Supplemental Method：

Patient recruitment

The diagnoses of scoliosis included a standing full-spine X-ray, 3-dimensional (3D) computed tomography (CT), and magnetic resonance imaging (MRI). To affirm the idiopathic nature of the diagnosis, strict inclusion criteria were employed:

i) no congenital vertebrate malformation, neuromuscular defect or other congenital anomalies at the time of enrollment;

ii) no clinical or molecular diagnosis of Marfan syndrome, Ehlers-Danlos syndrome, neurofibromatosis, osteogenesis imperfecta, or other scoliosis-related disease found during phenotype evaluation;

iii) no factors that could lead to secondary scoliosis or kyphosis, such as spine injury, lumbar tuberculosis, and discrepancy of lower limbs.

Next generation sequencing

For whole-exome sequencing (WES), three different capture kits were used to build paired-end libraries: xGEN targeted capture kit v2(IDT, Coralville, IA USA), VCRome SeqCap EZ Chice HGSC 96 Reactions (Roche, Pleasanton, CA, USA), SureSelect Human All Exon V6+UTR r2 core design (Agilent, Santa Clara, CA, USA). For whole-genome sequencing (WGS) , sequencing libraries were prepared using the KAPA

Wen W, *et al. J Med Genet* 2024; 61:666–676. doi: 10.1136/jmg-2023-109667

Hyper Prep kit (KAPA Biosystems, Kusatsu, Japan) according to the optimized manufacturer's protocol.

Variant-level and sample-level quality control (QC)

SNVs and indels underwent multiple layers of filtration:

1. Hard filter: We filtered out variants that meet any of the following criteria:
    a) genotype quality (GQ) < 20
    b) depth (DP) < 10
    c) quality by depth (QD) < 2
    d) strand odds ratio (SOR) > 9
    e) Variant allele balance < 25%
2. Population-based filter: We filtered out variants that deviated from the Hardy-Weinberg equilibrium ($p < 0.000001$) and variants with a missing rate > 10% in the case-control population.
3. Variant Quality Score Recalibration (VQSR) was performed using the standard GATK protocol with a sensitivity of 99%.

After variant-level QC, we performed ancestry estimation by principal component analysis (PCA) using plink (version 1.90) based on the combined genotypes of the in-house subjects and the 1000 Genomes Phase III population[1]. Samples that met any of the following criteria were excluded:

1. PCA outliers (> 2 standard deviation)
2. Overall call rate < 0.95
3. Average depth < 30X

4. Heterozygosity < 0.8

In addition, the relatedness among individuals was calculated using the identity-by-descent (IBD) analysis. For each pair of individuals with IBD > 0.8, we excluded the one with a lower call rate.

Variant annotation

The Ensembl Variant Effect Predictor (VEP, version 103) was employed to annotate the qualifying variants[2]. The LofTee (https://github.com/konradjk/loftee) and dbNSFP[3] plugins were used to generate bioinformatic predictions. The Genome Aggregation Database (gnomAD, https://gnomad.broadinstitute.org/, accession date: 2022/05/25) was used to annotate population frequencies for the variants. Rare variants with a gnomAD population-max allele frequency ≤ 0.1% and a cohort allele frequency ≤ 0.1% were selected to perform gene-based mtational burden analysis. The retained variants were further annotated with transcript-level information according to NBCI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/, accession date: 2022/05/25). When a variant is allocated to multiple RefSeq transcripts of one gene, the RefSeq transcript labeled as 'canonical' by ENSEMBL (https://www.ensembl.org/, accession date:2022/05/25) was selected.

Gene-based association analysis

Weighting of variants

A weight of 0-1 was assigned to each variant according to the variant type and bioinformatic predictions. REVEL[4] and CADD, two ensemble predictors, were used for the prediction of missense variants and in-frame indels respectively. The detailed weighting standards are provided in Supplemental Table 3.

Subset analysis

In addition to a global weighted burden test, we also performed on a subset of synonymous variants to calibrate the burden test. For the synonymous variants, a minimum variant count of n=3 in the case-control population was required for a gene to be included rather than n=5 for the all-variant-model.

