Candidate Gene Variant Effects on Language Disorders in Robinson Crusoe Island

Hayley S. Mountford [1], Pía Villanueva [2, 3], María Angélica Fernández [2], Zulema De Barbieri [2], Jean-Baptiste Cazier [4,] and Dianne F. Newbury [1]

1 Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, UK

2 Department of Speech Language and Hearing Sciences, Faculty of Medicine, University of Chile, Santiago, Chile

3 Institute of Biomedical Sciences, Human Genetics Division, Faculty of Medicine, University of Chile, Santiago, Chile

4 Centre for Computational Biology, University of Birmingham, Edgbaston, UK

Abstract

Robinson Crusoe Island is a geographically and socially isolated settlement located over 600km west of the Port of Valparíso, Chile. An unusually high incidence (30%) of the Chilean equivalent of developmental language disorder (TEL) has been reported in Islander children, with 90% of these affected children found to be direct descendants of a pair of original founder-brothers, therefore strongly suggesting a shared genetic basis.

Here we utilise whole-genome sequencing to investigate potential underlying variants in a panel of thirty-four genes known to play a role in language disorders, in seven TEL affected and ten unaffected islanders. We use this targeted approach to look for rare, shared variants that

may underlie the diagnosis of TEL in a Mendelian genetic model. We go on to test whether the overall burden of rare variants is enriched in individuals affected by TEL or with Islanders related to the founder-brother lineage.

In the absence of explanatory rare variants, we further investigate these candidate genes within a complex model of inheritance, where inheriting a small number of moderate impact common variants may increase susceptibility of developing TEL. We examine if any variants segregate with affection status or with founder-brother-related status, and therefore may increase risk of developing a language disorder. Finally, we perform a pooled, gene-based tests to evaluate relationships between combined variation across candidate genes and TEL affection status.

Here we report a comprehensive examination of genes directly implicated in language-related mechanisms to identify ‘low hanging fruit’ of causative monogenic Mendelian variants, and complex association model of increased susceptibility in developmental language disorder found on Robinson Crusoe Island.

Keywords

Developmental language disorder (DLD), language genetics, trastorno especifico de lenguaje (TEL), isolated population, language development.

Introduction

Developmental language disorder (DLD) is the term given to primary childhood language disorders, which are not explained by other neurobiological disorders such as autism spectrum disorder, developmental delay, or hearing loss (Bishop et al. 2017). DLDs are remarkably prevalent and estimated to occur in over 7% of UK school age children (Norbury et al. 2016). In real terms, this means there are two to three affected children in every classroom. Even with adequate access to speech therapy and educational support, half of children with delayed language do not fully catch up with their peers, continuing to struggle with language throughout their childhood and into their adult lives (Hulme and Snowling 2009). Children with language disorders often struggle academically, and have been shown to have an increased risk of poor mental health outcomes, and are more likely to be unemployed in adulthood (Conti-Ramsden and Botting 2008).

Despite the remarkably high prevalence of DLDs, little is understood of the underlying aetiology. It is well established that DLD has a strong familial component, supported by twin and heritability studies (Stromswold 1998; Bishop et al. 2006; Barry et al. 2007). These familial disorders, termed Mendelian disorders, result from inheriting either one (dominant) or two (recessive) copies of extremely rare and damaging variants. Which act to disrupt protein function. The most well-known and clear-cut examples of Mendelian inheritance in language disorders can be found in a motor disorder known as childhood apraxia of speech (CAS), also known as developmental verbal dyspraxia. CAS is considered to be a sub-category of DLD that specifically refers to difficulties in the fine motor control required to produce and coordinate sounds into complete words and sentences, characterised by difficulties in producing speech sounds, dysarthric speech and poor oral motor control. The first CAS case to be solved involved

a dominant mutation in the gene *FOXP2* was found to be shared by all CAS-affected members of a large multigenerational pedigree, known as the KE family (Lai et al. 2001). The p.Arg553His *FOXP2* mutation is fully-penetrant in the KE family, meaning that all carriers have CAS and non-carriers do not. A number of subsequent studies have identified additional *FOXP2* mutations in other unrelated individuals as the cause of CAS (MacDermot et al. 2005; Tomblin et al. 2009; Turner et al. 2013; Moralli et al. 2015; Liegeois et al. 2016; Reuter et al. 2017) providing further evidence of the gene's role in language. A small number of other genes have also been implicated in the CAS phenotype. One such gene is the protein transporter *ERC1* (Thevenon et al. 2013) which was identified by overlapping 12p13.33 deletions in five unrelated CAS cases.

The CAS phenotype is considered extremely rare, and very few children are found to carry causative variants in *FOXP2* or *ERC1*. It is a similar story with other Mendelian causes of DLDs, and there are very few examples of genes where a high impact familial variant is shared among affected family members. One such example is the transmembrane protein encoding gene *TM4SF20*, in which a deletion of the second last exon leads to DLD and white matter hyper-intensities (Wiszniewski et al. 2013). The heterozygous deletion, found in children of mainly of South East Asian descent, was reported to show near-complete penetrance, meaning that deletion carriers were extremely likely to have a DLD phenotype.

Clear-cut Mendelian causes of language disorders are still relatively rare and are the exception rather than the rule. More commonly, genes are implicated in comorbid, overlapping disorders such as dyslexia, autism spectrum disorder and intellectual disability syndromes, still with DLD as a prominent feature as part of a complex disorder. A recent illustrative example is the identification of chromatin modelling gene *CHD3* (Snijders Blok et al. 2018). Mutations which led to changes in the ATPase/helicase domains of this protein resulted in CAS,

accompanied by intellectual disability, and macrocephaly in 35 unrelated individuals. Mutations in *CHD3* result in a more global neurodevelopmental syndrome rather than a language specific phenotype.

Similarly, mutations in the glutamate-gated ion channel gene *GRIN2A* can result in dominant DLD and epilepsy, with or without intellectual disability (Endele et al. 2010; Carvill et al. 2013). The CAS and DLD phenotypes found seen in these *CHD3* and *GRIN2A* examples are considered to be a primary feature of the disorder, as opposed to a secondary deficit of a more complex disorder. They do, however, reflect a difference in opinion of what constitutes a primary language disorder compared to a secondary feature of a broader neurodevelopmental syndrome. As there are very few genes which result in primary language disorders, genes implicated in syndromic neurodevelopmental disorders represent a substantial increase in our understanding of the genetic aetiology of language.

The family-based studies described above can provide starting points in understanding the biological base of language disorders through the identification of rare Mendelian variants. It is, however, not the case that the Mendelian inheritance model fully explains the underlying genetics of all DLDs. As genetic technologies and knowledge develops, we are slowly building a picture of genetic susceptibility within a complex inheritance model; where a number of variants are inherited together, interacting in particular environmental and cellular circumstances to result in a language disorder phenotype. These ‘risk’ variants are likely to be much more common in the population (minor allele frequency (MAF) ≥5%) and confer a moderately damaging effect. This contrasts with Mendelian recessive and dominant variants which tend to be extremely rare in the population (MAF≤1%) and are much more damaging to the resulting protein.

When inherited in combination with other 'risk' variants, they can combine together to become damaging and result in the DLD phenotype. This is more akin to cancer or diabetes; where there are a number of more common risk variants that interact with environmental factors. These risk variants each confer a small increase in an individual's susceptibility to develop a particular disorder. Our understanding of the role of genetic risk factors in language disorders lags behind other neurological disorders such as schizophrenia and autism spectrum disorder (ASD), which are much better characterised.

One of the best characterised examples of risk variants in language are in the gene *CNTNAP2*, which was first identified as a candidate for DLD through its functional interaction with FOXP2 (Vernes et al. 2008). The *CNTNAP2* gene encodes a neurexin-family synaptic protein and has been found to be associated with DLD (Devanna et al. 2017), epilepsy (Zweier et al. 2009) and ASD (Alarcon et al. 2008; Arking et al. 2008; Bakkaloglu et al. 2008). Other language-implicated genes have been found to associate closely with related comorbid disorders, suggestive of overlapping disease aetiology. This underlying comorbidity appears to be the rule rather than the exception.

