Osteoarthritis and Cartilage 24 (2016) 655–663

Osteoarthritis and Cartilage

Targeted re-sequencing of linkage region on 2q21 identifies a novel functional variant for hip and knee osteoarthritis

M. Taipale † ‡, E. Jakkula † §, O.-P. Kämäräinen †, P. Gao †, S. Skarp † ‡, S. Barral ‖, I. Kiviranta ¶ #, H. Kröger †† ‡‡, J. Ott §§, G.-H. Wei †, L. Ala-Kokko ‖‖, M. Männikkö † ‡ *

† *Biocenter Oulu and Faculty of Biochemistry and Molecular Medicine, University of Oulu, Finland*
‡ *Center for Life Course Epidemiology and Systems Medicine, Faculty of Medicine, University of Oulu, Oulu, Finland*
§ *Public Health Genomics Unit, National Institute for Health and Welfare, Helsinki, Finland*
‖ *Gertrude H. Sergievsky Center, College for Physicians and Surgeons, Columbia University, New York, USA*
¶ *Department of Orthopaedics and Traumatology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland*
# *Department of Orthopaedics and Traumatology, Jyväskylä Central Hospital, Jyväskylä, Finland*
†† *Department of Orthopaedics and Traumatology, Kuopio University Hospital, Kuopio, Finland*
‡‡ *Bone and Cartilage Research Unit, University of Eastern Finland, Kuopio, Finland*
§§ *Institute of Psychology, Chinese Academy of Sciences, Beijing, China*
‖‖ *Connective Tissue Gene Tests, Allentown, PA, USA*

ARTICLE INFO

Article history:
Received 23 June 2015
Accepted 21 October 2015

Keywords:
Osteoarthritis
Hip
Knee
Linkage analysis
Next generation sequencing
Enhancer element

SUMMARY

Objective: The aim of the study was to identify genetic variants predisposing to primary hip and knee osteoarthritis (OA) in a sample of Finnish families.
Methods: Genome wide analysis was performed using 15 independent families (279 individuals) originating from Central Finland identified as having multiple individuals with primary hip and/or knee OA. Targeted re-sequencing was performed for three samples from one 33-member, four-generation family contributing most significantly to the LOD score. In addition, exome sequencing was performed in three family members from the same family.
Results: Genome wide linkage analysis identified a susceptibility locus on chromosome 2q21 with a multipoint LOD score of 3.91. Targeted re-sequencing and subsequent linkage analysis revealed a susceptibility insertion variant rs11446594. It locates in a predicted strong enhancer element region with maximum LOD score 3.42 under dominant model of inheritance. Insertion creates a recognition sequence for ELF3 and HMGA1 transcription factors. Their DNA-binding affinity is highly increased in the presence of A-allele compared to wild type null allele.
Conclusion: A potentially novel functional OA susceptibility variant was identified by targeted re-sequencing. This variant locates in a predicted regulatory site and creates a recognition sequence for ELF3 and HMGA1 transcription factors that are predicted to play a significant role in articular cartilage homeostasis.

Introduction

Epidemiological studies of family history and family clustering, twin studies, and exploration of rare genetic disorders related to osteoarthritis (OA) have shown that hereditary factors have a significant role in the development of this condition[1]. Estimated prevalence of hip and knee OA varies widely in epidemiological studies[2,3], but taken together, these conditions afflict several percent of working age and elderly men and women, causing also a substantial financial burden among all western populations.

Several genome wide linkage studies with families affected with hand and hip OA have been performed in order to find new susceptibility genes for OA. These scans have indicated suggestive linkage to most chromosomes, strengthening the hypothesis that the genetic etiology of this disease may be more complex than

* Address correspondence and reprint requests to: M. Männikkö, Center for Life Course Epidemiology and Systems Medicine, Faculty of Medicine, University of Oulu, Oulu, Finland, Aapistie 5, 90220 Oulu, Finland. Tel: 358-(0) 294-485751.
E-mail address: minna.mannikko@oulu.fi (M. Männikkö).

http://dx.doi.org/10.1016/j.joca.2015.10.019

previously thought. Also, studies concentrating primarily on hip OA families have revealed linkage to multiple chromosomal areas including 2q, 4q, 6p (women), 11q, 16p and 16q (women) (reviewed in[4]). In addition to numerous OA associating loci reported in larger population based studies, there are also previous reports of rare familial forms of OA where disease seems to be transmitted in families in a dominant manner[5,6].

Only large effect size alleles or variants can be detected by linkage analysis, while genome-wide association studies (GWAS) identify common small effect alleles. For different OA phenotypes several GWAS have been conducted[7–9]. It has become apparent that most of the OA associated alleles, such as the ones in *GDF5* and *DIO2*, contribute by modifying gene expression at transcriptional level[10,11]. However, only few genome-wide significant associations have been identified and they explain only a small proportion of the heritability. Recently, the focus in identifying disease causing sequence variants has been in the next-generation sequencing (NGS) technology, which allows identifying rare causal variants with a large effect size. NGS has successfully been used to identify causative variants in other diseases such as in prostate cancer[12] and schizophrenia[13].

Discovering new risk genes for genetically complex disease such as OA is challenging. This is partly due to the multiform etiology of the disease. No major disease causing variants have been identified so far, even though many variants in several genes have been implicated through previous linkage scans and following association analyses. Here we have conducted a family-based linkage analysis with subsequent targeted re-sequencing of the linkage region. To restrict the heterogeneity we have focused on 15 Finnish families with early-onset primary OA.

Patients and methods

Families

Initially ten independent families with early-onset OA were recruited for the genome wide linkage analysis. Additional five families were collected for the fine mapping, making the total number of studied families fifteen. Probands of each family were identified through the patient registers of Jyväskylä and Kuopio central hospitals, both located in Central Finland. All living family members of each proband were contacted and interviewed for the study. Ten families that were used for the whole genome scan included a total of 225 individuals, with the pedigree size ranging from 11 to 37 individuals. The inclusion criterion for the genome wide analysis was the presence of at least two affected individuals in each family who had hip and/or knee OA. The additional five families included a total of 54 individuals, with the pedigree size ranging from 8 to 14. At least three family members in each family were required to have hip and/or knee OA diagnosis while at least two of the affected individuals had to be first degree relatives. Taken together, 279 individuals in 15 families were analyzed, including 58 individuals with OA diagnoses of the hip and/or knee and 34 subjects that were considered healthy. OA status of 185 subjects was considered unknown, as described in detail later. The study was approved by local ethical committees and all subjects signed an informed written consent.

