Comprehensive profiling of transcriptional regulation in cartilage reveals pathogenesis of osteoarthritis

Wen Tian[1#], Shan-Shan Dong[1#], Feng Jiang[1#], Chang-Yi He[1], Chen Wang[1], Shou-Ye Hu[2], Ruo-Han Hao[1], Hui-Miao Song[1], Hui-Wu Gao[1], Jun-Qi Zhang[1], Ke An[1], Dong-Li Zhu[1], Zhi Yang[2], Yan Guo[1*], Tie-Lin Yang[1*]

[1]Key Laboratory of Biomedical Information Engineering of Ministry of Education, Biomedical Informatics & Genomics Center, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, P.R. China

[2]Honghui Hospital, Xi'an Jiaotong University, Xi'an, Shaanxi 710054, P.R. China

[#]These authors contributed equally

[*]Correspondence authors:

Tie-Lin Yang, Ph.D., Yan Guo, Ph.D.

E-MAIL: yangtielin@xjtu.edu.cn (T.-L.Y.), guoyan253@xjtu.edu.cn (Y.G.)

Conflict of Interest: The authors declare that they have no competing interests.

Abstract

Cartilage damage is a leading cause of osteoarthritis (OA) etiology, however, the underlying mechanism governing gene expression regulation in this progress is poorly understood. Here, we described a comprehensive profiling of transcriptional regulation of 234 primary human cartilage samples. Based on speculated chondrocyte subtype proportions, we demonstrated the heterogeneity of cell composition contributed to the variance of gene expression across samples. We identified 3,232 significant expression quantitative trait loci (eQTLs) for 2,999 genes. Cell type-interaction eQTL (ci-eQTL) analyses revealed the regulation relationship of 81 eQTL-gene pairs showing cell type dependency. Further, by integrating with genome-wide association studies (GWASs) data of OA, we nominated 55 candidate effector genes, of which, 26 were novel genes that have not been reported by previous studies. Finally, we established a multi-level functional variants-deciphering pipeline by integrating the cartilage-specific ATAC-seq data, and identified 112 causal eQTLs that display allele-specific open chromatin (ASoC), as well as 410 transcription factor binding-disruption (TBD) eQTLs. We experimentally verified that the T allele at rs11750646, which was located within a GWAS-eQTL co-localized region, increased the AR binding affinity to an open chromatin region and promoted the expression of an OA-related gene *PIK3R1*. Altogether, our findings provide new insights into the unique regulatory landscape of cartilage and elucidate potential mechanisms underlying the OA pathogenesis.

Key word: Osteoarthritis; Cartilage; eQTL; Fine-mapping

Introduction

Osteoarthritis (OA) is the most prevalent chronic joint disease, which causes pain, disability and loss of function, and affects over 500 million people worldwide[1]. Knee is the leading site of OA, followed by hand and hip[1,2]. In the process of OA, articular cartilage is both the primary victim and betrayer[2]. Figuring out the role of articular cartilage in OA and its specific mechanisms, especially underlying gene expression regulation, is critical to developing efficient therapies.

Genome-wide association studies (GWASs) have reported more than 100 susceptibility loci for OA[3-9], however, the majority of SNPs map to noncoding regions and the downstream effector genes in articular cartilage have not been well clarified. Expression quantitative trait loci (eQTLs) data provide a valuable tool for screening effector genes for noncoding SNPs. The first eQTL regulation landscape of OA articular cartilage has been constructed from 99 European patients[10], and identified a total of 1,453 genes with eQTLs, which allows the nomination of four effector genes of OA GWAS loci. However, the power to detect mechanistically informative expression signals rely on sufficient numbers of samples in relevant tissues[11], a larger sample size of articular cartilage is needed to obtain more statistically significant eQTLs. Furthermore, given the discrepancy of minor allele frequencies (MAF) in different populations, a cartilage eQTL resource from non-European population will provide additional information to elucidate the gene expression roles of GWAS variants.

Previous studies have shown that some of the eQTL effects are cell type-specific[12-16]. Chondrocytes have several subtypes[17-19], but their genetic regulatory heterogeneity remains unknown. As an effort to solve this issue, an approach based on the in silico speculated cell subtype proportions from bulk RNA-seq data has been developed to reveal the cell type dependency of eQTL effects[20] and has been broadly used for bulk RNA-seq data from several tissues[21-24].

In this study, we performed comprehensive characterization for transcriptional regulation of knee cartilage samples from 204 OA patients (Fig. 1). Based on the deconvoluted chondrocyte subtypes proportions from bulk RNA-seq data, we evaluated the contribution of cell composition to the overall gene expression variance across samples and its contribution to the differential expression between damaged and intact cartilage samples. We not only identified *cis*-eQTLs in this key tissue, but also evaluated the dependence of the *cis*-eQTLs effect on the chondrocyte subtypes. Next, we nominated the effector genes for OA GWAS loci and annotated their potential function using multiple lines of evidence. Finally, leveraging the cartilage-specific epigenomic data and our multi-level functional variants-deciphering pipeline, we conducted functional fine-mapping to reveal candidate causal variants that linked both to OA risk and gene expression regulation through transcription factors (TFs) binding site disruption or chromatin accessibility alteration. In all, our study represented a valuable eQTL resource in cartilage, and provide new insights

into how genetics contributes to the pathogenesis of OA.

Results

Contribution of chondrocyte composition to the expression variance of cartilage samples

We generated 235 qualified knee cartilage RNA sequencing (RNA-seq) data from 204 Chinese Han individuals undergoing joint replacement for OA with extensive quality controls (Fig. 1 and Supplementary Tables 1,2). We obtained a median of 14,809 expressed protein coding genes, 3,407 long non-coding RNA (lncRNA) genes and 1,915 genes of other types (expressed gene was defined as TPM ≥ 0.1, Supplementary Fig. 1a). The most of protein coding genes showed high expressions (TPM ≥ 10), while the majority of non-coding genes had low expressions (TPM < 1) (Supplementary Fig. 1b).

Using a liner mixed model-based method[25], we compared the expression between the paired damaged and intact cartilage samples from 31 donors. We observed 334 genes significantly high expressed in damaged cartilage, and 715 genes significantly high expressed in intact cartilage (FDR < 0.05, no fold change cutoff; Fig. 2a and Supplementary Table 4). Gene set enrichment analyses showed that the genes expressed lower in damaged cartilage were significantly enriched in the cell markers of chondrocyte subtypes, effector chondrocytes (EC) and regulatory chondrocytes (RegC) (Fig. 2b,c). Cell type composition differences have been proven to be a vital

factor for driving gene expression variance in bulk RNA-seq data[26]. Cartilage consists of multiple subtypes of chondrocyte[17-19]. To evaluate the contribution of chondrocyte composition to the overall variance of gene expression, we deconvoluted our bulk gene expression data (204 from damaged cartilage and 31 from intact cartilage) into chondrocyte subtype proportions based on the expression signature of each subtype extracted from scRNA-seq data[19] (Fig. 2d,e and Supplementary Table 5). By comparing the speculated cell proportions in the paired damaged and intact cartilage samples, our results showed that the damaged cartilage samples had significantly lower proportions in EC and RegC, and significantly higher proportion in prehypertrophic chondrocytes (preHTC) (Fig. 2f), which were consistent with the findings from scRNA-seq data[19] and enrichment analyses.

Next, for each gene, we estimated the proportion of expression variance that could be explained by each chondrocyte subtype. After adjusting multiple common biological factors and artificial factors, the variance could be largely explained by proportion of chondrocyte subtypes (Fig. 2g). In combination, the median of variance explained by all six chondrocyte subtypes was 21.3% (Fig. 2h), indicating the great impacts of chondrocytes composition heterogeneity on the expression variance.

Cis-eQTL mapping

Regular cis-eQTL

To identify *cis*-eQTLs in cartilage, we generated genome-wide genotype data from

peripheral blood samples of all participants (Supplementary Tables 1,3). The genotype data showed no significant stratification (Supplementary Fig. 2a), and the minor allele frequencies (MAF) were strongly correlated with that of Asians but weaker correlated with that of Europeans in the 1000 Genomes Project (Supplementary Fig. 2b.c). We then performed *cis*-eQTL mapping by integrating the chondrocyte subtype proportions as covariates to correct for the impact of cell type heterogeneity. Totally, we identified 2,999 genes (eGenes) with at least one genome variant (eVariant) significantly associated with its expression at 5% false discovery rate (FDR) level (Fig. 3a and Supplementary Table 6), including 2,449 (81.7%) protein coding genes, 385 (12.8%) lncRNAs and 165 (5.5%) of other types (Fig. 3b and Supplementary Table 7). Compared with non-eGenes, eGenes showed significantly higher expression ($P < 2.2\times10^{-16}$; Wilcoxon rank sum test; Fig. 3c) and lower expression variation coefficient ($P < 2.2\times10^{-16}$; Wilcoxon rank sum test; Fig. 3c).

Conditional analysis

To identify independent eQTL signals, we conducted conditional analysis and found that 216 of 2,999 eGenes had more than one independent eQTL signals (Fig. 3d and Supplementary Table 8). Importantly, 98 eGenes (45.4%) had at least one newly identified independent signal which could not be detected by regular eQTL analysis (Fig. 3e). For example, rs623263 showed significant association with the *SNAP91* expression only when conditional on the effect of rs145051363, which tagged the other independent signal (Fig. 3f). We compared the characteristics of the newly

identified independent lead eVariants (the most significant eVariant for each eGene) by conditional analysis with those significant lead eVariants by regular eQTL analysis. In general, the newly identified lead eVariants showed significantly lower MAF ($P = 1.7 \times 10^{-4}$; Wilcoxon rank sum test; Supplementary Fig. 3a) and longer distance to transcription start sites (TSS) of the target genes ($P = 2.6 \times 10^{-2}$; Wilcoxon rank sum test; Supplementary Fig. 3b), but no difference in effect size for the target genes ($P = 0.404$; Wilcoxon rank sum test; Supplementary Fig. 3c). Down sampling analysis showed that conditional analysis with our larger sample size could help uncover more independent association signals (Supplementary Fig. 4).

For the 3,232 unique independent lead eVariants, the largest part (1,606, 49.7%) was located in the intron regions, followed by the intergenic regions (747, 23.1%), promoter regions (607, 18.8%), and exon regions (272, 8.4%) (Fig. 3g). Lead eVariants were almost symmetrically concentrated on both sides of the TSS of their target genes (Fig. 3h). In total, 1,978 (61.2%) lead eVariants were located within ± 50 Kb from the TSS. However, there were also 754 lead eVariants (23.3%) located > 100 Kb from the TSS of the target genes, suggesting the long-range regulation models between eVariants and their target genes.

To assess eQTL reproducibility, we compared our results with a previously published cartilage study involving approximately 90 European individuals[10]. Despite the high direction consistency of regulatory effect of the shared variant-gene pairs, our data

also uncovered a large number of pairs that were specifically significant in our study (Supplementary Fig. 5). This was partly attributable to the larger sample size of our study (Supplementary Fig. 6). Besides, the difference in MAF between the Chinese and European populations was also a considerable factor. Among the 236,922 specific eVariants identified in our study, 13,568 (5.76%) and 4,300 (1.83%) with a MAF < 0.01 or < 0.001 in the 1000 Genomes Project European population, respectively (Supplementary Fig. 7). This meant that if the minimum minor allele count was set to 10, as used in this study, it would take about 1,000 and 10,000 samples to calculate their effects on the expression of *cis*-genes, respectively, which exceed the maximum sample size of any single tissue in the GTEx project (Muscle-Skeletal, 803 samples, v8)[27], posing a great challenge in terms of sample collection and highlighting the values of eQTL mapping in different populations.

Characteristics of eVariants

We assessed the functional characteristics of the identified eVariants using epigenomic data from human knee cartilage[28] and human mesenchymal stem cells (hMSCs) derived chondrocytes[29]. As expected, eVariants were significantly enriched in open chromatin regions with ATAC-seq peaks (Fig. 3i). Significant enrichments were also observed in multiple active histone markers (H3K4me3, H3K27ac, H3K4me1, and H3K36me3) (Fig. 3i). Specifically, we found that the eVariants exhibited stronger enrichment in super-enhancers (SEs) than that in typical enhancers (TEs) (Fig. 3j), suggesting a deeper implication of SEs in genetic regulation of gene

expression.

Genetically regulatory heterogeneity among chondrocyte subtypes

Genetic regulation of gene expression exhibited a broad cellular specificity[12-16]. To explore whether there was regulatory heterogeneity among chondrocytes subtypes, we performed cell type-interaction eQTL analyses (ci-eQTL), from which the possible driver or confirmable non-driver cell types for a single variant-gene regulatory pair would be revealed (Fig. 4a; Methods). After applying to three widely distributed chondrocyte subtypes (EC, preHTC, and RegC) in our deconvolution results (Fig. 2d,e), we identified 29, 23 and 29 genes with significant ci-eQTLs for EC, preHTC and RegC, respectively (Supplementary Tables 9-11). Of which, most of variant-gene regulatory pairs were aligned to the driver model (Fig. 4b).

