OPEN ACCESS

EPIDEMIOLOGICAL SCIENCE

# Linking chondrocyte and synovial transcriptional profile to clinical phenotype in osteoarthritis

Julia Steinberg ,[1,2,3] Lorraine Southam,[1,3] Andreas Fontalis,[4] Matthew J Clark,[4] Raveen L Jayasuriya,[4] Diane Swift,[4] Karan M Shah,[4] Roger A Brooks,[5] Andrew W McCaskie,[5] Jeremy Mark Wilkinson ,[4,6] Eleftheria Zeggini [1,3,7]

**Handling editor** Josef S Smolen

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/annrheumdis-2020-219760).

For numbered affiliations see end of article.

**Correspondence to**
Professor Eleftheria Zeggini, Institute of Translational Genomics, Helmholtz Zentrum Munchen Deutsches Forschungszentrum fur Gesundheit und Umwelt, Neuherberg 85764, Germany; eleftheria.zeggini@helmholtz-muenchen.de
Professor Jeremy Mark Wilkinson; j.m.wilkinson@sheffield.ac.uk

JMW and EZ contributed equally.

Received 19 December 2020
Revised 30 March 2021
Accepted 11 April 2021



**To cite:** Steinberg J, Southam L, Fontalis A, *et al*. *Ann Rheum Dis* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/annrheumdis-2020-219760

## ABSTRACT

**Objectives** To determine how gene expression profiles in osteoarthritis joint tissues relate to patient phenotypes and whether molecular subtypes can be reproducibly captured by a molecular classification algorithm.

**Methods** We analysed RNA sequencing data from cartilage and synovium in 113 osteoarthritis patients, applying unsupervised clustering and Multi-Omics Factor Analysis to characterise transcriptional profiles. We tested the association of the molecularly defined patient subgroups with clinical characteristics from electronic health records.

**Results** We detected two patient subgroups in low-grade cartilage (showing no/minimal degeneration, cartilage normal/softening only), with differences associated with inflammation, extracellular matrix-related and cell adhesion pathways. The high-inflammation subgroup was associated with female sex (OR 4.12, p=0.0024) and prescription of proton pump inhibitors (OR 4.21, p=0.0040). We identified two independent patient subgroupings in osteoarthritis synovium: one related to inflammation and the other to extracellular matrix and cell adhesion processes. A seven-gene classifier including *MMP13, APOD, MMP2, MMP1, CYTL1*, *IL6* and *C15orf48* recapitulated the main axis of molecular heterogeneity in low-grade knee osteoarthritis cartilage (correlation $\rho$=−0.88, p<$10^{-10}$) and was reproducible in an independent patient cohort ($\rho$=−0.85, p<$10^{-10}$).

**Conclusions** These data support the reproducible stratification of osteoarthritis patients by molecular subtype and the exploration of new avenues for tailored treatments.

## Key messages

**What is already known about this subject?**

- Osteoarthritis (OA) is a disease with both clinical and molecular heterogeneity.
- However, it remains unclear whether molecular subgroupings of patients vary between joint tissue types, how they relate to clinical characteristics and whether they can be reproducibly captured by a molecular classification algorithm.

**What does this study add?**

- We carried out the first in-depth characterisation of molecular heterogeneity using patient synovium and cartilage in the largest cohort to date.
- We detected two patient subgroups based on low-grade (largely intact) OA cartilage, which were associated with sex and proton pump inhibitor prescription. Patient subgroups in synovium were associated with inflammation and, separately, extracellular matrix and cell adhesion, and were independent of the low-grade OA cartilage subgroups.
- A seven-gene classifier reproducibly recapitulated both the discrete assignment of knee low-grade OA cartilage subgroups and the main continuous spectrum of variation within the tissue.

**How might this impact on clinical practice or future developments?**

- These data demonstrate that molecular tissue profile in OA is associated with patient clinical characteristics, that this profile can be characterised using a limited panel of genes, and support the case for precision medicine approaches in OA.

