RNA-Seq Metadata Schema Documentation

Version 1.0 (draft)
June 2017

SequencingRun

NOTES:

A sequencingRun is defined as a single NGS run (e.g. all lanes of in an Illumina flow cell). This metadata should apply to all samples in a single sequencing run. Samples of different types may be multiplexed and included in the same lane. The information collected here should be true for all samples in the sequencing run. Sample-specific information is collected in the fields specific to each sample type (i.e. RNAseqSamples).

FIELDS:

- **runID** (unique integer, required)
    - should be filled automatically from next available value
- **description:** (text, required)
    - brief description of the project(s).
- **investigators:** (multi-select, required)
    - PI of lab group
    - choose from
        - J. Kim
        - J. Eberwine
        - other - **you must list PI's in externalLab**
- **externalLab:** (text)
    - comma separated list of PI's not included in the "PI" field.
    - this field should be empty if "other" is not selected in "PI".
    - **[VALIDATE: (externalLab == null) if (PI not contain "other")]**
    - **[VALIDATE: (externalLab != null) if (PI contains "other")]**
- **submitDate** (date, required)
    - Date submitted to sequencing facility or date put on sequencer.
- **manufacturer** (drop-down, required)
    - Manufacturer of sequencing machine/technology
    - choose from
        - Illumina
        - Pacific Biosciences

- **machineModel** (drop-down, required)
  - Specific model number/name of sequencing machine
  - choose from
    - NextSeq 500
    - HiSeq 2500
    - PacBio RS II
    - HiSeq 1000
    - HiSeq 2000
    - MiSeq
    - MiSeq v2
    - GA II
    - GA IIx
- **sequencingKit** (drop-down, required)
  - comma separated list of sequencing kits used. Please include the catalog number so as to be explicit about the kits used.
    - Illumina CT-402-4001, PE-402-4002, FC-402-4021 - TruSeq Rapid SBS Kit v2
    - Illumina CT-402-4001, PE-402-4001, FC-402-4001 - TruSeq Rapid SBS Kit v1
    - Illumina FC-404-2002 - NextSeq
    - Illumina PE-401-4001, FC-401-4003 - TruSeq Kit v4
    - unknown
- **readLength** (integer, required)
  - length of reads
  - for paired end use length of a single read, i.e. 100x100 paired end would be "100".
- **paired** (drop-down, required)
  - Either Paired End or Single End.
  - choose from
    - Paired-end
    - Single-end
- **facility** (text, required)
  - the sequencing facility
  - choose from
    - inHouse
    - CSH
    - USC
    - UCSD
    - PGFI
    - other - **you must document the sequencing center in the notes field.**
- **facilityID** (text)
  - ID of the sequencing run used by the sequencing facility.
  - i.e. FGC1028 or PathBio 90513
  - Leave blank if unknown.

- **demultiplexer** (text, required)
    - the program used to convert the images to fastq files.
    - Be sure to include the program version number (i.e. Casava v1.8.2)
    - choose from
        - bcl2fastq2 v2.17
        - bcl2fastq v1.8.4
        - CASAVA v1.8.2
        - unknown
- **notes** (text)
    - Other information or comments about the full sequencing run not captured in the fields above.
    - For published data sets include a link to where you downloaded the data.

RNAseqSamples

NOTES:

1. Use "**n/a**" when values are "**Not Applicable**" (i.e. the field is not relevant to the specific sample) and use "**unknown**" when values are applicable but not known (i.e. there is a value but you either do not know the value or are not able to find the value through reasonable efforts). **Do not use "n/a" to mean "not available".**
2. For multi-day procedures (e.g. amplification), the date represents the day the procedure started.
3. The date **"Jan 1, 1970"** (zero in Unix time) is used to denote an unknown date value.
4. For attaching files to samples, see FileAttachments

FIELDS:

