A molecular quantitative trait locus map for osteoarthritis

Julia Steinberg1,2,3,4, Lorraine Southam1,3, Theodoros I. Roumeliotis3,5, Matthew J. Clark6, Raveen L. Jayasuriya6, Diane Swift6, Karan M. Shah6, Natalie C. Butterfield7, Roger A. Brooks8, Andrew W. McCaskie8, J. H. Duncan Bassett7, Graham R. Williams7, Jyoti S. Choudhary3,5, J. Mark Wilkinson6,9,11✉ & Eleftheria Zeggini1,3,10,11✉

Osteoarthritis causes pain and functional disability for over 500 million people worldwide. To develop disease-stratifying tools and modifying therapies, we need a better understanding of the molecular basis of the disease in relevant tissue and cell types. Here, we study primary cartilage and synovium from 115 patients with osteoarthritis to construct a deep molecular signature map of the disease. By integrating genetics with transcriptomics and proteomics, we discover molecular trait loci in each tissue type and omics level, identify likely effector genes for osteoarthritis-associated genetic signals and highlight high-value targets for drug development and repurposing. These findings provide insights into disease aetiopathology, and offer translational opportunities in response to the global clinical challenge of osteoarthritis.
Osteoarthritis is a severe, debilitating disease that affects the whole joint organ and is hallmark by cartilage degeneration and synovial hypertrophy. As of 2019, osteoarthritis is estimated to affect 528 Million people worldwide and be the 15th leading cause of years lived with disability. The lifetime risk of developing symptomatic knee and hip osteoarthritis is estimated to be 45 and 25%, respectively, and is on an upward trajectory commensurate with rises in obesity and the ageing population. Between 2010 and 2019, the global prevalence of osteoarthritis and the resulting years lived with disability have both risen by 27.5%. In 2013, osteoarthritis cost the United States $304 Billion and was the second most costly health condition treated at US hospitals, accounting for 4.3% of the combined costs for all hospitalisations. Osteoarthritis-associated reduced physical activity results in a standardised all-cause mortality ratio of 1.55 (95% confidence interval 1.41 to 1.70) for its sufferers versus the general population. Disease management focusses on alleviating pain, and in end-stage disease the only treatment is joint replacement surgery, emphasising the clear and urgent need to develop new therapies. To achieve this, we need to improve our understanding of the underlying molecular pathophysiology.

Epidemiological risk factors for the disease have been well-established and include older age, female sex, obesity, joint morphology and injury, and family history. The heritability of osteoarthritis has been estimated to range between 40 (for knee osteoarthritis) and 60% (for hip osteoarthritis). Genome-wide association studies (GWAS) have identified ~90 robustly-replicating risk loci. However, the molecular landscape of osteoarthritis-relevant tissue has not been similarly characterised by large-scale efforts such as the GTEx, ENCODE and Road-Map Epigenomics projects.

In this work, we perform a deep characterisation of the transcriptional and proteomic landscape of disease in chondrocytes and synoviocytes extracted from primary joint tissue of osteoarthritis patients. We define molecular quantitative trait loci, identify likely effector genes for GWAS signals, characterise molecular features of cartilage degradation and highlight drug development and repurposing opportunities through analysis of transcriptional signature changes.

Results

**Generation of human joint tissue molecular profiles.** We collected and characterised low-grade (intact; preserved) and high-grade (highly degraded; lesioned) cartilage, and synovial tissue from patients undergoing joint replacement for osteoarthritis (see Methods). The availability of paired low-grade and high-grade cartilage samples enables the comparison between these two disease states in affected primary tissue within the same individual, with high-grade cartilage showing more advanced cartilage degradation. All three tissues were profiled by RNA sequencing, and cartilage samples were additionally profiled by quantitative proteomics (Fig. 1 and Supplementary Fig. 1). We generated genome-wide genotype data from peripheral blood to define molecular quantitative trait loci (molQTLs) in each tissue type and omics level.

**Molecular QTLs in osteoarthritis tissues.** Identification of molecular QTLs can provide a better understanding of the transcriptional regulation of key cell types across disease stages. We identified cis expression QTLs (cis-eQTLs) for 1891 genes in at least one tissue (Fig. 2), with high correlation of effects across the tissues studied (Supplementary Fig. 2). The direction of effect was concordant across all cis-eQTLs detected in both low-grade and high-grade cartilage. We identified cis protein QTLs (cis-pQTLs) for 38 genes in at least one tissue, with a similarly strong correlation across low-grade and high-grade cartilage (Supplementary Figs. 3, 4 and Supplementary Note 1).

**Differential genetic regulation of gene expression.** We identified differential regulation of gene expression between high-grade and low-grade cartilage, with 172 variants showing strong evidence for an eQTL effect in one tissue grade (posterior probability >0.9), but not in the other (posterior probability <0.1), termed ‘differential eQTLs’ (Fig. 2, Supplementary Figs. 5, 6, and Supplementary Data 1). We detected 32 genes with differential eQTLs (Supplementary Data 1). These genes function in the regulation of gene expression (high-grade eQTL: HOXB2, IFITM3, EIF2B3, TRAF2, HLA, APBA1, HLA; low-grade eQTL: EARS2, TCEBP1, USP16), nervous system development (high-grade eQTL: HOXB2, CRLF1, EIF2B3, APBA1, HLA, NEGR1, AREGAPI11B; low-grade eQTL: SZT2, NRN1), response to stress (high-grade eQTL: IFITM3, EIF2B3, TRAF2, ICAM3; low-grade eQTL: PNKP, SZT2, REV1, USP16), immune response (high-grade eQTL: IFITM3, IL4I1, CRLF1, TRAF2, ICAM3, HLA), cell adhesion (high-grade eQTL: TRAF2, ICAM3, APBA1, HLA, NEGR1) and catabolic processes (high-grade eQTL: IL4I1, TRAF2, WDR91, HAAO; low-grade eQTL: USP16). For most of these processes, some genes with differential eQTLs show gain of genetic regulatory associations in high-grade cartilage, while others show loss of such associations compared to low-grade cartilage, suggesting a broader rewiring of regulatory processes (see Supplementary Note 2). Sixteen genes have differential eQTLs located in a regulatory region.

**Co-localisation of GWAS signals and molecular QTLs in disease tissue.** Having established these cell- and disease-stage specific maps of molecular QTLs, we used them to identify effector genes driving GWAS signals, the majority of which reside in non-coding sequence. Co-localisation analysis can indicate whether the same variant underpins association with both disease and gene expression levels. We found strong evidence for co-localisation of five osteoarthritis signals with cartilage molQTLs for ALDH1A2, NPC1, SMAD3, FAM53A and SLC4A2 (Table 1, Fig. 3, and Supplementary Fig. 7). In all five instances, the GWAS index variant is non-coding. In three cases (ALDH1A2, SMAD3 and SLC4A2), the likely effector gene is the one closest to the lead variant. For the NPC1 and FAM53A loci, the lead variants reside in introns of the TMEM241 and SLPB genes, 141 and 18 kb away from the likely effector gene, respectively.

