Genetic and Demographic Predictors of Latent Reading Ability in Two Cohorts

Hope Sparks Lancaster[1,2], Valentin Dinu[2], Jing Li[3], Jeffrey R Gruen[4,5], The GRaD Consortium

[1] Boys Town National Research Hospital, Center for Childhood Deafness Language and Learning, Omaha, NE, USA

[2] College of Health Solutions, Arizona State University, Tempe, AZ, USA

[3] School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA

[4] Department of Pediatrics, Yale University School of Medicine, New Haven, CT, USA

[5] Department of Genetics, Computational Biology and Bioinformatics, and the Investigative Medicine Program, Yale University School of Medicine, New Haven, CT, USA

RUNNING HEAD: Predictors of Reading Ability

Keywords: reading, GWAS, machine learning, ALSPAC, GRaD

Corresponding Author: Hope Sparks Lancaster

Email: hope.lancaster@boystown.org

Phone: 531-355-5086

Address:

Lied Learning and Technology Center

425 N 30th St

Office 234

Omaha, NE, USA 68131

Date submitted:

Introduction words: 1207

Discussion words: 1844

Abstract

Reading ability is a complex skill requiring multiple proficiencies (e.g., phonological awareness, decoding, and comprehension). Reading ability has genetic and environmental components that create the potential for significant gene-gene and gene-environment interactions, but the evidence for these interactions is limited. We used data from the Avon Longitudinal Study of Parents and Children and the Genes, Reading and Dyslexia Study to assess the contributions of genetic and demographic features to a continuous latent reading ability score. We then used this score as the phenotype on which to predicate genome-wide single nucleotide polymorph screening, followed by feature selection using an elastic net analysis. Results from the elastic net models showed that genetic and demographic features predicted reading ability for both cohorts. Five single nucleotide polymorphisms were associated with latent reading in the Avon Longitudinal Study of Parents and Children, as well as in the Genes, Reading and Dyslexia cohorts. For both cohorts, larger vocabularies were positively associated with latent reading ability. Genes within the neuron migration pathway were overrepresented in the Avon Longitudinal Study of Parents and Children cohort. We provide support that genes involved in early brain development have an impact on latent reading ability performance. Our findings also indicate high generalizability of genetic findings between cohorts, using our approach.

Efficient and adequate reading is the result of several neurological processes and pathways [1,2] that enable critical component skills, which include phonological awareness [3,4], decoding [5,6], fluency [7,8], and comprehension. There is substantial evidence that reading ability has a strong genetic component [9–11]. Understanding the genetics of reading ability will help to identify key developmental periods and underlying molecular processes. Although reading ability has a high heritability, ranging from 40 to 60% [12], the current knowledge about specific genes and genetic variants for reading is limited by small datasets and the traditional statistical approaches that have been used for genetic analysis. The overall goal of this study was to examine genetic associations with reading ability using an alternative statistical approach to identify genes not previously implicated by other approaches.

Reading development and performance has a complex etiology involving genetics, as well as environmental and demographic factors. Past research has established genetic contributions to reading disability (i.e., difficulties learning to read which cannot be explained by neurological or sensorial conditions, including dyslexia and other subtypes) [13–15] and performance on quantitative traits used to determine reading disability status (e.g., nonword reading [16]) or correlated with reading (e.g., multivariate rapid automatized naming/rapid alternating stimulus [17]). Recent investigations involving reading disability and related tasks have provided evidence for polygenicity, gene-gene interactions, [18] and functional biological pathways [19,20]. For example, *KIAA0319/TTRAP* and *DYX1C1* interact with *GRIN2B* in children with a reading disability when performing a short-term memory task [21] and DCDC2/KIAA0319 interact to diminish single word reading, nonword reading, spelling, phoneme deletion, and comprehension [22]. Additionally, pathway analysis suggests the effects of functional mechanisms such as neuron migration, neurite outgrowth, cortical morphogenesis, and ciliary structure and function [23] on reading. Reading development has also been linked to a number of other factors including biological sex, birth weight, gestational weeks, mother's highest education, and

language ability [24]. Mascheretti and colleagues [21] suggest that birth weight and gestational weeks are potential environmentally-related factors for dyslexia and that environmental factors may interact with each other and genetic risk. For example, they suggested that teacher quality and parental education may interact with genetic risk for dyslexia, either exacerbating or providing protection against dyslexia. Gu and colleagues [25] reported that two non-coding single nucleotide polymorphisms (SNPs; rs3779031, rs987456) within *CNTNAP2* interact with environmental factors in females but not males. In females, scheduled reading time interacted with rs987456 to reduce the risk of dyslexia. In summary, past research has revealed that the genetic contributions to reading development and performance are a complex system with multiple genes involved, as well as gene-gene and gene-environment interactions.

The most significant limitation to genetic studies of reading performance, reading disability, and dyslexia have been the small size of cohorts available for study. However, our review of published studies exploring genetic association with reading disability or reading ability identified two additional limitations. We searched the Genome Wide Association Study (GWAS) catalog for publications on reading or dyslexia and reviewed the included studies in Carrion-Castillo and colleagues [12] (see Table 1 for identified articles). Past research has (1) represented reading as either case-control (3/15 published studies, Table 1) or single task performance (6/15 published studies) and (2) relied on one-SNP-at-a-time statistical approaches (12 out of 15 published studies), examining reading disability or reading and language performance. In 7 out of 15 association studies, only one or two measures of reading or reading-related tasks were used to assign affected status as a binary variable or as a continuous variable. Additionally, 12 out of 15 studies published to date used the classical one-SNP-at-a-time to identify common variants. Neither of these conceptualizations fully captures the complexity of reading and neither accounts for possible measurement error (i.e., the difference between the “true” score and what we can observe or measure).

There are several methods to overcome these limitations. One method to account for possible measurement error is to create and use a latent reading ability score as the phenotype for genetic analysis. Three studies [26–28] have used either principal components (PCs) or a regression based composite score as their phenotype; however, these methods are not always adaptable across datasets or even iterations within a dataset. In contrast, confirmatory factor analysis can be used to create a latent reading ability score using multiple measures and is adaptable across datasets with different reading measures. For the statistical analyses, assessing genetic associations one at a time can result in data loss, especially in the small sample sizes that characterize the genetics of reading research, since multiple test correction methods must be applied. This statistical bottleneck is slowing the identification of potentially relevant genes and SNPs. It is possible to address this limitation by employing additional statistical models, such as elastic net, in conjunction with genome-wide association, to increase the number of informative SNPs. Due to these limitations, previous studies may have missed important genetic factors that contribute to or protect against reading impairments.

There is limited knowledge of how environmental and demographic factors interact with genetic factors because to date only three studies have examined gene-environment interactions in reading [25,29,30]. Due to constraints imposed by research design and statistical analysis, few studies have integrated genetic, environmental, and demographic data within the same analysis [19]. Beyond statistical constraints, another limiting factor in understanding the interactions between genes and environmental-demographic features is the demographic similarity between the most frequently used cohorts. Cohorts that capture a wider range of environmental-demographic features must be included, so that findings are generalizable beyond samples representing European descent. Increasing numbers of demographically varied cohorts are being utilized, as a result of recruitment of understudied groups within genetics [17]. Because our understanding of the genetics of reading is limited by prior statistical and cohort

constraints, we do not know how many genes are relevant for understanding the genetics of reading, which biological pathways are crucial, or how including environmental and demographic factors influence genetic associations. By combining machine learning (i.e., statistical models that learn from data) and confirmatory analytical methods, we can further our understanding of the genetics of reading ability and contribute to the refinement of the field's hypotheses. We have previously developed methods for combining statistical and machine learning approaches with biological domain knowledge to study the association between genetic and environmental factors and disease [31–33], and have applied them to study various disorders [19,34–43], including those involving reading ability [19].

In this study, we sought to address two limitations in the research exploring genetic contributions to reading ability: the focus on a reduced phenotype and over-reliance on certain statistical methods. We hypothesized that (1) the elastic net model would generate more replicated genetic markers associated with latent reading ability between datasets than traditional methods, (2) informative genetic markers would be overrepresented in certain biological processes, and (3) we would find informative positive and negative associations. Our approach was to use confirmatory factor analysis to create a latent reading ability score and combine machine learning with genome-wide association study (GWAS) approaches. We studied a robust and well-known population dataset, the Avon Longitudinal Study of Parents and Children (ALSPAC) [44], to test our initial hypotheses and then replicated our findings in an ethnically/racially diverse dataset, the Genetics of Reading and Dyslexia (GRaD) Study.

Methods

We obtained ethical approval for this study from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committee(s) (Arizona State University Institutional Review Board).

Datasets

Table 2 provides descriptive information for the ALSPAC and GRaD cohorts. Table 3 reports reading data for the ALSPAC and GRaD cohorts, including sample size for each variable, mean, and standard deviation.

ALSPAC. Our discovery cohort was the ALSPAC. The ALSPAC is a population-based birth cohort which has been extensively described in various studies [44–47]. The total sample size was 15,454 pregnancies, resulting in 15,589 fetuses, 14,901 of which were alive at 1 year of age. For this study, we used data from 8,071 participants who had participated in at least one of the ALSPAC "Focus at" sessions. These Focus sessions began at age 7 with Focus at 7 and collected behavioral and physiological data not easily assessed via questionnaire. We used data from Focus at 7, Focus at 8, and/or Focus at 9 and for whom genetic data were available. Measures included parent surveys and clinical data. The study website contains details for all the data that are available through a fully searchable data dictionary (http://www.bristol.ac.uk/alspac/researchers/our-data/). The inclusion criteria for this study were: (1) no diagnosis of autism spectrum disorder, (2) normal hearing status at Focus at 7, (3) nonverbal intelligence greater than 72 standard score on the Weschler Intelligence Scales for Children [48], (4) reading data from two Focus sessions, and (5) genetic data. Lastly, for twin-pairs one child was randomly selected for analysis to achieve data independence, which resulted in 186 children being removed from the analysis.

