Supplemental Table 1. Inter-marker r^2 values for 3' *EN2* polymorphisms with *rs1861972* and *rs1861973*.

Dataset[a]	Polymorphism	*rs1861972*[b]	*rs1861973*[b]
	rs1895091	0.004	0.003
	rs12533271	0.000	0.001
	rs1861958	0.029	0.033
	rs3071184	0.041	0.048
	rs10259822	0.005	0.007
	rs10233570	0.017	0.020
	rs11976901	0.016	0.017
	rs10243118	0.042	0.046

[a] 167 initial AGRE I dataset described in [1].
[b] Inter-marker pairwise r^2 values calculated by GOLD program, version 1.0 [2].

Supplemental Table 2. Association analysis for 8 additional 3' *EN2* polymorphisms in the AGRE I dataset.

Dataset	Polymorphism	Diagnosis	χ^{2b}	*P*-value [c]
AGRE I[a]	*rs1895091*	narrow	0.000	1.000
		broad	0.000	1.000
	rs12533271	narrow	2.286	0.131
		broad	5.444	0.020
	rs1861958	narrow	1.750	0.186
		broad	0.583	0.445
	rs3071184	narrow	1.373	0.241
		broad	1.561	0.211
	rs10259822	narrow	0.177	0.673
		broad	1.923	0.165
	rs10233570	narrow	0.081	0.777
		broad	0.114	0.735
	rs11976901	narrow	3.073	0.080
		broad	3.500	0.061
	rs10243118	narrow	0.847	0.357
		broad	0.777	0.378

[a] The 167 AGRE families described in [1].
[b] Global χ^2 values calculated by PDTPHASE^{sum} [3].
[c] *P*-value generated by PDTPHASE^{sum} (1df).

Supplemental Table 3. Comparison of LD values in AGRE and CEU datasets.

		AGRE[a]		CEU[b]	
SNPs	Distance (Kb)	r^2	D'	r^2	D'
rs6460013-rs3824067	1.747	0.010	1.000	0.009	1.000
rs6460013-rs1861973	1.749	0.017	1.000	0.022	1.000
rs6460013-rs3808331	2.580	0.391	0.822	0.579	1.000
rs6460013-rs1861958	5.700	0.213	0.695	0.240	0.800
rs3824067-rs1861973	0.200	0.069	0.882	0.092	1.000
rs3824067-rs3808331	0.833	0.017	1.000	0.016	1.000
rs3824067-rs1861958	3.953	0.023	0.997	0.009	0.570
rs1861973-rs3808331	0.831	0.029	1.000	0.039	1.000
rs1861973-rs1861958	3.749	0.033	0.903	0.065	1.000
rs3808331-rs1861958	3.120	0.119	0.400	0.109	0.414

[a] AGRE I (167 families; 750 subjects).
[b] CEU (30 trios; 90 subjects).

Supplemental Table 4. *Rs1861972* and *rs1861973* association analysis for all individuals regardless of ethnicity and the White non-Hispanic subset.

Polymorphism	Diagnosis	All ethnicities[a]		White non-Hispanic[b]	
		χ^{2c}	*P*-value[d]	χ^{2c}	*P*-value[d]
rs1861972	narrow	10.79	0.00102	9.47	0.00200
	broad	13.69	0.00022	11.32	0.00076
rs1861973	narrow	15.48	0.00008	15.18	0.00010
	broad	16.96	0.00004	16.02	0.00006

[a] *rs1861972 and rs1861973* association data from AGRE I, AGRE II, and NIMH datasets (518 families, 2336 individuals) previously described in [1].
[b] 489 families, 2266 subjects, 790 individuals diagnosed with autism (narrow), 938 individuals diagnosed with ASD (broad).
[c] Global χ^2 values calculated by PDTPHASE^{sum} [3].
[d] *P*-value generated by PDTPHASE^{sum} (1df).

Supplemental Table 5. *Rs1861972-rs1861973* haplotype association analysis for all individuals regardless of ethnicity and the White non-Hispanic subset.

