OPEN ACCESS

Original research

Exome sequencing of 1190 non-syndromic clubfoot cases reveals *HOXD12* as a novel disease gene

Wu-Lin Charng ,[1] Momchil Nikolov,[1] Isabel Shrestha,[1] Mark A Seeley,[2] Navya Shilpa Josyula,[3] Anne E Justice,[3] Matthew B Dobbs,[4] Christina A Gurnett[1]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/jmg-2024-109846).

[1]Department of Neurology, Washington University in Saint Louis School of Medicine, Saint Louis, Missouri, USA
[2]Department of Orthopaedics, Geisinger Medical Center, Danville, Pennsylvania, USA
[3]Department of Population Health Sciences, Geisinger, Danville, PA, USA
[4]Paley Orthopedic & Spine Institute, West Palm Beach, Florida, USA

Correspondence to
Dr Christina A Gurnett, Department of Neurology, Washington University in St Louis, St Louis, MO 63130, USA; gurnettc@wustl.edu

Received 2 January 2024
Accepted 20 March 2024

To cite: Charng W-L, Nikolov M, Shrestha I, *et al*. *J Med Genet* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jmg-2024-109846

ABSTRACT

Background Clubfoot, presenting as a rigid inward and downward turning of the foot, is one of the most common congenital musculoskeletal anomalies. The aetiology of clubfoot is poorly understood and variants in known clubfoot disease genes account for only a small portion of the heritability.

Methods Exome sequence data were generated from 1190 non-syndromic clubfoot cases and their family members from multiple ethnicities. Ultra-rare variant burden analysis was performed comparing 857 unrelated clubfoot cases with European ancestry with two independent ethnicity-matched control groups (1043 in-house and 56 885 gnomAD controls). Additional variants in prioritised genes were identified in a larger cohort, including probands with non-European ancestry. Segregation analysis was performed in multiplex families when available.

Results Rare variants in 29 genes were enriched in clubfoot cases, including *PITX1* (a known clubfoot disease gene), *HOXD12*, *COL12A1*, *COL9A3* and *LMX1B*. In addition, rare variants in posterior *HOX* genes (*HOX9–13*) were enriched overall in clubfoot cases. In total, variants in these genes were present in 8.4% (100/1190) of clubfoot cases with both European and non-European ancestry. Among these, 3 are *de novo* and 22 show variable penetrance, including 4 *HOXD12* variants that segregate with clubfoot.

Conclusion We report *HOXD12* as a novel clubfoot disease gene and demonstrate a phenotypic expansion of known disease genes (myopathy gene *COL12A1*, Ehlers-Danlos syndrome gene *COL9A3* and nail-patella syndrome gene *LMX1B*) to include isolated clubfoot.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Talipes equinovarus, or clubfoot, is one of the most common congenital musculoskeletal anomalies.
⇒ Nevertheless, variants in known disease genes, including *PITX1*, account for only a small portion of heritability.
⇒ Large-scale exome sequencing studies of non-syndromic clubfoot have not previously been reported.

WHAT THIS STUDY ADDS

⇒ *HOXD12* is a novel clubfoot disease gene in which primarily missense variants result in highly penetrant autosomal dominant inherited clubfoot.
⇒ Phenotypes of three known disease genes (the myopathy gene *COL12A1*, Ehlers-Danlos syndrome gene *COL9A3* and nail-patella syndrome gene *LMX1B*) are expanded to include isolated clubfoot.
⇒ Rare variants in limb-expressed posterior *HOX* genes (*HOX9–13*) are enriched in the clubfoot cohort, consistent with the expression and function of posterior *HOX* genes.
⇒ Rare and predicted deleterious variants in *PITX1*, *LMX1B*, *COL9A3*, *COL12A1*, *HOXD12* and other posterior *HOX* genes are present in only 8.4% (100/1190) of our clubfoot cases highlighting the genetic heterogeneity of this congenital disorder.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Given the genetic heterogeneity of clubfoot, exome sequencing is a reasonable strategy for clinical diagnosis though the yield is modest.
⇒ The variants in *PITX1*, *LMX1B*, *COL9A3*, *COL12A1*, *HOXD12* and other posterior *HOX* genes reported in this study are identified in both Caucasian and non-Caucasian (African-American, Asian, Hispanic/Latino and others) probands, suggesting our findings may apply to many populations.

INTRODUCTION

Talipes equinovarus (TEV (MIM: 119800)), or clubfoot, is a structural abnormality of leg, ankle and foot, resulting in a rigid, inward and downward turning of the foot.[1] Left untreated, the deformities lead to pain and disability. Current clubfoot casting and bracing treatments have improved outcomes, but the treatment course is prolonged and relapses are common. The prevalence of clubfoot is 0.5–2 individuals per 1000 live births,[2] making it one of the most common congenital disorders. The clubfoot phenotype may be unilateral or bilateral and it affects more males, with male-to-female ratio of 2:1 across multiple ethnicities.[3] Approximately 80% of patients have non-syndromic clubfoot without other malformations while the remaining 20% have syndromic clubfoot with another congenital disorder, such as distal arthrogryposis, congenital myotonic dystrophy or myelomeningocele.[4]

There is a family history in about 25% of non-syndromic cases and data suggesting both autosomal dominant with incomplete penetrance and recessive inheritance modes.[3 5 6] Results from

genome-wide association studies, candidate gene association studies and exome sequencing analysis have revealed multiple genetic factors associated with non-syndromic clubfoot risk,[1] including variants in genes involved in limb development, such as *HOXC* genes[7] and PITX1-TBX4 pathway,[8 9] scaffold protein for extracellular matrix and cytoskeleton *Filamin B* (*FLNB*)[10 11] and duplication of *SHOX*.[12] Nevertheless, variants in known disease genes account for only a small portion of heritability. Large-scale exome sequencing studies of non-syndromic clubfoot have not been reported. Therefore, in this study, we performed exome sequencing of 1293 individuals from a variety of ethnic backgrounds, including 1190 unrelated clubfoot probands and their family members.