GRaD. Following discovery analyses in the ALSPAC, we replicated our procedures using the GRaD Study database [49]. The GRaD Study is a multi-site, case-control study of reading disability in minority youth in the USA, Canada and Puerto Rico. Phenotype information and DNA were collected for 1,435 participants. To be included in the GRaD cohort, participants had to be African American or Hispanic American between the ages of 8 and 15 years with typical or disordered reading. Exclusion criteria were age outside the target range; non-minority race/ethnic status; foster care placement; preterm birth (<36 weeks); prolonged stay in NICU;

history of diagnosed or suspected significant developmental delay, behavioral problems, serious emotional/psychiatric disturbances, chronic neurological condition, vision loss, or hearing loss; and frequent school absences. We used data from all participants in the GRaD cohort for our analyses. We had complete behavioral data for 1,409 participants to create the reading ability score and 1,341 participants who passed genomic quality control for final data analyses.

Measures

ALSPAC. We used behavioral and demographic measures, including reading, language, and nonverbal intelligence measures, collected between the ages of 7 and 9. Reading skill was measured during Focus at 7 and Focus at 9 using a combination of word reading, spelling, and connected text tasks. At Focus at 7 years, children completed the single word reading subtest on the Wechsler Objective Reading Dimensions [50] and an experimenter derived spelling task [51]. Nonword repetition was assessed at Focus at 8 [52]. At Focus at 9, children completed single word reading, nonword reading [53], and spelling tasks similar to the ones presented during Focus at 7 years, but with new words/items. Additionally, at age 9, children completed the Neale Analysis of Reading Ability (NARA; Neale, 1997), which provided scores for reading rate, accuracy, and reading comprehension. The Wechsler Intelligence Scale for Children (WISC; Wechsler et al. 1992) (Focus at 8) yielded an estimate of nonverbal intelligence. Receptive language was assessed using the Wechsler Objective Language Dimensions Language Comprehension subtest (WOLD; Rust, 1996). Vocabulary was measured using the Wechsler Intelligence Scales for Children vocabulary subtest (WISC; Wechsler, Golombok, & Rust, 1992).

We selected biological sex, birth weight, maternal education, child ethnicity, bilingual language status, hearing function, and Attention Deficit Hyperactivity Disorder (ADHD) status, as our demographic measures. Biological sex and birth weight in grams were reported at birth. Maternal education was obtained at 32-weeks' gestation and measures the highest degree the mother had obtained by that point: Vocation, certificate of secondary education, O-levels, A-

levels, or College degree. Child's ethnicity was reported by mothers at 32 weeks' gestation. Bilingual language status was obtained via parent report at Focus at 8. Hearing functioning was measured via bone conduction at Focus at 7. ADHD) status was determined at age 7 using parent and teacher questionnaires.

GRaD. We used reading, language, and demographic data from the GRaD. All data was collected during one or two sessions. Reading was measured using multiple indicators, including the Test of Word Reading Efficiency (TOWRE; Torgesen et al. 2012), Woodcock-Johnson Tests of Achievement (3rd Edition; WJ-III; Woodcock et al. 2001), Clinical Test of Phonological Processing (CTOPP; Wagner et al. 2013), and Standardized Reading Inventory (SRI; Newcomer 1986). The TOWRE provided two timed measures of reading fluency – Sight Word Efficiency and Phonemic Decoding Efficiency. The TOWRE measures a child's ability to quickly and accurately read real words and pseudowords. The WJ-III provides three measures of reading – Letter-Word Identification, Word Attack, and Spelling. Letter-Word Identification and Word Attack are untimed analogous tasks to those of the TOWRE test, in which children read lists of known and pseudowords. Similar to the ALSPAC spelling task, the WJ-III Spelling task requires children to correctly spell an orally presented word. The CTOPP-Blending Words requires children to combine phonemes into a word. This task taps into phonological awareness. The Standardized Reading Inventory provides a comprehension and word recognition score. Receptive vocabulary was measured using the Peabody Picture Vocabulary Test [60].

We selected parent reported ADHD status, Hispanic/Latino status, child's race, and maternal education (either birth or, if relevant, adoptive) as demographic measures. ADHD status was defined as having received a psychiatric diagnosis of ADHD. Maternal education was obtained via parent questionnaire and was defined as (1) less than 7 years of school, (2) 7 to 9 years of school, (3) 10 to 11 years of school, (4) high school diploma or GED, (5) associate

degree, trade, or business school, (6) bachelor's degree, and (7) professional or advanced degree.

Genotyping

ALSPAC. ALSPAC samples were genotyped using the Illumina HumanHap550 quad chip genotyping platforms by 23andme, subcontracting the Wellcome Trust Sanger Institute, (Cambridge, UK) and the Laboratory Corporation of America (Burlington, NC, USA). The resulting raw genome-wide data were subjected to standard quality control methods. Individuals were excluded on the basis of gender mismatches; minimal or excessive heterozygosity; disproportionate levels of individual missingness (>3%) and insufficient sample replication (identity by descent (IBD) < 0.8). Population stratification was assessed by multidimensional scaling analysis and compared with Hapmap II (release 22) European descent (CEU), Han Chinese, Japanese and Yoruba reference populations; all individuals with non-European ancestry were removed. SNPs with a minor allele frequency of < 1%, a call rate of < 95%, or evidence for violations of Hardy-Weinberg equilibrium (P < 5E-7) were removed. Cryptic relatedness was measured as a proportion of IBD > 0.1. Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation. 9,115 participants and 500,527 SNPs passed these quality control filters.

GRaD. Saliva was collected using Oragene-DNA kits and DNA was then extracted with prepIT-L2P (OG-500; DNA Genotek Inc, Ottawa, Ontario, Canada). Participants were genotyped for >2 million SNPs using the Illumina Infinium Omni2.5-8 BeadChip at the Yale Center for Genome Analysis (Orange, CT, USA). Initial genotyping quality control and SNP genotyping calls were conducted using GenomeStudio (Illumina, San Diego, CA, USA) and standard Infinium genotyping data analysis parameters were used to optimize genotyping accuracy. Individuals were removed if they were missing more than 3% of their genotypes (n = 39), if there were discrepancies between reported and inferred sex based on X chromosome

heterozygosity (n = 52), and if IBD > 0.125 using REAP (n = 10)[61]. SNPs were removed from downstream analyses if they had minor allele frequency of less than 5% (n = 926,457), missingness greater than 5% (n = 22,849), Hardy-Weinberg equilibrium p<0.0001 (n = 116,259) or were not autosomal (n = 60,551). 1,331 participants and 1,265,623 SNPs passed these quality control filters.

Statistical Analysis

Creating reading ability score. To create the latent reading ability score, we used confirmatory factor analysis (CFA) to load all of the reading variables onto a single factor. We used lavaan [62] to fit and assess the reading ability model. Current practice is to use several model fit criteria instead of relying on a single measure. We assessed model fit using a combination of absolute, parsimonious, and comparative indices of model fit [63]. To determine goodness of fit, we evaluated (a) Tucker-Lewis Index (TLI), (b) root mean square error approximation (RMSEA), and (c) standardized root mean square residual (SRMR). We used the TLI to assess comparative or incremental fit. The TLI is a non-normed fit index that is analogous to the r-squared coefficient, with penalties for added parameters. Like the r-squared coefficient, higher values indicate better fit, with the traditional cutoff value for good fit at 0.90. Our index of parsimony was RMSEA. RMSEA ranges from 0 to 1 with values of < 0.08 representing acceptable fit and values < 0.05 representing close fit [64]. We report the 90% confidence interval for the RMSEA and the p-value for the closeness of fit test, which tests the null hypothesis that RMSEA is <= 0.05; this test should result in a nonsignificant p-value. Our index of absolute fit was SRMR, which represents the squared difference between observed and predicted correlations and for which values < 0.08 are considered acceptable.

After assessing the fit of the model, we extracted the lambda values associated with manifest paths that exceeded 0.20. The following equation was used to approximate an individual's reading ability score:

$$Reading\ Ability = \lambda_x z_x + \cdots + \lambda_i z_i$$

wherein each lambda was multiplied by the corresponding z-score for a measure and the products were summed together. By using this method, we can approximate the latent construct of reading ability instead of relying on a single reading measure. After approximating the reading ability, we assessed the distribution of scores and normalized if necessary. We used the reading ability score as the phenotype for SNP screening and as the outcome variable for the elastic net model.

Imputation for missing data. We used multiple imputation methods to deal with missing data. For the multiple imputation, we used mice [65]. For each imputation, we estimated scores for missing reading measures, fit the reading ability model to the imputed dataset, and approximated reading ability scores. We averaged reading ability scores for the five imputations and determined the variability. We used the average reading ability score as the phenotype for SNP screening and as the outcome variable for the elastic net model.

SNP screening. To constrain the high-dimensionality of the dataset, we used genome-wide association (GWA) to screen SNPs prior to multivariate modeling. GWA was completed in PLINK [66] and performed chromosome-by-chromosome based on criteria used in prior studies [67]. We selected up to 100 SNPs based on uncorrected *p*-values and SNPs with an FDR-BH of 0.1 or less to be included in the subsequent multivariate modeling. Because the GRaD dataset contains more than one ethnic group, we included 10 ancestry principal components. We used the standard settings in PLINK.

Machine learning. There are many options for machine learning. We selected a procedure that would allow for correlated features, more features than subjects, multiple data types, and multiple feature selection. We performed multivariate modeling using an elastic net model to link reading ability with the SNPs that survived the screening step, as well as demographic, environmental, and behavioral covariates. An elastic net is a regularized

regression model that enables simultaneous feature selection (in our case, variables are the SNPs) in a high-dimensional setting [68,69]. It adds two regularization terms to the loss function of an ordinary regression model: one L1-norm regularization whose effect is to force the regression coefficients of small effects to be exactly zero, thus enabling feature (e.g., SNP) selection; a second L2-norm regularization term insures highly correlated SNPs are selected. There are two tuning parameters corresponding to the two regularization terms to balance with the loss function. Tuning parameter selection is typically done using cross validation (see below).

We used 5-fold cross-validation to determine the best tuning parameters. In this process, the sample is split into five random groups, four of which are used to train the model and one for testing. This splitting repeats until every “fold” has served as the test set. Cross-validation was performed 10 times to select tuning parameters. After the best tuning parameters were identified, the model was refit using all the data to generate coefficients.

