nature REVIEWS GENETICS

SUPPLEMENTARY INFORMATION

In format as provided by the authors

Long-read human genome sequencing and its applications

Glennis A. Logsdon, Mitchell R. Vollger and Evan E. Eichler

https://doi.org/10.1038/s41576-020-0236-x

SUPPLEMENTARY INFORMATION

1. **Supplementary Note**
2. **Supplementary Figure**
3. **Supplementary References**

Supplementary Note

We performed limited meta-analysis with long-read datasets as part of the comparisons of read length, accuracy, and homopolymer analysis for this review. All datasets are publicly available, and we describe briefly how the data were analyzed.

Data Sources

Pacific Biosciences (PacBio) HG002 CLR data was retrieved from PacBio's public database (https://github.com/PacificBiosciences/DevNet/wiki/HG002-Structural-Variant-Analysis-with-CLR-data) at the following URL: https://downloads.pacbcloud.com/public/dataset/SV-HG002-CLR/m64013_190124_221354.subreads.bam. Library preparation was performed with the SMRTbell Express 2.0 kit and size-selected to be greater than 30 kbp with a BluePippin instrument.

PacBio CHM13 HiFi data[1] was retrieved from SRA accessions SRR9087597, SRR9087598, SRR9087599, and SRR9087600.

Oxford Nanopore Technologies (ONT) CHM13 whole-genome sequencing data was retrieved from the T2T consortium's GitHub repository (https://github.com/nanopore-wgs-consortium/chm13) at the following URL: https://s3.amazonaws.com/nanopore-human-wgs/chm13/nanopore/rel3/rel3.fastq.gz. This dataset is comprised of long reads (N50 = 35.2 kbp) generated on the PromethION at the University of California, Davis and ultra-long reads (N50 = 146.1 kbp) generated on the MinION and GridION at the University of Washington. The entire dataset was base called with Guppy 3.1.5 using the flip-flop model and separated into long- and ultra-long-read datasets based on the read IDs provided from each institution, which are available at the following URLs: https://s3.amazonaws.com/nanopore-human-wgs/chm13/nanopore/rel3/ids/uwashington.ids.gz and https://s3.amazonaws.com/nanopore-human-wgs/chm13/nanopore/rel3/ids/ucd.ids.gz.

Estimation of read length, accuracy, and homopolymer error

PacBio CLR and HiFi data were aligned to GRCh38 using pbmm2 (https://github.com/PacificBiosciences/pbmm2) v1.1.0 with the following parameters: `--preset SUBREAD`. ONT long- and ultra-long-read data were aligned to GRCh38 using minimap2[2] v2.17 with the following parameters: `-ax map-ont -L --eqx`. All data was filtered to exclude secondary, supplementary, and unmapped alignments using SAMtools[3] v1.9 and SAM flag 2308.

Read length was determined from the sequence length column in the SAM file for aligned reads, and read accuracy was calculated from the CIGAR string in the SAM file using the following formula: 100 * (number of matching bases)/(number of matching bases + number of mismatched bases + number of bases in insertions + number of bases in deletions). The code to reproduce these results is available in the GitHub repository listed below in the snakemake `length_and_qv.smk`.

