Reproduction of "Human Gut Microbiome Viewed Across Age and Geography"

Hermann Pauly

September 12, 2014

Abstract

In 2012, Yatsunenko et al. published their article "Human Gut Microbiome Viewed Across Age and Geography" in Nature magazine. They had sequenced and investigated the microbial content of fecal samples from 528 humans of different ages from different countries. They found signature patterns for western and non-western samples and differences for age groups in composition, taxonomy and metabolic functions of synthesised proteins. Here I worked on their data to reproduce their findings. I downloaded the data from public servers, followed the descriptions given in the methods supply paper, and compared my graphs and results with the published ones. After manual rebuilding of links between sample information and sequence data, I recreated most of the main results. Due to missing identification data, I was not able to reproduce the protein analysis.

Introduction

All higher creatures' bodies are populated by microorganisms. These microorganisms, beneficial symbionts and pathogenes alike, outnumber human body cells by one magnitude. A community of such microorganisms on a specified location is called a microbiome. There are a number of microbiomes on and inside the human body, for example the communites of skin, lung or mouth bacteriae. Not only do those organisms impact their host's individual health and influence its metabolism, they also provide a source for genetical information.
Yatsunenku et al. gathered and examined genomic data of microbiomes in 528 fecal samples from human participants hailing from urban areas in the USA and rural areas in Malawi (southeast Africa), and from indigenous people from the Amazonas region of Venezuela.
They acquired the taxonomies of microorganisms by sequencing only a small region, the V4 hypervariable region, of the ribosomal 16S rRNA found in the

feces. This V4 region is diverse enough to allow species identification and can be located easily because it is surrounded by "constant regions", which are strongly conserved across species. The authors used the taxonomic data to analyse the microbiome composition (relative abundances of taxa) and species richness in each sample.
They further pyrosequenced the whole gene content of 110 of the provided samples and analysed the biological functions of encoded proteins in the individual samples.
Using ANOVA post hoc testing, the authors found significant differences among samples from different ages, geographic locations and families (see [1]).
Here I work on their data to recreate calculations, data processing, and data visualisation following the descriptions given in the original paper to see if their findings and conclusions can be reproduced knowing nothing more than the published sequencer data.

Methods and results

Data aquisition

I downloaded the raw data and metadata files from MG-RAST[2], project ID numbers 98 (whole genome shotgun sequences) and 401 (16S rRNA V4 sequences), using the UNIX command line tool *wget*. Each project includes a Microsoft *.docx* spreadsheet, containing additional information (metadata) for each sample, including participant's country of origin, family and age. There are subfolders for each sample, containing the *FASTA* sequence data as well as *.stats* files with information about the sample's microbiome, which come from an analysis that is done automatically when sequence data is uploaded to MG-RAST. This information includes taxonomy of found bactererial species, their overall metabolic function if known, and counts of their appearance for each sample.
For the following analyses, only the metadata and *FASTA* files were used.

Preprocessing

Mapping sample IDs, sequence information, and metadata

To be able to group and compare samples by their host's properties (see 3.2.3 for a list of available metadata), a link between sequenced sample files and their respective metadata is required. The MG-RAST database provides a mapping file between sample data files and unique IDs with incomplete metadata, while Yatsunenko et al. provide a complete metadata file without mapping to the sample data files. I applied the "calc" module of the *LibreOffice* suite to manually combine the contents of both files to a complete mapping file. For this I sorted

both provided files' contents by their sample IDs and copy-pasted missing rows from the MG-RAST mapping file into the metadata table. Two ID entries from the mapping file were not found in the metadata table, what means that the corresponding samples' sequence information was not uploaded to MG-RAST. This conforms to the paper, where Yatsunenko et al. state that two samples could not be used in the analysis without giving reasons or sample IDs. Assuming they referred to the same two samples I removed these IDs from the metadata-mapping file. The result was saved as a tab-separated .csv file. On my system with German environment settings I had to convert decimal values from comma-separated values to international dot-separated format using the command line tool *sed*:

```
sed 's/,/\./g' 16s_mapping.csv > 16s_mapping_decimaldot.csv
```

The unique sequencer IDs differed from the sample IDs only by a numerical appendix, so I used the custom Python script *checkSampleIDmapping.py* to check if all samples were matched correctly. Calling the script produced the following output:

```
checking 16s_mapping_decimaldot.csv --> read 528 samples, 0 errors
checking wgs_mapping_decimaldot.csv --> read 110 samples, 0 errors
```

OTUs, α- and β diversity measures

As there are still many unknow microorganism species, a complete species-level taxonomy can not always be obtained. Therefore Yatsunenko et al. followed the QIIME standard protocol [3], assigning all organisms with at least 97% sequence identity to the same operational taxonomic unit (OTU) instead of trying to determine all species exactly.

