A Comprehensive Registry Framework for Oxford Nanopore Sequencing Experiments: Metadata Management, Quality Tracking, and Institutional Standardization

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

Background: Oxford Nanopore Technologies (ONT) sequencing generates complex metadata across instruments, chemistries, and basecalling configurations. Systematic tracking of experiment provenance and quality metrics is essential for protocol optimization, quality assurance, and reproducibility, yet standardized approaches for institutional registry management remain limited.

Methods: We developed a comprehensive experiment registry framework combining automated metadata extraction from MinKNOW output files and BAM headers, pattern-based inference for missing fields, and systematic validation protocols. Registry completeness was assessed using a weighted scoring system prioritizing critical fields (sample, chemistry, basecall model) and quality metrics (Q-score, N50). Provenance was tracked through event-sourced logging with Git-based versioning.

Results: The registry encompasses 165 validated ONT sequencing experiments spanning August 2020 to December 2025, achieving 100% "good" completeness status. Sample categories included plasmid sequencing (n=80, 48.5%), research projects (n=39, 23.6%), human genomics (n=16, 9.7%), and pharmacogenomics (n=13, 7.9%). Technical characterization revealed near-universal R10.4.1 chemistry adoption (95.2%), dorado basecaller dominance (82.4%), and preferential high-accuracy model usage (89.7%). Quality metrics across 150 experiments showed median Q-score of 14.0 (range: 2.9–26.4) and median N50 of 4,828 bp (range: 110–95,808 bp). Temporal analysis captured exponential growth in 2025, technology transitions from R10.4/guppy to R10.4.1/dorado, and application evolution from research toward plasmid sequencing and clinical pharmacogenomics.

Conclusions: Systematic metadata tracking enables comprehensive characterization of institutional nanopore sequencing operations. The registry framework—combining YAML storage, hierarchical metadata extraction, and event-sourced provenance—provides a template for managing long-read sequencing experiments. As clinical applications expand, such registries become critical infrastructure for quality benchmarking, protocol optimization, and regulatory compliance.

Keywords: Oxford Nanopore, long-read sequencing, metadata registry, quality control, provenance tracking, pharmacogenomics

1 Introduction

1.1 The Rise of Long-Read Sequencing

Oxford Nanopore Technologies (ONT) sequencing has transformed genomics research by enabling real-time, long-read DNA and RNA sequencing without the need for amplification or synthesis Jain et al. [2016]. Unlike short-read platforms that generate fragments of 150–300 base pairs, nanopore sequencing routinely produces reads exceeding 10,000 bases, with ultra-long protocols achieving reads surpassing 1 megabase Payne et al. [2019]. This capability has proven transformative for applications including *de novo* genome assembly, structural variant detection, full-length transcript isoform characterization, and direct detection of base modifications Logsdon et al. [2020].

The technology has evolved rapidly since its commercial introduction in 2014. Early R7 and R9 pore chemistries have given way to R10.4.1, which achieves modal raw read accuracy exceeding Q20 (99% accuracy) Oxford Nanopore Technologies [2023a]. Concurrently, basecalling algorithms have progressed from early hidden Markov models through recurrent neural networks to current transformer-based architectures, with the dorado basecaller replacing the legacy guppy software as of September 2022 Oxford Nanopore Technologies [2022]. Hardware platforms now span portable MinION devices through high-throughput PromethION systems capable of generating terabases of data per run.

1.2 The Metadata Challenge

This rapid technological evolution presents significant challenges for experiment management and reproducibility. A single nanopore sequencing experiment generates metadata spanning multiple domains: sample information (identity, preparation method, concentration), instrument parameters (device type, flow cell chemistry, pore version), basecalling configuration (software version, model accuracy tier, modification detection), and quality metrics (yield, read length distribution, accuracy estimates). Unlike mature short-read platforms with standardized metadata schemas, the ONT ecosystem lacks consensus approaches for comprehensive metadata capture and management.

The challenge is compounded by the platform's flexibility. The same MinION device might sequence bacterial isolates for species identification, human samples for clinical diagnostics, or synthetic constructs for biotechnology applications—each with distinct metadata requirements and quality expectations. Without systematic tracking, correlating sequencing outcomes with experimental parameters becomes difficult, hindering protocol optimization and troubleshooting.

1.3 Provenance and Reproducibility

Reproducibility in computational biology requires not only methodological transparency but also comprehensive provenance tracking Sandve et al. [2013]. For sequencing experiments, this encompasses the complete chain from sample preparation through data generation to analysis outputs. The FAIR principles (Findable, Accessible, Interoperable, Reusable) provide a framework for data management Wilkinson et al. [2016], yet implementing FAIR-compliant workflows for nanopore sequencing remains challenging given the diversity of experimental contexts and rapidly evolving technology stack.

Institutional sequencing facilities face particular challenges in maintaining experiment registries. High-throughput operations may generate dozens of experiments weekly across multiple instruments, each requiring metadata capture, quality assessment, and long-term archival. Manual curation approaches scale poorly and introduce transcription errors, while fully automated systems must accommodate the heterogeneity of experimental designs and naming conventions employed by diverse research groups.

1.4 Clinical and Regulatory Considerations

The expansion of nanopore sequencing into clinical applications—including infectious disease surveillance, pharmacogenomics, and cancer profiling—introduces additional requirements for metadata management Quick et al. [2016], Bowden et al. [2019]. Clinical Laboratory Improvement Amendments (CLIA) and equivalent international regulations mandate documented quality control procedures, instrument calibration records, and complete audit trails linking patient samples to reported results. Registries supporting clinical workflows must therefore capture not only technical metadata but also chain-of-custody information and quality benchmarks against validated reference standards.

Pharmacogenomics applications exemplify these requirements. Accurate genotyping of cytochrome P450 enzymes and other pharmacologically relevant genes directly impacts drug dosing decisions, necessitating rigorous quality thresholds and comprehensive documentation Relling and Evans [2015]. As nanopore platforms demonstrate sufficient accuracy for clinical variant calling, institutional frameworks for quality tracking become essential infrastructure rather than optional conveniences.

