# Supplementary Text

Supplementary materials for *Boer C.G., et al., Stratified hand phenotypes identifies WNT9A as novel gene associated with thumb osteoarthritis.*

**Contents**

## Supplementary methods

*Description of the Rotterdam Study (RS)*

The Rotterdam Study(RS) is a prospective population based cohort consisting of elderly inhabitants, 45 years and older, of the Omnoord district in the city of Rotterdam, the Netherlands[1]. The RS has been ongoing since 1990 to study the determinants of chronic disabling disease in the elderly. The Rotterdam Study I (RS-I) is the first cohort, of 7,983 persons living in the Ommoord district of Rotterdam in the Netherlands. All subjects were aged 55 years and older, recruitment of participants started in 1990. The Rotterdam Study II (RS-II) started in 1999 when 3,011 participants moved into the study since they became 55 years of age or moved into the study district. The Rotterdam Study III (RS-III) started in 2006 with all 3,932 participants aged 45 years and older from the study district not yet included in the study. The present study includes all participants for whom radiographs of the hand joints at baseline visit were present. During the visit to the study center (online supplementary Table S1), bilateral hand radiographs were made during the baseline visit, which were examined and scored by trained radiologists[2]. For both hands all distal interphalangeal joints (DIP), interphalangeal joints (PIP), metacarpophalangeal joints (MCP), thumb interphalangeal joint(IP), first carpometacarpal (CMC1) joint and the trapezioscaphoid (TS) joint were scored according to the Kellgren-Lawrence (KL) OA severity grading scale[3]. . For 1,609 participants, we were unable to score one or more joints, or genetic data was ,not available, which left us with a total of 8,691 participants to perform the study. The interobserver reliability for KL≥2 was: DIP: κ=0.60, PIP =0.61, MCP =0.63 and CMC1/TS=0.75[2].

The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC (registration number MEC 02.1015) and by the Dutch Ministry of Health, Welfare and Sport (Population Screening Act WBO, license number 1071272-159521-PG). The Rotterdam Study has been entered into the Netherlands National Trial Register (NTR; www.trialregister.nl) and into the WHO International Clinical Trials Registry Platform (ICTRP; www.who.int/ictrp/network/primary/en/) under shared catalogue number NTR6831. All participants provided written informed consent to participate in the study and to have their information obtained from treating physicians

*Description of the Framingham Heart Study(FHS)*

The original Framingham Study was a population-based sample of adults (ages 28-61 years) that began in 1948[4] . The Framingham Offspring Study is composed of children of the original Framingham Heart Study participants, and the children's spouses[5]. As part of an ancillary study in 1992 to 1995, Offspring (and their spouses) were contacted by mail and telephone call to participate in a visit to assess hand OA. About 1,800 individuals (ages 28-82 years) were examined, representing about 65% of those contacted. Of these individuals, 1,293 participants returned for another hand examination in 2002 to 2005[6]. As osteoarthritis often does not present until a later date, we have used data from the 2002 to 2005 visit. Of which 1,203 had genotyping data available for analysis. Individuals underwent bilateral poster-anterior hand radiographs, which were read by a trained musculoskeletal radiologist. The bilateral 2nd-5th DIP, 2nd-5th PIP, 1st-5th MCP, IP, thumb base (carpometacarpal) joint and wrist joints were scored according to KL-score with good inter-reader reliability (weighted $\kappa$=0.76).

*Patient and Public Involvement*

Patients were involved in the design of this study, through the Dutch Arthritis Association (DAA)through the founding of this research (DAA 2010_017). Patients and the general Public will be informed of the results through the dedicated website of the Dutch Arthritis Association (https://reumanederland.nl/), and via the Erasmus MC Rotterdam Osteoarthritis Research (ROAR) twitter account (@roar_NL).

*GWAS, discovery, replication and meta-analysis*

Genome-wide association(GWAS) methods of the discovery cohort have been described previously[7], briefly genotyped variants were imputed after quality control using the michigain imputation sever(HRC panel v.1.1[8]). Genetic dosages were used to investigate association with the stratified hand OA phenotypes using RVTESTS[9]. All performed GWAS, including replication analysis, were adjusted for age, sex and the first four genetic principal components. We did not include BMI in our model, as inclusion of heritable and causally associated covariates can introduce "collider bias"[10]. Variants were considered for replication if p-value$\leq 1*10^{-06}$. Meta-analysis between discovery (RS) and replication (FHS) was performed using inverse variance weighting (METAL[11]). Variants were considered "replicated and genome-wide significant if their replication p-value<0.05, had

the same direction of the beta and the meta-analysis p-value<$5*10^{-08}$ (Genome-wide significance threshold)[12]. Variants were considered genome-wide suggestive when they were replicated and the meta-analysis p-value<$1*10^{-06}$. Independence for each signal was determined by conditional-joint analysis(Jo-Co) in GCTA[13]. Manhattan, QQ-plots and heatmap plots were made in R[14] using the CRAN software packages qqman, gplots and RcolorBrewer. Images were saved in. eps format, font size, style and additional text were added/modified using Adobe Illustrator.

