Idiopathic Scoliosis Families Highlight Actin-Based and Microtubule-Based Cellular Projections and Extracellular Matrix in Disease Etiology

*Department of Orthopedics, University of Colorado Anschutz Medical Campus, Aurora, CO, †Musculoskeletal Research Center, Children’s Hospital Colorado, Aurora, CO, ‡Department of Cardiology, §Department of Pediatrics, and **Department of Cell and Developmental Biology, University of Colorado Anschutz Medical Campus, Aurora, CO, and ††Department of Molecular, Cellular and Developmental Biology, University of Colorado Boulder, Boulder, CO

ABSTRACT Idiopathic scoliosis (IS) is a structural lateral spinal curvature of \( \geq 10^\circ \) that affects up to 3% of otherwise healthy children and can lead to life-long problems in severe cases. It is well-established that IS is a genetic disorder. Previous studies have identified genes that may contribute to the IS phenotype, but the overall genetic etiology of IS is not well understood. We used exome sequencing to study five multigenerational families with IS. Bioinformatic analyses identified unique and low frequency variants (minor allele frequency \( \leq 5\% \)) that were present in all sequenced members of the family. Across the five families, we identified a total of 270 variants with predicted functional consequences in 246 genes, and found that eight genes were shared by two families. We performed GO term enrichment analyses, with the hypothesis that certain functional annotations or pathways would be enriched in the 246 genes identified in our IS families. Using three complementary programs to complete these analyses, we identified enriched categories that include stereocilia and other actin-based cellular projections, cilia and other microtubule-based cellular projections, and the extracellular matrix (ECM). Our results suggest that there are multiple paths to IS and provide a foundation for future studies of IS pathogenesis.

KEYWORDS idiopathic scoliosis exome sequencing actin cytoskeleton microtubules cilia extracellular matrix

Idiopathic scoliosis (IS) is a common disorder of the immature skeleton that affects up to 3% of the pediatric population, with girls often more severely affected than boys (Asher and Burton 2006). IS is defined clinically as a structural lateral spinal curvature of \( \geq 10^\circ \) with a rotatory component, documented by radiologic analysis and occurring in otherwise healthy children. Current therapeutic options are limited to physical therapy, bracing, and surgery. The only treatment option for severe progressive curves is an operative spinal instrumentation and fusion, a costly procedure with life-long implications. IS is known to have a genetic component, with higher concordance rates in monozygotic twins than in dizygotic twins and an increased risk for first-degree relatives compared to the general population (Cowell et al. 2017; Riseborough and Wynne-Davies 1973; Bonaiti et al. 1976; Czeizel et al. 1978; Kesling and Reinker 1997; Inoue et al. 1998; Aksenovich et al. 1999; Andersen et al. 2007; Ward et al. 2010; Tang et al. 2012).

Genetic studies for IS have been conducted for more than 30 years. Genome-wide linkage studies have resulted in the identification of large chromosomal regions rather than specific genes (Salehi et al. 2002; Chan et al. 2002; Justice et al. 2003; Miller et al. 2005; Gao et al. 2007; Ocaka et al. 2008; Gurnett et al. 2009; Raggio et al. 2009; Edery et al. 2011).
Several genome-wide association studies (GWAS) in different ethnic populations have identified potential genes and regions of interest (Sharma et al. 2011; Takahashi et al. 2011; Kou et al. 2013; Miyake et al. 2013; Ogura et al. 2015; Zhu et al. 2015; Sharma et al. 2015; Chettier et al. 2015; Zhu et al. 2017; Ogura et al. 2017). Both the LBX1 (Takahashi et al. 2011; Fan et al. 2012; Jiang et al. 2013; Gao et al. 2013; Londono et al. 2014; Chettier et al. 2015; Grauers et al. 2015; Zhu et al. 2015; Guo et al. 2016; Cao et al. 2016; Liu et al. 2017; Nada et al. 2018) and GPR126 (Kou et al. 2013; Xu et al. 2015; Karner et al. 2015; Qin et al. 2017) genes have been associated with IS in multiple studies. LBX1 is a transcription factor and core regulator of myoblast development (Masselink et al. 2017), while GPR126 functions within Schwann cells to control differentiation and myelination (Mogha et al. 2013). Recently, exome sequencing has identified genes that may contribute to the IS phenotype, including genes that encode components of the extracellular matrix (ECM), cytoskeleton, cilia, and centrosomes (Buchan et al. 2014a; Baschal et al. 2014; Patten et al. 2015; Li et al. 2016; Oliazadeh et al. 2017; Einarsdottir et al. 2017; Haller et al. 2016). Overall, these genetic findings lead to proposed mechanisms of IS pathogenesis stemming from alterations in the ECM, cilia, muscle, and axon myelination. Even with these findings, the biological basis of IS is not well understood and the genetic contributions are still poorly defined.

