# The complete human diploid reference genome of RPE-1 identifies the phased epigenetic landscapes from multi-omics data

Emilia Volpe[1], Luca Corda[1], Elena Di Tommaso[1], Franca Pelliccia[1], Riccardo Ottalevi[2], Danilo Licastro[3], Giulio Formenti[4], Mattia Capulli[5], Andrea Guarracino[6], Evelyne Tassone[1], Simona Giunta[1,7]

[1] Giunta Laboratory of Genome Evolution, Department of Biology and Biotechnologies Charles Darwin, University of Rome "Sapienza", Piazzale Aldo Moro 5, 00185 Rome, Italy

[2] Department of Bioinformatic, Dante Genomics Corp Inc., 667 Madison Avenue, New York, NY 10065 USA and S.s.17, 67100, L'Aquila, Italy

[3] Area Science Park, Padriciano 99, 34149 Trieste, Italy

[4] The Rockefeller University, 1230 York Avenue, 10065 New York, USA

[5] Department of Biotechnological and Applied Clinical Sciences, University of L'Aquila, L'Aquila, Italy

[6] Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA

[7] Lead Contact and Correspondence: Simona Giunta (simona.giunta@uniroma1.it)

Highlights

- To improve the standard practice of aligning sequencing reads originating from laboratory human diploid cell lines against a reference, we present a novel approach that leverages the matched, or 'isogenomic' reference genome for multi-omics analyses.

- To achieve this, we have generated a phased human diploid genome assembly of one of the most widely used non-cancer cell lines (RPE-1) with a stable diploid karyotype. We used this genome as a matched reference to analyze sequencing data from RPE-1.

- We demonstrate how the isogenomic reference genome improves read alignment quality, peak calling accuracy and faithful assignment of sequencing data to each haplotype, uncovering for the first time inter-haplotype variation within the same diploid genome at polymorphic loci such as centromeres.

- The RPE-1 genome represents a new telomere-to-telomere human diploid reference for the scientific community that will advance genetic and epigenetic research across fields using this cell line.

**SUMMARY**

Comparative analysis of recent human genome assemblies highlights profound sequence divergence that peaks within polymorphic loci such as centromeres. This raises the question about the adequacy of relying on human reference genomes to accurately analyze sequencing data derived from experimental cell lines. Here we propose a novel approach, referred to as 'isogenomic reference', that leverages a matched reference genome to perform multi-omics analyses. We have generated a new diploid genome assembly for the human retinal epithelial cells (RPE-1), a widely used non-cancer laboratory cell line with a stable diploid karyotype, that presents phased haplotypes and chromosome-level scaffolds that completely span centromeres. Using this assembly, we have characterized haplotype-resolved genomic variation unique to RPE-1, including a stable marker chromosome X with a 73.18 Mb segmental duplication of chromosome 10 translocated onto the microdeleted telomere t(X*q*;10*q*), specific to this cell line. Comparative analyses revealed sequence polymorphism within centromeric regions, including unexpected genetic and epigenetic diversity among haplotypes for all chromosomes. Using our assembly as reference, we re-analyzed both our own and publicly available sequencing, methylation and epigenetic data generated in RPE-1 which had previously been analyzed with non-matched and non-diploid reference genomes. Our results show that the isogenomic reference improves alignments with an increased mapping quality up to 85% while halving mismatches, resulting in significant changes in peaks calling related to centromeres. Our work represents a proof-of-concept, showcasing the use of matched reference genomes for multiomics analyses and, at scale, serves as the foundation for a call to comprehensively assemble experimentally relevant cell lines for a widespread application of isogenomic reference genomes.

KEYWORDS: Human reference; diploid genome; *de novo* assembly; genome reference; centromere assembly; laboratory cell line; multi-omics analysis; epigenetics; human polymorphism; experimental cell line; isogenomic reference.

## INTRODUCTION

Recent advances in DNA sequencing and genome assembly have led to the achievement of complete human genomes. These include the near-homozygous reference genome CHM13 [1] and personal genomes from individuals of diverse ancestry [2,3]. These, and others that have since been added [4–6], represent an important step toward a more equitable and fair representation of human genomic diversity [3,7]. Comparative analyses of the currently available complete assemblies have revealed an unexpected level of sequence variation [3,8]. Complete understanding of the divergence between individuals [3] and population-level variants [9–11], including their functional significance, is ongoing. In addition to allelic variation within the population, differences within an individual genome pertaining to the parental alleles can be reliably explored and characterized as more completely phased diploid genomes become available [2,12]. In contrast with previous evidence pointing to low haplotypes variation [13], the recent human diploid genome assembly HG002 shows remarkable sequence divergence between haplotypes, with non-synonymous amino acid changes in almost 48% protein-coding genes, that peaks at polymorphic loci [2]. This profound genomic diversity [3] underlies phenotypic traits [14–16] and significantly affects targeted gene editing and in turn, clinical outcomes [17]. The extent of sequence variation between genomes of different individuals highlights the need to reconsider our current approach based on reference genomes. We reasoned that the analysis of multi-omics sequencing data generated from laboratory cell lines presents challenges when using a non-matched reference, the so-called 'reference bias' [18]. These challenges are particularly important for faithful studies of our most polymorphic and divergent regions including centromeres [8]. Human centromeres consist of repetitive monomers of alpha-satellite DNA hierarchically organized into reiterating higher order repeats (HORs) arrays [19–21] that span from 0.3 up to 5 Mb in size (in CHM13) [1,22]. Centromeres' essential function in chromosome segregation is epigenetically supported by the presence of the centromere-specific histone H3 variant CENP-A [23,24]. Two features were recently identified that mark the kinetochore site within the active HOR region: a sudden drop in methylation, which prompted the recent naming as centromere dip region (CDR) [22], and an increased density of CENP-A nucleosomes estimated to reach ~1 every 4 H3 nucleosomes in the CDR compared to 1 every 20 canonical nucleosomes in the rest of the active region [25]. Centromeres are thought to be inherited as intact haplotypes [26] with reduced intervening meiotic recombination (Di Tommaso & Giunta, in press) that instead shuffles the rest of the maternal and paternal genomes to make the new, unique individual. Outside of meiosis, they appear to withstand high mutational burden [27,28], DNA damage [29–31] and crossing-over [32–34]. Yet their study has been greatly hindered by the inability to faithfully assemble and phase such complex repetitive DNA *via* short-read sequencing [35]. Advances in long-read sequencing technologies Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) high-fidelity (HiFi), along with improved computational tools for base calling, scaffolding, and assembly, enabled the resolution of the first complete human genome (CHM13). CHM13 included for the first time linear annotation of the most recalcitrant human repetitive loci, namely centromeric satellites, ribosomal DNA (rDNA) and other hard-to-assemble repeats [1]. In spite of these significant steps [1,22,36–39], reliable assembly of complete centromeres for all chromosomes remains a challenge [40]. Centromeres were in fact omitted from comparative analyses in the recent human Pangenome draft [3]. Recent genetic and epigenetic characterization of complete centromeres leveraged near-homozygous *ad hoc* cell lines [1,22,27], that do not allow for the

assessment of inter-haplotype polymorphism within the same diploid individual. The CHM13hTERT cell line was derived from a non-viable molar pregnancy of an anucleate oocyte fertilized by a single sperm. The absence of maternal DNA and the duplication of the paternal to give a single human haplotype eliminates complication from allelic variation and facilitates complete assembly [36,41,42]. A second genome used for human centromere reconstruction was also from an hTERT immortalized hydatidiform mole cell line (CHM1) [27]. To date, we lack a high quality human diploid reference genome from a human cell line that is amenable to laboratory experimentation and can be used to interrogate genetic and epigenetic changes between genomes and haplotypes, including previously unattainable allelic variation at centromeres [21,28]. The hTERT immortalized retinal pigment epithelial line (RPE-1) is one of the most commonly used non-cancer laboratory cell lines, counting thousands of scientific publications across fields from chromosome to cell biology [43]. Because of their stable diploid karyotype, and extensive sequencing data publicly available, we chose this cell line to generate a new telomere-to-telomere (T2T) human diploid genome assembly with phased haplotypes and used it to baseline variation at centromeres. RPE1 sequencing data currently available have been previously analyzed using haploid human reference genomes. Here, we used this assembly as diploid reference genome to improve multiomics analyses with reads from RPE-1 (Fig. 1a). We coined the term '*isogenomic*' in line with the words 'isogenic' or 'isogenetic', to refer to the same genetic make-up not just within coding regions, but across the whole genome. We demonstrate how the *isogenomic* reference genome better supports accurate multi-omics analyses from experiments done in the RPE-1 cell line, whole-genome and especially at centromeres and polymorphic loci with the highest diversity, and influences the resulting epigenomic landscape. Altogether, we present the complete human reference genome of a diploid experimentally-relevant laboratory cell line, including fully-phased haplotypes spanning the centromeres, to address reference bias and advance our understanding of human genetic and epigenetic variation at centromeres. Our work represents a proof-ofconcept that calls for a comprehensive catalog of complete genome assemblies for commonly used cells, including diploid embryonic and induced pluripotent stem cells (ESC and iPSC), primary and disease lines, for a widespread application of *isogenomic reference* genomes for faithful multi-omics analyses.

## RESULTS

### A telomere-to-telomere human diploid reference genome of the RPE-1 cell line

To improve the standard practice of aligning sequencing reads originated from laboratory cell lines against a human reference genome from a different individual, we present a new approach for multi-omics analyses using the matched or 'isogenomic reference genome' (Fig. 1a). First, (1) alignment against a haploid reference prevents the assessment of differences between haplotypes and to determine the phased epigenetic landscapes specific to that cell line. Also, despite improvement in accuracy and completeness of reference genomes [1], (2) alignment quality and (3) mapped reads accuracy can be compromised due to underlying human genomic variation. To address these points (Fig. 1a), we generated a *de novo* diploid assembly of one of the most commonly used non-cancer laboratory cell lines for multi-omics experiments, the human retinal epithelial cells RPE-1. We produced state-of-the-art sequencing data aiming at above-average read depth coverage, with ~46x coverage of HiFi reads, ~80x ONT long (>70 kb) and ~30x UltraLong (UL; >100 kb), as well as ~120x Illumina and ~30x of Hi-C linked reads (Table 1). ONT reads were exclusively generated using the V14 chemistry on R10.14.1 nanopores with >99% accuracy without additional polishing as needed previously [2]. To isolate native DNA from hTERT RPE-1 cells for ONT-UL, we improved the yield of high molecular weight with peaks up to 1 Mb (Supplementary Fig. 1, Methods). Once the suitable coverage was achieved (Supplementary Fig. 2), we compared two *de novo* whole-genome assemblers, Hifiasm [12,44] (Supplementary Fig. 3) and Verkko [45] (Supplementary Fig. 4), and proceeded with the latter due to contiguity in complex regions and computational resource compatibility (Supplementary Note 1). To achieve complete phasing of RPE-1 diploid genome in absence of parental information for trio-binning [2], we produced Hi-C data for RPE-1 and integrated them onto the unphased Verkko's output to obtain two haplotype blocks (Supplementary Fig. 5a-b). Hi-C reads were mapped onto the total assembly, haplotype 1 (Hap1) and haplotype 2 (Hap2), as well as onto the assembly with the unassigned reads to obtain the final 23 scaffolds for each haplotype and resolve all possible missed joins (Supplementary Note 2, Supplementary Fig. 5c). Using this final Hi-C contact map, we performed manual dual curation [2] (Methods) to generate 23 chromosome-level scaffolds resolved for each haplotype (Supplementary Fig. 6). Genome quality was evaluated using the reference-free Merqury *k*-mer analysis tool [46], BUSCO [47], Asmgene [48] and our own integrative assessment method (Corda & Giunta, unpublished). The RPE-1 N50, coverage and sequencing data quality is comparable to the latest benchmarks from a variety of consortia (i.e. HPRC, T2T, HGSVC) for producing high-quality assemblies (Table 1; Supplementary Fig. 7a-e). The final RPE-1 assembly (RPE1v1.0) was validated for high accuracy (Methods) estimated base accuracy >99.9999% (QV>60), specifically a QV of 64.1 on haplotype 1 (Hap1) and 61.8 on haplotype 2 (Hap2) (Table 1).

