Zengini, E., Hatzikotoulas, K., Tachmazidou, I., Steinberg, J., Hartwig, F. P., Southam, L., Hackinger, S., Boer, C. G., Styrkarsdottir, U., Gilly, A., Suveges, D., Killian, B., Ingvarsson, T., Jonsson, H., Babis, G. C., McCaskie, A., Uitterlinden, A. G., Van Meurs, J. B. J., Thorsteinsdottir, U., ... Zeggini, E. (2018). Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis. *Nature Genetics*, *50*(4), 549-558. https://doi.org/10.1038/s41588-018-0079-y

Peer reviewed version

Link to published version (if available):
10.1038/s41588-018-0079-y

Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis

Eleni Zengini[1,2*], Konstantinos Hatzikotoulas[3*], Ioanna Tachmazidou[3,4*], Julia Steinberg[3,5], Fernando P. Hartwig[6,7], Lorraine Southam[3,8], Sophie Hackinger[3], Cindy G. Boer[9], Unnur Styrkarsdottir[10], Arthur Gilly[3], Daniel Suveges[3], Britt Killian[3], Thorvaldur Ingvarsson[11,12,13], Helgi Jonsson[12,14], George C. Babis[15], Andrew McCaskie[16], Andre G. Uitterlinden[9], Joyce B. J. van Meurs[9], Unnur Thorsteinsdottir[10,12], Kari Stefansson[10,12], George Davey Smith[7,17,18], Jeremy M. Wilkinson[1], Eleftheria Zeggini[3#]

1. Department of Oncology and Metabolism, University of Sheffield, Sheffield S10 2RX, United Kingdom
2. 5th Psychiatric Department, Dromokaiteio Psychiatric Hospital, Athens 124 61, Greece
3. Human Genetics, Wellcome Trust Sanger Institute, Hinxton CB10 1HH, United Kingdom
4. GSK, R&D Target Sciences, Medicines Research Centre, Stevenage SG1 2NY, United Kingdom
5. Cancer Research Division, Cancer Council NSW, Sydney NSW 2011, Australia
6. Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas 96020-220, Brazil
7. Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol BS8 2BN, United Kingdom
8. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom
9. Department of Internal Medicine, Erasmus MC, Rotterdam, Netherlands
10. deCODE genetics/Amgen, Reykjavik 101, Iceland
11. Department of Orthopaedic Surgery, Akureyri Hospital, Akureyri 600, Iceland
12. Faculty of Medicine, University of Iceland, Reykjavik 101, Iceland
13. Institution of Health Science, University of Akureyri, Akureyri 600.0, Iceland
14. Department of Medicine, Landspitali, The National University Hospital of Iceland, Reykjavik 101, Iceland
15. National and Kapodistrian University of Athens, 2nd Department of Orthopaedic Surgery Konstantopouleio General Hospital, Athens, Greece
16. Division of Trauma & Orthopaedic Surgery; Department of Surgery; University of Cambridge; Box 180, Addenbrooke's Hospital, Cambridge CB2 0QQ, United Kingdom
17. School of Social and Community Medicine, University of Bristol, Bristol BS8 2BN, United Kingdom
18. National Institute for Health Research Bristol Biomedical Research Centre, University Hospitals Bristol NHS Foundation Trust and University of Bristol

*These authors contributed equally. #Correspondence to Eleftheria Zeggini: Eleftheria@sanger.ac.uk

Osteoarthritis is a common complex disease with huge public health burden. Here we perform a genome-wide association study for osteoarthritis using data across 16.5 million variants from the UK Biobank resource. Following replication and meta-analysis in up to 30,727 cases and 297,191 controls, we report 9 new osteoarthritis loci, in all of which the most likely causal variant is non-coding. For three loci, we detect association with biologically-relevant radiographic endophenotypes, and in five signals we identify genes that are differentially expressed in degraded compared to intact articular cartilage from osteoarthritis patients. We establish causal effects for higher body mass index, but not for triglyceride levels or genetic predisposition to type 2 diabetes, on osteoarthritis.

INTRODUCTION

Osteoarthritis is the most prevalent musculoskeletal disease and the most common form of arthritis[1]. The hallmarks of osteoarthritis are degeneration of articular cartilage, remodelling

of the underlying bone and synovitis[2]. A leading cause of disability worldwide, it affects 40% of individuals over the age of 70 and is associated with an increased risk of comorbidity and death[3]. The health economic burden of osteoarthritis is rising, commensurate with longevity and obesity rates, and there is currently no curative therapy. The heritability of osteoarthritis is ~50%, and previous genetic studies have identified 21 loci in total, traversing hip, knee and hand osteoarthritis with limited overlap[3]. Here we conduct the largest osteoarthritis genome wide association study (GWAS) to date, using genotype data across 16.5 million variants from UK Biobank. We define osteoarthritis based on both self-reported status and through linkage to Hospital Episode Statistics data, and on joint-specificity of disease (knee and/or hip) (Supplementary Fig. 1).

RESULTS

Disease definition and power to detect genetic associations

We compare and contrast the hospital diagnosed (n=10,083 cases) to self-reported (n=12,658 cases) osteoarthritis GWAS drawn from the same UK Biobank dataset (with non-osteoarthritis controls selected to be ~4x the number of cases to preserve power for common alleles while avoiding case:control imbalance causing association tests to misbehave for low frequency variants[4]) (Supplementary Tables 1-3, Supplementary Figs. 2-4; Methods). We find power advantages with the self-reported dataset, indicating that the increase in sample size overcomes the limitations associated with phenotype uncertainty. When evaluating the accuracy of disease definition, we find that self-reported osteoarthritis has modest positive predictive value (PPV=30%) and sensitivity (37%), but a high negative predictive value of 95% and high specificity, correctly identifying 93% of individuals who do not have osteoarthritis (Supplementary Table 4). In terms of power to detect genetic associations, the self-reported osteoarthritis dataset has clear advantages commensurate with its larger samples size (Figure 1). For example, for a representative complex disease-associated variant with minor allele frequency (MAF) 30% and allelic odds ratio 1.10, the self-reported and hospital diagnosed osteoarthritis analyses have 80% and 56% power to detect an effect at genome-wide significance (i.e., $P<5.0\times10^{-8}$), respectively (Supplementary Table 5).

We find nominally significant evidence for concordance between the direction of effect at previously reported osteoarthritis loci and the discovery analyses for hospital diagnosed osteoarthritis definitions (Supplementary Table 6, Supplementary Table 7, Supplementary note), indicating that a narrower definition of disease may provide better effect size estimates albeit limited by power to identify robust statistical evidence for association.

Heritability estimates across osteoarthritis definitions

We find that common-frequency variants explain 12% of osteoarthritis heritability when using self-reported status, and 16% when using hospital records (19% of hip osteoarthritis and 15% of knee osteoarthritis heritability) (Supplementary Table 8). Heritability estimates from self-reported and hospital records were not significantly different from each other (Supplementary Table 9). The concordance between self-reported and hospital diagnosed osteoarthritis was further substantiated by the high genetic correlation estimate of the two disease definitions (87%, $P=3.14\times10^{-53}$) (Supplementary Table 10). We find strong genome-wide correlation between hip osteoarthritis and knee osteoarthritis (88%, $P=1.96\times10^{-6}$), even though previously reported osteoarthritis loci are predominantly not shared between the two osteoarthritis joint sites. Based on this new observation of a substantial shared genetic aetiology, we sought replication of association signals across joint sites.

Identification of novel osteoarthritis loci

We took 173 variants with $P<1.0\times10^{-5}$ and MAF>0.01 forward to replication in an Icelandic cohort of up to 18,069 cases and 246,293 controls (Supplementary Fig. 1, Supplementary Tables 11-15; Methods). Given the number of variants, the replication significance threshold was $P<2.9\times10^{-4}$. Following meta-analysis in up to 30,727 cases and 297,191 controls, we

report six genome-wide significant associations at novel loci, and three further replicating signals just below the corrected genome-wide significance threshold (Table 1, Figure 2).