Clinical and radiological assessment

Probands in each family had gone through or were currently waiting for endoprosthetic hip and/or knee replacement due to OA. Their living adult relatives were first interviewed personally followed by a questionnaire in order to obtain adequate information about individual medical histories, possible OA diagnosis and former as well as current joint symptoms. Subjects with a history of a significant hip or knee joint trauma or obesity were excluded from the analyses or their OA status were defined as unknown to increase the power to detect actual genetic risk factors. Similarly families and individuals with a history, clinical findings or symptoms of any inflammatory joint disease, such as rheumatoid arthritis (RA) or any other local or systemic rheumatic disease/condition were excluded. All family members reporting joint symptoms in hip and/or knee joints were clinically evaluated for the presence of OA. Nearly all of the OA affected subjects had gone through or were currently waiting for hip and/or knee joint replacement. Standard radiological examination of symptomatic hip or knee joint(s) was performed and diagnosis of OA in this study was based on two main criteria: presence of common OA symptoms and typical radiological findings associated with OA observed in the standard joint X-ray projections. Common symptoms included significant joint pain (in movement and/or at rest), stiffness and possibly a decreased ability to move and use the affected joint. Diagnosis was confirmed if radiological evaluation of symptomatic joints showed at least moderately severe narrowing of joint space and one or more of the other OA specific changes: subchondral bone sclerosis, osteophytes or joint deformity. Even though standard clinical evaluation was performed, no systematic grading of the level of radiologically diagnosed OA with Kellgren–Lawrence or other scales was made. Studied subjects were considered healthy if they had never suffered from any prolonged joint symptoms concerning hip or knee joints; and in these cases, no radiological imaging of the hip or knee joints was systematically made. Pure radiological OA (rOA) without any symptoms was thus not considered a relevant variable in this study. Borderline cases with atypical symptoms or inadequate radiological findings were not included in either group in the analysis. A detailed overview of the characteristics of this study sample is presented in Table I.

Genotyping for linkage analysis

Genomic DNA was extracted from the white blood cells of the subjects using standard protocols. Genotyping for genome wide linkage scan was performed at the Finnish Genome Center (Helsinki, Finland) using the Applied Biosystems Linkage Mapping Set (MD 10) and an automated instrument (Megabase 1000; Molecular Dynamics).

In the first fine mapping additional markers (http://www.ncbi.nlm.nih.gov) spaced 2–5 cM apart were selected for the regions of interest on chromosomes 2 and 11 (10 and 7 markers, respectively). The genotyping was performed using standard fluorescence-based genotyping methodologies (ABI PRISM 3100 Genetic Analyzer, Applied Biosystems). After denaturation with formamide at 90°C for 2 min, products were separated by size and were detected using ABI PRISM 3100 genetic analyzer (Applied Biosystems). In the second fine mapping, 15 additional markers from chromosome 2 and 10 markers from chromosome 11 were genotyped to gain average ~1 cM spacing between markers. Genotyping was performed using GeneMapper Software version 4.0 (Applied Biosystems).

Statistics in linkage analysis

All genotypes were checked for Mendelian incompatibilities using PedCheck 1.1[14]. Allele frequencies were estimated at each locus from the data using observed and reconstructed genotypes of founders within the pedigrees. Age dependent penetrance models were used in linkage analysis (available on request). Both parametric and non-parametric multipoint LOD scores were calculated with Simwalk2 version 2.91[15]. Two-point LOD scores were

Table I
Characteristics of the studied families with the definitions of the OA outcomes

Family	*n*	Male	Female	Hip OA	Knee OA	Knee & Hip OA	Healthy	Unknown
1*	22	11	11	0	1	0	4	17
2	11	4	7	0	3	0	2	6
3	20	10	10	0	3	1	5	11
4	29	14	15	2	0	1	3	23
5	14	8	6	1	1	2	2	8
6	31	13	18	1	2	2	2	24
7	14	6	8	0	3	0	1	10
8	14	8	6	3	0	1	1	9
9	37	16	21	0	2	2	3	30
10	33	17	16	5	1	4	6	17
11	14	8	6	3	2	0	1	8
12	13	9	4	5	0	0	1	7
13	10	6	4	2	0	1	0	7
14	8	4	4	2	0	1	2	3
15	9	3	6	1	2	0	1	5
All	279	137	142	25	20	15	34	185

* Three affected individuals were identified in family 1, but DNA was available only from one.

calculated using the MLINK of the LINKAGE program package version 5.2[16]. The disease-allele frequency was set at 0.001. Genetic distances (cM) and microsatellite marker locations were specified according to database of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

Targeted re-sequencing and variant detection

One healthy (II-6) and two affected family members (II-9, III-26) of the same family [Fig. 1] were chosen for targeted re-sequencing. The entire 15.5 Mb segment of chromosome 2 containing the linkage peak and neighboring regions (126216105–141687523 GrCh37) was targeted by NimbleGen Sequence Capture 2.1M array (Roche) and the sequencing was performed by The Genome Analyzer IIx (Illumina). Alignment of sequence reads to reference and visualization, and variant detection, were done by a bioinformatics pipeline at the Institute of Molecular Medicine Finland (FIMM)[17].

Variant annotation and filtering

ANNOVAR software[18] and SNPnexus program (http://snp-nexus.org/)[19] was used for variant annotation and filtering. A custom pipeline with following criteria was used: (1) *in silico* pathogenicity using UCSC RefGene database to annotate variants as nonsense, splice, missense, synonymous, UTR, or noncoding was estimated; (2) variants with minor allele frequency (MAF) > 5% in the 1000 Genomes Project April 2012 release were removed; (3) the dbSNP version 137 was used in search of known SNPs, however, it was not used for filtering, since the database may include disease causing variants; (4) functional effects with whole-exome SIFT scores (version 2), whole-exome PolyPhen scores, whole-exome PolyPhen 2 scores built on HumanDiv database (for complex phenotypes), whole-exome MutationTaster scores (version 2) were predicted; (5) conserved regions and transcription factor binding sites using whole-exome GERP++ scores and phastCons 46-way alignments were screened, respectively; (6) non-coding variants that disrupt enhancers, repressor or promoter regions were identified using HMM (Hidden Markov models) predictions for skeletal muscle myoblasts (HSMM) and B-lymphocyte (GM12878) cell lines[20]; (7) DNase I hypersensitivity regions and segmental duplications were estimated.

