

UCSF

UC San Francisco Previously Published Works

Title

A PAX1 enhancer locus is associated with susceptibility to idiopathic scoliosis in females

Permalink

<https://escholarship.org/uc/item/86g9p29b>

Journal

Nature Communications, 6(1)

ISSN

2041-1723

Authors

Sharma, Swarkar
Londono, Douglas
Eckalbar, Walter L
et al.

Publication Date

2015

DOI

10.1038/ncomms7452

Peer reviewed

ARTICLE

Received 4 Aug 2014 | Accepted 30 Jan 2015 | Published 18 Mar 2015

DOI: 10.1038/ncomms7452

OPEN

A *PAX1* enhancer locus is associated with susceptibility to idiopathic scoliosis in females

Swarkar Sharma^{1,†}, Douglas Londono², Walter L. Eckalbar³, Xiaochong Gao¹, Dongping Zhang¹, Kristen Mauldin¹, Ikuyo Kou⁴, Atsushi Takahashi⁵, Morio Matsumoto⁶, Nobuhiro Kamiya^{1,†}, Karl K. Murphy³, Reuel Cornelia¹, TSRHC Scoliosis Clinical Group*, Japan Scoliosis Clinical Research Group*, John A. Herring^{7,8}, Dennis Burns⁹, Nadav Ahituv³, Shiro Ikegawa⁴, Derek Gordon² & Carol A. Wise^{1,8,10,11}

Idiopathic scoliosis (IS) is a common paediatric musculoskeletal disease that displays a strong female bias. By performing a genome-wide association study (GWAS) of 3,102 individuals, we identify significant associations with 20p11.22 SNPs for females ($P=6.89 \times 10^{-9}$) but not males ($P=0.71$). This association with IS is also found in independent female cohorts from the United States of America and Japan (overall $P=2.15 \times 10^{-10}$, OR=1.30 (rs6137473)). Unexpectedly, the 20p11.22 IS risk alleles were previously associated with protection from early-onset alopecia, another sexually dimorphic condition. The 174-kb associated locus is distal to *PAX1*, which encodes paired box 1, a transcription factor involved in spine development. We identify a sequence in the associated locus with enhancer activity in zebrafish somitic muscle and spinal cord, an activity that is abolished by IS-associated SNPs. We thus identify a sexually dimorphic IS susceptibility locus, and propose the first functionally defined candidate mutations in an enhancer that may regulate expression in specific spinal cells.

¹Sarah M. and Charles E. Seay Center for Musculoskeletal Research, Research Department, Texas Scottish Rite Hospital for Children, Dallas, Texas 75219, USA. ²Department of Genetics and Human Genetics Institute, Rutgers University, Piscataway, New Jersey 08854, USA. ³Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics, University of California San Francisco, San Francisco, California 94143, USA. ⁴Laboratory of Bone and Joint Diseases, Center for Integrative Medical Sciences, RIKEN, Tokyo 108-8639, Japan. ⁵Laboratory for Statistical Analysis, Center for Integrative Medical Sciences, RIKEN, Yokohama 230-0045, Japan. ⁶Department of Orthopaedic Surgery, School of Medicine, Keio University, Tokyo 108-8345, Japan. ⁷Department of Orthopaedics, Texas Scottish Rite Hospital for Children, Dallas, Texas 75219, USA. ⁸Department of Orthopaedic Surgery, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75390, USA. ⁹Department of Pathology, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75390, USA. ¹⁰McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75390, USA. ¹¹Department of Pediatrics, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75390, USA. † Present addresses: School of Biotechnology, Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir 182320, India (S.S.); Department of Sports Medicine, Tenri University, Tenri, Nara 632-0071, Japan (N.K.). Correspondence and requests for materials should be addressed to C.A.W. (email: carol.wise@tsrh.org).

*Full list of members and affiliations appears at the end of the paper.

Scoliosis is defined as a curvature of the thoracolumbar spine greater than 10° in the coronal plane. Scoliosis is frequently secondary to other underlying diseases or seen as part of the phenotypic spectrum of heritable syndromes, in particular, disorders of neuromuscular or connective tissue development. In other patients, so-called 'congenital' scoliosis arises from frank malformations of the vertebrae that are most often isolated segmentation defects. However, for more than 80% of patients, the origins of scoliosis are unknown, occurring in individuals who are otherwise healthy and bear no obvious structural deficiencies in the vertebral column and associated soft tissues¹. This 'idiopathic' scoliosis (IS) is the most common paediatric musculoskeletal disorder, affecting $\sim 3\%$ of children worldwide². The onset of IS typically coincides with the adolescent growth spurt. Affected individuals are at risk for increasing deformity until growth ceases, although patients with large curves ($>50^\circ$) may continue to worsen more slowly throughout adulthood^{1,3}. Severe IS warrants surgical correction to prevent later disfigurement and other co-morbidities including back pain and pulmonary dysfunction^{1,2}. Progression of the deformity can be rapid, prompting many states in the United States to require school screenings for early detection⁴.

Multiple biologic phenotypes are suspected in IS but have proven difficult to define, particularly as the involved structures, that is, the spine and associated soft tissues, appear superficially normal⁵. Functional and clinical assessments have associated IS with both neuropathologic/neuromuscular as well as connective tissue abnormalities^{5,6}. These studies hint at multiple subphenotypes in human IS but have not defined causality. Aetiologic understanding of IS has also suffered from a lack of tractable, genetically defined animal models that clearly recapitulate the phenotype. However, recent descriptions of IS-like phenotypes in both naturally occurring and genetically engineered teleosts hold out promise in this regard⁷. In particular, forward genetic and genome editing approaches in zebrafish (*Danio rerio*) have recently produced strains with an IS-like phenotype, suggesting that this system may be a powerful tool for modeling IS and defining its disease mechanisms^{8,9}.

IS is a sexually dimorphic disease¹⁰. Girls and boys exhibit a striking difference in the prevalence of progressive IS, with girls having approximately tenfold greater risk of progressive curves that require operative treatment¹¹. This dichotomy in female/male disease expression, and its correlation with the adolescent growth spurt have prompted investigations of hormonal influences in the development and progression of female IS⁶. Sexual dimorphism in IS has also been attributed to differences in genetic loading between males and females, with the least affected sex (males) requiring a stronger genetic load to acquire disease, a so-called Carter effect. One study of a cohort of multiplex families has provided epidemiological evidence for this genetic model¹².

Genetic influences in IS were first suggested almost a century ago⁵. Segregation studies support a complex inheritance model in which multiple genetic factors contribute greater than 80% of the overall disease risk in IS^{13,14}. Although most cases are sporadic, about 25% of IS patients report family history of IS, and more rare families with apparent Mendelian, or single-gene inheritance are described¹⁵. Early gene discovery efforts centred on applying traditional linkage mapping methods to search for causal genes in such extended pedigrees. These studies defined five IS candidate loci as noted in the Online Mendelian Inheritance in Man (OMIM): OS1 (OMIM 181800, chr19p13.3), OS2 (OMIM 607354, chr 17p11), OS3 (OMIM 608765, chr 8q12.1-12.2), OS4 (OMIM 612238, chr 9q31-q34) and OS5 (OMIM 612239, chr 17q25-qter). All but OS3 were defined in single extended pedigrees^{16,17-19}. Common single-nucleotide polymorphisms (SNPs) for one candidate gene within OS3, *CHD7*, were

significantly associated with IS in a cohort of 52 families¹⁹. Otherwise, the causal genes and mutations encoded within familial IS disease loci have not been forthcoming, most likely due to issues of genetic heterogeneity in IS that can confound traditional gene discovery approaches²⁰.

As with other complex genetic disorders, the overall genetic architecture of IS is expected to reflect genetic factors with varying frequencies and effect sizes that will be discoverable by sequence- and haplotype-based methods²¹. Population-based genome-wide association studies (GWAS), a method only recently applied to IS, have proven powerful and efficient for mapping common susceptibility loci for hundreds of complex human traits²². Three published GWAS of IS have begun to define IS susceptibility loci²³⁻²⁵. A locus on chromosome 10q24.1 is the most studied and was initially identified by GWAS of 1,033 East Asian (Japanese) cases and 1,473 matched controls²⁴. A recent combined analysis from multiple ethnic groups (that is, mostly East Asian and non-Hispanic white, NHW) provided further evidence for the locus, which is in the proximity of the *LBX1* gene (combined $P = 1.22 \times 10^{-43}$ for rs11190870)²⁶. *LBX1* encodes the ladybird homeobox 1 protein that is important for early muscle patterning as well as specification of dorsal horn neurons in developing spinal cord²⁷⁻²⁹. Causal mutations underlying this association and their potential effects on *LBX1* and/or other genes are not yet defined. A separate expanded analysis of the original East Asian GWAS yielded significant association with SNPs at a second locus, within the *GPR126* gene in chromosome 6, a result that was replicated in both East Asian and NHW cohorts²⁵. *GPR126* encodes G-protein-coupled receptor 126 that is critical in early neurologic development and Schwann cell myelination³⁰. It is interesting that SNPs in *GPR126* also have been associated with sitting height in humans³¹. In total, reported associations are estimated to explain less than 5% of the overall genetic contribution to disease risk. How these loci function in IS pathogenesis is as yet undefined^{25,26}.

