Detecting Genetics Modifiers of Spondyloepimetaphyseal Dysplasia with Joint Laxity (SEMDJL) in the Caucasian Afrikaner community

Article *in* Human Molecular Genetics · October 2018

DOI: 10.1093/hmg/ddy373

CITATION
1

READS
72

1 author:

Emile R. Chimusa
University of Cape Town
111 PUBLICATIONS **954** CITATIONS

SEE PROFILE

Detecting Genetics Modifiers of Spondyloepimetaphyseal Dysplasia with Joint Laxity (SEMDJL) in the Caucasian Afrikaner community

Emile R. Chimusa[1*], Peter Beighton[1], Judit Kumuthini[4], and Rajkumar S Ramesar[1*]

[1] Division of Human Genetics, Department of Pathology, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Private Bag, Rondebosch, 7700, Cape Town, South Africa.

[2]African Institute for Mathematical Sciences,7945 Muizenberg, Cape Town, South Africa.

[3]Centre for Proteomic and Genomic Research, First floor, St. Peter's St. Peter's Square St. Peter's Square Mall Observatory, Cape Town, 7925. South Africa.

***Correspondence E-mail:** ERC (emile.chimusa@uct.ac.za) and **RR (**raj.ramesar@uct.ac.za)

Abstract

Spondyloepimetaphyseal dysplasia with joint laxity (SEMDJL) is an autosomal-recessive skeletal dysplasia. A relatively large number of patients with SEMDJL have been identified in the Caucasian Afrikaans-speaking community in South Africa. We used a combination of Genome-Wide Human SNP Array 6.0 data and whole exomic data to potentially dissect genetics modifier associated with SEMDJL in Caucasian Afrikaans-speaking patients.

Leveraging the family-based association signal in prioritizing candidate mutations, we identified two potential modifiers genes, ***COL1A2*** and ***MATN1,*** and replicating previously identified mutation in ***KIF22***. Importantly, our findings of genetics modifiers genes and previously identified mutations are layered on the same sub-network implicated in syndromes characterised by skeletal abnormalities and intellectual disability, bone and connective tissue fragility. This study has potentially provided crucial insights in identifying the indirect modifying mutation(s) linked to the true causal mutation associated with SEMDJL. It is a critical lesson that one may use constructively especially when the pace of exomic sequencing of rare disorders continues apace.

Introduction

The Spondyloepimetaphyseal dysplasias (SEMD) are a heterogeneous group of genetic bone disorders which are defined in terms of the anatomical distribution of radiographically apparent changes in the skeleton [1]. The Spondyloepimetaphyseal dysplasias with joint laxity (SEMDJL) are classified as a subgroup of the former category on a basis of the additional syndromic feature of gross articular hypermobility [2]. Although very rare, several autonomous forms [2] of SEMDJL have been documented and phenotypic and genetic heterogeneity are well recognized. The prototype SEMDJL (MIM 271647) was delineated in the Afrikaner community of South Africa in 1980 and the manifestations were further documented [3,4,5]. By 2016 a total of 30 affected South Africans had been studied. The condition is an autosomal recessive trait and several of the affected families had common progenitor. It is likely that the determinant gene reached South Africa from Western Europe during the past 300 years [6]. It is hypothesized that the mutated gene reached the Afrikaner community from sea-borne European immigrants [6].

The skeletal involvement in this South African form of SEMDJL leads to defective growth and stunted stature, while the articular hypermobility predisposes to limb joint dislocation and progressive spinal malalignment. In addition, structural cardiac abnormalities are present in about 30% of affected individuals. Demise from cardio-respiratory failure before adulthood is usual but longer survival is not unprecedented. Bone density is normal and spontaneous fracturing does not occur. Children with the condition have a characteristic facial appearance and cleft or high palate is frequent. The skin may be somewhat extensible although not fragile.

A very similar SEMDJL phenotype has been recognized in a series of Japanese patients and it has emerged that the molecular determinant is *B3GAL*T [7]. Thereafter, the same mutation was identified in affected South African families [8]. This condition is conventionally designated SEMDJL-1, South African Type or with the eponym 'Beighton' appended.

The Leptodactylic form of SEMDJL is an autosomal dominant condition which is distinguished from the other types by narrowing and shortening of the fingers, hence the alternative designations "Lepto-SEMDJL" and "SEMDJL-2 (Hall Type)". This disorder results from a mutation in *KIF2*2 (MIM 603313) [9]. The Progeroid form of SEMDJL has features of the Ehlers-Danlos Syndrome plus an appearance of premature ageing. The term "progeroid" in the title pertains to this attribute. A propensity to skeletal fragility and fracturing is an additional manifestation. Loss-of-function mutations in *B4JLT7* underlies this condition (MIM 130070). A few other rare genetic skeletal – loose joint disorders are listed in the SEMDJL spectrum. Notable amongst these is the autosomal recessive form of the Larsen Syndrome, in which joint laxity and cardiac anomalies are significant (MIM 245600)[10]. Another is Kniest Dysplasia (MIM 156550) which is an autosomal dominant, short trunk dwarfing skeletal condition with joint rigidity, myopia, flat facies and palatal clefts [11,12]. It is a member of the COL2A1 group. Substantial clinical variability is observed in many form of SEMDJL, so that patients with the same mutation may develop a very severe form of SEMDJL [3]. The prediction of the evolution of the complications of SEMDJL are still not fully understood, except that in general, patients with a predominant epiphyseal component as part of a skeletal dysplasia will develop premature osteoarthritis, especially of the weight bearing large joints [6]. Among the factors that may explain these differences in SEMDJL expression are modifier genes.

Here, we dissect genetics modifiers variants associated with previously identified mutation with SEMDJL by leveraging the linkage disequilibrium, different biological knowledge-based Protein-Protein Interactions (PPI) network analysis and further filtering candidate mutations from whole exome sequencing data. We further conduct a family-based variance component association model on the Affymetrix Genome-Wide Human SNP Array 6.0 data of 5 affected Afrikaners with the SEMDJL phenotype and compared this to 165 unaffected Caucasian individuals from 1000 Genomes Project [13].

Results

We genotyped five family samples from a Caucasian South African cohort (Afrikaners) with SEMDJL, two siblings and three genetical related individuals (third cousins according to IBS). We sequenced the exome of one of the affected family members from the Afrikaner family (Materials and Methods). We conducted a principal component

analysis (PCA) using both 1000 Genomes samples from the Caucasian (CEU) and Afrikaner data (**Figure 1**). The Caucasian CEU was related to the Afrikaner 'population' (F_{ST} = 0.001). Since SEMDJL is neither a common disorder nor an infectious disease, following the observation of close relatedness between Afrikaner and CEU, 103 CEU samples were used as a control in our study (Materials and Methods).

No evidence of population differentiation

From the PCA, the Caucasian, CEU, are observed to be not divergent to the Afrikaner (F_{ST} = 0.001); we therefore used CEU samples as the parental proxy population for the 5 Caucasian Afrikaners from Genome-Wide Human SNP Array 6.0. We evaluated whether there is an excess of SNPs with large allele frequency differences (expressed as a chi-square statistic with one degree of freedom under a model of neutral genetic drift) between Caucasian Afrikaners and the CEU populations which would imply unusual population differentiation. An unusual extent of population differentiation can suggest the action of population-specific natural selection. We observed no common SNPs within chromosomal regions for which the evidence of unusual population differentiation was genome-wide significant between the Caucasian Afrikaner and the CEU populations. A few rare variants observed (Supplementary **Table S1**) have moderately unusual allele frequency differences. No evidence of unusual allele frequency differences may be explained by the fact that the migration of Caucasians (1497-1849) into South Africa occurred [14, 15] too recently for it to have a significant impact on their allele frequencies.