Genes associated with language disorders and comorbid phenotypes, as well as details of the methods used to identify them, are reviewed in Chen et al. (2017), Deriziotis and Fisher (2017) and Mountford and Newbury (2018).

In reality, both rare damaging and common variants are likely to contribute to DLD. This can make the identification of novel variants challenging. The detection of common variants requires large (tens of thousands) of individuals all phenotyped in the same way. Family-based studies genetic studies are performed on a much smaller scale, but variants identified in one

family tend to be extremely rare, and unlikely to replicate in other pedigrees. To narrow the genomic regions in which to look for the candidate variants, a pedigree needs to contain both affected and unaffected individuals, and include as many second-degree relatives as possible. Extensive pedigrees provide exceptional opportunities to narrow down the regions shared by affected individuals (and not shared by unaffected individuals), ultimately narrowing the search space that contains the causative variant.

An exceptional example of a large pedigree with DLD comes from Robinson Crusoe Island (RCI). RCI is geographically and socially isolated, located over 600km off the coast of Chile. Islander children have an exceptionally high occurrence rate (62.5%) of speech and language disorder. Half of the cases (56%) have language delay in isolation with no evidence of intellectual disability, or other neurological disorders that may affect language ability (Villanueva et al. 2008; De Barbieri et al. 2018). This specific type of language disorder is named *Trastorno Especifico de Lengauje* (TEL). The Chilean term which when translated to English means ‘language specific disorder’ (De Barbieri et al. 2018). The remainder present with more generalised developmental or neurological disorders in which language delay is a secondary feature of an overlapping related disorder (e.g. ASD, developmental delay).

RCI provides a unique cohort in which to study the genetics of language disorders; a geographically isolated population, founded in 1876 by 64 individuals. The current population has over 800 inhabitants, with a high consanguinity rate (14.9%) (Villanueva et al. 2014). This means that risk alleles that have been inherited from original founders may be greatly enriched in the current cohort, providing substantial power to detect contributory variants. The strength of this population is in the consanguinity and isolated nature of the population, despite relatively modest sample numbers. Previous gene dropping simulations reported in Villanueva et al (2015)

within the observed pedigree structure indicate that causal allele frequencies will be consistently and significantly elevated above expected for a range of allele frequencies and founder allele combinations, and even rare (1%) founder alleles will be greatly enriched (13%) in the current cohort.

A previous study of the RCI population used whole exome sequencing to identify a rare nonsynonymous p.Asp150Lys (rs144169475, chr4:g.47,907,320A>T hg19) variant in the nuclear transcription factor X-box binding-like 1 gene, *NFXL1* (Villanueva et al. 2015). This variant was found to be enriched in islanders with the TEL phenotype (39%) compared to those with typical language development (TLD) (10%) ($p=2.04x10^{-4}$) and accounted for 7% of the trait variance seen on the island (Villanueva et al. 2015). Although strongly associated with TEL, this variant was found to occur in language-typical Islanders and was not present in all TEL individuals. It therefore only explains part of the TEL occurrence in this population.

The p.Asp150Lys variant was reported by (Villanueva et al. 2015) was detected in 4.2% of the 320 Chilean and Columbian controls tested (27/640 alleles tested), and not detected at all in European controls (0/254 alleles tested). Villanueva and colleagues also showed that rare coding variants in *NFXL1* were enriched in a cohort of 117 unrelated cases from the UK-based Specific Language Impairment Consortia cohort. While this replication cohort is small in sample size, it provides further independent evidence for the role of *NFXL1* in language disorders.

Since this publication, additional population data has become available through the gnomAD database (Lek et al. 2016), showing the variant is present at an allele frequency of 5.047% in Latinos (AMR) (1772/35,112 alleles tested) and includes 57 homozygotes. The variant is still extremely rare in people of European decent (0.00078%) (1/126,384 alleles tested)

and across all gnomAD populations it is 0.65% (1825/281,822 alleles tested) and is therefore still considered to be extremely rare variant. As the Robinson Crusoe Islanders are predominantly of Chilean ancestry, then careful selection of relevant population specific allele frequencies are important when trying to understand the effect of a variant. As larger control sample sets become available, more accurate estimates of allele frequencies can be derived. In this instance, the allele frequency of the p.Asp150Lys variant is still enriched in Islanders compared to the Latino population in general - 11.3% compared to 5% respectively. The incidence of this risk allele is significantly increased among those Islanders affected by TEL (19.4%) (Villanueva et al. 2015). It is, however, clear that this variant alone does not explain the entire risk. Instead, we propose this variant is a genetic modifier, conferring a moderate increase in risk of TEL and may act alongside another, as yet undiscovered, rare damaging variant.

Given the pedigree structure and isolated nature of the RCI population, it is hypothesised that the speech and language disorders present on RCI are caused by rare (minor allele frequency (MAF) of ≤1%) genetic mutations with a high-risk effect, or by combinations of genetic variants that together confer a high risk. Although previous publications have presented exome sequence data in this population, a full assessment of the entire genome has yet to be made. In the current study, we use whole-genome sequence data from 7 TEL and 10 typical language development (TLD) islander individuals to fully assess the role of rare and damaging variants in genes implicated in language disorders, including those associated with both primary language disorders and overlapping syndromes. We will examine whether rare and common variants in these thirty-four language candidate genes are risk factors for TEL in this extensive pedigree through association analyses. This whole-genome sequencing approach provides complete capture of the gene regions, covering coding exons, untranslated and intronic regions in its

entirety, enabling a comprehensive assessment of existing candidate genes within this population.

Villanueva et al. (2014) ascertained near-complete genealogical records of the RCI, and using this resource found that 90% of islander children with TEL are directly descended from a pair of original founder-brothers. Based on this finding, we further investigate variants patterns in founder-brother-related individuals allowing us to evaluate the hypothesis that causative variants may be shared or over-represented in founder-brother-related TEL individuals, thus contributing to an increased incidence of language disorder. Both rare and common variants in these thirty-four language candidate genes will be tested for segregation, enrichment and association in founder-brother-related-islanders.

Materials and Methods

Participants were assessed in line with current diagnostic practices set by the Chilean Ministry of Education, and described in detail in De Barbieri et al. (2018). The ethics department of the University of Chile approved the project "Genetic analysis of language-impaired individuals from the Robinson Crusoe Island" – Project Number 001-2010. Informed consent was given by all participants and/or, where applicable, their parents. The test battery was performed in Chilean Spanish, by native speakers, who assessed phonological production (TEPROSIF-R) (Pavez G et al. 2008), expressive and receptive morphosyntax (Toronto Spanish Grammar Exploratory (TGE) test) (Pavez MM 2003), and non-verbal intelligence (Columbia Mental Maturity Scale) (Burgemeister et al. 1998). Children met the criteria for a diagnosis of TEL if they fell either 2SD below the expected score, or more than two years below the score expected for their age (TEPROSIF-R), below the 10^{th} percentile (TGE test), and had a neuro-

typical non-verbal IQ score above the 10th percentile. Individuals describe as having typical language development (TLD) scored more than 2SD below expected and more than two years below their expected age (TEPROSIF-R test), above the 10th percentile (TGE test), and above the 10th percentile on the non-verbal IQ test.

Adults were assessed on verbal fluency using the Barcelona test (Peña-Casanova et al. 1997), verbal comprehension using the Token test (De Renzi and Vignolo 1962), and non-verbal ability using the Raven progressive matrices (Raven 2003). TEL adults scored below the 10th percentile on either the verbal fluency or verbal comprehension, and above the 10th percentile in the non-verbal IQ score. TLD adults scored above the 10th percentile on all three measures. Diagnostic criteria and language assessments were set as reported in De Barbieri et al 2018.

Seventeen islanders were selected for sequencing based on the most distantly related and therefore most informative individuals representing affected (TEL) and unaffected (TLD) phenotypes. Whole-genome sequencing performed by Oxford Genomics using Nimblegen™ capture and sequenced on the Illumina HiSeq platform with 98.2% of bases covered to a minimum coverage of 10x and 88.53% covered to 20x. Quality control and sequence alignment (to build hs37d5) were performed by the Oxford Genomics service using their standard analysis pipeline.