1.5 Existing Approaches

Several tools address aspects of nanopore data management. MinKNOW, ONT's instrument control software, generates run reports and summary statistics but does not provide cross-experiment registry functionality. EPI2ME, ONT's cloud analysis platform, offers workflow execution and result aggregation but focuses on analysis rather than comprehensive metadata management. Third-party tools including NanoPlot De Coster et al. [2018] and PycoQC Léger and Leonardi [2019] provide quality visualization but operate on individual experiments without registry integration.

Laboratory information management systems (LIMS) offer general-purpose sample tracking but typically lack nanopore-specific metadata schemas and quality metrics. Custom database solutions developed by individual laboratories address local requirements but rarely achieve the standardization necessary for cross-institutional comparison or community adoption.

1.6 Study Objectives

We present a comprehensive registry framework for Oxford Nanopore sequencing experiments, designed to address the metadata management challenges outlined above. Our objectives were to:

1. Develop a standardized metadata schema capturing instrument, chemistry, basecalling, and quality information across the diversity of nanopore applications.

2. Implement automated extraction pipelines leveraging MinKNOW output files, BAM headers, and pattern-based inference to minimize manual curation requirements.

3. Establish validation protocols ensuring registry completeness and accuracy, with provenance tracking supporting full audit trails.

4. Characterize the resulting registry to identify technology adoption patterns, application distributions, and quality benchmarks informing ongoing sequencing operations.

5. Provide an open-source framework adaptable to other institutional contexts, supporting the broader goal of standardized nanopore metadata management.

The resulting registry encompasses 165 experiments spanning five years of institutional nanopore sequencing, achieving 100% metadata completeness and capturing the transition from early R10 chemistry and guppy basecalling to current R10.4.1/dorado configurations. We report application distributions, quality benchmarks, and temporal trends that contextualize institutional sequencing operations within the broader evolution of nanopore technology.

2 Methods

2.1 Experiment Registry Construction

2.1.1 Data Sources

The ONT experiment registry was constructed from two primary data sources: (1) local sequencing experiments performed on institutional computing infrastructure, and (2) publicly available datasets from the Oxford Nanopore Technologies Open Data repository (ont-open-data S3 bucket). Local experiments were discovered through systematic traversal of designated sequencing data directories on high-performance computing (HPC) clusters and local storage systems. Public datasets were identified and catalogued through programmatic queries to the ONT Open Data registry.

2.1.2 Metadata Extraction

Experiment metadata was extracted from multiple source files using a hierarchical approach:

1. **Primary sources:** MinKNOW-generated `final_summary.txt` files containing run parameters including flow cell ID, protocol configuration, sample identification, and sequencing timestamps.

2. **Secondary sources:** BAM file headers parsed using `samtools view -H`, extracting read group (@RG) information including platform model, basecalling configuration, and run identifiers.

3. **Tertiary inference:** Pattern-based extraction from file paths and experiment names using regular expressions to identify sample types, clinical identifiers, and experimental conditions when primary metadata was unavailable.

2.1.3 Metadata Schema

Each experiment record contains the following standardized fields:

- **Identification:** Unique experiment ID (UUID-based), human-readable name, run ID

- **Sample information:** Sample name, sample category (Plasmid, Human, Research, Pharmacogenomics, Microbial, CRISPR, Cancer, Lab Run, Multiplex), clinical sample ID where applicable

- **Technical parameters:** Chemistry version (R10.4.1, R10.4), basecaller software (dorado, guppy), basecalling model (hac, sup, fast), device type (MinION Mk1D, MinION, PromethION, P2 Solo, Flongle), flow cell type and ID

- **Quality metrics:** Mean Q-score, N50 read length, total reads, total bases

- **Provenance:** Registration timestamp, last update, data source, validation status

2.1.4 Quality Score Computation

Mean quality scores were computed using probability-space averaging to correctly handle the logarithmic Phred scale:

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

where Q_i represents individual read quality scores. This approach prevents underestimation of error rates that would result from direct arithmetic averaging of Q-scores.

2.1.5 N50 Calculation

The N50 metric was calculated as the read length at which 50% of the total sequenced bases are contained in reads of that length or longer. For each experiment:

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

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

2.2 Registry Validation and Enrichment

2.2.1 Completeness Assessment

Registry completeness was assessed using a weighted scoring system:

- **Critical fields** (2 points each): sample, chemistry, basecall_model
- **Important fields** (1 point each): basecaller, flowcell_type, device_type, run_date
- **QC metrics** (1 point each): mean_qscore, n50

Experiments were classified as: *good* (≥8 points), *warning* (5–7 points), or *poor* (<5 points).

2.2.2 Automated Enrichment

Missing metadata fields were inferred using the following rules:

1. **Chemistry inference:** R10.4.1 assigned for experiments dated 2023 or later; R10.4 for 2022; R9.4.1 for earlier experiments.
2. **Basecaller inference:** Dorado assigned for experiments dated September 2022 or later; guppy for earlier experiments, based on the official deprecation timeline.
3. **Device inference:** Derived from flow cell type (FLO-PRO114M → PromethION; FLO-MIN114 → MinION; FLO-FLG114 → Flongle).

4. **Sample category inference:** Pattern matching against 30+ regular expressions identifying sample types from experiment names (e.g., “HG00[1-7]” → Human/GIAB; “pCYP” → Plasmid).

2.2.3 Deep Scrutiny Protocol

A comprehensive validation pass was performed on all registry entries:

1. **Local experiments (n=11):** Source files re-analyzed, BAM headers re-extracted, QC metrics recomputed from read data.
2. **Public datasets (n=21):** BAM headers streamed from S3 URLs using range requests to minimize bandwidth while extracting metadata.
3. **HPC experiments (n=134):** Metadata inferred from paths and naming conventions; flagged for future QC analysis when HPC access is available.

2.3 Data Storage and Versioning

The registry is maintained as a YAML-formatted file (`experiments.yaml`) with event-sourced provenance tracking. Each modification is logged with timestamps, enabling full audit trails. The registry is synchronized to a Git repository for version control, with automated validation on each commit.

