Whole-genome sequencing and deep imputation identifies non-coding variants near *Engrailed-1* with large effects on bone mineral density and fracture

Hou-Feng Zheng [1,2] *, Vincenzo Forgetta [1,2]*, Yi-Hsiang Hsu [3-5]*, Karol Estrada [4-7]*, Alberto Rosello-Diez [8]*, Paul J Leo [9]*, Chitra L Dahia [10]*, Kyung Hyun Park-Min[11]*, Jonathan H Tobias [12,13]*, Charles Kooperberg [14]*, Aaron Kleinman[15], Unnur Styrkarsdottir[16], Ching-Ti Liu [17], Charlotta Uggla [18], Daniel S Evans [19], Carrie M Nielson [20,21], Klaudia Walter [22], Ulrika Pettersson-Kymmer [23,24], Shane McCarthy [22], Joel Eriksson [25], Tony Kwan [26], Mila Jhamai[6], Katerina Trajanoska[6,27], Yasin Memari [22], Josine Min [13], Jie Huang [22], Petr Danecek [22], Beth Wilmot [28,29], Rui Li [1,2], Wen-Chi Chou [3,4], Lauren E Mokry [2], Alireza Moayyeri [30,31], Melina Claussnitzer [3-5], Chia-Ho Cheng[3], Warren Cheung [26,32], Carolina Medina-Gómez [6,27,33], Bing Ge [26], Shu-Huang Chen [26], Kwangbom Choi [34], Ling Oei [6,27,33], James Fraser [35], Robert Kraaij [6,27,33], Matthew A Hibbs [34,36], Celia L Gregson [37,38], Denis Paquette [35], Albert Hofman [27,33], Carl Wibom [39], Gregory J Tranah[20,21], Mhairi Marshall [9], Brooke B Gardiner [9], Paul Auer[40], Li Hsu[14], Sue Ring[41], Joyce Y Tung[15], Gudmar Thorleifsson[42], Anke W Enneman[6] , Natasja M van Schoor[43], Lisette C.P.G.M. de Groot[44], Nathalie van der Velde[6,45], Beatrice Melin [39], John P Kemp [9,13], Claus Christiansen[46], Adrian Sayers [37], Yanhua Zhou [17], Sophie Calderari [47,48], Jeroen van Rooij [6,33], Chris Carlson [14], Ulrike Peters [14], Soizik Berlivet [35], Josée Dostie [35], Andre G Uitterlinden [6,27,33], Stephen R. Williams[49], Charles Farber[49], Daniel Grinberg[50-52], Andrea Z LaCroix[53], Jeff Haessler[14], Daniel I Chasman[4,54], Franco Giulianini[54], Lynda M Rose[54], Paul M Ridker[4,54], John A Eisman[55-57], Tuan V Nguyen[55,57], Jacqueline R Center[55,57], Xavier Nogues[58-60], Natalia Garcia-Giralt[58,59], Lenore L Launer[61], Vilmunder Gudnason[62,63], Dan Mellström[18], Liesbeth Vandenput[18], Magnus K Karlsson[64,65], Östen Ljunggren[66], Olle Svensson[67], Göran Hallmans[24], François Rousseau[68,69], Sylvie Giroux[69], Johanne Bussière[69], Pascal P Arp[6], Fjorda Koromani[6,27], Richard L Prince[70,71], Joshua R Lewis[70,71], Bente L Langdahl[72], A Pernille Hermann[73], Jens-Erik B Jensen[74], Stephen Kaptoge[30], Kay-Tee Khaw[75], Jonathan Reeve[76,77], Melissa M Formosa[78], Angela Xuereb-Anastasi[78], Kristina Åkesson[65,79], Fiona E McGuigan[79], Gaurav Garg[79], Jose M Olmos[80,81], Maria T Zarrabeitia[82], Jose A Riancho[80,81], Stuart H Ralston[83], Nerea Alonso[83], Xi Jiang[84], David Goltzman [85], Tomi Pastinen [26,32], Elin Grundberg [26,32], Dominique Gauguier [47,48], Eric S Orwoll [21,86], David Karasik [3], George Davey-Smith [13], AOGC Consortium, Albert V Smith[62,63], Kristin Siggeirsdottir[62], Tamara B Harris[87], M Carola Zillikens[6], Joyce BJ van Meurs[6,27], Unnur Thorsteinsdottir[16,63], Matthew T Maurano [88], Nicholas J Timpson [13], Nicole Soranzo [22], Richard Durbin [22], Scott G Wilson [31,70,89], Evangelia E Ntzani[90,91], Matthew A Brown [9], Kari Stefansson[63,92] David A Hinds[15], Tim Spector [31], L Adrienne Cupples [17,93], Claes Ohlsson [18,25], Celia Greenwood [2,94,95], Rebecca D Jackson [96]†, David W Rowe[84]†, Cynthia A Loomis [97]†, David M Evans [9,13]†, Cheryl L Ackert-Bicknell [34]†, Alexandra L Joyner [8]†, Emma L Duncan [9,98]†, Douglas P Kiel [3,4,99]†, Fernando Rivadeneira [6,27,33]†, J Brent Richards [1,2,31]† for the GEFOS and UK10K Consortia.

