Fastq filename format <-- please make sure that the sequence facility names our files in this (and only this) format:

LabOwner\_ arbitraryLibraryNumber\_GTAC\_index#\_index1\_index2\_<sequencer stuff>

BRENT\_1\_GTAC\_1\_AAGATTA\_GTAACCA\_S1\_R1\_001.fastq.gz | from 2/28/2020

There should be no white space or any other information in the fastq filename other than the information listed above. If this becomes too restrictive, or if it is not unique enough to easily identify a given sample in a run, then please **do not** make up your own format for a single library. Rather, add a line below the above filename specification and include the date from which **all** filenames will follow the new format.

The following conventions should be followed in every individual sheet as well as in filenames:

1. All dates should be in format: mm.dd.yy, all numeric. E.g 07.29.19
2. All names should be FIRSTINITIAL.LASTNAME, all caps. E.g. “M.BRENT”
3. Each table should be in a separate .csv or .xlsx file, named according to the following scheme:

* bioSample\_M.BRENT\_07.29.19
* rnaSample\_M.BRENT\_07.30.19
* s1cDNASample\_M.BRENT\_07.31.19
* s1s2Pooling\_M.BRENT\_07.31.19
* s2cDNASample\_M.BRENT\_08.01.19
* library\_M.BRENT\_08.02.19
* fastqFiles\_08.04.19 (may contain files for libraries prepared by multiple people)

NOTE: The above name/date format implies that there should be no more than a single day’s worth of work in the unique key (see below) of a given sheet. For example, the harvestDates in a given bioSamplesheet should all be the same.

The column headers for each file and what should be entered into them are listed below.

See <https://github.com/BrentLab/database_files.git> for instructions on pulling the database-files repository. All sheets should be stored in the correct respective subdirectory (i.e. bioSample or rnaSample). See the link above for instructions on cloning, commiting and pushing changes to the remote repo. All changes should be commited and pushed by the end of any work day in which you add or modify a sheet.

**Biosample Table**

If multiple aliquots of cells are taken at the same time from the same culture (e.g. tube or colony on an agarose plate) then they are aliquots of a single biosample, so only one entry is needed in the biosample table. The key for this table is harvestDate, harvester, bioSampleNumber and therefore these columns taken togehter must be unique both in the individual sheet, and when concatenated with the rest of the bioSample tables in database-files.

An example biosample table file can be found in:

database-files/bioSample/bioSample\_J.PLAGGENBERG\_06.05.19.csv

**harvestDate**

The date on which the cells were taken out of culture and growth was stopped.

**harvester**

Name of the person (e.g. J.PLAGGENBERG) who was in charge of the experiment. If more than one person was involved, just pick one.

**bioSampleNumber**

Sequential positive integers uniquely identifying the samples taken by a given harvester on a given date. Numbers can be reused as long as the harvester and date are different.

**experimentDesign**

This should be a short string with no spaces that refers to a design which is actually explained in some other document. E.g. ZEVNeg1\_10\_15\_20\_90\_SCGal\_plates. Do not use something that could be mistaken for a number, such as “20.20”. This doesn’t have to be so descriptive -- you could even use ZEV1, since it will be explained elsewhere. The other document should have the same name as the table entry. It should describe the experiment and list the relevant fields needed to keep all the treatments and samples separate. An example can be found [here](https://docs.google.com/document/d/15tPOry1pybXC8HmQlRcVrRV1j-pQ75Xj3Dhg9JdMz5U/edit).

**experimentObservations**

A reference to a document that describes any observations made while running the procedure. If everything was normal and no specific observations, it can just say “As expected”. Still, there should be a different file for each experiment.

**Other fields**

Other fields describing the sample within the experiment design should be included, such as:

* strain -- the actual collection of cells used to inoculate. Usually corresponds to a tube or well in the freezer. There can be multiple strains of a given genotype -- typically when multiple independent mutants are made. But also each time a new batch of the wild type laboratory strain is grown or obtained from a source it should be given a new strain number. Format should conform to literature standard and be consistent (e.g. ZEV is always ZEV and not sometimes ZeV).
* genotype -- the genotypes used should be defined in the experimental design doc. This must conform to a standard and be consistent.
* Media – string, no whitespace and consistent if the media type exists already in the database. If you need to check this, use queryDB.
* Treatment – A string, no whitespace, and consistent with other entries in this column. Use queryDB if you need to check to see if a treatment already exists in the database and use that format exactly. Could be any type of action that the timepoint is relative to, such as
  + estradiol
  + mockEstradiol -- ethanol vehicle only
  + conditionShift
  + glucoseTo2%
* timePoint – A decimal number in minutes (if whole minutes, there is no need to add the decimal point). The time between the treatment and the sample collection. Use -1 (neg1 as an integer) for samples collected immediately before the treatment is applied.
* inductionDelay -- for ZEV experiments, the time between resuspending the cells and adding the inducer or mock inducer. A decimal number in minutes (if whole minutes, there is no need to add the decimal point).
* replicate – A positive integer. if a ZEV experiment is replicated more than once each experiment will be given a replicate number.

