

1 Whole-exome sequencing in children with dyslexia

2 identifies rare variants in *CLDN3* and ion channel genes

3

4 Krzysztof Marianski¹, Joel B. Talcott², John Stein³, Anthony P. Monaco⁴, Simon E. Fisher^{5,6},
5 Dorothy V.M. Bishop⁷, Dianne F. Newbury^{8,9} and Silvia Paracchini¹

6

7 ¹*School of Medicine, University of St Andrews, St Andrews, UK*

8 ²*Institute of Health and Neurodevelopment, College of Health and Life Sciences, Aston
9 University, Birmingham, UK*

10 ³*Department of Physiology, University of Oxford, Oxford, UK*

11 ⁴*Office of the President emeritus, Tufts University, Medford, MA, USA*

12 ⁵*Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen,
13 Netherlands*

14 ⁶*Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen,
15 Netherlands*

16 ⁷*Department of Experimental Psychology, University of Oxford, Oxford, UK*

17 ⁸*Department of Medical and Biological Sciences, Oxford Brookes University, UK*

18 ⁹*Centre for Human Genetics, University of Oxford, Oxford, UK*

19

20

21 Correspondence to Silvia Paracchini sp58@st-andrews.ac.uk

22

23

24

25

26

27

28

29

30

31 **Abstract**

32 Dyslexia is a specific difficulty in learning to read that affects 5-10% of school-aged children
33 and is strongly influenced by genetic factors. While previous studies have identified common
34 genetic variants associated with dyslexia, the role of rare variants has only recently begun to
35 emerge from pedigree studies and has yet to be systematically tested in larger cohorts. Here,
36 we present a whole-exome sequencing (WES) study of 53 individuals with dyslexia, followed
37 by replication analysis in 38 cases with reading difficulties and 82 controls assessed with
38 reading measures. Our stringent bioinformatics filtering strategy highlighted five brain-
39 expressed genes carrying rare variants: *CACNA1D*, *CACNA1G*, *CLDN3*, *CNGB1*, and *CP*.
40 Notably, a specific variant (7-73769649-G-A) in the *CLDN3* gene was identified in six
41 independent cases, showing a four-fold higher frequency compared to population reference
42 datasets. *CACNA1D* and *CACNA1G* encode subunits of voltage-gated calcium channels
43 (VGCC) expressed in neurons, and variants in both genes have been implicated in
44 neurodevelopmental disorders such as autism spectrum disorder (ASD) and epilepsy.
45 Segregation analysis in available family members were consistent with patterns of dominant
46 inheritance with variable expressivity. In total, high-impact variants in the five genes of
47 interest were found in 26% (N = 14) of individuals of the discovery cohort. Overall, our
48 findings support the involvement of rare variants in developmental dyslexia and indicate
49 that larger WES studies may uncover additional associated genes.

50

51 **Introduction**

52 Dyslexia is a neurodevelopmental condition characterised by a specific difficulty in learning
53 to read, occurring in the absence of other causes such as sensory or neurological problems
54 or lack of educational opportunity (1). Regardless of culture and spoken-language, dyslexia
55 affects 5%–10% of children, with a male: female ratio ranging from about 3:1 to 5:1 (2), and
56 often co-occurs with other neurodevelopmental conditions such as developmental language
57 disorders (DLD) and attention deficit hyperactivity disorder (ADHD) (3,4). A strong genetic
58 component for dyslexia is supported by twin studies, which have shown heritability

59 estimates of up to 70%. Moreover, having a first degree relative with dyslexia is the most
60 consistently identified risk factor (5). These observations have motivated molecular studies,
61 but the identification of specific genes has, until recently, been hindered by difficulties in
62 assembling sufficiently powered sample sizes. Lack of national screening programs and
63 heterogeneous assessment and diagnostic criteria pose challenges in recruiting study
64 participants and combining them across studies (1). Initial molecular studies highlighted a
65 handful of genes, including *ROBO1*, *DCDC2*, *KIAA0319* and *DYX1C1* (6). However,
66 associations to these genes did not consistently replicate (1).
67 A breakthrough was made possible by a large (51,800 cases and 1,087,070 controls)
68 genome-wide association study (GWAS) conducted with self-reported dyslexia diagnoses by
69 customers of the direct-to-consumer company 23andMe (7). This study identified 42
70 independently statistically significant associated loci. About half of these had previously
71 been associated with cognitive phenotypes, while the other half suggested effects more
72 specific to dyslexia. A gene-based analysis reported association with 173 genes. This study
73 also facilitated the generation of dyslexia polygenic scores, which showed significant
74 associations with the reading-related measures available for some cohorts of the GenLang
75 consortium (<https://www.genlang.org/>), an international project aimed at dissecting the
76 genetics of language-related traits, including reading abilities. An independent quantitative
77 meta-GWAS of reading- and language-related measures in the GenLang cohorts (N ~34,000)
78 identified only a single genome-wide significant locus but highlighted significant genetic
79 correlations across language-related cognitive domains and traits derived from
80 neuroimaging data (8). The SNP-based heritability from these GWAS efforts ranged between
81 0.13 and 0.26. Although larger GWAS are expected to identify additional common variants
82 associated with dyslexia, it is likely that rare variants, *e.g.* those with a minor allele
83 frequency (MAF) < 1%, will also play a role.
84 Whole-exome (WES) and whole-genome sequencing (WGS) studies have demonstrated the
85 role of rare variants in most complex traits (9) including for neurodevelopmental conditions
86 like autism spectrum disorder (ASD) (10), schizophrenia (11), bipolar disorder (12) and
87 childhood apraxia of speech (CAS) (13–15). Rare variants have been reported also for
88 conditions that frequently co-occur with dyslexia, such as ADHD (16) and language

89 impairment (17–19). Rare variant analyses and WES screenings focussing specifically on
90 dyslexia are limited. *DYX1C1*, the first risk gene reported as a dyslexia candidate was found
91 to be disrupted by a translocation in one individual (20). Segregation analyses in large
92 pedigrees suggested the role of variants in the *CEP63* (21), *NCAN* (22), and *SEMA3C* (23)
93 genes. Large population-based cohorts like the UK Biobank, for which WES data have been
94 generated at scale, are not well-suited for genetic studies of dyslexia because data on dyslexia
95 diagnosis are not reliable and reading-related measures have not been collected (1).

96 Furthermore, it has been shown that rare variants detected in clinical cohorts might not
97 show signals in population samples because of incomplete penetrance (*i.e.* when not all
98 individuals with the variant exhibit the trait) or variable expressivity (when different degrees
99 of a phenotype are associated with the variant) (24).

