Decoding the genomic basis of osteoarthritis

Julia Steinberg[1,2,3], Lorraine Southam[1,3], Natalie C Butterfield[4], Theodoros I Roumeliotis[3,5], Andreas Fontalis[6], Matthew J Clark[6], Raveen L Jayasuriya[6], Diane Swift[6], Karan M Shah[6], Katherine F Curry[4], Roger A Brooks[7], Andrew W McCaskie[7], Christopher J. Lelliott[3], Jyoti S Choudhary[3,5], JH Duncan Bassett[4], Graham R Williams[4], J Mark Wilkinson[6,8,9], Eleftheria Zeggini[1,3,9]

[1] Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, 85764 Neuherberg, Germany
[2] Cancer Research Division, Cancer Council NSW, Sydney, New South Wales 2000, Australia
[3] Wellcome Sanger Institute, Hinxton CB10 1SA, United Kingdom
[4] Molecular Endocrinology Laboratory, Department of Metabolism, Digestion and Reproduction, Imperial College London, London W12 ONN, UK
[5] The Institute of Cancer Research, London SW7 3RP, UK
[6] Department of Oncology and Metabolism, University of Sheffield, Sheffield S10 2RX, UK
[7] Division of Trauma & Orthopaedic Surgery, Department of Surgery, University of Cambridge, Cambridge CB2 2QQ, UK
[8] Centre for Integrated Research into Musculoskeletal Ageing and Sheffield Healthy Lifespan Institute, University of Sheffield, Sheffield S10 2TN, UK
[9] These authors contributed equally

Correspondence to: eleftheria.zeggini@helmholtz-muenchen.de and j.m.wilkinson@sheffield.ac.uk

ABSTRACT

Osteoarthritis is a serious joint disease that causes pain and functional disability for a quarter of a billion people worldwide[1], with no disease-stratifying tools nor modifying therapy. Here, we use primary chondrocytes, synoviocytes and peripheral blood from patients with osteoarthritis to construct a molecular quantitative trait locus map of gene expression and protein abundance in disease. By integrating data across omics levels, we identify likely effector genes for osteoarthritis-associated genetic signals. We detect stark molecular differences between macroscopically intact (low-grade) and highly degenerated (high-grade) cartilage, reflecting activation of the extracellular matrix-receptor interaction pathway. Using unsupervised consensus clustering on transcriptome-wide sequencing, we identify molecularly-defined patient subgroups that correlate with clinical characteristics. Between-cluster differences are driven by inflammation, presenting the opportunity to stratify patients on the basis of their molecular profile for tailored intervention. We construct and validate a 7-gene classifier that reproducibly distinguishes between these disease subtypes. Finally, we identify potentially actionable compounds for disease modification and drug repositioning. Our findings contribute to both patient stratification and therapy development in this globally important area of unmet need.

Keywords
Osteoarthritis, translational genomics, drug targets, drug repurposing, electronic health record, RNA sequencing, proteomics, patient stratification, functional genomics

Osteoarthritis, the most common form of arthritis, is a severe, debilitating disease hallmarked by cartilage degeneration and synovial hypertrophy, and affects ~240 Million people worldwide[1]. Osteoarthritis is a heterogeneous disease[2]. The lifetime risk of developing symptomatic knee and hip osteoarthritis is estimated to be 45% and 25%, respectively[3,4], and is on an upward trajectory commensurate with rises in obesity and the ageing population. Older age, female sex, obesity, joint morphology and injury are well-established risk factors for osteoarthritis, and genome-wide association studies (GWAS) have identified ~90 robustly-replicating risk loci.

There is no cure for osteoarthritis. Disease management focusses on alleviating pain, and in end-stage disease the only treatment is joint replacement surgery. Two million arthroplasties are carried out in the European Union annually[5], emphasising the clear and urgent need to develop new therapies that alter the natural history of the disease rather than deal with its consequences. To achieve this, we need to improve our understanding of the underlying molecular mechanisms of osteoarthritis pathogenesis and progression from low- to high-grade disease, which remain poorly characterised. Successful future treatment of early disease needs to reflect the heterogeneity of osteoarthritis and will require the identification of biological endotypes that match to the relevant therapeutic modality.

Molecular hallmarks of primary tissue

To improve our understanding of the molecular profile of key osteoarthritis cell types, we collected low-grade (macroscopically intact) and high-grade (highly degraded) cartilage and synovial tissue samples from 115 patients undergoing joint replacement for osteoarthritis. All cartilage samples were collected from weight-bearing areas of the joint. All three tissues were profiled by RNA sequencing, and cartilage samples were also profiled by isobaric labelling proteomics (**Supplementary Figure 1**). After quality control, we detected RNA-level expression of 15,249 genes in cartilage and 16,004 genes in synovium. We detected and quantified the abundance of 1,677 proteins across all patients, and of 4,801 proteins in at least 30 patients. We generated genome-wide genotype data from peripheral blood, imputing to 10,249,108 autosomal sequence variants to discover molecular quantitative trait loci (QTLs) in each tissue and omics type.

To identify molecular signatures associated with disease severity, we tested paired samples of high- versus low-grade cartilage for differential gene expression and protein abundance across 83 and 99 patients, respectively. We detected gene expression differences for 2,557 genes, and protein abundance differences for 2,233 proteins at 5% false discovery rate (FDR) (**Figure 1a**, **Methods, Supplementary Figure 2**). We identified significant cross-omics differential expression (i.e. at both the RNA and the protein level) for 409 of these genes (**Supplementary Table 1**). We found strong evidence for concordant direction of expression changes across the two omics levels (**Figure 1b**), providing internal cross-validation for the approaches used.

COL1A2, which showed significantly higher expression in high-grade cartilage at both omics levels, encodes the pro-alpha2 chain of type I collagen that is a prominent feature of disordered fibrocartilage repair in late osteoarthritis but is absent from intact articular cartilage[6]. *COL9A1* demonstrated lower cross-omics expression in high-grade cartilage, and encodes one of the type IX collagen alpha chains of hyaline cartilage, which is severely

degenerated in osteoarthritis. *MMP19*, a matrix metalloproteinase involved in the breakdown of the extracellular matrix (ECM)[7], also demonstrated higher cross-omics expression in high-grade cartilage, in agreement with matrix disintegration processes playing a central role in cartilage degradation.

We generated genetically-modified mice with mutant alleles in orthologues of 7 further genes with significant expression differences between high- and low-grade cartilage. Adult mice from these 7 lines underwent detailed phenotyping (**Supplementary Methods**). We identified at least one abnormal joint phenotype at nominal significance for each of the 7 genes studied (**Supplementary Figure 3-4, Supplementary Note**), functionally validating their role in the pathogenesis of musculoskeletal disease.

To identify key biological processes driven by molecular differences between high- and low-grade cartilage, we carried out gene set enrichment analyses based on differentially expressed (DE) genes (see **Methods**). ECM-receptor interaction emerged as the primarily activated pathway across omics levels in high-grade compared to low-grade cartilage (**Figure 1c**, **Supplementary Table 2**, **Supplementary Note, Supplementary Figure 2**).

Molecular QTLs in osteoarthritis tissues

Identification of expression QTLs can help elucidate effector genes for genetic association signals, and provide a better understanding of the transcriptional regulation of key cell types in health and disease. We identified *cis* expression QTLs (*cis*-eQTLs) for 1,891 genes in at least one tissue, with high correlation across the tissues studied (**Supplementary Figure 5a-c**). For example, the direction of effect was concordant across all 92,758 *cis*-eQTLs detected in both low- and high-grade cartilage (Pearson r=0.98, $p<2.2x10^{-16}$). We identified *cis* protein QTLs (*cis*-pQTLs) for 38 genes in at least one tissue, with similarly strong correlation across low- and high-grade cartilage (Pearson r=0.99, $p<2.2x10^{-16}$, **Supplementary Figure 5d-e**). This provides a first in-depth map of genetically-determined gene and protein level regulation in osteoarthritis-relevant tissues.

To further identify differential regulation of gene expression between high- versus low-grade cartilage, we examined variants with strong evidence for an eQTL effect in one tissue (posterior probability >0.9), but not in the other (posterior probability <0.1). We found 172 variants with differential effects on gene expression for 32 genes (differential eQTLs; **Supplementary Table 3**, **Figure 2a**, **Figure 2b**). Sixteen genes had differential eQTLs located in a *cis*-acting regulatory region, and key genes in which this effect was observed were involved in development (transcription factor *HOXB2*), inflammation (*IL4I1*), and fibrosis (*CRLF1*). These genotype-dependent, divergent patterns of gene regulation between high- and low-grade cartilage underline the importance of cell type and disease stage when investigating regulatory variant function.

Resolving GWAS signals

The majority of osteoarthritis genetic risk variants reside in non-coding sequence, making it challenging to identify the gene through which they confer their effect. Colocalisation analysis using molecular QTLs (molQTLs) can help clarify the mechanisms driving a GWAS locus by indicating whether the same variant is causal for both association with disease and for association with gene expression levels. We found strong evidence for colocalisation of 5

osteoarthritis loci with cartilage molQTLs for *ALDH1A2*, *NPC1*, *SMAD3*, *FAM53A*, and *SLC44A2* (**Figure 2c-d**). In all five instances, the GWAS index variant is non-coding. In three cases (*ALDH1A2, SMAD3* and *SLC44A2*) the likely effector gene is that residing closest to the lead variant. For the *NCP1* and *FAM53A* loci, the lead variants reside in introns of the *TMEM241* and *SLBP* genes, 141 kb and 18 kb away from the likely effector gene, respectively. This work helps pinpoint the identity of causal genes for hitherto unsolved association signals. In addition, our findings highlight the importance of generating molQTL data in disease-relevant tissue, as, in this case, using eQTLs from cartilage outperforms using the rich GTEx resource[8] (which does not include cartilage; see **Supplementary Note**).

