Am. J. Hum. Genet. 72:621–635, 2003

Extensive Linkage Disequilibrium, a Common 16.7-Kilobase Deletion, and Evidence of Balancing Selection in the Human Protocadherin α Cluster

James P. Noonan,[1] Jun Li,[3] Loan Nguyen,[1] Chenier Caoile,[3] Mark Dickson,[3] Jane Grimwood,[3] Jeremy Schmutz,[3] Marcus W. Feldman,[2] and Richard M. Myers[1,3]

[1]Department of Genetics, Stanford University School of Medicine, and [2]Department of Biological Sciences, Herrin Laboratories, Stanford University, Stanford, CA; and [3]Stanford Human Genome Center, Stanford University School of Medicine, Palo Alto, CA

Regions of extensive linkage disequilibrium (LD) appear to be a common feature of the human genome. However, the mechanisms that maintain these regions are unknown. In an effort to understand whether gene density contributes to LD, we determined the degree of promoter sequence variation in a large tandem-arrayed gene family, the human protocadherin α cluster, on chromosome 5. These genes are expressed at synaptic junctions in the developing brain and the adult brain and may be involved in the determination of synaptic complexity. We sequenced the promoters of all 13 α protocadherin genes in 96 European Americans and identified polymorphisms in the promoters α1, α3, α4, α5, α7, α9, α11, and α13. In these promoters, 11 common SNPs are in extensive LD, forming two 48-kb haplotypes of equal frequency, in this population, that extend from the α1 through α7 genes. We sequenced these promoters in East Asians and African Americans, and we estimated haplotype frequencies and calculated LD statistics for all three populations. Our results indicate that, although extensive LD is an ancient feature of the α cluster, it has eroded over time. SNPs 3′ of α7 are involved in ancestral recombination events in all populations, and overall α-cluster LD is reduced in African Americans. We obtained significant positive values for Tajima's D test for all α promoter SNPs in Europeans ($D = 3.03$) and East Asians ($D = 2.64$), indicating an excess of intermediate-frequency variants, which is a signature of balancing selection. We also discovered a 16.7-kb deletion that truncates the α8 gene and completely removes the α9 and α10 genes. This deletion appears in unaffected individuals from multiple populations, suggesting that a reduction in protocadherin gene number is not obviously deleterious.

Introduction

The human brain is the most complex product of vertebrate evolution. The processing power of the brain is due to the density and sophistication of the synaptic connections that form during its development and that are modified throughout life. The genetic basis of the mechanisms that generate diversity and specificity in synaptogenesis, however, is not understood. Members of the cadherin superfamily of calcium-dependent cell-adhesion molecules are known to be major structural and functional components of synapses (Fannon and Colman 1996; Uchida et al. 1996; Tang et al. 1998; Bruses 2000; Angst et al. 2001). Protocadherins are members of this family and are distinguished from classical cadherins by their greater number of ectodomain repeats, as opposed to the five repeats in classical cadherins (Suzuki 1996). A large cluster of highly similar protocadherin genes was recently identified on human chromosome 5 (Wu and Maniatis 1999). This cluster consists of 53 tandem-arrayed single-exon genes (excluding pseudogenes) organized into three subclusters, designated as "α," "β," and "γ" (GenBank accession numbers NG_000016, NG_000017, and NG_000012, respectively; for a graphical representation, see the Human Genome Brower, at the UCSC Genome Bioinformatics Web site, under contig number NT_029289) (Wu et al. 2001). The first protocadherin-cluster cDNAs identified were isolated in a yeast two-hybrid screen for mouse proteins interacting with the Fyn tyrosine kinase (Kohmura et al. 1998). These proteins, initially designated as "CNR" (cadherin-related neuronal receptor), are a subset of the mouse α protocadherins and are expressed at synaptic junctions in the developing brain and the adult brain (Kohmura et al. 1998). Individual neurons, including neurons of similar function, appear to express distinct combinations of these CNR genes. Under the assumption that the other protocadherins in the cluster are similarly expressed, an enormous amount of combinatorial complexity could arise from cell-to-cell variation in expression of all 53 protocadherin-cluster

Received October 8, 2002; accepted for publication December 5, 2002; electronically published February 7, 2003.

Address for correspondence and reprints: Dr. Richard M. Myers, Department of Genetics, M-344, Stanford University School of Medicine, Stanford, CA 94305-5120. E-mail: myers@shgc.stanford.edu

0002-9297/2003/7203-0013$15.00

members. These protocadherins could provide the molecular code required to generate synaptic specificity in both brain development and memory formation.

Each protocadherin exon encodes an extracellular domain consisting of six ectodomain repeats, a transmembrane domain, and a short cytoplasmic tail. Downstream of the α and γ subclusters in the mouse and human genomes are three short exons that are spliced to each α and γ exon; these three exons encode a common, "constant" cytoplasmic domain (fig. 1*A*) (Tasic et al. 2002; Wang et al. 2002). Each α and γ protocadherin protein thus has two major isoforms: a short form, encoded by each single exon, and a long form, including the constant region (Wu and Maniatis 1999). The β protocadherins lack a constant region and are exclusively single-exon genes that encode short-form proteins. We recently participated in a collaborative effort to sequence and annotate the orthologous mouse protocadherin cluster, on chromosome 18 (Wu et al. 2001). The organization of each cluster in mouse and human is very similar, suggesting that protocadherin-cluster genes have highly conserved functions in mammalian brain development.

Each α, β, and γ exon has its own compact promoter, located ~200 bp upstream of the translation start site (fig. 1*B*) (Wu et al. 2001; Tasic et al. 2002; Wang et al. 2002). These promoters are well conserved between mouse and human protocadherin orthologs and in most cases show more sequence variation among paralogs in the same species than between orthologs across species (Wu et al. 2001). These conserved elements provide a convenient and potentially highly informative way to study natural human variation in closely related regulatory sequences. DNA sequence variation in these elements could contribute to variation in protocadherin expression and could therefore influence both normal brain function and neuropathology. The pattern and distribution of regulatory polymorphisms in this gene cluster could also provide insights into mechanisms governing regional variation in recombination frequency throughout the human genome.

Regions of extensive linkage disequilibrium (LD) appear to be common in the human genome (Daly et al. 2001; Patil et al. 2001; Reich et al. 2001; Stephens et al. 2001; Gabriel et al. 2002). If extensive LD exists, then it can reduce the number of markers required to map both Mendelian-trait and complex-trait loci in genomewide association studies (Devlin and Risch 1995; Risch and Merikangas 1996; Goddard et al. 2000; Johnson et al. 2001; Pritchard and Przeworski 2001). The mechanisms that maintain these regions of extensive LD are not well understood. LD is expected to decline monotonically with recombinational map distance (Hill and Robertson 1968). Regions of extensive LD that are seen in Europeans and Asians are often reduced in older, more diverse populations, such as those of African descent, which show a greater number of ancestral recombination events (Tishkoff et al. 1996, 2000; Kidd et al. 1998; Reich et al. 2001; Gabriel et al. 2002). Recombination rates also vary across the genome, but the molecular basis of this variation is unknown.

Recombination may be discouraged in regions containing many highly similar repeated genes, because of the deletions and duplications that could result from unequal-crossover events. If this is so, then there should be extensive LD in clusters of tandem-repeated genes. Two recent studies of the T-cell receptor α/δ (Moffatt et al. 2000) and β loci (Subrahmanyan et al. 2001) report extensive LD across these gene clusters; this LD is irregularly distributed in large blocks, concomitant with reduced haplotype diversity. It is not clear from these studies whether there is any selective pressure to maintain these blocks in the context of the tandem gene array. However, given the high sequence similarity of the tandem-arrayed, paralogous protocadherin-cluster genes, recombination in the cluster could be highly prone to errors and could consequently be repressed. Accordingly, deletion or duplication of one or more protocadherin-cluster genes may be associated with recombination events. Patterns of recombination events could therefore vary across the cluster, with extensive LD anchored around the more essential members.

To address these questions, we measured the degree of promoter sequence variation for the 13 genes in the human protocadherin α cluster (MIM 604966). We discovered a number of common SNPs with >30% minor-allele frequency in these promoters and found that these polymorphisms are in strong LD. We determined allele frequencies, estimated haplotype frequencies, and calculated LD statistics for these polymorphisms in European, East Asian, and African American population samples. Our results indicate that extensive LD is an ancient feature of the α-cluster promoters. We also discovered a common (11% allele frequency in Europeans) 16.7-kb deletion that truncates the α8 gene and completely removes the α9 and α10 genes.

