Research Article
Open Access
CrossMark

Comparative metagenomics with metabolic reconstruction

Abstract

Human microbiome communities consist of variety of bacterial, fungi and archaea, which are integral part of the human body. These communities greatly vary from one part of the body to the other and help us maintain healthy environment. Most microbiome interacts with the host and each other via metabolic products, but how it interacts with various human metabolic pathways still remains unknown. Slight changes in the microbial communities could be important indicator of potential disease and can be served as potential Biomarker. For this study two separate specimens of human saliva microbial DNA short read sequences retrieved from the HMP (Human Metabiome Project).[1] After appropriate quality control, both sequences aligned against the KEGG database for identification of genes and known metabolic pathways these communities were involved. The Human Metabolic Reconstruction pipeline[2] was primarily developed to analytical process to perform metabolic reconstruction of pathways involved in human microbiome. Further information on statistical significance on each metabolic pathway analyzed and compared against both control and test Specimens. Finally 3M[3] comparative visual analytics and manual curation capabilities were developed using Oracle Apex rapid web development technology.[4] As a conclusion of this case study various surprising facts uncovered between both specimens. With the help of KEGG BRITE Metabolic Hierarchy[5] pathways abundance/coverage with Orthology, Enzyme and chemical function visualized more effectively. Such type of comparative metagenomic studies performed on large pool of patient cohort can be beneficial to discover effective biomarker for the diagnosis and prognosis of various diseases..

Keywords: metagenomics, human, disease, HUMAnN, pipeline, visualization, annotation, microbial community, metabolic, reconstruction, kegg pathways

Volume 2 Issue 3 - 2015

Nilesh Saratkar
Department of Informatics Analytics & Business Intelligence, Quest Diagnostics Inc, USA

Correspondence: Nilesh Saratkar, Department of Informatics Analytics & Business Intelligence, Quest Diagnostics Inc. 1290 Wall Street, Lyndhurst – NJ, USA, Tel +19739302015, Email Nilesh.S.Saratkar@QuestDiagnostics.com

Received: February 24, 2014 | **Published:** June 10, 2015

Abbreviations: KEGG, kyoto encyclopedia of genes and genomes; HMP, human metabiome project; MG-RAST, metagenomics RAST); 3M, metagenomic metabolic manual annotation; DIAG, data intensive academic grid; HUMAnN, HMP unified metabolic analysis network

Introduction

Human mouth is the first place to receive food and saliva, hence it's potential target for external and internal microbial communities. All of these microbial communities compete for resources and survival. Exact nature of these communities and their interaction with human system related to diseases progression is still not completely understood. In search of answers to various questions such as "Who they are?" and "What they do?", metagenomic comparative analysis were performed. Two separate human specimens[6,7] of human saliva microbial short reads sequences retrieved from the HMP (Human Metabiome Project).[1] Building a generic metagenomic pipeline for the pathway data analysis as well as data visualization platform was the major objective of this study. Understanding functional nature of microbial communities such as "How abundant they are?" and "What they are capable of doing?" will be essential for the conclusion of this study.

With the reducing cost of Next Generation Sequencing, now it's possible to perform cost effective sequencing and analyze entire site specific microbial communities at once. Metabolic Reconstruction[2] is the NGS sequence based computational process for analysis of metabolic pathways and interactions within these localized communities. Reconstruction generates statistical significance of abundance and coverage of various metabolic pathways involved. Without actual information on RNA gene expression, it will be difficult to identify actual pathways involved, but using metagenomic sequence it is definitely possible to understand capability of these microbial communities. Distantly related orthologs species rarely demonstrates sequence identity, thus accuracy of alignment improved by NCBI blastxtool on KEGG (genes.pep) protein reference database.[5] With known functionality of involved genes and related annotation from KEGG, further investigation performed to uncover enzymes and related chemical functionalities.

