Severe childhood speech disorder: gene discovery highlights transcriptional dysregulation

Michael S. Hildebrand PhD, [1,21,*] # Victoria E. Jackson PhD, [2,3*] Thomas S. Scerri PhD, [2,3*] Olivia Van Reyk MSpPath,[4] Matthew Coleman BSc (Hons),[1] Ruth O. Braden MSpPath,[4,5] Samantha Turner PhD,[4] Kristin A. Rigbye BSc (Hons),[1] Amber Boys PhD,[6] Sarah Barton DPsych,[4] Richard Webster MD,[7] Michael Fahey MD PhD,[8] Kerryn Saunders MD,[8,9] Bronwyn Parry-Fielder BAppSci,[10] Georgia Paxton MD,[10] Michael Hayman MD,[10] David Coman MD,[11] Himanshu Goel MD, [12] Anne Baxter MD,[12] Alan Ma MD,[13] Noni Davis MD,[14] Sheena Reilly PhD,[4,15] Martin Delatycki MBBS PhD,[6] Frederique J. Liégeois PhD,[16] Alan Connelly PhD,[17] Jozef Gecz PhD,[18] Simon E. Fisher DPhil,[19,20] David J. Amor MBBS PhD,[10,21] Ingrid E. Scheffer MBBS PhD,[1,10,17,21*] Melanie Bahlo PhD,[2,3*] Angela T. Morgan PhD[4,5,*] ##

[1] Department of Medicine, The University of Melbourne, Austin Health, Heidelberg, Victoria 3084, Australia; [2] Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia; [3] Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3052, Australia; *[4] Speech and Language, Murdoch Children's Research Institute, Parkville, Victoria 3052, Australia; [5] Department of Audiology and Speech Pathology, University of Melbourne, Parkville, Victoria 3052, Australia; [6] Victorian Clinical Genetics Services, Parkville, Victoria 3052, Australia; [7]Department of Neurology, The Children's Hospital Westmead, Westmead, NSW 2145, Australia [8] Department of Paediatrics, Monash University, Clayton, Victoria 3168, Australia; [9] Monash Children's Hospital, Clayton ,Victoria 3168, Australia; [10] Department of Paediatrics, The Royal Children's Hospital, The University of Melbourne, Parkville, Victoria 3052, Australia; [11]The Wesley Hospital, Auchenflower, Queensland 4066, Australia; [12] Hunter Genetics, John Hunter Hospital, New Lambton*

*Heights, NSW 2305, Australia; [13] Clinical Genetics, The Children's Hospital Westmead, Westmead, NSW 2145, Australia; [14] Melbourne Children's Clinic, Melbourne, Victoria 3052, Australia [15] Griffith University, Mount Gravatt, Queensland 4122, Australia; [16]UCL Great Ormond Street Institute of Child Health, London WC1N 1EH, UK; [17] Florey Institute of Neuroscience and Mental Health, Parkville 3052, Victoria, Australia; [18]South Australian Health and Medical Research Institute, Robinson Research Institute and Adelaide Medical School, University of Adelaide, Adelaide, South Australia 5005, Australia; [19] Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen 6525 XD, The Netherlands; [20] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen 6525 EN, The Netherlands; [21] Murdoch Children's Research Institute, Parkville, Victoria 3052, Australia; * These authors contributed equally to this manuscript; # michael.hildebrand@unimelb.edu.au; ## angela.morgan@mcri.edu.au*

CORRESPONDING AUTHORS

A/Professor Michael S. Hildebrand; Epilepsy Research Centre, Level 2, Melbourne Brain Centre, 245 Burgundy St. Heidelberg, Victoria 3084, Australia; Telephone: +61 3 9035 7143; E-mail Address: michael.hildebrand@unimelb.edu.au

Professor Angela T. Morgan; Murdoch Children's Research Institute, 50 Flemington Rd., Parkville, Victoria 3052, Australia; Telephone: +61 3 8341 6458; E-mail Address: angela.morgan@mcri.edu.au

WORD AND CHARACTER COUNTS

Running Head: Genetic Basis of Speech Disorder; **Number of Characters:** 28 (Running Head); 81 (Title); **Number of Words:** 250 (Abstract); 3,913 (Main Body); **Number of Figures:** 4; **Number of Color Figures:** 4; **Number Tables:** 3; **Number of References:** 48

STUDY FUNDING

M.S.H., F.J.L., S.E.F., A.C., S.R., D.J.A., I.E.S., M.B. and A.T.M., are funded by a National Health and Medical Research Council (NHMRC) Centre of Research Excellence (1116976) Grant. M.S.H., F.J.L., S.E.F., S.R., D.J.A. and A.T.M. are funded by a NHMRC Project Grant (1127144). M.S.H, A.T.M., I.E.S. and M.B. are supported by the March of Dimes Grant Scheme. M.S.H. is funded by a NHMRC Career Development Fellowship (ID: 1063799). S.E.F. is funded by the Max Planck Society. I.E.S. is funded by a NHMRC Development (1153614) Grant and a Practitioner Fellowship (1006110). M.B. is funded by a NHMRC Senior Research Fellowship (ID: 1102971). A.T.M is funded by a NHMRC Development (1153614) Grant and a Practitioner Fellowship (1105008).

SEARCH TERMS

Speech Apraxia; Clinical Neurology; Developmental Disorders; Genetics; Pediatric

DISCLOSURES

Authors M.Hildebrand, V.Jackson, T.Scerri, O.Van Reyk, M.Coleman, R.Braden, S.Turner, K.Rigbye, A.Boys, S.Barton, R.Webster, M.Fahey, K.Saunders, B. Parry-Fielder, G. Paxton, M. Hayman, D. Coman, H. Goel, A. Baxter, A. Ma, N. Davis, S. Reilly, M. Delatycki, F. Liegeois, A. Connelly, J. Gecz, S. Fisher, D. Amor, M. Bahlo, A. Morgan reports no disclosures relevant to the manuscript. I.Scheffer has served on scientific advisory boards for UCB, Eisai, GlaxoSmithKline, BioMarin, Nutricia and Xenon Pharmaceuticals; editorial boards of the *Annals of Neurology*, *Neurology* and *Epileptic Disorders*; may accrue future revenue on pending patent WO61/010176 (filed: 2008): Therapeutic Compound; has received speaker honoraria from GlaxoSmithKline, Athena Diagnostics, UCB, BioMarin, and Eisai;

has received funding for travel from Athena Diagnostics, UCB, Biocodex, GlaxoSmithKline, Biomarin and Eisai.

ABSTRACT

Objective: Determining the genetic basis of speech disorders provides insight into the neurobiology of human communication. Despite intensive investigation over the past two decades, the etiology of most children with speech disorder remains unexplained. To test the hypothesis that speech disorders have a genetic etiology we performed genetic analysis of children with severe speech disorder, specifically childhood apraxia of speech (CAS).

Methods: Precise phenotyping together with research genome or exome analysis were performed on children referred with a primary diagnosis of CAS. Gene co-expression and gene set enrichment analyses were conducted on high confidence gene candidates.

Results: 34 probands ascertained for CAS were studied. In 11/34 (32%) probands, we identified highly plausible pathogenic single nucleotide (n=10, *CDK13, EBF3, GNAO1, GNB1, DDX3X, MEIS2, POGZ, SETBP1, UPF2, ZNF142*) or copy number (n = 1, 5q14.3q21.1 locus) variants in novel genes or loci for CAS. Testing of parental DNA was available for nine probands and confirmed that the variants had arisen *de novo*. Eight genes encode proteins critical for regulation of gene transcription, and analyses of transcriptomic data found CAS-implicated genes were highly co-expressed in the developing human brain.

Conclusion: We identify the likely genetic aetiology in 11 patients with CAS and implicate 9 genes for the first time. We find that CAS is often a sporadic monogenic disorder, and highly genetically heterogeneous. Highly penetrant variants implicate shared pathways in broad transcriptional regulation, highlighting the key role of transcriptional regulation in normal speech development. CAS is a distinctive, socially debilitating clinical disorder, and

understanding its molecular basis is the first step towards identifying precision medicine approaches.

INTRODUCTION

Childhood speech disorders are common, affecting 1 in 20 preschool children in the general population (1). The majority of children present with mild articulation (e.g., lisp) or phonological errors (e.g., 'f' for 'th') and typically resolve with or without intervention (2). By contrast, approximately 1 in 1000 patients present with persistent and intractable speech disorders such as childhood apraxia of speech (CAS) (3). These individuals typically have abnormal speech development from infancy, with a history of poor feeding, limited babbling, delayed onset of first words, and highly unintelligible speech into the preschool years when a diagnosis is usually made (3). Three core symptoms support a CAS diagnosis, in accordance with consensus-based criteria set by the American Speech-Language-Hearing Association: (1) inconsistent errors on consonants and vowels; (2) lengthened and disrupted co-articulatory transitions between sounds and syllables; and (3) inappropriate prosody. Lifelong impairment is seen with psychosocial impact, literacy deficits, restricted educational and employment outcomes (1).

Childhood apraxia of speech was not shown to have a genetic basis until 2001, with the seminal discovery that pathogenic variants in *FOXP2* [MIM:605317], a transcriptional repressor, causes rare cases of CAS (reviewed in (4)). Later, downstream target *FOXP2* genes such as *CNTNAP2* [MIM:604569] and closely related family member *FOXP1* [MIM:605515], were also implicated in speech and language dysfunction (4). Since then, disruptions of single genes (e.g., *GRIN2A* [MIM:138253] (5)), microdeletions (e.g., 2p16.1,

12p13.33 and 17q21.31 implicating *BCL11A* [MIM:606557], *ERC1* [MIM:607127] and *KANSL1* [MIM:612452] (6)), and larger deletions (e.g., 16p11.2 deletion, encompassing >25 genes) (7) have been associated with CAS. A recent genome sequencing study of 19 predominantly US probands with CAS uncovered causal variants in 8/19 (42%) cases (8), informing diagnosis and genetic counselling for families (9). Here, we sought to understand the genetic architecture of CAS by detailed molecular studies of a larger cohort of 34 patients with CAS. We investigated gene co-expression of identified variants with previously published CAS genes.

MATERIALS AND METHODS

Standard Protocol Approvals, Registrations, and Patient Consents

The Human Research Ethics Committee of The Royal Children's Hospital, Melbourne, Australia, approved this study [Project 37353]. Written informed consent was obtained from living subjects or their parents or legal guardians in the case of minors or those with intellectual disability.

Phenotyping

Inclusion criteria for probands included a primary clinical diagnosis of severe and persistent speech disorder in childhood (< 18 years); that is, not occurring in the setting of severe intellectual disability and where parents and clinicians reported the current primary clinical concern as speech production. Participants were recruited via medical and speech pathology clinicians, online parent support groups for apraxia or direct parent referral. The medical and developmental history of each proband and participating sibling was taken, with strenuous attempts to obtain all medical, speech and neuropsychological assessments to identify additional secondary comorbidities, including hearing impairment, motor deficits, epilepsy,

attention deficit hyperactivity disorder, autism spectrum disorder (see Tables 1, 2). Brain magnetic resonance imaging (MRI) results were obtained.

CAS was diagnosed where children met three operationally defined ASHA diagnostic criteria (7) scored based on single word transcriptions of the: Diagnostic Evaluation of Articulation and Phonology (10), a polysyllable word test (11), and a 5-minute conversational speech sample. Dysarthria was diagnosed in the presence of oral tone or co-ordination disturbance using an oral motor assessment, and dysarthric features identified during conversation using the Mayo Clinic Dysarthria rating scale (6, 7). Language, literacy and cognition were also assessed [See Table S1a]. Parents were assessed with an age-appropriate battery complementary to the child version. [Data available from Dryad (Table S1b): https://doi.org/10.5061/dryad.zkh189363].