Variant calling was performed using Platypus (Rimmer et al. 2014) and GATKv3.5-0 (Van der Auwera et al. 2013) using best practises. Bcftools 1.2 and htslib-1.2.1 (http://github.com/samtools/bcftools) were used to intersect high confidence variants called by both algorithms. PASS variants were filtered using VCFtools 0.1.14 (Danecek et al. 2011). Additional hard filtering was used to filter variants with a map quality (MQ) score of ≥40, total

allele count (AN) of ≥26, and total read depth (DP) of ≥140 across all 17 samples using Bcftools 1.7 and htslib-1.2.1 (http://github.com/samtools/bcftools). Variant calls, in VCF format, were split and left aligned using Bcftools to ensure one variant per line, and therefore compatibility with downstream applications.

The list of genes implicated in language disorders, and by proxy language, was obtained from the literature, previously reviewed in Chen et al. (2017), Deriziotis and Fisher (2017), and Mountford and Newbury (2018). Additionally, we included *CHD3* which was reported in (Snijders Blok et al. 2018), plus six novel candidate language genes (*KAT6A*, *MKL2*, *SETD1A*, *TNRC6B*, *WDR5*, and *ZXHF4*) recently reported in Eising et al. (2018). Complete canonical gene region coordinates, including the entire annotated 5' and 3' untranslated regions, were obtained from UCSC Genome Browser hg19 build (https://genome-euro.ucsc.edu) (supplementary Table 1), and variants falling within these thirty-four regions were extracted from the VCF using VCFtools 0.1.15 (Danecek et al. 2011).

Variant annotation was performed using Annovar (release 2018Apr16) (Wang et al. 2010) with dbSNP (avsnp150), splice site prediction (dbscSNV version 1.1), variant pathogenicity prediction (dbSNP35a), ExAC exomes allele frequency data (2015 release) (ALL n=125,748), gnomAD genome collection (v2.0.1) (ALL n=15,708, Admixed American (AMR) n=424, non-Finnish European (NFE) n=7,718), 1000 Genomes Project (1000g2015aug) allele frequencies (ALL, AMR, and EUR (European)), and Clinvar version 20180603 databases (hg19 build).

Variant filtering and prioritisation was performed using Linux command line to identify potentially causative variants that are shared between all TEL or all founder-brother-related individuals using both recessive and dominant inheritance models.

Rare variants with a minor allele frequency of ≤1% in gnomAD AMR whole-genome population data, and falling within a language gene region (coding, intronic and UTR) were tested for enrichment in affected (TEL or founder-brother-related individuals) using students T-test.

Segregation analysis was performed on all variants (both rare and common) to test for variants which showed complete segregation between TEL affected individuals but not with TLD individuals. This approach was also used to identify variants that were shared by all founder-brother-related individuals and not shared by non-founder-brother-related individuals.

To test for the effect of all variants (rare and common) acting in combination in an individual gene, a gene-based association test was performed using the SKAT test in RVTESTs (Zhan et al. 2016) on all variants falling within the gene region (coding, intronic and UTR). Association testing was performed on both SNVs and indels, excluding those in HWE $\leq 1x10^{-5}$ which was calculated in Plink 1.90 (Purcell et al. 2007). Thresholds for statistical significance were set by Bonferroni correction, to account for multiple testing.

A flow diagram of the workflow and statistical analyses is shown in Figure 1.

Results

Sequencing of Language Gene Regions

A total of thirty-four genes implicated in language disorders (and therefore language) were identified through a combination of current literature reviews (Chen et al. 2017; Deriziotis and Fisher 2017; Eising et al. 2018; Mountford and Newbury 2018; Snijders Blok et al. 2018). The combined thirty-four language gene regions spanned a target region of 12.6 Mbp. Variants annotated by Annovar as 'non-coding RNA' were excluded from analysis, which included all variants contained within in *ABCC19*, and it was therefore excluded from further analysis resulting in only 33 genes remaining. After variant calling, a total of 33,966 non-reference calls were identified within the selected gene regions (exonic, intronic, untranslated (UTR), and potential splice sites (±3bp)). This consisted of 29,510 single nucleotide variants (SNVs) and 4,458 small insertions and deletions (indels). Sequenced individuals had a median of 13,817 variants calls (range=12,902-16,610) (Table 1).

Rare Mendelian Variant Analysis

We first performed a search for variants contained with potential for a clear functional impact, therefore contained only within exons or potential splicing site regions. 137 (129 SNVs, 8 indels) variants were detected (median=41, range=33-52).

To further narrow these to potential variants of interest, variants predicted unlikely to have an amino acid of the protein (synonymous (74 variants) or nonframeshift indels (7 variants)) totalling 81 variants were dropped from the analysis. This left a total of 56 variants, consisting of one potential splice region variant, one stop-gain, one frameshift insertion and 53 missense variants (median=16, range=10-22) (Table 1).

Finally, to identify novel or extremely rare variants that may be impacting on the TEL phenotype, variants were excluded if they had a minor allele frequency (MAF) of more than 1% (MAF ≥0.01) in the gnomAD AMR population and more than 5% (MAF ≥0.05) in the gnomAD ALL dataset. A total of fifteen rare and nonsynonymous variants were identified in the 17 sequenced islanders (median=1, range=0-3), spanning ten different genes (Table 2).

To assess a potential impact of these prioritised variants, and therefore their potential pathogenicity, functional annotations were investigated. All 15 variants were missense SNVs, with no loss of function variants (stop-loss, stop-gain, or frameshift) being prioritised. Similarly, no potentially damaging splice site mutations were found to be rare (MAF≤0.01).

None of the prioritised rare nonsynonymous variants were found to segregate with either TEL status or were shared by the direct descendants of the founder brothers. These findings indicate there is no single high impact Mendelian variant in previously reported language genes that fully explains the TEL phenotype in the RCI population. The rare and nonsynonymous variants detected spread across many of the candidate genes and were found in both TEL and TLD individuals alike, often occurring in a single individual. Six rare nonsynonymous variants were found only in TEL individuals. These occurred in the *ROBO1*, *NFXL1*, *KIAA0319*, *ERC1*, *ATP2C2*, and *TNRC6B* genes. Six further rare variants were found only in TLD individuals. These were contained in *CNTNAP5*, *ROBO2*, *KAT6A*, *ZFHX4*, *MKL2* and *ATP2C2*. Finally, three variants were found in both TEL and TLD individuals with two found in *ROBO2* and one within *ZFHX4*.

All fifteen of the prioritised variants were missense, single base pair changes rather than indels, and therefore the a *in silico* missense pathogenicity prediction score could be used to

interpret potential pathogenicity. To investigate the functional impact of these missense mutations, variants were flagged as potentially damaging if they were predicted to be damaging by at least seven of ten variant prediction tools (SIFT, Polyphen2, Polyphen2_HDIV, LRT, MutationTaster, MutationAsseser, FATHMM, Provean, MetaSVM and MetalLR). Only one of the fifteen variants met these criteria, c.C256T (p.Arg86Trp) in *ATP2C2*, which was detected in a homozygous state in a single individual with typical language development (TLD-2). This variant is unreported (MAF=0.0000) in the AMR and ALL genome control populations and is extremely rare (MAF=0.00001658) in the ExAC ALL dataset (n=125,748). It has not previously been reported in a homozygous state. While this is a likely functional variant in a gene of interest, as it was identified in an unaffected individual and so is unlikely to play a role in the language disorder seen on RCI.

We can therefore conclude that there is no causal variant within a language-implicated gene that is solely responsible for the language phenotype seen in the RCI. Nonetheless, it remains possible that Islanders may have a genetic susceptibility resulting from combinations of rare and/or common variants each conferring a moderate effect size.