Diagnosis	Haplotype	All ethnicities[a]		White non-Hispanic[b]	
		χ^{2c}	*P*-value[d]	χ^{2c}	*P*-value[d]
narrow	A-C		0.000021		0.000039
	A-T		0.000092		0.000108
	G-C		0.000092		0.000108
	G-T		0.002303		0.003140
	global	28.48	0.0000007	27.11	0.000001
broad	A-C		0.000006		0.000023
	A-T		0.000147		0.000195
	G-C		0.000147		0.000195
	G-T		0.001215		0.002142
	global	29.75	0.0000003	27.47	0.000001

[a] *rs1861972-rs1861973* association data for AGRE I, AGRE II, and NIMH datasets (518 families, 2336 individuals) previously described in [1].
[b] 489 families, 2266 subjects, 790 individuals diagnosed with autism (narrow), 938 individuals diagnosed with ASD (broad).
[c] Global χ^2 values calculated by PDTPHASE^{sum} [3].
[d] *P*-value generated by PDTPHASE^{sum} (1df).

Supplemental Table 6. SNPs displaying strongest r^2 with *rs1861973* in HapMap datasets.

	SNPs	Distance (Kb)	r^2
CEU[a]	*rs1861973-rs7789835*	16.1	0.369
	rs1861973-rs10251163	10.2	0.342
	rs10949799-*rs1861973*	21.9	0.296
YRI[b]	*rs6460001-rs1861973*	13.5	0.338
	rs1861973-rs10488639	26.5	0.245
	rs6976308-rs1861973	25.6	0.242
JPT[c]	*rs1861973-rs7784116*	895.5	0.493
	rs1861973-rs6950584	897.6	0.267
	rs12719693-rs1861973	295.0	0.232
CHB[c]	*rs1861973-rs7789835*	16.1	0.592
	rs1861973-rs4716609	15.5	0.591
	rs1861973-rs4716599	7.3	0.496

[a] 70.3% of AGRE I and II datasets are of Western/Northern European descent.
[b] 1.5% of AGRE I and II datasets are of African descent.
[c] 1.9% of AGRE I and II datasets are of Asian descent.

Supplemental Table 7. Association analysis for *rs2361688*, *rs3824068*, and *rs12533271* in AGRE I White non-Hispanic subset.

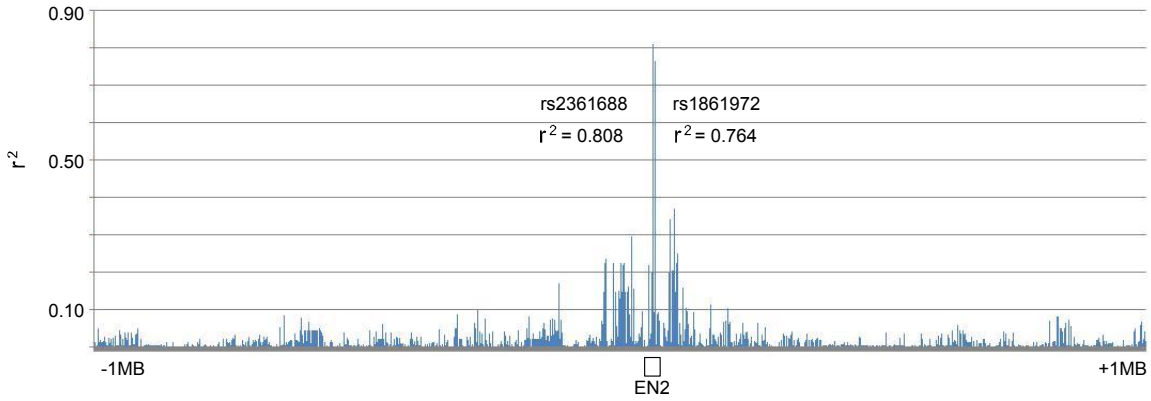
Dataset	Polymorphism	Diagnosis	All ethnicities[b]		White non-Hispanic[c]	
			χ^{2d}	*P*-value[e]	χ^{2d}	*P*-value[e]
AGRE I[a]	*rs2361688*	narrow	2.317	0.128	1.822	0.177
		broad	4.208	0.040	3.187	0.074
	rs3824068	narrow	4.372	0.036	4.490	0.034
		broad	2.664	0.103	2.290	0.130
	rs12533271	narrow	2.286	0.131	2.462	0.117
		broad	5.444	0.020	5.452	0.020

[a] The 167 AGRE families described in [1].
[b] *rs2361688* and *rs3824068* association data previously reported in Benayed *et al.* [5].
[c] 154 AGRE families, 686 subjects, 241 individuals diagnosed with autism (narrow), 298 individuals diagnosed with ASD (broad).
[d] Global χ^2 values calculated by PDTPHASE^{sum} [3].
[e] *P*-value generated by PDTPHASE^{sum} (1df).

