dbBact workshop

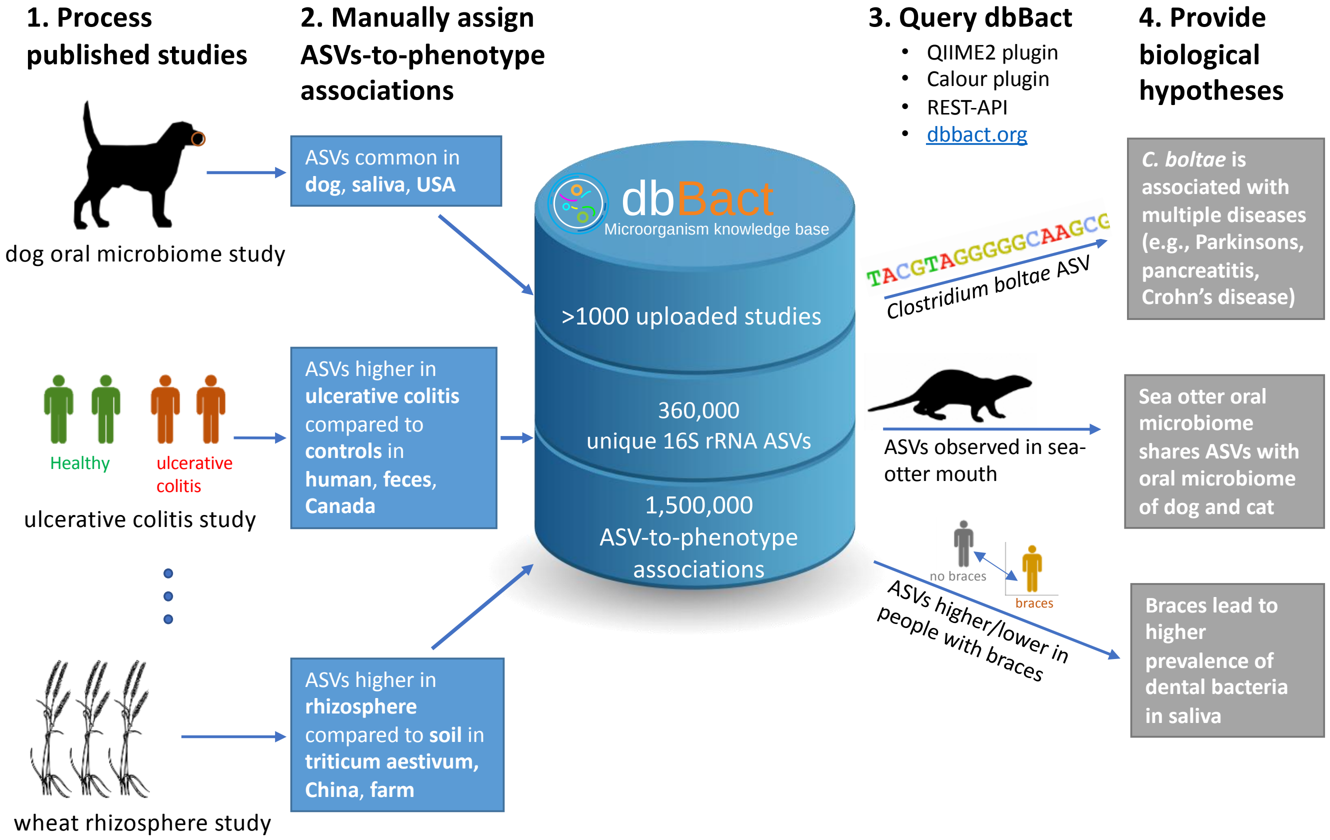
3/2023

General

- Feel free to ask questions/interrupt
- What is your background
- Outline:
    - What is dbBact - general concepts
    - Calour examples via jupyter notebook
    - Qiime2 plugin
    - Other?

