BIOST/STAT 571: Final Project (Coming up with an Idea)

What is a “New” Method?

- De novo frameworks?

- Adaption of prior frameworks
  - Translation to new context
  - Extensions of existing frameworks
  - Bells and whistles and Cute tricks

- A “new” method does not truly need to be “new”
  - Very little in statistics is truly new

How to Start Developing a “New” Method: Identifying a Problem

No universal approaches, but some options include the following:

- Motivation from data
  - Is there some characteristic of the data that the “usual” methods cannot handle?
  - Is there are question arising from the data that nobody has answered before?
  - Are there standard questions (from other data sets) for which methods do not exist?

- Motivation from previous methods
  - Under what situations do existing methods fail?
  - Are there situations that an existing approach cannot handle? Can we do better?
  - Can we apply/translate an existing method to a new context?
  - I found a cool trick. Can I try incorporating it into an existing method?

Starting from an Existing Method

- Suppose you have a method that is interesting

- What are some limitations of the method?
  - Does the approach not work for some outcomes?
  - Are there situations where the approach doesn’t work well? (e.g. low power, slow computation, etc.)

- Can we adapt the method to a different context?

Example 1: Extending a method (1)

- Hilbert Schmidt Independence Criterion is a strategy used for testing generalized measure of dependency between two sets of multivariate data
  - Popular approach in machine learning literature
  - Paper: Gretton et al. (2005) *International conference on algorithmic learning theory*

- Problem: what if our data are cluster correlated?

- Paper: Liu et al. (2021) *NeurIPS*.
  - Deals with the problem by developing a test that accommodates correlation
  - Restrictions:
    - Large sample size
    - Same number of observations in the data

Example 2: Translating a method to new context

- Variance component and kernel machine based testing is a standard approach for genome wide association studies (GWAS) and genetic sequencing studies
  - Refs: Wu et al (2011) *American Journal of Human Genetics*
  - This method was itself a translation of other work from gene expression literature

- Microbiome is an emerging field: can we apply this approach within the context of the microbiome?
  - Paper: Zhao et al (2015) *American Journal of Human Genetics*
  - Direct application of genetics work to microbiome field
  - Minor tweaks to tailor it to microbiome data
  - Some modest technical contributions to get better variance estimates

Example 3: Combining methods

- From example 2, Zhao et al. propose a test for global microbiome association analysis
  - Limitation: cannot handle multivariate outcomes
  - Maity et al. (2012) developed an approach that extends the variance component tests for multivariate data by extending the Wu (2011) work.

- Zhan et al (2018) *Genetic Epidemiology*
  - Extends the Zhao (2015) work to accommodate multivariate outcomes
    - Borrows directly from Maity (2012)
  - Adapts the Maity (2011) work to accommodate microbiome data
    - Tailored to microbiome data

Finding a Problem from Real Data

- There are many ways to come up with problems to solve

- Suggested approach: examination of a “complicated” data set

- Advantages to this:
  - Easy to motivate the work: importance
  - Natural data application

- Down-sides:
  - Real data can suck

Example Data Set: Longitudinal GvHD Microbiome Study

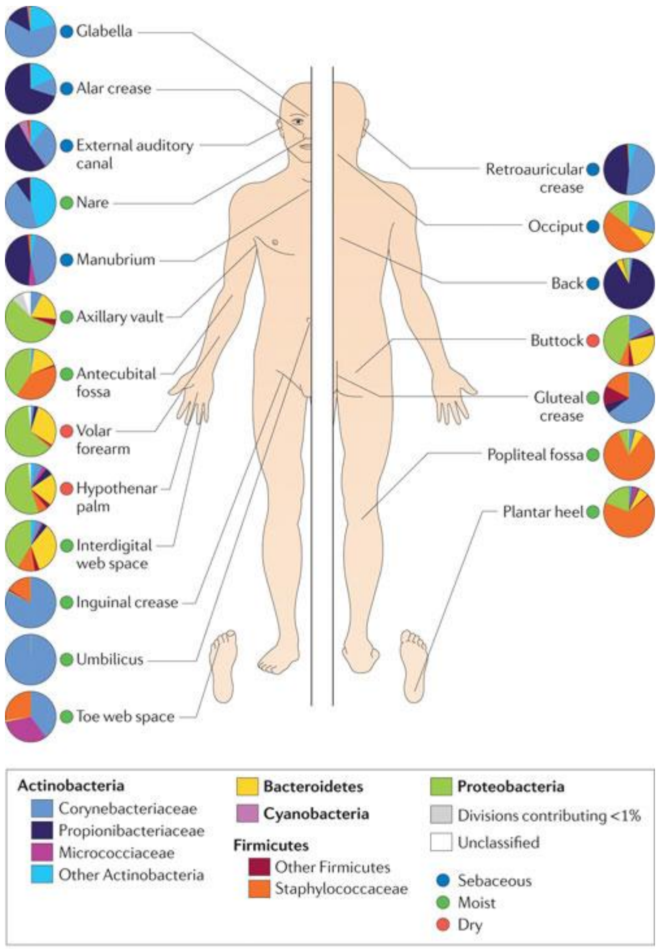
Bone Marrow Transplant and GvHD

- Bone marrow transplant is a standard therapy for many blood cancers, e.g. leukemia
  - Idea: transfer healthy blood-forming stem cells from a donor to you

- Graft-vs-host (GvHD) disease is a major complication:
  - The transferred (graft; from the donor) cells start attacking the body (host)
  - Results in considerable mortality

- Recently: evidence that gut microbiome may be closely related to development of GvHD

The Human Microbiome (Microbiota)



