Paulette Omeaku

CCTR691

October 3, 2025

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

Yes, the dataset GSE276609 was assigned the unique accession number “GSE276609” within GEO, which can be resolved using the URL (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE276609>).

1. **Are you aware that when you deposit a dataset in a data repository, you will need to provide discovery metadata in order to make the dataset findable, understandable and reusable to others?**  
   Yes, for GSE276609, the GEO entry includes discovery metadata like the dataset “Evaluation of breast cancer PDX tumor heterogeneity at single cell resolution II”, sample counts (117 samples), organism (Homo sapiens), experiment type (single cell RNA-seq), and related datasets (GSE235168, GSE161529).
2. **Are you aware that the data repository providing access to your dataset should make the metadata describing your dataset available in a format readable by machines as well as humans?**  
   Yes, GEO provides human-readable pages (HTML) for the series and samples. And metadata is available for download. However, the metadata aren’t really standardized machine-readable formats (e.g., JSON-LD, RDF) or explicit machine-actionable metadata.
3. **Are you aware that access to your dataset may need to be controlled and that metadata should include licence information under which the dataset can be reused?**  
   **Response:** Yes in principle, but for GSE276609 the evidence for a clearly stated reuse licence is limited. The dataset is marked “Public” (available since Sep 30 2024) in GEO, and raw data are submitted to SRA. However, a specific data usage licence (e.g., CC-BY, CC0) is **not** explicitly visible in the series metadata. Thus metadata include the access status (public) but not an explicit licence statement.  
   Example: The “Status: Public” is shown in the GEO record.  
   Note/Improvement: To fully meet the criterion, the metadata should clearly indicate a licence (for example “CC-BY 4.0” or equivalent) specifying reuse conditions. If restricted access samples exist, the metadata should clearly state the control mechanism or embargo.
4. **Are you aware that metadata should remain available over time, even if the dataset is no longer accessible?**  
   Yes, the GEO repository keeps metadata for datasets even if raw data becomes unavailable. For GSE276609, the series metadata is accessible through the GEO’s web interface.
5. **Are you aware that the metadata describing your dataset should use controlled vocabularies?**  
   **Yes,** metadata for GSE276609 describes the organism, experiment type, and sample metadata using standard terminologies (“Expression profiling by high throughput sequencing”, “Homo sapiens”), which are recognized by GEO. But, it’s not clear if the metadata uses controlled vocabularies (OBO terms, NCIT, EFO) for all sample features or variables.
6. **Are you aware that provenance information about the collection and/or generation of data should be included in the metadata?**  
   **Yes,** GSE276609 provides provenance details like the dataset “contains a total of 117 samples, includes new and re-analyzed samples, 25 are from GSE235168, the remaining 69 samples, re-analyzed, do not have raw data archived.” This shows how the data was generated and links to predecessor datasets, and submission.
7. **Are you aware that metadata describing your dataset should follow the specifications of a community-endorsed standard?**  
   Yes, the dataset is placed in GEO, which is a widely-used community repository for gene expression and sequencing data, and follows community conventions for sample and series metadata. The raw data are submitted to SRA, another community standard repository.
8. **Are you aware that your dataset should be deposited preferably in a file format that is open and supported by the data repository for long-term preservation?**  
   Yes, for GSE276609, raw sequencing data are submitted to SRA (which uses formats like FASTQ, BAM), and GEO provides processed data files (count matrices). These are standard formats within the community which support long-term preservation.
9. **Are you aware that keeping your dataset FAIR over time requires professional data curation and digital preservation?**  
   Yes, the dataset is archived in established repositories (GEO and SRA), which implies some level of professional curation and long-term preservation infrastructure. GEO is maintained by NCBI/NIH. But, the metadata for this specific series does **not** explicitly claim ongoing active curation or long-term guarantee of preservation.