Supplemental Materials: Identification and characterization of genetic risk shared across 24 chronic pain conditions in the UK Biobank.

Supplementary Note

Prior to running the genome-wide association study (GWAS) analysis, we prepared the data as follows. We used genotypes (EGAD00010001474 downloaded using the UKBB tool **ukbgene imp**), which included data that was both directly genotyped and imputed against the UK10K reference panel (Huang *et al.* 2015). Plink, version 1.9, (https://www.cog-genomics.org/plink/ (Chang *et al.* 2015).) provided SNP quality control filters that consisted of heterozygosity rate (| F het | > 0.2) (determined using the —het option in Plink), final call rate > 0.95 (—geno option in Plink), Hardy-Weinberg equilibrium, $P > 1.0x10^{-8}$ (–hwe option in Plink), and minor allele frequency > 0.01 (–maf option in Plink). Sample quality control filter removed mismatches between reported and genotyped sex (Category 100313).

For the GWAS analysis, we used Regenie (Mbatchou *et al.* 2021), a recently-published mixed-model method, particularly well-suited for GWAS of dichotomous outcomes (presence/absence of a given pain condition) with class imbalance (low case prevalence) and sample relatedness, i.e. high degree of familial relationship between participants in the study. In Regenie, the GWAS is run using Firth approximation-corrected logistic regression (Firth 1993). Briefly, the analysis was run in 2 steps: 1. a model-fitting step, in which genotyped SNPs are split into chunks and two levels of ridge regression were run to obtain a per-chromosome genetic predictor of the phenotype; 2. a test of association for all available (imputed) SNPs, also split into chunks, with covariates (described below) regressed out and predictors from the first step removed from phenotypes, using a leave-one-chromosome-out (LOCO) scheme. We used 339,444 genotyped SNPs in 100-SNP chunks in the first step and 11,359,143 imputed SNPs in 200-SNP chunks in the second step, with age, sex, and 10 PCs from genetic PCA (principal component analysis run on a subsample that was filtered using the above quality controls, as well as relatedness coefficient .05, using GCTA (Yang *et al.* 2011), the option –grm-cutoff). The PCs account for ancestral confounding, i.e. correlation between outcome of interest and ancestry. Including 10 is the standard practice in the genetics field as covariates in GWAS (Price *et al.* 2006).

The output of GWAS consists of effect sizes (regression coefficients) and P-values from regressions of each of the initial set of 33 chronic pain phenotypes and each of 11,359,143 imputed SNPs. For display of results, the SNPs' associations are shown in Manhattan plots (Supplementary Figure S1) as logarithmically-transformed P-values, ordered by chromosomal location and thresholded for statistical significance (with a standard genome-wide threshold of $P < 5x10^{-8}$). While we report these for each chronic pain condition, here, we are primarily interested in the correlations between the per-SNP associations with each pair of pain conditions.

Heritability

SNP heritability, h^2_{SNP}, is the variance in the phenotype explained by variance in genotypes (SNPs) of each condition (Zaitlen and Kraft 2012). We estimated h^2_{SNP} using linkage disequilibrium score regression (LDSC) (Bulik-Sullivan *et al.* 2015) which exploits and accounts for linkage disequilibrium, or non-independence between SNPs that lie near each other and are inherited together. Per-SNP statistics of association (chi-square) with the phenotype are regressed on the the total linkage disequilibrium score (measure of SNP correlation with other SNPs that is a function of both magnitude and number of correlations). The slope of this regression is h^2_{SNP}.

For case-control conditions, such as the ones used here, the h^2_{SNP} estimate is converted from the observed to liability scale, which assumes a continuous underlying latent risk, as described in (Lee *et al.* 2011). We excluded conditions whose genetic associations were not significantly heritable (less than or equal to 2 standard errors above zero, $h^2_{SNP} - 2 * SE \leq 0$).

Genetic correlations

To estimate genetic correlations, we used GWAS summary association statistics (standardized effect sizes) as input to the **ldsc** function in the Genomic SEM R package (https://github.com/GenomicSEM/GenomicSEM), which invokes LDSC (Bulik-Sullivan *et al.* 2015), modified by the authors of Genomic SEM to include a sampling matrix that corrects for sample overlap (Grotzinger *et al.* 2019). To estimate genetic correlations, LDSC uses the same method as described above in the Heritability section, except substituting the product of association statistics (chi-square) for the two phenotypes. The slope of this regression is the genetic correlation between the phenotypes.

Genomic SEM

The Genomic SEM framework enabled us to: a) model the underlying factor structure while accounting for the complex correlations within genetic segments and b) run GWASs on the resulting factors. Genomic SEM also produces a Q heterogeneity statistic (QSNP) that indexes single nucleotide polymorphisms (SNPs) unlikely to operate through the genomic factors, such as variants that have a disproportionately strong effect on one condition or directionally opposing effects on a subset of traits. Collectively, this allows for distilling SNPs associated with general pain etiology from trait-specific, genetic pathways.

EFA-CFA

To test our models, we used exploratory factor analysis, followed by confirmatory factor analysis in Genomic SEM, also known as the EFA-CFA approach. After running EFA using the **fa** function in the 'psych' R package, we used its output to inform model specification for CFA in Genomic SEM. We specified a condition to load on a factor in the CFA model if it had positive standardized EFA loadings > 0.2, with the highest loading dictating the factor onto which the indicator would load in CFA. If an indicator had 2 loadings within 0.1, both were included in the CFA model. We further specified residual covariances for conditions that are very similar conceptually and have definitional overlap (knee arthrosis and knee pain, hip arthrosis and hip pain, general chest pain/discomfort and chest pain during physical activity, headache and migraine). These residual covariances obviated one of the 3 factors (a headache-migraine factor), leaving a two-factor model. Polymyalgia rheumatica and ulcerative colitis were excluded at this step due to lack of positive loadings (greater than) 0.2 onto any factor in the EFA model. The correlated factors CFA served as the interim step between EFA and bifactor CFA, in which we specified all conditions to load onto one general factor and the prior model's correlated factors as now uncorrelated additional specific factors.

EFA-CFA model comparison with anatomic and etiologic models

The EFA-CFA model with the specific factor encompassing musculoskeletal conditions provided a better fit than either the anatomic or the etiologic model. In the anatomic model, the Leg/Foot, Pelvic, and Torso specific factors had largely non-significant loadings from their indicators, suggesting that shared variance for these conditions was explained primarily by the general factor. The 5 residual variances initially estimated to be negative also show that the model as specified is not an optimal fit for the data. The etiologic model, on the other hand, with 4 of 6 specific factor conditions having non-significant loadings, shows that this factor does not explain sufficient shared variance for conditions grouped by putative inflammatory etiology beyond what is explained by the general factor.