We identify association between rs2521349 and hip osteoarthritis (OR 1.13 (95% CI 1.09-1.17), P= 9.95x10^{-10}, effect allele frequency [EAF] 0.37). rs2521349 resides in an intron of *MAP2K6* on chromosome 17. *MAP2K6* is an essential component of the p38 MAP kinase mediated signal transduction pathway, involved in various cellular processes in bone, muscle, fat tissue homeostasis and differentiation[5]. The MAPK signalling pathway has been closely associated with osteoblast differentiation[6], chondrocyte apoptosis and necrosis[7], and reported to be differentially expressed in osteoarthritis synovial tissue samples[6-12]. In animal model studies, its activity has been found to be important in maintaining cartilage health and it has been proposed as a potential osteoarthritis diagnosis and treatment target[10,13,14].

rs11780978 on chromosome 8 is also associated with hip osteoarthritis with a similar effect size (OR 1.13 (95% CI 1.08-1.17), p=1.98x10^{-9}, EAF 0.39). This variant is located in the intronic region of the plectin gene, *PLEC*. We find rs11780978 to be nominally associated with the radiographically derived endophenotype of minimal joint space width (beta -0.0291, SE 0.0129, P=0.024) (Table 2; Methods). The direction of effect is consistent with the established clinical association between joint space narrowing and osteoarthritis. *PLEC* encodes plectin, a structural protein which interlinks components of the cytoskeleton. Functional studies in mice have shown an effect on skeletal muscle tissue and correlation with decreased body weight, size and postnatal growth[15].

rs2820436 is an intergenic variant located 24kb upstream of long non-coding RNA *RP11-392O17.1* and 142kb downstream of zinc finger CCCH-type containing 11B pseudogene *ZC3H11B*, and is associated with osteoarthritis across any joint site (OR 0.93 (95% CI 0.91-0.95), P=2.01x10^{-9}, EAF 0.65). It also resides within a region with multiple metabolic and anthropometric trait-associated variants, with which it is correlated (r^2 0.18-0.88).

rs375575359 resides in an intron of the zinc finger protein 345 gene, *ZNF345,* on chromosome 19. It was prioritised based on osteoarthritis at any joint site and is more strongly associated with knee osteoarthritis in the replication dataset (OR 1.21 (95% CI 1.14-1.30), P=7.54x10^{-9}, EAF 0.04). Similarly, rs11335718 on chromosome 4 was associated with osteoarthritis in the discovery and knee osteoarthritis in the replication stage (OR 1.11 (95% CI 1.07-1.16), P= 4.26x10^{-8}, EAF 0.10). We note that Bonferroni correction for the effective number of traits tested would mean that rs11335718 no longer reaches genome-wide significance with a meta-analysis P=4.26x10^{-8}. rs11335718 is an intronic variant in the annexin A3 gene, *ANXA3*. By meta-analysing the any site osteoarthritis phenotype across the discovery and replication datasets, we report P=2.6x10^{-5} and P=1.32x10^{-7} for rs375575359 and rs11335718, respectively (Supplementary Table 11). A recent mouse model study supports the involvement of a similar motif zinc finger protein expression (ZFP36L1) with osteoblastic differentiation[16].

rs3771501 (OR 0.94 (95% CI 0.92-0.96), P=1.66x10^{-8}, EAF 0.53) is associated with osteoarthritis at any site and resides in an intron of the transforming growth factor alpha gene, *TGFA*. *TGFA* encodes an epidermal growth factor receptor ligand and is an important integrator of cellular signalling and function. We detect association of rs3771501 with minimal joint space width (beta -0.0699, SE 0.0127, P=3.45x10^{-8}) (Table 2; Methods), i.e. the osteoarthritis risk increasing allele is also associated with decreased joint cartilage thickness in humans. A perfectly correlated variant in this gene has previously been associated with cartilage thickness, suggestively associated with hip osteoarthritis, and found to be differentially expressed in osteoarthritis cartilage lesions compared to non-lesioned cartilage[17]. Functional studies have revealed that TGFA regulates the conversion of cartilage to bone during the process of endochondral bone growth, that it is a dysregulated cytokine present in degrading cartilage in osteoarthritis and a strong stimulator of cartilage

degradation upregulated by articular chondrocytes in experimentally induced and human osteoarthritis[18-21]. The function of *TGFA* has also been associated with craniofacial development, palate closure and decreased body size[22].

rs864839 resides in the intronic region of the junctophilin 3 gene, *JPH3*, on chromosome 16, and was discovered based on the any joint site osteoarthritis analysis. It is more strongly associated with hip osteoarthritis in the replication dataset (OR 1.08 (95% CI 1.05-1.11), P=2.1x10^{-8}, EAF 0.71). By meta-analysing the any site osteoarthritis phenotype across the discovery and replication datasets, we report P=7.02x10^{-6} (Supplementary Table 11). JPH3 is involved in the formation of junctional membrane structure, regulates neuronal calcium flux and is reported to be expressed in pancreatic beta cells and in the regulation of insulin secretion.

rs116882138 is most strongly associated with hip and/or knee osteoarthritis in the discovery and with knee osteoarthritis in the replication dataset (OR 1.34 (95% CI 1.21-1.49), P=5.09x10^{-8}, EAF 0.02). It is an intergenic variant located 11kb downstream of the kinase activator 3B gene, *MOB3B*, and 16kb upstream of the equatorin, sperm acrosome associated gene, *EQTN*, on chromosome 9. We find rs116882138 to be nominally associated with acetabular dysplasia as measured with Center Edge-angle (beta -1.1388, SE 0.5276, P=0.031) (Table 2; Methods).

Finally, rs6516886 was prioritised based on the hip and/or knee osteoarthritis discovery analysis and is more strongly associated in the hip osteoarthritis replication dataset (OR 1.10 (95% CI 1.06-1.14), p=5.84x10^{-8}, EAF 0.75). rs6516886 is situated 1kb upstream of the RWD domain containing 2B gene, *RWDD2B*, on chromosome 21. *LTN1* (listerin E3 ubiquitin protein ligase 1), which resides at a distance of 28kb from the variant, has been reported to affect musculoskeletal development in a mouse model[23].

Functional analysis

We tested whether coding genes within 1Mb surrounding the novel osteoarthritis-associated variants were differentially expressed at 1% false discovery rate (FDR) in chondrocytes extracted from intact compared to degraded cartilage from osteoarthritis patients undergoing total joint replacement surgery using molecular phenotyping through quantitative proteomics and RNA sequencing (Table 3; Methods).

PCYOX1, located 209kb downstream of rs3771501, showed significant evidence of differential expression (1.21-fold higher post-normalisation in degraded cartilage at the RNA level, *q-value*=0.0047; and 1.17-fold lower abundance at the protein level, q=0.0042). This discrepancy could signal potential clinical relevance, as the gene product constitutes a candidate biomarker for osteoarthritis progression. Prenylcysteine oxidase 1, the protein product of this gene, catalyses the degradation of prenylated proteins[24], is a secreted protein, and has been identified in urinary exosomes[25]. Further investigation into the chondrocyte and peripheral secretome is warranted to assess the potential of this molecule as a biomarker for osteoarthritis progression. *PCYOX1* has been reported to be overexpressed in human dental pulp derived osteoblasts compared to osteosarcoma cells[26]. *FAM136A*, located 188kb upstream of the same variant (rs3771501), showed 1.13-fold lower transcriptional levels in chondrocytes from degraded articular cartilage (q=0.0066).

BACH1 and *MAP3K7CL*, located in the vicinity of rs6516886, showed evidence of differential transcription (1.26-fold, q=0.0019, and 1.37-fold, q=0.0021, higher transcription in degraded tissue, respectively). BACH1 is a transcriptional repressor of Heme oxygenase-1 (HO-1). Studies in Bach1 deficient mice have independently suggested inactivation of Bach1 as a novel target for the prevention and treatment of meniscal degeneration[27] and of osteoarthritis[28].

Finally, *PLAA* and *ZNF382* located proximal to rs116882138 and rs375575359, respectively, showed higher transcription levels in degraded compared to intact cartilage (1.15-fold, q=0.0027, and 1.31-fold, q=0.0031, respectively). *BOP1* located 451kb downstream of rs11780978 showed 1.17-fold lower levels of transcription in degraded tissue (q=0.003).

We examined evidence for expression quantitative trait loci (eQTLs) in GTEx tissues and found that none of the eQTLs identified at 5%FDR overlapped with the genes identified as differentially expressed between osteoarthritis intact and degraded cartilage (Supplementary Note; Supplementary Table 16).

Fine-mapping points to non-coding variants at all loci

For five of the new loci, the sum of probabilities of causality of all variants in the fine-mapped region was ≥0.95 (>0.99 for two signals), and for two further loci it was >0.93 (Supplementary Table 17; Methods). The majority of variants within each credible set have marginal posterior probabilities, while only a small number of variants have posterior probability of association (PPA) >0.1; these account for 25-92% of PPA across the different regions. The credible set of four signals can be narrowed down to 3 variants, one signal to 2 variants, and one signal to 1 variant with a probability of causality >0.1. For all 9 regions the variant identified as the most likely to be causal is non-coding (Supplementary Table 18, Supplementary note, Supplementary Fig. 5).

