Association between genes regulating neural pathways for quantitative traits of speech and language disorders

Penelope Benchek, Ph.D.¹*, Robert P Igo Jr., Ph.D.¹*, Heather Voss-Hoynes, Ph.D.¹*, Yvonne Wren, Ph.D.⁵, Gabrielle Miller, M.S. ², Barbara Truitt, M.S. ¹, Wen Zhang, Ph.D.⁷, Michael Osterman, MPH¹, Lisa Freebairn, M.S. ², Jessica Tag, M.A.², H. Gerry Taylor, Ph.D.³, E. Ricky Chan, Ph.D.¹, Panos Roussos, Ph.D.⁷, Barbara Lewis, Ph.D.², Catherine M. Stein, Ph.D.¹#, Sudha K. Iyengar, Ph.D.¹#

* These authors contributed equally as first authors of this work
# These authors contributed equally as senior authors of this work
¥ Recently deceased

¹ Department of Population & Quantitative Health Sciences, and ² Department of Psychological Sciences, Case Western Reserve University, Cleveland OH; ³ Department of Pediatrics, Case Western Reserve University, and Rainbow Babies & Children’s Hospital, University Hospital Case Medical Center, Cleveland, OH; ⁴ Cleveland Hearing and Speech Center, Cleveland, OH; ⁵ Bristol Dental School, Faculty of Health Sciences, University of Bristol, and Bristol Speech and Language Therapy Research Unit, North Bristol NHS Trust, Bristol, UK; ⁶ Nationwide Children’s Hospital Research Institute and Department of Pediatrics, The Ohio State University, Columbus, OH; ⁷ Department of Psychiatry, Friedman Brain Institute, and Department of Genetics and Genomic Science and Institute for Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY; ⁸ Mental Illness Research, Education, and Clinical Center (VISN 2 South), James J. Peters VA Medical Center, Bronx, NY

Corresponding authors:
Catherine M. Stein, Ph.D. or Sudha K. Iyengar, Ph.D.
Department of Population and Quantitative Health Sciences
Case Western Reserve University
2103 Cornell Rd
Cleveland, OH 44016 USA

Running Title: GWAS of communication disorder endophenotypes

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.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

Speech sound disorders (SSD) manifest as difficulties in phonological memory and awareness, oral motor function, language, vocabulary, reading and spelling. Families enriched for SSD are rare, and typically display a cluster of deficits. We conducted a genome-wide association study (GWAS) in 435 children from 148 families in the Cleveland Family Speech and Reading study (CFSRS), examining 16 variables representing 6 domains. Replication was conducted using the Avon Longitudinal Study of Parents and Children (ALSPAC). We identified 18 significant loci (combined p<10^{-8}) that we pursued bioinformatically. We prioritized 5 novel gene regions with likely functional repercussions on neural pathways, some which colocalized with differentially methylated regions in our sample. Polygenic risk scores for receptive language, expressive vocabulary, phonological awareness, phonological memory, spelling, and reading decoding associated with increasing clinical severity. In summary, neural genetic influence on SSD is primarily multigenic and acts on genomic regulatory elements, similar to other neurodevelopmental disorders.
INTRODUCTION

Communication disorders are highly prevalent in the United States with approximately one in twelve children ages 3-17 years demonstrating a disorder\(^1\). The most common difficulties are a speech problem (5%) or language problem (3.3%). Speech Sound disorders (SSD) include both errors of articulation or phonetic structure (errors due to poor motor abilities associated with the production of speech sounds) and phonological errors (errors in applying linguistic rules to combine sounds to form words). SSD have a prevalence of approximately 16% in children 3 years of age\(^2\), with an estimated 3.8% of children persisting with speech delay at 6 years of age\(^3\). More than half of these children encounter later academic difficulties in language, reading, and spelling\(^7-11\). Because of the relative rarity of persistent speech problems and their correlation with other communication domains, endophenotypes are key to the study of genetic underpinnings.

Vocabulary is core to speech acquisition\(^4\). Children with difficulties in speech sound development often have difficulties with oral language and later reading and spelling disability\(^2,5\).\(^8\). Thus, speech, language, reading, and spelling measures are highly correlated and often have common genetic associations\(^9,10\). Moreover, speech and other communication phenotypes follow a developmental trajectory, where some speech and language disorders resolve with age, whereas others persist; genetic influences on the less easily resolved manifestations are generally stronger\(^11,12\). Because of the common genetic underpinnings and pathologic associations between speech and other communication phenotypes, it is conceivable that genetic replication interweaves with different communication measures. Of 7 known GWASs, none overlap in their top results (at p<5x10\(^{-5}\), see Table 3\(^13\)), because they only focused on a limited number of phenotypes, or these measures were assessed at different ages (either pre-school or early school-age)\(^13-20\), they only present results from one or a few
measures and/or a binary trait; thus, the complexity of shared genetic influences is poorly understood. Most have not focused on children with SSD, particularly measures of articulation. Our sample represents a unique set of deeply phenotyped individuals with information on 6 domains that form the core of speech and language.

SSD are likely due to deficits in both motor ability and broader neural dysfunction. While motor deficits contribute to problems in speech production, abnormalities in other neural systems likely influence formation of phonological representation, which is common to SSD as well as reading and language impairment. We hypothesize that genetic regulation of these neural pathways is associated with variation common to speech, language, reading, and spelling ability. We conducted a GWAS in the Cleveland Family Speech and Reading Study (CFSRS), a cohort ascertained through a proband with SSD. We also conducted a methylome-wide study (i.e. MWAS) to determine the functional implications of these genetic associations, and replicated findings in a population-based cohort. We utilized a family-based cohort as our discovery sample because we hypothesized it would be enriched for disease-associated variants21,22. In these analyses, we identified new candidate genes for correlated communication endophenotypes, and bioinformatic annotation of these loci revealed that regulation of neural pathways is associated with variation in these measures.

SUBJECTS AND METHODS

Subject ascertainment – Cleveland Family Speech and Reading Study

From the Cleveland Family Speech and Reading Study (CFSRS)23-28, we examined 435 individuals from 148 families who had both DNA and endophenotype data available (Table 1). As previously described, families were ascertained through a proband with SSD identified from
caseloads of speech-language pathologists in the Greater Cleveland area and referred to the study; detailed inclusion criteria are provided in the Supplemental Methods. Diagnosis of CAS was confirmed by an experienced licensed speech-language pathologist upon enrollment into the study. Socioeconomic status was determined at the initial assessment based on parent education levels and occupations using the Hollingshead Four Factor Index of Social Class\textsuperscript{29}. This study was approved by the Institutional Review Board of Case Medical Center and University Hospitals and all parents provided informed consent and children older than 5 years provided assent.

