Discovery of Gene Expression Signatures Using Machine Learning: Social Isolation and Genetic Expression in Adolescence and Young Adulthood

Brandt Levitt[a], Lauren Gaydosh[b], Mike Shanahan[d], Steve Cole[c], Kathleen Mullan Harris[a]
[a] University of North Carolina at Chapel Hill, [b] Vanderbilt University, [c] University of California Los Angeles, [d] University of Zurich

Abstract

A large literature has identified social isolation as an important psychosocial determinant of health, but the biological mechanisms that explain this connection are relatively unknown. We address this gap using genome-wide transcriptome data from the National Longitudinal Study of Adolescent to Adult Health (Add Health) to examine whether social isolation operates through gene expression of innate and adaptive immune responses within the stress process system. We expand upon previous work that identified conserved immunological genes as important mediators of poor health outcomes in socially isolated individuals. We construct a quantitative measure of social isolation across multiple contexts. We use regression models and machine learning algorithms to develop a gene expression signature correlated with social isolation and seek to better understand this connection by analyzing genetic regulatory features and immunological cell subsets to identify causal patterns that explain the biological processes that make social isolation a risk factor for poor health.

Introduction

It has been well established that social connections are good for your health and social isolation is harmful to health. A large literature has shown that good social relationships are associated with improved physical and mental health (Berkman et al. 2000; Cohen and Janicki-Deverts 2009; George et al. 1989; House, Landis, and Umberson 1988; Penwell and Larkin 2010; Semen 1996; Smith and Christakis 2008; Umberson and Montez 2010; Yang et al. 2016). Social ties, embeddedness in social networks, and engagement in social life have been found to boost self-esteem, protect against illness, and facilitate coping with stress and injury or disease. Social isolation and the lack of social connections are detrimental to health (Cacioppo and Hawkley 2003; Holt-Lundstad et al. 2015). Living alone, having few social network ties, and having infrequent social contact are all markers of social isolation. The social stress model posits that the stress caused by social isolation carries negative consequences, primarily for mental health (Pearlin et al 1981).

Most research on the relationship between social isolation and health has focused on older and/or aging populations (e.g., Seeman et al. 1987; Holt-Lundstad et al. 2015). This focus makes sense because later life is often characterized by increasing social isolation and stressful transitions, including children moving away, retirement, bereavement, and the onset of chronic conditions. Social networks and social support are important for managing and coping with these stressful transitions. There has been less attention to the association between social relations and health and its biological linkages during the early stage of the life course, yet physiological response to stress related to the profound developmental, social, and emotional transitions young people experience in adolescence can be equally consequential for health trajectories set in early life (Yang et al. 2016). Indeed, recent public attention to the role of social isolation and lack of social networks in youth depression, violence, and victimization (i.e., bullying) demonstrates that research needs to start earlier in the life course to understand how adolescents' social connections shape their social affiliation behaviors in adulthood and matter for health across the life course (Hall-Lande et al. 2007; Faris and Felmlee 2011; Ladd and Ettekal 2013).

Social isolation is an important psychosocial determinant of health that may operate through gene expression of innate and adaptive immune responses within the stress process system. Social isolation is associated with health risk behaviors linked to stress, including smoking, physical inactivity, obesity, and poor sleep (Cacioppo et al. 2002; Hawkley, Thisted, and Cacioppo 2009; Holt-Lundstad et al. 2010; Theeke 2010). Biological risk factors are also elevated by social isolation, including higher blood pressure, inflammation, lipids profiles, and poorer immune functioning (Grant, Hamer and Stepoe 2009; Hawkley and Cacioppo 2010; Pressman et al. 2005). Indeed, multiple studies have documented worsened medical outcomes among those identified as being objectively or subjectively socially-isolated in a myriad of conditions including breast cancer metastasis and progression (Bower at al. 2016), hematopoetic stem cell transplantation (Knight et al. 2016), and lentivirus infection (Cole et al. 2015a).

