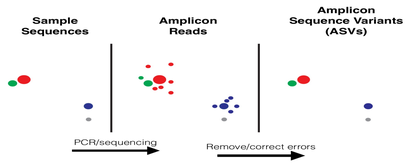
Metagenomic Storage and Analysis System Design

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.

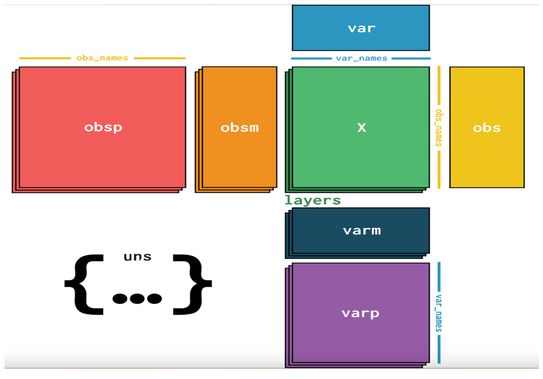


This visually shows how ASVs remove and correct errors from PCR

sampleID	ASV_1	ASV_2	ASV_3	ASV_4	ASV_5	A
1-D0AM	0.0	43.0	50.0	47.0	18.0	
1-D0PM	9.0	32.0	107.0	228.0	45.0	
1-D1AM	9.0	65.0	946.0	1865.0	487.0	2
1-D1PM	53.0	90.0	2962.0	7008.0	1077.0	4
1-D29AM	2729.0	9482.0	...	17.0	953.0	2350.0
...	...	...	...	...	...	...
M2-D9-1B	2911.0	11314.0	20850.0	65.0	614.0	5
M2-D9-2A	3834.0	19051.0	27270.0	123.0	828.0	7
M2-D9-2B	3104.0	15783.0	21064.0	100.0	586.0	5
M2-D9-3A	3626.0	18266.0	26375.0	107.0	725.0	7
M2-D9-3B	3341.0	12650.0	22036.0	97.0	634.0	6

339 rows × 1088 columns

ASV sequence counts data from different samples



General structure of anndata

Objectives

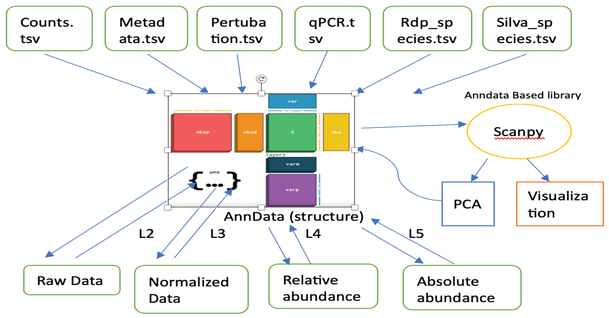
Our objectives are to design a system for data by gathering ASV sequences and sample metadata and store them on anndata platform. We also needed to save the data structure in a memory efficient way to save space due to the large data sets that we are analyzing. Having a data format, we will be able to transition the data for machine learning tasks.

Materials & Methods

Data Sets: (.tsv format)

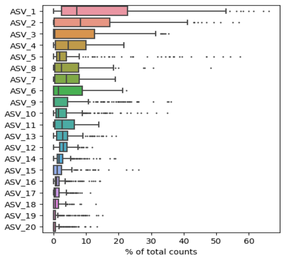
- ASV sample counts
- Metadata
- Perturbation intervals
- Silva species annotation
- qPCR measurements
- RDP species annotation

System Design Flow Chart

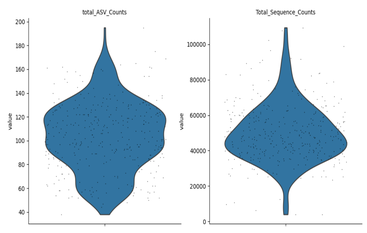


Data Visual Analysis

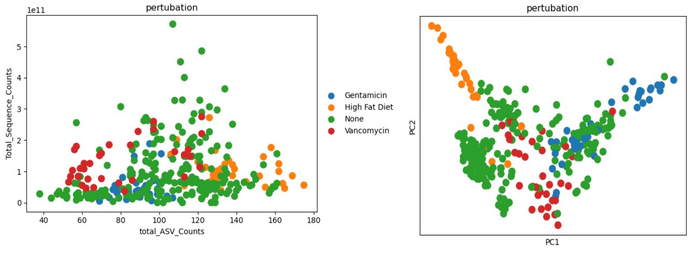
Most abundant ASV sample counts:



ASV sample counts showed in violin plot:

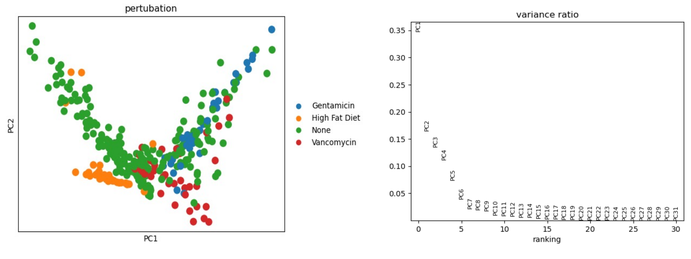


Absolute abundance scatter plot:



Normalized data PCA:

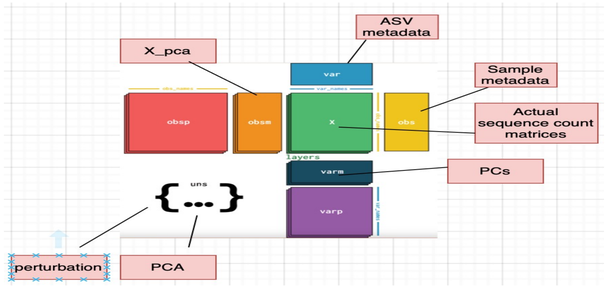
Principal Component Analysis(PCA) of actual data and variance ratio:



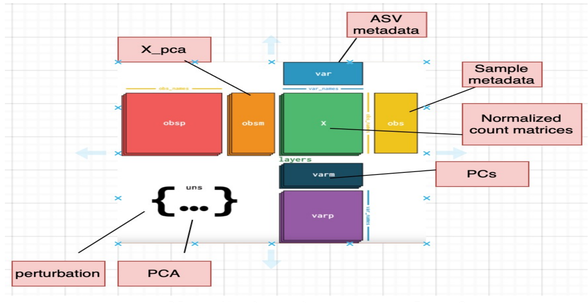
Results

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.

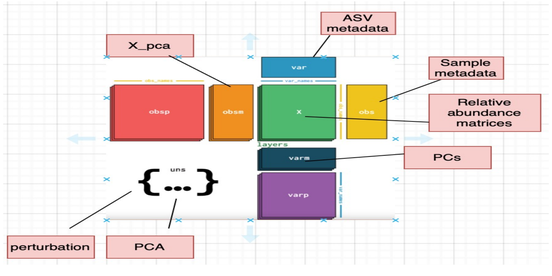
Final anndata system with all encapsulated data:



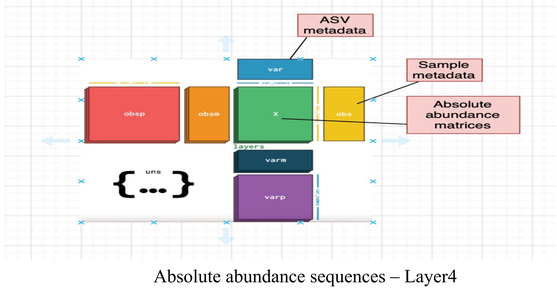
Actual sample sequences – Layer1



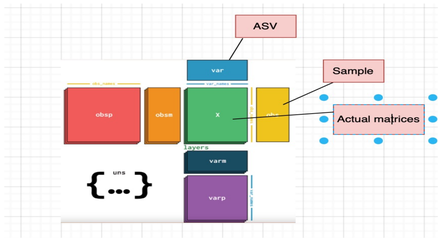
Normalized sample sequences – Layer2



Relative abundance sequences – Layer3



Absolute abundance sequences – Layer4



Raw data – layer5

```
adata
✓ 0.0s                                                    Python

AnnData object with n_obs × n_vars = 339 × 1088
    obs: 'subject', 'time', 'Gentamicin', 'High Fat Diet', 'Vancomycin', 'total_ASV_Counts', 'Total_Sequence_Counts', 'pertubation', 'measure
    var: 'rdp_sequence', 'silva_sequence', 'Sample_Counts', 'mean_counts', 'pct_dropout_by_counts', 'total_counts', 'rdp_kingdom', 'silva_kir
    uns: 'pertubation_colors', 'pca'
    obsm: 'X_pca'
    varm: 'PCs'
    layers: 'normalized_data', 'raw_data', 'absolute_abundance', 'relative_abundance'
```

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.