bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license.

A complete and near-perfect rhesus macaque reference genome: lessons from subtelomeric repeats and sequencing bias

Shilong Zhang[1#], Ning Xu[2, 3, 4, 5#], Yong Lu[2, 3], Yanhong Nie[2, 3], Zhengtong Li[1], Luciana de Gennaro[6], Lianting Fu[1], Zhendong Zhang[1], Jieyi Chen[1], Kaiyue Ma[1], Xiangyu Yang[1], Juan Zhang[1], Matthew T. Schmitz[7], Francesca Antonacci[6], Trygve E. Bakken[7], Mario Ventura[6], Adam M. Phillippy[8], Qiang Sun[2, 4, 5*], Yafei Mao[1, 9, 10*]

1. Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China
2. Institute of Neuroscience, Center for Excellence in Brain Science and Intelligence Technology, State Key Laboratory of Neuroscience, Chinese Academy of Sciences, Shanghai, China
3. Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China
4. Key Laboratory of Genetic Evolution & Animal Models, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China
5. University of Chinese Academy of Sciences, Beijing, China
6. Department of Biosciences, Biotechnology and Environment, University of Bari Aldo Moro, Bari, Italy
7. Allen Institute for Brain Science, Seattle, WA 98109, USA
8. Genome Informatics Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
9. Center for Genomic Research, International Institutes of Medicine, Fourth Affiliated Hospital, Zhejiang University, Yiwu, Zhejiang, China
10. Shanghai Key Laboratory of Embryo Original Diseases, International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

#These authors contributed equally to this work.

*Corresponding authors: Qiang Sun (qsun@ion.ac.cn) & Yafei Mao (yafmao@sjtu.edu.cn)

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license.

ABSTRACT

A truly complete, telomere-to-telomere (T2T), and error-free reference genome remains a foundational resource—and long-standing goal—for unbiased comparative and functional genomics. While recent T2T assemblies of humans and other primates have made substantial progress, most still contain thousands of base-level errors, particularly within highly repetitive regions. Here, we present T2T-MMU8v2.0, a near-perfect T2T assembly of the rhesus macaque (*Macaca mulatta*), representing the highest base-level accuracy reported in a primate genome to date. By employing an optimized ONT-only assembly strategy, we identify subtelomeric satellite-rich regions as the principal bottleneck to improving assembly quality, owing to technological biases in long-read platforms and limitations in current hybrid assembly frameworks. We discover 268 previously unannotated repeat families and resolve ~8 Mbp of SATR satellite arrays, with over 99-fold enrichment in historically misassembled subtelomeric regions. These satellites form four distinct genomic architectures, each with unique SATR satellite composition, segmental duplication organization, and epigenetic signatures, distinct from the subtelomeric architectures observed in hominid genomes. Notably, in contrast to the largely gene-poor subtelomeric regions in African hominids, the SATR architectures in macaques harbor 58 actively transcribed genes, supported by open chromatin and expression data, suggesting gene innovation within these repetitive regions. Functionally, T2T-MMU8v2.0 improves read mappability and accuracy across sequencing platforms, and results in a 19% improvement of transcription start site enrichment scores and 5,821 additional chromatin accessibility peaks on average, thereby enhancing variant detection, regulatory annotation, and transcriptomic resolution in population genetics or single-nucleus studies. Together, this work establishes a new benchmark for genomics, offers a roadmap for resolving complex repetitive regions, and reveals previously unrecognized features of subtelomeric genome structure and evolution.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

INTRODUCTION

An error-free, complete genome assembly represents the ultimate reference for unbiased comparative genomics and functional genetic analysis[1-4]. The advent of long-read sequencing technologies, combined with continual improvements in genome assembly algorithms, has enabled the generation of complete, telomere-to-telomere (T2T) assemblies[3,5-13]. Landmark accomplishments, such as the human T2T genome (T2T-CHM13) published in 2022, achieved exceptional accuracy with an estimated Phred quality value (QV) > 70, corresponding to fewer than one base error per million base pairs[6]. Similarly, the T2T macaque genome (T2T-MFA8) reached a comparable level of completeness and quality[8]. Despite these advances, current assemblies remain haploid representations and thus fail to capture the full diploid complexity of natural genomes.

The field is now advancing toward the more ambitious goal of fully phased, diploid, T2T assemblies in humans and other primates—a frontier actively being explored between 2023 and 2025[1,7,9,13]. Yet, current diploid T2T assemblies remain imperfect, often containing unplaced contigs and exhibiting lower consensus accuracy (mean QV ≈ 65) than the previous homozygous genomes (e.g., T2T-CHM13)[7,9]. At the heart of this challenge lies the difficulty of resolving extensive, highly identical repetitive regions—an issue that persists even in homozygous genomes[6,8,10]. The introduction of haplotype phasing further compounds this complexity, stretching the limits of current sequencing technologies and computational approaches[3]. Thus, recent diploid reference assemblies still contain numerous "error *k*-mers" (short sequences for which there is no underlying sequencing read support) or unexplained read mapping anomalies that suggest structural problems with the assembly.

We propose that reconstructing a complete and near-perfect haploid genome (estimated QV ≈ 100, Genome Continuity Inspector (GCI) = 100, and no unassembled/collapsed regions and unplaced contigs) represents a critical and necessary step toward the routine generation of near-perfect, complete diploid genomes. Such an assembly will resolve all previously unassembled or low-quality regions with high confidence, provide a controlled framework for identifying technical and algorithmic bottlenecks, and offer a benchmark for evaluating future assembly method performance. Insights gained from a complete and near-perfect homozygous genome will directly inform the development of future assembly tools, paving

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

the way for accurate and automated construction of diploid and even polyploid genomes at a single-laboratory scale.

Macaques, as one of the closest genetic relatives to humans, are important nonhuman primate (NHP) models in biomedical research, underpinning progress in genomics, neuroscience, and developmental biology[14-23]. However, the current macaque reference genome (Mmul_10 or T2T-MFA8) still harbors unresolved errors[8,19]. In this study, we present a complete and near-perfect T2T haploid genome assembly for a rhesus macaque with a GCI = 100 and zero error *k*-mers remaining in the consensus sequence ($k = 21$). This assembly enabled us to address four key questions: (1) Which genomic regions were previously misassembled or missing? (2) What are the remaining technological or algorithmic barriers to achieving complete and near-perfect assemblies? (3) What biological insights can be revealed through the accurate resolution of intractable regions? (4) Does a complete and near-perfect reference improve downstream applications—such as population genetics and single-nucleus omics—compared to existing the macaque reference?

RESULTS

Genome assembly of a rhesus macaque

We first established a pseudo-haploid rhesus macaque (*Macaca mulatta*, MMU2019108-1) cell line that is nearly entirely homozygous diploid, as previously described[8,24] (Supplementary Figures 1 and 2). We then generated 178 Gbp of PacBio high-fidelity (HiFi) reads, 125 Gbp of Oxford Nanopore Technology (ONT) ultra-long reads (≥ 50 kbp; N50 = 109 kbp), and 359 Gbp of Illumina short reads from this cell line (Supplementary Figure 3 and Supplementary Table 1). These datasets were used to assemble the genome with Hifiasm[12]. The resulting draft assembly spans approximately 3.09 Gbp. It exhibits high contiguity and accuracy, with a contig N50 of 146.63 Mbp, a Merqury QV[25] estimate of 54.23, a GCI[26] of 49.19, and a total of 55 contigs (Supplementary Tables 2 and 3).

Building on our previous T2T assembly strategy[8], we closed 19 remaining gaps in the initial draft and reconstructed a complete rhesus macaque genome (T2T-MMU8v1.0), achieving a N50 of 162.29 Mbp and a QV of 80.11 (Supplementary Tables 4–8). These improvements are predominantly concentrated in segmental duplications (SDs) and simple satellite regions, as expected. However, read-depth–based and *k*-mer–based evaluations revealed that 9 regions and 613 error *k*-mers ($k = 21$) in T2T-MMU8v1.0, respectively, still contain misassembled

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

sequences (Supplementary Tables 6 and 8). These error *k*-mers are predominantly enriched in subtelomeric regions (empirical $P = 0.039$, permutation test), highlighting persistent challenges in assembling subtelomeric highly repetitive sequences.

Complete and near-perfect genome assembly and sequencing platform bias

To achieve a complete and near-perfect T2T genome assembly, it is critical to understand why certain genomic regions remain refractory to current assembly strategies (Figure 1A). We re-examined the read depth, mapping quality, and sequence characteristics of these misassembled regions. Based on these characteristics, we reassembled these regions with a custom approach and identified four major sources of bias for sequencing errors.

We first identified nine regions with abnormal read depth, indicative of collapses in the T2T-MMU8v1.0 assembly (Supplementary Table 8). Based on sequence content and sequence similarity relationship among the unresolved regions, we classified these into two major types: (1) four regions harboring locally high-identity tandem repeat (TR) sequences (e.g., SATR satellites on chromosome (chr) 2 and chr6), and (2) five regions with high-identity TRs across different chromosomes (e.g., SATR satellites on chr12, chr13, and chr15). These findings suggest that all unresolved regions prior to manual finishing corresponded to complex satellite TR loci.

To resolve the first class of collapses (local TRs), we constructed a minimizer-based sparse *de Bruijn* graph from corrected ONT reads[27] using MBG[28] as implemented in Verkko[13], and mapped ONT reads to the local graph to trace full-length paths traversed by individual reads (Figure 1B). We then constructed a transition matrix for ONT read paths and applied a scoring function based on the Markov chain to infer the most likely ONT read path through the graph, yielding an optimized contig (Figure 1B). Using this ONT-only custom strategy, we successfully resolved four collapsed regions on chr2 and chr6 (total ~3.5 Mbp). For the second class of collapsed regions (inter-chromosomal high-identity TRs), we first partitioned ONT reads by chromosome of origin and then combined our custom ONT-only graph-based method with Hifiasm[29] (ONT), enabling us to resolve the collapsed subtelomeres (total ~15 Mbp) on chromosomes 12, 13, and 15 (Methods).

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

Following the resolution of large-scale collapses, as confirmed by uniform read depth, we focused on refining base-level errors. During polishing, we identified four distinct types of sequencing bias that underlie the remaining regions with error *k*-mers (Figure 1C).

