Metagenomic Storage and Analysis System Design

Irsath Azeez , Philippe Chlenski, and Prof. Itsik Pe'er

Columbia University School of Engineering

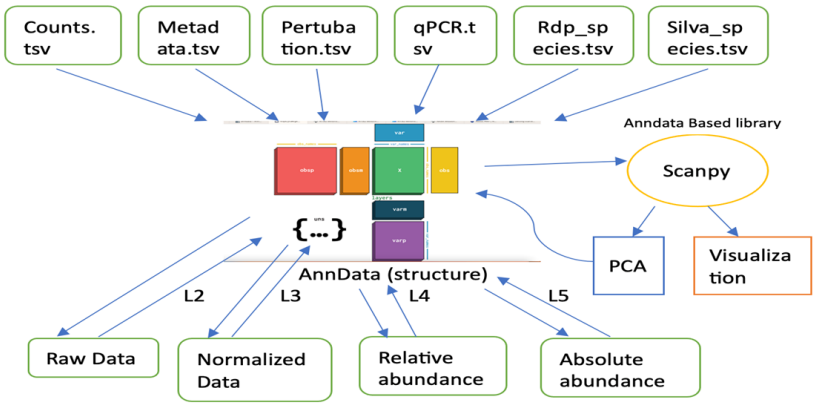
Introduction: The aim of our research is to collect Amplicon Sequence Variant (ASV) which is "used to analyze 16S rRNA gene sequence data"(Schloss, 2021). Once the sequence data and sample metadata are collected, we design a data structure to encapsulate all the data at one place for further analysis. The data is crucial for our research, thus, we have gathered all the ASV sequences data and metadata to store them on a disk using the anndata structure which provides an framework to store all the data in one platform. This anndata is interfaced with Scanpy which are toolkits used to analyze anndata. Scanpy "includes preprocessing, visualization, clustering, trajectory, inference and differential expression testing"(scanpy, 2023). We designed a data structure that contains feature selection, visualization, and mathematical calculations. We were also able to store the data on a disk in a memory efficient way since anndata provides the feature to store the data in categorical method.

Materials & Methods:

Data Sets(.tsv format):

ASV sample counts; Metadata Perturbation intervals; Silva species annotation; qPCR measurements; RDP species annotation

System Design Flow Chart



Results: The outcome of our project results in creating a system design by encapsulating Metagenomic sample ASV abundance, and adding featured taxonomical meta data on to the system of anndata. By accessing the data from the anndata system, we were able to calculate the Principal Component Analysis (PCA) as well as relative and absolute abundance of sample sequences. We were also able to store the results on the anndata system in layers for further analysis.

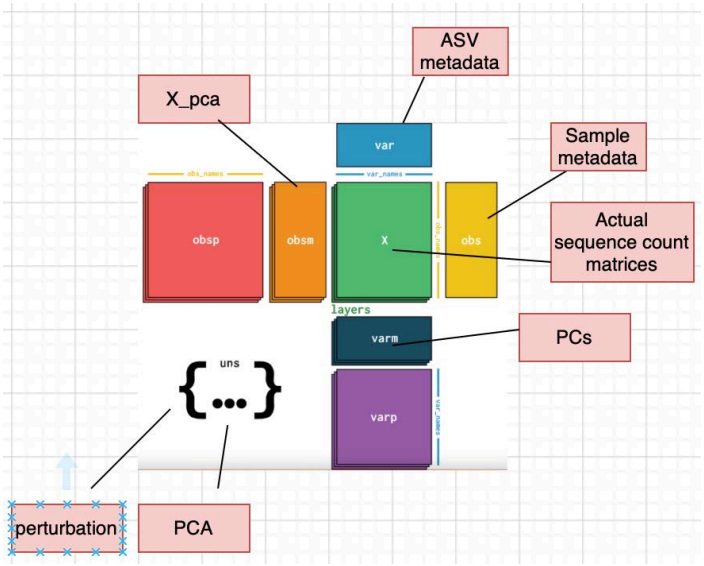


Figure 1: Completed anndata system design using ASV sample sequences

Conclusion: In conclusion, our focus on creating a system design structure to store our data to access and analyze them with user-friendly manner. Since the research associates with large abundant data, our consideration in saving the file with low memory space without losing any data in our dataset. As this system provide easy access to the dataset, we are able to save time and pre-process and visualize the data sets with convenience. This system design also help us to further analysis such as timeseries prediction analysis, clustering, etc. with the help of other python libraries involves in machine learning task. Our future analysis will involve using timeseries forecasting into supervised learning.

References:

Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angere, F. Alexander Wolf, anndata, bioRxiv 2021 Dec 19. doi: 10.1101/2021.12.16.473007

FA Wolf, P Angerer, FJ Theis, scanpy Genome biology 19, 1-5

Schloss, Patrick D. "Amplicon Sequence Variants Artificially Split Bacterial Genomes into Separate Clusters." mSphere vol. 6,4 (2021): e0019121. doi:10.1128/mSphere.00191-21

Callahan, Dr. Benjamin John. "This Illustration Shows the Precision of ASVs." WIKEMEDIA COMMONS, 14 Sept. 2020, https://commons.wikimedia.org/wiki/File:Amplicon_Sequence_Variants.png.

Gibson, Travis E., et al. "Intrinsic Instability of the Dysbiotic Microbiome Revealed through Dynamical Systems Inference at Scale." bioRxiv, 16 Dec. 2021, https://www.biorxiv.org/content/10.1101/2021.12.14.469105v1.full.pdf.