We estimated the sharing and specificity of all 81 significant cell type-interaction variant-gene pairs using multivariate adaptive shrinkage (mash)[30] analysis. Totally, we obtained 69 variant-gene pairs with local false sign rate (LFSR) < 0.05 in at least one cell type (Supplementary Table 12). We clustered the 69 pairs into 11 clusters according to their model type and mash posterior effect size (Fig. 4c), and observed similar pattern for EC and RegC, opposite pattern for preHTC, which were consistent with the correlation between their proportions that speculated from our samples (Supplementary Fig. 8). We found 7, 6 and 11 cell type-specific pairs that aligned to driver model for preHTC, RegC and EC, respectively. As examples, the expression of

the *IGFBP2* showed a significantly positive correlation with the count of T allele at rs12620516 (Fig. 4d). Consistently, with the proportion of preHTC increased, T allele carriers exhibited significantly higher *IGFBP2* expression, but no such pattern was observed in the EC and RegC (Fig. 4e). The *IGFBP2* was specifically expressed in preHTC in the scRNA-seq data of knee cartilage[19] (Supplementary Fig. 9), which could be the cause for cell type-specific eQTL effects[23]. Similarly, RegC and EC were the diver cell types of the regulatory pairs rs2356553~*MAST2* (Fig. 4f,g) and rs1408829~*TOR3A* (Fig. 4h,i), respectively. These findings demonstrate the genetically regulatory heterogeneity of gene expression among chondrocytes subtypes.

eQTL-GWAS integration nominates risk genes for OA

To date, GWASs have successfully discovered numerous genetic loci associated with OA, yet their underlying gene expression regulatory mechanisms remain elusive. Here, we performed co-localization analysis to integrate our cartilage eQTL data (regular, conditional and ci-eQTL results) and 28 publicly available GWAS summary statistics of OA from six studies (Supplementary Table 13)[4-9]. In total, we identified 55 genes for which the corresponding eQTL signals showed strong evidence (PPH4 $\geq$ 0.7) that shared a common causal variant with GWAS signals (Fig. 5a, Supplementary Fig. 10 and Supplementary Table 14). Specifically, our eQTL data supplemented 26 new genes to the candidate OA risk gene list (gene symbols were shown in red in Fig. 5b), which could not be nominated from eQTL data of any tissues in the GTEx project and

previous cartilage eQTL data (Fig. 5b and Supplementary Figs. 11-12), revealing the sufficient power of our relatively larger sample size in cartilage.

Our eQTL data linked OA GWAS signals to 14 functionally known OA related genes (Fig. 5b and Supplementary Table 15), including five genes showed differential expression between paired damaged and intact cartilage in our data or abnormal cartilage/chondrocytes related phenotypes in gene knockout mice (Fig. 5b and Supplementary Table 16). For example, *COLGALT2* showed a significant lower expression in our damaged cartilage samples ($\log_2$FC = -1.61, P = 6.19 × 10^{-4}, adjusted P = 0.023; Supplementary Table 15). Previous study has reported that *COLGALT2* initiates collagen glycosylation to organize and stabilizes collagen triple helix[31]. Particularly, our data nominated four functionally known genes (*IGF1R*, *PIK3R1*, *SERPINF1* and *WNT3*) that could not be nominated by eQTL data of GTEx tissues and previously reported cartilage tissue. For example, *IGF1R* is located on the cartilage cell membrane and could be activated by IGF1, and further inhibit cartilage matrix degradation and maintain the homeostasis of articular cartilage[32,33].

To reveal biological processes involved for more functionally unknown effector genes, we generated gene sets to annotate their potential functions based on a time series expression data of chondrocytes differentiation[34]. Briefly, we classified all differentially expressed genes into 12 clusters to shown the overall tendency during the chondrocyte differentiation (Supplementary Fig. 13a and Supplementary Table 17),

and a total of 23 co-localized genes belonged to these gene clusters (Fig. 5b). We performed gene ontology (GO) enrichment analyses to reveal the primary function of each cluster, and identified four clusters (cluster 3, 6, 7 and 12) enriched in biological processes related to chondrocytes differentiation (Supplementary Fig. 13b), suggesting the potential function of genes involved in these cluster. Especially, five newly functionally unknown effector genes (*FILIP1*, *OGFOD2*, *MMP23B*, *CTD-2287O16*.5 and *SLC44A2*) belonged to these four clusters (Fig 5b and Supplementary Fig. 13a). In addition, for another seven new effectors genes (*ALAS1*, *NRXN2*, *NSFP1*, *VAC14*, *ABCA9*, *MARVELD2* and *SLC35E2B*) belonged to three clusters (cluster 1, 9 and 10), although the top enriched biological processes of the gene clusters they belonged to were not directly associated with cartilage formation, their expressions were gradually decreased during chondrocyte differentiation (Supplementary Fig. 13a), suggesting their potential roles involved in OA. Taken together, leveraging our tissue match eQTL data, we provide a high confidence OA risk gene list for future research.

Functional fine-mapping reveal candidate causal variants underlying both OA risk and cartilage gene expression

To fill the mechanistic gap between the eVariants and their target genes expression, we established a multi-level functional variants-deciphering pipeline to evaluate whether they act by altering the chromatin accessibility of regulatory elements or disrupting the TF binding to the regulatory elements.

Allele-specific open chromatin (ASoC) eVariants

We first explored the impact of eVariants on the chromatin accessibility (Fig. 6a). Using the read coverage of heterozygous sites called from knee cartilage ATAC-seq data[35,36] (Supplementary Fig. 14; Methods), we identified 1,170 allele-specific open chromatin (ASoC) variants showing significant imbalance between the two alleles (nominal $P \leq 0.05$) (Supplementary Fig. 15 and Supplementary Table 18), which was limited by the low heterozygosity rate of variants in the ATAC-seq data sets. A total of 112 eVariants were annotated as ASoC variants and they associated to a total of the expression of 129 eGenes (Fig. 6b and Supplementary Table 18).

ASoC annotation of eVariants allowed us to interpret the underlying mechanism of OA GWAS loci. For example, the eVariant rs3810154 was associated with the expression of *SLC44A2* (Fig. 6c), which was assigned as an OA risk gene by co-localization analyses (Fig. 6d). Intriguingly, rs3810154 was located in an open region downstream *SLC44A2* and the local chromatin region around the G allele at rs3810154 showed significantly higher accessibility than that of A allele (44 reads for G allele; 20 reads for A; $P = 3.69 \times 10^{-3}$; binomial test; Fig. 6e), suggesting that rs3810154 regulates *SLC44A2* expression by altering the chromatin accessibility. Our soft-clustering and functional enrichment analyses revealed that *SLC44A2* belonged to a gene cluster with increased expression during chondrocytes differentiation (Fig. 6f, Supplementary Fig. 13), indicating its potential important role in cartilage

development (Supplementary Fig. 13b).

TF binding-disruption (TBD) eVariants

We and others had revealed the impacts of genetic variants on disrupting the TF binding on DNA and the consequent gene expression alteration[37-42]. As shown in Fig. 3i, eVariants were significantly enriched in the H3K4me3-defined promoters and H3K27ac-definded enhancers, which load TF binding motifs. We further investigated the effect of these eVariants on the TF binding. Due to the lack of ChIP-seq data of TFs in cartilage tissues, we established an allele-sensitive TF footprint analysis method, which considered the TF binding specificity to both the reference allele and alternative allele of each variant in motif scan step (Fig. 7a; Methods), to speculate the binding sites for 441 expressed TFs (median TPM $\geq$ 1) in our damaged cartilage samples (Supplementary Fig. 16a and Supplementary Table 19). We identified 222 to 39,416 non-overlapped unique binding sites for each TF (Supplementary Fig. 16b and Supplementary Table 19). Comparing to the traditional footprint analysis, our allele sensitive method could detect a total of 22,627 new footprints for all 673 motifs used for analysis (Supplementary Fig. 17). To validate accuracy of the footprint results, we used the CTCF ChIP-seq data from cartilage samples, and the high validation rates were achieved for all five motifs of CTCF (Supplementary Fig. 18), indicating the reliability of our method.

Next, we used a TF motif information content-based method to investigate the

potential motif disruption effects of eVariants (Fig. 7a; Methods). A total of 1087 variants were defined as candidate TBD variants (Supplementary Table 20). Further, we performed a TF expression-interaction eQTL analysis (TFi-eQTL) to assess whether the TF for which the motif was disrupted by a variant was the true mediator of the eQTL effect (Fig. 7a; Methods). Briefly, we constructed a linear-mixed regression model with an interaction term of genotype dosage and TF expression to detect the effect of the eVariant on the co-expression of TF and target gene. As a result, we identified 410 TBD eVariants that might regulate the expression of 369 eGenes through the disruption of binding sites of 231 TFs with a nominal significance threshold (Fig. 7b and Supplementary Table 21). These results allowed us to explain the underlying regulation mechanisms of candidate causal eVariants on seven co-localized effector genes (*PIK3R1*, *GNL3*, *hsa-mir-8072*, *TRIOBP*, *SLC44A2*, *RNF212* and *PCCB*) (Supplementary Table 20). As a representative example, rs11750646 was located within a co-localized signal region of hip OA and cartilage eQTL (Fig. 7c). The T allele at rs11750646 significantly increased the expression of *PIK3R1* ($P = 4.53 \times 10^{-8}$; Fig. 7d). It was potential to disrupt the motifs of five TFs (AR, FOXA2, FOXA3, FOXM1 and FOXO1) that left footprints in an open region downstream the *PIK3R1* (Fig. 7e,f and Supplementary Table 20). TFi-eQTL results showed that rs11750646 most significantly altered the co-expression relationship between *AR* and *PIK3R1* ($P = 2.81 \times 10^{-4}$; Fig. 7g and Supplementary Table 21), suggesting that *AR* was the true regulator affected by this SNP.

We further performed functional experiments to validate our hypothesis derived from our data analyses. Using human chondrocyte-like cell line SW1353, a commonly used cell model for OA *in vitro*, the dual-luciferase reporter assay revealed that rs11750646-T contained sequence showed significant enhancement ability on target gene *PIK3R1*, while rs11750646-C contained sequence showed slightly inhibiting ability for target gene expression (Fig. 7h,i). Further, in line with in silico motif analysis results, our ChIP-PCR followed by Sanger sequencing results proved that AR preferentially bound to the T allele at rs11750646 *in vivo* in SW1353 cells (Fig. 7j). Knockdown of AR significantly repressed *PIK3R1* expression (Fig. 7k). Dual-luciferase reporter assay showed that the gene expression-promoting role of rs11750646-T contained fragment was significantly abolished after AR knockdown, while rs11750646-C contained fragment had a relatively weaker change (Fig. 7l). The *PIK3R1* encoded the regulatory subunit 1 of phosphoinositide-3-kinase (PI3K), which have been reported to regulate the chondrocyte apoptosis through PI3K/Akt pathway and thus involved in development of OA[43]. Collectively, our integrated computational and functional results suggested that AR could allele-preferentially bind the protective T allele (Beta = -0.0521 in GWAS) at rs11750646 to reinforce *PI3KR1* expression and thus associated with OA (Fig. 7m). In addition, it was noteworthy that the binding-preferred allele for AR at this SNP was T, which is an alternative allele that is not presented in human reference genome sequence version hg19 (Fig. 7f) and this AR footprint only identified with our allele-sensitive footprint analysis method, demonstrating its great value in uncovering biological mechanisms of genomic

variants.

Discussion

In this study, we presented the largest sample size of human knee cartilage reported to date. We found the chondrocyte subtype composition heterogeneity vastly contributed to the expression variance across samples, which also partially explained the differential expression between damaged and intact cartilage samples. We reported the eQTLs for cartilage gene and investigated their cell type dependency. By integrating with OA GWAS data and epigenetic data of cartilage samples, we extended the list of OA risk genes and prioritized the potential causal variants linked both to the OA susceptibility and gene expression regulation.

For the high-throughput bulk RNA-seq generated expression data, the reported expression levels are the average number of all cell types in each sample. Thus, cell composition is an important factor that may contribute to the expression variance of heterogeneous tissues. Although chondrocytes cell type composition of cartilage tissue is relatively simple when compared to other tissues, our results still illustrated the difference of chondrocyte subtypes proportion explained a notable part of variance of gene expression across samples. Importantly, we revealed that the differentially expressed genes between damaged and intact cartilage samples using bulk RNA-seq data significantly enriched in the marker gene list of cell types (EC and RegC) with significant composition difference between damaged and intact cartilage. According

to previous study, the main distribution area of EC and RegC is the cartilage superficial zone[19], which suffered abrasion during the development of OA, so it is hard to say whether the cell type composition change is the cause of OA. Our results will prompt the need to consider the non-negligible impact of cell composition on the differential expression analysis and the interpretation of results.

As an efficient approach, eQTL analysis has been widely used to investigate the genetic impact on the gene expression of human tissues or cell types. To date, the GTEx project has generated eQTL data resources for 49 different human tissues. For some easily accessible samples, such as blood, a considerable number of eQTL studies have been reported and obtained a great number of samples. All of these have greatly facilitated the discovery of disease risk genes. As for OA research, Julia et al. published the only eQTL so far for articular cartilage, which is the main tissue related to OA[10]. In our present study, we doubled the sample size and reported a larger number of significant eQTLs for cartilage samples, and significantly complemented the previous research. We also successfully distinguished the independent eQTL signal from each other for 216 genes and further reported the capacity of conditional analysis to uncover the eQTL signals that have been missed in the regular eQTL analysis, indicating a new value for conditional analysis.

Genetic regulatory heterogeneity among chondrocyte subtypes remains poorly understood. By introducing the interaction item of genome dosage and cell proportion

proxy to the regression model, it enabled the cell type dependency test for eQTLs without using single cell expression data or bulk expression data from sorted cell of single type[20]. An increasing number of studies based on this strategy have revealed specifically genetic regulation effect on gene expression between different cell types[20,21,44]. In this study, we indeed demonstrated the genetic regulatory heterogeneity between three chondrocyte subtypes, much like the heterogeneity reported in immune cell subpopulations[45]. We believe it would stimulate future research in this aspect.

Integration of eQTL and GWAS is a powerful method for understanding the molecular mechanism influencing complex traits. The tissue non-matched eQTL data from the GTEx project have broadly been used for co-localization analyses to reveal the gene expression regulation role of OA GWAS loci[5,6], but the tissue specific genetic regulation effect may introduce inevitable false nomination of true effector genes. While efforts have been made to nominate effector genes using cartilage expression data derived annotation information, limitations still remained challenging to overcome. For example, Hollander et al. reported the transcriptome-wide allelic imbalance data in OA cartilage to link the GWAS variants to gene expression[46], but only variants in exons could be evaluated with this method. Further, existing OA studies have limited the nomination of risk genes using colocalization analysis because of their relatively small sample size. Here, by integrating our cartilage eQTL data and OA GWAS data, we nominated 26 novel OA risk genes that could not be

found by other studies. Our data presented in this study helped to make a big step in OA pathogenic mechanism analysis.

