Supplementary Materials

Genome-wide analysis of 944,133 individuals provides insights into the etiology of hemorrhoidal disease

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

Table of Contents

METHODS ***4***

Histology of hemorrhoidal plexus **4**

Study cohorts and patients' material **5**

23andMe 5
UK Biobank (UKBB) 5
Estonian Genome Project of University of Tartu (EGCUT) 6
Michigan Genomics Initiative (MGI) 6
Genetic Epidemiology Research on Aging 6
German case-control cohort 7
The Trøndelag Health Study (HUNT) 7
Danish Blood Donor Study 8
Danish National Patient Registry 8
Hemorrhoidal tissue 9

Genotyping, quality control and genotype imputation of cohorts included in this study **9**

23andMe 9
UK Biobank 11
Estonian Genome Project of University of Tartu (EGCUT) 12
Michigan Genomics Initiative (MGI) 13
Genetic Epidemiology Research on Aging (GERA) 14
German case-control cohort 14
The Trøndelag Health Study (HUNT) 15
Danish Blood Donor Study 17

GWAS association analysis for discovery cohorts **18**

23andMe 18
UK Biobank (UKBB), Estonian Genome Project of University of Tartu (EGCUT), Michigan Genomics Initiative (MGI), Genetic Epidemiology Research on Aging (GERA) 19

GWAS meta-analysis across discovery cohorts **19**

Annotation of HEM GWAS risk loci and gene mapping **20**

Bayesian fine-mapping analysis **20**

Heritability analysis via linkage disequilibrium score regression (LDSC) **21**

Genome-wide pleiotropy analysis **21**

Tissue and pathway enrichment analyses **21**

Polygenic risk scores (PRS) analysis **22**

Phenome-wide association studies (PheWAS) **22**

Cross-trait analyses **23**

RNA library preparation and RNA-sequencing. **23**

Mapping and quality assessment of RNA-Seq data **24**

Gene signature-based determination of anal canal zones **24**

Differential gene expression analysis **25**

Identification and characterization of enriched co-expression modules **25**

ABO blood group analysis **26**

Fluorescence Immunohistochemistry **26**

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

***In silico* variant protein analysis 27**

Site-directed ANO1 mutagenesis and whole-cell electrophysiology 29

SUPPLEMENTARY FIGURES..... 31

Online Supplementary Figure S1. *Histological analysis of the anorectum in four different species.* 31

Online Supplementary Figure S2. *Suggested integrated model that summarizes the contemporary thinking on the pathophysiology of HEM (figure and legend are mainly taken from Figure 2 in Nikolaos Margetis' review[92]).* 33

Online Supplementary Figure S3. Schematic overview of the study workflow. 36

Online Supplementary Figure S4. *Quantile-quantile (QQ) plot of GWAS meta-analysis results.* 37

Online Supplementary Figure S5. *Regional association plots of HEM GWAS risk loci.* 50

Online Supplementary Figure S6: *Previously reported associations of HEM risk loci with other traits and diseases, clustered by biological areas.* 51

Online Supplementary Figure S7. *Gene set enrichment analyses of HEM genes.* 52

Online Supplementary Figure S8. *Gene signature-based determination of anal canal zones.* 53

Online Supplementary Figure S9. *ANO1 Alignment of TM4-5 and ANO1 structure.* 54

Online Supplementary Figure S10. F608S mutant of ANO1 has high instantaneous current but slow voltage-dependent activation and deactivation kinetics *in vitro*. 57

Online Supplementary Figure S11. *Sushi repeat-containing protein (SRPX) structure und alignment.* 59

Online Supplementary Figure S12. *ABO blood groups and HEM risk in UKBB and GERA.* 60

Online Supplementary Figure S13. *Immunohistochemistry for selected HEM candidate proteins.* 61

Online Supplementary Figure S14. *Graphical abstract of the study.* 63

SUPPLEMENTARY REFERENCES..... 64

Zheng T, *et al*. *Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

METHODS

Histology of hemorrhoidal plexus

For histologic examination and phylogenetic comparison of the hemorrhoidal plexus, formalin fixed anorectal specimens were obtained from *Homo sapiens*, *Gorilla gorilla gorilla*, baboon (*Papio anubis*), and mouse (10-week old male C57BL/6JRj mouse).

Human tissue was retrieved from a healthy donor (female, 54 years) who was recruited by the body donation program of the Institute of Anatomy, Kiel University. The donors had previously given written consent to the use of their samples for teaching and research purposes; the donors were free from diseases related to the gastrointestinal tract and the anorectum. The gorilla specimen comes from a 43-year-old female western lowland gorilla from Nuremberg Zoo (Germany). The animal had to be euthanized due to a terminal metastatic adenocarcinoma of the uterus. Rectum and surrounding tissue were removed during post-mortem examination four hours after death, cut and fixed in 10% neutral buffered formalin. Rectum samples from the baboon were taken from a ten year old male olive baboon kept at the German Primate Center Göttingen and included in a study authorized by the governmental veterinary authority, i.e. the Lower Saxony State Office for Consumer Protection (Food Safety Ref. No. 33.19-42502-04-18/3036 according to the regulations of the German Welfare Act and the European Directive 2010/63/EU on the protection of animals used for experimental and other scientific purpose). The rectal specimen was collected during routine necropsy following a standardized necropsy protocol and fixed in 10% buffered formalin.

All tissue samples were taken from the anal canal at the level of the hemorrhoidal plexus, dehydrated, embedded in paraffin wax, cut into sections (6 µm) and processed for hematoxylin-eosin and Azan stainings. The findings were evaluated and documented with a Keyence microscope (BZ-X800) using the integrated stitching tool BZ-X800 Analyser software version 1.1.1.8.

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

Study cohorts and patients' material

23andMe

The 23andMe study dataset contains participants drawn from the research participant base of the personal genetics company, 23andMe, Inc[1]. Genetic data and comprehensive phenotypic information from health surveys were available for 402,845 unrelated individuals of European ancestry. Study participants were divided into HEM cases and controls based on their self-completed HEM health questionnaires, resulting in 174,785 HEM cases and 228,060 controls in the current 23andMe GWAS. Demographic data of 23andMe samples are reported in **online supplementary table S1**. Participants provided informed consent and participated in the research online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services (E&I Review). The full GWAS summary statistics for the 23andMe discovery data set will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please visit https://research.23andme.com/dataset-access for more information and to apply to access the data.

UK Biobank (UKBB)

The UKBB is a large population-based study in the United Kingdom with extensive phenotypic and genotype data from approximately 500,000 participants[2]. Each individual underwent cognitive, physical assessment and sampling for DNA collection when enrolled, and health-related information was collected including data from their electronic health records (EHRs). The diagnoses in the EHRs are coded in the terminology of the International Statistical Classification of Diseases and Related Health Problems (ICD) terminology. For this GWAS study we included 408,592 individuals of European ancestry (self-reported "white" and of genetic Caucasian descent). Of these, 23,856 samples met our criteria for HEM cases (either ICD10 code I84 or ICD9 code 455 in the medical records). The other part of the cohort (n=384,736) served as study controls. The demographic data of the individuals are reported in **online supplementary table S1**. UKBB received ethical approval from the competent Research Ethics Committee (REC reference 11/NW/0382) and the project's Application ID is 31435.

Estonian Genome Project of University of Tartu (EGCUT)

The Estonian Biobank is a population-based cohort of the Estonian Genome Center at the University of Tartu (EGCUT), Estonia, with a current size of app. 200,000 participants aged over 18[3]. The whole project is conducted according to the Estonian Gene Research Act and all participants have signed the broad informed consent. Upon recruitment, the biobank participants filled out a detailed questionnaire, covering lifestyle, diet and clinical diagnoses (described by ICD10 codes). In this study, individuals with any entry of the ICD10 code for HEM (I84) were included as HEM cases. Further, we selected 30,441 controls with genome-wide data as study controls, resulting in 6,956 HEM cases and 30,441 population controls. The demographic data of the individuals are reported in **online supplementary table S1**. This study has been reviewed and approved by the Estonian Committee on Bioethics and Human Research.

Michigan Genomics Initiative (MGI)

The Michigan Genomics Initiative (MGI) is a longitudinal cohort of participants in Michigan Medicine, USA[4]. MGI participants were recruited primarily through surgical procedures at Michigan Medicine and gave consent for link their EHRs and genetic data for research purposes. We used a current data freeze of 40,000 European individuals for GWAS analysis. Of these, 4,539 HEM cases were defined based on a review of EHRs (either ICD10 code I84, ICD10-CM code K64 or ICD9 code 455). The rest of the cohort with genome-wide data was defined as study controls (n=35,338). The demographic data of the individuals are reported in **online supplementary table S1**. This study has been reviewed and approved by the Michigan Institutional Review Board.

Genetic Epidemiology Research on Aging

The Genetic Epidemiology Research on Aging (GERA) Cohort comprises more than 100,000 adults who are members of the Kaiser Permanente Medical Care Plan, Northern California Region (KPNC), USA. The health status of participants in the GERA cohort was assessed using EHRs collected at Kaiser Permanente's facilities in Northern California from January 1, 1995 to March 15, 2013. HEM cases (n=8,813) were those in which at least two ICD9 code diagnoses of HEM (ICD9 code 455) were recorded on separate days. Their genome-wide data were compared with those of the

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

remaining cohort (n=46,780) as controls. The demographic data of the individuals are reported in **online supplementary table S1**. The GERA data access was applied for on the dbGaP website (dbGaP Study Accession: phs000674.v3.p3) and the study was approved by the dbGap Access Review Committee.

German case-control cohort

Initiated by the Department of General and Thoracic Surgery and the biobank PopGen[5] of the Medical Faculty of Kiel University, Kiel, Germany, a cohort of HEM patients with symptomatic hemorrhoids and the need of invasive treatment was newly established. Between January 2016 and December 2017, individuals with a prior diagnosis of high-grade hemorrhoids were identified based on the medical records of five hospitals and practices in the North German region using German procedural codes (OPS-301 by German Institute for medical Documentation and Information). The main inclusion criteria were the need for hemorrhoidectomy or invasive treatment (rubber band ligation, sclerotherapy) on high grade hemorrhoidal disease, verified by DRG-code (Diagnosis related Groups). Patients receiving exclusively conservative treatment were not included in this study as the aim was to recruit patients with a strong phenotype of advanced hemorrhoidal disease. The cohort included 1,007 patients undergoing surgical/invasive treatment of a high grade hemorrhoidal disease. In total, 1,144 cases and 2,740 controls were available for PRS analysis (section **Polygenic risk score (PRS) analysis, Methods**). The demographic data of the individuals are reported in **online supplementary table S1**. The study protocol was approved by the ethics committee (ref: A156/03-1/15) of the Medical Faculty of Kiel University and written informed consent was obtained from all study participants.

The Trøndelag Health Study (HUNT)

The Trøndelag Health Study (HUNT) is a large population-based cohort from the county Nord-Trøndelag in Norway. All residents in the county, aged 20 years and older, have been invited to participate. Data was collected through three cross-sectional surveys, HUNT1 (1984-1986), HUNT2 (1995-1997) and HUNT3 (2006-2008), and has been described in detail previously[6], with the fourth survey recently completed (HUNT4, 2017-2019). All genotyped participants have signed a written informed consent regarding the use of data from questionnaires, biological samples and linkage to other registries for research purposes. Cases were defined as having

an ICD10 K64 diagnosis and the reminder of the cohort were used as controls. In total, 977 cases and 68,314 controls were available for PRS analysis (section **Polygenic risk score (PRS) analysis, Methods**). The demographic data of the individuals are reported in **online supplementary table S1.**

Danish Blood Donor Study

The Danish Blood Donor Study (DBDS) is a large prospective cohort of nation-wide Danish blood donors (n=56,397) and comprises both extensive phenotype data as well as genome-wide genotyping data[7, 8]. HEM cases are defined using the ICD-8 code 455 or ICD-10 codes I84 or K64, resulting in 1,754 cases in the DBDS cohort as registered in the National Patient Registry. In total, 1,754 cases and 54,643 controls were available for genome-wide polygenic risk score (PRS) analysis. The demographic data of the individuals are reported in **online supplementary table S1.** This study was approved according to the Danish Blood Donor study protocol (ref: 1700407) as a part of "Genetics of healthy ageing and specific diseases among blood donors".

Danish National Patient Registry

The Danish National Patient Registry (DNPR) is a population-wide registry containing all diagnoses made in hospitals in Denmark from 1977 to 2018 and includes more than 8 million patients. The diagnoses in the registry are coded in the terminology of the International Statistical Classification of Diseases and Related Health Problems (ICD) 8th Revision (1997-1993) or 10th Revision (1994-2018) terminology. All patients with a hemorrhoid disease code in the disease registry were identified. In the ICD-8 period patients with 'Hemorrhoids' are recognized using the code 455. As the ICD-10 code for hemorrhoids changed in 2013, we combined patients diagnosed with 'Hemorrhoids' (ICD-10: I84) from 1994-2012 and patients diagnosed with 'Hemorrhoids and perianal venous thrombosis' (K64) from 2013-2018. Information about drugs administered in hospitals is available for more than 1.6 million patients in the period 2006-2016 and is defined using the Anatomical Therapeutic Chemical (ATC) Classification System of the WHO. We integrated data from two different electronic medication modules corresponding to the administrative databases for hospital internal drug consumption from two health regions of Denmark (Capital Region and Region Zealand): OPUS-medicin and Elektronisk patient medicinering[9].

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

The demographic data of the individuals are reported in **online supplementary table S1**. This DNPR study has been approved by the Danish Data Protection Agency, Copenhagen (ref: FSEID-00003092, FSEID-00003724 and 3-3013-1731/1).

Hemorrhoidal tissue

A group of 38 individuals undergoing surgery for hemorrhoids (n=20; cases) and anal fissures (n=18; controls) was included in this study. Hemorrhoidal tissue samples were obtained either by Milligan Morgan open hemorrhoidectomy or by stapled hemorrhoidopexy for grade (Goligher) 3 and 4 hemorrhoids. Approximately 1cm³ of hemorrhoidal tissue was obtained from Milligan-Morgan-specimens just above the dentate line. In hemorrhodopexy patients, approximately 1cm³ of biopsies were taken from the "doughnut" tissue at 3 o'clock in the prone position. Healthy hemorrhoid tissue samples were taken from adjacent zones (1-2 cm above the dentate line) of anal fissures. Clinical and demographic data for the sampled individuals are listed in **online supplementary table S1**. This study was approved by the bioethical committee of medical faculty, University Hospital Schleswig-Holstein Kiel, Germany. All participants provided written informed consent.

Genotyping, quality control and genotype imputation of cohorts included in this study

23andMe

DNA extraction and genotyping were performed on saliva samples by National Genetics Institute (NGI), a CLIA licensed clinical laboratory and a subsidiary of Laboratory Corporation of America. Samples had been genotyped on one of four genotyping platforms. The V1 and V2 platforms were variants of the Illumina HumanHap550+ BeadChip, including about 25,000 custom SNPs selected by 23andMe, with a total of about 560,000 SNPs. The V3 platform was based on the Illumina OmniExpress+ BeadChip, with custom content to improve the overlap with the V2 array, with a total of about 950,000 SNPs. The V4 platform is a fully custom array, including a lower redundancy subset of V2 and V3 SNPs with additional coverage of lower-frequency coding variation, and about 570,000 SNPs. Samples that failed to reach 98.5% call rate were re-analyzed. Individuals who repeatedly failed

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

analyses were recontacted by the 23andMe customer service to provide additional samples.

For GWAS quality control (QC) analysis, we limited participants to a set of individuals with ≥97% European descent, determined by analysis of local ancestry[10]. In brief, the algorithm initially partitions phased genomic data into short windows of about 100 SNPs. Within each window, a support vector machine (SVM) was used to classify each haplotype into one of 31 reference populations. SVM classifications were translated into a hidden Markov model (HMM) that takes into account switch errors and incorrect assignments and reports probabilities for each reference population in each window. Finally, simulated admixed individuals were used to recalibrate the HMM probabilities so that the reported assignments are consistent with the simulated admixture ratios. Reference population data was derived from public datasets (the Human Genome Diversity Project, HapMap, and 1000 Genomes) and from 23andMe customers who reported having four grandparents from the same country. For each analysis, a maximal set of unrelated individuals was selected using a segmental identity-by-descent (IBD) estimation algorithm[11]. Individuals were identified as related if they shared more than 700 cM IBD, including regions where the two individuals share either one or both genomic segments identical-by-descent. This degree of relatedness (about 20% of the genome) corresponds approximately to the expected minimum proportion between cousins and first-degree cousins in an outbred population. SNPs deviating from Hardy-Weinberg equilibrium ($P<10^{-20}$), having a call rate <95%, or with large discrepancies in allele frequency compared to the European 1000 Genomes reference data were excluded. SNPs with large differences in allele frequency (chi squared $P<10^{-15}$) were identified by computing a 2x2 table of allele counts for European 1000 Genomes samples and 2000 randomly sampled 23andMe customers of European ancestry.