Rare Variant Burden Analysis

To examine whether an overall burden of rare variants across the entire gene regions (inclusive of exons, introns, UTRs and potential splice variants) were contributing directly to TEL affection status, or were enriched in founder-brother related individuals, rare variants (≤1% in AMR genome population data set) were subset across the full gene regions. A total of 4,998 rare (≤1% in gnomAD AMR) variants were detected across the 33 language genes (12.6Mb) in 17 sequenced islanders, with a mean of 435 rare variants (1 rare variant every 29Kb)

(median=340, range=266-732) (Table 3). Interestingly, the TLD group (n=10) were found to harbour a higher number of rare variants across the thirty-three language genes than the TEL group (n=7). The TEL group had a mean of 390.5 rare variants, compared to 466.1 found in the TLD group. A student's t-test indicated that this group mean difference was not significant (p=0.3558).

To investigate whether direct descendants of the founder-brothers carry a higher number of rare variants within the language genes and are therefore at a higher risk of developing language difficulties, a test of variance was performed on the founder-brother-related group (n=14) compared to the non-founder-brother-related group (n=3). The founder-brother-related individuals carried a higher mean number of variants (mean=456.79) compared to the non-founder brother related individuals (mean=333.33) although this did not reach significance (p=0.0743). These findings therefore indicate that combinations of rare variants across the 33 genes are unlikely to be responsible for the developmental language disorder seen on RCI.

Variants Segregation Analysis

Initial exploration of the genome data focussed on rare variants. To explore the role of common variants, we assessed all 33,966 SNVs and indels from thirty-three language genes. In the absence of co-segregating rare variants, we applied a wider analysis that included all variants across the language-candidate genes. All 33,966 SNVs and indels from thirty-three language genes were assessed for co-segregation in the founder-brother-related individuals (TEL n=6, TLD=8).

Three hundred and twenty-seven variants were found to be homozygous across all six TEL founder-brother-related individuals and 54 variants were found to be heterozygous in all six

TEL founder-related individuals. However, all these variants were also found to occur in TLD individuals. One variant, an intronic SNV in *CNTNAP5* (rs9309831), was found to occur in a homozygous state in all 6 TEL cases. However, this variant was also found in a homozygous state in three TLD participants and in a heterozygous state in the remaining 5 TLD participants. Upon further investigation, this variant was found to have a minor allele frequency of ≥0.9. Five intronic variants (rs779979, rs779980, rs7605310, rs13402327 and rs2565748), all in *CNTNAP5*, were found to be heterozygous in all TEL cases, but were also observed in a heterozygous state in 4 of the 8 TLD participants.

These five intronic variants identified in *CNTNAP5* may be inherited together in a haplotype block, increasing susceptibility of TEL. To investigate if these variants fall within a shared region, we produced a genotype grid in order to visualise the region (Figure 2). The genotype grid shows there is no clear region shared between the TEL founder-brother related individuals, and there are a large number of non-segregating variants between the five variants. The lack of clear genotype segregation in the *CNTNAP5* region suggests it is not associated with TEL affection status in founder-brother related individuals.

Gene-based Analysis of Common and Rare Variants

Finally, to assess the possibility of a complex genetic mechanism, we performed gene-based association analyses of all variants (common and rare) across the language candidate genes. Collapsing variants into gene-regions allows for signals that may not be directly covered by the sequencing (upstream variants of a gene) and will improve detection of this signal by combining all variants into a gene-based test. Therefore, a statistically significant gene-based test is a strong indication of genetic contribution to TEL in this population, at a whole-gene level.

Thirty-three genes containing 33,966 identified variants were tested using the kernel based SKAT test of association in the RVTESTs package (Zhan et al. 2016).

When using the TEL status as a categorical variable, no genes (Table 4) reached statistical significance (Bonferroni corrected p=0.00073), and only one gene, *NOP9*, was nominally significant (p=0.0407).

To assess whether complex patterns of variants are enriched in original founder-brother-related individuals, we repeated the SKAT test in founder-brother-related against non-founder-brother-related individuals. Again, no gene was significantly associated after a Bonferroni correction (p=0.00073). Two genes (*DYX1C1* and *SETBP1*) had p-values <0.01 (p=0.0097 and p=0.008 respectively), with a further four genes (*FOXP1*, *MKL2*, *RBFOX2* and *SETD1A*) were found to be nominally significant at $p \leq 0.05$.

Discussion

Robinson Crusoe Island is an isolated population of admixed Chilean and European ancestry reports to have an unusually high incidence of language disorder, termed TEL in Chile. Near-complete genealogical records indicate that 90% of affected Islander children are direct descendants of a pair of original founder-brothers, strongly suggesting a genetic founder effect (Villanueva et al. 2008; Villanueva et al. 2011; Villanueva et al. 2014; Villanueva et al. 2015).

This study utilised whole-genome sequencing to comprehensively investigate variation across thirty-three genes previously implicated in language-related phenotypes in TEL affected (n=7) and TLD unaffected (n=10) Islanders.

Based on the high degree of relatedness and therefore shared genetics of the Islanders, we postulated that a rare (MAF ≤1% in gnomAD AMR) high impact variant may underlie TEL affection on RCI. Fifteen rare nonsynonymous variants in eleven genes were identified in the sequenced individuals. However, no single variant segregated with TEL status. Interestingly, the only variant with a potential pathogenicity prediction was found in a homozygous state in TLD-2 an unaffected individual. No single pathogenic variant within the language gene regions was detected.

A wider investigation of rare variants (MAF ≤1% in the gnomAD AMR population controls) across the entire gene regions (including introns) did not find any evidence of increased variant burden in TEL individuals compared to those with TLD (p=0.3558). In fact, the TLD group were found to carry a higher number of rare variants, although this was not statistically significant. Interestingly, a moderate enrichment of rare variants was observed in the founder-brother-related individuals. As the sample size is highly limited, this may be artificially inflated by the small number of non-founder-brother-related individuals sequenced (n=3).

In the absence of co-segregating rare variants, we extended our investigation and considered all variants observed across the candidate genes (n=33,966). Again, no variant was found to fully segregate with affection status. Interestingly, six intronic *CNTNAP5* variants were found to occur in a homozygous (rs9309831) or heterozygous (rs779979, rs779980, rs7605310, rs13402327 and rs2565748) state in all TLD individuals investigated. However, these variants were also observed in TLD individuals and genotype analysis showed there was no clear haplotype that segregated with TLD affection within the founder-brother-related individuals.

Finally, we assess all variants within a gene-based burden analysis. Again, no association was found to either TLD status or founder-brother-relatedness. Several genes reached nominal significance, potentially suggesting a multi-loci complex genetic signal however the small number of non-founder-related individuals may increase statistical bias.

In conclusion therefore, our results indicate either a single hit variant or a general burden of variants within known candidate language genes do not explain the risk of TEL in the RCI Islanders. No single variant was found to co-segregate with language or founder-brother-related status and no gene showed any evidence of increased burden in relation to TEL.

Interestingly, the *NFXL1* gene, which was identified as a risk-gene within the RCI population by Villanueva et al. (2015), did not show association to TEL in the current study. The p.Asp150Lys risk variant was not included in the rare variant analyses presented here as it has a gnomAD frequency greater than 1% and, as previously reported (Villanueva et al. 2015), the variant does not show complete segregation with TEL status. It should be noted that the cohort of islanders sequenced in this paper is much smaller than that genotyped for p.Asp150Lys in the Villanueva (2015) paper, which represents a much more complete characterisation of this variant. Information regarding the frequency of this variant across control population is now more extensive and is estimated at MAF=0.0504 (AMR) and MAF=0.006476 (ALL) clearly indicating the variant is more common in people with Latino ancestry. Nonetheless, the previous research stands and even with these updated population data, this variant remains enriched among Islanders and TEL individuals and, as such, still represents an excellent candidate risk variant in the RCI population.

The lack of association to the *NFXL1* variant in the current study is likely explained by the small sample size. Only 3/7 TEL cases carried the *NFXL1* variant compared to 3/10 TLD controls, by chance fewer than expected. At a genome wide level, the p.Asp150Lys *NFXL1* variant was found to account for 7% of the trait variance seen on RCI (Villanueva et al. 2015), suggesting this variant may be a modifier to a dominant model, or play a role in a complex susceptibility model. Our data do not contradict this finding but show a need for improved sensitivity through increased statistical power.