We also made efforts to fill the mechanism gap between genomic variants and gene expression. We first revealed that some eVariants were involved in the alteration of chromatin accessibility. As a prerequisite for genomic regulatory functions, open chromatin allowed the binding and assembling of transcription machines, including TFs, co-factors and RNA polymerase. Increasing number of studies have reported the chromatin accessibility quantitative trait loci (caQTLs)[47-50], through which could link the GWAS loci to chromatin accessibility regulation and further to gene expression alteration. However, there are no such resources for cartilage samples. As a compromise, we generated an allele-specific open chromatin SNP set using the publicly available ATAC-seq data from cartilage samples. Although it limited the detectable number of variants that affect chromatin accessibility, we still found the variant rs3810154 linked both to OA susceptibility and *SLC44A2* expression via the alteration of chromatin accessibility. We believe future caQTL data in cartilage would promote the discovery of variants that regulate gene expression through the regulation of chromatin accessibility.

Second, we established an allele-sensitive TF footprint analysis method and then identified footprints for 441 expressed TFs in cartilage samples. Here, we verified one example. The variant rs11750646, associating both with OA susceptibility and

PIK3R1 expression, had a disruption effect on the DNA binding of AR. Notably, this footprint of AR was only identified when using the alternative allele-replaced genome sequence for footprint analysis, it was due to the strong conservation of the alternative allele T in the motif. The allele T is the minor allele at rs11750646, the frequency in our samples was 0.1866, and its frequency in 1000 genome European samples was 0.3598, indicating it is a very common allele, emphasizing the great value of our allele sensitive footprint analysis method. Although more footprints were identified, our result still underestimated the value of allele sensitive footprint analysis method because in the current study, we only considered the variants genotyped in our samples, which only represented a small part of the variants across whole genome sequence.

In summary, we performed comprehensive profiling for transcriptional regulation in knee cartilage from OA patients, reported the largest cartilage eQTL dataset to date, and nominated new risk genes for OA. We revealed potential causal variants for both OA risk and gene regulation based on epigenetic feature resources identified from cartilage specific data. Our study provides a valuable resource to elucidate regulatory mechanisms for genes underlying OA etiology.

Methods

Sample collection

All patients clinically diagnosed with OA and required total knee replacement surgery

from 2019–2022 in Xi'an Honghui Hospital were eligible for this study after signing the informed consent. During total knee replacement surgery, we isolated the knee damaged or/and intact cartilage with scalpels after the knee joint was removed from the donors and froze it in liquid nitrogen rapidly followed by transferred to a -80□ refrigerator for storage. Besides, we collected 2-3 ml peripheral blood from the participants. The summary of clinical information for donors, including age, sex was shown in Supplementary Table S1.

DNA isolation, genotyping and imputation

DNA was extracted from whole blood using FlexGen Blood DNA Kit (CWBIO) according to the manufacturer's instructions. DNA purity and concentration were measured by NanoDrop, and DNA integrity was assessed using gel electrophoresis.

Samples were genotyped using Illumina Infinium Asian Screening Array-24 v1.0 (ASA) containing 659,184 variants in total. TOP strand-based genotypes were reported in plink format using GenomeStudio software (Illumina) and then converted to reference strand-based genotypes using PLINK (v1.9). Subsequently, we performed initial quality control at variants and sample level following the pipelines used by the Genotype-Tissue Expression (GTEx) consortium[27] and that introduced in ref,[51] with a little modification. Briefly, at variants level, keeping variants with consistent genome position and allele with that of 1000 Genomes Project, with genotype missing rate across all samples < 0.02, MAF > 0.01 and the Hardy–Weinberg Equilibrium $P > 1 \times$

10^{-6}. At sample level, keeping samples with genotype missing rate across all variants < 0.02, with matched sex tag between clinical record and genotype-imputed result, with heterozygosity within 3 standard deviations from the mean heterozygosity.

Genotype imputation was performed using IMPUTE2 (v2.3.2)[52] with 1000 Genome Project genotype data (phase 3) as reference panel. To increase the imputation accuracy, we changed the default values of the parameter -k and –buffer to 100 and 300, respectively. After imputation, we performed another round of quality control, keeping variants with quality score > 0.3, genotype missing rate across all samples < 0.05, MAF > 0.05 and the Hardy–Weinberg Equilibrium $P > 1 \times 10^{-6}$, keeping samples with genotype missing rate across all variants < 0.05.

We performed principal component analysis (PCA) using smartpca[53] to detect the stratification of population.

Cartilage RNA isolation, sequencing and data processing

Cartilage RNA was extracted with TRIzol. RNA integrity (RIN) was assessed using the Fragment Analyzer 5400 (Agilent Technologies). Samples with RIN ≥ 5.6 were used for sequencing library preparation using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB) according to manufacturer's instructions. RNA sequencing was performed on an Illumina Novaseq 6000 platform and 150 bp paired-end reads were generated.

Clean RNA-seq reads were mapped to human reference genome hg19 (only contained autosomes, sex chromosomes and mitochondrial chromosomes) with STAR aligner (v2.7.9a)[54] based on the GENCODE v19 (ref.[55]) annotation, the parameters setting was same as the GTEx v8 pipeline. Gene-level read counts and TPM values were generated using RNA-SeQC (v2.3.5)[56] with default parameters. The read mapping rate and base mismatch rate were calculated from the output of SAMtools (v1.9)[57] stats sub-command. The proportions of genomic origin of sequencing reads were produced using QualiMap (v2.2.2-dev)[58]. Samples are needed to meet the following metrics as used by the GTEx consortium[27]: read mapping rate ≥ 0.2, base mismatch rate ≤ 0.01, intergenic mapping rate ≤ 0.3, and rRNA mapping rate ≤ 0.3. In this study, all samples passed this mapping quality control step. Then, we conducted expression outlier filtration with the method described in ref.[59]. Briefly, we calculated the pairwise expression correlation coefficients using the log transformed read count of all genes, let the r_{ij} denote the expression correlation coefficient between sample *i* and sample *j*, we calculated

$$\bar{r}_i = \sum_j r_{ij}/n$$

the average correlation coefficient of sample *i* with all others of the total *n* samples. Lower $\bar{r}_i$ represented a lower quality. Then we calculated

$$D_i = (\bar{r}_i - \bar{\bar{r}})/median(|\bar{r}_i - \bar{\bar{r}}|)$$

to provide a sense of distance from the grand correlation mean $\bar{\bar{r}}$. Samples with $D <$

-5 were considered as outliers and removed. Finally, samples that passed both genotype and RNA-seq quality control were used for the following analysis, which contained a total of 245 samples, of which, 204 were damaged cartilage samples and 31 were intact cartilage samples.

Differential expression analyses

We performed differential expression analyses using the differential expression for repeated measures (dream) R package from variancePartition[25]. For the comparison between damaged and intact cartilage, we produced the gene-level read counts matrix for 31 donors from whom we obtained paired damaged and intact cartilage samples. Genes with count per million (CPM) value < 0.1 in more than half samples were removed, then gene-level read counts were normalized with trimmed means of *m* values (TMM) method applied by edgeR (v3.40.2)[60] and voom-transformed with voomWithDreamWeights function from variancePartition R package. Besides cartilage type (damaged or intact), donor ID, sex, age, first two principal components of genotype, RNA integrity number (RIN), and sequencing batch were also included in the differential expression analysis model to correct the effects of technical and non-cartilage type-related biological factors. FDR ≤ 0.05 was considered as the threshold for the identification of differentially expressed genes.

Gene set enrichment analysis

Gene set enrichment analysis was performed for differentially expressed genes

between paired damaged and intact cartilage samples using GSEA software[61,62]. The marker gene list of seven chondrocyte subtypes were calculated using FindAllMarkers function from R package Seurat (v4.1.3)[63] with the scRNA-seq data and cell labels from ref.[19].The enrichment *P*-values were calculated based on 100,000 times permutation.

Deconvolution of bulk RNA-seq data into chondrocyte subtype composition

Processed single-cell RNA-seq data of knee cartilage[19] was downloaded from the GEO database (GSE104782). UMI counts matrix where the columns were named by cell type (HomC, HTC, RegC, PreHTC, FC, EC and ProC) and the rows were named by gene symbols was used as input for CIBERSORTx[64] web server to produce a chondrocyte subtype signature matrix. Then the signature matrix was applied to the expression matrix of the bulk RNA-seq expression data to speculated estimate the proportions of chondrocyte subtypes using the CIBERSORTx web server. In bulk RNA-seq expression matrix, the columns were named by sample ID, rows were named by gene symbols, and genes with the same symbols were removed. S-mode was turned on for CIBERSORTx to correct the technical bias between single-cell and bulk RNA-seq data.

Variation partition of gene expression

To get the contributor of gene expression variation at sample level, we applied variancePartition (v1.28.4)[65], which is based on linear-mixed regression model, to

measure the impacts of biological and technical confounders on gene expression variance, including donor ID, sex, age, cartilage type (damaged or intact), genetic background (first two principal component of genotype), speculated proportions of six chondrocyte subtypes (HomC, HTC, RegC, PreHTC, FC, EC, ProC was excluded because of the extremely low fraction across all samples), RNA integrity number (RIN), and sequencing batch. Genes with count per million (CPM) value < 0.1 in more than half samples were removed, then gene-level read counts were normalized with trimmed means of *m* values (TMM) method applied by edgeR (v3.40.2)[60] and voom-transformed with voomWithDreamWeights function from variancePartition R package. Participants identity (represented with donor ID) and sequencing batch were modeled as random effects and the others confounders were modeled as fixed effects.

Cis-eQTL mapping

Expression data preprocessing

We followed the *cis*-eQTL mapping pipeline from the GTEx consortium[27]. Expression data from damaged cartilage samples were used for eQTL analysis. Genes with TPM > 0.1 and read count ≥ 6 in at least 20% samples were selected. Further, genes on sex and mitochondrial chromosomes were removed. Filtered expression data were normalized using TMM to correct the library size difference between samples, then the rank-based inverse normal transformation (RINT) was applied for each gene across all samples, this method could increase the analysis power when test the significance of correlation with large sample size than other normalization methods

did[66].

Covariates file generation

PEER factors[67] were calculated to estimate the hidden confounders of gene expression. Following the advice from the GTEx consortium, 30 PEER factors were used for covariates as our sample size is between 150 and 250. Integrated with sex, age, the first two principal components of genotype and the proportions of seven chondrocyte subtypes, we produced a covariate file used for *cis*-eQTL mapping.

Regular cis-eQTL mapping

All variants within 1 Mb of transcription start site (TSS) of each gene were used for regular *cis*-eQTL mapping with linear regression model implemented by FastQTL (v2.0)[68] and nominal *P*-value was calculated for each variant-gene pair. Beta distribution-adjusted empirical *P*-values (from adoptive permutation results by setting --permute 1000 10000 for FastQTL) of top variant of each gene were used for multiple testing correction (FDR estimation) using the method described by Storey and Tibshirani[69], which was implemented in R package qvalue, and FDR ≤ 0.05 was used as a threshold to identify genes with significant eQTL (eGenes). To assess the association significance of variants with eGenes, first, we got the mean of Beta distribution-adjusted empirical *P*-values of two genes closest to the 0.05 FDR threshold (one > 0.05 and the other one ≤ 0.05) as genome-wide empirical *P*-value threshold (p_t), then the nominal *P*-value threshold for each gene was calculated using

the two beta distribution parameters of the minimum *P*-value distribution obtained from the adoptive permutation procedure for the gene based on the genome-wide empirical *P*-value ($F^{-1}(p_t)$, F^{-1} denote the inverse cumulative distribution of beta distribution). For each gene, variants with a nominal *P*-value below the gene-level nominal *P*-value threshold were considered statistically significant (eVariants).

Conditional analysis

To get the independent *cis*-eQTL signal for each eGene, a two-step conditional analysis (a forward stepwise regression followed by a backward selection) was performed using cis_independent mode of tensorQTL[70] with the regular *cis*-eQTL analysis result as input, the lead variant for each independent eQTL signal was reported. For visualization of each post-conditional independent eQTL signal of single gene, we integrated the genotype dosage of the others lead variants of independent signal into the covariates file and re-conducted the *cis*-eQTL analysis with FastQTL for the gene to get a new *P*-value for each variant.

Epigenetic data processing

ATAC-seq data

The ATAC-seq data of knee cartilage samples from 16 donors were downloaded from GEO database (GSE108301)[36]. The reads were mapped to human reference genome hg19 with Bowtie2 (v2.3.5)[71] for each sample, reads with mapping quality < 30 or duplicated were filtered. Peak calling was performed with MACS2 (v2.2.7.1)[72,73] as

described before[38]. Peaks presented only in one sample were removed and the others were merged to produce a consistent peak set.

Histone modification ChIP-seq data

ChIP-seq data for four histone modifications (H3K4me1, H3K4me3, H3K27ac, and H3K27me3) of adult knee cartilage samples were downloaded from from GEO database (GSE111850)[28]. The reads were mapped to human reference genome hg19 with Bowtie2 (v2.3.5) for each sample, reads with mapping quality < 30 or duplicated were removed. Then the bam files from three biological replications of each modification type were merged to a single bam file. Peak calling was performed with MACS2 (v2.2.7.1), for H3K4me1 and H3K27me3, the broad peaks were called with --broad-cutoff set to 0.05, for H3K4me3 and H3K27ac, the narrow peak were called with -q set to 0.05.