To discover new genetic risk factors for IS, we performed a two-stage GWAS in 3,102 individuals. Our results define a new susceptibility locus encoding associated SNPs that, surprisingly, are also associated with androgenic alopecia (AGA), or male pattern baldness. We find that the locus is specifically associated with female IS, suggesting that it contributes to the sexually dimorphic expression of the disease. By functional fine-mapping assays in zebrafish, we further define a sequence in the associated locus with enhancer activity that is abolished by IS-associated SNPs. Altogether, our results identify the first functionally characterized candidate mutations for IS susceptibility and expand our understanding of the role of non-coding regulatory elements in the disease. Our findings also suggest hypotheses to explain disease pathogenesis and provide the first insights into its puzzling sexual dimorphism.

Results

Association with common variants near the *PAX1* gene.

Research subjects included in the two-stage GWAS were ascertained in Paediatric Orthopedic Clinics at the Texas Scottish Rite Hospital for Children (TSRHC). Affected individuals met standard criteria for a diagnosis of IS and had spinal deformity measuring at least 15° by the Cobb angle method (Fig. 1a,b). The GWAS I-715 included 715 trios from 702 trio families (parents and affected offspring, $N = 1,876$ total) a portion of which was previously described²³. The GWAS II case-control study included 482 independent IS-affected cases and 744 controls of self-reported NHW ethnicity that were independent of GWAS I-715. Further ascertainment, genotyping, quality control and statistical methods are described in the Methods section. In the first stage,

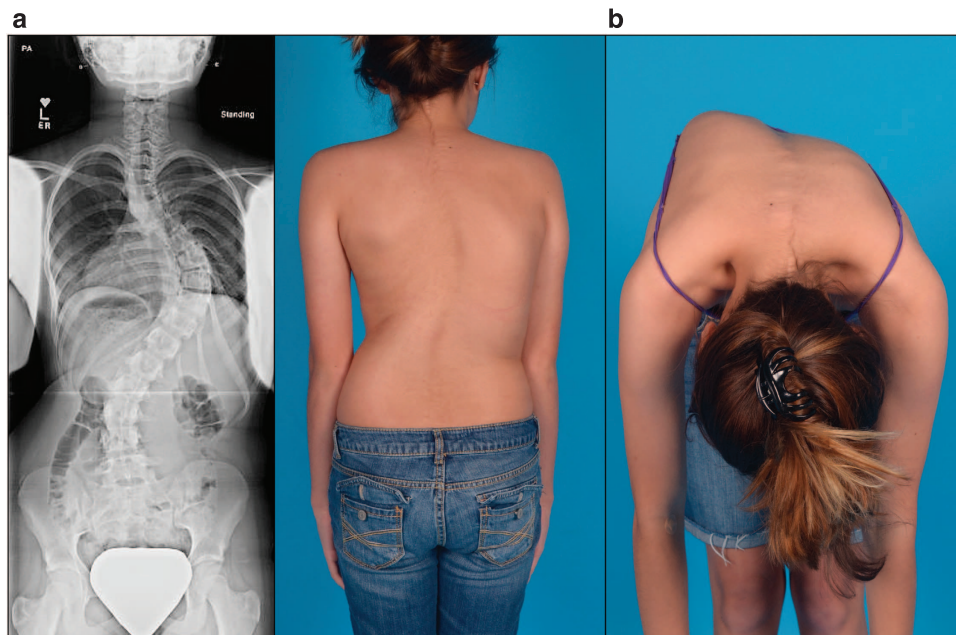


Figure 1 | Idiopathic scoliosis in an adolescent female. (a) Lateral radiograph reveals prototypical right thoracic spinal curvature. Shoulder imbalance is evident in clinical photograph. (b) Prominent rib hump is evident on forward bending.

we genotyped GWAS II cases and controls using the Illumina HumanOmniExpress-12 v1.0 beadchip containing 730,498 markers. We used PLINK v1.07 (ref. 32) to perform quality assurance (QA) and test the data for association using the Cochran Armitage trend test (CATT). This analysis replicated the previously reported associations near *LBX1* and *GPR126* (refs 24–26). However, we found strongest results with SNPs rs6137473 (CATT, $P = 5.58 \times 10^{-7}$, odds ratio (OR) = 1.53, 95% confidence interval (CI) = 1.29–1.80 for risk allele G) and rs169311 (CATT, $P = 1.25 \times 10^{-6}$, OR = 1.51, 95% CI = 1.28–1.78 for risk allele A) in a region of chromosome 20p11.22 between the *PAX1* and *FOXA2* genes (Fig. 2). SNPs in the 20p11.22 inter-genic region remained the most significantly associated with IS after imputing additional genotypes for all of chromosome 20 (Supplementary Fig. 1). For the second stage, we expanded a prior GWAS²³ from 419 to 715 parent-offspring trios ('GWAS I-715', 1,876 individuals). The ethnic composition of this cohort is given in Supplementary Fig. 2. All trios were analysed together using *TDT-HET*, a transmission disequilibrium test that allows for locus heterogeneity, and is robust to population stratification³³. Results of the two stages in the 20p11.22 region (chr20: 21,815,192–21,988,830) were combined using a set association method as implemented in Sumstat³⁴ and TDT-HET³³. The set association method performs multi-locus association by estimating sum statistics for an increasing number of SNPs to evaluate their joint effect on disease. Sumstat and TDT-HET were used because to generate a single P -value calculated for the entire region in each study, thereby minimizing the number of tests performed. Previous work has shown that these methods have sufficient power to identify SNPs acting in additive and/or multiplicative manners to increase disease risk³⁵. The P -value obtained by applying either Sumstat or TDT-HET represents the global significance of the candidate chromosomal region. Combining the two stages in this way yielded increased evidence for association with IS in the 20p11.22 region, with a combined Fisher's $P = 1.33 \times 10^{-8}$ (Fig. 2 and Table 1) and ORs depicted in Supplementary Fig. 3.

Comparing our results to the National Human Genome Research Institute (NHGRI) GWAS catalogue²², we found that

the chromosome 20 IS locus was previously associated with early-onset male pattern baldness (AGA). Similar to IS, AGA displays sexual dimorphism, that is, it is biologically unequal in males and females. However, unlike IS, disease progression in AGA (extent of hair loss) is generally more severe in males than in females³⁶. We identified chromosome 20p11.22 SNPs that were previously associated with AGA and that were genotyped in our GWAS^{37–39}. In this comparison, SNPs that were associated with IS and AGA displayed the opposite direction of effect for the two disorders (Supplementary Table 1). This observation suggested that sequences in the region conferring susceptibility to IS have a protective effect in AGA. To test whether the association we observed was sex-specific, we re-evaluated association with SNPs in the 20p11.22 locus after stratification by sex, that is, separating males and females. This analysis yielded evidence for association with IS in females but not males, with a combined Fisher's $P = 6.88 \times 10^{-9}$ in the former data set (Table 1 and Supplementary Tables 2 and 3).

To confirm the chromosome 20p11.22 association with IS, we genotyped SNP rs6137473, as it was highly correlated with other top-associated SNPs in the region, in an independent cohort of 216 IS patients ascertained from various paediatric orthopaedic clinics throughout the United States of America and 336 population-matched controls. This variant was significantly associated with IS in females ($P = 2.4 \times 10^{-4}$; OR = 1.67, 95% CI = 1.27–2.21) but not males ($P = 0.726$; OR = 0.90, 95% CI = 0.49–1.64; Table 2 and Supplementary Table 4). Data provided from a GWAS of female Japanese subjects²⁴ also replicated association with the 20p11.22 region (rs6137473 $P = 3.7 \times 10^{-3}$; summary $P = 3 \times 10^{-3}$; Table 2 and Supplementary Table 5). Results from all four studies were combined and confirmed the association of rs6137473 with IS for females ($P = 2.15 \times 10^{-10}$, OR = 1.30, 95% CI = 1.19–1.41; Table 2).