Materials and methods

Both HMP specimen sequences show in the Table 1 have demonstrated similar species abundance, which helped eliminate most common features and focus on critical variations' within the pathways. NGS pipeline consists of six sections including sequence acquisition, cleansing, alignment, reconstruction, visualization and annotation. As a result of reconstruction process, the HuMAnN pipeline[8] provided pathway abundance score and pathway coverage score for differential analysis. Both statistical indicators further utilized for visual analytics. Most backend processing including HMP sequence Acquisition, KEGG BRITE Hierarchy Alignment and HUNAnN Reconstruction is done within the DIAG computing platform,[2] while storage, visualization and annotation accomplished with the customized 3M Application hosted on Oracle Apex portal[4] (Figure 1).

Table 1 Sequence data volume

HMP Specimen#	HMP Sequence	Trimmed Filtered	Blast Input	Blast output	HUMAnN Output
4473347	5005604	4053766	440500	50875	355
4473378	8351741	4397314	675002	79669	253

Figure 1 End to end data processing flow.

Quality control

Existence of incomplete and noisy NGS sequences within the specimens will have adverse impact on data analysis and final interpretations.[8] Hence short read sequences from FASTQ format were trimmed for repeated and low quality reads using **TrimBWAstyle.pl**.[9,10] Incomplete sequence filtered out based up on length with the help of **remove_bad_seqs.py**.[10] Qualified sequences with length greater than 75bp were selected for further analysis. As a result of this rigorous cleansing operation around 81% short reads sequences were selected from the Specimen#1, while only 53% of the short reads from the Specimen#2 were retained.

Alignment

The KEGG database[5] is the unique form of pre-curated database and provides identification of gene and related organism. This linkage was very critical for orthologous function, enzymatic reaction studied during the metabolic reconstruction. Hence HUMAnN pipeline required blast alignment for the orthologs against annotated genome from the KEGG genome. pepprotein database.[5] The blastx hits on the KEGG reference database provided identity/similarity score along with the Organism and gene identifiers for the further metabolic reconstruction. The reconstruction also required aligned data in the tab delimited format, hence blastx parameters were tuned accordingly. As a result of quality control and alignment with the KEGG database, overall sequence volume further reduced significantly.

Reconstruction

The HUMAnN (HMP Unified Metabolic Analysis Network) pipeline[11] is designed to identify metabolic pathway or module abundance (presence/absence) and their relative metabolic pathway coverage. Typically reconstruction begins with the aligned gene identifiers and achieve objective with following five steps. Visit http://huttenhower.sph.harvard.edu/humann for the additional details on HUMAnN pipeline.[2]

i. Reads weighted for the quality of the matches to calculate abundance of orthologs gene family.

ii. The MinPath algorithm is used to identify gene family to the metabolic pathway and modules.

iii. Based up on taxonomic composition filter out false positive pathways identified by MinPath.

iv. Fill up the gap in the pathway produced due to sequencing error or low abundance gene.

Finally assign coverage score and abundance score to the metabolic pathway and module.

Pathway Hierarchy

Interpretation of metabolic pathway functionality is not possible without integrating reconstruction data with pathway hierarchy retrieved from KEGG Brite.[5] Hence curated information including pathway ID (koID), modules, orthologues (KID) and enzyme (EC#) also uploaded into the 3M Visualization Applicatio.[3] The 3M application also enhanced to customize pathway hierarchy as needed. Depend up on the experiment requirement the reconstruction step can be repeated for each Specimens such as control or test and visually analyze pathways involved in relative significance (Figure 2).

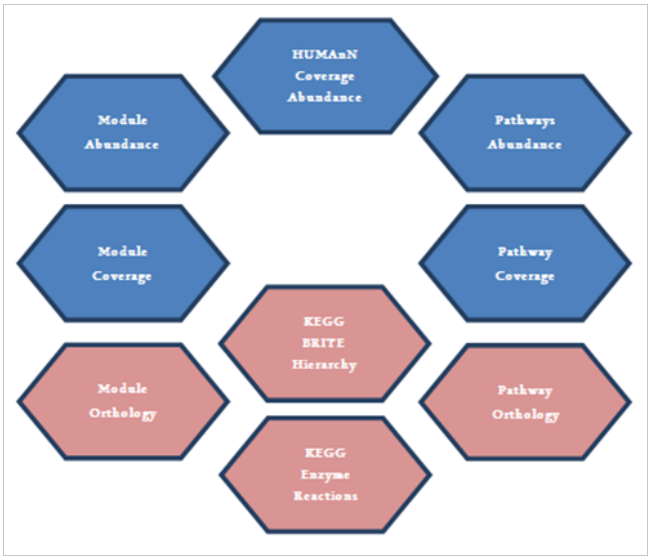


Figure 2 End to end data processing flow.

