Genetic Identification of a Common Collagen Disease in Puerto Ricans *via* Identity-by-Descent Mapping in a Health System

G.M. Belbin[1,2,3], J. Odgis[2], E.P. Sorokin[4], M-C. Yee[5], S. Kohli[1], B.S. Glicksberg[2,3,6], C.R. Gignoux[4], G.L. Wojcik[4], T. Van Vleck[1], J.M. Jeff[1] , M. Linderman[2,3,†], C. Schurmann[1], D. Ruderfer[7,8,9,††], X. Cai[2], A. Merkelson[1], A.E. Justice[10], K.L. Young[10], M Graff[10], K.E. North[10], U. Peter[11,12], R. James[13], L. Hindorff[14], R. Kornreich[2], L. Edelmann[2], O. Gottesman[1,†††], E.E.A. Stahl[1,3,6,7], J.H. Cho[1,2,15], R.J.F. Loos[1,16], E.P. Bottinger[1††††], G.N. Nadkarni[1] , N. S. Abul-Husn[1,2,3,†††], E.E. Kenny[1,2,3,9*]

[1] The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[2] Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[3] The Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[4] Department of Genetics, Stanford University School of Medicine, Stanford, CA

[5] Carnegie Institution for Science, Department of Plant Biology, Stanford, CA.

[6] Harris Center for Precision Wellness, Icahn School of Medicine at Mt Sinai, New York, NY, USA.

[7] Broad Institute, Cambridge, MA, USA.

[8] Division of Psychiatric Genomics, Icahn School of Medicine at Mt Sinai, New York, NY, USA.

[9] Center for Statistical Genetics, Icahn School of Medicine at Mt Sinai, New York, NY, USA.

[10] Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[11] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

[12] Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA

[13] National Institute on Minority Health and Health Disparities, National Institutes of Health, Bethesda, MD, USA

[14] National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

[15] Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, NY

[16] The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[†] Present address: Department of Computer Science, Middlebury College, Middlebury, VT, USA

[††] Present address: Departments of Medicine, Psychiatry and Biomedical Informatics, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA

[†††] Present address: Regeneron Pharmaceuticals, 777 Old Saw Mill River Rd, Tarrytown, NY, USA

[††††]Present address: Berlin Institute of Health, Kapelle-Ufer 2, 10117 Berlin, Germany

46 *Correspondence to: eimear.kenny@mssm.edu

47

Abstract

Achieving confidence in the causality of a disease locus is a complex task that often requires supporting data from both statistical genetics and clinical genomics. Here we describe a combined approach to identify and characterize a genetic disorder that leverages distantly related patients in a health system and population-scale mapping. We utilize genomic data to uncover components of distant pedigrees, in the absence of recorded pedigree information, in the multi-ethnic Bio*Me* biobank in New York City. By linking to medical records, we discover a locus associated with genetic relatedness that also underlies extreme short stature. We link the gene, *COL27A1*, with a little-known genetic disease, previously thought to be rare and recessive. We demonstrate that disease manifests in both heterozygotes and homozygotes, indicating a common collagen disorder impacting up to 2% of individuals of Puerto Rican ancestry, leading to a better understanding of the continuum of complex and Mendelian disease.

Introduction

During the past two decades major advances in deciphering the genetic basis of human disease have resulted in thousands of disorders that are now understood at a genetic level[1,2]. This progress has led to the integration of genomic sequencing in clinical care, especially for the diagnosis of rare genetic disease[3,4], and clinical sequencing is increasingly offered to patients with known or suspected genetic disorders. In the past few years, large national and international efforts[5–7] have emerged to enable patients and health systems to share knowledge of rare genetic disorders and improve genetic testing, resulting in improved healthcare management and outcomes for patients. In

parallel, many large regional and national biobank efforts[8–10] are underway to enable the broad integration of genomics in health systems for genetic identification of disease[11]. Such efforts have recently revealed clinically actionable variants[11,12] and genetic disorders segregating at higher frequencies in general patient populations than previously suspected. The increased promulgation of genomics in health systems represents an opportunity to improve diagnostic sensitivity for more precise therapeutic intervention and better health outcomes[13].

Despite this progress, most genetic diseases are still under-diagnosed[14] or misdiagnosed.[15–17] A number of barriers exist for wholesale genetic testing and diagnoses, including incomplete standardized guidelines for interpreting genetic evidence of disease[18], variable penetrance or expressivity of phenotype[19], and that the causal variant may be missed or mis-assigned during testing[20]. The latter is a particularly pernicious problem in non-European populations due to systematic biases in large genomic and clinical databases[21,22]. These challenges have led several research groups to attempt to genetically identify disease by examining patient health patterns using data from the Electronic Health Record (EHR)[23,24]. EHRs contain comprehensive information on medical care throughout a patient's life, including medications, medical billing codes, physician notes and generated reports (i.e. pathologic, genetic and radiologic reports). EHRs have been used to clinically characterize well-known genetic disorders, but have been of limited success for the vast cadre of less-characterized or unknown disorders[25].

The gold standard of genetic disorder diagnosis involves testing both patient and family members to confirm Mendelian segregation of the suspected underlying pathogenic

variant[26–28]. However, as genomic data becomes more ubiquitous in health systems, it can be used to detect genetic relationships in the absence of known family and pedigree information. Specifically, components of pedigrees can be uncovered within the general population; particularly those that have experienced recent founder effects. Pairs of individuals who are related share genetic homology in the form of long genomic haplotypes. These haplotypes are considered to be identical-by-descent (IBD) if they are inherited from a common ancestor without any intervening recombination. The chance of any two people sharing a tract of their genome IBD decays exponentially, with a ~50% reduction in the chance of sharing per generation. However, when IBD sharing does occur, the length of an IBD segment can remain long even between distantly related individuals. In practice, long tracts of IBD (>3cM) can be accurately detected using genetic data between individuals with a common ancestor from the past 4-50 generations[29]. Detection of IBD haplotypes can allow for the identification of distantly related patients with a genetic disorder driven by a locus inherited from a founding ancestor who brought the disease mutation into a population[30–37]. This is the principle underlying population-scale disease mapping approaches that combine IBD sharing and statistical association to discover novel disease loci, so called IBD-mapping.

By detecting genetic relatedness, as inferred by IBD sharing, we hypothesized that we may be able to detect hereditary forms of disease in an EHR-linked biobank. With over 38000 participants, the Bio*Me* biobank, at the Icahn School of Medicine at Mount Sinai, New York City (NYC), is one of the most diverse cohorts ascertained at a single urban medical center under a uniform study protocol. Participants are largely from the local

Upper East Side, Harlem and Bronx communities, and represent broad ancestral, ethnic, cultural, and socioeconomic diversity**.** We initially focused on adult height, which is easily measurable, stable over the adult life course, and one of the most abundantly recorded clinical parameters in EHRs. Height is known to be highly heritable and polygenic[38,39], however, extremes of short stature can be caused by rare variants in single genes with large effect sizes[40]. Although, many genetic syndromes are known to cause short stature, most of the time no definitive etiology underlying short stature is found in patients. Here we used loci associated with genetic relatedness as measured by IBD to map a locus underlying extreme short stature in the Bio*Me* biobank, and linked it to a known, but little characterized, collagen disorder previously thought to be rare. By interrogating a large global diversity panel, we demonstrated that this variant is actually common in Puerto Rican populations. Furthermore, we leveraged the EHR to show significant musculoskeletal disease in both heterozygous and homozygous patients, indicating the disease is not simply a recessive disorder as had previously been thought. Finally, we showed how this work can generate broad insights for sustainable adoption and large-scale dissemination of genomic medicine.

Results

Detecting Patterns of Diversity, Founder Effects and Relatedness in the BioMe Biobank from New York City

The Bio*Me* biobank comprises a highly diverse cohort, with over 65% of participants self-reporting as Black/African-American or Hispanic/Latino, and over 35% born outside mainland US, representing more than 110 countries of origin. First we estimated

patterns of direct relatedness in a subset of Bio*Me* participants genotyped on the Illumina OmniExpress array (N=11212) by detecting pairwise identity-by-state using RELATEAdmix[41], a method that accounts for admixture in populations (**Figure 1- figure supplement 1**). We observed that 701 individuals had primary (parent-child, sibling) or secondary (avuncular, grandparental) relationship with another participant in BioMe, and we removed these individuals from all downstream analysis. Next we devised a strategy to divide the diverse Bio*Me* biobank into population groups for downstream analysis. We combined genotype data for Bio*Me* participants (N=10511) with 26 global populations from the 1000 Genomes project (N=2504)[42] and two additional panels of Native American (N=43)[43] and Ashkenazi Jewish populations (N=100) (see Methods). Using a common set of 174468 SNPS we performed principal component analysis[44] (PCA; **Figure 1 – figure supplement 2**). Based on both self-reporting and patterns of genetic diversity observed in Bio*Me* participants, we stratified individuals into four broad population groups. The first group self-reported as European American, but were also genetically identified as Ashkenazi Jewish (AJ; N=808) as they clustered distinctly with an AJ reference panel and separately from other European ancestry groups in PCA space (**Figure 1- figure supplement 3**). The other three groups we defined using self-reported race/ethnicity categories, African-American (AA; N=3080), Hispanic/Latino (H/L; N=5102) and European-Americans with no AJ genetic ancestry (Non-AJ EA; N=1270) (**Figure 1- figure supplement 3**). An additional 251 individuals who reported 'Mixed' (N=89) or 'Other' (N=162) ethnicity were excluded from further analysis.

To evaluate signatures of distant relatedness Bio*Me* biobank participants, we estimated sharing of genomic tracts IBD >3cM between every pair of individuals using the

GERMLINE software[45]. The minimum length of 3cM was chosen based on reports of elevated type I error in call rates of smaller lengths[46,47]. It is known that population-level rates of distant relatedness are observed to be particularly elevated after population bottlenecks (*i.e.* in founder populations)[48]. We summed the length of all IBD-tracts shared between a given pair of individuals if they shared more than one tract and examined the distribution of pairwise sharing at a population level. We observed elevated levels of distant relatedness in both the AJ (median summed length of IBD sharing within population=44.7cM; 95% C.I. = 44.66-44.82cM) and HL (16.2cM; 16.18-16.22cM) populations, compared to AA (3.77cM; 3.76-3.77cM) or non-AJ EA (4.5cM; 4.45-4.55cM) populations (Figure 1A). This is congruent with previous reports of founder effects in both AJ populations[49] and in some H/L populations[50].

Hispanic or Latina is a broad ethnic label encompassing myriad populations with origins in Northern, Southern or Central America, century-long roots in New York City, and genetic ancestry from Africa, Europe and the Americas. To explore the signature of a founder effect in the Bio*Me* H/L population, we leveraged self-reported and genetic information about sub-continental ancestry. By self-reporting, the H/L participants in Bio*Me* were born in New York City (NYC) (40%), Puerto Rico (24%), Dominican Republic (19%), Central/South America (12%), Mexico (2%) or other Caribbean Island (2%) (**Figure 1B**). We examined IBD tract length distributions within H/L sub-continental populations and observed that the founder effect was predominantly driven in the Puerto Rican-born group (**Figure 1C**). We assembled a cohort of Puerto Ricans including Bio*Me* participants who were either born in Puerto Rico or, were born in NYC and had 2 parents or 3-4 grand parents who were born in Puerto Rico (N=1245).

Approximately 5086 NYC-born H/L individuals did not have recorded parental or grandparental country-of-origin, therefore we also devised a selection strategy using PCA analysis. We identified Bio*Me* H/L participants on the cline between the African and European reference panels in PCA space coincident with Puerto Rican-born individuals. We excluded those on the same cline with ancestry from the Dominican Republic or another Caribbean Island **(Figure 1 – figure supplement 4),** and counted the remainder (N=1571) in the Puerto Rican group. In total, we estimated 2816 H/L in the BioMe discovery cohort were of Puerto Rican ancestry, and focused the downstream analysis on this group as the largest founder population in Bio*Me*.

Detecting a Locus Shared Identical-by-Descent Underlying Extreme Short Stature in Puerto Ricans

Next we tested the hypothesis that rare, recessive disease variants may have arisen to appreciable frequency in the Puerto Rican founder population. We linked genomic data to clinical data in the Electronic Health Record (EHR) of the Mount Sinai Health System. We focused on height, a stable and ubiquitous health measure. Clinically, rare instances of growth failure or 'short stature' may be caused by a large heterogeneous group of genetic disorders (i.e. skeletal dysplasias)[51]. We first extracted measures of height for the Puerto Rican adult population of Bio*Me* (mean age=55.3, standard deviation (s.d.)=16.1). After making exclusions based on age (>= 18 years old for women, >=22 years old for men, and < 80 years old in both sexes), mean height measurements (mean height=5' 8.2", s.d.=3.2" for men; mean height=5' 2.8" s.d.=2.8" for women) were consistent with those reported for Puerto Rican populations in a recent global study on height[52]. We noted that 56 Puerto Ricans met the clinical definition of

short stature[53] (range of short stature 5'1"-4'0" in men, and 4'8"-3'8" in women) defined as 2 standard deviations below the population-specific mean for men and women separately (**Figure 2 – figure supplement 1**).