Genetic testing

Genomic DNA was extracted from blood using a Qiagen QIAamp DNA Maxi Kit (Valencia, CA) according to the manufacturer's instructions. Only saliva samples were available for some patients, and DNA was extracted using a *prepIT*•L2P *kit (DNA Genotek Inc, Ontario, Canada) according to* the manufacturer's instructions. Probands underwent chromosomal microarray testing on Illumina platforms (Illumina, San Diego, CA), with the reportable effective resolution of arrays being 200Kb. Results were analysed with Karyostudio software version 1.3 or 1.4 (Illumina), using genome reference sequence either NCBI36/hg18 (v1.3 pre 2013) or GRCh37/hg19 (v1.4 2013 onwards).

Variant discovery for the majority of probands was performed using trio, or parent-child pair (where one parent was unavailable for testing) designs. There were three exceptions to this: Proband 25, whose monozygotic twin was also sequenced (quad design, twin also affected);

Proband 26 whose mother, maternal grandmother, and sister were sequenced; and Proband 9 who was analysed as a singleton, as no parental DNA were available.

Whole exome sequencing (WES) was performed on 64 individuals from 23 families: 24 probands (includes the monozygotic twin pair); 38 parents; and the sister and grandmother of proband 26. Genomic DNA was sonicated to approximately 200 base pair (bp) fragments and adaptor-ligated to make a library for paired-end sequencing. Following amplification and barcoding, the libraries were hybridized to biotinylated complementary RNA oligonucleotide baits from the Agilent SureSelect XT Human All Exon +UTR v5 (75Mb) (Agilent Technologies, Santa Clara, CA) and purified using streptavidin-bound magnetic beads. Amplification was performed prior to sequencing on the Illumina HiSeq 2000 system to average 50-fold depth (San Diego, CA). Exome sequencing was run on a research basis at the Australian Genome Research Facility, Victorian Comprehensive Cancer Centre, Melbourne.

Whole genome sequencing WGS) was conducted on 24 individuals from 10 families: 10 probands and 14 parents. Illumina TruSeq DNA Nano (Santa Clara, CA) genome preparation was completed according to the manufacturer's instructions prior to sequencing on the Illumina X Ten (San Diego, CA) to average 30-fold depth. Genome sequencing was run on a research basis at the Kinghorn Centre for Clinical Genomics, Garvin Institute of Medical Research, Sydney.

The total number of individuals (both unaffected and affected) that had WES or WGS in this study was 88. In the follow up of candidate variants, targeted Sanger sequencing including additional family members who had not undergone WES/WGS, was carried out, to allow further segregation analysis.

Variant analysis and validation

We searched for Loss of Function (LoF) and predicted damaging variants exome- or genome-wide. Read pairs were mapped to the hg19 reference genome using Burrow-Wheeler Aligner (BWA-MEM, bwa v. 0.7.15) (12). Reads were sorted using SAMtools (v 1.7) and duplicates marked using Genome Analysis Toolkit (GATK) v4.0.11.0 (13). Base quality score recalibration was performed and variants called using HaplotypeCaller, on a per-sample basis, as implemented by GATK. Genotype calling and quality filtering were performed separately in the exome and genome sequencing batches, as follows: Per sample gvcf files were merged and genotypes were jointly called across all samples using GATK's GenotypeGVCFs tool. Variants with excess heterozygosity (Z-score>4.5) were removed, then Variant Quality Score Recalibration (VQSR) was carried out for SNVs and indels separately, and a truth sensitivity filter of 99.7 was used to flag variants for exclusion. Single nucleotide variants (SNVs) were filtered to exclude those flagged by VQSR or any of the following hard filters: low quality by depth (QD<2); evidence of strand bias (FisherStrand, FS>60 or StrandOddsRatio, SOR>3); evidence of differences between alternate and reference alleles for read mapping qualities (MQRankSum<-12.6) or position bias (ReadPosRankSum<-8). Indels were filtered to exclude any of the following: those flagged by VQSR; QD<2; FS>200; SOR>10; ReadPosRankSum<-20.

Analysis was restricted to (i) variants either not present in gnomAD or present with a mean allelic frequency < 0.05%, and (ii) not present in unaffected family members from our sequenced cohort. Only variants with read depth>10 and genotype quality >20 were

considered. Identified variants were annotated using variant effect predictor (VEP v93.3) using assembly version GRCh37.p13 and categorised based on the following series of annotations.

Predicted Loss of Function (LoF) Candidates were defined using VEP annotations meeting three criteria: 1. Annotated as splice acceptor variant; splice donor variant; frameshift variant; stop lost; stop gained; start lost; 2. In a gene intolerant to LoF variation (ExACpLI >= 0.9 or LoFtool < 0.1); 3. At least one of the following: a) Predicted to be damaging by Combined Annotation Dependent Depletion (CADD) Phred score >= 20; or b) Predicted to affect splicing (Ada Boost score >= 0.6 or random forest score >= 0.6, using the dbscSNV VEP plugin). For frameshift variants, the variant was only required to be in a LoF intolerant gene.

Predicted Damaging Candidates: Missense variants that met at least three criteria: 1. Predicted 'probably damaging' or 'possibly damaging' by PolyPhen-2; 2. Predicted 'deleterious' or 'deleterious low confidence' by SIFT (sorting intolerant from tolerant); 3. Predicted damaging with CADD Phred score >= 20; 4. Missense Tolerance Ratio (MTR) significantly different from 1 (MTR FDR < 0.05); 5. Predicted to affect splicing (Ada Boost score >= 0.6 or random forest score >= 0.6).

Other Notable Candidates: Missense variants which did not meet the above criteria, but were in genes with biological relevance to speech based on the literature, were also identified as candidates. All candidates were inspected by eye in Integrative Genome Viewer (IGV1.3).

Criteria for Reporting Rare or Novel Variants: We report a set of "high confidence" candidate variants, categorised as either predicted LoF or damaging candidates, and classified

as “pathogenic” according to the ACMG guidelines (14). For probands without a high confidence variant, we report “low confidence” candidate variants; these comprise all identified LoF candidates classified as “likely pathogenic”, or of uncertain significance (ACMG guidelines), and a subset of missense variants, in genes of biological relevance to speech based on the literature. ACMG guidelines strictly only apply to known disorder-causing genes (14).

Rare variant validation: Variants of interest were validated using PCR and Sanger sequencing. Gene variants were amplified using gene-specific primers (oligonucleotide sequences available on request) designed to the reference human gene transcripts (NCBI Gene). Amplification reactions were cycled using a standard protocol on a Veriti Thermal Cycler (Applied Biosystems, Carlsbad, CA) at 60ºC annealing temperature for 1 minute. Bidirectional sequencing of all exons and flanking regions was completed with a BigDye™ v3.1 Terminator Cycle Sequencing Kit (Applied Biosystems), according to the manufacturer’s instructions. Sequencing products were resolved using a 3730xl DNA Analyzer (Applied Biosystems). All sequencing chromatograms were compared to the published cDNA sequence; nucleotide changes were detected using Codon Code Aligner (CodonCode Corporation, Dedham, MA).

Interrogation of short tandem repeats

We also examined whether any proband had expanded short tandem repeats (STRs) at any known pathogenic locus [Data available from Dryad (Table S2): https://doi.org/10.5061/dryad.zkh189363]. Genome and exome sequenced samples were examined separately using two short tandem repeat detection tools, Expansion Hunter v.2.5.5 and exSTRa. For each locus we looked for evidence of outlying samples in terms of STR

length by inspecting plots of estimated STR size (ExpansionHunter), and empirical cumulative distribution function (eCDF) plots of the number of repeated bases observed for each sample.

Gene Co-Expression Networks and Gene Set Enrichment Analyses

Normalised brain expression values (reads per kilobase of exon model per million mapped reads [RPKM]) from the BrainSpan Developmental Transcriptome dataset (15) (Gencode v10 summarised to genes) were used for the gene co-expression analyses. Samples were restricted to include those from all available brain regions, from fetal and infancy periods only (8 post conception weeks [pcw], to 10 months after birth; data for included samples are available [Dryad (Table S3): https://doi.org/10.5061/dryad.zkh189363]. Following sample restriction, genes were removed if they had expression values missing from >50% of samples, expression values of 0 RKPM for ≥50% samples, or variance of expression across all samples <0.5. 15,392 genes, across 280 samples from 24 individuals, remained in the filtered data set. Finally, expression values were $\log_2$ transformed.

Using the log transformed expression values, a matrix of weighted correlations was generated, with weights determined as $1/\sqrt{n}$, where n is the number of samples contributed by the respective individual. Correlation plots were visualised using the corrplot R package (Version 0.84, available at https://github.com/taiyun/corrplot), with genes ordered by hierarchical clustering, using the median linkage method. Networks of the most highly co-expressed genes were constructed using the qgraph R package (16). Using the distribution of pairwise correlations of all 15,392 genes in the dataset, a threshold of $|\rho| > 0.647$ was determined, corresponding to the absolute correlation value which the 5% most highly

correlated genes exceeded. Networks were then constructed with edges drawn between genes with absolute pairwise correlations above this threshold.

Finally, we determined whether these genes were more highly co-expressed than would be expected for a random set of genes. Given the very large number of combinations of gene sets possible, selected from the full set of 15,392, we utilised a Monte Carlo sampling approach to approximate the distribution of the median $|\rho|$ for all sets of genes. To this end, we randomly sampled 5000 sets of genes, the same size as our high confidence set, and calculated the median $|\rho|$ for each random gene set. We derived an empirical cumulative distribution function (eCDF) based on these medians, to which we compared the observed median $|\rho|$ of our high confidence candidates. Replication of all co-expression analyses was undertaken using independent samples (Supplemental Methods).

Gene set enrichment analyses were undertaken using g:Profiler (17), and utilising Gene Ontology molecular function, cellular component and biological processes databases, and KEGG and Reactome pathways (18, 19). A Bonferroni corrected p-value <0.05 was used to determine significant over-representation of our candidate genes in a pathway.

Data Availability

Data not available in this article is available on Dryad at: https://doi.org/10.5061/dryad.zkh189363.

RESULTS

Phenotypic data

34 probands (16 male), with a median age of 8 years (range 2years 9months to 16years 10months), including one monozygotic twin pair, were studied (Table 1, Figure 1, Figure 2). Feeding difficulties during infancy or during transition to solids were reported in 16 individuals. Early speech milestones were delayed in 33/34 individuals. 32 children had CAS, either in isolation (n=13), or co-occurring with other speech disorders of dysarthria (n=6), phonological delay or disorder (n=18), or articulation disorder (n=4) (Table 1). Two children (2, 31) ascertained for CAS had phonological disorders on testing, rather than CAS. Oral motor co-ordination and range of movement deficits occurred in 26. Poor performance during single non-speech oromotor movements reflected impaired lingual movements (e.g., reduced tongue elevation and lateralization), labial-facial movements (e.g., poor lip rounding), and mandibular control (e.g., reduced jaw excursion and stability). Impaired double non-speech oromotor movements (e.g., "smile and kiss") were also seen, typified by impaired transition, precision of movements and groping (overt struggle, effort or excessive excursion of the articulators) (Table 2). In seven children, expressive language could not be evaluated due to poor compliance (n=1) or severity of verbal impairment (n=6).

