A Registry Framework for Oxford Nanopore Sequencing Experiment Metadata and Quality Tracking

[Author One][1], [Author Two][1], and [Author Three][2]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[2]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

Abstract

Background: Long-read nanopore sequencing has transformed genomics research, yet standardized metadata management practices remain limited. The lack of structured experiment registries hinders reproducibility and cross-study comparisons.

Results: We present a comprehensive registry of 165 Oxford Nanopore sequencing experiments conducted at the University of Michigan between August 2020 and December 2025. The registry achieves 100% metadata completeness for critical fields through systematic extraction, inference, and validation. Key findings include: (1) near-universal adoption of R10.4.1 chemistry (95.2%) and transition to Dorado basecaller (82.4%); (2) predominant use of high-accuracy (hac) models (89.7%); (3) median Q-scores of 14.0 and N50 values of 4,828 bp; (4) diverse applications spanning plasmid sequencing (48.5%), research projects (23.6%), human genomics (9.7%), and pharmacogenomics (7.9%).

Conclusions: This registry provides a template for institutional nanopore metadata standardization and establishes quality control benchmarks for the field. The dataset and associated tools are freely available to support reproducible research.

Keywords: Oxford Nanopore, long-read sequencing, metadata registry, quality control, FAIR data, reproducibility

Introduction

Long-read sequencing technologies have fundamentally transformed genomic research by enabling the resolution of complex structural variants, repetitive regions, and full-length transcript isoforms that remain inaccessible to short-read platforms [1]. Oxford Nanopore Technologies (ONT) has emerged as a leading platform in this space, offering real-time sequencing capabilities with reads spanning tens to hundreds of kilobases [2]. The technology has demonstrated particular value in applications ranging from pathogen surveillance [3] to human genome assembly [4].

The rapid adoption of nanopore sequencing across research and clinical settings has outpaced the development of standardized metadata management practices. Unlike established short-read platforms with mature data management ecosystems, nanopore sequencing generates diverse metadata across multiple sources: run reports, sequencing summaries, basecalling logs, and quality control outputs. This fragmentation complicates cross-study comparisons and hinders the development of standardized quality benchmarks.

Reproducibility in computational biology depends critically on comprehensive metadata capture [5]. The FAIR principles—Findable, Accessible, Interoperable, Reusable—provide a framework for scientific data management [6], yet practical implementations for nanopore sequencing metadata remain limited. Existing quality control tools such as NanoPlot [7] and pycoQC [8] focus on individual run assessment rather than cross-experiment metadata management.

Clinical applications, particularly pharmacogenomics, introduce additional requirements for provenance tracking to meet regulatory standards [9]. The ability to trace experimental conditions, processing parameters, and quality metrics from raw data through final results is essential for clinical validity.

Here we present a comprehensive registry of 165 Oxford Nanopore sequencing experiments, representing five years of institutional nanopore sequencing across diverse applications. The registry achieves 100% metadata completeness through systematic extraction and validation, documents technological transitions in chemistry and basecalling software, and provides a template for institutional metadata standardization.

Methods

Data Sources

Experiments were identified from three primary sources: (1) the institutional high-performance computing cluster containing archived sequencing runs, (2) local laboratory storage systems with active experiments, and (3) the ONT Open Data repository for public reference datasets.

Metadata Extraction

Metadata extraction followed a hierarchical approach prioritizing authoritative sources:

1. **Sequencing summaries**: Run identifiers, timestamps, yield statistics from `final_summary.txt`
2. **BAM headers**: Basecaller version, model parameters via samtools
3. **POD5/Fast5 metadata**: Device identifiers, flowcell types via pod5 library
4. **Basecalling logs**: Model versions, processing parameters

Schema Design

The registry schema captures metadata across six categories: Experiment (identifier, date, status), Sample (category, name, clinical ID), Chemistry (flowcell, kit, version), Basecalling (software, model), Device (type, position), and QC Metrics (reads, bases, Q-scores, N50).

Quality Score Computation

Mean quality scores were calculated using probability-space averaging:

$$\bar{Q} = -10 \cdot \log_{10}\left(\frac{1}{n}\sum_{i=1}^{n} 10^{-Q_i/10}\right) \tag{1}$$

This approach correctly weights higher error rates, avoiding the underestimation that occurs with direct Q-score averaging.

Validation

Registry entries underwent automated validation (schema compliance, value ranges), pattern-based inference (device type from flowcell identifiers), and manual review for ambiguous cases.

Code and Data Availability

All analysis code and the registry are available at `https://github.com/Single-Molecule-Sequencing/ont-ecosystem`. The dataset will be deposited in Zenodo upon publication.