Exome sequencing and variant detection

In addition to the targeted re-sequencing three affected individuals from family 10 were selected for exome sequencing in Beijing Genome Institute, Hong Kong. For exome capture and target sequencing SureSelect 51M Capture Kit (Agilent Technologies) and IlluminaHiSeq2000 100PE Platform were used, respectively. The paired-end sequence reads were aligned with Burrows-Wheeler Alignment tool[21]. Initial variant calling was performed using Genome Analysis ToolKit (GATK)[22]. Computational analysis of the whole exome data was done using the 64-bit Red Hat Enterprise Linux Server release 6/Hippu, IT Center for Science Ltd (CSC). Variants were filtered by GATK and single nucleotide variants (SNVs) with mapping quality (MQ) < 40, Haplotype score > 13 and genotype quality (GQ) < 20 were excluded. Small insertions or deletions

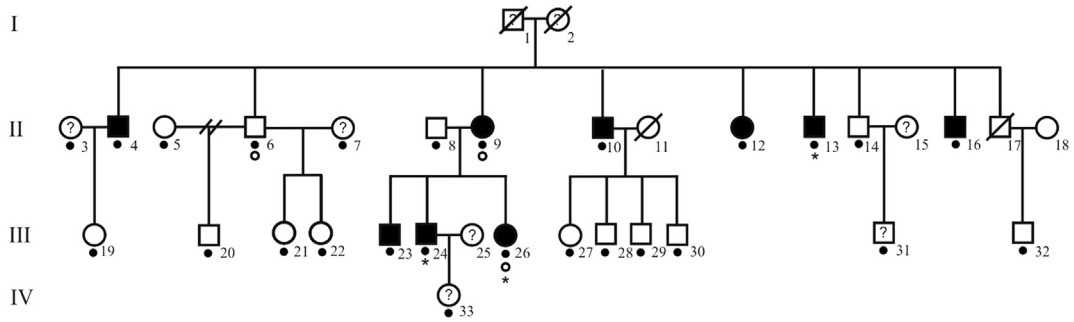

Fig. 1. Pedigree structure of family 10. A black dot beneath an individual indicates that DNA was available for genetic analysis. Individuals marked with question marks have unknown clinical status. Asterisk denotes individuals that were exome sequenced and open circles targeted re-sequenced individuals.

(indels) with quality by depth (QD) < 2.0, fisher strand >200.0 and Read Position Rank Sum Test < −20.0 were excluded. Comparison of the exome data between individuals was done using BEDTools[23]. Annotation of variants with ANNOVAR software was performed as described previously.

Variant validation for Pseudomarker analysis

Altogether 32 variants that located within regulatory regions (n = 20, targeted re-sequencing) or gained harmful prediction (n = 12, exome sequencing) were genotyped in family 10 (Suppl. Table I). Genotyping was performed by Sanger sequencing using ABI PRISM 3500xl Genetic Analyzer (Applied Biosystems) and Variant Reporter Software 2 (Applied Biosystems). Primer sequences are available on request. Pseudomarker version 2.0 was used to evaluate the evidence for linkage and association in the candidate region pointed out by the whole genome linkage analysis. Pseudomarker is able to analyze linkage and association both separately and jointly and analyze different pedigree structures, trios, cases and controls[24,25]. Recessive, dominant and model based models of Pseudomarker were used in analysis. Two variants, rs200661871 and rs11446594, with highest LOD scores, were genotyped among all 15 families.

Bioinformatic analyses with Enhancer Element Locator

Enhancer Element Locator (EEL)[26] software was used to predict weather rs200661871 and rs11446594 directly affect DNA-binding motifs for certain transcription factors. Surrounding sequences of rs200661871 and rs11446594 were analyzed with DNA-binding specificity data set for human transcription factors[27].

Results

Clinical and radiological findings

In fifteen families, 25 hip OA, 20 knee OA and 15 subjects with OA in both joints were identified from a total of 279 studied subjects. The average age at the disease onset in this study was 50 years. Thirty-four subjects were classified as healthy and 185 individuals as unknown in the analyses. The pedigree of the largest family (family 10) is presented in Fig. 1.

Genome-wide linkage analysis

In the whole genome scan, a total of ten hip and knee OA families were analyzed. Two-point linkage analysis identified four interesting loci on chromosomes 2, 11, 13 and 20 with logarithm of odds (LOD) scores above 1.5. In multipoint analysis a maximum overall LOD score of 2.14 occurred on chromosome 2 between markers D2S112 and D2S142 indicating the same region as in the two point analysis. The second highest peak was detected on chromosome 11 with a maximum multipoint LOD score of 1.82. There were no peaks exceeding the threshold of 1.5 for the multipoint analysis in other chromosomes (data not shown).

The two most interesting loci on chromosomes 2 and 11 with LOD scores over 1.5 in both two-point and multipoint analyses, were further analyzed by fine mapping, including five additional families. After the second fine mapping only one family (Family 10, Fig. 1) contributed significantly to the LOD score, whereas other families contributed only slightly. The fine mapping confirmed the susceptibility locus on chromosome 2 with a multipoint LOD score of 3.91 under a recessive model of inheritance. The same region was indicated in two-point linkage analysis resulting in a maximum two-point LOD score of 3.63 at a recombination fraction of 0.00 for locus D2S1260 [Fig. 2]. The linkage signal for chromosome 11 disappeared after fine mapping. No statistically significant results were obtained using dominant model of inheritance in either chromosomal locus.