Nature Reviews | Microbiology

All the microbes that colonize a person

- 90% bacteria

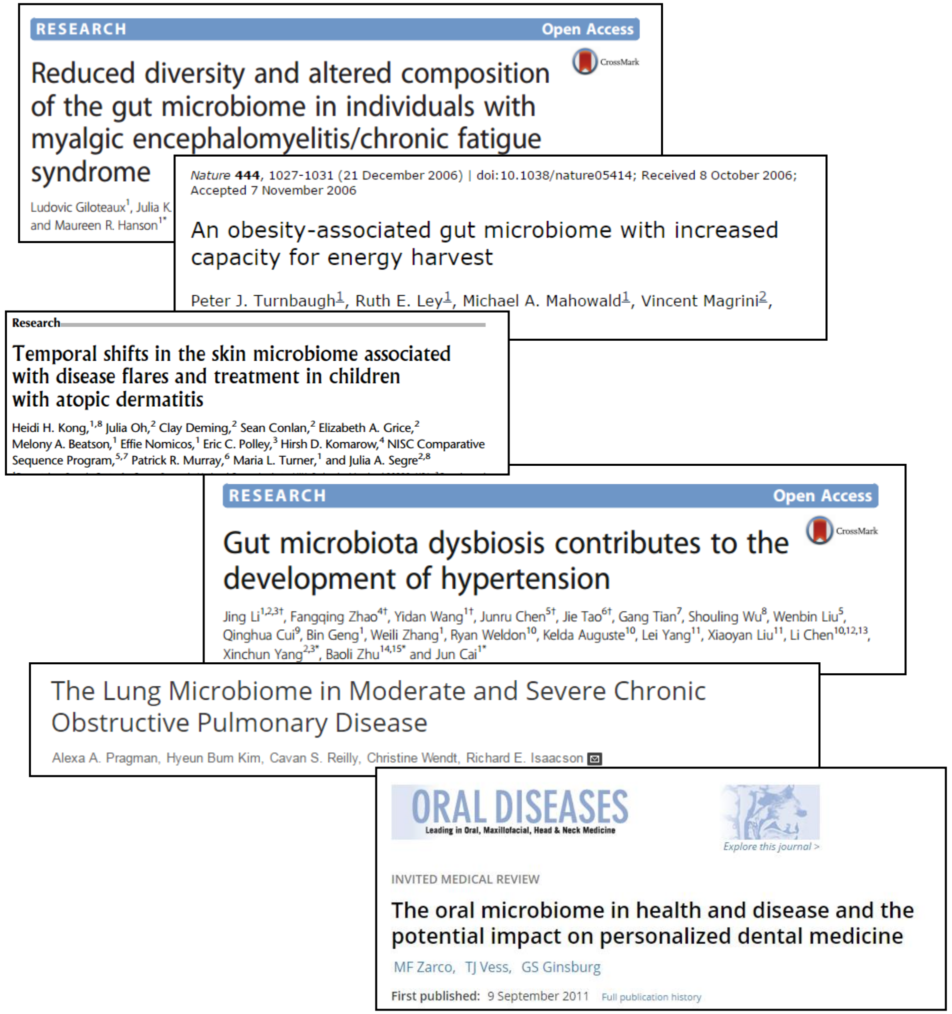
Humans contain as many bacterial cells as human cells

- 100x more bacterial genes than human genes

Found at nearly all body sites

- Composition varies by site and health status

Microbiome in Health and Human Disease



RESEARCH Open Access

Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome

Ludovic Giloteaux[1], Julia K.
and Maureen R. Hanson[1*]

Nature **444**, 1027-1031 (21 December 2006) | doi:10.1038/nature05414; Received 8 October 2006; Accepted 7 November 2006

An obesity-associated gut microbiome with increased capacity for energy harvest

Peter J. Turnbaugh[1], Ruth E. Ley[1], Michael A. Mahowald[1], Vincent Magrini[2],

Research

Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis

Heidi H. Kong,[1,8] Julia Oh,[2] Clay Deming,[2] Sean Conlan,[2] Elizabeth A. Grice,[2] Melony A. Beatson,[1] Effie Nomicos,[1] Eric C. Polley,[3] Hirsh D. Komarow,[4] NISC Comparative Sequence Program,[5,7] Patrick R. Murray,[6] Maria L. Turner,[1] and Julia A. Segre[2,8]

RESEARCH Open Access

Gut microbiota dysbiosis contributes to the development of hypertension

Jing Li[1,2,3†], Fangqing Zhao[4†], Yidan Wang[1†], Junru Chen[5†], Jie Tao[6†], Gang Tian[7], Shouling Wu[8], Wenbin Liu[5], Qinghua Cui[9], Bin Geng[1], Weili Zhang[1], Ryan Weldon[10], Kelda Auguste[10], Lei Yang[11], Xiaoyan Liu[11], Li Chen[10,12,13], Xinchun Yang[2,3*], Baoli Zhu[14,15*] and Jun Cai[1*]

The Lung Microbiome in Moderate and Severe Chronic Obstructive Pulmonary Disease

Alexa A. Pragman, Hyeun Bum Kim, Cavan S. Reilly, Christine Wendt, Richard E. Isaacson ✉

ORAL DISEASES
Leading in Oral, Maxillofacial, Head & Neck Medicine

Explore this journal >

INVITED MEDICAL REVIEW

The oral microbiome in health and disease and the potential impact on personalized dental medicine

