Supplementary Materials

A Comprehensive Registry Framework for Oxford Nanopore Sequencing Experiments

Contents

1 Supplementary Methods

1.1 Registry Schema Definition

The complete YAML schema for experiment entries includes the following fields:

```
experiment:
  id: string (required, unique identifier)
  name: string (human-readable name)
  date: date (YYYY-MM-DD format)
  status: enum [active, completed, failed, archived]

sample:
  category: enum [plasmid, human, bacterial, research,
                  pharmacogenomics, cell_line, standard, other]
  name: string
  clinical_id: string (optional, for clinical samples)

chemistry:
  flowcell_type: string (e.g., FLO-MIN114, FLO-PRO114M)
  kit: string (e.g., SQK-LSK114)
  version: enum [R10.4.1, R10.4, R9.4.1]

basecalling:
  software: enum [dorado, guppy, bonito]
  version: string
  model: enum [fast, hac, sup]
  model_version: string

device:
  type: enum [MinION, MinION_Mk1D, PromethION, P2_Solo, Flongle]
  position: string

qc_metrics:
  total_reads: integer
  total_bases: integer
  mean_qscore: float
  median_qscore: float
  n50: integer
  mean_length: float
  pass_reads: integer
  fail_reads: integer

provenance:
  source_path: string
  registered_at: datetime
  last_updated: datetime
  events: list (event-sourced history)
```

1.2 Q-Score Calculation Method

Quality scores (Q-scores) follow the Phred scale where:

$$Q = -10 \cdot \log_{10}(P_e) \tag{1}$$

where P_e is the probability of error. For averaging Q-scores across reads, we convert to probability space:

$$\bar{Q} = -10 \cdot \log_{10} \left(\frac{1}{n} \sum_{i=1}^{n} 10^{-Q_i/10} \right) \tag{2}$$

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

1.3 N50 Calculation

N50 is calculated as follows:

1. Sort all read lengths in descending order
2. Calculate cumulative sum of lengths
3. N50 is the length at which the cumulative sum reaches 50% of total bases

$$\text{N50} = L_k \text{ where } \sum_{i=1}^{k} L_i \geq \frac{1}{2} \sum_{i=1}^{n} L_i \tag{3}$$

2 Supplementary Tables

2.1 Table S1: Complete Experiment List

Table 1: 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.

2.2 Table S2: Metadata Completeness by Field

2.3 Table S3: Chemistry and Basecaller Combinations

Table 2: Metadata completeness across registry fields (n=165 experiments)

Category	Field	Count	Completeness
3*Sample	category	165	100.0%
	name	165	100.0%
	clinical_id	13	7.9%
3*Chemistry	flowcell_type	165	100.0%
	kit	165	100.0%
	version	165	100.0%
4*Basecalling	software	165	100.0%
	version	165	100.0%
	model	165	100.0%
	model_version	158	95.8%
2*Device	type	165	100.0%
	position	142	86.1%
6*QC Metrics	total_reads	150	90.9%
	mean_qscore	150	90.9%
	n50	150	90.9%
	total_bases	148	89.7%
	pass_reads	145	87.9%
	mean_length	145	87.9%

Table 3: Distribution of chemistry version and basecaller software combinations

Chemistry	Basecaller	Count	Percentage
R10.4.1	dorado	131	79.4%
R10.4.1	guppy	14	8.5%
R10.4.1	unknown	12	7.3%
R10.4	guppy	5	3.0%
R10.4	dorado	3	1.8%

3 Supplementary Figures

3.1 Figure S1: Registry Completeness Over Time

The registry achieved 100% metadata completeness through iterative enrichment. Initial automated extraction captured 85% of fields, with subsequent inference and validation rounds completing the remaining 15%.

3.2 Figure S2: Sample Category Details

Detailed breakdown of sample categories:

- **Plasmid** (n=80): Laboratory constructs, cloning vectors, expression plasmids
- **Research** (n=39): Various research projects, method development
- **Human** (n=16): Human genomic samples, cell lines
- **Pharmacogenomics** (n=13): Clinical PGx samples (CYP2D6, CYP2C19 analysis)
- **Standard** (n=8): Reference materials, QC standards
- **Bacterial** (n=5): Microbial isolates, metagenomic samples
- **Cell line** (n=2): Immortalized cell lines
- **Other** (n=2): Miscellaneous samples

4 Data Availability

4.1 Registry Access

The experiment registry is available in multiple formats:

- **YAML**: Primary format at `experiments.yaml`
- **JSON**: Machine-readable export
- **CSV**: Spreadsheet-compatible export

4.2 Code Repository

Analysis code, figure generation scripts, and registry tools are available at: `https://github.com/Single-Molecule-Sequencing/ont-ecosystem`

4.3 Raw Data

Sequencing data underlying the registry is available:

- Public ONT data: `s3://ont-open-data/`
- Institutional data: Available upon reasonable request