MOHCCN Whole Genome and Transcriptome Sequencing (WGTS) Guideline _v1

Table of Content

1. Introduction: WGTS Data Processing

Whole Genome and Transcriptome Sequencing (WGTS) projects typically comprise a series of distinct activities (WGTS Project Process), many of which draw on protocols and software deployed as standardized assays (**Figure 1**):

1. **Cohort selection:** implementation of MOHCCN project scope and objectives
2. **Sample Processing**: accessioning, cancer cell enrichment, and nucleic acid extraction
3. **Library preparation and sequencing:** conversion of specimen-derived nucleic acid into a format that can be read by a DNA/RNA sequencer
4. **Primary analysis:** Libraries are loaded onto a DNA/RNA sequencer and the resultant reads are retrieved and quality control performed
5. **Secondary analysis:** Technical execution of bioinformatics pipelines to detect variation in sequencing reads. This step commonly generates read alignments to a reference sequence from which are derived tables of mutations, copy number changes, gene expression levels, methylation levels, etc.
6. **Tertiary analysis:** Biological interpretation, statistical analysis, and summarization of genomic variation. This step commonly generates heatmaps, oncoprints, signature scores, significantly altered genes, etc. More information on the details of Primary, Secondary, and Tertiary analyses are described by Oliver, Hart, and Klee. Clin Chem. 2015. PMID: 25451870.
7. **Data handling and sharing:** Loading of data and analysis into suitable shared databases utilized "FAIR" (findable, accessible, interoperable, and reusable) data principles.

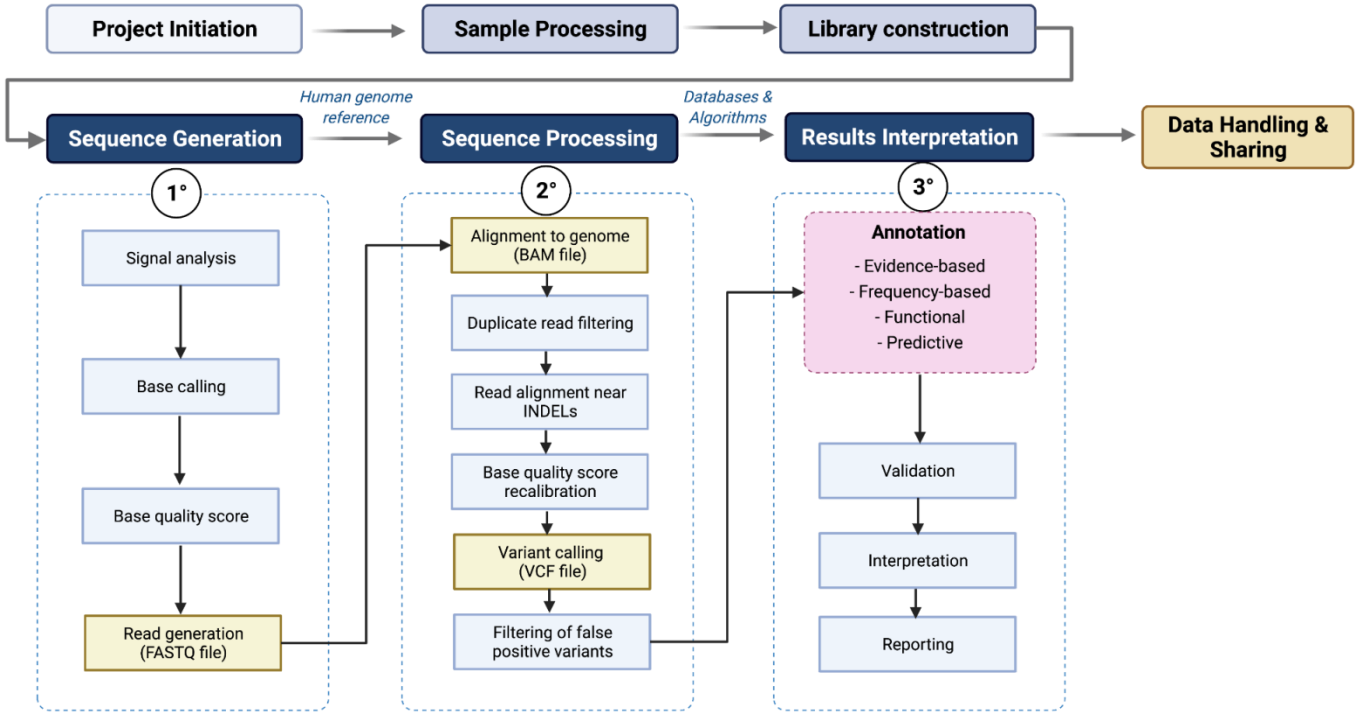

Figure 1. Activities within a WGS Project Process. Adapted from Oliver, Hart, and Klee. Clin Chem. 2015. PMID: 25451870.