Material and Methods

Samples

We used DNA isolated from lymphoblastoid cell lines derived from members of 143 families having at least two children affected with autism. These families were recruited as part of our ongoing search for genetic factors contributing to this disorder (Spiker et al. 1994; Risch et al. 1999; Li et al. 2002). All samples were collected with the approval of the appropriate institutional review boards and with informed consent from the participants. For population studies, we used the HD50CAU (European), HD50AA (African Amer-

ican), HD32 (Chinese), HD07 (Japanese), HD13 (Southeast Asian), and HD12 (African) panels from the Coriell Cell Repository. Mbuti and Biaka Pygmy genomic DNA samples were generously provided by Drs. Peter Underhill and Luca Cavalli-Sforza.

Protocadherin Promoter PCR and Sequencing

PCRs were performed in 10 μl with 1 μM of each primer, 2–5 U of Ampli*Taq* Gold DNA polymerase (Applied Biosystems), 1 × GeneAmp PCR Gold buffer (Applied Biosystems), 2 mM of $MgCl_2$, 2.5 mM of each dNTP, and 25 ng of genomic DNA. Reactions were performed on GeneAmp 9700 thermal cyclers (Applied Biosystems) by denaturing at 95°C for 10 min, followed by 35 cycles at 94°C for 30 s, 58°C for 30 s, and 72°C for 30 s. Primers were designed using Primer3, with the modification of several search parameters (Beasley et al. 1999). All primers were synthesized at Invitrogen, were desalted, and were resuspended at a concentration of 10 μM. PCR products were treated with exonuclease and were sequenced as described elsewhere (Li et al. 2002). Products were sequenced in both directions by the forward and reverse primers used in the PCR. Sequencing products were run on Applied Biosystems 377 and 3700 sequencers. Sequence traces were evaluated by using Phred, Phrap, and Consed (Ewing and Green 1998; Ewing et al. 1998). Potential heterozygotes were flagged by Polyphred and were verified by manual inspection of each trace. All polymorphisms were verified on both forward and reverse sequences. For each promoter except α4 and α13, we defined a 401-bp window—from positions −350 to 50, relative to the translation start site of the associated gene (fig. 1*B*)—in which the average Phred quality score was 40 or greater in each sequenced sample. This window was 360 bp and 398 bp for the α4 and α13 promoters, respectively. We used only the sequence within this window in all our analyses. Samples giving ambiguous or low-quality genotypes were resequenced.

Mapping of the Deletion Endpoints by Real-Time and Conventional PCR

We developed a real-time, PCR-based method to assay variation in copy number at a particular locus among multiple individuals. In brief, amplicons were designed against the putatively hemizygous locus and a control locus of known normal copy number. The PCR kinetics at the control locus was used to control for sample-to-sample differences in genomic DNA purity and concentration. Three concentrations of each genomic DNA sample (40, 20, and 10 ng) were assayed in triplicate, using each pair of real-time–PCR primers. We used two regions, upstream of the α3 and α8 promoters, as our control loci. PCRs were prepared as follows: in 20 μl, we combined 4 μl of genomic DNA, 2 U of the Stoffel fragment of Ampli*Taq* polymerase (Applied Biosystems), 1 × Stoffel buffer, 0.4 μM of each primer, 0.25 × SYBR Green (Molecular Probes), and an amount of $MgCl_2$ (1.5–3.5 mM) optimized for each primer pair. PCRs were performed on the iCycler thermal cycler (BioRad), as follows: 95°C for 10 min, followed by 40 cycles at 95°C for 30 s, 60°C for 30 s, and 72°C for 30 s. We used the iCycler analysis software for PCR baseline subtraction and exported the data to Excel (Microsoft) for analysis. For each reaction, we normalized the relative fluorescence value for each cycle against the average of the last three cycles (38–40) of the trace. We then calculated a threshold cycle number (C_t) as the point at which the PCR reached 16% of its maximum value. C_t values were calculated for each amplicon–genomic DNA dilution pair. C_t values for the control and test amplicons for the three dilutions of each DNA sample were plotted against each other, and the offset between two samples along the control-amplicon axis and test-amplicon axis was measured. An offset of 0.8–1.2 along the test-amplicon axis was taken to indicate a copy number difference of 1 between the two samples at that locus. A deletion involving the α9 promoter was identified by comparing the copy number at the α9 locus between individuals showing violation of Mendelian inheritance of the α9 promoter SNP and individuals heterozygous for this SNP. We then mapped the endpoints of this deletion by using multiple PCR amplicons flanking α8 and α10, and, once the junction was defined to a sufficiently narrow region, we designed primers to amplify across this junction.

After determining the endpoints of the deletion, we developed a conventional PCR genotyping assay for the deletion allele. PCRs were performed in 10 μl with 1 μM of the forward primer (5′-GTGATTCGGGGTAATTTGGATTTT-3′), 1 μM of the reverse primer (5′-ACAAATTCATGGCATTGGTGTTT-3′), 2–5 U of Ampli*Taq* Gold DNA polymerase, 1 × GeneAmp PCR Gold buffer, 2.5 mM of $MgCl_2$, 2.5 mM of each dNTP, and 25 ng of genomic DNA. PCRs were performed on GeneAmp 9700 thermal cyclers by denaturing at 95°C for 10 min, followed by 35 cycles at 94°C for 30 s, 58°C for 30 s, and 72°C for 30 s. Products were run on 2% SeaKem LE (BioWhittaker) agarose gels in 1 × Tris-acetate-EDTA and were visualized with ethidium-bromide staining. Heterozygosity was determined by performing a parallel PCR assay on each sample with primers for the α9 promoter (α9F, 5′-CAGGGATAAGAAAACCACAATCAA-3′; α9R, 5′ CCACATTGCGAGGATCAGAAG-3′). PCR conditions were as for the deletion allele, except that the $MgCl_2$ concentration was reduced to 2 mM. The deletion-allele PCR product is 554 bp, and the α9 PCR product is 469 bp.

Statistical Analyses

All statistical genetic measures except for LD measures were calculated using procedures implemented in the Arlequin software package (Schneider et al. 2000). All haplotype estimations were made using the expectation-maximization procedure under default parameters (Excoffier and Slatkin 1995). Diversity indices were also calculated using default settings. To calculate Tajima's D value, Arlequin uses two estimates (θ_s and θ_π) of the population mutation parameter $\theta = 2M\mu$, where $M = 2N$ for diploid populations of size N and μ is the mutation rate. The θ_s value is estimated from the number of polymorphic sites (S) and the sample size (Watterson 1975). The θ_π value is estimated from the mean number of pairwise differences between haplotypes in the sample (Tajima 1983). The basis of Tajima's D value is the difference between these two estimates (Tajima 1989; Schneider et al. 2000). Under neutral expectation, $\theta_s = \theta_\pi$, and $D = 0$. Fu's F_s test compares the number of haplotypes observed with the number of haplotypes expected in a random sample under an infinite-sites model without recombination (Fu 1997). The statistical significance of values obtained from Tajima's D and Fu's F_s tests was based on values from 1,000 random samples of the same sample size and polymorphism level as the actual data. The minimum-spanning (MS) network among all haplotypes was calculated in Arlequin and was depicted by building a rooted tree in TreeView, using the chimpanzee haplotype as the outgroup. We then added alternative branches to this tree as indicated by Arlequin. LD statistics were calculated using 2LD (see the KCL, Institute of Psychiatry, Section of Genetic Epidemiology and Biostatistics, Web site) and LD Shell, a multilocus LD calculator (Zapata et al. 2001). We calculated D' (Lewontin 1964) and χ^2 likelihood-ratio P values for every pair of SNPs (table 3).

Promoter Cloning, Cell Culture, and Luciferase Assays

We amplified and cloned the following promoter sequences into the pGL3 firefly luciferase-expression vector (Promega): α3 TAAC and α3 GGGT, −423 to −13 (relative to the translation start site); α9 G and α9 A, −397 to −17; and α3 mouse, −362 to −10. These fragments contain all known conserved sequence elements, including the promoter. Human α3 and α9 promoter sequences were amplified from individuals homozygous for each variant. The mouse α3 promoter was amplified from BAC DNA (GenBank accession number AC020968). There is no clear, single α9 ortholog in the mouse.