We found that RPE-1v1.0 chromosomes length was comparable to previous genome assemblies CHM13 (T2T-CHM13v2.0), HG38 (GRCh38.p14) and HG002 (HG002v0.7) paternal and maternal haplotypes, except for a copy of RPE-1 chromosome X of haplotype 1 with a total length of 227.21 Mb that deviates from the expected 155 Mb (Fig. 1b). The final RPE-1v1.0 total genomic content, after manual curation and translocation validation, was 3.06 Gbp and 2.99 Gbp for Hap1 and Hap2. We ascribed the difference in size between the haplotypes to a large structural variant (SV)

present in RPE-1 cells (Fig. 1a; Supplementary Fig. 8a). In agreement with this, T2T-RPE-1 assembly shows a total of 49 gaps for haplotype 1 (Hap1) and 25 gaps for haplotype 2 (Hap2) respectively, which did not localize within repetitive regions – implying complete centromeres, but pertain to chromosomes 10 and X due to the SV (Supplementary Fig. 8b-d). The RPE-1 karyotype confirmed the presence of the aberrant chromosome X with a large segmental duplication of chromosome 10 translocated onto chromosome X to form t(Xq;10q), still retaining two intact copies of chromosome 10 and one intact copy of chromosome X (Supplementary Fig. 8e). Notably, this is not a *de novo* somatic rearrangement but corresponds to a stable marker chromosome for the RPE-1 cell line that we [30,32] and others [43,49–51] had previously observed by cytological and sequencing analyses. The chromosome contacts between X and 10 were also flagged on the dual Hi-C contact map for both haplotypes (Supplementary Fig. 5a) and read coverage increased compatibly with a duplicated 10*q* (Supplementary Fig. 9a-c; Supplementary Note 3). Also, the interlinked triplex bandage of chromosome X *q*-arm fused to chromosome 10 *q*-arm is present on the DeBruijn graph of the unphased assembly (Supplementary Fig. 4a), but not upon downsampling of Verkko with HiFi and ONT-UL over 100 kb alone (Supplementary Note 4b). We believe this is the first report of a major long-range chromosomal translocation automatically called from a genome assembler on a diploid human genome.

Ongoing chromosomal rearrangements in laboratory cell lines can represent a challenge in defining a consensus reference. To confirm that RPE-1 cells are a karyotypically stable diploid cell line, we cytogenetically analyzed variations across the cell population, in different batches and during passages in culture. We used Chromomycin A3 staining to perform R-banding that allows us to identify each chromosome thanks to their specific banding pattern (Supplementary Fig. 8a, Methods). As expected, we did not observe overt karyotypic variations at the cytogenetic level, including stable presence of the t(Xq;10q) marker chromosome in all batches in this study (Supplementary Fig. 8e) and across our previous work using this cell line [30,32]. RPE-1 showed relatively stable ploidy ~46 chromosomes (Supplementary Fig. 9a), with no evident gross chromosomal rearrangements (GCRs) or other cytologically visible chromosomal changes through passages and comparing batches of RPE-1 acquired from different sources (Supplementary Fig. 9b-d). In more than 100 metaphases, we saw no evidence of tetraploid or pseudo-tetraploid karyotypes, implying that RPE-1 cells have a stable diploid asset maintained in culture (see “Discussion”). Altogether, the resulting final RPE-1 diploid genome (RPE1v1.0) represents a new haplotype-resolved ultrahigh quality reference with faithful representation of a major SV (see “Comparative characterization of RPE-1 inter-genomes and inter-haplotypes variation”).

## Comparative characterization of RPE-1 inter-genomes and inter-haplotypes variation

Recent analyses of a completely phased diploid genome suggest that differences between the two haplotypes of the same individual carry variation within coding regions [2] that increase in polymorphic loci like centromeres which limit meiotic crossover [52,53]. To establish RPE-1 specific variation, both compared to other genomes and between haplotypes, we characterized the specific genomic make-up of RPE-1 for each phased haplotype (Fig. 2a). First, we wanted to define the haplotype with the translocated X and map the precise breakpoint events in the

t(Xq;10q) at base pair resolution. A previous attempt determining chromosomal haplotypes inferred by bulk DNA sequencing reported high likelihood of the RPE-1 translocated segments belonging to the same haplotype [51]. Having the complete RPE-1 assembly, we instead developed a multi-step pipeline for the manual curation of the SV starting from the fully phased haplotypes (Fig. 2b; Supplementary Fig. 10a-c). With this approach (Supplementary Note 4, Methods), we identified a 73.18 Mb segmental duplication of chromosome 10 *q*-arm most likely to belong to haplotype 2 due to read alignment score, fused to the telomeric region of haplotype 1's chromosome X. Validation of the breakpoint also identified an ensuing deletion of 3603 bp mapping proximally to the X's telomere associated with the chromosomal fusion (Fig. 2b, Supplementary Note 4). The RPE-1v1.0 final release includes the sequences of the haplotypespecific SV at base pair resolution.

Next, we assessed genome variation between the two RPE-1 haplotypes and between each haplotype and the CHM13 to evaluate inter-genomic differences using SyRI [54]. The output was used firstly to evaluate and visualize the distribution of SVs among the genome. As expected, we observed that RPE-1 and CHM13 are syntenic throughout most of the genome with 60% of wholegenome variants falling within repetitive and polymorphic loci (Supplementary Fig. 11a). Between RPE-1 haplotypes, we found high true-positive variation rates including a total of 62 inversions, 1646 translocations and 2565 highly divergent regions (HDRs) (Supplementary Fig. 11b, Table 2), with the greatest differences occurring in the centromeric and pericentromeric regions (~40%) (Fig. 2c, Supplementary Fig. 11c-d). Notably, both RPE-1 haplotypes carried a similar number of SVs when compared against CHM13: 50 inversions, 876 translocations and 2589 HDRs for RPE1 Hap1, and 76 inversions 1602 translocations and 2636 HDRs for Hap 2 (Supplementary Fig. 11e, Table 2). HDRs associated with syntenic regions corresponded to the copy gain duplication from chromosome 10 and the DNA sequence that belongs to the Major Histocompatibility Complex (MHC), one of the most gene-dense and polymorphic stretches of human DNA [55]. Since the greatest differences between genomes and haplotypes were found to occur in the centromeric and pericentromeric regions (Fig. 2c, Supplementary Fig. 11b, Table 2), we focused our study on these genomic loci.

Having resolved a diploid genome gave us an opportunity to uncover intra-individual variation between the two haplotypes of RPE-1 (inter-haplotype differences, Fig. 2a). Furthermore, our diploid genome contains completely resolved centromeres that we localized along all chromosomes *via* HumAS-HMMER and using RepeatMasker as a comparison [56] (Methods), allowing us to comprehensively study their inter-haplotypes variability for the first time. Centromere alpha-satellite arrays in RPE-1 range from 0.8 (chromosome 21, haplotype 2) to 8 Mb (chromosome 7, haplotype 1) (Fig. 2d-e). Given previously reported recombination suppression in these regions [52], we anticipated observing substantial variation between centromeres of homologous chromosomes and likely, haplotypes. In line with this, RPE-1 centromeres of the same chromosome pairs show large differences in their size (Fig. 2d-e) and HOR structures inferred by pairwise sequence identity heatmaps (Fig. 2d) using StainedGlass (v6.7.0) [38] (Methods). Indeed, we found that centromeres of all chromosomes show a difference between the two haplotypes, with 39% of them showing different HOR organization. Moreover, for the majority of the chromosomes we observed changes in their centromere size and position (Fig. 2d-e). We observed no or only modest shifts in the centromere position for most

chromosomes (~300 kb for chromosomes 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 20 and X), a proxy that implies complete assembly of the chromosome *p-arm* and annexed telomere upstream of the centromere locus. However, some centromeres changed localization between haplotypes (with up to 2 Mb for chromosome 4, 8, 15, 16, 18, 21, and 22) (Supplementary Fig. 12), yet this may be due to inability to annotate centromeres and/or reliably anchor the start of the centromere with respect to the rest of the chromosome. The complete diploid genome currently available, HG002, showed similar inter-haplotype variation for all chromosomes (data not shown), indicating that our finding may be more widely applicable to human centromere polymorphism in diploid genomes in spite of the caveat that centromeric regions alignment/annotation potentially affecting their comparison. We also assessed inversions within the centromeres comparing the two haplotypes [40]. The large centromeric inversion on chromosome 1 of CHM13 [22] was not present to the same extent in either RPE-1 haplotypes (Supplementary Fig. 12a) or in HG002 maternal and paternal haplotypes [2], while the rest of the centromere and flanking pericentromere structure remains similar across genomes (RPE-1, HG002, CHM13) and haplotypes (Supplementary Fig. 12b), suggesting that CHM13 chromosome 1 inversion may be a rare polymorphism or specific to that line. However, we found other such haplotype-specific inversions within the centromeres of RPE-1 and HG002, for chromosomes 5 and 6 for the maternal genome and for chromosome 3 of haplotype 2. Next, we evaluated RPE1 inter-haplotype variation by calculating the ratio of centromere sizes. We found a distribution between 0.5- and 1.5-fold differences, with chromosomes 7, 9 and 11 having the greatest size divergence between haplotypes (Fig. 2e). Interestingly, the centromere of chromosomes 11 and 21 show more than 1.5 increase in size when comparing the RPE-1 haplotypes, suggesting that they may carry higher divergence or be more challenging to assemble and validate (Fig. 2e). This finding also begs the question on the biological impact of size divergence. For chromosome 21, this has been associated with ensuing chromosome missegregation which during embryonic development can lead to Down's Syndrome [57]; with chromosome 7, biological relevance remains to be established. Interestingly, 40% of centromeres show a completely different structure between haplotypes, with particular focus on chromosomes 6, 4, 11 and X (Fig. 2d). Chromosome 6 of haplotype 1 showed two highly homozygous regions compared to haplotype 2 (Fig. 2d), an inter-haplotypes structural divergence we also observed for chromosome 22 and X (Fig. 2d). These observations indicate that structurally, the centromere of one haplotype may be anchored within a single active HORs, while the other has two due to a possible recent expansion for one haplotype on this genome as showed previusly [22]. This is reminiscent of the metastable epialleles for chromosome 17 [58]. Altogether, our data present evidence and quantify the extent of the polymorphism between genomes, also reflected in haplotype variation, for all RPE-1 centromeres.

