CHD7 Gene Polymorphisms Are Associated with Susceptibility to Idiopathic Scoliosis

Xiaochong Gao, Derek Gordon, Dongping Zhang, Richard Browne, Cynthia Helms, Joseph Gillum, Samuel Weber, Shonn Devroy, Saralove Swaney, Matthew Dobbs, Jose Morcuende, Val Sheffield, Michael Lovett, Anne Bowcock, John Herring, and Carol Wise

Idiopathic scoliosis (IS) is the most common spinal deformity in children, and its etiology is unknown. To refine the search for genes underlying IS susceptibility, we ascertained a new cohort of 52 families and conducted a follow-up study of genomewide scans that produced evidence of linkage and association with 8q12 loci (multipoint LOD 2.77; $P =$ .0028). Further fine mapping in the region revealed significant evidence of disease-associated haplotypes ($P < 1.0 \times 10^{-4}$) centering over exons 2–4 of the *CHD7* gene associated with the CHARGE (c̲oloboma of the eye, h̲eart defects, a̲tresia of the choanae, r̲etardation of growth and/or development, g̲enital and/or urinary abnormalities, and e̲ar abnormalities and deafness) syndrome of multiple developmental anomalies. Resequencing *CHD7* exons and conserved intronic sequence blocks excluded coding changes but revealed at least one potentially functional polymorphism that is overtransmitted ($P = .005$) to affected offspring and predicts disruption of a caudal-type (cdx) transcription-factor binding site. Our results identify the first gene associated with IS susceptibility and suggest etiological overlap between the rare, early-onset CHARGE syndrome and common, later-onset IS.

Idiopathic scoliosis (IS [MIM %181800]) is defined by the presence of lateral deformity of the spine, with otherwise normal vertebral bodies and without coexisting diagnoses. With a prevalence of 2%–3% in school-aged children, IS is the most common pediatric spinal deformity and poses a significant health burden in the pediatric population. Onset of the disease is typically during the period of rapid growth at adolescence, and two clear risk factors are remaining growth potential and female sex.[1] Inheritance of IS is generally complex, although some families with apparent Mendelian transmission have been described.[2] We previously ascertained multigenerational pedigrees and performed a genomewide scan in one, family IS9.[3] More recently, chromosomal breakpoint–mapping studies and genomewide scans in single families with IS and in collections of families with IS have been reported.[4–8] Although each of these studies has tentatively identified several chromosomal regions that may contribute to disease, all results taken together have not clearly converged on any single region, and IS susceptibility loci have remained elusive.

In the present study, we sought to refine the search for IS susceptibility loci and to identify contributing genes, by following up results of genomewide scans in a new set of multiplex families with IS. All participating research subjects were ascertained under a protocol approved by the University of Texas Southwestern Medical Center Institutional Review Board. We had previously ascertained family IS14, as described elsewhere.[3] We ascertained 52 additional families through probands who had received a confirmed diagnosis of IS and had reported additional family history of IS. Scoliosis was previously diagnosed by the presence of lateral deformity of the spine of at least 10°, with otherwise normal vertebral bodies, and by the exclusion of trauma or coexisting disorders that often involve scoliosis, such as cerebral palsy, spinal muscular atrophy, Marfan syndrome, Ehlers-Danlo syndrome, Charcot-Marie-Tooth disease, neurofibromatosis, spina bifida, etc. Additionally, patients with phenotypes that fit the diagnosis of IS but with left thoracic curves or with abnormal neurological signs were typically screened by magnetic resonance imaging of the neuraxis, to rule out conditions such as syringomyelia. Of the families, >95% were referred through collaborating pediatric orthopedic surgeons at Texas Scottish Rite Hospital for Children (TSRHC) in Dallas. Other than probands, most (46 of 71) of the affected individuals were also current or former patients given diagnoses and treated at TSRHC. All medical records and x-rays were reviewed by a single orthopedic surgeon (J.H.). For the purposes of this study, we included as affected only those individuals meeting the above criteria with radiographic observation of a curve of at least 15°, to reduce possible false positives. Also for the purposes of this study, we excluded any families with IS with a history

From the Seay Center for Musculoskeletal Research (X.G.; D.Z.; R.B.; J.G.; S.W.; S.D.; S.S.; C.W.), Orthopaedics, Texas Scottish Rite Hospital (J.H.), and Department of Orthopaedic Surgery (J.H.; C.W.) and McDermott Center for Human Growth and Development (C.W.), University of Texas Southwestern Medical Center, Dallas; Department of Genetics, Rutgers, The State University of New Jersey, Piscataway (D.G.); Departments of Genetics (C.H.; M.L.; A.B.) and Orthopaedic Surgery (M.D.), Washington University School of Medicine, St. Louis; and Departments of Orthopaedic Surgery and Rehabilitation (J.M.) and Pediatrics (V.S.), University of Iowa School of Medicine, Iowa City

Received November 3, 2006; accepted for publication February 6, 2007; electronically published March 12, 2007.

Address for correspondence and reprints: Dr. Carol Wise, Seay Center for Musculoskeletal Research, Texas Scottish Rite Hospital for Children, 2222 Welborn Street, Dallas, TX 75219. E-mail: carol.wise@tsrh.org

Am. J. Hum. Genet. 2007;80:957–965. © 2007 by The American Society of Human Genetics. All rights reserved. 0002-9297/2007/8005-0015$15.00
DOI: 10.1086/513571

of disease that could involve scoliosis—for example, Marfan syndrome or Duchenne muscular dystrophy—or other diagnosed musculoskeletal deformities. Other connecting family members, either reporting negative history of IS or without documenting records and x-rays, were collected and treated as having an unknown diagnosis. All families included in this study were of European descent. We note that family IS9, of which we had previously performed a genomewide linkage scan,[3] was excluded from analyses presented here. For the cohort of 53 multiplex families (the 52 additional families and IS14), 130 individuals were affected with IS, with curve severities ranging from 15° to 113° and averaging 40.5°; age at first presentation was ~11.5 years, and 86% of families were ascertained via female probands. Because our goal was to identify genes underlying IS susceptibility, and because a range of curve severities is common within families, we did not attempt to distinguish quantitative differences but instead included everyone with curves $\geqslant 15°$ in a single liability class.