To test for recently arisen, recessive variants underlying clinical short stature in Puerto Ricans, we implemented a previously published pipeline for 'IBD mapping'[31,54] (**Figure 2 – figure supplement 2**). We first clustered participants into 'cliques' of 3 or more individuals whom, at a given genomic region, shared overlapping homologous IBD tracts of at least 0.5cM in length. Membership in a clique indicates the sharing of a recent common ancestor at that locus, from which the homologous IBD tract was jointly inherited. Clustering of IBD into cliques in the Puerto Rican population (N=2816) yielded 1434421 IBD-cliques after quality control filters (see methods). The site frequency spectrum of IBD-cliques (**Figure 2 – figure supplement 3**) demonstrates an expected exponential distribution of clique sizes (of 3-77 haplotypes), representing a class of rare IBD haplotypic alleles (allelic frequency 0.0005-0.0137). To test whether any cliques of IBD haplotypes were significantly associated with height we performed genome-wide association of height as a continuous trait under a recessive model using PLINKv1.9[55,56], including the first five PCA eigenvectors as covariates (see Methods). We restricted analysis to homozygous IBD-haplotypes that were observed among at least 3 individuals (480 cliques in total). Adjusting for 480 tests (Bonferonni adjusted threshold $p<1x10^{-4}$) one IBD-clique achieved a genome-wide significant signal at the locus 9q32 (IBD-clique frequency=0.012; β=-3.78; $p<2.57x10^{-11}$) (**Figure 3A**), spanning a large mapping interval chr9:112MB-120MB. The clique contains 59 individuals, 56 of whom are heterozygous and 3 are homozygous for the associated IBD haplotype.

The three individuals driving the recessive signal, two women and one man, were less than 2.5 s.d. shorter (height reduction range 6”-10”) than the population mean for height in the Puerto Rican cohort (**Figure 3B**). The IBD-haplotypes driving the signal spanned a genic region with several candidate loci, and the minimum shared boundary overlapped a single gene, *COL27A1,* which encodes for Collagen Type XXVII, Alpha 1 (**Figure 3C**). We performed whole genome sequencing (WGS) of the three homozygous individuals, and an additional short-statured individual that we observed to possess a homozygous IBD haplotype that was both directly upstream of and highly correlated with the top IBD-clique. Individuals were sequenced to a depth of 4-18X coverage (**Supplementary file 1**). Examination of variants that were observed in at least 6 copies between the four individuals revealed a single candidate coding allele, a missense mutation in Collagen Type XXVII, Alpha 1 (*COL27A1*, g.9:*116958257.C>G,* NM_032888.1, p.G697R*,* rs140950220*)* (**Supplemental file 2**). *In silico* analysis suggest that this glycine residue is highly conserved, and that a molecular alteration to arginine at this position is predicted to be damaging (SIFT score=0.0; PhyloP score=2.673; GERP NR score=5.67). These findings are consistent with a recent report implicating the same *COL27A1* variant as causal for the rare orthopedic condition Steel syndrome in a Puerto Rican family[57]**.** First described in 1993, the main clinical features of Steel syndrome include short stature, bilateral hip and radial head dislocations, carpal coalition (fusion of the carpal bones), scoliosis, *pes cavus* (high arches), and dysmorphic features[58]**.**

To confirm the link between the IBD haplotype and the putative causal variant, we calculated the concordance between the IBD haplotype and carrier status of the *COL27A1*.pG697R variant by genotyping all of the homozygotes and carriers of the top IBD-clique in the recessive model (N=59), along with a panel of age- and sex-matched controls (N=59). This demonstrated 100% concordance between the *COL27A1*.pG697R variant and the significant IBD-haplotype in homozygotes (**Supplementary file 3**). We note that two Puerto Rican participants in the phase 3 1000 Genomes Project reference panel (1KGP) were carriers of the *COL27A1*.pG697R variant, raising the possibility that we may have been able to detect this association using more a traditional SNP association approach. Therefore, we performed genome-wide association in the same Puerto Rican cohort (N=2622) by first imputing the 1KGP panel and re-running the recessive test as described above (n=10007795 imputed and genotyped SNPs with an INFO score of > 0.3 and at least two observations of homozygotes). The recessive model appeared to be well calibrated (λ=1.02), however, we observed no genome-wide significant signal (**Figure 2 - figure supplement 4**). Association with the *COL27A1*.pG697R variant was the 11775th most significant association (MAF=0.014: β=-3.0; p<0.001). Upon examination of the correlation between the imputed *COL27A1*.pG697R and the true carrier status of homozygotes, we noted a concordance of only 66.67%, indicating that the IBD haplotype was a better tag of the true *COL27A1*.pG697R homozygous state compared to 1KGP imputation in the Puerto Rican cohort (**Supplemental file 3)**.

The association between *COL27A1*.pG697R and clinical short stature was replicated using an independent cohort of 1775 individuals, from Bio*Me*, of self-reported Puerto

Rican ancestry, that were genotyped on the Illumina Infinium Multi-Ethnic Genotype Array (MEGA) as part of the Polygeneic Architecture using Genomics and Epidemiology (PAGE) Study. The *COL27A1*.pG697R (rs140950220) variant was directly genotyped on MEGA, and an association of the variant under a recessive model resulted in a strong signal of association (allele frequency=0.017; β=-3.5; s.e.=0.70; $p<4.87x10^{-07}$). The replication analysis revealed 51 additional Bio*Me* carriers and two individuals that were homozygous for the variant. Both carrier and affected status was confirmed *via* independent genotyping and Sanger sequencing (see Methods). The two homozygous participants were both short statured (2.4 and 3.6 s.d from the sex specific population mean).

Evidence from Electronic Health Records Supports Suspected Cases of Steel Syndrome

To determine whether there was any clinical evidence to validate the link between the *COL27A1*.pG697R variant and Steel syndrome, a clinical expert manually reviewed the electronic health records (EHR), including clinical diagnoses, surgical procedures, and radiology reports, of the five participants (3 women, 2 men, age range 34-74 years) homozygous for the *COL27A1*.pG697R variant. Of note, there was no evidence that any of the five patients had a clinical diagnosis of Steel syndrome. In all five individuals, however, we found EHR-documented evidence of several previously described Steel syndrome characteristics, including developmental dysplasia of the hip (or congenital hip dysplasia), carpal coalition, scoliosis, and cervical spine anomalies (**Table 1**)[58,59]**.** The incidence of cervical spine anomalies, including cord compression and spine

surgeries, was higher than previously reported (four out of five patients). There was also evidence of other significant musculoskeletal complications, including lumbar and thoracic spine anomalies in three patients, knee replacements in two patients (both under age 50), and joint degeneration or arthritis in four patients. Together, these data help further our understanding of Steel syndrome-associated characteristics and potential complications that can occur later in life.

Functional Investigation of COL27A1.pG697R

To understand the biological mechanism underlying Steel syndrome, we investigated the functional role of the *COL27A1* gene. *COL27A1* is a fibrillar collagen, which are a class of collagens that contribute to the structural integrity of the extracellular matrix[60]. Enrichment of *COL27A1* RNA expression in vertebrae, as well as long bones, eyes, and lungs has previously been observed in embryonic mice[60]. A mouse deletion of 87 amino acids of the *COL27A1* homolog exhibited severe chondroplasia consistent with clinical features observed in homozygotes[61], a similar musculoskeletal phenotype was observed in knockdown of the *col27a1a* and *col27a1b* genes in zebrafish[62]. Type alpha-1 collagen genes, of which *COL27A1* is a member, contain a conserved Gly-Xaa-Yaa repeat in their triple helical domain[63]. Therefore, we hypothesized that the *COL27A1*.pG697R variant may similarly disrupt stability of the *COL27A1* triple helix.

To test this hypothesis we modeled the effect of a glycine-to-arginine substitution in the structure of a prototypical collagen peptide[64]. We observed that the glycine residues occupied the center of the crowded triple helix, and that substitution for a bulkier arginine would likely destabilize helix formation through steric hindrance (**Figure 2 –**

figure supplement 5). These data provide support for a functional model of the pathogenicity of *COL27A1.*G697R through destabilization of the triple helix, which may occur within developing spinal chords, long bones, and other tissues, resulting in the observed clinical features in homozygotes. We note that many other collagen disorders, including Ehlers-Danlos syndrome[65–67], Alport syndrome[68,69] and Osteogenesis Imperfecta[70,71], are driven by molecular alterations of a glycine in the triple helix of the underlying collagen genes. However, all of these disorders are inherited under an autosomal dominant mode, in contrast to Steel syndrome, which has only been reported as a recessive disease. This analysis raises the question of whether some and/or milder clinical features of Steel syndrome may be present in carriers.

Assessing the Health Records of COL27A1.pG697R Carriers Reveals Evidence of Musculoskeletal Disease

To test for clinical features of Steel syndrome in *COL27A1*.pG697R carriers, we performed two analyses using EHR data. The first was a test for associated medical billing codes (ICD9s) with *COL27A1*.pG697R carrier status, or Phenome-Wide Association Study (PheWAS)[72,73]. PheWAS analysis is often performed using a general linear model (GLM), however standard implementations often do not account for scenarios where there is a large imbalance between per-test number of cases and controls, rare variants/ICD9s or the presence of elevated distant relatedness. Therefore, in addition to the GLM, we also ran three other score based tests; (i) that use saddlepoint approximation (SPATest)[74] to account for case:control imbalance; (ii) a linear mixed model (GCTA)[75] to account for distant relatedness; and (iii) a test that

incorporates a bias-reduction for small numbers of observations (Firth test)[76]. Each test was run using ICD9 codes in all individuals of Puerto Rican ancestry (N=106 *COL27A1*.pG697R carriers and N=4480 non-carriers). The ICD9 code was set as the outcome variable and *COL27A1*.pG697R as the primary predictor variable, including age, sex and the first five PCAs as covariates in all tests. To avoid spurious associations, we restricted the analysis to diagnosis codes with at least 3 observations (n=367 ICD9 codes) amongst carriers.

Results of the GLM test are shown in **Figure 3** and **Table 2**. Of the five significantly associated ICD9 codes (False Discovery Rate (FDR)<0.05), three involved the musculoskeletal system 730.08 ($p_{GLM}<7.1x10^{-6}$; odds ratio (OR)=34.5; 95% Confidence Interval (CI)=7.4-162), 721.0 ($p_{GLM}<6.6x10^{-5}$; OR=5.4; CI=2.4-12.3), and 716.98 ($p_{GLM}<4.4x10^{-4}$; OR=5.8; CI=2.2-15.3). ICD9 730.08 encodes for "acute osteomyelitis, other specified sites". Manual review of chart records for these patients revealed that this code referred to vertebral osteomyelitis in the three carriers with the ICD9 code. ICD9 721.0 encodes for cervical spondylosis without myelopathy. Cervical spondylosis refers to degenerative changes of the cervical spine, which can eventually progress to encroach on the cervical canal, causing myelopathy (spinal cord injury). A third diagnosis code, 716.98, encodes for "arthropathy, unspecified, or involving other specified sites". Manual review of chart records for these patients revealed that this code referred to knee arthropathy in all four patients. Finally, two other ICD9 codes were significantly associated with the *COL27A1*.pG697R variant; 622.10 ($p_{GLM}<1x10^{-4}$; OR=5.4; CI=2.3-12.6), which encodes for cervical dysplasia, and 789.1 ($p_{GLM}<2.1x10^{-4}$; OR=11.6; CI=3.2-42.2), which encodes for hepatomegaly. Presently, it is unclear

366 whether these two are related to a *COL27A1*.pG697R carrier phenotype, or are
367 spurious associations.

368 We observed over inflation in the distribution of the PheWAS test statistic, measured by
369 lambda (λ), for all four score based models (λ_{GLM}=1.59; $\lambda_{SPATest}$=1.20; λ_{GCTA}=1.36;
370 λ_{Firth}=2.09), indicating that no single model fully accounts for the confounding effects of
371 distant relatedness, case:control imbalance and rare variants/ICD9s (**Figure 3 – figure**
372 **supplement 1**). The code linked to vertebral osteomyelitis (730.08) was the top signal
373 in all tests ($p_{SPATest}<1.4x10^{-4}$; $p_{GCTA}<7.9x10^{-10}$; $p_{Firth}<1.5x10^{-9}$), but only remains
374 significant after genomic control adjustment in one of the tests ($p_{GCTA\_adjusted}<4.6x10^{-5}$).
375 Neither codes linked to cervical spondylosis (721.0; $p_{SPATest}<3.0x10^{-3}$ (rank=3rd);
376 $p_{GCTA}<3.3x10^{-3}$ (8th)) or knee arthropathy (716.98; $p_{SPATest}<0.022$ (21st); $p_{GCTA}<3.5x10^{-3}$
377 (9th); $p_{Firth<}0.001$ (35th))) were significant after genomic control correction. Therefore,
378 while PheWAS analysis provided preliminary support of Steel syndrome-associated
379 clinical features in carriers, best practices for PheWAS models for rare variants/ICD9
380 codes, and in the presence of population structure, remains an open problem for the
381 genomics community. It is also possible that some relevant clinical features of Steel
382 syndrome might be poorly captured by or absent from medical billing codes.

