International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 3, No. 6 (2014): 538-542
Research Article
Open Access

ISSN: 2320-0634

Metadata-Driven Management, Analysis, and Visualization of GPCR data using NGS approach

Aman Chandra Kaushik and Shakti Sahi*

School of Biotechnology, Gautam Buddha University, Greater Noida, Uttar Pradesh, India.

*Corresponding author: Shakti Sahi; e-mail: shaktis@gbu.ac.in

Received: 21 October 2014 Accepted: 04 November 2014 Online: 17 November 2014

ABSTRACT

Next generation sequencing (NGS) technology is better method for genome and transcriptome sequencing. NGS technologies are relatively easy and error free compared to the Sanger method. With the help of NGS we can identify gene structure as well as transcriptome sequencing. A typical metadata driven management, analysis and visualization of G-Protein coupled receptor (GPCR) dataset in different-different species is reported here. We considered the assignment of GPCR reads to gene families using BLAST for identification of genes and introduced a clustering method which reduces the complexity of metagenome dataset. We report that the clustering method is more accurate than the direct assignment for studies of the *Homo sapiens* GPCRs and other GPCRs in general. Along with the advent of next-generation sequencing platforms, several high-performance sequence analysis pipelines will be helpful for the detection of type 2 diabetes.

Keywords: GPCRs, NGS, Dataset, BLAST, Type 2 diabetes.

INTRODUCTION

Study of GPCR genes in different species and their frequencies of genetic variations]using next generation sequencing can provide DNA-protein association on a genome scale.. Next generation sequencing technologies are powerful in detecting the genomic locations of DNA as well as DNA binding proteins], NGS technologies offer opportunities and challenges for sequencing studies [1-10], typical metadata driven management analysis and visualization of GPCR dataset in different species [11,12]. In the present paper we paper considered the assignment of GPCR reads to gene families using BLAST [13] for identification of i genes and introduced a clustering method which reduces the complexity of metagenome Operational Taxonomic Units (OTU) datasets [14]. On the basis of simulated metagenome a metadata set for GPCR's was generated. We show that the clustering method is more accurate than the direct assignment, and generate heat map plot according to GPCR's in different species.

MATERIALS AND METHODS

Metadata associated with the OTU data, refers to information about the GPCR sequences data, in this work a description of the GPCRs and species from which the sequence data were generated is used as metadata. Library (nomenclature) represents all of the GPCR's sequences generated from multiple libraries. In this work one-to-one relationship between libraries and GPCR's was analysed. Metadata for each library includes the anatomical position, GPCR environment description, acquisition methods, Metadata needs to be imported only once.

Metadataset- Meta dataset also have quality information associated with the read and predictions about potential GPCRs within the genome in different species.

Hierarchy & OTU- This option creates two panels on workspace screen, the upper panel shows the hierarchy, and the lower panel shows the OTUs. The hierarchy panel is exploration of the dataset in tree

format, and OTU panel shows lateral view of the data from GPCRs.

Library- Percentage of library tends to be more useful, and important since the total number of counts received from multiple library is beyond the control. Using percent of library allows fair comparison of libraries, otherwise libraries would have very large number of counts [15].

RESULTS AND DISCUSSION

Studies of the *Homo sapiens* GPCRs, and other GPCRs in general, along with the advent of next-generation sequencing platforms and several high-performance sequence analysis pipelines are helpful for the detection of type 2 diabetes. In this work we aim to develop a clustering method with the help of all available information to accurately align as many GPCRs as possible for Meta dataset.

An overview of the dominant organisms that exist in the dataset was generated using OTU stacked bar graph (Figure 1), This figure shows only the top 10 taxa of the orphan GPCRs, and total value of the first OTU in the column is 31.35, and total value of the 10th OUT in the column is 1.41. X axis represents the OTU percent of total whereas Y axis represents the library name. Different colors represent the source organisms or species of GPCRs. Red and brown appear to be dominant colors in this OTU Stacked Bar Chart and correspond to *Homo sapiens*.