Putative enhancers in the chr20p11.22 susceptibility locus. *PAX1* is a key regulator of sclerotome formation and vertebral development⁴⁰. Naturally occurring missense and deletion mutations in *Pax1* are well-described in the 'undulated' and

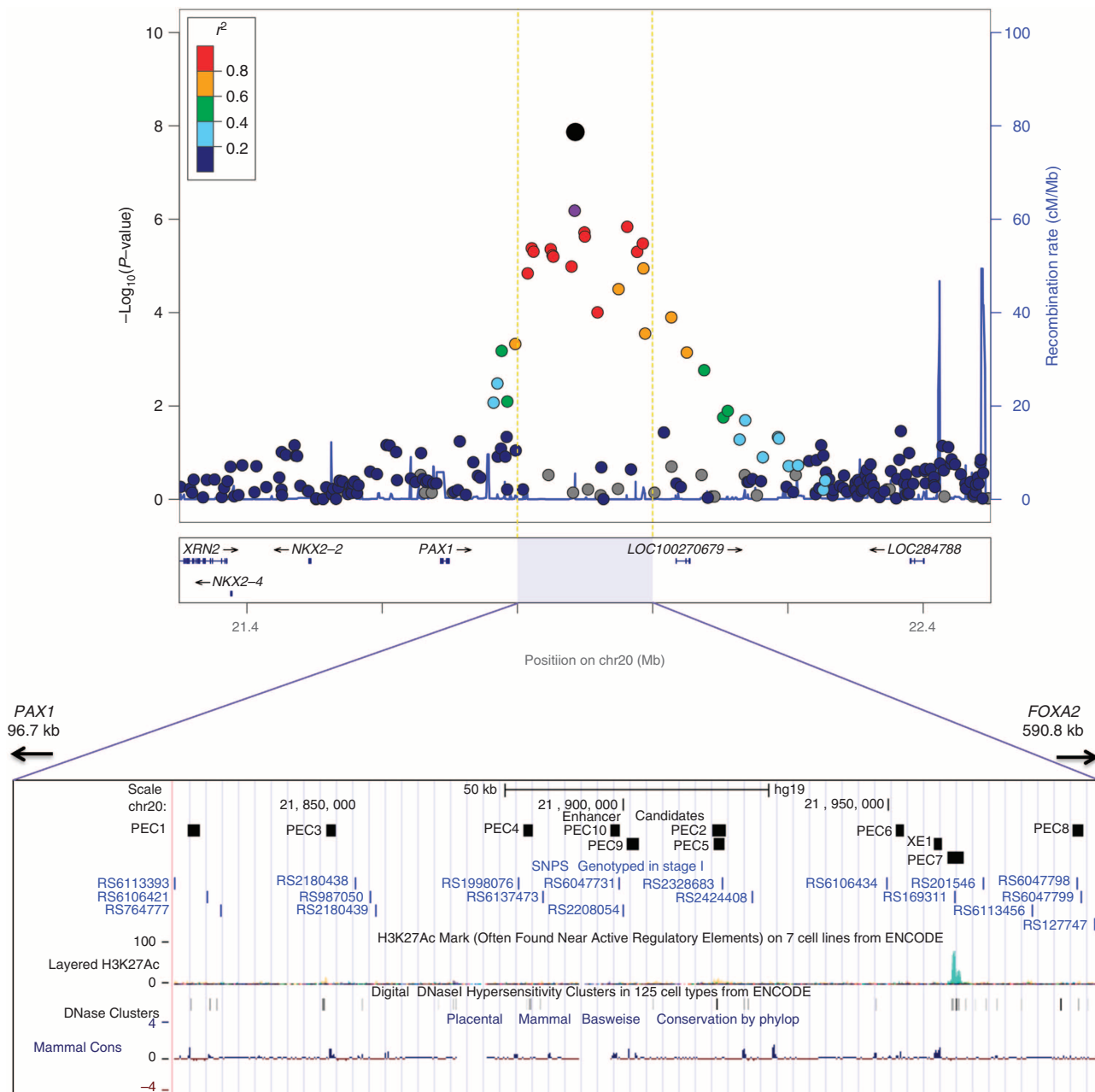


Figure 2 | Chromosome 20p11.22 regional association plot. $-\log_{10}(P\text{-value})$ of association with IS is plotted versus chromosomal position for each genotyped SNP for the second-stage GWAS (GWAS II). Each SNP is colour coded to reflect relative correlation (r^2 value) with major SNP rs6137473 (purple dot). A physical map of the associated region is shown below the x-axis. The summary statistic P -value for the combined TSRHC GWAS (stages I and II) was calculated for the interval encompassed by yellow dotted lines as described in the text. This P -value is depicted by the black circle. Detailed view (shown below) of sequence conservation in the region highlighted in purple. Custom UCSC genome browser view with hg19 chromosome 20 positions of PEC and Xel sequence blocks and genotyped SNPs is shown. In the layered H3K27ac ENCODE track, the green represents HSMM (human skeletal muscle myoblast) cells.

'scoli' mouse strains, so-called because of their distinct tail deformities and varying spinal malformations including scoliosis^{40–42}. The orthologous 20p11.22 IS-associated locus overlaps a regulator of murine *Pax1* expression in the developing spine (sclerotome) as identified by murine enhancer trap and reporter assays⁴³. In the latter study, deletions of the region including *Pax1* caused a corresponding reduction in gene expression in the sclerotome at embryonic day 11.5 (E11.5). Moreover, an ~1.5-kb sequence 'Xel' encoded in the deleted region recapitulated a *Pax1* expression pattern⁴³. We hypothesized that variants in the IS-associated region may

affect *PAX1* regulatory elements. Using comparative genomics and ENCODE data for biologically relevant cell lines, human skeletal muscle myoblast and human embryonic stem cells (H1-hESC)⁴⁴, we analysed the associated region (chr20:21,815,192–21,988,830; hg19) for sequences representing putative regulatory elements. We identified ten candidate regions (*PAX1* Enhancer Candidates 'PEC1', 'PEC2' and so on), including sequences orthologous to the previously described 'Xel' enhancer (Fig. 2 and Supplementary Table 6). To test the ability of each candidate to function as an enhancer, that is, to drive gene expression, we cloned each PEC including Xel into the E1b-GFP-

Table 1 | GWAS summary statistics.

| | Stage I | | | Stage II | | | Combined | |
|---------|--------------------|--------------------|------|------------------|---------------------|------|-----------------------|------|
| | N (cases/controls) | P-value | OR | Trios (subjects) | P-value | OR | P-value | OR |
| Total | 447/737 | 2×10^{-5} | 1.44 | 715 (1,876) | $3 \times 10^{-5*}$ | 1.23 | 1.33×10^{-8} | 1.31 |
| Females | 371/533 | 1×10^{-5} | 1.54 | 600 (1583) | $3 \times 10^{-5*}$ | 1.21 | 6.88×10^{-9} | 1.35 |
| Males | 76/204 | 0.631 | 1.08 | 115 (293) | 0.543 | 1.25 | 0.71 | 1.15 |

GWAS, genome-wide association study; OR, average odds ratio.

Summary statistics obtained using Sumstat³⁴ and TDT-HET³³ were applied to SNPs in the chromosome 20 region spanning genomic positions 21,815,192 to 21,988,830 in both data sets. Combined P-values were calculated Fisher's Combined P-value method⁶⁵.

*Actual permutation P-value was 0. The value in the table is the upper bound of the 95% confidence interval determined by the BINOM programme.

Table 2 | rs6137473 association results in all data sets.

| Study | Risk allele | Females | | Males | | Combined |
|-----------|-------------|------------------------|------------------|-------|------------------|-----------------------|
| | | P | OR (95% CI) | P | OR (95% CI) | P |
| Stage I | G | 2.36×10^{-7} | 1.64 (1.36–1.98) | 0.67 | 1.09 (0.75–1.58) | 5.58×10^{-7} |
| Stage II | G | 0.19 | 1.16 (0.93–1.45) | 0.45 | 1.21 (0.74–1.97) | 0.13 |
| TSRHC III | G | 2.40×10^{-4} | 1.67 (1.27–2.21) | 0.73 | 0.90 (0.49–1.64) | 1.50×10^{-3} |
| Japan | G | 3.70×10^{-3} | 1.18 (1.06–1.33) | NA | NA | NA |
| Combined | | 2.15×10^{-10} | 1.30 (1.19–1.41) | 0.81 | 1.08 (0.82–1.41) | 3.12×10^{-8} |

CI, confidence interval; NA, data not available; OR, odds ratio; TSRHC, Texas Scottish Rite Hospital for Children.

Results of allelic association tests performed using Cochran Armitage trend test are shown for the case-control studies (Stage I, TSRHC III and Japan). TDT-HET P-values are shown for the family-based study (Stage II).

