MOHCCN SEQUENCING METADATA POLICY V1

Table of Contents

1. Introduction

The Marathon of Hope Cancer Centres Network (MOHCCN) aims to create a "gold-standard" cohort of clinical cancer specimens with a well-annotated, uniformly generated, and consistently quality-controlled dataset (clinical and genomic) from 15,000 (15k) cases collected from across Canada over 5 years. Not only does the MOHCCN aim to build a pan-Canadian Cancer Network and to produce immediate clinical impact by identifying actionable targets through molecular profiling, but it also proposes to generate in-depth molecular profiling data from cancer patient cohorts to address important scientific questions. This richly clinically annotated molecular dataset, starting with standardized clinical information, treatment response data, and whole-genome and transcriptome profiles (WGTS), will serve as an invaluable resource for cancer biology discovery.

The MOHCCN WGTS Guideline states that "Raw and processed WGTS data, QC, and variant reports are the final information products that are deposited and shared." This Sequencing Metadata Policy describes what specific files and metadata need to be submitted to CanDIG for a complete Gold Cohort Case (Tier A or B, see definitions in the Gold Cohort Standards Policy).

The MOHCCN will follow the standards used by the International Nucleotide Sequence Database Collaboration (INSDC). INSDC members – DDBJ (Japan), ENA (EU) and NCBI (USA) – as well as the European Genome-phenome Archive (EGA), which inherits from ENA, use the same base standard for sequencing metadata.

An overview of the various schemas and their relationships from the ENA:

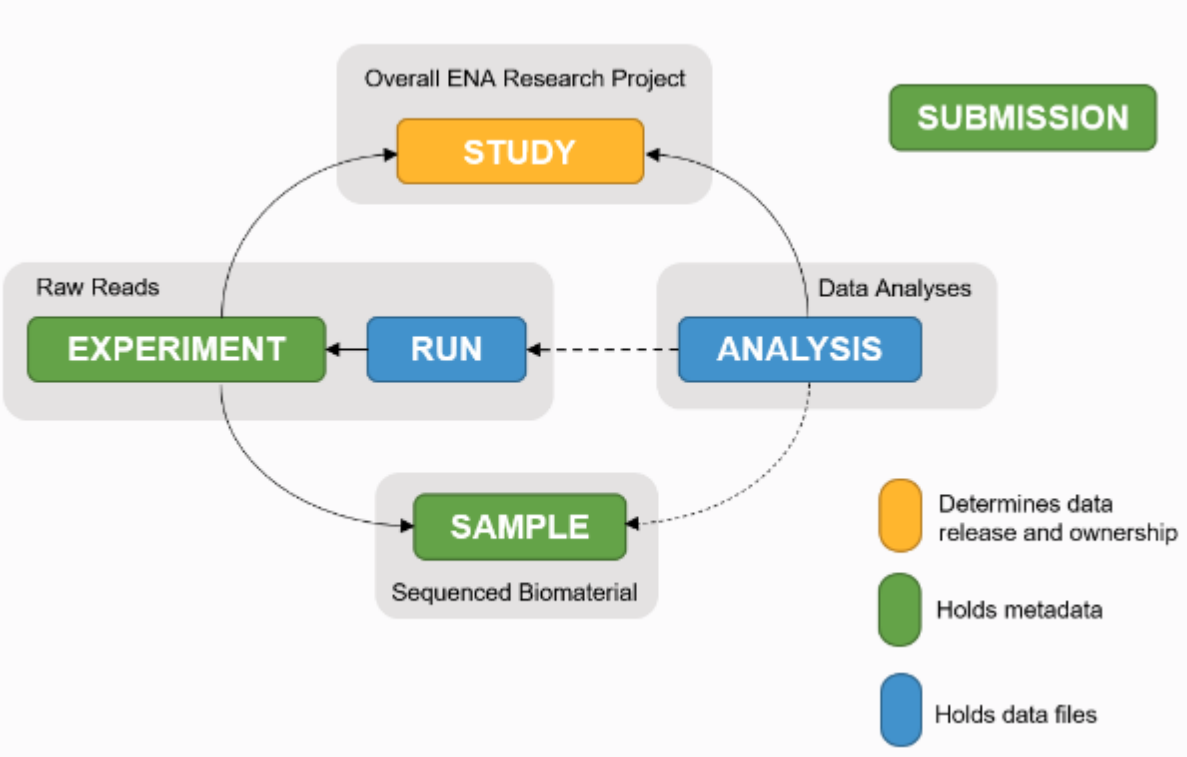

Figure 1: ENA overview

For MOHCCN, the study and sample information is contained in the clinical data model, so this Policy specifies use of the Run, Experiment, and Analysis schemas. For reference, the canonical format of the metadata schemas is XML. See XML documentation in the ENA webin-xml GitHub repo (these are the same files that get pushed to the ENA ftp site).

2. Required Sequencing Files

A Gold Cohort Case requires the following:

- Raw sequence data, paired end (lossless .bam or .fastq.gz) for tumour and normal sample (WGS)
- Raw sequence data, paired end (lossless .bam or .fastq.gz) for tumour sample (WTS)
- Alignments for WGS (.bam and .bai)
- Alignments for WTS (.bam and .bai)
- Gene expression matrices, raw counts (csv or equivalent)
- Mutation calls (.vcf and .tbi)

Outputs from the other analyses are optional, e.g:

- Copy number variants
- Structural variants
- Fusion calls

3. Required Sequencing Metadata

The sequencing metadata will require two Experiments schemas, one for WGS and one for WTS, with a Run schema for the set of raw reads from each Experiment. Each set of processed files requires an Analysis schema. See MOHCCN_sequencing_metadata.xlsx - Google Sheets for a list of fields (objects) and allowable values. These are a subset of the SRA / EGA files relevant to MOHCCN. CanDIG will provide a JSON schema specific to the MOHCCN data as well as validation tools for use during data preparation and ingest.

Document Revision History

Developed by	Reviewed by	Endorsed by	Effective Date	Policy Version	Summary of revisions
TWG & DPSC	Steering Committee	Network Council	August 1, 2025	V1	n/a

Authors

Name	Institution	Title
Marco Marra (TWG Co-Chair)	BC Cancer	Distinguished Scientist / Professor, Medical Genetics and Michael Smith Labs at UBC
Trevor Pugh (TWG Co-Chair)	U of T	Director / Senior Investigator
Ian Watson (TWG Co-Chair)	McGill	Associate Professor
Sorana Morrissy	U of Calgary	Assistant Professor
Thomas Belbin	MUN	Associate Professor
Lincoln Stein (DPSC Co-Chair)	OICR	Head, Adaptive Oncology
Steven Jones (DPSC Co-Chair)	BCGSC	Director of Bioinformatics
Guillaume Bourque	McGill	Professor
Jeff Bruce	UHN	Scientific Associate
Jennifer Chan	U of Calgary	Director / Associate Professor
Karen Cranston	CanDIG	Technical Project Manager
Daniel Gaston	Dalhousie	Lead, Bioinformatician
Benjamin Haibe-Kains	U of Toronto	Associate Professor
Martin Hirst	UBC	Senior Scientist
Anne-Marie Mes-Masson	U of Montreal	Associate Scientific Director
Jessica Nelson	BCGSC	Projects Team Leader
Carolyn-Ann Robinson	U of Calgary	Senior Research Associate
Enrique Sanz-Garcia	UHN	Assistant Professor / Staff Medical Oncologist
Lillian Siu	UHN	Senior Scientist
Dominique Trudel	CHUM	Pathologist / Associate Clinical Professor
Tran Truong	UHN	Director of Data & Technology
Ian Watson	McGill	Associate Professor
Emily Van de Laar	UHN	Project Team Lead
Ma'n Zawati	McGill	Assistant Professor