*Denotes shared contribution
†Denotes shared contribution

Affiliations

1. Departments of Medicine, Human Genetics, Epidemiology and Biostatistics, McGill University, Montreal, Canada.
2. Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montreal, Canada.
3. Institute for Aging Research, Hebrew SeniorLife, Boston, USA.
4. Harvard Medical School, Boston, USA.
5. Broad Institute of MIT and Harvard, Boston, USA.
6. Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands
7. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, USA.
8. Department of Developmental Biology, Sloan-Kettering Institute, New York, NY, USA.
9. The University of Queensland Diamantina Institute, Translational Research Institute, Princess Alexandra Hospital, Brisbane, Australia.
10. Tissue Engineering, Regeneration and Repair Program, Hospital for Special Surgery, New York, USA.
11. Rheumatology Divison, Hospital for Special Surgery New York, New York, USA.
12. School of Clinical Science, University of Bristol, Bristol, United Kingdom.
13. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom.
14. Fred Hutchinson Cancer Research Center, Seattle, USA.
15. Department of Research, 23andMe, Inc. Mountain View, USA.
16. Department of Population Genomics, deCODE Genetics, Reykjavik, Iceland.
17. Department of Biostatistics, Boston University School of Public Health, Boston, USA.
18. Internal Medicine and Clinical Nutrition, Institute of Medicine, Sahlgrensk Academy, University of Gothenburg, Gothenburg, Sweden.
19. California Pacific Medical Center, San Francisco, USA.
20. Department of Public Health and Preventative Medicine, Oregon Health & Science University, Portland, USA.
21. Bone & Mineral Unit, Oregon Health & Science University, Portland, USA.
22. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom.
23. Departments of Pharmacology and Clinical Neurosciences, Umeå University, Umeå, Sweden.
24. Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden.
25. Centre for Bone and Arthritis Research, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.
26. McGill University and Genome Quebec Innovation Centre, Montreal, Canada.
27. Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands.
28. Oregon Clinical and Translational Research Institute, Oregon Health & Science University, Portland, USA.
29. Department of Medical and Clinical Informatics, Oregon Health & Science University, Portland, USA.
30. Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom.
31. Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom.
32. Department of Human Genetics, McGill University, Montreal, Canada.
33. Netherlands Genomics Initiative (NGI)-sponsored Netherlands Consortium for Healthy Aging (NCHA), Leiden, The Netherlands.
34. The Jackson Laboratory, Bar Harbor, Maine, USA.
35. Department of Biochemistry, McGill University, Montreal, Canada.
36. Department of Computer Science, Trinity University, San Antonio, Texas, USA.
37. Musculoskeletal Research Unit, University of Bristol, Bristol, United Kingdom.

38. Medical Research Council Epidemiology Unit, University of Southhampton, Southhampton, United Kingdom.
39. Departments of Oncology and Radiation Sciences, Umeå University, Umeå, Sweden.
40. School of Public Health, University of Wisconsin, Milwaukee, USA.
41. School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom
42. Department of Statistics, deCODE Genetics, Reykjavik, Iceland.
43. Department of Epidemiology and Biostatistics and the EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands.
44. Department of Human Nutrition, Wageningen University, Wageningen, The Netherlands.
45. Department of Internal Medicine, Section Geriatrics, Academic Medical Center, Amsterdam, The Netherlands.
46. Nordic Bioscience, Herlev, Denmark.
47. Cordeliers Research Centre, INSERM UMRS 1138, Paris, France.
48. Institute of Cardiometabolism and Nutrition, University Pierre & Marie Curie, Paris, France.
49. Departments of Medicine (Cardiovascular Medicine), Centre for Public Health Genomics, University of Virginia, Charlottesville, Virginia.
50. Department of Genetics, University of Barcelona, Barcelona, Spain.
51. U-720, Centre for Biomedical Network Research on Rare Diseases (CIBERER), Barcelona, Spain.
52. Department of Human Molecular Genetics, The Institute of Biomedicine of the University of Barcelona (IBUB), Barcelona, Spain.
53. Women's Health Center of Excellence Family Medicine and Public Health, University of California - San Diego, San Diego, USA.
54. Division of Preventive Medicine, Brigham and Women's Hospital, Boston, USA.
55. Osteoporosis & Bone Biology Program, Garvan Institute of Medical Research, Sydney, Australia.
56. School of Medicine Sydney, University of Notre Dame Australia, Sydney, Australia.
57. St. Vincent's Hospital & Clinical School, NSW University, Sydney, Australia.
58. Musculoskeletal Research Group, Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona, Spain.
59. Cooperative Research Network on Aging and Fragility (RETICEF), Institute of Health Carlos III, Spain.
60. Department of Internal Medicine, Hospital del Mar, Universitat Autònoma de Barcelona, Barcelona, Spain.
61. Neuroepidemiology Section, National Institute on Aging, National Institutes on Health, Bethesda, USA.
62. Icelandic Heart Association, Kopavogur, Iceland.
63. Faculty of Medicine, University of Iceland, Reykjavik, Iceland.
64. Clinical and Molecular Osteoporosis Research Unit, Department of Clinical Sciences, Lund University, Lund, Sweden.
65. Department of Orthopaedics, Skåne University Hospital Malmö, Sweden.
66. Department of Medical Sciences, University of Uppsala, Uppsala, Sweden.
67. Department of Surgical and Perioperative Sciences, Umeå Unviersity, Umeå, Sweden.
68. Department of Molecular Biology, Medical Biochemistry and Pathology, Université Laval, Québec City, Canada.
69. Axe santé des populations et pratiques optimales en santé, Centre de recherche du CHU de Québec, Québec City, Canada.
70. Department of Endocrinology and Diabetes, Sir Charles Gairdner Hospital, Nedlands, Australia.
71. Department of Medicine, University of Western Australia, Perth, Australia.

