Isogenomic reference genomes by *de novo* assembly of experimentally-relevant human diploid laboratory cell lines

Volpe E.[1], Ottalevi R.[2], Corda L.[1], Formenti G.[3], Licastro D.[4], Tassone E.[1], Giunta S.[1]

1. Laboratory of Genome Evolution, Dept of Biology & Biotechnology Charles Darwin, Sapienza University of Rome, Rome 00185 Italy.
2. Dante Labs company l'Aquila 67100 Italy
3. Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA NY 10065 USA
4. Area Science Park Trieste 34132 Italy

ABSTRACT:

The rapid development in sequencing technologies provided us with the first haploid human reference genome without gaps, the CHM13, and a near-complete human diploid genome, the HG002 in 2022. The analysis of these assemblies have highlighted significant divergence between individuals' repetitive DNA and other highly polymorphic loci. In line with these observations, it has become apparent that omics studies must be supported by a matched genome reference, for which we coined the term '*isogenomic*'. Here, we propose a novel approach using *isogenomic* referencing for omics analyses and present a near-complete human genome assembly of the diploid Retinal Pigment Epithelial (RPE-1) cell line, commonly used in molecular and cell biology laboratories around the world. Our RPE-1 reference genome provides both a valuable resource to the wider scientific community, and also enables *isogenomic* referencing to analyze data from RNA-seq, ChIP-seq, SV calling and many other applications performed in RPE-1 cells. We show how *isogenomic* referencing reduces analyses' bias and improves previous data from mismatched genomes.

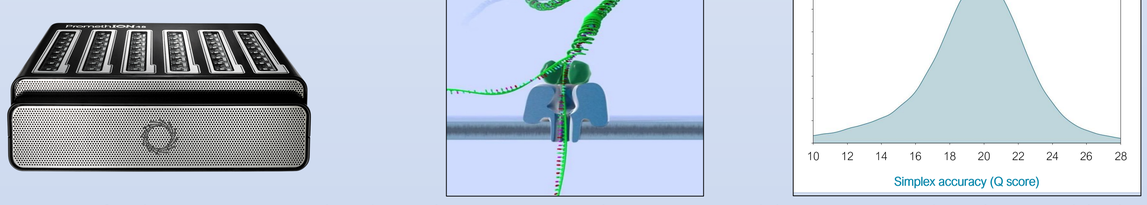


Fig.1 Oxford Nanopore technology improves read accuracy with new v14 chemistry. This increase in Qscore (20+), allowed us to make near error-free de novo diploid assembly. We used Ultra-long DNA kit to make the libraries and sequenced using PromethION.

GENOME GLOBAL FEATURES:

We used GenomeScope to evaluate the global characteristics of the RPE-1 genome, based on raw unassembled reads. This evaluation highlights total reads length, low heterozygosity level, and high level of unique k-mer from raw reads. These features are useful to select the parameters' of choice for downstream analysis given the expectation of high level of homozygosity between the partially-phased haplotypes hereby presented.

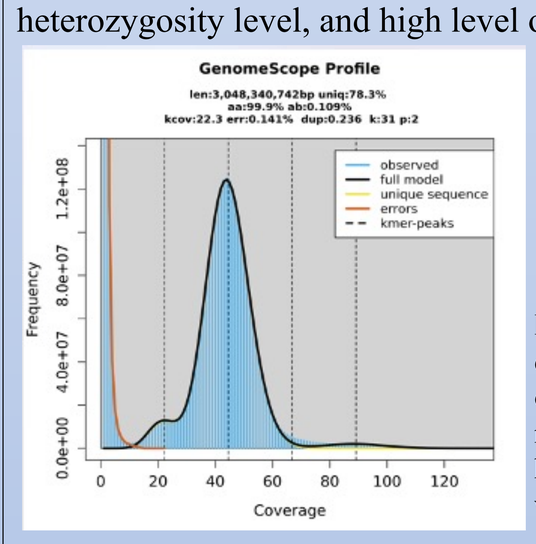


Fig. 2 K-mer graph with estimating coverage depth of raw DNA reads (x axis) and the number of times a k-mer is observed by number of k-mers with that coverage (y axis). Kcov = heterozygous coverage peak. The k-mer database was created with PacBio HiFi reads.

DE NOVO DIPLOID NEAR COMPLETE HUMAN ASSEMBLY:

We present our near complete *de novo* human genome assembly from the experimentally-relevant human diploid cell line RPE-1 **(1)**. We generated PacBio HiFi libraries with 46x coverage, and UL-ONT reads (>100 kbp) with 27x coverage. We used hybrid assembler Verkko **(2)**, combining features of Long (HiFi) and UL (ONT) reads.