Statistic methods

After variant filtration, the number of cases/controls carrying at least one qualifying variant in each gene was calculated and compared using a two-sided Fisher's Exact test. P-values were adjusted for multiple testing using the Benjamini-Hochberg (BH) procedure. A conservative Bonferroni-corrected gene-level exome-wide significance threshold of P = 0.05/ (1 model × 19,337 genes) = 2.6×10^{-6} was used.

RNA Sequencing and QC

Total RNA was extracted using TriZol, and RNA quality was assessed using the NanoDrop 2000 Kit and the Agilent 2100 Bioanalyzer. RNA libraries were prepared using the TruSeq Stranded Total RNA Library Prep Kit (Illumina), which involved poly-T mRNA selection using oligo-dT beads, fragmentation, and reverse transcription to cDNA. The prepared libraries were then sequenced on an Illumina NovaSeq 6000 system, generating paired-end reads of 150 bp.

The initial acquisition of raw data, comprised of FASTQ-formatted sequences, was subjected to preprocessing using proprietary Perl scripts. In this initial quality control stage, we established a cleaned dataset (termed 'clean reads') by eliminating the following categories of reads:

1. Reads possessing adapter sequences;
2. Reads featuring more than three undefined nucleotides (N);

3. Reads where over 20% of the nucleotides registered a Phred quality score (Qphred) of 5 or lower.

Subsequently, key parameters, such as Q20, Q30, and GC content, were calculated for the resultant clean data. To further refine the dataset, these 'clean reads' were aligned to the SILVA rRNA database (https://www.arb-silva.de/), facilitating the elimination of any rRNA sequences. All subsequent downstream analyses were conducted exclusively on this rRNA-free clean dataset.

Polygenic risk analysis

To investigate the combined effect of rare and common variants, we employed a Polygenic Risk Score (PRS) model based on the 20 most significant unrelated SNPs, with the p-value threshold of $5*10^{-8}$, derived from the GWAS summary statistics of a large-scale AIS meta-GWAS study focused on the Japanese population[5] (Supplemental Table 12).

Using the PRScise (v2.3.5) with recommended parameters, we optimized and calculated the PRS for our cohort of 1696 WGS samples. This calculation involved integrating the weighted effect of each SNP, providing a quantitative measure of genetic risk. The resultant PRS data were analyzed to discern patterns and associations, particularly focusing on the distribution differences.

1 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015). https://doi.org:10.1038/nature15393

2 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016). https://doi.org:10.1186/s13059-016-0974-4

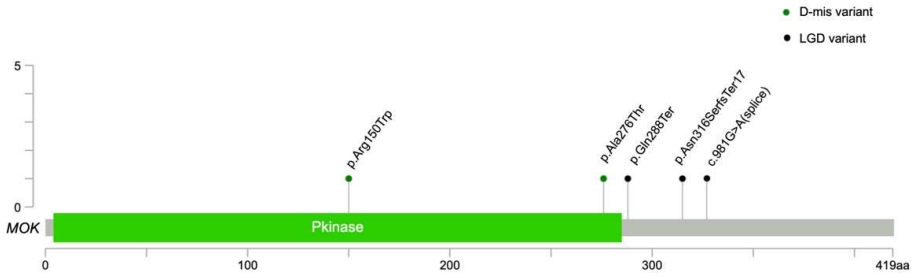
3 Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med* **12**, 103 (2020). https://doi.org:10.1186/s13073-020-00803-9

4 Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885 (2016). https://doi.org:10.1016/j.ajhg.2016.08.016