## INTRODUCTION

Our understanding of the molecular mechanisms that underlie the observed epidemiological and clinical heterogeneity in osteoarthritis (OA) is incomplete. The accessibility of primary disease tissues at the point of joint replacement surgery provides the opportunity to stratify patients based on tissue-specific molecular profiles. Such stratification may help provide mechanistic insights into the molecular processes underlying the disease and subsequently develop novel tailored treatments. By examining expression profiles in cartilage, two previous studies in small cohorts have identified two subgroups of patients: microarray data from 23 patients suggested subgroups with gene expression differences related to inflammatory response, leucocyte activation, regulation of cytokine production and chemokine activity[1]; RNA sequencing data from 44 patients also suggested two subgroups, but with differences related to oxidative stress, innate-immune responses, Wnt signalling and chemokine signalling rather than classical inflammation.[2]

Several questions emerge from these studies: Do molecular profiles in different disease-relevant tissues define the same patient subgroups? Is the molecular subgrouping associated with any clinical characteristics? Is the molecular subgrouping

**Table 1** Characteristics of patients in discovery and validation cohorts

| Variable | Discovery total | Discovery cohort 1 | Discovery cohort 2 | Discovery cohort 3 | Discovery cohort 4 | Validation |
|---|---|---|---|---|---|---|
| No of patients (after QC) | **113 (106)** | 12 (11) | 20 (17) | 11 (10) | 70 (68) | **60 (60)** |
| Osteoarthritis joint | **Knee or hip** | Knee | Knee | Hip | Knee | **Knee** |
| Low-grade OA cartilage samples after QC | **87** | 11 | 16 | 10 | 50 | **60** |
| High-grade OA cartilage samples after QC | **95** | 10 | 16 | 10 | 59 | – |
| Synovium samples after QC | **77** | – | 16 | – | 61 | – |
| Females, n (%) after QC | **63 (59)** | 2 (19) | 12 (71) | 8 (73) | 41 (60) | **27 (45)** |
| Age, average years (range) after QC | **69 (38–88)** | 69 (50–88) | 70 (54–79) | 61 (44–88) | 70 (38–84) | **72 (63–85)** |

OA, osteoarthritis; QC, quality control.

truly categorical or better reflected by a continuous spectrum of variation? Can a molecular classification algorithm reproducibly recapitulate both categorical subgrouping and the main continuous spectrum of variation? Here, we examine these questions through genome-wide transcriptional profiling of multiple primary patient tissues (low-grade OA cartilage, high-grade OA cartilage and synovium), integrating information from electronic health records, and substantially increasing the cohort size and hence power (doubling the size of the discovery sample and almost tripling the total number of patients with genome-wide data compared with past studies).

## MATERIAL AND METHODS

### Patients

We analysed tissue samples from 113 patients undergoing total joint replacement surgery (table 1). All patients gave informed written consent prior to participation (online supplemental methods). Matched low-grade and high-grade OA cartilage samples (ie, largely intact vs degraded tissue, respectively; see online supplemental methods) were collected from each patient and synovial lining samples from 90 knee OA patients (table 1). All cartilage samples were collected from weight-bearing areas of the joint, ensuring that any differences observed between low-grade and high-grade OA cartilage reflected disease severity rather than differential mechanical loading. Cartilage scoring, isolation of chondrocytes and synoviocytes, and RNA extraction are described in online supplemental methods.

For knee OA patients, we obtained clinical characteristics and prescribed drugs from the electronic patient records (online supplemental methods).

### RNA sequencing

Multiplexed libraries were sequenced on the Illumina HiSeq 2000 or HiSeq 4000 (75bp paired-end reads). After quality control, we applied salmon 0.8.2,[3] and tximport[4] to obtain gene-level scaled transcripts per million estimates (online supplemental methods).

### Molecular subgroups

We applied limma-voom[5] to remove heteroscedasticity from gene expression data, permuted Surrogate Variable Analysis (pSVA)[6] to remove technical variation and explicitly regressed out effects of known sample collection and sequencing batches. We then applied ConsensusClusterPlus[7] to identify discrete molecularly defined subgroups ('clusters') of patients, with a sensitivity analysis to verify that clustering was similar when restricting to knee OA patients only (online supplemental text). We tested differential gene expression between clusters using limma,[8] and applied SPIA[9] and GOseq[10] to identify associated biological processes (online supplemental methods).

