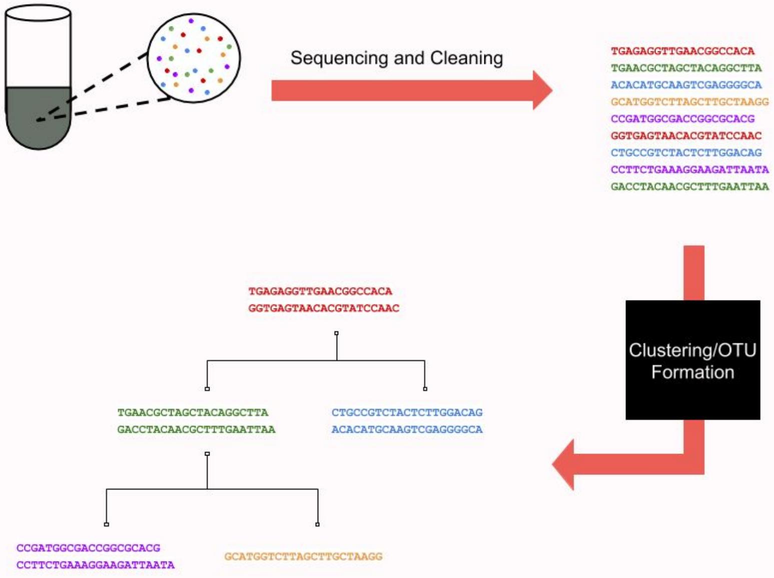
# Applying NLP Techniques to 16s rRNA Sequence Data

Camellia Hilker

Metis - Summer 2019

# Current method for microbiome evaluation

- Processed using pre-packaged pipelines such as QIIME[1,2]
- Perform QC on data (denoising and confidence score)
- Cluster reads based on OTUs
- Provides abundance and diversity metrics, draws phylogenetic tree
- Output files are not in a processable format



Sequencing and Cleaning

TGAGAGGTTGAACGGCCACA
TGAACGCTAGCTACAGGCTTA
ACACATGCAAGTCGAGGGGCA
GCATGGTCTTAGCTTGCTAAGG
CCGATGGCGACCGGCGCACG
GGTGAGTAACACGTATCCAAC
CTGCCGTCTACTCTTGGACAG
CCTTCTGAAAGGAAGATTAATA
GACCTACAACGCTTTGAATTAA