**Communication Measures in CFSRS**

We examined diadochokinetic rates using the *Robbins and Klee Oral Speech Motor Control Protocol*\textsuperscript{30} or *Fletcher Time-by-Count Test of Diadochokinetic Syllable Rate*\textsuperscript{31}. The merged variable is referred to as DDK. Expressive vocabulary was assessed with the *Expressive One Word Picture Vocabulary Test-Revised* (EOWPVT\textsuperscript{32}) and receptive vocabulary with the *Peabody Picture Vocabulary Test- Third Edition* (PPVT\textsuperscript{33}), and phonological memory with the *Nonsense Word Repetition (NSW)*\textsuperscript{34}, *Multisyllabic Word Repetition (MSW)*\textsuperscript{34}, and *Rapid Color Naming*\textsuperscript{35} task. In addition to examining the total number of words correct for the MSW and NSW, we also examined the percent phonemes correct for both of these tasks (NSW-PPC and MSW-PPC, respectively). Phonological awareness was assessed using the *Elision* subtest of the *Comprehensive Test of Phonological Processing – 2\textsuperscript{nd} Edition*\textsuperscript{36}. Reading was assessed using the *Woodcock Reading Mastery Test-Revised, Word Attack subtest (WRMT-AT)* and *Word Identification Subtest (WRMT-ID)*, *the Reading Comprehension subtest (WIAT-RC)* and *Listening Comprehension subtest (WIAT-LC)* of the *Wechsler Individual Achievement Test*\textsuperscript{37}. Spelling was assessed on the *Test of Written Spelling-3 (TWS)* using the total score\textsuperscript{38}. Expressive and receptive language were assessed using the *Test of Language Development*
(TOLD\textsuperscript{39}) and \textit{Clinical Evaluation of Language Fundamentals-Revised (CELF\textsuperscript{40})}, referred to as the CELF-E (expressive) and CELF-R (receptive), respectively. Additional details about these measures are provided in the Supplemental Methods. For each of our tests we selected the first available assessment for each individual (Supplemental Table 1).

\textbf{GWAS analysis}

Genotyping methods and quality control (QC) are described in the Supplemental Methods. Principal components (PC) obtained from principal component analysis (PCA) and the genetic relationship matrix (GRM) were generated using genotyped markers that met QC criteria. We used PC-AiR and PC-Relate from the Bioconductor package GENESIS\textsuperscript{41} to generate our PCs and GRM, respectively. PC-AiR accounts for sample relatedness to provide ancestry inference that is not confounded by family structure, while PC-Relate uses the ancestry representative PCs from PC-AiR to provide relatedness estimates due only to recent family (pedigree) structure.

To examine cross-trait correlation, we used GCTA\textsuperscript{42} to run a bivariate REML analysis for each pair of tests and tested for genetic correlations equal to 0. GCTA’s bivariate REML analysis estimates the genetic variance of each test and the genetic covariance between the two tests that can be captured by all SNPs\textsuperscript{43}. Here we included all SNPs with MAF $\geq 0.01$. The genetic variance/covariance calculated was adjusted for sex and the first two PCs.

We used RVTests, version 2.0\textsuperscript{44} to run our GWAS. We specifically relied on RVTest’s Grammar-gamma test\textsuperscript{45}, which performs a linear mixed model association test while allowing for genotype dosages and accounting for family structure using the Genetic Relationship Matrix.
Because each of our tests were age-normed we included only sex and the first two PCs as covariates in our regression models.

In addition, we generated endophenotype-based polygenic risk scores (PRS) in the European subset of the CFSRS where genotype data, as well as clinical group data (no disorder, SSD only, language impairment (LI) only, SSD+LI, CAS) were available. Risk scores were derived from association statistics from our CFSRS GWASs (see GWAS methods section for details) and were constructed using PLINK 1.946 (clump and score functions). Additional details are in the Supplemental Methods. These polygenic risk scores were used to examine the hypothesis that an increase in PRS score would associate with more complex clinical phenotypes when comparing SSD only versus SSD+LI and CAS.

### Statistical analysis of Methylome-wide data

**Methylome-wide association study (MWAS)**

Quality control analysis of methylation data is detailed in Supplemental Methods. We tested for association between CpG beta values and endophenotypes using the linear mixed model approach of GRAMMAR-Gamma45 as implemented in RVtests44. Because our phenotypes were age-normed, we did not adjust for age, but rather for sex and one to four PCs. We also examined methylation-QTLs (meQTL) as described in the Supplemental Methods.

### Replication dataset – ALSPAC

To replicate our GWAS findings, we obtained data from the Avon Longitudinal Study of Parents and Children (ALSPAC). The ALSPAC study was a prospective population-based birth cohort of babies born from > 14,000 pregnancies between April 1991-December 1992, who were followed prospectively with a wide battery of developmental tests, parental questionnaires, child-
completed questionnaires, and health outcomes\textsuperscript{47-49}. Pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 were invited to take part in the study. The study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary). Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and Institutional Review Board of Case Medical Center and University Hospitals. Because this was a birth cohort, all children were included, regardless of diagnosis. We obtained both parental report data on speech development in the children, and also communication measures similar to those that we analyzed (see Communication Measures above and Supplemental Table 3). As this was a longitudinal study, different measures were given at different ages, and when the same domain was tested at two different ages, the identical measure was not used. At some ages, only random subsets were selected, so the sample size available from each age is not the same. In Supplemental Table 4, we list the measures given in the CFSRS battery along with the most similar measure given in ALSPAC.

GWAS in ALSPAC data

QC analyses of ALSPAC data are described in Supplemental Methods. Because of the format of data that were provided, we used slightly different methods for statistical analyses. Genetic association testing was performed using linear regression in Hail 0.1. Covariates adjustments included sex and the first two PCs. Age was not a consideration as ALSPAC is a longitudinal birth cohort study and age differences were negligible for any given measure.

Functional annotation and results integration
In this analysis, we considered CFSRS the discovery sample, since families were ascertained through a child with SSD, and used ALSPAC as the replication sample. We identified associated loci with SNPs significant at $p<10^{-5}$ in CFSRS and $p<0.05$ in ALSPAC, with effects in the same direction.

**Functional annotation**

Because the majority of our findings are intergenic and/or fall in noncoding regions, we relied on annotation tools FUMA and HaploReg to characterize which genes our variants might affect, as well as variants’ functionality. We utilized FUMA for mapping genes to our variants based on genomic proximity, eQTL evidence and chromatin interactions evidence. Default settings in FUMA were used, with the exception of tissue specificity. We hypothesized that gene expression and regulation would be most relevant in brain and neural tissues, as well as muscles related to speech. In FUMA we focused on eQTL and chromatin interaction evidence in our target tissues (brain, muscle and esophagus). HaploReg v.4.1 was used to examine the chromatin state evidence predicting whether the variant fell in a promoter or enhancer region. In HaploReg we focused on chromatin state evidence in our target tissues (brain and muscle).