### Associations between molecular clusters and clinical characteristics

We tested for association between cluster assignment and clinical characteristics using a generalised linear model and applied a Bonferroni multiple-testing correction for the effective number of tests ($p<0.0047$, online supplemental methods). As sensitivity analyses, we successively added the following covariates to the models: sex, age, OA joint, body mass index (online supplemental methods).

### In-depth characterisation of molecular heterogeneity

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).[11] We investigated the correspondence between the discrete clusters and the continuous spectrum of variation identified by the MOFA and carried out extensive sensitivity analyses to verify robustness of the MOFA results (online supplemental methods).

### Low-grade OA cartilage classifier

We used the 'Prediction analysis for microarrays for R' (PAMR)[12] package to construct a classifier which used a smaller subset of genes to recapitulate the main axis of molecular heterogeneity in knee low-grade OA cartilage (online supplemental methods). We validated the classifier using an independent publicly-available dataset from 60 knee OA patients[2] (table 1, online supplemental methods).

## RESULTS

After quality checks, we analysed transcriptomic profiles of 259 tissue samples from 106 patients (table 1).

### Do molecular profiles in different tissues define the same patient subgroups?

Using consensus clustering, we identified two robust patient clusters in synovium, each of which further formed two subclusters (figure 1A). We also identified two robust patient clusters within low-grade, but not high-grade OA, cartilage (figure 1B, online supplemental figure 1). Cartilage clustering was independent of the synovium clusters (Fisher's $p>0.66$), and not associated with patient cohort nor with sequencing batches ($\chi^2$ test, $p>0.96$).

Signalling Pathway Impact Analysis[9] showed that the differences between the two synovium patient clusters relate to inflammation, while differences between the sub-clusters relate to the extracellular matrix and to cell adhesion (figure 1C,D, online supplemental figure 2, table 1 and text). The differences between low-grade OA cartilage clusters were also strongly associated with inflammation, extracellular matrix-related and cell adhesion pathways (figure 1E, online supplemental table 1).