Peak files, which were generated by ENCODE project[29] using pseudo-replications-based method, for six histone modification (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3) of hMSCs derived chondrocytes were download from GEO database under the access number GSE188124, GSE187652, GSE187044, GSE187055, GSE187641, and GSE187471, respectively. Then the genome coordinates were transformed from hg38 to hg19 using liftOver with hg38ToHg19.over.chain.gz annotation file. The corresponding bam files were downloaded from ENCODE portal. The alignment coordinates were transformed

to hg19 using CrossMap[74] with hg38ToHg19.over.chain.gz annotation file and then the bam files for replication of each modification type were merged.

Super-enhancers identification

We performed the super-enhancers identification based on the H3K27ac ChIP-seq data using ROSE2 with default settings. The H3K27ac peaks used for indicating the enhancer position and the merged bam files used for enhancer signal strength estimation.

eVariants enrichment analyses

All lead variants of conditional independent eQTL signals were LD-pruned with a threshold $r^2 < 0.8$ to produce a query set. Then for each query variant, we extracted at most 100 random variants from the SNPsnap dataset[75] as a background set, which accounted for MAF deviation (± 0.05), distance to gene deviation (± 0.5), gene density deviation (± 0.5) and LD buddy deviation (± 0.5). Then, the epigenetic enrichment analysis was performed using fisher's exact test.

Cell type-interaction eQTL (ci-eQTL) analysis

As described in ref.[23], to measure the contribution of a single chondrocyte subtype to eQTL effect, the significant variant-gene pairs in regular *cis*-eQTL result were used for ci-eQTL analysis. We modeled following two linear-mixed regression models with R package lme4 (v1.1-30)[76]:

$$E \sim Beta_g \times G + covariates \#(1)$$

$$E \sim Beta_i \times G{:}C + Beta_g \times G + Beta_c \times C + covariates \#(2)$$

where *E* denotes the RINT normalized gene expression value. *G* denotes the genotype dosage. *C* denotes the proportion of specific chondrocyte subtype (ranging from 0 to 1). *Beta* denotes the effect size of each item. Covariates included in the models were the same as that used in regular *cis*-eQTL analysis. Sex was modeled as random effect and the other covariates were modeled as fixed effect.

The difference of those two models was compared with analysis of variance (ANOVA) implemented by R package pbkrtest (v0.5.1)[77] and a χ^2 *P*-value was calculated to indicate the statistical significance. We extracted the lead variant with the smallest χ^2 *P*-value for each tested gene and corrected the χ^2 *P*-value with Benjamini–Hochberg FDR control method[78]. Genes for which the lead variants with FDR < 0.05 and *P*-value < 5×10^{-5} of genotype dosage and cell proportion interaction item were considered as statistically significant cell type-interaction eGenes. We divided the significant variant-gene pairs into driver and non-driver model based on the effect direction of $Beta_g$ from formula (1) and $Beta_i$ from formula (2). The same effect direction meant that the eQTL effect increased as the cell fraction increased (driver model), suggesting the eQTL effect may come from the cell type used for ci-eQTL analysis (Fig. 4a). In contrast, the opposite effect direction meant that the eQTL effect decreased as the cell fraction increased (non-driver model), suggesting the eQTL effect didn't come from the cell type used for ci-eQTL analysis (Fig. 4a). The cell

proportions of EC, preHTC and RegC were used for ci-eQTL analyses because of their broader distribution than that of the other four subtypes (Figs. 2d,e).

Sharing and specificity estimation of the ci-eQTL effect

To assess whether the cell type-interaction variant-gene pairs were shared or specific across multiple chondrocytes subtypes (EC, preHTC and RegC). We applied the multivariate adaptive shrinkage method (mash)[30], which was implemented in R package mashr (v0.2.57), based on the workflow used for the GTEx data (https://github.com/stephenslab/gtexresults). The effect size of interaction item of genotype dosage and cell proportion ($Beta_i$) and the standard error ($s.e._{Betai}$ calculated form formula 2) were used as inputs for mashr. We extracted the variant with the smallest *P*-value across all three cell types, and all tested variants for each gene and produced two matrices which saved their effect sizes and standard errors for the gene in three cell types, respectively. In addition, two matrices with the same format for a 200,000 randomly selected variant-gene pairs set also be produced. A mash model was built using the exchange effects mode to estimate the prior. The model was then applied to all significant variant-gene pairs for each cell type to get the posterior mash effect sizes and the local false sign rate (LFSR). Pairs with LFSR < 0.5 were considered statistically significant.

Co-localization analyses

We downloaded a total of 28 OA GWAS summary data from six published studies[4-9]

(Supplementary Table 13). For each GWAS summary data, we produced a list of 200-Kb signal window centered on each genome-wide significant variant ($P < 5 \times 10^{-8}$). The final signal region list for co-localization analysis was produced after merging the overlapping signal windows. Then, the GWAS association *P*-value, MAF and LD structure (calculated from UK biobank samples) of variants within each region were integrated for further co-localization analysis.

To get *cis*-eQTL signal regions for co-localization analysis, for eQTL signal of each gene from our regular *cis*-eQTL analysis, we merged the 200-Kb signal windows centered on each significant eVariant of genome-wide significant eGenes or 200-Kb signal windows centered on the top variant which with *P*-value < 0.05 for genes that did not reach genome-wide significance. Then, the eQTL association *P*-value, MAF (calculated from our samples) and LD structure (calculated from our samples) of variants within each region were integrated for co-localization analysis. For *cis*-eQTL signal from conditional analysis and ci-eQTL analysis, the information of variants within 1-Mb windows of TSS were used for co-localization analysis.

Co-localization analysis was performed using R package coloc (v5.1.0.1)[79], which implemented a Bayesian test to assess whether two association signals shared a consistent causal variant by reporting the posterior probabilities of following five hypotheses: H0, No association with either trait; H1, Association with trait 1, not with trait 2; H2, Association with trait 2, not with trait 1; H3, Association with trait 1 and

trait 2, two independent SNPs; H4, Association with trait 1 and trait 2, one shared SNP. The results were denoted as PPH0, PPH1, PPH2, PPH3, PPH4, higher PPH4 suggested higher probability that a single variant effecting both two traits. After preparation of input files, we performed co-localization analysis using coloc.abf function with default setting. Genes of which the eQTL signal showed PPH4 ≥ 0.7 with GWAS signal were called co-localized genes and considered as candidate OA risk genes.

To verify that our eQTL data have the capacity to nominate more risk genes for OA GWAS signals, we performed co-localization analysis using the eQTL signals of co-localized genes from all tissues in the GTEx project v8. We also verified that our cartilage eQTL data could nominate more OA risk genes than that using cartilage eQTL data from ref. [10].

Processing of time-series gene expression data of chondrocytes differentiation

Soft-clustering of time-series expression data

To assess the functional roles of colocalized genes, we performed clustering and enrichment analysis on a chondrogenic differentiation time series RNA-seq data (GSE128554)[34]. We supposed the genes with similar expression change patterns have similar functions. First, the gene-level read counts and TPM values of each sample were obtained with the same method used for quantification of genes of our cartilage samples. Then, pairwise differentially expressed gene identification was performed

using R package DESeq2 (v1.28.1)[80]. All genes with $P \leq 0.05$ in at least one comparison were used for clustering analysis.

We performed soft-clustering of all differentially expressed genes based on the mean TPM values at each time point using R package Mfuzz (v2.48.0)[81]. This method extracted the patterns of pre-specified n clusters and then assigned a membership score μ_{ij} ($0 < \mu_{ij} < 1$) to gene i to indicate the degree of membership of this gene for cluster j. We chose to cluster the genes into 12 classes by visual observation. Finally, we considered genes with membership score ≥ 0.6 as final members for each cluster.

GO enrichment analyses

GO enrichment was performed by R package clusterProfiler (v3.16.4)[82] with default setting. GO terms with P-value ≤ 0.5 were considered statistically significant

TF binding-disruption analysis

Allele-sensitive TF footprint analysis

Footprint analysis was limited to the expressed TFs which were defined as the genes with at least one DNA binding motif included in JASPAR 2022 core non-redundant set[83] or HOCOMOCO v11 (ref.[84]) database and with a median TPM ≥ 1 across all damaged cartilage samples.

TF footprint analysis involved a step to search TF motifs in the ATAC signal valleys

using sequence-match-based method. We supposed there was a possible scenario where the reference genome-stored allele of a SNP may strongly disrupt a TF motif but the alternative allele may help to form a sequence context suitable for TF binding. To account for this, we produced an alternative allele--replaced genome sequence based on hg19 using SNPsplit (v0.6.0)[85] by replacing the reference allele to the alternative allele of all variants presented in genotype data of our samples. Both the original reference genome sequence and newly generated allele-replaced genome sequence were used for TF footprint analysis.

We performed TF footprint analysis using TOBIAS (v0.13.3)[86] with default settings, TOBIAS was specially developed to account for the ATAC-seq data specific bias. All 16 bam files from ATAC-seq pipeline described above were merged to produce a single input to increase the detection power. The search regions were limited to the peaks presented in at least two samples. All motifs corresponding to the expressed TFs were used for motif search. The results reported by TOBIAS were the genome coordinates of the motifs and the speculated binding status.

TF binding-disruption evaluation

To investigate the potential impacts of eVariants on the TF binding, we applied a TF motif information content-based method. Briefly, for each motif of expressed TF, we calculated the information content (IC) of each position based on the position frequency matrix (PFM). For each position of the motif, the total possible IC was

fixed

$$IC_{total} = log_2 n = log_2 4 = 2$$

where *n* denote the number of all possible base could be presented in this position, thus *n* is 4 for any TF motif. Then, let the $P(N)$ denote the probability of a certain base in the position, we calculated the sum of the uncertainty of all bases (A, C, G, T):

$$U = -\sum_{N=A}^{T} P(N) \times log_2 P(N)$$

Finally, the final information of the position was calculated:

$$IC_{final} = IC_{total} - U$$

The variants located in a position that with $IC_{final} \geq 0.5$ for a certain motif, and the $|P(ref) - P(alt)| \geq 0.3$ were considered as candidate TF binding-disruption variants.

TF expression-interaction (TFi-eQTL) eQTL analysis

To verify whether the TF for which the motif was disrupted by a variant was the true mediator of the eQTL effect. we performed TFi-eQTL analysis that was similar to the ci-eQTL analysis. Briefly, we modeled following two linear-mixed regression models with lme4 (v1.1-30)[76]:

$$E_{target} \sim Beta_g \times G + covariate \ \#(3)$$

$$E_{target} \sim Beta_{tfi} \times G{:}E_{tf} + Beta_g \times G + Beta_{tf} \times E_{tf} + covariates \ \#(4)$$

where E_{target} denote the RINT normalized expression value of target gene. E_{tf} denote the RINT normalized expression of TF. *Beta* denotes the effect size of each

item. Covariates included in the models were the same as that used in the regular eQTL analysis plus the expression value of TF. Sex was modeled as random effect and the other covariates were modeled as fixed effect. The difference of those two models was compared with analysis of variance (ANOVA) implemented by R package pbkrtest (v0.5.1)[77] and a χ^2 *P*-value was calculated to indicate the statistical significance. For each TF-target pair for test, the motif and variant with smallest *P*-value was retained.

Identification of allele-specific open chromatin (ASoC) SNPs

We adopted the pipeline described in ref.[35] to identify the variants that showed imbalanced chromatin accessibility with a little modification(Supplementary Fig. 13). Briefly, BAM files from 16 ATAC-seq data of cartilage samples[36] were calibrated by WASP[87] and used as inputs for GATK (v4.2.6.1)[88]. Following the GATK Best Practice recommendation, HaplotypeCaller sub-command was used for variants calling based on human hg19 genome sequence and corresponding dbSNP version 138 annotation. Then, the construction and application of InDels and SNPs recalibration model were performed in tandem using VariantRecalibrator and ApplyVQSR sub-command, respectively. In addition, the databases used for VariantRecalibrator as true training set were hg19 versions of Mills and 1000G gold standard InDels list (priority = 12), HapMap v3.3 and 1000G_omni v2.5 (priority = 12), the database used as non-true training set was 1000G phase 1 high confidence SNP list (priority = 10) and dbSNP 138 (priority = 2), the database used as known set was dbSNP 138 (priority = 2).

Heterozygous variants with tranche level > 99.0% and located within ATAC-seq peaks were retained. Due to the low read coverage in individual samples, we added up the read depth corresponding to each allele across all heterozygous samples after checking the consistency of allelic effect direction (Supplementary Fig. 14). Finally, we did the binomial test for variant sites with minimum read depth count (DP) ≥ 20 using R function binom.test with a null hypothesis that the two alleles showed the same effect on chromatin accessibility (namely the parameter p was set to 0.5). FDR control was applied using R function p.adjust.

Cell culture

SW1353 cell line was obtained from Zeye (Shanghai, China). Cells were cultured in DMEM medium supplemented with 10% fetal bovine serum (Biological Industries, Israel), penicillin (100 U/ml) and streptomycin (0.1 mg/ml) in 5% CO_2 at 37℃ incubator.