Genetic Underpinnings

Social processes may impact biology through modulation of neuroeffector processes such as the hypothalamus-pituitary-adrenal axis and the sympathetic nervous system. Previous studies have determined a subset of genes that are up- or down-regulated in individuals experiencing social isolation that may serve as a useful biomarker for these poor behavioral, health, and medical outcomes (Cole et al. 2015b). This 53-gene panel correlated to poor health outcomes has been termed the "conserved transcriptional response to adversity" (CTRA). The CTRA advances our understanding of this social phenotype by ascribing multiple biological mechanisms to social adversity including the depression of anti-inflammatory pathways, promotion of proinflammatory cellular signals, and impairment of adaptive immune responses. Further, these genetic phenomena have been shown to be most prevalent in a specific immunological cell subset of leukocytes including monocytes and natural killer cells. Indeed, real and perceived social isolation remains a serious concern for the health and well-being of individuals in all stages of the life course and its genetic underpinnings are beginning to be understood (Qualter et al. 2015; Goossens et al. 2015).

The studies that identified the CTRA as a prognostic indicator for poor health outcomes were limited in scope and performed with a relatively small sample size due to their exploratory nature. With these constraints, analysis was focused on a small subset of genes linked to inflammatory outcomes. However, in this project, we use data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a large, nationally-representative population-based sample to refine and expand upon existing work. Add Health interviewed a sample of >20,000 adolescents in 1995 who have been followed into adulthood with its most recent interview in 2016-18 when the cohort was in their late 30s. Multilevel and longitudinal social, behavioral, environmental, biological and genomic data have been collected over the 20+ years of observation. This project takes advantage of the particularly large sample size of Add Health participants who have undergone transcriptomic sequencing in the most recent wave of data collection in 2016 (N=1,132). Interrogation of transcript abundance in these individuals permits the discovery of gene expression patterns that correlate with life course measures of social isolation. Importantly, this study design allows for the discovery of novel genetic associations through an unsupervised clustering of genes associated with social isolation while remaining agnostic to current notions of which biological pathways are most important to the isolation phenotype. An understanding of how these genetic mechanisms affect health would be useful for prognostic, diagnostic, and preventative interventions and will further advance the field of social genomics.

Research Objective

We measure social isolation objectively based on the structural or quantitative aspect of social relations in both adolescence and young adulthood, reflecting the degree of social contacts that individuals have for social support and assistance in times of stress (Berkman and Glass 2000; Yang et al. 2016). We exploit the multilevel design of Add Health to create an index of social isolation across multiple domains of young people's lives, including connections and activities with friends, parents, schoolmates, romantic partners, and within religious institutions and the community. These domains for social contact serve to promote social engagement and have been used in recent studies identifying genetic variants linked to social isolation in a broad-based population sample (Day, Ong and Perry 2018).

Utilizing this social isolation index and transcriptional profiling of our longitudinal sample, we define a gene signature pattern that can be used to prognostically identify individuals who experienced social isolation in adolescence and young adulthood. We then use this set of genes associated with social isolation to interrogate the biology of the social isolation phenotype by identifying the cis regulatory genetic elements (i.e., factors that bind the promoter and alter the abundance of their target gene RNA) common to many of the genes. Further, we use transcript origin analysis to determine if this genetic profile is preferentially expressed in certain immunological cell subsets most responsive to social stress and biologically relevant to inflammatory processes (Irwin and Cole 2011). Thus, our paper explores a fundamental social exposure of isolation from human contact that is vitally important to human health by making use of a unique, large sample with transcriptional data to define a novel gene signature pattern that may explain the genetic and biological mechanisms by which social isolation is related to health.

Data and Measures

We use data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a nationally representative study of adolescents in grades 7-12 in 1994-95 (Wave I) in the U.S. who have been followed with four additional waves of interviews in 1996 (Wave II), 2001-02 (Wave III), 2008-09 (Wave IV), and 2016-18 (Wave V) when the cohort was aged 32-42. Wave V data collection is ongoing through 2018, but a preliminary subset of the Wave V sample (Sample 1, N~3800) was released in 2017 representing about 1/3rd of the eventual Wave V sample that will be released in 2019. At Wave V venous blood was collected in a PAXgene tube for transcriptome-wide profiling and mRNA on 1132 participants from Sample 1 in Wave V have been analyzed.