72. Department of Endocrinology and Internal Medicine, Aarhus University Hospital, Aarhus C, Denmark.
73. Department of Endocrinology, Odense University Hospital, Odense C, Denmark.
74. Department of Endocrinology, Hvidovre University Hospital, Hvidovre, Denmark.
75. Clinical Gerontology Unit, Unviersity of Cambridge, Cambridge, UK.
76. Medicine and Public Health and Primary Care, University of Cambridge, Cambridge, UK.
77. Institute of Musculoskeletal Sciences, The Botnar Research Centre, University of Oxford, Oxford, United Kingdom.
78. Department of Applied Biomedical Science, Faculty of Health Sciences, University of Malta, Msida, Malta.
79. Department of Clinical Sciences Malmö, Lund University, Malmö, Sweden.
80. Department of Medicine and Psychiatry, University of Cantabria, Santander, Spain.
81. Department of Internal Medicine, Hospital U.M. Valdecilla- IDIVAL, Santander, Spain.
82. Department of Legal Medicine, University of Cantabria, Santander, Spain.
83. Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, Western General Hospital, University of Edinburgh, Edinburgh, UK.
84. Department of Reconstructive Sciences, College of Dental Medicine, University of Connecticut Health Center, Farmington, Connecticut, USA.
85. Department of Medicine and Physiology, McGill University, Montreal, Canada.
86. Department of Medicine, Oregon Health & Science University, Portland, USA.
87. Laboratory of Epidemiology, National Institute on Aging, National Institutes of Health, Bethesda, USA.
88. Department of Genome Sciences, University of Washington, Seattle, USA.
89. School of Medicine and Pharmacology, University of Western Australia, Crawley, Australia.
90. Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece.
91. Department of Health Services, Policy and Practice, Brown University School of Public Health, Providence, USA.
92. CEO, deCODE Genetics, Reykjavik, Iceland.
93. Framingham Heart Study, Framingham, USA.
94. Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada.
95. Department of Oncology, McGill University, Montreal, Canada.
96. Department of Medicine, Division of Endocrinology, Diabetes and Metabolism, The Ohio State University, Columbus, Ohio USA.
97. The Ronald O Perelman Department of Dermatology and Department of Cell Biology, New York University School of Medicine, New York USA.
98. Department of Diabetes and Endocrinology, Royal Brisbane and Women's Hospital, Brisbane, Australia.
99. Beth Israel Deaconess Medical Center, Boston, USA.

Correspondence to:
Brent Richards
Pavillon H-413, Jewish General Hospital
3755 Cote Ste Catherine
Montreal, QC, Canada, H3T 1E2
brent.richards@mcgill.ca

Abstract

The extent to which low-frequency (minor allele frequency [MAF] between 1-5%) and rare (MAF ≤1%) variants contribute to medically relevant complex traits in the general population is largely unknown. Bone mineral density (BMD) is highly heritable and is a major predictor of osteoporotic fractures. Previous efforts have identified BMD-associated common genetic variants[1–8], and rare, population-specific, coding variants[9]. Here we identify novel non-coding genetic variants with large effects on BMD in 27 population-based cohorts of European ancestry (n_{total} = 53,236). This large-scale meta-analysis was derived from whole-genome sequencing from the UK10K consortium (n=2,882), whole-exome sequencing (n= 3,549), deep imputation of genotyped samples using a combined UK10K/1000Genomes reference panel (n=26,534), and *de-novo* replication genotyping (n= 20,271). We identified a low-frequency non-coding variant near *EN1* that was not previously identified, with an effect size 4-fold larger than that of the mean of previously reported common variants for lumbar spine BMD[8] (rs11692564[T], MAF = 1.7%, replication effect size = +0.20 standard deviations [SD], *Pmeta* = 1.7×10^{-14}). Three additional novel variants encompassing *EN1* also achieved genome-wide significance for lumbar spine or femoral neck BMD. rs11692564[T] was also associated with a decreased risk of fracture (OR = 0.85 [95% CI: 0.80-0.89]; P = 2.0×10^{-11}; I^2 = 0.00) in (n_{total} = 508,253 [n_{cases} = 98,742 and $n_{controls}$ = 409,511]). Using an $En1^{Cre/flox}$ mouse model, we observed that conditional loss of *En1* results in low bone mass, likely as a consequence of high bone turn-over. We also identified a novel low-frequency non-coding variant with an effect on forearm BMD which is 2.2-fold larger than that of any previously reported common variants near *WNT16* (rs148771817[T], MAF = 1.1%, replication effect size = +0.39 SD, P_{meta} = 1.1×10^{-11}). Lastly, we observed an excess of association signals arising from deleterious coding and conserved non-coding variants. These findings provide evidence that low-frequency non-coding variants have large effects on BMD and fracture in the general population, thereby providing rationale for whole-genome sequencing-based approaches to study the genetic architecture of complex medically-relevant traits in the general population.

A major goal of the Human Genome Project has been to enable the discovery of genetic determinants of common diseases. This framework has facilitated the identification of thousands of common variants with generally small effects on disease through genome-wide association studies (GWAS)[10]. More recently, discoveries have focused on rare coding variants identified through whole-exome sequencing initiatives[11]. The effect of low-frequency and rare non-coding variants upon common diseases, and their underlying traits has been recently explored in an isolated population[12,13], but has not been well-studied to date in the general population. The UK10K has generated a large whole-genome sequence-based resource to address this question in the general European-ancestry population, which is 10-fold larger than the standard 1000 Genomes Project reference.

Osteoporosis, diagnosed largely through measurement of bone mineral density (BMD), is a common systemic skeletal disease characterized by an increased propensity to fracture[14]. Between one-third to one-half of women will experience an osteoporotic fracture in their lifetime[15] and direct annual costs exceed $19 billion in the United States alone[16]. The narrow-sense heritability of BMD has been estimated to be ~85%, and genome-wide association studies (GWAS) have successfully identified numerous loci associated with BMD which in total explain ~5% of the genetic variance for this trait[17]. However, these studies have been largely unable to assess the role of low frequency (MAF 1-5%) and rare (MAF ≤ 1%) genetic variation, since their methods relied on testing common variants (MAF ≥ 5%). A recent sequencing-based study identified a nonsense variant associated with BMD using 4,931 Icelandic subjects with low BMD and 69,034 population-based controls[9]. This

rare coding variant (MAF 0.1-0.2%), which disrupts the function of *LGR4*, is confined to the Icelandic population.