Genomic DNA was isolated from whole blood by standard procedures. We performed a genomewide scan in the previously ascertained family IS14 (fig. 1), using polymorphic microsatellite loci evenly spaced at 10–15 cM intervals. Genotyping was performed as described elsewhere, with use of an ABI 377 sequence analysis system.[3] For all analyses of polymorphisms described here and below, allele frequencies were calculated from the data with use of the method implemented in the RECODE program. For the genomewide scan of family IS14, two-point LOD scores were calculated by the MLINK program in the LINKAGE package, with the use of a disease frequency of .01.[9] Nonparametric LOD (NPL) scores of the same data were generated using Genehunter.[10]

Results produced LOD scores $\geqslant 1$ for regions of chromosomes 1p, 10q, and 8q (fig. 2). The latter linkage peak was detected between markers *D8S1477* and *D8S2324,* apparently near the inversion breakpoint described elsewhere in a family with IS.[8] Maximum NPL scores were also obtained for these chromosomes, although the chromosome 8 peaks were not completely overlapping for the two methods. We elected to follow up these results and other published findings[3–7] by testing linkage between IS and microsatellite loci in the 52 additional multiplex families with 123 affected individuals (fig. 3). Genotyping was performed as described above. Because of the uncertainty in inheritance model and penetrance for IS, we initially applied genetic model–free methods, using "affecteds-only" analyses that use the statistical method of Kong and Cox (KAC).[11] This statistic is normally distributed under the null hypothesis of no linkage. Transmission disequilibrium was measured using the transmission/disequilibrium test (TDT), with allowance for errors (TDTae) as implemented in the TDTAE program.[12,13] The TDTae method is robust to missing parental genotype data or to errors that may be introduced when genotyping microsatellite loci. In this analysis, we used the multiplicative model for the TDTae; that is, the genotype relative risk (GRR) for the homozygous genotype was constrained to be the square of the GRR for the heterozygote genotype, making the statistic equivalent to the original TDT when both parents are genotyped.[12,13] Results of the TDTae are reported with correction for tests at multiple alleles. The false-discovery rate (FDR) method[14] was applied to the final data set as further correction for tests at multiple loci, as described below. We obtained the strongest results for 8q12 loci (full results for other chromosomes not shown). This revealed positive evidence of linkage between IS and chromosome 8q loci in the more proximal region that had provided modest evidence of linkage in family IS14 (maximum KAC LOD 2.77; $P = .0028$ at *D8S1136* in the 52 follow-up families) (fig. 4). The TDT unexpectedly revealed some evidence of association between IS and alleles of both *D8S1136* (TDTae $P = .001$) (fig. 4) and the next proximal marker, *D8S1113* (TDTae $P = .016$), although the latter result was not significant ($P < .05$) after correction for multiple tests.

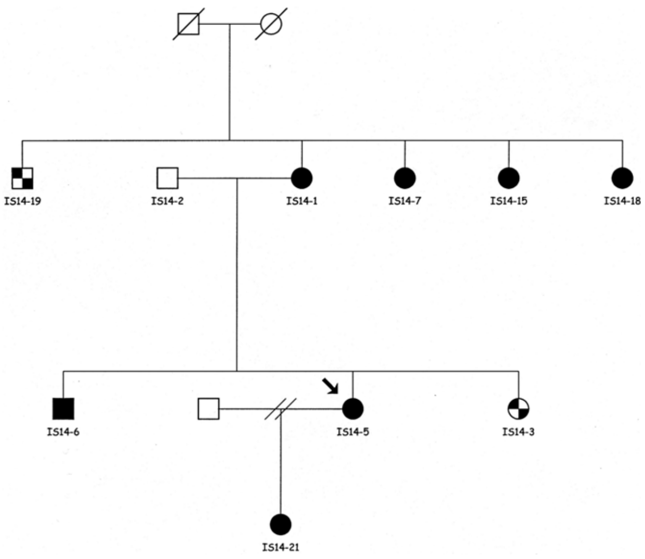

Figure 1. Pedigree of family IS14. Blackened symbols indicate affected individuals, and the arrow indicates the proband. Checkerboard patterns denote individuals with mild scoliosis (<15° Cobb angle) who were scored as "unknown" in subsequent analyses. We note that individual IS14–19 was originally reported as affected[3] but, upon reevaluation, was considered to be of unknown affection because of borderline curve measurements.

The peak of linkage and association we observed occurred about 11 Mb distal to the *SNTG1* gene (MIM *608714) disrupted in the 8q inversion breakpoint described elsewhere.[8] This distinction encouraged us to consider other candidate genes, and we initially selected genes in the 4-cM region between *D8S1113* and *D8S1136.* One of these was the chromodomain helicase DNA-binding protein 7 gene (*CHD7* [MIM *608892]). Missense, stop, and splicing mutations within coding exons of *CHD7* have been identified in ~60% of patients with the syn-

