

Multivariate genome-wide association analysis of quantitative reading skill and dyslexia improves gene discovery

Running title: Multivariate genome-wide association study of reading skill

Hayley S. Mountford PhD ¹, Else Eising PhD ², Pierre Fontanillas PhD ³, Adam Auton PhD ³, 23andMe Research Team, Evan K. Irving-Pease PhD ^{4,5}, Catherine Doust PhD ¹, Timothy C. Bates PhD ¹, Nicholas G. Martin PhD ⁶, Simon E. Fisher PhD ^{2,7} and Michelle Luciano PhD ¹

¹ School of Philosophy, Psychology and Language Sciences, University of Edinburgh, UK

² Language and Genetics Department, Max Planck Institute for Psycholinguistics, Netherlands

³ 23andMe, Inc., Sunnyvale, California, USA

⁴ Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, Denmark

⁵ Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Denmark

⁶ Genetic Epidemiology Laboratory, QIMR Berghofer Medical Research Institute, Australia

⁷ Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

Corresponding author: Hayley S. Mountford, email: hayley.mountford@ed.ac.uk and Michelle Luciano, email: michelle.luciano@ed.ac.uk, 7 George Square, Edinburgh, EH8 9JZ, UK, telephone: +44 (0)131 650 3063

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Abstract

The ability to read is an important life skill and a major route to education. Individual differences in reading ability are influenced by genetic variation, with a heritability of 0.66 for word reading, estimated by twin studies. Until recently, genomic investigations were limited by modest sample size. Here we use a multivariate genome-wide association study (GWAS) method, MTAG, to leverage summary statistics from two independent GWAS efforts, boosting power for analyses of reading ability; GenLang meta-analysis of word reading ($N = 27\,180$) and the 23andMe, Inc., study of dyslexia ($N_{\text{cases}} = 51\,800$, $N_{\text{controls}} = 1\,087\,070$). We increase effective sample size to $N = 102\,082$, representing the largest genetic study of reading ability, to date. We identified 35 independent genome-wide significant loci, including 7 regions not previously reported. Single-nucleotide polymorphism (SNP) based heritability was estimated at 24%. We observed clear positive genetic correlations with cognitive and educational measures. Gene-set analyses implicated neuronal synapses and proneural glioblastoma pathways, further supported by enrichment of neuronally expressed genes in the developing embryonic brain. Polygenic scores of our multivariate results predicted between 2.29-3.50% of variance in reading ability in an independent sample, the National Child Development Study cohort ($N = 6\,410$). Polygenic adaptation was examined using a large panel of ancient genomes spanning the last ~15k years. We did not find evidence of selection, suggesting that reading ability may not have been subject to recent selection pressure in Europeans. By combining existing datasets to improve statistical power, these results provide novel insights into the biology of reading.

Introduction

Reading is a key academic skill and an important component of education. Difficulties with reading are associated with poorer life outcomes, lower socioeconomic status, and can greatly impact quality of life¹. Unravelling the biological basis of this complex trait is essential in understanding the aetiology of reading, and why some people struggle with reading throughout their lives. A recent twin-based meta-analysis found that reading ability is highly heritable at 0.66^{2,3}, although prior estimates from classical twin models varied widely (0.03-0.83). Allelic variation in a number of genes have been associated with reading-related traits and dyslexia⁴, although with mixed support from replication studies. The largest genome-wide association study (GWAS) of quantitative reading skill⁵ (meta-analysis of 33 959 individuals from 19 cohorts) by the GenLang Consortium identified a single locus associated with word reading (rs11208009, $P = 1.10 \times 10^{-8}$) containing three candidate genes (*DOCK7*, *ANGPTL3* and *USP1*). Here, we improve statistical power using a much larger GWAS of dyslexia⁴ in a multivariate GWAS to improve gene discovery in investigations of reading ability.

Dyslexia, a common neurodevelopmental learning difference occurring in 5-10% of school age children⁶, is characterised by difficulties with accurate and/or fluent word reading, and poor spelling. Some diagnostic definitions of dyslexia extend to reduced performance on measures of verbal memory and processing speed, and/or emphasise a discrepancy between reading and other cognitive abilities⁷. Dyslexia tends to cluster within families⁸ and shows high heritability in twin-studies (0.4-0.6)⁹. Doust et al (2022)⁴ performed the largest GWAS of this trait to-date, using 23andMe, Inc. self-reported dyslexia diagnosis in 51 800 cases and 1 087 070 controls. Forty-two significantly associated regions were identified, including 27 not previously reported in studies of educational attainment or cognitive traits. Noteworthy for the current study, strong genetic correlations have been observed between dyslexia and quantitative measures of reading and spelling, ranging from -0.7 to -0.75 with CI 95% spanning -0.6 to -0.86⁵. These strong genetic correlations are consistent with the view that dyslexia is, to a large degree, representative of the low extreme of normal varying reading ability^{10,11} rather than being a qualitatively distinct phenotype.

GWAS for dyslexia⁴ and quantitative reading skill⁵ clearly demonstrate the importance of large sample size to detect genetic effects that underlie reading traits. Eising et al's (2022)⁵ GWAS meta-analysis of varied reading and spelling traits in up to 34 000 individuals made many important discoveries including revealing genetic correlations with other traits, significant SNP heritability (e.g., estimated at $h^2_{\text{snp}} = 0.19$ (SE = 0.02) for word reading), and clarifying the genetic structure among measures of reading, spelling, language, general cognitive ability, and educational attainment. However, it was relatively underpowered to detect single SNP variant associations that exceeded genome-wide thresholds. It is clear from studies of the genetics of reading^{5, 12}, and also from other neurodevelopmental traits such as autism spectrum disorder (ASD)¹³ and attention deficit hyperactivity disorder (ADHD)¹⁴ with case sample sizes of 18 381 and 38 691, respectively, that analysing large sample sizes is key to improving resolution of associated variants. For neurodevelopmental measures of literacy collected in childhood, it has historically been challenging to gather sufficient sample size, primarily due to the fact that reading skill is not perceived as medically relevant. Meta-analysis approaches are appropriate for addressing this issue.

One of the alternative methods to collecting and phenotyping new cohorts for a trait of interest, multi-trait analysis of GWAS approach (MTAG)¹⁵ which uses the shared genetic architecture of related phenotypes (multivariate) to increase gene discovery power. For example, Grove and colleagues¹³ used MTAG to increase their ASD GWAS power by adding GWAS for schizophrenia, educational attainment, and major depression. This showed stronger evidence for previously reported regions, and seven novel regions shared with educational attainment or depression. More recently, multivariate analyses were used across five psychiatric traits (ASD, ADHD, bipolar disorder, schizophrenia, and depression)¹⁶. Again, this increased the number of associated loci identified for each individual trait, particularly bipolar disorder, which increased from 8 genome-wide significant loci to 54. In addition to statistical power gain, this multivariate associated approach permits to delineate the shared and separate genetic profiles of related conditions.

Given the strong genetic correlation between dyslexia and word-reading skills⁴, we applied the multivariate method to boost sample size of the word reading GWAS and identify novel associated

loci. We note that minimal sample overlap is expected, given that the GWAS of dyslexia includes a single sample of adult predominantly US-based 23andMe, Inc. research participants and the GWAS of word reading is a meta-analysis of cohorts of children and young adults with a minority from the US. Moreover, MTAG uses bivariate LD score regression to control for sample overlap between cohorts.

By increasing effective sample size through multivariate GWAS, we aimed to achieve the following. Firstly, improved power to detect novel loci associated with reading, uncovering biological pathways at the gene and functional level, and linking to biological mechanisms that underlie reading ability through the availability of developmental brain gene-expression datasets¹⁷. Secondly, obtaining sufficient power to generate reliable polygenic scores to predict reading variation in independent cohorts. Doust et al⁴ generated a polygenic score from the dyslexia GWAS which significantly predicted variance in reading performance measures in independent samples, for example, accounting for 3.6% in nonword reading in the Brisbane Adolescent Cohort and 5.6% of variance in word recognition in a US cohort enriched for individuals with reading difficulties (Colorado Learning Disabilities Research Centre)⁴. Finally, we used a novel method to determine if the polygenic scores for word reading showed evidence of selection in the past 15 000 years^{18, 19}.

Materials / Subjects and Methods

Ethics approval statement

The study made use of existing data sets and was granted ethical approval by the University of Edinburgh School of Philosophy, Psychology and Language Sciences research ethics committee (PPLSREC 29-1819/8).

Multivariate GWAS

GWAS summary statistics for quantitative measures of word reading were available in the GenLang meta-analysis of 27 180 (male = 13 874, female = 13 202, no information = 104) participants of European ancestry only across 18 studies⁵. Participants were children or young adults ranging from 5-26 years⁵. Measures depended on the contributing cohort and are detailed in Eising *et al*⁵, with the Test of Word Reading Efficiency (TOWRE) and Wide Range Achievement Test (WRAT) most common. Summary statistics without genomic-control correction were used.

Summary statistics for self-reported dyslexia diagnosis (23andMe, Inc.) were reported in Doust *et al* (2022)⁴. Participants responded “Yes” to “have you been diagnosed with dyslexia” ($N_{\text{cases}} = 51\,800$, female = 30 287, male = 21 513), and $N_{\text{controls}} = 1\,087\,070$ responded “No” (female = 641 016, male = 446 054). All participants were over 18 years of age (mean 49.6 (cases) and 51.7 years (controls)). Participants with non-European ancestry were excluded prior to analysis. Summary statistics without genomic-control correction were used⁴.

Both sets of summary statistics were annotated with rsIDs (build hg19), formatted for MTAG, and Z scores were calculated in R. Variants with imputation quality <0.8 or minor allele frequency of <0.01 were excluded prior to analysis. Multivariate GWAS was performed with MTAG¹⁵ using default settings and false discovery rate (FDR) calculation. Associations were visualised using ggplot2²⁰. Individual regions were visualised using LocusZoom (<http://locuszoom.org>). FUMA version v1.5.0 was used to annotate the associated regions that met the genome-wide significance threshold²¹ ($P \leq 5 \times 10^{-8}$, $R^2 < 0.6$, and <250kb maximum distance between LD blocks to merge into one genomic locus).

Heritability and Genetic Correlation

SNP-based heritability (h^2_{snp}) was estimated using LDSC v1.0.1²². European LD reference panel was obtained from <https://alkesgroup.broadinstitute.org/LDSCORE>. GWAS summary statistics were obtained from the Complex-Traits Virtual Genetics Lab (CTG-VL) platform (<https://vl.genoma.io>), except for ASD¹³ and ADHD¹⁴ which were downloaded from the Psychiatric Genomics Consortium repository (<https://www.med.unc.edu/pgc/download-results>). Genetic correlations were performed using LD-Score v1.0.1 within the CTG-VL platform and considered significant at a Bonferroni corrected threshold of $P \leq 3.478 \times 10^{-5}$ from 1438 tests.

Due to the difference in sample size between the Dyslexia and GenLang univariate summary statistics, we cross-checked genetic correlations between each set of summary statistics and the MTAG output to check that the contributions were approximately equivalent, assessed by false discovery rate (FDR).

