Appendix Section B(2): CONCEPT PAPER TEMPLATE

# DUNEDIN MULTIDISCIPLINARY HEALTH AND DEVELOPMENT STUDY

# (The Dunedin Study)

## CONCEPT PAPER TEMPLATE

## (July 2024)

# DUNEDIN STUDY CONCEPT PAPER

**Provisional Paper Title:** Linking gene expression to longitudinal aging through archival blood single-nuclei sequencing

**Proposing Author:** Evan Macosko, Mukund Murthy

**Author's Email:** murthy@broadinstitute.org; emacosko@broadinstitute.org

**P.I. Sponsor:** Avshalom CAspi
(if the proposing author is a student or colleague of an original PI)

**Today's Date: 10 September 2024**

---

## Objective of the project:

The primary objective of this project is to perform single-nucleus RNA sequencing (snRNA-seq) of archival frozen blood samples collected from the Dunedin cohort at Phase 26 and to analyze the resulting gene expression signatures. This project is the deliverable product of the February 2022 Memorandum of Understanding between Prof Richie Poulton and Prof Evan Macosko.

We hope to

1. demonstrate that meaningful biological signals can be recovered from 26-year-old frozen blood, showcasing the viability and reliability of using long-term archived samples for modern transcriptome analyses.
2. demonstrate that the identified gene expression signatures are associated with longitudinal aging measured from the Dunedin Study, namely Pace of Aging, a uniformly weighted combination of slopes of 19 biomarkers between ages 26 and 45. By studying the relationship between the 26 year old transcriptomic data and Pace Of Aging/DunedinPACE, we aim to show that gene expression ascertained in archival frozen blood can be used to uncover molecular mechanisms underlying health and aging.

## Data analysis methods:

*Quality Control*

Our project will involve several key analyses. Following sequencing, which has been completed, we will conduct extensive quality control (QC) measures to filter out low-quality nuclei and study members' blood samples by assessing percent intronic versus percent mitochondrial reads per nucleus. We will normalize and scale the gene expression data to ensure comparability across samples and cluster the data to visualize immune cell types.

Next, we will employ Covarying Neighborhood Analysis (CNA)[1] to measure the association between clinical phenotypes and the distribution of individual participants' samples across transcriptomic neighborhoods.

*Differential Expression Analyses*

For differential expression analysis, we will apply the limmaTrend method to identify genes with significant expression changes associated with clinical phenotypes. Additionally, we will use the TRanscriptome-wide Analysis of Differential Expression (TRADE)[2] framework to estimate the magnitude of perturbational effects on the entire transcriptome, providing a transcriptome-wide impact (TWI) metric to identify clinical phenotypes with significant effects on participant transcriptomes. To ensure the robustness of our TWI estimates, we will jackknife the estimates to approximate error bounds.

First, we will direct one class of differential expression analyses towards objective one – demonstrating the viability and reliability of using long-term archived samples for modern transcriptome analyses. We will validate differential expression signatures for **smoking and cortisol traits** by comparing phenotype-perturbed genes with known biomarkers and gene sets from the literature and databases to assess whether the biological signals we have recovered are valid.

Next, we will target the second objective: given that age of 26 represents a relatively phenotypically-homogenous age before many aging-related declines become apparent, analyzing differential expression at this age with respect to Pace of Aging / DunedinPACE allows us to identify molecular signatures predictive of long-term health outcomes. By focusing on Pace of Aging / DunedinPACE, we can explore how gene expression patterns at age 26 correlate with rates of aging, providing insights into the predictive value of these molecular signatures.

Finally, we will utilize pathway and enrichment analyses, employing gene function relational databases such as SigCOM-LINCS[3], Gene Ontology, and MSigDB[4], along with Gene Set Enrichment Analysis (GSEA)[5] to explore biological pathways emphasized in transcriptional signatures belonging to Pace Of Aging / DunedinPACE.

**Variables needed at which ages:**

Sex, CurrentSmk26, cortis26np, Pace of Aging/DunedinPACE

**Significance of the project (for theory, research methods or clinical practice):**

**For Theory:** This project contributes to the understanding of how young adult gene expression signatures can predict long-term health outcomes, particularly aging. It provides insights into the molecular pathways mediating these processes, with the potential to advance our understanding of aging biology.

**For Research Methods**: The development and validation of the Blood-Seq protocol for archival frozen blood samples demonstrate the feasibility of using historical samples for

modern transcriptome analyses. This method can be applied to other longitudinal cohort studies, expanding the utility of archived samples.

**How the paper will contribute to Māori health advancement and/or equitable health outcomes**

Phase 26 frozen blood from Maori Study members was not shipped to the United States.
We did not perform our Blood-seq measurements on Maori samples. However, the project aims to uncover genomic biomarkers of aging which are related to metabolic health. The identification of early genomic diagnostic markers and therapeutic targets can eventually help in developing personalized health interventions earlier in adult life. For example, the New Zealand Maori population has a higher prevalence of metabolic syndromes, so biomarkers at age 26 of accelerated aging may may be predictive of future metabolic disease and in this way could contribute to Maori health advancement.

**References cited:**

1. Reshef, Y. A. *et al.* Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics. *Nat. Biotechnol.* 40, 355–363 (2022).
2. Nadig, A. *et al.* Transcriptome-wide characterization of genetic perturbations. *bioRxivorg* 2024.07.03.601903 (2024).
3. Evangelista, J. E. *et al.* SigCom LINCS: data and metadata search engine for a million gene expression signatures. *Nucleic Acids Res.* 50, W697–W709 (2022).
4. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425 (2015).
5. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550 (2005).
6. Understanding genetic risk factors for metabolic disease in Maori and Pacific. https://www.hrc.govt.nz/resources/research-repository/understanding-genetic-risk-factors-metabolic-disease-maori-and.