Datasets with time points of compositional data

For all datasets you have a corresponding paper about it. Please note that features/species may be just “counts” or “percentage”, so to make data compositional, you need to normalize it.

BASIC: postpartum depression

The only dataset you can not freely share with anyone. The data for depression isn't fully open yet, but you can refer to the Axfors cohort description: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6830667/ How to read the files:

For the metadata table. There are 6 columns in total, including:

- **Individual_ID**. The IDs for the BASIC individuals.
- **TimePoint**. Sampling time point for the individuals. Each individuals has 3 timepoints, namely Trimester2 (corresponding to week 20), Trimester3 (corresponding to week 30) as well as PostpartumWeek6 (corresponding to postpartum).
- **EPDS**. The score of EPDS, ranging from 0 to 30.
- **Dichotomous_EPDS**. Dichotomous EPDS scores based on the threshold of 12. (Individuals with EPDS < 12 will be regarded as Healthy and 0 will be assigned in this column, individuals with 12 <= EPDS < 30 will be regarded as Depressed and 1 will be assigned)
- **Sample_ID**. The ID of sequenced fecal samples, can be used to match with the samples in Species_Profile.csv file. NA means no fecal sample is available for the individauls at specific time point.
- **ReadsNumber**. Empirically, if a fecal microbiome sample contains less than 500k reads then we will regard this sample as failed and will not include it in downstream analysis. If a fecal microbiome sample contains more than 500k reads but less than 2 million reads we will consider it’s likely that all species in the sample were covered/sequenced. Samples containing more than than 2 million reads are rather reliable. ***Therefore the easiest thing that you could do regarding reads number is to exclude samples with less than 500k reads based on this column***.

For the species profile table:

- Rows are samples, can be matched with the **Sample_ID** column in the metadata table.
- Columns are features (or species).

GMAP: food allergies

similar structure, used in my paper as well, attached to the email, has described here https://link.springer.com/content/pdf/10.1186/s40168-022-01322-y.pdf

At term or preterm delivery 1:

Supplementary files 5 and 6 here have patient data and vaginal microbiome data for pregnant women at several time-points in pregnancy, with the outcome being if they deliver at term or preterm. May be too little data but easier to access than the next one https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-016-0223-9#Sec16

At term or preterm delivery 2:

If you get here:
https://pretermbirthdb.org/mod/studydata
And click on the link that says SDY2187 it will take you here
https://www.immport.org/shared/study/SDY2187 It should give a good collection of vaginal swabs including if the outcome of pregnancy was preterm or full-term, from this paper: (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10829755/), which is in itself a meta-analysis of several datasets. But this download requires some plugins that are not working very well in Linux.