Depth dropout is a pronounced issue in PacBio HiFi[6,13,30], with 268 regions (~16.68 Mbp) detected in total. By analyzing the current assembly without large-scale collapses, we found that dropout is not primarily attributable to read misalignment or misassignment, but rather reflects a true absence of HiFi reads generated by the PacBio platform (Supplementary Figure 4). This limitation likely stems from the platform's consensus-based read generation. The affected regions are typically tandemly arranged GC-rich repeats (116 regions spanning ~8.5 Mbp) suggesting that current chemistry and consensus algorithms struggle to meet quality criteria in such sequence contexts in the PacBio platform.

Within these dropout regions, we observe additional sequencing biases that hinder base-level accuracy. The most prominent of these are well-characterized homopolymer errors—inaccurate base calling in regions of consecutive identical nucleotides[5,30-33] (Supplementary Figure 5). We also find strand bias, where base calling differs between plus- and minus-strand[29], and sequence context bias, involving systematic miscalls associated with specific sequence motifs[31] (Supplementary Figures 5 and 6). These biases may act independently or synergistically, compounding the challenge of polishing assemblies to high base-pair accuracy.

The near-homozygous nature of the cell line enables more confident resolution of these problematic regions. To address them, we employed a local reference-anchored *k*-mer graph and short-read recruitment strategy (Figure 1C; Methods). Notably, this approach revealed 190 TR loci where ONT reads consistently exhibit a GGG deletion, while Illumina-derived *k*-mers align without discrepancy (Figure 1D and Supplementary Figure 6), indicating a likely systematic ONT base-calling error in these sequence contexts. These platform-specific biases (especially within complex repetitive regions) have been underappreciated in prior studies and remain a major obstacle to achieving complete and near-perfect genome assemblies.

Following the above efforts, we generated a final genome assembly (T2T-MMU8v2.0) characterized by uniform read depth, a GCI of 100 (totally continuous alignment even with

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

strict filtering), and a Merqury *k*-mer–based QV estimate of 100 (QV100, $k = 21$; QV77, $k = 31$) (Table 1, Supplementary Figures 7 and 8, and Supplementary Tables 9–17; Methods).

We resolved ~302 Mbp and ~25 Mbp of novel sequence absent from the Mmul_10 reference and the previously published T2T-MFA8v1.1 assembly, respectively. These regions, largely localized to subtelomeric domains (empirical $P = 0$, permutation test), are highly repetitive—comprising ~94% SDs and ~81% TRs with associated palindromic structures (Figures 1E–1G). Strikingly, 57% of the newly assembled sequences represent previously uncharacterized repeats[6,9,34-37], underscoring that large portions of the subtelomeric regions remained systematically inaccessible in prior references, including complete T2T assemblies (Figure 1G). However, as expected, a genomic region composed of near-identical sequences, most notably ribosomal DNA (rDNA) arrays, remains refractory to current assembly strategies (Supplementary Figure 9). As a result, the rDNA region remains unresolved in T2T-MMU8v2.0, consistent with all previously reported T2T genomes[6,8,9].

Structure and epigenetic pattern of subtelomeric regions in macaques

Building on the above efforts, we found that extensive satellite TRs with palindromic configurations remain a major barrier to achieving a complete and near-perfect genome assembly. A comprehensive understanding of the genomic distribution, sequence composition, and structural organization of these satellite repeats in T2T-MMU8v2.0 may help inform generalized strategies for constructing complete and near-perfect genomes across species in the future (Supplementary Figures 10–13 and Supplementary Table 18).

Focusing on subtelomeric regions at p-arms of T2T-MMU8v2.0 (previously prone to misassembly), we observed that 18% of annotated repeats are dominated by the SATR satellite family, with the remaining 62% ($n = 93$) corresponding to previously uncharacterized repeats (total 268 in the T2T-MMU8v2.0) absent from both current primate T2T references and the Dfam database[6,9,34-37] (Supplementary Figure 12 and Supplementary Table 18).

We identified three major SATR families, including SATR1 (3.44 Mbp), SATR1v[34] (4.48 Mbp), and SATR2 (0.22 Mbp), which are dispersed across the entire genome but exhibit ~99-fold enrichment in subtelomeric regions at p-arms (empirical $P = 0$, permutation test) (Figures 2A–2C and Supplementary Table 19). Specifically, 63.44% of SATR satellite arrays

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

(5.16 Mbp) are localized to 17 subtelomeric loci, while 27.2% (2.21 Mbp) are distributed across 11 interstitial loci (Figures 2B and 2C). Notably, 99.9% of SATR1v arrays (n = 2,066) participate in complex mixed arrays with SATR1 or SATR2, with only two existing as standalone SATR1v arrays (Figure 2B). Similarly, around 90% of SATR1 arrays (n = 2,178) are found within mixed arrays. With respect to available T2T ape and human genomes, the macaque genome harbors ~3.5-fold more SATR1 and ~36-fold more SATR1v satellites, but ~42% fewer SATR2 satellites[8,9] (Figure 2A).

Based on the satellite composition and sequence organization of the SATR flanking regions, we identified four distinct structural architectures in T2T-MMU8v2.0 (Figure 2D, Supplementary Figures 14–22, and Supplementary Table 20). Each of these architectures contains at least one SATR core, composed of SATR arrays. Among these, three involve SATR1–SATR1v mixed arrays (n = 19), while the fourth consists of SATR1–SATR2 mixed arrays (n = 7) (Figures 2B and 2D and Supplementary Figure 16). Additionally, two loci on chr11 feature the combination of SATR1–SATR1v and SATR1–SATR2 mixed arrays. Notably, 98% and 81% of subtelomeric, and 90% and 31% of interstitial SATR architectures remained unresolved in Mmul_10 and T2T-MFA8v1.1, respectively.

Among the 19 SATR1–SATR1v loci, we further classified three structural subtypes (Figure 2D and Supplementary Table 20). At 17 subtelomeric loci, 12 exhibit a configuration with a solo core, while five display expanded multi-core structures (Figure 2D). In both cases, each locus is consistently flanked by two composite repeat elements (Figures 2D and 2E). Notably, each SATR1–SATR1v mixed array at subtelomeric loci is embedded within a stereotyped structure comprising a central SATR1–SATR1v core flanked by two spacer sequences (Figure 2D). In contrast, the other two interstitial loci harbor only the SATR1–SATR1v core and lack both spacers and composite repeats (Figure 2D and Supplementary Figure 16).

To further characterize these structures, we analyzed the sequence composition of the SATR1–SATR1v mixed arrays and their associated composite elements. Each core unit at an SATR1–SATR1v mixed array consists of one variable-length SATR1v array (GC-rich), a conserved 589-bp inverted SATR1 array, an uncharacterized repeat (rnd6; AT-rich), a second conserved 191-bp forward SATR1 array, and another uncharacterized repeat (rnd155) (Figures 2E and 2F and Supplementary Figure 23). The left-flanking composite element (comp1) at a SATR1–SATR1v mixed array spans approximately 120–200 kbp and comprises

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

two SATR1v arrays, three SATR1 arrays, and other repeats (Figure 2F). The right-flanking composite element (comp2; 80–150 kbp) at a SATR1–SATR1v mixed array typically consists of a SATR1 array and other repeats (Figure 2F).

In contrast, the two SATR1–SATR2 mixed arrays are confined to interstitial regions and are absent from subtelomeric loci (Figure 2D). Unlike SATR1–SATR1v mixed arrays, the SATR1–SATR2 mixed arrays lack flanking composite repeats but are bordered by two spacers (Figure 2E). The motif of SATR1–SATR2 mixed cores is variable, comprising SATR1 arrays of varying lengths (ranging from 24 bp to ~35 kbp) interspersed with uniformly sized SATR2 satellites (~377 bp each) (Figures 2E and 2F and Supplementary Figure 23). The SATR1–SATR2 core unit is markedly more heterogeneous and less expanded than that of SATR1–SATR1v mixed arrays.

In addition to analyzing sequence content, structural organization, and array architecture, we next examined the GC content and DNA methylation profiles of the core unit in SATR mixed arrays and their spacer sequences. Core units in both SATR1–SATR1v and SATR1–SATR2 mixed arrays exhibit global hypermethylation (mean methylation level: 82% for SATR1–SATR1v core unit and 74% for SATR1–SATR2 core unit) (Figure 2G); however, they differ markedly in GC content (Figure 2G). Within the core units of the SATR1–SATR1v mixed arrays, SATR1v elements show markedly elevated GC content (mean GC = 81%), whereas SATR1 elements show a lower GC content close to the genome-wide average (mean GC = 44%) (Figure 2G). In contrast, the core units of the SATR1–SATR2 mixed arrays exhibit uniformly GC content (mean = 52%, s.d. = 0.05) across the entire array, comparable to that of SATR1–SATR1v mixed arrays (mean = 60%, s.d. = 0.23) (Figure 2G). These findings suggest that SATR1v is the primary contributor to the high GC content observed in SATR1–SATR1v mixed arrays.

Spacer sequences in subtelomeric regions (~110–200 kbp) are highly conserved on the sequence level, with an average pairwise identity of 97.88% (ranging from 95.46% to 99.98%), whereas those in interstitial regions (~70–230 kbp) are more divergent, averaging 91.76% identity (ranging from 88.94% to 99.73%) (Figures 2H and 2I, Supplementary Figures 24–26, and Supplementary Table 21). Notably, these spacers consistently form flanking inverted SDs at the boundaries of SATR1–SATR1v and SATR1–SATR2 mixed arrays (Figures 2H and 2I), except for interstitial spacers on chr2 (Supplementary Figures 21

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

and 24). This finding explains a genetic basis for the enrichment of SDs in subtelomeric regions of the macaque genome—a pattern that contrasts with humans and other great apes, in which SDs are predominantly concentrated in pericentromeric regions[8,9] (Supplementary Figures 27 and 28). The sequence identity between subtelomeric SD spacers is ~6% higher than that of interstitial SD spacers ($P = 1.57\times10^{-17}$, two-sided Mann–Whitney U test), which might indicate either a recent expansion or recombination between subtelomeric regions (Figure 2I). In addition, synteny analysis revealed that the SD spacers are structurally divergent, harboring large SVs within them (Supplementary Figure 29).