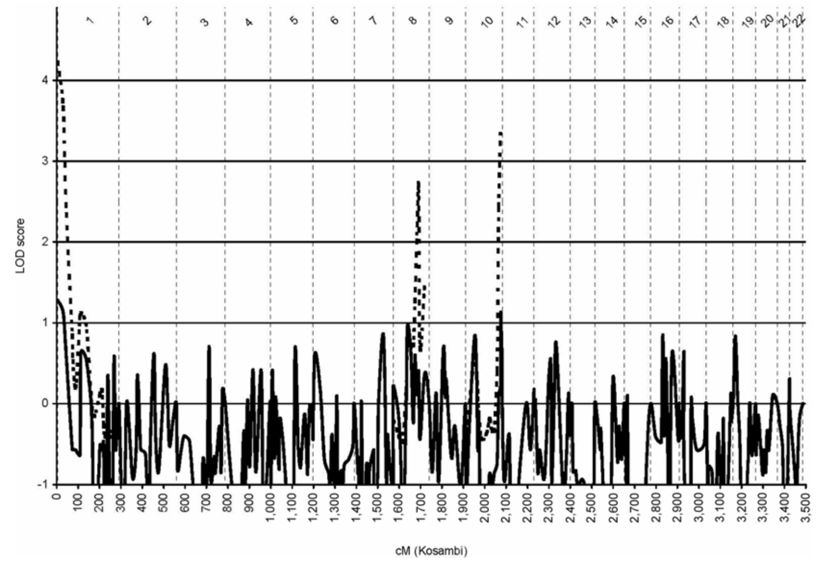

Figure 2. Results of genomewide scan in family IS14. Distance across chromosomes is plotted on the *X*-axis versus results of linkage analysis along the *Y*-axis. Resulting LOD scores for parametric analyses in which we considered only affected individuals and dominant inheritance are plotted as solid lines for each chromosome. Maximal results were obtained from chromosomes 1, 8, and 10. The top three NPL scores also occurred for chromosomes 1, 8, and 10 and are overlaid and plotted as dashed lines.

drome of c̲oloboma of the eye, h̲eart defects, a̲tresia of the choanae, r̲etardation of growth and/or development, genital and/or urinary abnormalities, and e̲ar abnormalities and deafness (CHARGE [MIM #214800]). Infant mortality in CHARGE syndrome is high, but life expectancy has improved as the epidemiology has become better understood.[15–18] In surviving individuals, a high prevalence (>60%) of later-onset scoliosis was recently reported in a series of adolescent and adult patients with CHARGE syndrome.[19] Given this, we hypothesized that milder variants in *CHD7* could underlie IS susceptibility. To test this hypothesis, we performed a fine-mapping study in the region of *CHD7*. In the first analysis, 15 SNP loci evenly spaced throughout the *CHD7* gene were genotyped in the 53 pedigrees, including family IS14. Ten additional SNPs were subsequently selected from within the ~93-kb region producing evidence of association with IS (see below). All polymorphic markers were selected using publicly available information except three, which were selected using the Applied Biosystems Genome Browser. SNP genotyping was performed by amplifying 20 ng genomic DNA in Taqman allele-discrimination assays (Applied Biosystems). Custom Taqman probes[20] for each allele were designed using Primer Express v2.0 software (Applied Biosystems) in accordance with recommended guidelines.[21] Genotyping and analysis were performed using an ABI Prism 7900HT system. Consistency with Hardy-Weinberg equilibrium was verified for all SNP genotypes. Genetic distances for SNPs were interpolated from physical locations (assembly hg18) given in public databases. Two of the SNPs were not sufficiently polymorphic and were dropped, leaving 23 SNPs for all further analyses.

For our single-locus analyses, we considered three genetic model–free methods: (i) the affected sib pair (ASP) method, as implemented in the ANALYZE program[22]; (ii) the haplotype-based haplotype relative risk test (HHRR), as implemented in the ANALYZE program[23]; and (iii) the TDTae,[12,13] as implemented in the TDTAE program. Each of the three statistics complements the others: the ASP statistic tests for linkage and does not use information on linkage disequilibrium (LD), the HHRR statistic is a test of association (i.e., it tests whether there is preferential transmission of a given allele to affected offspring across families), and the TDTae is a test of linkage in the presence of association that also provides estimates of GRRs. Although we did not observe any genotyping errors in these data, we used the TDTae statistic nonetheless, since it is also robust to missing parental genotype information.[13] We used the multiplicative model specification with the TDTae method and restricted our attention to those markers with observed minor-allele frequencies >0.05.

For multipoint analyses, as with the two-point analyses, we used two genetic model–free methods: (i) the affected relative pair method, $Z_{\rm lr}$, as implemented in the GENEHUNTER-PLUS program,[10,11] and (ii) the multilocus TDT statistic, as implemented in the TRANSMIT program (v. 2.1).[24] The $Z_{\rm lr}$ method tests for linkage, whereas the TDT tests for linkage in the presence of association. The formatting of all pedigree data for multipoint analyses was facilitated through use of the Mega2 program.[25] For the $Z_{\rm lr}$ method, we performed multipoint linkage analysis, using all 23 markers. However, we note that these results should be viewed with some caution, since there is intermarker LD, and linkage statistics may inflate the false-

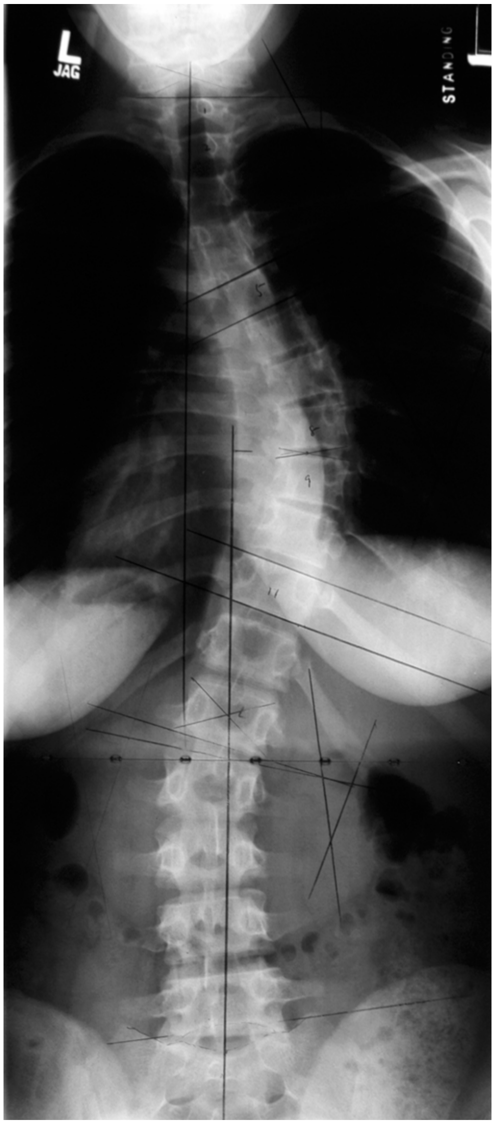

Figure 3. IS in a representative proband from the 52-family set. Standing posteroanterior radiograph reveals a right thoracic curve in an otherwise healthy adolescent female.