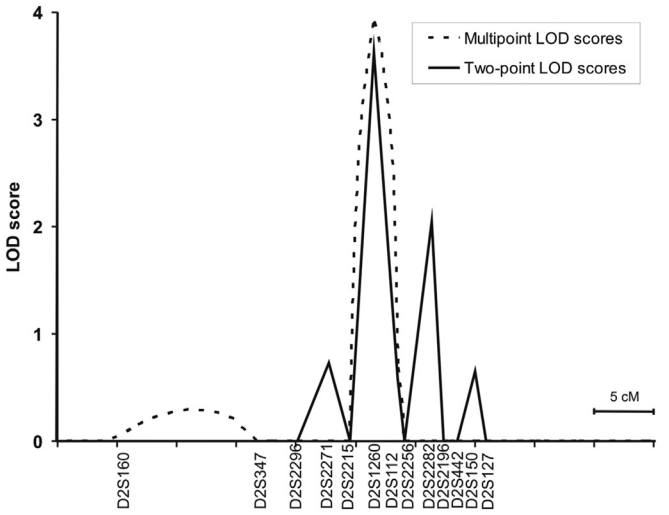

Fig. 2. Multipoint and two-point LOD scores from the second fine mapping on chromosome 2 under a recessive mode of inheritance.

Targeted re-sequencing and prioritized candidate variants

The 15.5 Mb region identified in linkage analysis on chromosome 2 was re-sequenced using NGS technology. The average 20-fold coverage for target region was >89%. Sequencing resulted in 26,078 high quality variant calls (reference sequence GRCh37). Two affected family members shared 4006 variant calls. Out of these 4006 variant calls 475 had MAF ≤ 0.05 or the frequency was unknown according to 1000 Genomes Project 2012 April release on European subjects, and 63 (1.6%) were not previously annotated in dbSNP (http://www.ncbi.nlm.nih.gov/snp). Only 16 were exonic but they all were synonymous or benign according to computational pathogenicity estimates as described above. Seventeen variants were identified to target enhancers, repressor or promoter regions according to HMM predictions for HSMM or GM12878 cell lines. In addition, targeted sequencing identified 659 indels, which were shared by the two affected family members and were not found in the healthy relative. All the indels were intronic or intergenic and three of them with MAF ≤ 0.05 or unknown frequency located in predicted regulatory regions according to HMM predictions for HSMM or GM12878 cell lines. General statistics of the targeted re-sequencing and variant prioritization are shown in Table II.

Exome sequencing and prioritized candidate variants

Variant calling resulted to an average of 77,755 SNVs and 8852 indels. Average sequencing depth in target area was 67X and the average coverage of target region was 92%. Filtering by quality resulted in average of 70,333 SNVs and 8776 indels. Comparison by BEDTools showed that the three affected family members shared 39,039 SNVs and 4797 indels. Functional annotation by ANNOVAR revealed five SNVs and one indel with MAF ≤ 0.01 or private exonic variants with pathogenic annotations. In addition six rare or private SNVs were found on active promoter sites or strong enhancer areas.

Table II
General statistics of the sequencing run and variant annotation of the targeted re-sequencing and exome sequencing

Coverage statistics		Variant statistics	
Targeted re-sequencing			
Total number of mapped bases	11,267 512	Variants detected	26,078
Target 20-fod coverage		SNVs shared by the affected	4006
III-26	88.26%	MAF < 0.05	475
II-9	91.12%	Novel variants	63
II-6	88.26%	Regulatory region	17
		Indels shared by affected	659
		Regulatory region	3
Exome sequencing			
The average coverage of target region		SNVs shared by the affected	39,039
III-26	91,84%	MAF < 0.01	94
IV-33	92,12%	Novel variants	212
II-13	92,42%	Harmful	11
		Indels shared by the affected	4797
		MAF < 0.01	19
		Novel variants	326
		Harmful	1

Numbering of the individuals is according to Figure 1.

General statistics of the exome sequencing and variant prioritization are shown in Table II.

Pseudomarker and EEL analysis

None of the selected 12 variants with harmful prediction from exome sequencing were found in all affected family members of family 10, and thus were not analyzed further (data not shown). Predicted regulatory region variants from targeted re-sequencing were analyzed using Pseudomarker analysis software. The results are shown in Table III. Moderate evidence for linkage was found for two heterozygous indels rs200661871 and rs11446594 located at 132127004 and 134570742 (hg19), respectively, with maximum LOD score 2.34 under a dominant model of inheritance. According to phastCons 46-way alignments and chromHMM predictions of ANNOVAR program rs200661871 locates on a transcription factor binding site and rs11446594 interrupts a strong enhancer element. Furthermore, we conducted bioinformatics analysis using EEL algorithm[26], and found that rs11446594 is a motif-disruptor for the transcription factors ELF3 and HMGA1. EEL prediction also showed that ELF3 and HMGA1 indicate highly increased DNA-binding affinity to the rs11446594 A allele than null allele, suggesting that the variant A at rs11446594 may create a recognition DNA sequence for these transcription factors. These two variants were genotyped in all 15 families (Table IV). In the following Pseudomarker analysis insertion rs11446594 showed strong linkage (LOD score 3.42) under a dominant model of inheritance, but only nominal evidence of association ($P = 0.02$) to OA. Insertion rs200661871 indicated some evidence of linkage (LOD score 1.62), but did not show significant association ($P = 0.92$) when studied in all families.

Discussion

In the present study linkage analysis and targeted re-sequencing identified a novel susceptibility locus for OA on chromosome 2q21 with two candidate variants, insertions rs11446594 and rs200661871. Chromosome 2 has been among the most studied chromosomes that are likely to harbor OA susceptibility genes. Several different studies have indicated linkage to chromosome 2 (reviewed in[4,28]). However, the OA susceptibility locus identified in this study does not overlap with these earlier findings. The highest LOD score 3.91 in genome wide linkage analysis indicating strong linkage to OA was achieved using a recessive model of inheritance. However, the highest LOD score 3.42 in targeted linkage analysis using Pseudomarker software was achieved using a dominant model of inheritance, and all the affected family members in four-generation family 10 were heterozygous for rs11446594 and

