MIDUS
Midlife in the United States
A National Study of Health & Well-Being

MIDUS 3

Genomics Project (P6)

Gene Expression Documentation

February 2024

University of Wisconsin Institute on Aging and UCLA Social Genomics Core Laboratory. NIA Grant #U19 AG051426

Gene Expression Data Documentation

Summary: Gene expression profiling of MIDUS 3 (M3) biomarker study participants was conducted in 2018-2022 and resulted in 3 files of data on expression of genes, two of them involved in assessing the Conserved Transcriptional Response to Adversity (CTRA) RNA profile [1-5] and biological aging (the Senescence-Associated Secretory Phenotype/SASP; DNA Damage Response/DDR; and cell senescence, p16INK4A/CDKN2A), and one providing transcript abundance values for all assayed genes (not just those involved in CTRA):

- M3_P6_RNA_Scores_N747_20240216.sav
  - An SPSS data file that contains CTRA and CDKN2A and SASP and DDR composite scores along with relevant sample quality metrics & RNA covariates
  - Appropriate for most users
  - Available through the MIDUS Colectica Portal (http://midus.colectica.org/)
- M3_P6_RNA_Genes_N747_20240214.sav
  - An SPSS data file that contains expression values for 51 individual CTRA indicator genes, along with relevant sample quality metrics & RNA covariates
  - Appropriate for use by investigators with experience in statistical genetics/genomics
  - Information about accessing this data file can be found here: MIDUS Genomic Repository
- M3_P6_RNA_GeneExpressionLog2Matrix_N672_20240222.txt
  - A text data file that contains expression values for 60,675 distinct gene transcripts (complete genome-wide transcriptome data)
  - Rows denote genes; columns denote samples
  - Data represent log2-transformed transcript abundance values (gene transcript counts per million total human transcriptome-mapped RNA sequencing reads), with values normalized to hold constant 11 standard reference genes and floored at 1 transcript per million (0 log2) to suppress spurious variability
  - Appropriate for use by investigators with experience in statistical genetics/genomics
  - Information about accessing this data file can be found here: MIDUS Genomic Repository

The remainder of this document contains general information about the RNA Gene Expression dataset along with technical and other details about how these data are used to generate CTRA scores.

Variables in the SPSS files are named according to MIDUS conventions (see the Naming and Coding Conventions for MIDUS Refresher and MIDUS 3 survey data https://midus-study.github.io/public-documentation/M3P1/Documentation/MR_M3_NamingAndCodingConventions_20220531.pdf), and the variable names for the gene expression data begin with the unique 3-character set C6R. Variable names for gene transcripts include the gene name. Both SPSS files contain the following common variables to facilitate use of the data:

- Administrative Variables:

    - C6RAVAIL – categorical flag variable indicating if gene expression data is available, and if not, why not.
- Technical variables:
    - C6RPLATE - the batch of samples in which the RNA was assayed
    - C6RMAPPEDREADS - indicates sample RNA quantity (ideally > 5000000)
    - C6RMAPPEPCT - indicates sample RNA quality (% of total RNA sequencing reads that were successfully mapped to the human genome; ideally >= 80% for assay plates 1-30)
    - C6RAVGR - sample transcriptome profile correlation with other samples
- Transcript Variables: there are 59 gene transcript variables organized into the following three groups in the order they are described in below.
    - There are 8 gene transcript variables that mark major leukocyte subsets and are often used as covariates to control for variation in blood cell pool composition. These are also included in the "M3_RNA_Scores" file and appear in the data file adjacent to the Technical variables. The gene name is included in the variable name:
        - C6RCD3E, C6RCD3D, C6RCD4, C6RCD8A, C6RCD14, C6RCD19, C6RFCGR3A, C6RNCAM1
    - The remaining 51 gene transcript variables represent individual CTRA indicator genes that have been used in previous research, including 19 pro-inflammatory genes and 32 genes involved in interferon (IFN)-mediated antiviral responses and antibody production.
        - Pro-inflammatory genes: C6RCXCL8, C6RFOS, C6RFOSB, C6RFOSL1, C6RFOSL2, C6RIL1A, C6RIL1B, C6RIL6, C6RJUN, C6RJUNB, C6RJUND, C6RNFKB1, C6RNFKB2, C6RPTGS1, C6RPTGS2, C6RREL, C6RRELA, C6RRELB, C6RTNF
        - Interferon/antibody genes: C6RGBP1, C6RIFI16, C6RIFI27, C6RIFI27L1, C6RIFI27L2, C6RIFI30, C6RIFI35, C6RIFI44, C6RIFI44L, C6RIFI6, C6RIFIH1, C6RIFIT1, C6RIFIT2, C6RIFIT3, C6RIFIT5, C6RIFITM1, C6RIFITM2, C6RIFITM3, C6RIFITM4P, C6RIFITM5, C6RIFNB1, C6RIGLL1, C6RIRF2, C6RIRF7, C6RIRF8, C6RJCHAIN, C6RMX1, C6RMX2, C6ROAS1, C6ROAS2, C6ROAS3, C6ROASL

For the text file containing transcriptome-wide data, rows are labeled by Gene Symbols (or by Ensembl Stable Gene Identifier number, if no Gene Symbol is available), and columns are labeled by MIDUS Participant IDs. Note that this data matrix does NOT contain the technical variables noted above in the SPSS files (sample quality metrics, cell abundance covariates). Users of this genome-wide text file data set can incorporate those technical variables into their analyses by merging data files.