2. WGTS General Workflow

This workflow outlines the standard scenario for completing paired normal and tumour WGTS in a MOHCCN case (**Figure 2**). Tissues are first selected via primary checkpoints for adequate quantity and cellularity for nucleic acid extraction. Nucleic acids meeting quality and quantity requirements are used for subsequent whole transcriptome and whole genome library construction, proceeding in parallel. Transcriptome libraries and whole-genome libraries with complete deep sequencing to MOHCCN standard coverage (Whole genome sequencing of tumor (80X minimum, 100X preferred) and matched normal (30X minimum, 40X preferred, Total RNA sequencing using a ribodepletion protocol (80M reads minimum, 100M reads preferred) results in a complete MOHCCN WGTS case (refer to table 1 in section 3.2.1 for minimum and target thresholds). Coverage metrics refer to de-duplicated coverage, and not just the raw coverage estimation obtained by dividing yield and the reference genome length. Any case that fails to proceed through all the technical checkpoints may be discontinued as a MOHCCN case.

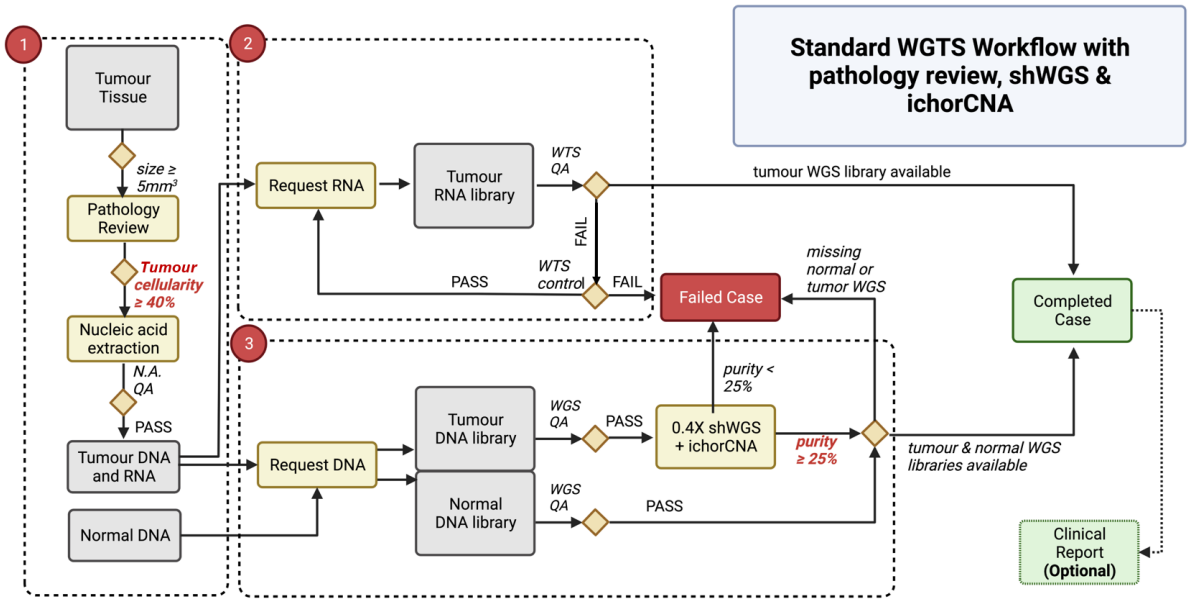

Figure 2. Example WGTS workflow from tissue to complete a case in MOHCCN. (See section 3.2.1 for more details)

3. WGTS Data Guidelines

3.1 Introduction

The MOHCCN 15k Gold Cohort will consist of a mixture of newly generated WGTS data for MOHCCN (following MOHCCN-specific standards) and legacy data. MOHCCN recognizes the value of including legacy data, however, a higher degree of technical and data quality variation and the additional efforts to reprocess and harmonize old with new data creates new challenges for MOHCCN data workflows.