MF Zarco, TJ Vess, GS Ginsburg

First published: 9 September 2011 Full publication history

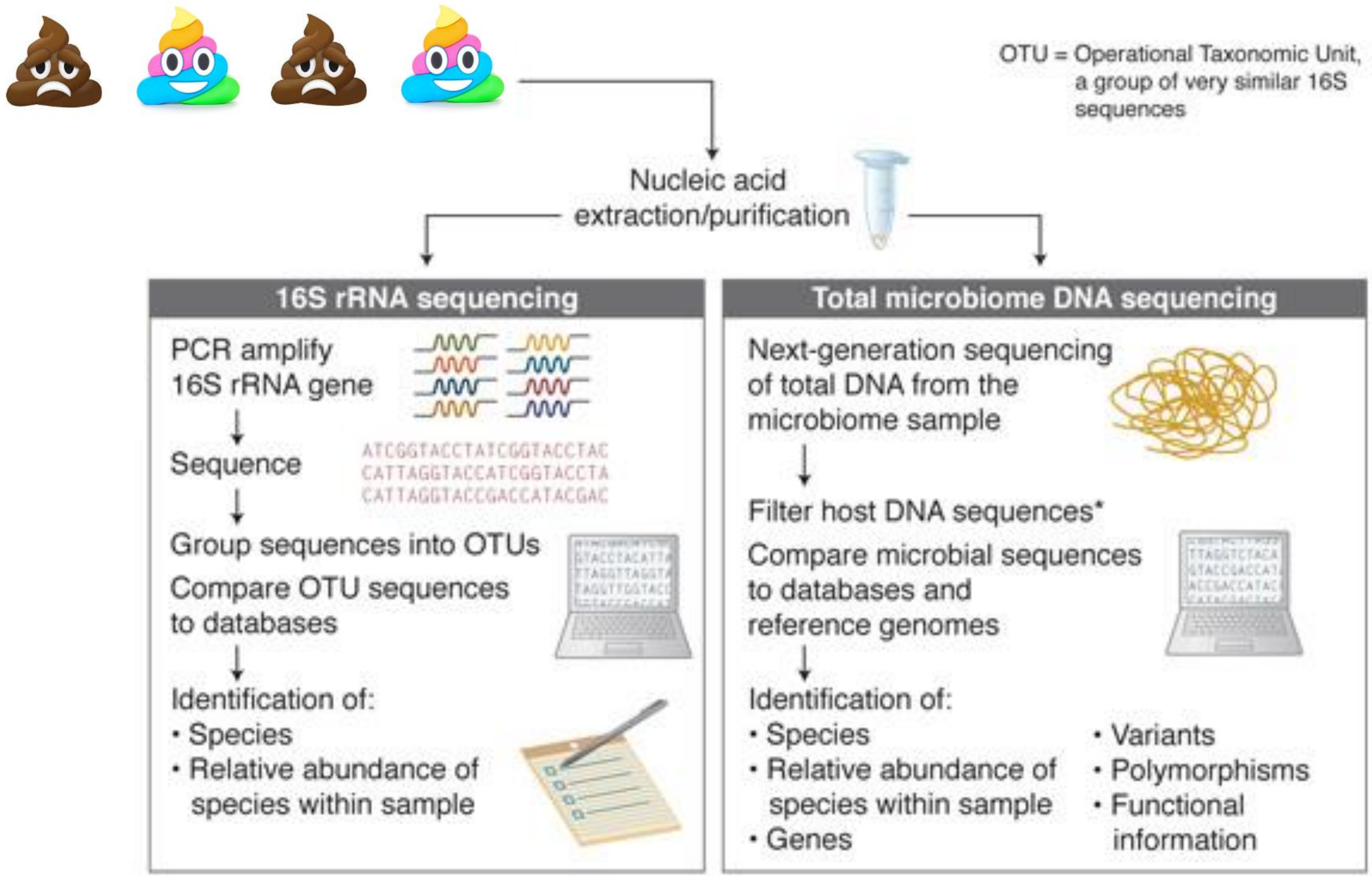
- **<u>Exposures</u>**
- Diet/Exercise
- Drugs/Alcohol/Smoking
- Treatment

- **<u>Outcomes (?)</u>**
- Asthma
- Cancer
- Diabetes
- Treatment Efficacy

Typical Gut Microbiome Experiment

- Get poo samples from individuals (e.g. healthy and affected subjects):

Typical Gut Microbiome Experiment



OTU = Operational Taxonomic Unit, a group of very similar 16S sequences

Nucleic acid extraction/purification

16S rRNA sequencing

PCR amplify 16S rRNA gene

↓

Sequence

↓

Group sequences into OTUs

Compare OTU sequences to databases

↓

Identification of:

- Species
- Relative abundance of species within sample

Total microbiome DNA sequencing

Next-generation sequencing of total DNA from the microbiome sample

↓

Filter host DNA sequences*

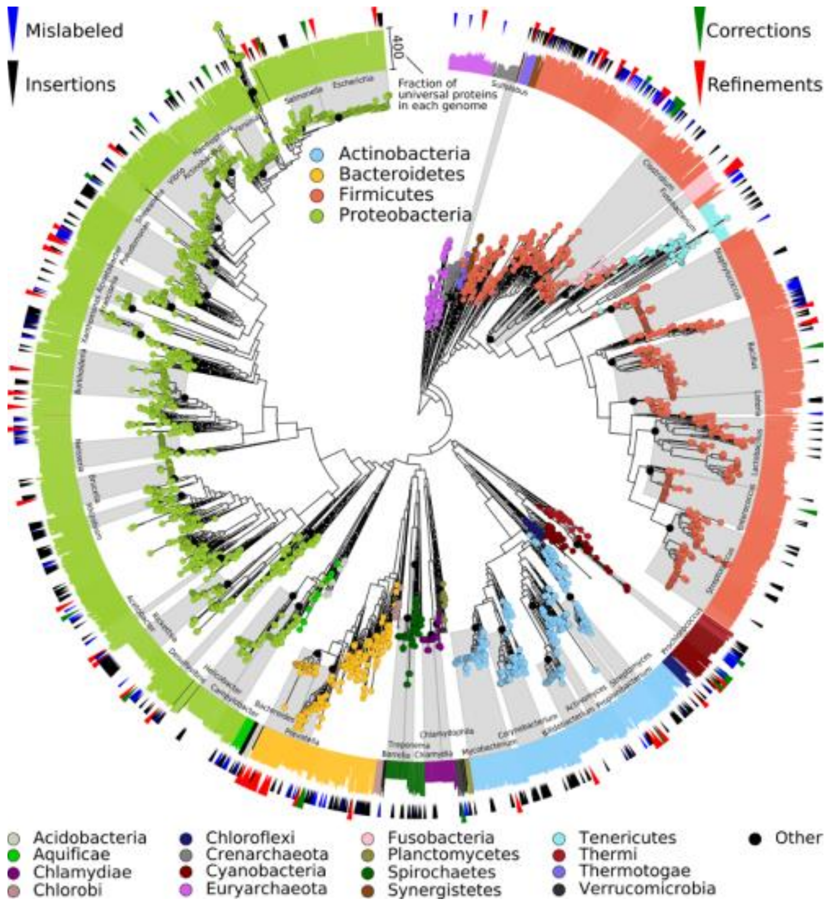
Compare microbial sequences to databases and reference genomes

↓

Identification of:

- Species
- Relative abundance of species within sample
- Genes
- Variants
- Polymorphisms
- Functional information

Microbiome Data at a Single Time Point



- **Microbiome data**
  - Taxon (e.g. species) is unit of analysis
  - Sequence reads quantifying taxa

- **High dimensional**
  - Many taxa
  - Count data
  - Zero Inflated
  - Over-dispersed
  - Compositional

- **Biological structure**
  - Phylogeny
  - Co-occurrence

Microbiome vs. GvHD Data Set

- Followed approximately patients from before transplant to 100 days after transplant

- Regular stool collection for each patient over time:
    - Microbiome profiling (multivariate data)

- Collection of hematologic markers: Not necessarily at same time as stool

- Demographic information

- **Objective:** study the relationship between microbiome and GvHD related variables

- Available online

IMPORTANT!!!

