Last updated on September 25th, 2018

EMP datatables in Babase

There are 3 views that contain data related to the EMP collaboration project. These views are emp_tissue, emp_dna, and emp_library. For the regular user the views should contain all of the information needed to link sequencing results with sample metadata. These 3 views get data from each other and a range of support tables. Some of these tables are part of the general ABRP database (mainly the genetic inventory schema), whilst others were created as support tables for the EMP project. The tables are discussed in detail in the Appendix for tables. If you have any questions regarding these views and/or tables please contact David Jansen (david.awam.jansen@gmail.com) or Beth Archie (beth.archie@gmail.com).

The original emp_collaboration table is obsolete and no longer in use.

Table of contents

Last updated on September 25th, 2018

Tables

tissue (genetic_inventory)	Data on all fecal samples as collected in the field
dna (genetic_inventory)	Information on all DNA samples/extractions of the ABRP project
emp_barcode	The barcode map of the primer plates used by EMP
emp_contamination_statuses	All possible contamination statuses
emp_dna_support	A support table to create the emp_dna view
emp_extract_dna_concentrations	A table with all available DNA concentrations after extraction
emp_library_support	A support table to create the emp_library view
emp_post_pcr_dna_concentrations	A table with all available DNA concentrations after PCR
emp_primerplate_plates	Overview of primer plates that were used for the loading plates.
emp_sample_statuses	An overview of possible sample statuses
emp_sequencing_support	An overview of data related to sequencing runs
emp_tissue_support	An support table to create the emp_tissue view

EMP sample statuses

The possible statuses of the sample in the well (see appendix EMP sample statuses for details).

Column descriptions

sample_status

This column indicated the status of the sample.

description

A description of the sample status.

EMP contamination statuses

The possible contamination statuses of the sample in the well (see appendix EMP contamination statuses for details).

Column descriptions

contamination_status

This column indicates if the sample is suspected to be contaminated. If contaminated the value will indicate at what stage in sample processing the sample was contaminated. (see appendix EMP contamination statuses for details).

description

A detailed description of the contamination status.

Last updated on September 25th, 2018

Notes the use of on sample and contamination statuses

All the views (and some of the supporting tables) report both a sample and contamination status (hereafter status). It is important for users to note that the these statuses are not necessarily the same at all stages of the data generation, from tissue sample storage to DNA extraction, to the libraries. The statuses in emp_tissue reflect the statuses of the fecal samples. The statuses in emp_dna reflect the statuses of the archived DNA extracts. The statuses of the samples in emp_library reflect the statuses of the PCR products that were sequenced. For projects that are using the sequencing data from 2018, it is recommended to use the sample status in emp_library. If you are trying to use the archived dna for future projects the recommendation is to use emp_dna in combination with the dna table (in the genetic inventory schema). The sample statuses are ordered to ease sorting and selecting. For most analysis users will want to restrict data to below 2. If technical replicates are desired that sorting can include sample status 2. The details of the value in the sample_status can be found here and for the contamination here.

Views

There are 3 views that combine general project data and data specific for the EMP project.

EMP_TISSUE

All of the data related to the fecal samples used in the EMP project can be found in the *emp_tissue* view. The *emp_tissue* view extracts subsets of data from 3 different tables: tissue (from the *genetic_inventory* schema); prep (from the *fecal* schema); and *emp_tissue_support*. The *emp_tissue_support* is specific to the EMP sequencing data and project.

Query to see the emp_tissue view

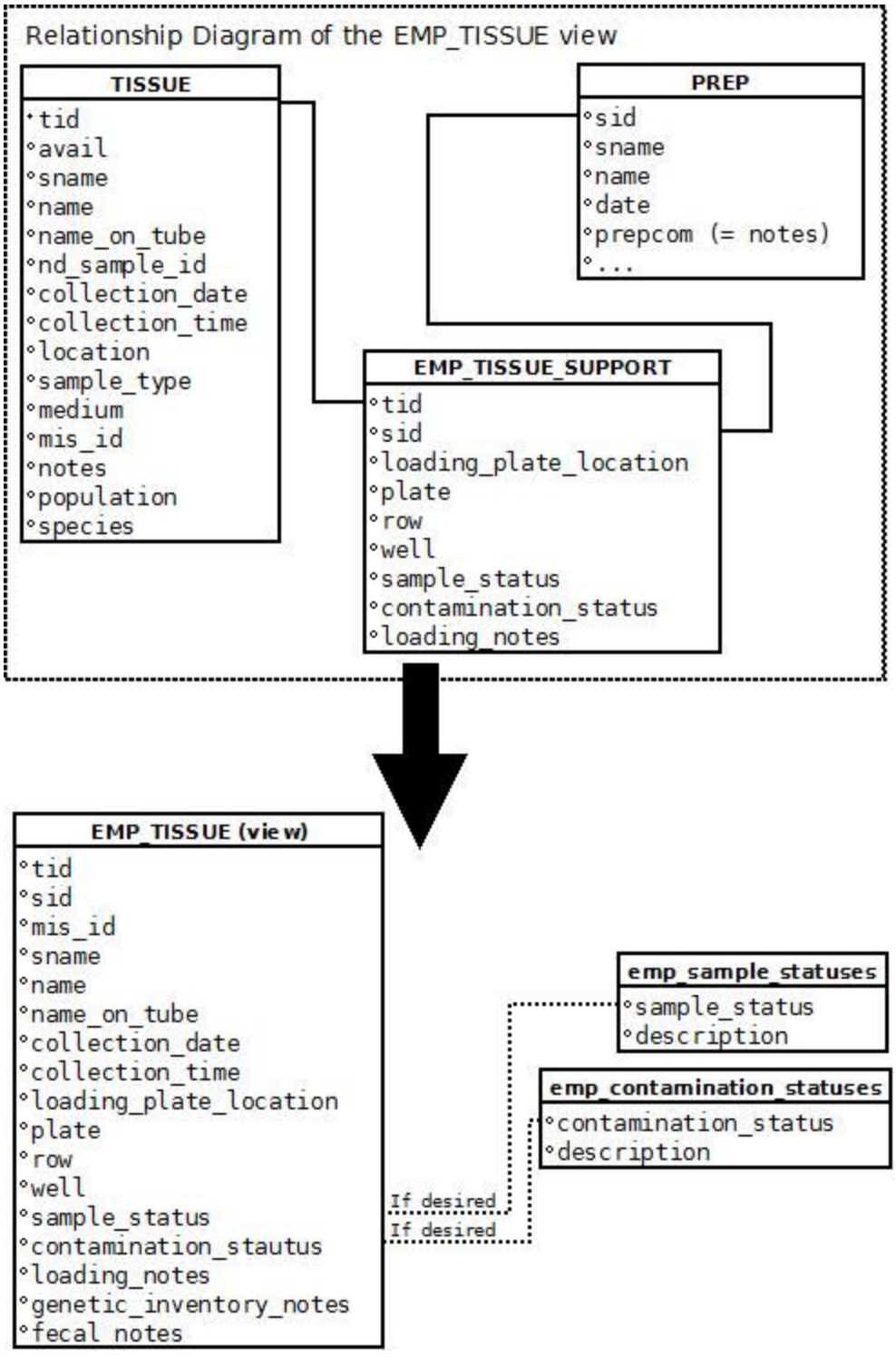
```
SELECT *
FROM genetic_sequencing.emp_tissue;
```