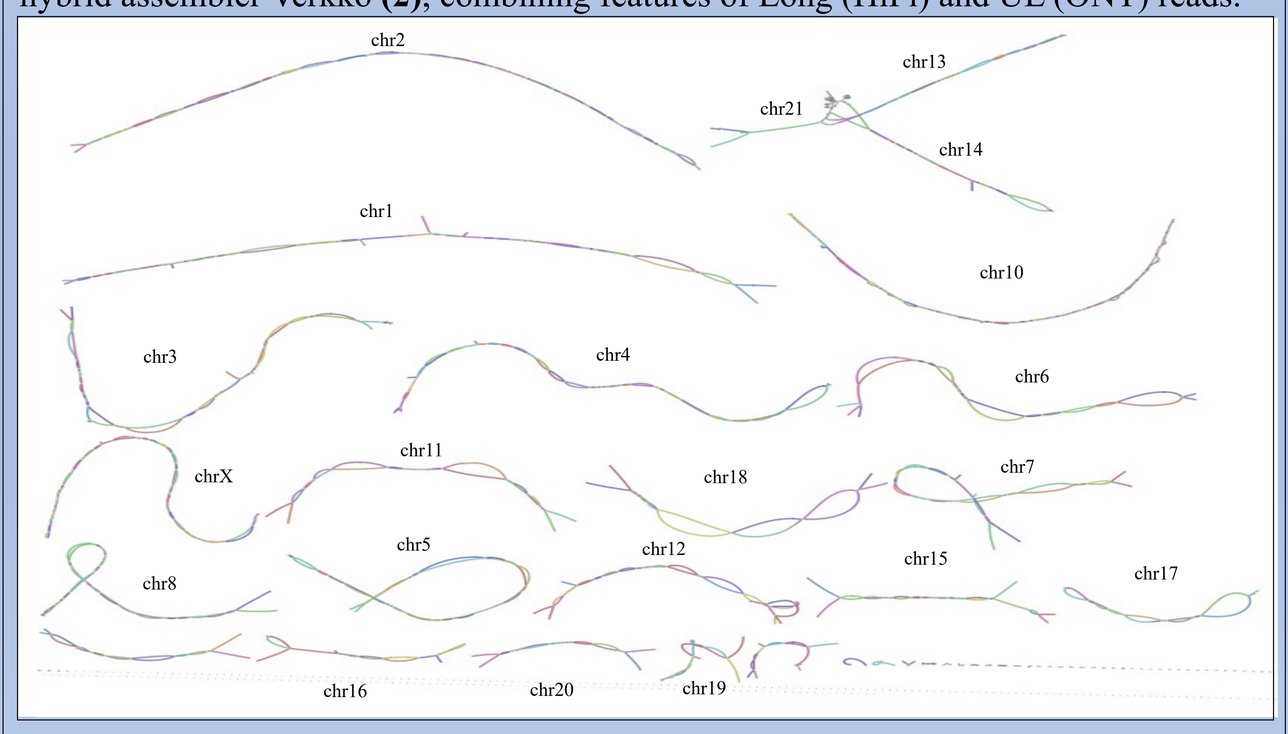


Fig. 3 Fasta file graphical representation of the *de novo* diploid assembly of RPE-1. We used Mashmap **(3)** for genome-to-genome alignment to allow us to identify chromosomes. The rDNA repetitive regions on the long arm of chromosomes Chr21, Chr13, Chr14 remain unresolved. Chr10, Chr6 and Chr4 are near-complete. We made a manual curation to identify unitigs paths.

Research in the Giunta Lab is supported by:

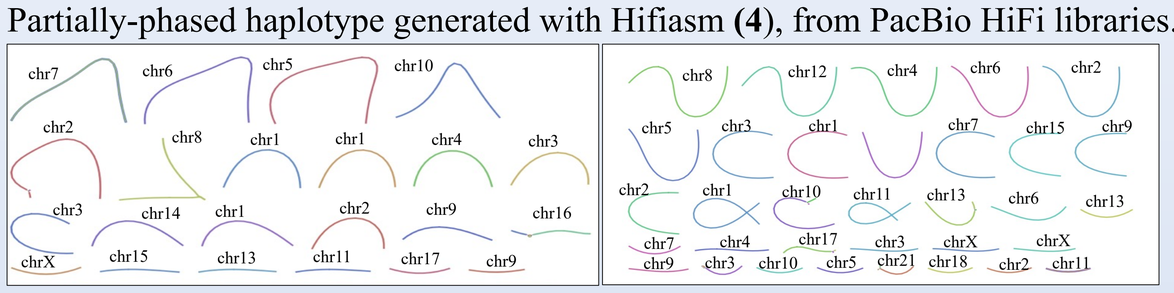
Rita Levi Montalcini Award
MINISTERO DELL' ISTRUZIONE, DELL'UNIVERSITÀ E DELLA RICERCA