**Locus prioritization**

In order to further prioritize and synthesize our findings, we annotated associated loci as described above, including annotation of associated effects of these loci in the literature, and incorporated supportive findings from our MWAS. We summarize findings in Table 2, and generated a simple locus priority score as the number of times a locus included an enhancer and/or promoter, included an eQTL, was previously associated with a communication disorder and/or neuropsychiatric disorder, showed eQTL or chromatin state evidence specific to brain
and/or neural tissues, mapped to a gene that was a \textit{FOXP2} target in brain tissue\textsuperscript{51-53}, and a meQTL in that region (at $p < 5 \times 10^{-5}$) with an associated methylation site (at $p < 0.05$) with the same phenotype as the associated GWAS loci. We applied the EpiXcan pipeline\textsuperscript{54} to identify eQTLs with our associated SNPs that are differentially expressed in the dorsolateral prefrontal cortex (DLPFC)\textsuperscript{55} (Supplemental Methods).

\textbf{RESULTS}

The CFSRS sample included 435 subjects from 148 families (Table 1). Of these, 27\% had SSD only, 4\% had LI only, 16\% had SSD+LI without CAS, and 11\% had CAS (Table 1). Of the subjects in the ALSPAC sample, the prevalence of speech problems by parental report varied from 4\%-6\% (Supplemental Table 3).

\textbf{Genetic correlation analysis reveals new relationships among endophenotypes}

Genetic correlation analysis revealed that while many of the patterns of correlation were consistent with phenotypic correlations we have previously reported\textsuperscript{10}, polygenic correlations enable a deeper understanding of these measures, which will inform examination of replication of association effects both within the CFSRS data set and with measures from ALSPAC (Figure 1). For example, while previous studies have demonstrated a strong genetic correlation between reading and spelling measures, polygenic correlation analysis additionally reveals correlations between those skills and Elision. Not surprisingly, expressive and receptive language, as measured on the CELF, are strongly correlated with vocabulary (EOWPVT and PPVT) in addition to reading (WRMT-AT and WRMT-ID). Vocabulary is also strongly correlated with listening comprehension (WIAT-LC).
Most significant findings from GWAS reveal 5 new candidate genes

The majority of associated SNPs (p<10^{-5}) were intergenic, with a lesser number of intronic SNPs (Supplemental Figure 2). Noncoding regions harboring a significant proportion of risk alleles is consistent with previous findings related to neuropsychiatric disease and behavioral traits^{56}. We focused on SNPs that had a p-value<1x10^{-5} in CFSRS with replication with a related trait in ALSPAC (p<0.05), or Fisher combined p-values < 1x10^{-7}, that had functional relevance based on our gene priority score (Table 2).

Among the 5 prominent loci, all had enhancers or promoters for muscle, brain, and/or neuronal progenitor cells, 4 out of 5 had significant methylation and meQTL effects, and 3 out of 5 had eQTLs for brain and/or skeletal-muscle tissue (Figure 2, Supplemental Table 5). EpiXcan analysis suggested that the SNP in the chromosome 1 \textit{IFI6} region is associated with expression in the DLPF cortex (Elision p=0.018, TWS p=0.008; Supplemental Tables 6 and 7). The first region on chromosome 14, including \textit{NFKBIA} and \textit{PPP2R3C}, shows significant chromatin interaction mapping in adult cortex tissue. \textit{NFKBIA}, which codes for a component of the NF-\kappaB pathway, is associated with neurogenesis, neuritogenesis, synaptic plasticity, learning and memory^{57}. The second region on chromosome 14 includes \textit{PP2R3C}, which is within the topologically associating domain (TAD) boundary of the \textit{NFKBIA} locus in Hippocampus and DLPFC. EpiXcan analysis showed \textit{NFKBIZ}, a gene in the same pathway as \textit{NFKIBA}, is also associated with expression in the DLPFC (Elision p=0.000452, TWS p=0.004939; Supplemental Tables 6 and 7).

Replication of previous communication disorder loci

...
A TP2C2 was associated with WRMT-ID (p=7.6x10^{-8}), WRMT-AT (p=4.6x10^{-5}), and Elision (p=4.6x10^{-5}), consistent with prior literature^{58} (Supplemental Figures 3 and 4). Similarly, CYP19A1 was associated with WRMT-AT (p=2.8x10^{-5}), Elision (p=3.3x10^{-4}), and WRMT-ID (p=5.0x10^{-4}), validating a previous association^{59}.CNTNAP2 was associated with CELF-R (p=5.2x10^{-6}), and DDK (p=2.9x10^{-5}), replicating a previous association^{58}. While SNPs within ROBO1 and ROBO2 were not significantly associated with our measures, SNPs in the intergenic region were associated with WRMT-ID (p=3.6x10^{-6}); ROBO1 was originally associated with dyslexia while ROBO2 was originally associated with expressive vocabulary^{20,60}. Finally, SNPs within the DCDC2-KIAA0319-TTRAP and in FOXP2 regions were associated with various traits at p<0.01. Within the ALSPAC cohort, a different pattern of replication emerged (Supplemental Figure 5), with sometimes different SNPs and/or different phenotypes than those associated with CFSRS.

In addition, we examined loci (genes and/or SNPs) associated in recently published GWAS studies of language and reading^{13-20} (Supplemental Table 8); we restricted our examination to the CFSRS data, since the ALSPAC data were included in some of the published studies. In these analyses, we often observed cross-trait replication, with most genes originally associated with dyslexia, and associated with other traits in our sample. These included ZNF385D^{14}, which was associated with all CFSRS traits at p<0.005, CDH13^{19}, associated with all CFSRS traits at p<0.005, GRIN2B^{15}, associated with TWS, EOWPVT, and Elision at P<0.0005 and all CFSRS traits at P<0.05, NKAIN^{15}, associated with CELF-R at 9.7 x 10^{-5} (rs16928927 p=1x10^{-4}) and WIAT-RC (p=4x10^{-4}), and MACROD2^{17} associated with all CFSRS traits at p<0.005).

Polygenic risk scores are associated with increasing clinical severity
In Figure 3, we illustrate polygenic risk scores (PRS) for 6 endophenotypes representing the major domains (receptive language, expressive vocabulary, phonological awareness, phonological memory, spelling, and reading decoding), by quintile, across the clinical subgroups (all endophenotypes are illustrated in Supplemental Figure 6). Generally, we found that polygenic load, indicated by increasing risk scores, was associated with clinical severity \( p < 1 \times 10^{-8} \) by ANOVA, with typical children having the lowest scores, followed by children with SSD-only, and children with SSD+LI and CAS having the greatest scores. The exception to this trend is receptive language, where the genetic load is greatest for children with LI, for whom receptive language is a focal deficit. Thus, in general, an increase in PRS score is associated with greater clinical severity.

