *T. gondii*, our results suggest that the 2015T isolate may harbor three gene copies related to ROP8 and ROP2A, while both the 2000B isolate and the Xia et al. assembly contain six. These findings highlighted a substantial degree of copy number variation at a virulence-related locus of the *T. gondii* ME49 genome. To assess whether the additional gene copies were potentially functional, we mapped *T. gondii* ME49 RNA-seq data obtained from SRA (SRX6428507) to the resolved sequence of the SV from isolate 2000B and found evidence that ROP8-1, ROP8-2, ROP8-3, ROP2A-A, ROP2A-2B and ROP2A-1 were potentially expressed (Figure S8).

Graph-based re-genotyping of 2015T, 2000B and 2020T

To establish a unified variant calling framework for the detection of high-quality isolate- distinguishing variants, we constructed a variation graph with vg [54], using the 2015T assembly as the reference and integrating the set of 670 small and large 2000B and 2020 T candidate post-filtering variant loci. To produce graph-based variant calls, we re-aligned the generated Nanopore sequencing reads to the graph using GraphAligner [55], followed by re-genotyping with vg. For stringent quality control, we removed all positions for which the re-genotyping process did not yield a "reference" call for isolate 2015T, and all positions at which the frequency of the reference allele in isolate 2015T in the aligned reads was below 90%, leading to the removal of 487 loci. We found this step to be empirically necessary to reduce the rate of false-positive variant calls associated with homopolymer and tandem repeat regions, and with systematic differences

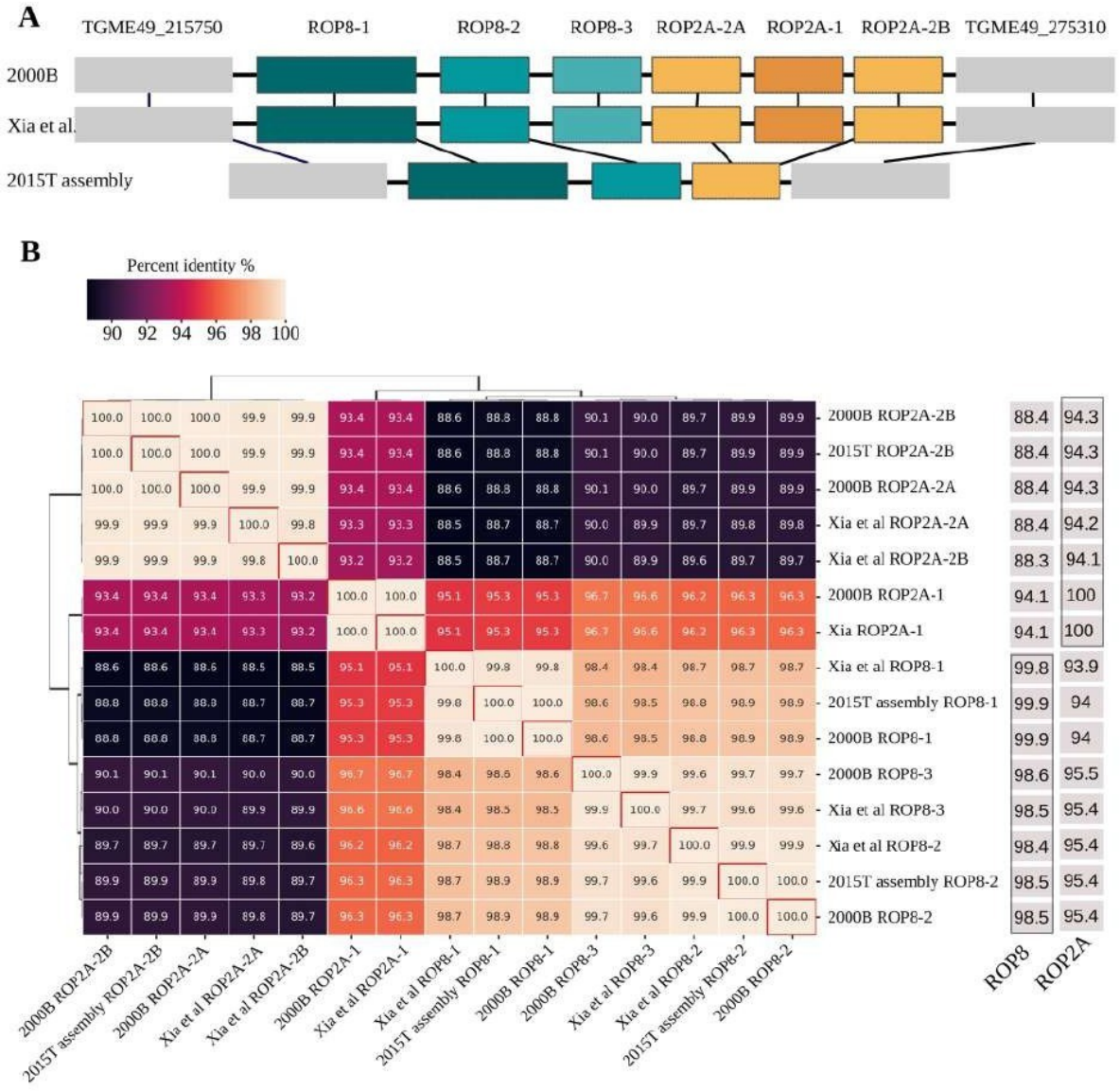
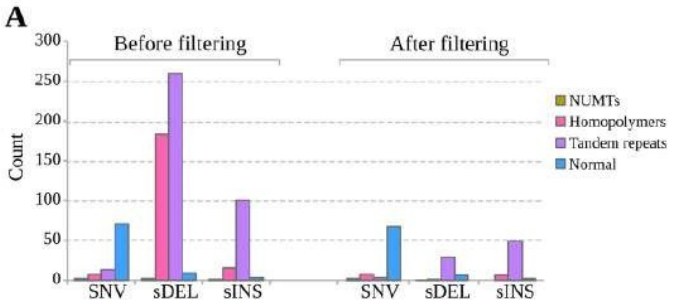


Fig. 4 Structural and sequence variation in the ROP8–ROP2A locus. (**A**) Structural comparison of the ROP8–ROP2A locus across three assemblies: 2015T, 2000B, and the Xia et al. ME49 assembly. The copy number of ROP8 and ROP2A gene copies vary among isolates, with 2020T (not shown) carrying the same configuration as 2000B, as determined by graph-based genotyping. (**B**) Pairwise sequence identity heatmap of ROP8 and ROP2A gene copies across the 2015T, 2000B, and Xia et al. assemblies. Percent identity values are shown, with hierarchical clustering highlighting relationships among paralogs. Sequence identity to canonical ROP8 and ROP2A ME49 reference genes from ToxoDB is shown in the rightmost columns

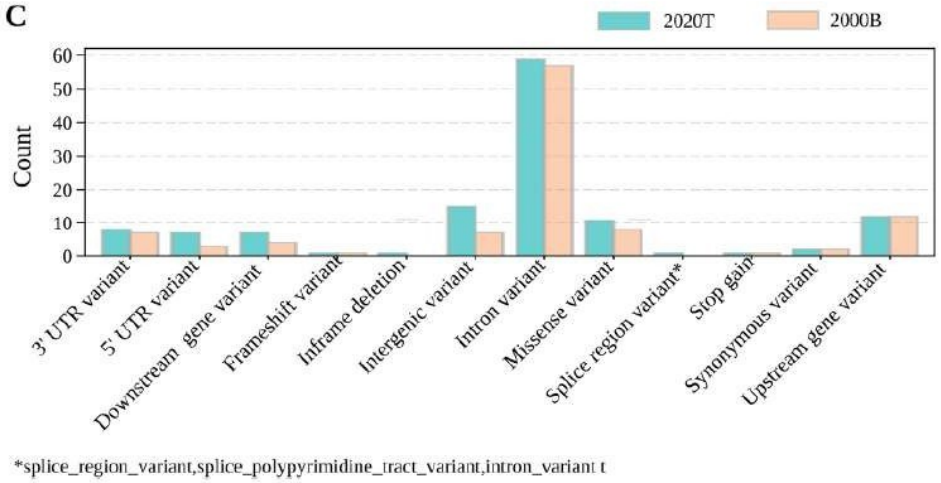
of sequencing read quality between the sequenced isolates; of note, to reduce systematic differences in read quality between the samples during genotyping, we did not use HERRO error correction for the 2015T reads for graph-based genotyping. As a final step of quality control, we manually assessed 32 instances in which the post-filtering graph-based genotypes deviated from the candidate variant callsets used to construct the graph by visual inspection in IGV, and removed 6 loci for which the graph genotype was likely incorrect or for which the true underlying genotype remained ambiguous (see Fig. 5A for variant calls by sequence context before and after filtering). Graph-based genotyping recovered previously undetected SNPs (in one of the samples) and rescued filtered variants, improving the accuracy of genotype determination (see Figure S8).

Our final callset included 177 loci that were genotyped for all 3 isolates; the callset was dominated by



B

Variant type	2020T	2000B
SNV	62	41
sDEL	28	26
sINS	35	35
DEL	2	3
INS	1	1



*splice_region_variant,splice_polypyrimidine_tract_variant,intron_variant t

Fig. 5 Summary of variant filtering and functional annotation. (**A**) Distribution of small variants by sequence context (NUMTs, homopolymers, tandem repeats, and normal regions) before and after filtering. Filtering removed spurious calls in repetitive regions, particularly among small deletions (sDEL). (**B**) Number of detected non-reference alleles in isolates 2000B and 2020T, stratified by variant type. (**C**) Functional annotation of small variants using Ensembl-VEP

SNVs ($n = 79$), followed by small insertions ($n = 57$), and included a total of five structural variants (Supplementary Table 7). The concordance between the graph-based final callsets and the pre-graph candidate variant callsets was generally high; of note, however, the graph-based approach enabled the detection of 16 non-reference alleles not comprised in the candidate variant callset, including the detection of the large insertion in the ROP8-ROP2A region in isolate 2020 T (which was also supported by manual investigation of 2020 T reads aligned to the reference modified by the inclusion of the ROP8-ROP2A SV; data not shown). In total, we found 106 non-reference alleles in isolate 2000B and 128 non-reference alleles in isolate 2020 T (Fig. 5B); when assessing inter-isolate genetic distances based on SNVs, we found that isolates 2000B and 2015T exhibited the highest degree of relatedness (Figure S9).

Functional annotation of small variants

We used Ensembl-VEP [56] to annotate the identified small variants in isolates 2020 T and 2000B for potential functional impact. Almost 50% of small variants were located in introns in both isolates (60 of 128 variants for 2020 T and 57 of 106 variants for 2000B), followed by intergenic variants in 2020 T and variants classified as "upstream gene variant" in 2000B (Fig. 5C). 10 variants in 2000B and 14 variants in 2020 T were predicted to alter amino acid sequence; in almost all cases, the predicted changes were driven by missense variation, whereas gained stop codons only accounted for a total of 2 instances, 1 specific for each isolate. In 2000B, the stop-gained variant was found in the 6-phosphogluconate dehydrogenase protein; deletion mutants of this gene were found to be associated with severely attenuated virulence in vivo [57]. In 2020 T, the detected stop-gained event was located in the AP2 domain transcription factor

Table 1 Overview of non-synonymous small variants in isolates 2000B and 2020T

Contig	Position	Type	Gene	Predicted effect	2020T	2000B	Predicted gene product
Contig_11	4,845,879	SNV	TgME49_XII0073200	Missense variant	1	0	Hypothetical protein
Contig_11	1,499,544	SNV	TgME49_XII0024800	Missense variant	1	1	Hypothetical protein
Contig_20	185,661	SNV	TgME49_VI0008200	Missense variant	1	0	Guanylate cyclase organizer (UGO)
Contig_20	2,140,215	SNV	TgME49_VI0036200	Stop gained	0	1	6-Phosphogluconate dehydrogenase
Contig_20	3,137,120	SNV	TgME49_VI0050600	Missense variant	1	0	Microneme-like protein
Contig_20	1,235,798	SNV	TgME49_VI0022100	Missense variant	1	1	Conoid gliding protein (Cgp)
Contig_24	2,197,273	SNV	TgME49_III0038200	Missense variant	1	1	Cleft lip and palate transmembrane protein 1 (Clptm1)
Contig_30	2,193,723	SNV	TgME49_IX0037400	Missense variant	1	0	Heat repeat-containing Protein
Contig_30	491,285	INS	TgME49_IX0011000	Frameshift variant	0	1	Hypothetical protein
Contig_42	4,341,951	SNV	TgME49_VIIa0066400	Missense variant	0	1	Phosphatidate cytidylyltransferase
Contig_43	3,772,043	SNV	TgME49_VIII0061200	Missense variant	1	1	Hypothetical protein
Contig_43	3,493,540	SNV	TgME49_VIII0057400	Missense variant	1	1	Chromodomain helicase DNA binding protein Chd1/Swi2/Snf2
Contig_43	8,031,232	DEL	TgME49_VIII0121700	Inframe deletion	1	0	Hypothetical protein
Contig_46	6,422,637	SNV	TgME49_X0090900	Missense variant	1	0	Hypothetical protein
Contig_46	5,243,975	SNV	TgME49_X0073000	Missense variant	1	1	Ulk kinase
Contig_46	1,929,612	SNV	TgME49_X0029600	Missense variant	0	1	PolyA polymerase
Contig_46	6,778,532	SNV	TgME49_X0096700	Stop gained	1	0	Ap2 domain transcription factor Ap2x-8
Contig_46	1,046,388	SNV	TgME49_X0016800	Frameshift variant	1	0	Hydrolase, Nudix family protein

Table 2 Overview of structural variants in isolates 2000B and 2020T

Contig	Position	Type	SV length	Gene	Predicted effect	2020T	2000B	Predicted gene product
Contig_46	7,310,767	INS	12,896	-	Expansion in ROP8-ROP2A locus	1	1	-
Contig_2	2,235,315	DEL	520	TgME49_IV0036400	Exon removal of a highly repetitive gene	0	1	Hypothetical protein
Contig_42	608,620	DEL	52	TgME49_VIIa0014000	Shortening of an AT-rich repeat region in intron	0	1	Histidine acid phosphatase superfamily protein
Contig_35	1,340,294	DEL	1180	TgME49_IX0117400	Exon removal of a single exon gene	1	0	Hypothetical protein
Contig_35	699,700	DEL	203	TgME49_IX0109000	Frameshift deletion	1	1	Pt repeat containing protein

AP2X-8, a member of the ApiAP2 family. This family plays a critical role in the biology of *T. gondii*, particularly in regulating gene expression that governs life cycle transitions and developmental stages [58]. Of 18 non-synonymous small variants detected across both isolates, six were located in hypothetical proteins; other affected genes included e.g. Microneme-like protein and Conoid gliding protein (Table 1).

Functional annotation of structural variants

In addition to the large insertion in the ROP8-ROP2A region (see above), the graph callset included four additional deletions (Table 2), ranging in size from 52 to 1,180 bp. Three of these deletions were located in or affected exonic regions of predicted genes, including (i) a deletion specific to 2000B, occurring in the predicted overlapping hypothetical proteins TgME49_IV0036400.1 and TgME49_IV0036400.2 and leading to the loss of exon 4 in both cases; of note, the exons of TgME49_IV0036400.2 exhibited a complex repeat structure with a high degree of inter-exon homology (Figure S10); (ii) the likely inactivation of a hypothetical protein-encoding gene (TgME49_IX0117400), located within a cluster of orthologous genes predicted by Companion, due to the deletion of its single exon; and (iii) a deletion associated with a frameshift event in a "PT repeat"-annotated gene (TgME49_IX0109000); of note, this deletion was present in both 2020 T and 2000B. The fourth detected deletion was located in an intronic region of a "histidine acid phosphatase superfamily"-annotated gene (TgME49_VIIa0014000) and was associated with the shortening of an AT-rich repeat region. In conclusion, in addition to differences at the level of small variants, the sequenced isolates also exhibited genetic differences at the structural level.

Discussion

We used Oxford Nanopore long-read sequencing to generate a new *T. gondii* ME49 reference genome and to characterize genetic variations between three independently maintained *T. gondii* ME49 isolates. Our analyses demonstrated that our new reference genome exceeded

other available *T. gondii* ME49 reference genomes in terms of contiguity and accuracy, as well as the existence of significant within-strain variation at the level of both large and small variants. Furthermore, functional annotation pointed to potential effects of the detected variants on virulence and life cycle transition phenotypes, suggesting a potentially confounding role in some of the experimental systems in which *T. gondii* ME49 is used as a model organism.

In our analysis, we identified both small and large genetic variants, including a notable copy number variation (CNV) in the ROP8-ROP2A locus, a potentially virulence-associated region in *T. gondii* [59, 60]. Among the three sequenced isolates, 2015T harbored only three copies, whereas 2000B (based on manual construction of the locus) and 2020 T (based on graph genotyping) contained six copies. A previous study showed that this locus is expanded in *T. gondii*, while it is entirely absent in its close relatives *Neospora caninum* and *Hammondia hammondi* [61]. Short-read sequencing data from reference *T. gondii* strains (GT1, ME49, VEG) indicated that these harbor six copies of ROP8/ROP2A genes [61], suggesting that this configuration is ancestral and that the reduced copy number in 2015T likely resulted from a deletion event. Additionally, we identified four large deletions, of which one was shared between isolates 2000B and 2020 T, and 172 small variants, the majority of which were SNVs.

While we did not carry out any functional studies, a phenotypic impact of at least some of the detected variants seems plausible. First, the ROP8-ROP2A structural variant was associated with a difference in the total number of ROP8-ROP2A genes from three to six, with three of the additional gene copies being potentially expressed. Although a complete deletion of this locus in the RH strain did not show any apparent effects on tachyzoite growth in vitro [62], the ROP2 family (of which ROP8 is a member) is generally considered a key virulence factor [60, 63, 64]. ROP2A and ROP8 were found to interact with ROP18 and GRA7, two proteins that disrupt immunity-related GTPases and act synergistically to regulate acute virulence in mice [59]. Additionally, previous studies reported that ROP8 was upregulated during the tachyzoite-to-bradyzoite transition in the M4 strain [65], while ROP2A was upregulated in association with sulfadiazine resistance in type II strains [66]. While further functional studies are necessary to determine the impact of the ROP8-ROP2A structural variant, the difference in the combined ROP8-ROP2A gene dosage may plausibly affect *T. gondii* virulence and/or life cycle transitions. Second, additional structural variants were associated with the deletion of entire exons in genes *TgME49_IV0036400* and *TgME49_IX0117400* annotated as encoding hypothetical proteins, a predicted frameshift mutation in *TgME49_IX0109000* hypothetical protein, and the disruption of a tandem repeat stretch in an intron in the gene *TgME49_VIIa0014000* annotated as the histidine acid phosphatase superfamily protein. If the corresponding hypothetical proteins are expressed, a phenotypic effect of these variants seems likely. Notably, histidine acid phosphatase superfamily protein was predicted to play a role in host invasion based on ToxoNet, a high-confidence *T. gondii* protein-protein interaction map [67]. However, the functional impact of an intronic deletion in this gene remains unclear. Lastly, among the 18 non-synonymous variants we detected, several affect proteins involved in metabolic processes (e.g., 6-phosphogluconate dehydrogenase [57], phosphatidate cytidylyltransferase [68], NUDIX hydrolase [69], transcriptional regulation and chromatin remodeling (e.g., Ap2 domain transcription factor Ap2x-8 [58], Chromodomain Helicase DNA Binding Protein Chd1/Swi2/Snf2 [70], cellular signaling (e.g., Guanylate Cyclase Organizer [71], Ulk Kinase [72], and parasite motility (e.g. Conoid Gliding Protein [73]. These findings suggest that the detected variations could have functional consequences across multiple biological pathways, including metabolism, gene regulation, host-cell interaction, and intracellular signaling. Furthermore, six of the detected small non-synonymous variants were annotated as affecting genes encoding hypothetical proteins, highlighting that the substantial proportion of the *T. gondii* proteome that remains uncharacterized is a challenge for functional interpretation across variant classes.

While we unambiguously detected the presence of genetic variation, we did not attempt to determine when these variants arose in the propagation history of the sequenced isolates or whether their emergence was driven by neutral genetic drift or selection. In terms of selective pressure, the propagation environments of isolates 2015T and 2020T were similar to each other, because they were maintained in vitro in cell culture. By contrast, isolate 2000B was maintained in vivo and therefore under immune system pressure from the murine host. Additionally, 2015T and 2020 T were acquired relatively recently from ATCC (the exact acquisition date for 2015T is unknown). Notably, however, pairwise genetic distances, measured by shared SNV alleles, were relatively similar across all isolates, with 2015T and 2000B unexpectedly exhibiting the highest pairwise genetic similarity. It is possible—though not testable within the scope of this study—that a proportion of isolate-distinguishing genetic variants were already present when the isolates were obtained from ATCC. In conclusion, it is possible that isolates 2015T and 2020T do not share a relatively recent common ancestor, and the evolutionary relationship between the characterized isolates remains unresolved.

Key methodological innovations of our study in the field of *T. gondii* genomics include the use of high-quality Oxford Nanopore R10 long-read sequencing and the utilization of a pangenomic approach for joint genotyping. With recent improvements in base calling and flow cell chemistry, Nanopore technology has achieved a raw-read accuracy > 99% [74], and near-perfect genome assemblies have already been demonstrated for bacterial genomes [37]. In the context of our study, high-quality Oxford Nanopore sequencing, particularly with R10.4.1 chemistry and HERRO error correction, enabled the assembly of the most contiguous and complete ME49 genome to date; compared to the Xia et al. assembly, which was also based on Nanopore sequencing data but generated with older R9 flow cells, this was particularly evident from the increased proportion of perfectly aligned ToxoDB genes, indicating higher consensus accuracy. Joint graph-based genotyping was instrumental in the reliable detection of isolate-distinguishing genetic variation, including the identification of the ROP8-ROP2A structural variant in isolate 2020 T, the rescue of individual variants that would have otherwise been removed (e.g., due to quality or coverage filters), and the detection of potential false-positive variant calls in isolates 2000B and 2020 T by filtering for positions with a reference allele frequency of < 90% in isolate 2015T. However, even with the most recent Nanopore flow cell chemistry, base calling algorithms, and sophisticated filtering approaches, distinguishing between true- and false-positive variants in homopolymeric and tandem repeat regions remained a challenge, due to an increased error rate of Nanopore sequencing in these regions [75]. While rigorous filtering was therefore essential to ensure the most reliable variant call, homopolymeric and tandem repeat regions are known to exhibit increased mutation rates in many species [76, 77] and it is likely that our filtering approach led to the removal of some true-positive variants in these regions.

In addition to representing the highest-quality assembly for *T. gondii* strain ME49, our reference genome is also among the highest-quality *T. gondii* assemblies for any strain currently available [46, 78]. Reference genomes play important roles in e.g. the analysis of RNA-seq data and short-read-based variant calling [35], and ME49 is often used as a reference for *T. gondii* type II strains [79, 80]. The reference genome produced by us can therefore contribute to improved transcriptomic and population genetic characterization of *T. gondii*. However, some limitations remain. First, approximately 80 kb of sequence content from the Xia et al. assembly were not recovered in our assembly; further investigation is needed to determine whether this represents truly missing sequence or an assembly artifact in Xia et al. Second, due to its highly repetitive structure, the rDNA region remained challenging to assemble [81]. Spanning the complete locus would require, based on an estimate of 110 rDNA tandem array copies per *T. gondii* genome [82], a read of approximately 1 Mbp in size; while such read lengths can be achieved with Nanopore sequencing, they typically require further protocol optimization for obtaining ultra-long reads [83, 84]. Based on the longest rDNA-containing read in our dataset, we could, however, confirm that *T. gondii* ME49 rDNA arrays follow the - generally less common - "L-type" conformation, in which the 5 S gene is linked to (8 S–5.8 S–28 S) rDNA array units [81]. In contrast to an earlier study [82], we found that the 5 S gene was encoded on the opposite strand relative to the other rDNA genes. Furthermore, we found that the *T. gondii rDNA* repeat unit measured approximately 8.3 kb in size, consistent with reports for strains P and RH (GenBank accession X75453.1 and X75429.1, respectively), and detected an inversion in the rDNA repeat array, which may also exist in humans [85]. Third, our assembly did not contain a resolved mitochondrial genome; this may be due to the as-yet incompletely understood, and possibly fragmented, genomic architecture of mitochondrial DNA in *T. gondii* [86]. Last, we did not phenotypicaly characterize the *T. gondii* isolated we sequenced in this study. We acknowledge that this is an important direction for a future research and a requirement for better understanding the functional impact of the variants detected.

Conclusion

In conclusion, we generated a new reference genome for *T. gondii* ME49 and demonstrated the existence of genetic divergence of potential phenotypic importance within the ME49 strain at the level of both large and small genetic variants. In addition to the specific findings for *T. gondii* ME49, our study highlights the importance of continuous monitoring of the genetic identity and stability of model organisms and shows how this aim can be achieved with modern high-accuracy long-read sequencing platforms.

Methods

Parasite cell culture

T. gondii ME49 isolates 2015T and 2020 T were maintained by serial passaging of tachyzoites in human foreskin fibroblasts (HFFs) Hs27. HFFs were cultured in Iscove Modified Dulbecco Media (IMDM) containing 10% fetal bovin serum (FBS) at 37 °C in a 5% CO2 incubator. The isolates were obtained from ATCC under the same catalog number (50611) (refer to the Supplementary Notes for provenance details about the isolates).

Animal experiments

CD1 mice were obtained from Charles River Breeding Laboratories and were used to maintain and propagate ME49 *T. gondii* (approved by the German

authorities (LAVE, formerly known as LANUV), permit no. 81-02.04.04.04.04.40.23.VG055) to obtain cysts for infection studies (approved by the German authorities (LAVE, formerly known as LANUV), permits no. 81-02.04.04.04.04.2021.A060, 81-02.0.0.04.04.2020.A113, 81-02.04.04.04.04.2018.A406). All animal experiments were conducted in strict accordance with the German Animal Welfare Act. All applicable international, national, and institutional guidelines for the care and use of animals were followed. The cysts used in this study were surplus cysts not needed for propagation or used in the infection studies mentioned above.

Following the approved protocol, mice were infected by i.p. injection of the equivalent of 10 *T. gondii* ME49 cysts in a volume of 200 µl PBS/g body weight: For isolation of cysts, animals were sacrificed by cervical dislocation and cysts were isolated from the brain as described below. Pelleted cysts were washed with PBS and used for analysis (this study) or counted and lysed with Trypsin-EDTA to obtain free bradyzoites. Lysis was stopped with the addition of heat-inactivated (56 °C, 30 min) fetal calf serum (FCS, PAN-Biotech, Germany). After a final wash with PBS followed by centrifugation (660 x g, 10 min, RT), ME49 bradyzoites were resuspended in PBS and injected.

T. gondii cyst preparation

T. gondii ME49 2000B cysts were isolated from the brains of CD1 mice via Ficoll-Paque gradient centrifugation 11 to 19 weeks post-infection. Briefly, mouse cerebrum tissue was homogenized by sequentially passing through progressively smaller cannulas (smallest gauge: 23G). The homogenate was first centrifuged at 130 x g for 5 min at room temperature (RT). The resulting pellet was resuspended in 20 ml of sterile PBS. To separate cysts, 10 ml of Ficoll-Paque Plus (GE Healthcare, USA) was carefully layered below the PBS suspension, followed by centrifugation at 1,250 x g for 25 min at room temperature, without using brakes. The cysts were then collected, washed in PBS, and stored at either 4 °C (for short-term nucleic acid extraction) or −80 °C (for long-term nucleic acid extraction) until further processing.

HMW DNA extraction

T. gondii ME49 tachyzoites (2015T and 2020 T) were cultured in a confluent HFF monolayer for approximately 7 days, until the monolayer was fully infected and the tachyzoites had lysed the host cells to emerge. The medium containing the released tachyzoites was collected for DNA extraction. For 2000B, DNA was extracted from the collected cysts. Extraction for isolate 2020 T and 2000B was done using Phenol-Chloroform Extraction (see below). In contrast, extraction for the 2015T isolate was performed using one of the following methods in an effort to obtain high-quality, non-fragmented DNA for genome assembly (Supplementary Table 1).

