Cross-species gene modules emerge from a systems biology approach to osteoarthritis

Alan James Mueller1, Elizabeth G. Canty-Laird1,2, Peter D. Clegg1,2 and Simon R. Tew1,2

Complexities in degenerative disorders, such as osteoarthritis, arise from multiscale biological, environmental, and temporal perturbations. Animal models serve to provide controlled representations of the natural history of degenerative disorders, but in themselves represent an additional layer of complexity. Comparing transcriptomic networks arising from gene co-expression data across species can facilitate an understanding of the preservation of functional gene modules and establish associations with disease phenotypes. This study demonstrates the preservation of osteoarthritis-associated gene modules, described by immune system and system development processes, across human and rat studies. Class prediction analysis establishes a minimal gene signature, including the expression of the Rho GDP dissociation inhibitor ARHGDIB, which consistently defined healthy human cartilage from osteoarthritic cartilage in an independent data set. The age of human clinical samples remains a strong confounder in defining the underlying gene regulatory mechanisms in osteoarthritis; however, defining preserved gene models across species may facilitate standardization of animal models of osteoarthritis to better represent human disease and control for ageing phenomena.

npj Systems Biology and Applications (2017) 3:13 ; doi:10.1038/s41540-017-0014-3

ARTICLE

INTRODUCTION

Disorders of cartilage and joints account for a high incidence of disability1 and are prevalent co-morbidities of the ageing population.2 Debilitating in their own right musculoskeletal disorders contributes significantly to the global burden of disease, being the fourth most prevalent disorder3 and have a wider impact on the rehabilitation of co-occurring pathologies (obesity, stroke, and cardiovascular disease), thereby representing a major health policy issue.

Osteoarthritis (OA), considered a chronic, degenerative condition of multiple tissues that comprise a joint,4 results in the destruction of cartilage, the friction-free interface, leading to considerable functional impairment. The major cell population of cartilage, chondrocytes, account for the unique extracellular matrix (ECM), which confers compression resistance and gliding characteristics of normal cartilage.5 Despite considerable efforts to characterize the nature of degenerate cartilage the pathophysiological process is not fully understood, disease-associated genetic variants are limited, and there are no disease-modifying therapeutics available.6

Disease complexity, arising from multiscale perturbations, makes a mechanistic understanding of OA difficult. OA is a complex disease because it involves multiple tissues, environmental factors, behaviors, signaling pathways and genes. For example, numerous genetic risk loci, epigenetic effects, inflammation associated with ageing7 and obesity8 and biomechanical factors contribute to joint degeneration. Although heritable factors account for 50% of an individual’s risk of developing OA, only 16 disease risk loci have been consistently identified9 with candidate genes such as GDF5 and SMAD3 harboring the most promising risk alleles;10 overall, multiple risk alleles are likely to contribute to OA susceptibility. Additionally, OA is dynamic, being progressive and chronic, and so is likely to involve the dysregulation of several biological systems over multiple time-scales. As with other multifactorial diseases (e.g., neurological disorders), analysis of individual components cannot adequately explain the properties of the whole system (the contributing tissues) as novel properties emerge with increasing complexity of the system.11

Animal models of multifactorial disorders are used to provide a controlled representation of subsets of human disease and aim to reproduce the natural history and progression. Rodent models of cartilage pathophysiology are frequently employed and include surgical-induced (destabilization of the medial meniscus) and chemical-induced (monioiodoacetate joint injection) OA. The rat is frequently used in the study of OA; however, there is no single standardized in vivo model.12 Gene expression studies arising from these models are often poorly controlled, underpowered, combine joint tissues, and use multiple different gene expression analysis platforms, making comparison across studies problematic. Overall, animal models that better represent human OA are required.13

Weighted gene co-expression network analysis14 is a systems biology methodology that considers the connectivity between genes based upon their co-expression. This method facilitates investigation of the global network properties of a transcriptome and provides functional insights into the organization of a co-expression network by utilizing the concept of scale-free networks.15 As the co-expression of genes encodes the downstream protein interactions, the study of transcriptional co-expression patterns can reveal emergent functional properties of the cellular system under investigation. Weighted gene co-expression
network analysis (WGCNA) has been used widely to define candidate genes for human disorders including prognostic signatures for cancers, and has demonstrated preservation of functional gene modules between human and mouse brains. In this context network, nodes represent genes that are expressed in a sample. Edges connect nodes based upon their weighted co-expression across samples. WGCNA assumes that all nodes are connected and the connections have different strengths; highly connected genes within a network can be gathered as modules, with “hubs” being the most highly connected genes within a module. The modularity of networks is inherent to cell biology, and biological phenomena arise from molecular interactions organized into functional modules. The network topology (or architecture of these module structures) can be compared across networks to assess conservation of modules in different conditions or between species.

The system under consideration in this study was the chondrocyte, either as whole cartilage or isolated cells. Transcriptomic profiling from different environments and conditions provided information on perturbations to that system. The study sought to establish, from publically available gene expression data, a comprehensive analysis of the gene–gene co-expression networks from transcriptomic profiles of different chondrocyte phenotypes in human and rat. By performing this analysis on human and rat data, an understanding of the preservation of network module topology would inform the validity of rodent in vivo models of OA. Additionally, by establishing a subnetwork of genes associated with the phenotype of interest, osteoarthritic cartilage, rational therapeutic and diagnostic targets may be proposed for future study. This study demonstrates high preservation of modules across species associated with physiological functions in addition to modules associated with inflammatory mediators and system development that are characteristic of a subset of human osteoarthritic cartilage samples. Importantly, genes with class discrimination potential were established that may serve to define early cartilage degeneration.

