EJHG Open

European Journal of Human Genetics (2011) 19, 687–695
© 2011 Macmillan Publishers Limited All rights reserved 1018-4813/11
www.nature.com/ejhg

ARTICLE

Genome-wide analysis of genetic susceptibility to language impairment in an isolated Chilean population

Pia Villanueva*[,1,2,3,5], Dianne F Newbury*[,4,5], Lilian Jara[1], Zulema De Barbieri[2], Ghazala Mirza[4], Hernán M Palomino[3], María Angélica Fernández[2], Jean-Baptiste Cazier[4], Anthony P Monaco[4] and Hernán Palomino[1]

Specific language impairment (SLI) is an unexpected deficit in the acquisition of language skills and affects between 5 and 8% of pre-school children. Despite its prevalence and high heritability, our understanding of the aetiology of this disorder is only emerging. In this paper, we apply genome-wide techniques to investigate an isolated Chilean population who exhibit an increased frequency of SLI. Loss of heterozygosity (LOH) mapping and parametric and non-parametric linkage analyses indicate that complex genetic factors are likely to underlie susceptibility to SLI in this population. Across all analyses performed, the most consistently implicated locus was on chromosome 7q. This locus achieved highly significant linkage under all three non-parametric models (max NPL=6.73, P=4.0×10^{-11}). In addition, it yielded a HLOD of 1.24 in the recessive parametric linkage analyses and contained a segment that was homozygous in two affected individuals. Further, investigation of this region identified a two-SNP haplotype that occurs at an increased frequency in language-impaired individuals (P=0.008). We hypothesise that the linkage regions identified here, in particular that on chromosome 7, may contain variants that underlie the high prevalence of SLI observed in this isolated population and may be of relevance to other populations affected by language impairments.

European Journal of Human Genetics (2011) **19,** 687–695; doi:10.1038/ejhg.2010.251; published online 19 January 2011

Keywords: Specific language impairment (SLI); Robinson Crusoe Island; linkage; language

INTRODUCTION

Specific language impairment (SLI) is a profound deficit in the acquisition of language despite adequate intelligence and opportunity, in the absence of any possible medical aetiology.[1] This disorder is a common developmental condition affecting between 5% and 8% of pre-school children, and thus places a heavy burden upon health-related and educational services.[2] It is well documented that SLI has a strong genetic basis (reviewed by Stromswold[3]). However, it is proposed that susceptibility to this disorder is complex in nature involving multiple genes, in combination with environmental factors.[4] The genetic basis of complex disorders are notoriously difficult to characterise, as the contributing factors can vary greatly between affected individuals and may be masked by undetermined environmental effects. This is reflected in the fact that, to date, only four genetic loci[5–7] and three associated candidate genes[8,9] have been described for SLI (OMIM no. 606711 (SLI1), OMIM no. 606712 (SLI2), OMIM no. 607134 (SLI3), OMIM no. 612514 (SLI4), OMIM no. 612514 (*CNTNAP2*, SLI4) OMIM no. 613082 (*ATP2C2*, in SLI1) and OMIM no. 610112 (*CMIP* in SLI1)).

Isolated founder populations can provide an important resource in the identification of causal genes underlying complex disorders.[10] Such populations are derived from a small number of relatively recent ancestors and thus are relatively homogeneous, a point which can greatly assist gene mapping processes.[11] Furthermore, one may postulate that loci identified in founder populations may hold more relevance to the general population than those yielded by the study of rare monogenic forms of impairment. In 2008, Villanueva *et al*[12] described a Chilean founder population with an increased incidence of SLI (known as TEL in Spanish-speaking countries). This population inhabit the Robinson Crusoe Island, which forms part of the Juan Fernandez Archipelago, 677 km to the west of Chile, South America. Robinson Crusoe Island is the only inhabited island in the archipelago and has 633 residents. The most recent colonisation dates to the late nineteenth century when the island was repopulated by a group of eight families. A total of 77% of the current population has at least one of the colonising surnames supporting a high degree of consanguinity. Linguistic profiling of the colonising children indicated that 35% met current criteria for SLI (expressive or comprehensive language >2SD below that expected for their age), 27.5% had language deficits associated to other pathologies (eg, delayed psychomotor development, intellectual deficit or auditory impairment) and 37.5% displayed normal language skills.[12] In contrast, the frequency of SLI

[1]Human Genetics Division, Faculty of Medicine, University of Chile, Santiago, Chile; [2]School of Speech and Hearing Services, Faculty of Medicine, University of Chile, Santiago, Chile; [3]Department of Paediatric Dentistry and Dentomaxillary Orthopaedics, Faculty of Dentistry, University of Chile, Santiago, Chile; [4]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

*Correspondence: Professor P Villanueva, Faculty of medicine, University of Chile, Av. Independencia 1027, Independencia, Santiago, CP 8380453, Chile. Tel: +56 2 978 6606; Fax: +56 2 978 6608; E-mail: piavillahk@gmail.com

or Dr DF Newbury, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Headington, Oxford, OX3 7BN, UK. Tel: +44 1865 287510; Fax: +44 1865 287501; E-mail: dianne@well.ox.ac.uk

[5]These authors contributed equally to this work.

Received 27 April 2010; revised 17 November 2010; accepted 3 December 2010; published online 19 January 2011

in the non-colonising children (3.8%), coincided with that reported for mainland Chile (~4%).[13] Genealogical reconstruction indicated that 75% of known affected individuals were descended from a single pair of founder brothers.[12] This population therefore represents a rare resource, which may be valuable in the identification of genetic loci contributing to susceptibility to SLI.

In this study, we perform genome-wide loss of heterozygosity mapping and parametric and non-parametric linkage analysis of the Robinson Crusoe population. We identify five regions (on chromosomes 6, 7, 12, 13 and 17) that meet genome-wide significance, and several loci, which are consistently implicated across alternative analyses. We hypothesise that these regions may contain variants that underlie the high prevalence of SLI observed in this isolated population.

SUBJECTS AND METHODS

This work was approved by the ethics department of the University of Chile. Informed consent was given by all participants and/or, where applicable, their parents.

DNA was extracted from EDTA whole blood samples collected from all available SLI and language-normal probands and their immediate families (125 individuals from 34 families, Table 1) using a standard chloroform extraction protocol.