Intracellular RNA was harvested, exposed to quality checks and sequenced. The resulting data was filtered for read quality and mapped to the human reference genome. RNA abundance for each gene was then calculated and used for further analytical exercises. Future analyses will dissect the RNA abundance counts for each gene into their respective transcripts that may identify unique and interesting effects of the different isoforms for each gene. The RNA data used in this paper includes 1132 participants and identifies 60,000 unique transcripts that map to known genes, in aggregate. A representative set of 5000 of these genes were identified for which there are sequencing data present for all participants; thus ensuring that any genetic expression signature is robust enough to be applied across a substantial spectrum of genetic diversity. These 5000 selected genes each include an individual abundance score based on the intensity of the signal for each participant that can then be used to determine an association between that gene and the social isolation phenotype.

Add Health was designed to study the effects of the social contexts of adolescent life on the health and behavior of adolescents and their outcomes in adulthood. The innovative design allows us to measure the extent to which young people are socially integrated/isolated within the multiple contexts of their lives, including the family, peer, school, and community contexts (Harris 2010). For participants with transcriptome data, we construct a binary indicator of social isolation within each of four contexts: family, peers, school and the community. Isolation is indicated at approximately the bottom quartile of the sample distribution on i) the number of activities with friends (3 or less out of 10 activities); ii) number of activities with parents (3 or less out of 20 activities); iii) school cohesion index (0 or 1 on index total of 3 based on feeling close to people at school, feeling a part of the school, and feeling happy to be at school); and iv) no religious service attendance during the year. We then sum the dichotomous isolation indicators in each context to construct a cumulative index of social isolation ranging from 0 to 4. Higher scores indicate greater isolation (e.g., less engagement with parents, friends, school, and religious institutions).

We construct a similar social isolation index during young adulthood at Wave IV (ages 24-32), focusing on relationships with romantic partners, friends and within the community. While our preliminary analysis presented here is based only on the social isolation index during adolescence, we plan to also analysis gene expression signatures with the isolation phenotype in young adulthood, as well as a longitudinal cumulative isolation measure for the period from adolescence to young adulthood. Analyses use several key demographic covariates including age, sex, and race (non-Hispanic white, non-Hispanic black, non-Hispanic Asian, Native American, and Hispanic).

Analysis Plan

We conduct descriptive analysis examining the social isolation index by sex, race, and age. Principal component analyses are performed to disentangle underlying population structure from legitimate variations in social isolation scores. The contribution of the CTRA genes are analyzed by implementing an ordinal regression model between the categories of social isolation (0-4) and the relative expression levels in these 53 genes.

The contribution of individual leukocyte classes to this gene expression pattern are analyzed by transcriptome origin analysis. Some transcripts are more likely to be expressed in high abundance by certain immune cell subsets and can thus be used as a marker of increased monocyte, NK cell, B cell, T cell or neutrophil richness. Using these cell type-specific marker sets, the relative contribution of each cell to the social isolation phenotype are determined by comparing the correlation of gene expression indicative of each cell with increased CTRA and social isolation scores.

Novel gene sets that associate with social isolation are determined and characterized to expand upon those represented in the CTRA. The statistical power of this study allows a greater number of transcripts to be connected with social isolation than in previous studies and increases the usefulness of this as a biomarker for the trait and its poor health outcomes. Machine learning approaches, specifically support vector machines with recursive feature elimination, are used to determine which transcripts are most predictive of the social isolation score. The identified genes are further analyzed to characterize the cis regulatory elements common to their promoters. In this way, transcription factors that underlie the genetic regulation of these transcripts and associate with social isolation might be uncovered. We plan to use the current sample as a training data set for discovery of gene expression signatures associated with social isolation detailed here, and additional Wave V RNA sample available later this year as a validation data set.

Preliminary Results

We begin with bivariate analyses of social isolation in the Add Health analytic sample (N=1132) by demographic characteristics. Significant variation in social isolation by demographic factors such as age, gender, and race/ethnicity, may necessitate stratification of the sample in further gene expression signature analyses. Figure 1 shows the distribution of social isolation scores on these covariates. Figure 1a includes a breakdown of the number of individuals at each level of social isolation by race/ethnicity using five mutually exclusive categories of non-Hispanic white, non-Hispanic black, Hispanic, Asian, and Native American. Figure 1b represents the same data showing the proportionate distribution of individuals at each social isolation level. Non-Hispanic White participants and Asians tend to have a smaller percentage in the most isolated categories compared to the other race and ethnic groups.