383 To evaluate the preliminary evidence from the PheWAS analysis, we performed a
384 second analysis of EHR data that focused on a comprehensive manual chart review to
385 examine for evidence of Steel syndrome characteristics in the *COL27A1*.pG697R
386 carriers in the same manner as performed for homozygotes. We limited the analysis to
387 carriers below the age of 55 (N=34; mean age 41.8 years) to reduce confounding from

388 age-related related symptoms of spine and joint pain. We also selected 31 age and sex
389 matched Puerto Rican non-carriers for comparison (mean age 40.6 years). Utilizing the
390 same criteria used to characterize Steel syndrome cases, we found no evidence of
391 clinical short stature or hip dislocation in carriers, but did observe a trend of elevated
392 rates of major joint and spine degradation (**Table 1**). In general, 38% (13/34) of carriers
393 showed evidence of spine degeneration varied from severe (multiple level cord
394 compression and neurological symptoms necessitating corrective surgery) to moderate
395 (lower back pain with no neurological symptoms managed with physical therapy and/or
396 pain medication) compared to 13% (4/31) of non-carriers (Fishers exact test $p<0.03$).
397 Specifically, we found an increased risk of cervical stenosis in 15% (5/34) of carriers
398 compared to 0% (0/31) of controls ($p<0.05$). Although not reaching statistically
399 significance, we show a trend of 2-fold higher rates of scoliosis (24%; $p<0.35$), arthritis
400 (38%; $p<0.1$), and lumbar spine degradation (29%; $p<0.25$) in carriers compared to non-
401 carriers and previous published reports in similar age groups[77,78]. Together these data
402 suggest an appreciable burden of joint and spine degradation in *COL27A1*.pG697R
403 carriers (**Table 1**).

404 *Worldwide Frequency and Demographic History of COL27A1.pG69R*

405 Having genetically identified and clinically characterized a previously little-known
406 disease variant, we next investigated which populations were at risk for harboring the
407 allele. We assessed the carrier frequencies of the *COL27A1*.pG697R variant in global
408 panel of 51745 individuals from Africa (N=376), the Americas (N=45685), Asia
409 (N=5311), Europe (N=209), the Middle East (N=163) and Oceania (N=28) genotyped on

MEGA in the PAGE Study. This included; 13050 in the Multi-Ethnic Cohort (MEC)[79] Study; 12327 in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL)[80]; 12852 in the Women's Health Initiative (WHI) Study[81]; 13044 additional Bio*Me* biobank participants (including the 1775 Puerto Ricans on MEGA described above); and a Global Reference Panel from Stanford University including the Human Genome Diversity Panel[82] (N=986, see Methods). Combined, the PAGE and Bio*Me* dataset represented 57316 individuals from 112 global populations (**Supplementary file 4**). The *COL27A1*.pG697R C allele was present in 183 copies (173 heterozygous carriers and 5 homozygous cases). We estimate the carrier rate of *COL27A1*.G697R to be 1:51 in Puerto Rican-born individuals (minor allele frequency (MAF)=1.1%); 1:9 in individuals born on the island of St. Thomas (MAF=11%); 1:346 in Hispanic/Latinos in the US (MAF=0.29%) and 1:72 in Bio*Me* Hispanic/Latino populations from New York City (MAF=0.7%); and 1:746 in individuals born in the Dominican Republic (MAF=0.067%) (**Figure 4A)**. We note that only 9 people were assayed from St. Thomas, so the high carrier frequency estimate could be biased by small sample size. Finally, the variant is present in only 4 copies in the 60,706 exomes in the ExAC database[83], likely due to differences in the populations comprising both datasets.

To predict what other populations might be at risk for Steel syndrome, we explored the locus-specific demographic history in carriers of the *COL27A1*.pG697R risk haplotype. First, by visual inspection, we were able to discriminate a single haplotype of 107.5kb in length that contained 55 SNPs, which uniquely tagged the *COL27A1*.pG697R variant (R^2=1). This haplotype was present only in individuals born in Puerto Rico (N=25), the Dominican Republic, (N=2), Columbia (N=1), New York City (N=40) and St. Thomas

(N=1). Genotyping determined that only the haplotype carriers from Puerto Rico, New York City and St. Thomas also carried the *COL27A1*.pG697R variant (N=56 in total). Second, we inferred continental ancestry along the genomes of the three Puerto Rican homozygotes in the discovery cohort. Local ancestry inference is the task of assigning continental ancestry to genomic segments in an individual with recent ancestors from multiple continents. For the Puerto Rican homozygotes, we estimated local haplotypic similarity with a reference panel of African, European and Native American genomes using RFMix[84]. Examination of local ancestry on the background of the IBD haplotype in all three homozygous individuals revealed all to be homozygous for Native American ancestry, suggesting the *COL27A1*.pG697R arose on a Native American haplotype (**Figure 4B**).

To test whether the disease variant arose via genetic drift or selection, we used the IBDNe software[85] to estimate the historical effective population size (N_e) of the Puerto Rican discovery cohort (N=2816) (**Figure 4C**). The IBDNe software calculates the effective population size of a given population over past generations by modeling the distribution of IBD tract lengths present in the contemporary population. The analysis suggested evidence of a strong bottleneck in Puerto Ricans approximately 9-14 generations ago, with the smallest effective population sized dating approximately 12 generations ago (estimated N_e=2580, 95% C.I 2320-2910). This is consistent with the timing of European immigration and slave trading on the Island, resulting in admixture and population bottlenecking, followed by demographic growth post-contact[50,86]. Finally, to see if there was evidence that the locus had undergone a recent selective sweep we calculated the integrated haplotype score (iHS)[87,88] across chromosome 9 in phased

456 genotype data for Bio*Me* Puerto Rican samples, but did not observe evidence of
457 selection at the locus (**Figure 4 – figure supplement 1**). Together, this evidence
458 suggests that the *COL27A1*.pG697R variant arose in the ancestral Native American
459 populations that peopled the Caribbean, which underwent a strong bottleneck during the
460 period of colonization, which may help explain the prevalence of this disease in
461 amongst contemporary Puerto Rican populations.

462 **Discussion**

463 Here we describe a new approach to utilize genomic data in health systems for
464 identifying and characterizing genetic disorders, the cornerstone of which is the ability to
465 identify related individuals in the absence of recorded pedigree or genealogy
466 information. By linking medical records of distantly related patients, identified by shared
467 tracts of genetic homology identical-by-descent (IBD), we discovered a recessive
468 haplotype on 9q32 conferring extreme short stature. Whole genome sequencing
469 revealed that a mutation (Gly697Arg) in the *COL27A1* gene had been previously
470 implicated as the genetic variant underlying Steel syndrome[57–59]. Population screening
471 indicated that the disease variant is more common than previously thought in people
472 with Puerto Rican ancestry, and in some other Caribbean populations, and very rare or
473 absent elsewhere in the world. Extensive analysis of clinical records confirms almost all
474 features of the recessive disorder in cases, and reveals potential complications that can
475 occur later in life. An agnostic survey of the medical records of carriers, supplemented
476 by manual chart review, indicates evidence of joint and spine degradation in
477 heterozygotes. Biochemical modeling suggests that *COL27A1.*G697R disrupts a

478 conserved triple helix domain of the alpha-1 collagen in a mechanism similar to
479 dominant forms of other collagen disorders[63]. Taken together, this study indicates that a
480 single mutation in the *COL27A1* gene underlies a common collagen disorder impacting
481 up to 2% of people of Puerto Rican ancestry.

482 This is consistent with our finding, supported by previous work[50], demonstrating a
483 founder effect in Puerto Rican populations. Despite segregating at an estimated carrier
484 rate of 1:51, the *COL27A1.pG697R* variant was first described very recently[57]. This
485 suggests that there may be other highly penetrant disease variants segregating at
486 appreciable frequencies in Puerto Rican populations[89–94], and other understudied
487 founder populations, the discovery of which could lead to new disease variants and
488 biology. Indeed, although *COL27A1* was first implicated as the Steel syndrome disease
489 locus in an extended family from Puerto Rico recently[56], other variants in *COL27A1*
490 have since been linked to Steel syndrome in Indian[95] and Emerati[96] families revealing
491 additional clinical features of the disease such as hearing loss. In our own health
492 system, approximately 190,000 patients of Puerto Rican descent are treated annually[97].
493 We estimate that up to 80 may have the severe homozygous form of the disorder and
494 that the milder heterozygous form could be found in up to 1200 patients. A search of
495 progress notes, discharge summaries, and operative reports of over 4 million patients in
496 the Mount Sinai data warehouse discovered mentions of the text term “Steel Syndrome”
497 in 42 patient records. However, all of these patients were on dialysis for end stage renal
498 disease, indicating that this mention was a misspelling of vascular Steal Syndrome,
499 which is common in dialysis patients. This suggests that Steel syndrome might be
500 largely undiagnosed. Attempts are currently being made to re-contact Bio*Me*

501 participants with suspected Steel syndrome, and a genetic test is now available at
502 Mount Sinai (website: http://sema4genomics.com/products/test-catalog/).

503 This study highlights the benefits of incorporating statistical and population genetics
504 approaches in medical genetic research. First, we demonstrated that leveraging distant
505 relationships *via* IBD mapping was better powered for discovery of the *COL27A1* variant
506 compared to a more typical GWAS approach (i.e. genotype, imputation, and SNP
507 association). As sample sizes increase in health systems and biobanks, the odds of a
508 new individual being a direct or distant relative of someone already in the database
509 increases exponentially[98], enabling the detection of shared haplotypes harboring rarer
510 causal variants and better-powered IBD mapping studies. Second, we inferred that
511 *COL27A1*.pG697R variant arose on a Native American haplotype, and we estimate that
512 the allele may have segregated at a carrier frequency of 25-30% in pre-Columbian
513 Taíno populations and/or been driven to it's current frequency by a bottleneck that
514 occurred during the early days of colonization in Puerto Rico. Therefore this study not
515 only helps estimate population attributable risk of *COL27A1*.pG697R in Puerto Rican
516 populations, but also to predict other populations potentially at risk, including other
517 Caribbean and Taíno populations. Targeted population screening of *COL27A1*.pG697R
518 could potentially provide personalized health management, surveillance for associated
519 complications, guidelines for intervention (particularly in newborns[59]), and improved
520 reproductive choices.

521 This work also highlights some of the current challenges in the emerging field of
522 genomic medicine. We demonstrated that evidence from EHRs could be readily
523 extracted and retrospectively used to characterize clinical features of a musculoskeletal

524 genetic disorder. However, features of many other genetic disorders may not be
525 detectable via routine clinical exam, lab tests and radiologics, and may not be amenable
526 to such an approach. Furthermore, statistical methods for population-scale disease
527 variant discovery, which were predominantly developed for cohorts collected for genetic
528 research, may not be optimally calibrated for discovery in patient populations
529 encountered in health systems. Finally, many genetic disorders are very rare, or have
530 more complex genetic underpinnings, which would reduce power for detection using the
531 strategy we have described. However, recent efforts, such as the Precision Medicine
532 Initiative, that focus on the broad adoption of genomics in medicine, combined with
533 international efforts to catalog rare genetic diseases, are primed to increase the rate of
534 incidental genetic diagnosis of disease.

535 In summary, this work demonstrates the utility of biobanks for exploring full medical
536 phenomes, and highlights the importance of documenting a wider spectrum of genetic
537 disorders, in large and diverse populations of humans. In particular, this method
538 provides a bridge between classical medical genetic methods and those employed in
539 population-level GWAS. Here we note that the *COL27A1* variant is very rare in current
540 large-scale genomic databases used for clinical research. Thus traditional association
541 strategies and ascertainment bias focused on populations of European descent would
542 have failed to identify and characterize this disorder and its public health burden. As
543 ours and other recent studies have demonstrated, EHR-embedded research will be
544 increasingly important for disentangling the pathology of rare genetic disorders, and
545 understanding the continuum of complex and Mendelian disease. As studies grow in
546 size, and healthcare systems learn to leverage the wealth of information captured in the

EHR, there is a need to provide relevant medical information to any patient entering the clinic anywhere in the world. Methods like that described here allow for precision medicine with a truly global outlook.

Materials and Methods

BioMe Biobank Program

Study participants were recruited from the Bio*Me* Biobank Program of The Charles Bronfman Institute for Personalized Medicine at Mount Sinai Medical Center from 2007 onward. The Bio*Me* Biobank Program (Institutional Review Board 07–0529) operates under a Mount Sinai Institutional Review Board-approved research protocol. All study participants provided written informed consent. Of the approximately 38000 participants currently enrolled in Bio*Me,* N=10511 unrelated are genotyped on the Illumina Infinium OmniExpress (OMNI) array. 5102 of these participants self-report as “Hispanic or Latino”, 3080 as “African American or African”, 2078 as “White or Caucasian”, 89 as “Mixed” and 162 as “Other”. Country of origin information is available for all but N=5 participants, with N=6553 reporting being born in the United States and the remaining N=3953 report being born outside of the US. Parental and grandparental country of origin information is only available for a small subset of individuals genotyped on the OMNI array (N=43). An additional N=10471 participants were genotyped on the Illumina Infinium Multi Ethnic Genotyping array (MEGA) v1.0. Of these, approximately 4704 self-reported “Hispanic or Latino” ethnicity, 3143 self-reported as “African American or African”, 22 self-reported as “White/Caucasian”, 708 self-reported as “Mixed” and 1894 self-reported as “Other”. Country of birth information was available for all but a small

number of participants (N=228), with 5190 reporting being born in the United States, and the remaining 5053 self-reporting being born elsewhere. Parental and grandparental country of origin information is available for 4323 individuals genotyped on the MEGA array.