To compare samples, some kind of measure must be available. They then applied α (intra-sample) and β (inter-sample) diversity as measures. To calculate α-diversity in each sample they used the relative abundances of OTUs, for pairwise β-diversity they applied the UniFrac measure $1 - f$, where f is the fraction of the phylogeny tree shared by all organisms in both samples.

Using the supported data and the manually created metadata mapping table, I followed the QIIME workflow to create tables of OTU abundances and analyse them for α- and β diversity.

The first step required is to categorize the sequenced microorganisms into groups by phylogenetic distance. This step is called "picking". To do this, I used the QIIME tool *pick_closed_reference.py*, which compares the samples to an existing similarity tree. As picking reference, Yatsunenko et al. used the GreenGenes database from 2011-02-04, which supplies a phylogenetical tree

for microorganisms [4]. The picking was done on each sample individually, ten samples in parallel at any time, by using the custom Python script *picking.py*.

The QIIME tools for α-rarefaction and β-diversity-analysis require a single biom table as input file, so I combined the picking results. A try to combine all 528 sample tables exceeded the computational power available, so I used the custom Python script *combine.py* which creates a shell script that repeatedly calls the QIIME tool *merge_otu_tables.py* to iteratively add sample tables to a common table. The QIIME command *biom* was applied to convert a copy of this table to a tab-separated file (`combined.csv`) for easy use with `R`.

The following steps were all done with QIIME tools. To assess species richness from the samples, the "rarefaction" technique is used. A given population (here: microbiome inside a fecal sample) is subsampled to calculate the overall species richness in the population while keeping the sample size as small as possible while still acquiring representative read samples. I used *alpha_rarefaction.py* to create an overview of suitable rarefaction depths.
The read numbers per fecal sample ranged from 305,631 to 5,826,936, with a mean of 1,932,291 and a median of 1,884,081. To guarantee that all samples are represented, and to avoid that small samples would supply all reads available, I subsampled 290,603 reads from each sample, the same rarefaction depth as was used in the original publication. I repeated the rarefactioning ten times, using the tool *multiple_rarefaction_even_depth.py*.

From those rarefied OTU tables I calculated the α- and β-diversity. I calculated α-diversity using the tool *alpha_diversity.py* with the number of observed species as a disance measure and merged the results into a single table (`observed_species.csv`) with the tool *collate_alpha.py*. To measure β-diversity I used UniFrac distance, the percentage of shared branches on the phylogenetic tree, compared against the GreenGenes reference tree. With the tool *beta_diversity.py* I calculated β-diversity distance matrices for all ten rarefied tables, each with weighted and unweighted UniFrac distances as a measure. The resulting files are called `(unweighted_)unifrac_rarefaction_290000_N.txt`.

Accessing the data from `R`

I loaded the data files created by the previous steps 3.2.1 and 3.2.2 and programmed methods to access elements of the data conveniently.

```
theCountries <- c("Malawi", "USA", "Venezuela")
theColours <- c("red", "blue", "green")
names(theColours) <- theCountries
alphaTable <- read.delim("local_copy/observed_species.csv")
betaTable <- read.delim("local_copy/unweighted_unifrac_rarefaction_290000_1.txt")
rownames(betaTable) <- betaTable[,1]; betaTable <- betaTable[,-1]
theMetadata <- read.delim("16s_mapping_decimaldot.csv")
```

```
 # sort tables by id-string-heads to make up for inconsistencies in id-string-tails
theMetadata <- theMetadata[order(rownames(theMetadata)),]
beta.order <- order(colnames(betaTable))
betaTable <- betaTable[beta.order,beta.order]
rownames(theMetadata) <- theMetadata$X.SampleID
 # get sample ids with label of specific value
getGroupIDs <- function(label, value) {
  theMetadata[theMetadata[,label] %in% value,]$X.SampleID
}
 # get sample ids with label in numeric range
getRangeIDs <- function(label, lower, upper) {
  values <- theMetadata[,label]
  theMetadata[values >= lower & values <= upper,]$X.SampleID
}
 # get beta variance data for samples with given ids
getBetaGroup <- function(label=NULL, value=NULL, range=FALSE, ids=NULL) {
  if (length(ids) == 0) {
    if (any(is.na(c(label,value)))) { # no input at all
      print("either label and value or ids required")
      return(c())
    }
    if (!range) ids <- getGroupIDs(label, value)
    else ids <- getRangeIDs(label, value[1], value[2])
  }
  nums <- which(rownames(theMetadata) %in% ids)
  result <- betaTable[nums, nums]
  rownames(result) <- colnames(result)
  result
}
```