100 In the current proof-of-principle WES study, we analysed 53 unrelated individuals with
101 dyslexia. Replication analysis was performed in an independent cohort of 38 cases with
102 reading difficulties (RD) and 82 typically developing (TD) controls, assessed with the same
103 quantitative test as part of a twin study on language-related problems and co-occurring
104 conditions. Our findings highlight rare variants in five genes *CACNA1G*, *CACNA1D*, *CLDN3*,
105 *CNGB1*, and *CP*. Notably, a specific *CLDN3* variant was identified in six independent cases
106 and was not observed in the controls. The availability of family members and quantitative
107 assessments facilitated segregation analysis and enabled the evaluation of the impact of
108 variants on reading-related measures.

109

110 **Material and Methods**

111 **Discovery cohort**

112 We analysed a discovery sample of 53 unrelated cases selected from two existing cohorts
113 recruited for genetic studies of dyslexia and which have been previously described (Table 1)
114 (25). The first cohort includes 290 families (689 siblings, 580 parents), most with at least one
115 child diagnosed with dyslexia and a second child presenting reading difficulties. The second
116 cohort comprises 592 singletons with dyslexia or reading problems. Both cohorts were
117 recruited in the South of England at Dyslexia Research Centre clinics in Oxford and Reading
118 or the Aston Dyslexia and Development Unit in Birmingham. All children (proband and their

119 siblings) were individually assessed with the following core tasks: single-word reading (SWR)
120 and single-word spelling accuracy (SPELL)(26), single non-word reading (NWR) (27),
121 phonological awareness (PA) (28), irregular word reading (IWR) (27), orthographic coding
122 (ORTH) (29), as well as measures of verbal (VIQ) and performance (PIQ) IQ (30) (See
123 supplementary Table 1). Parents in the family cohort were not formally assessed but some of
124 them provided information about their own and their family's history of dyslexia or reading
125 problems by questionnaire during an interview. In both cohorts, probands were excluded if
126 they had been formally diagnosed with co-occurring developmental conditions such as
127 language impairment, autism or ADHD.

128 For this study, individual cases were prioritised based on low SWR scores and availability of a
129 high-quality DNA sample. In total, 53 unrelated cases were selected, 33 from the family
130 cohort and 20 from the singleton cohort. WES data were also generated for available family
131 members (N = 14) of variant carriers in the top five genes for segregation analysis.

132 The study was approved by the Oxfordshire Psychiatric Research Ethics Committee and the
133 Aston University Ethics Committee, with informed consent obtained from all the participants
134 and/or their caregivers.

135

136 **Replication and control cohorts**

137 The replication and control cohorts were derived from a UK twin cohort (N = 194 twin pairs)
138 recruited to study language difficulties (31). From this cohort, replication cases and controls
139 were selected based on the following criteria.

140 Replication cases with reading difficulties (RD) were required to have a PIQ above or equal to
141 -1 SD and at least one score more than 1 SD below the expected level on any of the
142 following tests: single word reading (SWR), non-word reading (NWR) (32), text reading
143 accuracy (ACC), comprehension (COMP) or reading rate (RR) from the NARA-II (33). See
144 Supplementary Table 1 for more details on the tests.

145 Typically developing (TD) controls were selected based on a PIQ of no more than 1 SD below
146 the expected level for their age and a SWR score above the mean. Children with a
147 Development and Well-Being Assessment (DAWBA) diagnosis of ASD (34) were removed
148 from both groups.

149 If both twins in a pair met the criteria for inclusion, one twin was randomly selected for the
150 study. This process resulted in a replication group of 38 cases with RD and 82 TD controls
151 with good reading scores. Descriptive statistics and mean phenotypic scores for these two
152 groups are shown in Table 1. WES and phenotypic data were available for their twin siblings
153 for follow up analyses. The study was approved by the Berkshire NHS Research Ethics
154 Committee.

155

156 Siblings and family members of the dyslexia and RD cases were assigned to a category of
157 mild reading difficulties if they scored below 1 SD on at least one reading measure or had an
158 average score across all reading measures at least 0.25 SD below the mean.

159

160 **Bioinformatics analysis**

161 WES data for all samples were generated using Illumina technology (NovaSeq 6000, Q30 \geq
162 80%, with 50X coverage). Raw sequences were trimmed using Trimmomatic (35). Trimmed
163 reads were mapped to the human reference genome (GRCh38) using bwa-mem
164 (<https://github.com/lh3/bwa>). Picard tools (<http://broadinstitute.github.io/picard/>) were
165 used for read-groups replacement, removal of PCR duplicates, and base recalibration. The
166 pre-processed reads were indexed using SAMtools (<https://www.htslib.org/>) and called with
167 DeepVariant v1.4.0 (36). VCF files were annotated with ANNOVAR (37) to identify high-
168 impact variants.

169 In accordance with the ACMG guidelines (38), high-impact variants were defined as rare
170 variants (PM2) predicted to damage protein function (PP3). Accordingly, variants were
171 removed if they had a minor allele frequency (MAF) \geq 1% (based on gnomAD v3.1.2 for Non-
172 Finnish Europeans), or if they were annotated as synonymous, intronic or intergenic variants.
173 To be retained, variants had to be predicted as damaging by all five predictive algorithms
174 (SIFT, PolyPhen2, LRT, MutationTaster and FATHMM) or already annotated as pathogenic or
175 likely pathogenic in ClinVar for any disorders (<https://www.ncbi.nlm.nih.gov/clinvar/>).

176 Genes were selected as candidates if high-impact variants were detected in three or more
177 independent cases in the discovery cohort (in accordance with Genomics England guidelines
178 (<https://panelapp.genomicsengland.co.uk/>)). To warrant further follow-up, the gene also had

179 to carry at least one high-impact variant in the RD replication cases and no such variants in
180 the TD replication controls.

181 Genomic localisation of the variants of interest was based on UniProt and InterPro data.

182 Exonic locations were accessed from UCSC Table Browser hg38.refGene.

183

184 **Post-WES analyses**

185 The overlap between the genes identified as candidates in the discovery cohort and those

186 found to be associated in the dyslexia GWAS by Doust et al. (2022) was tested with the

187 GeneOverlap package in R (<https://rdrr.io/bioc/GeneOverlap/>).

188

189 Gene-set enrichment was performed in FUMA (GENE2FUNC tool <https://fuma.ctglab.nl/>)

190 using Gene Ontology classifications and expression data from GTExv8.0. A list of unique

191 genes was compared against all protein-coding genes, with Benjamini-Hochberg (false

192 discovery rate, FDR) correction applied for multiple testing. Enrichment within brain regions

193 at different developmental stages was tested using the BrainSpan dataset.

194

195 Brain expression levels for top genes were extracted from GTExv8.0 (39) and reported as the

196 maximum observed normalised gene expression across 13 brain regions: amygdala, anterior

197 cingulate cortex, basal ganglia, cerebellar hemisphere, cerebellum, cortex, frontal cortex,

198 hippocampus, hypothalamus, nucleus accumbens, putamen, spinal cord, and substantia

199 nigra.