1. Monarch HMW DNA Extraction Kit (NEB, Cat# T3050L): DNA was extracted using the Monarch HMW DNA Extraction Kit according to the manufacturer's protocol.
2. Size Selection via Semi-Selective DNA Precipitation after extraction: DNA was size-selected to deplete short fragments using a semi-selective DNA precipitation method, following the August 2021 version of the protocol from Oxford Nanopore Technologies [87].
3. Ultra-Long DNA Sequencing: For ultra-long DNA sequencing: The Ultra-Long DNA Sequencing Kit (SQK-ULK114) was used according to the November 2022 ONT protocol [84].

Phenol-chloroform extraction for gDNA 2020T and 2000B samples

Genomic DNA was purified using a phenol-chloroform extraction protocol. Briefly, *T. gondii* cells (5 × 10^7) were washed in PBS and resuspended in a digestive solution containing 500 µl TNE, 50 µl 10% SDS, and 7.5 µl Proteinase K (10 mg/ml). The suspension was incubated overnight at 56 °C with shaking. Following lysis, 500 µl of chloroform/isoamyl alcohol (24:1) was added, mixed vigorously and centrifuged at 13,000 rpm for 5 min at room temperature. The supernatant was transferred to a new tube, and the extraction was repeated. DNA was precipitated with 500 µl isopropanol, incubated for 5 min at room temperature, and centrifuged at 8,000 rpm for 10 min. The DNA pellet was washed three times with 500 µl of 70% ethanol, dried for 10 min at room temperature, and resuspended in 100 µl TE buffer (incubated for 1 h at 56 °C, then overnight at 4 °C).

Oxford nanopore sequencing

Data acquisition for the three different *T. gondii* isolates was performed using Oxford Nanopore sequencing (R10.3, R10.4 or R10.4.1) from multiple samples with the aim of achieving at least 20X coverage for variant calling and sufficient coverage for generating a high quality genome assembly. For 2015T, base-calling was performed using Dorado v0.7.3 [40] with the command dorado duplex. Error correction on simplex reads was then carried out using HERRO [41] independently for the data generated by each sequencing run. After correction, simplex and duplex reads were combined. For 2020 T and 2000B, all sequencing runs were basecalled using Guppy v6.1.5 [42] For details about the sequencing and basecalling of each sample, see Supplementary Table 1.

Toxoplasma gondii ME49 genome assembly

To eliminate human DNA contamination, the sequencing data were filtered by aligning the reads against a combined *T. gondii* ME49 (GenBank: JACEHA000000000.1) (Xia et al., 2021), human (GRCh38 Primary Assembly, RefSeq: GCF_000001405.26) and mouse (GRCm39 Genome Assembly, RefSeq: GCF_000001635.27) reference genomes using minimap2 v2.28-r1209 [88]. Reads mapping to the human or mouse genomes were removed, and the remaining reads were extracted from the FASTQ file using Seqkit v0.15.0 [89] with the command seqkit grep -v. The filtered FASTQ files from all 2015T sequencing runs were combined and filtered to remove mitochondrial reads. This filtering was performed by aligning the reads against a FASTA file containing mitochondrial sequence blocks and mitochondrial cytochrome genes [86]. Reads with 40% or more of their length covered by mitochondrial sequence blocks were removed from the FASTQ file using a custom Python script. The remaining sequencing reads were then used to assemble the genome with Flye v2.9.4-b1799 [43] using the --nano-corr parameter. One circular contig likely associated with contaminant DNA from the bacterial species *Facklamia ignava* was removed.

Gene annotation of the 2015T assembly

To annotate the new assembly, the Companion web tool v2.2.0 was employed [45], with default settings and utilizing the *T. gondii* ME49 reference available in the Companion database. The output GFF3 file was then modified to ensure compatibility with subsequent gene annotation tasks, using gffread with the options -O -F -E --force-exons. Transcript biotype information, including rRNA, tRNA, pseudogenic transcript, and protein-coding (mRNA), was incorporated into the GFF3 file using a custom script. Finally, the GFF3 file was compressed using bgzip and indexed with Tabix.

Comparison of the 2015T assembly to the Xia et al. Assembly

To compare the 2015T assembly with the Xia et al. assembly, we aligned the two genomes using nucmer from the MUMmer4 package [47] with default parameters, followed by dnadiff. Chromosome mapping between assemblies was performed using minimap2 with parameters -x asm20 -m 10,000 -z 10000,50 -r 50,000 --end-bonus = 100 --secondary = no -a -t 20 -- eqx -Y -O 5,56 -E 4,1 -B 5. Dot plots were generated using D-Genesis [90] for both nuclear and apicoplast sequences to visualize large-scale structural differences. To obtain reference gene sequences for *T. gondii* from ToxoDB v48, gene coordinates were extracted from the ToxoDB-68_TgondiiME49.gff annotation file and converted to BED format. The corresponding sequences were retrieved from the ToxoDB-68_TgondiiME49_Genome fasta file using bedtools getfasta. The final dataset contained 8,778 gene sequences. Gene sequences were then mapped to the assembly using minimap2 map-ont, and a custom Python script was used to count the number of mapped genes and calculate the fraction of perfectly mapped genes. Additionally, assembly anomalies were identified by mapping the reads back to the assembly and inspecting the alignments in IGV v2.17.4 [49] (see Supplementary Note: Identification of Assembly Anomalies).

Analysis of the rDNA region

The breakpoint in the 2015T assembly occurred in Chromosome IX, within a repetitive rDNA region that was misassembled in the Xia et al. assembly. The longest read from our dataset that mapped to this position (JACEHA010000011.1:1376356–1526114) in the Xia et al. assembly was 207,813 bp in length and spanned positions 1,387,207 to 1,515,512. The read was extracted using seqkit and aligned, using minimap2 map-ont, to the longest available rRNA sequences for the 28 S, 5.8 S, and 18 S genes of *T. gondii* ME49 (NCBI: XR_001974492.1, XR_001974334.1, and XR_001974441.1), as well as to the 5 S rRNA sequence from the RH strain (GenBank: 5XXB_3). The alignment coordinates of these sequences on the longest read were visualized using a custom Python script. Similarly, the sequence from the Xia et al. assembly at position (JACEHA010000011.1:1276356–1626114) was extracted and aligned to the rRNA sequences, with the alignment coordinates visualized using the custom Python script.

Calling of SNPs and indels

2020 T and 2000B isolate FASTQ files were filtered to remove contamination from human or mouse sources, employing the same approach that was also used for the 2015T reads. Single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) in 2000B and 2020 T were called separately for each sample using Medaka (v2.0.1) [51] with the medaka_variant script and using the 2015T assembly as a reference genome. Variant calling was carried out using the Medaka model r104_e81_sup_variant_g610, which was empirically found to reduce the number of false-positive variant calls in isolate 2000B compared to more recent Medaka models. Variants were annotated and classified using the varType command of SnpSift v5.1 [91]. Only variants located on nuclear chromosomes were considered. We applied a stringent filtering approach to Medaka-identified variants based on sequencing depth, genotype quality, and deletion overlap. Variants were retained if the sequencing depth (DP) was ≥ 10 and the genotype quality (GQ) was ≥ 10. Additionally, for SNPs, we assessed the proportion of deletions at the variant position; SNPs at positions where deletions

accounted for ≥ 50% of the total depth were removed. Additionally, we removed small variants that overlapped with structural variants (SVs), occurred at the very end or beginning of the chromosome where assembly quality is potentially suboptimal, or that were located at previously identified erroneous or suspicious positions.

Calling SVs

For each sample, Sniffles v2.2 [52] was used to identify structural variants. To ensure the VCF file conformed to the VCF version 4.1 specifications [92], a custom script was developed to adjust SV calls by decrementing the variant position and updating the reference and alternate alleles for insertions and deletions, accordingly. All called SVs were then manually evaluated in IGV v2.17.4. Variants were retained if they were clearly supported by long reads spanning the entire variant region, such that the structural change was directly observable in the read alignments or supported by consistent and extensive read clipping. SVs were excluded if they were located on the apicoplast contig included from the Xia et al. assembly, corresponded to translocations resulting from incomplete chromosome assembly, or lacked strong read support. In addition, SVs were also excluded if they overlapped regions potentially enriched for assembly artifacts—such as contig ends—or if they fell within nuclear mitochondrial DNA segments (NUMT) regions, in which misaligned mitochondrial-origin reads commonly produced spurious signals. Filtering results for all SV are summarized in Supplementary Table (Sniffles manual filtering).

ROP8-ROP2A locus annotation

Our objective was to characterize the number and order of ROP2A and ROP8 gene copies within the RO8-ROP2A locus of three assemblies: Xia et al. assembly, our 2015T assembly, and the sequence of the locus in isolate 2000B. Inspecting Companion-based gene annotations for this locus, we found that Companion classified all ROP2A sequences as pseudogenes, likely driven by the incompleteness of the ROP2A reference sequence obtained from ToxoDB. We thus employed BLASTn and Augustus [53] in addition to Companion for the characterization of the ROP8-ROP2A. We first used BLASTn v2.14.1 + to obtain candidate ROP8 and ROP2A sequences from the assemblies. We mapped the reference ROP2A (*TGME49_215785*) and ROP8 (*TGME49_215775*) sequences from ToxoDB (release 68) to the assemblies, extracted the sequence coordinates of the BLAST hits, and merged the coordinates of ROP8 and ROP2A hits whenever they overlapped; finally, we extracted the sequences corresponding to the merged coordinate sets from the assemblies. In addition, we ran Augustus (using --species = toxoplasma --strand = both --genemodel = complete --codingseq = on) on the assembly sequences and extracted the sequences of the detected genes. To create a merged set of candidate gene sequences, we combined the BLAST- and Augustus-based sets, merging BLAST- and Augustus-derived sequences whenever they overlapped on the assembly they were extracted from. In a final step, we cross-checked the merged set against Companion to ensure completeness; no additional sequences were added during this step.

To determine sequence homology between the extracted candidate gene sequences, and to determine homology between the candidate gene sequences and ROP2A and ROP8 reference sequences, we computed a multiple sequence alignment (MSA) using Clustal Omega v1.2.4 [93], including (i) the candidate gene sequences from the investigated assemblies; (ii) the ToxoDB reference sequences for ROP2A and ROP8; and (iii) the Augustus-predicted gene sequences from the investigated assemblies (i.e., Augustus-based sequences prior to merging with BLAST hits), as these were informative about the positions of start and stop codons within the extracted sequences. We projected start and stop codon positions from Augustus- and ToxoDB-derived sequences into the MSA and obtained MSA interval (positions 409–2122) corresponding to the union of the implied open reading frames; pairwise sequence similarities between the extracted candidate gene sequences were computed based on this MSA interval. In addition, we projected the start and stop positions of ROP2A and ROP8 ToxoDB sequences into the MSA and intersected the implied MSA coordinates (positions 409–1700); assignment of the candidate gene sequences to either ROP2A or ROP8 was based on pairwise sequence similarity between the individual candidate gene sequences and ROP2A/ROP8 ToxoDB reference sequences over this interval. Of note, the end of the latter range (MSA position 1700) corresponded to the end of the partial ROP2A sequence, which is not complete.

Similarity matrices and heatmaps were generated using a custom Python script. To assess which of the identified candidate gene sequences were potentially functional, we mapped Toxoplasma ME49 RNA-seq sequencing data retrieved from the NCBI Sequence Read Archive (SRA) using prefetch (SRX6428507) and converted them to FASTQ format with fasterq-dump. Paired-end reads from SRR9667927 and SRR9667928 were merged to create a combined dataset. The merged reads were then mapped to the stitched 2000B sequence using STAR v2.7.10a [94] with the following parameters: --alignIntronMax 10,000 and --alignMatesGapMax 1000 --outMultimapperOrder Random.

Graph construction and genotyping from the graph

A genome graph was constructed using the vg toolkit v1.63.1 (75) based on a catalog of previously identified SNPs, indels, and structural variants (SVs). Variant calls from Medaka and Sniffles2 were merged using bcftools concat with duplicate removal, followed by position-wise merging across samples using a modified version of the merge_vcfs.py script from PanGenie [95]. The script was adapted to process a haploid genome and handle genotype formats from Sniffles2 and Medaka. The final variant catalog was compressed and indexed with bgzip and tabix. The genome graph was built with vg construct, retaining all alternative paths (-a), and indexed (-L) in xg-format (-x). Long-read sequencing data were aligned to the graph using GraphAligner v1.0.18- [55] employing an identity threshold of 0.75 (--precise clipping 0.75). Alignments were filtered for a minimum mapping quality of 30 (vg filter -q 30), and variant genotyping was performed with vg pack (-Q 20), followed by vg call (-d 1).

Identification of low-complexity genomic regions and NUMTs

Tandem repeats in the 2015T assembly FASTA file were identified using GMATA (v2.3) [96] with the following parameters: motif unit lengths ranging from 2 to 6 nucleotides and a minimum of 5 repeat units. Homopolymers were identified using the homopolymer_finder.py.

script from the umiVar toolkit [97] with a minimum homopolymer length of 4 bases (-l4). The resulting positions were converted into BED format using an AWK script, which grouped consecutive positions into continuous intervals. Positions on apicoplast sequences were excluded from further analysis. NUMTs were detected using RepeatMasker (v4.1.7-p1) [98] with a custom library of mitochondrial sequences [86]. Positions detected on apicoplast sequences, as well as those corresponding to low-complexity regions identified by RepeatMasker's default settings, were excluded. The remaining NUMT coordinates were saved in BED format.

Variant annotation using ensembl variant effect predictor (VEP)

Variants were functionally annotated using Ensembl VEP [56] with the 2015T assembly as the reference genome. The annotation was performed using the Companion-generated GFF3 assembly annotation file generated and the --pick parameter to retain a single, most relevant consequence per variant.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-12372-5.

Supplementary Material 1.
Supplementary Material 2.
Supplementary Material 3.
Supplementary Material 4.
Supplementary Material 5.
Supplementary Material 6.
Supplementary Material 7.
Supplementary Material 8.
Supplementary Material 9.

Acknowledgements

Not applicable.

Authors' contributions

YG performed the genome assembly, developed the analysis strategy. YG and AD wrote the manuscript. YG and MN jointly developed the variant analysis and graph-based genotyping pipeline. LH, DW, JM, KB, and MH prepared parasite samples and generated sequencing data. US, DD, and KP contributed to project discussions and provided feedback on the experimental design and interpretation of results. AD supervised the project and provided critical input throughout. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This study was funded by the Jürgen Manchot Foundation.

Data availability

The raw Oxford Nanopore sequencing data generated and analyzed during this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA1241696. The de novo genome assembly of the Toxoplasma gondii ME49 isolate 2015T has been deposited in GenBank under accession number JBMUND000000000. All custom scripts and computational workflows used in this study are openly available. The genome filtering and assembly pipeline is accessible at [https://github.com/YomnaGohar/T.-gondii_filter_assemble], and the variant calling and graph-based genotyping pipeline is available at [https://github.com/YomnaGohar/ToxoVar]. Detailed instructions for reproducing the main figures are provided at [https://github.com/YomnaGohar/Intra-Strain-Genetic-Heterogeneity-in-Toxoplasma-gondii-ME49].

Declarations

Ethics approval and consent to participate

The study was approved by the § 15 Animal Ethics Committee of the Landesamt für Verbraucherschutz und Ernährung of the state of North-Rhine Westphalia [LAVE], formerly known as Landesamt für Natur, Umwelt und Verbraucherschutz [LANUV], which is responsible for all animal experimentation approvals in North Rhine-Westphalia, including the ethical review and authorization process; permit no. 81-02.04.40.23.VG055). Obtaining of cysts for infection studies was also approved by the German authorities (LAVE, formerly known as LANUV) under permit numbers 81-02.04.2021.A060, 81-02.04.2020.A113, and 81-02.04.2018.A406.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 12 June 2025 / Accepted: 21 November 2025
Published online: 03 December 2025

References

1. Bittihn P, Hasty J, Tsimring LS. Suppression of beneficial mutations in dynamic microbial populations. Phys Rev Lett. 2017;118:028102.
2. Carroll SM, Xue KS, Marx CJ. Laboratory divergence of *Methylobacterium extorquens* AM1 through unintended domestication and past selection for antibiotic resistance. BMC Microbiol. 2014;14:2.
3. Bush ZD, Naftaly AFS, Dinwiddie D, Albers C, Hillers KJ, Libuda DE. Comprehensive detection of structural variation and transposable element differences between wild type laboratory lineages of *C. elegans*. 2023. Available from: http://biorxiv.org/lookup/ https://doi.org/10.1101/2023.01.13.523974 [cited 8 Apr 2025]
4. Quevedo R, Smirnov P, Tkachuk D, Ho C, El-Hachem N, Safikhani Z, et al. Assessment of genetic drift in large pharmacogenomic studies. Cell Syst. 2020;11:393-401.e2.
5. Almodovar AJO, Luther RJ, Stonebrook CL, Wood PA. Genomic structure and genetic drift in C57BL/6 congenic metabolic mutant mice. Mol Genet Metab. 2013;110:396–400.
6. Selman C, Swindell WR. Putting a strain on diversity. EMBO J. 2018;37:e100862.
7. Fanning SL, Appel MY, Berger SA, Korngold R, Friedman TM. The immunological impact of genetic drift in the B10.BR congenic inbred mouse strain. J Immunol. 2009;183:4261–72.
8. López-Cortegano E, Chebib J, Jonas A, Vock A, Künzel S, Tautz D, et al. Variation in the spectrum of new mutations among inbred strains of mice. Mol Biol Evol. 2024;41:msae163.
9. Houot B, Svetec N, Godoy-Herrera R, Ferveur J-F. Effect of laboratory acclimation on the variation of reproduction-related characters in *Drosophila melanogaster*. J Exp Biol. 2010;213:2322–31.
10. Gloria-Soria A, Soghigian J, Kellner D, Powell JR. Genetic diversity of laboratory strains and implications for research: the case of Aedes aegypti. PLoS Negl Trop Dis. 2019;13:e0007930.
11. Catoiu EA, Phaneuf P, Monk J, Palsson BO. Whole-genome sequences from wild-type and laboratory-evolved strains define the alleleome and establish its hallmarks. Proc Natl Acad Sci U S A. 2023;120:e2218835120.
12. Bradley MD, Neu D, Bahar F, Welch RD. Inter-laboratory evolution of a model organism and its epistatic effects on mutagenesis screens. Sci Rep. 2016;6:38001.
13. Wassenaar TM, Zimmermann K. How industrial bacterial cultures can be kept stable over time. Lett Appl Microbiol. 2020;71:220–8.
14. Weber KP, De S, Kozarewa I, Turner DJ, Babu MM, De Bono M. Whole genome sequencing highlights genetic changes associated with laboratory domestication of C. elegans. Lehner B. Editor PLoS ONE. 2010;5:e13922.
15. Andrews PW. Human pluripotent stem cells: genetic instability or stability? Regen Med. 2021;16:113–5.
16. He Z, Wilson A, Rich F, Kenwright D, Stevens A, Low YS, et al. Chromosomal instability and its effect on cell lines. Cancer Reports. 2023;6:e1822.
17. Briske-Anderson MJ, Finley JW, Newman SM. The influence of culture time and passage number on the morphological and physiological development of Caco-2 cells. Exp Biol Med (Maywood). 1997;214:248–57.
18. Wenger SL, Senft JR, Sargent LM, Bamezai R, Bairwa N, Grant SG. Comparison of established cell lines at different passages by karyotype and comparative genomic hybridization. Biosci Rep. 2004;24:631–9.
19. Jutzeler KS, Platt RN, Diaz R, Morales M, Clec'h WL, Chevalier FD et al. Abundant genetic variation is retained in many laboratory schistosome populations. 2024. Available from:http://biorxiv.org/lookup/ https://doi.org/10.1101/2024.10.21.619418 [cited 20 Mar 2025]
20. Arslan A, Fang Z, Wang M, Tan Y, Cheng Z, Chen X, et al. Analysis of structural variation among inbred mouse strains. BMC Genomics. 2023;24:97.
21. Vincent MS, Uphoff S. Bacterial phenotypic heterogeneity in DNA repair and mutagenesis. Biochem Soc Trans. 2020;48:451–62.
22. Lujan SA, Williams JS, Pursell ZF, Abdulovic-Cui AA, Clark AB, Nick McElhinny SA, et al. Mismatch repair balances leading and lagging strand DNA replication fidelity. PLoS Genet. 2012;8:e1003016.
23. Chang HHY, Pannunzio NR, Adachi N, Lieber MR. Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat Rev Mol Cell Biol. 2017;18:495–506.
24. Stevens L, Martínez-Ugalde I, King E, Wagah M, Absolon D, Bancroft R, et al. Ancient diversity in host-parasite interaction genes in a model parasitic nematode. Nat Commun. 2023. https://doi.org/10.1038/s41467-023-43556-w.
25. Woolhouse MEJ, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet. 2002;32:569–77.
26. Nast R, Choepak T, Lüder CGK. Epigenetic control of IFN-γ host responses during infection with *Toxoplasma gondii*. Front Immunol. 2020;11:581241.
27. Kim K, Weiss LM. Toxoplasma: the next 100 years. Microbes Infect. 2008;10:978–84.
28. Sibley LD, Khan A, Ajioka JW, Rosenthal BM. Genetic diversity of *Toxoplasma gondii* in animals and humans. Phil Trans R Soc Lond B Biol Sci. 2009;364:2749–61.
29. Howe DK, Sibley LD. *Toxoplasma gondii*: analysis of different laboratory stocks of the RH strain reveals genetic heterogeneity. Exp Parasitol. 1994;78:242–5.
30. Khan A, Behnke MS, Dunay IR, White MW, Sibley LD. Phenotypic and gene expression changes among clonal type I strains of *Toxoplasma gondii*. Eukaryot Cell. 2009;8:1828–36.
31. Dubey JP, Shen SK, Kwok OC, Frenkel JK. Infection and immunity with the RH strain of *Toxoplasma gondii* in rats and mice. J Parasitol. 1999;85:657–62.
32. Frenkel JK, Dubey JP, Hoff RL. Loss of stages after continuous passage of *Toxoplasma gondii* and *Besnoitia jellisoni* *. J Protozool. 1976;23:421–4.
33. Lindsay DS, Dubey JP, Blagburn BL, Toivio-Kinnucan M. Examination of tissue cyst formation by *Toxoplasma gondii* in cell cultures using bradyzoites, tachyzoites, and sporozoites. J Parasitol. 1991;77:126–32.
34. Fukumoto J, Yamano A, Matsuzaki M, Matsuzaki M, Kyan H, Masatani T, et al. Molecular and biological analysis revealed genetic diversity and high virulence strain of Toxoplasma gondii in Japan. PLoS One. 2020. https://doi.org/10.1371/journal.pone.0227749.
35. Lorenzi H, Khan A, Behnke MS, Namasivayam S, Swapna LS, Hadjithomas M, et al. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic *Toxoplasma gondii* genomes. Nat Commun. 2016;7:10147.
36. Guan P, Sung W-K. Structural variation detection using next-generation sequencing data. Methods. 2016;102:36–49.
37. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, 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;19:823–6.
38. Nayeri T, Sarvi S, Daryani A. Effective factors in the pathogenesis of *Toxoplasma gondii*. Heliyon. 2024;10:e31558.
39. Howe DK, Sibley LD. *Toxoplasma gondii* comprises three clonal lineages: correlation of parasite genotype with human disease. J Infect Dis. 1995;172:1561–6.
40. Dorado. Available from: https://github.com/nanoporetech/dorado [cited 25 Mar 2025]
41. Stanojevic D, Lin D, Nurk S, Florez De Sessions P, Sikic M. Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads. 2024 [cited 12 Mar 2025]. Available from: http://biorxiv.org/lookup/ https://doi.org/10.1101/2024.05.18.594796
42. Oxford Nanopore Technologies. Guppy protocol. Available from: https://nanoporetech.com/document/Guppy-protocol [cited 2025 Mar 30]
43. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6.
44. Williamson DH, Denny PW, Moore PW, Sato S, McCready S, Wilson RJM (Iain), editors. The in vivo conformation of the plastid DNA of Toxoplasma gondii: implications for replication1 1Edited by N.-H. Chua. Journal of Molecular Biology. 2001;306:159–68.
45. Haese-Hill W, Crouch K, Otto TD. Annotation and visualization of parasite, fungi and arthropod genomes with companion. Nucleic Acids Res. 2024;52:W39–44.
46. Xia J, Venkat A, Bainbridge RE, Reese ML, Le Roch KG, Ay F, et al. <article-title update="added">Third-generation sequencing revises the molecular