All Island inhabitants between 3 and 8 years, 11 months of age (n=66) were subjected to a linguistic battery, which included tests of phonology (Test para Evaluar Procesos de Simplificación Fonológica (TEPROSIF)[14]) and expressive and receptive morphosyntax (Toronto Spanish Grammar Exploratory test[15]). Any child who performed >2SD below that expected for their age was classified as having SLI. Exclusion criteria included non-verbal IQ (Columbia Mental Maturity Scale) below the 80th percentile, hearing disability, motor or structural abnormalities (Oral Motor and Speech Examination[16]) and a co-morbid diagnosis of autism, emotional difficulties, or neurological disorders (as assessed by medical history). Following proband ascertainment, available family members were assessed for the presence of SLI. Individuals who fell outside the age-range of available standardised tests (3 and 8 years 11 months) were assessed through a family history interview[17] and tests of verbal fluency (Barcelona test[18]), verbal comprehension (Token test[19]), non-verbal intelligence (Raven's progressive matrices[20]) and auditory screening. The identification and classification of probands formed part of the descriptive study by Villanueva *et al* (2008). As this previous manuscript was in Spanish, detailed assessment descriptions are provided as Supplementary Material.

Table 1 Sample structure

	N	*SLI (%)*	*Language normal (%)*
Probands	34	12 (35)	22 (65)
Sibs	22	5 (23)	17 (77)
Half-sibs	6	4 (67)	2 (33)
Parents	61	21 (34)	40 (66)
Total	**123**	**42 (34)**	**81 (66)**
Male	55	17 (39)	38 (47)
Female	68	25 (61)	43 (53)

Abbreviation: SLI, specific language impairment.
A total of 123 samples were analysed. These included 42 language impaired individuals and 81 language normal individuals.
Percentage of SLI and language normal probands, sibs, half-sibs, parents and totals are given as a percentage of the total number of the appropriate group.
Percentage of males and females are given as a percentage of the language group.
Values in bold are the total number of samples.

The present study considers only families derived from colonising families of the Robinson Crusoe Island (ie, at least one ancestor related to a founder member).

Genotyping

DNA was quantified by a pico-green assay (Quant-iT, http://www.invitrogen.com). In total, 125 samples were genotyped on the Illumina HumanLinkage-12 panel following the multi-sample Infinium II assay (http://www.illumina.com). These beadchips allow the genotyping of 6090 genome-wide single nucleotide polymorphisms (SNPs) and simultaneously analyse 12 DNA samples.

Quality control procedures

All genotypes were called within Beadstudio (Version 3, Illumina Inc., San Diego, CA, USA). Any SNP with a gentrain score below 0.9 was manually inspected and if, necessary, the clusters adjusted. A total of 18 samples were duplicated across arrays. Any SNP with a gentrain score below 0.5 (n=27), a call rate below 0.97 (n=4) or a minor allele frequency below 2.5% (n=2) was excluded from further analyses.

All called genotypes were subjected to a haplotypic error detection algorithm in MERLIN.[21] All identified unlikely genotypes ($P<0.001$) were re-examined and, if necessary, excluded. Probabilities of Hardy–Weinberg Equilibrium (HWE) were calculated within PEDSTATS[22] and any SNP with a HWE-p <0.001 (2 of 5666 SNPs examined) was identified for cautious treatment in the remaining analyses.

Allele-sharing between individuals was examined using the Graphical Representation of Relationships (GRR).[23] This software calculates mean Identity by State (IBS) values for all possible pairs of samples and clusters individuals accordingly. Any individual found to cluster outside the expected IBS values were further examined. This error checking stage identified two DNA samples that had been mislabelled and were therefore excluded.

Generation of linkage pedigrees

Genealogical information was collated from birth and marriage certificates, family names and parent and relative interviews. Known relationships between

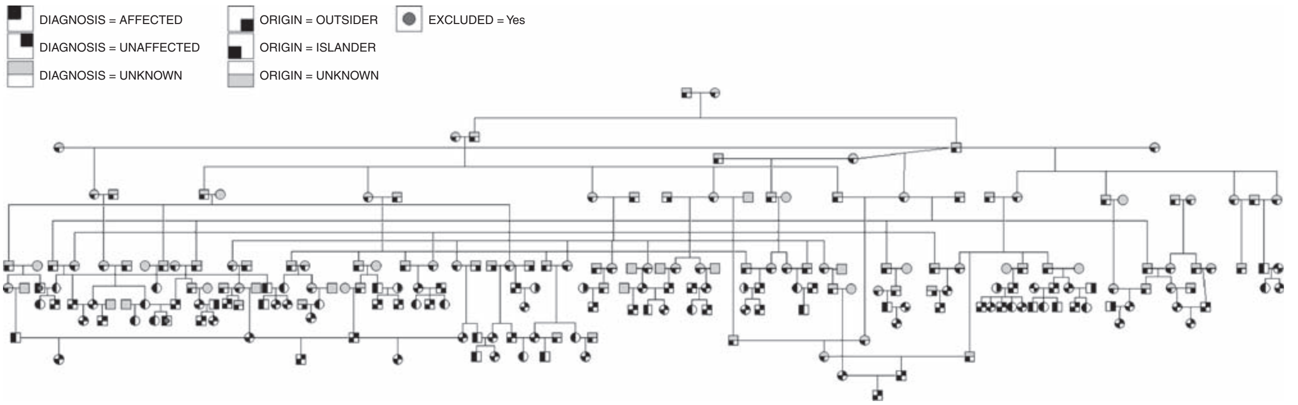

Figure 1 Descendants of founder brothers. The majority of affected individuals were found to be descended from a single pair of founder brothers.