**Alignment quality improves with RPE-1 genome compared to non-matched references**

The haplotype variation and divergence between genomes that we found using the RPE-1 reference is case-in-point to support the need for matching the reads generated from multi-omics experiments with the reference genome used to align them. Non-aligned content to widely-used human reference HG38 has been shown in all human genomes analyzed [59]. While for small genomes it has been possible to study genomic variation by whole-genomes comparison [60,61], for large human genomes, matched sequence-reference has been previously considered but remained largely unexplored [62]. Recent efforts addressed reference bias and its negative impact on read mapping by increasing representation of individuals' genomes variation [3,18,63]. We reasoned that alignment and analyses of multi-omics data may not always be successfully supported using a non-matched single reference, or a collection of them, particularly for regions that show considerable variation (Fig. 2). To validate the improvement in read alignment quality, we considered two main parameters, 1) mapping quality and 2) edit distance to the reference, defined as the number of changes necessary to make an aligned read equal the reference, excluding clipping (Supplementary Note 4). To this aim, we tested the alignment quality of short and long reads against the RPE-1 genome, using the complete CHM13 assembly as control (Fig. 3a). The alignment was performed using BWA and Minimap2 for both genomes (Methods). Because CHM13 is a haploid genome, we chose a different approach to show not only the differences between completely matched and non-matched genomes, but also to highlight the different alignment between diploid and haploid genomes. We used a complete diploid RPE-1 genome with both haplotypes. With mapping filters, only 16.25% of the alignment file was retained that contains information from polymorphic loci of RPE-1 genome (Supplementary Fig. 13a). We proceeded to align ~46x coverage HiFi reads with 99.99% base calling confidence for both RPE1 haplotypes and CHM13 (Fig. 3a-b) without filters, in order to avoid losing any information. Notably, isogenomic alignment showed a reduction in secondary alignments and mismatches for all chromosomes. different mapping for each chromosome. The total coverage of HiFi reads is equally distributed between each RPE-1 haplotype (20x) but secondary alignment and multimapping was much higher mapping to the CHM13 genome, further increasing within repetitive regions (Fig. 3b, Supplementary Fig. 13b). We observed no drop in coverage on both haplotypes compared to the expected 20x coverage, as opposed to CHM13 that showed relatively good alignment coverage but with higher mismatched bases (Fig. 3c-d). In addition to improvements in quality and fidelity of alignment, the isogenomic RPE-1 reference was tested for chromosome 6 and chromosome X that presented the highest diversity in the centromere structure between these two cell lines (Fig. 2d). When compared in read alignment quality, this genomic polymorphism led to a missing alignment for CHM13, while a complete non-drop alignment for RPE-1 was obtained (Fig. 3b). These data point to an improvement in coverage, quality, fidelity and, consequenty, a reduction in incorrectly aligned reads upon use of the isogenomic reference for sequencing alignment. Even an as complete reference genome as CHM13 failed to support fidelity of alignment at the same extent for whole-genome – and faired poorly for polymorphic regions (Fig. 3b). The improvement is especially significant at polymorphic loci and centromeres where inter-genomic diversity interferes with the capacity of reads to find a high-fidelity match within a reference genome (HDR; Fig. 3e-h). To confirm and validate these improvements, we used whole-genome sequencing (WGS) reads generated with a different batch of RPE-1 cells. We tested both Illumina short reads and HiFi long reads. We decided to use

Illumina as short reads allow an even lower number of mismatches. Here, we decided to use the three haplotypes separately to evaluate the mapping quality and mismatches distribution, making a complete comparable comparison with the CHM13 haploid genome, and allowed Illumina and HiFi to map independently against the RPE-1 Hap1, Hap2, CHM13. The differences in distribution of our parameters clearly shows improvements with isogenomic reference, using either Hap1 or Hap2 of RPE-1, versus CHM13 reference but the trends were not statistically significant for the whole-genome alignments (Fig. 3, Supplementary Fig. 13c). Hence, we decided to evaluate the differences for these parameters extracting HDRs regions that we found with SyRI, and comparing haplotypes to each other. Within the HDR, mapping quality and mismatches significantly changed when we used RPE1 instead of CHM13, with an increase of 3.5 for quality mapping (Fig. 3e-f) and decrease of 2.1 for mismatched bases (Fig. 3g-h). This was true for all chromosomes analyzed (Fig. 3f) and for both HiFi and Illumina reads (Fig. 3g-h). In particular, the HDR regions that were identified between Hap2 and CHM13 showed a complete loss in proper mapping for CHM13 and correlated with higher edit distances (Supplementary Fig. 13d). Hap2 showed the same trend also with short reads alignment, confirming the previous data obtained with the syntenic region founder tool (Supplementary Fig. 13e, Methods). Altogether, we proved that the reference bias decreases while unique position mapping increases with a proper reference genome highlighting the importance of isogenomic mapping, which represents a critical step for any omics experiment to generate the alignment file (Fig. 3a). The same level of accuracy is not supported by the latest complete human reference genome CHM13, in spite being the most complete and validated genome available to date. Particularly, having a near non-mismatched alignment represents an innovative resource in SVs analysis, avoiding variant overestimation, and a crucial point for correct phasing evaluation in protein DNA interaction experiments [64].

**Publicly available experimental dataset re-analyzed using isogenomic reference change the epigenetic landscape of human centromeres**

To verify whether the isogenomic reference genome can improve the analysis of epigenetic data in highly polymorphic regions, we used publicly available CUT&RUN dataset from RPE-1 cells and we compared the results by using different haploid reference genomes: our RPE-1 haplotypes (RPE1v1.0 Hap1 and Hap2), CHM13 (T2T-CHM13v2.0), HG38 (GRCh38.p14) and the HG002 paternal and maternal haplotypes. Due to displaying the highest level of polymorphism across genomes and between haplotypes (Fig. 2) and high quality in alignment with the isogenomic reference (Fig. 3), we focused on the analysis of centromeric chromatin enrichment and spread of the centromere-specific H3 histone variant, CENP-A. Recent data have shown differential enrichment loci during the cell cycle [65], changes under different treatment conditions [31] and, more recently, shifts in the region of CENP-A enrichment called through non-isogenomic alignment [66]. Based on our evidence, we used the RPE-1 reference re-assess and improve multiomics analyses of human centromeres aligning a CUT&RUN publicly available dataset (Methods) performed on RPE-1 cells (Fig. 4a). Since the epigenetic centromere identity is defined and maintained by the presence of the histone variant CENP-A, we used experiments conducted on RPE-1 to explore the power of isogenomic referencing for the investigation of the structure and organization of the centromeres at high level of precision. Mapping of CENP-A-bound reads show dramatic differences in occupancy and spread at nearly all centromeres between the RPE-1

haplotypes and HG38 (Fig. 4b-c, Supplementary Fig. 14a-d). In line with lack of complete annotation, HG38 shows incomplete peak calling on chromosomes 1, 9 and 16 which all have a large array of HSat2 and HSat3 [1], likely due to incomplete centromeres' organization and annotation in the HG38 genome which impacts the ability to distinguish between centromeric and pericentromeric regions. This is one of the main reasons why centromeres have been historically omitted from such analyses. Consequently, we focused on the most recent and complete human reference, CHM13, which has been used to align centromere reads and study the epigenetic landscape of these loci [66] and contain complete linear sequence for all centromeres [1]. The use of isogenomic reference can also highlight variation when comparing the peaks of the centromere marker proteins within the RPE-1 phased haplotypes and the CHM13 reference genome, thus we extended our comparison to the phased diploid assembly of HG002 [2]. First, we decided to investigate the structure of the active centromere by monitoring the presence and the spread of highly significant CENP-A peaks (q-value <0.00001) (Fig. 4b, Methods). We also investigated the kinetochore's positioning through the co-localization of the high-confidence CENP-A peaks (MAPQ >20, q-value <0.00001) and cytosine methylation (5mC) positioning to identify the hypomethylated CDR (Methods). Highly significant CENP-A peaks were used to highlight variations between all reference genomes, whilst the DNA methylation variation and high confidence CENP-A peaks were assessable only across RPE-1 haplotypes. The centromere of chromosome 2 shows the lowest variation in kinetochore position between the two RPE-1 haplotypes and, interestingly, exhibits the lowest variation in size for the CDR (Fig. 4b). Moreover, chromosome 2 displayed the highest conservation of CENP-A mapping among the reference genomes, and this is furtherly reflected by the high conservation in size of the active centromere (Fig. 4b). On the contrary, chromosome 7 presented the highest, near-2-fold expansion, in CENPA occupancy both in regards to RPE-1 haplotypes and to HG002 haplotypes, highlighting an underlying variation at the sequence level that in turn, changes the epigenetic landscape of the centromere (Fig. 4b, Supplementary Fig. 14e). Surprisingly, when assessing the CDRs, we noticed that these differences in chromosome 7 are actually not reflected in the kinetochore's size of the RPE-1 haplotypes (Fig. 4b). The centromeres of chromosome X and 7 both exhibited an interesting 440-500kb respective range of distance between the CDRs of the two haplotypes (Fig. 4b). One can appreciate the power of the isogenomic referencing especially when focusing on the reliable detection and calling of the CDRs which would have been missed otherwise because it falls within a region which would have been hidden when mapping against the other assemblies. Also, the absence of the underlying sequence and structural expansion in other haplotypes would have made it impossible to identify this enrichment without matched-reference. Importantly, the alignment between the high-confidence CENP-A peak and the CDR would be shifted using a reference genome in all cases observed (Fig. 4b-c). The precision of mapping also enabled us to call enrichment spot for CENP-A identifying the site of the kinetochore to be usually preferentially localized toward the left or right boundary of the CDR (Fig. 4b), implying a structural organization that favors chromosome segregation (see "Discussion"). Interestingly, our reference showed that location of the CENP-A high-confidence peaks change in position within the CDR, while generally retaining the same span (Fig. 4b-c, Supplementary Fig. 14e-g). Altogether, we have highlighted a significant variation displayed between human cell lines when aligning against the different available reference genomes that affects our interpretation of the underlying biological phenomena. In addition, the isogenomic reference genome used for the alignment within repetitive DNA reduced the requirements for multimapping strategies as the genome contains a

closer and more comparable copy number with the repetitive array. Copy number variations are more easily normalized and interpreted, as well as reads depth and coverage, using the isogenomic reference genome. Our method facilitates the characterization of more faithful genomic coordinates and matched sequence variation between the sequencing data and the reference that allowed all reads to be correctly and precisely assigned.

## DISCUSSION

Here we present the first, to our knowledge, human reference genome of a diploid experimentallyrelevant laboratory cell line, including fully-phased haplotypes spanning the centromeres, RPE1v1.0. We demonstrate how the use of this *isogenomic* reference genome better supports accurate multi-omics analyses from experiments done in the RPE-1 cell line, especially for regions with the highest diversity, compared to other reference genomes. We coined the term
'*isogenomic*' in line with the words 'isogenic' or 'isogenetic' to refer to the same genetic make-up across the whole genome. The isogenomic reference improves alignment quality and faithful peak calling, and enables to dissect how sequence changes, between haplotypes and across genomes of different individuals, influence the resulting epigenomic landscape. We leveraged our diploid genome with phased haplotypes to assess four differences that we observed within human centromeres: (1) in sequence; (2) in size; (3) in structure, and (4) in position. In line with rapid emergence of new alpha-satellite HOR structure that underlies divergence in centromeres between genomes [27], we found differences in one or more of these 4 parameters also between haplotypes for all chromosomes. The change in position of the centromere however may not reflect a biologically-relevant shift in kinetochore location along the chromosomes but reflect the current inability to faithful align centromeres from different haplotypes. For acrocentric chromosomes, for instance, we found >1 Mb shift between haplotypes in RPE-1 (and when compared to CHM13 and HG002 haplotypes). While this could be a byproduct of misassembly and/or divergence along the petit arm before approaching the centromere, it is tempting to speculate that copy number variation in the ribosomal DNA (rDNA) specific to that genome/haplotype is actually causing these misalignments. In spite of the positioning caveat for some chromosomes we and others found [27], we were able to show striking difference in size, structure and sequence of active alpha-satellite HOR arrays to various extent for all chromosomes, which in turn leads to substantial changes in the epigenetic enrichment of proteins onto the centromeric DNA. Essentially, RPE-1 reads from centromeric chromatin map differently depending on the reference genome used without necessarily reflecting a *bona fide* biological change. This is noteworthy considering the abundant literature showing epigenetic datasets calling CENP-A localization [30,31,65–69], and other centromeric proteins [32,70,71], on reference genomes. Historically, reference genomes did not contain linear annotation of centromeres. Even more recently, widely used reference assemblies such as HG38, were incomplete or misassembled across the primary constriction spanning the alpha-satellite HOR arrays and the flanking divergent pericentromeric repeats. Here, we chose to compare the improvements of isogenomic mapping not only with HG38, but also the latest version of the most complete human assemblies available, CHM13 and HG002 maternal and paternal haplotypes, respectively offering a haploid and a diploid comparison. CHM13 [1,2,45] gave more consistent and precise mapping