karyotype for *Toxoplasma gondii* and identifies emerging copy number variants in sexual recombinants. Genome Res. 2021;31:834–51.
47. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. Darling AE, editor. PLoS Comput Biol. 2018;14:e1005944.
48. Kissinger JC. ToxoDB: accessing the *Toxoplasma gondii* genome. Nucleic Acids Res. 2003;31:234–6.
49. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
51. Medaka. Available from: https://github.com/nanoporetech/medaka [cited 25 Mar 2025]
52. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42:1571–80.
53. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics. 2008;24:637–44.
54. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36:875–9.
55. Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biol. 2020;21:253.
56. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17:122.
57. Guo Q, Guo X, Ji N, Shen B, Zhong X, Xiao L, et al. Role of 6-phosphogluconate dehydrogenase enzyme 1 in growth and virulence of *Toxoplasma gondii* and development of attenuated live vaccine. Microb Biotechnol. 2023;16:1957–70.
58. Jeninga M, Quinn J, Petter M. ApiAP2 transcription factors in apicomplexan parasites. Pathogens. 2019;8:47.
59. Alaganan A, Fentress SJ, Tang K, Wang Q, Sibley LD. *Toxoplasma* GRA7 effector increases turnover of immunity-related GTPases and contributes to acute virulence in the mouse. Proc Natl Acad Sci USA. 2014;111:1126–31.
60. Labesse G, Gelin M, Bessin Y, Lebrun M, Papoin J, Cerdan R, et al. ROP2 from *Toxoplasma gondii*: a virulence factor with a Protein-Kinase fold and no enzymatic activity. Structure. 2009;17:139–46.
61. Adomako-Ankomah Y, Wier GM, Borges AL, Wand HE, Boyle JP. Differential Locus Expansion Distinguishes *Toxoplasmatinae* Species and Closely Related Strains of Toxoplasma gondii. Johnson P, editor. mBio. 2014;5:e01003-13.
62. Pernas L, Boothroyd JC. Association of host mitochondria with the parasitophorous vacuole during *Toxoplasma* infection is not dependent on Rhoptry proteins ROP2/8. Int J Parasitol. 2010;40:1367–71.
63. El Hajj H, Demey E, Poncet J, Lebrun M, Wu B, Galéotti N, et al. The ROP2 family ofToxoplasma gondii rhoptry proteins: proteomic and genomic characterization and molecular modeling. Proteomics. 2006;6:5773–84.
64. Sibley LD, Qiu W, Fentress S, Taylor SJ, Khan A, Hui R. Forward genetics in *Toxoplasma gondii* reveals a family of Rhoptry kinases that mediates pathogenesis. Eukaryot Cell. 2009;8:1085–93.
65. Buchholz KR, Fritz HM, Chen X, Durbin-Johnson B, Rocke DM, Ferguson DJ, et al. Identification of tissue cyst wall components by transcriptome analysis of *in vivo* and in vitro Toxoplasma gondii bradyzoites. Eukaryot Cell. 2011;10:1637–47.
66. Doliwa C, Xia D, Escotte-Binet S, Newsham EL, Sanya JS, Aubert D, et al. Identification of differentially expressed proteins in sulfadiazine resistant and sensitive strains of Toxoplasma gondii using difference-gel electrophoresis (DIGE). Int J Parasitology: Drugs Drug Resist. 2013;3:35–44.
67. Swapna LS, Stevens GC, Sardinha Da Silva A, Hu LZ, Brand V, Fusca DD et al. ToxoNet: A high confidence map of protein-protein interactions in Toxoplasma gondii reveals novel virulence factors implicated in host cell invasion. Available from: https://doi.org/10.1101/2021.09.14.460186
68. Kong P, Ufermann C-M, Zimmermann DLM, Yin Q, Suo X, Helms JB, et al. Two phylogenetically and compartmentally distinct CDP-diacylglycerol synthases cooperate for lipid biogenesis in *Toxoplasma gondii*. J Biol Chem. 2017;292:7145–59.
69. Sharma A, Yogavel M, Sharma A. Structural and functional attributes of malaria parasite diadenosine tetraphosphate hydrolase. Sci Rep. 2016;6:19981.
70. Gissot M, Walker R, Delhaye S, Huot L, Hot D, Tomavo S. Toxoplasma gondii Chromodomain Protein 1 Binds to Heterochromatin and Colocalises with Centromeres and Telomeres at the Nuclear Periphery. Waller RF, editor. PLoS ONE. 2012;7:e32671.
71. Bisio H, Lunghi M, Brochet M, Soldati-Favre D. Phosphatidic acid governs natural egress in *Toxoplasma gondii* via a guanylate cyclase receptor platform. Nat Microbiol. 2019;4:420–8.
72. Subauste CS. Interplay between *Toxoplasma gondii*, autophagy, and autophagy proteins. Front Cell Infect Microbiol. 2019;9:139.
73. Li W, Grech J, Stortz JF, Gow M, Periz J, Meissner M, et al. A splitCas9 phenotypic screen in *Toxoplasma gondii* identifies proteins involved in host cell egress and invasion. Nat Microbiol. 2022;7:882–95.
74. Kim BY, Gellert HR, Church SH, Suvorov A, Anderson SS, Barmina O et al. Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. Jiggins CD, editor. PLoS Biol. 2024;22:e3002697.
75. Delahaye C, Nicolas J. Sequencing DNA with nanopores: troubles and biases. Andrés-León E. Editor PLoS ONE. 2021;16:e0257521.
76. Dechering K. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. Nucleic Acids Res. 1998;26:4056–62.
77. Fan H, Chu J-Y. A brief review of short tandem repeat mutation. Genom Proteom Bioinform. 2007;5:7–14.
78. Berná L, Marquez P, Cabrera A, Greif G, Francia ME, Robello C. Reevaluation of the *Toxoplasma gondii* and *Neospora caninum* genomes reveals misassembly, karyotype differences, and chromosomal rearrangements. Genome Res. 2021;31:823–33.
79. Ramaprasad A, Mourier T, Naeem R, Malas TB, Moussa E, Panigrahi A et al. Comprehensive Evaluation of Toxoplasma gondii VEG and Neospora caninum LIV Genomes with Tachyzoite Stage Transcriptome and Proteome Defines Novel Transcript Features. Blader IJ, editor. PLoS ONE. 2015;10:e0124473.
80. Su C, Khan A, Zhou P, Majumdar D, Ajzenberg D, Dardé M-L, et al. Globally diverse *Toxoplasma gondii* isolates comprise six major clades originating from a small number of distinct ancestral lineages. Proc Natl Acad Sci U S A. 2012;109:5844–9.
81. Hall AN, Morton E, Queitsch C. First discovered, long out of sight, finally visible: ribosomal DNA. Trends Genet. 2022;38:587–97.
82. Guay J-M, Huot A, Gagnon S, Tremblay A, Levesque RC. Physical and genetic mapping of cloned ribosomal DNA from *Toxoplasma gondii*: primary and secondary structure of the 5S gene. Gene. 1992;114:165–71.
83. Cahyani I, Tyson J, Holmes N, Quick J, Moore C, Loman N et al. *FindingNemo*: A Toolkit for DNA Extraction, Library Preparation and Purification for Ultra Long Nanopore Sequencing. 2024. Available from: http://biorxiv.org/lookup/ https://doi.org/10.1101/2024.08.16.608306 [cited 30 Mar 2025]
84. Oxford Nanopore Technologies. Ultra-Long DNA, Sequencing Kit V14 (SQK-ULK114). Ultra-Long DNA Sequencing Kit V14 (SQK-ULK114). 2024. Available from: https://nanoporetech.com/document/ultra-long-dna-sequencing-kit-sqk-ulk114 [cited 24 Mar 2025]
85. Kim J-H, Dilthey AT, Nagaraja R, Lee H-S, Koren S, Dudekula D, et al. Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. Nucleic Acids Res. 2018;46:6712–25.
86. Namasivayam S, Sivaranjani Namasivayam, Namasivayam S, Rodrigo de Paulo Baptista, Baptista RP, Xiao W et al. A novel fragmented mitochondrial genome in the protist pathogen Toxoplasma gondii and related tissue coccidia. bioRxiv. 2020.
87. Oxford Nanopore Technologies. Size selection of HMW DNA by semi-selective DNA precipitation [Internet]. Size selection of HMW DNA by semi-selective DNA precipitation.. Available from: https://nanoporetech.com/document/extraction-method/size-selection2#change-log [cited 2025 Mar 25]
88. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
89. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One. 2016;11:e0163962.
90. Cabanettes F, Klopp C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. PeerJ. 2018;6:e4958.
91. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Gene. 2012;3. Available from: http://journal.frontiersin.org/article/ https://doi.org/10.3389/fgene.2012.00035/abstract [cited 12 Mar 2025]
92. The Variant Call Format (VCF). Version 4.1 Specification.. Available from: https://samtools.github.io/hts-specs/VCFv4.1.pdf [cited 25 Mar 2025]
93. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal Omega. Mol Syst Biol. 2011;7:539.

94. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
95. Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nat Genet. 2022;54:518–25.
96. Wang X, Wang L. GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. Front Plant Sci. 2016;7. Available from: http://journal.frontiersin.org/Article/ https://doi.org/10.3389/fpls.2016.01350/abstract [cited 12 Mar 2025]
97. umiVar. Available from: https://github.com/imgag/umiVar [cited 30 Mar 2025]
98. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013. Available from: http://www.repeatmasker.org

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.2 Supplementary material for Publication 1

3.2.1 Supplementary notes

3.2.1.1 Propagation history of the sequenced T. gondii ME49 isolates

2015T was obtained from ATCC (the exact date is unkown) and thawed in 2015. After thawing, it was cultured for two passages and then cryopreserved. In January 2022, it was thawed again and maintained in continuous cell culture, with passaging occurring twice per week. Samples for sequencing were collected between June and November 2024. Exact collection dates are provided in Supplementary Table 1. **2020T** was obtained from ATCC in April 2020, thawed and cultured in May 2020, and has been continuously maintained in cell culture since that time, with biweekly passages. Two samples were collected for sequencing in December 2021 and May 2022 (Supplementary Table 1). **2000B** was acquired in the early 1990s; the exact source is unknown. The isolate has been continuously propagated in mice at the Institute of Medical Microbiology and Hospital Hygiene, University Hospital Düsseldorf, for at least 30 years. Reinfection typically occurred every 3–6 months, depending on experimental needs. Periods of frequent passaging were associated with higher animal usage, while lower demand resulted in less frequent propagation. Sample collection for sequencing happend between January 2022 until July 2023. Exact collection dates are provided in Supplementary Table 1.

3.2.1.2 Mitochondrial reads confound assembly at NUMT-Containing Locus

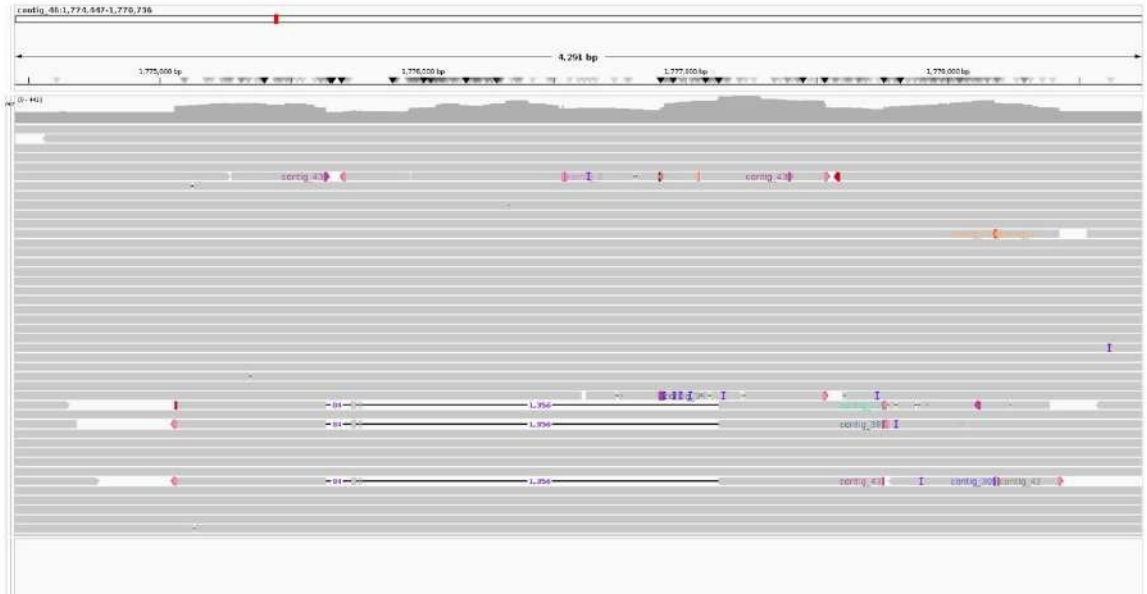
During an initial genome assembly attempt that did not exclude mitochondrial reads, we identified an anomalous region that aligned with one of the identified anomalous loci in the Xia et al. ME49 assembly (see later). This artifact arises due to the presence of a large nuclear mitochondrial DNA segment (NUMT), which causes mitochondrial reads to erroneously align to the nuclear genome and induces a misassembly. This issue is clearly visible in the IGV screenshots in the following Figure: (A) shows the region before mitochondrial read filtering, where abnormal read alignment and coverage patterns are evident, while (B) shows the same region after

filtering for mitochondrial reads and reassembly, where the assembly anomaly is resolved.

A



B



3.2.1.3 Identification of Assembly Anomalies

The entire Xia et al. assembly was, based on the Nanopore sequencing data generated by Xia et al. that was used to generate the assembly (SRA: SRX8862416), manually inspected in IGV v2.17.4 to assess assembly quality. The goal was to systematically identify anomalous positions that could indicate potential errors in the assembly. Anomalous positions were identified based on three primary criteria:

1. Read depth anomalies: Sudden increases or decreases in read depth, particularly regions with sharp depth drops or unexpected coverage spikes.
2. Evidence for structural variants (SVs) : Positions showing split-read signatures, or large insertions/deletions indicative of structural variations.
3. Patterns associated with collapsed repeats: Regions with stretches of positions containing multiple alleles, suggesting possible misassemblies due to repeat compression.

Positions were then classified into two categories:

- Erroneous positions: Positions where all reads mapping to the site exhibited the same anomaly, such as a uniform pattern of insertions or a complete lack of spanning reads at a high-depth site.
- Suspicious positions: Positions where some but not all reads exhibited anomalies, indicating possible inconsistencies or mixed signals in the sequencing data (e.g., at least one spanning read was present)

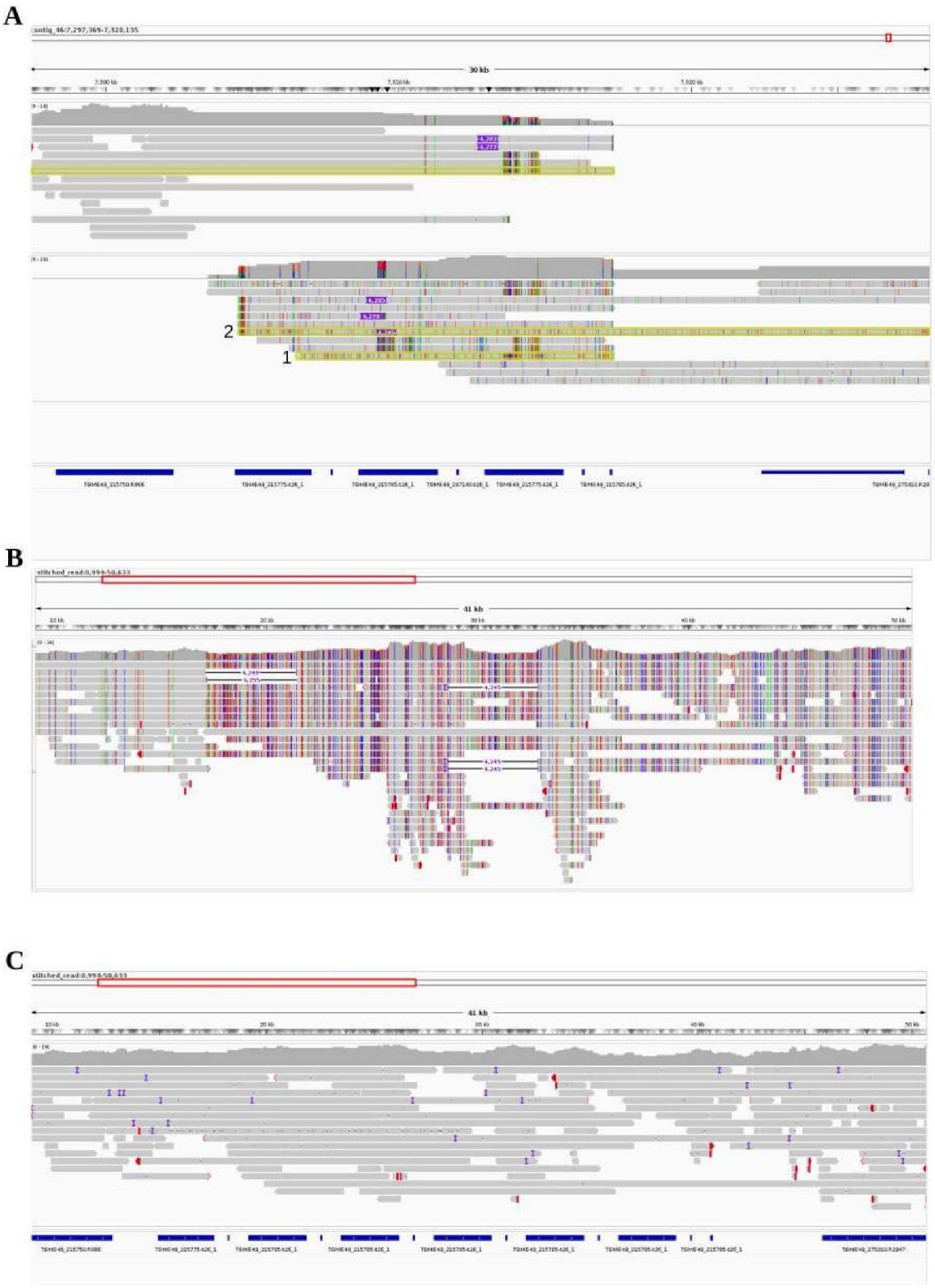
These positions were extracted with coordinates selected to cover the full extent of the anomalous regions. After creating the 2015T assembly, we used LiftOver [1](version 469, installed from Bioconda) to project these identified positions from the Xia et al. assembly onto the 2015T assembly. To facilitate coordinate projection, we implemented a Snakemake v7.32.4 pipeline, first mapping the new assembly to the JACEHA000000000.1 reference with minimap2 using the -cx asm20 -m 10000 -z 10000,50 -r 50000 --end-bonus=100 --secondary=no --eqx -Y -O 5,56 -E 4,1 -B 5 preset to generate a PAF alignment file. This file was converted to chain files using

paf2chain [2] for use in coordinate conversion, and LiftOver was subsequently performed with the UCSC liftOver tool with -minMatch=0.4. Each position in the 2015T assembly was then reviewed in IGV to verify alignment consistency and read depth, ensuring that no unexpected variations or mismatches were present. A table containing the coordinates of the anomalous positions on the Xia et al. assembly and their corresponding positions on the 2015T assembly is provided in Supplementary Table 2. A final scan of the 2015T assembly in IGV confirmed the absence of new anomalies and validated that all regions exhibited consistent read alignment and expected depth, indicating a more accurate and homogeneous assembly.

3.2.1.4 Manual curation of ROP8-ROP2A sequence in 2000B sample

We defined two anchor genes—genes that are neither ROP2A nor ROP8 and that exist uniquely in the genome. The identified anchors were TGME49_215750, located left of the ROP2A-ROP8 locus, and TGME49_275310, located right of the locus. No single read from sample 2000B spanned both anchor genes in the 2015T assembly, i.e. no single read was informative about the true structure of this locus in the 2000B sample. To reconstruct the sequence of the locus in isolate 2000B, we identified a pair (A, B) of overlapping reads, selected according to the following criteria: i) read A contained the left anchor; ii) read B contained the right anchor; iii) A and B overlapped in the reference genome with similar mismatch pattern at the overlapping region. To find suitable read pairs, reads mapping upstream of position 7,302,329 in 2015T (left anchor) and those spanning position 7,302,329 to 7,322,405 in 2015T (right anchor) were extracted using SAMtools, and a pair of reads, meeting the criteria stated above, was selected: 9350f71a-a210-444c-97d9-ab2c695af5e3 (left) and cb2b7796-e42a-46b4-b204-7b3e511adecc (right) (panel A in the following figure). The second read was reverse-complemented, and the corrected sequence was used in analysis. The two reads were stitched together using Minimap2 (-cx ava-ont), selecting the alignment from the PAF file based on the expected 10,000 bp overlap. Bases up to position 17,115 in 9350f71a-a210-444c-97d9-ab2c695af5e3 were concatenated with

cb2b7796-e42a-46b4-b204-7b3e511adecc starting from position 108. Reads from the 2000B sample mapping to the ROP8-ROP2A locus in the 2015T assembly were extracted and aligned to the stitched sequence using Minimap2 (-ax map-ont) (panel B in the following figure). The stitched sequence was then polished in two rounds with Medaka (r104_e81_sup_variant_g610_model_pt.tar.gz). Secondary alignments and small indels (<10bp) were filtered during visualization(panel C in the following figure).

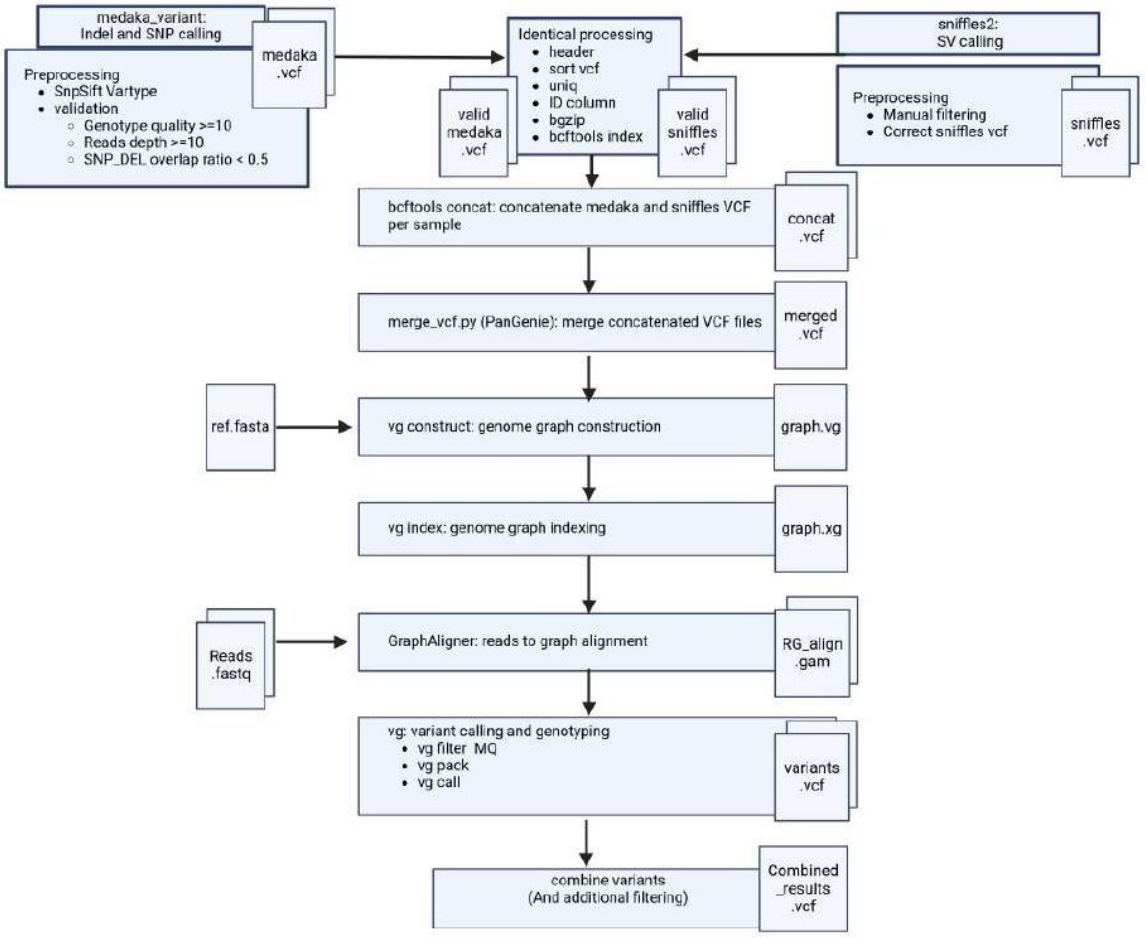


3.2.1.5 References

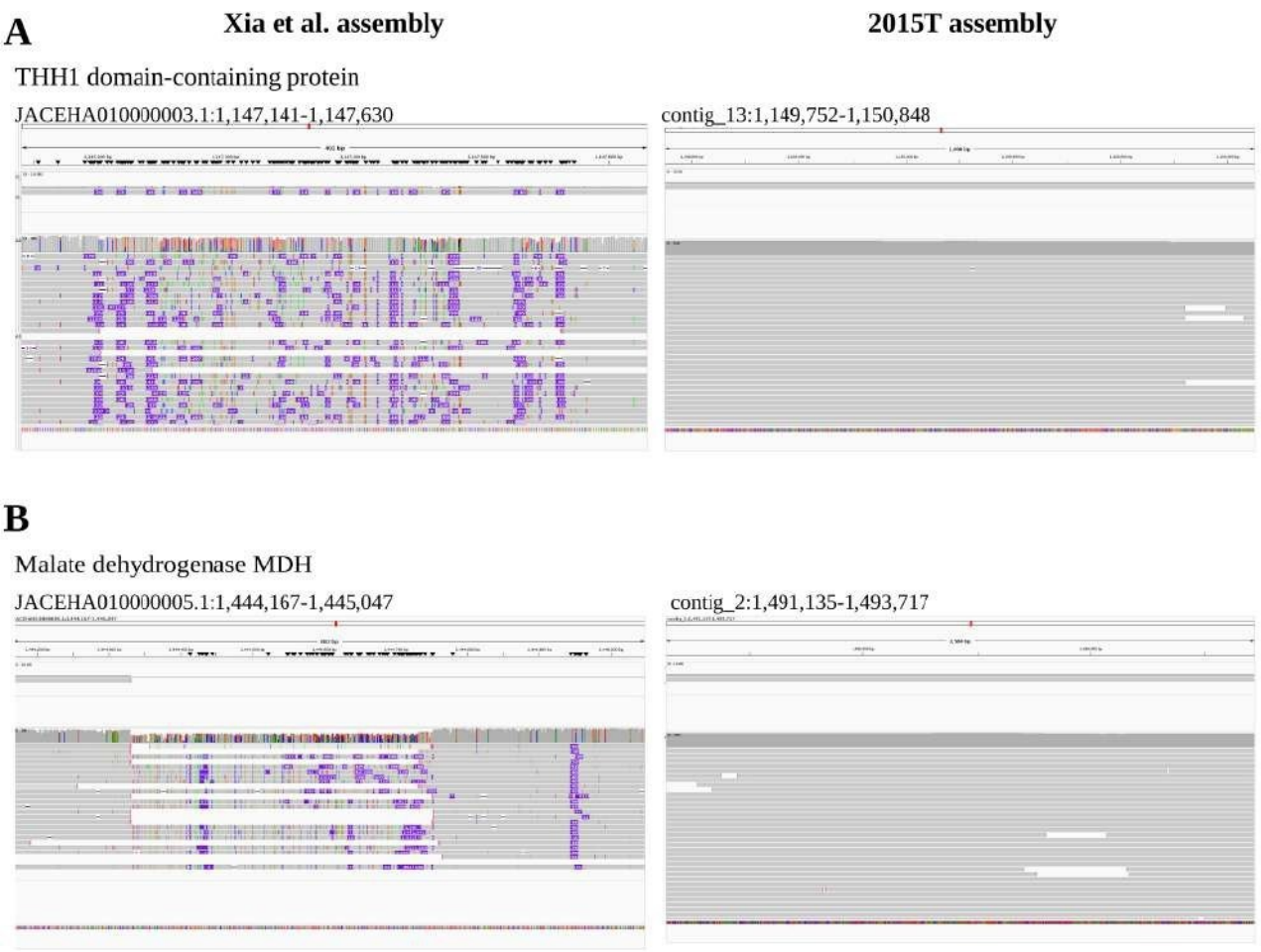
1. Hinrichs AS. The UCSC Genome Browser Database: update 2006. Nucleic Acids Research. 2006;34:D590–8.