Transfection and dual-luciferase reporter assays

A 876 bp putative open chromatin region containing rs11750646 (C alle or T alle), and a 1500 bp fragment surrounding *PI3KR1* transcription start site were inserted into pGL3 Basic vector (Promega, USA) (Supplementary Table 21). The luciferase reporter plasmid and phRLTK containing Renilla luciferase were co-transfected into cells using Lipo8000 Transfection Reagent (Beyotime, China). After 48 h, the relative luciferase signal was computed as the ratio of firefly luciferase activity to Renilla

signal, and relative activity was normalized by pGL3-promoter plasmid. All data came from four replicate wells, and statistical analyses were performed by a two-tailed *t* test.

shRNA expression plasmid constructs and shRNA knockdown

For short hairpin RNA (shRNA) knockdown of *AR*, two independent oligonucleotides targeting *AR* were cloned into linearized miR30 backbone. The shRNA and negative control (NC) sequences are shown (Supplementary Table 22). A total of 2.5 μg of each plasmid were independently transfected into 80% confluence of SW1353 cells using Lipo8000 Transfection Reagent (Beyotime, China) according to the manufacturer's instruction. After 48 hours of transfection, total RNA was isolated to detect the mRNA expression by quantitative real-time PCR (qRT-PCR). Moreover, *AR* shRNA plasmids were independently co-transfected with the expression plasmids including rs11750646 (C alle or T alle) with *PI3KR1* promoter used in the luciferase reporter assay. The transfection and measuring of luciferase activity are the same as indicated in dual-luciferase reporter assay section. All data came from at least three replicate wells, and statistical analyses were performed by a two-tailed *t* test.

Cell line RNA isolation, reverse transcription, and quantitative real-time PCR

Total RNA was extracted using RNA fast200 Kit (Fastagen, China). 1 μg of total RNA was then reverse transcribed to cDNA using SuperScripts II First-Strand cDNA Synthesis Kit (Invitrogen, Waltham, MA). qRT-PCR was conducted with the

QuantiTect SYBR® Green PCR Kit (Qiagen, Germany) on Bio-Rad instrument. The $2^{-\Delta\Delta CT}$ method was used for data analysis. The qPCR primers are listed in Supplementary Table 22.

Chromatin immunoprecipitation assay

Chromatin immunoprecipitation was performed using Simple-ChIP® Enzymatic Chromatin IP Kit (Magnetic Beads) (Cell Signaling Technology, USA). Pre-cleared chromatin was immunoprecipitated with monoclonal antibodies against AR (ab108341, Abcam, USA), normal rabbit IgG or antibody against acetyl Histone H3 (included in the kit). The total cell chromatin extract was used as input. Protein-DNA crosslinks were precipitated by using ChIP-Grade Protein G Magnetic Beads. And DNA is purified using DNA purification spin columns (included in the kit). The target DNA fragments were analyzed by Sanger sequencing of PCR fragments harboring SNP (Supplementary Table 22).

Data visualization

ChIP-seq signals were visualized using R package Gviz (v1.32.0)[89]. Heatmap was produced using R package ComplexHeatmap (v2.4.3)[90]. Other results were visualized using R package ggplot2 (v3.3.6).

Data availability

Raw RNA-seq data reported in this paper have been deposited in the Genome

Sequence Archive (GSA)[91] in National Genomics Data Center[92], China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human: HRA006224) that are accessible at https://ngdc.cncb.ac.cn/gsa-human with reasonable request. Individual-level genotype data are available upon reasonable request: T.-L.Y. (yangtielin@xjtu.edu.cn). Access to individual-level genotypes from samples recruited within mainland China is subject to the policies and approvals from the Human Genetic Resource Administration, Ministry of Science and Technology of the People's Republic of China. All differentially expressed genes, lead variants of eQTL analyses (including regular eQTL analysis, conditional analysis, ci-eQTL analysis, and TFi-eQTL analyses), GWAS-eQTL co-localization, and allele-specific open chromatin results are presented as Supplementary Tables.

Processed knee cartilage single-cell RNA data can download from supplementary file at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104782. Knee cartilage eQTL summary data in ref.[10] is available at https://personal.broadinstitute.org/ryank/Southam_FunGen_eQTL_LowGradeCartilage.FastQTL_perm_nom_info.ForMSK-KP_16Jan2021.zip for low-grade (intact) cartilage and https://personal.broadinstitute.org/ryank/Southam_FunGen_eQTL_HighGradeCartilage.FastQTL_perm_nom_info.ForMSK-KP_16Jan2021.zip for high-grade (damaged)

cartilage. The GTEx v8 eQTL summary data which contain associations of all tested variant-gene pairs is available at https://console.cloud.google.com/storage/browser/gtex-resources. Position Frequency Matrices (PFMs) of JASPAR 2022 core non-redundant motif set of vertebrates is available at https://jaspar.genereg.net/download/data/2022/CORE/JASPAR2022_CORE_vertebrates_non-redundant_pfms_jaspar.txt. Position Frequency Matrices (PFMs) of HOCOMOCO v11 motif set is available at https://hocomoco11.autosome.org/final_bundle/hocomoco11/core/HUMAN/mono/HOCOMOCOv11_core_HUMAN_mono_jaspar_format.txt. 1000 Genomes Project Phase 3 haplotype reference is available at https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.tgz. Download links of GWAS summary data of osteoarthritis are presented in Supplementary Table 13.

Code availability

Code used in this study is available at https://github.com/tianwen0003/OA-eQTL. The GTEx consortium modified version of FastQTL and corresponding auxiliary scripts were available at https://github.com/francois-a/fastqtl. Scripts used for gene-level filtration and PEER factor calculation are available at https://github.com/broadinstitute/gtex-pipeline/archive/refs/tags/gtex_v8.tar.gz. Pipeline code for mashR is available at https://github.com/stephenslab/gtexresults. Additional code may be available upon reasonable request from the corresponding

authors.

Acknowledgements

We thank the patients who donated material for these studies. This work was supported by the National Natural Science Foundation of China (32170616, 32370653, and 82170896); Science Fund for Distinguished Young Scholars of Shaanxi Province (2021JC-02); Innovation Capability Support Program of Shaanxi Province (2022TD-44); Key Research and Development Project of Shaanxi Province (2022GXLH-01-22, 2023-YBSF-180), and the Fundamental Research Funds for the Central Universities. This study was also supported by the High-Performance Computing Platform and Instrument Analysis Center of Xi'an Jiaotong University.

Author contributions

Conceptualization: T.-L.Y. and Y.G.; methodology: W.T., F.J., C.W., R.-H.H. and S.-S.D.; data analysis: W.T. and C.-Y.H.; experiment: C.W. and K.A.; visualization: W.T. and C.W.; resource: S.-Y.H., H.-M.S., H.-W.G.and Z.Y.; public data collection: W.T. and C.W.; writing-original draft: W.T. and C.W.; writing-review and editing: T.-L.Y. and Y.G.; project administration and supervision: T.-L.Y. and Y.G.

Competing interests

The authors declare no competing interests.

References

1. Long, H. *et al.* Prevalence Trends of Site-Specific Osteoarthritis From 1990 to 2019: Findings From the Global Burden of Disease Study 2019. *Arthritis & Rheumatology* **74**, 1172-1183 (2022).
2. Hunter, D.J. & Bierma-Zeinstra, S. Osteoarthritis. *The Lancet* **393**, 1745-1759 (2019).
3. Cibrián Uhalte, E., Wilkinson, J.M., Southam, L. & Zeggini, E. Pathways to understanding the genomic aetiology of osteoarthritis. *Human Molecular Genetics* **26**, R193-R201 (2017).
4. Zengini, E. *et al.* Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis. *Nature Genetics* **50**, 549-558 (2018).
5. Tachmazidou, I. *et al.* Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nat Genet* **51**, 230-236 (2019).
6. Boer, C.G. *et al.* Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations. *Cell* **184**, 4784-4818 e17 (2021).
7. Boer, C.G. *et al.* Genome-wide association of phenotypes based on clustering patterns of hand osteoarthritis identify WNT9A as novel osteoarthritis gene. *Annals of the rheumatic diseases* **80**, 367-375 (2021).
8. Donertas, H.M., Fabian, D.K., Valenzuela, M.F., Partridge, L. & Thornton, J.M. Common genetic associations between age-related diseases. *Nat Aging* **1**, 400-412 (2021).
9. Jiang, L., Zheng, Z., Fang, H. & Yang, J. A generalized linear mixed model association tool for biobank-scale data. *Nat Genet* **53**, 1616-1621 (2021).
10. Steinberg, J. *et al.* A molecular quantitative trait locus map for osteoarthritis. *Nat Commun* **12**, 1309 (2021).
11. Gamazon, E.R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics* **50**, 956-967 (2018).
12. Bryois, J. *et al.* Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nature Neuroscience* **25**, 1104-1112 (2022).
13. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
14. van der Wijst, M.G.P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat Genet* **50**, 493-497 (2018).
15. Schmiedel, B.J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715 e16 (2018).
16. Ota, M. *et al.* Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184**, 3006-3021 e17 (2021).
17. Dreier, R. Hypertrophic differentiation of chondrocytes in osteoarthritis: the developmental aspect of degenerative joint disorders. *Arthritis Research & Therapy* **12**, 216 (2010).
18. Prein, C. *et al.* Structural and mechanical properties of the proliferative zone of the developing murine growth plate cartilage assessed by atomic force microscopy. *Matrix Biology* **50**, 1-15 (2016).
19. Ji, Q. *et al.* Single-cell RNA-seq analysis reveals the progression of human osteoarthritis. *Ann Rheum Dis* **78**, 100-110 (2019).
20. Westra, H.J. *et al.* Cell Specific eQTL Analysis without Sorting Cells. *PLoS Genet* **11**, e1005223 (2015).

21. Kim-Hellmuth, S. *et al.* Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**(2020).
22. Elorbany, R. *et al.* Single-cell sequencing reveals lineage-specific dynamic genetic regulation of gene expression during human cardiomyocyte differentiation. *PLoS Genet* **18**, e1009666 (2022).
23. Sheng, X. *et al.* Mapping the genetic architecture of human traits to cell types in the kidney identifies mechanisms of disease and potential treatments. *Nature Genetics* **53**, 1322-1333 (2021).
24. Patel, D. *et al.* Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue. *Transl Psychiatry* **11**, 250 (2021).
25. Hoffman, G.E. & Roussos, P. Dream: powerful differential expression analysis for repeated measures designs. *Bioinformatics* **37**, 192-201 (2021).
26. Sturm, G., List, M. & Zhang, J.D. Tissue heterogeneity is prevalent in gene expression studies. *NAR Genomics and Bioinformatics* **3**(2021).
27. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
28. Ferguson, G.B. *et al.* Mapping molecular landmarks of human skeletal ontogeny and pluripotent stem cell-derived articular chondrocytes. *Nature Communications* **9**, 3634 (2018).
29. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
30. Urbut, S.M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* **51**, 187-195 (2019).
31. Bann, J.G., Peyton, D.H. & Bächinger, H.P. Sweet is stable: glycosylation stabilizes collagen. *FEBS Letters* **473**, 237-240 (2000).
32. Middleton, J., Manthey, A. & Tyler, J. Insulin-like growth factor (IGF) receptor, IGF-I, interleukin-1 beta (IL-1 beta), and IL-6 mRNA expression in osteoarthritic and normal human cartilage. *Journal of Histochemistry & Cytochemistry* **44**, 133-141 (1996).
33. Tanaka, N. *et al.* The attenuation of insulin-like growth factor signaling may be responsible for relative reduction in matrix synthesis in degenerated areas of osteoarthritic cartilage. *BMC Musculoskeletal Disorders* **22**, 231 (2021).
34. Vail, D.J., Somoza, R.A., Caplan, A.I. & Khalil, A.M. Transcriptome dynamics of long noncoding RNAs and transcription factors demarcate human neonatal, adult, and human mesenchymal stem cell-derived engineered cartilage. *Journal of Tissue Engineering and Regenerative Medicine* **14**, 29-44 (2020).
35. Zhang, S. *et al.* Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. *Science* **369**, 561-565 (2020).
36. Liu, Y. *et al.* Chromatin accessibility landscape of articular knee cartilage reveals aberrant enhancer regulation in osteoarthritis. *Scientific Reports* **8**, 15499 (2018).
37. Wang, N.N. *et al.* Enhancer variants on chromosome 2p14 regulating SPRED2 and ACTR2 act as a signal amplifier to protect against rheumatoid arthritis. *Am J Hum Genet* **110**, 625-637 (2023).
38. Wang, C. *et al.* Lineage-selective super enhancers mediate core regulatory circuitry during adipogenic and osteogenic differentiation of human mesenchymal stem cells. *Cell Death & Disease* **13**, 866 (2022).

39. Gao, P. *et al.* Biology and Clinical Implications of the 19q13 Aggressive Prostate Cancer Susceptibility Locus. *Cell* **174**, 576-589.e18 (2018).
40. Chen, X.F. *et al.* An Osteoporosis Risk SNP at 1p36.12 Acts as an Allele-Specific Enhancer to Modulate LINC00339 Expression via Long-Range Loop Formation. *Am J Hum Genet* **102**, 776-793 (2018).
41. Dong, S.S. *et al.* An Intronic Risk SNP rs12454712 for Central Obesity Acts As an Allele-Specific Enhancer To Regulate BCL2 Expression. *Diabetes* **70**, 1679-1688 (2021).
42. Thynn, H.N. *et al.* An Allele-Specific Functional SNP Associated with Two Systemic Autoimmune Diseases Modulates IRF5 Expression by Long-Range Chromatin Loop Formation. *J Invest Dermatol* **140**, 348-360 e11 (2020).
43. Fan, Z. *et al.* MiR-155 promotes interleukin-1β-induced chondrocyte apoptosis and catabolic activity by targeting PIK3R1-mediated PI3K/Akt pathway. *Journal of Cellular and Molecular Medicine* **24**, 8441-8451 (2020).
44. Donovan, M.K.R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K.A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat Commun* **11**, 955 (2020).
45. Aguirre-Gamboa, R. *et al.* Deconvolution of bulk blood eQTL effects into immune cell subpopulations. *BMC Bioinformatics* **21**, 243 (2020).
46. den Hollander, W. *et al.* Annotating Transcriptional Effects of Genetic Variants in Disease-Relevant Tissue: Transcriptome-Wide Allelic Imbalance in Osteoarthritic Cartilage. *Arthritis & Rheumatology* **71**, 561-570 (2019).
47. Kosoy, R. *et al.* Genetics of the human microglia regulome refines Alzheimer's disease risk loci. *Nat Genet* **54**, 1145-1154 (2022).
48. Liang, D. *et al.* Cell-type-specific effects of genetic variation on chromatin accessibility during human neuronal differentiation. *Nature Neuroscience* **24**, 941-953 (2021).
49. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat Genet* **50**, 424-431 (2018).
50. Currin, K.W. *et al.* Genetic effects on liver chromatin accessibility identify disease regulatory variants. *The American Journal of Human Genetics* **108**, 1169-1189 (2021).
51. Wang, T. *et al.* A pipeline for RNA-seq based eQTL analysis with automated quality control procedures. *BMC Bioinformatics* **22**, 403 (2021).
52. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
53. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909 (2006).
54. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
55. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**, 1760-1774 (2012).
56. Graubert, A., Aguet, F., Ravi, A., Ardlie, K.G. & Getz, G. RNA-SeQC 2: efficient RNA-seq quality control and quantification for large cohorts. *Bioinformatics* **37**, 3048-3050 (2021).
57. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
58. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292-294 (2015).