compared to HG38 due to complete centromeres, yet using RPE-1v1.0 reference for RPE-1 reads substantially improved peaks positioning, size and enrichment confidence compared to all other genomes. It also gave a phased epigenetic landscape that adds valuable information to interpret the findings in a diploid state. Thus, our work addresses the reference bias [18] empirically demonstrating how sequence changes negatively impact the ability to define, size and position of the active centromeric chromatin and, in turn, precisely locate kinetochore sites (marked by concomitant hypomethylation with CENP-A enrichment). To this end, we find a convincing overlap between methylation dip in the CDR and high confidence CENP-A peaks only using the isogenomic reference RPE-1v1.0 to align RPE-1 reads from chromatin immunoprecipitation experiments but not using any other reference genome – however complete. Reanalyzing isogenomically previously published datasets [66], we found variation in kinetochore position between haplotypes in all chromosomes. These remarkable inter-haplotype differences are in line with recent observations of high kinetochore plasticity between genomes [27]. We determined the enrichment ratio for chromosomes 2, 7 and X which shows higher, equal or lower CENP-A levels between haplotypes, underscoring this heterogeneity. Interestingly, the CDR ratio showed lower levels of variation between haplotypes, suggesting a minimal functional size required. Because of the precision in our mapping, we observed a trend for CENP-A (high confidence peaks) to be positioned proximal to the beginning or end of the CDR, often different between the two haplotypes of the same chromosome. CENP-A has been previously shown to be peppered along the active HORs, and its density increases from an estimate of 1 every 20 to every 4 canonical nucleosomes. Our data imply that the such CENP-A enrichment positioned at the CDR boundary – left or right, marks the exact sites of kinetochore attachment for that haplotype of each specific chromosome. Having this level of haplotypes-resolved sequence-specific precision will transform technologies like genome editing and chromosome-specific aneuploidy [72].

The dramatic acceleration of the human genomic field with the emergence of long read sequencing technologies, along with improved computational tools for assembly, base calling, scaffolding, haplotype phasing, and the concerted efforts of several genome assembly consortia (i.e. HPRC, T2T, HGSVC) paved the way to make genome assemblies accessible [16]. Our achievement in generating a complete diploid assembly by a single research laboratory underscores a change in time, where it will soon be possible, rapid and progressively more affordable to assemble the right reference for accurate analysis of sequencing data. Only five years ago, the enduring quality of HG38 assembly patchwork was celebrated as valuable references [73]; today, we are in a position to avoid the reference biases due to sequence variation [74]. Thus, while the human Pangenome draft [3] represents an important step toward a more equitable and fair representation of human genomic diversity [3], our work expands from the idea of representing as many human variants as possible toward the concept of aligning to a "personalized" reference sequence that already incorporates the relevant individual's variants.
This has been first proposed for RNA sequencing data in a 2012 workshop by Yuan & Qin [62]. Here, we applied the idea to align reads "isogenomically" to the most similar sequence for any individual or DNA sample to determine genetic and epigenetic landscape with accuracy in experimentally-relevant laboratory cell line, RPE-1. We selected RPE-1, because, besides being widely used with over 2000 publications [43], their stable karyotype offers a reference genome that can in principle support experimental data generated in any laboratory. Laboratory cell lines have been thought to be subjected to and tolerate higher level of genomic rearrangements due to

prolonged exposure to genotoxicity during culturing conditions *ex vivo* compared to newly derived line from human individuals at early passages. Instead, our data suggest that they may have achieved a more stable genomic asset than previously thought. Accordingly, the hTERT-HG002 line freshly isolated from an individual displays whole genome duplication resulting in a small percentage of tetraploid subclones [2], while we did not observe tetraploidization in RPE-1 or any other karyotype variation except lower chromosomes number in a small percentage of cells. Because we only ever see a decrease, and not an increase, in ploidy, this is likely due to common rolling of chromosomes away from the metaphase plate during the experiment. The RPE-1v1.0 assembly can serve as a benchmark of RPE-1 genetic and epigenetic landscape to enable monitoring of changes under a variety of experimental conditions and over time in culture. Nonetheless, in addition to being a useful genome, the RPE-1v1.0 assembly offers a time 0 snapshot of this cell to understand whole-genome and local mutation rates in experimentallyrelevant human laboratory cell lines. Furthermore, we confirmed improvements in alignment quality and reduction in mismatches and secondary alignments using three different batches of RPE-1: RPE-1 bought from ATCC, validated RPE-1 used to generate the RPE-1v1.0 assembly, and a vial from a laboratory in a different continent [72] (Methods), compared to aligning reads to the CHM13 reference. Interestingly, our data show better alignments and lower mismatches using long sequencing reads, while with short reads the improvement was less significant, which begs the question of the scalability of its application for next generation sequencing (NGS). Thus, although we cannot rule out that all RPE-1 data generated under different culture conditions will benefit, our evidence supports the validity of the isogenomic genome for analyzing RPE-1 sequencing, and, at scale, any other cell line isogenomically.

Altogether, our study opens important opportunities pertaining to advancing our understanding of the extent of the reference bias, its influence on downstream analysis and the biological relevance of improved alignment. While further work will be necessary to clarify and improve our isogenomic referencing method for wider applications, our proof-of-concept calls for a comprehensive catalog of complete genome assemblies for commonly used cell lines for a widespread application of *isogenomic reference* genomes for faithful multi-omics analysis changing downstream analyses, interpretations and discoveries. To this end, we initiated a sister Human Pangenome Reference Consortium (HPRC), referred to the Human Pangenom“eX” Consortium, an open initiative that focuses on “eX”perimentally relevant human cell lines. Our aim is to build a comprehensive catalog of complete genome assemblies for commonly used cell lines, including diploid embryonic and induced pluripotent stem cells (ESC and iPSC), primary and cancer cell lines. Finally, integrating the human pangenome graph that represents human diversity with genome assemblies derived from historic experimentally-amenable cell will provide important information about the differences between reference genomes freshly derived from individuals and those generated from laboratory cell lines. The RPE-1 genome represents a new telomere-to-telomere human diploid reference for the scientific community that will advance genetic and epigenetic research across fields using this cell line and a first step toward “personalized” genomes – for the benefit of individuals and of epigenomic experiments alike.

## METHODS

### Cell lines

The hTERT-immortalized retinal pigment epithelial (hTERT RPE-1) cells used to generate the RPE-1v1.0 assembly were validated by MSKCC; a second batch of cells was purchased from ATCC (CRL-4000); a third batch of cells was obtained from a laboratory in a different continent [72]. RPE-1 cells are near-diploid human cells of female origin with 46 chromosomes. Cells were cultured in DMEM/F12 medium supplemented with 10% fetal bovine serum (Gibco 10270-106), 100 µg/ml streptomycin, 100 U/ml penicillin (Euroclone ECB3001D) and 2 mM glutamine (Euroclone ECB3000D). Cells were grown at 37°C in a humidified atmosphere of 5% $CO_2$, and tested negative for mycoplasma contamination.

**Metaphase spread preparation** hTERT RPE-1 cells at 70% confluence were harvested by trypsinization after 2-hour treatment with colcemid (100 ng/ml, Roche 10295892001) to arrest cells in mitosis, washed with PBS and incubated with a pre-warmed hypotonic solution (0.07 M KCl) for 30 minutes at 37°C. After centrifugation, swollen cells were fixed with a solution of methanol-acetic acid (3:1) and washed twice with the same fixative solution. Fixed cells were dropped onto clean, wet slides and air dried overnight. All centrifugation steps were performed at 1,200 rpm for 5 minutes at room temperature (RT).

### Karyotype analysis by R-banding with Chromomycin A3

To confirm the chromosomal structure of the assembly, a karyotype for hTERT RPE-1 cells was generated. Metaphase spreads from hTERT RPE-1 cells were washed with a phosphate buffer (0.07 M $NaH_2PO_4$, 0.07 M $Na_2HPO_4$, 1 mM NaCl, pH 6.8) followed by 2-hour incubation with Chromomycin A3 (0.6 mg/ml, Sigma-Aldrich C2659) at RT in a dark, moist chamber. Slides were then washed with NaCl-HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) (0.15 M NaCl, 5 mM HEPES), stained for 15 minutes in a methyl green solution (0.1 mM, Sigma-Aldrich) and washed twice with NaCl-HEPES. The antifading (Vectashield H-1000; Vector Laboratories) isopropilgallate 1:300 was added to the slides before storing them at 4°C for 3 days in a dark, moist chamber. Images were acquired using a Thunder fluorescent widefield microscope (Leica) at a 100X magnification.

**DNA extraction** hTERT RPE-1 cells at 70-80% confluence were harvested by trypsinization, washed with PBS and centrifuged at 1,000 rpm for 5 minutes at RT. Cell pellets were prepared in individual aliquots of 1.5 x $10^6$ cells, frozen in dry ice and stored at -80°C until further use. High Molecular Weight (HMW) DNA was extracted from hTERT RPE-1 dry cell pellets using the Monarch HMW DNA extraction kit for cells & blood (New England Biolabs, NEB T3050L) and for tissue (NEB T3060L), following manufacturer's instructions and introducing few technical improvements. HMW DNA was quantified by Nanodrop and Qubit with a high sensitivity kit (Thermo Scientific Q33230). Native DNA size distribution assessed using Femto Pulse (Agilent).

### Library preparation and sequencing

Sequencing data were obtained using two complementary long-read sequencing technologies for the assembly of hTERT RPE-1 cells: Pacific Biosciences (PacBio) High-Fidelity (HiFi) long reads

and Oxford Nanopore Technologies (ONT) long (>70 kb) and Ultra-Long (UL; >100 kb) reads. Additionally, Illumina and Hi-C (Arima Genomics) sequencing technologies were used. HMW DNA from 6 x $10^6$ cells was used to generate PacBio HiFi libraries with the SMRTbell express template prep kit 2.0 (PacBio 100-938-900). Size selection was performed with Megaruptor 2 (Diagenode) with default settings, and fraction sized 15-20 kb as determined by
Femto Pulse were sequenced on the Sequel IIe platform with SMRT Cells 8M (PacBio 101-389001) using the chemistry 10.0 with 2-hour pre-extension, 2-hour adaptive loading and 30-hour movie collection time, to reach a coverage of 46x in PacBio HiFi reads. Circular Consensus Sequencing (CCS) reads were obtained using SMRT-Link
(https://github.com/WenchaoLin/SMRT-Link) with default parameters.
HMW DNA from 6 x $10^6$ cells was used to generate UL-ONT libraries with the UL DNA sequencing kit (ONT SQK-ULK114) following manufacturer's instructions. Ninety µl of library was loaded in a R10.4.1 (FLO-PRO114M) flow cell and sequenced on the PromethION 24, with two nucleases washes and reloads after 24 and 48 hours of sequencing to reach a total coverage of 80x in ONT reads (average 70 kb) and of ~30x in 100 kb + ONT reads with a Q score >20. These ONT data were base-called with Guppy (v5.0.11). In parallel, HMW DNA from 6 x $10^6$ cells was used to generate additional UL-ONT libraries with the same UL DNA sequencing kit and protocol. Ninety µl of library was loaded in a R10.4.1 (FLO-PRO114M) flow cell and sequenced on the PromethION 2 Solo, with two nucleases washes and reloads after 24 and 48 hours of sequencing to reach a total coverage of 10x in ONT reads (average 19 kb) with a Q score >35. These ONT data were base-called with Guppy (v6.5.7).
HMW DNA from 1.5 x $10^6$ cells was used to generate Illumina libraries with the Nextera XT DNA Library Preparation kit (Illumina FC-131-1024) following manufacturer's instructions. These libraries were sequenced on the Illumina Nova-seq 6000 to reach a coverage of 120x in Illumina reads. Raw data were processed with Fastp (https://github.com/OpenGene/fastp) [75] to trim the adapters and FastQC (https://github.com/s-andrews/FastQC) [76] to evaluate the quality of the reads.
HMW DNA from 2 x $10^6$ cells was used to generate Hi-C libraries with the Arima High Coverage HiC+ kit (Arima Genomics A101030), and the Arima Library Prep kit v2 (Arima Genomics A303011) following manufacturer's instructions. Hi-C libraries were sequenced on the Illumina Nova-Seq 6000 to reach a coverage of ~30x in Hi-C reads. Raw data were processed with Cutadapt (https://github.com/marcelm/cutadapt) [77] to trim the adapters.