2. Guarracino A. AndreaGuarracino/paf2chain: v0.1.0 [Internet]. Zenodo; 2023 [cited 2025 Apr 3]. Available from: https://zenodo.org/record/8108447

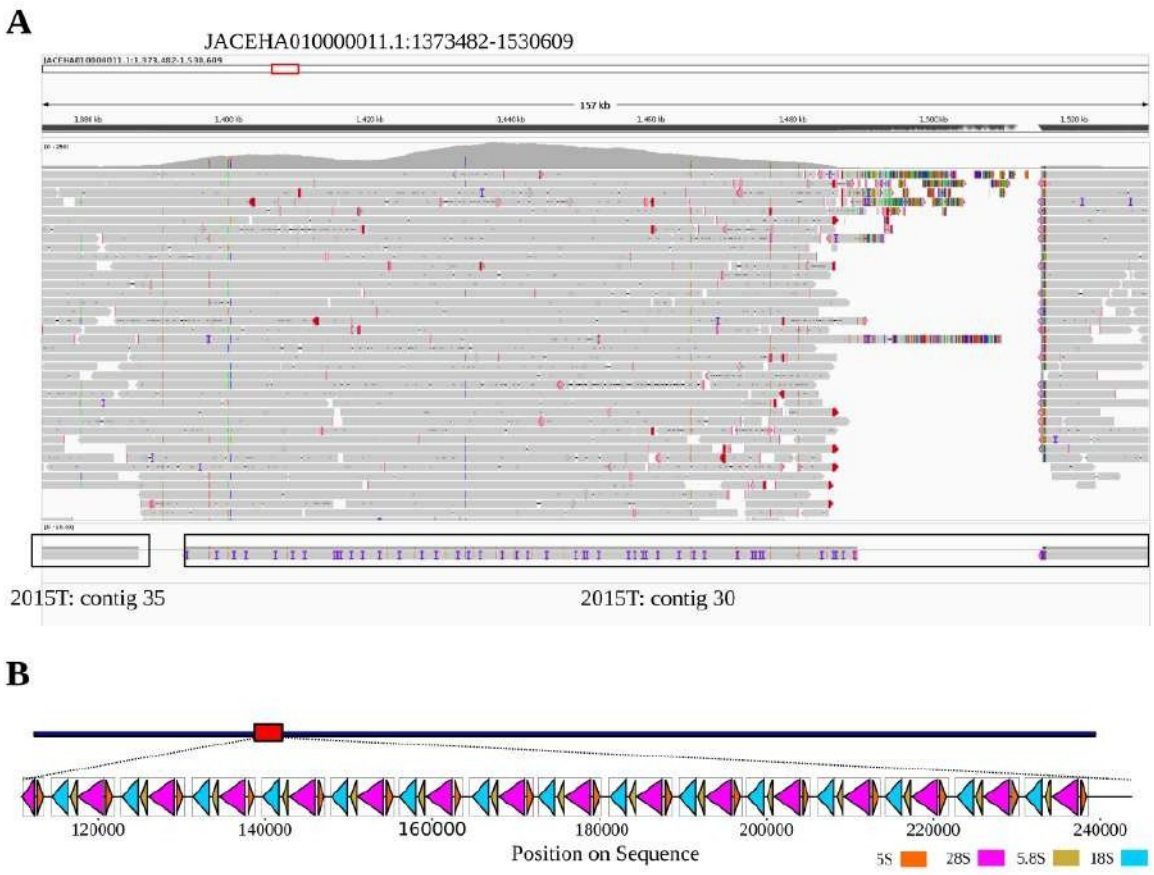
3.2.2 Supplementary figures



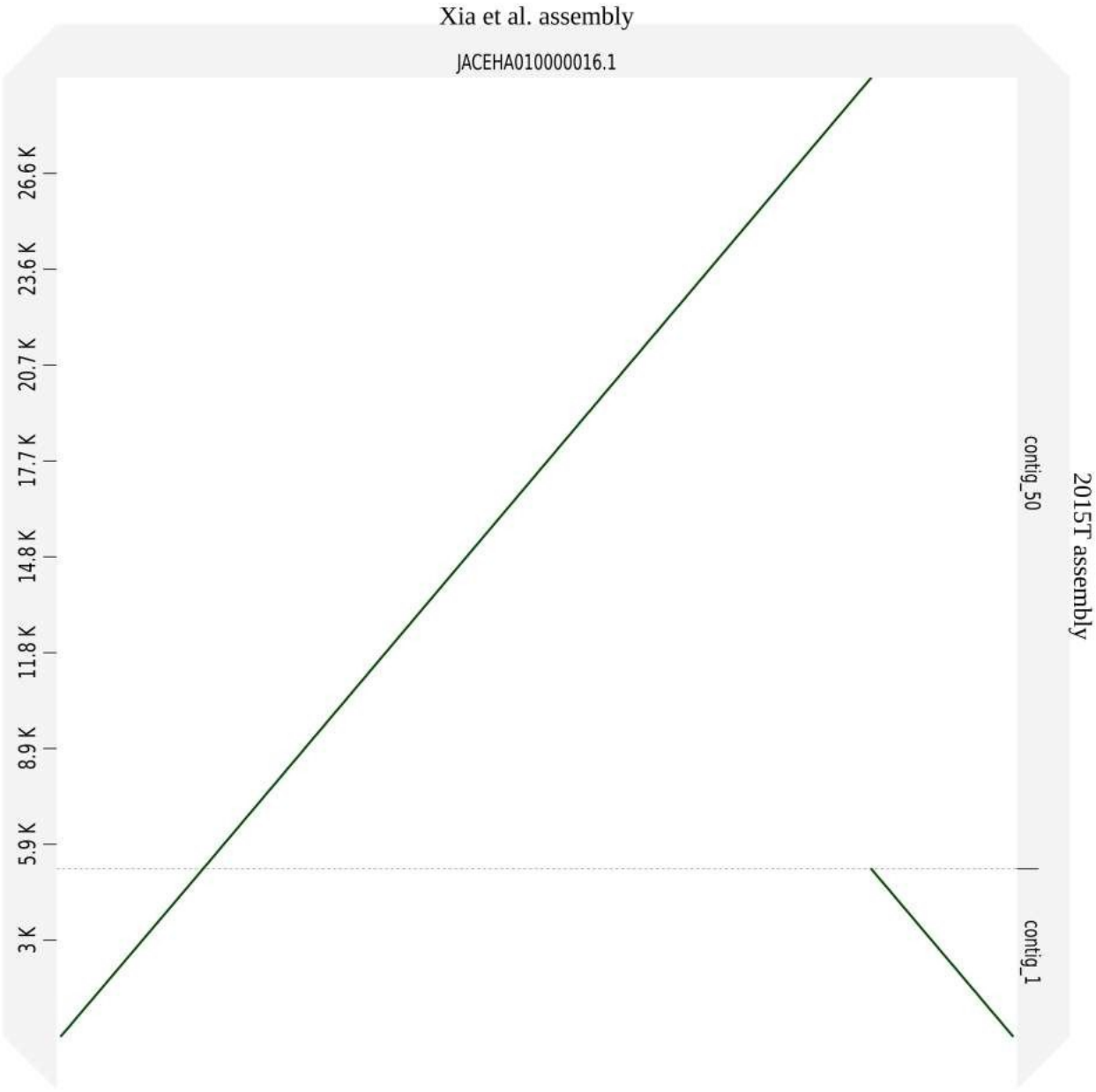
Supplementary Figure S1:
Flowchart of the Snakemake pipeline used in the variant analysis. Created in https://BioRender.com



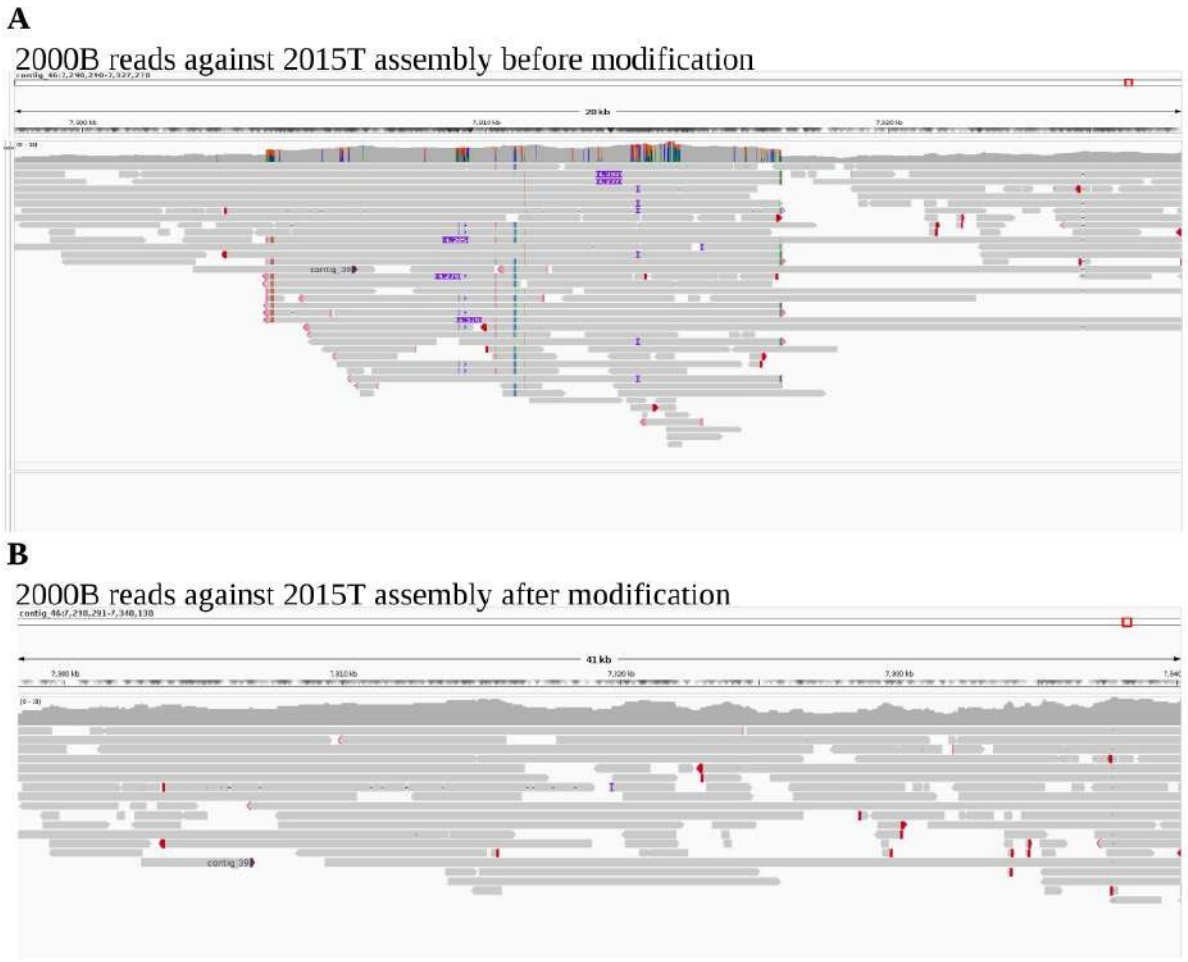
Supplementary Figure S2: Improved alignment of sequencing reads and ToxoDB reference gene sequence in the 2015T assembly compared to the Xia et al. assembly.
In each IGV panel, the top track shows the gene from ToxoDB, and the bottom track displays the long reads used to construct each assembly.
(A) Read alignment over the *THH1 domain-containing* gene *(TGME49_222310)*. The Xia et al. assembly exhibits a high density of mismatches and indels, whereas the 2015T assembly shows clean and consistent alignment. (B) Read alignment over the *Malate dehydrogenase (MDH)* gene *(TGME49_318430)*. In the Xia et al. assembly, extensive mismatches and gaps lead to incomplete mapping of the gene sequence. In contrast, the 2015T assembly enables smooth and concordant alignments, reflecting improved sequence accuracy.



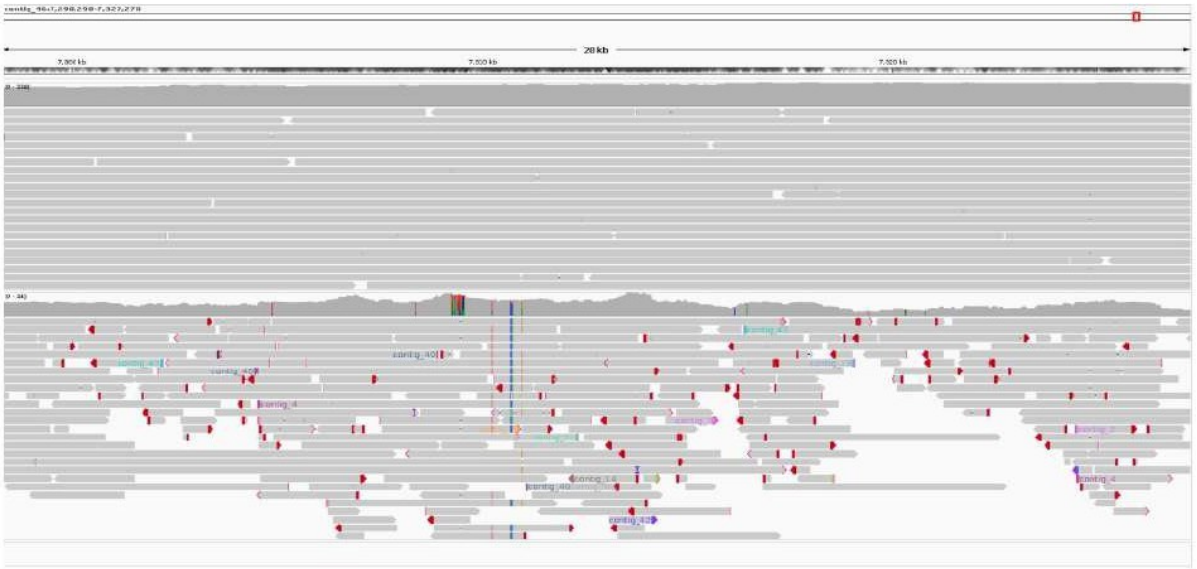
Supplementary Figure S3. Evidence of a likely misassembled rDNA repeat region in the Xia et al. assembly.
(A) IGV screenshot showing the alignment of reads used to construct the Xia et al. assembly (JACEHA010000011.1:137,482–1,530,609). A marked increase in read depth is observed in this region, accompanied by a lack of spanning reads, suggesting a potential misassembly. The bottom track shows the alignment of 2015T assembly contigs (contig_35 and contig_30). (B) Schematic representation of rDNA gene blocks annotated in the same region of the Xia et al. assembly.



Supplementary Figure S4. Dot plot comparison of the apicoplast genome between the Xia e al. assembly and the 2015T assembly.
A dot plot generated using D-Genies shows alignment between the apicoplast contig from the Xia et al. assembly (JACEHA010000016.1) and the 2015T assembly. A single contig from the 2015T assembly (contig_1) aligns to the Xia et al. contig in two segments: once in the forward direction and once in the reverse complement. This pattern is consistent with the presence of an inverted repeat in apicoplast genome.



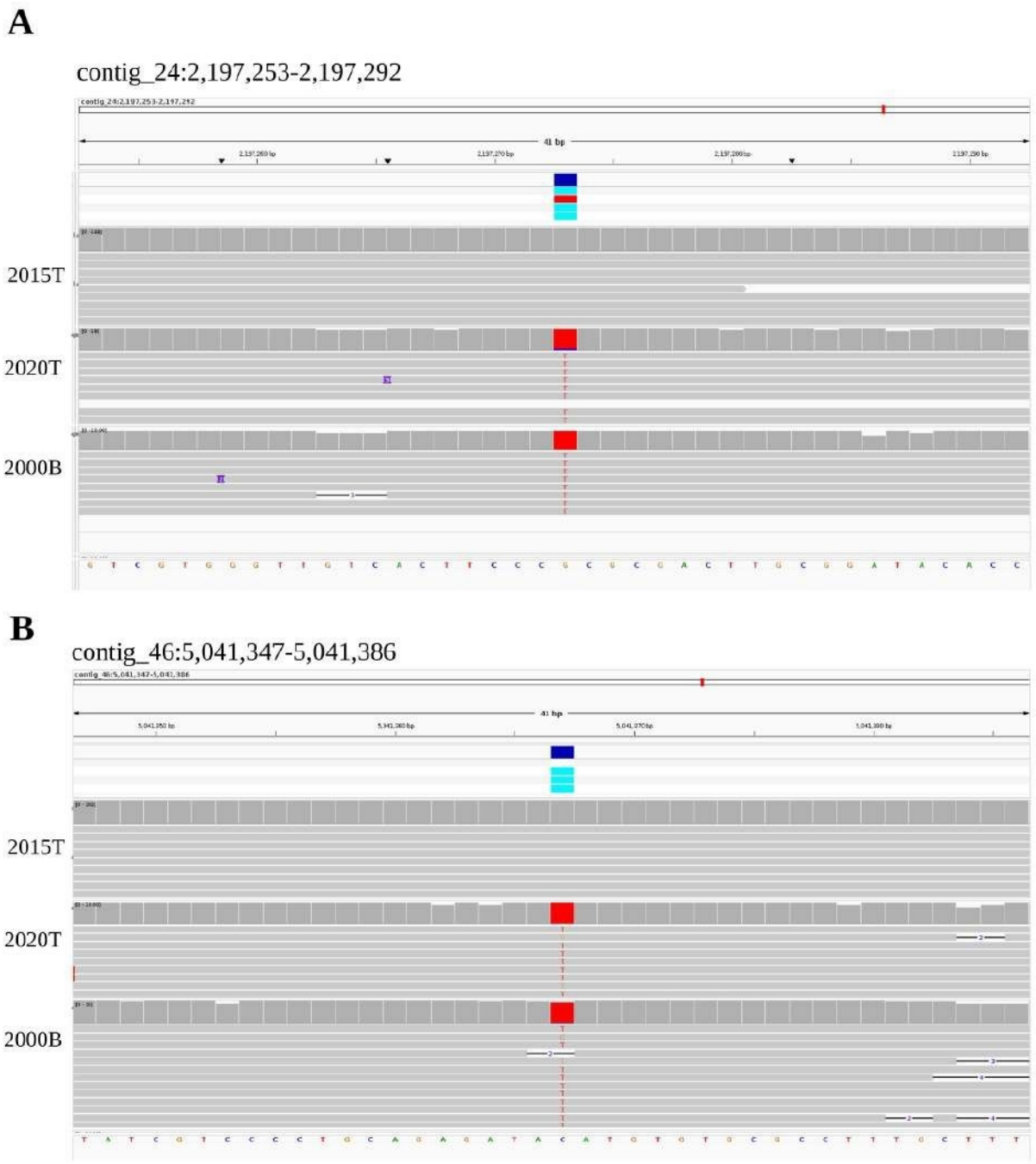
Supplementary Figure S5. Resolution of a structural variant in the 2000B isolate through reference modification.
(A) IGV snapshot showing 2000B reads aligned to the original 2015T assembly, displaying a region with abnormal coverage and extensive read clipping, consistent with a misrepresented structural variant. (B) Alignment of the same 2000B reads to the modified reference shows a uniform coverage profile with minimal clipping, indicating that the structural variant was correctly resolved. For clean visualization, secondary alignments and reads with mapping quality 0 (MAPQ0) were filtered out, and the coverage allele-fraction threshold was set to 0.5.



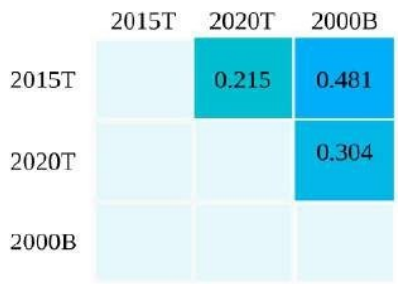
Supplementary Figure S6. Evidence for a structural variant in 2020T relative to the 2015T reference.
IGV screenshots showing read alignments of 2015T (top panel) and 2020T (bottom panel) to the 2015T assembly. While the 2015T reads align cleanly with uniform coverage, the 2020T reads show a mixed base pattern, where subsets of reads consistently support different genotypes across the same region. This pattern is similar to that observed in the 2000B alignment (Figure S5A) and is consistent with a collapsed repeat or unresolved structural variation in the reference assembly. For clean visualization, secondary alignments and reads with mapping quality 0 (MAPQ0) were filtered out, and the coverage allele-fraction threshold was set to 0.5.



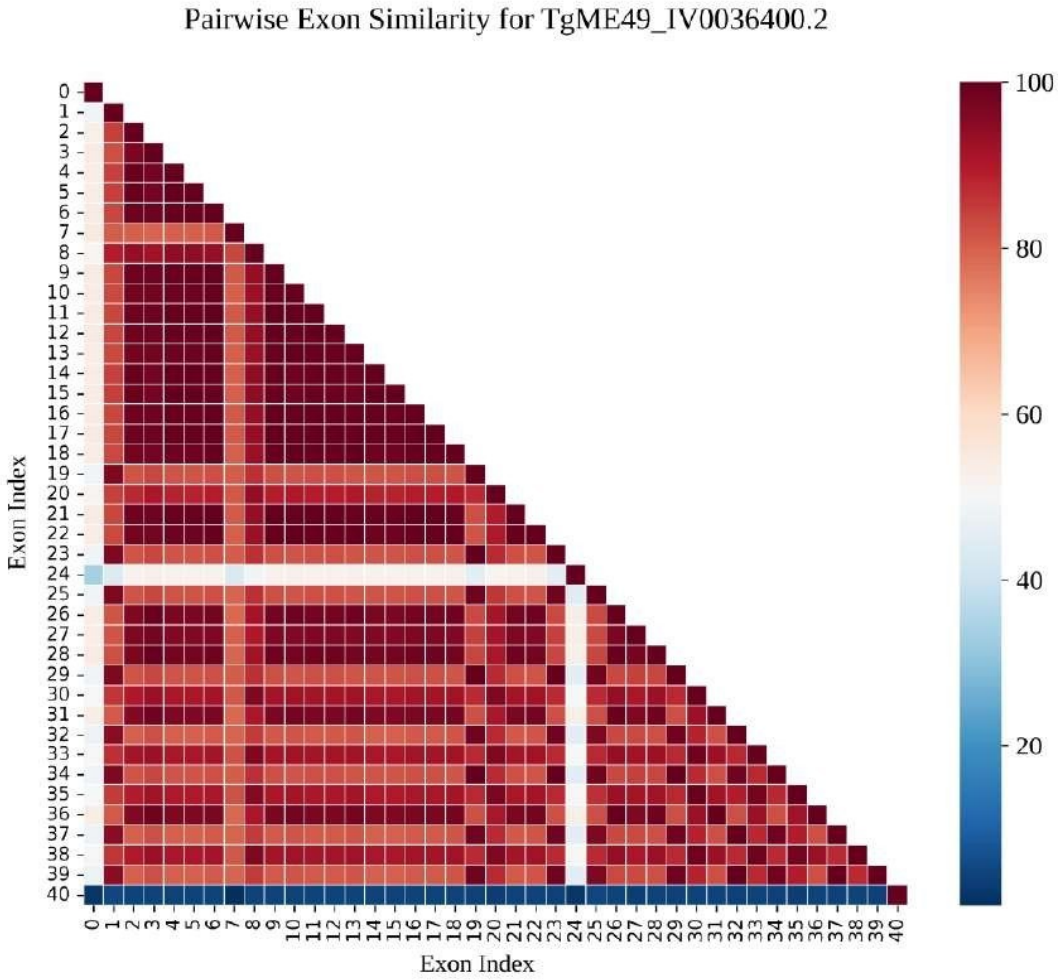
Supplementary Figure S7. RNA-seq read coverage across the *ROP8–ROP2A* locus in the 2000B.
IGV screenshot showing RNA-seq alignments from *T. gondii* ME49 mapped to the structurally resolved *ROP8–ROP2A* region of isolate 2000B. The coverage profile indicates detectable expression of ROP8-1, ROP8-2, ROP8-3, ROP2A-A, ROP2A-2B and ROP2A-1. The alignment is visualized with secondary alignments filtered out.



Supplementary Figure S8. Examples of genotype correction through graph-based variant calling.
(A) An example of a variant that was initially called only in isolate 2020T but missed in 2000B, despite clear evidence of its presence in both isolates. The graph-based genotyping approach correctly recovered the variant in both samples. (B) A case where a true variant in 2020T was erroneously filtered out during the Medaka-based filtering step. The variant was subsequently rescued by the graph-based genotyper, restoring the correct genotype.



Supplementary Figure S9. Pairwise genetic similarity between isolates based on SNVs. Matrix showing inter-isolate genetic similarity calculated from shared single-nucleotide variants (SNVs). The highest level of similarity was observed between isolates 2000B and 2015T, indicating a closer genetic relationship relative to the other isolate pairs.



Supplementary Figure S10. Pairwise exon similarity matrix for *TgME49_IV0036400.2*

Heatmap showing the percent identity between all exon pairs of the gene *TgME49_IV0036400.2*. Exons were aligned in all-vs-all fashion, and similarity was calculated as percent identity over alignment length. The exon indices on both axes start at 0 and follow the exon order in the GTF annotation.

3.2.3 Supplementary tables

Due to their size, the supplementary tables are not included in print and can be found on the enclosed CD

3.3 Publication 2: “KIR*BLOOM: Accurate KIR genotyping using a new copy number-aware integrated genotype likelihood framework”

Available on: BioRxiv

DOI: https://doi.org/10.64898/2026.04.15.718735

Authorship: first author

Contribution to the publication:

This study was initiated through joint discussions with AD, who proposed the core idea of the method. Building on this, I led the development of the KIR*BLOOM algorithm, including its implementation and iterative refinement. I designed and conducted the computational experiments and performed the analyses. The manuscript was written jointly with AD. AD supervised the study and provided continuous input on the methodology and interpretation. ADG contributed at an early stage by helping to define the modeling strategy and evaluation metrics.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

KIR*BLOOM: Accurate KIR genotyping using a new copy number-aware integrated genotype likelihood framework

Yomna Gohar[1,2], Alberto Descalzo Garcia[1], Katherine M. Kichula[3], Paul J. Norman[3], Alexander Dilthey [1,2]

[1] Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[2] Center for Digital Medicine, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[3] Department of Biomedical Informatics, and Department of Immunology and Microbiology, University of Colorado School of Medicine

Abstract

Killer-cell immunoglobulin-like receptor (KIR) genes, key modulators of natural killer (NK) cell activity, play critical roles in immune response and disease susceptibility. Accurate KIR genotyping from short-read sequencing data remains challenging because of high sequence similarity among genes, extensive copy number variation, and substantial allelic diversity. Here, we present KIR*BLOOM, a likelihood-based approach for KIR genotyping from short-read data that models read depth and sequencing error across alternative genotype configurations. KIR*BLOOM first identifies KIR-relevant read pairs, maps them to a KIR allele database, and reduces the candidate allele space by excluding alleles unlikely to be present. It then infers gene copy number and selects alleles under the inferred copy-number constraints. Finally, variant calling is used to refine CDS sequences and identify potential novel alleles. We evaluated performance on 45 whole-genome sequencing samples with haplotype-resolved assemblies from the HPRC or HGSVC, using Immuannot-derived annotations as ground truth. KIR*BLOOM achieved 99.85% precision, 99.92% recall, and a Jaccard index of 99.77% for copy-number inference. At five-digit allele resolution, it achieved 92.73% precision, 92.69% recall, and an 87.29% Jaccard index, outperforming T1K, GraphKIR, and Geny. Together, these results demonstrate that KIR*BLOOM enables highly accurate KIR genotyping from short-read sequencing data.

Introduction

Killer-cell immunoglobulin-like receptors (KIRs) are a family of receptors expressed on natural killer (NK) cells and a subset of T cells (Campbell and Purdy 2011) that interact primarily with major histocompatibility complex (MHC) class I molecules on host cells (Vilches and Parham 2002). These interactions are central to immune surveillance, enabling NK cells to balance activation and inhibition, thereby detecting and eliminating cells with intracellular abnormalities such as viral infection or malignant transformation (Benson and Caligiuri 2014; Trowsdale et al. 2015). KIR receptors are encoded by KIR genes located on chromosome 19q13.4. Since their discovery in the

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

early 1990s, 17 KIR genes have been identified. They include eight inhibitory receptors (KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL5A, KIR2DL5B, KIR3DL1, KIR3DL2, and KIR3DL3), seven activating receptors (KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS4, KIR2DS5, KIR3DS1, and the dual-function KIR2DL4), and two pseudogenes (KIR2DP1 and KIR3DP1) (Béziat et al. 2017). Individuals can carry diverse combinations of 8 to 17 KIR genes. In addition, KIR genes exhibit extensive allelic polymorphism and gene copy number variation, ranging from complete absence to up to four copies of a single gene per individual, resulting in hundreds of observed haplotype structures. Despite this diversity, KIR haplotypes are broadly classified into group A and group B, which differ in gene content and functional balance. Group A haplotypes are relatively conserved and enriched for inhibitory receptors, typically containing a single activating gene (KIR2DS4), whereas group B haplotypes are more diverse and harbor multiple activating KIR genes, contributing to greater variability in immune responses (Hollenbach et al. 2012; Hsu et al. 2002). Genetic variation within the KIR locus has been associated with a wide range of clinically relevant phenotypes, including susceptibility to infectious diseases (Khakoo et al. 2004; Martin et al. 2002), autoimmune disorders (Hollenbach et al. 2009), and cancer (Purdy and Campbell 2009; Liu et al. 2015). In pregnancy, KIR receptors play a critical role in regulating uterine NK cell interactions with fetal trophoblasts, thereby influencing placental development and the risk of disorders such as pre-eclampsia (Kulkarni et al. 2008).

Genotyping KIR genes remains particularly challenging because the KIR locus combines extensive allelic polymorphism, high sequence homology between genes, and substantial structural variation (Roe et al. 2020). While a central database of KIR alleles exists (IPD-KIR) (Robinson et al. 2019) currently containing 2,286 alleles, However, even with this reference, precise genotype inference remains difficult because many alleles are distinguished by only small sequence differences, including synonymous substitutions within coding regions or variants confined to intronic or other non-coding regions. Closely related genes such as KIR3DL1/KIR3DS1 and KIR2DL5A/KIR2DL5B (Song et al. 2023), for example, share considerable sequence similarity, increasing the risk of ambiguous read assignment. In addition, about 40% of the alleles in the database are incomplete. Together, these factors, including the extensive copy number variation of KIR genes described above, underscore the need for specialized and accurate genotyping approaches.

Among the currently available tools for allele-level KIR genotyping from short-read sequencing data, PING (Marin et al. 2021) and T1K (Song et al. 2023) are among the best established, with additional newly developed methods such as Geny (Zhou et al. 2024a) and GraphKIR (Lin et al. 2023) also contributing to the available set of approaches. PING is a KIR-specific workflow that extracts KIR reads and then estimates gene copy number and allele genotypes, whereas T1K is developed for both

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

HLA and KIR alleles inference by estimating allele abundances from read alignments and supports RNA-seq, whole-genome, and whole-exome sequencing data. However, both approaches retain important limitations. PING has been reported to require substantial manual parameter tuning and to be sensitive to cohort-dependent thresholds and copy-number normalization assumptions. T1K is efficient and broadly applicable, but recent evaluations indicate that complex KIR copy-number states are not modeled robustly, which can lead to undercalling of gene copies, particularly when copy number exceeds two. In addition, newer tools such as Geny and GraphKIR have not yet been extensively benchmarked across diverse real-world short-read datasets. Together, these limitations restrict accurate fine-mapping of KIR-associated phenotypes to their underlying genotypes. Consequently, there remains a clear need for improved tools for accurate high-resolution KIR genotyping.