DISCUSSION

Communication disorders are genetically complex, manifested by a variety of deficiencies in articulation, vocabulary, receptive and expressive language, phonological awareness, reading decoding and comprehension, and spelling. This GWAS ascertained children through an earlier-presenting clinical disorder and examined several key communication measures, and is thus one of the first studies of its kind. This study is also novel in that it is the first GWAS to include a measure of phonological awareness, as well as a motor speech measure. By analyzing several endophenotypes together, we can draw conclusions about the common genetic basis across these seemingly dissimilar skills. Here, we have identified five new candidate regions, some containing multiple genes, that have connections to neurological function and regulation of neurological pathways. We also found that increased polygenic load is associated with more severe communication disorders. Finally, by examining genetic correlations among these traits, we conclude that different domains of communication have
some common genetic influences. All of these aspects together add new clarity regarding the
genetic underpinnings of speech and language skills.

First, the novel candidate genes that we have identified all have roles in neurological function as
evidenced by expression levels of those genes in brain and/or neural tissue, and associations
with other communication and/or psychiatric phenotypes. This commonality between
communication traits and brain and neural pathways was also demonstrated by a mouse study
of vocalization\textsuperscript{61}, and pleiotropy between brain, learning, and psychiatric phenotypes was
recently demonstrated by a large GWAS of brain phenotypes\textsuperscript{62}. Existence of enhancers,
promoters, and methylation effects in the associated regions further emphasizes the importance
of regulatory effects on these traits. Deletions spanning \textit{SETD3} and \textit{CCNK} have been
associated with syndromic neurodevelopmental disorders\textsuperscript{63} and variants in \textit{SETX}, within this
same family of genes, have been associated with CAS\textsuperscript{64}. In addition, \textit{CCNK} is in the \textit{FOXP2}
pathway in brain tissue\textsuperscript{51-53}. \textit{NFKBIA} is involved in regulation of the NF-\kappa B pathway, which is
involved a number of brain-related processes including neurogenesis, neuritogensis, synaptic
plasticity, learning, and memory\textsuperscript{65}. \textit{PPP2R3C} has been associated with schizophrenia\textsuperscript{66}. \textit{IFI6}
expression has been associated with autism\textsuperscript{67} and overexpression of \textit{IFI6} in the brain is present
in chronic neurodegeneration\textsuperscript{68}. Finally, \textit{DACT1} may be involved in excitatory synapse
organization and dendrite formation during neuronal differentiation\textsuperscript{69} and is mainly expressed
within the first two trimesters of pregnancy, just before the first evidence of speech processing is
observed in preterm neonates\textsuperscript{70}. Interestingly, \textit{SETD3}, \textit{NFKBIA}, and \textit{IFI6} are all also tied to the
immune system, and a recent study identified an excess of T cells in brains of individuals with
autism\textsuperscript{71}. 

\textsuperscript{14}
Second, understanding the genetic architecture across these endophenotypes is essential for understanding how loci are associated with different measures in different study cohorts or across the developmental trajectory. Strong genetic correlations are observed between spelling, reading comprehension and decoding, expressive and receptive language, vocabulary, and phonological awareness. The strongest replications were for a variety of measures collected in CFSRS with ALSPAC from older youth. Consistent with these findings, we previously demonstrated that spelling at later ages has a higher estimated heritability than spelling at school-age\textsuperscript{11}. Measures administered in older youth may also be more sensitive to variations in clinical manifestation of SSD. Examination of the ALSPAC measures suggests that many of those administered at younger ages may have tapped different domains than intended, or may have been less sensitive to later emerging reading and spelling skills. Methods of cohort ascertainment may also be important in comparing our findings to those of other studies. Our families were ascertained through a child with SSD whereas other studies ascertained subjects through LI or dyslexia. These different ascertainment schemes affect both the available measures, as well as the distribution of scores and power to detect association. Since both LI and dyslexia emerge later than SSD, longitudinal studies that ascertain through a proband with SSD will be able to capture variants associated with all three disorders, as there is high comorbidity. In addition to the plethora of studies ascertaining children at a variety of ages, which has an impact on the heritability of traits\textsuperscript{10}, these studies use a wide variety of measures, even for the same endophenotype. Moreover, these studies have been conducted in populations that speak different languages of varying orthographic transparency, which makes them difficult to compare. As noted by Carrion-Castillo et al.\textsuperscript{13}, most of the novel loci identified through GWAS have been unique to each study, and these aforementioned issues may explain that lack of replication. Thus, examination of the genetic correlation matrix is essential for interpretation of results across studies, as it is nearly impossible to analyze the same exact traits, as we have demonstrated with our replication study cohort (ALSPAC).
Third, we replicated candidate genes that had been previously primarily associated with reading and/or language impairment: CNTNAP2, ATP2C2, and CYP19A1. These analyses extend previous findings to show that these genes are associated with articulation (CNTNAP2) and phonological awareness (ATP2C2 and CYP19A1). This further illustrates the pleiotropic nature of these genes. While we did not observe association with SNPs within the coding regions of ROBO1 and ROBO2, we did observe significant associations with SNPs between these two genes, which may have regulatory influences on ROBO1/ROBO2. We also replicated (p<5x10^-3) loci identified in recent GWAS of reading and/or language traits. Similar to another association study between FOXP2 variants and language72, we did not observe statistically significant association between FOXP2 and measures in CFSRS, though there was replication of some traits at a less stringent (p<0.01) level72.

Finally, our analysis of polygenic risk scores shows strong associations between these risk scores and clinical outcomes of increasing severity. Because of the strong significance of these findings, this suggests that the genetic architecture of communication disorders maybe largely polygenic, which may additionally explain the lack of replication and/or genome-wide significance. While other studies have examined polygenic risk scores associated with language15,73, ours is the first to examine polygenic risk associated with other communication endophenotypes. It is noteworthy that our associated SNPs fell outside of gene coding regions but resided in regulatory regions, even having potential regulatory effects themselves. This further illustrates the genetic complexity of communication disorders; perhaps the search for single gene dysfunction is misplaced, and rather regulatory functions are more relevant.
This study has several limitations. The sample size of the CFSRS cohort was modest, potentially reducing power. There was not clear correspondence between measures obtained in ALSPAC with those in CFSRS, necessitating consideration of cross-trait replication. We restricted analyses in both cohorts to individuals of European descent because of low sample size in other ethnic groups, reducing generalizability.

In summary, this first GWAS of communication measures ascertained through families with SSD identified five new candidate genes, all with potential relevance in central nervous system function. Polygenic risk is strongly associated with more severe speech and language outcomes. Careful consideration of genetic correlation among domains of verbal and written language shows that these loci have general effects on communication, not specific to any single domain, suggesting a common genetic architecture. Further research is needed to more closely examine the impact of regulatory variants on these outcomes.