METHODS

Patients

1190 unrelated patients with non-syndromic clubfoot were recruited at St. Louis Children's Hospital in St. Louis, Missouri, USA, and Shriners Hospital, St Louis. The gender ratio of male-to-female cases was approximately 2:1, with 822 male and 368 female cases among our exome sequenced cohort. Probands included 936 with European ancestry, 29 Asian, 51 African-American, 24 Hispanic/Latino, 53 multiracial, and 97 others or not reported. 35% (416/1190) probands have a family history for clubfoot, among which 13.8% (164/1190) has a first-degree relative with clubfoot. The clubfoot diagnosis required rigid hindfoot equinus, hindfoot varus, midfoot supination and midfoot cavus deformities. Syndromic cases were excluded. However, it is possible that some cases may have been recruited prior to the recognition of a syndromic status, although we had long-term follow-up on the majority of cases. DNA was isolated from blood or saliva with DNA Genotek kits. The in-house controls consist of unrelated individuals of European ancestry with Alzheimer's disease, amyotrophic lateral sclerosis or male infertility.

Exome sequencing and annotations

Exome libraries were prepared with either Agilent's SureSelect Human All Exon kits V5 or IDT xGen Exome Panel V1 capture and then sequenced at the McDonnell Genome Institute on Illumina HiSeq 2000/4000 or NovaSeq 6000 with paired-end reads. Raw sequencing data were aligned to the human genome reference (GRCh37) using BWA-MEM and marked for duplication with Picard MarkDuplicates. The following data processing was performed according to GATK Best Practices[13] for GATK V.3. Variant joint calling was performed for all cases and in-house control using GenotypeGVCFs. We kept variants with depth ≥10, genotype quality ≥20 and allele balance for heterozygous calls between 0.2 and 0.8. Based on variant quality scores recalibrated with VariantRecalibrator, single-nucleotide variants fell above 99.7 and indels fell above 99.0 were kept. Multiallelic-sites were split and left-realigned. The final VCF was annotated using ANNOVAR[14] with Gencode V.19, Combined Annotation-Dependent Depletion (CADD V.1.3),[15] MCAP (V.1.3),[16] REVEL,[17] Polyphen2,[18] ClinVar,[19] InterVar[20] and gnomAD exome (V.2.1.1).[21]

Data cleaning for gene burden analysis

Exomes with genotype call rate <90% and inconsistent gender between genotypes and clinical database were excluded. Individual relatedness was evaluated using KING[22] and identity-by-descent calculation using PLINK V.1.9.[23] For each pair of relativeness more closely related than second degree, we retained the exome with higher genotyping rate. 857 unrelated clubfoot cases and 1043 in-house controls were anchored to non-Finnish European population in 1000 genome phase III data in principal component analysis.

Gene/Gene group/region-specific burden analysis

Genes on sex chromosomes were not included and gender was not considered in this analysis to maximise the sample size for both cases and controls. Sites with genotype call rate of <90% in cases and both controls and indels >10 base pairs were excluded. The Testing Rare vAriants using Public Data (TRAPD) method[24] was adapted to perform gene burden analysis between 857 unrelated non-syndromic clubfoot probands of European ancestry and 2 independent controls (1043 in-house controls and 56 885 gnomAD non-Finnish European controls) in the autosomal dominant mode. Multiple filters were applied to select deleterious/likely deleterious variants: (1) ultra-rare variant allele count (AC) ≤3 for both clubfoot and control cohorts, (2) removal of benign/likely benign variants in ClinVar or InterVar,[19 20] (3) pathogenic prediction scores (Polyphen2 not Benign category, CADD phred ≥20, M-CAP score >0.025, REVEL ≥0.25),[15–18] (4) use missense variants for genes with low pLI score (the probability of being loss-of-function intolerant) and consider all variants for genes with high pLI score from gnomAD (online supplemental table S1). In the comparison between cases and in-house controls, we added a gnomAD minor allele frequency filter (≤0.002%) to exclude rare variants with higher frequency in gnomAD database. For control variants, we used synonymous variants with the same frequency criteria along with removal of pathogenic/likely pathogenic variants in ClinVar or InterVar[19 20] (online supplemental table S1). Genes with enrichment of both deleterious and control (synonymous) variants in clubfoot cases were excluded and not considered as enriched.

For gene group (*HOX*) burden analysis, the same filters were applied and the qualified variants were collapsed within the gene groups. For region-specific (*COL12A1*) burden analysis, the same filters were applied, and we collapsed qualified variants within the shared region, and those within the unique region encoding the long form.

Screen for additional variants in candidate genes

Additional variants in *PITX1*, *LMX1B*, *COL9A3*, *COL12A1*, *COL15A1*, *HOXD12* and other posterior *HOX* genes (*HOX9–13*) were identified in the entire clubfoot cohort, including individuals of non-European ancestry and individuals with genotyping rate <90%. Indels >10 base pairs were also included. For low pLI genes, we did not include frameshift indels. To include additional probands in which segregation analysis could be performed, we used a less stringent filters as in TRAPD (AC ≤3, inhouse AC ≤clubfoot AC and gnomAD MAF $\leq 5\times10^{-4}$; REVEL >0.1, other cut-offs for prediction scores are the same).

Sanger sequencing for segregation analysis

Genomic DNA extracted blood or saliva samples were amplified using suitable primers to cover the region of interest. The purified PCR amplicons were used in Sanger sequencing performed by Azenta Life Sciences, South Plainfield, New Jersey, USA.