To investigate the role of rare and low-frequency genetic variation on BMD in a general population of European descent, we first undertook whole genome sequencing in 2,882 subjects from two cohorts in the UK10K project and whole-exome sequencing in 3,549 subjects from five cohorts (**Supplementary Table 1**) with BMD phenotypes. We then used a novel imputation reference panel generated by the UK10K and 1000Genomes consortia to impute variants that were missing, or poorly captured, from previous GWAS studies in 26,534 subjects (**Supplementary Table 1 and Extended Data Fig. 1a**). The UK10K and 1000Genomes reference panel, which in total contained 3,781 and 379 European individuals with whole genome sequences from UK10K and 1000Genomes Projects, respectively enabled improved imputation, particularly of low frequency variants, when compared to the 1000Genomes reference panel alone[18]. We then undertook *de-novo* replication genotyping of lead variants in 13 cohorts for BMD, comprising 20,271 individuals of European descent.

We meta-analyzed association results from all discovery cohorts (n_{total} = 32,965, **Supplementary Table 1**) for BMD measured at the forearm, femoral neck and lumbar spine, the sites where osteoporotic fractures are most prevalent. We tested bi-allelic single-nucleotide variants (SNVs) with MAF ≥ 0.5% for association, declaring genome-wide statistical significance at $P \leq 1.2 \times 10^{-8}$ (accounting for all independent SNVs above this MAF threshold; **Supplementary Methods**)[19]. We found little evidence for genomic inflation (**Supplementary Table 2,** $\lambda = 1.04$, averaged across sites), suggesting that population stratification did not unduly influence results. The

sequence kernel association test (SKAT)[20], a method for region-based analysis of rare genetic variation was used to assess association of regions containing SNVs with MAF ≤5% and ≤1%. Cohort-specific single-SNV and SKAT associations were combined using fixed effects meta-analysis[21] and skatMeta[22], respectively. All summary-level meta-analytic results are available for unrestricted download (www.gefos.org). Genome-wide significant loci were then tested for their relationship with fracture in up to 508,253 individuals. Finally, functional genomics as well as cellular and animal models were utilized to investigate the relevance of these novel genetic associations to bone physiology.

Association signals near Engrailed-1

Through meta-analysis of sequenced and imputed single-SNV association tests from the discovery cohorts (**Supplementary Table 1**), we identified a novel locus at 2q14.2 harboring variants associated with lumbar spine BMD (lead low-frequency SNV rs11692564[T], MAF = 1.7%, effect size = +0.24 SD, $P = 4.1 \times 10^{-9}$, **Fig. 1** and **Table 1**). The direction of effect for rs11692564 was consistent across all discovery cohorts (**Extended Data Fig. 2**) and the mean information score (a measure of imputation quality) for the imputed cohorts was 0.71 (**Supplementary Table 3**). This variant is located 53 kilobase pairs (kb) downstream from engrailed homeobox-1 (*EN1*), which to our knowledge, has not previously been associated with any osteoporosis-related traits in humans The rs11692564 variant was not present on HapMap imputation panels, nor on genotyping chips, underlining the importance of developing more comprehensive imputation reference panels, such as that provided by the combined UK10K/1000Genomes panel used in this study.

To validate whole-genome sequencing genotypes at rs11692564, which were used in the imputation reference panel, we genotyped 1,853 ALSPAC subjects, using the KASP™ genotyping platform, all of whom had been whole-genome sequenced. We found all genotypes to be perfectly concordant (**Supplementary Table 4**). Next we genotyped 245 monozygotic twins from the TwinsUK cohort and found all genotypes to be perfectly concordant with their whole-genome sequenced co-twin (**Supplementary Table 5**). To validate the imputation of rs11692564, we directly genotyped this variant on samples from the TwinsUK imputed cohort (N = 3,601), and observed that the association strengthened and its statistical significance improved for the lumbar spine and femoral neck BMD as compared to imputed results (lumbar spine: imputed effect size = 0.22 SD P = 0.05, genotyped effect size = 0.31 SD P = 0.004) **(Supplementary Table 6**). Next, we compared these KASP™-generated genotypes to a second genotyping platform (Sequenom) to assure accuracy of replication genotypes and observed a concordance rate of 99.4% and non-reference discordant rate of 1.7%. We then sought additional evidence for the association at rs11692564 by performing additional *de novo* genotyping in 16,233 independent individuals and found a similarly large effect size in this population (effect size = +0.20 SD, $P = 2.76 \times 10^{-6}$). Meta-analysis of the discovery and replication cohorts provided strong statistical evidence for association ($P_{combined\text{-}meta} = 1.65 \times 10^{-14}$) (**Table 1**).

We previously reported a common variant, rs1878526 (MAF = 28%), 507kb away from rs11692564 at the *EN1* locus, to be genome-wide significant for lumbar spine BMD[8]. However, these two variants represent independent signals as they exhibit low linkage disequilibrium (LD) ($r^2 < 0.001$) and conditioning on rs1878526 within the

discovery setting did not change the association signal at rs11692564 ($P_{discovery} = 4.1 \times 10^{-9}$ to $P_{conditional} = 1.81 \times 10^{-9}$) (**Supplementary Table 7**).