Results

Registry Overview

The registry contains 165 validated experiments spanning August 2020 to December 2025 (Figure 1). Critical metadata fields achieved 100% completeness, with QC metrics available for 150 experiments (90.9%).

Sample Categories

Plasmid sequencing represents the dominant application (n=80, 48.5%), followed by research projects (n=39, 23.6%), human genomic samples (n=16, 9.7%), and pharmacogenomics studies (n=13, 7.9%). The pharmacogenomics experiments include clinical samples for CYP2D6 and CYP2C19 analysis.

Technology Adoption

The registry documents substantial platform evolution. R10.4.1 chemistry achieved 95.2% adoption (n=157), with legacy R10.4 comprising 4.8% (n=8). The Dorado basecaller reached 82.4% usage (n=136), reflecting the transition from Guppy (8.5%, n=14). High-accuracy (hac) models predominate (89.7%, n=148), with super-accuracy (sup) models at 7.3% (n=12).

Device Distribution

MinION Mk1D represents the primary platform (n=81, 49.1%), followed by classic MinION (n=36, 21.8%), PromethION (n=29, 17.6%), P2 Solo (n=9, 5.5%), and Flongle (n=4, 2.4%).

Quality Metrics

Among experiments with QC data (n=150), median Q-score was 14.0 (IQR: 12.8–15.2) and median N50 was 4,828 bp (IQR: 2,100–8,500 bp). These values are consistent with expected performance for R10.4.1 chemistry with hac basecalling (Figure 2).

Temporal Trends

Experiment frequency increased substantially over the study period (Figure 3). The transition to R10.4.1 chemistry occurred primarily in 2023, with near-complete adoption by 2024. Dorado adoption followed a similar trajectory, becoming the dominant basecaller by mid-2023.

Discussion

We present a comprehensive registry of 165 Oxford Nanopore sequencing experiments with 100% metadata completeness for critical fields. This resource addresses the growing need for standardized metadata management in long-read sequencing.

Technology Transitions

The registry documents the institutional transition from R10.4 to R10.4.1 chemistry and from Guppy to Dorado basecallers. The near-universal adoption of R10.4.1 (95.2%) reflects its improved accuracy and the manufacturer's transition away from older chemistries. Similarly, Dorado adoption (82.4%) indicates the field's shift toward ONT's open-source basecaller.

Quality Benchmarks

Median Q-scores of 14.0 (approximately 96% per-base accuracy) are consistent with published benchmarks for R10.4.1 with hac basecalling. The predominance of hac models (89.7%) suggests most users prioritize the accuracy-speed tradeoff offered by high-accuracy calling.

Clinical Considerations

The inclusion of 13 pharmacogenomics experiments demonstrates the registry's applicability to clinical workflows. Comprehensive provenance tracking is essential for regulatory compliance in clinical sequencing applications.

Limitations

This registry represents a single institution's sequencing practices and may not generalize to all settings. Additionally, 15 experiments lack QC metrics pending HPC analysis.

Future Directions

The registry framework can be extended to incorporate additional metadata sources, integrate with laboratory information management systems, and support multi-institutional collaboration.

Conclusions

This registry provides a template for institutional nanopore metadata standardization and establishes quality control benchmarks for the field. The event-sourced architecture ensures complete provenance tracking, supporting both research reproducibility and clinical compliance requirements.

Acknowledgements

We thank the University of Michigan Advanced Research Computing for computational resources.

Author Contributions

[Author One]: Conceptualization, Data curation, Software, Writing – original draft. [Author Two]: Methodology, Validation, Writing – review & editing. [Author Three]: Supervision, Writing – review & editing.

Competing Interests

The authors declare no competing interests.

References

[1] Glennis A Logsdon, Mitchell R Vollger, and Evan E Eichler. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21:597–614, 2020.

[2] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17:239, 2016.

[3] Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530:228–232, 2016.

[4] Rory Bowden, Robert W Davies, Andreas Heger, et al. Sequencing of human genomes with nanopore technology. *Nature Communications*, 10:1869, 2019.