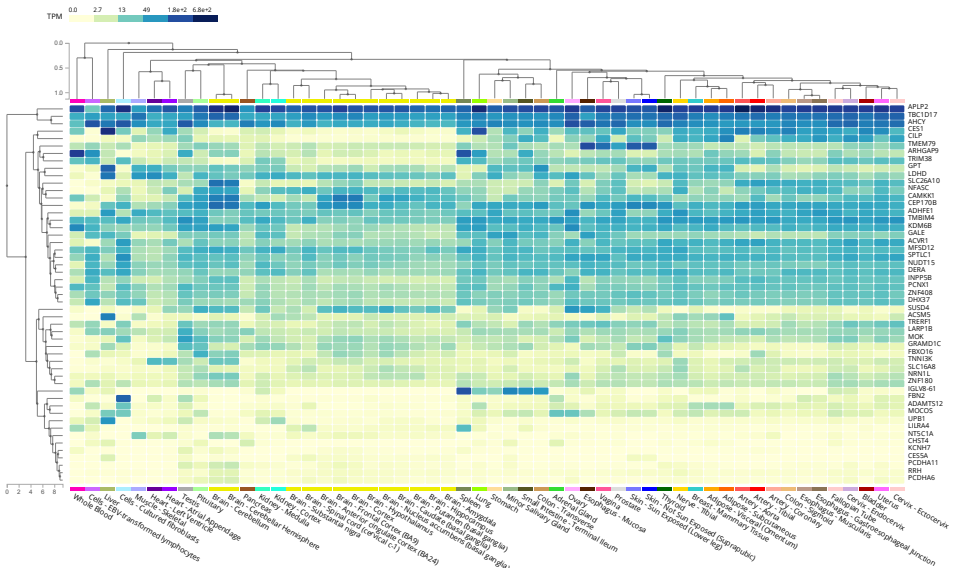
5 Kou, I. *et al.* Genome-wide association study identifies 14 previously unreported susceptibility loci for adolescent idiopathic scoliosis in Japanese. *Nature communications* **10**, 3685 (2019).

Supplemental Figures and legends

Supplemental Figure 1: Mutational spectrum of D-mis and LGD variants in MOK identified from AIS cases.

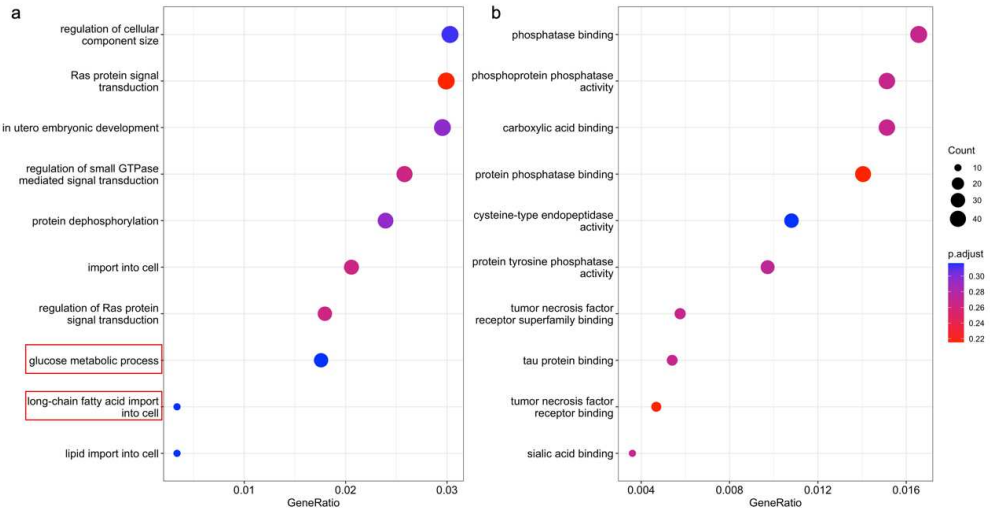


Supplemental Figure 2: Tissue-specific heatmap of RNA expression of candidate genes.