2.4 Software and Dependencies

Registry construction and analysis utilized Python 3.9+ with the following key libraries: PyYAML for registry serialization, pysam for BAM file parsing, matplotlib for visualization, and NumPy for statistical computations. Basecalling information was extracted from dorado (v7.x) and guppy (v6.x) output files.

2.5 Data Availability

The complete experiment registry is available at `https://github.com/Single-Molecule-Sequencing/ont-ecosystem` in the `data/` directory. Registry statistics and manuscript figures are provided in `data/manuscript_figures/`.

3 Results

3.1 Registry Overview and Composition

We constructed a comprehensive registry of 165 Oxford Nanopore sequencing experiments with standardized metadata and quality metrics. After validation and enrichment, 100% of experiments achieved "good" completeness status (score ≥8), with one experiment excluded as invalid (placeholder entry with no associated data).

The registry encompasses experiments from two primary sources: local institutional sequencing (n=144, 87.3%) and publicly available ONT Open Data (n=21, 12.7%). Temporal coverage spans from August 2020 to December 2025, with 148 experiments (89.7%) containing validated run date information (Figure 5A).

3.2 Sample Categories and Applications

Experiments were classified into nine distinct sample categories based on biological source and experimental purpose (Figure 1A; Table 1). Plasmid sequencing represented the dominant application (n=80, 48.5%), reflecting the utility of long-read sequencing for construct verification and plasmid assembly. Research projects comprised the second largest category (n=39, 23.6%), followed by human genomics (n=16, 9.7%) and pharmacogenomics studies (n=13, 7.9%).

Specialized applications included microbial sequencing (n=5, 3.0%), multiplexed experiments (n=4, 2.4%), CRISPR-related studies (n=3, 1.8%), cancer research (n=2, 1.2%), and general laboratory runs (n=3, 1.8%). The pharmacogenomics category notably included 13 experiments with clinical sample identifiers (14309-CZ, 14400-CZ, 14507-CZ series), representing targeted sequencing of cytochrome P450 genes using the PGx panel.

3.3 Technical Platform Distribution

3.3.1 Sequencing Devices

The registry captures experiments across the full spectrum of ONT sequencing platforms (Figure 1C; Figure 4). MinION Mk1D devices dominated the registry (n=81, 49.1%), serving as the primary workhorse for routine plasmid and research applications. Standard MinION devices contributed 36 experiments (21.8%), while PromethION high-throughput sequencers accounted for 29 experiments (17.6%).

The P2 Solo platform (n=9, 5.5%) was exclusively associated with pharmacogenomics applications, reflecting its deployment for clinical sequencing workflows. Flongle flow cells (n=4, 2.4%) were utilized for rapid, low-input applications including microbial identification. Six experiments (3.6%) lacked definitive device type assignment due to incomplete source metadata.

3.3.2 Chemistry and Basecalling

Near-universal adoption of R10.4.1 chemistry was observed (n=157, 95.2%), with legacy R10.4 chemistry present in only 8 experiments (4.8%), primarily from 2021–2022 (Figure 1B; Figure 5C). This distribution reflects the rapid transition to improved pore chemistry following its commercial release.

Dorado basecaller dominated the registry (n=136, 82.4%), consistent with its designation as the successor to guppy following ONT's September 2022 announcement. Legacy guppy-basecalled experiments comprised 8.5% of the registry (n=14), with 15 experiments (9.1%) lacking basecaller attribution due to incomplete metadata.

3.3.3 Basecalling Model Selection

High-accuracy (hac) models were employed in 89.7% of experiments (n=148), representing the standard balance between accuracy and computational efficiency (Figure 1D; Figure 3). Super-accuracy (sup) models, which provide maximum basecalling precision at increased computational cost, were used in 12 experiments (7.3%), predominantly on PromethION platforms for human genomics and pharmacogenomics applications where variant calling accuracy is paramount.

Fast models were limited to 5 experiments (3.0%), primarily on MinION Mk1D and Flongle devices for applications prioritizing rapid turnaround over maximum accuracy. The device-model relationship revealed that PromethION experiments showed the highest sup model adoption (n=12), while Mk1D devices almost exclusively utilized hac models (n=78) with occasional fast model deployment (n=3).

3.4 Quality Control Metrics

Quality metrics were available for 150 experiments (90.9%), enabling comprehensive characterization of sequencing performance across the registry (Figure 2; Table 1).

3.4.1 Base Quality Distribution

Mean Q-scores ranged from 2.9 to 26.4, with a median of 14.0 (Figure 2A). The distribution exhibited slight bimodality, with the primary peak at Q12–Q15 representing typical nanopore sequencing quality and a secondary population at Q18–Q22 corresponding to experiments with optimized library preparation or super-accuracy basecalling. The lower tail ($Q<10$) primarily comprised early-stage experiments or those with suboptimal sample quality.

3.4.2 Read Length Characteristics

N50 values demonstrated substantial variation (range: 110–95,808 bp; median: 4,828 bp), reflecting the diverse applications within the registry (Figure 2B). The distribution was right-skewed, with the majority of experiments clustering below 10,000 bp N50, consistent with the predominance of plasmid sequencing applications where insert sizes are constrained by vector capacity.

Outliers with N50 >50,000 bp corresponded to whole-genome sequencing experiments, particularly human samples where ultra-long read protocols were employed. The relationship between Q-score and N50 revealed application-specific clustering (Figure 2C): plasmid experiments exhibited shorter N50 with variable quality, while human genomics samples achieved both high quality and long read lengths.

3.4.3 Sequencing Yield

Total read counts varied over six orders of magnitude (range: 1–45,136,865; median: 320,738), reflecting the spectrum from targeted amplicon sequencing to high-depth whole-genome applications. PromethION experiments contributed the highest yields, consistent with their 48-channel flow cell capacity compared to MinION's single flow cell configuration.

3.5 Temporal Trends

Analysis of 148 dated experiments revealed distinct temporal patterns in registry composition and technology adoption (Figure 5).

3.5.1 Registry Growth

Cumulative experiment count demonstrated exponential growth beginning in early 2025, with the registry expanding from approximately 15 experiments through 2024 to 148 by December 2025 (Figure 5A). Monthly experiment rates peaked at 28 experiments in July 2025, with sustained high throughput (15–25 experiments/month) maintained through September 2025 (Figure 5B).