A similar analysis is shown for gender in Figures 1c and 1d and for age in adolescence (Wave I) in Figures 1e and 1f. Males show a slightly higher level of social isolation with a higher percentage in isolation categories 3 and 4 compared to females which is consistent with published literature (Yang et al. 2016). Age differences are minor except for those aged 19 and 20 years old at the beginning of the study who have somewhat higher isolation levels, but Figure 1e indicates the relatively small sample size of these older high school students. In summary, analyses of social isolation scores by demographic covariates show modest differences by race/ethnicity, gender and age that do not warrant statistical adjustment in further gene expression analyses.

To create a baseline gene signature prior to machine learning algorithm training, an ordinal regression model was estimated for each gene for each participant, using each gene as an independent variable and the social isolation score for that person as an outcome. The 100 genes that most consistently and robustly correlate with social isolation score were selected as candidate hits for more detailed analysis. Figure 2a -2h show the top eight hits with their normalized RNA expression levels grouped by individuals at each level of the social isolation score. Each color on the violin plots correspond with a different social isolation score and the y-axis indicates predicted social isolation level in arbitrary units based upon the formula derived in the regression model and the relative gene expression level. Some of these genes were down-regulated while others were up-regulated as a correlate of increasing social isolation scores. The regression model accounts for this and generates a score in an increasing fashion regardless of the directionality of the RNA expression levels. The distribution of the gene expression values is quite extensive, which is reflective of the noisy nature of transcriptomic data. However, there are clear trends in the median expression levels for each gene shown. Median predicted social isolation scores as calculated by the regression model and gene expression levels are shown for the top 20 genes without their distributions in Figures 3a-3d. Several of these genes yield a multiple fold change in expression level across levels of social isolation.

A cursory examination of the location of these top 100 genes indicates that we were not enriching for genes in sex chromosomes as a function of the slightly increased social isolation scores among females. Figure 4 shows the number of the top 100 genes from each chromosome, and it is clear that the mitochondrial genome and X-chromosome do not yield an unexpectedly high number of candidate genes.

Any individual gene among the top 100 candidate genes produces a prediction of social isolation, however, the statistical certainty is lacking. We therefore grouped the top hits into gene sets of varying sizes and used these gene signature patterns to predict social isolation scores. Figures 5a and 5b show the degree to which these gene expression signatures predict social isolation. In every case, the signatures do a better job of predicting social isolation scores than using a single gene. When 30 or fewer genes comprise the signature, it is difficult to distinguish between participants at 0 and 1 levels of social isolation. However, the remaining levels of social isolation are easier to differentiate. In all cases, individuals at the highest level of social isolation are easy to separate from the remainder of the sample. When 40 or more genes comprise the signature, a more robust distinction can be made between each level of social isolation, and this relationship remains linear until it reaches the highest level of social isolation. Thus, these gene expression signatures may serve as an indicator that individuals have experienced social isolation in adolescence.

In sum, our preliminary findings demonstrate feasibility for the machine learning prediction of gene signature by validating that the social isolation phenotype is not meaningfully altered by statistical covariates such as age, race, and gender by using non-automated forms of statistical analysis to create a precursor gene expression signature that correlates with social isolation scores.

Individual genes within this precursor gene expression signature were analyzed to verify that no single gene can be used to predict social isolation by itself and that the gene signature does not preferentially associate with sex chromosomes. We furthermore find that genetic signatures of 40 or more genes robustly distinguish individuals at different levels of social isolation.

Future Directions

These gene signatures will be tested against a validation RNA data set of Wave V Add Health participants (available later this year) to test the statistical limits of its predictive power on a naive population. In addition, an analogous gene expression signature will be determined by support vector machine algorithms and compared to this non-automated gene expression signature in predictive power. Further, the gene signatures will be studied to identify common cis regulatory element binding sites in the promoter regions of their genes and in intronic regions that might inform their regulation and splicing which modify the abundance and transcript identity of RNA arising from each gene. Finally, these gene signatures will be associated with distinctive transcripts arising from certain immune cell subsets to determine whether they are more likely to be expressed in cell types known to be modified by social isolation and act as effectors in immune disease. This last step will provide new understanding of potential links between social isolation, gene expression and biological processes leading to health. We expect to complete these steps by the time of the PAA meetings in April.