59. Wright, F.A. *et al*. Heritability and genomics of gene expression in peripheral blood. *Nature Genetics* **46**, 430-437 (2014).
60. McCarthy, D.J., Chen, Y. & Smyth, G.K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**, 4288-4297 (2012).
61. Subramanian, A. *et al*. GSEA: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
62. Mootha, V.K. *et al*. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267-273 (2003).
63. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
64. Newman, A.M. *et al*. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37**, 773-782 (2019).
65. Hoffman, G.E. & Schadt, E.E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).
66. Bishara, A.J. & Hittner, J.B. Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychol Methods* **17**, 399-417 (2012).
67. 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**, 500-7 (2012).
68. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-85 (2016).
69. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440-9445 (2003).
70. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biology* **20**, 228 (2019).
71. Langmead, B. & Salzberg, S.L. Bowtie2: Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
72. Zhang, Y. *et al*. MACS: Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
73. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nature Protocols* **7**, 1728-1740 (2012).
74. Zhao, H. *et al*. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006-1007 (2013).
75. Pers, T.H., Timshel, P. & Hirschhorn, J.N. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* **31**, 418-20 (2015).
76. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1 - 48 (2015).
77. Halekoh, U. & Højsgaard, S. A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. *Journal of Statistical Software* **59**, 1 - 32 (2014).
78. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*

57, 289-300 (1995).

79. Giambartolomei, C. *et al*. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).

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

81. Kumar, L. & Futschik, M.E. Mfuzz: a software package for soft clustering of microarray data. *Bioinformation* **2**, 5 (2007).

82. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**, 284-287 (2012).

83. Castro-Mondragon, J.A. *et al*. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50**, D165-D173 (2022).

84. Kulakovskiy, I.V. *et al*. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research* **46**, D252-D259 (2017).

85. Krueger, F. & Andrews, S. SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes [version 2; peer review: 3 approved]. *F1000Research* **5**(2016).

86. Bentsen, M. *et al*. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* **11**, 4267 (2020).

87. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J.K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**, 1061-3 (2015).

88. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303 (2010).

89. Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. in *Statistical Genomics: Methods and Protocols* (eds. Mathé, E. & Davis, S.) 335-351 (Springer New York, New York, NY, 2016).

90. Gu, Z. Complex heatmap visualization. *iMeta* **1**, e43 (2022).

91. Chen, T. *et al*. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. *Genomics, Proteomics & Bioinformatics* **19**, 578-583 (2021).

92. Members, C.-N. & Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2023. *Nucleic Acids Research* **51**, D18-D28 (2023).

Figure Legends

Figure 1. Overview of the study

We collected damaged and intact knee cartilage samples for bulk RNA-seq and blood samples for DNA genotyping. We then deconvoluted bulk gene expression data into chondrocyte subtype proportions based on single-cell RNA-seq data of knee cartilage samples. We performed *cis*-eQTL mapping and followed by conditional analysis to identify independent eQTL signals for each eGene. We also performed cell type-interaction eQTL analyses to reveal the genetically regulatory heterogeneity of gene expression between chondrocyte subtypes. By integrating our eQTL data and OA GWAS summary data, we nominated risk genes for OA and annotated their function using multiple lines of evidences. Finally, using our established multi-level functional variants-deciphering pipeline, we performed functional fine-mapping based on cartilage epigenetic features that called from cartilage ATAC-seq data to nominate the candidate causal eVariants which may regulate gene expression through alteration of chromatin accessibility and TF binding.

Figure 2. Expression analyses results

a, The differentially expressed genes (DEGs) (FDR $\leq$ 0.05, no fold change cutoff) between paired damaged and intact cartilage samples when not corrected the impact of cell composition heterogeneity. The x axis is log-converted fold change (damaged vs intact) of expression; The y axis is statistical significance in $-\log_{10}$(adjusted *P*-value). **b,c**, Gene set enrichment analyses (GSEA) result showed the genes

expressed higher in intact cartilage samples (*n* = 715 as showed in Fig. f) enriched in EC (b) and RegC markers (c). The FDR q-value was calculated based on the *P*-values obtained from 100,000 permutations. **d**, Chondrocytes subtype proportions speculated by CIBERSORTx per sample. Each column is a different sample. The y axis is the accumulated cell proportions. **e**, Distribution of proportions of each chondrocyte subtype across all damaged and intact cartilage samples. The y axis is the sample density; The x axis is the cell proportion. **f**, Comparison of speculated chondrocytes subtype proportions between paired damaged and intact cartilage sample. The *P*-values were calculated using paired Student's t-test. The x axis is different cell type; The y axis is the cell proportion. **g**, Distribution of gene expression variance explained per gene by different confounders. The x axis is different factors; The y axis is percentage (%) of total variance explained. Biological factors are colored in light blue; artificial factors are colored in orange; chondrocyte subtype proportions are colored in brown. **h**, Distribution of gene expression variance explained per gene by all chondrocyte subtype proportions. The y axis is percentage (%) of total variance explained.

Figure 3. *cis*-eQTL results

a, Manhattan plot shows the association strength of lead variants of all eQTL signals of all tested gene (*n* = 18,756). The y axis is the statistical significance in $-\log_{10}$(*P*-value). **b**, Biotype distribution of eGenes. **c**, Comparison of expression and variable coefficients distribution of significant eGenes and non-eGenes. The x axis is

log-converted mean expression (top panel) and variable coefficients (bottom panel). *P*-value was calculated using Wilcoxon rank sum test. **d** Number of genes with certain number of independent signals in conditional analysis. **e**, Distribution of genes with or without newly identified independent signal (not significant in regular *cis*-eQTL analysis) after conditional analysis. **f**, *SNAP91* has two independent eQTL signals tagged by rs145051363 (middle) and rs623263 (bottom), respectively. In regular *cis*-eQTL analysis, only the rs145051363 tagged signal is significant, the rs623263 tagged signal is not significant (top). **g**, Genome distribution of lead eVariants of all independent eQTL signals. **h**, Distance of lead variants of all independent eQTL signals from the transcription start site (TSS) of the associated eGenes and corresponding nominal statistical significance in $-\log_{10}$(*P*-value). **i**, Enrichment of LD-pruned ($r^2 < 0.8$) lead eVariants in different epigenetic features. Data are showed as odds ratio (OR) ± 95% confidence intervals. The utilized epigenetic feature data are from both original adult human knee cartilage samples and hMSCs derived chondrocytes. ATAC-seq data is absent from hMSCs derived chondrocytes. H3K36me3 and H3K9me3 ChIP-seq data are absent from adult human knee cartilage. **j**, Enrichment of LD-pruned ($r^2 < 0.8$) lead eVariants in H3K27ac-defined SEs and TEs from both adult human knee cartilage and hMSCs derived chondrocytes.

Figure 4. Genetically regulatory heterogeneity between chondrocyte subtypes

a, Schematic outline of ci-eQTL analysis. For significant variant-gene pairs in regular *cis*-eQTL analysis, an interaction term of genotype dosage and cell type proportion is

modeled in the linear-mixed model to test the cell type dependency of certain eQTL effect. Driver model means the eQTL effect emerging with increase of cell proportion. Non-driver model means that eQTL effect disappearing with increase of cell proportion. **b**, Number of significant genes discovered in ci-eQTL analyses that aligned to driver and non-driver model for three chondrocyte subtypes. **c**, Heatmap shows the detailed mash results of all significant ci-eQTL results. The color depth represents the mash-reported posterior effect size estimates of interaction item of genotype dosage and cell proportion. Red represents the diver model; Blue represents non-Driver model. Each row is a variant-gene pair. The rows and columns were clustered based on the similarity. **d**, Association between the genotypes of variant rs12620516 and *IGFBP2* gene expression. The y axis is the normalized expression of *IGFBP2* in damaged cartilage samples; the x axis is the genotype at the rs12620516. The whiskers extend to the 5th and 95th percentiles; the outliers are represented by the dots. The *P*-value was reported by FastQTL nominal procedure. **e**, ci-eQTL results of rs12620516~*IGFBP2* pair across three different chondrocyte subtypes. The y axis is the normalized expression of *IGFBP2* in damaged cartilage samples; the x axis is the cell proportion. Each point represents one sample and colored by genotype types. The *P*-values indicate the statistical significance of interaction term in linear-mixed regression model. LFSR was estimated by mash. **f**, Association between the genotypes of variant rs2356553 and *MAST2* gene expression. **g**, ci-eQTL results of rs2356553~*MAST2* pair across three different chondrocyte subtypes. The y axis is the normalized expression of *MAST2* in damaged cartilage samples; the x axis is the cell

proportion. Each data point represents one sample and colored by genotype types. The *P*-values indicate the statistical significance of interaction term in linear-mixed regression model. LFSR was estimated by mash. **h**, Association between the genotypes of variant rs1408829 and *TOR3A* gene expression. **i**, ci-eQTL results of rs1408829~*TOR3A* pair across three different chondrocyte subtypes. The y axis is the normalized expression of *TOR3A* in damaged cartilage samples; the x axis is the cell proportion. Each data point represents one sample and colored by genotype types. The *P*-values indicate the statistical significance of interaction term in linear-mixed regression model. LFSR was estimated by mash.

Figure 5. Co-localization analysis nominate risk genes for osteoarthritis

a, Co-localization result of OA GWAS loci with cartilage eQTLs. The numbers next to the triangles represent the highest value of probability of a single shared variant for both GWAS and eQTL signal (PPH4) for each type of OA site (see more details in Supplementary Table 13). Genes nominated with PPH4 ≥ 0.7 in at least one GWAS data were showed. **b**, Function annotation of co-localized genes from multiple lines of evidences.

Figure 6. Allele-specific open chromatin (ASoC) variant links both to OA risk and cartilage gene expression

a, The schematic diagram of an allele specific open chromatin variant. **b**, Number of eVariant that is ASoC variant and eGenes associated with ASoC eVaraints under

different statistical significance threshold. **c**, Association between the genotypes of variant rs3810154 and *SLC44A2* gene expression. The y axis is the normalized expression of *SLC44A2* in damaged cartilage samples; the x axis is the genotype at the rs3810154. The whiskers extend to the 5th and 95th percentiles; Each dot is a different sample. The *P*-value was reported by FastQTL nominal procedure. **d**, Co-localization of OA GWAS locus and eQTL signal associated to *SLC44A2*. The y axis is the statistical significance of OA GWAS and cartilage eQTL for *SLC44A2* gene in $-\log_{10}$(*P*-value). The x axis is the chromosome coordinate. **e**, Variant rs3810154 located within an open chromatin region marked by an ATAC-seq peak downstream the *SLC44A2* and altered the local chromatin accessibility. The sequencing reads with G allele piled up a significantly higher peak than that piled up by A allele-contained reads. **f**, *SLC44A2* expression at each time point during the chondrocyte differentiation. The y axis is the expression value in TPM; The x axis is the time point. Each time point has three biological replications; The data are showed as mean ± sd.

Figure 7. TF binding-disruption (TBD) variants links both to OA risk and cartilage gene expression

a, Schematic diagram for TF binding-disruption analysis module in our multiple-level functional variants-deciphering pipeline. Three main steps included. For the allele-sensitive TF footprint analysis step, we introduced both the original reference genome sequence and the allele-replaced genome sequence corresponding to the signal valley regions as inputs for TF motif search. In this example scenario, only the

alternative allele-replaced genome sequence matched to the TF motif. **b**, The number of eVariant that is TBD variant, eGenes associated with TBD eVaraints and TFs whose binding sites were affected under different statistical significance threshold. **c**, Co-localization of OA GWAS locus and eQTL signal associated to *PIK3R1*. The y axis is the statistical significance of OA GWAS and cartilage eQTL for *PIK3R1* gene in -$\log_{10}$(*P*-value). The x axis is the chromosome coordinate. **d**, Association between the genotypes of variant rs11750646 and *PIK3R1* gene expression. The y axis is the normalized expression of *PIK3R1* in damaged cartilage samples; the x axis is the genotype at the rs11750646. The whiskers extend to the 5th and 95th percentiles; Each dot is a different sample. The *P*-value was reported by FastQTL nominal procedure. **e**, rs11750646 overlapped with footprints of AR, FOXA2, FOXA3, FOXM1 and FOXO1, that were identified in an open chromatin region marked by an ATAC peak downstream the *PIK3R1*. **f**, Reference allele C at rs11750646 disrupts the motif (ID: ANDR_HUMAN.H11MO.0.A) of AR. The motif coordinate was shown on the top. The symbol (+) represents the genome sequence shown is positive strand. **g**, TFi-eQTL analysis reveal rs11750646 modified the regulatory role of *AR* on *PIK3R1* expression. The y axis is the normalized expression of *PIK3R1* in damaged cartilage samples; The x axis is the normalized expression of *AR* in damaged cartilage samples. The *P*-value indicates the statistical significance of interaction term of TF expression and target expression in linear-mixed regression model. **h**, Schematic diagram of dual luciferase reporter vector. **i**, Detection of activity effect on open chromatin region containing rs11750646 (T and C allele) compared with pGL3-*PIK3R1*-promoter in

SW1353 cells by dual-luciferase reporter assay. Luciferase signal was computed as the ratio of firefly luciferase activity to Renilla signal and relative activity was normalized by pGL3-*PIK3R1*-promoter. *P*-values are calculated using two-tailed Student's *t*-test. **j**, AR preferred the binding to the T allele at rs11750646 confirmed by ChIP followed by Sanger sequencing in SW1353cells. **k**, Effect of *AR* knockdown on *PIK3R1* expression in SW1353 cells. Two independent shRNAs (shRNA1 and shRNA2) and shRNA-NC (negative control) were used. RT-qPCR analysis of gene expression normalized to *GAPDH* expression level. *P*-values are calculated using two-tailed Student's *t*-test. **l**, Dual-luciferase reporter assay containing rs11750646-C or -T allele plasmids co-transfected with *AR* shRNAs in SW1353 cells. The shRNA-NC is used as the negative control. Luciferase signals are normalized to Renilla signals. *P*-values are calculated using two-tailed Student's *t*-test. **m**, The OA risk allele C at rs11750646 disrupts binding of AR to a regulatory element, resulting in decreased *PIK3R1* expression and further increasing the risk of osteoarthritis.