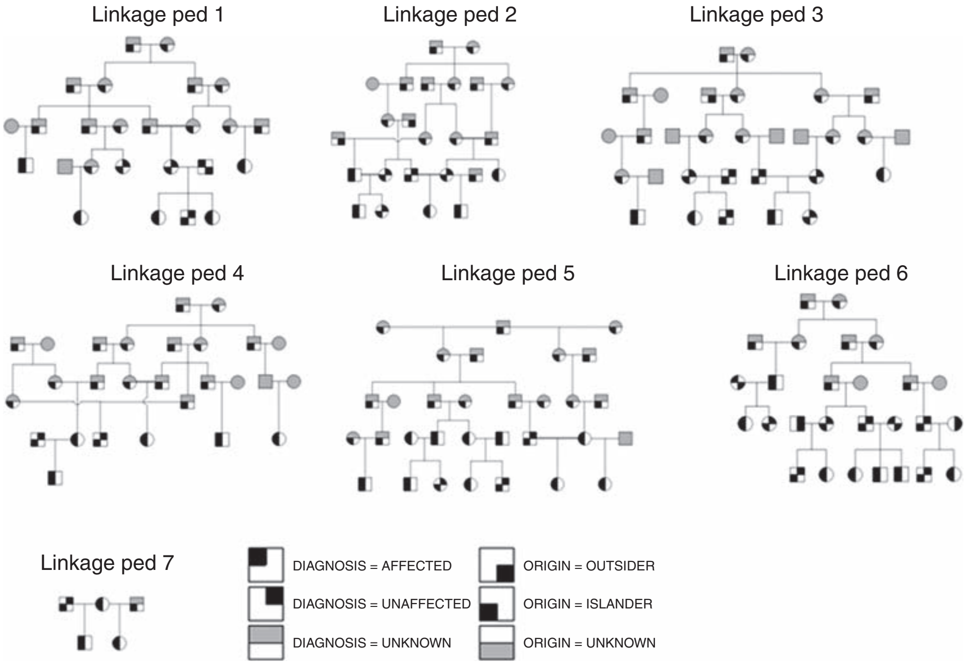

Figure 2 Small pedigrees used for linkage analyses. The larger pedigree shown in Figure 1 was broken into seven smaller pedigrees of maximum size 24 bits for linkage analyses.

identified nuclear families and the relevant pair of founder brothers were reconstructed and examined within the Progeny software (www.progenygenetics.com) (Figure 1).

Homozygosity mapping

Genotype data from all affected individuals were analysed for loss of heterozygosity within PLINK.[24] Sliding windows of 20-SNP genotypes were examined for runs of homozygosity. In all, 42 affected individuals from 23 nuclear families were examined including 2 affected sib-pairs, an affected trio of siblings and 3 affected half-sib-pairs. Previous studies have found that runs of homozygosity <4 Mb are common in outbred individuals.[25] Segments were therefore defined as homozygous tracts if 10 homozygous SNPs were found to extend across a region greater than 4 Mb in size.

Linkage analyses

Genotype data were stored in the Integrated Genotyping System (IGS) database. Individuals were classified as affected or unaffected on the basis of linguistic testing as described in 'subjects'. Data were analysed for linkage within MERLIN (autosomes) and the MERLIN extension, MINX (X chromosome).[21] As linkage packages were unable to analyse genome-wide data for the 242-bit pedigree as a whole, it was broken into sub-pedigrees. This segmentation was manually performed on the basis of closest shared ancestor. Seven extended families of 20–24 bits (where a bit is defined as 2×the number of non-founders−the number of founders) were defined and included 41 affected individuals and 63 of the 123 genotyped individuals (Figure 2). Although some individuals were present across multiple sub-pedigrees, all affected-relative pairs were only represented once. Genotype data for unaffected individuals were used for haplotype analyses (described below).

Parametric linkage analyses were performed under dominant and recessive models of linkage assuming a disease frequency of 35% (as described in the Robinson Crusoe population) and full penetrance. As the model of inheritance for SLI is unknown (and not expected to be monogenic in most instances) we also performed non-parametric analyses. Although explicit input parameters are not necessary for the completion of non-parametric analyses, expected allele frequencies must be specified. In this study, because of the isolated nature of the population, we had no directly appropriate control data and therefore performed three non-parametric analyses using alternative allele frequency estimation strategies. First, we used allele frequencies of all genotyped individuals (n=123). These individuals are derived directly from the population under study and can therefore be expected to provide representative expected allele frequencies. Nonetheless, these data are derived from related individuals and can therefore lead to a bias. We therefore repeated the analyses using allele frequency data from genotyped founder individuals of the generated sub-pedigrees (ie, those who marry into the pedigree, n=9). Although this reduces the dependence between individuals, it relies upon a small number of data points. We therefore also performed linkage analysis using allele frequency data from 60 unrelated CEPH individuals. The Y chromosome SNP data of the Robinson Crusoe population indicated that the founder males were European in origin (data not shown). Non-parametric results are reported as NPL scores and threshold levels for genome-wide significance are in line with that suggested by Kruglyak and Lander.[26] Namely, NPL scores of >3.8 ($P=7.4\times10^{-4}$) are described as suggestive linkage, NPL scores >4.08 ($P=2.2\times10^{-5}$) as significant and NPL scores >4.99 ($P=3.0\times10^{-7}$) as highly significant. Using a Bonferroni multiple testing correction for the three non-parametric analyses run, these thresholds equate to $P=2.46\times10^{-4}$, $P=7.3\times10^{-6}$ and $P=1.0\times10^{-7}$, respectively. In this instance, we expect the Bonferroni correction to be over-conservative because of the high-expected correlation between the three analyses.

Haplotype analyses

Haplotypes were reconstructed for the chromosome 7 region of linkage within nuclear 2-generation families using MERLIN.[21] Two-SNP sliding windows were visually inspected for allele combinations that co-segregated with affection status. All haplotypes that were found to have odds ratios of >2.0 or <0.5 (n=5) were analysed for association within PLINK using all genotyped cases and controls under a linear model.[24] In these analyses, no correction was made for the relationships between cases and controls. Association analyses of simulated data-sets yielded a distribution of empirical P-values that fit well with those expected under the theoretical model indicating that, in this particular case, the relationships between individuals do not inflate the significance of the results obtained (data not shown). Measures of linkage disequilibrium (LD) were calculated within haploview.[27]

RESULTS

Pedigree reconstructions confirmed that of the 44 affected individuals from whom we had DNA, 37 (84%) were descendants of a pair of