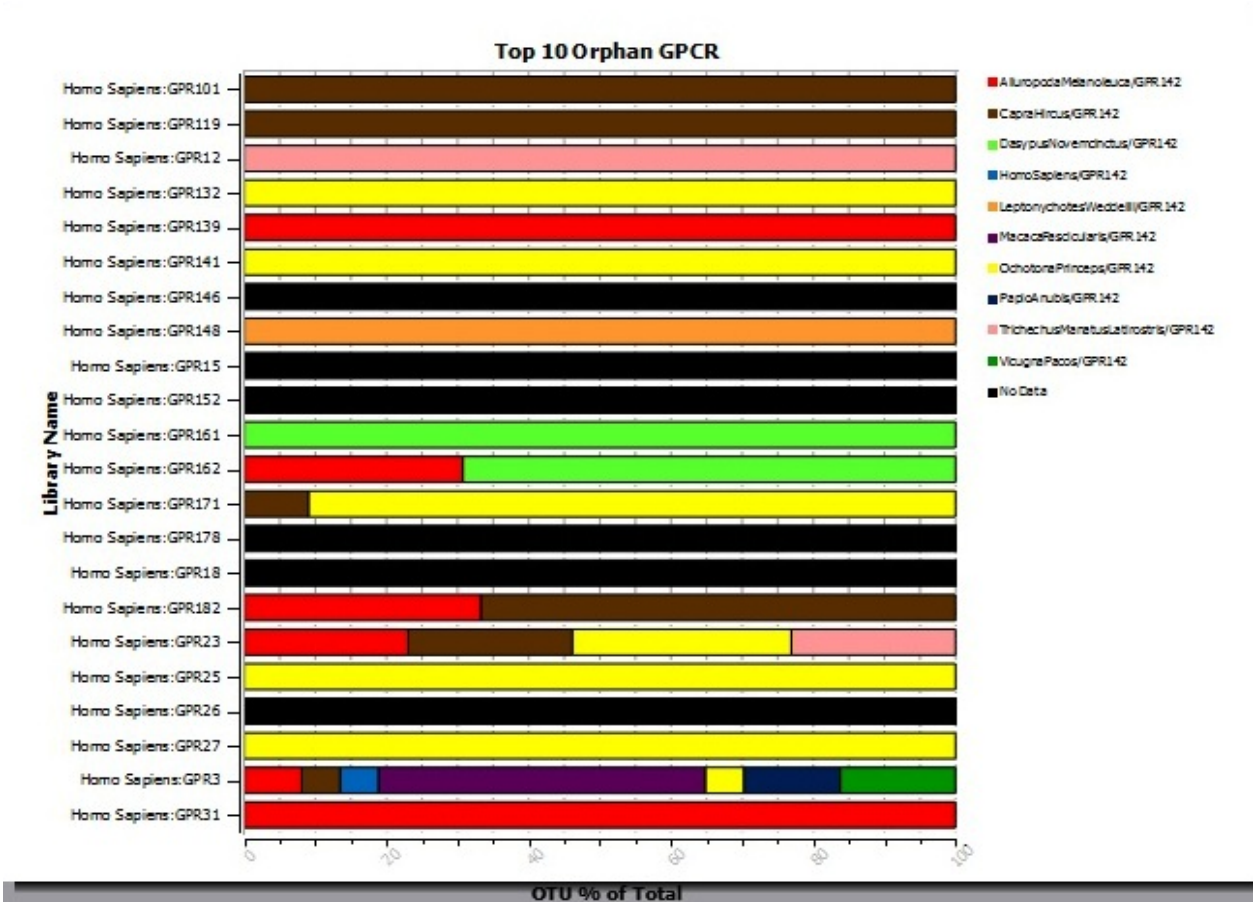

Figure 1. OTU Stacked Bar Chart of the top 10 most prevalent taxa. The hierarchy panel is exploration of the dataset in tree format, and OTU panel shows lateral view of the data from GPCRs.

