Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D1ZC9V

Title: Transcriptional Dysregulation in the Pathogenesis of NAFLD

Creator: Baran Ersoy - **ORCID:** 0000-0003-4848-5097

Affiliation: Weill Cornell Medicine

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: PA-20-185

Grant: https://grants.nih.gov/grants/guide/pa-files/PA-20-185.html

Template: NIH-Default DMSP

Project abstract:

Obesity-induced lipotoxicity, hepatic glucose production (HGP) and insulin resistance are the primary pathophysiological defects that predispose to non-alcoholic fatty liver disease (NAFLD). Because current management options remain limited, identification of new regulatory mechanisms that govern the maladaptive response to overnutrition will serve to identify novel opportunities for pharmacologic intervention. Metabolic responses to nutritional state are controlled by transcriptionally regulated pathways. Our *long-term goal* is to define how aberrant transcriptional activity can be leveraged for therapeutic purposes. The *objective of this research* is to determine the mechanisms by which the maladaptive activation of Y box-binding protein 1 (Ybx1) culminates in metabolic disease. Our *preliminary data* using mass spectrometry-based proteomics of hepatic nuclei identified Ybx1, a DNA and RNA-binding protein (RBP), as the largest aberration from health. Obesity-induced nuclear localization of Ybx1 by 9-fold was linked to phosphorylation by obesity-activated casein kinase 1 and 2 (CK1/2). Using three independent liver-specific mouse models (shRNA, gRNA and genetic ablation -*YbxLKO*), we demonstrated that Ybx1 strongly promotes hepatic steatosis and HGP by forming a complex with C/EBPα to drive the expression of lipogenic and gluconeogenic factors *Pparγ2* and *Pck1, respectively.* In addition to its regulatory control over gene expression, Ybx1 inhibited the clearance of excess lipids via

mRNA processing of RXR*α* and LDLR, which promote b-oxidation and clear low-density lipoproteins (LDL), respectively. On the other hand, improved insulin sensitivity of *YbxLKO* mice was attributable to specific Ybx1-miRNA interactions. Based on extensive preliminary data, our central hypothesis is that obesity-induced activation of hepatic Ybx1 impairs lipid and glucose homeostasis by both transcriptional and post-transcriptional mechanisms. This will be tested in three specific aims: 1) To identify the mechanism underlying diet-induced activation of Ybx1; 2) To establish the Ybx1-C/EBPα complex as a critical factor in the pathogenesis of NAFLD; 3) To elucidate the Ybx1-RNA interactions that impair lipid and glucose homeostasis. In Aim 1, the roles of CK1/2 will be tested using inhibitors, siRNA and phospho-deficient Ybx1 constructs. The role of Ybx1 in the pathogenesis, progression and reversal of NAFLD will be tested in *Ybx1LKO* mice fed a diet rich in fructose, palmitate and cholesterol. Aim 2 will utilize gel shift and luciferase reporter assays as well as ChIP-seq analyses to establish the lipogenic and gluconeogenic control by the Ybx1-C/EBPα complex. Co-dependency will be tested by knockdown and overexpression of Ybx1 and/or C/EBPα in primary hepatocytes. Aim 3 will test Ybx1-RNA interactions using recombinant protein, RNA constructs as well as RBP-eCLIP analysis. The roles of mRNA and miRNA processing by Ybx1 will be confirmed in hepatocytes using knockdown and locked nucleic acid (LNA), respectively. This is *significant* because Ybx1 represents the most severe nuclear aberration in NAFLD. These studies are expected to establish Ybx1 as a tractable target for the management of NAFLD.

Start date: 09-01-2023

End date: 08-31-2028

Last modified: 07-08-2024

Copyright information:

Transcriptional Dysregulation in the Pathogenesis of NAFLD

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)

In this proposed project, data will be generated via the following methods: Phenotypic analysis of animal models and cell culture systems. Data will be saved and retained in the form of spectrometry analyses, immunoblot images, colorimetric assays, light and epifluorescent microscopy images, real-time quantitative polymerase chain reaction (PCR), tissue pathology images, chromosome immunoprecipitation sequence analysis (ChIP-seq), ultraviolet crosslinked immunoprecipitation sequence analysis (CLIP-seq) and mass spectrometry of proteomics. In cell culture systems, this data will be collected from a minimum of 3 independent experiments, with each independent experiment consisting of two or more groups. In mouse models, this data will be collected from a minimum of 10 mice per group. There will be 6 groups of mice: Conditional Ybx1 knockout mice will be injected with AAV8-TBG-iCRE or AAV8-TBG-LacZ for control at 4, 16 or 32 weeks of age. The total size of the data collected is projected to be 300 GB.

