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

MIDUS Refresher

Gene Expression Documentation (P6)

2018 Interim Data Set

January 2021

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

Gene Expression: CTRA and Aging Score Data Documentation

Summary: Gene expression profiling of MIDUS-Refresher (MR) biomarker study participants was conducted in 2017-2018 and resulted in 2 files of data on expression of genes involved in the Conserved Transcriptional Response to Adversity (CTRA) [1-5] and in biological aging (the senescence marker $p16^{INK4a}$/*CDKN2A* and composite scores capturing the senescence-associated secretory phenotype/SASP and DNA damage response/DDR) [6-8]:

- MR_P6_RNAScores_N863_20210106
  - Contains CTRA and CDKN2A and SASP and DDR composites scores along with relevant sample quality metrics & RNA covariates;
  - Appropriate for most users
  - Available through the MIDUS Colectica Portal (http://midus.colectica.org/)

- MR_P6_RNAGenes_N863_20200605
  - Contains expression values for 51 individual CTRA indicator genes, along with relevant sample quality metrics & RNA covariates;
  - Note, this file does not contain individual indicator genes used to create the cell senescence composites scores. Those will be released separately at a later date.
  - Appropriate for use by investigators with experience in statistical genetics/genomics.
  - Information about accessing this datafile can be found here: MIDUS Genomic Repository.

The remainder of this document contains general information about the RNAScores dataset, along with technical and other details about how the scores are generated.

Variables are named according to MIDUS conventions (see the Naming and Coding Conventions included with the MIDUS Refresher Survey documentation), thus the variable names for the gene expression data begin with the unique 4 character set RA6R. Variable names for gene transcripts include the gene name. Both files contain the following common variables to facilitate use of the data:

- Administrative Variables:
  - RA6RAVAIL – categorical flag variable indicating if gene expression data is available, and if not, why not.
- Technical variables:
  - RA6RPLATE - the batch of samples in which it was assayed
  - RA6RRIN - indicates sample RNA integrity
  - RA6RAVGR - sample transcriptome profile correlation with other samples
- CTRA Score Variables: there are two sets of 3 CTRA score variables; one set is centered while the other is z-scored. Details are provided below.
- Aging Score Variables: the data file also contains data on expression of the *CDKN2A* senescence indicator gene ($p16^{INK4a}$) and two sets of 3 Aging score variables; one set is centered while the other is z-scored. Details are provided below.
- Transcript Variables: 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. The gene name is included in the variable name:
  - RA6RCD3E, RA6RCD3D, RA6RCD4, RA6RCD8A, RA6RCD14, RA6RCD19, RA6RFCGR3A, RA6RNCAM1

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 R Biomarker Project (P4) Blood, Urine, Saliva documentation which is available at ICPSR (https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/36901) or via the MIDUS Colectica Portal (http://midus.colectica.org/). The portal houses interactive codebooks for all the publicly available MIDUS projects. The Portal includes search and explore functions, links to documentation, and a custom download function. A link to the portal is also available on the MIDUS website (http://midus.wisc.edu/) under QuickLinks.

IMPORTANT: These data represent a preliminary data release for the MIDUS Refresher sample only. It is customary in genomics research to normalize data across an entire dataset in order to mitigate technical variations related to assay batch. When additional data are collected from the MIDUS 3 cohort, these data may be incorporated into a combined data set and re-normalized. As such, these gene expression values are subject to change as additional gene expression profiling data are collected.

Generating Composite Scores

***Background on CTRA Scores*:** 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 interferon (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 [9]. Conceptually, the CTRA indicator composite is simple: compute the average expression of pro-inflammatory genes (call it Inflam), compute the average expression of antiviral/antibody-related genes (IFNAb) and take their difference: CTRA composite = Inflam – IFNAb [1-5]. However, averaging gene expression values brings some extra complications and those are discussed below (see Indicator gene composite scores).