RESULTS

Construction of rat and human co-expression networks

Global co-expression networks were constructed from rat (115 arrays) and human (129 arrays) gene expression data using 5982 genes with common annotation by WGCNA. Overall, 12 modules were defined for the human network (Fig. 1a, b) and 20 modules for the rat (Fig. 2a, b), inclusive of a module of unassigned genes for each species. All further characterizations were undertaken on these modules. An alphanumeric code for each module (H-human, R-rat) is provided for reference, Supplementary Figs. 4a and 5a. The genes assigned to each module are listed Supplementary Data SD1 and SD2.

Consensus modules established between species

To establish concordance and divergence in the network organization across the two species, consensus modules, modules shared by both networks, were derived from each species network using the weighted average of the two correlation matrices. In all, there were six consensus modules (C1–C6), Figs. 1b and 2b and Supplementary Figs. 4b and 5b. Functional annotation of the consensus modules included: “muscle system process” (C1), “oxidative phosphorylation” (C2), “cell cycle” (C3), “system development” (C4), and “immune system process” (C5). Unassigned genes were placed in a single module (C6, colored “gray”). See Supplementary Data SD36–41 for full annotations and gene lists.

Module–trait relationships define subsets of whole-cartilage samples in both species

To understand the possible functions of these modules and their relevance to chondrocyte dysregulation, the association between each module eigengene (ME) and traits of interest were established. Traits groups were defined as samples derived from similar experimental conditions (e.g., monolayer expansion of chondrocytes), anatomical sites (e.g., growth plate zones), and phenotypic groups (e.g., different ages); MEs were representative genes for each module. Whole-cartilage clinical samples for human were obtained at surgery (humans, age range: pre-adolescent–82 years old) and defined as healthy or osteoarthritic; rat whole cartilage was obtained following an in vivo surgical intervention in rats (e.g., joint destabilization model of OA) or from normal tissue post mortem. In both species, gene expression profiles of whole cartilage did not show strong statistical associations with network modules based upon the published phenotypic groups (e.g., healthy or osteoarthritic cartilage).

Whole-cartilage samples were assigned to new groups (“Clinical Group 1–3”, humans, or “Intervention Group 1–2”, rats) based upon co-clustering of gene expression profiles by multidimensional scaling. These new groups demonstrated associations with species-specific modules (Supplementary Figs. 4a and 5a).

Functional annotation of trait-associated modules

The unique ME expression profiles for trait groups are presented in Fig. 3a–d for rat and in Figs. 3e–g for human. Age-associated rat modules (Fig. 3a) were significantly enriched with functional annotations for RNA metabolic process, immune system process, and muscle system process. MEs associated with in vivo interventions (Fig. 3d) had annotations for immune system process, skeletal system development, sterol biosynthetic process, and cell adhesion. Functional annotations for in vitro studies and growth plate zone-associated modules (Fig. 3b, c) are found in Supplementary Table 1 and Supplementary Data SD3–SD21. Human chondrocyte condensation and differentiation profiles (Fig. 3e) were associated with modules annotated for multicellular organismal development, cell differentiation, and muscle system process. Whole-cartilage samples from clinical samples (Fig. 3g) divided into groups associated with modules annotated for positive regulation of the immune system and cell differentiation (“Clinical Group 2”) or with oxidative phosphorylation (normal articular cartilage). Further functional annotations for human modules are found in Supplementary Table 2 and Supplementary Data SD22–SD32.

Subset of human clinical samples associated with immune system and differentiation modules

Further consideration was given to trait associations between subsets of whole-cartilage samples showing the greatest separation from other cartilage samples (“Clinical Group 2”) and the H2 and H4 modules (immune system process and cell differentiation annotations, respectively). These modules were negatively correlated with normal articular cartilage samples (Fig. 1b and Supplementary Fig. 4a). The H2 and H4 modules demonstrated the greatest overlap with the C4 and C5 consensus modules (Supplementary Fig. 4b). Differential expression analysis of the eigengenes for these modules with clinical sample groups found a significant difference across all groups with the greatest difference in expression found between “Clinical Group 2” and other whole-cartilage subsets (Fig. 1c).

Rat module–trait relationships reveal distinct associations with age and growth plate zones

In vivo joint interventions in rats could not be readily discriminated into experimental groups (e.g. surgical sham or joint...
Cross-species module preservation statistical analysis

To establish how well the modules defined in the larger reference set (rat, 20 modules) were preserved and reproducible in the test network (human, 12 modules), module preservation statistics were calculated for each reference-test module pair using a series of permutation tests for measures of module density and connectivity. Thirteen rat modules were shown to have well-defined human counterparts (summary Z-score >5), while some appeared specific to the rat network (Supplementary Table 1 and Supplementary Data SD35). Modules associated with physiological processes (RNA metabolic process, cell cycle, immune system process, and skeletal system development) were strongly preserved between species (summary Z-score >10). The module of unassigned genes (R20) also demonstrated high preservation; however, this contained the majority of the genes (Supplementary Fig. 5b). A "toy" module of 100 randomly assigned genes collected from all possible genes (R21, "gold"; Supplementary Table 1) demonstrated no evidence of preservation across species.