200

201

202 **Results**

203 WES analysis in this discovery cohort (N=53) identified 580 high-impact variants (as defined

204 in Methods) across 512 genes (Supplementary Table 2). On average, each individual carried

205 11 high-impact variants.

206

207 To investigate overlaps with previous findings from analyses of common variants, we

208 compared the identified genes with the 173 genes reported in gene-based analysis of the

209 dyslexia GWAS by Doust and colleagues (7). There was no statistically significant overlap ($P =$
210 0.16) and only seven of the GWAS-identified genes carried high-impact variants in our
211 current study: *CUX2*, *HSPB2*, *INA*, *MITF*, *SCN5A*, *SGCD*, and *WDR38*. Among these, *INA* was
212 the only gene that carried high-impact variants in more than one independent case ($N = 2$;
213 Supplementary Table 2).

214

215 **Table 1. Descriptive statistics of the cohorts.**

Cohort	N (M)	Age	PIQ	SWR	NWR	SPELL	IWR	ORTH	PA
Discovery	53 (40)	12	0.17	-2.04	-1.52	-2.30	-2.39	-1.79	-1.25
Replication			PIQ	SWR	NWR	ACC	COMP	RR	
RD	38 (22)	8	-0.10	-0.71	-0.80	-1.05	-1.00	-0.67	
TD	82 (36)	8	0.47	0.97	0.73	0.34	0.27	0.62	
T-stat			3.77	11.21	10.23	11.23	10.12	8.22	
<i>P</i>			1.55E-3	1.61E-19	3.50E-17	2.02E-19	8.84E-17	2.09E-12	

216 Phenotypic scores are represented as mean z-scores, which have been standardized based on population
217 distributions with mean=100 and SD=15. N: sample size; M: male; Age: mean age at assessment in years; PIQ:
218 performance IQ, SWR: single word reading; NWR: non-word reading; SPELL: single word spelling accuracy; IWR:
219 irregular word reading; ORTH: orthographic coding; PA: phonemic awareness; ACC: NARA-II accuracy; COMP:
220 NARA-II comprehension; RR: NARA-II reading rate; T-stat: absolute t-test statistic; P-value: Bonferroni adjusted
221 for six tests.

222

223

224 The 512 genes identified in our analysis did not include any previously reported candidate
225 genes associated with dyslexia through common (*i.e.* *KIAA0319* and *DCDC2*) or rare (*i.e.*
226 *CEP63*, *DYX1C1*, *NCAN*, *ROBO1*, and *SEMA3C*) variants. However, our annotations for disease
227 associations highlighted two different variants in *ZGRF1*, a gene previously implicated in CAS
228 (Peter et al., 2016), in two independent cases. Gene pathway enrichment analysis showed
229 that the top associated pathways were action potential for the biological processes (BP, $p_{\text{adj}} =$
230 3.55E-6) category, membrane protein complex for the cellular component (CC, $p_{\text{adj}} = 5.23E-$
231 14) category, and adenyly nucleotide binding for the molecular function (MF, $p_{\text{adj}} = 4.56E-14$)
232 category (Supplementary Tables 3–5).

233

234 Analyses of gene expression profiling data did not show enrichment for brain expression in
235 any specific regions when testing GTEx v8 general tissue types (Supplementary Table 6) or
236 BrainSpan datasets (Supplementary Tables 7 and 8).

237

238 There were 22 genes that presented high-impact variants in at least three independent cases
239 (Supplementary Table 9). The gene with the highest number of high-impact variants was
240 *CFTR*, with six different variants identified in seven individuals. This was followed by *CLDN3*
241 which had the same 7-73769649-G-A variant reported in five unrelated individuals.

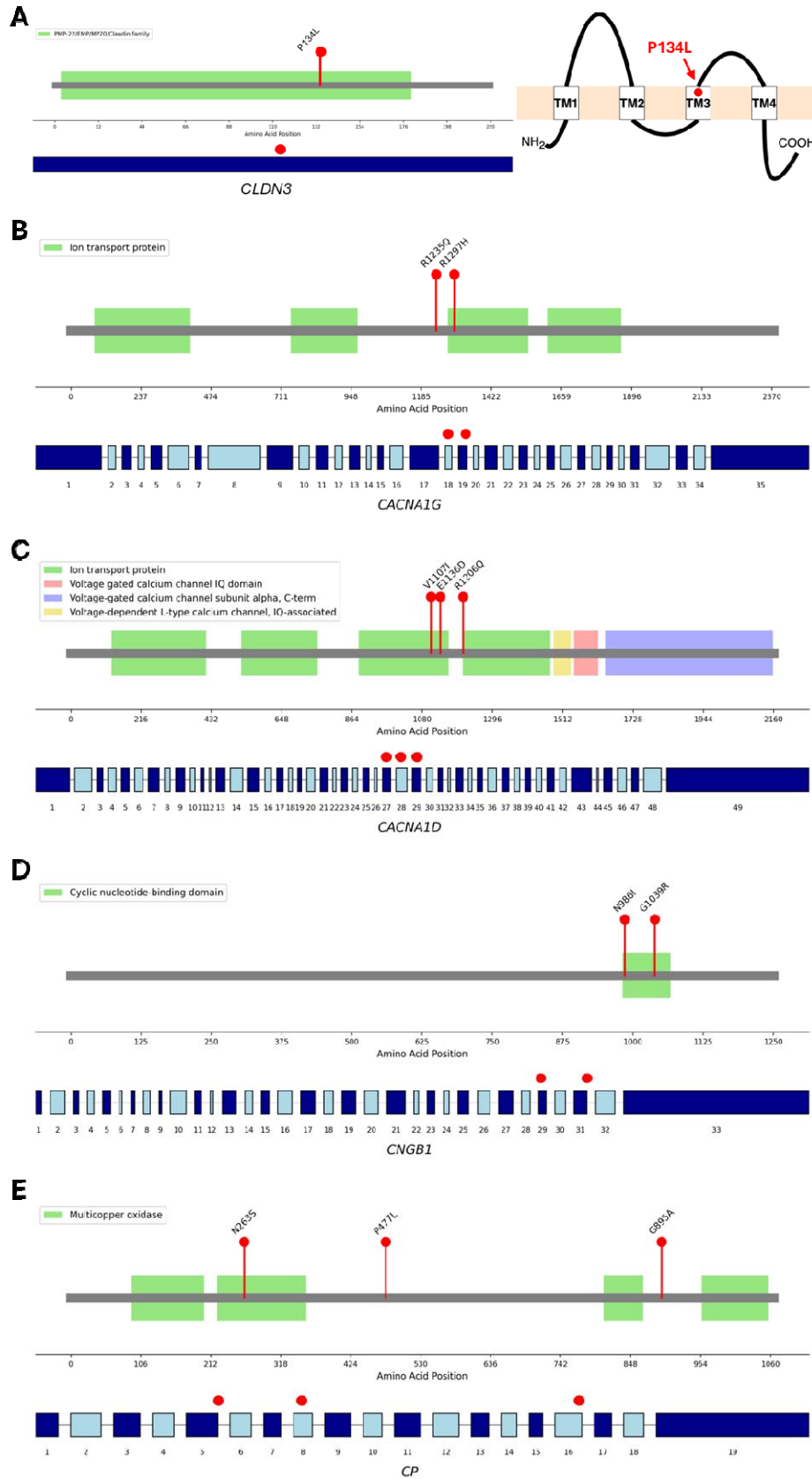
242 The top associations in gene pathways enrichment analysis for these genes were membrane
243 depolarization (BP, $p_{\text{adj}} = 0.01$; Supplementary Table 10), transporter complex (CC, $p_{\text{adj}} =$
244 $4.26\text{E-}3$; Supplementary Table 11), and high voltage-gated calcium channel activity (MF, $p_{\text{adj}} =$
245 $2.63\text{E-}3$; Supplementary Table 12). This group of genes did not exhibit specific patterns of
246 brain expression (Supplementary Tables 13–15).