**Molecular hallmarks of cartilage degradation.** To identify molecular signatures associated with disease severity, we characterised gene expression and protein abundance differences between high-grade and low-grade cartilage in the largest sample set to date. We detected significant expression differences for 2557 genes and abundance differences for 2233 proteins at 5% false discovery rate (FDR) (Supplementary Figs. 8, 9, and Supplementary Notes 3, 4). A total of 409 genes (Supplementary Data 2) demonstrated significant differential expression at both the RNA and protein levels, lending robust cross-omics evidence for their involvement in disease progression. In keeping with previous smaller-scale reports, extracellular matrix (ECM)-receptor interaction was the primarily activated pathway in high-grade compared to low-grade cartilage (Supplementary Note 5, Supplementary Fig. 10, and Supplementary Data 3).

We used an independent, previously published analysis of RNA sequencing data from 35 osteoarthritis patients to replicate the observed molecular differences between high-grade and low-grade cartilage (within the power constraints of the smaller
Of the differentially expressed genes and proteins, 65.9 and 68.3% showed a concordant direction of effect in the replication data, respectively (both Fisher’s $p < 10^{-10}$, Supplementary Data 4). This concordance increased to 77.9% for genes with cross-omics concordant differential expression in this study, indicating additional robustness afforded by cross-

![Fig. 1 Study design leveraging multi-omics profiling from osteoarthritis patient tissues.](Image)

We examined the molecular characteristics of osteoarthritis by profiling mRNA and proteins from low-grade cartilage, high-grade cartilage and synovium tissue of over 100 patients undergoing total joint replacement for osteoarthritis, and combining these data with patient genotypes. We identified genetic variants influencing mRNA or protein levels, several of which co-localise with genetic risk variants for osteoarthritis. We also identified molecular markers of cartilage degeneration, creating a gene expression profile of degeneration, and shortlisting existing drugs or compounds that reverse this profile in cell experiments.

![Fig. 2 Molecular QTLs in osteoarthritis disease tissue.](Image)

49% of detected eQTLs are not tissue-specific. An example of differential QTL effect: a molecular QTL present in high-grade, but not low-grade cartilage (posterior probability $m > 0.9$ and $m < 0.1$, respectively), or vice versa. The boxplots show expression (residuals after regressing the normalised expression data on the 15 PEER factors, sex and array) at 25th to 75th percentiles, with centre at the median and whiskers extend to 1.5 times the interquartile range. NES FastQTL normalised effect size, $P$ FastQTL association $P$ value, $n$ number of individuals included in the analysis for each genotype.

Genes with $\geq 5$ differential eQTL variants (all genes see Supplementary Data 1).
omics data integration. We found significantly higher concordance where replication power was highest (88.5% for genes with cross-omics higher expression in high-grade cartilage, compared to 66.7% for genes with cross-omics lower expression in high-grade cartilage, Fisher’s $p = 8.6 \times 10^{-6}$, Supplementary Data 4).

**Differentially expressed genes with genetic associations.** Ninety-one of the genes with significantly different expression profiles between high- and low-grade cartilage were also associated with genetic risk of osteoarthritis (i.e. among the 238 genes with gene-level significant association in a recent meta-analysis$^8$; Table 1).}

### Table 1. Osteoarthritis GWAS signals with high posterior probability for co-localisation with molecular QTLs.

| GWAS variant$^a$ | rs10502437 | rs11732213 | rs12901372 | rs1560707 | rs4775006 |
|------------------|------------|------------|------------|------------|------------|
| Osteoarthritis phenotype | All        | Hip/Knee   | Hip        | All        | Knee       |
| Risk allele frequency | 0.6        | 0.81       | 0.53       | 0.37       | 0.41       |
| Odds ratio        | 1.03       | 1.06       | 1.08       | 1.04       | 1.06       |
| $P$ value          | $2.50 \times 10^{-8}$ | $8.81 \times 10^{-10}$ | $3.46 \times 10^{-11}$ | $1.35 \times 10^{-13}$ | $8.40 \times 10^{-10}$ |
| Risk allele        | A          | T          | C          | T          | A          |
| Gene              | NPC1       | FAM53A     | SMAD3      | SLC44A2    | ALDH1A2    |
| Risk allele effect on gene expression | decrease | decrease | decrease | increase | increase |

*Signals are denoted by their lead variants.

---

**Fig. 3.** *GWAS and molecular QTL $P$ values in regions with co-localisation of the associations.* GWAS and molecular QTL $P$ values in regions with co-localisation of the associations. PP4 posterior probability for co-localisation. a) P4 NPC1 eQTLs in high-grade cartilage (b) SLC44A2 eQTLs in high-grade cartilage (c) FAM53A eQTLs in low-grade cartilage (d) ALDH1A2 pQTLs in low-grade cartilage (e) SMAD3 eQTLs in high-grade cartilage. For NPC1 and SMAD3, Supplementary Fig. 7 shows co-localisation with low-grade cartilage molecular QTLs.
Supplementary Data 5). A total of 54 genes also showed concordant molecular differences in the independent replication dataset, providing further evidence for the potential involvement of these genes in osteoarthritis disease processes. For example, variants in \( ALDH1A2 \) are associated with knee osteoarthritis, and we found significantly higher \( ALDH1A2 \) gene expression and lower protein abundance in high-grade cartilage (suggesting a potential role for post-transcriptional regulation for this gene). For \( SLC39A8 \), the GWAS signal was fine-mapped with posterior probability of 0.999 to a single missense variant predicted to be deleterious by PolyPhen-2\(^{16} \), and the gene demonstrated higher expression levels in high-grade cartilage.

Candidate therapeutic compounds and drug targets. Having characterised the molecular differences between low-grade and high-grade cartilage in osteoarthritis, we sought to identify compounds capable of reversing these changes.

Using in vitro drug screen data from ConnectivityMap\(^{17} \), we identified 19 compounds that induced strong opposing gene expression signatures, reducing the expression of genes with cross-omics higher expression in high-grade cartilage (Table 2, Supplementary Data 6, and Supplementary Note 6). Several of these compounds foster known biological relevance to osteoarthritis, including the oestrogen receptor agonists diethylstilbestrol and alpha-estradiol, consistent with established epidemiological data showing an association between osteoarthritis and oestrogen deficiency\(^{18} \). Although studies of oestrogen therapy for osteoarthritis have been inconclusive\(^{19,20} \), this screening approach could facilitate further refinement of existing drug groups and allow more focussed investigational molecule development.

Further notable examples of compounds with the potential to reverse molecular changes include IB-MECA, VEGF-receptor-2-kinase-inhibitor-IV and nornicotine. IB-MECA is used as an anti-inflammatory drug in rheumatoid arthritis, and has been shown to prevent cartilage damage, osteoclast/osteophyte formation and bone destruction in a rat model of chemically-induced osteoarthritis\(^{21} \). VEGF-receptor-2-kinase-inhibitor-IV is a VEGF receptor inhibitor. VEGF modulates chondrocyte survival during development, is essential for bone formation and skeletal growth, and dysregulated in osteoarthritis\(^{22} \). A VEGFR2 kinase inhibitor has demonstrated therapeutic potential in mice\(^{23} \). Nornicotine is a demethylated derivative of nicotine found in plants, and there is a well-established epidemiological inverse relationship between smoking and osteoarthritis\(^{8,24} \). Thus, synthetic nicotine derivatives may have a candidate role in osteoarthritis prevention.