References

Berkman LF, Glass T, Brissette I, & Seeman TE (2000) From social integration to health: Durkheim in the new millennium. *Social science & medicine (1982)* 51(6):843-857.

Bower JE, Shiao SL, Sullivan P, Lamkin DM, Atienza R, Mercado F, Arevalo J, Asher A, Ganz PA, Cole SW. Prometastatic Molecular Profiles in Breast Tumors From Socially Isolated Women. JNCI Cancer Spectr. 2018 Jul;2(3):pky029.

Cacioppo JT & Hawkley LC (2003) Social isolation and health, with an emphasis on underlying mechanisms. *Perspectives in biology and medicine* 46(3 Suppl):S39-52.

Cole SW, Capitanio JP, Chun K, Arevalo JM, Ma J, Cacioppo JT. Myeloid differentiation architecture of leukocyte transcriptome dynamics in perceived social isolation. Proc Natl Acad Sci U S A. 2015 Dec 8;112(49):15142-7.

Cole SW, Levine ME, Arevalo JM, Ma J, Weir DR, Crimmins EM. Loneliness, eudaimonia, and the human conserved transcriptional response to adversity. Psychoneuroendocrinology. 2015 Dec;62:11-7.

Day FR, Ong KK, Perry JRB. Elucidating the genetic basis of social interaction and isolation. Nat Commun. 2018 Jul 3;9(1):2457

Faris, Robert and Diane Felmlee. 2011. “Status Struggles: Network Centrality and Gender Segregation in Same- and Cross-Gender Aggression.” *American Sociological Review* 76(1):48-73.

Goossens L, van Roekel E, Verhagen M, Cacioppo JT, Cacioppo S, maes M, Boomsma DI. The genetics of loneliness: Linking evolutionary theory to genome-wide genetics, epigenetics and social science. Perspectives on Psychological Science 2015 10(2):213-226.

Harris KM (2010) An integrative approach to health. *Demography* 47(1):1-22.

Holt-Lunstad J, Smith TB, Layton JB. Social relationships and mortality risk: a meta-analytic review. PLoS Med. 2010 Jul 27;7(7):e1000316.

Holt-Lunstad J, Smith TB, Baker M, Harris T, Stephenson D. Loneliness and social isolation as risk factors for mortality: A meta-analytic review. Perspectives on Psychological Science 2015 10(2):227-237.

House JS, Landis KR, & Umberson D (1988) Social relationships and health. *Science (New York, N.Y.)* 241(4865):540-545.

Irwin MR, Cole SW. Reciprocal regulation of the neural and innate immune systems. Nat Rev Immunol. 2011 Aug 5;11(9):625-32.

Knight JM, Rizzo JD, Logan BR, Wang T, Arevalo JM, Ma J, Cole SW. Low Socioeconomic Status, Adverse Gene Expression Profiles, and Clinical Outcomes in Hematopoietic Stem Cell Transplant Recipients. Clin Cancer Res. 2016 Jan1;22(1):69-78.

Pearlin, Leonard I., Elizabeth G. Menaghan, A. Lieberman Morton, and Joseph T. Mullan. 1981. “The stress process.” *Journal of Health and Social Behavior* 22 (4):337-356.

Qualter P, Vanhalst J, Harris R, Van Roekel E, Lodder G, Bangee M, Maes M, Verhagen M. Loneliness across the life span. Perspectives on Psychological Science 2015 10(2):250-264.

Seeman, Teresa E., George A. Kaplan, Lisa Knudsen, Richard Cohen, and Jack Guralnik. 1987. “Social network ties and mortality among the elderly in the Alameda County Study.” *American Journal of Epidemiology* 126 (4):714–723.

Smith, Kirsten P. and Nicholas A. Christakis. 2008. “Social Networks and Health.” *Annual Review of Sociology* 34:405-429.

Yang, Claire Yang, Courtney Boen, Karen Gerken, Ting Li, Kristen Schorpp, and Kathleen Mullan Harris. 2016. “Social relationships and physiological determinants of longevity across the human life span.” *Proceedings of the National Academy of Sciences* 113(3): 578-583.