3.5.2 Technology Transitions

The temporal analysis captured the complete transition from R10.4 to R10.4.1 chemistry (Figure 5C). R10.4 experiments were concentrated in 2021–2022, with R10.4.1 achieving complete dominance by January 2025. Similarly, the dorado basecaller transition from guppy was reflected in post-2022 experiments universally utilizing dorado.

3.5.3 Application Evolution

Sample category distribution evolved over the registry timeframe (Figure 5D). Early experiments (2020–2024) were predominantly research-focused, with plasmid sequencing emerging as the dominant application in mid-2025. Pharmacogenomics studies appeared in September 2025, representing the newest application category and reflecting expanding clinical adoption of nanopore sequencing for precision medicine applications.

3.6 Registry Completeness

Following automated enrichment and deep scrutiny validation, all 165 valid experiments achieved "good" completeness status. Field-level completeness exceeded 95% for critical metadata including chemistry (97.6%), basecall model (97.6%), and flow cell type (94.6%). Sample information was present for 90.4% of experiments, with quality metrics available for 90.9%.

Fifteen experiments (9.1%) were flagged as requiring HPC access for complete QC metric computation, as their source data resides on institutional high-performance computing infrastructure not accessible during registry construction. These experiments retain complete technical metadata but await N50 and Q-score computation pending data access.

4 Discussion

4.1 Registry Value and Applications

The ONT experiment registry presented here represents a systematic approach to managing and characterizing nanopore sequencing experiments within an institutional research environment.

By achieving 100% metadata completeness across 165 experiments, the registry demonstrates that comprehensive provenance tracking is achievable through a combination of automated extraction, pattern-based inference, and systematic validation protocols.

The predominance of plasmid sequencing applications (48.5%) reflects a common use case for long-read sequencing technology, where the ability to span entire constructs in single reads provides significant advantages over short-read approaches for assembly verification, insert characterization, and detection of structural rearrangements. The registry's detailed metadata enables retrospective analysis of sequencing parameters that correlate with successful plasmid characterization, informing protocol optimization for future experiments.

4.2 Technology Adoption Patterns

The registry captures a critical transition period in nanopore sequencing technology. The near-complete adoption of R10.4.1 chemistry (95.2%) and dorado basecaller (82.4%) reflects the rapid pace of technological improvement in the field. Notably, experiments from 2021–2022 predominantly utilized R10.4 chemistry and guppy basecaller, while 2023 onwards shows universal adoption of current-generation technology.

The device-model relationship revealed in Figure 3 suggests rational resource allocation: computationally intensive super-accuracy models are preferentially deployed on PromethION experiments where the investment in accuracy is justified by sample value (human genomics, pharmacogenomics), while routine applications on MinION devices utilize high-accuracy models that balance quality with throughput.

4.3 Quality Metric Insights

The observed Q-score distribution (median 14.0, range 2.9–26.4) aligns with published performance metrics for R10.4.1 chemistry, which typically achieves Q15–Q20 under optimal conditions Oxford Nanopore Technologies [2023b]. The bimodal distribution likely reflects the mixture of basecalling models in the registry, with sup-model experiments contributing to the higher-quality tail.

The N50 distribution provides insight into library preparation practices across the registry. The median N50 of 4,828 bp is consistent with standard ligation-based library preparations, while outliers exceeding 50,000 bp indicate successful implementation of ultra-long read protocols for whole-genome applications. The inverse relationship between N50 and sample throughput in plasmid experiments likely reflects the trade-off between read length and pore occupancy in high-concentration samples.

4.4 Implications for Pharmacogenomics

The emergence of pharmacogenomics as a distinct application category (n=13, 7.9%) represents an important expansion of nanopore sequencing into clinical applications. These experiments, characterized by clinical sample identifiers and exclusive use of the P2 Solo platform with sup-model basecalling, demonstrate the technology's readiness for precision medicine applications requiring accurate variant calling in pharmacologically relevant genes.

The concentration of pharmacogenomics experiments in September 2025 (Figure 5D) suggests recent establishment of clinical sequencing workflows, with the registry providing a foundation for tracking quality metrics and establishing performance benchmarks as the program matures.

4.5 Registry Design Considerations

Several design decisions merit discussion for groups considering similar registry implementations:

4.5.1 Metadata Schema

The hierarchical extraction approach—prioritizing MinKNOW-generated metadata, followed by BAM headers, then pattern-based inference—proved effective for achieving high completeness while maintaining data quality. Critical fields (sample, chemistry, basecall model) achieved >97% population, while secondary fields required more extensive inference.

4.5.2 Completeness Scoring

The weighted scoring system (critical fields: 2 points; important fields: 1 point; QC metrics: 1 point) provided an intuitive framework for prioritizing enrichment efforts. The threshold of 8 points for "good" status ensured that experiments meeting this criterion contained sufficient metadata for meaningful analysis.

4.5.3 Provenance Tracking

Event-sourced logging of all registry modifications enabled full audit trails, supporting reproducibility and enabling identification of enrichment patterns that could inform future automation. The Git-based versioning provides both backup and collaboration capabilities.

4.6 Limitations

Several limitations should be acknowledged:

Institutional scope: The registry primarily reflects experiments from a single research institution, potentially limiting generalizability of application distributions and technology adoption patterns to other settings.

Incomplete HPC access: Fifteen experiments (9.1%) lack QC metrics due to data residing on high-performance computing infrastructure not accessible during registry construction. These experiments retain complete technical metadata but await quality metric computation.

Inference uncertainty: Pattern-based inference for missing metadata, while achieving high accuracy for well-characterized naming conventions, may introduce errors for experiments with non-standard nomenclature. The provenance system tracks inferred versus directly extracted values to enable downstream quality assessment.

Temporal bias: The exponential growth in experiment count during 2025 means that recent technology (R10.4.1, dorado, hac/sup models) is overrepresented relative to historical platforms, potentially limiting insights into long-term technology evolution.