Differential eigengene network analysis shows strong preservation of network structure across species

Differential eigengene network analysis (Fig. 4a-f) was used to define the overall preservation of the correlation of consensus ME pairs across the two species networks. To assess the overall preservation of modules and connectivity across the two data sets, eigengene networks were prepared based upon correlations between each pair of consensus MEs. This analysis sets out to establish whether consensus modules C4 and C5, associated with whole-cartilage subsets in both the rat and human, were conserved in the global network structure. There was strong evidence for eigengene network preservation between rat and human (density, D(Preservhuman/rat) = 0.85). Consensus MEs from the human data were defined by three main groups, or meta-modules (Fig. 4b). The first (M1) consisted of the C2 and C3 modules (blue and turquoise), the second (M2) of the C4 and C5 modules (yellow and brown), and the third (M3) contained the C1 module (green). This configuration was approximated in the rat data, particularly the preservation of the M2 meta-module (Fig. 4a). This demonstrated that in addition to the C4 and C5 modules being (i) present in both species, (ii) associated with gene expression profiles of subsets of whole cartilage, (iii) the organization of these functional units was also preserved across the networks and were highly correlated in their expression in cartilage from two species.

Meta-module M2 associated with cell differentiation and immune system related to distinct cartilage samples

Overall, the M2 meta-module represented 165 genes differentially expressed in subsets of cartilage samples from both rat and human, thus representing a biologically relevant superset of genes. Collectively, these genes were significantly enriched (after false discovery rate correction) for functional annotations including: "immune system process"—CTSC, CCR1, MMP9 (p = 3.1e−16), "response to wounding"—TLR7, GAP43 (p = 1.8e−8), "cell adhesion"—TNF, CAM1, COL18A1 (p = 1.5e−5), and "system development"—SOX17, SOX18, DMP1, IL18 (p = 3.7e−6). The M2 meta-module was significantly enriched for evidence of protein–protein interactions using STRING (p = 0). These modules were also enriched for distinct canonical signaling pathways: C4—"PI3K-Akt signaling pathway" (q = 2.9e−4) and "ECM interaction" (q = 1.3e−4); C5—"immune system" (q = 1.3e−4); and "osteoclast differentiation" (q = 3.4e−7) using over-representation analysis.

Systems development module C4 associated with osteophytic and hypertrophic samples across species

The structure and the nature of the most highly connected genes in the M2 modules were considered further. The C4 consensus...
module was associated with the H4 human module (Supplementary Fig. 4b), but are split across the R8 and R9 modules in the rat (Supplementary Fig. 5b). There was low to moderate correlation of module membership kME values between the rat and human modules (R8: \( r = 0.27 \), \( p = 4.7e^{-3} \); R9: \( r = 0.39 \), \( p = 3.2e^{-07} \)) indicating some preservation of hub modules. Consensus hub genes for the C4 module included CCL12, CTSK, DMP1, ACP5, MMP9, and COL1A1; however angiogenesis-associated genes EMCN (endomucin) and KDR (kinase insert domain receptor) were both found to be the most highly connected hubs in both species (Supplementary Fig. 6a, b). A subset of human OA cartilage samples (“Clinical Group 2”, \( r = 0.28 \), \( p = 1e^{-3} \)), osteophytic cartilage, and developing chondrocyte samples were associated with the H4 module; in the rat data the equivalent modules (R8 and R9) were associated with hypertrophic chondrocytes (R8, \( r = 0.37 \), \( p = 4e^{-05} \)) and a subset of OA intervention models, “Intervention Group 2” (R9, \( r = 0.37 \), \( p = 6e^{-05} \)). In summary, rat and human modules associated with the C4 module were also associated with subgroups of cartilage samples associated with degenerate whole cartilage, osteophytic cartilage, or hypertrophic chondrocytes.

Consensus module C5 reveals conserved network of immune system genes in cartilage

The C5 module described both the rat R5 and R11 and human H2 and H4 modules and was highly conserved. In the rat the R5 module was positively associated with cartilage samples derived from adult/early-aged rats and a subset of intervention studies modelling OA (“Intervention Group 2”). In the human, the H2 module was associated with a subset of clinical cartilage samples (“Clinical Group 2”, \( r = 0.44 \), \( p = 1e^{-07} \)). Across species, the consensus module hub genes were comparable (CD53, ALOX5AP, NCKAP1L, IGF5B6, CYBB, LCPI, LAPTMS); however, the connectivity pattern (those genes with the highest degree of the modules differed between the species (Fig. 5a, b). The C5 consensus module demonstrated significant enrichment of protein–protein interactions indicating that the shared module had a functional significance (Fig. 5c). Overall, genes with membership of the H2 module had more sparse connections than the comparable module in the rat. There was moderate correlation of module membership values (kME) for the genes in the R5 module with the human H2 module (\( r = 0.44 \), \( p = 1.7e{-12} \)) sufficient to indicate that module hubs identified in the rat were also likely to be hubs in the equivalent human module. Top hub genes are given in Supplementary Data SD35. Together with the description of the C4 module, these findings suggested that expression of genes in the M2 meta-module was consistent with a degenerate or dysregulated chondrocyte phenotype in both rat and human whole cartilage.
Cross-species gene modules in osteoarthritis
AJ Mueller et al.

Published in partnership with the Systems Biology Institute npj Systems Biology and Applications (2017) 13
Expression of Bcl6 across rat age groups demonstrated a reciprocal trend with cartilage from adult and aged rats showing higher expression relative to neonatal and young samples \( (p = 0.034) \). Other modules also showed evidence of low classification errors, but had not demonstrated strong associations with OA samples. With healthy cartilage arising from predominantly young individuals, the association between age and cartilage health confounded genes with classification potential.