ACKNOWLEDGMENTS

We would like to thank the families who have so generously participated in this study for many years. This research was supported by the Genomics Core Facility of the CWRU School of Medicine’s Genetics and Genome Sciences Department. This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This work was supported by NIH grant R01DC000528 awarded to Dr. Lewis and R01DC012380 awarded to Dr. Iyengar. We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and
nurses. The UK Medical Research Council and Wellcome (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 Dr. Sudha Iyengar will serve as guarantor for the contents of this paper. GWAS data for ALSPAC was generated at the Genotyping Facilities at Wellcome Sanger Institute.

DATA AVAILABILITY

Data from the Cleveland Family Speech and Reading study are not available for broad genetic data sharing because of IRB restrictions. Please contact the corresponding author, Dr. Sudha Iyengar, to request data, which will require an IRB application.

CONFLICT OF INTEREST

The authors have no conflicts of interest to report.

MATERIALS AND CORRESPONDANCE

Please contact Dr. Sudha Iyengar, ski@case.edu, regarding access to summary statistics.
### Table 1. Characteristic table for CFSRS GWAS sample

| Characteristic                  | Value |
|---------------------------------|-------|
| **N**                           | 435   |
| Number of Families              | 148   |
| Age range [2.5, 64]             |       |
| Female N (%)                    | 194 (45%) |
| Speech Disorder Subgroup N (%)  |       |
| CAS                             | 47 (11%) |
| SSD + LI (no CAS)               | 70 (16%) |
| SSD only                        | 119 (27%) |
| Lang only                       | 17 (4%) |
| No CAS/SSD/Lang                 | 177 (41%) |
| Missing                         | 5 (1%) |
| Hollingshead SES                |       |
| 1 (lowest)                      | 3 (1%) |
| 2                               | 30 (7%) |
| 3                               | 67 (15%) |
| 4                               | 167 (38%) |
| 5 (highest)                     | 147 (34%) |
| Missing                         | 21 (5%) |

*Sample considered is the union of all samples across the 16 tests. Specific test sample sizes and age ranges are shown in supplemental Table 1.*
### Table 2. Annotation of most significant loci with replication in CFSRS and ALSPAC

| Locus (Chr location) | Gene(s) | # Associated SNPs | # independently associated SNPs (after conditional analysis) | Gene priority score | Expression in brain/neural tissue | Associated with Communication and/or psych phenotype | Associated with multiple CFSRS traits | Parameter (geophagia, muscle, brain, neural) | Enhancer (geophagia, muscle, brain, neural) | eQTL (geophagia, muscle, brain, neural) | Target of FOXP2 (brain) | Methylation / eQTL |
|----------------------|---------|------------------|---------------------------------------------------------------|--------------------|-----------------------------------|-----------------------------------------------|---------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|-----------------------------------------------|------------------|
| 1:307352871          | LINC01645/MATN1 | 1                 | 1                                                             | 3                  | 1                                 | 1                                             | 1                                      | 0                                           | 1                                           | 0                                           | 0                                           | 0                                             |
| 1:345402735          | LGN, DCX, DCX9 | 5                 | 1                                                             | 4                  | 1                                 | 1                                             | 1                                      | 0                                           | 1                                           | 0                                           | 0                                           | 0                                             |
| 1:1499198760         | LINC000324     | 1                 | 1                                                             | 5                  | 1                                 | 1                                             | 1                                      | 0                                           | 1                                           | 1                                           | 0                                           | 1                                             |
| 1:155928378          | IFIL1, AIM2    | 23                | 1                                                             | 5                  | 0                                 | 1                                             | 0                                      | 0                                           | 1                                           | 1                                           | 1                                           | 1                                             |
| 2:143789805          | LIP1, B2MU     | 4                 | 1                                                             | 4                  | 1                                 | 1                                             | 1                                      | 0                                           | 1                                           | 0                                           | 0                                           | 0                                             |
| 2:165940713          | STK39C, E156   | 4                 | 1                                                             | 3                  | 1                                 | 1                                             | 1                                      | 0                                           | 0                                           | 1                                           | 0                                           | 0                                             |
| 3:15428868            | CNTN6, CNTN4   | 2                 | 1                                                             | 3                  | 1                                 | 1                                             | 1                                      | 0                                           | 0                                           | 0                                           | 0                                           | 0                                             |
| 3:29051236            | MOP, MIP, MP9  | 1                 | 1                                                             | 2                  | 1                                 | 1                                             | 1                                      | 0                                           | 1                                           | 1                                           | 0                                           | 1                                             |
| 4:23279733            | LINC02051, LINC0275 | 9   | 1                                                             | 2                  | 1                                 | 1                                             | 1                                      | 0                                           | 0                                           | 1                                           | 0                                           | 0                                             |
| 4:73572756            | ADAMTS3, COLX18 | 7                | 1                                                             | 4                  | 1                                 | 1                                             | 1                                      | 0                                           | 1                                           | 0                                           | 0                                           | 0                                             |
| 4:7755158             | SHC00M3        | 1                 | 1                                                             | 3                  | 1                                 | 0                                             | 0                                      | 1                                           | 1                                           | 0                                           | 0                                           | 0                                             |
| 5:72144005            | TNPO1          | 1                 | 1                                                             | 4                  | 1                                 | 1                                             | 0                                      | 1                                           | 1                                           | 0                                           | 1                                           | 1                                             |
| 5:13701351            | NF3A           | 1                 | 1                                                             | 4                  | 1                                 | 1                                             | 1                                      | 0                                           | 0                                           | 0                                           | 0                                           | 1                                             |
| 5:170132906           | KCNIP1         | 2                 | 1                                                             | 1                  | 0                                 | 0                                             | 0                                      | 0                                           | 0                                           | 0                                           | 0                                           | 0                                             |
| 5:37520367            | MIR8056, MIR6593 | 15   | 1                                                             | 4                  | 1                                 | 1                                             | 1                                      | 0                                           | 1                                           | 1                                           | 0                                           | 1                                             |
| 7:13630438            | SPAM1          | 10                | 1                                                             | 4                  | 1                                 | 1                                             | 1                                      | 0                                           | 1                                           | 1                                           | 0                                           | 1                                             |
| 7:154705153           | ORP6, PAX1P-AS2 | 1             | 1                                                             | 5                  | 1                                 | 1                                             | 0                                      | 1                                           | 1                                           | 0                                           | 1                                           | 0                                             |
| 8:11438954            | PTGR1, L2N4F   | 0                 | 1                                                             | 6                  | 1                                 | 1                                             | 0                                      | 1                                           | 1                                           | 1                                           | 1                                           | 1                                             |
| 10:64027240           | MARC8R         | 2                 | 1                                                             | 4                  | 1                                 | 1                                             | 0                                      | 0                                           | 1                                           | 1                                           | 1                                           | 1                                             |
| 12:10102703           | SLC18A3        | 2                 | 1                                                             | 1                  | 0                                 | 0                                             | 0                                      | 0                                           | 0                                           | 0                                           | 0                                           | 0                                             |
| 12:120437681**        | ODC1, ODC20056, C12orf6 | 1       | 1                                                             | 3                  | 1                                 | 1                                             | 1                                      | 0                                           | 0                                           | 0                                           | 0                                           | 0                                             |
| 12:21130783           | FAN1, ADE1, F13Q1 | 1  | 1                                                             | 0                  | 1                                 | 0                                             | 0                                      | 0                                           | 0                                           | 0                                           | 0                                           | 0                                             |
| 13:26231049           | POI1, GSK1     | 18                | 1                                                             | 1                  | 0                                 | 0                                             | 0                                      | 1                                           | 1                                           | 0                                           | 1                                           | 0                                             |
| 13:79830132           | LINC00351, RIM26 | 10   | 1                                                             | 2                  | 1                                 | 1                                             | 0                                      | 0                                           | 1                                           | 1                                           | 0                                           | 0                                             |
| 14:35583476            | PSMA6, RFX8A    | 26               | 1                                                             | 7                  | 1                                 | 1                                             | 0                                      | 1                                           | 1                                           | 1                                           | 1                                           | 1                                             |
| 15:45920764           | DAC1, L1N0C9000 | 7                 | 1                                                             | 5                  | 1                                 | 1                                             | 0                                      | 1                                           | 1                                           | 0                                           | 1                                           | 1                                             |
| 14:93195374            | UGMN           | 1                 | 1                                                             | 4                  | 1                                 | 1                                             | 0                                      | 1                                           | 1                                           | 0                                           | 1                                           | 1                                             |
| 14:99598366**         | SEPRAL2, SEPRAL4 | 5     | 1                                                             | 5                  | 0                                 | 1                                             | 1                                      | 1                                           | 1                                           | 1                                           | 0                                           | 1                                             |
| 15:49585970           | BC1L16, SETD3 | 1                 | 1                                                             | 8                  | 1                                 | 1                                             | 1                                      | 1                                           | 1                                           | 1                                           | 1                                           | 1                                             |
| 16:77231207           | MG18           | 1                 | 1                                                             | 7                  | 1                                 | 1                                             | 0                                      | 1                                           | 1                                           | 1                                           | 1                                           | 1                                             |
| 18:4026367            | DUBP1          | 1                 | 1                                                             | 3                  | 1                                 | 1                                             | 0                                      | 1                                           | 1                                           | 0                                           | 1                                           | 1                                             |
| 18:40227933           | R12T, SYT4     | 1                 | 1                                                             | 1                  | 0                                 | 1                                             | 0                                      | 0                                           | 0                                           | 0                                           | 0                                           | 0                                             |
| 18:305620355          | MAL1, LINC01526 | 1             | 1                                                             | 3                  | 0                                 | 1                                             | 0                                      | 1                                           | 1                                           | 1                                           | 0                                           | 1                                             |