Motivated by this need, we developed KIR*BLOOM (bounded likelihood optimization for observed mixture), a method that explicitly accounts for ambiguously mapping reads and first infers KIR gene copy number and then uses this information to constrain a likelihood function for identifying the combination of KIR alleles that best explains the observed read mixture. KIR*BLOOM was designed to be cohort-agnostic and therefore does not require manual parameter tuning or predefined thresholds. Moreover, it focuses on recovering the correct coding sequence (CDS) of KIR alleles and further refines allele sequences through variant calling.

Results

Development of high accuracy KIR genotyping method

We developed KIR*BLOOM, a method for inferring KIR gene copy numbers and alleles from short read sequencing data. We formulate KIR genotyping as an optimization problem over copy-number vectors over alleles of multiple genes; For possible copy-number configurations across candidate alleles, we seek the configuration that best explains the observed read alignments. For any proposed copy-number vector, KIR*BLOOM constructs, for each allele, a coverage profile and error profile (reflecting mismatches, Insertion, deletions and clipping in aligned reads) by probabilistically assigning reads to alleles, thereby accounting for ambiguous mappings in the highly homologous KIR locus. The resulting allele-specific coverage and error profiles are then scored by comparing them to a background expectation estimated from a stable reference region on chromosome 17 (Prodanov et al. 2025), and statistical optimization is carried out to find the copy number vector associated with the best overall coverage and error profile (Figure 1).

KIR*BLOOM begins by extracting KIR-candidate read pairs from WGS. Then it generates all possible read alignment to a KIR*BLOOM allele database from the KIR*BLOOM reference. The

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

KIR*BLOOM reference combines (i) a KIR*BLOOM allele database which is derived from IPD-KIR, completing CDS-only alleles by inferring missing segments from the closest related alleles, and further expanded with additional alleles extracted from T2T and GRCh38 and (ii) the T2T and GRCh38 human genomes in which the KIR locus is masked, thereby discouraging spurious mappings to KIR alleles (Methods, Supplementary Notes). The resulting alignments are processed through three stages: (1) allele filtering to reduce the solution space and runtime, (2) copy-number estimation using a likelihood model that selects the gene copy-number vector that yields a read-depth profile most consistent with the expectation, and (3) exon-level (five-digit) allele inference, in which the inferred copy numbers impose explicit bounds on the downstream allele-resolution step; during exon-level inference, the same likelihood framework already used for copy number estimation is used together with the error profile rather than depth alone (Supplementary Figure 1).

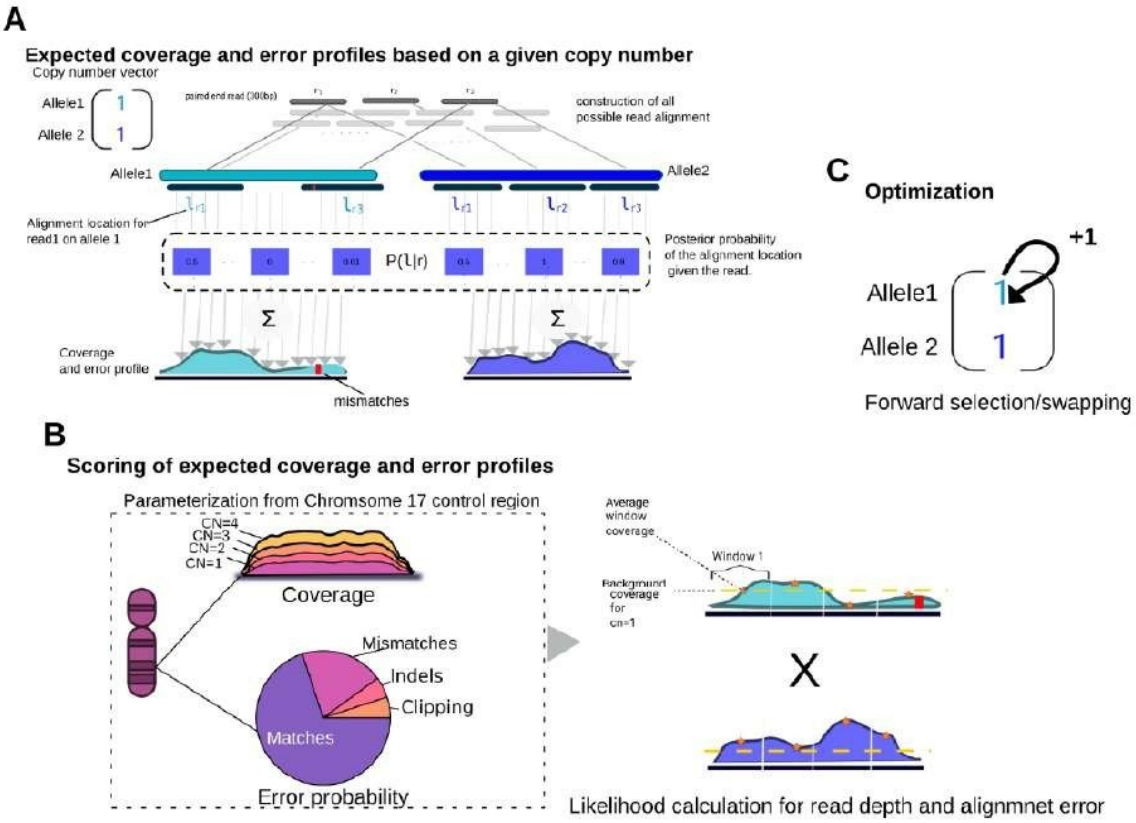


Figure 1. Overview of the three main steps of the KIR*BLOOM algorithm.

(A) For a proposed allele copy-number configuration, KIR*BLOOM probabilistically assigns ambiguously mapping reads across candidate alleles and constructs allele-specific expected coverage and alignment-error profiles.

(B) These profiles are scored against background expectations estimated from a stable control region on chromosome 17, incorporating both read depth and alignment-error patterns, including mismatches, insertions, deletions, and clipping.

(C) KIR*BLOOM then optimizes the copy-number configuration by forward selection and swapping to

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

identify the solution that best explains the observed read alignments. Together, these three steps form the core inference framework used for copy-number and downstream allele genotyping.

Performance on gene copy number

We ran KIR*BLOOM on short-read whole-genome sequencing data from 97 samples with available haplotype-resolved assemblies from the HPRC and HGSVC resources (Supplementary Table 1). Of these, 52 samples were used for method development and 45 for independent evaluation; all presented evaluation results are based on the 45 independent samples. The assemblies (Supplementary Table 1) were used to derive ground-truth genotypes with Immuannot (Zhou et al. 2024b). For comparison, we evaluated three currently available methods, T1K, Geny, and GraphKIR, on the 45 independent evaluation samples. For each method, we assessed agreement between the predicted and reference gene copy numbers (CN) using absolute CN accuracy, precision, recall, and the Jaccard index (Methods). Absolute CN accuracy was defined, for each gene, as the proportion of samples in which the predicted copy number exactly matched the ground truth for each gene (i.e., for each gene, an exact-match indicator of 1 for a match and 0 otherwise, summed over all samples and divided by the total number of samples). Metrics were computed separately for each gene, except for KIR2DL5A and KIR2DL5B, whose copy numbers were aggregated and evaluated jointly as KIR2DL5.

KIR*BLOOM and T1K achieved 100% gene detection accuracy, correctly identifying the presence or absence of every gene in the reference set across all evaluated samples. GraphKIR also achieved 100% accuracy for all genes except KIR2DS1, for which accuracy dropped to 84.4%. Geny showed the lowest overall gene detection performance, with an average accuracy of 98.3% and a minimum per-gene accuracy of 93% for KIR2DS1 (Supplementary Table 2). For exact copy number prediction for each gene, KIR*BLOOM showed very high copy-number prediction performance, achieving an overall absolute accuracy of 99.58%, compared with 93.61% for T1K, 93.4% for Geny, and 97.50% for GraphKIR (Figure 2; Supplementary Figure 2). Across 867 gene copy-number calls, KIR*BLOOM made only 2 errors (2 false positive) and missed one call (1 false negative), which we investigated in detail. One error involved overestimation of the copy number of KIR2DS4 by one. This occurs due to a KIR3DL1–KIR2DS4 fusion gene (Hung et al. 2024) which is absent from the reference dataset (IPD-KIR) and couldn't be detected by immuannot. For the other two errors, there was no clear reason. We further assessed whether excluding novel alleles derived from the GRCh38 and T2T assemblies affected the results; however, the results remained identical.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

A

	KIR*BLOOM	T1K	GraphKIR	Geny
Absolute CN accuracy	99.58	93.61	97.5	93.4
Precision	99.84	94.86	97.76	94.34
Recall	99.91	97.38	99.04	97.17
Jaccard	99.76	92.44	97.03	91.9

B

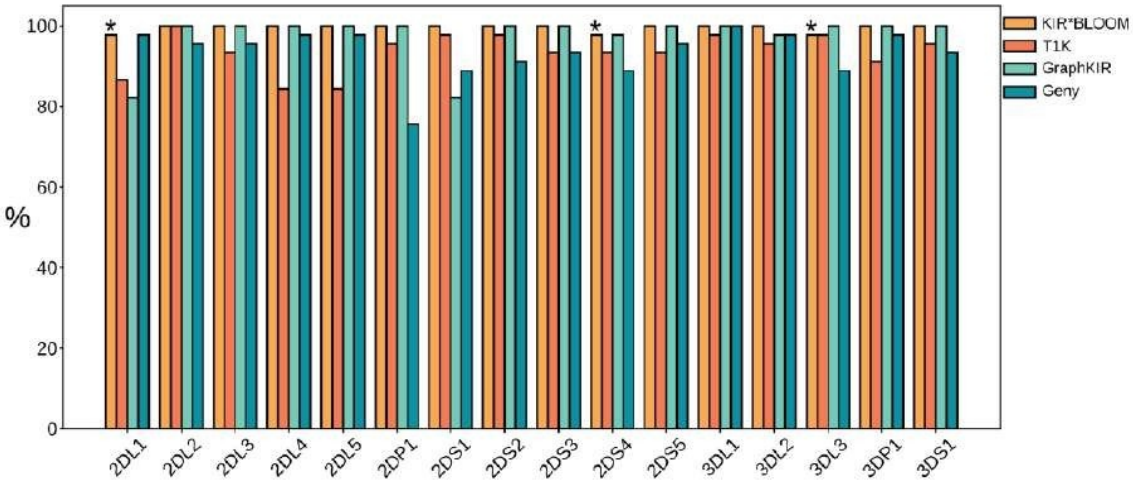


Figure 2. Evaluation of gene copy-number inference performance.

(A) Summary of overall copy-number performance for KIR*BLOOM, T1K, GraphKIR, and Geny. Performance metrics were computed separately for each gene and then averaged across genes.

(B) Gene-stratified absolute copy-number accuracy for each method, shown as the percentage of exact copy-number matches relative to the ground truth. An asterisk indicates genes for which KIR*BLOOM did not achieve 100% performance.

Performance on five-digit allele inference

We next evaluated performance for five-digit allele inference, which reflects prediction of the correct coding-sequence (CDS). We assessed our method using allele match, recall, precision, and the Jaccard index (Methods). Allele match was defined per sample and per gene as 1 if all alleles were predicted correctly with their corresponding copy numbers, and 0 otherwise. KIR*BLOOM achieved the best average five-digit typing performance using the KIR*BlOOM allele database (Methods). For allele match specifically, KIR*BLOOM achieved 90.95%, notably exceeding T1K and Geny, which achieved 80.71% and 73.47, respectively, while GraphKIR showed the lowest performance at 62.52% (Table 1).

Table1: Comparison of KIR*BLOOM, T1k, GraphKIR and Geny on 45 evaluation samples.

	KIR*BLOOM		T1K		GraphKIR		Geny	
	Before VC	After VC	Before VC	After VC	Before VC	After VC	Before VC	After VC
Allele match	88.09	90.95	78.16	80.71	62.52	-	73.47	-

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Jaccard	83.06	87.29	77.8	80.76	57.33	-	72.33	-
Precision	90.05	92.73	85.53	87.31	70.33	-	82.42	-
Recall	89.98	92.69	89.09	90.89	70.8	-	84.59	-

Stratified by gene, KIR*BLOOM outperformed GraphKIR across all genes and all evaluated metrics. Compared with T1K after variant calling and correction, KIR*BLOOM achieved higher precision (13 genes), allele match (16 genes), recall (9 genes), and jaccard index (15 genes) (Figure 3). T1K showed a higher jaccard index for two genes, KIR3DP1 and KIR2DL2, by 2.7% and 7%, respectively. However, although T1K performed better for KIR3DP1, this gene, together with KIR3DL3, had the highest number of errors overall for all evaluated methods. Compared with Geny, KIR*BLOOM outperforms it in 15 genes for precision, 14 genes for recall, 15 genes for allele match, and 15 genes for the Jaccard index. For the Jaccard index, Geny outperformed KIR*BLOOM for the genes KIR2DL4 and KIR2DL5A by 1% and 2.3%, respectively. For genes KIR2DL3, KIR2DS1, KIR2DS3, KIR3DS1, and KIR3DL1, KIR*BLOOM achieved perfect performance (100%) across all evaluated metrics. In contrast, T1K achieved perfect performance across all evaluated metrics only for KIR2DL2. For this gene, KIR*BLOOM made a single error involving a novel allele, where it failed to identify the correct variant. GraphKIR achieved perfect performance for KIR3DS1, like KIR*BLOOM, whereas Geny did not achieve perfect performance for any gene.

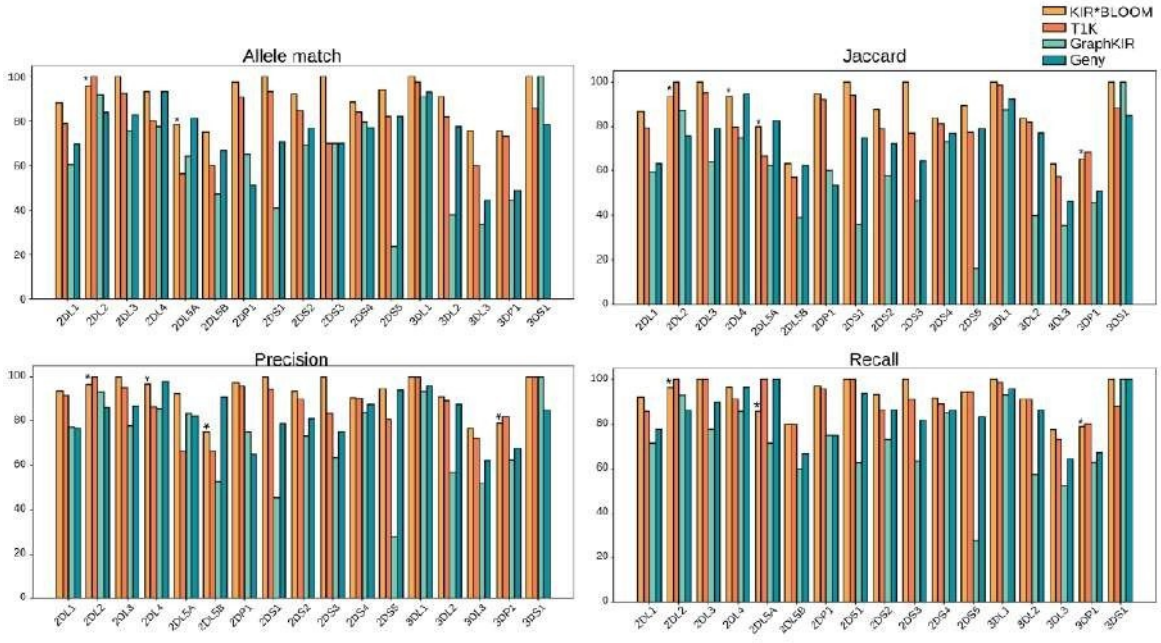


Figure 3. Gene-wise performance of KIR inference at the 5-digit level.

An asterisk indicates genes for which KIR*BLOOM was outperformed by at least one other method.

We then evaluated the performance of the different algorithms when using only an IPD-KIR-based database (i.e., not including the alleles in the extended KIR*BLOOM allele database) and without variant calling; here, KIR*BLOOM outperforms all tools on all metrics, with the exception of recall,

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

where T1K achieves the best overall result (90.89% compared to 88.81% for KIR*BLOOM) (Supplementary Table 3). Still based on the IPD-KIR database, we then evaluated the effect of variant calling, which is only supported by KIR*BLOOM and T1K. In this comparison, KIR*BLOOM outperformed T1K on all metrics including recall. Finally, comparing KIR*BLOOM using an IPD-only database to its full extended database showed an average performance increase of 1% across all metrics. Overall, the inclusion of additional alleles increases true positives by eight. The additional alleles contributed most notably to improved performance for KIR3DL3. This improvement was primarily driven by the addition of a complete KIR3DL3*005 to the KIR*BLOOM allele database; in-depth investigation on the development set (which was not used in any validation experiments) showed that the IPD-KIR version of this allele lacked intron 1. We therefore added a corrected version that included intron 1. This increased the number of true positives by five. Another added allele that contributed to the improvement was KIR2DL5B*00601, which contained a novel intronic variation compared with the IPD-KIR version. Together, these results indicate that the improved performance of KIR*BLOOM reflects an improved performance of the underlying genotyping algorithm as well as that of an improved database.

Next, we characterized the effect of the variant calling step, which may, in principle, (i) enable the detection of novel alleles not present in the database; (ii) correct errors made by allele selection step; (iii) introduce false-positive variants. Across the analyzed samples, 39 alleles carried a novel coding sequence (CDS) (Supplementary Table 4). Of these, KIR*BLOOM method correctly recovered 10, whereas T1K recovered 17. Although T1K recovered more novel CDS sequences than our method, both approaches resolved fewer than half of all novel CDS sequences. For KIR*BLOOM, recovery was particularly difficult for genes with copy numbers greater than one were the method only reports unambiguous allele call (Methods). Overall, variant calling reduced the number of false negatives for both methods. For KIR*BLOOM, the number of false negatives across all loci decreased from 89 to 73, corresponding to an approximately 18% reduction, and performance increased across all evaluation metrics by an average of 3%. For T1K, false negatives decreased from 109 to 92, corresponding to an approximately 16% reduction. However, For T1K, this improvement was partly offset by unnecessary sequence changes introduced into alleles whose CDS sequences had already been correctly assigned. For example, this could occur for homozygous alleles where sequence modifications were applied based on the estimated allele frequencies (Methods).

The IPD-KIR database includes alleles with (i) undefined intron sequences; (ii) incomplete exon sequences; or (iii) both. When constructing the KIR*BLOOM allele database, the missing regions were interpolated from closely related alleles (see Methods and Supplementary Notes). We investigated whether KIR*BLOOM could recover the alleles' complete CDS in cases in which the

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

evaluation samples carried these incomplete alleles, because these interpolated regions may not fully represent the true sequence, alignments to such alleles may be suboptimal. We therefore assessed whether the CDS of these incomplete alleles could still be identified correctly. Across all samples, 8 occurrences representing 5 alleles involved partially or completely missing exons cases. KIR*BLOOM correctly recovered the CDS sequence in all 8 instances, whereas T1K recovered 2 instances corresponding to KIR2DS4*00104. GraphKIR and Geny missed them all (considering the predicted allele name and not CDS). Among alleles with complete CDS sequence but missing intronic sequence, there were 74 instances across the samples. KIR*BLOOM recovered the correct CDS sequence in 60 instances, compared with 51 instances for T1K, 19 instances for GraphKIR and 30 instances for Geny (Supplementary Table 5).

Analysis of KIR bloom error mode

KIR*BLOOM produced 74 false positives and 43 false negatives. In nearly all cases, each false negative was accompanied by a corresponding false positive, indicating that most errors arose from allele misassignment rather than complete omission. There were three exceptions: one resulted from underestimation of the KIR2DL1 copy number by one, and two involved switching between KIR2DL5A and KIR2DL5B alleles. In the latter two cases, a false positive was still generated for each false negative, but not within the same gene. In addition, two false positive calls arose from copy number overestimation: one for KIR3DL3 and one for KIR2DS4. The KIR2DS4 overcall is explained by a KIR3DL1–KIR2DS4 fusion that is absent from the IPD-KIR reference and therefore was not detected by Immuannot.

To understand the causes of the 73 false negatives produced by KIR*BLOOM, we classified them into four categories: loss of the true allele during solution-reduction filtering, failure to recover a novel coding sequence (CDS) by variant calling, ambiguous genotype configurations, and other gene-specific causes.

The most frequent error class was failure to rescue novel alleles by variant calling (Figure 4), which accounted for 39.7% (n = 29) of all false negatives. In KIR*BLOOM, a variant is called only when at least 75-85% of aligned reads support the sequence change (Methods). This stringent threshold can prevent recovery of novel CDS sequences when reads from the novel allele are distributed across multiple selected alleles, reducing the apparent support for the variant below the calling threshold. This problem was particularly pronounced in KIR3DP1, which had the highest number of novel CDS alleles (n = 9). In these cases, the presence of a novel allele often also led to incorrect selection of the second, non-novel allele. The second most common error class was ambiguous genotype configurations, which represented 32% of all false negatives (n=24). An ambiguous genotype occurs

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

when two different pairs of alleles produce an identical diploid base-level sequence, while the heterozygous sites distinguishing the two configurations are separated by distances greater than the maximum fragment length, preventing read-based phasing (Supplementary Figure 3). Consequently, both allele pairs explain the data equally well and reads align to both without alignment errors on the exons. KIR*BLOOM therefore selects the allele pair with the highest likelihood, even when the improvement over the alternative combination is only marginal. This pattern was especially overrepresented in KIR3DL3, where it explained 18 of the 20 false negatives observed for this gene. It was also seen in KIR3DP1 (n=2) and KIR3DL2 (n=4), accounting for 10.5% and 50% of their false negatives, respectively.

The remaining false negatives were mainly caused by allele loss during filtering and another set of heterogeneous causes which mostly reflected limitations of the reference representation rather than a systematic weakness of the inference model. In several cases, incomplete reference sequences, particularly missing intronic regions in the IPD-KIR database, led to suboptimal read mapping and caused the correct allele either to be filtered out or not selected during inference. This was observed for KIR2DS2*005, KIR3DP1*004 (n=3), KIR2DL5A*01201, KIR3DL2*00601, and KIR3DL2*01004. Additional errors occurred when the presence of a novel allele led to loss of the second allele in the genotype. This was most frequent in KIR3DP1 (n=5) and was also observed in KIR3DL3 (n=1) and KIR2DL1 (n=1). The remaining false negatives arose from several sources. One was caused by copy-number misassignment during copy-number inference, as mentioned earlier. Another involved KIR2DL1, where true structural variation in exon composition made the genotype inherently ambiguous. Two additional cases reflected confusion between the highly similar paralogous genes KIR2DL5A and KIR2DL5B, with one misassignment in each direction. Finally, two false negatives, in KIR2DL1 and KIR2DS5, had no clear explanation.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

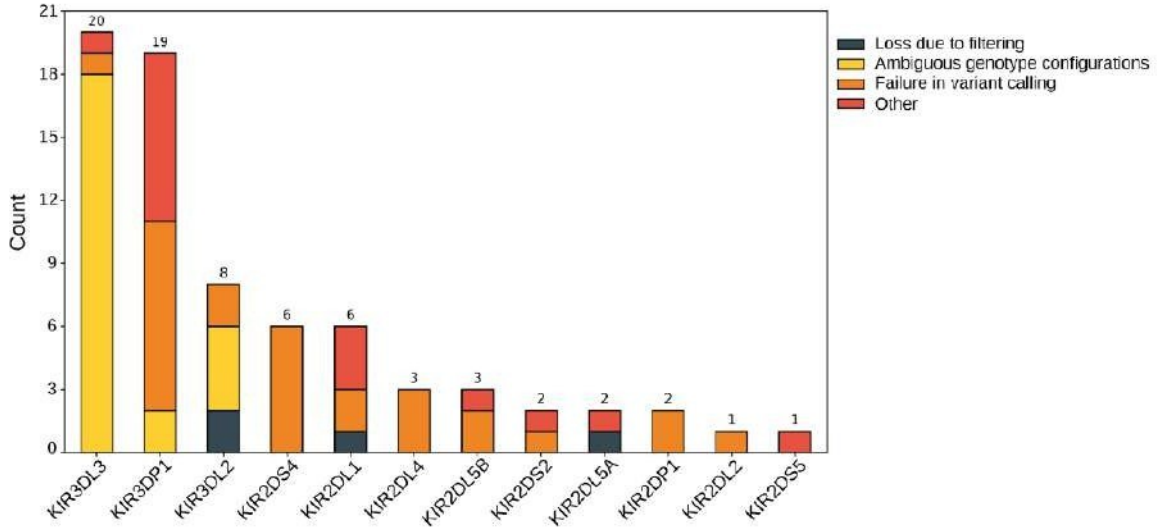


Figure 4. Gene-wise stratification of KIR*BLOOM false negatives by source.

Discussion

Variations in KIR genes shapes NK-cell immunity. Accurate genotyping of the KIR locus remains technically challenging because the region combines extensive gene-content variation, high allelic polymorphism, copy-number variation, and strong sequence homology between related genes, and these challenges are further compounded by reference databases that do not yet fully represent the breadth of KIR allelic diversity. In this study, we developed KIR*BLOOM and evaluated its performance on whole-genome sequencing data from 97 real-world samples with available assemblies. We show that KIR*BLOOM addresses several of the major challenges of KIR genotyping by enabling accurate inference of gene presence and absence, showing an excellent performance for gene copy-number inference, and outperforming existing methods in CDS-level sequence prediction, particularly through the use of variant calling to recover correct CDS sequences. Nevertheless, two important limitations remain, both largely driven by incomplete reference representation: genotype ambiguity and the presence of novel alleles.

The strong performance of KIR*BLOOM was achieved in part through the combination of an extended reference database and variant calling. Interpolating regions missing from the IPD-KIR reference using closely related alleles in the extended database improved reference completeness. This enabled correct prediction in cases where the missing regions included CDS sequence, which was a factor limiting the performance of T1K, and also improved detection of alleles with missing sequence in non-coding regions by facilitating read mapping and allowing such alleles to remain under consideration during inference. Nevertheless, this approach does not fully resolve all cases,

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

because when the inferred reference sequence remains substantially different from the true allele, alignment can still be suboptimal, leading to failure in correct allele prediction. Variant calling is an important extension that remains absent from many KIR genotyping methods, although it is implemented in T1K. Here, we applied stringent criteria for sequence correction to minimize false variant calls, particularly in genes with copy number greater than one, where reads from a novel allele may be distributed across multiple inferred alleles and generate misleading evidence for sequence changes. We therefore prioritized specificity and called variants only when supported by strong evidence. Although this conservative strategy reduced recovery of true novel variants to approximately 25%, it limited erroneous sequence corrections and preserved the overall reliability of allele inference.

The predominant source of error was genotype ambiguity, which had the greatest impact on KIR3DL3, KIR3DL2, and KIR3DP1. KIR3DL3 and KIR3DP1 also showed comparatively weak performance across other methods, suggesting that these cases reflect a broader difficulty of KIR genotyping from short-read data. To some extent, this ambiguity is intrinsic to short reads, because closely related alleles may not always be separated by distinguishing variants within the sequenced fragments. In principle, some ambiguities could be resolved by borrowing information from linked intron variants. In practice, however, this strategy was of limited value because intronic regions often contained substantial sequence variation not represented in the reference database, resulting in a high density of mismatches that made reliable phasing difficult. This problem was further compounded by the fact that many alleles in the database lack complete intronic sequence information. Consequently, adding intronic sequence alone is unlikely to fully resolve these ambiguous genotypes. A more promising strategy may be to incorporate cohort-level information for statistical phasing, as implemented in PING, which we plan to explore in future work.