Validation of the EFA-CFA model

To validate our findings for the EFA-CFA model, we split the genome into odd and even autosomes. At this step, we excluded conditions whose genetic associations were not significantly heritable (less than or equal to 2 standard errors above zero, $h^2_{SNP} - 2 * SE \leq 0$) in both odd and even autosomes: arthropathy of carpometacarpal joint, diabetic neuropathy, Crohn's disease, fibromyalgia, prostatitis, seropositive rheumatoid arthritis, and urinary colitis. This exclusion does not imply that these conditions are not heritable or are genetically unrelated to the common factors, as some conditions may be selectively related to genes on odd or even autosomes. However, the exclusions helped ensure that the conditions included in the factor model had broad polygenic representation across odd and even autosomes independently. Then we ran EFA on the odd set and used its output to specify loadings in Genomic SEM on the same set of chromosomes for CFA. We ran the same model that we used for CFA in the even chromosomes in Genomic SEM and compared the fit indices. The odd chromosome model yielded a CFI of 0.88 and SRMR of 0.12 and the even one a CFI of 0.90 and SRMR of 0.13. The factor structure for this validation model was similar to the main whole-genome model, except there were 2 specific factors: one with loadings from arthropathies, back pain, Carpal tunnel, enthesopathies of the lower limb, CWP, hip pain, hip arthrosis, knee pain, knee arthrosis, leg pain, neck pain, enthesopathies, pain in joint, and the other with loadings from back pain, chest pain (baseline), chest pain during physical activity, cystitis, gastritis, CWP, gout, oesophagitis, rheumatoid arthritis, and stomach pain. All loadings were significant, except for cystitis on the second specific factor. Given these comparable metrics in the training, validation, and whole genome datasets we concluded that using EFA and CFA on the same dataset did not result in substantial overfitting.

Factor GWAS

SNP association effects on the general factor and the specific musculoskeletal factor were calculated by adding the genotypic score for each SNP to the genetic correlation matrix output by LDSC for 24 conditions, estimating a new matrix of correlations and fitting the model with additional paths from the SNP to each of the factors (Supplementary Figure S6).

QSNP heterogeneity test