In addition to SNP only modes, we ran SNP plus demographic feature models. For ALSPAC, we included nonverbal IQ, vocabulary, receptive language score, ADHD status, birth weight, bilingual language status, and mother’s highest education in the multivariate model. For GRaD, we selected ADHD status, Hispanic/Latino status, child’s race, mother’s highest education (either birth or, if relevant, adoptive), and vocabulary. However, for GRaD, we were only able to use child’s race and vocabulary score (measured by Peabody Picture Vocabulary), due to large missingness for the other variables (ADHD missing = 266, Hispanic/Latino missing = 138, mother’s education missing = 215). We conducted this analysis to determine the impact of SNPs on reading when controlling for demographic, environmental, and behavioral contributions.

Pathway enrichment and network analysis. We mapped informative SNPs from the elastic net to genes using g:SNPense on g:Profiler [70]. After mapping SNPs to genes, we

performed enrichment analysis using g:GOSt on g:Profiler. g:Profiler was selected over similar tools because recent comparisons of the available tools showed that g:Profiler has the most up-to-date repository of pathways and draws from multiple curated sources (e.g., KEGG, Reactome). However, enrichment analysis alone only identifies those pathways that are overrepresented in a gene list but it cannot delineate how these pathways interact. Therefore, we used Cytoscape to explore how the pathways were connected [71]. Cytoscape performs network analysis on biological pathways and produces visualizations and network statistics.

Results

Latent Reading Ability Score Creation

ALSPAC. We fitted a single factor model to the reading variables after z-score transformation. Two measures of goodness of fit indicated highly acceptable fit (TLI = 0.925, SRMR = 0.033), whereas chi-square (χ^2(df = 27) = 1307.97, p < .001) and RMSEA (RMSEA = 0.124, 95% confidence interval = 0.118, 0.131, *p* = < .001) did not meet the guidelines for fit. All manifest paths were significant and had standardized path values over 0.70 (range = 0.745, 0.929). There were no negative variances, and the model explained a large amount of variance in reading performance (average variance = 0.667). These results indicate that the single factor model fit the data and could be used to create a reading ability score.

After computing the latent reading ability score, we investigated the distribution. Because of the nature of the score, the mean was zero and the standard deviation was one. Reading ability scores ranged from -3.46 to 2.07 and inspection of the plot indicated slight right skew. We investigated the reading ability scores for children with a reading disorder compared to children with typical reading ability. A reading disorder was defined as failing three of the reading variables. Children with a reading disorder had lower scores compared to children with typical reading (Supplemental Figure 1). These results indicated that the latent reading ability score functioned as desired, by capturing the full range of reading ability. We proceeded with the

genetic analyses using this latent reading ability score. Additionally, we imputed the latent reading ability score for children for whom reading data were missing using multiple imputation.

GRaD. We fitted a single factor model using TOWRE, CTOPP-Blending Words, WJ-III Word Reading, WJ-III Word Attack, WJ-III Spelling, and SRI – Passage Comprehension, and SRI WR measures. Twenty-six children had incomplete data. We used robust modeling to account for missing data. Two goodness-of-fit metrics were highly acceptable (TLI = 0.909, SRMS = 0.031), whereas chi-square (χ^2(df = 20) = 730.21, p < .001) and RMSEA (RMSEA = 0.159, 95% confidence interval = 0.149, 0.168, p < .001) did not indicate good fit. The standardized path values ranged from 0.528 (CTOPP – Blending Words) to 0.913 (SRI WR). The model accounted for a large amount of variance in reading ability (average variance = 0.754). There were no negative variances. These results indicate that the single factor model fit the data and could be used to create a latent reading ability score.

We inspected the distribution of the latent reading ability score, which had a mean of 0 and standard deviation of 1. There was no evidence of non-normality in the histogram (Supplemental Figure 1).

ALSPAC

Genome-wide association and multivariate modeling. We performed a GWAS predicated on latent reading ability and imputed reading ability scores in the ALSPAC cohort. The latent reading ability analysis will be called “Analysis_1” and the imputed reading ability analysis will be called “Analysis_2” henceforth. After multiple test correction, Analysis_1 had four significant SNPs and Analysis_2 had 24 significant SNPs when using an FDR-BH of 0.05. Analysis_2 had an additional 67 SNPs with an FDR-BH of less than 0.1, all on chromosome 17 (Supplemental Table 1). Table 4 contains the list of SNPs that were significant after multiple test correction and genomic information for Analysis_1 and Analysis_2.

For the elastic net model, we used 149 SNPs for Analysis_1 and 250 SNPs for Analysis_2. Seventy-one and 67 SNPs were identified as informative (i.e., had a beta value greater than the absolute value of 0.01) for Analysis_1 and Analysis_2, respectively. Table 5 reports genomic information and beta weights from the elastic net model for informative SNPs. Supplemental Table 1 provides the lists of SNPs used in both Analysis_1 and Analysis_2. Across the two analyses, there were seven SNPs commonly selected. For Analysis_1, 24 SNPs were positively associated with latent reading ability (i.e., predicted better reading) and 47 SNPs were negatively associated with latent reading ability (i.e., predicted worse reading). For Analysis_2, 17 SNPs were positively associated with imputed reading ability, 50 were negatively associated. SNPs were selected from across the genome with the majority on chromosome 6 for Analysis_1 and on chromosome 15 for Analysis_2.

We used a linear regression model to assess the fit of Analysis_1 and Analysis_2. Analysis_1 fit the data significantly better than a null model ($F(1, 83) = 8.22$, $p < .0001$) and explained roughly 14 percent of the variance in reading ability (Adjusted $R^2 = 0.141$). The positively associated SNPs were located within 16 genes with the majority representing intron variants. The negatively associated SNPs were located within 21 genes with most variant effects being intronic. SNPs mapped to *DNAAF4* had positive and negative associations with reading ability. *RAPGEF2* and *GRIN2B* were negatively associated with reading ability.

Analysis_2 fit the data significantly better ($F(1, 75) = 10.07$, $p < .0001$) than a null model and explained roughly 14 percent of the variance in the reading score (Adjusted $R^2 = 0.138$). The positively associated SNPs were located within 11 genes while the negatively associated SNPs were located within 24 genes. Most of these variants were intron or non-coding variants for both positive and negative associations.

We compared the results from both analyses to identify which SNPs replicated internally and mapped the replicated SNPs to genes. The seven SNPs common to both Analysis_1 and

Analysis_2 mapped to six genes, which were *KIAA0319* (chr 6), *FOXP2* (chr 7), *DRD2* (chr 11), *CYP19A1* (chr 15), *DNAAF4* (chr 15), and *ATP2C2* (chr 16). These SNPs were the SNPs included from previous genome-wide studies on dyslexia and reading traits, thus replicating previous research.

Pathway and network analysis. For Analysis_1, there were two significantly overrepresented biological pathways (GO:0010996, response to auditory stimulus, p = .0315; KEGG:04015, Rap1 signaling pathway, p = .02395). For Analysis_2, we replicated the significance of response to auditory stimulus (GO:0010996, p = 0.0186), but not the Rap1 signaling pathway. We were unable to perform the network analysis due to the limited number of pathways identified.

Addition of demographic features. We investigated the impact of nonverbal IQ, birthweight, mother's highest education, vocabulary at age 8, receptive language, and bilingual language status in the presence of genetic features. Our elastic net model selected all the demographic features as informative in the presence of genetic features. It yielded positive beta weights for nonverbal IQ, vocabulary, mother's highest education, and bilingual language status, indicating that higher values on these features were associated with higher latent reading ability scores. Birthweight had a negative association with latent reading ability; however, further examination of this relationship did not indicate a negative trend but rather a near zero correlation (Pearson's r = 0.017). Examining the beta weights indicate that birthweight and bilingual language status had the lowest associations with latent reading ability, which may indicate that these factors were selected due to their relationship with other factors in the model and not directly with latent reading ability. Relationships between demographic features and latent reading ability are presented in Supplemental Figure 2.

The addition of demographic features also increased the number of SNPs selected as informative, with the model selecting 35 positively associated SNPs and 52 negatively

associated SNPs. Twenty-seven of the positively associated SNPs mapped to 19 genes across the genome, including *DCDC2*, *DNAAF4*, *GRIN2B,* and *KIAA0319*. Thirty-four of the negatively associated SNPs mapped to 23 genes across the genome, including *DCDC2, FOXP2, RAPGEF2, SNTG1,* and *DNAAF4*.

We fit a standard linear regression model using the selected features from the elastic net. A model with demographic features fit the data significantly better ($F(1, 106) = 16.28$, $p < .0001$) than a null model and explained roughly 33 percent of the variance in the reading score (Adjusted $R^2 = 0.329$). Additionally, comparing the model with demographic features to Analysis_1 without these features, showed that the model with demographic features fit the data significantly better ($F(1, 23) = 38.68$, $p < .001$). There were no "significantly" overrepresented biological pathways, although neuron migration had the lowest adjusted p-value ($p = 0.269$).

Replication in GRaD

We performed a GWAS using the reading score generated by confirmatory factor analysis within the GRaD dataset. No SNPs remained significant after multiple test correction.

For the elastic net model, we used the top 100 SNPs before multiple test correction and 12 SNPs from the ALSPAC Analysis_1 list for which we were able to find matches in the GRaD dataset. Forty-eight SNPs were positively associated with latent reading ability and 36 SNPs were negatively associated with latent reading ability. After removing markers that did not begin with "rs", there were 40 SNPs positively and 34 SNPs negatively associated with latent reading ability. The 40 positive SNPs mapped to 18 genes, while the 34 negative SNPs mapped to 15 genes. Overall, these SNPs mapped to 33 unique genes.

We also included biological sex, child's ethnicity/race, and vocabulary score within the model. Including the demographic features increased the number of SNPs selected as informative to 91, with 52 positively associated and 41 negatively associated. After removing SNPs that did not begin with "rs", the positively associated SNPs mapped to 18 genes, while the

negatively associated SNPs mapped to 16 genes. Table 6 presents the 55 informative SNPs that mapped to known genes for the analysis that included demographic features. Biological sex, child's race, and vocabulary scores had small associations with latent reading ability. Girls had higher latent reading ability scores than boys (Cohen's *d* = 0.097). Children who were African American had a small difference on their latent reading ability scores compared to children who were not African American (Cohen's *d* = 0.14) Additionally, higher vocabulary scores were related to higher latent reading ability scores (Pearson's r = 0.709). See Supplemental Figure 2 for a visual depiction of latent reading ability score and demographic feature relationships.

