BIOST/STAT 571: Final Project (Part 1)

February 10, 2021

Final Project

- **Objective:** Develop a “new” method or approach with the aim of publishing in a statistical or applied journal

- **Well Regarded Journals** (not comprehensive list!!)
  - Statistical Methods: Biometrika, JASA T&M, JRSS-B, Biometrics
  - Applied Methods: JASA A&CS, Biometrics (Practice), Biostatistics, Statistics in Medicine, JRSS-C (Applied Statistics)
  - Computational: CSDA, Journal of Statistical Computation and Simulation, JCGS
  - Informatics: Bioinformatics, Genetic Epidemiology, PLoS Journals, Neuroimage
  - Applied Journals: Journal of Clinical Oncology, American Journal of Human Genetics, Nature Journals

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

What Goes into Development of a New Method

1. The problem that is being solved

2. Methods and statistical framework for solving problem

3. Justification and Evaluation of the method
   - Why are you solving this problem?
   - Why do you need a “new” or different approach?
   - Can you study the theoretical properties of your approach?
   - Computational considerations and algorithms?
   - Under what circumstances will your approach work? When will it fail?
   - What happens with real data?

4. Iterate through 1-3

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?

Building Your Method

- Focus on specific aspects of the problem that you want to address

- Depends on what you're trying to do:
  - Better model
  - Better algorithms
  - Better theorys
  - Etc.

Justification and Evaluation a Method

- Why someone should care (most important part)
  - Sometimes, honest lies
- Theory and properties
  - Asymptotics usually
  - Finite sample theory rarely
- Simulations
- Data Applications
- Generally: No method universally wins, just want to show that yours *can* win and issue guidance

Paper Structure

- No definitive structure for papers: depends on context and the journal
- A Rough typical structure: (Not necessarily section headings!)
  - Introduction
  - Methods and theory
  - Results
  - Discussion
- Good idea: follow structure of relevant papers
- Bad Idea: follow structure of relevant papers
- Main idea: how would you explain and justify your approach to others?

Introduction

- By far the most important part of the paper
- What background material is necessary?
- What is the problem that you are solving?
- Why is the problem important?
- What related work has already been done?
- What is the approach that you are taking?

Methods

- (Sometimes) prior related work
- Proposed approaches and models
- Algorithms for implementation
- Theory:
  - Justification for your approach
  - Theoretical comparisons of your method to existing approaches
- This may take multiple sections in a paper

Results

- Empirical evaluation of your method (should back up what you say in the intro)
  - Comparisons with existing approaches and relevant metrics for evaluation
  - Comparisons of different options of your approach (e.g. should one use CV vs. AIC)
  - Sensitivity analysis

- Simulation scenarios (sometimes in methods) and simulation results

- Real data applications:
  - Show that your method works on real data
  - What insights does your method provide that are new or unusual?
  - For your final project:
    - Apply your method to real data
    - Do not need to give significant insights (or even work well, e.g. poor type I error control due to sample size)
    - Need to explain what you're seeing in terms of behavior of your method

Discussion/Conclusions

- What did you do in this paper and what did you show?
- Recap of when your method wins and when it loses.
- What are the options that go into your method? Recommendations for which to use?
- What are things that you would have liked to investigate further but are outside of the scope? Future research?
- What are things that others are likely to pick on? Pre-empt their comments.

Details of your Final Project

- **Groups:**
  - Tentative groups assigned in next couple days
  - Finalized groups by weekend

- **Grading:**
  - The paper will be evaluated on the basis of originality, scholarship (including appropriate literature citations), clarity, organization and relevance to class goals.
  - Not all group members may receive same score: you will be asked for relative contributions (HW4)
  - Creativity and thoroughness of evaluation count

- **Due:** 5pm on Thursday, March 18, 2021
  - E-mailed to: Instructor, TA's and ALL group members

Deliverable: The Paper

- Ideally: something publishable or that is close to publishable as a methods paper

- Length:
  - No restriction as long as it is complete as a paper
  - Expect about +/- 15-20 pages double spaced (not including any figures)

- Format:
  - No set format, but probably good idea to follow usual structure
  - Template: available next week

- Note:
  - HW: can you do this?
  - Project: what can you do?

