Pipeline for Identifying Genetic and Demographic Predictors of Latent Reading Ability: Demonstration in ALSPAC Cohort

Hope Sparks Lancaster[1], Valentin Dinu[2], Amit Arora[2], Jing Li[3], Jeffrey R Gruen[4,5]

[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: Pipeline to Identify Predictors of Reading

Keywords: reading, GWAS, machine learning, ALSPAC

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:

Manuscript word count: 4,749

Abstract

Purpose: Reading ability is a complex skill utilizing multiple proficiencies and that develops through interactions between genetic and environmental factors. This study presents an alternative analytic pipeline to identify key genetic and demographic contributors to reading ability.

Methods: We analyzed data from the Avon Longitudinal Study of Parents and Children (ALSPAC; N = 3 232) using a multi-step analytical pipeline. To reduce measurement error, we generated a latent reading ability score. We selected single nucleotide polymorphisms (SNPs) based on existing literature and genome-wide association studies (GWAS). We applied elastic net regression to identify informative predictors in two models, a SNP-only model and a SNP-plus demographic, environmental, and behavioral variables model. We compared the SNP-based heritability estimates and R^2 from the fitted models. We also performed pathway enrichment analysis on the informative SNPs.

Results: The traditional GWAS identified one genome-wide significant SNP on chromosome X and produced a moderate heritability estimate of .23 (SE = 0.07). We included 148 SNPs in the elastic net models. The SNP-only model identified 61 informative SNPs (R^2 = .12), whereas the SNP-plus model identified 96 informative SNPs (R^2 = .32). The SNP-plus model also showed that several behavioral characteristics positively predicted latent reading ability. Enrichment analysis revealed overrepresentation of several biological pathways among the informative SNPs.

Conclusions: This study shows that our analytic pipeline can identify important genetic and demographic predictors of reading ability, providing a powerful alternative to traditional methods and contributing to a deeper understanding of the factors that drive reading development.

Efficient and adequate reading depends on several neurocognitive processes and pathways[1] that support critical component skills. These skills include working memory, phonological awareness,[2] decoding,[3] rapid automatized naming (also referred to as fluency)[4], and language comprehension[3]. Substantial evidence shows that both genetic and non-genetic factors contribute to reading ability[5–7]. Although research on specific genetic variants related to reading continues to evolve, recent studies have made significant progress in this area[8,9]. Behavioral work has also highlighted the influence of various factors on reading development, including demographic characteristics, environmental exposures, and cognitive skills[10].

Despite this progress, most genetic studies have used traditional approaches with relatively small samples, which limits the identification of relevant genetic predictors. Many studies have also lacked the capacity to integrate genetic and non-genetic data in comprehensive analyses. Three recurring methodological limitations appear in much of the prior work. First, studies often rely on simplified phenotypes that differ across projects. Second, many analyses examine one variable at a time, which limits the ability to detect complex relationships. Third, researchers frequently omit non-genetic factors from analyses that focus on genetic data.

This study aims to demonstrate how alternative analytic methods can address these limitations. By integrating genetic, demographic, environmental, and behavioral data, we seek to identify meaningful associations with reading ability and offer a more comprehensive understanding of its underlying mechanisms.

Shifting from a Reduced Phenotype to a Multidimensional Phenotype

Many genetic studies use only one or two measures of reading or reading-related tasks to represent reading ability. For example, multiple studies have utilized word reading[11,12] or single word spelling tasks[13] as phenotype proxies for reading ability. However, these limited measures fail to capture the full complexity of reading as a multidimensional skill. To address this issue, some researchers have applied multi-trait genetic association methods, such as multi-trait analysis of genome-wide association (MTAG)[8,14]. However, MTAG analyses typically

rely on harmonized scores from individual tasks, which may introduce and even amplify measurement error. This discrepancy between observed scores and the underlying true ability could contribute to the lack of replication across genetic studies in different cohorts.

One effective strategy for reducing measurement error involves using latent scores as phenotypes in genetic analysis. A few studies[12,15,16] have taken this approach by employing principal component analysis (PCA) or composite scores. However, PCA presents several limitations. First, its results may not align with theoretical models and often lack clear interpretability. Second, researchers must decide how many components to retain, which introduces subjectivity. Third, each orthogonal component requires separate interpretation, which can complicate analysis.

Confirmatory factor analysis (CFA) offers a stronger alternative. CFA provides greater control over measurement error, produces interpretable and theoretically grounded scores, and supports replication within and across datasets. Like PCA, CFA allows researchers to compare the fit of multiple models and select the one that best suits the data. In this study, we used CFA to generate a latent reading ability score. This approach enabled us to create a phenotype that is both adaptable and grounded in theory, while also remaining sensitive to the structure of the data.

Moving from One-At-A-Time to Many-At-Once

Previous studies have established genetic contributions to performance on reading-related quantitative traits, such as nonword reading[9]. Like other complex phenotypes, reading ability reflects the combined influence of many genes. However, most genetic studies have used one-variant-at-a-time statistical approaches, such as genome-wide association studies (GWAS). These methods assess each genetic variant independently and apply strict multiple testing corrections, which often result in data loss—especially in the small sample sizes common in reading genetics research. Because of these stringent thresholds, most studies fail to identify

any genetic variants, such as single nucleotide polymorphisms (SNPs), that survive correction. This limitation has slowed progress in uncovering the genetic architecture of reading.

Troung et al.[17] found only one SNP that passed multiple test correction for rapid automatized naming, a skill that predicts later reading ability. Similarly, Eising et al.[18] conducted a large meta-GWAS of multiple reading-related traits and identified only one SNP that survived multiple test correction. One-variant-at-a-time approaches also cannot account for interactions or shared variance across genetic markers. Research on developmental dyslexia provides evidence of gene-gene interactions. For instance, variants within READ1 and DCDC2/KIAA0319 have been shown to jointly affect single word reading, nonword reading, spelling, phoneme deletion, and comprehension[19,20]. These findings underscore the importance of analyzing multiple variants together to capture additional predictive value and biological insight.

To address these limitations, researchers can apply statistical models that evaluate many features simultaneously. Machine learning offers one solution, as it can identify informative predictors while accounting for correlations across variables. Elastic net regression, a type of regularized machine learning model, provides an effective tool for this purpose[21,22]. Elastic net models perform feature selection in high-dimensional data and can detect informative genetic variants by leveraging their shared structure and interactions. In this study, we combine a traditional GWAS with an elastic net model to improve the identification of genetic predictors associated with reading ability.

Incorporating Multiple Data Types

There is evidence for gene-environment interplay for reading development[23,24]. Behavioral research has shown that factors such as biological sex, birth weight, gestational age, maternal education level, and language ability relate to reading outcomes[25]. For example, Mascheretti and colleagues[25] reported that maternal smoking during pregnancy, birth weight, and socioeconomic status can influence how the DYX1C1 gene affects susceptibility to

developmental dyslexia. These findings suggest that reading development arises from a complex interaction between genetic, environmental, and demographic factors.

Despite this complexity, most traditional genetic studies include non-genetic variables only as covariates. This practice limits our ability to explore how genetic and non-genetic factors may jointly contribute to reading development. Research designs and analytic approaches often lack the flexibility to support fully integrated models. As a result, few studies have examined genetic, demographic, and environmental data within the same analysis. To address this gap, we have developed methods that integrate multiple data types using machine learning techniques. In particular, we have applied elastic net models to examine interactions between genetic and environmental factors across various health-related outcomes[26–29]. We have also used these approaches in prior work on reading ability[30]. By applying elastic net modeling to integrate diverse predictors, we aim to advance understanding of how multiple influences contribute to reading ability and to refine hypotheses for future research.