Family-based variance component association test

To account for both population stratification and hidden relatedness, we applied the variance component-based statistic on five SEMDJL individuals and 103 unaffected trio samples from CEU using 766,876 SNPs. This analysis identified six variants (**Table 1** and **Figure 2**) with low minor allele frequencies. Additionally, we conducted an additive, dominant and recessive association test (**Table 1**). Low-frequency variants (1-5%), may often reach genome-wide significance in standard tests, such as mixed linear model association (**Figure 2**) or logistic regression, due to the imperfect asymptotic distribution of those tests in the case of low-frequency variants [16]. Therefore, we performed a Fisher's exact test for rare SNPs, to check whether the specific SNPs would still be genome-wide significant. The result in **Table 1** suggests that the variant *rs16833849* (mixed model p-value = 6.54e-09, Fisher's exact p-value = 1.91e-08, MAF = 0.03) is still genome-wide significant. The closest gene to this SNP is ***MATN1***. This gene encodes a von Willebrand factor involved in the formation of filamentous networks in the extracellular matrices of various tissues. Notably, mutations in this gene have been associated with a variety of inherited chondrodysplasias [17].

Whole exome sequencing (WES) analysis

The implementation of newer sequencing technologies is still required to search for rare risk variants that are not well captured by SNP genotyping [18-22]. WES analysis was conducted using a single SEMDJL affected family member. The GATK unified genotyper implemented in Genome Analysis Toolkit (GATK) [18] was used based on the recommended settings to detect different variants – 683,538 single nucleotide variants (SNVs) and indels (mutation class that includes both insertions, deletions) were identified. Despite the prior knowledge of mutations in ***B3GALT6*** reported to be causative of the disorder, markers in this region were in poor quality, and could not pass the quality control (QC) test**.** We filtered variants for rarity, amino acid substitution on the structure and function, evolutionary conservation, predicted functional significance, natural selection on codons, and deleteriousness based on a comparative genomics data set of 32 vertebrate species [19-21] and by prioritizing the association signals obtained from family-based tests [21]. During this filtering process, all variants were annotated using the dbSNP137 and 1000 Genome databases [13,18]. Because SNPs that arise by recurrent mutations, identical alleles are presumed to be identical by descent (IBD) in the population-genetics sense [23-24]. Such IBD segments tend to be extremely short, perhaps covering only a single polymorphism or, at most, a few kilobases of DNA [24, 25]. Therefore, the Genome-Wide Human SNP Array 6.0 was used to search for the longest region (in minimal length of 103 307 kb) of each interesting variant, in which there is at least one allele identical among all SEMDJL samples [24-25]. As expected this yields several long regions of IBD among the five SEMDJL samples.

To further highlight potential causal homozygous mutations or to exclude those that are non-disease-relevant, we used the estimated shared regions of IBD among affected SEMDJL samples and the summary statistics from the family-based association tests from the Affymetrix Genome-Wide Human SNP Array 6.0 data. **Table 2** displays the three novel candidates in genes ***MATN1*** and ***COL1A2*** (**Table 3**), segregated with disease status and were absent in unaffected CEU reference exomes. In addition, a previously identified homozygous mutation in ***KIF22*** is identified (**Table 3**) in the exome sequencing data of a single affected family member. It is worth to note that these identified changes in ***MATN1*** and ***COL1A2*** might be influence from the main SEMDJL causal mutation variants**.** The identified mutation of ***MATN1*** is supported and is consistent with the association result obtained from the Affymetrix Genome-Wide Human SNP Array 6.0 data (**Table 1**). The homozygous mutation in the ***KIF22*** gene was also previously identified in another type of SEMDJL, the leptodactylic (lepto-SEMDJL) form in eight affected Korean individuals [11]. Lepto-SEMDJL is well-known as an autosomal-dominant skeletal dysplasia manifesting with short stature, joint laxity with dislocation(s), limb malalignment, and spinal deformity [11]. It is worth to observe that the modification of the KIF22 association implies that the association of KIF22 is dependent on genotypes in COL1A2 and MATN. Moreover, SEMDJL type I was reported to be caused by homozygous mutations in the ***B3GALT6*** gene on chromosome 1p36 in 5 unrelated Japanese families, a Japanese family from Singapore [12] and in 3 unrelated Iranian families [26]. Relative to the identified ***B3GALT6*** mutations, we recently conducted in-depth mutation analysis by sequencing of ***B3GALT6*** in eight affected South African families, and identified a mixture of homozygous, and compound heterozygous mutations, indicating non-shared haplotypes among the affected South African families [2]. Since, our current exome sequencing data of one of these eight affected South

African families [2] and filtering candidate mutation analysis were based on causal homozygous mutations using an estimated shared region of IBD, logit models in KGGSeq [23-24] and ANNOVAR [27], therefore it was expected that the ***B3GALT6*** gene could not be isolated even if variants in ***B3GALT6*** were available or passed QC. We performed imputation fine mapping and computed the linkage disequilibrium (LD) between the identified candidate variants with mutations in the previously identified ***B3GALT6*** gene (Materials and Methods). **Table 3** shows the LD pattern between our identified candidate variants with mutations and previously identified ***B3GALT6.***

To see how our findings and these previously identified mutations are layered in a biological network, we queried these genes from a comprehensive human Protein-Protein Interaction (PPI) network [28]. **Figure 3B** is the extracted sub-network of interactions of our two novel mutations in ***COL1A2*** and ***MATN1,*** and previously identified causal mutations in ***B3GALT6*** and ***KIF22*** genes and all are observed to be connected. The ***COL1A2*** gene is the hub of the sub-network (**Figure 3A**). The mutations in ***COL1A2*** are known to be associated with osteogenesis imperfecta, Ehlers-Danlos syndrome, idiopathic osteoporosis, and atypical Marfan syndrome [29-30]. We have characterized the sub-network in **Figure 3A**, with relevant biological pathways, biological processes, molecular functions and human phenotypes (**Table 3**), the current identified variant genes in which mutations occurred and previously identified variants are shown to be in the same sub-network and to contribute to the same biological pathways. The biological pathways (**Table 3**) associated with the sub-network in **Figure 3A** are enriched with relevant molecular functions, implicated in syndromes characterised by skeletal abnormalities and intellectual disability, bone and connective tissue fragility. Furthermore, the results in **Table 3** demonstrate a possible implication of our identified central sub-network in **Figure 3A** to significantly play roles in causing SEMDJL, and potentially indicate that other anatomical systems may be implicated in these disorders. Furthermore, the fraction of ancestral allele and derived alleles were computed using a custom script. The proportion of ancestral alleles was analysed in the SEMDJL single exome patient, 5 genotyped patients and European controls from 1000 Genomes (Materials and Methods). **Figure 3B** indicates that the SEMDJL patients from exome and genotyped data have a lower proportion of ancestral alleles whereas the European control from 1000 Genomes has a higher proportion of ancestral alleles at most candidate mutation genes. This result indicates that the SEMDJL population has a higher proportion of derived alleles and potentially pathogenic, variations are observed.