AIRC

Ministero dell'Università e della Ricerca
PON RICERCA E INNOVAZIONE 2014 - 2020

DE NOVO ASSEMBLY PARTIALLY-PHASED HAPLOTYPE:

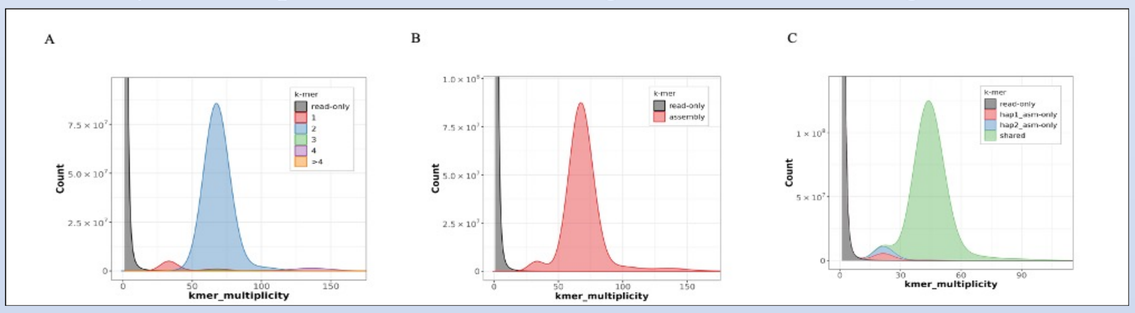
Partially-phased haplotype generated with Hifiasm **(4)**, from PacBio HiFi libraries.



Haplotype one — Haplotype two

DE NOVO ASSEMBLY QUALITY AND COMPLETENESS:

To evaluate our reference-free assembly, we compared sets of k-mers derived from unassembled reads, and quantified their presence and frequency in the *de novo* genome assembly. We also took into consideration the completeness and redundancy in the representation of human genes, estimated using BUSCO tool.



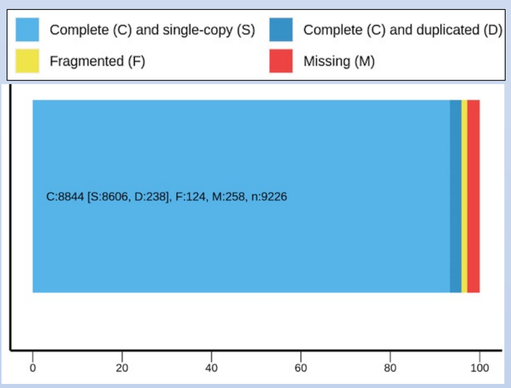
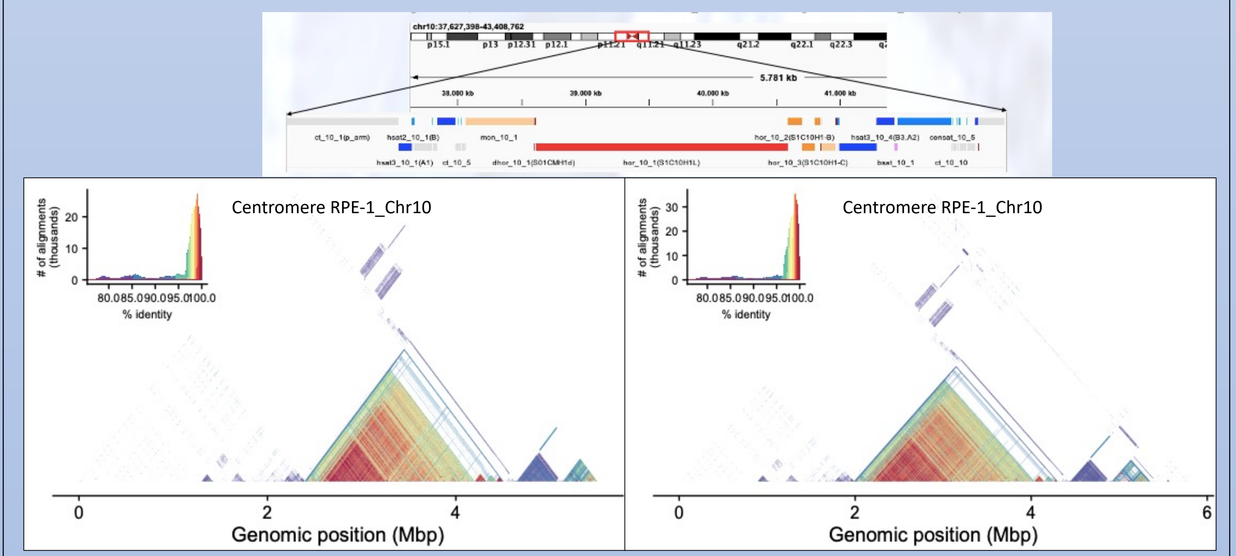