We cultured undifferentiated P19 cells and induced differentiation as described elsewhere (Bain et al. 1998). Differentiated P19 cells were plated in 6-well or 12-well tissue-culture plates. Plates were treated at 37°C for 1 h with a solution containing 100 mM of sodium bicarbonate, 150 mM of sodium chloride, and 2 μg/ml of laminin (Sigma). These cells were cultured in Neurobasal medium (Invitrogen) plus 2 μg/ml of laminin, 1 × B27 supplement, and 250 μM of L-glutamine (Invitrogen). Dual-luciferase assays were performed by using the Dual-Luciferase Reporter System (Promega). We transfected each construct, in triplicate, into differentiated and undifferentiated P19 cells by using the Effectene transfection reagent (Qiagen). We cotransfected pRL-TK (Promega), which constitutively expresses the *Renilla* luciferase gene, as a transfection control. Transfected cells were cultured for 48 h and were lysed in 1 × passive lysis buffer, according to the manufacturer's instructions. Lysates were stored at −80°C until assay. Luciferase assays were measured on a Wallac Victor2 plate luminometer (Perkin-Elmer). Promoter strength for each construct was expressed as the ratio of firefly to *Renilla* luciferase activity.

dbSNP

We have deposited all our polymorphism data into dbSNP (National Center for Biotechnology Information) under assay ID numbers ss5607025–ss5607061. Complete genotype data for these SNPs are available for all Coriell samples. Five polymorphisms reported here were already in dbSNP (α9 A205G [rs251352], α11 T150C [rs192231], α13 C177A [rs59479], β12 C107T [rs2910326], and γB7 C88A [rs2240701]). We have added our genotyping data to these records.

Results

Polymorphism Discovery and Population Distribution

To characterize the degree of protocadherin promoter variation, we sequenced all 13 α protocadherin promoters in genomic DNA samples derived from 24 European American children affected with autism. To compare observed and predicted haplotypes, we also sequenced the promoters in genomic DNA from the parents and siblings of these individuals. We discovered a total of 11 common SNPs in the α1, α3, α4, α5, α7, α9, α11, and α13 promoters (fig. 1*A*). We later determined that three of these SNPs (α9 A205G, α11 T150C, and α13 C177A) had previously been discovered and deposited in dbSNP (see the "Material and Methods" section). Each allele for each promoter SNP from α1 to α7 is present at equal frequency in this sample. These SNPs also appear to be in complete LD and to form two haplotypes of equal frequency across 48 kb, on the basis of direct observation of haplotypes in the families of these individuals and on the basis of haplotypes that we predicted using the expectation-maximization algorithm (see the "Material and Methods" section).

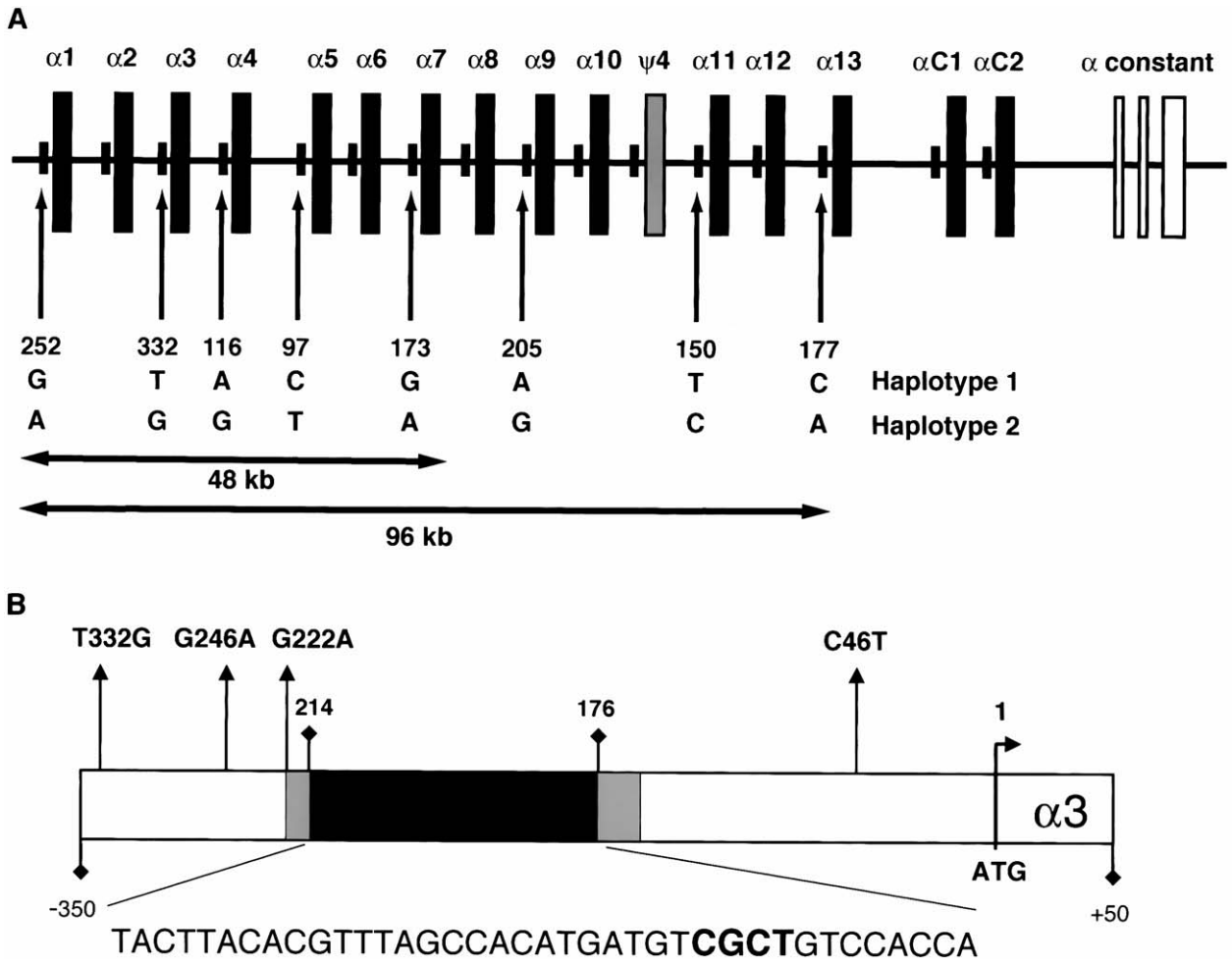

Figure 1 Structure and organization of the human protocadherin α genes. *A,* The human protocadherin α cluster. Each black box represents a single short-form exon; smaller black boxes represent promoter sequences (not to scale). Positions of common promoter polymorphisms and the two major haplotypes that they comprise are shown. SNPs are numbered relative to the translation start site of each gene. Exons of the α constant region are positioned at the 3′ end of the α cluster as indicated (*white boxes*). αC1 and αC2 are more similar to each other than to the other members of the α cluster. Exon sizes and distributions are approximately to scale (except for constant-region exons). *B,* Protocadherin promoter structure. Each exon has a core promoter element (*black*) upstream of the translation start site. This core promoter is highly conserved among paralogs and orthologs. Nucleotides in boldface are almost completely conserved in all mouse and human α, β, and γ protocadherin-cluster promoters. Gray shading indicates regions conserved between specific mouse and human orthologous promoters.

To determine whether this apparently binary haplotype distribution is a general feature of the α cluster in all populations, we determined genotypes for each polymorphism that we discovered by the sequencing of promoters from additional samples, including 27 Europeans, 28 African Americans, 25 East Asians, and 7 sub-Saharan (non-Pygmy) Africans. All common polymorphic sites and alleles present in our initial samples are also present in the European, East Asian, African, and African American individuals (table 1; African data not shown). The α4 A116G and α13 C177A SNPs are present in Africans, but several of these samples failed in the resequencing reactions. We therefore excluded these samples from our haplotype estimates. Despite the issue of admixture, the African American Coriell samples can be considered as equivalent to the African samples (Vigilant et al. 1991; Cavalli-Sforza et al. 1994).

