RNA-Seq Metadata Schema

Version 1.0
July 2017

Patient	Type	isRequired
patientID	integer	yes
surgeon	text	yes
sex	text	yes
race	text	yes
dateOfSurgery	date	yes
age	text	yes
historicalDiagnosis	text	yes
workingDiagnosis	text	yes
pathologicalDiagnosis	text	no
medicationID	integer	yes
FMAID	integer	yes
anatomicalLocation	text	yes
secondaryAnatomicalLocation	text	no
pathology	text	yes

Medications	Type	isRequired
medicationID	integer	yes
medication	text	yes
dose	text	yes
route	text	yes
frequency	text	yes

In addition to the pre-sequencing metadata items listed here, we also capture the wet-lab protocol used for sample preparation as either a PDF or Word document.

SequencingRun	Type	isRequired
runID	integer	yes
description	text	yes
investigators	text	yes
externalLab	text	no
submitDate	date	yes
manufacturer	text	yes
machineModel	text	yes
sequencingKit	text	yes
readLength	integer	yes
paired	text	yes
facility	text	yes
facilityID	text	no
demultiplexer	text	yes
notes	text	no

Sample (pre-sequencing)	Type	isRequired
sampleID	text	yes
harvestBy	text	yes
harvestDate	date	yes
harvestTime	text	yes
firstStrandAmplifiedBy	text	no
firstStrandAmplifiedDate	date	no
amplifiedBy	text	yes
amplifiedDate	date	yes
libraryBy	text	yes
libraryDate	date	yes
species	text	yes
strain	text	yes
strainOther	text	no
ageHarvestedUnit	text	yes
ageHarvestedUnitOther	text	no
ageHarvestedValue	float	yes
ageCultured	integer	no
harvestMethod	text	yes
harvestQuant	text	no
harvestCompoundType	text	yes
isHarvestCompoundUncaged	text	no
region	text	yes
subregion	text	no
subregionOther	text	no
cellClass	text	no
cellClassOther	text	no
cellType	text	no
cellTypeOther	text	no
supplier	text	no
sourceType	text	yes
sourceUnit	text	yes
sourceUnitOther	text	no
sourceRegion	text	no
sourceRegionOther	text	no
patientID	integer	no
cellID	text	no
dendID	text	no
sourceWellID	text	no
treatment	text	yes
replicate	text	no
sourceAmplificationMethod	text	yes
sourceAmplificationRounds	integer	yes
barcode	text	yes
barcodeSeq	auto-complete	yes
flowcellLane	text	no
spikeInType	text	yes
spikeInDilution	integer	no
libraryInputAmount	text	yes
libraryKit	text	yes
strandSpecific	text	yes
libraryConstPCRCycles	text	yes
amplificationProtocol	text	yes
libraryProtocol	text	yes
contaminantsFile	text	yes
dbGaP Release	text	no
notes	text	no

For post-sequencing, we capture provenance about the primary analysis pipeline. The fields below illustrate the metadata we track for various stages of our pipeline (i.e. the PennSCAP-T pipeline). This post-sequencing metadata is largely the version numbers and program parameters.

- BLAST – for QC
- TRIM – trimming prior to aligning
- STAR – aligning and includes running samtools for some post-aligning processing of the BAM output
- VERSE – gene counting after aligning

Sample (post-sequencing)	Type
sampleID	text
pipeline	text
STAR.star.version	text
STAR.samtools.version	text
STAR.species (genome)	text
STAR.readLength	integer
STAR.SE	boolean
VERSE.verse.version	text
VERSE.transcriptome	text
VERSE.stranded	boolean
VERSE.ID_attribute	text
VERSE.introns	boolean
VERSE.intergenic	boolean
VERSE.lines-sines	boolean
BLAST.blastn.version	text
BLAST.parseBlast.py.version	text
BLAST.species	text
BLAST.readLength	integer
BLAST.numReads	integer
BLAST.kmer	text
TRIM.trimReads.py.version	text
TRIM.minLen	integer
TRIM.phredThresh	integer
TRIM.removeN	boolean
TRIM.numAT	integer
TRIM.SE	boolean
TRIM.ContamFile	text