OmniExpress Genotyping and QC

Genotyping of 12749 Bio*Me* participants was performed on the Illumina Infinium OmniExpress plus HumanExome array. Calling was performed using the GenomeStudio software. A total of 1093 individuals were removed prior to zCall due to plate failure (N=672), unambiguous discordance between genetic and EHR recorded sex (N=693), a call rate of <98% (N=834), or deviances in levels of heterozygosity (N=773 in total). This was defined as having either an inbreeding coefficient outside the range -0.1 and 0.3 for common alleles (MAF >1%), or between 0.4 and 0.9 for rare alleles (MAF <1%). Additional quality control of 11656 individuals was performed using PLINK1.7[55] (RRID:SCR_001757). An individual with a call rate of <99% was also excluded (N=1), along with intentional genetic duplicates (PiHat >0.8, N=444). Site-level quality control consisted of the removal of SNPs with a call rate of <95% (n=42217), and the removal of sites that were significantly out of Hardy-Weinberg equilibrium ($p<1x10^{-5}$; n=39660) when calculated for self-reported EA, AA and H/L separately. Palindromic sites and those that deviated considerably from the 1000 Genomes project allele frequencies (<40% *versus* >60%) were also removed to ensure uniform stranding across datasets. After QC steps, 11212 participants and 866864 SNPs remained for downstream analysis.

591 *IBS Relatedness Estimates and Estimation of Inbreeding Co-efficient*

592 Pairwise IBS-based relationship estimates were derived for Bio*Me* participants
593 (N=11212) using the RELATEAdmix software[41], which accounts for inflation of IBS
594 statistic due to admixture linkage disequillibrium in admixed populations**.** To include
595 allele frequency information and global ancestry proportions from ancestral populations
596 relevant for each admixed population in the analysis. These were estimated using
597 ADMIXTURE[99] (RRID:SCR_001263) H/L samples were merged with the Utah
598 Residents (CEPH) with Northern and Western European Ancestry (CEU; N=100),
599 Yoruba in Ibadan, Nigeria (YRI; N=100) and Native American (NA; N=43; including
600 Nauha, Ayamaran, Mayan and Quechan individuals) and used as input for the
601 ADMIXTURE software, which was run unsupervised at k=3. ADMIXTURE analysis
602 confirmed that NA reference panel comprised > 99% proportion Native American
603 genetic ancestry (**Figure 4 – figure supplement 2**). European-American (EA)
604 individuals were merged with the CEU and a panel of self-reported Ashkenazi Jewish
605 individuals genotyped on OMNI from Bio*Me* (AJ; N=100) and run unsupervised at k=2.
606 AA samples were merged with the CEU and YRI reference panels and run
607 unsupervised at k=2. After intersecting with reference panels 99296 SNPs were used as
608 the input for RELATEAdmix..

609 *Principal Component Analysis*

610 Principal Component Analysis (PCA) was performed using the SMARTPCAv10210
611 software from the EIGENSOFTv5.0.1 (RRID:SCR_004965)[44] in 10511 unrelated Bio*Me*
612 participants. Regions containing the Human Leukocyte Antigen (chr6: 27000000-

35000000 (NCBI37/hg19)), Lactase gene (chr2:135000000-137000000 (NCBI37/hg19)) and a common inversion (chr8:6000000-16000000 (NCBI37/hg19)), all of which are regions known to confound PCA analysis were removed from the genotype data prior to analysis. Data were merged with a reference panel of 2504 individuals from Phase 3 of the 1000 Genomes project (RRID:SCR_006828)[42] that was constructed by extracting OmniExpress sites from whole-genome sequence data. Following this, a further two other relevant reference panels were added: a the NA (N=43), and a AJ (N=100) panels described above. A total of 174468 SNPs remained after intersecting the data with these reference panels.

Identity-by-Descent Tract Inference and Clustering

Phased genotype data were filtered to MAF > 0.01 and converted to PLINK format using the FCGENE software[100] (we avoided using PLINK software for the conversion process in order to retain the phase information). Recombination maps from HapMap II (Build GRCh37/hg19) were intersected with the genotyped sites (n=490510 SNPs). GERMLINE (RRID:SCR_001720)[45] was used to infer tracts of identity by descent > 3cM across all pairs of Bio*Me* individuals (N=11212) using the following flags*: "-min_m 3 -err_hom 0 -err_het 2 -bits 25 –haploid".* IBD haplotypes that fell within or overlapped with centromeres, telomeres and regions of low complexity were removed from the GERMLINE output using an in-house Ruby script. Additional quality control measures consisted of the exclusion of regions of the genome where the depth of IBD-sharing (that is, the number of pairwise IBD-haplotypes that contain a given locus of the

genome) exceeded 4 standard deviations from the genome-wide mean (**Figure 2 – figure supplement 6**).

IBD clustering to identify 'cliques' of three or more IBD haplotypes shared between multiple individuals was then performed using the efficient connect-component-based clustering version of the Dash Associated Shared Haplotypes algorithm (DASH)[101], using the default parameters. As a further quality control measure IBD-sharing 'cliques' inferred by DASH that exhibited excessive sharing (which we defined as clique membership that exceeded 4 s.d. above the genome-wide mean) were removed (**Figure 2 –figure supplement 7**). Data was outputted from DASH in PLINK tped format, and alleles were encoded as; homozygote member in a clique as "2", heterozygote member as "1" and everyone else not a member in the clique encoded as "0".

Population-level IBD sharing

We calculated the length of any pairwise IBD tract (or sum of the lengths if a pair of individuals shared more than one tract IBD) for each IBD sharing pair within each population to obtain an estimate of the mean and variance of pairwise sharing per population. To compare the tract length distribution between populations (of size *N*), we first binned pairwise IBD tracts by length bin in 0.01cM increments. We then summed the number of pairwise IBD tracts falling into each length bin (*x*), and divided this number by the number of possible pairwise IBD sharing for each population: N*(N-1)/2.

Height Measurement and Transformation

655 A self-reported measurement of height in feet and inches was recorded for each
656 participant at enrollment into the Bio*Me* program. Raw height data were stratified on the
657 basis of sex for all individuals who were inferred to be of Puerto Rican ancestry
658 (N=2816). Height data was then log transformed and converted to age-adjusted Z-
659 scores. Participants were excluded on the basis of age reported at the point of
660 enrollment, with a minimum cut-off of 18 years old for females (N=0) and 22 for males
661 (N=0), and a maximum cut off of 79 years old for both sexes (N=194) leaving a total of
662 n=2622 PR.

663 *Association of IBD-cliques with height under a recessive model*

664 Association of IBD clique membership with height as a continuous trait was performed
665 under a recessive model using PLINKv1.9[56] using the *"--linear recessive*" flag. Age and
666 sex adjusted Z-scores for height were used as the outcome variable. IBD clique
667 membership was used as the primary predictor variable and the first five PCA
668 eigenvectors were used as covariates. The model was run across a total of 2622 PR
669 ancestry individuals and a total of 480 IBD-cliques where at least 3 individuals were
670 homozygous for the IBD haplotype.

671 *Genome Wide Association of Imputed Data under a Recessive Model*

672 Genotype data for all of the Bio*Me* individuals ascertained on the Illumina OMNI
673 Express array (N=11212) were phased together using SHAPEIT2[102,103]. Imputation was
674 subsequently performed in 5MB chunks using IMPUTE2 (RRID:SCR_013055)[104] *via* the
675 flags *'-Ne 20000 -buffer 250 -filt_rules_l 'ALL<0.0002' 'ALL>0.9998"* with a reference

676 panel derived from Phase 3 data from the 1000 Genomes project. A total of 46538253
677 SNPs were imputed from 828109 directly genotyped SNPs.
678
679 We ran a recessive GWAS on the same 2622 inferred Puerto Rican ancestry individuals
680 used in our recessive IBD-mapping model. The association was run over hard-called
681 data using the PLINKv1.9 software using the *"--linear recessive*" flag. Age and sex
682 adjusted Z-scores for height were used as the phenotypic outcome and the first five PC
683 eigenvectors were used as covariates. Analysis was restricted to SNPs with >=2
684 observations of individuals homozygous for the minor allele (as the only 2 of the 3
685 homozygotes had been imputed correctly), and SNPs with an INFO score of >=0.3
686 (n=10007795 SNPs in total).

687 *Whole Genome Sequencing*

688 Genomic libraries were prepared from DNA obtained for the four IBD homozygous
689 individuals. DNA was sheared to 300 bp on a Covaris E220, libraries were made using
690 the NEBNext Ultra DNA Library Prep kit for Illumina. The libraries were submitted for
691 Whole Genome Sequencing (WGS) at the Mount Sinai Genomic Core using the Illumina
692 HiSeq 2500 system, performed by the Genomics Core Facility of the Icahn Institute for
693 Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai**.** Reads
694 were aligned to the NCBI37/hg19 reference genome and variants were called using the
695 sequence analysis pipeline by Linderman *et al*[105] Variant calls and coverage at every
696 site at the genomic interval spanned by the candidate IBD haplotype (chr9:112000000-
697 118000000bp (NCBI37/hg19)) were obtained using the *"-out_mode EMIT_ALL_SITES*"

698 flag in GATKv3.2-72 (RRID:SCR_001876)**.** For summary statistics of whole genome
699 sequencing (WGS) see **Supplement file 1**.

700 *In Silico analysis and validation of COL27A1.*pG967R

701 WGS variant calls were annotated with allele frequency information and *in silico*
702 prediction scores for SIFT, PhyloP, GERP generated using snpEffv3.0
703 (RRID:SCR_005191) as part of the sequence analysis pipeline published by Linderman
704 *et al*[105]. We identified all genomic variants that were present in at least 6 copies across
705 the four IBD-homozygotes and that lay within the shared boundary of the IBD haplotype.
706 Using this criteria, only one rare, coding variant was found to be shared between all four
707 homozygotes, namely a point mutation in the gene *COL27A1* (g.9:*116958257.C>G,*
708 NM_032888.1, p.G697R*,* rs140950220*)* which was present in 7 copies (with 3
709 individuals being homozygous, and the fourth being heterozygous). The rs140950220
710 G/C allele status was validated by Sanger sequencing of exon 7 in the *COL27A1* gene
711 in all four individuals. We also validated *COL27A1.*pG697R status in individuals carrying
712 the significant IBD-clique at 9q32 using the Fluidigm SNPType assay adhering to the
713 standard protocol. All individuals carrying at least one copy of the top IBD-haplotype
714 (N=59) were genotyped for the rs140950220 variant in addition to a panel of age and
715 sex matched Puerto Rican ancestry controls (N=59).

716 *Genotyping COL27A1.pG697R in a Multi-Ethnic Population of PAGE*

717 We estimated the frequency of the *COL27A1.pG697R (*dbSNP=rs140950220) variant in
718 the Population Architecture using Genomics and Epidemiology (PAGE) study. The

719 PAGE study comprises a diverse global reference panel from five studies. African-
720 American and Hispanic/Latino women from the Women's Health Initiative (WHI), a
721 multi-center cohort study investigating post-menopausal women's health in the US and
722 recruited women at 40 centers across the US. Self-identified Hispanic/Latinos from four
723 sites in San Diego, CA, Chicago, IL, Bronx, NY, and Miami, FL as part of the Hispanic
724 Community Health Study / Study of Latinos (HCHS/SOL). African American, Japanese
725 American, and Native Hawaiian participants from the Multiethnic Cohort (MEC)
726 prospective cohort study recruiting men and women from Hawaii and California. The
727 Global Reference Panel (GRP) created by Stanford University contributed samples
728 including; a population sample of Andean individuals primarily of Quechuan/Aymaran
729 ancestry from Puno, Peru; a population sample of Easter Island (Rapa Nui), Chile;
730 individuals of indigenous origin from Oaxaca, Mexico, Honduras, Colombia, the Nama
731 and Khomani KhoeSan populations of the Northern Cape, South Africa; the Human
732 Genome Diversity Panel in collaboration with the Centre Etude Polymorphism Humain
733 (CEPH) in Paris; and the Maasai in Kinyawa, Kenya (MKK) dataset from the
734 International Hapmap Project hosted at Coriell. Finally, the Bio*Me* biobank in the Mount
735 Sinai health system, New York City, contributed African-American, Hispanic/Latino, and
736 participants who reported as mixed or other ancestry to the PAGE study, ~50% of whom
737 were born outside New York City and for whom country-of-birth information was
738 available. In all, participants in the PAGE Study represent a global reference panel of
739 112 populations ranging from 4-17773 individuals in size (**Supplement file 4**). Samples
740 in the PAGE study were genotyped on the Illumina Multi-Ethnic Genotyping Array
741 (MEGA), which included direct genotyping of the rs140950220 variant. A total of 53338

742 PAGE and GRP samples were genotyped on the MEGA array at the Johns Hopkins
743 Center for Inherited Disease Research (CIDR), with 52878 samples successfully
744 passing CIDR's QC process. Genotyping data that passed initial quality control at CIDR
745 were released to the Quality Assurance / Quality Control (QA/QC) analysis team at the
746 University of Washington Genetics Coordinating Center (UWGCC). The UWGCC
747 further cleaned the data according to previously described methods[106] and returned
748 genotypes for 51520 subjects. A total of 1705969 SNPs were genotyped on the
749 MEGA. The *COL27A1*.pG697R variant passed the following filters; (1) CIDR technical
750 filters, (2) SNPs with missing call rate >= 2%, (3) SNPs with more than 6 discordant
751 calls in 988 study duplicates, (4) SNPs with greater than 1 Mendelian errors in 282 trios
752 and 1439 duos, (5) SNPs with a Hardy-Weinberg $p<10^{-10}$, (6) positional duplicates.