We further identified 36 genes for which knock-down or overexpression has the potential to reduce the molecular differences between high-grade and low-grade cartilage (Supplementary Data 6), for example knock-down of \( IL11 \). \( IL11 \) is a cytokine with a key role in inflammation, and several therapeutics that inhibit \( IL11 \) signalling are in development against a range of inflammatory and fibrotic diseases\(^{25,26} \). Variation in \( IL11 \) is associated with increased risk of hip osteoarthritis\(^{8} \), and the gene is upregulated in osteoarthritis knee tissue\(^{27} \), showing the most significant upregulation in high-grade cartilage in the independent replication dataset\(^{15} \) (22.8-fold higher expression, FDR = \( 1.5 \times 10^{-20} \)). These findings provide strong supportive evidence for downregulation of \( IL11 \) as a potential therapeutic intervention for osteoarthritis.

### Table 2 Compounds with strongest evidence for inducing gene expression signatures that counter differences between high-grade and low-grade cartilage, based on data from ConnectivityMap\(^{17} \) and the clue.io platform.

| Name          | Description                                      | DE targets\(^b\)         |
|---------------|--------------------------------------------------|--------------------------|
| Emetine       | protein synthesis inhibitor                      | RPS2 (P+)                |
| Rucaparib     | PARP inhibitor                                    | PARP2 (R-)               |
| Alpha-estradiol| oestrogen receptor agonist                       | KCNMA1 (R-, P-)          |
| VEGF-receptor-2-kinase-inhibitor-IV | VEGFR inhibitor                             |                          |
| IB-MECA       | adenosine receptor agonist, granulocyte colony stimulating factor agonist | GSK3B (R+, P+), CDK2 (P-) |
| Diethylstilbestrol | oestrogen receptor agonist, chloride channel blocker |                          |
| KIN01-220     | Aurora kinase inhibitor                          | GSK3B (R+, P+), CDK2 (P-) |
| SB-216763     | glycosyn thase kinase inhibitor                  | IMPDH2 (P-)              |
| RHO-kinase-inhibitor-III[rockout] | ROCK inhibitor                              |                          |
| Nornicotine   | acetylcholine receptor agonist                   |                          |

\(^a\)The 10 compounds with strongest evidence based on median of cell lines are shown; full results see Supplementary Data 6.

\(^b\)Drug targets as listed in ConnectivityMap with differential expression between high-grade and low-grade cartilage on RNA (R) or protein (P) level; ‘+’ and ‘−’ indicate higher or lower expression high-grade cartilage, respectively.

Discussion

To date, deep molecular profiling of cells of relevance to osteoarthritis has been absent from large-scale efforts such as GTEx, ENCODE and RoadMap, in part because of the challenges associated with low cellularity and high ECM content of cartilage\(^{28,29} \). Here, we address this gap by the systematic study of paired low-grade and high-grade cartilage and synovium tissues from 115 patients with osteoarthritis to provide the first deep molecular QTL map of cell types directly involved in disease. These data are made publically available (see Data availability). We identify genotype-dependent, divergent patterns of gene regulation between diseased and healthy cartilage, and between cartilage and synovium, that underline the biological specificity of disease stage and cell type when investigating regulatory variant function.

The co-localisation of GWAS signals and molecular QTLs in disease tissue helps pinpoint the identity of causal genes for hitherto unresolved association signals. A previous study\(^8 \) examined the co-localisation of osteoarthritis GWAS signals with eQTLs from the GTEx resource\(^9 \), which does not include cartilage. Several osteoarthritis GWAS signals were found to co-localise with eQTLs in only 1 or 2 of 53 tissues (e.g. \( ALDH1A2 \): ovary and tibial artery, \( SMAD3 \): skeletal muscle; \( SLC44A2 \): adrenal gland), without clear transferability of results to disease-relevant tissue. Here, we provide robust evidence that these three GWAS signals co-localise with molecular QTLs in primary cartilage. This availability of molQTL data from chondrocytes and synoviocytes offers a resource that will enable the resolution of further genetic association signals emerging from ongoing large-scale efforts in osteoarthritis (e.g. \( https://www.genetics-osteoarthritis.com \)) and further traits of musculoskeletal relevance, in which these cell types are of importance.

The strong correlation between RNA and protein level signatures within cell type and disease state and the contrast...
between disease states provides both strong internal consistency and with previous observations of ECM remodelling by chondrocytes in an inflammatory environment as a key modifiable molecular signature in the degeneration process. We further demonstrate the external validity of these hallmark signatures by showing good concordance with publically-available RNA-seq data from a smaller, independent, human osteoarthritis cohort, particularly for genes whose expression is increased in high-grade cartilage.

On a related theme, our identification of an association between 91 established osteoarthritis variants and disease state, cell-specific gene expression and protein profiles highlights the value of integrating multi-omics data with genetic association summary statistics to identify likely effector genes for GWAS signals, and hence targets for development of new therapies. The subsequent ConnectivityMap analyses built upon this theme to identify molecules with potential clinical application to reverse the molecular signatures characteristic of diseased chondrocytes, an in-silico complementary approach to conventional compound library screening. Here, we identify 19 molecules and 36 genes with high confidence for validation as candidate drugs in experimental models, such as that recently applied by Shi et al. to demonstrate that the small molecule BNTA stimulates expression of ECM components while suppressing inflammatory mediators in human osteoarthritis cartilage, an effect mediated through superoxide dismutase.

Although these data demonstrate the context-specificity of molecular indicators of disease, this work also has limitations. Cartilage remains a difficult tissue to study at the molecular level, given the low cell accessibility. The analyses were performed on extracted tissues characterised by macroscopic grading using the International Cartilage Repair Society (ICRS) scoring system and although all tissues were collected from weight-bearing areas (i.e. mechanically-loaded), a degree of cell disease state heterogeneity is inevitable in the pooled sample collection. However, the collection of both disease states from the same individual eliminates inter-individual variation as a source of confounding. Finally, although these data represent the largest cohort of its depth and kind in osteoarthritis, they are observational in nature. Future biological experiments (e.g. by gene knock-out or over-expression) will be key to prove which gene expression changes are causal in disease development and progression. The genes with convergent evidence from this study as well as osteoarthritis GWAS are presented to help such research and accelerate the pathway to translation.

In summary, by integrating multiple layers of omics data, we have generated a first molecular QTL map for osteoarthritis-relevant tissues and have helped resolve genetic association signals by identifying likely effector genes. We demonstrate how integrating multi-omics data in primary human complex disease tissue can serve as a valuable approach that moves from basic discovery to accelerated translational opportunities. Our findings identify drug repurposing opportunities and allow novel investigational avenues for therapy development, responding to the urgent clinical need of patients suffering from osteoarthritis.

### Methods

#### Study participants
We collected tissue samples from 115 patients undergoing total joint replacement surgery in 4 cohorts: 12 knee osteoarthritis patients (cohort 1; 2 women, 10 men, age 50–88 years, mean 68 years); 20 knee osteoarthritis patients (cohort 2; 14 women, 6 men, age 54–82 years, mean 70 years); 13 hip osteoarthritis patients (cohort 3; 8 women, 5 men, age 44–84 years, mean 62 years); 70 knee osteoarthritis patients (cohort 4; 42 women, 28 men, age 38–84 years, mean 70 years).

All patients provided written, informed consent prior to participation in the study. Matched low-grade and high-grade cartilage samples were collected from each patient, while synovial lining samples were collected from patients in cohorts 2 and 4. All cartilage samples were collected from weight-bearing areas of the joint to ensure that any differences observed between low- and high-grade cartilage reflect disease progression stage rather than differential biomechanical stress.