*# Associated SNPs include those associated in CFSRS (p < 10^-5) and ALSPAC (p<0.05) or Fisher combined (p < 10^-7)

**CFSRS P = 1.3 * 10^-5 and ALSPAC P = 5.8 * 10^-5 (Fisher P = 1.6 * 10^-8).
FIGURES (attached separately)

Figure 1. Genetic correlation matrix across traits in CFSRS. Figure 1 shows cross-trait correlation results for each pair of tests using GCTA’s bivariate REML analysis. Cross-trait correlation was tested under the null hypothesis of 0 correlation. Circles shown are for results significant at $P<0.05$, with increasing diameter/color corresponding with increasing correlation (circles omitted otherwise).

Figure 2. Locus zoom plots for most significant findings. Figure 2 shows association results for the top loci. P-values displayed are for CFSRS and are for the test for which the top SNP was observed. Circles show P-values for SNP associations and triangles show P-values for methylation associations (specifically those for which the top SNP is an meQTL). The larger plot shows the top SNP for each region +/- 200 kb. The window highlights the region that spans significant association results ($P\leq1\times10^{-5}$ in CFSRS. A. IFI16 region (window spans chr1:159001292-159028378) rs855865 was associated with NSW in CFSRS ($p=7\times10^{-6}$) and with vocabulary (WISC-V) in ALSPAC ($p=0.01$). This region also includes an meQTL (rs12124059, $p=4\times10^{-8}$) for methylation marker cg07196514, and this methylation marker was also associated with NSW ($p=0.018$). B. NFKBIA region (window spans chr14:35770806-35846092). rs57645874 was associated with Elision in CFSRS ($p=1\times10^{-6}$) and with reading accuracy (NARA-A) in ALSPAC ($p=0.02$). This region also contains an meQTL, rs4981288, for cg07166546 ($p=2\times10^{-50}$), and this methylation marker was associated with Elision ($p=3\times10^{-5}$), TWS ($p=0.0005$) and WRMT-ID ($p=0.002$). C. DACT1 region (window spans chr14:59210335-59221002). rs856379 was associated with MSW in CFSRS ($p=3\times10^{-6}$) and with nonword reading (ALSPACread) in ALSPAC ($p=0.036$). This SNP is an meQTL for methylation marker cg13972423 ($p=3\times10^{-5}$). D. SETD3 region (window spans chr14:99858970-99942692). rs1257267 was associated with WRMT-AT in CFSRS ($p=6.58\times10^{-6}$) and with nonsense word
repetition (CNrep5) in ALSPAC (p=0.05). While only 1 SNP replicated between CFSRS and
ALSPAC, 14 additional SNPs showed association in CFSRS at p<10^{-5}. This SNP is an meQTL
for cg18949721 (p=4x10^{-12}), which was also associated with WRMT-AT (p=0.003). E. MON1B
region (window spans chr16:77231207-77248555). rs4888606 was associated with MSW in
CFSRS (p=9 \times 10^{-6}) and with nonword reading (ALSPACread) in ALSPAC (p=0.046). While
only 1 SNP replicated between CFSRS and ALSPAC, 18 additional SNPs showed association
in CFSRS at p<10^{-5}. This SNP falls in an intron of MON1B and is an meQTL for cg06128999
(p=4x10^{-23}) and cg05007098 (p=1x10^{-15}), which were also associated with MSW (p=0.045 and
p=0.12, respectively). Functional annotation is in Supplemental Figure 2.
Figure 3. Polygenic risk scores across major domains. We constructed polygenic risk scores for 587 individuals who were both genotyped and had clinical subgroup information available. Polygenic risk scores are displayed by quantile across the clinical subgroups for six endophenotypes representing the major domains (A Receptive language; B Expressive vocabulary; C Phonological awareness; D Phonological memory; E Spelling; F Reading decoding).