Study Purpose

This study aimed to demonstrate a pipeline for identifying informative genetic, demographic, and environmental contributors to reading ability. We tested two primary hypotheses. First, we predicted that combining a multidimensional reading phenotype with elastic net models and traditional genetic approaches would identify more informative markers than using traditional methods alone. Second, we hypothesized that the genetic markers identified through this approach would show enrichment in functional biological processes related to brain development. To test these hypotheses, we used data from the Avon Longitudinal Study of Parents and Children (ALSPAC)[31,32], a well-established population-based cohort. This dataset allowed us to evaluate how our alternative methods can reveal key genetic and non-genetic features associated with reading ability.

Material and Methods

Subjects, Recruitment, and DNA Collection

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a population-based birth cohort that has been described extensively in previous research[31,32]. The cohort includes 15 447 pregnancies, which resulted in 15 658 fetuses. Of these, 14 901 children were alive at one year of age. Beginning at age seven, ALSPAC invited participants to annual "Focus" sessions, which collected a wide range of behavioral measures, including assessments of reading and language skills.

For this study, we used data from parent surveys and behavioral assessments related to reading, language, hearing, and cognition. These data were collected during the Focus sessions conducted at ages seven, eight, and nine. A searchable data dictionary and variable search tool, available on the ALSPAC website (http://www.bristol.ac.uk/alspac/researchers/our-data/), provides full details of the available dataset.

After completing quality control and genotype imputation procedures, ALSPAC included genomic data for 8 237 participants. However, not all participants had the necessary behavioral data for this analysis. For our final sample, we included 3 232 participants with both complete genetic and behavioral data.

Additional details about recruitment, inclusion and exclusion criteria, behavioral assessments, demographic characteristics, and DNA collection procedures are provided in Supplementary Methods and in Supplemental Tables S1 and S2.

Statistical Analysis

Our analytic pipeline included four primary steps: (1) generating a latent reading ability phenotype using confirmatory factor analysis, (2) screening single nucleotide polymorphisms (SNPs) based on prior knowledge and genome-wide association studies (GWAS), (3) identifying informative SNPs using elastic net modeling, and (4) linking informative SNPs to functional biological processes. We also compared the amount of variance explained by traditional GWAS and elastic net models. To estimate this, we used SNP-based heritability for GWAS and R^2 values for the elastic net models.

We applied this pipeline in two stages. First, we modeled the predictive value of SNPs alone. Then, we included demographic, environmental, and behavioral variables to evaluate the combined effects of genetic and non-genetic features. Commented code for all data analyses is provided in the Supplementary Information.

Confirmatory factor analysis. To construct a latent reading ability phenotype, we used confirmatory factor analysis (CFA) with the lavaan package in R[33]. We selected reading-related variables that included single word reading at ages 7 and 9, nonword reading at age 9, spelling at ages 7 and 9, and the Neale Analysis of Reading Ability (NARA)[34] measures for rate, accuracy, and comprehension. We applied robust modeling procedures to account for missing data. A regression-based equation was then used to compute the latent reading ability score.

SNP screening. To constrain the high-dimensionality of the genomic data, we screened SNPs using two methods: (1) review of prior literature and (2) genome-wide association analyses. We examined candidate gene, GWAS, and family-based studies published between 1990 and 2019 that investigated reading ability or reading disability. From this review, we selected 50 SNPs that were reported at least twice to be associated with reading and that were available in the ALSPAC dataset.

We conducted GWAS using PLINK2[35], analyzing each chromosome separately following methods established in previous work[36]. We specified a generalized linear model with latent reading ability as the outcome variable. We accounted for genetic similarity[*] by including the top three principal components. From the GWAS results, we selected the top 100 SNPs based on uncorrected p-values across all chromosomes and included any SNPs with a false discovery rate (FDR-BH) of 0.1 or less for the elastic net analysis.

---

[*]Genetic similarity is a more accurately descriptive term for the genetic relatedness to describe similarity between participants and reference panels due do common genetic ancestors. Genetic similarity can describe study population as well as control for genetic background.

SNP-based heritability. To estimate SNP-based heritability, we combined summary statistics across autosomes from the GWAS. We used linkage disequilibrium score regression (LDSC)[37] to process the summary statistics and calculate heritability estimates. This procedure involved merging alleles and estimating linkage disequilibrium (LD) scores using the 1000 Genomes Project European HapMap3 reference data.

Elastic net modeling. We used elastic net regression to identify informative predictors of reading ability. Elastic net models apply regularization to perform variable selection in high-dimensional settings[21,22]. This approach incorporates two penalty terms into the regression loss function. The first term (L1 norm) forces coefficients of less influential predictors to zero, effectively performing feature selection. The second term (L2 norm) allows the model to retain groups of correlated predictors, such as linked SNPs.

We used five-fold cross-validation to determine the optimal tuning parameters for both penalties. During each iteration, we trained the model on four folds and tested it on the fifth, repeating the process until every fold had served as the test set. We repeated this cross-validation ten times to ensure consistent parameter selection. After identifying the best parameters, we refit the final model using the full dataset to estimate the regression coefficients.

We defined informative features as those with absolute beta values greater than 0.019. We entered these features into linear regression models to estimate the variance explained in reading ability. We then compared the resulting models to a null model and to each other. In Model 1, we included only the SNPs selected during the screening stage. In Model 2, we added demographic, environmental, and behavioral variables. These variables included biological sex, race or ethnicity, nonverbal IQ, birth weight, maternal education level, receptive vocabulary at age 8, receptive language ability, and bilingual language status.

Enrichment analysis. To examine the biological relevance of the SNPs identified through elastic net modeling, we mapped them to genes using the g:SNPense tool in g:Profiler[38]. We then performed enrichment analysis using the g:GOSt function within g:Profiler.

We selected g:Profiler based on recent evaluations that found it to have the most up-to-date and comprehensive repository of biological pathways, drawing from curated sources such as Gene Ontology and KEGG. We submitted the list of informative SNPs to g:GOSt for functional profiling and interpretation of enriched biological processes.

Results

Latent Reading Ability Score

We used confirmatory factor analysis (CFA) to create a latent reading ability score by loading all reading-related variables onto a single factor. The base model showed mixed goodness-of-fit statistics (see Supplementary Table S3). We improved the model fit statistically by adding correlated error terms between observed variables, based on both theoretical rationale and modification indices. However, a comparison between the extracted latent scores from the base and improved models revealed no difference in the final scores, $t(3231) = -1.203e{-}15$, $p = 1$.

Figure 1 presents standardized path loadings from the base model. All observed variables loaded significantly onto the latent factor, with standardized coefficients ranging from 0.745 to 0.929. The model accounted for a substantial portion of variance in reading ability, with an average explained variance of 0.667 (Supplementary Table S3). We found no negative variance estimates. These results indicate that the single-factor model without correlated errors provided a strong enough fit to generate the latent reading ability score.