The structural, organizational, genetic, and epigenetic architecture of macaque subtelomeric regions is markedly distinct from the pCht arrays (satellite TRs) described in nonhuman African great apes[9,38] (gorillas and the *Pan* lineage). The pCht arrays consist of uniform 34-bp satellite monomers and fixed-length SD spacers that display hypomethylation relative to surrounding satellite DNA[9,38]. Comparing to pCht arrays, we did not observe substantially reduced DNA methylation in macaque subtelomeric SD spacers (SATR spacer mean = 76%, pCht spacer mean = 43%) relative to the adjacent satellite arrays (SATR mean = 86%, pCht mean = 82%) (Supplementary Figure 30). These differences highlight the evolutionary divergence of subtelomeric satellite architecture and suggest lineage-specific chromatin regulation across primates.

In addition to sequence-based analyses, we performed fluorescence *in situ* hybridization (FISH) to validate the composition and chromosomal organization of SATR1 arrays. Among the 21 macaque chromosomes examined, four subtelomeric regions lacked SATR1 arrays entirely, as confirmed by FISH, while 16 exhibited clear SATR1 signals (Figure 2J). Of these, 11 harbored non-expanded arrays, whereas five showed substantial expansions. The strongest FISH signals were observed at the subtelomeres of chromosomes 12, 13, and 15, corresponding to the most highly expanded SATR1 arrays (Figure 2J). Notably, we did not observe large blocks of heterochromatin at subtelomeric regions in macaques, in contrast to the prominent subtelomeric heterochromatin observed in gorillas and the *Pan* lineage. We speculate that differences in subtelomeric structure and epigenetic regulation may underlie these cytogenetic distinctions. However, the molecular mechanisms driving the emergence, expansion, and turnover of subtelomeric regions and their biological functions in primates remain poorly understood.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

Overall, the complete T2T-MMU8v2.0 assembly enables the first genome-wide fine-scale resolution of subtelomeric satellite architectures—regions that are frequently misassembled by current tools or collapsed in current macaque genomes. This high-resolution view not only offers a valuable benchmark for future algorithm development but also advances our understanding of the structural complexity in NHP genomes.

Novel genes in the SATR satellite-associated regions

Using publicly available long-read and short-read RNA-seq data[8,39], we annotated 21,925 protein-coding genes and 185,700 transcripts (31,423 novel) in the T2T-MMU8v2.0 assembly (Supplementary Table 22). Comparative synteny analysis with the current macaque reference genome, we identified 269 genes and 210 genes present in T2T-MMU8v2.0 but absent from Mmul_10 and T2T-MFA8v1.1, respectively (Supplementary Table 23), likely reflecting unresolved regions or incomplete annotation in the previous assemblies (Figure 3A). With respect to Mmul_10, these novel genes are ~10-fold enriched on chrY (empirical $P = 0$, permutation test) and ~48-fold enriched within SATR architectures (empirical $P = 0$, permutation test).

T2T-MMU8v2.0 enabled complete resolution of several previously collapsed, highly duplicated gene families corresponding to the SATR architectures. We totally identified 6 genes and 225 additional copies in these regions, including *ZNF669* (*ZNF669L*, 41 copies, 10 supported by Iso-Seq), *SH3TC1* (*SH3TC1L*, 76 copies, 35 supported), *PFKP* (*PFKPL*, 9 copies, 2 supported), *PITRM1* (*PITRM1L*, 18 copies, 11 supported), *USP17* (*USP17L*, 68 copies, none supported), as well as the primate-specific *FAM90A5*[40,41] (13 copies, none supported) (Figure 3A and Supplementary Table 23). Among primates, *ZNF669* has undergone a lineage-specific duplication in macaques that yielded 41 truncated *ZNF669L* copies in T2T-MMU8v2.0[9,19,42] (Figure 3B). Notably, each subtelomeric SATR architecture harbors at least one *ZNF669L* copy within the ~10–20 kbp boundary (17 copies in total), yet only 4 exhibit both transcriptional support from full-length Iso-Seq data and epigenetic marks of active transcription (Figure 3C and Supplementary Figure 31).

In addition, we identified a previously unannotated duplicated gene, *SH3TC1*, located within SD spacers of all interstitial SATR1–SATR2 mixed arrays[43] (Figure 3A). Similar to *ZNF669L*, the 76 newly annotated *SH3TC1L* duplicates are truncated and exhibit elevated copy number specifically in macaques relative to other primates (Figures 3D and 3E and

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

Supplementary Table 24). All *SH3TC1L* copies (n = 76) harbor canonical CpG islands at their predicted transcription start sites (TSSs) (Figure 3F); however, only 46% (n = 35) exhibit hypomethylation at the TSSs, alongside support from full-length Iso-Seq transcripts, suggesting potential transcriptional activity and regulatory competence (Figure 3G and Supplementary Figure 32).

Together, these findings demonstrate that the T2T-MMU8v2.0 assembly not only recovers previously missing genes but also uncovers extensive, transcriptionally active duplicated copies, particularly within subtelomeric and interstitial SATR architectures. Remarkably, many of these duplicated copies (e.g., *ZNF669L*, *SH3TC1L*, and *PITRM1L*) show transcriptional activity across all 15 tissues examined, including the prefrontal cortex, spleen, and ovary (Methods), implying their function to macaque-specific traits.

Performance improvements on read alignment and single-nucleus omics analysis

Given the enhanced resolution of previously unassembled genomic regions and gene annotation, we next evaluated the utility of the T2T-MMU8v2.0 assembly for downstream applications in population genomics and single-nucleus multi-omics.

To assess improvements in read mappability, we aligned Illumina, HiFi, ONT, and Iso-Seq full-length non-chimeric (FLNC) reads from multiple rhesus macaque individuals[8] to both the T2T-MMU8v2.0 assembly and the current reference genome, Mmul_10. T2T-MMU8v2.0 consistently demonstrated superior mapping performance, with reduced proportions of unmapped reads, lower mismatch rates, and reduced non-primary alignments across all sequencing platforms (Figures 4A–4D). These improvements underscore the enhanced accuracy of T2T-MMU8v2.0 and its suitability as a reference for population-scale variant discovery and transcriptome analyses.

To assess the impact of T2T-MMU8v2.0 on functional genomic analyses, we reanalyzed 1.04 Tbp of brain single-nucleus RNA-seq (snRNA-seq) and ATAC-seq (snATAC-seq) data[44] (46,204 cells in total) using T2T-MMU8v2.0 and Mmul_10 as references, respectively. While snRNA-seq showed only modest gains in read mappability to the transcriptome (a 0.2% improvement), alignment to T2T-MMU8v2.0 resulted in a 0.35% increase in snATAC high-quality fragments and a 19% improvement in TSS enrichment scores, recovering an average of 5,821 additional chromatin accessibility peaks (Figure 4E). Although these

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

differences had minimal impact on broad cell-type classification (Figure 4F), they significantly affected the detection of 20 protein-coding genes that are expressed in neurons or oligodendrocytes (Supplementary Figure 33 and Supplementary Table 25). All of these genes were misassembled in Mmul_10, resulting in their misrepresentation in expression analyses (Supplementary Figure 34). Additionally, 11,606 more chromatin accessibility peaks were identified using T2T-MMU8v2.0, highlighting the improved resolution for functional genomics analyses.

For example, the *ATAD3* locus, which is collapsed in Mmul_10, is fully resolved in T2T-MMU8v2.0, unveiling intact flanking regions where two genes—*VWA1* and *TMEM278*—are correctly assembled but entirely missing in Mmul_10 (Figures 4G–4I). Notably, *VWA1*, which encodes a von Willebrand factor A domain–containing protein with key roles in brain development[45-47], is broadly expressed across multiple brain cell types yet was undetectable in analyses based on Mmul_10 (Figures 4G–4I). In addition, three putative regulatory elements showing chromatin accessibility in excitatory and inhibitory neurons, oligodendrocytes, and astrocytes were annotated only in T2T-MMU8v2.0 (Figures 4G and 4H). These findings highlight the critical importance of complete and accurate genome assemblies for resolving cell-type-specific gene expression and regulatory landscapes.

DISCUSSION

An error-free, complete genome assembly represents the ultimate standard for unbiased comparative and functional genomics; yet, achieving such accuracy remains a substantial challenge[1-3]. Here, we developed a customized strategy and performed manual curation to produce a complete, near-perfect genome assembly of a rhesus macaque (T2T-MMU8v2.0). Our approach integrates QV, GCI, and read depth to generate a fully assembled genome that is devoid of error *k*-mers, unplaced contigs, and unassembled regions.

Since the release of the first complete human genome (T2T-CHM13) in 2022[6], an increasing number of assemblies have been labeled as "complete", "T2T", or "gapless."[8-10] However, the definitions of "complete," "error-free," and "QV100" remain under discussion, especially given the challenges in correctly estimating assembly quality at this high level of accuracy[2,48]. Rhie et al. proposed a set of quantitative metrics to define genome quality and completeness, yet these criteria are often inconsistently applied or overlooked[25]. To our knowledge, based on five years of development and the integration of multiple T2T genome evaluation tools,

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

the T2T-MMU8v2.0 assembly represents one of the highest standards of completeness and base-level accuracy currently achievable[6,25,26,49-53] (Table 1).

Nevertheless, we acknowledge that these quality metrics are ultimately proxies for the true accuracy of an assembly. Despite cross-platform validation and the use of orthogonal evaluation frameworks, no current approach can fully determine absolute correctness of a genome. For example, *k*-mer–based QV estimates cannot capture all error types, such a long, homopolymer length errors or phasing errors within near-identical repeats. These limitations are exacerbated for smaller values of *k*. Therefore, we refer to this assembly as complete and near-perfect rather than error-free, emphasizing its empirical quality while acknowledging the inherent limitations of current benchmarking systems.

Based on the analysis of the T2T-MMU8v2.0 genome, our study highlights subtelomeric satellite arrays as among the most structurally complex and challenging regions to assemble. Identifying the optimal ONT read path within the initial ONT-only *de Bruijn* graph was effective for resolving these loci, though the method remains semi-automated. We propose that both this approach and the resulting genome-wide map of subtelomeric organization may inform future improvements in genome assembly algorithms. Importantly, our goal was not to develop a generalized assembler, as the scalability of this approach to non-primate or complex genomes remains unclear. Instead, our study focuses on the technological and genetic insights that can be gained from a complete, near-perfect genome, particularly in rapidly evolving, satellite-rich regions.

