nature genetics

Article

https://doi.org/10.1038/s41588-023-01444-5

Converging evidence from exome sequencing and common variants implicates target genes for osteoporosis

Received: 8 February 2022

Accepted: 9 June 2023

Published online: 9 August 2023

Check for updates

Sirui Zhou[1,2,3,15], Olukayode A. Sosina[4,15], Jonas Bovijn[4,15], Laetitia Laurent[1,15], Vasundhara Sharma[1], Parsa Akbari[4], Vincenzo Forgetta[1,5], Lai Jiang[1], Jack A. Kosmicki[4], Nilanjana Banerjee[4], John A. Morris[6], Erin Oerton[7], Marcus Jones[4], Michelle G. LeBlanc[4], Regeneron Genetics Center*, Vincent Idone[8], John D. Overton[4], Jeffrey G. Reid[4], Michael Cantor[4], Goncalo R. Abecasis[4], David Goltzman[9], Celia M. T. Greenwood[1,2,10], Claudia Langenberg[7,11], Aris Baras[4], Aris N. Economides[4], Manuel A. R. Ferreira[4], Sarah Hatsell[8], Claes Ohlsson[12,13], J. Brent Richards[1,2,3,5,14,16] ✉ & Luca A. Lotta[4,16]

In this study, we leveraged the combined evidence of rare coding variants and common alleles to identify therapeutic targets for osteoporosis. We undertook a large-scale multiancestry exome-wide association study for estimated bone mineral density, which showed that the burden of rare coding alleles in 19 genes was associated with estimated bone mineral density ($P < 3.6 \times 10^{-7}$). These genes were highly enriched for a set of known causal genes for osteoporosis (65-fold; $P = 2.5 \times 10^{-5}$). Exome-wide significant genes had 96-fold increased odds of being the top ranked effector gene at a given GWAS locus ($P = 1.8 \times 10^{-10}$). By integrating proteomics Mendelian randomization evidence, we prioritized *CD109* (cluster of differentiation 109) as a gene for which heterozygous loss of function is associated with higher bone density. CRISPR–Cas9 editing of *CD109* in SaOS-2 osteoblast-like cell lines showed that partial CD109 knockdown led to increased mineralization. This study demonstrates that the convergence of common and rare variants, proteomics and CRISPR can highlight new bone biology to guide therapeutic development.

Osteoporosis is a common and costly disease leading to substantial morbidity and disability[1]. Existing therapies for this condition are associated with several side-effects that have led to a 50% decrease in their use[2]. Alternatives to bisphosphonates, such as denosumab[3] and romosozumab[4], may also have notable adverse effects. Therefore, new therapies are required.

Human genetics is among the most reliable methods to identify and validate drug targets that impact clinical care[5–7]. While genome-wide association studies (GWAS) have identified thousands of common (typically noncoding) genetic variants associated with disease[8,9], whole-exome sequencing (WES) association studies have become an effective way to pinpoint drug targets since they can more reliably implicate effector genes, and their direction of effect is often clear[10–14]. However, the identification of robust associations between rare coding alleles and complex traits requires sequencing of hundreds of thousands of individuals. As a result, few studies to date have used data from both GWAS and WES to determine whether their convergence can help identify effector genes and possible drug targets for complex traits.

A full list of affiliations appears at the end of the paper. ✉e-mail: brent.richards@mcgill.ca

Here, we utilize WES data from nearly 300,000 multiancestry participants from the UK Biobank (UKB) to identify genes whose perturbation by rare coding alleles influences ultrasound-derived heel estimated bone mineral density (eBMD), a strong predictor of osteoporosis and fracture[15], representing the largest WES study for this trait to date (Fig. 1). We then combined WES with GWAS findings to nominate a further set of prioritized genes. Next, we integrated evidence from protein quantitative trait loci (pQTLs) Mendelian randomization (MR), followed by biological validation using CRISPR–Cas9 in vitro experiments, to elucidate the functional effects of some of the identified candidate genes. Taken together, these studies show that large-scale WES data used in conjunction with GWAS can pinpoint high-confidence new candidate therapeutic targets for common, complex diseases.

Results

Gene burden associations with eBMD in 300,000 exomes

We performed WES in nearly 300,000 people from the UKB cohort (Supplementary Table 1) and, for each gene in the genome, estimated associations with eBMD for the burden of rare nonsynonymous and/or predicted loss-of-function (pLOF) variants (Methods). In the larger European-ancestry subset of UKB ($n = 278,807$), we identified 17 genes where the burden of rare nonsynonymous or pLOF alleles was associated with eBMD at exome-wide significance ($P < 3.6 \times 10^{-7}$; Table 1 and Fig. 2a). These associations did not arise from common genetic variants since these WES analyses were designed to be independent of eBMD-associated fine-mapped common alleles (Methods). The association estimates for these genes were consistent among 13,125 individuals of African, East Asian or South Asian ancestry from UKB, with no strong evidence of heterogeneity across ancestries (Supplementary Table 2). An exome-wide multiancestry meta-analysis identified two additional genes (*WNT5B* and *KREMEN1*) at exome-wide significance (Table 1, Fig. 2b and Supplementary Table 3), providing 19 exome-wide significant genes in total. Of the 17 genes discovered in the European-ancestry-only analysis, 16 remained significant in the multiancestry meta-analysis, the only exception being *CYP19A1*, which fell just short of the threshold in the multiancestry meta-analysis ($P = 7.4 \times 10^{-7}$). Supplementary Tables 2 and 3 show associations in each ancestry for all exome-wide significant genes; Supplementary Table 4 shows all variants in the *WNT5B*, *KREMEN1* and *CYP19A1* gene burden tests that were observed in only one ancestry.

To complement the gene burden analysis, we also estimated the association with eBMD of individual rare (minor allele frequency (MAF) < 1%) nonsynonymous or pLOF variants. In the European-ancestry subset, we found 15 associated variants ($P < 5 \times 10^{-8}$; Supplementary Table 5), independent of eBMD-associated common alleles (Methods). Three of these variants were in genes not discovered in the gene burden analysis (*FAM20C*, *TCIRG1* and *VASN*; Supplementary Table 5), and their association with eBMD was consistent in the multiancestry analysis (Supplementary Table 6).

Of the 19 genes identified in the gene burden analysis, 3 (*LRP5*, *SOST* and *WNT1*) were part of a set of 56 expert-curated and validated genes implicated in bone mineral density by Mendelian genetics or pharmacological validation (that is positive control genes for osteoporosis; Supplementary Note 1 and Supplementary Table 7)[16,17], corresponding to a 65-fold enrichment compared with what is expected by chance (odds ratio (OR), 65; 95% confidence interval (CI), 11, 237; Fisher's exact test $P = 2.5 \times 10^{-5}$). Of the remaining 16 genes, only five (*MEPE*[18], *DLX3* (ref. 19), *CYP19A1* (ref. 20), *INSC*[11] and *SHBG*[11]) have been previously implicated in rare-variant studies of bone density-related phenotypes in humans, either by population-based exome sequencing or Mendelian genetics studies (Supplementary Note 2).

We tested whether the rare coding variants in these 19 exome-wide significant genes are also associated with fracture (77,223 fracture cases and 358,509 controls) and osteoporosis (20,871 cases and 428,313 controls) and found negative correlations between the effect sizes for eBMD and fracture/osteoporosis (Extended Data Fig. 1), as expected.

Convergent evidence from common and rare genetic variants to identify high-confidence genes

Next, we examined whether independent evidence from WES and GWAS converged on the same genes. As GWAS implicates common variants and loci as opposed to specific genes, we used a validated machine-learning method, called the Effector Index (Ei), to map trait-associated common variants at eBMD GWAS loci to likely effector genes[21]. For each gene at a GWAS locus, Ei provides a probability of causal implication between zero and one, where a value of one represents the strongest evidence of causality. Five genes outside GWAS loci were not scored by Ei (Table 2). Of the 19 exome-wide significant genes, 16 (all but *SLC5A3*, *ZNF367* and *DLX3*) were previously mapped to eBMD lead single nucleotide polymorphisms (SNPs) by the GWAS catalog based on a physical distance criterion (Supplementary Table 8).

The 14 WES-identified genes located in GWAS loci had a median Ei score of 0.88 (interquartile range (IQR), 0.77, 0.91; Fig. 3a,b), compared with a median Ei score of only 0.37 (interquartile range, 0.17, 0.58; Fig. 3b) for the remaining GWAS-loci genes. Among the 14 WES-identified genes, a higher Ei score was correlated with higher statistical strength of association in the gene burden analysis (Fig. 3c).

Taking a locus-centric view, exome-wide significant genes had 96-fold higher odds of being the gene with the highest Ei score at their respective loci when compared with non-exome-wide significant genes (OR, 96.2; 95% CI: 17.6, 1,002.2; Fisher's exact test $P = 1.83 \times 10^{-10}$; Table 2 and Fig. 4). In particular, among the 14 exome-wide significant genes, 11 also had the highest Ei score at their respective loci (Table 2, Fig. 3d,e, Fig. 4 and Supplementary Note 3). Two exome-wide significant genes, *KREMEN1* and *LIF*, were located in the same GWAS locus and were associated independently with eBMD in gene burden analyses (Supplementary Table 9). These two genes had the highest and second-highest Ei scores in this locus (Table 2 and Fig. 4). A distinct GWAS effector gene prioritization method, the gene-level polygenic priority score (PoPS)[22], yielded similar results to the Ei (Extended Data Fig. 2, Supplementary Note 4 and Supplementary Table 10). These findings highlight a notable convergence of common and rare-variant associations for bone density.

We also hypothesized that the Ei and WES could be jointly leveraged to identify additional eBMD effector genes and looked for genes that met a 1% false discovery rate (FDR) threshold but fell short of exome-wide significance in the gene burden analysis (corresponding to $P < 1.49 \times 10^{-5}$ but $\geq 3.6 \times 10^{-7}$; termed '1% FDR group'). We found that 50% of genes (four of eight genes) in the '1% FDR group' had an Ei score >0.75 (compared with 11% of genes in GWAS loci that did not meet that threshold; Fig. 3f), indicating an enrichment for GWAS effector genes in this group. We thus propose these four genes (*EYA2*, *SMAD7*, *SNX8*, *WLS*; Extended Data Fig. 3, Supplementary Table 11 and Supplementary Note 5) as additional eBMD effector genes.

MR of circulating protein abundances with eBMD

Next, we leveraged large-scale proteomics data to provide further evidence implicating specific genes and their protein products in bone mineral density. We used two-sample MR[23] to identify circulating proteins genetically associated with eBMD[24]. First, we identified cis-SNPs associated with 863 circulating protein levels from two proteomic GWAS, the INTERVAL study[25] and the AGES study[26]. Both studies measured circulating proteins using the SomaScan platform, and included 3,301 and 3,200 European-ancestry individuals, respectively. MR analyses revealed that genetically predicted concentrations of 39 circulating proteins from INTERVAL ($P < 9.2 \times 10^{-5}$, corresponding to a Bonferroni correction for 548 proteins tested in INTERVAL) and 45 circulating proteins from AGES ($P < 6.5 \times 10^{-5}$, corresponding to a Bonferroni correction for 775 proteins tested in AGES) were associated with eBMD. In total, there were 62 unique associated proteins, of which 22 were found in both INTERVAL and AGES (Supplementary Table 12).

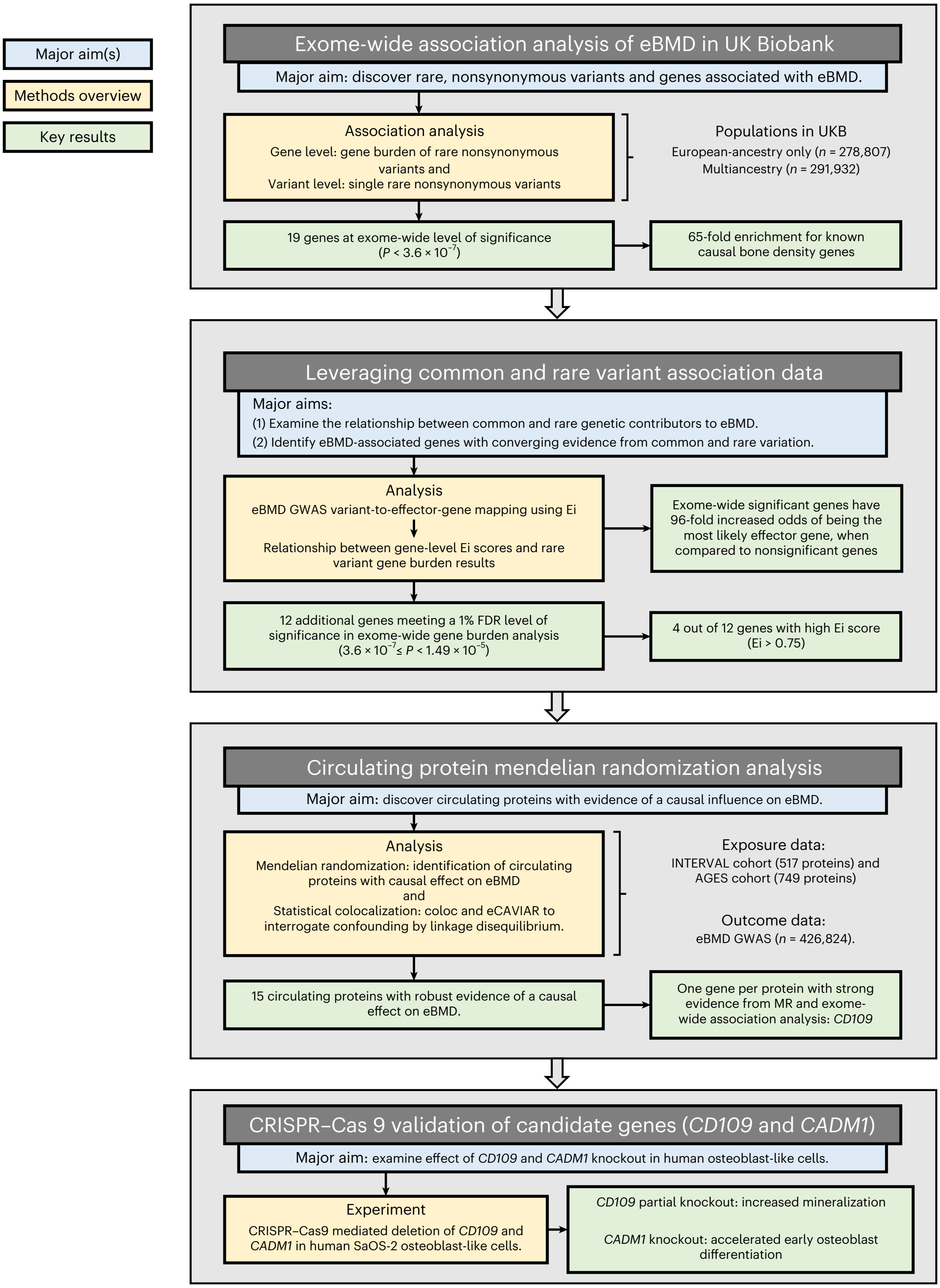

Major aim(s)
Methods overview
Key results
Exome-wide association analysis of eBMD in UK Biobank
Major aim: discover rare, nonsynonymous variants and genes associated with eBMD.
Association analysis
Gene level: gene burden of rare nonsynonymous variants and
Variant level: single rare nonsynonymous variants
Populations in UKB
European-ancestry only (n = 278,807)
Multiancestry (n = 291,932)
19 genes at exome-wide level of significance (P < 3.6 × 10^−7)
65-fold enrichment for known causal bone density genes
Leveraging common and rare variant association data
Major aims:
(1) Examine the relationship between common and rare genetic contributors to eBMD.
(2) Identify eBMD-associated genes with converging evidence from common and rare variation.
Analysis
eBMD GWAS variant-to-effector-gene mapping using Ei
Relationship between gene-level Ei scores and rare variant gene burden results
Exome-wide significant genes have 96-fold increased odds of being the most likely effector gene, when compared to nonsignificant genes
12 additional genes meeting a 1% FDR level of significance in exome-wide gene burden analysis (3.6 × 10^−7 ≤ P < 1.49 × 10^−5)
4 out of 12 genes with high Ei score (Ei > 0.75)
Circulating protein mendelian randomization analysis
Major aim: discover circulating proteins with evidence of a causal influence on eBMD.
Analysis
Mendelian randomization: identification of circulating proteins with causal effect on eBMD
and
Statistical colocalization: coloc and eCAVIAR to interrogate confounding by linkage disequilibrium.
Exposure data:
INTERVAL cohort (517 proteins) and AGES cohort (749 proteins)
Outcome data:
eBMD GWAS (n = 426,824).
15 circulating proteins with robust evidence of a causal effect on eBMD.
One gene per protein with strong evidence from MR and exome-wide association analysis: CD109
CRISPR–Cas 9 validation of candidate genes (CD109 and CADM1)
Major aim: examine effect of CD109 and CADM1 knockout in human osteoblast-like cells.
Experiment
CRISPR–Cas9 mediated deletion of CD109 and CADM1 in human SaOS-2 osteoblast-like cells.
CD109 partial knockout: increased mineralization
CADM1 knockout: accelerated early osteoblast differentiation

Fig. 1 | Overview of this study, describing methods and results for each major aim. Study overview.