- **sampleID** (text, required)
  - for demultiplexing with bcl2fastq v2.17, sample names must be a valid linux filename containing only a-z, A-Z, 0-9, underscore (_) and dash (-), with no other special characters.
  - this should be string that will uniquely identify the sample within the project. If this sample is a technical replicate, then use the same sample ID as the original replicate and append a "b, c, d" to the sampleID so that it's still unique.
  - limit to as few characters as possible while still unique and descriptive.
  - i.e. CM123b, UHRR100ng, Ms_Dend1, etc.
  - **[VALIDATE: sampleID is legal Linux filename]**
- **harvestBy** (drop-down, required)
  - first initial and complete last name.
  - researcher who collected the sample from the source material be it culture dish, tissue, etc.
  - in the case of controls such as a water control, the researcher who initiated the sample.
  - choose from
    - J. Singh
    - S.A. Fisher
    - Y.H. Kim
    - Agilent 740000 - (UHRR) purchased from Agilent, catalog number 740000
    - Life Tech AM6050 - (HBR) purchased from Life Technologies, catalog number AM6050
    - Ambion AM6051
    - CSH - Cold-Spring Harbor
    - National workshop
    - GEO - put link to data download in Notes field
- **harvestDate** (date, required)
  - date sample was collected from the source material.
  - date sample was purchased, if relevant

- **harvestTime** (drop-down)
    - the time of day when the sample was harvested from the organism or culture
    - choose from
        - morning – between 7am and 12pm
        - afternoon – between 12pm and 5pm
        - night – between 5pm and 7am
        - unknown
- **firstStrandAmplifiedBy** (drop-down)
    - first initial and complete last name.
    - only use this field when first strand amplification is done at a separate time than the rest of the amplification, even if by the same person.
    - choose from
        - J. Singh
        - T. Bell
        - S. Middleton
        - n/a
- **firstStrandAmplifiedDate** (date)
    - date the first strand amplification was performed.
    - leave empty if firstStrandAmplifiedBy is set to n/a.
    - **[VALIDATE: (firstStrandAmplifiedDate == null) if (firstStrandAmplifiedBy = "n/a")]**
    - **[VALIDATE: (firstStrandAmplifiedDate >= harvestDate)]**
- **amplifiedBy** (drop-down, required)
    - this is amplification prior to library construction.
    - first initial and complete last name.
    - choose from
        - J. Wang
        - J. Singh
        - T. Bell
        - S. Kim
        - Y.H. Kim
        - CSH - Cold-Spring Harbor
        - National workshop
        - Unknown
- **amplifiedDate** (date, required)
    - date the sample was amplified.
    - **[VALIDATE: (amplifiedDate > firstStrandAmplifiedDate) if (firstStrandAmplifiedDate != null)]**
    - **[VALIDATE: (amplifiedDate >= harvestDate)]**
- **libraryBy** (text, required)
    - first initial and complete last name.
    - choose from
        - [same list as "amplifiedBy"]
- **libraryDate** (date, required)
    - date of library creation
    - **[VALIDATE: (libraryDate >= amplifiedDate)]**

- **species** (drop-down, required)
    - The species from which the source material is derived.
    - choose from
        - human - select "n/a" for strain
        - mouse
        - rat
        - danio rerio
        - drosophila melanogaster
        - C. elegans
        - water
- **strain** (drop-down, required)
    - strain of source animal
    - if this is a transgenic strain then include the reference in notes
    - choose from
        - C57/BL6
        - C57/BL6/N
        - Sprague Dawley
        - CD-1
        - wild type
        - Tu/TLF
        - n/a - select this for human species
        - unknown
        - other - **you must fill out strainOther**
    - **[VALIDATE: (strain == "n/a") if (species == "human")]**
- **strainOther** (text)
    - this should only be filled out if strain is set to "other"
    - **[VALIDATE: (strainOther == null) if (strain != "other")]**
    - **[VALIDATE: (strainOther != null) if (strain == "other")]**
- **ageHarvestedUnit** (drop-down, required)
    - Age or developmental state. Please be as specific as possible.
    - choose from
        - days old (P)
        - weeks old
        - months old
        - years old
        - embryonic (E)
        - hours post fertilization
        - days post fertilization
        - weeks post fertilization
        - months post fertilization
        - Hamburger Hamilton Stage
        - adult (use for HBR and UHRR samples)
        - unknown
        - other