RESULTS

Rare variants in 29 genes are enriched in non-syndromic clubfoot cases

Exome sequencing was performed on 1293 participants, including 1190 probands with clubfoot and 103 additional

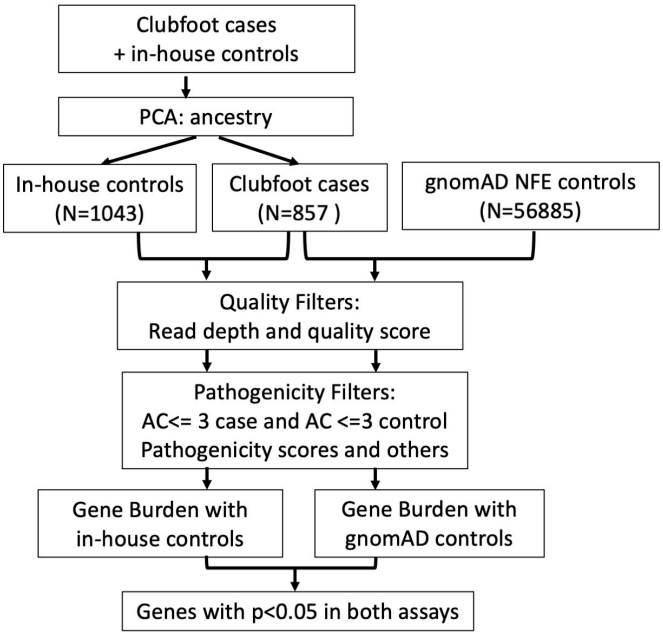


Figure 1 Flow chart of gene burden analysis. During data cleaning (see 'Methods' section), exomes of clubfoot cases and in-house controls went through principal component analysis (PCA) to retain data anchored to non-Finish European (NFE) population in 1000 genome phase III data. Testing Rare vAriants using Public Data (TRAPD) method was used in gene burden analysis between 857 unrelated isolated clubfoot probands and 2 independent controls (1043 in-house controls and 56 885 gnomAD NFE controls) in the autosomal dominant mode. Multiple filters were applied in order to select the deleterious/likely deleterious variants in each gene (see 'Methods' section). AC, allele count.

members of 74 families. Among 74 families, 4 families are sequenced as trios, 17 families have more than 2 family members being sequenced and 53 families have 2 affected members being sequenced. To identify new disease genes for clubfoot, we evaluated the enrichment of rare variants in clubfoot cases when compared with two different sources of control data (in-house and gnomAD controls) to limit false discovery. We analysed our data using TRAPD,[24] which was developed to perform Fisher's exact test for rare variant gene burden analysis between cases and publicly available control dataset. To select for deleterious variants in each gene, we applied four filters, including ultra-rare AC, ClinVar classification, pathogenic prediction scores and variant types (see 'Methods' section) (online supplemental table 1). The enrichment of deleterious variants in 857 unrelated isolated clubfoot probands of European ancestry was compared with 1043 in-house controls (both anchored to 1000 genomes data) and 56 885 gnomAD non-Finnish European controls in the autosomal dominant mode (figure 1). Three genes (*RBM47*, *OGFOD1*, *SH3TC2*) were enriched for both synonymous (control) variants and deleterious variants in our clubfoot cohort and were therefore excluded from our gene burden analysis.

From this analysis, we identified 29 genes with enrichment of rare deleterious variants in non-syndromic clubfoot cases compared with both control datasets ($p<0.05$ in both analyses) (table 1). *PITX1*, a clubfoot disease gene,[9] was one of the enriched genes, serving as a positive control in this analysis. *COL12A1*, *LMX1B*, *COL9A3*, *SLC26A2*, *HSPG2* and *MESD* are OMIM disease genes associated with multiple developmental defects including limb abnormalities (online supplemental table 2). Moreover, the limb defects caused by *LMX1B, SLC26A2* and *HSPG2* specifically include clubfoot.[25–28] Although rare variants in *HOXD12* have not been associated with human disease, *Hoxd12* null mice exhibit limb defects.[29 30] The remaining genes function in a wide variety of cellular processes (online supplemental table 2).

Additional rare variants in enriched genes in clubfoot cohort

Based on their dominant disease inheritance and limb-related phenotypes/functions, we identified additional variants in *PITX1*, *LMX1B*, *COL9A3*, *COL12A1* and *HOXD12* in our entire clubfoot cohort (n=1190), which also included cases of non-European ancestry who were excluded from the discovery rare variant gene burden analysis. To identify additional probands in which segregation analysis could be performed, we used less stringent filters than the TRAPD analysis (see 'Methods' section). Using this strategy, we identified 7 *PITX1* candidate disease variants in 8 individuals among 1190 probands (online supplemental table 3). These variants are conserved across species and located throughout the PITX1 protein, with no obvious motif enrichment (online supplemental figures 1 and 2). In total, we identified 10 candidate variants in *LMX1B*, 16 in *COL9A3*, 22 in *COL12A1* and 9 in *HOXD12* (online supplemental table 3).

HOXD12 is a novel disease gene for non-syndromic clubfoot

We identified 9 variants in *HOXD12* in 10 cases (online supplemental table 3). *HOXD12* encodes a homeodomain-containing transcription factor and is one of the posterior *HOX* paralogs expressed early in limb development.[31] Four variants segregate with clubfoot as a dominant condition with complete penetrance and two probands carry the same *HOXD12* variant (figure 2). For the remaining variants, there were no additional affected family members to test for segregation. The amino acids corresponding to the candidate variants are conserved across species (online supplemental figure 3) and located in two clusters within the N-terminal region and around the C-terminal homeobox domain (figure 2). The *HOXD12* variants in clubfoot cases are all missense except one long non-frameshift indel (online supplemental figure 3). Notably, loss-of-function alleles are less likely to contribute to clubfoot pathogenesis because of the low pLI (pLI=0.00032) and because several nonsense or frameshift variants found in clubfoot cases had higher minor allele frequencies in our in-house controls (data not shown). This also suggests that the segregating missense variants may cause disease through dominant negative effects on protein function.