MR analysis for a particular protein can, however, yield false positive results when the SNP-protein and SNP-eBMD associations are driven by distinct causal variants that are correlated with each other through linkage disequilibrium (LD)[27]. To identify associations supported by evidence of a shared genetic association, we performed Bayesian colocalization analyses as implemented in coloc[28] and eCAVIAR[29], as the latter allows for more than one causal variant at a locus. Of the 39 prioritized circulating proteins from INTERVAL, 15 were found to colocalize with eBMD using either coloc (posterior probability for colocalization >0.7) or eCAVIAR (SNP-level colocalization posterior probability (CLPP) > 0.01). Of these 15, 12 had consistent MR findings using the AGES cohort data (Supplementary Table 13).

Of the 15 circulating proteins with evidence of association from MR and colocalization analyses, 3 genes (*CD109*, *VTN* and *MRC2*) also had evidence of association with eBMD in the exome-wide gene burden analysis ($P < 0.003$, Bonferroni correction for 15 genes; Table 3 and

Table 1 | Exome-wide gene burden association results for eBMD

Gene (genomic coordinate)	Variants contributing to burden test	Genetic exposure, variant type; frequency cutoff	Beta (95% CI) per allele in s.d. units of eBMD	Beta (95% CI) per allele in g cm^{-2} units of eBMD	*P*	AAF, fraction of 1	Genotype counts, RR|RA|AA genotypes
European-ancestry exome analysis							
LRP5 (11:68312714)	1,100	pLOF plus deleterious missense (1/5); AAF < 1%	−0.11 (−0.13, −0.1)	−0.01 (−0.02, −0.01)	1.4×10^{-47}	0.0240	265,444|13,345|18
MEPE (4:87834714)	56	pLOF; AAF < 1%	−0.36 (−0.41, −0.31)	−0.04 (−0.05, −0.04)	2.7×10^{-41}	0.0020	277,683|1,122|2
WNT1 (12:48978650)	239	pLOF plus any missense; AAF < 1%	−0.12 (−0.14, −0.1)	−0.01 (−0.02, −0.01)	7.66×10^{-22}	0.0092	273,699|5,095|13
INSC (11:15112469)	232	pLOF plus deleterious missense (5/5); AAF < 1%	0.09 (0.07, 0.11)	0.01 (0.01, 0.01)	3.28×10^{-20}	0.0154	270,206|8,587|14
GREM2 (1:240492968)	110	pLOF plus any missense; AAF < 1%	−0.16 (−0.2, −0.13)	−0.02 (−0.02, −0.02)	2.22×10^{-18}	0.0041	276,529|2,274|4
FBXW5 (9:136940927)	582	pLOF plus any missense; AAF < 1%	0.06 (0.05, 0.07)	0.01 (0.01, 0.01)	1.5×10^{-15}	0.0268	263,887|14,903|17
WNT16 (7:121325415)	257	pLOF plus any missense; AAF < 1%	−0.12 (−0.16, −0.09)	−0.02 (−0.02, −0.01)	2.03×10^{-14}	0.0053	275,870|2,933|4
LIF (22:30243650)	176	pLOF plus any missense; AAF < 1%	−0.09 (−0.11, −0.06)	−0.01 (−0.01, −0.01)	1.16×10^{-13}	0.0105	272,995|5,794|18
SMAD6 (15:66703258)	169	pLOF plus deleterious missense (5/5); AAF < 0.1%	−0.2 (−0.26, −0.15)	−0.02 (−0.03, −0.02)	4.68×10^{-13}	0.0018	277,783|1,024|0
SHBG (17:7630172)	206	pLOF plus deleterious missense (1/5); AAF < 1%	0.09 (0.06, 0.11)	0.01 (0.01, 0.01)	1.39×10^{-12}	0.0094	273,598|5,192|17
CD109 (6:73696215)	146	pLOF; AAF < 1%	0.18 (0.12, 0.24)	0.02 (0.02, 0.03)	1.18×10^{-9}	0.0015	277,945|862|0
SOST (17:43755341)	45	pLOF plus deleterious missense (5/5); AAF < 1%	0.45 (0.3, 0.59)	0.05 (0.04, 0.07)	1.47×10^{-9}	0.0003	278,663|144|0
SLC5A3 (21:34095198)	289	pLOF plus deleterious missense (1/5); AAF < 1%	−0.07 (−0.09, −0.05)	−0.01 (−0.01, −0.01)	4.7×10^{-9}	0.0102	273,123|5,668|16
CADM1 (11:115176473)	230	pLOF plus deleterious missense (1/5); AAF < 1%	−0.08 (−0.11, −0.05)	−0.01 (−0.01, −0.01)	7.49×10^{-8}	0.0065	275,177|3,628|2
ZNF367 (9:96388236)	59	pLOF plus deleterious missense (5/5); AAF < 0.1%	−0.41 (−0.57, −0.26)	−0.05 (−0.07, −0.03)	1.58×10^{-7}	0.0002	278,678|129|0
DLX3 (17:49991516)	155	pLOF plus any missense; AAF < 1%	0.08 (0.05, 0.11)	0.01 (0.01, 0.01)	1.88×10^{-7}	0.0067	275,093|3,702|12
CYP19A1 (15:51210807)	168	pLOF plus deleterious missense (5/5); AAF < 0.1%	−0.17 (−0.24, −0.11)	−0.02 (−0.03, −0.01)	2.69×10^{-7}	0.0012	278,117|690|0
Additional genes from multiancestry exome analysis							
WNT5B (12:1631354)	139	pLOF plus deleterious missense (5/5); AAF < 0.1%	0.2 (0.13, 0.28)	0.02 (0.02, 0.03)	9.89×10^{-8}	0.0009	291,390|540|2
KREMEN1 (22:29073130)	146	pLOF plus deleterious missense (5/5); AAF < 0.1%	0.15 (0.09, 0.21)	0.02 (0.01, 0.03)	1.87×10^{-7}	0.0016	290,966|966|0

Table shows genes for which the burden of rare nonsynonymous and/or pLOF variants is associated with eBMD at the exome-wide level of significance ($P<3.6\times10^{-7}$). Genomic coordinates are based on Genome Reference Consortium Human Build 38. Beta is reported relative to the alternative allele. AAF is shown as a fraction of 1 (not percentage). All statistical tests were two-sided, and unadjusted *P* values are presented. RR, reference-reference genotype; RA, reference-alternative heterozygous genotype; AA, alternative-alternative homozygous genotype; Missense (1/5), missense variant predicted to be deleterious by at least one out of five in silico prediction algorithms; Missense (5/5), missense variant predicted to be deleterious by five out of five in silico prediction algorithms.

Supplementary Table 13). *CD109* and *MRC2* displayed concordant directions of effect on eBMD (for example, lower protein concentration and pLoF for a particular gene associated with the same direction of effect on eBMD) (Table 3).

Prioritization of *CD109*

Most therapies act by inhibiting the function, or level, of a target protein. Thus, the *CD109* gene was of particular interest since lower genetically predicted circulating concentration of CD109 protein was associated with higher eBMD. Specifically, each standard deviation lower genetically predicted CD109 concentration was associated with a 0.056 s.d. (or 0.0078 g cm^{-2}) higher eBMD ($P=6.4\times10^{-37}$), with strong evidence of colocalization (posterior probability of shared genetic signal of 0.96 in coloc and CLPP of 0.024 in eCAVIAR). MR estimation of the effect of CD109 protein level on eBMD may be biased by the potential binding efficacy of the aptamer-based proteomic assay, particularly since the *cis*-pQTL for *CD109* in INTERVAL (rs6903575) is in LD with a missense variant (rs10455097, $r^2=0.996$) (ref. 25). However, a variant in high LD with rs6903575 (rs57799429; $r^2=0.99$) has also been reported to be a cis-pQTL for CD109 abundance measured using the

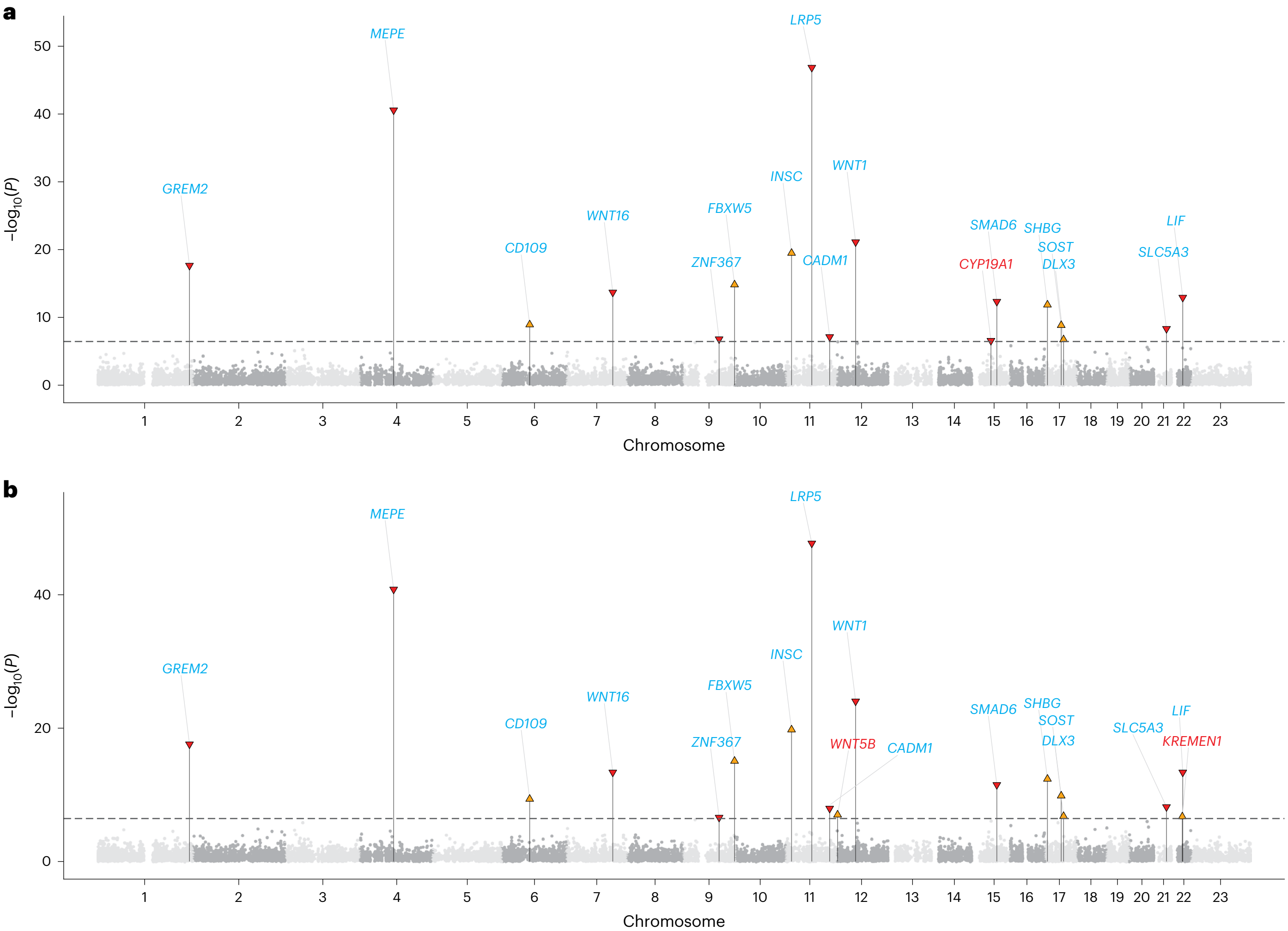


Fig. 2 | Association of rare coding variant burden with eBMD in the exome-wide gene burden analysis. a, European-ancestry analysis results. **b**, Multiancestry analysis results. The dotted line corresponds to the exome-wide level of statistical significance threshold ($P < 3.6 \times 10^{-7}$). Genes in blue are identified in both the European-ancestry analysis and the multiancestry analysis. Genes in red are genes identified only in either the European-ancestry analysis or the multiancestry analysis. Triangles represent the effect direction on eBMD, with downward-facing red triangles representing association with lower eBMD and upward-facing orange triangles association with higher eBMD.

antibody-based Olink assay[30] (SNP-protein association $P = 4.4 \times 10^{-17}$), and MR using this pQTL yielded similar results (beta = −0.047 s.d., $P = 5.9 \times 10^{-37}$). The convergence of results using pQTLs derived from Olink- and SomaLogic-based protein assays may therefore reduce the probability that these findings represent a false positive.

In an entirely independent line of human genetic evidence, that is our exome-discovery analysis, the burden of rare pLOF variants in *CD109* was associated with higher eBMD (per allele beta in SDs of eBMD, 0.18; 95% CI, 0.12–0.24; $P = 1.2 \times 10^{-9}$). Several nonsynonymous and/or pLOF variants in *CD109* were found individually to be associated with eBMD (Supplementary Table 14). This included strong evidence of association for the frameshift Ser1394fs variant (rs766189794; 0.24 s.d. units higher eBMD per allele; $P = 1.7 \times 10^{-7}$), and the missense Phe343Leu variant (rs147944841; 0.11 s.d. units higher eBMD per allele; $P = 1.8 \times 10^{-6}$; Extended Data Fig. 4 and Supplementary Table 14). As predicted by AlphaFold[31], implemented in DECIPHER[32], the Phe343Leu variant might lead to disruption of the two hydrogen bonds with amino acid position 237 and 239, both located in the macroglobulin domain MG3. Finally, when testing whether GWAS evidence also pointed to *CD109* using the Ei SNP-to-gene mapping approach (which is independent of the proteomics evidence presented above), we found strong evidence to support *CD109* as a likely causal gene for common-variant signals at this locus (Ei = 0.96). Taken together, these findings strongly implicate *CD109* as a modulator of eBMD where loss of function may lead to higher bone density in humans.

To assess whether rare pLOF variants in *CD109* were associated with other health traits, we also performed a phenome-wide analysis across 1,108 health phenotypes, and did not find any statistically significant associations besides those with ultrasound bone density measures after multiple-test correction (Supplementary Table 15).

CD109 influences mineralization in osteoblast-like cells

We further characterized the role of *CD109* using CRISPR–Cas9. Stable knockout clones using two single guide RNAs (sgRNAs) were generated in SaOS-2 osteoblast-like cells. These sgRNAs were designed to generate indels in the fifth exon of the *CD109* gene to create truncated proteins. We then selected five clones to assess the degree of reduction in CD109 protein and to assess whether this reduction led to changes in an assay of mineralization from these osteoblast-like cells. Sequencing of exon 5 after CRISPR–Cas9 editing demonstrated that three clones had deletions and one clone had an insertion, whereas one clone had three different deletions (Extended Data Fig. 5a and

Table 2 | Genes associated with eBMD at exome-wide significance and their evidence from common-variant GWAS, predicted by Ei

Exome-wide significant gene	Ei score	Positive control	Gene with highest Ei score at GWAS locus (Ei score)	Ei top gene at GWAS locus concordant with WES
LRP5	0.87	Yes	*TPCN2* (0.89)	No
MEPE	0.87	No	*MEPE* (0.87)	Yes
WNT1	0.83	Yes	*WNT1* (0.83)	Yes
INSC	0.89	No	*INSC* (0.89)	Yes
GREM2	0.91	No	*GREM2* (0.91)	Yes
FBXW5	No Ei	No	–	–
WNT16	0.91	No	*WNT16* (0.91)	Yes
LIF	0.70	No	*KREMEN1* (0.71)	No
SMAD6	No Ei	No	–	–
SHBG	No Ei	No	–	–
CD109	0.96	No	*CD109* (0.96)	Yes
SOST	0.75	Yes	*SOST* (0.75)	Yes
SLC5A3	No Ei	No	–	–
CADM1	0.92	No	*CADM1* (0.92)	Yes
ZNF367	No Ei	No	–	–
DLX3	0.03	No	*COL1A1* (0.85)	No
CYP19A1	0.91	No	*CYP19A1* (0.91)	Yes
KREMEN1	0.71	No	*KREMEN1* (0.71)	Yes
WNT5B	0.93	No	*WNT5B* (0.93)	Yes

'No Ei' indicates genes not located in eBMD GWAS loci. 'Positive control' indicates whether a gene is among a subset of 56 expert-curated genes implicated in bone mineral density by Mendelian genetics or pharmacological validation. The estimates for each gene in the exome-wide gene burden association analysis of eBMD are shown in Table 1.

Supplementary Note 7). With immunoblotting (western blot), we observed a decrease in CD109 protein levels of around 48–69% in the four clones that contained single indels in exon 5 compared with the control cells with wild-type *CD109*, and a complete knockout in the clone with three deletions (analysis of variance (ANOVA) $P = 9.3 \times 10^{-5}$) (Fig. 5a and Extended Data Fig. 5b,c).