- **ageHarvestedUnitOther** (text)
  - this should only be filled out if ageHarvestedUnit is set to "other"
  - **[VALIDATE: (subregionOther == null) if (subregion != "other")]**
  - **[VALIDATE: (subregionOther != null) if (subregion == "other")]**
- **ageHarvestedValue** (float, required)
  - Age or developmental state.
  - This pertains to the unit selected in ageHarvestedUnit
  - Can be an integer or a decimal, such as "47" (DO) or (P)"18.5"
  - Use "-1" if ageHarvestedUnit does not have an accompanying number (embryo, adult, unknown, etc.)
  - **[VALIDATE: (ageHarvestedValue == null) if (ageHarvestedUnit == "unknown" or "immortal" or "adult" or "neonatal" or "embryo")]**
  - **[VALIDATE: (ageHarvestedValue != null) if (ageHarvestedUnit != "unknown" or "immortal" or "adult" or "neonatal" or "embryo")]**
- **ageCultured** (integer)
  - length of cell culture IN DAYS
  - this field is required when sourceType is "culture"
  - enter -1 if sourceType is culture and the ageCultured is unknown.
  - enter 0 if sourceType is not "culture"
  - **[NOTES: we really want this to be a text field allowing for "unknown", "n/a" or an integer > 0]**
  - **[VALIDATE: (ageCultured == 0) if (sourceType != "culture")]**
  - **[VALIDATE: ((ageCultured == -1) or (ageCultured > 0)) if (sourceType == "culture")]**
- **harvestMethod** (drop-down, required)
  - choose from
    - FACS
    - pipette - this is a standard pulled-pipette used to collect cell bodies
    - bulk pipette - this is a larger pipette used to collect bulk material
    - micro pipette - this is a special smaller than usual pipette for intracellular use
    - Trizol
    - Fluidigm C1
    - in situ transcription
    - homogenization
    - isoraft
    - HBR: modified Ambion ToTALLY RNA Kit (Cat #1910) - This is the harvest protocol used by Life Tech to extract HBR (see https://tools.thermofisher.com/content/sfs/manuals/sp_6050.pdf)
    - unknown
    - N/A - use this only when no nucleic material at all was harvested (i.e. water control)
- **harvestQuant** (text)
  - quantity of (unamplified) RNA harvested, in the form 10pg, 1ng, etc.
  - for approximate values prefix with ~ (i.e. ~100ng)

- **harvestCompoundType** (drop-down, required)
  - choose from
    - TIVA
    - TISA v1
    - TISA v3
    - TISA v4a
    - APRA
    - n/a
- **isHarvestCompoundUncaged** (drop-down)
  - Required if harvestCompoundType is not n/a.
  - This should be blank if harvestCompoundType is n/a.
  - choose from
    - Yes
    - No
    - Unknown
    - leave blank
  - **[VALIDATE: (harvestCompoundUncaged == null) if (harvestCompoundType == "n/a")]**
  - **[VALIDATE: (harvestCompoundUncaged != null) if (harvestCompoundType != "n/a")]**
- **region** (drop-down, required)
  - Either an organ or a type of cell line
  - choose from
    - brain
    - heart
    - oviduct
    - pancreas
    - bone marrow
    - skin
    - embryonic stem cell line
    - melanoma cancer cell line
    - melanocyte
    - prostate cancer cell line
    - mixed
    - HeLa - this is the cellClass and region
    - n/a
    - unknown

- **subregion** (drop-down)
    - choose from
        - cortex
        - hippocampus
        - hypothalamus
        - parietal cortex
        - frontal cortex
        - temporal cortex
        - visual cortex
        - hippocampus CA1
        - hippocampus CA2
        - hippocampus CA3
        - hippocampus stratum oriens
        - hippocampus
        - olfactory epithelium
        - PI
        - dorsal raphe LW
        - dorsal raphe VM
        - amygdala
        - cerebelum
        - blastomere
        - dermal foreskin
        - n/a
        - unknown
        - other - **you must fill out subregionOther**
- **subregionOther** (text)
    - this should only be filled out if subregion is set to "other"
    - **[VALIDATE: (subregionOther == null) if (subregion != "other")]**
    - **[VALIDATE: (subregionOther != null) if (subregion == "other")]**