- Today: Go over interesting aspects of the data

- I want you to develop a method, NOT do a data analysis
    - These data are only meant to serve as a “context” for you to motivate methods
    - You do NOT need to address ALL aspects of the data
    - You do NOT need to fully understand the context in this case
    - You can IGNORE certain issues in the data while addressing others

What I am Providing to You:

- Microbiome data: GvHD_Microbiome_Data_571.csv
- Taxonomic Tree: Taxonomic Tree.csv
- Covariate information: GvHD_Covariates_571.csv
  - Also sometimes called meta data
- Data Dictionary for covariates: Dictionary.csv
- Additional biomarkers: GvHD_Biomarkers_571.csv

The Microbiome Data

Microbiome Data

patientID	sample_day	agvhday	agvhgrd	agvhgut	Escherichia_Shigella	Phocaeicola vulgatus	Phocaeicola dorei	Enterococcus rivorum	Enterocloster bolteae/clostridioformis	Enterococcus faecalis	Lactobacillus gasseri/johnsonii/paragasseri
0	-9	19	3	2	0	0	0	0	0	0	0
0	11	19	3	2	0	263	0	23	0	51	0
0	14	19	3	2	0	0	0	0	0	0	0
0	17	19	3	2	0	0	0	56	0	0	0
0	25	19	3	2	0	1601	0	3997	0	54	0
0	32	19	3	2	19867	2878	0	246	287	0	0
0	40	19	3	2	44	1035	0	0	489	0	357
0	45	19	3	2	0	594	0	150	154	0	3755
0	54	19	3	2	0	641	36	1146	102	0	6706
0	62	19	3	2	373	1226	54	6881	102	0	622
0	64	19	3	2	0	35	0	6079	74	40	78
0	73	19	3	2	0	478	0	8075	340	0	0
0	79	19	3	2	602	336	0	3567	70	64	25
0	91	19	3	2	0	712	0	61	121	0	328
0	100	19	3	2	0	23304	0	16302	0	7079	139
1	-6	NA	0	0	0	4506	0	0	104	0	0
1	5	NA	0	0	0	12220	0	0	386	0	0
1	10	NA	0	0	0	6144	0	40	1980	23	0
1	18	NA	0	0	0	1041	0	0	366	1	0
1	27	NA	0	0	0	7332	0	0	119	0	0

Key Characteristics of the Microbiome Data

- Counts

- Sparsity

- High dimensionality at each time point (multivariate data)
    - n = 229 and p > 850

- Structured (taxonomy)

- What makes this data set special: longitudinal collection

Count Data

- Each person has a microbial community, each sample just takes a few members of this community
    - Each person is a forest, and we capture a bunch of animals (members) from the forest

- Z_{ij} = # of taxon j in subject i
    - # of tigers, # of bears, etc. in the i-th forest

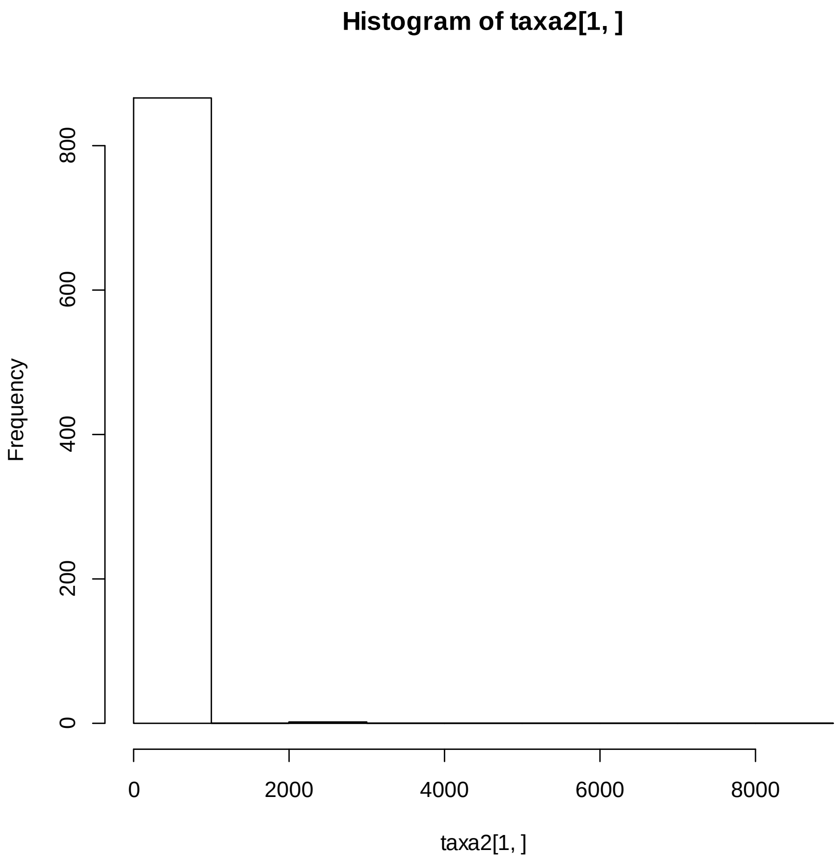
- Points:
    - Total number counts differs per person
        - Total number of animals captured in each forest is different, just due to chance
    - Over-dispersion