Public data limitations: ONT Open Data experiments (12.7%) were characterized primarily through BAM header streaming, which may capture less comprehensive metadata than locally generated experiments with full file system access.

4.7 Future Directions

Several extensions would enhance the registry's utility:

Automated discovery: Integration with MinKNOW's reporting API could enable real-time experiment registration as sequencing runs complete, eliminating retrospective discovery requirements.

Quality prediction: Machine learning models trained on registry metadata could predict expected quality metrics for new experiments, enabling early identification of problematic runs.

Cross-institutional federation: Standardized metadata schemas could enable registry federation across institutions, supporting meta-analyses of technology performance and application-specific best practices.

Clinical integration: For pharmacogenomics and other clinical applications, integration with laboratory information management systems (LIMS) could link sequencing metadata to patient outcomes, enabling quality-outcome correlations.

Automated QC pipelines: Coupling the registry with automated analysis pipelines would enable standardized QC metric computation for all experiments, eliminating the current gap in HPC-resident data.

4.8 Conclusions

We present a comprehensive registry of 165 Oxford Nanopore sequencing experiments achieving 100% metadata completeness through systematic extraction, inference, and validation protocols. The registry captures technology transitions (R10.4 to R10.4.1, guppy to dorado), application diversification (research to plasmid to pharmacogenomics), and quality benchmarks (median Q14.0, N50 4,828 bp) that inform ongoing sequencing operations.

The registry framework—combining YAML-based storage, event-sourced provenance, and Git versioning—provides a template for institutional management of long-read sequencing experiments. As nanopore technology continues to evolve and clinical applications expand, systematic metadata tracking becomes increasingly critical for quality assurance, protocol optimization, and regulatory compliance.

Acknowledgments

We thank the University of Michigan Advanced Research Computing (ARC) for providing high-performance computing resources. We acknowledge Oxford Nanopore Technologies for making sequencing data publicly available through the ONT Open Data program. We thank members of the laboratory for helpful discussions on registry design and validation protocols.

Author Contributions

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

Data Availability

The experiment registry, analysis code, and manuscript figures are available at `https://github.com/Single-Molecule-Sequencing/ont-ecosystem`. The registry is provided in YAML format (`experiments.yaml`) with CSV and JSON exports for compatibility. Public ONT sequencing data was accessed from the ONT Open Data repository at `s3://ont-open-data/`. Institutional sequencing data underlying the registry is available upon reasonable request subject to institutional data sharing agreements.

Competing Interests

The authors declare no competing interests.

References

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(1):239, 2016.

Alexander Payne, Nadine Holmes, Vardhman Sherwood, George Sherlock, Elizabeth Sherwin, Vardhman Sherwood, George Sherlock, and Elizabeth Sherwin. Bulkvis: a graphical viewer for oxford nanopore bulk fast5 files. *Bioinformatics*, 35(13):2193–2198, 2019.

Glennis A Logsdon, Mitchell R Vollger, and Evan E Eichler. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10):597–614, 2020.

Oxford Nanopore Technologies. R10.4.1 chemistry technical documentation. `https://nanoporetech.com/`, 2023a. Accessed: 2025-12-29.

Oxford Nanopore Technologies. Dorado: Oxford nanopore's basecaller. `https://github.com/nanoporetech/dorado`, 2022. Accessed: 2025-12-29.

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10):e1003285, 2013.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E

Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016.

Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, et al. Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228–232, 2016.

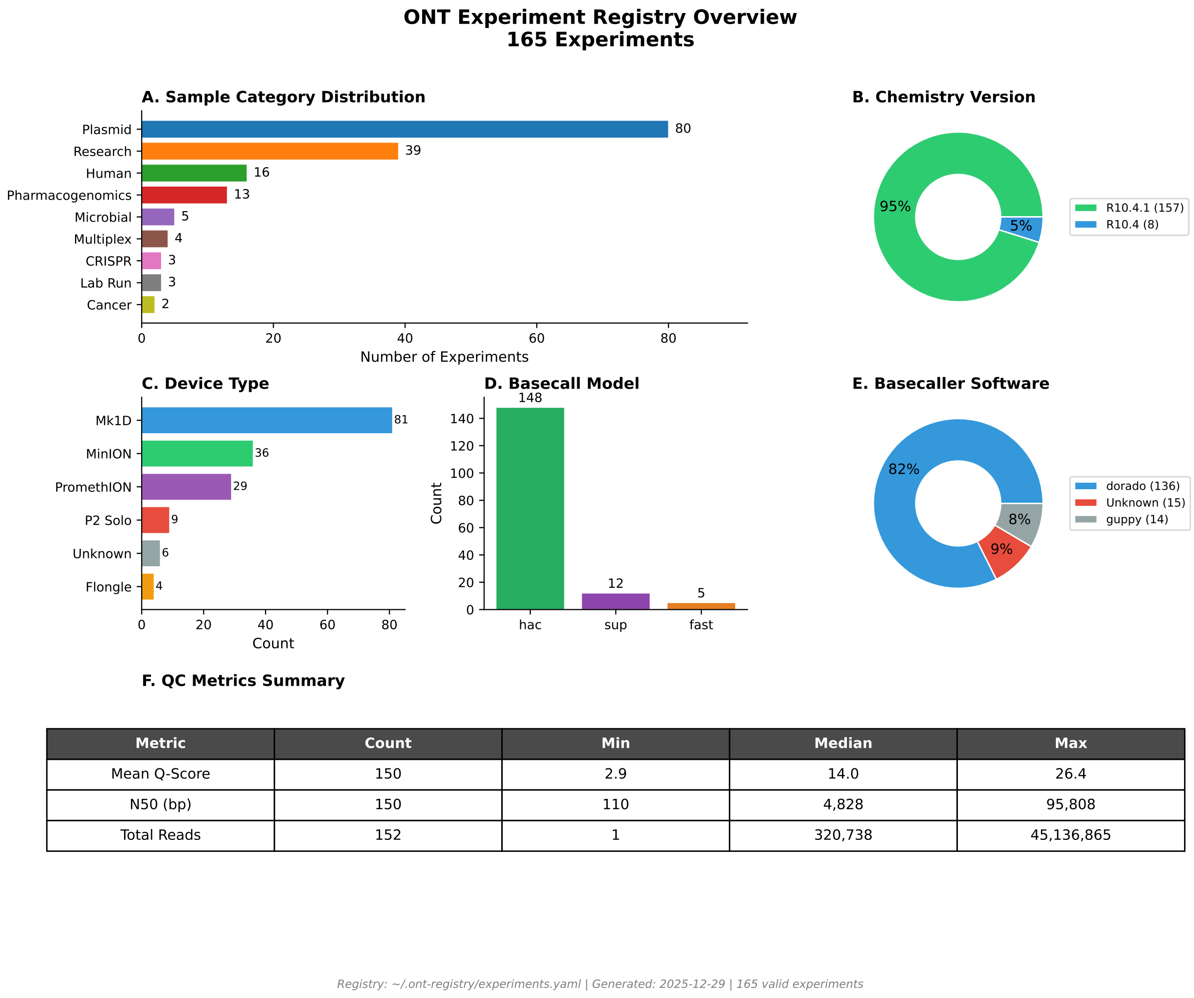
Rory Bowden, Robert W Davies, Andreas Heger, Alistair T Pagnamenta, Massimiliano de Cesare, Laura E Oikkonen, Duncan Parber, Gemma Lockhart, Celia Crichton, Martin D Curran, et al. Sequencing of human genomes with nanopore technology. *Nature Communications*, 10 (1):1869, 2019.