Fig. 1: Overview of the study

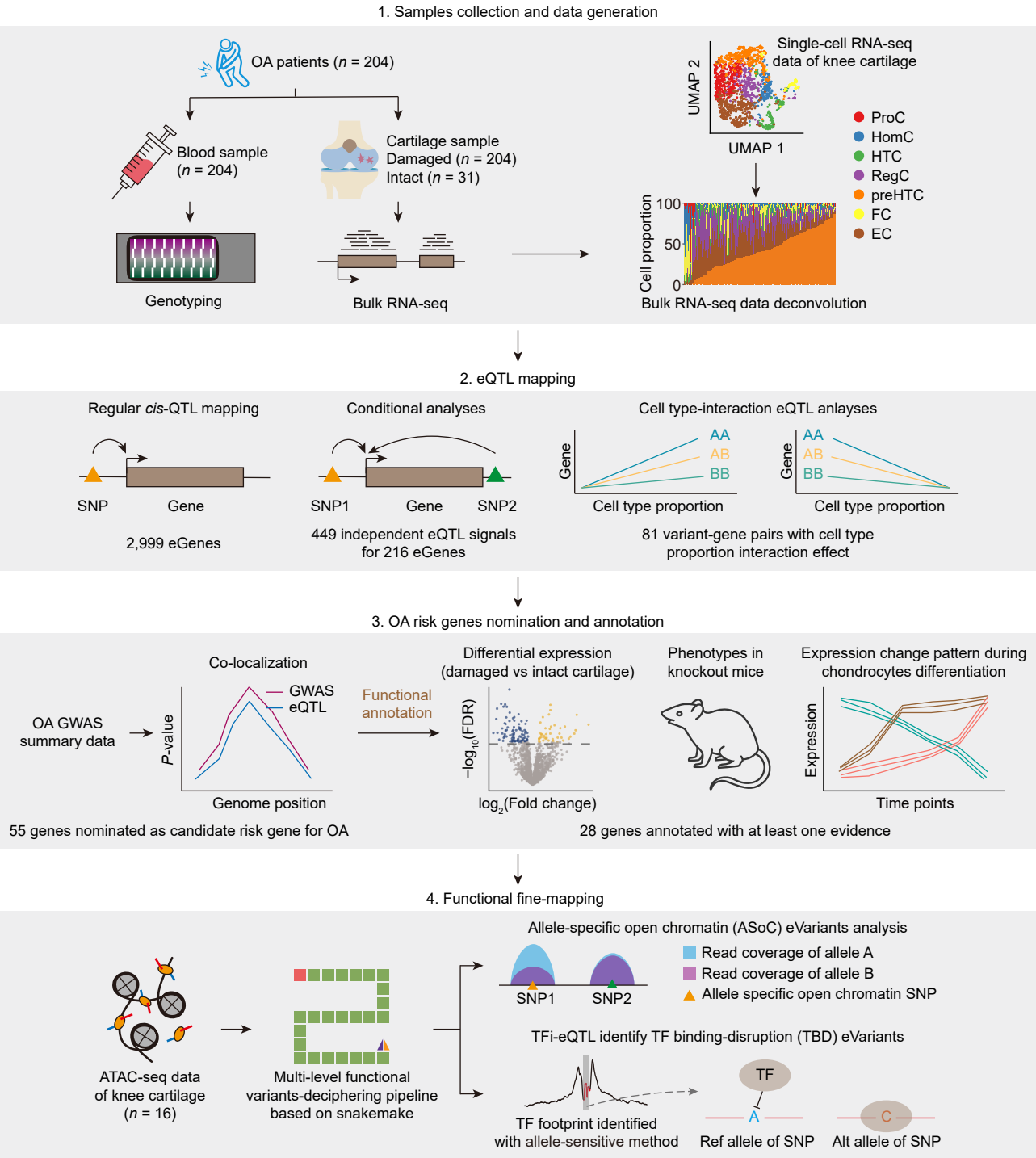

Fig. 2: Expression analyses results

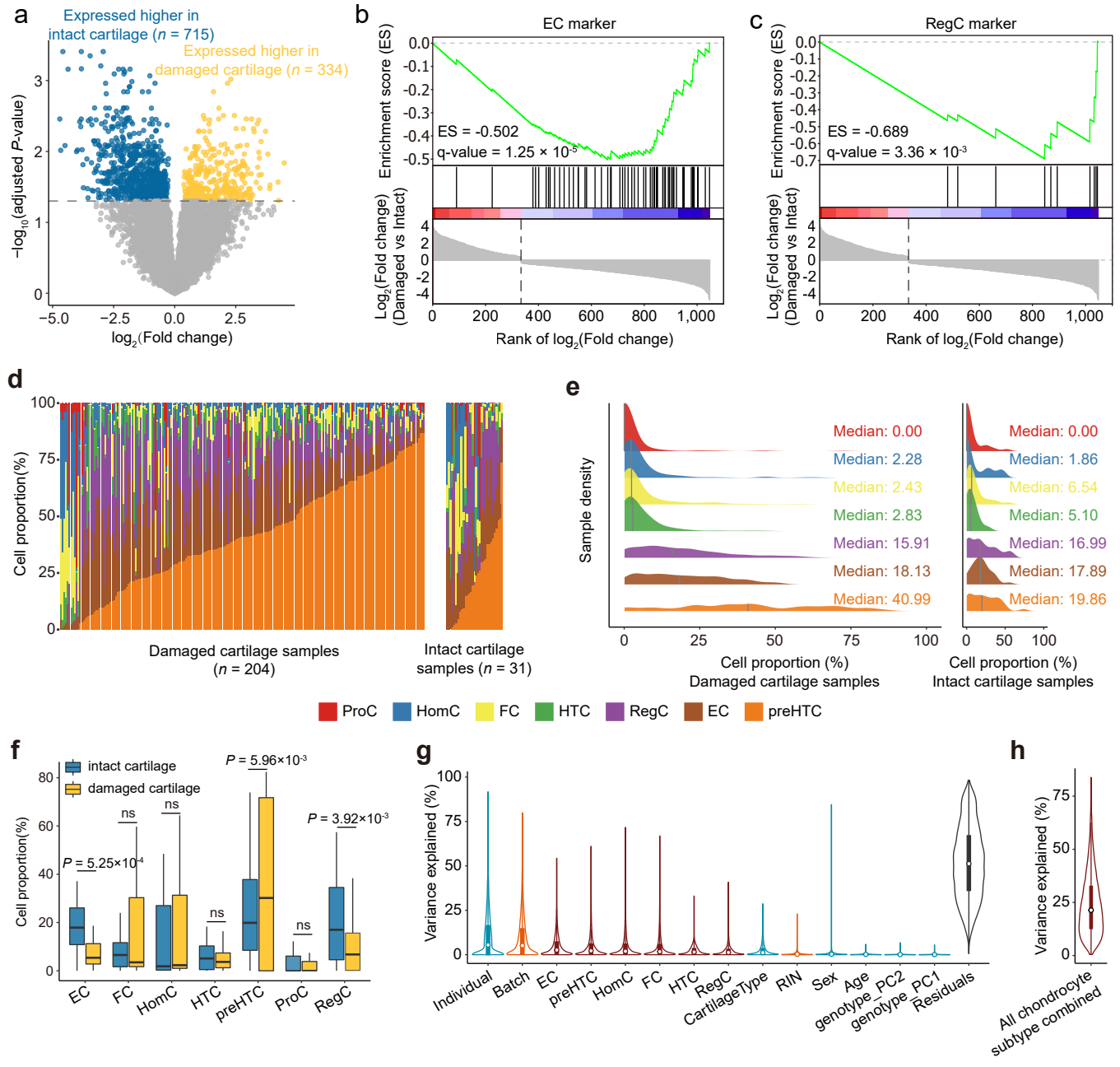

Fig. 3: *cis*-eQTL results

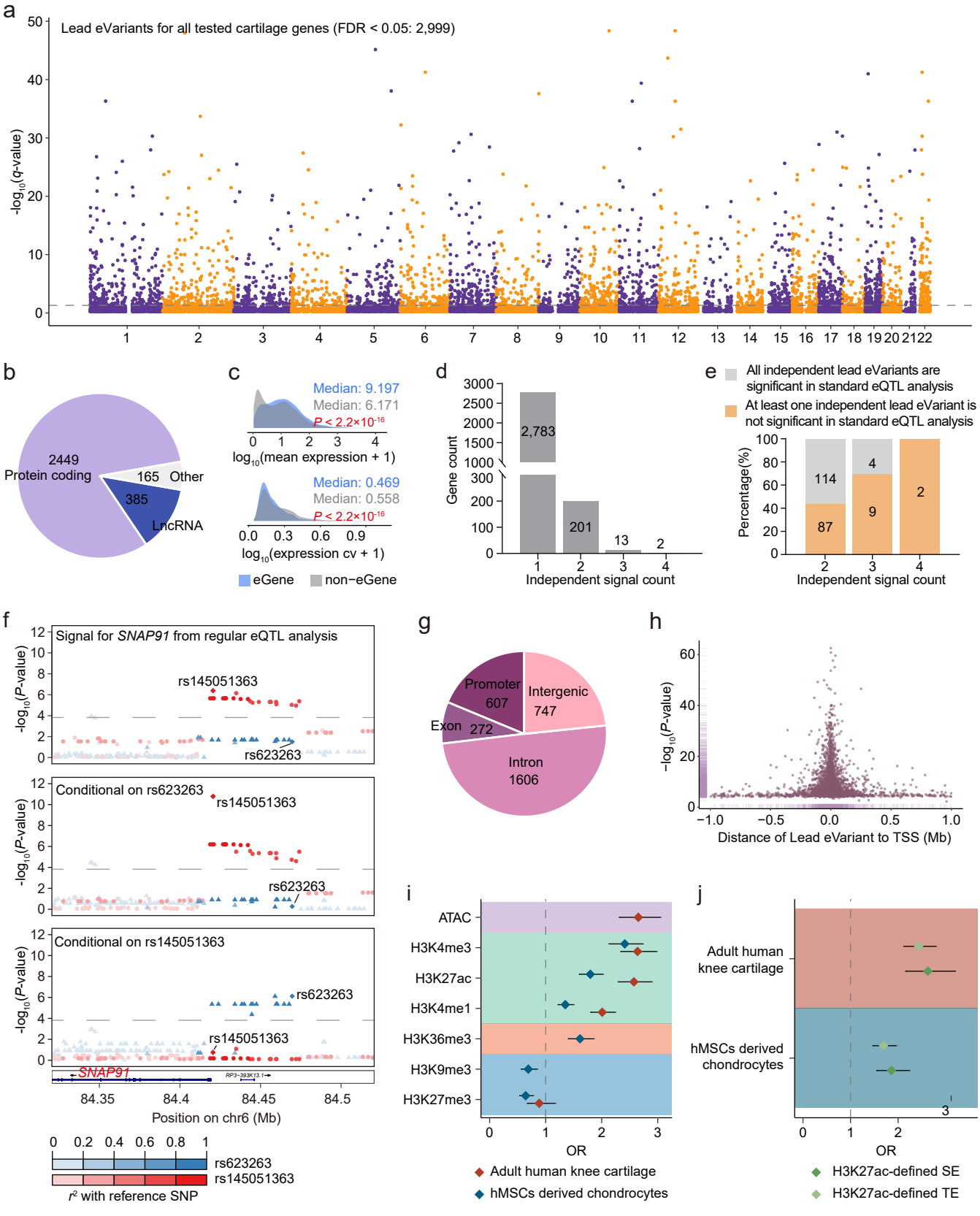

Fig. 4: Genetically regulatory heterogeneity between chondrocyte subtypes

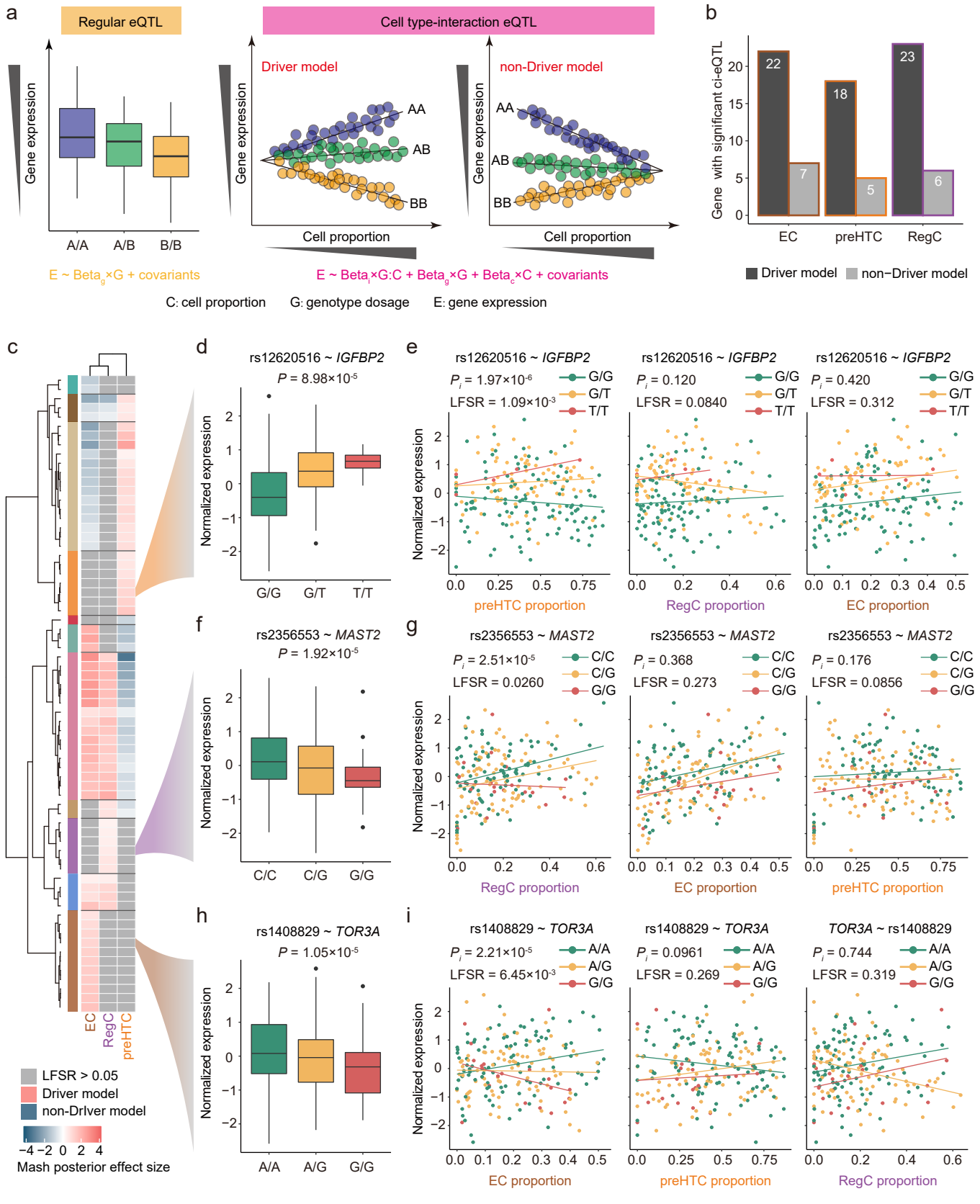