247

248 We then conducted a replication analysis for the 22 genes in 38 individuals with reading
249 difficulties (RD replication cases) and in 82 typically developing (TD) controls (Table 1). In the
250 RD replication cases, a gene was considered replicated if at least one high-impact variant
251 was present, even if this involved a different variant from that identified in the discovery
252 cohort. Additionally, our replication criteria required that no variants identified in the
253 discovery and RD cases could be present in the TD group. Under these criteria, five genes
254 showed evidence of replication: *CACNA1D*, *CACNA1G*, *CLDN3*, *CNGB1* and *CP*. An additional
255 7-73769649-G-A rare allele in *CLDN3* was identified in the RD cases (Table 2).

256 For three genes (*CACNA1G*, *CACNA1D*, and *CNGB1*), high-impact variants tended to cluster in
257 adjacent exons and, consequently, in spatially close regions of the resulting protein (Figure
258 1).

It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).



260 **Figure 1. Schematic showing high-impact variants in five replicated genes.** The location of
261 high-impact variants (red dots) is shown in relation to exons (lower tracks) and protein
262 domains (upper track) for the A) *CLDN3*, B) *CACNA1G*, C) *CACNA1D*, D) *CNGB1* and E) *CP*
263 genes. A) A 2D structure of the single-exon *CLDN3* gene shows that the 7-73769649-G-A
264 variant leads to the P134L amino acid change in the third transmembrane (TM3) domain of
265 the protein.

266

267

268 All high impact variants occurred in heterozygous status. One individual carried high impact
269 variants in three of these genes (*CNGB1*, *CLDN3* and *CACNA1D*) and two individuals carried
270 high impact variants in two genes (*CACNA1D* and *CLDN3*; *CACNA1G* and *CLDN3*).

271 We compared reading scores between the carriers and non-carriers of the *CLDN3* variant
272 which showed highest recurrence in the discovery cohort (Supplementary Table 16). On
273 average the carriers of this variant exhibited lower scores on all six reading-related
274 measures. The individual in the RD replication cases who carried the 7-73769649-G-A rare
275 allele exhibited a high PIQ (z-score = 1.13), an extremely low reading scores (*i.e.* SWR z-score
276 = -3; average reading z-scores = -2.23). Their dizygotic twin sibling shared the same variant
277 and did not meet criteria for assignment to either the RD or the TD group. This sibling
278 presented a high PIQ score (z-score = 1.6) and all five reading scores were below the mean
279 (average z-score = -0.68), meeting criteria for mild reading problems.

280

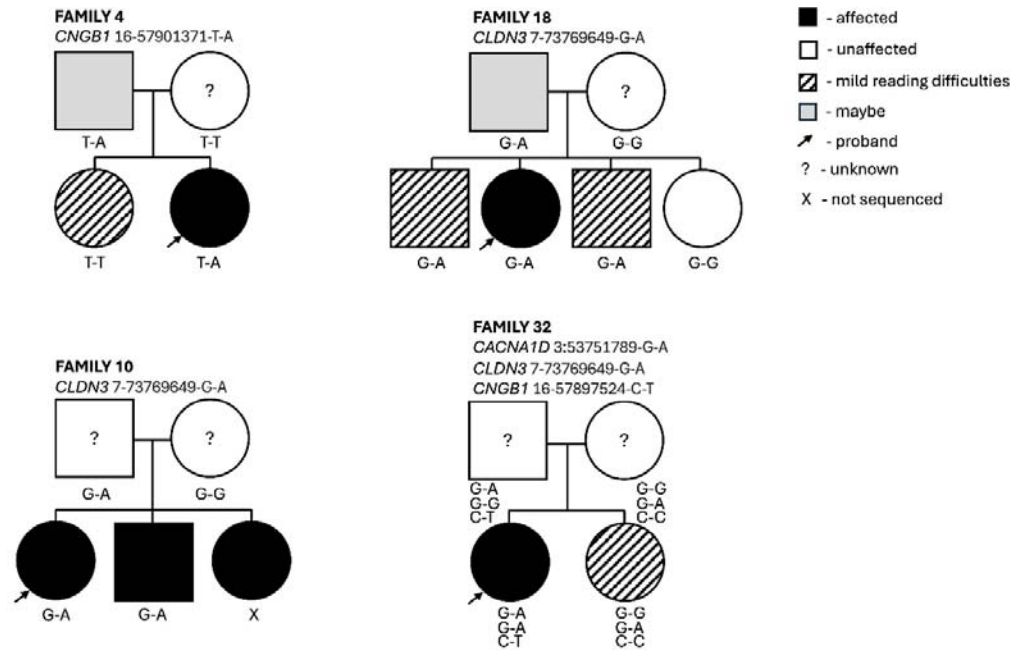
281 Finally, we evaluated the segregation patterns of variants in the five replicated genes, for
282 cases in the discovery cohort where family members were available. Siblings had been
283 assessed using the same battery of tests as the cases. Parents had not been assessed
284 systematically, but for some of them self-reported information for dyslexia diagnoses or
285 reading difficulties were available.

286

287 No data on family members were available for carriers of variants in the *CACNA1G* and *CP*
288 genes. Only one family was available for *CNGB1* (16-57901371-T-A), two families were
289 available for *CLDN3* (7-73769649-G-A), and one family had variants in three genes (*CNGB1*
290 16-57901371-T-A, *CLDN3* 7-73769649-G-A and *CACNA1D* 3-53751789-G-A; Figure 2).