DISCUSSION

This study demonstrates that added value may be gained from re-analysis of small transcriptomic studies using a network-based systems biology approach to establish conservation and divergences of transcriptional subnetworks between chondrocyte phenotypes in humans and a rodent model species. This study provides a comprehensive reference framework for chondrocytes in both physiological and pathological contexts. This has relevance to an understanding of the degenerative processes that underlie OA in humans and develops a clearer understanding of the utility of rodent models. Gene co-expression network analysis has become an increasingly useful tool to understand the regulatory context in which single genes may operate and has been applied to many aspects of human biology, but has not been used to determine the preservation of cross-species modules for OA.

A data-merging approach was employed in this study to agglomerate raw expression data from multiple sources. There is a strong rationale for the integration of microarray expression data including the increase in statistical power, re-use of rare specimens and animal data in silico, exploration of variance and noise in the data, and the definition of biological markers and prognostic signatures not evident in a small analysis.\(^{18}\) There is no definitive methodology for tackling microarray data merging, although integration at an interpretative level, using summary statistics, is a common approach. Data collection, curation, establishing common gene identifiers, and standardizing data-handling methods are essential.\(^{19}\) Integration is confounded by microarray platforms from different vendors being inherently incomparable.\(^{20}\) We chose to work with the probe-level, raw expression data from Affymetrix arrays using 25-mer probes to allow greater flexibility in our analysis; a Z-score global normalization strategy was used to avoid “over-smoothing” the expression data.

The validity of rodent models of OA at a systems level has not been established; this study sought to define shared mechanisms of cartilage degeneration between rats and humans. Two consensus modules (termed C4 and C5, comprising the M2 meta-module) were associated with perturbed cartilage phenotypes in human clinical samples and subsets of rodent OA models and sham surgical interventions. The C5 module was associated with immune system processes with a many highly connected module hub genes preserved between the species. Of these conserved hub genes CDS3, ALOX5AP, and NCKAP1L have previously been identified, using a co-expression network analysis, as key drivers in other rodent inflammatory conditions,\(^{21}\) suggesting that degenerative processes in cartilage are likely to be associated with inflammatory regulatory networks already defined in other disease processes. Although a pro-inflammatory molecular mechanism associated with OA progression is clear,\(^{22}\) there is no definitive evidence that DNA polymorphisms in inflammatory genes are a risk factor for OA.\(^{23}\) This study reveals that inflammatory gene networks are conserved across species and that modules contain genes widely described as having an association with OA.

We show that a differentiation and systems development module is preserved across species and associated with subsets of cartilage samples. Specifically, the C4 consensus module was associated with skeletal system development, cellular differentiation, ECM annotations, and PI3K-Akt signaling. The presence of genes with known angiogenesis (EMCN, KDR), chondrogenesis, OA, and cartilage knockout phenotypes—including DMP1,\(^{24}\) CTSK,\(^{25}\) MMP9, and ACP5\(^{26}\)—in a single consensus module demonstrates the utility of a network-based systems biology approach to an understanding of a multi-gene disorder across species.

OA is a multifactorial and complex disease with diagnosis by imaging modalities usually in the late stages of joint degeneration. Critically, this degeneration occurs over a considerable duration over which intervention could occur; a lack of disease-modifying therapeutics and poor characterization of pre-osteoarthritic disease states means that early intervention is not possible.\(^{27}\) Clinical information from public repositories was limited with age, sex, and only a general description of cartilage health available. Additionally, no information on co-morbidities (e.g., obesity) is provided in public repositories for these samples. Notably, a range of expression profiles from osteoarthritic and ostensibly normal cartilage was apparent, and these samples did not group according to definitions of cartilage health made by gross appearance.

The M2 meta-module had the greatest overlap with the H4 module in the human network. A class prediction approach was used to define a gene signature, using member genes of the H4 module, to discriminate osteoarthritic cartilage from healthy samples. Classification analysis identified a G-protein signaling gene ARHGDIIB (Rho GDP dissociation inhibitor beta, or RhoGDI2) as the highest scoring gene using an independent cohort; a variable signature of fewer than 25 genes could be used to consistently distinguish healthy from osteoarthritic cartilage. This indicated that a module of highly connected genes established as being shared across species was also discriminatory between healthy and OA cartilage samples in an independent data set. Notably, expression of ARHGDIIB protein has been found to be differentially abundant between healthy and late-stage OA synovial fluid.\(^{28}\) Subgroups of clinical samples showing differential associations with modules may represent different stages in cartilage degeneration, including pre-OA states,\(^{29}\) and should be investigated further as potential diagnostic and prognostic markers.