Visualization

Actual comparative metagenomics visualization process began after uploading both specimens information into the custom built Metagenomic Metabolic Manual (3M) Annotation Application.[3] This web application can be accessed with (dev/dev) credential at the Apex

Citation: Saratkar N. Comparative metagenomics with metabolic reconstruction. *MOJ Proteomics Bioinform.* 2015;2(3):95–100.
DOI: 10.15406/mojpb.2015.02.00049

URL http://apex.oracle.com/pls/apex/f?p=689131. For cost effective solution and ease of maintenance visualization and annotation screens were developed based up on Oracle Apex[4] rapid development framework. Access to the Apex web development framework is free for the non-production application like 3M. Since this web development technology is also included within the actual Oracle database license, it will be sustainable cost effective alternative for the web development.

Results

With the help of pathway significance and curated KEGG pathway hierarchy, now user can perform data visualization from various aspects. Following are some of the observations identified during visualization of metagenomic specimen received after reconstruction. For simplicity user selection of control and test specimen data for visualization will be retained during the active user session.

Top 10 Pathways

Visualization of top 10 pathways from both Specimens indicated identical pathway abundance, except "One carbon pool by folate" on Specimen1 while "DNA Replication" on Specimen2 was major differences among both specimens.

Sided by side comparison

As per the pathway coverage is concern, gap between both Specimens widens drastically. Especially Sulfur Relay System (ko04122) in Specimen1 appeared as the largest gap in the coverage might explain dominant Sulfur metabolism function of the Specimen1 community.

Highest differences

Significant differences among specimens observed within the C5-Branched dibasic acid metabolism (ko00660) pathway. Similar difference also noticed within the Lipoic acid metabolism (ko00785) pathways, while flagellar assembly (ko02040) and Sulfur relay system (ko04122) pathways were indicator of behavior of the microbial communities.

Least differences

As per the least difference in the pathway coverage, observed 10 pathways without any significant difference. Abundance comparison also found Alzheimer's disease (ko05010) pathways unrelated to the microbial community. During annotation phase, 3M annotation can easily identify and eliminate this pathway. Since both Specimen collected from Human saliva sample[7] possibility of human genomic sequences contamination cannot be ruled out.

Metagenomic Metabolic Manual Annotation(3M)

Welcome: DEV Logo

Visualization | Annotation

Home | Pie Charts | Bar Chart | Diffrence | Compare | Missing | Coorelation | Hierarchy

Sample1 4473347-PathwayAbundance Sample2 4473378-PathwayAbundance

Top Diffrence

PATHWAY	SAMPLE1	SAMPLE2	DIFF
ko00660-C5-Branched dibasic acid metabolism	.0189049	.00922175	.00968315
ko03030-DNA replication	.0161485	.0226766	.0065281
ko00908-Zeatin biosynthesis	.00755702	.0124052	.00484818
ko00900-Terpenoid backbone biosynthesis	.0135747	.0183761	.0048014
ko00473-D-Alanine metabolism	.0227109	.0274287	.0047178
ko03430-Mismatch repair	.0201628	.0247895	.0046267
ko00540-Lipopolysaccharide biosynthesis	.0137939	.0181539	.00436