positive rates in the presence of missing parental genotype information. The maximum Z_{lr} statistic of 2.63 (P = .004) occurred for marker *rs4738813* at position 10.361. The remaining markers all had Z_{lr} statistics on the order of 2.35 (P = .009) (full results not shown). For the multilocus TDT method, we considered two- and four-locus TDT statistics. Haplotypes and their frequencies were estimated via maximum likelihood, as implemented in TRANSMIT. Because of computation constraints, we computed all consecutive four-locus haplotype TDT statistics in a "sliding window" fashion. That is, we computed each multilocus TDT statistic, using ordered SNPs 1–4, then SNPs 4–7, then SNPs 7–10, etc. The last set of four loci we considered were SNPs 19–22. We also computed a two-locus TDT statistic, using SNPs 22 and 23. To be consistent with the single-locus TDT analyses, we considered only haplotypes whose estimated frequency was >5%. We chose the maximum of the set of TDT statistics for each set of observed haplotypes in the four (or two) loci. *P* values were computed for the maximum TDT statistic in each set of four SNPs, by creating 50,000 bootstrap samples and computing the proportion of bootstrap samples in which the maximum TDT statistic exceeded that of the maximum TDT statistic for the observed data. To combine *P* values for SNPs that were in more than one set of loci (e.g., SNPs 4, 7, and 10), we computed the average of transformed *P* values. For example, if the maximum TDT statistic *P* value is p_1 for the first set of SNPs containing SNP 4 and is p_2 for the second set of SNPs containing SNP 4, then the transformed *P* value is $[-\log(p_1) + -\log(p_2)]/2$.

Altogether, these analyses produced a total of 70 TDT *P* values. To correct for the multiple tests performed, we used a variation of the FDR method that allows for correlated data.[14] Specifically, for the 70 TDTs performed (single-locus and multilocus), we determined the FDR threshold by sorting the test *P* values p_i for i between 1 and 70 and relabeling the sorted *P* values as $p_{(i)}$, so that $p_{(1)} \leqslant p_{(2)} \ldots \leqslant p_{(70)}$. When we let $t_i = \min[\alpha, 70 \times \alpha/(71 - i)^2]$ (where α is the significance level—in this case, 0.05), then we declared those *P* values $p_{(i)}$ that satisfied the property $p_{(i)} \leqslant t_i$ to be significant after correction for multiple testing. The FDR threshold for TDT analyses was computed to be 1.9×10^{-4}.

Results of the three single-locus analyses (i.e., ASP, HHRR, and TDTae) are given graphically in figure 5*B* and are listed in table 1. Examination of all 23 SNPs revealed a peak of association encompassing exons 2–4. The strongest evidence of association was obtained for SNP marker

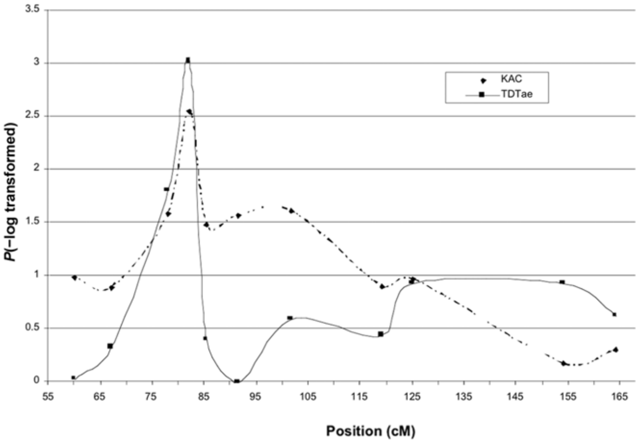

Figure 4. Analyses of linkage and transmission disequilibrium for 8q microsatellite loci in 52 families with IS. Polymorphic microsatellites spaced at 5–10 cM were genotyped in all members of the 52 families. The method of KAC[11] was used to compute linkage (*dashed line*), and family-based association was measured using the TDTae (*solid line*). For reporting consistency, results are shown as *P* values (−log transformed) versus position (in cM) for the two methods.