Definition of the emp_tissue view

```
WITH unique_fecal_notes AS (
        SELECT prep.sid,
          prep.prepcom
         FROM fecal.prep
        GROUP BY prep.sid, prep.prepcom
       )
 SELECT tissue.tid,
    emp.sid,
    tissue.mis_id,
    tissue.sname,
    tissue.name,
    tissue.name_on_tube,
    tissue.collection_date,
    tissue.collection_time,
    emp.loading_plate_location,
    emp.plate,
    emp."row",
    emp.well,
    emp.sample_status,
    emp.contamination_status,
    emp.loading_notes,
    tissue.notes AS genetic_inventory_notes,
    fecal.prepcom AS fecal_notes
```

```
  FROM emp_tissue_support emp
  LEFT JOIN genetic_inventory.tissue tissue ON tissue.tid = emp.tid
    LEFT JOIN unique_fecal_notes fecal ON fecal.sid = emp.sid
ORDER BY emp.plate, emp."row", emp.well;
```

Relational diagram of the emp_tissue view



Last updated on September 25th, 2018

Columns in the emp_tissue view

Column	From	Description
tid	genetic_inventory.tissue	Unique identifier for the fecal sample
sid	fecal.prep	Unique identifier for the fecal sample used for hormone analysis
mis_id	genetic_inventory.tissue	Indicator of the mis-identification status of the fecal sample
sname	genetic_inventory.tissue	The short name of the individual from whom this fecal sample was collected
name	genetic_inventory.tissue	The name of the individual from whom this fecal sample was collected
name_on_tube	genetic_inventory.tissue	Label on the fecal sample tube
collection_date	genetic_inventory.tissue	Collection date of the fecal sample
collection_time	genetic_inventory.tissue	Collection time of the fecal sample
loading_plate_location	tissue_support	The location (i.e. plate, row, and well) where the fecal sample aliquot was allocated to in the loading plate
plate	tissue_support	The plate the fecal sample aliquot was allocated to
row	tissue_support	The row the fecal sample aliquot was allocated to
well	tissue_support	The well the fecal sample aliquot the was allocated to
sample_status	tissue_support	The status of the fecal sample aliquot (i.e. normal sample, technical replicate, or contaminated)
contamination_status	tissue_support	The contamination status of the fecal sample aliquot
loading_notes	tissue_support	Notes on the allocating of the fecal sample aliquots
genetic_inventory_notes	genetic_inventory.tissue	Notes on the fecal sample from genetic_inventory
fecal_notes	fecal.prep	Notes on the fecal sample from fecal.prep

Column descriptions

Columns with an asterisk (*) come from the tissue table in the genetic_inventory schema.

sid

The sid is the sample id of the fecal samples that were used for hormone analysis. Using the sid, information regarding the original fecal sample can be pulled from the fecal.prep table. For some of the older fecal samples (sid < 3000) multiple tubes can have the same sid as there are cases where the original fecal sample was aliquoted into more than one tube of freeze-dried fecal powder. Thus, in some cases the name_on_tube column (see below) have an additional character/number, e.g., A or B or .1 or .2. However, each fecal sample tube has a unique tid (see below). The value is NULL for blank and/or contaminated sample aliquots.

tid*

The tid is the tissue sample id. This is the tube-specific and unique id for the fecal sample. The tid forms the relationship to the tissue table. The value is NULL for contaminated samples.

sname (Short Name)*

The sname is the abbreviated name for the focal animal which the fecal sample was collected from. The sname is only available for fecal samples of known animals. The value is NULL for blank and/or contaminated samples.

name*

The full name of the animal the fecal sample was collected from (see sname above). This value is NULL for contaminated wells.

name_on_tube*

The name_on_tube is the number that is written on the tube that contains the fecal powder from the focal fecal sample. In most cases this is the same as the sid. As noted under sid, a single sid can be assigned to multiple tubes. In most cases where a single sid is represented by multiple tubes, the name_on_tube will have an additional character/number, e.g., A or B or .1 or .2. However, this nomenclature is not consistent, so sid or tid should be used to query unique samples instead of name_on_tube. This value is NULL for blank and/or contaminated samples.

collection_date*

The date (year-month-day) the fecal sample was collected. This value is NULL for blank and/or contaminated samples.

collection_time*

The time the fecal sample was collected. Time is not available for all fecal samples.

mis_id*

The mis-identification status of the DNA extract. A value greater than 0 indicates that there may be some identification issues with the sample. See descriptions in the *mis_ids* table in the *genetic_inventory* schema by query SELECT * FROM genetic_inventory.mis_ids. Only use the column if you have a good understanding of the genetic inventory system in Babase. Otherwise use the sample_status and contamination statuses columns as the mis_id statuses are integrated in them.

sample_status

The status of the fecal sample aliquot. Any value greater than 0 indicates whether the archived DNA sample is a blank, technical replicate, or had any other issues during processing. See appendix EMP sample statuses for details and here for notes on the use of the status of samples during data analysis/selection.

contamination_status

This column indicates whether the fecal sample aliquot is suspected to be contaminated or problematic in some way. Any value greater than 0 indicates the type of contamination. See appendix EMP contamination statuses and here or notes on the use of the status of samples during data analysis/selection. on the use of the status of samples during data analysis/selection.

loading_plate_location

This column indicates the plate location that the fecal sample aliquot was first allocated on prior to being sent to EMP for DNA extraction and sequencing. A value of e.g., “Plate 1, A.1” indicates that the sample can be found on plate 1, row A, and in well 1. For the ease of sorting and selecting, this information is also presented separately in the columns: plate; row; and well (see below).

plate

Numeric value of the plate the fecal sample aliquot was allocated on.

row

Alphabetic character (A-H) corresponding to the row the fecal sample aliquot was allocated on.

well

Numeric value corresponds to the well (1-12) where the fecal sample aliquot was allocated in.

genetic_inventory_notes

This column contains notes from the *genetic_inventory.tissue* table. The values are notes related to the fecal sample itself. Possible values are e.g., “DAMP ON ARRIVAL, FREEZE-DRIED AGAIN 13-A” and “SPE AGAIN ON 02/12/03”.

loading_notes

This column contains notes collected during the allocation of the fecal sample aliquots onto the 96-well plates for sequencing by the EMP project. Possible values are e.g., “SAMPLE EXHAUSTED” and “WELL CONTAMINATED DO NOT USE”.

