**Methods for Food Tree from FL100 ASA24 Recalls**

**Elizabeth Chin**

**February 18, 2020**

The food tree was inspired by Johnson et al.1 The tree was generated using GraPhlAn2 using python 2.7. Data preparation was conducted using R v3.6.0. The code and data (excluding subject recall files) are available at <https://github.com/ebeth-chin/Food_Tree>.

Though the food tree is hierarchical and foods with similar characteristics are grouped together, the distances between all levels is the same for each food as the taxonomy is built from the FoodCode, which does not contain any quantitative information about food relatedness.

**1. Dietary Data Processing**

*1.1 Formatting Foods*

Code: [00\_format\_foods](https://github.com/ebeth-chin/Food_Tree/blob/master/src/00_format_foods.R) & [01\_find\_overlap](https://github.com/ebeth-chin/Food_Tree/blob/master/src/01_find_overlap.Rmd)

Dietary intake was collected from healthy adults in the Nutritional Phenotyping Study3 using the Automated Self-Administered 24-hour Dietary Assessment Tool (ASA24).4 Participants completed a total of four recalls: one training recall in the presence of study personnel, and three at-home recalls. Quality control (QC) of dietary data was performed by a registered dietitian as described in Bouzid et al. A total of n = 75 participants used ASA24-2014 and n = 319 participants used ASA24-2016, which use USDA Food and Nutrient Database for Dietary Studies (FNDDS) versions 4.1 and 2011-2012, respectively. Each food output by the ASA24 system has a FoodCode that is derived from FNDDS. A grand total of n = 2431 unique FoodCodes were reported by study participants.

*1.2 Formatting FoodCodes*

Code: [02a\_foodcode\_curation](https://github.com/ebeth-chin/Food_Tree/blob/master/src/02a_foodcode_curation.Rmd) & [02b\_cleanup\_foodcodes](https://github.com/ebeth-chin/Food_Tree/blob/master/src/02b_cleanup_foodcodes.R)

Some FoodCodes were identical for some foods with differing details, regardless of the ASA24 version. In most cases, the details described differences in oil/fat additions; for these types of differences, the only one version of the food was retained on the tree. For example, FNDDS FoodCode: 75226060 was used for the ASA24 output foods “Peppers, red, cooked, fat added in cooking” and “Peppers, red, cooked, fat added in cooking W/ VEGETABLE OIL NFS.” Note that this is different from Johnson et al., where each version of a food is retained as a separate leaf (i.e. those two versions of “Peppers” would be considered separate foods).

For two foods, the details were additions of different food products and these foods were kept as separate leaves on the tree. These foods were:

1. FoodCode: 92101900, used for the ASA24 output foods, “Coffee, Latte W/ WHOLE MILK” and “Coffee, Latte W/ SOY MILK”
2. FoodCode: 32105010, used for the ASA24 output foods, “Egg omelet or scrambled egg with cheese W/ VEGETABLE OIL” and “Egg omelet or scrambled egg with ham or bacon and cheese”

Data Files: [duplicated\_foods\_qc.xlsx](https://github.com/ebeth-chin/Food_Tree/blob/master/data/01_find_overlap/duplicated_foods_qc.xlsx) & [foodcodes\_duplicated\_with\_jitter\_formatted.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/02_foodcode_curation/foodcodes_duplicated_with_jitter_formatted.txt)

For one of the duplicated foods in each set, .1 was added to the end of the FoodCode to allow the food to be identified as a unique food on the tree while also retaining the correct taxonomy (explained in section 2.1 below). Therefore, the FoodCode for “Coffee, Late W/ SOY MILK” was converted to 92101900.1.

Some foods reported in ASA24-2014 had the same description or highly similar description as a food reported in ASA24-2016, but with a different FoodCode. For these cases, the ASA24-2016 food description and FoodCode were used for the ASA24-2014 food. This prevented the same food as appearing as separate leaves (i.e. appearing twice) on the food tree. A total of 28 ASA24-2014 FoodCodes were collapsed into ASA24-2016 FoodCodes.

