Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D1401B3ec0

Title: Placental and breastmilk microRNAs in relation to early-life growth and metabolism

Creator: Elizabeth Goehring - **ORCID:** 0000-0003-0074-2840

Affiliation: Emory University (emory.edu)

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: PA-20-188

Grant: K99HD104991

Template: NIH-Default DMSP

Project abstract:

Gestational and early childhood exposures, including maternal metabolic status and breastfeeding, are increasingly recognized as critical in the development of childhood and life-long obesity. The proposed research seeks to comprehensively assess the impact of maternal metabolic status on early childhood growth and metabolism, through the gene regulatory dynamics of placental and breastmilk microRNAs. This research has the potential to advance our understanding of the contribution of microRNAs and early-life exposures on growth and metabolism, and to more broadly contribute the Developmental Origins of Health and Disease.

Start date: 02-01-2024

End date: 02-01-2027

Last modified: 07-08-2024

Copyright information:

language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Placental and breastmilk microRNAs in relation to early-life growth and metabolism

Data Type

Types and amount of scientific data expected to be generated in the project: *Summarize the types and estimated amount of scientific data expected to be generated in the project.*

Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)

This project will produce gene expression data and hormone concentration data from RNA-sequencing and immunoassay, respectively. Demographic, health endpoint and survey data as well as biospecimen collection will be performed in the Atlanta Metropolitan area with from lactating mothers and their infants over a 6 to 7 month period. Data for sequencing analysis will be collected from 50 research participants at 3 time points, generating a sequencing data for approximately 150 samples. Data for hormone data will be collected from 50 research participants at 2 time points, generating approximately 700 observations for 100 samples. Raw sequencing data files will be processed to generate CSV files containing gene-level counts for all participants, to enable statistical analysis.

Scientific data that will be preserved and shared, and the rationale for doing so: *Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.*

Raw sequencing read files and metadata will be accessible through dbGAP.

Processed count data and metadata will be provided at the Gene Expression Omnibus.

All analysis codes and de-identified files needed to reproduce analyses will be shared through Emory Dataverse. Each dataset created will be accompanied by a R markdown file to document the preprocessing steps, basic statistics, and profile figures such as the intended sample heterogeneity shown in PCA, and correlation heatmaps.

Metadata, other relevant data, and associated documentation: Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

For gene expression and hormone data, metadata (including detailed methodology and procedures used to collect the data, data labels, definitions of variables, and any other information necessary to reproduce and understand the data) will be provided within the Material and Methods section or additional Supplementary information of each peer-reviewed publication. For RNA sequencing data, associated metadata will accompany the data sharing through Emory Dataverse.

To facilitate the interpretation and reuse of the data, a README file and data dictionary will be generated and deposited into Emory Dataverse along with all shared datasets. The README file will include methods descriptions (e.g., software, databases, libraries, services), and Protocol DOIs issued from protocols.io. The data dictionary will define and describe all variables in the dataset.

Related Tools, Software and/or Code

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

No new tools, software, or code will be developed during the project.

All RNAseq analyses are performed using publicly available software packages available through CRAN and bioconductor.

Standards

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources, and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist

As noted in (1) above, we will, when possible, make use of public data repositories for storage of all sequencing data and results, applying onotologies and formats required by those repositories. We will use a standardized ontology in line with those used in other large population studies such as NHANES. For any provided data, detailed data dictionaries, metadata and relevant references will be provide which define all variables.

Data Preservation, Access, and Associated Timelines

Repository where scientific data and metadata will be archived: Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived.

Raw sequencing data and associated metadata from this study will be accessible through dbGAP.

Processed data in the form of a count matrix and associated metadata, will be provided at the Gene Expression Omnibus.

How scientific data will be findable and identifiable: Describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

Findable: All public archives utilize a variety of search features for data to be findable based on metadata criteria.

Identifiable: Data stored in public archives such as dbGAP, GEO, or SRA are given uniquely identifiable accession numbers. We will include those accession numbers in all publications arising from this research.

When and how long the scientific data will be made available: Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

RNA sequencing data will be shared within 6 months of data completion and will remain available in perpetuity in public archives.

Access, Distribution, or Reuse Considerations

Factors affecting subsequent access, distribution, or reuse of scientific data: NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing.

There are no anticipated factors or limitations that will affect the access, distribution or reuse of the scientific data generated by the proposal.

Whether access to scientific data will be controlled: State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

Sequencing data made accessible in dbGAP is controlled and requires that investigators requesting the data have training in the conduct of human subject research and have received appropriate IRB approvals or waivers to work with such data. Researchers requesting access to read-level human data will need to request such access through dbGAP.

Protections for privacy, rights, and confidentiality of human research participants:
If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

As noted above, sequence read-level data on human participants will be made available through dbGAP, given the potential for brief in confidentiality associated with that data. Accompanying exposure and covariate data will provided in such as way that it cannot be linked to individual participants. We will follow the Safe Harbor standards in line with Section 164.514(c) of the HIPAA privacy rule.

Oversight of Data Management and Sharing

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

Elizabeth Kennedy (PI, ORCID: 0000-0003-0074-2840) will be responsible for the day-to-day oversight of lab data management activities and data sharing. Broader issues of DMS Plan compliance oversight and reporting will be handled by the PI as part of general Emory University stewardship, reporting, and compliance processes.

Planned Research Outputs

Dataset - "RNA-sequencing reads"

RNA sequencing reads and metadata generated from breastmilk exosomes for 50 study participants over three time periods.

Dataset - "Hormone concentration data"

Salivary hormone concentration data from 50 infants at 1 and 6 months of age.

Text - "Code, metadata and de-identified, individual level data"

Information necessary to perform analyses related to this study

Dataset - "Sequence counts"

Processed count matrix and associated metadata for sequencing performed on breastmilk exosomes from 50 participants across 3 time points.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
RNA-sequencing reads	Dataset	2026-01-31	Restricted	NCBI dbGaP		None specified	None specified	No	Yes
Hormone concentration data	Dataset	2026-01-31	Open	Emory Dataverse		None specified	None specified	No	No
Code, metadata and de-identified, individual level ...	Text	Unspecified	Open	Emory Dataverse		None specified	None specified	No	No
Sequence counts	Dataset	2026-01-31	Open	Gene Expression Omnibus		None specified	None specified	No	No