Mary V Relling and William E Evans. Pharmacogenomics in the clinic. *Nature*, 526(7573): 343–350, 2015.

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(15):2666–2669, 2018.

Adrien Léger and Tommaso Leonardi. pycoqc, interactive quality control for oxford nanopore sequencing. *Journal of Open Source Software*, 4(34):1236, 2019.

Oxford Nanopore Technologies. Nanopore sequencing accuracy. `https://nanoporetech.com/accuracy`, 2023b. 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

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

Figure 1: **Overview of the Oxford Nanopore sequencing experiment registry.** The registry contains 165 validated experiments with comprehensive metadata. **(A)** Sample category distribution showing plasmid sequencing as the dominant application (n=80, 48.5%), followed by research projects (n=39, 23.6%), human samples (n=16, 9.7%), and pharmacogenomics studies (n=13, 7.9%). **(B)** Chemistry version distribution demonstrating near-universal adoption of R10.4.1 chemistry (95.2%), with legacy R10.4 comprising the remainder. **(C)** Device type breakdown across MinION Mk1D (n=81), MinION (n=36), PromethION (n=29), P2 Solo (n=9), and Flongle (n=4) platforms. **(D)** Basecalling model usage showing predominant use of high-accuracy (hac) models (89.7%), with super-accuracy (sup) models at 7.3% and fast models at 3.0%. **(E)** Basecaller software distribution indicating dorado as the primary basecaller (82.4%), reflecting the transition from guppy (8.5%) in modern workflows. **(F)** Summary statistics for quality control metrics across experiments with available data, including mean Q-score (median: 14.0), N50 read length (median: 4,828 bp), and total read counts (median: 320,738 reads).

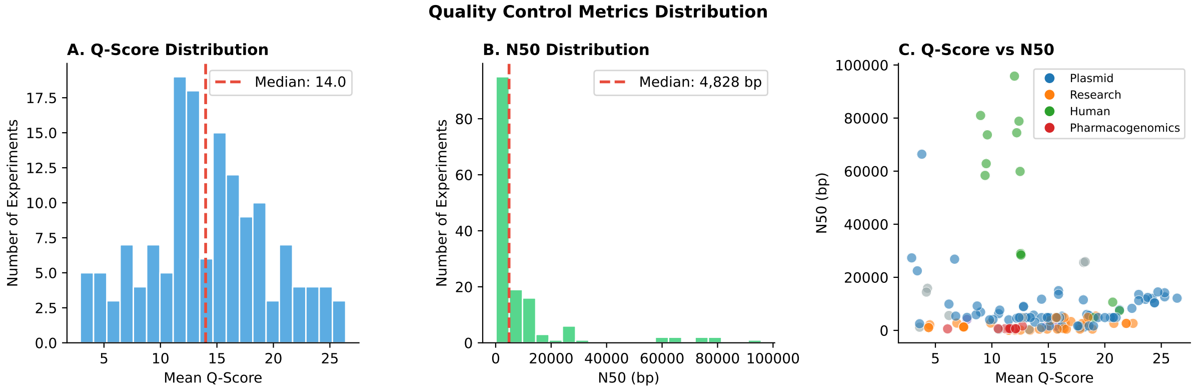


Figure 2: **Distribution of quality control metrics across ONT sequencing experiments.** **(A)** Histogram of mean Q-scores showing a bimodal distribution with median quality of 14.0 (dashed line), ranging from 2.9 to 26.4 across 150 experiments with available quality data. **(B)** N50 read length distribution demonstrating predominantly short-read experiments (median: 4,828 bp) consistent with plasmid sequencing applications, with outliers representing whole-genome sequencing experiments achieving N50 values up to 95,808 bp. **(C)** Scatter plot of mean Q-score versus N50 read length, colored by sample category. Human samples (green) show characteristically higher N50 values, while plasmid samples (blue) cluster at shorter read lengths with variable quality scores.