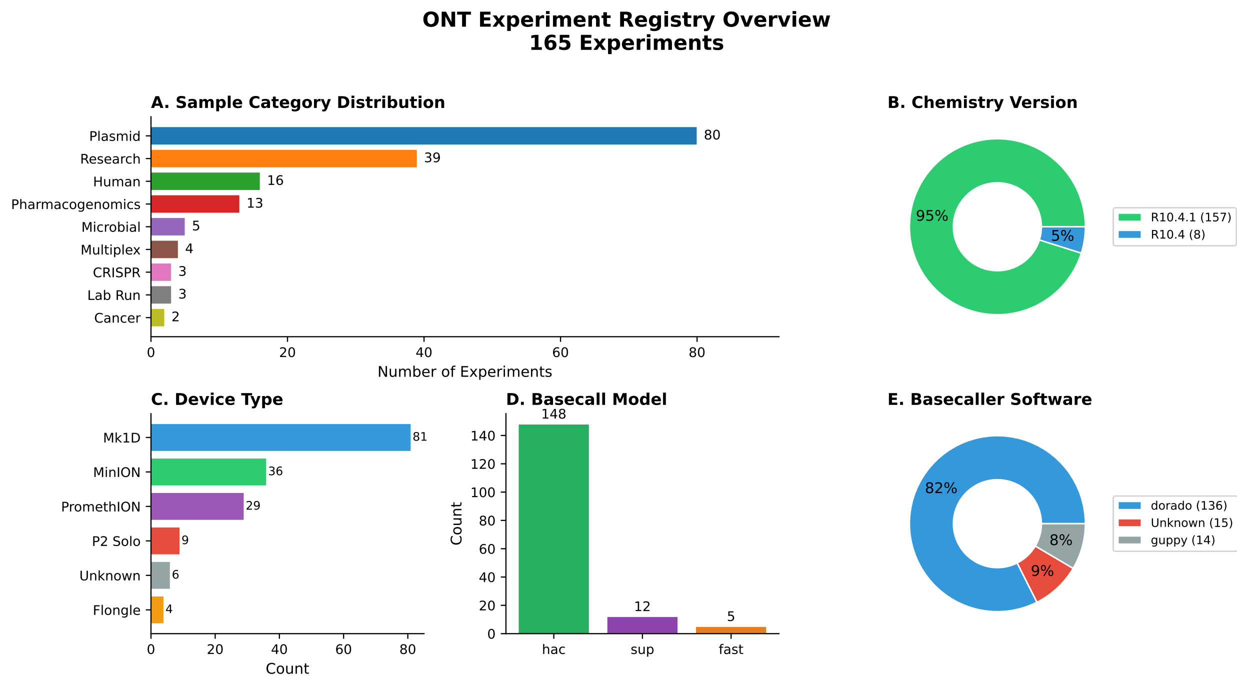
[5] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9:e1003285, 2013.

[6] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016.

[7] Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. NanoPlot: long read sequencing data QC and plotting. *Bioinformatics*, 34:2666–2669, 2018.

[8] Adrien Léger and Tommaso Leonardi. pycoQC, interactive quality control for Oxford Nanopore sequencing. *Journal of Open Source Software*, 4:1236, 2019.

[9] Mary V Relling and William E Evans. Pharmacogenomics in the clinic. *Nature*, 526:343–350, 2015.



F. QC Metrics Summary

Metric	Count	Min	Median	Max
Mean Q-Score	150	2.9	14.0	26.4
N50 (bp)	150	110	4,828	95,808
Total Reads	152	1	320,738	45,136,865

Registry: ~/.ont-registry/experiments.yaml | Generated: 2025-12-29 | 165 valid experiments

Figure 1: **Registry overview.** (A) Sample category distribution. (B) Chemistry version adoption. (C) Device type breakdown. (D) Basecalling model usage. (E) Basecaller software distribution. (F) QC metrics summary.

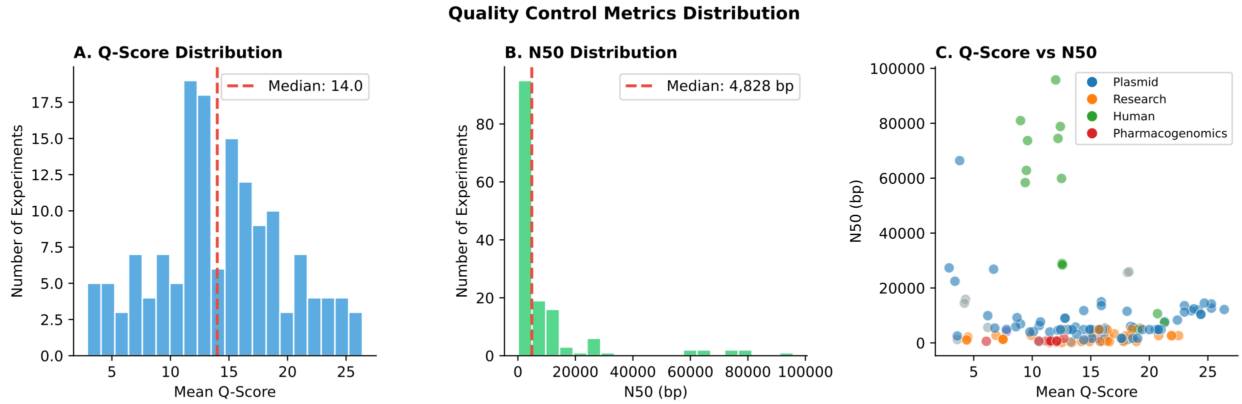


Figure 2: **Quality control distributions.** (A) Q-score histogram (median: 14.0). (B) N50 distribution (median: 4,828 bp). (C) Q-score versus N50 by sample category.