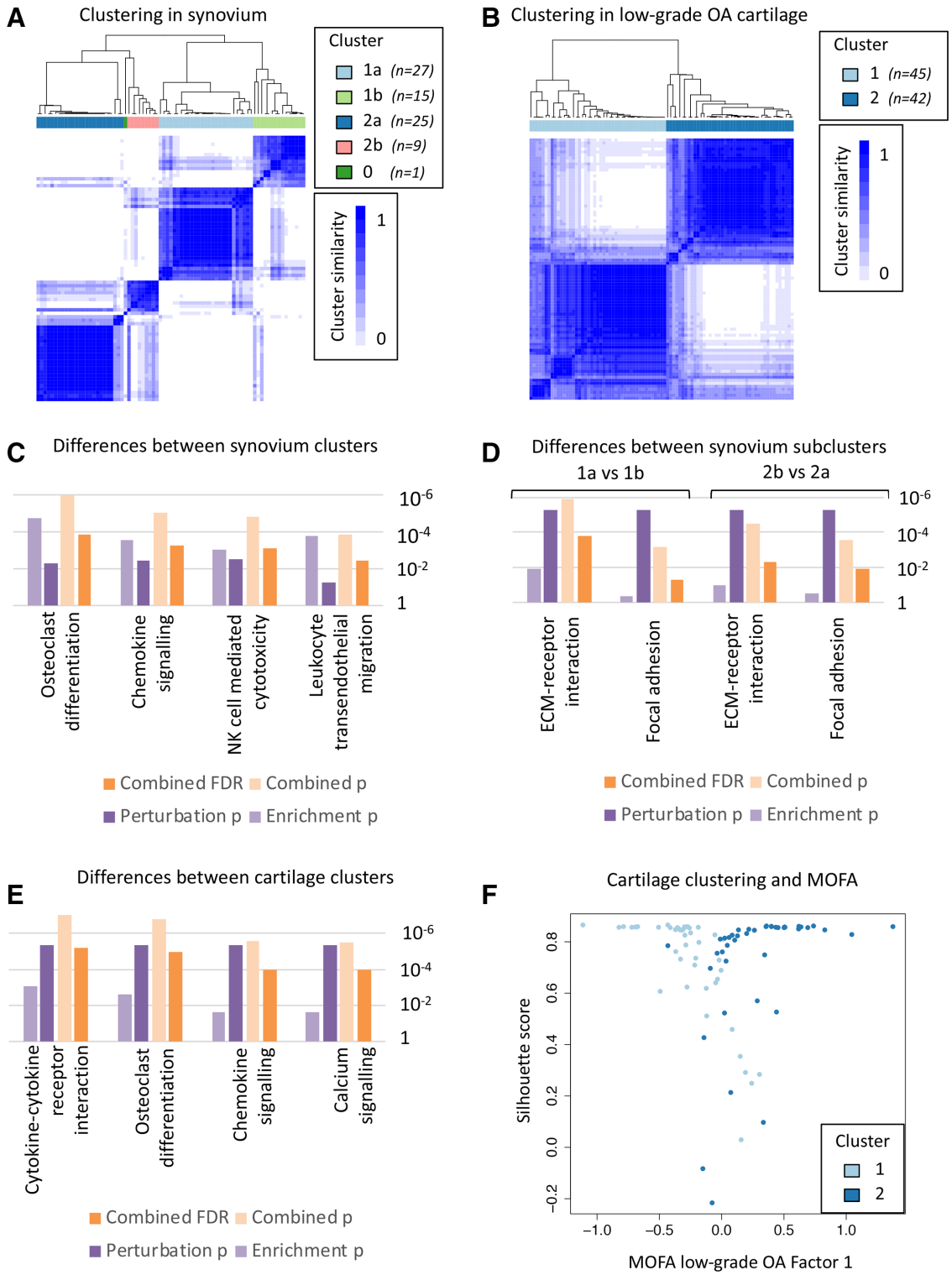


**Figure 1** Distinct molecularly defined patient clusters identified in low-grade OA cartilage and synovium tissue. (A) Two clusters of patients based on consensus clustering of synovium RNA data. Each cluster formed two subclusters, with one outlier sample. (B) Two clusters of patients based on consensus clustering of low-grade OA cartilage RNA data. (C) Gene expression differences between synovium clusters show several significant (false discovery rate <5%) associations related to inflammation and osteoclast differentiation using Signalling Pathway Impact Analysis (SPIA). Here and below, P: p values based on enrichment of genes; perturbation of the pathway based on gene log-fold differences; or combining enrichment and perturbation. The associations shown are robust across several gene-level differential expression cut-offs (online supplemental table 1). (D) Gene expression differences between the synovium subclusters within each cluster show similar pathway associations, including to ECM–receptor interaction and focal adhesion pathways. (E) Gene expression differences between low-grade OA cartilage clusters show significant associations with inflammation and osteoclast differentiation pathways. (F) An analysis of low-grade OA cartilage samples using MOFA identifies a continuous spectrum of variation between samples, which corresponds to the identified clusters. Samples with intermediate MOFA factor 1 scores have lower Silhouette scores, showing more uncertainty in cluster assignment. For synovium, see online supplemental figure 3. ECM, extracellular matrix; FDR, false discovery rate; MOFA, Multi-Omics Factor Analysis; OA, osteoarthritis.

### Is the molecular subgrouping associated with clinical characteristics?

We found that women were more likely to be members of the cartilage cluster characterised by higher inflammation (OR=4.12, p=0.0024; online supplemental table 2 and text). Patients in the high-inflammation cluster were also more likely to be prescribed proton pump inhibitors (PPIs; OR=4.21, p=0.0040; online supplemental text). We did not detect significant associations between synovium clustering and clinical characteristics (online supplemental table 2 and text).

### Is the molecular clustering categorical or continuous?

MOFA[11] identified continuous axes of variation within synovium and low-grade OA cartilage tissue that correspond strongly with cluster assignment (figure 1F, online supplemental figures 3 and 4, online supplementary text). In low-grade OA cartilage, the first MOFA factor (ie, the main axis of variation) explained 28% of variation in gene expression levels; the gene expression weights for this first factor and the log-fold-differences between clusters had very high correlation (Pearson correlation r=0.91, $p<10^{-15}$; online supplemental figure 4). These findings were also recapitulated in synovium (r 0.83–0.96 for gene weights for the first two MOFA factors and the log-fold-differences between synovium clusters and subclusters; online supplemental figure 4). This suggests that the variation within these tissues is better represented as a continuous spectrum.