Table III
Results of the Pseudomarker single-point linkage analysis

Position (hg19)	snp137	MAF	Gene	Location	Recessive LOD score		Dominant LOD score	
					Model-free	Model based*	Model-free	Model based
127166786	rs191705727	0.004	CNTNAP5, GYPC	Intergenic, GERP score† 3.54	0.558	0.551	0.067	0.064
127414803	rs115260878	0.03	GYPC	Intronic, strong enhancer, active promoter	0.558	0.551	0.067	0.064
127415578	rs112991427	0.03	GYPC	Intronic, active promoter	0.557	0.551	0.065	0.062
127417098	rs116012614	0.03	GYPC	Intronic, strong enhancer	0.558	0.551	0.067	0.064
127420383	rs112745348	0.04	GYPC	Intronic, strong enhancer	0.558	0.551	0.067	0.064
127832275	rs138530867	NA§	BIN1	Intronic, strong enhancer	0.023	0.026	0.000	0.000
127832682	rs141817721	0.003	BIN1	Intronic, strong enhancer	0.558	0.551	0.067	0.064
128146026	rs192154491	0.01	MAP3K2, PROC	Intergenic, active promoter	0.558	0.551	0.067	0.064
129864234	rs141185315	0.05	HS6ST1, LOC389033	Intergenic, weak enhancer	0.750	0.751	0.312	0.367
130691676	NA	NA	LOC389033	Intronic, active promoter	0.558	0.551	0.067	0.064
130939163	rs146359515	0.01	SMPD4	Exonic, active promoter, tfbs score‡ 885	0.558	0.551	0.067	0.064
132127004	rs200661871	NA	WTH3DI, LINC01120	Intergenic, tfbs score 793	1.806	1.707	2.346	1.868
134570742	rs11446594	0.38*	NCKAP5, MIR3679	Intergenic, strong enhancer	1.806	1.707	2.341	1.907
134575035	rs116663953	0.02	NCKAP5, MIR3679	Intergenic, weak enhancer	1.806	1.705	0.342	0.269
137134616	rs148906434	0.01	CXCR4, THSD7B	Intergenic, weak enhancer	1.806	1.705	0.342	0.269
137986076	rs115528720	0.03	THSD7B	Intronic, weak enhancer	1.806	1.705	0.343	0.368
138709574	rs186666211	0.004	THSD7B, HNMT	Intergenic, weak enhancer	1.806	1.705	0.343	0.368
138714163	NA	NA	THSD7B, HNMT	Intergenic, strong enhancer	1.806	1.705	0.343	0.368
138723086	rs145115612	0.0040	HNMT	Intronic, weak promoter	1.806	1.705	0.335	0.472
141311237	rs143690440	0.01	LRP1B	Intronic, tfbs score 804	1.806	1.705	0.988	1.049

*1000 Genomes Project October 2014 release on European subjects. At the time of Annovar analysis, MAF of variant was unknown according to 1000 Genomes Project April 2012 release on European subjects.
* Model based analyses include liability classes.
† GERP = the genomic evolutionary rate profiling.
‡ tfbs = transcription factor binding site.
§ NA = not available.

Table IV
Results of the Pseudomarker association analysis

Position (hg19)	snp137	MAF	Gene	Location	LOD score	*P* value*
132127004	rs200661871	NA	*WTH3DI. LINC01120*	Intergenic	1.62	0.92
134570742	rs11446594	NA	*NCKAP5. MIR3679*	Intergenic	3.42	0.02

*non-corrected.

rs200661871 supporting the fact that the observed results are true rather than false-positive[29]. These results demonstrate the complex inheritance pattern of OA with multiple risk loci and some of them acting in a dominant-like fashion and others being more recessive-like. A non-Mendelian inheritance pattern has been confirmed in several studies (reviewed in Peach *et al.* 2005)[1].

Variant rs11446594 resides between the genes coding NCK-associated protein 5 (*NCKAP5)* and microRNA 3679 (*MIR3679)*, and based on UCSC genome browser is located within a histone H3K27ac and DNase I hypersensitivity site (ENCODE) [Fig. 3]. H3K27ac separates active enhancers from poised/inactive enhance elements[30] while DNase I hypersensitivity sites are chromosomal areas that associate with active gene loci and regulatory elements[31]. Enhancer elements are short *cis*-regulatory regions of DNA sequence which bind *trans*-acting transcription factors to enhance gene transcription (reviewed in Maston *et al.* 2006)[32]. *Cis*-regulatory elements may lay hundreds or thousands of kilobases away from target genes and regulate transcription over a long physical distance. It has been suggested that small indels or point mutations in *cis*-regulatory elements may be the cause of the disease, for example in cardiovascular disease[33]. In addition, SNPs associated with complex diseases such as celiac disease and type 1 diabetes have been demonstrated to affect genes in *trans*[34]. Compared to protein-coding mutations changes in the non-coding *cis*- and *trans*-regulatory regions have minor effects and they are thus able to cluster in populations, and therefore predispose to common complex diseases[35].

According to EEL algorithm variant rs11446594 showed increased DNA-binding affinity to ELF3 and HMGA1 transcription factors. The exact target gene(s) regulated by this binding is not yet known, but both factors have an important catabolic role in cartilage homeostasis. ELF3, also known as ESE1, belongs to ETS-domain family of transcription factors, which are phosphorylated trans-acting proteins regulating epithelial cell differentiation, gut development and apoptosis[36]. In OA cartilage, enhanced stimulation of ELF3 has been shown to lead to increased expression of MMP13 (matrix metalloproteinase 13)[36], the major enzyme responsible for degradation of cartilage matrix collagen[37]. HMGA1 (high-mobility group A1) is a small nuclear protein that regulates for example transcription, embryogenesis, cell cycle and DNA repair[38]. HMGA1 expression is increased in OA and is required for full activation of IGFBP-3 (Insulin-like growth factor-binding protein 3). Activated IGFBP-3 further blocks IGF-1 (insulin growth factor 1) binding to the IGF receptor and thus decreasing the amount of effective IGF-1