What is dbBact

- Collect and use biological observations about bacteria



1. Process published studies

dog oral microbiome study

ulcerative colitis study

wheat rhizosphere study

2. Manually assign ASVs-to-phenotype associations

ASVs common in **dog, saliva, USA**

ASVs higher in **ulcerative colitis** compared to **controls** in **human, feces, Canada**

ASVs higher in **rhizosphere** compared to **soil** in **triticum aestivum, China, farm**

>1000 uploaded studies

360,000 unique 16S rRNA ASVs

1,500,000 ASV-to-phenotype associations

3. Query dbBact

- QIIME2 plugin
- Calour plugin
- REST-API
- dbbact.org

Clostridium boltae ASV

ASVs observed in sea-otter mouth

ASVs higher/lower in people with braces

4. Provide biological hypotheses

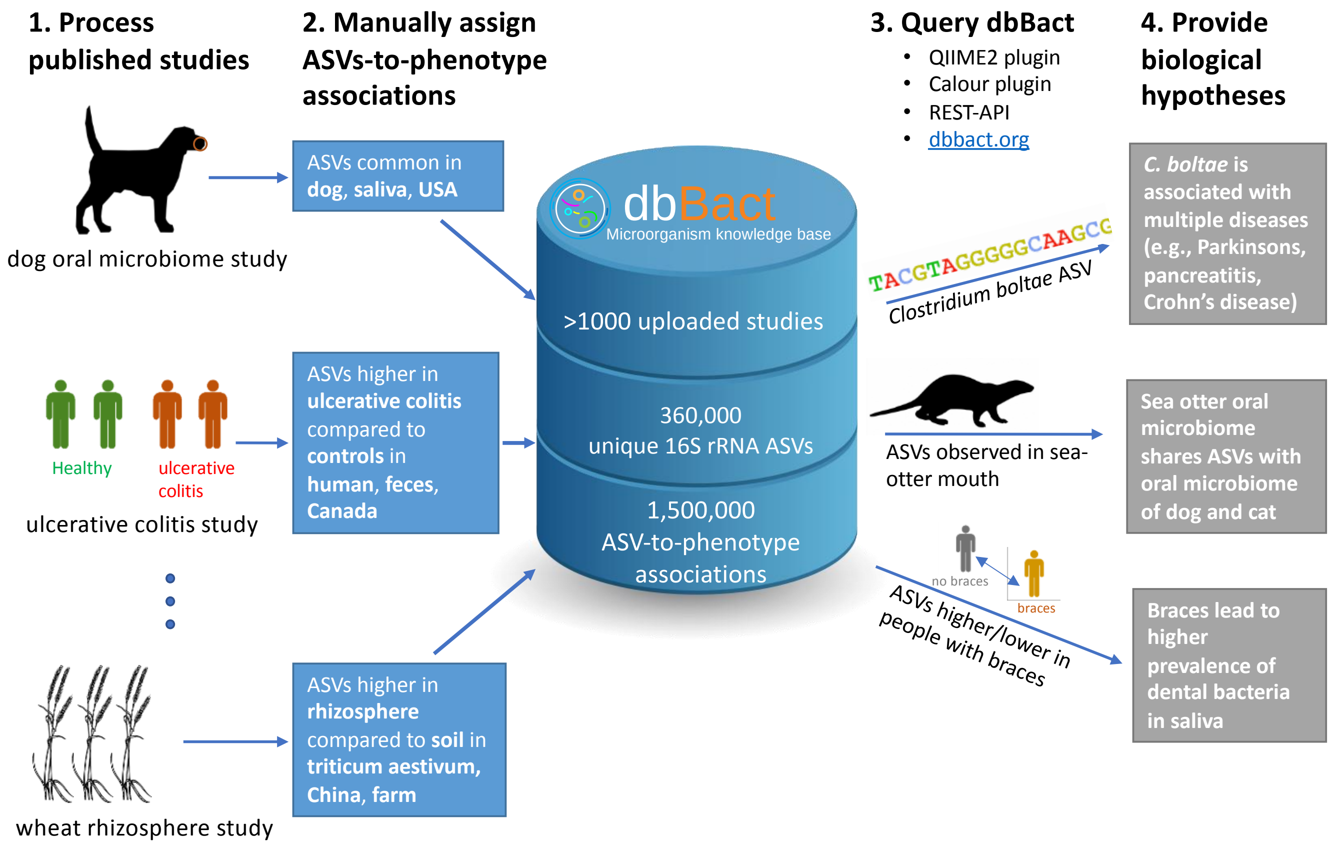
C. boltae **is associated with multiple diseases (e.g., Parkinsons, pancreatitis, Crohn's disease)**