Whole-genome sequencing is a useful method for identifying the variants underpinning genetic disorders and provides an unprecedented range of genetic information in one test. As whole-genome sequencing becomes more common, the major bottleneck is the analysis of the huge volume of data generated from a sequencing run and narrowing variants of interest from the vast numbers of non-reference variant calls detected in each individual. This means, that even with extensive phenotypic and familial information, it can be difficult to narrow the cause of disease. One practical approach to thinning the number of potential causative variants is to combine the usage of candidate genes already implicated in the disorder, in combination with pedigree information to look for segregating variants. In large, complex pedigrees like Robinson Crusoe Island, prioritising variants that segregate with affection status can increase power to detect causative variants. These methods allow for the identification of ‘low hanging fruit’: rare, nonsynonymous variants that segregate with disease and are in genes already implicated in language. Whole-genome sequence analysis, particularly in a large and complex family like RCI, can be challenging and technically non-trivial, therefore this approach is a sensible first place to start.

Similar family and genealogy-based methodologies have been successfully applied to other related population in which language disorders are common. (Wiszniewski et al. 2013) identified an ancestral deletion of exon 3 in the gene *TM4SF20* that is present in ~1% of people with South East Asian ancestry. They found that this deletion strongly increased the risk of language disorders with or without white matter hyperintensity. Similarly, (Kornilov et al. 2016) reported the association of *SETBP1* with DLD in a Russian population from an isolated village where they found a remarkably high occurrence rate of ~30%.

We have shown that the TEL seen on Robinsons Crusoe Island is not caused by a single shared Mendelian mutation in known language candidate genes and have comprehensively tested for 'low hanging fruit'. Founder-brother related individuals may have a subtle 'risk' profile from a small number of moderate effect variants as part of a complex model however we did not detect a robust association with any of the thirty-three language genes tested. Therefore, the underlying cause of TEL on RCI is likely due to a Mendelian variant in a novel gene that is yet to be associated with language, or alternatively a complex susceptibility model that we lacked the power to detect.

Acknowledgements

We would like to thank all the families, professionals and individuals who participated in this research. In particular, we are extremely grateful to the inhabitants of Robinson Crusoe Island who have agreed to participate in this study. We would also like to thank Mr. Felipe Paredes, the mayor of the Ilustre Municipalidad de Juan Fernández for his infinite assistance and

patience in the development of this research. Also to the authorities of schools of medicine and dentistry for giving us the necessary permits to travel to the island of Juan Fernandez.

The Robinson Crusoe project was funded by the Medical Research Council [MR/J003719/1]. The collection of DNA samples and characterisation of the Robinson Crusoe population was funded by Vicerrectoría de Investigación, Universidad de Chile (www.uchile.cl), UCHILE DID TNAC 01-02/01, UCHILE DI MULT 05-05/02 grants. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Interest Statement

The authors declare no conflicts of interest

References

Alarcon M, Abrahams BS, Stone JL, Duvall JA, Perederiy JV, Bomar JM, Sebat J, Wigler M, Martin CL, Ledbetter DH et al. 2008. Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. Am J Hum Genet. 82(1):150-159.

Arking DE, Cutler DJ, Brune CW, Teslovich TM, West K, Ikeda M, Rea A, Guy M, Lin S, Cook EH et al. 2008. A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. Am J Hum Genet. 82(1):160-164.

Bakkaloglu B, O'Roak BJ, Louvi A, Gupta AR, Abelson JF, Morgan TM, Chawarska K, Klin A, Ercan-Sencicek AG, Stillman AA. 2008. Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders. The American Journal of Human Genetics. 82(1):165-173.

Barry JG, Yasin I, Bishop DVM. 2007. Heritable risk factors associated with language impairments. Genes, Brain and Behavior. 6(1):66-76.

Bishop DVM, Adams CV, Norbury CF. 2006. Distinct genetic influences on grammar and phonological short-term memory deficits: evidence from 6-year-old twins. Genes, Brain and Behavior. 5(2):158-169.

Bishop DVM, Snowling MJ, Thompson PA, Greenhalgh T, Catalise-consortium T. 2017. Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. Journal of Child Psychology and Psychiatry. 58(10):1068-1080.

Burgemeister B, Blue L, Lorge I. 1998. Escala de madurez mental de Columbia. Madrid. Publicaciones de Psicología Asociada.

Carvill GL, Regan BM, Yendle SC, O'Roak BJ, Lozovaya N, Bruneau N, Burnashev N, Khan A, Cook J, Geraghty E. 2013. GRIN2A mutations cause epilepsy-aphasia spectrum disorders. Nature genetics. 45(9):1073-1076.

Chen XS, Reader RH, Hoischen A, Veltman JA, Simpson NH, Francks C, Newbury DF, Fisher SE. 2017. Next-generation DNA sequencing identifies novel gene variants and pathways involved in specific language impairment. Sci Rep. 7:46105. eng.

Conti-Ramsden G, Botting N. 2008. Emotional health in adolescents with and without a history of specific language impairment (SLI). J Child Psychol Psychiatry. 49(5):516-525.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. Bioinformatics. 27(15):2156-2158.

De Barbieri Z, Fernández MA, Newbury DF, Villanueva P. 2018. Family aggregation of language impairment in an isolated Chilean population from Robinson Crusoe Island. International journal of language & communication disorders. 53(3):643-655.

De Renzi A, Vignolo LA. 1962. Token test: A sensitive test to detect receptive disturbances in aphasics. Brain: a journal of neurology.

Deriziotis P, Fisher SE. 2017. Speech and language: Translating the genome. Trends in Genetics. 33(9):642-656.

Devanna P, Chen XS, Ho J, Gajewski D, Smith SD, Gialluisi A, Francks C, Fisher SE, Newbury DF, Vernes SC. 2017. Next-gen sequencing identifies non-coding variation disrupting miRNA-binding sites in neurological disorders. Mol Psychiatry. eng.

Eising E, Carrion-Castillo A, Vino A, Strand EA, Jakielski KJ, Scerri TS, Hildebrand MS, Webster R, Ma A, Mazoyer B. 2018. A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. Molecular psychiatry.1.

Endele S, Rosenberger G, Geider K, Popp B, Tamer C, Stefanova I, Milh M, Kortüm F, Fritsch A, Pientka FK. 2010. Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. Nature genetics. 42(11):1021-1026.

Hulme C, Snowling MJ. 2009. Developmental disorders of language learning and cognition. John Wiley & Sons.

Kornilov SA, Rakhlin N, Koposov R, Lee M, Yrigollen C, Caglayan AO, Magnuson JS, Mane S, Chang JT, Grigorenko EL. 2016. Genome-Wide Association and Exome Sequencing Study of Language Disorder in an Isolated Population. Pediatrics. 137(4).

Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature. 413(6855):519-523.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 536(7616):285-291.

Liegeois FJ, Hildebrand MS, Bonthrone A, Turner SJ, Scheffer IE, Bahlo M, Connelly A, Morgan AT. 2016. Early neuroimaging markers of FOXP2 intragenic deletion. Sci Rep. 6:35192.

MacDermot KD, Bonora E, Sykes N, Coupe AM, Lai CS, Vernes SC, Vargha-Khadem F, McKenzie F, Smith RL, Monaco AP et al. 2005. Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits. Am J Hum Genet. 76(6):1074-1080.

Moralli D, Nudel R, Chan MTM, Green CM, Volpi EV, Benítez-Burraco A, Newbury DF, García-Bellido P. 2015. Language impairment in a case of a complex chromosomal rearrangement with a breakpoint downstream of FOXP2. Molecular cytogenetics. 8(1):36.

Mountford HS, Newbury DF. 2018. The genomic landscape of language: Insights into evolution. Journal of Language Evolution. 3(1):49-58.

Norbury CF, Gooch D, Wray C, Baird G, Charman T, Simonoff E, Vamvakas G, Pickles A. 2016. The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. J Child Psychol Psychiatry. 57(11):1247-1257.

Pavez G, Maggiolo M, Coloma T, González M. 2008. Test para evaluar procesos de simplificación fonológica: TEPROSIF-R. Santiago: Ediciones Universidad Católica de Chile.

