Jan/20/2020

This document was created to briefly describe the tables available in the database.

Overall note: All `\*\_ag` tables have two mandatory fields: key and uploaded\_date.

-`key` is a unique identifier for each table's row.

-`uploaded\_date` is the date in which the row was uploaded.

Any table with `obsolete` naming on it is not under use and are kept to preserve previous information in case of need.

List of tables:

**alpha\_diversity\_ag** (Obsolete: This version uses OTU based pipeline. Use

*asv\_alpha\_diversity\_ag* for DADA2 based pipeline).

- It has the alpha diversity for each sampleid in terms of simpson\_reciprocal as well as shannon entropy.

**antibiotics\_antibacterial\_multicenter\_ag**

- This table has the antibiotics information for patients investigated in the multicenter study (Peled JU et al., NEJM in press) and some extra patients.

**antibiotics\_antimicrobial\_ordername\_map\_ag**

- This table translates antimicrobial ordernames (as we obtain on our request) to a clean version (order\_name\_clean).

**antibiotics\_category\_map\_ag**

- It describes the category of each order\_name. I have not been using this table, but it is kept alive for legacy and in case we need this reference.

**antibiotics\_medication\_obsolete\_ag**

- It has drug information for some patients. This table is labeled as obsolete as we have been using `antibiotics\_antibacterial\_multicenter\_ag` instead. It does contain MRN from a few patients not in `antibiotics\_antibacterial\_multicenter\_ag`.

**antibiotics\_obsolete\_ag**

- Obsolete version of antibiotics information.

**antibiotics\_route\_simple\_map\_ag**

- It classifies route of drug use into three categories: iv, oral or other.

**antibiotics\_ying\_antibacterial\_ag**

- An old version of antibacterial table. I had got this data from Ying and uploaded it.

**asv\_alpha\_diversity\_ag**

- All tables that start with `asv\_` indicates data analyzed via DADA2. This table shows the alpha diversity from each oligoid. Note: the same sample id may be sequenced more than once. Make sure to filter only one data for it (e.g. the one with highest coverage). sampleid is the id for the sample collection. Oligoid is the ide for sequencing. The castori center standard is using `sampleid` + `..` + `pool#`.  
 NOTE: I uploaded a field called path\_pool. It tells the path of where I ran the pipeline. In order to avoid mixing data, I lock from table `asv\_counts\_ag` any new upload from a sampleid that is already in the table. For `asv\_alpha\_diversty\_ag`, I accept a new upload if it comes from a different path. The reason is that I can track independent sequencing that would have exactly the same sampleid. One solution, in case of Castori samples is to add `..pool####` on the oligos files. This has yet to be implement.

**asv\_annotation\_blast\_ag**

- All tables that start with `asv\_` indicates data analyzed via DADA2. This table shows annotation based on blast. It only output one hit for ASV in case of multiple hits. Check `asv\_annotation\_blast\_detailed\_ag` for more information.

`asv\_key` links to `asv\_key` in other `asv\_\*` tables or `key` in asv\_sequences\_ag.

Notes: When you observe NA in at a taxonomy level, look at higher level. E.g. ASV 636 family is classified as NA, and order is Clostridiales. Consider it equivalent to `unclassified Clostridiales`.

**asv\_annotation\_blast\_color\_ag**

- All tables that start with `asv\_` indicates data analyzed via DADA2. This table has a color map for ASV keys. Each genus has one unique color that is spreading around a `base` color for the group. It facilitates making plots. This table is useful, but it is not dynamically updated yet.

`asv\_key` links to `asv\_key` in other `asv\_\*` tables or `key` in asv\_sequences\_ag.

**asv\_annotation\_blast\_detailed\_ag**

- All tables that start with `asv\_` indicates data analyzed via DADA2. It shows detailed blast output for annotation. Some ASVs had more than one equally good annotation. In case more than one species match was returned, the field `name` will contain them separated by ";".

`asv\_key` links to `asv\_key` in other `asv\_\*` tables or `key` in asv\_sequences\_ag.

**asv\_annotation\_gtdb\_ag**

- ASV annotation based on GTDB database.

`asv\_key` links to `asv\_key` in other `asv\_\*` tables or `key` in asv\_sequences\_ag.

**asv\_annotation\_silva\_ag**

- ASV annotation based on Silva database.

`asv\_key` links to `asv\_key` in other `asv\_\*` tables or `key` in asv\_sequences\_ag.

**asv\_counts\_ag**

- All tables that start with `asv\_` indicates data analyzed via DADA2. This table shows counts for each ASV per sampleid. Note: the same sample id may be sequenced more than once. Make sure to filter only one data for it (e.g. the one with highest coverage). Sampleid is the id for the sample collection. Oligoid is the id for sequencing. The castori center standard is using `sampleid` + `..` + `pool#`.

`asv\_key` links to `asv\_key` in other `asv\_\*` tables or `key` in asv\_sequences\_ag.