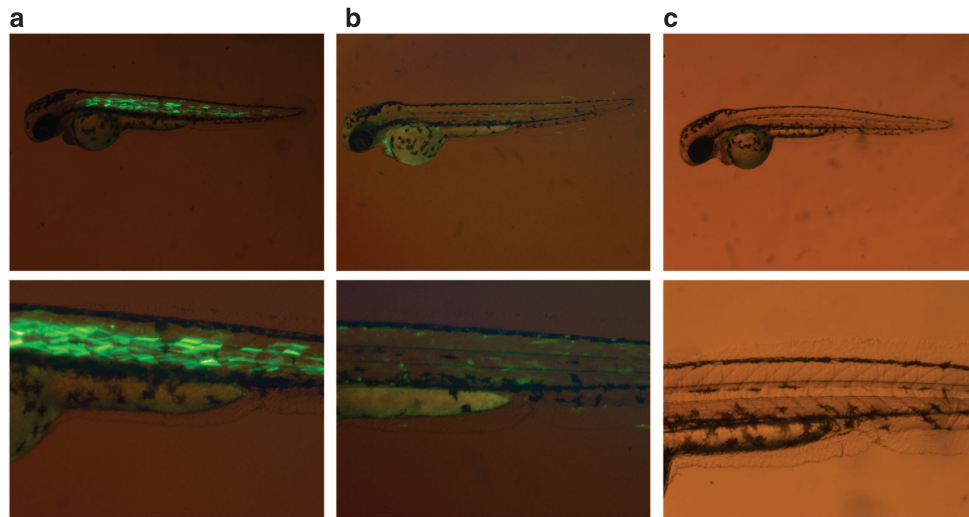


Figure 3 | Functional enhancer assays in zebrafish. (a) Representative zebrafish (of at least 100 tested for each construct) injected with Xe1-E1b-GFP-Tol2 enhancer construct showing expression predominantly in the somitic muscle. (b) Representative zebrafish injected with PE7-E1b-GFP-Tol2 enhancer construct showing expression predominantly in the somitic muscle. (c) Representative zebrafish injected with PEC7risk-E1b-GFP-Tol2 enhancer construct leads to loss of expression throughout. Corresponding pictures at higher magnification are shown below.

Tol2 enhancer assay vector and successfully tested seven of them in zebrafish as previously described⁴⁵. As expected, Xe1 showed enhancer activity in the developing spine, specifically corresponding to somitic muscle (Fig. 3a). One other candidate, PEC7, also displayed functional enhancer activity. Similar to Xe1, PEC7-driven expression of the green fluorescent protein (GFP) reporter was largely restricted to somitic muscle, with weaker expression in spinal cord and heart (Fig. 3b). Thus, Xe1 and PEC7 clearly harboured enhancer activity, possibly in developing spinal muscle, although more detailed studies are needed to precisely define the temporo-spatial effects of these enhancers on gene expression in humans. It is worth noting that PEC7 enhancer expression did not completely overlap the characterized *pax1b* expression in zebrafish at 48 h post-fertilization (h.p.f.)⁴⁶. This

could be due to the testing of human sequences in zebrafish. However, previous work has shown that human enhancer sequences can function as active enhancers in zebrafish, even without homologous sequences in zebrafish^{47–50}. In addition, the expression pattern of *pax1a* has yet to be determined in zebrafish, and we cannot exclude the possibility that PEC7 controls the expression of other genes.

Disruption of enhancer activity by IS-associated SNPs. To search for disease alleles potentially underlying the association with IS, the Xe1 and PEC7 regions were re-sequenced in 48 cases that were enriched for chromosome 20 risk alleles as identified in the original GWAS. Although no variants were detected in Xe1 in these individuals, a haplotype of five variants including top-

associated SNP rs169311 was identified in PEC7 (Supplementary Table 7). Using Haploreg⁵¹, we noted that top SNP rs169311 is predicted to alter binding sites for component of myogenesis protein 1 and vitamin D receptor, and three other PEC7 SNPs are also predicted to alter transcription factor-binding sites (Supplementary Table 7). We compared the ability of the risk haplotype to drive zebrafish reporter gene expression to that of the wild-type sequence. Four independent replicate experiments confirmed that the associated haplotype completely abolished enhancer activity as detected by this assay, suggesting that the IS susceptibility haplotype confers a loss of function for PEC7 enhancer activity (Fig. 3c and Supplementary Data set 1).

PAX1 expression in spinal myofibers post-somitogenesis.

In the developing mouse embryo, Pax1 expression is well-described in somitogenesis, beginning at E8.5 in the stage III somite and continuing to E12.5, becoming restricted to cells surrounding the vertebrae, intervertebral disc anlagen and precursors to the connective tissue around spinal nerve and dorsal root ganglia^{40,41}. As vertebral structures and segmentation appear normal in IS, we hypothesized a role for PAX1 in spinal development post-somitogenesis. Accordingly, we examined Pax1 protein in mouse spinal tissues at seven developmental time points from E13.5 to postnatal day 84 (P84). As shown in Fig. 4a–c, Pax1 immunohistochemical staining was essentially negative post-somitogenesis at E13.5. However, at E16.5, we

observed a striking pattern in developing myofibers, with more modest staining in other cells types. Weaker but persistent staining was evident in myofibers up to stage P84. Pax1 was essentially absent in other spinal cell types (Supplementary Fig. 4). These data demonstrate a potential role for Pax1 in spine development post-somitogenesis, possibly in paraspinous muscles.

Discussion

We present evidence of a new IS susceptibility locus in an ~100-kb region of chromosome 20p11.22 downstream of *PAX1*. Using a functional fine-mapping approach, we potentially narrow the locus to an ~1.5-kb domain with enhancer activity that is disrupted by disease-associated variants. The *PAX1*-encoding region was originally associated with spinal development through studies of the naturally occurring *undulated* mouse strains. The original *undulated* (*un*) strain, first described in 1947, carries a missense mutation in *Pax1* (ref. 41). *Un/un* mice display a curved spine with malformations of individual vertebrae including the vertebral bodies and intervertebral discs. Three additional strains, *scoliosis* (*sco*) or *undulated intermediate* (*un-i*), *undulated-extensive* (*un^{ex}*), *undulated short-tail* (*un^s*), harbour partial or complete deletions of *PAX1*, with the latter including all of the gene and displaying the most severe phenotype^{41,42}. In early mouse development, Pax1 displays expression restricted to specific structures including the sclerotome that will give rise to the axial spine (vertebrae, ribs, connective tissues and skin).