Another useful way to generate an overview of the organisms in the Meta dataset was based on using pie chart (Figure 2) which helps to generate graphical depictions of the taxonomic hierarchy of 22 total libraries. Different wedges colors shows percentage of most prevalent phylum in the data set. Different species shows different wedges colors in form of percentage.

Beta Diversity (Morisita-Horn)- With the help of Beta Diversity GPCRs libraries can be represented in form of Morisita-Horn heat map. These heat maps give an estimate of the similarity of the GPCRs from different organisms into two anatomical positions (Fgiure 3). Morisita-Horn is an often used metric that gives insight into similarity or dissimilarity index of different sets of GPCRs. Meta datasets by looking at the patterns of all of the different OTUs simultaneously. Anatomical positions with Morisita-Horn values near one imply that the GPCRs Meta dataset of taxonomy patterns are very similar and appear in black color. While anatomical positions with Morisita-Horn values near zero imply that the GPCRs Meta dataset of taxonomy patterns are very different and appear in red color (Figure3). Based on this dataset the Homo sapiens GPCRs is more similar across subjects than the others.

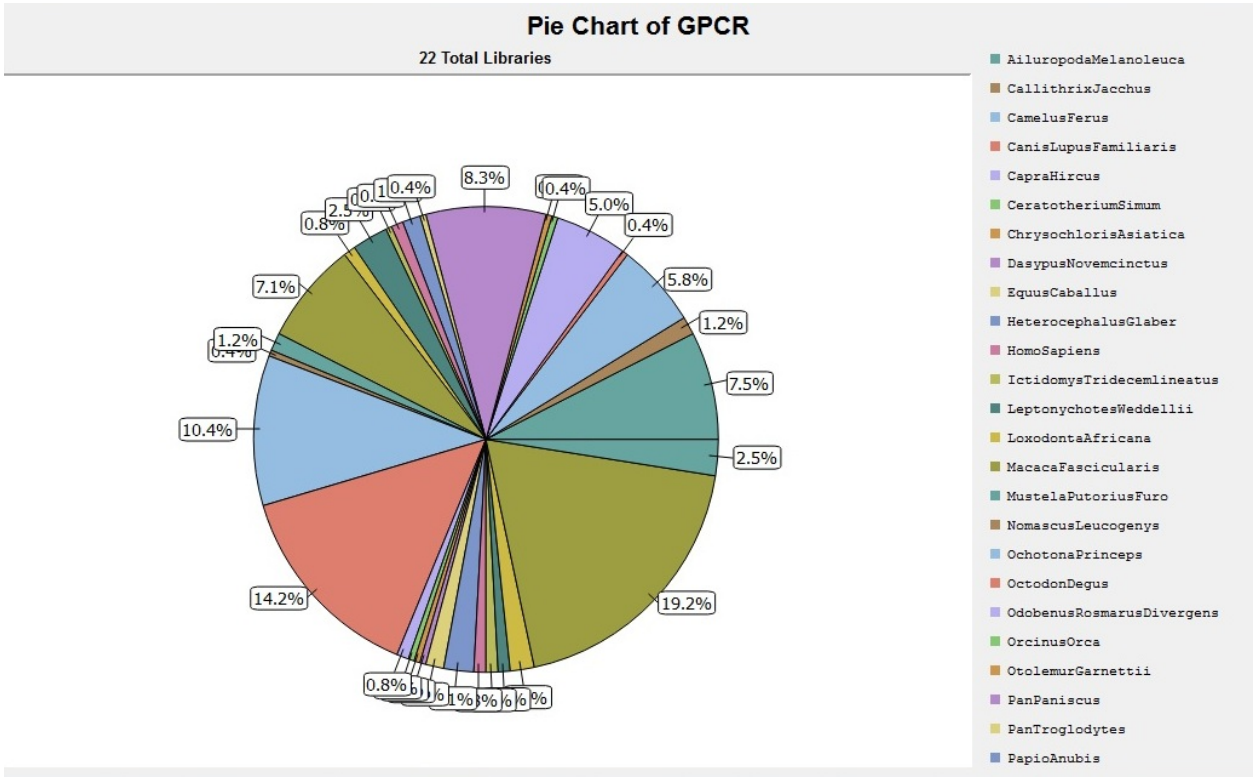

Figure 2. Taxonomy Pie chart of the Phyla, Meta dataset using pie chart which helps to generate graphical depictions of the taxonomic hierarchy of 22 total libraries.

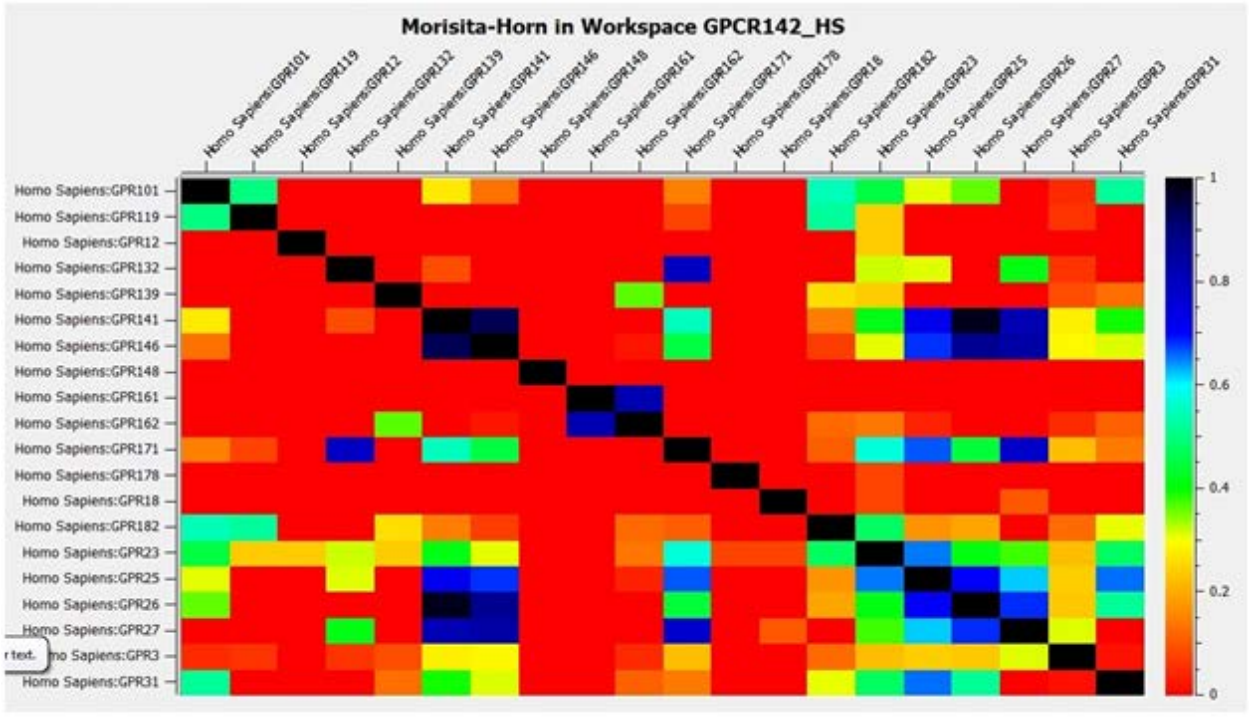

Figure 3. Morisita-Horn Heatmap Plot, Beta Diversity viewing GPCRs libraries in form of Morisita-Horn heat map.