Rare variants in limb-expressed posterior *HOX* genes are enriched in clubfoot cases

Animal studies have shown that posterior or 5' *HOX* (*HOX9–13*) genes are highly expressed in limb and critical for limb development.[31] Based on our discovery of *HOXD12* as a novel disease gene and our prior identification of posterior *HOXC* gene deletions in clubfoot,[7 32] we sought to determine whether there is an overall enrichment of rare and deleterious variants in posterior *HOX* genes in clubfoot cases. Except for *HOXD12*, there are insufficient variants for any single *HOX* genes to demonstrate statistically significant enrichment in our gene burden analysis (table 2), therefore we collapsed rare and deleterious variants across all posterior *HOX* genes and performed similar burden analysis. We used non-posterior *HOX* (*HOX1–8*) genes which are not expressed in limb and autism genes as negative controls for comparison. Our data show that rare variants in posterior *HOX* (*HOX9–13*) genes are enriched ($p<0.05$) in clubfoot cases when compared with the two independent controls and there is

Table 1 Genes with ultra-rare variants enriched in non-syndromic clubfoot cases when compared with both gnomAD controls and in-house controls

		Enrichment versus gnomAD			Enrichment versus in-house controls		
Gene	VAR_TYPE	P_DOM	CASE (n=857)	CONTROL (n=56 885)	P_DOM	CASE (n=857)	CONTROL (n=1043)
ATP6V0D2	Missense	3.68E-06	7	37	4.12E-02	4	0
COL15A1	Missense	1.05E-04	10	143	2.04E-02	10	3
PHLDB1	Missense	3.15E-04	8	106	1.68E-03	8	0
ZFYVE28	Missense	3.34E-04	8	107	2.75E-02	8	2
PITX1	All	4.12E-04	6	59	8.34E-03	6	0
MESD	All	1.95E-03	4	32	4.12E-02	4	0
RHOH	Missense	1.95E-03	4	32	4.12E-02	4	0
LMX1B	All	3.02E-03	5	61	3.59E-02	6	1
GLDC	Missense	3.21E-03	9	191	2.55E-02	11	4
PC	Missense	3.43E-03	9	193	2.75E-02	8	2
ATG2A	Missense	4.04E-03	6	95	3.59E-02	6	1
WDR7	All	4.34E-03	7	129	1.82E-02	7	1
COL12A1	All	5.49E-03	13	372	2.14E-02	14	6
G6PC1	Missense	6.85E-03	5	75	4.12E-02	4	0
SUCLG2	Missense	6.85E-03	5	75	4.12E-02	4	0
CATSPERG	Missense	7.31E-03	3	24	1.86E-02	5	0
COL9A3	Missense	7.87E-03	7	145	2.75E-02	8	2
HPS6	Missense	7.98E-03	5	78	1.86E-02	5	0
DNAH1	Missense	1.05E-02	13	405	3.42E-02	13	6
ANKRD27	Missense	1.07E-02	6	118	3.59E-02	6	1
BLTP3A	Missense	1.22E-02	4	56	4.12E-02	4	0
KDM5B	Missense	1.24E-02	7	159	9.10E-03	8	1
HOXD12	Missense	1.50E-02	3	32	1.86E-02	5	0
VPS37D	All	1.54E-02	2	11	4.12E-02	4	0
MAPK7	Missense	1.68E-02	4	62	4.12E-02	4	0
SLC26A2	Missense	1.97E-02	5	99	1.86E-02	5	0
KRT28	Missense	2.04E-02	4	66	4.12E-02	4	0
HSPG2	Missense	2.19E-02	16	595	4.64E-02	17	10
XKR4	All	3.14E-02	4	76	4.12E-02	4	0

CASE, count in cases; CONTROL, count in controls; gnomAD TRAPD, TRAPD assay using gnomAD non-Finish European data; in-house TRAPD, TRAPD assay using in-house Caucasian controls; P_DOM, p value in dominant mode of TRAPD; TRAPD, Testing Rare vAriants using Public Data; VAR_TYPE, type of variants included in this burden analysis.

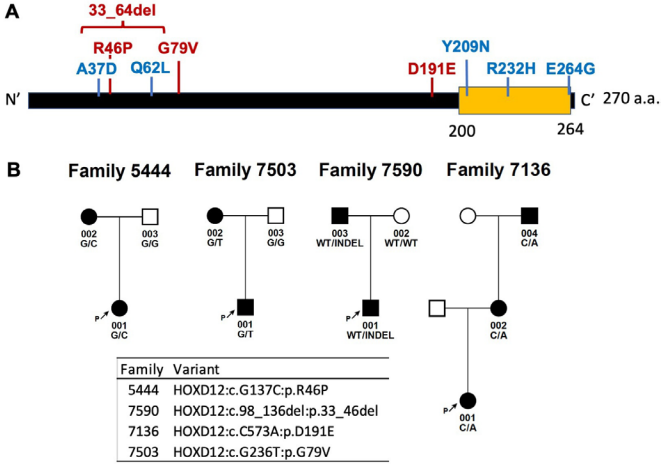


Figure 2 Localisation and segregation of *HOXD12* variants. (A) Localisation of *HOXD12* variants identified in the study along the protein. *HOXD12* variants co-segregated with clubfoot in multiplex families are marked in red. *HOXD12* variants in singletons are marked in blue. Yellow motif represents homeobox domain. (B) *HOXD12* variants co-segregate with clubfoot in four multiplex families with complete penetrance. a.a., amino acid.

no enrichment in negative control gene sets (table 2). We then extended our analysis to the entire clubfoot cohort, including non-European ancestry cases, and identified 39 candidate variants in posterior *HOX* genes (other than *HOXD12*) (online supplemental table 4). Segregation analysis identified a *de novo* variant in *HOXD11* and two variants in *HOXC11* that segregate with clubfoot in two multiplex families (online supplemental figure 4). In addition, variants in *HOXC9*, *HOXC12*, *HOXD11* potentially segregate with clubfoot in three families (online supplemental figure 4). Overall, our data provides additional evidence for dysfunction of multiple posterior *HOX* genes in clubfoot aetiology.