**Whole genome assembly, manual curation and phasing pipelines**

Before assembling the genome of hTERT RPE-1 cells, global genomic features such as heterozygosity, repeat content and size were evaluated with GenomeScope 2.0
(https://github.com/tbenavi1/genomescope2.0) [78] from unassembled HiFi raw sequencing reads. Two different assembly algorithms were then employed, Hifiasm (https://github.com/chhylp123/hifiasm) [12,44] and Verkko (https://github.com/marbl/verkko) [45]. The assembly was first generated with Hifiasm using HiFi data with base-calling accuracy of 99.99%. This first assembly was compared to the assembly generated with Verkko both using HiFi together with ONT UL reads only, and using HiFi and all ONT reads regardless of size, with integration of Hi-C linkage data for manual curation and complete haplotypes phasing of the human diploid genome. The VGP mapping pipeline [79] was run to map the Hi-C reads against the haplotype 1

and 2 independently and against the diploid assembly. The final diploid assembly.fasta was merged with the unassigned.reads.fasta generated by Verkko. On PretextView (https://github.com/wtsi-hpag/PretextView), the contact map derived from the last alignment was used, and the correct position was assigned for each contig. The subsequently dual manual curation was done as described in Rapid-curation-2.0 (https://github.com/Nadolina/Rapidcuration-2.0) achieving 23 chromosome-level scaffolds for each haplotype. The algorithm MashMap (https://github.com/marbl/MashMap) [80] was then used to perform a genome-to-genome alignment, with the CHM13 assembly as a reference, in order to identify each chromosome.
MUMmer (v4.x) (https://github.com/mummer4/mummer) [81] and GSAlign (https://github.com/hsinnan75/GSAlign) [82] were used to make dot plots of each chromosome and determine whether the orientation of the chromosomes in the assembly was correct.

**Evaluation of de novo genome assembly**

To assess the quality of the final RPE-1v1.0 assembly, reference-based assemblers such as Cross-stitch and MaSuRCA were purposely avoided, in order to obtain an unbiased genome evaluation completeness and quality without using a genome comparison. The quality and gene completeness were evaluated using the reference-free Merqury *k*-mer analysis tool (https://github.com/marbl/merqury) [46], BUSCO (https://github.com/WenchaoLin/BUSCO-Mod) [47] and Asmgene [48] with default parameters. The Bandage (https://github.com/rrwick/Bandage) [83] visualization on the De Bruijn graph [84] is used to represent continuous whole-chromosome contigs and unassigned reads. The quality of the assembled genome was evaluated with Gfastats (https://github.com/vgl-hub/gfastats) [85] and QUAST (https://github.com/ablab/quast) [86]. We assessed the assemblies for uniform read depth across chromosomes via IGV [87] and NucFreq (https://github.com/mrvollger/NucFreq) as well as read mapping bias.

**Identification of structural variants**

Minimap2 (https://github.com/lh3/minimap2) [48,88] performed a genome-to-genome alignment choosing the parameter -x asm5 for genomes with low sequence divergence. SyRI (https://github.com/schneebergerlab/SyRI) [54] was used to search variants between: haplotype 1 vs. haplotype 2 of the RPE1v.1.0 assembly; CHM13 vs. haplotype 1; and CHM13 vs. haplotype 2. SyRI outputs the complete information about structural variants and genomic rearrangements, such as syntenic regions, copy number variation and highly diverged regions between two or more genomes. Final visualization was obtained with Plotsr (https://github.com/schneebergerlab/plotsr) [89].

**Multi-step identification and curation of breakpoint in phased haplotypes**

To map the precise t(Xq;10q) breaking point characteristic of hTERT RPE-1 cells [43], we performed a stepwise process (1) detecting changes in read coverage for both haplotypes; (2) read alignment on the diploid genome to identify haplotypes with translocated X; (3) isolation of reads spanning the breakpoint on chromosome X, (4) manual curation of breakpoint at base pair resolution, and (5) validation of the new fasta file through reads re-alignment. HiFi and ONT read alignments were first evaluated against haplotype 1 and 2 of our *de novo* RPE-1 genome (RPE1v1.0) using Minimap2. The reads were then aligned against the diploid genome to understand the haplotype with the translocated chromosome X. After isolation of the reads to

confirm the breakpoint sequence, the final marker chromosome X with the t(Xq;10q) was manually curated. The resulting fasta file was then validated by re-alignment of the reads that span the breakpoint for >99% identity (Supplementary Note 4).

**Annotation and analysis of centromere repetitive regions**

Centromere estimation was done intersecting the outputs derived from RepeatMasker (https://github.com/rmhubley/RepeatMasker) [56] and HumAS-HMMER (https://github.com/enigene/HumAS-HMMER) to identify the position of the centromeres in the *de novo* RPE-1v1.0 assembly. RepeatMasker was used with default parameters, and HumASHMMER as described in [22]. To generate heatmaps showing the variation between centromeres, StainedGlass (v6.7.0) (https://github.com/mrvollger/StainedGlass) [90] was run with the following parameters: window=1000 and mm_f=10000.

**OMICS data analysis**

CUT&RUN reads from the publicly available dataset of hTERT RPE-1 cells (GSE132193) were assessed for quality using FastQC (v0.11.9) (https://github.com/s-andrews/FastQC). CENP-A (SRA: SRR9201843) CUT&RUN reads and WGS Input (SRA: SRR9201844) of RPE-1 cells were aligned against multiple genomes: our *de novo* RPE-1 phased haplotypes, the T2T-CHM13 whole-genome assembly v2.0, the HG38 (GRCh38.p14) and the HG002 paternal and maternal haplotypes. Bowtie2 (v2.4.4) (https://github.com/BenLangmead/bowtie2) [91] was used with the following parameters: bowtie2 --end-to-end -x [index_reference] [read1.fastq] [read2.fastq] for paired-end data. The parameter end-to-end allows to search for alignments involving all the read characters instead of performing a local alignment. The resulting alignment files were filtered using SAMtools (v1.12) (https://github.com/samtools/samtools) [92] with FLAG score 2308 to avoid secondary and supplementary mapping of the reads, hence, preventing mapping biases in highly identical regions. Significantly enriched CENP-A peaks were determined using MACS3 (v3.0.0b1) (https://github.com/macs3-project/MACS), calculating the ratio between immunoprecipitated samples and background (WGS Input) with these parameters: -f BAMPE -B -g 3.03e9 -q 0.00001. High-confidence CENP-A positions were also identified using the mapping scores (MAPQ: 20, corresponding to the probability that the correct mapping to another location is 0.01) to identify reads that aligned uniquely to low-frequency repeat variants. The package karyoploteR (https://github.com/bernatgel/karyoploteR) [93] was used to generate the CENP-A density visualization across all chromosomes. Highly significant CENP-A peaks were computed using the internal function kpPlotDensity using a 500 kbp window size and peaks called with a q-value <0.00001. For CENP-A high-confidence peaks, the window size was 500 kbp and the q-value <0.00001.

**Methylation processing with Dorado**

ONT reads were aligned to the RPE-1v1.0 reference genome with Minimap2 (v2.26) (https://github.com/lh3/minimap2) [48,88] using a *k*-mer size of 15 and a window size of 10. To discriminate the ONT reads of each haplotype, a fasta reference file of the RPE-1 genome containing the merge of the fasta sequences of the two haplotypes was first generated. The BAM files were filtered using SAMtools (v1.12) (https://github.com/samtools/) [92] with FLAG 2308 to remove secondary and supplementary alignments and with MAPQ >10 to preserve confidence alignments. To retrieve the read IDs of a specific haplotype, the BAM file was converted in BED

using BEDtools (v2.30.0) (https://github.com/arq5x/bedtools2) [94]. All the raw POD5 files were then filtered using pod5-file-format (v0.2.2) to create the haplotype-specific POD5 file. Using Dorado (v0.3.0) (https://github.com/nanoporetech/dorado), we downloaded the simplex basecalling model (dna_r10.4.1_e8.2_400bps_sup@v4.1.0) and the modified basecalling model (dna_r10.4.1_e8.2_400bps_sup@v4.2.0_5mCG_5hmCG@v2) to call the methylation profile with default parameters. Dorado outputs directly the modified base calls in the SAM/BAM files that produces. It adds the MM (base modification methylation) and ML (base modification probability) tags that represent the relative position and probability that particular canonical bases have the requested modified bases. The output obtained with Dorado was then processed using modbam2bed (v0.9.4) (https://github.com/epi2me-labs/modbam2bed) with -e -m 5mC –cpg -f 0.80 parameters to get a bedMethyl file with the annotation and the position of each single modification and its relative coverage. Considering only the 5mC filtered using a -f 0.80 parameter, the DNA methylation profile was plotted using the internal function kpPlotDensity with a 10000 kbp window size.

**DATA AVAILABILITY**

Release of the RPE-1v1.0 reference genome has been approved for public access by Geron Corp. All scripts and resources generated and/or used in this project are publicly available at (https://github.com/GiuntaLab/RPE1).

## ACKNOWLEDGEMENTS

We thank Valentina Liguori and all members of the Giunta lab, as well as Alessandro Paiardini, Beniamino Trombetta, Fulvio Cruciani at University of Rome Sapienza, for helpful discussions and insights. We are especially indebt with Luigi Faino at Sapienza for help with HPC troubleshooting, key computational insights and advice. We are thankful to T2T and HPRC members for the kind assistance, especially Sergey Koren with Verkko, Nadolina Brajuka and members of Erich Jarvis lab for HiC data integration. We acknowledge Karen Miga, Glennis Logsdon and Jennifer Gerton for sharing information and useful discussions. We are grateful to Nicolas Altemose and Altemose lab members, as well as Riccardo Paone, Isabella Baldini, Juan Caceres and Noemi Di Sabatino at Dante Labs, and Margherita de Gaspari at Area Science Park, for generous sharing of insight, space and resources. We thank Alessandra Bonito Oliva at Rockefeller University, as well as Sai Swaroop Chittoor, Giulio Marano, Veronica Marselli and Alessia Spurio, for critical reading of the manuscript and providing feedback.
All computing was possible thanks to CINECA, and the Terastat2 HPC, with important support from Umberto Ferraro Petrillo and Edoardo Bompiani of the Statistics Department, University of Rome Sapienza.
This work was possible thanks to the Italian Foundation for Cancer Research (AIRC Start-Up Grant 2020 ID. # 25189) and the Rita Levi-Montalcini program from the Italian Ministry of University and Research (MIUR) to S.G, and the PON Ph.D. Fellowship to E.V. by the MIUR, Italy.

## AUTHOR CONTRIBUTIONS

S.G. conceived the idea underlying this work; S.G. and E.V. designed the study with key input from L.C.. E.V., and D.L. generated all sequencing data, with supervision from R.O. and M.C. for PacBio data. E.V. and L.C. performed all computational analyses. E.D.T., S.G. and F.P. designed and performed cytogenetic experiments, while E.V. performed all other experiments. S.G supervised experiments and analyses with key support from E.T. and G.F.. E.D.T. and S.G. designed and assembled all figures with help from L.C., E.V. and A.G.. S.G. wrote the manuscript together with E.T., L.C., E.V., E.D.T. with key editing from G.F. and A.G.. All authors have read and approved the manuscript.