With the discovery of multiple genes that may contribute to IS etiology, it has become apparent that IS is likely a polygenic disease (Kruse et al. 2012; Haller et al. 2016). In this study, we completed exome sequencing in five multigenerational families to identify variants present in all affected family members, and then used Gene Ontology (GO) term enrichment analyses of these variant lists to identify functional categories that we predict are important for IS etiology.

**MATERIALS AND METHODS**

**Subjects**

Individuals were enrolled as described previously (Baschal et al. 2014). A diagnosis of IS required a standing anteroposterior spinal radiograph showing ≥10° curvature by the Cobb method with pedicle rotation, and no congenital deformity or other co-existing genetic disorder (Shands and Eisberg 1955; Kane 1977; Armstrong et al. 1982). If two or more individuals in a family were affected with IS using these criteria, those individuals were classified as having familial IS.

Written informed consent was obtained from study subjects who were enrolled in accordance with protocols approved by the Johns Hopkins School of Medicine Institutional Review Board and the University of Colorado Anschutz Medical Campus Institutional Review Board (Colorado Multiple Institutional Review Board, Study #06-1161 and 07-0417). All procedures involving human participants were performed in accordance with the ethical standards of these institutional review boards, the 1964 Declaration of Helsinki and its later amendments, or comparable ethical standards.

We collected blood samples from all participants and extracted genomic DNA using the QIAGEN Gentra Puregene Blood Kit or standard phenol-chloroform purification protocols (Sambrook et al. 1989; Moore and Dowhan 2002). In the event a blood sample was not available, a saliva sample was collected using the Oragene OGR-250 kit and extracted according to the manufacturer's protocol.

Five multigenerational IS families with European ancestry were selected for exome sequencing. These families were selected based on the number of affected individuals in a family, scoliosis curve severity, and the degree of relationship to the proband. Where possible, distantly related individuals in each family were selected for exome sequencing.

Exome sequencing was completed for three to four individuals from each family, for a total of 16 sequenced individuals from these five families. Clinical details and family relationships are presented in Table 1, and pedigrees are provided in File S1. In some cases, pedigrees were simplified to protect the identity of the study participants. Individuals selected for exome sequencing were affected with IS, with the exception of one individual from Family B. In this family, we sequenced one unaffected individual (II-2) who is expected to carry any variants that contributed to the IS phenotype within the family. In addition to her grandchildren who were sequenced in the study (IV-3 and IV-5), her father (I-1), brother (II-5), and two nieces (III-6 and III-7) are affected with IS, but DNA was not available for these individuals.

**Exome Sequencing**

Exome capture was completed using 1 μg of genomic DNA from 16 individuals across five families using the Illumina TruSeq Exome kit. Samples were sequenced with a 2 × 100 bp run on the Illumina HiSeq 2000 at the University of Colorado Denver Genomics and Microarray Core Facility with three samples multiplexed per lane.

**Bioinformatic Filtering**

The reads were aligned to the human genome assembly GRCh38 using GSNAP (Genomic Short-read Nucleotide Alignment Program, version 2014-12-17) (Wu and Nacu 2010) and variants were identified by FreeBayes (v1.0.1-2-g0ch2697) (Garrison and Marth 2012). The candidates were filtered by SnpEff (version 4.1g) (Cingolani et al. 2012) and custom scripts to retain only non-synonymous SNPs, coding indels, and variants affecting splice sites. These were also stripped of known artifacts and variants whose frequency was greater than 5% in the ExAC database (r0.3) (Lek et al. 2016). If the variant was annotated in the dbNSFP database (version 3.0) (Liu et al. 2011; Liu et al. 2016), it was retained only if at least one of the prediction algorithms (SIFT, Polyphen2, LRT, MutationTaster) scored it as “damaging,” signifying that the resulting change to the protein had a predicted functional consequence.

Within each family, variants were required to be shared by all individuals that underwent exome sequencing. If any individuals in the family were missing data at the position of interest, the variant was retained in the list as long as the alleles were shared by the remaining family members. This filter eliminated variants that did not segregate with the IS phenotype in each family, which reduced the number of false positive variants in our results.

Gene values for pLI and pRec were obtained from the Functional Gene Constraint download from ExAC (fordist_cleaned_exac_r03_march16_z_pli_rec_null_data.txt, modified 2016-01-13, 5:00:00 PM) (Samocha et al. 2014; Lek et al. 2016).

**Functional Category Annotations**

Functional annotations for the GO terms “actin cytoskeleton” and “microtubule cytoskeleton” listed in File S2 were obtained from DAVID on 2018-05-21. Annotations for “extracellular matrix” were obtained from the Core Matrisome list (Naba et al. 2012).