We identified 32 total haplotypes in the α cluster, 19 of which are exclusive to African Americans. Of these exclusive haplotypes, 14 include the α3 C212G SNP, which is polymorphic only in Africans and African Americans. Five haplotypes (1–5; see table 1) account for 71% of all chromosomes in our sample. Europeans have 9 α-cluster haplotypes, 0 of which are exclusive to this population, whereas East Asians have 10 α-cluster haplotypes, 4 of which are exclusive. Low-frequency haplotypes

Table 1

Estimated Common α-Cluster Haplotype Frequencies

	Frequency in[b]				Allele at											
Haplotype[a]	All (80)	E (27)	EA (25)	AA (28)	α1 G252A	α3 T332G	α3 A246G	α3 A222G	α3 C212G[c]	α3 C46T	α4 A116G	α5 C97T	α7 G173A	α9 A205G	α11 T150C	α13 C177A
Human:																
1	27.7	44.4	24.0	14.5	G	T	A	A	C	C	A	C	G	A	T	C
2	24.5	20.4	46.0	9.1	A	G	G	G	C	T	G	T	A	G	C	A
3	8.2	16.7	8.0	0	A	G	G	G	C	T	G	T	A	A	T	C
4	7.5	3.7	0	18.2	A	G	G	G	C	T	G	T	A	G	C	C
5	3.1	7.4	2.0	0	A	G	G	G	C	T	G	T	A	Δ	C	C
6	2.5	1.9	0	5.5	G	T	A	A	C	C	A	C	G	A	C	C
7	2.5	0	8.0	0	G	T	A	A	C	C	A	C	G	A	C	A
8	2.5	0	0	7.3	G	G	A	G	G	T	G	T	A	G	C	C
9	1.9	1.9	4.0	0	G	T	A	A	C	C	A	C	G	G	C	A
10	1.9	1.9	0	3.6	A	G	G	G	C	T	G	T	A	G	T	C
11	1.9	0	0	5.5	G	T	A	A	C	C	A	C	G	G	C	C
12	1.3	1.9	2.0	0	A	G	G	G	C	T	G	T	G	A	T	C
13	1.3	0	0	3.6	G	G	G	G	C	T	G	T	A	G	C	C
14	1.3	0	0	3.6	G	G	A	G	G	T	G	C	A	G	C	A
15	1.3	0	0	3.6	G	G	A	G	G	T	G	T	A	G	C	A
17	.6	0	2.0	0	A	G	G	G	C	T	G	T	G	Δ	C	C
24	.6	0	0	1.8	G	G	A	G	G	T	G	T	A	Δ	C	A
31	.6	0	0	1.8	A	G	A	G	C	T	G	T	A	Δ	T	C
Chimpanzee					G	T	A	A	C	T	ns	T	A	G	C	A

	Allele Frequency at[d]											
Sample	α1 G252A	α3 T332G	α3 A246G	α3 A222G	α3 C212G	α3 C46T	α4 A116G	α5 C97T	α7 G173A	α9 A205G	α11 T150C	α13 C177A
E:												
Allele 1	.48	.48	.48	.48	1.00	.48	.48	.48	.50	.65	.65	.78
Allele 2	.52	.52	.52	.52	.00	.52	.52	.52	.50	.28	.35	.22
Δ										.07		
EA:												
Allele 1	.38	.38	.38	.38	1.00	.38	.38	.38	.42	.46	.38	.42
Allele 2	.62	.62	.62	.62	.00	.62	.62	.62	.58	.50	.62	.58
Δ										.04		
AA:												
Allele 1	.58	.33	.56	.33	.77	.31	.31	.35	.31	.33	.25	.75
Allele 2	.42	.67	.44	.67	.24	.69	.69	.66	.69	.64	.75	.25
Δ										.036		

Note.—E = Europeans; EA = East Asians; AA = African Americans; ns = not sequenced.

[a] Only haplotypes with worldwide frequency >1% are shown, in addition to rare haplotypes carrying the 16.7-kb deletion allele (designated by "Δ"; see fig. 4).

[b] The number of individuals genotyped from each population is indicated in parentheses.

[c] The G allele (underlined) of the α3 C212G SNP is exclusive to Africans and African Americans.

[d] In each SNP's name, allele 1 is given first.

in all samples are the product of either multiple ancestral recombination events or rare SNPs. The two most frequent haplotypes, 1 and 2, are also the major haplotypes in Europeans and East Asians, and the most frequent haplotype in African Americans (haplotype 4) is identical to haplotype 2 except for a recombination event between α11 T150C and α13 C177A.

Promoter polymorphisms from α1 through α7 form two haplotypes of estimated equal frequency in unaffected Europeans (table 1). The estimated haplotype frequencies in East Asians are slightly different, but the major haplotypes are identical. Ancestral recombination events involving the α9 and α11 promoter SNPs are evident in both populations, indicating an erosion of LD 3′ of the α7 promoter. The greater haplotype diversity in African Americans reflects the accumulation of mutations, as well as accumulated recombination events, in this population (Underhill et al. 2001). The genotype frequencies for all polymorphisms meet Hardy-Weinberg expectations in all populations (data not shown).

We sequenced the α3, α5, and α9 promoters in 16 genomic DNA samples derived from Mbuti and Biaka Pygmies. All alleles present in these promoters in other populations are present in this sample, indicating that these SNPs are extremely old (Mountain and Cavalli-Sforza 1994; Underhill et al. 2001). We also sequenced the orthologous α1, α3, α5, α7, α9, α11, and α13 promoters in genomic DNA derived from 10 chimpanzees, to determine the ancestral state at each of the polymorphic loci that we discovered (table 1). No polymorphism was evident at these sites in these chimpanzees. The α-cluster promoter sequences in chimpanzees agree with one of the two alleles at all SNP sites in humans. The two major haplotypes in the α cluster therefore emerged after the human-chimpanzee divergence but may predate the divergence of Pygmies from other human populations.

In Europeans and East Asians, the α promoter SNPs reveal a block of apparently complete LD, spanning 48 kb from the α1 promoter through the α7 promoter (tables 2*A* and 2*B*). Across this region, $D' = 1$ (calculated relative to α1 G252A), with $P < .0001$ by likelihood-ratio χ^2 test. The α9 and α11 promoter SNPs show reduced but still significant LD with the α1–α7 block in both populations. African Americans also show considerable LD in the α1–α7 region ($D' > 0.9$; $P < .0002$), but considerably less than that seen in European and East Asian populations (table 2*C*). The LD seen in these populations is also apparent in African Americans, but it has been broken into two smaller, discontinuous regions, by recombination or mutation in the α3 promoter. These common α promoter polymorphisms define an ancient, apparently binary haplotype structure, spanning the α cluster, that is slowly being degraded by recombination and new mutation. We excluded SNPs for which the minor-allele frequencies were too low to give meaningful results (Lewontin 1995). To determine whether the α-cluster LD is a feature of the entire protocadherin cluster, we also sequenced most of the remaining β and γ promoters in our discovery set of 24 autistic European Americans. We discovered polymorphisms in several of these promoters, but we observed no significant LD (data not shown) (for dbSNP accession numbers, see the "Material and Methods" section). The existence of two major haplotypes and associated extensive LD seems to be restricted to the α protocadherins.

The two major α-cluster haplotypes, 1 and 2, differ at every common SNP and are present at nearly equal frequency worldwide. This suggests that selection may actively maintain two common allele states at one or more positions in the α cluster. Alternatively, the haplotype distribution in the α cluster could be the signature of an ancient selective event. We therefore applied Tajima's *D* test for selection to all α promoter polymorphisms in the European, East Asian, and African American samples (Tajima 1989). A significant positive value for Tajima's *D* test indicates an excess of intermediate-frequency variants, as compared with expected frequencies under neutrality, and constitutes evidence of balancing selection or population subdivision (Tajima 1989; Kreitman 2000). We obtained significant positive values for Tajima's *D* test in Europeans and Asians, both for promoter SNPs from α1 through α7, which are in complete LD, and for all α promoter SNPs (table 3). An independent set of European samples (34 parents of children affected with autism) also gave a similar value of *D* for the α1–α7 promoter SNPs. We obtained a positive *D* value for African Americans, but this result was not statistically significant. We also applied another test statistic, F_s (Fu 1997). A positive value for F_s indicates a deficiency of rare alleles, and both European and East Asian populations gave significant positive F_s values for the α1–α7 promoter SNPs (table 3). Recombination, however, tends to decrease the value of F_s (Fu 1997). We found that the inclusion of promoter polymorphisms 3′ of α7, which have been involved in ancestral recombination events, decreased the F_s value for the α promoter SNPs in all populations. There is clearly an excess of polymorphic sites with two high-frequency alleles across the α cluster, apparently centered around the α3 promoter, which has four polymorphisms (five in African Americans).