Ninety-one of the genes with significantly different expression profiles between high-and low-grade cartilage were found to also be associated with genetic risk of osteoarthritis in a recent GWAS meta-analysis[9] (see **Methods**, **Supplementary Table 4**). For *ALDH1A2*, in which the GWAS signal and cartilage eQTLs colocalise, the risk-associated variants increase gene expression, in agreement with the higher gene expression levels we observe in high-grade cartilage. For *SLC39A8*, the GWAS signal is fine-mapped to a single missense variant with posterior probability of 0.999 and the gene demonstrates higher expression levels in high-grade cartilage. These findings highlight the value of integrating multi-omics data with genetic association summary statistics to identify likely effector genes for GWAS signals.

Patient stratification

Better stratification of patients by molecular endotype can provide opportunities for tailored therapeutic intervention. Primary tissue samples offer the opportunity to stratify patients on the basis of their molecular profiles. Here, we applied a clustering analysis to identify discrete subgroups across our patient tissue samples. Based on RNA sequencing data, we identified 2 clusters in synovium (42 and 34 samples, respectively), each of which further formed 2 sub-clusters (**Figure 3a**, **Supplementary Figure 6a,c**). We identified 2 clusters within low-grade cartilage (45 and 42 samples, respectively; **Figure 3b**, **Supplementary Figure 6a,c**), and no clear sub-clustering within high-grade cartilage (**Supplementary Figure 6a,b**). The identified cartilage clustering was independent of the synovium clusters (Fishers p-value >0.66).

Gene expression analysis showed large differences between the synovium clusters and sub-clusters, with over 5,000 genes differentially expressed at 5% FDR (**Supplementary Figure 7**). The differences between the two clusters relate to inflammation, while differences between the sub-clusters relate to the extracellular matrix and to cell adhesion (**Figure 3c-d**, **Supplementary Table 5**). Gene expression analysis also identified strong differences between the two low-grade cartilage clusters, with over 7,500 genes differentially expressed at 5% FDR. This clustering is also robustly associated with inflammation, extracellular matrix-related and cell adhesion pathways (**Figure 3e**, **Supplementary Table 5**).

When comparing our results to two smaller studies[10,11] that analysed data from gene expression arrays or RNA sequencing in low-grade cartilage, we find that there is consistent evidence for the clustering of patients based on inflammation-related molecular profiles (**Supplementary Note**, **Supplementary Table 6**). The presence of an inflammatory endotype axis within osteoarthritis provides an opportunity for patient selection for clinical trials of inflammation-modulating investigational therapies in appropriately selected patients.

Disease endpoints and discrete endotypes represent underlying processes that may be more sensitively captured by continuous axes of variation within the molecular data rather than binary or categorical classifiers. Such an approach may help define disease trajectories earlier on in the natural history of osteoarthritis. To evaluate this, we applied multi-omics factor analysis (MOFA)[12], an integrative method that can discover drivers of variability between samples or patients (latent factors) that is akin to a cross-data principal component analysis. The first two factors (axes of variation) were strongly associated with immune system processes and the extracellular matrix (see **Methods**, **Supplementary Note**), in keeping with the biological pathways identified to play an important role above. We also found the continuous axes of variation within low-grade cartilage and synovium to correspond strongly with cluster assignment (**Figure 3f**, **Supplementary Figure 7c**,**8**, **Supplementary Note**). This is consistent with variation within tissues being better captured as a continuous spectrum rather than as discrete clusters.

7-gene classifier predicts clustering

Based on the above findings, we sought to develop a molecular tool based on the expression of a small number of genes that can predict cartilage cluster assignment for osteoarthritis patients. We used a soft-thresholding centroid-based method, PAMR[13], to develop a gene expression-based classifier capable of distinguishing between the two cartilage clusters. We identified 7 genes, the expression levels of which could be combined to predict cluster assignment for each patient sample (**Figure 4a**, **Supplementary Figure 9**): *MMP1*, *MMP2*, and *MMP13*, known to be involved in cartilage degradation[14]; *IL6*, a pro-inflammatory cytokine; *CYTL1*, a cytokine-like gene, loss of which has been found to augment cartilage destruction in surgical osteoarthritis mouse models[15]; *APOD*, a component of high-density lipoprotein found to be strongly up-regulated by retinoic acid[16], which is in turn regulated by *ALDH1A2*[17], an osteoarthritis risk locus[9,18]; and *C15orf48*, with currently unknown function. Notably, the posterior probabilities for cluster assignment output by the classifier captured the main continuous spectrum of variation in this tissue (**Figure 4b**).

To validate the 7-gene classifier, we obtained an independent gene expression dataset of low-grade cartilage samples from 60 knee osteoarthritis patients undergoing joint replacement surgery[11]. The samples had been assigned into two groups, reflecting differences in complement activation and innate immunity. This group assignment corresponded to the 7-gene classifier cluster assignment for 73% of the samples (32 out of 44 samples with available data). In addition, the posterior probabilities for cluster assignment had good correspondence to the main continuous spectrum of variation (**Figure 4c**). These findings indicate strong agreement and support the predictive potential of the 7-gene classifier in this independent dataset.

Clinical profiles of molecular clusters

We investigated whether the stratification of patients into different tissue-based transcriptional profile clusters was associated with clinical characteristics. We compiled information on sex, age, height, weight, body mass index and pre-operative American Society of Anesthesiologists (ASA) grade[19], and electronic health records information on prescribed medications at the time of joint replacement surgery. Cartilage cluster

assignment was associated with patient sex (OR=4.12, p=0.0024), with women more likely to be members of the cluster characterised by higher inflammation. One explanation for this observation may be the lower concentration of oestrogen and androgens, which have established anti-inflammatory effects, in post-menopausal women[20-22]. This is in line with the disproportionate increase in the incidence of osteoarthritis in women after the menopause. Patients in the high-inflammation cluster were also more likely to be prescribed proton pump inhibitors (OR=4.21, p=0.0040; **Supplementary Table 7**). Several further clinical characteristics were associated with cluster assignment at nominal significance: patients in the high-inflammation cluster were more likely to be prescribed a higher number of drugs (OR=1.21 per additional drug, p=0.023) and to be older (OR=1.06 per year, p=0.0036).

Although the mechanisms of these associations remain unclear, there is an established association between osteoarthritis, multimorbidity and polypharmacy[23,24]. The association of molecularly-defined patient clusters with clinical characteristics lends initial evidence to support the integration of omics biomarkers to drive precision medicine approaches in osteoarthritis, and indicates that post-menopausal women and patients with polypharmacy may provide an opportunity for targeted disease-modifying interventions.

Candidate therapeutic compounds

We aimed to identify compounds with the potential to reverse the spectrum of molecular differences between high- and low-grade patient cartilage based on existing *in vitro* drug screen data. We used ConnectivityMap[25], a dataset of 2,684 gene expression perturbations induced by compounds across 9 human cell lines, to assess each perturbation profile against our differentially expressed genes. We identified 19 compounds that induced strong opposing gene expression signatures to the differences between high- and low-grade cartilage, reducing the expression of genes with cross-omics higher expression in high-grade cartilage (**Table 1**, **Supplementary Table 8**). These include oestrogen receptor agonists diethylstilbestrol and alpha-estradiol, the latter of which targets *KCNMA1*, coding for the pore-forming alpha subunit of a calcium-sensitive potassium channel, and demonstrating significantly lower gene expression and protein abundance in high-grade cartilage. These findings are consistent with our clinical classifier, molecular clustering, and with established epidemiological data showing an association between osteoarthritis and oestrogen deficiency[26]. Although studies of oestrogen therapy for osteoarthritis have been largely inconclusive[27,28], identification of cartilage-specific oestrogen-mediated pathways, such as through *KCNMA1*, may allow more focussed investigational molecule development.

Several further drugs with molecular signatures potentially capable of reversing those observed in our differential expression analysis have known links to osteoarthritis (**Table 1**): IB-MECA (an adenosine receptor agonist used as an anti-inflammatory drug in rheumatoid arthritis)[29], VEGF-receptor-2-kinase-inhibitor-IV, RHO-kinase-inhibitor-III[rockout] (a rho associated kinase inhibitor), and nornicotine (an acetylcholine receptor agonist extracted from tobacco and related to nicotine)[30]. In a rat model of chemically-induced osteoarthritis, IB-MECA prevented cartilage damage, osteoclast/osteophyte formation, and bone destruction[31]. *VEGF* modulates chondrocyte survival during development and is essential for bone formation and skeletal growth. However, dysregulation of *VEGF* expression in the adult joint is a feature of osteoarthritis[32]. Conditional knock-down of *Vegf* attenuates

surgically-induced osteoarthritis in mice, with intra-articular anti-VEGF antibodies as well as oral administration of the VEGFR2 kinase inhibitor Vandetanib suppressing osteoarthritis progression[33]. In a rat model of osteoarthritis, a rho kinase inhibitor was found to reduce knee cartilage damage[34]. Finally, there is a well-established effect of smoking on osteoarthritis[9,35]. Together, these results identify candidate compounds that warrant investigation, and provide evidence for the validity of this approach.

In addition to signatures induced by compounds, ConnectivityMap contains gene expression profiles induced by *in vitro* gene knock-down or over-expression. We identified 36 genes for which the experimental perturbation induces changes in the opposite direction to molecular differences between high-grade and low-grade cartilage (**Supplementary Table 8**), notably including knock-down of *IL11*. Variation in *IL11* is associated with increased risk of hip osteoarthritis[9], and the gene is up-regulated in osteoarthritis knee tissue[36], with a similar trend observed here. IL11 is a cytokine with a key role in inflammation, and monoclonal anti-IL11 antibodies have been developed for use in several diseases. These findings provide strong supportive evidence for down-regulation of *IL11* as a potential therapeutic intervention for osteoarthritis.