Hearing was normal in all except one child, who wore a hearing aid for unilateral low frequency sensorineural hearing loss. Two children had a history of severe recurrent otitis media necessitating grommet insertion. 10/34 (29%) patients had dysmorphic features (Table 1). Nineteen children had an IQ assessment showing average (n=1), low average (n=3), borderline (n=5), and extremely low average (n=5) FSIQ (Table 2). All but four children were attending mainstream schools. For five children, a full scale IQ (FSIQ) could not be calculated because of significant variability in performance across subscales. Sixteen children did not have IQ testing, largely due to young age (under age 5 years) or the family declined. Other features included: mild autism spectrum disorder (n=5), ADHD (n=2), difficulties with concentration (n=6), Tourette's syndrome (n=1), behavioural problems (n=5), and anxiety

and mood-related symptoms (n=2). Gross motor (n=24) and fine motor delays (n=26) were common with a slower trajectory in learning to ride a bike, balance appropriately, draw, write and cut compared to typical peers. Body praxis or Developmental Co-ordination Disorder diagnoses were reported in just two children. One 16 year old adolescent with a repaired cleft lip and palate had severe CAS with unintelligible speech not attributable to the cleft. Several children had a history of seizures; two had epilepsy, with one on valproate, two had febrile seizures, and a further two had unconfirmed seizures. Six probands had MRI brain abnormalities including: mild thinning of the corpus callosum (case 3), non-specific frontal gliosis (case 4), foci of white matter hyperintensity in bilateral parietal and posterior fossa (case 17), right medial frontal gyrus (case 18), 1 small focus of subcortical hyperintensity (case 30) and delayed frontal lobe myelination (case 20). 23/34 children had delayed independent toileting. All cases were receiving or had received speech therapy.

Copy Number Analysis and Short Tandem Repeats

Chromosomal microarray testing was performed in all patients. Only one proband (case 6) had a significant finding with a *de novo* mosaic deletion of approximately 9.2 megabases on chromosome 5q14.3q21.1 in about 75% of cells (genomic coordinates GRCh37/Hg19 chr5:90,779,680-99,959,810) (Figure 3, Table 3). We additionally searched for evidence of expansions of known pathogenic STRs. Most disorders caused by expanded STRs affect the nervous system and often include speech problems such as dysarthria. We found no evidence for an expanded STR in any patient.

Exome and Genome Sequence Analysis We identified candidate variants in 21/34 (62%) patients (Table 3, Figure 2, Figure 3). We found twelve high confidence variants - five were missense, three frameshift, three nonsense (stop gain) in 10 genes (*CDK13* [MIM: 603309], *EBF3* [MIM: 607407], *GNAO1* [MIM: 139311], *GNB1* [MIM: 139380], *DDX3X* [MIM:

300160], *MEIS2* [MIM: 601740], *POGZ* [MIM: 614787], *SETBP1* [MIM: 611060], *UPF2* [MIM: 605529], *ZNF142* [MIM: 604083]), and a large mosaic deletion (5q14.3q21.1) by chromosomal microarray. Nine high confidence variants were confirmed *de novo* dominant, one pair were recessively inherited (compound heterozygous) and, for one, inheritance could not be assessed by segregation analysis as the proband was adopted. All variants were novel, except for one of the compound heterozygous variants, according to the gnomAD database (Table 3.a, Figure 2). The six nonsense or frameshift variants were all in genes (*DDX3X*, *EBF3*, *GNB1*, *MEIS2*, *SETBP1*, *UPF2*) intolerant to LoF variation, according to ExACpLI and/or LoFtool scores. The five missense variants were all predicted to be damaging by three *in silico* tools (SIFT, PolyPhen and CADD). All twelve variants were classified as pathogenic according to ACMG guidelines (14).

In 9/34 (26%) probands, we found very rare (<0.05%) missense variants predicted to be damaging by multiple *in silico* tools (Table 3) [full list of predicted damaging candidates are available from Dryad (Table S5): https://doi.org/10.5061/dryad.zkh189363]. This list included variants in *BRWD3* [MIM: 300553], *UBA6* [MIM: 611361], *PTBP2* [MIM: 608449], *ZKSCAN1* [MIM: 601260], *TENM4* [MIM: 610084] and *ASTN2* [MIM: 612856] (Table 3.b). We also identified rare variants in *GRIN2A* [MIM: 138253], implicated in epilepsy-aphasia syndromes (5), and *KIRREL3* [MIM: 607761] in non-syndromic intellectual disability (*KIRREL3*); but these variants did not meet our strict criteria for predicted damaging candidates.

In a further four probands, we identified five novel or very rare LoF variants in genes predicted to be intolerant to variation, which were classified as of uncertain significance for CAS (Table 3.c). These variants are all predicted to be amongst the most damaging in these

probands; however, none of these genes have been implicated in CAS or neurodevelopmental disorders to date.

Gene Co-Expression During Brain Development

Using brain expression (RNA-seq) data from BrainSpan, we examined co-expression of our ten high confidence candidate genes (Figure 4a). The median absolute correlation between our ten high confidence candidate genes was $|\rho| = 0.463$, and 10 out of the 45 pairwise correlations were amongst the top 5% most highly correlated gene pairs genome-wide ($|\rho| > 0.647$, Figure 4.b). Using a Monte Carlo sampling approach, we found evidence that this set of genes was more highly co-expressed than expected by chance [$P = 0.006$, Data available from Dryad (Figure S1): https://doi.org/10.5061/dryad.zkh189363]. This suggests that these genes form part of a common pathway impacted in CAS, empirically captured by our results. When expanding the co-expression analyses to include the eight candidate genes for CAS in Eising et al. (8), we found strong overlap in co-expression patterns between these genes and our ten high confidence candidates [Figure 4.c.; Data available from Dryad(Figure S2): https://doi.org/10.5061/dryad.zkh189363]. This set of 18 genes had a median correlation that was significantly higher than expected [median $|\rho| = 0.463$, $P = 2 \times 10^{-4}$;Data available from Dryad (Figure S3): https://doi.org/10.5061/dryad.zkh189363], giving evidence of even better capture of our hypothesised biological network/pathway, and providing the first evidence of validation of the Eising et al. results (8).

Gene set enrichment analyses of our ten novel genes highlighted that there was an over-representation of genes (*CDK13*, *DDX3X*, *EBF3*, *MEIS2*, *POGZ*, *SETBP1*, *UPF2* and *ZNF142*) involved in DNA binding [GO:0003677; Data available from Dryad (Table S6, Figure S4 a): https://doi.org/10.5061/dryad.zkh189363]. The remaining two genes (*GNAO1*

and *GNB1)* are part of the heterotrimeric G-protein complex [GO:0005834; Data available from Dryad (Table S6, Figure S4 b): https://doi.org/10.5061/dryad.zkh189363].

DISCUSSION

We describe the molecular genetic architecture of CAS, a rare and debilitating disorder, in the largest cohort of children studied to date. We identified pathogenic variants in one third (11/34) of the cohort, newly implicating 9 genes (*CDK13*, *EBF3*, *GNAO1*, *GNB1*, *DDX3X*, *MEIS2*, *POGZ*, *UPF2*, *ZNF142*) and providing the first confirmation of the tenth (*SETBP1*) (8). We expand the phenotypic spectra for these genes, to included speech difficulties in the absence of, or with mild, intellectual disability. All except *ZNF142* have been previously reported with more severe phenotypes of syndromic or non-syndromic intellectual disability (*CDK13 (20), DDX3X (21), EBF3 (22), GNB1 (23), GNAO1 (24), MEIS2 (25), POGZ (26), SETBP1 (27), UPF2 (28)*). Broad speech and language deficits were noted, but not precisely phenotyped, in these single gene studies. A further two genes (*CHD1*, *NR2F1*), located within a contiguous gene deletion at 5q14.3-21.1 that includes 18 genes, are also potential candidates. *CHD1* has been linked to CAS in a previous report, and is part of a gene family of chromatin remodellers linked to neurodevelopmental disorders (e.g. *CHD2*, *CHD3* and *CHD8) (29)*, while *NR2F1* is associated with an optic atrophy and intellectual disability syndrome for which a variety of speech and language phenotypes (e.g. speech delay, expressive language deficits) have been described (30).

Our gene set enrichment analyses show that eight of these ten genes code for DNA binding proteins and play a role in transcriptional regulation. Using RNA-seq data from the brain, we empirically determined that these same eight genes are also strongly co-expressed in the developing brain, across multiple brain regions. Furthermore, we found evidence of co-

expression between the candidate genes reported here, and genes previously implicated in CAS, by the Eising et al. study (8). These findings suggest there is at least one distinct network of co-expressed genes emerging from molecular screening of CAS, characterised by similar function and patterns of expression in the brain. Similar observations of gene co-expression networks have been made for other disorders, such as the epileptic encephalopathies (31), leading to identification and then validation of candidate genes. This approach may also be productive to identify molecular determinants for CAS in future studies. Understanding why and how mutations of genes in this network result in CAS requires *in vitro* and *in vivo* functional studies.

Beyond our ten high confidence candidate genes, variants of unknown significance were identified in a further 10 genes (Table 3.b&c). *ASTN2, BRWD3, GRIN2A, KIRREL3* and *PTBP2* have been implicated in neurodevelopmental disorders (5, 32-35). Our remaining variants of unknown significance occur in genes associated with brain development and dysfunction. The protein encoded by *TENM4* plays a role in establishing neuronal connectivity during development, and mutations cause essential tremor (36). *ZKSCAN1* encodes a transcription factor that regulates expression of the $GABA_A$ receptor GABRB3 subunit essential for fast inhibitory neurotransmission in brain. *AAK1* [MIM: 616405] has established roles in dendritic arborization and spine development. *PHKA1* [MIM: 311870] causes glycogen storage disease type IX [MIM: 300559], an X-linked recessive metabolic disorder characterised by exercise-induced muscle weakness. Homozygous mutations in *ATP7B* [MIM: 606882] cause Wilson disease [MIM: 277900], a disorder characterised by excess storage of intracellular hepatic copper and neurologic abnormalities; however, these patients usually present in adolescence or later.

These disparate protein functions highlight the challenges associated with determining the significance of gene variants discovered in genome-wide screens of large cohorts, particularly for neurodevelopmental speech and language disorders (8), as is it well known that benign variants will also be found. Many were missense variants; definitively determining the pathogenicity of this variant class is often challenging. In interpreting their significance, we applied the convention of using the ACMG guidelines (14); however, these guidelines are more difficult to apply to genes for a novel phenotype that has not yet been studied extensively with next-generation sequencing, and they may be too conservative. Ongoing observations of phenotype-genotype correlations will be critical to determining the relevance of each variant, together with large curated databases of clinical and molecular information.

In this comprehensively phenotyped cohort of children with CAS, we describe a range of co-occurring neurodevelopmental features (Figure 1, Table 1, Table 2). Feeding challenges were common in the early years and the trajectory of speech development was delayed and aberrant, consistent with previous reports (9). Our data support the concept that CAS is often part of a more wide-ranging neurodevelopmental disorder, rather than isolated speech impairment (3, 8). All probands had additional deficits, that could involve a range of domains, including motor skills, cognition, attention, behavior, emotional regulation, toileting or social skills. There were no obvious differences between the phenotypes of children with solved molecular genetic diagnoses compared to those with uncertain or no genetic findings.

A novel finding was the high rate of co-occurrence of delays in fine and gross motor skills in our CAS cohort. Children had challenges with learning specific motor skills beyond speech, such as riding a bike, or learning to write. Gross and fine motor skills resolved earlier than the persisting speech deficits however, and only two children had formal diagnoses of motor

dyspraxia or DCD. Deficits in implicit motor learning (procedural learning) have long been proposed as a potential root cause for CAS (37) and other specific speech or language deficits (38). In CAS, the procedural deficit hypothesis proposes that children fail to automatize the ability to sequence sounds into words and words into phrases with little cognitive effort (37). Further to motor planning and programming deficits however, co-occurring neuromuscular tone involvement was seen in some children, or even ataxia in one, suggesting additional cerebellar or other common motor pathway deficits for at least one subgroup. Whilst there is increasing evidence linking motor ability with speech outcomes (39), whether motor skills are causative for, or simply correlate with speech outcomes, is yet to be elucidated. Attention issues were also noted in 8 probands and one child had Tourette's syndrome; these conditions have also been linked to the procedural learning hypothesis. A number of children had cognitive involvement, with more generalised learning deficits, beyond implicit learning. As acknowledged earlier, many of the genes identified here have been linked to intellectual disability (ID) and/or other health and medical conditions, including epilepsy and autism, and as such, these co-morbidities could play a role in the aetiology of CAS. Although not all children with epilepsy, ID, autism, ADHD or DCD present with CAS, so rather, we posit that there are several neurobiological subtypes of CAS that are more closely correlated with some neurodevelopmental conditions than others.

