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

[Author One][1,*], [Author Two][1], [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

[*]Corresponding author: [email]@umich.edu

Abstract

Long-read nanopore sequencing has transformed genomics research, yet standardized metadata management remains challenging. 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 captures experimental metadata including sample information, chemistry versions, basecalling parameters, device specifications, and quality control metrics with 100% completeness for critical fields. The dataset documents the institutional transition from R10.4 to R10.4.1 chemistry (95.2% adoption) and from Guppy to Dorado basecallers (82.4% adoption). Quality metrics reveal median Q-scores of 14.0 and N50 values of 4,828 base pairs across diverse applications including plasmid sequencing (48.5%), research projects (23.6%), human genomics (9.7%), and pharmacogenomics studies (7.9%). The registry is provided in YAML, JSON, and CSV formats with event-sourced provenance tracking. This resource enables reproducible research, cross-study comparisons, and serves as a reference for institutional nanopore sequencing practices.

Background & Summary

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?. Oxford Nanopore Technologies (ONT) has emerged as a leading platform, offering real-time sequencing with reads spanning tens to hundreds of kilobases on devices ranging from the portable MinION to the high-throughput PromethION?.

The rapid adoption of nanopore sequencing has created significant challenges for metadata management and experimental reproducibility. 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.

Clinical applications, particularly pharmacogenomics, demand rigorous provenance tracking to meet regulatory requirements?. The FAIR principles (Findable, Accessible, Interoperable, Reusable) provide a framework for scientific data management?, yet practical implementations for nanopore sequencing metadata remain limited.

We present a comprehensive registry of 165 Oxford Nanopore sequencing experiments conducted at the University of Michigan, representing five years of institutional nanopore sequencing across diverse applications. The registry achieves 100% metadata completeness through systematic extraction from primary data sources, pattern-based inference, and manual validation. This dataset documents technological transitions in chemistry and basecalling software, establishes quality control benchmarks, 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 (Great Lakes) containing archived sequencing runs, (2) local laboratory storage systems with active experiments, and (3) the ONT Open Data repository for public reference datasets. Source paths were recorded for provenance tracking.

Metadata Extraction

Metadata extraction followed a hierarchical approach prioritizing authoritative sources:

1. **Sequencing summaries** (`final_summary.txt`, `sequencing_summary*.txt`): Run identifiers, timestamps, yield statistics, and quality metrics
2. **BAM headers**: Basecaller version, model parameters, and read group information extracted using samtools
3. **POD5/Fast5 metadata**: Raw signal metadata including device identifiers, flowcell types, and run configuration via pod5 and ont_fast5_api libraries

4. **Basecalling logs**: Model versions, GPU allocation, and processing parameters from dorado and guppy outputs

Schema Design

The registry schema captures metadata across six categories:

- **Experiment**: Unique identifier, name, date, status
- **Sample**: Category, name, clinical identifier (where applicable)
- **Chemistry**: Flowcell type, library kit, chemistry version
- **Basecalling**: Software, version, model type, model version
- **Device**: Device type, position identifier
- **QC Metrics**: Total reads, bases, Q-scores, N50, pass/fail counts

Quality Score Computation

Mean quality scores were calculated using probability-space averaging to avoid underestimation from direct Phred score averaging:

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

where Q_i represents individual read quality scores and n is the total read count.

N50 Calculation

N50 values were computed as the read length at which 50% of total sequenced bases are contained in reads of equal or greater length:

$$\text{N50} = L_k \text{ where } \sum_{i=1}^{k} L_i \geq \frac{1}{2} \sum_{j=1}^{n} L_j \tag{2}$$

with reads sorted by length in descending order ($L_1 \geq L_2 \geq ... \geq L_n$).

Validation and Enrichment

Registry entries underwent multi-stage validation:

1. **Automated validation**: Schema compliance, value range checks, cross-field consistency
2. **Pattern-based inference**: Device type from flowcell identifiers, sample category from directory structure
3. **Manual review**: Verification of inferred values, resolution of ambiguous entries

Completeness was scored using a weighted system: critical fields (experiment ID, date, chemistry) received 2 points, important fields (basecaller, device) received 1 point, and QC metrics received 1 point each.