Genotype data were imputed using the September 2013 1000 Genomes Phase1 reference haplotypes[12]. Phasing and imputation was performed separately for the data of each genotyping platform. Phasing was performed with a phasing tool, Finch, developed internally by 23andMe, which implements the Beagle haplotype graph-based phasing algorithm[13] and which was modified to separate the steps of constructing the haplotype graph and phasing. Finch extends the Beagle model to

allow genotyping errors and recombination events to handle cases where there are no consistent paths through the haplotype graph for the individual to be phased. From a representative sample of genotyped individuals, haplotype graphs were generated for European and non-European samples for each 23andMe genotyping platform. Subsequently, an out-of-sample phasing of all genotyped individuals against the corresponding graph was performed. In preparation for imputation, the phased chromosomes were divided into segments of no more than 10,000 genotyped SNPs, with overlaps of 200 SNPs. Each phased segment was imputed against all-ethnicity 1000 Genomes haplotypes (excluding monomorphic and singleton sites) using Minimac2[14], using 5 rounds and 200 states for parameter estimation. For the X chromosome, we created separate haplotype graphs for the non-pseudoautosomal region and each pseudoautosomal region, and these regions were separately phased. Then we imputed males and females together using Minimac2, as for the autosomes, and treated males as homozygous pseudo-diploids for the non-pseudoautosomal region. After QC and genotype imputation a total of 7,024,410 SNPs with imputation quality score Rsq>0.8 and minor allele frequency (MAF) >1% in 174,785 cases and 228,060 controls were available for association analysis.

UK Biobank

DNA samples were genotyped on custom UK Biobank (UKBB) arrays. 408,951 individuals from UKBB were genotyped for 825,927 variants using a custom Affymetrix UK Biobank Axiom Array, and 49,626 individuals were genotyped for 807,411 variants using a custom Affymetrix UK BiLEVE Axiom Array chip from the UK BiLEVE study[15], which is a subset of UKBB.

All SNPs were subjected to quality control (QC): checks, such as deviations from Hardy-Weinberg equilibrium ($P<10^{-5}$), batch and plate effects, sex effects, and array effects across control replicates. The SNPs that failed call rate <0.95 were set to missing for all individuals. The QC was performed centrally for each sample tested for heterozygosity and missing rates. Samples with excessive relatedness (>10 suspected third-degree relatives) were excluded. Full details of the QC of the genetic data performed centrally by UK Biobank are available in the original publication[2]. To identify sample outliers (i.e. subjects of non-Europeans ancestry), we performed

principal component analysis (PCA) with FlashPCA2[16]. PCA revealed no non-European ancestry outliers. Genotypes of 408,592 UKBB participants with European ancestry (self-reported "white" and genetic Caucasian) were used after QC. Of these, 23,856 samples satisfied our criteria for being HEM cases (either ICD10 code I84 or ICD9 code 455 in medical records) and the remainder of the cohort (n=384,736) served as study controls.

Genetic variants were imputed centrally by UKBB using IMPUTE4[2] and a reference panel that merged the UK10K and 1000 Genomes Phase 3 panel as well as the Haplotype Reference Consortium (HRC) panel[2]. After QC and genotype imputation, a total of 9,572,556 SNPs with an imputation quality score INFO>0.8 and MAF >1% in 23,856 cases and 384,736 controls were available for association analysis.

Estonian Genome Project of University of Tartu (EGCUT)

The Estonian cohort originates from the population-based biobank of the Estonian Genome Project of University of Tartu (EGCUT). The EGCUT project has been conducted according to the Estonian Gene Research Act and all participants have signed the broad informed consent. The current cohort size is about 200,000 aged 18 years and older, which is very close to the age distribution in the adult Estonian population. Subjects were recruited by general practitioners and doctors in hospitals. The persons who visited the general practitioner's practices or hospitals were selected at random. Each participant completed a computer assisted personal interview during 1-2 hours in a doctor's office, which included personal data (place of birth, place(s) of living, nationality etc.), genealogical data (family history, three generations), educational and occupational history and lifestyle data (physical activity, dietary habits, smoking habits, alcohol consumption, women's health, quality of life). Diseases were defined according to the ICD10 coding. Illumina Human CoreExome, OmniExpress, 370CNV BeadChip and Illumina Global Screening Array (GSA) arrays were used for genotyping.

QC included filtering based on sample call rate (<98%), heterozygosity (> mean ± 3SD), genotype and phenotype sex discordance, cryptic relatedness (IBD >20%) and outliers of European ancestry based on a multidimensional scaling (MDS) analysis

including 210 HapMap reference samples[17]. SNP QC included testing for call rate (<99%), MAF (<1%) and extreme deviation from Hardy–Weinberg equilibrium ($P<10^{-4}$).

Pre-phasing was performed using SHAPEIT2[18]. Genotype imputation was performed using the Estonian-specific reference panel[19] and IMPUTE2[20] with default parameters. After QC and genotype imputation, a total of 7,462,975 SNPs with imputation quality score INFO>0.8 and minor allele frequency (MAF) >1% in 6,956 cases and 30,441 controls were available for association analysis.

Michigan Genomics Initiative (MGI)

DNA samples were genotyped on custom Illumina HumanCoreExome v12.1 bead chips. Samples were excluded if they exhibited (1) a calling rate < 99%, (2) an estimated contamination > 2.5% (BAF Regress)[21] or (3) deviating sex information if the derived sex did not match the self-reported gender. Variants were excluded if they (1) deviated from Hardy-Weinberg equilibrium ($P_{HWE}<10^{-5}$), (2) had a calling rate < 99%. After quality control, 392,323 polymorphic variants were kept in the following analyses. Next, we estimated the pair-wise relationship of the samples using the software KING[22] and we limited the dataset within a subset of individuals without first- or second-degree relationship. The genetic ancestry of the samples were derived by projecting the principal components of the samples onto that of the Human Genome Diversity Project (HGDP) reference panel (938 unrelated individuals)[23]. Principal component analysis was performed using PLINK v1.90[24], including a subset of LD pruned variants ($r^2 < 0.5$) with MAF >1% shared between the HGDP reference and the MGI data. We retained only samples of recent European ancestry (defined as samples that fell into a circle around the center of the reference HGDP populations in the PC1 versus PC2 space).

Genotype imputation was conducted using the Haplotype Reference Consortium (HRC) panel and the Michigan Imputation Server[25]. After QC and genotype imputation, a total of 6,536,218 SNPs with imputation quality score Rsq>0.8 and MAF >1% in 4,539 cases and 35,338 controls were available for association analysis.

Genetic Epidemiology Research on Aging (GERA)

DNA samples were collected from participants of the Genetic Epidemiology Research on Aging (GERA) cohort and genotyped on high-density custom designed Affymetrix Axiom arrays. Genetic variants with >5% of missing data, MAF <1% in either disease sets or in controls or deviating from Hardy-Weinberg equilibrium ($P<10^{-5}$) were excluded. Samples with >2% missing data and overall increased/decreased heterozygosity rates were removed. For robust duplicate/relatedness testing (IBS/IBD estimation) and population structure analysis, a pruned subset of 144,799 independent SNPs was used. Pair-wise percentage IBD values were computed using PLINK. By definition, Z0: P(IBD=0), Z1: P(IBD=1), Z2: P(IBD=2), Z0+Z1+Z2=1, and PI_HAT: P(IBD=2) + 0.5 * P(IBD=1) (proportion IBD). One individual (the one showing greater missingness) from each pair with PI_HAT>0.1875 was removed. To identify sample outliers (i.e. subjects of non-Europeans ancestry), we performed principal component analysis (PCA) using the smartpca program[26], based on a set of 144,799 "high-performing" markers after exclusion of SNPs that had an r^2 value greater than 0.5, were within 5 MB of each other, within the MHC region, had a call rates lower than 99.5% and that were located in regions with inversions on chromosomes 8p23 and 17q21.

Genotype data were pre-phased with SHAPE-IT v2.5[18], and then imputed with IMPUTE2 v2.3.1[27] using the 1000 Genomes Phase 3 data as a reference panel. After QC and genotype imputation, a total of 6,897,996 SNPs with imputation quality score INFO>0.8 and MAF >1% in 8,813 cases and 46,780 controls were available for association analysis.

German case-control cohort

DNA samples were genotyped using Illumina's Global Screening Array version 1.0. Patients with a reported "migration background" were excluded. 3,505 eligible patients were contacted by their treating physician by mail. The initial submission rate was 40%. After consent to participate, the Popgen Biobank sent a study kit with a questionnaire on clinical and socio-demographic characteristics and a set of blood tubes, so that a blood sample could be collected at the family doctor's office and returned to the study center. In addition, a subset of study participants were asked to

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

complete a comprehensive questionnaire on their dietary habits and usual physical activity. Patients were excluded from the study in the absence of informed consent/blood sample or after withdrawal of consent.

Variants that had >2% missing data, a minor allele frequency <0.1% in either of the different disease sets or in controls, had different missing genotype rates in affected and unaffected individuals ($P_{Fisher}<10^{-5}$) or deviated from Hardy-Weinberg equilibrium ($P_{HWE}<10^{-5}$) were excluded. Samples that had >2% missing data and overall increased/decreased heterozygosity rates (with an average marker heterozygosity of ±5 s.d. away from the sample mean) were removed. For robust duplicate/relatedness testing (IBS/IBD estimation) and population structure analysis, we used a pruned subset of 100,596 independent SNPs (MAF>0.05) SNPs excluding X- and Y-chromosomes, SNPs in LD (leaving no pairs with $r^2>0.2$), and 11 high-LD regions as described by Price *et al.[28]*. Pair-wise percentage IBD values were computed using PLINK2. By definition, Z0: P(IBD=0), Z1: P(IBD=1), Z2: P(IBD=2), Z0+Z1+Z2=1, and PI_HAT: P(IBD=2) + 0.5 * P(IBD=1) (proportion IBD). One individual (the one showing greater missingness) from each pair with PI_HAT>0.1875 was removed. To identify sample outliers (i.e.subjects of non-Europeans ancestry), we performed principal component analysis (PCA) with FlashPCA2[16], on the basis of a set of 100,596 independent markers (described above).

Genotype imputation was conducted using the Haplotype Reference Consortium (HRC) panel and the Sanger Imputation Service[25]. After QC and genotype imputation, a total of 7,117,385 SNPs with imputation quality score INFO>0.8 and MAF >1% in 1,144 cases and 2,740 population controls were available for association analysis.

The Trøndelag Health Study (HUNT)

DNA was extracted from whole blood from HUNT2 and HUNT3. Genotyping was a research collaboration between researchers from the Norwegian University of Science and Technology (NTNU) and the University of Michigan. Each individual with a DNA sample of an appropriate DNA concentration was selected for genotyping. Samples

were taken at random and genotyped in batches. All genotyping was performed at the Genomics-Core Facility (GCF) at NTNU.

Genotype quality control and genotype imputation were conducted by the K.G. Jebsen Center for Genetic Epidemiology, Department of Public health and Nursing, Faculty of Medicine and Health Sciences, NTNU. In total, DNA from 71,860 HUNT samples was genotyped using one of three different Illumina HumanCoreExome arrays: HumanCoreExome12 v1.0, HumanCoreExome12 v1.1 and UM HUNT Biobank v1.0. Samples were excluded if they did not achieve a 99% call rate, had a contamination >2.5% as estimated with BAF Regress[29], had large chromosomal copy number variants, a lower call rate of a technical duplicate pair and a twin pair, gonosomal constellations other than XX and XY, or whose derived sex was inconsistent with the reported sex. Samples that passed quality control were analyzed in a second round of genotype calling following the Genome Studio quality control protocol described elsewhere[30]. Genomic position, strand orientation and the reference allele of genotyped variants were determined by aligning their probe sequences against the human genome (Genome Reference Consortium Human genome build 37 and revised Cambridge Reference Sequence of the human mitochondrial DNA; http://genome.ucsc.edu) using BLAT. Variants were excluded if their probe sequences could not be perfectly mapped to the reference genome, cluster separation was <0.3, Gentrain score was <0.15, showed deviations from Hardy-Weinberg equilibrium in unrelated samples of European ancestry with *P*-value <0.0001), their call rate was <99%, or another assay with higher call rate genotyped the same variant. Ancestry of all samples was inferred by projecting all genotyped samples onto top principal components of the Human Genome Diversity Project (HGDP) reference panel (938 unrelated individuals; downloaded from http://csg.sph.umich.edu/chaolong/LASER/)[23, 31], using PLINK v1.90. Recent European ancestry was defined for samples that fell into an ellipsoid spans European populations of the HGDP panel. The different arrays were harmonized by reducing them to a set of overlapping variants and excluding variants that had frequency differences >15% between data sets, or that were monomorphic in one data set and had a MAF >1% in another data set. The resulting genotype data were phased using Eagle2 v2.3[32].

Imputation was performed on the 69,716 samples of recent European ancestry using Minimac3 (v2.0.1, http://genome.sph.umich.edu/wiki/Minimac3)[33] with default settings (2.5 Mb reference based chunking with 500kb windows) and a customized Haplotype Reference consortium release 1.1 (HRC v1.1) for autosomal variants and HRC v1.1 for chromosome X variants[25]. The customized reference panel represented the merged panel of two reciprocally imputed reference panels: (1) 2,201 low-coverage whole-genome sequences samples from the HUNT study and (2) HRC v1.1 with 1,023 HUNT WGS samples removed before merging. After QC and genotype imputation, over 24.9 million SNPs with imputation quality score R^2 >=0.3 in 977 cases and 68,314 population controls were available for association analysis.

Danish Blood Donor Study

DNA samples were genotyped at deCode genetics, Iceland, using Illumina's Global Screening Array as described elsewhere[7]. Details on genotype quality control and imputation are available in Hansen et al., 2019[7]. First- and second-degree relatives were excluded from the analysis. The phenotypic data used in this project includes sex, age, self-reported BMI and selected diagnoses from the Danish National Patient Registry. Participant where classified as having HEM using the ICD-8 code 455 or ICD-10 codes I84 or K64 from the National Patient Registry, resulting in the identification of 1,754 HEM cases in the DBDS cohorts.

The DBDS Genomic Consortium is represented by the following scientists: Andersen Steffen, Department of Finance, Copenhagen Business School, Copenhagen, Denmark; Banasik Karina, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; Brunak Søren, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; Burgdorf Kristoffer, Department of Clinical Immunology, Copenhagen University Hospital, Copenhagen, Denmark; Erikstrup Christian, Department of Clinical Immunology, Aarhus University Hospital, Aarhus, Denmark; Hansen Thomas Folkmann, Danish Headache Center, department of Neurology Rigshospitalet, Glostrup, Denmark; Hjalgrim Henrik, Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark; Jemec Gregor, Department of Clinical Medicine, Sealand University hospital, Roskilde, Denmark;

Jennum Poul, Department of clinical neurophysiology at University of Copenhagen, Copenhagen, Denmark; Johansson Per Ingemar, Department of Clinical Immunology, Copenhagen University Hospital, Copenhagen, Denmark; Nielsen Kasper Rene, Department of Clinical Immunology, Aalborg University Hospital, Aalborg, Denmark; Nyegaard Mette, Department of Biomedicine, Aarhus University, Denmark; Mie Topholm Bruun, Department of Clinical Immunology, Odense University Hospital, Odense, Denmark; Pedersen Ole Birger, Department of Clinical Immunology, Naestved Hospital, Naestved, Denmark; Petersen Mikkel, Department of Clinical Immunology, Aarhus University Hospital, Aarhus, Denmark; Sørensen Erik, Department of Clinical Immunology, Copenhagen University Hospital Copenhagen, Denmark; Ullum Henrik, Department of Clinical Immunology, Copenhagen University Hospital, Copenhagen, Denmark; Werge Thomas, Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Copenhagen University Hospital, Roskilde, Denmark; Gudbjartsson Daniel, deCODE genetics, Reykjavik, Iceland; Stefansson Kari, deCODE genetics, Reykjavik, Iceland; Stefánsson Hreinn, deCODE genetics, Reykjavik, Iceland; Þorsteinsdóttir Unnur, deCODE genetics, Reykjavik, Iceland.

GWAS association analysis for discovery cohorts

23andMe

For comparisons between cases and controls, association test results were performed by logistic regression analysis assuming additive allelic effects. For tests using imputed data, imputed allelic dosages were used rather than best-guess genotypes. Age, biological sex, BMI, the top five principal components from principal component analysis (to account for potential residual population structure) as well as indicators for genotype platforms (to account for genotype batch effects) were included as covariates in the regression analysis. The association test *P*-value was computed using a likelihood ratio test. Results for the X chromosome are computed similarly, with male genotypes coded as if they were homozygous diploid for the observed allele. For chromosome X association analysis, haplotypic allele calls in males outside pseudoautosomal regions (PAR) are converted to homozygous calls by doubling the haplotypic allele (assuming inactivation of large parts of one of the two female X chromosomes[34] and sex was used as a covariate for association testing.

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

Association summary statistics were adjusted for an estimated genomic control inflation factor λ_{GC}=1.200.

