Oxford Nanopore Technologies

EPI2ME

Open access Oxford Nanopore datasets for reproducible benchmarking, sequence exploration, and testing

Oxford Nanopore have published more than 20 representative datasets that are freely available for the community to learn about and explore our highly accurate, information-rich reads at any scale and for both DNA and RNA sequencing

More information can be found at: **epi2me.nanoporetech.com/dataindex/**

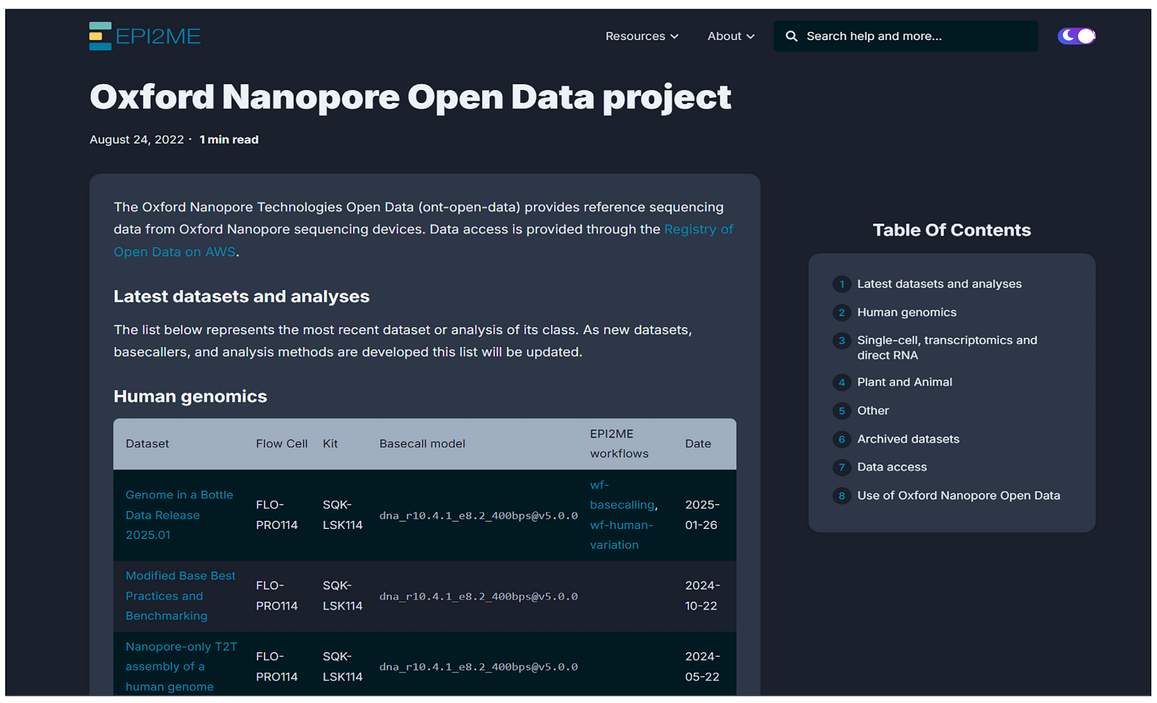

Fig. 1 Index page on the EPI2ME™ website of datasets created and published by Oxford Nanopore.

Comprehensive and representative datasets

The Oxford Nanopore Open Data Project aims to provide representative datasets to support a variety of use cases. This includes exploration of the unique characteristics of Oxford Nanopore sequencing data, validation and reproduction of internal performance benchmarks, demonstration of EPI2ME workflows, and the development or benchmarking of novel analytical tools and methodologies. All datasets are freely accessible and can be downloaded anonymously, without the need for user registration, from a public Amazon Web Services (AWS) S3 bucket: **s3://ont-open-data**. Each dataset typically includes raw signal data, basecalled reads, and accompanying analysis outputs. Comprehensive metadata is also provided, covering sample preparation, sequencing parameters, and illustrative analytical results, including EPI2ME workflow outputs.