Clustering/OTU Formation

TGAGAGGTTGAACGGCCACA
GGTGAGTAACACGTATCCAAC

TGAACGCTAGCTACAGGCTTA
GACCTACAACGCTTTGAATTAA

CTGCCGTCTACTCTTGGACAG
ACACATGCAAGTCGAGGGGCA

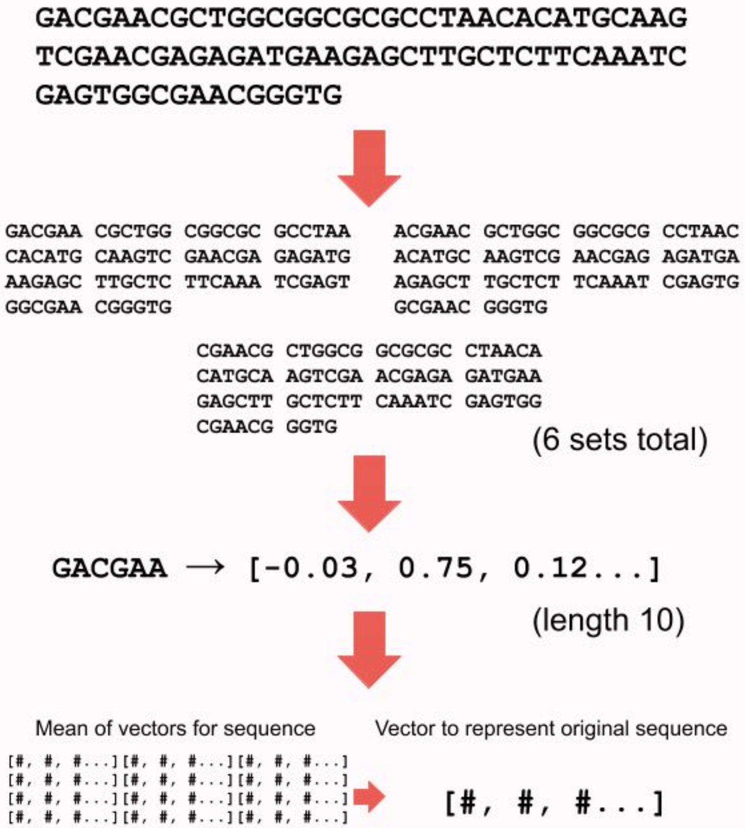
CCGATGGCGACCGGCGCACG
CCTTCTGAAAGGAAGATTAATA

GCATGGTCTTAGCTTGCTAAGG

***Operational Taxonomic Units (OTUs)** - obtained clusters of bacteria that have 97-99% similar sequences

# Proposed NLP inspired method

- Gensim[4] Word2Vec embeddings
- Train using GreenGenes database
- Sequences are broken into ‘words’ of six nucleotide bases (letters)
- A 10-dimensional vector is created for each word based on skip-grams window of 2
- Each sequence is represented by the Average of all word vectors associated with that sequence



GACGAACGCTGGCGGCGCGCCTAACACATGCAAG
TCGAACGAGAGATGAAGAGCTTGCTCTTCAAATC
GAGTGGCGAACGGGTG

GACGAA CGCTGG CGGCGC GCCTAA
CACATG CAAGTC GAACGA GAGATG
AAGAGC TTGCTC TTCAAA TCGAGT
GGCGAA CGGGTG

ACGAAC GCTGGC GGCGCG CCTAAC
ACATGC AAGTCG AACGAG AGATGA
AGAGCT TGCTCT TCAAAT CGAGTG
GCGAAC GGGTG

CGAACG CTGGCG GCGCGC CTAACA
CATGCA AGTCGA ACGAGA GATGAA
GAGCTT GCTCTT CAAATC GAGTGG
CGAACG GGTG

(6 sets total)

GACGAA → [-0.03, 0.75, 0.12...]

(length 10)

Mean of vectors for sequence

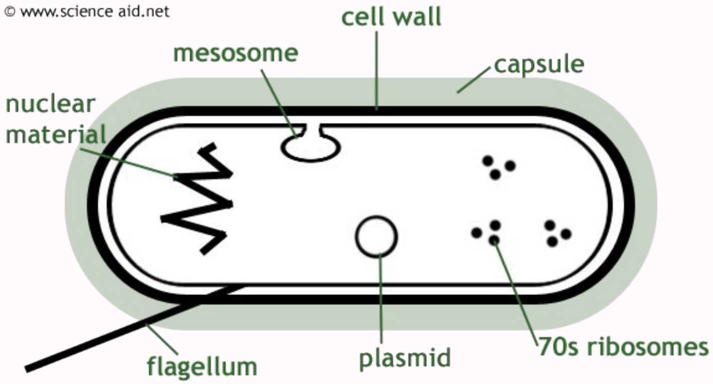
[#, #, #...][#, #, #...][#, #, #...]
[#, #, #...][#, #, #...][#, #, #...]
[#, #, #...][#, #, #...][#, #, #...]
[#, #, #...][#, #, #...][#, #, #...]

Vector to represent original sequence

[#, #, #...]