Genome-Wide Association for SNP Screening

To manage the high dimensionality of genetic data, we used genome-wide association analysis in PLINK2 to screen SNPs before applying multivariate modeling. The analysis used the latent reading ability score as the phenotype. We identified one genome-wide significant SNP after correcting for multiple comparisons: rs181384543 (X:22779030, FDR = 0.007), located within the gene *PTCHD-1AS*.

After combining 101 SNPs identified through GWAS screening with 50 SNPs selected from prior research, we extracted genotype data for 148 SNPs from the ALSPAC dataset (Supplementary Table S4). Three SNPs were unavailable. Using autosomal summary statistics, we estimated a moderate SNP-based heritability of 0.23 (SE = 0.07) for the latent reading ability phenotype.

Elastic Net Models

We fit two elastic net models to identify informative predictors of reading ability. The first model included only SNP data. The second included SNPs along with demographic, environmental, and behavioral variables. Due to the requirement for complete data in elastic net modeling, our analytic sample decreased (ALSPAC: n = 3 647). Informative SNPs are reported in Tables 1 and 2, and model test statistics appear in Supplementary Table S5.

Model 1: SNP-only model. This model identified 62 informative SNPs with absolute beta values greater than 0.019. Thirty-one SNPs had positive coefficients, indicating associations with better reading ability, while 31 had negative coefficients, indicating associations with lower reading ability (Table 1). The positive SNPs mapped to 19 genes and the negative SNPs to 14 genes. Some genes, including *ATCN7* and *DNAAF4*, had both positive and negative associated SNPs.

Most identified SNPs were intronic or non-coding variants. Several SNPs aligned with previous findings in the literature, including variants in *KIAA0319* (chromosome 6), *DRD2* (chromosome 11), *CYP19A1* (chromosome 15), and *DNAAF4* (chromosome 15), supporting replication of prior results from ALSPAC and other cohorts such as the Colorado Reading Disabilities study.

We used linear regression to evaluate model fit and variance explained. The model accounted for 12 percent of variance in latent reading ability, performing significantly better than a null model (Supplementary Table S5; $F(1, 53) = 10.32$, $p < .0001$, Adjusted $R^2 = 0.1193$).

Seven SNP coefficients could not be estimated due to perfect multicollinearity, suggesting redundancy that may allow for model simplification.

Model 2: SNP-plus demographic, environmental, and behavioral features model. To assess the combined influence of genetic and non-genetic factors, we added biological sex, race or ethnicity, nonverbal IQ, birth weight, mother's highest education, receptive vocabulary at age 8, receptive language, and bilingual status to the elastic net model (Table 2). Figure 2 visualizes the relationships between these demographic features and the latent reading ability score.

This model identified 96 informative SNPs, 49 with positive coefficients and 48 with negative coefficients. These SNPs mapped to 31 unique genes across the genome. Twenty-seven of those genes included positive SNPs, and 24 included negative SNPs. Eleven genes included both positive and negative associations, such as *DCDC2*, *KIAA0319*, *GRIN2B*, *CYP19A1*, and *DNAAF4*, all of which have prior associations with dyslexia and reading development. Chromosomes 5 and 17 contributed 24 percent of all informative SNPs, all selected through the GWAS screening stage.

Six demographic features were identified as informative based on beta values exceeding 0.019 in absolute value. These included biological sex, race or ethnicity, nonverbal IQ, receptive language, vocabulary, and mother's education level. All six features had positive coefficients, indicating that higher values were associated with higher latent reading scores. These findings align with prior studies linking cognitive and environmental variables to reading outcomes.

The model explained 32 percent of the variance in latent reading ability (Supplementary Table S5; $F(1, 85) = 18.93$, $p < .0001$, Adjusted $R^2 = 0.3191$, $n = 3254$). It outperformed both the null model ($F(1, 85) = 18.93$, $p < .0001$) and the SNP-only model from ALSPAC ($F(1, 32) = 29.69$, $p < .001$). Seventeen SNP coefficients were not estimated due to multicollinearity and were likely redundant with other features in the model.

We also examined correlations between demographic features and latent reading ability. Vocabulary (r = .42), nonverbal IQ (r = .31), receptive language (r = .29), and mother's education (r = .25) showed moderate positive correlations. Biological sex (r = .07) and race or ethnicity (r = .03) showed weak associations. These weaker correlations suggest that sex and ethnicity may have been selected in the model due to their shared variance with more predictive features.

Enrichment Analysis

We conducted functional profiling of the informative SNPs using the g:GOSt tool in g:Profiler. The results appear in Table 3. For the SNP-only model, we identified 23 significantly enriched biological pathways (adjusted g:SCS $p < 0.05$). Eight of these pathways were identified as likely drivers, including the postsynaptic membrane (GO:0045211, p = .0001) and learning or memory processes (GO:0007611, p = .0226).

In the SNP-plus model, we found 14 enriched pathways, of which six were selected as drivers. These included the plasma membrane region (GO:0098590, p = .0025) and the regulation of dendrite development (GO:0050773, p = .0103). These pathways reflect neurobiological processes relevant to cognitive and language development, supporting the interpretation that our pipeline successfully identified biologically meaningful predictors of reading ability.

Discussion

We examined the genetic contributions to reading ability by integrating confirmatory factor analysis, data imputation, genome-wide association studies (GWAS), multivariate elastic net modeling, and pathway enrichment analysis. We hypothesized that combining (1) a multidimensional reading phenotype and (2) elastic net modeling with traditional genetic approaches would yield more informative genetic markers than using conventional methods alone. We also expected that the identified genetic variants would be overrepresented in biological processes related to brain development.

Our results support these hypotheses. By constructing a latent reading ability score and applying complementary analytic techniques, we addressed several limitations of prior genetic studies on reading. Specifically, we identified one genome-wide significant SNP in the ALSPAC cohort, explained an equal or greater proportion of variance in reading ability than traditional approaches, and linked informative markers to biological pathways involved in brain development and cognitive function. Together, these findings highlight the value of applying integrated, alternative statistical pipelines within small but deeply phenotyped cohorts to advance hypothesis generation in the study of complex traits.

Impact of Latent Trait on Genome-wide Association

Using a latent trait to represent reading ability allowed us to reduce measurement error and increase statistical power. This approach enabled us to identify one genome-wide significant association (rs181384543) after correcting for multiple comparisons. This result demonstrates that refining the phenotype can enhance the detection of genetic signals, even in relatively small samples.

Our GWAS indicated that the C allele of rs181384543, located on chromosome X within the *PTCHD1-AS* gene (Patched Domain-Containing Protein 1, chrXp22.1), was associated with better general reading ability. Previous research has linked disruptions in *PTCHD1* to autism spectrum disorder and intellectual disability[39–41]. Although no prior studies have associated rs181384543 specifically with reading ability or disability, Doust and colleagues reported that several SNPs implicated in autism and intellectual disability also correlated with self-reported dyslexia. These findings support the possibility of shared genetic mechanisms across neurodevelopmental and cognitive traits.

Despite its statistical significance in the GWAS, rs181384543 was not selected by our elastic net models. This may suggest that the SNP contributes little to multivariate interactions or is redundant when modeled alongside other variants. The lack of selection also highlights

how univariate significance does not always translate into predictive value in multivariate models.

This result emphasizes the potential importance of sex-linked alleles in reading development. Given the location of rs181384543 on the X chromosome, future studies should investigate how sex-linked genetic variation influences reading outcomes. To confirm the relevance of this SNP, replication in larger and more diverse cohorts will be necessary.