In summary, we provide novel insights into the aetiology of CAS. We show that CAS is highly genetically heterogeneous, often occurring as a sporadic monogenic disorder. Inheritance is most frequently *de novo* dominant, although recessive and mosaic variants can also arise. One-third of patients have pathogenic variants, implicating shared pathways in transcriptional regulation. These findings highlight the key role of transcriptional regulation in normal speech development.

ACKNOWLEDGEMENTS

We thank the families for their participation in this study. Tim Green (Epilepsy Research Centre) is acknowledged for performing DNA extractions.

APPENDIX 1

Name	Location	Role	Contribution
Michael S. Hildebrand, PhD	The University of Melbourne, Australia	Author	Designed and conceptualized study; directed project; generated data; analyzed data; wrote manuscript
Victoria E. Jackson, PhD	The Walter and Eliza Hall Institute of Medical Research, Australia	Author	Generated data; analysed data; interpreted data; wrote manuscript
Thomas S. Scerri, PhD	The Walter and Eliza Hall Institute of Medical Research, Australia	Author	Generated data; analysed data; interpreted data; wrote manuscript
Olivia Van Reyk, MSpPath	Murdoch Children's Research Institute, Australia	Author	Generated data; analyzed data
Matthew Coleman, BSc (Hons)	The University of Melbourne, Australia	Author	Generated data; analyzed data
Ruth O. Braden, MSpPath	Murdoch Children's Research Institute, Australia	Author	Generated data; analyzed data
Samantha Turner, PhD	Murdoch Children's Research Institute, Australia	Author	Generated data; analyzed data
Kristin A. Rigbye, BSc (Hons)	The University of Melbourne, Australia	Author	Generated data; analyzed data
Amber Boys, BSc Hons	Victorian Clinical Genetics Services, Australia	Author	Generated data; analyzed data
Sarah Barton, DPsych	Murdoch Children's Research Institute, Australia	Author	Generated data; analyzed data
Richard Webster, MD	The Children’s Hospital Westmead, Australia	Author	Generated data; analyzed data
Michael Fahey, MD	Monash University, Australia	Author	Generated data; analyzed data

PhD			
Kerryn Saunders, MD	Monash University, Australia	Author	Generated data; analyzed data
Bronwyn Parry-Fielder, BAppSci	The Royal Children's Hospital, Australia	Author	Generated data; analyzed data
Georgia Paxton, MD	The Royal Children's Hospital, Australia	Author	Generated data; analyzed data
Michael Hayman, MD	The Royal Children's Hospital, Australia	Author	Generated data; analyzed data
David Coman, MD	The Wesley Hospital, Australia	Author	Generated data; analyzed data
Himanshu Goel, MD	John Hunter Hospital, Australia	Author	Generated data; analyzed data
Anne Baxter, MD	John Hunter Hospital, Australia	Author	Generated data; analyzed data
Alan Ma, MD	The Children's Hospital Westmead, Australia	Author	Generated data; analyzed data
Noni Davis, MD	Melbourne Children's Clinic, Australia	Author	Generated data; analyzed data
Sheena Reilly, PhD	Griffith University, Australia	Author	Generated data; analyzed data
Martin Delatycki, MBBS PhD	Victorian Clinical Genetics Services, Australia	Author	Generated data; analyzed data
Frederique J. Liégeois, PhD	UCL Great Ormond Street Institute of Child Health, UK	Author	Generated data; analyzed data
Alan Connelly, PhD	Florey Institute of Neuroscience and Mental Health, Australia	Author	Generated data; analyzed data
Jozef Gecz, PhD	University of Adelaide, Australia	Author	Generated data; analyzed data
Simon E. Fisher, PhD	Max Planck Institute for Psycholinguistics, The Netherlands	Author	Generated data; analyzed data
David J. Amor MBBS PhD	Murdoch Children's Research Institute, Australia	Author	Generated data; analyzed data
Ingrid E. Scheffer, MBBS PhD	The University of Melbourne, Australia	Author	Designed and conceptualized study; directed project; wrote manuscript
Melanie Bahlo, PhD	The Walter and Eliza Hall Institute of Medical Research, Australia	Author	Designed and conceptualized study; directed project; wrote manuscript
Angela T. Morgan, PhD	Murdoch Children's Research Institute, Australia	Author	Designed and conceptualized study; directed project;

			generated data; analyzed data; interpreted data; wrote manuscript

REFERENCES

1. Reilly S, McKean C, Morgan A, Wake M. Identifying and managing common childhood language and speech impairments. BMJ 2015;350:h2318.
2. Morgan A, Ttofari Eecen K, Pezic A, Brommeyer K, Mei C, Eadie P, et al. Who to Refer for Speech Therapy at 4 Years of Age Versus Who to "Watch and Wait"? J Pediatr 2017;185:200-204.e201.
3. Morgan AT, Webster R. Aetiology of childhood apraxia of speech: A clinical practice update for paediatricians. J Paediatr Child Health 2018 Oct;54(10):1090-1095.
4. Graham SA, Fisher SE. Understanding language from a genomic perspective. Annu Rev Genet 2015;49:131-160.
5. Carvill GL, Regan BM, Yendle SC, O'Roak BJ, Lozovaya N, Bruneau N, et al. GRIN2A mutations cause epilepsy-aphasia spectrum disorders. Nat Genet 2013;45(9):1073-1076.
6. Morgan AT, van Haaften L, van Hulst K, Edley C, Mei C, Tan TY, et al. Early speech development in Koolen de Vries syndrome limited by oral praxis and hypotonia. Eur J Hum Genet 2018;26(1):75-84.
7. Mei C, Fedorenko E, Amor DJ, Boys A, Hoeflin C, Carew P, et al. Deep phenotyping of speech and language skills in individuals with 16p11.2 deletion. Eur J Hum Genet 2018;26(5):676-686.
8. Eising E, Carrion-Castillo A, Vino A, Strand EA, Jakielski KJ, Scerri TS, et al. A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. Mol Psychiatry 2019; 24(7):1065–1078.
9. Morgan AT, Fisher SE, Scheffer IE, Hildebrand MS. FOXP2-Related Speech and Language Disorders. In: Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJH, et al., editors. GeneReviews(R). Seattle: University of Washington, Seattle; 2017.
10. Dodd B, Hua Z, Crosbie S, Holm A, Ozanne A. Diagnostic Evaluation of Articulation & Phonology (DEAP). London, UK: Pearson Assessment; 2002.
11. Gozzard H, Baker E, McCabe P. Requests for clarification and children's speech responses: changing 'pasghetti' to 'spaghetti'. Child Language Teaching and Therapy 2008;24(3):249-263.
12. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25(14):1754-1760.
13. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297-1303.
14. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015;17(5):405-423.
15. Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, et al. Transcriptional landscape of the prenatal human brain. Nature 2014;508(7495):199.

16. Epskamp S, Cramer AO, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: Network visualizations of relationships in psychometric data. J Stat Softw 2012;48(4):1-18.
17. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc 2019;14(2):482-517.
18. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27-30.
19. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2017;46(D1):D649-D655.
20. Sifrim A, Hitz MP, Wilsdon A, Breckpot J, Turki SH, Thienpont B, et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. Nat Genet 2016;48(9):1060-1065.
21. Beal B, Hayes I, McGaughran J, Amor DJ, Miteff C, Jackson V, et al. Expansion of phenotype of DDX3X syndrome: six new cases. Clin Dysmorphol 2019;28(4):169-174.
22. Harms FL, Girisha KM, Hardigan AA, Kortum F, Shukla A, Alawi M, et al. Mutations in EBF3 disturb transcriptional profiles and cause intellectual disability, ataxia, and facial dysmorphism. Am J Hum Genet 2017;100(1):117-127.
23. Petrovski S, Kury S, Myers CT, Anyane-Yeboa K, Cogne B, Bialer M, et al. Germline de novo mutations in GNB1 cause severe neurodevelopmental disability, hypotonia, and seizures. Am J Hum Genet 2016;98(5):1001-1010.
24. Nakamura K, Kodera H, Akita T, Shiina M, Kato M, Hoshino H, et al. De Novo mutations in GNAO1, encoding a Galphao subunit of heterotrimeric G proteins, cause epileptic encephalopathy. Am J Hum Genet 2013;93(3):496-505.
25. Verheije R, Kupchik GS, Isidor B, Kroes HY, Lynch SA, Hawkes L, et al. Heterozygous loss-of-function variants of MEIS2 cause a triad of palatal defects, congenital heart defects, and intellectual disability. Eur J Hum Genet 2018; 5;5(10):018-0281.
26. Stessman HAF, Willemsen MH, Fenckova M, Penn O, Hoischen A, Xiong B, et al. disruption of POGZ is associated with intellectual disability and autism spectrum disorders. Am J Hum Genet 2016;98(3):541-552.
27. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet 2010;42(6):483-485.
28. Johnson JL, Stoica L, Liu Y, Zhu PJ, Bhattacharya A, Buffington SA, et al. Inhibition of Upf2-dependent nonsense-mediated decay leads to behavioral and neurophysiological abnormalities by activating the immune response. Neuron EPub 2019 Oct 1.
29. Lamar KMJ, Carvill GL. Chromatin Remodeling Proteins in Epilepsy: Lessons From CHD2-Associated Epilepsy. Front Mol Neurosci 2018;11:208-208.
30. Bosch DG, Boonstra FN, Gonzaga-Jauregui C, Xu M, de Ligt J, Jhangiani S, et al. NR2F1 mutations cause optic atrophy with intellectual disability. Am J Hum Genet 2014;94(2):303-309.
31. Oliver KL, Lukic V, Thorne NP, Berkovic SF, Scheffer IE, Bahlo M. Harnessing gene expression networks to prioritize candidate epileptic encephalopathy genes. PloS one 2014;9(7):1-11.
32. Doan RN, Bae BI, Cubelos B, Chang C, Hossain AA, Al-Saad S, et al. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell 2016;167(2):341-354.

33. Field M, Tarpey PS, Smith R, Edkins S, O'Meara S, Stevens C, et al. Mutations in the BRWD3 gene cause X-linked mental retardation associated with macrocephaly. Am J Hum Genet 2007;81(2):367-374.
34. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 2009;459(7246):569-573.
35. Quintela I, Barros F, Fernandez-Prieto M, Martinez-Regueiro R, Castro-Gago M, Carracedo A, et al. Interstitial microdeletions including the chromosome band 4q13.2 and the UBA6 gene as possible causes of intellectual disability and behavior disorder. Am J Med Genet A 2015;167A(12):3113-3120.
36. Hor H, Francescatto L, Bartesaghi L, Ortega-Cubero S, Kousi M, Lorenzo-Betancor O, et al. Missense mutations in TENM4, a regulator of axon guidance and central myelination, cause essential tremor. Hum Mol Genet 2015;24(20):5677-5686.
37. Vargha-Khadem F, Gadian DG, Copp A, Mishkin M. FOXP2 and the neuroanatomy of speech and language. Nat Rev Neurosci 2005;6(2):131-138.
38. Ullman MT, Pierpont EI. Specific language impairment is not specific to language: the procedural deficit hypothesis. Cortex 2005;41(3):399-433.
39. Visscher C, Houwen S, Scherder EJ, Moolenaar B, Hartman E. Motor profile of children with developmental speech and language disorders. Pediatrics. 2007;120(1):e158-e163.
40. Khan K, Zech M, Morgan AT, Amor DJ, Skorvanek M, Khan TN, et al. Recessive variants in ZNF142 cause a complex neurodevelopmental disorder with intellectual disability, speech impairment, seizures, and dystonia. Genet Med 2019; 21:2532-2542.