Table 2 Homozygous segments shared between more than two affected individuals

Chromosome	*Start*	*End*	*Size*	*Number of SNPs*	*Number of Inds*	*Homozygous individuals*
2	169 542 195	173 937 368	4 395 173	9	2	relationship unknown
4	73 731 890	78 761 621	5 029 731	11	2	relationship unknown
6	71 779 542	77 471 874	5 692 332	15	3	unrelated
6	77 572 235	77 572 235	0	1	2	
6	87 364 428	87 532 681	168 253	3	2	
6	88 115 604	92 044 752	3 929 148	17	3	unrelated
6	92 098 625	93 639 259	1 540 634	9	2	
7	108 674 847	114 462 759	5 787 912	13	2	relationship unknown
8	18 755 221	19 196 467	441 246	3	2	
8	19 559 214	23 746 576	4 187 362	16	3	1 sib pair and 1 unrelated
9	83 685 047	93 408 941	9 723 894	20	2	relationship unknown
10	3 791 413	9 111 974	5 320 561	22	2	relationship unknown
11	54 867 814	54 867 814	0	1	3	
11	55 360 988	59 674 738	4 313 750	12	4	1 sib pair and 2 unrelated
11	59 957 022	60 616 462	659 440	2	2	
13	41 575 238	46 431 276	4 856 038	14	2	relationship unknown
14	20 899 244	24 950 428	4 051 184	14	2	unrelated
14	94 718 410	98 165 036	3 446 626	8	2	
14	98 298 832	99 813 174	1 514 342	8	3	unrelated
14	100 345 436	101 474 494	1 129 058	4	2	
15	36 837 208	36 837 208	0	1	2	
15	37 016 395	37 119 086	102 691	5	3	
15	37 318 605	43 474 548	6 155 943	16	4	1 sib pair and 2 unrelated
15	81 012 306	88 150 562	7 138 256	20	2	relationship unknown
16	15 723 647	19 953 169	4 229 522	18	2	unrelated
19	50 678 730	57 149 118	6 470 388	19	2	relationship unknown
20	52 260 700	58 384 823	6 124 123	18	2	relationship unknown
21	32 754 546	35 223 308	2 468 762	5	2	
21	35 233 892	38 156 688	2 922 796	18	3	unrelated
21	38 599 459	39 524 326	924 867	3	2	

Abbreviation: SNPs, single nucleotide polymorphisms.
Start and end positions give positions of the extremities of overlapping segments between all individuals (in bp, B36). Boxed segments are contiguous.

founder brothers (Figure 1), 3 (7%) were not related to the founder brothers and 4 (9%) had unknown ancestry. Following quality control, genotypes were available for 6009 SNPs (5666 autosomal) with an average spacing of one SNP every 490 kb. The average genotype call rate was 99.9%. The minimum SNP genotype rate was 94.3% and the minimum SNP heterozygosity was 4%. Two individuals (both affected) were excluded from the analyses yielding genotype data for 123 individuals with an average individual genotype rate of 99.9% and a minimum individual genotype rate of 99.2%. The genotype mismatch rate across duplicated samples was 0.0027% and two SNPs were found to have a Hardy–Weinberg *P*-value of <0.001.

Of the 42 affected individuals examined, 28 showed at least one tract of homozygosity. Across all affected individuals, an average of 13.1 Mb (median, 5.4 kb) of the genome consisted of homozygous tracts. In individuals whose parents were known to be first or second cousins ($n=6$), this figure increased to 26.3 Mb (median, 28.9 kb). No chromosome region was found to be homozygous in all affected individuals, but two chromosome regions were homozygous in four (10%) affected individuals. These comprised of a 4 Mb region of chromosome 11 and a 6 Mb region of chromosome 15, both of which were homozygous in a sib-pair and two additional unrelated individuals (Table 2). In total, 18 chromosome regions contained overlapping segments of homozygosity (Table 2).

No chromosome region reached parametric genome-wide significance (HLOD$>$3, Figure 3). Maximum HLODs were observed on chromosome 8 for the dominant model (rs1390950, HLOD=2.4, Figure 3) and chromosome 1 for the recessive model (rs1906255, HLOD=1.52, Figure 3). Under the recessive model, chromosome 15 gave HLOD scores marginally above 1 (maximum HLOD=1.05) in a region that was homozygous in four affected individuals (Table 3).

Non-parametric linkage analyses identified five chromosome regions (chromosomes 6q, 7, 12, 13 and 17) that reached the threshold for genome-wide significance (NPL$>$4.08, $P<2.2\times10^{-5}$). Three of which (chromosomes 6q, 7 and 12) were highly significant (NPL$>$4.99, $P<3.0\times10^{-7}$) (Table 3, Figure 3), even after an over-conservative Bonferroni correction for the three non-parametric tests performed. The linkages to chromosomes 6q and 12 were only observed in a single non-parametic analysis whereas those to chromosomes 7, 13 and 17 were consistent across all non-parametric analyses performed (Figure 3, Table 3).

The most consistently implicated region was on chromosome 7. This locus achieved highly significant linkage under all three non-parametric models (max NPL=6.73, $P=4.0\times10^{-11}$) and contained a region, which gave a HLOD of 1.24 in the recessive parametric linkage analyses and a segment that was found to be homozygous in two affected individuals (Table 3, Figure 4). Linkage analyses within each of the sub-pedigrees, revealed that four families were contributing to linkage at this locus (linkage peds 3, 5, 6 and 7 (Figure 2), data not shown). Segregation analyses of two-SNP sliding window across this