Genomic SEM produces a Q heterogeneity statistic (QSNP) that indexes single nucleotide polymorphisms (SNPs) unlikely to operate through the genomic factors, such as variants that have a disproportionately strong effect on 1 condition or directionally opposing effects on a subset of traits. Collectively, this allowed for distilling SNPs associated with general pain etiology from trait-specific, genetic pathways. To conduct a heterogeneity Q test (Huedo-Medina *et al.* 2006; Grotzinger *et al.* 2019), we specified a less restrictive model, in which for every SNP, path coefficients were estimated from the SNP to the individual phenotypes (independent pathways model). We formally assessed the difference between the 2 models – common and independent pathways – using the chi-square difference test. If significant, the SNP's effects on the individual pain conditions were interpreted to be inadequately modelled by the factor approach. Because this test was calculated for every imputed SNP, we used the standard whole-genome correction for multiple testing, $p < 5x10^{-8}$, as the threshold for Q significance. The resulting QSNPs were considered to be associated with pain conditions independent of common factors.

Additionally, we tested for SNPs in linkage disequilibrium (LD), i.e. correlated with QSNPs, using Plink, version 1.9, –indep-pairwise 500 50 0.6, corresponding to 500 kilobases, 50 SNPs, 0.6 r^2 threshold. After removing the QSNPs and SNPs in LD with them, the SNPs that remained were deemed associated with the common factors and retained for functional annotation.

Pathway analysis

Starting with the list of genes mapped using all 3 approaches in FUMA – positional, eQTL, and chromatin interactions – we used GeneSCF (Subhash and Kanduri 2016) to assign gene ontology (GO) pathway IDs to them and submitted the resulting list of GO IDs with the corresponding gene association P values to REVIGO Supek *et al.* (2011), which reduces redundancies and aids in classification and interpretation of biological pathways. For REVIGO, we specified the options: resulting list "Small (0.5)", values associated with GO terms represent "P value", remove obsolete GO terms "Yes", species "Homo sapiens (9606)", and semantic similarity measure "SimRel (default)". The resulting list cutoff 0.5 is the default similarity threshold which marks any pathways with a similarity score greater than that value as redundant, thereby shortening the list. REVIGO provides "uniqueness", which measures the extent to which a given term is an outlier when semantically compared to the complete list of submitted terms, and "dispensability", the extent to which a term is semantically close to other terms, based on semantic distance and association statistics. The term's "frequency" reflects the percentage of proteins assigned to it in the reference database (UniProt, https://www.uniprot.org/proteomes/UP000005640), such that higher frequency is an attribute of a more general term.

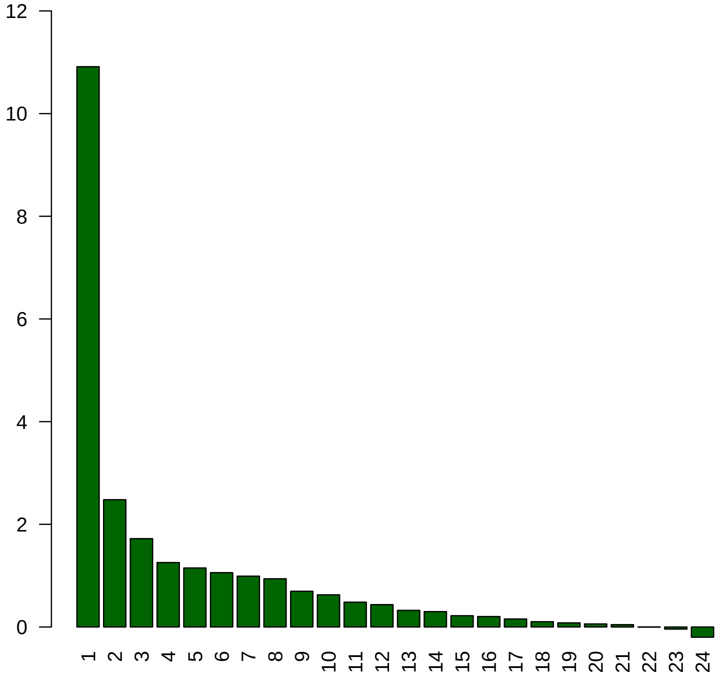
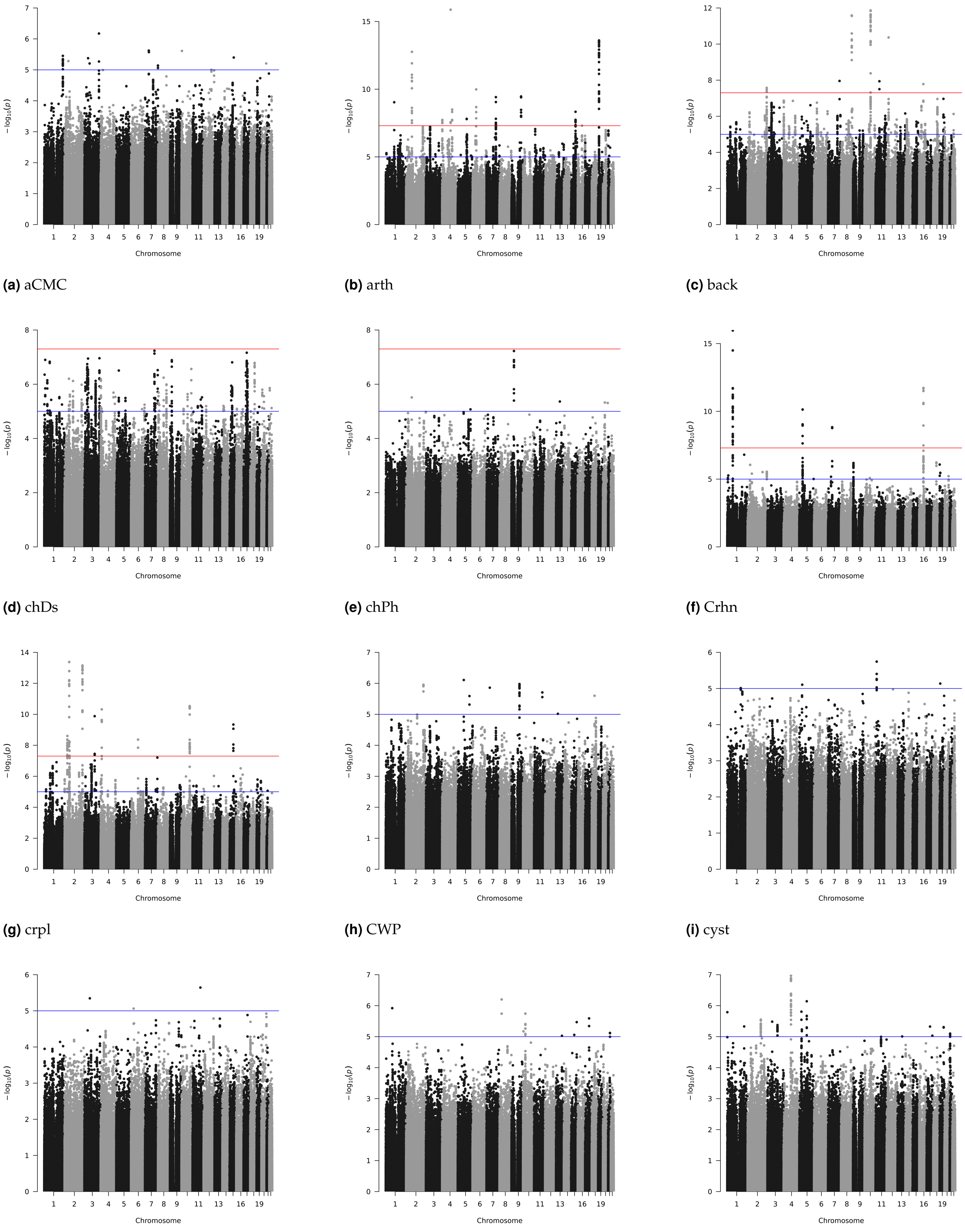


