The data for the Module 2 Capstone:

For the Module 2 Capstone you will execute the Mothur workflow on data that has been collected from the Richmond, IN area by Earlham and IUE faculty and students. You will also analyze these data to *test a hypothesis*. These data are all part of the same project, but are two independent experimental designs. The 2012 and 2013 data are part of a single design, and the 2014 data are different and represent a 'next step' from the 12/13 data. The following is an introduction to the data so that you can identify an appropriate question to test. Note, there are many possible good questions/hypotheses - those that are good will be easy to discuss and explain to your parents. Luckily, this project is connected to both agriculture, the economy, and conservation, so there are many angles that you can use to give context to your question.

We will design our study and execute the workflow before we start discussing data analysis...depending on your data, it may take awhile to get the workflow going - and even longer for the cluster to do your dirty work.

Due Wednesday in class (in hard copy):

1. Group name(s) (you can work in pairs or singly) - we will do some 'ideation' in class Tuesday. I encourage you to collaborate with people who you have not yet partnered with.
2. Your question and hypothesis (including a rationale for your question...why is it an interesting question to ask?)
3. The data you will use to test your hypothesis/prediction
4. The '.files' file that you will give to Mothur to run your data.

Preview: your capstone, when officially introduced, will be a 'mini research paper'. That is, think of a full scientific research paper, with all the requisite sections (intro, methods, results, discussion, references), but in brief. The focus will be on getting the workflow through and analyzing data - and putting everything in context (big picture).

Experiment 1:

**Context:** In 2012 and 2013 a group of Earlham and IUE students and faculty ran an experiment assessing how bacterial communities varied across agricultural fields differing in the crop that was planted (corn or soybean) and the soil management regime (conventional or no-till tillage). These data are fully described in [Smith et al. (2016)](https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.2553). The take-home result from that study was that soil management seemed to have a dramatic effect on community composition.

**Experimental design:** In each year we sampled 20 local agricultural fields. In 2012 half of these were planted in corn and the other half in soy; similarly, half were traditionally tilled (the soil was turned over) and the other half were not tilled or minimally tilled (conservation tilling). Note, there are possible benefits to both tilling and not, but there are also conservation implications (tilling leads to increased erosion and nutrient/soil loss). In 2013 we sampled the same fields (though two were planted in wheat and were not sampled); the majority of fields were rotated to the alternate crop (though one was planted in corn again). We had no control over what the farmers planted and so the design is not perfectly balanced between years.

**Associated Files and Naming Convention:**

All of the fields have a numeric assignment (1-20) that is carried over between regardless of what was planted. In the descriptive file below you can see that there are separate columns for "crop" and "till" as well as the name of the family farmers. We also did chemical analyses on the soil; if you don't know what some of the analyses mean, you can look them up or ask to be sure. Most units are listed as either ppm (parts per million) or pph (parts per hundred - i.e., percent).

* The location of each sample is viewable in GoogleEarth with this file: 2012fielddata\_photosge.kmz
* The basic field data and chemical data are in: 20122013BasicFieldData.xlsx

At the following URL you will find all of the fastq files associated with this project.

<http://cluster.earlham.edu/metagenomes/FieldTwo/>

They are named as follows:

* [130819\_M01529\_0018\_2012\_1.R1.fastq](http://cluster.earlham.edu/metagenomes/FieldTwo/130819_M01529_0018_2012_1.R1.fastq)

In this name the relevant part is 2012\_1, which says that this is field 1 from 2012 (for simplicity, this is the R1 file, there is an similarly named R2 file. If you were to want the same field from 2013, it is: [130819\_M01529\_0018\_2013\_1.R1.fastq](http://cluster.earlham.edu/metagenomes/FieldTwo/130819_M01529_0018_2013_1.R1.fastq)

Experiment 2:

**Context:** In 2014 a group of Earlham students and faculty (Chris Smith, Brent Kramer, and Emily Sells) conducted a follow-up experiment. The goal of this experiment was to examine the direct effect of soil tillage on microbial communities with different histories of soil disturbance.

**Experimental design:** We selected two sites with very different histories of disturbance. The first was the unplanted margin of a heavily tilled agricultural field owned by Earlham College and leased to a local farmer. The second was an adjacent mature forest, also owned by Earlham. Ten 1m2 plots were chosen at each site and these were randomly assigned to one of two treatments, tillage or not. The tilled plots were tilled by hand with a roto-tiller to a depth of 10cm. Soil from all plots was sampled using a soil corer to a depth of 10cm (i.e., the core contained soil from 0-10cm, but the top litter was excluded). Soil samples were taken at three times.

* Pre-treatment
* 1-day Post treatment
* 10-days Post treatment

Soil from the Pre-treatment sample was divided and half was sent for chemical analysis at the A&L Great Lakes Lab, a commercial lab for agricultural soil analysis. DNA was isolated from the soil of all sites and all sampling times - a total of 2 sites \* 10 plots \* 3 samples = 60 samples. 16s rRNA amplicons were amplified using the 515F and 806R primer set for bacteria where each of the 60 samples had a unique barcode. Amplicon libraries were combined at equimolar ratios and the combined set of samples was sequenced at the Center for Genomics and Bioinformatics at IU-Bloomington on the Illumina MiSeq platform using 150 bp paired-end sequencing.

**Associated Files and Naming Convention:**

The sites are named according to their site and number. Sites labeled “C” are those in the crop site, while those labeled “F” are from the forest site. The number is their plot number, which is then associated with the treatment (till or not) and the sample time (pre, post1 or post2). The files below have relevant information, site location, soil chemistry (from the pre sample), and soil moisture at each sample event.

* The location of each sample is viewable in GoogleEarth: 2014SampleLatLong.kml
* The soil chemistry: SoilChemData2014\_reduced.txt
* The soil moisture data: SoilMoisture2014.txt

All of the sequences from 2014 can be found at:

<http://cluster.earlham.edu/metagenomes/FieldThree/>

The raw sequencer output files have the relevant sample labels within them. These raw Illumina output files are in FASTQ format. Below is one example file name for reference.

* GSF723-**Smith-Post-C-1**\_S21\_L001.fastq