We expect to generate the following data file types and formats during this project: Carl Zeiss microscopic image file (.CZI); images (.TIFF); tabular (.CSV); ChIP-seq and CLIP-seq (FASTA); Proteomics (mzML); Data analysis

Raw data files will be analyzed to generate .XLSX and .PZFX and .PPT files.

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.*

As the proposed experiments become published in scientific journals, the cleaned, item-level spreadsheet data for all variables will also be shared openly, along with example quantifications and transformations from initial raw data. Final files used to generate specific analyses to answer the Specific Aims and related results will also be shared. The rationale for sharing only cleaned data is to foster ease of data reuse.

If any data collection arising from the proposed studies are not shared by publication within one year after the proposed award period ends, the unpublished data will be similarly shared in public domains for ease of access.

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.

As the proposed experiments become published in scientific journals, a README file and data dictionary will be generated and deposited into a repository along with all shared datasets to facilitate the interpretation and reuse of the data. The README file will include method description, instrument settings, RRIDs of resources such as antibodies, model organisms, cell lines, plasmids, and other tools (e.g., software, databases, services), and Protocol DOIs issued from protocols.io. The data dictionary will define and describe all variables in the dataset.

If any data collection arising from the proposed studies are not shared by publication within one year after the proposed award period ends, the unpublished data will be similarly shared in public domains for ease of access.

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.

The raw data generated via imaging will be converted to TIFF file and will not require specialized software for visualization. The raw data generated via ChIP-seq and CLIP-seq will be in the compressed (.GZ) FASTQ (.FQ) format .FQ.GZ. Statistical programs such Python or R can be used to analyze the raw data present in the FASTQ file and converted into FASTA file.

Python and R are free software environment for statistical computing and graphics. Scripts produced through the course of the research will be made publicly available on the lab's GitHub repository, and will be provided as Supplementary files for any publications through a Zenodo-GitHub link. Code will be available no later than when a publication has been accepted. In addition to Zenodo-GitHub link, all data produced by the grant proposal will be recorded in and shared via the electronic notebook system LabArchives, which complies with the DMS requirements (https://www.labarchives.com/nih-2023-data-sharing-requirement/).

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

In accordance with FAIR Principles for data, we will use open file formats (e.g. JPEG, MP4, CSV, TXT, PDF, HTML, etc.) and persistent unique identifiers (PIDs) such as RRIDs for resources (e.g., organisms, plasmids, antibodies, cell lines, software tools, and databases) and DOIs for protocols using protocols.io. The bioimaging community has not yet agreed on a single standard data format that is

generated by all acquisition systems, but we will use OME-Files for data that will be preserved and shared.

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; see Selecting a Data Repository)

All sequence analyses, including RNA-seq, ChIP-seq and CLIP-seq data will be deposited into the "Gene Expression Omnibus" (GEO) of NCBI. All proteomics-related data will be deposited into the "PRIDE" of EMBL-EBI. All other data described above in the “data to be shared” section will be deposited into Zenodo. In addition, all data produced by the grant proposal will be recorded in and shared via the electronic notebook system LabArchives, which complies with the DMS requirements (https://www.labarchives.com/nih-2023-data-sharing-requirement/). International Mouse Phenotyping Consortium (IMPC) does not allow data submission. Therefore, mouse phenotypic analyses will also be stored in Zenodo. The proposed studies do not include human samples and clinical approaches.

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.

We will use Persistent Unique Identifiers (PIDs) to improve data findability across all dissemination outputs. PIDs used will include ORCID iDs for people, DOIs for outputs (e.g., datasets, protocols), Research Resource IDentifiers (RRIDs) for resources, and Research Organization Registry (ROR) IDs and funder IDs for places, as much as possible to make data identifiable and findable. We will also use indexed metadata, such as MeSH terms with a unique URL to make scientific data easily findable. We will keep our ORCID Records up to date with DOIs for our datasets and publications, ROR, and funder IDs to increase findability.

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.

All scientific data generated from this project will be made available as soon as possible, and no later than the time of publication or one year after the end of the funding period, whichever comes first. The duration of preservation and sharing of the data will be a minimum of 10 years after the funding period.

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. See Frequently Asked Questions for examples of justifiable reasons for limiting sharing of data.

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).

Controlled access will not be used. The data that is shared will be shared by unrestricted download.

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).

Proposed research will not produce human data.

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).

Lead PI Baran A. Ersoy, ORCID: 0000-0003-4848-5097, will be responsible for the day-to-day oversight of lab/team data management activities and data sharing. Broader issues of DMS Plan compliance oversight and reporting will be handled by the PI team as part of general stewardship, reporting, and compliance processes.