UK Biobank (UKBB), Estonian Genome Project of University of Tartu (EGCUT), Michigan Genomics Initiative (MGI), Genetic Epidemiology Research on Aging (GERA)

For each individual case-control data set, association testing was performed using a linear mixed model (LMM) under an additive genetic model for all measured and imputed genetic variants in dosage format using BOLT-LMM[35] (UKBB, GERA) or SAIGE[36] (MGI). Within association analysis, we adjusted for the following covariates: sex, age, BMI (available for UKBB and GERA), the top ten principal components from principal component analysis and a binary indicator variable for genotyping platform (e.g. UKBB Axiom Array vs. UK BiLEVE Axiom Array) to account for the different genotyping chips. For GWAS data set from EGCUT, association testing was carried out with EPACTS [https://github.com/statgen/EPACTS], adjusting for age, sex, binary indicator variable for genotyping platform and top four principal components from principal component analysis. For chromosome X association analysis, see text above. The genomic control inflation factors for UKBB, EGCUT, MGI and GERA were λ_{GC}=1.0966, 1.0263, 0.9822 and 0.9541, respectively. For GWAS meta-analysis across discovery cohorts (23andme, UKBB, EGCUT, MGI and GERA).

GWAS meta-analysis across discovery cohorts

Prior to GWAS meta-analysis, separate GWAs analyses for discovery cohorts were performed either via logistic regression or mixed linear model association analysis using BOLT-LMM[35] or SAIGE[36] including sex, age, BMI (where available), top principle components (PCs) from principal component analysis (PCA; to control for potential residual population stratification) and genotyping array (if relevant) as covariates**.** File-level QC of the five individual GWAS summary statistics and meta-level QC from discovery cohorts were carried out using the R package “EasyQC” (v9.2)[37]. In short, the QC process verified data integrity and harmonized both SNP marker IDs and allele coding across the datasets. We only included markers with imputation quality metrics (INFO or Rsq)>0.8 and MAF>1% in the meta-analysis.

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

Markers with deviating allele frequency (difference >20% from the Haplotype Reference Consortium (HRC) genome reference panel v1.1 comprising 32,488 reference individuals of European ancestry[25]) were removed along with indels and multi-allelic markers. The resulting summary statistics of the five discovery cohorts (with a total of 218,920 HEM cases and 725,213 controls) were meta-analyses via fixed-effect meta-analysis based on METAL's inverse-variance weighted approach[38]. We used the generally accepted threshold of 5×10^{-8} for meta-analysis *P*-values to define statistical significance ($P_{Meta}<5\times10^{-8}$). Genome-wide summary statistics of our analyses are publicly available through our web browser (http://hemorrhoids.online) and have been submitted to the European Bioinformatics Institute (www.ebi.ac.uk/gwas) under accession number GCST90014033.

Annotation of HEM GWAS risk loci and gene mapping

We used independent computational pipelines for the functional annotation of GWAS meta-analysis results, using FUnctional Mapping and Annotation of Genome-Wide Association Studies (FUMA v1.3.5)[39], Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT)[40], and Multi-marker Analysis of GenoMic Annotation (MAGMA, also implemented in FUMA)[41]. The 102 newly identified genome-wide significant risk loci were defined in FUMA (using default parameters and eQTL databases including GTEx v7) as non-overlapping genomic regions that extend a linkage disequilibrium (LD) window ($r^2 = 0.6$) around each lead SNP association signal with $P_{Meta}<5\times10^{-8}$. Annotation of these regions with FUMA resulted in 712 transcripts mapped to risk loci (415 positional and 562 eQTL candidates), while 217 genes were identified using DEPICT, and 255 in MAGMA independent gene-based tests, bringing the total of non-redundant HEM candidate genes to 819 (**online supplementary table S7**). Regional association plots of all 102 risk loci were generated using LocusZoom[42].

Bayesian fine-mapping analysis

A Bayesian fine-mapping analysis was carried out using FINEMAP[43] for the 102 genome-wide significant risk loci in order to calculate the posterior inclusion probability (PIP) for each lead SNP as causal and to determine a credible set for each risk locus, i.e. a minimum set of variants containing all causal variants with certainty ≥0.95%. As

input for fine-mapping we extracted all genetic variants located within the 102 risk loci (as defined by FUMA) and calculated the local LD structure using genotypes from UKBB samples (**online supplementary table S1**) serving as a reference population.

Heritability analysis via linkage disequilibrium score regression (LDSC)

Narrow-sense heritability (h^2_{SNP}) for HEM and the genetic correlation (r_g) between HEM and other traits were estimated using LD score regression, as implemented in the online platform CTG-VL[44]. This platform integrates summary statistics of 1,387 traits from multiple resources such as UKBB, the Psychiatric Genomics Consortium (PGC) and the Genetic Investigation of ANthropometric Traits (GIANT) consortium. Significantly correlated pairs of traits were reported after FDR correction for multiple comparisons at α=0.05.

Genome-wide pleiotropy analysis

We conducted cross-phenotype association analysis based on subsets (ASSET) methodology[45] across association summary statistics from diverticular disease[46], irritable bowel syndrome (IBS)[47] and HEM to identify shared risk loci. The subset-based meta-analysis (SBM) method maintains similar type-I error rates as for standard meta-analysis and identifies the best subset of non-null studies while in parallel accounting for multiple-hypothesis testing and shared individuals. This method offers a substantial power increase (sometimes approaching between 100-500%)[45] compared to standard univariate meta-analysis approaches, where the (heterogeneous) effect of a specific SNP is not exclusively restricted to a single phenotype. Under the assumption that association signals from shared risk loci based on positional overlap are tagging same causal variant for different phenotypes, the SBM approach improves power compared to standard fixed-effects meta-analysis methodology.

Tissue and pathway enrichment analyses

Gene-set and tissue-specific enrichment analyses (respectively GSEA and TSEA) of HEM genes were carried out using integrated default pipelines in FUMA, and DEPICT implemented in the CTG-VL platform[44],[40]. HEM gene lists were derived from three

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

alternative approaches including positional and/or eQTL mapping in FUMA, MAGMA gene-based analyses (also implemented n FUMA), and DEPICT functional annotations, and teste against Gene Ontology (GO) terms and 30 GTEx v7 general tissue types. Statistical significance was defined using $P_{\text{Benjamini-Hochberg}}<0.05$.

Polygenic risk scores (PRS) analysis

The analysis of polygenic risk scores (PRS) was performed on the basis of a pruning and thresholding approach, using the *P* value and LD-driven clumping procedure as implemented in PRSice-2[48]. Effect estimates and corresponding standard errors from GWAS meta-analysis results were used as the base dataset to generate weights over a range of *P* values (0.5 to 5×10^{-8}) and r^2 0.1 LD thresholds, with the most appropriate thresholds selected as those that include SNPs with the highest Nagelkerke's R^2 value. The selected model was then applied to the QCed genetic datasets from the German case-control cohort, HUNT and DBDS, respectively. Logistic regression was used to test HEM PRS distribution in cases and controls, taking into account sex, age, BMI and the top 10 PCs from PCA. For HUNT and DBDS we also studied HEM prevalence across PRS percentile distributions. PRSs were binned into percentiles and HEM prevalence from the top 5% of PRS distribution was compared to the reminder of the population in a logistic regression adjusting for sex, age, BMI and the top 10 PCs. Additional analyses were performed to evaluate the relationship between HEM PRS and age at diagnosis (Spearman's correlation test) and need for invasive treatments (number of surgeries and/or rubber-band ligation; tested with linear regression correcting for sex, age, BMI and the top 10 PCs).

Phenome-wide association studies (PheWAS)

For each of the 102 GWAS risk loci, we queried the lead SNP and its LD proxies ($r^2>0.8$, from 1000 Genomes Project samples of European descent) using PhenoScanner v2[49], and manually inspecting the GWAS catalog[50] and GWAS ATLAS[51]. Only genome-wide significant associations ($P<5\times10^{-8}$) were taken into account, and those from GWAS ATLAS were collapsed by trait categories and plotted with the R package "ggforce" into an alluvial diagram.

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

Cross-trait analyses

Traits genetically correlated with HEM (from LDSC analyses) were tested for their prevalence in HEM patients vs controls in UKBB and DNPR. In UKBB, we derived the ICD10 diagnoses from data-fields “41202” (primary diagnosis) and “41204” (secondary diagnosis), self-reported medical conditions from data-field “20002”, and self-reported medication use from data-field “20003”. Differential prevalence was tested using a logistic regression model adjusted for sex and age, including FDR correction for multiple comparisons. For DNPR, a previously published method[52] was used to identify diseases that significantly co-occur more often with HEM diagnoses. Each combination of pair-wise disease co-occurrences was compared to a comparison group matched by sex, age, type of hospital encounter and week of discharge. The relative risk (RR) is used to evaluate the strength of the correlation between significant disease pairs (disease A followed by disease B and *vice versa*). Here, we have used this method to evaluate the temporal co-occurrence of selected diseases and medications with the HEM diagnosis in the DNPR, including FDR correction for multiple comparisons.

RNA library preparation and RNA-sequencing.

The RNA-Sequencing (RNA-Seq) libraries were prepared from 20 ng of total RNA from freshly frozen tissue extracted with the mirVana miRNA Isolation Kit according to the manufacturer’s protocol (Ambion). The NEXTFLEX Combo-Seq Kit (Perkin Elmer) was used to generate combined mRNA and microRNA libraries following manufacturer’s instructions. In short, poly-A-tailed RNA species were reverse transcribed to generate DNA:RNA duplexes whose RNA molecules were specifically sheared by RNase H, resulting in RNA fragments containing 5’-monophosphate and a 3’-hydroxyl groups. These mRNA fragments were 3’-polyadenylated together with small RNAs and then 5’ 4N adapters were ligated to their 5’ ends. Finally, first strand synthesis followed by PCR amplification were used to add sequences required for Illumina sequencing. The generated RNA libraries were quality-controlled using the Agilent 2200 TapeStation (Agilent Technologies), randomized and then deep sequenced (5 samples per lane), 1x50bp using the Illumina HiSeq 4000 platform.

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

Mapping and quality assessment of RNA-Seq data

The sequenced reads were demultiplexed and obtained as fastq files for each sample. Data pre-processing, quality control, mapping to genome (build hg38) and transcriptome annotation (miRBase v21, Ensembl 83) were performed using the exceRpt[53] pipeline. More precisely, reads were trimmed for 3' adapter sequences, 4N nucleotides at 5' end and low-quality bases (<Q20). The trimmed sequences shorter than 15 bp were discarded and only high-quality reads were then mapped to genome (with minimum sequence match of 15 nucleotides), annotated and quantified. All RNA-Seq libraries were quality controlled for library size (>10M of mapped reads), transcriptome genome ratio (> 0.95) and outliers for number of detected unique genes and microRNAs (< Q1−1.5 IQR). Low abundant gene-level and microRNA arm level counts that were expressed below 0.1 RPM in less than 25% of the samples per trait were removed from downstream analyses. The generated quality-controlled counts and raw sequencing reads have been deposited at NCBI Gene Expression Omnibus (GEO)[54] under the accession number GSE154650.

Gene signature-based determination of anal canal zones

Histologically, the anal canal can be divided into three zones according to the epithelial lining. The upper part is of the mucosal type (intestinal) and the lower part is of the squamous keratinized (anoderm), while the middle part, where the epithelium varies, is called the anal transitional zone[55, 56]. Due to the gradient nature of the anal canal epithelium, keratinocyte and sebocyte marker gene signatures from the xCell catalog[57] were used to discriminate the different histological zones. More specifically, the quality-controlled gene counts were normalized using the variance stabilizing transformation (VST) implemented in the DESeq2 R package[58]. The normalized gene counts were then ranked according to their expression level using the rank() function from the base R package and submitted to single sample gene set enrichment analysis (ssGSEA)[59] implemented in the GSVA R package[60]. The obtained normalized enrichment scores (NES) of keratinocytes and sebocytes were used to cluster samples into 3 groups (in accordance to the number of histological zones) by employing the base R function kmeans() with k=3 and nstart=20 as parameters. The obtained clusters were assigned to histological zones by the relative abundance of keratinocytes and sebocytes (i.e. sebum-producing epithelial cells), and

further confirmed by the expression levels of previously defined marker genes, including *KRT4*, *KRT8*, *KRT13* and *KRT20[56, 61]* genes (**online supplementary figure S8**). Multidimensional scaling (MDS) analysis using Spearman's rank correlation distance (1-correlation coefficient) was performed on VST normalized expression data and was used to explore the results.

Differential gene expression analysis

The quality-controlled count data were further analyzed using edgeR[62] workflow for differential expression analysis. Negative binomial generalized log-linear models were fitted to the trimmed mean of M-values (TMM) normalized count data of HEM genes using glmFit() function with trended dispersion estimates and the offsets for GC-content correction generated by EDASeq (default parameters). The glmLRT() function was used to calculate log-likelihood-ratio statistics and *P*-values of differential expression. The models were adjusted for BMI and histological zones of anal canal (see the previous paragraph). The nominal *P*-values were corrected for multiple testing according to Benjamini and Hochberg. Transcripts with an FDR corrected *P*-value < 0.05 and a $\log_2$ fold change > 0.5 (in either direction) were considered to be significantly differentially expressed.

Identification and characterization of enriched co-expression modules

Weighed gene co-expression network analysis of hemorrhoid-specific tissue was performed using the automated WGCNA[63] pipeline implemented in the CEMiTool[64] R package. The quality-controlled and VST normalized data (36,342 genes in 20 samples) was used to calculate signed scale-free topology overlap matrix, which was subsequently used to define gene co-expression modules in an unsupervised manner. More specifically, Pearson correlation coefficients for each gene-gene comparison (including miRNAs) were used to calculate adjacencies defined as following: $a_{ij} = |0.5 + 0.5 \times cor(x_i, x_j)|^{\beta}$, where x_i and x_j are expression values of i^{th} and j^{th} genes and where β is a soft threshold power based on scale-free topology, which was identified by employing pickSoftThreshold() function from the WGCNA R package. The generated adjacencies were then used to compute topological overlap measures (TOM) and their dissimilarity measures (1-TOM) were further used for

average linkage hierarchical clustering and dynamic tree cutting (cutoff value of 0.995) to identify gene co-expression modules. Each gene co-expression module contained a minimum of 50 genes and was summarized into eigengene, which is the first principal component of their expression values. Highly similar modules were identified by correlation of their eigengenes (>0.7 Pearson's r) and merged together. The intramodular connectivity of each gene was measured by Pearson's correlation of module eigengene and its expression value. The top 10% of genes having the highest connectivity values were considered as being module hub genes (central nodes in the scale-free network). A Fisher's exact test was used to identify modules with significantly (P_{FDR}< 0.05) overrepresented in HEM genes. The ClusterProfiler[65] R package was used to identify gene ontology (GO) terms "biological process" pathway enrichments for HEM-significant modules.

ABO blood group analysis

The association between ABO blood types and HEM was tested on 408,592 and 55,593 individuals from UKBB and GERA, respectively. We first imputed ABO blood group information individually based on genotypes at the *ABO locus* on chromosome 9q34.2. We extracted the genotypes of three SNPs: rs8176719, rs41302905 and the adjacent rs8176747 and inferred blood group status based on these SNPs as previously described[66]. Next, the risk of HEM was assessed on samples from each blood groups of the ABO blood group system ("A", "B", "AB" and "O"). An association test based on logistic regression was employed to test for significant HEM association for each of the four blood groups, adjusting for sex, age, BMI and the top 10 PCs from PCA. FDR correction was applied for multiple testing.

Fluorescence Immunohistochemistry

Fluorescence immunohistochemistry was performed as previously described[46, 67]. Briefly, anorectal specimens were fixed in 4% paraformaldehyde for 24 hours. Paraffin-embedded tissue sections were pre-treated with citrate buffer and primary antibodies were incubated overnight. Used primary and secondary antibodies are listed in **online supplementary table S13**. All antibodies were diluted in antibody diluent (ThermoFisher Scientific). Nuclei were counterstained with DAPI (Roche, Mannheim, Germany). Image acquisition was performed on a fluorescence inverted

microscope (Axiovert 200 M, Zeiss, Gottingen, Germany) coupled to an AxioCam MR3 camera (Zeiss) using Axiovision software (version 4.7, Zeiss).

In silico variant protein analysis

To construct a first hypothetical model of whether *SRPX* and *ANO1* missense lead variants (shown in red in **figure 1**) are likely to interfere with functionally active domains at the protein level, we conducted protein domain analyses for SRPX and ANO1.

ANO1 (also TMEM16A) is an anion channel protein that enables the passive flow of Cl anions through the membrane as a result of increased intracellular Ca^{2+} levels. The decrease in an anion flow occurs over time after prolonged stimulation eventually leads to complete desensitization to saturated Ca^{2+}. In addition to elevated Ca-levels, ANO1 function is regulated by the PIP2 (Phosphatidylinositol(4,5)-bisphosphate) signal lipid which binds at the cytoplasmic membrane interface[68]. The Interaction with PIP2 has been shown to slow down the ANO1 regulatory process, probably by hindering the gradual collapse of the ion conduction pore[69].