Overall, KIR*BLOOM achieved strong performance across the 97 real-world WGS samples analyzed in this study. The main remaining limitation appears to be incomplete representation of KIR diversity in current reference resources rather than the inference framework itself. Continued expansion of the IPD-KIR database, together with advances in long-read sequencing, should enable more complete characterization of full-length KIR haplotypes, including intronic and structurally complex alleles, across larger and more diverse cohorts. As long-read sequencing becomes increasingly scalable and cost-effective, these data should substantially improve reference completeness and reduce many of the ambiguities that still limit high-resolution genotyping.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Methods

KIR genotyping model

Our method performs KIR genotyping from short-read sequencing data based on a likelihood model that evaluates read depth and sequencing error on KIR alleles under alternative KIR genotype configurations. As a first step, KIR*BLOOM identifies relevant read pairs, maps them against the KIR*BLOOM allele database, and heuristically reduces the solution space by identifying and removing KIR alleles that are likely absent from the genotyped sample. Within the reduced solution space, KIR*BLOOM determines the likely copy number of each gene, followed by selection of actual KIR alleles, in a manner bounded by the inferred copy numbers. In the following, we first describe the core statistical model of KIR*BLOOM that is used for copy number inference and allele selection; the heuristic steps for reducing the solution space and the method for identifying KIR-related read pairs are described later.

Let R denote the set of KIR read pairs, let $B = \{b_1, \ldots, b_n\}$ denote the set of KIR alleles from the KIR*BLOOM reference database remaining after heuristic filtering, and let c be a copy number vector of length n, where the i-th element represents the copy number of KIR allele b_i, defined to be between 0 and 4. Our aim is to find a configuration of c that maximizes the likelihood of the coverage and read error profiles of the alleles in B, where we account for alignment ambiguity of the read pairs in R by considering the probabilities of their alignment locations conditional on c.

In the following, we first describe the initial KIR read mapping step and how we probabilistically distribute individual read pairs r over the alleles in B, conditional on c. We then describe how we aggregate read alignments at the level of the KIR alleles in B, obtaining a coverage and read error profile for each allele in B, conditional on c.. Last, we describe how assess consistency between the KIR alleles' coverage and read error profiles that result from a particular instantiation of c on the one hand and sample genome-wide sequencing depth and sequencing error rate on the other hand using a likelihood model, and how we explore the space of possible configurations of c to maximize this likelihood.

Computation of the alignment probability distribution (A_r) for an individual read pair r

Let $L_r = \{(x,y) | x \in L_{(r_1)}, y \in L_{(r_2)}\}$ be the set of possible paired-end mapping locations of read r against the set B of KIR alleles remaining after heuristic filtering; L_r is generated by mapping KIR reads against the KIR*BLOOM database in single-ended mode, considering all possible pairs of individual read alignments, and by filtering these for correct strandedness, insert size etc.; for details.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

see section "Alignment of KIR-related read pairs" for details. The read likelihood P(r|l), representing the probability of observing r if it originated from location $l \in L_r$ on KIR allele $b(l)$ is defined as:

$$P(r|l) = \frac{1}{T_{b(l)}} \cdot P_M^m \cdot P_X^x \cdot P_I^i \cdot P_D^d \cdot P_C^c$$

where m, x, i, d, c represent the number of match, mismatch, insertion, deletion and clipping operations in alignment l, respectively, and P_M, P_X, P_I, P_D, P_C denote the probabilities of observing a match, mismatch, insertion, deletion, and clipping event, respectively (see section "Parameter estimation"). $T_{b(l)}$ is the length of allele b, included as a normalization term to account for differences in allele lengths.

We then define the probability that read pair r emanated from l as:

$$A_r(l) = P(l|r;c) = \frac{\tilde{c}_{b(l)} \times P(r|l)}{\sum_{l' \in L_r} \tilde{c}_{b(l')} \times P(r|l')} \quad \text{Eq. 1}$$

, with $\tilde{c}_{b(l)} = \begin{cases} \epsilon & , c_{b(l)} = 0 \\ c_{b(l)}, & c_{b(l)} > 0 \end{cases}$

Setting ϵ to a small value (> 0) allows for the assignment of reads to alleles with copy number 0. $A_r(l)$ for alignments on alleles $b(l)$ with copy number 0 will always be small, unless r has no other possible alignment locations. Reads aligning to alleles with copy number 0 are considered inconsistent with the current configuration of c and therefore penalized during later stages of inference. The specific value of ϵ varies during different stages of inference (see below).

Computation of allele alignment profiles P_b

Denote the allele alignment profile for allele b as $P_b = (P_b^{cov}, P_b^{err})$, where P_b^{cov} and P_b^{err} are two vectors of length T_b that represent the coverage and sequencing error for all positions on b. Formally, for position i of allele b, let $L_{b,i}$ denote the set of alignment locations of all reads that overlap position i. Each $l \in L_{b,i}$ has an associated read pair $r(l)$, and the probability l is the alignment location of read pair $r(l)$ is $A_{r(l)}(l)$ (see previous section).

Coverage at positions i is defined as

$$P_{b,i}^{cov} = \sum_{l \in L_{b,i}} A_{r(l)}(l)$$

That is, $P_{b,i}^{cov}$ is the expected value of reads covering position i, where the expectation is taken over A_r for all relevant reads. Similarly, cumulative read error is computed as

$P_{b,i}^{err} = \sum_{l \in L_{b,i}} A_{r(l)} * 1_{(\text{alignment_carries_an_error_at_i})}$, where $1_{(\text{alignment_carries_an_error_at_i})}$ is an indicator function that is 1 if and only if the aligment l carries an error at position i.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Likelihood function for scoring allele alignment profiles

KIR*BLOOM uses a likelihood model to assess consistency between the alleles' coverage and read error profiles, conditional on c, and genome-wide estimates of sequencing depth (adjusted for copy number) and sequencing error.

Define $P_B = \{P_b : b \in B\}$ as the set of allele alignment profiles of the alleles in B. The likelihood of these, computed conditional on c, is defined as

$$L(c; P_B) = \prod_{b \in B} \left(L_{cov}(c_b; P_b) x L_{err}(c_b; P_b) \right),$$

where $L_{cov}(c_b; P_b)$ and $L_{err}(c_b; P_b)$ represent coverage and read error likelihoods, respectively; the specific functional form of the utilized likelihood depends on the stage of inference (copy number inference or allele selection; for details, see section "KIR genotype inference").

The coverage likelihood $L_{cov}(c_b; P_b)$ for an allele b is computed in a window-based manner. Windows are defined separately for intronic and exonic positions and ignoring the UTR; for full details, see section "Window construction". Let w denote a window, defined as a set of (exclusively intronic or exonic) positions on allele b, and let W_b be the set of all windows associated with b. By definition, the union of the positions in W_b coveres the non-UTR positions of b; UTRs sequences are ignored because they are generally less well-represented in IPD-KIR. Let is_exon(w) be an indicator function that equals 1 if and only if w is an exonic window, and 0 otherwise. The average coverage of a window w is defined as

$$\overline{P_b^{cov}}(w) = \frac{\sum_{i \in w} P_{b,i}^{cov}}{|w|}$$

The coverage likelihood $L_{cov}(c_b; P_b)$ is defined as

$$L_{cov}(c_b; P_b) = \prod_{\{w \in W_b : \text{is_exon}(w)=1\}} \left(\text{P}_{\text{Poisson}} \left(\overline{P_b^{cov}}(w) \middle| \lambda = \text{expectedCoverage}(c_b) \right) \right) * \prod_{\{w \in W_b : \text{is_exon}(w)=0\}} \left(\text{intron_weight}(w) * \text{P}_{\text{Poisson}} \left(\overline{P_b^{cov}}(w) \middle| \lambda = \text{expectedCoverage}(c_b) \right) \right)$$

That is, $L_{cov}(c_b; P_b)$ is computed over the windows of b, treating their average coverages as independent, and separately for intronic and exonic windows. The assumed expected coverage for each window depends on the underlying allele's copy number c_b; if $c_b > 0$, expectedCoverage(c_b) is defined as $c_b * D$, where D is the estimated haploid genome-wide sequencing depth (see section "Parameter estimation"); if $c_b = 0$, expectedCoverage(c_b) is defined

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

as $0.01 * D$, penalizing read alignments to alleles with copy number 0 while always retaining a non-zero likelihood for such alleles. Finally, intron_weight(w) determines whether the inference process uses coverage information from intronic windows; it is defined as 0 or 1, depending on the stage of inference (section "KIR genotype inference").

For alleles with $c_b > 0$, the error likelihood $L_{err}(c_b; P_b)$ for an allele b is computed over the exonic positions of the allele. Let E_b denote the set of exonic positions, and define

$$I_b = \{i \in E_b : \frac{P_{b,i}^{err}}{P_{b,i}^{cov}} > 0.2\}$$

; that is, I_b represent the set of exonic positions with error rate > 0.2. We then define $L_{err}(c_b; P_b)$ as

$$L_{err}(c_b; P_b) = \prod_{i \in I_b} P_{\text{Binomial}}\left(P_{b,i}^{err} | P_{b,i}^{cov}, P_X + P_I + P_D\right)$$

where P_X, P_I, and P_D denote the empirical probabilities of observing a mismatch, insertion, and deletion event, respectively (see section "Parameter estimation"). For alleles with $c_b = 0$, $L_{err}(c_b; P_b)$ is defined as 1. Of note, our error model implicitly assigns a likelihood of 1 to all exonic positions with an error rate ≤0.2; this is motivated by increasing robustness against background sequencing error and also to reduce the overall number of positions considered for alleles with $c_b > 0$, as each considered position implicitly penalizes the presence of the allele.

KIR genotype inference

KIR genotype inference is carried out in two stages, copy number inference and 5-digit allele selection.

- During copy number inference, sequencing errors are ignored ($L_{err}(c_b; P_b) := 1$), intronic positions contribute to L_{cov} (i.e. $intron_weight(w) = 1$), and $\epsilon = 10^{-4}$.

  To optimize $L(c; P_B)$ over the space of possible c, we employ a forward selection strategy. We initialize c^0 as a vector of length $|B|$, with $c_i^0 = 0$ for all $i \in \{1..|B|\}$. During iteration k, we then evaluate, for all $i \in \{1..|B|\}$, $L(d^i; P_B)$, where $d^i = c^{k-1}$, except for $d_i^i = c_i^{k-1} + 1$. c^k is then set to the d^i that achieved the highest likelihood. The process terminates when the maximum log likelihood of iteration k does not increase compared to iteration $k-1$.

  To accelerate inference, all copy number inference steps are executed on a reduced subset of alleles $B' \subseteq B$, selected based on an EM algorithm that estimates an overall abundance F_b for each allele $b \in B$ in the set of all reads R. B' is constructed from B in a gene-wise manner,

iv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint
ch was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is
made available under a CC-BY 4.0 International license.

retaining, from all alleles B_g belonging to KIR gene g, the allele b_g^* that achieved the highest abundance F_b (i.e. $b_g^* = \max_{b \in B_g} F_b$), as well as all other alleles $b_g' \in B_g$ with $\frac{F_{b_g'}}{F_{b_g^*}} \geq 0.1$.

- During 5-digit allele selection, sequencing errors are taken into account (using the definition of $L_{err}(c_b; P_b)$ given in the previous section), introns are not used in the process of computing L_{cov} (i.e. $intron_weight(w) = 0$), and $\epsilon = 10^{-300}$.

  Like copy number inference, allele selection operates on a reduced subset of alleles $B'' \subseteq B$; B'' is constructed from B by removing all elements corresponding to genes with estimated copy number 0. Let c^{AS} denote the copy number vector of length $|B''|$ used for allele selection; we initialize c^{AS} by setting, for all genes with estimated copy number 1, the element corresponding to the allele with the highest EM-estimated abundance to 1; for genes with estimated copy number > 1, the elements corresponding to the alleles selected during copy number inference are set to the corresponding estimated copy number; and all other elements of c^{AS} are set to 0.

  To optimize $L(c^{\mathrm{AS}}; P_{B''})$ over the space of possible c^{AS}, we employ a multi-level allele swapping strategy; copy number per gene is held constant over all steps.

  First, we determine the set of genes G^* for which alleles currently present in c^{AS} may not represent the true genotype. To do so, we check, for all alleles b belonging to g (i.e. for all $b \in B''_g$) whether i) there is at least one exonic position with an error rate ≥ 0.2 (i.e. whether $|I_b| > 0$) or whether ii) there is at least one window $w \in W_b$ for which $\overline{P_b^{cov}}(w)$ deviates significantly (two-sided Poisson test, $p = 0.01$) from $\mathrm{expectedCoverage}\left(c_b^{\mathrm{AS}}\right)$; if any of the two conditions is met for alleles b, the corresponding gene g is added to G^*.

  Second, $L(c^{\mathrm{AS}}; P_{B''})$ is optimized using one-allele swapping between alleles belonging to different 5-digit allele groups. Swapping is implemented in a gene-wise manner, i.e. during each iteration, we sequentially process all $g \in G^*$. Within each gene g, we process all pairs (b^1, b^2) of alleles in g (i.e. $(b^1, b^2) \in \left\{B''_g\right\}^2$) for which $b^1 \neq b^2$, for which b^1 has a current copy number > 0, and for which b^1 and b^2 belong to different 5-digit allele groups; we then decrease the copy number of b^1 by 1 and increase the copy number of b^2 by 1 (i.e., we swap a copy of b^1 for a copy of b^2) and evaluate the likelihood. After having processed all genes, we change c^{AS} to the configuration that achieved the highest likelihood (or leave c^{AS} unmodified, if no swapping operation was associated with an increase in likelihood). One-

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

allele swapping terminates if one complete iteration over all $g \in G^*$ did not result in any swaps.

Third, $L(c^{AS}; P_{B''})$ is optimized using two-allele swapping. Like one-allele swapping, two-allele swapping is implemented in a gene-wise manner and only between alleles belonging to different 5-digit allele groups. Within each gene g, we process all pair of pairs $\big((b^1, b^2), (b^3, b^4)\big)$ of alleles in g (i.e. $(b^1, b^2) \in \{B''_g\}^2, (b^3, b^4) \in \{B''_g\}^2$) for which $b^1 \neq b^2$; $b^3 \neq b^4$; $b^1$and b^2 have a current copy number > 0; b^3 and b^4 have a current copy number of 0; and for which the sets of 5-digit allele groups associated with (b^1, b^2) and (b^3, b^4) are not identical. For each pair of pair, we decrease the copy numbers of b^1 and b^2 by 1, increase the copy numbers of b^2 and b^3 by 1, and evaluate the likelihood. After having processed all genes, we change c^{AS} to the configuration that achieved the highest likelihood (or leave c^{AS} unmodified, if no swapping operation was associated with an increase in likelihood). Two-allele swapping terminates if one complete iteration over all $g \in G^*$ did not result in any swaps.

EM algorithm

We use the EM algorithm to reduce the size of the allele set considered for copy number inference, and also for seeing the allele selection step. Briefly, each KIR allele $b \in B$ is assumed to have an abundance F_b over all read pairs $r \in R$, and each read pair r is assumed to have a hidden variable that indicates the KIR allele that read pair r emanates from. We define $P(r|b)$ as identical to the highest-likelihood mapping location on allele b, i.e. $P(r|b) = \max_{l \in \{l' \in L_r:\, b(l')=b\}} P(r|l)$. The probability that read r emanated from allele b is then defined as $P(b|r) = \frac{F_b \times P(r|b)}{\sum_{b' \in B} F_{b'} \times P(r|b')}$, and the overall likelihood is $L(R) := \prod_{r \in R} F_b \times P(r|b)\ F_b$ is iteratively maximized using EM, i.e. by computing $P(b|r)$ based on the current estimate of F, followed by updating F based on $P(b|r)$ for all $r \in R$, until the increase in $\log L(R)$ falls below 0.01.

Parameter estimation

A copy-number–stable background region of 2Mbp on GRCh38 chromosome 17 (specific coordinates: chr17:74Mb-76Mb) (Prodanov et al. 2025) is used for parameter estimation at multiple stages of the pipeline. Alignment event probabilities are estimated by counting matches (m), mismatches (x), insertions (i), deletions (d), and clipped bases (c) from primary paired-end read alignments and by normalizing these counts by their total to obtain P_M, P_X, P_I, P_D, P_C, respectively. The chromosome 17 control region is also used for estimating haploid genome-wide sequencing depth (parameter D) as average per-base sequencing depth, divided by 2.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Identification of KIR-related read pairs

KIR*BLOOM supports BAM and CRAM files aligned to 1000 Genomes GRCh38 reference genome, as well as raw reads in FASTQ format; if FASTQ format is used, reads are first aligned to the 1000 Genomes reference using bwa-mem2.

Based on GRCh38 read alignments, a first set of candidate KIR read pairs is obtained by extracting all pairs of reads for which at least one member read aligns into the KIR region on chromosome 19 (coordinates) or onto the alternative LRC/KIR haplotypes, as well as read pairs with at least one unaligned member read.

The candidate KIR read set is further refined by aligning, using minimap2, the extracted pairs against a combined reference genome that comprises i) the B38 reference with all KIR gene locations masked; ii) the T2T reference genome with all KIR gene locations masked; iii) the KIR*BLOOM allele database. Read pairs with properly paired primary alignments to the KIR*BLOOM reference database represent high-confidence KIR reads; these therefore constitute the final set of read pairs used for KIR genotyping. Masking of the two reference genomes was carried out based on manually curated KIR gene annotations (see Section "Annotation of KIR loci in GRCh38 and T2T references"). Combining the T2T and GRCh38 references during the refinement stage reduces nonspecific mappings to repetitive elements that would otherwise be propagated into the KIR genotyping step.

Alignment of KIR-related read pairs

For each read pair $r = (r_1, r_2)$ns sets $L_{(r_2)}$ and [illegible] for r_1 and r_2 by mapping them in single-ended mode against the KIR*BLOOM reference database, using minimap2 in short-read mode (-ax sr --secondary-seq --MD --eqx The set of possible paired-end alignment locations is then defined as [illegible], where only members that satisfy the following constraints are retained: (i) both alignments map to the same KIR allele, (ii) the alignments have inward-facing orientations on opposite strands, and (iii) the inferred insert size lies within the interval $[\mu-4\sigma, \mu+4\sigma]$, where μ and σ the considered set of alignment locations for read [illegible]generated by filtering B to KIR alleles present in the reduced allele set B.

Heuristic reduction of the KIR allele solution space

After identifying KIR-related reads, the first step of KIR*BLOOM is to heuristically reduce the search space by identifying KIR alleles that are likely not present in the genotyped sample, based on a two-step filtering strategy. The KIR alleles that survive the heuristic filters become the allele set B that goes into the main genotyping model.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

During the first filtering step, we remove alleles b that have at least one "core" position with alignment depth 0 when considering the set of all possible KIR read pair alignment locations $L_r^{\text{ALL_ALLELES}}$ for all reads $r \in R$. The "core" positions of an allele are defined as positions that lie i) outside of the UTR region of the allele and b) outside of the first or last 15% of the allele's sequence; note that the "alignment depth" criterion used here includes non-match alignment operations.

During the second filtering step, we focus on positions that discriminate between KIR alleles of the same gene, and we rank (and then remove) alleles according to their read support at such positions. Formally, for gene g, let the set Z denote the columns in the MSA of the alleles of g that discriminate between alleles (i.e., which contain more than one base) and which are consistently annotated with respect to the intron/exon status of the included alleles at this position in the MSA. To score allele b of gene g, we iterate through the set Z, removing all columns at which the sequence of b that was interpolated (during the construction of the KIR*BLOOM reference database; see section "Construction of the KIR*BLOOM allele reference database"); at which b carries a "gap" character; or which are part of the UTR. Let the set Z^b denote the columns in the MSA that remain after filtering and that are used for scoring allele b; furthermore, let Z^b_{intron} and Z^b_{exon} denote the intronic and exonic subsets of Z^b, respectively. Alleles with 0 remaining positions for evaluation (i.e. alleles with $|Z^b| = 0$) are always retained. For the other alleles, we count, at each position $z \in Z^b$, the number of "match" characters across all possible read alignments at the considered position (i.e., based on the set $L_r^{\text{ALL_ALLELES}}$ for all reads $r \in R$ that overlap the position), and carry out a statistical hypothesis test against the null hypothesis that the observed number of matches can be explained by sequencing error alone (one-sided Poisson with $\lambda = 1$ and significance threshold depending on the relative location of the position within b, see below). Positions at which the null hypothesis is rejected are counted as "supported", and positions at which the null hypothesis is not rejected are counted as "not supported". For each allele, we separately compute the total number of unsupported exonic and intronic positions. If there is a single allele that has a lower count of unsupported exonic positions than all other alleles within the set of considered alleles with $|Z^b| > 1$, this allele is retained, and all other alleles with at $|Z^b| > 1$ are removed. Otherwise, we consider the set of alleles that achieve the same minimum number of unsupported exonic positions and retain i) all alleles with $|Z^b_{\text{intron}}| = 0$ and ii) all alleles with $|Z^b_{\text{intron}}| > 1$ that have the minimum number of unsupported intron positions. The significance threshold used for the Poisson test is $p<0.05$ for all positions that fall into the central 70% region of each allele, and p<0.1 otherwise.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Annotation of KIR loci in the GRCh38 and T2T references

We annotated KIR loci in both the GRCh38 and T2T reference genomes based on available reference annotations (RefSeq for GRCh38 and JHU RefSeq v110 for T2T), Immuannot, and KIR allele mappings generated by aligning the IPD-KIR database with interpolated introns and UTRs (see Section "Construction of the KIR-BLOOM allele reference database") to the reference genomes. For each reference genome, we took the union of KIR loci identified by these three sources. For each locus, gene labels were determined based on Immuannot annotations, while locus start and end positions were defined using the KIR allele mappings. In addition, the annotation results for each locus were manually curated.

Construction of the KIR*BLOOM allele reference database

The KIR*BLOOM allele database was constructed using the IPD-KIR database (December 2025 release) together with novel alleles identified by Immuannot from the T2T and GRCh38 genomes (see Section "Annotation of KIR loci in the GRCh38 and T2T references"). The IPD-KIR database contains 2,286 alleles available only as coding sequences (CDS); for these, intronic sequences were interpolated from the most closely related intron-and-exon-complete alleles. Finally, UTR sequences for all alleles were completed, based on either i) the UTR sequence of the allele that intron interpolation was based on, if this allele contained a UTR and if the allele with incomplete UTR was originally only available as CDS; or ii) based on inserting UTR sequence from an arbitrary UTR-complete allele. Details of the KIR*BLOOM allele reference construction are provided in the Supplementary Notes.

Window construction

The coverage component $L_{cov}(c_b; P_b)$ of the KIR*BLOOM likelihood is computed in a window-based manner rather than at single-base resolution to reduce sensitivity to local read coverage and to account for statistical non-independence between neighboring positions. Windowing is carried out separately for each gene g and, for each allele, exonic and intronic regions are partitioned separately, and UTR regions are excluded. The exonic portion of each allele $b \in B_g$ is partitioned into contiguous windows of length 150bp for all windows except for the last window; if the last window has a size of less than 100bp, it is combined with the previous window. If, after having partitioned all alleles of gene g, there is a difference between the alleles of g with respect to the number of windows, windowing is repeated with a fixed target number of windows per allele, equal to the minimum window count achieved during the first iteration. During this second iteration, once the number of 150bp windows has achieved its allowed maximum, the last windows is extended until all relevant positions are included. Intronic positions are partitioned analogously.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Variants correction

To refine the inferred allele sequences, we applied a variant-correction procedure based on read support at exonic sites. First, for each inferred allele, we calculated the fraction of non-reference observations at each exonic position as the sum of mismatches, insertions, and deletions divided by the total depth at that site. Positions with an error fraction greater than 0.75 were flagged as candidate correction sites, and the most strongly supported alternative allele in the read pileup was recorded. Second, for genes with two inferred alleles, we used the gene-specific multiple sequence alignment to identify ungapped exonic positions at which the two alleles differed. At each such site, we compared the read support for the two allele-specific bases. When more than 85% of the support favored one base, the alternate allele was corrected to match the supported base, effectively collapsing unsupported heterozygous differences into a homozygous call. All candidate corrections were written as VCF records for downstream sequence updating.

Development and validation data

Whole-genome sequencing data were obtained from two publicly available sources. First, pre-aligned CRAM files were downloaded from the European Nucleotide Archive (ENA) FTP repository. These datasets correspond to samples for which HGSVC3 assemblies are available . All such CRAM files were aligned against the GRCh38 full analysis set plus decoy and HLA sequences reference genome, which was obtained from *https://42basepairs.com/download/s3/1000genomes/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa*. Second, additional whole-genome sequencing datasets were obtained from samples generated by the Human Pangenome Reference Consortium (HPRC). For these samples, paired-end whole-genome sequencing data (fastq files) were downloaded individually from the NCBI by querying the Sequence Read Archive (SRA) for available WGS paired-end sequencing runs.

For samples available as CRAM files, reads had already been aligned using bwa-mem2 (v2.2.1); therefore, reads mapping to KIR-associated regions were directly extracted from these alignments. For the HPRC datasets, raw sequencing reads were first aligned using bwa to GRCh38 full analysis set reference genome, including decoy and HLA sequences, after which KIR-related reads were similarly extracted. The primary KIR locus on chromosome 19 (chr19:54,025,634–55,084,318) was selected and extended by ±1 kb to capture flanking regions. In addition, alternative reference sequences corresponding to the leukocyte receptor complex (LRC) were obtained from Genome Reference Consortium resources (https://www.ncbi.nlm.nih.gov/grc/human) by selecting alternative contigs whose names contained the string LRC-KIR. Finally, a non-KIR region on chromosome 17

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

(chr17:74,000,000–76,000,000) was included as a background reference region for downstream analyses (Prodanov et al. 2025).

KIR read identification

To identify reads for KIR genotyping, we constructed a joint reference genome consisting of the T2T and GRCh38 reference genomes, in which KIR loci were masked, together with the KIR*BLOOM allele database. Input reads were mapped against this joint reference genome using BWA in paired-end mode. Read pairs for which both reads had a primary alignment to the KIR-BLOOM reference database were retained for downstream KIR genotyping. Combining the T2T and GRCh38 references at this stage reduces nonspecific mappings to repetitive elements that would otherwise be propagated into the KIR genotyping step.

Constructing the ground truth

Ground-truth KIR genotypes were derived from haplotype-resolved genome assemblies available for each sample using Immuannot and IPD-KIR release 2.15.0. After running Immuannot with default parameters on the two phased haplotypes of each sample, we parsed the resulting maternal and paternal GTF files to extract the consensus CDS annotation for each KIR gene. For every annotated KIR copy, we recorded the consensus CDS sequence together with its reported CDS distance, and then combined the maternal and paternal annotations to construct the ground-truth CDS set for each sample. From these annotations, we also derived the ground-truth gene copy number by counting the total number of annotated KIR copies across the two haplotypes. These CDS-resolved and copy-number truth sets were then used as the reference for downstream performance evaluation.

Performance evaluation

Performance was evaluated separately at the gene copy-number level and at the allele level for each sample and gene. For copy-number evaluation, the predicted and true copy numbers were compared directly. The number of correctly inferred copies was counted as true positives, defined as the minimum of the true and predicted copy number for a given gene. Any excess predicted copies were counted as false positives, and any missing copies were counted as false negatives. An exact copy-number match was recorded when the predicted copy number was identical to the true copy number. At the allele level, after variant calling, allele-level performance was assessed using the CDS sequence of allele calls from while preserving allele multiplicity. For each gene, correctly predicted CDS copies were counted as true positives, missing true CDS copies as false negatives, and additional predicted CDS copies as false positives. This ensured that partial recovery of a duplicated CDS was scored proportionally; for example, if a CDS was present at copy number two in the truth set but

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

predicted at copy number one, this contributed one true positive and one false negative. Exact allele match was defined as complete agreement in both CDS identity and copy number.

For both copy-number and allele-level evaluations, precision was calculated as TP / (TP + FP), recall as TP / (TP + FN), and the Jaccard index as TP / (TP + FP + FN). All metrics were reported as percentages. Exact-match performance was calculated as the proportion of evaluated sample–gene combinations for which the prediction matched the truth exactly.