Figure S1 Scree plot for 24 pain conditions (all autosomes), obtained using exploratory factor analysis on genetic correlations.



(a) aCMC

(b) arth

(c) back

(d) chDs

(e) chPh

(f) Crhn

(g) crpl

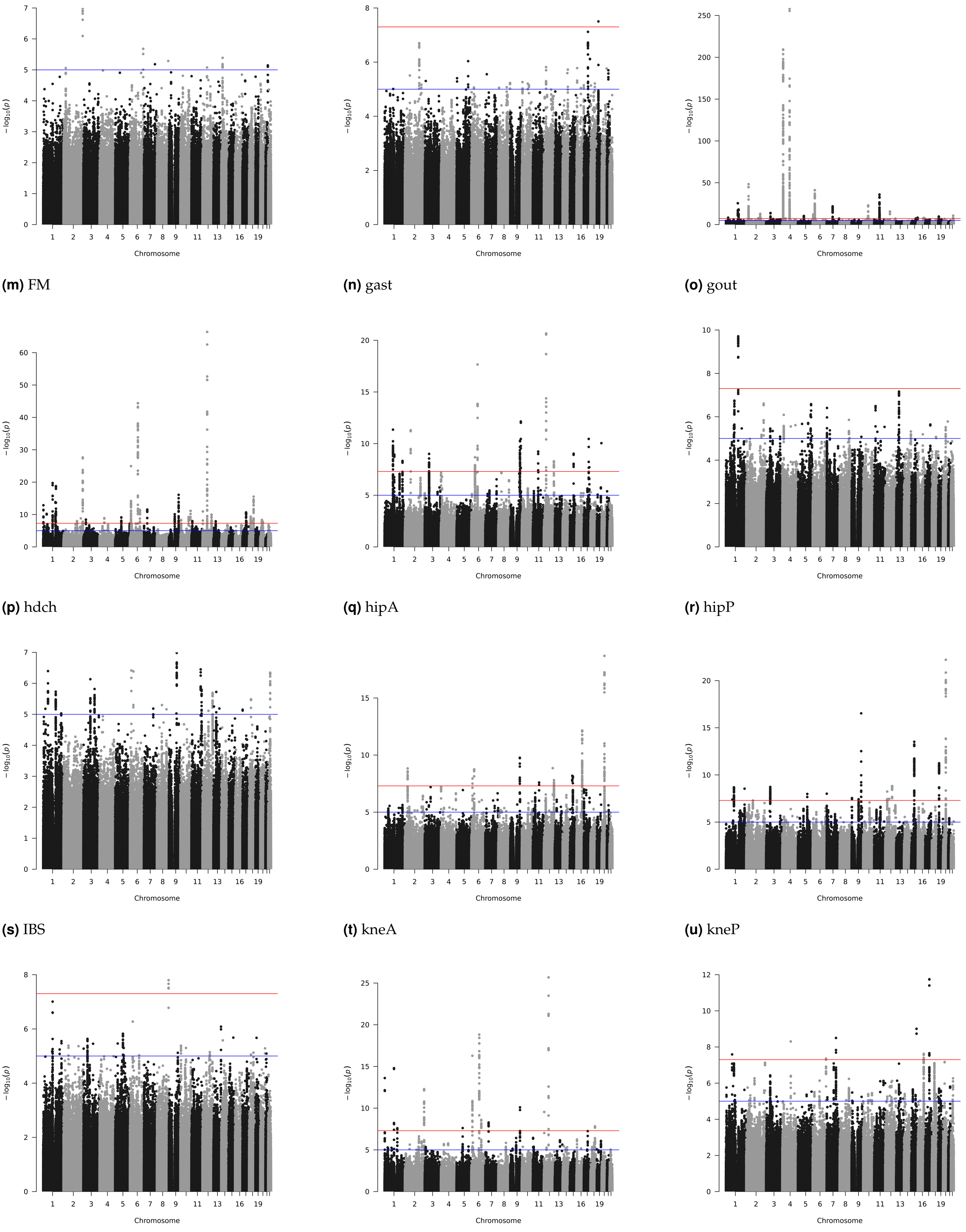
(h) CWP

(i) cyst

(j) dbNr

(k) enLL

(l) enth



(m) FM

(n) gast

(o) gout

(p) hdch

(q) hipA

(r) hipP

(s) IBS

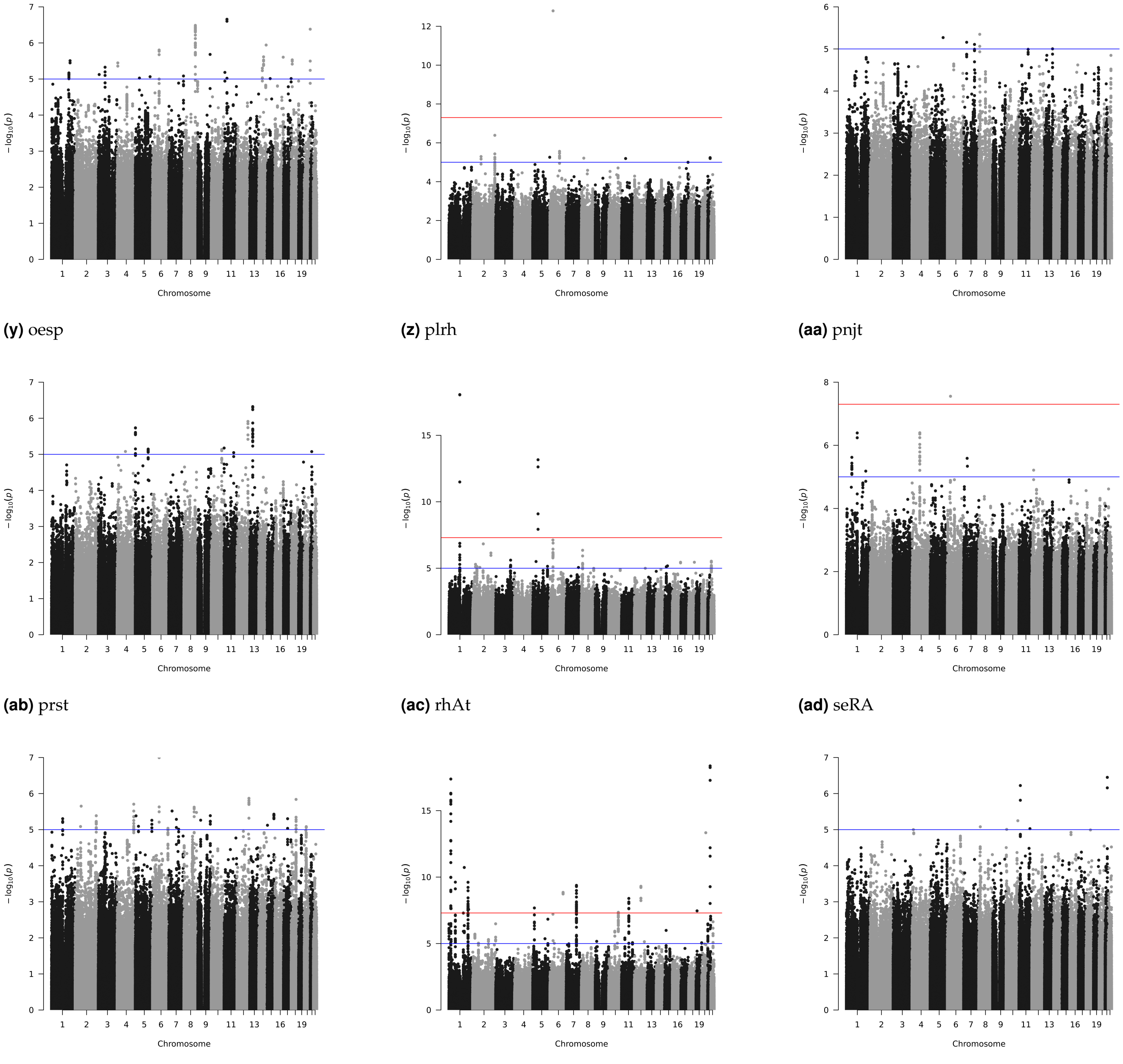
(t) kneA

(u) kneP

(v) legP

(w) mgrn

(x) neck



(y) oesp

(z) plrh

(aa) pnjt

(ab) prst

(ac) rhAt

(ad) seRA

(ae) stmP

(af) ulcC

(ag) urCl

Figure S2 Manhattan plots for 33 pain conditions. Condition definitions are in Table 1, and details are in Supplementary Table 1).