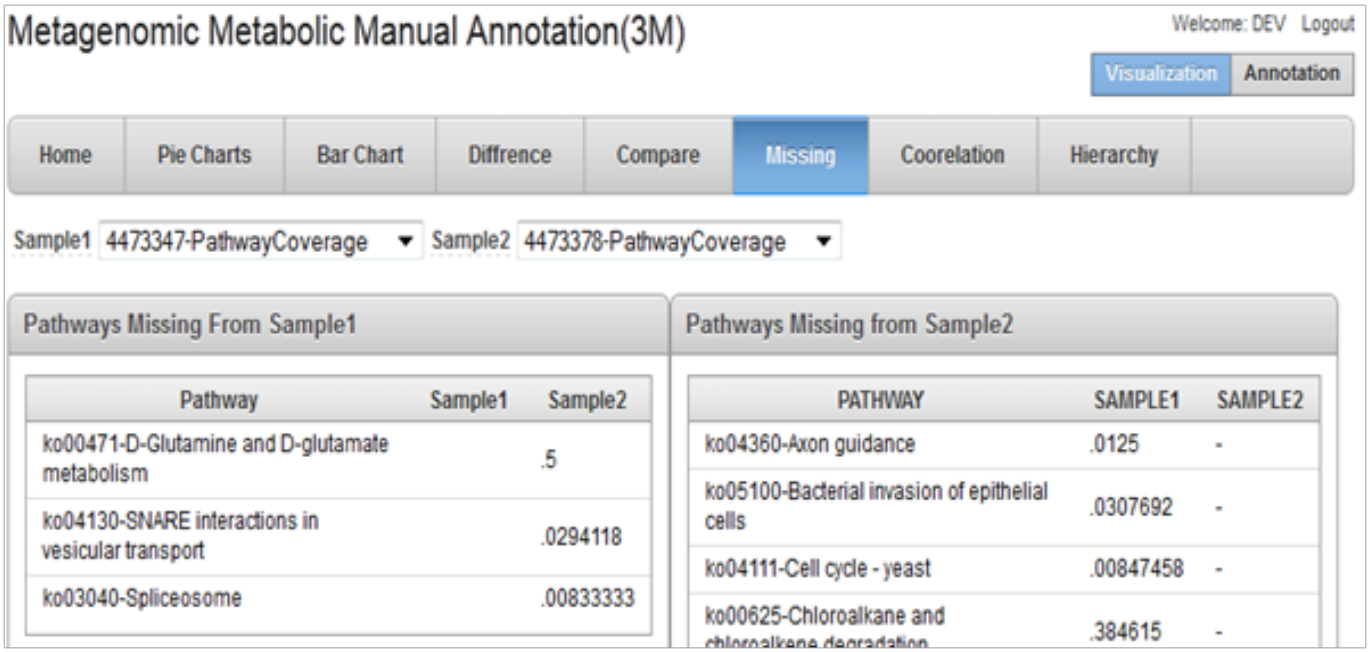
Bottom Diffrence

PATHWAY	SAMPLE1	SAMPLE2	DIFF
ko05145-Toxoplasmosis	.00000061	.00000058	.00000003
ko04610-Complement and coagulation cascades	.00000122	.00000029	.00000093
ko00522-Biosynthesis of 12-, 14- and 16-membered macrolides	.00000142	.00000005	.00000137
ko00622-Xylene degradation	.00000096	.00000285	.00000189
ko04075-Plant hormone signal transduction	.00000294	.00000021	.00000273
ko00402-Benzoxazinoid biosynthesis	.00000406	.00000085	.00000321
ko00940-Phenylpropanoid biosynthesis	.00000251	.00000591	.0000034
ko00140-Steroid hormone biosynthesis	.0000492	.00004496	.00000424

Citation: Saratkar N. Comparative metagenomics with metabolic reconstruction. *MOJ Proteomics Bioinform.* 2015;2(3):95–100.
DOI: 10.15406/mojpb.2015.02.00049