The MR_P6_RNAScores file contains CTRA composite scores (as well as their separate inflammation and antiviral/antibody sub-components, labeled Inflam and IFNAb), accompanied by 2 technical quality metrics for each sample (RNA Integrity Number and Profile average correlation 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.). This file provides the information that most users would need. More background on the contents of this file is given below.

***Background on Aging Scores*:** Cell senescence is one major biological mechanism by which our bodies "grow old" functionally and become more vulnerable to disease or death [10, 11]. Senescence involves reduced cellular capacity for growth and development, and this typically occurs in response to DNA damage. Cell senescence is associated with increased transcription of genes involved in the cellular DNA Damage Response (DDR; [7]) and increased expression of genes involved in inflammation and growth inhibition (a profile known as the Senescence-Associated Secretory Phenotype, or SASP [6, 8]. The MR_P6_AgingScores file contains data on two SASP summary scores (average expression of 10 SASP indicator genes originally identified by Campisi [6] and subsequently expanded to 57 indicator genes as a result of additional research [8], a DDR summary score comprised of the mean expression of 30 DDR genes [7], and expression levels for the senescence indicator gene, *CDKN2A* (which encodes the cell senescence mediator protein p16INKa) [7]. More background on the computation of these scores is given below in Indicator Gene Composite scores.

Gene-specific data: For those with experience analyzing gene expression data, expression values for each of the individual indicator genes involved in the CTRA composites have been generated. These data are more sensitive and thus have been made available through a separate mechanism (see MIDUS Genomic Repository). More background on the contents of this file is given below in Indicator Gene Composite scores.

Indicator gene composite scores.

The computation of a multi-gene composite score is simple in principle: compute the average expression value across all the indicator genes. However, average gene expression values vary substantially across genes (e.g., by 1000-fold or more). To prevent the average from being dominated by a few highly expressed genes, it is customary to log2 transform gene expression values and center them before averaging (i.e., subtract each log2 expression value from the mean value for that gene, so each gene has an average expression value of 0). This produces a "centered" composite score (e.g., in the case of a CTRA composite score, labeled cCTRA). Gene expression values can also be quite heteroscedastic, and to prevent the average from being dominated by a few highly variable genes, analysts often additionally z-score standardize gene expression data before averaging (i.e., divide the centered gene expression value by the standard deviation of values for that gene). This leads to a "z-transformed composite" (e.g., zCTRA). Likewise, cDDR and zDDR (or cSASP10 and zSASP10, etc.) are overall composite scores formed by averaging centered gene expression values or z-score standardized gene expression values for the DDR, SASP10, etc. indicator gene sets. In the case of the bipolar CTRA score, which involves both a positive component (Inflam) and a negative component (IFNAb), positive and negative subcomponent scores are first formed by averaging expression values for each component (e.g., 19 pro-inflammatory indicator genes forming the Inflam score; and 32 interferon/antibody-related indicator genes forming the IFNAb score), and then subtracting the latter from the former [9]. For those interested in separate effects on Inflam and IFNAb gene sets, values of those 2 sub-components are also supplied (in both centered and z-standard metrics).

Why are there 2 sets of composite scores (c and z)?

Unlike many of the measures used in behavioral science, gene expression data are highly heteroscedastic. The magnitude of variation in RNA transcript abundance varies by > 10-fold across genes (even after log2 transformation, as done here) [12-16]. This means that a composite score comprised of many genes will often be dominated numerically by a small fraction of those genes with the largest variability across samples. Depending on the purpose of the analysis, this may or may not be desirable. The heteroscedasticity is a real biological phenomenon; it is not due to differential measurement error and may reflect important physiological information. Or it may not, depending on the purposes of the analysis. When it is helpful to retain the heteroscedasticity is an open question. Until that question is resolved for any specific application, it may be helpful to have both types of composite score available. The c composite retains the heteroscedasticity; the z composite eliminates it.