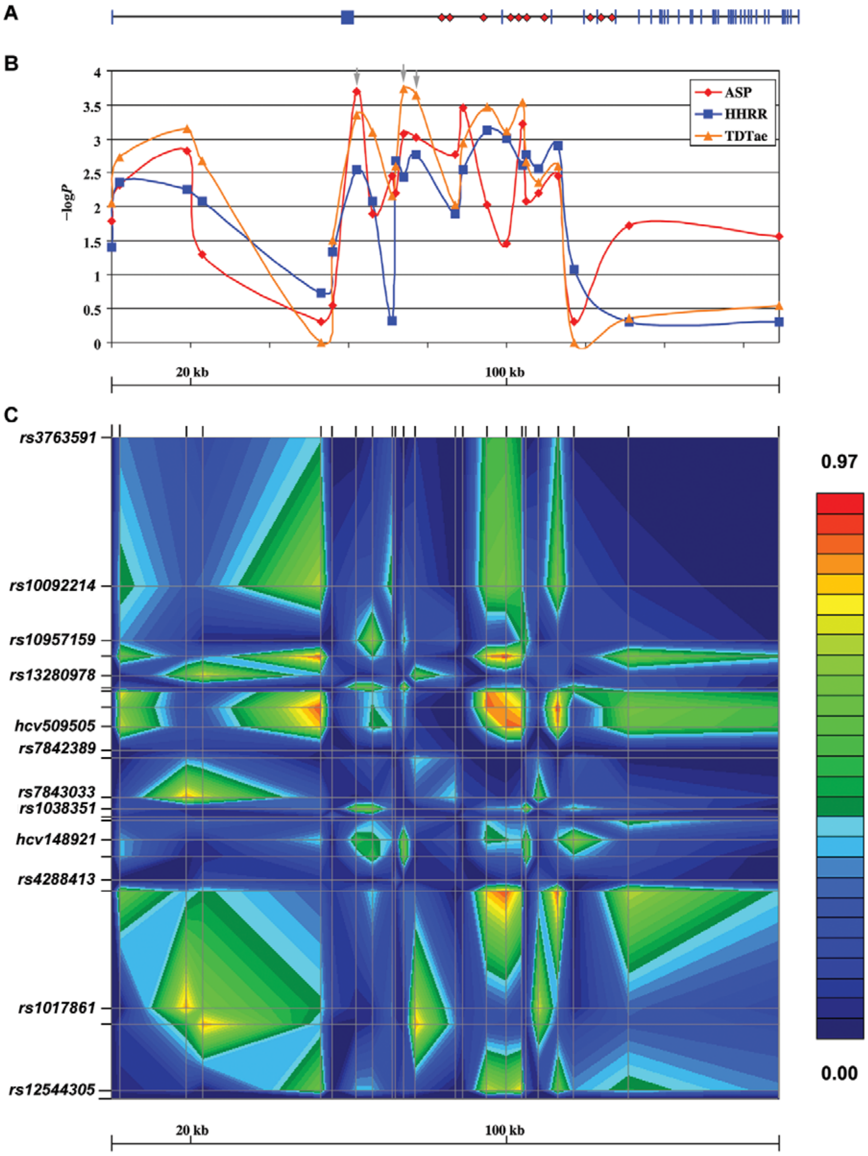

Figure 5. Fine-mapping results for the *CHD7* gene. *A, CHD7* genomic region is shown with exons indicated in blue and intronic conserved sequence blocks shown in red. *B,* Plot of single-point linkage and transmission disequilibrium *P* values for 23 SNPs in the *CHD7* gene. $-\log_{10}P$ values are plotted along the *Y*-axis versus physical position along the *X*-axis for each SNP. Results for the ASP, HHRR, and TDTae statistics are shown by diamonds (*red*), squares (*blue*), and triangles (*orange*), respectively. The TDTae method is more significant than the ASP method for markers in a region of high pairwise LD, as expected, given that the TDT method was originally developed to increase evidence of linkage when marker and trait loci are in high LD. *C,* Graphical representation of the pairwise LD (Δ^2) values for all 23 SNPs. This panel may be thought of as a "heat map" of the pairwise LD values (Δ^2). Both the horizontal and vertical axes represent the 23 SNPs shown by tick marks and ordered as in table 1, and places where the axes intersect indicate the pairwise Δ^2 values for those marker pairs. In this panel, pairs of markers with larger Δ^2 values (close to or equal to 1, indicating complete LD) are denoted in red, whereas pairs with smaller Δ^2 values (closer to or equal to 0, indicating linkage equilibrium) are denoted in blue, as illustrated by the vertical color bar. Panels A, B, and C have been aligned so that results for each of the markers correspond among the three panels.

11 (*rs1038351*; TDTae $P = .00018$) (fig. 5 and table 1), which was slightly more significant in this test than SNP marker 12. Two-point LOD score analysis also produced supporting evidence of linkage (maximum LOD 2.72; $P = 2.0 \times 10^{-4}$ for SNP marker 7, *rs7000766*) (table 1). Multilocus analyses revealed significant overtransmission of overlapping haplotypes, with strongest results for haplotypes containing SNPs 14–19 (fig. 6). We also computed levels of LD, as determined by the squared correlation coefficient[26] Δ^2, for all pairs of the 23 SNPs, using the fine-mapping pedigree data. These coefficients were computed using the method implemented in the GOLD software.[27]

Pairwise estimates of LD for all 23 SNPs revealed the highest LD within the region defined by SNPs 15–19, indicated by yellow or red in figure 5*C*. Specifically, the set of three consecutive SNP markers 15, 16, and 17 all displayed pairwise correlation (Δ^2) values close to 1. The pairwise Δ^2 values for the pairs 15–16, 15–17, and 16–17 were 0.892, 0.897, and 0.773, respectively, with a minimum χ^2 value of 53.97 (1 df) ($P = 2.0 \times 10^{-13}$). We note that these three SNP markers are also among those (along with SNPs 11 and 12) that showed the strongest results with the HHRR and TDTae statistics (fig. 5*B* and table 1). Given that marker-marker LD is often used as a surrogate for disease locus–marker disequilibrium,[28,29] we concluded that all the multilocus data together provided the strongest evidence of association with IS in a region encompassing SNPs 14–19, followed by a region encompassed by SNPs 7–12.