**Cohorts 1, 2, 4 (knee osteoarthritis).** This work was approved by Oxford NHS REC C (10/H0608/20 and 15/SC/0132), and samples were collected under Human Tissue Authority license 12182, Sheffield Musculoskeletal Biobank, University of Sheffield, UK.

We confirmed a joint replacement for osteoarthritis, with no history of significant knee surgery (apart from meniscectomy), knee infection, or fracture, and no malignancy within the previous 5 years. We further confirmed that no patient used glucocorticoid use (systemic or intra-articular) within the previous 6 months, or any other drug associated with immune modulation. For cohort 1, from each patient, we obtained one sample with high OARSI grade signifying high-grade degeneration (‘high-grade sample’), and one cartilage sample with low OARSI grade signifying healthy tissue or low-grade degeneration (‘low-grade sample’).

For cohorts 2 and 4, cartilage samples were scored macroscopically using the ICRS scoring system. From each patient, we obtained one sample of ICRS grade 3 or 4 signifying high-grade degeneration (‘high-grade sample’), and one cartilage sample of ICRS grade 0 or 1 signifying healthy tissue or low-grade degeneration (‘low-grade sample’). For cohorts 2 and 4, we also collected synovial membrane from the suprapatellar region of the knee joint.

Finally, from all patients in cohorts 1, 2, and 4, we also obtained a blood sample to extract DNA for genotyping.

**Cohort 3 (hip osteoarthritis).** Samples were collected under National Research Ethics Service approval reference 11/EE/0011, Cambridge Biomedical Research Centre Human Research Tissue Bank, Cambridge University Hospitals, UK.

We confirmed osteoarthritis disease status by examination of the excised femoral head. From each patient, we obtained a cartilage sample showing a fibrillated or fissured surface signifying high-grade degeneration (‘high-grade sample’), one cartilage sample showing a smooth shiny appearance signifying healthy tissue or low-grade degeneration (‘low-grade sample’).

#### Isolation of chondrocytes.

For cohorts 1, 2, and 4, we followed a previously established protocol to isolate chondrocytes with the details as follows. Osteochondral samples were transported in Dulbecco’s modified Eagle’s medium (D-MEM/F-12 (1:1)) (Life Technologies) supplemented with 2 mM glutamine (Life Technologies), 100 U/ml penicillin, 100 μg/ml streptomycin (Life Technologies), 2.5 μg/ml amphotericin B (Sigma-Aldrich) and 50 μg/ml ascorbic acid (Sigma-Aldrich) (serum free media). Half of each sample was then taken forward for chondrocyte extraction. Cartilage was removed from the bone, dissected and washed twice in 1xPBS. Tissue was digested in 3 mg/ml collagenase type I (Sigma-Aldrich) in serum free media overnight at 37 °C on a shaker. The resulting cell suspension was passed through a 70 μm cell strainer (Fisher Scientific) and centrifuged at 400×g for 10 min. Subsequently, the cell pellet was washed twice in serum free media and centrifuged at 400×g for 10 min. The resulting cell pellet was resuspended in serum free media. Cells were counted using a haemocytometer and the viability checked using trypan blue exclusion (Invitrogen). The optimal cell number for spin column extraction from cells was between 4×10^6 and 1×10^7. Cells were then pelleted and homogenised.

For cohort 3, the extraction of chondrocytes has previously been described in the majorities of these samples, with the remaining samples following the same protocol. The protocol was based on that for cohorts 1, 2, and 4 and highly similar as described in the following. Each cartilage portion was minced with a scalpel and placed in 20 ml of Dulbecco’s modified Eagle medium (Invitrogen) containing 10% foetal bovine serum (Invitrogen) and 6 mg/ml bovine collagenase A (Sigma). The tissue culture flasks were incubated overnight to digest the cartilage pieces. The resulting cell suspension was passed through a 30 μm filter (Millipore) and centrifuged at 400×g for 10 min. The cell pellet was then re-suspended in 1 ml of PBS and counted on a haemocytometer following 1:1 mixing with trypan blue to determine cell viability.

#### Isolation of synoviocytes.

We followed a previously established protocol to process synovial samples, with details as follows. Synovial samples were transported in serum free media, as described above. The synovial membrane was dissected out from underlying tissue then trypsinised for 1 h. Tissue was then digested in 1 mg/ml Collagenase Blend H (Sigma Aldrich) in serum free media overnight at 37°C on a shaker. The resulting cell suspension was passed through a 100 μm cell strainer (Fisher Scientific) and centrifuged at 400×g for 10 min. Subsequently, the cell pellet was washed twice in serum free media and centrifuged at 400×g for 10 min. The resulting cell pellet was resuspended in serum free media. Cells were counted using a haemocytometer and the viability checked using trypan blue exclusion (Invitrogen). The optimal cell number for spin column extraction from cells was between 4×10^6 and 1×10^7. Cells were then pelleted and homogenised.
DNA, RNA and protein extraction. DNA, RNA, and protein extraction was carried out using Qiagen AllPrep DNA/RNA/Protein Mini Kit following the manufacturer's instructions for cohorts 1 and 2. For cohort 3 as previously described35 and recapitulated in the Supplementary Methods. Samples were frozen at −80 °C (cohorts 1, 2, 4) or −70 °C (cohort 3) prior to assays.

RNA sequencing. We performed a gene expression analysis on samples from 113 patients (Supplementary Data 7). We purified poly-A tailed RNA (mRNA) from total RNA using Illumina v1.5 TruSeq RNA Sample Prep v2 kits. We then fragmented the mRNA using metal ion-catalysed hydrolysis and synthesised a random-primed cDNA library. The resulting double-strand cDNA was used as the input to a standard Illumina library prep, whereby ends were repaired to produce blunt ends by a combination of fill-in reactions and exonuclease activity. We performed A-tailing to allow samples to be pooled, by adding a 3′-A tail to the blunt ends and ligation to Illumina Paired-end Sequencing adaptors containing unique index sequences. Due to better performance, the 10-cycle PCR amplification of libraries was carried out using KAPA HiFi Polymerase. A post-PCR Agilent Bioanalyzer was used to quantify samples, followed by sample pooling and size-selection of pools using the Agilent 2100 Bioanalyzer. The multiplexed libraries were sequenced on the Illumina HiSeq 2000 for cohort 1 and HiSeq 4000 for cohorts 2–4 (75 bp paired-ends). Sequenced data underwent initial analysis and quality control (QC) on reads as standard. The sequencing depth was similar across samples, with 90% of samples passing final QC (see below) having 87.2–129.2 million reads.

Proteomics. Proteomics analysis was performed on cartilage samples from 103 patients (Supplementary Data 7). For cohort 1, all steps of protein digestion, 6-plex TMT labelling, peptide fractionation and LC-MS analysis on the Dionex Ultimate 3000 UHPLC system coupled with the high-resolution LTQ Orbitrap Velos mass spectrometer (Thermo Scientific), were previously described25 and are recapitulated in the Supplementary Methods. The sample preparation protocol formed the basis of processing for cohorts 2–4 using 10-plex TMT labelling and an Orbitrap Fusion Tribrid Mass Spectrometer (Thermo Scientific) with otherwise only minor alterations as described in the Supplementary Methods.