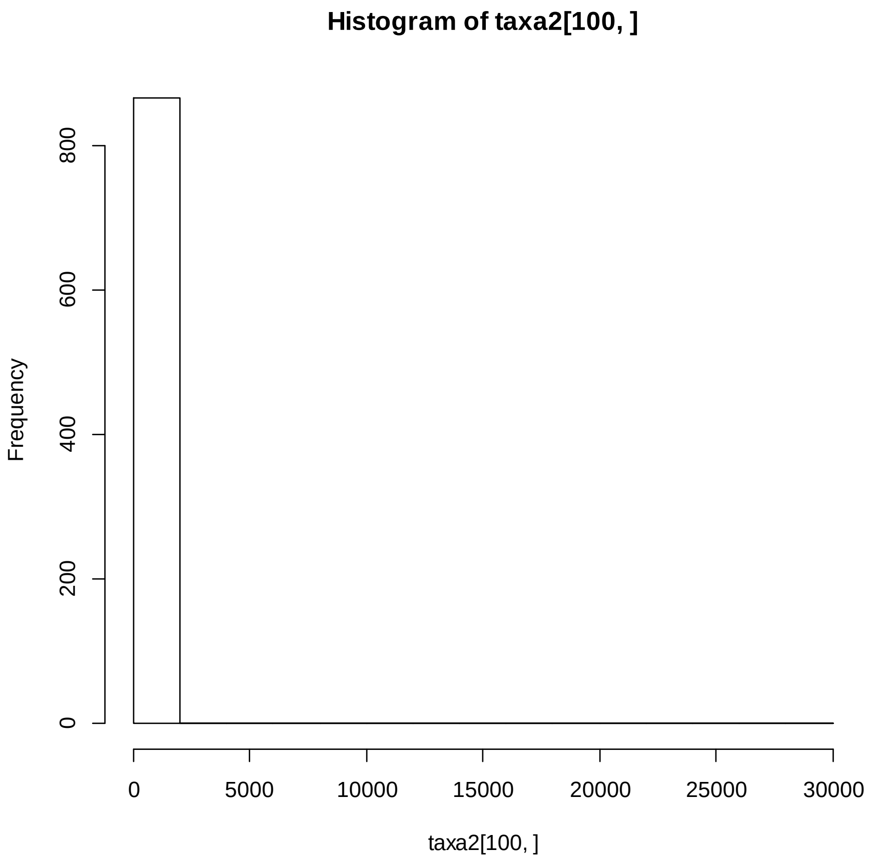
Counts: Statistically Interesting Issues

- Count data are discrete – interesting distributions

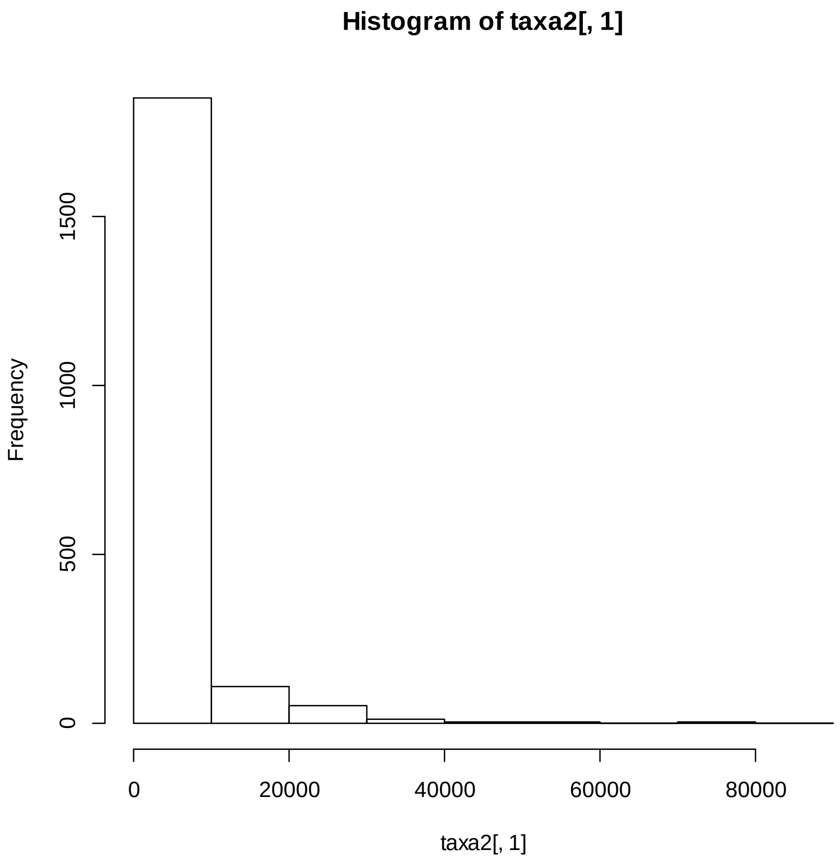
- Total counts vary across individual
  - Standard approaches:
    - Divide counts by total count for each individual: then the data for each person are the relative abundance (proportion) of each taxon
      - i.e. percent of captured animals that are bears, pct lions, pct tigers, etc.
      - Sort of continuous now – still overdispersed
      - Data are “compositional” now – total for each subject is 1
    - “Rarefy” the data: pick the sample with lowest count, then subsample counts from the others samples such that the total count for each sample is the same
    - Use an off-set in subsequent statistical modeling
  - Dealing with compositionality has spurred a lot of research recently
    - Compositionality can be ignored if modeling a single taxon