FIGURE LEGENDS

Figure 1 Summary of Phenotypic Overlap in CAS Cohort

FSIQ: Full Scale Intelligence Quotient < 70.

Figure 2 Families with High Confidence Variants

Families 1-6 analysed by Whole Exome Sequencing. Pedigrees from 6 families showing segregation of 7 high confidence variants. Sequence chromatograms showing confirmed *de novo* variants in the probands of families 1, 2, and 4, and confirmed compound heterozygous variants in the proband of family 3. Sanger sequencing was not performed for the variant in family 5, and the proband in family 6 had a large deletion as shown in Figure 3. Families 7-11 analyzed by Whole Genome Sequencing. Pedigrees from 5 families showing 5 high confidence variants. Sequence chromatograms showing confirmed *de novo* variants in the

probands of families 7, 8, 10 and 11. The proband in family 9 was adopted and her biological parents were unavailable for testing.

Figure 3 Large Mosaic Deletion in Family 6

Illumina Karyostudio image showing the Illumina Infinium Global Screening Array-24v1.0 SNP data for chromosome 5. The Smoothed Log R (representing copy number) is depicted as a red line, and the B allele frequency (representing genotyping) is depicted as blue dots. The mosaic 9.2 Mb deletion of chromosome 5q14.3q21.1 is observed as a negative shift in the Smoothed Log R and a change in the genotyping at 5q14.3 to q21.1. The deletion is present in approximately 75% of cells.

Figure 4 Gene Regulation Network for Speech Development

a. Gene co-expression matrix for the 10 high confidence candidate genes. Pairwise Spearman Correlations between genes shown, based on 280 samples from 24 individuals (8 weeks post conception to 10 months after birth) from the BrainSpan resource. Genes ordered by hierarchical clustering, using the median linkage method.

b. Network of gene co-expression. Nodes represent genes; edges represent gene-pair correlations, that exceed the threshold for the top 5% most highly correlated gene pairs genome-wide ($|\rho| > 0.64$).

c. Gene co-expression matrix for the 10 high confidence candidate genes and the Eising et al. (8) genes.

Table 1. Medical and Neurodevelopmental Features of CAS Cohort

Case	Age	Sex (M/F)	Core Speech Phenotype	Gross-motor delays	Fine- motor delays	Vision impaired	Hearing loss	MRI findings	Seizures	Other NDD	Toileting delays	Dysmorphic features	Other medical
1	8;11	F	CAS, dysarthria	Y	Y	N	N	N	Febrile seizures	N	Y	N	NR
2	11;5	M	Severe phon.	N	N	Glasses	N	NA	N	N	Y	Clinodactyly 5th fingers	Asthma, eczema
3	5;0	F	CAS, phon. delay and disorder	Y	Y	N	N	Mild thinning posterior CC, reduced WM	N	Attention deficits	N	N	NR
4	6;7	F	CAS, phon. Delay, arti.disorder	Y	Y	N	N	Non-specific frontal gliosis	Bilateral temporal discharges at 6y	Attention deficits	N	Retrognathia	NR
5	4;8	M	CAS, dysarthria	Y	Y	N	N	NA	N	Behavioural problems due to speech frustration	Y	N	Ataxia
6	8;9	F	CAS, phon.delay, artic.disorder	Y	Y	N	N	N	N	NA	Y	Narrow palepbral fissures, arched eyebrows, low columnella, hypoplastic alar nasae.	NR
7	11;3	F	CAS	Y	Y	N	N	NA	N	Learning deficits	N	High nasal root, prominent nose, thin upper lip	Atrial SD
8	5;1	F	CAS, phon. delay and disorder	Y	Y	N	N	NA	N	Learning deficits	Y	Brachycephaly, flat midface, anterverted nares, cupid's bow upper lip	NR
9	16;10	F	CAS	Y	Y	Glasses	N	NA	N	Mild ASD, Auditory processing deficits	Y	Arched eyebrows, sparse laterally, cleft lip and palate repair	NR
10	9;1	F	CAS, dysarthria, phon. delay	Y	Y	Glasses	N	N	NR	Mild ASD[#]	Y	Brachycephaly, small mouth, thin upper lip	Mastocytosis, L hemiplegia
11	4y	M	CAS	Y	Y	N	N	NA	N	Mild ASD	Y	Cupid's bow upper lip, hypoplastic columnella	Cystoscopy + retrograd pyelogram), L pelvic kidney w/o sig. reflux
12	8	M	CAS	Y	Y	N	N	N	N	Mild ASD, ADHD	Y	NR	NR
13	6;9	M	CAS, phon. delay and disorder	Y	Y	N	N	NA	N	ADHD, Tourettes	Y	NR	NR
14	6;11	M	CAS, phonological delay	N	N	N	N	N	N	NA	N	N	Coeliac HLA DQ8 haplotype
15	7;9	M	CAS, phon. delay	N	N	N	N	NA	N	N	N	Triangular face, anteverted ears, broad nasal root.	NR
16	4;4	M	CAS, phon.delay and disorder	Y	Y	N	N	NA	N	Attention deficits	Y	N	NR
17	11;1	M	CAS, phon. delay	Y	Y	N	N	Multiple foci hyper-intensity	N	Mild ASD, ADD, anxiety &	Y	N	NR

								subcortical WM		depression			
18	14;1	F	CAS, dysarthria, artic.disorder	Y	Y	N	N	WM hyperintensity below R MFG	N	Attentional & emotional deficits, anxiety& depression	Y	N	Overbite, braces
19	2;9	M	CAS	N	N	N	N	NA	N	N	Y	N	NR
20	11;11	F	CAS, dysarthria, phon. delay and disorder, artic. disorder	Y	Y	N	N	Delayed frontal lobe myelination	N	Motor dyspraxia	Y	N	NR
21	6;8	M	CAS	Y	Y	N	R low freq. SNHL	N	N; discharges in sleep^	ID	N	Broad forehead, mild hypertelorism	NR
22	3;11	M	CAS	Y	N	N	N	N	4 febrile seizures	N	Y	N	NR
23	5;9	F	CAS, phon. delay and disorder, artic. disorder	N	Y	N	N	NA	N	N	Y	N	NR
24	5	M	CAS	Y	Y	N	N	N	N	DCD, behavioural deficits	Y	NR	Peanut allergy
25 (a)	4	M	CAS	N	Y	N	N	NA	N	NA	Y	NA	Tongue- tie
25 (b)	4	M	CAS	N	Y	N	N	NA	N	NA	Y	NA	Tongue tie
26	5	M	CAS	Y	Y	N	N	NA	N	NA	N	NR	NR
27	4;8	M	CAS, phon. delay and disorder, artic. disorder	N	N	N	N	NA	N	N	N	N	NR
28	8;0	F	CAS, phon.delay and disorder	Y	Y	N	N	NA	N	Attention deficits	Y	Large upturned earlobes, brachydactyly, 2,3 toe syndactyly, metacarpal & metatarsal shortening	Central obesity, insulin resistance
29	6;5	F	CAS	Y	Y	N	N	N	N	N	N	N	NR
30	7;8	F	CAS, dysarthria, phon. delay	N	Y	N	N	1 small focus subcortical hyper-intensity	2 normal EEGs	Mild ASD, Migraine, behavioural deficits	N	N	Obesity; sleep issues
31	4;0	M	Phonological delay, phon. disorder	Y	Y	N	N	NA	Jerking, 2 normal EEGs	N	Y	Glabellar flame naevus, full nasal root and tip, prominent tongue	NA
32	5;3	M	CAS	N	Y	N	N	NA	N	Learning deficits	Y	N	NR
33	4;10	F	CAS, phon. delay	Y	N	N	N	NA	N	N	N	N	Gluten intolerant

Y: Yes, N: No, NA: Not assessed, NR: Not reported, freq.:frequency, SNHL: sensorineural hearing loss, MRI: Magnetic resonance imaging, post.: posterior; CC: Corpus callosum, R: right, L: left, MFG: Medial frontal gyrus, WM: white matter, EEG: electroencephalogram, EAS: Epilepsy aphasia syndrome, ASD: autism spectrum disorder, NDD: neurodevelopmental disorder, ADHD: Attention deficit hyperactivity disorder, DCD: developmental coordination disorder, Atrial SD: Atrial septal defect, Lg.: large, SP: speech pathology, BMI: Body mass index, Phon.:, phonological, Artic.: articulation, ^not sufficient to cause EAS. # diagnosis reported to be 'debatable' by parent

Table 2. Extended linguistic phenotype and educational outcomes of CAS Cohort

Case	Oral motor impairment	History of feeding issues	Language - receptive	Language - expressive	Reading deficits	Spelling deficits	Speech pathology	Intelligence quotient (IQ)#	Education setting
1	Y	Y	Severe	Severe	Y	Y	Y	BDLN (FSIQ)	Specialist
2	N	Y	Mild	Severe	Low Average	Below Average	Y	Low AVG (FSIQ)	Mainstream
3	Y	N	Mild	Mild	Y	NA	Y	BDLN (FSIQ), Low AVG (Verbal IQ), Low AVG (NVIQ)	Mainstream
4	Y	Y	Severe	Severe	NA	NA	Y	Ext low (FSIQ), Ext low (Verbal IQ) Ext low (Performance Score)^	School for Deaf (because child was signing, but is not deaf)
5	Y	Y	Above average	NA - speech too severe to test	NA	NA	Y	NA	Not yet at school
6	Y	Y	Average	Severe	NA	NA	Y	Unable to calculate FSIQ (clinician concluded moderate impairment)	Mainstream then specialist
7	Y	N	Mild	Average	Y	Y	Y	Low AVG (FSIQ)	Mainstream
8	Y	N	Mild	Severe	NA	NA	Y	BDLN (FSIQ)	Mainstream kindergarten
9	Y	Y	Mod-severe	NA - speech too severe to test	Lower extreme	Y	Y	NA	Specialist
10	Y	Y	Severe	Severe	Y	Y	Y	BDLN (FSIQ), BDLN (Verbal Scale), Ext low (Performance Scale), BDLN (Process. Speed)	Mainstream
11	Y	N	Moderate	Severe	Y	Y	Y	NA	Not yet at school
12	Y	Y	NA	NA	NA	NA	Y	NA	Mainstream
13	NA	N	Average	Severe	High Average	High average	Y	Low AVG (FSIQ), AVG (Process. speed), BDLN (Working memory), AVG (Percept. reasoning), Low AVG (Verbal comp.)	Mainstream
14	Y	Y	Moderate	Severe	Average	Average	Y	NA	Mainstream
15	Y	N	Average	Moderate	Y	Y	Y	NA	Mainstream
16	Y	Y	Severe	Severe	NA	NA	Y	NA	Mainstream
17	N	N	Moderate	Moderate	Y	Y	Y	Ext low (FSIQ), Ext low (Verbal), Ext low (Process. speed) BDLN (NV)	Mainstream
18	Y	N	Mild	Severe	Y	Y	Y	Unable to calculate FSIQ. Low AVG (Verbal comp.), Ext low (Percept. reasoning), Ext low (Process. speed), BDLN (Working memory)	Mainstream
19	Y	Y	Above average	NA - speech too severe to	NA - too	NA - too young	Y	NA	Not yet at school

				test	young				
20	Y	Y	Severe	NA - speech too severe to test	Y	Y	Y	Ext low avg (FSIQ)	Mainstream
21	Y	N	Severe	Severe	NA	Y	Y	Ext low avg (FSIQ)	Mainstream
22	Y	N	Average	NA - speech too severe to test	NA	NA	Y	NA	Not yet at shool
23	Y	N	Average	Severe	NA	NA	Y	AVG (Verbal), Superior (NV), AVG (Process. Speed)	Mainstream
24	NA	N	Moderate	Moderate	NA	NA	Y	Unable to calculate FSIQ BDLN (Verbal), Low AVG-AVG (NV)	Mainstream
25 (a)	Y	Y	Mild	Moderate	NA	NA	Y	NA (PPVT WNL)	Mainstream kindergarten
25 (b)	Y	Y	Mild	Moderate	NA	NA	Y	NA (PPVT WNL)	Mainstream kindergarten
26	NA	Y	Average	Average	NA	NA	Y	NA	Mainstream Kindergarten (repeating kinder due to speech)
27	N	N	Average	Mild	NA	NA	Y	NA	Mainstream kindergarten
28	Y	N	Moderate	Severe	Y	Y	Y	BDLN (FSIQ)	Mainstream
29	NA	N	Severe	Average	NA	NA	Y	NA	Mainstream
30	Y	N	Average	Mild	Y	NA	Y	BDLN (Verbal Comp.), Low AVG (Percept. reasoning), Low AVG (Working Memory), AVG (Process. Speed)	Mainstream
31	N	Y	Average	NA - speech to severe to test	NA	NA	Y	Average	Mainstream
32	Y	N	Moderate	Severe	NA	NA	Y	Ext Low (FSIQ), Ext low (Verbal Comp.), Low AVG (Visual spatial, Fluid Reasoning, Working Memory), BDLN (Process. Speed)	Mainstream kindergarten
33	Y	N	Average	Mild	NA	NA	Y	NA	Mainstream kindergarten