Details about the technical and gene transcript variables are provided below.

The blood samples used for the gene expression profiling were obtained, along with other samples, as part of a fasting blood draw completed in the morning of the second day of the Biomarker visit. The sample was collected using a BD Vacutainer CPT Tube. Details about collecting and processing this sample are included in the MIDUS 3 Biomarker Project (P4) Blood, Urine, Saliva documentation which is available at ICPSR (https://www.icpsr.umich.edu/web/ICPSR/studies/38837) or via the MIDUS Colectica Portal (http://midus.colectica.org/). The Portal houses interactive documentation for all the publicly available MIDUS projects. The Portal includes search and explore functions, and a custom data extract function. A link to the portal is also available on the MIDUS website (http://midus.wisc.edu/) under QuickLinks.

Generating CTRA Composite Scores

***Background*:** The CTRA is a gene expression program that is up-regulated by sympathetic nervous system activity and involves increased expression of pro-inflammatory genes and decreased expression of genes involved in IFN antiviral responses and antibody production [1-5]. It is one biological pathway through which psychosocial factors might impact physical health (particularly infectious diseases and chronic illnesses fueled by inflammation). There are multiple ways to measure the CTRA (see below for more detail), but one frequently used approach involves a composite score that contrasts average expression of a pre-specified set of genes involved in inflammation with average expression of a pre-specified set of genes involved in antiviral and antibody responses [6]. The "RNA_Genes" file contains gene-specific RNA expression data for 51 genes that have previously been used as indicators of the CTRA profile (19 inflammatory and 32 interferon/antibody) and were used to form the CTRA composite scores in the "RNA_Scores" data file.

If you are interested only in the composite scores, and don't need "item-level" data on each of the 51 individual indicator genes, you will find it easier to use the "RNA_Scores" file. The "RNA_Genes" file is intended for use by individuals experienced in gene expression analysis who may want to employ complex statistical techniques that require "item-level" data on individual genes.

In both SPSS files (RNA_Genes and RNA_Scores), data on the 51 CTRA indicator genes are accompanied by 3 technical quality metrics for each sample (total number of RNA sequencing reads that could be successfully mapped to the human transcriptome; % of total RNA sequencing reads that could be successfully mapped to the human transcriptome; and the Profile Average Correlation of each sample with other samples) and a set of 8 gene transcripts that are often used as covariates to adjust for the effects of varying prevalence of distinct leukocyte subsets within the analyzed blood samples (e.g., differential prevalence of monocytes, B cells, NK cells, CD4+ and CD8+ T cells, etc.).

CTRA indicator gene composite scores

Conceptually, the CTRA indicator composite is simple to understand: compute the average expression of pro-inflammatory genes (call it Inflam), compute the average expression of antiviral/antibody-related genes (IFN) and take their difference: CTRA composite = Inflam – IFN [1-6]. However, gene expression data have several characteristics that complicate their use as "items" for computation of simple average "scale scores." The average level of a gene's expression can vary by 1000-fold or more across genes, as can the range of variation in a given gene's expression. In addition, the distribution of expression values for some genes is markedly non-normal, due to skew, outliers, and frequent 0 abundance values. Gene expression data also show correlations across genes, and the pattern of these inter-gene correlations (or "covariance structure") is often complex (i.e., it is not homogenous across distinct pairs of genes, even within a biologically specific sub-component such as the Inflam or IFN gene set) [7-11]. These distributional characteristics complicate efforts to summarize the general activity of a set of genes by simply averaging over their individual expression values. Alternative approaches for summarizing general activity might include analyzing the association between each gene and an external variable of interest and then pooling those association measures (e.g., by meta-analysis) or explicitly modeling the distributional heterogeneity across genes within the framework of a statistical model (e.g., using a structural equation model that treats genes as heterogenous indicators of an underlying construct, or a mixed effect linear model that treats the contrast-weighted genes as 51 separate repeated measurements on each individual and tests their average association with a predictor of interest while

explicitly modeling the cross-gene heterogeneity in variance and covariance of statistical residuals [7-11]). The gene-specific expression data in this file can be used for such analyses.