The ANO1 protein functions as a homodimer, with each subunit consisting of ten transmembrane helices and its own anion conduction pore (**online supplementary figure S9**) which is composed of helices 3-7 and contains a conserved Ca^{2+} binding site[70, 71]. Ion flow through the pore is made possible by local structural rearrangements that open the channel in response to Ca^{2+} binding[70].

The variant F608S is located at the beginning of the transmembrane helix 5, i.e. near the cytoplasmic interface (**online supplementary figure S9**). Although helix 5 is part of the ion conduction pore, the sidechain of F608 points in the opposite direction to the dimer interface and is located near the predicted PIP2 binding residues. Adjacent K609 forms a stabilizing salt bridge with E594 in the TM4-TM5 linker which is conserved in all members of the TMEM16 protein family. Mutation of this salt bridge results in a rapid Ca^{2+} desensitization, similar to a direct mutation of the predicted PIP2 binding residues[69].

F608 and its sequential and structural neighbors are conserved among ANO1 orthologs (**online supplementary figure S9**). The variant causes a change from the aromatic and very hydrophobic phenylalanine to the smaller and polar/hydrophilic serine. All members of the TMEM16 superfamily conserved a non-polar residue at this

position, suggesting that the polar sidechain of the serine may cause a structural conflict in the region. The variant could interfere with the K609-E594 salt bridge which stabilizes the PIP2 binding. F608S may thus interfere with the PIP2 binding and consequently accelerates ANO1 degradation, similar to the rapid desensitization that was demonstrated by mutational analyses of basic amino acids in the vicinity and the salt bridge[69].

The SRPX (also DRS, ETX1, SRPX1) variant rs35318931 causes a Ser413Phe exchange at the C-terminal domain of unknown function (**online supplementary figure S11**). The protein is further composed of three Sushi domains, and one HYR domain. Sushi domains are components involved in extracellular protein-protein interactions and are often found in complement control proteins[72]. The HYR (hyalin repeat) domain is predicted to contribute to cell adhesion since the domain enables the hyalin protein to bind to the receptor[73]. The SRPX C-terminal domain is a phylogenetically widespread protein domain that is well-conserved in vertebrates (**online supplementary figure S11**) and also in many bacteria, and has been named the DUDES domain (DRO1-URB-DRS-Equarin-SRPX)[74]. Protein structural analyses assign a thioredoxin-like fold to this domain, although the location of potential functional cysteines seem unique for SRPX and SRPX2 proteins[75]. Therefore, the conserved structural core allows fold recognition, but the lack of suitable structural templates including loops and termini complicates in-silico functional prediction for SRPX (**online supplementary figure S11**). SRPX was originally identified as a tumor suppressor[76] and, in this context, to the induction of apoptosis[77] and downregulation of glucose metabolism via Lactate dehydrogenase-B[78]. Proteomics studies found SRPX expression in the extracellular matrix (ECM) of different tissues (lung[79], cartilage[80] and colon and liver[81]) and is upregulated in the ECM during cardiac remodeling[82]. Further, SRPX was also shown to interact with PELO at the actin cytoskeleton[83].

Other members of the DUDES protein family were shown to localize in the extracellular matrix, e,g, SRPX2 in brain[84], equarin in chick lens[85]. CCDC80 is a remote homologous that binds activated JAK2 and is consequently more abundant in the extracellular matrix. JAK2-binding was also detected by the paralog SRPX2, and interaction is therefore also predicted for SRPX[86]. CCDC80 is composed of three DUDES domains, that are independently able to bind JAK2, assuming the SRPX

DUDES domain is responsible for protein association with the ECM. The ECM provides structural integrity for tissues, and involves in cell differentiation, activation and migration. HEM tissue is less stable and show abnormalities in the ECM collagen composition (compared to healthy tissue[87]).

The variant Ser413Phe locates at the beginning of strand 3 in the central beta sheet. The preceding loop is highly variable among homologs[75] but the following strand is one of the best conserved regions within the protein family, including an invariant F414. The change from the polar and small amino acid serine to the larger, aromatic and hydrophobic phenylalanine potentially destabilizes the domain structure due to its location adjacent the conserved hydrophobic core of the protein fold.

The SRPX domain structure was derived from the UniProt database and by search against the NCBI Conserved Domains Database (CDD). SRPX and SRPX2 protein sequences were derived from UniProt, Ensembl and the consensus sequence of pfam13778 from the CDD. Sequence alignments were conducted using Muscle. The sequence alignment was visualized using JalView applying the Clustal coloring scheme. Protein sequence identifiers UniProt or Ensembl: SRPX: human, P78539; mouse, Q9R0M3; cow, F1MQX1; zebrafish, Q58ED3; xenopus tropicalis, ENSXETT00000018780.4. SRPX2: human, O60687; mouse, Q8R054; cow, Q5EA25; zebrafish, E7F8X0, xenopus tropicalis, ENSXETT00000014699.4.

The structure-based alignment for modeling the SRPX C-terminal domain of unknown function (DUF4174/pfam13778, 332-451) is based on secondary structure predictions, structural alignments of two templates (PDBs 3drn/chain A, 3cmi/chain A) and multiple sequence alignment including the consensus sequence of pfam13778. Structural models of SRPX and ANO1 were visualized using PyMOL.

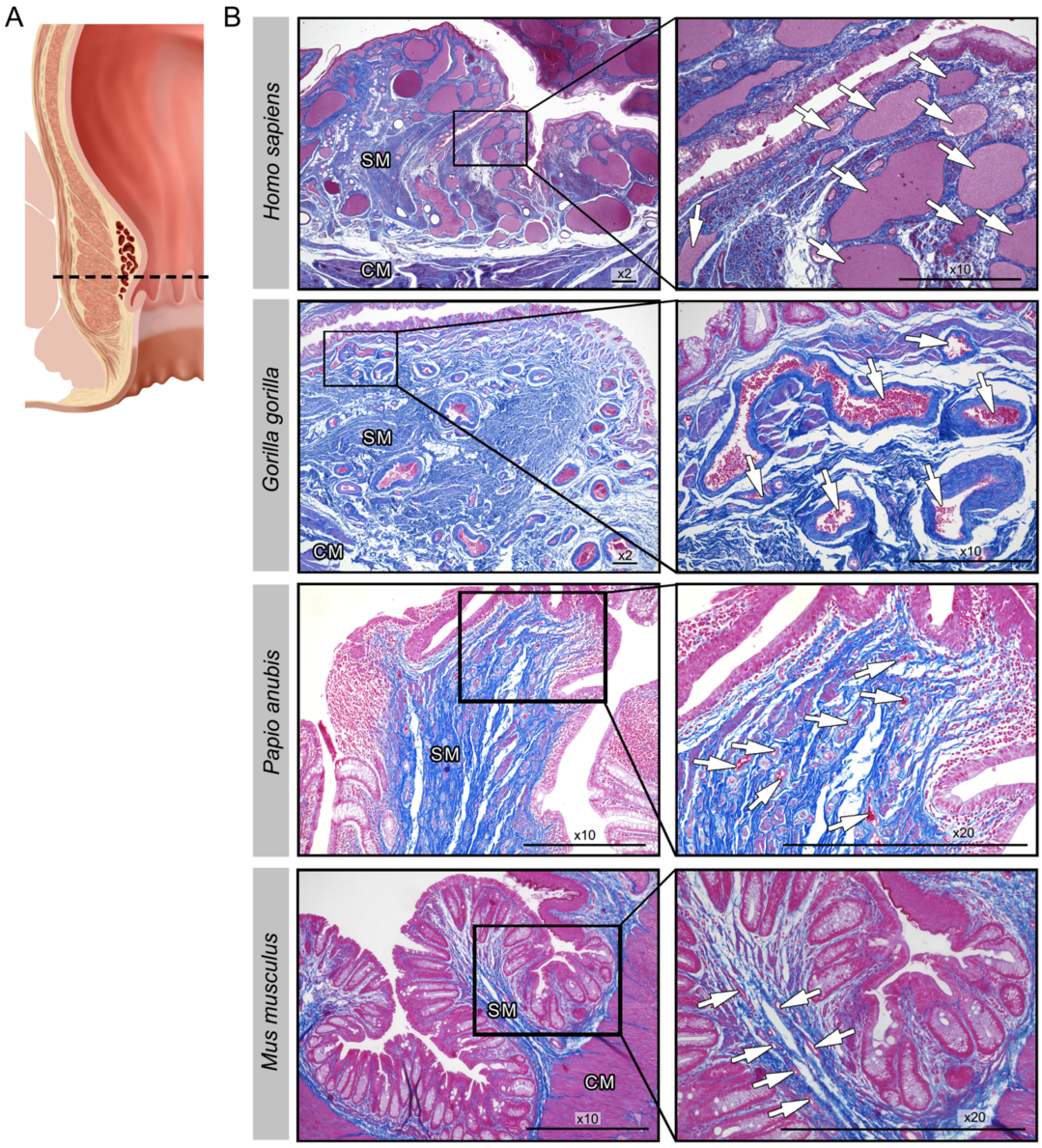
Site-directed ANO1 mutagenesis and whole-cell electrophysiology

F608S (F671S in NM_018043) variant was introduced into the full-length human ANO1 gene with exon 0 (123 bp[88, 89]) and exon b (66 bp[90]) by a single nucleotide change (c.2012 T→C), using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies, Santa Clara, CA, USA) according to the manufacturer's instructions. The integrity of the construct and presence of the F608S mutation were verified by DNA sequencing. The primer sequences were: (forward) 5'-cttccgcagggaggagta-3' and (reverse) 5'-cagcaggaaagccttggagatcagcctctcctc-3'.

HEK293 cells were co-transfected with pEGFP-C1 plus either wild-type ANO1 or F608S-ANO1 by Lipofectamine 3000 (Thermo Fisher Scientific, MA). Ca^{2+}-activated Cl^- currents were recorded by whole-cell electrophysiology as previously described by Strege et al.[91]

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

SUPPLEMENTARY FIGURES

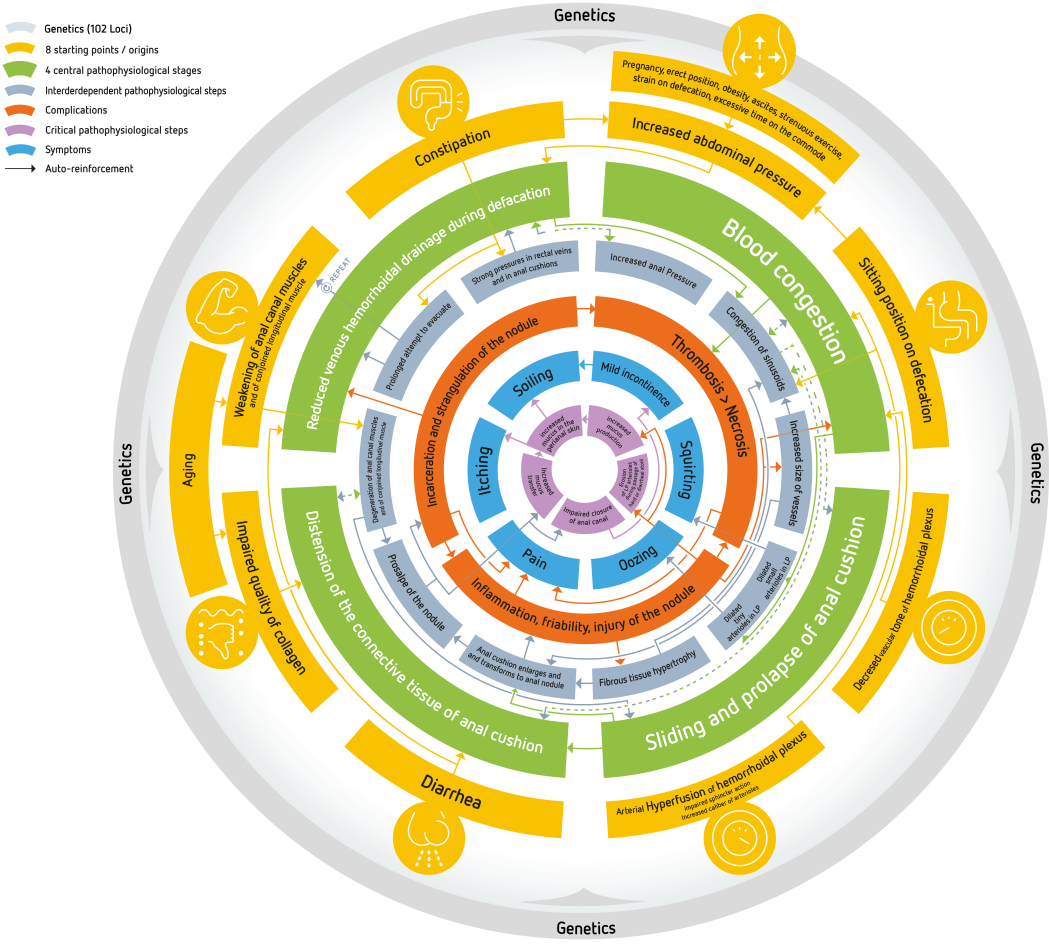


Online Supplementary Figure S1. ***Histological analysis of the anorectum in four different species.***

The left panel **(A)** shows the section plane of the anal canal at the level of the hemorrhoidal plexus. Panel **(B)** shows the hemorrhoidal plexus of 4 different species: *Homo sapiens* (top row), *Gorilla gorilla* (second row), baboon (*Papio anubis; third row*), and mouse (10-week-old male C57BL/6JRj mouse; bottom row). While the human anorectum shows a well-developed hemorrhoidal plexus with densely packed blood

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

vessels of large diameters, the gorilla sample displays a rudimentary hemorrhoidal plexus with fewer and smaller blood vessels. Both baboon and mouse samples exhibit only small-sized and scattered blood vessels which resemble normal vascularization patterns of the regular rectal mucosa. Azan staining with visualization of connective tissue (blue) as well as cell nuclei, erythrocytes and smooth muscle (all purple red). Magnifications for human and gorilla (left 2x, right 10x), for baboon and mouse (left 10x, right 20x). Scale bars: 500 μm. White arrows: hemorrhoidal/submucosal blood vessels, SM = submucosa, CM = circular muscle layer/internal anal sphincter.

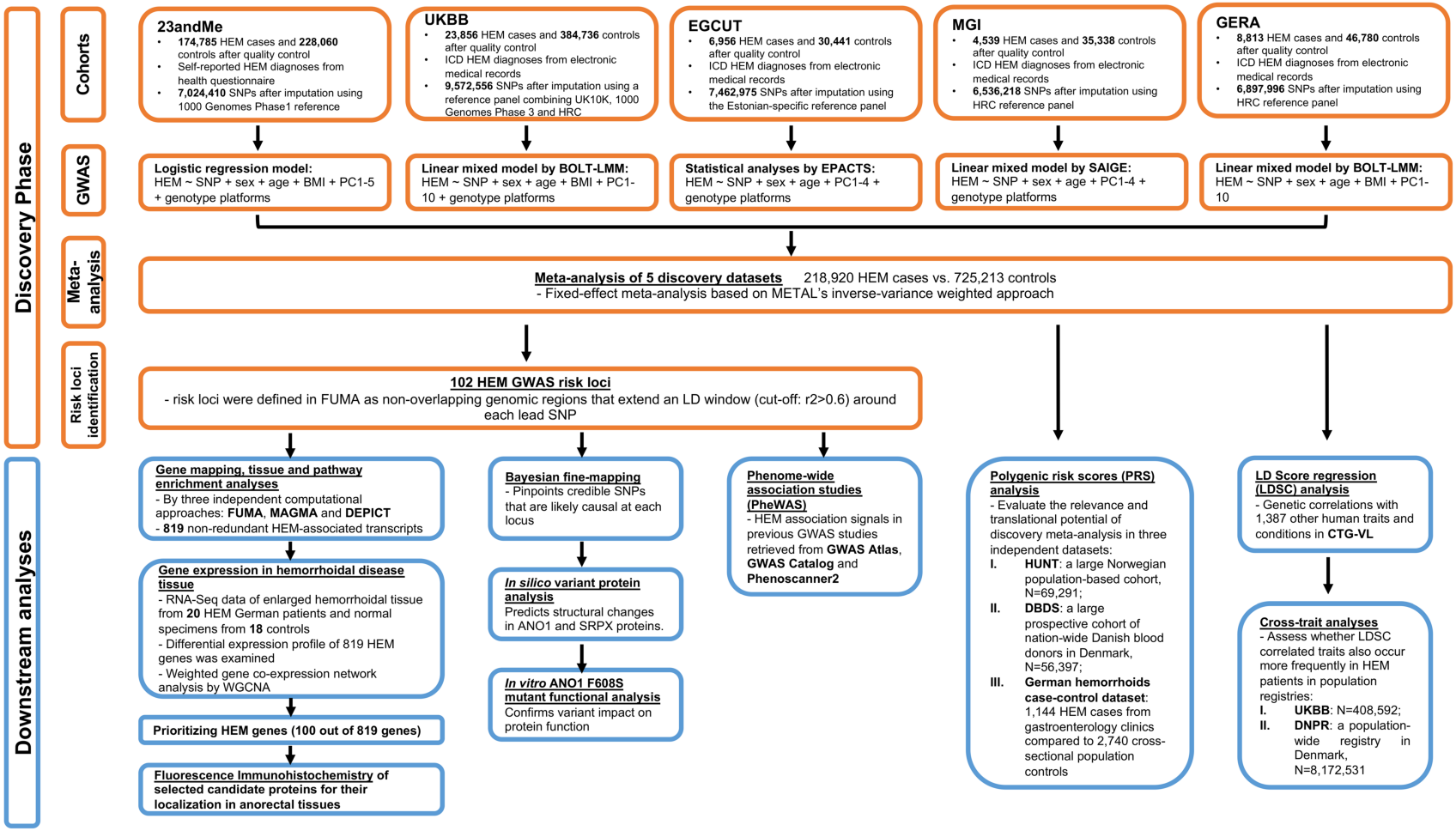


Online Supplementary Figure S2. ***Suggested integrated model that summarizes the contemporary thinking on the pathophysiology of HEM (figure and legend are mainly taken from Figure 2 in Nikolaos Margetis' review[92]).***

HEM is a complex and multifactorial disease, most likely resulting from separate origins and combinations thereof. Different origins and causes have been suggested (orange), which force the hemorrhoidal plexus in different abnormal directions and probably converge in four central pathophysiological events (green). Different consecutive pathophysiological stages (grey) connect the primary causes and the 4 central events. These pathophysiological stages are interconnected, interdependent, and mutually reinforcing, creating a vicious cycle. This multidirectional network is continuously auto-reinforced, as shown by the arrows, and over time provides only

one outcome, with the hemorrhoids deteriorating. Ultimately, symptoms (blue) and complications (red) occur. For further details we refer to Margetis' review[92].