753 *Structural modeling of the COL27A1.PG697R missense variant*

754 We downloaded X-ray crystal coordinates (1CAG from Bella *et al*[64].; www.pdb.org) on
755 January 21, 2017. Visualization and modeling of the missense variant were performed
756 in PyMol (www.pymol.org; RRID:SCR_000305).

757

758 *Phenome-Wide Association Study*

759 To test for clinical symptoms of Steel syndrome in *COL27A1*.pG967R carriers, we
760 performed a Phenome-Wide Association Study (PheWas) with EHR-derived ICD9 billing
761 codes as the phenotypic outcome. In the association model, for each individual ICD9
762 codes were encoded as "1" if the ICD9 was present in their EHR, and "0" if the ICD9
763 code was absent. Carrier status for *COL27A1*.pG697R was used as the primary

764 predictor variable, with heterozygous individuals encoded as “1”, non-carriers encoded
765 a “0” and homozygotes excluded from the analysis. We restricted the analysis to
766 carriers of *COL27A1*.pG697R (n=106) and non-carriers (n=4480) who either reported
767 being born in Puerto Rico or who were US-born, self-identified as H/L and overlapped
768 with Puerto Rican born individuals in principal component analysis. Age, sex and the
769 first 5 principal components were included as covariates in our model. The regression
770 was performed using four methods; Generalized Linear Models (GLM) using the glm()
771 function in Rv3.2.1; a score test based on the saddlepoint approximation (SPATest)
772 using the SPAtest() function in Rv3.2.1; a score test using a base adjustment for rare
773 variants (Firth test) using the logistf() function in Rv3.2.1; and a linear mixed model
774 using the GCTAv1.24.2 software with a genetic relationship matrix constructed from
775 281666 SNPs shared between the OMNI and MEGA arrays (MAF >= 1%). To adjust for
776 multiple tests, raw p-values were adjusted for false discovery rate using the p.adjust()
777 function in R, and only those below an FDR adjusted p-value of 0.05 were reported as
778 significant.

779 *Clinical review of patient records*

780 Information from inpatient, outpatient, emergency and private practice settings housed
781 in the Mount Sinai health system since 2004 was reviewed by two clinical experts
782 independently. This data includes laboratory reports, radiological data, pathology
783 results, operative and inpatient/outpatient progress notes, discharge summaries,
784 pharmacy, and nurses reports. The clinical experts examined for clinical features similar
785 to those reported for Steel syndrome cases in Flynn et al 2010[59], including

786 developmental dysplasia of the hip (or congenital hip dysplasia), carpal coalition,
787 scoliosis, and joint and spine anomalies. Both clinical experts reviews patient records
788 independently and compared notes to resolve discrepancies. They reviewed the records
789 of the 34 youngest *COL27A1*.pG697R carriers (mean age 42), and compared their
790 findings to 31 randomly selected age and sex matched Puerto Rican non-carriers, and
791 also to published reports of population prevalences of key clinical features for similar
792 age groups where available.

793 *Local Ancestry Estimation*

794 Due to the process of recombination, individuals from populations that have undergone
795 recent admixture can exhibit a mosaic of genetic ancestry along their genome. Their
796 genetic ancestry at a given genomic segment (referred to as local ancestry), can be
797 inferred from genotype data with the use of non-admixed reference panels of known
798 continental ancestry. We calculated local ancestry in the three homozygous Puerto
799 Rican individuals genotyped on OMNI by first extracting the intersecting sites of the
800 Affymetrix 6.0 array (n=593729 SNPs in total) and merging them with 3 ancestral
801 reference panels. These reference panels consisted of the CEU and YRI samples from
802 the 1000 Genomes Project in addition to the Native American reference panel described
803 previously that were used as a proxy for European, African and Native American
804 ancestral source populations, respectively. RFMix[84] was used to infer local ancestry.

805 *Calculation of Historical Effective Population Size in Puerto Ricans*

806 To investigate evidence of a founder effect in Puerto Ricans we ran the IBDNe
807 software[85] in 2816 Puerto Ricans from the discovery effort using the cleaned set of
808 pairwise IBD-haplotypes inferred using GERMLINE. IBDNe was run using the default
809 parameters, including an assumed generation time of 25 years.

810 **Acknowledgements**

811 We would like to thank Noah Zaitlen, Alexander Gusev, George Diaz, Rounak Dey, and
812 Alicia Martin for their helpful suggestions in preparing this manuscript. This work was
813 supported by funds from several grants; The Population Architecture Using Genomics
814 and Epidemiology (PAGE) program is funded by the National Human Genome
815 Research Institute, with co-funding from the National Institute on Minority Health and
816 Health Disparities, supported by U01HG007416, U01HG007417, U01HG007397,
817 U01HG007376, and U01HG007419. The PAGE consortium thanks the staff and
818 participants of all PAGE studies for their important contributions. The complete list of
819 PAGE members can be found at http://www.pagestudy.org. Genotyping services were
820 provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded
821 through a federal contract from the National Institutes of Health to The Johns Hopkins
822 University, contract number HHSN268201200008I. Genotype data quality control and
823 quality assurance services were provided by the Genetic Analysis Center in the
824 Biostatistics Department of the University of Washington, through support provided by
825 the CIDR contract. High performance computing was supported in part through the
826 computational resources and staff expertise provided by Scientific Computing at the
827 Icahn School of Medicine at Mount Sinai and through the Office of Research

Infrastructure under award number S10OD018522. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Hamosh, A., Scott, A.F., Amberger, J., Valle, D., and McKusick, V.A. (2000). Online Mendelian Inheritance in Man (OMIM). Hum. Mutat. *15*, 57–61.

2. McCarthy, J.J., McLeod, H.L., and Ginsburg, G.S. (2013). Genomic medicine: a decade of successes, challenges, and opportunities. Sci. Transl. Med. *5*, 189sr4.

3. 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 - Supplementary information. Nat. Rev. Genet. *12*, 745–755.

4. Manolio, T. a, Chisholm, R.L., Ozenberger, B., Roden, D.M., Williams, M.S., Wilson, R., Bick, D., Bottinger, E.P., Brilliant, M.H., Eng, C., et al. (2013). Implementing genomic medicine in the clinic: the future is here. Genet. Med. *15*, 258–267.

5. Manolio, T.A., Abramowicz, M., Al-Mulla, F., Anderson, W., Balling, R., Berger, A.C., Bleyl, S., Chakravarti, A., Chisholm, R.L., Dissanayake, V.H.W., et al. (2015). Global implementation of genomic medicine : We are not alone. Sci. Transl. Med. *7*, 1–8.

6. Philippakis, A.A., Azzariti, D.R., Beltran, S., Brookes, A.J., Brownstein, C.A., Brudno, M., Brunner, H.G., Buske, O.J., Carey, K., Doll, C., et al. (2015). The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. Hum. Mutat. *36*, 915–921.

7. Chong, J., Yu, J.-H., Lorentzen, P., Park, K., Jamal, S.M., Tabor, H.K., Rauch, A., Saenz, M.S., Boltshauser, E., Patterson, K.E., et al. (2015). Gene discovery for Mendelian conditions via social networking: de novo variants in KDM1A cause developmental delay and distinctive facial features. bioRxiv 028241.

8. Collins, F.S., and Varmus, H. (2015). A New Initiative on Precision Medicine. N. Engl. J. Med. *372*, 793–795.

9. Ashley, E.A. (2015). The Precision Medicine Initiative. JAMA *313*, 2119.

10. Collins, R. (2012). What makes UK Biobank special? Lancet *379*, 1173–1174.

11. Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N., O'Dushlaine, C., Van Hout, C. V, Staples, J., Gonzaga-Jauregui, C., et al. (2016).

858 Distribution and clinical impact of functional variants in 50,726 whole-exome sequences
859 from the DiscovEHR study. Science *354*, aaf6814.

860 12. Feldman, G.L. (2016). 2016 ACMG Annual Meeting presidential address: the
861 practice of medical genetics: myths and realities. Genet. Med. *18*, 957–959.

862 13. Ashley, E.A. (2016). Towards precision medicine. Nat. Rev. Genet. *17*, 507–522.

863 14. Abul-Husn, N.S., Manickam, K., Jones, L.K., Wright, E.A., Hartzel, D.N., Gonzaga-
864 Jauregui, C., O'Dushlaine, C., Leader, J.B., Lester Kirchner, H., Lindbuchler, D.M., et al.
865 (2016). Genetic identification of familial hypercholesterolemia within a single U.S. health
866 care system. Science *354*, aaf7000.

867 15. Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A.,
868 Wang, M., Buhay, C., et al. (2014). Molecular Findings Among Patients Referred for
869 Clinical Whole-Exome Sequencing. JAMA *312*, 1870.

870 16. Taylor, J.C., Martin, H.C., Lise, S., Broxholme, J., Cazier, J.B., Rimmer, A.,
871 Kanapin, A., Lunter, G., Fiddy, S., Allan, C., et al. (2015). Factors influencing success of
872 clinical genome sequencing across a broad spectrum of disorders. Nat Genet *47*, 717–
873 726.

874 17. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith,
875 J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The
876 Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities.
877 Am. J. Hum. Genet. *97*, 199–215.

878 18. Amendola, L.M., Jarvik, G.P., Leo, M.C., McLaughlin, H.M., Akkari, Y., Amaral,
879 M.D., Berg, J.S., Biswas, S., Bowling, K.M., Conlin, L.K., et al. (2016). Performance of
880 ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical
881 Sequencing Exploratory Research Consortium. Am. J. Hum. Genet. *99*, 247.

882 19. Katsanis, N. (2016). The continuum of causality in human genetic disorders.
883 Genome Biol. *17*, 233.

884 20. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P.,
885 Margulies, D.M., Loscalzo, J., and Kohane, I.S. (2016). Genetic Misdiagnoses and the
886 Potential for Health Disparities. N. Engl. J. Med. *375*, 655–665.

887 21. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature
888 *538*, 161–164.

889 22. Petrovski, S., Goldstein, D.B., Need, A., Goldstein, D., Bustamante, C., Burchard,
890 E., Vega, F., Price, A., Patterson, N., Plenge, R., et al. (2016). Unequal representation
891 of genetic variation across ancestry groups creates healthcare inequality in the
892 application of precision medicine. Genome Biol. *17*, 157.

23. Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W.A., Li, R., Manolio, T. a, Sanderson, S.C., Kannry, J., Zinberg, R., Basford, M. a, et al. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet. Med. *15*, 761–771.

24. Green, R.C., Goddard, K.A.B., Jarvik, G.P., Amendola, L.M., Appelbaum, P.S., Berg, J.S., Bernhardt, B.A., Biesecker, L.G., Biswas, S., Blout, C.L., et al. (2016). Clinical Sequencing Exploratory Research Consortium: Accelerating Evidence-Based Practice of Genomic Medicine. Am. J. Hum. Genet. *98*, 1051–1066.

25. Blair, D.R., Lyttle, C.S., Mortensen, J.M., Bearden, C.F., Jensen, A.B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E.V., Brunak, S., et al. (2013). A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. Cell *155*, 70–80.

26. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen--the Clinical Genome Resource. N. Engl. J. Med. *372*, 2235–2242.

27. Kirkpatrick, B.E., Riggs, E.R., Azzariti, D.R., Miller, V.R., Ledbetter, D.H., Miller, D.T., Rehm, H., Martin, C.L., and Faucett, W.A. (2015). GenomeConnect: Matchmaking Between Patients, Clinical Laboratories, and Researchers to Improve Genomic Knowledge. Hum. Mutat. *36*, 974–978.

28. Gahl, W.A., Mulvihill, J.J., Toro, C., Markello, T.C., Wise, A.L., Ramoni, R.B., Adams, D.R., and Tifft, C.J. (2016). The NIH Undiagnosed Diseases Program and Network: Applications to modern medicine. Mol. Genet. Metab. *117*, 393–400.

29. Browning, S.R., and Browning, B.L. (2011). Identity by Descent Between Distant Relatives: Detection and Applications. Annu. Rev. Genet. *46*, 120920150949000.

30. Houwen, R.H., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L.A., and Freimer, N.B. (1994). Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nat. Genet. *8*, 380–386.

31. Kenny, E.E., Gusev, A., Riegel, K., Lütjohann, D., Lowe, J.K., Salit, J., Maller, J.B., Stoffel, M., Daly, M.J., Altshuler, D.M., et al. (2009). Systematic haplotype analysis resolves a complex plasma plant sterol locus on the Micronesian Island of Kosrae. Proc. Natl. Acad. Sci. U. S. A. *106*, 13886–13891.

32. Henden, L., Freytag, S., Afawi, Z., Baldassari, S., Berkovic, S.F., Bisulli, F., Canafoglia, L., Casari, G., Crompton, D.E., Depienne, C., et al. (2016). Identity by descent fine mapping of familial adult myoclonus epilepsy (FAME) to 2p11.2???2q11.2. Hum. Genet. 1–9.

928 33. Qi, L., Cornelis, M.C., Kraft, P., Stanya, K.J., Kao, W.H.L., Pankow, J.S., Dupuis, J.,
929 Florez, J.C., Fox, C.S., Paré, G., et al. (2010). Genetic variants at 2q24 are associated
930 with susceptibility to type 2 diabetes. Hum. Mol. Genet. *19*, 2706–2715.

931 34. Traherne, J.A., Jiang, W., Valdes, A.M., Hollenbach, J.A., Jayaraman, J., Lane, J.A.,
932 Johnson, C., Trowsdale, J., and Noble, J.A. (2016). KIR haplotypes are associated with
933 late-onset type 1 diabetes in European-American families. Genes Immun. *17*, 8–12.