Gene-based analyses

Gene set analysis identified *UQCC1* and *GDF5*, located in close vicinity to each other on chromosome 20, as key genes with consistent evidence for significant association with osteoarthritis across phenotype definitions (Supplementary Table 19, Supplementary note). *UQCC1* and *GDF5* were significantly associated with four and three of the five osteoarthritis definitions, respectively. *GDF5* codes for growth differentiation factor 5, a member of the TGFbeta superfamily, and there is accruing evidence that it play a central role in skeletal health and development[29-32]. Pathway analyses identified significant associations between self-reported osteoarthritis and anatomical structure morphogenesis (P=4.76x10^{-5}), ion channel transport (P=8.98x10^{-5}); hospital diagnosed hip osteoarthritis and activation of MAPK activity (P=1.61x10^{-5}); hospital diagnosed knee osteoarthritis and histidine metabolism (P=1.02x10^{-5}); and between hospital diagnosed hip and/or knee osteoarthritis and recruitment of mitotic centrosome proteins and complexes (P=8.88x10^{-5}) (Supplementary Table 20, Supplementary Fig. 6).

Genetic links between osteoarthritis and other traits

Established clinical risk factors for osteoarthritis include increasing age, female sex, obesity, occupational exposure to high levels of joint loading activity, previous injury, smoking status and family history of osteoarthritis. We estimated the genome-wide genetic correlation between osteoarthritis and 219 other traits and diseases and identified 35 phenotypes with significant (5% FDR) genetic correlation with osteoarthritis across definitions, with large overlap between the identified phenotypes (Supplementary Fig. 7, Figure 3, Supplementary Table 21; Methods).

The phenotypes with significant genetic correlations (rg) fall into the following broad categories: obesity, BMI and related anthropometric traits (rg>0); type 2 diabetes (rg>0); educational achievement (rg<0); neuroticism, depressive symptoms (rg>0), and sleep duration (rg<0); mother's, father's, or parents' age at death (rg<0); reproductive phenotypes, including age at first birth (rg<0) and number of children ever born (rg>0); smoking, including age of smoking initiation (rg<0) and ever smoker (rg>0), and lung cancer (rg>0) (Figure 3, Supplementary Table 21). The four phenotypes with significant genetic correlation in all analyses were: years of schooling, waist circumference, hip circumference and BMI.

We find a nominally significant positive genetic correlation with rheumatoid arthritis, which did not pass multiple-testing correction for self-reported and hospital diagnosed osteoarthritis (rg=0.14-0.19, FDR 10%-12%). Of musculoskeletal phenotypes, lumbar spine bone mineral density showed positive genetic correlation with hospital diagnosed hip and/or knee osteoarthritis (rg=0.2, FDR=3%) but did not reach significance in other analyses.

Disentangling causality

We undertook Mendelian randomisation (MR) analyses[33] to strengthen causal inference regarding modifiable exposures that could influence osteoarthritis risk (Supplementary Tables 22-25; Methods). Each kg/m^2 increment in body mass index was predicted to increase risk of self-reported osteoarthritis by 1.11 (95% CI: 1.07-1.15, P= 8.3x10^{-7}). This result was consistent across MR methods (OR ranging from 1.52 to 1.66) and disease definition (OR ranging from 1.66 to 2.01). Consistent results were also observed for other obesity-related measures, such as waist (OR: 1.03 per cm increment; 95% CI: 1.02-1.05, P=5x10^{-4}) and hip circumference (OR: 1.03 per cm increment; 95% CI: 1.01-1.06, P=0.021). OR for type 2 diabetes liability and triglycerides were consistently small in magnitude across estimators and osteoarthritis definitions; given that analyses involving those traits were well-powered (Supplementary Table 26), these results are compatible with either a weak or no causal effect. Results for years of schooling were not consistent across estimators, and there was evidence of directional horizontal pleiotropy, thus hampering any causal interpretation (Figure 4). For lumbar spine bone mineral density, there was evidence of a causal effect with OR per standard deviation increment of 1.28 (95% CI: 1.11-1.47, P=0.002) for hip and/or knee osteoarthritis. This effect appeared to be site-specific, with OR of 1.29 (95% CI: 1.06-1.57, P=0.014) for knee osteoarthritis, while OR for hip osteoarthritis ranged from 0.71 to 1.57. There was also some evidence for a site-specific causal effect of height on knee osteoarthritis (OR: 1.13 per standard deviation increment; 95% CI: 1.02-1.25, P=0.023), which was consistent across estimators. One-sample MR analyses corroborated these findings, with obesity-related phenotypes presenting strong statistical evidence after multiple testing correction (Supplementary Table 27). These analyses did not detect reliable effects of smoking or reproductive traits on osteoarthritis (Supplementary Tables 28 and 29).

DISCUSSION

In order to improve our understanding of the genetic aetiology of osteoarthritis, we have conducted a study combining genotype data in up to 327,918 individuals. We identify 6 novel, robustly replicating loci associated with osteoarthritis, and three which fall just under the corrected genome-wide significance threshold; this constitutes a substantial increase in the number of known osteoarthritis loci. Taken together, all established osteoarthritis loci account for 26.3% of trait variance (Supplementary Fig. 8). The key attributes of this study are sample size and the homogeneity of the UK Biobank dataset, coupled with independent replication, independent association with clinically-relevant radiographic endophenotypes, and functional genomics follow up in primary osteoarthritis tissue. We have further capitalised on the wealth of available genome-wide summary statistics across complex traits to identify genetic correlations between osteoarthritis and multiple molecular, physiological and behavioural phenotypes, followed by formal Mendelian randomization analyses to assess causality and disentangle complex cross-trait epidemiological relationships.

The vast majority of novel signals are at common frequency variants and confer small to modest effects, in line with a highly polygenic model underpinning osteoarthritis risk. We identify one low-frequency variant association with osteoarthritis (MAF 0.02) with a modest effect size (combined OR 1.34). Even though well-powered to detect them, we find no evidence for a role of low frequency variation of large effect in osteoarthritis susceptibility (Supplementary Table 5). The power of this study is very limited for low frequency variants with OR<1.50, and for rare variants. We estimate the requirement of up to 40,000 osteoarthritis cases and 160,000 controls to recapitulate the effects identified in this study at

genome-wide significance based on the sample size-weighted effect allele frequencies and replication cohort odds ratio estimates (Table 1, Supplementary Table 30).

We integrated functional information with statistical evidence for association to fine-map the location of likely causal variants and genes. All predicted most likely causal variants reside in non-coding sequence: 6 are intronic and 3 are intergenic. We are able to refine the association signal to a single variant in one occasion, and to variants residing within a single gene in three instances, although the mechanism of action could be mediated through other genes in the vicinity.

We empirically find self-reported osteoarthritis definition to be a powerful tool for genetic association studies, for example as evidenced by the fact that the established *GDF5* osteoarthritis locus reaches genome-wide significance in the self-reported disease status analyses only. Published epidemiological studies investigating osteoarthritis via self-report[34,35] and validation of self-reported status against primary care records has yielded similar conclusions[34]. We also find very high genetic correlation between self-reported and hospital diagnosed osteoarthritis, and similar variant-based heritability estimates, corroborating the validity of self-reported osteoarthritis status for genetic studies. However, we also note that the hospital diagnosed osteoarthritis analyses have higher heritability and yield stronger evidence for effect direction concordance at established loci, indicating that larger sample size would afford the power required to convincingly detect them. Hospital diagnosed data for osteoarthritis potentially capture a different patient demographic compared to self-reported data (Supplementary note). Based on the results of this study, we deduce that there is no gold standard for osteoarthritis definition in genetics studies, and identify advantages in employing both methods of defining disease to maximize discovery power across the board.

We identify strong genome-wide correlation between hip and knee osteoarthritis, indicating a substantial shared genetic aetiology that has been hitherto missed. We therefore sought replication of signals across these highly correlated phenotypes and identify multiple instances of signals detected in the larger discovery analysis of osteoarthritis, and independently replicated in joint-specific definitions of disease. Indeed, when looking at the replication phenotypes, we find no instances of confirmed replication where the replication phenotype is not captured within the accompanying discovery phenotype definition. Further analysis in larger sample sets with precise phenotyping will help distinguish signal specificity.