Cilia genes were annotated using a pipeline that initially focused on two cilia gene databases, SYSCILIA Gold Standard (http://syscilia.org [van Dam et al. 2013]) and the Centrosome and Cilium Database (CCDB [Gupta et al. 2015]). We also included annotations from two cilia-related Cellular Component GO terms (“cilium” and “microtubule cytoskeleton”). All potential cilia genes were manually investigated through searches in UniProt (The UniProt Consortium 2017) and PubMed (“cilia” + [gene name]). Genes were classified as encoding cilia
Table 1 Clinical information for subjects who underwent exome sequencing. Five families underwent exome sequencing, with three to four individuals sequenced in each family. Gender, curve measurement (Cobb angle), relationship to proband, degree relationship to proband, and sequencing coverage across the exome are presented for each individual. Double curves are represented by a slash. Pedigrees are provided in Supplemental File 1.

<table>
<thead>
<tr>
<th>Family</th>
<th>Individual</th>
<th>Gender</th>
<th>Curve (Degrees)</th>
<th>Relationship to Proband</th>
<th>Degree Relationship to Proband</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>III-4</td>
<td>F</td>
<td>52/81</td>
<td>proband</td>
<td>proband</td>
<td>91X</td>
</tr>
<tr>
<td>A</td>
<td>III-5</td>
<td>F</td>
<td>76/65</td>
<td>sister</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>III-1</td>
<td>M</td>
<td>25</td>
<td>half cousin (paternal)</td>
<td>1st degree</td>
<td>51X</td>
</tr>
<tr>
<td>A</td>
<td>II-4</td>
<td>F</td>
<td>unknown</td>
<td>aunt (paternal)</td>
<td>2nd degree</td>
<td>49X</td>
</tr>
<tr>
<td>B</td>
<td>IV-3</td>
<td>F</td>
<td>34/28</td>
<td>proband</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>IV-5</td>
<td>M</td>
<td>56</td>
<td>cousin (maternal)</td>
<td>3rd degree</td>
<td>44X</td>
</tr>
<tr>
<td>B</td>
<td>II-2</td>
<td>F</td>
<td>0</td>
<td>grandmother (maternal)</td>
<td>2nd degree</td>
<td>63X</td>
</tr>
<tr>
<td>C</td>
<td>IV-1</td>
<td>F</td>
<td>36/20</td>
<td>proband</td>
<td></td>
<td>49X</td>
</tr>
<tr>
<td>C</td>
<td>III-6</td>
<td>F</td>
<td>45</td>
<td>first cousin once removed (maternal)</td>
<td>4th degree</td>
<td>52X</td>
</tr>
<tr>
<td>C</td>
<td>IV-2</td>
<td>M</td>
<td>35/70</td>
<td>second cousin (maternal)</td>
<td>5th degree</td>
<td>38X</td>
</tr>
<tr>
<td>D</td>
<td>III-1</td>
<td>F</td>
<td>32/47</td>
<td>proband</td>
<td></td>
<td>52X</td>
</tr>
<tr>
<td>D</td>
<td>III-5</td>
<td>F</td>
<td>78/42</td>
<td>cousin (paternal)</td>
<td>3rd degree</td>
<td>44X</td>
</tr>
<tr>
<td>D</td>
<td>II-5</td>
<td>F</td>
<td>12/8</td>
<td>aunt (paternal)</td>
<td>2nd degree</td>
<td>49X</td>
</tr>
<tr>
<td>E</td>
<td>IV-1</td>
<td>F</td>
<td>60/60</td>
<td>proband</td>
<td></td>
<td>55X</td>
</tr>
<tr>
<td>E</td>
<td>III-3</td>
<td>F</td>
<td>50</td>
<td>aunt (paternal)</td>
<td>2nd degree</td>
<td>40X</td>
</tr>
<tr>
<td>E</td>
<td>II-3</td>
<td>F</td>
<td>22/35</td>
<td>great aunt (paternal)</td>
<td>3rd degree</td>
<td>38X</td>
</tr>
</tbody>
</table>

proteins (“yes”) if the gene was listed in at least one of the cilia gene databases, or had published evidence localizing the protein product to the cilium. Genes were classified as potential cilia genes (“potential”) if published evidence suggested the protein interacted with the cilium, but was not confirmed to be related to cilia function. Genes were not classified as cilia genes if no published evidence suggested ciliary involvement, even if they were associated with the GO term “cilium”. Final determinations were made through review of the identified articles. Genes annotated as “stereocilia” were identified from the “stereocilium” GO term and also through review of articles identified in the literature.

GO Term Enrichment Analyses

**PANTHER:** Gene Ontology (GO) term analysis was completed using the PANTHER website (Mi and Thomas 2009; Mi et al. 2013; Mi et al. 2017). A combined variant list was generated that contained all of the variants that were identified in any of our five families (variants were filtered as described above). The corresponding gene names were used as the input for PANTHER (270 variants in 246 genes), which resulted in 251 PANTHER IDs. Statistical overrepresentation tests were performed on 2017-11-28 using the PANTHER website, with PANTHER overrepresentation test release 20170413 and Gene Ontology database release 2017-10-23. These tests used a Bonferroni multiple testing correction for the Cellular Component GO term class. The p-values were also obtained without Bonferroni correction and those results are indicated in the text and tables where appropriate.