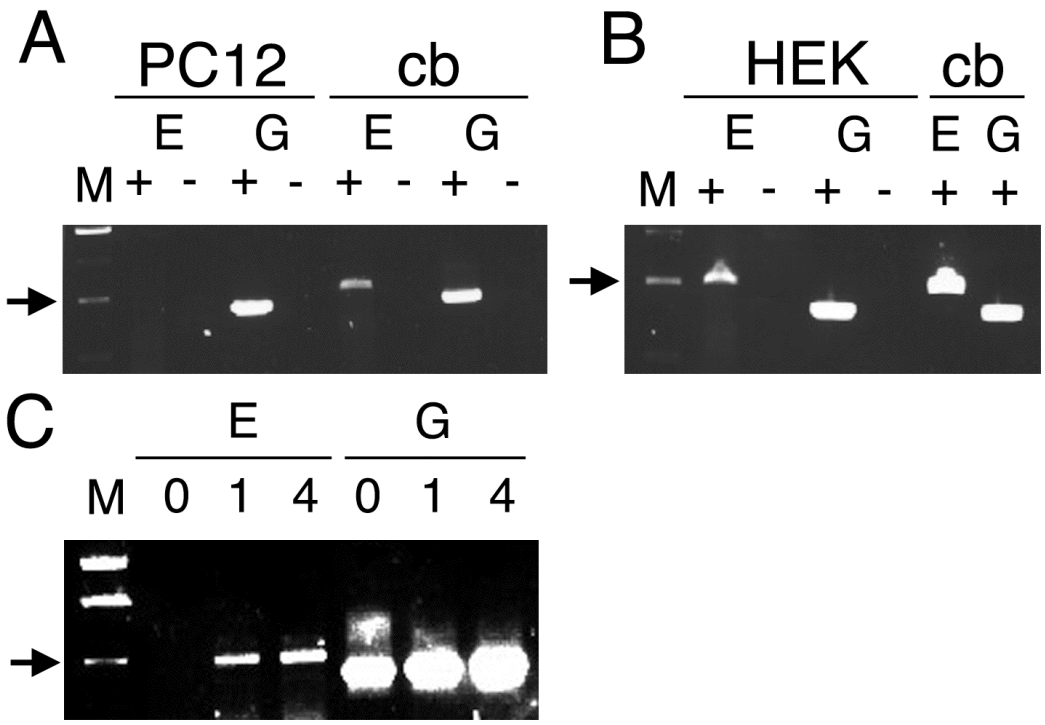
Supplemental Table 8. *Rs1861972* and *rs1861973* transcription factor bioinformatics.

SNP	Allele	Factor[a]	Sequence[b]	Score[c]
rs1861972	A	NF1	GCCA<u>A</u>TG	2.00
		NFY	CCA<u>A</u>T	1.99
		C/EBP	CCA<u>A</u>T	1.97
	A and G	Ets	CCTGC	1.96
rs1861973	C	Sp1	CTG<u>C</u>CC	2.00
		Ets	CCTG<u>C</u>	1.96
	C and T	MBF-1	CCAAAA	1.80