Differences decrease with age

Yatsunenko et al. observed a change of microbiome composition from infant-specific to adult configuration by comparing the composition of each child's microbiome against the microbiome composition of all adults from the same country. As it is not completely clear how the distance to all adults was calculated I used the mean of distances of the child to each adult.

```
adultNames <- setdiff( rownames(theMetadata), getRangeIDs("Age", 0, 18) )
plot(NULL, NULL, xlim=c(0, 18), ylim=c(0.35, 0.85), xlab="age", ylab="UniFrac distance")
country <- list(
    "USA" = getGroupIDs("Country", "USA"),
    "Malawi" = getGroupIDs("Country", "Malawi"),
    "Venezuela" = getGroupIDs("Country", "Venezuela")
    )
```

```
adults <- list(
    "USA" = intersect(country[["USA"]], adultNames),
    "Malawi" = intersect(country[["Malawi"]], adultNames),
    "Venezuela" = intersect(country[["Venezuela"]], adultNames)
    )
children <- list(
    "USA" = intersect(country[["USA"]], getRangeIDs("Age", 0, 18)),
    "Malawi" = intersect(country[["Malawi"]], getRangeIDs("Age", 0, 18)),
    "Venezuela" = intersect(country[["Venezuela"]], getRangeIDs("Age", 0, 18))
    )
for (country in theCountries) {
    for (child in children[[country]]) {
        betadiv <- getBetaGroup(ids=c(child, adults[[country]]))
        x <- theMetadata[theMetadata$X.SampleID==child,]$Age
        y <- sum(betadiv[child,]) / (nrow(betadiv) - 1)
        points(x, y, col=theColours[[country]], pch=20)
    }
}
legend(x=14, y=1.0, legend=theCountries, text.col=theColours)
```

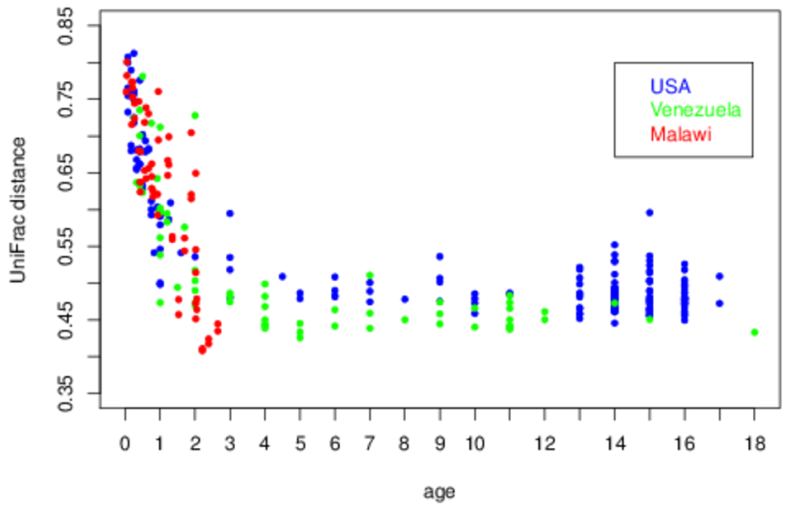

Figure 1: The UniFrac distances start with high values at early ages, show a strong decline until approximately three years of age, and stay steadily low until adulthood.

PCoA analysis of β-diversity

Partitioning Around Medoids (PAM) is a method of clustering data points into a given number k of groups. It initially assigns a data point as center for each of the groups and then minimises the sum of distances from all points to their nearest center by iteratively swapping points and centers and reassigning all data points to the nearest new center. Yatsunenko et al. used the β-distance matrix as a one-dimensional dissimilarity measures for the PAM algorithm and chose k=3 clusters to refind the samples' countries of origin in the microbiome composition of adults.