Pavez MM. 2003. Test exploratorio de gramática española de A. Toronto Aplicación en Chile Santiago: Ediciones Universidad católica de Chile.

Peña-Casanova J, Guardia J, Bertran-Serra I, Manero R, Jarne A. 1997. Shortened version of the Barcelona test (I): subtests and normal profiles. Neurología (Barcelona, Spain). 12(3):99-111.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81(3):559-575.

Raven J. 2003. Raven progressive matrices. Handbook of nonverbal assessment. Springer; p. 223-237.

Reuter MS, Riess A, Moog U, Briggs TA, Chandler KE, Rauch A, Stampfer M, Steindl K, Gläser D, Joset P. 2017. FOXP2 variants in 14 individuals with developmental speech and language disorders broaden the mutational and clinical spectrum. Journal of medical genetics. 54(1):64-72.

Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, Consortium WGS, Wilkie AO, McVean G, Lunter G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 46(8):912-918.

Snijders Blok L, Rousseau J, Twist J, Ehresmann S, Takaku M, Venselaar H, Rodan LH, Nowak CB, Douglas J, Swoboda KJ et al. 2018. CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nat Commun. 9(1):4619. eng.

Stromswold K. 1998. Genetics of spoken language disorders. Hum Biol. 70(2):297-324.

Thevenon J, Callier P, Andrieux J, Delobel B, David A, Sukno S, Minot D, Anne LM, Marle N, Sanlaville D. 2013. 12p13. 33 microdeletion including ELKS/ERC1, a new locus associated with childhood apraxia of speech. European Journal of Human Genetics. 21(1):82.

Tomblin JB, O'Brien M, Shriberg LD, Williams C, Murray J, Patil S, Bjork J, Anderson S, Ballard K. 2009. Language features in a mother and daughter of a chromosome 7;13 translocation involving FOXP2. J Speech Lang Hear Res. 52(5):1157-1174.

Turner SJ, Hildebrand MS, Block S, Damiano J, Fahey M, Reilly S, Bahlo M, Scheffer IE, Morgan AT. 2013. Small intragenic deletion in FOXP2 associated with childhood apraxia of speech and dysarthria. American journal of medical genetics Part A. 161(9):2321-2326.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics. 43(1):11.10. 11-11.10. 33.

Vernes SC, Newbury DF, Abrahams BS, Winchester L, Nicod J, Groszer M, Alarcon M, Oliver PL, Davies KE, Geschwind DH et al. 2008. A functional genetic link between distinct developmental language disorders. N Engl J Med. 359(22):2337-2345.

Villanueva P, de Barbieri Z, Palomino HM, Palomino H. 2008. [High prevalence of specific language impairment in Robinson Crusoe Island. A possible founder effect]. Rev Med Chil. 136(2):186-192.

Villanueva P, Fernandez MA, De Barbieri Z, Palomino H. 2014. Consanguinity on Robinson Crusoe Island, an isolated Chilean population. J Biosoc Sci. 46(4):546-555.

Villanueva P, Newbury DF, Jara L, De Barbieri Z, Mirza G, Palomino HM, Fernandez MA, Cazier JB, Monaco AP, Palomino H. 2011. Genome-wide analysis of genetic susceptibility to language impairment in an isolated Chilean population. Eur J Hum Genet. 19(6):687-695.

Villanueva P, Nudel R, Hoischen A, Fernandez MA, Simpson NH, Gilissen C, Reader RH, Jara L, Echeverry MM, Francks C et al. 2015. Exome sequencing in an admixed isolated population indicates NFXL1 variants confer a risk for specific language impairment. PLoS Genet. 11(3):e1004925.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38(16):e164.

Wiszniewski W, Hunter JV, Hanchard NA, Willer JR, Shaw C, Tian Q, Illner A, Wang X, Cheung SW, Patel A et al. 2013. TM4SF20 ancestral deletion and susceptibility to a pediatric disorder of early language delay and cerebral white matter hyperintensities. Am J Hum Genet. 93(2):197-210.

Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. 2016. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics. 32(9):1423-1426.

Zweier C, de Jong EK, Zweier M, Orrico A, Ousager LB, Collins AL, Bijlsma EK, Oortveld MA, Ekici AB, Reis A et al. 2009. CNTNAP2 and NRXN1 are mutated in autosomal-recessive Pitt-Hopkins-like mental retardation and determine the level of a common synaptic protein in Drosophila. Am J Hum Genet. 85(5):655-666.

Appendices

Tables

Table 1 shows the numbers of variants found in the language gene regions by whole-genome sequencing, if the individual is directly related to the pair of original founder brothers, and whether they carry the p.Asp150Lys *NFXL1* variant. **A)** Results for the Typical Language Delay (TLD) control group, and **B)** results for the language disorder affected (TEL) group. **C)** The median number of variants for each level of filtering (Median), and the range across all 17 individuals.

A

	Typical Language Delay Controls									
Number of variants identified in:	**TLD-1**	**TLD-2**	**TLD-3**	**TLD-4**	**TLD-5**	**TLD-6**	**TLD-7**	**TLD-8**	**TLD-9**	**TLD-10**
Full regions	13,764	14,124	13,817	16,610	15,433	14,036	13,388	12,902	12,261	13,804
Coding	40	33	45	54	52	33	32	44	34	42
Nonsynonymous variants	15	14	18	20	19	14	10	15	10	17
Rare (≤1%) nonsynonymous variants	1	1	1	1	2	0	1	1	0	0
Founder brother related	Y	Y	Y	Y	Y	Y	Y	Y	N	N
NFXL1 p.Asp150Lys carrier	0/1	0/1	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0

B

	Language Disorder Affected Cases						
Number of variants identified in:	**TEL-1**	**TEL-2**	**TEL-3**	**TEL-4**	**TEL-5**	**TEL-6**	**TEL-7**
Full regions	14,320	12,918	12,864	15,761	14,129	14,169	12,492
Coding	35	44	51	47	41	32	37
Nonsynonymous variants	16	14	18	22	16	11	17
Rare (≤1%) nonsynonymous variants	1	1	2	3	2	0	0
Founder brother related	Y	Y	Y	Y	Y	Y	N
NFXL1 p.Asp150Lys carrier	0/1	0/1	0/1	0/0	0/0	0/0	0/0

C

Number of variants identified in:	**Total**	**Median**	**Range**
Full regions	33,968	13,817	12,902-16,610
Coding	137	41	33-52
Nonsynonymous variants	56	16	10-22
Rare (≤1%) nonsynonymous variants	15	1	0-3

Table 2 shows prioritised rare, nonsynonymous variants identified in the thirty-three language genes in 17 Robinson Crusoe Islanders, 7 with a diagnosis of TEL and 10 unaffected individuals (TLD). The table includes the genomic location (chromosome and position) of the variants, the gene in which it falls, the resulting coding and amino acid changes, and the transcript and corresponding exon. The rsID column indicates the dbSNP identifier, and no record is indicated by a full stop. The Pred. lists the combined *in silico* missense pathogenic prediction score (maximum score of 10). gnomAD MAF indicates the sum of minor allele frequencies list in the gnomAD database for the Latino (AMR) and combined (ALL) populations. No data, meaning an allele has not been detected in the control populations, is indicated by a full stop. Finally, the genotype of individual for each variant is indicated as wild-type (0/0), heterozygous (1/0) or homozygous (1/1).