291 Overall, the observed segregation patterns were consistent with dominant effects with

292 variable expressivity as all individuals carrying high-impact variants and for whom data were
293 available exhibited some form of reading difficulties.
294



295
296 **Figure 2. Segregation analysis.** Four families were available for segregation analysis of the
297 variants in the five candidate genes. Overall, segregation patterns are consistent with
298 dominant effects with variable expressivity. The phenotypes of the probands and their
299 siblings were defined through a battery of cognitive tests aimed at assessing reading
300 abilities. Parents were not formally assessed, and their phenotypic status was based on self-
301 reported information. Parents who answered "maybe" when asked if they had dyslexia are
302 indicated in grey. A question mark indicates that no information was available for the
303 parents. Mild reading difficulties (striped pattern) in the children were defined as having
304 either at least one of the reading scores lower than 1 SD from the mean or an average score
305 across all reading measures at least 0.25 SD below the mean.

306 **Table 2.** Genes meeting our prioritisation criteria

Gene	Variant	gnomAD AF	CADD score	Discovery N=53	Replication N=38	Control N=82	Max brain Expression**	Nervous system region
CLDN3	7-73769649-G-A	0.0078	33	5	1	0	2.03	Cerebellar Hemisphere
	All variants			5	1	1*		
CACNA1G	17-50600739-G-A	0.0034	28	3	1	0	3.86	Cerebellum
	17-50601149-G-A	0.000031	34	1	0	0		
	All variants			4	1	0		
CACNA1D	3-53747393-G-A	0.0000155	34	1	0	0	1.7	Cerebellum
	3-53749301-G-T		28.8	1	0	0		
	3-53751789-G-A	0.0000465	34	1	0	0		
	All variants			3	1*	0		
CNGB1	16-57897524-C-T	0.0061	34	2	1	0	2.48	Hypothalamus
	16-57901371-T-A	0.0011	31	1	0	0		
	All variants			3	2*	0		
CP	3-149178609-C-G	0.0023	27.4	1	1	0	0.87	Spinal cord
	3-149199783-G-A	0.0042	28.3	1	0	0		
	3-149207611-T-C	0.0006	25.8	1	0	0		
	All variants			3	1	1*		

307 * variant in the same gene but different from those identified in the discovery cohort; ** maximum observed normalised gene expression across 13 brain regions; CADD =
308 Combined Annotation Dependent Depletion.

309 Discussion

310 We report the results of a WES study of dyslexia conducted in a discovery cohort (N = 53)
311 and followed up in a replication cohort (N = 38 RD cases and N = 82 TD controls),
312 representing the largest analysis of this kind conducted to date. All cases and their siblings
313 were assessed in person with a battery of standardised tests. Our findings highlight five
314 replicated genes with high-impact variants detected in 26% (N = 14) individuals of the
315 discovery cohort.

316 The observation with highest recurrence was a specific rare variant (7-73769649-G-A) in the
317 *CLDN3* gene which was observed in a total of six unrelated cases across the discovery and RD
318 cohorts (N_{total} = 91). This corresponds to an allele frequency of 3.3%, representing a four-fold
319 increase compared to the frequency of 0.8% reported for this variant in the gnomAD
320 samples of European ancestry. Moreover, this variant was not observed in the TD controls,
321 minimising a potential population structure effect caused by a recruitment bias within the
322 UK.

323 *CLDN3* is a single exon gene spanning 1274 bases in total at 7q11.23, within the critical
324 region deleted in Williams syndrome, a developmental disorder characterized by learning
325 difficulties, distinctive facial features, and cardiovascular problems (41). This gene encodes
326 Claudin-3, a protein expressed in different epithelia with roles for the formation of tight
327 junctions and the blood-brain-barrier (42). Depletion of *Cldn3* has been shown to lead to
328 neural tube defects using a chick embryos model system. Targeted analysis of *CLDN* genes in
329 152 patients with spinal neural tube defects (NTDs) identified eleven variants, including a
330 A128T missense variant in *CLDN3*. Overexpression of the *CLDN3* A128T variant in chick
331 embryos resulted in a significant increase in NTDs in this model (43). The 7-73769649-G-A
332 variant identified in this study results in a P134L amino acid change, which, like A128T, is
333 located in the third transmembrane (TM3) domain of Claudin-3. In both Claudin-3 and
334 Claudin-5, the TM3 domain is critical for tight junctions assembly, as demonstrated by
335 studies in HEK293 cells (44). These observations, combined with a high CADD score (= 33;
336 Table 2), predict that the 7-73769649-G-A variant is likely to impact the function of the
337 *CLDN3* protein.

338 Among the five replicated genes were two members of the voltage-gated calcium channel
339 (VGCC) family, *CACNA1D* and *CACNA1G*. VGCC genes are critical for proper brain function,

340 mediating essential calcium-dependent processes such as gene transcription,
341 neurotransmitter release, and neurite outgrowth (45). Both *CACNA1D* and *CACNA1G*, which
342 show high level of expression in the cerebellum, have been previously implicated in
343 neurodevelopmental disorders. *CACNA1D* variants have been associated with ASD, global
344 developmental delay and other neurodevelopmental conditions (46,47). *CACNA1G* variants
345 have been associated with infantile-onset syndromic cerebellar ataxia (48), severe
346 neurodevelopmental delay, epilepsy (47) and intractable seizures (49). Furthermore, we
347 observed high-impact variants in another VGCC gene, *CACNA1C*, in two independent cases
348 from the discovery cohort (Supplementary Table 2). No variants in *CACNA1C* were observed
349 in the TD group.

350

351 For the remaining two replicated genes, there is less evidence for a role in
352 neurodevelopment. The *CNGB1* gene encodes the cyclic nucleotide-gated channel beta 1
353 protein, which is a subunit of the rod photoreceptor cyclic nucleotide-gated channels. These
354 channels ensure a flow of ions into the rod photoreceptor outer segment in response to
355 light-induced changes in intracellular cyclic GMP levels. *CNGB1* variants are associated with
356 retinitis pigmentosa, a degenerative eye disease that affects photoreceptor function (50). *CP*
357 encodes ceruloplasmin, a major copper-carrying protein in the blood. Variants in the *CP*
358 gene can lead to aceruloplasminemia, a rare genetic disorder characterised by iron
359 accumulation in the brain and other organs which can result in neurological and systemic
360 symptoms (51). Elevated levels of iron in the cortical speech motor network have been
361 reported in people who stutter (52).

362

363 In terms of intragenic localisation, the variants detected for *CACNA1D*, *CACNA1G* and *CNGB1*
364 clustered within restricted regions of each gene. Specifically, the three *CACNA1D* (one novel
365 and two ultra-rare) variants, cluster within exons 26-28, while all two *CNGB1* variants cluster
366 within the cyclic nucleotide-monophosphate (c-NMP) binding domain (Figure 1).