First, we find that the majority of error *k*-mers and unresolved sequences during the upgrade to QV100 are concentrated in subtelomeric satellite regions (Figure 1). Through targeted reassembly and polishing, we show that the ONT-only strategy offers a viable solution for resolving these regions[29] and reveal sequence biases intrinsic to current long-read platforms (Figure 1). Second, our comparative analyses demonstrate that subtelomeric architecture differs markedly among primate species[9], including variation in satellite array composition, structural organization, and DNA methylation patterns (Figures 2 and 3). These regions are among the most dynamic and lineage-specific in the genomes, differences that would likely remain undetected without high-quality T2T assemblies[6,8,9,34,35,38,54-56]. Our findings also expose a fundamental limitation in human-centric genomics: overreliance on the human reference genome as a benchmarking tool can obscure substantial sequence diversity and may

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

impede the development of more broadly applicable assembly strategies. Finally, the T2T-MMU8v2.0 genome provides a robust platform for population genetics and single-cell multi-omics analysis, and serves as a high-confidence reference for comparative genomics, including cross-species structural variant discovery (Figure 4). As such, complete and near-perfect assemblies will be instrumental in advancing tool development, improving variant detection, and deepening our understanding of genome organization and evolution.

Despite these advances, several limitations remain. First, the current assembly represents a haploid genome; to our knowledge, achieving a complete and near-perfect assembly in a fully diploid genome remains an unsolved technical challenge due to the added difficulty of haplotype phasing. Second, rDNA arrays are still represented as consensus sequences in the T2T-MMU8v2.0 assembly, as current technologies are not yet able to resolve long, near-identical tandem arrays.

In summary, our study advances both sequencing methodology and biological understanding by generating a complete genome assembly with near-perfect accuracy. This work establishes a practical foundation for improved genome assembly practices and opens new avenues for investigating the structure, evolution, and function of highly repetitive genomic regions.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

TABLE and FIGURES

Table 1. The quality evaluation of representative T2T assemblies.

Assembly	Common name	Scientific name	# Chr	# Unassigned contigs	Genome size (Gbp)	N50 (kbp)	GCI	# Error 21-mers	QV (*k*=21)	Citation
T2T-MMU8v2.0	Rhesus macaque	*Macaca mulatta*	22	0	3.12	162.29	100	0	100	This study
T2T-MFA8v1.1	Crab-eating macaque	*Macaca fascicularis*	22	0	3.06	162.13	81.75	4,802	71.27	Zhang et al., 2025
T2T-CHM13v2.0	Human	*Homo sapiens*	24	0	3.12	150.62	97.11	3,003	73.38	Nurk et al., 2022
mGorGor1_v2.0_pri	Gorilla	*Gorilla gorilla*	25	0	3.55	150.80	76.30	11,701	68.03	Yoo et al., 2025
mPanPan1_v2.0_pri	Bonobo	*Pan paniscus*	25	5	3.24	147.03	77.72	15,609	66.39	Yoo et al., 2025
mPanTro3_v2.0_pri	Chimpanzee	*Pan troglodytes*	25	0	3.18	146.29	78.70	8,108	69.15	Yoo et al., 2025
mPonAbe1_v2.0_pri	Sumatran orangutan	*Pongo abelii*	25	4	3.26	146.20	43.51	30,090	63.56	Yoo et al., 2025
mPonPyg2_v2.0_pri	Bornean orangutan	*Pongo pygmaeus*	25	0	3.22	143.26	63.73	9,686	68.43	Yoo et al., 2025
mSymSyn1_v2.0_pri	Siamang	*Symphalangus syndactylus*	26	0	3.26	146.71	90.62	4,482	71.84	Yoo et al., 2025
T2T-mhaESCv1.1 (no chrY)	Mouse	*Mus musculus*	20	0	2.77	141.30	69.79	45,441	61.08	Liu et al., 2024

*The GCI score was assessed using ONT alignments, excluding gaps induced by rDNA arrays (Methods). Mitochondrial chromosomes were excluded.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

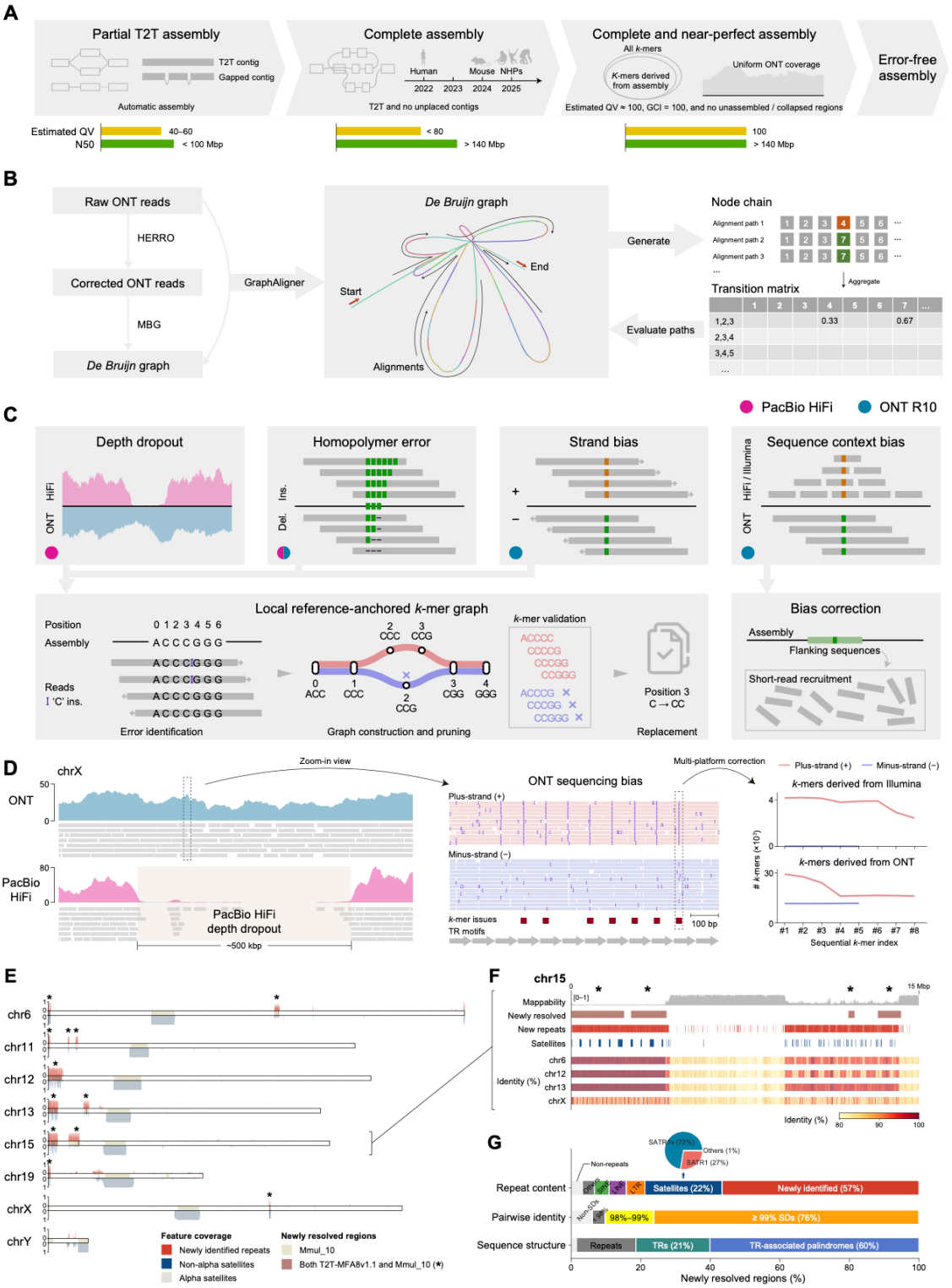


Figure 1. Assembly strategy and platform-specific biases in generating the T2T-MMU8v2.0 assembly. (A) Schematic overview of the stepwise assembly workflow, progressing from an initial draft generated by automated assemblers to the complete and near-perfect T2T-MMU8v2.0 assembly. **(B)** Custom path-finding strategy based on a corrected ONT-derived *de Bruijn* graph. The initial *de Bruijn* graph was generated corrected

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

ONT reads with MBG, and each raw ONT read was aligned to the pruned graph. Three sequential nodes covered by a single ONT read were defined as the current state, and the fourth node as the next step. A transition matrix incorporating read-path information was then used to score enumerated candidate paths. **(C)** Representative illustrations of four types of base-pair level errors, highlighting the importance of addressing technology-driven artifacts during QV100 polishing. The schematic illustration below depicts a local reference-anchored *k*-mer graph used to resolve sequencing bias. Systematic context bias was corrected using a short-read recruitment approach. Ins, insertion; del, deletion. **(D)** Sequencing bias and polishing strategy on chrX. Left: A ~500 kbp PacBio HiFi depth dropout region. Middle: Strand-specific ONT alignments within the dropout region reveal sequencing bias and *k*-mer distortion, with annotated TR motifs and affected *k*-mers. Right: Strand-specific *k*-mer support from Illumina and ONT reads confirms accurate homopolymer lengths in the final consensus. **(E)** The ideogram of newly resolved sequences in T2T-MFA8v2.0. Colored blocks show newly resolved sequences to Mmul_10 (tan) and those to both T2T-MFA8v1.1 and Mmul_10 (brown, indicated with asterisks). The coverage of newly identified repeats (red) and satellites (grey for alpha satellites and blue for non-alpha satellites) in 50-kbp windows are shown along the ideogram. **(F)** Repetitive landscape of newly resolved regions on chr15. Shown from top to bottom: *k*-mer mappability ($k = 31$), regions newly assembled compared to both T2T-MFA8v1.1 and Mmul_10, novel repeat families, non-alpha satellite annotations, and regions with $\geqslant$80% sequence identity to other representative chromosomes. **(G)** Composition and structure of newly resolved sequences relative to both T2T-MFA8v1.1 and Mmul_10. Bar plots show the proportions of repeat content, pairwise identity, and structural categories. The accompanying pie chart summarizes the satellite repeat family composition of these regions.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