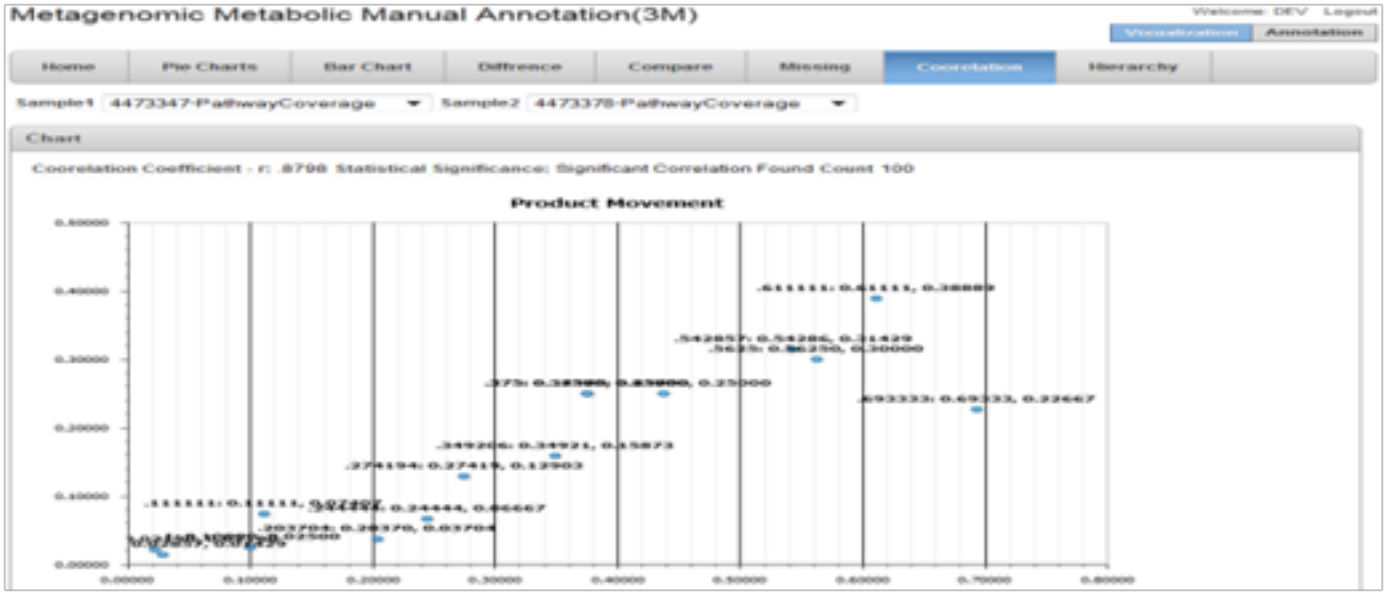
Missing pathways

No missing pathways appeared during abundance analysis from either side of the Specimens. While from pathway coverage point of view, observed three missing pathways from Specimen1 but found in Specimen2, including D-Glutamine and D-glutamate metabolism (ko00471), SNARE interactions in vesicular transport (ko04130) and Spliceosome (ko03040). However there are 37 pathways missing from Specimen2 mostly related to human metabolic pathways.



Correlationship

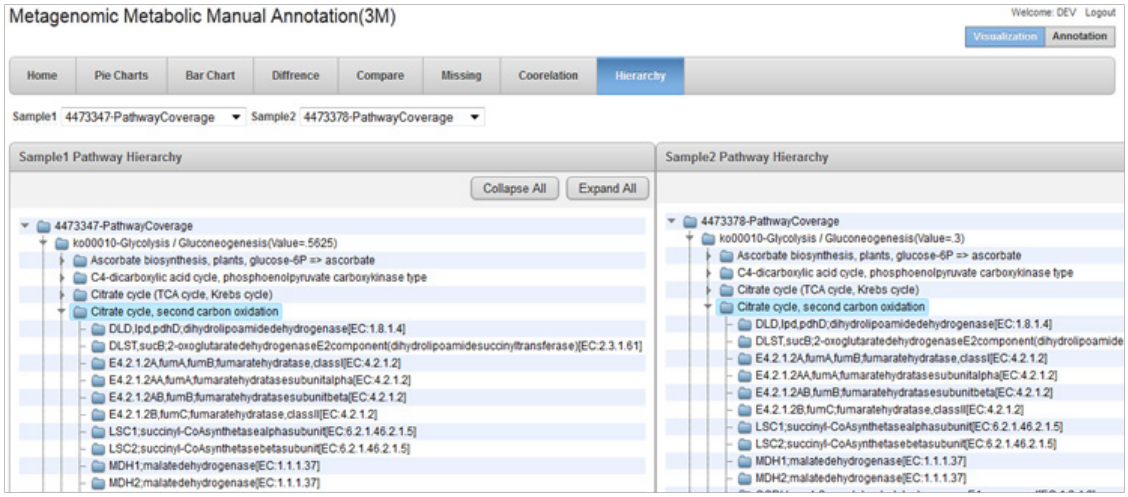
The visualization 3M application also designed to calculate correlation coefficient-r and compare it against t-table for the possible relationship between Specimen1 and Specimen2 scores.[6,7] After establishing relationship between matching pair among 147 metabolic pathways, calculated coefficient r=0.9754 indicating significant correlation between both Specimens pathway abundance. But pathway coverage with 100 metabolic pathways could able to calculate r coefficient of 0.8798, indicating less correlation compared to the pathway abundance.