OTU Heat Map Plot-. The similarity of the GPCRs from different organisms with the help of OUT heat maps also classifies into two anatomical positions of four organisms. Anatomical positions with OUT alues near 100 implyi that the GPCRs Meta dataset of taxonomy patterns are very similar and appear in black color. While anatomical positions with OTU values near zero imply that the GPCRs Meta dataset of taxonomy patterns are very different and appear in red color (Figure 4). Based on this dataset the Homo sapiens GPCRs is more similar across subjects than the others.

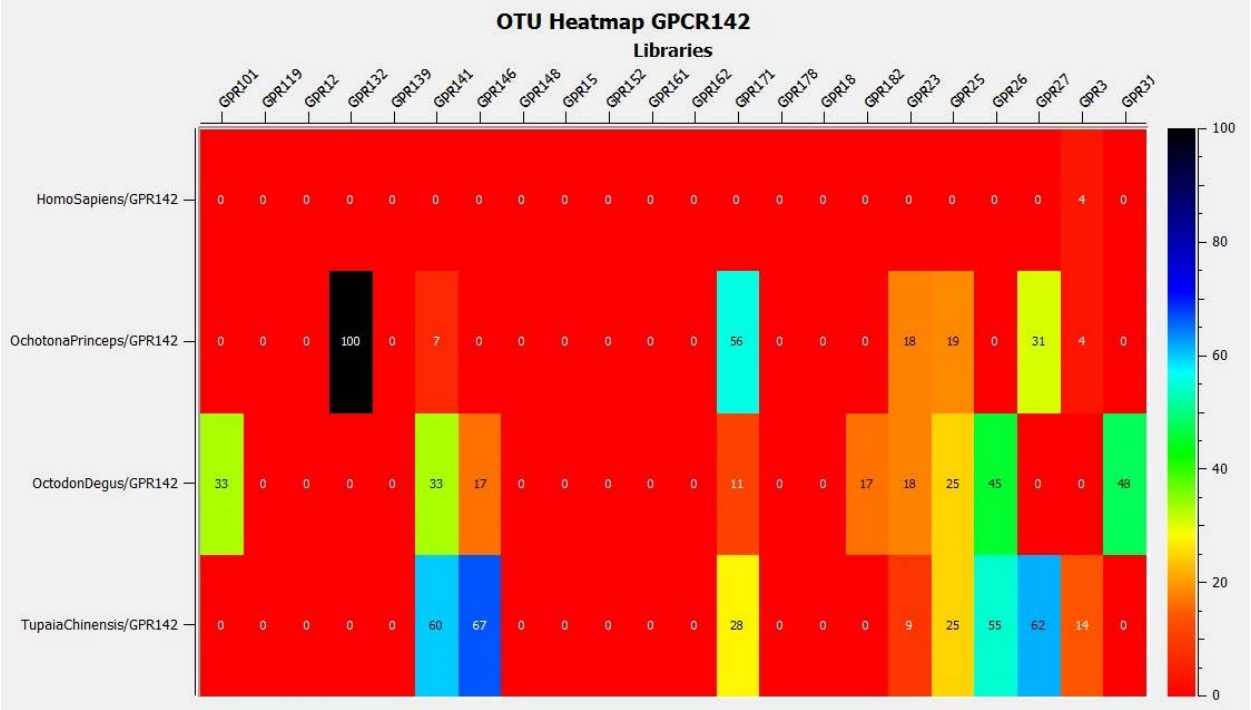

Figure 4. OTU Heat map Plot of GPCR142, where Y axis represents some selective species.