934 35. Shaw, M., Yap, T.Y., Henden, L., Bahlo, M., Gardner, A., Kalscheuer, V.M., Haan,
935 E., Christie, L., Hackett, A., and Gecz, J. (2015). Identical by descent L1CAM mutation
936 in two apparently unrelated families with intellectual disability without L1 syndrome. Eur.
937 J. Med. Genet. *58*, 364–368.

938 36. Ko, J.M., Zhang, P., Law, S., Fan, Y., Song, Y.Q., Zhao, X.K., Wong, E.H.W., Tang,
939 S., Song, X., Lung, M.L., et al. (2014). Identity-by-descent approaches identify regions
940 of importance for genetic susceptibility to hereditary esophageal squamous cell
941 carcinoma. Oncol. Rep. *32*, 860–870.

942 37. Lalli, M.A., Cox, H.C., Arcila, M.L., Cadavid, L., Moreno, S., Garcia, G., Madrigal, L.,
943 Reiman, E.M., Arcos-Burgos, M., Bedoya, G., et al. (2014). Origin of the PSEN1 E280A
944 mutation causing early-onset Alzheimer's disease. Alzheimer's Dement. *10*, S277–
945 S283.

946 38. Visscher, P.M., McEvoy, B., and Yang, J. (2010). From Galton to GWAS:
947 quantitative genetics of human height. Genet. Res. (Camb). *92*, 371–379.

948 39. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F.,
949 Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of
950 variants clustered in genomic loci and biological pathways affect human height. Nature
951 *467*, 832–838.

952 40. Durand, C., and Rappold, G.A. (2013). Height matters-from monogenic disorders to
953 normal variation. Nat. Rev. Endocrinol. *9*, 171–177.

954 41. Moltke, I., and Albrechtsen, A. (2014). RelateAdmix: A software tool for estimating
955 relatedness between admixed individuals. Bioinformatics *30*, 1027–1028.

956 42. The 1000 Genomes Project Consortium (2015). A global reference for human
957 genetic variation. Nature *526*, 68–74.

958 43. Mao, X., Bigham, A.W., Mei, R., Gutierrez, G., Weiss, K.M., Brutsaert, T.D., Leon-
959 Velarde, F., Moore, L.G., Vargas, E., McKeigue, P.M., et al. (2007). A genomewide
960 admixture mapping panel for Hispanic/Latino populations. Am. J. Hum. Genet. *80*,
961 1171–1178.

962 44. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and
963 Reich, D. (2006). Principal components analysis corrects for stratification in genome-
964 wide association studies. Nat. Genet. *38*, 904–909.

965 45. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman,
966 J.M., and Pe'Er, I. (2009). Whole population, genome-wide mapping of hidden
967 relatedness. Genome Res. *19*, 318–326.

968 46. Gusev, A., Palamara, P.F., Aponte, G., Zhuang, Z., Darvasi, A., Gregersen, P., and
969 Pe'er, I. (2012). The architecture of long-range haplotypes shared within and across
970 populations. Mol. Biol. Evol. *29*, 473–486.

971 47. Chiang, C.W.K., Ralph, P., and Novembre, J. (2016). Conflation of Short Identity-by-
972 Descent Segments Bias Their Inferred Length Distribution. G3 (Bethesda). *6*, 1287–
973 1296.

974 48. Browning, S.R., and Thompson, E.A. (2012). Detecting rare variant associations by
975 identity-by-descent mapping in case-control studies. Genetics *190*, 1521–1531.

976 49. Need, A.C., Kasperaviciute, D., Cirulli, E.T., and Goldstein, D.B. (2009). A genome-
977 wide genetic signature of Jewish ancestry perfectly separates individuals with and
978 without full Jewish ancestry in a large random sample of European Americans. Genome
979 Biol. *10*, R7.

980 50. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux,
981 C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al. (2013).
982 Reconstructing the Population Genetic History of the Caribbean. PLoS Genet. *9*,.

983 51. Durand, C., and Rappold, G. a (2013). Height matters-from monogenic disorders to
984 normal variation. Nat. Rev. Endocrinol. *9*, 171–177.

985 52. eLife (2016). A century of trends in adult human height. Elife *5*, 1–29.

986 53. Cohen, P., Rogol, A.D., Deal, C.L., Saenger, P., Reiter, E.O., Ross, J.L.,
987 Chernausek, S.D., Savage, M.O., Wit, J.M., and 2007 ISS Consensus Workshop
988 participants (2008). Consensus statement on the diagnosis and treatment of children
989 with idiopathic short stature: a summary of the Growth Hormone Research Society, the
990 Lawson Wilkins Pediatric Endocrine Society, and the European Society for Paediatric
991 Endocrinology Workshop. J. Clin. Endocrinol. Metab. *93*, 4210–4217.

992 54. Vacic, V., Ozelius, L.J., Clark, L.N., Bar-Shira, A., Gana-Weisz, M., Gurevich, T.,
993 Gusev, A., Kedmi, M., Kenny, E.E., Liu, X., et al. (2014). Genome-wide mapping of IBD
994 segments in an Ashkenazi PD cohort identifies associated haplotypes. Hum. Mol.
995 Genet. *23*, 4693–4702.

996 55. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D.,
997 Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A tool set for
998 whole-genome association and population-based linkage analyses. Am. J. Hum. Genet.
999 *81*, 559–575.

1000 56. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J.
1001 (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets.
1002 Gigascience *4*, 7.

1003 57. Gonzaga-Jauregui, C., Gamble, C.N., Yuan, B., Penney, S., Jhangiani, S., Muzny,
1004 D.M., Gibbs, R.A., Lupski, J.R., and Hecht, J.T. (2015). Mutations in COL27A1 cause
1005 Steel syndrome and suggest a founder mutation effect in the Puerto Rican population.
1006 Eur. J. Hum. Genet. *23*, 342–346.

1007 58. Steel, H.H., Piston, R.W., Clancy, M., and Betz, R.R. (1993). A syndrome of
1008 dislocated hips and radial heads, carpal coalition, and short stature in Puerto Rican
1009 children. J. Bone Jt. Surg. - Ser. A.75 ()(pp 259-264), 1993.Date Publ. 1993. *75*, 259–
1010 264.

1011 59. Flynn, J.M., Ramirez, N., Betz, R., Mulcahey, M.J., Pino, F., Herrera-Soto, J. a,
1012 Carlo, S., and Cornier, A.S. (2010). Steel syndrome: dislocated hips and radial heads,
1013 carpal coalition, scoliosis, short stature, and characteristic facial features. J. Pediatr.
1014 Orthop. *30*, 282–288.

1015 60. Pace, J.M., Corrado, M., Missero, C., and Byers, P.H. (2003). Identification,
1016 characterization and expression analysis of a new fibrillar collagen gene, COL27A1.
1017 Matrix Biol. *22*, 3–14.

1018 61. Plumb, D.A., Ferrara, L., Torbica, T., Knowles, L., Mironov, A., Kadler, K.E., Briggs,
1019 M.D., and Boot-Handford, R.P. (2011). Collagen XXVII Organises the Pericellular Matrix
1020 in the Growth Plate. PLoS One *6*, e29422.

1021 62. Christiansen, H.E., Lang, M.R., Pace, J.M., and Parichy, D.M. (2009). Critical Early
1022 Roles for col27a1a and col27a1b in Zebrafish Notochord Morphogenesis, Vertebral
1023 Mineralization and Post-embryonic Axial Growth. PLoS One *4*, e8481.

1024 63. Persikov, A. V., Pillitteri, R.J., Amin, P., Schwarze, U., Byers, P.H., and Brodsky, B.
1025 (2004). Stability related bias in residues replacing glycines within the collagen triple
1026 helix (Gly-Xaa-Yaa) in inherited connective tissue disorders. Hum. Mutat. *24*, 330–337.

1027 64. Bella, J., Eaton, M., Brodsky, B., and Berman, H.M. (1994). Crystal and molecular
1028 structure of a collagen-like peptide at 1.9 A resolution. Science *266*, 75–81.

1029 65. McGrory, J., Costa, T., and Cole, W.G. (1996). A novel G499D substitution in the
1030 alpha 1(III) chain of type III collagen produces variable forms of Ehlers-Danlos
1031 syndrome type IV. Hum. Mutat. *7*, 59–60.

66. Tromp, G., De Paepe, A., Nuytinck, L., Madhatheri, S., and Kuivaniemi, H. (1995). Substitution of valine for glycine 793 in type III procollagen in Ehlers-Danlos syndrome type IV. Hum. Mutat. *5*, 179–181.

67. Anderson, D.W., Thakker-Varia, S., Tromp, G., Kuivaniemi, H., and Stolle, C.A. (1997). A glycine (415)-to-serine substitution results in impaired secretion and decreased thermal stability of type III procollagen in a patient with Ehlers-Danlos syndrome type IV. Hum. Mutat. *9*, 62–63.

68. Knebelmann, B., Deschenes, G., Gros, F., Hors, M.C., Grünfeld, J.P., Zhou, J., Tryggvason, K., Gubler, M.C., and Antignac, C. (1992). Substitution of arginine for glycine 325 in the collagen alpha 5 (IV) chain associated with X-linked Alport syndrome: characterization of the mutation by direct sequencing of PCR-amplified lymphoblast cDNA fragments. Am. J. Hum. Genet. *51*, 135–142.

69. Zhou, J., Hertz, J.M., and Tryggvason, K. (1992). Mutation in the alpha 5(IV) collagen chain in juvenile-onset Alport syndrome without hearing loss or ocular lesions: detection by denaturing gradient gel electrophoresis of a PCR product. Am. J. Hum. Genet. *50*, 1291–1300.

70. Starman, B.J., Eyre, D., Charbonneau, H., Harrylock, M., Weis, M.A., Weiss, L., Graham, J.M., and Byers, P.H. (1989). Osteogenesis imperfecta. The position of substitution for glycine by cysteine in the triple helical domain of the pro alpha 1(I) chains of type I collagen determines the clinical phenotype. J. Clin. Invest. *84*, 1206–1214.

71. Shapiro, J.R., Stover, M.L., Burn, V.E., McKinstry, M.B., Burshell, A.L., Chipman, S.D., and Rowe, D.W. (1992). An osteopenic nonfracture syndrome with features of mild osteogenesis imperfecta associated with the substitution of a cysteine for glycine at triple helix position 43 in the pro alpha 1(I) chain of type I collagen. J. Clin. Invest. *89*, 567–573.

72. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205–1210.

73. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. *31*, 1102–1110.

74. Dey, R., Schmidt, E.M., Abecasis, G.R., and Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. bioRxiv.

1068 75. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for
1069 genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.

1070 76. Wang, X. (2014). Firth logistic regression for rare variant association tests. Front.
1071 Genet. *5*, 187.

1072 77. Kebaish, K.M., Neubauer, P.R., Voros, G.D., Khoshnevisan, M.A., and Skolasky,
1073 R.L. (2011). Scoliosis in adults aged forty years and older: prevalence and relationship
1074 to age, race, and gender. Spine (Phila. Pa. 1976). *36*, 731–736.

1075 78. Reginster, J.Y. (2002). The prevalence and burden of arthritis. Rheumatology
1076 (Oxford). *41 Supp 1*, 3–6.

1077 79. Kolonel, L.N., Henderson, B.E., Hankin, J.H., Nomura, A.M., Wilkens, L.R., Pike,
1078 M.C., Stram, D.O., Monroe, K.R., Earle, M.E., and Nagamine, F.S. (2000). A multiethnic
1079 cohort in Hawaii and Los Angeles: baseline characteristics. Am. J. Epidemiol. *151*, 346–
1080 357.

1081 80. LaVange, L.M., Kalsbeek, W.D., Sorlie, P.D., Avilés-Santa, L.M., Kaplan, R.C.,
1082 Barnhart, J., Liu, K., Giachello, A., Lee, D.J., Ryan, J., et al. (2010). Sample Design and
1083 Cohort Selection in the Hispanic Community Health Study/Study of Latinos. Ann.
1084 Epidemiol. *20*, 642–649.

1085 81. (1998). Design of the Women's Health Initiative clinical trial and observational study.
1086 The Women's Health Initiative Study Group. Control. Clin. Trials *19*, 61–109.

1087 82. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer,
1088 J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human
1089 genome diversity cell line panel. Science *296*, 261–262.

1090 83. Lek, M., Karczewski, K.J., Samocha, K.E., Banks, E., Fennell, T., O, A.H., Ware,
1091 J.S., Hill, A.J., Cummings, B.B., Birnbaum, D.P., et al. (2016). Analysis of protein-coding
1092 genetic variation in 60,706 humans. bioRxiv *536*, 030338.

1093 84. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: A
1094 discriminative modeling approach for rapid and robust local-ancestry inference. Am. J.
1095 Hum. Genet. *93*, 278–288.

1096 85. Browning, S.R., and Browning, B.L. (2015). Accurate Non-parametric Estimation of
1097 Recent Effective Population Size from Segments of Identity by Descent. Am. J. Hum.
1098 Genet. *97*, 404–418.

1099 86. Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J.K., Muzzio, M., Rodriguez-
1100 Flores, J.L., Kenny, E.E., Gignoux, C.R., Maples, B.K., Guiblet, W., et al. (2013).
1101 Reconstructing Native American migrations from whole-genome and whole-exome data.
1102 PLoS Genet. *9*, e1004023.