Genotyping. We used Illumina HumanCoreExome-12v1-1 for genotyping cohort 1 and Illumina InfiniumCoreExome-24v1-1 for genotyping cohort 2–4 patients.

Quantification of RNA levels. We used samtools v1.3.17 and biobambam v0.191.18 to convert cram to fastq files after exclusion of reads that failed QC. We applied FastQC v0.11.5 to check sample quality30 and excluded nine samples (Supplementary Data 7).

We obtained transcript-level quantification using salmon 0.8.20 (with-gcBias and --seqBias flags to account for potential biases) and the GRCh38 cDNA assembly release 87 downloaded from Ensembl [http://ftp.ensembl.org/pub/release-87/fasta/homo_sapiens/cdna/]. We used topquartile41 to convert transcript-level to gene-level to gene-level scaled transcripts per million (TPM) estimates, with estimates for 39,037 genes based on Ensembl gene IDs.

We excluded four samples due to low mapping rate (<80%), three samples due to non-European ancestry recorded in the clinic, 18 samples due to low RIN (<5), two samples as duplicates, eight samples due to abnormal gene read density plots (detected separately in cartilage and synovium for three cartilage and two samples as duplicates, eight samples due to low RIN (<5), we excluded samples from four patients due to non-European ancestry (Supplementary Data 7). The final dataset included low-grade and high-grade cartilage samples each from 99 patients, with 4801 proteins was observed in ≥20% of samples, and 1677 proteins in all samples, in line with the resolution depth of the isobaric labelling method employed. 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 excluded samples from four patients due to non-European ancestry (Supplementary Data 7). The final dataset included low-grade and high-grade cartilage samples each from 99 patients, with 4801 proteins was observed in ≥20% of samples, and 1677 proteins in all samples, in line with the resolution depth of the isobaric labelling method employed. 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.

Genotype analysis and QC. Genotypes were called using GenCall (Illumina) and mapped to GRCh37/hg19 using online tools (http://www.well.ox.ac.uk/~wrayner/strand/index.html). QC was carried out using the same method for both arrays. Brieﬂy, we performed a pre-ﬁltering step to exclude samples and variants with a call rate <90%. Sample QC included identity checks correlating the array genotypes to Fluidigm genotypes obtained at sample reception (no samples had a concordance <0.95). We excluded samples based on call rate <98%, heterozygosity >0.7 or high allele frequency in less than one thinning interval. In addition, we used PLINK (http://www.csg.sfu.ca更能津／wrayner/tools/; v4.2.7; the resulting dataset contained 111 patients and 389,511 variants. We imputed QC samples and variants with poor imputation quality (R^2 < 0.5) and pHWE < 10^{-4}. We excluded two patients due to absence of RNA and protein data. The resulting final dataset contained 10,249,108 autosomal variants and 109 patients.

Identification of cis-eQTLs. For each gene, we considered genetic variants within 1 Mb of the transcription start site (TSS) (definition see below), and followed a similar method to GTEX47 as follows.

For each tissue, we included only genes with ≥1 count per million in at least 20% samples and we normalised between samples using TMM (weighted trimmed mean of M values)46 implemented in edgeR. To facilitate cartilage comparisons problems and exclusion of the previous two steps (exclusions of low expressed genes and the between sample normalisation) were performed with high-grade and low-grade cartilage samples combined. For each tissue separately, we then normalised across samples using an inverse normalisation transformation for each gene. We inferred hidden factors associated with cohort, sequencing batch, or other technical differences using applied probabilistic estimation of expression residuals (PEER)30 separately to each tissue (PEER C + + with standard parameters from the R version, i.e. iteration = 1000, bound = 0.001, variance = 0.00001, Alpha a = 0.001, Alpha b = 0.1, Eps a = 0.1, Eps b = 10). We used the GTEX modified version of FastQTL (https://github.com/francois-a-fastqtl/v6p), which allows for minor allele count filtering, reporting of MAF and calculation of FDR. We determined the TSS for each gene using empirical transcript level expression information from synovium, high-grade and low-grade cartilage samples (see below) and defined the cis-mapping region to be 1 Mb in either direction from the TSS. We restricted the analysis to variants with minor allele count of at least 10 in a given tissue. Nominal p-values for each gene were based on linear mixed-effects models using 15 PEER factors for the given tissue, sex and genotype array as covariates. We then employed the adaptive permutation scheme with the—permute 1000 10000 option to generate empirical p values. Genes with significant eQTLs (‘eGenes’) were defined at the 3% Storey–Tibshirani FDR using the q values generated from the empirical p values35. For each eGene, significant eQTLs were defined as variants with nominal p value below the nominal p value threshold for that gene generated in FastQTL.

The normalised effect size (NES) of the eQTL is reported for the alternate allele according to GRCh37/hg19.

Identification of cis-pQTLs. We followed a similar protocol as for cis-eQTL analysis, also considering genetic variants within 1 Mb of the TSS (definition see below). For low-grade and high-grade cartilage samples, we included 1677 genes that were measured across all samples. We normalised across samples using an inverse normalisation transformation for each gene separately in each tissue. To account
for possible technical variation, we used PEER50 (with parameters as for the eQTL analysis above) and included 26 PEER factors, sex and genotype array as covariates using the GTEx modified version of FastQTL (https://github.com/francosi-a/fastqtl; v6p). We used the TSS established for the eQTL analysis, yielding a unique mapping for 1461 proteins, which were then taken forward. For each protein, we considered variants within a 1 Mb region in either direction from the TSS, restricting further to minor allele count of 10 or higher. We then followed the same procedure as for cis-eQTLs to identify variant-protein pairs with significant cis-pQTL effects.

We carried out a protein–protein network analysis for the proteins with significant pQTL effects using the STRING v11.0 database23 (https://string-db.org; see Supplementary Note 1).

Sensitivity analysis for cis-eQTLs and cis-pQTLs. For both eQTLs and pQTLs, we verified that the results were robust by carrying out a sensitivity analysis including patient age and osteoarthritis joint (knee or hip) as covariates in addition to the PEER factors, sex and array (see Supplementary Note 1).

Transcription start site (TSS) definition. To determine which transcript to use to define the TSS for each gene, we established the most abundant transcript for each gene. To have the same definition across all tissues, we analysed cartilage and synovium tissues jointly, considering 16,886 genes that passed quantification QC in at least one tissue (based on 15,249 genes in cartilage and 16,004 genes in synovium). For each transcript, we calculated the expression in each sample as scaled TPM using tximport41. For each gene, we then obtained the most abundant transcript in each tissue in each patient, and calculated the proportion of samples in which each transcript was the most abundant. The transcript that was the most abundant in the largest proportion of samples was used to define the TSS for the gene. For genes in which more than one transcript was the most abundant, we chose one of the most abundant transcripts at random to define the TSS. Across all genes, 47.9% had the same most abundant transcript in at least 90% of samples in both cartilage and synovium; 71% of genes had the same most abundant transcript in 60% of samples in cartilage and synovium. We mapped the most abundant transcript (using the ENST identifier) to GRCh37 using ftp://ftp.ensembl.org/pub/ghrch37/release-87/gtf/homo_sapiens/Homo_sapiens.GRCh37.87.chr.gtf.gz. For 93 transcripts, ENSG identifiers differed between the builds, therefore genes were identified by a composite ID in the form (GeneName@b37). We excluded 1940 transcripts with missing start or end positions (largely in patched genome build regions), and ~500 transcripts mapped to chromosomes X, Y or mitochondrial DNA, we established the TSS for the 13,180 autosomal genes included in the cartilage and 13,708 genes included in the synovium molQTL analysis.