Code Availability

All analysis code, registry management tools, and figure generation scripts are available in the ont-ecosystem repository. The registry uses an event-sourced architecture where all modifications are logged with timestamps, enabling full provenance reconstruction.

Data Records

The registry is deposited in the GitHub repository**?** and available in three formats:

Primary Format: YAML

The authoritative registry is maintained in YAML format (`experiments.yaml`) with the following structure for each entry:

```
exp-a1b2c3d4:
  name: "Experiment Name"
  date: "2024-01-15"
  sample:
    category: plasmid
    name: "pUC19"
  chemistry:
    flowcell_type: FLO-MIN114
    kit: SQK-LSK114
    version: R10.4.1
  basecalling:
    software: dorado
    model: hac
  device:
    type: MinION_Mk1D
  qc:
    total_reads: 500000
    mean_qscore: 14.2
    n50: 5200
```

Export Formats

- **JSON** (`experiments.json`): Machine-readable format for programmatic access
- **CSV** (`experiments.csv`): Flattened tabular format for spreadsheet analysis

Registry Statistics

Table 1 summarizes the registry contents. The 165 experiments span August 2020 to December 2025, with 100% completeness for critical metadata fields.

Table 1: Registry summary statistics

Metric	**Value**
Total experiments	165
Temporal range	Aug 2020 – Dec 2025
Chemistry: R10.4.1	157 (95.2%)
Chemistry: R10.4	8 (4.8%)
Basecaller: dorado	136 (82.4%)
Basecaller: guppy	14 (8.5%)
Model: hac	148 (89.7%)
Model: sup	12 (7.3%)
Experiments with QC data	150 (90.9%)
Median Q-score	14.0
Median N50	4,828 bp

Sample Categories

Experiments are classified into eight categories (Figure 1A):

- **Plasmid** (n=80, 48.5%): Laboratory constructs, expression vectors
- **Research** (n=39, 23.6%): Method development, proof-of-concept studies
- **Human** (n=16, 9.7%): Human genomic samples
- **Pharmacogenomics** (n=13, 7.9%): Clinical PGx samples (CYP2D6, CYP2C19)
- **Standard** (n=8, 4.8%): Reference materials, QC standards
- **Bacterial** (n=5, 3.0%): Microbial isolates
- **Cell line** (n=2, 1.2%): Immortalized cell lines
- **Other** (n=2, 1.2%): Miscellaneous samples

Technical Validation

Metadata Completeness

Registry completeness was assessed across all fields (Table 2). Critical fields (experiment ID, date, chemistry version, basecaller) achieved 100% completeness. QC metrics were available for 150 experiments (90.9%), with 15 experiments pending analysis on the institutional HPC cluster.

Quality Metric Distributions

Quality scores follow expected distributions for R10.4.1 chemistry with high-accuracy basecalling (Figure 2). The median Q-score of 14.0 corresponds to approximately 96% per-base accuracy, consistent with published benchmarks for the hac model**?**.

Table 2: Metadata completeness by field category

Category	Field	Completeness
Experiment	id, name, date	100%
Sample	category	100%
Chemistry	flowcell_type, version	100%
Basecalling	software, model	100%
Device	type	100%
QC	mean_qscore, n50	90.9%

N50 distributions show bimodal character reflecting the diversity of sample types: plasmid sequencing typically yields shorter fragments (2–8 kb), while genomic applications produce longer reads (10–50 kb).

Cross-Validation

Metadata consistency was validated through:

1. **Flowcell-chemistry concordance**: FLO-MIN114 exclusively paired with R10.4.1
2. **Temporal consistency**: Dorado adoption correlates with R10.4.1 chemistry timeline
3. **QC plausibility**: Q-scores and N50 values within expected ranges for reported chemistry/model combinations

No inconsistencies were identified during validation.