Phenotypic expansion of nail-patella syndrome gene *LMX1B* to include non-syndromic clubfoot

LMX1B encodes a LIM-homeodomain transcription factor essential for limb, kidney and eye development. Pathogenic variants in *LMX1B* cause autosomal dominant nail-patella syndrome (NPS (MIM: 161200)), which includes nail dysplasia, patellar abnormalities, clubfoot, nephropathy and glaucoma.[26 33] Among the 10 *LMX1B* candidate variants (online supplemental table 3), 3 probands had additional affected family members with DNA available for testing. Variants in *LMX1B* segregate with clubfoot in all three families with dominant inheritance (online

Table 2 Enrichment of ultra-rare variants in *HOX* genes in non-syndromic clubfoot cases

Gene/Pathway		VAR_TYPE	Enrichment versus gnomAD			Enrichment versus in-house control		
			P_DOM	CASE (n=857)	CONTROL (n=56 885)	P_DOM	CASE (n=857)	CONTROL (n=1043)
Limb *HOX* gene	*HOXA9*	Missense	7.71E-02	2	29	4.27E-01	2	1
	HOXA10	All	4.97E-01	1	45	1	0	1
	HOXA11	All	3.12E-01	1	24	4.51E-01	1	0
	HOXA13	All	2.70E-01	1	20	4.51E-01	1	0
	HOXB9	Missense	4.25E-01	1	36	4.51E-01	1	0
	HOXC9	Missense	4.58E-01	1	40	6.99E-01	1	1
	HOXC10	Missense	9.96E-02	2	34	2.03E-01	2	0
	HOXC11	Missense	1.29E-01	2	40	4.27E-01	2	1
	HOXC12	Missense	5.20E-01	1	48	6.99E-01	1	1
	HOXC13	All	3.32E-01	1	26	6.99E-01	1	1
	HOXD9	Missense	1	0	20	1	0	1
	HOXD10	Missense	1	0	33	1	0	2
	HOXD11	Missense	1	0	25	NA	0	0
	HOXD12	Missense	1.50E-02	3	32	1.86E-02	5	0
	Combined (*HOX9–13*)		1.96E-03	16	452	2.92E-02	17	9
Non-limb *HOX* genes (*HOX1–8*) (combined)			7.71E-01	10	806	7.66E-01	15	22
Autism gene set			5.16E-01	125	8296	5.58E-01	136	167

CASE, count in cases; CONTROL, count in controls; gnomAD TRAPD, TRAPD assay using gnomAD non-Finish European data; in-house TRAPD, TRAPD assay using in-house Caucasian controls; Limb *HOX*, posterior *HOX* genes (*HOX9–13*); NA, not available; Non-limb *HOX*, non-posterior *HOX* genes (*HOX1–8*); P_DOM, p value in dominant mode of TRAPD.

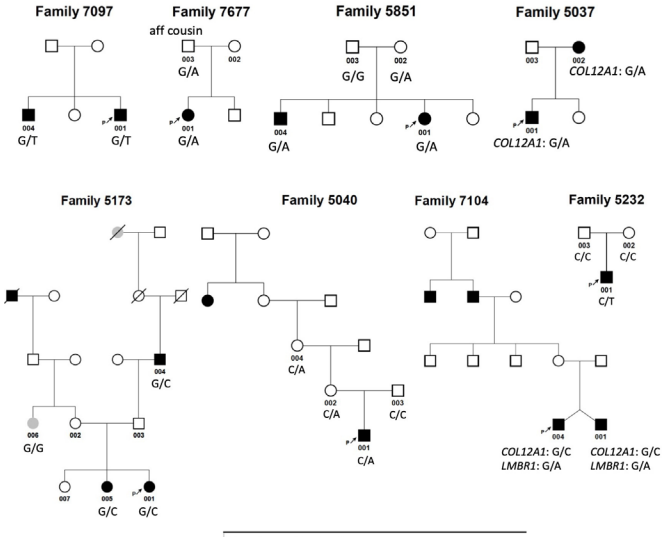
supplemental figure 5). In one of family, a history of NPS was discovered retrospectively, although three members of that family had non-syndromic clubfoot without other evidence of NPS. The other families had no clinical evidence of NPS. Our data suggest that individuals with pathogenic variants in *LMX1B* may present with non-syndromic clubfoot without associated NPS phenotypes.

Phenotypic expansion of collagen disease genes *COL9A3* and *COL12A1* to include non-syndromic clubfoot

Of the 29 clubfoot enriched genes in our burden analysis, 3 are non-fibrillar collagen genes (*COL9A3*, *COL12A1* and *COL15A1*). *COL9A3* and *COL12A1* are OMIM disease genes with limb phenotypes.[34–37] *COL9A3* variants have been identified in a handful of patients with autosomal dominant multiple epiphysial dysplasia-3 (MIM: 600969)[34 36] with limb defects and in recessive Stickler syndrome (MIM: 620022).[38] We identified 16 *COL9A3* candidate variants (online supplemental table 3), including 6 missense variants resulting in glycine substitutions in the Gly-X-Y repeat domain. Three variants potentially segregate with incomplete penetrance because they are inherited from unaffected parents with clubfoot family history (online supplemental figure 6). In addition, one *COL9A3* missense variant is *de novo* in a proband without family history (online supplemental figure 6).

COL12A1 encodes the alpha chain of type XII collagen.[39] Deleterious variants in *COL12A1* have been described in a few patients of autosomal recessive Ullrich congenital muscular dystrophy 2 (MIM: 616470)[37] and Bethlem myopathy 2 (MIM: 616471).[35 37] Twenty-two candidate variants in *COL12A1* were identified in 21 people in our clubfoot cohort, including 2 nonsense and 21 missense variants, with one of these resulting in a glycine substitution (online supplemental table 3). Variants in *COL12A1* segregate with clubfoot in five multiplex families, potentially in two families and one proband carries a *de novo* *COL12A1* variant (figure 3).