Data files: [foodcodes2014\_convert\_to\_16\_input.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/02_foodcode_curation/foodcodes2014_convert_to_16_input.txt)

Two foods reported in ASA24-2014 had FoodCodes that would result in incorrect taxonomy given their FoodCodes. We assigned these foods the same FoodCode as a similar food reported in ASA24-2016 and added .1 to the end. The foods were as follows:

1. 2014 FoodCode: 41430310, Protein diet powder with soy and casein. New FoodCode: 95230030.1, derived from 2016 FoodCode 95230030, Protein powder, NFS
2. 2014 FoodCode: 41435110, High protein bar, candy-like, soy and milk base. New FoodCode: 53729000.1, derived from 2016 FoodCode 53729000, Nutrition bar or meal replacement bar, NFS

Data files: [foodcodes2014\_convert\_to\_16\_input.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/02_foodcode_curation/foodcodes2014_convert_to_16_input.txt)

During recall QC, there were some instances where the food “Milk, almond, ready-to-drink” was used to replace “Milk, NFS” (likely due to subjects being unable to find “Almond Milk” in the ASA24 system and writing it in), but the FoodCode for “Milk, almond, ready-to-drink” (FoodCode: 11350000) was not updated during this QC process, leaving the FoodCode for “Milk, NFS” (FoodCode 11100000) in place. The correct FoodCode for “Milk, Almond” was assigned during data preparation for the food tree.

Data files: [foodcodes\_duplicated\_with\_jitter\_formatted.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/02_foodcode_curation/foodcodes_duplicated_with_jitter_formatted.txt)

**2. Creation of the Food Tree for *All Foods* Reported by FL100 Subjects**

Note that two Food Trees were created. The first tree (this section) uses all unique QC’d FoodCodes from all ASA24 recalls, regardless of whether the recalls passed QC and were used for the eventual OTU table. This tree was initially created in 02/2020.

A second tree (section 5) was created using only the foods reported in recalls retained after QC (section 3) and was created in 04/2020. The second tree was necessary because QIIME2 does not allow the number of leaves on the tree to differ from the number in the OTU file. A full tree is being retained for supplemental purposes.

*2.1 Creating Food Taxonomy Levels*

Code: [03\_make\_newick\_tree](https://github.com/ebeth-chin/Food_Tree/blob/master/src/03_make_newick_tree.Rmd)\_all\_foods

The list of FoodCodes and Food Descriptions from all FL100 recalls (i.e. even from subjects/recalls that were eventually not used in the OTU table) were used as input to create the newick tree and accompanying taxonomy file. A newick tree was generated using custom R scripts (Johnson et al.). The Taxonomy was assigned using the taxonomy levels created by Johnson et al.1 The taxonomy levels are determined by the FNDDS FoodCode: each FoodCode is an eight-digit number, with the first two numbers (levels 1 and 2) corresponding to food groups assigned by the USDA.5 Levels 3-5 were manually curated by Johnson et al.,1 using the FoodCode numbers as a guide for grouping foods together. Level 6 is the ASA24 food description.

Data file: [NodeLabelsMCT.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/03_make_newick_tree_all_foods/NodeLabelsMCT.txt)

Newick Tree: [fl100\_newick\_tree\_all\_foods.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/03_make_newick_tree_all_foods/fl100_newick_tree_all_foods.txt)

Taxonomy: [fl100\_newick\_taxonomy\_all\_foods.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/03_make_newick_tree_all_foods/fl100_newick_taxonomy_all_foods.txt)

An example of the FoodCode Levels, corresponding to the number sequence in the FoodCode, and the corresponding taxonomy is below:

|  |  |
| --- | --- |
| **FoodCode Level** | **Taxonomy** |
| 1 | Milk\_and\_Milk\_Products |
| 11 | Milks\_and\_milk\_drinks |
| 110 | Milk\_human |
| 111 | Milk\_fluid |
| 1111 | Milk\_fluid\_cow\_goat\_buttermilk |