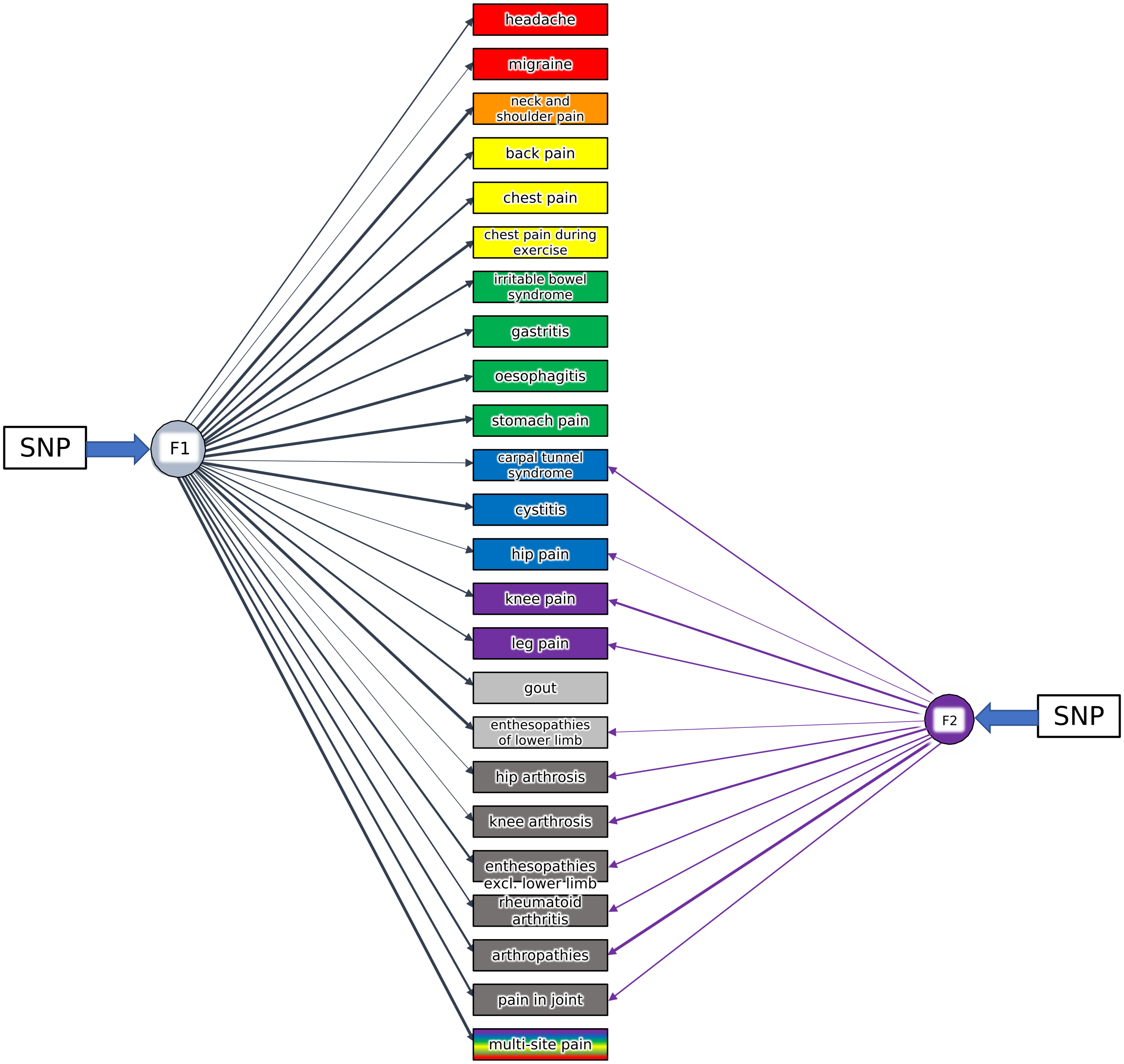


Figure S3 The EFA-CFA model for 24 pain conditions, Figure 2A, with SNP effects. The pathway coefficients for each SNP were estimated for both factors in factor genome-wide association study (GWAS) in GenomicSEM. (More information on all conditions in Supplementary Table 1)

Literature Cited

Bulik-Sullivan, B. K., P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, *et al.*, 2015 Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics **47**: 291–295.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, *et al.*, 2015 Second-generation plink: rising to the challenge of larger and richer datasets. Gigascience **4**: s13742–015.

Firth, D., 1993 Bias reduction of maximum likelihood estimates. Biometrika **80**: 27–38.

Grotzinger, A. D., M. Rhemtulla, R. de Vlaming, S. J. Ritchie, T. T. Mallard, *et al.*, 2019 Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nature human behaviour **3**: 513–525.

Huang, J., B. Howie, S. McCarthy, Y. Memari, K. Walter, *et al.*, 2015 Improved imputation of low-frequency and rare variants using the uk10k haplotype reference panel. Nature communications **6**: 1–9.

Huedo-Medina, T. B., J. Sánchez-Meca, F. Marin-Martinez, and J. Botella, 2006 Assessing heterogeneity in meta-analysis: Q statistic or i^2 index? Psychological methods **11**: 193.

Lee, S. H., N. R. Wray, M. E. Goddard, and P. M. Visscher, 2011 Estimating missing heritability for disease from genome-wide association studies. The American Journal of Human Genetics **88**: 294–305.

Mbatchou, J., L. Barnard, J. Backman, A. Marcketta, J. A. Kosmicki, *et al.*, 2021 Computationally efficient whole-genome regression for quantitative and binary traits. Nature genetics **53**: 1097–1103.

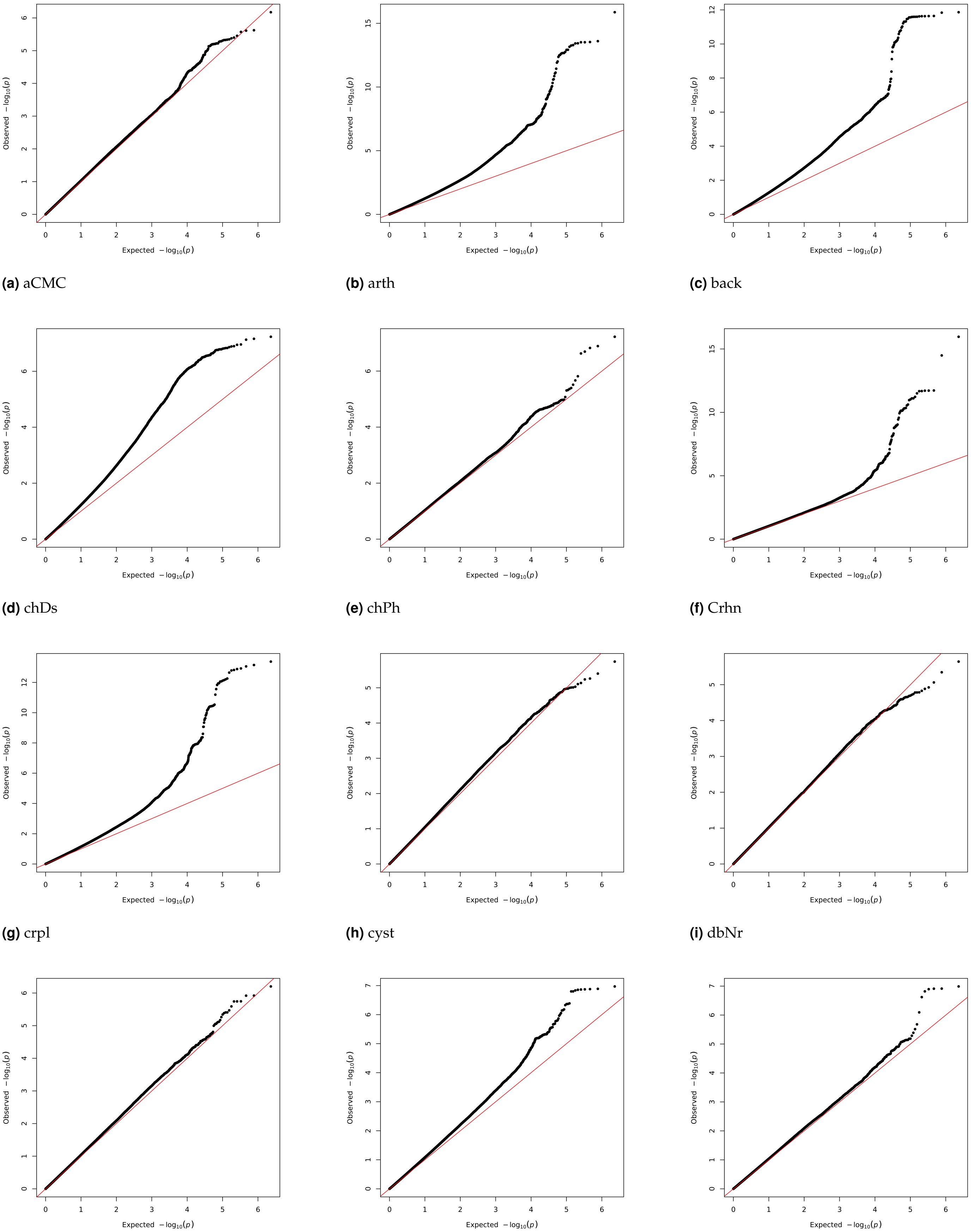
Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics **38**: 904–909.

Subhash, S. and C. Kanduri, 2016 Genescf: a real-time based functional enrichment tool with support for multiple organisms. BMC bioinformatics **17**: 1–10.

Supek, F., M. Bošnjak, N. Škunca, and T. Šmuc, 2011 Revigo summarizes and visualizes long lists of gene ontology terms. PloS one **6**: e21800.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 Gcta: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics **88**: 76–82.

Zaitlen, N. and P. Kraft, 2012 Heritability in the genome-wide association era. Human genetics **131**: 1655–1664.