Pathway hierarchy

Side by side comparison of the pathway abundance also observed existence of significant amount of differences in the pathways among Specimen1 and Specimen2.[6,7] Few noticeable pathways observed within the Specimen1 only related to the Human metabolic pathways. Similar conclusion also interpreted within the Pathway coverage hierarchy for both Specimens.

Citation: Saratkar N. Comparative metagenomics with metabolic reconstruction. *MOJ Proteomics Bioinform.* 2015;2(3):95–100.
DOI: 10.15406/mojpb.2015.02.00049



Manual annotation

The 3M visualization application[3] enhanced to enable user to perform curation on the reconstruction data as well as pathway hierarchy. Hence visualization supported various Annotation capabilities including upload HUMAnN pipeline data, modify, delete, uploaded data, maintain KEGG Pathway related Orthology and maintain relationship between module and related KEGG Orthology and enzyme functions. With the help of "HUMAnN Data" tab, user can able to inactivate specific unrelated pathways abundance and coverage score. Thus user can effectively eliminate any outlier pathways from further analysis.

Table 2 Top 10 Metagenomic species distribution

Species	Specimen1 population SRR062371- 4473347.3	Specimen2 population SRR062402-4473378.3
Firmicutes	2022849	3359405
Proteobacteria	1000866	1077421
Bacteroidetes	658657	2514329
Actinobacteria	534319	265861
Fusobacteria	83655	217717
Chordata	36145	41367
Fibrobacteres	31905	44676
Cyanobacteria	27592	46302
Ascomycota	24399	42809
Spirochaetes	23074	42138

Discussion

Even though both human saliva specimens contain similar representation of the microbial communities, both Specimens are different from the comparative metagenomic potential point of view. However top three species including Firmicutes, Proteobacteria and Bacteriodetes were predominantly found within both specimens. Most of the pathways mentioned earlier in the result section are possible genomic pathways and may not be actually expressed in the microbiome. Only proteomics and transcriptomic studies may able to conclude actual pathway abundance and coverage within cellular environment. This information might help us in the future to understand functional nature of the community and possibly prepare comparative metagenomic profile and possible biomarker indicative of potential diseases such as Periodontal Disease.[12]

Interpretations

Further study of top ten pathways identified abundance of most common pathways such as Ribosome, lipopolysaccharides, Valine, Leuicine, Isoluicine, Thymine, Alanine and Peptidoglycan biosynthesis pathways. Abundance of Sulfur Relay System (ko04122) pathway within Specimen1 also leads us to possible microbial community function of sulfur metabolism, cell proliferation, apoptosis, and DNA repair. Abundance of Flagellar assembly (ko02040) within Specimen1 also indicates that it contains abundant bacterial communities which utilize flagellum for its locomotion. Analysis of the missing pathways remains inconclusive, but it clearly suggested that we did not acquire sufficient amount of Specimen2 short read sequences to perform metabolic reconstruction. Even though pathway abundance correlates nicely (r=0.9754), but pathway coverage remains less correlated (r=0.8798). Surprisingly only three missing pathway coverage from both specimens strongly suggest existence of certain pathways pattern within the human saliva specimens. Such type of pathway profiles on ecosystem can be very useful for clinical diagnosis and prognosis.