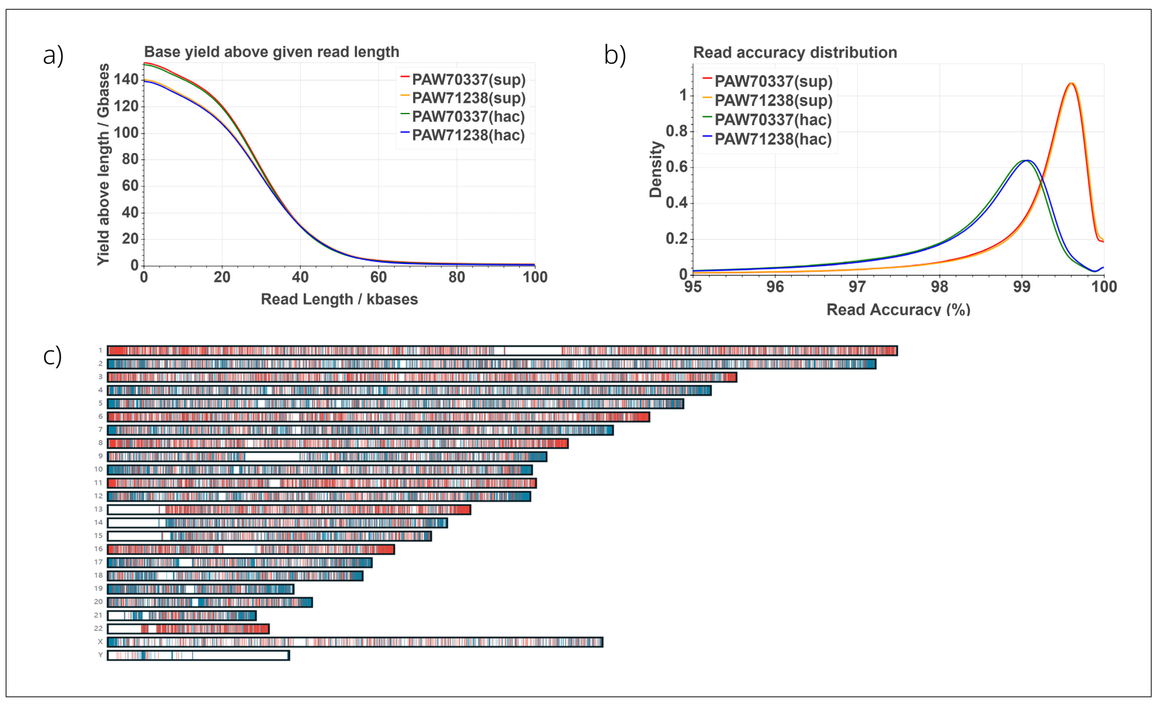

Fig. 2 a) Sequencing yields of HG002 HAC and SUP over two flow cells b) read accuracy of HG002 HAC and SUP over two flow cells c) karyogram of chromosomal hotspots of structural variation from report output by wf-human-variation for HG002 (red: insertion, blue: deletions).

Reproducible benchmarks demonstrate performance

One of the latest releases is the Genome in a Bottle sample dataset covering HG001-07 including the Ashkenazi and Han Chinese Trios. The samples were prepared using our popular SQK-LSK114 protocol with dna_r10.4.1_e8.2_400bps@v5.0.0 basecaller models both HAC and SUP with 5mCG_5hmCG modifications. The quality and quantity of data represents what a user could expect from a routine sequencing experiment and facilitates reproducible benchmarking. **Fig. 2a** shows that the average sequencing yield per flow cell is ~140 GB and **Fig. 2b** shows that average read accuracy is >99%. The dataset can also be used to test out some of our workflows for example wf-human-variation for exploring most types of variants including structural variation (**Fig. 2c**) and our latest release: wf-trio for joint variant merging and pedigree phasing.