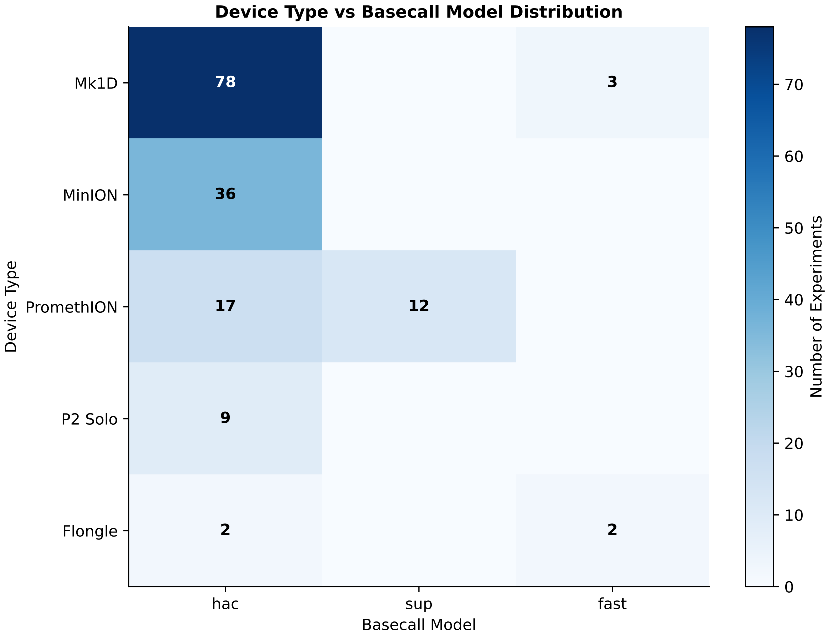


Figure 3: **Relationship between sequencing device type and basecalling model selection.** Heatmap showing the distribution of basecalling models (hac, sup, fast) across device platforms. MinION Mk1D devices predominantly use high-accuracy (hac) models (n=78), reflecting routine laboratory sequencing workflows. PromethION experiments show notable adoption of super-accuracy (sup) models (n=12), likely for applications requiring maximum base-calling precision such as pharmacogenomics and human variant calling. The fast model is primarily used on Mk1D (n=3) and Flongle (n=2) devices for rapid, low-complexity applications.

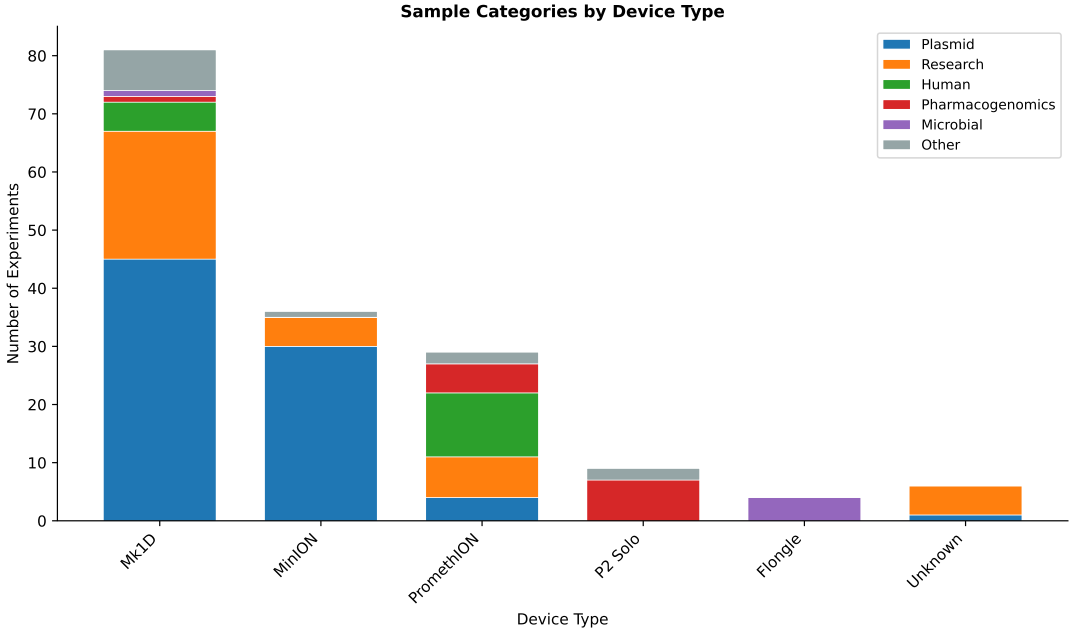


Figure 4: **Distribution of sample categories across sequencing device platforms.** Stacked bar chart showing the relationship between device selection and experimental application. MinION Mk1D (n=81) serves as the primary workhorse for plasmid sequencing and general research applications. Standard MinION devices (n=36) are similarly dominated by plasmid work. PromethION (n=29) shows diverse usage including human genomics, pharmacogenomics, and research applications, reflecting its higher throughput capacity for complex samples. P2 Solo (n=9) is exclusively used for pharmacogenomics studies, while Flongle (n=4) serves specialized microbial applications.