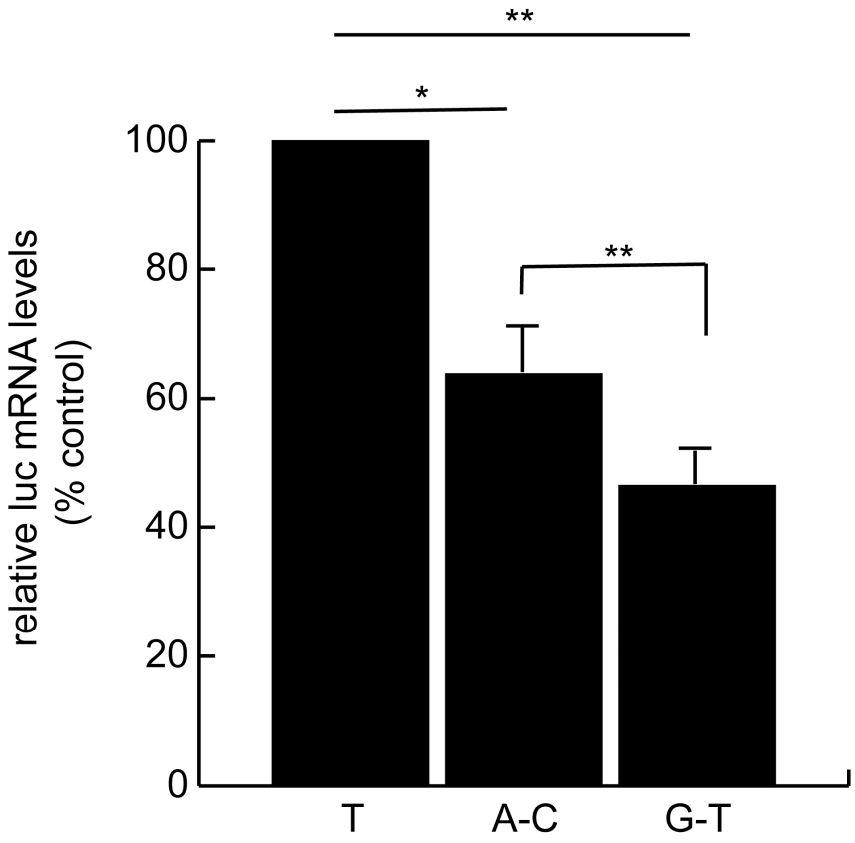
[a] TESS predicted transcription factor binding site.
[b] *EN2* sequence of predicted binding site with polymorphic allele underlined.
[c] TESS La/ score (0-2); log-likelihood score/base pair length of site with 2.0 being the maximum score.



Supplemental Figure 1. *ENGRAILED 2* LD map for White non-Hispanic subset. Inter-marker r^2 values with *rs1861973* are shown, including 26 *EN2* polymorphisms typed in a White non-Hispanic subset of the AGRE I dataset (154 families) and 3120 CEU SNPs within 2Mb of *EN2*. Only *rs1861972* and *rs2361688* display high r^2 values (>.75) with *rs1861973*, but *rs2361688* is not associated with ASD in the White non-Hispanic subset (Supplemental Table 6), identifying *rs1861972* and *rs1861973* as candidate risk alleles.