Discussion

Osteoarthritis is a globally important condition of huge public health relevance. As a heterogeneous disease, it requires patient stratification for successful therapy development and translation. Here, we have leveraged the accessibility of primary disease tissue to create a comprehensive molecular portrait of key cell types relevant to osteoarthritis. We have combined genome-wide genotyping with RNA sequencing and quantitative proteomics to construct the first deep molecular quantitative trait locus map of cell types directly involved in disease. We have identified molecular drivers of disease grade and have stratified patients on the basis of their omics profile. We have built and independently replicated a 7-gene classifier that captures patient heterogeneity and distinguishes between two patient clusters with distinct clinical characteristics. By integrating multiple layers of omics data, we have helped resolve genetic association signals by identifying likely effector genes, and have highlighted opportunities for new targets and for repositioning. Our findings identify drug repurposing opportunities and allow the identification of novel investigational avenues for patient stratification, disease severity, and therapy development, responding to the global challenge of osteoarthritis and the clear unmet clinical need.

Acknowledgements

We thank the study participants who made this work possible by their generous donation of samples. This work was funded by the Wellcome Trust (206194). M.J.C. was funded through a Centre for Integrated Research into Musculoskeletal Ageing grant (MRC 148985). R.A.B. and the Human Research Tissue Bank are supported by the NIHR Cambridge Biomedical Research Centre. J.H.D.B. and G.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). Mutant mice were generated via Wellcome Trust grant WT098051. We thank members of the Sanger Institute Mouse Pipelines teams (Mouse Informatics, Molecular Technologies, Genome Engineering Technologies, Mouse Production Team, Mouse Phenotyping) and the Research Support

Facility for the provision and management of the mice. The authors are grateful to Dr. Iris Fischer for helpful edits.

Author Contributions

Study design: E.Z., J.M.W, J.S. Collection of knee samples: M.J.C., R.L.J., D.S., K.S., J.M.W. Collection of hip samples: R.A.B., A.W.M. Review of patient electronic health record data: A.F., J.M.W. Proteomics assays: T.I.R., J.S.C. Mouse Resources: C.J.L. Mouse experiments: N.B., K.F.C., S.M.P., J.H.D.B., G.R.W. Molecular QTL and colocalisation analyses: L.S. Differential expression analyses: J.S., L.S. Pathway association, tissue clustering, MOFA, drug repurposing, and statistical mouse data analyses: J.S. Writing - original draft: J.S, L.S., N.B, J.M.W., E.Z. Writing - comments and review: all authors.

Declaration of Interests

The authors declare no competing interests.

Online Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Study participants

We collected tissue samples from 115 patients undergoing total joint replacement surgery (102 knee and 13 hip osteoarthritis patients from 4 cohorts, **Supplementary Methods**). 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. Full details for confirmation of joint replacement for osteoarthritis and cartilage tissue scoring are listed in **Supplementary Methods**. We followed previously established protocols to isolate chondrocytes[37,38], and synoviocytes[39], summarised in **Supplementary Methods**.

Cohorts 1, 2, 4 (knee osteoarthritis)

This work was approved by Oxford NHS REC C (10/H0606/20 and 15/SC/0132), and samples were collected under Human Tissue Authority license 12182, Sheffield Musculoskeletal Biobank, University of Sheffield, UK.

We obtained information on patient clinical characteristics (age, height, weight, body mass index (BMI), American Society of Anaesthesiologists (ASA) grade[19]) from the electronic patient records. For each patient, a list of drugs prescribed on the date of sample collection was also compiled from the electronic patient record and cross referenced with the patient medical history.

Cohort 3 (hip osteoarthritis)

Samples were collected under National Research Ethics approval reference 11/EE/0011, Cambridge Biomedical Research Centre Human Research Tissue Bank, Cambridge University Hospitals, UK.

DNA, RNA and protein extraction

DNA, RNA, and protein extraction was carried out using Qiagen AllPrep DNA/RNA/Protein Mini Kit following manufacturer's instructions, with small variations for cohort 3 as previously described[38]. 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 Table 9**). We purified poly-A tailed RNA (mRNA) from total RNA using Illumina's TruSeq RNA Sample Prep v2 kits. After fragmentation and standard Illumina library prep (see **Supplementary Methods**), multiplexed libraries were sequenced on the Illumina HiSeq 2000 for cohort 1 and HiSeq 4000 for cohorts 2-4 (75bp paired-ends). Sequenced data underwent initial analysis and quality control 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 Table 9**). 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 described[37]. 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 (e.g. 0.05% SDS in protein pellet re-suspension for cohort 1, with 0.1% SDS in cohorts 2-4, see **Supplementary Methods**).

Genotyping

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

Statistical Analyses

Quantification of RNA levels

We used samtools v1.3.1[40] and biobambam v0.0.191[41] to convert cram to fasq files after exclusion of reads that failed QC. We applied FastQC v0.11.5 to check sample quality[42] and excluded 9 samples (**Supplementary Table 9**). We quantified expression levels using salmon v0.8.2[43] and the GRCh38 cDNA assembly release 87, obtaining gene-level scaled transcripts per million (TPM) estimates from tximport 1.4.0[44] (details see **Supplementary Methods**). We excluded 43 samples due to low mapping rate (<80%), non-European ancestry, low RIN (<5), duplicates and abnormal gene read density plots (see **Supplementary Methods**, **Supplementary Table 9**). The final gene expression dataset included 259 samples (**Supplementary Figure 1**; 87 patients' low-grade and 95 high-grade cartilage samples with 15,249 genes that showed counts per million (CPM) of ≥1 in ≥40 samples, and 77 patients' synovium samples with 16,004 genes that showed CPM ≥1 in ≥20 samples).

Quantification of protein levels

We used SequestHT in Proteome Discoverer 2.1 for protein identification and quantification (details see **Supplementary Methods**), searching all spectra against a UniProt fasta file with 20,165 reviewed human entries. We only used peptides uniquely belonging to protein groups for quantification. We excluded samples from 4 patients due to non-European ancestry (**Supplementary Table 9**). The final dataset included low-grade and high-grade cartilage samples each from 99 patients, with 4,801 proteins was observed in ≥30% of samples, and 1,677 proteins in all samples. 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 quality control

Genotypes were called using GenCall (Illumina) and mapped to GRC37/hg19, with quality control (QC) including checks for identity, sex, call rate, heterozygosity rate, Hardy-Weinberg equilibrium, relatedness, and European ancestry (**Supplementary Methods**). The resulting dataset containing 111 patients and 504,235 overlapping variants across both arrays. We imputed up to HRC panel v1.1[45] using the Michigan imputation server[46] with Eagle2 phasing. After post-imputation variant QC (**Supplementary Methods**), we excluded

two patients due to absence of RNA and protein data. The resulting final dataset contained 109 patients and 10,249,108 autosomal variants.

Differential RNA expression between low-grade and high-grade cartilage

We tested differential expression of 15,249 genes between low-grade and high-grade cartilage using paired samples from 83 patients. To identify robust results, we carried out multiple analyses using different software packages, as recommended in a landmark survey of best practices[47], applying limma[48], edgeR[49], and DESeq2[50]. We tested 5 analysis designs with different options to account for technical variation, including SVAseq[51], see **Supplementary Methods**. In each analysis design and method, we used a 5% False Discovery Rate (FDR) threshold to correct for multiple testing. This yielded 2,557 genes with significant differential expression between low-grade and high-grade cartilage across all analysis designs and testing methods (2,418 with uniquely corresponding Ensembl gene ID and gene name, see **Supplementary Methods**).

Differential protein abundance between low-grade and high-grade cartilage

We performed differential analysis for 4,801 proteins that were measured in ≥30% of patients, applying limma[48] to paired samples from 99 patients. Significance was defined at 5% FDR to correct for multiple testing, yielding 2,233 proteins with significant differential abundance (2,019 proteins with uniquely corresponding Ensembl gene ID and gene name). Paired samples from any patient were always assayed in the same multi-plex and we used a sensitivity analysis to confirm adjustment for patient effects captured between-plex batch effects (**Supplementary Methods**).

Pathway associations for differences between low-grade and high-grade cartilage

To identify the biological processes with significant molecular differences between low-grade and high-grade cartilage, we carried out gene set enrichment analyses based on the differential expression (DE) on RNA, protein, and cross-omics levels, using several FDR DE thresholds for robustness checks (5%, 1%, 0.5%, 0.1%; **Supplementary Methods**).
We applied Signalling Pathway Impact Analysis (SPIA)[52] to test for association with KEGG signaling pathways. SPIA combines enrichment p-values with perturbation impact on the pathway based on log-fold-changes of the DE genes, with low p-values when both over-representation and pathway impact p-values are low. Significance of pathway association was defined as a threshold of 5% FDR applied to the combined p-values in each analysis. We also tested enrichment in Gene Ontology terms using GOseq[53], separately for genes with higher or lower expression in high-grade compared to low-grade cartilage (**Supplementary Methods**). Significance was defined as a threshold of 5% FDR in each analysis. The results showed broad agreement with the results of the SPIA analysis (**Supplementary Table 2**).

Identification of *cis*-eQTLs and *cis*-pQTLs

We followed a similar method to GTEx[8,54], see **Supplementary Methods** for details. Briefly, for eQTLs, we processed low-grade and high-grade cartilage together for exclusion of genes with low expression and between-sample normalisation, then normalised each gene across samples within the same tissue. We applied PEER[55] to infer hidden factor due to any technical differences, using 15 PEER factors in each tissue, sex, and genotype array as covariates in eQTL analyses. We applied the GTEx modified version of FastQTL[56] to detect

eQTLs. We determined the transcription start site (TSS) for each gene using empirical transcript level expression information and defined the *cis*-mapping region to be 1Mb in either direction from the TSS. Genes with significant eQTLs ("eGenes") and proteins with significant pQTLs (genetic variants associated with protein abundance levels) ("pGenes") were defined at the 5% False Discovery Rate (FDR) threshold for q-values obtained from permutations. 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 GRC37/hg19.

We followed a similar protocol for pQTLs, using 26 PEER factors for each tissue. For low-grade and high-grade cartilage, we included 1677 proteins that were measured across all samples. We used the TSS established for the eQTL analysis, taking forward 1461 proteins with a unique mapping.

For both eQTLs and and pQTLs, we verified that the results were robust by carrying out a sensitivity analysis including age and joint as covariates (**Supplementary Methods**).