To illustrate the relationships among these 32 α-cluster haplotypes, we constructed an MS network (fig. 2). The effect of balancing selection is reflected in such a network as the presence of two major lineages separated by many mutational or recombinational steps (Marjoram and Donnelly 1994). Conversely, population expansions would yield starlike genealogies owing to the retention of new lineages (Donnelly 1996). There are clearly two major lineages of α-cluster haplotypes in our network. Haplotypes 1 and 2 are at either end of the

Table 2

LD in the Protocadherin α Cluster

A. Europeans

	Pairwise D' between									
SNP	α3 T332G	α3 A246G	α3 A222G	α3 C46T	α4 A116G	α5 C97T	α7 G173A	α9 A205G	α11 T150C	α13 C177A
α1 G252A	1**	1**	1**	1**	1**	1**	1**	.876671*	.729219**	ns
α3 T332G		1**	1**	1**	1**	1**	1**	.876671*	.729219**	ns
α3 A246G			1**	1**	1**	1**	1**	.876671*	.729219**	ns
α3 A222G				1**	1**	1**	1**	.876671*	.729219**	ns
α3 C46T					1**	1**	1**	.876671*	.729219**	ns
α4 A116G						1**	1**	.876671*	.729219**	ns
α5 C97T							1**	.876671*	.729219**	ns
α7 G173A								.879845*	.738607**	ns
α9 A205G									.916828**	1**
α11 T150C										1**

B. East Asians

	Pairwise D' between									
SNP	α3 T332G	α3 A246G	α3 A222G	α3 C46T	α4 A116G	α5 C97T	α7 G173A	α9 A205G	α11 T150C	α13 C177A
α1 G252A	1**	1**	1**	1**	1**	1**	1**	.771891*	.515592*	.459908*
α3 T332G		1**	1**	1**	1**	1**	1**	.771891*	.515592*	.459908*
α3 A246G			1**	1**	1**	1**	1**	.771891*	.515592*	.459908*
α3 A222G				1**	1**	1**	1**	.771891*	.515592*	.459908*
α3 C46T					1**	1**	1**	.771891*	.515592*	.459908*
α4 A116G						1**	1**	.771891*	.515592*	.459908*
α5 C97T							1**	.771891*	.515592*	.459908*
α7 G173A								.725915	.587768*	.544496**
α9 A205G									1**	.918871**
α11 T150C										1**

C. African Americans

	Pairwise D' between										
SNP	α3 T332G	α3 A246G	α3 A222G	α3 C212G	α3 C46T	α4 A116G	α5 C97T	α7 G173A	α9 A205G	α11 T150C	α13 C177A
α1 G252A	1**	.76256**	1**	1*	1**	1**	1**	1**	ns	ns	ns
α3 T332G		.718221*	1**	ns	1**	1**	.913026**	1**	ns	ns	ns
α3 A246G			.718221*	ns	.696466*	.718221*	.73707*	.696466*	ns	ns	ns
α3 A222G					1**	1**	.913026**	1**	ns	ns	ns
α3 C212G					ns	ns	ns	ns	ns	ns	ns
α3 C46T						1**	1**	1**	ns	ns	ns
α4 A116G							.913026**	1**	ns	ns	ns
α5 C97T								1**	ns	ns	ns
α7 G173A									ns	ns	ns
α9 A205G										ns	ns
α11 T150C											ns

NOTE.—Pairwise D' values and χ^2 likelihood-ratio P values were calculated using 2LD (see the "Material and Methods" section). ns = Not significant.

* $P < .05$.

** $P < .005$.

Table 3

Diversity Statistics and Results of Tajima's *D* and Fu's F_s Tests for the α1, α3, α4, α5, α7, α9, α11, and α13 Promoter SNPs

Population and Region	No. of Chromosomes	No. of Haplotypes	S[a]	θ_s[a]	π_n[b]	θ_π	D	P	F_s	P
European:										
α1–α7	54	4	9	1.98 ± .83	.211 ± .118	4.1475	3.03578	.001	8.04180	.013
α1–α13	46	9	12	2.73 ± 1.07	.172 ± .094	5.4560	2.99635	.001	2.85951	.157
Asian:										
α1–α7	50	4	9	2.01 ± .85	.199 ± .111	3.9029	2.63786	.0046	7.27703	.018
α1–α13	46	8	13	2.96 ± 1.14	.173 ± .094	5.4754	2.58899	.0053	3.83598	.087
African American:										
α1–α7	56	11	14	3.05 ± 1.14	.218 ± .121	4.2851	1.20914	.12243	.74536	.351
α1–α13	52	20	17	3.76 ± 1.35	.177 ± .096	5.6056	1.52058	.07297	−3.88951	.111
AT parents:[c]										
α1–α7	68	3	8	1.67 ± .71	.207 ± .116	4.0593	3.6919	.00008	10.93173	.0048

[a] For details, see the "Material and Methods" section.
[b] Nucleotide diversity (per-site heterozygosity) as calculated in Arlequin. SDs are shown.
[c] Parents of children affected with autism.

network and are separated by a long branch of 10 steps, owing to their allelic differences at every common α promoter SNP. Both haplotype groups have haplotypes present in all three populations, as well as haplotypes exclusive to one population. There are population-specific differences between the groups. Many of the low-frequency haplotypes exclusive to African Americans are in group 2, except for haplotypes 19, 20, and 32, which are in group 1. The radiations within each group may be the signature of population growth, imposed concurrent with or subsequent to the effects of balancing selection.

α8–α10Δ Deletion-Allele Discovery, Structure, and Population Distribution

While genotyping the members of families with affected children, for each α promoter polymorphism, we identified three families that showed violation of Mendelian inheritance for the α9 A205G promoter SNP. Children in these families lacked the G allele when one parent appeared to be homozygous for that allele. One explanation for such a result is that the parent or child is actually hemizygous, carrying an α9 promoter deletion on one of his or her homologs. To test this hypothesis, we developed a real-time–PCR assay to determine genomic copy number at the α9 locus (see the "Material and Methods" section). In this assay, children with apparent α9 promoter deletions had twofold less α9 target copy number than did unaffected control individuals. We mapped both endpoints of the deletion to within 900 bp, using this method. We then amplified the junction fragment by PCR and determined the precise structure of the junction by sequencing the PCR product. We found a 16.7-kb deletion, which we designated as "α8–α10Δ," extending from position 1231 of the α8 coding region to 764 bases 3′ of the α10 stop codon (figs. 3*A* and 3*B*). Three nucleotides (CCA) of unknown origin are at the deletion junction, but, otherwise, the flanking sequences are intact (fig. 3*B*). There is an in-frame stop codon in the α10-α11 intergenic region directly 3′ of the deletion junction, so translation of the shortened α8 transcript would produce a truncated α8 protein that lacks transmembrane or cytoplasmic domains. Although transcription of a truncated α8 mRNA is possible, the absence of a polyadenylation signal makes it unlikely for stable α8 transcript to be produced. Therefore, this deletion effectively eliminates expression of α8–α10.

To determine whether the α8–α10Δ allele is associated with autism, we genotyped 542 individuals from 143 multiplex families with autism, by PCR. We also genotyped 48 European samples from unrelated, unaffected individuals. We found no significant difference in the frequency of the α8–α10Δ allele between affected and unaffected individuals (table 4). We also saw no significant allele-frequency difference between mothers and fathers of affected children or between parents of affected children and unaffected individuals. We also failed to see any significant ($P = .08$) transmission disequilibrium for the deletion allele (table 5) (Spielman et al. 1993). Our failure to detect association between the α8–α10Δ allele and autism does not exclude it as a candidate mutation for other neurological disorders.

It seems unlikely that the deletion of three well-conserved genes is a neutral mutation. If the α8–α10Δ allele were restricted to Europeans, then it could have arisen recently. In an attempt to understand when the α8–α10Δ allele arose, we determined the α8–α10Δ allele frequencies in the major human populations. The α8–α10Δ allele shows worldwide distribution, present in European, East Asian, African, and African American populations (table 4). This widespread population distribution suggests that