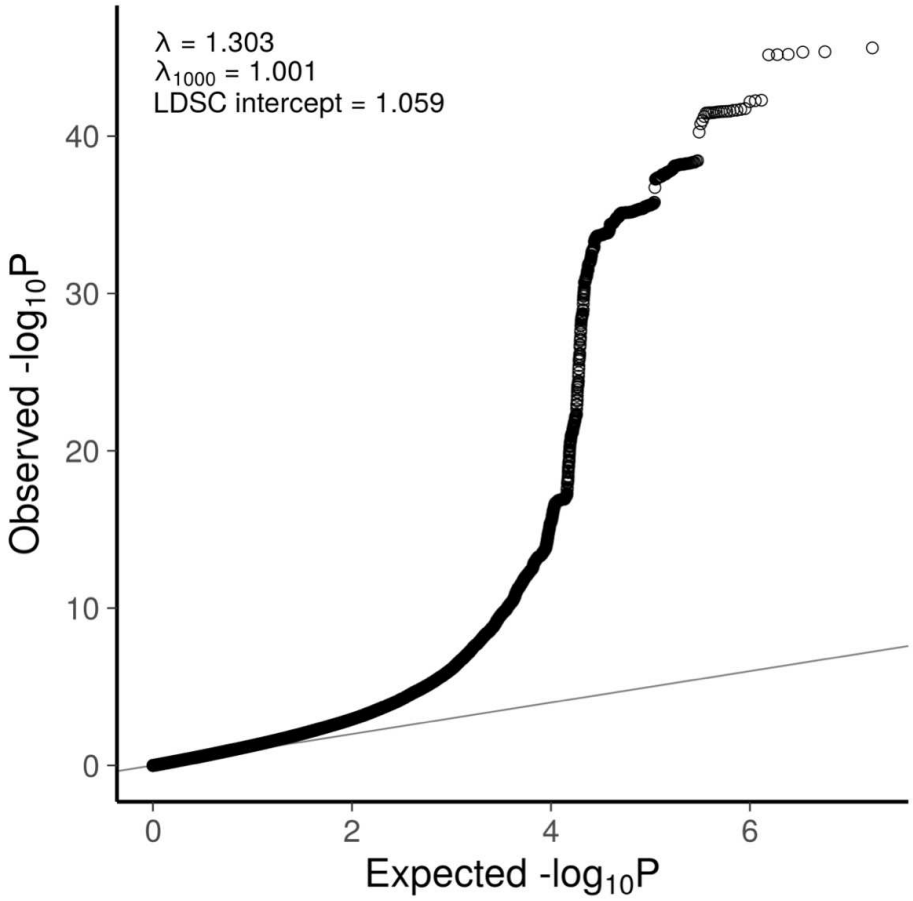
Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868


Discovery Phase
Downstream analyses
Cohorts
GWAS
Meta-analysis
Risk loci identification
23andMe
174,785 HEM cases and 228,060 controls after quality control
Self-reported HEM diagnoses from health questionnaire
7,024,410 SNPs after imputation using 1000 Genomes Phase1 reference
UKBB
23,856 HEM cases and 384,736 controls after quality control
ICD HEM diagnoses from electronic medical records
9,572,556 SNPs after imputation using a reference panel combining UK10K, 1000 Genomes Phase 3 and HRC
EGCUT
6,956 HEM cases and 30,441 controls after quality control
ICD HEM diagnoses from electronic medical records
7,462,975 SNPs after imputation using the Estonian-specific reference panel
MGI
4,539 HEM cases and 35,338 controls after quality control
ICD HEM diagnoses from electronic medical records
6,536,218 SNPs after imputation using HRC reference panel
GERA
8,813 HEM cases and 46,780 controls after quality control
ICD HEM diagnoses from electronic medical records
6,897,996 SNPs after imputation using HRC reference panel
Logistic regression model: HEM ~ SNP + sex + age + BMI + PC1-5 + genotype platforms
Linear mixed model by BOLT-LMM: HEM ~ SNP + sex + age + BMI + PC1-10 + genotype platforms
Statistical analyses by EPACTS: HEM ~ SNP + sex + age + PC1-4 + genotype platforms
Linear mixed model by SAIGE: HEM ~ SNP + sex + age + PC1-4 + genotype platforms
Linear mixed model by BOLT-LMM: HEM ~ SNP + sex + age + BMI + PC1-10
Meta-analysis of 5 discovery datasets 218,920 HEM cases vs. 725,213 controls
- Fixed-effect meta-analysis based on METAL's inverse-variance weighted approach
102 HEM GWAS risk loci
- risk loci were defined in FUMA as non-overlapping genomic regions that extend an LD window (cut-off: r2>0.6) around each lead SNP
Gene mapping, tissue and pathway enrichment analyses
- By three independent computational approaches: FUMA, MAGMA and DEPICT
- 819 non-redundant HEM-associated transcripts
Gene expression in hemorrhoidal disease tissue
- RNA-Seq data of enlarged hemorrhoidal tissue from 20 HEM German patients and normal specimens from 18 controls
- Differential expression profile of 819 HEM genes was examined
- Weighted gene co-expression network analysis by WGCNA
Prioritizing HEM genes (100 out of 819 genes)
Fluorescence Immunohistochemistry of selected candidate proteins for their localization in anorectal tissues
Bayesian fine-mapping
- Pinpoints credible SNPs that are likely causal at each locus
In silico variant protein analysis
Predicts structural changes in ANO1 and SRPX proteins.
In vitro ANO1 F608S mutant functional analysis
Confirms variant impact on protein function
Phenome-wide association studies (PheWAS)
- HEM association signals in previous GWAS studies retrieved from GWAS Atlas, GWAS Catalog and Phenoscanner2
Polygenic risk scores (PRS) analysis
- Evaluate the relevance and translational potential of discovery meta-analysis in three independent datasets:
I. HUNT: a large Norwegian population-based cohort, N=69,291;
II. DBDS: a large prospective cohort of nation-wide Danish blood donors in Denmark, N=56,397;
III. German hemorrhoids case-control dataset: 1,144 HEM cases from gastroenterology clinics compared to 2,740 cross-sectional population controls
LD Score regression (LDSC) analysis
- Genetic correlations with 1,387 other human traits and conditions in CTG-VL
Cross-trait analyses
- Assess whether LDSC correlated traits also occur more frequently in HEM patients in population registries:
I. UKBB: N=408,592;
II. DNPR: a population-wide registry in Denmark, N=8,172,531

Online Supplementary Figure S3. Schematic overview of the study workflow.

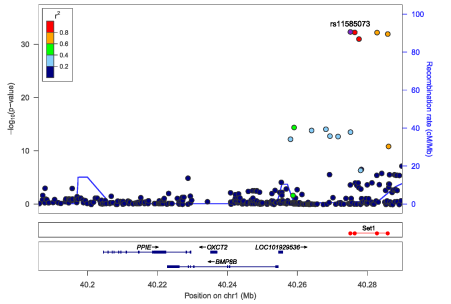
The flowchart shows the study design and analytic strategy of both the discovery phase and the downstream analyses, which includes the study aims, cohorts and numbers of samples of each analytic stage. HEM: hemorrhoids disease. BMI: body mass index, UKBB:UK Biobank, EGCUT: Estonian Genome Center at the University of Tartu, MGI: Michigan Genomics Initiative, GERA: Genetic Epidemiology Research on Aging, HUNT: The Trøndelag Health Study, DBDS: Danish Blood Donor Study, DNPR: Danish National Patient Registry. QC: quality control. IBD: identity by descent. ICD: International Classification of Diseases. Rsq: R square. MAF: minor allele frequency. SNP: Single nucleotide polymorphisms. PC: principal component. LD: linkage disequilibrium. HRC: haplotype reference consortium.

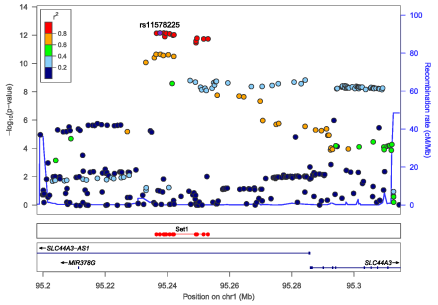
Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868

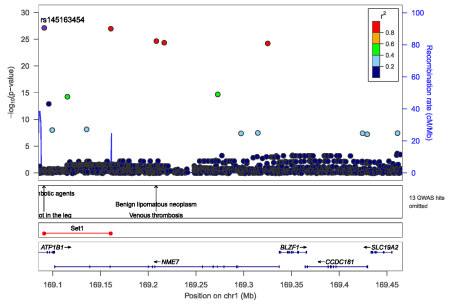


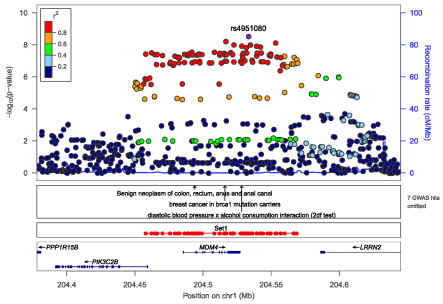
Online Supplementary Figure S4. ***Quantile-quantile (QQ) plot of GWAS meta-analysis results.***

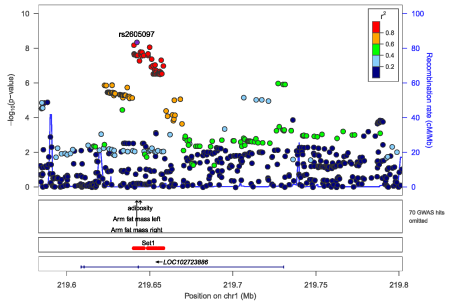
Only markers that passed the imputation quality score R^2>0.8 and MAF>1% were used for the plot. The genomic inflation factor lambda (λ**)** is defined as the ratio of the medians of the sample χ2 test statistics and the 1-d.f. χ2 distribution (0.455)[93]. Lambda inflation statistics are influenced by the sample size. To facilitate comparison with other studies, λ_{1000} converts a given lambda from n cases and m controls so that the value corresponds to an analysis with 1000 cases and 1000 controls. Although genomic inflation was observed (λ=1.303) this was probably due to polygenicity rather than population stratification as determined by linkage disequilibrium score regression analysis (LDSC, intercept=1.059)[94].

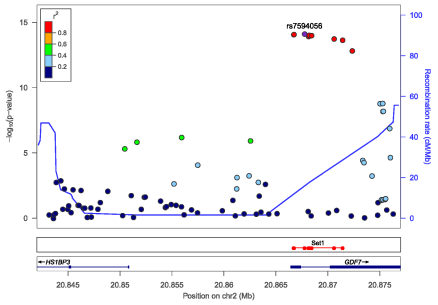

rs11585073
PPIE
CXCT2
LOC101929536
BMP8B
Set1
Position on chr1 (Mb)
Recombination rate (cM/Mb)

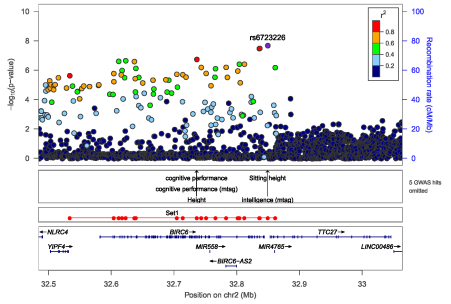

rs11578225
SLC44A3-AS1
MIR378G
SLC44A3
Set1
Position on chr1 (Mb)
Recombination rate (cM/Mb)

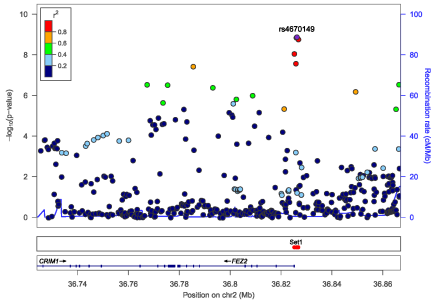

rs145163454
Benign lipomatous neoplasm
Venous thrombosis
13 GWAS hits omitted
ATP1B1
NME7
BLZF1
CCDC181
SLC19A2
Set1
Position on chr1 (Mb)
Recombination rate (cM/Mb)

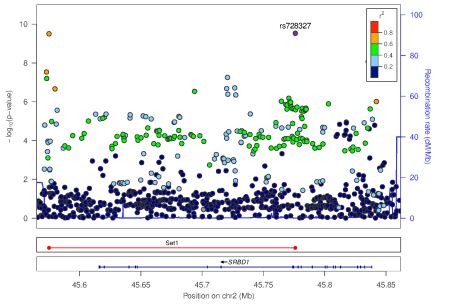

rs4951080
Benign neoplasm of colon, rectum, anus and anal canal
breast cancer in brca1 mutation carriers
diastolic blood pressure x alcohol consumption interaction (2df test)
7 GWAS hits omitted
PPP1R15B
PIK3C2B
MDM4
LRRN2
Set1
Position on chr1 (Mb)
Recombination rate (cM/Mb)

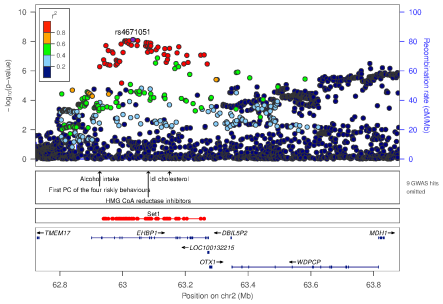

rs2605097
adiposity
Arm fat mass left
Arm fat mass right
70 GWAS hits omitted
LOC102723886
Set1
Position on chr1 (Mb)
Recombination rate (cM/Mb)

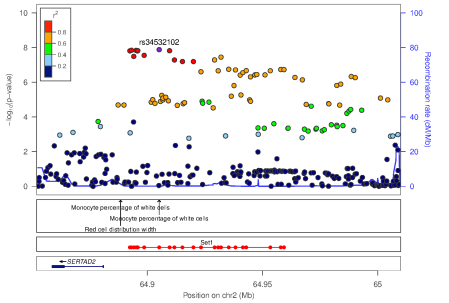

rs75940056
HS1BP3
GDF7
Set1
Position on chr2 (Mb)
Recombination rate (cM/Mb)

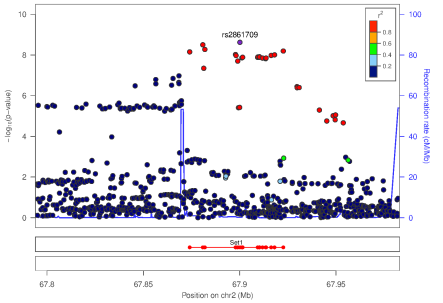

rs6723226
cognitive performance
cognitive performance (mtag)
Sitting height
Height
Intelligence (mtag)
6 GWAS hits omitted
NLRC4
BIRC6
TTC27
YIPF4
MIR558
MIR4765
LINC00486
BIRC6-AS2
Set1
Position on chr2 (Mb)
Recombination rate (cM/Mb)

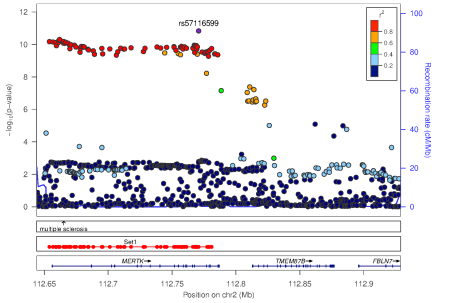

rs4670149
CRIM1
FEZ2
Set1
Position on chr2 (Mb)
Recombination rate (cM/Mb)