Which composite score should be used?

For most applications, that answer is not known at this time. Either can be considered valid based on current knowledge. It probably makes sense to try both and see how your substantive results vary. In many cases there will be little difference (e.g., zCTRA and cCTRA composite scores correlate r > .95 in this sample). However, in other cases the differences may be material (e.g., zSASP10 correlates only around r = .68 with cSASP10 in this sample).

Ultimately, the relative validity of the c vs. z indicator scores depends on their respective correlations with a valid and technically independent measure of physiology. For example, the "CTRA" is a physiological pattern involving increased expression of pro-inflammatory genes in general, and decreased expression of Type I interferon (antiviral) and antibody-related genes in general. The specific "51-gene CTRA indicator gene composite" provided here is one way of measuring that 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 types of 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 analyzed here. But these scores do represent one easily used approach for measuring the CTRA. The same general principles apply to other composite scores. The SASP10, SASP57, DDR, and CDKN2A scores all represent different ways of measuring cell senescence, and none of them captures perfectly the functional characteristic of a cell being resistant to proliferation and development.

Another approach to analyzing multi-gene composites

As noted above, the reason separate c and z summary scores are provided is to address the complications that arise from heteroscedasticity. However, neither of these approaches is ideal, for biological reasons explained below. Both score types have advantages and disadvantages, and the ideal RNA-based measurement of the CTRA, DDR, or SASP patterns would not involve the computation of any single-number composite score at all; it would treat the multiple (heteroscedastic) indicator genes as a set of multiple distinct indicator variables, each of which provides some information about the CTRA or DDR or SASP pattern, but is also influenced by a wide variety of other influences (e.g., microbial exposures, other physiological processes, other biobehavioral processes, medical conditions, etc.). In this setting of heterogenous "nuisance" variance, standardizing the total variance across genes may have the unintended effect of increasing variability in the magnitude of "true score" (e.g., CTRA-related, or senescence-related) covariation between the indicator genes and the external predictor variable/s of interest.

In addition to the effects of heteroscedasticity, 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 composite) [12-16].

In the presence of heterogenous variance and covariance among genes, single-number z and c composite scores provide sub-optimal power to detect true associations that would be observed across a set of individual indicator genes. Using these scores may lead analysts to miss some associations that are truly present and might have been identified with a more sensitive analysis (i.e., false negative results), but will generally not increase the risk of a false positive error. It is possible to analyze data on multiple indicator genes in other ways that do not incur the problems involved in summary score computation (e.g., by treating them as separate repeated measurements on each individual and testing their average association with a predictor of interest while explicitly modeling the cross-gene heterogeneity in variance and covariance of statistical residuals [12-16]). These more complex analyses require gene-specific expression values which are provided in the "Genes" series of ancillary files (e.g., "MR_P6_RNAGenes") which will be made available via a different method as noted above.

<u>How were the gene expression values derived?</u>

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 instrument targeting > 10 million single-strand 65-nucleotide reads/sample). cDNA library preparation was carried out in 96-sample batches indicated by the RA6RPLATE 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 [17]. Raw read counts for each gene were normalized to transcript rates per million total mapped reads (transcripts per million; TPM), log2 transformed (with data floored at 0 log2 = 1 TPM*), and subject to a standard endpoint quality control screen to exclude aberrant data (r < .85 correlation of sample-

specific transcriptome profile with other profiles). Data represent log2-transformed 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 composite scores based on microarray gene expression profiling, which typically included data from these genes.