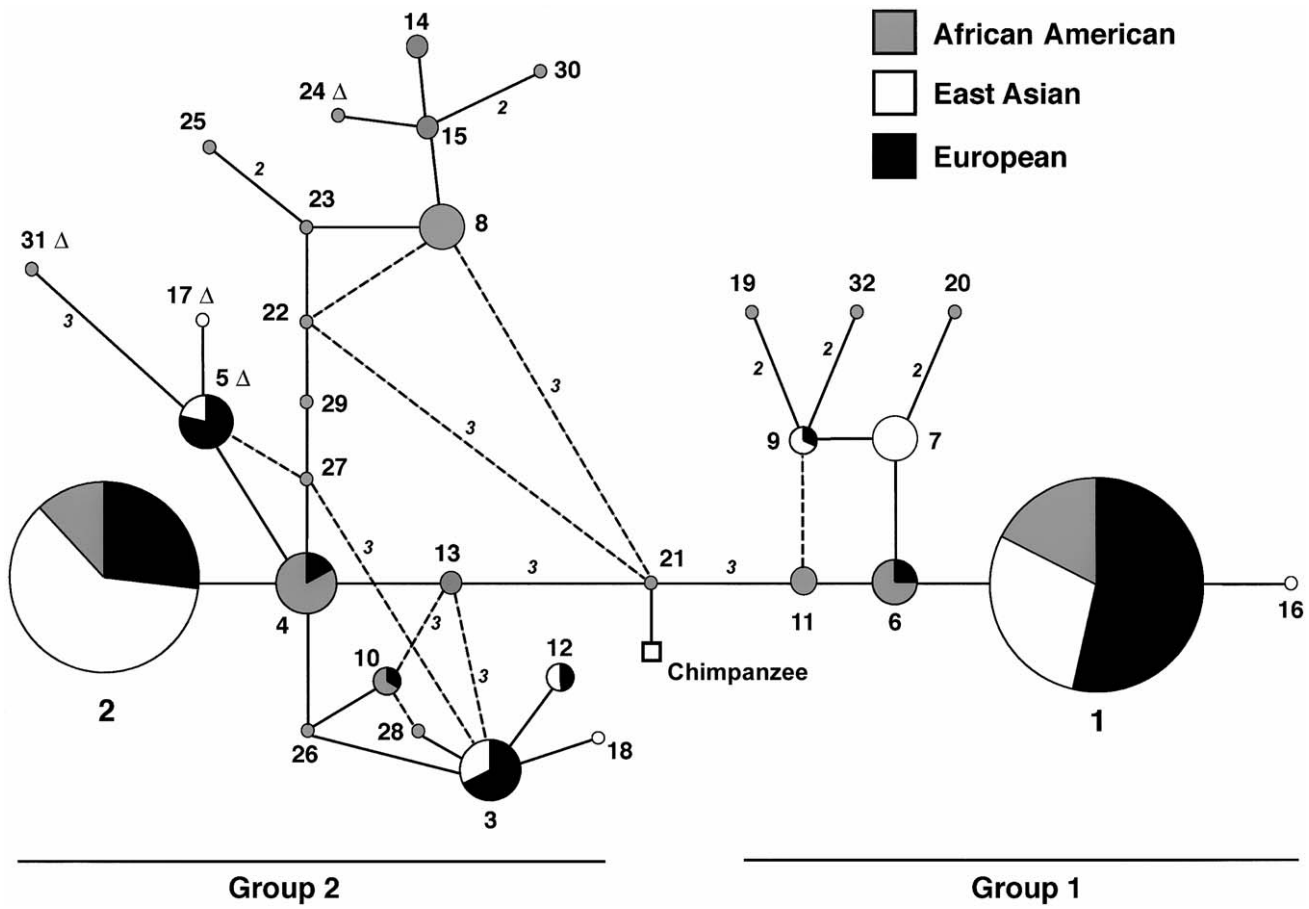

Figure 2 MS network for predicted α-cluster haplotypes. The size of each node is proportional to the haplotype frequency in the sample. Colors within each node represent the relative distribution of each haplotype among the three populations in the present study. Branch lengths represent one nucleotide substitution, except as otherwise labeled. Dashed lines represent alternative relationships among haplotypes. For clarity, in some cases, branches are not drawn to scale.

it is both ancient and neutral. The distribution of deletion genotypes, including the distribution of α8–α10Δ homozygotes, meets Hardy-Weinberg expectation in all populations (table 4). Our results indicate that the α8–α10Δ allele arose in the ancestral African population, before the migrations out of Africa that founded the European and East Asian populations. Current estimates for the isolation of Europeans from their African ancestors range from 40,000 to 60,000 years ago (Harpending and Rogers 2000; Thomson et al. 2000; Underhill et al. 2000). Given its worldwide distribution, the α8–α10Δ allele is at least that old.

We attempted to estimate an upper limit for the age of the α8–α10Δ allele by genotyping a mixed sample of 23 Mbuti and Biaka Pygmies. These are two of the most ancient human populations, and polymorphisms that are found both in Mbuti and Biaka Pygmies and in all other human populations are very old (Cavalli-Sforza et al. 1994). We did not detect the deletion allele in any member of this sample; the deletion either is not present or is present at a low frequency in Pygmies. In the MS network of all α-cluster haplotypes (fig. 2), all haplotypes bearing an α8–α10Δ allele are assigned to group 2, indicating a common ancestry with haplotype 2, one of the two major α-cluster haplotypes that are present in all populations. In Europeans, the deletion allele shows nearly complete LD with the flanking α7 and α11 promoter SNPs ($D' > 0.92$; $P < .0001$). On the basis of the MS network, the α8–α10Δ allele entered the European population on haplotype 5. This haplotype is two nucleotide substitutions removed from haplotype 2. We have determined full α-cluster haplotypes for a small number of non-European individuals carrying the α8–α10Δ allele, and the rarity of these haplotypes makes their placement in the MS network uncertain. Haplotypes 17 and 31, however, appear to be derivatives of haplotype 5 (fig. 2), although the MS algorithm places haplotype 24 on a separate lineage.

Effect of Neuronal Differentiation and Polymorphism on Promoter Strength

To determine whether α promoter SNPs affect promoter function, we cloned the variants for the α3 and α9 promoters into the pGL3 luciferase-expression vector (see the "Material and Methods" section) and assayed for differences in promoter strength as measured by luciferase activity. We transfected these constructs into mouse P19 embryonic carcinoma cells, which can be differentiated into neuronlike cells by use of retinoic acid (Bain et al. 1998) (see the "Material and Methods" section). The relative strengths of both human α3 promoter variants increased ~28-fold between undifferentiated and differentiated P19 cells ($P = 1.45e-6$ for α3 TAAC, and $P = 1.24e-6$ for α3 GGGT, both by single-factor ANOVA) (fig. 4). Human α9 promoter strength also increased substantially on differentiation, although there was a significant difference in this increase between the α9 promoter variants: the strength of the α9 G promoter variant increased 21.9-fold on differentiation, whereas the strength of the α9 A promoter variant increased 15.7-fold ($P = .0005$). The strength of the α9 G promoter variant was also less than that of the α3 GGGT promoter variant in differentiated ($P = .001$) and undifferentiated ($P = .003$) P19 cells, indicating that a functional difference exists between these paralogous promoters. These results indicate that significant determinants for cell-type–specific transcription regulation of these genes are contained within the promoter.

Table 4

α8–α10Δ Allele and Genotype Frequencies in Selected Populations

Population	No. of Individuals				Frequency		
	α9/α9	α9/Δ	Δ/Δ	Total	α9	Δ	P[a]
Affected individuals[b]	111	30	2	143	.881	.119	.987
Europeans	39	8	1	48	.890	.110	.459
African Americans	46	4	0	50	.96	.04	.768
East Asians	28	2	0	30	.97	.033	.850
Africans[c]	6	3	0	9	.83	.17	…
Pygmies	23	0	0	23	1.00	0	…

[a] Meeting Hardy-Weinberg expectations.
[b] Unrelated European American children with autism.
[c] The deletion-allele frequency in this limited number of African samples may not reflect the true allele frequency in Africans.

The α9 promoter SNP falls within the α9 promoter element, whereas the α3 promoter is highly polymorphic (fig. 1*B* and table 1). In our assays, the α3 GGGT promoter variant was 1.2-fold stronger than the α3 TAAC variant in differentiated P19 cells, but this result was not significant ($P = .13$ by single-factor ANOVA). The α9 G promoter variant, however, is ~1.6-fold stronger than the α9 A variant in differentiated P19 cells ($P = .02$). This result, combined with the different responses of

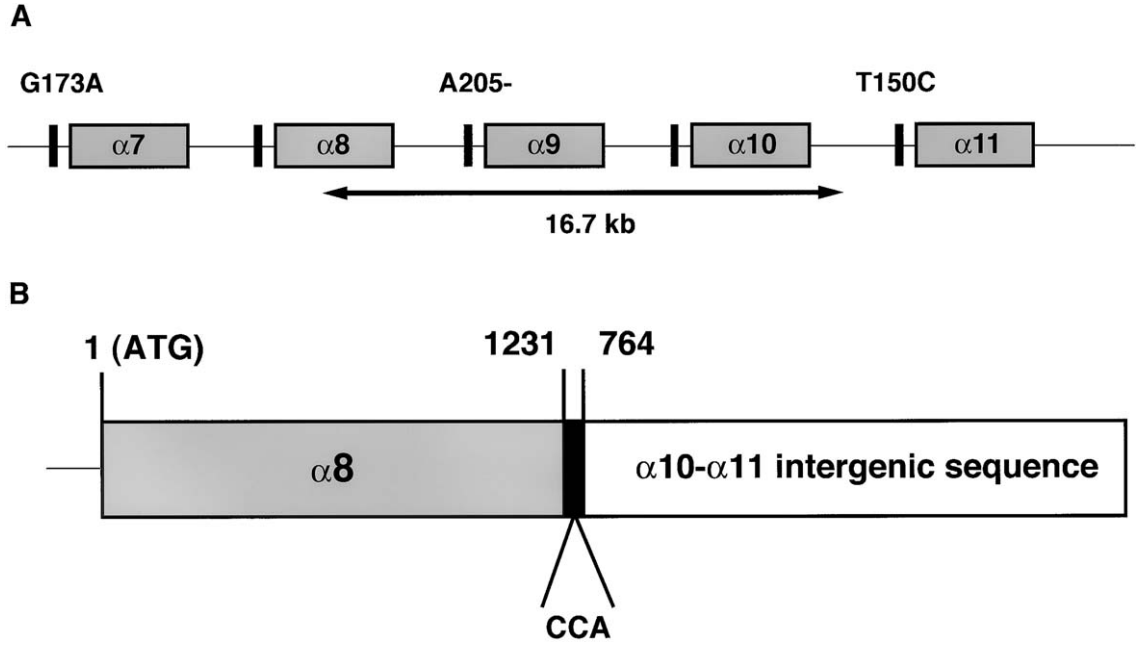

Figure 3 The α8–α10 deletion. *A*, The deletion interval. The region from α7 to α11 is shown, with known promoter polymorphisms positioned as indicated. Intergenic regions are not shown to scale. *B*, Structure of the α8–α10 deletion junction. Three nucleotides of unknown origin link position 1231 of α8 to a site 764 bases 3′ of the α10 stop codon.