**DAVID:** DAVID 6.8 was used to identify the significant GO terms and clusters in our dataset (Huang da et al. 2009b, 2009a). The same input gene list was used for DAVID as for PANTHER (246 genes), which resulted in 242 DAVID IDs. Functional annotation clustering was used on our dataset with custom settings and the GORTERM_CC_All annotation category. DAVID clustering is based on the genes in the dataset of interest that are shared among the GO terms. The clustering stringency settings were relaxed from default values for this analysis (both Initial Group Membership and Final Group Membership were set to 2). Additionally, the EASE threshold was set to 0.1, to match the default threshold used by DAVID for single GO term chart analyses. The EASE score can be interpreted like a p-value. Default values were used for the other settings (Similarity Term Overlap = 3, Similarity Threshold = 0.50, Multiple Linkage Threshold = 0.5). The enrichment score for the cluster is based on the EASE score for each term member (negative log scale), where higher numbers represent clusters that are more enriched.

**BiNGO:** The Biological Networks Gene Ontology tool (BiNGO) is an open-source Java tool and Cytoscape plugin that allows for the determination of GO terms that are significantly overrepresented in a set of genes and provides a visual representation of the results (Maere et al. 2005; Shannon et al. 2003). The same input gene list was used for BiNGO as for PANTHER and DAVID (246 genes), which resulted in 196 BiNGO IDs. An overrepresentation binomial test was used, with visualization of the overrepresented categories before correction. We downloaded the ontology file go.obo from the GO website (data-version: releases/2017-11-25, CVS version 38972) (Ashburner et al. 2000; The Gene Ontology Consortium 2017). The BiNGO analysis used the reference set whole annotation, with the downloaded go.obo file and namespace Cellular_component. A significance level of 0.05 was used, with multiple testing correction, and first degree relatives of significant GO terms were added to the visualization to give a clearer view of how the significant GO terms were related. A custom color gradient was used for the nodes, where blue is more highly significant (P = 1.0E-03) and teal is less significant (P = 0.05). White represents the non-significant first degree relatives. The size of the node is proportional to the number of genes in the input dataset that are annotated to that GO term. Initially, the Prefuse Force Directed Layout was used, but the terms were subsequently manually rearranged in an attempt to minimize the overlap and to create clusters or groups of related terms. GIMP v2.8 (http://gimp.org) was used to add text labels for the overall groups.

**Availability of Data Statement**

The authors affirm that all data necessary for confirming the conclusions of this article are represented fully within the article and its supplementary files, including the complete lists of filtered variants for each family. Supplemental material available at Figshare: https://doi.org/10.25387/g3.6104780.

**RESULTS**

We studied five multigenerational families of European ancestry affected with familial IS (Table 1). Affected individuals were diagnosed with IS
Table 2 Functional annotation clusters identified by DAVID. Functional annotation clustering analysis was completed as described in the Methods. In brief, the GOTERM_CC_ALL annotation category was used, with an input of the 246 genes identified in our dataset. Five clusters were identified, and higher enrichment scores represent clusters that are more enriched. Detailed results, including underlying genes for each GO term, are presented in Supplemental File 4.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Enrichment Score</th>
<th>GO Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>2.47084939</td>
<td>stereocilium, stereocilium bundle, actin-based cell projection, cluster of actin-based cell projections, brush border, nonmotile primary cilium, primary cilium, microvillus</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.549760389</td>
<td>extracellular matrix component, complex of collagen trimers, proteinaceous extracellular matrix</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>1.470776063</td>
<td>cell projection, cilium, cell projection part, cilary part</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>1.268556732</td>
<td>intrinsic component of plasma membrane, integral component of plasma membrane, plasma membrane part</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>1.058855585</td>
<td>synapse, neuron part</td>
</tr>
</tbody>
</table>

Based on radiographic findings of a spinal curvature ≥10°. Exome sequencing, followed by variant detection and filtering, was completed for 16 individuals (3 to 4 individuals per family). Sequencing coverage ranged from 36 to 91X, with an average of 53X. Filtering resulted in a list of variants for each family, where each variant was an insertion/deletion, nonsense, missense, or splice-site change with a predicted functional consequence and was present at a minor allele frequency (MAF) of 5% or less in the ExAC database. We used an MAF cut-off of 5% due to the relatively high prevalence of IS in the general population (up to 3% when defining scoliosis as a Cobb angle ≥10°). In addition, to be considered for further study, we required variants to be present in all sequenced individuals in the family, as described in detail in the Methods section.