Variance Explained of Latent Reading Ability

Previous GWAS studies have reported SNP-based heritability estimates ranging from 0.13 for nonword repetition to 0.26 for nonword reading[18]. Our estimate of SNP-based heritability for latent reading ability 0.23 falls within this range and aligns with prior findings. The elastic net model that included only SNPs explained 12 percent of the variance in latent reading ability, which places it at the lower end of the range found in previous research.

The inclusion of child-level demographic, environmental, and behavioral features to the model increased the variance explained to 33 percent. This substantial improvement highlights the critical role of non-genetic factors—such as maternal education, language skills, and cognitive ability—in shaping reading development. In addition to their direct impact, some of these behavioral features may reflect shared genetic architecture with reading ability. For instance, recent genetic studies have identified overlapping associations between reading skills and broader cognitive or educational traits[9,18,42]. It is likely that features such as vocabulary and receptive language were selected in part due to these shared genetic influences.

Despite the improved explanatory power of the combined model, a considerable portion of variance remains unexplained. This underscores the importance of including a broader range of features in future analyses. In this study, we intentionally limited the number of predictors to demonstrate the utility of the pipeline in a focused and interpretable way. However, future applications could incorporate larger sets of genetic variants and a more diverse array of non-genetic features.

Additional non-genetic features could include nonlinguistic measures, such as finger tapping, rhythm judgement, and visual attention, because previous studies suggest a link between motor abilities,[43] auditory perception,[44] and visual abilities[45] with reading and dyslexia. Expanding the scope of predictors would allow for a more comprehensive understanding of the complex factors that shape reading development.

Links with Theoretically Relevant Biological Functions

Our enrichment analyses identified several overrepresented biological processes involved in brain development, including those related to learning, memory, and dendritic regulation. These findings are consistent with prior genetic research on reading ability, which has linked structural brain differences to reading performance[46,47], as well as studies demonstrating genetic correlations between reading and other cognitive traits[9]. Since reading is fundamentally a cognitive skill, we expected biological processes related to brain function and development to emerge in the enrichment results.

Importantly, many of the SNPs driving these results were located in genes previously associated with reading ability and reading disorders. One notable example is the postsynaptic membrane pathway (GO:0045211), which was significantly overrepresented in both the SNP-only and SNP-plus models. This suggests that the biological effects of many informative SNPs may operate within postsynaptic regions of neurons, which play critical roles in learning and synaptic plasticity.

Nine SNPs from the SNP-plus model mapped to this cellular component and were located in genes such as *KIAA0319*, *ATP2C2*, *GRIN2B*, *LIN7A*, *TLN1*, *KCNJ10*, and *TNM2*. Several of these genes—*KIAA0319*,[48,49] *ATP2C2*,[50] and *GRIN2B*[51]—have well-established links to reading-related traits. Additionally, *ATP2C2,*[52,53] and *LIN7A*[54,55] have been implicated in language impairment, further reinforcing the relevance of our findings.

Because these enrichment results reflect known theoretical frameworks and replicate findings from prior studies, they strengthen the validity of our analytic pipeline. Future research

can apply this approach to investigate reading and other complex traits, using enrichment analyses to generate biologically meaningful hypotheses that bridge genotype and phenotype.

Limitations and Future Directions

This study's primary strength lies in its use of statistical procedures designed to mitigate power loss due to multiple testing corrections in small genetic cohorts. We also demonstrated that a latent reading ability model with acceptable fit can perform comparably to one with near-perfect fit. The consistency of the latent scores across models suggests that researchers can prioritize theoretical coherence and practicality over statistical perfection, especially when modeling complex traits across diverse datasets. This insight has important implications for replicating genetic findings in studies that use different instruments or populations.

Despite these strengths, the study faced several limitations related to sample size, dataset constraints, and analytic methods. Although ALSPAC is one of the largest available cohorts for studying child behavioral genetics, its size remains modest by genomic standards. Small sample sizes are a common limitation in studies of reading ability and reduce the power to detect subtle genetic effects. Furthermore, because the ALSPAC data were collected in the 1990s, we had no control over the selected behavioral measures or the genotyping arrays. These constraints limited the range and specificity of available features and SNP coverage. As such, our findings should be interpreted as hypothesis-generating rather than conclusive.

Methodological limitations also affected our analyses. We ran genome-wide association (GWA) screening chromosome by chromosome due to computational limitations, rather than across the whole genome simultaneously. This approach may have introduced overcorrection during principal component adjustment for genetic similarity. Future studies should replicate our pipeline using updated methods for computing genetic similarity on a genome-wide scale.

In addition, elastic net models do not rely on traditional significance testing to determine the relevance of predictors. In prior studies using similar methods, researchers often input several thousand SNPs[56]. In contrast, we limited our input to fewer than 150 SNPs per analysis

to highlight interpretability and methodological transparency. The exclusion of rs181384543—a genome-wide significant SNP in our GWAS—from the elastic net model confirms that feature selection in this context is driven by joint predictive utility rather than univariate significance. Because elastic nets capitalize on intercorrelations among variables, some selected SNPs exhibited multicollinearity, resulting in singularities in subsequent regression models. Still, most features retained in our models have strong support from existing literature and theory, reinforcing the credibility of our pipeline.

Future research should expand the number of input features, including a wider range of genetic markers and non-genetic variables such as motor skills, rhythm perception, and attention measures. This expansion will help refine models of reading ability and further validate the utility of our pipeline in more diverse contexts.

Conclusions

This study demonstrates how researchers can leverage relatively small but richly phenotyped cohorts to explore the complex genetic and environmental architecture of reading ability. By integrating traditional GWAS with machine learning techniques such as elastic net modeling, we developed a pipeline that successfully identified informative predictors, replicated known genetic associations, and revealed biologically meaningful pathways.

Our findings reinforce the view that reading ability is a multifactorial trait shaped by both genetic and non-genetic influences. We showed that constructing a latent reading phenotype improves power in traditional GWAS, and that elastic net models can detect relevant patterns even without relying on p-values. These models did not generate arbitrary or spurious results but instead produced findings aligned with current theoretical and empirical knowledge.

This analytic approach offers a flexible framework for hypothesis development that can be applied to other complex traits in developmental science. Future research teams can adapt this pipeline to their own data to deepen our understanding of reading development and its

broader cognitive and biological underpinnings. By doing so, researchers can continue to refine models of learning that are grounded in both statistical rigor and developmental relevance.

Acknowledgements

We 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. Thank you to Xiaonan Liu for his help with data analyses on the F32 project. Thank you to everyone who reviewed or provided feedback on a version of this paper.

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. The first author was supported by a National Institutes of Health F32 postdoctoral training grant (1F32HD089674-01A1; PI: HSL). JRG was supported by the generous support of the Manton Foundation, the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01 NS043530, P50 HD027802, P50 HD052120), Kurt and Tammy Mobley, Anna Nordberg, and Tim and Cindy Wollaeger.

Conflict of Interests

The authors declare no other competing interests.

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). Consent for biological samples has been collected in accordance with Human Tissue Act (2004). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time.

Availability of Data and Material

Data from this study are available through ALSPAC upon approval by the executive board.

Code Availability

Code is available in Supplementary Information and at https://www.medrxiv.org/content/10.1101/2021.08.24.21262573v1.

Authors Contributions

HSL conceptualized 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, and JG.