Fig. 5: Co-localization analysis nominate candidate risk genes for osteoarthritis

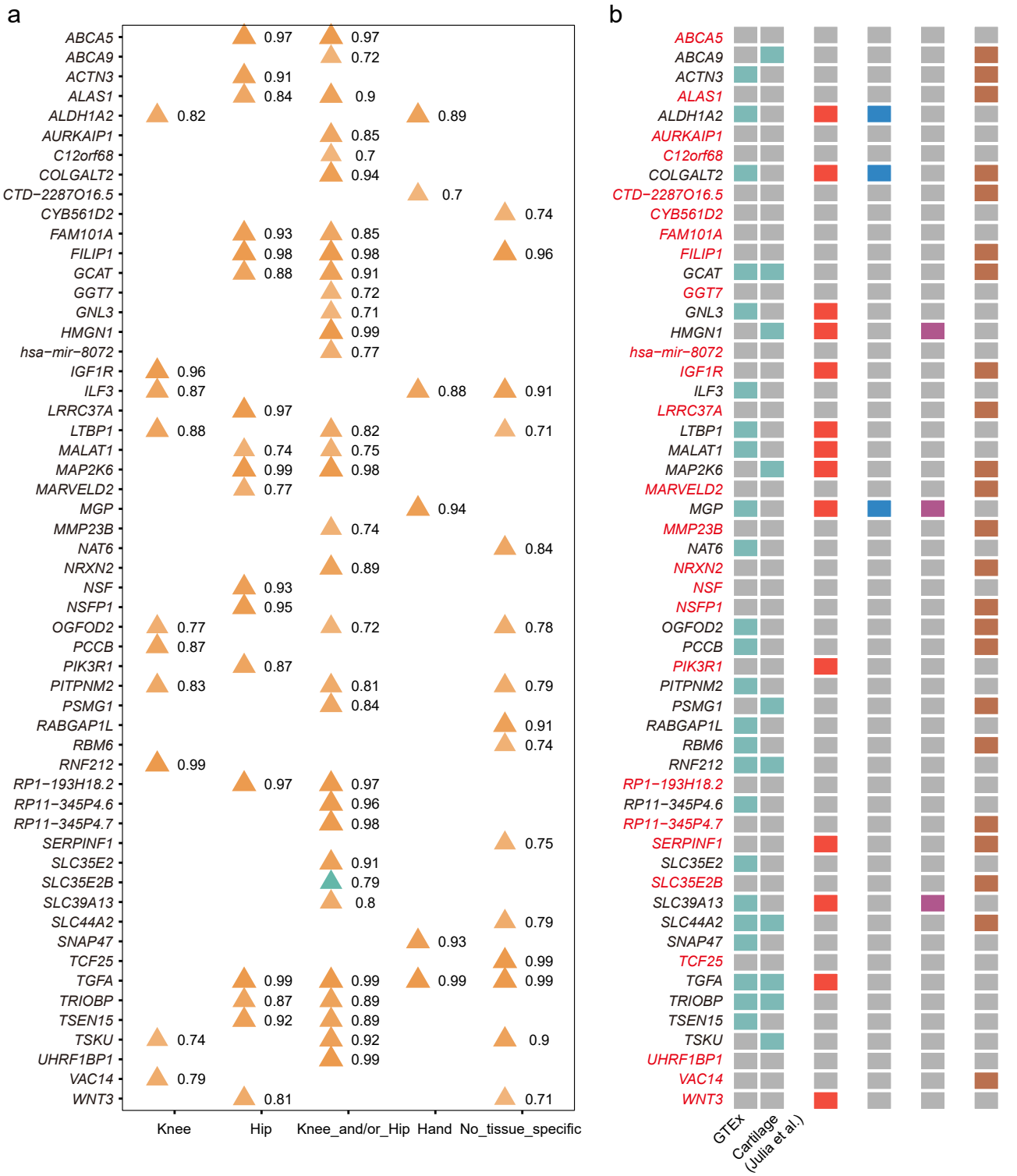

Fig. 6: Allele-specific open chromatin (ASoC) variant links both to OA risk and cartilage gene expression

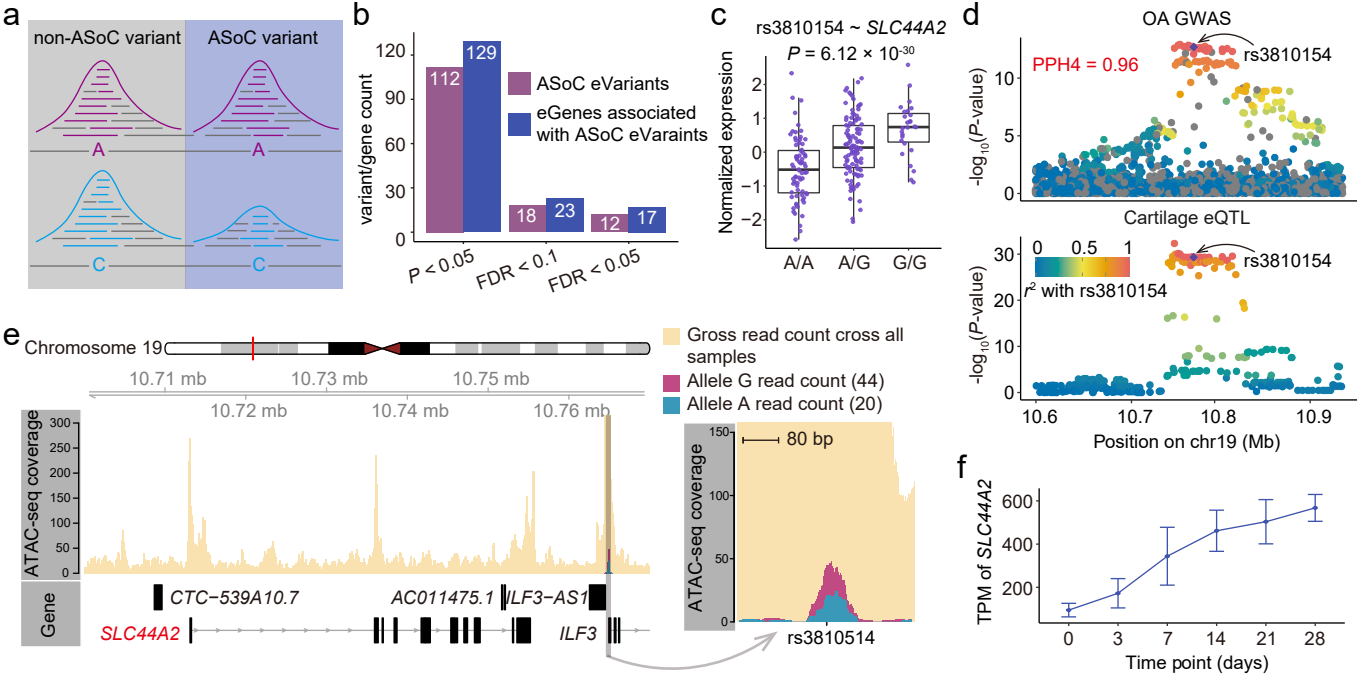
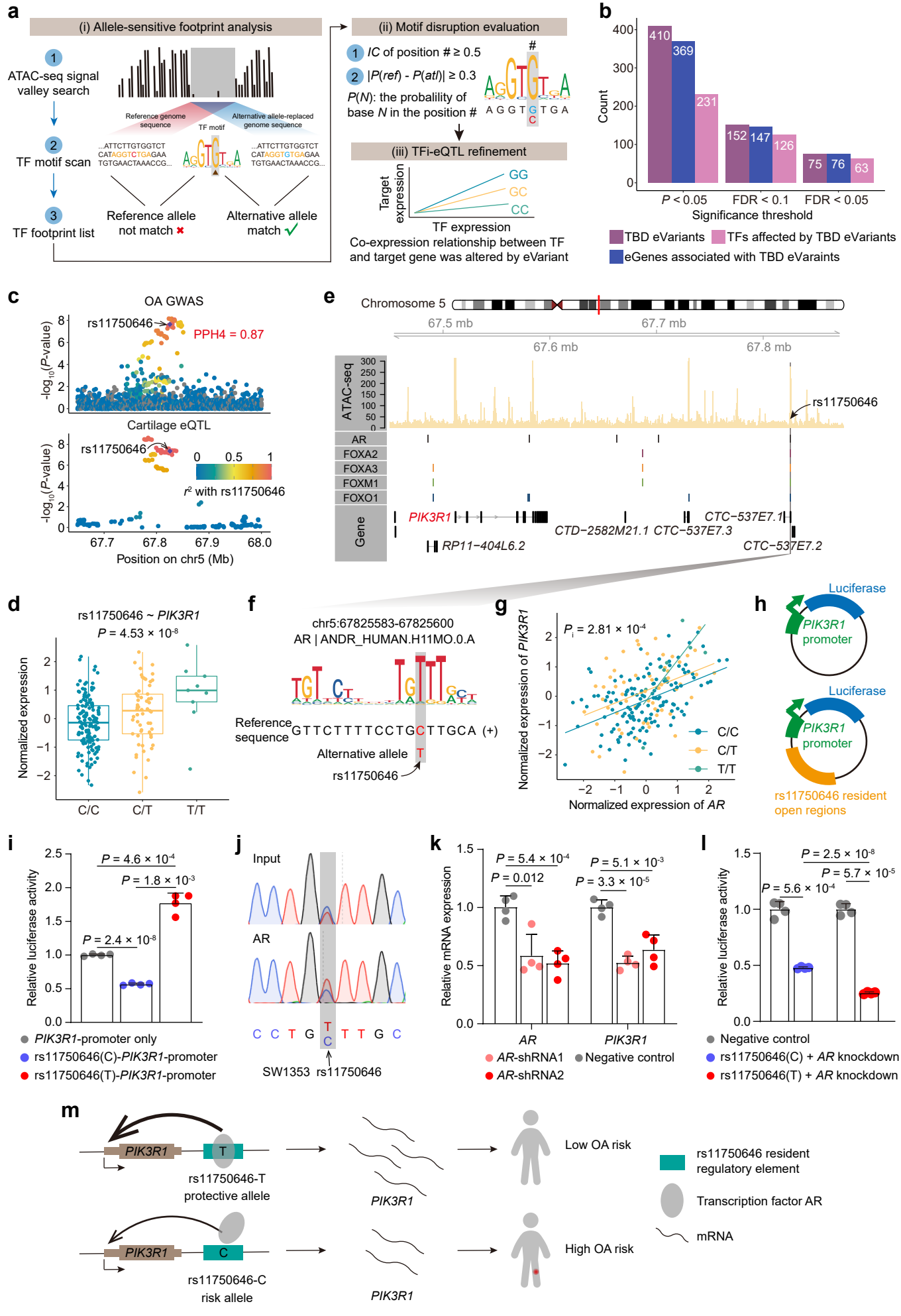
Fig. 7: TF binding-disruption (TBD) variants links both to OA risk and cartilage gene expression