Discussion

This present study is innovative in two key ways: (1) The data supports accumulation of multiple genetics modifiers in core genes with major effect, and (2) The methodological approach used could serve as a model to explore novel and known genetic modifiers of monogenic and complex conditions. In this paper, we presented an application of a combined family data of Caucasian Afrikaners with SEMDJL from Genome-Wide Human SNP Array 6.0 and the exome sequencing of a single-family member in identifying candidate causal mutations. Since SEMDJL is known to be a rare non-communicable disease, we used the CEU samples from 1000 Genomes

Project as proxy control in both the association analysis, in isolating causative mutations and the fraction of ancestral alleles at identified and previously known candidate mutation genes. We evaluated whether there is an excess of SNPs with large allele frequency differences (expressed as a chi-square statistic under a model of neutral genetic drift) between the Caucasian Afrikaner (SJL) population and the CEU population. Our results showed no evidence of common SNPs within chromosomal regions for which the evidence of unusual population differentiation was genome-wide significant between Caucasian Afrikaner and CEU population. This may be explained by the fact that the migration of Caucasians into South Africa [14, 15] occurred too recently for it to have a significant unusual differentiation on their allele frequencies. Our study applied a family-based mixed variance component model to account for both population stratification and hidden relatedness in the family-based association test of 900k genotyping SNPs of five SEMDJL Caucasian Afrikaners against the trio CEU samples as control group. We identified six variants of low minor allele frequencies (1-5%). We addressed the issue of rare variants in a mixed model [16], by applying Fisher's exact test which allowed us to demonstrate that there were five nearly rare variants among the identified six variants of which were no longer genome-wide significant although they achieved significant association p-values. In general, the analysis of rare or low-frequency variants poses challenges in any dataset of smaller to larger sample size [16].

From sequencing data of eight Caucasian Afrikaner families with SEMDJL, we previously identified and validated three variants in ***B3GALT6,*** missense mutations from targeted sequencing. Given observed clinical variability in severity and differing phenotypic modifications of SEMDJL [3], we have analysed the whole exomic sequence of one member of these affected families to investigate for genetics modifiers. We utilised the family-based association results as prior knowledge within the estimated identity-by-descent and different knowledge-based filtering analyses to efficiently discover and narrow down candidate mutations causing SEMDJL. Our study identified two candidate mutations in ***COL1A2*** and ***MATN1*** which might potential be affected by the change the main known mutation in B3GALT6 [26,29-30]. Several variants in these genes are in fairly linkage disequilibrium with variants in ***B3GALT6*** (**Table 4**). Our finding replicated a previously identified mutation in ***KIF22*** to be potentially associated with SEMDJL. We also show that our identified candidate mutation genes ***COL1A2***, ***MATN1*** and ***KIF22*** are in a direct biological interaction (**Figure 3A**) with ***B3GALT6***. Importantly, mutations of the ***MATN1*** are known to be associated with a variety of inherited chondrodysplasias [31, 32] and those in ***COL1A2*** are associated with osteogenesis imperfecta, Ehlers-Danlos syndrome, idiopathic osteoporosis, and atypical Marfan syndrome [28 29]. A study by Byung-Joo et al. 2012 [11], demonstrated that the mechanism by which the mutations in ***KIF22*** lead to disease is likely related to the dysfunction of ***KIF22.*** Importantly, our study has identified a sub-network (**Figure 3A**) of interactive variant genes (that also include ***B3GALT6***) through which SEMDJL might have resulted [2, 3]. We have characterized the identified sub-network with biological pathways, biological processes and molecular functions that are implicated in syndromes characterised by skeletal abnormalities, intellectual disability, and connective tissue fragility. Moreover, we have discovered that most mutations happened in previously identified variant genes and these variant genes are interacting in the same sub-network in **Figure 3A**.

This indicates that this sub-network may play critical roles in the cause of SEMDJL from differing populations with some environmental implications that differentiate the SEMDJL trait. It has shown that derived alleles are more often minor alleles (< 50% allele frequency) and more often associated with risk than ancestral alleles [33], our analysis of the fraction of ancestral alleles in single exome patient with SEMDJL, 5 genotyped patients and European control from 1000 Genomes (**Figure 3B**) yielded expected results with regards to low proportion of ancestral at identified candidate SEMDJL genes. The SEMDJL patient population presented with a higher proportion of derived alleles. This study has potentially provided crucial insights in identifying the causal mutation(s) associated with SEMDJL in combining SNPs genotyping and exomic sequencing data and revealed an enriched sub-network of interacting variant genes in which mutations happened resulting in SEMDJL trait, and perhaps contributing to a range of phenotypic manifestations which might be difficult to tease out from the core syndromic phenotype.

Materials and Methods

Sampling and genotyping

DNA samples were shipped to Affymetrix (http://www.affymetrix.com) for genotyping using the Affymetrix Genome-Wide Human SNP Array 6.0, containing 766, 876 SNPs of 5 samples previously sequenced in [2, 3]. The samples consisted of two siblings and three closely related individuals belonging to the Afrikaans-speaking Caucasian community of South Africa. Blood samples were collected with the subject's informed consent, and the use of DNA samples for population genetics research was approved by institutional review board (IRB; REC REF 643/2013). Quality of hybridization intensity (CEL) files was assessed with the Dynamic Model (DM) algorithm, and only individuals for which the call rate was greater than 90% (QC > 90) were included in downstream genotype calling. Mean CEL file QC call rates were all > 97.5% for the remaining individuals. Probe specific intensities for each sample were normalized and summarized using the Affymetrix quantile normalization protocol. Genotypes were called using the Birdseed algorithm distributed with Affymetrix Power Tools [34, 35]. We performed quality-control filters and removed 142,746 SNPs that failed the Hardy-Weinberg exact test $P < 0.000001$ and had a call rate < 95% across all samples using PLINK [36]. We retained 766, 876 SNPs. genotypes (CEU, Caucasian) data from 1000 Genomes Project (http://www.1000genomes.org/) were downloaded. We conducted a PCA using both CEU and Afrikaner data. Because the Caucasian CEU are related to Afrikaner ($F_{ST} = 0.001$), since SEMDJL is not a common infectious disease, and following a recent recommendation on Mendelian diseases analysis in [37], the Caucasian CEU was used as proxy control in this study. A custom python script was used to convert probe_id to dbSNP_id and align HapMap3 and our data to the same strand.

Unusual difference in allele frequency

To minimize deviation from the normality assumption, SNPs with minor allele frequencies (MAF) < 0.05 are excluded. Thus, at a given locus *i*, the difference$(p_i^k - p_i^l)$between observed variant allele frequencies of two populations, *k* and *l*, can be approximated as a normal distribution under neutral drift with mean 0 and variance σ^2 as in [37, 38]. Using the implemented χ^2 approach in [38,39], we test unusual difference in allele frequency from populations k and l. An excess of large values of the χ^2 statistic indicate deviations from the null model equation (equation a and b), suggesting the action of natural selection [38,39]. We applied this method to the data from the Afrikaner and European (CEU) populations. All gene annotations were obtained using GeneCards [40].