Gene-based and Gene-set Analysis

Gene-based associations were calculated using MAGMA v1.08²³ using SNP2GENE within the FUMA interface (<https://fuma.ctglab.nl/>) for 18 842 genes and were considered significant at a Bonferroni corrected threshold of $P \leq 2.65 \times 10^{-6}$. Gene-set analyses of biological pathways defined by gene ontology (GO) pathways and curated gene-sets were examined using MAGMA. GO terms and gene-sets containing fewer than 20 genes were excluded from analysis, so a total of 9 113 biological pathways were tested and a Bonferroni corrected threshold of $P \leq 5.49 \times 10^{-6}$ was applied.

Functional Mapping, Annotation and Partitioned Heritability

Fine mapping and annotations were performed using the Variant Effect Predictor (VEP) online tool (<http://grch37.ensembl.org/>) on the list of candidate SNPs present with $R^2 \leq 0.6$ with an independent

significant SNP generated by FUMA, and includes tagged SNPs extracted from the 1000 genomes reference panel. Variants were considered potentially damaging if they were annotated as probably damaging by PolyPhen2 and deleterious by SIFT. Expression QTL analysis was performed using PsychENCODE eQTLs, BRAINEAC, eQTLcatalogue BrainSeq, and GTEx v8 Brain databases within FUMA.

MAGMA, within FUMA, was used to test for enrichment of tissue-specific annotations. For this, we used bulk RNA-seq expression profiles from 53 tissue types from GTEx v8, and BrainSpan RNA-sequencing from 29 ages spanning 11 developmental stages.

To interrogate cell- and region-specific resolution, we accessed single-cell RNA-seq (scRNA) data via the Cell Type function within FUMA. Expression data from human embryonic ventral mid-brain (6-11 weeks post gestation) (GSE76381), human embryonic prefrontal cortex (8-26 weeks post gestation) (GSE104276), and human adult and fetal cortex (GSE67835) datasets were tested.

We partitioned SNP heritability using stratified LDSC, as described by Finucane *et al*²⁴, to determine if significantly more SNPs clustered within tissue-specific chromatin modification patterns than expected by chance, based on the proportion of SNPs that map within these types of genomic regions. Annotations were based on data from the Roadmap Epigenomics project and Enhancing GTEx project (ENTEx). LD scores, regression weights and European allele frequencies were obtained from <https://alkesgroup.broadinstitute.org/LDSCORE>.

Polygenic Score Analysis

Word-reading polygenic scores were calculated for SNPs based on increasing P value thresholds spanning the full range. Scores were calculated for the National Child Development Study (NCDS); a large UK birth cohort study born in 1958²⁵, with extensive genetic and longitudinal reading measures as described in Bridges *et al*²⁶. Individuals genotyped on one of the two lowest resolution arrays (Illumina 15k Custom chip and Affymetrix 500k) were excluded prior to analysis due to lack of available data from chromosome X, resulting in a cohort of N = 6 410 with imputed genotypes. Six

measures of functional reading were used, consisting of composite measures at ages 7, 11, 16 and across all time points, and binary measures of difficulties at ages 23 and 33²⁶. Polygenic scores were estimated using PRSice2 v2.3.5²⁷, and sex, array and 10 principal components were used as covariates, as previously described²⁶.

Polygenic Selection Analysis

We sought to identify evidence of polygenic selection for reading ability using a large panel of 1015 imputed ancient genomes, sampled from across West Eurasia^{19, 28}. This dataset represents a dense transect of ancient individuals sampled over the last 15k years with local ancestry contributing to present day Europeans. We ascertained statistically independent SNPs associated with reading ability by filtering our genome-wide summary statistics to only retain positions imputed with high confidence in the ancient dataset. LD-clumping was performed using Plink 1.90b4 with a window size of 250kb, maximum R^2 threshold of 0.05, and maximum $P \leq 5 \times 10^{-8}$ using the 1000 Genomes Project phase 3 European populations (GBR, FIN, TSI) as the reference panel. We then inferred allele frequency trajectories and selection coefficients using CLUES²⁹ and exported the posterior likelihood densities. Finally, we modelled the polygenic selection gradient for reading ability with PALM³⁰ using imputed ancestral data generated and method described by Barrie *et al*¹⁸.

Results

Multivariate GWAS of Reading Ability

The genetic correlation between the univariate summary statistics of word reading and dyslexia (Europeans only, without GC correction) was -0.71 (SE = 0.05, $Z = -15.06$), and indicative of a high degree of shared genetic aetiology enabling multivariate GWAS analysis with MTAG¹⁵. The analysis produced an equivalent sample size of 102 082 individuals for reading ability and used 5 449 985 DNA variants that were shared between the two sets of univariate summary statistics. This provided 87% power to detect additive trait variance of up to 0.04% ($N = 102\,082$, $\alpha = 5 \times 10^{-8}$). We identified 35 genome-wide significant ($P \leq 5 \times 10^{-8}$) independent loci ($r^2 < 0.6$, and <250kb maximum distance between LD blocks to merge into one genomic locus) (figure 1a), containing 86 independent significant SNPs, independent from each other at an R^2 of 0.1. Twenty-six of these regions were previously reported in the dyslexia GWAS⁴, 6 regions can be considered novel but were present in the uncorrected dyslexia summary statistics, and 1 region is novel (figure 1, table 1). The region previously reported as significant in the GenLang word reading meta-analysis⁵ was not associated in the present study. Quantile-Quantile (Q-Q) plots indicated (supplementary figure 1) that the data were appropriately controlled for population stratification, as markers showing low association with word reading did not deviate from the expected quantile. Similarly, LDSC lambda genomic control statistics (1.28), intercept (0.84, SE = 0.01) and ratio (<0) support an absence of genomic inflation. The maximum FDR for MTAG was moderate at 0.06, indicating that 6% of associated regions may be false positives. Constituent univariate GWAS χ^2 statistics were calculated by MTAG for dyslexia ($\chi^2 = 1.66$) and GenLang word reading ($\chi^2 = 1.08$), and multivariate GWAS ($\chi^2 = 1.35$). Individual LocusZoom plots for each region are presented in supplementary figure 2:1-35. Summary statistics for SNPs reaching suggestive significance ($P \leq 1 \times 10^{-5}$) are presented in supplementary table 1.

[Figure 1 and Table 1 about here]

The most significantly associated loci were consistent with regions previously reported in the univariate dyslexia GWAS⁴. The top locus, chr3q22.2 (rs13082684, $P = 1.87 \times 10^{-17}$) containing

PPP2R3A (table 1, supplementary figure 2-14) is consistent with the top SNP of the dyslexia GWAS, although the association has become more highly significant in the MTAG analysis. Both the second and third top loci, chr20q13.13 (rs348258) (supplementary figure 2-34) and chr17q12 (rs1075218) (supplementary figure 2-30), also show more significant associations than the dyslexia source GWAS. However, for both loci, the top SNPs have changed and fall within different genes.

The lead SNP identified by the GenLang word-reading meta-GWAS⁵, rs11208009, did not reach genome-wide significance in the present multivariate study ($P = 2.71 \times 10^{-6}$). However, it fell within a region that reached suggestive significance (chr1:62900811-63199936) at $P = 1.96 \times 10^{-7}$ in which rs1168114 ($LD = 0.636$) was now the lead SNP. This suggestive locus overlaps completely with the original study, therefore including candidate genes *DOCK7*, *ANGPTL3*, and *USP1* (supplementary figure 3).

Novel Regions Associated with Reading Ability

Our multivariate analysis detected seven novel regions that did not previously reach genome-wide significance in either univariate GWAS for word reading or dyslexia (table 2). One of the lead SNPs in these regions, rs362307, has been previously linked to a range of phenotypes, the most relevant of which include educational attainment³¹, cognitive ability³², and a “worry” phenotype key to neuroticism^{33, 34}. The remaining six regions all contained lead SNPs detected by the Doust *et al* dyslexia univariate GWAS⁴ and present in the GWAS Catalogue, but at suggestive significance threshold of $P \leq 1 \times 10^{-5}$. Interestingly, SNP rs2158266 in region chr16:72333576-72697419 containing gene *AC004158.2*, was previously associated with phoneme awareness, considered a cognitive marker of dyslexia³⁵, although we note that the discovery cohort from that prior study was included in the GenLang word reading GWAS, and therefore contributed also to the current investigation.

[Table 2 about here]

Multivariate GWAS of Dyslexia

We also examined the dyslexia output of MTAG which produced an effective sample size of 1 228 832 individuals and identified 80 independent loci that met the genome-wide significance threshold (figure 1b). Of these 80 associated loci, 41 were genome-wide significant loci in the original univariate dyslexia GWAS, and 19 were reported as suggestively significant⁴. The 7 novel loci presented in our reading ability multivariate GWAS, including lead SNP rs362307 in *HTT*, were also significantly associated in our dyslexia multivariate analysis. Finally, this analysis identified 13 novel loci, not previously associated with dyslexia. These novel regions are presented in supplementary table 2. Summary statistics for SNPs reaching suggestive significance ($P \leq 1 \times 10^{-5}$) are presented in supplementary table 3. Subsequent genetic and biological analyses are focussed on the reading-ability multivariate summary statistics due to the substantial increase in power for follow-up analyses, as compared to the original univariate GWAS.

Heritability and Genetic Correlations of Reading Ability

LDSC analysis of the reading-ability multivariate GWAS revealed a high SNP-based heritability estimate (h^2_{snp}) of 0.24 (SE = 0.01), an increase on the previous estimate of 0.19 (SE = 0.02) previously reported for word reading⁵. Modest evidence of population stratification was indicated by lambda genomic-control estimate of 1.28, while mean χ^2 and genomic correction having been applied by MTAG as reflected by the intercept ≤ 1 (0.84) and ratio < 0 .

Genetic correlations were estimated between multivariate reading ability and 1438 traits including recently published summary statistics for ASD¹³ and ADHD¹⁴. Statistically significant genetic correlations (Bonferroni corrected threshold of $P \leq 3.48 \times 10^{-5}$) were found for 305 traits. A subset of relevant significant correlations is presented in figure 2, with full results in Supplementary table 4.

Reading ability showed strongly positive correlations (r_g) with measures of intelligence, specifically cognitive performance (0.57, SE = 0.02) (SSGAC) and verbal-numerical reasoning (UKBiobank) (0.61, SE = 0.02). Consistent with previous findings, higher reading ability was positively genetically

correlated with academic achievement and education level, including educational attainment (0.30, SE = 0.02) (SSGAC), completing college or a university degree (0.32, SE = 0.02) and completing GCSEs or equivalent (0.58, SE = 0.02). Reading ability showed negative genetic correlations with achieving either vocational qualifications (NVQ or HND or HNC: -0.57, SE = 0.05; CSEs or equivalent: -0.48, SE = 0.04) or not completing any qualifications (-0.39, SE = 0.02).