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

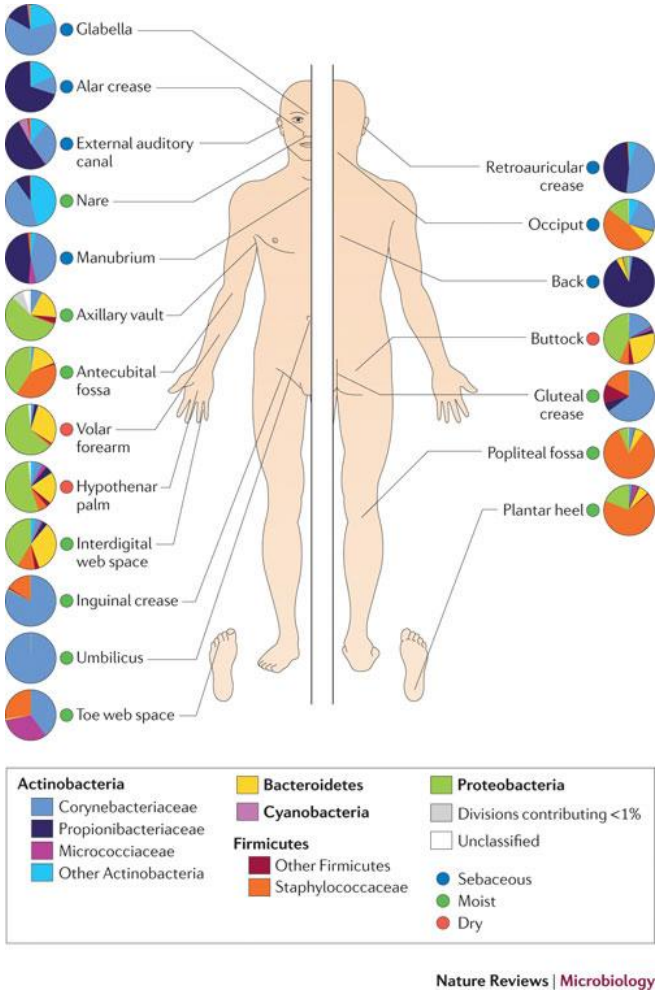
Bone Marrow Transplant and GvHD

- 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)



All the microbes that colonize a person

- 90% bacteria

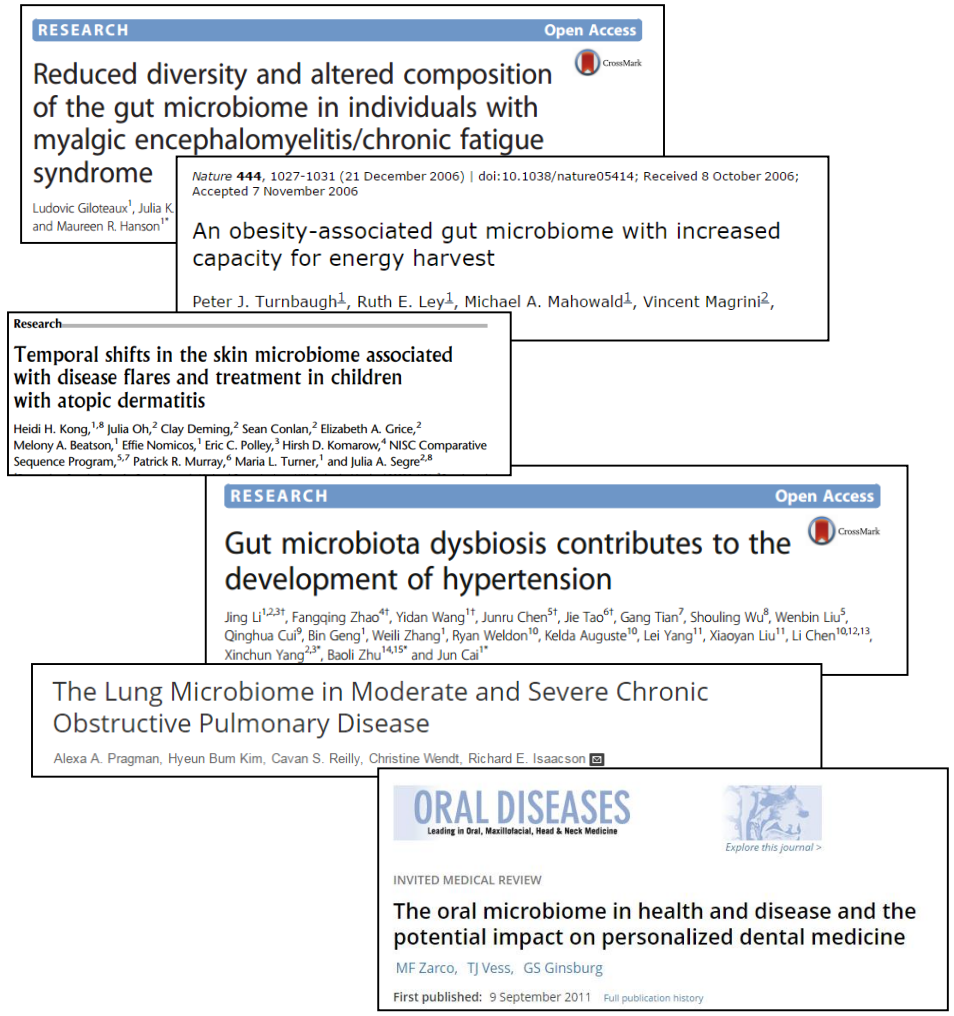
Humans contain as many bacterial cells as human cells