Therefore, the taxonomy for FoodCode 11111000, “Milk, cow’s, fluid, whole” is:

L1\_Milk\_and\_Milk\_Products;L2\_Milks\_and\_milk\_drinks;L3\_Milk\_fluid;L4\_Milk\_fluid\_cow\_goat\_buttermilk;L5\_;Milk\_cows\_fluid\_whole

*2.2 Generating the Food Tree Figure*

Code: [make\_graphlan.sh](https://github.com/ebeth-chin/Food_Tree/blob/master/src/make_graphlan.sh)

The resulting newick tree was used as input to GraPhlAn. The annotation files used for color-coding the food groups was sourced from the GitHub of the first author of the Johnson et al. manuscript: <https://github.com/abbycole/food_graphlan/tree/master/data>.

Plot files: [fl100\_food\_tree\_all\_foods.png](https://github.com/ebeth-chin/Food_Tree/blob/master/plots/fl100_food_tree_all_foods.png) & [fl100\_food\_tree\_all\_foods\_annot.png](https://github.com/ebeth-chin/Food_Tree/blob/master/plots/fl100_food_tree_all_foods_annot.png)

**3 Recall QC**

Code: [04a\_qc\_recalls](https://github.com/ebeth-chin/Food_Tree/blob/master/src/04a_qc_recalls.Rmd), [04b\_fix\_recall\_foodcodes](https://github.com/ebeth-chin/Food_Tree/blob/master/src/04b_fix_recall_foodcodes.R)

Recalls failing the QC process were excluded from analyses. A “fail” for the QC includes incomplete recalls and unlikely reported total intake (kcal, determined by Joanne Arsenault, RD) The in-person training recall was excluded from analyses; only the at-home recalls were included for analyses. Subjects with at least two recalls were included for analyses (n = 350).

**4 Generating Food OTU Abundance Table**

Code: [04c\_food\_otu\_abundance](https://github.com/ebeth-chin/Food_Tree/blob/master/src/04c_food_otu_abundance.Rmd)

The total amount of each food in grams (ASA24 variable: FoodAmt) was averaged over the total number of recalls per subject (two or three at home recalls and used as input to create an OTU abundance table. The total number of unique FoodCodes reported in the resulting QC’d subjects/recalls was 2125.

Data file: [fl100\_otu\_abundance.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/04_get_abundance/fl100_otu_abundance.txt)

**5. Creation of the Food Tree for *Foods in QC’d Recalls***

[Code: 05\_make\_newick\_tree\_recall\_foods.Rmd](https://github.com/ebeth-chin/Food_Tree/blob/master/src/05_make_newick_tree_recall_foods.Rmd)

A separate food tree was created using only the FoodCodes reported in the QC’d recalls (n = 2125). Taxonomy for these FoodCodes follows what is described in section 2.1. The tree figure was generated as described in section 2.2

Figure files: [fl100\_food\_tree\_from\_recalls.png](https://github.com/ebeth-chin/Food_Tree/blob/master/plots/fl100_food_tree_from_recalls.png) & [fl100\_food\_tree\_from\_recalls\_annot.png](https://github.com/ebeth-chin/Food_Tree/blob/master/plots/fl100_food_tree_from_recalls_annot.png)

Newick Tree:[fl100\_newick\_tree\_from\_recalls.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/05_make_newick_tree_foods_from_recalls/fl100_newick_tree_from_recalls.txt)

Taxonomy: [fl100\_newick\_taxonomy\_from\_recalls.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/05_make_newick_tree_foods_from_recalls/fl100_newick_taxonomy_from_recalls.txt)

**6. Newick tree w/o water**

[07\_make\_newick\_tree\_recall\_foods\_nowater.Rmd](https://github.com/ebeth-chin/Food_Tree/blob/master/src/07_make_newick_tree_recall_foods_nowater.Rmd)

A food tree was created using the FoodCodes reported in the QC’d recalls *excluding* foods in L2\_Water\_noncarbonated (FoodCodes (FoodDescription): 94000100 (water, tap), 94100100 (water, bottled, unsweetened), 94000000 (water as an ingredient)). I also noticed that the FoodCode for Meat, NFS (20000000) was getting read in scientific notation by R, so L2-5 were missing for this food.