Last updated on September 25th, 2018

EMP_DNA

All the data related to the DNA extracted in the emp study can be found in the *emp_dna* view. The *emp_dna* view extracts subsets of data from 4 different locations: *dna* table (from the *genetic_inventory* schema); *emp_tissue* view (from the *genetic_sequencing* schema); *emp_extract_dna_concentrations* table (from *genetic_sequencing* schema); and *emp_dna_support* table (from the *genetic_sequencing* schema).*emp_tissue*, *emp_extract_dna_concentrations* and *emp_dna_support* are specific to the EMP sequencing data and project.

Query of the emp_dna view

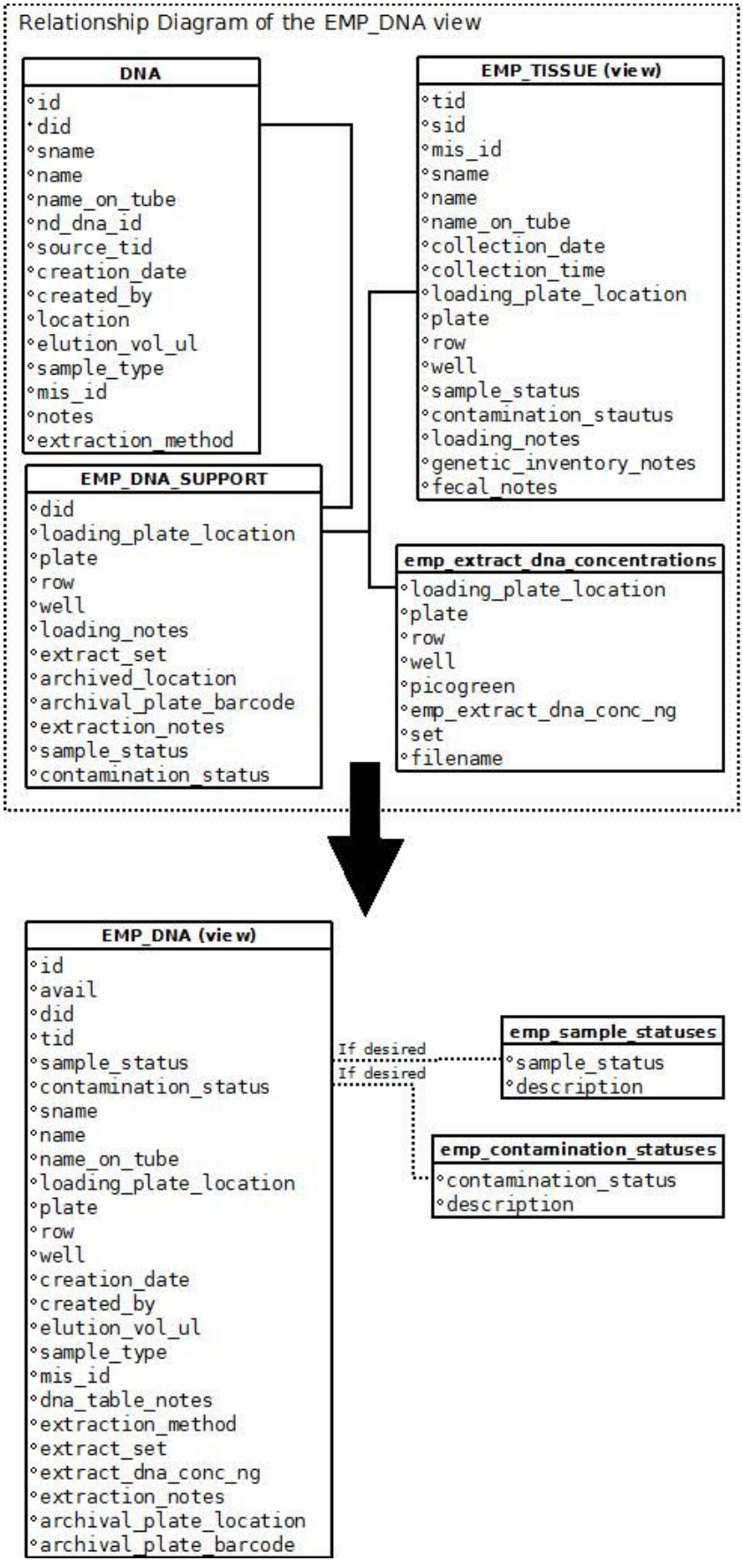
```
SELECT *
FROM genetic_sequencing.emp_dna;
```

Definition of the emp_dna view

```
SELECT dna.id,
    dna.avail,
    dna.did,
    emp.tid,
    dna_support.sample_status,
    dna_support.contamination_status,
    dna.sname,
    dna.name,
    dna.name_on_tube,
    dna_support.loading_plate_location,
    dna_support.plate,
    dna_support."row",
    dna_support.well,
    dna.creation_date,
    dna.created_by,
    dna.elution_vol_ul,
    dna.sample_type,
    dna.mis_id,
    dna.notes,
    dna.extraction_method,
    dna_support.extract_set,
    conc.extract_dna_conc_ng,
    dna_support.extraction_notes,
    dna.location AS archived_location,
    dna_support.archival_plate_barcode
   FROM emp_dna_support dna_support
     LEFT JOIN genetic_inventory.dna dna ON dna.did::double precision = dna_support.did::double precision
     LEFT JOIN emp_tissue tissue ON tissue.loading_plate_location::text = dna_support.loading_plate_location
     LEFT JOIN emp_extract_dna_concentrations conc ON conc.loading_plate_location = dna_support.loading_plate_location
  ORDER BY dna_support.plate, dna_support."row", dna_support.well;
```