The *CHD7* gene spans 188 kb and contains one noncoding (exon 1) and 37 coding exons (fig. 5). The SNP loci we found to be associated with IS susceptibility are clearly contained within an ~116-kb region encompassing exons 2–4 of the *CHD7* gene. We searched for potential functional elements in this region and extending out to exon 7, by comparing publicly available reference sequences across vertebrate species. Reference human *CHD7* genomic sequence was compared with other vertebrate (mouse, rat, rabbit, dog, armadillo, elephant, opossum, and chicken) *CHD7* sequences, with use of the University of California–Santa Cruz (UCSC) Genome Browser conservation track. Regions showing evidence of sequence conservation between multiple vertebrate species were analyzed further for potential variation, with the use of SNPBLAST, available at the National Center for Biotech-

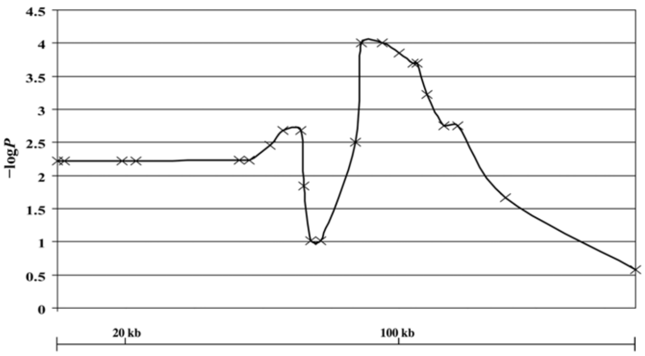

Figure 6. Maximum multilocus TDT results for each set of four SNPs. For SNPs that appear in more than one set of overlapping windows, we report the average of the log-transformed *P* values for the two maximum multilocus TDT statistics. *P* values were computed using a bootstrap sample of 50,000.

Table 1. Results of Two-Point Analyses for the 23 SNPs in the *CHD7* Gene

	Location		ASP	Overtransmitted	Associated *P* (−log*P*)[b]			GRR[c]	
SNP	(bp)	Locus[a]	LOD	Allele	ASP	HHRR	TDTae	R_1	R_2
1	61,758,225	*rs4738813*	.988	C	.016 (1.80)	.039 (1.41)	.009 (2.05)	1.900	3.609
2	61,760,363	*rs12544305*	1.461	G	.005 (2.30)	.004 (2.40)	.002 (2.70)	2.373	5.627
3	61,777,438	*rs9643371*	1.912	T	.002 (2.70)	.006 (2.22)	.0007 (3.15)	2.478	6.139
4	61,781,267	*rs1017861*	.584	G	.05 (1.30)	.008 (2.10)	.002 (2.70)	2.084	4.342
5	61,811,105	*rs13256023*	.000	T	.50 (.30)	.184 (.74)	1.000 (.00)	1.000	1.000
6	61,814,698	*rs4288413*	.071	A	.284 (.55)	.046 (1.34)	.030 (1.52)	1.755	3.079
7	61,820,387	*rs7000766*	2.718	G	.0002 (3.70)	.003 (2.53)	.0005 (3.30)	2.701	7.294
8	61,824,902	*hcv148921*	1.084	A	.013 (1.89)	.008 (2.10)	.0008 (3.10)	2.196	4.820
9	61,829,834	*rs1483207*	1.578	G	.004 (2.40)	.486 (.31)	.007 (2.15)	2.222	4.933
10	61,830,452	*rs1483208*	1.353	A	.006 (2.22)	.002 (2.70)	.003 (2.52)	2.284	5.216
11	61,832,862	*rs1038351*	2.145	T	.0008 (3.10)	.004 (2.40)	.0002 (3.70)	3.059	9.355
12	61,835,069	*rs7843033*	2.089	C	.001 (3.00)	.002 (2.70)	.0002 (3.70)	2.994	8.961
13	61,845,746	*rs7002806*	1.857	T	.002 (2.70)	.013 (1.89)	.009 (2.05)	2.049	4.200
14	61,847,748	*rs7842389*	2.491	T	.0004 (3.40)	.003 (2.52)	.001 (3.00)	2.518	6.341
15	61,853,914	*rs7017676*	1.197	A	.009 (2.05)	.0007 (3.15)	.0003 (3.52)	2.860	8.182
16	61,858,259	*hcv509505*	.712	G	.035 (1.46)	.001 (3.00)	.0008 (3.10)	2.455	6.028
17	61,862,485	*rs4392940*	2.269	A	.0006 (3.22)	.002 (2.70)	.0003 (3.52)	2.909	8.460
18	61,863,611	*rs4237036*	1.242	T	.008 (2.10)	.002 (2.70)	.002 (2.70)	2.340	5.476
19	61,866,609	*rs13280978*	1.354	T	.006 (2.22)	.003 (2.53)	.004 (2.40)	2.105	4.431
20	61,871,613	*rs4301480*	1.570	A	.004 (2.40)	.001 (3.00)	.003 (2.52)	2.498	6.240
21	61,874,997	*rs10957159*	.000	G	.50 (.30)	.084 (1.08)	1.000 (.00)	1.000	1.000
22	61,889,433	*rs10092214*	.940	A	.019 (1.72)	.50 (.30)	.434 (.36)	1.181	1.395
23	61,926,943	*rs3763591*	.800	T	.027 (1.57)	.50 (.30)	.288 (.54)	1.289	1.660

[a] Each locus and the corresponding overtransmitted allele are shown.

[b] Associated *P* values and −log*P* values are shown for the ASP method measuring linkage and for the TDTae and HHRR methods measuring family-based association.

[c] Two-point GRRs for each locus are shown, for which we assumed a log-additive model of inheritance for the disease. The values presented in the table are maximum-likelihood estimates of these values for each of the 23 SNPs. In this analysis, the GRR for the homozygous genotype was constrained to be the square of the GRR for the heterozygote genotype, making our statistic equivalent to the original TDT statistic when both parents are genotyped.

nology Information (NCBI) Web site. Similarity to consensus transcription-factor binding site sequences was identified using TFSEARCH.[30] In these analyses, results were restricted to those that surpassed a threshold score of 85.0 in searches of vertebrate species. This identified 16 blocks of relatively high sequence conservation, with coding exons 2–7 comprising 6 of these blocks (fig. 5).