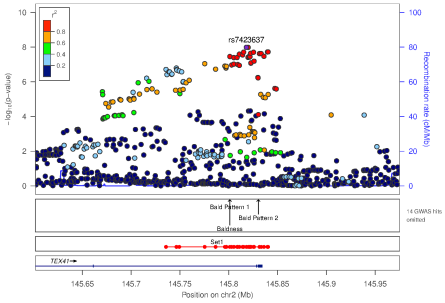

rs728327
SRBD1
Position on chr2 (Mb)

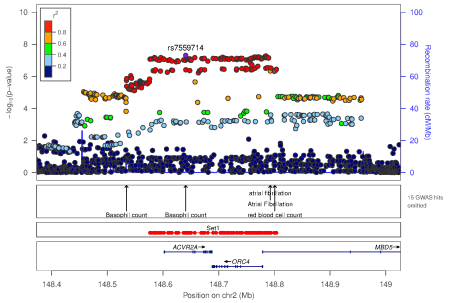

rs4671051
Alcohol intake
First PC of the four risky behaviours
ldl cholesterol
HMG CoA reductase inhibitors
TMEM17
EIF6P1
DBILSP2
LOC100132215
OTX1
WDPCP
MOH1
Position on chr2 (Mb)

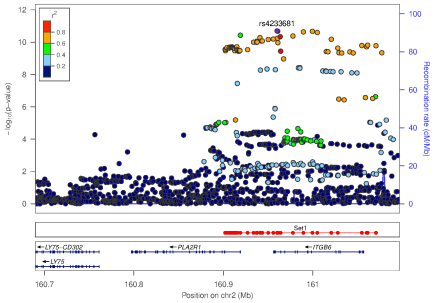

rs34532102
Monocyte percentage of white cells
Red cell distribution width
SERTAD2
Position on chr2 (Mb)

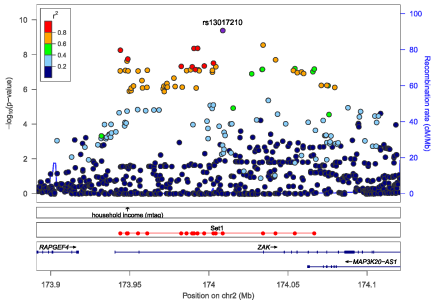

rs2861709
Position on chr2 (Mb)

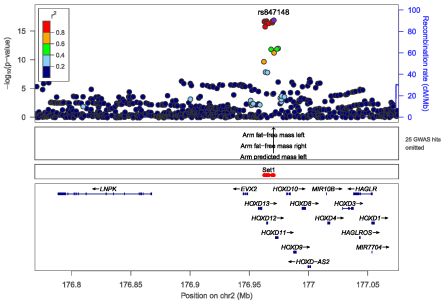

rs57116599
multiple sclerosis
MERTK
TMEM87B
FBLN7
Position on chr2 (Mb)

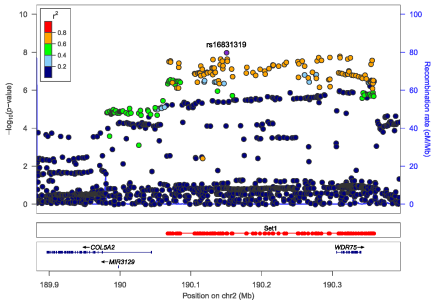

rs7423637
Bald Pattern 1
Bald Pattern 2
Baldness
TEX41
Position on chr2 (Mb)

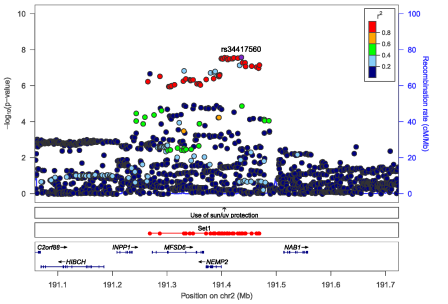

rs7559714
Basophil count
Basophil count
atrial fibrillation
Atrial Fibrillation
red blood cell count
ACVR2A
ORC4
MBD5
Position on chr2 (Mb)

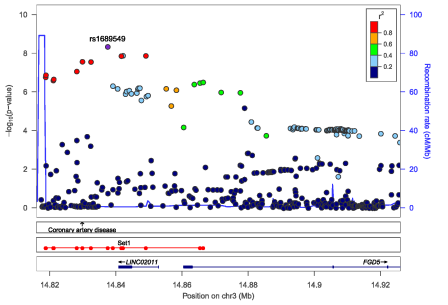

rs4233681
LY75-CD302
LY75
PLA2R1
ITGB6
Position on chr2 (Mb)

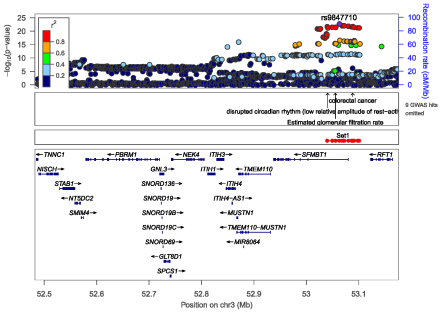

rs13017210
household income (misc)
Set1
RAPGEF4
ZAK
MAP3K20-AS1
Position on chr2 (Mb)
Recombination rate (cM/Mb)

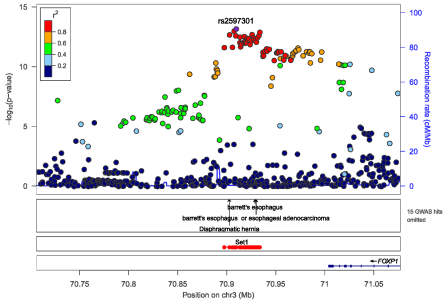

rs847148
Arm fat-free mass left
Arm fat-free mass right
Arm predicted mass left
26 GWAS hits omitted
Set1
LNPK
EVX2
HOXD10
MIR196
HAGLR
HOXD13
HOXD9
HOXD3
HOXD12
HOXD4
HOXD1
HOXD11
HAGLROS
HOXD8
MIR7704
HOXD-AS2
Position on chr2 (Mb)
Recombination rate (cM/Mb)

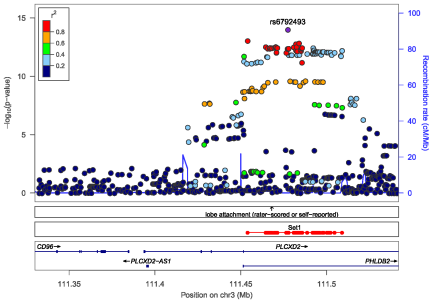

rs16831319
Set1
COL5A2
MIR3129
WDR75
Position on chr2 (Mb)
Recombination rate (cM/Mb)

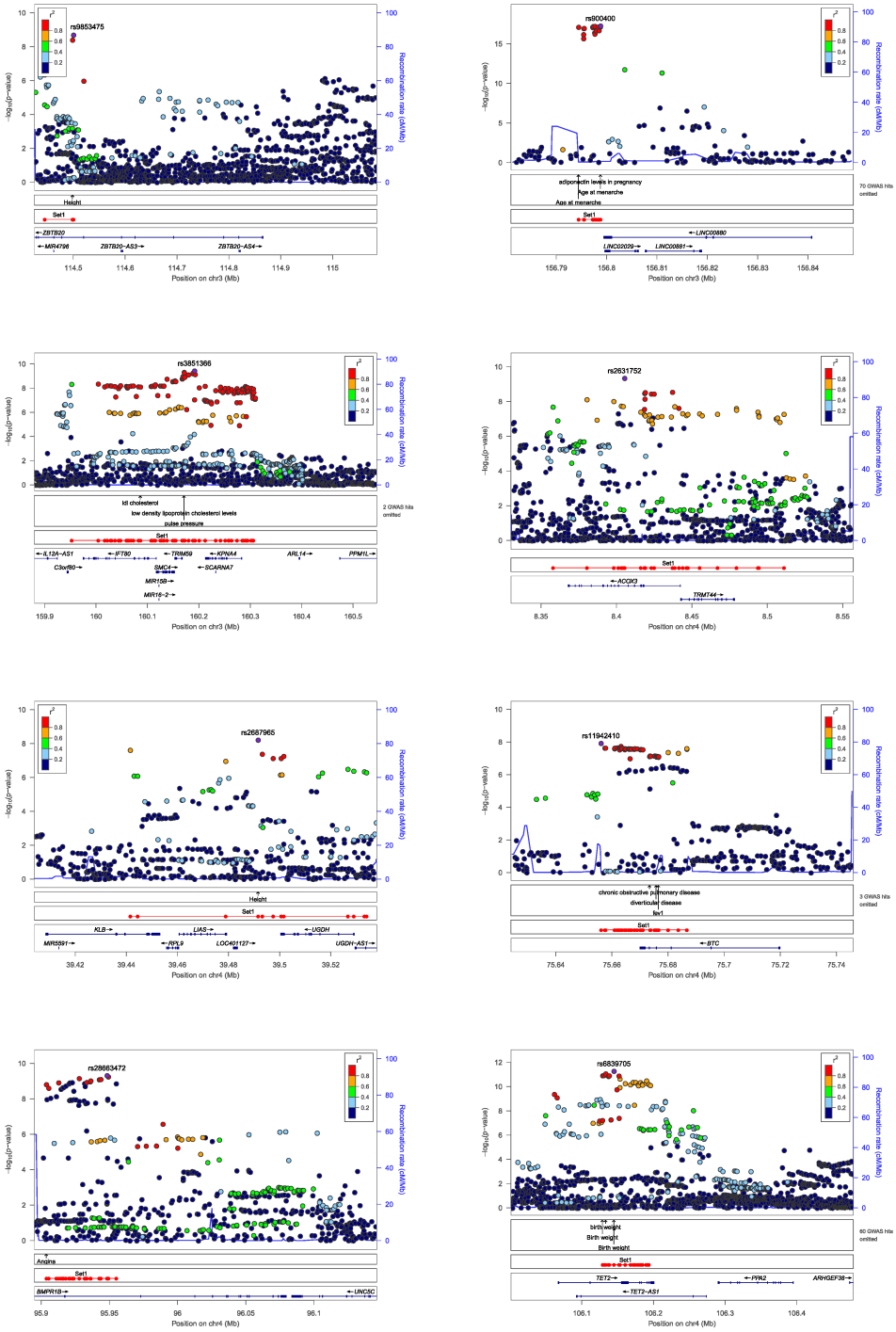

rs34417560
Use of sun/uv protection
Set1
C2orf69
HIBCH
INPP1
MFSD6
NEMP2
NAB1
Position on chr2 (Mb)
Recombination rate (cM/Mb)

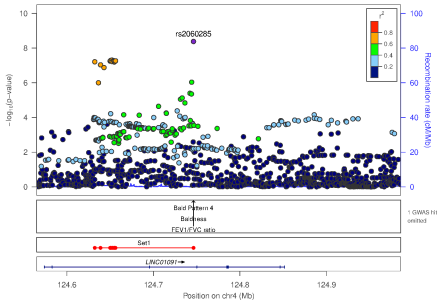

rs1689549
Coronary artery disease
Set1
LINC02011
FGD5
Position on chr3 (Mb)
Recombination rate (cM/Mb)

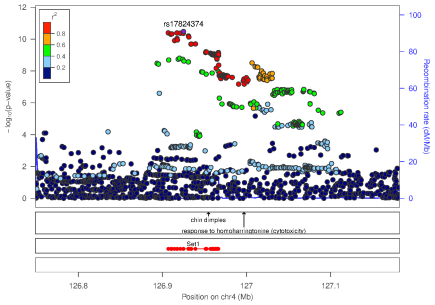

rs9847710
colorectal cancer
disrupted circadian rhythm (low relative amplitude of rest–acti
Estimated glomerular filtration rate
9 GWAS hits omitted
Set1
TNNC1
PBRM1
NEK4
ITK3
SFMBT1
RFT1
NISCH
GNL3
ITIH1
TMEM110
STAB1
ITIH4
NT5DC2
SNORD136
SNORD19
ITIH4-AS1
SMIM4
SNORD19B
MUSTN1
SNORD19C
TMEM110-MUSTN1
SNORD69
MIR8064
GLTSD1
SPCS1
Position on chr3 (Mb)
Recombination rate (cM/Mb)

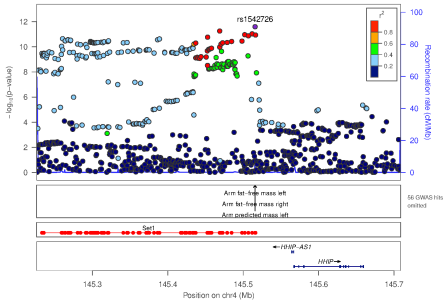

rs2597301
barrett's esophagus
barrett's esophagus or esophageal adenocarcinoma
Diaphragmatic hernia
16 GWAS hits omitted
Set1
FGKP1
Position on chr3 (Mb)
Recombination rate (cM/Mb)

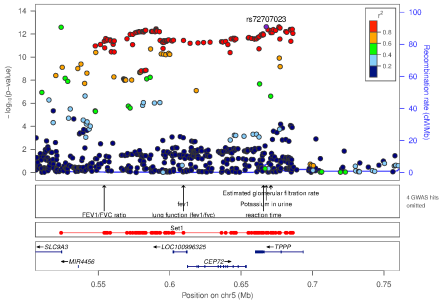

rs6792493
lobe attachment (rater−scored or self−reported)
Set1
CD96
PLCXD2
PLCXD2-AS1
PHLDB2
Position on chr3 (Mb)
Recombination rate (cM/Mb)

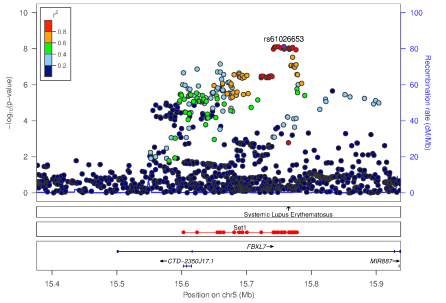

rs9853475
rs900400
rs3851366
rs2631752
rs2687965
rs11942410
rs28863472
rs6839705
Position on chr3 (Mb)
Position on chr4 (Mb)
Recombination rate (cM/Mb)

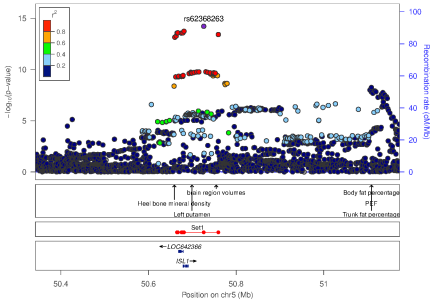

rs2060285
Bad Pattern 4
Badness
FEV1/FVC ratio
Set1
LINC01091
Position on chr4 (Mb)
Recombination rate (cM/Mb)

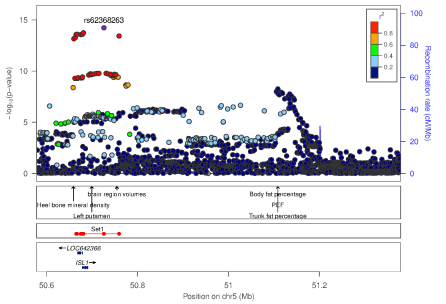

rs17824374
chin dimples
response to formothermotolerone (cytotoxicity)
Set1
Position on chr4 (Mb)
Recombination rate (cM/Mb)

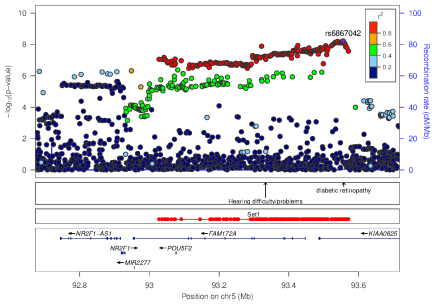

rs1542726
Arm fat-free mass left
Arm fat-free mass right
Arm predicted mass left
Set1
HNRP-AS1
HNRP
Position on chr4 (Mb)
Recombination rate (cM/Mb)

