Supplementary Information 2 for methods

GENOTYPING AND IMPUTATION

Methods description for study populations at deCODE genetics (Iceland, Denmark, Sweden, Norway)

Whole-genome sequencing

Preparation of samples. Three different sample preparation kits were employed. TruSeq Nano from Illumina (Method A), TruSeq PCR-Free from Illumina (Method B), NEBNext Ultra™ II PCR-free from New England Biolabs (Method C). Methods A and B: In short, either 50 ng (Method A) or 1 μg (Method B) of genomic DNA, isolated from either frozen blood samples or buccal swabs, was fragmented to a mean target size of 350-450 bp using the Covaris E220 instrument. End repair, generating blunt ended fragments was performed followed by size selection using different ratios of AMPure XP magnetic purification beads. 3'-Adenylation and ligation of indexed sequencing adaptors containing a T nucleotide overhang was performed, followed either by AMPure purification alone (Method B) or purification followed by PCR enrichment (10 cycles) using appropriate primers (Method A). The quality and concentration of all sequencing libraries was assessed using either the Agilent 2100 Bioanalyzer (12-samples) or the LabChip GX (96-samples) instrument from Perkin Elmer. Sequencing libraries were diluted and stored at −20 °C. Further quality control of sequencing libraries was done by multiplexing and pooling 96 samples and sequencing each pool on an Illumina MiSeq instrument to assess optimal cluster densities, library insert size, duplication rates and library diversities.

Method C: In short, 0.5-1 μg of genomic DNA was fragmented to a mean target size of 450-500 bp using a Covaris LE220plus instruments and 96-well TPX-AFA plates (Covaris Inc). End repair and A-tailing was performed in a single step followed by ligation of unique dual

Saevarsdottir S, *et al. Ann Rheum Dis* 2022; 81:1085–1095. doi: 10.1136/annrheumdis-2021-221754

Supplementary Information 2 for methods

indexed sequencing adaptors (IDT for Illumina) and two rounds of SPRI-bead purification (0.6X) using the Hamilton STAR NGS liquid handler. Quality (concentration and insert size) of sequencing libraries was determined using the LabChip GX (96-samples) instrument (Perkin Elmer). Sequencing libraries were pooled appropriately for sequencing using Hamilton STARlet liquid handlers.

All steps in the workflows described above were monitored using an in-house laboratory information management system (LIMS) with barcode tracking of all samples and reagents.

DNA whole genome sequencing. Paired-end sequencing-by-synthesis (SBS) was performed on Illumina sequencers, HiSeq 2000/2500, HiSeq X and NovaSeq6000 instruments, respectively. Readlengths depended on the instrument and/or sequencing kit being employed and varied from 2x101 cycles to 2x151 cycles of incorporation and imaging. Real-time analysis involved conversion of image data to base-calling in real-time. For the HiSeq instruments, the sequencing libraries were hybridized to the surface of the flowcells using the Illumina cBot™, with a single sample per lane. For the NovaSeq6000, on-board clustering was performed using appropriately pooled samples. Paired-end sequencing on the S4 flowcell (v1.0 chemistry) was performed with a readlenght of 2x151 cycles of incorporation and imaging, in addition to 2*8 index cycles.

Reference. As previously described in detail[1], the reference sequences used to map reads are based on the human genome assembly GRCh38, not including alternate assemblies (GCA_000001405.15_GRCh38_no_alt_analysis_set.fna on ftp.ncbi.nlm.nih.gov) in addition to sequences determined to represent common contaminants in our sequencing pipeline. These sequences are the sequences of the bacteriophage PhiX (ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/enterobacteria_phage_phix174_sensu_lato_uid1

Supplementary Information 2 for methods

4015/NC_001422.fna), the bacteria Ralstonia Pickettii (Ralstonia_pickettii_12D_uid58859) and two sequences from the human microbiome (Coprobacillus_D7_uid32495/NZ_EQ999972.fna, Coprobacillus_D7_uid32495/NZ_EQ999922.fna).

Sequence alignment[1]. The raw sequences where aligned against the reference described above with BWA version 0.7.10 mem[2]. The sequences in the BAM files were realigned around indels with GenomeAnalysisTKLite/2.3.9[3] using a public set of known indels and a set of indels previously discovered in the Icelandic data[1]. PCR duplicates were marked with Picard tools 1.117.