Mixed model for family data association tests

The association testing was performed on the full data set of a family member of two siblings and three closely related samples of South African Caucasian Afrikaner with Spondyloepimetaphyseal dysplasia with joint laxity against a total of 103 unaffected Caucasian individuals from 1000 Genomes Project [13], which contained trios and related individuals. To account for both population stratification and hidden relatedness that can result from the genealogy, we applied a mixed model for family data describes in [41,42], like family-based variance component association test as in [42], implemented in a custom R script. This corrects for these relationships during the association mapping. We first computed a pair-wise relatedness matrix from both datasets, which represents the structure of all samples. We estimated the contribution of the sample structure to the SEMDJL phenotype using a variance component mixed model, resulting in an estimated covariance matrix of phenotype that models the effect of genetic relatedness on the SEMDJL phenotype. We ran our custom R script on SEMDJL phenotype data using the estimated covariance matrix to detect possible association. To account for rare variants that the variance component model could not address adequately, we separately performed the Fisher's exact test using PLINK [36], which is known to be appropriate for rare SNPs [30].

To cautiously account for the lack of flexibility to handle multiple predictors, the contribution of the sample structure to the phenotype and possible inflation in using Fisher's exact test, we only apply this test to few low-frequency variants (MAF < 0.05) that show to be significant from the mixed models in GWAS, as this will aid in identifying associations. This is due to the imperfect asymptotic distribution of those mixed models in the case of low-frequency variants. Given m SNPs (from typed or imputed data sets) for association with SEMDJL, we expected around m x 0.05 to have p-value less than 0.05 in each data set, respectively. We therefore used a genome-wide significance level of α = (0.05 / 2 x m).

Exome sequencing analysis

The SEMDJL subject were used for commercial WES (Otogenetics, Norcross, GA) and analyzed as previously reported [2, 3]. The SureSelect Human All Exon Kit, 38 Mb, from Agilent Technologies was used for filtering

candidate mutation in one of previous SEMD-JL affected South African families in [2, 3]. Exomes were then sequenced on an Illumina GAII using one lane each with a 90-bp paired-end protocol. Briefly, paired-end reads of 100 bp from a single-family member patient were aligned by bwa-0.6.1 to the UCSC hg19 reference genome using default settings. We applied Samtools [43-44] and Picards (http://picard.sourceforge.net/) to process SAM/BAM files and mark duplicates as recently recommended [17]. We conducted local realignment around indels and base quality score recalibration, and variants were called by a unified genotyper in GATK tool [18]. We applied a custom python script and ANNOVAR [27] to annotate variants and to search important known SNPs and indels from dbSNP (http://www.ncbi.nlm.nih.gov/project/SNP/, build 137) and the 1000 Genome database (ftp://www.1000genome.org). The variants with a read depth greater than 10 and genotype quality score greater than 30 were filtered for further analysis. Using Picards, we removed Polymerase chain reaction (PCR) duplicates.

Filtering variant and candidate mutations

Low-quality SNP calls and indels were filtered out using the following criteria described previously [18, 19, 21, 22]: 1) consensus quality score ≥ 20; 2) sequencing depth ≥ 4 and ≤500; 3) copy number ≤ 2; 4) distance between two adjacent SNPs ≥ 5 bp. Given that common genetic variations were unlikely to be the causal mutations of SEMDJL, all identified SNPs and indels were further filtered against the 1000 exome data of 30 CEU individuals from the 1000 Genome Project (ftp://www.1000genome.org) and their SNP data available in the dbSNP database (http://www.ncbi.nlm.nih.gov/project/SNP/, build 137). Although hard-filtering (MAF>1% in dbSNP, 1000 Genomes) can exclude more than 80% of non-synonymous single-nucleotide variations (nsSNVs), there are still ~700 variants left for examination. We used prediction from different logit models in KGGSeq [22-23] and ANNOVAR [27] to further exclude more than 500 (55% of) nsSNVs. We thus, applied the Mutation Taster, LRT, PhyloP and SIFT implemented in KGGSeq to prioritize nsSNVs in the SEMDJL-disease patient. We provided to KGGSeq the association result obtained from Affymetrix Genome-Wide Human SNP Array 6.0 and the estimated IBD among the 5 affected SEMDJL Caucasian Afrikaners. We used the suggested cut-off as follows: PolyPhen2: probably damaging (P) if the score is between 0.909 and 1, and "possibly damaging" if the score is between 0.447 and 0.908, and "benign" is the score is between 0 and 0.446. Mutation Taster, LRT, PhyloP and SIFT are similar and all predict possible impact of an amino acid substitution on the structure and function of a human protein. The score ranges from 0 to 1. There are four possible predictions; A: disease causing automatic, D: disease causing, N: polymorphism or P: polymorphism automatic. For SIFT, a small score indicates a high chance for a substitution to damage the protein function.

Fine-mapping and Characterization of enriched sub-networks

We performed fine mapping based on the imputation using Minimac [45] and computed the linkage disequilibrium between the identified candidate variants with mutations and previously identified *B3GALT6* following the method

describes below. Assuming sets of SNPs $S_a = \{s_i\}_{i=1,2,,m}$ and $S_b = \{s_j\}_{j=1,2,\ldots,n}$, $s_i \neq s_j$, for i = 1, 2, . . . , m, and for j = 1, 2, . . . ,n are assigned to genes a and b; the pairwise LD of SNPs between a and b are independent and are computed using the r2 measure [46]. The distribution of the LD_{sisj} is not normal, thus from ($s_i \neq s_j$) we compute the average z-transforms of LD from all possible combinations of pairs of SNPs between genes a and b. The z-transforms of LD are normally distributed with mean 0 and variance 1 [47]. We compute the combined LD between two genes a and b as follows,

$$r_{\mathrm{ab}} = \tanh\left(\frac{\sum_{i\neq j}^{\mathrm{nxm}} \tanh^{-1}\left(\mathrm{LD}_{s_i s_j}\right)}{\mathrm{nxm}}\right)$$

Here we aim to identify a sub-network of interactive variant genes of which our current finding mutations and previously identified mutations occurred and investigate the association between that sub-network with known human pathways. In addition, examine how these variants genes in the sub-network are associated with human phenotypes and what are their biological processes and molecular functions. First, we use a comprehensive human Protein-Protein Interaction (PPI) network (http://cbg.garvan.unsw.edu.au/pina/) [28] to analyse the set of our finding variants genes and previously identified variants to see how they are layered in a biological network and, thus extract a sub-network. Secondly, we conducted an enrichment analysis on the obtained sub-network using Enrichr (http://www.amppharm.mssm.edu/Enrichr/#) to identify enriched pathways associated with our set of variant genes, their biological processes, their molecular functions and the associated human phenotypes.

Fraction of ancestral allele

The separate exome patient, all 5 samples genotyped from affymetrix 6.0 and and European VCF from 1000 Genomes were used to analyse the fraction of ancestral alleles. We downloaded the SNP ancestral alleles from the Ensembl [48], and further checked the SNPs for those present in the dbSNP database. The single exome SEMDJL, all 5 SEMDJL samples genotyped from affymetrix 6.0 and European VCF files was further independently annotated using the VCFtools 'fillOaa' script [33].

We first computed, for each SNP, the fraction of ancestral allele, which was calculated by dividing the number of times the alternative allele matched with ancestral allele by the total number of copies of all alternative alleles across all samples for that SNP. The fraction of derived allele is equivalent to 1 minus the fraction of ancestral allele. Previous work has shown that derived alleles are more often minor alleles (< 50% allele frequency) and more often associated with risk than ancestral alleles [33]. Furthermore, we computed the fraction of ancestral alleles for all known and candidate mutation SEMDJL genes. To this end, we aggregated the fraction of ancestral alleles at SNP-based level to gene, considering all SNPs located within the gene's downstream or upstream region

[33].