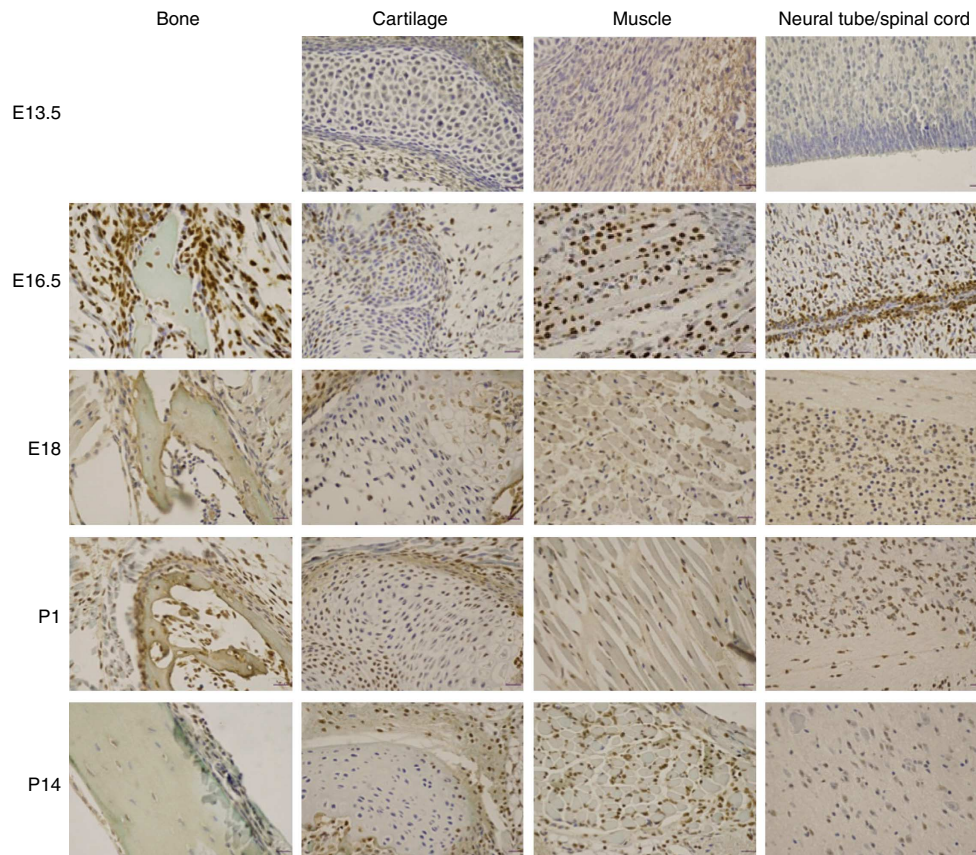


Figure 4 | Pax1 immunohistochemistry in developing mouse spinal tissues. Spines and surrounding soft tissue harvested at E13.5, E16.5, E18, P1 and P14 are shown in each row. Representative images for bone, cartilage, muscle and nerve are shown in the columns, with the exception of bone at E13.5, as it was not well-formed at that time point. Muscle was also not well-formed at E13.5 and had a soft tissue appearance. In contrast, strong nuclear staining (brown) was clearly seen in osteoblasts, myoblasts and ependymal cells at E16.5, with weaker chondrocyte staining. More modest staining was evident at E18 and P1 in all tissues. By P14, Pax1 staining persisted in muscle but was essentially absent from bone and cartilage (note positive staining in muscle surrounding cartilage and adjacent to bone). Very weak staining was observed in a subset of spinal cord cells, possibly glia, at P14. Each experimental time point was repeated at least once. Scale bar, 20 μ M.

Genomic studies have delineated intervals downstream of *Pax1* harbouring *cis*-regulatory activity consistent with this pattern⁴³. In particular, transposon-based deletion mapping and reporter gene assays defined the ~148-kb region 3' of *Pax1* as necessary to drive somitic gene expression (that is, in the dorsal sclerotome) during early mouse development. Furthermore, the mouse *Xe1* enhancer encoded in this region was shown to be sufficient to drive a similar expression pattern⁴³. Our data using zebrafish transgene assays confirmed the enhancer activity of the human *Xe1* orthologue and revealed another element in the region, *PEC7*, with potential somitic enhancer activity that was disrupted by IS-associated sequence variants. This observation strongly suggests that *PEC7* itself functions in IS susceptibility, a hypothesis that may be tested in model systems by targeted mutagenesis.

The spinal anomalies observed in *Pax1* mouse mutants recapitulates a congenital scoliosis phenotype, although mutations in human *PAX1* are not clearly correlated with this condition⁵². By contrast, the structure of the axial spine in IS appears normal. Consequently, we reasoned that either the association we observed reflects functional effects on a gene other than *PAX1*, or an effect on the participation of *PAX1* in a post-somitogenesis developmental role. In support of the latter hypothesis, we found that *Pax1* is strongly expressed in developing myofibers of the mouse spine at day E16.5, well after the somites are formed, and that expression persisted at least to early adulthood, albeit at reduced levels. It is intriguing to consider that *PAX1* may participate in multiple roles in spinal development that are dictated by the spatio-temporal control of its expression. Such a scenario predicts that deleterious mutations that occur in *PAX1*-coding sequences, versus those that occur in its regulatory elements, will give rise to distinct but potentially overlapping phenotypes. This phenomenon is well-described for other early developmental genes. For example, coding mutations in the sonic Hedgehog gene produce the severe multi-system disorder holoprosencephaly, whereas mutations in the sonic Hedgehog *cis*-regulator *ZRS* cause limb-specific pre-axial polydactyly^{53,54}. Whether congenital scoliosis and IS may be aetiologically connected in humans is unclear but has been suggested from familial aggregation studies^{55–57}. An aetiological relationship between the two diagnoses is also supported by the recent discovery that *ptk7* mutant zebrafish model both IS- and CS-like phenotypes in zygotic and maternal-zygotic genetic backgrounds, respectively⁸. The latter study also illustrates the distinct morphologies that may arise from the differential expression of key developmental genes.

Our investigation of the chromosome 20p11 locus provides the first genetic evidence to explain the puzzling sexual dimorphism that is a hallmark of IS. Besides susceptibility to progression, the pattern, onset and flexibility of deformity also differ between boys and girls¹⁰. Various hypotheses have been proposed to explain male/female differences in IS, including the existence of X-linked genetic risk factors and effects on circulating hormones. Neither mechanism has been clearly supported, although investigations have been limited^{6,58}. Our identification of a female-specific IS susceptibility locus suggests an underlying mechanism that is sensitive to the female milieu at the time of adolescence. Although we did not find evidence for oestrogen receptor-binding sites within the *PEC7* enhancer locus itself, it is interesting to postulate that this locus increases risk of IS via downstream hormonal interactions. We note in this regard that the next-nearest gene, *FOXA2*, is implicated in sexually dimorphic gene expression via cooperation with androgen and oestrogen receptor⁵⁹. It is possible that *PEC7* regulates *FOXA2*. However, we did not detect *Foxa2* expression in embryonic or postnatal mouse spine (data not shown) and consider it an unlikely candidate for IS

susceptibility. *PAX1* is also expressed in the adult scalp³⁷. Whether variants in *PEC7* affect this expression and drive association with early-onset male pattern baldness requires further study, but the overlapping genetic association suggests a possible correlation between the two sexually dimorphic conditions.

We provide evidence that *PAX1* is expressed in paraspinous muscles, and to a lesser extent in spinal cord, at time points well after the initial patterning of the axial skeleton has been completed. This suggests a later developmental role for *PAX1*, possibly in the growth or maintenance of these tissues. It is interesting that other IS risk loci identified by GWAS occur in non-coding regions near or within genes involved in muscle and nerve biogenesis. In this regard, we expect that the functional fine-mapping method applied in the present study should prove fruitful for identifying the alleles and functional elements driving association at other IS risk loci. Our results also imply that additional sex-specific genetic loci may await discovery, and draw attention to the need to consider males and females in separate liability classes.

Methods

Study subjects. All research subjects included in GWAS I-715, GWAS II and TSRHC III provided written informed consent to participate in the study as approved by the Institutional Review Board of the University of Texas Southwestern Medical Center. For GWAS I and II, patients or former patients were sequentially ascertained in Orthopedic Clinics at TSRHC and enrolled into an idiopathic scoliosis (IS) registry. Parents and other affected family members were also ascertained when possible. For TSRHC III, 216 NHW cases were included from cases ascertained in TSRHC clinics or through collaborating orthopaedic surgeons in the United States of America at Shriners Hospital for Children, Lexington, KY (T. Milbrandt, V. Talwalkar, H.J. Iwinski, R. Muchow); Hasbro Children's Hospital, Providence, RI (C.P. Ebersson); University of Massachusetts Memorial Medical Center, Worcester, MA (A. Lapinsky); Children's Hospital of Wisconsin, Milwaukee, WI (J.C. Tassone, X.C. Liu) and Akron Children's Hospital, Akron, OH (W. Schrader); OrthoArizona Phoenix, AZ (R. Shindell); Indiana University-Purdue University Indianapolis, Indianapolis, IN, (R. Loder); University of Oklahoma Health Sciences Center, Oklahoma City, OK (J. Davey). All affected subjects in these cohorts met criteria for a positive diagnosis of IS: lateral deviation from the midline greater than 15° as measured by the Cobb angle method from standing spinal radiographs, axial rotation towards the side of the deviation and exclusion of relevant co-existing diagnoses. Blood samples were obtained by venipuncture. In some cases, saliva samples were self-collected using the Oragene DNA kit (DNA Genotek, Inc.). Control individuals were ascertained from within the local Texas population or non-orthopaedic clinics at TSRHC. A diagnosis of scoliosis, or family history of scoliosis, was excluded by questionnaire. To reduce the possibility of biases due to population stratification in the TSRHC III cohort, we previously genotyped 384 ancestry informative markers and performed multi-dimensional scaling analysis of identity-by-state distances to identify outliers²³.

The Japanese cohort consisted of 1,050 Japanese females with adolescent IS recruited from eight collaborating hospitals (Japan Scoliosis Clinical Research Group: JSCRG) between February 2009 and January 2011. All subjects with adolescent IS underwent clinical and radiologic examinations for IS as previously described²⁴. Subjects with definitive family history of Mendelian inheritance were excluded from the study. The control subjects consisted of 1,474 Japanese females, including healthy volunteers from the Midosuji Rotary Club, Osaka, Japan, and individuals who were registered in the BioBank Japan Project but were genotyped in GWAS of other diseases, as previously described⁶⁰. Informed consent was obtained from all Japanese subjects and from parents of minor subjects according to a protocol approved by the ethics committees of the University of Texas Southwestern Medical Center, University of Kentucky, University of Massachusetts Medical School, Children's Hospital of Wisconsin, Akron Children's Hospital, Indiana University-Purdue University, Oklahoma University Health Sciences Center and RIKEN.