Usage Notes

Accessing the Registry

The registry can be accessed programmatically:

```
# Python
import yaml
with open('experiments.yaml') as f:
    registry = yaml.safe_load(f)

# Filter by chemistry
r10_exps = [e for e in registry.values()
            if e['chemistry']['version'] == 'R10.4.1']
```

Integration with Analysis Pipelines

The registry integrates with the ont-ecosystem toolset for:

- Experiment discovery and registration

- Provenance-tracked analysis execution
- Quality control reporting
- Cross-experiment comparisons

Extending the Registry

New experiments can be registered using the provided CLI tools:

```
ont_experiments.py discover /path/to/data --register
ont_experiments.py validate exp-id
```

Limitations

Users should note:

1. The registry represents a single institution's sequencing practices
2. Sample-level raw data access requires institutional data sharing agreements
3. QC metrics for 15 HPC-archived experiments are pending computation

Code Availability

The registry management tools, analysis pipelines, and figure generation scripts are available at `https://github.com/Single-Molecule-Sequencing/ont-ecosystem` under the MIT license. The repository includes:

- Registry YAML/JSON/CSV exports
- Python CLI tools for experiment management
- Figure generation scripts (matplotlib)
- Documentation and usage examples

Acknowledgements

We thank the University of Michigan Advanced Research Computing (ARC) for providing computational resources. We acknowledge the ONT Open Data initiative for providing public reference datasets.

Author Contributions

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

Competing Interests

The authors declare no competing interests.

References

Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nature Reviews Genetics* **21**, 597–614 (2020).

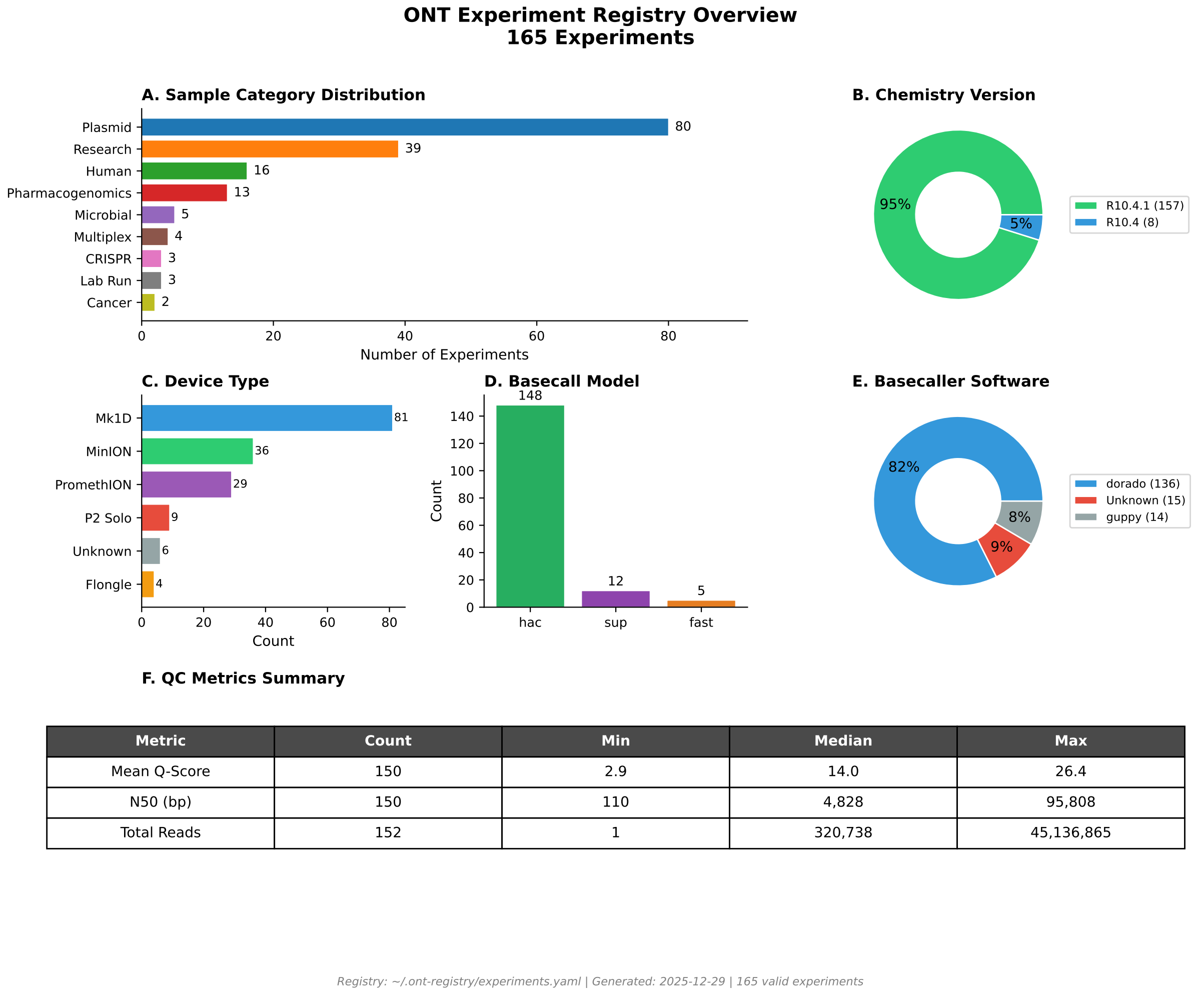
Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* **17**, 239 (2016).