Benchmark methods

T1K v1.0.9-r239 was benchmarked using the IPD-KIR 2.15 database. To harmonize allele resolution with the evaluation framework, T1K was run with --alleleDigitUnits 5, producing 5-digit genotype calls in genotype.tsv. These calls were subsequently refined with t1k_copynumer.py script, and the resulting copy-number-corrected 5-digit alleles were used for evaluation before variant correction. For evaluation after variant correction, the T1K allele.tsv file and VCF output were used to reconstruct 7-digit representative alleles. Briefly, the corrected 5-digit calls were mapped to the corresponding 7-digit representative alleles reported by T1K, after which sequence variants from the VCF were applied. For alleles with copy number greater than one, variants were assigned to ceil(Freq × copy number) copies. When multiple variants were present, this assignment was performed iteratively for each variant, with copies selected arbitrarily because haplotype-specific phase information was not available. GraphKIR was run using the supplied index (IPD-KIR 2.10). Geny was also run with the available database (IPD-KIR 2.12). In both cases, the resulting allele and copy-number predictions were used for downstream benchmarking after converting reported allele names to their corresponding IPD-KIR 2.15 nomenclature using Allelelist_history.txt from IPD-KIR.

Reference

1. Benson DM, Caligiuri MA. 2014. Killer Immunoglobulin-like Receptors and Tumor Immunity. *Cancer Immunology Research* **2**: 99–104.
2. Béziat V, Hilton HG, Norman PJ, Traherne JA. 2017. Deciphering the killer-cell immunoglobulin-like receptor system at super-resolution for natural killer and T-cell biology. *Immunology* **150**: 248–264.
3. Campbell KS, Purdy AK. 2011. Structure/function of human killer cell immunoglobulin-like receptors: lessons from polymorphisms, evolution, crystal structures and mutations. *Immunology* **132**: 315–325.
4. Hollenbach JA, Ladner MB, Saeteurn K, Taylor KD, Mei L, Haritunians T, McGovern DPB, Erlich HA, Rotter JI, Trachtenberg EA. 2009. Susceptibility to Crohn's disease is

:xiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint ich was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

mediated by KIR2DL2/KIR2DL3 heterozygosity and the HLA-C ligand. *Immunogenetics* **61**: 663–671.

5. Hollenbach JA, Nocedal I, Ladner MB, Single RM, Trachtenberg EA. 2012. Killer cell immunoglobulin-like receptor (KIR) gene content variation in the HGDP-CEPH populations. *Immunogenetics* **64**: 719–737.
6. Hsu KC, Chida S, Geraghty DE, Dupont B. 2002. The killer cell immunoglobulin-like receptor (KIR) genomic region: gene-order, haplotypes and allelic polymorphism. *Immunological Reviews* **190**: 40–52.
7. Hung T-K, Liu W-C, Lai S-K, Chuang H-W, Lee Y-C, Lin H-Y, Hsu C-L, Chen C-Y, Yang Y-C, Hsu JS, et al. 2024. Genetic complexity of killer-cell immunoglobulin-like receptor genes in human pangenome assemblies. *Genome Res* **34**: 1211–1223.
8. Khakoo SI, Thio CL, Martin MP, Brooks CR, Gao X, Astemborski J, Cheng J, Goedert JJ, Vlahov D, Hilgartner M, et al. 2004. HLA and NK Cell Inhibitory Receptor Genes in Resolving Hepatitis C Virus Infection. *Science* **305**: 872–874.
9. Kulkarni S, Martin MP, Carrington M. 2008. The Yin and Yang of *HLA* and *KIR* in human disease. *Seminars in Immunology* **20**: 343–352.
10. Lin H-Y, Chuang H-W, Hung T-K, Wang T-J, Lin C-J, Hsu JS, Hsu C-L, Yang Y-C, Chen P-L, Chen C-Y. 2023. Graph-KIR: Graph-based KIR Copy Number Estimation and Allele Calling Using Short-read Sequencing Data. http://biorxiv.org/lookup/doi/10.1101/2023.11.29.568665 (Accessed February 21, 2026).
11. Liu LL, Pfefferle A, Yi Sheng VO, Björklund AT, Béziat V, Goodridge JP, Malmberg K-J. 2015. Harnessing adaptive natural killer cells in cancer immunotherapy. *Molecular Oncology* **9**: 1904–1917.
12. Marin WM, Dandekar R, Augusto DG, Yusufali T, Heyn B, Hofmann J, Lange V, Sauter J, Norman PJ, Hollenbach JA. 2021. High-throughput Interpretation of Killer-cell Immunoglobulin-like Receptor Short-read Sequencing Data with PING ed. F. Ay. *PLoS Comput Biol* **17**: e1008904.
13. Martin MP, Gao X, Lee J-H, Nelson GW, Detels R, Goedert JJ, Buchbinder S, Hoots K, Vlahov D, Trowsdale J, et al. 2002. Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nat Genet* **31**: 429–434.
14. Prodanov T, Plender EG, Seebohm G, Meuth SG, Eichler EE, Marschall T. 2025. Locityper enables targeted genotyping of complex polymorphic genes. *Nat Genet* **57**: 2901–2908.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

15. Purdy AK, Campbell KS. 2009. Natural killer cells and cancer: Regulation by the killer cell Ig-like receptors (KIR). *Cancer Biology & Therapy* **8**: 2209–2218.
16. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. 2019. IPD-IMGT/HLA Database. *Nucleic Acids Research* gkz950.
17. Roe D, Vierra-Green C, Pyo C-W, Geraghty DE, Spellman SR, Maiers M, Kuang R. 2020. A Detailed View of KIR Haplotype Structures and Gene Families as Provided by a New Motif-Based Multiple Sequence Alignment. *Front Immunol* **11**: 585731.
18. Song L, Bai G, Liu XS, Li B, Li H. 2023. Efficient and accurate KIR and HLA genotyping with massively parallel sequencing data. *Genome Res* **33**: 923–931.
19. Trowsdale J, Jones DC, Barrow AD, Traherne JA. 2015. Surveillance of cell and tissue perturbation by receptors in the LRC. *Immunological Reviews* **267**: 117–136.
20. Vilches C, Parham P. 2002. KIR: Diverse, Rapidly Evolving Receptors of Innate and Adaptive Immunity. *Annu Rev Immunol* **20**: 217–251.
21. Zhou Q, Ghezelji M, Hari A, Ford MKB, Holley C, Sahinalp SC, Numanagić I. 2024a. Geny: a genotyping tool for allelic decomposition of killer cell immunoglobulin-like receptor genes. *Front Immunol* **15**: 1494995.
22. Zhou Y, Song L, Li H. 2024b. Full-resolution HLA and KIR gene annotations for human genome assemblies. *Genome Res* **34**: 1931–1941.

Availability of data and materials

KIR*BLOOM source code is available in https://github.com/YomnaGohar/KIR-BLOOM
Construction of KIR*BLOOM allele database is available at
https://github.com/YomnaGohar/Interpolating-Complete-Sequences-for-KIR-Alleles

Authors' contributions

Y.G. and A.D. developed the KIR*BLOOM method and wrote the manuscript. Y.G. implemented the method and performed the analyses. A.D. supervised the study. A.D.G. contributed at an early stage of the project by helping to identify the optimal modeling strategy and define the evaluation metrics. K.M.K. performed the PING benchmarking analyses. P.J.N. contributed through scientific discussion and advice. All authors reviewed and approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.64898/2026.04.15.718735; this version posted April 20, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Funding

Manchot Foundation, Netzwerk Universitätsmedizin (BMBF) Grant No. 01KX2121, BMBF Award 031L0184B

Conflict of interest

Alexander Dilthey is a co-founder of Peptide Groove, LLP, a company that commercializes statistical HLA typing approaches.

3.4 Supplementary material for Publication 2

3.4.1 Supplementary notes

Supplementary

Reference Construction

For accurate KIR genotyping, KIR-related reads must first be extracted, which requires the use of a reference genome. In this study, the reference consisted of two components: (1) a human reference genome in which KIR loci were masked, and (2) a database of curated KIR allele sequences.

For component (2), we downloaded the IPD-KIR database (December 2025 release), which contains both complete and partial allele sequences, including entries represented only by coding DNA sequences (CDS). The first step involved constructing full-length alleles by interpolating missing intronic and untranslated regions (UTRs) for CDS-only entries, as described below.

For component (1), we annotated KIR loci in both the GRCh38 and T2T (CHM13) human genome assemblies using three complementary approaches. First, curated public annotations were employed: RefSeq annotations for GRCh38 and the JHU RefSeq v110 annotations generated with Liftoff v5.2 for T2T. Second, we applied the **Immuannot** tool to both assemblies to obtain KIR-specific feature predictions. Third, all KIR alleles from the IPD-KIR database were mapped to both assemblies to delineate KIR locus boundaries precisely.

All candidate KIR loci were manually inspected to resolve discrepancies between annotation sources, such as overlapping or conflicting gene labels (see Supplementary File X for examples). When inconsistencies arose, Immuannot annotations were prioritized based on their concordance with allele mappings. For each confirmed locus, masking coordinates were defined as the minimum start and maximum end positions across all aligned alleles, thereby ensuring inclusion of variable UTR regions and preventing residual KIR-related sequences from attracting spurious mappings. This process produced masked reference genomes that retained intergenic and non-KIR regions while comprehensively excluding all KIR loci.

Novel KIR alleles identified during annotation were extracted using a custom Python script based on the GTF file generated by Immuannot. These novel alleles, verified through manual inspection and sequence comparison, are provided in **Supplementary Table Novel_alleles.ods**.

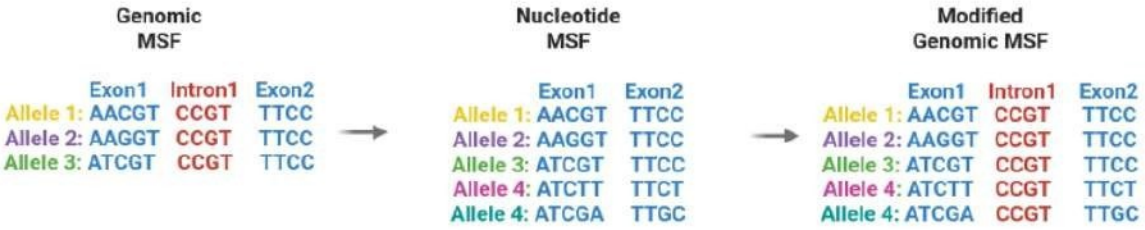
Novel alleles identified during the annotation step were incorporated into the corresponding multiple sequence alignment (MSA) files for each KIR gene. Each MSA contained the full set of alleles from the IPD-KIR database, including those with interpolated intronic and untranslated regions (UTRs). Novel alleles were added to the existing MSAs using **MAFFT** with the `--add` and `--reorder` options. Following integration, the 5′ and 3′ UTRs of all alleles within each gene were visually inspected and standardized to ensure uniform UTR lengths, except in cases where extensive structural heterogeneity or numerous alignment gaps exist.

All KIR allele sequences across all genes, together with the GRCh38 and T2T (CHM13) reference assemblies in which KIR loci were masked, were concatenated into a single FASTA file. This comprehensive reference was used throughout the study for read mapping, allele identification, and genotyping analyses.

1. Modify IPD-KIR

The complete sequences of all KIR alleles are not fully known, as some alleles are represented only by their cDNA sequences, which include exons but lack introns, untranslated regions (UTRs), and, in some cases, even terminal exons are partially or completely missing. To achieve a comprehensive

set of KIR alleles enabling genotyping at a seven-digit resolution, we interpolated the missing sequences from closely related alleles with complete sequences. To accomplish this, we utilized MSF files from IPD-KIR (December 2024). These MSF files contain multiple sequence alignments for alleles of each KIR gene. There are three types of MSF files: protein MSF, genomic MSF, and nucleotide MSF. The genomic MSF contains the multiple sequence alignments of alleles with complete sequences, except for a few cases (e.g., KIR3DL3*005). The nucleotide MSF contains the multiple sequence alignments of the exons of all known alleles of that gene. The aim is to produce a modified genomic MSF that includes both the complete alleles and the exons of the incomplete alleles, with introns and UTRs (if present) interpolated from closely related alleles, as illustrated in Figure 1S. This step is crucial for enabling the alignment of sequencing reads from whole-genome sequencing (WGS) experiments.



This process was achieved through the following steps:

1. **Transforming Annotations to Account for Gaps in the MSA**:
   We transformed the annotations from the XML file, which contains the coordinates of features, into multiple sequence alignment (MSA) annotations to account for gaps introduced during alignment. This transformation ,using the following function, ensures that the annotations accurately reflect the positions within the MSA.

```
function get_msa_position(msa_sequence, original_position):
  original_idx = 0
  msa_idx = 0
  while original_idx < original_position:
    if msa_sequence[msa_idx] is not '-':
      original_idx += 1
    msa_idx += 1
  while msa_idx < length(msa_sequence) and msa_sequence[msa_idx] is '-':
    msa_idx += 1
  return msa_idx
```

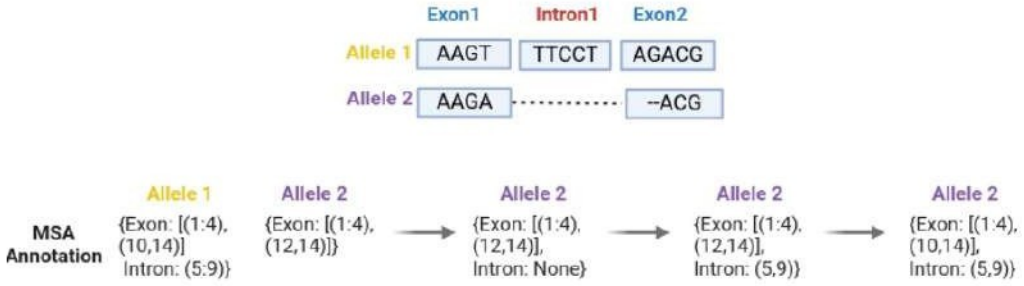
   An example of this process is depicted in Figure 2S.



2. **Ensuring No Overlaps Between Features in the MSA Genomic Files**:
   To ensure there are no overlaps between features such as terminal exons and untranslated

regions (UTRs), we standardized the annotations by harmonizing the feature counts and the coordinates of the features across all alleles. For each allele, we ensured the same number of features by adding `None` as placeholders for any missing features at a given position. The feature annotations were then modified: we adjusted the end coordinates of the features to match the maximum end coordinate observed across all alleles at that index. Subsequently, we adjusted the start coordinates so that each feature began one base after the previous feature's end coordinate. These steps are crucial for two reasons: first, to ensure that gaps at the very start or end of features are included in the ranges, and second, to allow the detection of any overlaps between features.

All genes had no overlaps between their alleles' features in the genomic MSA except for KIR2DL3*010 and KIR3DL1*05901 where they had an overlap between last exon and UTR. This overlap was modified using a custom python script. An example of the standardization process is depicted in Figure S3.



Interpolation of Missing Exons in Genomic MSA Using Closely Related Alleles

Some alleles in the genomic msf file contained missing features; however, it is unclear whether this reflects true biological absence or incomplete annotation in the database. Alleles with any identified missing features were therefore not used to complete missing UTR or intronic regions in the nucleotide sequence files. These alleles are reported in the following table.

Gene	allele	feature/s missing or incomplete
2DL5A	KIR2DL5A*00102	Missing intron(s)
	KIR2DL5A*00104	Missing intron(s)
	KIR2DL5A*00105	Missing intron(s)
	KIR2DL5A*0050102	Missing intron(s)
	KIR2DL5A*01201	Missing intron(s)
	KIR2DL5A*01202	Missing intron(s)
2DL5B	KIR2DL5B*0020102	Missing intron(s)
	KIR2DL5B*0020104	Missing intron(s)
	KIR2DL5B*0020106	Missing intron(s)
	KIR2DL5B*0020201	Missing intron(s)
	KIR2DL5B*003	Missing intron(s)
	KIR2DL5B*00602	Missing intron(s)
	KIR2DL5B*00603	Missing intron(s)

	KIR2DL5B*0070102	Missing intron(s)
	KIR2DL5B*0080102	Missing intron(s)
	KIR2DL5B*00802	Missing intron(s)
	KIR2DL5B*009	Missing intron(s)
	KIR2DL5B*010	Missing intron(s)
	KIR2DL5B*011	Missing intron(s)
	KIR2DL5B*01301	Missing intron(s)
	KIR2DL5B*01302	Missing intron(s)
	KIR2DL5B*01303	Missing intron(s)
2DS4	KIR2DS4*0010103	Missing intron(s)
2DL4	KIR2DL4*0010302	Incomplete exon1 (1 to 28)
	KIR2DL4*00104	Incomplete exon1 (1 to 28)
	KIR2DL4*00901	Incomplete exon1 (1 to 28)
	KIR2DL4*0080202	Incomplete exon1 (1 to 28)
	KIR2DL4*0080103	Incomplete exon1 (1 to 28)
	KIR2DL4*010	Incomplete exon1 (1 to 28)
3DL3	KIR3DL3*0020201	Missing exon1
	KIR3DL3*005	Missing intron(s)

For alleles that have partially or completely missing exons (which occurs at the terminal sequences), the missing sequences were interpolated from closely related complete alleles of the same gene. We define a complete allele as one that has no missing or partial exons or introns. The following pseudocode describes how a closely related allele is identified:

```
function sequence_difference(seq1, seq2):
    difference = 0
    for each character c1, c2 in seq1 and seq2:
        if c1 ≠ c2:
            difference += 1
    return difference


function find_closest_allele(target_allele, nucleotide_msa_dict,
complete_sequence_annotations):
    target_sequence = nucleotide_msa_dict[target_allele]
    min_difference = infinity
    closest_allele = None

    for each allele, sequence in nucleotide_msa_dict:
        if allele ≠ target_allele:
```

```
        diff = sequence_difference(target_sequence, sequence)
        if diff < min_difference and allele is in complete_sequence_annotations:
            min_difference = diff
            closest_allele = allele
    return closest_allele, min_difference
```

To accurately identify missing or partial exons, we utilized the standardization process described in the previous section. During this process, `None` was added as a placeholder for any missing features (whether exons, introns, or UTRs) across all alleles. This standardization allowed us to systematically detect missing exons, as the presence of `None` indicated that a feature was absent in the sequence. For partial sequences, the identification was based on information derived from the XML file, which includes a `status` attribute for each exon. This attribute specifies whether an exon is complete or partial. However, it is important to note that this `status` attribute is only available for exons; UTRs and introns do not have this attribute in the XML file.

To ensure accurate interpolation of a completely missing exon, we needed to determine whether the missing exon was due to a true biological absence or simply a gap in the data. This determination was also based on the XML file, which contains cDNA sequence data. If the cDNA sequence does not start at position 1, or if there is a gap between the end position of one exon and the start position of the next, it suggests that the absence is due to a gap in the data rather than a biological missingness.

Interpolation of Missing Introns and UTRs in Incomplete Alleles Using Closely Related Genomic MSAs

For the alleles in the nucleotide file, which contains the MSA of exons for all alleles—including those in the genomic MSA and those missing introns and UTRs—we interpolated the missing introns and UTRs (if present) from the closest alleles in the modified genomic MSA resulting from the previous section. For each allele, the closest match was found. The interpolation process followed the same steps as those used for interpolating the missing sequences in the genomic MSA.

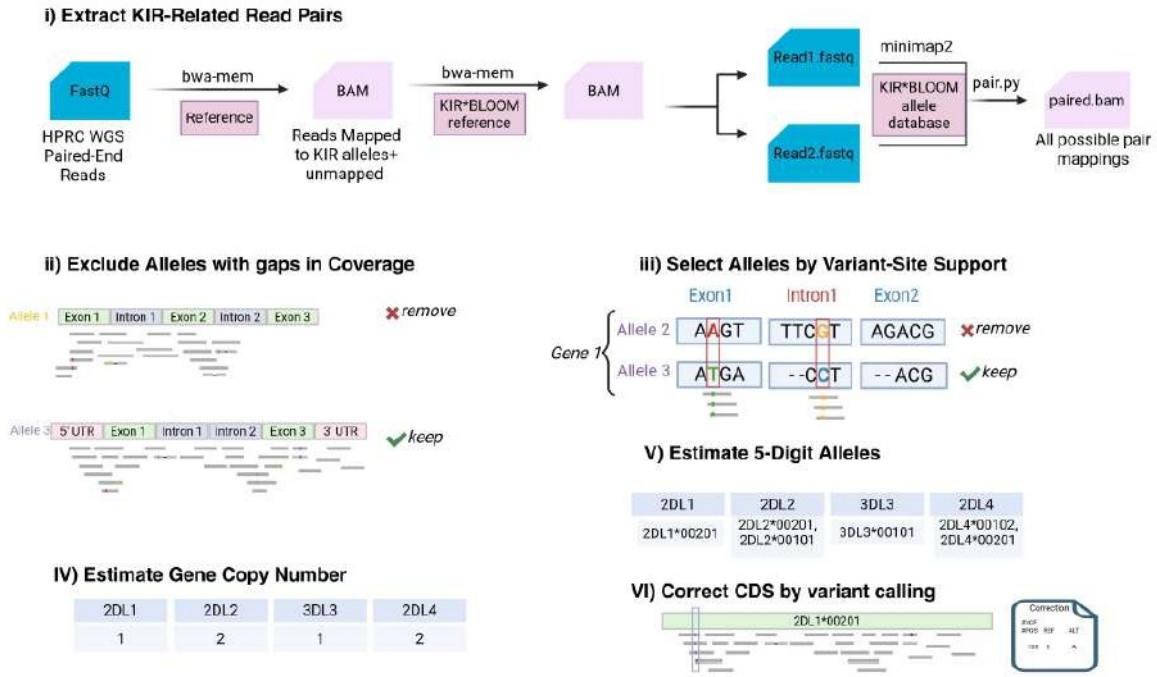
Briefly, for alleles that are incomplete, we standardized the sequence annotations by adding `None` as a placeholder for the missing exons. We then iterated over the features of the complete sequence of the closest related allele, ordered by their starting positions. If the feature was a UTR or an intron, the sequence was derived from the closely related allele. If the feature was an exon, we checked whether it was missing or partial in the allele. If it was missing, we determined whether the missing exon was due to true biological absence or simply missing data. If the absence was due to missing data, we interpolated the exon from the closely related allele. If the exon was partially missing, we also interpolated the missing part from the closely related allele. If no missingness or partial sequence was present, we inserted the exon as it is.

Importantly, the sequence added during interpolation was the multiple aligned sequence. To ensure consistency, we verified that the multiple sequence alignment of the exons in the closest related allele was identical in both the genomic and nucleotide MSF files before incorporating the modified sequence.

3.4.2 Supplementary figures

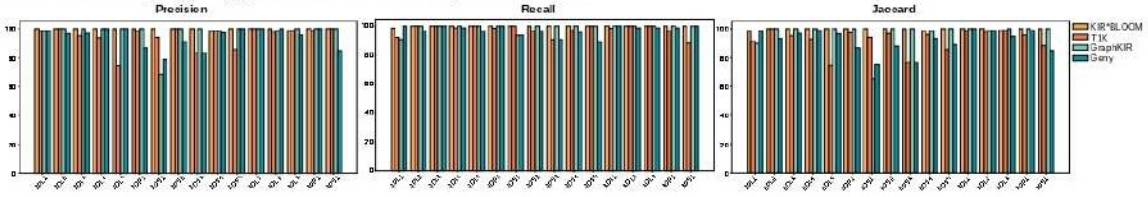
Supplementary Figure 1

Overview of KIR*BLOOM pipeline



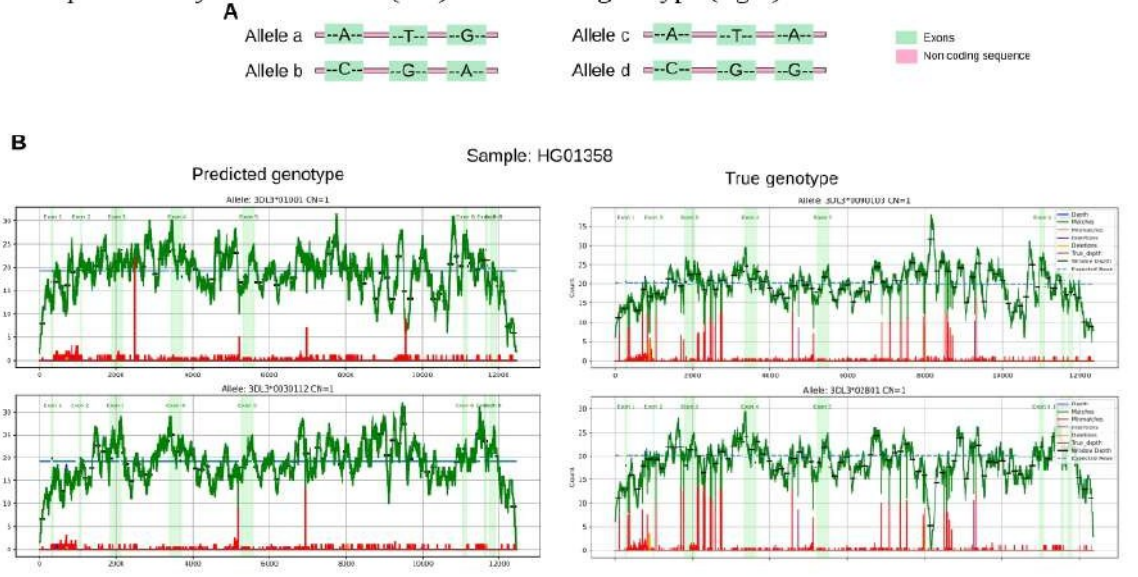
Supplementary Figure 2

Evaluation of gene copy-number inference performance.



Supplementary Figure 3

Schematic representation of an ambiguous genotype (top). Coverage and error profiles for the alleles predicted by KIR*BLOOM (left) and the true genotype (right) for KIR3DL3.



3.4.3 Supplementary tables

Due to their size, the supplementary tables are not included in print and can be found on the enclosed CD

4 Discussion

The work presented in this dissertation addresses a central challenge in modern genomics: how to accurately characterize genomic variation, whether acquired during laboratory culturing or encoded as natural polymorphism, under different technological constraints. Although sequencing has transformed genome analysis, inference remains shaped by read length, error profile, reference representation, and locus complexity. At first glance, an intracellular parasite and a human immune receptor seem unrelated, yet both confront the same limitation: standard pipelines fail when variation lies in repetitive, duplicated, or incompletely represented regions.

This dissertation explored the problem in two complementary systems. In T. gondii, Oxford Nanopore long-read sequencing was used to generate the most contiguous and complete ME49 reference assembly to date, providing a substantially improved genomic foundation for the species. Building on this assembly, variant calling and graph-based genotyping were then applied to compare three independently maintained ME49 isolates, revealing that laboratory strains are not genetically stable and detecting both small variants and a 13 kb expansion in the repetitive ROP8–ROP2A locus that increased copy number from three to six, a structural variant that would have been undetectable using short-read sequencing alone. In the human KIR locus, where most available data exist only as Illumina short reads, current methods are insufficient for several reasons. The locus is notoriously difficult to genotype due to its highly homologous paralogous genes, extensive copy number variation, complex haplotype structure, and the presence of fusion genes, all of which confound short-read alignment and allele assignment. Compounding these challenges, the allele database remains incomplete, limiting the ability of catalog-based methods to recognize novel alleles. KIR*BLOOM was therefore developed to extract gene-content and allele-level genotypes from short reads using a copy-number-aware likelihood framework, achieving near-perfect gene copy number detection and substantially improved allele-level genotyping accuracy that exceeds the performance of currently available short-read KIR genotyping tools such as T1K.

Together, the two studies address opposite sides of the same problem: generating long reads to directly resolve structural variants in T. gondii, versus developing statistical models to recover complex genotypes from imperfect short reads at the KIR locus. In T. gondii, long reads were essential for the ROP8–ROP2A locus, where short reads

cannot distinguish the order or copy number of near-identical paralogs. Equally critical was the improved ME49 assembly, contiguous across nearly all chromosomes, because a fragmented reference obscures variant calling. De novo assembly was therefore not a technical add-on but the foundation for comparative genomics, and joint genotyping across isolates was subsequently performed using a pangenome graph. A remaining challenge is that long reads are not error-free, and homopolymers and tandem repeats still produce systematic artifacts that require stringent filtering and graph-based re-genotyping to distinguish from true isolate-specific variation.

For KIR, the constraints were different. Long reads would be ideal, but discarding the vast existing short-read cohorts that constitute most of the available KIR data is impractical. KIR*BLOOM therefore treats ambiguous mappings as informative rather than forcing deterministic read assignment, infers copy number from depth, and constrains allele choice accordingly. Some limitations remain intrinsic to short-read data: variants separated by distances greater than the fragment length cannot be reliably phased, and incomplete reference entries lead to suboptimal allele calls. KIR*BLOOM therefore also required an improved reference, constructed by completing missing regions in the allele database, incorporating alleles from GRCh38 and T2T, and masking KIR sequences in the genome to prevent spurious mapping. Genotyping was then performed by evaluating alternative allele combinations under explicit copy-number constraints rather than aligning to a single linear haplotype.