References

1. Partanen M, Kim DHC, Rauscher A, Siegel LS, Giaschi DE. White matter but not grey matter predicts change in reading skills after intervention. Dyslexia. 2021;27(2):224-244. doi:10.1002/dys.1668
2. Kirby JR, Parrila RK, Pfeiffer SL. Naming speed and phonological awareness as predictors of reading development. J Educ Psychol. 2003;95(3):453-464. doi:10.1037/0022-0663.95.3.453
3. Gough PB, Tunmer WE. Decoding, reading, and reading disability. Remedial Spec Educ. 1986;7(1):6-10. doi:10.1177/074193258600700104
4. Adlof SM, Catts HW, Little TD. Should the simple view of reading include a fluency component? Read Writ. 2006;19(9):933-958. doi:10.1007/s11145-006-9024-z
5. The Brainstorm Consortium, Anttila V, Bulik-Sullivan B, et al. Analysis of shared heritability in common disorders of the brain. Science. 2018;360(6395):eaap8757. doi:10.1126/science.aap8757
6. Andreola C, Mascheretti S, Belotti R, Ogliari A, Marino C, Battaglia M. The heritability of reading and reading-related neurocognitive components: A multi-level meta-analysis. Neurosci Biobehav Rev. 2021;121:175-200.
7. Little CW, Haughbrook R, Hart SA. Cross-study differences in the etiology of reading comprehension: A meta-analytical review of twin studies. Behav Genet. 2017;47(1):52-76. doi:10.1007/s10519-016-9810-6
8. Eising E, Mirza-Schreiber N, de Zeeuw EL, et al. Genome-wide analyses of individual differences in quantitatively assessed reading- and language-related skills in up to 34,000 people. Proc Natl Acad Sci. 2022;119(35):e2202764119. doi:10.1073/pnas.2202764119

9. 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 Res Hum Genet. 2020;23(1):23-32. doi:10.1017/thg.2020.7

10. Aaron PG, Joshi RM, Gooden R, Bentum KE. Diagnosis and treatment of reading disabilities based on the component model of reading: An alternative to the discrepancy model of LD. J Learn Disabil. 2008;41(1):67-84. doi:10.1177/0022219407310838

11. Price KM, Wigg KG, Feng Y, et al. Genome-wide association study of word reading: Overlap with risk genes for neurodevelopmental disorders. Genes Brain Behav. 2020;19(6):e12648. doi:10.1111/gbb.12648

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

13. Gialluisi A, Andlauer TFM, Mirza-Schreiber N, et al. Genome-wide association scan identifies new variants associated with a cognitive predictor of dyslexia. Transl Psychiatry. 2019;9(1):1-15. doi:10.1038/s41398-019-0402-0

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. Mol Psychiatry. 2021;26(7):3004-3017. doi:10.1038/s41380-020-00898-x

15. Luciano M, Gow AJ, Pattie A, Bates TC, Deary IJ. The influence of dyslexia candidate genes on reading skill in old age. Behav Genet. 2018;48(5):351-360. doi:10.1007/s10519-018-9913-3

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

17. Truong DT, Adams AK, Paniagua S, Frijters JC, Boada R, Hill DE. Multivariate genome-wide association study of rapid automatised naming and rapid alternating stimulus in Hispanic American and African-American youth. J Med Genet. 2019;56(8):557-566.

18. Eising E, Mirza-Schreiber N, de Zeeuw EL, et al. Genome-wide analyses of individual differences in quantitatively assessed reading- and language-related skills in up to 34,000 people. Proc Natl Acad Sci. 2022;119(35):e2202764119. doi:10.1073/pnas.2202764119

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

20. Trezzi V, Forni D, Giorda R, et al. The role of READ1 and KIAA0319 genetic variations in developmental dyslexia: testing main and interactive effects. J Hum Genet. 2017;62(11):949-955. doi:10.1038/jhg.2017.80

21. Zou H. The Adaptive Lasso and Its Oracle Properties. J Am Stat Assoc. 2006;101(476):1418-1429. doi:10.1198/016214506000000735

22. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol. 2008;70(5):849-911. doi:10.1111/j.1467-9868.2008.00674.x

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

24. McGrath LM, Pennington BF, Willcutt EG, Boada R, Shriberg LD, Smith SD. Gene × Environment interactions in speech sound disorder predict language and preliteracy outcomes. Dev Psychopathol. 2007;19(4):1047-1072. doi:10.1017/S0954579407000533

25. Mascheretti S, Andreola C, Scaini S, Sulpizio S. Beyond genes: A systematic review of environmental risk factors in specific reading disorder. Res Dev Disabil. 2018;82:147-152. doi:10.1016/j.ridd.2018.03.005

26. Saul M, Dinu V. Family Rank: a graphical domain knowledge informed feature ranking algorithm. Bioinformatics. 2021;37(20):3626-3631. doi:10.1093/bioinformatics/btab387

27. 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

28. Li C, Dinu V. miR2Pathway: A novel analytical method to discover MicroRNA-mediated dysregulated pathways involved in hepatocellular carcinoma. J Biomed Inform. 2018;81:31-40. doi:10.1016/j.jbi.2018.03.013

29. 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. Behav Genet. 2019;49(4):399-414. doi:10.1007/s10519-019-09957-8

30. Lancaster HS, Liu X, Dinu V, Li J. Identifying interactive biological pathways associated with reading disability. Brain Behav. 2020;10(8):e01735. doi:10.1002/brb3.1735

31. 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. Int J Epidemiol. 2013;42(1):111-127. doi:10.1093/ije/dys064

32. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol. 2013;42(1):97-110. doi:10.1093/ije/dys066

33. Rosseel Y. lavaan: An R Package for structural equation modeling. J Stat Softw. 2012;48:1-36. doi:10.18637/jss.v048.i02

34. Neale MD, McKAY MF, Childs GH. The Neale Analysis of Reading Ability — Revised. Br J Educ Psychol. 1986;56(3):346-356. doi:10.1111/j.2044-8279.1986.tb03047.x

35. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4(1):s13742-015-0047-0048. doi:10.1186/s13742-015-0047-8

36. 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. Nat Genet. 2018;50(8):1112-1121. doi:10.1038/s41588-018-0147-3

37. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291-295. doi:10.1038/ng.3211

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

39. Filges I, Röthlisberger B, Blattner A, et al. Deletion in Xp22.11: PTCHD1 is a candidate gene for X-linked intellectual disability with or without autism. Clin Genet. 2011;79(1):79-85. doi:10.1111/j.1399-0004.2010.01590.x

40. Halewa J, Marouillat S, Dixneuf M, et al. Novel missense mutations in PTCHD1 alter its plasma membrane subcellular localization and cause intellectual disability and autism spectrum disorder. Hum Mutat. 2021;42(7):848-861. doi:10.1002/humu.24208

41. Noor A, Whibley A, Marshall CR, et al. Disruption at the PTCHD1 Locus on Xp22.11 in Autism Spectrum Disorder and Intellectual Disability. Sci Transl Med. 2010;2(49):49ra68-49ra68. doi:10.1126/scitranslmed.3001267

42. Procopio F, Keser E, Knyspel J, Malanchini M, Rimfeld K, Plomin R. The genetic specificity of cognitive tests after controlling for general cognitive ability. Behav Genet. 2025;55(2):103-113. doi:10.1007/s10519-025-10213-5