Acknowledgements

We would like to acknowledge all participants who have volunteered to participate in our research, for their roles in ensuring the genotyping of the southern African individuals in this study. Computations were performed using facilities provided by the Centre for high performance computing (https://users.chpc.ac.za/, South Africa).

Conflict of Interest Statement

Disclosure: The authors declare no conflict of interest.

Ethics approval and consent to participate

This investigation received ethical approval from the institutional review board (IRB; REC REF 643/2013).

References

1. Bonafe, L., Cormier-Daire, V., Hall, C., Lachman, R., Mortier, G., Mundlos, S., Nishimura, G., Sangiorgi, L., Savarirayan, R., Sillence, D. and Spranger, J., (2015). Nosology and classification of genetic skeletal disorders: 2015 revision. Am. J. Med. Genet. Part A, 167:2869-2892.
2. Beighton, P. and Kozlowski, K., (1980). Spondylo-epi-metaphyseal dysplasia with joint laxity and severe, progressive kyphoscoliosis. Skeletal radiology, 5:205-212.
3. Beighton P., Kozlowski K., Gericke G., Wallis G., Grobler L. (1983) Spondylo-epimetaphyseal dysplasia with joint laxity and severe, progressive kyphoscoliosis. S. Afr. Med. J. 64: 772-775.
4. Kozlowski, K. and Beighton, P., (1984), September. Radiographic features of spondylo-epimetaphyseal dysplasia with joint laxity and progressive kyphoscoliosis. In RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren). Georg Thieme Verlag Stuttgart· New York, 337-341.
5. Beighton, P., Gericke, G., Kozlowski, K. and Grobler, L., (1984). The manifestations and natural history of spondylo-epi-metaphyseal dysplasia with joint laxity. Clinical genetics, 26:308-317.
6. Torrington M., Beighton P. (1991) The ancestry of spondyloepimetaphyseal dysplasia with joint laxity (SEMDJL) in South Africa. Clin. Genet. 39: 210-213.
7. Christianson, A.L. and Beighton, P., 1996. Spondyloepimetaphyseal dysplasia with joint laxity (SEMDJL) in three neonates. Genetic counseling (Geneva, Switzerland), 7(3),219-225.
8. Vorster AA, Beighton P, Ramesar RS. (2014). Spondyloepimetaphyseal dysplasia with joint laxity (Beighton type); mutation analysis in eight affected South African families. Clin. Genet. 87:492-495.
9. Boyden, E.D., Campos-Xavier, A.B., Kalamajski, S., Cameron, T.L., Suarez, P., Tanackovich, G., Andria, G., Ballhausen, D., Briggs, M.D., Hartley, C. and Cohn, D.H., (2011). Recurrent dominant

mutations affecting two adjacent residues in the motor domain of the monomeric kinesin KIF22 result in skeletal dysplasia and joint laxity. The American Journal of Human Genetics, 89:767-772.

10. Sulko J., Czarny-Ratajczak M., Wozniak A., Latos-Bielenska A., Kozlowski K. (2007) Novel amino acid substitution in the Y-position of collagen type II causes spondyloepimetaphyseal dysplasia congenita. Am. J. Med. Genet. 137A: 292-297.
11. Byung-Joo M., Namshin K., Chung T., Ok-Hwa K., Nishimura G., Chin Youb C.,Song HR., Kim HW., Lee HR., Kim J.,3 Tae-Hoon K., Myung-Eui S., San-Deok Y., Do-Hwan K., Seung-Bok L., Jong-Il K., Jeong-Sun S., Ji-Yeob C.,Daehee K., Dongsup K., Woong-Yang P., and Tae-Joon C. (2012) Whole-Exome Sequencing Identifies Mutations of KIF22 in Spondyloepimetaphyseal Dysplasia with Joint Laxity, Leptodactylic Type. Am. J. Hum. Genet. 89:760-766.
12. Malfait, F., Kariminejad, A., Van Damme, T., Gauche, C., Syx, D., Merhi-Soussi, F., Gulberti, S., Symoens, S., Vanhauwaert, S., Willaert, A., Bozorgmehr, B., Kariminejad, M. H., Ebrahimiadib, N., Hausser, I., Huysseune, A., Fournel-Gigleux, S., De Paepe, A. Defective initiation of glycosaminoglycan synthesis due to B3GALT6 mutations causes a pleiotropic Ehlers-Danlos-syndrome-like connective tissue disorder (2013). Am. J. Hum. Genet. 92: 935-945.
13. Laura C, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, et al. (2012) The 1000 Genomes Project: data management and community access. Nature methods 9: 459-462.
14. Felipe Fernandez-Armesto, Pathfinders (2006) A Global History of Exploration. New York: Norton, pp. 177178.ISBN 0-393-06259-7.
15. David Hatcher Childress (2011) A Hitchhiker's Guide to Armageddon. SCB Distributors, Gardena, California:ISBN 1935487507.
16. Chimusa ER, Noah Zaitlen, Michelle Daya, Marlo Mller, Nicola J Mulder, Alkes L Price, Eileen G Hoal (2013) Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. Hum. Mol. Genet. 23:796-809.
17. Bae JW, Cho CH, Min WK, Kim UK (2012) Associations between matrilin-1 gene polymorphisms and adolescent idiopathic scoliosis curve patterns in a Korean population. Mol. Biol. Rep.39:5561-5567.
18. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K,Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA.(2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res.20(9):1297-303.
19. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., Liu, X., (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet. 24, 2125-2137.
20. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J., (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310-315.
21. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC (2012). A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res. 40(7):e53.
22. Li, M.X., Kwan, J.S., Bao, S.Y., Yang, W., Ho, S.L., Song, Y.Q. and Sham, P.C., (2013). Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing

studies. PLoS genetics, 9(1), p.e1003143.

23. Musunuru, K., Pirruccello, J.P., Do, R., Peloso, G.M., Guiducci, C., Sougnez, C., Garimella, K.V., Fisher, S., Abreu, J., Barry, A.J. and Fennell, T., (2010). Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. New England J. of Med., 363:2220-2227.
24. Krawitz, P.M., Schweiger, M.R., Rödelsperger, C., Marcelis, C., Kölsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S. and Isau, M., (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nature genetics, 42(10), p.827.
25. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 40(10):53-60.
26. Nakajima, M., Mizumoto, S., Miyake, N., Kogawa, R., Iida, A., Ito, H., Kitoh, H., Hirayama, A., Mitsubuchi, H., Miyazaki, O., Kosaki, R., Horikawa, R., and 19 others (2013). Mutations in B3GALT6, which encodes a glycosaminoglycan linker region enzyme, cause a spectrum of skeletal and connective tissue disorders. Am. J. Hum. Genet. 92: 927-934.
27. Wang K, Li M, Hakonarson H.(2010) ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data Nucleic Acids Research, 38:e164.
28. Wu J, Vallenius T, Ovaska K, Westermarck J, Mkel T, et al. (2009) Integrated network analysis platform for protein-protein interactions. Nat. Meth. 6, 75-77.
29. Retief, E., Parker, M.I. and Retief, A.E., (1985). Regional chromosome mapping of human collagen genes alpha 2 (I) and alpha 1 (I)(COLIA2 and COLIA1). Human genetics, 69:304-308.
30. Wenstrup RJ, Cohn DH, Cohen T, Byers PH (1988) Arginine for glycine substitution in the triple-helical domain of the products of one alpha 2(I) collagen allele (COL1A2) produces the osteogenesis imperfecta type IV phenotype. J. Biol. Chem 263 (16): 34-40. PMID 2897363.
31. Jenkins RN, Osborne-Lawrence SL, Sinclair AK, Eddy RL, Byers MG, Shows TB, Duby AD (1990). Structure and chromosomal location of the human gene encoding cartilage matrix protein. J. Biol. Chem. 265 (32): 24-31.
32. Deák F, Piecha D, Bachrati C, Paulsson M, Kiss I (1997). Primary structure and expression of matrilin-2, the closest relative of cartilage matrix protein within the von Willebrand factor type A-like module superfamily. J. Biol. Chem. 272 (14):68-74.
33. Gorlova, O.Y., Ying, J., Amos, C.I., Spitz, M.R., Peng, B. and Gorlov, I.P., 2012. Derived SNP alleles are used more frequently than ancestral alleles as risk-associated variants in common human diseases. J. of bioinf. and comp. biol., 10:1241008.
34. Brettschneider J, Collin F, Bolstad BM, Speed TP. Quality assessment for short oligonucleotide microarray data. Technometrics. 2008;50:41-64.
35. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. and Hornik, K., (2004). Bioconductor: open software development for computational

