FAIR AWARE Assignment

1. Are you aware that a data(set) should be assigned a globally unique persistent and resolvable identifier when deposited with a data repository?

**Based on my understanding of an identifier, this dataset has a globally unique identifier according to the GEO databased, specifically GSE276609. Each of the samples also have unique identifiers, for example, GSM8502984. This is the most well-organized GEO database I’ve ever seen – here are unique identifiers/links for not only the GEO series but also GEO sample, SRA experiment, biological sample and its corresponding project.**

1. Are you aware that when you deposit a data(set) in a data repository, you will need to provide discovery metadata to make the data(set) findable, understandable and reusable to others?

**Each “GSM” data set has corresponding metadata that is well organized and gives a brief, yet thorough, understanding of the linked data. For example, GSM8502984 corresponds to BCM-2147\_109176, Basal-like triple-negative breast cancer patient-derived xenograft**.

1. Are you aware that the data repository providing access to your data(set) should make the metadata describing your data(set) available in a format readable by machines as well as humans?

**Yes, this database provides a SOFT, MINiML, and TXT formatted metadata that is readable by machines as well as humans.**

1. Are you aware that access to your data(set) may need to be controlled and that metadata should include license information under which the data(set) can be reused?

**After having gone through the metadata files, I could not identify any licensing information, however, there was information listed regarding the authors and associated PMID to the publication, which I believe could act, in part, as a “license” binding the dataset to a publication. However, licensing information in accordance with the definition of the FAIR AWARE website was not (at least easily) identifiable in the metadata or the rest of the database page on GEO. Perhaps the authors just wanted to make this a publicly available database and did not see the need to providing licensing information.**

1. Are you aware that metadata should remain available over time, even if the data(set) is no longer accessible?

**Yes, this database provides a SOFT, MINiML, and TXT formatted metadata that should remain available overtime. The fact that the authors provide all three formats minimizes the probability that machines will move beyond the readability of all three file times in the near and distant future.**

1. Are you aware that the metadata describing your data(set) should use controlled vocabularies?

**Controlled vocabularies, according to the FAIR AWARE website, facilitate the interoperability of metadata usage. Vocabulary can correspond to varying degrees of sophistication which is why it’s important to utilize existing databases of vocabulary that can make the metadata easily interpretable regardless of someone’s educational background. I would even change their definition from saying vocabulary varies based on “sophistication” to “experience speaking the English language” because there are many sophisticated scientists that don’t necessarily speak English well from around the world (slightly egocentric of the FAIR AWARE people on this one). Going through the metadata as well as the “Overall Design” and “Summary” sections, I believe the authors did a good job of making the information easily digestible and thoroughly describe each dataset (i.e. the loupe browser documents, the CancerCellsOnly data, etc.).**

1. Are you aware that provenance information about the collection and/or generation of data should be included in the metadata?

**Data provenance is a fancy term of “lineage” or, more aptly, how your data was generated and if it has undergone any modification. The metadata provides a series summary motivating the experiments that were done by the authors (all listed with corresponding contact information), and a “Series\_overall\_design” section that outlines the how datasets that were made available were generated. In this data, they reanalyzed data from two other datasets in “GSE235168” and “GSE161529”. Specifically, for the GSE161529 dataset, the authors explain the lineage of how the data was generated such that they requested the original authors of that dataset to run the raw, original data through CellRanger v6 and share the generated output files with their analysis team. This level of detail that another team was responsible for generating the CellRanger output files is very pertinent and can help someone, like myself, who has questions about how that data was generated to aim those questions to the correct analysis team if I’d like to reuse or generate similar data.**

1. Are you aware that metadata describing your data(set) should follow the specifications of a community-endorsed standard?

**Community-endorsed standards ensure readability by individuals in the same domain of research as the research study that was conducted. As someone with domain knowledge, I can verify that the readability of the summaries in the database and the metadata is very high and easy to digest. I do have one recommendation however, and this might just be a personal thing – but I like it when authors provide sections in their actual manuscript when they refer to where data or specific analyses were utilized in their papers. For example, in the metadata, it states that “we provide 5 Loupe Cell Browser files that contain the merged dataset as presented in the paper (.cloupe files).” As a reader, I’d like to know exactly which part of the paper this is in reference to especially when most scRNAseq or RNAseq papers are just figures and figures of graphs that correspond to different analyses/datasets.**

1. Are you aware that your data(set) should be deposited preferably in a file format that is open and supported by the data repository for long-term preservation?

**All of the datasets (.cloupe .TAR .txt) files are presented in formats that are open and supported by GEO. This question actually made me look into why some data offers FTP means of download while others are HTTPS. It seems FTP is for institutions with legacy tools that require FTP to pull data – very interesting since I didn’t realize that could still be prevalent. Knowing this, that means this dataset has even more accessibility than I previously imagined. Based on my readings, it seems all modern browsers are compatible with pulling data using HTTPS or FTP so it doesn’t make a difference for me, but it could in older institutions around the world.**

1. Are you aware that keeping your data(set) FAIR over time requires professional data curation and digital preservation?

**Data stewardship entails proper maintenance and organization of data to maintain preservation over time. Using a trusted digital repository (TDR) is essential to this process and GEO is a TDR.**