Two of the newly identified signals, indexed by rs11780978 and rs2820436, reside in regions with established metabolic and anthropometric trait associations. osteoarthritis is epidemiologically associated with increased BMI, and the association is stronger for knee osteoarthritis. In line with this, we find higher genetic correlation between BMI and knee osteoarthritis (rg=0.52, P=2.2x10^{-11}), than with hip osteoarthritis (rg=0.28, P=4x10^{-4}). BMI is also known to be genetically correlated with education phenotypes, depressive symptoms, reproductive and other phenotypes; hence, some of the genetic correlations for osteoarthritis observed here could be mediated through BMI. However, for the education and personality/psychiatric phenotypes, the strength of genetic correlations observed here for osteoarthritis is substantially higher than the genetic correlations observed for BMI (e.g. hospital diagnosed osteoarthritis and years of schooling rg=-0.45, P=5x10^{-27}, while BMI and years of schooling have rg=-0.27, P=9x10^{-32}; hospital diagnosed osteoarthritis and depressive symptoms rg=0.49, P=6x10^{-7}, while BMI and depressive symptoms have rg=0.10, P=0.023). Epidemiologically, lower educational levels are known to be associated particularly with risk of knee osteoarthritis, even when adjusting for BMI[36].

Mendelian randomization provided further insight into the nature of the genetic correlations we observed. In the case of BMI and other obesity-related measures, there was evidence for a causal effect of those phenotypes on osteoarthritis. This result corroborates the findings

from conventional observational studies[37], which are prone to important limitations (such as reverse causation and residual confounding) regarding causal inference[38]. For all other exposure phenotypes, there was no convincing evidence for a causal effect on osteoarthritis risk, suggesting that the genetic correlations detected by LD score regression may be mostly due to horizontal pleiotropy, although for some phenotypes the MR analyses were underpowered (Supplementary Table 26). In the case of triglycerides and liability to type 2 diabetes, the Mendelian randomization analyses had sufficient power to rule out non-small causal effects, suggesting that these phenotypes have at most weak effects on osteoarthritis risk.

Crucially, structural changes in the joint usually precede the onset of symptoms for osteoarthritis. Articular cartilage is an avascular, aneural tissue. It provides tensile strength, compressive resilience and a low-friction articulating surface. Chondrocytes are the only cell type in cartilage. The mode of function of non-coding DNA is linked to context-dependent gene expression regulation, and identification of the causal variants and the genes they affect requires experimental analysis of genome regulation in the right cell type. Our functional analysis of genes in osteoarthritis-associated regions and pathways identified differentially expressed molecules in chondrocytes extracted from degraded compared to intact articular cartilage. Cartilage degeneration is a key hallmark of osteoarthritis pathogenesis and regulation of these genes could be implicated in disease development and progression.

Osteoarthritis is a leading cause of disability worldwide and carries a substantial public health and health economics burden. Here, we have gleaned novel insights into the genetic aetiology of osteoarthritis, and have implicated genes with translational potential[10,13,14,27,28]. The cohorts contributing to this study were composed of European-descent populations. Going forward, large-scale whole genome sequencing studies of well-phenotyped individuals across diverse populations will capture the full allele frequency and variation type spectrum, and afford us further insights into the causes of this debilitating disease.

URLs
Quanto, http://biostats.usc.edu/Quanto.html; genotyping and quality control of UK Biobank, http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf; genotype imputation of UK Biobank, http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/imputation_documentation_May2015.pdf; GRCh38 cDNA assembly release 87, http://ftp.ensembl.org/pub/release-87/fasta/homo_sapiens/cdna/

ACKNOWLEDGEMENTS
This research has been conducted using the UK Biobank Resource under Application Number 9979. This work was funded by the Wellcome Trust (WT098051). We are grateful to Roger Brooks, Jyoti Choudhary and Theodoros Roumeliotis for their contribution to the functional genomics data collection. The Human Research Tissue Bank is supported by the NIHR Cambridge Biomedical Research Centre.

AUTHOR CONTRIBUTIONS
Association analyses: EleniZ, KH, IT, LS, JS, SH, AG
Mendelian randomisation: FH, GDS
Functional genomics sample collection: AMcC, JMW, EZ
Functional genomics analyses: JS, LS
Endophenotype analyses: CGB, AGU, JBJvM
Replication analyses: US, TI, HJ, UT, KS
Bioinformatics: AG, DS, BK
Student supervision: KH, GCB, GDS, JMW, EZ
Manuscript writing: EleniZ, KH, IT, JS, FH, LS, CGB, US, DS, JBJvM, GDS, JMW, EZ

REFERENCES

1. Loeser, R.F., Goldring, S.R., Scanzello, C.R. & Goldring, M.B. Osteoarthritis: a disease of the joint as an organ. *Arthritis Rheum* **64**, 1697-707 (2012).
2. Felson, D.T. *et al.* Osteoarthritis: new insights. Part 1: the disease and its risk factors. *Ann Intern Med* **133**, 635-46 (2000).
3. Uhalte, E.C., Wilkinson, J.M., Southam, L. & Zeggini, E. Pathways to understanding the genomic aetiology of osteoarthritis. *Hum Mol Genet* **26**, R193-R201 (2017).
4. Ma, C., Blackwell, T., Boehnke, M., Scott, L.J. & Go, T.D.i. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet Epidemiol* **37**, 539-50 (2013).
5. Broome, D.T. & Datta, N.S. Mitogen-activated protein kinase phosphatase-1: function and regulation in bone and related tissues. *Connect Tissue Res* **57**, 175-89 (2016).
6. Rodríguez-Carballo, E., Gámez, B. & Ventura, F. p38 MAPK Signaling in Osteoblast Differentiation. *Front Cell Dev Biol* **4**, 40 (2016).
7. Wei, L., Sun, X.J., Wang, Z. & Chen, Q. CD95-induced osteoarthritic chondrocyte apoptosis and necrosis: dependency on p38 mitogen-activated protein kinase. *Arthritis Res Ther* **8**, R37 (2006).
8. Wang, Q. *et al.* Bioinformatics analysis of gene expression profiles of osteoarthritis. *Acta Histochem* **117**, 40-6 (2015).
9. Prasadam, I. *et al.* Osteoarthritic cartilage chondrocytes alter subchondral bone osteoblast differentiation via MAPK signalling pathway involving ERK1/2. *Bone* **46**, 226-35 (2010).
10. Prasadam, I. *et al.* Inhibition of p38 pathway leads to OA-like changes in a rat animal model. *Rheumatology (Oxford)* **51**, 813-23 (2012).
11. Prasadam, I. *et al.* ERK-1/2 and p38 in the regulation of hypertrophic changes of normal articular cartilage chondrocytes induced by osteoarthritic subchondral osteoblasts. *Arthritis Rheum* **62**, 1349-60 (2010).
12. Zhang, Y., Pizzute, T. & Pei, M. A review of crosstalk between MAPK and Wnt signals and its impact on cartilage regeneration. *Cell Tissue Res* **358**, 633-49 (2014).
13. Namdari, S., Wei, L., Moore, D. & Chen, Q. Reduced limb length and worsened osteoarthritis in adult mice after genetic inhibition of p38 MAP kinase activity in cartilage. *Arthritis Rheum* **58**, 3520-9 (2008).
14. Zhang, R., Murakami, S., Coustry, F., Wang, Y. & de Crombrugghe, B. Constitutive activation of MKK6 in chondrocytes of transgenic mice inhibits proliferation and delays endochondral bone formation. *Proc Natl Acad Sci U S A* **103**, 365-70 (2006).
15. Castañón, M.J., Walko, G., Winter, L. & Wiche, G. Plectin-intermediate filament partnership in skin, skeletal muscle, and peripheral nerve. *Histochem Cell Biol* **140**, 33-53 (2013).
16. Tseng, K.Y., Chen, Y.H. & Lin, S. Zinc finger protein ZFP36L1 promotes osteoblastic differentiation but represses adipogenic differentiation of mouse multipotent cells. *Oncotarget* **8**, 20588-20601 (2017).
17. Castano-Betancourt, M.C. *et al.* Novel Genetic Variants for Cartilage Thickness and Hip Osteoarthritis. *PLoS Genet* **12**, e1006260 (2016).
18. Usmani, S.E. *et al.* Transforming growth factor alpha controls the transition from hypertrophic cartilage to bone during endochondral bone growth. *Bone* **51**, 131-41 (2012).
19. Appleton, C.T., Usmani, S.E., Bernier, S.M., Aigner, T. & Beier, F. Transforming growth factor alpha suppression of articular chondrocyte phenotype and Sox9 expression in a rat model of osteoarthritis. *Arthritis Rheum* **56**, 3693-705 (2007).
20. Appleton, C.T., Usmani, S.E., Mort, J.S. & Beier, F. Rho/ROCK and MEK/ERK activation by transforming growth factor-alpha induces articular cartilage degradation. *Lab Invest* **90**, 20-30 (2010).
21. Usmani, S.E. *et al.* Context-specific protection of TGFα null mice from osteoarthritis. *Sci Rep* **6**, 30434 (2016).