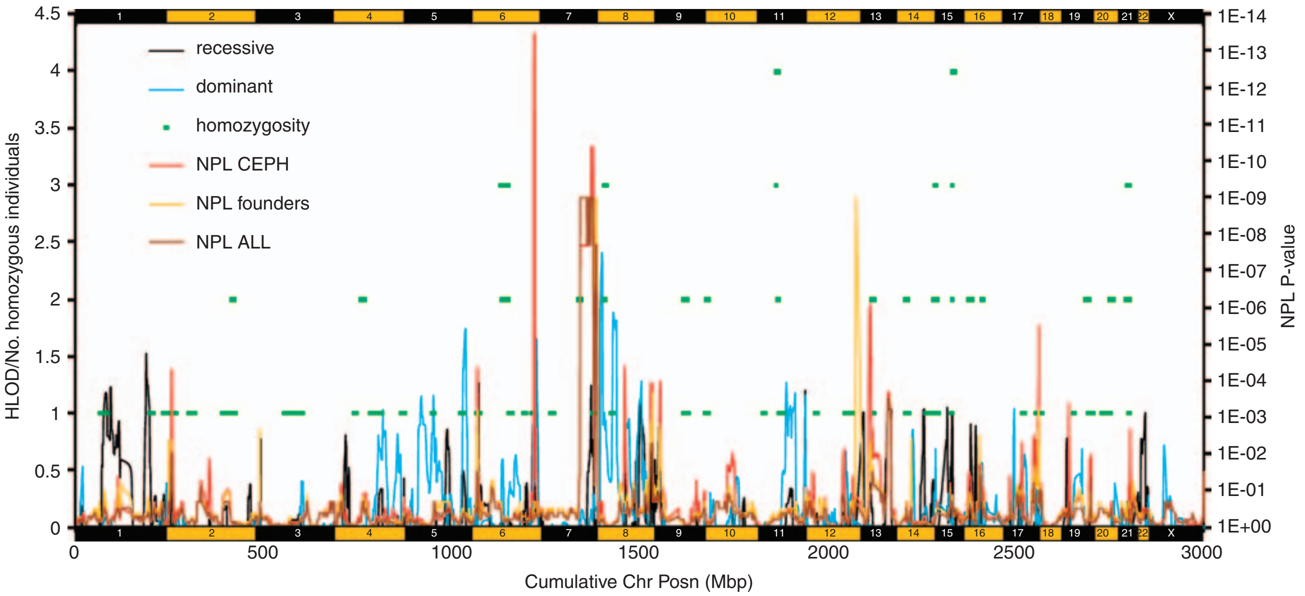

Figure 3 Genome-wide linkage analyses. Traces are shown for parametric analyses using both dominant and recessive models with full penetrance and three non-parametric models utilising expected allele frequencies derived from CEPH population, from genotyped founders in the sub-pedigrees and from all genotyped individuals. Traces are also shown for identified stretches of homozygosity (where the X-axis represents the number of individuals found to be homozygous across the region).

region identified five 2-SNP combinations that were present in at least 90% of affected individuals. Further investigations in all genotyped individuals, indicated that one of these haplotypes (rs727714/rs969356, AG) occurred at a significantly lower frequency in unaffected individuals than affected (Supplementary Table 1). The AG genotype of the rs727714/rs969356 haplotype was present in 98% of cases and 76% of controls and had an allele frequency of 67% in cases and 48% in controls (P=0.008). This association remains marginally significant (P=0.04) after the application of a Bonferroni correction. This haplotype covered 74 kb of sequence and coincided with the non-parametric (All) peak of linkage. It lay 2.5 Mb proximal to the SNPs with the highest NPL score in the two alternative non-parametric analyses (rs1524341 and rs1024676, D′=0.21–0.23, Tables 3 and 4) and was ∼2.6 Mb distal to the region of parametric linkage and 3 Mb proximal to a segment of homozygosity. Investigation of the LD structure indicated that the rs727714/rs969356 haplotype showed moderate (D′>0.4 and LOD>2) long-range LD with surrounding variants (Table 4), which may provide an alternative explanation for the association observed. One of the two haplotype SNPs (rs727714) falls in exon 3 of the *NOBOX* gene creating a synonymous base substitution.

As expected, given the density of the panel used in this study, single SNP association across the entire region of linkage on chromosome 7 did not identify any significant associations (min$P$ across linkage region=0.02, Figure 4). As single SNPs, rs727714 and rs969356 yielded association P-values of 0.04 and 0.13, respectively.

DISCUSSION

In this paper, we perform genome-wide analyses of an isolated Chilean population affected by Specific Language Impairment (SLI). Homozygosity mapping and parametric linkage analyses did not identify any chromosome segments that co-segregate with SLI in this population, suggesting that a completely penetrant monogenic aetiology is unlikely. This hypothesis is further supported by the observed nature of the language impairments. Affected individuals do not present with a specific core phenotype as may be predicted under a monogenic model, but instead show extensive heterogeneity in the severity and nature of impairment between affected individuals, as is typical of complex genetic forms of SLI.

The most consistent region of linkage extended across 48 Mb of chromosome 7q (chromosome position 111 285 062–158 710 965). This region reached a maximum NPL score of 6.73 (P=4.0×10^{-11}) and achieved genome-wide significance in all three non-parametric analyses performed and overlapped with a peak of parametric linkage (recessive model max HLOD=1.24) and two segments of homozygosity. Although these are not independent observations and a number of alternative analyses were performed, the reliability of the linkage in this region is consistent with that expected from a true positive.

Segregation analyses identified a two-SNP haplotype that was found at a marginally increased frequency in cases than controls (P=0.008). This haplotype fell across the *NOBOX* (OMIM no. 610934) and *TPK1* (OMIM no. 606370) genes. *NOBOX* is a homeobox gene, which is preferentially expressed in oocytes, but not reported to be expressed in brain.[28] *TPK1* encodes the thiamine pyrophosphokinase 1 enzyme, which catalyses the conversion of thiamine to thiamine pyrophosphate. Thiamine (or vitamin B1) is essential for the metabolism of carbohydrates into glucose and acts as a co-enzyme in the production of acetylcholine. Thiamine deficiency forms part of numerous disorders including ataxia, confusion and impaired memory.[29] Interestingly, a recent study suggested a link between thiamine deficiency and syntactic and lexical disorder.[30] The chromosome 7 peak also overlaps with the AUTS1 locus of linkage to autism[31] and includes both the *FOXP2* and *CNTNAP2* genes, both of which have previously been associated with language disorders.[9,32] The genotyping panels utilised in this study were optimised for linkage investigations and thus involve a relatively sparse map of SNPs (∼1SNP every 500 kb). The fine mapping of these regions is therefore required to enable the identification of candidates in an unbiased manner. We found that the two-SNP haplotype on chromosome 7 showed moderate long-range