Note: Same sample may be sequenced more than once! Be careful to select for only one sampleid. Use this to select the sampleid for the run with more coverage depth and at least 200 counts:

asv\_counts\_ag = data.table(asv\_counts\_ag );

oligos\_id\_filtered = unique(asv\_counts\_ag[count\_total>200][order(sampleid,-count\_total)][!duplicated(sampleid)]$oligos\_id);

asv\_counts\_ag = asv\_counts\_ag[count\_total>200][order(sampleid,-count\_total)][oligos\_id %in% oligos\_id\_filtered];

asv\_counts\_ag[, count\_relative := count/count\_total];

Version (b1, b2, …): I am adding b# to indicate different batches of demultiplex + DADA2 + annotation. I use it to keep track of variability in the data.

b1: demultiplex with 0 primer mismatch + capping 100k/reads + annotation not properly prioritizing taxid.  
b1p1: demultiplex with 0 primer mismatch + capping 100k/reads + annotation priority for smallest taxid.  
b2: demultiplex with 1 primer mismatch + capping 100k/reads + annotation priority for smallest taxid.

**asv\_dominant\_ag**

- All tables that start with `asv\_` indicates data analyzed via DADA2. This table provides the most common ASV and most common genus (according to BLAST annotation) per run (oligo\_id). It also provides the genus for most common ASV. In some cases, the genera of the most common ASV is different than the most common genus.

**asv\_sequences\_ag**

- All tables that start with `asv\_` indicates data analyzed via DADA2. This table shows the raw sequence for each asv. The key in this table is used as reference and represent `asv\_key` in other `asv\_\*` tables.

**comorbidity\_index\_ag**

- The commorbidity index of patients from all four institutions studied in the multicenter study (Peled et al., NEJM in press).

**counts\_ag** (Obsolete: This version uses OTU based pipeline. Use asv\_counts\_ag for DADA2 based pipeline).

- It has the counts for each sampleid and OTU from the OTU based pipeline.

**counts\_replicates\_ag** (Obsolete: This version uses OTU based pipeline. Use asv\_counts\_ag for DADA2 based pipeline)

- It is similar to counts\_ag, but simplified and it also contains pool\_path. It was created to be able to compare technical replicates in the data.

**drugs\_not\_antibacterial\_ag**

- I created this table based on Eric Litmann's file data (`tasks1\_table\_Jan2019.RDS`). Tsoni curated a list of non-antibiotic drug on this (`msk.non.antibiotic.drugs.csv`). It shows usage of non-antibiotic drugs.

**enterotypes\_ag** (Obsolete: This version uses OTU based pipeline. Use it until I create an `asv\_\*` version of it)

- The enterotype classification of each sampleid.

**frozen\_set\_ag**

- The frozen set of patients and sampleids used in the multicenter study (Peled et al., NEJM)

**frozen\_set\_cgvhd\_ag**

- The frozen set of patients and sampleids used in the cGVHD study (Markey et al., under revision).

**frozen\_set\_christoph\_ag**

- The frozen set of patients and sampleids used in the Enterococcys study (Stein-Thoeringer et al., Science 2019). This frozen\_set was created when a few analysis had already been done for other samples. TO DO: create an expanded version of it with the lab of the samples used in each figure.

**frozen\_set\_multivariate\_clean\_obsolete\_ag;**

- This table is an exact copy of multivariate\_clean\_frozen\_set\_ag. The naming `frozen\_set\_\*` was not be the best naming because it suggests this to be a frozen set type of table.

**frozen\_set\_nofmt\_ag**

- I excluded any patient from frozen\_set that had been unber treatment arm in FMT investigation.

**frozen\_set\_obsolete\_may172019\_dukeupdatedays\_ag**

- Obsolete version of frozen\_set\_ag. We obtained more samples from Duke that had to be added to it.

**frozen\_set\_obsolete\_sep122018\_ag**

- Obsolete version of frozen\_set\_ag.

**healthy\_volunteers\_ag**

- List of sampleids from healthy volunteers.

**hospitalization\_ag**

- It shows admission and discharge dates for MSK patients.

**hospitalization\_obsolete\_in\_may092019\_ag**

- Obsolete table for hospitalization

**inpatient\_classification\_ag**

- It classifies samples as inpatient/outpatient according to data of collection and `hospitalization\_ag`.

**jennifer\_blood\_test\_ag**

- I uploaded this table for Jennifer.

**jennifer\_demographics\_ag**

- I uploaded this table for Jennifer.

**jennifer\_drug\_exposure\_ag**

- I uploaded this table for Jennifer.

**multivariate\_clean\_frozen\_set\_ag**

- This table has the multivariate parameters for each institution in the multicenter study (Peled et al., NEJM). It has source type, conditioning, gender and age per patient.

**nutrition\_ag**

- It contains nutrition information of patients (at the time it was created, it was a renamed from nutrition\_feb2019\_ag, which was the most updated table of nutrition demographics).