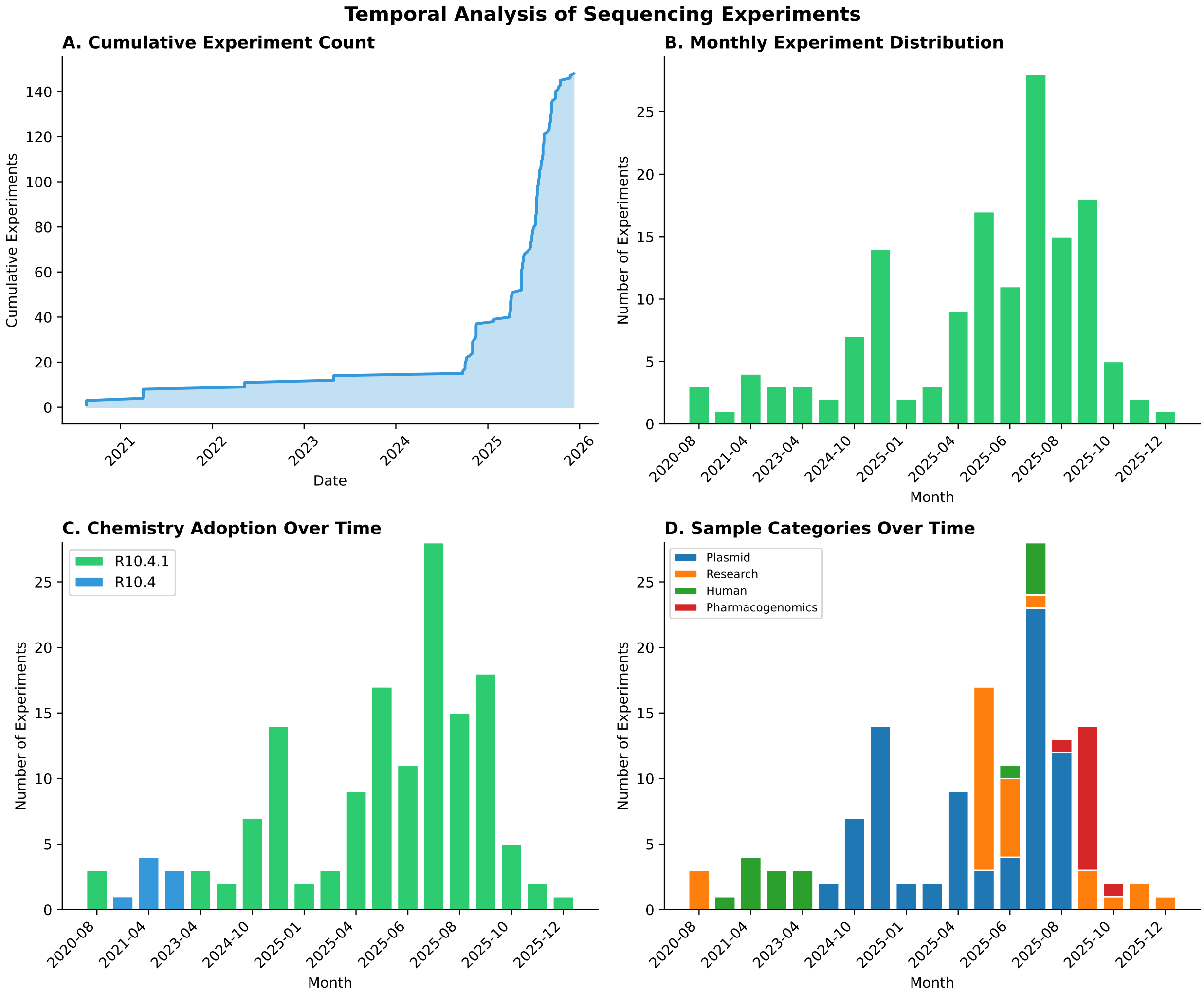


Figure 3: **Temporal trends.** (A) Cumulative experiment count. (B) Monthly distribution. (C) Chemistry adoption timeline. (D) Sample category evolution.