367

368 We did not detect variants in genes previously identified in GWAS, WES, or candidate gene
369 association studies related to dyslexia, and none of the five genes highlighted in our study

370 have been previously implicated in dyslexia. This underscores the highly polygenic nature of
371 the condition.

372

373 However, our annotations highlighted two unrelated individuals with distinct variants in
374 *ZGRF1* (Supplementary Table 2), one of which was the same variant (4-112585555-C-T)
375 previously reported in CAS, a language-related condition (40). The same 4-112585555-C-T
376 variant was present also in a TD control and therefore studies of larger cohorts are required
377 to clarify the role of this gene in language-related conditions.

378 The gene with highest number of recurrent variants observed in the discovery cohort was
379 *CFTR*, which was not retained as a gene of interest because one of the variants detected in
380 the discovery sample was also present in the TD group. Variants in *CFTR* are known to cause
381 cystic fibrosis, and its role in neurodevelopment is not well established. However, early
382 studies reported a significant linkage signal for language impairment at the *CFTR* locus (53).
383 These two examples highlight the importance of including controls with relevant phenotypic
384 measures to highlight potential false positives. Conversely, the likelihood of false negatives
385 was increased by the relatively small size of our sample, requiring stringent filtering criteria
386 for pathogenicity predictions. WES analyses in larger cohorts will enable a more systematic
387 evaluation of the role of rare variants to dyslexia, such as for the spectrum of *CFTR* variants.

388

389 The detailed cognitive assessment of cases and their family members allowed us to further
390 evaluate the phenotypic effects of the top five genes and variants. Although limited to four
391 families, the segregation patterns were consistent with dominant effects for the three genes
392 that were analysed. All variant carriers with available phenotypic data exhibited some
393 degree of reading problems. In family 4, the *CNGB1* risk variant is transmitted from the
394 father, who self-reported reading problems, to the proband but not to the sibling. The
395 sibling presents with mild reading difficulties suggesting that other risk factors contribute
396 might be present in this family. The patterns in family 10 and 18 are consistent with a
397 dominant effect of the *CLDN3* 7-73769649-G-A variant on dyslexia and/or reading
398 difficulties. In family 18, the variant is transmitted from the father with self-reported reading
399 difficulties to three siblings with either dyslexia (PIQ z-score = 1.5; SWR z-score = -2.2;

400 average reading measures z-score = -2.51) or mild reading problems. One of the siblings with
401 a mild phenotype has a high PIQ (z-score = 1.7) and an average z-score across the reading
402 measures of -0.54. The other sibling does not have a PIQ measure but presents with a high
403 verbal IQ (z-score = 1.6) and an average z-score across the reading measures of -0.4. The
404 fourth sibling does not have the variant and is a good reader (SWR z-score = 1.3). Finally, in
405 family 32, the proband carries three risk variants and presents exceptionally low reading
406 scores (average across six reading tasks = -2.5 SD) despite having a PIQ above the mean (z-
407 score = 0.1). The sibling, who shares only the *CLDN3* variant out of the three, presents a mild
408 phenotype characterised by a high PIQ (z-score = 2.3) and low z-scores on specific reading
409 tasks (IWR = -0.61, ORTH = -1.16).

410 A large discrepancy between PIQ and reading scores were observed for the carrier of the
411 *CLDN3* variant identified in the RD replication cases (PIQ z-score = 1.13; average reading z-
412 scores = -2.23) and the carrier's sibling (PIQ z-score = 1.6; average reading z-scores = -0.68)
413 who shared the same variant. Overall, such patterns support a role for these rare variants in
414 dyslexia and varying degrees of reading difficulties.

415 The phenotypic variability suggests that additional factors modulate the effects of these
416 variants. It is worth noting that, in the discovery cohort, the other two carriers of the
417 *CLDN3* variant for whom we had no family data, also carried a high impact variant in either
418 *CACNA1D* or *CACNA1G*. This suggests that the effect of the *CLDN3* variant on reading
419 abilities could potentially be modulated by other rare risk variants. The phenomenon of
420 variable expressivity, as well as incomplete penetrance, has been reported for most complex
421 traits, including psychiatric and cognitive phenotypes (24). These phenomena are likely to
422 reflect the interaction of genetic, environmental, and lifestyle factors in influencing
423 phenotypic outcomes. Typically, rare variants discovered in clinical cohorts with specific
424 phenotypic characteristics tend to be associated with milder manifestations in the general
425 population, likely due to the absence of such interacting risk factors.

426

427 In conclusion, this study advances our understanding of potential roles of rare genetic
428 variants in dyslexia. We identified five novel candidate genes. These include *CLDN3* which
429 plays several key functions in neurodevelopmental processes. Notably, a specific variant in

430 *CLDN3* was identified in six independent cases. Two other genes, *CACNA1D* and *CACNA1G*,
431 members of the VGCC gene-family, are involved in neuronal excitability, and have previously
432 been linked to other neurodevelopmental conditions. While our findings do not overlap with
433 previously reported dyslexia-associated genes, they align with the established role of VGCC
434 genes in psychiatric and neurodevelopmental conditions. We extend the significance of this
435 pathway to dyslexia, suggesting that VGCCs may contribute to its aetiology and linking these
436 genes to shared mechanisms across neurodevelopmental traits. Additionally, genes such
437 as *CFTR* and *CACNA1C*, which were excluded based on our stringent filtering criteria, may
438 represent potential candidates for future investigations.

439 These findings underscore the importance of systematic WES or WGS analysis in larger, well-
440 characterised cohorts to identify additional genetic factors and further advance our
441 understanding of the neurobiology underlying dyslexia.

442

443 **Code availability**

444 Code used in the current study to analyse the data can be found at

445 https://github.com/kmarianski/DyslexiaWES_CLDN3

446

447 **Acknowledgements**

448 We would like to thank the participants, their families, recruiters and health-care staff

449 involved in collection of the data.□

450

451 **Funding**

452 KM is supported by a Medical Research Scotland scholarship [PhD-50010-2019]. This work
453 was supported by Action Medical Research Action/The Chief Scientist Office (CSO), Scotland
454 grant [GN2614] and a Royal Society Grant [UF100463]. Bioinformatic analysis was conducted
455 on the UK's Crop Diversity Bioinformatics HPC which is funded by the BBSRC
456 [BB/S019669/1]. The recruitment of the Discovery cohort was supported by Wellcome Trust
457 Grants [076566/Z/05/Z and 075491/Z/04] and a Waterloo Foundation Grant [797–1720].

458 Recruitment and analysis of the Replication cohorts was supported by Wellcome Trust

459 Programme Grants [082498], and European Research Council Advanced Grant [694189]. SEF
460 is supported by the Max Planck Society.