3.2 Newly generated WGTS data for MOHCCN

For all newly generated WGTS data for MOHCCN, samples are sequenced at a minimum 30X normal and 80X tumour coverage. The coverage thresholds are for de-duplicated coverage, not raw coverage.

Raw and processed WGTS data, QC, and variant reports are the final information products that are deposited and shared internally at each site and externally with the network via MOHCCN approved data sharing mechanisms (i.e., DHDP/CanDIG).

3.2.1. Comparison of WGTS QC Parameters across MOHCCN.

WGTS QC parameters for prospective sequencing cases in MOHCCN Genomic Centres (**Table 1**).

QC GATES	MOHCCN CONSORTIA				
	BC2C	PM2C	MOH-Q	PR2C	ACC
Sample receipt					
Initial sample QC	Visual inspection of container integrity & sample volume	Visual inspection of container integrity & sample volume	Visual inspection of container integrity & sample volume	Visual inspection of container integrity & sample volume	Visual inspection of container integrity & sample volume
Receipt temperature appropriate	Receipt temperature appropriate	Receipt temperature appropriate	Receipt temperature appropriate	Receipt temperature appropriate.	Receipt temperature appropriate
The appropriate amount of material for tissues and blood samples	The appropriate amount of material for tissues and blood samples	The appropriate amount of material for tissues and blood samples		The appropriate amount of material for tissues and blood samples	The appropriate amount of material for tissues and blood samples
Container IDs match SSF	Container IDs match SSF	Container IDs match SSF	Container IDs match MGC sample manifest	Container ID match sample submission worksheet	Container ID match sample submission worksheet
DNA volume	25-40uL	>12uL (RUO)	25-100uL	20-60uL	25-100uL
RNA volume	10uL	>12uL (RUO)	10-20uL	20uL	10-20uL
Qubit upper range	120 ng (1X dsDNA HS) 250 ng (Agilent RNA 6000 Nano Kit or PerkinElmer RNA Assay LabChip GX)	120 ng (1X dsDNA HS) 200 ng (1X RNA HS)	120 ng (1X dsDNA HS)	l125 ng/ul (1XdsDNA HS and 1X RNA HS)	120 ng (1X dsDNA HS) 200 ng (1X RNA HS)
Qubit lower range	0.1 ng (1X dsDNA HS)	0.1 ng (1X dsDNA HS) 4 ng (1X RNA HS)	0.1 ng (1X dsDNA HS)	1 ng/ul (1XdsDNA HS and 1X RNA HS)	0.1 ng (1X dsDNA HS) 4 ng (1X RNA HS)

QC GATES	MOHCCN CONSORTIA				
	BC2C	**PM2C**	**MOH-Q**	**PR2C**	**ACC**
	5 ng (Agilent RNA 6000 Nano Kit or PerkinElmer RNA Assay LabChip GX)				
Extraction					
DNA positive controls	Mouse lung FFPE tissue	Sigma Aldrich (Cat# ERMAD442K)		None	None
RNA positive controls	HeLaS3	Fisher Scientific (Cat# AM7155M)		None	None
Appropriate volume	>=25 uL	>=11 uL		>=25uL	>=25uL
Fluorometric quantification	DNA (Buffy coat or whole blood) >=250ng DNA (Flash Frozen tissue) >=250 ng DNA (FFPE) >=100 ng RNA (Flash Frozen tissue/cells) >=50 ng RNA (FFPE) >400 ng	DNA (Buffy coat or whole blood) >=200ng DNA (Flash Frozen tissue) >=50 ng DNA (FFPE) >=100 ng RNA (Flash Frozen tissue/cells) >=60 ng RNA (FFPE) >220 ng		Normal op minimums, and absolute min vals for "precious" samples (add 14% for quant) DNA (PCR-free preps) >=500ng [abs min 100ng] DNA FFPE >= 100ng [abs min 30ng] RNA fresh frozen >= >=50ng [abs min 10ng] RNA FFPE >= 150ng [abs min 20ng] RNA (FFPE) >85-100ng	DNA (Buffy coat or whole blood) >=200ng DNA (Flash Frozen tissue) >=50 ng DNA (FFPE) >=100 ng RNA (Flash Frozen tissue/cells) >=50 ng RNA (FFPE) >200 ng
Library preparation					
Total RNA integrity	Caliper or Agilent Bioanalyzer	Fragment Analyzer or Tapestation	Tapestation	TapeStation	