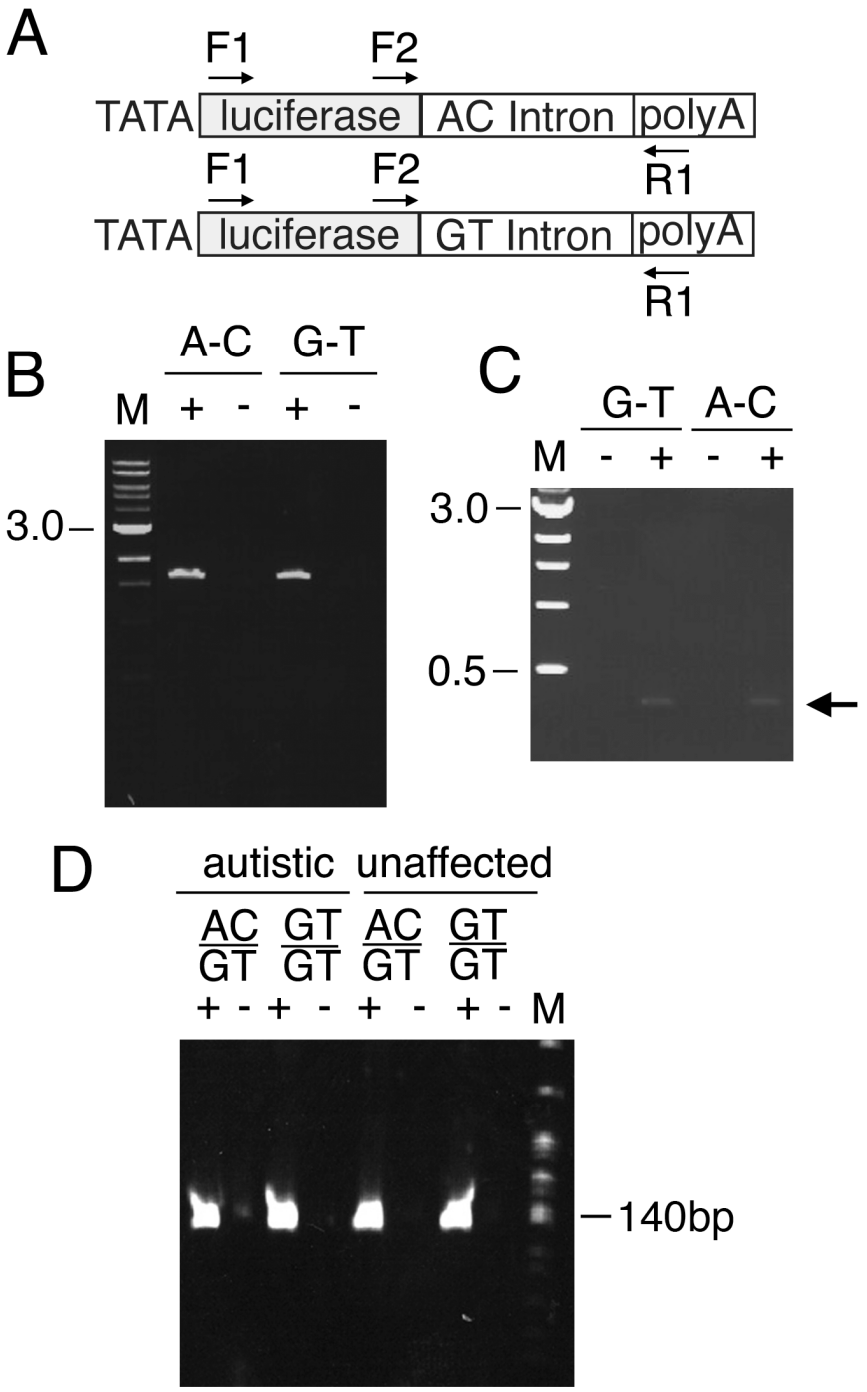


Supplemental Figure 2. *EN2* expression in HEK293T, PC12 and cerebellar granule cells (**A-C**). To investigate whether *EN2* is expressed in the cell types used for our transfection analysis, RTPCR experiments were performed for *EN2* and *GAPDH* on cDNA generated from PC12 (**A**), HEK293T (**B**) cell lines and cerebellar granule cells isolated from P6 pups and cultured for 0, 1 or 4 days *in vitro* (**C**). HEK293T and PC12 cell lines were originally generated from human and rat cells respectively so adult human and rat cerebella were used as positive controls. The expected sizes for *EN2* and *GAPDH* are 1002bp and 836bp. *En2* is expressed in granule cells cultured for 1 and 4 days *in vitro*, HEK293T cells and as expected rat and human cerebella. No *En2* expression was detected in PC12 cells. *GAPDH* was expressed in all cells and tissues. Arrow: 1.0 kb. Cb, cerebellum; E, *EN2*; G, *GAPDH*; M, 1kb ladder; -, minus RT; +, plus RT; 0, 1 and 4, days *in vitro*.



Supplemental Figure 3. Luc q-RTPCR for *rs1861972-rs1861973* A-C and G-T haplotypes. To investigate whether the destabilized version of the luc protein underestimated the effect of the A-C haplotype, the TATA-Luc-Intron A-C (A-C) and G-T (G-T) constructs were transiently transfected into HEK293T cells and luc transcript levels were quantitated by qPCR. Relative luc mRNA levels were normalized to *Renilla reniformis* and expressed as percent of control, pgl3 promoter vector (T).

* P=.02, ** P<.008, two tailed paired Student's T test (n=3).



Supplemental Figure 4. Associated A-C *rs1861972-rs1861973* haplotype does not cause cryptic splicing. (**A**) To investigate whether the risk allele affects the splicing of the *EN2*

intron, the SV40 minimal promoter constructs diagrammed above were transiently transfected into HEK293T cells and RTPCR experiments were performed. The position of the forward (F1, F2) and reverse (R1) primers is depicted. (**B, C**) The RTPCR results are shown for two different primer sets: F1, R1 (**B**) and F2, R1 (**C**). For both experiments only the predicted PCR product (F1, R1: 1758bp; F2, R1: 342bp) was observed, indicative of proper splicing. M, 1kb ladder. (**D**) To investigate if the risk allele results in cryptic splicing of the *EN2* intron *in vivo*, RT-PCR was performed with primers to exon 1 and 2 and RNA isolated from human post-mortem cerebellar tissue. Regardless of affection status and genotype, only the PCR product (134bp) indicative of correct splicing was observed. M, pBR322 DNA MspI digest.

