**Appendices**

|  |  |  |  |
| --- | --- | --- | --- |
| Tool | Version | Reference | Used for |
| python | >=3.10 | https://www.python.org/downloads/release/python-3100/ | To run all the scripts |
| urllib.request | 3.9 | https://docs.python.org/3/library/urllib.request.html | Download the VCF gzip files |
| gzip | 3.12 | https://docs.python.org/3/library/gzip.html | Unzip the VCF files |
| requests | 2.31.0 | https://requests.readthedocs.io/en/latest/ | Download and save the reference genome |
| numpy | 1.23.1 | https://numpy.org/ | For preprocessing the SBS file. To calculate a cutoff value for data normalization. And used for multinomial randomization |
| pandas | 1.5.0 | https://pandas.pydata.org/ | Used for all the matrices operations |
| pyarrow | 14.0.1 | https://arrow.apache.org/docs/python/index.html | To write the large SBS and NMF files as parquet files to save space. |
| fastparquet | 2023.10.1 | https://fastparquet.readthedocs.io/en/latest/ | To write the large SBS and NMF files as parquet files to save space. |
| scikit learn | 1.3.1 | https://scikit-learn.org/stable/ | Calculating optimal column assignments between two DataFrames based on the linear sum assignment and calculating Jensen-Shannon distance |
| seaborn | 0.13.0 | https://seaborn.pydata.org/ | Used to create all the plots/figures. |
| matplotlib | 3.7.1 | https://matplotlib.org/ | Used for the backend of all the plots/figures. |
| statkit | 0.2.3 | https://gitlab.com/hylkedonker/statkit | To calculate the perplexity of the NMF result. |
| pdoc | 14.1.0 | https://pdoc.dev/ | Auto-generates documentation |
| mubelnet | Commit 7ed677cd | https://gitlab.com/hylkedonker/mubelnet | Performs the deep Bayesian unsupervised clustering models and trains them. |

*Table 1: A table of all the used* *libraries/tools/plugins for this research. And what they were used for.*