The Manhattan plot in Figure 5 shows logarithmically transformed p-values, with higher peaks representing lower, and lower peaks representing significant p-values. The horizontal lines represent p-value of 0.08, 0.09, and 0.01. Inclusion of the p-value=1 line is intended to highlight taxa that are approaching significance analysis. The X axis represents the alphabetical position by number of each OTU name in the two part format. The first significant peak position 0.8 corresponds to *Homo sapiens* GPR141, , the second significant peak position 0.7 corresponds to *Homo sapiens* GPR146, , the third peak position 1.0 corresponds *to Homo sapiens* GPR148, and the 4th significant peak position 0.8 corresponds to *Homo sapiens* GPR162,. , The 5th significant peak position 0.8 corresponds to *Homo sapiens* GPR171, The 6th significant peak position 0.7 corresponds *to Homo sapiens*GPR182, and the 7th significant peak position 0.5 corresponds to *Homo sapiens* GPR23, respectively. The peaks with high alpha index are regarded as higher proportion and relative abundance in the dataset and vice versa for lower peaks.

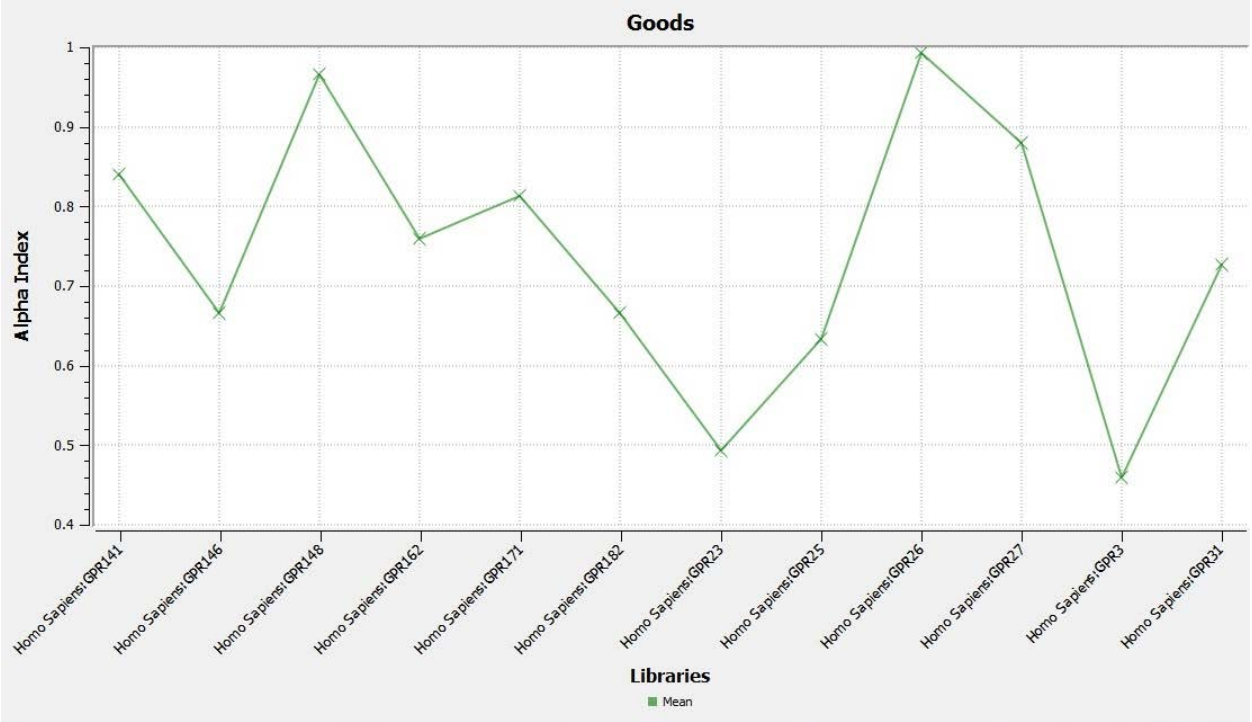

Figure 5. Two parts results displayed as a Manhattan plot.