Last updated on September 25th, 2018

Relational diagram of the emp_dna view



Last updated on September 25th, 2018

Columns in the emp_dna view

Column	From	Description
id	genetic_inventory.dna	Row identifier
avail	genetic_inventory.dna	Whether the archived DNA is available
did	genetic_inventory.dna	Unique identifier for the DNA extract
tid	genetic_inventory.dna	Unique identifier for the fecal sample
sample_status	emp_dna_support	The status of the archived DNA
contamination_status	emp_dna_support	The contamination status of the archived DNA
sname	genetic_inventory.dna	The short name of the individual
name	genetic_inventory.dna	Collection time of the fecal sample
name_on_tube	genetic_inventory.dna	Label on the archived DNA tube
loading_plate_location	emp_dna_support	The location (i.e. plate, row, and well) of the DNA extract allocation
plate	emp_dna_support	The plate DNA extract was allocated to
row	emp_dna_support	The row DNA extract the was allocated to
well	emp_dna_support	The well DNA extract the was allocated to
created_date	genetic_inventory.dna	The date of when DNA extract was extracted
created_by	genetic_inventory.dna	The initials of the person who extracted the DNA sample
elution_vol_ul	genetic_inventory.dna	Volume in microliters of the DNA extract when first created
sample_type	genetic_inventory.dna	The type of DNA extracted
mis_id	genetic_inventory.tissue/genetic_inventory.dna	Indicator of the mis-identification status of the extract
notes	genetic_inventory.dna	Any notes that come from the DNA table
extraction_method	genetic_inventory.dna	The method used to extract the DNA
extraction_set	emp_dna_support	The set of the DNA extracted
extract_dna_conc_ng	emp_extract_dna_concentrations	The DNA concentration in nanograms after extraction
extraction_notes	emp_dna_support	Notes on the extract
archival_plate_location	emp_dna_support	The location (i.e. plate, row, and well) of the archived DNA

archival_plate_barcode	emp_dna_support	The barcode of the archival plate

Column descriptions

Columns marked with a * come from the tissue table in the genetic_inventory schema.

id*

This is an automatically generated sequential number that uniquely identifies the row in the DNA table.

avail*

Indicator of the availability of the archived DNA.

did*

Unique identifier for the DNA extract, generated by Babase project managers.

tid (source_tid)*

The tid is the tissue id. This is the unique id for the specific tube of fecal sample used to generate this DNA extract. This id matches the tissue information in the tissue table. For contaminated wells this value is NULL.

sample_status

The status type of the archived DNA sample. Any value greater than 0 indicates whether the archived DNA sample is a blank, technical replicate, or had any other issues during processing. See appendix EMP sample statuses for details and here for notes on the use of the status of samples during data analysis/selection.

contamination_status

This column indicates whether the archived DNA sample is suspected to be contaminated or problematic in some way. Any value greater than 0 indicates the type of contamination. See appendix EMP contamination statuses and here or notes on the use of the status of samples during data analysis/selection. on the use of the status of samples during data analysis/selection.

sname (Short Name)*

The short name of the individual the fecal sample was collected from. The sname is only available for samples of known individuals. For blank and contaminated wells this value is NULL.

name*

The name of the individual the fecal sample was collected from. For contaminated wells this value is NULL.

name_on_tube*

The name_on_tube is the label that is written on the tube used for the archived DNA sample.

loading_plate_location

The location of the sample allocation. This is the location that the fecal sample aliquot was allocated in the plates sent to EMP for DNA extraction and sequencing. The information in this column is also presented in the next 3 columns for ease of sorting and selecting. With the exception of plate 8, all H.12 wells are blank. In plate 8, well H.12 was assigned a sample and plate 8 does not have a blank.
plate: Numeric value of the plate.

row: Alphabetic character between A and H for the row of on the plate of the sample
well: Numeric value for the well of the sample on the plate.

created_date*

The date on which the DNA was extracted.

created_by*

The initials of the person that extracted the DNA.

elution_vol_ul*

Volume in microliters of the sample when first extracted.

sample_type*

The type of DNA extracted.

mis_id*

The mis-identification status of the DNA extract. Any value greater than 0 in this column indicates there may be some identification issues with the sample. See descriptions in the *mis_ids* table in the *genetic_inventory* schema by query SELECT * FROM genetic_inventory.mis_ids. Only use the column if you have a good understanding of the genetic inventory system in Babase, otherwise use the sample_status and contamination statuses columns as the mis_id statuses are integrated in them.

notes

Notes that come from the *dna* table. These notes are taken pre DNA extraction and may be related to issues such as the fecal sample's storage conditions and availability.

extraction_method*

The method used to extract the DNA. All of the EMP DNA extractions were done using a modified MoBio PowerSoil HTP Kit Extraction. Modifications were as follows: 950 µL of PowerBead Solution was used instead of 750 µL. Once C1 and PowerBead Solution were added, plates were incubated for 10 minutes in a water bath set to 60°C. All plates were centrifuged at 3100 rpm (the Qiagen protocol specifies 3500 rpm, but samples were spun for longer to account for this). Samples were eluted in 100µL C6 elution buffer just once.

extraction_set

The set of the DNA extracted. The EMP DNA extractions were completed in randomized pairs, and the extraction_set value indicates which plates were paired.

extract_dna_conc_ng

The DNA concentration of the DNA extract after extraction. DNA extracts were quantified using picogreen. The picogreen value was divided by 750 to approximate the amount of nanograms extracted.

extraction_notes

Notes taken during the DNA extraction. May be related to contamination status of the well/plate.

archival_plate_location* (location in dna table)

The location at which the DNA sample was archived. In most cases this will be the same as the loading_plate_location. The archival_plate_location is the location in the *dna* table.

archival_plate_barcode

The unique, computer-generated barcode on the archival plate.