Y: Yes, N: No, NA: Not assessed; BDLN: Borderline (70-79), AVG: average (90-109), Low AVG (80-89), Ext Low (69 and below); FSIQ: Full Scale IQ, NVIQ: Non-verbal IQ, Comp.: comprehension, Process.: processing; Percept.: perceptual, Ext: extremely; ^ Results from 3 years prior were less severe: i.e. Borderline (FSIQ 76), Low average (Verbal IQ), Borderline (Performance Score);*wide discrepancy in performance in nonverbal subtests and unable to complete verbal subtests due to severe speech impairment; PPVT: Peabody picture vocabulary test used as limited proxy for NVIQ. [#]IQ performance severity descriptors were converted to the same synonymous terms across tools for ease of comparison.

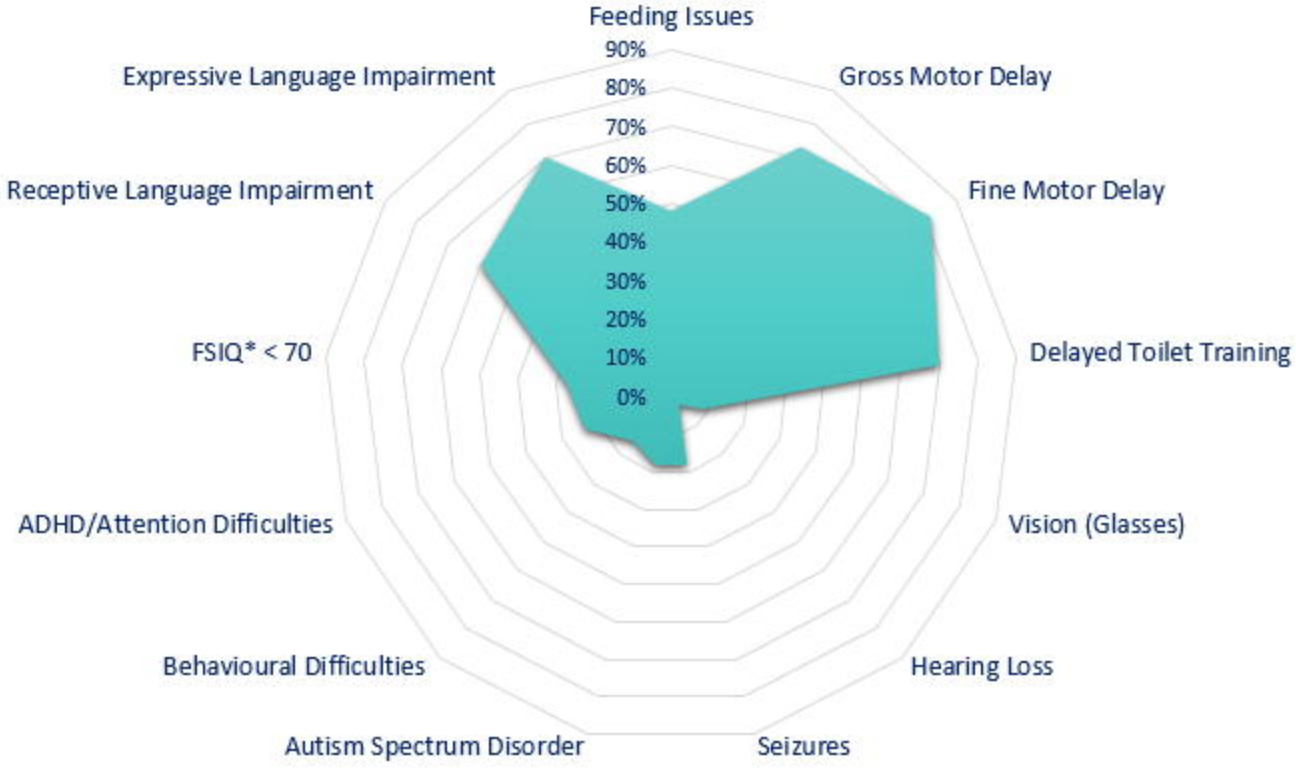
Table 3. Gene Variants in CAS Cohort

Case	Sex (M/F)	Method	Chr:Pos	Gene	DNA Variant	Protein Change	Effect	In Silico Predictions[$]	gnomAD Count [~]	Inheritance	ACMG score	Reference
a)	*High confidence variants - pathogenic variants according to ACMG guidelines*											
1	F	WES	18: 42531970	*SETBP1*	c.2665C>T	p.R889*	Nonsense	ExACpLI = 1; LoFtool = 0.0297; CADD= 38	0	*De novo*	PP3, PP4, PM2, PM4, PS2, PS3, PVS1, Class 5 Pathogenic	Eising, E., et al. (2018) #(7)
2	M	WES	10:12021068	*UPF2*	c.1940delA	p.F648Sfs*23	Frameshift	ExACpLI = 1	0	*De novo*	PP4, PM2, PM4, PS2, PS3?, PVS1, Class 5 Pathogenic	Johnson, JL., et al (2019)*(28)
3	F	WES	10:219507541, 10:219505483	*ZNF142*	c.3698G>T, c.4498C>T	p.C1233F, p.R1500W	Missense, Missense	SIFT = Del(0)/ Del(0) ; PolyPhen = Dam (0.998) / Dam (0.998) ; CADD= 31 / 26	0 1	Compound heterozygous	PP3, PP4, PM3, PS3?, PVS1, Class 5 Pathogenic	Khan K., et al 2019 *(40)
4	F	WES	16:56388880	*GNAO1*	c.980C>G	p.T327R	Missense	SIFT = Del(0); PolyPhen = Dam (1); CADD= 28.3	0	*De novo*	PP3, PP4, PM2, PS2, PVS1, Class 5 Pathogenic	-
5	M	WES	10:131666059	*EBF3*	c.872T>A	p.L291*	Nonsense	ExACpLI =0.999; LoFtool = 0.0389; CADD= 39	0	*De novo*	PP3, PP4, PM2, PM4, PS2, PVS1, Class 5 Pathogenic	-
6	F	WES & CMA		5q14.3q21.1 deletion	NA		LOH	NA	0	*De novo* mosaic	PP4, PM2, PS2, PVS1, Class 5 Pathogenic	-
7	F	WGS	7:40102433	*CDK13*	c.2609A>G	p.Y870C	Missense	SIFT = Del(0); PolyPhen = Dam (0.996); CADD= 32; MTR FDR = 0.031	0	*De novo*	PP3, PP4, PS2, PM2, PVS1, Class 5 Pathogenic	-
8	F	WGS	1:151379435	*POGZ*	c.2497C>A	p.H833N	Missense	SIFT = Del(0); PolyPhen = Dam (0.968); CADD= 28.2	0	*De novo*	PP3, PP4, PS2, PM2, PVS1, Class 5 Pathogenic	-
9	F	WGS	15:37242564	*MEIS2*	c.934_937delTTAG	p.L312Rfs*11	Frameshift	ExACpLI = 0.99; LoFtool = 0.091	0	Parents unavailable	PP3, PP4, PM2, PM4, PVS1, Class 5 Pathogenic	-
10	F	WGS	X:41205635	*DDX3X*	c.1470delA	p.S492Afs*4	Frameshift	ExACpLI = 1; LoFtool = 0.0555	0	De novo	PP3, PP4, PM2, PM4, PS2, PS3?, PVS1, Class 5 Pathogenic	Beal, B., et al 2019 *(21)
11	M	WGS	1:1721901	*GNB1*	c.632G>A	p.W211*	Nonsense	ExACpLI = 1; CADD= 40	0	De novo	PP3,PP4, PM2, PS2, PVS1, Class 5 Pathogenic	-
b)	*Predicted damaging variants classified as likely pathogenic, or with uncertain significance (ACMG guidelines)*											
12	M	WES	1:97216982	*PTBP2*	c.74G>C	p.R25T	Missense & splice region	SIFT = Del(0); PolyPhen = PosDam (0.641); CADD= 32; MTR FDR = 0.043; Ada = 0.981; RF = 0.886	0	De novo	PP3, PM2, PS2, Class 4 likely pathogenic	-
14	M	WES	16:9858387	*GRIN2A*	c.3014A>G	p.K1005R	Missense	CADD=21.8	0	Inherited from affected father	PP1, PP4, PM2, Class 3 uncertain significance	-

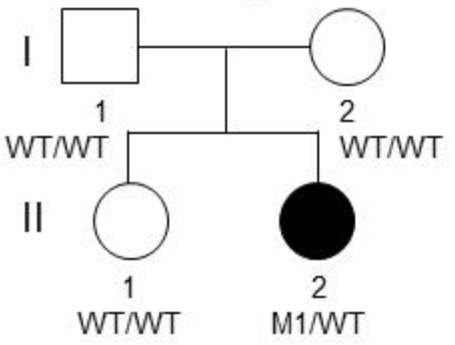
15	M	WES	11: 126294626	*KIRREL3*	c.2186G>T	p.S729I	Missense	CADD=23.7	1	Unconfirmed – father unavailable	PP1, PP4, Class 3 uncertain significance	-
16	M	WES	11:78614398, 11:78574177	*TENM4*	c.664G>A, c.1085C>T	p.G222R, p.A362V	Missense, Missense & splice region	PolyPhen = PosDam (0.877) / Dam (0.977); CADD= 24/32; Ada = NA/0.997; RF = NA/0.956	19 5	Compound heterozygous	PP3, PM3, Class 3 uncertain significance	-
17	M	WES	X:79958990	*BRWD3*	c.2824A>G	p.M942V	Missense	SIFT = Del(0.01); CADD= 23.5; MTR FDR = 0.034	0	X-linked hemizygous	PP3, PM2, Class 3 uncertain significance	-
18	F	WES	9:119204816	*ASTN2*	c.3361G>A	p.V1121M	Missense	SIFT = Del(0); PolyPhen = Dam (0.961); CADD= 33	0	Unconfirmed – father unavailable	PP3, PM2, Class 3 uncertain significance	-
19	M	WGS	3:67571051	*SUCLG2*	c.425T>C	p.V142A	Missense	SIFT = Del(0); PolyPhen = PosDam (0.733); CADD= 27.1	1	Unconfirmed - father unavailable	PP3, PP4, Class 3 uncertain significance	-
			4:68501247	*UBA6*	c.1766T>C	p.L589S	Missense	SIFT = Del(0); PolyPhen = Dam (0.979); CADD= 27.6	0	Unconfirmed - father unavailable	PP3, PM2, Class 3 uncertain significance	
20	F	WGS	7:99627930	*ZKSCAN1*	c.731A>G	p.Q244R	Missense	PolyPhen = PosDam (0.877); CADD= 24	0	De novo	PP3, PS2, PM2, Class 4 likely pathogenic	-
	c)	*Predicted LoF variants classified as likely pathogenic, or with uncertain significance (ACMG guidelines)*										
12	M	WES	21:46309189	*ITGB2*	c.1877+2T>C	NA	splice donor	LoFtool = 0.0333; CADD= 25.6; Ada = 0.999; RF = 0.652	0	Inherited from affected father	PP1, PM2, Class 3 uncertain significance	-
13	M	WES	2:69734646	*AAK1*	c.2071G>T	p.E691*	Nonsense	ExACpLI = 1; CADD= 38	0	De novo	PS2, PM2, PM4, Class 4 likely pathogenic	-
14	M	WES	13:52532497	*ATP7B*	c.2304dupG	p.M769Hfs *26	frameshift	LoFtool = 0.034; CADD= 34	32	Inherited from affected father	PP1, PM4, Class 3 uncertain significance	-
			10:121602918	*MCMBP*	c.847delG	p.D283Ifs* 21	frameshift	ExACpLI = 1	0	Inherited from affected father	PP1, PM2, PM4, Class 3 uncertain significance	-
22	M	WES	X:71855117	*PHKA1*	c.1601delT	p.L534Rfs *5	frameshift	LoFtool = 0.0318	0	X-linked hemizygous	PM2, PM4, Class 3 uncertain significance	-

*Only 22 reported here as no variants met criterion for remaining probands in cohort.