In terms of neurodevelopmental traits, ADHD¹⁴ showed a negative correlation with reading ability (-0.43, SE = 0.03). This was weaker than that shown by the dyslexia study alone (0.53, SE = 0.12)⁴, and while not reported by Eising *et al*⁵, we calculated r_g for the univariate word-reading GWAS as -0.40 (SE = 0.05). In contrast, ASD did not show a significant correlation (-0.08, SE = 0.04, $P = 3.15 \times 10^{-2}$) with reading ability.

Major depressive disorder showed a negative correlation (-0.13, SE = 0.03), as did measures linked to poorer wellbeing including manic episodes, risk taking and mood swings. Measures of pain and use of pain medications were also negatively correlated, including toothache (-0.39, SE = 0.06), pain all over the body (-0.45, SE = 0.06) and back pain (UKBB) (-0.24, SE = 0.03). A moderately strong positive correlation was seen between reading ability and no pain experienced in the last month (0.34, SE = 0.03). Reading ability showed negative genetic correlations with measures of lower socio-economic status and less desirable workplace conditions, and positive correlations with higher socioeconomic and health measures.

As the two constituent univariate summary statistics were derived from GWAS efforts with different sample sizes ($N = 27\,180$ versus $N_{\text{cases}} = 51\,800 / N_{\text{controls}} = 1\,087\,070$) and type of measure (quantitatively assessed continuous measure versus self-reported binary), we investigated whether the multivariate GWAS for reading ability was overpowered by signal from the dyslexia study. To do this we estimated genetic correlations between multivariate reading ability and univariate dyslexia ($r_g = -0.98$, SE = 0.00, $Z = -334.76$, $P = 0.00$), and between multivariate reading-ability and univariate GenLang word reading ($r_g = 0.84$, SE = 0.03, $Z = 27.63$, $P = 4.66 \times 10^{-168}$). Whereas these correlations suggested a bias from the dyslexia GWAS, there was clear evidence that the multivariate reading-

ability results were indeed capturing the genes associated with quantitative reading ability with the genetic correlation almost reaching 0.90.

[Figure 2 about here]

Gene and Gene-set Associations

Gene-based analysis of the multivariate reading ability summary statistics identified 103 genes, meeting the Bonferroni-derived α level for 18,842 tests ($P < 2.65 \times 10^{-6}$). Most of these genes associated with reading ability ($N = 79$) were also present in associated regions detected by GWAS, while 24 fell outside of associated regions from the SNP-based screen (supplementary table 5). The overall number of associated genes ($N = 103$) was lower than the 173 detected by Doust *et al* (2022) in the gene-based testing of the original dyslexia study⁴. Fourteen genes were statistically associated in the present study but not in Doust *et al* (supplementary table 5). Only three of these genes fall within regions that met genome-wide significance in the present multivariate reading ability GWAS. Eising *et al* (2022) did not report a gene-based association analysis. However, when we perform this using FUMA, only one gene meets the Bonferroni significance threshold (*AC079354.1*, $P = 7.4 \times 10^{-7}$), with candidate genes *DOCK7* and *USP1* showing the next strongest associations just above the Bonferroni threshold⁵.

MAGMA gene-set analysis detected enrichment of two biological pathways from 9 113 curated gene sets and gene ontology (GO) terms, tested with a Bonferroni-derived threshold ($P < 5.49 \times 10^{-6}$). The Gene-Set Enrichment Analysis term for Verhaak glioblastoma proneural (genes correlated with proneural type of glioblastoma multiforme tumours) containing 170 genes showed significant association (Beta = 0.35, SE = 0.08, $P = 2.08 \times 10^{-6}$). Similarly, the GO term for cellular compartment (GOCC) synapse containing 1 390 genes showed significant association (Beta = 0.12, SE = 0.03, $P = 2.71 \times 10^{-6}$). See supplementary table 6 for the full gene-set enrichment analysis.

Variant Mapping and Functional Annotation

Candidate SNPs ($N = 7\,548$) found in LD ($R^2 \leq 0.6$) with one of the independent significant SNPs, including tagged SNPs from the 1000 Genomes reference panel, underwent variant annotation using the Variant Effect Predictor (VEP) online tool (<http://grch37.ensembl.org/>). A total of 14 985 individual annotations were identified for candidate SNPs, due to multiple allelic variants per SNP. Intronic variants were most common (58%) with coding variants making up 0.68% (supplementary figure 4a). Of the coding variants, 51% were missense and 8% were stop-gains (supplementary figure 4b).

Six variants predicted as damaging by SIFT and PolyPhen and with CADD scores > 25 , an indication for possible deleterious effects of the variants, were found: rs11142 (chr1:109897103) in *SORT1*, rs1983864 (chr10:100017453, tag SNP) in *LOXL4*, rs1064213 (chr2:198950240, tag SNP) in *PLCL1*, rs1130146 (chr 20:47859217, tag SNP) in *DDX27*, and rs10891314 (chr11:111916647) in *DLAT* with two allelic variants, G/A and G/T (supplementary table 7). Two of the variants with CADD scores > 25 (but not annotated by SIFT or PolyPhen) were predicted to result in stop gain changes: rs3764090 (chr13:50008301) in *AL136218.1* and rs2424922 (chr20:31386449, T/A) in *DNMT3B*.

At the gene level, 591 genes were contained within genome-wide significant regions (supplementary table 8). Sensitivity to loss-of-function was annotated with probability of loss-intolerance scores (pLI) and sensitivity to non-coding variation in regulatory sequences was annotated with non-coding residual intolerance scores (ncRVIS). One-hundred and twenty-five genes (21.2%) were predicted as loss-of-function intolerant by $pLI \geq 0.9$, and seven (1.2%) were predicted as less tolerant to non-coding variation by $ncRVIS \geq 2.0$. Three genes (*SIK2*, *PTPN14* and *XYLT1*) were predicted as intolerant by both metrics ($pLI \geq 0.9$ and $ncRVIS \geq 2.0$).

Ninety-seven genes ($N = 97/591$) located within associated regions showed evidence of association with expression QTLs (eQTL) in brain tissue ($P \leq 5.94 \times 10^{-5}$) (supplementary table 8). The strongest

eQTL associations were for *DHRS11* ($P = 6.36 \times 10^{-77}$), *CYB56I* ($P = 1.41 \times 10^{-69}$) and *SETDB2* ($P = 7.10 \times 10^{-52}$).

Functional Enrichment using Partitioned Heritability and Gene Property Analysis

To examine the tissue-specific expression profiles of genes implicated in reading ability, we used MAGMA gene property analysis within FUMA. Using RNA-seq data from the Genotype-Tissue Expression (GTEx) project, we found significant enrichments of genes associated with reading ability in brain tissue: 11 brain regions tested showed significantly higher expression levels of reading ability associated genes, particularly the cerebellum, cerebellar hemisphere and frontal cortex (supplementary figure 5, supplementary table 9) at a Bonferroni threshold of $P < 4.03 \times 10^{-4}$ corrected for 124 tests (GTEx and BrainSpan combined). As MAGMA analysis corrects for the average expression level in the dataset, a significant association indicates that genes associated with reading ability have a higher expression in that tissue relative to the average expression within the dataset.

Next, we tested for enrichment within the BrainSpan data set, consisting of RNA-seq from 11 developmental stages (supplementary figure 6, supplementary table 10) and 29 ages (supplementary figure 7, supplementary table 11) of human brains. No associations met Bonferroni correction for 124 tests ($P < 4.03 \times 10^{-4}$).

To further investigate enrichment of gene expression within the developing human brain, we tested for associations with specific cell types in single cell RNA-seq (scRNA) data using MAGMA within FUMA. Embryonic ventral midbrain from 6-11 post conception weeks (pcw) embryos (GSE76381) revealed enrichment in three cell types at Bonferroni corrected threshold of $P < 6.58 \times 10^{-4}$ for 76 tests (figure 3a, supplementary table 12). These were GABAergic neurons (Gaba, $P = 1.07 \times 10^{-5}$), neuroblast GABAergic neurons (NbGaba, $P = 3.79 \times 10^{-4}$), and red nucleus neurons (RN, $P = 2.03 \times 10^{-4}$). Embryonic prefrontal cortex scRNA-seq data from pcw 4-26 (GSE104276) showed significant enrichment in GABAergic neurons at 26 pcw ($P = 5.00 \times 10^{-7}$), neurons at 26 pcw ($P = 2.85 \times 10^{-4}$), and neurons at 16 pcw ($P = 5.59 \times 10^{-5}$), meeting the Bonferroni corrected threshold (76 tests, $P < 6.58 \times 10^{-4}$).

(figure 3b, supplementary table 13). Finally, scRNA-seq data from foetal and adult human cortex (GSE67835) showed significant expression in adult cortex neurons ($P = 1.10 \times 10^{-5}$) (76 tests, $P < 6.58 \times 10^{-4}$) (figure 3c, supplementary table 14).

Heritability partitioning by LDSC identified statistically significant enrichment of variants associated with reading ability in chromatin signatures annotated in foetal and adult brain tissues obtained from the Roadmap Epigenomics project and Enhancing GTEx project (ENTEEx) (figure 4, supplementary table 15). Out of 489 chromatin signatures tested, twenty-nine annotations were significantly enriched (Bonferroni corrected threshold $P < 1.02 \times 10^{-4}$), from foetal brain ($N = 6$), adult brain ($N = 21$) and primary neuronal cultured cell lines ($N = 2$), across a range of chromatin signatures of (active) enhancers and promoters and actively transcribed regions.

[Figure 3 about here]

[Figure 4 about here]

Polygenic Scores Prediction in NCDS

Polygenic scores from the multivariate GWAS of reading ability were computed in an independent cohort across five developmental stages, plus an overall composite measure. The reading-ability PGS explained between 2.29 and 3.50% of variance in reading ability in the NCDS cohort. Predictions for composite measures of reading proficiency were 2.78% at age 7 years ($N = 5\,712$), 2.61% at age 11 years ($N = 5\,528$), and 2.29% at age 16 years ($N = 4\,809$). Binary measures of struggling with reading scored highest at 3.5% at age 23 ($N_{\text{cases}} = 167$, $N_{\text{controls}} = 5\,288$) and 2.55% at age 33 ($N_{\text{cases}} = 203$, $N_{\text{controls}} = 5\,497$). The prediction for the overall composite measure of reading proficiency across all ages was 3.32% ($N = 3\,089$) (supplementary figures 8-13; supplementary table 16).