A total of 270 variants in 246 genes were identified across the five families after this filtering process, with a range of 19 to 89 variants identified in each family (File S2). Eight genes were shared by two families (XRR1A, ANKR11A, ANKR30B, DMD1, ERN2, MYBPC3, NPY4R and TNFRSF10C) and no genes were shared by three or more families. Even with our stringent familial analysis that required the identified variants to be present in every sequenced family member, a large number of potentially causal variants remained in each family.

To further understand and interpret these large variant lists, we completed Gene Ontology (GO) term enrichment analyses. These analyses allowed us to investigate the hypothesis that IS is caused by variants in multiple genes, and that these genes will show enrichments of certain functional annotations or pathways. We used the programs PANTHER, DAVID, and BiNGO to complete these analyses and explore this hypothesis in relation to our data.

We completed PANTHER GO term overrepresentation analysis, which allowed us to compare the number of genes in our dataset annotated with each GO term to the number expected by chance. The input was our combined gene list from all five families (246 genes). The only cellular component GO term in our dataset that passed significance with a Bonferroni multiple testing correction was “stereocilium” (P = 0.0209 with Bonferroni correction, P = 1.55E-05 without Bonferroni correction, 11.70 fold enrichment, n = 6 genes), which is an actin-based cellular projection that is important for both hearing and balance.

Two major themes emerged when we examined the other enriched GO terms that were significant without Bonferroni multiple testing correction (File S3). The prominent theme from the PANTHER results are terms related to cellular projections, which represent 18 of the top 20 most significant GO terms without Bonferroni correction. The two terms in the top 20 that are not specifically related to cellular projections are “phagocytic vesicle lumen” (P = 6.23E-04, 55.89 fold enrichment, n = 2 genes) and “intrinsic component of plasma membrane” (P = 6.20E-03, 1.59 fold enrichment, n = 32 genes). Both actin-based and microtubule-based cellular projections are included in the top 20 GO terms. Actin-based cell projections in this list include two terms related to the stereocilium and two terms related to general actin-based cellular projections. The microtubule-based cellular projection terms include seven cilia-related terms, including “cilary part” (P = 4.69E-03, 2.42 fold enrichment, n = 12 genes) and “axoneme part” (P = 8.01E-04, 9.86 fold enrichment, n = 4 genes). The term “kinocilium” (P = 5.35E-03, 18.63 fold enrichment, n = 2 genes) is also present in the top 20 PANTHER list. Kinocilia are a specialized type of cilium present in vestibular hair cells, and work with stereocilia to sense and respond to spatial orientation. The term “neuron projection” (P = 2.15E-03, 1.87 fold enrichment, n = 25 genes) is also present in the top 20 list, and includes projections that are composed of both actin and microtubules. This cell projection theme is also carried beyond the top 20 GO terms.

The second major theme from our PANTHER results (beyond the top 20 terms) is the extracellular matrix (ECM), with specific terms of “proteinaceous ECM” (P = 1.36E-02, 2.28 fold enrichment, n = 10 genes) and “ECM component” (P = 1.62E-02, 3.44 fold enrichment, n = 5 genes). Additionally, our PANTHER results include multiple terms related to collagen and collagen trimers, which are major components of the ECM. Overall, we identified themes of related GO terms from our PANTHER results, including cellular projections and ECM.

The DAVID Functional Annotation Clustering Tool is an unbiased method for clustering significantly enriched GO terms. This tool works by identifying the genes in our dataset that are shared among the GO terms. We applied this algorithm to our data, again using the Cellular Component GO terms. DAVID identified clusters in our data that loosely correlate with stereocilium/actin-based cell projections/primary cilium (Annotation Cluster 1), extracellular matrix/collagen (Cluster 2), cell projection/cilium (Cluster 3), plasma membrane components (Cluster 4), and neuron part/synapse (Cluster 5) (Table 2 and File S4).

We used BiNGO visualization software to further investigate and understand these clusters. Overall, we found that there were nine overall “groups” of GO terms that are loosely clustered (Figure 1). The raw data are reported in File S6. We manually annotated these nine groups as GO terms related to muscle, ECM/collagen, microtubule cytoskeleton, cilium, stereocilium, neuron/synapse, plasma membrane, mitochondrial membrane, and nucleus. The genes annotated to the significant GO terms included in each of the nine overall groups are listed in Table 3. Of note, only one gene underlies the significant result in three of the nine groups (muscle, mitochondrial membrane, and nucleus). However, the significant results in groups related to our major themes (ECM/collagen, microtubule cytoskeleton, cilium, stereocilium, and neuron/synapse) are driven by at least four genes. These overall results and themes are consistent with the manual clustering we used for our PANTHER results and the algorithmic clustering used by DAVID.