-, minus RT; +, plus RT.

Supplemental Materials and Methods

Genotyping, association and LD analysis

Prior to association analysis, each polymorphism was assessed for deviations from Hardy-Weinberg equilibrium using genotype data from all parents and standard formulae. The DNA from the MZ twins was used as a genotyping internal control, with complete genotypic concordance observed for all 5 MZ cotwins in AGRE I. Genotypes were checked for Mendelian inconsistencies using the PEDCHECK program version 1.1 [4] and all identified Mendelian errors were corrected by re-genotyping individual samples. 58 out of the 5960 genotypes (0.97%) could not be resolved. The inter-marker linkage disequilibrium coefficients (D' and r^2) between the 3' polymorphisms and *rs1861972* and *rs1861973* were calculated in the AGRE I dataset using the parental genotypes and the GOLD (version 1.0) program [2]. All single association analyses were performed using PDTPHASE (version 2.404) as described previously [3,5]. PDTPHASE can calculate two global scores: the $PDTPHASE^{sum}$ (which sums the level of significance from all families) and the $PDTPHASE^{ave}$ (which gives equal weight to all families in a data set). Since most families in our study have similar size as well as structure and the observed χ^2 distribution and *P*-values were similar for both PDT scores, only $PDTPHASE^{sum}$ are reported.

For *rs1895091*, *rs12533271*, *rs1861958*, *rs3071184*, *rs10259822*, *rs10233570*, *rs11976901*, and *rs10243118,* a ligase detection reaction (LDR) and the Luminex^{TM} 100 flow cytometry platform was used [6]. The forward and reverse primers for PCR as well as the allele specific and common primers for the LDR are listed for each polymorphism:

rs1895091 (F: CCGCATTTTTGTGTGTTGAA; R: ATCAGCTTCGCCTGTGCT; Luminex: A: CCCAGGTGGAATGATGAGTTTTA, C: CCCAGGTGGAATGATGAGTTTTC, common: TAAGCATGTCTGAGTCCAGAGG), *rs12533271* (F: CCGCATTTTTGTGTGTTGAA; R: ATCAGCTTCGCCTGTGCT; Luminex: A: GCTATTCCAGTGACTACATTTTTTA, T: GCTATTCCAGTGACTACATTTTTTT, common: AAATTTTGCGAAGTTCAACAGGTC), *rs1861958* (F: CCGCATTTTTGTGTGTTGAA; R: ATCAGCTTCGCCTGTGCT; Luminex: A: GGCGCTTTCCCCTGGGGACA, G: GGCGCTTTCCCCTGGGGACG, common: GGAACTTTTTACTCCACTGAGAC), *rs3071184* (F: CCGCTCATCAGGAAGTGTTT; R: GTCCTCAGACCCTTGAAACG; Luminex: Ins: TTCCAAACTTTCATGGCTTAAAAAAA, Del: TTTTCCAAACTTTCATGGCTTAAAAA, common: TTTTTTTTCTTCTCCCTGTAAAAGAA), *rs10259822* (F: CCGCTCATCAGGAAGTGTTT; R: GTCCTCAGACCCTTGAAACG; Luminex: A: GTTCCCAGGCGGAAGCGGGA, G: TTCCCAGGCGGAAGCGGGG, common: CCTCTGGGCTGGGCCTCTG), *rs10233570* (F: CCGCTCATCAGGAAGTGTTT; R: GTCCTCAGACCCTTGAAACG; Luminex: C: CGCATTCCCCGCGAACCCC, T: ACGCATTCCCCGCGAACCCT, common: GACTTCTGAACAGTTCAGAAAGTT), *rs11976901* (F: CCGCTCATCAGGAAGTGTTT; R: GTCCTCAGACCCTTGAAACG; Luminex: A: TAAAAACCAAACAAAAAATACTGGCA, G: AAAAACCAAACAAAAAATACTGGCG, common: CAACTTCTATTGCCGTATTGGCC), *rs10243118* (F:

CCGCTCATCAGGAAGTGTTT; R: GTCCTCAGACCCTTGAAACG; Luminex: C: CACAGCTTTGTAAGGTAATGAGC, T: GCACAGCTTTGTAAGGTAATGAGT, common: TCCCCGTCCTCAGAGGTGTTC). PCR was conducted in a 20µl reaction using 0.4 µM of each F and R primer, 0.125mM dNTP, 31.25 mM KCl and 10 mM Tris-HCl (pH 8.8), $MgCl_2$ (1.75mM for *rs1895091*, *rs12533271*, and *rs1861958* and 0.75mM for *rs3071184*, *rs10259822*, *rs10233570*, *rs11976901*, and *rs10243118*). Standard cycling conditions were used: 94°C, 4min, 1x; 94°C, 30s, T_m°C, 30s, 74°C, 40s, 35 x; 74°C, 10min, 1x (T_m =59°C for *rs1895091*, *rs12533271*, and *rs1861958,* T_m =62°C for *rs3071184*, *rs10259822*, *rs10233570*, *rs11976901*, and *rs10243118*). For the LDR, 1µl of PCR product, 40 U of Taq DNA ligase (NEB), 1X Taq DNA ligase buffer, and 0.15 pmoles of each of the allele specific primers and common primer were used. PCR cycling conditions for the LDR reaction were as follows: 95°C for 1min (1 cycle); 95°C for 15sec, 58°C for 2min (21 cycles).

Constructs

To test for a functional difference between the haplotypes, the *EN2* intron was isolated from human BAC # RP1160D5 (A-C haplotype) obtained from BACPAC Resources center using *AccI* and *KpnI* and cloned in the multiple cloning site of pBluescript II KS + vector (Stratagene, La Jolla, CA). It was then isolated from pBluescript II KS+ using *KpnI* and *EcoRI* and subcloned in a *NotI* site of a pTimer-1 vector (Clontech, Mountain View, CA) with appropriate adapters. The resultant intron flanked by 37 bp of exon 1 and 120 bp of exon 2 was then subcloned into *XbaI* site of pGL3-basic vector (Promega, Madison, WI) and pGL3-promoter vector using appropriate adapters. The A-C haplotype

was converted to the G-T haplotype by two successive rounds of site directed mutagenesis using the QuickChange XL kit (Stratagene). Nucleotide changes were confirmed by sequencing both DNA strands. 5.5kb of *EN2* promoter was isolated from the same human BAC described above using *SacI* and *SgrAI* restriction enzymes and cloned into the *SacI* site of pGL3-basic vector 5' of luciferase coding sequence using appropriate adapters.

***EN2* expression analysis.** RNA from PC12 cells, HEK293T cells, P6 granule cells cultured for 0, 1, and 4 days *in vitro* as well as rat and human cerebellum was isolated using the *mir*Vana^{Tm} PARIS^{Tm} kit (Ambion, Austin, TX). First-strand cDNA was generated using 2µg of RNA, random hexamers (250ng), an *EN2*-specific reverse primer that was conserved across species (100ng) and SuperScript^{Tm} II Reverse Transcriptase (20U) (Invitrogen, Carlsbad, CA). Primers were designed using primer3 (http://primer3.sourceforge.net). Primer sequences and PCR conditions are listed below.

The following reverse primer was used for all RT-PCR assays: CTACTCGCTGTCCGACTTGCC. The following species-specific forward primers were used (PC12 cells, rat cerebellum: ATGGAGGAGAAGGATTCCAAGTCC; HEK293T cells, human post-mortem sample: ATGGAGGAGAATGACCCCAAGCC; granule cells: ATGGAGGAGAAG GATTCCAAGCC). GAPDH primers were as follows: PC12 cells, rat cerebellum: F: AACGGATTTGGCCGTATCGGA, R: TTGCTGTTGAAGTCACAG GAGAC; HEK293T cells: F: TCGCTGTTGAAGTCAGAGGAG AC, R: AACGGATTTGGTCGTATTGGG; granule cells: F:

AACGGATTTGGCCGTATTGGG, R: TTGCTGTTGAAGTCGCAGGAGAC. GAPDH RT-PCR assays were performed using the 2X GoTaq polymerase mix (Promega), 2.5mM dNTPs, 100ng of each primer and the following cycling conditions: primer annealing - 65ºC for 30sec; extension time - 1min, 30 cycles. *EN2* RT-PCR was performed using the Advantage® GC genomic Polymerase mix (6 Units) (Clontech) and the following amplification conditions: primer annealing- 60ºC to 56ºC with 1 degree decrement/cycle for 15 seconds; extension time - 1min; followed by primer annealing - 56ºC for 15 seconds; extension time - 1min, 25 cycles.