Table 3 Chromosome regions highlighted by linkage analysis

Chr	*Max rec HLOD*	*SNP (chr posn) of max rec HLOD*	*Max dom HLOD*	*SNP (chr posn) of max dom HLOD*	*CEPH max NPL score (P-value)*	*SNP (chr posn) of max CEPH NPL score*	*Founder max NPL score (P-value)*	*SNP (chr posn) of max founder NPL score*	*All max NPL score (P-value)*	*SNP (chr posn) of max all NPL score*	*No. Homo Runs*
1	1.23	rs2129975 (92054668)					1.26 (0.10)	rs792321–rs481387 (74454556–74483038)			1
1	1.52	rs1906255 (186438670)					1.05 (0.15)	rs2039759 (191870092)			
2					3.90 (5.0×10^{-5})	rs3102960 (8369948)	2.63 (0.004)	rs2001660 (9530892)	2.37 (0.009)	rs2001660 (9530892)	
4			1.03	rs11098966 (129566420)							1
5			1.15	rs1553578 (40704941)							
5			1.15	rs1200485 (72420814)							1
5			1.74	rs270664 (158489316)							1
6	1.25	rs6926835 (12335531)			3.96 (4.0×10^{-5})	rs761116 (9566315)	3.23 (0.0006)	rs761116 (9566315)	1.96 (0.03)	rs2876143 (8928623)	1
6			1.65	rs880900 (167059469)	7.56 (3.2×10^{-14})	rs927450–rs675162 (160102086–160696315)					1
7	1.24	rs1476640-rs768055 (141058779–141059520)			6.73 (4.0×10^{-11})	rs1524341–rs1024676 (146337622–146346794)	6.10 (9.9×10^{-10})	rs1524341–rs1024676 (146337622–146346794)	6.00 (9.9×10^{-10})	rs928916–rs969356 (141954283–143804256)	2
8			2.40	rs1390950 (11633238)							3
8			1.88	rs749540 (41600253)							1
8					3.95 (4.0×10^{-5})	rs268564 (71646063)					
8	1.09	rs1353277 (116148616)	1.28	rs727581 (116719666)							
8					3.69 (0.0001)	rs1375062–rs768803 (142034963–143203532)	3.04 (0.0013)	rs1868280–rs1375062 (141965436–142034963)	2.59 (0.005)	rs4246828–rs9071 (144240466–145721314)	
9					3.72 (0.0001)	rs717081 (20277139)	3.61 (0.0002)	rs1532310–rs1532309 (592986–593192)	1.21 (0.11)	rs263580 (17029312)	
11			1.27	rs1945906 (81238725)							
11	1.20	rs3345 (131346779)	1.16	rs2044727 (131644201)	1.44 (0.08)	rs3345 131346779	1.25 (0.10)	rs3345 (131346779)			
12					1.13 (0.13)	rs937538–rs7960480 (132120315–132288239)	6.14 (9.9×10^{-10})	rs595241–rs632610 (132164652–132170791)			
13	1.01	rs1572372-rs3847993 (19738004–20193194)									
13					4.78 (8.0×10^{-7})	rs980285 (37937932)	2.66 (0.004)	rs980285 (37937932)	1.89 (0.03)	rs1544295 (42396689)	2

Table 3 (Continued)

Chr	Max rec HLOD	SNP (chr posn) of max rec HLOD	Max dom HLOD	SNP (chr posn) of max dom HLOD	CEPH max NPL score (P-value)	SNP (chr posn) of max CEPH NPL score	Founder max NPL score (P-value)	SNP (chr posn) of max founder NPL score	All max NPL score (P-value)	SNP (chr posn) of max all NPL score	No. Homo Runs
13	1.04	rs2044348 (88700438)			3.49 (0.0002)	rs996297–rs979969 (84905498–87095102)	3.47 (0.0003)	rs407218–rs979969 (85344625–87095102)	3.43 (0.0003)	rs407218–rs979969 (85344625–87095102)	
14	1.04	rs961700-rs1015023 (67401190-67431984)									
15	1.05	rs1557874 (23139577)			1.57 (0.06)	rs2596156 (31529470)	1.42 (0.08)	rs2596156 (31529470)	1.25 (0.11)	rs2596156 (31529470)	4
17			1.04	rs2215054 (12779734)							
17					4.49 (3.0×10^{-6})	rs1046875–rs1046896 (78278715–78278822)	3.26 (0.0006)	rs1046875–rs1046896 (78278715–78278822)	1.31 (0.10)	rs1046875–rs1046896 (78278715–78278822)	1
22	1.00	rs1540297 (34440411)									

Abbreviations: Chr, chromosome; posn, position; max, maximum; SNP, single nucleotide polymorphism.
Results are given for any region that achieved a genome-wide suggestive NPL score>3.8 (P=0.0007) or a parametric HLOD>1. Information is also given regarding any homozygous stretches identified in linkage regions (No. Homo Runs). The positions of homozygous tracts are given in Table 2.
Non-parametric linkage analyses were completed using; (i) the CEPH allele frequencies (CEPH columns), (ii) allele frequencies calculated from the genotype data of the founder members (founder columns) and (iii) allele frequencies calculated from the all genotyped individuals (All columns). Only NPL-scores of greater than 1.0 are reported.
Parametric linkage analyses were completed under a dominant (dom) and recessive (rec) model. Only HLOD scores>1.0 are reported.

linkage disequilibrium with a number of SNPs indicating that further information would be required to narrow the linkage peak. Higher density SNP arrays would also enable the detection of smaller runs of homozygosity.

We did not observe any linkage to chromosomes 16 or 19, which have previously been implicated in SLI.[5,6,33] Again, this may be caused by the low density of markers investigated in the present study. Alternatively, as the loci on chromosome 16 and 19 were identified by a quantitative genome screen of language-related measures, this may reflect differences in study design. As the Chilean quantitative linguistic data was collected only for subjects within a restricted age range (3 and 9 years), the current study utilised a binary affection status. This is similar to the approach applied by Bartlett *et al* (2002, 2004) in their genome screen for SLI in which they identified a region of linkage on chromosome 13 (SLI3), which overlaps with that found by the present study. This region has also been linked to autism,[34] a result which was strengthened by the selection of families on the basis of linguistic data.[35] Our chromosome 13 linkage consisted of two adjacent peaks. The distal peak (34–48 Mb) overlapped with a segment of homozygosity and achieved a maximum NPL score of 4.8 (P=8.0×10^{-7}) using CEPH allele frequencies. The proximal peak (83–94 Mb) reached an NPL of 3.5 (P=0.0002) under all non-parametric analyses performed and coincided with an area of marginal linkage under a recessive parametric model.