*Hand osteoarthritis Cluster Analysis*

Cluster analysis was performed on all hand joints in the RS cohorts (n=8,691). The goal was to organize the observed data into meaningful clusters using hierarchical clustering of a Euclidean distance matrix using Ward's method. For each radiographic measurement separately (joint space narrowing, osteophyts, and KL-score) we performed normalization of the data through scaling the data across the joints and calculated a distance matrix based on Euclidean distances. Next, we used Ward's agglomerative hierarchical clustering method to generate tree diagrams. Where the vertical axis denoted the linkage distance. We used multidimensional scaling (MDS) to further detect biologically interpretable clusters between the joints groups. For the multidimensional scaling we used 2 dimensions. We scaled the data and calculated a distance matrix based on Euclidean distances to be used in the MDS. These cluster analysis were also performed on the radiographic KLscore of each measured joint of the hand separately, without grouping joints per type. Clusters were determined by comparing the results from all cluster analysis. The following clusters were recognized: finger KLsum included all DIP and PIP joints, excluding the IP joints, as these cluster either more with the MCP in the KL-grade, or the DID/PIP dependent on the radiographic feature examined (online supplementary Figure S1 and S2). The Finger KLsum score includes in total the KL grade of 16 joints and can range from 0 to 64, where a score of 64 means that the maximum KL grade (4) was assigned to all joint included in the KLsum score. The thumb KLsum included the TS and the CMC1 joints, the IP joint was excluded as this did not cluster with CMC1 or TS joints in any of the radiographic features examined. The Thumb KLsum score includes in total the KL grade of 4 joints can range from 0 to 16, where a score of 16 means that the maximum KL grade (4) was assigned to all joints included in the KLsum score. The hand KLsum included all DIP, PIP, MCP, IP and CMC1 joints, the IP and TS joint

were excluded, as these do not consistently cluster with the rest of the joints in the radiographic features examined. The Hand KLsum score includes in total the KL grade of 30 joints and can range from 0 to 120, where a score of 120 means that the maximum KL grade (4) was assigned to all joints included in the KLsum score. All cluster analysis were performed in R.

*Lookup in DECODE and UKbiobank osteoarthritis GWAS*

The SNVs identified in the discovery GWAS (RS) and replicated in the replication cohort (FHS) were also examined for association with clinical osteoarthritis in a meta-analysis of the Icelandic DECODE population cohort and the united kingdom based UKbiobank population cohort[15,16]. Information on osteoarthritis was derived from a national Icelandic hip or knee arthroplasty registry, electronic health records (using ICD10 codes), and a dedicated hand osteoarthritis database[15,16].

*Variant functional annotation*

A locus was defined as the region 500kb upstream and 500kb downstream from the lead SNP. For each lead variant SNPs in high LD ($r^2 \geq 0.8$) were determined and annotated using annotation provided by FUMA and HaploregV4. 1[17,18]. All variants were and gene regulatory region annotation were provided by the SNP2GENE tool from FUMA, Haploreg V4 annotation and from the ROADMAP and ENCODE projects[19,20]. Intersection of the variant with gene regulatory elements as predicted by histone post-translational modifications, were made by the ROADMAP project[19]. CTCF-protein binding Chip-seq peaks in primary osteoblast cells were generated by ENCODE[20] and visualized via UCSC genome-browser[21]. Annotation of the variant location, number of proteins bound and transcription factor(TF) binding motifs change was done via Haploreg V1.4. as described previously[18]. TF binding to gene promoter locations were taken from the ENCODE Transcription Factor Binding Site Profiles dataset[20] accessed through harmonizome[22].

*Human embryonic cartilage ATAC-seq*

Intersection of SNVs with open chromatin regions in human embryonic cartilage was done using ATAC-seq data. For a detailed description of this human embryonic ATAC seq dataset see[23]. We have acquired chromatin accessibility from human embryonic cartilage ATAC-seq datasets at E59 of gestation[23] to investigate if our lead SNV and variants in high

LD co-localized with these open chromatin regions. All variants in LD with the lead variant (hg19 coordinates) were intersected with E59 ATAC-seq peaks from four cartilage tissues (proximal femur, distal femur, proximal tibia and distal tibia) using the UCSC Genome Browser Table Browser tool. ATAC-seq peaks from these tissues were also used to map regulatory elements in the Wnt locus on human chromosome 1.