There are two splice variants of *COL12A1*, long form (collagen XIIA) and short form (collagen XIIB), with distinct spatial and temporal expression patterns.[39] The long form is predominantly expressed in early embryonic development and then restricted to dense connective tissues while the short form becomes predominant at later stages.[39] To determine whether the location of



Family	Variant
7097	COL12A1:c.C4377A:p.N1459K
7104	COL12A1:c.C467G:p.S156X
5851	COL12A1:c.C9110T:p.P3037L
5173	COL12A1:c.C3245G:p.T1082R
5037	COL12A1:c.C967T:p.Q323X
5040	COL12A1:c.G616T:p.D206Y
7677	COL12A1:c.C758T:p.S253F
5232	COL12A1:c.G958A:p.V320I

Figure 3 Pedigrees of segregated *COL12A1* families. Variants in *COL12A1* co-segregate with clubfoot in five multiplex families, potentially in two families and one proband carries a *de novo* *COL12A1* variant.

variants within the different splice variants impacts clubfoot risk, we performed burden analysis for variants within the shared region and those within the unique region encoding the long form. Variants in the unique region are enriched in clubfoot cases when compared with gnomAD controls ($p<0.05$), although this was not significant when compared with our smaller in-house control cohort (online supplemental table 5).

Another collagen gene with rare variants enriched in our clubfoot cases is *COL15A1*, which has not yet been linked to any human disease. We also screened for additional *COL15A1* variants in the entire clubfoot cohort, and identified 17 rare variants in 17 individuals: one proband also has a variant in *COL9A3* (online supplemental table 6), another proband also has variants in *PITX1* and *LMX1B* (online supplemental table 6), two variants do not segregate and one is likely inherited from his father whose half-brother had clubfoot (online supplemental table 7). Therefore, we do not have strong evidence to support a role of *COL15A1* in clubfoot aetiology with current data.

In summary, from exome sequence data including 1190 probands with non-syndromic clubfoot and 103 additional members of 74 families, we identified *HOXD12* as a novel clubfoot disease gene and expanded the disease phenotypes of *LMX1B*, *COL9A3* and *COL12A1* to include isolated clubfoot. Our data provide additional evidence supporting a role for posterior *HOX* genes in clubfoot aetiology. Reported variants are identified in both Caucasian and non-Caucasian probands, suggesting our findings may apply to general populations.

DISCUSSION

The genetic aetiology of clubfoot is heterogeneous

Limb development requires complex spatial and temporal coordination of bone, tendon and muscle which may explain the genetic heterogeneity underlying the pathogenesis of clubfoot. This genetic heterogeneity is highlighted by the clubfoot disease gene, *PITX1*, in which variants were found in only 8 out of 1190 probands in this cohort (online supplemental table 3). Moreover, rare and predicted deleterious variants in *PITX1*, *LMX1B*, *COL9A3*, *COL12A1*, *HOXD12* and other posterior *HOX* genes are present in only 8.4% (100/1190) of our clubfoot cases (online supplemental tables 3 and 4), highlighting the genetic heterogeneity of this congenital phenotype. Many of the variants we identified segregate with incomplete penetrance, therefore identification of genetic modifiers and additional risk factor genes will require even larger datasets. While genome-wide association studies have not shown evidence of common variants of large effect, more work remains to be done to generate polygenic risk scores for clubfoot.

Posterior *HOX* genes and clubfoot aetiology

We report here the first description of *HOXD12* as a novel disease gene for clubfoot based on its enrichment in our rare variant burden analysis and segregation in four multiplex families. Interestingly, *Hoxd12* knockout mice have multiple limb defects, including abnormal morphology or length of carpal, phalanx, radius, ulna, fibula, tibia, metacarpal, metatarsal bones and digits, with majority of these being in the recessive condition.[29 30] *In vitro* reporter or binding assay may provide a better understanding for how these *HOXD12* variants alter the function of protein, as haploinsufficiency is unlikely due to the low pLI score.

Several studies have previously linked dysfunction in posterior *HOX* genes to clubfoot, including microdeletions of *5' HOXC* genes,[7] and common variants in *HOXD13*.[40] *HOX* genes are located in 4 clusters (*HOXA*, *HOXB*, *HOXC*, *HOXD*) and each can be further divided into 13 gene paralogs (*HOX1–13*) with similar expression patterns and functional redundancy among paralogs.[31] The co-linear arrangement of these clusters on the chromosomes correspond to their expression in anterior-posterior body axis for all *HOX* genes and the proximo-distal axis of limb for the posterior *HOX* genes (*HOX9–13*).[31] In our current analysis, we identified an overall enrichment of ultra-rare variants in posterior *HOX* genes in clubfoot cases, consistent with the expression and function of these genes.

Collagen genes and clubfoot aetiology

The collagen superfamily consists of 28 genes characterised by triple-helical domains with multiple Gly-X-Y triplet repeats (X is often a proline and Y is often a hydroxyproline) that function as extracellular matrix proteins.[41] Fibrillar collagens have one major triple-helical domain while non-fibrillar collagens have multiple. Previously, differential composition of extracellular matrix proteins, including the major collagen type I and III as well as collagen XII (COL12A1),[42] were observed between contracted tissue of clubfoot, non-contracted tissue of clubfoot and control samples. Genetically, common SNPs in *COL9A1* were reported to associate with congenital clubfoot.[43]

In our burden analysis, rare and deleterious variants in three non-fibrillar collagen genes, *COL9A3*, *COL12A1* and *COL15A1*, were enriched in clubfoot cases. Both *COL9A3* and *COL12A1* belong to fibril-associated collagens with interrupted triple helices (FACITs), which associate with fibril collagen and may affect the interaction between fibril collagen and other matrix proteins, while *COL15A1* is a member of multiple triple helix domains with interruptions (multiplexin).[39 41] *COL9A3* is an autosomal dominant disease gene responsible for multiple epiphysial dysplasia with limb defects.[34 36] Interestingly, the previously described variants in *COL9A3* all consist of deletions, splice site or nonsense whereas 14 of 16 variants in our clubfoot cohort were missense variants, including 6 glycine substitutions, suggesting a genotype-phenotype correlation. Common genetic variants in another alpha chain, *COL9A1*, are associated with clubfoot.[43] Therefore, abnormalities in alpha chains of the major collagen component for hyaline cartilage may lead to clubfoot and hyaline cartilage proteins may be candidates for future analysis.