Polygenic Selection Analysis

The polygenic selection analysis examined if there was evidence of selection on alleles associated with reading ability seen through the past 15 000 years of human history. Essentially, this looks for differences in allele frequencies in variants associated with a trait, between the ancient ancestral population(s) and the present day. Such shifts in frequency indicate that selection acted to change the allele frequency of that variant in response to environmental pressures. From the 86 significant independent variants within 35 loci associated with reading ability in the multivariate GWAS, 42 were retained after clumping and present at high quality in the previously generated imputed ancestral data set^{18, 19, 28}. Overall analyses of these SNPs identified no evidence of directional selection acting on reading ability over the past 15 000 years ($P = 0.89$) (figure 4). Of the 42 SNPs submitted for analysis, eight SNPs individually showed statistically significant evidence of directional selection, after accounting for the number of tests (Bonferroni threshold of $P < 1.12 \times 10^{-3}$ for 42 tests) (supplementary table 17).

rs10774624 in gene *LINC02356* showed the most pronounced increase in allele frequency over time and showed a positive effect on reading ability. The variant is also associated with a wide range of traits including arterial disease³⁶ and rheumatoid arthritis³⁷. Two further SNPs associated with positive effects on reading ability were rs1317140 in gene *TRAIP* (proxy to rs13316065), which is also associated with body mass index³⁸ and urate measurement³⁹, and rs7204631 with no reported association. However, both these SNPs showed a decrease in allele frequency over time.

The strongest signal in alleles associated with poorer reading ability was rs17687323 on chromosome 17 with no prior associations with other traits and showed a reduction in allele frequency over time.

The remaining four SNPs also associated with poorer reading performance, were at increased frequency in modern populations; rs72916919 (no reported associations with other traits), rs9879531 in gene *NCK1* (linked to blood urea nitrogen measurement⁴⁰), rs787995 (no association), and rs9880211 in *STAG1* (previously associated with body height⁴¹). It is plausible that these eight variants individually showed modest directional selection because of their contributions to traits other than reading ability, given that our overall analyses showed no evidence of selection for the latter through the past 15 000 years.

[Figure 4 about here]

Discussion

We performed a multivariate GWAS using MTAG on reading ability using summary statistics from the two largest reading-related cohorts to increase the effective population size to 102 082, which is substantially larger than previous studies. We detected 35 independent loci associated with reading ability, including 7 novel loci that were not significantly associated in either of the original univariate GWAS studies. The majority of loci ($N = 26$) were previously reported in the univariate dyslexia GWAS⁴. The most significantly associated independent SNPs in the multivariate GWAS; rs13082684 (*PPP2R3A*), rs348258 (*STAU1*) and rs1075218 (*MRM1*), were consistent with loci reported in the dyslexia GWAS⁴.

The most significantly associated SNP identified by the GenLang word-reading meta-GWAS⁵, rs11208009, did not reach genome-wide significance in the present multivariate study ($P = 2.71 \times 10^{-6}$). It fell within a region that reached suggestive significance (chr1:62900811-63199936, $P = 1.96 \times 10^{-7}$), fully overlapping with the previously reported region.

We also used the multivariate approach to identify novel loci for dyslexia, increasing the effective population size to 1 228 832. We detected 80 loci using the multivariate approach including 13 novel loci, boosting 19 regions that were previously suggestive, and seven that were novel and present in both of our multivariate analyses. However, we focussed the follow-up analyses on the reading ability output, and not on dyslexia output, because the vast increase in effective sample size compared to the respective univariate analysis, providing substantially more power than in the original GenLang study.

Annotation of associated variants indicated only six genes (seven variants) that contained potentially deleterious coding variants associated with reading ability, as predicted damaging by SIFT, PolyPhen and CADD. At the gene level, of the 591 genes contained within the associated regions, only 125 (21.2%) were predicted as loss-of-function intolerant, and seven (1.2%) were predicted as less tolerant to non-coding variation. Three genes (*SIK2*, *PTPN14* and *XYLT1*) were predicted as intolerant by both

metrics. *SIK2* was previously reported in the prior dyslexia GWAS⁴, where as *PTPN14* and *XYLT1* are novel. eQTL associations revealed a large number of genes (N = 97) with evidence of expression associations in brain tissue, similar to those identified in the prior dyslexia GWAS, the strongest associations were identified in genes *DHRS11* and *CYB56L*.

We identified seven novel loci that were not previously significantly associated with either word reading or dyslexia in the original univariate GWASs. The novel lead SNP, rs362307 within *HTT*, has been associated with a range of cognitive traits relevant to reading ability in prior studies, including cognitive function³² and educational attainment³¹. Genomic structural equation modelling and genetic correlation analyses have shown that reading-related traits have substantial genetic overlaps with educational attainment and IQ^{4, 5}, and therefore SNPs involved in general cognition are likely to be important across traits. The remaining six loci were identified in the univariate dyslexia study at a sub-genome-wide association threshold⁴, but are present in the GWAS Catalogue⁴². One of these regions, chr16:72333576-72697419 in gene *AC004158.2* (rs2158266) was previously reported by Gialluisi and colleagues as suggestively associated with phoneme awareness, deficits of which are considered a marker of dyslexia³⁵. That study included 3 093 individuals of European ancestry from the Neurodys, CLDRC and UKdys cohorts, which also formed part of the GenLang meta-GWAS, meaning that these findings are not fully independent.

We predicted a maximum of 3.5% of trait variance in reading measures in the NCDS using the multivariate word reading polygenic scores derived from our multivariate reading ability GWAS. This is an improvement on the previous predictions using the dyslexia GWAS on measures of word reading in similar population-based cohorts where 2.9% of variance was explained in adolescents and ~2% in adults⁴. Predictions across longitudinal measures at five ages within the NCDS ranged from 2.29-3.5% suggesting that this part of the genetic contribution of reading is stable through schooling and into adulthood. Other studies have used phenotypically related PGS to predict reading outcomes. For example, Selzam *et al* (2017)⁴³ used a years-in-education PGS⁴⁴ to predict 5% of reading variance at 14 years of age. Future studies might test whether our reading ability PGS explains incremental

variance in reading variation above an educational attainment PGS, which is a broader phenotype encompassing cognitive and non-cognitive factors.

Genetic correlations between the multivariate reading-ability trait and other phenotypes showed a high degree of similarity to the profiles of genetic correlations seen for both the univariate GWAS studies of word reading and dyslexia. Key findings were consistent with the univariate GWAS, such as strong genetic correlations with educational attainment, completing a college or university degree, and a negative correlation with manual jobs. Notably, the strongest correlation for the multivariate reading-ability GWAS was verbal-numerical reasoning in the UKBiobank (0.61, SE = 0.02), which was also among the strongest correlations for Dyslexia at -0.5 (SE = 0.03)⁴.

Correlations with other neurodevelopmental traits were reasonably consistent for the univariate and multivariate reading-ability GWASs. The strongest correlation of the dyslexia GWAS was with ADHD at 0.53 (SE = 0.01)⁴, and although this correlation was lower in absolute magnitude for the multivariate reading-ability analysis (-0.43, SE = 0.03), it remains consistent with the literature linking ADHD with reading and language outcomes^{9, 45}. The observation that genetic correlation with ADHD is stronger with dyslexia than with reading ability overall may indicate that dyslexia diagnosis is influenced by more than just reduced reading performance, and that these other considerations overlap with indicators of ADHD. Consistent with previous findings that there is no significant genetic correlation between autism spectrum disorder (ASD) and reading ability^{4, 5}, no significant correlation was found in the multivariate reading-ability investigation carried out here.

Gene-set analysis revealed two enriched biological pathways implicated in reading ability: the gene-ontology term for cellular compartment (GOCC) neural synapses, and for Verhaak glioblastoma proneural (genes correlated with proneural type of glioblastoma multiforme tumours). Both gene-sets hint at essential neuronal mechanisms.

Our analysis of expression patterns of associated genes in the developing human brain offered further evidence for a role in early developmental processes, implicating GABAergic and red neurons in embryonic prefrontal cortex and midbrain, as well as cortex neurons in adults. These findings echo the

cell type analysis reported in the GenLang paper, where an enrichment in red nucleus neurons and a trend towards enrichment in fetal GABAergic neurons was observed⁵. Price and colleagues reported evidence supporting neuronal migration/axon guidance as potential pathways using a candidate gene-set approach for known neurodevelopmental genes in a hypothesis-driven association analysis of word reading which included the GenLang meta-GWAS⁴⁶. More recently, the same research group implicated glutamatergic (excitatory) and GABAergic (inhibitory) neurons in the adult cortex in word reading, using a subset of the GenLang cohorts (N = 5 054)¹⁷. The sample used in the present study overlaps with that of the prior work, which may contribute to the consistency between the two sets of results. Despite the small GWAS sample size used by Price and colleagues, their work offers support for the GABAergic inhibitory system as a future focus for connecting genetics to neuronal mechanisms.

Polygenic selection analysis found no significant selection observed from ancestral populations suggesting that the genetic influences on reading skills were not specifically selected for or against in the transition between hunter gatherer and farmers in Europeans. This finding may be considered unsurprising since reading is many thousands of years old but has only very recently become widespread and has no obvious selection pressure or effect on reproductive fitness. However, because reading processes are highly dependent on brain circuits that evolved in support of spoken language, it was still possible that we may have detected signals related to language evolution. The consistency of the PGS through the past 15k years of history in northern Europe suggests it has not been affected by any major social or societal changes that have taken place in history such as the transition to farming, although it is important to note that our PGS accounts for only a small proportion of heritable reading ability. We identified eight SNPs that showed individually significant changes through recent history; five showed evidence of negative selection over time and three showed signs of positive selection, although their direction of effect on reading-ability were mixed. Considering that reading ability is likely a neutral trait with regard to biological fitness, we speculate that the patterns observed for these eight SNPs are most likely due to selection pressures acting on pleiotropic traits.

Analysis that examines further back through evolution (30 million years ago to 50 000 years ago) was performed in both the Eising *et al*⁵ and Doust *et al*⁴ papers. Based on findings of human brain structure evolution⁴⁷, five annotations reflecting aspects of human evolution were examined. Doust *et al* found no evidence of enrichment for annotations related to human evolution. Eising *et al* found evidence of an enrichment in archaic deserts; long regions in the human genome where there is an absence of Neanderthal admixture, suggesting these regions may be intolerant to gene flow and therefore harbouring variants essential to *Homo sapiens*. The findings suggested that these archaic desert regions could contain genetic variations that contribute more to reading and language traits in modern humans than expected by chance.

In terms of limitations of the current study, we note that the false discovery rate (FDR) was moderate at 6%, primarily because of the size imbalance between the constituent summary statistics. Genetic correlation between the multivariate reading ability phenotype was marginally stronger with dyslexia than GenLang word reading. When considered along with the MTAG modest FDR, they suggest the observed patterns of genetic associations may be influenced by the larger dyslexia GWAS.

In summary, through implementation of multivariate GWAS analysis combining work on quantitative measures with self-report of disorder, we have produced the largest genetic study of reading ability (effective N = 102 082) to date. Our findings account for 24% (h^2_{snp}) of the observed variability of reading ability in our datasets. We identified seven novel loci associated with quantitative reading ability and implicated early brain developmental processes in the biological underpinnings of reading.

Acknowledgements

We thank the participants and employees of 23andMe, Inc, and the participants and research members of the GenLang Consortium. We are grateful to the Centre for Longitudinal Studies (CLS), UCL Social Research Institute, for the use of the NCDS data and to the UK Data Service for making them available. However, neither CLS nor the UK Data Service bear any responsibility for the analysis or interpretation of these data.