QC GATES	MOHCCN CONSORTIA				
	BC2C	PM2C	MOH-Q	PR2C	ACC
Batch controls (DNA)	HL60	NA12878 (Coriell Institute)	NA12878 (Coriell Institute)	none	
Batch controls (RNA)	UHR (500 ng & matching sample amount)	AM7852 (Thermo Fisher)	We can add NA12878 RNA control if requested	UHR (matching sample amount, per run)	UHR (matching sample amount, per run)
Negative controls	DEPC water & PCR brew control	Water NTC	We don't add any usually	none	none
Average Size Distribution	225bp (FFPE) to 600 bp by Caliper/Agilent Bioanalyzer	300 bp (FFPE) to 700 bp by Fragment Analyzer or Tapestation	250 bp to 700 bp by Fragment Analyzer	350-500bp	200bp-600bp
Adapter contamination	<10% of area under the curve between lower and upper markers: adapter peak between 130-140bp	<10% of area under the curve between lower size marker and 170bp; adapter peak between 130-140bp	adapter peak between 100-140bp	<1%	<1%
RNA-seq final library yield (Qubit)	>3.0 nM in 10 uL (200-500 bp)	>4 nM (0.8 ng/uL)	>2 nM in 50 ul	>1nM	>1nM
PCR-free WGS (qPCR)	>2.5 nM in 15 uL		>2nM in 25 ul	>1nM	>1nM
Low-pass sequencing (QC)					
Library quantification	qPCR for TruSeq genome libraries	qPCR for TruSeq genome libraries		Not yet applicable, was planned only for FFPE but prostate (only FFPE cohort to date) has few aneuploidies from which derive estimated purity	
Instrument	MiSeq Nano QC before NovaSeq (RUO)	MiSeq QC before NovaSeq (RUO). NextSeq 550			
% Bases Over Q30	>75 at 2x150	>80 at 2x150 bp			
Min Clusters (PF)	>50% of target	>500K/lane			
Spike-in controls	pCR-TOPO4 tracking	PhiX (~0.1-2%)			

QC GATES	MOHCCN CONSORTIA				
	BC2C	**PM2C**	**MOH-Q**	**PR2C**	**ACC**
	plasmid (~0.1%)				
Shallow whole genome sequencing (QC) + ichorCNA					
Library quantitation		qPCR for TruSeq genome libraries			
Instrument		MiSeq QC before NovaSeq (RUO). NextSeq 550			
% Bases Over Q30		>80 at 2x150 bp			
Min Clusters (PF)		100,000 - 4,000,000			
Mean Coverage		0.1X - 0.4X			
ichorCNA % Tumour Purity		≥25%			
ichorCNA solution is accurate/logical		Yes (Trained Personnel)			
Full-depth sequencing					
Instrument	HiSeqX (Clinical) or NovaSeq 6000 (RUO)	NextSeq550 or NovaSeq	NovaSeq 6000	NovaSeq 6000	NovaSeq
% Bases Over Q30	>75 at 2x150 bp	>75 at 2x150 bp	>75 at 2x150 bp	>75 at 2x150 bp	>75 at 2x150 bp
Min Reads Delivered (PF)	18B reads per NovaSeq S4 flowcell	S4 Flow Cell = 1.6B/lane	2.4B per lane	10B reads per S4 flowcell passing filters (Illumina spec)	
Sequencing control	PhiX Control (1%)	PhiX Control (0.1%)	PhiX Control (1%)	PhiX Control (1%)	PhiX Control (1%)
WGS Minimum Coverage (Deduplicated)	80x T, 30x N We calculate the insert size from the sequenced data by getting the mean of the insert sizes of non-chasity-failed, primary-aligned, properly paired reads. We use some simple custom code for this.	80x T, 30x N Calculated by Picard CollectWgsMetrics.	80x T, 30x N Calculated by Picard CollectWgsMetrics.	80x T, 30x N Coverage would be the total length as reported by "samtools stat" with the remove-overlaps options enabled, divided by 2.9e9.	80x T, 30x N Calculated by Picard CollectWgsMetrics.