We replicated five SNPs between GRaD and ALSPAC analyses. These SNPs were rs79439102 (*LSAMP*), rs7681750 (*RAPGEF2*), rs10046 (*CYP19A1*), rs77641439 (*DNAAF4*), and rs12606138 (*NEDD4L*). Of the replicated SNPs, three had the same direction in the GRaD and ALSPAC cohorts (rs77641439, rs79439102, and rs10046), whereas the other two (rs12606138 and rs7681750) had positive associations in the GRaD cohort but negative associations in the ALSPAC cohort. Additionally, *RAPGEF2:*rs7681750 was not considered informative after adding demographic features for the ALSPAC cohort, although a different SNP on *RAPGEF2* was selected in both ALSPAC analyses (*RAPGEF2*:rs55703414). There were no significantly overrepresented pathways using the genes from the GRaD results.

Discussion

We investigated the genetic contributions to reading ability using a combination of confirmatory factor analysis, data imputation, GWAS, multivariate elastic net models, and pathway analysis. We were able to overcome a common limitation of genetic studies of reading ability, namely lack of significant findings after multiple testing correction, and identified several genes that are informative for reading ability. We replicated multiple genetic associations, from previous studies, across our complete data and imputed analyses, and between the ASLPAC

and GRaD cohorts. We identified novel SNPs within known loci and novel loci. Our results support the polygenic understanding of reading [14] and suggest that several domain general genes are involved in reading development. We also demonstrated that certain demographic features are associated with reading ability alongside genetic features.

Genetic Associations

Our ALSPAC results consistently selected SNPs from previous literature from *DNAAF4, RAPGEF2, DCDC2, KIAA0319, FOXP2, GRIN2B*, *SNTG1*, *SUCLA2*, and others as informative to understanding reading (dis)ability. We replicated SNPs between the ALSPAC and GRaD cohorts from *LSAMP* (limbic system associated membrane protein; chr3q13.31), *RAPGEF2* (Rap guanine nucleotide exchange factor 2; chr 4q32.1), *CYP19A1* (cytochrome P450 subfamily A member 1; chr15q21.2), *DNAAF4* (Dynein axonemal assembly factor 4; chr15q21.3), and *NEDD4L* (Ubiquitin protein ligase NEDD4-like; chr18q21.31). These genes all have some role in brain and neuron development and have been implicated in related cognitive processes. *CYP19A1*:rs10046 and *DNAAF4*:rs77641139 are located within DYX1C1 (15q15.1 to 15q21.3), the first susceptibility locus for dyslexia, and both genes have been linked to reading and language disorders [72]. *CYP19A1* and *DNAAF4* help to regulate estrogen, with *CYP19A1* regulating estrogen signaling, and *DNAAF4* influencing neuronal differentiation, survival, and plasticity by regulating estrogen receptors. Disruptions in *CYP19A1* in animal models have resulted in cortical disorganization [73]. *DNAAF4:*rs77641439 is within 11bp of the previously identified *DNAAF4*:rs57809907, suggesting that it was not selected due to a false positive but rather because this region influences reading ability. Variants for *CYP19A1*:rs10046 and *DNAAF4*:rs77641139 can affect regulatory elements of these genes as they can alter the mRNA either in the 3' UTR or through nonsense mediated decay, providing a possible method by which they influence our reading ability phenotype. SNPs from *RAPGEF2* and *NEDD4L* have also been previously associated with reading disability [19,74] and related cognitive skills (e.g.,

working memory) [75]. Like *CYP19A1* and *DNAAF4*, *RAPGEF2* and *NEDD4L* are expressed in the brain and are involved in functional biological pathways that help with brain development. *RAPGEF2* helps with the formation of connections for the corpus callosum, anterior commissure, and the hippocampal commissure during brain development. *NEDD4L* is a ubiquitin protein ligase which binds and regulates membrane proteins to aid in internalization and turnover. However, the SNPs replicated are both within the introns for these genes, so how exactly they influence reading ability is unclear. Lastly, we identified a novel SNP in *LSAMP*. *LSAMP* mediates selective neuron growth and axon targeting, which contributes to the guidance of axons. SNPs from *LSAMP* have been associated with self-reported educational attainment and mathematical ability[74], two skills related to reading ability. *LSAMP*:rs79439102 is an intron variant, so we are uncertain how it might influence reading ability; although, there are a growing number of intron variants associated with behavioral phenotypes (for example see [76]) or it may have been selected because it is correlated with another SNP.

All of these genes are implicated in brain development, specifically in neural organization and neuron connections and many have a more general role in development. There is growing awareness that domain general genes have an important role to play in the genetics of reading [77,78]. Additionally, four out of five of these genes play a role in the neuron migration pathway, which is often hypothesized to be causal to reading disabilities [79,80]. These findings build on prior work which showed that neuron migration was overrepresented in genetic findings for typical and atypical readers [19,81]. There is still much to understand about how the selected SNPs influence reading ability through gene expression and function over time, during brain development.

We replicated two novel SNPs from a known loci and three novel loci. Our analysis was able to identify these novel SNPs and loci because of the elastic net model and latent reading ability score. One potential concern is that these novel SNPs/loci may have been selected

because of less stringent criteria for p-values than in previous studies. Machine learning methods, such as elastic net, bypass significance testing thus providing a way to overcome false positives. In other studies, using similar analysis methods, researchers input several thousand SNPs. Here, we utilized less than 200 per analysis. Additionally, not all SNPs with the same or similar p-value were selected as informative in our analysis. For example, only two SNPs from *AC104041.1* were selected as informative, although six SNPs from this gene were inputted into the model with an FDR p-value of .035. This demonstrates that the elastic net model is informed more by how the inputted features work together than by significance testing. Our latent reading ability score is what enabled us to replicate results across datasets. This score was better able to represent the same construct (i.e., reading) than a single task across participants, and reduced measurement error, which could have influenced SNP selection.

Demographic Features

We were able to replicate the importance of several demographic features in predicting reading ability. Our results demonstrated that higher vocabulary, better receptive language, higher mother's education, and higher nonverbal IQ scores were associated with better reading ability. These associations have been identified by several research studies [19,24,82]. Our previous study using a case-control design suggested that these features may function as protective factors against developing a reading disorder [19]; however, more research is needed to examine the interaction between genetic and demographic features. The inclusion of demographic features more than doubled the amount of variance explained in reading ability, from 14% to 33%, although there was still a significant amount of unexplained variance. This finding suggests that there are other important features associated with reading ability that we were unable to include in our model. Additional features could include nonlinguistic measures, such as finger tapping, rhythm judgement, and visual attention, where previous studies suggest a link

between motor abilities [83], auditory perception [84,85], and visual abilities [86,87] with reading and dyslexia.

Limitations and Future Directions

The strengths of this study are the statistical procedures to overcome multiple test correction loss in small sample sizes, ability to compensate for missing data, and external replication in an ethnically/racially diverse sample. Despite these strengths, there are a few notable limitations, including the relatively small sample sizes of the cohorts for genetic analyses, as well as key statistical limitations. The ALSPAC sample is one of the largest samples for investigating behavioral genetics in children, but it is still considered small by genetics standards. This is especially true, as evidenced by the work of large consortiums in recent years. The sample size limitation is unavoidable, and a common limitation for genetic studies examining reading ability. Partly, this is due to the cost involved in resources and infrastructure required to collect in-person behavioral data. Despite this limitation, the ALSPAC cohort is a rich database with a large enough sample size for developing hypotheses. We were able to diminish this limitation by using imputation to increase the amount of useable data and by employing a replication dataset. Future research studies can use methods such as online data collection, meta-GWA across samples, and data pooling to address the sample size issue.

Our statistical limitations include inconsistent goodness of fit indices for our latent reading ability, inability to determine the influence of all selected SNPs or investigate interactions, and interpretation of directionality. Our chi-square and RMSEA indices were not within the “acceptable” range. This finding was not unexpected because these estimates are often misestimated when a sample size is greater than 100 (chi-square) or there are less than ten observed variables (i.e., RMSEA). Because our sample sizes exceed 1000 and we used eight variables per cohort, we must rely on TLI and SRMR, both of which were within acceptable limits. Therefore, we can be reasonably confident that our model fit the data adequately. We

cannot explain the role of every selected SNP, because many SNPs did not map to genes. For example, rs2957954 had a high beta weight in Analysis_1 but is an intergenic marker. Therefore, we cannot adequately explain the role this SNP might have on reading ability. There is some evidence that such intergenic markers might have a role in the evolution and development of communication in humans [88], so this is an avenue for future research. Although all of the informative SNPs were non-coding or non-functional in our queried databases, it is, nevertheless, possible that some could cause alternative splicing or regulate gene expression of other genes. We could not investigate interactions between SNPs or between SNPs and demographic features, because the elastic net model we used only considers cumulative effects. Future research studies should investigate gene-gene and gene-environment interactions to better understand how these sets of features work together for reading development. A final limitation to the elastic net model is that the directionality of the selected features may not map exactly onto the relationship in the data. Directionality in the elastic net is determined by the algorithm for best fit to the data, but deeper exploration outside of the model is needed to understand the true relationship. In general, the directionality in the model mirrors actual directionality (e.g., positive beta weight for vocabulary from the model and true positive association); however, there are a few instances where the direction the elastic model selected was not reflected within the data (e.g., birth weight and reading; child's race in GRaD and reading). Therefore, our results provide guidance for what factors are important to consider, but all relationships should be explored in greater depth.

Conclusions

Our findings reinforce the understanding that reading ability is a complex disorder with genetic and demographic associations. For both cohorts, we found that SNPs from more than 20 genes were selected as informative. This suggests that reading ability is influenced by several genetic factors, each contributing a small effect. This polygenic hypothesis is becoming

a prominent hypothesis for understanding reading ability. Although the genetic associations are each individually small, together they influence the development of brain structures and connections between neurons, resulting in good or poor reading. We identified three novel loci using our methods and these loci should be further explored in other datasets. In addition to these genetic markers, child characteristics such as vocabulary, nonverbal IQ, and language, and external factors, such as maternal educational attainment, also predict reading ability. Therefore, reading ability is not only the product of genetic markers, but also additional skills and factors, all of which could potentially serve as targets for early interventions to improve reading. Our findings provide evidence for genetic markers that replicate in ethnically/racially diverse samples, expanding our understanding of the genetics of reading beyond English speaking, European-Caucasian children. Therefore, our findings suggest that there is large generalizability of genetic factors for reading across research cohorts.