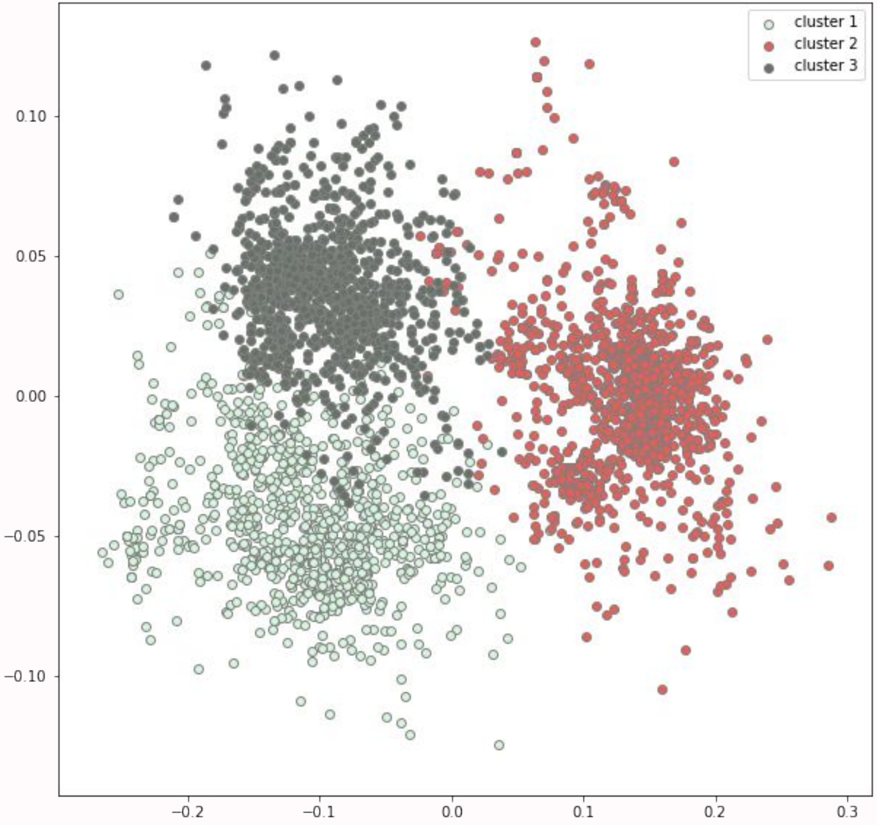
# Human Microbiome Project -T2D Study

- Study collected fecal samples from patients with type 2 diabetes
- 10-90k sequence reads per sample*
- 20 patients, ~700MB of sequence data*
- Two phyla occupy 90% of gut microbiome[9]:
  - Bacteroides
  - Firmicutes
- Levels of phyla may vary between diabetic/obese and healthy individuals[6]


© www.science aid.net

# Word2Vec Clustering

- Gut bacteria can be classified into 3 Enterotypes[5]
  - Bacteroides
  - Prevotella
  - Ruminococcus
- sklearn Agglomerative clustering
- Clustered on 10-dimensional sequence vectors
- Plotted using two-dimensional PCA