I used the *pam* function from the `R` package *cluster* [5] to repeat the process.

```
adultNames <- setdiff( rownames(theMetadata), getRangeIDs("Age", 0, 18) )
betaAdults <- getBetaGroup(ids=adultNames)
betaChldrn <- getBetaGroup("Age", c(0, 18), range=TRUE)
clu <- pam(betaAdults, diss=TRUE, k=3, keep.diss=TRUE)
clu2 <- pam(betaChldrn, diss=TRUE, k=3, keep.diss=TRUE)
clusplot(clu, col.p=c("red","blue","green")[theMetadata[names(clu$clustering),]$Country], ma
legend(x=-0.48, y=0.3, legend=theCountries, text.col=theColours)


 # note: color order is different, because the pamobject$clustering vector has a different
contTable <- table(clu$clustering,
theMetadata[names(clu$clustering),]$Country)
 # correct order
contTable <- cbind(contTable[,3], contTable[,1], contTable[,2])
pam.result <- sum(diag(contTable) / sum(contTable))
print(pam.result)
```

```
## [1] 0.8431
```

I also clustered the children's distance measures with the same method.

SVM classificator analysis of β-diversity

The PCoA plot suggested that discrimination of samples by microbiome diversity is possible. To improve the assignment I used the implementation of support vector machines (SVM) in the `R` package *e1071* [6]. Support vector machines classify multi-dimensional datasets by finding a hyperplane that separates the classes among (a subset of) their features. They solve the problem of seemingly inseperable datasets by applying a "kernel function" that adds additional dimensions to the data's used features.

I trained a SVM with default radial kernel function to discriminate the dataset by different subsets of the OTUs that showed the greatest variance and compared the predicted samples with their countries of origin, using 20-fold cross validation.

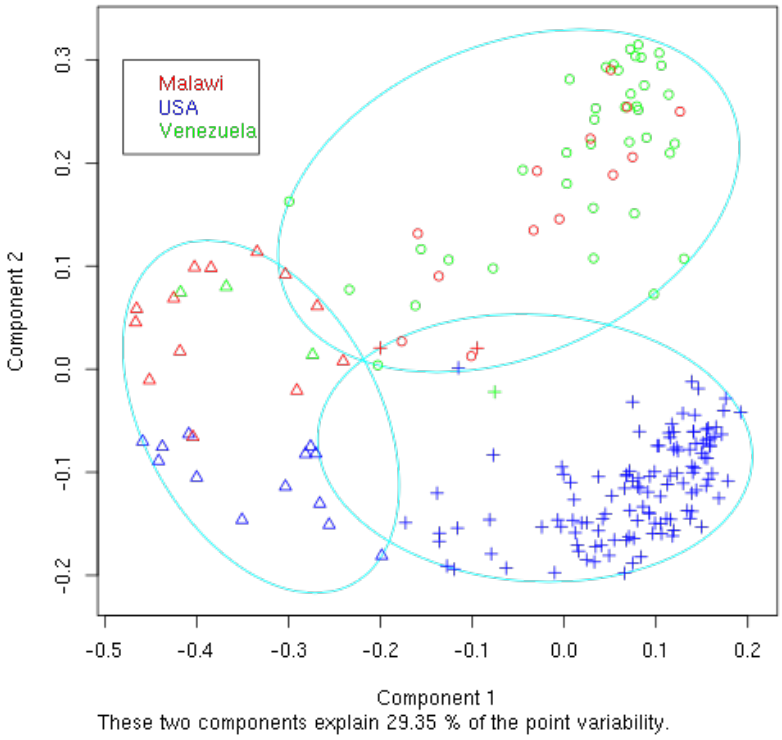

Figure 2: PCoA of beta-diversity among adults. Western (US) Samples can be separated linearly with five errors.

```
require(e1071)
species <- read.delim("local_copy/combined.csv", sep="\t", header=TRUE)
rownames(species) <- species[,1]
species <- species[,-1] # species level OTU table
species <- t(species)   # now rows = samples, cols = OTU counts
species <- species[order(colnames(species)),] # sort by id
speciesAdult <- species[which(rownames(theMetadata) %in% adultNames),] # select adult sample
reps <- 20              # number of repetitions for cross validation
L <- nrow(speciesAdult)
N <- L / reps           # number of samples per validation run
featureSizes <- c(1:5, 10, 20, 30, 50, 90, seq(100, 1000, len=10), ncol(speciesAdult)) #
strength <- matrix(0, nrow=reps, ncol=length(featureSizes))
for (j in 1:length(featureSizes)) {
    allOfThem <- 1:L    # available samples
    for (i in 1:reps) {
```