EMP_LIBRARY

Query of the emp_library view

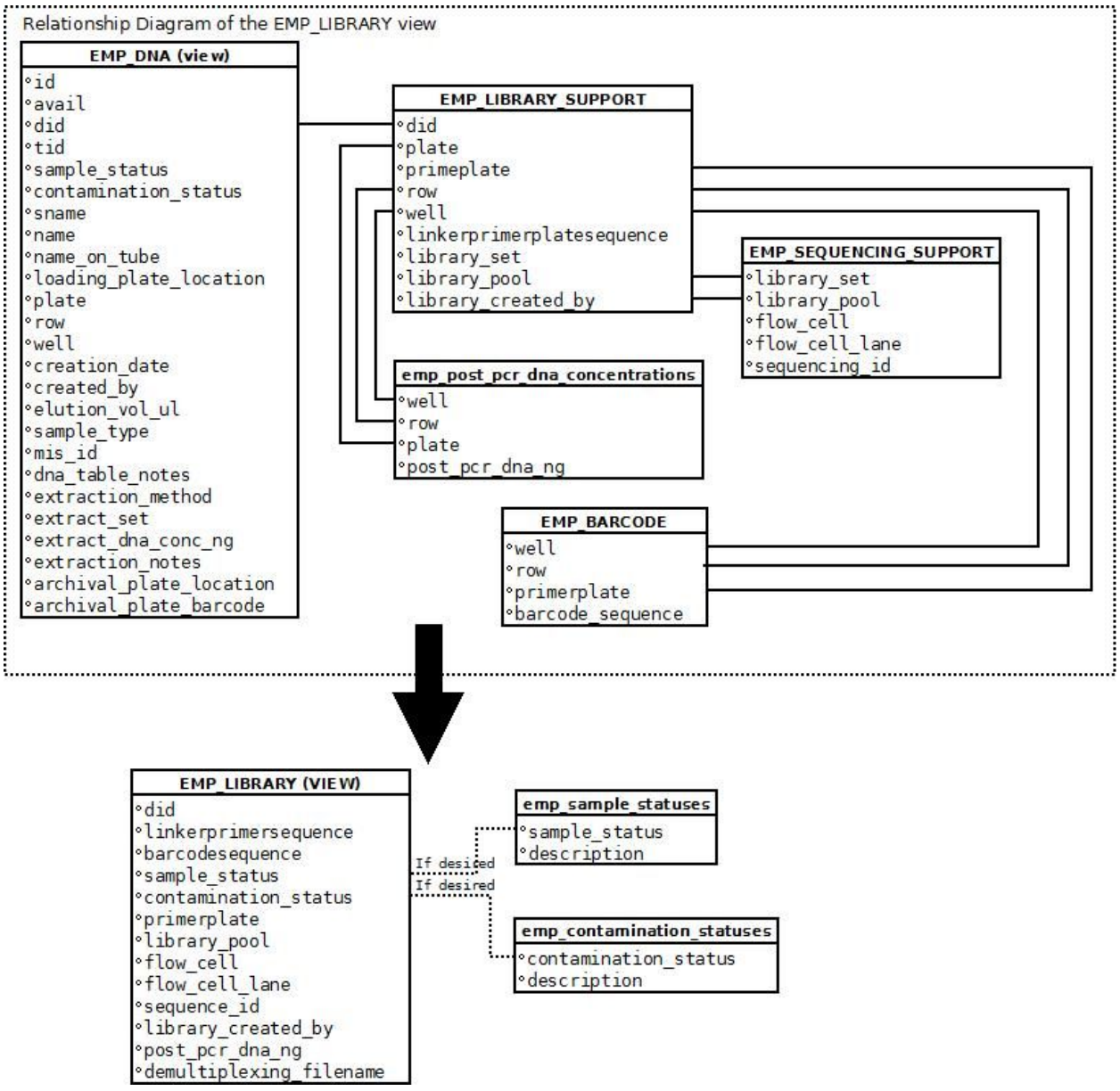
```
SELECT *
FROM genetic_sequencing.emp_library;
```

Definition of the emp_library view

```
SELECT dna.did,
    library.linkerprimersequence,
    barcode.barcodesequence,
    library.sample_status,
    library.contamination_status,
    library.primerplate,
    sequencing.library_pool,
    sequencing.flow_cell,
    sequencing.flow_cell_lane,
    sequencing.sequencing_id,
    library.library_created_by,
    pcr.post_pcr_dna_ng,
    concat(sequencing.sequencing_id, '_', barcode.barcodesequence, '.fastq') AS demultiplexing_filename
   FROM emp_library_support library
      JOIN emp_barcode barcode ON barcode.primerplate = library.primerplate::double precision AND barcode."row" = library."row"
AND barcode.well = library.well::double precision
     JOIN emp_dna_support dna ON dna.plate = library.plate AND dna."row" = library."row" AND dna.well = library.well
       LEFT JOIN emp_post_pcr_dna_concentrations pcr ON pcr.plate = library.plate AND pcr."row" = library."row" AND pcr.well =
library.well
      LEFT JOIN emp_sequencing_support sequencing ON library.library_pool = sequencing.library_pool::text and library.repeated =
sequencing.repeated
  ORDER BY library.plate, library."row", library.well;
```

Last updated on September 25th, 2018

Relational diagram of the emp_library view



Last updated on September 25th, 2018

Columns in the emp_library view

Column	From	Description
did	emp_dna	Identifier for the dna extract
linkerprimersequence	emp_library_support	The sequence of the linker primer for the library
barcodesequence	emp_library_support	The unique barcode that belongs to a unique well in the library pool
sample_status	emp_library_support	The status of the DNA extract
contamination_status	emp_library_support	The contamination status of the DNA extract
primerplate	emp_library_support	The number of the primer plate
library_set	emp_library_support	The set of the libraries that was sequenced as a set
library_pool	emp_library_support	An identifier of the library pool
flow_cell	emp_sequencing_support	An identifier of the flow cell
flow_cell_lane	emp_sequencing_support	An identifier of the lane used in the flow cell
sequecing_id	emp_sequencing_support	An unique identifier for the flow cell and flow cell lane combination
created_by	emp_library_support	The initials of the person who created the library
post_pcr_dna_ng	emp_post_dna_concentrations	The DNA concentration of the amplified DNA after PCR
demultiplexing_filename	emp_sequencing_support	Name of the fastq file generated by the demultiplexing

Column descriptions

did

This is a unique identifier of the DNA extract.

linkerprimersequence

The Illumina linker and primer sequence used during sampling. The same linker primer sequence is used for all library samples.

barcodesequence

The barcode sequence used for library sequencing. Barcode sequences are unique to each library sample within each library pool, but may be repeated between pools.

sample_status

The status type of the sequenced DNA sample. Any value greater than 0 indicates whether the archived DNA sample is a blank, technical replicate, or had any other issues during processing. See appendix EMP sample statuses for details and here for notes on the use of the status of samples during data analysis/selection.

contamination_status

This column indicates whether sequenced DNA sample is suspected to be contaminated or problematic in some way. Any value greater than 0 indicates the type of contamination. See appendix EMP contamination statuses and here or notes on the use of the status of samples during data analysis/selection. on the use of the status of samples during data analysis/selection.

primerplate

The number of the primer plate used when creating the library.

library_set

The set of library pools that were generated at the same time.

library_pool

All of the plates in the same library_pool were added to the same library.

flow_cell

The identifier used for the flow cell used for the sequencing of the library.

flow_cell_lane

The identifier used for the flow cell lane used in the flow cell.

sequencing_id

A unique identifier created by the sequencing facility for any unique combination of flow cell and flow cell lane.

created_by

The person who created the libraries.

post_pcr_dna_ng

The concentration of DNA (ng) in the DNA sample after PCR, but before the PCR product cleanup step was run. This is not available for all plates.

demultiplexing_filename

Name of the fastq file generated by the demultiplexing. It's a combination of the sequencing_id and the barcode of the sample. Each fastq file name is unique.