Sea otter oral microbiome shares ASVs with oral microbiome of dog and cat

Braces lead to higher prevalence of dental bacteria in saliva

What is dbBact

- Collect and use biological observations about bacteria



1. Process published studies

2. Manually assign ASVs-to-phenotype associations

3. Query dbBact

- QIIME2 plugin
- Calour plugin
- REST-API
- dbbact.org

4. Provide biological hypotheses

dog oral microbiome study

ASVs common in **dog, saliva, USA**

ulcerative colitis study

ASVs higher in **ulcerative colitis** compared to **controls** in **human, feces, Canada**

wheat rhizosphere study

ASVs higher in **rhizosphere** compared to **soil** in **triticum aestivum, China, farm**

dbBact Microorganism knowledge base

>1000 uploaded studies

360,000 unique 16S rRNA ASVs

1,500,000 ASV-to-phenotype associations

Clostridium boltae ASV

C. boltae **is associated with multiple diseases (e.g., Parkinsons, pancreatitis, Crohn's disease)**

ASVs observed in sea-otter mouth

Sea otter oral microbiome shares ASVs with oral microbiome of dog and cat

ASVs higher/lower in people with braces

Braces lead to higher prevalence of dental bacteria in saliva

What is “bacteria”

- We use ASV (Amplicon Sequence Variants) rather than taxonomy as a unique
- Advantages:
    - The full information from each experiment (unlike taxonomy which is limited)
    - Objective (rather than taxonomy that is subjective and subject to change)
- Disadvantages:
    - How to link different experiments/regions/read lengths etc.

What is “bacteria” - problems with taxonomy

- We don’t have enough names for all the bacteria out there (only ~30,000 names compared to ~300,000 ASVs in EMP)
- Taxonomy is subjective (should we use SILVA or GreenGenes? Versions change) and arbitrary (what is the difference between genus and species? Different in different clades).
- The fact that we have only genus level identification does not mean this is the maximal information from the ASV
  - For example, our ASV can be 100% match to 2 species in a given genus, so taxonomy will give us only genus level, but there are hundreds of more species in the same genus that the ASV does not match.

What is “bacteria” - harmonizing reads

- There is a set of commonly used sequencing regions (V12, V34, V45, ?)
- We use a constant starting point for each region (based on the primer), with programs to automatically trim the primers, so each read starts at one of the 3 options (V1f, V3f, V4f)
- Read length varies between experiments. We store the full sequence available for each experiment, and harmonize by looking at the minimal length between the database and query sequence.
- Inferring between regions (i.e. two sequences belong to the same bacteria over different regions) is based on an external database of full length 16S sequences (SILVA).
    - This is done at query time (so the database stores only to original ASVs).
    - We can choose if we want to get information also for the linked sequences.
    - It is subjective (depends on the database used), can be noisy (i.e. if there are chimeras in the database) and incomplete (we don’t have 16S sequences of all bacteria).
    - It is not 1-1 (so even if the database is complete, we can get wrong answers)
    - BUT: works nicely (we’ll see example later)

What is “biological observations”

- dbBact annotations are based on re-analysis of raw reads from each experiment
- Current annotation types are:
    - COMMON - bacteria present in >0.5 of samples of a given type in the experiment
    - DOMINANT - bacteria with mean frequency > 1% in samples of a given type in the experiment
    - HIGHER IN XX compared to YY - bacteria showing a difference in distribution between two groups in the experiment
    - CONTAMINANT - bacteria suspected as contaminants in the experiment (i.e. present in blanks and not well-to-well spillage)
    - OTHER - for example pathogen, etc.