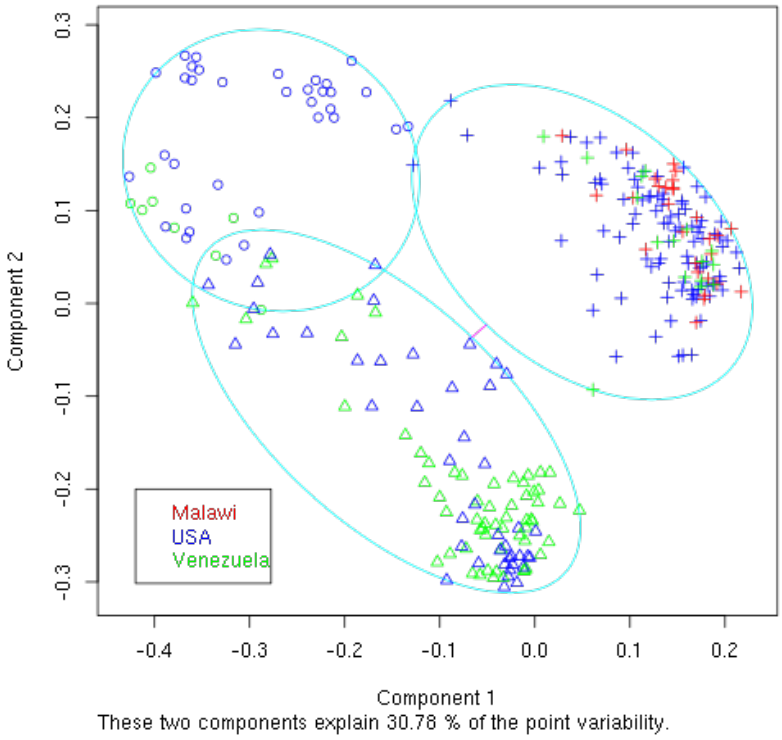

Figure 3: PCoA of beta-diversity among children shows no clear grouping.

```
        testSet <- sample(allOfThem, N)
        allOfThem <- allOfThem[-testSet]
        training <- c(1:L)[-testSet]
        variances <- apply(speciesAdult[training,], 2, var)
        topVariables <- order(variances, decreasing=TRUE)
        topVariables <- topVariables[1:featureSizes[j]]
        trnNames <- rownames(speciesAdult)[training]
        tstNames <- rownames(speciesAdult)[testSet]
        model <- svm(x=speciesAdult[training,topVariables], y=theMetadata[trnNames,]$Country
        prediction <- predict(model, speciesAdult[testSet,topVariables])
        contingency <- table(prediction, theMetadata[tstNames,]$Country)
        strength[i,j] <- sum(diag(contingency)) / sum(contingency)
    }
}
boxplot(strength, xlab="number of features", ylab="prediction success", axes=FALSE)
abline(h=pam.result, col="blue")
```

```
axis(side=1, at=1:length(featureSizes), labels=as.integer(featureSizes), las=2)
axis(side=2, at=seq(from=0, to=1.2, by=0.2))
```