The specific "51-gene CTRA indicator gene composite" described above is one way of measuring the CTRA physiological pattern at the level of RNA. There are other ways of measuring the CTRA physiological pattern using RNA (e.g., using different sets of indicator genes, or using bioinformatic measures of pro-inflammatory and interferon-related transcription factor activity) or using other biological measures (e.g., protein-based measures of immune cell development and differentiation; biological assays of inflammation and/or antiviral responses). The CTRA is not equivalent to (or defined by) the specific set of 51 gene transcripts provided here. But a contrast computed across these genes does represent one easily used approach for measuring the CTRA.

How were the gene expression values derived?

Participants in the MIDUS biomarker project provided blood samples from which peripheral blood mononuclear cells were isolated and stored. RNA was later extracted from these stored white blood cells, tested for suitable RNA yield and RNA integrity, and subject to transcriptome profiling by RNA sequencing. To maximize measurement precision and accommodate as many samples as possible, the RNA sequencing approach used a highly efficient mRNA-targeted approach (cDNA library preparation by Lexogen QuantSeq FWD 5' gene counting assay, with sequencing on an Illumina HiSeq 4000 or NovaSeq instrument targeting > 10 million single-strand 65-nucleotide reads/sample). cDNA library preparation was carried out in 96-sample batches indicated by the C6RPLATE nominal variable (controlling for this factor in statistical analyses may help reduce residual variance). Sequence reads were mapped to the consensus human transcriptome and quantified on a per-gene basis using the STAR aligner [12]. Raw read counts for each gene were pre-normalized to transcript rates per million total mapped reads (transcripts per million; TPM), normalized to hold constant the average expression level of 11 standard reference genes [13], log2 transformed (with data floored at 0 log2 = 1 normalized TPM*), and subject to a standard endpoint quality control screen to exclude poor quality data (samples yielding < 5 million RNA sequencing reads or an elevated number of sequencing reads that do not map to the human transcriptome). Data represent log2-transformed normalized TPM values for all samples that passed endpoint quality screening. The 5' gene counting assay was indicated by the condition of the archival samples and does not allow for resolving different isoforms of a given gene (it assays only 65 nucleotides at the 5' end of each transcript). This approach is highly efficient for quantifying the total abundance of all transcripts encoded by a given gene, which is typically of primary interest in behavioral science and disease pathogenesis research.

* Note that for some genes, all or almost all observations are 0 transcript counts per million total transcripts (TPM). These minimally or un-detected gene transcripts are included here to provide data that is most consistent with previous CTRA composite scores based on microarray gene expression profiling, which typically included data from these genes.

References

1. Cole, S.W., *Functional genomic approaches to psychophysiology*, in *Handbook of Psychophysiology*, J.T. Cacioppo, L.G. Tassinary, and G.G. Berntson, Editors. 2016, Cambridge University Press: Cambridge. p. 354-376.
2. Cole, S.W., *Human social genomics.* PLoS Genet, 2014. **10**(8): p. e1004601.

3. Cole, S.W., *Nervous system regulation of the cancer genome.* Brain Behav Immun, 2013. **30 Suppl**: p. S10-8.
4. Irwin, M.R. and S.W. Cole, *Reciprocal regulation of the neural and innate immune systems.* Nat Rev Immunol, 2011. **11**(9): p. 625-632.
5. Slavich, G.M. and S.W. Cole, *The emerging field of human social genomics.* Clin Psychol Sci, 2013. **1**(3): p. 331-348.
6. Fredrickson, B.L., et al., *A functional genomic perspective on human well-being.* Proc Natl Acad Sci U S A, 2013. **110**(33): p. 13684-9.
7. Nelson-Coffey, S.K., et al., *Kindness in the blood: A randomized controlled trial of the gene regulatory impact of prosocial behavior.* Psychoneuroendocrinology, 2017. **81**: p. 8-13.
8. Kohrt, B.A., et al., *Psychological resilience and the gene regulatory impact of posttraumatic stress in Nepali child soldiers.* Proc Natl Acad Sci U S A, 2016. **113**(29): p. 8156-61.
9. Kitayama, S., et al., *Work, meaning, and gene regulation: Findings from a Japanese information technology firm.* Psychoneuroendocrinology, 2016. **72**: p. 175-181.
10. Fredrickson, B.L., et al., *Psychological well-being and the human conserved transcriptional response to adversity.* PLoS One, 2015. **10**(3): p. e0121839.
11. Cole, S.W., et al., *Myeloid differentiation architecture of leukocyte transcriptome dynamics in perceived social isolation.* Proc Natl Acad Sci U S A, 2015. **112**(49): p. 15142-7.
12. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21. PMC3530905
13. Eisenberg, E. and E.Y. Levanon, *Human housekeeping genes, revisited.* Trends Genet, 2013. **29**(10): p. 569-74.