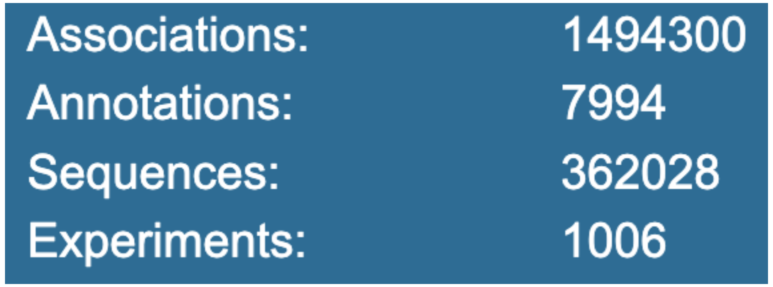
What is “biological observations”

- The terms describing the samples are derived from multiple ontologies integrated into dbBact (ENVO, GAZ, UBERON, DOID, EFO, NCBITAX, PATO, TO, HSAPDV), and an additional dbBact ontology for all terms not present in this ontology. Possible to add additional ontologies if needed.
- Using ontologies enables utilizing the tree structures in them (i.e. Crohns disease and ulcerative colitis are both inflammatory bowel diseases etc.)
- Presents interesting statistical questions.

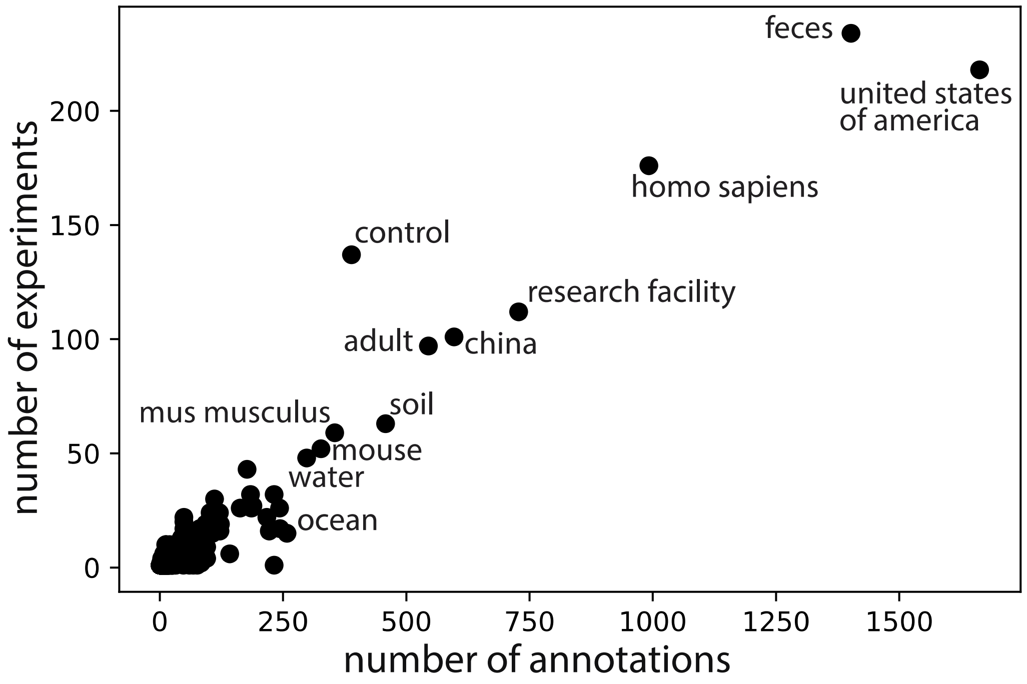
dbBact terminology

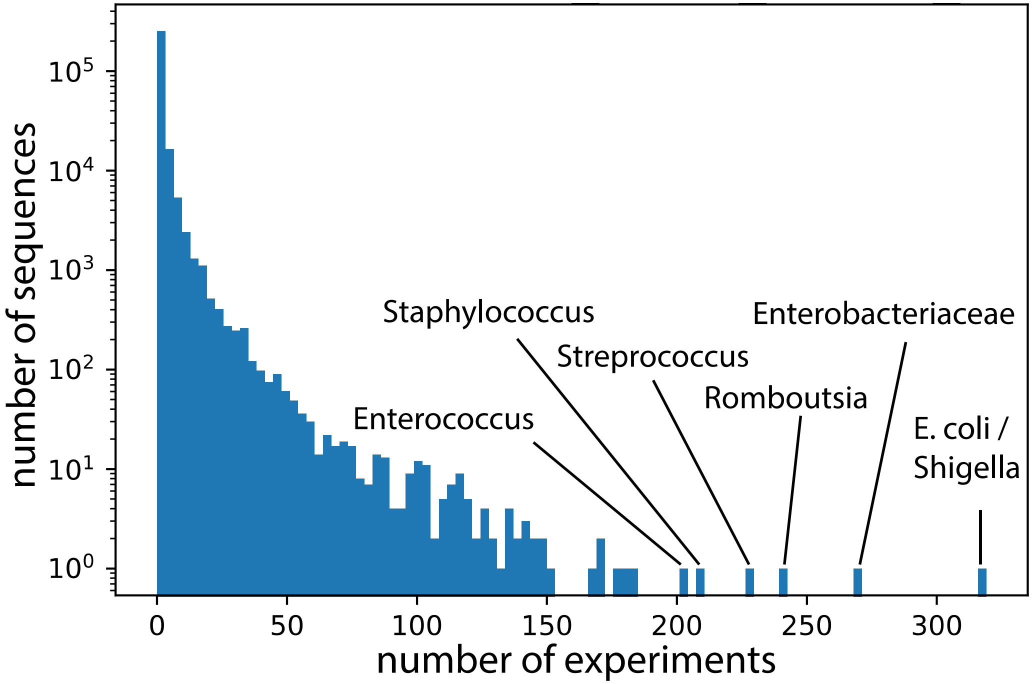
- *Experiment* - a 16S rRNA ASV dataset from a single study.
- *Sequence* - an ASV for a bacteria.
- *Term* - an ontology derived word describing a set of samples (i.e. feces/ulcerative colitis/paris etc.).
- *Annotation* - a biological observation derived from an *experiment*, associating a set of *sequences* with a set of *terms* describing these ASVs (i.e. “HIGHER in ulcerative coliis COMPARED to controls IN feces, homo sapiens, paris, adult).

Current dbBact stats



Associations:	1494300
Annotations:	7994
Sequences:	362028
Experiments:	1006





Adding data to dbBact

- Currently almost all experiments were added by the dbBact team.