After 14 days of treatment with osteogenic factors, there were differences in mineralization for the five edited clones when compared with the control (ANOVA $P = 4.3 \times 10^{-11}$; Fig. 5c). The four clones with knockdown of CD109 had significant increases in mineralization when compared with the wild-type control (Fig. 5b,c). Interestingly, increasing levels of knockdown seemed to correlate with a more modest impact on mineralization and, for the clone with complete knockout of CD109, there was a 61% decrease in mineralization ($P = 0.03$).

CRISPR–Cas9 of *CADM1*

In addition to *CD109*, we also targeted *CADM1* for further functional follow-up given convergent evidence from several sources. First, in our gene burden analysis, rare pLOF plus predicted deleterious missense variants in *CADM1* were associated with eBMD at the exome-wide level of statistical significance (beta, −0.08; 95% CI, −0.11, −0.05; $P = 7.5 \times 10^{-8}$). The effect was in the same direction but larger when considering only rare pLOF variants (beta, −0.21; 95% CI, −0.35, −0.06; $P = 0.004$; Supplementary Table 16). Second, *CADM1* was one of the six exome-identified genes (*GREM2*, *WNT16*, *CD109*, *CADM1*, *CYP19A1* and *WNT5B*) that had a very high Ei score from GWAS data (Ei >0.9). CADM1 is also a cell-surface protein that is expressed in both mouse and human osteoblasts[33], indicating a possible role in bone biology.

We used CRISPR–Cas9 to create double-strand breaks in the first exon of *CADM1* to construct *CADM1*-edited SaOS-2 osteoblast-like cells. We then tested the effect of *CADM1* deletion on differentiation and mineralization for these osteoblast-like cells. Using CRISPR–Cas9, we obtained a decrease of around 98–99% of CADM1 protein level on the surface of the three edited cell lines compared with control cells (Extended Data Fig. 6). When testing the role of cell-surface CADM1 in osteoblast-like cell differentiation, we found increased levels of alkaline phosphatase activity (Extended Data Fig. 7) and early bone markers (Extended Data Fig. 8) in *CADM1*-edited cells; however, no change in late bone markers (Extended Data Fig. 9) or mineralization (Extended Data Fig. 10) was observed, which suggests that CADM1 influences early osteoblast-like cell differentiation (Supplementary Note 6).

Discussion

Identifying effector genes and their direction of effect on a phenotype is required for functional dissection of genetic associations and subsequent drug development. Recent evidence has illustrated the power of WES at scale to identify and confidently attribute genetic association signals, provide directionality of association, inform biology and identify therapeutically modifiable pathways for complex traits[10–14,34,35]. However, since large-scale WES datasets have only recently been generated, the synergistic potential of using these data in conjunction with evidence from common-variant GWAS for gene discovery and target prioritization has not been exploited for many complex traits.

Here, we first pinpointed new bone density genes using a large-scale WES dataset. This approach found 19 genes with robust rare nonsynonymous and/or pLOF variant associations, which were enriched for known effector genes. Only 8 of these 19 genes had been previously implicated in bone density through studies of rare genetic variation in humans, suggesting that large-scale WES in the general population can identify new genes. Importantly, inclusion of individuals of African, East Asian and South Asian ancestries yielded two genes that were not found in the European-ancestry analysis, emphasizing the importance of performing multiancestry analyses with WES data. Second, we combined WES association results with a causal gene prediction method for common-variant GWAS called the Ei to identify a further set of implicated genes. Next, we used MR and proteomics to provide further evidence for the role of several genes in bone density. Finally, functional CRISPR follow-up was able to validate two genes, *CD109* and *CADM1*. These genes were prioritized based on multiple orthogonal lines of evidence, which increases confidence in their role in osteoporosis.

Our study highlights the complementarity and convergence of common and rare-variant association evidence to implicate new genes in common diseases using large datasets. This study also incorporated previous findings of rare variants in Mendelian bone disorders, including sclerosteosis (caused by rare variants in *SOST*) and osteoporosis pseudoglioma syndrome (*LRP5*), in the discovery of such evidence. In addition to the 19 genes identified using an exome-wide discovery approach, we were able to use the combined evidence from WES and GWAS to prioritize four additional genes. While such converging evidence is helpful, it is important to emphasize that associations derived from rare coding variants, which often impart loss of function, provide key insights into the direction of effect of loss of gene function on bone density. WES analysis in conjunction with Ei and other tools can also help prioritize likely effector genes at GWAS loci, particularly in situations where GWAS signals are driven by cis-acting variants. Furthermore, five genes identified using WES were not at known eBMD GWAS loci, underscoring the new contributions that can be made from the analysis of rare genetic variation.

A WES analysis of 3,994 health-related traits in UKB[11] identified five genes (*INSC*, *LRP5*, *MEPE*, *WNT1* and *SHBG*) associated with at least one of the 15 phenotypes from bone-densitometry of the heel in UK Biobank

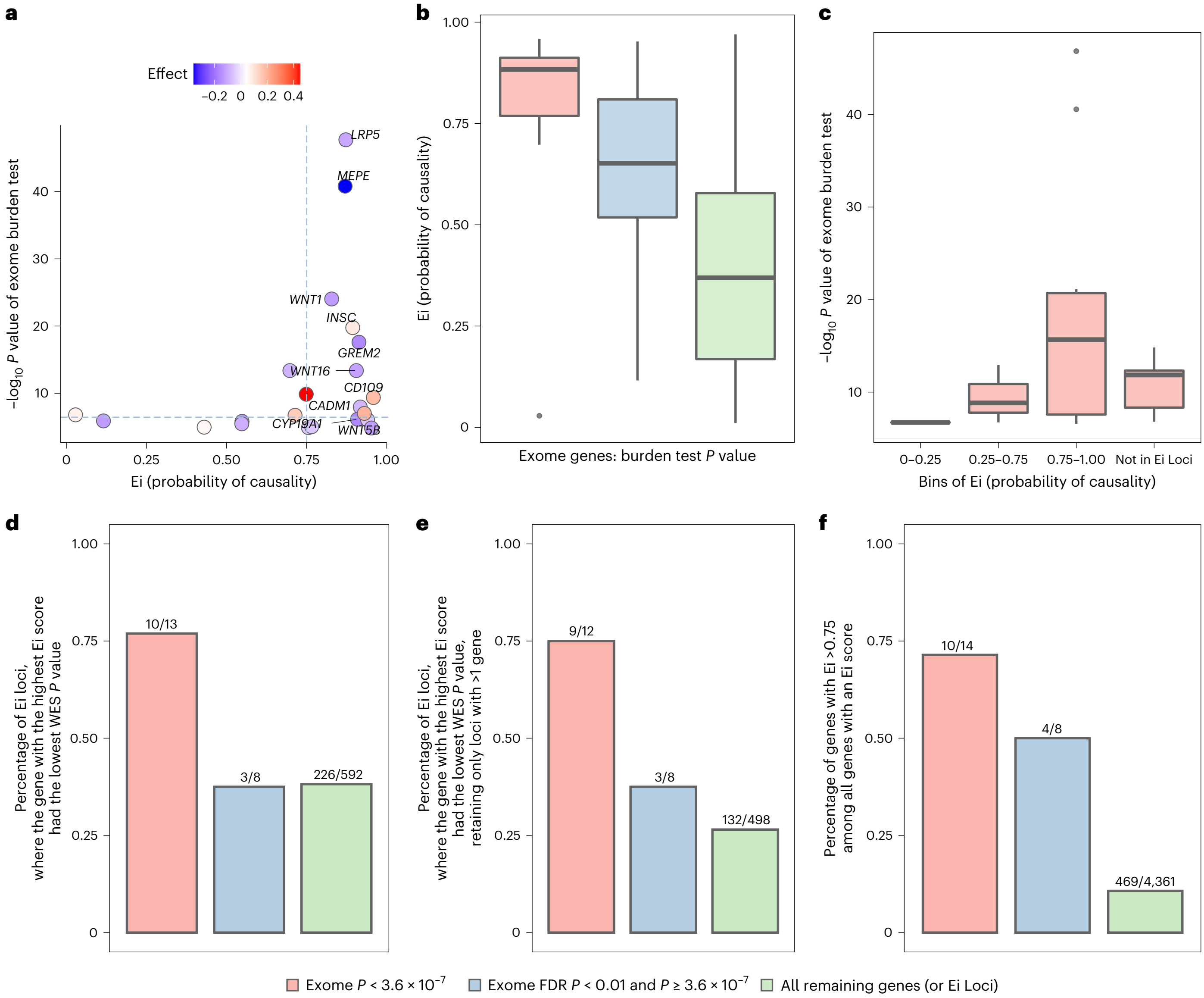


Fig. 3 | Convergent evidence from common and rare genetic variants to identify high-confidence target genes. **a**, Relationship between Ei and multiancestry gene burden analysis P value, among genes with a burden analysis FDR $P < 0.01$ and an assigned Ei score (that is, located in GWAS loci). Named genes are genes reaching the exome-wide statistical significance threshold ($P < 3.6 \times 10^{-7}$, horizontal dashed line) and with Ei >0.75 (vertical dashed line). For the purposes of interpretation, we designated genes with Ei >0.75 as genes having strong evidence of causality in GWAS. Note that *CYP19A1* was significant only in the European-ancestry cohort. The color of each dot indicates the burden test effect size and its direction. **b**, Comparing Ei scores for multiancestry exome-wide significant genes ($n = 14$ genes that have Ei scores), genes with a burden analysis FDR $P < 1\%$ but $P \geq 3.6 \times 10^{-7}$ ($n = 8$ genes that have Ei scores) and all genes not in the two previous categories ($n = 4{,}361$ that have Ei scores). Genes that are not in GWAS loci do not have Ei scores. Box plot shows IQR and median, whisker shows 1.5 IQR of the upper quartile/lower quartile. Each 0.25 unit increase in Ei score was associated with 7.4-fold higher odds (95% CI: 3.1, 17.5, $P = 5.6 \times 10^{-6}$) of a gene being exome-wide significant (red box) in the gene burden analysis. **c**, Multiancestry burden analysis P values of 19 exome-wide significant genes, grouped by their Ei score. Of the 19 exome-wide significant genes, 5 are not located in GWAS loci. Box plot shows IQR and median, whisker shows 1.5 IQR of the upper quartile/lower quartile. **d**, The percentage of loci (y axis) where the gene with the highest Ei score also had the lowest multiancestry gene burden analysis P value at that locus, among: (1) loci with exome-wide significant genes; (2) loci with genes with a burden analysis FDR $P < 1\%$ but $P \geq 3.6 \times 10^{-7}$ and (3) all loci with genes not in the two previous categories. The numbers used to calculate the percentage for each category are indicated on the bar plot. All 613 GWAS loci were included. **e**, The percentage of multigene loci (y axis) where the gene with the highest Ei score also had the lowest multiancestry gene burden analysis P value at that locus, among: (1) loci with exome-wide significant genes; (2) loci with genes with a burden analysis FDR $P < 1\%$ but $P \geq 3.6 \times 10^{-7}$ and (3) all loci with genes not in the two previous categories. The numbers used to calculate the percentage for each category are indicated on the bar plot. A total of 95 GWAS (Ei) loci were excluded from the analysis since they contained only one gene. The genes per loci for each category were mutually exclusive. **f**, The percentage of genes (y axis) with Ei >0.75 among multiancestry exome-wide significant ($P < 3.6 \times 10^{-7}$) genes, genes with a burden analysis FDR $P < 1\%$ but $P \geq 3.6 \times 10^{-7}$ and all genes not in the two previous categories. The numbers used to calculate the percentage for each category are indicated on the bar plot.

(maximum sample size: 246,314). Our eBMD-specific analyses using a refined bone density phenotype, and a larger sample size, yielded 14 additional genes, highlighting the value of the refined phenotyping and trait-centric analytical approach reported here.

We highlighted two genes (*CD109* and *CADM1*) that were supported by evidence from WES, GWAS and proteomics. In our human genetic studies, pLOF variants in *CD109* and common variants associated with lower CD109 circulating protein were associated with higher

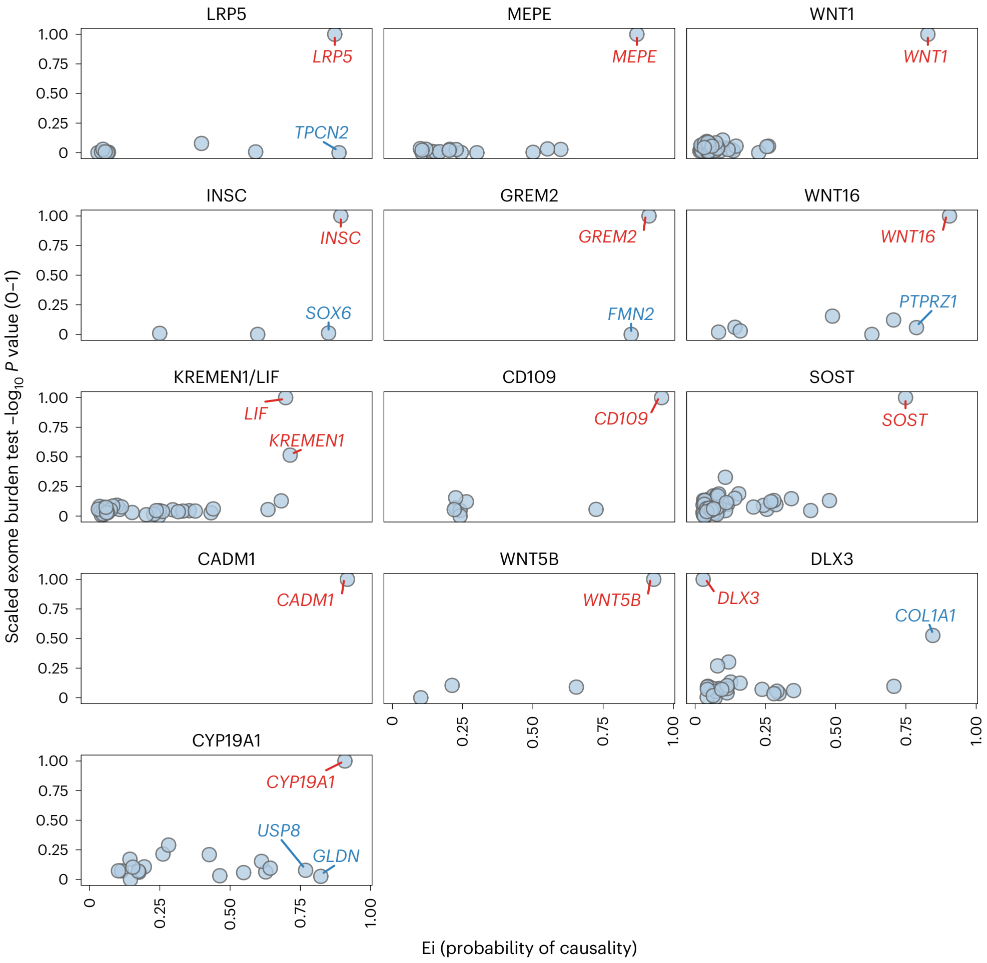


Fig. 4 | Gene burden associations and Ei scores for exome-wide significant genes. Loci are shown if they were identified in eBMD GWAS and included a gene that was identified in our exome-wide rare-variant burden analysis ($P < 3.6 \times 10^{-7}$). Each circle represents a gene at a particular locus. The *y* axis indicates the exome burden analysis $-\log_{10}$ *P* value, scaled between 0 and 1, to allow for visual comparison across loci; the *x* axis indicates the Ei score. Genes highlighted in red text are exome-wide significant genes. Genes highlighted in blue text are other genes with Ei >0.75.

Table 3 | Genes with evidence from circulating protein MR and gene burden testing

Gene	pQTL	Circulating protein MR beta[a]	Circulating protein MR s.e.	Circulating protein MR *P* value	Coloc posterior probability of shared variants	eCAVIAR SNP-level CLPP	Gene burden genetic exposure, variant type; frequency cutoff	Gene burden beta per allele in s.d. units of eBMD	Gene burden *P* value
CD109	rs6903575	0.056	0.004	6.41×10^{-37}	0.96	0.024	pLOF; AAF <1%	0.181	4.3×10^{-10}
VTN	rs704	0.016	0.002	1.09×10^{-18}	1.000	0.503	pLOF; AAF <1%	−0.113	8.6×10^{-4}
MRC2	rs146385050	−0.075	0.011	1.29×10^{-11}	0.95	0.034	pLOF plus deleterious missense (1/5); AAF <0.1%	−0.034	2.7×10^{-3}

[a]Beta corresponds to the effect on eBMD (in standard deviation units), per 1 s.d. decrease in blood protein level. Gene burden estimates are from the multiancestry analysis. All statistical tests were two-sided, and unadjusted *P* values are presented.

eBMD. CRISPR–Cas9-induced partial knockdown of CD109 protein led to an increase in mineralization in SaOS-2 cells. In addition, we also observed that complete knockout of CD109 protein in a set of edited cells led to decreased mineralization, which suggests that the relationship between the degree of CD109 knockdown and mineralization might be nonlinear. We did not observe any individuals who are homozygous for *CD109* pLOF mutations in our data; therefore, the association with bone phenotypes of complete loss of function of the gene could not be estimated in our human genetics analysis. Overall, individuals with partial loss of function due to heterozygous pLOFs in *CD109* had increased eBMD, and osteoblast-like cell lines with partial loss of function of *CD109* exhibited greater mineralization. Taken together, the evidence presented here suggests that partial inhibition of CD109 results in increased bone mineralization in humans.