Last updated on September 25th, 2018

Appendix for tables

TISSUE (genetic inventory)

This table contains one row for every tissue sample that is or ever has been in our possession.

Columns in the tissue table

Column	Description
tid	Unique identifier for the tissuesample
avail	Indicator of the availability of the tissue sample
sname	The short name of the individual from whom this tissue sample was collected
name	The name of the individual from whom this tissue sample was collected
name_on_tube	Label on the tissue sample tube
nd_sample_id	Sample id used at the University of Notre Dame
collection_date	Collection date of the tissue sample
collection_time	Collection time of the tissue sample
location	Location where the tissue sample is stored.
sample_type	Type of tissue sample
medium	Medium the sample is stored in
mis_id	Indicator of the mis-identification status of the tissue sample
notes	Comments or miscellaneous information about this tissue sample
species	The species of the tissue sample
population	The population were the sample was collected

Column descriptions

Columns with an asterisk (*) come from the tissue table in the genetic_inventory schema.

tid*

The tid is the tissue sample id. This is the tube-specific and unique id for the fecal sample. The tid forms the relationship to the tissue table. The value is NULL for contaminated samples.

avail

Indicating the availability of the tissue sample.

sname (Short Name)*

The sname is the abbreviated name for the focal animal which the fecal sample was collected from. The sname is only available for fecal samples of known animals. The value is NULL for blank and/or contaminated samples.

name*

The full name of the animal the fecal sample was collected from (see sname above). This value is NULL for contaminated wells.

name_on_tube*

The name_on_tube is the number that is written on the tube that contains the fecal powder from the focal fecal sample.

nd_sample_id

Sample id used at the University of Notre Dame.

collection_date*

The date (year-month-day) the fecal sample was collected. This value is NULL for blank and/or contaminated samples.

collection_time*

The time the fecal sample was collected. Time is not available for all fecal samples.

location

Storage location of the tissue sample.

sample_type

Type of tissue sample.

medium

Medium sample is stored in.

mis_id*

The mis-identification status of the DNA extract. Any value greater than 0 in this column indicates there may be some identification issues with the sample. See descriptions in the *mis_ids* table in the *genetic_inventory* schema by query SELECT * FROM genetic_inventory.mis_ids. Only use the column if you have a good understanding of the genetic inventory system in BabaseBabase., otherwise use the sample_status and contamination statuses colums as the mis_id statuses are integrated in them.

notes

Comments or miscellaneous information about this tissue sample.

species

Species the sample was collected from.

population

Population and/or study site the sample was collected from.

Last updated on September 25th, 2018

DNA (genetic inventory)

Columns in the dna table

Column	Description
id	Row identifier
avail	Whether the archived DNA is available
did	Identifier for the dna extract
tid	Identifier for the tissue sample
sname	The short name of the individual from whom this fecal sample was collected
name	Collection time of the fecal sample
name_on_tube	Label on the archived DNA tube
created_date	The date of when DNA extract was extracted
created_by	The initials of the person that extracted the DNA sample
elution_vol_ul	Volume in microliters of the DNA extract when first created
sample_type	The type of DNA extracted
mis_id	Indicator of the mis-identification status of the extract
notes	Any notes that come from the DNA table
extraction_method	The method used to extract the DNA

Column descriptions

Columns marked with a * come from the genetic_inventory schema.

id*

This is an automatically generated sequential number that uniquely identifies the row in the DNA table.

avail*

Indicator of the availability of the archived DNA.

did*

Unique identifier for the DNA extract, generated by Babase project managers.

tid (source_tid)*

The tid is the tissue id. This is the unique id for the specific tube of fecal sample used to generate this DNA extract. This id matches the tissue information in the tissue table. For contaminated wells this value is NULL.

sname (Short Name)*

The short name of the individual the tissue sample was collected from. The sname is only available for samples of known individuals. For blank and contaminated wells this value is NULL.

name*

The name of the individual the tissue sample was collected from. For contaminated wells this value is NULL.

name_on_tube*

The name_on_tube is the label that is written on the tube used for the archived DNA sample.

created_date*

The date on which the DNA was extracted.

created_by*

The initials of the person that extracted the DNA.

elution_vol_ul*

Volume in microliters of the sample when first extracted.

sample_type*

The type of DNA extracted

mis_id*

The mis-identification status of the DNA extract. Any value greater than 0 in this column indicates there may be some identification issues with the sample. See descriptions in the *mis_ids* table in the *genetic_inventory* schema by query SELECT * FROM genetic_inventory.mis_ids. Only use the column if you have a good understanding of the genetic inventory system in Babase., otherwise use the sample_status and contamination statuses colums as the mis_id statuses are integrated in them.

extraction_method*

The method used to extract the DNA. All of the EMP DNA extractions were done using a modified MoBio PowerSoil HTP Kit Extraction. Modifications were as follows: 950 µL of PowerBead Solution was used instead of 750 µL. Once C1 and PowerBead Solution were added, plates were incubated for 10 minutes in a water bath set to 60°C. All plates were centrifuged at 3100 rpm with times adjusted to reflect speed difference. Samples were eluted in 100µL C6 elution buffer just once.

archival_plate_location* (location in dna table)

The location at which the DNA sample was archived. In most cases this will be the same as the loading_plate_location. The archival_plate_location is the location in the *dna* table.

Last updated on September 25th, 2018

EMP_BARCODE

This table is the plate maps of the barcodes as used by EMP.

Note: The manufacturer had accidentally duplicated one of the barcodes CACCGAAATCTG (plate 11), and that effectively shifted the order of all the barcodes down by one. We don't use plate 11 for that reason, and that final unassigned barcode (TAGAGTGTAACA) isn't used anywhere

Columns in the emp_barcode table

Column	Description
primerplate	The number of the primerplate
row	The row of the primerplate
well	The well of the primerplate
barcodesequence	The unique barcode that belongs to a unique well in the library pool

Column descriptions

primerplate

The number of the primer plate used when creating the library.

row

Alphabetic character between A and H corresponding to the primerplate row the barcode.

well

Numeric value corresponds to the actual well number (1 - 12) on the primerplate of the barcode.

barcodesequence

The barcode sequence used for library sequencing. Barcode sequences are unique to each library sample within each library pool, but may be repeated between pools.

Last updated on September 25th, 2018

EMP_CONTAMINATION_STATUSES

The possible contamination statuses of the sample in the well

Columns in the emp_contamination_statuses table

Column	Description
contamination_status	The contamination status
description	The description of the contamination

Column descriptions

contamination_status

This column indicates if the sample is suspected to be contaminated. If contaminated the value will indicate at what stage of sample processing the sample was contaminated.

description

A detailed description of the contamination status.