Fig. 3 Module eigengene heatmaps for associated traits. Module eigengenes (left-side vertical bars, rows; R—rat, H—human) derived from network modules with strong trait associations are presented as heatmaps of module eigengene expression. Supporting module—trait correlations and results of statistical tests are presented in Supplementary Figs. 4a (human) and 5a (rat). Samples associated with traits of interest are colored above the heatmap (top bar, columns) according to the associated key; samples with no strong module association are colored gray. Dendrograms indicate sample clusters. Module eigengene expression is indicated as high (red) or low (green), see color key. a–d Rat module eigengenes show distinct expression signatures for several trait groups (age, growth plate zones, in vitro chondrocyte studies, and in vivo joint interventions). Patterns of expression differ between cartilage samples derived from different age groups (a) and between different in vivo intervention groups (d); e–g equivalent heatmaps are presented for human trait groups (development and differentiation studies, in vitro studies including both two-dimensional and three-dimensional chondrocyte cultures, and whole cartilage clinical samples). The difference in expression profile between the clinical subgroups and normal articular cartilage may be visualized in (g). Gene ontology functional annotations for all modules are provided in Supplementary Data SD3–32.
Although a strong risk factor, ageing alone does not cause OA, but promotes OA in conjunction with a number of other risk factors, e.g., obesity, through multifaceted mechanisms that are incompletely understood. Age is a confounding variable in this analysis, with the majority of human studies profiling OA cartilage from aged patients. In the class prediction analysis from independent data ostensibly normal cartilage from aged donors is shown to cluster more readily with OA samples. We attempted to tease apart this association by using genes associated with aged cartilage for further class prediction analysis. The senescence and age-associated gene BCL6 was identified, but high expression was found in both young and healthy cartilage. Ageing chondrocytes do exhibit an associated decline in synthetic capacity (so-called "chondrocyte senescence") but retain the ability to produce pro-inflammatory cytokines. Functionally, this correlates with the M2 meta-module where consensus hubs...
Cross-species gene modules in osteoarthritis

AJ Mueller et al.

Network Legend

- **node**
  - degree

- **edge**
  - weight

PPIN legend

**Known interactions**
- From curated databases
- Experimentally determined

**Predicted interactions**
- Gene neighborhood
- Gene fusions
- Gene co-occurrence

**Other**
- Text mining
- Co-expression
- Protein homology

**Biological process:**
immune response
(q=2.8e-14)

**Pathway description:**
Osteoclast differentiation
(q=5.5e-09)

**PPI enrichment:**
p=0
are associated with immunomodulation or cell differentiation. Therefore, in human samples it is not possible to distinguish module associations with the pathophysiological process of OA from ageing per se in part because of the bias in healthy tissue arising only from young individuals.

A module containing the hallmarks of functional cartilage (R12, Col2a1, Acan, Comp, Hapln) was found in the rat co-expression network associated with sham controls (isotonic saline injection into normal joints). These genes were not assigned to a distinct module in the human, and COL2A1 was not differentially expressed across samples. If the assumption is made that the R12 module represents a healthy, functional cartilage profile, the absence of an equivalent module in human data may indicate that reported cartilage control tissue in most human transcriptomic surveys is not functionally normal or expression of COL2A1 is not the best indicator of human cartilage degeneration. The lack of age-matched healthy and OA human cartilage gene expression profiles is a notable obstacle to further interpretation of the results.

Given the complexity of this type of analysis, the key points of this study may be reduced to the following: (i) the use of ARHGDB as a potential OA marker or therapeutic target should be explored further, given the existing OA association, with BCL6 expression validated as a marker of chondrocyte senescence or age-associated cartilage degeneration; (ii) the accessibility of normal human cartilage needs to be evaluated as there is a dearth of healthy samples and clinical data in public repositories; (iii) sham surgery and surgical destabilization of the joint in rat models of OA may be poor comparators, given the co-clustering of samples in this study; (iv) community-based approaches in OA research are essential to developing appropriate standardized in vivo models in particular full experimental disclosure is lacking in many of the rat studies; (v) hub genes are the fragile points in a network; a number of these are indicated for conserved modules with OA associations. These should be considered as novel knockout targets in the mouse as part of age-matched longitudinal studies; (vi) further validation of co-expression networks with phenotypic and quantitative traits should be undertaken to elucidate causal mechanisms.

To conclude, two highly correlated consensus modules are conserved across species when cartilage gene expression profiles are considered. Inflammation and differentiation status of the resident chondrocytes are shown to be strongly associated with a dysregulated cartilage phenotype in both humans and rats. While evidence for an association with a number of established OA genes is present, demonstration that these OA-associated genes are co-expressed has not previously been shown. We found that some elements of human OA are conserved in rodent models, but suitably matched prospective studies of sufficient power across species are required to maximize translational impact and utility in the discovery of disease-modifying therapeutics to target multiple disease-associated networks.

METHODS

An overview of the general approach used for data collection and analysis is provided in Supplementary Figs. 1 and 2. Gene expression profiles were selected from curated public-access repositories Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo)34 and ArrayExpress (http://www.ebi.ac.uk/arrayexpress/).35 To be included in initial analysis studies had to: (a) be performed in the rat (Rattus norvegicus) or human, (b) profile chondrocytes from tissue or in vitro culture systems, (c) provide adequate phenotypic information, (d) provide complete raw data for a minimum of three biological replicates, (e) be performed on Affymetrix microarray platforms (Affymetrix Inc., USA) using 25-mer oligo probe sets. All studies released up to December 2015 were considered. All raw data were imported into and analyzed using R.36 A quality control and pre-processing pipeline was applied to each autonomous study, and these assessed for systematic technical issues. Expression data were background-corrected using the RMA algorithm,57 with cyclic loess normalization method applied across each data set. Probe sets were re-annotated with the appropriate Ensembl gene identifier. Expression data for each gene were aggregated and collapsed into a single-gene measurement consisting of the maximum mean expression value using the “collapseRows” function in the WGCNA.38 The output of this workflow was a normalized matrix of expression values consisting of one summarized gene per row. Intersection of data sets by common gene identifiers was performed such that all data sets contained the same gene identifiers. The matrix of merged data sets was termed a “meta-set”. The rat meta-set consisted of 115 arrays (10,159 common annotations) and the human meta-set consisted of 129 arrays (11,392 common annotations). A z-score normalization was applied to each species meta-set using the inSilicoMerging R package.39 A complete description of data sources and retained samples is provided in Supplementary Data SD43 (human) and SD44 (rat).