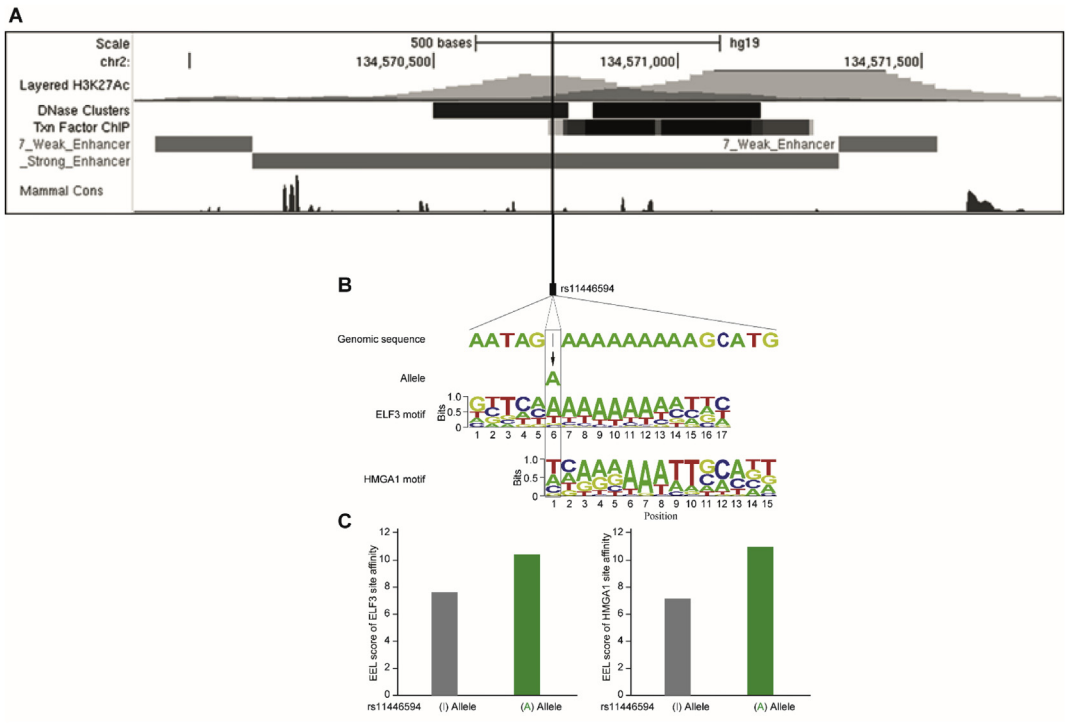

Fig. 3. (A) A shot from UCSC genome browser to illustrate the location of non-coding insertion rs11446594 in relation to predicted regulatory regions in HSMM cell line. The location of rs11446594 is highlighted by vertical line. It is located on strong enhancer element predicted by chromHMM. (B) Genomic sequence surrounding the SNP rs11446594 is aligned to a LOGO graphic representation of ELF3 and HMGA1 transcription factor DNA-binding motifs. Note that rs11446594 physically maps to an ELF3 and HMGA1 binding site, respectively, as shown in the positional weight matrixes (PWMs) of ELF3 and HMGA1 DNA-binding motif generated by protein binding microarray[27]. (C) The affinity of ELF3 and HMGA1 binding to rs11446594 A and null alleles is predicted by EEL algorithm[26]. Note that the variant A allele at rs11446594 may create a recognition sequence for Elf3 and HMGA1 transcription factors, respectively.

enabling to induce the expression of collagens and proteoglycans[39]. ELF3 and HMGA1 transcription factors are important regulators of articular cartilage homeostasis, and may regulate several OA related genes in addition to MMP13 and IGFBP-3. Further studies are needed to unravel the mechanism how binding of rs11446594 affects the activity of these transcription factors and further their target genes.

At the time of discovery the MAF of rs11446597 was unknown according to 1000 Genomes Project April 2012 release on European subjects. However, according to updated database of 1000 Genomes Project October 2014 release on European subjects, the MAF of rs11446597 is 0.38 and, this insertion appears to be a relatively common variant. It has been suggested that there are multiple common and small affecting variants which together predispose to OA[40] and common SNPs with functional evidence have previously been associated with OA[10].

The second variant rs200661871 locates between a pseudogene *WTH3D* and the long intergenic non-protein coding RNA 1120 gene *(LINC01120)*, and resides in putative transcription factor binding site. According to the SNP-nexus program rs200661871 was predicted to locate near OCT-1 transcription factor binding site. OCT-1 is involved in the synovial cell activity of RA patients[41]. These two variants or nearby genes have not been previously associated with OA nor have known cartilage related functions. In general, only little is known about these genes. Previously variants in the *NCKAP* gene have been associated with other complex diseases such as bipolar disorder[42].There is no previous information about MIR3679 but variants in miRNA binding regions have been associated with complex common diseases. A link between genetic risk factors and miRNA function has previously been shown in lumbar disc degeneration, where the level of target gene expression was reduced in susceptibility allele carriers[43]. Based on *in silico* bioinformatic analyses variants rs11446594 and rs200661871 may affect the regulation of genes responsible for OA. However, it is possible that these genes are not neighboring ones but other genes residing in a greater distance.

The low number of sequenced family members might have been a limitation to the study, but earlier studies on Mendelian diseases have shown that it is possible to identify a causative variant using linkage analysis followed by NGS of only few affected family members[44,45]. In complex diseases such as OA finding the disease causing variant is likely to be more complicated. However, NGS has been successfully used in identification of causative variants in complex diseases including prostate cancer[12] and schizophrenia[13]. One could also criticize the fact, that hip and knee OA cases are pooled in this analysis. It has been suggested that there are joint specificity in the OA process since several genes have shown opposed changes in gene expressions in different joints[46]. However, certain sequence variants have been associated not only with OA of different joints but in different degenerative diseases. For example a variant in the growth differentiation factor 5 gene (*GDF5*) has been associated with both OA and disc degeneration[47,48]. Separating the two outcomes would have diminished the statistical linkage power in this analysis. Moreover, we were not able to predict regulatory regions using data from chondrocytic cell lines thus data from HSMM and GM12878 cell lines were used instead. This may affect the predictions given the fact that transcription factor binding patterns are heterogeneous and differ between tissue types[49]. In addition, to best of our knowledge the identified variants have not been previously reported to associate with OA or any other disease in GWA studies[50]. Finally, the *P*-values obtained in the association analyses were modest, and hence they would not have survived correction for multiple comparisons.

In summary, we have identified two novel susceptibility variants (rs11446594 and rs200661871) for OA on chromosome 2q21, in Finnish families. The insertion rs11446594 with strong linkage is predicted to reside within strong enhancer element between genes *NCKAP5* and *MIR3679*, and creates a putative recognition sequence for ELF3 and HMGA1 transcription factors. These transcription factors are predicted to play a significant role in articular cartilage homeostasis. The second insertion rs200661871 with modest linkage locates between the *WTH3D* and *LINC01120* genes on an OCT-1 transcription factor binding site. Our finding strengthens the assumption that chromosome 2 comprises multiple OA associated variants. Further studies based on our findings may lead to identification of new biological networks and pathways behind the OA.

Author contributions

MT = Design of the study, data preparation, statistical analysis, interpretation of the findings and drafting the manuscript.