All non-standard columns (those included in ‘other fields’ or simply columns added by the experimenter) should be described in the experimental design.

**RNA Sample Table**

If multiple aliquots of RNA are made from a single tube or well, they are the same RNA sample, so only one entry is needed. An example RNA sample table can be found in:

database-files/rnaSample/rnaSampLe\_J.PLAGGENBERG\_05.30.19.csv

**harvestDate,** **harvester**, **bioSampleNumber**

The entries in these three columns should uniquely identify the biosample from which the RNA was prepared. They should exactly match the corresponding entries for one and only one row of a biosample table.

**rnaDate**

The date (see date format at top of document) on which the total rna was prepared and ready for use.

**rnaPreparer**

Name of the person (see name format at top of document) who was in charge of the total RNA isolation. If more than one person was involved, just pick one.

**rnaSampleNumber**

Sequential positive integers uniquely identifying the total RNA samples prepared by a given preparer on a given date.

**rnaPrepMethod**

One of the following. If anything other than these is entered, it should be added to the list below with an explanation. For each entry, there should be a detailed protocol document with the same name. Any time the protocol is changed, a new document should be made up describing the differences relative to an older document. Older documents should never be deleted.

* DirectZol -- DirectZol kit from Zymo, per instructions
* RiboPure0.5X -- RiboPure Yeast kit from Ambion, using 0.5 times volume and reagents in instructions.
* RiboPure0.25X -- RiboPure Yeast kit from Ambion, using 0.25 times volume and reagents in instructions.
* RiboPure0.125X -- RiboPure Yeast kit from Ambion, using 0.125 times volume and reagents in instructions.
* ComboA -- uses a combination of RNA prep methods, homogenize RNA following the Directzol kit from Zymo, and purify RNA using the RiboPure Yeast kit from Ambion (using 0.25 times volume and reagents).
* ComboB -- uses a combination of RNA prep methods, homogenize RNA following the RiboPure Yeast kit from Ambion (using 0.25 times volume and reagents), and purify RNA using the Directzol kit from Zymo.
* TRIzol

**rnaPrepProtocol**

The name of a file in the protocols directory under the experimenter directory. Can be reused in multiple files -- no need to rename as long as the protocol is unchanged.

**roboticRNAPrep**

“TRUE” (with that capitalization) if the robot was used, “FALSE” if not.

**Other fields**

Option other fields that may describe the quality of the RNA.

* RIBOSOMAL\_BAND (TRUE or FALSE)
* RIBOSOMAL\_BAND\_SHAPE (straight, smile, or NA)
* SMALL\_RNA\_BANDS (TRUE or FALSE)
* RIN -- a value assigned by the Bioanalyzer between 0 and 10.

**First Strand cDNA sample Table**

If multiple aliquots of the first strand product made in a single pool or well are combined into different pools for second strand synthesis, they are the same first strand sample so only one entry is needed.

**rnaDate, rnaPreparer**, **rnaSampleNumber**

The entries in these three columns should uniquely identify the RNA sample from which the first strand cDNA was prepared. The formats of rnaDate and rnaPreparer should be the same as the date and name columns in all other date and name columns (see top of document). The rnaSampleNumber must be a sequential integer. They should exactly match the corresponding entries for one and only row of an RNA sample table.

**s1cDNADate**

The date on which the first strand cDNA was prepared and ready for use. See date format at the top of this document.

**s1cDNAPreparer**

Name of the person (see format at top of document) who was in charge of the first strand cDNA preparation (try not to split these, or we will have to add more fields!). If more than one person was involved, just pick one.

**s1cDNASampleNumber**

Sequential positive integers uniquely identifying the single-stranded cDNA or cDNA-RNA hybrid samples prepared by a given preparer on a given date.

**polyAIsolationProtocol**

Entries, to be defined, indicating the kit or protocol used for isolating polyA+ rna from total RNA or for depleting unwanted RNAs such as ribosomal RNAs. Should be a single string of letters with no whitespace. Add any entries used below, with an explanation.

* NEBNextPoly(A)E7490 -- NEBNext Poly(A) mRNA Magnetic Isolation Module, thermo catalog number E7490
* None -- if no polyA isolation was performed, as in BRBSeq

**s1Protocol**

Entries, to be defined, indicating the kit or protocol used for generating the first strand. Should be a single string of letters with no whitespace. Add any entries used below, with an explanation.

* NEBUltraIIE7771L -- The NEBNext Ultra II RNA First Strand Synthesis Module, Thermo catalog number E7771L
* …..

**roboticS1Prep**

“TRUE” (with this capitalization) if the robot was used; “FALSE” otherwise.

**s1PrimerSeq**

The sequence (all caps) of the primer if an indexed primer is being used, or “random” if random primers are used.

**Second Strand cDNA Table**

There should be one entry per tube full of double-stranded DNA. An example file can be found at database-files/s2cDNASample/s2CDNASample\_J.PLAGGENBERG\_05.30.19.csv.

**s2cDNADate**

The date (see format at top of page) on which the second strand cDNA was prepared and ready for use.