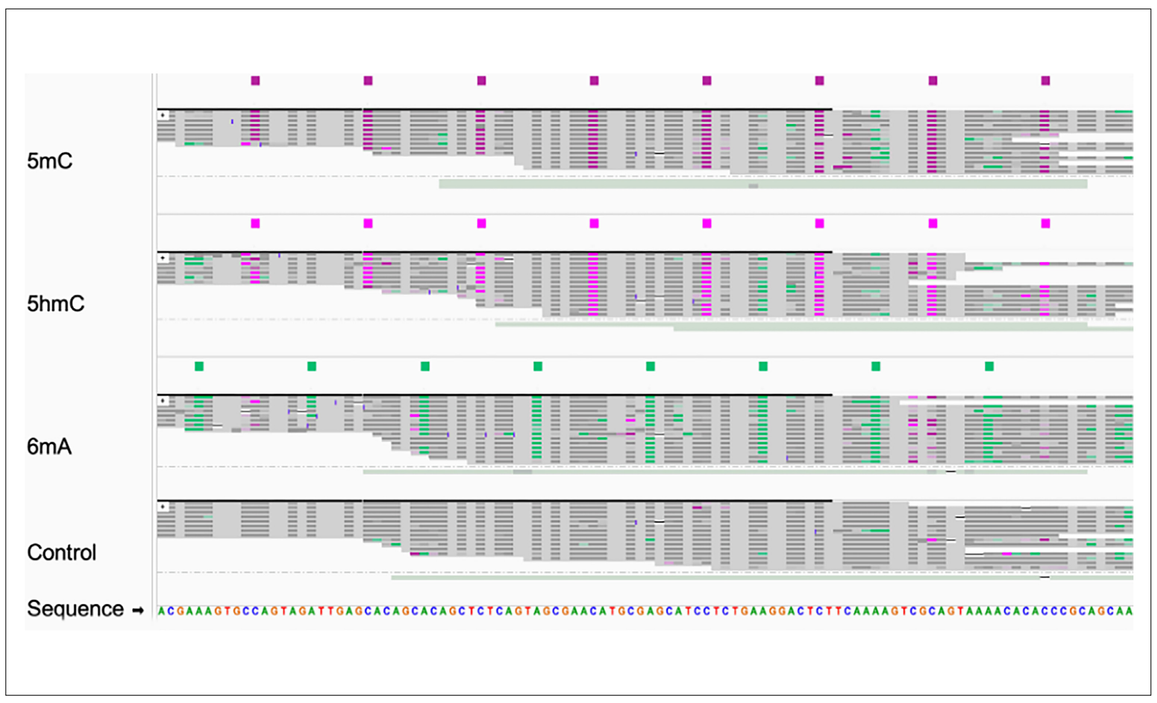

Fig. 3 Modified base validation reads displaying in Integrative Genomics Viewer (IGV).

Explore and learn about modification patterns

Datasets are published with relevant modifications where applicable. Two newly published datasets are dedicated to exploring modifications. First a dataset of synthetic oligonucleotides, each containing canonical (unmodified) or modified bases within all distinct 5-mer sequence contexts for three DNA modifications: 5mC, 5hmC, and 6mA (**Fig. 3**). Accompanying tutorials provide a guide to analyses. Benchmarking results demonstrate high accuracy in detecting modified bases, with precision and recall metrics. The second dataset is of RNA modifications, again derived from synthetic oligonucleotides, it includes modifications such as m6A, m5C, pseudouridine, and inosine. Special emphasis is placed on m6A within DRACH sequence contexts. This resource supports researchers in evaluating model performance and advancing RNA modification analysis.