EJ = Design of the study, data collection, statistical analysis and interpretation of the findings.

OPK = Data collection, interpretation of the findings and drafting the manuscript.

PG = Statistical analysis and interpretation of the findings.

SS = Data preparation and statistical analysis.

SB = Statistical analysis and interpretation of the findings.

IK = Data collection and design of the study.

HK = Data collection and design of the study.

JO = Design of the study.

GHW = Statistical analysis, interpretation of the findings and drafting the manuscript.

LAK = Data collection and design of the study.

MM = Design of the study, interpretation of the findings and drafting the manuscript.

All authors reviewed the draft manuscript and approved the final version. MT (mari.taipale@oulu.fi) and MM (minna.mannikko@oulu.fi) take responsibility for the integrity of the work as a whole.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments and funding

We would like to thank all the patients and their family members for participating in the study. We thank Tero Hiekkalinna, Ph.D., for his valuable help and advice in Pseudomarker analyses. Laboratory technicians Anu Myllymäki and Aira Erkkilä are thanked for their excellent technical assistance. Partial support by China NSFC grant 3147 0070 (JO) is gratefully acknowledged.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.joca.2015.10.019.

References

1. Peach CA, Carr AJ, Loughlin J. Recent advances in the genetic investigation of osteoarthritis. Trends Mol Med 2005;11:186–91, http://dx.doi.org/10.1016/j.molmed.2005.02.005.
2. Jordan JM, Helmick CG, Renner JB, Luta G, Dragomir AD, Woodard J, *et al.* Prevalence of knee symptoms and radiographic and symptomatic knee osteoarthritis in African Americans and Caucasians: the Johnston County Osteoarthritis Project. J Rheumatol 2007;34:172–80.
3. Roux CH, Saraux A, Mazieres B, Pouchot J, Morvan J, Fautrel B, *et al.* Screening for hip and knee osteoarthritis in the general

population: predictive value of a questionnaire and prevalence estimates. Ann Rheum Dis 2008;67:1406–11, http://dx.doi.org/10.1136/ard.2007.075952.
4. Williams CJ. The genetics of osteoarthritis. Expert Rev Clin Immunol 2007;3:503–16, http://dx.doi.org/10.1586/1744666X.3.4.503.
5. Jakkula E, Melkoniemi M, Kiviranta I, Lohiniva J, Räinä SS, Perälä M, *et al*. The role of sequence variations within the genes encoding collagen II, IX and XI in non-syndromic, early-onset osteoarthritis. Osteoarthritis Cartilage 2005;13:497–507, http://dx.doi.org/10.1016/j.joca.2005.02.005.
6. Mabuchi A, Nakamura S, Takatori Y, Ikegawa S. Familial osteoarthritis of the hip joint associated with acetabular dysplasia maps to chromosome 13q. Am J Hum Genet 2006;79:163–8, http://dx.doi.org/10.1086/505088.
7. arcOGEN ConsortiumarcOGEN Collaborators. Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. Lancet 2012;380:815–23, http://dx.doi.org/10.1016/S0140-6736(12)60681-3.
8. Evangelou E, Kerkhof HJ, Styrkarsdottir U, Ntzani EE, Bos SD, Esko T, *et al*. A meta-analysis of genome-wide association studies identifies novel variants associated with osteoarthritis of the hip. Ann Rheum Dis 2014;73:2130–6, http://dx.doi.org/10.1136/annrheumdis-2012-203114.
9. Styrkarsdottir U, Thorleifsson G, Helgadottir HT, Bomer N, Metrustry S, Bierma-Zeinstra S, *et al*. Severe osteoarthritis of the hand associates with common variants within the ALDH1A2 gene and with rare variants at 1p31. Nat Genet 2014;46:498–502, http://dx.doi.org/10.1038/ng.2957.
10. Miyamoto Y, Mabuchi A, Shi D, Kubo T, Takatori Y, Saito S, *et al*. A functional polymorphism in the 5' UTR of GDF5 is associated with susceptibility to osteoarthritis. Nat Genet 2007;39:529–33, http://dx.doi.org/10.1038/2005.
11. Bos SD, Bovee JV, Duijnisveld BJ, Raine EV, van Dalen WJ, Ramos YF, *et al*. Increased type II deiodinase protein in OA-affected cartilage and allelic imbalance of OA risk polymorphism rs225014 at DIO2 in human OA joint tissues. Ann Rheum Dis 2012;71:1254–8, http://dx.doi.org/10.1136/annrheumdis-2011-200981.
12. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, *et al*. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet 2012;44:685–9, http://dx.doi.org/10.1038/ng.2279.
13. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, *et al*. Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nat Genet 2011;43:864–8, http://dx.doi.org/10.1038/ng.902.
14. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 1998;63:259–66, http://dx.doi.org/10.1086/301904.
15. Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 1996;58:1323–37.
16. Lathrop GM, Lalouel JM, Julier C, Ott J. Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 1984;81:3443–6.
17. Sulonen AM, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, *et al*. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol 2011;12:R94, http://dx.doi.org/10.1186/gb-2011-12-9-r94.
18. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164, http://dx.doi.org/10.1093/nar/gkq603.
19. Dayem Ullah AZ, Lemoine NR, Chelala C. A practical guide for the functional annotation of genetic variations using SNPnexus. Brief Bioinform 2013;14:437–47, http://dx.doi.org/10.1093/bib/bbt004.
20. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, *et al*. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011;473:43–9, http://dx.doi.org/10.1038/nature09906.
21. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60, http://dx.doi.org/10.1093/bioinformatics/btp324.
22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al*. The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303, http://dx.doi.org/10.1101/gr.107524.110.
23. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2, http://dx.doi.org/10.1093/bioinformatics/btq033.
24. Hiekkalinna T, Schäffer AA, Lambert B, Norrgrann P, Göring HH, Terwilliger JD. PSEUDOMARKER: a powerful program for joint linkage and/or linkage disequilibrium analysis on mixtures of singletons and related individuals. Hum Hered 2011;71:256–66, http://dx.doi.org/10.1159/000329467.
25. Gertz EM, Hiekkalinna T, Digabel SL, Audet C, Terwilliger JD, Schaffer AA. PSEUDOMARKER 2.0: efficient computation of likelihoods using NOMAD. BMC Bioinformatics 2014;15(47), http://dx.doi.org/10.1186/1471-2105-15-47. 2105–15-47.
26. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, *et al*. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 2006;124:47–59, http://dx.doi.org/10.1016/j.cell.2005.10.042.
27. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, *et al*. Diversity and complexity in DNA recognition by transcription factors. Science 2009;324:1720–3, http://dx.doi.org/10.1126/science.1162327.
28. Valdes AM, Spector TD. The contribution of genes to osteoarthritis. Rheum Dis Clin North Am 2008;34:581–603, http://dx.doi.org/10.1016/j.rdc.2008.04.008.
29. Hiekkalinna T, Göring HH, Terwilliger JD. On the validity of the likelihood ratio test and consistency of resulting parameter estimates in joint linkage and linkage disequilibrium analysis under improperly specified parametric models. Ann Hum Genet 2012;76:63–73, http://dx.doi.org/10.1111/j.1469-1809.2011.00683.x.
30. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, *et al*. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA 2010;107:21931–6, http://dx.doi.org/10.1073/pnas.1016071107.
31. Gross DS, Garrard WT. Nuclease hypersensitive sites in chromatin. Annu Rev Biochem 1988;57:159–97, http://dx.doi.org/10.1146/annurev.bi.57.070188.001111.
32. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum.-Genet 2006;7:29–59, http://dx.doi.org/10.1146/annurev.genom.7.080505.115623.
33. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, *et al*. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 2010;466:714–9, http://dx.doi.org/10.1038/nature09266.
34. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, *et al*. Systematic identification of trans eQTLs as