DESCRIPTION OF SUPPLEMENTAL DATA

Supplemental Methods: Describes behavioral phenotypes in detail and detailed methods for genetic methylation analysis

Supplemental Tables and Figures:

Supplemental Table 1. Descriptive statistics for CFSRS measures
Supplemental Table 2. Results of methylation analysis of candidate gene regions
Supplemental Table 3. Descriptive statistics for ALSPAC sample
Supplemental Table 4. Correspondence between CFSRS and ALSPAC measures
Supplemental Table 5. Annotation of functional implications of most significant loci from GWAS
Supplemental Table 6. PsychEncode EpiXcan method using Meta-analysis results of Elision GWAS
Supplemental Table 7. PsychEncode EpiXcan method using Meta-analysis results of TWS GWAS
Supplemental Table 6 – Association results from regions identified from published GWAS of reading and language phenotypes

Supplemental Figure 1. Distribution of associated SNPs

Supplemental Figure 2. Functional annotation corresponding to Figure 3.

Supplemental Figure 3. Clustering of Significant Variants (P < 0.01) among Known Speech Genes across CFSRS Tests

Supplemental Figure 4. LocusZoom plots of candidate genes where at least one trait had a SNP significant at p < 10^{-4}

Supplemental Figure 5. Clustering of Significant Variants (P < 0.01) among Known Speech Genes across ALSPAC Tests

Supplemental Figure 6. Polygenic Risk score across all individual measures

REFERENCES

1 Almost 8 percent of US children have a communication or swallowing disorder, 2015).
2 Catts, H. W., Adlof, S. M., Hogan, T. P. & Weismer, S. E. Are specific language impairment and dyslexia distinct disorders? Journal of speech, language, and hearing research : JSLHR 48, 1378-1396, doi:10.1044/1092-4388(2005/096) (2005).
3 Shriberg, L., Tomblin, J. & McSweeny, J. Prevalence of speech delay in 6-year-old children and comorbidity with language impairment. Journal of Speech, Language, and Hearing Research 42, 1461-1481 (1999).
4 McLeod, S. B., E. in Children's speech: an Evidence-based approach to assessment and intervention (ed S. McLeod, Baker, E) 181-184 (Pearson Education, 2017).
5 Lemons, C. J. & Fuchs, D. Phonological awareness of children with Down syndrome: its role in learning to read and the effectiveness of related interventions. Research in developmental disabilities 31, 316-330, doi:10.1016/j.ridd.2009.11.002 (2010).
6 Al Otaiba, S., Puranik, C., Zilkowski, R. & Curran, T. Effectiveness of Early Phonological Awareness Interventions for Students with Speech or Language Impairments. The Journal of special education 43, 107-128, doi:10.1177/0022466908314869 (2009).
7 Larivee, L. C., HW. Early reading achievement in children with expressive phonological disorders.  
*Am J Speech Lang Pathol* 8, 119-128 (1999).

8 Scarborough, H. in *Specific Reading Disabilities: A view of the spectrum* (ed BK; Accardo Shapiro, 
P; Capute, AJ) 75-119 (York Press, 1990).

9 Lewis, B. A. *et al.* The Genetic Bases of Speech Sound Disorders: Evidence From Spoken and 
Written Language. *J Speech Lang Hear. Res* 49, 1294-1312 (2006).

10 Stein, C. M. *et al.* Pleiotropic effects of a chromosome 3 locus on speech-sound disorder and 
reading. *Am J Hum Genet* 74, 283-297 (2004).

11 Lewis, B. A. *et al.* Heritability and longitudinal outcomes of spelling skills in individuals with 
histories of early speech and language disorders. *Learning and individual differences* 65, 1-11, 
doi:10.1016/j.lindif.2018.05.001 (2018).

12 Stevenson, J., Graham, P., Fredman, G. & McLoughlin, V. A twin study of genetic influences on 
reading and spelling ability and disability. *Journal of child psychology and psychiatry, and allied 
disciplines* 28, 229-247, doi:10.1111/j.1469-7610.1987.tb00207.x (1987).

13 Carrion-Castillo, A. *et al.* Evaluation of results from genome-wide studies of language and 
reading in a novel independent dataset. *Genes Brain Behav* 15, 531-541, doi:10.1111/gbb.12299 
(2016).

14 Eicher, J. D. *et al.* Genome-wide association study of shared components of reading disability 
and language impairment. *Genes Brain Behav* 12, 792-801, doi:10.1111/gbb.12085 (2013).

15 Gialluisi, A. *et al.* Genome-wide association scan identifies new variants associated with a 
cognitive predictor of dyslexia. *Translational psychiatry* 9, 77, doi:10.1038/s41398-019-0402-0 
(2019).

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

17 Harlaar, N. *et al.* Genome-wide association study of receptive language ability of 12-year-olds. *J 
Speech Lang Hear Res* 57, 96-105, doi:10.1044/1092-4388(2013/12-0303) (2014).

18 Kornilov, S. A. *et al.* Genome-Wide Association and Exome Sequencing Study of Language 
Disorder in an Isolated Population. *Pediatrics* 137, doi:10.1542/peds.2015-2469 (2016).

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

20 St Pourcain, B. *et al.* Common variation near ROBO2 is associated with expressive vocabulary in 
infancy. *Nature communications* 5, 4831, doi:10.1038/ncomms5831 (2014).

21 Morris, N., Elston, R. C., Barnholtz-Sloan, J. S. & Sun, X. Novel approaches to the analysis of 
family data in genetic epidemiology. *Front Genet* 6, 27, doi:10.3389/fgene.2015.00027 (2015).

22 Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. 
*Nat Rev Genet* 12, 465-474, doi:10.1038/nrg2989 (2011).

23 Lewis, B. & Freebairn, L. Speech production skills of nuclear family members of children with 
phonology disorders. *Speech and Language* 41, 45-61 (1998).

24 Lewis, B., Freebairn, L. & Taylor, H. Follow-up of children with early expressive phonology 
disorders. *Journal of Learning Disabilities* 33, 433-444 (2000).

25 Lewis, B. A. *et al.* Literacy outcomes of children with early childhood speech sound disorders: 
impact of endophenotypes. *J Speech Lang Hear. Res* 54, 1628-1643 (2011).

26 Lewis, B. A. *et al.* Family pedigrees of children with suspected childhood apraxia of speech. 
*Journal of Communication Disorders* 37, 157-175 (2004).

27 Lewis, B. A., Freebairn, L. A., Hansen, A. J., Iyengar, S. K. & Taylor, H. G. School-age follow-up of 
children with childhood apraxia of speech. *Language, speech, and hearing services in schools* 35, 
122-140 (2004).
Lewis, B. A. et al. Speech and language skills of parents of children with speech sound disorders.
Am J Speech Lang Pathol 16, 108-118 (2007).

Hollingshead, A. (Department of Sociology, Yale University, New Haven, CT. 06520, 1975).

Robbins, J. & Klee, T. Clinical assessment of oropharyngeal motor development in young children. Journal of Speech and Hearing Research 52, 271-277 (1987).