Filtering of BAM files[1]. The sequencing data of each individual was organized into one BAM file per lane on the flowcell. The sequencing data generated for each individual from HiSeqX machines largely derives from single lanes while the sequencing data coming from other machines is derived from multiple lanes. In the following, BAM files created before and after the merging per individual will be referred to as BM- and AM-BAM files, respectively. A pileup using samtools BM-BAM files was performed at all sites where the sample is homozygous according to their chip genotypes. These pileups were combined for each sample and a mismatch rate was calculated as the number of bases not matching the chip genotype divided by the number of chip typed bases. BM-BAM files with mismatch rates above 2% or if they failed any of the following criteria were excluded: 1) Mean base quality <25, 2) Percent marked duplicate >50, 3) Mean N per read >30, 4) Percent mapping quality below 20 >11, 5) Percent reads unmapped >40, 6) Percent both reads in a pair unmapped >40, 7) Percent first read in a pair unmapped >40, 8) Percent second read in a pair unmapped >40. BM-BAM files that passed all the criteria were merged into single AM-BAM files based on unique individual and sample type source combinations.

Supplementary Information 2 for methods

Sequence variant calling. For the Swedish, Danish and Norwegian cohort, AM-BAM files as described above were collected from samples of Northwestern-European origin and for the Icelandic cohort (50,306 AM-BAM files of Icelandic origin). Duplicate samples were discarded based on sequencing yield. Only samples with a genome-wide average coverage of over 20X were considered. Samples that were contaminated (according to the read_haps tool[4]) were discarded. This left 17,683 samples from the Swedish, Danish and Norwegian cohort and 49,962 Icelandic samples for variant calling. Joint variant calling was performed using Graphtyper (version 1.4)[5], separately on the two datasets.

Chip genotyping and long range phasing.

Swedish, Danish and Norwegian cohort. Chip genotypes for 604,064 samples of Northwestern-European origin were collected. The majority of the samples were chipped using chips from the Illumina Global Screening Array (gsa) family of chips (n=434,595) with the remaining samples having chip data coming from the older OmniExpress (omni) family of chips. Individual genotype arrays were discarded if the total yield was below 98%. Sample duplicates were discarded thus: A fingerprint of ~160 common variants was created for each array. If two or more arrays had a fingerprint similarity above a certain threshold, a kinship analysis with KING (version 2.1.5)[6] was performed. For duplicates, a sample with sequencing data associated was kept, if possible. Otherwise the array with the highest yield was kept. After removal of low yield chips and duplicates, chips for 570,100 samples were left for long range phasing. Only variants which had previously been seen as polymorphic in in-house sequencing data were analysed. For each array family (omni or gsa) variants were excluded if they showed an allele frequency bias between the array family subtypes in the Norwegian sample cohort. Variants with a missing rate greater than 2% or MAF less than 0.5% were discarded. The samples were phased using Eagle (version 2.4.1)[7] on a per

Supplementary Information 2 for methods

genotyping array family and chromosome basis. Missing genotypes and non-overlapping variants were then imputed between the two chip families (omni, gsa) using a viterbi algorithm giving the most probable haplotype reconstruction using haplotype sharing in a Hidden Markov Model based on a Li and Stephens model[8] similar to the one used in IMPUTE2[9]. Finally the two chiptype-specific genotype sets were combined after imputation and phased again using Eagle, resulting in a panel of 570,100 samples with chip genotypes.

Icelandic cohort. Chip genotypes for 166,467 Icelandic samples were collected. The majority of the samples were chipped using chips from the Illumina OmniExpress family (n=129,108) with the remaining samples having chip data coming from the older HumanHap family of chips. Individual genotype arrays were discarded if the total yield was below 98%, leaving chips for 166,364 samples for long range phasing. Only variants which had previously been seen as polymorphic in inhouse sequencing data were analyzed. For a given array family (omni or humanhap) variants were excluded if they showed an allele frequency bias between the array family subtypes. The chip genotypes were then phased using the algorithm described by Kong et al.[10], resulting in a panel of 166,281 samples with LRP chip genotypes.

Phasing and imputation

Sequence variants passing filters were phased, creating a haplotype reference panel, using the long-range phased chip data. The haplotype reference panel was then used to impute each sequence variant into all chip genotyped samples, again using the long-range phased chip data. The imputation consists of estimating, for each haplotype, haplotype sharing with haplotypes in the haplotype reference panel, giving haplotype weights for each haplotype. These weights, along with allele probabilities for each haplotype in the haplotype reference

Supplementary Information 2 for methods

panel allow imputation with a Li and Stephens model[8] similar to the one used in IMPUTE2[9]. Estimation of haplotype weights is based on long-range phased chip haplotypes.