putative drivers of known disease associations. Nat Genet 2013;45:1238–43, http://dx.doi.org/10.1038/ng.2756.
35. Sakabe NJ, Savic D, Nobrega MA. Transcriptional enhancers in development and disease. Genome Biol 2012;13(238), http://dx.doi.org/10.1186/gb-2012-13-1-238. 2012–13-1-238.
36. Otero M, Plumb DA, Tsuchimochi K, Dragomir CL, Hashimoto K, Peng H, *et al*. E74-like factor 3 (ELF3) impacts on matrix metalloproteinase 13 (MMP13) transcriptional control in articular chondrocytes under proinflammatory stress. J Biol Chem 2012;287:3559–72, http://dx.doi.org/10.1074/jbc.M111.265744.
37. Billinghurst RC, Dahlberg L, Ionescu M, Reiner A, Bourne R, Rorabeck C, *et al*. Enhanced cleavage of type II collagen by collagenases in osteoarthritic articular cartilage. J Clin Invest 1997;99:1534–45, http://dx.doi.org/10.1172/JCI119316.
38. Gasparini G, De Gori M, Paonessa F, Chiefari E, Brunetti A, Galasso O. Functional relationship between high mobility group A1 (HMGA1) protein and insulin-like growth factor-binding protein 3 (IGFBP-3) in human chondrocytes. Arthritis Res Ther 2012;14:R207, http://dx.doi.org/10.1186/ar4045.
39. Iwanaga H, Matsumoto T, Enomoto H, Okano K, Hishikawa Y, Shindo H, *et al*. Enhanced expression of insulin-like growth factor-binding proteins in human osteoarthritic cartilage detected by immunohistochemistry and in situ hybridization. Osteoarthritis Cartilage 2005;13:439–48. S1063-4584(04)00281-X.
40. Panoutsopoulou K, Southam L, Elliott KS, Wrayner N, Zhai G, Beazley C, *et al*. Insights into the genetic architecture of osteoarthritis from stage 1 of the arcOGEN study. Ann Rheum Dis 2011;70:864–7, http://dx.doi.org/10.1136/ard.2010.141473.
41. Wakisaka S, Suzuki N, Takeno M, Takeba Y, Nagafuchi H, Saito N, *et al*. Involvement of simultaneous multiple transcription factor expression, including cAMP responsive element binding protein and OCT-1, for synovial cell outgrowth in patients with rheumatoid arthritis. Ann Rheum Dis 1998;57:487–94.
42. Smith EN, Bloss CS, Badner JA, Barrett T, Belmonte PL, Berrettini W, *et al*. Genome-wide association study of bipolar disorder in European American and African American individuals. Mol Psychiatry 2009;14:755–63, http://dx.doi.org/10.1038/mp.2009.43.
43. Song YQ, Karasugi T, Cheung KM, Chiba K, Ho DW, Miyake A, *et al*. Lumbar disc degeneration is linked to a carbohydrate sulfotransferase 3 variant. J Clin Invest 2013;123:4909–17. 69277.
44. Sobreira NL, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, Stevens EL, *et al*. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS Genet 2010;6:e1000991, http://dx.doi.org/10.1371/journal.pgen.1000991.
45. Zhao Y, Zhao F, Zong L, Zhang P, Guan L, Zhang J, *et al*. Exome sequencing and linkage analysis identified tenascin-C (TNC) as a novel causative gene in nonsyndromic hearing loss. PLoS One 2013;8:e69549, http://dx.doi.org/10.1371/journal.pone.0069549.
46. Xu Y, Barter MJ, Swan DC, Rankin KS, Rowan AD, Santibanez-Koref M, *et al*. Identification of the pathogenic pathways in osteoarthritic hip cartilage: commonality and discord between hip and knee OA. Osteoarthritis Cartilage 2012;20:1029–38, http://dx.doi.org/10.1016/j.joca.2012.05.006.
47. Evangelou E, Chapman K, Meulenbelt I, Karassa FB, Loughlin J, Carr A, *et al*. Large-scale analysis of association between GDF5 and FRZB variants and osteoarthritis of the hip, knee, and hand. Arthritis Rheum 2009;60:1710–21, http://dx.doi.org/10.1002/art.24524.
48. Williams FM, Popham M, Hart DJ, de Schepper E, Bierma-Zeinstra S, Hofman A, *et al*. GDF5 single-nucleotide polymorphism rs143383 is associated with lumbar disc degeneration in Northern European women. Arthritis Rheum 2011;63:708–12, http://dx.doi.org/10.1002/art.30169.
49. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res 2012;22:1748–59, http://dx.doi.org/10.1101/gr.136127.111.
50. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, *et al*. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 2014;42:D1001–6, http://dx.doi.org/10.1093/nar/gkt1229.