Declarations

Not Applicable.

Funding

The UK Medical Research Council Wellcome Trust Grant (ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and they will serve as guarantors for the contents of this paper. A comprehensive list of grants funding (PDF, 459KB, http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf) is available on the ALSPAC website. GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. We acknowledge funding from the National Institutes of Health (Grant ref: R01 NS043530 awarded to JRG; P50 HD027802 awarded to JRG), and the Manton Foundation for support of the GRaD Study. The

first author was supported by a National Institutes of Health F32 postdoctoral training grant (1F32HD089674-01A1; PI: HSL).

Competing Interests

The authors have no competing interests to declare related to this paper.

Ethics Approval

We obtained ethical approval for this study from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committee(s) (Arizona State University Institutional Review Board). Informed consent and assent were obtained by research staff for each dataset. The GRaD dataset also had ethical approval from all recruitment sites (University of Colorado-Boulder, University of Denver, Tufts University, University of New Mexico, Kennedy Krieger Institute/John Hopkins University, Hospital for Sick Children-Toronto, and Yale University).

Consent to Participate

Not Applicable.

Consent to Publish

Not Applicable

Availability of Data and Material

Please note that the study website contains details of all the data that are available through a fully searchable data dictionary (http://www.bristol.ac.uk/alspac/researchers/our-data/). Data from this study are available through ALSPAC upon approval by the executive board. Summary level data for the GRaD Study will be made available upon request.

Code Availability

Code is available in Supplemental A.

Authors Contributions

HSL conceived and designed the study under the mentorship of JL and VD, analyzed and interpreted the data, drafted the manuscript, and revised it based on feedback from JL, VD, JG,

and GRaD Consortium. HSL approved the final version of the manuscript on behalf of all the authors.

Acknowledgements

The authors are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the entire ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. We are indebted to the members of the GRaD Study: the participants and their families, the teams who performed recruiting and testing, as well as the project managers at each site. We would also like to thank the Yale W.M. Keck Biotechnology Resource Laboratory's DNA Sequencing Resource for Sanger sequencing services. Thank you to Xiaonan Liu for his help with data analyses on the F32 project.

References

1. Martin A, Schurz M, Kronbichler M, Richlan F. Reading in the brain of children and adults: A meta-analysis of 40 functional magnetic resonance imaging studies. *Human Brain Mapping*. 2015;36(5):1963-1981. doi:10.1002/hbm.22749
2. Partanen M, Kim DHC, Rauscher A, Siegel LS, Giaschi DE. White matter but not grey matter predicts change in reading skills after intervention. *Dyslexia*. Published online 2020:1-21. doi:10.1002/dys.1668
3. Knoop-van Campen CAN, Segers E, Verhoeven L. How phonological awareness mediates the relation between working memory and word reading efficiency in children with dyslexia. *Dyslexia*. 2018;(February):1-14. doi:10.1002/dys.1583
4. Kirby JR, Parrila RK, Pfeiffer SL. Naming speed and phonological awareness as predictors of reading development. *Journal of Educational Psychology*. 2003;95(3):453-464. http://search.proquest.com.proxy.library.vanderbilt.edu/docview/614384952?accountid=14816 LA - English
5. Nunes T, Bryant P, Barros R. The development of word recognition and its significance for comprehension and fluency. *Journal of Educational Psychology*. 2012;104(4):959-973. doi:10.1037/a0027412
6. Gough PB, Tunmer WE. Decoding, reading, and reading disability. *Remedial and Special Education*. 1986;2:6-10.
7. van Setten ERHH, Hakvoort BE, van der Leij A, Maurits NM, Maassen BAMM. Predictors for grade 6 reading in children at familial risk of dyslexia. *Annals of Dyslexia*. 2018;68(3):181-202. doi:10.1007/s11881-018-0162-1
8. Adlof SM, Catts HW, Little TD. Should the simple view of reading include a fluency component? *Reading and Writing*. 2006;19(9):933-958. doi:10.1007/s11145-006-9024-z

9. The Brainstorm Consortium. Analysis of shared heritability in common disorders of the brain. *Science*. 2018;360(6395). doi:10.1126/science.aap8757

10. Harlaar N, Spinath FM, Dale PS, Plomin R. Genetic influences on early word recognition abilities and disabilities: A study of 7-year-old twins. *Journal of Child Psychology and Psychiatry and Allied Disciplines*. 2005;46(4):373-384. doi:10.1111/j.1469-7610.2004.00358.x

11. Fisher SE, Marlow AJ, Lamb J, et al. A quantitative-trait locus on chromosome 6p influences different aspects of developmental dyslexia. *American Journal of Human Genetics*. 1999;64(1):146-156. doi:10.1086/302190

12. Carrion-Castillo A, Franke B, Fisher SE. Molecular genetics of dyslexia: An overview. *Dyslexia*. 2013;19(4):214-240. doi:10.1002/dys.1464

13. Facoetti A, Gori S, Vicari S, Menghini D. Introduction to the special issue: Developmental dyslexia: From genes to remediation. *Neuropsychologia*. 2019;130(June):1-2. doi:10.1016/j.neuropsychologia.2019.06.003

14. Gialluisi A, Andlauer TFM, Mirza-Schreiber N, et al. Genome-wide association study reveals new insights into the heritability and genetic correlates of developmental dyslexia. *Molecular Psychiatry*. 2020;(Dd). doi:10.1038/s41380-020-00898-x

15. Landi N, Perdue M V. Neuroimaging genetics studies of specific reading disability and developmental language disorder: A review. *Language and Linguistics Compass*. 2019;13(9). doi:10.1111/lnc3.12349

16. Doust C, Gordon SD, Garden N, et al. The Association of Dyslexia and Developmental Speech and Language Disorder Candidate Genes with Reading and Language Abilities in Adults. *Twin Research and Human Genetics*. 2020;23(April):23-32. doi:10.1017/thg.2020.7

17. Truong DT, Adams AK, Paniagua S, et al. Multivariate genome-wide association study of rapid automatised naming and rapid alternating stimulus in Hispanic American and African-American youth. *Journal of Medical Genetics*. 2019;56(8):557-566. doi:10.1136/jmedgenet-2018-105874

18. Price KM, Wigg KG, Feng Y, et al. Genome-Wide Association Study of Word Reading: Overlap with Risk Genes for Neurodevelopmental Disorders. *Genes, brain, and behavior*. Published online 2020:e12648. doi:10.1111/gbb.12648

19. Lancaster HS, Liu X, Dinu V, Li J. Identifying interactive biological pathways associated with reading disability. *Brain and Behavior*. Published online 2020.

20. Mascheretti S, Bureau A, Trezzi V, Giorda R, Marino C. An assessment of gene-by-gene interactions as a tool to unfold missing heritability in dyslexia. *Human genetics*. 2015;134(7):749-760. doi:10.1007/s00439-015-1555-4

21. Mascheretti S, Bureau A, Battaglia M, et al. An assessment of gene-by-environment interactions in developmental dyslexia-related phenotypes. *Genes, Brain and Behavior*. 2013;12(1):47-55. doi:10.1111/gbb.12000

22. Powers NR, Eicher JD, Miller LL, et al. The regulatory element READ1 epistatically influences reading and language, with both deleterious and protective alleles. *Journal of Medical Genetics*. 2016;53(3):163-171. doi:10.1136/jmedgenet-2015-103418

23. Newbury DF, Monaco AP, Paracchini S. Reading and language disorders: The importance of both quantity and quality. *Genes*. 2014;5(2):285-309. doi:10.3390/genes5020285

24. Mascheretti S, Andreola C, Scaini S, Sulpizio S. Beyond genes: A systematic review of environmental risk factors in specific reading disorder. *Research in Developmental Disabilities*. 2018;82(March):147-152. doi:10.1016/j.ridd.2018.03.005

25. Gu H, Hou F, Liu L, et al. Genetic variants in the CNTNAP2 gene are associated with gender differences among dyslexic children in China. *EBioMedicine*. 2018;34:165-170. doi:10.1016/j.ebiom.2018.07.007

26. Luciano M, Gow AJ, Pattie A, Bates TC, Deary IJ. The Influence of Dyslexia Candidate Genes on Reading Skill in Old Age. *Behavior Genetics*. 2018;48(5):351-360. doi:10.1007/s10519-018-9913-3

27. Luciano M, Evans DM, Hansell NK, et al. A genome-wide association study for reading and language abilities in two population cohorts. *Genes, Brain and Behavior*. 2013;12(6):645-652. doi:10.1111/gbb.12053

28. Gialluisi A, Newbury DF, Wilcutt EG, et al. Genome-wide screening for DNA variants associated with reading and language traits. *Genes, Brain and Behavior*. 2014;13(7):686-701. doi:10.1111/gbb.12158

29. Becker N, Vasconcelos M, Oliveira V, et al. Genetic and environmental risk factors for developmental dyslexia in children: systematic review of the last decade. *Developmental Neuropsychology*. 2017;42(7-8):423-445. doi:10.1080/87565641.2017.1374960

30. Jerrim J, Vignoles A, Lingam R, Friend A. The socio-economic gradient in children's reading skills and the role of genetics. *British Educational Research Journal*. 2015;41(1):6-29. doi:10.1002/berj.3143

31. Dinu V, Zhao H, Miller PL. Integrating domain knowledge with statistical and data mining methods for high-density genomic SNP disease association analysis. *Journal of biomedical informatics*. 2007;40(6):750-760. doi:10.1016/j.jbi.2007.06.002

32. Saul M, Dinu V. Family Rank: A graphical domain knowledge informed feature ranking algorithm. *Bioinformatics (Oxford, England)*. Published online May 19, 2021. doi:10.1093/bioinformatics/btab387

33. Brown JR, Stafford P, Johnston SA, Dinu V. Statistical methods for analyzing immunosignatures. *BMC bioinformatics*. 2011;12:349. doi:10.1186/1471-2105-12-349