**nutrition\_obsolete\_Jul2017\_ag**

- It contains nutrition (obsolete)

**nutrition\_obsolete\_Jan2018\_ag**

- It contains nutrition (obsolete)

**nutrition\_obsolete\_jun2018\_ag**

- It contains nutrition (obsolete)

**nutrition\_demographics\_ag**

- It contains nutrition information of patients (at the time it was created, it was a renamed from nutrition\_demographics\_feb2019\_ag, which was the most updated table of nutrition demographics).

**nutrition\_demographics\_ag\_obsolete\_jan2018\_ag**

- It contains nutrition patients demographic information (obsolete)

**nutrition\_demographics\_ag\_obsolete\_Jul2017\_ag**

- It contains nutrition patients demographic information (obsolete)

**otu\_ag** (Obsolete: This version uses OTU based pipeline. Look at asv\_\* for DADA2 based pipeline).

It contains OTU annotation according to modified green genes.

**otu\_blast\_ag** (Obsolete: This version uses OTU based pipeline. Look at asv\_\* for DADA2 based pipeline).

It contains OTU annotation according to blast.

**otu\_blast\_not\_passed\_ag** (Obsolete: This version uses OTU based pipeline. Look at asv\_\* for DADA2 based pipeline).

It details blast OTU annotation.

**otu\_colormap\_genus\_shade\_ag** (Obsolete: This version uses OTU based pipeline. Look at asv\_\* for DADA2 based pipeline).

It show OTU colors. One color per genus.

**otu\_colormap\_yt\_pallete2\_ag**  (Obsolete: This version uses OTU based pipeline. Look at asv\_\* for DADA2 based pipeline).

It show OTU colors based on Ying's pallete.

**patient\_allo\_ag**

List of clinical information for MSKCC allo patients

**patient\_allo\_excluded\_ag**

One patient was left the protocol. This table was created to keep track of this.

**patient\_allo\_obsolete\_jul2017\_a**g

Obsolete allo patients table

**patient\_auto\_ag**

List of clinical information for MSKCC auto patients

**patient\_auto\_jul2017\_ag**

Obsolete auto patients table

**patient\_duke\_ag**

List of clinical information for Duke patients. This version is more detailed but does not have all patients. Look at patient\_duke\_simplified\_ag to a list with more patients but that only bring HCT, last\_contact and vital\_status information.

**patient\_duke\_engraftment\_day\_ag**

It has the engraftment date for Duke patients.

**patient\_duke\_simplified\_ag**

It is a simplified version of patient\_duke\_ag. It contains only fields related with death/censor. It was created for multicenter analysis of OS because we had observed trends with statistical significance and needed to increase our patient list to get this info.

**patient\_hokkaido\_ag**

List of clinical information for Hokkaido patients.

**patient\_regensburg\_ag**

List of clinical information for Regensburg patients

**patient\_regensburg\_supplement\_ag**

It has the engraftment date for Regensburg patients.

**patient\_standardized\_competing\_risks\_ag**

Organize standardized table of competing risk for four institutions. It has relapse, GVHD, GVHD\_grade, GVHD\_onset, last contact and vital status.

**patients\_cgvhd\_ag**

Cgvhd patients from Kate.

**patients\_msk\_comorbidity\_index\_ag**

Comorbidity index from MSK

**samples\_castori\_ag**

A postgres version of tblSamples from access database.

**samples\_dominant\_otu\_ag**

It describes the dominant OTU for each sampleid. OTU method is obsolete!

**samples\_duke\_ag**

Info for Duke samples. Check table samples\_duke\_clean\_ag instead.

**samples\_duke\_clean\_ag**

Info for Duke samples: sampleid, pid, timepoint, collection\_date

**samples\_duke\_clean\_obsoleteapr042019\_ag**

Info for Duke samples (obsolete)

**samples\_hokkaido\_ag**

Info for Hokkaido samples

**samples\_regensburg\_ag**

Info for Regensburg samples

**shotgun\_sample\_lookup\_ag (obsolete)**

A version of E. Littmann shotgun\_sample\_lookup. It tracks shotgun data we have sequenced. This table is obsolete. Angel tracked more samples we had sequenced. Look at shotgun\_lookup\_ad instead.

**tblactions\_ag**

A version of E. Littmann tblactions. I had to create it because I wasn't able to read that table.

It has some "action" information about samples. E.g. Test for resistance, 16S qPCR and other.

This is the list of `action` values: % Cdiff 16S qPCR Cdiff 16s Cdiff 16s/gm Cdiff 16s/uL TcdB U16s/gm U16s/uL VanA

**tsne\_ag**

tsne coordinates used in multicenter study.

**tsne\_frozen\_obsolete\_sep122018\_ag**

An obsolete version of tsne for frozen set.

**vre\_status\_ag**

VRE status in a sample.