## References


1. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
2. Jarvis, E. D. *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).
3. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023). 4. Chao, K.-H., Zimin, A. V., Pertea, M. & Salzberg, S. L. The first gapless, referencequality, fully annotated genome from a Southern Han Chinese individual. *G3 Bethesda Md* **13**, jkac321 (2023).
5. Yang, C. *et al.* The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res.* **33**, 745–761 (2023).
6. He, Y. *et al.* T2T-YAO: A Telomere-to-telomere Assembled Diploid Reference Genome for Han Chinese. *Genomics Proteomics Bioinformatics* S1672-0229(23)00100–6 (2023).
7. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
8. Yang, X. *et al.* Characterization of large-scale genomic differences in the first complete human genome. *Genome Biol.* **24**, 157 (2023).
9. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
10. Jónsson, H. *et al.* Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4**, 170115 (2017).
11. Seplyarskiy, V. B. *et al.* Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science* **373**, 1030–1035 (2021).
12. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
13. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
14. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
15. Pollen, A. A., Kilik, U., Lowe, C. B. & Camp, J. G. Human-specific genetics: new tools to explore the molecular and cellular basis of human evolution. *Nat. Rev. Genet.* **24**, 687–711 (2023).
16. Eichler, E. E. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *N. Engl. J. Med.* **381**, 64–74 (2019).
17. Cancellieri, S. *et al.* Human genetic diversity alters off-target outcomes of therapeutic gene editing. *Nat. Genet.* **55**, 34–43 (2023).
18. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
19. Hayden, K. E. Human centromere genomics: now it's personal. *Chromosome Res.* **20**, 621–633 (2012).
20. Schueler, M. G. & Sullivan, B. A. Structural and functional dynamics of human centromeric chromatin. *Annu. Rev. Genomics Hum. Genet.* **7**, 301–313 (2006).

21. Willard, H. F. Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.* **37**, 524–532 (1985).
22. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
23. McKinley, K. L. & Cheeseman, I. M. The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17**, 16–29 (2016).
24. Sullivan, K. F. A solid foundation: functional specialization of centromeric chromatin. *Curr. Opin. Genet. Dev.* **11**, 182–188 (2001).
25. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
26. Langley, S. A., Miga, K. H., Karpen, G. H. & Langley, C. H. Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. *eLife* **8**, e42989 (2019).
27. Logsdon, G. A. *et al. The variation and evolution of complete human centromeres*. http://biorxiv.org/lookup/doi/10.1101/2023.05.30.542849 (2023) doi:10.1101/2023.05.30.542849.
28. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
29. Black, E. M. & Giunta, S. Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes* **9**, 615 (2018).
30. Giunta, S. *et al.* CENP-A chromatin prevents replication stress at centromeres to avoid structural aneuploidy. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2015634118 (2021).
31. Saayman, X., Graham, E., Nathan, W. J., Nussenzweig, A. & Esashi, F. Centromeres as universal hotspots of DNA breakage, driving RAD51-mediated recombination during quiescence. *Mol. Cell* **83**, 523-538.e7 (2023).
32. Giunta, S. & Funabiki, H. Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1928–1933 (2017).
33. Giunta, S. Centromere Chromosome Orientation Fluorescent in situ Hybridization (CenCO-FISH) Detects Sister Chromatid Exchange at the Centromere in Human Cells. *Bio-Protoc.*
**8**, e2792 (2018).
34. Pascarella, G. *et al.* Recombination of repeat elements generates somatic complexity in human genomes. *Cell* **185**, 3025-3040.e6 (2022).
35. Miga, K. H. Centromere studies in the era of ‘telomere-to-telomere’ genomics. *Exp. Cell Res.* **394**, 112127 (2020).
36. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using singlemolecule sequencing. *Nature* **517**, 608–611 (2015).
37. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* gr.263566.120 (2020) doi:10.1101/gr.263566.120.
38. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).

39. Logsdon, G. A. & Eichler, E. E. The Dynamic Structure and Rapid Evolution of Human Centromeric Satellite DNA. *Genes* **14**, 92 (2022).
40. Porubsky, D. *et al.* Inversion polymorphism in a complete human genome assembly. *Genome Biol.* **24**, 100 (2023).
41. Steinberg, K. M. *et al.* Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res.* **24**, 2066–2076 (2014).
42. Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
43. https://www.atcc.org/products/crl-4000. hTERT RPE-1. https://www.atcc.org/products/crl-4000.
44. Cheng, H. *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
45. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
46. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
47. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
48. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
49. Janssen, A., van der Burg, M., Szuhai, K., Kops, G. J. P. L. & Medema, R. H. Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations. *Science* **333**, 1895–1898 (2011).
50. Soto, M. *et al.* p53 Prohibits Propagation of Chromosome Segregation Errors that Produce Structural Aneuploidies. *Cell Rep.* **19**, 2423–2431 (2017).
51. Tourdot, R. W., Brunette, G. J., Pinto, R. A. & Zhang, C.-Z. Determination of complete chromosomal haplotypes by bulk DNA sequencing. *Genome Biol.* **22**, 139 (2021). 52. Choo, K. H. Why is the centromere so cold? *Genome Res.* **8**, 81–82 (1998).
53. Balzano, E. & Giunta, S. Centromeres under Pressure: Evolutionary Innovation in Conflict with Conserved Function. *Genes* **11**, 912 (2020).
54. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
55. Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annu. Rev. Genomics Hum. Genet.* **14**, 301–323 (2013). 56. Smit, Hubley and Green. https://repeatmasker.org. (2013).
57. Antonarakis, S. E. Parental origin of the extra chromosome in trisomy 21 as indicated by analysis of DNA polymorphisms. Down Syndrome Collaborative Group. *N. Engl. J. Med.* **324**, 872–876 (1991).
58. Maloney, K. A. *et al.* Functional epialleles at an endogenous human centromere. *Proc. Natl. Acad. Sci.* **109**, 13704–13709 (2012).

59. Levy-Sakin, M. *et al.* Genome maps across 26 human populations reveal populationspecific patterns of structural variation. *Nat. Commun.* **10**, 1025 (2019).
60. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27**, 2369–2376 (1999).
61. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).
62. Yuan, S. & Qin, Z. Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression. *IEEE Int. Conf. Bioinforma. Biomed. Workshop IEEE Int. Conf. Bioinforma. Biomed.* **2012**, 718–724 (2012).
63. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017).
64. Altemose, N. *et al.* DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome wide. *Nat. Methods* **19**, 711–723 (2022).
65. Nechemia-Arbely, Y. *et al.* DNA replication acts as an error correction mechanism to maintain centromere identity by restricting CENP-A to centromeres. *Nat. Cell Biol.* **21**, 743–754 (2019).
66. Mahlke, M. A., Lumerman, L., Ly, P. & Nechemia-Arbely, Y. Epigenetic centromere identity is precisely maintained through DNA replication but is uniquely specified among human cells. *Life Sci. Alliance* **6**, e202201807 (2023).
67. Nechemia-Arbely, Y. *et al.* Human centromeric CENP-A chromatin is a homotypic, octameric nucleosome at all cell cycle points. *J. Cell Biol.* **216**, 607–621 (2017).
68. Gamba, R. *et al.* Enrichment of centromeric DNA from human cells. *PLoS Genet.* **18**, e1010306 (2022).
69. Cappelletti, E. *et al.* The localization of centromere protein A is conserved among tissues. *Commun. Biol.* **6**, 963 (2023).
70. Fachinetti, D. *et al.* A two-step mechanism for epigenetic specification of centromere identity and function. *Nat. Cell Biol.* **15**, 1056–1066 (2013).
71. Thakur, J. & Henikoff, S. CENPT bridges adjacent CENPA nucleosomes on young human α-satellite dimers. *Genome Res.* **26**, 1178–1187 (2016).
72. Bosco, N. *et al.* KaryoCreate: A CRISPR-based technology to study chromosomespecific aneuploidy by targeting human centromeres. *Cell* **186**, 1985-2001.e19 (2023).
73. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
74. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
75. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
76. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

77. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
78. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
79. Larivière, D. *et al.* Scalable, accessible, and reproducible reference genome assembly and evaluation in Galaxy. 2023.06.28.546576 Preprint at https://doi.org/10.1101/2023.06.28.546576 (2023).
80. Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**, i748–i756 (2018).
81. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944 (2018).
82. Lin, H.-N. & Hsu, W.-L. GSAlign: an efficient sequence alignment tool for intra-species genomes. *BMC Genomics* **21**, 182 (2020).
83. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
84. Limasset, A., Cazaux, B., Rivals, E. & Peterlongo, P. Read mapping on de Bruijn graphs. *BMC Bioinformatics* **17**, 237 (2016).
85. Formenti, G. *et al.* Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics* **38**, 4214–4216 (2022).
86. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
87. Thorvaldsdóttir, H., Jt, R. & Jp, M. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, (2013).
88. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
89. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926 (2022).
90. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinforma. Oxf. Engl.* **38**, 2049–2051 (2022).
91. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
92. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
93. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinforma. Oxf. Engl.* **33**, 3088–3090 (2017).
94. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).

## Figure Legends

**Fig. 1. Human diploid reference genome of RPE-1 enables faithful multi-omics data analysis.** (a) Schematic overview of workflow from *in vitro* experiments to multi-omics sequencing data aligned using the Isogenomic Reference Genome (top). The advantages of this model are in 1) ability to assess the variation between haplotypes, 2) improved quality of alignment wholegenome and within polymorphic loci and 3) faithful peaks calling to map the phased epigenome (bottom). (b) RPE-1v1.0 diploid genome assembly (Haplotype 1, Hap1 and Haplotype 2, Hap2) shows comparable chromosomes length to CHM13 (haploid), HG002 (diploid, Mat and Pat), and HG38, except for the RPE-1-specific marker chromosome X (Fig. 2).

**Fig. 2. Sequence and structure variation between RPE-1 haplotypes.** (a) Study of RPE-1 haplotype variation using StainedGlass that builds heatmaps based on pairwise sequence identity as a proxy to assess the repeats structure of human centromeric HORs (Method). (b) Multi-step identification and curation of RPE-1 specific structural variant in phased haplotypes at base pair resolution (Methods); t(X;10)(q28;q21.2) originates after a translocation event where a 73.18 Mb segmental duplication of the *q*-arm of the Hap 2's chromosome 10 fused to the microdeleted telomeric region of Hap 1's chromosome X. (c) RPE-1 diploid genome representation with pairwise whole-genome comparison using SyRI highlights structural variations between haplotypes, peaking at centromeric and pericentromeric regions. (d) StainedGlass from eight chromosomes (from left to right: chromosome 2, 4, 5, and 6, top row; chromosome 11, 14, 16, and X, bottom row) showing examples of the major differences we observed between the two haplotypes: variation in centromere sequence, structure, position, and size. (e) Centromeres of the two RPE-1 haplotypes show differences in size, demonstrated also by the ratio between Hap1 and Hap2.

**Fig. 3. Reads alignment improvement consequently to an isogenomic mapping.** (a) Read alignment comparing primary alignment (black) and alternate alignment or mismatch (red) for RPE-1 Haplotype 1, RPE-1 Haplotype 2 and CHM13. (b) The HiFi read-depth profile is shown for all chromosomes, where black represents the coverage of the most frequent base pair at that position while red is the second most frequent. Using the RPE-1 haplotypes as reference improves the primary alignments compared to CHM13. (c) Reads from a second batch of RPE-1 from another laboratory show lower levels of mismatches whole-genome using RPE-1 haplotypes as the reference compared to CHM13, in all chromosomes (d). RPE-1 reads as in (c) aligned to High Diversity Regions (HDRs) show statistically significant increase in Mapping Quality for HiFi (e) and Illumina reads (f), and lower edit distance (NM Value) for HiFi (g) and Illumina (h), when using RPE-1 haplotypes as the reference with respect to using CHM13 assembly. P values are from the student t-test (ns $p > 0.05$, * $p \leq 0.05$, ** $p < 0.01$, *** $p < 0.001$, and **** $p < 0.0001$).