43. Orban P, Lungu O, Doyon J. Motor sequence learning and developmental dyslexia. Ann N Y Acad Sci. 2008;1145(1):151-172. doi:10.1196/annals.1416.016

44. Caccia M, Lorusso ML. The processing of rhythmic structures in music and prosody by children with developmental dyslexia and developmental language disorder. Dev Sci. 2021;24(1):e12981. doi:10.1111/desc.12981

45. Valdois S, Roulin JL, Line Bosse M. Visual attention modulates reading acquisition. Vis Res Internet. 2019;165(February):152-161. doi:10.1016/j.visres.2019.10.011

46. Price KM, Wigg KG, Nigam A, et al. Identification of brain cell types underlying genetic association with word reading and correlated traits. Mol Psychiatry. 2023;28(4):1719-1730. doi:10.1038/s41380-023-01970-y

47. Carrión-Castillo A, Paz-Alonso PM, Carreiras M. Brain structure, phenotypic and genetic correlates of reading performance. Nat Hum Behav. 2023;7(7):1120-1134. doi:10.1038/s41562-023-01583-z

48. Thomas T, Khalaf, Shiva, Grigorenko EL. A systematic review and meta-analysis of imaging genetics studies of specific reading disorder. Cogn Neuropsychol. 2021;38(3):179-204. doi:10.1080/02643294.2021.1969900

49. Zhao H, Chen Y, Zhang B ping, Zuo P xiang. KIAA0319 gene polymorphisms are associated with developmental dyslexia in Chinese Uyghur children. J Hum Genet. 2016;61(8):745-752. doi:10.1038/jhg.2016.40

50. Müller B, Schaadt G, Boltze J, et al. 2C2 and 1C1 are putative modulators of dyslexia-related MMR. Brain Behav. 2017;7(11):e00851. doi:10.1002/brb3.851

51. Ludwig KU, Roeske D, Herms S, et al. Variation in GRIN2B contributes to weak performance in verbal short-term memory in children with dyslexia. Am J Med Genet B Neuropsychiatr Genet. 2010;153B(2):503-511. doi:10.1002/ajmg.b.31007

52. Martinelli A, Rice ML, Talcott JB, et al. A rare missense variant in the ATP2C2 gene is associated with language impairment and related measures. Hum Mol Genet. 2021;30(12):1160-1171. doi:10.1093/hmg/ddab111

53. Smith BL, Mcgregor KK, Demille D. Phonological development in lexically precocious 2-year-olds. Appl Psycholinguist. 2006;27(3):355-375. doi:10.1017/S0142716406060310

54. Matsumoto A, Mizuno M, Hamada N, et al. LIN7A depletion disrupts cerebral cortex development, contributing to intellectual disability in 12q21-deletion syndrome. Bardoni B, ed. PLoS ONE. 2014;9(3):e92695. doi:10.1371/journal.pone.0092695

55. Benítez-Burraco A, Jiménez-Romero MS, Fernández-Urquiza M. Delving into the genetic causes of language impairment in a case of partial deletion of NRXN1. Mol Syndromol. 2023;13(6):496-510. doi:10.1159/000524710

56. Abdellaoui A, Verweij KJH. Dissecting polygenic signals from genome-wide association studies on human behaviour. Nat Hum Behav. 2021;5(6):686-694. doi:10.1038/s41562-021-01110-y

Figures

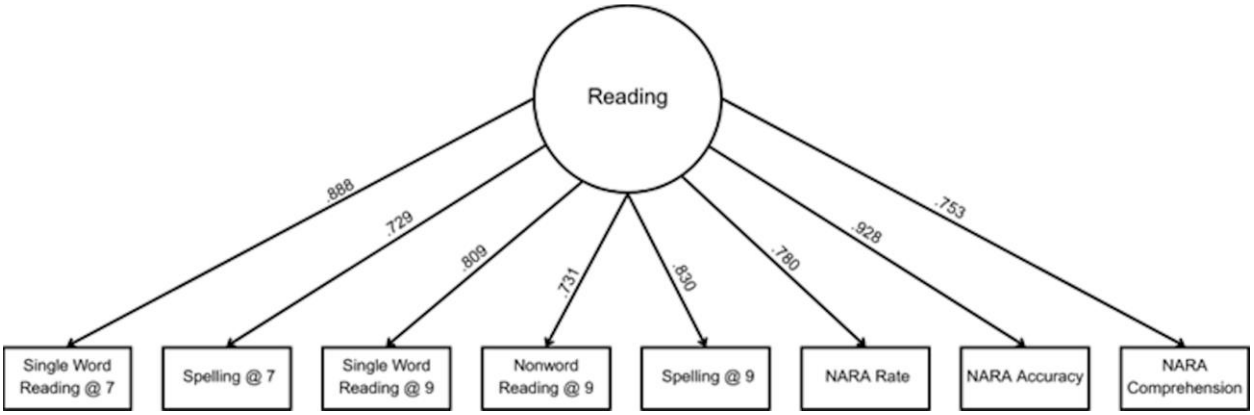

Figure 1. Manifest path values for Latent Reading Ability Base Model for ALSPAC. All paths were significant. NARA = Neale Analysis of Reading Ability (Neale, 1997).