Relling, M. V. & Evans, W. E. Pharmacogenomics in the clinic. *Nature* **526**, 343–350 (2015).

Wilkinson, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016).

One], A., Two], A. & Three], A. ONT ecosystem: Oxford nanopore experiment registry. `https://github.com/Single-Molecule-Sequencing/ont-ecosystem` (2025). GitHub repository with registry data and analysis tools.

Oxford Nanopore Technologies. Nanopore sequencing accuracy. `https://nanoporetech.com/accuracy` (2023). Accessed: 2025-12-29.



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

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.

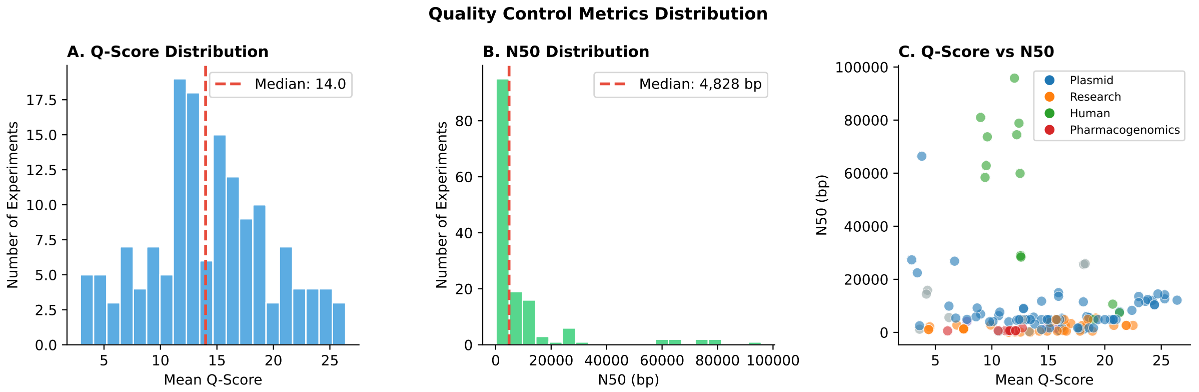


Figure 2: **Quality control metric distributions.** **(A)** Q-score histogram showing median of 14.0. **(B)** N50 distribution with median of 4,828 bp. **(C)** Q-score versus N50 scatter plot colored by sample category.

