Insight data brainstorming



<ftp://ftp.microbio.me/AmericanGut/ag-July-29-2016/>

ag-July-29-2016/02-filtered/

fecal-sequences.fna

OTU (observed taxanomic unit)

One cancer + gut microbe

Colorectal cancer: uncontrolled growth of cells within the large intestine (colon).

Colon cancer/bowl cancer

>50 yrs old

Diet high in fat

*Fusobacterium*

Fusobacterium are present at increased levels in colorectal cancer patients and can trigger inflammatory responses associated with [tumour](https://www.britannica.com/science/tumor) growth and progression.

abnormal digestion and elimination

Some bacterial species have been identified and suspected to play a role in colorectal carcinogenesis, such as *Streptococcus bovis*, *Helicobacter pylori*, *Bacteroides fragilis*, *Enterococcus faecalis*, *Clostridium septicum*, *Fusobacterium* spp. and *Escherichia coli*. The potential pro-carcinogenic effects of these bacteria are now better understood.

**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4716055/**

<http://vegetablepharm.blogspot.com/2014/12/american-gut-and-ubiome-compared.html> Where to get AmGut data

About available data

1. User data: Tim Steele blog, already downloaded
2. Amerigut: ftp://ftp.microbio.me/AmericanGut/latest/11-packaged/

Definition of healthy adults (fecal samples)

* Age: 20-69
* BMI: 18.5 – 30
* No history of : Inflammatory Bowel Disease, Diabetes, Antibiotic use in the past year

Source: ftp://ftp.microbio.me/AmericanGut/latest/11-packaged/readme.txt

Parent partition/data\_dictionary

OTU: ftp://ftp.microbio.me/AmericanGut/latest/11-packaged/fecal/100nt/all\_participants/all\_samples/10k/ ag\_10k\_fecal.biom

Metadata: XXX.txt (the information about samples)

1. Training data

Publication:

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155362>

**Week 2: Data acquisition**

1. WGSS

* WGSS DC (52 CRC + 52 controls) + metaHIT (cit 12, 13) + HMP phase I (cit 14)
* mOTUs (no differences btw CRC and controls)

1. Abundancy (not many correlations)

Use WGSS or OTUs

1. Pathways/modules/genes

In pop F (cit 8) strong association with KEGG genes (relative abundance) (NOT presence)

In WGSS DC, no association (NOT presence)

1. Four genes

K00605

K01667

K12267

K01421

Three modules

M00036

M00011

M00017

Seven pathways

Ko00020

Ko05014

Ko00300

Ko00281

Ko00910

Feature choices:

Do I put more weights on the several genes proved to have correlations (only these?)

http://www.ebi.ac.uk/ena/data/view/PRJEB12449

<https://www.ncbi.nlm.nih.gov/pubmed/24316595>

(Older paper of DC data in 16S)

Population F

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4299606/pdf/msb0010-0766.pdf>

**Data analysis log**

STEP1

1. Picked top several species in the abundance table that is most related to CRC vs control
2. Change all probablilties to ratio\*100
3. Yielded 15 features (WARNING: a few samples have 0s for all)
4. ## Replace control = 0, CRC = 1,

## Replace species = sp1, sp2, sp3

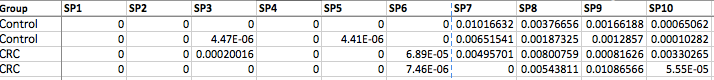
## Deleted the last two rows of CRC (1) and control (0) samples for validation purposes

1. 100\*Cylindrospermopsis raciborskii is 0 on average, get rid (SP6) in the species\_abund\_table

STEP1 redo

1. Pick the top 10 species contributing to the model according to (Zeller et al., 2014) Name SP1-SP10
2. Rank sum of species abundance, pick the sums >1.00 species SP11-SP27
3. Saved to Pop\_F\_data\_summary.xlsx -> 27Species\_new\_list

The four samples taken out



1-88 control 6 flase positive

89-141 CRC (53) 19 false negative

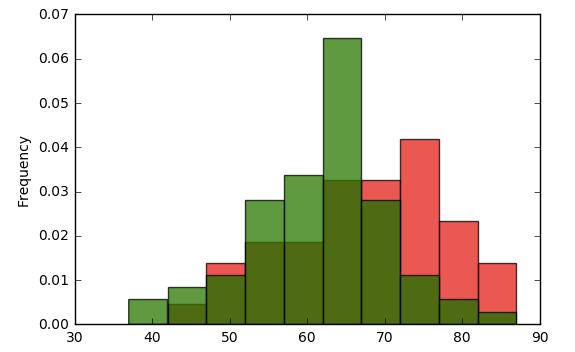
Logistics regression with only 10 species taken into consideration

mean (Pred2==True) = 0.8865

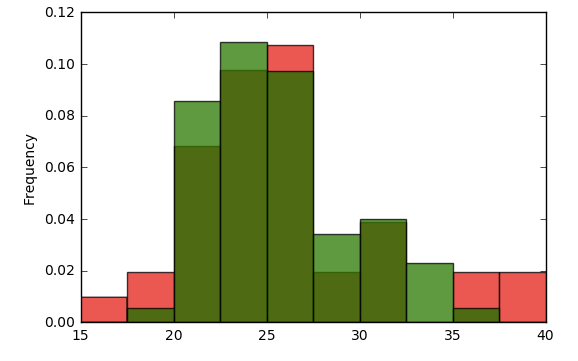
Logistics regression with 27 species taken into consideration

Mean(Pred4==True) = 0.8865248

No difference between the 10 species model and the 27 species model



Age, CRC = red, Control = green



Sensitivity vs specificity (performance of a binary classification test, classification function)

Sensitivity (true positive rate): the proportion fo positives that are correctly identified

Specificity (true negative rate)

A fecal occult blood test (FOBT) checks for hidden (occult) blood in the stool (feces).

https://github.com/lordluen

<https://repo.continuum.io/archive/Anaconda2-4.3.0-Linux-x86_64.sh>

**File upload with html**

**http://www.instantshift.com/2015/10/27/free-html5-css3-jquery-upload-file-forms/**

Template 1: <http://codepen.io/rajeshodayanchal/full/prByE>

<https://designmodo.com/create-upload-form/>

<http://codepen.io/jackarmley/pen/edjxB/>

<http://codepen.io/robcopeland/full/nyGCE/>’