While there are rare reports of *COL12A1* causing both autosomal dominant and recessive myopathies,[35 37] we identified more rare variants in our clubfoot cases than have previously been reported in myopathies, suggesting that it may play an even more important role in clubfoot pathogenesis. Interestingly, we found an enrichment of variants in the unique region of the long *COL12A1* isoform that is predominantly expressed in early embryonic stage[39] in cases when compared with gnomAD controls ($p<0.05$). This may explain the earlier onset clubfoot phenotype, as individuals with Ullrich congenital muscular dystrophy and Bethlem myopathy have progressive disease with onset in childhood or as adults. Larger studies are needed to replicate our result.

The third collagen gene identified in our clubfoot burden analysis is *COL15A1*. However, we do not have enough evidence to strongly support the role of *COL15A1* in clubfoot aetiology.

Other potential candidates from gene burden analysis

Since the gnomAD database only provides the information of ACs for each variant, we do not have the information of compound heterozygosity to explore a recessive disease model.

Nevertheless, several enriched genes in our burden analysis are autosomal recessive disease genes, including *SLC26A2* and *HSPG2*, which have clubfoot as one of the phenotypes (online supplemental table 2). Diseases caused by dysfunction of *SLC26A2*[28] or *HSPG2*[25] range from mild to severe phenotypes depending on the protein domains, variant types and effect of the variant. Additional study is needed to investigate the combination of a less deleterious or more common allele with a deleterious and rare allele, as well as if these genes may contribute to a milder, or incompletely penetrant clubfoot phenotype with autosomal dominant inheritance.

Limitations of this study

Although this study included exome data from 1190 individuals with clubfoot and their family members, this is a relatively small dataset for a rare variant gene burden analysis. Thus, we may have insufficient power to detect clubfoot genes with few rare variants under our selection conditions. To address this limitation, less stringent variant filters may be applied to increase the number of variants. However, we performed our analysis under the hypothesis that extremely rare variants would segregate with complete or nearly complete penetrance within families, as shown in *HOXD12*. An alternative approach is to collapse variants in related genes for a gene group burden analysis, as demonstrated by the enrichment of rare variants in posterior *HOX* genes in clubfoot cases. Another limitation of our study is that we excluded variants on the sex chromosomes, which may provide insight into the skewed gender ratio for clubfoot. However, we would need an alternative source of controls with sex information as well as a different gene burden method to complete that analysis.

Overall, our rare variant burden analysis and segregation analysis in multiplex families supports *HOXD12*, *COL12A1*, *LMX1B* and *COL9A3* as clubfoot disease genes. Our results indicate that congenital clubfoot is genetically heterogeneous, with *HOXD12* being a new causative clubfoot disease gene and phenotypic expansion of known musculoskeletal disease genes to include non-syndromic clubfoot. Clinical genetic testing of patients with clubfoot can therefore be expected to yield important information regarding recurrence risk and phenotypic spectrum.

Acknowledgements We thank the patients and their family members who participated in this study. We also thank Clubfoot Cuties for their generous support of this work.

Contributors Conceptualisation: W-LC, CAG; data curation: W-LC; formal analysis: W-LC; funding acquisition: MBD, CAG; investigation: W-LC, MN, IS, CAG; methodology: W-LC, MS, NSJ, AEJ, CAG; resources: MBD, CAG; software: W-LC; validation: MN, IS; visualisation: W-LC, MN, CAG; writing—original draft: W-LC, CAG; writing—review and editing: W-LC, MN, IS, MS, AEJ, MBD, CAG; guarantor: CAG.

Funding Research reported in this publication was supported by National Institute of Arthritis and Musculoskeletal and Skin Diseases (R01AR067715), Eunice Kennedy Shriver National Institutes of Child Health and Human Development of the National Institutes of Health (P01HD084387), Washington University Institute of Clinical and Translational Sciences grant UL1 TR002345 from the National Center for Advancing Translational Sciences of the National Institutes of Health, the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award Number P50HD103525 to the Intellectual and Developmental Disabilities Research Center at Washington University. W-LC was supported by the National Institute of Mental Health of the National Institutes of Health (T32-MH014677).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the institutional review board of Washington University in St. Louis (IRB project number 201102118). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