22. Miettinen, P.J. *et al.* Epidermal growth factor receptor function is necessary for normal craniofacial development and palate closure. *Nat Genet* **22**, 69-73 (1999).
23. Chu, J. *et al.* A mouse forward genetics screen identifies LISTERIN as an E3 ubiquitin ligase involved in neurodegeneration. *Proc Natl Acad Sci U S A* **106**, 2097-103 (2009).
24. Wang, M. & Casey, P.J. Protein prenylation: unique fats make their mark on biology. *Nat Rev Mol Cell Biol* **17**, 110-22 (2016).
25. Gonzales, P.A. *et al.* Large-scale proteomics and phosphoproteomics of urinary exosomes. *J Am Soc Nephrol* **20**, 363-79 (2009).
26. Palmieri, A. *et al.* Comparison between osteoblasts derived from human dental pulp stem cells and osteosarcoma cell lines. *Cell Biol Int* **32**, 733-8 (2008).
27. Ochiai, S. *et al.* Oxidative stress reaction in the meniscus of Bach 1 deficient mice: potential prevention of meniscal degeneration. *J Orthop Res* **26**, 894-8 (2008).
28. Takada, T. *et al.* Bach1 deficiency reduces severity of osteoarthritis through upregulation of heme oxygenase-1. *Arthritis Res Ther* **17**, 285 (2015).
29. Capellini, T.D. *et al.* Ancient selection for derived alleles at a GDF5 enhancer influencing human growth and osteoarthritis risk. *Nat Genet* **49**, 1202-1210 (2017).
30. Daans, M., Luyten, F.P. & Lories, R.J. GDF5 deficiency in mice is associated with instability-driven joint damage, gait and subchondral bone changes. *Ann Rheum Dis* **70**, 208-13 (2011).
31. Miyamoto, Y. *et al.* A functional polymorphism in the 5' UTR of GDF5 is associated with susceptibility to osteoarthritis. *Nat Genet* **39**, 529-33 (2007).
32. Southam, L. *et al.* An SNP in the 5'-UTR of GDF5 is associated with osteoarthritis susceptibility in Europeans and with in vivo differences in allelic expression in articular cartilage. *Hum Mol Genet* **16**, 2226-32 (2007).
33. Smith, G.D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1-22 (2003).
34. Prieto-Alhambra, D. *et al.* An increased rate of falling leads to a rise in fracture risk in postmenopausal women with self-reported osteoarthritis: a prospective multinational cohort study (GLOW). *Ann Rheum Dis* **72**, 911-7 (2013).
35. Baldwin, J.N. *et al.* Self-reported knee pain and disability among healthy individuals: reference data and factors associated with the Knee injury and Osteoarthritis Outcome Score (KOOS) and KOOS-Child. *Osteoarthritis Cartilage* (2017).
36. Callahan, L.F. *et al.* Associations of educational attainment, occupation and community poverty with knee osteoarthritis in the Johnston County (North Carolina) osteoarthritis project. *Arthritis Res Ther* **13**, R169 (2011).
37. Hussain, S.M. *et al.* How Are Obesity and Body Composition Related to Patellar Cartilage? A Systematic Review. *J Rheumatol* **44**, 1071-1082 (2017).
38. Fewell, Z., Davey Smith, G. & Sterne, J.A. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol* **166**, 646-55 (2007).

FIGURE LEGENDS

Figure 1

Power to detect association in the discovery stage. Odds ratios (ORs) and 95% confidence intervals as a function of minor allele frequencies (MAF). Newly reported variants are denoted in diamonds, while known variants are denoted in circles. The curves indicate 80% power at the genome-wide significance threshold of $P<5.0\times10^{-8}$, for the number of cases and controls of each trait at the discovery stage.

Figure 2

Regional association plots for the nine novel osteoarthritis loci. The y axis represents the negative logarithm (base 10) of the variant *P*-value and the x axis represents the position on the chromosome, with the name and location of genes and nearest genes shown in the

bottom panel. The variant with the lowest *P*-value in the region after combined discovery and replication is marked by a purple diamond. The same variant is marked by a purple dot showing the discovery *P*-value. The colours of the other variants indicate their r^2 with the lead variant.

Figure 3
Heat map of genetic correlations between osteoarthritis phenotypes in UK Biobank and 35 traits grouped in 10 categories from GWAS consortia. Symbols and hues depict the FDR *q-values* and strength of the genetic correlation (darker shade denotes stronger correlation), respectively. Red and blue indicate positive and negative correlations, respectively. RP: reproductive; SL: sleep.

Figure 4
Two-sample Mendelian randomization estimates and 95% confidence intervals of the effect of obesity-related measures, triglyceride levels, years of schooling (in standard deviations units) and type 2 diabetes liability (in ln(odds ratio) units) on different definitions of osteoarthritis.
HD: hospital diagnosed. IVW: inverse-variance weighting. MBE: mode-based estimate. MBE (1): tuning parameter φ=1. MBE (0.5): tuning parameter φ=0.5.

Table 1

Association summary statistics for the nine signals. [a]Imputation accuracy was assessed by IMPUTE-infoscore. [b]Heterogeneity *P-values* are derived from Cochran's Q-test. [c]The number of cases required for 80% power is calculated on the basis of the replication study OR estimate, sample size-weighted effect allele frequency across the discovery and replication studies, and assuming 4 controls per case. [d]two sided *P*-value. EA, effect allele; EAF, effect allele frequency; OR, odds ratio; n, sample size.

rsID	EA	discovery phenotype	discovery EAF	discovery OR	discovery OR lower 95% CI	discovery OR upper 95% CI	discovery *P-value*[d]	discovery n cases/ controls	imputation accuracy score discovery [a]	replication phenotype	replication EAF	replication OR	replication OR lower 95% CI	replication OR upper 95% CI	replication *P-value*[d]	replication n cases/ controls	imputation accuracy score replication [a]	overall OR	overall OR lower 95% CI	overall OR upper 95% CI	overall *P-value*[d]	hetero-geneity *P*-value [b]	overall n cases/ controls
rs2820436	C	Hospital diagnosed osteoarthritis	0.66	0.92	0.9	0.96	$6.45x10^{-6}$	10,083/ 40,425	Directly typed	osteoarthritis at any site	0.64	0.94	0.91	0.97	$8.71x10^{-5}$	18,069/ 246,293	0.99972	0.93	0.91	0.96	$2.01x10^{-9}$	0.5739	28,152/ 286,718
rs3771501	G	Self-reported osteoarthritis	0.53	0.94	0.91	0.96	$3.81x10^{-6}$	12,658/ 50,898	0.991707	osteoarthritis at any site	0.54	0.95	0.92	0.98	0.001069	18,069/ 246,293	0.999808	0.94	0.92	0.96	$1.66x10^{-8}$	0.4825	30,727/ 297,191
rs11335718	A	Self-reported osteoarthritis	0.11	1.12	1.07	1.17	$1.12x10^{-6}$	12,658/ 50,898	0.968932	Knee osteoarthritis	0.11	1.1	1.02	1.2	0.014675	4,672/ 172,791	0.998899	1.11	1.07	1.16	$4.26x10^{-8}$	0.792	17,330/ 223,689
rs11335718	A	Self-reported osteoarthritis	0.11	1.12	1.07	1.17	$1.12x10^{-6}$	12,658/ 50,898	0.968932	osteoarthritis at any site	0.11	1.06	1.01	1.11	0.013023	18,069/ 246,293	0.998899	1.09	1.06	1.13	$1.32x10^{-7}$	0.1254	30,727/ 297,191
rs11780978	A	Hospital diagnosed hip	0.4	1.16	1.08	1.23	$6.24x10^{-6}$	2,396/ 9,593	0.983752	Hip osteoarthritis	0.39	1.11	1.05	1.16	$4.55x10^{-5}$	5,714/ 199,421	0.999673	1.13	1.08	1.17	$1.98x10^{-9}$	0.2424	8,110/ 209,014
rs116882138	A	Hospital diagnosed hip and/or knee osteoarthritis	0.02	1.4	1.22	1.6	$2.96x10^{-6}$	6,586/ 26,384	Directly typed	Knee osteoarthritis	0.02	1.27	1.07	1.5	0.006552	4,672/ 172,791	0.998087	1.34	1.21	1.49	$5.09x10^{-8}$	0.3988	11,258/ 199,175
rs116882138	A	Hospital diagnosed hip and/or knee osteoarthritis	0.02	1.4	1.22	1.6	$2.96x10^{-6}$	6,586/ 26,384	Directly typed	Hip and/or knee osteoarthritis	0.02	1.13	0.99	1.29	0.069018	9,429/ 199,421	0.998087	1.25	1.14	1.38	$2.93x10^{-6}$	0.03456	16,015/ 222,805
rs2521349	A	Hospital diagnosed hip osteoarthritis	0.38	1.18	1.11	1.26	$6.85x10^{-7}$	2,396/ 9,593	0.996479	Hip osteoarthritis	0.37	1.1	1.05	1.16	0.000103	5,714/ 199,421	0.999925	1.13	1.09	1.18	$9.95x10^{-10}$	0.1151	8,110/ 209,014
rs864839	T	Self-reported osteoarthritis	0.72	1.08	1.05	1.12	$6.21x10^{-7}$	12,658/ 50,898	0.97115	Hip osteoarthritis	0.7	1.07	1.02	1.13	0.008275	5,714/ 199,421	0.997756	1.08	1.05	1.11	$2.01x10^{-8}$	0.7886	18,372/ 250,319
rs864839	T	Self-reported osteoarthritis	0.72	1.08	1.05	1.12	$6.21x10^{-7}$	12,658/ 50,898	0.97115	osteoarthritis at any site	0.7	1.02	0.99	1.06	0.18218	18,069/ 246,293	0.997756	1.05	1.03	1.08	$7.02x10^{-6}$	0.0121	30,727/ 297,191
rs375575359	C	Self-reported osteoarthritis	0.03	1.2	1.12	1.29	$9.96x10^{-8}$	12,658/ 50,898	0.823533	Knee osteoarthritis	0.05	1.15	1.02	1.29	0.025177	4,672/ 172,791	0.992996	1.21	1.14	1.3	$7.54x10^{-9}$	0.25	17,330/ 223,689
rs375575359	C	Self-reported osteoarthritis	0.03	1.2	1.12	1.29	$9.96x10^{-8}$	12,658/ 50,898	0.823533	osteoarthritis at any site	0.05	1.03	0.96	1.1	0.47234	18,069/ 246,293	0.992996	1.12	1.06	1.18	$2.6x10^{-5}$	0.000403	30,727/ 297,191
rs6516886	T	Hospital diagnosed hip and/or knee osteoarthritis	0.75	1.13	1.08	1.19	$5.36x10^{-8}$	6,586/ 26,384	0.998499	Hip osteoarthritis	0.76	1.06	1	1.12	0.055135	5,714/ 199,421	0.999981	1.1	1.06	1.14	$5.84x10^{-8}$	0.06276	12,300/ 225,805