34. Bixia X, Ao L, Dinu V, Nowak NJ, Zhao H, Li P. Analytical and clinical validity of whole-genome oligonucleotide array comparative genomic hybridization for pediatric patients with mental retardation and developmental delay. *American journal of medical genetics Part A*. 2008;146A(15):1942-1954. doi:10.1002/ajmg.a.32411

35. Li C, Liu L, Dinu V. Pathways of topological rank analysis (PoTRA): a novel method to detect pathways involved in hepatocellular carcinoma. *PeerJ*. 2018;6:e4571. doi:10.7717/peerj.4571

36. Li C, Dinu V. miR2Pathway: A novel analytical method to discover MicroRNA-mediated dysregulated pathways involved in hepatocellular carcinoma. *Journal of biomedical informatics*. 2018;81:31-40. doi:10.1016/j.jbi.2018.03.013

37. Day SE, Coletta RL, Kim JY, et al. Next-generation sequencing methylation profiling of subjects with obesity identifies novel gene changes. *Clinical epigenetics*. 2016;8:77. doi:10.1186/s13148-016-0246-x

38. Huentelman MJ, Muppana L, Corneveaux JJ, et al. Association of SNPs in EGR3 and ARC with Schizophrenia Supports a Biological Pathway for Schizophrenia Risk. *PloS one*. 2015;10(10):e0135076. doi:10.1371/journal.pone.0135076

39. Bradley BS, Loftus JC, Mielke CJ, Dinu V. Differential expression of microRNAs as predictors of glioblastoma phenotypes. *BMC bioinformatics*. 2014;15:21. doi:10.1186/1471-2105-15-21

40. Gallitano AL, Tillman R, Dinu V, Geller B. Family-based association study of early growth response gene 3 with child bipolar I disorder. *Journal of affective disorders*. 2012;138(3):387-396. doi:10.1016/j.jad.2012.01.011

41. Briones N, Dinu V. Data mining of high density genomic variant data for prediction of Alzheimer's disease risk. *BMC medical genetics*. 2012;13:7. doi:10.1186/1471-2350-13-7

42. Dinu V, Miller PL, Zhao H. Evidence for association between multiple complement pathway genes and AMD. *Genetic epidemiology*. 2007;31(3):224-237. doi:10.1002/gepi.20204

43. Peter B, Dinu V, Liu L, et al. Exome Sequencing of Two Siblings with Sporadic Autism Spectrum Disorder and Severe Speech Sound Disorder Suggests Pleiotropic and Complex Effects. *Behavior genetics*. 2019;49(4):399-414. doi:10.1007/s10519-019-09957-8

44. Boyd A, Golding J, Macleod J, et al. Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology*. 2013;42(1):111-127. doi:10.1093/ije/dys064

45. Eicher JD, Powers NR, Miller LL, et al. Genome-wide association study of shared components of reading disability and language impairment. *Genes, Brain and Behavior*. 2013;12(8):792-801. doi:10.1111/gbb.12085

46. Paracchini S, Steer CD, Buckingham LL, et al. Association of the KIAA0319 dyslexia susceptibility gene with reading skills in the general population. *American Journal of Psychiatry*. 2008;165(12):1576-1584. doi:10.1176/appi.ajp.2008.07121872

47. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology*. 2013;42(1):97-110. doi:10.1093/ije/dys066

48. Wechsler D, Golombok S, Rust J. *WISC-III UK Wechsler Intelligence Scale for Children: UK Manual*.; 1992.

49. Li M, Malins JG, DeMille MMC, et al. A molecular-genetic and imaging-genetic approach to specific comprehension difficulties in children. *npj Science of Learning*. 2018;3(1). doi:10.1038/s41539-018-0034-9

50. Rust J, Golombok S, Trickey G. *WORD. Weschler Objective Reading Dimensions*. Psychological Corp; 1993.

51. Bryant P, Nunes T, Barros R. The connection between children's knowledge and use of grapho-phonic and morphemic units in written text and their learning at school. *British Journal of Educational Psychology*. 2014;84(2):211-225. doi:10.1111/bjep.12030

52. Gathercole SE, Willis CS, Baddeley AD, Emslie H. The children's test of nonword repetition: A test of phonological working memory. *Memory*. 1994;2(2):103-127. doi:10.1080/09658219408258940

53. Nunes T, Byrant P, Olsson J. Learning Morphological and Phonological Spelling Rules: An Intervention Study. *Scientific Studies of Reading*. 2003;7(3):289-307. doi:10.1207/S1532799XSSR0703

54. Neale MD, McKay MF, Childs GH. The Neale Analysis of Reading Ability - Revised. *British Journal of Educational Psychology*. 1986;56(3):346-356. doi:10.1111/j.2044-8279.1986.tb03047.x

55. Rust J. *Weschler Objective Language Dimensions Manual*.; 1996.

56. Torgesen JK, Wagner RK, Rashotte CA. *Test of Word Reading Efficiency–Second Edition,*. 2nd ed. Pro-Ed; 2012. Accessed May 20, 2020. https://www.proedinc.com/Products/13910/towre2-test-of-word-reading-efficiencysecond-edition-complete-kit.aspx

57. Woodcock R, McGrew KS, Mather N. *Woodcock Johnson III Tests of Achievement*. Riverside Publishing; 2001.

58. Wagner RK, Torgesen JK, Rashotte CA, Pearson NA. *Comprehensive Test of Phonological Processing 2*. 2nd ed. Pearson Assessments; 2013.

59. Newcomer PL. *Standardized Reading Inventory.* Pro-Ed; 1986.

60. Dunn LM, Dunn DM. Peabody Picture Vocabulary Test - 4th edtion. *Summary*. Published online 2007.

61. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating Kinship in Admixed Populations. *The American Journal of Human Genetics*. 2012;91(1):122-138. doi:10.1016/j.ajhg.2012.05.024

62. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*. 2012;48(2):1-36.

63. Byrne BM. *A Primer of LISREL: Basic Applications and Programfor Confirmatory Factor Analytic Models*. Springer Science & Business Media; 1989.

64. Browne MW, Cudeck R. Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*. 1992;21(2):230-258. doi:10.1177/0049124192021002005

65. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software2*. 2011;45(3):1-67.

66. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4(1):1-16. doi:10.1186/s13742-015-0047-8

67. Lee LC, Su MT, Cho YC, Lee-Chen GJ, Yeh TK, Chang CY. Multiple epigenetic biomarkers for evaluation of students' academic performance. *Genes, Brain and Behavior*. 2019;18(5):1-10. doi:10.1111/gbb.12559

68. Fan J, Lv J. Sure independene screening for ultrahigh dimensional feature space. *Journal of Royal Statistical Society Series B (Statistical Methodology)*. 2008;70(5):849-911.

69. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006;101(476):1418-1429. doi:10.1198/016214506000000735

70. Reimand J, Arak T, Adler P, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic acids research*. 2016;44(W1):W83-W89. doi:10.1093/nar/gkw199

71. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303

72. Kere J. Molecular genetics and molecular biology of dyslexia. *Wiley Interdiscip Rev Cogn Sci*. 2011;2(4):441-448. doi:10.1002/wcs.138

73. Anthoni H, Sucheston LE, Lewis BA, et al. The Aromatase Gene CYP19A1: Several Genetic and Functional Lines of Evidence Supporting a Role in Reading, Speech and Language. *Behavior Genetics*. 2012;42(4):509-527. doi:10.1007/s10519-012-9532-3

74. Lee JJ, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*. 2018;50(8):1112-1121. doi:10.1038/s41588-018-0147-3

75. Maeta K, Hattori S, Ikutomo J, et al. Comprehensive behavioral analysis of mice deficient in Rapgef2 and Rapgef6, a subfamily of guanine nucleotide exchange factors for Rap small GTPases possessing the Ras/Rap-associating domain. *Molecular Brain*. 2018;11(1). doi:10.1186/s13041-018-0370-y

76. Gialluisi A, Andlauer TFM, Mirza-Schreiber N, et al. Genome-wide association scan identifies new variants associated with a cognitive predictor of dyslexia. *Translational Psychiatry*. 2019;9(1):77. doi:10.1038/s41398-019-0402-0

77. Landi N, Frost SJ, Mencl WE, et al. The COMT Val/Met polymorphism is associated with reading-related skills and consistent patterns of functional neural activation. *Developmental Science*. 2013;16(1):13-23. doi:10.1111/j.1467-7687.2012.01180.x

78. Plomin R, Kovas Y. Generalist Genes and Learning Disabilities. *Psychological Bulletin*. 2005;131(4):592-617. doi:10.1037/0033-2909.131.4.592

79. Kere J. The molecular genetics and neurobiology of developmental dyslexia as model of a complex phenotype. *Biochemical and Biophysical Research Communications*. 2014;452(2):236-243. doi:10.1016/j.bbrc.2014.07.102

80. Bieder A, Yoshihara M, Katayama S, et al. Dyslexia Candidate Gene and Ciliary Gene Expression Dynamics During Human Neuronal Differentiation. *Molecular Neurobiology*. 2020;57(7):2944-2958. doi:10.1007/s12035-020-01905-6

81. Luciano M, Gow AJ, Pattie A, Bates TC, Deary IJ. The Influence of Dyslexia Candidate Genes on Reading Skill in Old Age. *Behavior Genetics*. 2018;48(5):351-360. doi:10.1007/s10519-018-9913-3

82. Younger JW, Lee KW, Demir-Lira OE, Booth JR. Brain lateralization of phonological awareness varies by maternal education. *Developmental Science*. 2019;22(6):1-13. doi:10.1111/desc.12807

83. Orban P, Lungu O, Doyon J. Motor sequence learning and developmental dyslexia. *Annals of the New York Academy of Sciences*. 2008;1145:151-172. doi:10.1196/annals.1416.016

84. Hari R, Helenius P. Deficit of temporal auditory processing in dyslexia adults. *Neuroscience Letters*. 1996;205:138-140.