- 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



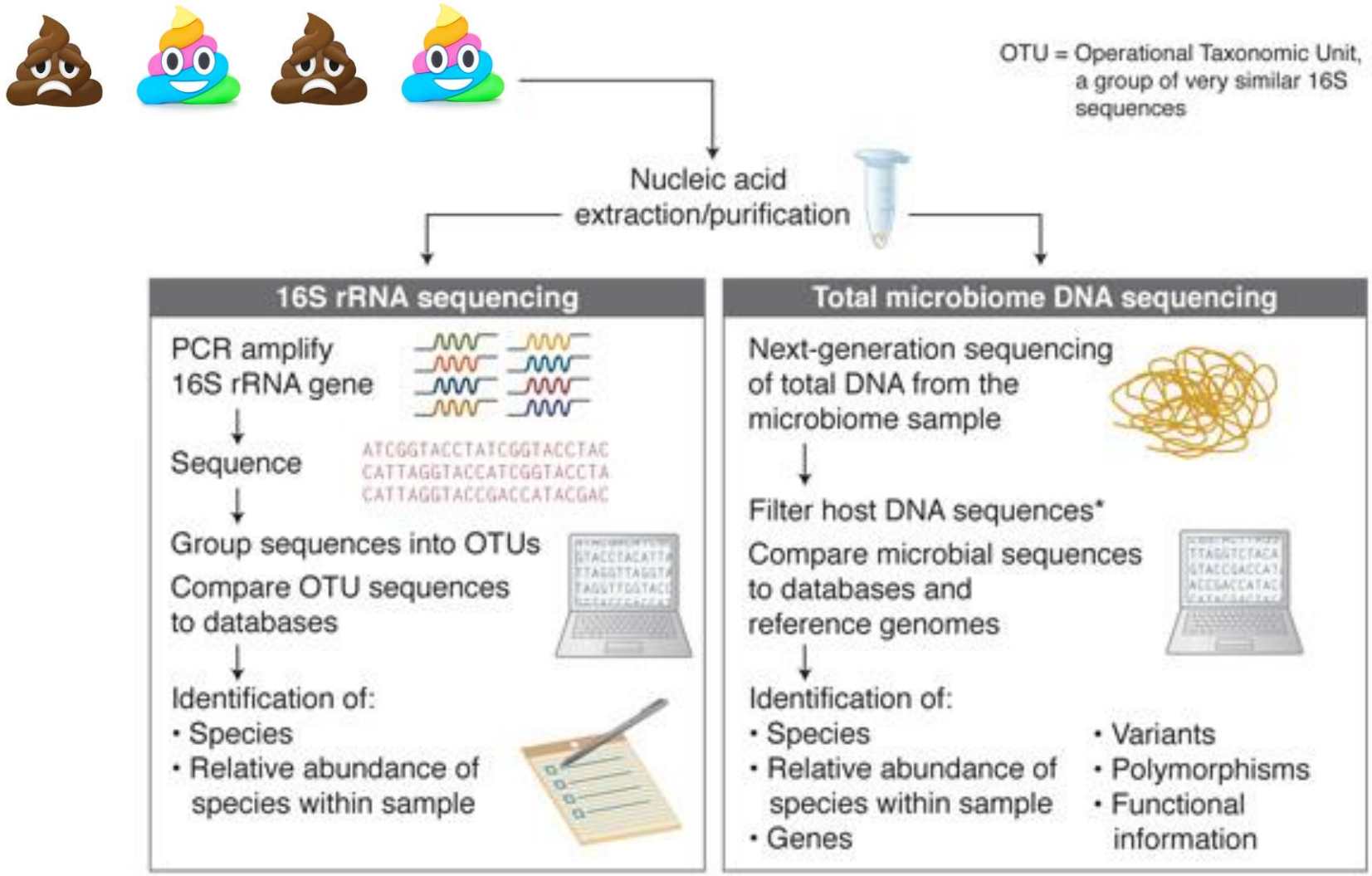
- **<u>Exposures</u>**
- Diet/Exercise
- Drugs/Alcohol/Smoking
- Treatment