biology and bioinformatics. Genome biology, 5(10), p.R80.

36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, et al. (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559-575.
37. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. and Shendure, J., (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics, 12:745.
38. Price, A., Helgason, A., Palsson, S., Stefansson, H., Clair, D., Andreassen, O., Reich, D., Kong, A. & Stefansson, K. (2009a). The impact of divergence time on the nature of population structure: An example from iceland. PLoS Genet. 5(6),e1000505.
39. Chimusa, E.R., Meintjies, A., Tchanga, M., Mulder, N., Seoighe, C., Soodyall, H. and Ramesar, R., (2015). A genomic portrait of haplotype diversity and signatures of selection in indigenous southern African populations. PLoS genetics, 11(3), p.e1005052.
40. Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Iny Stein, T., Bahir, I., Belinky, F., Morrey, C.P., Safran, M. and Lancet, D., (2013). MalaCards: an integrated compendium for diseases and their annotation. Database, ,bat018, 2013.
41. Wang X, Morris JN, Zhu X and Elston RC (2013) A variance component based multi-marker association test using family and unrelated data. BMC Genetics 2013, 14:17.
42. Jamsen KM, Zaloumis SG, Scurrah KJ, Gurrin LC (2012) Specification of Generalized Linear Mixed Models for Family Data using Markov Chain Monte Carlo Methods. J. Biomet. Biostat. S1:003. doi: 10.4172/2155-6180.S1-003.
43. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer,N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
44. Danecek, P. and McCarthy, S.A., 2017. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics, 33:2037-2039.
45. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. and Abecasis, G.R., 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature genetics, 44(8), p.955.
46. Kristin, C., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nature Reviews Genet. 3, 299-309.
47. Choi, S. (1977). Tests of equality of dependent correlation coefficients. Biometrika 64 (3): 645-647.
48. Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I., Birney, E., 2008. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. 18:1829-1843.

Figure legends

Figure 1: A) Principal component analysis of the Afrikaner's 5 affected case and 165 control individuals from CEU HapMap 3 Population as distinct groups (Materials and Methods) within five putative ancestral populations, the first principal component shows a close relationship between the Afrikaner and CEU, where most of the Caucasian Afrikaner are pooled toward the CEU. (B and D) the proportion of identity by state and identity by descent in the Caucasian Afrikaner and CEU. (C and E) show the distribution of individual minor allele frequency and heterozygosity in the Caucasian Afrikaner and CEU populations, respectively.

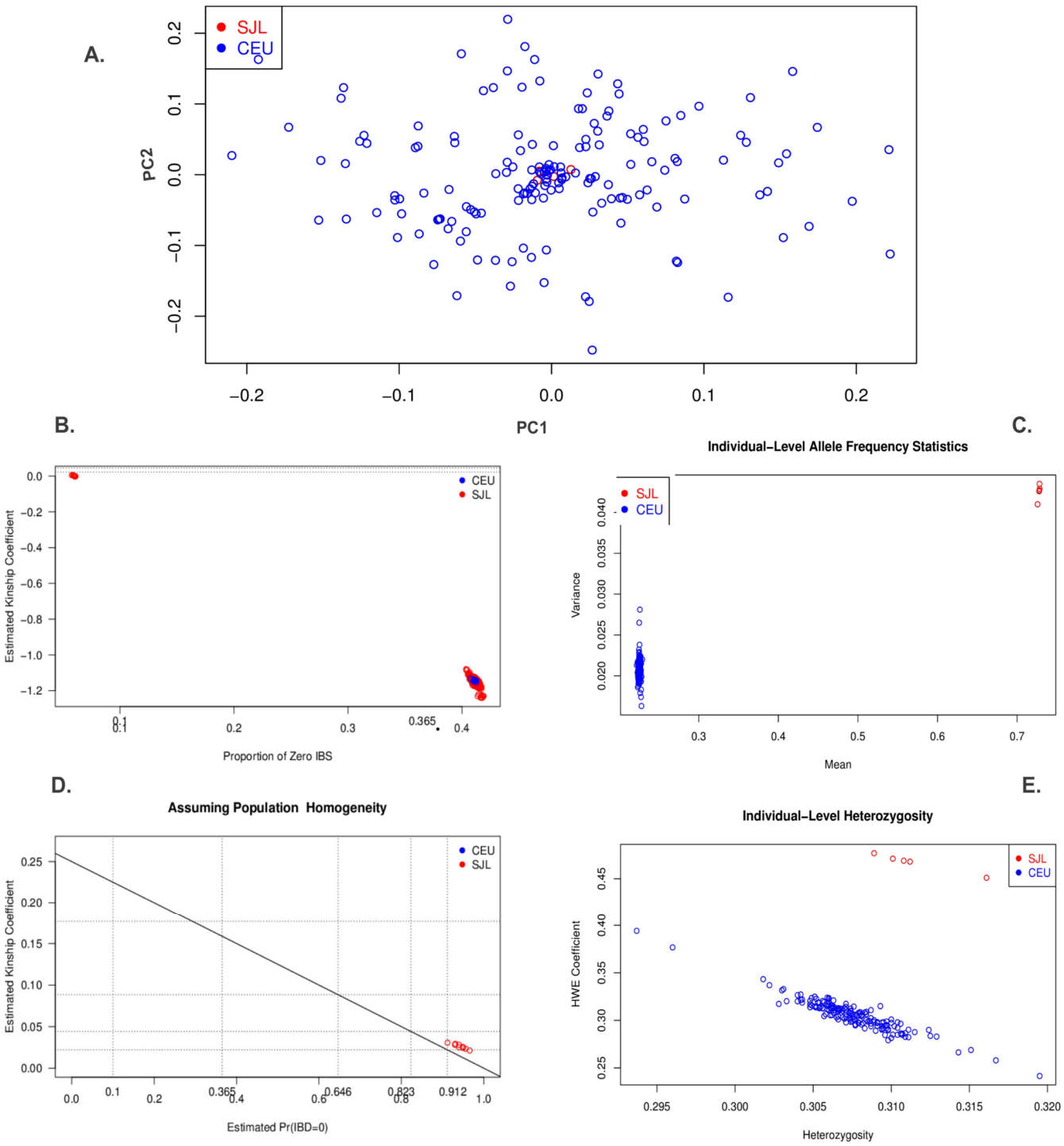

Figure 2: Manhattan plot of family-based association analysis (Materials and Methods) with the Spondyloepimetaphyseal dysplasia with joint laxity (SEMDJL) in 5 Caucasian Afrikaner individuals affected with SEMDJL and 165 unaffected samples from CEU HapMap3 using 900K SNP genotyping on Affymetrix 6.0.