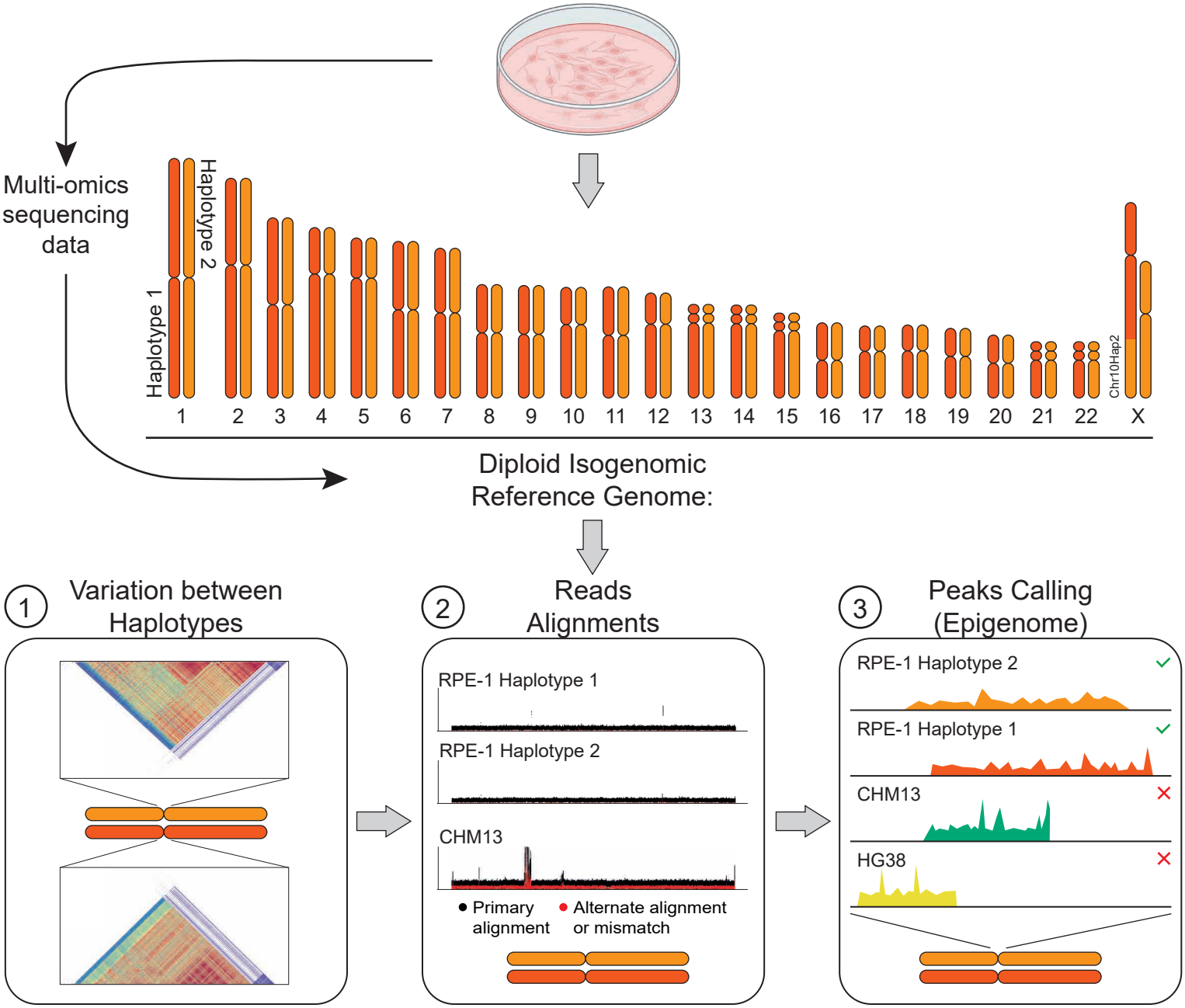
**Fig. 4. Isogenomic reference genomes improves epigenomic and methylation landscapes.** (a) Graphical representation of the workflow used for the peak calling showing accurate mapping using the phased isogenomic reference. (b) CENP-A CUT&RUN experiments enrichment on centromeres of chromosomes 2, 7, and X aligned to different genomes as reference as indicated, shows a difference in position and size of peaks. On the bottom, the chromosome ideogram showing the centromere position in CHM13 annotation. DNA methylation profile shows the

positioning of the CDRs of both haplotypes (boxes). (c) The spread of CENP-A occupancy is shown for CHM13, RPE-1 (haplotype 1 and haplotype 2), and HG38. Differences in CENP-A occupancy are visible in all chromosomes between haplotypes and when using different reference genomes. For the RPE-1 haplotypes, HumAS-HMMER annotation is shown (black rectangle).

a

# Isogenomic Reference Genome

New approach using the matched genome reference to analyze multi-omics sequencing data