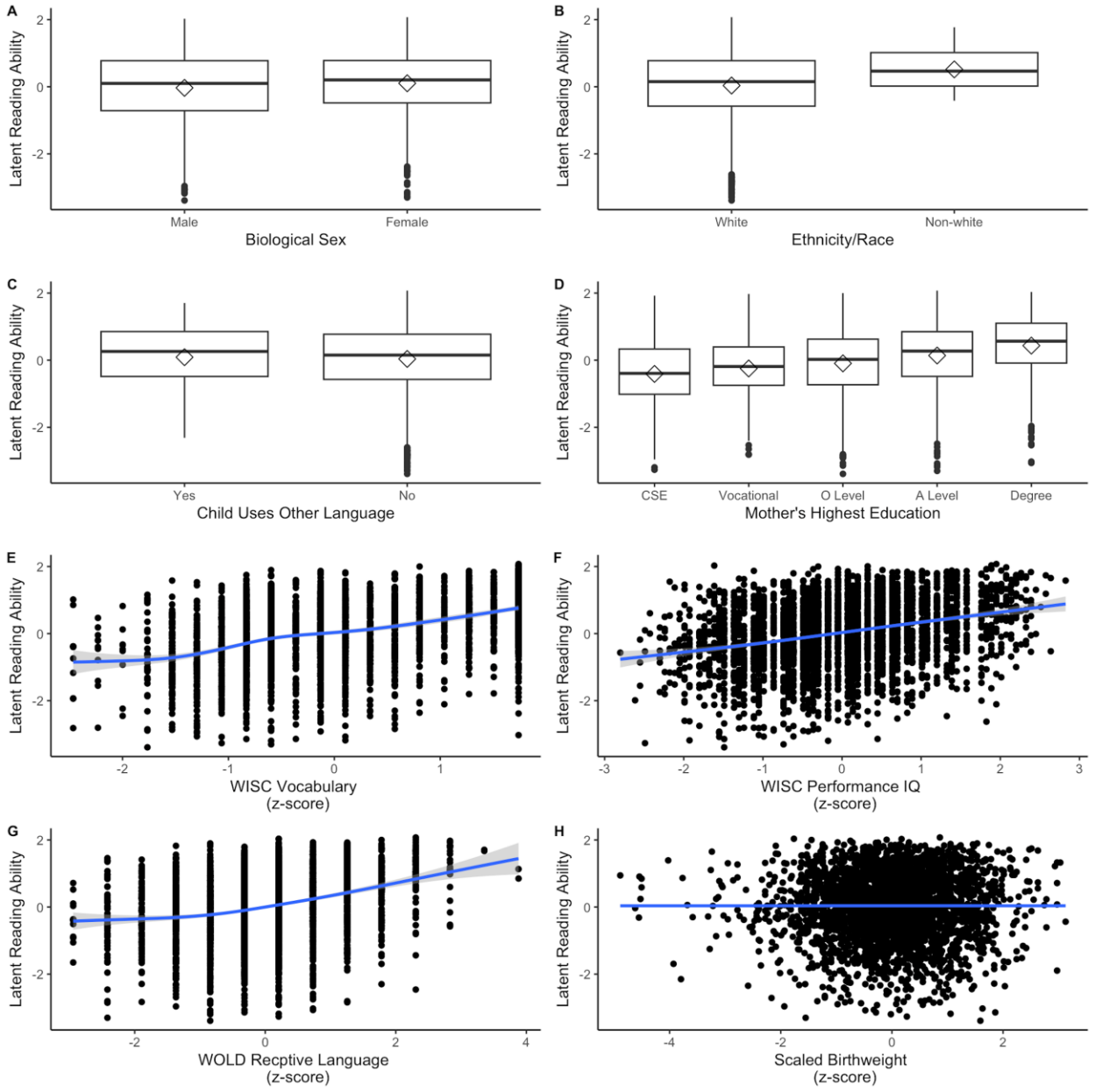

Figure 2. Plots of relationships between demographic variables and latent reading ability score. Plots for the relationships between (A) biological sex, (B) ethnicity / race, (C) use of language other than English, (D) highest level of maternal education, (E) receptive vocabulary, (F) nonverbal intelligence, (G) receptive language, and (H) scaled preferred birthweight in grams and latent reading ability in the ALSPAC cohort.

Table 1

Elastic net results for general reading ability in the ALSPAC - Read SNP only models

SNP	Chr	Gene(s)	Coefficient	Variant type(s)
rs12086013	1	KCNJ10	0.038	intron
rs12086013.1	1	KCNJ10	0.037	
rs12118079	1	.	-0.046	
rs12118079.1	1	.	-0.032	
rs1984411	1	.	-0.025	
rs1984411.1	1	.	-0.029	
rs2904122	1	EPHB2	-0.035	intron
rs2904122.1	1	EPHB2	-0.032	
rs11901326	2	COL6A3	0.099	intron, non-coding transcript
rs13408091	2	SSB	0.035	NMD, intron, non-coding transcript
rs617829	2	HDAC4	0.065	intron, non-coding transcript
rs6544443	2	LOC105374506	-0.046	intron
rs6758123	2	SSB	0.069	intron, non-coding transcript
rs7586252	2	.	-0.057	
rs1486369	3	ZNF385D	-0.090	NMD, intron
rs200564558	3	ZNF385D	-0.281	NMD, frameshift, intron
rs2358932	3	.	-0.093	
rs7624403	3	ATXN7	0.022	intron, non-coding transcript
rs832197	3	ATXN7	-0.054	intron, non-coding transcript
rs41508146	4	LINC02432	0.039	intron, non-coding transcript exon, non-coding transcript
rs55703414	4	RAPGEF2	-0.042	intron, non-coding transcript
rs6536395	4	RAPGEF2	-0.041	intron, non-coding transcript
rs1035418	5	TENM2	0.077	intron
rs11750607	5	.	-0.030	
rs12404907	5	.	-0.033	
rs1508818	5	.	-0.030	
rs154205	5	.	0.034	
rs35244816	5	.	-0.022	

rs416750	5	CHD1-DT	0.026	intron
rs709390	5	CHD1-DT	0.032	intron, non-coding transcript
rs1427079	6	.	0.055	
rs16889556	6	KIAA0319	0.027	intron
rs6935076	6	KIAA0319	0.021	intron
rs10231255	7	.	-0.056	
rs11767365	7	RAMP3	-0.101	3' UTR
rs13307587	7	ENSG00000229618	0.072	intron, non-coding transcript
rs57301765	7	.	0.056	
rs67554174	7	ENSG00000234710	0.041	intron, non-coding transcript
rs7798197	7	HDAC9	0.038	3' UTR
rs4292642	8	.	0.091	
rs5025174	8	CSMD1	0.068	intron
rs10814272	9	TLN1	0.049	intron
rs2295794	9	TLN1	0.024	intron
rs10795024	10	LOC105376360, LOC124902565	-0.068	stop lost, terminator codon, 2KB upstream, upstream transcript
rs1163203	10	.	-0.030	
rs1079727	11	DRD2	-0.021	intron, non-coding transcript
rs7130517	11	NFRKB	-0.071	NMD, intron
rs1054442	12	DDN, ENSG00000258283	0.067	3' UTR, intron, non-coding transcript
rs2268119	12	GRIN2B	-0.033	NMD, intron
rs67309388	12	.	0.052	
rs73144951	12	LIN7A	-0.121	NMD, intron
rs61997099	14	.	0.095	
rs7154388	14	LOC124903373	-0.094	500B downstream, downstream transcript
rs1075938	15	DNAAF4	-0.075	5' UTR
rs1902586	15	CYP19A1, MIR4713HG	0.046	intron, non-coding transcript
rs57809907	15	DNAAF4, DNAAF4-CCPG1	-0.035	3' UTR, NMD, intron, non-coding transcript, stop gained
rs77641439	15	DNAAF4, DNAAF4-CCPG1	0.057	3' UTR, NMD, intron, missense, non-coding transcript

rs6502712	17	OR3A2	-0.026	intron
rs7220982	17	ASPA, SPATA22	0.020	intron
rs1468647	18	LINC03069, LINC01905, LINC01539	0.090	intron, non-coding transcript
rs17689314	21	.	-0.097	

n = 3647; Chr = chromosome NMD = nonsense mediated decay transcript variant

Table 2

Elastic net results for informative SNPs for general reading ability in the ALSPAC - Read SNP + demographics model

SNP	Chr	Gene(s)	Coefficient	Variant type(s)
rs12118079	1	.	-0.044	
rs12118079.1	1	.	-0.039	
rs12404907	1	.	-0.032	
rs1984411	1	.	-0.025	
rs1984411.1	1	.	-0.023	
rs2904122	1	EPHB2	-0.030	intron
rs2904122.1	1	EPHB2	-0.027	
rs12086013	1	KCNJ10	0.023	intron
rs12086013.1	1	KCNJ10	0.021	
rs7586252	2	.	-0.058	
rs11901326	2	COL6A3	0.084	intron, non-coding transcript
rs617829	2	HDAC4	0.047	intron, non-coding transcript
rs6544443	2	LOC105374506	-0.106	intron
rs7577992	2	LOC105374506	0.038	intron
rs13408091	2	SSB	0.065	NMD, intron, non-coding transcript
rs6758123	2	SSB	0.064	intron, non-coding transcript
rs2358932	3	.	-0.056	
rs7624403	3	ATXN7	0.076	intron, non-coding transcript
rs1486369	3	ZNF385D	-0.047	NMD, intron
rs200564558	3	ZNF385D	-0.808	NMD, frameshift, intron
rs2630806	3	ZNF385D	-0.028	NMD, intron, missense
rs1369973	4	LINC02432	-0.045	intron, non-coding transcript
rs41508146	4	LINC02432	0.049	intron, non-coding transcript exon, non-coding transcript
rs6537043	4	LINC02432	-0.030	intron, non-coding transcript
rs7685977	4	LINC02432	0.090	intron, non-coding transcript
rs13123795	4	RAPGEF2	0.024	intron, non-coding transcript
rs36018887	4	RAPGEF2	0.045	intron, non-coding transcript
rs55703414	4	RAPGEF2	-0.080	intron, non-coding transcript
rs6536395	4	RAPGEF2	-0.094	intron, non-coding transcript