To identify variants underlying the association with IS susceptibility, we resequenced these exons and flanking regions in 25 affected probands and 44 parental controls. To optimize the probability of detecting risk alleles, we selected probands who were homozygous for the majority of overtransmitted alleles for SNPs 2–20. Selected regions of the *CHD7* gene were amplified from DNA samples of the 25 affected cases and 44 parental controls via PCR (primer sequences and PCR conditions available upon request). Amplicons were analyzed by direct DNA capillary sequencing with use of a 3730 XL (Applied Biosystems) instrument. Chromatograms were searched for heterozygous variants with sequencing analysis 5.1.1 software. All sequences were aligned using the Sequencher program (Genecodes) and were compared with human reference sequence (hg18) from publicly available databases. This revealed two rare coding SNPs, one of which predicted a nonsynonymous change in parental controls, but this change was not transmitted to affected offspring. We also identified two previously described intronic SNPs (*rs7836586* and *rs4540437*),[15] but we could not ascribe obvious function to the transmitted or nontransmitted alleles (table 2).

We next searched the 10 remaining conserved sequence blocks *in silico* for similarity to known functional elements. One region, sequence block 3, was found to harbor the highest density of predicted transcription-factor binding sites. In particular, 30 independent consensus sequences for caudal-type (cdxA) sites were predicted in the ~700 bp comprising conserved block 3 (table 3). The caudal homeobox transcription factors are required for anterior/posterior positional cues and appropriate embryonic axial development in model organisms.[31,32] Resequencing this block in case patients and parental controls revealed a polymorphism, *rs4738824,* which predicts disruption of a possible binding site for caudal-type (cdx) homeodomain-containing transcription factors. Specifically, in this polymorphism, an A nucleotide that appears to be perfectly conserved across nine vertebrate species is replaced by a G nucleotide. We analyzed SNP *rs4738824* in the remaining families and found significant overtransmission of the G allele predicted to disrupt cdx binding (TDTae $P = .005$). Furthermore, the set of consecutive SNP markers 14–19 all displayed high pairwise Δ^2 values (>0.6) with this SNP, which lies between the original SNPs 14 and 15 (full results not shown). Given the convergence of linkage, LD, and sequence conservation, it is intriguing to speculate that SNPs such as *rs4738824* may confer functional effects on CHD7. However, it is equally possible that the associated variants we have detected are in LD with true causal alleles; further mapping and functional studies of CHD7 in additional IS collections are important to elucidate this.

Table 2. Polymorphisms Observed by Resequencing in 25 Patients Affected with IS and 44 Parental Controls

SNP and Genotype	Frequency[a]	Exon or Intron	Variant[b]	dbSNP Accession Number
1:		Exon 2	M340V	*ss68756179*
AA	67			
AG	2			
GG	0			
2:		Exon 2	P544P	*ss68756180*
CC	64			
CG*	4			
GG	0			
3:		Intron 2	c. 1665+34	*rs7836586*
AA	55			
AG	13			
GG	0			
4:		Intron 2	c.1666−3238	*rs4738824*
AA	0			
AG	13			
GG	54			
5:		Intron 4	c.2238+39	*rs4540437*
AA	59			
AG	13			
GG	00			

[a] We show genotype frequencies for the total set of resequenced individuals, without distinguishing related versus unrelated chromosomes, except for SNPs 1 and 2. For these, we note that SNP *ss68756179* was observed twice, in two unrelated controls, whereas SNP *ss68756180* (denoted with an asterisk) was observed in four related cases.

[b] Variants observed in the CHD7 mRNA or predicted protein are shown.

Table 3. Analyses of Conserved Sequence Blocks in the *CHD7* Gene

Block and Predicted Transcription-Factor Binding Sites[a]	No. of Sites[b]	Location[c] (bp)	Size (bp)
1:		61,844,311–61,844,737	426
Cdx	11		
SRY	5		
2:		61,845,377–61,845,721	344
Cdx	7		
Nkx-2	5		
3:		61,852,850–61,853,569	719
Cdx	30		
SRY, GATA-n	7		
4:		61,856,975–61,857,402	427
Cdx	8		
SRY	3		
5:		61,858,961–61,859,577	616
Cdx	15		
SRY, Oct-1	6		
6:		61,860,005–61,860,527	522
Cdx	19		
Nkx-2, Oct-1	3		
7:		61,867,600–61,868,85	1,250
Cdx	30		
SRY, GATA-n, C/EBPn	8		
8:		61,875,850–61,876,200	350
Cdx	14		
9:		61,878,250–61,878,500	250
Cdx	6		
Nkx-2	3		
10:		61,882,400–61,883,120	720
Cdx	21		
SRY	6		

[a] The two most abundant transcription-factor binding sites, as predicted by TFSEARCH,[28] are shown. For block 8, no transcription-factor binding sites were predicted more than once, other than Cdx.

[b] The number of independent consensus sequences identified for each transcription-factor binding site.

[c] Locations of DNA sequence conservation, as identified using the UCSC conservation track, are shown.

Idiopathic forms of scoliosis have been described for centuries, but the etiology has remained a clinical conundrum. Our results are the first description of a responsible gene and provide an initial insight into underlying disease mechanisms. Haploinsufficiency of CHD7 protein during embryogenesis has been proposed to explain the CHARGE syndrome in the presence of *CHD7* coding mutations.[17] We likewise hypothesize that a relative reduction of functional CHD7 in the postnatal period, particularly during the adolescent growth spurt, may disrupt normal growth patterns and predispose an individual to spinal deformity. This may be mediated at the transcriptional level by interaction of *CHD7 cis*-acting elements with factors such as cdx. Beyond the evidence of linkage, we elected to study the *CHD7* gene as a candidate for IS because of its previous association with a disease involving scoliosis. One implication of our findings is that variation in genes responsible for other rare disorders involving nonstructural scoliotic spinal deformity may likewise contribute to IS susceptibility. Our results also suggest that downstream targets of *CHD7* may be important players in IS pathogenesis.