CONCLUSION

Studies of the Homo sapiens GPCRs, and other community GPCRs in general, along with the advent of next-generation sequencing platforms, have dramatically increased the scope of genomics related work. In the present work we have carried out metanalysis of orphan GPCR datasets and visualized the results using OTU heat maps, Manhattan plot and Morista-Horn heat maps. Different GPCR metadatasets were created using data from different organisms as well as orphan GPCRs from *Homo sapiens.* The results indicated that the different orphan GPCRs from *Homo sapiens* were closely related to each other as compared to GPCRs from other organisms. These studies can provide an insight into the mechanism of action of orphan GPCRs and will be helpful in drug designing processes

REFERENCES

1. Bao E. Seed (2011): efficient clustering of next-generation sequences. Bioinformatics;14(18):2502–2509.
2. Du X, Kim YJ, Lai S, Chen X, Lizarzaburu M, Turcotte S, Fu Z, Liu Q, Zhang Y, Motani A, Oda K, Okuyama R, Nara F, Murakoshi M, Fu A, Reagan JD, Fan P, Xiong Y, Shen W, Li L, Houze J, Medina JC (2012), Phenylalanine derivatives as GPR142 agonists for the treatment of type II diabetes: Bioorg Med Chem Lett.;22(19):6218-23.
3. Süsens U, Hermans-Borgmeyer I, Urny J, Schaller HC (2005), Characterisation and differential expression of two very closely related G-protein-coupled receptors, GPR139 and GPR142, in mouse tissue and during mouse development: Neuropharmacology. 2006 Mar;50(4):512-20. Epub.
4. Huang W, Li L, Myers JR, Marth GT. ART (2012): a next generation sequencing read simulator. Bioinformatics; 14(4):593–594. doi: 10.1093/bioinformatics/btr708.
5. Shokralla S, Spall JL, Gibson JF, Hajibabaei M. (2012); Next-generation sequencing technologies for enviromental DNA research. Molecular Ecology;14(8):1794–1805.
6. Leinonen R, Sugawara H, Shumway M. (2011); The sequence read archive. Nucleic acids research;14(1):D19–D2.1.
7. Lizarzaburu M, Turcotte S, Du X, Duquette J, Fu A, Houze J, Li L, Liu J, Murakoshi M, Oda K, Okuyama R, Nara F, Reagan J, Yu M, Medina JC (2012), Discovery and optimization of a novel series of GPR142 agonists for the treatment of type 2 diabetes mellitus: Bioorg Med Chem Lett. 2012 Sep 15;22(18):5942-7. doi: 10.1016/j.bmcl.2012.07.063. Epub.
8. MacLean D, Jones JDG, Studholme DJ. (2009); Application of next-generation sequencing technologies to microbial genetics. Nature Reviews Microbiology;14(4):287–296.
9. Miller J, Koren S, Sutton G. (2010); Assembly algorithm for next-generation sequencing data.Genomics;14(6):315–327. doi: 10.1016/j.ygeno.2010.03.001.
10. Ross JS, Cronin M. (2011) Whole cancer genome sequencing by next-generation methods. Am J Clin Pathol;14(4):527–539.
11. Kakarala KK, Jamil K (2014), Sequence-structure based phylogeny of GPCR Class A Rhodopsin receptors, Mol Phylogenet Evol.
12. Fredriksson R, Höglund PJ, Gloriam DE, Lagerström MC, Schiöth HB (2003), Seven evolutionarily conserved human rhodopsin G protein-coupled receptors lacking close relatives: FEBS Lett.;554(3):381-8.
13. Needleman SB, Wunsch CD. (1970); A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol;14(3):443–453. doi: 10.1016/0022-2836(70)90057-4.
14. Schloss PD, Handelsman J. Introducing DOTUR a computer program for defining operational taxonomic units and estimating species richness. Applied and environmental microbiology.2005;14(3):1501–1506.
15. Robertson CE, Harris JK, Wagner BD, Granger D, Browne K, Tatem B, Feazel LM, Park K, Pace NR, Frank DN (2013). Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. Bioinformatics 29(23):3100-1.

*****