Differential gene regulation in low-grade and high-grade cartilage. To identify differential regulation of gene expression between high- and low-grade cartilage, we used Meta-Tissue v6.54 (downloaded from http://metatissue.cs.ucl.ac.uk:8080/metatissue/index.html), which implements METASOFT55. The m value calculated by METASOFT for gene-variant pair in each tissue provides a posterior probability (m value) of an effect in that tissue. Consequently, we aimed to identify eQTLs present in both tissues (m > 0.9, defined as cis-eQTLs), and eQTLs defined in one but not the other (defined as trans-eQTLs). We note that there were no cis-eQTLs present in both tissues (m > 0.9) with opposing direction of effect.

Meta-Tissue restricts covariates input to the same values for each patient across tissues, while different PEER covariates were provided for each tissue in the FastQTL analysis. Hence, for each tissue, we obtained residuals from regressing the normalised expression data on the 15 PEER factors, sex and array, then used the residuals as input for Meta-Tissue. We included genotype dosages based on both the low- and high-grade results for each analysis. We ran METASOFT using the default settings provided in the output script from Meta-Tissue. For 32 genes with differential QTLs, we considered gene annotations from several sources: using Gene Ontology biological process terms, through gene annotations in GeneCards (accessed on 31 May 2020), and from the summaries and functions sections of the Ensembl gene ID and gene name pages (https://www.genecards.org, accessed 3 June 2020). We also carried out a formal enrichment analysis of the 32 genes using GoSeq (see Supplementary Note 2).

Co-localisation between molecular QTLs (molQTLs) and osteoarthritis GWAS associations. To examine co-localisation between molQTLs and GWAS associations, we used gene-centric summary statistics from the largest osteoarthritis meta-analysis to date, based on UK Biobank and arcOGEN data9. We analysed all 64 genome-wide significant signals using colo57, separately for each tissue and omics level.

In the co-localisation analysis for each signal, we considered the region spanning 100 kb on either side of the index variants. If that region overlapped any genes with significant cis-eQTLs or cis-pQTLs, we extended the region to encompass all variants included in the molQTL analysis for these genes. To formally obtain a posterior probability for co-localisation, we used colo.fast (https://github.com/tobyjohnson/git/blob/526120453b6ec39c71604eece03a77e57353/R/colo.fast), a Bayesian statistical test which implements the coloc27 method. We used the default settings for colo.fast. We considered a 80% posterior probability of GWAS and molQTL shared association at a single variant (PP4 ≥ 0.8) to indicate evidence of co-localisation.

Differential RNA expression between high-grade and low-grade cartilage. We tested differential expression of 15,249 genes between high-grade and low-grade cartilage using paired samples from 83 patients. To detect robust gene expression differences, we carried out analyses using different software packages as recommended in a landmark survey of best practices8, applying limma59, edgeR60 and DESeq261. We also tested five analysis designs with different options to account for technical variation, including SVAsed55. In particular, we tested for differential expression using (i) a paired analysis of intact and degraded samples (i.e. specifying patient ID as covariate); (ii) a paired analysis of intact and degraded samples, with ten additional covariates accounting for technical variation identified by SVAsed55; (iii) a paired analysis of intact and degraded samples, with ten RNA sequencing batches as covariates; (iv) an unpaired analysis of intact and degraded samples; (v) an unpaired analysis of intact and degraded samples, with 19 additional covariates accounting for technical variation identified by SVAsed. We tested for differential expression using the following R packages: (i) limma59 (with fitInT and fitEx), after applying limma-voom62 to remove heteroscedasticity; (ii) DESeq261, separately with and without outlier filtering/replacement (mninReplacementsForReplace = Inf, cooksCutOff = FALSE options); (iii) edgeR60, using the likelihood ratio test (glmFit and glmLRT functions), and separately using the F test (glmQFLet and glmQFTest functions).

Here and elsewhere, we used Ensembl38p10 to identify genes with uniquely corresponding Ensembl gene ID and gene name (13,737 of 15,249 genes in the RNA data).

In each analysis design and method, we used a 5% FDR threshold to correct for multiple testing. As the final step, we applied a conservative approach and considered a gene ‘significantly differentially expressed’ between low-grade and high-grade cartilage if it showed significant differential expression across all analysis designs and testing methods (2557 genes, including 2418 with uniquely corresponding Ensembl gene ID and gene name). As discussed in the Supplementary Note 3, for each analysis design, the number of significant genes at 5% FDR was similar across all tests. Furthermore, we have performed a sensitivity analysis testing differential expression based on the patients in cohort 4 only and verified that results were highly consistent with the analysis of all patients (see Supplementary Note 3).

We used the pheatmap v1.0.12 function in R to plot the gene expression for the 20 genes with the highest absolute log-fold differences between high-grade and low-grade cartilage (selected from 290 genes with significant and concordant cross-omics differences, see Supplementary Methods).

Protein expression abundance between high-grade and low-grade cartilage. We performed differential analysis for 4801 proteins that were measured in ≥30% of patients, applying limma27 to paired samples from 99 patients. Significance was defined at 5% FDR to correct for multiple testing, yielding 2333 proteins with significant differential abundance (2019 proteins with uniquely corresponding Ensembl gene ID and gene name).

As batch effects in proteomics data can be pervasive44,45, paired samples from any patient were always assayed in the same 6-plex (cohort 1) or 10-plex (cohorts 2–4). As a sensitivity analysis, we carried out a differential abundance analysis for the proteomics data with explicit adjustment for the plexes, to confirm that adjustment for patient effects was captured between-plex batch effects. For each protein, we calculated the log2 of normalised abundance values plus 1. We then obtained residuals from linear regression of these values on the 13 batches used in the proteomics data. These residuals were quantile normalised. We then used limma to test for differential abundance between low-grade and high-grade cartilage using the quantile normalised residuals, applying a paired design and the same method as in the main analysis. The results were highly similar to the main analysis (see Supplementary Note 4). We used the pheatmap v1.0.12 function in R to plot the abundance of the 20 proteins with the highest absolute log-fold differences between high-grade and low-grade cartilage (selected from 290 genes with significant and concordant cross-omics differences, see Supplementary Methods).

Replication of differences between high-grade and low-grade cartilage. We used an independent dataset from the RAAK study (35 osteoarthritis patients, of which 28 knee, 7 hip osteoarthritis) to replicate the molecular differences between high-grade and low-grade cartilage. The RAAK dataset and analysis were taken from a previous publication15. Briefly, mRNA samples were sequenced on Illumina HiSeq 2000/4000 (paired-end 2 x 100 bp RNA-sequencing), aligned using
From the recent UK Biobank and arcOGEN GWAS meta-analysis\(^8\), we obtained Identi
Supplementary Note 5 and Supplementary Data 3).