Acknowledgments

We thank the families for their participation; J. Birch, C. Johnston, L. Karol, K. Rathjen, B. S. Richards, D. Sucato, J. Purvis, for patient referrals; N. Ritter, L. Edwards, and L. Macleod, for help with the families; and J. Bainbridge-Smith for help in analysis. For help with genotyping and sequencing, we thank the McDermott Center Sequencing Core at University of Texas Southwestern Medical Center and the Division of Human Genetics Core Genotyping Facility at Washington University School of Medicine in St. Louis. Support from the Scoliosis Research Society, Fondation Cotrel, and TSRHC Research Fund grant 12-96-381 (to C.W.) is gratefully acknowledged.

Web Resources

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

Applied Biosystems Genome Browser, http://marketing.appliedbiosystems.com

dbSNP, http://www.ncbi.nlm.nih.gov/SNP/ (for *CHD7* variants predicting M340V [accession number *ss68756179*] and P544P [accession number *ss68756180*])

NCBI, http://www.ncbi.nlm.nih.gov/

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for IS, *SNTG1, CHD7,* and CHARGE syndrome)

RECODE, http://watson.hgen.pitt.edu/register/

UCSC Genome Browser, http://genome.ucsc.edu/cgi-bin/hgGateway

References

1. Herring JA (ed) (2002) Tachdjian's pediatric orthopaedics. Vol 1. WB Saunders, Philadelphia, PA
2. Emery AEH, Rimoin DL (1990) Principles and practice of molecular genetics. Churchill Livingstone, New York, NY
3. Wise CA, Barnes R, Gillum J, Herring JA, Bowcock AM, Lovett M (2000) Localization of susceptibility to familial idiopathic scoliosis. Spine 25:2372–2380
4. Salehi LB, Mangino M, De Serio S, De Cicco D, Capon F, Semprini S, Pizzuti A, Novelli G, Dallapiccola B (2002) Assignment of a locus for autosomal dominant idiopathic scoliosis (IS) to human chromosome 17p11. Hum Genet 111: 401–404
5. Chan V, Fong GC, Luk KD, Yip B, Lee MK, Wong MS, Lu DD, Chan TK (2002) A genetic locus for adolescent idiopathic scoliosis linked to chromosome 19p13.3. Am J Hum Genet 71:401–406
6. Justice CM, Miller NH, Marosy B, Zhang J, Wilson AF (2003) Familial idiopathic scoliosis: evidence of an X-linked susceptibility locus. Spine 28:589–594
7. Miller NH, Justice CM, Marosy B, Doheny KF, Pugh E, Zhang J, Dietz HC 3rd, Wilson AF (2005) Identification of candidate gene regions for familial idiopathic scoliosis. Spine 30:1181–1187
8. Bashiardes S, Veile R, Allen M, Wise CA, Dobbs M, Morcuende JA, Szappanos L, Herring JA, Bowcock AM, Lovett M (2004)

SNTG1, the gene encoding γ1-syntrophin: a candidate gene for idiopathic scoliosis. Hum Genet 115:81–89

9. Lathrop GM, Lalouel JM, Julier C, Ott J (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81:3443–3446
10. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363
11. Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188
12. Gordon D, Heath SC, Liu X, Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am J Hum Genet 69:371–380
13. Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM, Gordon D, Ott J (2004) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. Eur J Hum Genet 12:752–761
14. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125:279–284
15. Vissers LE, van Revenswaaij CM, Admiraal R, Hurst JA, de Vries BB, Janssen IM, van der Vliet WA, Huys EH, de Jong PJ, Hamel BC, et al (2004) Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. Nat Genet 36:955–957
16. Lalani SR, Safiullah AM, Fernbach SD, Harutyunyan KG, Thaller C, Peterson LE, McPherson JD, Gibbs RA, White LD, Hefner M, et al (2006) Spectrum of *CHD7* mutations in 110 individuals with CHARGE syndrome and genotype-phenotype correlation. Am J Hum Genet 78:303–314
17. Jongmans MC, Admiraal RJ, van der Donk KP, Vissers LE, Baas AF, Kapusta L, van Hagen JM, Donnai D, de Ravel TJ, Veltman JA, et al (2006) CHARGE syndrome: the phenotypic spectrum of mutations in the CHD7 gene. J Med Genet 43:306–314
18. Sanlaville D, Etchevers HC, Gonzales M, Martinovic J, Clement-Ziza M, Delezoide AL, Aubry M-C, Pelet A, Chemouny S, Cruaud C, et al (2006) Phenotypic spectrum of CHARGE syndrome in fetuses with CHD7 truncating mutations correlates with expression during human development. J Med Genet 43:211–217
19. Doyle C, Blake K (2005) Scoliosis in CHARGE: a prospective survey and two case reports. Am J Med Genet A 133:340–343
20. Livak KJ (1999) Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. Genet Anal 14:143–149
21. de Kok JB, Wiegerinck ET, Giesendorf BA, Swinkels DW (2002) Rapid genotyping of single nucleotide polymorphisms using novel minor groove binding DNA oligonucleotides (MGB probes). Hum Mutat 19:554–559
22. Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56:777–787
23. Terwilliger JD, Ott J (1992) A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered 42:337–346
24. Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177
25. Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE (2005) Mega2: data-handling for facilitating genetic linkage and association analyses. Bioinformatics 21:2556–2557
26. Hill WG, Weir BS (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. Am J Hum Genet 54:705–714
27. Abecasis GR, Cookson WO (2000) GOLD—graphical overview of linkage disequilibrium. Bioinformatics 16:182–183
28. Nielsen DM, Ehm MG, Weir B (1998) Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet 63:1531–1540
29. Gordon D, Simonic I, Ott J (2000) Significant evidence for linkage disequilibrium over a 5-cM region among Afrikaners. Genomics 66:87–92
30. Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, et al (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. Nucleic Acids Res 26:362–367
31. Subramanian V, Meyer BI, Gruss P (1995) Disruption of the murine homeobox gene *Cdx1* affects axial skeletal identities by altering the mesodermal expression domains of *Hox* genes. Cell 83:641–653
32. Lohnes D (2003) The Cdx1 homeodomain protein: an integrator of posterior signaling in the mouse. Bioessays 25:971–980