Fig 4. *De novo* assembly evaluation with Merqury **(5)** and BUSCO **(6)**. Top: K-mer multiplicity found in the reads dataset compare with the k-mer count in the assembly (A), and evaluation of k-mers completeness in the assembly (B). The first peak represents heterozygous k-mer, and the second one the homozygous sequences of the assembly. (C) Merqury graph shows the two heterozygous peaks (red and blue) and the green peak, with k-mers shared between the two haplotypes. Bottom: BUSCO graph shows genes completeness for diploid human genome produced with Verkko, for a total of 95.9% of complete BUSCO genes.

ASSEMBLY OF REPETITIVE REGIONS OF THE GENOME:

Here we present an example of near-complete chromosome10 with centromere repetitive region completely assembled but missing part of the *q* arm telomere end.



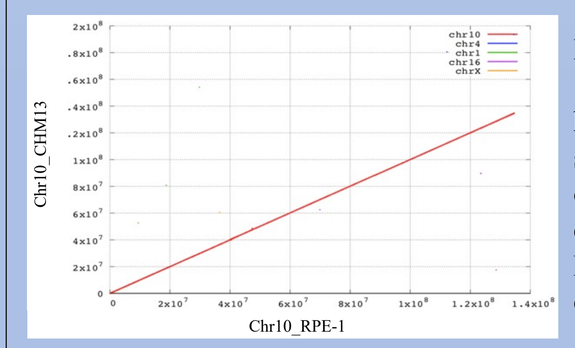


Fig. 5 Centromere organization in chromosome 10. Top: Centromere annotation. Middle: Identity heatmaps, created with Stainedglass **(7)**, which show the base composition in centromere regions of chromosome 10 in RPE-1 cell line (left) and chromosome 10 in CHM13 (right). Bottom: Dotplot of genome-to-genome alignment of chromosome 10 (x=RPE-1, y=CHM13).

STRUCTURAL VARIANTS IDENTIFICATION:

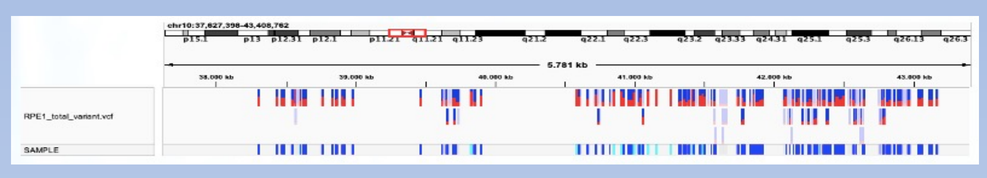


Fig. 6 Alignment of RPE-1 ONT reads to CHM13 with Winnowmap **(8)** and SVs calling with Sniffles **(9)**.The visualization of Variant Calling File on IGV shows Single Nucleotide Polymorphisms and Structural Variants in highly repetitive regions of the genome. Chr10 is shown as an example in this Figure.

METHODS & REFERENCES:

1- Isogenomic reference genomes by *de novo* assembly of experimentally-relevant human diploid laboratory cell lines Volpe et al., 2023
2-Verkko, Telomere-to-telomere assembly of diploid chromosomes with Verkko Rautiainen M., et al., *Nat Biotechnol* , 2023
3- Mashmap, A fast adaptive algorithm for computing whole-genome homology maps Jain C., et al., *Bioinformatics* 2018
4-Hifiasm Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm Cheng H., et al., *Nat. Methods* 2021
5- Merqury, Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies Rhie et al., *Genome Biol* 2020
6- BUSCO BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs Simao F., et al., *Bioinformatics* 2015
7- Stainedglass, StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps Vollger et al., *Bioinformatics* 2
8- Winnowmap, Weighted minimizer sampling improves long read mapping Jain C., et al., 2020
9- Sniffles Accurate detection of complex structural variations using single-molecule sequencing Sedlazeck F. J., et al., *Nat Methods* 2018