Sequence variant phasing consists of iteratively imputing the phase in each sequenced sample using the other sequenced samples and the estimated phase from last iteration. The imputed genotypes, along with the original genotypes are weighted together to estimate new allele probabilities for the haplotypes. The imputation part is the same as described above.

Method description for other study populations

UK Biobank.

The 500,000 UK Biobank samples were genotyped with a custom-made Affymetrix chip, UK BiLEVE Axiom and the Affymetrix UK Biobank Axiom array and imputed using the Haplotype Reference Consortium and UK10K haplotype resources.[11]

FINNGEN

FINNGEN was genotyped using the Axiom array (https://www.finngen.fi/en). The Swedish, Danish and Norwegian samples were genotyped with various Illumina SNP chips, long-range phased using Eagle2[7] and imputed with a phased haplotype reference panel created from 17,408 whole-genome sequenced individuals from Northwestern Europe, including 8,635 Danish, 3,329 Norwegian, and 3,704 Swedish samples. Genotype samples and variants with less than 98% yield were excluded. All sequencing and genotyping of the Icelandic, Danish, Swedish and Norwegian cohorts was done at deCODE genetics.

Saevarsdottir S, *et al. Ann Rheum Dis* 2022; 81:1085–1095. doi: 10.1136/annrheumdis-2021-221754

Supplementary Information 2 for methods

ANCESTRY ANALYSIS

Genetic ancestry analysis to identify groups of similar ancestry was performed for the Danish, Swedish and Norwegian sample sets separately. First ADMIXTURE v1.23[12] was run in supervised mode with 1000 Genomes populations CEU, CHB, and YRI[13] as training samples and Danish, Swedish or Norwegian individuals as test samples, and test samples with less than 0.9 assigned CEU ancestry were excluded. Remaining test samples were projected onto 20 principal components (PCs) calculated from an in-house European reference panel. The UMAP R package [14] was used to reduce the test sample coordinates to two dimensions. Additional European samples not in the original reference were also embedded into the UMAP space to help identify the ancestries represented in clusters. A polygon informed by visual inspection was drawn to include all samples with very similar ancestries to the main Danish, Swedish or Norwegian clusters.

ASSOCIATION TESTING

We used logistic regression in the Icelandic, Swedish, Danish, Norwegian and UK datasets separately to test for association of RA overall, seropositive and seronegative RA, with sequence variants, using software developed at deCODE genetics.[15-17] In the Icelandic analysis, we adjusted for sex, county of origin, current age or age at death, blood sample availability for the individual, and an indicator function for the overlap of the lifetime of the individual with the time-span of phenotype collection. In the Danish, Swedish and Norwegian analysis, we adjusted for sex, age, chip-typed and/or sequenced status and 20 principal components. In the UK analyses, we adjusted for sex, age, and 40 principal components. We used LD-score regression to account for distribution inflation due to cryptic relatedness and population stratification.[18] After this adjustment, we do not observe inflation in the test

Supplementary Information 2 for methods

statistics for rare variants compared to common variants (both are adjusted with the same inflation factor). This can be seen in QQ plots of the test statistics for each cohort (Supplementary Information 3). Likewise, the genomic inflation adjustment for each cohort would also remove any inflation due to case-control imbalance. The genomic inflation factors, estimated using LD score regression, are $\lambda g = 1.037$ for all RA, $\lambda g = 1.033$ for seropositive RA and $\lambda g = 1.009$ for seronegative RA. We did not adjust the meta-analysis for this inflation, but all of our association results remained significant after this adjustment. SNP-based heritability (observed scale) was estimated using LD score regression[18]. In these analyses, we used results for about 1.2 million well imputed variants, and for LD information we used precomputed LD scores for European populations (downloaded from:https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2). The heritability estimates (total observed scale h2) for RA overall is 0.128 (0.0118), for seropositive RA 0.192 (0.0216) and for seronegative RA 0.0989 (0.0192).