Value	Description
0	Not contaminated
0.5	Sample had unusual properties that could affect the microbial community; e.g. sample was not completely freeze dried or was part of a storage experiment so exposed to different environmental conditions.
1	There is a known problem with the identity of the animal the sample was collected from. This includes all cases where the sample was genotyped and the genotype of the sample did not match the known genotype of the animal in BabaseBabase. These samples all have a mis_id score of 2 in the tissue table. It also includes all samples that had a comment such as *“This could also be x”*.
2	The sample is collected from an animal with an unknown sname (e.g. an animal not from one of our study groups).
3	Unexplained duplicate sid. We strongly suspect that one of each pair of duplicates has a typo in the SID, but we do not know which sample has the typo. Eventually, we will check these for typos see e.g. the code David developed for sid = 17679.
4	Sample was contaminated before loading into plates for DNA extraction. This may have happened during a prior hormone extraction or other sample use before DNA extraction. This includes switched tubes and caps or comments like *"lids switched, fixed now"* as this could contaminate the samples.
5	Sample well was contaminated while we were loading fecal powder into DNA extraction plates.
6	Sample well was contaminated during storage prior to DNA extraction or during the DNA extraction.
7	DNA extract was contaminated during PCR or library preparation.
8	DNA extract was contaminated during DNA archiving.

Last updated on September 25th, 2018

EMP_DNA_SUPPORT

Columns in the emp_dna_support view

Column	Description
did	Identifier for the DNA extract
sample_status	The status of the archived DNA
contamination_status	The contamination status of the archived DNA
loading_plate_location	The location of the DNA extract allocation
plate	The plate DNA extract was allocated to
row	The row DNA extract the was allocated to
well	The well DNA extract the was allocated to
notes	Any notes that come from the DNA table
extraction_set	The set of the DNA extracted
extract_dna_conc_ng	The DNA concentration of the extract after extraction
extraction_notes	Miscellaneous notes about the extract
archival_plate_barcode	The barcode on the archival plate

Column descriptions

did

Unique identifier for the DNA extract, generated by Babase project managers.

sample_status

The status type of the archived dna sample.. Any value greater than 0 indicates whether the archived DNA sample is a blank, technical replicate, or had any other issues during processing. See appendix EMP sample statuses for details and here for notes on the use of the status of samples during data analysis/selection.

contamination_status

This column indicates whether the archived dna sample is suspected to be contaminated or problematic in some way. Any value greater than 0 indicates the type of contamination. See appendix EMP contamination statuses and here or notes on the use of the status of samples during data analysis/selection. on the use of the status of samples during data analysis/selection.

loading_plate_location

The location of the sample allocation. This is the location that the fecal sample aliquot was allocated in the plates sent to EMP for DNA extraction and sequencing. The information in this column is also presented in the next 3 columns for ease of sorting and selecting. With the exception of plate 8, all H.12 wells are blank. In plate 8 well H.12 was assigned a sample and plate 8 does not have a blank.

Last updated on September 25th, 2018

plate

Numeric value of the plate the fecal sample aliquot was allocated on.

row

Alphabetic character between A and H corresponding to the plate row the fecal sample aliquot was allocated on.

well

Numeric value corresponds to the actual well number (1 - 12) where the fecal sample aliquot was allocated in.

extraction_set

The set of the dna extracted. The EMP DNA extractions were completed in randomized pairs, and the extraction_set value indicates which plates were paired.

extract_dna_conc_ng

The DNA concentration of the DNA extract after extraction. DNA extracts were quantified using picogreen. The picogreen value was divided by 750 to approximate the amount of nanograms extracted.

extraction_notes

Notes taken during the DNA extraction. May be related to contamination status of the well/plate.

archival_plate_barcode

The unique, computer-generated barcode on the archival plate.

Last updated on September 25th, 2018

EMP_EXTRACT_DNA_CONCENTRATION

Columns in the emp_extact_dna_concentration table

Column	Description
loading_plate_location	The location of the DNA extract allocation
plate	The plate DNA extract was allocated to.
row	The row DNA extract the was allocated to.
well	The well DNA extract the was allocated to.
picogreen	The picogreen value of the DNA extraction
extract_dna_conc_ng	The DNA concentration of the extract after extraction
filename	The name of the file that contains the picogreen values

Column descriptions

loading_plate_location

This column indicates the plate location the fecal sample aliquot was first allocated to prior of being sent to EMP for DNA extraction and sequencing. A value of e.g., “Plate 1, A.1” indicates that the sample can be found on plate 1, row A, and in well 1. For the ease of sorting and selecting, this information is also presented separately in the columns: plate; row; and well (see below).

plate

Numeric value of the plate the fecal sample aliquot was allocated on.

row

Alphabetic character between A and H corresponding to the plate row the fecal sample aliquot was allocated on.

well

Numeric value corresponds to the actual well number (1 - 12) where the fecal sample aliquot was allocated in.

picogreen

The picogreen values of the DNA extract after extraction.

extract_dna_conc_ng

The DNA concentration of the DNA extract after extraction. DNA extracts were quantified using picogreen. The picogreen value was divided by 750 to approximate the amount of nanograms extracted.

filename

Filename of the file with the picogreen results

Last updated on September 25th, 2018

EMP_LIBRARY_SUPPORT

Columns in the emp_library_support table

Column	Description
did	Unique identifier of the dna sample extract
linkerprimersequence	The linkerprimersequence of the library
barcodesequence	The unique barcode that belongs to a unique well in the library pool.
primerplate	The number of the primerplate
library_set	The set of the libraries that was sequenced as a set
library_pool	An identifier of the library pool
created_by	The initials of the person who created the library
sample_status	The status of the sequenced DNA
contamination_status	The contamination status of the sequenced DNA
repeated	A boolean variable that indicates if this was a repeated sample

Column descriptions

did

This is a unique identifier of the dna extract.

linkerprimersequence

The Illumina linker and primer sequence used during sampling. The same linkerprimersequence is used for all library samples.

barcodesequence

The barcode sequence used for library sequencing. Barcode sequences are unique to each library sample within each library pool, but may be repeated between pools.

primerplate

The number of the primer plate used when creating the library.

library_set

The set of library pools that were generated at the same time.

library_pool

All of the plates in the same library_pool were added to the same library.

created_by

Who or whom created the libraries.

sample_status

The status type of the sequenced dna sample. Any value greater than 0 indicates whether the archived DNA sample is a blank, technical replicate, or had any other issues during processing. See appendix EMP sample statuses for details and here for notes on the use of the status of samples during data analysis/selection.

contamination_status

This column indicates whether the sequenced dna sample is suspected to be contaminated or problematic in some way. Any value greater than 0 indicates the type of contamination. See appendix EMP contamination statuses and here or notes on the use of the status of samples during data analysis/selection. on the use of the status of samples during data analysis/selection.

repeated

This column indicates whether these are the details belong to the library that was repeated in a second sequencing run. The samples that are not contaminated were given a sample_status of 2.5.