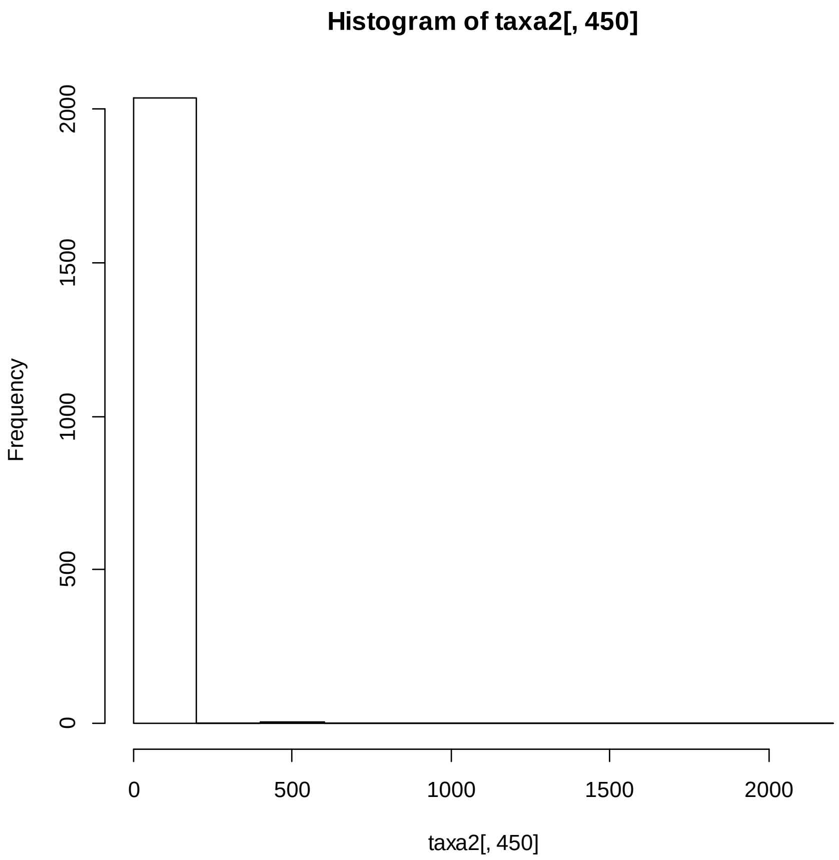
Counts for 2 Samples



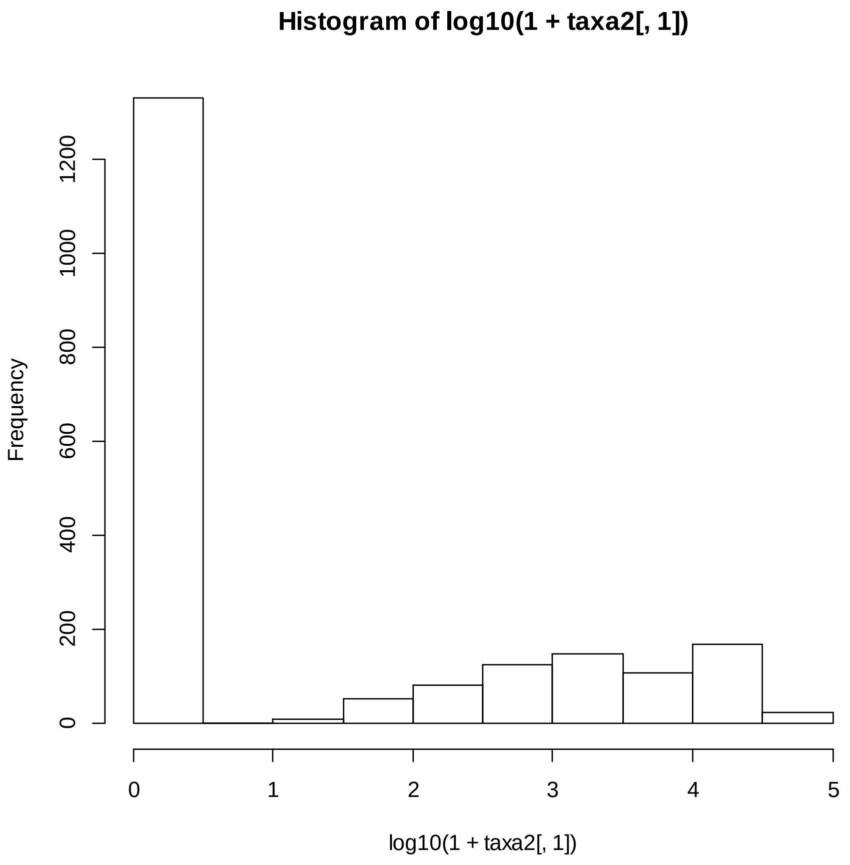


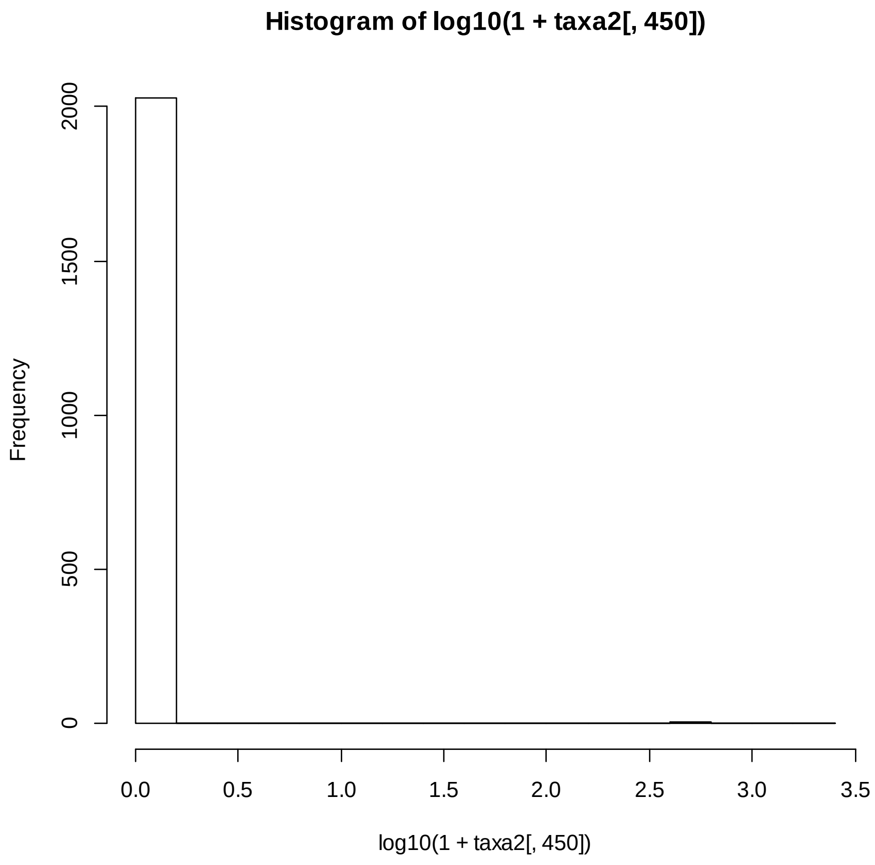
Counts for 2 Taxa





Counts for 2 taxa (logged)