QC GATES	MOHCCN CONSORTIA				
	BC2C	**PM2C**	**MOH-Q**	**PR2C**	**ACC**
WGS Mean Insert Size	>150bp	>150bp	>150bp	>150bp	>150bp
WGS % Duplication Rate	≤50% (we don't fail on dups but using this metric makes sense)	≤50%	≤50% (Lucigen usually give 5-10% dup, but we don't fail on dups)	None (using PCR-free almost exclusively, so not really an issue)	
WTS Clusters per sample	>=80,000,000 (Warning only >95,000,000 but using 80M should be fine)	>80,000,000	>80,000,000 (try to reach 100M)	80M, but it's usually well over 100M	>80,000,000
WTS rRNA contamination	< 10 %	<35%	No set threshold (but <10% would be acceptable)	<20% historically	< 10 %
WTS %mapped to coding	> 25 % exonic	>5%	Not set threshold (but > 25-30% be acceptable)	Not set	>5%
WTS Mean Insert Size	No threshold set. Can use > 150bp	>150	No threshold set. Can use > 150bp	>150bp	>150bp
Informatics pipeline & variant interpretation					
Callability (exonic space)	Not yet applicable	≥75% of target bases above 30x T, 30x N	Not yet applicable	95% of "mappable genome" at 14x in N	
Inferred Tumour Purity	>35% to report genomic findings and full report. RNA expression outliers are always reported	≥30%	≥30%	>= 30%	>= 30%
Trimming	NA	Minimum base quality Q>20	Minimum base quality Q>25	Soft clipping only (Dragen feature), reads clipped to < 20nt are replaced with token 10Ns.	Minimum base quality Q>20
Final report generation					
QC Passed at all stages	Yes for clinical BAM/VCF (Clinical director	Yes (Geneticist)	Yes	Yes	Yes

QC GATES	MOHCCN CONSORTIA				
	BC2C	PM2C	MOH-Q	PR2C	ACC
	delegate)				
Calls are accurate/logical	Not yet applicable	Yes (Geneticist)	Not yet applicable	Not yet applicable	Not yet applicable

Color Legend: *different between all sites (**green**), MOH-Q specific difference (**blue**), PM2C specific difference (**yellow**), BC2C specific difference (**red**).*

3.3 Comparison of WGTS Data Processing Pipelines across MOHCCN Consortia (Table 2).

Experiment	WGTS analysis/pipeline step	MOHCCN Consortia				
		BC2C	PM2C	MOH-Q	PR2C	ACC
WGS Tumour ‖ Normal	**Primary Analysis**					
	Raw Data Files	BCL	BCL	BCL	BCL	BCL
		FASTQ	FASTQ	FASTQ	FASTQ	FASTQ
	QC Tools	QC0 Assessment	FastQC	FastQC	DRAGEN FastQC	FastQC
			BAMQC		BAMQC	
			Sample Fingerprinting		GenomeID match for tumor/normal (kind of hashcode based on highly polymorphic exonic SNPs)	
	Aligned Data Files	Raw BAM (mmp2)	Raw BAM (bwa-mem)	Raw BAM (bwa-mem)	Aligned BAM (dragmap), including merged runs	
			Aligned BAM (mapped reads only)			
	QC Tools		Picard		Dragen mapping metrics CSV	
	Processed Data Files	FASTQ (150 bp concat)	None		None	
		FASTQ (125 bp trimmed)				
WGS Tumour + Normal	**Secondary Analysis**					
	SNVs	VCF	VCF	VCF	VCF	VCF
			MAF	MAF		

Experiment	WGTS analysis/pipeline step	MOHCCN Consortia				
		BC2C	PM2C	MOH-Q	PR2C	ACC
	Tools	MuTect2	MuTect2	Mu Tect2	DRAGEN	
		Strelka2		Strelka2		
				Vardict		
				Varscan2		
	CNVs	SEG	SEG	VCF-CNV	VCF - CNV	VCF - CNV
		ploidy/purity estimates	ploidy/purity estimates		VCF - LOH	
		HRD				
		HOMD				
		DTA				
	Tools	cnaseq (requires tumour content)	Sequenza	Sequenza/PURPLE	DRAGEN	
		apollloh (requires tumour content)				
	Structural Variants	VCF	VCF	VCF	VCF	VCF
		Mavis.csv	Mavis.csv			
	Tools	DELLY	DELLY	GRIDSS	DRAGEN	
		MANTA				
		MAVIS	MAVIS			
	Tertiary Analysis					
	TMB	TMBur	None			