The biological implications follow directly from this improved variant resolution. For T. gondii, the detection of structural variation in virulence-associated genes such as those at the ROP8–ROP2A locus challenges the long-standing assumption that long-maintained laboratory strains remain genetically identical to their original isolates. This has direct relevance for experimental reproducibility, and argues that periodic genomic monitoring of widely used reference strains should become standard practice. For KIR, the implication is parallel but operates at a different scale: because accurate genotyping is a prerequisite for immunogenetic association studies in infection, autoimmunity, cancer, pregnancy outcome, and transplantation, improving short-read KIR calls unlocks decades of accumulated cohort data without requiring re-sequencing, potentially enabling discoveries that have been blocked by genotyping limitations rather than statistical power.

Both studies have limitations that point toward future work. In the T. gondii study, the number of isolates was small, and the study did not determine when the detected

variants arose or whether they were driven by drift, selection, or pre-existing heterogeneity within the original stock. Functional validation was also not performed, so the phenotypic consequences of the detected variants remain unresolved. Future work should combine long-read sequencing with transcriptomic, proteomic, and experimental phenotyping approaches to test whether structural and coding variants affect parasite growth, differentiation, virulence, or host interaction. Expanded sampling of additional ME49 isolates and other T. gondii strains would also help determine how common laboratory-acquired structural variation is and whether similar loci are recurrently affected. For KIR, the main limitations arise from incomplete reference databases and intrinsic short-read ambiguity. Some allele combinations cannot be resolved when distinguishing variants are too far apart to be phased by short-read fragments. Novel alleles also remain difficult, especially in genes with copy number greater than one, because reads from novel sequences may be distributed across multiple candidate alleles, obscuring variant calling. Future improvements could include cohort-level statistical phasing, and further expansion of the allele database. As long-read sequencing becomes more widely available, it will be valuable both for direct KIR genotyping and for improving the reference resources that short-read methods depend on. A key next step will be the application of KIR*BLOOM to large existing short-read cohorts, enabling genome-wide association studies (GWAS) that link KIR allele- and copy-number-level variation to disease susceptibility, treatment response, and other clinically relevant phenotypes at a scale that has so far been inaccessible due to the limitations of existing genotyping tools. In conclusion, these studies converge on a single principle: in genomics, detection is not limited by the reads alone. The reference sets the observable space of variation, and improving it is as important as the algorithmic and sequencing choices. Long reads made it possible to identify previously hidden differences among *T. gondii* laboratory isolates only because they were paired with an improved assembly that could represent those differences, and a revised statistical framework allowed accurate KIR typing from existing short-read data only because it was paired with an extended allele database that defined the space of recognizable variation. The reliability of every downstream analysis, including variant calls, genotypes, association tests, and functional inferences, ultimately depends on whether the analytical strategy is matched to both the architecture of the genome and the limitations of the data, and on whether the reference is rich enough to make the underlying biology visible.

5 References

1. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA. 1977;74:560–4. https://doi.org/10.1073/pnas.74.2.560

2. NovaSeq 6000 System | Key specifications and performance parameters [Internet]. [cited 2026 May 27]. https://emea.illumina.com/systems/sequencing-platforms/novaseq/specifications.html. Accessed 27 May 2026

3. Gupta N, Verma VK. Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality. In: Arora PK, editor. Microbial Technology for the Welfare of Society [Internet]. Singapore: Springer Singapore; 2019 [cited 2026 May 27]. p. 313–41. https://doi.org/10.1007/978-981-13-8844-6_15

4. Rhoads A, Au KF. PacBio Sequencing and its Applications. Genomics, Proteomics & Bioinformatics. 2015;13:278–89. https://doi.org/10.1016/j.gpb.2015.08.002

5. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, Proteomics & Bioinformatics. 2016;14:265–79. https://doi.org/10.1016/j.gpb.2016.05.004

6. The UK Biobank Whole-Genome Sequencing Consortium, Manuscript Writing Group, Carss K, Halldorsson BV, Hou L, Liu J, et al. Whole-genome sequencing of 490,640 UK Biobank participants. Nature. 2025;645:692–701. https://doi.org/10.1038/s41586-025-09272-9

7. The All of Us Research Program Genomics Investigators, Manuscript Writing Group, Bick AG, Metcalf GA, Mayo KR, Lichtenstein L, et al. Genomic data in the All of Us Research Program. Nature. 2024;627:340–6. https://doi.org/10.1038/s41586-023-06957-x

8. 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;20:246. https://doi.org/10.1186/s13059-019-1828-7

9. Attias M, Teixeira DE, Benchimol M, Vommaro RC, Crepaldi PH, De Souza W. The life-cycle of Toxoplasma gondii reviewed using animations. Parasites Vectors. 2020;13:588. https://doi.org/10.1186/s13071-020-04445-z

10. Kim K, Weiss LM. Toxoplasma: the next 100 years. Microbes and Infection. 2008;10:978–84. https://doi.org/10.1016/j.micinf.2008.07.015

11. Frenkel JK, Dubey JP, Hoff RL. Loss of Stages after Continuous Passage of *Toxoplasma gondii* and *Besnoitia jellisoni* *. The Journal of Protozoology. 1976;23:421–4. https://doi.org/10.1111/j.1550-7408.1976.tb03799.x

12. Howe DK, Sibley LD. Toxoplasma gondii: Analysis of Different Laboratory Stocks of the RH Strain Reveals Genetic Heterogeneity. Experimental Parasitology. 1994;78:242–5. https://doi.org/10.1006/expr.1994.1024

13. Lorenzi H, Khan A, Behnke MS, Namasivayam S, Swapna LS, Hadjithomas M, et al. Local admixture of amplified and diversified secreted pathogenesis determinants

shapes mosaic Toxoplasma gondii genomes. Nat Commun. 2016;7:10147. https://doi.org/10.1038/ncomms10147

14. Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. Andrés-León E, editor. PLoS ONE. 2021;16:e0257521. https://doi.org/10.1371/journal.pone.0257521

15. Benson DM, Caligiuri MA. Killer Immunoglobulin-like Receptors and Tumor Immunity. Cancer Immunology Research. 2014;2:99–104. https://doi.org/10.1158/2326-6066.CIR-13-0219

16. Béziat V, Hilton HG, Norman PJ, Traherne JA. Deciphering the killer-cell immunoglobulin-like receptor system at super-resolution for natural killer and T-cell biology. Immunology. 2017;150:248–64. https://doi.org/10.1111/imm.12684

17. Prager I, Watzl C. Mechanisms of natural killer cell-mediated cellular cytotoxicity. Journal of Leukocyte Biology. 2019;105:1319–29. https://doi.org/10.1002/JLB.MR0718-269R

18. Fierst JL. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Front Genet [Internet]. 2015 [cited 2026 May 27];6. https://doi.org/10.3389/fgene.2015.00220

19. The Cost of Sequencing a Human Genome [Internet]. [cited 2026 May 27]. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost. Accessed 27 May 2026

20. Quail MA, Swerdlow H, Turner DJ. Improved Protocols for the Illumina Genome Analyzer Sequencing System. CP Human Genetics [Internet]. 2009 [cited 2026 May 27];62. https://doi.org/10.1002/0471142905.hg1802s62

21. NGS Workflow Steps | Illumina sequencing workflow [Internet]. [cited 2026 May 27]. https://emea.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-workflow.html. Accessed 27 May 2026

22. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics and Bioinformatics. 2021;3:lqab019. https://doi.org/10.1093/nargab/lqab019

23. Liu B, Conroy JM, Morrison CD, Odunsi AO, Qin M, Wei L, et al. Structural variation discovery in the cancer genome using next generation sequencing: Computational solutions and perspectives. Oncotarget. 2015;6:5477–89. https://doi.org/10.18632/oncotarget.3491

24. Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, et al. Variant calling and benchmarking in an era of complete human genome sequences. Nat Rev Genet. 2023;24:464–83. https://doi.org/10.1038/s41576-023-00590-0

25. Stapleton JA, Kim J, Hamilton JP, Wu M, Irber LC, Maddamsetti R, et al. Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing. Tang H, editor. PLoS ONE. 2016;11:e0147229. https://doi.org/10.1371/journal.pone.0147229

26. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012;13:36–46. https://doi.org/10.1038/nrg3117

27. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62. https://doi.org/10.1038/s41587-019-0217-9

28. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36:338–45. https://doi.org/10.1038/nbt.4060

29. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21:597–614. https://doi.org/10.1038/s41576-020-0236-x

30. Ni P, Nie F, Zhong Z, Xu J, Huang N, Zhang J, et al. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. Nat Commun. 2023;14:4054. https://doi.org/10.1038/s41467-023-39784-9

31. Gouil Q, Keniry A. Latest techniques to study DNA methylation. Blewitt M, editor. Essays in Biochemistry. 2019;63:639–48. https://doi.org/10.1042/EBC20190027

32. Doshi R, Kinnear E, Chatterjee S, Guha P, Liu Q. Reliable investigation of DNA methylation using Oxford nanopore technologies. Sci Rep. 2025;15:15900. https://doi.org/10.1038/s41598-025-99882-0

33. Moustakli E, Christopoulos P, Potiris A, Zikopoulos A, Mavrogianni D, Karampas G, et al. Long-Read Sequencing and Structural Variant Detection: Unlocking the Hidden Genome in Rare Genetic Disorders. Diagnostics. 2025;15:1803. https://doi.org/10.3390/diagnostics15141803

34. Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data. 2020;7:399. https://doi.org/10.1038/s41597-020-00743-4

35. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5. https://doi.org/10.1038/s41592-020-01056-5

36. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6. https://doi.org/10.1038/s41587-019-0072-8

37. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol. 2023;41:1474–82. https://doi.org/10.1038/s41587-023-01662-6

38. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–53. https://doi.org/10.1126/science.abj6987

39. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31:1119–25. https://doi.org/10.1038/nbt.2727

40. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5. https://doi.org/10.1126/science.aal3327

41. Earl D, Bradnam K, St. John J, Darling A, Lin D, Fass J, et al. Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Res. 2011;21:2224–41. https://doi.org/10.1101/gr.126599.111

42. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2. https://doi.org/10.1093/bioinformatics/btv351

43. Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biol. 2019;20:159. https://doi.org/10.1186/s13059-019-1774-4

44. Human Genome Project Fact Sheet [Internet]. [cited 2026 May 27]. https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project. Accessed 27 May 2026

45. International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Lander ES, Linton LM, Birren B, Nusbaum C, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. https://doi.org/10.1038/35057062

46. Genome Reference Consortium [Internet]. [cited 2026 May 27]. https://www.sanger.ac.uk/data/genome-reference-consortium/. Accessed 27 May 2026

47. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P, et al. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. GigaScience [Internet]. 2017 [cited 2026 May 27];6. https://doi.org/10.1093/gigascience/gix038

48. Frequently Asked Questions - Genome Reference Consortium [Internet]. [cited 2026 May 27]. https://www.ncbi.nlm.nih.gov/grc/help/faq/. Accessed 27 May 2026

49. Kissinger JC. ToxoDB: accessing the Toxoplasma gondii genome. Nucleic Acids Research. 2003;31:234–6. https://doi.org/10.1093/nar/gkg072

50. Ramaprasad A, Mourier T, Naeem R, Malas TB, Moussa E, Panigrahi A, et al. Comprehensive Evaluation of Toxoplasma gondii VEG and Neospora caninum LIV Genomes with Tachyzoite Stage Transcriptome and Proteome Defines Novel Transcript Features. Blader IJ, editor. PLoS ONE. 2015;10:e0124473. https://doi.org/10.1371/journal.pone.0124473

51. Berná L, Marquez P, Cabrera A, Greif G, Francia ME, Robello C. Reevaluation of the *Toxoplasma gondii* and *Neospora caninum* genomes reveals misassembly, karyotype differences, and chromosomal rearrangements. Genome Res. 2021;31:823–33. https://doi.org/10.1101/gr.262832.120

52. Xia J, Venkat A, Bainbridge RE, Reese ML, Le Roch KG, Ay F, et al. Third-generation sequencing revises the molecular karyotype for *Toxoplasma gondii* and identifies emerging copy number variants in sexual recombinants. Genome Res. 2021;31:834–51. https://doi.org/10.1101/gr.262816.120

53. Errors in DNA Replication | Learn Science at Scitable [Internet]. [cited 2026 May 27]. http://www.nature.com/scitable/topicpage/dna-replication-and-causes-of-mutation-409. Accessed 27 May 2026

54. Brown TA. Mutation, Repair and Recombination. Genomes 2nd edition [Internet]. Wiley-Liss; 2002 [cited 2026 May 27]. https://www.ncbi.nlm.nih.gov/books/NBK21114/. Accessed 27 May 2026

55. Health (US) NI of, Study BSC. Understanding Human Genetic Variation. NIH Curriculum Supplement Series [Internet] [Internet]. National Institutes of Health (US); 2007 [cited 2026 May 27]. https://www.ncbi.nlm.nih.gov/books/NBK20363/. Accessed 27 May 2026

56. Trost B, Loureiro LO, Scherer SW. Discovery of genomic variation across a generation. Human Molecular Genetics. 2021;30:R174–86. https://doi.org/10.1093/hmg/ddab209

57. Mohiuddin M, Kooy RF, Pearson CE. De novo mutations, genetic mosaicism and human disease. Front Genet. 2022;13:983668. https://doi.org/10.3389/fgene.2022.983668

58. Li MJ, Yan B, Sham PC, Wang J. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. Briefings in Bioinformatics. 2015;16:393–412. https://doi.org/10.1093/bib/bbu018

59. Vallejos-Vidal E, Reyes-Cerpa S, Rivas-Pardo JA, Maisey K, Yáñez JM, Valenzuela H, et al. Single-Nucleotide Polymorphisms (SNP) Mining and Their Effect on the Tridimensional Protein Structure Prediction in a Set of Immunity-Related Expressed Sequence Tags (EST) in Atlantic Salmon (Salmo salar). Front Genet. 2020;10:1406. https://doi.org/10.3389/fgene.2019.01406

60. Rangadurai A, Szymanski ES, Kimsey I, Shi H, Al-Hashimi HM. Probing conformational transitions towards mutagenic Watson–Crick-like G·T mismatches using off-resonance sugar carbon R1ρ relaxation dispersion. J Biomol NMR. 2020;74:457–71. https://doi.org/10.1007/s10858-020-00337-7

61. Hunter WN, Brown T, Kneale G, Anand NN, Rabinovich D, Kennard O. THE STRUCTURE OF GUANOSINE-THYMIDINE MISMATCHES IN B-DNA AT 2.5 ANGSTROMS RESOLUTION: 113d [Internet]. 1993 [cited 2026 May 27]. https://doi.org/10.2210/pdb113d/pdb

62. Kimsey IJ, Petzold K, Sathyamoorthy B, Stein ZW, Al-Hashimi HM. Visualizing transient Watson–Crick-like mispairs in DNA and RNA duplexes. Nature. 2015;519:315–20. https://doi.org/10.1038/nature14227

63. Calculated consequences [Internet]. [cited 2026 May 27]. https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html? Accessed 27 May 2026

64. Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, et al. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. Genome Res. 2013;23:749–61. https://doi.org/10.1101/gr.148718.112

65. Burssed B, Zamariolli M, Bellucco FT, Melaragno MI. Mechanisms of structural chromosomal rearrangement formation. Mol Cytogenet. 2022;15:23. https://doi.org/10.1186/s13039-022-00600-6

66. Yeeles JTP, Poli J, Marians KJ, Pasero P. Rescuing Stalled or Damaged Replication Forks. Cold Spring Harbor Perspectives in Biology. 2013;5:a012815–a012815. https://doi.org/10.1101/cshperspect.a012815

67. Hastings PJ, Ira G, Lupski JR. A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. Matic I, editor. PLoS Genet. 2009;5:e1000327. https://doi.org/10.1371/journal.pgen.1000327

68. Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet. 2009;41:849–53. https://doi.org/10.1038/ng.399

69. Sakofsky CJ, Malkova A. Break induced replication in eukaryotes: mechanisms, functions, and consequences. Critical Reviews in Biochemistry and Molecular Biology. 2017;52:395–413. https://doi.org/10.1080/10409238.2017.1314444

70. Lieber MR. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annu Rev Biochem. 2010;79:181–211. https://doi.org/10.1146/annurev.biochem.052308.093131

71. Chang HHY, Pannunzio NR, Adachi N, Lieber MR. Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat Rev Mol Cell Biol. 2017;18:495–506. https://doi.org/10.1038/nrm.2017.48

72. Sfeir A, Symington LS. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? Trends in Biochemical Sciences. 2015;40:701–14. https://doi.org/10.1016/j.tibs.2015.08.006

73. Wang H, Xu X. Microhomology-mediated end joining: new players join the team. Cell Biosci. 2017;7:6. https://doi.org/10.1186/s13578-017-0136-8

74. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17:224–38. https://doi.org/10.1038/nrg.2015.25

75. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. arXiv; 2012 [cited 2026 May 27]. https://doi.org/10.48550/ARXIV.1207.3907

76. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303. https://doi.org/10.1101/gr.107524.110

77. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7. https://doi.org/10.1038/nbt.4235

78. Junjun R, Zhengqian Z, Ying W, Jialiang W, Yongzhuang L. A comprehensive review of deep learning-based variant calling methods. Briefings in Functional Genomics. 2024;23:303–13. https://doi.org/10.1093/bfgp/elae003

79. nanoporetech/medaka [Internet]. Oxford Nanopore Technologies; 2026 [cited 2026 May 27]. https://github.com/nanoporetech/medaka. Accessed 27 May 2026

80. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42:1571–80. https://doi.org/10.1038/s41587-023-02024-y

81. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. 2020;21:35. https://doi.org/10.1186/s13059-020-1941-7

82. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191

83. Introduction_to_how_ONT's_medaka_works [Internet]. [cited 2026 May 27]. https://epi2me.nanoporetech.com/notebooks/Introduction_to_how_ONT%27s_medaka_works.html. Accessed 27 May 2026

84. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. Pangenome Graphs. Annu Rev Genom Hum Genet. 2020;21:139–62. https://doi.org/10.1146/annurev-genom-120219-080406

85. Miga KH, Wang T. The Need for a Human Pangenome Reference Sequence. Annu Rev Genom Hum Genet. 2021;22:81–102. https://doi.org/10.1146/annurev-genom-120120-081921

86. Li Y, Yao J, Sang H, Wang Q, Su L, Zhao X, et al. Pan-genome analysis highlights the role of structural variation in the evolution and environmental adaptation of *Asian honeybees* [Internet]. 2023 [cited 2025 Mar 30]. https://doi.org/10.1101/2023.06.15.545041

87. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. Nat Rev Genet. 2020;21:243–54. https://doi.org/10.1038/s41576-020-0210-7

88. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36:875–9. https://doi.org/10.1038/nbt.4227

89. Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biol. 2020;21:253. https://doi.org/10.1186/s13059-020-02157-2

90. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. Nucleic Acids Research. 2019;gkz950. https://doi.org/10.1093/nar/gkz950

91. IPD-KIR Database [Internet]. [cited 2026 May 28]. https://www.ebi.ac.uk/ipd/kir/about/. Accessed 28 May 2026

92. Dilthey AT, Mentzer AJ, Carapito R, Cutland C, Cereb N, Madhi SA, et al. HLA*LA—HLA typing from linearly projected graph alignments. Berger B, editor. Bioinformatics. 2019;35:4394–6. https://doi.org/10.1093/bioinformatics/btz235

93. Prodanov T, Plender EG, Seebohm G, Meuth SG, Eichler EE, Marschall T. Locityper enables targeted genotyping of complex polymorphic genes. Nat Genet. 2025;57:2901–8. https://doi.org/10.1038/s41588-025-02362-4

94. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science. 2021;374:abg8871. https://doi.org/10.1126/science.abg8871

95. Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nat Genet. 2022;54:518–25. https://doi.org/10.1038/s41588-022-01043-w

96. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51. https://doi.org/10.1038/nrg2986

97. Hill D, Dubey JP. Toxoplasma gondii: transmission, diagnosis and prevention. Clinical Microbiology and Infection. 2002;8:634–40. https://doi.org/10.1046/j.1469-0691.2002.00485.x

98. Robert-Gangneux F, Dardé M-L. Epidemiology of and Diagnostic Strategies for Toxoplasmosis. Clin Microbiol Rev. 2012;25:264–96. https://doi.org/10.1128/CMR.05013-11

99. On a Leishman body infection (or related organisms) of the gondi. International Journal for Parasitology. 2009;39:863–4. https://doi.org/10.1016/j.ijpara.2009.02.001

100. Sibley LD, Khan A, Ajioka JW, Rosenthal BM. Genetic diversity of *Toxoplasma gondii* in animals and humans. Phil Trans R Soc B. 2009;364:2749–61. https://doi.org/10.1098/rstb.2009.0087

101. Su C, Khan A, Asis Khan, Zhou P, Majumdar D, Ajzenberg D, et al. Globally diverse Toxoplasma gondii isolates comprise six major clades originating from a small

number of distinct ancestral lineages. Proceedings of the National Academy of Sciences of the United States of America. 2012;109:5844–9. https://doi.org/10.1073/pnas.1203190109

102. Farrell A, Coleman BI, Benenati B, Brown KM, Blader IJ, Marth GT, et al. Whole genome profiling of spontaneous and chemically induced mutations in Toxoplasma gondii. BMC Genomics. 2014;15:354. https://doi.org/10.1186/1471-2164-15-354

103. Saldarriaga Cartagena AM, Aparicio Arias A, Cristaldi C, Ganuza A, Gonzalez MM, Corvi MM, et al. Toxoplasma gondii RAD51 recombinase is required to overcome DNA replication stress and its inactivation leads to bradyzoite differentiation. DNA Repair. 2025;152:103882. https://doi.org/10.1016/j.dnarep.2025.103882

104. Liao H, Ji F, Helleday T, Ying S. Mechanisms for stalled replication fork stabilization: new targets for synthetic lethality strategies in cancer treatments. EMBO Rep. 2018;19:EMBR201846263. https://doi.org/10.15252/embr.201846263

105. Guirouilh-Barbat J, Lambert S, Bertrand P, Lopez BS. Is homologous recombination really an error-free process? Front Genet [Internet]. 2014 [cited 2026 May 28];5. https://doi.org/10.3389/fgene.2014.00175

106. Syn G, Anderson D, Blackwell JM, Jamieson SE. Toxoplasma gondii Infection Is Associated with Mitochondrial Dysfunction in-Vitro. Front Cell Infect Microbiol. 2017;7:512. https://doi.org/10.3389/fcimb.2017.00512

107. Fox BA, Ristuccia JG, Gigley JP, Bzik DJ. Efficient Gene Replacements in *Toxoplasma gondii* Strains Deficient for Nonhomologous End Joining. Eukaryot Cell. 2009;8:520–9. https://doi.org/10.1128/EC.00357-08

108. Jiang H, Jiang J. Balancing act: the complex role of NK cells in immune regulation. Front Immunol. 2023;14:1275028. https://doi.org/10.3389/fimmu.2023.1275028

109. Ljunggren H-G, Kärre K. In search of the 'missing self': MHC molecules and NK cell recognition. Immunology Today. 1990;11:237–44. https://doi.org/10.1016/0167-5699(90)90097-S

110. Kärre K, Ljunggren HG, Piontek G, Kiessling R. Selective rejection of H–2-deficient lymphoma variants suggests alternative immune defence strategy. Nature. 1986;319:675–8. https://doi.org/10.1038/319675a0

111. Raulet DH, Vance RE. Self-tolerance of natural killer cells. Nat Rev Immunol. 2006;6:520–31. https://doi.org/10.1038/nri1863

112. Alari-Pahissa E, Grandclément C, Jeevan-Raj B, Leclercq G, Veillette A, Held W. Activation by SLAM Family Receptors Contributes to NK Cell Mediated "Missing-Self" Recognition. Sandberg JK, editor. PLoS ONE. 2016;11:e0153236. https://doi.org/10.1371/journal.pone.0153236

113. Blunt MD, Khakoo SI. Activating killer cell immunoglobulin-like receptors: Detection, function and therapeutic use. Int J Immunogenetics. 2020;47:1–12. https://doi.org/10.1111/iji.12461

114. Stewart CA, Laugier-Anfossi F, Vély F, Saulquin X, Riedmuller J, Tisserant A, et al. Recognition of peptide–MHC class I complexes by activating killer immunoglobulin-like receptors. Proc Natl Acad Sci USA. 2005;102:13224–9. https://doi.org/10.1073/pnas.0503594102

115. Cichocki F, Miller JS, Anderson SK. Killer Immunoglobulin-Like Receptor Transcriptional Regulation: A Fascinating Dance of Multiple Promoters. J Innate Immun. 2011;3:242–8. https://doi.org/10.1159/000323929

116. Béziat V, Traherne JA, Liu LL, Jayaraman J, Enqvist M, Larsson S, et al. Influence of KIR gene copy number on natural killer cell education. Blood. 2013;121:4703–7. https://doi.org/10.1182/blood-2012-10-461442

117. Pelak K, Need AC, Fellay J, Shianna KV, Feng S, Urban TJ, et al. Copy Number Variation of KIR Genes Influences HIV-1 Control. Emerman M, editor. PLoS Biol. 2011;9:e1001208. https://doi.org/10.1371/journal.pbio.1001208

118. Martin MP, Qi Y, Gao X, Yamada E, Martin JN, Pereyra F, et al. Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. Nat Genet. 2007;39:733–40. https://doi.org/10.1038/ng2035

119. Hilton HG, Guethlein LA, Goyos A, Nemat-Gorgani N, Bushnell DA, Norman PJ, et al. Polymorphic HLA-C Receptors Balance the Functional Characteristics of *KIR* Haplotypes. The Journal of Immunology. 2015;195:3160–70. https://doi.org/10.4049/jimmunol.1501358

120. Yawata M, Yawata N, Draghi M, Little A-M, Partheniou F, Parham P. Roles for HLA and KIR polymorphisms in natural killer cell repertoire selection and modulation of effector function. The Journal of Experimental Medicine. 2006;203:633–45. https://doi.org/10.1084/jem.20051884

121. Orr MT, Lanier LL. Natural Killer Cell Education and Tolerance. Cell. 2010;142:847–56. https://doi.org/10.1016/j.cell.2010.08.031

122. Anfossi N, André P, Guia S, Falk CS, Roetynck S, Stewart CA, et al. Human NK Cell Education by Inhibitory Receptors for MHC Class I. Immunity. 2006;25:331–42. https://doi.org/10.1016/j.immuni.2006.06.013

123. Xiong S, Sharkey AM, Kennedy PR, Gardner L, Farrell LE, Chazara O, et al. Maternal uterine NK cell–activating receptor KIR2DS1 enhances placentation. J Clin Invest. 2013;123:4264–72. https://doi.org/10.1172/JCI68991

124. Hiby SE, Apps R, Sharkey AM, Farrell LE, Gardner L, Mulder A, et al. Maternal activating KIRs protect against human reproductive failure mediated by fetal HLA-C2. J Clin Invest. 2010;120:4102–10. https://doi.org/10.1172/JCI43998

125. Ahlenstiel G, Martin MP, Gao X, Carrington M, Rehermann B. Distinct KIR/HLA compound genotypes affect the kinetics of human antiviral natural killer cell responses. J Clin Invest. 2008;JCI32400. https://doi.org/10.1172/JCI32400

126. Du J, Lopez-Verges S, Pitcher BN, Johnson J, Jung S-H, Zhou L, et al. CALGB 150905 (Alliance): Rituximab Broadens the Antilymphoma Response by Activating Unlicensed NK Cells. Cancer Immunology Research. 2014;2:878–89. https://doi.org/10.1158/2326-6066.CIR-13-0158

127. Terszowski G, Klein C, Stern M. KIR/HLA Interactions Negatively Affect Rituximab- but Not GA101 (Obinutuzumab)-Induced Antibody-Dependent Cellular Cytotoxicity. The Journal of Immunology. 2014;192:5618–24. https://doi.org/10.4049/jimmunol.1400288