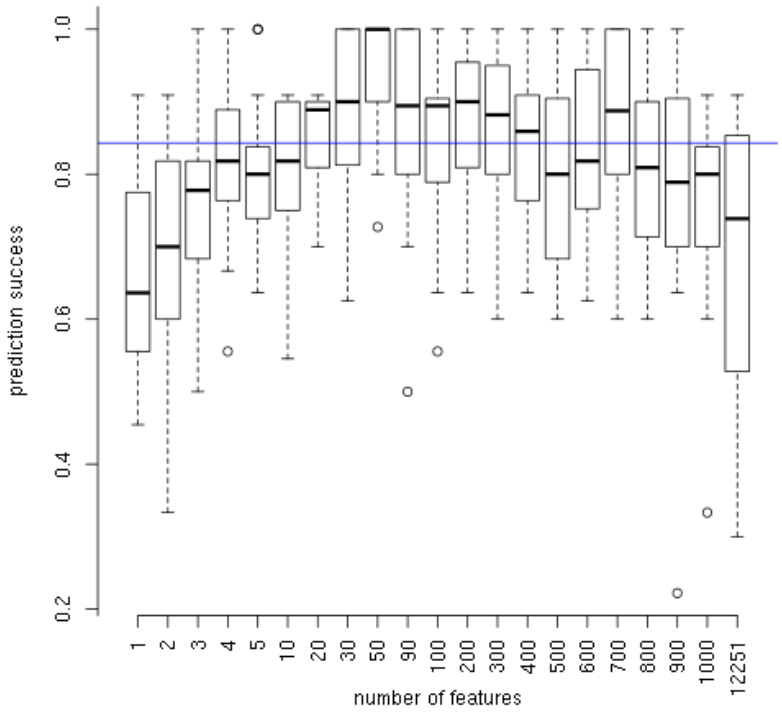

Figure 4: Barplots of SVM-classifier prediction of the 16S dataset. Abscissa describes the number of features used. The features were pre-selected by greatest variance. A blue horizontal line shows the prediction success of PAM clustering for comparison.

Microbiomes get more diverse with age

Yatsunenko et al. found that the number of OTUs inside the fecal samples increased with age. To verify this I calculated the means of OTU counts found in the ten rarefaction repetitions for each sample and plotted them against each samples' age.

```
rarefaction <- 188517
alpha <-alphaTable[alphaTable$sequences.per.sample==rarefaction,][,4:ncol(alphaTable)]
```

```
counts <- colMeans(data.matrix(alpha))
names(counts) <- colnames(alpha)
x <- theMetadata[names(counts),]$Age
cols <- c("red", "blue", "green")[theMetadata[names(counts),]$Country]
plot(x, counts, col=cols, pch=20, ylab="Number of OTUs", xlab="Age", lab=c(20, 6, 7))
legend(x=55, y=500, legend=theCountries, text.col=theColours)
```

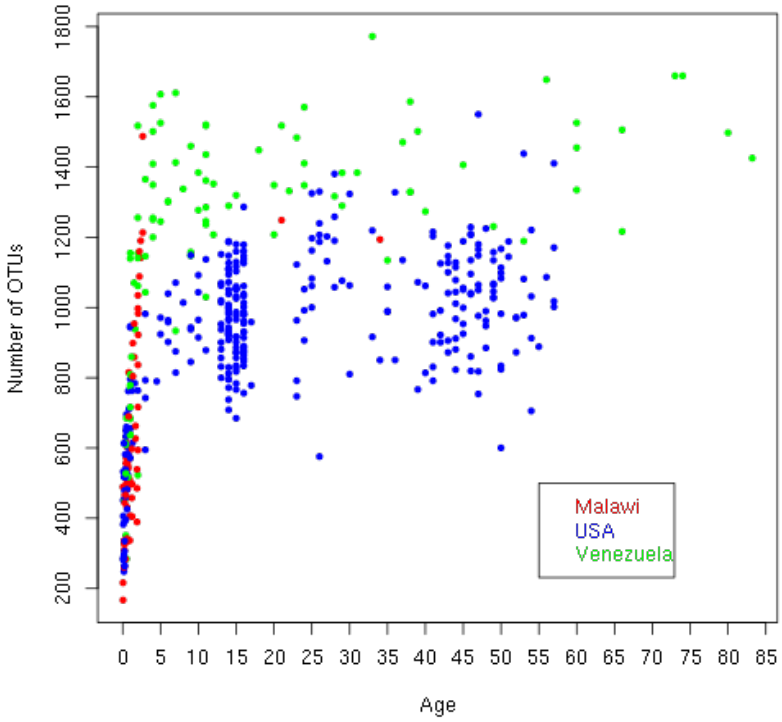

Figure 5: The number of OTUs per sample increases with age. Missing entries due to NA values for age (abcisssa) of most African participants.

Processing of whole genome shotgun sequence data

Yatsunenko et al. sequenced the whole gene sequences (WGS) of 110 samples out of the 528 that were taken, without making clear if there was a method of choosing or if the 110 samples were picked at random. They compared the proteins encoded in the microbiotas' sequences from these samples against the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Clusters of Orthologous Groups (COG)

protein databases to analyse the encoded proteins for their functions and found distinct patterns that distinguished microbiomes from US citizens from non-US citizens concerning metabolic functions.