**s2cDNAPreparer**

Name (see format at top of page) of the person who was in charge of the second strand cDNA preparation (try not to split these, or we will have to add more fields!). If more than one person was involved, just pick one.

**s2cDNASampleNumber**

Sequential positive integers uniquely identifying the double-stranded cDNA samples prepared by a given preparer on a given date.

**Protocol**

Entries, to be defined, indicating the kit or protocol used for generating the second strand. Should be a single string of letters with no whitespace. Add any entries used below, with an explanation.

* NEBNextUltraIIE7550L -- NEBNext Ultra II Directional RNA Second Strand Synthesis Module Thermo catalog number E7550L
* …...

**PooledSecondStrand**

“TRUE” (with that capitalization) if the second strand sample contains multiple first strand samples, “FALSE” otherwise.

**roboticS2Prep**

“TRUE” (with that capitalization) if the robot was used, “FALSE” if not.

**First strand pooling table**

Each 1st strand sample may go into multiple pools for 2nd strand synthesis and each second strand sample may pool many 1st strand samples. Since this is potentially a many-to-many relationship, we need a separate table containing one entry for (1stStrand, 2ndStrand) pair. If there is no pooling, this will just be 1-to-1 1st strand to 2nd strand.

**s1cDNADate, s1cDNAPreparer,** **s1cDNASampleNumber**

The entries in these three columns should exactly match the corresponding entries for one and only row of a first strand cDNA sample table. Date and name columns should conform to the formats at the top of the document. Sample number should be a positive integer. In the simplest, these columns can be cut and pasted from that table.

**s2cDNADate**, **s2cDNAPreparer**, **s2cDNASampleNumber**

The entries in these three columns should exactly match the corresponding entries for one and only row of a second strand cDNA sample table. Date and name columns should conform to the formats at the top of the document. Sample number should be a positive integer. In the simplest, these columns can be cut and pasted from that table.

**Library Table**

Each entry denotes one library, even if multiple aliquots are made and sequenced at different times. A library means something with Illumina sequencing primers, indices or barcodes as needed, and ready to pool and put into a lane.

**s2cDNADate**, **s2cDNAPreparer**, **s2cDNASampleNumber**

The entries in these three columns should exactly match the corresponding entries for one and only row of a second strand cDNA sample table. Date and name columns should conform to the formats at the top of the document. Sample number should be a positive integer. In the simplest, these columns can be cut and pasted from that table.

**libraryDate**

The date (see format at top of document) on which the sequencing library was prepared and ready for use. Provide the date when the library was completed.

**libraryPreparer**

Name (see format at top of document) of the person who was in charge of the sequencing library preparation. If more than one person was involved, just pick one.

**librarySampleNumber**

Sequential positive integers uniquely identifying the libraries prepared by a given preparer on a

given date.

**index1Name**

E.g. “Index1\_2”.

**index1Sequence**

The sequence of index one (all caps), used to help distinguish samples pooled into a single lane. “NA” if no indices are used.

**index2Name**

E.g. “SICindex\_06”.

**index2Sequence**

The sequence of index two (all caps), used to help distinguish samples pooled into a single lane. “NA” if zero or one indices are used.

**libraryProtocol**

Entries, to be defined, indicating the kit or protocol used for generating the library. Should be a single string of letters with no whitespace.

Add any entries used below, with an explanation.

E7760 - NEBNext Ultra II Directional RNA Library Prep kit for Illumina

**roboticLibraryPrep**

“TRUE” (with that capitalization) if the robot was used, “FALSE” if not.

**fastq File Table**

One entry for each fastq file generated by sequencing a given library in a given lane. If the same library is sequenced multiple times, each run will have a different entry. Deconvolution of pooled libraries based on Illumina indices or other indices or barcodes may be needed.

**libraryDate,** **libraryPreparer**, **librarySampleNumber**

The entries in these three columns should exactly match the corresponding entries for one and only row of a library table. Date and name columns should conform to the formats at the top of the document. Sample number should be a positive integer. In the simplest, these columns can be cut and pasted from that table.

**runNumber**

The run number assigned by the sequencing core for the sequence run (not the spike-in run). This should be a positive integer

**laneNumber**

The lane within the run. If run only contains one separate sample, enter 1. This should be a positive integer.

**sequencerModel**

* MiSeq
* MiniSeq (all of our spike-ins are MiniSeq)
* NextSeq (all of our full runs are NextSeq)
* …..

**flowcellType**

* V3
* Standard
* Nano
* MiniSeq (the only type of flow cell used during spike-ins)
* HighOutput (used for NextSeq when there are 400M read runs)
* MidOutput (used for NextSeq when there are 130M read runs)

**purpose**

* Rebalancing
* spikein
* fullRNASeq
* fullgDNASeq
* fullChIPSeq
* ….

**tapestationConc**

A decimal number indicating the nanomolar reading from the tapestation.

**volumePooled**

The volume that was put into the pool that became the lane. A decimal number.

**readsObtained**

The number of reads obtained. A decimal number.

**fastqFileName**

The name (a string, no whitespace) of the fastqc file containing the reads. Do not include the full path, only the filename with extension.