CD109 is a cell-surface glycoprotein expressed in osteoblasts and has been shown to play a role in osteoclastogenesis[36]. A re-analysis

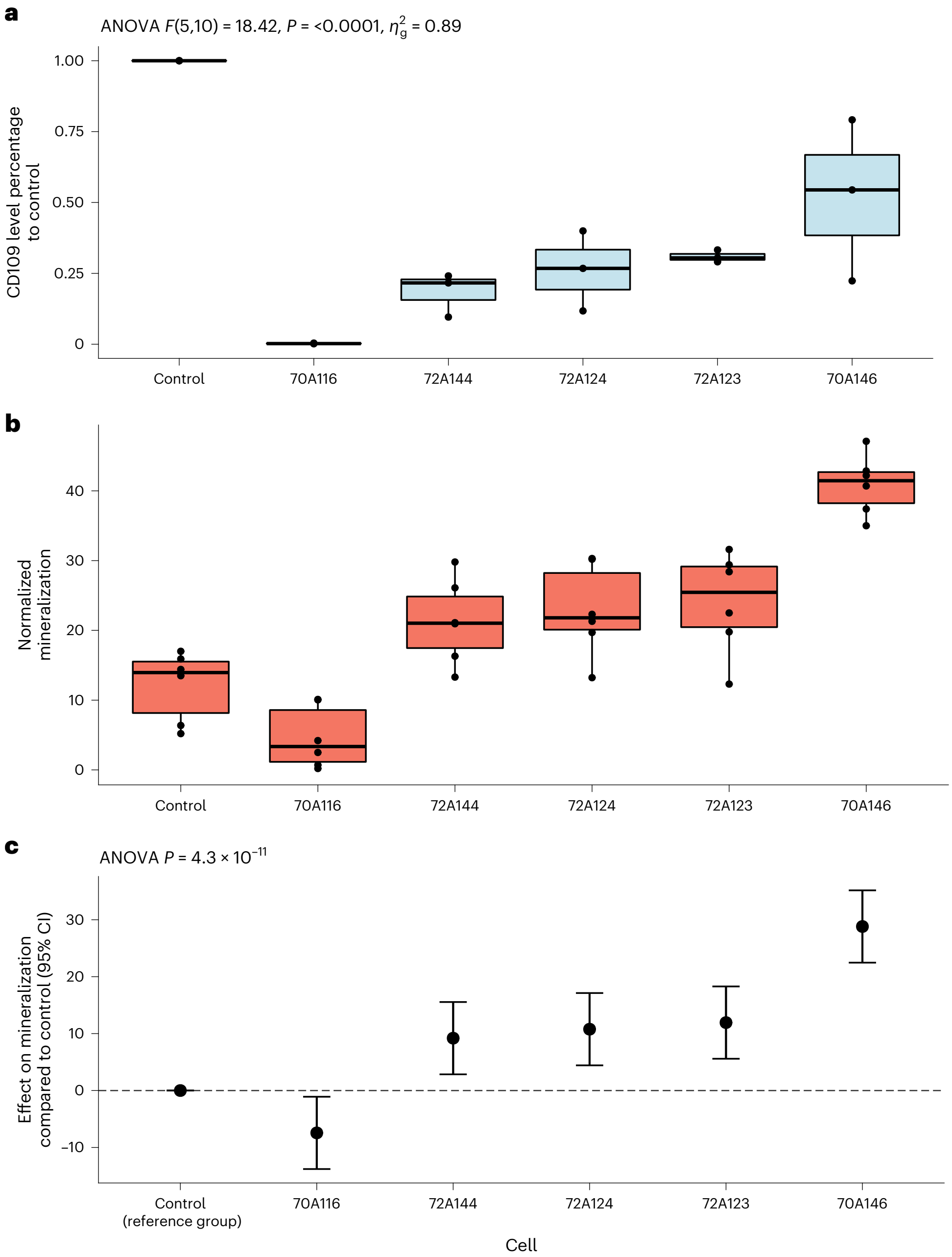


Fig. 5 | *CD109* knockdown in SaOS-2 cells impacts mineralization. **a**, Percentage of CD109 protein level compared with controls in five edited cells. Three independent experiments by western blot were performed for the five clones and controls. One-way repeated measures ANOVA was used to compare the mean relative CD109 level percentage across clones, compared with control. 70A116 shows complete knockout of CD109; the other four cell lines show knockdown of CD109. Box plots in **a** and **b** show IQR and median; whiskers show 1.5 IQR of the upper quartile/lower quartile. All statistical tests were two-sided. **b**, Mineralization levels of five edited cells, normalized against total proteins expressed in the edited cells. Six independent paired experiments were performed for edited cells and wild-type control. **c**, Effect of CD109 editing on mineralization, relative to controls, using generalized least squares controlling for repeated experiments. All four CD109 knockdown cells showed significant increase in mineralization compared with controls (72A144: 1.76-fold, $P = 0.008$; 72A124: 1.86-fold, $P = 0.002$; 72A123: 1.98-fold, $P = 9 \times 10^{-4}$; 70A146: 3.39-fold, $P = 6.5 \times 10^{-10}$), and the CD109 knockout cells showed significant decrease in mineralization compared with controls (70A116: 0.38-fold, $P = 0.03$). $n = 6$ independent experiments per cell. ANOVA was used to compare between the estimates and estimates from null model. Data are presented as mean values ±1.96. All statistical tests were two-sided, and unadjusted *P* values are presented. η^2_g, generalized η squared.

of four male *Cd109* knockout mice found that *Cd109* deficiency may induce a high-turnover, osteoporosis-like phenotype[37]. However, the International Mouse Phenotyping Consortium (IMPC, www.mouse-phenotype.org) found a significant increase in bone mineral density ($P = 9 \times 10^{-4}$) and bone mineral content ($P = 3.3 \times 10^{-4}$) when examining seven male *Cd109* knockout mice[38]. These observations demonstrate heterogeneous effects of *Cd109* on bone in mice, which may be due to the use of separate substrains of C57BL6 mice, resulting in different penetrance and expressivity of mutational effects[39]. Furthermore, in a study by Mii et al.[37], reduced bone mass may have arisen secondary to psoriasis-like skin inflammation, which is linked to bone loss[40], so this model may not be generalizable.

For *CADM1*, high-throughput mouse knockout programs have shown that *Cadm1* knockout mice have decreased femur bone mineral content (BMC) (−3.9 s.d.), vertebral BMC (−3.5 s.d.) and bone strength (−2.0 s.d.) reduction in maximum load required to fracture bones) compared with a wild-type population[41,42]. Previously published *Cadm1* knockout mouse models have also shown a decrease in BMD[43] and a reduction in both bone mass and strength[44]. In addition, IMPC reported decreased BMD ($P = 0.01$) and BMC ($P = 0.001$) in 14 *Cadm1* knockout mice compared with 1,594 wild-type mice[38]. These observations are concordant with our human genetic evidence. In this study, we found that removal of CADM1 protein on the cell surface of osteoblast-like cells via CRISPR–Cas9 resulted in an increase of *RUNX2* mRNA level in the early stages of osteoblast differentiation. RUNX2 is one of the first transcription factors expressed by osteoblast cells and is required for their differentiation as well as for the proper function of mature osteoblasts[45]. It is likely that this increase in *RUNX2* induced higher expression of type 1 collagen (*COL1A1*, *COL1A2*) and *ALPL* mRNA levels at early stages of osteoblast differentiation, as well as the augmentation of later stage alkaline phosphatase activity[46]. Our results suggest that CADM1 could negatively regulate RUNX2, and the absence of CADM1 at the cell-surface accelerated the differentiation of mature osteoblast into osteocytic cells, leading to higher sclerostin mRNA level, independent of RUNX2, contrary to previous evidence[47]. Others have reported that loss of neuronal expression of *Cadm1* also leads to reduced bone mass[44], which highlights the need for further work into the cell types and tissues meditating the effect of CADM1 on bone.

Our study has several limitations. First, we did not assess rare variants outside of protein-coding genes, given our exome-centric analytical design. Second, we did not undertake formal replication in an independent sample, but we applied strict statistical significance thresholds and used several other approaches to validate our results (for example, proteomics MR and CRISPR experiments). Finally, in depth in vivo characterization in animal models may provide additional evidence to further elucidate the biological functions of the identified genes.

In summary, these findings demonstrate that exome sequencing at large scale can be a powerful approach to identify genes for bone density and complex traits more generally. We showed that many of these WES-identified genes have concordant evidence from common-variant GWAS, using the Ei as a method to map associated common variants from GWAS to high-yield target genes. Further, we demonstrated that functional dissection of two candidate genes, *CD109* and *CADM1*, using a CRISPR–Cas9 approach can rapidly shed light on biological pathways influencing bone density. These results provide empirical evidence enabling the efficient design of genomic studies to identify and validate effector genes and potential drug targets for bone density and other complex traits.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01444-5.

References

1. Compston, J. E., McClung, M. R. & Leslie, W. D. Osteoporosis. *Lancet* **393**, 364–376 (2019).
2. Jha, S., Wang, Z., Laucis, N. & Bhattacharyya, T. Trends in media reports, oral bisphosphonate prescriptions, and hip fractures 1996–2012: an ecological analysis. *J. Bone Miner. Res.* **30**, 2179–2187 (2015).
3. Noble, J. A., McKenna, M. J. & Crowley, R. K. Should denosumab treatment for osteoporosis be continued indefinitely? *Ther. Adv. Endocrinol. Metab.* **12**, 20420188211010052 (2021).
4. Kerschan-Schindl, K. Romosozumab: a novel bone anabolic treatment option for osteoporosis? *Wien. Med. Wochenschr.* **170**, 124–131 (2020).
5. King, E. A., Wade Davis, J. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).
6. Cook, D. et al. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13**, 419–431 (2014).
7. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
8. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun. Biol.* **2**, 9 (2019).
9. MacArthur, J. et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
10. Akbari, P. et al. Sequencing of 640,000 exomes identifies *GPR75* variants associated with protection from obesity. *Science* **373**, eabf8683 (2021).
11. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
12. Wang, Q. et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
13. Verweij, N. et al. Germline mutations in *CIDEB* and protection against liver disease. *N. Engl. J. Med.* **387**, 332–344 (2022).
14. Akbari, P. et al. Multiancestry exome sequencing reveals *INHBE* mutations associated with favorable fat distribution and protection from diabetes. *Nat. Commun.* **13**, 4844 (2022).
15. Njeh, C. F., Boivin, C. M. & Langton, C. M. The role of ultrasound in the assessment of osteoporosis: a review. *Osteoporos. Int.* **7**, 7–22 (1997).
16. Rivadeneira, F. & Makitie, O. Osteoporosis and bone mass disorders: from gene pathways to treatments. *Trends Endocrinol. Metab.* **27**, 262–281 (2016).
17. Morris, J. A. et al. An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2018).
18. Surakka, I. et al. *MEPE* loss-of-function variant associates with decreased bone mineral density and increased fracture risk. *Nat. Commun.* **11**, 4093 (2020).
19. Dong, J. et al. DLX3 mutation associated with autosomal dominant amelogenesis imperfecta with taurodontism. *Am. J. Med. Genet. A* **133A**, 138–141 (2005).
20. Bilezikian, J. P., Morishima, A., Bell, J. & Grumbach, M. M. Increased bone mass as a result of estrogen therapy in a man with aromatase deficiency. *N. Engl. J. Med.* **339**, 599–603 (1998).
21. Forgetta, V. et al. An effector index to predict target genes at GWAS loci. *Hum. Genet.* **141**, 1431–1447 (2022).
22. Weeks, E. M. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* https://doi.org/10.1038/s41588-023-01443-6 (2023).
23. Davey Smith, G., Ebrahim, S., Smith, G. D. & Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
24. Dastani, Z. et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.* **8**, e1002607 (2012).
25. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
26. Emilsson, V. et al. Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).
27. Zheng, J. et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).

28. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
29. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
30. Maik, P. et al. Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).
31. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
32. Firth, H. V. et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
33. Inoue, T. et al. Cell adhesion molecule 1 is a new osteoblastic cell adhesion molecule and a diagnostic marker for osteosarcoma. *Life Sci.* **92**, 91–99 (2013).
34. Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
35. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814 (2016).
36. Wang, Y., Inger, M., Jiang, H., Tenenbaum, H. & Glogauer, M. CD109 plays a role in osteoclastogenesis. *PLoS ONE* **8**, e61213 (2013).
37. Mii, S. et al. CD109 deficiency induces osteopenia with an osteoporosis-like phenotype in vivo. *Genes Cells* **23**, 590–598 (2018).
38. Dickinson, M. E. et al. High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508–514 (2016).
39. Simon, M. M. et al. A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome Biol.* **14**, R82 (2013).
40. Lee, J. W., Min, C., Bang, C. H., Kwon, B. C. & Choi, H. G. Psoriasis is associated with an increased risk of osteoporosis: follow-up and nested case–control studies using a national sample cohort. *Osteoporos. Int.* **32**, 529–538 (2021).
41. Bassett, J. H. et al. Rapid-throughput skeletal phenotyping of 100 knockout mice identifies 9 new genes that determine bone strength. *PLoS Genet.* **8**, e1002858 (2012).
42. Freudenthal, B. et al. Rapid phenotyping of knockout mice to identify genetic determinants of bone strength. *J. Endocrinol.* **231**, R31–R46 (2016).
43. Nakamura, S. et al. Negative feedback loop of bone resorption by NFATc1-dependent induction of Cadm1. *PLoS ONE* **12**, e0175632 (2017).
44. Yan, X., Kononenko, N. L., Bruel, A., Thomsen, J. S. & Poy, M. N. Neuronal cell adhesion molecule 1 regulates leptin sensitivity and bone mass. *Calcif. Tissue Int.* **102**, 329–336 (2018).
45. Komori, T. et al. Targeted disruption of Cbfa1 results in a complete lack of bone formation owing to maturational arrest of osteoblasts. *Cell* **89**, 755–764 (1997).
46. Kim, Y. J., Lee, M. H., Wozney, J. M., Cho, J. Y. & Ryoo, H. M. Bone morphogenetic protein-2-induced alkaline phosphatase expression is stimulated by Dlx5 and repressed by Msx2. *J. Biol. Chem.* **279**, 50773–50780 (2004).
47. Sevetson, B., Taylor, S. & Pan, Y. Cbfa1/RUNX2 directs specific expression of the sclerosteosis gene (SOST). *J. Biol. Chem.* **279**, 13849–13858 (2004).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Quebec, Canada. [2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Quebec, Canada. [3]Department of Human Genetics, McGill University, Montréal, Quebec, Canada. [4]Regeneron Genetics Center, Regeneron Pharmaceuticals Inc, Tarrytown, NY, USA. [5]Five Prime Sciences Inc, Montréal, Québec, Canada. [6]New York Genome Center, New York, NY, USA. [7]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK. [8]Regeneron Pharmaceuticals Inc, Tarrytown, NY, USA. [9]Research Institute of the McGill University Health Centre, Montréal, Québec, Canada. [10]Gerald Bronfman Department of Oncology, McGill University, Montréal, Québec, Canada. [11]Computational Medicine, Berlin Institute of Health, Charité University Medicine Berlin, Berlin, Germany. [12]Centre of Bone and Arthritis Research, Department of Internal Medicine and Clinical Nutrition, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [13]Department of Drug Treatment, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden. [14]Department of Twin Research, King's College London, London, UK. [15]These authors contributed equally: Sirui Zhou, Olukayode A. Sosina, Jonas Bovijn, Laetitia Laurent. [16]These authors jointly supervised this work: J. Brent Richards, Luca A. Lotta. *A full list of members appears in the Supplementary Information. ✉e-mail: brent.richards@mcgill.ca

Regeneron Genetics Center

Olukayode A. Sosina[4,15], Jonas Bovijn[4,15], Parsa Akbari[4], Jack A. Kosmicki[4], Nilanjana Banerjee[4], Marcus Jones[4], Michelle G. LeBlanc[4], John D. Overton[4], Jeffrey G. Reid[4], Michael Cantor[4], Goncalo R. Abecasis[4], Aris Baras[4], Aris N. Economides[4], Manuel A. R. Ferreira[4] & Luca A. Lotta[4,16]

Methods

UKB cohort

The UKB is a cohort study of people between 40 and 69 years of age, recruited via 22 testing centers in the UK in 2006–2010. A total of 291,932 participants (278,807 of European ancestry and 13,125 of African, East Asian or South Asian ancestry) with available WES and eBMD data were included in the analyses (Supplementary Table 1).