Before the raw sequences can be matched with databases, they need to be prepared. Genomic information from human cells must be removed as well as sequences containing probable sequencing errors and duplicate sequences. The quality filtering was done using the following rules: (1) a read must be longer than 59 bases (2) a read must not contain more than two base positions given as N or any NN pair (3) a read must not be identical in more than 97% with any other read in the sample. If any of the rules is violated, the read is discarded. Human DNA was identified by BLASTing the sequences against human genome databases. The authors stated that "[this] preprocessing was done using custom Perl scripts and publicly available software tools".
The custom scripts were not available for download, the tools were not named, and an e-mail to the corresponding author concerning the methods was not answered.

I thererefore created a Python script *faster_filter.py* to fulfill the filtering rules, but even with the naive approach at finding duplicates, the computational time required for a single sample was exceptionally high, so it served only as a proof of concept.
The authors did not write which, if any, further preprocessing steps they took. The metadata file on MG-RAST contains information to link the WGS sequences to the IDs of the 16S rRNA reads used in analyses 3.2.1 to 3.2.7 and to the respective metadata, but the IDs given in the WGS data differ from those expected by the metadata spreadsheet.

Thus it was not possible for me to reconstruct any of the WGS heatmaps and clusters given in the original publication.

Software used

The metadata mapping files were created using *LibreOffice* software suite version 4.2.4.2. The QIIME workflow was followed using the *QIIME* suite of software tools, version 1.8.0 as described by Caporaso et al. [3]. Data analysis and visualisation was done using the `R` statistical sofware version 3.1.0 [7]. The custom Python scripts mentioned in the scope of this document can be found on my github page (https://github.com/hermann-p/yatsunenko-2012-microbiome).

Discussion

Using data from Yatsunenko et al. I tried to reproduce results they published in Nature magazine 2012. They searched for patterns inside human gut microbiomes found in feces. Their work contains results from two datasets, a set of taxonomic data obtained from sampling a region of the 16S rRNA in the samples and a set of proteine data from sequencing all non-human DNA inside the samples. In addition to the main publication, further results and supporting information, that was helpful in the reproduction, are provided online.

The analysis of 16S data was very well documented. After recreating the metadata file (3.2.1), the graphs and results deduced from the 16S data (3.2.4 - 3.2.7) were virtually identical to those published by Yatsunenko et al., in the limits of randomness involved in the rarefactioning step.
The conclusions drawn from those results were plausible. The PAM clustering of adults' microbiomes to distinguish western (US) samples from non-US samples turned out to be inferior to prediction by SVM (3.2.6).
One weakness in reproducability was the fact, that no complete usable metadatafile is provided. Rebuilding a working one by hand was possible with the information provided, but it is error prone and all further analysis depended on the connection of samples to their metadata. When using participants' ages as plotting coordinates, it became apparent that age information for the Malawian population was incomplete. 30 of 107 Malawian samples had no age information with them, thus the visualisation differed from the published images. This leads to the conclusion, that metadata was manually edited or transferred to the MG-RAST servers after publication.

On contrast to the 16S data analysis, the steps for the WGS analysis were only vaguely documented. Custom scripts were not provided, applied tools were not named. In addition, the genomic data cannot be mapped to the samples, making them useless for reproduction efforts. Nevertheless, although the WGS analysis is a work- and computationally intensive part, it only provided one of the many main results.

In conclusion it can be said, that the work had a high degree of reproducability. For all analysis that was done, the results and images of Yatsunenko et al. could

be recreated. Any problems that occured originated in the way the data was published.

References

[1]T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, J. I. Gordon, *Human Gut Microbiome Viewed Across Age and Geography*, Nature, **2012**.

[2]MG-RAST, metagenomics analysis server, *http://metagenomics.anl.gov*.

[3]J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, *QIIME Allows Analysis of High-Throughput Community Sequencing Data*, Nature Methods, **2010**.

[4]The GreenGeenes Database Consortium, The GreenGenes Database, *http://greengenes.secondgenome.com*.

[5]M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik, in cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4, **2013**.

[6]D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, in e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-1, **2012**.

[7]R Core Team, in R: A Language and Environment for Statistical Computing v3.0.2; R Foundation for Statistical Computing, http://www.R-project.org/, **2013**.