In addition to the linkages on chromosome 7 and 13, we also observed significant linkage (NPL>4.08 ($P<2.2\times10^{-5}$)) to chromosome 17 and highly significant linkages (NPL>4.99 ($P<3.0\times10^{-7}$)) to chromosomes 6q and 12 (Figure 3, Table 3). However, these peaks were only observed under a single non-parametric model and not in models using alternative expected allele frequencies. It is therefore likely that these divergent results may be driven by differences in the allele frequencies of the control populations used and illustrate the importance of correctly estimating allele frequencies, especially for markers that are in linkage disequilibrium.[36] Indeed, we found that the correlation of expected allele frequencies between the three different control groups was moderate (0.41–0.70 across all SNPs) and was lower than average across the conflicting regions of linkage on chromosome 6 and 12 (as low as 0.29 and 0.09, respectively), but remained moderate across the region of linkage on chromosome 7 (0.48–0.67). Importantly, simulation studies indicate that although allele frequency misspecification can lead to false positives, this artefact is not expected to affect the power to detect true linkages.[37] Thus, although the loci on chromosome 6 and 12 reached a threshold of highly significant linkage, as these were observed with only one non-parametric analysis, we must recognise the possibility that they represent false positives, especially given the high number of tests performed. Instead, a more fruitful avenue of investigation may be provided by the examination of regions found to be consistently implicated across all three analyses performed, even in cases where this linkage did not reach genome-wide significance (eg, chromosome 2, 6p, 8, 9, 15 and 17. Table 3, Supplementary Figure 1).

In conclusion, this study has applied a genome-wide approach to identify loci which may contain genes underlying susceptibility to SLI in an isolated population. This study represents the first step in the detection of genetic variants that underlie the increased frequency of language impairments in this population. It is envisaged that the fine mapping of the identified loci will allow the detection of associated polymorphisms. It is likely that the variants identified by the further study of this population will have a significant role in furthering our understanding of the genetic basis of language impairments and language development.

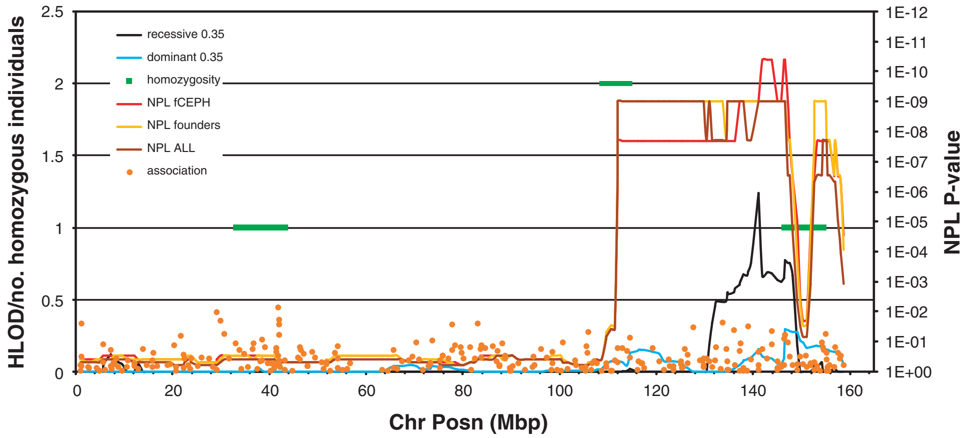

Figure 4 Chromosome 7. Chromosome 7 represented the most consistently linked locus across analyses. Traces are shown for parametric analyses using both dominant and recessive models with full penetrance, three non-parametric models utilising expected allele frequencies derived from CEPH population, from genotyped founders in the sub-pedigrees and from all genotyped individuals. Traces are also shown for identified stretches of homozygosity (where the *y* axis represents the number of individuals found to be homozygous across the region) and association of *P*-values (relative to the secondary *y* axis).

Table 4 **SNPs that are in LD with associated haplotype**

		D′ nuc	*D′ Confidence*	*LOD nuc*	*R*^{2} *nuc*	*D′*	*D′ Confidence*	*LOD*	*R*^{2}		
SNP	*Chr posn*	*Fams*	*Intervals*	*Fams*	*Fams*	*Fam5*	*Intervals*	*Fam5*	*Fam5*	*Genes*	*Peak of Linkage?*
rs2040587	114 462 759	0.23	0.04–0.44	0.66	0.04	0.50	0.27–0.66	2.99	0.14	300 kb downstream of *FOXP2*	
rs1962522	134 436 889	0.48	0.25–0.65	2.77	0.17	0.39	0.20–0.54	2.69	*0.14*	*AGBL3*	
rs10488598	136 238 383	0.56	0.28–0.76	2.58	0.11	0.53	0.21–0.73	1.74	*0.10*	*CHRM2*	
rs1371463	137 933 620	0.46	0.26–0.62	3.32	0.15	0.31	0.10–0.49	1.39	*0.07*	*SVOPL*	
rs1464798	138 982 540	0.64	0.36–0.81	3.25	0.15	0.62	0.39–0.77	4.26	*0.20*	*HIPK2*	
rs1476640	141 058 779	0.10	0.00–0.31	0.17	0.01	0.27	0.05–0.48	0.74	*0.04*	*WEE2*	Peak of parametric linkage
rs768055	141 059 520	0.80	0.47–0.93	3.61	0.17	0.76	0.34–0.91	2.04	*0.12*	*WEE2*	Peak of parametric linkage
rs727714	143 729 925	1.00	0.88–1.00	18.1	0.67	1.00	0.92–1.00	23.1	*0.70*	*NOBOX* (exon 3)	Associated haplotype SNP
rs969356	143 804 256	1.00	0.89–1.00	19.8	0.71	1.00	0.91–1.00	21.5	*0.65*	*TPK1*	Associated haplotype SNP
rs802200	145 736 404	0.42	0.23–0.58	3.00	0.17	0.39	0.18–0.55	2.21	*0.13*	*CNTNAP2*	
rs1524341	146 337 622	0.23	0.05–0.41	0.87	0.05	0.21	0.04–0.37	0.78	*0.04*	*CNTNAP2*	Peak of non-parametric linkage
rs1024676	146 346 794	0.22	0.03–0.43	0.52	0.03	0.23	0.03–0.45	0.51	*0.03*	*CNTNAP2*	Peak of non-parametric linkage
rs4431523	147 228 099	0.45	0.23–0.63	2.66	0.12	0.38	0.15–0.57	1.7	*0.07*	*CNTNAP2*	

Abbreviations: Chr, chromosome; posn, position; SNP, single nucleotide polymorphism.
Any SNP that has a D′ > 0.4 and a pairwise LOD > 2.0 with the associated haplotype is shown. Measures of LD were evaluated both in nuclear families and in the extended pedigree as shown in Figure 1.
The associated haplotype was formed from SNPs rs727714 and rs969356. These two SNPs gave the maximum NPL score of the non-parametric linkage analyses using allele frequencies from all genotyped individuals. The peak of linkage in the non-parametric analyses using allele frequencies from founder and CEPH individuals fell across SNPs rs1524341 whereas the peak of parametric linkage fell at SNPs rs1476640 and rs760885.
All SNPs are intronic unless otherwise stated.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The authors 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 Leopoldo Gonzalez 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. We would also like to thank Simon Fisher for his comments on the original manuscript. This work was specifically funded by a John Fell Fund award from Oxford University (http://www.ox.ac.uk) and was supported by core genotyping and statistical support facilities funded by a Wellcome Trust (http://www.wellcome.ac.uk) core award grant (grant no. 075491). 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.