rs6516886	T	Hospital diagnosed hip and/or knee osteoarthritis	0.75	1.13	1.08	1.19	$5.36x10^{-8}$	6,586/ 26,384	0.998499	Hip and/or knee osteoarthritis	0.76	1.05	1	1.11	0.043467	9,429/ 199,421	0.999981	1.09	1.06	1.13	$1.42x10^{-7}$	0.01914	16,015/ 222,805

Table 2

Association of the 9 osteoarthritis loci with radiographically-derived osteoarthritis endophenotypes. [a]For minimal joint space width, proxy variant rs2150403 (r2=0.99 with rs6516886) was used. [b]Sample size=13,013. [c]Sample size=6,880. [d]Sample size cases=639. [e]Sample size controls=4,339. [f]two sided p-value. EA: effect allele; EAF: effect allele frequency; SE: standard error; N/A: not available.

	minimal joint space width[b]					**Center edge-angle**[c]					**alpha angle (cam deformity)**[d,e]				
rsID	**EA**	**EAF**	**Beta**	**SE**	***P*-value**[f]	**EA**	**EAF**	**Beta**	**SE**	***P*-value**[f]	**EA**	**EAF**	**Beta**	**SE**	***P*-value**[f]
rs2820436	A	0.317	-0.0146	0.0135	0.2817	A	0.317	-0.0104	0.1301	0.9363	A	0.318	0.0165	0.0675	0.8073
rs3771501	A	0.484	-0.0699	0.0127	3.454E-08	A	0.474	0.1943	0.1199	0.1051	A	0.4779	-0.0144	0.0626	0.8176
rs11335718	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
rs11780978	A	0.389	-0.0291	0.0129	0.02419	A	0.386	0.078	0.1239	0.5291	A	0.3866	0.0035	0.0644	0.9564
rs2521349	A	0.398	0.0229	0.0128	0.07404	A	0.391	0.0998	0.1236	0.4192	A	0.3921	-0.0262	0.0644	0.6835
rs864839	N/A	N/A	N/A	N/A	N/A	T	0.702	-0.0206	0.1325	0.8766	T	0.7026	-0.0081	0.0691	0.907
rs375575359	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
rs116882138	N/A	N/A	N/A	N/A	N/A	A	0.0137	-1.1388	0.5276	0.0309	A	0.0135	0.1814	0.2607	0.4865
rs6516886[a]	T	0.272	-0.0222	0.0143	0.1206	A	0.265	-0.1491	0.1373	0.2773	A	0.263	0.0544	0.0713	0.4458

Table 3

Genes in the osteoarthritis-associated signals with significantly different gene expression and/or protein abundance in intact compared to degraded articular cartilage. logFC: log2-fold change based on normalised values (increase means higher value in degraded cartilage); FDR: Bonferroni-Hochberg false discovery rate; - denotes that proteomics data were not available.

Index variant	Gene	Position (chr:start-end)	Distance from index variant (kb)	Proteomics logFC	Proteomics FDR *q-value*	RNAseq logFC	RNAseq FDR *q-value*
rs3771501	*PCYOX1*	2:70484518- 70508323	209.3	-0.27	0.0042	0.27	0.0047
rs3771501	*FAM136A*	2:70523107- 70529222	188.4	-	-	-0.20	0.0066
rs6516886	*BACH1*	21:30566392- 31003071	172.7	-	-	0.32	0.0019
rs6516886	*MAP3K7CL*	21:30449792-30548210	56.1	-	-	0.41	0.0021
rs11780978	*BOP1*	8:145486055-145515082	451.2	-	-	-0.26	0.0030
rs116882138	*PLAA*	9:26904081-26947461	366	-0.07	0.601	0.20	0.0027
rs375575359	*ZNF382*	19:37095719-37119499	233.8	-	-	0.39	0.0031

ONLINE METHODS

Accuracy of self-reported data

We evaluated the classification accuracy of self-reported disease status by estimating the sensitivity, specificity, positive predictive values (PPV) and negative predictive values (NPV) in the self-reported and hospital diagnosed disease definition datasets. We performed a sensitivity test to evaluate the true positive rate by calculating the proportion of individuals diagnosed with osteoarthritis that were correctly identified as such in the self-reported analysis, and a specificity test to evaluate the true negative rate by calculating the proportion of individuals not diagnosed with osteoarthritis that were correctly identified as such in the control set. The number of individuals overlapping between the self-reported (n_{SR}=12,658) and hospital-diagnosed (n_{HD}=10,083) datasets was n_{OVER}=3,748. The total number of individuals was n_{TOT}=138,997. $\text{Sensitivity} = \frac{n_{OVER}}{n_{HD}}$; $\text{Specificity} = \frac{n_{TOT}-(n_{HD}+n_{SR}-n_{OVER})}{n_{TOT}-n_{HD}}$; $\text{PPV} = \frac{n_{OVER}}{n_{SR}}$; $\text{NPV} = \frac{n_{TOT}-(n_{HD}+n_{SR}-n_{OVER})}{n_{TOT}-n_{SR}}$.

Discovery GWAS

UK Biobank's scientific protocol and operational procedures were reviewed and approved by the North West Research Ethics Committee (REC Reference Number: 06/MRE08/65). The 1st UK Biobank release of genotype data includes ~150,000 volunteers between 40-69 years old from the UK, genotyped at approximately 820,967 single nucleotide polymorphisms (SNPs). 50,000 samples were genotyped using the UKBiLEVE array and the remaining samples were genotyped using the UK Biobank Axiom array (Affymetrix; see URLs). The UK Biobank Axiom is an update of UKBiLEVE and the two arrays share 95% of their content. In total, after sample and SNP quality control (QC), which was carried out centrally, 152,763 individuals and 806,466 directly typed SNPs remained. Phasing, imputation and derivation of principal components were also carried out centrally. Briefly, the combined UK10K/ 1000 Genomes Project haplotype reference panel was used to impute untyped variants through the IMPUTE3 program (see URLs). Following imputation, the number of variants reached 73,355,667 in 152,249 individuals. We performed additional quality control (QC) checks. We excluded samples with call rate ≤97%. We checked samples for gender discrepancies, excess heterozygosity, relatedness, ethnicity and we removed possibly contaminated and withdrawn samples. Following QC, the number of individuals was 138,997. We excluded 528 SNPs that had been centrally flagged as subject to exclusion due to failure in one or more additional quality metrics.