ORCID iD

Wu-Lin Charng http://orcid.org/0000-0003-0661-4081

REFERENCES

1 Sadler B, Gurnett CA, Dobbs MB. The genetics of isolated and syndromic clubfoot. *J Child Orthop* 2019;13:238–44.
2 Smythe T, Kuper H, Macleod D, *et al*. Birth prevalence of congenital talipes equinovarus in low- and middle-income countries: a systematic review and meta-analysis. *Trop Med Int Health* 2017;22:269–85.
3 Lochmiller C, Johnston D, Scott A, *et al*. Genetic epidemiology study of idiopathic talipes equinovarus. *Am J Med Genet* 1998;79:90–6.
4 Gurnett CA, Boehm S, Connolly A, *et al*. Impact of congenital talipes equinovarus etiology on treatment outcomes. *Dev Med Child Neurol* 2008;50:498–502.
5 de Andrade M, Barnholtz JS, Amos CI, *et al*. Segregation analysis of idiopathic talipes equinovarus in a Texan population. *Am J Med Genet* 1998;79:97–102.
6 Rebbeck TR, Dietz FR, Murray JC, *et al*. A single-gene explanation for the probability of having idiopathic talipes equinovarus. *Am J Hum Genet* 1993;53:1051–63.
7 Alvarado DM, McCall K, Hecht JT, *et al*. Deletions of 5′ *HOXC* genes are associated with lower extremity malformations, including clubfoot and vertical talus. *J Med Genet* 2016;53:250–5.
8 Alvarado DM, Aferol H, McCall K, *et al*. Familial isolated clubfoot is associated with recurrent chromosome 17Q23.1Q23.2 microduplications containing TBX4. *Am J Hum Genet* 2010;87:154–60.
9 Gurnett CA, Alaee F, Kruse LM, *et al*. Asymmetric lower-limb malformations in individuals with homeobox *PITX1* gene Mutation. *Am J Hum Genet* 2008;83:616–22.
10 Yang H, Zheng Z, Cai H, *et al*. Three novel Missense mutations in the filamin B gene are associated with isolated congenital talipes equinovarus. *Hum Genet* 2016;135:1181–9.
11 Quiggle A, Charng W-L, Antunes L, *et al*. Whole exome sequencing in individuals with idiopathic clubfoot reveals a recurrent filamin B (*FLNB*) deletion. *Clin Orthop Relat Res* 2022;480:421–30.
12 Sadler B, Haller G, Antunes L, *et al*. Rare and *de Novo* duplications containing *SHOX* in clubfoot. *J Med Genet* 2020;57:851–7.
13 Van der Auwera GA, Carneiro MO, Hartl C, *et al*. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;43:11.
14 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
15 Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative Pathogenicity of human genetic variants. *Nat Genet* 2014;46:310–5.
16 Jagadeesh KA, Wenger AM, Berger MJ, *et al*. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 2016;48:1581–6.
17 Ioannidis NM, Rothstein JH, Pejaver V, *et al*. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 2016;99:877–85.
18 Adzhubei IA, Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–9.
19 Landrum MJ, Lee JM, Benson M, *et al*. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;46:D1062–7.
20 Li Q, Wang K. Intervar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am J Hum Genet* 2017;100:267–80.
21 Karczewski KJ, Francioli LC, Tiao G, *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–43.
22 Manichaikul A, Mychaleckyj JC, Rich SS, *et al*. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010;26:2867–73.
23 Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–75.

24 Guo MH, Plummer L, Chan Y-M, *et al*. Burden testing of rare variants identified through exome sequencing via publicly available control data. *Am J Hum Genet* 2018;103:522–34.
25 Arikawa-Hirasawa E. Impact of the heparan sulfate proteoglycan perlecan on human disease and health. *Am J Physiol Cell Physiol* 2022;322:C1117–22.
26 McIntosh I, Dreyer SD, Clough MV, *et al*. Mutation analysis of *LMX1B* gene in nail-Patella syndrome patients. *Am J Hum Genet* 1998;63:1651–8.
27 Rieubland C, Jacquemont S, Mittaz L, *et al*. Phenotypic and molecular characterization of a novel case of dyssegmental dysplasia, Silverman-Handmaker type. *Eur J Med Genet* 2010;53:294–8.
28 Rossi A, Superti-Furga A. Mutations in the diastrophic dysplasia sulfate transporter (DTDST) gene (*Slc26A2*): 22 novel mutations, mutation review, associated skeletal phenotypes, and diagnostic relevance. *Hum Mutat* 2001;17:159–71.
29 Cho K-W, Kim J-Y, Cho J-W, *et al*. Point Mutation of *Hoxd12* in mice. *Yonsei Med J* 2008;49:965–72.
30 Davis AP, Capecchi MR. A mutational analysis of the 5′ *HoxD* genes: dissection of genetic interactions during limb development in the Mouse. *Development* 1996;122:1175–85.
31 Pineault KM, Wellik DM. Hox genes and limb musculoskeletal development. *Curr Osteoporos Rep* 2014;12:420–7.
32 Alvarado DM, Buchan JG, Frick SL, *et al*. Copy number analysis of 413 isolated talipes equinovarus patients suggests role for transcriptional regulators of early limb development. *Eur J Hum Genet* 2013;21:373–80.
33 Dreyer SD, Zhou G, Baldini A, *et al*. Mutations in *LMX1B* cause abnormal skeletal patterning and renal dysplasia in nail patella syndrome. *Nat Genet* 1998;19:47–50.
34 Bönnemann CG, Cox GF, Shapiro F, *et al*. A Mutation in the alpha 3 chain of type IX collagen causes autosomal dominant multiple epiphyseal dysplasia with mild myopathy. *Proc Natl Acad Sci U S A* 2000;97:1212–7.
35 Hicks D, Farsani GT, Laval S, *et al*. Mutations in the collagen XII gene define a new form of extracellular matrix-related myopathy. *Hum Mol Genet* 2014;23:2353–63.
36 Paassilta P, Lohiniva J, Annunen S, *et al*. COL9A3: a third locus for multiple epiphyseal dysplasia. *Am J Hum Genet* 1999;64:1036–44.
37 Zou Y, Zwolanek D, Izu Y, *et al*. Recessive and dominant mutations in *COL12A1* cause a novel EDS/myopathy overlap syndrome in humans and mice. *Hum Mol Genet* 2014;23:2339–52.
38 Nixon TRW, Alexander P, Richards A, *et al*. Homozygous type IX collagen variants (*Col9A1*, *Col9A2*, and *Col9A3*) causing recessive stickler syndrome-expanding the phenotype. *Am J Med Genet A* 2019;179:1498–506.
39 Chiquet M, Birk DE, Bönnemann CG, *et al*. Collagen XII: protecting bone and muscle integrity by organizing collagen fibrils. *Int J Biochem Cell Biol* 2014;53:51–4.
40 Wang L, Jin C, Liu L, *et al*. Analysis of association between 5′ *HOXD* gene and idiopathic congenital talipes equinovarus. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* 2005;22:653–6.
41 Ricard-Blum S. The collagen family. *Cold Spring Harb Perspect Biol* 2011;3:a004978.
42 Eckhardt A, Novotny T, Doubkova M, *et al*. Novel contribution to clubfoot pathogenesis: the possible role of extracellular matrix proteins. *J Orthop Res* 2019;37:769–78.
43 Liu L-Y, Jin C-L, Cao D-H, *et al*. Analysis of association between *COL9A1* gene and idiopathic congenital talipes equinovarus. *Yi Chuan* 2007;29:427–32.