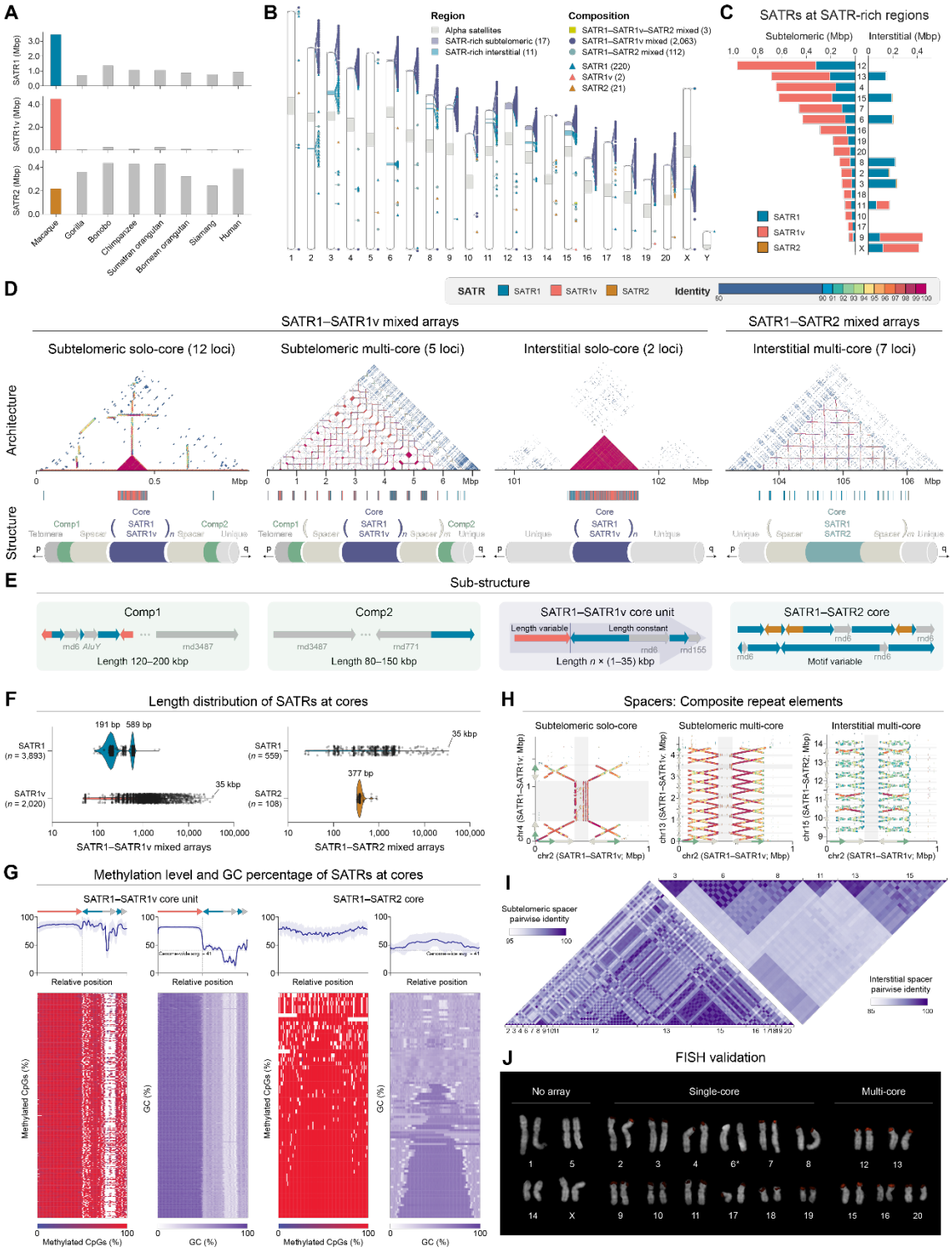
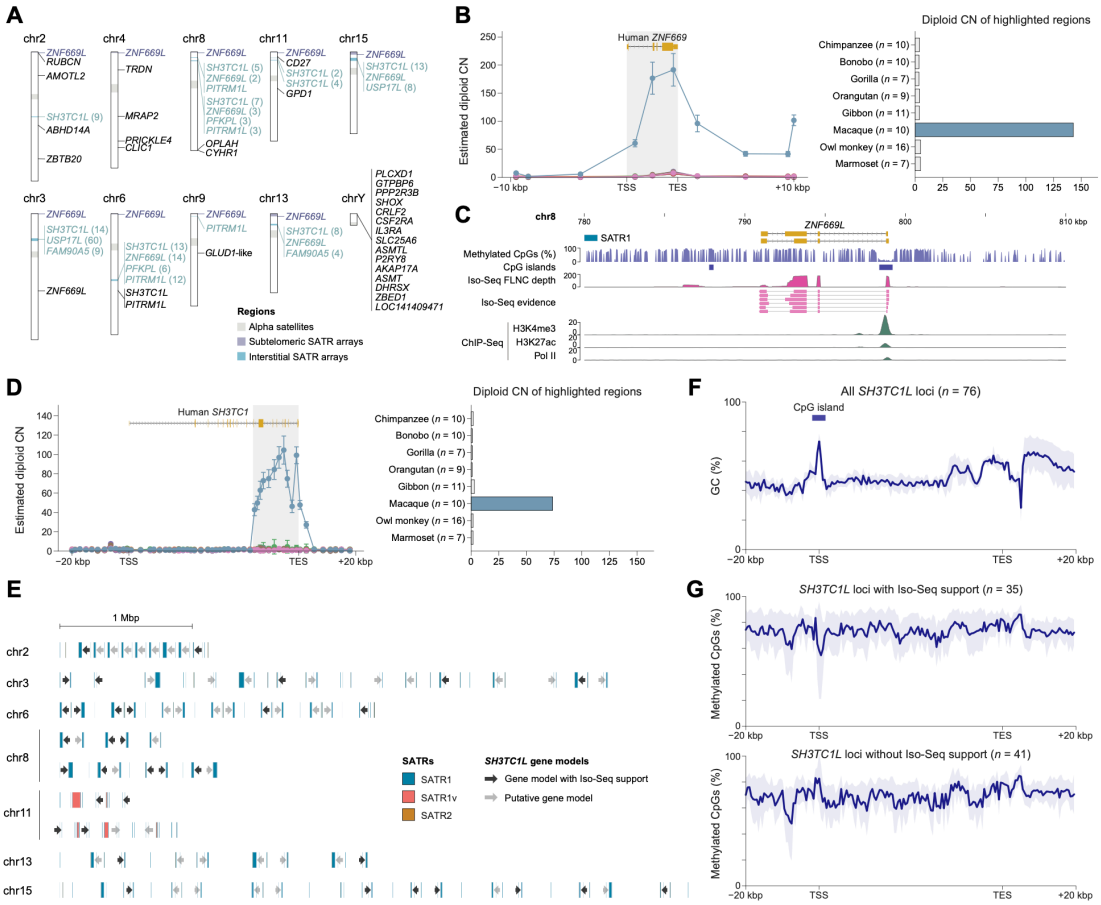


Figure 2. Organization, composition, and genomic distribution of SATR satellite families in T2T-MM8v2.0. (A) The total length of SATR1, SATR1v, and SATR2 satellite families across representative primate genomes. **(B)** Ideogram of the T2T-MMU8v2.0 with chromosomal localization of SATR arrays. Arrays composed exclusively of SATR1 (blue), SATR1v (red), or SATR2 (bronze) are indicated by triangles, respectively. Mixed arrays are shown as colored shapes: SATR1–SATR1v (dark blue circles), SATR1–SATR2 (light blue

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

circles), and SATR1–SATR1v–SATR2 (yellow rectangles). SATR-rich genomic regions are indicated in the ideogram. **(C)** Length distribution of SATR arrays in SATR-rich regions. The left panel shows array lengths in subtelomeric regions, and the right panel shows the corresponding lengths in interstitial regions. Colors represent SATR1 (blue), SATR1v (red), and SATR2 (bronze). **(D)** Structural organization and sequence identity patterns of SATR mixed arrays. Top: heatmaps show sequence identity across four distinct array configurations — (1) subtelomeric SATR1–SATR1v solo-core mixed arrays, (2) subtelomeric SATR1–SATR1v multi-core mixed arrays, (3) interstitial SATR1–SATR1v solo-core mixed arrays, and (4) interstitial SATR1–SATR2 multi-core mixed arrays. Bottom: schematic models of the corresponding structures. Subtelomeric SATR1–SATR1v mixed arrays are flanked by conserved composite repeats and SD spacers, arranged as: telomere – left composite repeats – SD spacer – mixed arrays (core) – SD spacer – right composite repeats (from p-arm to q-arm). Interstitial SATR1–SATR1v solo-core mixed arrays lack SD spacers and flanking composite repeats. SATR1–SATR2 interstitial multiple mixed arrays are flanked by SD spacers without the aforementioned composite repeats. Telomeres (dark grey), unique sequences (light grey), composite repeats (green), SD spacers (tan), SATR1–SATR2 mixed arrays (light blue), and SATR1–SATR1v mixed arrays (dark blue) are indicated. **(E)** Repeat composition of left/right composite repeats, SATR1–SATR1v core units, and SATR1–SATR2 core units. Non-SATR repeats are shown in grey. **(F)** Length distribution of SATR1, SATR1v, and SATR2 arrays within core units. **(G)** DNA methylation and GC content profiles across core units. Top: aggregated methylation and GC percentage across individual core units. The lines represent the mean values, with shaded areas indicating mean ± s.d. Bottom: heatmaps showing methylation (blue to red) and GC content (white to purple) across individual core units. **(H)** Dot plots showing high sequence identity between SD spacers. X-axis represents composite repeats (green arrows) and SD spacers (tan arrows) and mixed arrays (grey shadow) from a representative SATR1–SATR1v mixed array; y-axis shows alignments to other mixed arrays (grey shadow) configurations with composite repeats (green arrows) and SD spacers (tan arrows) across subtelomeric and interstitial loci. Colors indicate estimated sequence identity. **(I)** Pairwise sequence identity heatmap of subtelomeric SD spacers (left; scale bar: 95%–100%) and interstitial SD spacers (right; scale bar: 85%–100%). **(J)** FISH validation of SATR arrays. Strong fluorescent signals (red) correspond to 512-bp SATR1 probes hybridizing to expanded SATR1 loci across multiple chromosomes. Chr6 (indicated by an asterisk) harbors a single SATR1 core, but barely exhibits FISH signal, suggesting potential individual polymorphism or reduced probe accessibility.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.