GWAS genotyping, quality control, imputation and association testing. The GWAS II data set was derived from 1,201 individuals (457 cases and 744 controls). We chose the sample sizes to have 80% power to detect genetic association with population data (cases and controls) at the 5E-08 significance level (that is, 0.05 genome-wide significance) using the following parameters: genetic model-based design, disease prevalence of 0.03, heterozygote relative risk of 1.5, homozygote relative risk of 2.25, disease allele frequency of 0.50, linear trend test statistic with additive weights (0, 1, 2 for the three possible genotypes at a SNP). Using these specifications, we compute a minimum sample size of 932 cases and 932 controls. The combined Stage I and Stage II cohorts therefore exceeded the sample sizes

required for our specified power. DNA samples from cases and controls were distributed non-sequentially into plates and genotyped using the Illumina HumanOmniExpress-12 v1.0 beadchip (Illumina, Inc.) containing 730,498 markers. Twenty-nine technical replicates were included and were 100% concordant. We used PLINK v1.07 (ref. 32) to perform QA on these data. A total of 17 individuals were removed from the analyses. Fifteen had genotyping call rates <95%. Two individuals had ambiguous gender information according to the X-chromosome inbreeding coefficient⁶¹. SNP markers were evaluated for the presence of heterozygous haploid (HH) genotypes and missingness. All markers on chromosome Y with HH genotypes were removed. Markers on chromosome X with HH genotypes were kept. Markers with more than 5% of missing data were removed. A total of 2,828 markers (0.39%) failed QA. After all QA filters were applied, a total of 1,184 individuals (447 cases and 737 controls) and 727,670 markers were included in the analyses (GWAS II). In addition, we checked for population stratification confounder effects by applying the genomic control (GC) method⁶². We obtained a correction factor (λ) of 1.0992, potentially indicating a slight inflation of the GWAS results due to population stratification. To study the region in more detail, we used microarray genotypes to impute 1000 Genomes (<http://www.1000genomes.org>) chromosome 20 variants with MACH⁶³ (<http://www.sph.umich.edu/csg/abecasis/MACH/tour/imputation.html>). Genotypes were tested for association using CATT as implemented in the PLINK software³². Genotyping and quality control for GWAS I-715 were as previously described^{23,24}. Nine individual samples were removed with genotyping call rates <95%. Of 345,111 SNPs, 1,514 (0.44%) had call rates less than 95% and were removed, leaving 343,597 (99.56%) of SNPs. SNP markers corresponding to X and Y chromosomes were also removed from the analysis. The Japanese GWAS was analysed as described previously^{24,25}.

Meta-analysis of Stages I and II, TSRHC III and Japan sets assumed a fixed effect model and was carried out using the inverse variance weighting method. Where applicable, meta-analyses of case-control and trio designs were performed as described in Kazeem and Farral⁶⁴.

Set association method. We applied the set association method, as implemented in the Sumstat³⁴ and TDT-HET³³ software, to the region on chromosome 20 spanning 21,815,192–21,988,830 corresponding to a linkage disequilibrium (LD) block harbouring top-associated SNPs from the second stage GWAS. We used this method to obtain a single global *P*-value for the chromosomal region in each data set (Stage I, Stage II and Japan). Our approach is similar to the analysis of variance approach for testing equality of means across multiple categories. In this analogy, we consider SNPs as categories. We applied TDT-HET to pedigree data from Stage II and Sumstat to case-control data from Stage I and the Japan data set. We obtained three *P*-values this way, each corresponding to a single data set (Stage I, Stage II and Japan). Because each data set was independent, we used Fisher's Combined *P*-value method⁶⁵ to compute a single *P*-value for the chromosomal region.

Association analyses were performed using 100,000 permutations. None of the markers in the selected regions failed either the Hardy-Weinberg proportions test in the unaffected sample or presented significant differences in genotype calling between cases and controls at the 0.1% level. In Supplementary Table 3, the actual permutation *P*-value for the chromosome 20 region in GWAS I-702 (Stage II) was 0. We used the value $P = 3 \times 10^{-5}$ since it is the upper limit of the 95% confidence interval, as determined by the method implemented in the BINOM programme⁶⁶. An interesting result from our association analyses is that we observed a *P*-value of 0.19 in Stage II pedigree data (affected subjects restricted to females) for the single marker rs6137473. When we applied the TDT-HET method to the markers in the candidate region for the same pedigrees, our multi-locus *P*-value became highly significant. In addition, *P*-values obtained with female population data are also highly significant for the same marker. We conjecture that this result is due to differences in sample size, especially given that, for family-based association, trios must contain at least one heterozygous parent to be informative. When we checked the number of trios with female-only affected subjects for marker rs6137473, we computed a total of 201 informative trios out of the total 600. We performed comparative power calculations for case-control females, and for trios. We used parameter settings from both Table 1 and Table 2. Using these values, we computed a power value of 70% for the case-control design using the linear trend test (weights equal to number of disease alleles in respective genotype) at a 10^{-5} significance level. On the other hand, we computed a power of 12% for the TDT design at the same significance level. This decrease in power is substantial, and provides one possible explanation as to why we did not observe a clear replication with the TDT statistic at the single-locus level.

TSRHC III genotyping. We genotyped SNP rs6137473 in the 216 unrelated cases and 336 unrelated controls using Taqman genotyping. Results were in agreement with Hardy-Weinberg equilibrium ($P > 0.05$). We performed allelic tests of association and CATT as implemented in the PLINK software³².

Candidate enhancer identification. Using the UCSC Genome Browser (<https://genome.ucsc.edu/cgi-bin/hgGateway>), PhastCons (<http://compugen.bscb.cornell.edu/phast/help-pages/phastCons.txt>) conserved elements and PhyloP (<http://compugen.bscb.cornell.edu/phast/help-pages/phyloP.txt>) scores, we analysed

chr20:21,815,192–21,988,830 (hg19) for evolutionary conserved sequences. In addition, we used enhancer mark data (H3K27ac and H3K4me1) from the ENCODE project for human skeletal muscle myoblast cells and human embryonic stem cells (H1-hESC), as well as DNase hypersensitivity data to identify putative regulatory elements within the chromosome 20 region. These two methods of analysis led to the selection of ten regions that were conserved between human and other placental mammals or were predicted regulatory elements based on ENCODE data.

Re-sequencing study. To identify individuals for re-sequencing we used the top Stage I CATT-associated SNP markers on chromosome 20 (rs6137473, rs6106434, rs169311). Two of these SNPs (rs6106434 and rs169311) flank the Xel1 enhancer region. We determined all possible genotype patterns of the three markers. To prioritize cases and controls for re-sequencing, we selected the multi-locus genotype pattern (MLGP) that (i) showed the largest difference in frequency between cases and controls, (ii) had at least five samples in cases and controls and (iii) for which the frequency in cases was greater than in controls. The MLGP that met these criteria was GG/AA/AA, corresponding to the homozygous minor alleles at rs6137473, rs6106434 and rs169311, respectively. Forty-eight NHW IS cases with this MLGP were selected for re-sequencing. PCR primers were designed to flank PEC7 and Xel1 regions. PCR fragments were sequenced using the Sanger method.

Zebrafish assays. Male and female strain AB zebrafish were used in all experiments. Enhancer candidate sequences were cloned by PCR from human genomic DNA (Roche) into the E1b-GFP-Tol2 enhancer assay vector containing an E1b minimal promoter followed by GFP⁶⁷ and verified by sequencing. For the PEC7 risk haplotype, PCR was carried out on an individual encompassing this haplotype and sequence verified. Each construct was injected at two different days and Xel1, PEC7-reference and PEC7 risk haplotype at four different days following standard procedures^{68,69} into at least 100 zebrafish embryos along with Tol2 mRNA⁷⁰ to facilitate genomic integration. GFP expression was observed and annotated up to 48 h.p.f using a Leica M165 FC microscope. An enhancer was considered positive if at least 10% of all fish surviving to 48 h.p.f. showed a consistent expression pattern after subtracting out percentages of tissue expression in fish injected with the empty enhancer vector. For each construct, at least 100 fishes were analysed for GFP expression up to 48 h.p.f. All animal work was approved by the University of California San Francisco Institutional Animal Care and Use Committee.