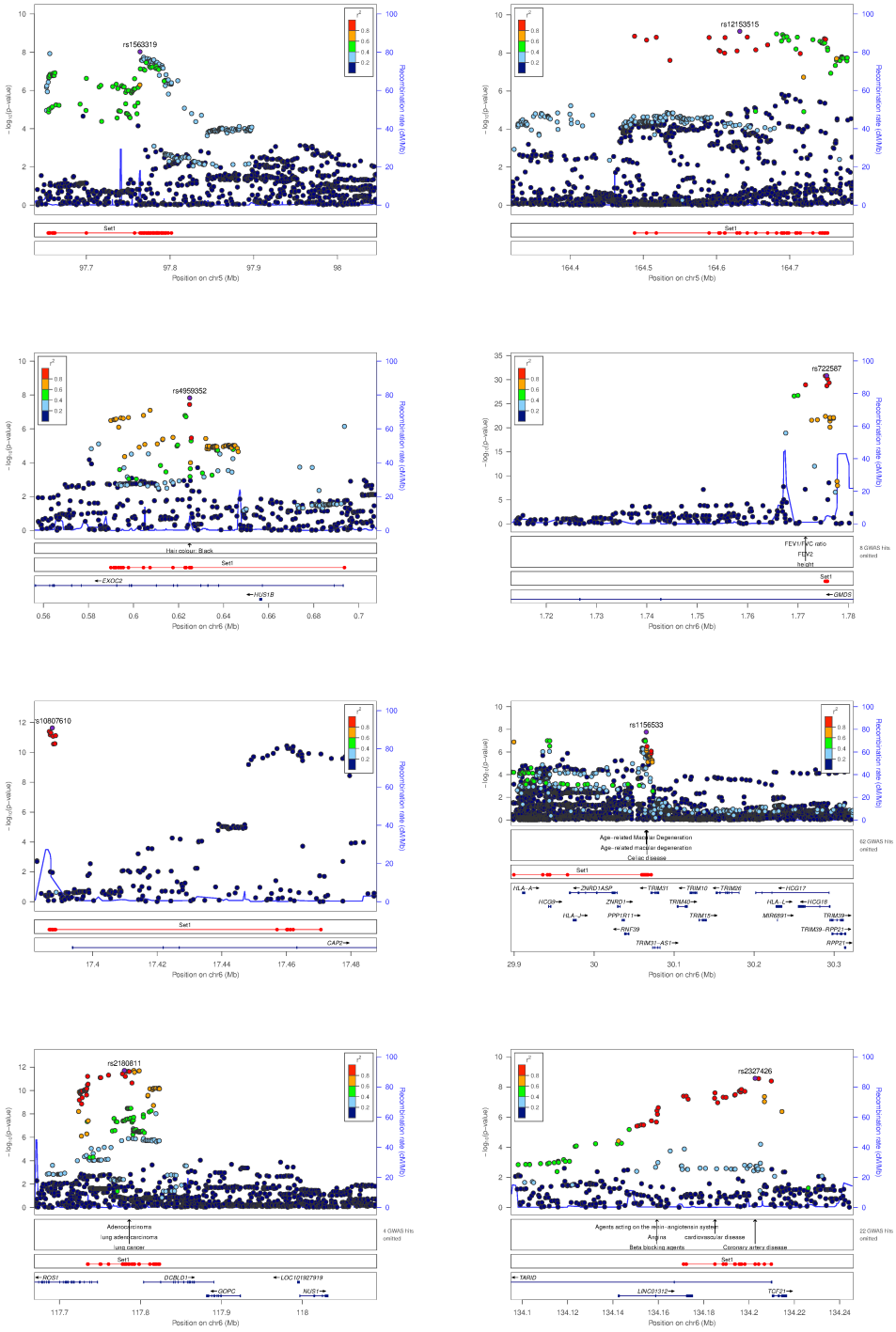

rs72707023
FEV1/FVC ratio
Set1
SLC9A3
LOC100996325
TPPP
MIR4456
CEP72
Position on chr5 (Mb)
Recombination rate (cM/Mb)

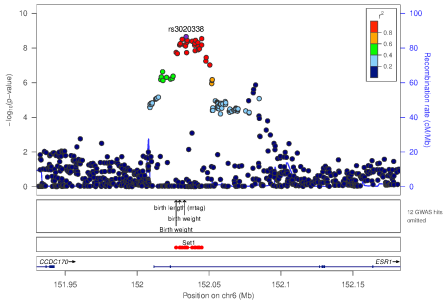

rs61026653
Systemic Lupus Erythematosus
Set1
FBXL7
CTD-2350J17.1
MIR597
Position on chr5 (Mb)
Recombination rate (cM/Mb)

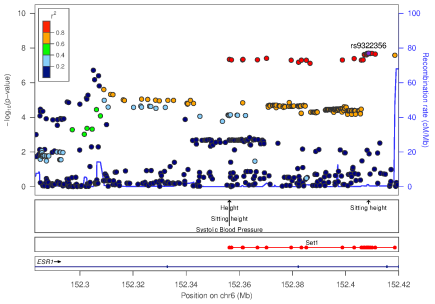

rs62368263
brain region volumes
Heel bone mineral density
Left putamen
Body fat percentage
PEF
Trunk fat percentage
Set1
LOC642366
ISL1
Position on chr5 (Mb)
Recombination rate (cM/Mb)

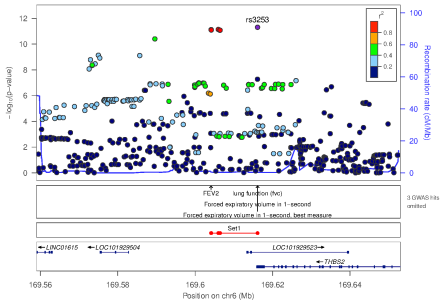

rs62368263
brain region volumes
Heel bone mineral density
Left putamen
Body fat percentage
PEF
Trunk fat percentage
Set1
LOC642366
ISL1
Position on chr5 (Mb)
Recombination rate (cM/Mb)

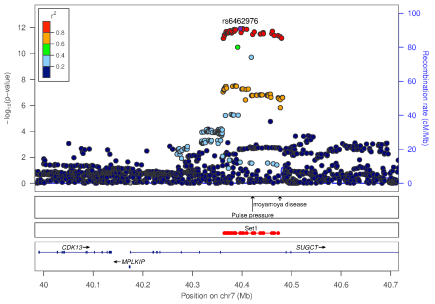

rs6867042
diabetic retinopathy
Hearing difficulty/problems
Set1
NR2F1-AS1
FAM172A
KIAA0825
NR2F1
POU5F2
MIR2277
Position on chr5 (Mb)
Recombination rate (cM/Mb)

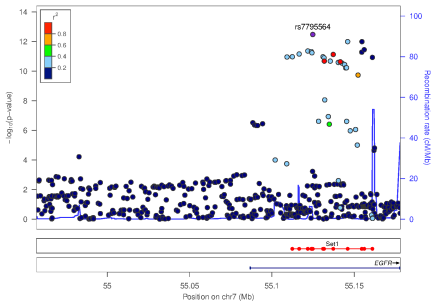

rs1563319
rs12153515
rs4959352
rs722587
s10807610
rs1156533
rs2180811
rs2327426
Position on chr5 (Mb)
Position on chr6 (Mb)
Recombination rate (cM/Mb)

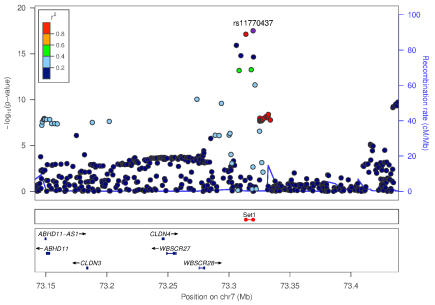

rs3020338
birth length (mtag)
birth weight
Birth weight
CCDC170
ESR1
Position on chr6 (Mb)

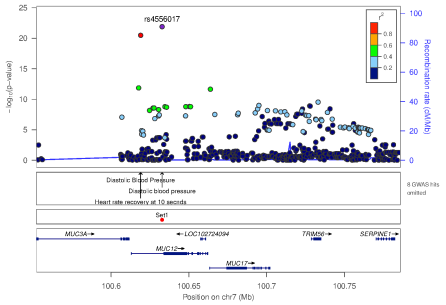

rs9322356
Height
Sitting height
Systolic Blood Pressure
ESR1
Position on chr6 (Mb)

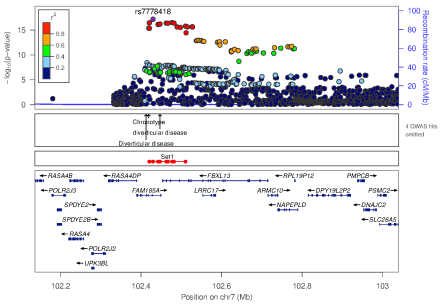

rs3253
FEV2 lung function (fvc)
Forced expiratory volume in 1-second
Forced expiratory volume in 1-second, best measure
LINC01615
LOC101929504
LOC101929523
TYRO3
Position on chr6 (Mb)

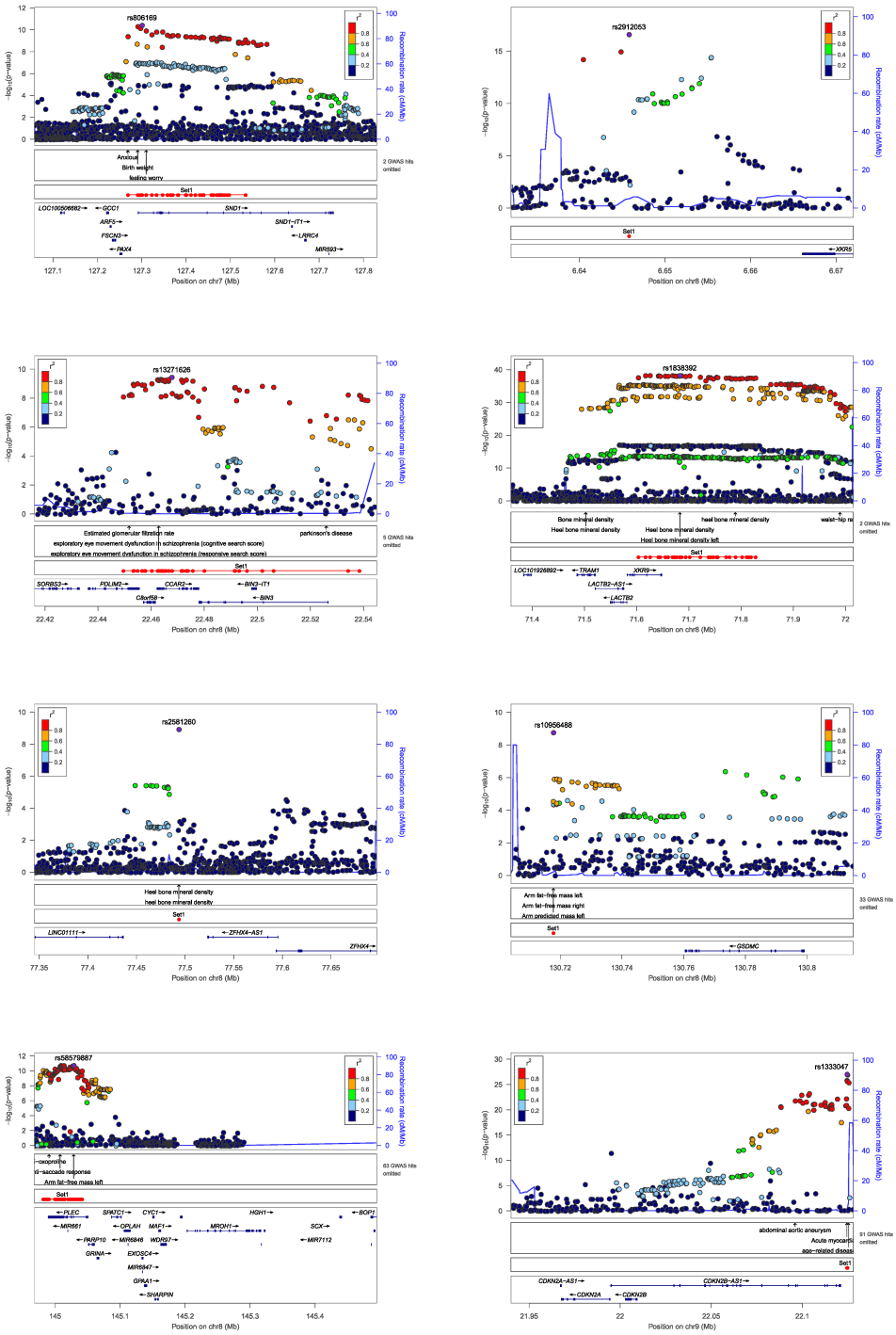

rs6462976
moyamoya disease
Pulse pressure
CDK13
MPLKIP
SUGCT
Position on chr7 (Mb)

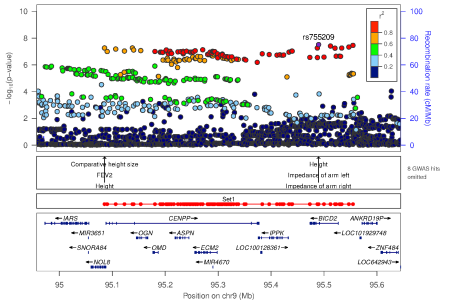

rs7795564
EGFR
Position on chr7 (Mb)

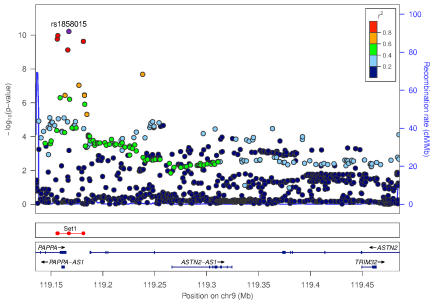

rs11770437
ABHD11-AS1
ABHD11
CLDN4
WBSCR27
CLDN3
WBSCR28
Position on chr7 (Mb)

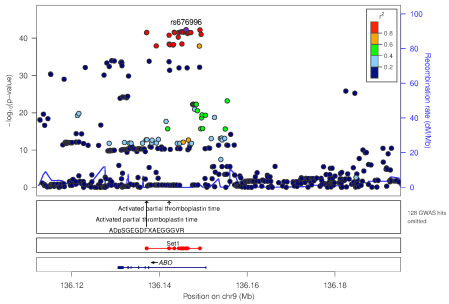

rs4556617
Diastolic Blood Pressure
Diastolic blood pressure
Heart rate recovery at 10 secnds
MUC3A
LOC102724094
MUC12
TRIM50
SERPINE1
MUC17
Position on chr7 (Mb)

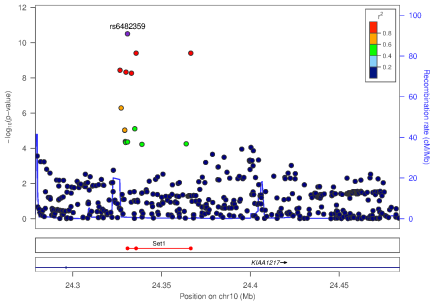

rs7778418
Chronotype
diverticular disease
Diverticular disease
RASA4B
RASA4DP
TBXL13
RPL19P12
PMPCB
POLR2J3
FAM185A
LRRC17
ARMC10
DPY19L3P2
PSMC2
SPDYE2
NAPEPLD
DNAJC2
SPDYE2B
SLC26A5
RASA4
POLR2J2
UPK3BL
Position on chr7 (Mb)

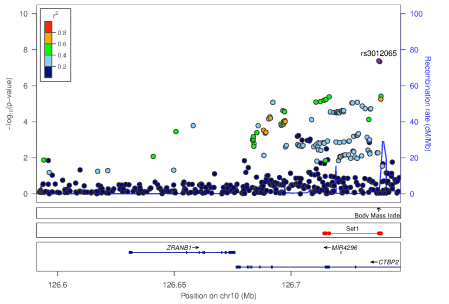

rs806169
rs2912053
rs13271626
rs1838392
rs2581260
rs10956488
rs58579887
rs1333047
Recombination rate (cM/Mb)
Position on chr7 (Mb)
Position on chr8 (Mb)
Position on chr6 (Mb)
Position on chr9 (Mb)

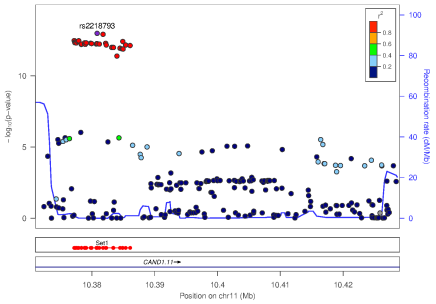

rs755209
Position on chr9 (Mb)

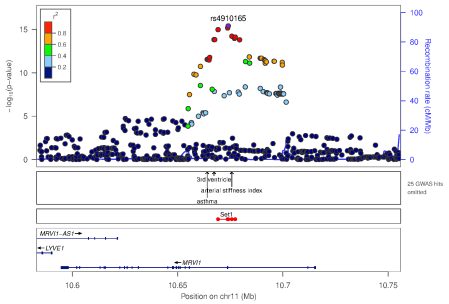

rs1858015
Position on chr9 (Mb)

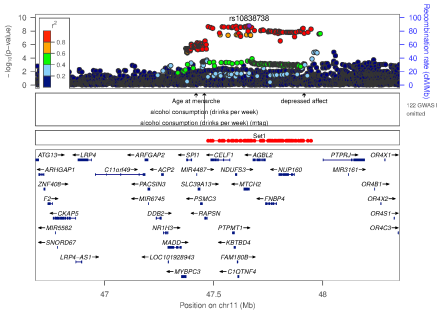

rs676996
Position on chr9 (Mb)