We verified robustness of the MOFA results using extensive sensitivity analyses (including restricting analysis to knee OA patients only or explicitly regressing out age and sex effects; online supplemental text).

### Can a classifier reproducibly recapitulate both categorical clustering and the main axis of variation?

We used a soft-thresholding centroid-based method, PAMR,[12] to construct a tool that can recapitulate the clustering and main axis of heterogeneity in low-grade OA cartilage. As clinical and research applications would likely differ between OA joints, we restricted the analysis to patients with knee OA. The resulting tool predicts probabilities of knee low-grade OA cartilage cluster assignment based on the expression levels of seven genes (figure 2A, online supplemental figure 5, Data availability): *MMP1*, *MMP2* and *MMP13*, which are involved in cartilage degradation[13]; *IL6*, a proinflammatory cytokine; *CYTL1*, a cytokine-like gene, loss of which has been found to augment cartilage destruction in surgical OA mouse models[14]; *APOD*, a component of high-density lipoprotein found to be strongly upregulated by retinoic acid,[15] which is in turn regulated by *ALDH1A2*,[16] an OA risk locus[17 18] and *C15orf48*, of currently unknown function. Notably, the probabilities for cluster assignment generated by the classifier captured the main continuous spectrum of variation in this tissue (Spearman's correlation $\rho=-0.88$, $p<10^{-10}$; figure 2B). We validated the seven-gene classifier in an independent gene expression dataset of low-grade OA cartilage samples from 60 knee OA patients undergoing joint replacement surgery.[2] The posterior probabilities for cluster assignment had good correspondence to the main continuous spectrum of variation in the validation samples, supporting the predictive potential of the seven-gene classifier ($\rho=-0.85$, $p<10^{-10}$; figure 2C).

We also found that the seven-gene classifier had improved generalisability compared with a classifier developed in previous work[2]: the majority of genes in the previously developed classifier showed either discordant expression differences between the clusters in our larger dataset or high false discovery rates (>30%; online supplemental table 3 and text).

## DISCUSSION

Our findings indicate that molecular heterogeneity in OA cartilage and synovium is associated with similar biological processes (including inflammation), but molecularly defined patient clusters differ between tissues, potentially reflecting differences in tissue-specific dominant disease processes.

The clustering in low-grade OA cartilage agrees with two previous smaller studies.[1 2]

We also identified an association between the cartilage high-inflammation cluster and female sex, which is consistent with the