1 Law J, Boyle J, Harris F, Harkness A, Nye C: Prevalence and natural history of primary speech and language delay: findings from a systematic review of the literature. *Int J Lang Commun Disord* 2000; **35**: 165–188.

2 Harel S, Greenstein Y, Kramer U *et al*: Clinical characteristics of children referred to a child development center for evaluation of speech, language, and communication disorders. *Pediatr Neurol* 1996; **15**: 305–311.

3 Stromswold K: The heritability of language: a review and meta analysis of twin adoption and linkage studies. *Language* 2001; **77**: 647–723.

4 Bishop DV: Genetic and environmental risks for specific language impairment in children. *Philos Trans R Soc Lond B Biol Sci* 2001; **356**: 369–380.

5 SLIC: A genomewide scan identifies two novel loci involved in specific language impairment. *Am J Hum Genet* 2002; **70**: 384–398.

6 SLIC: Highly significant linkage to the SLI1 locus in an expanded sample of individuals affected by specific language impairment. *Am J Hum Genet* 2004; **74**: 1225–1238.

7 Bartlett CW, Flax JF, Logue MW *et al*: A major susceptibility locus for specific language impairment is located on 13q21. *Am J Hum Genet* 2002; **71**: 45–55.

8 Newbury DF, Winchester L, Addis L *et al*: CMIP and ATP2C2 modulate phonological short-term memory in language impairment. *Am J Hum Genet* 2009; **85**: 264–272.

9 Vernes SC, Newbury DF, Abrahams BS *et al*: A functional genetic link between distinct developmental language disorders. *N Engl J Med* 2008; **359**: 2337–2345.

10 Sabatti C, Service SK, Hartikainen AL *et al*: Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* 2009; **41**: 35–46.

11 Gulcher J, Kong A, Stefansson K: The genealogic approach to human genetics of disease. *Cancer J* 2001; **7**: 61–68.

12 Villanueva P, de Barbieri Z, Palomino HM, Palomino H: High prevalence of specific language impairment in Robinson Crusoe Island. A possible founder effect. *Rev Med Chil* 2008; **136**: 186–192.

13 De Barbieri Z, Maggiolo L, Y A: Trastornos de la comunicación oral en niños que asisten a control de salud en un consultorio de atención primaria. *Rev Chil Pediatr* 1999; **70**: 36–40.

14 Maggiolo M, Pavez M: *Test para evaluar los procesos fonológicos de simplificación (TEPROSIF)*. Santiago: Escuela de Fonoaudiología, Facultad de Medicina, Universidad de Chile, 2000.

15 Pavez M: *Test exploratorio de Gramática española de A. Toronto*. Aplicación en Chile, Santiago: Ediciones Universidad católica de Chile, 2003.

16 Villanueva P: *Pauta de Examen en Habla y motricidad Orofacial*. Santiago: Escuela de Fonoaudiología, Facultad de Medicina, Universidad de Chile, 2000.

17 Tallal P, Ross R, Curtiss S: Familial aggregation in specific language impairment. *J Speech Hear Disord* 1989; **54**: 167–173.

18 Benito-Cuadrado MM, Esteba-Castillo S, Bohm P, Cejudo-Bolivar J, Pena-Casanova J: Semantic verbal fluency of animals: a normative and predictive study in a Spanish population. *J Clin Exp Neuropsychol* 2002; **24**: 1117–1122.

19 De Renzi E, Vignolo L: The Token Test: a sensitive tests to detect receptive disturbances in aphasics. *Brain* 1962; **85**: 665–678.

20 Raven J, Raven J, VCourt J: *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Harcourt assessment, 2003.

21 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30**: 97–101.

22 Wigginton JE, Abecasis GR: PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* 2005; **21**: 3445–3447.

23 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: GRR: graphical representation of relationship errors. *Bioinformatics* 2001; **17**: 742–743.

24 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.

25 McQuillan R, Leutenegger AL, Abdel-Rahman R *et al*: Runs of homozygosity in European populations. *Am J Hum Genet* 2008; **83**: 359–372.

26 Lander E, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11**: 241–247.

27 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.

28 Rajkovic A, Pangas SA, Ballow D, Suzumori N, Matzuk MM: NOBOX deficiency disrupts early folliculogenesis and oocyte-specific gene expression. *Science* 2004; **305**: 1157–1159.

29 Langlais PJ, Zhang SX, Savage LM: Neuropathology of thiamine deficiency: an update on the comparative analysis of human disorders and experimental models. *Metab Brain Dis* 1996; **11**: 19–37.

30 Friedmann N, Fattal I, Fattal-Valevski A: The effect of thiamine deficiency in infancy on the development of syntactic and lexical abilities. *Procedia Social and Behavioural Sciences* 2010; **6**: 168–169.

31 IMGSAC: A full genome screen for autism with evidence for linkage to a region on chromosome 7q. *Hum Mol Genet* 1998; **7**: 571–578.

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

33 Falcaro M, Pickles A, Newbury DF *et al*: Genetic and phenotypic effects of phonological short-term memory and grammatical morphology in specific language impairment. *Genes Brain Behav* 2008; **7**: 393–402.

34 CLSA: An autosomal genomic screen for autism. *Am J Med Genet* 2001; **105**: 609–615.

35 CLSA: Incorporating language phenotypes strengthens evidence of linkage to autism. *Am J Med Genet* 2001; **105**: 539–547.

36 Boyles AL, Scott WK, Martin ER *et al*: Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. *Hum Hered* 2005; **59**: 220–227.

37 Freimer NB, Sandkuijl LA, Blower SM: Incorrect specification of marker allele frequencies: effects on linkage analysis. *Am J Hum Genet* 1993; **52**: 1102–1110.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)