We thank Drs David Hill and Charley Xia for their assistance with the multivariate analyses and genetic correlations, and to Jonny Flint for assistance with the polygenic score analysis.

SEF and EE are supported by the Max Planck Society (Germany). EE is also supported by a Veni grant of the Dutch Research Council (NWO; VI.Veni.202.072). HSM is supported by the Biotechnology and Biological Sciences Research Council [BB/T000813/1].

The following members of the 23andMe Research Team contributed to this study:

Stella Aslibekyan, Adam Auton, Elizabeth Babalola, Robert K. Bell, Jessica Bielenberg, Jonathan Bowes, Katarzyna Bryc, Ninad S. Chaudhary, Daniella Coker, Sayantan Das, Emily DelloRusso, Sarah L. Elson, Nicholas Eriksson, Teresa Filshtein, Pierre Fontanillas, Will Freyman, Zach Fuller, Chris German, Julie M. Granka, Karl Heilbron, Alejandro Hernandez, Barry Hicks, David A. Hinds, Ethan M. Jewett, Yunxuan Jiang, Katelyn Kukar, Alan Kwong, Yanyu Liang, Keng-Han Lin, Bianca A. Llamas, Matthew H. McIntyre, Steven J. Micheletti, Meghan E. Moreno, Priyanka Nandakumar, Dominique T. Nguyen, Jared O'Connell, Aaron A. Petrakovitz, G. David Poznik, Alexandra Reynoso, Shubham Saini, Morgan Schumacher, Leah Selcer, Anjali J. Shastri, Janie F. Shelton, Jingchunzi Shi, Suyash Shringarpure, Qiaojuan Jane Su, Susana A. Tat, Vinh Tran, Joyce Y. Tung, Xin Wang, Wei Wang, Catherine H. Weldon, Peter Wilton, Corinna D. Wong.

Conflict of Interest

PF, AA and the 23andMe Research Team are employed by and hold stock or stock options in 23andMe, Inc. All other authors declare no conflicts of interest.

Supplementary information is available at Molecular Psychiatry's website

References

1. Ritchie SJ, Bates TC. Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychol Sci* 2013; **24**(7): 1301-1308.
2. Davis OS, Band G, Pirinen M, Haworth CM, Meaburn EL, Kovas Y *et al*. The correlation between reading and mathematics ability at age twelve has a substantial genetic component. *Nat Commun* 2014; **5**: 4204.
3. Andreola C, Mascheretti S, Belotti R, Ogliari A, Marino C, Battaglia M *et al*. The heritability of reading and reading-related neurocognitive components: A multi-level meta-analysis. *Neurosci Biobehav Rev* 2021; **121**: 175-200.
4. Doust C, Fontanillas P, Eising E, Gordon SD, Wang Z, Alagoz G *et al*. Discovery of 42 genome-wide significant loci associated with dyslexia. *Nat Genet* 2022; **54**(11): 1621-1629.
5. Eising E, Mirza-Schreiber N, de Zeeuw EL, Wang CA, Truong DT, Allegrini AG *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 U S A* 2022; **119**(35): e2202764119.
6. Pennington BF, Bishop DV. Relations among speech, language, and reading disorders. *Annu Rev Psychol* 2009; **60**: 283-306.
7. Protopapas A. Evolving Concepts of Dyslexia and Their Implications for Research and Remediation. *Front Psychol* 2019; **10**: 2873.
8. Snowling MJ, Melby-Lervag M. Oral language deficits in familial dyslexia: A meta-analysis and review. *Psychol Bull* 2016; **142**(5): 498-545.
9. Gialluisi A, Andlauer TFM, Mirza-Schreiber N, Moll K, Becker J, Hoffmann P *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.
10. Bates TC, Castles A, Luciano M, Wright MJ, Coltheart M, Martin NG. Genetic and environmental bases of reading and spelling: A unified genetic dual route model. *Read Writ* 2007; **20**(1-2): 147-171.
11. Haworth CMA, Kovas Y, Harlaar N, Hayiou-Thomas ME, Petrill SA, Dale PS *et al*. Generalist genes and learning disabilities: a multivariate genetic analysis of low performance in reading, mathematics, language and general cognitive ability in a sample of 8000 12-year-old twins. *J Child Psychol Psyc* 2009; **50**(10): 1318-1325.
12. Luciano M, Evans DM, Hansell NK, Medland SE, Montgomery GW, Martin NG *et al*. A genome-wide association study for reading and language abilities in two population cohorts. *Genes Brain Behav* 2013; **12**(6): 645-652.
13. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H *et al*. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* 2019; **51**(3): 431-444.

14. Demontis D, Walters GB, Athanasiadis G, Walters R, Therrien K, Nielsen TT *et al*. Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nat Genet* 2023; **55**(2): 198-208.
15. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA *et al*. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* 2018; **50**(2): 229-237.
16. Wu Y, Cao H, Baranova A, Huang H, Li S, Cai L *et al*. Multi-trait analysis for genome-wide association study of five psychiatric disorders. *Transl Psychiatry* 2020; **10**(1): 209.
17. Price KM, Wigg KG, Nigam A, Feng Y, Blokland K, Wilkinson M *et al*. Identification of brain cell types underlying genetic association with word reading and correlated traits. *Mol Psychiatry* 2023.
18. Barrie W, Yang Y, Irving-Pease EK, Attfield KE, Scorrano G, Jensen LT *et al*. Elevated genetic risk for multiple sclerosis emerged in steppe pastoralist populations. *Nature* 2024; **625**(7994): 321-328.
19. Irving-Pease EK, Refoyo-Martinez A, Barrie W, Ingason A, Pearson A, Fischer A *et al*. The selection landscape and genetic legacy of ancient Eurasians. *Nature* 2024; **625**(7994): 312-320.
20. Wickham H. *ggplot2 : Elegant Graphics for Data Analysis*. 2nd 2016. edn. Springer-Verlag: New York, 2016.
21. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017; **8**(1): 1826.
22. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C *et al*. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47**(3): 291-295.
23. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015; **11**(4): e1004219.
24. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015; **47**(11): 1228-1235.
25. University College London. National Child Development Study. In: UCL Institute for Education CfLS (ed). 13th release edn2023.
26. Bridges EC, Rayner NW, Mountford HS, Bates TC, Luciano M. Longitudinal Reading Measures and Genome Imputation in the National Child Development Study: Prospects for Future Reading Research. *Twin Res Hum Genet* 2023: 1-11.

27. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* 2019; **8**(7).
28. Allentoft ME, Sikora M, Refoyo-Martinez A, Irving-Pease EK, Fischer A, Barrie W *et al.* Population genomics of post-glacial western Eurasia. *Nature* 2024; **625**(7994): 301-311.
29. Stern AJ, Wilton PR, Nielsen R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *Plos Genet* 2019; **15**(9).
30. Stern AJ, Speidel L, Zaitlen NA, Nielsen R. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *Am J Hum Genet* 2021; **108**(2): 219-239.
31. Okbay A, Wu Y, Wang N, Jayashankar H, Bennett M, Nehzati SM *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat Genet* 2022; **54**(4): 437-449.
32. Davies G, Lam M, Harris SE, Trampush JW, Luciano M, Hill WD *et al.* Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat Commun* 2018; **9**(1): 2098.
33. Hill WD, Weiss A, Liewald DC, Davies G, Porteous DJ, Hayward C *et al.* Genetic contributions to two special factors of neuroticism are associated with affluence, higher intelligence, better health, and longer life. *Mol Psychiatry* 2020; **25**(11): 3034-3052.
34. Nagel M, Watanabe K, Stringer S, Posthuma D, van der Sluis S. Item-level analyses reveal genetic heterogeneity in neuroticism. *Nat Commun* 2018; **9**(1): 905.
35. Gialluisi A, Andlauer TFM, Mirza-Schreiber N, Moll K, Becker J, Hoffmann P *et al.* Genome-wide association scan identifies new variants associated with a cognitive predictor of dyslexia. *Transl Psychiatry* 2019; **9**(1): 77.
36. van Zuydam NR, Stiby A, Abdalla M, Austin E, Dahlstrom EH, McLachlan S *et al.* Genome-Wide Association Study of Peripheral Artery Disease. *Circ Genom Precis Med* 2021; **14**(5): e002862.
37. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014; **506**(7488): 376-381.
38. Zhu Z, Guo Y, Shi H, Liu CL, Panganiban RA, Chung W *et al.* Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *J Allergy Clin Immunol* 2020; **145**(2): 537-549.
39. Gill D, Cameron AC, Burgess S, Li X, Doherty DJ, Karhunen V *et al.* Urate, Blood Pressure, and Cardiovascular Disease: Evidence From Mendelian Randomization and Meta-Analysis of Clinical Trials. *Hypertension* 2021; **77**(2): 383-392.

40. Wuttke M, Li Y, Li M, Sieber KB, Feitosa MF, Gorski M *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* 2019; **51**(6): 957-972.
41. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014; **46**(11): 1173-1186.
42. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* 2023; **51**(D1): D977-D985.
43. Selzam S, Dale PS, Wagner RK, DeFries JC, Cederlof M, O'Reilly PF *et al.* Genome-Wide Polygenic Scores Predict Reading Performance Throughout the School Years. *Sci Stud Read* 2017; **21**(4): 334-349.
44. Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 2016; **533**(7604): 539-542.
45. Verhoef E, Demontis D, Burgess S, Shapland CY, Dale PS, Okbay A *et al.* Disentangling polygenic associations between attention-deficit/hyperactivity disorder, educational attainment, literacy and language. *Transl Psychiatry* 2019; **9**(1): 35.
46. Price KM, Wigg KG, Eising E, Feng Y, Blokland K, Wilkinson M *et al.* Hypothesis-driven genome-wide association studies provide novel insights into genetics of reading disabilities. *Transl Psychiatry* 2022; **12**(1): 495.
47. Tilot AK, Khramtsova EA, Liang D, Grasby KL, Jahanshad N, Painter J *et al.* The Evolutionary History of Common Genetic Variants Influencing Human Cortical Surface Area. *Cereb Cortex* 2021; **31**(4): 1873-1887.

Figure Legends

Figure 1: Manhattan plot of the multivariate GWAS of a) reading ability and b) dyslexia. The y axis indicates the $-\log_{10} P$ value for association. The threshold for genome-wide significance ($P < 5 \times 10^{-8}$) is represented by a dashed grey line. Significant loci that were previously reported in the GenLang word reading GWAS are represented in red, and those reported in the dyslexia GWAS are shown in purple.