- **cellClass** (drop-down)
    - choose from
        - neuron
        - astrocyte
        - cardiomyocyte
        - brown adipose cell
        - embryonic stem cell
        - fibroblast
        - Hela
        - hepatocyte
        - macrophage
        - microglia
        - neoblast
        - primitive endoderm cell
        - stem cell
        - zygote
        - n/a - **use this if sourceUnit is bulk or tissue**
        - unknown
        - other - **you must fill out cellClassOther**
    - **[VALIDATE: (cellClass == "n/a") if (sourceUnit == "bulk") or (sourceUnit == "tissue"))]**
- **cellClassOther** (text)
    - this should only be filled out if cellClass is set to "other"
    - **[VALIDATE: (cellClassOther == null) if (cellClass != "other")]**
    - **[VALIDATE: (cellClassOther != null) if (cellClass == "other")]**
- **cellType** (drop-down)
    - choose from
        - pyramidal
        - olfactory sensory neuron
        - fast spiking
        - 5-HT
        - interneuron
        - dilp2+
        - Marker identified
            - for cells that have been selected for a specific protein/other molecular marker.
            - fill out "cellTypeOther" with the specific marker(s)
        - n/a
        - unknown
        - other - **you must fill out cellTypeOther**
- **cellTypeOther** (text)
    - this should only be filled out if cellType is set to "other" or "Marker identified"
    - **[VALIDATE: (cellTypeOther == null) if (cellType != "other" || cellType != "marker identified")]**
    - **[VALIDATE: (cellTypeOther != null) if (cellType == "other" || cellType == "marker identified")]**

- **supplier** (drop-down)
  - this is only relevant to the subset of samples purchased from an outside supplier, otherwise leave blank
  - choose from
    - GTEx
    - Agilent
    - Life Tech
    - Ambion
    - Zyagen
    - Unknown Supplier - for samples known to come from a commercial source, but that source is unknown.
- **sourceType** (drop-down, required)
  - choose from
    - culture - **you must fill out ageCultured**
    - primary culture
    - in vivo
    - biopsy
    - acute slice
    - fixed section
    - tissue
    - dissociated tissue
    - plasmids
    - Mixed cell lines (This includes UHRR)
    - purchased from Zyagen
    - purchased from Ambion AM6051
    - purchased from UCSD
    - n/a - if you are unsure about the sourceType then select "unknown"
    - unknown
- **sourceUnit** (drop-down, required)
  - this is the unit that is sequenced.
  - choose from
    - single cell
    - pooled single cells - put the number of cells pooled in sourceUnitOther
    - subcellular
    - tissue
    - coverslip
    - culture dish
    - bulk purified total RNA
    - n/a - if you are unsure about the sourceUnit then select "unknown"
    - unknown
    - other - **you must fill out sourceUnitOther**