552

553 **Figure 3. Discovery and characterization of novel genes in T2T-MMU8v2.0. (A)**
554 Ideograms of eight representative chromosomes from T2T-MMU8v2.0, highlighting newly
555 annotated genes within SATR satellite-associated and previously misassembled genomic
556 regions. **(B)** Copy number estimation at the human *ZNF669* locus shows multiple *ZNF669L*
557 copies (left), which are absent in other primate genomes (right). **(C)** Epigenetic and
558 transcriptomic features for a *ZNF669L* copy (~20 kb flanking of the comp2 on chr8),
559 including DNA methylation, Iso-Seq support, and epigenetic signals. ChIP-seq tracks are
560 normalized by BPM (bins per million mapped reads). **(D)** Copy number estimation at the
561 human *SH3TC1* locus reveals the expansion of truncated *SH3TC1L* copies (left), not
562 expanded in other primates (right). **(E)** Genomic localization of *SH3TC1L* copies with (black
563 arrows) and without (grey arrows) Iso-Seq FLNC transcript support, embedded within
564 SATR1 (blue), SATR1v (red), and SATR2 (bronze) arrays. **(F)** GC-content profile across
565 *SH3TC1L* loci shows well-defined CpG islands at the transcription start sites (TSSs). **(G)**
566 Differential methylation patterns at CpG islands in *SH3TC1L* copies: transcriptionally active
567 copies (with Iso-Seq support) show hypomethylation, while inactive copies exhibit
568 hypermethylation, supporting the regulatory potential of active copies.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

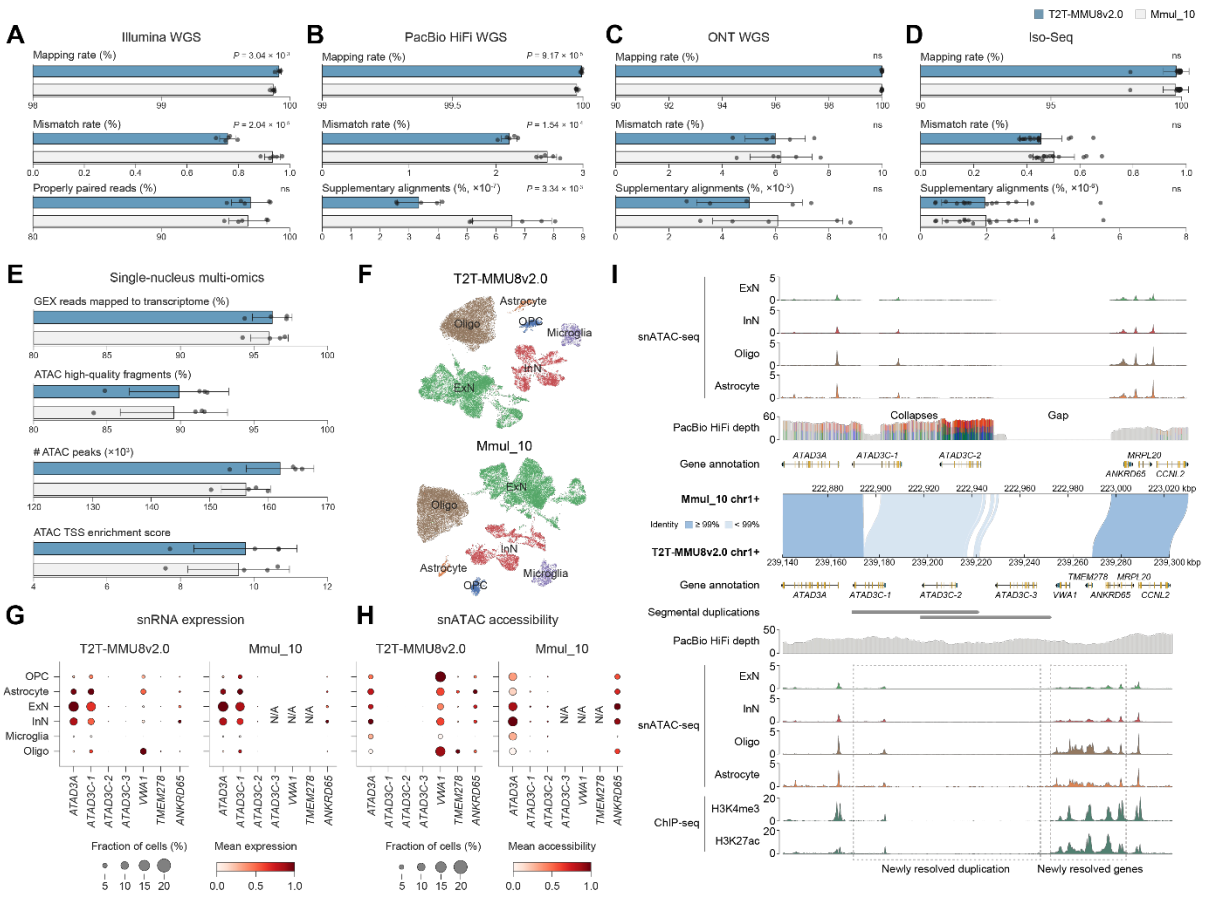


Figure 4. Performance comparison of T2T-MMU8v2.0 and Mmul_10 in population genetics and single-nucleus multi-omics analyses. (A–D) Quality control metrics for Illumina WGS (n = 5), PacBio HiFi WGS (n = 5), ONT WGS (n = 5), and RNA-Seq (Iso-Seq; n = 15) using T2T-MMU8v2.0 (blue) and Mmul_10 (grey) as references. P values were calculated using two-sided Welch's t-tests. Error bar, mean ± s.d. ns, not significant. **(E)** Quality control metrics for single-nucleus multi-omics data (n = 4) using T2T-MMU8v2.0 (blue) and Mmul_10 (grey) as references. **(F)** UMAP embedding of cell types identified using T2T-MMU8v2.0 (top) and Mmul_10 (bottom) as references. **(G)** Gene expression profiles in previously misassembled regions, analyzed using snRNA-seq with T2T-MMU8v2.0 (blue) and Mmul_10 (grey) as references. **(H)** Chromatin accessibility in previously misassembled regions, analyzed using snATAC-seq with T2T-MMU8v2.0 (blue) and Mmul_10 (grey) as references. **(I)** Comparative visualization of snATAC-seq and ChIP-seq signals in previously misassembled genomic regions, using Mmul_10 (top) and T2T-MMU8v2.0 (bottom). Collapsed and gap regions in Mmul_10 are indicated by PacBio HiFi read depth. Positions with mismatch frequencies above 20% are colored, with each color representing a different substitution. Newly assembled segmental duplications are marked by grey bars, and newly resolved duplications and genes are highlighted with grey dashed boxes.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

587 snATAC-seq tracks are normalized by CPM (counts per million), while ChIP-seq tracks are
588 normalized by BPM (bins per million mapped reads). ExN, excitatory neuron; InN, inhibitory
589 neuron; Oligo, oligodendrocyte; OPC, oligodendrocyte progenitor cell.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

Data availability

The T2T-MMU8v2.0 assembly is deposited in the National Centre for Biotechnology Information (NCBI) under accession number GCA_049350105.2. The raw PacBio HiFi, ONT, and Illumina WGS data of MMU2019108-1 and MMU1003063 (chrY) are deposited in the NCBI under BioProject accession number PRJNA1229433. T2T-MMU8 assembly, annotations, and the UCSC track hub are available at GitHub (https://github.com/zhang-shilong/T2T-MMU8) under a CC0 1.0 license. Previously published sequencing data used in this study, including PRJNA269593, PRJNA530776, PRJNA602326, PRJNA953340, PRJNA976699, PRJNA976700, PRJNA976701, PRJNA976702, PRJNA986878, PRJNA986879, PRJNA1004471, PRJNA1037719, PRJNA1041301, PRJNA1097000, and PRJCA018217, are available from the NCBI or the National Genomics Data Center.

Code availability

Custom scripts used in this study are available at GitHub (https://github.com/zhang-shilong/T2T-MMU8).

Acknowledgments

We thank Nancy F. Hansen for discussions on this manuscript. This work was supported, in part, by National Natural Science Foundation of China grants (32370658 to Y.M. and 82021001 to Q.S.); by Natural Science Foundation of Chongqing, China (CSTB2024NSCQ-JQX0004), the Computational Biology Program (24JS2840300) of Science and Technology Commission of Shanghai Municipality (STCSM), and Shanghai Jiao Tong University 2030 Initiative (WH510363003/016) to Y.M.; by National Key Research and Development Program of China (2022YFF0710901), Biological Resources Program of Chinese Academy of Sciences (KFJ-BRP-005), and National Science and Technology Innovation 2030 Major Program (2021ZD0200900) to Q.S.; and by the Intramural Research Program of the National Human Genome Research Institute, U.S. National Institutes of Health to A.M.P. The computations in this study were run on the π 2.0 cluster supported by the Center for High Performance Computing at Shanghai Jiao Tong University.