To define osteoarthritis cases, we used the self-reported status questionnaire and the Hospital Episode Statistics data (Supplementary Table 3; Supplementary note). We conducted five osteoarthritis discovery GWAS and one sensitivity analysis, and the case strata were: self-reported osteoarthritis at any site n=12,658; sensitivity analysis (a random subset of the self-reported cohort equal to the sample size of the hospital diagnosed cohort) n=10,083; hospital-diagnosed osteoarthritis at any site based on ICD10 and/or ICD9 hospital records codes n=10,083; hospital-diagnosed hip osteoarthritis n=2,396; hospital-diagnosed knee osteoarthritis n= 4,462; and hospital-diagnosed hip and/or knee osteoarthritis n=6,586. We applied exclusion criteria to minimise misclassification in the control datasets to the extent possible (using approximately 4x the number of cases for each definition) (Supplementary Table 2, Supplementary Fig. 1). We restricted the number of controls used and did not utilise the full set of available genotyped control samples from UK Biobank in order to guard against association test statistics behaving anti-conservatively in the presence of stark case: control imbalance for alleles with minor allele count (MAC) <400[4] (analogous to MAF ~0.02 in the self-reported and hospital diagnosed osteoarthritis datasets). For the control set, we excluded all participants diagnosed with any musculoskeletal disorder, or with relevant symptoms or signs, such as pain and arthritis, and selected older participants to ensure we decrease the number of controls that might be

diagnosed with osteoarthritis in the future, while keeping the number of males and females balanced (Supplementary Table 1).

At the SNP level, we further filtered for Hardy Weinberg equilibrium (HWE) $P \leq 10^{-6}$, MAF≤0.001 and info score<0.4 (Supplementary Fig. 1). We tested for association using the frequentist likelihood ratio test (LRT) and method ml in SNPTEST v2.5.2[39] with adjustment for the first 10 principal components in order to control for population structure. Power calculations were carried out using Quanto v1.2.4 (see URLs).

Replication

Two hundred independent and novel variants with $P<1.0\times10^{-5}$ in the discovery analyses were taken forward for *in silico* replication in an independent cohort from Iceland (deCODE) using fixed effects inverse-variance weighted meta-analysis in METAL[40]. One hundred and seventy three variants were present in the replication cohort. The remaining 27 variants had ambiguous alleles, i.e. incompatible due to alignment issues, and were not included in further analyses. The significance threshold for declaring association in the replication study was hence $0.05/173=2.9\times10^{-4}$.The deCODE dataset comprised four osteoarthritis phenotypes: any osteoarthritis site (18,069 cases and 246,293 controls), hip osteoarthritis (5,714 cases and 199,421 controls), knee osteoarthritis (4,672 cases and 172,791 controls), and hip and/or knee osteoarthritis (9,429 cases and 199,421 controls). We performed meta-analyses (across osteoarthritis definitions) using summary statistics from the UK Biobank osteoarthritis analyses and deCODE. We use $P \leq 2.8\times10^{-8}$ as the threshold corrected for the effective number of traits to report genome wide significance.

Replication cohort: The information on hip, knee and vertebral osteoarthritis was obtained from Landspitali University Hospital electronic health records, Akureyri Hospital electronic health records and from a national Icelandic hip or knee arthroplasty registry[41]. Secondary osteoarthritis (e.g. Perthes disease, hip dysplasia), post-trauma osteoarthritis (e.g. ACL rupture) and those also diagnosed with rheumatoid arthritis were excluded from these lists. Only those diagnosed with osteoarthritis after the age of 40 were included. Hand osteoarthritis patients were drawn from a database of 9,000 hand osteoarthritis patients that was initiated in 1972[42]. The study was approved by the Data Protection Authority of Iceland and the National Bioethics Committee of Iceland. Informed consent was obtained from all participants.

Association with osteoarthritis-related endophenotypes

The 9 replicating genetic loci were examined for association in radiographic osteoarthritis endophenotypes. This was done for 3 phenotypes: minimal Joint space width (mJSW), and two measures of hip shape deformities known as strong predictors for osteoarthritis: acetabular dysplasia (measured with Center Edge-angle), and cam deformity (as measured with alpha angle). For mJSW association statistics for the variants were looked-up in a previously published GWAS, which joint analyzed data from the Rotterdam Study I (RS-I), Rotterdam Study II (RS-II), TwinsUK, SOF and MrOS using standardized age, gender and population stratification (four principal components) adjusted residuals from linear regression[17]. For the two hip shape phenotypes, CE-angle and alpha angle were measured as previously published. CE-angle was analyzed as a continuous phenotype. We conducted GWAS on a total of 6880 individuals from the Rotterdam Study I (RS-I), Rotterdam Study II (RS-II), Rotterdam Study III (RS-III) and CHECK[43] datasets using standardized age, gender adjusted residuals from linear regression. For cam-deformity individuals with an alpha-angle >60° were defined as a case (n=639), while all others were controls (4339). The GWAS was done on individuals from RS-I, RS-II and CHECK, using age, sex and principal components to adjust for population stratification as covariates. The results of the separate cohorts were combined in a meta-analysis using inverse variance weighting with METAL[40]. Genomic control correction was applied to the standard errors and *P-values* before meta-analysis.

Functional genomics

Patients and samples: We collected cartilage samples from 38 patients undergoing total joint replacement surgery: 12 knee osteoarthritis patients (cohort 1; 2 women, 10 men, age 50-88 years); 17 knee osteoarthritis patients (cohort 2; 12 women 5 men, age 54-82 years); 9 hip osteoarthritis patients (cohort 3; 6 women, 3 men, age 44-84 years). We collected matched intact and degraded cartilage samples from each patient (Supplementary Note). Cartilage was separated from bone and chondrocytes were extracted from each sample. From each isolated chondrocyte sample, we extracted RNA and protein. All patients provided full written informed consent prior to participation. All sample collection, RNA and protein extraction steps are described in detail in[44].

Proteomics: Proteomics analysis was performed on intact and degraded cartilage samples from 24 individuals (15 from cohort 2, 9 from cohort 3). LC-MS analysis was performed on the Dionex Ultimate 3000 UHPLC system coupled with the Orbitrap Fusion Tribrid Mass Spectrometer. To account for protein loading, abundance values were normalised by the sum of all protein abundances in a given sample, then log2-transformed and quantile normalised. We restricted the analysis to 3917 proteins that were quantified in all samples. We tested proteins for differential abundance using limma[45] in R, based on a within-individual paired sample design. Significance was defined at 1% Benjamini-Hochberg False Discovery Rate (FDR) to correct for multiple testing. Of the 3732 proteins with unique mapping of gene name and Ensembl ID, we took forward 245 proteins with significantly different abundance between intact and degraded cartilage at 1% FDR.

RNA sequencing: We performed a gene expression analysis on samples from all 38 patients. Multiplexed libraries were sequenced on the Illumina HiSeq 2000 (75bp paired-end read length). This yielded bam files for cohort 1 and cram files for cohorts 2 and 3. The cram files were converted to bam files using samtools 1.3.1[46] and then to fastq files using biobambam 0.0.191[47], after exclusion of reads that failed QC. We obtained transcript-level quantification using salmon 0.8.2[48] (with --gcBias and --seqBias flags to account for potential biases) and the GRCh38 cDNA assembly release 87 downloaded from Ensembl (see URLs). We converted transcript-level to gene-level count estimates, with estimates for 39037 genes based on Ensembl gene IDs. After quality control, we retained expression estimates for 15994 genes with counts per million of 1 or higher in at least 10 samples. Limma-voom[49] was used to remove heteroscedasticity from the estimated expression data. We tested genes for differential expression using limma[45] in R (with lmFit and eBayes), based on a within-individual paired sample design. Significance was defined at 1% Benjamini-Hochberg false discovery rate (FDR) to correct for multiple testing. Of the 14408 genes with unique mapping of gene name and Ensembl ID, we took forward 1705 genes with significantly different abundance between intact and degraded cartilage at 1% FDR.

Fine-mapping

We constructed regions for fine-mapping, by taking a window of at least 0.1 centimorgans either side of each index variant. The region was extended to the furthest variant with r^2>0.1 with the index variant within a 1Mb window. LD calculations for extending the region were based on whole-genome sequenced EUR samples from the combined reference panel of UK10K[50] and 1000 Genomes Projects[51]. For each region we implemented the Bayesian fine-mapping method CAVIARBF[52], which uses association summary statistics and correlations among variants to calculate Bayes' factors and posterior probabilities of each variant being causal. We assumed a single causal variant in each region and calculated 95% credible intervals, which contain the minimum set of variants that jointly have at least 95% probability of including the causal variant. We also applied the extended CAVIARBF method that uses functional annotation scores to upweigh variants according to their predicted functional scores. To this end, we downloaded pre-calculated CADD[53] and Eigen[54] scores from their equivalent websites. We observed better separation of severe-consequence genic variants with the CADD score and better separation of regulatory variants with the Eigen score, and

therefore created a combined score, where splice acceptor, splice donor, stop lost, stop gained, missense and splice region variants were assigned their CADD-Phred score and the rest their Eigen-Phred score.