Splicing RT-PCR

A reverse primer mapping to the SV40 polyA sequence was used for the RT-PCR assay: TGGTTTGTCCAAACTCAT CAA. Two different forward primers mapping 5' and 3' within the luciferase coding sequence were used, respectively: F1: TGTTTGTGGACGAAGTACC G. F2: TGCACATATCGAGGTGGACATC. RT-PCR was performed using the Advantage® GC genomic Polymerase mix, 100ng of each primer, 2.5mM of dNTPs (primer annealing - 63ºC for 30 seconds; extension time - 3min 30sec, 30 cycles).

For the human post-mortem samples, the following RT-PCR conditions were used: 2µl of +RT or -RT product, 0.5µl Advantage GC genomic polymerase mix (Clontech), 1.5 M GC Melt, 1.1 mM Magnesium acetate, 0.5µM each of forward (exon1: ACTCGGACAGCTCGCAAGC) and reverse (exon2: CGGGTTCTTCTTTGGTTTTCG) primers and 2mM dNTPs in total reaction volume of 25µl. The PCR cycling conditions

were as follows: one cycle at 94ºC for 4min, 18 cycles each at 94ºC for 30sec, 63-58ºC for 15sec (with 1 degree decrement every 3 cycles) and 72ºC for 4min, and final 25 cycles of 94ºC for 30sec, 58ºC for 15sec and 72ºC for 4min.

qRT-PCR

qPCR primers specific for the luciferase coding sequence in pGL3 vectors are: F. TGCACATATCGAGGTGGACATC, R. GCCAACCGAACGG ACATTT. qPCR primers specific for Renilla luciferase coding sequence in phRL-null vector are F. CCTCACCGCTTGGTTCGA, R. CGTGGCCCACAAAGATGATT.

Electrophoretic mobility shift assays.

EMSA probe sequences for *rs1861972* are: A allele: CTCCCTGCCAATGGCCTTGCC ; G allele: CTCCCTGCCAGTGGCCTTGCC; T allele: GGCAAGGCCATTGGCAGGGAG ; C allele: GGCAAGGCCACTGGCAGGGAG; mutant CTCCCTGACACGGGCCTTGCC. For *rs1861973*, probes sequences are: C allele: AGCGACCCTGCCCAAAACCTG; T allele: AGCGACCCTGTCCAAAACCTG
G allele: CAGGTTTTGGGCAGGGTCGCT; A allele: CAGGTTTTGGACAGGGTCGCT; mutant: CAGGTTTTCTAGAATGTCGCT.

REFERENCES

1. Gharani N, Benayed R, Mancuso V, Brzustowicz LM, Millonig JH (2004): Association of the homeobox transcription factor, ENGRAILED 2, with autism spectrum disorder. *Mol Psychiatry* 9:474-484.
2. Abecasis GR, Cookson WO (2000): Gold--graphical overview of linkage disequilibrium. *Bioinformatics* 16:182-183.
3. Dudbridge F (2003): Pedigree disequilibrium tests for multilocus haplotypes. *Genet Epidemiol* 25:115-121.
4. O'Connell JR, Weeks DE (1998): Pedcheck: A program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* 63:259-266.
5. Benayed R, Gharani N, Rossman I, Mancuso V, Lazar G, Kamdar S, *et al.* (2005): Support for the homeobox transcription factor gene ENGRAILED 2 as an autism spectrum disorder susceptibility locus. *Am J Hum Genet* 77:851-868.
6. Iannone MA, Taylor JD, Chen J, Li MS, Rivers P, Slentz-Kesler KA, *et al.* (2000): Multiplexed single nucleotide polymorphism genotyping by oligonucleotide ligation and flow cytometry. *Cytometry* 39:131-140.