WES in UKB

Sample preparation and sequencing of the UKB samples were performed at the Regeneron Genetics Center as described[10,34,35] and briefly summarized below. A modified version of the xGen exome design available from Integrated DNA Technologies was used for target DNA capture. Sequencing was performed using 75 bp paired-end reads on Illumina NovaSeq instruments. Sequencing had a coverage depth sufficient to provide >20× coverage over 90% of targeted bases in 99% of samples. Variant calling and annotation were based on the GRCh38 Human Genome reference sequence and Ensembl v.85 gene definitions using the snpEff software. Variants were annotated according to the most deleterious functional effect in this order (of descending deleteriousness): frameshift, stop-gain, stop-loss, splice acceptor, splice donor, in-frame indel, missense and other annotations. pLOF variants included (1) insertions or deletions resulting in a frameshift; (2) insertions, deletions or single nucleotide variants resulting in the introduction of a premature stop codon or in the loss of the transcription start site or stop site and (3) variants in donor or acceptor splice sites. Missense variants were classified for predicted functional impact using a number of in silico prediction algorithms that predicted deleteriousness (SIFT[48], PolyPhen2 (HDIV)[49], PolyPhen2 (HVAR)[49,50], LRT[51] and MutationTaster[52]). For each gene, the AAF and functional annotation of each variant determined inclusion into seven gene burden exposures as previously described[10]: (1) pLOF variants with AAF <1%; (2) pLOF or missense variants predicted deleterious by five out of five algorithms with AAF <1%; (3) pLOF or missense variants predicted deleterious by five out of five algorithms with AAF <0.1%; (4) pLOF or missense variants predicted deleterious by at least one out of five algorithms with AAF <1%; (5) pLOF or missense variants predicted deleterious by at least one out of five algorithms with AAF <0.1%; (6) pLOF or any missense with AAF <1%; (7) pLOF or any missense variants with AAF <0.1%. SNP array genotyping and imputation was performed in the UKB as previously described[53].

Phenotype definition in UKB

eBMD of the heel was derived from quantitative ultrasound speed of sound and broadband ultrasound attenuation using a previously described model[17]. This pipeline yielded a high quality eBMD phenotype while maximizing the number of participants, compared with using the heel bone-densitometry variables provided by UKB directly[11] (UKB Field IDs: 78, 3144, 3146, 3147, 3148, 4101, 4103, 4104, 4105, 4106, 4120, 4122, 4123, 4124, 4125). eBMD is used as a surrogate of BMD because of its high correlation with dual-energy X-ray absorptiometry (DXA)-derived BMD (Pearson's correlation $r = 0.69$) (ref. 54) and its strong association with risk of osteoporotic fracture[55]. Before analysis, we performed rank-inverse normal transformation of the eBMD phenotype, by sex and within each ancestry.

Exome-wide association analyses in UKB

We estimated the association of genetic variants or their gene burden with eBMD by fitting mixed-effects regression models using REGENIE v.1.0.6.8 (ref. 56). REGENIE accounts for relatedness, polygenicity and population structure by approximating the genomic kinship matrix using predictions of individual trait values which are based on genotypes from across the genome. Then, the association of genetic variants or their burden is estimated conditional upon that polygenic predictor along with other covariates. Covariates in association models included age, age squared, sex, age-by-sex interaction term, age squared-by-sex interaction term, experimental batch-related covariates, 10 common-variant derived principal components, and 20 rare-variant derived principal components. To ensure that rare coding variant or gene burden associations were statistically independent of eBMD-associated common genetic variants, we further adjusted exome association analyses for sentinel common variants (MAF ≥ 1%) identified by fine-mapping genome-wide associations of common alleles with eBMD as previously described[10]. Meta-analyses between subgroup results were performed using fixed-effect inverse-variance weighted models. The exome-wide level of statistical significance for the gene burden analysis was defined as $P < 3.6 \times 10^{-7}$, a Bonferroni correction at the type I error rate of 0.05 that assumes 20,000 genes and accounts for the seven variant selection models used per gene[10]. In a secondary analysis, we estimated the association with eBMD of individual nonsynonymous and/or pLOF variants (minor allele frequency <1% and minor allele count ≥25) identified by exome sequencing. The threshold of $P < 5 \times 10^{-8}$, which is a Bonferroni correction based on 1 million effective number of independent tests at the type I error rate of 0.05, was used to identify exome-wide significant single variants as described[10].

For all secondary analyses involving FDR-corrected results, we obtained FDR-adjusted *P* values by first preselecting for each gene, gene burden exposures with the strongest associations (lowest *P* value) and then correcting for multiple testing using the Benjamini–Hochberg approach across all genes in this subset. Hence, the reported FDR threshold of 1% (corresponding to an unadjusted *P* value threshold of 1.49×10^{-5}) is applied to 18,866 genes, after selecting the best gene burden exposure per gene. This translates to an FDR threshold of 2.05%, if we had applied the FDR correction to the overall analysis, and not a preselected subset.

Fine-mapping of GWAS common variants. We identified eBMD-associated common variants by performing a genome-wide association study based on imputed genetic variants. Imputation was based on the HRC reference panel[57] supplemented with UK10K (ref. 58). Genome-wide association analyses were performed in the UKB by fitting mixed-effects linear regression models using REGENIE v.1.0.6.8 (ref. 56). Within each ancestry, fine-mapping was performed using the FINEMAP[59] software at genomic regions harboring genetic variants associated with eBMD at the genome-wide significance threshold of $P < 5 \times 10^{-8}$. LD was estimated using genetic data from the exact set of individuals included in each ancestry-specific genome-wide association analysis.

Test of association with fracture and osteoporosis. We tested the association with fracture and osteoporosis in UKB for genes that met the exome-wide level of statistical significance in the gene burden analysis of eBMD. Fracture cases were defined as individuals with a history of electronic health record-coded or self-reported fracture (not including fractures of the skull, facial bones, hands or toes, where possible), and individuals with a history of any type of fracture were excluded from the control group. Osteoporosis cases were defined as individuals with a history of electronic health record-coded or self-reported osteoporosis, and individuals with a self-reported history of osteopenia were further excluded from the control group.

Test of enrichment for positive control genes for osteoporosis

To evaluate the ability of WES to detect effector genes for osteoporosis, we identified a set of positive control genes for this disease. Fifty-six protein-coding genes that are either known drug targets for osteoporosis or whose perturbation causes a Mendelian form of osteoporosis or bone mass disease, resulting in changes to bone density, bone mineralization or bone mass, were included as positive control genes[17]. We used a Fisher's test to estimate the enrichment for positive control genes among the exome-wide significant genes in the gene burden analysis.

Ei for eBMD effector genes

The development of Ei was described in full in a recent publication[21]. The goal of the Ei is to generate a probability of causality for each protein-coding gene at a GWAS locus, assigning a score from zero to one. GWAS loci were defined by 500 kb around the lead GWAS SNP following LD clumping. Protein-coding genes with at least 50% of their gene body located in a GWAS locus were included[21], and overlapping GWAS loci were merged. In short, to generate Ei scores for eBMD, positive control genes for 12 diseases and traits (type 2 diabetes, low density lipoprotein cholesterol level, adult height, calcium level, hypothyroidism, triglyceride level, eBMD, glucose level, red blood cell count systolic blood pressure, diastolic blood pressure and direct bilirubin level) were selected. GWAS followed by fine-mapping was performed for each disease, and genomic annotations at GWAS loci were used as features to predict positive control genes. This was achieved by first training a gradient boosted trees algorithm (XGBoost) to generate the probability of causality for genes in GWAS loci for 11 diseases and traits (excluding eBMD), and then applying this trained algorithm to derive Ei scores from eBMD GWAS data.

Generalized linear models implemented in R/Rstudio were used to assess the association of the Ei score with the odds of being an exome-wide significant gene.

We used a further, complementary gene prioritization method called PoPS[22] to identify effector genes for eBMD from GWAS data. The generation of the PoP scores for eBMD were described by Weeks et al.[22].

Test of enrichment for Ei prioritized genes within loci identified using exome-wide gene burden results for osteoporosis

Based on the data from Table 2, we generated 2 × 2 contingency tables comparing genes prioritized by Ei with genes identified from the exome-wide analyses per locus. We then aggregated the data across these loci and tested for enrichment using a stratified Fisher's exact test approach[60]. Estimation of the OR and its CI were then based on the conditional maximum likelihood estimate and estimation of the exact confidence bounds using the tail approach for discrete distributions, respectively.

Two-sample MR

We performed two-sample MR analyses to identify circulating proteins that influence eBMD. Two-sample MR uses genetic variants strongly and specifically associated with circulating protein levels (pQTLs) as instrumental variables to estimate the causal relationship between a given protein and an outcome (in this case eBMD). This approach is less affected by confounding and reverse causality than observational epidemiology biomarker studies. The MR framework is based on three main assumptions. First, the SNPs are robustly associated with the exposure. Second, the SNPs are not associated with factors that confound the relationship between the exposure and the outcome. Third, the SNPs have no effect on the outcome that is independent of the exposure (that is, a lack of horizontal pleiotropy). Of these, the most challenging to assess is the third assumption since the biological mechanistic effect of SNPs on outcomes like eBMD is most often not known. However, in the case of circulating proteins, SNPs that are associated with the protein level and close to the gene that encodes the protein are more likely to have an effect via the protein level by influencing the transcription or translation of the gene into the protein. Such SNPs are called cis-SNPs and may help to reduce potential bias from horizontal pleiotropy[27,61].

To select genetic instruments for circulating proteins, we used summary-level data from two proteomic GWAS studies that both measured serum protein levels on the SOMAlogic platform. For the primary analysis, we used as source of pQTL data the INTERVAL[25] study, which included the measurement of 1,478 serum proteins in 3,301 individuals. In a replication analysis, we used the AGES[26] study, which included measurement of 4,137 serum proteins in 3,200 individuals. We selected proteins for inclusion in our analysis if they had cis-acting associated SNPs ('cis-SNPs'), because such instruments may be less likely to be affected by horizontal pleiotropy[62]. The cis-SNPs from INTERVAL were independent, genome-wide significant SNPs ($P < 1.5 \times 10^{-11}$, the multiple-testing corrected genome-wide significance threshold previously adopted in INTERVAL[25]) within 1 Mb of the transcription start site (TSS) of the gene encoding the protein. To select these cis-SNPs, we used PLINK and the 1000 Genomes Project European population reference panel (1KG EUR) to clump and select independent SNPs ($r^2 < 0.001$, distance 1,000 kb) for each protein. The cis-SNPs from AGES were the sentinel cis-SNPs (genome-wide significant SNPs of $P < 5 \times 10^{-8}$ and with the lowest P value for each protein) within 300 kb of the corresponding protein-coding gene[26]. The association of each cis-SNP with eBMD (that is the outcome in our MR analysis) was taken from our recent eBMD GWAS, including 426,824 white British individuals[17]. Palindromic cis-SNPs with MAF > 0.42 (as recommended by the TwoSampleMR R package) were removed before MR to prevent allele-mismatches. For cis-SNPs that were not present in the eBMD GWAS, SNPs in LD ($r^2 > 0.8$) and with MAF < 0.42 were selected as proxies. For the alignment of SNP proxies, MAF > 0.3 was used as a threshold for removal of palindromic SNPs.

After matching of the cis-SNPs of proteins with eBMD GWAS and the removal of palindromic SNPs, 550 SOMAmer reagents (517 proteins) from INTERVAL (including 515 matching cis-SNPs and 59 LD-proxy cis-SNPs; Supplementary Table 17) and 749 circulating proteins from AGES (including 706 unique matching cis-SNPs, 41 LD-proxy cis-SNPs, and two cis-SNPs each for two proteins; Supplementary Table 18) were included in the MR analyses.

MR analyses were performed using the TwoSampleMR[63] package in R, using the Wald ratio ($\beta_{\mathrm{eBMD}}/\beta_{\mathrm{protein}}$) to estimate the effect of each circulating protein on eBMD. For any proteins with multiple independent cis-SNPs, the inverse-variance weighted (IVW) method was used to meta-analyze their combined effects[64]. A Bonferroni correction was used to control for the number proteins tested in INTERVAL and AGES independently.

Colocalization of eBMD and protein abundance

All proteins with MR estimates that met the prespecified significance threshold were investigated further using colocalization analyses. Such analyses are useful to interrogate the potential impact of confounding by LD. This would occur if the exposure and eBMD shared associated variants reflecting different biological signals, which were shared only due to LD. Specifically, for each of these MR significant proteins, a Bayesian analysis implemented in the coloc R package was performed to estimate the posterior probability (PP) that the same causal signal in the 1-Mb genomic locus centered on the cis-SNP affects both circulating protein and eBMD[28]. This analysis was performed in the INTERVAL cohort due to lack of genome-wide SNP-level data from AGES. Coloc assumes that there is a single causal variant driving the association with both traits, which may not hold true in all instances. We therefore also tested the colocalization of the same 1-Mb genomic loci using eCAVIAR[29], which is not subject to this limitation. A maximum of two causal signals were determined for eCAVIAR analysis. We performed eCAVIAR analysis using 1KG EUR reference as the LD panel for the proteomic GWAS and 50,000 random white British individuals from UKB—a subset of the eBMD GWAS cohort[17]—as the LD panel for the eBMD GWAS. A posterior probability of H4 (sharing same signal) >0.7 for coloc or a combined CLPP (SNP-level colocalization posterior probability) score >0.01 for eCAVIAR were used to determine colocalization[29].

CRISPR–Cas9 of CD109 and CADM1 in SaOS-2 cells

For *CD109*, two different guide RNAs (sgRNA) with high MIT Specificity Score (76 and 79) targeting the fifth exon of *CD109* (Supplementary Table 19), as well as a nontargeting guide, were cloned in the PX458 plasmid (pSpCas9(BB)-2A-GFP) from F. Zhang[65] (Addgene plasmid, catalog no. 48138). The construct plasmids were purified using the QIAGEN Plasmid Plus Maxi kit (QIAGEN, catalog no. 12963) according to the manufacturer's instructions. SaOS-2 human osteblastic cells (ATCC, catalog no. HTB-85) were obtained and cultured as described in the Supplementary Note 7. The SaOS-2 cells were then transfected with one of the three different plasmids generated using TransIT LT1 transfection reagent (Mirus catalog no. MIR2304) with a reagent-to-DNA ratio of 3:1. At 48 h posttransfection, GFP-positive cells were sorted by fluorescence-activated cell sorting in a single cell model (Supplementary Note 7). After 21 days, colonies were expanded and then seeded on Lab Tek 8-well chamber slides (40,000 cells per well) to assess CD109 expression using immunofluorescence. At 48 h postseeding, cells were stained with an antibody against CD109 (R&D systems, catalog no. MAB4385; 1/200) followed by goat anti-mouse IgG Alexa Fluor 488 secondary antibody (Abcam, catalog no. ab150113; 1/1,000). DNA from identified clones were extracted using the QIAGEN DNeasy blood and tissue kit (QIAGEN catalog no. 69504) and amplified by PCR using primers designed against regions of CD109 flanking the sgRNA target sequences to generate an amplicon of 312 bp (Supplementary Table 19). PCR products were sequenced using Sanger Sequencing (Genome Quebec) and indels were identified using ICE analysis software (Synthego Performance Analysis, ICE Analysis. 2019. v.3.0. Synthego). Clones with an indel score of about 100% were selected for western blot (Supplementary Note 7) and mineralization.

For mineralization quantification, cells were cultured to 90% confluence in a six-well plate and then treated with osteogenic factors (ascorbic acid 50 μg ml^{-1} and β-gycerophosphate 10 mM). Fresh medium containing osteogenic factors was added every 2–3 days over 13 days. At day 14, mineralization was quantified using the osteogenesis assay kit according to manufacturer's instructions (Millipore, catalog no. ECM815). The alizarin red concentration (μM) was normalized to the protein content as assessed in the medium of each culture (Pierce BCA Protein assay kit; Thermo Fisher, catalog no. 23227). The mineralization quantification of each edited cell culture compared with wild-type clones was repeated six times.

To estimate the effect of edited clones on mineralization compared with wild-type controls, we fit the data to a linear mixed model using generalized least squares ('gls' in 'nlme' R package)[66]. The model is fit by maximizing the log-likelihood. Repeated experiments were used as a grouping factor and correlation structure (compound symmetry) was applied to observations within the same experiment. The *P* value of the linear mixed model was calculated by comparing with the null model, using ANOVA.