Functional enrichment analysis
We used genome-wide summary statistics to test for enrichment of functional annotations. We used GARFIELD[55] with customized functional annotations, making use of the functional genomics data we generated in primary articular chondrocytes using RNA sequencing and quantitative proteomics. We defined differentially expressed genes separately at the RNA (transcriptional) level and at the protein level when comparing intact to degraded cartilage (1% FDR). We extended each differentially regulated gene by 5kb each side. Using GARFIELD's approach, we calculated the effective number of independent annotations to be 1.995, which led to an adjusted p-value significance level of 0.025. We tested for enrichment using variants with $P<1.0x10^{-5}$ and no analysis surpassed the corrected significance threshold.

LD regression
We used LDHub[56] [accessed 23-27 January] to estimate the genome-wide genetic correlation between each of the osteoarthritis definitions and 219 other human traits and diseases. In each analysis, we extracted variants with rsIDs (range 11,999,363-15,561,966) and uploaded the corresponding association summary statistics to LDHub, yielding 896,076-1,172,130 variants overlapping with LDHub. We corrected for multiple testing by defining significance at 5% Benjamini-Hochberg False Discovery Rate (FDR) for each of the five osteoarthritis analyses.

Mendelian randomization analysis
We used Mendelian randomization (MR) to assess the potential causal role of the phenotypes identified in the LD score regression analysis on osteoarthritis. We also included birth weight and height (Supplementary Table 22). In all analyses, the primary outcome variable was self-reported osteoarthritis. We used data from hospital records (which were available for a much smaller number of individuals) as sensitivity analyses and to identify potential site-specific effects.

Data sources: Genetic instruments were identified from publicly-available summary GWAS results through the TwoSampleMR R package, which allows extracting the data available in the MR-Base database[57]. Only results that combined both sexes were extracted. Preference was given to studies restricted to European populations to minimise the risk of bias due to population stratification; however, for a few traits those were either not available or corresponded to much smaller studies (Supplementary Table 22). However, this is unlikely to substantially bias the results because all studies employed correction methods, and even multi-ethnic studies are mostly composed of European populations. The exception was for number of children ever born and age of the individual when his/her first child was born: given that the GWAS of reproductive traits by Barban and colleagues[58] was not available in MR-Base, we extracted summary association results for the variants that achieved genome-wide significance directly from the paper, and used coefficients from each sex in sex-specific analyses. The search was performed on June 19, 2017. For each trait, all genetic instruments achieved the conventional levels of genome-wide significance (i.e., $P<5.0x10^{-8}$) and were mutually independent (i.e. $r^2<0.001$ between all pairs of instruments).

Two-sample MR: For the exposure phenotypes with at least one genetic instrument available, we used two-sample MR analysis to evaluate their causal effect on osteoarthritis risk. The exceptions were smoking and reproductive traits, which were performed using one-sample MR only due to the need of performing the analysis within specific subgroups. All summary association results used for two-sample MR are shown in Supplementary Table

23, and Supplementary Table 24 provides an overall description of each set of genetic instruments. We applied the following methods:

- Ratio method: for exposure phenotypes with only one genetic instrument available, MR was performed using the ratio method, which consists of dividing the instrument-outcome by the instrument-exposure regression coefficient. The standard error of the ratio estimate can be calculated by dividing the instrument-outcome standard error by the instrument-exposure regression coefficient. Confidence intervals and *P-value* were calculated using the Normal approximation.
- Inverse-variance weighting (IVW): this method allows combination of the ratio estimates from multiple instruments into a single, pooled estimate. We used a multiplicative random effects version of the method, which incorporates between-instrument heterogeneity in the confidence intervals.
- MR-Egger regression: this method yields consistent causal effect estimates even if all instruments are invalid, provided that horizontal pleiotropic effects are uncorrelated with instrument strength (i.e. the Instrument Strength Independent of Direct Effects – InSIDE – assumption holds).
- Weighted median: this method allows consistent causal effect estimation even if the InSIDE assumption is violated, provided that up to (but not including) 50% of the weights in the analysis come from invalid instruments.
- Mode-based estimate (MBE): the weighted MBE relies on the ZEro Modal Pleiotropy Assumption (ZEMPA), which postulates that the largest subgroup (or the subgroup that carries the largest amount of weight in the analysis) of instruments that estimate the same causal effect estimate is composed of valid instruments. This allows consistent causal effect estimation even if the majority of instruments are invalid. The stringency of the method can be regulated by the φ parameter. We tested two values of φ: φ=1 (ie, the default) and φ=0.5 (half of the default, or twice as stringent).

For exposure phenotypes with more than 1 but less than 10 genetic instruments, only the IVW method was applied. This was because the remaining methods are typically less powered and require a relatively large number of genetic instruments to provide reliable results. The degree of weak instrument bias (which corresponds to regression dilution bias in two-sample MR) for the IVW and MR-Egger methods was quantified using the $\frac{F_{XG}-1}{F_{XG}}$ and I^2_{GX} statistics, respectively. Both range from 0% to 100%, and $100\left(1-\frac{F_{XG}-1}{F_{XG}}\right)\%$ and $100(1-I^2_{GX})\%$ can be interpreted as the amount of dilution in the corresponding causal effect estimates. Given that only genome-wide significant variants were selected as instruments, the $\frac{F_{XG}-1}{F_{XG}}$ statistic will necessarily be high (approximately 95%, at least). However, the I^2_{GX} statistic depends both on instrument strength and heterogeneity between instrument-exposure associations, which implies that regression dilution bias in MR-Egger can be substantial even if instruments are individually strong. Indeed, for some traits the I^2_{GX} statistic was very low (Supplementary Table 24). Therefore, all MR-Egger regression analyses were corrected for regression dilution bias using a Simulation Extrapolation (SIMEX) approach.

Horizontal pleiotropy tests: We additionally assessed the robustness of our findings to potential violations of the assumption of no horizontal pleiotropy by applying two tests of horizontal pleiotropy. One of them was the MR-Egger intercept, which can be interpreted as the average instrument-outcome coefficient when the instrument-exposure coefficient is zero. If there is no horizontal pleiotropy, the intercept should be zero. Therefore, the intercept provides an indication of overall unbalanced horizontal pleiotropy . The second test was Cochran's Q test of heterogeneity, which relies on the assumption that all valid genetic instruments estimate the same causal effect.

Power calculations: We performed power calculations to estimate the power of our two-sample MR analysis to detect odds ratios of 1.2, 1.5 and 2.0 (Supplementary note).

One-sample MR: UK Biobank data were used to perform one-sample MR using the same genetic instruments than in the two-sample MR (Supplementary Note).

DATA AVAILABILITY

All RNA sequencing data have been deposited to the European Genome/Phenome Archive (cohort 1: EGAD00001001331; cohort 2: EGAD00001003355; cohort 3: EGAD00001003354).

METHODS-ONLY REFERENCES

39. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
40. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
41. Franklin, J., Ingvarsson, T., Englund, M. & Lohmander, S. Association between occupation and knee and hip replacement due to osteoarthritis: a case-control study. *Arthritis Res Ther* **12**, R102 (2010).
42. Styrkarsdottir, U. *et al.* Whole-genome sequencing identifies rare genotypes in COMP and CHADL associated with high risk of hip osteoarthritis. *Nat Genet* **49**, 801-805 (2017).
43. Wesseling, J. *et al.* CHECK (Cohort Hip and Cohort Knee): similarities and differences with the Osteoarthritis Initiative. *Ann Rheum Dis* **68**, 1413-9 (2009).
44. Steinberg, J. *et al.* Integrative epigenomics, transcriptomics and proteomics of patient chondrocytes reveal genes and pathways involved in osteoarthritis. *Sci Rep* **7**, 8935 (2017).
45. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
47. Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine* **9**, 13-13 (2014).
48. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Meth* **14**, 417-419 (2017).
49. Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**, R29 (2014).
50. UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
51. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
52. Chen, W. *et al.* Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**, 719-36 (2015).
53. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
54. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J.D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* **48**, 214-20 (2016).

55. Iotchkova, V. *et al.* Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. *Nat Genet* **48**, 1303-1312 (2016).
56. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).
57. Hemani, G. *et al.* MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv* (2016).
58. Barban, N. *et al.* Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat Genet* **48**, 1462-1472 (2016).