Author contributions

Y.M. and Q.S. conceived the project. N.X., Y.L., Y.N., and Q.S. generated the cell line for sequencing and assembly. S.Z. performed all data analysis. L.d.G., F.A., and M.V. performed FISH experiments. S.Z., Z.L., L.F., Z.Z., J.C., K.M., X.Y., J.Z., M.T.S., and T.E.B.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

624 performed the genome assembly validation and application experiments. A.M.P. consulted on
625 experimental design and validation. S.Z. and Y.M. drafted the manuscript. All authors read
626 and approved the manuscript.
627
628 **Competing interests**
629 The authors have declared no competing interests.
630
631 **References**

632 1. Jarvis, E.D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., Tracey, A.,
633 Thibaud-Nissen, F., Vollger, M.R., Porubsky, D., et al. (2022). Semi-automated
634 assembly of high-quality diploid human reference genomes. Nature *611*.
635 10.1038/s41586-022-05325-5.
636 2. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-
637 Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete
638 and error-free genome assemblies of all vertebrate species. Nature *592*.
639 10.1038/s41586-021-03451-0.
640 3. Li, H., Durbin, R., Li, H., and Durbin, R. (2024). Genome assembly in the telomere-
641 to-telomere era. Nature Reviews Genetics *25*. 10.1038/s41576-024-00718-w.
642 4. Mao, Y., Zhang, G., Mao, Y., and Zhang, G. (2022). A complete, telomere-to-
643 telomere human genome sequence presents new opportunities for evolutionary
644 genomics. Nature Methods *19*. 10.1038/s41592-022-01512-4.
645 5. Logsdon, G.A., Vollger, M.R., Eichler, E.E., Logsdon, G.A., Vollger, M.R., and
646 Eichler, E.E. (2020). Long-read human genome sequencing and its applications.
647 Nature Reviews Genetics *21*. 10.1038/s41576-020-0236-x.
648 6. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A.,
649 Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete
650 sequence of a human genome. Science *376*. 10.1126/science.abj6987.
651 7. Yang, C., Zhou, Y., Song, Y., Wu, D., Zeng, Y., Nie, L., Liu, P., Zhang, S., Chen, G.,
652 Xu, J., et al. (2023). The complete and fully-phased diploid genome of a male Han
653 Chinese. Cell Research *33*. 10.1038/s41422-023-00849-5.
654 8. Zhang, S., Xu, N., Fu, L., Yang, X., Ma, K., Li, Y., Yang, Z., Li, Z., Feng, Y., Jiang,
655 X., et al. (2025). Integrated analysis of the complete sequence of a macaque genome.
656 Nature *640*. 10.1038/s41586-025-08596-w.
657 9. Yoo, D., Rhie, A., Hebbar, P., Antonacci, F., Logsdon, G.A., Solar, S.J., Antipov, D.,
658 Pickett, B.D., Safonova, Y., Montinaro, F., et al. (2025). Complete sequencing of ape
659 genomes. Nature *641*. 10.1038/s41586-025-08816-3.
660 10. Liu, J., Li, Q., Hu, Y., Yu, Y., Zheng, K., Li, D., Qin, L., and Yu, X. (2024). The
661 complete telomere-to-telomere sequence of a mouse genome. Science *386*.
662 10.1126/science.adq8191.
663 11. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., Cheng, H., Concepcion,
664 G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly
665 using phased assembly graphs with hifiasm. Nature Methods *18*. 10.1038/s41592-
666 020-01056-5.
667 12. Cheng, H., Asri, M., Lucas, J., Koren, S., Li, H., Cheng, H., Asri, M., Lucas, J.,
668 Koren, S., and Li, H. (2024). Scalable telomere-to-telomere assembly for diploid and

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

polyploid genomes with double graph. Nature Methods *21*. 10.1038/s41592-024-02269-8.

13. Rautiainen, M., Nurk, S., Walenz, B.P., Logsdon, G.A., Porubsky, D., Rhie, A., Eichler, E.E., Phillippy, A.M., Koren, S., Rautiainen, M., et al. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nature Biotechnology *41*. 10.1038/s41587-023-01662-6.
14. Liu, Z., Cai, Y., Wang, Y., Nie, Y., Zhang, C., Xu, Y., Zhang, X., Lu, Y., Wang, Z., Poo, M., and Sun, Q. (2018). Cloning of Macaque Monkeys by Somatic Cell Nuclear Transfer. Cell *172*. 10.1016/j.cell.2018.01.020.
15. Luo, X., Liu, Y., Dang, D., Hu, T., Hou, Y., Meng, X., Zhang, F., Li, T., Wang, C., Li, M., et al. (2021). 3D Genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. Cell *184*. 10.1016/j.cell.2021.01.001.
16. Liu, Z., Cai, Y., Liao, Z., Xu, Y., Wang, Y., Wang, Z., Jiang, X., Li, Y., Lu, Y., Nie, Y., et al. (2019). Cloning of a gene-edited macaque monkey by somatic cell nuclear transfer. National Science Review *6*. 10.1093/nsr/nwz003.
17. Wu, J., Li, Y., Huang, Y., Liu, L., Zhang, H., Nagy, C., Tan, X., Cheng, K., Liu, Y., Pu, J., et al. (2023). Integrating spatial and single-nucleus transcriptomic data elucidates microglial-specific responses in female cynomolgus macaques with depressive-like behaviors. Nature Neuroscience *26*. 10.1038/s41593-023-01379-4.
18. Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R., Remington, K.A., Strausberg, R.L., Venter, J.C., Wilson, R.K., et al. (2007). Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. Science. 10.1126/science.1139247.
19. Warren, W.C., Harris, R.A., Haukness, M., Fiddes, I.T., Murali, S.C., Fernandes, J., Dishuck, P.C., Storer, J.M., Raveendran, M., Hillier, L.W., et al. (2020). Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. Science *370*. 10.1126/science.abc6617.
20. Mao, Y., Harvey, W.T., Porubsky, D., Munson, K.M., Hoekzema, K., Lewis, A.P., Audano, P.A., Rozanski, A., Yang, X., Zhang, S., et al. (2024). Structurally divergent and recurrently mutated regions of primate genomes. Cell *187*. 10.1016/j.cell.2024.01.052.
21. Hao, S., Zhu, X., Huang, Z., Yang, Q., Liu, H., Wu, Y., Zhan, Y., Dong, Y., Li, C., Wang, H., et al. (2024). Cross-species single-cell spatial transcriptomic atlases of the cerebellar cortex. Science *385*, eado3927. 10.1126/science.ado3927.
22. Rogers, J., Gibbs, R.A., Rogers, J., and Gibbs, R.A. (2014). Comparative primate genomics: emerging patterns of genome content and dynamics. Nature Reviews Genetics *15*. 10.1038/nrg3707.
23. Lei, Y., Liu, Y., Wang, M., Yuan, N., Hou, Y., Ding, L., Zhu, Z., Wu, Z., Li, C., Zheng, M., et al. (2025). Single-cell spatial transcriptome atlas and whole-brain connectivity of the macaque claustrum. Cell *188*. 10.1016/j.cell.2025.02.037.
24. Yang, H., Liu, Z., Ma, Y., Zhong, C., Yin, Q., Zhou, C., Shi, L., Cai, Y., Zhao, H., Wang, H., et al. (2013). Generation of haploid embryonic stem cells from Macaca fascicularis monkey parthenotes. Cell Research *23*. 10.1038/cr.2013.93.
25. Rhie, A., Walenz, B.P., Koren, S., Phillippy, A.M., Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology *21*. 10.1186/s13059-020-02134-9.
26. Chen, Q., Yang, C., Zhang, G., and Wu, D. (2024). GCI: a continuity inspector for complete genome assembly. Bioinformatics *40*. 10.1093/bioinformatics/btae633.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

27. Stanojević, D., Lin, D., Nurk, S., Sessions, P.F.d., and Šikić, M. (2024). Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads. bioRxiv. 10.1101/2024.05.18.594796.
28. Rautiainen, M., and Marschall, T. (2021). MBG: Minimizer-based sparse de Bruijn Graph construction. Bioinformatics *37*. 10.1093/bioinformatics/btab004.
29. Cheng, H., Qu, H., McKenzie, S., Lawrence, K.R., Windsor, R., Vella, M., Park, P.J., and Li, H. (2025). Efficient near telomere-to-telomere assembly of Nanopore Simplex reads. bioRxiv. 10.1101/2025.04.14.648685.
30. Mc Cartney, A.M., Shafin, K., Alonge, M., Bzikadze, A.V., Formenti, G., Fungtammasan, A., Howe, K., Jain, C., Koren, S., Logsdon, G.A., et al. (2022). Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. Nature Methods *19*. 10.1038/s41592-022-01440-3.
31. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology *36*. 10.1038/nbt.4060.
32. Vollger, M.R., Logsdon, G.A., Audano, P.A., Sulovari, A., Porubsky, D., Peluso, P., Wenger, A.M., Concepcion, G.T., Kronenberg, Z.N., Munson, K.M., et al. (2020). Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. Annals of Human Genetics *84*. 10.1111/ahg.12364.
33. Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology *37*. 10.1038/s41587-019-0217-9.
34. Hoyt, S.J., Storer, J.M., Hartley, G.A., Grady, P.G.S., Gershman, A., Lima, L.G.d., Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., et al. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science *376*. 10.1126/science.abk3112.
35. Altemose, N., Logsdon, G.A., Bzikadze, A.V., Sidhwani, P., Langley, S.A., Caldas, G.V., Hoyt, S.J., Uralsky, L., Ryabov, F.D., Shew, C.J., et al. (2022). Complete genomic and epigenetic maps of human centromeres. Science *376*. 10.1126/science.abl4178.
36. Wheeler, T.J., Clements, J., Eddy, S.R., Hubley, R., Jones, T.A., Jurka, J., Smit, A.F.A., and Finn, R.D. (2013). Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Research *41*. 10.1093/nar/gks1265.
37. Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A., and Wheeler, T.J. (2016). The Dfam database of repetitive DNA families. Nucleic Acids Research *44*. 10.1093/nar/gkv1272.
38. Yoo, D., Munson, K.M., and Eichler, E.E. (2025). Epigenetic and evolutionary features of ape subterminal heterochromatin. bioRxiv. 10.1101/2025.05.29.656835.
39. Mao, Y.-X., Li, Y., Yang, Z., Xu, N., Zhang, S., Wang, X., Yang, X., Sun, Q., Mao, Y., Mao, Y.-X., et al. (2023). Comparative transcriptome analysis between rhesus macaques (Macaca mulatta) and crab-eating macaques (M. fascicularis). Zoological Research *45*. 10.24272/j.issn.2095-8137.2023.322.
40. Bosch, N., Cáceres, M., Cardone, M.F., Carreras, A., Ballana, E., Rocchi, M., Armengol, L., and Estivill, X. (2007). Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events. Human Molecular Genetics *16*. 10.1093/hmg/ddm209.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