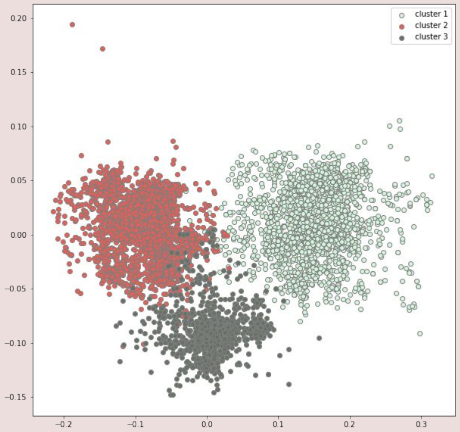


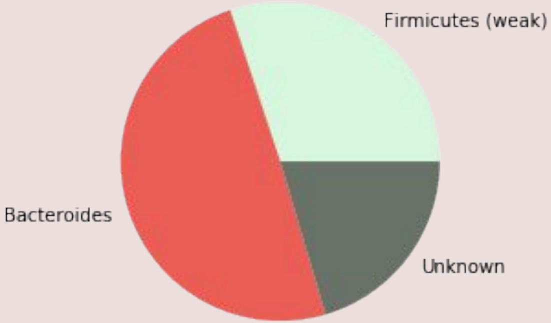
## Caucasian Female

## Age-52

## Non-Obese

## Insulin Sensitive



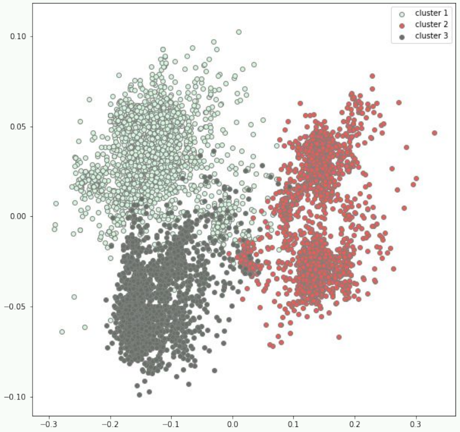


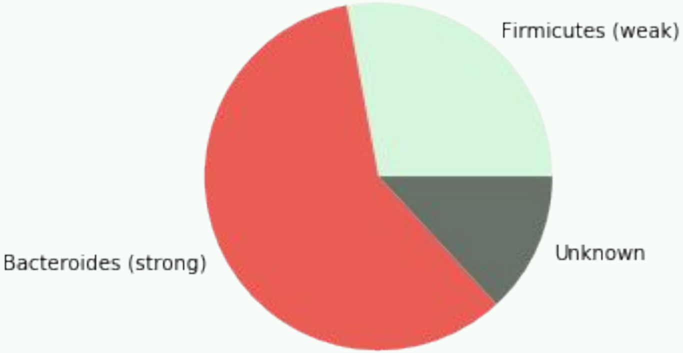
## Caucasian Female

## Age-62

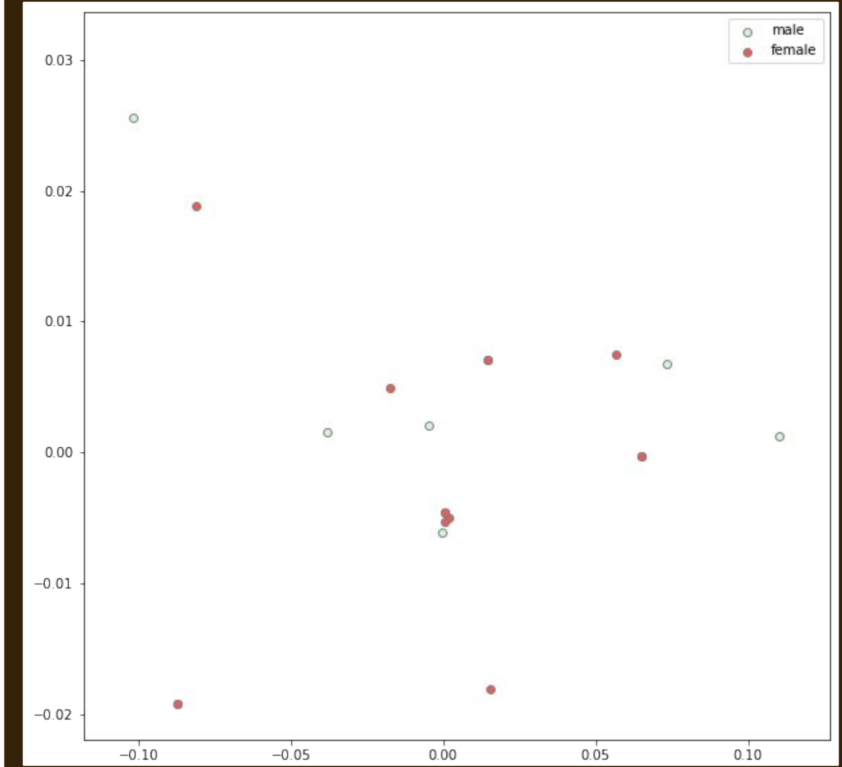
## Obese

## Insulin Resistant

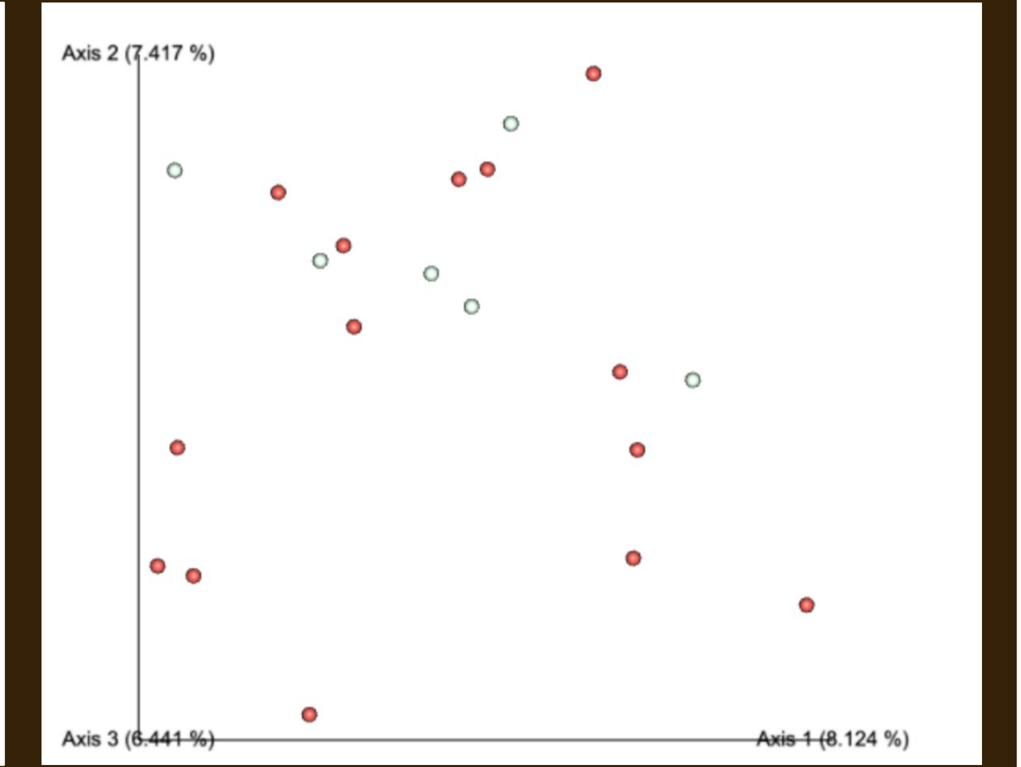




## Clustering Patients based on Overall Average Word Vector (NLP)



## Clustering Patients based on Jaccard distance (QIIME)



# How does the NLP method compare to QIIME?

- NLP word Count: 4,581
- “Word” Minimum count = 90
  - Word must appear in 90 sequences
- [CALCULATE EQUIVALENT METRICS HERE]

- QIIME Feature Count: 2,023
- Sequence depth = 2,000
  - Feature must be present in 2,000 sequences
- Average Number of OTUs = 160
- Average Pielou’s evenness index = 0.72

Questions?

# Works Cited

1. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857 (2019).
2. Caporaso, J Gregory et al. “QIIME allows analysis of high-throughput community sequencing data.” Nature methods vol. 7,5 (2010).
3. Woloszynek, Stephen et al. “16S rRNA sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses.” PLoS computational biology vol. 15,2 (2019).
4. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New …, (2010).
5. Arumugam, Manimozhiyan et al. “Enterotypes of the human gut microbiome.” Nature vol. 473,7346 (2011): 174-80. doi:10.1038/nature09944
6. Shen, Jian, Martin S. Obin, and Liping Zhao. "The gut microbiota, obesity and insulin resistance." Molecular aspects of medicine 34.1 (2013): 39-58.
7. Redel, Henry et al. “Quantitation and composition of cutaneous microbiota in diabetic and nondiabetic men.” The Journal of infectious diseases vol. 207,7 (2013): 1105-14. doi:10.1093/infdis/jit005
8. Devaraj, Sridevi et al. “The human gut microbiome and body metabolism: implications for obesity and diabetes.” Clinical chemistry vol. 59,4 (2013): 617-28. doi:10.1373/clinchem.2012.187617
9. Suau, A et al. “Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut.” Applied and environmental microbiology vol. 65,11 (1999): 4799-807.