Future enhancements

a. NCBI Blastx alignment definitely improved accuracy of alignment, but also introduced performance implications within the reconstruction pipeline. In the future pipeline can be enhanced with bowtie2.

b. Most of the alignment jobs were submitted to DIAG HPC cluster[13] via "qsub" command, In future pipeline throughput can be enhanced by Hadoop distributed data and processing platform.

c. For simplicity, 3M visualization application only supported two specimens (test and control) for the comparative study. In the future 3M application can be enhance to perform multi specimen analysis.

d. The reconstruction pipeline integrates backend (DIAG pipeline) and frontend (Oracle Apex) 3M applications and exchange data via tab delimited files. In the future direct database connectivity to visualization tool will reduce manual overhead.

e. The 3M application can be enhanced to perform series of iterative automated and manual annotation until reaching out to the desired result quality.

Citation: Saratkar N. Comparative metagenomics with metabolic reconstruction. *MOJ Proteomics Bioinform.* 2015;2(3):95–100.
DOI: 10.15406/mojpb.2015.02.00049

Conclusion

Microbial species distribution shown in the Table #2 contains various probiotic species such as Firmicutes, Proteobacteria, Actinobacteria, Bacteroidetes and Fusobacteria. No significant source of pathogenic species such as Streptococcus-pyogenes found within either of the human saliva specimens. Various studies[12] linked periodontal/Gum disease to the Streptococcus-pyogenes bacteria within sub-gingival plaque communities. Invasion of periodontal pathogenic species into the blood stream has been associated with tooth decay, chronic vascular disease and stroke.[12] As periodontal disease progress, patient's treatment options also confines. Thus complicated and costly procedures can be avoided by detecting shift in the microbial community profile early in the disease progression.

During specimen collection phase of Specimen2, HMP could not able to filter out human cells from the microbial cells. Once again this study emphasize on importance of the microbial specimen cell preparation and filtration procedures. Lesson learned and pathway gaps identified from this study can be further curated by 3M Annotation functionality and reprocessed via HUMAnN pipeline. Finally this study concludes with small step towards the potential application of comparative metagenomic to define reference ranges for healthy human subjects. Such type of reference data bases should help researchers distinguish pathogenic state of the human microbial communities. Further research in the functional nature of microbial communities may lead to the future state of diagnosis, prognosis and personalized medicine biomarker.

Acknowledgements

I would like to express sincere gratitude towards the Johns Hopkins Prof. Joshua Orvis, who encourage me to think out of box and take challenging tasks like this one. I am also thankful for the DIAG team for making such a great computational resources available for most of the analytical data processing. Without support from the HUMAnN team at the Hutten hower Lab Department of Biostatistics, Harvard School of Public Health this study will not be possible. Finally special thanks to HMP, MG-RAST, KEGG and Oracle.

Conflict of interest

The author declares no conflict of interest.

References

1. The Human Microbiome Project (HMP) NIH Human Microbiome Project, powered by GOLD.
2. HUMAnN. *(Meta HUMAnN: The HMP Unified Metabolic Analysis Network)*. The Huttenhower Lab Department of Biostatistics, Harvard T.H. Chan School of Public Health.
3. Featured Application - Metagenomic Metabolic Manual (3M) Annotation Application.
4. Oracle Application Express (Apex).
5. KEGG: Kyoto Encyclopedia of Genes and Genomes.
6. MG-RAST Specimen1 - SRR062371 – 4473347.
7. MG-RAST Specimen2 - SRR062402 – 4473378.3.
8. Abubucker S, Segata N, Goll J, et al. Metabolic Reconstruction for Metagenomic Data and It's Application to the Human Microbiome. *PLOS Comput Biol*. 2012;8(6):e1002358.
9. Joseph Fass. BWA: TrimBWAstyle.pl. The Bioinformatics Core at UC Davis Genome Center. 2010.
10. Curtis Huttenhower. Genomic Data Manipulation BIO 508 Spring 2012 Problem 06. *Genomes and Sequencing*. 2012.
11. Curtis Huttenhower. *Meta'omic Analysis with MetaPhlAn, HUMAnN and LEfSe*. Department of Biostatistics, Harvard School of Public Health; 2012.
12. Ettinger G, MacDonald K, Reid G, et al. The influence of the human microbiome and probiotics on cardiovascular health. *Gut Microbes*. 2014;5(6):719–728.
13. DIAG (Data Intensive Academic Grid)