							Genome MAF		ExAC MAF	TEL - low language, normal IQ							Control - normal language, normal IQ									
Chr	**Position**	**Gene**	**AA change**	**Exon**	**rsID**	**Pred.**	**AMR**	**ALL**	**ALL**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	125660609	*CNTNAP5*	c.C3584T p.T1195M	NM_130773 exon22	rs34165507	2	0.0095	0.0123	0.0214	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/0
3	77614248	*ROBO2*	c.T1826C p.M609T	NM_002942 exon12	rs764591005	4	.	.	8.29E-06	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/0	0/0	0/0
3	77623752	*ROBO2*	c.C2074T p.P692S	NM_002942 exon14	rs187796411	1	0	0.0026	0.001	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0
3	77657063	*ROBO2*	c.G3251T p.G1084V	NM_002942 exon21	rs375469648	4	0	0.0003	5.81E-05	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0
3	78717171	*ROBO1*	c.G1720A p.V574I	NM_001145845 exon13	rs141178745	1	0	0.0005	0.0005	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
4	47916175	*NFXL1*	c.C46T p.R16W	NM_001278623 exon2	.	2	.	.	.	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
6	24596420	*KIAA0319*	c.G347A p.R116Q	NM_001168376 exon2	rs147251392	0	0	6.46E-05	7.43E-05	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
8	41792377	*KAT6A*	c.A3361G p.I1121V	NM_006766 exon17	rs534998448	0	.	.	2.56E-05	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0
8	77690516	*ZFHX4*	c.A3166G p.T1056A	NM_024721 exon4	.	1	.	.	.	0/0	1/0	0/0	0/0	0/0	0/0	0/0	1/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
8	77761294	*ZFHX4*	c.T3575A p.L1192Q	NM_024721 exon7	.	0	.	.	.	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0
12	1599274	*ERC1*	c.G3145C p.E1049Q	NM_178039 exon18	rs759970105	0	.	.	8.25E-06	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
16	14334126	*MKL2*	c.C864A p.H288Q	NM_001308142 exon10	rs75963814	1	0.0095	0.0115	0.0103	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0
16	84438779	*ATP2C2*	c.C256T p.R86W	NM_001286527 exon3	rs200841107	7	.	.	1.66E-05	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
16	84442069	*ATP2C2*	c.A386C p.K129T	NM_001286527 exon4	rs375393459	4	.	.	2.50E-05	0/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
22	40662984	*TNRC6B*	c.G2750C p.G917A	NM_001162501 exon5	rs143708410	3	0.0024	0.0032	0.0041	0/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0

Table 3 shows the total number of rare variants found in the sequenced Robinson Crusoe Islanders across the entire gene region of the thirty-three language genes. **A**. The total number of rare variants per individual. **B**. The mean number of rare variants per group; TEL compared to TLD, and founder-brother-related compared to non-founder-brother-related, and the Students t-test to test for a statistical difference in means between the two groups.

A

	TEL Affected						
Sample	**TEL-1**	**TEL-2**	**TEL-3**	**TEL-4**	**TEL-5**	**TEL-6**	**TEL-7**
No. Rare Variants	445	340	320	732	291	340	266
Founder brother related	Yes	Yes	Yes	Yes	Yes	Yes	No

Total variants = 4998
Total mean variants = 362.32
Standard deviation = 164.52

	TLD Controls									
Sample	**TLD-1**	**TLD-2**	**TLD3**	**TLD-4**	**TLD-5**	**TLD-6**	**TLD-7**	**TLD-8**	**TLD-9**	**TLD-10**
No. Rare Variants	345	686	293	722	645	400	470	366	406	328
Founder brother related	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No

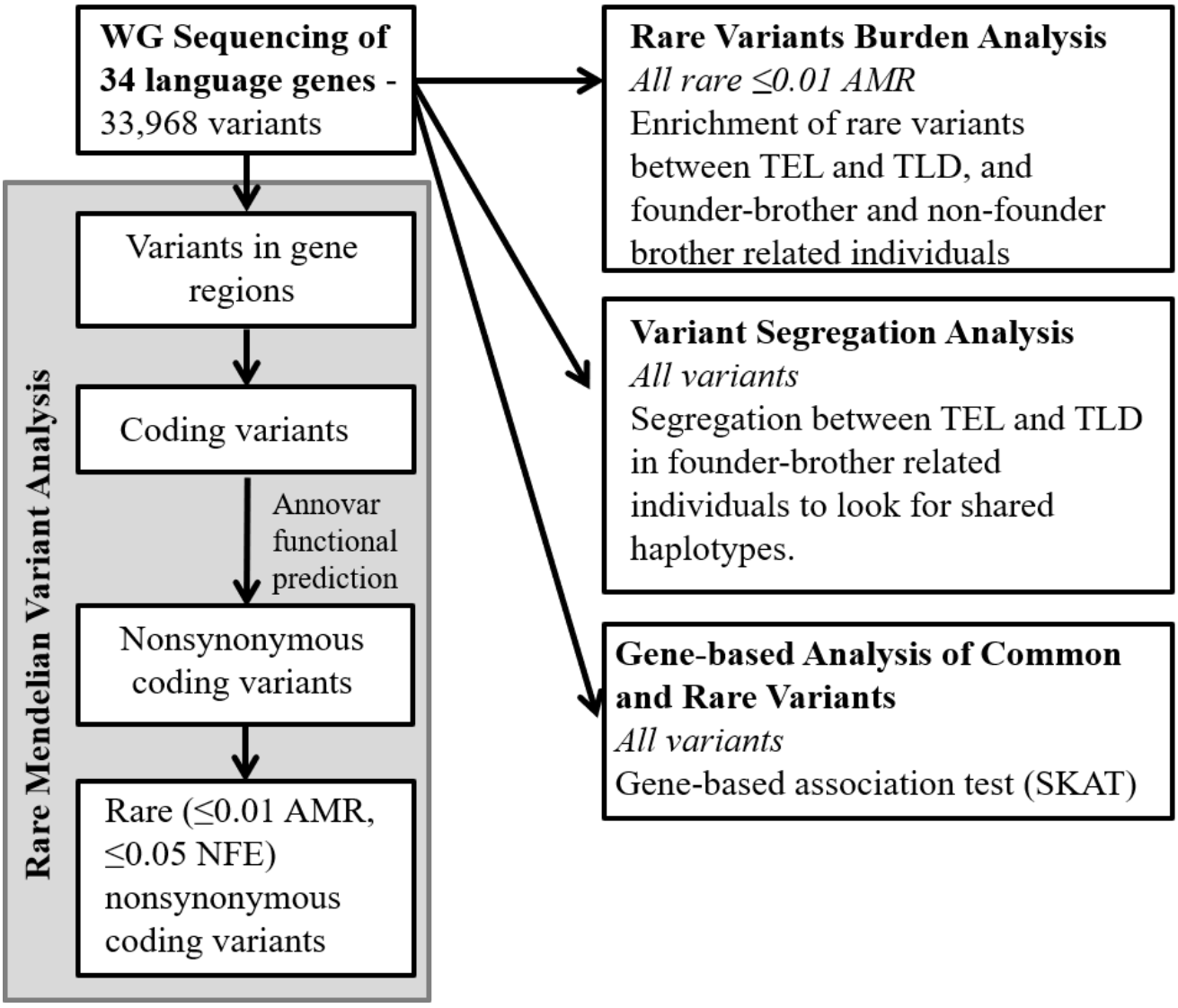
B

TEL n=7, TLD n=10	Founder brother related n=14, unrelated n=3
TEL affected mean = 390.571	Founder related mean = 456.786
TLD unaffected mean = 466.1	Founder unrelated mean = 333.333
Students t-test p= 0.35575	Students t-test p= 0.07431

Table 4 shows the gene-based association testing results using the SKAT test (RVTESTS) collapsed into gene regions. The number of variants tested within each gene, and permuted p values for each measure are reported. The greyed highlights represent nominally significant gene results.