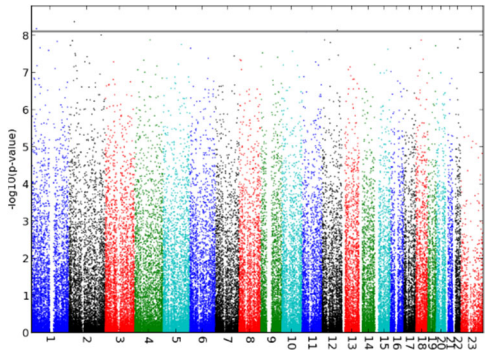

Figure 3: (A) Biological network of 3 identifying causative gene mutations, COL1A2 (7q21.3), MATN1 (1p35) and KIF22 (16p11.2) in SEMDJL patient. The interactions were obtained from the comprehensive human Protein-Protein Interaction network downloaded from the Protein Interaction Network Analysis platform. The plot shows that the sub-networks of interactions with COL1A2 (7q21.3), MATN1 (1p35) and KIF22 (16p11.2) overlap each other, consistent with the fact that the SNPs in each of these loci (COL1A2, MATN1 and KIF22) were in LD in the Caucasian Afrikaner. (B) Proportion of ancestral alleles in the SEMDJL patients and CEU control population indicating the higher proportion of ancestral alleles in CEU control population as compared to the SEMDJL patient population. This indicates a higher prevalence of derived alleles for the patient population as compared to the control population.

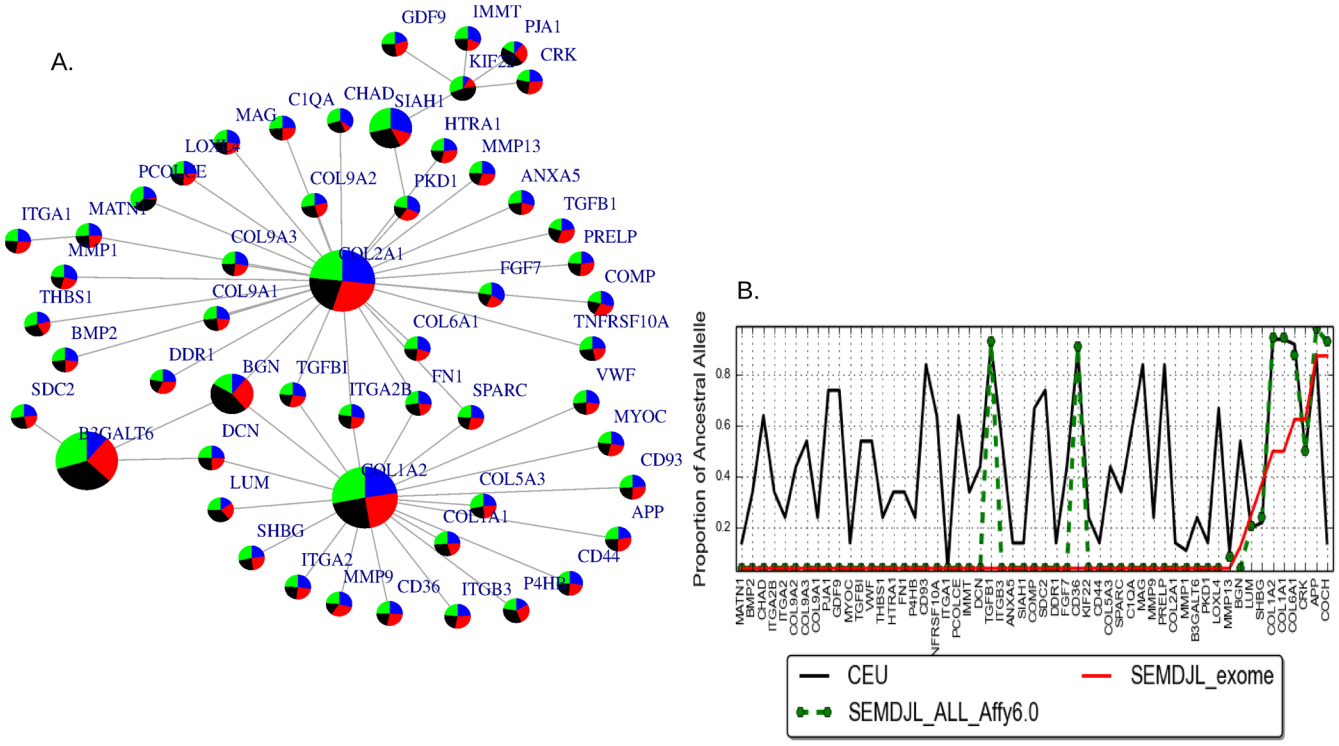

Table legends

Table 1: Top 10 genetic markers obtained from the family-based association analysis with the Spondyloepimetaphyseal dysplasia with joint laxity (SEMDJL) in 5 Caucasian Afrikaner individuals affected with SEMDJL and 165 unaffected samples from CEU HapMap3 using 900K SNP genotyping on affymetrix 6.0**.** A1, A2 are reference, derived alleles.

SNP	CHR	POS	MAF Afrikaner	MAF CEU	Closest Gene	Region	A1	A2	OR	P variance component	P Fisher	P family-based	P Recessive	P Dominant
rs3770890	2	36657992	0.03	0.04	*CRIM1*	p22.2	A	C	0.90	4.22e-09	9.99e-05	1e-05	8.96e-10	3.91e-06
rs16833849	1	31000821	0.03	0.04	*MATN1*	p35.2	C	T	0.92	6.54e-09	1.91e-08	1e-05	8.96e-10	3.91e-06
rs3849055	5	123190627	0.03	0.04	*KRT18*	q23.2	A	G	0.91	1.74e-08	9.99e-05	1e-05	8.96e-10	3.91e-06
rs11574837	17	43341110	0.03	0.04	*LOC100133991*	q21.31	A	G	0.92	2.18e-08	9.99e-05	1e-05	8.96e-10	3.91e-06
rs9504044	6	4306353	0.03	0.04	*PECI*	p25.1	A	G	0.92	2.19e-08	9.99e-05	1e-05	8.96e-10	3.91e-06
rs7098006	10	71196975	0.03	0.04	*ATP5G1*	q22.1	C	T	0.92	2.63e-08	9.99e-05	1e-05	9.24e-10	4.03e-06
rs10959895	9	11527480	0.03	0.04	*PTPRD*	p23	A	G	0.92	2.91e-08	9.99e-05	1e-05	8.968e-10	5.55e-06
rs12311948	12	107925779	0.03	0.04	*BTBD11*	q23.3	C	T	0.92	3.38e-08	9.99e-05	1e-05	9.24e-10	4.03e-06
rs16836954	2	137600865	0.03	0.04	*THSD7B*	q22.1	G	T	0.92	5.54e-08	9.99e-05	1e-05	8.968e-10	3.91e-06
rs2163932	4	21098882	0.03	0.04	*KCNIP4*	p15.31	C	T	0.92	8.35e-08	9.99e-05	1e-05	8.968e-10	5.55e-06