Figure 2: Genetic correlations of reading ability with selected relevant phenotypes. Significant ($P \leq 3.561 \times 10^{-5}$) genetic correlations (r_g) between multivariate analysis of reading ability and other selected phenotypes

Figure 3: MAGMA gene property analyses of reading ability associated genes with single cell gene expression data from a) embryonic ventral midbrain from 6-11 post gestational weeks (pgw), b) embryonic prefrontal cortex from 4 – 26 pgw, c) human fetal and adult cortex. Neuronal cell lineages in a) are: DA0-1 – dopaminergic neurons, Endo – endothelial cells, Gaba – GABAergic neurons, Mgl – microglia, NbGaba - neuroblast GABAergic, NbM – medial neuroblast, NbML1-5 – mediolateral neuroblasts, NProg – neuronal progenitor, OMTN - oculomotor and trochlear nucleus, OPC – oligodendrocyte precursor cells, Peric – pericytes, ProgBP – progenitor basal plate, ProgFPL – progenitor medial floorplate, ProgM – progenitor midline, Rgl1-3 – radial glia-like cells, RN – red nucleus, Sert – serotonergic

Figure 4: Partitioned heritability enrichment analysis of chromatin signatures. SNP-based heritability of the multivariate reading-ability GWAS is significantly enriched in brain enhancers, promoters and transcribed regions. 489 annotation of tissue-specific chromatin signatures were used

to analyse the GWAS results with LDSC heritability partitioning. Only brain-related annotations are shown. P values are plotted on the y axis as $-\log_{10}$.

Figure 5: No evidence for directional selection of reading ability associated SNPs. Stacked line plot of the ancient ancestry PALM analysis, showing the contribution of SNPs to reading ability over time. SNPs are shown as stacked lines, the width of each line being proportional to the population frequency of the positive effect allele, weighted by its effect size. When a line widens over time the positive effect allele has increased in frequency, and vice versa. SNPs are sorted by the magnitude and direction of selection, with positively selected SNPs at the top, negatively selected SNPs at the bottom, and neutral SNPs in the middle. SNPs are coloured by their corresponding P -value in a single locus selection test. The asterisk on the scale bar marks the Bonferroni corrected significance threshold, and nominally significant SNPs are shown in yellow and labelled by their rsIDs. The Y-axis shows the scaled average polygenic score (PGS) in the population, ranging from 0 to 1, with 1 corresponding to the maximum possible average PGS (i.e. when all individuals in the population are homozygous for all positive effect alleles) and the X-axis shows time in units of thousands of years before present (kyr BP).

Table 1: Loci associated with reading ability detected by multivariate GWAS (MTAG), with the lead SNP indicated. Genes which were significant in the gene-based test are indicated in bold, and those unavailable to test are indicated by *. Regions previously reported are indicated by A (Doust et al 2022) or B (Eising et al 2022). Where the previous lead SNP or genes have changed in the current analysis, these are indicated in bold. ^s indicates regions that were called as one locus in Doust et al 2022, and two in the present analysis.

Table 2: Novel regions associated with reading ability from multivariate GWAS using MTAG. Table shows associated region, lead SNP, minimum P-value, and the number of significant SNPs within the regions. The closest gene to the Lead SNP is listed, as well as other genes falling within the associated region. Relevant known associations output from FUMA are listed for both the SNP and the gene, listed as SNP: gene.

Tables

Cytoband	Lead SNP	Independent Significant SNPs at locus	Effect allele	Frequency	Effect	SE	MTAG P	Nearest Gene(s) to lead SNP	Previous Lead SNP	Previous P	Previous Gene(s)
chr1p32.1	rs12723322	rs890292	T	0.125	-0.039	0.006	5.00 x 10 ⁻¹⁰	<i>[C1orf87]</i>	rs12737449 ^A	1.40 x 10 ⁻¹¹	<i>[C1orf87]</i>
chr1p13.3	rs3853500		T	0.287	0.031	0.005	3.75 x 10 ⁻⁹	<i>[SORT1]</i>	rs2091329 ^A	1.94 x 10 ⁻⁹	<i>CYB561D1-[/-AMIGO1</i>
chr1q21.3	rs11264291		T	0.437	0.026	0.005	1.58 x 10 ⁻⁸	<i>PMVK--[/-PBXIP1</i>	rs4845687 ^A	1.08 x 10 ⁻⁹	<i>KCNN3--[/-PMVK</i>
chr1q41	rs17043436		G	0.156	0.033	0.006	3.12 x 10 ⁻⁸	<i>[ESRRG]*</i>	rs35570426 ^A	4.10 x 10 ⁻⁸	<i>[ESRRG]</i>
chr2p22.1	rs906549	rs2217363	T	0.345	-0.030	0.005	3.15 x 10 ⁻¹¹	<i>SLC8A1---[/-AC010739.1*</i>	rs906549 ^A	1.40 x 10 ⁻⁹	<i>SLC8A1---[/-C2orf91</i>
chr2q22.3	rs2381977	rs10432338; rs1376383	T	0.392	0.028	0.005	1.1 x 10 ⁻⁹	<i>AC062032.1*---[/-RP11-279E17.1*</i>	rs497418 ^A	3.00 x 10 ⁻⁹	<i>[/-ACVR2A</i>
chr2q32.2	rs719166		G	0.161	-0.032	0.006	1.82 x 10 ⁻⁸	<i>AC092638.2*---[/-RP11-305O20.1*</i>	rs719166 ^B	1.46 x 10 ⁻⁷	
chr2q33.1	rs72916919	rs7571545; rs1560277	T	0.463	0.029	0.004	3.15 x 10 ⁻¹¹	<i>[RFTN2]</i>	rs72916919 ^A	4.10 x 10 ⁻¹²	<i>[RFTN2]</i>
chr2q33.1	rs11692189	rs2084735	G	0.309	-0.026	0.005	2.51 x 10 ⁻⁸	<i>AC018717.1*---[/-SATB2</i>	rs6435017 ^A	4.92 x 10 ⁻⁹	<i>[SATB1]</i>
chr3p24.3	rs12493567		G	0.434	0.024	0.004	3.80 x 10 ⁻⁸	<i>[TBC1D5]</i>	rs1374197 ^A	2.33 x 10 ⁻⁹	<i>[TBC1D5]</i>
chr3p21.31	rs13316065		T	0.304	-0.027	0.005	2.16 x 10 ⁻⁸	<i>[TRAIP]</i>	rs2624839 ^{AS}	3.00 x 10 ⁻⁹	<i>[SEMA3F]</i>

chr3p21.31	rs6800021		G	0.398	0.025	0.004	1.64 x 10 ⁻⁸	[<i>RP11-493K19.3</i>]*			
chr3p12.1	rs4856600		C	0.350	0.025	0.005	3.54 x 10 ⁻⁸	[<i>CADM2</i>]	rs10511073 ^A	4.61 x 10 ⁻¹⁰	[/--- <i>CADM2</i>
chr3q22.3	rs13082684	rs34153877; rs7644471; rs13066401; rs35304385; rs28489326; rs7626676	G	0.215	0.043	0.005	1.87 x 10 ⁻¹⁷	[<i>PPP2R3A</i>]	rs13082684 ^A	1.00 x 10 ⁻¹⁶	[<i>PPP2R3A</i>]
chr4p16.3	rs362307		T	0.063	-0.047	0.008	1.91 x 10 ⁻⁸	[<i>HTT</i>]	na	5.71 x 10 ⁻⁷	
chr4q31.3	rs6840360	rs6832872; rs361170; rs6831644	G	0.483	0.029	0.004	6.40 x 10 ⁻¹¹	[<i>PET112</i>]	rs4696277 ^A	4.80 x 10 ⁻¹¹	[<i>FAM160A1</i>]
chr5q34	rs31730	rs2591580	T	0.254	-0.030	0.005	2.47 x 10 ⁻⁹	<i>RPL7P20</i> *--[/--- <i>RP11-67M9.1</i> *	rs41012 ^A	1.34 x 10 ⁻¹⁰	[/]
chr6p22.3	rs4282415	rs10807604	G	0.325	0.026	0.005	3.37 x 10 ⁻⁸	<i>ATXN1</i> ---[/-- <i>STMND1</i>	rs2876430 ^A	3.69 x 10 ⁻⁸	<i>ATXN1</i> ---[/-- <i>STMND1</i>
chr7q11.22	rs3735260		G	0.062	-0.046	0.008	2.90 x 10 ⁻⁸	[<i>AUTS2</i>]	rs3735260 ^A	4.71 x 10 ⁻⁸	[<i>AUTS2</i>]
chr9q34.11	rs9696811	rs4295766; rs10988208; rs10988229	T	0.276	0.036	0.005	3.70 x 10 ⁻¹⁴	<i>PTPA</i> *--[/-- <i>RP11-</i> <i>247A12.7</i> *	rs9696811 ^A	1.00 x 10 ⁻¹⁴	<i>PTPA</i> --[/-- <i>IER5L</i>
chr10q24.2	rs11189513	rs55917128; rs10883026	G	0.306	0.030	0.005	2.97 x 10 ⁻¹⁰	[<i>R3HCC1L</i>]	rs10786387 ^A	1.10 x 10 ⁻¹⁰	<i>CRTAC1</i> --[/-- <i>R3HCC1L</i>
chr11p14.1	rs532395	rs653452; rs564832; rs573434; rs11605215; rs704660; rs4141920	T	0.286	-0.031	0.005	2.01 x 10 ⁻¹⁰	[<i>MPPED2:RP4-</i> <i>710M3.2</i> *]	rs676217 ^A	1.10 x 10 ⁻¹¹	<i>KCNA4</i> ---[/-- <i>FSHB</i>
chr11q23.1	rs10891314	rs55945818	G	0.326	-0.030	0.005	1.31 x 10 ⁻¹⁰	[<i>DLAT</i>]	rs138127836 ^A	1.70 x 10 ⁻¹³	[<i>PPP2R1B</i>]

chr12q24.12	rs1265564	rs7970490; rs10774624; rs6490029	C	0.431	0.031	0.004	2.78 x 10 ⁻¹²	[CUX2]	rs7310615^A	1.10 x 10 ⁻¹⁰	[SH2B3]
chr13q14.2	rs7996852	rs7327520	T	0.310	0.027	0.005	1.84 x 10 ⁻⁸	[SETDB2]	rs7328782 ^B	1.24 x 10 ⁻⁷	
chr13q21.2	rs9538299		G	0.333	-0.026	0.005	3.50 x 10 ⁻⁸	<i>POLR3KPI*--[]--</i> <i>RP11-105A24.2*</i>	rs9538299 ^B	2.63 x 10 ⁻⁷	
chr16p12.13	rs2015573		C	0.338	-0.026	0.005	2.40 x 10 ⁻⁸	<i>CTD-3229J4.1*--[]--</i> <i>CTA-481E9.4*</i>	rs6498749^B	2.51 x 10 ⁻⁷	
chr16q22.2	rs2158266		T	0.171	0.032	0.006	4.56 x 10 ⁻⁸	<i>[AC004158.2]*</i>	rs11646282^B	4.50 x 10 ⁻⁷	
chr17q12	rs1075218	rs11263775; rs9906189; rs62070737	T	0.395	-0.036	0.005	3.82 x 10 ⁻¹⁵	[MRM1]	rs34349354^A	8.20 x 10 ⁻¹⁵	[GGNBP2]
chr17q12	rs12453682	rs1565920; rs57448077	T	0.313	-0.031	0.005	6.48 x 10 ⁻¹¹	<i>NEUROD2-[]--</i> <i>PPP1R1B</i>	rs12453682 ^A	2.90 x 10 ⁻¹²	<i>NEUROD2-[]--</i> <i>PPP1R1B</i>
chr17q23.3	rs72841392	rs8066571	G	0.205	-0.033	0.005	3.44 x 10 ⁻¹⁰	[TANC2]	rs72841395^A	5.40 x 10 ⁻⁹	[TANC2]
chr18q11.2	rs7505854		G	0.435	-0.025	0.004	2.68 x 10 ⁻⁸	<i>[AQP4-</i> <i>AS1*:CHST9]</i>	rs7505854 ^B	3.20 x 10 ⁻⁷	
chr20q11.21	rs4911109	rs10875486; rs293566; rs6087992	C	0.350	0.029	0.005	2.82 x 10 ⁻¹⁰	[DNMT3B]	rs4911257^A	7.50 x 10 ⁻¹⁴	[DNMT3B]
chr20q13.13	rs348258	rs6067019; rs13038202; rs6125596; rs6012558; rs79186842	T	0.291	0.038	0.005	2.02 x 10 ⁻¹⁵	[STAUI]	rs11393101^A	2.20 x 10 ⁻¹⁶	[ARFGEF2]