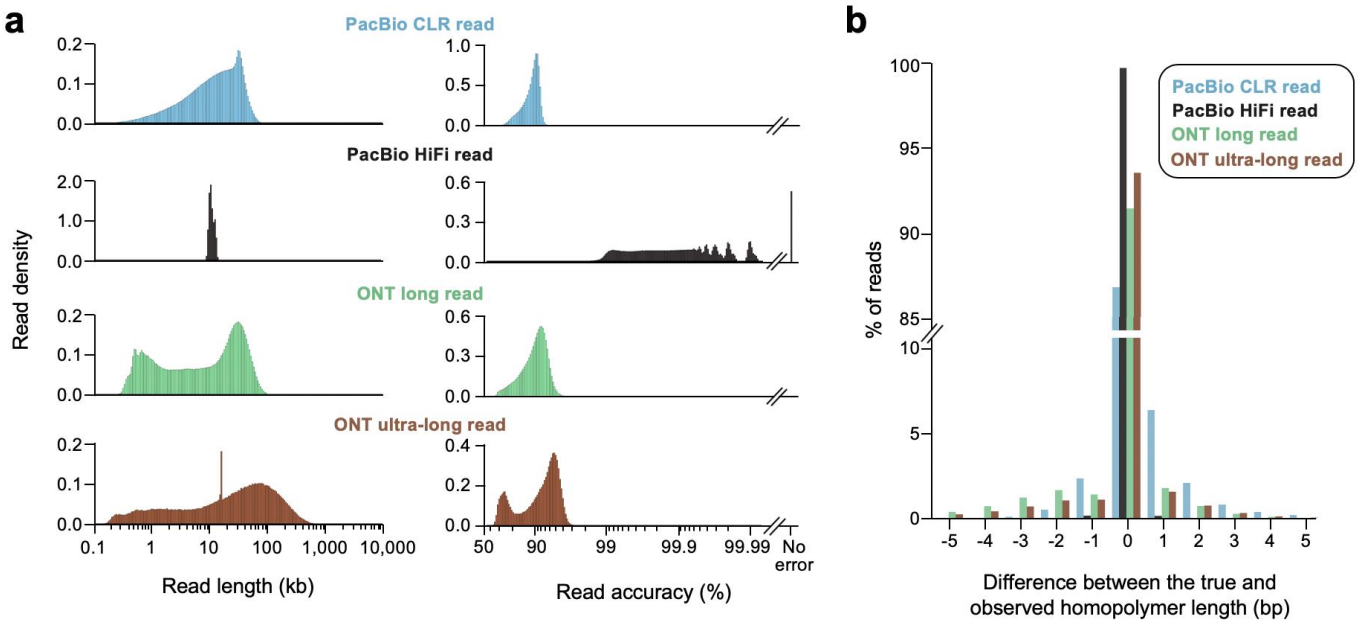
Differences in homopolymer lengths were calculated by first tabulating the homopolymers in the reference genome assembly using the script `Find_homopolymers.py` and then comparing the length of these homopolymers to the aligned reads using the snakemake `homopolymer.smk`. Both programs are available in the GitHub repository listed below. Specifically, the CIGAR operations over homopolymers were used to detect inaccuracies in length and then encoded into a matrix of counts where the row value was the reference genome assembly homopolymer length and the column value was the observed read homopolymer length.

The datasets generated in the analyses above were then visualized using Matplotlib[4] v2.1.2 and seaborn[5] v0.10.0 in a Jupyter notebook called `plots_notebook.ipynb` to create Figure 3c,d.

These analyses were also used to calculate the read lengths, accuracies, and homopolymer profiles in Figure S1 with two exceptions: 1) the reads were aligned to the 2019 T2T consortium CHM13 genome assembly v0.7 from Ref. 6, available at the following URL: https://s3.amazonaws.com/nanopore-human-wgs/chm13/assemblies/chm13.draft_v0.7.fasta.gz, and 2) accuracy and homopolymer estimations were restricted to the X chromosome due to the high level of curation performed on this chromosome.

We provide an online code repository for the analyses described above at the following URL: https://github.com/mrvollger/long_read_nrg.

Supplementary Figure



Supplementary Figure 1. Read length, accuracy, and homopolymer errors of PacBio and ONT long-read data types when aligned to the 2019 T2T CHM13 genome assembly. a) Read length distributions and base accuracy of PacBio and ONT long-read data types. Plots showing the read length and accuracy distributions for PacBio CLR (light blue), PacBio HiFi (black), ONT long (green), and ONT ultra-long (brown) reads. Read accuracy was estimated by aligning raw reads from each data type to the 2019 T2T consortium CHM13 assembly[6] and counting differences in alignment between the reads and the highly curated ChrX. PacBio HiFi reads have a visibly higher read accuracy distribution when aligned to the CHM13 T2T assembly than GRCh38 (**Figure 3**), since these reads are derived from the CHM13 human genome. **b)** Homopolymer accuracy in PacBio and ONT long-read data types. Plot showing the homopolymer accuracy for PacBio CLR (light blue), PacBio HiFi (black), ONT long (green), and ONT ultra-long (brown) reads. Homopolymer error was estimated by aligning raw reads from each data type to the 2019 T2T consortium CHM13 assembly[6] and comparing the differences between the observed homopolymer length in the reads and the true homopolymer length in the ChrX assembly. Homopolymers ≥5 bases were assessed for accuracy. See **Supplementary Note** for details on these analyses.

Supplementary References

1. Vollger, M. R. *et al.* Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann. Hum. Genet.* (2019) doi:10.1111/ahg.12364.
2. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
3. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
4. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science Engineering* **9**, 90–95 (2007).
5. Michael Waskom *et al. seaborn: v0.5.0 (November 2014)*. (Zenodo, 2014). doi:10.5281/zenodo.12710.
6. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *bioRxiv* 735928 (2019) doi:10.1101/735928.