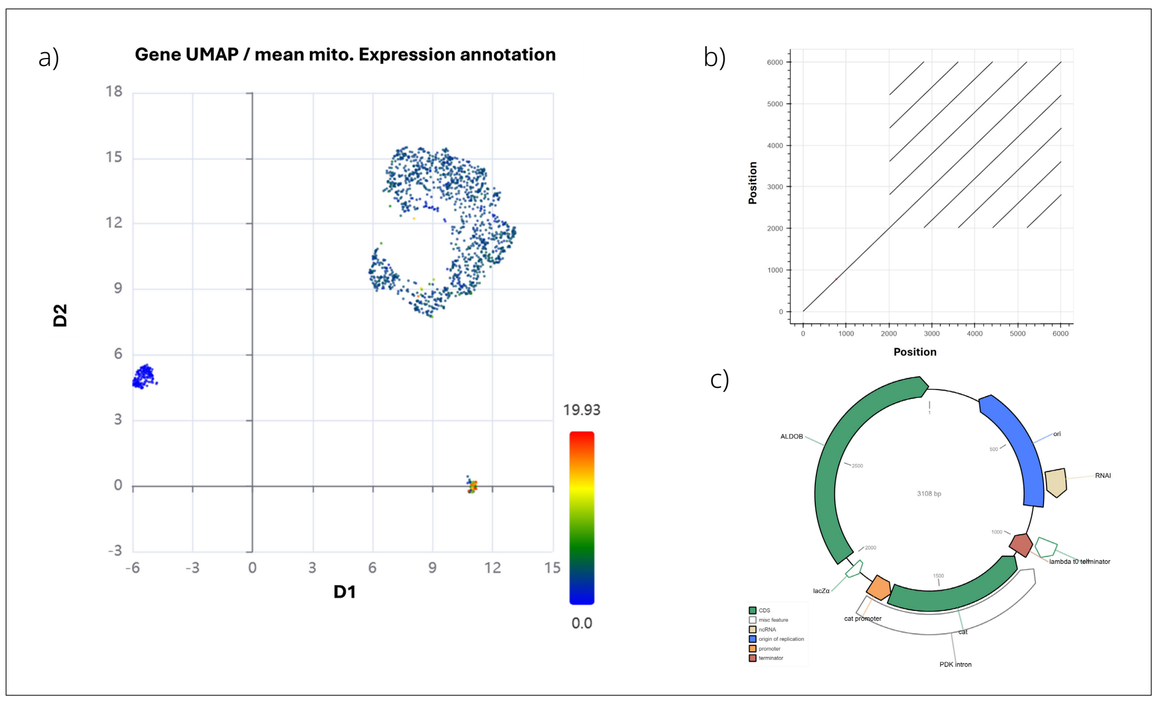

Fig. 4 a) Mitochondria gene expression UMAP plot from the wf-single-cell b) wf-clone-validation self alignment dot-plot c) plannotate plot of generic protein coding plasmid.

Test EPI2ME workflows and analysis tools

Application-specific EPI2ME datasets are available, including a comprehensive single-cell cDNA sequencing dataset for human 293T and Jurkat cell lines, generated using 10x Genomics GEM-X kits—to test the EPI2ME wf-single-cell (**Fig. 4a**). A plasmid dataset is also available, consisting of 96 unique plasmids ranging from 2–7 Kb in size, featuring complex inserts such as duplications and repetitive regions (**Fig. 4b**), as well as generic protein coding plasmids (**Fig. 4c**) ideal for testing out wf-clone-validation. An archive of over 20 curated Oxford Nanopore datasets is currently accessible, with additional releases planned throughout 2026.

Information correct at time of publication. May be subject to change.

Oxford Nanopore Technologies, the Wheel icon, AmPORE-TB, EPI2ME, GridION, MinION, MinKNOW, PromethION, P2 Solo, and P2 are registered trademarks or the subject of trademark applications of Oxford Nanopore Technologies plc in various countries. Information contained herein may be protected by copyright, patents or patents pending of Oxford Nanopore Technologies plc. All other brands and names contained are the property of their respective owners. © 2026 Oxford Nanopore Technologies plc. All rights reserved. Oxford Nanopore Technologies products are RUO. Products labelled/branded as Oxford Nanopore Diagnostics may be RUO or may be regulated as in-vitro diagnostic devices in some jurisdictions, please check individual product labelling.

PO_1299(EN)_V2_19May2026