Pax1 immunohistochemistry. All animal work was approved by the University of Texas Southwestern Medical Center Institutional Animal Care and Use Committee. Male and female *BL6* mice were euthanized at different developmental stages (E13.5, E16, E18, P1, P14, P28, P84) by asphyxiation with CO₂. Spines were excised and fixed with 10% buffered formalin for 3 days and then processed in a Leica ASP300S tissue processor and embedded in paraffin. Four-micrometre sections were cut in series in the frontal plane and loaded onto 3-aminopropyl-triethoxysilane-coated slides. Antigen retrieval was performed with 0.25% Trypsin (Sigma-Aldrich) and endogenous peroxidase was blocked with 3% H₂O₂ (Thermo Fisher Scientific Inc.). Slides were incubated with rabbit polyclonal antibody (1:15,000, Sigma-Aldrich catalogue number SAB2101727) directed against a PAX1-specific peptide and detected by goat anti-rabbit IgG-HRP (1:500, EMD Millipore Corporation) and 3,3'-diaminobenzidine (Dako) as chromogenic substrate. After Pax1 staining, sections were counterstained with haematoxylin and Fast Green (Sigma-Aldrich) and mounted with Cytoseal XYL (Thermo Fisher Scientific Inc.). Haematoxylin and eosin staining were also performed separately for pathological characterization. Slides were viewed with an Olympus BX40 microscope.

References

- Hresko, M. T. Clinical practice. Idiopathic scoliosis in adolescents. *New Engl. J. Med.* **368**, 834–841 (2013).
- Herring, J. A. *Tachdjian's Pediatric Orthopaedics* (WB Saunders, 2008).
- Little, D. G., Song, K. M., Katz, D. & Herring, J. A. Relationship of peak height velocity to other maturity indicators in idiopathic scoliosis in girls. *J. Bone Joint Surg. Am.* **82**, 685–693 (2000).
- Richards, B. S. & Vitale, M. G. Screening for idiopathic scoliosis in adolescents. An information statement. *J. Bone Joint Surg. Am.* **90**, 195–198 (2008).
- Sharma, S. & Wise, C. in *The Genetics and Development of Scoliosis* (ed. Dunwoodie, K. A.) 167–190 (Springer, 2010).
- Lowe, T. G. *et al.* Etiology of idiopathic scoliosis: current trends in research. *J. Bone Joint Surg. Am.* **82-A**, 1157–1168 (2000).
- Gorman, K. F., Tredwell, S. J. & Breden, F. The mutant guppy syndrome curveback as a model for human heritable spinal curvature. *Spine (Phila Pa 1976)* **32**, 735–741 (2007).
- Hayes, M. *et al.* ptk7 mutant zebrafish models of congenital and idiopathic scoliosis implicate dysregulated Wnt signalling in disease. *Nat. Commun.* **5**, 4777 (2014).
- Buchan, J. G. *et al.* Kinesin family member 6 (kif6) is necessary for spine development in zebrafish. *Dev. Dyn.* **243**, 1646–1657 (2014).

10. Raggio, C. L. Sexual dimorphism in adolescent idiopathic scoliosis. *Orthop. Clin. North Am.* **37**, 555–558 (2006).
11. Karol, L. A., Johnston, 2nd C. E., Browne, R. H. & Madison, M. Progression of the curve in boys who have idiopathic scoliosis. *J. Bone Joint Surg.* **75-A**, 1804–1810 (1993).
12. Kruse, L. M., Buchan, J. G., Gurnett, C. A. & Dobbs, M. B. Polygenic threshold model with sex dimorphism in adolescent idiopathic scoliosis: the Carter effect. *J. Bone Joint Surg. Am.* **94**, 1485–1491 (2012).
13. Tang, N. L. *et al.* Genetic epidemiology and heritability of AIS: A study of 415 Chinese female patients. *J. Orthop. Res.* **30**, 1464–1469 (2012).
14. Ward, K. *et al.* Polygenic inheritance of adolescent idiopathic scoliosis: a study of extended families in Utah. *Am. J. Med. Genet. A.* **152A**, 1178–1188 (2010).
15. Weinstein, S. L. Natural history. *Spine (Phila Pa 1976)* **24**, 2592–2600 (1999).
16. Chan, V. *et al.* A genetic locus for adolescent idiopathic scoliosis linked to chromosome 19p13.3. *Am. J. Hum. Genet.* **71**, 401–406 (2002).
17. Salehi, L. B. *et al.* Assignment of a locus for autosomal dominant idiopathic scoliosis (IS) to human chromosome 17p11. *Hum. Genet.* **111**, 401–404 (2002).
18. Ocaka, L. *et al.* Assignment of two loci for autosomal dominant adolescent idiopathic scoliosis to chromosomes 9q31.2-q34.2 and 17q25.3-qtel. *J. Med. Genet.* **45**, 87–92 (2008).
19. Gao, X. *et al.* CHD7 gene polymorphisms are associated with susceptibility to idiopathic scoliosis. *Am. J. Hum. Genet.* **80**, 957–965 (2007).
20. Altschuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
21. Antonarakis, S. E., Chakravarti, A., Cohen, J. C. & Hardy, J. Mendelian disorders and multifactorial traits: the big divide or one for all? *Nat. Rev. Genet.* **11**, 380–384 (2010).
22. Hindorf, L. A., J.H., Mehta, J. P. & Manolio, T. A. A Catalog of Published Genome-Wide Association Studies. Available at <http://www.genome.gov/gwastudies>. Accessed September 18, 2012.
23. Sharma, S. *et al.* Genome-wide association studies of adolescent idiopathic scoliosis suggest candidate susceptibility genes. *Hum. Mol. Genet.* **20**, 1456–1466 (2011).
24. Takahashi, Y. *et al.* A genome-wide association study identifies common variants near LBX1 associated with adolescent idiopathic scoliosis. *Nature Genet.* **43**, 1237–1240 (2011).
25. Kou, I. *et al.* Genetic variants in GPR126 are associated with adolescent idiopathic scoliosis. *Nat. Genet.* **45**, 676–679 (2013).
26. Londono, D. *et al.* A meta-analysis identifies adolescent idiopathic scoliosis association with LBX1 locus in multiple ethnic groups. *J. Med. Genet.* **51**, 401–406 (2014).
27. Brohmann, H., Jagla, K. & Birchmeier, C. The role of Lbx1 in migration of muscle precursor cells. *Development* **127**, 437–445 (2000).
28. Muller, T. *et al.* The homeodomain factor lbx1 distinguishes two major programs of neuronal differentiation in the dorsal spinal cord. *Neuron* **34**, 551–562 (2002).
29. Matise, M. A dorsal elaboration in the spinal cord. *Neuron* **34**, 491–493 (2002).
30. Monk, K. R., Oshima, K., Jors, S., Heller, S. & Talbot, W. S. Gpr126 is essential for peripheral nerve development and myelination in mammals. *Development* **138**, 2673–2680 (2011).
31. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
32. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
33. Londono, D. *et al.* TDT-HET: a new transmission disequilibrium test that incorporates locus heterogeneity into the analysis of family-based association data. *BMC Bioinformatics* **13**, 13 (2012).
34. Hoh, J., Wille, A. & Ott, J. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res.* **11**, 2115–2119 (2001).
35. Hoh, J. & Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.* **4**, 701–709 (2003).
36. Price, V. H. Androgenetic alopecia in women. *J. Investig. Dermatol. Symp. Proc.* **8**, 24–27 (2003).
37. Hillmer, A. M. *et al.* Susceptibility variants for male-pattern baldness on chromosome 20p11. *Nature Genet.* **40**, 1279–1281 (2008).
38. Richards, J. B. *et al.* Male-pattern baldness susceptibility locus at 20p11. *Nature Genet.* **40**, 1282–1284 (2008).
39. Li, R. *et al.* Six novel susceptibility Loci for early-onset androgenetic alopecia and their unexpected association with common diseases. *PLoS Genet.* **8**, e1002746 (2012).
40. Monsoro-Burq, A. H. Sclerotome development and morphogenesis: when experimental embryology meets genetics. *Int. J. Dev. Biol.* **49**, 301–308 (2005).
41. Wallin, J. *et al.* The role of Pax-1 in axial skeleton development. *Development* **120**, 1109–1121 (1994).
42. Adham, I. M. *et al.* The scoliosis (sco) mouse: a new allele of Pax1. *Cytogenet. Genome Res.* **111**, 16–26 (2005).
43. Kokubu, C. *et al.* A transposon-based chromosomal engineering method to survey a large cis-regulatory landscape in mice. *Nat. Genet.* **41**, 946–952 (2009).
44. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
45. Oksenberg, N., Stevison, L., Wall, J. D. & Ahituv, N. Function and regulation of AUTS2, a gene implicated in autism and human evolution. *PLoS Genet.* **9**, e1003221 (2013).
46. Liu, X., Wang, H., Li, G., Huang, H. Z. & Wang, Y. Q. The function of DrPax1b gene in the embryonic development of zebrafish. *Genes Genet. Syst.* **88**, 261–269 (2013).
47. Booker, B. M., Murphy, K. K. & Ahituv, N. Functional analysis of limb enhancers in the developing fin. *Dev. Genes Evol.* **223**, 395–399 (2013).
48. Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L. & McCallion, A. S. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312**, 276–279 (2006).
49. Navratilova, P. *et al.* Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev. Biol.* **327**, 526–540 (2009).
50. McGaughey, D. M. *et al.* Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. *Genome Res.* **18**, 252–260 (2008).
51. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
52. Giampietro, P. F. *et al.* An analysis of PAX1 in the development of vertebral malformations. *Clin. Genet.* **68**, 448–453 (2005).
53. Belloni, E. *et al.* Identification of Sonic hedgehog as a candidate gene responsible for holoprosencephaly. *Nat. Genet.* **14**, 353–356 (1996).
54. Laurell, T. *et al.* A novel 13 base pair insertion in the sonic hedgehog ZRS limb enhancer (ZRS/LMBR1) causes preaxial polydactyly with triphalangeal thumb. *Hum. Mutat.* **33**, 1063–1066 (2012).
55. Purkiss, S. B., Driscoll, B., Cole, W. G. & Alman, B. Idiopathic scoliosis in families of children with congenital scoliosis. *Clin. Orthop. Relat. Res.* **401**, 27–31 (2002).
56. Bulman, M. P. *et al.* Mutations in the human delta homologue, DLL3, cause axial skeletal defects in spondylocostal dysostosis. *Nat. Genet.* **24**, 438–441 (2000).
57. Maisenbacher, M. K. *et al.* Molecular analysis of congenital scoliosis: a candidate gene approach. *Hum. Genet.* **116**, 416–419 (2005).
58. Justice, C. M., Miller, N. H., Marosy, B., Zhang, J. & Wilson, A. F. Familial idiopathic scoliosis: evidence of an X-linked susceptibility locus. *Spine (Phila Pa 1976)* **28**, 589–594 (2003).
59. Li, Z., Tuteja, G., Schug, J. & Kaestner, K. H. Foxa1 and Foxa2 are essential for sexual dimorphism in liver cancer. *Cell* **148**, 72–83 (2012).
60. Cha, P. C. *et al.* A genome-wide association study identifies three loci associated with susceptibility to uterine fibroids. *Nat. Genet.* **43**, 447–450 (2011).
61. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics* (Longman Group Ltd., 1996).
62. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
63. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
64. Kazeem, G. R. & Farrall, M. Integrating case-control and TDT studies. *Ann. Hum. Genet.* **69**, 329–335 (2005).
65. Fisher, R. A. *Statistical Methods for Research Workers* 239 (Oliver and Boyd, 1925).
66. Ott, J. *Analysis of Human Genetic Linkage* (Johns Hopkins, 1999).
67. Li, Q. *et al.* A systematic approach to identify functional motifs within vertebrate developmental enhancers. *Dev. Biol.* **337**, 484–495 (2010).
68. Westerfield, M. *The Zebrafish Book* (Univ. Oregon, 2007).
69. Nusslein-Volhard, C. A. R. *Zebrafish* (Oxford Univ., 2002).
70. Kawakami, K. Transposon tools and methods in zebrafish. *Dev. Dyn.* **234**, 244–254 (2005).