Weighted gene co-expression network analysis

To establish universal gene identifiers and facilitate comparison across species, rat gene identifiers from Affymetrix probes were re-annotated with human Ensembl gene orthologs using biomart, an R interface with the Biomart database (http://www.biomart.org).40 Only identifiers that were common to both meta-sets were retained and genes with a global variance less than 0.3 were removed to reduce noise and computational demands. The data met the assumptions of a scale-free network. The general approach employed to develop co-expression modules is described by Miller et al.41 and is outlined in graphical workflows in Supplementary Figs. 1 and 2. A consensus network represents a single network arising from multiple sources of data constructed from the weighted average of correlation matrices from both the human and rat in this study. By definition, consensus modules are the branches of a clustering tree developed from a consensus gene dissimilarity, comparable to the single-network approach; consensus modules contain genes that are closely related in both networks, i.e., the modules are present in both networks.41 Consensus network and module generation were performed in WGCNA (version 1.49) with the following changes to the default settings for consensus network generation: β = 0.7, deepSplit = 1, cutHeight = 0.25, and a minimum module size of 30 genes. This was consistent with the parameters used for single-species network generation. Only consensus module colors are equivalent across the species. Modules were characterized based upon a representative gene termed the “ME defined as the first principal component of the module gene expression profile. MEs were used for module–trait association analysis, differential eigengene network analysis, and for differential gene expression analysis. Difference
Cross-species gene modules in osteoarthritis

AJ Mueller et al.

Classifier
Test 1: AUC = 0.86
Test 2: AUC = 1
Test 3: AUC = 0.992
Test 4: AUC = 0.875
Test 5: AUC = 0.896
Test 6: AUC = 0.88
Test 7: AUC = 0.765
Test 8: AUC = 1
Test 9: AUC = 0.872
Test 10: AUC = 0.875

PC1 (59.5%) vs. PC2 (40.5%)

age: young, mid, aged
class: healthy, OA

Gene Expression

Aged Condition

Ankrd

Ddx7
in expression across trait groups was tested using a Kruskall-Wallis one-way analysis of variance. A gene's module membership \( k_{\text{ME}} \) is defined as the Pearson correlation between each gene and each ME genes with high \( k_{\text{ME}} \) values were considered "hub" genes and were highly co-expressed within a subnetwork. How well these hubs were preserved across species was determined by correlating gene expression profiles from a subnetwork. To determine the significance of this preservation, we employed to determine the signifi-
can connectivity preservation relative to a module of randomly assigned genes where values \( S > Z < 10 \) represent moderate preservation, while \( Z > 10 \) indicated high preservation. The composite statistic summarized density-based and connectivity-based preservation statistics (Eq. 1):

\[
Z_{\text{summary}} = \frac{Z_{\text{density}} + Z_{\text{connectivity}}}{2}
\]

Density-based measures assessed whether module nodes remained densely connected in a test network; connectivity-based measures defined whether intranode connectivity patterns in the reference network were similar to those in the test network. A separate summary \( p \) value for module preservation, given as the median of the log- \( p \) values for the associated permutation \( Z \) statistics, was calculated. Permutation tests, where the module labels of the test network were randomly permuted, were employed to determine the significance of the observed preserva-
tions test statistics. A module of randomly assigned genes, "gold" (R21) module, was prepared as a sham module to evaluate bias in the module preservation across species. The reader is referred to other sources for glossaries of terms associated with WGCNA. 16, 42

The gene expression profile for a consensus module of highly co-
expressed genes could be summarized by a single representative gene, the eigengene (described as the first right-singular vector of the standardized expression profile for each module), i.e., a module could be characterized by a single representative gene. 41 Eigengene networks for single (species-
pecific) co-expression networks were prepared using the correlations exhibited by pairs of eigengenes from different modules where the connection strength (adjacency) between eigengenes \( E_i \) and \( j \) (Eq. 2):

\[
\alpha_{E_{ij}} = \frac{1 + \text{cor}(E_i, E_j)}{2}
\]

The study considered the correlation preservation between all pairs of consensus MEs across the two species networks, \( A_{\text{Eigenv}(\text{human})} \) and \( A_{\text{Eigenv}(\text{rat})} \), where \( A_{\text{Eigenv}(\text{s})} \) is the adjacency matrices for data set \( s \) defined in Eq. 2. A preservation network \( \text{Preserv}_{\text{human}, \text{rat}} = \text{Preserv}(A_{\text{Eigenv}(\text{human})}, A_{\text{Eigenv}(\text{rat})}) \) was prepared in which adjacencies are defined as:

\[
\text{Preserv}_{\text{human}, \text{rat}} = 1 - \frac{\text{cor}(E_{ij}, E_{ij}^\text{human}) - \text{cor}(E_{ij}, E_{ij}^\text{rat})}{2}
\]

where \( E_{ij} \) is the eigengene of the \( i \)-th module in the data set \( s \). High values of \( \text{Preserv}_{\text{human}, \text{rat}} \) indicated robust preservation, across the two networks, of the correlation between module eigengenes \( I \) and \( J \). The scaled connectivity \( C_o \) or degree, for the \( i \)-th module (Eq. 4) is described as the mean connection strength with all other eigengenes; the scaled connectivity of the preservation network is given by:

\[
C_{\text{Preserv}_{\text{human}, \text{rat}}} = 1 - \frac{\sum_{s=1}^{S} \text{cor}(E_{ij}^s, E_{ij}^\text{human}) - \text{cor}(E_{ij}^s, E_{ij}^\text{rat})}{2(N - 1)}
\]