461

462 **References**

- 463 1. Erbeli F, Rice M, Paracchini S. Insights into dyslexia genetics research from the last two
464 decades. *Brain Sciences* 2022, Vol 12, Page 27. 2021 Dec;12(1):27–27.
- 465 2. Arnett AB, Pennington BF, Peterson RL, Willcutt EG, DeFries JC, Olson RK. Explaining the
466 sex difference in dyslexia. *Journal of Child Psychology and Psychiatry and Allied*
467 *Disciplines*. 2017 Jun, 58(6), 719-727
- 468 3. Daucourt MC, Erbeli F, Little CW, Haughbrook R, Hart SA. A Meta-Analytical review of the
469 genetic and environmental correlations between reading and Attention-
470 Deficit/Hyperactivity Disorder symptoms and reading and math. *Scientific Studies of*
471 *Reading*. 2020 Jan 2;24(1):23–56.
- 472 4. Snowling MJ, Hayiou-Thomas ME, Nash HM, Hulme C. Dyslexia and Developmental
473 Language Disorder: comorbid disorders with distinct effects on reading comprehension.
474 *Journal of Child Psychology and Psychiatry*. 2020;61(6):672–80.
- 475 5. Snowling MJ, Melby-Lervåg M. Oral language deficits in familial dyslexia: A meta-analysis
476 and review. *Psychological Bulletin*. 2016;142(5):498–545.
- 477 6. Paracchini S, Scerri TS, Monaco AP. The genetic lexicon of dyslexia. *Annual Review of*
478 *Genomics and Human Genetics*. 2007/04/21 ed. 2007;8:57–79.
- 479 7. Doust C, Fontanillas P, Eising E, Gordon SD, Wang Z, Alagöz G, et al. Discovery of 42
480 genome-wide significant loci associated with dyslexia. *Nature Genetics*. 2022
481 Nov;54(11):1621–9.
- 482 8. Eising E, Mirza-Schreiber N, de Zeeuw EL, Wang CA, Truong DT, Allegrini AG, et al.
483 Genome-wide analyses of individual differences in quantitatively assessed reading- and
484 language-related skills in up to 34,000 people. *Proceedings of the National Academy of*
485 *Sciences*. 2022 Aug 30;119(35):e2202764119.
- 486 9. Momozawa Y, Mizukami K. Unique roles of rare variants in the genetics of complex
487 diseases in humans. *Journal Human Genetics*. 2021 Jan;66(1):11–23.
- 488 10. Wang T, Zhao PA, Eichler EE. Rare variants and the oligogenic architecture of autism.
489 *Trends in Genetics*. 2022 Sep 1;38(9):895–903.
- 490 11. Howrigan DP, Rose SA, Samocha KE, Fromer M, Cerrato F, Chen WJ, et al. Exome
491 sequencing in schizophrenia-affected parent–offspring trios reveals risk conferred by
492 protein-coding de novo mutations. *Nature Neuroscience*. 2020 Feb;23(2):185–93.

- 493 12. Forstner AJ, Fischer SB, Schenk LM, Strohmaier J, Maaser-Hecker A, Reinbold CS, et al.
494 Whole-exome sequencing of 81 individuals from 27 multiply affected bipolar disorder
495 families. *Translational Psychiatry*. 2020 Feb 4;10(1):57.
- 496 13. Eising E, Carrion-Castillo A, Vino A, Strand EA, Jakielski KJ, Scerri TS, et al. A set of
497 regulatory genes co-expressed in embryonic human brain is implicated in disrupted
498 speech development. *Molecular Psychiatry*. 2019 Jul;24(7):1065–78.
- 499 14. Hildebrand MS, Jackson VE, Scerri TS, Van Reyk O, Coleman M, Braden RO, et al. Severe
500 childhood speech disorder: Gene discovery highlights transcriptional dysregulation.
501 *Neurology*. 2020;94(20):e2148–67.
- 502 15. Kaspi A, Hildebrand MS, Jackson VE, Braden R, van Reyk O, Howell T, et al. Genetic
503 aetiologies for childhood speech disorder: novel pathways co-expressed during brain
504 development. *Molecular Psychiatry*. 2023 Apr;28(4):1647–63.
- 505 16. Rajagopal VM, Duan J, Vilar-Ribó L, Grove J, Zayats T, Ramos-Quiroga JA, et al.
506 Differences in the genetic architecture of common and rare variants in childhood,
507 persistent and late-diagnosed attention-deficit hyperactivity disorder. *Nature Genetics*.
508 2022 Aug;54(8):1117–24.
- 509 17. Chen XS, Reader RH, Hoischen A, Veltman JA, Simpson NH, Francks C, et al. Next-
510 generation DNA sequencing identifies novel gene variants and pathways involved in
511 specific language impairment. *Scientific Reports*. 2017;7:46105–46105.
- 512 18. Devanna P, Chen XS, Ho J, Gajewski D, Smith SD, Gialluisi A, et al. Next-gen sequencing
513 identifies non-coding variation disrupting miRNA-binding sites in neurological disorders.
514 *Molecular Psychiatry*. 2018;23(5):1375–84.
- 515 19. Villanueva P, Nudel R, Hoischen A, Fernández MA, Simpson NH, Gilissen C, et al. Exome
516 sequencing in an admixed isolated population indicates NFXL1 variants confer a risk for
517 specific language impairment. Abrahams BS, editor. *PLoS Genetics*. 2015
518 Mar;11(3):e1004925–e1004925.
- 519 20. Taipale M, Kaminen N, Nopola-Hemmi J, Haltia T, Myllyluoma B, Lyytinen H, et al. A
520 candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat
521 domain protein dynamically regulated in brain. *Proceedings of the National Academy of
522 Sciences*. 2003;100(20):11553–8.
- 523 21. Einarsdottir E, Svensson I, Darki F, Peyrard-Janvid M, Lindvall JM, Ameer A, et al.
524 Mutation in CEP63 co-segregating with developmental dyslexia in a Swedish family.
525 *Human Genetics*. 2015;134(11–12):1239–48.
- 526 22. Einarsdottir E, Peyrard-Janvid M, Darki F, Tuulari JJ, Merisaari H, Karlsson L, et al.
527 Identification of NCAN as a candidate gene for developmental dyslexia. *Scientific
528 Reports*. 2017 Aug 24;7(1):9294.