Pathway associations for differences between high-grade and low-grade cartilage. To identify the biological processes with significant molecular differences between high-grade and low-grade cartilage, we carried out gene set enrichment analyses based on the differential expression on RNA, protein and cross-omics levels. We tested for association of the differentially expressed (DE) genes on RNA and/or protein levels at 5% FDR, with robustness checks using more stringent FDR thresholds (1, 0.5 and 0.1%). We restricted this analysis to genes with unique mapping between Ensembl gene ID and Gene Name in Ensembl\(^{38,10}\), and Ensembl gene ID and Entrez ID in HUGO (https://www.
genenames.org/ accessed 05/03/2018; 13,094 genes measured on RNA level, 4390 genes measured on protein level, 4387 genes measured on both RNA and protein levels).

We applied Signalling Pathway Impact Analysis (SPIA)\(^{36}\) to test for association with KEGG signalling pathways. SPIA combines enrichment p values with perturbation impact on the pathway based on log-fold differences of the DE genes; perturbation impact values are obtained by bootstrapping. Enrichment and perturbation p values were combined using a normal inversion method which only gives low p values when both over-representation and pathway impact p values are low (function option combine = ‘norminv’). Significance of pathway association was defined as a threshold of 5% FDR applied to the combined p values in each analysis. For the analysis of genes DE on both RNA and protein levels, we carried out tests using the log-fold differences from the RNA data (based on the limma analysis with paired samples and SVAeq covariates), and separately, from the protein data. We also tested enrichment in Gene Ontology terms using GOseq\(^{35}\), separately for genes with higher or lower expression in high-grade compared to low-grade cartilage. We accounted for gene length (pw function options ‘hg19’ and ‘geneSymbol’). Significance was defined as a threshold of 5% FDR in each analysis. The results showed broad agreement with the results of the SPIA analysis (see Supplementary Note 5 and Supplementary Data 3).

Identification of genes with osteoarthritis GWAS gene-level association. From the recent UK Biobank and arcOGEN GWAS meta-analysis\(^{46}\), we obtained the results of a gene-level analysis for each of the four osteoarthritis phenotypes (self-reported plus hospital diagnosed, hospital diagnosed knee, hospital diagnosed hip, hospital diagnosed knee, hospital diagnosed hip), as described in the GWAS paper. Briefly, this analysis used MAGMA v1.066\(^{67}\) and was based on the mean SNP log−p-value in the gene, accounting for LD.

To calculate the effective number of tests across phenotypes, we calculated the correlation matrix between the gene p values for the four osteoarthritis phenotypes, and obtained the eigenvalues of this matrix. The effective number of tests \(N_{\text{eff}}\) for phenotypes was then calculated as \(N_{\text{eff}} = N - \sum I(\lambda > 1)\)\(^{-1}\), where \(N = 4\) is the number of phenotypes, and \(\lambda\) denotes the eigenvalues. Across the Pearson and Spearman correlation matrices, we obtained \(N_{\text{eff}} < 2.65\). With 18,449 genes per phenotype, the significance threshold for gene-level p values across genes and phenotype was therefore \(0.05/18,449 \times 2.65\) = 0.0005 (\(2.0 \times 10^{-6}\)).

After accounting for the effective number of tests across phenotypes and genes using a Bonferroni correction, 320 of 18,449 genes showed significant association with at least one phenotype. Of these genes, 238 genes were compared between low-grade and high-grade cartilage on at least one omics level and had uniquely corresponding Ensembl gene ID and gene name.

ConnectivityMap analysis. To identify opportunities for drug repurposing, we used ConnectivityMap\(^{13}\) to identify compounds and perturbation classes (PCLs) that could possibly reverse the differences identified between high-grade and low-grade cartilage. Using the online interface clue.io (accessed 3 March 2019), we submitted the 148 genes with significantly higher expression on both RNA and protein level to calculate a ‘tau’ connectivity score to gene expression signatures experimentally induced by various perturbations in nine cell lines. A positive tau score indicates similarity between the gene expression signature of a perturbation and the submitted query (i.e. upregulation of the genes with higher expression in high-grade compared to low-grade cartilage). A negative tau score indicates that gene expression signature of a perturbation opposes the submitted query (i.e. downregulation of the genes with higher expression in high-grade compared to low-grade cartilage). Recommended thresholds for further consideration of results are tau of at least 90, or below −90, respectively (https://clue.io/connectomepa
connectivity_scores, accessed 3 March 2019). A total of 2857 compound and 171 PCL perturbations were evaluated in clue.io. We shortlisted perturbations where both the summary tau and the median tau across cell lines were higher than 90 or lower than −90 for PCLs, with more conservative thresholds of higher than 95 or lower than −95 for compounds. The clue.io platform also contained perturbation data from 3799 gene knock-down and 2160 over-expression experiments (with 2111 genes in both, i.e. 3848 genes total). These data were used to shortlist genes where both the summary and median tau were higher than 95 or lower than −95.

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

Data availability

The RNA sequencing data reported in this paper have been deposited to the EGA (accession numbers EGAD00001005215, EGAD00001003355, EGAD00001003354, EGAD00001001331). The proteomics data from this paper have been deposited to PRIDE (accession numbers PXD014666, PXD006673, PXD030114). The genotype data reported in this paper have been deposited to the EGA (accession numbers EGAD00001001746, EGAD0000100285, EGAD00001001292, EGAD000010000722). Data from the 1000 Genomes Project are publicly available (https://www.
internationalgenome.org). We also used publicly available data on osteoarthritis differential gene expression from the RAAK Study (https://git.lumc.nl/
rcoutinhodealmeida/miRNAmRNA, accessed 20 June 2020).

Further data including the TSS information, all significant mQTLs and full co-localisation results, can be obtained online from https://hmgbox.helmholtz-muenchen.
de/d1fc16f5a6724152b789/. The full molecular QTL data and molecular differences between high-grade and low-grade cartilage are available through the Downloads page of the Musculoskeletal Knowledge Portal (mskp.org).

Code availability

All software used in this study is available from free repositories or from manufacturers as referenced in the Methods section.

Received: 9 July 2020; Accepted: 3 February 2021; Published online: 26 February 2021

References

1. GRD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 396, 1204–1222 (2020).

2. Murphy, L. et al. Lifetime risk of symptomatic knee osteoarthritis. Arthritis Rheum. 59, 1207–1217 (2008).

3. Murphy, L. B. et al. One in four people may develop symptomatic hip osteoarthritis in his or her lifetime. Osteoarthr. Cartil. 18, 1372–1379 (2010).

4. Murphy, L. B., Cisternas, M. G., Pasta, D. J., Helmsick, C. G. & Yelin, E. H. Medical expenditures and earnings losses among US dults with arthritis in 2013. Arthritis Care Res. 70, 869–876 (2018).

5. Toto, C. M. & Moore, B. J. Statistical Brief #204, National Inpatient Hospital Costs: The Most Expensive Conditions by Payor, 2013. (Agency for Healthcare Research and Quality, 2016).

6. Nüesch, E. et al. All cause and disease specific mortality in patients with knee or hip osteoarthritis: population based cohort study. BMJ 342, d1165 (2011).

7. Spector, T. D. & MacGregor, A. J. Risk factors for osteoarthritis: genetics. Osteoarthr. Cartil. 12, 39–44 (2004).

8. Tachmazidou, I. et al. Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. Nat. Genet. 51, 230–236 (2019).

9. GTEx, Consortium et al. Genetic effects on gene expression across human tissues. Nature 550, 204–212 (2017).

10. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

11. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).