Furthermore, we divided the study population into two subgroups of similar size (Denmark-Iceland-Finland with 15,976 cases and 554,675 controls; and Norway-Sweden-UK with 15,337 cases and 440,702 controls) and calculated the genetic correlation between the RA subsets. The correlation between seropositive and seronegative RA was (in meta-analysis comparing seropositive RA in one subgroup with seronegative RA in the other subgroup, and vice versa) rg 0.87, se 0.13, P=4.50E-12 (see Supplementary Table 9 for further information).

We used a fixed-effects inverse variance meta-analysis[19] to combine results from the six study groups. Variants with imputation information below 0.8 were excluded. Genome-wide significance was determined using class-based Bonferroni significance thresholds adjusting for all 64 million variants tested, maintaining an unadjusted significance threshold of 8×10^{-10}. Sequence variants were split into five classes based on their genome annotation, with significance threshold for each class based on the number of variants in that class (e.g. lower

Supplementary Information 2 for methods

thresholds for loss of function (high impact) and missense variants (moderate impact), as previously described[20]. The adjusted significance thresholds are 1.3×10^{-7} for variants with high impact (splice donor, splice acceptor, stop gained, frameshift, stop lost, initiator codon), 2.6×10^{-8} for variants with moderate impact (missense, splice region, stop retained, inframe indels), 2.4×10^{-9} for low-impact variants (synonymous, 5' UTR, 3' UTR, up- and downstream), 1.2×10^{-9} for other low-impact variants in DNase I hypersensitivity sites (intronic, intergenic, regulatory-region) and 5.92×10^{-10} for all other variants not in DNase I hypersensitivity sites (intronic, intergenic, regulatory-region).

The primary signal at each genomic locus was defined as the sequence variant with the lowest Bonferroni-adjusted *P*-value using the adjusted significance thresholds described above and the results are presented in Table 2 and Supplementary Tables 2-3. Conditional analysis was used to identify possible secondary signals within 500 kB from the primary signal (excluding the HLA-locus). This was done using genotype data for the Icelandic, Swedish, Norwegian, Danish and UK datasets and an approximate conditional analysis implemented in the GCTA software[21] for the Finnish summary data. Adjusted *P*-values and odds ratios were combined using a fixed-effects inverse variance method. Threshold for secondary signal was adjusted *P*-value less than ten times the class-specific significance threshold. The conditional analysis was done separately in RA overall, the seropositive and seronegative RA subsets. For significant associations in our study, we searched for functionally important variants, affecting protein coding, mRNA expression or protein levels, using a multi-omics approach (see below), in order to identify candidate causal genes, and highlight those with strongest effect on RA risk. For replication of reported variants, multiple testing was accounted for using Bonferroni correction.

Novelty of the GWAS significant loci was evaluated based on the GWAS catalog (https://www.ebi.ac.uk/gwas/home), with significance threshold set at $1.0\text{x}10^{-8}$ and searching

Supplementary Information 2 for methods

for all variants correlated (r^2>0.8) with the primary and secondary signals with RA or its subsets.

FUNCTIONAL EVALUATION OF IDENTIFIED SEQUENCE VARIANTS

We performed a systematic variant annotation of the RA associations to identify candidate causal genes, through identification of coding variants and variants affecting mRNA expression, as described below. Candidate causal gene refers to the gene that mediates the effect on the disease at a given locus with the highest probability, based on either the variants or highly correlated variants (r^2>0.8) being coding in the candidate causal gene (Supplementary Table 4, including likelihood of affecting the protein function, using dbNSFP 4.1c)[22], having the strongest and significant effect on mRNA expression of the candidate causal genes (top cis-eQTL, see transcriptomics chapter below) or having significant effect on levels of the protein encoded by the candidate causal gene (pQTL, see proteomics chapter below). The results of this multi-omics analysis are summarized in Figure 2.

TRANSCRIPTOMICS

We tested whether the sequence variants that associated with RA were in strong linkage disequilibrium (R^2>0.8) with top cis-eQTL variants for genes expressed in whole blood from 13,175 Icelanders and in adipose tissue from 700 Icelanders.[23] For gene expression, association was tested using a generalized linear regression, assuming an additive genetic effect and quantile-normalized gene expression estimates, adjusting for measurements of sequencing artefacts, demographic variables, and hidden covariates.[24] We also looked up the association with expression in other tissue types in GTEx (https://gtexportal.org) and in 15 published accessible data-sources (listed in Supplementary Table 5).