85. Caccia M, Lorusso ML. The processing of rhythmic structures in music and prosody by children with developmental dyslexia and developmental language disorder. *Developmental Science*. Published online 2020. doi:10.1111/desc.12981

86. Valdois S, Roulin JL, Line Bosse M. Visual attention modulates reading acquisition. *Vision Research*. 2019;165(February):152-161. doi:10.1016/j.visres.2019.10.011

87. Facoetti A, Zorzi M, Cestnick L, et al. The relationship between visuo-spatial attention and nonword reading in developmental dyslexia. *Cognitive Neuropsychology*. 2006;23(6):841-855. doi:10.1080/02643290500483090

88. Tilot AK, Khramtsova EA, Liang D, et al. The Evolutionary History of Common Genetic Variants Influencing Human Cortical Surface Area. *Cerebral Cortex*. Published online 2020:1-15. doi:10.1093/cercor/bhaa327

89. Davis OSP, Band G, Pirinen M, et al. The correlation between reading and mathematics ability at age twelve has a substantial genetic component. *Nature Communications*. 2014;5(May). doi:10.1038/ncomms5204

90. Field LL, Shumansky K, Ryan J, Truong D, Swiergala E, Kaplan BJ. Dense-map genome scan for dyslexia supports loci at 4q13, 16p12, 17q22; suggests novel locus at 7q36. *Genes, Brain and Behavior*. 2013;12(1):56-69. doi:10.1111/gbb.12003

91. Harlaar N, Meaburn EL, Hayiou-Thomas ME, et al. Genome-wide association study of receptive language ability of 12-year-olds. *Journal of Speech, Language, and Hearing Research*. 2014;57(1):96-105. doi:10.1044/1092-4388(2013/12-0303)

92. Nudel R, Simpson NH, Baird G, et al. Genome-wide association analyses of child genotype effects and parent-of-origin effects in specific language impairment. *Genes, Brain and Behavior*. 2014;13(4):418-429. doi:10.1111/gbb.12127

93. Scerri TS, Darki F, Newbury DF, et al. The dyslexia candidate locus on 2p12 is associated with general cognitive ability and white matter structure. *PLoS One*. 2012;7(11):e50321. doi:10.1371/journal.pone.0050321

94. St Pourcain B, Cents RAM, Whitehouse AJO, et al. Common variation near ROBO2 is associated with expressive vocabulary in infancy. *Nature Communications*. 2014;5:4831. doi:10.1038/ncomms5831

Table 1. Previously published genetic association studies for reading traits

Authors (Year) [ref]	**Phenotype**	**Reading Measure(s)**	**Association Method**
Carrion-Castillo et al (2020) [12]	Language and reading skill	Word reading fluency, nonword reading fluency, phonological awareness, rapid automatized naming	Multivariate association
Davis et al (2014) [89]	Reading and math	Test of Word Reading Efficiency at 11years	Genome wide association, bivariate twin analysis, bivariate population analysis
Eicher et al (2013) [45]	Language impairment (2/3 failed measures), Reading disability (3/5 failed measures)	Phoneme deletion at 7years, Single word reading at 7years, Single word reading at 9years, Nonword reading at 9years, Reading comprehension at 9years	Genome wide association
Field et al (2013) [90]	Dyslexia (clinical diagnosis)	Phonological coding dyslexia	Linkage analysis, Genome wide association
Gialluisi et al (2014) [28]	Language and reading skill	Word reading, spelling, phonological decoding, phoneme	Genome wide association meta-analysis

		awareness, orthographic coding, nonword repetition	
Gialluisi et al (2019) [76]	Dyslexia	Word reading, spelling, nonword reading, phonological awareness	Genome wide association meta-analysis
Harlaar et al (2014) [91]	Receptive Language		Genome wide association
Luciano et al (2013) [27]	Language and reading skill	Reading and spelling composite, word reading nonword reading, spelling, reading comprehension	Genome wide association, gene set, meta-analysis
Luciano et al (2018) [26]	Reading skill	Two tests of word reading, self-reported lifetime reading survey	Genome wide association meta-analysis
Nudel et al (2014) [92]	Language impairment		Genome wide association
Paracchini et al (2008) [46]	Reading skill	Reading difficulties, Statutory Assessment Test, word reading, spelling, phoneme awareness, reading	Single marker association, haplotype analysis

		accuracy, nonword reading	
Price et al (2020) [18]	Word reading	Word reading	Genome wide association, genome wide association meta-analysis
Scerri et al (2012) [93]	General cognitive ability	Reading correlation with verbal IQ, reading correlation with performance IQ	Single SNP association
St Pourcain et al (2014) [94]	Expressive language ability		Genome wide association meta-analysis
Truong et al (2019) [17]	Reading related skills	Rapid automatized naming objects, rapid automatized naming letters, rapid alternating stimulus	Multivariate association

Table 2. Demographic information by cohort for children included in genetic analyses.

	ALSPAC	GRaD
N	7977	1341
Age [a]	---	11;5 (2;3)
Male [b]	4036	711
Race/Ethnicity		
White	4965	---
Non-white	185	---
African American	---	812
Hispanic	---	480
Missing	445	---
ADHD	78	131
Bilingual	136	---
Birthweight (grams)	n = 4014 m = 3444.63 (521.95)	---
Mother's education		
CSE	568	---
Vocational	427	---
O Levels	1819	---
A Levels	1479	---
Degree	932	---
Less than 7 years	---	62
7 - 9 years	---	87
10 -11 years	---	97
High school diploma/GED	---	455
Associate's/Trade/Business	---	182

Bachelor's	---	199
Professional or Advanced degree	---	117
Missing	370	236
Vocabulary		
WISC - Vocabulary	n = 4185, m = 11.52 (4.29)	---
PPVT	---	94.76 (15.58)
Receptive language	n = 4205 m = 7.60 (1.91)	---

Notes. [a] Because we used behavioral data from three time points, we do not report mean age for the ALSPAC sample. [b] Missing sex for 6 participants

Table 3. Reading data for both cohorts.

	ALSPAC	GRaD
Phonological Awareness		
Phoneme deletion	20.78 (9.22) 5513	---
Nonword repetition	7.36 (2.45) 5537	---
CTOPP-Blending Words	---	9.22 (2.62) 1323
Single Word Reading		
Single word reading at 7	29.12 (8.87) 5518	---
Single word reading at 9	7.73 (2.23) 5546	---
TOWRE - Sight Word Efficiency	---	94.45 (13.80) 1336
WJ-III - Letter-Word Identification	---	95.23 (14.56) 1334
Nonword Reading		
Nonword reading at 9	5.36 (2.43) 5542	---
TOWRE - Phonemic Decoding Efficiency	---	93.28 (15.61) 1336
WJ-III - Word Attack	---	94.39 (11.82) 1333

Spelling		
Spelling at 7	26.79 (12.23) 5445	---
Spelling at 9	10.48 (3.27) 5537	---
WJ-III - Spelling	---	94.34 (16.81) 1333
Connected Text		
NARA - rate	106.15 (12.25) 4972	---
NARA - accuracy	105.06 (13.19) 4893	---
NARA - comprehension	101.32 (11.41) 4983	---
SRI - word recognition	---	7.07 (4.13) 1323
SRI - comprehension	---	7.48 (3.91) 1324

Notes. CTOPP = Clinical Test of Phonological Processing; TOWRE = Test of Word Reading Efficiency; WJ-III = Woodcok-Johnson Tests of Achievement 3rd Edition; NARA = Neale Analysis of Reading Ability; SRI = Standardized Reading Inventory. Mean (Standard Deviation), Number of children with data

Table 4. Significant SNPs after multiple test correction identified by GWA for general reading ability in ALSPAC cohort

SNP	Chr	Position	Gene(s)	Beta value	FDR p-value
			Analysis 1		
rs181384543	X	22779030	PTCHD-1AS	-2.91	0.007
rs191271648	X	22690028	PTCHD-1AS	-2.88	0.057
rs185462130	X	22728643	PTCHD-1AS	-5.62	0.057
rs185003220	X	22733372	PTCHD-1AS	-2.78	0.057
			Analysis 2		
rs148283392	15	81870924	AC104041.1	-3.16	0.035
rs28532251	15	81826579	AC104041.1	-3.13	0.035
rs76099038	15	81854231	AC104041.1	-3.23	0.035
rs80179050	15	81893828	AC104041.1	-3.22	0.035
rs75707864	15	81807607	AC104041.1	-3.15	0.035
rs113517011	15	81761276	AC104041.2	-1.82	0.047
rs9911708	17	.	.	-0.16	0.014
rs9904090	17	.	.	-0.16	0.014
rs11867791	17	.	.	-0.16	0.014
rs16975961	17	.	.	-0.16	0.014
rs7222670	17	.	.	-0.16	0.014
rs12449913	17	.	.	-0.16	0.017
rs9908077	17	.	.	-0.15	0.036
rs9897225	17	.	.	-0.15	0.036
rs41381246	17	.	.	-0.15	0.036
rs9895773	17	.	.	-0.15	0.037

rs11077508	17	.	.	-0.15	0.037
rs1792694	18	56117088	AC006305.1, LINC01905, LINC01539	-0.16	0.052
rs1468647	18	56100396	AC006305.1, LINC01905, LINC01540	-0.16	0.052
rs1789594	18	56103335	AC006305.1, LINC01905, LINC01541	-0.16	0.052
rs1792705	18	56103062	AC006305.1, LINC01905, LINC01542	-0.15	0.052
rs143101843	18	74720538	ZNF407	-45.49	0.052
rs181384543	X	22779030	PTCHD-1AS	-2.73	0.031

Notes. SNP = single nucleotide polymorph. Chr = chromosome, FDR = false discovery rate. A period means that the SNP was imputed and therefore could not be mapped to a specific location.

Table 5. *Elastic net results for general reading ability in the ALSPAC* .