*Gene prioritization*

Candidate genes in the locus were defined as: all genes annotated to be (partially) located within the locus (500 kb upstream and downstream of the lead SNV).All genes fitting this definition were considered as potential causal gene, in total we analyzed 18 genes (supplementary table 3). Genes were prioritized based on several lines of evidence: 1)eQTL-analysis consisted of two separate hip OA cartilage datasets (n=29 and n=87) of which genotypes and RNA-seq data were available[24,25]. For each dataset an eQTL analysis was performed for each lead SNP with all genes in the locus. To increase detection power, we then meta-analyzed the result from both datasets together using a weighted meta-analysis based on p-values, sample size and direction of effect in METAL[11]. 3)The 3D chromatin structure of the locus was examined using Capture Hi-C data from human mesenchymal stem cells (hMSCs)[26].We visualized the chromatin interactions between the lead SNV and possible causal SNV(s) with promoter regions of *WNT9A/WNT3A/JMJD4* and *SNAP47* and the rest of the investigated locus.

*RNA-sequencing for differential expression*

Differential gene expression between OA lesioned and Preserved cartilage: Post-RNA isolation (Qiagen RNeasy Mini Kit, RIN >7) of 40 knee (15 paired preserved (P) and OA lesioned (OAL), 7 P only and 3 OAL only) and 28 hip (six paired P and OAL, 14 P only and 2 OAL only) cartilage samples (supplementary table 3), paired-end 2×100 bp RNA library sequencing (Illumina TruSeq RNA-Library Prep Kit, Illumina HiSeq2000) resulted in an average of 10 million fragments per sample. Reads were aligned using GSNAP against GRCh37/hg19, in which SNPs from the Genome of the Netherlands consortium with a minor allele frequency (MAF) >1% were masked to prevent alignment bias. Number of

fragments per gene were used to assess quantile-adjusted conditional maximum likelihood (edgeR, R-package)[27]. Subsequently, differential gene expression analysis was performed pairwise between P and OAL samples for which we had RNA of both (n=21).

*Expression quantitative trait loci analysis*

The eQTL-analysis consisted of two separate OA cartilage datasets (n=29 and n=87) of which genotypes and RNA-seq data were available. RNA-expression and eQTL analysis of the first dataset of 29 samples has been previously described here[28]**.** The second dataset consisting of 87 samples were collected and RNA was extracted as previously described[25]. Post RNA isolation, multiplexed libraries were sequenced on the Illumina HiSeq 2000 (75bp paired-end read length). Sample QC was carried out using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and transcript-level quantification was performed using salmon[29] based on the GRCh38 cDNA assembly [http://ftp.ensembl.org/pub/release-87/fasta/homo_sapiens/cdna/]. Transcript-level estimates were summarized to gene-level estimates (scaled transcripts per million) based on Ensembl gene IDs using tximport[30]. Only genes with ≥1 count per million in ≥20% samples were kept, with 87 low-grade cartilage samples and 15,249 genes post QC. All 87 individuals were genotyped using Illumina HumanCoreExome. Genotypes were called using GenCall and mapped to GRC37/hg19. Following sample and variant QC, we imputed up to HRC panel v1.1 using the Michigan imputation server (https://imputationserver.sph.umich.edu/index.html). We followed the GTEx approach for eQTL analysis[31]. Briefly, we normalized gene expression between samples using weighted trimmed mean of M-values implemented in edgeR[27]. For each gene, expression across samples was normalized using an inverse normal transformation. To identify cis-eQTLs within 1Mb either direction of a gene transcription start site, we used the GTEx modified version of FastQTL (https://github.com/francois-a/fastqtl; v6p), including 15 Probabilistic Estimation of Expression Residuals (PEER) factors[32], sex and genotype array as covariates. We generated empirical p-values, with a 5% Storey-Tibshirani FDR cut-off to identify genes with a significant eQTL[33]. The normalised effect size (NES) of each eQTL is reported for the alternate allele.

*Methylation quantitative trait loci analysis*

We used cartilage CpG methylation and genotype data that had been generated previously using Illumina's Infinium HumanMethylation450 array and HumanOmniExpress array, respectively[34]. Methylation and genotype data were generated from 87 patients who had undergone knee or hip joint arthroplasty: 57 knee OA patients, 14 hip OA patients and 16 control patients who had undergone hip replacement due to a neck-of-femur (NOF) fracture. If the SNP reported as associated with OA in the GWAS was directly genotyped on the HumanOmniExpress array, that SNP data was used by us. If the SNP was not, we searched for and, where possible, used a proxy SNP that was in perfect or high LD (pairwise r2>0.7) with the association SNP. Proxies were derived from a candidate list using LDlink's LDproxy tool[35] and European population data. Where multiple proxies were identified, the one with the highest r2 relative to the association SNP was chosen. For each locus, we covered a 1Mb region encompassing 500kb upstream and 500kb downstream of the association SNP. For each CpG within the 1Mb, linear regression was used to measure the relationship between methylation in the form of M-values and genotype (0, 1 or 2 copies of the minor allele) at the OA association SNP or its proxy. Age, sex and joint site/condition were added into the model as covariates. Methylation status is reported using β-values (ranging from 0 for no methylation to 1 for 100% methylation). mQTL calculations were performed using Matrix eQTL[36] implementing a false discovery rate (FDR) estimation that is based on the Benjamini–Hochberg FDR procedure[37] and which accounts for the number of tests performed.