Supplementary Information 2 for methods

PROTEOMICS

We tested whether disease associating sequence variants associated with levels of 4,789 proteins in plasma (pQTL), measured on the SomaLogic® platform (SomaLogic, Inc.) in 35,559 Icelanders with genetic information and biological samples available, as previously described (Supplementary Table 7).[23 25 26] Multiple testing was accounted for using Bonferroni correction. For each pQTL variant we calculated the count and ranks of the associated proteins; provided in Supplementary Table 7 in the columns "Rank Proteins per pQTL", "Count Proteins per pQTL", "Rank pQTL per Protein" and "Count pQTL per Protein".

NETWORK ANALYSIS

We used the Ingenuity Pathway Analysis (IPA) software (QUIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis) to evaluate and illustrate whether identified candidate causal genes have experimental evidence for direct interaction between the proteins coded by those genes or indirect interaction (e.g. one affecting the level of another), supporting biological connection, illustrated as a network (Supplementary Figure 3).

PUBLICLY AVAILABLE SOFTWARE

Publicly available software that was used in conjunction with the algorithms in the sequencing processing pipeline (whole-genome sequencing, association testing, RNA-sequence mapping and analysis) is listed below (URLs):

BWA 0.7.10 mem, https://github.com/lh3/bwa 396

GenomeAnalysisTKLite 2.3.9, https://github.com/broadgsa/gatk/ 397

Supplementary Information 2 for methods

Picard tools 1.117, https://broadinstitute.github.io/picard/ 398

SAMtools 1.3, http://samtools.github.io/ 399

Bedtools v2.25.0-76-g5e7c696z, https://github.com/arq5x/bedtools2/ 400

Variant Effect Predictor, https://github.com/Ensembl/ensembl-vep 401

Read_haps, https://github.com/DecodeGenetics/read_haps

IPA software (QUIAGEN Inc.),

https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis

In-silico prediction of missense variants, https://sites.google.com/site/ jpopgen/dbNSFP

Supplementary Information 2 for methods

REFERENCES

1. Jonsson H, Sulem P, Kehr B, et al. Whole genome characterization of sequence diversity of 15,220 Icelanders. Sci Data 2017;4:170115.
2. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754-60.
3. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297-303.
4. Eggertsson HP, Halldorsson BV. read_haps: using read haplotypes to detect same species contamination in DNA sequences. Bioinformatics 2021;37(15):2215-17.
5. Eggertsson HP, Jonsson H, Kristmundsdottir S, et al. Graphtyper enables population-scale genotyping using pangenome graphs. Nat Genet 2017;49(11):1654-60.
6. Manichaikul A, Mychaleckyj JC, Rich SS, et al. Robust relationship inference in genome-wide association studies. Bioinformatics 2010;26(22):2867-73.
7. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 2016;48(11):1443-48.
8. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 2003;165(4):2213-33.
9. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5(6):e1000529.
10. Kong A, Masson G, Frigge ML, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet 2008;40(9):1068-75.
11. Welsh S, Peakman T, Sheard S, et al. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. BMC Genomics 2017;18(1):26.
12. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19(9):1655-64.
13. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature 2015;526(7571):68-74.
14. McInnes L, Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv e-prints 2018:1802.03426.
15. Gudbjartsson DF, Helgason H, Gudjonsson SA, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 2015;47(5):435-44.
16. Gudbjartsson DF, Sulem P, Helgason H, et al. Sequence variants from whole genome sequencing a large group of Icelanders. Sci Data 2015;25(2):150011.
17. Steinthorsdottir V, Thorleifsson G, Sulem P, et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet 2014;46(3):294-8.
18. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015;47(3):291-5.
19. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22(4):719-48.
20. Sveinbjornsson G, Albrechtsen A, Zink F, et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. Nat Genet 2016;48(3):314-7.
21. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 2012;44(4):369-75, S1-3.
22. Liu X, Li C, Mou C, et al. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med 2020;12(1):103.
23. Saevarsdottir S, Olafsdottir TA, Ivarsdottir EV, et al. FLT3 stop mutation increases FLT3 ligand level and risk of autoimmune thyroid disease. Nature 2020;584(7822):619-23.
24. Stegle O, Parts L, Durbin R, et al. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol 2010;6(5):e1000770.
25. Suhre K, Arnold M, Bhagwat AM, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun 2017;8:14357.
26. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. Nature 2018;558(7708):73-79.