To identify *cis*-eQTLs active exclusively in low-grade or high-grade cartilage, we used Meta-Tissue[57] and METASOFT[58], as described in the **Supplementary Methods**. The m-value calculated by METASOFT for gene-variant pair in each tissue provides a posterior probability of an effect in that tissue. Consequently, we aimed to identify eQTLs present in one tissue (defined as m>0.9), and absent in the other (defined as m<0.1). We note that there were no *cis*-eQTLs present in both tissues (m>0.9) with opposing direction of effect. To identify variants located in regulatory regions, we used Ensembl Variant Effect Predictor (http://grch37.ensembl.org/Homo_sapiens/Tools/VEP/).

Colocalisation between molQTLs and osteoarthritis GWAS associations

To examine colocalisation between molQTLs and GWAS associations, we used genome-wide summary statistics from the largest osteoarthritis meta-analysis to date, based on UK Biobank and arcOGEN data[9]. We analysed all 64 genome-wide significant signals using coloc[59], separately for each tissue and omics level (**Supplementary Methods**). We considered a 80% posterior probability of GWAS and molQTL shared association at a single variant ("PP4≥0.8") to indicate evidence of colocalisation.

Identification of genes with osteoarthritis GWAS gene-level association

From the recent UK Biobank and arcOGEN GWAS meta-analysis[9], we obtained the results of a gene-level analysis for each of the four osteoarthritis phenotypes (self-reported plus hospital diagnosed, hospital diagnosed knee or hip, hospital diagnosed knee, hospital diagnosed hip), as described in the GWAS paper. Briefly, this analysis used MAGMA v1.06[60] and was based on the mean SNP log-p-value in the gene, accounting for LD. After accounting for the effective number of tests across phenotypes and genes using a Bonferroni correction (**Supplementary Methods**), 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.

Sample clustering

For RNA data, we normalised each tissue separately, using limma-voom[61] to remove heteroscedasticity from scaled TPM values, followed by pSVA[62] and regressing out RNA

sequencing batches and clinical batches (**Supplementary Methods**). For the proteomics data, we regressed out batches from the log2-transformed normalised abundance values, then quantile normalised the residuals as analogous step to the differential expression analysis; all analyses were also done without quantile normalisation, with no appreciable difference in results.

For each tissue and omics level, we applied ConsensusClusterPlus[63], a consensus clustering method that splits samples into a discrete number of groups, so that samples within a group are more similar to each other than to samples outside the group (with standard settings, see **Supplementary Methods**). The final number of clusters was chosen based on the Consensus Cumulative Distribution Function plots, the Delta Area Plot, and a visual investigation of the Consensus Matrices, as advised in the manual. Results were confirmed via additional analysis using a distance metric based on Pearson correlation.

Differential gene expression between tissue clusters

To follow up the clustering results for low-grade cartilage and synovium, we tested gene differential expression between sets of samples based on cluster assignment (applying limma to the normalised expression values underlying the clustering, i.e. gene expression after voom, pSVA, and regression of batch covariates). The differential expression analysis was followed up by gene set enrichment analyses using SPIA and GOseq, with 8 gene differential expression FDR thresholds to assess robustness of the association (5%, 0.5%, $5x10^{-3}$, …, $5x10^{-7}$). In each analysis, gene set association was defined at the 5% FDR threshold. As before for GOseq, genes with positive and negative log-fold-change between clusters were analysed separately.

Multi-omics factor analysis (MOFA) and correspondence to sample clustering

To test for patient heterogeneity using a method that can detect both discrete clustering and a continuous spectrum of variation, we used multi-omics factor analysis (MOFA)[12]. MOFA can integrate data across omics levels and across tissues to discover drivers of variability between samples or patients. MOFA was run i) jointly on all RNA and protein data; ii) jointly on RNA data across all three tissues; iii) on RNA and protein data within each tissue (see **Supplementary Methods** for settings). MOFA identifies a factor score for each sample or patient, calculates the variance explained by each factor in each omics level and tissue, calculates weights of genes on each factor from each omics level and tissue, carries out a gene set enrichment for each factor in each omics level and tissue based on gene weights.

We further investigated the correspondence between the continuous spectrum of variation identified by the MOFA and the discrete clusters identified by ConsensusClusterPlus by calculating the correlation between RNA gene weights on MOFA factors and gene expression differences between or within clusters (**Supplementary Methods**, **Supplementary Note**).

Construction of classifier to reflect sample clustering

We applied PAMR[13], a soft-thresholding centroid-based method, to identify a smaller subset of genes which could distinguish the low-grade cartilage clusters. To train a classifier, we

restricted the analysis to 1,063 genes with high expression levels and 83 samples with cluster silhouette score >0.2 (**Supplementary Methods**).

Based on the resulting 7-gene classifier, we used the pamr.predict function to predict cartilage-Cluster1 and cartilage-Cluster2 probabilities for all 87 low-grade samples, with an agnostic prior setting of 0.5 for both clusters. We also calculated Spearman correlations for the PAMR cluster probabilities and MOFA Factor1 for low-grade cartilage.

Replication of clustering classifier

We obtained RNA expression data from low-grade cartilage tissue of 60 knee osteoarthritis patients undergoing joint replacement (27 women, 33 men, age range 63-85 years), sequenced on Illumina HiSeq 2500, with transcript quantification using kallisto and quality control as described previously[11]. We obtained gene-level data and removed batch effects as for the discovery data (**Supplementary Methods**). We applied the pamr.predict function to predict cartilage-Cluster1 and cartilage-Cluster2 probabilities for all 60 samples using the trained 7-gene classifier, with an agnostic prior setting of 0.5 for both clusters. We also applied MOFA (with the same parameters and options as above) to the data post batch effect removal. Finally, we calculated Spearman correlations for the PAMR cluster probabilities and MOFA factor 1. The original publication also included a division of samples into 2 groups using non-negative matrix factorisation based on known biological networks. This assignment was compared to a cluster assignment based on PAMR 7-gene classifier posterior probabilities.

Associations between tissue cluster assignment and clinical data

We tested for association between low-grade cartilage dichotomous cluster assignment (high-inflammation cartilage-Cluster1 versus low-inflammation cartilage-Cluster2) and clinical characteristics using a generalised linear model (via the glm function in R with option family="binomial"). To consider the association of tissue clusters with drug prescription, drugs were grouped by pharmacological mechanism into 58 categories by two clinical experts (AF & JMW). We restricted the analysis to 9 drug categories, each with at least 20 patients who were also assigned a low-grade cartilage or synovium cluster (**Supplementary Table 7**). We calculated the effective number if tests across clinical characteristics and the 9 drug categories as N_{eff}<10.54 (**Supplementary Methods**), and thus used a Bonferroni-corrected threshold of p<0.05/10.54=0.0047 to define significance. We carried out a sensitivity analysis including sex or sex and age as covariates to verify the robustness of associations detected for low-grade cartilage clustering (**Supplementary Methods**, **Supplementary Note**).

For association with synovium cluster assignment, we carried out analogous tests for the two clusters, with the same Bonferroni-corrected significance threshold.

ConnectivityMap analysis

To identify opportunities for drug repurposing, we used ConnectivityMap[25] to identify compounds and perturbagen classes (PCL) that could possibly reverse the differences identified between high-grade and low-grade cartilage. Using the online interface clue.io (accessed 03/03/2019), 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 9 cell lines. A positive tau score indicates similarity between the gene expression signature of a perturbation and the submitted query (i.e. up-regulation 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. down-regulation 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/connectopedia/connectivity_scores, accessed 03/03/2019). A total of 2837 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 perturbagen classes, 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.

Data Availability

The RNA sequencing data reported in this paper have been deposited to the EGA (accession numbers EGAS00001002255 (https://wwwdev.ebi.ac.uk/ega/studies/EGAS00001002255), EGAD00001003355, EGAD00001003354, EGAD00001001331). The proteomics data reported in this paper have been deposited to PRIDE (accession numbers PXD014666, PXD006673, PXD002014). The genotype data reported in this paper have been deposited to the EGA (accession numbers EGAD00010001746, EGAD00010001285, EGAD00010001292, EGAD00010000722).