Acknowledgements

We thank the patients, their families, other study subjects, and referring physicians for their participation in this research. We also thank J. Rios for critical review of the manuscript, and J. Brandon for administrative assistance. This study was supported by the National Institutes of Health (NICHD award R01HD052973), the TSRHC Research Fund project 867, Crystal Charity Ball, Scoliosis Research Society and the Cain Foundation (to C.A.W.), Scoliosis Research Society (to N.K.) and by NICHD award R01HD059862 and NHGRI awards R01HG005058 and R01HG006768 (to N.A.). N.A. is also supported in part by NIGMS award number GM61390, NIDDK award number R01DK090382 and NINDS award number R01NS079231.

Author contributions

S.S. and X.G. performed GWAS I-715 and GWAS II and S.S. analysed the results. S.S. performed and analysed the TSRHC III genotyping. D.L. and D.G. performed statistical analyses of GWAS I-715, GWAS II and all combined results and advised the statistical design of the project. J.A.H. provided clinical evaluation for GWAS I-715, GWAS II and TSRHC III and advised the clinical design of the project. A.T. and M.M. provided clinical evaluation for the Japan GWAS. I.K. performed statistical analyses of the Japan GWAS. S.I. designed the Japan GWAS and evaluated the results. N.K. and R.C. provided mouse tissue slides, D.Z. performed the immunohistochemistry analysis, and D.B. and D.Z. interpreted the results. K.M. and D.Z. re-sequenced candidate enhancers (PECs) and cloned into reporter constructs. W.L.E. and N.A. identified candidate enhancers *in silico* and designed zebrafish reporter assays; W.L.E. and K.K.M. performed reporter assays and characterized results, W.L.E. and N.A. interpreted results. Members of the TSRHC Scoliosis Clinical Group and Japanese Scoliosis Clinical Research Group identified patients and provided samples for the study. C.A.W. designed the study, interpreted results and drafted the manuscript.

Additional information

Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunications>.

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions>.

How to cite this article: Sharma, S. *et al.* A *PAX1* enhancer locus is associated with susceptibility to idiopathic scoliosis in females. *Nat. Commun.* 6:6452 doi: 10.1038/ncomms7452 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

TSRHC Scoliosis Clinical Group

L. Karol¹², K. Rathjen¹², D. Sucato¹², J. Birch¹², C. Johnston III¹², B. S. Richards¹², T. Milbrandt¹³, V. Talwakar¹³, H. Iwinski¹³, R. Muchow¹³, J.C. Tassone¹⁴, X.-C. Liu¹⁴, R. Shindell¹⁵, W. Schrader¹⁶, C. Ebersson¹⁷, A. Lapinsky¹⁸, R. Loder¹⁹, J. Davey²⁰

Japan Scoliosis Clinical Research Group

Naobumi Hosogane²¹, Yoji Ogura²¹, Yohei Takahashi²¹, Atushi Miyake²¹, Kota Watanabe²¹, Kazuhiro Chiba²¹, Yoshiaki Toyama²¹, Katsuki Kono²², Noriaki Kawakami²³, Taichi Tsuji²³, Koki Uno²⁴, Teppei Suzuki²⁴, Manabu Ito²⁵, Hideki Sudo²⁵, Shohei Minami²⁶, Toshiaki Kotani²⁶, Haruhisa Yanagida²⁷, Hiroshi Taneichi²⁸, Ikuho Yonezawa²⁹, Kazuo Kaneko²⁹

¹²Department of Orthopedic Surgery, Texas Scottish Rite Hospital for Children, Dallas, Texas, USA. ¹³Department of Orthopaedic Surgery, Shriners Hospitals for Children, Lexington, Kentucky, USA. ¹⁴Department of Orthopaedic Surgery, Children's Hospital of Wisconsin, Milwaukee, Wisconsin, USA. ¹⁵OrthoArizona, Phoenix, Arizona, USA. ¹⁶Departments of Orthopedics, Sports Medicine, and Surgical Services, Akron Children's Hospital, Akron, Ohio, USA. ¹⁷Pediatric Orthopaedics and Scoliosis, Hasbro Children's Hospital, Providence, Rhode Island, USA. ¹⁸University of Massachusetts Memorial Medical Center, Worcester, Massachusetts, USA. ¹⁹Indiana University-Purdue University Indianapolis, Indianapolis, Indiana, USA. ²⁰University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, USA. ²¹Department of Orthopaedic Surgery, Keio University, Tokyo, Japan. ²²Department of Orthopaedic Surgery, Eiju General Hospital, Tokyo, Japan. ²³Department of Orthopaedic Surgery, Meijo Hospital, Nagoya, Japan. ²⁴Department of Orthopaedic Surgery, National Hospital Organization, Kobe Medical Center, Kobe, Japan. ²⁵Department of Orthopaedic Surgery, Hokkaido University, Sapporo, Japan. ²⁶Department of Orthopaedic Surgery, Seirei Sakura Citizen Hospital, Sakura, Japan. ²⁷Department of Orthopaedic Surgery, Fukuoka Children's Hospital, Fukuoka, Japan. ²⁸Department of Orthopaedic Surgery, Dokkyo Medical University, Mibu, Japan. ²⁹Department of Orthopedic Surgery, Juntendo University School of Medicine, Tokyo, Japan.