We also identified an additional association signal, arising from rs55983207 (MAF = 4%), 17 kb downstream of rs11692564 ($r^2 = 0.001$) to be associated with femoral neck BMD from the combined meta-analysis ($P_{meta} = 7.2 \times 10^{-15}$ **Table 1**). A haplotype containing both effect alleles was not observed from within the UK10K whole genome-sequenced cohort (total number of haplotypes = 7,562).

In addition to rs11692564, we also observed from the combined meta-analysis two additional genome-wide significant variants for lumbar spine BMD near *EN1*, rs6542457 (MAF = 6.7%) and rs188303909 (MAF = 1.9%), which are 391kb downstream and 67kb upstream from rs11692564, respectively (**Fig. 1b** and **Table 1**). Variant rs188303909 was in moderate LD with rs11692564 ($r^2 = 0.47$), and conditional analysis demonstrated that these two association signals were not independent (**Supplementary Table 7**). On the other hand, rs6542457 was in low LD with rs11692564 ($r^2 = 0.002$), and in conditional analyses these were found to be independent association signals **(Supplementary Table 7**). Overall, the *EN1* locus harbors multiple non-coding variants associated with lumbar spine and a single variant associated with femoral neck BMD. All three genome-wide significant variants for lumbar spine BMD (**Table 1**) co-localize solely with *EN1* in a sub-region of high interaction frequency within a single topologically-associated domain (TAD)[23] (**Fig. 1a**).

158 The mean effect size of previously reported genome-wide significant SNPs (MAF ≥
159 5%) from the largest GWAS meta-analysis to date for lumbar spine and femoral BMD
160 was 0.048 SD per effect allele and the largest effect size was 0.1 SD[8]. Hence, the
161 observed effect size at rs11692564 is 4-fold larger than this mean and twice that of the
162 largest previously reported effect (**Figure 1c**)[8]. For all genome-wide significant
163 variants, we observed larger effect sizes across decreasing MAF bins (**Fig. 2a**).
164 **Figure 2a** does not aim to describe the full allelic architecture of BMD, since we
165 lacked statistical power to observe small effects arising from low-frequency and rare
166 variants, however, it does demonstrate that variants of large effect size and low MAF
167 do underlie BMD in general European ancestry populations.
168
169 An increase in BMD is associated with a decrease in risk of bone fracture. We
170 therefore tested the association of rs11692564[T] (the low-frequency allele at *EN1*
171 associated with the largest increase in BMD) in 18 cohorts comprising 508,253
172 individuals (98,742 cases and 409,511 controls, **Supplementary Table 8**).
173 rs11692564[T] was strongly associated with a decreased risk of fracture (OR = 0.85
174 [95% CI: 0.80-0.89]; $P = 2.0\times10^{-11}$; $I^2 = 0.00$) (**Table 2** and **Supplementary Table 9**).
175 **Table 2** also shows clear associations between other variants near *EN1* and risk of
176 fracture. The fracture association at rs11692564 was 2.9-fold larger than the mean of
177 fracture associations detected in the largest GWAS to date, and 2.0-fold larger than
178 the largest previously identified fracture association.[8]

179 **Relevance of Engrailed homeobox-1 to bone physiology**
180 *EN1* encodes a homeobox gene central to mouse limb development[24], which has been
181 shown to be involved in Wnt signaling interaction with Dkk1[25]. Studies of calvarial

bone development and fracture healing of long bones in mice have shown that perinatal *En1*$^{-/-}$ mutants display osteopenia and enhanced skull bone resorption[26]. whereas in normal adult mice *En1* is up-regulated in the bone callus post fracture[22]. Investigating the functional role of *EN1*, we detected *En1* expression during osteoblastogenesis in developing and mature cultured murine calvarial osteoblasts, but not in marrow-derived osteoclasts, or in human primary osteoclast cultures (**Figure 3a and Extended Data Fig. 3**). To determine where *En1* is active in adult bones, we analyzed vertebrae from *En1*$^{lacZ/+}$ knock-in mice[27] and detected LacZ expression in proliferative and hypertrophic chondrocytes, osteogenic cells in the periosteum and trabecular bone surface, and in osteocytes of cortical and trabecular bone (**Fig. 3b and Extended Data Fig. 4**).

Using *En1*$^{Cre/+}$; *R26*$^{lox\text{-}STOP\text{-}lox\text{--}EYFP}$ reporter mice to genetically tag cells for which the *En1* promoter was active at any point within a cell lineage, we confirmed that *En1*expression was only observed in osteogenic lineages (**Extended Data Fig. 4**). Since most *En1*$^{-/-}$ animals die soon after birth, we generated *En1*$^{Cre/flox}$ self-deleted *En1* (*sdEn1*) conditional mutants[28] (n = 5) and demonstrated by µCT that mutants have lower trabecular bone volume fraction (BV/TV), trabecular number, and trabecular thickness in both the lumbar L5 vertebrae (**Fig. 3c and 3d and Extended Data Fig. 5**) and the femur (**Extended Fig. 5**) as compared to littermate controls (n = 6). A decrease in femoral cortical thickness was also observed (**Extended Fig. 5**). By histomorphometry (**Fig. 3c**), we observed that the *sdEn1* mice had a statistically higher proportion of osteogenic and osteoclastic cells compared to littermate controls (**Fig. 3d and Supplementary Table 10**). The driving force for the low bone mass would appear to be an increase in osteoclastic activity induced by *En1* null osteogenic cells. This in turn initiates the expected coupled increase in mineralizing bone

formation (**Fig. 3b & 3d**) mediated by an increased number of osteogenic cells and thus conforms to a high turnover osteoporosis-like phenotype, although *dynamic* histomorphometry and evidence from bone turn-over markers would be required to confirm an increased rate of bone formation (**Extended Data Fig. 4**). Lastly, genetic evidence from homologous regions in mice also supported a role for *En1* in bone, as the homologous region contained a QTL peak for femur BMD (**Supplementary Table 11**)[29]. These findings, together with an earlier study focusing on *En1* function in calvarial bone development[26] implicate this gene as an important mediator in skeletal biology.