DOI: 10.1073/pnas.1511085112 PMCID: PMC4725506

Figure 1 Social isolation scores by Demographic Characteristics

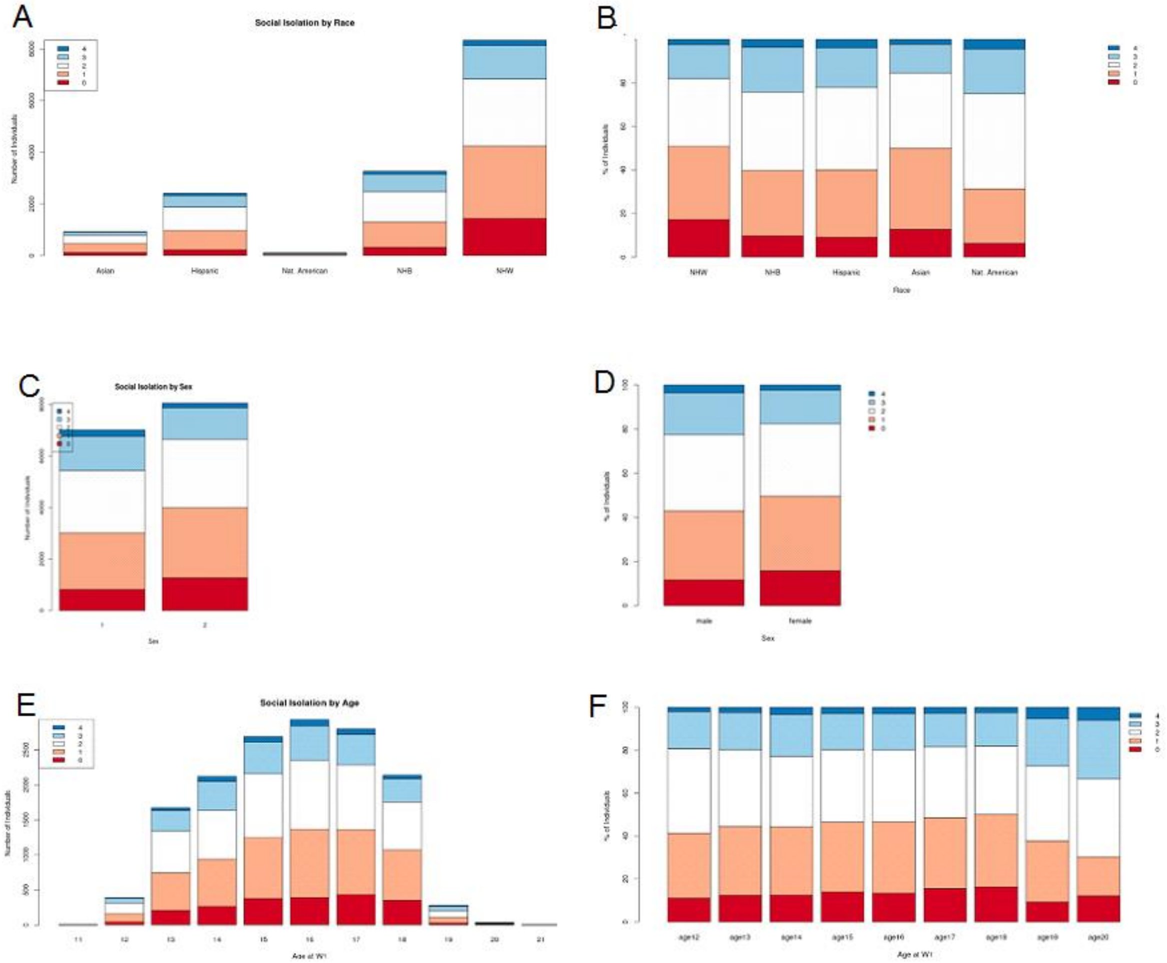

Figure 1. Social isolation scores by race (A and B), gender (C and D) and age at first interview (E and F). A, C, and E show total counts of participants while B, D, and F show percentage of participants in each social isolation category. Social isolation scores 0 and 1 are shades of red (low social isolation), 2 is white and 3 and 4 are shades of blue (high social isolation).