Code Availability

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

Figures

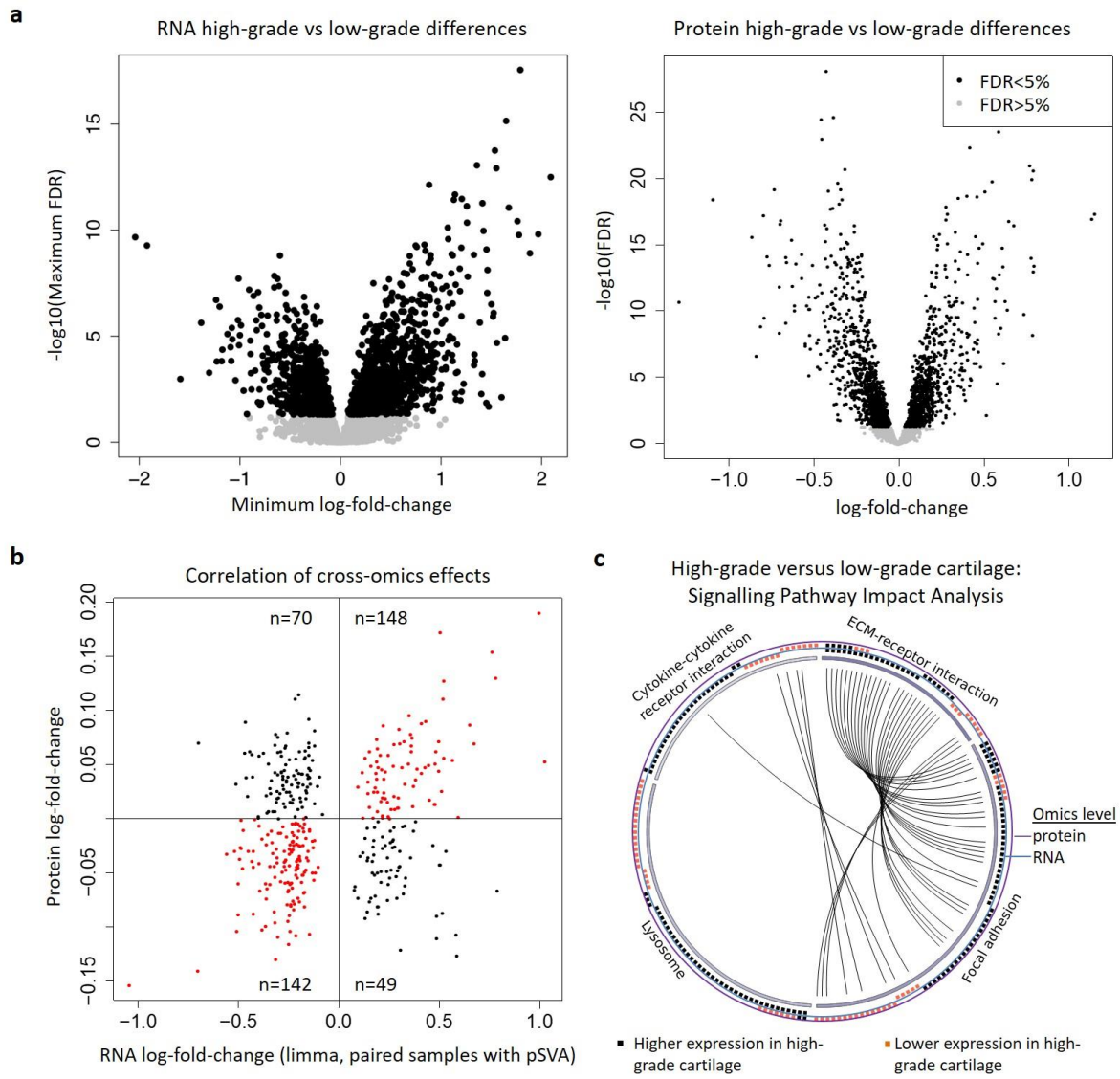


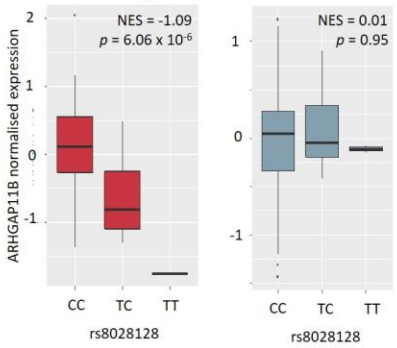
Figure 1. Molecular differences between high-grade and low-grade cartilage

a) Wide-spread RNA-level (left) and protein-level (right) differences between high-grade and low-grade cartilage. The RNA plot shows conservative results based on different approaches (see **Methods**).

b) RNA- and protein-level log-fold-changes for 409 genes with significant cross-omics differences between high-grade and low-grade cartilage (see **Supplementary Figure 2a** for all genes). The direction of difference agrees for 290 of the 409 genes (71%; binomial $p<1.0x10^{-17}$).

c) Signalling Pathway Impact Analysis (SPIA) identified biological pathways associated with differences between high-grade and low-grade cartilage. Pathways with significant results at 5% FDR based on RNA-level changes are shown, all activated in high-grade cartilage. Boxes on the outside circles represent individual genes, with arches connecting the same gene across pathways. See also **Supplementary Figure 2b** and full results in **Supplementary Table 2**.

a



b

Cartilage differential eQTLs

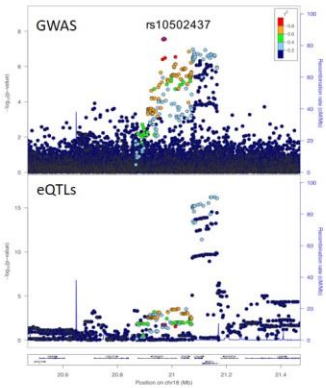
Gene	Cartilage tissue with QTL effect	Number of differential eQTLs	Number differential eQTLs in regulatory regions
HOXB2	High-grade	61	15
MT1DP	Low-grade	17	12
AC114737.3	High-grade	10	3
IFITM3	High-grade	9	3
PMS2P5	High-grade	9	2
IL4I1	High-grade	7	1
CRLF1	High-grade	6	1
TMEM237	High-grade	6	3
PCYT1A	High-grade	6	0

c

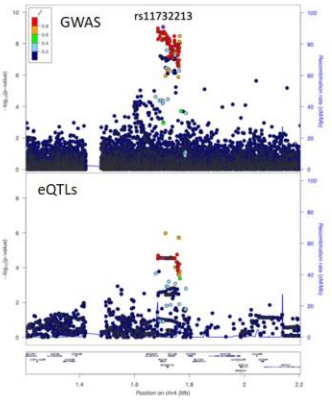
GWAS variant	rs10502437	rs11732213	rs12901372	rs1560707	rs4775006
OA Phenotype	All	Hip/Knee	Hip	All	Knee
AF	0.6	0.81	0.53	0.37	0.41
OR	1.03	1.06	1.08	1.04	1.06
p-value	2.50×10^{-8}	8.81×10^{-10}	3.46×10^{-11}	1.35×10^{-13}	8.40×10^{-10}
Risk allele	A	T	C	T	A
Gene	*NPC1*	*FAM53A*	*SMAD3*	*SLC44A2*	*ALDH1A2*
Risk allele effect	-	-	-	+	+

d

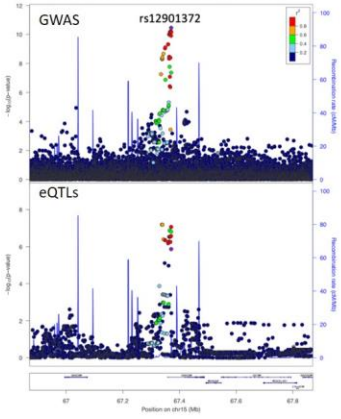
NPC1
high-grade cartilage eQTLs, PP4 = 0.94



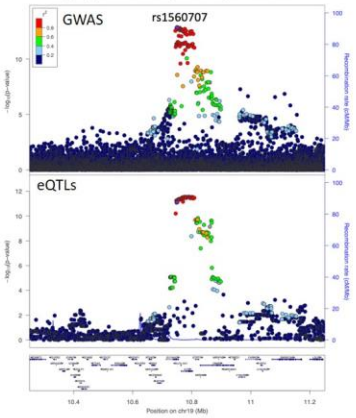
FAM53A
low-grade cartilage eQTLs, PP4=0.84



SMAD3
high-grade cartilage eQTLs, PP4=0.98



SLC44A2
high-grade cartilage eQTLs, PP4=0.94



ALDH1A2
low-grade cartilage pQTLs, PP4 = 0.92

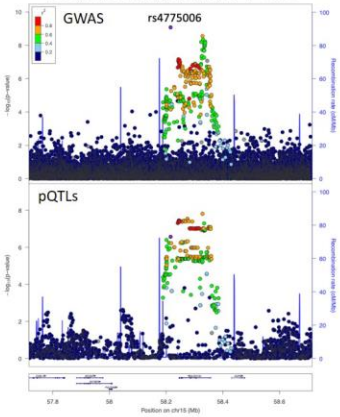


Figure 2. Molecular QTLs in osteoarthritis disease tissue

a) An example of differential QTL effect: an association present in high-grade, but not low-grade cartilage. The boxplots show normalised expression at 25th, 50th and 75th percentiles, and whiskers extend to 1.5 times the interquartile range.

b) Genes with at least 5 differential eQTL variants, i.e. posterior probability for presence of an eQTL effect is high in high-grade cartilage (m>0.9) and low in low-grade cartilage (m<0.1), or vice versa. Full results see **Supplementary Table 3**.

c) Osteoarthritis GWAS signals with high posterior probability for colocalisation with molecular QTLs. Risk allele effect: "+" for increase of expression with risk allele, "-" for decrease.

d) GWAS and molecular QTL p-values in regions with colocalisation of the associations. Plots show 1Mb regions centered around the GWAS index SNPs (purple), with one point per genetic variant. PP4: posterior probability for colocalisation. For rs10502437 and rs12901372, colocalisation between GWAS and low-grade cartilage molecular QTLs is shown in **Supplementary Figure 5f**.