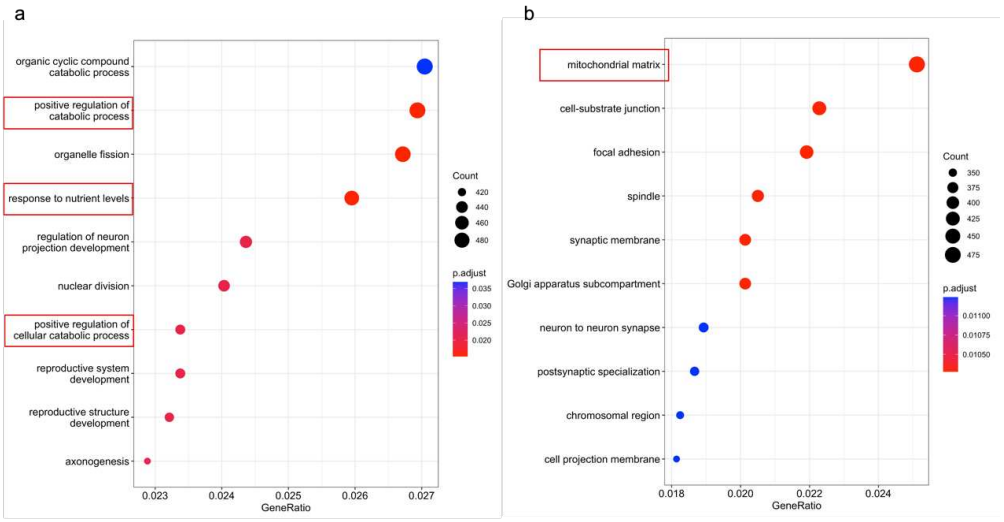
The heatmap displays the expression levels of the top 50 genes identified in the gene-based burden analysis across multiple tissues. The color gradient from light yellow to dark blue represents the transcripts per million (TPM) values, with higher TPM values indicating higher expression levels. The expression data was obtained from the GTEx database.

Supplemental Figure 3: Pathway enrichment analysis of differentially expressed genes within AIS subgroups.



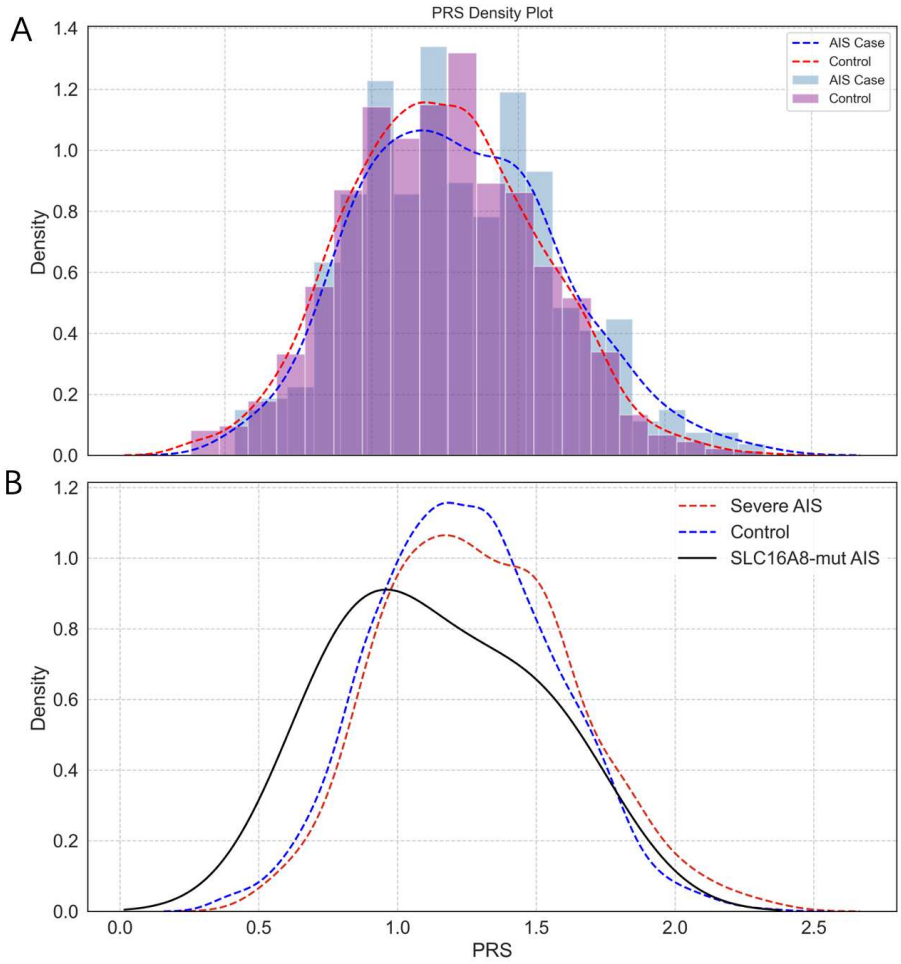
The dotplot displays the top 10 results of the Gene Ontology Biological Process (a) and Cellular Components (b) enrichment analyses between AIS subgroups based on the genomic variants in carbohydrate metabolism. The dots represent the enriched pathways, with the size of the dots indicating the number of genes involved in the pathway and the color indicating the statistical significance of the enrichment. The metabolism pathways are highlighted in a red frame.