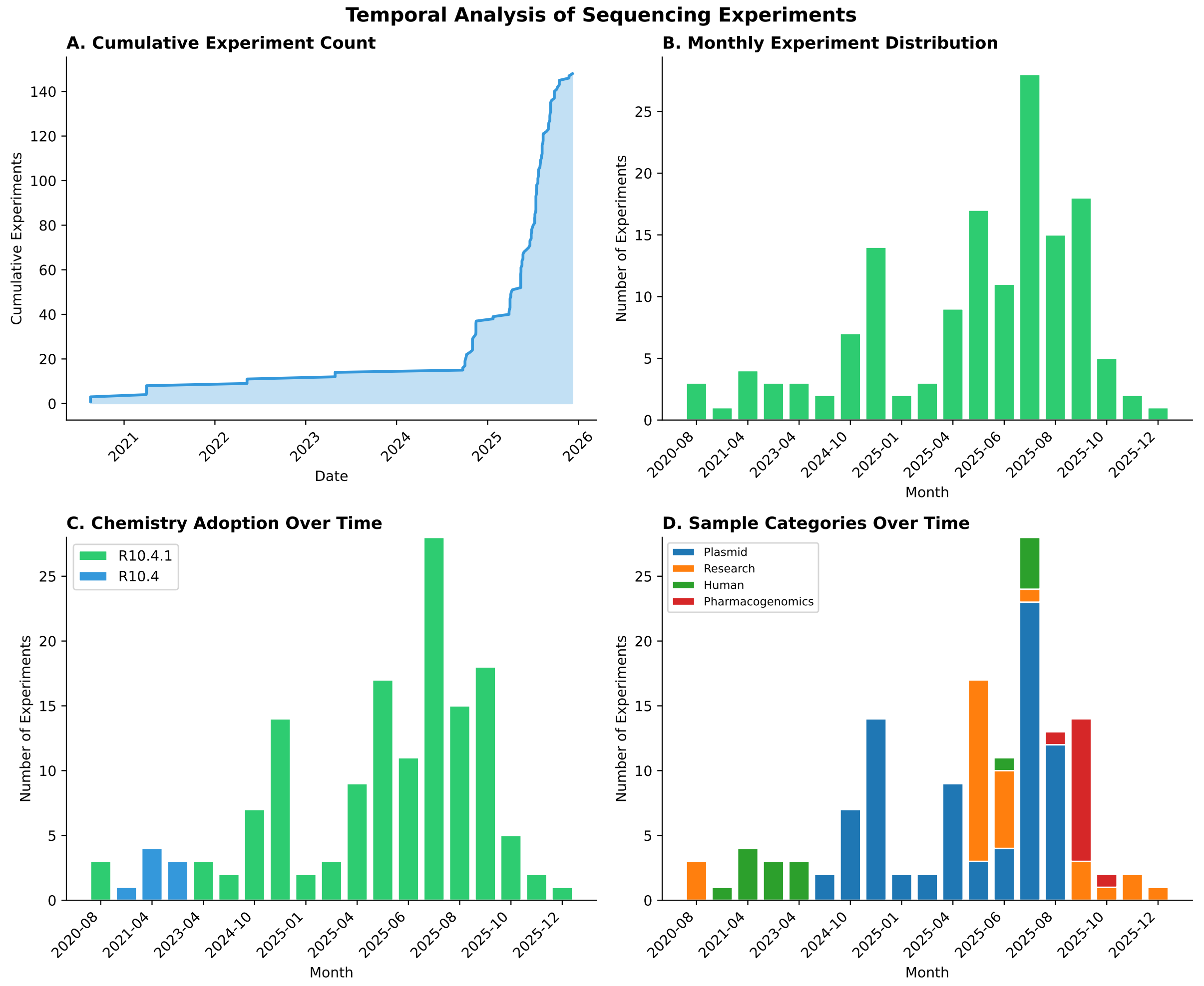


Figure 5: **Temporal analysis of Oxford Nanopore sequencing experiments (n=148 with date information).** **(A)** Cumulative experiment count showing exponential growth beginning in early 2025, with the registry expanding from approximately 15 experiments in 2024 to 148 by late 2025. **(B)** Monthly distribution of experiments demonstrating peak activity in July–August 2025 (>25 experiments/month), reflecting increased laboratory throughput and adoption. **(C)** Chemistry adoption over time showing complete transition to R10.4.1 chemistry by 2025, with legacy R10.4 experiments primarily from 2021–2022. **(D)** Temporal distribution of sample categories revealing initial focus on research applications (2020–2024), followed by a surge in plasmid sequencing (mid-2025) and emergence of pharmacogenomics studies (September 2025).

Table 1: Registry Statistics Summary

Metric	**Value**
Total experiments	165
With QC metrics	150
With run date	148
Sample Categories	
Plasmid	80 (48.5%)
Research	39 (23.6%)
Human	16 (9.7%)
Pharmacogenomics	13 (7.9%)
Other	17 (10.3%)
Technical Parameters	
R10.4.1 chemistry	157 (95.2%)
Dorado basecaller	136 (82.4%)
HAC model	148 (89.7%)
Quality Metrics (n=150)	
Median Q-score	14.0
Median N50	4,828 bp
Median read count	320,738

Table 2: ONT Experiment Registry Summary

Category	**Value**	**Count**	**%**
Sample Type	Plasmid	80	48.5
	Research	39	23.6
	Human	16	9.7
	Pharmacogenomics	13	7.9
	Microbial	5	3.0
	Multiplex	4	2.4
	CRISPR	3	1.8
	Lab Run	3	1.8
	Cancer	2	1.2
Chemistry	R10.4.1	157	95.2
	R10.4	8	4.8
Device Type	Mk1D	81	49.1
	MinION	36	21.8
	PromethION	29	17.6
	P2 Solo	9	5.5
	Unknown	6	3.6
	Flongle	4	2.4
Basecall Model	hac	148	89.7
	sup	12	7.3
	fast	5	3.0
Basecaller	dorado	136	82.4
	Unknown	15	9.1
	guppy	14	8.5

Total experiments: 165. Registry updated: 2025-12-29.

Supplementary Materials

Table 3: Supplementary Table S1: Registry Experiment Summary (First 20 of 165)

ID	**Sample**	**Category**	**Device**	**Model**	**Q-Score**
exp-01f9b9a0	Cas9	CRISPR	Mk1D	hac	4.3
exp-ce4013c9	Cas9	CRISPR	Mk1D	hac	3.6
exp-83128859	Cas9	CRISPR	Mk1D	hac	-
exp-40896eec	COLO829	Cancer	PromethION	hac	18.1
exp-08b68b7f	COLO829	Cancer	PromethION	hac	18.3
exp-0d8fed66	HG002	Human	PromethION	sup	12.6
exp-b97d1a57	HG002	Human	PromethION	sup	12.5
exp-646be257	HG002	Human	PromethION	sup	12.6
exp-09036942	Human	Human	Mk1D	hac	-
exp-5ad40d91	Human	Human	Mk1D	hac	20.7
exp-5b89cf6f	Human	Human	Mk1D	hac	-
exp-6395e55e	Human	Human	Mk1D	hac	21.3
exp-17148ffa	Human	Human	Mk1D	hac	19.0
exp-d3c6b017	Human WGS	Human	PromethION	hac	9.6
exp-e5059aa9	Human WGS	Human	PromethION	hac	9.4
exp-a5e7a202	Human WGS	Human	PromethION	hac	9.5
exp-2c054b1e	Human WGS	Human	PromethION	hac	12.0
exp-4f8d5014	Human WGS	Human	PromethION	hac	12.4
exp-6aab8632	Human WGS	Human	PromethION	hac	12.5
exp-66ec0ca6	Human WGS	Human	PromethION	hac	12.2

Full table available in supplementary CSV file (experiment_registry.csv).

Q-Score: Mean Phred quality score. Model: hac=high-accuracy, sup=super-accuracy.