As cilia have been described as a potential factor in IS etiology (Grimes et al. 2016; Patten et al. 2015; Buchan et al. 2014b; Andersen et al. 2017; Einarsdottir et al. 2017), we further investigated the cilia connections identified in our dataset. Cilia components and functions
are still under active exploration by many groups, and as a result the cilia-related GO term annotations are not always up to date. Therefore, we chose to use a different pipeline for annotating our cilia genes to generate a more inclusive list. We compared our gene list to the cilia gene databases (SYSCILIA [van Dam et al. 2013] and CCDB [Gupta et al. 2015]) and cilia-related GO terms, and also incorporated confirmatory literature searches. These annotations are included in File S2.

We then used literature searches to annotate the location within the cilium for each cilia protein identified in our dataset (Table 4 and File S5). Overall, we have found that 6.3 to 8.8% of the genes identified in our five IS families encode proteins that localize to the cilium.

We next investigated our results at the level of individual IS families, rather than combining the results from all five families together. We compared each family’s gene list with the three functional categories identified in our analyses (stereocilia and other actin-based cellular projections, cilia and other microtubule-based projections, and the ECM, File S2). Three families had at least one gene annotated to each of these three functional categories (Families A, B, and E). Family C had genes annotated to actin cytoskeleton and microtubule cytoskeleton, but did not have an ECM gene. Family D had genes annotated to microtubule cytoskeleton and ECM, but did not have an actin cytoskeleton gene.

When these results are examined together, three themes emerge. Our IS families show enrichments of variants in genes related to stereocilia and other actin-based cellular projections, cilia and other microtubule-based projections, and the ECM. We have identified these three functional categories of genes in both the combined family dataset and in individual families with IS.

**DISCUSSION**

In this study, we report exome sequencing results for five multigenerational IS families. We identified a total of 270 variants in 246 genes across all 5 families. No genes were shared by all families, indicating that there is not a single Mendelian cause for IS within this sample group. We performed GO term enrichment analyses, with the hypothesis that certain functional annotations or pathways would be enriched in the 246 genes identified in our IS families. Using complementary programs to complete these analyses, we identified enriched categories that include stereocilia and other actin-based cellular projections, cilia and other microtubule-based cellular projections, and the extracellular matrix (ECM). These results suggest that familial IS is a polygenic disorder with multiple genes and pathways involved in the pathogenesis of the disease.

Figure 1 BiNGO visualization of Cellular Component GO terms. BiNGO was used to determine and visualize GO terms that are significantly overrepresented in our dataset, as described in detail in the Methods. Blue represents more significant GO terms, teal is \( P = 0.05 \), and white are non-significant first degree relatives. The size of the node is proportional to the number of genes in the dataset that are annotated to that GO term. The data used to generate this figure are presented in Supplemental File 6.

Genes related to actin-based cellular projections were the first enriched category in our dataset. These actin-based structures include stereocilia, neuronal projections, microvilli, and other actin-related genes that are not specific to cellular projections. The only GO term
that passed significance with Bonferroni multiple testing correction in our data were "stereocilium" (PANTHER, P = 2.09E-02, n = 6 genes). Stereocilia are actin-based cellular projections that are important for both the auditory system (hearing) and the vestibular system (balance and spatial orientation). Multiple studies support an association between IS and vestibular dysfunction (Hawasli et al. 2015). Individuals with IS have been described to have differences in vestibular function and/or anatomy (Hawasli et al. 2015) and spatial orientation. Multiple studies support an association between IS and vestibular dysfunction (Hawasli et al. 2015). These functional studies in individuals with IS, compared to controls, although the functional significance of these findings from our group and others are supportive of a functional role for cilia in IS etiology.

Clinical observations also suggest a plausible role for cilia dysfunction in the pathogenesis of IS. Ciliopathies are a diverse class of human diseases that result from defects in both primary and motile cilia (Waters and Beales 2011). Importantly, ciliopathy patients, including those with Joubert, Jeune, Bardet-Biedl, Alström, and primary ciliary dyskinesia (PCD), show an increased incidence of scoliosis (Braganti et al. 2010; O’Brien et al. 2015; Ramirez et al. 2004; Marshall et al. 2011; Engesaeth et al. 1993; Schlösser et al. 2017; Knowles et al. 2013). Furthermore, 55 of the 303 genes from the SYSCILIA annotated list of cilia genes (van Dam et al. 2013) were associated with human syndromes that have clinical reports of scoliosis (Oliazadeh et al. 2017). This overlap suggests shared mechanisms in the etiologies of ciliopathies and IS. Additionally, cilia are crucial for both bone and cartilage development through their

Table 3 BiNGO GO term annotations and gene lists. BiNGO was used to determine and visualize GO terms that are significantly overrepresented in our dataset, as described in detail in the Methods and Results. Nine overall groups of clustered GO terms were manually annotated. This table lists the number of significant GO terms included in the overall group, the number of genes annotated to those significant GO terms, and the corresponding gene names. In addition, three significant GO terms were not clustered with the nine overall groups: non-membrane-bounded organelle, intracellular non-membrane-bounded organelle, and organelle part. The non-redundant list of genes annotated to those three GO terms are included, as are the genes from our five families that were not annotated by BiNGO