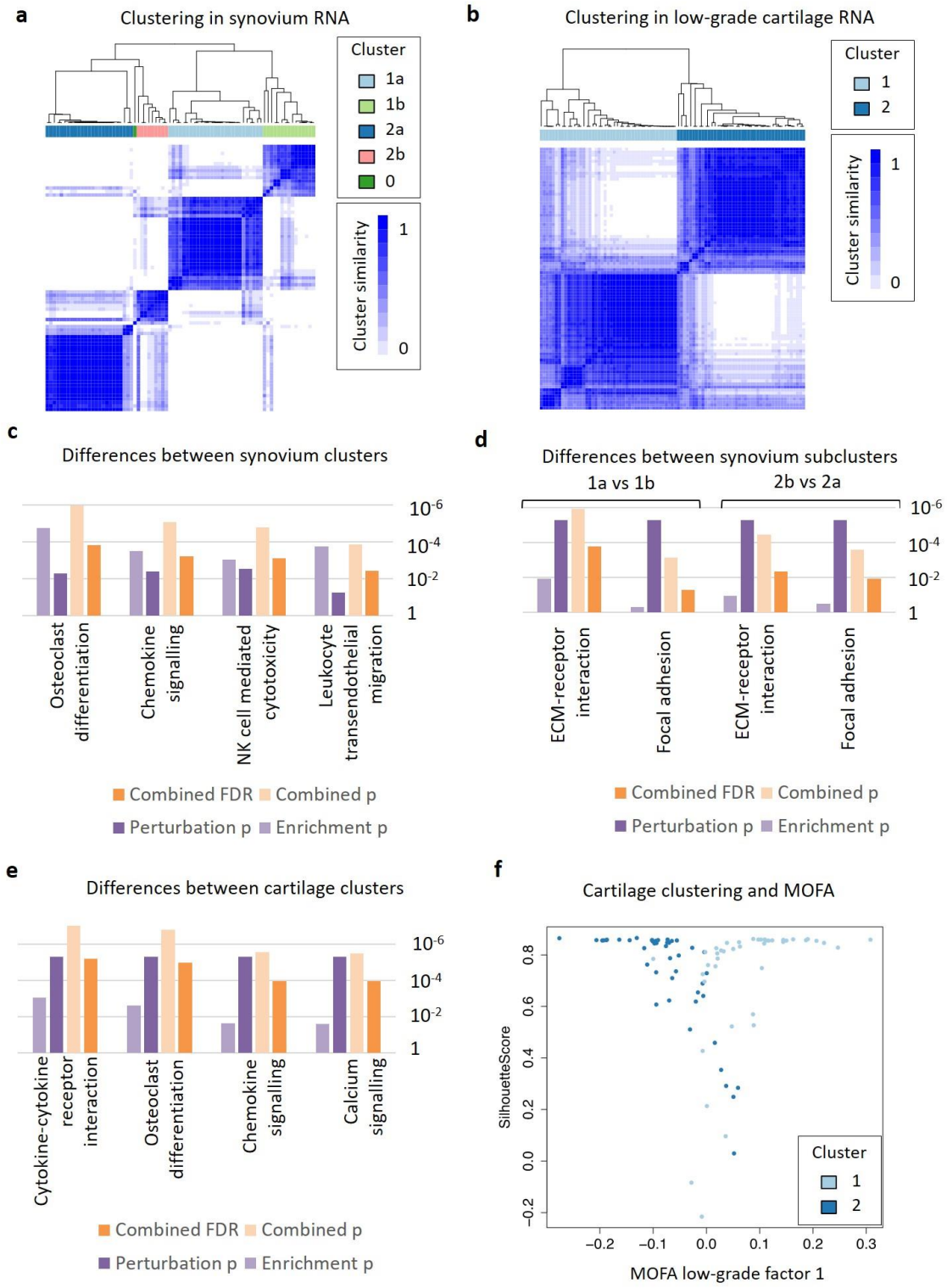


Figure 3. Distinct clusters identified in low-grade cartilage and synovium tissue

a) Synovium tissue samples from patients are separated into two clusters based on RNA data (synovium-Cluster1 and synovium-Cluster2). Each cluster formed 2 sub-clusters (synovium-Cluster1a and synovium-Cluster1b; separately synovium-Cluster2a and synovium-Cluster2b). Cluster 0: one outlier sample.

b) Low-grade cartilage tissue samples from patients are separated into two clusters based on RNA data (cartilage-Cluster1 and cartilage-Cluster2).

c) Gene expression differences between synovium clusters shows several significant associations related to inflammation and osteoclast differentiation.
d) Gene expression differences between the synovium sub-clusters within each cluster show similar pathway associations, including to ECM-receptor interaction and focal adhesion pathways.
e) Gene expression differences between low-grade cartilage clusters show several significant pathway associations, including inflammation and osteoclast differentiation.
f) An analysis of low-grade cartilage samples using MOFA identifies a continuous spectrum of variation between samples. Samples with high MOFA Factor 1 scores are mostly in cartilage-Cluster1 and those with low MOFA Factor 1 scores mostly in cartilage-Cluster2. Samples with intermediate MOFA Factor 1 scores have lower Silhouette Scores, showing more uncertainty in cluster assignment. For synovium, see **Supplementary Figure 7c**.

Enrichment p: SPIA p-value for over-representation analysis of genes; Perturbation p: SPIA p-value for perturbation of the pathway based on gene log-fold-changes; Combined p: SPIA p-value from combining enrichment and perturbation p-values.

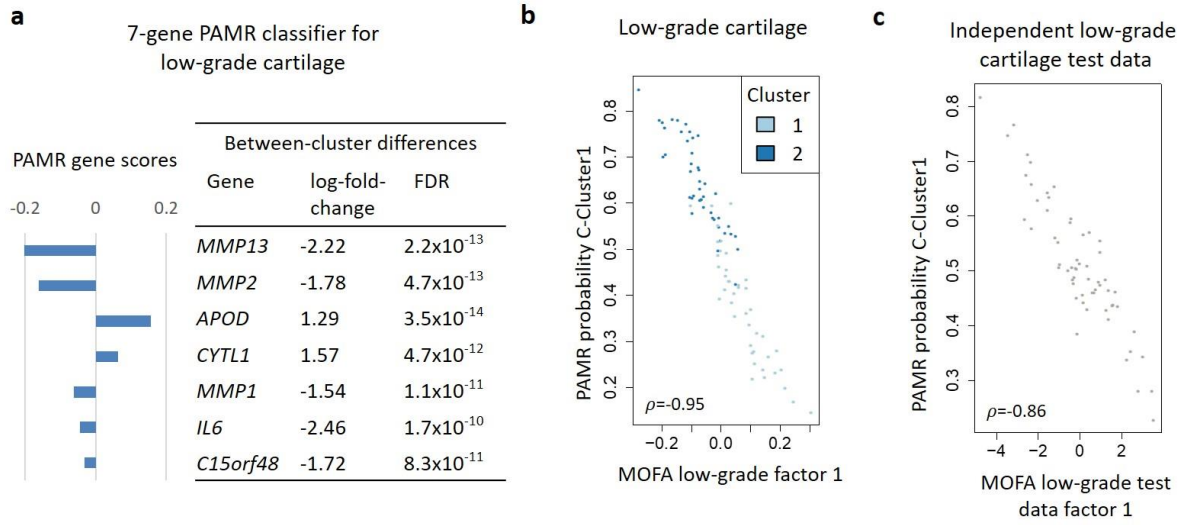


Figure 4. Variation within low-grade cartilage can be recovered using a 7-gene classifier

a) Using PAMR, we constructed a 7-gene classifier to predict cluster assignment for low-grade cartilage samples. The barplot shows the PAMR score for each gene, the right panel the differential expression of the genes between the two low-grade cartilage clusters. See also **Supplementary Figure 9** for classifier performance.

b) The PAMR posterior probabilities for cluster assignment are highly correlated with MOFA Factor 1 scores for low-grade cartilage samples, capturing the continuous spectrum of variation between samples. Inset: Spearman correlation.

c) In an independent set of 60 low-grade cartilage samples from 60 osteoarthritis patients undergoing total-knee-replacement, the posterior probabilities for cluster assignment from the 7-gene classifier are well-correlated with the continuous spectrum of variation in these samples, as quantified by MOFA Factor 1 in an *ab initio* analysis. Inset: Spearman correlation.

Tables

Name	Description	DE targets
Emetine	protein synthesis inhibitor	RPS2 (P+)
Rucaparib	PARP inhibitor	PARP2 (R-)
Alpha-estradiol	estrogen receptor agonist	KCNMA1 (R-, P-)
VEGF-receptor-2-kinase-inhibitor-IV	VEGFR inhibitor	
IB-MECA	adenosine receptor agonist, granulocyte colony stimulating factor agonist	
Diethylstilbestrol	estrogen receptor agonist, chloride channel blocker	
KIN001-220	Aurora kinase inhibitor	
SB-216763	glycogen synthase kinase inhibitor	GSK3B (R+, P+), CDK2 (P-)
RHO-kinase-inhibitor-III[rockout]	ROCK inhibitor	IMPDH2 (P-)
Nornicotine	acetylcholine receptor agonist	

Table 1. Compounds with strongest evidence for inducing gene expression signatures that counter differences between high-grade and low-grade cartilage.
Results based on data from ConnectivityMap[25]. DE targets: drug targets as listed in ConnectivityMap with RNA (R) or protein (P) differences between high-grade and low-grade cartilage, "+" and "-" indicate higher or lower expression high-grade cartilage, respectively. The 10 compounds with lowest median tau scores are shown; the full list of compounds is in **Supplementary Table 8**.

Supplementary information

Supplementary Figures

Supplementary Figure 1. Large-scale multi-omics characterisation of osteoarthritis disease tissue: study approach

a) 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 and information from electronic health records (EHRs). 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. We generated mouse lines of several markers of cartilage degeneration, extensively profiling the bone and cartilage phenotypes of the mutant mice. Finally, we identified patient heterogeneity based on the molecular data, constructed and replicated a 7-gene probabilistic classifier to capture the heterogeneity, and identified associations with the patients' clinical characteristics extracted from EHRs.

b) Number of patients with data for each tissue and omics type after quality control. All 109 patients also have genome-wide genotype data. See **Supplementary Table 9** for patient-level details.

Supplementary Figure 2. Molecular differences between low-grade and high-grade cartilage

a) RNA- and protein-level log-fold-changes for all genes measured on both omics levels. The x-axis shows gene expression differences between low-grade and high-grade cartilage as quantified by limma in an analysis including the technical covariates as identified by pSVA as well as pairing samples from the same patients (see **Supplementary Methods**). The genes highlighted black or red were significant on both RNA- and protein-level, see also **Figure 1b**.

b) Signalling Pathway Impact Analysis (SPIA) identified biological pathways associated with low-grade/high-grade differences. All pathways shown are activated in high-grade compared to low-grade cartilage. Enrichment p: p-value from over-representation analysis of genes; Perturbation p: p-value for perturbation of the pathway based on gene log-fold-changes; Combined p: p-value from combining enrichment and perturbation p-values. Pathways with significant results at 5% FDR based on RNA-level changes are shown.

Supplementary Figure 3. Mouse models of implicated genes display osteoarthritis-relevant abnormal joint phenotypes.

a) Overview of the abnormal mouse joint phenotypes displayed by mouse models for 7 genes with differential expression between low-grade and high-grade cartilage. For each of the genes (in rows), 9 phenotypes were assayed on the lateral tibial plateau (LTP) and medial tibial plateau (MTP). For each joint parameter (in columns), the plot shows the ratio of the mean value of each mutant strain to the mean value of the wild-type background strain, where the differences were significant ($p<0.05$).