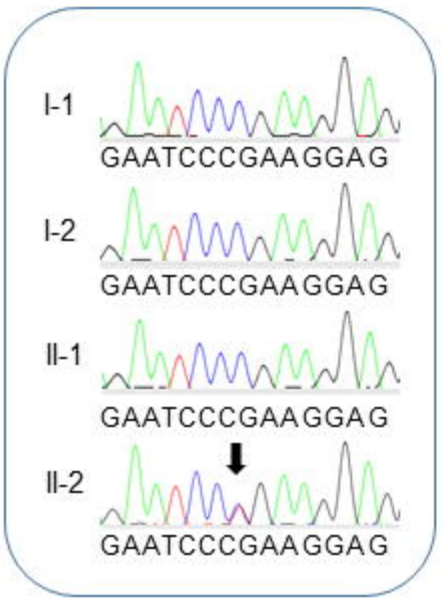
CAS, childhood apraxia of speech; WGS, whole genome sequencing; WES, whole exome sequencing; CMA, chromosomal microarray; LOH, loss of heterozygosity; NA, not applicable; ND, none detected. All coordinates correspond to the Homo sapiens (human) genome assembly GRCh37 (hg19) from Genome Reference Consortium. All variants were confirmed by Sanger sequencing. ^ Identical twins, $ *In silico* pathogenicity predictions reported, only if in support of pathogenicity: SIFT (sorting intolerant from tolerant), scores <0.05 reported, Del="Deleterious"; PolyPhen-2, scores >0.15 reported, Dam="Damaging", PosDam="Possibly Damaging"; CADD (Combined Annotation Dependent Depletion), Phred-scaled scores >= 20 reported; MTR (Missense Tolerance Ratio), FDR < 0.05 reported; Ada (Ada Boost prediction for effect on splicing), score >= 0.6 reported; RF (random forest algorithm for effect on splicing) score >= 0.6 reported; ExACpLI (The Exome Aggregation Consortium (ExAC) probability of intolerance to LoF), scores >0.9 reported; LoFTool, scores <0.1 reported, ¯ Number of alleles for variant from gnomAD, # Published with additional families described by collaborators, * Collaborative paper with additional families under review.


Feeding Issues
Gross Motor Delay
Fine Motor Delay
Delayed Toilet Training
Vision (Glasses)
Hearing Loss
Seizures
Autism Spectrum Disorder
Behavioural Difficulties
ADHD/Attention Difficulties
FSIQ* < 70
Receptive Language Impairment
Expressive Language Impairment
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%