Table 2: Whole Exomic Sequence identifies gene mutations associated with the Spondyloepimetaphyseal dysplasia with joint laxity (SEMDJL) in 5 Caucasian Afrikaner individuals using different filter-based approaches (Materials and Methods). PolyPhen2: probably damaging if the score is between 0.909 and 1, and "possibly damaging" if the score is between 0.447 and 0.908, and "benign" is the score is between 0 and 0.446. Mutation Taster, LRT, PhyloP and SIFT are similar and all predict possible impact of an amino acid substitution on the structure and function of a human protein. The score ranges from 0 to 1. There are four possible predictions; A: disease causing automatic, D: disease causing, N:polymorphism or P: polymorphism automatic. For SIFT, a small score indicates a high chance for a substitution to damage the protein function. **In the last column provides a pattern of Linkage Disequilibrium (LD) between our identified candidate variants and *B3GALT6.***

Gene	CHR	Exomic Function	dbSNP137	PhyloP	SIFT, Prediction	PolyPhen2	LRT, Prediction	Mutation Taster, Prediction	LD with *B3GALT6*
MATN1	1	Synonymous SNV	***rs20566,rs7515295 rs2088473,rs7542828 rs7540540,rs1408149 rs7551867***	0.99, **C**	1.2e-04, T	0.98	0.99, **D**	0.98, **D**	**0.75**
COL1A2	7	Nonsynonymous SNV	***rs42524***	0.99, **C**	7.1e-05, **T**	0.99	0.99, **D**	0.99,**D**	**0.56**
KIF22	16	Synonymous SNV	***rs2450399***	0.99, ***C***	5.4e-03, T	0.71,	0.95, **N**	0.87,**P**	**0.52**

Table 3: Enrichment-test based on the obtained sub-network in Figure 3 using Enrichr (http://www.amppharm.mssm.edu/Enrichr/#) to identify enriched pathways associated with set of genes in sub-network in Figure 3, characterizing their biological processes, their molecular functions and the associated human phenotypes.

Term	Overlap	P-value	Adjusted P-value	Z-score	Overlap Genes with sub-network in Figure 3
Associated Pathway					
Hs_Senescence_and_Autophagy_WP615_29697	6/55	5.3e-06	6.8e-05	-1.9	***COL1A2;COL1A1;SPARC;THBS1;TGFB1;FN1***
Hs_Inflammatory_Response_Pathway_WP453_21632	4/27	8e-05	0.0008	-1.2	***COL1A2;COL1A1;THBS1;FN1***
Hs_Endochondral_Ossification_WP474_28899	4/60	0.001	0.009	-1.4	***MMP9;COL2A1;MMP13;TGFB1***
Hs_Matrix_Metalloproteinases_WP129_20953	3/28	0.001	0.009	-0.8	***MMP9;MMP13;MMP1***
Hs_Regulation_of_Actin_Cytoskeleton_WP51_20783	4/129	0.019	0.07	-1.2	***FGF7;ITGA1;CRK;FN1***
Associated Biological Process					
Extracellular matrix organization (GO:0030198)	32/359	0	0	-2.38	***COL2A1;DCN;TGFB1;COL9A1;APP;COL9A2;COL9A3;CD44;COMP;TGFB3;COL6A1;THBS1;FN1;MATN1;ITGA1;ITGA2;SPARC;COL5A3;BGN;COL1A2;COL1A1;ITGA2B***
Extracellular structure organization (GO:0043062)	32/360	0	0	-2.38	***MMP9;LUM;COL2A1;DCN;ITGB3;MMP1;SDC2;TGFB1;COL9A1;COL6A1;THBS1;FN1;P4HB;BMP2;MATN1;ITGA1;ITGA2;SPARC;COL1A2;COL1A1;ITGA2B***
Extracellular matrix disassembly (GO:0022617)	14/116	0	2e-14	-2.18	***COL9A1;COL9A2;COL9A3;CD44;MMP9;COL1A2;COL6A1;COL2A1;COL1A1;DCN***
Blood coagulation (GO:0007596)	19/472	1e-14	2.64e-12	-2.37	***MAG;KIF22;ITGA1;ITGA2;ITGB3;SPARC;ANXA5;MMP1;TGFB1;COL1A2;COL1A1***
Collagen catabolic process	11/74	2e-	3.23e-	-	***COL9A1;COL9A2;COL9A3;MMP9;COL1A2;COL6A1;COL2A1;COL1A1;COL5A3;MMP13;MMP1***

(GO:0030574)		14	12	2.15	
Associated Molecular function					
Collagen binding (GO:0005518)	16/62	0	0	-2.26	***MMP9;LUM;ITGA1;ITGA2;DCN;SPARC;COL5A3;MMP13;PCOLCE;DDR1;VWF;CD44;C***
Extracellular matrix structural constituent (GO:0005201)	12/68	0	1e-14	-2.2	***COL9A1;COL9A2;BGN;COL9A3;LUM;COMP;MATN1;COL1A2;COL2A1;COL1A1;COL5A3;PRELP***
Extracellular matrix binding (GO:0050840)	8/51	6.78e-11	3e-09	-2.3	***BGN;TGFBI;ITGA2;ITGB3;SPARC;DCN;THBS1;ITGA2B***
Glycosaminoglycan binding (GO:0005539)	11/190	2.91e-10	1e-08	-2.33	***APP;BGN;FGF7;CD44;COMP;DCN;COL5A3;THBS1;PCOLCE;PRELP;FN1***
Platelet-derived growth factor binding (GO:0048407)	4/11	2e-07	7e-06	-2.4	***COL1A2;COL6A1;COL2A1;COL1A1***
Associated Human Phenotype					
Osteoarthritis (HP:0002758)	9/42	4e-10	1.7e-07	-2.02	***KIF22;COL9A1;COL9A2;COL9A3;COMP;COL1A2;COL2A1;COL1A1;MMP13***
Platyspondyly (HP:0000926)	10/72	1.7e-09	3e-07	-1.9	***KIF22;COL9A1;COL9A2;COL9A3;B3GALT6;COMP;COL1A2;COL2A1;COL1A1;MMP13***
Autosomal dominant inheritance (HP:0000006)	25/1134	2e-07	1e-05	-2.28	***KIF22;COL2A1;DCN;ITGB3;TGFB1;COL9A1;COL9A2;APP;COL9A3;COMP;TGFBI;PKD1;COL6A1;MYOC;FN1;COCH;P4HB;BMP2;ITGA2;MMP13;VWF;CD36;COL1A2;COL1A1;ITGA2B***
Abnormality of the vertebral endplates (HP:0005106)	6/18	5.4e-08	5.5e-06	-2.01	***KIF22;COL9A1;B3GALT6;COMP;COL2A1;MMP13***
Mild short stature (HP:0003502)	5/12	3.2e-07	1.9e-05	-1.9	***COL9A2;COL9A3;COMP;COL1A2;COL1A1;COL1A2***

Epiphyseal dysplasia (HP:0002656)	5/14	6e-07	2.6e-05	-1.82	***COL9A1;COL9A2;COL9A3;COMP;COL2A1***
Irregular epiphyses (HP:0010582)	5/11	2e-07	1e-05	-1.67	***KIF22;COL9A2;COL9A3;COMP;COL2A1;COL1A2***