SNP	Chr	Gene(s)	Coefficient
		Analysis 1	
rs183880474	3	FHIT	-0.944
rs79439102	3	LSAMP	-0.837
rs17140364	3	HGD	-0.067
rs4676819	3	HGD	-0.025
rs192775885	4	YTHDC1	-1.721
rs11099761	4	LRBA	-0.175
rs55703414	4	RAPGEF2	-0.046
rs7681750	4	RAPGEF2	-0.042
rs139080150	4	LNX1,LNX1-AS1,AC058822.1	0.165
rs7673630	4	LINC02432	0.060
rs161731	5	LINC02062	0.057
rs60198643	6	UTRN	-0.815
rs7798197	7	HDAC9	0.061
rs34346046	7	AOAH	0.087
rs112364229	8	SNTG1	-0.061
rs77659190	8	SNTG1	-0.038
rs57575663	8	SNTG1	-0.058
rs75276483	8	SNTG1	-0.052
rs60604540	8	SNTG1	-0.034
rs144540231	8	ARHGEF10,AC019257.8	0.390
rs5025174	8	CSMD1	0.057
rs877365	9	TLN1	0.072
rs11021850	11	GALNT18	-0.098
rs2268119	12	GRIN2B	-0.041
rs143145485	13	SUCLA2	-0.229
rs147686493	13	SUCLA2	-0.223
rs17127713	14	SAMD4A	-0.210
rs1075938	15	DNAAF4	-0.039
rs80033521	16	ROGDI	-2.309
rs144280335	17	GCGR	-0.660
rs7220982	17	ASPA,SPATA22	0.048
rs182289205	18	AP005328.1	-0.799
rs150818953	18	AP005328.1	-0.778
rs151045653	18	ZNF236-DT	-2.218
rs149592081	18	TCF4,TCF4-AS1	0.621
rs459962	21	SAMSN1,SAMSN1-AS1	0.030
rs151110013	X	MID1	-0.080

rs4830707	X	AC003666.1	0.031
rs183845081	X	FHL1	0.432
rs1163203	.	.	-0.028
rs10231255	.	.	-0.051
rs28735279	.	.	-0.075
rs13184015	.	.	-0.079
rs72843127	.	.	-0.103
rs186659202	.	.	-0.280
rs185046620	.	.	-0.287
rs189820864	.	.	-0.456
rs114157996	.	.	-0.531
rs146478707	.	.	-0.534
rs138900347	.	.	-0.546
rs191544303	.	.	-0.841
rs73634038	.	.	-0.914
rs137926990	.	.	-1.694
rs114585831	.	.	-1.781
rs7972680	.	.	-1.942
rs138653067	.	.	-2.013
rs2957954	.	.	2.083
rs191884462	.	.	0.264
rs62602151	.	.	0.218
rs142791311	.	.	0.142
rs77236383	.	.	0.124
rs4292642	.	.	0.089
rs467650	.	.	0.060
rs57301765	.	.	0.038
		Analysis 2	
rs184240058	1	AGTRAP	-0.705
rs6424160	1	IFNLR1	-0.213
rs117172756	2	KCNE4	-1.357
rs57993969	2	SLC16A14	-0.084
rs957509	3	MAGI1	0.067
rs75807148	3	SPATA16	0.168
rs148902155	5	NSD1	-1.508
rs2143340	6	TDP2	-0.070
rs793862	6	DCDC2	0.010
rs188975652	6	TENT5A	0.222
rs2158591	7	NXPH1	-0.070
rs68006848	7	VPS41	-0.060
rs923875	7	FOXP2	-0.020
rs7788833	7	HDAC9	0.085
rs6990556	8	LINC02235	0.017
rs17187458	11	SCUBE2	-0.207
rs181139150	11	LRP5	-0.945

rs149226354	11	CLMP	-0.120
rs4901526	14	SAMD4A	-0.188
rs148918681	15	TEX9	-1.508
rs28532251	15	AC104041.1	-1.414
rs76099038	15	AC104041.1	-0.078
rs8034835	15	CYP19A1,MIR4713HG	0.010
rs6564903	16	CMIP,AC092135.1	-0.023
rs439984	17	OR3A2	-0.058
rs72813059	17	MYO1D,AC079336.4	-0.673
rs184413500	17	ASIC2	-0.310
rs17687054	17	AC005747.1	0.056
rs12948443	17	AC005747.1	0.022
rs114322121	18	LINC01895	-2.132
rs141908149	18	LINC01895	-0.117
rs5918826	X	OPHN1	0.264
rs183965210	X	MID1	-0.419
rs185185408	.	.	-1.945
rs138026430	.	.	-1.929
rs185066499	.	.	-1.910
rs143618695	.	.	-1.730
rs113623735	.	.	-1.665
rs149743247	.	.	-1.389
rs188864561	.	.	-1.153
rs74022314	.	.	-1.025
rs5976372	.	.	-0.739
rs188463859	.	.	-0.686
rs4708270	.	.	-0.457
rs57438272	.	.	-0.302
rs79166645	.	.	-0.150
rs800956	.	.	-0.148
rs76784138	.	.	-0.146
rs9911708	.	.	-0.130
rs12187304	.	.	-0.100
rs7099623	.	.	-0.081
rs6547089	.	.	-0.077
rs141291309	.	.	-0.077
rs7729009	.	.	-0.068
rs1355077	.	.	-0.066
rs5976373	.	.	-0.041
rs1839897	.	.	0.080
rs145901808	.	.	0.084
rs118045883	.	.	0.231
rs80317159	.	.	0.239
			Shared
			Analysis 1 Analysis 2

rs16889556	6	KIAA0319	0.042	0.048
rs936146	7	FOXP2	-0.024	-0.019
rs1079727	11	DRD2	-0.033	-0.027
rs57809907	15	DNAAF4,DNAAF4-CCPG1	-0.048	-0.048
rs1902586	15	CYP19A1,MIR4713HG	0.050	0.021
rs77641439	15	DNAAF4,DNAAF4-CCPG1	0.081	0.061
rs11860694	16	ATP2C2	-0.020	-0.026

Notes. NMD = nonsense mediated decay transcript variant

Variant type(s)
intron, non-coding transc
intron
intron, non-coding transc
intron, non-coding transc
3' UTR
NMD, intron, non-coding
intron
intron
intron, non-coding transc
intron, non-coding transc
intron, non-coding transc
intron
3' UTR
intron, non-coding transc
NMD, intron, non-coding
NMD, intron
NMD, intron
NMD, intron
NMD, intron
NMD, intron
intron
intron
intron
intron
NMD, intron, non-coding
NMD, intron, non-coding
intron, non-coding transc
5' UTR
3' UTR, NMD, intron, mis
intron
intron
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transc
NMD, intron, non-coding
intron, non-coding transc
intron

intron, non-coding transc
5' UTR, intron

NMD, intron, non-coding
intron
intron, non-coding transc
intron
intron, non-coding transc
intron
NMD, intron
intron, non-coding transc
intron
intron
intron
intron, non-coding transc
NMD, intron, non-coding
intron, non-coding transc
intron, non-coding transc
intron, non-coding transc
NMD, intron

intron
intron, non-coding transc
intron
intron, non-coding transc
intron, non-coding transc
intron, non-coding transc
intron, non-coding transc
intron, non-coding transc
intron, non-coding transc
intron, non-coding transc
intron
intron
intron, non-coding transc
intron, non-coding transc
intron, non-coding transc
3' UTR

intron
NMD, intron
intron, non-coding transc
3' UTR, NMD, intron, nor
intron, non-coding transc
3' UTR, NMD, intron, mis
intron, non-coding transc

Table 6. *Elastic net results for mapped informative SNPs for the GRaD da*

SNP	Chr	Gene(s)	Coefficient
rs73025663	1	GORAB-AS1	-0.086
rs73025679	1	GORAB-AS1	-0.021
rs143469658	1	GORAB-AS1	-0.021
rs10863512	1	SLC30A10	-0.082
rs2338774	2	AC096570.1	0.038
rs73920150	2	AC096570.1	0.048
rs55924242	2	AC096570.1	0.058
rs56096244	2	AC096570.1	0.056
rs111561745	2	AC096570.1	0.043
rs895547	2	GLI2	-0.120
rs73119726	3	ADAMTS9-AS2	0.097
rs608715	3	PTPRG	0.100
rs4301023	3	CADM2	-0.016
rs7651996	3	CADM2	-0.016
rs79439102	3	LSAMP	-0.105
rs1680073	4	CTBP1-DT	0.035
rs1128427	4	MAEA	0.149
rs7681750	4	RAPGEF2	0.048
rs2115146	5	EGFLAM	0.053
rs1896658	5	EGFLAM	0.052
rs4917151	7	ABCA13	-0.083
rs7787626	7	AC091685.1	-0.039
rs58574981	7	GUSBP6	-0.029
rs7787626	7	GUSBP6	-0.039
rs1581538	7	SEM1	-0.164
rs62504391	8	CSMD1	0.219
rs61496947	8	MYOM2	0.023
rs58311037	8	MYOM2	0.023
rs73699003	8	RIMS2	0.153
rs4750032	10	CELF2	0.052
rs11601105	11	AC131571.1	0.018
rs11031434	11	AC131571.1	0.027
rs1232202	11	AC131571.1	0.027
rs11601105	11	ELP4	0.018
rs11031434	11	ELP4	0.027
rs1232202	11	ELP4	0.027
rs10844267	12	AC010186.4	0.034
rs35561167	12	AC010186.4	0.035
rs12809846	12	AC010186.4	0.035
rs6488031	12	AC092821.3	0.066
rs10844267	12	AC092821.3	0.034
rs35561167	12	AC092821.3	0.035
rs12809846	12	AC092821.3	0.035

rs1915358	12	ACSS3	-0.084
rs17577911	13	AL356295.1	-0.148
rs10142928	14	KLHL28	0.167
rs77641439	15	DNAAF4, DNAAF4-CCPG1	0.142
rs10046	15	CYP19A1, MIR4713HG	-0.017
rs61433139	16	TK2	-0.114
rs12606138	18	NEDD4L	0.033
rs7234617	18	CCBE1	-0.003
rs66548292	22	EFCAB6	-0.042
rs77571308	22	EFCAB6	-0.046
rs12628638	22	EFCAB6	-0.050
rs73422423	22	EFCAB6	-0.046

Notes. NMD = nonsense mediated decay

taset when including demographic features.

Variant type
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
NMD, intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron
intron
intron
intron, non-coding transcript
NMD, missense, non-coding transcript exon
intron
NMD, intron
NMD, intron
NMD, intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
NMD, intron, non-coding transcript
intron, non-coding transcript
intron
intron
intron
NMD, intron
NMD, intron, non-coding transcript
NMD, intron, non-coding transcript
3' UTR, NMD, intron, non-coding transcript
NMD, intron, non-coding transcript
NMD, intron, non-coding transcript
3' UTR, NMD, intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript

intron, non-coding transcript
intron, non-coding transcript
intron
3' UTR, NMD, intron, missense, non-coding trans
3' UTR, intron, non-coding transcript
NMD, intron
NMD, intron, non-coding transcript
intron
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript
intron, non-coding transcript