Table 5

Transmission Disequilibrium Test Results for the α8–α10Δ Allele in Children with Autism

Test	No. of Transmissions			No. of Families	χ^2	P
	Δ	α9	Total			
Paternal transmission	16	12	28	14	.57	.45
Maternal transmission	24	14	38	19	2.63	.10
Overall	40	26	66	33	2.97	.08

these variants to neuronal induction (fig. 4), suggests that these two alleles are likely to result in different levels of α9 transcript in neurons.

Discussion

We have discovered several α protocadherin promoter SNPs that define an extensive region of LD in the α cluster. This LD is due to reduced recombination, which is reflected in the limited haplotype diversity evident even in genetically diverse populations. This is consistent with the hypothesis that recombination should be infrequent in tandem gene arrays. However, the ancestral recombination events apparent in African populations within the region of extensive LD seen in Europeans and East Asians appear to have occurred without gene loss. Our results are also consistent with reports of extensive LD in other regions of the genome (Daly et al. 2001; Patil et al. 2001; Reich et al. 2001; Stephens et al. 2001; Gabriel et al. 2002). What distinguishes the α protocadherin cluster from these other regions is the abundance of intermediate-frequency variants and the unusual α-cluster haplotype structure.

Europeans and East Asians have two major α promoter haplotypes. Since European and East Asian populations were founded recently (40,000–60,000 years ago) by populations that left Africa and underwent bottleneck events at some point in their history, they are less genetically diverse than older African populations (Bowcock et al. 1994; Cavalli-Sforza et al. 1994; Tishkoff et al. 1996; Hammer et al. 1997; Jorde et al. 1997; Harpending and Rogers 2000; Ingman et al. 2000). The two major haplotypes in the ancestral African population are the haplotypes most likely to be represented in the descendant European and East Asian founders. African populations are expected to have greater polymorphism and haplotype diversity, as well as reduced LD, all of which have been observed in multiple studies (Tishkoff et al. 1996, 1998, 2000; Kidd et al. 1998; Reich et al. 2001; Gabriel et al. 2002). African Americans in the present study have reduced α-cluster LD, as well as a larger number of low-frequency polymorphisms and haplotypes.

The two major α-cluster haplotypes, however, are present at high frequency in African Americans. This result, together with the MS network of all α promoter haplotypes, indicates that, during early human history, two haplotype groups arose across the entire α cluster. These groups were clearly established before the migrations out of Africa that founded the European and East Asian populations and have persisted despite subsequent population expansion. These groups also arose after the human-chimpanzee divergence, since chimpanzees are monoallelic at every α-promoter-SNP site. Mbuti and Biaka Pygmy populations carry both alleles for the common α3, α5, and α7 promoter SNPs, suggesting that the two groups of α promoter haplotypes may predate the divergence between Pygmies and other populations.

The positive values for Tajima's D test that we obtained for the α promoter SNPs in both Europeans and East Asians suggest that balancing selection is operating to maintain both alleles of one or more of these polymorphisms—or the alleles of polymorphisms closely linked to them—at high frequencies. However, given the increased α promoter haplotype diversity in African Americans and the nonsignificant D value in this population, it seems more likely that the nearly equal allele frequencies of many α promoter SNPs are the signature of a past episode of balancing selection, occurring early in human history. This selective event could have generated the original, binary α-cluster haplotype structure, which recombination has since eroded. The rate of this erosion would be proportional to the age of the population.

The many regions of limited haplotype diversity that are seen throughout the genome may be the signature of balancing selection (Harpending and Rogers 2000). However, a recent survey of polymorphism in 313 human genes found that 281 of these genes had a negative

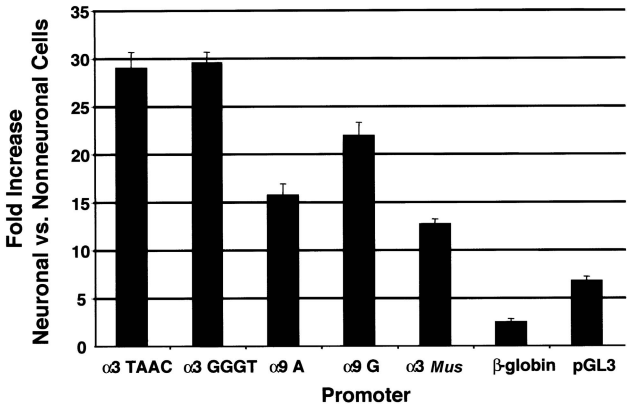

Figure 4 Fold increase in α3 and α9 promoter strength on neuronal differentiation of mouse P19 cells by use of retinoic acid. Results shown are the average of three independent experiments.

D value, irrespective of the amount of haplotype diversity in the region (Stephens et al. 2001). Balancing selection may therefore operate only at a limited number of loci. A classic example of balancing selection is found at the major histocompatibility complex (MHC) loci (Hughes and Nei 1988; Satta et al. 1994). The selective advantage of heterozygotes over homozygotes is thought to maintain high levels of polymorphism in functional MHC genes. This advantage is believed to derive from the ability of heterozygotes to present a greater range of antigens to T cells, as compared with homozygotes. Evidence of balancing selection has also recently been found in the 5′ *cis*-regulatory region *CCR5,* the principal coreceptor for HIV-1 (Bamshad et al. 2002). The structure and population distribution of the *CCR5*-regulatory-region haplotypes are strikingly similar to those of the protocadherin α promoter haplotypes reported here. The *CCR5*-regulatory-region haplotypes form two major groups with an estimated divergence time of ~2.1 million years. Tajima's *D* value was significant and positive in this region in non-African populations, which also showed reduced *CCR5*-regulatory-region haplotype diversity, as compared with African populations.

Balancing selection may operate in *CCR5* to maintain regulatory sequence diversity in defense against pathogens (Bamshad et al. 2002). A similar mechanism may be operating in the protocadherin α cluster. If protocadherin-cluster genes provide a reservoir of genetic complexity for brain development, then it could be advantageous to maintain allelic diversity in protocadherin regulatory and coding regions. This could confer diversity in expression patterns of protocadherin genes in neurons and in interactions of protocadherin proteins at synapses. However, we did not observe an excess of heterozygotes for any α promoter SNP in any population, suggesting that balancing selection either is too weak to significantly affect genotype frequencies or is no longer operating. We have not determined the ratio of nonsynonymous to synonymous coding changes in the protocadherin α genes. The α9 promoter variants show a clear difference both in their overall strength and in their responses to neuronal differentiation. The promoters confer much of the cell-type specificity of these genes (fig. 4). At some point in human evolution, it may have been advantageous to have two alleles of one or more α promoters with different strengths. Alleles at linked SNPs would reach high frequency owing to hitchhiking, and the low recombination rate would retain the observed haplotype structure (Nachman 2001).

The α8–α10Δ allele is not the product of an unequal crossover between α8 and α10 coding sequences. The deletion junction does not involve any region of α8–α10 sequence similarity, and the additional 3 bp at the junction site could be a transposition remnant. Because the α8–α10Δ allele is in Hardy-Weinberg equilibrium and, to our knowledge, exists in all populations except Mbuti and Biaka Pygmy, it is probably neutral. The α8–α10Δ allele arose on one of the two major α-cluster haplotypes at some point before the divergence of African, European, and East Asian populations. Its 11% frequency in Europeans may reflect a high initial frequency in the European founder populations or may be due to selection operating on the surrounding haplotype.

The deletion of multiple conserved protocadherin genes suggests that some level of functional redundancy has evolved over the course of protocadherin-cluster evolution in humans. The hypothesized role that these proteins play in the generation of synaptic complexity is likely to depend on the complexity of their regulation. The exact number of genes may not be crucial, so long as the overall complexity generated by the cluster is maintained. However, it is still possible that the α8–α10 deletion has a deleterious effect in some genetic or environmental context. The loss of three protocadherins could subtly perturb synaptogenesis, having relatively minor effects that might not appear until late in life. The α8–α10Δ allele is therefore a candidate mutation for neuropathologies that have a limited effect on reproductive fitness, such as depression or bipolar disorder. Until the phenotypic effect of gene loss in the α8–α10 region is thoroughly investigated, we are left with the signature of balancing selection provided by the α promoter SNPs and a mystery as to the biological significance of the α8–α10 deletion.