(4)

(where \( N \) denotes the number of MEs; this value is found to be close to 1 if there is preservation of the correlation between the \( i \)-th eigengene and all other eigengenes across the two networks. The density of the eigengene network \( D_{\text{Preserv}_{\text{human}, \text{rat}}} \) (Eq. 5), defined as the average scaled connectivity, is given by:

\[
D_{\text{Preserv}_{\text{human}, \text{rat}}} = 1 - \frac{\sum_{s=1}^{S} \sum_{j=1}^{J} \text{cor}(E_{ij}^s, E_{ij}^\text{human}) - \text{cor}(E_{ij}^s, E_{ij}^\text{rat})}{2N(N - 1)}
\]

(5)

Values of \( D_{\text{Preserv}_{\text{human}, \text{rat}}} \) that are large, approaching 1, indicate strong preservation of correlation between all the eigengene pairs across the two networks (human and rat). Procedures to detect modules in networks could be applied to eigengene networks to find modules of highly positively correlated eigengenes, term "meta-modules". 41

Module--trait relationships

To determine whether modules were associated with chondrocyte phenotypes or traits the MEs were correlated with a binary matrix coding, the membership of an individual sample to a phenotypic trait or experimental group (1 = member, 0 = non-member). Multidimensional scaling plots of each meta-set was used to define clusters of samples rather than using the phenotypic data from the published data set to define sample groups.

Network visualization and annotation

The module network structure, consisting of nodes (genes filtered for high module membership, \( k_{\text{ME}} \)) and edges (weighted intramodular connections based on the topological overlap matrix) were represented graphically using Cytoscape (v3.3.0, January 2016). 43 Only nodes with high degree were retained for clarity. Enrichment of protein–protein interaction networks was assessed using STRING v10 (http://string-db.org/). Pathway enrichment analysis was undertaken for each consensus module using the ConsensusPathwayDB platform (release 31 September 2015) (http://cpdb.molgen.mpg.de). Modules were functionally annotated using DAVID (https://david.ncifcrf.gov/). 44

Class prediction analysis

Class prediction analysis was performed using the pamr package implemented in R 48 (see Supplementary Fig. 3). This method employs a "nearest shrunken centroids" approach to determine cohorts of genes that best characterize classes from high dimensional data. The average gene expression for every gene in a class is divided by the within-class SD for the gene; this is the standardized centroid for each class. The gene expression profile of a new (test) sample is compared to the class centroid in nearest centroid classification; the predicted class for the new sample is the nearest class centroid by squared distance. The nearest shrunken centroid modification "shrinks", by a threshold value, all class centroids toward an overall centroid; the threshold is defined by a 10-fold cross-validation for a range of threshold values. Genes from modules with important trait
associations were used as the selected features for class prediction where the two classes were “healthy” or “osteoarthritic” cartilage. Classification training was performed on gene expression data (Illumina) from an independent data set\(^59\) profiling healthy \((n = 7)\) and osteoarthritic \((n = 33)\) cartilage. This was repeated for each of ten randomized test and training sets. Receiver operator characteristic (ROC) curves and area under the curve analysis was undertaken using the ROC package in R for each gene signature.\(^50\)

Code availability

Code contributing to the analysis presented here is available in Supplementary Methods 1 with supporting processed and annotated data files for rat and human.

ACKNOWLEDGEMENTS

We acknowledge the helpful discussions with Professor Francesco Falconi (University of Liverpool) and Dr. Marta Milo (University of Sheffield) in devising this study. Critical reviews by Dr. Blandine Poulet and Dr. Mandy Peppers (both University of Liverpool) guided the development of the final manuscript. This study was funded by a Biotechnology and Biological Sciences Research Council (BBSRC) Doctoral Training Grant (BB/H531951/1).

AUTHOR CONTRIBUTION

A.J.M., E.G.C.-L., S.R.T. and P.D.C. devised the study. A.J.M. executed and interpreted the study and wrote the manuscript. All authors contributed to revisions of the manuscript and approved the final version.