Supplemental Figure 4: Pathway enrichment analysis of differentially expressed genes between AIS cases and CVM controls.



The dotplot displays the top 10 results of the Gene Ontology Biological Process (a) and Cellular Components (b) enrichment analyses towards AIS and control. The dots represent the enriched pathways, with the size of the dots indicating the number of genes involved in the pathway and the color indicating the statistical significance of the enrichment. The metabolism pathways are highlighted in a red frame.

Supplemental Figure 5: Polygenic Risk Score (PRS) distribution of different groups**.**



a) The histogram shows the PRS distribution among AIS cases and controls, bule for AIS and purple for controls, with corresponding density curves in dashed lines. b) A visualization comparing the PRS distribution among AIS cases, SLC16A8-mutated AIS cases and controls.

Supplemental Figure 6: Inheritance Pattern of the Validated Trio

The pedigree validation of AIS18012500102-WGS-1. The location of candidate variant (SLC16A8: p.Gly296Asp) is chr22:38081151 in hg38, highlighted in the red frame. The panels are sequencing alignment results illustrated by Integrative Genomics Viewer (IGV).

Wen W, *et al. J Med Genet* 2024; 61:666–676. doi: 10.1136/jmg-2023-109667

Supplemental Tables

Supplemental Table 1: Clinical characteristic of AIS patient
Uploaded as separate Excel-file.

Supplemental Table 2: Sequencing information of case and control datasets

Uploaded as separate Excel-file.

Supplemental Table 3: Standard of mask level and corresponding weight value

Uploaded as separate Excel-file.

Supplemental Table 4: Pathway-based gene-set burden analysis results, including the enriched pathways, their associated p-values, and the involved genes

Uploaded as separate Excel-file.

Supplemental Table 5: Nonsynonymous rare variants in genes involved in carbohydrate metabolism identified in AIS cases, including the allele frequency and annotation information

Uploaded as separate Excel-file.

Supplemental Table 6: Comparison of phenotypic characteristics between AIS cases with and without rare variants in genes involved in carbohydrate metabolism

Uploaded as separate Excel-file.

Supplemental Table 7: candidate lactate transporter genes

Uploaded as separate Excel-file.

Supplemental Table 8: Nonsynonymous rare variants in lactate transporter genes identified in AIS cases

Uploaded as separate Excel-file.

Supplemental Table 9: Information about AIS patients and CVM controls included in the expression analysis, including their clinical and genetic characteristics

Uploaded as separate Excel-file.

Supplemental Table 10: Enrichment results of the expression differential analysis between AIS subgroups based on Gene Ontology

Uploaded as separate Excel-file.

Supplemental Table 11: Enrichment results of the expression differential analysis between AIS and control groups based on Gene Ontology

Uploaded as separate Excel-file.

Supplemental Table 12: Reported significant AIS-associated SNPs and prioritized genes

Uploaded as separate Excel-file.

Wen W, *et al. J Med Genet* 2024; 61:666–676. doi: 10.1136/jmg-2023-109667

Supplemental Table 13: The selected SNP from GWAS summary statics for PRS model
Uploaded as separate Excel-file.