Acknowledgments

We are grateful to members of the Myers Lab for discussions and support. This work was supported by the Stanford Genome Training Program (National Institutes of Health training grant 5 T32 HG00044 [to J.P.N.]).

Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

Arlequin, http://lgb.unige.ch/arlequin/

dbSNP Home Page, http://www.ncbi.nlm.nih.gov/SNP/ (for ss5607025–ss5607061)

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for protocadherin α, β, and γ subclusters [accession numbers NG_000016, NG_000017, and NG_000012, respectively] and BAC DNA [accession number AC020968])

KCL, Institute of Psychiatry, Section of Genetic Epidemiology and Biostatistics, http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.stm (for 2LD)

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for the human protocadherin α subcluster [MIM 604966])

Primer3, http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
UCSC Genome Bioinformatics, http://genome.ucsc.edu/ (for the human protocadherin cluster)

References

Angst BD, Marcozzi C, Magee AI (2001) The cadherin superfamily: diversity in form and function. J Cell Sci 114:629–641

Bain G, Yao M, Huettner JE, Finley MFA, Gottlieb DI (1998) Neuronlike cells derived in culture from P19 embryonal carcinoma and embryonic stem cells. In: Banker G, Goslin K (eds) Culturing nerve cells, 2nd ed. MIT Press, Cambridge, pp 189–212

Bamshad MJ, Mummidi S, Gonzalez E, Ahuja S, Dunn DM, Watkins WS, Wooding S, Stone AC, Jorde LB, Weiss RB, Ahuja SK (2002) A strong signature of balancing selection in the 5′ *cis*-regulatory region of *CCR5*. Proc Natl Acad Sci USA 99:10539–10544

Beasley A, Myers RM, Cox DR, Lazzeroni LC (1999) Statistical refinement of primer design parameters. In: Innis MA, Gelfand DH, Sninsky JJ (eds) PCR applications. Academic Press, New York, pp 55–72

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368: 455–457

Bruses JL (2000) Cadherin-mediated adhesion at the interneuronal synapse. Curr Opin Cell Biol 12:593–597

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322

Donnelly P (1996) Interpreting genetic variability: the effects of shared evolutionary history. Ciba Found Symp 197:25–40

Ewing B, Green P (1998) Base-calling of automated sequencer traces using *phred*. II. Error probabilities. Genome Res 8: 186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. Genome Res 8:175–185

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Fannon AM, Colman DR (1996) A model for central synaptic junctional complex formation based on the differential adhesive specificities of the cadherins. Neuron 17:423–434

Fu Y-X (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915–925

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296: 2225–2229

Goddard KAB, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216–234

Hammer MF, Spurdle AB, Karafet T, Bonner MR, Wood ET, Novelletto A, Malaspina P, Mitchell RJ, Horai S, Jenkins T, Zegura SL (1997) The geographic distribution of human Y chromosome variation. Genetics 145:787–805

Harpending H, Rogers A (2000) Genetic perspectives on human origins and differentiation. Annu Rev Genomics Hum Genet 1:361–385

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170

Ingman M, Kaessmann H, Pääbo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713

Johnson GCL, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RCJ, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SCL, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237

Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J, Harpending HC (1997) Microsatellite diversity and the demographic history of modern humans. Proc Natl Acad Sci USA 94:3100–3103

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC, Bonne-Tamir B, Lu R, Goldman D, Lee C, Nam YS, Grandy DK, Jenkins T, Kidd JR (1998) A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Hum Genet 103:211–227

Kohmura N, Senzaki K, Hamada S, Kai N, Yasuda R, Watanbe M, Ishii H, Yasuda M, Mishina M, Yagi T (1998) Diversity revealed by a novel family of cadherins expressed in neurons at a synaptic complex. Neuron 20:1137–1151

Kreitman M (2000) Methods to detect selection in populations with applications to the human. Annu Rev Genomics Hum Genet 1:539–559

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67

——— (1995) The detection of linkage disequilibrium in molecular sequence data. Genetics 140:377–388

Li J, Tabor HK, Nguyen L, Gleason C, Lotspeich LJ, Spiker D, Risch N, Myers RM (2002) Lack of association between HoxA1 and HoxB1 gene variants and autism in 110 multiplex families. Am J Med Genet 114:24–30

Marjoram P, Donnelly P (1994) Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics 136: 673–683

Moffatt MF, Traherne JA, Abecasis GR, Cookson WOCM (2000) Single nucleotide polymorphism and linkage disequilibrium within the TCR α/δ locus. Hum Mol Genet 9:1011–1019

Mountain J, Cavalli-Sforza LL (1994) Inference of human evo-

lution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc Natl Acad Sci USA 91:6515–6519

Nachman MW (2001) Single nucleotide polymorphisms and recombination rate in humans. Trends Genet 17:481–485

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BTN, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SPA, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199–204

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Risch N, Spiker D, Lotspeich L, Nouri N, Hinds D, Hallmayer J, Kalaydjieva L, et al (1999) A genomic screen of autism: evidence for a multilocus etiology. Am J Hum Genet 65:493–507

Satta Y, O'hUigin C, Takahata N, Klein J (1994) Intensity of natural selection at the major histocompatibility complex loci. Proc Natl Acad Sci USA 91:7184–7188

Schneider S, Roessli D, Excoffier L (2000) Arlequin, version 2.000: a software for population genetics data analysis. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Geneva

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Spiker D, Lotspeich L, Kraemer HC, Hallmayer J, McMahon W, Peterson PB, Wong DL, Dimiceli S, Ritvo E, Cavalli-Sforza LL, Ciaranello RD (1994) Genetics of autism; characteristics of affected and unaffected children from 37 multiplex families. Am J Med Genet 54:27–35

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, et al (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293: 489–492

Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA (2001) Sequence variation and linkage disequilibrium in the human T-cell receptor β (TCRB) locus. Am J Hum Genet 69:381–395

Suzuki ST (1996) Protocadherins and diversity of the cadherin superfamily. J Cell Sci 109:2609–2611

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

——— (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tang L, Hung CP, Schuman EM (1998) A role for the cadherin family of cell-adhesion molecules in hippocampal long-term potentiation. Neuron 20:1165–1175

Tasic B, Nabholz CE, Baldwin KK, Kim Y, Rueckert EH, Ribich SA, Cramer P, Wu Q, Axel R, Maniatis T (2002) Promoter choice determines splice site selection in protocadherin α and γ pre-mRNA splicing. Mol Cell 10:21–33

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci USA 97:7360–7365

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M, Pääbo S, Watson E, Risch N, Jenkins T, Kidd KK (1996) Global patterns of linkage disequilibrium at the *CD4* locus and modern human origins. Science 271:1380–1387

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origins of myotonic dystrophy mutations. Am J Hum Genet 62:1389–1402

Tishkoff SA, Pakstis AJ, Stoneking M, Kidd JR, Destro-Bisol G, Sanjantila A, Lu RB, Deinard AS, Sirugo G, Jenkins T, Kidd KK, Clark AG (2000) Short tandem-repeat polymorphism/*Alu* haplotype variation at the PLAT locus: implications for modern human origins. Am J Hum Genet 67:901–925

Uchida N, Honjo Y, Johnson KR, Wheelock MJ, Takeichi M (1996) The catenin/cadherin adhesion system is localized in synaptic junctions bordering transmitter release zones. J Cell Biol 135:767–779

Underhill PA, Passarino G, Lin AA, Shen P, Lahr MM, Foley RA, Oefner PJ, Cavalli-Sforza LL (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet 65:43–62

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi Q, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y-chromosome sequence variation and the history of human populations. Nat Genet 26:358–361

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507

Wang X, Su H, Bradley A (2002) Molecular mechanisms governing *Pcdh-*γ gene expression: evidence for a multiple promoter and *cis*-alternative splicing model. Genes Dev 16:1890–1905

Watterson G (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276

Wu Q, Maniatis T (1999) A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell 97:779–790

Wu Q, Zhang T, Cheng JF, Kim Y, Grimwood J, Schmutz J, Dickson M, Noonan JP, Zhang MQ, Myers RM, Maniatis T (2001) Comparative DNA sequence analysis of mouse and human protocadherin clusters. Genome Res 11:389–404

Zapata C, Carollo C, Rodriguez S (2001) Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci. Ann Hum Genet 65: 395–406