12. Steinberg, J. et al. Integrative epigenomics, transcriptomics and proteomics of patient chondrocytes reveal genes and pathways involved in osteoarthritis. Sci. Rep. 7, 8935 (2017).

13. Karlsson, C. et al. Genome-wide expression profiling reveals new candidate genes associated with osteoarthritis. Osteoarthr. Cartil. 18, 581–592 (2010).
14. Ramos, V. F. M. et al. Genes involved in the osteoarthritis process identified through genome wide expression analysis in articular cartilage, the RAALK Study. PLoS ONE 9, e103056 (2014).

15. Coutinho de Almeida, R. et al. RNA sequencing data integration reveals an miRNA interactome of osteoarthritic cartilage. *Ann. Rheum. Dis.* 78, 270–277 (2019).

16. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* https://doi.org/10.1002/9780471142905.hg0720s76 (2013).

17. Subramanian, A. et al. A next generation connectivity map: 1,000 platform and the first 1,000,000 profiles. *Cell* 171, 1437–1452.e1417 (2017).

18. Roman-Blas, J. A., Castañeda, S., Largo, R. & Herrero-Beaumont, G. Osteoarthritis associated with estrogen deficiency. *Arthritis Res. Ther.* 11, 241 (2009).

19. de Klerk, B. M. et al. Limited evidence for a protective effect of unopposed oestrogen therapy for osteoarthritis of the hip: a systematic review. *Rheumatol. 48*, 104–112 (2009).

20. Watt, F. E. Hand osteoarthritis, menopause and menopausal hormone therapy. *Maturitas* 83, 13–18 (2016).

21. Bar-Yehuda, S. et al. Induction of an anti-inflammatory effect and prevention of cartilage damage in rat knee osteoarthritis by CF101 treatment. *Arthritis Rheumatol.* 60, 3061–3071 (2009).

22. Yuan, Q., Sun, L., Li, J.-J. & An, C.-H. Elevated VEGF levels contribute to the pathogenesis of osteoarthritis. *BMJ Musculoskelet. Disord.* 15, 43 (2014).

23. Nogue, M. et al. Vascular endothelial growth factor in cartilage development and osteoarthritis. *Sci. Rep.* 7, 13027 (2017).

24. Kong, L., Wang, L., Meng, F., Cao, J. & Shen, Y. Association between smoking and the risk of knee osteoarthritis: a systematic review and meta-analysis. *Osteoarthr. Cartil.* 28, 1695–1700 (2020).

25. Chou, C. H. et al. Insights into osteoarthritis progression revealed by analyses of both knee tibiofemoral compartments. *Osteoarthr. Cartil.* 23, 571–580 (2015).

26. Ruetten, A., Neumann, S., Wiederanders, B. & Huber, R. Comparison of different methods for preparation and characterization of total RNA from cartilage samples to uncover osteoarthritis in vivo. *BMC Res. Notes* 3, 7 (2010).

27. Le Bleu, H. K. et al. Extraction of high-quality RNA from human articular cartilage. *Anal. Biochem.* 518, 134–138 (2017).

28. Shi, Y. et al. A small molecule promotes cartilage extracellular matrix generation and inhibits osteoarthritis development. *Nat. Commun.* 10, 1914 (2019).

29. Malondono, M. & Nam, J. The role of changes in extracellular matrix of cartilage with the presence of inflammation on the pathology of osteoarthritis. *Biomed. Res. Int.* 2013, 284873 (2013).

30. Mainil-Varlet, P. et al. Histological assessment of cartilage repair: a report by the RAAK Study. *J. Bone Joint Surg. Am.* 85-A Suppl. 2, 45–57 (2003).

31. Mankin, H. J., Dorfman, H., Lippiello, L. & Zarins, A. Biochemical and II. Correlation of morphology with biochemical and metabolic data. *J. Bone Joint Surg. Am.* 331 (2011).

32. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq accounting for selection bias. *Genome Biol.* 11, R4 (2010).

33. Giannhartoumi, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 9, e1003491 (2013).

34. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17, 13 (2016).

35. Ritchie, M. E. et al. Imma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e17 (2015).

36. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4289–4297 (2012).

37. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data using DESeq2. *Genome Biol.* 15, 550 (2014).

38. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* 42, e161 (2014).

39. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock variance estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* 7, 500–507 (2012).

40. Ongen, H., Bul, A., Brown, A. A., Dermitzakis, E. T. & Deloue, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 32, 1479–1485 (2016).

41. Sorensen, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* 100, 9440–9445 (2003).

42. Sklarcky, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613 (2019).

43. Sul, J. H., Han, B., Ye, C., Choi, T. & Eskim, E. Effectively identifying eQTLs from multiple tissues using quantized mixed model and meta-analytic approaches. *PLoS Genet.* 9, e1003491 (2013).

44. Han, B. & Eskim, E. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet.* 8, e1002555 (2012).

45. McCarthy, M. D., Smyth, G. K. & Osb, A. Gene ontology analysis for RNA-seq accounting for selection bias. *Genome Biol.* 11, R4 (2010).

46. Gregori, J. et al. Batch effects correction improves the sensitivity of significance tests in spectral counting based comparative discovery proteomics. *J. Proteome. 75*, 3938–3951 (2012).

47. Kelkiewicz, J. et al. Detection of batch effects in liquid chromatography-mass spectrometry metabolomic data using guided principal component analysis. *Talanta* 130, 442–448 (2014).

48. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875 (2005).

49. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169 (2015).

50. Tarca, A. L. et al. A novel signaling pathway impact analysis. *Bioinformatics* 25, 75–82 (2008).

51. de Leuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* 11, e1004219 (2015).

Acknowledgements
We thank the study participants who made this work possible by their generous donation of samples. This research was conducted using the UK Biobank Resource under application number 9979. The authors are grateful to Dr. Iris Fischer for helpful edits. We are grateful to the RAAK study for making their data available. This work was funded by the Wellcome Trust (206194). M.I.C. was funded through a Medical Research Council Centre for Integrated Research into Musculoskeletal Ageing grant (149865). R.A.B. and the Human Research Tissue Bank are supported by the NIHR Cambridge Biomedical Research Centre. J.H.D.B. and C.R.W. are funded by a Wellcome Trust Strategic Award (101123), a Wellcome Trust Joint Investigator Award (110140 and 110141) and a
European Commission Horizon 2020 Grant (666869, THYRAGE). A.W.M. receives funding from Versus Arthritis; Tissue Engineering and Regenerative Therapies Centre (21156).

**Author contributions**
Study design: E.Z., J.M.W. and J.S. Collection of knee samples: M.J.C., R.L.J., D.S., K.S. and J.M.W. Collection of hip samples: R.A.B. and A.W.M. Proteomics assays: T.I.R. and J.S.C. Molecular QTL and co-localisation analyses: L.S. Differential expression analyses: J.S. and L.S. Pathway association and drug repurposing analyses: J.S. Writing—original draft: J.S., L.S., N.B., J.H.D.B., G.R.W., J.M.W. and E.Z. Writing—comments and review: all authors.

**Funding**
Open Access funding enabled and organized by Projekt DEAL.

**Competing interests**
The authors declare no competing interests.

**Additional information**
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21593-7.

Correspondence and requests for materials should be addressed to J.M.W. or E.Z.

**Peer review information** Nature Communications thanks Caroline Ospelt and the other anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at http://www.nature.com/reprints

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021, corrected publication 2021