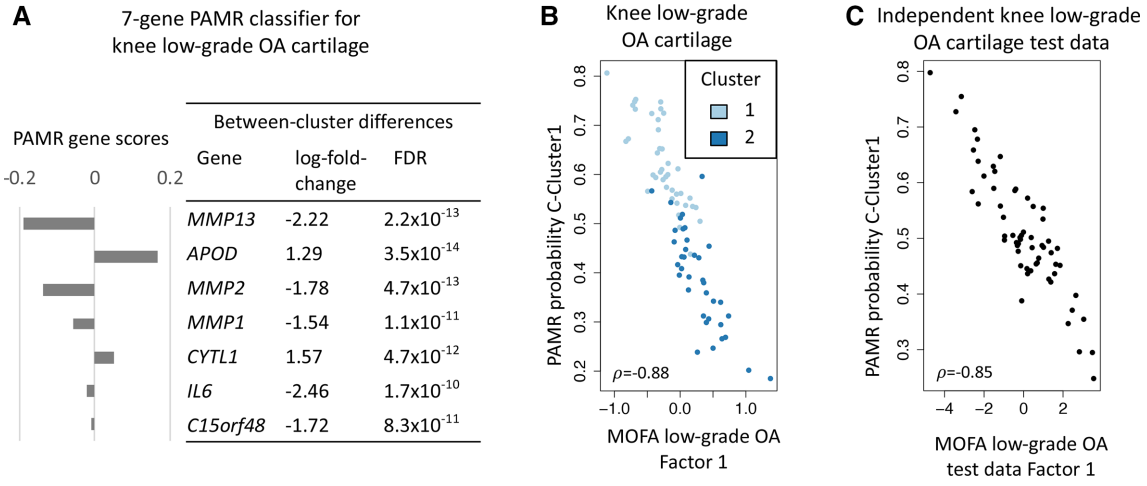


**Figure 2** Clustering and main axis of variation within knee low-grade OA cartilage can be recapitulated using a seven-gene classifier. (A) PAMR scores for each gene in the seven-gene knee OA classifier (the difference between the standardised centroids of the two clusters) and the differential expression of the genes between the two low-grade OA cartilage clusters. See online supplemental figure 5 for classifier performance. (B) The PAMR posterior probabilities for cluster assignment are highly correlated with MOFA factor 1 scores for knee low-grade OA cartilage samples, capturing the main continuous spectrum of variation between samples. Inset: Spearman correlation, $p<10^{-10}$. (C) In an independent set of 60 low-grade OA cartilage samples from 60 knee OA patients, the posterior probabilities for cluster assignment from the seven-gene classifier are well correlated with the continuous spectrum of variation in these samples, as quantified by the first MOFA factor in an ab initio analysis. Inset: Spearman's correlation, $p<10^{-10}$. IL, interleukin; MOFA, Multi-Omics Factor Analysis; OA, osteoarthritis.

disproportionate increase in the incidence of OA in women after the menopause. This association might be explained by the lower concentration of oestrogen and androgens (which have established anti-inflammatory effects) in postmenopausal women.[19 20] We speculate that our observed association between the high-inflammation cluster and PPI use could be explained by over the counter use of non-steroidal anti-inflammatory drugs for which PPIs are commonly coprescribed. We did not see discrete subgrouping in high-grade OA cartilage, perhaps indicating that there is less clear variation in molecular profiles in cartilage with advanced degeneration.

Our MOFA results further confirmed that the main axis of variation was related to inflammation in both synovium and cartilage. The seven-gene classifier generated using PAMR was able to place knee OA patients along the inflammatory endotype axis of variation and confirmed that such classification reflects continuous variation rather than categorical clustering, with validation in independent data. This finding has implications for the development of therapeutic strategies for OA, providing empirical evidence that responses might be expected to be heterogeneous along an axis of variation, rather than discrete. However, further study will be required to determine to what extent the inflammation axis is present in earlier disease stages, is stable across time or differs with disease activity, which cartilage scoring system is best suited to detect this axis, and whether the classifier can be applied to or modified for peripheral tissue (eg, saliva or blood). We anticipate that, looking ahead, this approach could underpin tailored therapy development and help improve patient care.

**Author affiliations**
[1]Institute for Translational Genomics, Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt, Neuherberg, Germany
[2]Daffodil Centre, The University of Sydney, a joint venture with Cancer Council NSW, Sydney, New South Wales, Australia
[3]Wellcome Sanger Institute, Hinxton, UK
[4]Department of Oncology and Metabolism, The University of Sheffield, Sheffield, UK
[5]Division of Trauma & Orthopaedic Surgery, University of Cambridge, Cambridge, UK
[6]Centre for Integrated Research into Musculoskeletal Ageing and Sheffield Healthy Lifespan Institute, University of Sheffield, Sheffield, UK
[7]Translational Genomics, Klinikum rechts der Isar der Technischen Universitat Munchen, Munchen, Germany

**Twitter** Julia Steinberg @SteinbergJulia

**Acknowledgements** We thank the study participants who made this work possible by their generous donation of samples. The authors are grateful to Dr Iris Fischer for helpful edits.

**Contributors** Study design: EZ, JMW and JS. Collection of knee samples: MJC, RLJ, DS, KMS and JMW. Collection of hip samples: RAB and AWM. Review of patient electronic health record data: AF and JMW. Patient stratification, Multi-Omics Factor Analysis, differential expression, pathway association and statistical analyses: JS. Writing – original draft: JS, LS, JMW and EZ. Writing – comments and review: all authors.

**Funding** This work was funded by the Wellcome Trust (206194). MJC was funded through a Centre for Integrated Research into Musculoskeletal Ageing grant (MRC 148985). RAB and the Human Research Tissue Bank are supported by the NIHR Cambridge Biomedical Research Centre. AM receives funding from Versus Arthritis; Tissue Engineering and Regenerative Therapies Centre (21156).