<table>
<thead>
<tr>
<th>Overall Group</th>
<th># Significant GO Terms</th>
<th># Genes</th>
<th>Genes Annotated to the Overall Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muscle</td>
<td>2</td>
<td>1</td>
<td>MYBPC3</td>
</tr>
<tr>
<td>ECM/ Collagen</td>
<td>6</td>
<td>8</td>
<td>COL6A3, SERPINA1, TNXB, FLRT2, DST, COL6A2, COL6A5, TINAG</td>
</tr>
<tr>
<td>Microtubule Cytoskeleton</td>
<td>6</td>
<td>12</td>
<td>DNAH8, DNAH6, DNHD1, DST, KIF15, MACF1, SULT1C2, C2CD3, AKAP9, MCM3, NPHP4, CEP290</td>
</tr>
<tr>
<td>Cilium</td>
<td>4</td>
<td>5</td>
<td>PCDH15, C2CD3, CEP290, DNAH8, NPHP4</td>
</tr>
<tr>
<td>Sterocilium</td>
<td>3</td>
<td>4</td>
<td>LOXHD1, PCDH15, USH1C, LRP2</td>
</tr>
<tr>
<td>Neuron/ Synapse</td>
<td>7</td>
<td>15</td>
<td>EPHB6, SELPGL, CSF3R, NPY4R, LRP2, AQP2, ATP12A, EPH8, GRID2, EPHA7, GRD2, FLRT2, STAB1, PCDH4, GRID2, ABCCC1, DST, MGA, SEMA4C, TNFRSF10C, PCDHA11, ATRN, CPEB1, PTPRB, DLG1, GJB4, CCKBR, NEDD4, SCN1A, LSC26A7, GFR1</td>
</tr>
<tr>
<td>Plasma Membrane</td>
<td>4</td>
<td>31</td>
<td>EPB1B, DLG1, GRID2, SEMA4C, EPHA7, LRRK2, SEMA3A, CHAT, PCDH15, USH1C, ARID1A, LOXHD1, SAC5, CEP290</td>
</tr>
<tr>
<td>Mitochondrial Membrane</td>
<td>1</td>
<td>1</td>
<td>LRR2K</td>
</tr>
<tr>
<td>Nucleus</td>
<td>2</td>
<td>1</td>
<td>MLH1</td>
</tr>
<tr>
<td>Three Significant GO Terms</td>
<td>3</td>
<td>47</td>
<td>HIST1H2BJ, CHD9, NOP2, CHD6, NAT10, ZNF22, RRB1, PPM1E, AKAP12, ORC4, MRPL3, CLSPN, LRRFIP1, FARP1, DDX11, KRT2, ELMOD3, DDX54, EBNA1BP2, TADA2A, TS1, VCL, ECE2, GOLGA2, JPH3, QALNT8, TET1, CDC25A, ERN2, NDUF58, MMRN1, MED20, PYYE, UGCRCC2, GOLGA8A, POM121, SLCA2A7, ATXN3, MAN2A2, UBN1, DHX38, SRR1, ATRIP, BSGAT2, DNAJC11, SON, QARS</td>
</tr>
<tr>
<td>That Are Not in the Nine</td>
<td></td>
<td></td>
<td>FASTKD1, ULK4, DDX60L, FAM216A, C16orf71, C6orf232, EVA1A, HENT1, EPG5, IAHI, C12orf58, TET2, RNF208, JOSD2, KIF7, PALD1, ATAD3C, PATL2, FCSD1, THAP3, POM121L2, KIAA1257, CCDC428, METTL20, DXO, NUMT2G, TRABD2A, TRUB1, TRIM51, STAMBPL1, DCCDC28, BCO1, C16orf52, FAM83C, LKR1, HELQ, FBXO39, TYW1B, DNAJC13, FAM111B, BPIFB1, NOL4L, S5T, KANSL3, RHCC3, GNB3, CEP162, KBTBD8, CCDC173</td>
</tr>
<tr>
<td>Overall Groups</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Annotated by BiNGO</td>
<td>N/A</td>
<td>49</td>
<td></td>
</tr>
</tbody>
</table>
roles in mechanosensation, and are also responsible for directional ECM production and endochondral ossification at the growth plate (Seeger-Nukpezah and Golemis 2012). However, it is important to note that IS appears to specifically affect the spine, while most currently defined ciliopathies affect multiple tissues and organ systems. These distinct phenotypes may be explained by differences in either mutation severity or tissue-specific gene expression, which will need to be explored in functional studies.