Experiment	WGTS analysis/pipeline step	MOHCCN Consortia				
		BC2C	PM2C	MOH-Q	PR2C	ACC
WGS Normal (Germline)	**Secondary Analysis**					
	SNPs	VCF	None	VCF-ensemble	VCF	VCF
	Tools	Haplotype	None	Strelka2, verdict, varscan2	DRAGEN	
	QC Tools		None			
	CNVs	control-freec	None	loh	VCF	
RNAseq	**Primary analysis**					
	Raw Data Files	FASTQ	FASTQ	FASTQ	FASTQ	FASTQ
	QC Tools		FASTQC	FASTQC		FASTQC
	Aligned Data Files	BAM	BAM	BAM	BAM	BAM
		.junction	.junction	.junction		
	QC Tools	guigolab bamstats, RSeQC (https://rseqc.sourceforge.net/)	RNASeqQC	RseQC, RNASeqQC		RNASeqQC
	Tools	STAR-RSEM	STAR	STAR	DRAGEN	STAR
	Secondary analysis					
	Gene expression Files	gene expression matrix (TPM, FPKM)	gene expression matrix (TPM, FPKM)	Gene expression matrix (TPM, FPKM)	gene expression matrix (TSV)	gene expression matrix (TPM, FPKM)
	Tools	RSEM	RSEM		DRAGEN	RSEM
	Fusion genes	Fusions.txt	Fusion.txt	Fusion.txt	Fusion report (CSV)	Fusion.txt

Experiment	WGTS analysis/pipeline step	MOHCCN Consortia				
		BC2C	PM2C	MOH-Q	PR2C	ACC
	Tools	STAR fusion	STAR fusion	STAR fusion	DRAGEN Gene fusion detection	STAR fusion
		Arriba	Arriba	Arriba		
		chimerscan		Annofuse		
		deFUSE				
		MAVIS	MAVIS			
	Variant calling			VCF	VCF	VCF
	Tools				DRAGEN expressed SNV calling	
	Tertiary Analysis					
	Immune Inference	CIBERSORT	None		None	
	TCR repetoire/diversity	MIXCR	None		None	
	Moffitt score	Note: PAAD-only	None		None	
	WPS	custom	None		None	
	Spearman correlation comparator check	custom	None		None	
	Drug target (DTA)	custom	None		None	
	HRD	custom	None		None	
	Mutation Signature	custom	None		None	
RNAseq + WGS Tumour	**Secondary Analysis**					

Experiment	WGTS analysis/pipeline step	MOHCCN Consortia				
		BC2C	PM2C	MOH-Q	PR2C	ACC
	Microbial detection	Custom analysis	None		None	
WGTS T + N + RNAseq	**Secondary Analysis**					
	MHC prediction		None		None	
	Tools	Optitype				
		Targeted Alignment				
	Neoantigen prediction		None		None	
	Tools	NetMHCPan				

Appendix 1

Protocol Recommendations for ChIP-seq (IHEC data standards) (Table 3).

The Technology working group recommends and invites to consider to IHEC protocols.

Stage	Metric (Threshold)	Metric Value
ChIP		
1. Chromatin preparation	Sheared chromatin fragment length	~150 – 500 bp
2. Antibody	1. Western-blot verification for histone recognition specificity	Verified
	2. ELISA or dot blots with modified histone tail peptides to verify specificity for the target modification	Verified
	3. Control Antibody to validate ChIP efficiency	Verified
3. ChIPed material	Amount and size of recovered DNA	Verified by Agilent Bioanalyzer or agarose gel electrophoresis
Library construction	PCR amplification cycles	~10 cycles with Illumina procedures
	Minimum positions unique of multiplexed adapters	At least 2
ChIP-seq		
Sequencing depth	Number of aligned reads	30-50 million
	- For transcriptional activation	~30 million may be adequate
Read length	Read length	36 bases
Replication	Number of biological replicates	At least 2

Table 3: Protocol recommendations for ChIP-seq adapted from International Human Epigenome Consortium (IHEC) data standards [https://ihec-epigenomes.org/research/reference-epigenome-standards/

Document revision history

Developed by	Reviewed by	Endorsed by	Effective Date	Policy Version	Summary of revisions
TWG	Steering Committee	Network Council	February 2,2023	V1	n/a