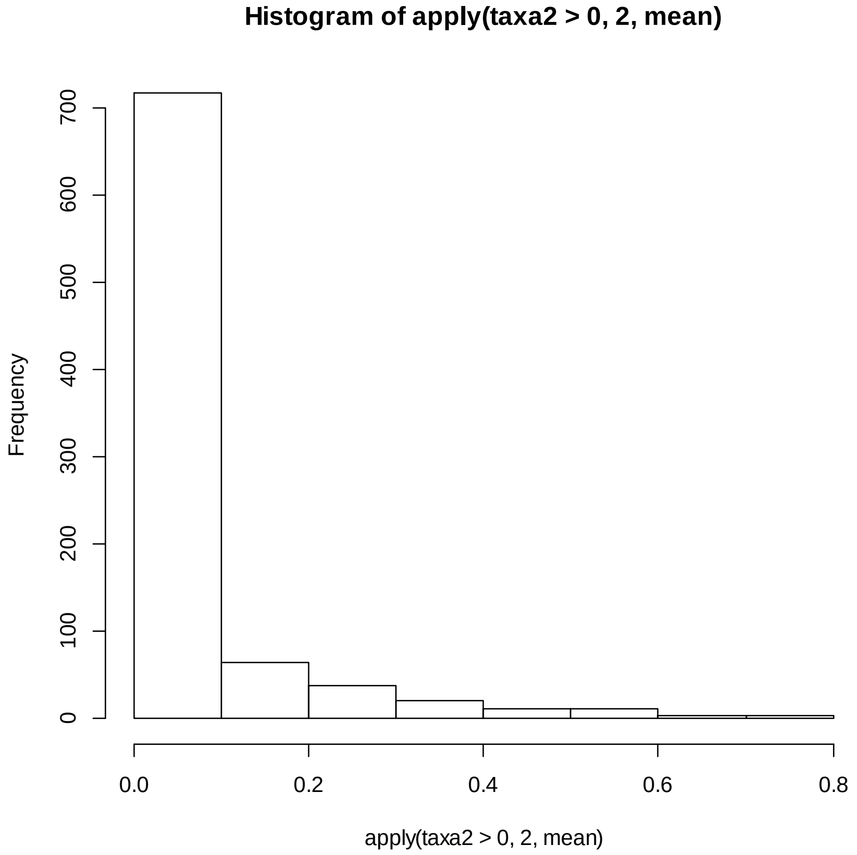
Sparsity: Lots of Zeros

- Issue: lots of taxa not detected in particular samples

- Lots of zeros can make modeling hard or reduce power
  - Cannot take logs
  - Sparsity reduces variance and therefore power in many cases

- Options:
  - Omit taxa found in only a few people (<3-5% of samples), then ignore zero counts
  - Add a small constant to all the data then transform
  - Analyze at higher level of taxonomic tree
  - Model the zeros, e.g. zero-inflated models (Statistically interesting!!!)

Histogram of Prevalances (Across Samples) of Different Taxa



High-Dimensionality

- Lots of taxa!

- Standard approach:
  - Analyze association between each taxon and outcome
  - Adjust the p-values (e.g. Bonferroni or FDR adjustment; “p.adjust” in R)

- Possibly interesting approaches:
  - Variable selection as an alternative
  - Other high dimensional modeling approaches
    - May want to deal with compositionality

Taxonomic Tree

Phylum	Class	Order	Family	Genus	Species_group	Species
Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Escherichia_Shigella	Escherichia_Shigella	Escherichia_Shigella
Bacteroidetes <Bacteroidetes>	Bacteroidia	Bacteroidales	Bacteroidales	Phocaeicola	Phocaeicola	Phocaeicola vulgatus
Bacteroidetes <Bacteroidetes>	Bacteroidia	Bacteroidales	Bacteroidales	Phocaeicola	Phocaeicola	Phocaeicola dorei
Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus	Enterococcus	Enterococcus rivorum
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Enterocloster	Enterocloster	Enterocloster bolteae/clostridioformis
Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus	Enterococcus	Enterococcus faecalis
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	Lactobacillus	Lactobacillus gasseri/johnsonii/paragasseri
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia	Blautia	Blautia wexlerae
Bacteroidetes <Bacteroidetes>	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides	Bacteroides fragilis
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia	Blautia	Blautia caecimuris
Bacteroidetes <Bacteroidetes>	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides	Bacteroides uniformis
Bacteroidetes <Bacteroidetes>	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides	Parabacteroides	Parabacteroides merdae
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Streptococcus	Streptococcus thermophilus

The Covariates

Covariates

sub_ID	txage	race	sex	donrel	donsex	donage	status	celltxl	relday	agvhday	agvhgrd	agvhskn	agvhlvr	agvhgut	cgvhday	cgvhgrd
0	65.87635	Caucasian	Female	Not Related	Male	34.28107	Relapse	PBSC	NA	19	3	0	0	2	114	Clinical
1	69.60301	Caucasian	Male	Sibling	Male	68.93675		PBSC	173	NA	0	0	0	0	NA	
2	65.4095	Caucasian	Male	Not Related	Female	14.98895	Remission	PBSC	27	NA	0	0	0	0	NA	
4	59.96767	Caucasian	Female	Not Related	Male	28.05909	Relapse	PBSC	NA	NA	0	0	0	0	395	Clinical
5	43.75315	Pacific Islande	Female	Not Related	Male	53.45538		BM	NA	28	2	0	0	1	253	Clinical
6	43.85209	Caucasian	Male	Not Related	Male	22.63076	Remission	PBSC	NA	25	2	3	0	0	NA	Normal
7	59.26321	Caucasian	Female	Not Related	Male	38.73873	Relapse	PBSC	1102	30	2	2	0	1	99	Clinical
8	54.05751	Caucasian	Male	Not Related	Female	60.13194	Remission	PBSC	96	20	2	0	0	1	NA	Subclinical
9	59.64431	Caucasian	Female	Not Related	Female	27.21501	Relapse	PBSC	NA	7	2	2	0	1	125	Clinical