Taken together, these findings suggest that *EN1* plays an important role in bone physiology and that low-frequency non-coding variants mapping near *EN1* have large effects on BMD and risk of fracture in the general European population.

Association signal at *CPED1/WNT16*

We also identified a novel SNV at 7q31.31 within the intron of *CPED1* (rs148771817[T], MAF = 1.2%, effect size = +0.47 SD, $P_{discovery} = 9.31 \times 10^{-9}$) associated with forearm BMD (**Table 1, Supplementary Table 12, and Extended Data Fig. 6**). We replicated the association at rs148771817 in 2,539 independent individuals and found a similar effect size (effect size = +0.41 SD, $P = 5.5 \times 10^{-4}$), and combined meta-analysis of the discovery and replication cohorts further improved statistical evidence for association (effect = +0.46 SD, $P = 1.1 \times 10^{-11}$) (**Table 1**). This variant had an effect size 2.2-fold larger than the mean of previously reported effect sizes for common variants associated with forearm BMD (**Extended Data Fig. 6**)[30]**.**

We have previously identified BMD-associated genetic variants at *WNT16*, a gene neighboring *CPED1,* (**Extended Data Fig. 6**) and demonstrated that knock-out of *Wnt16* in mice confers a 50% decrease in bone strength ($P = 6.5 \times 10^{-13}$).[30,31] We have recently shown that osteoblast-derived Wnt16 represses osteoclastogenesis[32]. As a result, we undertook conditional analysis of rs148771817 upon the lead common variant for forearm BMD at this locus, rs7776725. The rs148771817 variant remained associated after conditioning, albeit with lower statistical significance (effect size =0.35 SD, $P=1.3 \times 10^{-7}$,**Extended Data Fig. 6d**) upon meta-analysis of the discovery and replication cohorts. Similarly, conditional analysis of the common variant upon rs148771817 revealed little change in the effect size or the statistical significance (**Supplementary Table 7**). We also note that rs148771817 is in low LD with rs7776725 ($r^2 = 0.024$) and acknowledge that both variants may be causal. However, our data does not permit us to distinguish if one or both of these variants have distinct biologic effects.

While rs148771817 is intronic in *CPED1*, we found that DNA accessibility at this region, as measured by DNase I hypersensitivity data from ENCODE, was moderately correlated with DNA accessibility at the *WNT16* promoter in 305 cell types[24] (maximum $r^2= 0.4$, $P= 2.2 \times 10^{-15}$, **Supplementary Table 13**), whereas correlation to the promoter of *CPED1* was lower (maximum $r^2 = 0.1$, $P = 0.06$). Moreover, analysis of chromosome conformation capture Hi-C interaction frequencies from human H1 embryonic stem cells shows elevated interaction frequency between rs148771817 and *WNT16* (**Extended Data Fig. 6**), though we also observed stronger interactions between these loci and their immediate neighboring regions.

Functional enrichment of genetic associations

We assessed whether association signals were enriched for deleterious coding SNVs (defined using Variant Effect Predictor[33] see **Supplementary Methods**) or SNVs with increased GERP++ scores (a measure of evolutionary constraint)[34]. These two groups of SNVs were matched to control SNVs by MAF and distance to gene (**Supplementary Methods and Supplementary Table 14**), followed by LD pruning ($r^2 < 0.2$). We then compared the proportion of deleterious coding and evolutionarily constrained SNVs passing a false-discovery rate (FDR) q-value of 0.05 to the same proportion for their respective control SNVs. We observed enrichment of association signal across the spectrum of positive GERP++ thresholds, which was comparable to deleterious coding variants (**Fig. 2b**). As expected, no such enrichment was seen for synonymous variants (**Fig. 2b and Supplementary Table 15**).

Additional genetic association results

In total, we have identified multiple variants associated with BMD, including 3 genome-wide significant loci for forearm BMD, 14 for femoral neck and 19 for lumbar spine (**Supplementary Tables 12 and 16-18).** A common variant not on previous HapMap imputation panels, near the *SOX6* gene was also identified (rs11024028, MAF = 19%) (**Table 1**), and was found to be an independent signal from a previously reported signal at this locus (rs7108738, $r^2 = 0.002$)[8]. Consistent with recent experience,[35–37] region-based collapsing methods did not identify any convincing novel associations that were not already identified as genome-wide significant through single SNV associations. This included collapsing variants below 1% and 5% MAF thresholds, including all variants, only variants with increased GERP++ scores or those from protein-coding regions (**Supplementary Table 19 and**

Extended Data Figures 7 & 8). Finally, conditioning upon genome-wide significant signals and stratifying the population by sex did not identify novel loci (data not shown).

Conclusions

We have identified low-frequency, non-coding genetic variants of large effect that are present in the general population and associate with BMD and fracture. These variants have effect sizes up to four-fold larger than the mean effect described for common variants associated with BMD and approximately three-fold larger than those for fracture. Our study illustrates that larger reference panels, covering relevant ethnicities, will facilitate the discovery of low frequency and rare variants. This was facilitated here by a large imputation reference panel (UK10K and 1000 Genomes) offering 11-fold more European samples than the 1000 Genomes reference panel alone, resulting in more accurate genotype calls of such variants. Although we did not identify *coding* low-frequency or rare variants associated with BMD at a genome-wide significant level, we did observe that deleterious coding variants were enriched for association as a group. This suggests the existence of as yet undiscovered coding variants influencing BMD. Importantly, we have also generated new functional evidence for a central role of the engrailed homeobox-1 gene in regulation of bone mineral density and outlined *En1* as a critical protein in bone biology. In summary, our findings demonstrate the utility of whole-genome sequencing-based discovery and deep imputation to enable the identification of novel genetic associations. These discoveries provide an improved understanding of the pathophysiology of osteoporosis and suggest that more comprehensive sets of whole-genome sequenced individuals, covering relevant ethnicities, will enable accurate imputation and thus

facilitate discovery of low frequency and rare variants influencing complex traits and common disease.