For *CADM1*, we aimed to generate a *CADM1* knockout by editing the first exon of *CADM1*. To reduce the likelihood of off-target effects, three different guide RNAs (sgRNA) with high MIT score (88, 89 and 90) (ref. 67) (Supplementary Note 7 and Supplementary Table 19) targeting the first exon of *CADM1* were cloned in the PX458 plasmid as described previously. The construct plasmids were purified using the QIAGEN filter midi prep kit (QIAGEN, catalog no. 12243) according to the manufacturer's instructions. Cells were then transfected with one of the three different plasmids generated, or with the intact plasmid as a control, following methods described previously for *CD109*. After 19 days, cells were stained with an antibody against CADM1 conjugated to phycoerythrin (PE) (MBL Life Science, catalog no. CM004-05; 1/1,000) and the negatives cells were sorted. PCR primers (Supplementary Note 7 and Supplementary Table 19) were designed against regions of *CADM1* flanking the three sgRNA target sequences to generate an amplicon of 331 bp. PCR products of the negative clones were sequenced using MiSeq (Genome Quebec). Western blots and real-time qPCR (Supplementary Note 7) were used to assess CADM1 protein levels and mRNA levels in SaOS-2 *CADM1* knockout and controls, respectively.

Alkaline phosphatase activity was used as a measure of osteoblast differentiation (Supplementary Note 7). Data were analyzed for the CRISPR experiments using GraphPad Prism (v.5.04; GraphPad Software). Statistically significant differences ($P < 0.05$) were determined by unpaired *t*-tests, or by one-way ANOVA followed by a Bonferroni post hoc correction for multiple testing. If parametric conditions were not met, a Kruskal–Wallis test followed by Dunn's post hoc test was used.

Ethics

The use of UKB data was based on ethical approval from Northwest Multi-Centre Research Ethics Committee, and informed consent was obtained from all participants before participation. Data were accessed under applications 26041 and 24268.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Please address general correspondence to J.B.R. (brent.richards@mcgill.ca); for enquiries about exome sequencing and analysis, please contact L.A.L. (luca.lotta@regeneron.com). Individual-level exome sequencing, genotype and phenotype data is available to approved researchers via UKB at: https://www.ukbiobank.ac.uk/enable-your-research. Summary statistics of the following dataset are publicly available and can be accessed at INTERVAL: http://www.phpc.cam.ac.uk/ceu/proteins and eBMD GWAS: http://www.gefos.org/?q=content/data-release-2018 pQTL-only summary statistics of the AGES data is available at: https://www.science.org/doi/suppl/10.1126/science.aaq1327/suppl_file/aaq1327_excel_tables.xlsx. Source data are provided with this paper.

Code availability

REGENIE can be found at https://github.com/rgcgithub/regenie. UKB exome data was analyzed using REGENIE v.1.0.6.8 (Methods). All other data analysis was performed using R (v.3.6.3), RStudio (v.1.4.1717) and eCAVIAR. R packages including twoSampleMR (v.0.4.26), coloc (v.3.2.1) nlme (v.3.1-144), tidyverse (v.1.3.0), ggpubr (v.0.2.5) and ggplot2 (v.3.3.3) were used for analysis and plotting.

References

48. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
49. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
50. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* https://doi.org/10.1002/0471142905.hg0720s76 (2013).
51. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).
52. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575–576 (2010).
53. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
54. Kemp, J. P. et al. Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat. Genet.* **49**, 1468–1475 (2017).
55. Moayyeri, A. et al. Quantitative ultrasound of the heel and fracture risk assessment: an updated meta-analysis. *Osteoporos. Int.* **23**, 143–153 (2012).
56. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

57. Loh, P. R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
58. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **14**, 8111 (2015).
59. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
60. Jung, S. H. Stratified Fisher's exact test and its sample size calculation. *Biometrical J.* **56**, 129–140 (2014).
61. Zhou, S. et al. A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat. Med.* **27**, 659–667 (2021).
62. Swerdlow, D. I. et al. Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *Int. J. Epidemiol.* **45**, 1600–1616 (2016).
63. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
64. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
65. Ran, F. A. et al. Genome engineering using the CRISPR–Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
66. Pinheiro, J., Bates, D. & R Core Team *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-144 https://CRAN.R-project.org/package=nlme (2020).
67. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17**, 148 (2016).

Acknowledgements

This project used UKB exome sequencing data under the project number 26041 and 24268. The Richards research group is supported by the Canadian Institutes of Health Research (CIHR: 365825; 409511, 100558), the McGill Interdisciplinary Initiative in Infection and Immunity (MI4), the Lady Davis Institute of the Jewish General Hospital, the Jewish General Hospital Foundation, the Canadian Foundation for Innovation, the NIH Foundation, Genome Québec, the Public Health Agency of Canada, McGill University, and the Fonds de Recherche Québec Santé (FRQS). J.B.R. is supported by a FRQS Mérite Clinical Research Scholarship. S.Z. is supported by a CIHR fellowship and a FRQS postdoctoral scholarship. Support from Calcul Québec and Compute Canada is acknowledged. TwinsUK is funded by the Welcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. These funding agencies had no role in the design, implementation or interpretation of this study.

Author contributions

S.Z., O.A.S., J.B., J.B.R. and L.A.L. designed the study, participated in the acquisition, analysis and interpretation of data, and drafted the manuscript. L.L., V.S., P.A., V.F., L.J., J.A.K., N.B., J.A.M., E.O., M.J., M.G.L., V.I., J.D.O., J.G.R., M.C., G.R.A., D.G., C.M.T.G., C.L., A.B., A.N.E., M.A.R.F., S.H. and C.O. participated in the acquisition, analysis or interpretation of the data, and reviewed the manuscript for important intellectual content. All authors reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

J.B.R. has served as an advisor to GlaxoSmithKline and Deerfield Capital. He is the founder of 5 Prime Sciences (5prime.com). Regeneron Genetics Center and Regeneron co-authors (O.A.S., J.B., P.A., J.A.K., N.B., M.J., M.G.L., V.I., J.D.O., J.G.R., M.C., G.R.A., A.B., A.N.E., M.A.R.F., S.H., L.A.L.) receive salary and own stocks or stock-options from Regeneron Pharmaceuticals Inc. This research received funding from Regeneron Pharmaceuticals Inc. E.O. is currently an employee at AstraZeneca. The remaining authors declare no competing interests.

Additional information

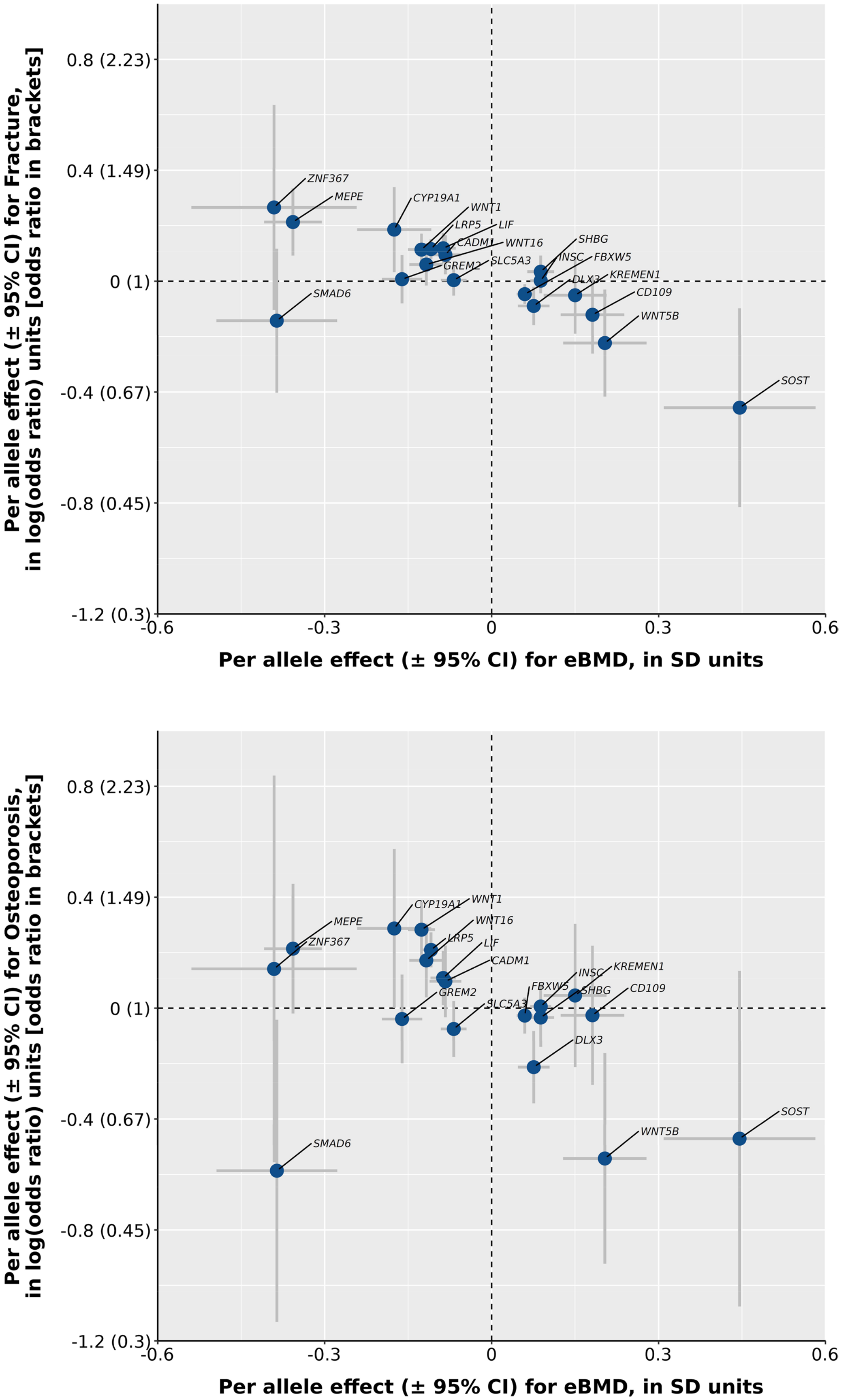
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01444-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01444-5.

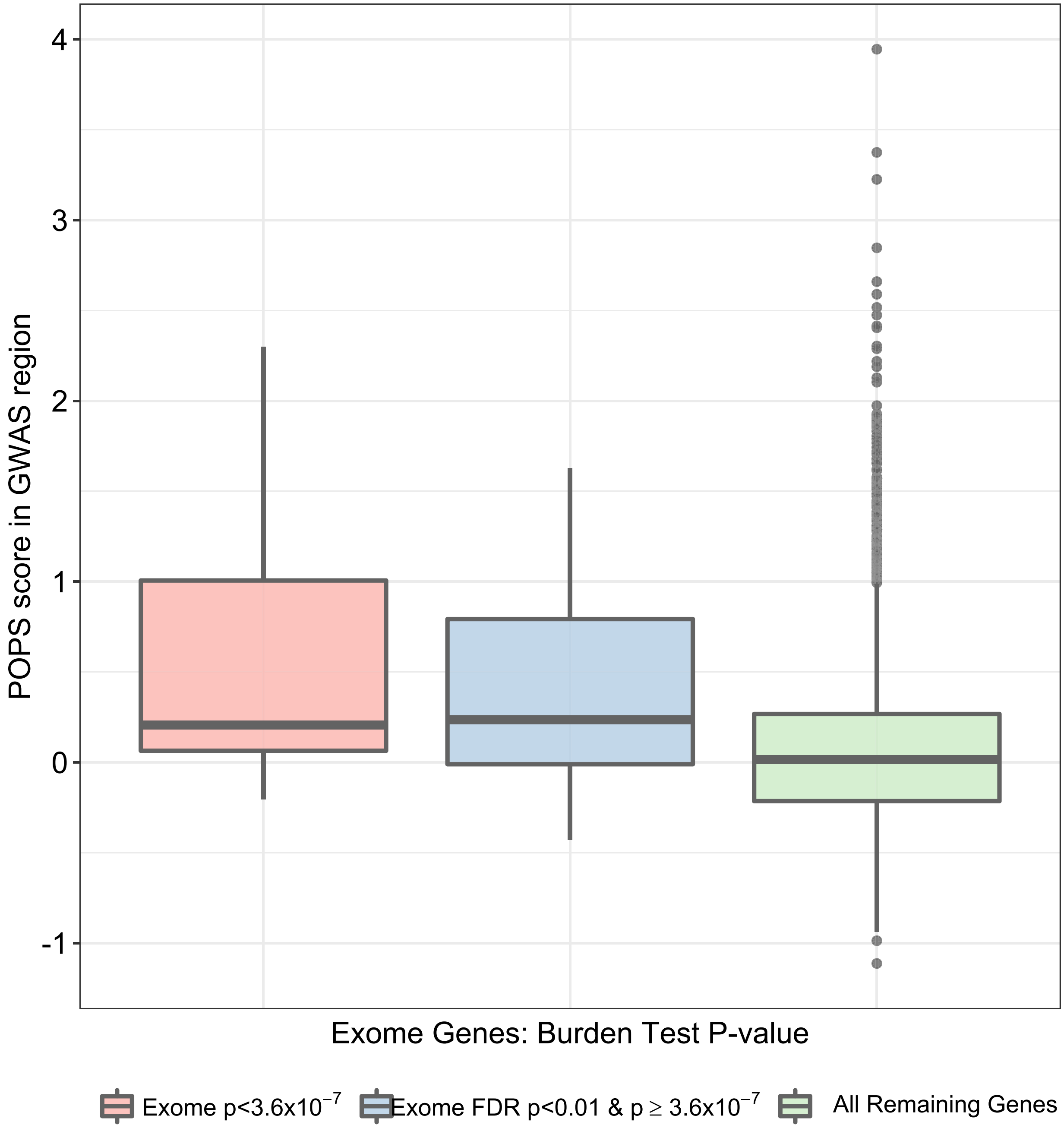
Correspondence and requests for materials should be addressed to J. Brent Richards.

Peer review information *Nature Genetics* thanks Jonathan Tobias, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

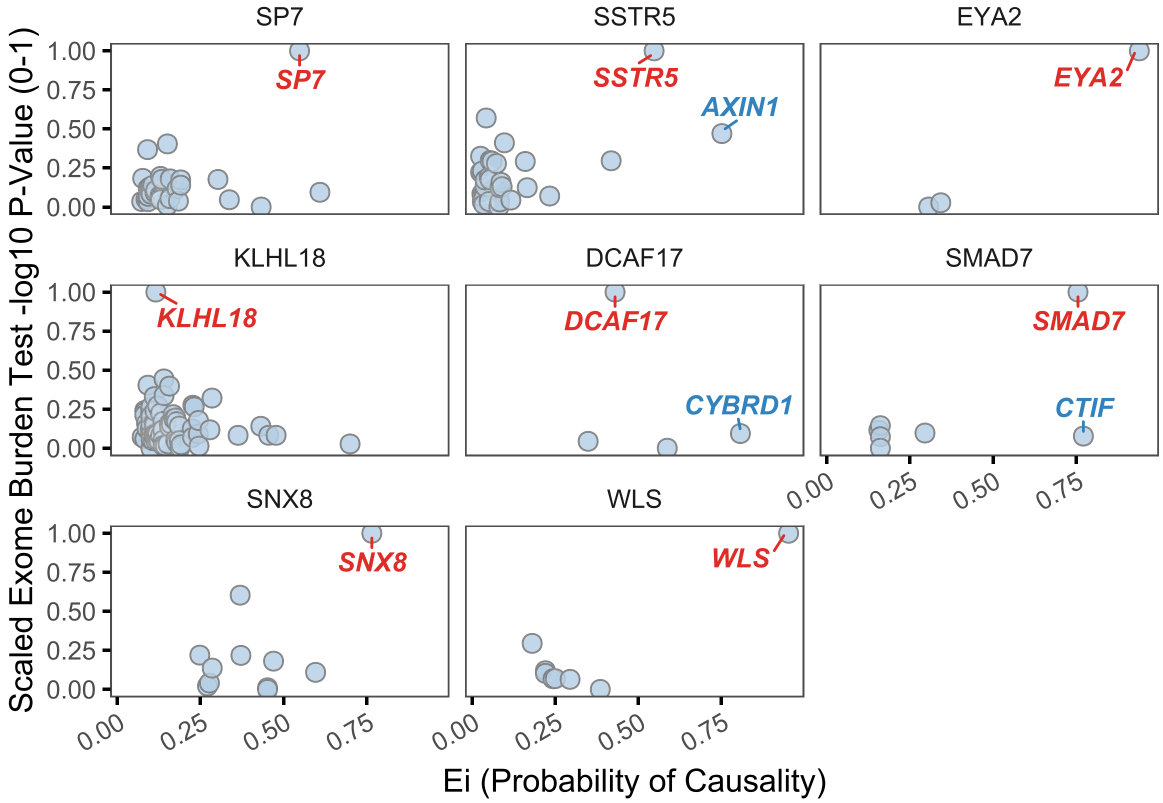


Extended Data Fig. 1 | Associations with fracture and osteoporosis for rare coding variants associated with eBMD. Genes were included in this analysis if they were associated with eBMD at exome-wide significance. For each gene, we used the gene burden exposure with the strongest association (lowest *P*-value) with eBMD. Estimates (point estimates in blue, with 95% confidence intervals as gray lines) for the association with eBMD are shown on the *x*-axis, and estimates for the association with fracture (upper panel) or osteoporosis (lower panel) are on the *y*-axis. The Spearman's rank correlation coefficient of effect sizes was −0.70 (*P* = 0.001; eBMD vs. fracture) and −0.49 (*P* = 0.035; eBMD vs. osteoporosis). SD, standard deviation; CI, confidence interval.