References

1. Cole SW. (2013). Nervous system regulation of the cancer genome. *Brain, Behavior and Immunity*, 30 Suppl, S10-18. https://doi.org/10.1016/j.bbi.2012.11.008
2. Cole SW. (2014). Human social genomics. *PLoS Genetics*, 10(8), e1004601. https://doi.org/10.1371/journal.pgen.1004601
3. Cole SW. (2016). Functional genomic approaches to psychophysiology. In: Cacioppo JT, Tassinary LG, Berntson GG, editors, *Handbook of psychophysiology* (4th ed., pp. 354-376). Cambridge University Press.
4. Irwin MR, Cole SW. (2011). Reciprocal regulation of the neural and innate immune systems. *Nature Reviews Immunology*, 11(9), 625-632. https://doi.org/10.1038/nri3042
5. Slavich GM, Cole SW. (2013). The emerging field of human social genomics. *Clinical Psychological Science*, 1(3), 331-348. https://doi.org/10.1177/2167702613478594
6. Campisi J. (2005). Senescent cells, tumor suppression, and organismal aging: Good citizens, bad neighbors. *Cell*, 120(4), 513-522. https://doi.org/10.1016/j.cell.2005.02.003
7. Carroll JE, Cole SW, Seeman TE, Breen EC, Witarama T, Arevalo JMG, . . . Irwin MR. (2016). Partial sleep deprivation activates the DNA damage response (DDR) and the senescence-associated secretory phenotype (SASP) in aged adult humans. *Brain, Behavior and Immunity*, 51, 223-229. https://doi.org/10.1016/j.bbi.2015.08.024
8. Coppé JP, Desprez PY, Krtolica A, Campisi J. (2010). The senescence-associated secretory phenotype: The dark side of tumor suppression. *Annual Review of Pathology*, 5, 99-118. https://doi.org/10.1146/annurev-pathol-121808-102144
9. Fredrickson BL, Grewen KM, Coffey KA, Algoe SB, Firestine AM, Arevalo JM, . . . Cole SW. (2013). A functional genomic perspective on human well-being. *Proceedings of the National Academy of Sciences*, 110(33), 13684-13689. https://doi.org/10.1073/pnas.1305419110
10. Finch CE. (2007). The biology of human longevity: Inflammation, nutrition, and aging in the evolution of life spans. Academic Press. https://doi.org/10.1016/B978-0-12-373657-4.X5000-4
11. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. (2013). The hallmarks of aging. *Cell*, 153(6), 1194-1217. https://doi.org/10.1016/j.cell.2013.05.039
12. Cole SW, Capitanio JP, Chun K, Arevalo JMG, Ma J, Cacioppo JT. (2015). Myeloid differentiation architecture of leukocyte transcriptome dynamics in perceived social isolation. *Proceedings of the National Academy of Sciences*, 112(49), 15142. https://doi.org10.1073/pnas.1514249112
13. Fredrickson BL, Grewen KM, Algoe SB, Firestine AM, Arevalo JMG, Ma J, Cole SW. (2015). Psychological well-being and the human conserved transcriptional response to adversity. *PLoS ONE*, 10(3), e0121839. https://doi.org/10.1371/journal.pone.0121839
14. Kitayama S, Akutsu S, Uchida Y, Cole SW. (2016). Work, meaning, and gene regulation: Findings from a Japanese information technology firm. *Psychoneuroendocrinology*, 72, 175-181. https://doi.org/10.1016/j.psyneuen.2016.07.004

15. Kohrt BA, Worthman CM, Adhikari RP, Luitel NP, Arevalo JM, Ma J, . . . Cole SW. (2016). Psychological resilience and the gene regulatory impact of posttraumatic stress in Nepali child soldiers. *Proceedings of the National Academy of Sciences*, 113(29), 8156-8161. https://doi.org/10.1073/pnas.1601301113
16. Nelson-Coffey SK, Fritz MM, Lyubomirsky S, Cole SW. (2017). Kindness in the blood: A randomized controlled trial of the gene regulatory impact of prosocial behavior. *Psychoneuroendocrinology*, 81, 8-13. https://doi.org/10.1016/j.psyneuen.2017.03.025
17. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, . . . Gingeras TR. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21. https://doi.org/10.1093/bioinformatics/bts635