The third functional category enriched in our data are the ECM (P = 1.36E-02, n = 10 genes), which is clearly evident in our BiNGO clustering (Figure 1). The ECM is a dense network of fibrous proteins that provides structural support to cells and provides tissue organization, and has been identified as important for IS etiology by multiple groups. IS patients have a significant burden of rare variants in the fibrinogen genes (FBN1 and FBN2) [Buchan et al. 2014a], which are important components of the ECM. Additionally, our group previously reported similar results for rare variants in the ECM component perlecan (HSPG2) [Baschal et al. 2014]. Finally, in a larger scale IS study, Haller et al. identified an increased burden of rare variants in ECM genes in IS patients, with a specific association to the musculoskeletal collagen genes. This study found that the risk of IS increases proportionally with the number of rare variants in ECM genes [Haller et al. 2016]. In addition, scoliosis is often a phenotypic element of diseases caused by defects in the ECM, such as Marfan syndrome and several of the Ehlers-Danlos syndromes. Therefore, our results support and expand upon the previously published literature that implicates an involvement of the ECM in IS etiology.

Actin and microtubules are both essential components of neuronal projections, and therefore neuronal projections fit into both the first and second functional categories identified from our data. The GO term “neuron projection” was significantly enriched in our dataset (P = 2.15E-03, n = 25 genes) and this category is evident in our BiNGO clustering (Figure 1). These genes are related to both actin-based and microtubule-based neuronal projections, encoding proteins located in the axon growth cone and dendritic spine, and also encoding proteins involved in axon guidance and neuronal growth. The axon growth cone tip and dendritic spines are actin-rich structures, while the axon growth cone central domain is a microtubule-based structure. Neuronal growth relies heavily on microtubules, and both actin and microtubules are important for the axon guidance process. Previous studies have identified neuronal genes that may be important for IS. A GWAS for IS identified significant associations near the CHLI and DSCAM genes, which both encode axon guidance proteins (Sharma et al. 2011). GPR126, replicated in multiple IS genetic studies (Kou et al. 2013; Xu et al. 2015; Karner et al. 2015; Qin et al. 2017), is important for normal myelination of axons but does not have an obvious role in the cytoskeleton (Mogha et al. 2013). Damaging variants in these genes may lead to mild neurological defects that give rise to IS in some patients.

IS is now believed to be a complex polygenic disease, where multiple variants, acting in a polygenic fashion, are needed to develop the phenotype. We propose two main mechanisms that could underlie the polygenic nature of IS. The first is that more than one variant in a single cellular structure or pathway could be required for IS development (e.g., multiple variants in cilia genes are required in one family and multiple variants in ECM genes are required in another family). A second possibility is that a combination of variants in several cellular structures or pathways could be required for IS development (e.g., at least one variant is needed in a stereocilia gene, cilia gene, and ECM gene). Our data seems support the second possibility, as three of the five families have at least one variant in each of the three enriched functional categories identified in this study (File S2). This suggests that a combination of variants in all three functional categories may be important for the development and progression of IS. This interpretation is based on only five families, and will need to be investigated further in larger familial studies. Overall, we believe that there are multiple mechanisms that can lead to IS and that our identification of variants in three enriched functional categories supports this theory.

To summarize, we identified three interconnected categories of GO terms that are enriched in our dataset of five IS families: actin-based cellular projections, microtubule-based cellular projections, and the ECM. The ECM is important for both stereocilia (actin-based cellular projections) and cilia (microtubule-based cellular projections), tying together the three functional categories identified in our dataset (Seeger-Nukpezah and Golemis 2012; Collins and Ryan 2014; Deans 2013). Previous studies support the involvement of each of these individual categories of genes and GO terms in IS etiology. Functional studies of the variants identified in these families will ultimately be needed to strengthen the link between IS and these associated genetic changes. Our data provides additional genetic support for the involvement of these functional categories in IS etiology. Once the genetic causes of IS are known, IS may be classified into multiple sub-diseases, each with a different genetic cause and different disease pathogenesis. This information will hopefully lead to further clinical insights into IS, including risk of curve progression or response to bracing, which could eventually lead to personalized disease treatments.

ACKNOWLEDGMENTS

Exome sequencing was completed using funds donated by the LARRK Foundation. NHM’s laboratory is supported by NIH/NIAMS R01AR068292. CGP is supported by NIH/NIGMS R01GM099820. Additional funding was provided by Rose Brown Endowment Funds and Children’s Hospital of Colorado Research Institute. Our research used the REDCap database and the Colorado Clinical and Translational Sciences Institute, which are supported by Colorado CTSA Grants UL1TR002535, KL2TR002534, and TL1TR002533. Contents are the authors’ sole responsibility and do not necessarily represent official NIH views. Exome sequencing for this project was completed at the University of Colorado Denver Genomics and Microarray Core Facility with Bifeng Gao and Katrina Diener. The authors thank G. Devon Tranah for his assistance in acquiring pLI and pRec scores listed in File S2. On behalf of all authors, the corresponding author states that there are no conflicts of interest.


Cingolani, P., A. Platts, L. Wang le M., Coon, T. Nguyen et al., 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80–92. https://doi.org/10.4161/fly.19695


Communicating editor: B. Andrews