Figure legends

Figure 1 | Association signals near Engrailed-1 for lumbar spine BMD.

a, Evidence for a topological domain that includes associated variants and *EN1*. Chromatin interaction data from chromosome conformation capture Hi-C performed in H1 embryonic stem (hES) cells[23] of a 2 Mb region encompassing the *EN1* gene reveals that this gene borders a topological associated domain (TAD, grey border). A sub-region of higher chromatin interaction frequencies as well as CTCF chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) performed in MCF-7 cells[38] (annotation track) suggests the existence of a smaller interacting region which contains *EN1*, as well as three genome-wide significant variants for lumbar spine BMD from combined meta-analysis (in red).

b, Association signals at the *EN1* locus. The y-axis denotes the $-\log_{10}$ association P-value from discovery meta-analysis (lines at genome-wide suggestive [$P = 1.2 \times 10^{-5}$] and genome-wide significant [$P = 1.2 \times 10^{-8}$] thresholds, respectively). The P-value for variants that underwent replication via *de novo* genotyping is represented as a circle for the discovery P-value and a triangle for the combined discovery and replication meta-analytic P-value.

c, Allele frequency versus absolute effect size (in standard deviations) for lumbar spine BMD of all previously identified genome-wide significant variants (blue)[8] and three novel variants in *EN1* (red), rs11692564, rs188303909, and rs6542457. The blue line denotes the mean of effect sizes for previously reported variants. Note that **Table**

1 also demonstrates genome-wide significant findings for variants near *ENI1* with femoral neck BMD.

Figure 2 | Genome-wide features of association signals.

a, Relationship between MAF and observed effect sizes in BMD discovery cohorts. Box plots of the effect sizes of genome-wide significant SNVs ($P < 1.2x10^{-8}$), pruned for LD ($r^2 < 0.2$) by MAF bin. Grey bars represent the values of beta not observed and for which we lack statistical power to observe (at $\alpha \leq 1.2x10^{-8}$ and power ≥ 0.8). No SNVs achieved genome-wide significance for MAF bin 0.5 - 1% for femoral neck and lumbar spine. *P*-values displayed per phenotype are calculated from the non-parametric trend test across MAF bins.
b, Proportion of SNVs passing an FDR q-value 0.05 across different types of annotation features for all three BMD sites in discovery cohorts. Left-most panel shows a lack of enrichment of signal from synonymous SNVs (green circles) compared to matched control variants (red circles), as expected. Second panel shows enrichment of signal from deleterious (green) SNVs (defined using Variant Effect Predictor on UK10K variants as frameshift, inframe deletions, inframe insertions, initiator codon, missense, splice-acceptor, splice-donor, stop-gain and stop-loss variants) compared to all other variants (red). Remaining panels show enrichment of signal across a range of GERP scores for each BMD site, where green denotes SNVs above the threshold and red denotes variants below the threshold. After LD pruning (r^2>0.2), variants above the GERP threshold were matched (**Supplementary Table 14**) to variants below the threshold by distance to gene and MAF. Data is presented in table format in **Supplementary Table 15**.

Figure 3 | Mouse *En1* Functional Experiments.

a, Quantitative expression of *En1* and its temporal pattern through RNA-seq in cultured calvarial murine osteoblasts across day 2 through day 18 of osteoblast development (top panel). Confirmation of the expression of *En1* in a separate RT-PCR experiment of cultured calvarial murine osteoblasts. Expression was not detected in osteoclasts matured from bone marrow derived precursor cells (Positive controls for osteocalcin in osteoblasts and RANK in osteoclasts are also shown) (bottom panel).

b, Representative sections from lumbar vertebra 2 (LV2) show the growth plate and bone marrow (GP and BM, left), cortical bone (CB, middle), and trabecular bone (TB, right) at 40x magnification from $En1^{lacZ/+}$ adult mice (n = 2) stained for β-gal activity (LacZ blue, *En1*+ cells) and alkaline phosphatase (AP, red late chondrocytes and actively calcifying tissues). In the periosteum (PO), all the LacZ+ cells were AP+; some AP- BM cells expressed LacZ. Some AP- proliferative chondrocytes in the GP expressed lacZ+, whereas most AP+ hypertrophic chondrocytes expressed LacZ (see also Extended Data Figure 12). Some AP- osteocytes (Ocy) in CB and TB were LacZ+.

c, Histomorphometry images of lumbar vertebrae 5 (LV5) demonstrate decreased trabecular bone volume and increased bone surface area occupied by osteoclast cells (as assessed by TRAP staining) when comparing $En1^{Cre/flox}$ (self-deleted *En1*, *sdEn1*) mutants and $En1^{flox/+}$ control mice (left panels). Thickness map of the 3D reconstructed μCT images of the L5 shows the spatial distribution of the mineral density in a control and an *sdEn1* animal (right panels).

d, Micro-CT (μCT) and histomorphometry measures within *sdEn1 (n = 5)* and controls ($En1^{lox/+}$, n = 6). By μCT, *sdEn1* mutants exhibit decreased L5 trabecular

number (Tb.N) and thickness (Tb.Th), as well as deceased bone volume fraction (BV/TV). Using histomorphometry, *sdEn1* mutants exhibit increased osteoclastic area (TRAP/BS).