Figure 2. Distribution of calculated social isolation scores for top 8 gene candidates.

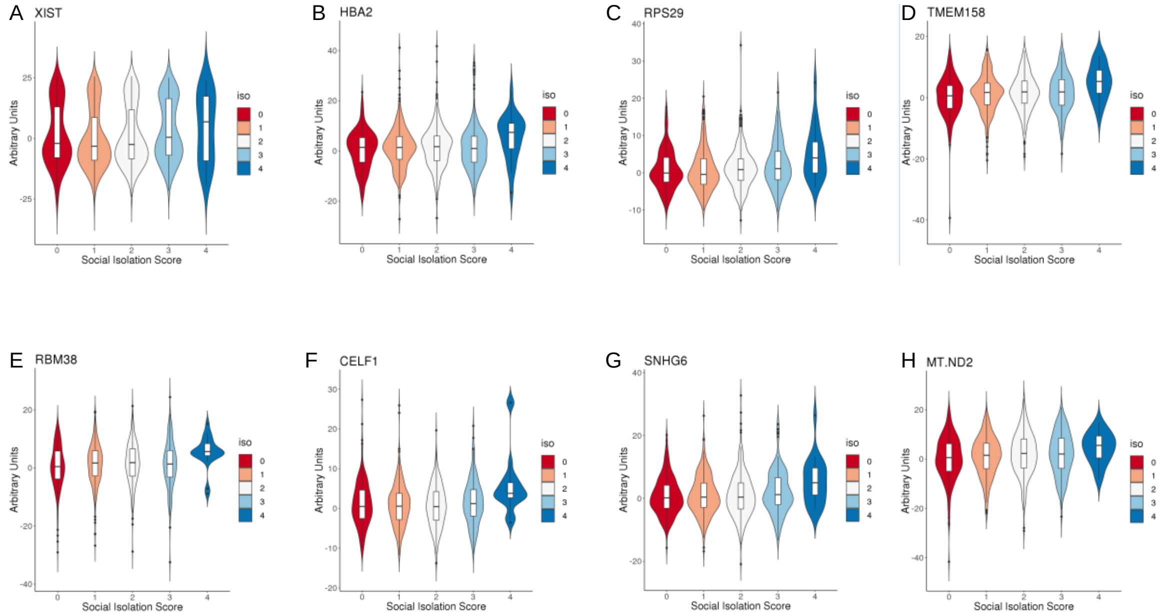

Figure 2. Predicted social isolation from regression model for top gene candidates. Social isolation scores 0 and 1 are shades of red (low social isolation), 2 is white, and 3 and 4 are shades of blue (high social isolation). Each social isolation score category contains a violin plot along with a box and whisker plot indicating the median, the 25th percentile, and 75th percentile. Each panel is a different gene showing A XIST, B HBA2, C RPS29, D TMEM158, E RBM36, F CELF1, G SNHG6 and H MT.ND2.

Figure 3. Chromosomes associated with top 100 gene candidates.

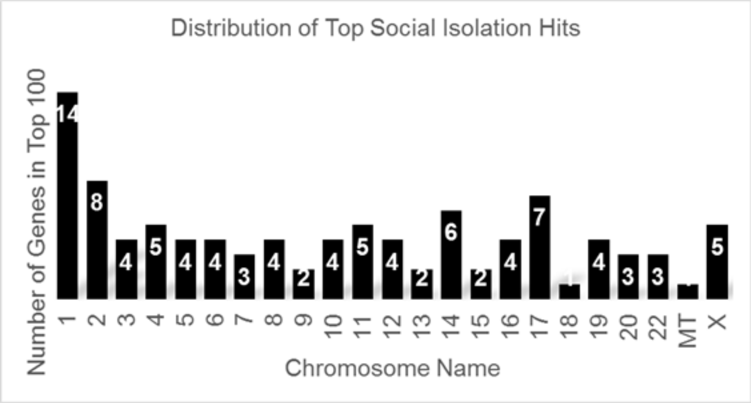

Figure 3. Distribution of the number of gene candidates in the top 100 by chromosome with which they are associated.