(a) aCMC

(b) arth

(c) back

(d) chDs

(e) chPh

(f) Crhn

(g) crpl

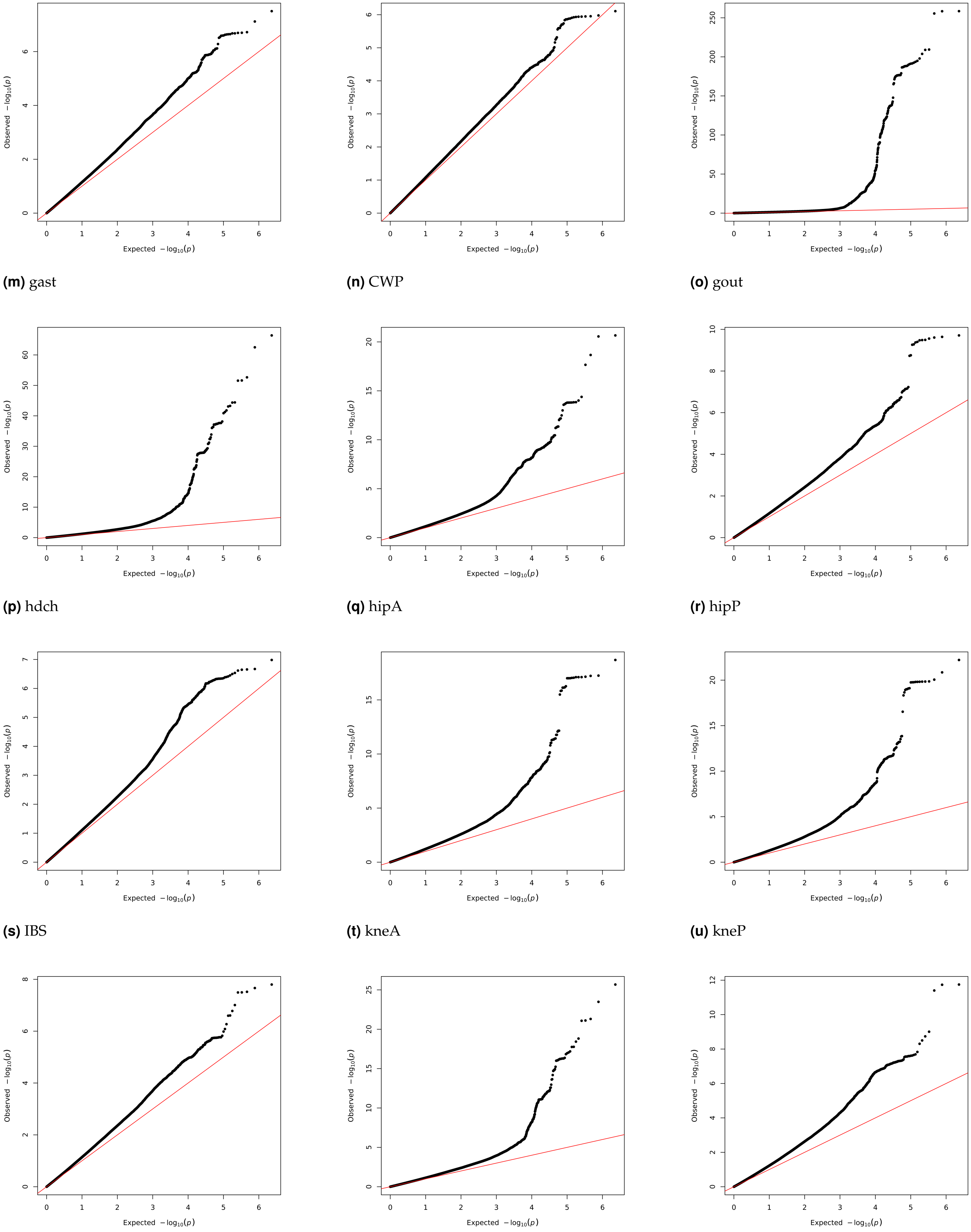
(h) cyst

(i) dbNr

(j) enLL

(k) enth

(l) FM



(m) gast

(n) CWP

(o) gout

(p) hdch

(q) hipA

(r) hipP

(s) IBS

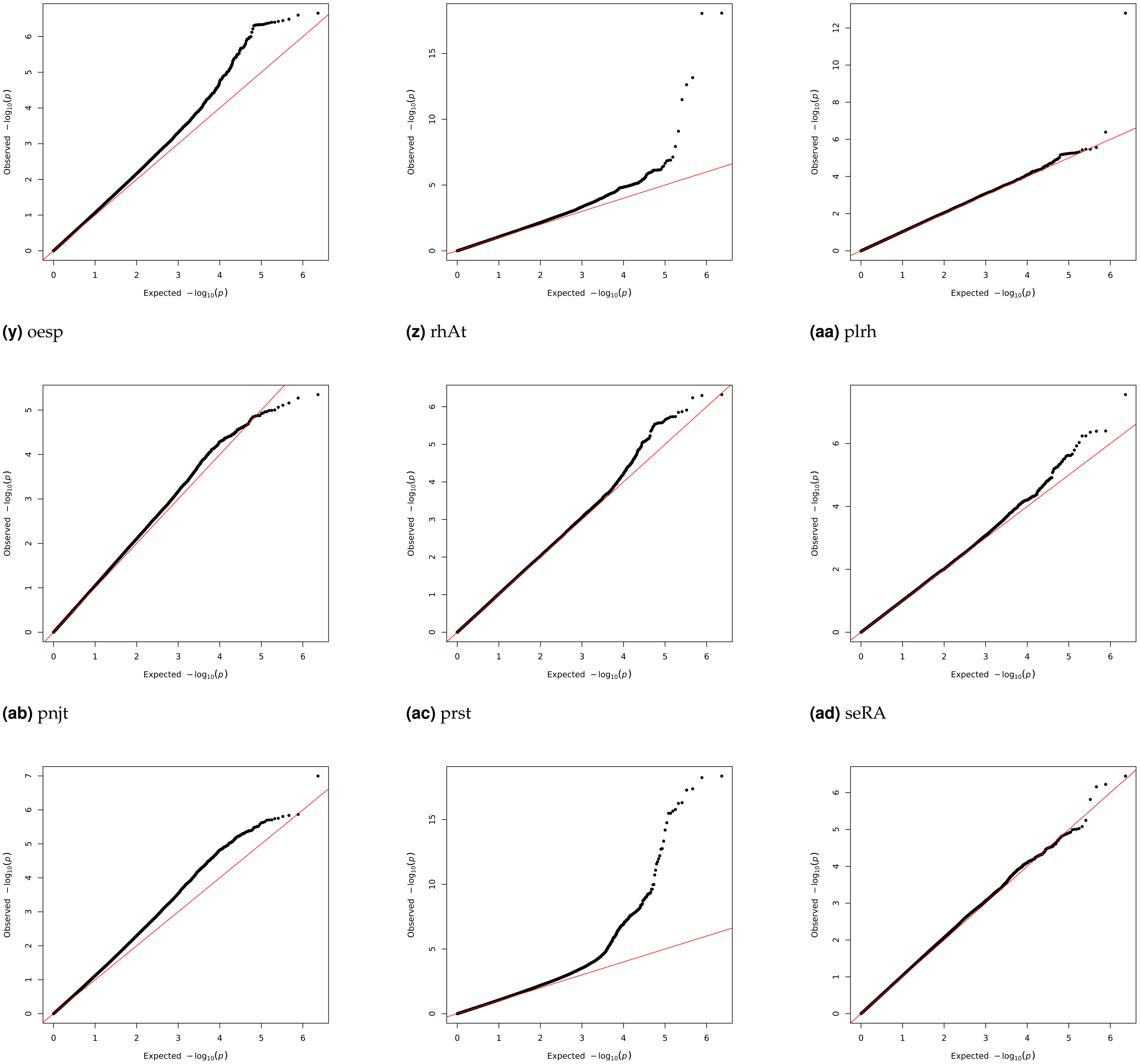
(t) kneA

(u) kneP

(v) legP

(w) mgrn

(x) neck



(y) oesp

(z) rhAt

(aa) plrh

(ab) pnjt

(ac) prst

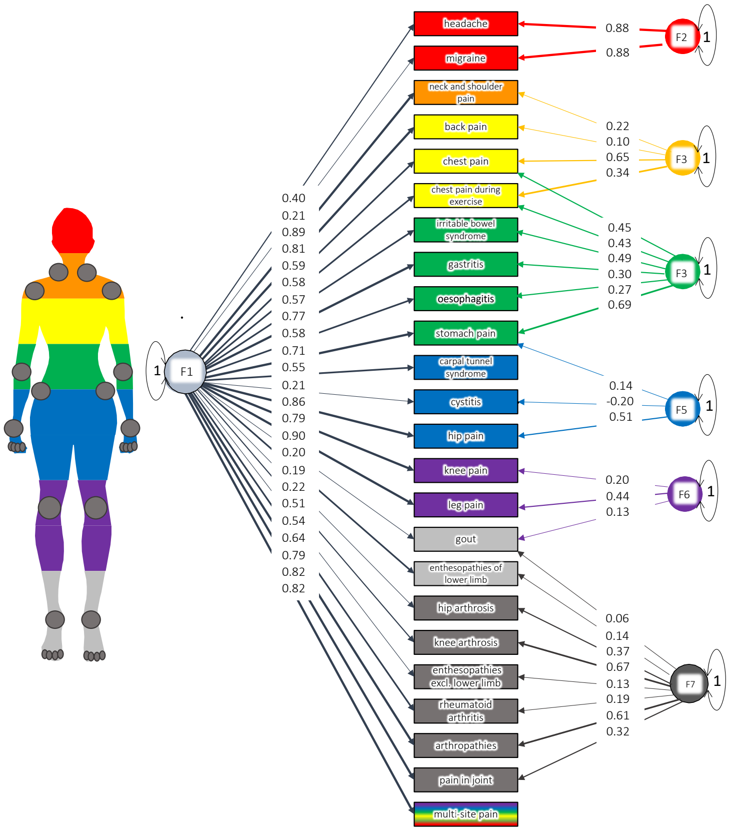
(ad) seRA

(ae) stmP

(af) ulcC

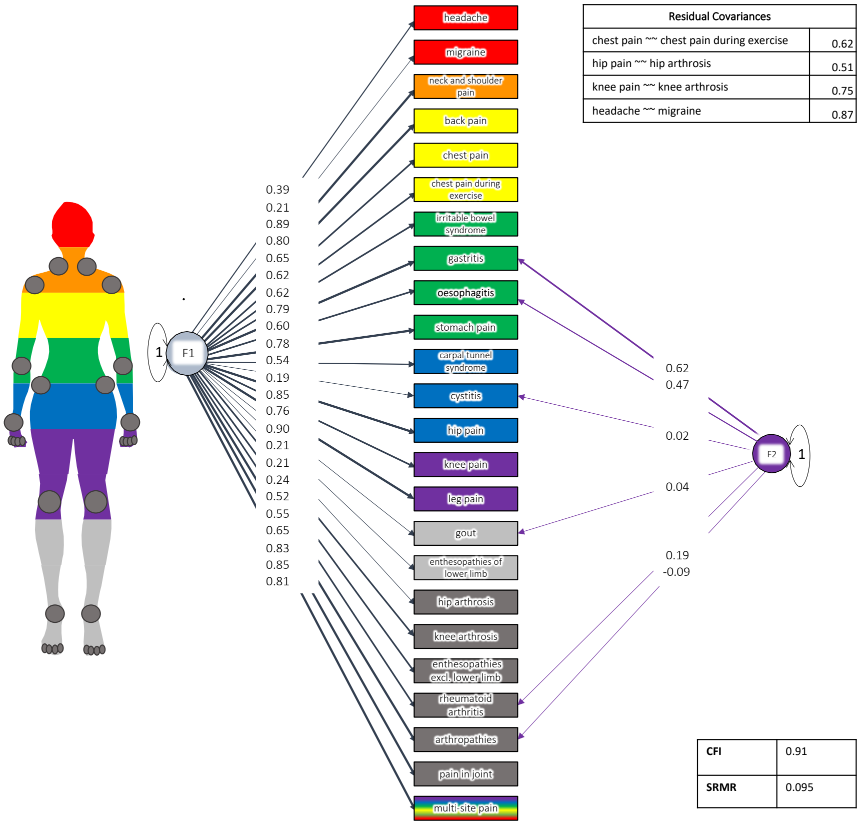
(ag) urCl

Figure S4 Quantile-quantile (QQ) plots for 33 pain conditions. Condition definitions are in Table 1, and details are in Supplementary Table 1).



CFI	0.87
SRMR	0.087

Figure S5 Structural equation model with hypothesis-driven anatomic groupings for 24 pain conditions. Factors: General (F1), Cranial (F2), Gastrointestinal (F3), Torso (F4), Pelvic (F5), Leg/Foot (F6) and Joint (F7). CFI, comparative fit index; SRMR, standardized root mean squared residual. All shown loadings are significant at $\alpha = 0.05$ except for all the conditions defining F6 "Leg/Foot": gout, knee pain, and leg pain; all the conditions defining F5 "Pelvic": cystitis, hip pain, and stomach pain; as well as enthesopathies of the lower limb and gout with F7 "Joint". (More information on all conditions in Supplementary Table 1).



Residual Covariances	
chest pain ~~ chest pain during exercise	0.62
hip pain ~~ hip arthrosis	0.51
knee pain ~~ knee arthrosis	0.75
headache ~~ migraine	0.87

CFI	0.91
SRMR	0.095

Figure S6 Structural equation model with hypothesis-driven etiologic groupings for 24 pain conditions. Factors: General (F1), Inflammatory (F2). CFI, comparative fit index; SRMR, standardized root mean squared residual. All loadings are significant at $\alpha = 0.05$, except arthropathies, rheumatoid arthritis, cystitis, and gout on the "Inflammatory" factor. (More information on all conditions in Supplementary Table 1).