rs11750607	5	.	-0.048	
rs1508818	5	.	-0.053	
rs154205	5	.	0.036	
rs35244816	5	.	0.041	
rs10040066	5	CHD1-DT	-0.028	intron
rs112393710	5	CHD1-DT	-0.021	intron, non-coding transcript
rs113868271	5	CHD1-DT	-0.030	intron, non-coding transcript
rs1472622	5	CHD1-DT	0.022	intron
rs416750	5	CHD1-DT	0.055	intron
rs709390	5	CHD1-DT	0.069	intron, non-coding transcript
rs447044	5	CHD1-DT	-0.037	intron
rs1035418	5	TENM2	0.071	intron
rs1427079	6	.	0.047	
rs2274305	6	DCDC2	-0.045	missense
rs3765502	6	DCDC2	0.042	intron
rs7765678	6	DCDC2	-0.053	intron
rs807724	6	DCDC2	0.037	intron
rs16889506	6	KIAA0319	-0.019	intron
rs6935076	6	KIAA0319	0.047	intron
rs10231255	7	.	-0.063	
rs57301765	7	.	0.053	
rs13307587	7	ENSG00000229618	0.074	intron, non-coding transcript
rs67554174	7	ENSG00000234710	0.058	intron, non-coding transcript
rs7782412	7	FOXP2	-0.038	NMD, intron, non-coding transcript
rs936146	7	FOXP2	-0.052	NMD, intron
rs7798197	7	HDAC9	0.038	3' UTR
rs11767365	7	RAMP3	-0.106	3' UTR
rs7837231	8	.	0.041	
rs5025174	8	CSMD1	0.058	intron
rs4292642	8		0.043	
rs10814272	9	TLN1	-0.025	intron
rs2295794	9	TLN1	0.062	intron
rs1163203	10	.	-0.025	

rs10795024	10	LOC105376360, LOC124902565	-0.055	stop lost, terminator codon, 2KB upstream, upstream transcript
rs2141786	11	TMEM45B	0.025	intron, non-coding transcript exon, non-coding transcript
rs55900082	11	TMEM45B	-0.111	intron, non-coding transcript
rs1054442	12	DDN, ENSG00000258283	0.081	3' UTR, intron, non-coding transcript
rs2192973	12	GRIN2B	-0.101	NMD, intron
rs2216128	12	GRIN2B	0.098	NMD, intron
rs2268119	12	GRIN2B	-0.031	NMD, intron
rs5796555	12	GRIN2B	0.031	NMD, intron
rs73144951	12	LIN7A	-0.108	NMD, intron
rs61997099	14	.	0.093	
rs7154388	14	LOC124903373	-0.105	500B downstream, downstream transcript
rs7158782	14	LOC124903373	0.029	intron
rs10046	15	CYP19A1, MIR4713HG	-0.186	3' UTR, intron, non-coding transcript
rs1902586	15	CYP19A1, MIR4713HG	0.073	intron, non-coding transcript
rs8034835	15	CYP19A1, MIR4713HG	0.086	intron, non-coding transcript
rs8034835.1	15	CYP19A1, MIR4713HG	0.120	
rs57809907	15	DNAAF4, DNAAF4-CCPG1	-0.088	3' UTR, NMD, intron, non-coding transcript
rs77641439	15	DNAAF4, DNAAF4-CCPG1	0.029	3' UTR, NMD, intron, missense, non-coding transcript
rs11860694	16	ATP2C2	-0.031	intron, non-coding transcript
rs12150649	17	.	0.040	
rs1084906	17	OR3A2	0.037	intron
rs230467	17	OR3A2	0.044	intron
rs230472	17	OR3A2	0.027	intron
rs231684	17	OR3A2	-0.106	intron
rs231688	17	OR3A2	-0.083	intron
rs2676612	17	OR3A2	0.054	intron

rs368810	17	OR3A2	-0.056	intron
rs370361	17	OR3A2	-0.023	intron
rs804230	17	OR3A2	0.051	intron
rs804231	17	OR3A2	0.028	intron
rs11873029	18	DYM	-0.020	intron
rs1468647	18	LINC03069, LINC01905, LINC01539	0.087	intron, non-coding transcript
rs1299348	18	MC5R	-0.032	2KB upstream, upstream transcript
rs8094327	18	NEDD4L	0.031	NMD, intron, non-coding transcript
rs17689314	21	.	-0.081	

n = 3254; Chr = chromosome; NMD = nonsense mediated decay transcript variant

Table 3

Enrichment analysis results for ALSPAC - Read by elastic net model

SNP - Only Model		
Term name	**Term ID**	**Adjusted p-value**
postsynaptic membrane	GO:0045211	0.0001
regulation of neuronal synaptic plasticity	GO:0048168	0.0007
amyloid-beta binding	GO:0001540	0.0410
peptidyl-lysine deacetylation	GO:0034983	0.0422
G protein-coupled receptor complex	GO:0097648	0.0131
multicellular organismal process	GO:0032501	0.0149
learning or memory	GO:0007611	0.0226
synaptic membrane	GO:0097060	0.0009
plasma membrane region	GO:0098590	0.0012
regulation of long-term neuronal synaptic plasticity	GO:0048169	0.0032
synapse	GO:0045202	0.0038
regulation of synaptic plasticity	GO:0048167	0.0064
dendritic spine	GO:0043197	0.0077
neuron spine	GO:0044309	0.0084
Rap1 signaling pathway	KEGG:04015	0.0255
multicellular organism development	GO:0007275	0.0396
postsynapse	GO:0098794	0.0138
Alcoholism	KEGG:05034	0.0164
cell junction	GO:0030054	0.0168
system process	GO:0003008	0.0236
anatomical structure development	GO:0048856	0.0285
system development	GO:0048731	0.0349
cognition	GO:0050890	0.0455
SNP + Demographics Model		
Term name	**Term ID**	**Adjusted p-value**
plasma membrane region	GO:0098590	0.0025
system process	GO:0003008	0.0088
regulation of plasma membrane bounded cell projection organization	GO:0120035	0.0175
regulation of dendrite development	GO:0050773	0.0103
animal organ development	GO:0048513	0.0319
peptidyl-lysine deacetylation	GO:0034983	0.0497
postsynaptic membrane	GO:0045211	0.0040
synaptic membrane	GO:0097060	0.0226
dendrite development	GO:0016358	0.0156
regulation of cell projection organization	GO:0031344	0.0205
system development	GO:0048731	0.0212
anatomical structure development	GO:0048856	0.0264
nervous system development	GO:0007399	0.0288
dendrite morphogenesis	GO:0048813	0.0429