COMPETING INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Cross, M. et al. The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. Ann. Rheum. Dis. 73, 1323–1330 (2014).
2. van Dijk, G. et al. Comorbidity, limitations in activities and pain in patients with osteoarthritis of the hip or knee. BMC Musculoskelet. Disord. 9, 95 (2008).
3. Murray, C. et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the global burden of disease study 2010. Lancet 380, 2197–2223 (2012).
4. Goldberg, M. Chondrogenesis, chondrocyte differentiation, and articular cartilage metabolism in health and osteoarthritis. Ther. Adv. Musculoskelet. Dis. 4, 269–285 (2012).
5. Eyre, D. Articular cartilage and changes in arthritis: collagen of articular cartilage. Arthritis Res. 4, 30–35 (2002).
6. Hunter, D. J. Pharmacologic therapy for osteoarthritis: the era of disease modification. Nat. Rev. Rheumatol. 7, 13–22 (2011).
7. Greene, M. A. & Loeser, R. F. Aging-related inflammation and intracellular metabolism: new targets in OA. Arthritis Res. Ther. 17, 971–979 (2009).
8. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets-10 years on. Nucleic Acids Res. 39, D1005–D1010 (2011).
9. Rustici, G. et al. ArrayExpress update-trends in database growth and links to data analysis tools. Nucleic Acids Res. 41, D987–D990 (2013).
10. R Core Team. R: A Language and Environment for Statistical Computing v3.2.2 (2011).
11. Chu, C. R., Williams, A. A., Coyle, C. H. & Bowers, M. E. Early diagnosis to enable early treatment of pre-osteoarthritis. Arthritis Res. Ther. 14, 212–212 (2012).
12. Hunter, D. J. Pharmacologic therapy for osteoarthritis: the era of disease modification. Nat. Rev. Rheumatol. 7, 13–22 (2011).
13. Greene, M. A. & Loeser, R. F. Aging-related inflammation and intracellular metabolism: new targets in OA. Arthritis Res. Ther. 17, 971–979 (2009).
14. Barret, T. et al. NCBI GEO: archive for functional genomics data sets-10 years on. Nucleic Acids Res. 39, D1005–D1010 (2011).
15. Rustici, G. et al. ArrayExpress update-trends in database growth and links to data analysis tools. Nucleic Acids Res. 41, D987–D990 (2013).
16. R Core Team. R: A Language and Environment for Statistical Computing v3.2.2 (2011).
17. Chintu, M. et al. Integration of biological networks and gene expression data using inSilicoDb and inSilicoMerging R/Bioconductor packages. BMC Bioinformatics 13, 257–264 (2012).
18. Loeser, R. F. Aging and osteoarthritis. Curr. Opin. Rheumatol. 23, 492–496 (2011).
19. Hugel, T., Geurts, J., Nesucz, C., Mueller-Gerbli, M. & Valderrabano, V. Aging and osteoarthritis: an inevitable encounter? J. Aging Res. 2012, 950192 (2012). Article ID.
20. Shvarts, A. et al. A senescence rescue screen identifies BCL6 as an inhibitor of anti-proliferative p19(ARF)–p53 signaling. Genes Dev. 16, 681–686 (2002).
21. Martin, J. & Buckwalter, J. Aging, articular cartilage chondrocyte senescence and osteoarthritis. Biogerontology 3, 257–264 (2002).
22. Loeser, R. F. Aging and osteoarthritis: the role of chondrocyte senescence and aging changes in the cartilage matrix. Osteoarthritis Cartilage 17, 971–979 (2009).
23. Aspden, R. M. Obesity punches above its weight in osteoarthritis. Nat. Rev. Rheumatol. 7, 65–68 (2011).
24. Panouktsopoulou, K. & Zeggini, E. Advances in osteoarthritis genetics. J. Med. Genet. 50, 715–724 (2013).
25. Reardon, L. & Loughlin, J. The genetics and functional analysis of primary osteoarthritis susceptibility. Expert Rev. Mol. Med. 15, e2 (2013).
26. Knox, S. From ‘omics’ to complex disease: a systems biology approach to gene–environment interactions in cancer. Cancer Cell. Int. 10, 11 (2010).
27. Poole, R. et al. Recommendations for the use of preclinical models in the study and treatment of osteoarthritis. Osteoarthritis Cartilage 18, 510–516 (2010). Supplement 3.
28. Felson, D. T. Osteoarthritis: priorities for osteoarthritis research: much to be done. Nat. Rev. Rheumatol. 10, 447–448 (2014).
29. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).
30. Zhao, W. et al. Weighted gene coexpression network analysis: state of the art. J. Biopharm. Stat. 20, 281–300 (2010).
31. Miller, J., Horvath, S. & Geschwind, D. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc. Natl Acad. Sci. USA 107, 12698–12703 (2010).
32. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. Nature 402, C47–C52 (1999).
33. Sarmah, C. K. & Samarasainge, S. Microarray data integration: frameworks and a list of underlying issues. Curr. Bioinformatics 5, 280–289 (2010).
34. Ramasamy, A., Mondry, A., Holmes, C. & Altman, D. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med. 5, e184 (2008).
35. R Core Team. R: A Language and Environment for Statistical Computing v3.2.2 (2011).
36. Cline, M. et al. Integration of biological networks and gene expression data using inSilicoDb and inSilicoMerging R/Bioconductor packages. BMC Bioinformatics 13, 257–264 (2012).
37. Rustici, G. et al. ArrayExpress update-trends in database growth and links to data analysis tools. Nucleic Acids Res. 41, D987–D990 (2013).
38. R Core Team. R: A Language and Environment for Statistical Computing v3.2.2, Fire Safety (R Foundation for Statistical Computing, 2016).
39. Irizarry, R. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249–264 (2003).
40. Chintu, M. et al. Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. 2, 2366–2382 (2007).
41. van Mering, C. et al. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 31, 258–261 (2003).
42. Szklarczyk, D. et al. STRING v10 protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43, D447–D452 (2015).
43. Kambarov, A., Wierling, C., Lehbrach, H. & Hervig, R. ConsensusPathDB—a database for integrating human functional interaction networks. Nucleic Acids Res. 37, D623–D628 (2009).
44. Huang, D. W., Sherman, B. & Lempicki, R. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57 (2009).
45. Shibutani, R., Hatse, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA 99, 6567–6572 (2002).
46. Ramos, Y. et al. Genes involved in the osteoarthritis process identified through genome wide expression analysis in articular cartilage; the RAAK study. PLoS ONE 9, (2014).
50. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).

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

© The Author(s) 2017

Supplementary Information accompanies the paper on the *npj Systems Biology and Applications* website (doi:10.1038/s41540-017-0014-3).