41. Soto, D.C., Uribe-Salazar, J.M., Kaya, G., Valdarrago, R., Sekar, A., Haghani, N.K., Hino, K., La, G., Mariano, N.A.F., Ingamells, C., et al. (2025). Human-specific gene expansions contribute to brain evolution. Cell *0*. 10.1016/j.cell.2025.06.037.
42. Iskow, R.C., Gokcumen, O., Abyzov, A., Malukiewicz, J., Zhu, Q., Sukumar, A.T., Pai, A.A., Mills, R.E., Habegger, L., Cusanovich, D.A., et al. (2012). Regulatory element copy number differences shape primate expression profiles. Proceedings of the National Academy of Sciences *109*. 10.1073/pnas.1205199109.
43. Gokcumen, O., Tischler, V., Tica, J., Zhu, Q., Iskow, R.C., Lee, E., Fritz, M.H.-Y., Langdon, A., Stütz, A.M., Pavlidis, P., et al. (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. Proceedings of the National Academy of Sciences *110*. 10.1073/pnas.1305904110.
44. Zemke, N.R., Armand, E.J., Wang, W., Lee, S., Zhou, J., Li, Y.E., Liu, H., Tian, W., Nery, J.R., Castanon, R.G., et al. (2023). Conserved and divergent gene regulatory programs of the mammalian neocortex. Nature *624*. 10.1038/s41586-023-06819-6.
45. Deschauer, M., Hengel, H., Rupprich, K., Kreiß, M., Schlotter-Weigel, B., Grimmel, M., Admard, J., Schneider, I., Alhaddad, B., Gazou, A., et al. (2021). Bi-allelic truncating mutations in VWA1 cause neuromyopathy. Brain *144*. 10.1093/brain/awaa418.
46. Pagnamenta, A.T., Kaiyrzhanov, R., Zou, Y., Da'as, S.I., Maroofian, R., Donkervoort, S., Dominik, N., Lauffer, M., Ferla, M.P., Orioli, A., et al. (2021). An ancestral 10-bp repeat expansion in VWA1 causes recessive hereditary motor neuropathy. Brain *144*. 10.1093/brain/awaa420.
47. Nagy, S., Pagnamenta, A.T., Cali, E., Braakman, H.M.H., Wijntjes, J., Kusters, B., Gotkine, M., Elpeleg, O., Meiner, V., Lenberg, J., et al. (2024). Autosomal recessive VWA1-related disorder: comprehensive analysis of phenotypic variability and genetic mutations. Brain Communications *6*. 10.1093/braincomms/fcae377.
48. Wang, P., and Wang, F. (2023). A proposed metric set for evaluation of genome assembly quality. Trends in Genetics *39*. 10.1016/j.tig.2022.10.005.
49. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature *617*. 10.1038/s41586-023-05896-x.
50. Huang, N., and Li, H. (2023). compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics *39*. 10.1093/bioinformatics/btad595.
51. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*. 10.1093/bioinformatics/btv351.
52. Li, K., Xu, P., Wang, J., Yi, X., Jiao, Y., Li, K., Xu, P., Wang, J., Yi, X., and Jiao, Y. (2023). Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. Nature Communications *14*. 10.1038/s41467-023-42336-w.
53. Chen, Y., Zhang, Y., Wang, A.Y., Gao, M., Chong, Z., Chen, Y., Zhang, Y., Wang, A.Y., Gao, M., and Chong, Z. (2021). Accurate long-read de novo assembly evaluation with Inspector. Genome Biology *22*. 10.1186/s13059-021-02527-4.
54. Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.P., et al. (2022). Segmental duplications and their variation in a complete human genome. Science. 10.1126/science.abj6965.
55. Jeong, H., Dishuck, P.C., Yoo, D., Harvey, W.T., Munson, K.M., Lewis, A.P., Kordosky, J., Garcia, G.H., Yilmaz, F., Hallast, P., et al. (2025). Structural

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

polymorphism and diversity of human segmental duplications. Nature Genetics *57*. 10.1038/s41588-024-02051-8.

56. Ma, K., Yang, X., and Mao, Y. (2025). Advancing evolutionary medicine with complete primate genomes and advanced biotechnologies. Trends in Genetics *41*. 10.1016/j.tig.2024.11.001.
57. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*. 10.1093/bioinformatics/bty560.
58. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 10.48550/arXiv.1303.3997.
59. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Auwera, G.A.V.d., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 10.1101/201178.
60. Jain, C., Rhie, A., Hansen, N.F., Koren, S., Phillippy, A.M., Jain, C., Rhie, A., Hansen, N.F., Koren, S., and Phillippy, A.M. (2022). Long-read mapping to repetitive reference sequences using Winnowmap2. Nature Methods *19*. 10.1038/s41592-022-01457-8.
61. Hu, J., Wang, Z., Liang, F., Liu, S.-L., Ye, K., and Wang, D.-P. (2024). NextPolish2: A Repeat-aware Polishing Tool for Genomes Assembled Using HiFi Long Reads. Genomics, Proteomics & Bioinformatics *22*. 10.1093/gpbjnl/qzad009.
62. Rautiainen, M., Marschall, T., Rautiainen, M., and Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biology *21*. 10.1186/s13059-020-02157-2.
63. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*. 10.1093/bioinformatics/btq033.
64. Astashyn, A., Tvedte, E.S., Sweeney, D., Sapojnikov, V., Bouk, N., Joukov, V., Mozes, E., Strope, P.K., Sylla, P.M., Wagner, L., et al. (2024). Rapid and sensitive detection of genome contamination at scale with FCS-GX. Genome Biology *25*. 10.1186/s13059-024-03198-7.
65. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*. 10.1093/bioinformatics/bty191.
66. Vollger, M.R. (2021). mrvollger/NucFreq. Zenodo. 10.5281/zenodo.5570288.
67. He, Y., Luo, X., Zhou, B., Hu, T., Meng, X., Audano, P.A., Kronenberg, Z.N., Eichler, E.E., Jin, J., Guo, Y., et al. (2019). Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. Nature Communications *10*. 10.1038/s41467-019-12174-w.
68. Rautiainen, M. (2024). Ribotin: automated assembly and phasing of rDNA morphs. Bioinformatics *40*. 10.1093/bioinformatics/btae124.
69. Smit, A.F.A., Hubley, R., and Green, P. (2013-2015). RepeatMasker Open-4.0. http://www.repeatmasker.org.
70. Bao, W., Kojima, K.K., Kohany, O., Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA *6*. 10.1186/s13100-015-0041-9.
71. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., Flynn, J.M., Hubley, R., Goubert, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences *117*. 10.1073/pnas.1921046117.
72. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research *27*. 10.1093/nar/27.2.573.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

73. Numanagić, I., Gökkaya, A.S., Zhang, L., Berger, B., Alkan, C., and Hach, F. (2018). Fast characterization of segmental duplications in genome assemblies. Bioinformatics *34*. 10.1093/bioinformatics/bty586.
74. Sweeten, A.P., Schatz, M.C., and Phillippy, A.M. (2024). ModDotPlot—rapid and interactive visualization of tandem repeats. Bioinformatics *40*. 10.1093/bioinformatics/btae493.
75. Jain, C., Koren, S., Dilthey, A., Phillippy, A.M., and Aluru, S. (2018). A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics *34*. 10.1093/bioinformatics/bty597.
76. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology *14*. 10.1371/journal.pcbi.1005944.
77. Yang, Z., Zhang, S., Logsdon, G.A., and Mao, Y. (2025). VAMPIRE: Analyzing variation and motif pattern in tandem repeats. bioRxiv. 10.1101/2025.06.15.659631.
78. Pockrandt, C., Alzamel, M., Iliopoulos, C.S., and Reinert, K. (2020). GenMap: ultra-fast computation of genome mappability. Bioinformatics *36*. 10.1093/bioinformatics/btaa222.
79. Shumate, A., and Salzberg, S.L. (2021). Liftoff: accurate mapping of gene annotations. Bioinformatics *37*. 10.1093/bioinformatics/btaa1016.
80. Pardo-Palacios, F.J., Arzalluz-Luque, A., Kondratova, L., Salguero, P., Mestre-Tomás, J., Amorín, R., Estevan-Morió, E., Liu, T., Nanni, A., McIntyre, L., et al. (2024). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nature Methods. 10.1038/s41592-024-02229-2.
81. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics *14*. 10.1093/bib/bbs017.
82. Vollger, M.R., and Harvey, W.T. (2023). mrvollger/fastCN-smk: v0.2. Zenodo. 10.5281/zenodo.8136270.
83. Ning, C., Wu, X., Zhao, X., Lu, Z., Yao, X., Zhou, T., Yi, L., Sun, Y., Wu, S., Liu, Z., et al. (2024). Epigenomic landscapes during prefrontal cortex development and aging in rhesus. National Science Review *11*. 10.1093/nsr/nwae213.
84. Langmead, B., Salzberg, S.L., Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*. 10.1038/nmeth.1923.
85. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research *44*. 10.1093/nar/gkw257.
86. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology *9*. 10.1186/gb-2008-9-9-r137.
87. Lopez-Delisle, L., Rabbani, L., Wolff, J., Bhardwaj, V., Backofen, R., Grüning, B., Ramírez, F., and Manke, T. (2021). pyGenomeTracks: reproducible plots for multivariate genomic datasets. Bioinformatics *37*. 10.1093/bioinformatics/btaa692.
88. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*. 10.1093/bioinformatics/bts635.
89. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G.P.D.P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*. 10.1093/bioinformatics/btp352.

bioRxiv preprint doi: https://doi.org/10.1101/2025.08.04.668424; this version posted August 4, 2025. 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-NC-ND 4.0 International license.

90. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*. 10.1093/gigascience/giab008.
91. Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N., Mumbach, M.R., Pierce, S.E., Corces, M.R., et al. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature Biotechnology *37*. 10.1038/s41587-019-0206-z.
92. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nature Communications *8*. 10.1038/ncomms14049.
93. Wolf, F.A., Angerer, P., Theis, F.J., Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biology *19*. 10.1186/s13059-017-1382-0.
94. Bredikhin, D., Kats, I., Stegle, O., Bredikhin, D., Kats, I., and Stegle, O. (2022). MUON: multimodal omics analysis framework. Genome Biology *23*. 10.1186/s13059-021-02577-8.
95. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., Raychaudhuri, S., et al. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods *16*. 10.1038/s41592-019-0619-0.
96. Slowikowski, K., Arevalo, J., and Manning, J. (2024). slowkow/harmonypy: harmonypy version 0.0.10. Zenodo. 10.5281/zenodo.12693230.