- 529 23. Carrion-Castillo A, Estruch SB, Maassen B, Franke B, Francks C, Fisher SE. Whole-genome
530 sequencing identifies functional noncoding variation in SEMA3C that cosegregates with
531 dyslexia in a multigenerational family. *Human Genetics*. 2021 Aug 1;140(8):1183–200.
- 532 24. Kingdom R, Wright CF. Incomplete penetrance and variable expressivity: from clinical
533 studies to population cohorts. *Frontiers in Genetics*. 2022 Jul 25;13:920390.
- 534 25. Scerri TS, Macpherson E, Martinelli A, Wa WC, Monaco AP, Stein J, et al. The DCDC2
535 deletion is not a risk factor for dyslexia. *Translational Psychiatry*. 2017 Jul 25;7(7):e1182–
536 e1182.
- 537 26. Elliot CD Murray, DJ, and Pearson, LS. *British Abilities Scales*. NFER-Nelson, Windsor, UK.
538 1983.
- 539 27. Castles A, Coltheart M. Varieties of developmental dyslexia. *Cognition*. 1993;47(2):149–
540 80.
- 541 28. Frederickson N Frith, U, Reason, R. *Phonological assessment battery (PhAB)*. Windsor:
542 NFER-Nelson. 1997.
- 543 29. Olson R Forsberg, H, Wise, B, & Rack, J. Measurement of word recognition, orthographic,
544 and phonological skills. In G. R. Lyon (Ed.), *Frames of reference for the assessment of*
545 *learning disabilities: New views on measurement issues*. Paul H Brookes Publishing Co.
546 1994.
- 547 30. Wechsler D, Golombok J, Rust J. *WISC-III UK Wechsler intelligence scale for children: UK*
548 *manual*. Sidcup, UK: The Psychological Corporation; 1992.
- 549 31. Wilson AC, Bishop DVM. Resounding failure to replicate links between developmental
550 language disorder and cerebral lateralisation. *PeerJ*. 2018;6:e4217.
- 551 32. Torgesen JK, Rashotte CA, Wagner RK. *TOWRE: Test of word reading efficiency*. Pro-ed
552 Austin, TX; 1999.
- 553 33. Neale MD. *Neale Analysis of Reading Ability: Reader*. Third Edition. ACER Press,
554 Australian Council for Educational Research Limited, 19 Prospect Hill Road, Camberwell,
555 Melbourne, Victoria 3124, Australia; 1999.
- 556 34. Goodman R, Ford T, Richards H, Gatward R, Meltzer H. The Development and Well-Being
557 Assessment: description and initial validation of an integrated assessment of child and
558 adolescent psychopathology. *Journal of Child Psychology and Psychiatry*. 2000/08/18 ed.
559 2000;41(5):645–55.
- 560 35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
561 data. *Bioinformatics*. 2014 Aug 1;30(15):2114–20.

- 562 36. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and
563 small-indel variant caller using deep neural networks. *Nature Biotechnology*. 2018
564 Nov;36(10):983–7.
- 565 37. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from
566 high-throughput sequencing data. *Nucleic Acids Research*. 2010 Sep;38(16):e164.
- 567 38. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines
568 for the interpretation of sequence variants: a joint consensus recommendation of the
569 American College of Medical Genetics and Genomics and the Association for Molecular
570 Pathology. *Genetics in Medicine*. 2015 May;17(5):405–24.
- 571 39. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue
572 Expression (GTEx) project. *Nature Genetics*. 2013 Jun;45(6):580–5.
- 573 40. Peter B, Wijsman EM, Nato AQ, University of Washington Center for Mendelian
574 Genomics, Matsushita MM, Chapman KL, et al. Genetic candidate variants in two
575 multigenerational families with childhood apraxia of speech. Cai T, editor. *PLoS ONE*.
576 2016 Apr 27;11(4):e0153864.
- 577 41. Morris CA, Mervis CB. Williams Syndrome. In: Cassidy and Allanson's Management of
578 Genetic Syndromes . John Wiley & Sons, Ltd; 2021 . p. 1021–38.
579 <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119432692.ch63>
- 580 42. Günzel D, Yu ASL. Claudins and the modulation of tight junction permeability.
581 *physiological review*. 2013 Apr;93(2):525–69.
- 582 43. Baumholtz AI, De Marco P, Capra V, Ryan AK. Functional validation of CLDN variants
583 identified in a neural tube defect cohort demonstrates their contribution to neural tube
584 defects. *Frontiers in Neuroscience*. 2020
585 [https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2020.00664/f](https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2020.00664/full)
586 [ull](https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2020.00664/full)
- 587 44. Rossa J, Ploeger C, Vorreiter F, Saleh T, Protze J, Günzel D, et al. Claudin-3 and Claudin-5
588 protein folding and assembly into the tight junction are controlled by non-conserved
589 residues in the transmembrane 3 (TM3) and extracellular loop 2 (ECL2) segments *.
590 *Journal of Biological Chemistry*. 2014 Mar 14;289(11):7641–53.
- 591 45. Simms BA, Zamponi GW. Neuronal voltage-gated calcium channels: structure, function,
592 and dysfunction. *Neuron*. 2014 Apr 2;82(1):24–45.
- 593 46. Hofer NT, Tuluc P, Ortner NJ, Nikonishyna YV, Fernández-Quintero ML, Liedl KR, et al.
594 Biophysical classification of a CACNA1D de novo mutation as a high-risk mutation for a
595 severe neurodevelopmental disorder. *Molecular Autism*. 2020;11(1):4.
- 596 47. Rajakulendran S, Hanna MG. The role of calcium channels in epilepsy. *Cold Spring Harb*
597 *Perspect Med*. 2016 Jan 4;6(1):a022723.

- 598 48. Barresi S, Dentici ML, Manzoni F, Bellacchio E, Agolini E, Pizzi S, et al. Infantile-onset
599 syndromic cerebellar ataxia and CACNA1G mutations. *Pediatric Neurology*. 2020
600 Mar;104:40–5.
- 601 49. Kunii M, Doi H, Hashiguchi S, Matsuishi T, Sakai Y, Iai M, et al. De novo CACNA1G variants
602 in developmental delay and early-onset epileptic encephalopathies. *Journal of the*
603 *Neurological Sciences*. 2020 Sep;416:117047.
- 604 50. Alshamrani AA, Raddadi O, Schatz P, Lenzner S, Neuhaus C, Azzam E, et al. Severe
605 retinitis pigmentosa phenotype associated with novel CNGB1 variants. *American Journal*
606 *of Ophthalmology Case Reports*. 2020 Sep;19:100780.
- 607 51. Yoshida K, Furihata K, Takeda S, Nakamura A, Yamamoto K, Morita H, et al. A mutation in
608 the ceruloplasmin gene is associated with systemic hemosiderosis in humans. *Nature*
609 *Genetics*. 1995 Mar;9(3):267–72.
- 610 52. Cler GJ, Krishnan S, Papp D, Wiltshire CEE, Chesters J, Watkins KE. Elevated iron
611 concentration in putamen and cortical speech motor network in developmental
612 stuttering. *Brain*. 2021 Nov 29;144(10):2979–84.
- 613 53. O'Brien EK, Zhang X, Nishimura C, Tomblin JB, Murray JC. Association of Specific
614 Language Impairment (SLI) to the Region of 7q31. *The American Journal of Human*
615 *Genetics*. 2003 Jun 1;72(6):1536–43.
- 616
- 617