b

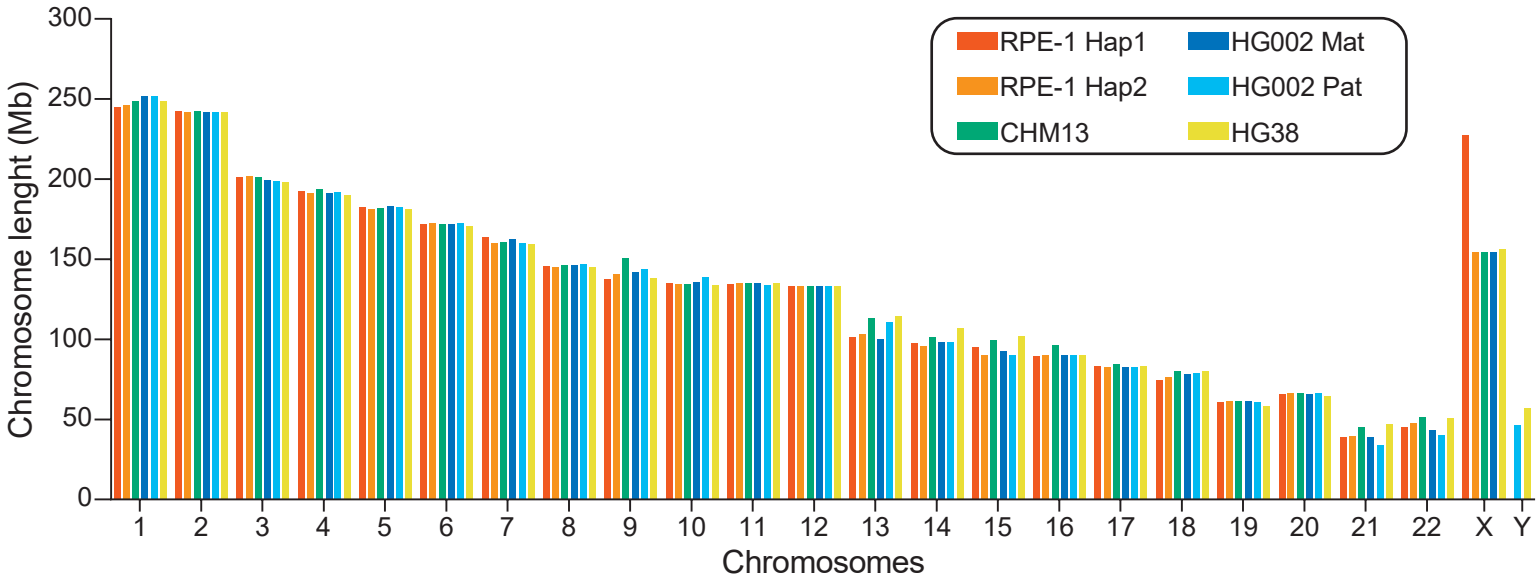



Figure 1

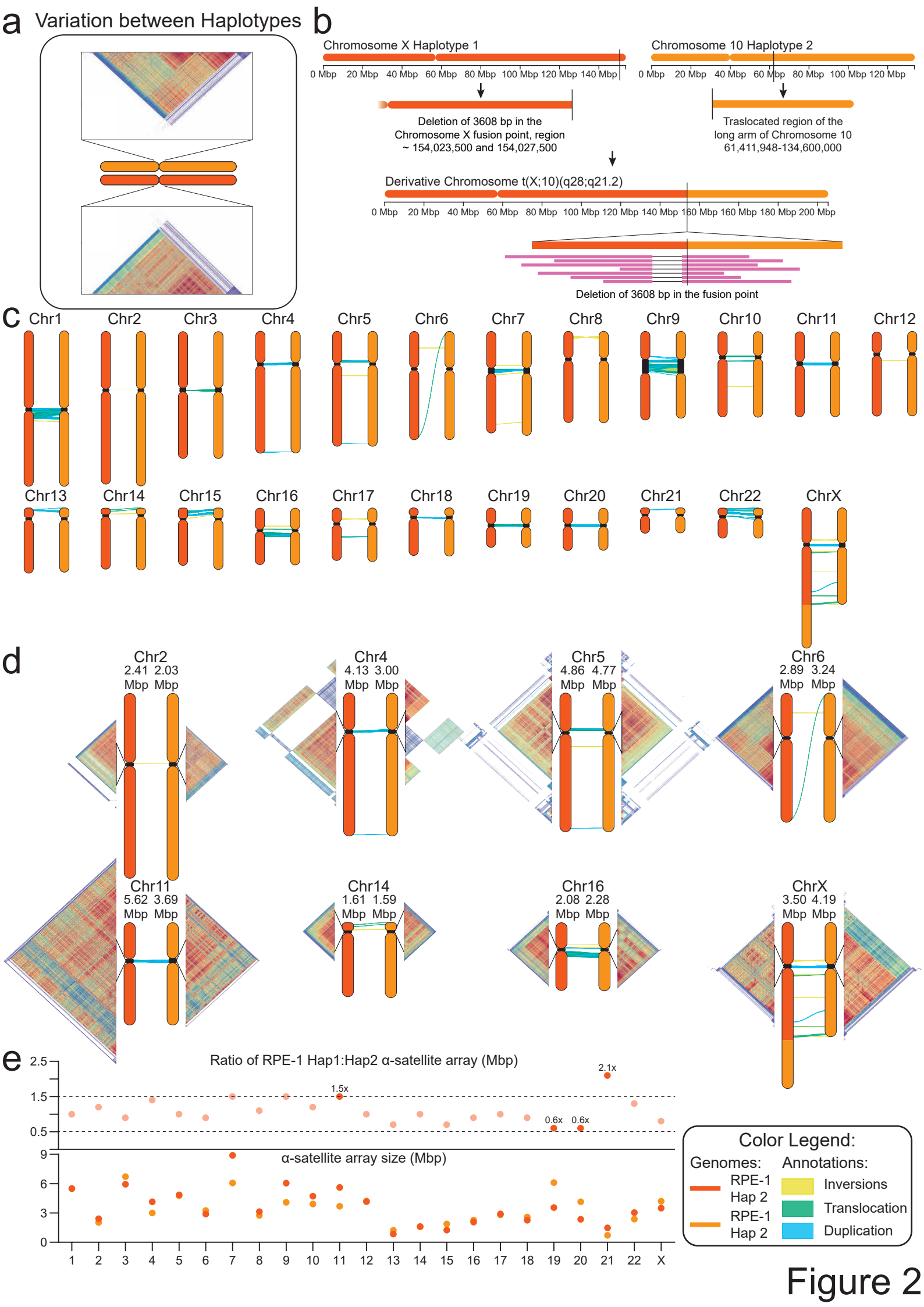

a
Variation between Haplotypes
b
Chromosome X Haplotype 1
0 Mbp 20 Mbp 40 Mbp 60 Mbp 80 Mbp 100 Mbp 120 Mbp 140 Mbp
Chromosome 10 Haplotype 2
0 Mbp 20 Mbp 40 Mbp 60 Mbp 80 Mbp 100 Mbp 120 Mbp
Deletion of 3608 bp in the Chromosome X fusion point, region ~ 154,023,500 and 154,027,500
Traslocated region of the long arm of Chromosome 10 61,411,948-134,600,000
Derivative Chromosome t(X;10)(q28;q21.2)
0 Mbp 20 Mbp 40 Mbp 60 Mbp 80 Mbp 100 Mbp 120 Mbp 140 Mbp 160 Mbp 160 Mbp 180 Mbp 200 Mbp
Deletion of 3608 bp in the fusion point
c
Chr1 Chr2 Chr3 Chr4 Chr5 Chr6 Chr7 Chr8 Chr9 Chr10 Chr11 Chr12
Chr13 Chr14 Chr15 Chr16 Chr17 Chr18 Chr19 Chr20 Chr21 Chr22 ChrX
d
Chr2 2.41 Mbp 2.03 Mbp
Chr4 4.13 Mbp 3.00 Mbp
Chr5 4.86 Mbp 4.77 Mbp
Chr6 2.89 Mbp 3.24 Mbp
Chr11 5.62 Mbp 3.69 Mbp
Chr14 1.61 Mbp 1.59 Mbp
Chr16 2.08 Mbp 2.28 Mbp
ChrX 3.50 Mbp 4.19 Mbp
e
Ratio of RPE-1 Hap1:Hap2 α-satellite array (Mbp)
2.5
1.5
0.5
1.5x
2.1x
0.6x
0.6x
α-satellite array size (Mbp)
9
6
3
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Color Legend:
Genomes:
RPE-1 Hap 2
RPE-1 Hap 2
Annotations:
Inversions
Translocation
Duplication

Figure 2

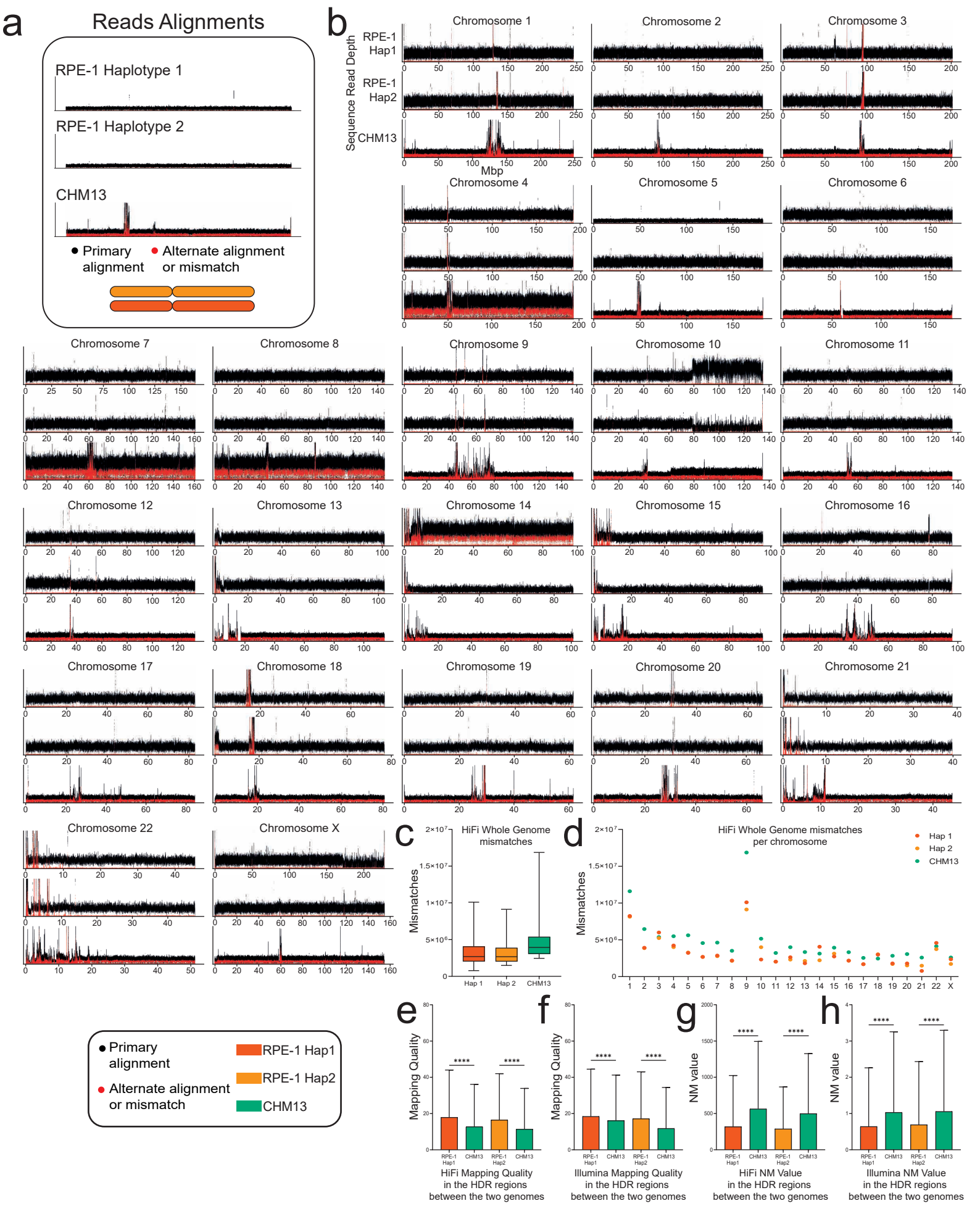

a
Reads Alignments
RPE-1 Haplotype 1
RPE-1 Haplotype 2
CHM13
Primary alignment
Alternate alignment or mismatch
b
Sequence Read Depth
RPE-1 Hap1
RPE-1 Hap2
CHM13
Mbp
Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
Chromosome 13
Chromosome 14
Chromosome 15
Chromosome 16
Chromosome 17
Chromosome 18
Chromosome 19
Chromosome 20
Chromosome 21
Chromosome 22
Chromosome X
c
HiFi Whole Genome mismatches
Mismatches
Hap 1
Hap 2
CHM13
d
HiFi Whole Genome mismatches per chromosome
Mismatches
Hap 1
Hap 2
CHM13
Primary alignment
Alternate alignment or mismatch
RPE-1 Hap1
RPE-1 Hap2
CHM13
e
Mapping Quality
HiFi Mapping Quality in the HDR regions between the two genomes
f
Mapping Quality
Illumina Mapping Quality in the HDR regions between the two genomes
g
NM value
HiFi NM Value in the HDR regions between the two genomes
h
NM value
Illumina NM Value in the HDR regions between the two genomes

Figure 3

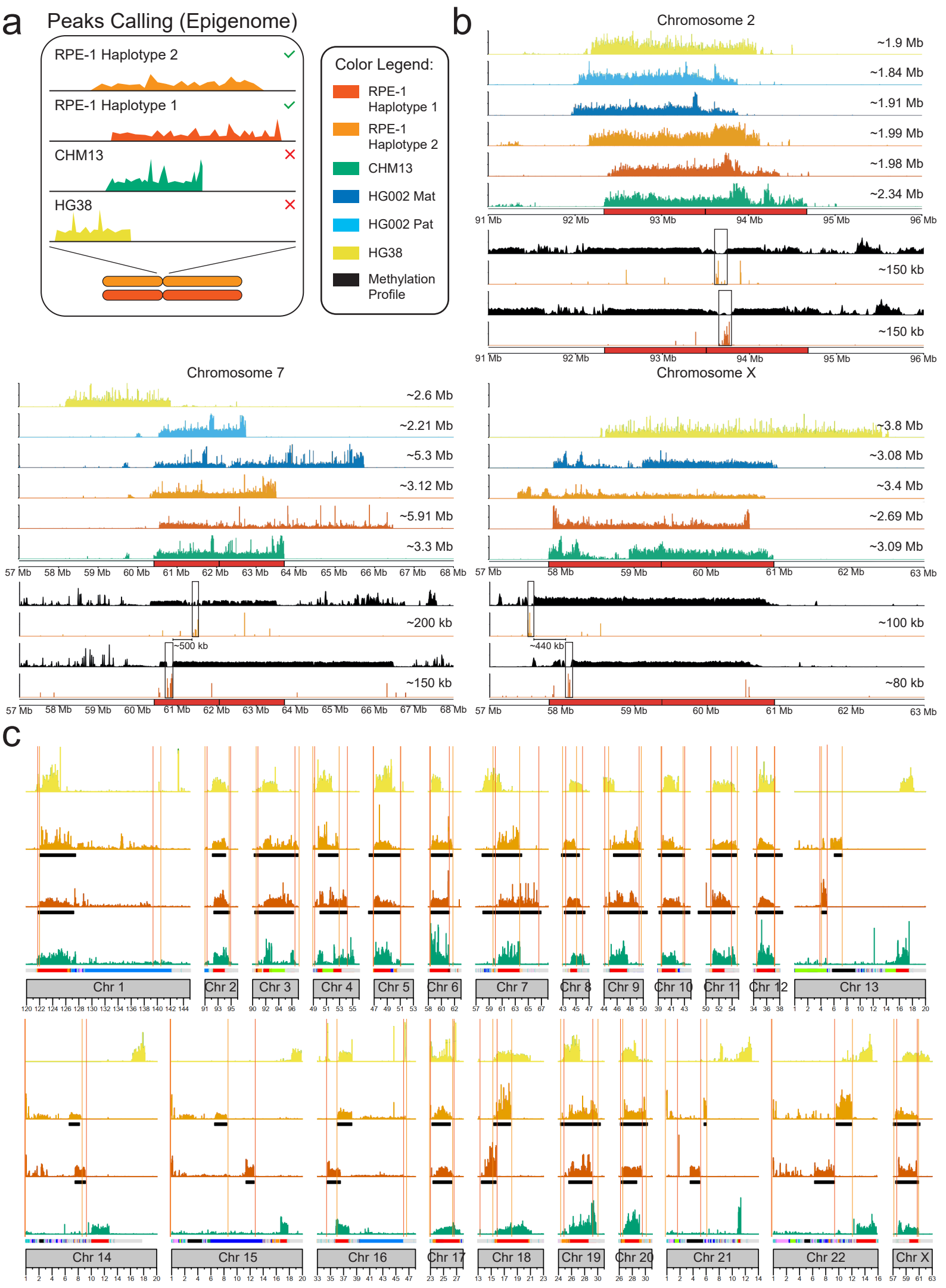

a
Peaks Calling (Epigenome)
RPE-1 Haplotype 2
RPE-1 Haplotype 1
CHM13
HG38
Color Legend:
RPE-1 Haplotype 1
RPE-1 Haplotype 2
CHM13
HG002 Mat
HG002 Pat
HG38
Methylation Profile
b
Chromosome 2
~1.9 Mb
~1.84 Mb
~1.91 Mb
~1.99 Mb
~1.98 Mb
~2.34 Mb
91 Mb 92 Mb 93 Mb 94 Mb 95 Mb 96 Mb
~150 kb
~150 kb
Chromosome 7
~2.6 Mb
~2.21 Mb
~5.3 Mb
~3.12 Mb
~5.91 Mb
~3.3 Mb
57 Mb 58 Mb 59 Mb 60 Mb 61 Mb 62 Mb 63 Mb 64 Mb 65 Mb 66 Mb 67 Mb 68 Mb
~200 kb
~500 kb
~150 kb
Chromosome X
~3.8 Mb
~3.08 Mb
~3.4 Mb
~2.69 Mb
~3.09 Mb
57 Mb 58 Mb 59 Mb 60 Mb 61 Mb 62 Mb 63 Mb
~100 kb
~440 kb
~80 kb
c
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr X

Figure 4