- **sourceUnitOther** (text)
    - this should only be filled out if sourceUnit is set to "other" or "pooled single cells"
    - for "pooled single cells", fill with an integer value for the number of cells pooled.
    - **[VALIDATE: (sourceUnitOther == null) if ((sourceUnit != "other") and (sourceUnit != "pooled single cells"))]**
    - **[VALIDATE: (sourceUnitOther != null) if ((sourceUnit == "other") or (sourceUnit == "pooled single cells"))]**
- **sourceRegion** (drop-down)
    - Region of the cell.
    - Each branch coming directly off the soma is a unique dendritic branch. All secondary and tertiary branches are considered the same dendrite.
    - choose from
        - full cell - FACS and similar techniques used to capture an entire cell.
        - whole cell - for neurons, this is soma and ONLY proximal dendrites (i.e. distal dendrites were likely missed during collection).
        - soma - this is just the soma without any proximal dendrites.
        - dendrite - only one dendrite in the sample.
        - multiple dendrites - this means more than one dendrite in the sample. Put the number of dendrites in sourceRegionOther.
        - whole cell plus - **ONLY USE THIS OPTION FOR NEURONS.** This is soma and ALL dendrites (i.e. care was taken to make sure all dendrites were collected).
        - no cell - no actual cell was collected, such as from released RNA, culture media, or water controls.
        - n/a - **use this if sourceUnit is bulk or tissue**
        - unknown
        - other - **you must fill out sourceRegionOther**
    - **[VALIDATE: (sourceRegion == "n/a") if (sourceUnit == "bulk") or (sourceUnit == "tissue"))]**
- **sourceRegionOther** (text)
    - this should be filled out if sourceRegion is set to "other"
    - put the number of dendrites here if "multiple dendrites" is selected in sourceRegion.
    - **[VALIDATE: (sourceRegionOther == null) if (sourceRegion != "other")]**
    - **[VALIDATE: (sourceRegionOther != null) if (sourceRegion == "other")]**
- **patientID** (text)
    - Patient specific identifier
    - i.e. SCE-mRNA-HD-xxx
    - Project specific, so leave empty if not relevant.
- **cellID** (text)
    - Unique ID for a specific cell.
    - The value should be: HarvesterInitials-Integer (e.g. SAF-1)
    - MUST BE UNIQUE FOR ALL SAMPLES COLLECTED BY ONE HARVESTER, IRRESPECTIVE OF THE PROJECT. For example, all cells harvested by Stephen would be uniquely identified as SAF-1, SAF-2, SAF-3, etc. Future projects would continue in the same sequence (e.g. SAF-4, SAF-5, etc).
    - Comma separated if multiple cellID's for one sample.

- **dendID** (text)
  - Unique ID for each dendrite when multiple dendrites from the same cell are collected separately (i.e. unique per cell).
  - Semi-colon separated if multiple dendID's for one sample.
  - IF THE SAMPLE IS NOT A DENDRITE THEN LEAVE BLANK (i.e. "soma" and "whole cell plus" should have no value here).
  - **[VALIDATE: (dendID == null) if not ((sourceRegion == "dendrite") or (sourceRegion != "multiple dendrites"))]**
- **sourceWellID** (text)
  - Well in source plate, i.e. for FACS samples.
  - The value should be: *plateNumber.columnNumber.rowCharacter* (e.g. 1.5.A, 2.12.C)
  - The plateNumber is specific to a single harvest event. For example, a patient might have multiple plates.
- **treatment** (text, required)
  - any alterations of the cells prior to sampling
  - any one-off alterations to a protocol
  - if no treatment then enter "n/a"
- **replicate** (text)
  - REQUIRED if this sample is a TechnicalReplicates of any previous sample. Leave blank if this sample is not a replicate of a previous sample.
  - It is important that this description explain at what point this sample deviated from the other replicates (ex after amplification, library construction, etc). The sample should have the same name as replicate, with a "b, c, d, etc" appended to sample name.
- **sourceAmplificationMethod** (drop-down, required)
  - how was the source material amplified prior to library construction
  - choose from
    - aRNA
    - PCR
    - SMARTer Ultra Low RNA Kit
    - SMARTer Ultra Low RNA Kit, Advantage 2
    - none - if you are unsure about the amplificationMethod then select "unknown"
    - unknown
- **sourceAmplificationRounds** (integer, required)
  - rounds of the above amplificationMethod
  - i.e. for 3x aRNA amplificationRounds = 3
  - **[VALIDATE: (sourceAmplificationRounds > 0) if ((sourceAmplificationMethod != "none") and (sourceAmplificationMethod != "unknown"))]**
- **barcode** (drop-down, required)
  - Illumina index for barcode. The value in this field dictates the value of BarcodeSeq.
  - **[POPULATE: from barcodeValues table keys]**
- **barcodeSeq** (auto-complete, required)
  - This field is automatically populated by BarcodeIdx. This is not an editable field.
  - Select "From External Lab" if the barcode has not been provided to us.
  - **[POPULATE: from barcodeValues using value of barcodeIdx]**