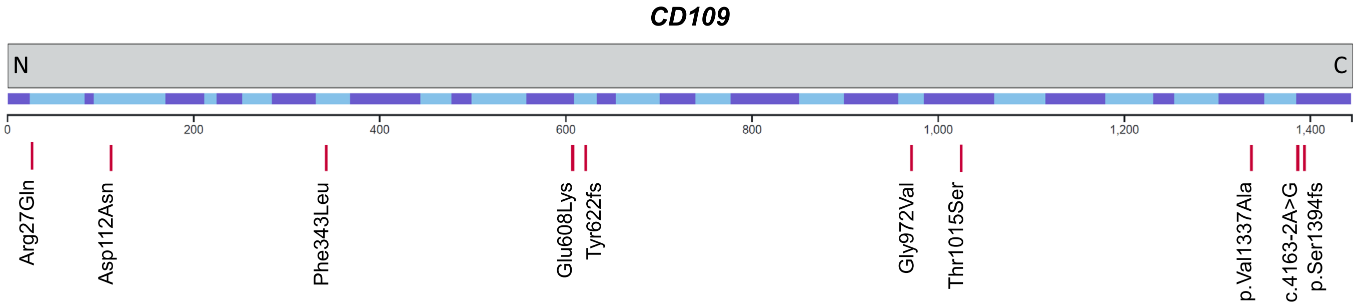
Extended Data Fig. 2 | PoP scores of genes in 857 eBMD GWAS loci. Plots display 17 multi-ancestry exome-wide significant genes, 8 genes with a burden test FDR $P < 1\%$ but $P \geq 3.6 \times 10^{-7}$, and 4,899 genes not in the two previous categories. Box plots show IQR and median; whisker shows 1.5 IQR of the upper quartile/lower quartile.

https://doi.org/10.1038/s41588-023-01444-5

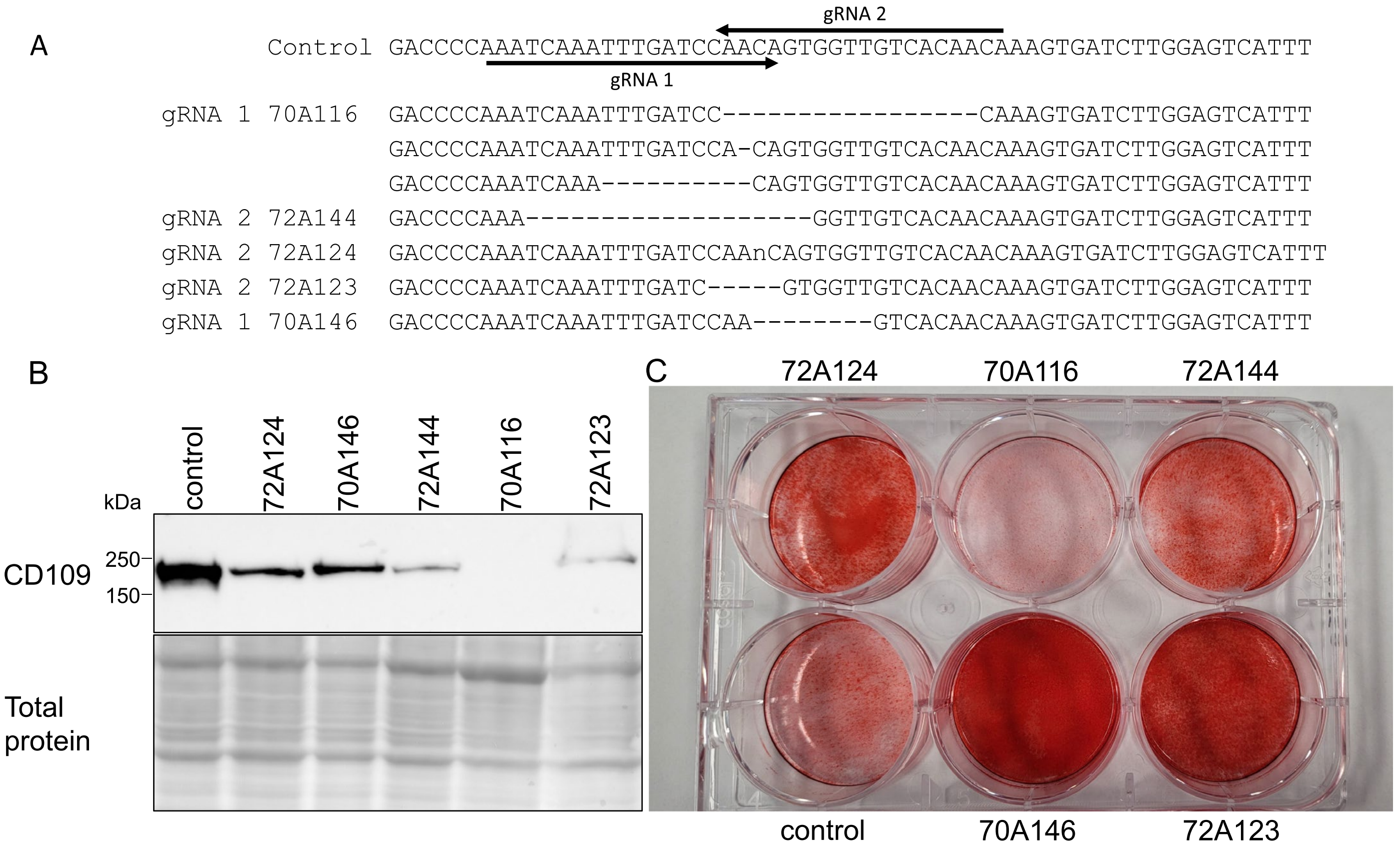


Extended Data Fig. 3 | Gene burden associations and Ei scores for genes with FDR $P < 0.01$ but not reaching exome-wide significance. Loci are shown if they were identified in eBMD GWAS and included a gene that was identified in our cross-ancestry exome-wide rare-variant burden analysis ($P < 1.49 \times 10^{-5}$ but $\geq 3.6 \times 10^{-7}$). Each dot represents a gene in a particular locus. The *y*-axis indicates the exome burden test $-\log_{10}$ *P*-value, scaled between 0–1; the *x*-axis indicates the Ei score. Genes highlighted in red are genes with FDR $P < 0.01$, and genes highlighted in blue are other genes with Ei > 0.75.

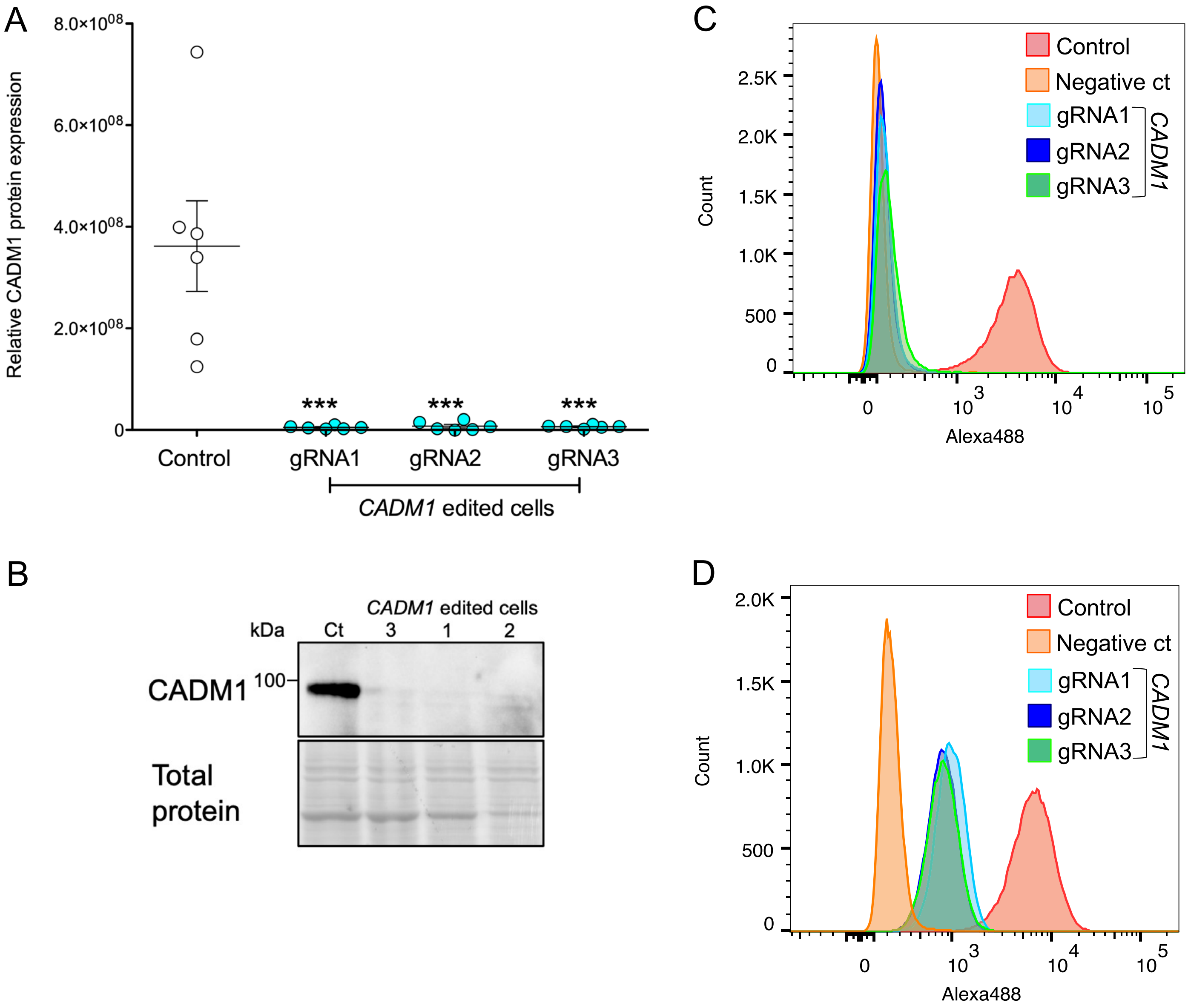
https://doi.org/10.1038/s41588-023-01444-5



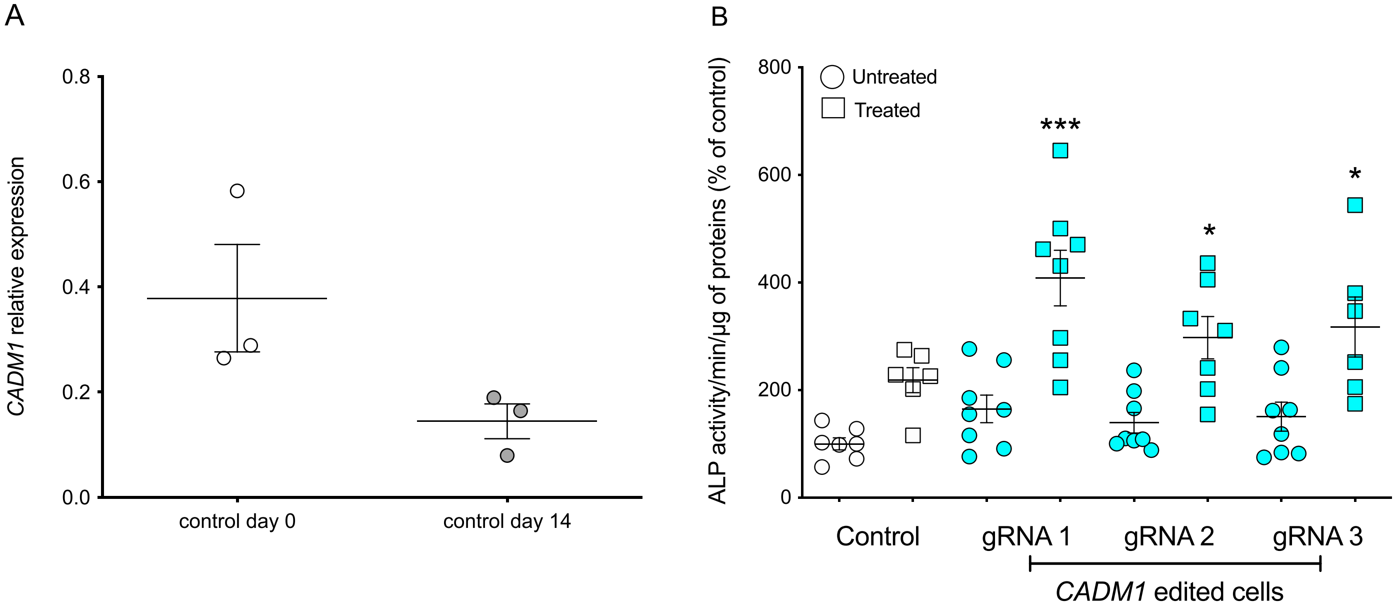
Extended Data Fig. 4 | Distribution of rare nonsynonymous variants in *CD109* with evidence of association with eBMD in exome-wide association analysis. Shown from top to bottom are the CD109 protein (with N and C-terminals indicated), a diagrammatic representation of the *CD109* exons (shown as alternating blue and purple blocks), and the distribution of rare (alternative allele frequency <1%, minor allele count > 25) nonsynonymous variants with evidence of association ($P < 0.05$) with eBMD. c.4163-2 A > G is a splice acceptor variant.


A
gRNA 2
Control GACCCCAAATCAAATTTGATCCAACAGTGGTTGTCACAACAAAGTGATCTTGGAGTCATTT
gRNA 1
gRNA 1 70A116 GACCCCAAATCAAATTTGATCC-----------------CAAAGTGATCTTGGAGTCATTT
GACCCCAAATCAAATTTGATCCA-CAGTGGTTGTCACAACAAAGTGATCTTGGAGTCATTT
GACCCCAAATCAAA----------CAGTGGTTGTCACAACAAAGTGATCTTGGAGTCATTT
gRNA 2 72A144 GACCCCAAA-------------------GGTTGTCACAACAAAGTGATCTTGGAGTCATTT
gRNA 2 72A124 GACCCCAAATCAAATTTGATCCAAnCAGTGGTTGTCACAACAAAGTGATCTTGGAGTCATTT
gRNA 2 72A123 GACCCCAAATCAAATTTGATC-----GTGGTTGTCACAACAAAGTGATCTTGGAGTCATTT
gRNA 1 70A146 GACCCCAAATCAAATTTGATCCAA--------GTCACAACAAAGTGATCTTGGAGTCATTT
B
control
72A124
70A146
72A144
70A116
72A123
kDa
CD109
250
150
Total protein
C
72A124
70A116
72A144
control
70A146
72A123

Extended Data Fig. 5 | *CD109*-edited SaOS-2 cells. a, Sanger sequence of the five edited cells. Indel scores (obtained with Synthego Performance Analysis, ICE Analysis. 2019 v.3.0. Synthego) of the five clones were: 70A146 (96), 72A123 (100), 72A144 (98), 72A124 (97) and 70A116 (84). Two different sgRNAs were used to induce double strand breaks in exon 5 of *CD109* as shown above. Deletion of 8, 5 and 19 nucleotides were obtained in clones 70A146, 72A123 and 72A144, respectively, whereas a single nucleotide insertion was observed in clone 72A124. Clone 70A116 had three different deletions type (1, 10 and 17 nucleotides). **b**, Bands from representative western blots of CD109 (190 kDa; upper panel) and total protein (lower panel) of three independent experiments from wild-type control and *CD109*-edited cells. Full-length blots are provided as Source Data. **c**, A mineralization staining example of the five edited cells from one of the six experiments, where darker red indicates a higher mineral content.