Data Dictionary

Field	Group	Description
AGVHDAY	Acute GVH:	Day
AGVHGRD	Acute GVH:	Overall grade (0-4,5,9)
AGVHGUT	Acute GVH:	Gut (0-4,5,9)
AGVHLVR	Acute GVH:	Liver (0-4,5,9)
AGVHSKN	Acute GVH:	Skin (0-4,5,9)
CELLTXL	Cells:	Type(s) of cells infused at transplant (BM, PBSC, CORD)
CGVHDAY	Chronic GVH:	Day
CGVHGRD	Chronic GVH:	Grade (Clinical, Subclinical,Abnormal,Normal)
DONAGE	Donor:	Donor age at transplant (not necessarily age at BM\PBSC Collection)
DONREL	Donor:	Donor relation specific (Sibling, Parent, Child...)
DONSEX	Donor:	Donor sex
RACE	Demographic:	Patient race
RELDAY	Relapse:	Day
SEX	Demographic:	Patient sex
STATUS	Diagnosis:	Status at/pre-transplant (Remission,Relapse)
TBIDOSE	Transplant:	Total Body Irradiation dose
TX	Transplant:	Number
TXAGE	Transplant:	Age in years

Some Scientific Questions (Longitudinally Speaking…)

- Which taxa are associated with GvHD?

- Which taxa are associated with development of GvHD?

- Which taxa are associated with time to development of GvHD?

- Which taxa are associated with grade of GvHD?

- Which taxa interact with other variables?

Remember: These are questions that *motivate* methods beyond the question themselves

Standard Analyses

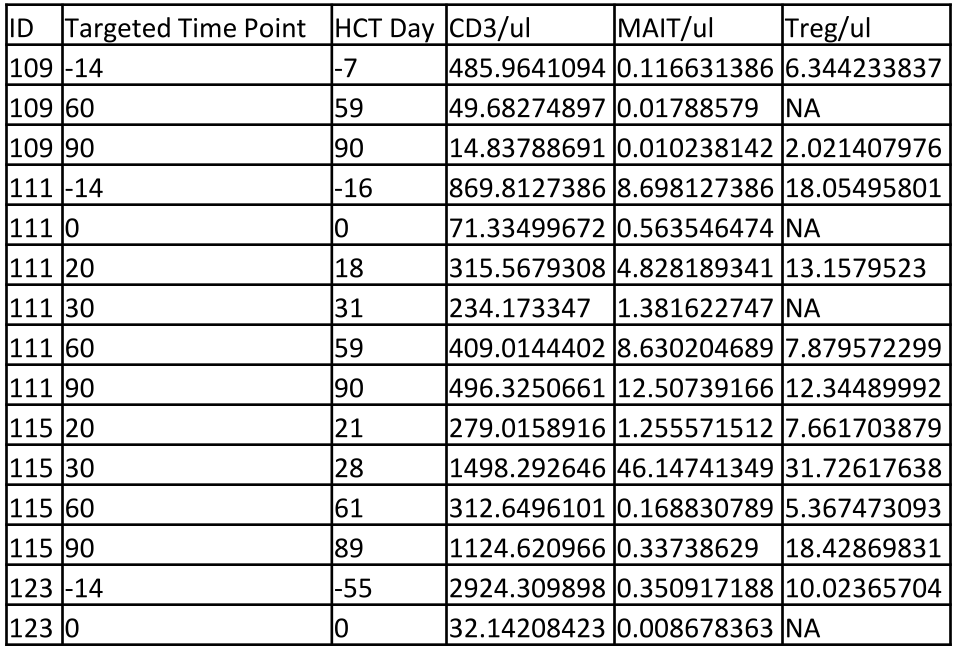
- Standard Analysis:
  - Regress (transformed) taxon abundance on variables of interest using LMM or GEE with assuming Gaussian outcome (probably more standard)

    OR

  - Regress variable(s) of interest on taxon abundance

Biomarker Data

Biomarkers



ID	Targeted Time Point	HCT Day	CD3/ul	MAIT/ul	Treg/ul
109	-14	-7	485.9641094	0.116631386	6.344233837
109	60	59	49.68274897	0.01788579	NA
109	90	90	14.83788691	0.010238142	2.021407976
111	-14	-16	869.8127386	8.698127386	18.05495801
111	0	0	71.33499672	0.563546474	NA
111	20	18	315.5679308	4.828189341	13.1579523
111	30	31	234.173347	1.381622747	NA
111	60	59	409.0144402	8.630204689	7.879572299
111	90	90	496.3250661	12.50739166	12.34489992
115	20	21	279.0158916	1.255571512	7.661703879
115	30	28	1498.292646	46.14741349	31.72617638
115	60	61	312.6496101	0.168830789	5.367473093
115	90	89	1124.620966	0.33738629	18.42869831
123	-14	-55	2924.309898	0.350917188	10.02365704
123	0	0	32.14208423	0.008678363	NA

Biomarker data

- Hematologic (blood) biomarkers measured on a subset of individuals

- Can treat individual markers in the same way as other data, but some caveats:
    - Irregular distributions?
    - Captured at different time points than microbiome?
    - Interest in multivariate analysis of all markers?

- Same principle: this is methods development
    - You do not *have* to use these data unless interesting for your method
    - You do not need to capture all aspects of the data: e.g. missing values if your method is not concerned with missingness

Deriving Some (Applied) Statistical Ideas from the Data

- Improved distributions for modeling individual microbes
  - Zero inflation
  - Over-dispersion
  - Count/continuous

- Incorporating hierarchical taxonomic structure into the analysis
  - Joint analysis of multiple microbes in a group

- Compositionality concerns:
  - Matters if you are regressing outcomes on ALL taxa… e.g. variable selection methods

- GvHD is something that arises in the course of the study
  - Joint modeling of longitudinal and time to event outcome (deep area)

- Dealing with sparsity of the data
    - Are zeros truly zero? Can we borrow information from other time points (and samples) to impute zeros?

- Prediction of outcomes from longitudinal data
    - Can we use a “longitudinal profile” to predict eventual GvHD?

- What if there is a lagged effect? What if microbiome at previous time points predicts outcome at current time point?
    - Lagged or transition model?

- Joint modeling of multiple longitudinal outcomes (e.g. biomarkers) in relation to one or more taxa?

- Figuring out what to do with biomarkers that are not measured at the same time points

- Data visualization
  - Multi-dimensional scaling plots taking into account longitudinal data

- Identification of samples that are outliers from all the others

- Clustering of the data based on their longitudinal profiles
  - Remember that the data are multivariate at each time point and unbalanced