Figure 4. Median predicted social isolation scores for top 20 gene candidates

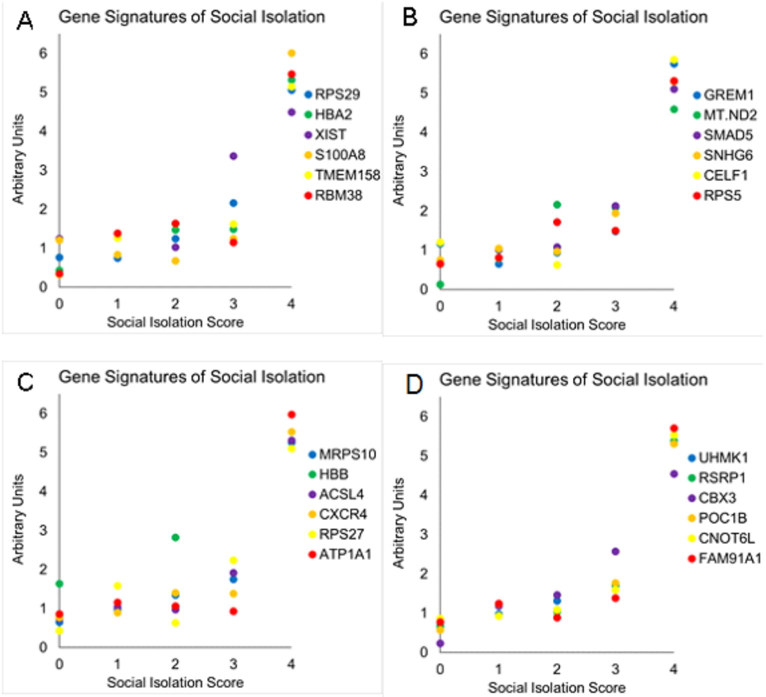

Figure 4. Predicted vs actual social isolation score. X axis shows actual social isolation scores from observed data. Y axis shows the predicted isolation score value based upon the gene expression level for each of the top 20 genes. The figure is divided into panels of genes for ease of presentation of results.

Figure 5. Gene signatures correlated to social isolation score.

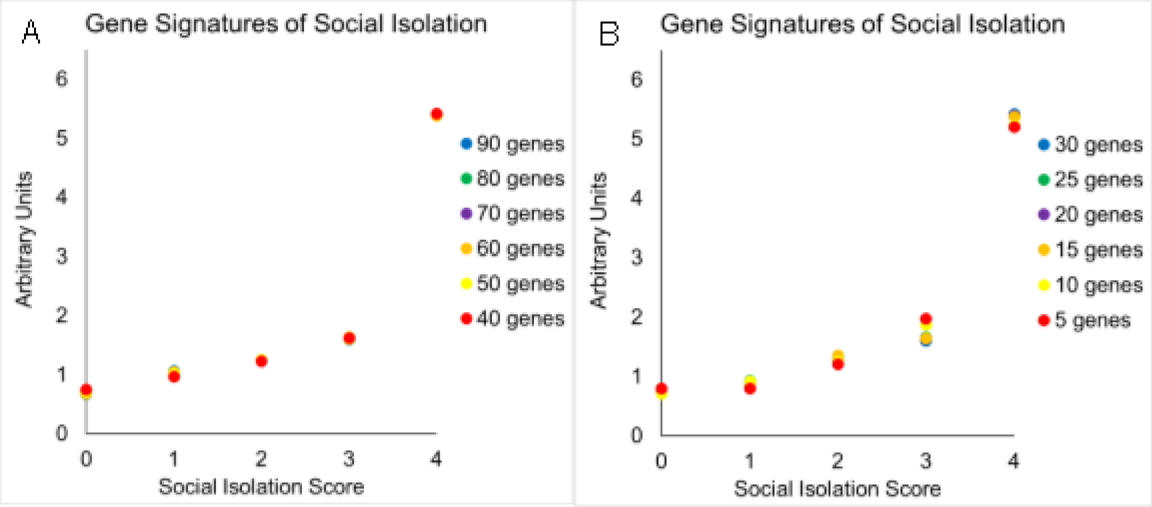

Figure 5. Gene expression signatures correlated to social isolation score according to the number of top candidate genes used in each signature. Signatures comprised of 40 or more genes are shown in panel A and fewer than 40 shown in panel B.