Borders around boxes show phenotype difference significant after multiple-testing correction for the effective number of tests within each line ($p<0.00568$). BMC: bone mineral content; BMD: bone mineral density; vol: volume. Technical details of each mouse mutant line and plots of all individual values for abnormal phenotypes at $p<0.05$ see **Supplementary Figure 4**.

b) Iodine contrast enhanced micro computed tomography (ICE-µCT) detects differences in articular cartilage volume and thickness (red volumes), and subchondral bone morphology (blue volumes); scale bar 100µm. *Matn4*$^{-/-}$ mice show decreased subchondral BV/TV (bone volume/tissue volume) and trabecular thickness (black arrow) compared to wild-type controls, whereas *Pdlim1*$^{-/-}$ mice display increased articular cartilage thickness (black arrow).

c) Joint surface replication (JSR) detects damage to the articulating surfaces of the tibial plateaux; scale bar 100um. *Htra3*$^{-/-}$ mice show increased articular cartilage surface damage (black arrows, C) compared to wild-type controls.

d) Subchondral X-ray microradiography (scXRM) detects changes in BMC within the subchondral region. White boxes represent the subchondral regions analysed; scale bar 1mm. *Matn4*$^{-/-}$ mice show decreased subchondral BMC compared to wild-type controls.

e) All abnormal phenotypes displayed by *Matn4*$^{-/-}$, *Pdlim1*$^{-/-}$ and *Htra3*$^{-/-}$ ($p<0.00568$). The boxplots show phenotype values for 100 mice from the background strain, with error bars for the 25%-75% interquartile range. Red diamonds: phenotypes of mutant mice. BV/TV: bone volume / tissue volume.

Supplementary Figure 4. Mouse models of implicated genes display osteoarthritis-relevant abnormal joint phenotype

a) Technical details for each mouse mutant line, including targeting method and allele name.

b) The figure shows all abnormal phenotypes displayed at $p<0.05$ which were not shown in the main figure. The boxplots show phenotype values for 100 mice from the background strain, with error bars for the 25%-75% interquartile range. Red diamonds: phenotypes of the mouse lines.

Supplementary Figure 5. Molecular QTLs in osteoarthritis disease tissue

a) eQTL overlap between tissues, for a total of 1,891 genes with a least one eQTL (left) and 219,709 eQTL gene-variant pairs (right). 49% of detected eQTLs are not tissue-specific.

b) High correlation of eQTL normalized effect sizes (NES) between low-grade and high-grade cartilage. Inset: Spearman correlation ρ=0.94 between NES effect sizes across all eQTLs.

c) High correlation of eQTL normalized effect sizes (NES) between low-grade cartilage and synovium (left), and between high-grade cartilage and synovium (right). Inset: Spearman correlation between NES effect sizes across all eQTLs.

d) pQTL overlap between tissues, for a total of 38 genes with a least one pQTL (left) and 3,211 pQTL protein-variant pairs (right).

e) High correlation of pQTL NES between low-grade and high-grade cartilage. Inset: Spearman correlation between NES effect sizes across all eQTLs.

f) Plots of GWAS and low-grade cartilage eQTL p-values for regions surrounding rs10502437 and rs12901372. For both GWAS signals, we observed colocalisation with eQTLs in both low-grade and high-grade cartilage, and the plots for high-grade cartilage are shown in **Figure 2d**. Each plot shows 1Mb region centered around the GWAS index SNP (purple); each point represents a genetic variant. Top panels show GWAS p-values, bottom panels QTL p-values for the indicated gene. Here and in **Figure 2d**, LD between variants was calculated using UK Biobank. PP4: posterior probability for colocalisation.

Supplementary Figure 6. Technical details of the clustering analysis of samples within tissues

a) Cluster consensus plots for clustering in low-grade cartilage, synovium, and high-grade cartilage based on RNA data. The x-axis shows the number k of clusters, the y-axis the cluster consensus value (higher values showing stronger clustering). For clustering in low-grade cartilage and synovium, but not high-grade cartilage, the cluster consensus value is above 0.8 for both clusters when k=2.
b) High-grade cartilage tissue samples from patients do not show a separation into two clusters by ConsensusCluster analysis based on RNA data.
c) Cluster tracking plots for low-grade cartilage and synovium based on RNA data. Each column is a sample, coloured by the cluster assignment when separating samples into k=2,…,10 clusters (k values in rows).
d) Low-grade cartilage tissue samples from patients do not show a separation into clusters by ConsensusCluster analysis based on protein data. k: number of clusters.
e) High-grade cartilage tissue samples from patients do not show a separation into clusters by ConsensusCluster analysis based on protein data. k: number of clusters.

Supplementary Figure 7. Distinct clusters identified in low-grade cartilage and synovium tissue

a) Gene expression differences between low-grade and high-grade cartilage do not depend on the low-grade cartilage cluster. Plots show log-fold-changes for all genes based on the analysis of all patients (x-axis) versus log-fold-changes for all genes based on the analysis of all patients with low-grade cartilage in only one of the two clusters (y-axis). In each within-cluster analysis, 99% of the genes significant in the all-patient analysis had the same direction of effect. Inset: Spearman correlation of log-fold-differences, $p<1.0\text{x}10^{-10}$.
b) Gene expression differences between the synovium sub-clusters within each cluster are highly correlated. Plot shows log-fold-changes of each gene in the comparison of sub-clusters within the larger (x-axis) and smaller (y-axis) cluster. Over 99% of the genes with significant differences between synovium-Cluster1a and synovium-Cluster1b also had directionally concordant differences between synovium-Cluster2b and synovium-Cluster2a, and over 80% were also significant at 5% FDR, and vice versa (i.e. genes with higher expression in synovium-Cluster1a compared to synovium-Cluster1b also had higher expression in synovium-Cluster2b compared to synovium-Cluster2a).
c) An analysis of synovium samples using MOFA identifies a continuous spectrum of variation between samples. This variation corresponds well to the clustering: MOFA

synovium factor 1 captures differences between sub-clusters, while synovium factor 2 captures variation between clusters.

Supplementary Figure 8. Multi-omics factor analysis (MOFA) RNA gene weights are correlated with gene expression differences between tissue clusters

a) Correlation between MOFA low-grade cartilage Factor 1 gene weights for RNA data and gene expression differences between low-grade cartilage clusters. logFC: log-fold-change. Inset: Spearman correlation, $p<10^{-15}$. Genes with significant differential expression between low-grade and high-grade cartilage are coloured red.
b) Gene expression differences between low-grade cartilage samples in the same clusters, divided by MOFA low-grade cartilage factor 1 values >0 and <0, are correlated with gene expression differences between clusters. logFC: log-fold-change. Inset: Spearman correlation, $p<10^{-10}$. Left shows results for low-grade cartilage-Cluster1, right for cartilage-Cluster2. Genes with significant differential expression between low-grade and high-grade cartilage are coloured red.
c) Correlation between MOFA synovium Factor 1 and 2 gene weights for RNA data and gene expression differences between synovium clusters and subclusters. logFC: log-fold-change. Inset: Spearman correlation ρ, $p<10^{-15}$.

C-Cluster: cartilage-Cluster.

Supplementary Figure 9. PAMR 7-gene low-grade cartilage classifier performance
PAMR diagnostic plots for a classifier of low-grade cartilage based on RNA. Left: Sample classification error based on the PAMR internal threshold and the corresponding number of genes in the classifier. The top panel shows the overall error estimate, the bottom panel error rates separately for cartilage-Cluster1 and cartilage-Cluster2. The optimal selection as used in the paper included 7 genes and an internal threshold of 6.67 (vertical line). Right: False Discover Rate (FDR) for between-cluster differences for the genes in the classifier as calculated by PAMR. C-Cluster: low-grade cartilage-Cluster.

Supplementary Information

Supplementary Methods and Supplementary Notes. This file contains the Supplementary Methods, details of mouse models and identification of abnormal joint phenotypes, a note on resolving GWAS signals, additional details for patient stratification analysis and the 7-gene classifier for low-grade cartilage.

Supplementary Tables

Supplementary Table 1. Genes with significant cross-omics differences between high-grade and low-grade cartilage.

Supplementary Table 2. Pathways and gene sets associated with significant RNA-level and/or protein-level differences between high-grade and low-grade cartilage.

Supplementary Table 3. Detailed list of differential eQTLs, i.e. variants with high posterior probability for presence of eQTL effect in high-grade cartilage (m>0.9) and low for presence in low-grade cartilage (m<0.1), or vice versa.

Supplementary Table 4. Genes with significantly different expression profiles between high-and low-grade cartilage that were also found to be associated with genetic risk of osteoarthritis in a recent GWAS (multiple-testing corrected significance threshold of $p<1.02x10^{-6}$).

Supplementary Table 5. Pathways associated with gene expression differences between low-grade cartilage clusters or between synovium clusters.

Supplementary Table 6. Expression differences between low-grade cartilage clusters for genes highlighted in previous cartilage clustering analyses.

Supplementary Table 7. Full association results between patient clinical characteristics and low-grade sample cluster assignment or synovium cluster assignment, including individual drugs assigned to drug classes.
The nine drug classes which were tested for association are shown (see **Methods**).

Supplementary Table 8. Comparison of perturbations by compounds, gene knockdown or overexpression, to differences between high-grade and low-grade cartilage.
The comparison was based on data from ConnectivityMap and genes with higher expression in high-grade than in low-grade cartilage on both RNA- and protein-level, see **Methods**.

Supplementary Table 9. List of all assayed patient tissue samples with detailed information including cohort, batch, and quality control exclusions.

References

1 Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388**, 1545-1602, doi:10.1016/S0140-6736(16)31678-6 (2016).

2 Mobasheri, A. *et al.* The role of metabolism in the pathogenesis of osteoarthritis. *Nat. Rev. Rheumatol.* **13**, 302, doi:10.1038/nrrheum.2017.50 (2017).