# References


1. Ikram MA, Brusselle GGO, Murad SD, et al. The Rotterdam Study: 2018 update on objectives, design and main results. Eur J Epidemiol 2017;32:807-50.
2. Dahaghin S, Bierma-Zeinstra SM, Ginai AZ, Pols HA, Hazes JM, Koes BW. Prevalence and pattern of radiographic hand osteoarthritis and association with pain and disability (the Rotterdam study). Ann Rheum Dis 2005;64:682-7.
3. Kellgren JH, Lawrence JS. Radiological Assessment of Osteo-Arthrosis. Ann Rheum Dis 1957;16:494-502.
4. Dawber TR, Meadors GF, Moore FE. Epidemiological Approaches to Heart Disease: The Framingham Study *. Am J Public Health Nations Health 1951;41:279-86.
5. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol 1979;110:281-90.
6. Haugen IK, Englund M, Aliabadi P, et al. Prevalence, incidence and progression of hand osteoarthritis in the general population: the Framingham Osteoarthritis Study. Ann Rheum Dis 2011;70:1581-6.
7. den Hollander W, Boer CG, Hart DJ, et al. Genome-wide association and functional studies identify a role for matrix Gla protein in osteoarthritis of the hand. Ann Rheum Dis 2017;76:2046-53.
8. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016;48:1279-83.
9. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics 2016;32:1423-6.
10. Day FR, Loh PR, Scott RA, Ong KK, Perry JR. A Robust Example of Collider Bias in a Genetic Association Study. Am J Hum Genet 2016;98:392-3.
11. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;26:2190-1.
12. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic epidemiology 2008;32:381-5.
13. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88:76-82.
14. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria., 2014. at http://www.R-project.org/.)
15. Styrkarsdottir U, Lund SH, Thorleifsson G, et al. Meta-analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1 and 13 more new loci associated with osteoarthritis. Nat Genet. United States2018:1681-7.
16. Styrkarsdottir U, Thorleifsson G, Helgadottir HT, et al. Severe osteoarthritis of the hand associates with common variants within the ALDH1A2 gene and with rare variants at 1p31. Nat Genet 2014;46:498-502.
17. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun 2017;8:1826.
18. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res 2016;44:D877-81.
19. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317-30.
20. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57-74.
21. Casper J, Zweig AS, Villarreal C, et al. The UCSC Genome Browser database: 2018 update. Nucleic Acids Res 2018;46:D762-d9.

22. Rouillard AD, Gundersen GW, Fernandez NF, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford) 2016;2016.
23. Richard D, Liu Z, Cao J, et al. Evolutionary Selection and Constraint on Human Knee Chondrocyte Regulation Impacts Osteoarthritis Risk. Cell 2020.
24. Steinberg JS, S.; Butterfield, N.C.; Roumeliotis,T.I.; Fontalis A.; Clark, M.J.; Jayasuriya, R.L.; Swift,D.; Shah,K.M.; Curry,K.F.; Brooks,R.A.; McCaskie, A.W.; Lelliott, C.J.;Choudhary, J.S.; Bassett, J.H.D.; Williams,G.R.; Wilkinson, J.M.; Zeggini, E. Decoding the genomic basis of osteoarthritis. bioRxiv 8358502019.
25. Steinberg J, Ritchie GRS, Roumeliotis TI, et al. Integrative epigenomics, transcriptomics and proteomics of patient chondrocytes reveal genes and pathways involved in osteoarthritis. Sci Rep 2017;7:8935.
26. Wang Y, Song F, Zhang B, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol 2018;19:151.
27. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139-40.
28. den Hollander W, Pulyakhina I, Boer C, et al. Annotating Transcriptional Effects of Genetic Variants in Disease-Relevant Tissue: Transcriptome-Wide Allelic Imbalance in Osteoarthritic Cartilage. Arthritis Rheumatol2019:561-70.
29. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 2017;14:417-9.
30. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res 2015;4:1521.
31. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580-5.
32. 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 2012;7:500-7.
33. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100:9440-5.
34. Rushton MD, Reynard LN, Young DA, et al. Methylation quantitative trait locus analysis of osteoarthritis links epigenetics with genetic risk. Hum Mol Genet 2015;24:7432-44.
35. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015;31:3555-7.
36. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 2012;28:1353-8.
37. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res 2001;125:279-84.