		SKAT Association Test Permuted P value	
Gene	**No. of Variants**	**TEL**	**Founder-Brother Status**
ARHGEF39	15	0.730112	0.578505
ATP2C2	664	0.52521	0.203707
AUTS2	2661	0.961538	0.408514
BCL11A	177	0.873362	0.881289
CHD3	41	0.632511	0.573808
CMIP	946	0.680272	0.355643
CNTNAP2	8008	0.979432	0.220699
CNTNAP5	3164	0.668896	0.05785
DCDC2	802	0.453721	0.254134
DOCK4	1251	0.779423	0.228042
DYX1C1	30	0.925255	0.0097
ERC1	1956	0.979432	0.168808
FLNC	49	0.623053	0.456313
FOXP1	1390	0.845309	0.03645
FOXP2	481	0.275482	0.122056
GRIN2A	1733	0.585187	0.266962
GRIN2B	1361	0.207107	0.07865
KAT6A	218	0.983284	0.670584
KIAA0319	347	0.118494	0.0951
MKL2	348	0.965251	0.0343
NDST4	776	0.996016	0.464419
NFXL1	50	0.744039	0.194515
NOP9	18	0.0407	0.286312
RBFOX2	449	0.390168	0.026
ROBO1	3303	0.961538	0.338305
ROBO2	2289	0.904977	0.483285
SEMA6D	139	0.639475	0.733113
SETBP1	325	0.910747	0.008
SETD1A	53	0.88879	0.0424
TM4SF20	80	0.588928	0.253887
TNRC6B	290	0.462535	0.11708
WDR5	179	0.352983	0.3397
ZFHX4	373	0.0565	0.653796

Figures (on individual pages)


WG Sequencing of 34 language genes - 33,968 variants
Rare Mendelian Variant Analysis
Variants in gene regions
Coding variants
Annovar functional prediction
Nonsynonymous coding variants
Rare (≤0.01 AMR, ≤0.05 NFE) nonsynonymous coding variants
Rare Variants Burden Analysis
All rare ≤0.01 AMR
Enrichment of rare variants between TEL and TLD, and founder-brother and non-founder brother related individuals
Variant Segregation Analysis
All variants
Segregation between TEL and TLD in founder-brother related individuals to look for shared haplotypes.
Gene-based Analysis of Common and Rare Variants
All variants
Gene-based association test (SKAT)

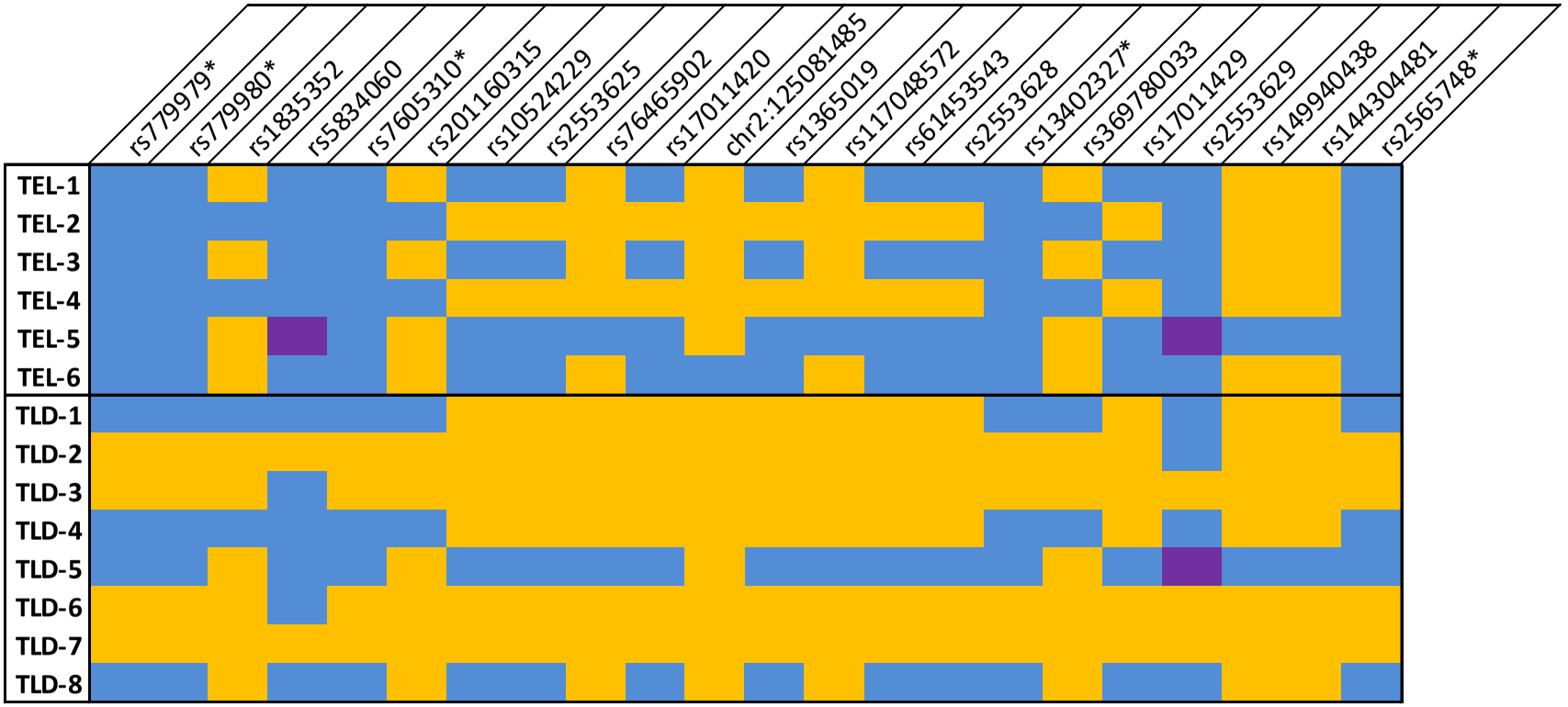

rs779979*
rs779980*
rs1835352
rs5834060
rs7605310*
rs201160315
rs10524229
rs2553625
rs76465902
rs17011420
chr2:125081485
rs1365019
rs117048572
rs61453543
rs2553628
rs13402327*
rs369780033
rs17011429
rs2553629
rs149940438
rs144304481
rs2565748*
TEL-1
TEL-2
TEL-3
TEL-4
TEL-5
TEL-6
TLD-1
TLD-2
TLD-3
TLD-4
TLD-5
TLD-6
TLD-7
TLD-8

Figure Captions (as a list)

Figure 1 shows the analysis work flow from whole-genome sequencing variant calls in 17 TEL and TLD Robinson Crusoe Islanders. The variant filtering steps for the Rare Mendelian Variant Analysis are highlighted in grey (left). Rare variant burden, segregation and gene-based association analyses are described in the black boxes (right).

Figure 2 shows the genotype grid for the founder-brother related TEL and TLD individuals across the chr2:125,078,104-12,5086,667 region of *CNTNAP5*. The five prioritised variants from the segregation analysis are indicated by an asterisk. The genotype for each variant in the region are indicated as wild-type (orange), heterozygous (blue) and homozygous (purple). The genotype analysis shows affected individuals do not share a clearly defined single haplotype containing all five markers, and it therefore is unlikely to represent a causative region.

Supplementary Tables

Supplementary Table 1

Chrom	Start	End	Gene
2	124782864	125672863	*CNTNAP5*
2	60684329	60780633	*BCL11A*
2	228226874	228244022	*TM4SF20*
3	71003865	71633140	*FOXP1*
3	75986645	77699114	*ROBO2*
3	78646388	79817059	*ROBO1*
4	115748927	116035032	*NDST4*
4	47849258	47916680	*NFXL1*
6	24171983	24358280	*DCDC2*
6	24544332	24646383	*KIAA0319*
7	111366164	111846462	*DOCK4*
7	114055052	114333827	*FOXP2*
7	128470483	128499328	*FLNC*
7	145813453	148118088	*CNTNAP2*
7	69063905	70257885	*AUTS2*
8	41786997	41909505	*KAT6A (MYST3)*
8	77593515	77779521	*ZXHF4*
9	137001210	137025094	*WDR5*
9	35659341	35665278	*ARHGEF39 (C9orf100)*
12	1100404	1605099	*ERC1*
12	13714410	14133022	*GRIN2B*
14	24769098	24774374	*NOP9 (C14orf21)*

15	48009584	48066420	*SEMA6D*
15	55722506	55800432	*DYX1C1 (DNAAF4)*
16	14165196	14360630	*MKL2*
16	30968615	30995981	*SETD1A*
16	81478775	81745367	*CMIP*
16	84402133	84497793	*ATP2C2*
16	9847265	10276263	*GRIN2A*
17	7788123	7816075	*CHD3*
18	42260138	42457379	*SETBP1*
21	15646120	15673692	*ABCC13*
22	36134783	36424585	*RBFOX2*
22	40573929	40731812	*TNRC6B*