3 Murphy, L. *et al.* Lifetime risk of symptomatic knee osteoarthritis. *Arthritis Care Res.* **59**, 1207-1213, doi:10.1002/art.24021 (2008).

4 Murphy, L. B. *et al.* One in four people may develop symptomatic hip osteoarthritis in his or her lifetime. *Osteoarthr. Cartil.* **18**, 1372-1379, doi:10.1016/j.joca.2010.08.005 (2010).

5 Eurostat. *Surgical operations and procedures performed in hospitals*, <http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_co_proc2&lang=en> (2018).

6 Miosge, N. *et al.* Light and electron microscopicin situhybridization of collagen type I and type II mRNA in the fibrocartilaginous tissue of late-stage osteoarthritis. *Osteoarthr. Cartil.* **6**, 278-285, doi:10.1053/joca.1998.0121 (1998).

7 Stracke, J. O. *et al.* Matrix metalloproteinases 19 and 20 cleave aggrecan and cartilage oligomeric matrix protein (COMP). *FEBS Lett.* **478**, 52-56, doi:10.1016/s0014-5793(00)01819-6 (2000).

8 GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

9 Tachmazidou, I. *et al.* Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nat. Genet.* **51**, 230-236, doi:10.1038/s41588-018-0327-1 (2019).

10 Fernández-Tajes, J. *et al.* Genome-wide DNA methylation analysis of articular chondrocytes reveals a cluster of osteoarthritic patients. *Ann. Rheum. Dis.* **73**, 668-677, doi:10.1136/annrheumdis-2012-202783 (2014).

11 Soul, J. *et al.* Stratification of knee osteoarthritis: two major patient subgroups identified by genome-wide expression analysis of articular cartilage. *Ann. Rheum. Dis.* **77**, 423-423, doi:10.1136/annrheumdis-2017-212603 (2018).

12 Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124, doi:10.15252/msb.20178124 (2018).

13 Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6567-6572, doi:10.1073/pnas.082099299 (2002).

14 Murphy, G. & Lee, M. H. What are the roles of metalloproteinases in cartilage and bone damage? *Ann. Rheum. Dis.* **64**, iv44-iv47, doi:10.1136/ard.2005.042465 (2005).

15 Tanaka, T., Narazaki, M. & Kishimoto, T. IL-6 in inflammation, immunity, and disease. *Cold Spring Harb. Perspect. Biol.*, doi:10.1101/cshperspect.a016295 (2014).

16 López-Boado, Y. S., Tolivia, J. & López-Otín, C. Apolipoprotein D gene induction by retinoic acid is concomitant with growth arrest and cell differentiation in human breast cancer cells. *J. Biol. Chem.* **269**, 26871-26878, <http://www.jbc.org/content/269/43/26871.abstract> (1994).

17 Shepherd, C. *et al.* Functional characterization of the osteoarthritis genetic risk residing at *ALDH1A2* identifies rs12915901 as a key target variant. *Arthritis Rheumatol.* **70**, 1577-1587, doi:10.1002/art.40545 (2018).

18 Styrkarsdottir, U. *et al.* Severe osteoarthritis of the hand associates with common variants within the *ALDH1A2* gene and with rare variants at 1p31. *Nat. Genet.* **46**, 498-502, doi:10.1038/ng.2957s (2014).

19 Owens, William D., M.D., Felts, James A., M.D. & Spitznagel, Edward L., Ph.D. ASA physical status classifications: A study of consistency of ratings. *Anesthesiology* **49**, 239-243 (1978).

20 Bianchi, V. E. The anti-inflammatory effects of testosterone. *J. Endocr. Soc.* **3**, 91-107, doi:10.1210/js.2018-00186 (2018).

21 Gubbels Bupp, M. R. Sex, the aging immune system, and chronic disease. *Cell. Immunol.* **294**, 102-110, doi:10.1016/j.cellimm.2015.02.002 (2015).

22 Martín-Millán, M. & Castañeda, S. Estrogens, osteoarthritis and inflammation. *Joint Bone Spine* **80**, 368-373, doi:10.1016/j.jbspin.2012.11.008 (2013).

23 Aoki, T., Yamamoto, Y., Ikenoue, T., Onishi, Y. & Fukuhara, S. Multimorbidity patterns in relation to polypharmacy and dosage frequency: a nationwide, cross-sectional study in a Japanese population. *Sci. Rep.* **8**, 3806, doi:10.1038/s41598-018-21917-6 (2018).

24 Doos, L., Roberts, E. O., Corp, N. & Kadam, U. T. Multi-drug therapy in chronic condition multimorbidity: a systematic review. *Fam. Pract.* **31**, 654-663, doi:10.1093/fampra/cmu056 (2014).

25 Subramanian, A. *et al.* A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437-1452.e1417, doi:10.1016/j.cell.2017.10.049 (2017).

26 Roman-Blas, J. A., Castañeda, S., Largo, R. & Herrero-Beaumont, G. Osteoarthritis associated with estrogen deficiency. *Arthritis Res. Ther.* **11**, 241, doi:10.1186/ar2791 (2009).

27 de Klerk, B. M. *et al.* Limited evidence for a protective effect of unopposed oestrogen therapy for osteoarthritis of the hip: a systematic review. *Rheumatology (Oxford)* **48**, 104-112, doi:10.1093/rheumatology/ken390 (2009).

28 Watt, F. E. Hand osteoarthritis, menopause and menopausal hormone therapy. *Maturitas* **83**, 13-18, doi:10.1016/j.maturitas.2015.09.007 (2016).

29 pubchem.ncbi.nlm.nih.gov. *IB-Meca*, <https://pubchem.ncbi.nlm.nih.gov/compound/ib-meca> (2019).

30 pubchem.ncbi.nlm.nih.gov. *Nornicotine*, <https://pubchem.ncbi.nlm.nih.gov/compound/nornicotine> (2019).

31 Bar-Yehuda, S. *et al.* Induction of an antiinflammatory effect and prevention of cartilage damage in rat knee osteoarthritis by CF101 treatment. *Arthritis Rheumatol.* **60**, 3061-3071, doi:10.1002/art.24817 (2009).

32 Yuan, Q., Sun, L., Li, J.-J. & An, C.-H. Elevated VEGF levels contribute to the pathogenesis of osteoarthritis. *BMC Musculoskelet. Disord.* **15**, 437, doi:10.1186/1471-2474-15-437 (2014).

33 Nagao, M. *et al.* Vascular endothelial growth factor in cartilage development and osteoarthritis. *Sci. Rep.* **7**, 13027, doi:10.1038/s41598-017-13417-w (2017).

34 Takeshita, N. *et al.* Alleviating effects of AS1892802, a rho kinase inhibitor, on osteoarthritic disorders in rodents. *J. Pharmacol. Sci.* **115**, 481-489, doi:10.1254/jphs.10319FP (2011).

35 Kong, L., Wang, L., Meng, F., Cao, J. & Shen, Y. Association between smoking and risk of knee osteoarthritis: a systematic review and meta-analysis. *Osteoarthr. Cartil.* **25**, 809-816, doi:10.1016/j.joca.2016.12.020 (2017).

36 Chou, C. H. *et al.* Insights into osteoarthritis progression revealed by analyses of both knee tibiofemoral compartments. *Osteoarthr. Cartil.* **23**, 571-580, doi:10.1016/j.joca.2014.12.020 (2015).

37 Steinberg, J. *et al.* Integrative epigenomics, transcriptomics and proteomics of patient chondrocytes reveal genes and pathways involved in osteoarthritis. *Sci. Rep.* **7**, 8935, doi:10.1038/s41598-017-09335-6 (2017).

38 Steinberg, J. *et al.* Widespread epigenomic, transcriptomic and proteomic differences between hip osteophytic and articular chondrocytes in osteoarthritis. *Rheumatology (Oxford)* **57**, 1481-1489, doi:10.1093/rheumatology/key101 (2018).

39 Hawtree, S., Muthana, M., Wilkinson, J. M., Akil, M. & Wilson, A. G. Histone deacetylase 1 regulates tissue destruction in rheumatoid arthritis. *Hum. Mol. Genet.* **24**, 5367-5377, doi:10.1093/hmg/ddv258 (2015).

40 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

41 Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13, doi:10.1186/1751-0473-9-13 (2014).

42 Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. (2010).

43 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, doi:10.1038/nmeth.4197 (2017).

44 Soneson, C., Love, M. & Robinson, M. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 1; referees: 2 approved]. *F1000Research* **4**, 1521, <http://f1000r.es/6a7> (2015).

45 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).

46 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284-1287, doi:10.1038/ng.3656 (2016).

47 Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13, doi:10.1186/s13059-016-0881-8 (2016).

48 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47, doi:10.1093/nar/gkv007 (2015).

49 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**, 4288-4297, doi:10.1093/nar/gks042 (2012).

50 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

51 Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161, doi:10.1093/nar/gku864 (2014).

52 Tarca, A. L. *et al.* A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75-82, doi:10.1093/bioinformatics/btn577 (2008).

53 Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14, doi:10.1186/gb-2010-11-2-r14 (2010).

54 Brown, A. A. *et al.* Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet.* **49**, 1747-1751, doi:10.1038/ng.3979 (2017).

55 Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500-507, doi:10.1038/nprot.2011.457 (2012).

56 Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-1485, doi:10.1093/bioinformatics/btv722 (2016).

57 Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLOS Genet.* **9**, e1003491, doi:10.1371/journal.pgen.1003491 (2013).

58 Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLOS Genet.* **8**, e1002555, doi:10.1371/journal.pgen.1002555 (2012).

59 Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genet.* **10**, e1004383, doi:10.1371/journal.pgen.1004383 (2014).

60 de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLOS Comput. Biol.* **11**, e1004219, doi:10.1371/journal.pcbi.1004219 (2015).

61 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29, doi:10.1186/gb-2014-15-2-r29 (2014).

62 Parker, H. S. *et al.* Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics* **30**, 2757-2763, doi:10.1093/bioinformatics/btu375 (2014).

63 Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573, doi:10.1093/bioinformatics/btq170 (2010).