- It is wiki-like, so anyone can add experiments and annotations.
    - Users can flag suspicious experiments/annotations.
    - The truth will float
- Adding new experiments/annotations is currently supported using the Calour dbBact plugin
- We will not focus on adding new data in this workshop, but will be happy to help as needed

Let’s start analyzing!

- Several ways to use dbBact for analysis (from simplest to more complex):
  - dbBact website (www.dbbact.org)
  - EZCalour (full GUI version of Calour)
  - Qiime2 plugin
  - Calour plugin (python)
  - dbBact REST-API server (api.dbbact.org)
    - Documentation at http://api.dbbact.org/docs
- We will mostly go over the dbBact paper examples
  - Using Jupyter notebooks (python) with Calour and dbBact plugin

Let’s start analyzing!

- More useful links:
  - Calour: https://github.com/biocore/calour
  - dbBact-calour plugin: https://github.com/amnona/dbbact-calour
  - Qiime2 dbBact plugin: https://library.qiime2.org/plugins/q2-dbbact/36/
  - dbBact preprint: https://www.biorxiv.org/content/10.1101/2022.02.27.482174v2.abstract
  - Example notebooks: https://github.com/amnona/dbbact-paper

We want feedback!

- Questions? Bugs? Feature suggestions?
- Contact:
  amnonim@gmail.com
- OR add an issue in the dbbact-server github:
  https://github.com/amnona/dbbact-server/issues

THANKS!!!