chrXq27.3	rs6626462		T	0.365	0.027	0.005	1.13 x 10 ⁻⁸	<i>TMEM257</i> ---[]-- <i>AL445258.1</i> *	rs5904158^A	3.30 x 10 ⁻⁹	<i>TMEM257</i> ---[]- -- <i>CXorf51B</i>
-----------	-----------	--	---	-------	-------	-------	-------------------------	---	------------------------------	-------------------------	---

Table 1

Associated region	Lead SNP	P	Number of GWAS sig SNPs	Top SNP gene(s)	Other genes in region	Relevant known associations (SNP: gene)
chr2:193965614-194320930	rs719166	1.82 x 10 ⁻⁸	11	<i>AC092638.2</i>	<i>RP11-305O20.1</i>	Dyslexia (not GWsig) : Dyslexia (not GWsig)
chr4:3142660-3273010	rs362307	1.91 x 10 ⁻⁸	3	<i>HTT</i>	<i>MSANTD1</i>	Educational attainment, general cognitive ability, type 2 diabetes, household income, predicted visceral adipose tissue, automobile speeding propensity, brain morphology, worrying too long after an embarrassing experience, worry/vulnerability (special factor of neuroticism), walking pace, drinks per week, years in education : as for SNPs
chr13:50002976-50101159	rs7996852	1.84 x 10 ⁻⁸	101	<i>SETDB2</i>	<i>CADB39L, PHF11</i>	None : Dyslexia (not GWsig), mean spheric corpuscular volume, mean corpuscular volume, AFP levels
chr13:59485233-59706143	rs9538299	3.50 x 10 ⁻⁸	113	<i>POLR3KP1</i>	<i>RP11-105A24.2</i>	Dyslexia (not GWsig) : Age at first sexual intercourse, externalizing behaviour, chronotype, intelligence, dyslexia (not GWsig)
chr16:17886842-18050926	rs2015573	2.40 x 10 ⁻⁸	55	<i>CTD-3229J4.1</i>	<i>CTA-481E9.4</i>	None : Dyslexia (not GWsig), self-reported math ability, smoking initiation, cognitive empathy, disruptive behaviour, externalizing behaviour
chr16:72333576-72697419	rs2158266	4.56 x 10 ⁻⁸	150	<i>AC004158.2</i>	<i>LINC01572, RNU6-385P, AC092289.1</i>	None : Dyslexia (not Gwsig), intelligence, height, smoking initiation, risk-taking, bmi, age at first sexual intercourse, phoneme awareness, schizophrenia
chr18:24627620-24652744	rs7505854	2.68 x 10 ⁻⁸	14	<i>CHST9</i>	<i>AQP4-AS1</i>	Dyslexia (not GWsig) : Dyslexia (not GWsig)