1103 87. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent
1104 positive selection in the human genome. PLoS Biol. *4*, e72.

1105 88. Szpiech, Z.A., and Hernandez, R.D. (2014). Selscan: An efficient multithreaded
1106 program to perform EHH-based scans for positive selection. Mol. Biol. Evol. *31*, 2824–
1107 2827.

1108 89. Anikster, Y., Huizing, M., White, J., Shevchenko, Y.O., Fitzpatrick, D.L., Touchman,
1109 J.W., Compton, J.G., Bale, S.J., Swank, R.T., Gahl, W. a, et al. (2001). Mutation of a
1110 new gene causes a unique form of Hermansky-Pudlak syndrome in a genetic isolate of
1111 central Puerto Rico. Nat. Genet. *28*, 376–380.

1112 90. Cornier, A.S., Staehling-Hampton, K., Delventhal, K.M., Saga, Y., Caubet, J.-F.,
1113 Sasaki, N., Ellard, S., Young, E., Ramirez, N., Carlo, S.E., et al. (2008). Mutations in the
1114 MESP2 Gene Cause Spondylothoracic Dysostosis/Jarcho-Levin Syndrome. Am. J.
1115 Hum. Genet. *82*, 1334–1341.

1116 91. Daniels, M.L.A., Leigh, M.W., Davis, S.D., Armstrong, M.C., Carson, J.L., Hazucha,
1117 M., Dell, S.D., Eriksson, M., Collins, F.S., Knowles, M.R., et al. (2013). Founder
1118 mutation in RSPH4A identified in patients of Hispanic descent with primary ciliary
1119 dyskinesia. Hum. Mutat. *34*, 1352–1356.

1120 92. Al-Zaidy, S.A., Malik, V., Kneile, K., Rosales, X.Q., Gomez, A.M., Lewis, S.,
1121 Hashimoto, S., Gastier-Foster, J., Kang, P., Darras, B., et al. (2015). A slowly
1122 progressive form of limb-girdle muscular dystrophy type 2C associated with founder
1123 mutation in the SGCG gene in Puerto Rican Hispanics. Mol. Genet. Genomic Med. *3*,
1124 92–98.

1125 93. Arnold, S.E., Vega, I.E., Karlawish, J.H., Wolk, D.A., Nunez, J., Negron, M., Xie,
1126 S.X., Wang, L.S., Dubroff, J.G., McCarty-Wood, E., et al. (2013). Frequency and
1127 clinicopathological characteristics of presenilin 1 Gly206Ala mutation in puerto rican
1128 hispanics with dementia. J. Alzheimer's Dis. *33*, 1089–1095.

1129 94. Lee, J.H., Kahn, A., Cheng, R., Reitz, C., Vardarajan, B., Lantigua, R., Medrano, M.,
1130 Jiménez-Velázquez, I.Z., Williamson, J., Nagy, P., et al. (2014). Disease-related
1131 mutations among Caribbean Hispanics with familial dementia. Mol. Genet. Genomic
1132 Med. *2*, 430–437.

1133 95. Kotabagi, S., Shah, H., Shukla, A., and Girisha, K.M. (2017). Second family
1134 provides further evidence for causation of Steel syndrome by biallelic mutations in
1135 *COL27A1*. Clin. Genet.

1136 96. Gariballa, N., Ben-Mahmoud, A., Komara, M., Al-Shamsi, A.M., John, A., Ali, B.R.,
1137 and Al-Gazali, L. (2017). A novel aberrant splice site mutation in *COL27A1* is
1138 responsible for Steel syndrome and extension of the phenotype to include hearing loss.
1139 Am. J. Med. Genet. Part A *173*, 1257–1263.

97. Humes, K.R., Jones, N. a., and Ramirez, R.R. (2011). Overview of race and hispanic origin: 2010. Office 23.

98. Henn, B.M., Hon, L., Macpherson, J.M., Eriksson, N., Saxonov, S., Pe'er, I., and Mountain, J.L. (2012). Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One *7*,.

99. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

100. Roshyara, N.R., and Scholz, M. (2014). fcGENE: A Versatile Tool for Processing and Transforming SNP Datasets. PLoS One *9*, e97589.

101. Gusev, A., Kenny, E.E., Lowe, J.K., Salit, J., Saxena, R., Kathiresan, S., Altshuler, D.M., Friedman, J.M., Breslow, J.L., and Pe'er, I. (2011). DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. Am. J. Hum. Genet. *88*, 706–717.

102. Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

103. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. *10*, e1004234.

104. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

105. Linderman, M.D., Brandt, T., Edelmann, L., Jabado, O., Kasai, Y., Kornreich, R., Mahajan, M., Shah, H., Kasarskis, A., and Schadt, E.E. (2014). Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Med. Genomics *7*, 20.

106. Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. Genet. Epidemiol. *34*, 591–602.

Web Resources

BioMe OmniExpress data:

https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000888.v1.p1

BioMe MEGA data:

https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000925

PAGE MEGA data:

https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000356

Location of Native American panels:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130711_native_american_admix_train/

Software used in the analysis:

SMARTPCA: https://github.com/DReichLab/EIG

ADMIXTURE: https://www.genetics.ucla.edu/software/admixture/download.html

RelateAdmix: http://www.popgen.dk/software/index.php/RelateAdmix#Download

RFMix: https://sites.google.com/site/rfmixlocalancestryinference/

GERMLINE: http://www.cs.columbia.edu/~gusev/germline/

DASH: http://www1.cs.columbia.edu/~gusev/dash/

PyMol: www.pymol.org

IBDNe: http://faculty.washington.edu/browning/ibdne.html

EHR-Documented Evidence	Literature Homozygous children	Literature Homozygous Adults	BioMe Homozygous Adults	BioMe Heterozygous Adults <55yoa	BioMe Controls Adults <55yoa
N (N of females)	27 (9)	7(6)	5 (3)	34(20)	31(23)
Mean age (years)	12.8	35.4	51.6	41.8	40.6
EHR-Documented Medical History	**N(%)**	**N(%)**	**N(%)**	**N(%)**	**N(%)**
Height >= 2 s.d. from pop mean	27 (100)	7 (100)	5 (100)	0 (0)	0 (0)
Congenital hip dislocation			4	0	0
Leg length discrepancy			1	0	0
Total	27(100)	7(100)	5 (100)	0 (0)	0 (0)
Elbow contractures	24 (89)	7 (100)	1 (20)	0 (0)	0 (0)
Wrist deformity			1	2	1
Lunotriquetial fusion			1	0	0
Carpel tunnel			0	3	3
Total	24 (89)	6 (86)	2 (40)	5(15)	4(13)
Scoliosis	12 (44)	6 (86)	2 (40)	8 (24)	4 (13)
Pes cavus	12 (44)	0 (0)	0 (0)	0 (0)	0 (0)
Cervical stenosis			3	5	0
Cervical discitis			1	0	0
Cervical spondylosis			2	7	3
Cervical cord compression			3	3	1
Total	3 (9)	0 (0)	4 (80)	7 (21)	3 (10)
EHR-Documented Medical History			**N(%)**	**N(%)**	**N(%)**
Lumbar spine	-	-	1 (20)	10 (29)	5 (16)
Thoracic spine	-	-	1 (20)	5(15)	4(13)
Osteoporosis or osteopenia	-	-	3 (60)	3(9)	1 (3)
Arthritis or degenerative changes	-	-	3 (60)	13(38)	6 (19)
Hip replacement	-	-	3 (60)	0 (0)	0 (0)
Knee replacement	-	-	2 (40)	0 (0)	0 (0)
Cervical spine	-	-	3 (60)	2 (6)	1 (3)
Lumbar spine	-	-	1 (20)	1 (3)	0 (0)
Thoracic spine	-	-	2 (40)	1 (3)	1 (3)

Table 1 Clinical characteristics of five BioMe participants homozygote, thirty-four carriers and thirty-one non-carriers of the *COL27A1*.pG697R variant using evidence documented in Electronic Health Records (including billing and procedural codes, laboratory, radiologic, and progress notes) compared to features previously reported in Flynn et al, 2010, and Steel et al, 1993 for 27 children and 7 adults with Steel syndrome.

Disease Category	ICD9 Code	Short description	Number of diagnoses among carriers N (%)	Number of diagnoses among non-carriers N (%)	Odds Ratio (5% confidence intervals)	P-value
Neoplasms	622.10	Dysplasia of cervix, not otherwise specified	6 (5.7)	62 (1.4)	5.4 (2.3-12.6)	1.0 x 10-4
Musculoskeletal	716.98	Arthropathy unspecified, involving other unspecified sites	4 (3.8)	48 (1.1)	5.8 (2.2-15.3)	4.4 x 10-4
Musculoskeletal	721.00	Cervical spondylosis without myelopathy	5 (4.5)	74 (1.6)	5.4 (2.4-12.3)	6.6 x 10-5
Musculoskeletal	730.08	Acute osteomyelitis involving other specified sites	3 (2.8)	6 (0.1)	34.5 (7.4-162)	7.1 x 10-6
Digestive	789.10	Hepatomegaly	3 (2.8)	17 (0.4	11.6 (3.2-42.2)	2.1 x 10-4

Table 2 Top five significantly PheWAS associated ICD9 codes in *COL27A1*.pG697R carriers (N=106) compared to non-carriers (N=4480)

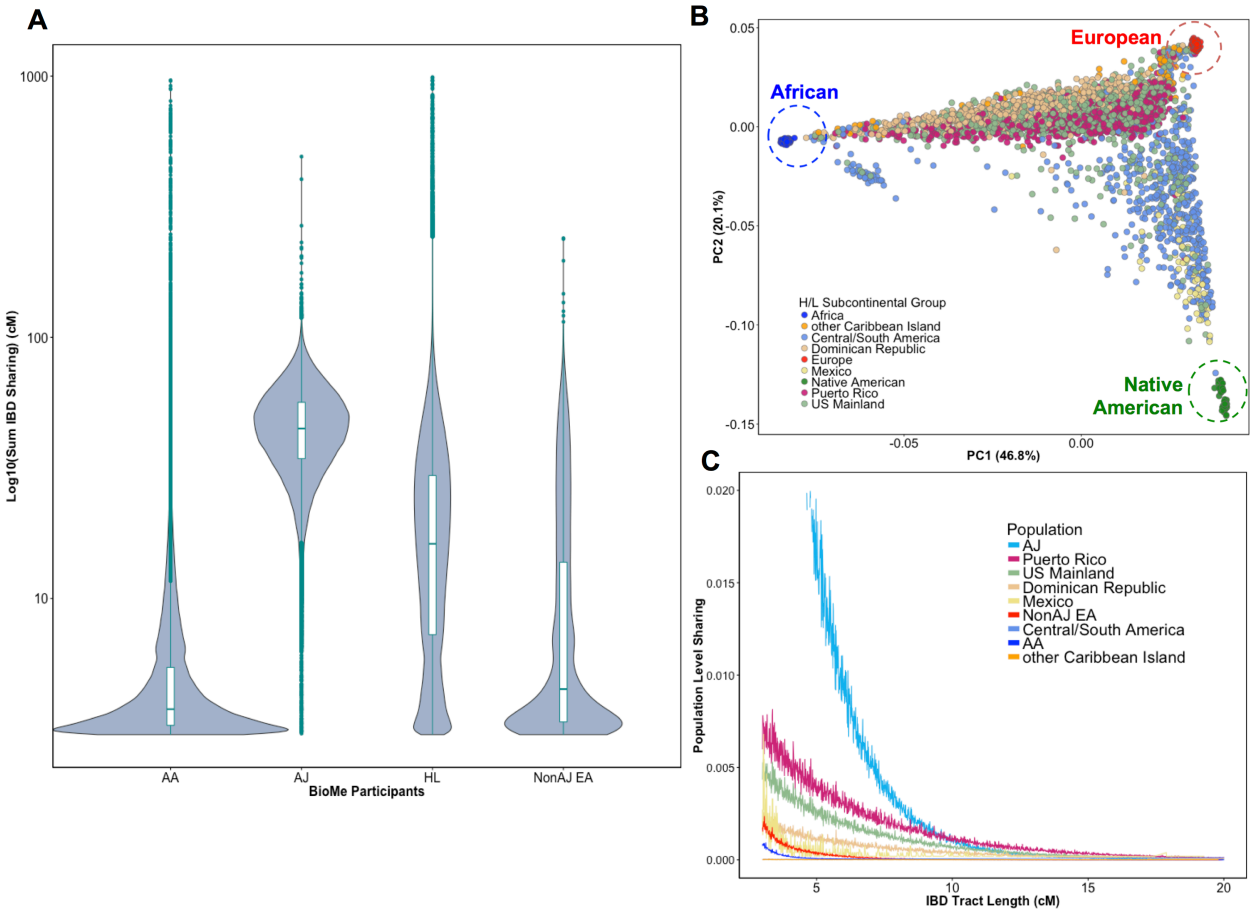

Figure 1 (A) Distribution of the pairwise sum of Identity-by-Descent (IBD) sharing (cM) between four broadly defined Bio*Me* populations, namely; African American (AA), Ashkenazi Jewish (AJ), Hispanic/Latino (H/L) and Non-Jewish European American (Non-AJ EA). (B) Sub-continental diversity in self-reported H/L participants in Bio*Me.* Afro-Caribbean participants fall between European (red) and African (blue) continental reference panels, Mexican and Central/South American H/L participants fall between European and Native American (green) reference panels, mainland US-born participants reside on either cline. (C) The tract length distribution of IBD sharing among Bio*Me* populations, normalized by population size. The y-axis represents the proportion population-level sharing (*x / (N*(N-1)/2)*), where *x* is the sum of the number of pairwise

shared IBD tracts and N is the number of individuals per population. The AJ population exhibits the highest level of population level sharing, followed by Puerto Rican born H/L.