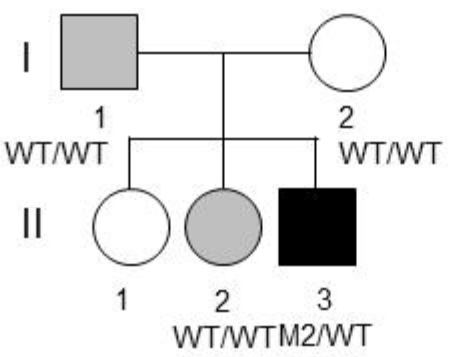
Family 1



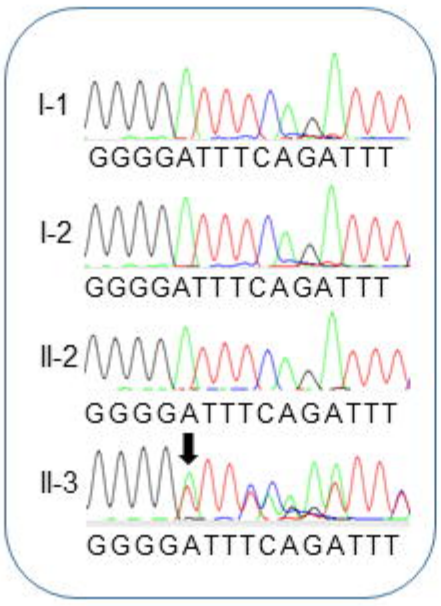
M1:
***SETBP1* c.2665C>T**



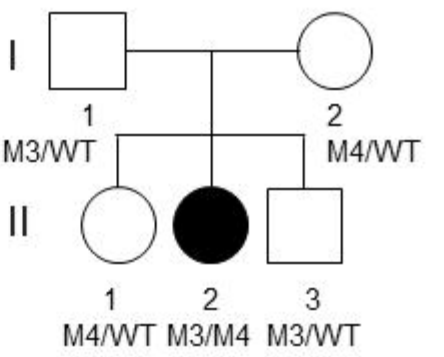
Family 2



M2:
***UPF2* c.1940delA**

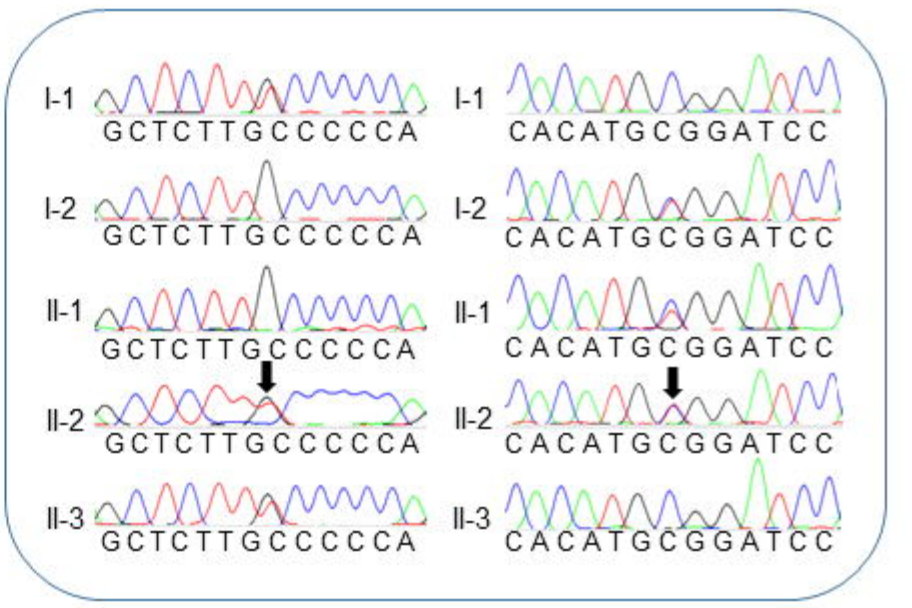


Family 3

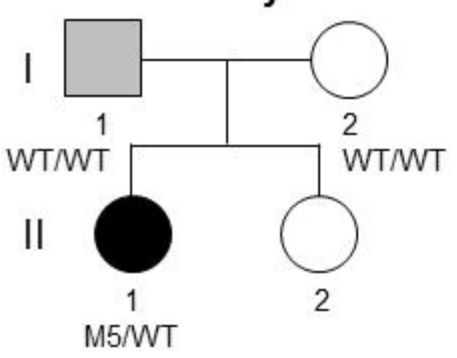


M3:
***ZNF142* c.3698G>T**

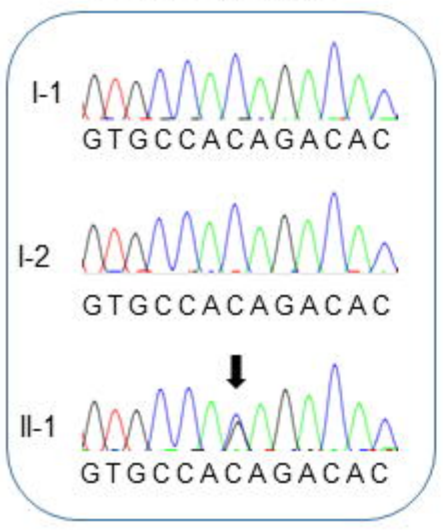
M4:
***ZNF142* c.4498C>T**



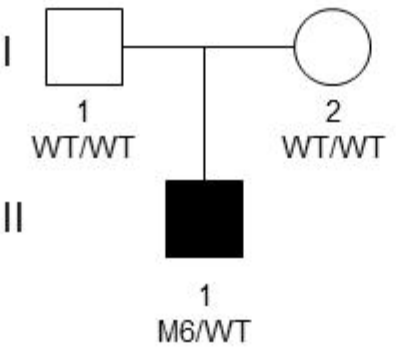
Family 4



M5:
***GNAO1* c.980C>G**

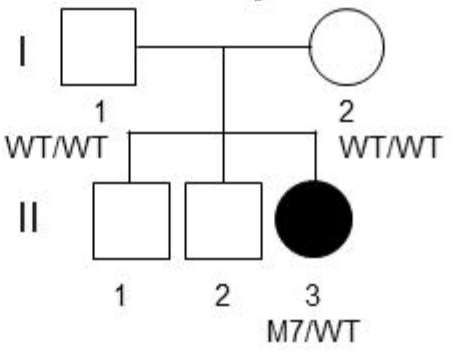


Family 5



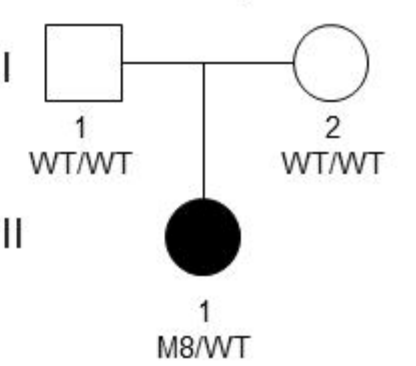
M6:
***EBF3* c.872T>A**

Family 6

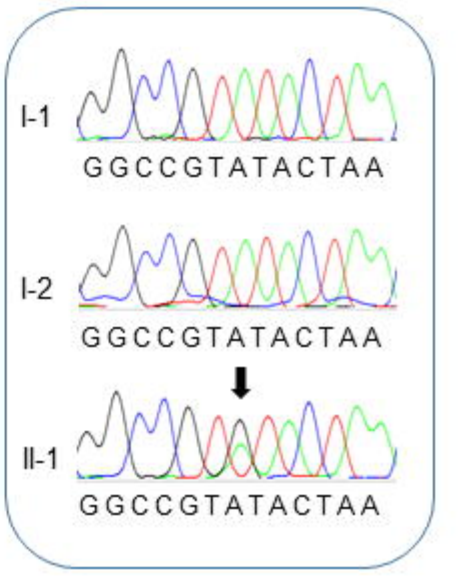


M7:
5q14.3q21.1 deletion

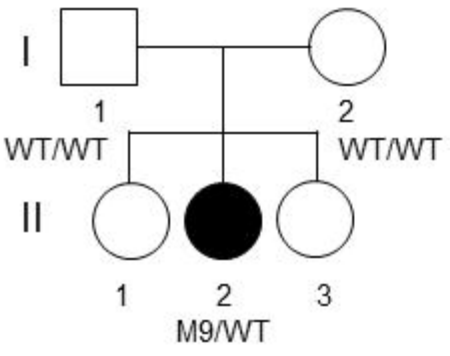
Family 7



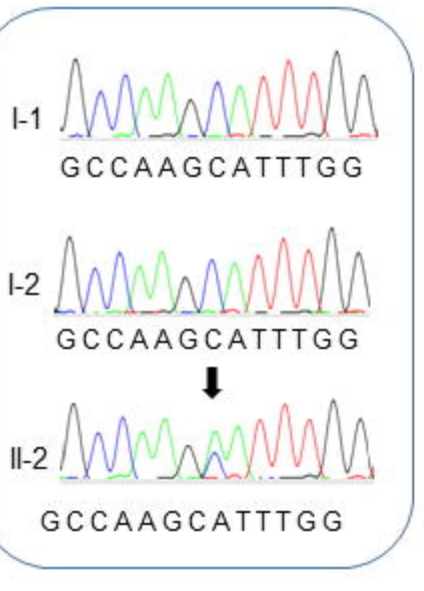
M8:
***CDK13* c.2609A>G**



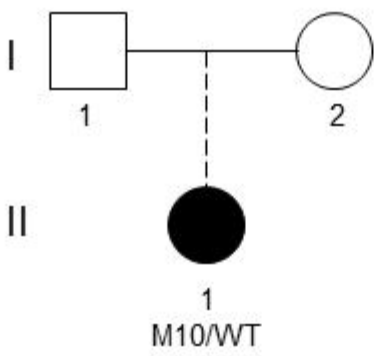
Family 8



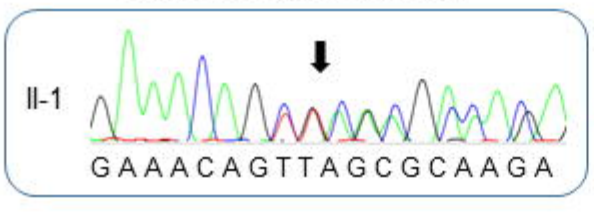
M9:
***POGZ* c.2497C>A**



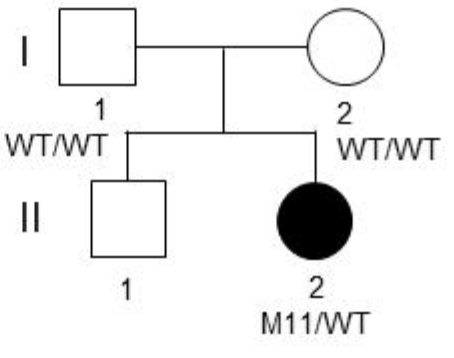
Family 9



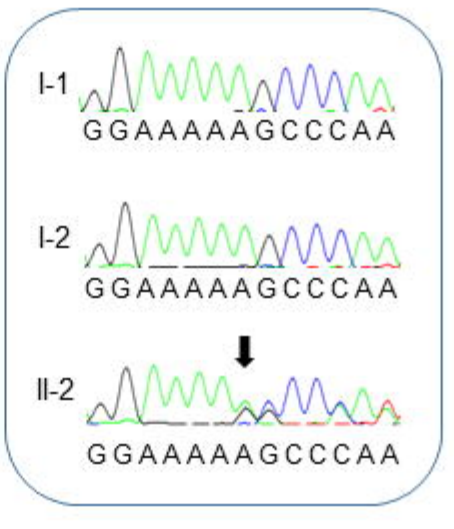
M10:
***MEIS2* c.934_937delTTAG**



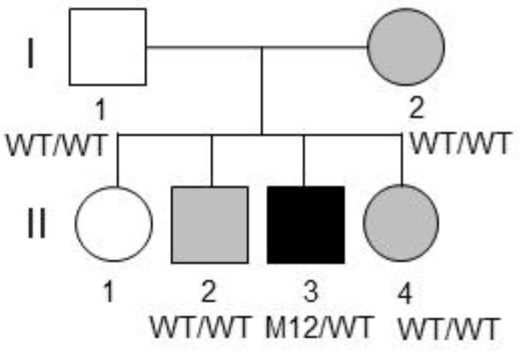
Family 10



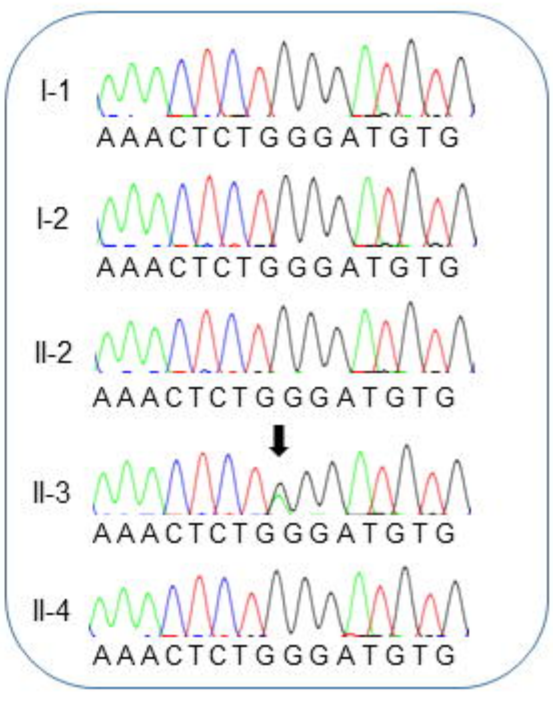
M11:
***DDX3X* c.1470delA**

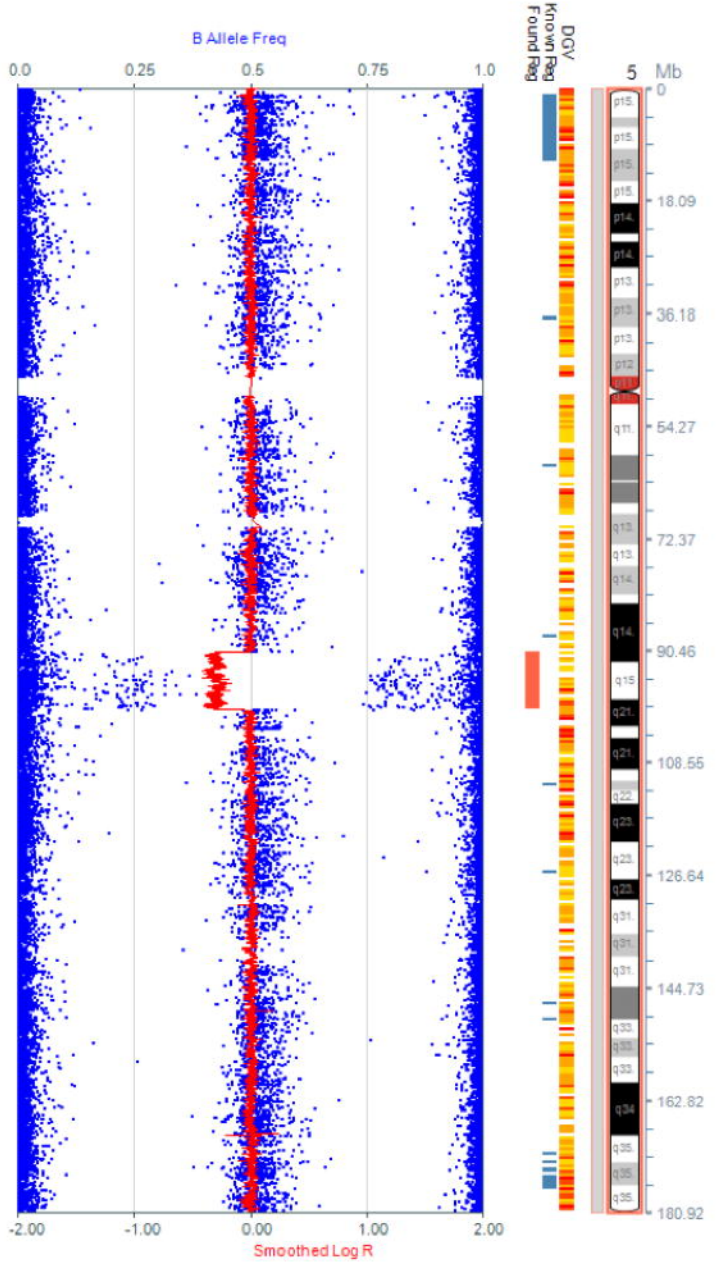


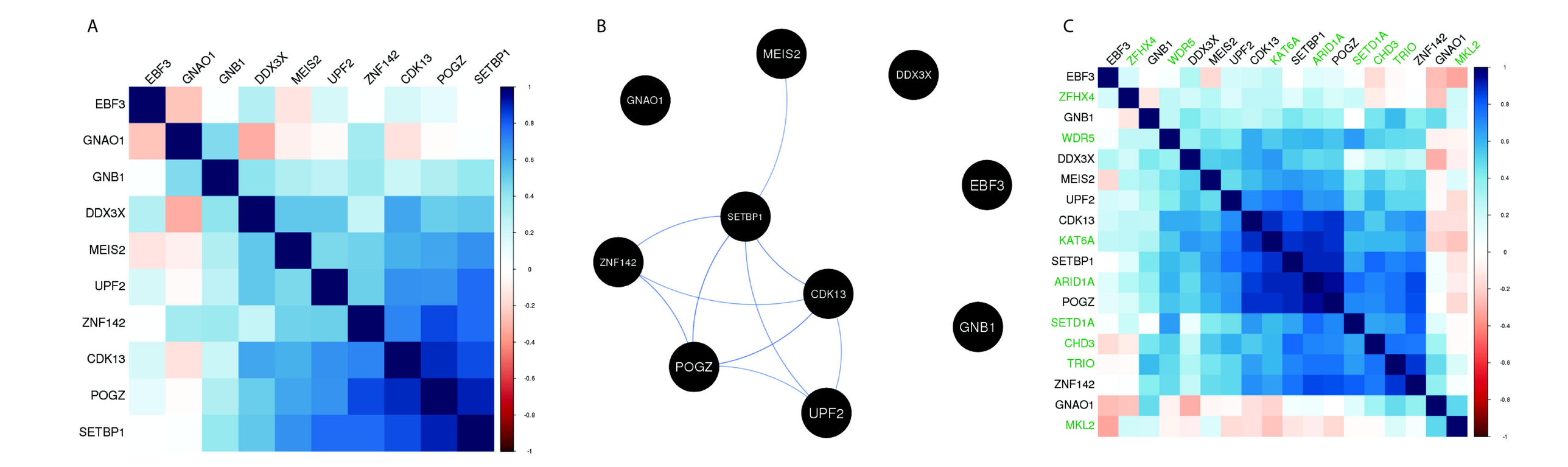
Family 11



M12:
***GNB1* c.632G>A**




B Allele Freq
0.0
0.25
0.5
0.75
1.0
-2.00
-1.00
0.00
1.00
2.00
Smoothed Log R
DGV
Known Reg
Found Reg
5
Mb
0
18.09
36.18
54.27
72.37
90.46
108.55
126.64
144.73
162.82
180.92


A
EBF3
GNAO1
GNB1
DDX3X
MEIS2
UPF2
ZNF142
CDK13
POGZ
SETBP1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
B
MEIS2
GNAO1
DDX3X
EBF3
SETBP1
ZNF142
CDK13
GNB1
POGZ
UPF2
C
EBF3
ZFHX4
GNB1
WDR5
DDX3X
MEIS2
UPF2
CDK13
KAT6A
SETBP1
ARID1A
POGZ
SETD1A
CHD3
TRIO
ZNF142
GNAO1
MKL2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1