- **lane** (text)
  - Enter the lane number, if known, else leave blank.
- **spikeInType** (drop-down, required)
  - External control added.
  - choose from
    - ERCC (LifeTech 4456740, Mix 1)
    - SIRV
    - Ambion C1, C4, C7
    - none
- **spikeInDilution** (integer)
  - concentration of spike-in added to sample. Using "1 to X" this value is the X.
  - leave blank if no spike-in.
  - **[VALIDATE: (spikeInDilution == null) if (spikeInType == "none")]**
  - **[VALIDATE: (spikeInDilution > 0) if (sourceRegion != "none")]**
- **libraryInputAmount** (text, required)
  - the input in ng or ug of DNA or RNA into library prep (i.e. the raw product from extraction, no adapters)
  - Should be either formatted like 200ng or 1.5ug or "unknown"
  - If you are unsure then enter "unknown"
- **libraryKit** (drop-down, required)
  - specific kit used to generate library. list the catalog number to be as explicit as possible. We need to be able to determine if the kit changes for future samples.
  - choose from
    - Illumina TruSeq Stranded (RS-122-2101/2)
    - Illumina Nextera
    - Illumina TruSeq DNA nano
    - Illumina TruSeq RNA Sample Prep Kit v2 (RS-122-2001/2)
    - Illumina TruSeq RNA Sample Prep Kit v1 (RS-122-1001/2)
    - Illumina mRNA-seq (Illumina # 1004814)
    - PacBio SMRTbell barcoded prep kit (100-465-800, 100-466-000, 100-465-900)
    - Unknown Small RNA
    - Unknown Illumina DNA-Seq
    - unknown
    - none
- **strandSpecific** (drop-down, required)
  - was this library prepared with a strand specific library kit
  - choose from
    - TRUE
    - FALSE

- **libraryConstPCRCycles** (drop-down, required)
    - the number of PCR cycles during library construction amplification
    - choose from
        - 1
        - 2
        - ...
        - 20
        - unknown
- **amplificationProtocol** (drop-down, required)
    - this is amplification prior to library construction.
    - The drop-down list is the list of all files of type "amplificationProtocol"
    - **[POPULATE: from files with type "amplifiedProtocol"]**
- **libraryProtocol** (drop-down, required)
    - The drop-down list is the list of all files of type "libraryProtocol"
    - **[POPULATE: from files with type "libraryProtocol"]**
- **contaminantsFile** (drop-down, required)
    - this is the contaminants file we use for trimming. All adapter and priming sequences are included in this file.
    - choose from
        - Default - Illumina TruSeq index adaptors and aRNA primers
        - TIVA
        - Nextera
        - Smartseq Illumina
        - Smartseq Nextera
        - Smartseq Nextera_v2
- **dbGaP Release** (drop-down)
    - this includes the dbGaP accession number for the project
        - version information is stripped from the dbGaP accession number (e.g. "phs000833.v3.p1" becomes " phs000833")
        - in the case of projects with multiple dbGaP upload and release sets, a release number is appended to the end of the dbGaP accession number (e.g. " phs000833 r1" and " phs000833 r2" for release one and release two, respective).
    - choose from
        - phs000833 r1
        - phs000833 r2
        - phs000833 r3
- **notes** (text)
    - Sample-specific information not captured in other fields.

Patient

FIELDS:

- **patientID** (integer, required)
- **surgeon** (text, required)
- **sex** (text, required)
- **race** (text, required)
- **dateOfSurgery** (date, required)
- **age** (text, required)
- **historicalDiagnosis** (text, required)
- **workingDiagnosis** (text, required)
- **pathologicalDiagnosis** (text)
- **medicationID** (integer, required)
- **FMAID** (integer, required)
- **anatomicalLocation** (text, required)
- **secondaryAnatomicalLocation** (text)
- **pathology** (text, required)

Medications

FIELDS:

- **medicationID** (integer, required)
- **medication** (text, required)
- **dose** (text, required)
- **route** (text, required)
- **frequency** (date, required)

Analysis

FIELDS:

- **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)