References

1. Richards, J. *et al.* Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet* **371,** 1505–1512 (2008).
2. Styrkarsdottir, U. *et al.* Multiple genetic loci for bone mineral density and fractures. *N. Engl. J. Med.* **358,** 2355–2365 (2008).
3. Styrkarsdottir, U. *et al.* New sequence variants associated with bone mineral density. *Nat. Genet.* **41,** 15–17 (2008).
4. Rivadeneira, F. *et al.* Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat. Genet.* **41,** 1199–1206 (2009).
5. Duncan, E. L. *et al.* Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. *PLoS Genet.* **7,** e1001372 (2011).
6. Koller, D. L. *et al.* Genome-wide association study of bone mineral density in premenopausal European-American women and replication in African-American women. *J. Clin. Endocrinol. Metab.* **95,** 1802–1809 (2010).
7. Xiong, D.-H. H. *et al.* Genome-wide Association and Follow-Up Replication Studies Identified ADAMTS18 and TGFBR3 as Bone Mass Candidate Genes in Different Ethnic Groups. *Am. J. Hum. Genet.* **84,** 388–398 (2009).
8. Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* 1–13 (2012). doi:10.1038/ng.2249
9. Styrkarsdottir, U. *et al.* Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. *Nature* **497,** 517–20 (2013).

10. Hindorff, L. a *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–9367 (2009).

11. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nature Genetics* **44,** 623–630 (2012).

12. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* (2015). doi:10.1038/ng.3247

13. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nat. Genet.* (2015). doi:10.1038/ng.3243

14. Kanis, J. A. Diagnosis of osteoporosis. *Osteoporos. Int.* **7 Suppl 3,** S108–16 (1997).

15. Kanis, J. A. *et al.* Long-term risk of osteoporotic fracture in Malmo. *Osteoporos. Int.* **11,** 669–674 (2000).

16. Burge, R. *et al.* Incidence and economic burden of osteoporosis-related fractures in the United States, 2005-2025. *J. Bone Miner. Res.* **22,** 465–475 (2007).

17. Richards, J. B., Zheng, H.-F. & Spector, T. D. Genetics of osteoporosis from genome-wide association studies: advances and challenges. *Nat. Rev. Genet.* **13,** 672–672 (2012).

18. Huang, J., Howie, B. & Marchini, J. Imputation Performance of 4,000 genomes from the UK10K Project. in *American Society of Human Genetics Annual Meeting* (2013).

19. Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.* **38,** 281–290 (2014).

20. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* **89,** 82–93 (2011).

21. Magi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11,** 288 (2010).

22. Voorman, A. A., Brody, J. & Lumley, T. Package ‘ skatMeta ’. (2013).

23. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

24. Loomis, C. A. *et al.* The mouse Engrailed-1 gene and ventral limb patterning. *Nature* **382,** 360–363 (1996).

25. Adamska, M., MacDonald, B. T., Sarmast, Z. H., Oliver, E. R. & Meisler, M. H. En1 and Wnt7a interact with Dkk1 during limb development in the mouse. *Dev. Biol.* **272,** 134–144 (2004).

26. Deckelbaum, R. a, Majithia, A., Booker, T., Henderson, J. E. & Loomis, C. a. The homeoprotein engrailed 1 has pleiotropic functions in calvarial intramembranous bone formation and remodeling. *Development* **133,** 63–74 (2006).

27. Matise, M. P. & Joyner, a L. Expression patterns of developmental control genes in normal and Engrailed-1 mutant mouse spinal cord reveal early diversity in developing interneurons. *J. Neurosci.* **17,** 7805–7816 (1997).

28. Sgaier, S. K. *et al.* Genetic subdivision of the tectum and cerebellum into functionally related regions based on differential sensitivity to engrailed proteins. *Development* **134,** 2325–2335 (2007).

29. Ackert-Bicknell, C. L. *et al.* Mouse BMD quantitative trait loci show improved concordance with human genome-wide association loci when recalculated on a new, common mouse genetic map. *J. Bone Miner. Res.* **25,** 1808–1820 (2010).

30. Zheng, H.-F. F. *et al.* WNT16 Influences Bone Mineral Density, Cortical Bone Thickness, Bone Strength, and Osteoporotic Fracture Risk. *PLoS Genet.* **8,** e1002745 (2012).

31. Medina-Gomez, C. *et al.* Meta-Analysis of Genome-Wide Scans for Total Body BMD in Children and Adults Reveals Allelic Heterogeneity and Age-Specific Effects at the WNT16 Locus. *PLoS Genet.* **8,** e1002718 (2012).

32. Movérare-Skrtic, S. *et al.* Osteoblast-derived WNT16 represses osteoclastogenesis and prevents cortical bone fragility fractures. *Nat. Med.* **20,** 1279–88 (2014).

33. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).

34. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (2010).

35. Ladouceur, M., Dastani, Z., Aulchenko, Y. S., Greenwood, C. M. T. & Richards, J. B. The Empirical Power of Rare Variant Association Methods: Results from Sanger Sequencing in 1,998 Individuals. *PLoS Genet.* **8,** e1002496 (2012).

36. Morrison, A. C. *et al.* Whole-genome sequence–based analysis of high-density lipoprotein cholesterol. *Nat. Genet.* **45,** 899–901 (2013).

37. Tang, H. *et al.* A large-scale screen for coding variants predisposing to psoriasis. *Nat. Genet.* **46,** 45–50 (2014).

38. Li, G. *et al.* Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* **148,** 84–98 (2012).