Fletcher, D. (C.C. Publications, Inc., Tigard, OR, 1977).

Gardner, M. (Academic Therapy Publications, Novato, CA, 1990).

Dunn, L. & Dunn, L. (American Guidance Service, Inc, Circle Pines, MN, 1997).

Catts, H. Speech production/phonological deficits in reading disordered children. Journal of Learning Disabilities 19, 504-508 (1986).

Denkla, M. & Rudel, R. Rapid 'automatized' naming (R.A.N.): dyslexia differentiated from other learning disabilities. Neuropsychologia, 471-479 (1976).

Wagner, R. T.; J; Rashotte, C; Pearson, NA. (Pearson, London, England, 2013).

Wechsler, D. (The Psychological Coporation, San Antonio, TX, 1991).

Larsen, S. H., D. (The Psychological Corporation, San Antonio, TX, 1994).

Newcomer, P. & Hammill, D. Test of language development - Primary, Second Edition. (Pro-Ed., 1988).

E, S., Wiig, E. & Secord, W. Clinical evaluation of language fundamentals-Revised. (The Psychological Corporation, 1987).

GENESIS: GENetic ESTimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. R package version (2019).

Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum Genet 88, 76-82 (2011).

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics (Oxford, England) 28, 2540-2542, doi:10.1093/bioinformatics/bts474 (2012).

Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics (Oxford, England) 32, 1423-1426, doi:10.1093/bioinformatics/btw079 (2016).

Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. Nature genetics 44, 1166-1170, doi:10.1038/ng.2410 (2012).

Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575 (2007).

Fraser, A. et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. International journal of epidemiology 42, 97-110, doi:10.1093/ije/dys066 (2013).

Golding, J., Pembrey, M. & Jones, R. ALSPAC—the Avon Longitudinal Study of Parents and Children. I. Study methodology. Paediatric and perinatal epidemiology 15, 74-87, doi:10.1046/j.1365-3016.2001.00325.x (2001).

Boyd, A. et al. Cohort Profile: the 'children of the 90s'–the index offspring of the Avon Longitudinal Study of Parents and Children. International journal of epidemiology 42, 111-127, doi:10.1093/ije/dys064 (2013).

Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. Nature communications 8, 1826-1826, doi:10.1038/s41467-017-01261-5 (2017).
51 MacDermot, K. et al. Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits. *Am J Hum Genet* **76**, 1074-1080 (2005).

52 Spiteri, E. et al. Identification of the transcriptional targets of FOXP2, a gene linked to speech and language, in developing human brain. *American journal of human genetics* **81**, 1144-1157, doi:10.1086/522237 (2007).

53 Vernes, S. C. et al. High-throughput analysis of promoter occupancy reveals direct neural targets of FOXP2, a gene mutated in speech and language disorders. *American journal of human genetics* **81**, 1232-1250, doi:10.1086/522238 (2007).

54 Zhang, W. et al. Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nature communications* **10**, 3834, doi:10.1038/s41467-019-11874-7 (2019).

55 Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, doi:10.1126/science.aat8464 (2018).

56 Goriounova, N. A. & Mansvelder, H. D. Genes, Cells and Brain Areas of Intelligence. *Front Hum Neurosci* **13**, 44-44, doi:10.3389/fnhum.2019.00044 (2019).

57 Zhang, Y. & Hu, W. NFκB signaling regulates embryonic and adult neurogenesis. *Front Biol (Beijing)* **7**, 10.1007/s11515-11012-11233-z, doi:10.1007/s11515-012-1233-z (2012).

58 Newbury, D. F. & Monaco, A. P. Genetic advances in the study of speech and language disorders. *Neuron* **68**, 309-320 (2010).

59 Anthoni, H. et al. The aromatase gene CYP19A1: several genetic and functional lines of evidence supporting a role in reading, speech and language. *Behav Genet* **42**, 509-527, doi:10.1007/s10519-012-9532-3 (2012).

60 Hannula-Jouppi, K. et al. The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia. *PloS Genet* **1**, e50 (2005).

61 Ashbrook, D. G. et al. Born to Cry: A Genetic Dissection of Infant Vocalization. *Front Behav Neurosci* **12**, 250-250, doi:10.3389/fnbeh.2018.00250 (2018).

62 Zhao, B. et al. Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nature genetics* **51**, 1637-1644, doi:10.1038/s41588-019-0516-6 (2019).

63 Fan, Y. et al. De Novo Mutations of CCNK Cause a Syndromic Neurodevelopmental Disorder with Distinctive Facial Dysmorphism. *American journal of human genetics* **103**, 448-455, doi:10.1016/j.ajhg.2018.07.019 (2018).

64 Worthey, E. A. et al. Whole-exome sequencing supports genetic heterogeneity in childhood apraxia of speech. *Journal of neurodevelopmental disorders* **5**, 29, doi:10.1186/1866-1955-5-29 (2013).

65 Lanzillotta, A. et al. NF-κB in Innate Neuroprotection and Age-Related Neurodegenerative Diseases. *Front Neurol* **6**, 98-98, doi:10.3389/fneur.2015.00098 (2015).

66 Gusev, A. et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature genetics* **50**, 538-548, doi:10.1038/s41588-018-0092-1 (2018).

67 El-Ansary, A. & Al-Ayadhi, L. GABAergic/glutamatergic imbalance relative to excessive neuroinflammation in autism spectrum disorders. *J Neuroinflammation* **11**, 189-189, doi:10.1186/s12974-014-0189-0 (2014).

68 Nazmi, A. et al. Chronic neurodegeneration induces type I interferon synthesis via STING, shaping microglial phenotype and accelerating disease progression. *Glia* **67**, 1254-1276, doi:10.1002/glia.23592 (2019).
Okerlund, N. D. et al. Dact1 is a postsynaptic protein required for dendrite, spine, and excitatory synapse development in the mouse forebrain. J Neurosci 30, 4362-4368, doi:10.1523/JNEUROSCI.0354-10.2010 (2010).

Le Guen, Y. et al. A DACT1 enhancer modulates brain asymmetric temporal regions involved in language processing. bioRxiv, 539189, doi:10.1101/539189 (2019).

DiStasio, M. M., Nagakura, I., Nadler, M. J. & Anderson, M. P. T lymphocytes and cytotoxic astrocyte blebs correlate across autism brains. Ann Neurol 86, 885-898, doi:10.1002/ana.25610 (2019).

Mueller, K. L. et al. Common Genetic Variants in FOXP2 Are Not Associated with Individual Differences in Language Development. PLoS One 11, e0152576, doi:10.1371/journal.pone.0152576 (2016).

Nudel, R. et al. Language deficits in specific language impairment, attention deficit/hyperactivity disorder, and autism spectrum disorder: An analysis of polygenic risk. Autism research: official journal of the International Society for Autism Research, doi:10.1002/aur.2211 (2019).