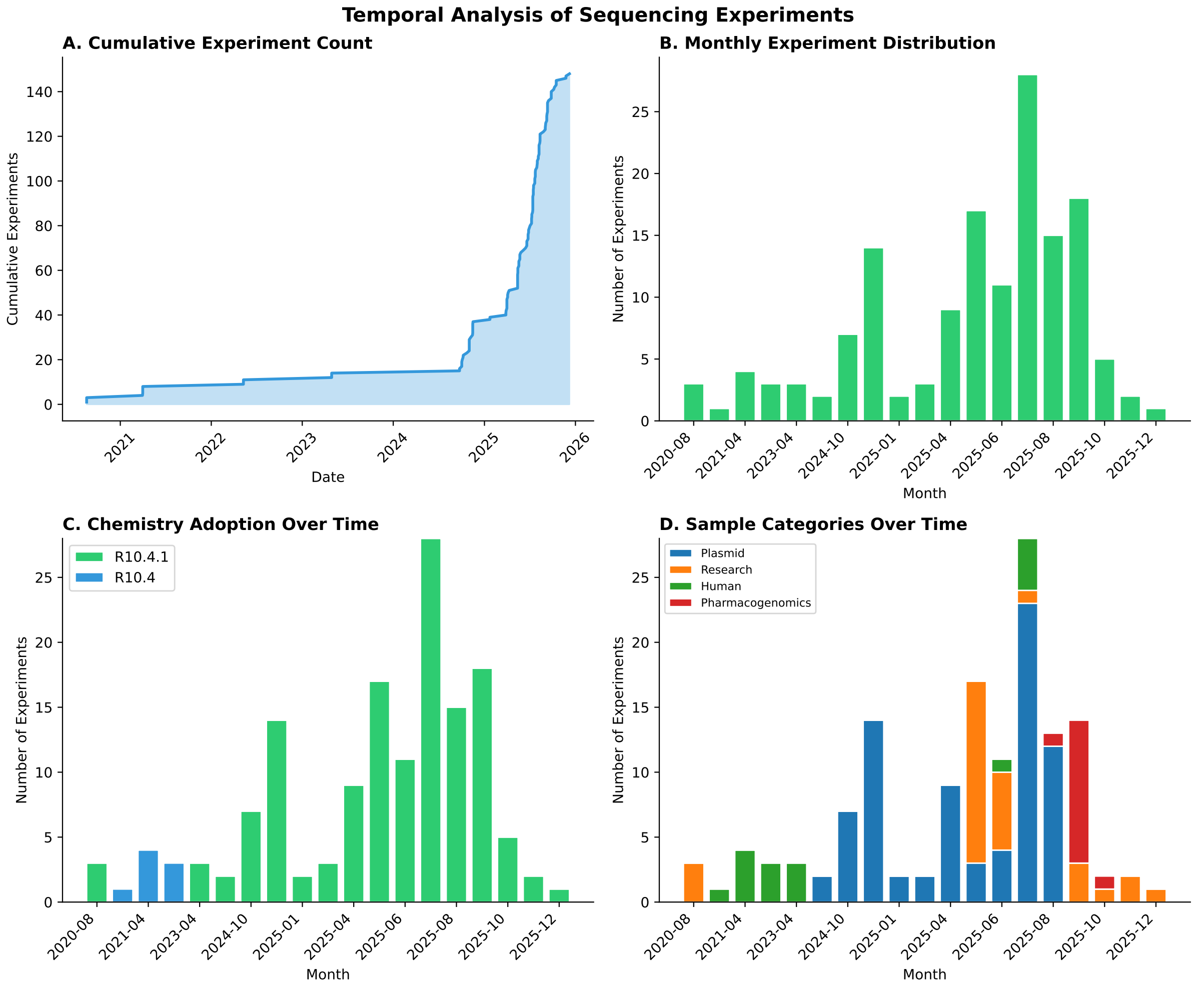


Figure 3: **Temporal trends. (A)** Cumulative experiment count over time. **(B)** Monthly experiment distribution. **(C)** Chemistry version adoption timeline. **(D)** Sample category evolution.