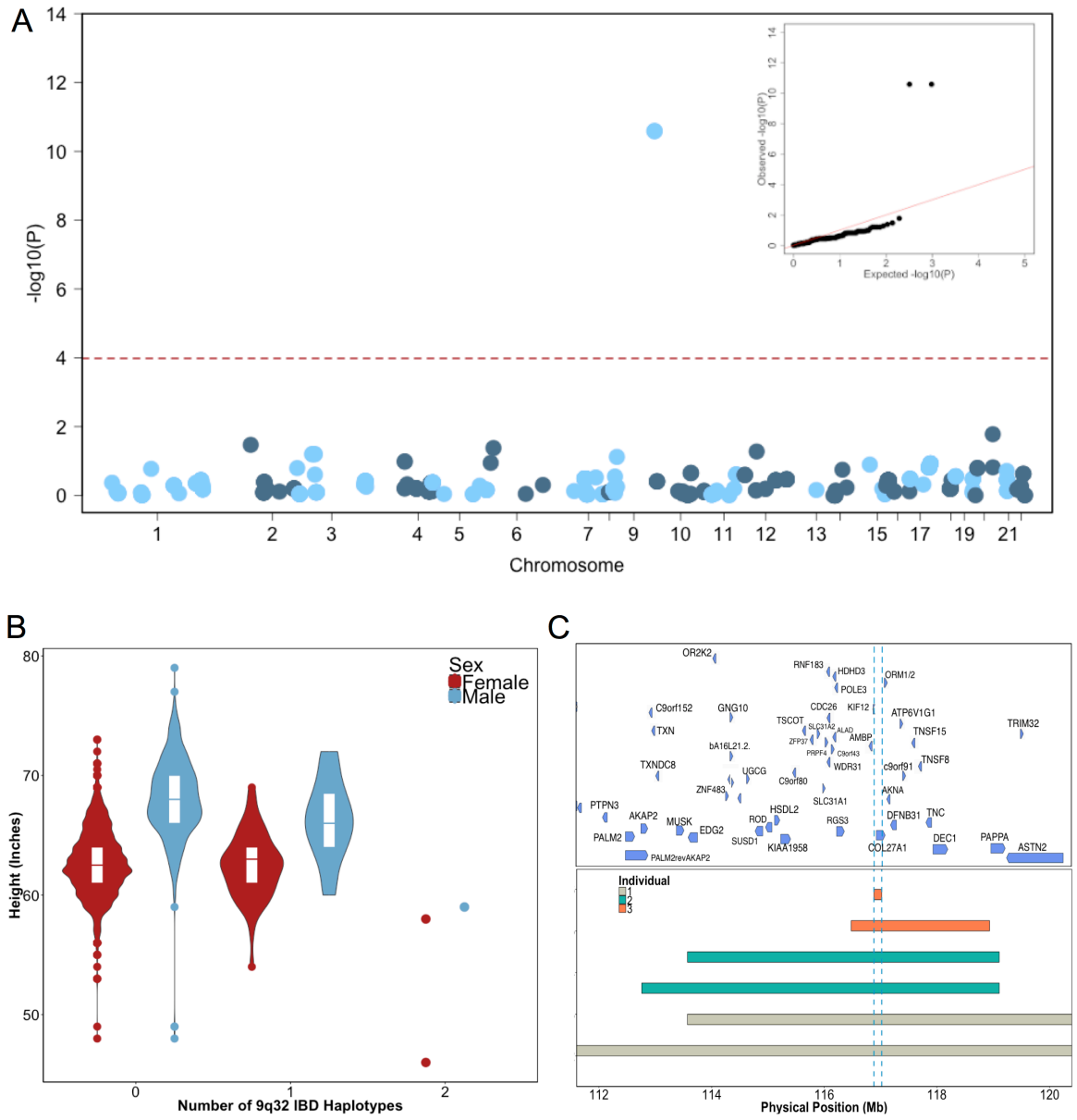

Figure 2 (A) Identity-By-Descent (IBD) mapping for height in Bio*Me* Puerto Ricans using a recessive model. Analysis was restricted to IBD-cliques where at least three individuals were homozygous. Only one IBD-clique achieved Bonferonni significance (at 9q32). (B) Distribution of height among Puerto Rican individuals who carry either 0,1 or 2 copies of the IBD-haplotype reveals a large recessive effect. Homozygous individuals (those carrying 2 copies of the IBD-haplotype) are on average 6-10" shorter than the population mean for Puerto Rican ancestry individuals. (C) The minimum shared

boundary of the significant IBD-haplotype between the three homozygous individuals (represented by the dashed blue line). The top panel depicts known genes at the 9q32 locus. The minimum shared boundary of the IBD overlaps the gene *COL2*

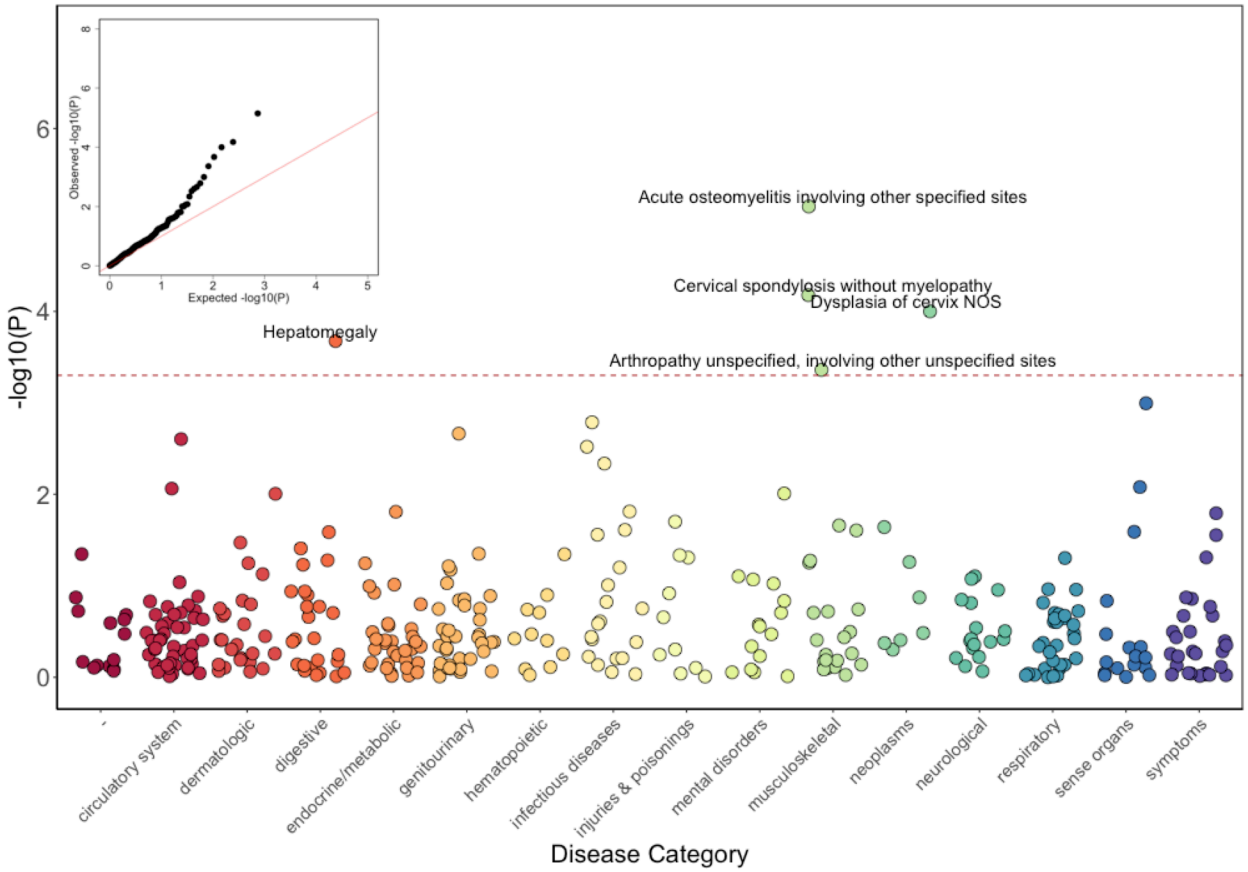

Figure 3 Phenome-Wide Association Study (PheWAS) of *COL27A1*.pG697R carriers vs ICD9 billing codes derived from the Electronic Health Records (EHR) under a general linear model (GLM). Five billing codes achieve significance (FDR adjusted p<0.05). Three of the five significant ICD9 codes are in category of musculoskeletal disorders.

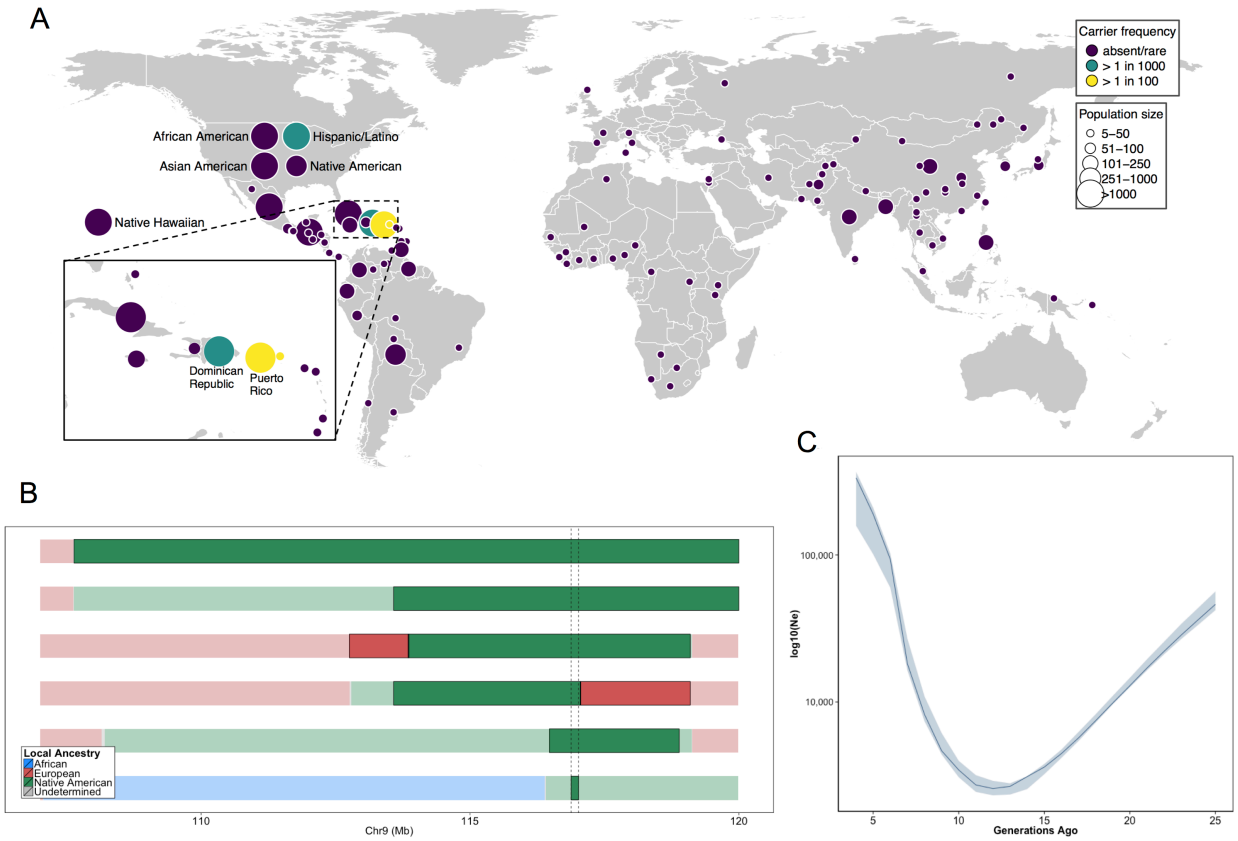

Figure 4. (A) Global carrier frequency of *COL27A1*.pG697R in a multi-ethnic database of over 57,000 individuals representing 112 populations. The variant is absent or very rare in most populations (purple), at 1:746 and 1:346 carrier frequency amongst individuals from the Dominican Republic and Hispanic/Latino's in the United States (green), and at 1:51 and 1:9 carrier frequency amongst individuals from Puerto Rico and St. Thomas (yellow). (B) Joint analysis of identity-by-descent and local ancestry haplotypes in three individuals homozygous for the *COL27A1*.pG697R variant. A large 15cM interval on chromosome 9 is shown with local ancestry inferred as African (blue), European (red) and Native American (green), with shading to indicating the boundaries of the IBD haplotypes. The location of *COL27A1* is indicated by the dashed line (C) Effective population size of the Puerto Rican discovery population (N=2816) over the past 4-25 generations inferred from the tract length distribution of IBD haplotypes

1252 suggests that the ancestral population underwent a bottleneck approximately 9-14
1253 generations ago. 95% confidence intervals are represented by blue ribbon

1254