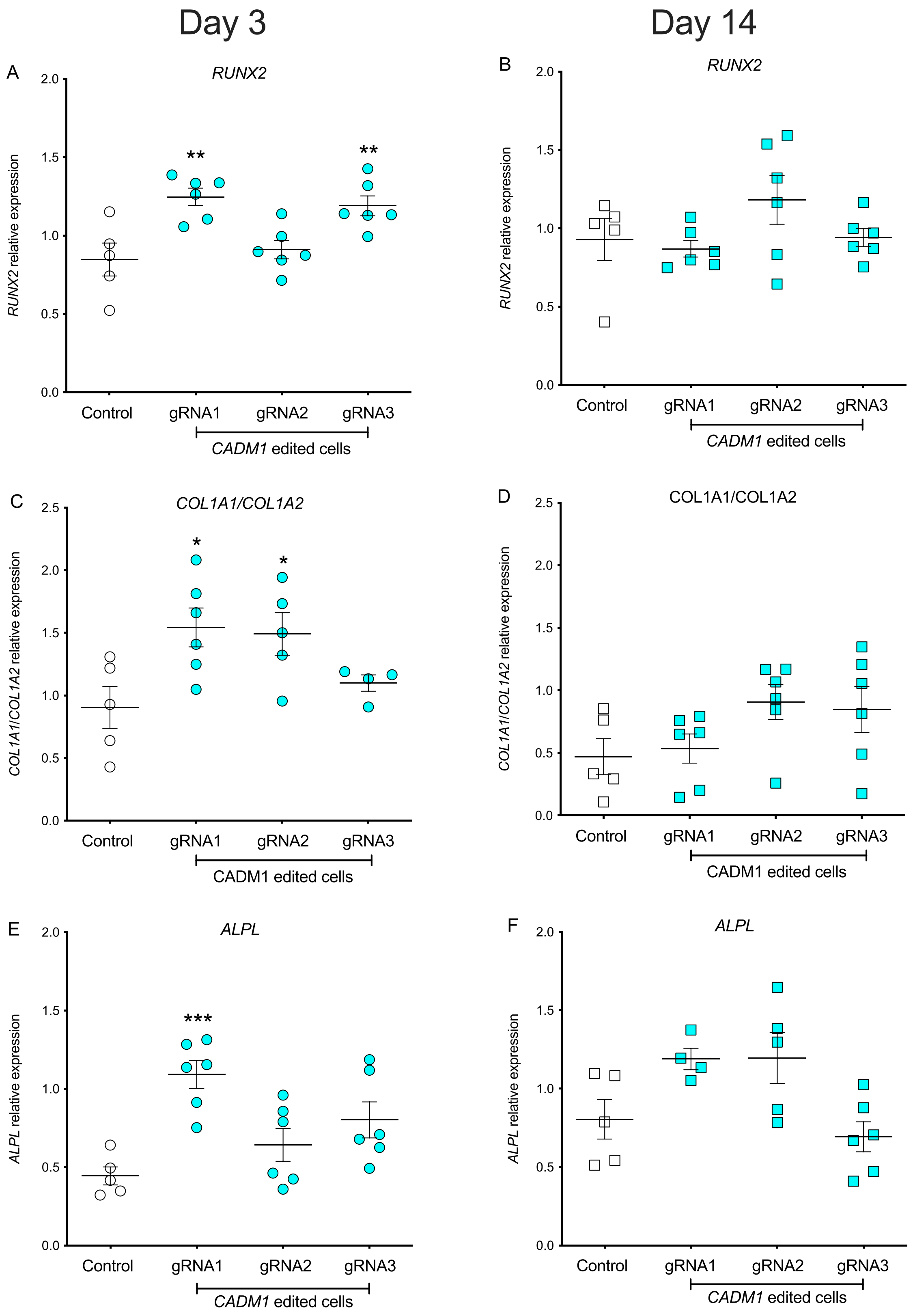


Extended Data Fig. 6 | Targeting *CADM1* exon 1 with CRISPR/Cas9-induced double stranded breaks decreased CADM1 protein level in SaOS-2 cells. **a**, CADM1 protein level quantification in control cells and *CADM1*-edited cells (gRNA1, gRNA2 and gRNA3). Data are presented as mean values +/− standard error of the mean (s.e.m.) of $n = 6$ independent experiments. ***$P = 4.6 \times 10^{-5}$, $P = 5.1 \times 10^{-5}$ and $P = 4.9 \times 10^{-5}$, respectively, compared to control cells determined by one-way ANOVA and Bonferroni post-hoc tests. **b**, Bands from representative western blots of CADM1 (upper panel) and total protein (lower panel) of at least six independent experiments from different cell line passages. Full-length blots are provided as Source Data. **c**, Staining of CADM1 protein using anti-CADM1 monoclonal antibody at the cell surface, showing almost complete knockout of CADM1 using three gRNAs. **d**, Staining of the intracellular CADM1 protein, showing partial knockout of CADM1 using three gRNAs (8.8-, 11- and 10.7-fold decrease compared to controls).



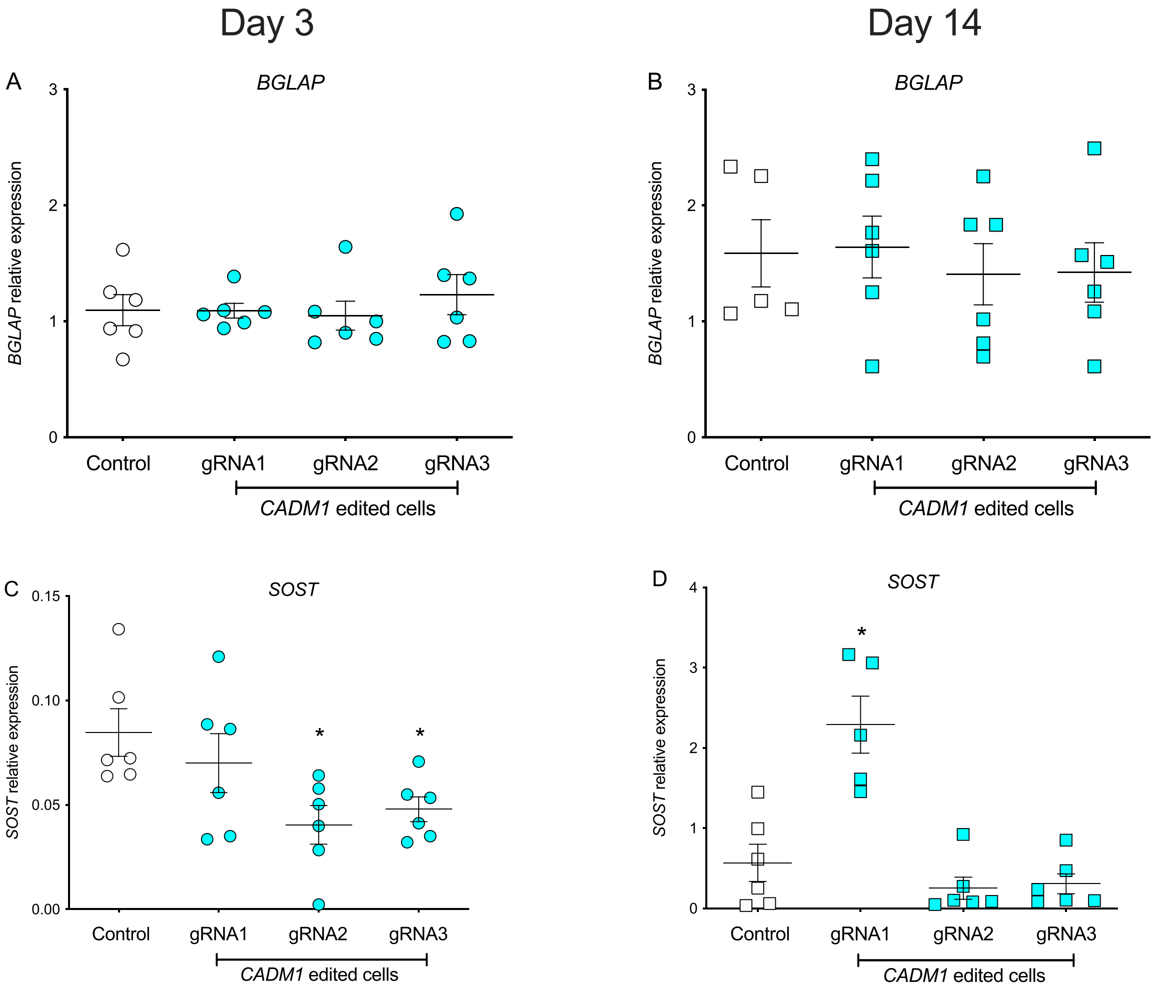
Extended Data Fig. 7 | CADM1 is expressed in early SaOS-2 cell differentiation and influences alkaline phosphatase activity. a, Relative expression of *CADM1* (mRNA level) to two reference genes *PPIA* and *HPRT1* at day 0 and day 14 in SaOS-2 *CADM1* wild-type cells. Data are presented as mean values +/− s.e.m. of *n* = 3 independent experiments. **b**, In *CADM1*-edited cells, the absence of CADM1 at the cell surface increases the activity of alkaline phosphatase after osteogenic treatment. Data are presented as mean values +/− s.e.m. of *n* = 8 independent experiments. Significant changes were shown between treated and untreated edited cells by gRNA1 (****P* = 3 × 10^{-5}) and between treated and untreated edited cells by gRNA2, gRNA3, respectively (**P* = 0.0176 and *P* = 0.0161), determined by one-way ANOVA and Bonferroni post-hoc tests.

https://doi.org/10.1038/s41588-023-01444-5

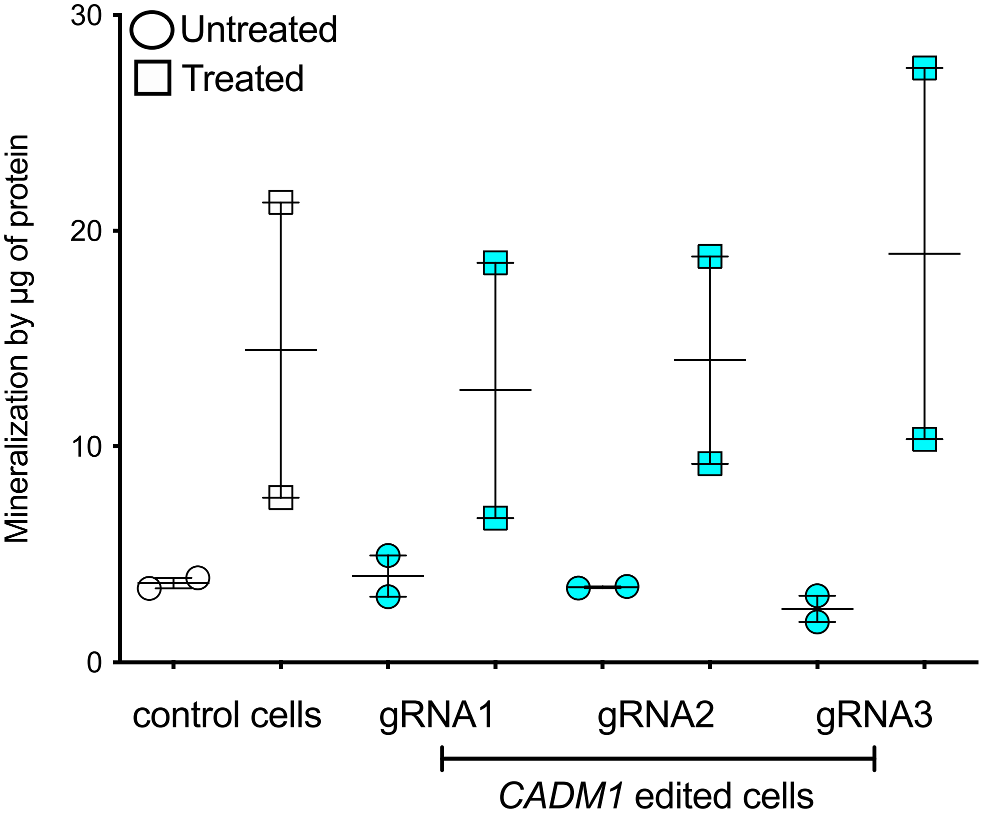


Extended Data Fig. 8 | Expression of early bone markers in *CADM1*-edited SaOS-2 cells. a, **b**, Expression of *RUNX2* mRNA on day 3 ($P = 0.027$ and $P = 0.0096$) and day 14 in *CADM1*-edited SaOS-2 cells. **c**, **d**, Ratio of *COL1A1*/*COL1A2* expression on day 3 ($P = 0.0229$ and $P = 0.0494$) and day 14 in *CADM1*-edited SaOS-2 cells. **e**, **f**, Expression of *ALPL* on day 3 ($P = 0.0006$) and day 14 in *CADM1*-edited SaOS-2 cells. Data are presented as mean values +/− s.e.m. of $n = 6$ independent experiments. Statistical differences compared to control cells were determined by one-way ANOVA and Bonferroni post-hoc tests.

https://doi.org/10.1038/s41588-023-01444-5



Extended Data Fig. 9 | Expression levels of late bone markers in control and *CADM1*-edited cells. a, **b**, Expression of *BGLAP* mRNA on day 3 and day 14 in *CADM1*-edited SaOS-2 cells. **c**, **d**, Expression of *SOST* mRNA on day 3 ($P = 0.0212$ and $P = 0.0480$) and day 14 ($P = 0.0356$) in *CADM1*-edited SaOS-2 cells. Data are presented as mean values +/− s.e.m. of $n = 6$ independent experiments. Statistical differences compared to control cells were determined by one-way ANOVA and Bonferroni post-hoc tests or Kruskal-Wallis and Dunn's multiple comparison test.



Extended Data Fig. 10 | Mineralization of *CADM1*-edited cells after 14 days of treatment with an osteogenic medium. The *y*-axis shows the mineralization levels of three edited cells, normalized against total proteins expressed in the edited cells. Data are presented as mean values +/− s.e.m. of *n* = 2 independent experiments.

nature portfolio

Corresponding author(s): J. Brent Richards & Luca A. Lotta

Last updated by author(s): Jun 8, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	*Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.*
Data analysis	UK Biobank exome data was analyzed using REGENIE v1.0.6.8 (as described in the methods section), all other data analysis were using R (version 3.6.3), RStudio (Version 1.4.1717), eCAVIAR; R packages including twoSampleMR (0.4.26), coloc (3.2.1) nlme(3.1-144), tidyverse(1.3.0), ggpubr(0.2.5), ggplot2(3.3.3) were used for analysis and plotting.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Individual-level sequence data described in this manuscript have been deposited with UK Biobank and will be available to researcher upon approval at: https://www.ukbiobank.ac.uk/enable-your-research
Regenie can be found at https://github.com/rgcgithub/regenie
Summary statistics of the following dataset are publicly available and can be accessed at:

INTERVAL: http://www.phpc.cam.ac.uk/ceu/proteins/
eBMD GWAS: http://www.gefos.org/?q=content/data-release-2018
pQTL-only summary statistics of the AGES data is available at: https://www.science.org/doi/suppl/10.1126/science.aaq1327/suppl_file/aaq1327_excel_tables.xlsx

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   □ Behavioural & social sciences   □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Maximum sample available from UK Biobank (~300,000 exomes) were used.
Data exclusions	All available samples that passed genotype and phenotype QC were included in association analyses. Phenotype selection and QC was performed as described in methods section "Phenotype definition in UKB". Variant level QC was performed as described in methods section " Whole exome sequencing in UKB".
Replication	For CD109 functional analysis, successful replication were performed using 6 repeated measures of mineralization in edited and control cells.
Randomization	No randomization was used in this study
Blinding	No blinding was used in this study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
□	☒	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used	anti-CADM1 antibody (MBL #CM004-5; clone 3E1; lot #040; 1/1,000) anti-CD109 antibody (CST #24765; clone E4I2V; lot #1; 1/500) mouse anti-CD109 monoclonal antibody (R&D System #MAB4385; clone #496920; lot #CBLF0120011; 1/200) goat anti-chicken (Abcam #ab97135; 1/10,000) goat anti rabbit (Abcam #ab6721; 1/5,000) anti-chicken IgY Alexa 488 (Abcam #ab150169; 1/2,000) goat anti-mouse IgG Alexa Fluor 488 (Abcam #ab150113; 1/1,000)
Validation	Validation studies were performed by the commercial vendor (use the catalog and lot number listed above to access this data on the websites of the vendors)

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	SaOS-2 cells were obtained from ATCC (#ATCC HTB-85)

Authentication	SaOS2 were directly purchased from ATCC (lot number 63360718)
Mycoplasma contamination	Cell lines were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Human research participants

Policy information about studies involving human research participants

Population characteristics	UK Biobank population, description can be found in manuscript (table S1) and UK Biobank data showcase
Recruitment	Please refer to UK Biobank data showcase
Ethics oversight	This project used UK Biobank exome sequencing data under the project number 26041 and 24268. UK Biobank received MREC approval, details below: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	500,000 cells/tube were washed in PBS containing 2% FBS then blocked with human FC Block (BD Bioscience #564219) for 10 min at room temperature. CADM1 protein at the cell surface was stained with a chicken anti-CADM1 monoclonal antibody (MBL international #CM004-3; 1/1,000) for 30 min at 4°C followed by the secondary antibody anti-chicken IgY Alexa 488 (Abcam #ab150169; 1/2,000) for 30 min at 4°C. To stain intracellular CADM1 protein, cells were fixed and permeabilized for 20 min at 4°C using a commercially available kit for permeabilization (BD Bioscience #554714). Cells were kept in the permeabilization buffer throughout the experiment to maintain them in a permeabilized state. Cells were then processed in the same way as for CADM1 membrane staining. Finally, cells were resuspend in the appropriate buffer, filtered on a 70-µm strainer and analysed using a BD FACSCanto II instrument.
Instrument	FACSCanto II
Software	BD FACSDiva 8.0.2
Cell population abundance	A minimun of 25K cells have been analysed.
Gating strategy	Single cells were selected by forward scatter and side scatter. Positive population were determined by using the unstained SaOS-2 cells as negative control and the stained non edited cells as positive control.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.