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

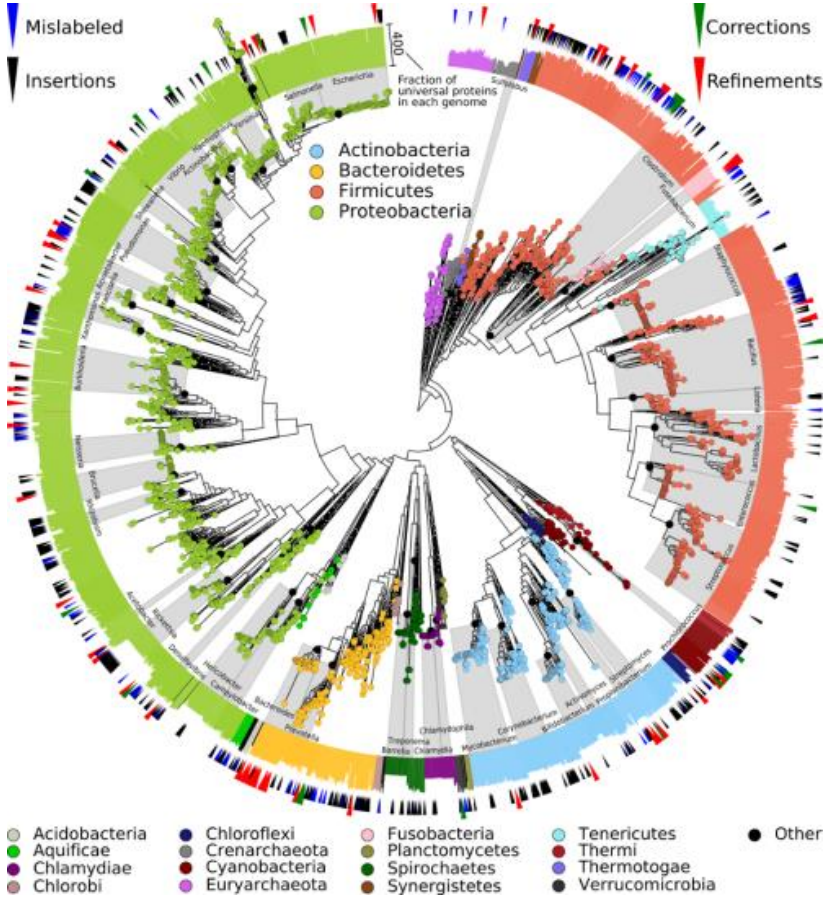
Typical Gut Microbiome Experiment

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

Typical Gut Microbiome Experiment



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 in next couple days

Friday

- We will look more closely at the data set
- We will talk about potentially problematic aspects of the data
- We will discuss usual approaches to the data analysis