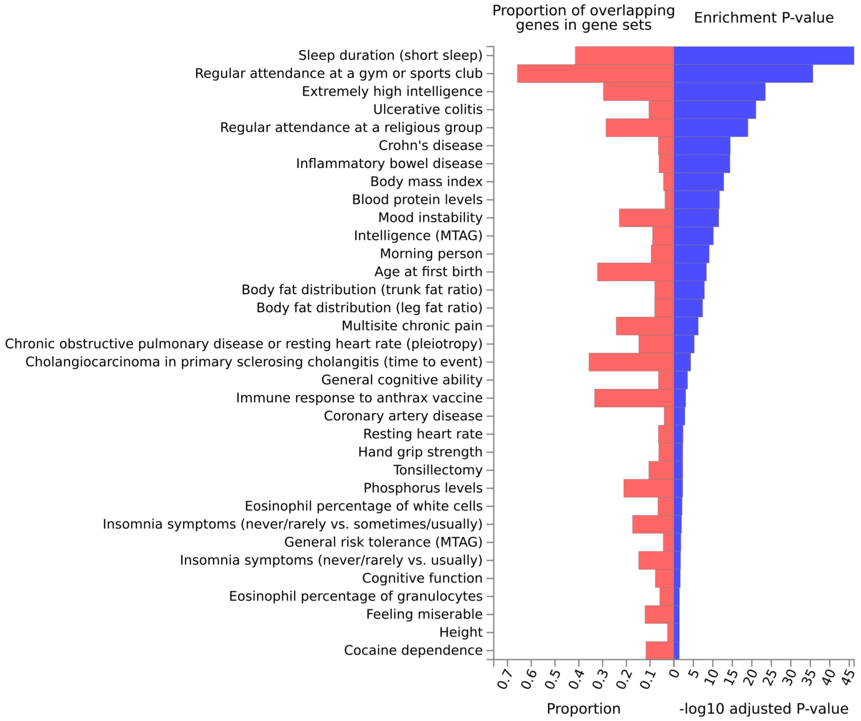
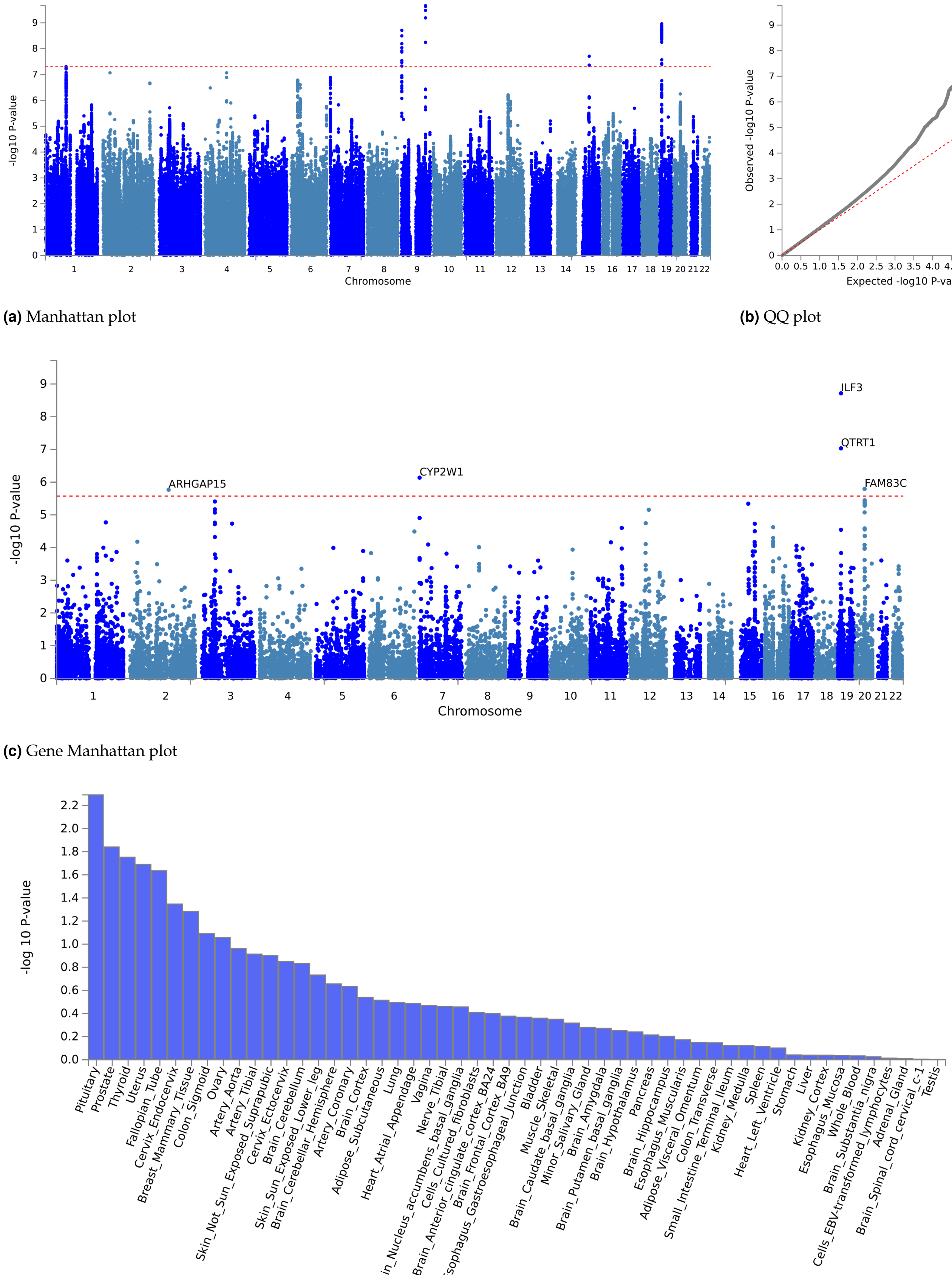


Figure S7 FUMA plot of overlap of F1 genes with genes reported in previously published genome-wide association studies (GWAS).



(a) Manhattan plot

(b) QQ plot

(c) Gene Manhattan plot

(d) 53 specific tissues

Figure S8 Genome-wide association study (GWAS) results for the musculoskeletal pain factor (F2). SNP Manhattan (a) and quantile-quantile, QQ, (b) plots for F2 GWAS. (c) Gene-based genome-wide association Manhattan plot, with the top 31 associated genes labelled. (d) Gene property analysis for association between factor GWAS gene effects and gene expression levels in 53 specific tissues from GTEx, version 8.

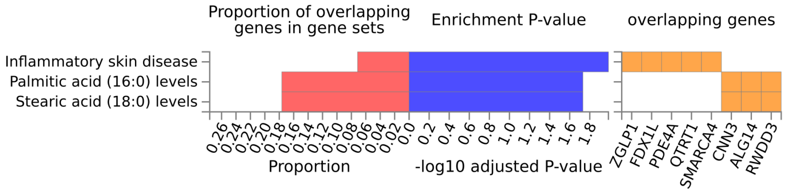


Figure S9 FUMA plot of overlap of F2 genes with genes reported in previously published genome-wide association studies (GWAS).