Last updated on September 25th, 2018

EMP_PLATE_STATUSES

Columns in the emp_plate_statuses table

Column	Description
plate	Unique identifier of the dna sample extract
pilot	Boolean variable that indicates if the plate was part of the emp pilot
replated	Boolean variable that indicates if the plate was replated
plate_notes	Notes on the issue with the plates

Column descriptions

plate

The number of the plate

pilot

The number of plates was done as a pilot to test dna extraction methods. Several issues where found with the dna of these plates. They were therefore redone.

replated

During shipping, storages and the pilot study some of the allocated plates were contaminated or lost. They were redone.

plate notes

Some notes on the issues with the plates.

Last updated on September 25th, 2018

EMP_POST_PCR_DNA_CONCENTRATIONS

This table contains the dna concentrations post-PCR. It may not be available for all plates.

Columns in the emp_post_pcr_dna_concentrations table

Column	Description
plate	The number of the loading_plate with the DNA extraction
row	The row of the loading_plate with the DNA extraction
well	The well of the loading_plate with the DNA extraction
post_pcr_dna_ng	Post PCR DNA concentration

Column descriptions

plate

The number of the loading_plate with the DNA extraction.

row

The row of the loading_plate with the DNA extraction.

well

The well of the loading_plate with the DNA extraction.

post_pcr_dna_ng

The concentration of DNA (ng) in the DNA sample after PCR, but before the PCR product cleanup step was run. This is not available for all plates.

Last updated on September 25th, 2018

EMP_PRIMERPLATE_PLATES

A table that indicates which plates were assigned to which primerplates.

Columns in the emp_primerplate table

Column	Description
primerplate	The number of the primer plate that was used to generate the libraries
plate	The number of the loading_plate with the DNA extractions

Column descriptions

primerplate

The number of the primer plate that was used to generate the libraries.

plate

The number of the loading_plate with the extracted DNA.

Last updated on September 25th, 2018

EMP_SAMPLE_STATUSES

The possible statuses of the sample in this specific well.

Columns in the emp_sample_statuses table

Column	Description
contamination_status	The sample status of the well
description	The description of the status

Column descriptions

sample_status

This column indicates the status of the sample in this specific well.

description

A description of the sample status.

Value	Description
0	normal sample
1	blank
2	technical replicate
2.5	These are the non contaminated samples of the 2nd sequencing run of the repeated library.
3	user beware: potentially contaminated or problematic sample, read the contamination_status and the notes
4	contaminated
5	plate was lost

Last updated on September 25th, 2018

EMP_SEQUENCING_SUPPORT

The possible statuses of the sample in this specific well.

Columns in the emp_sequencing_support table

Column	Description
library_pool	All of the plates in the same library_pool were added to the same library
library_set	The set of library pools that created at the same time
flow_cell	The identifier for the flowcell that was used in the sequencing run
flow_cell_lane	The lane that was used in the flowcell
sequencing_id	A unique identifier for the unique combination of the flowcell and flowcell_lane

Column descriptions

library_pool

All of the plates in the same library_pool were added to the same library.

library_set

The set of library pools that created at the same time.

flow_cell

The identifier for the flowcell that was used in the sequencing run.

flow_cell_lane

The lane that was used in the flowcell

sequencing_id

A unique identifier for the unique combination of the flowcell and flowcell_lane. This unique identifier is created by the sequencing facility.

Last updated on September 25th, 2018

EMP_TISSUE_SUPPORT

The *emp_tissue_support* is specific to the EMP sequencing data and project. It provides additional information on the fecal sample aliquots. Regular users should not use this table, but use the emp_tissue view.

Columns in the emp_tissue_support table

Column	Description
tid	Unique identifier for the fecal sample
sid	Unique identifier for the fecal sample used for hormone analysis
loading_plate_location	The location where the fecal sample aliquot was allocated to in the loading plate
plate	The plate the fecal sample aliquot was allocated to.
row	The row the fecal sample aliquot was allocated to.
well	The well the fecal sample aliquot the was allocated to.
sample_status	The status of the fecal sample aliquot (i.e. normal sample, technical replicate or contaminated)
contamination_status	The contamination status of the fecal sample aliquot
loading_notes	Notes related to the allocating of the fecal sample aliquots

Column descriptions

tid

The tid is the tissue sample id. This is the tube-specific and unique id for the fecal sample. The tid forms the relationship to the tissue table. The value is NULL for contaminated samples.

sid

The sid is the sample id of the fecal samples that were used for hormone analysis. Using the sid, information regarding the original fecal sample can be pulled from the fecal.prep table. For some of the older fecal samples (sid < 3000) multiple tubes can have the same sid as there are cases where the original fecal sample was aliquoted into more than one tube of freeze-dried fecal powder. Thus, in some cases the name_on_tube column (see below) have an additional character/number, e.g., A or B or .1 or .2. However, all fecal sample tubes have a unique tid (see below). The value is NULL for blank and/or contaminated sample aliquot.

sample_status

The status type of the fecal sample aliquot. Any value greater than 0 indicates whether the archived DNA sample is a blank, technical replicate, or had any other issues during processing. See appendix EMP sample statuses for details and here for notes on the use of the status of samples during data analysis/selection.

contamination_status

This column indicates whether the fecal sample aliquot is suspected to be contaminated or problematic in some way. Any value greater than 0 indicates the type of contamination. See appendix EMP contamination statuses and here or notes on the use of the status of samples during data analysis/selection. on the use of the status of samples during data analysis/selection.

loading_plate_location

This column indicates the plate location the fecal sample aliquot was first allocated to prior of being sent to EMP for DNA extraction and sequencing. A value of e.g., “Plate 1, A.1” indicates that the sample can be found on plate 1, row A, and in well 1. For the ease of sorting and selecting, this information is also presented separately in the columns: plate; row; and well (see below).

plate

Numeric value of the plate the fecal sample aliquot was allocated on.

row

Alphabetic character between A and H corresponding to the plate row the fecal sample aliquot was allocated on.

well

Numeric value corresponds to the actual well number (1 - 12) where the fecal sample aliquot was allocated in.

loading_notes

This column contains notes collected during the allocation of the fecal sample aliquots onto the 96-well plates for sequencing by the EMP project. Possible values are e.g., “SAMPLE EXHAUSTED” and “WELL CONTAMINATED DO NOT USE”.