Hannah Stadtler

FAIR Aware Assessment

Bioinformatics Fall 2024

The dataset we are meant to analyze was uploaded to NCBI>GEO and is titled “Evaluation of Breast Cancer PDX Tumor Heterogeneity at Single Cell Resolution II.” Below are the 10 questions associated with the FAIR Data assessment.

1. *Data sets should be assigned a globally unique persistent and resolvable identifier when deposited with a data repository* 
   1. This data set has its own unique GEO accession number (GSE276609), so yes, this dataset has been assigned a globally unique identifier.
2. *You will need to provide discovery metadata in order to make the dataset findable, understandable, and reusable to others*
   1. Metadata is available for this dataset, including information on the organism associated with the dataset (homo sapiens), a summary of the dataset, and a summary of the overall design, all which seems like metadata that will/can persist even if the actual data has been removed, which is consistent with FAIR principles.
3. *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*
   1. It’s definitely readable by humans! Per the FAIR assessment help icons, it seems that the data repository should ensure that the data is readable by machines regardless of if the researcher did this step or not, so I’m going to say that this principle has been achieved as well (at the very least by uploading to a reputable data repository!).
4. *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*
   1. Yes, underneath the “status” section, there is a blurb that indicates that this data is private and should not be shared or distributed without permission, indicating that access ability has been defined and license information is present.
5. *Metadata should remain available over time, even if the data(set) is no longer accessible*
   1. This is hard to answer because the data are still available, and this dataset was just uploaded, but hopefully metadata will remain available if this dataset were every taken down.
6. *Metadata describing your data(set) should use controlled vocabularies*
   1. Although I cannot say for certain that everything associated with this dataset uses a controlled vocabulary, because that feels a little subjective, it is important to note that where the researchers have acronyms that may not be widely known, they define the acronym, which increases interoperability and FAIRness of the dataset. For example, PDXs is explained to be “patient-derived xenografts” which can point users in the right direction if they were previously unfamiliar with PDXs. However, “ER+” might not be known to others outside of those familiar with triple negative breast cancer and could be better defined like PDX was.
7. *Provenance information about the collection and/or generation of data should be included in the metadata*
   1. Provenance information is available for this dataset, as there is a section that lists the contributors to this data, as well as information on submission date, last update, and other pertinent information that allows this data to be better tracked and reused. Also, the overall design includes information on the origins of the data that were changed within this dataset, which is also a good FAIR practice.
8. *Metadata describing your data(set) should follow the specifications of a community-endorsed standard*
   1. Per the FAIR assessment help icons, uploading data to a repository that’s familiar with your type of data will help ensure that metadata follows specifications of a community-endorsed standard. It seems like GEO is very familiar with this type of data, and would ensure that the metadata is structured appropriately, so I would say that this has been achieved.
9. *The data(set) should be deposited preferably in a file format that is open and supported by the data repository for long-term preservation*
   1. It seems like uploading to GEO allows for this long-term preservation!
10. *Keeping your data(set) FAIR over time requires professional data curation and digital preservation* 
    1. Similar to #5 of this assessment, it’s hard to say if this will be achieved as perhaps one day in the future GEO will completely crash and this dataset (even the metadata?) may be lost. However, hopefully with consistent data curation and digital preservation, that will be avoided, and this principle will also have been fulfilled!

Overall, I would say that this dataset does a very good job at endorsing FAIR data principles, with the only area of improvement being perhaps better defining some of the other more technical terms in the summary and overall design to increase interoperability.