Newick Tree: [fl100\_newick\_tree\_nowater.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/05_newicktree_nowater/fl100_newick_tree_nowater.txt)

Taxonomy: [fl100\_newick\_taxonomy\_nowater.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/05_newicktree_nowater/fl100_newick_taxonomy_nowater.txt)

Data (not on github, be sure to read the README data/05\_newicktree\_nowater): data/04\_get\_abundance/recalls\_average\_g\_nowater.txt, data/05\_newicktree\_nowater/fl100\_otu\_abundance\_nowater\_cleaned.txt

**07. Tree w/o water and removed 7 subjects**

[08\_newicktree\_w\_filteredsamples.Rmd](https://github.com/ebeth-chin/Food_Tree/blob/master/src/08_newicktree_w_filteredsamples.Rmd)

Removed 7 subjects ("5022" "6066" "6072" "6082" "6091" "7063" "8012") as per Mary’s email 06/12/2020. The tree includes foods only in the QC’d recalls, excluding water.

Newick Tree: [fl100\_newick\_tree\_filt.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/06_filter_samples/fl100_newick_tree_filt.txt)

Taxonomy: [fl100\_newick\_taxonomy\_filt.txt](https://github.com/ebeth-chin/Food_Tree/blob/master/data/06_filter_samples/fl100_newick_taxonomy_filt.txt)

Data (not on github): data/06\_filter\_samples/fl100\_otu\_abundance\_filt.txt, data/06\_filter\_samples/recalls\_average\_g\_filt.txt

**08. Fiber (g) OTU**

[09\_fiber\_otu\_abundance.Rmd](https://github.com/ebeth-chin/Food_Tree/blob/master/src/09_fiber_otu_abundance.Rmd)

Created an OTU table using average g of fiber (averaged over 2 or 3 QC’d recalls, using the ASA24 output variable FIBE) instead of the g of food (FoodAmt). The OTU table has been filtered to remove water variables and the 7 subjects listed in section 7, so it is compatible with the tree output in section 7 (n foods = 2107, n subjects = 343).

Data (not on github): data/07\_fiber\_otu/fl100\_fiber\_otu.txt, data/07\_fiber\_otu/avg\_fiber\_g.txt

**99. Remove whitespace from the outputs**

*We attempted to remove whitespace to troubleshoot issues with loading the food tree and otu tables into QIIME. It turns out that removal of whitespace was not the issue. This code and the outputs should not to be used for analyses. We have retained them in the project folder for the purpose of thorough documentation and will archive the files when the analysis has been finalized. These files have not been pushed to github:*

*src/06\_remove\_whitespace\_from\_outputs.R*

*data/04\_get\_abundance/fl100\_otu\_abundance\_nows.txt*

*data/05\_make\_newick\_tree\_foods\_from\_recalls/fl100\_newick\_tree\_from\_recalls\_nows.txt*

**References**

1. Johnson, A. J. *et al.* Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host Microbe* **25**, 789-802.e5 (2019).

2. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **2015**, (2015).

3. Baldiviez, L. M. *et al.* Design and implementation of a cross-sectional nutritional phenotyping study in healthy US adults. *BMC Nutr.* **3:79**, (2017).

4. National Cancer Institute. Automated Self-Administered 24-Hour (ASA24®) Dietary Assessment Tool. Available at: https://epi.grants.cancer.gov/asa24/. (Accessed: 21st March 2019)

5. Rhodes, D. G. *et al.* 2015-2016 Food and Nutrient Database for Dietary Studies Documentation. (2018). Available at: https://www.ars.usda.gov/ARSUserFiles/80400530/pdf/fndds/2015\_2016\_FNDDS\_Doc.pdf. (Accessed: 2nd February 2020)