**Competing interests** None declared.

**Patient and public involvement statement** The biobank under which this project was conducted is overseen by a steering committee that includes two lay members who reviewed this project proposal prior to its initiation. The lay committee members had the opportunity to comment upon and make edits to the study design, as did the Sheffield Lay Advisory Panel for Bone Research (LAPBR). The conduct of the biobank and its outputs are also reviewed by the biobank lay committee members.

**Patient consent for publication** Not required.

**Ethics approval** The work was approved by the National Research Ethics Service (15/SC/0132).

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** The RNA sequencing data reported in this paper have been deposited to the EGA (accession numbers EGAS00001002255, EGAD00001003355, EGAD00001003354, EGAD00001001331). Further data including results from consensus clustering, Multi-Omics Factor Analysis and PAMR analyses, as well as the scripts for PAMR classifier construction and application to test data, can be obtained online from https://hmgubox.helmholtz-muenchen.de/d/f5be29c5123244359f73/.



**ORCID iDs**
Julia Steinberg http://orcid.org/0000-0002-0585-2312
Jeremy Mark Wilkinson http://orcid.org/0000-0001-5577-3674
Eleftheria Zeggini http://orcid.org/0000-0003-4238-659X

## REFERENCES

1 Fernández-Tajes J, Soto-Hermida A, Vázquez-Mosquera ME, *et al*. Genome-Wide DNA methylation analysis of articular chondrocytes reveals a cluster of osteoarthritic patients. *Ann Rheum Dis* 2014;73:668–77.
2 Soul J, Dunn SL, Anand S, *et al*. Stratification of knee osteoarthritis: two major patient subgroups identified by genome-wide expression analysis of articular cartilage. *Ann Rheum Dis* 2018;77:423.
3 Patro R, Duggal G, Love MI, *et al*. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;14:417–9.
4 Soneson C, Love MI, Robinson MD. Differential analyses for RNA-Seq: transcript-level estimates improve gene-level inferences. *F1000Res* 2015;4:1521.
5 Law CW, Chen Y, Shi W, *et al*. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;15:R29.
6 Parker HS, Leek JT, Favorov AV, *et al*. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics* 2014;30:2757–63.
7 Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 2010;26:1572–3.
8 Ritchie ME, Phipson B, Wu D, *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43:e47.
9 Tarca AL, Draghici S, Khatri P, *et al*. A novel signaling pathway impact analysis. *Bioinformatics* 2009;25:75–82.
10 Young MD, Wakefield MJ, Smyth GK, *et al*. Gene ontology analysis for RNA-Seq: accounting for selection bias. *Genome Biol* 2010;11:R14.
11 Argelaguet R, Velten B, Arnol D, *et al*. Multi-Omics factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;14:e8124.
12 Tibshirani R, Hastie T, Narasimhan B, *et al*. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* 2002;99:6567–72.
13 Murphy G, Lee MH. What are the roles of metalloproteinases in cartilage and bone damage? *Ann Rheum Dis* 2005;64 Suppl 4:iv44–7.
14 Tanaka T, Narazaki M, Kishimoto T. Il-6 in inflammation, immunity, and disease. *Cold Spring Harb Perspect Biol* 2014;6:a016295.
15 López-Boado YS, 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* 1994;269:26871–8.
16 Shepherd C, Zhu D, Skelton AJ, *et al*. Functional characterization of the osteoarthritis genetic risk residing at *ALDH1A2* identifies rs12915901 as a key target variant. *Arthritis Rheumatol* 2018;70:1577–87.
17 Styrkarsdottir U, Thorleifsson G, Helgadottir HT, *et al*. Severe osteoarthritis of the hand associates with common variants within the *ALDH1A2* gene and with rare variants at 1p31. *Nat Genet* 2014;46:498–502.
18 Tachmazidou I, Hatzikotoulas K, Southam L, *et al*. Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nat Genet* 2019;51:230–6.
19 Gubbels Bupp MR. Sex, the aging immune system, and chronic disease. *Cell Immunol* 2015;294:102–10.
20 Martín-Millán M, Castañeda S, Estrogens CS. Estrogens, osteoarthritis and inflammation. *Joint Bone Spine* 2013;80:368–73.