Table 2

Figures

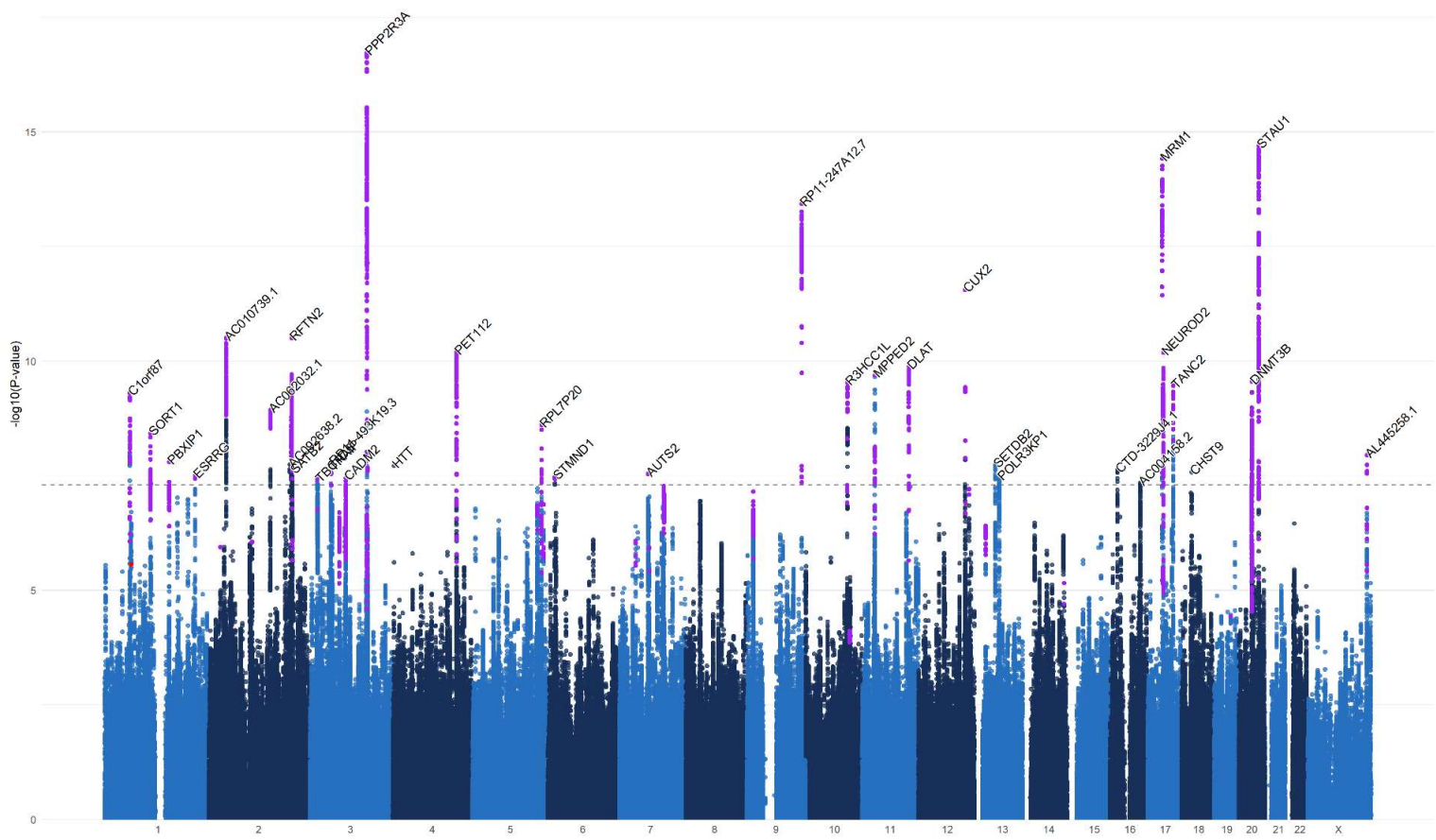


Figure 1a

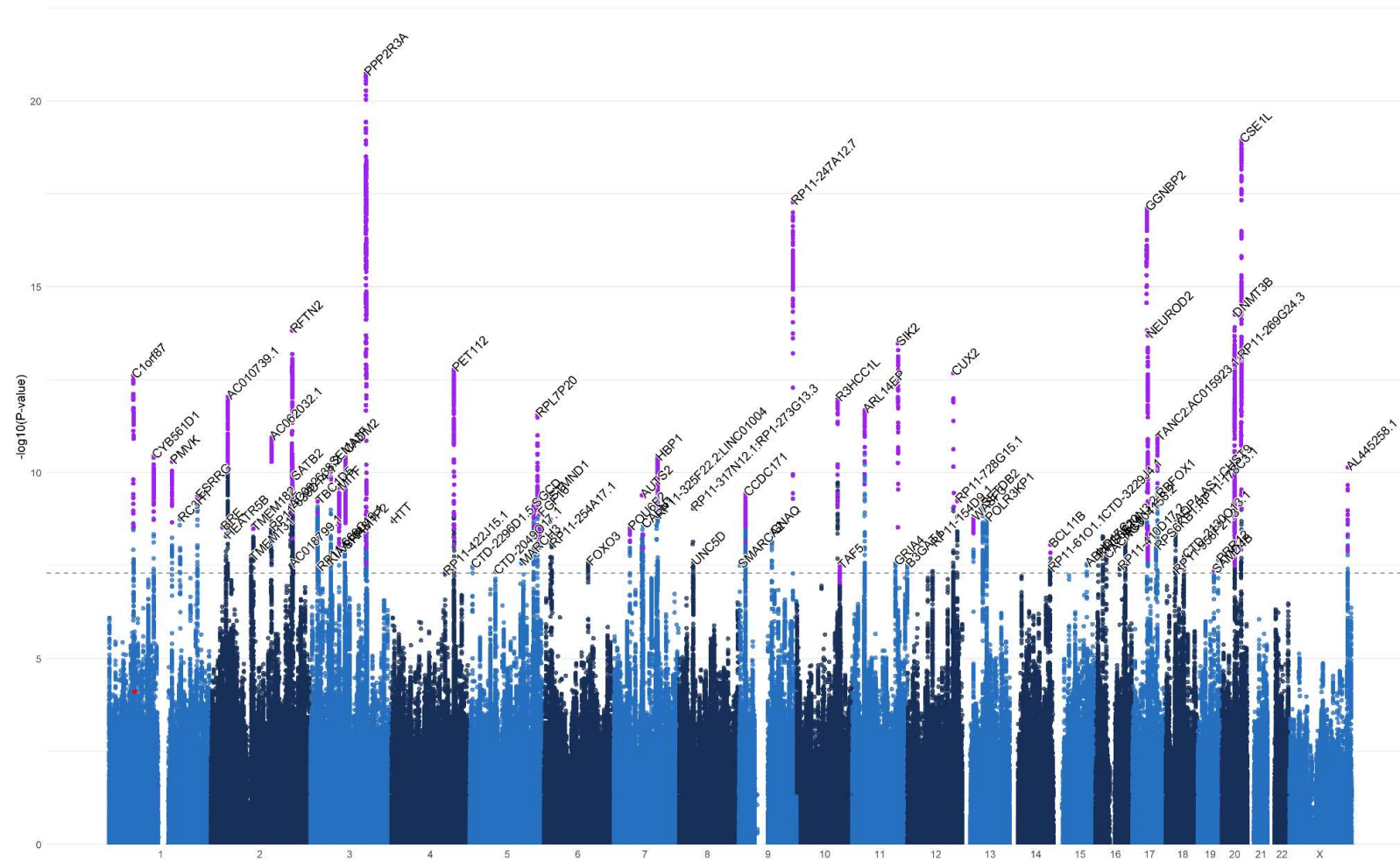


Figure 1b



Figure 2

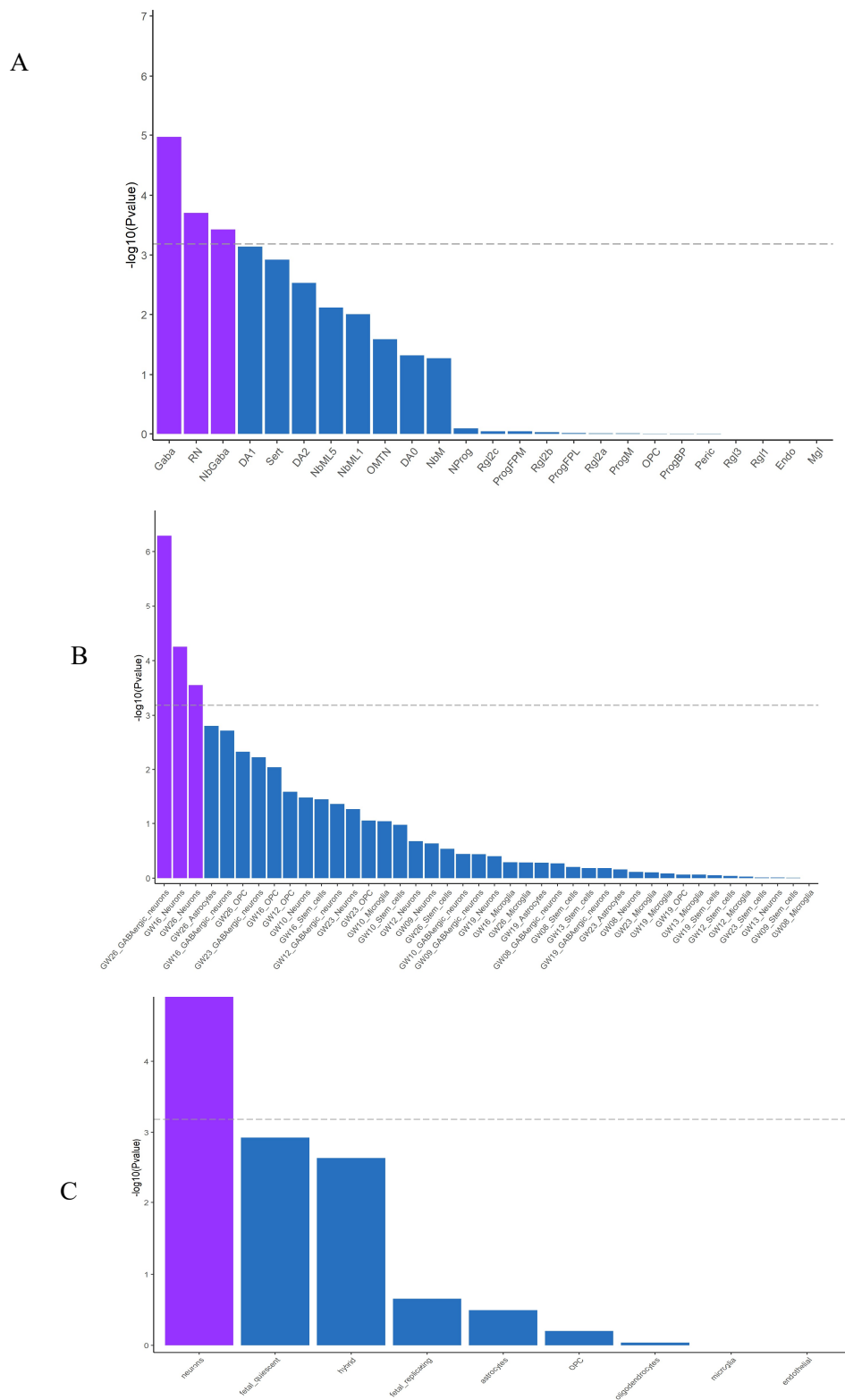


Figure 3

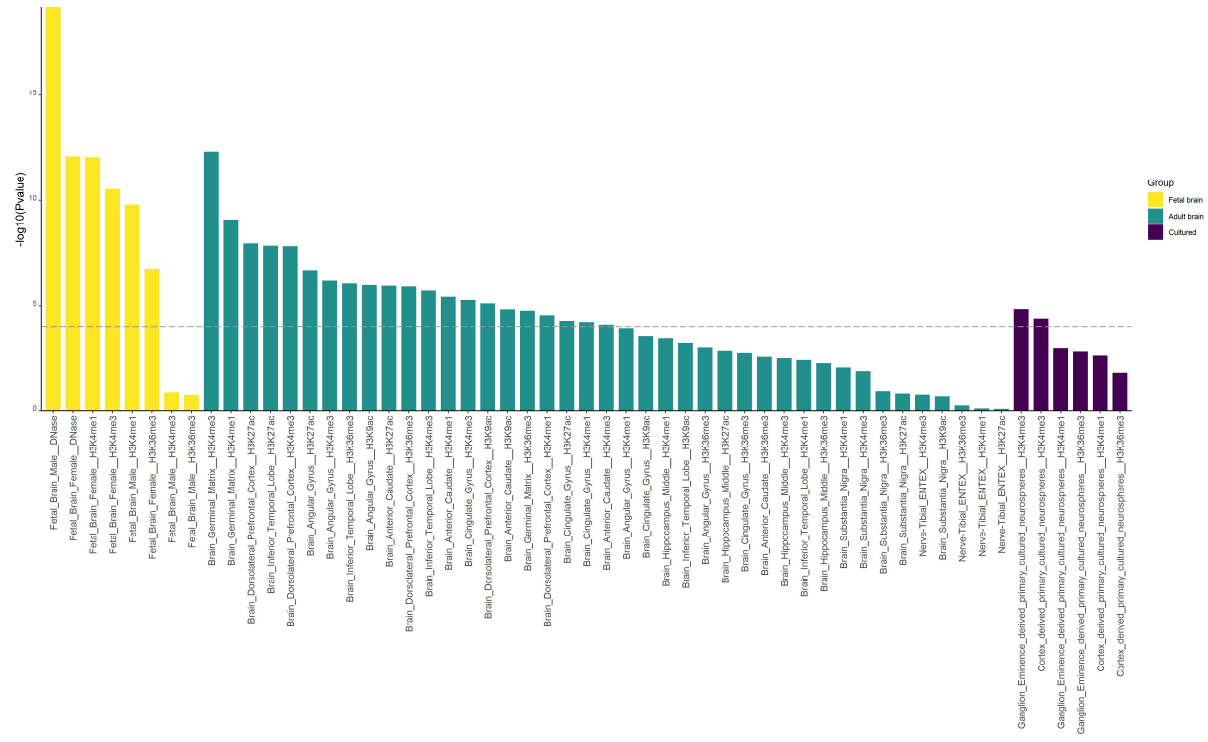


Figure 4

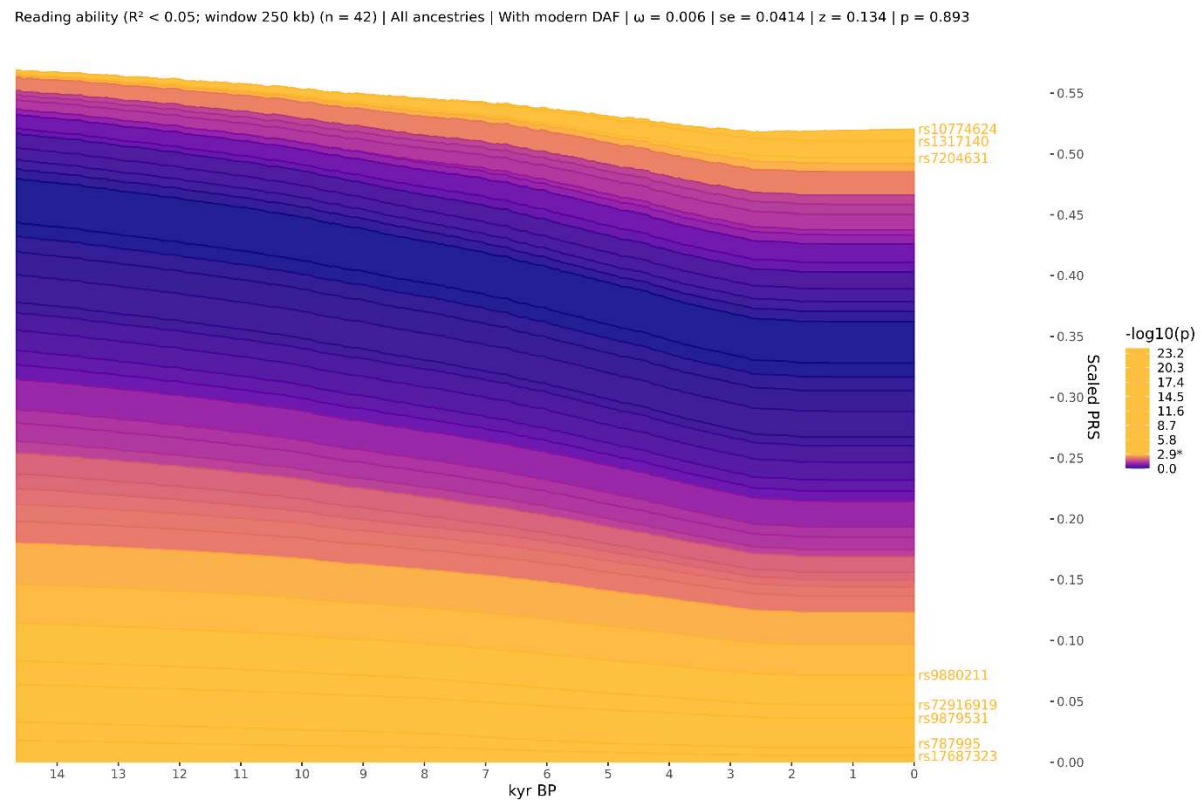


Figure 5

Availability of Data and Materials

The univariate GWAS summary statistics for word reading are available to download from the GenLang website <https://www.genlang.org/downloads.html>. The full GWAS summary statistics for the 23andMe discovery data set will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please visit <https://research.23andme.com/collaborate/#dataset-access/> for more information and to apply to access the data. The multivariate summary statistics for reading ability generated by this study are available through 23andMe to qualified researchers, as described.