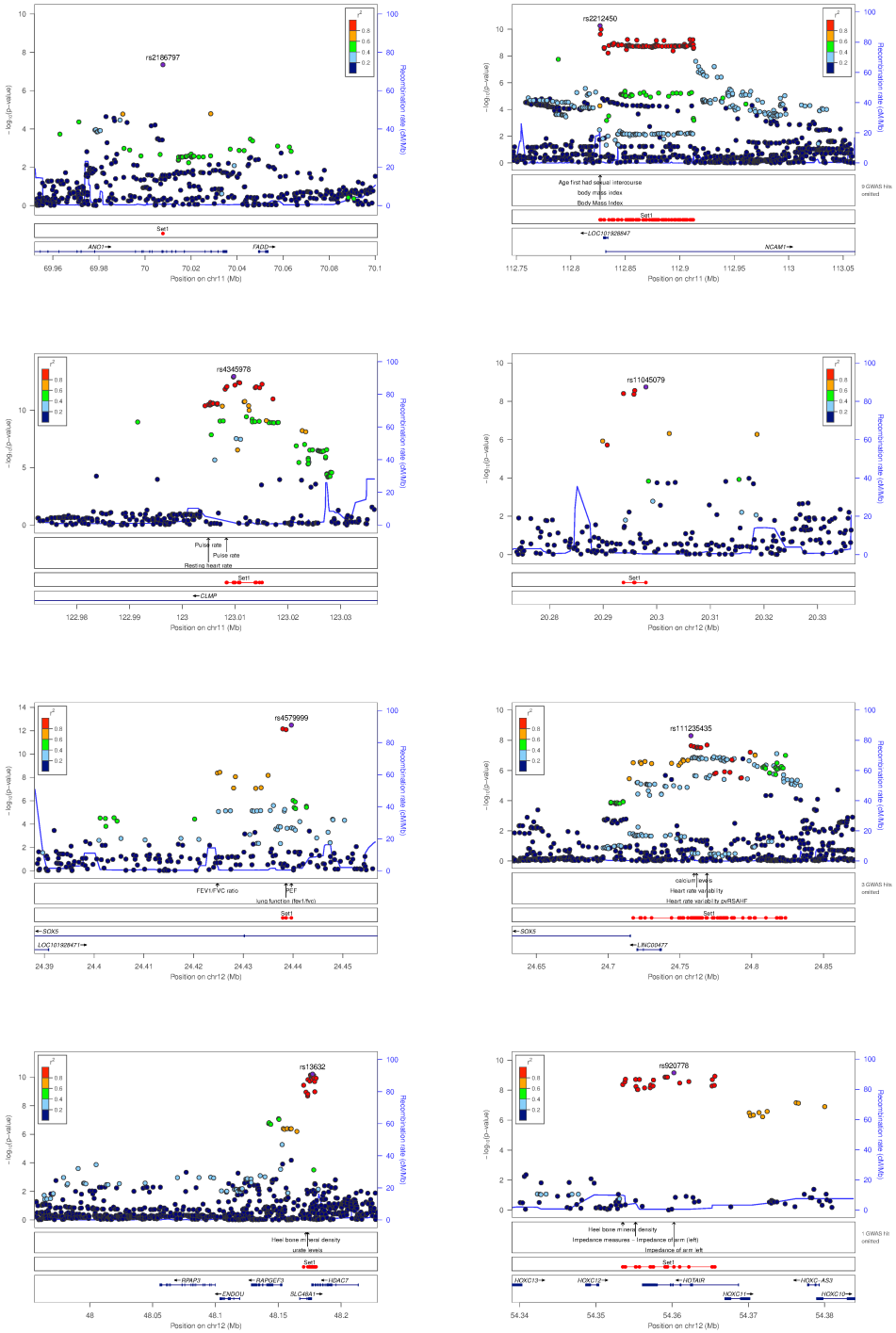

rs6482359
Position on chr10 (Mb)

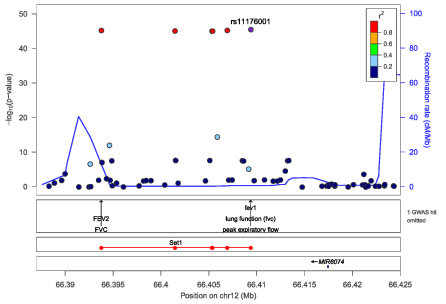

rs3012065
Position on chr10 (Mb)

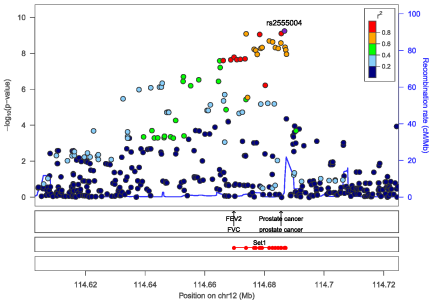

rs2218793
Position on chr11 (Mb)

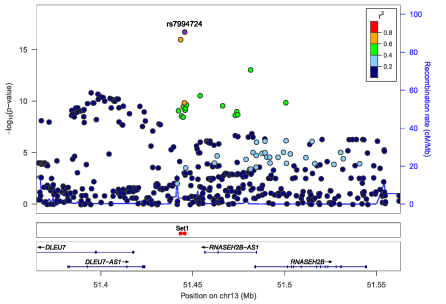

rs4910165
Position on chr11 (Mb)

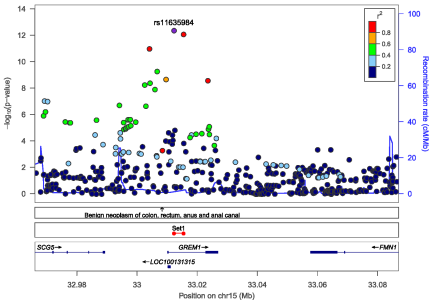

rs10838738
Position on chr11 (Mb)

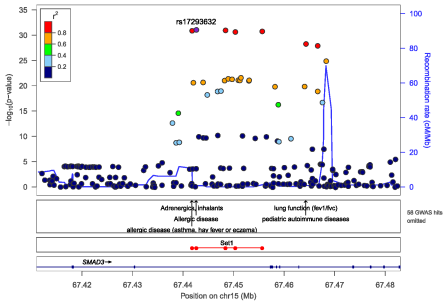

rs2186797
rs2212450
rs4345978
rs11045079
rs4579999
rs111235435
rs13632
rs920778
Position on chr11 (Mb)
Position on chr12 (Mb)
Recombination rate (cM/Mb)

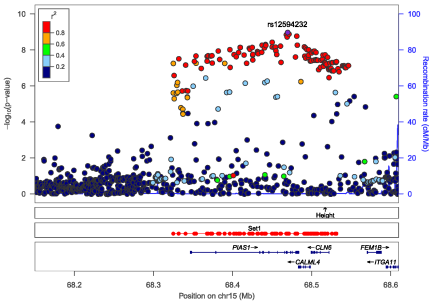

rs11176001
−log10(p-value)
Recombination rate (cM/Mb)
fev1
lung function (fvc)
peak expiratory flow
FEV2
FVC
1 GWAS hit omitted
Set1
MIR8074
Position on chr12 (Mb)

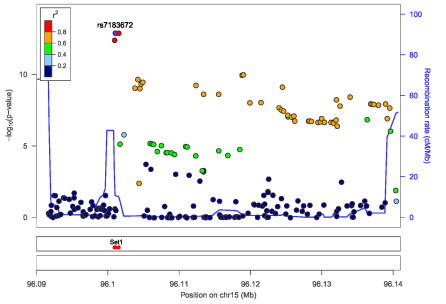

rs2555004
−log10(p-value)
Recombination rate (cM/Mb)
FEV2
FVC
Prostate cancer
prostate cancer
Set1
Position on chr12 (Mb)

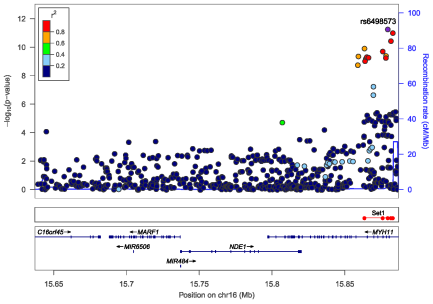

rs7994724
−log10(p-value)
Recombination rate (cM/Mb)
Set1
DLEU7
DLEU7-AS1
RNASEH2B-AS1
RNASEH2B
Position on chr13 (Mb)

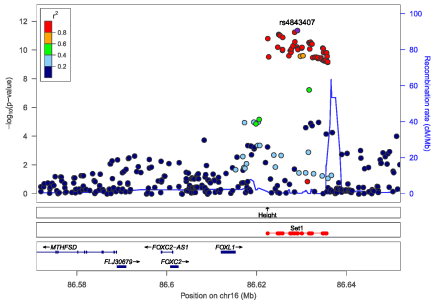

rs11635984
−log10(p-value)
Recombination rate (cM/Mb)
Benign neoplasm of colon, rectum, anus and anal canal
Set1
SCG5
GREM1
FMN1
LOC100131315
Position on chr15 (Mb)

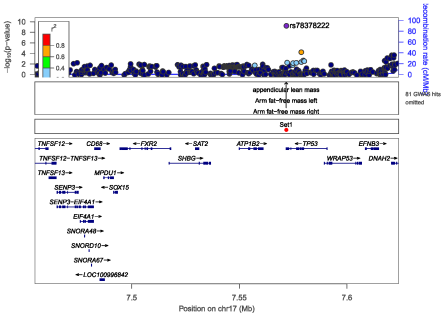

rs17293632
−log10(p-value)
Recombination rate (cM/Mb)
Adrenergics, inhalants
Allergic disease
allergic disease (asthma, hay fever or eczema)
lung function (fev1/fvc)
pediatric autoimmune diseases
66 GWAS hits omitted
Set1
SMAD3
Position on chr15 (Mb)

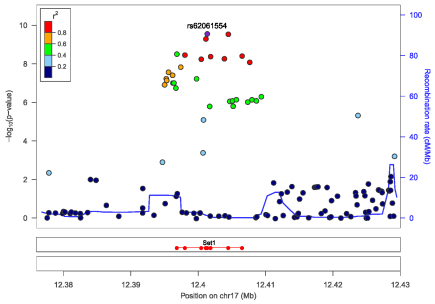

rs12594232
−log10(p-value)
Recombination rate (cM/Mb)
Height
Set1
PIAS1
CLN6
FEM1B
CALML4
ITGA11
Position on chr15 (Mb)

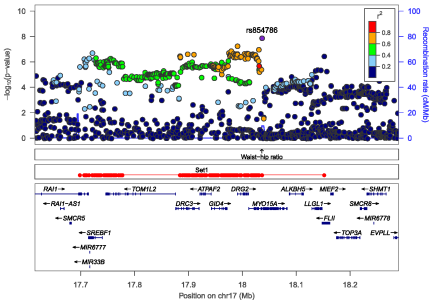

rs7183672
−log10(p-value)
Recombination rate (cM/Mb)
Set1
Position on chr15 (Mb)

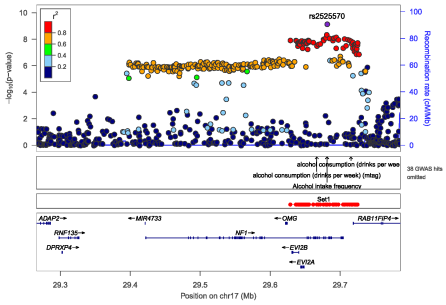

rs6498573
−log10(p-value)
Recombination rate (cM/Mb)
Set1
C16orf45
MARF1
MYH11
MIR8506
NDE1
MIR484
Position on chr16 (Mb)

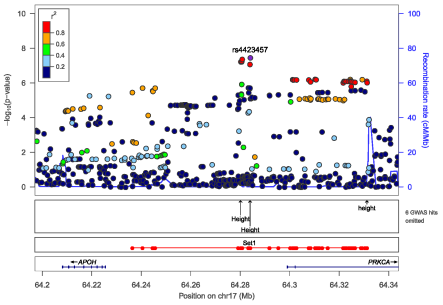

rs4843407
Height
Set1
MTHFSD
FOXC2-AS1
FOXL1
FLJ30679
FOXC2
Position on chr16 (Mb)
Recombination rate (cM/Mb)

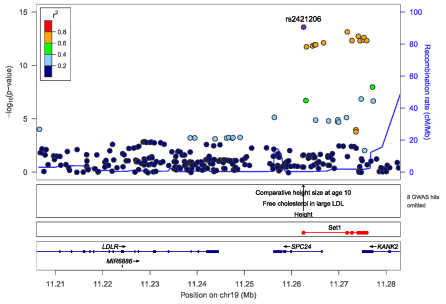

rs78378222
appendicular lean mass
Arm fat-free mass left
Arm fat-free mass right
81 GWAS hits omitted
Set1
TNFSF12
CD58
FXR2
SAT2
ATP1B2
TP53
EFNB3
TNFSF12-TNFSF13
SHBG
WRAP53
DNAH2
TNFSF13
MPDU1
SENP3
SQX15
SENP3-EIF4A1
EIF4A1
SNORA48
SNORD10
SNORA67
LOC100996842
Position on chr17 (Mb)
Recombination rate (cM/Mb)

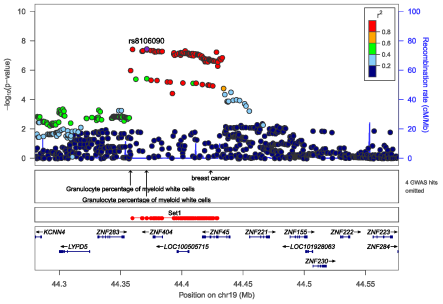

rs62061554
Set1
Position on chr17 (Mb)
Recombination rate (cM/Mb)

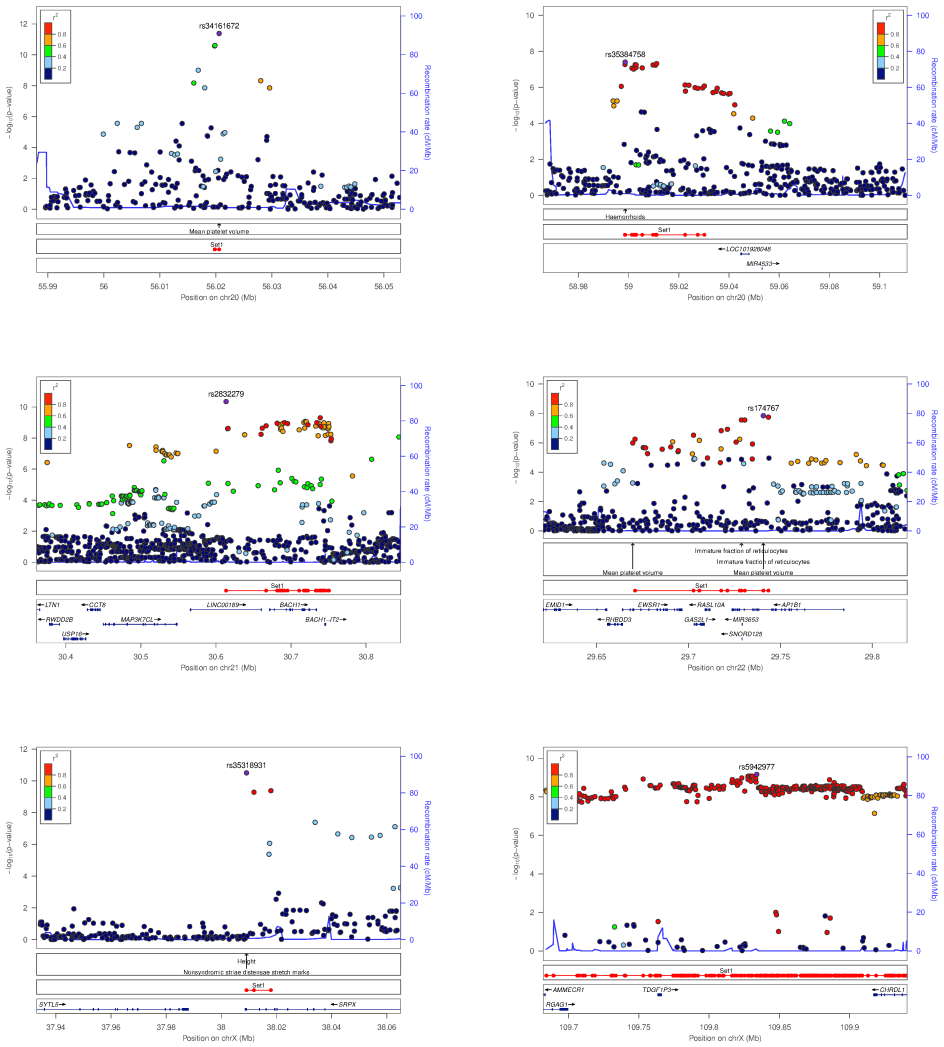

rs854786
Waist-hip ratio
Set1
RAI1
TOM1L2
ATPAF2
DRG2
ALKBH5
MIEF2
SHMT1
RAI1-AS1
DRC3
GID4
MYO15A
LLGL1
SMCR8
SMCR5
FLII
MIR6778
SREBF1
TOP3A
EVPLL
MIR6777
MIR33B
Position on chr17 (Mb)
Recombination rate (cM/Mb)


rs2525570
alcohol consumption (drinks per week)
alcohol consumption (drinks per week) (mtag)
Alcohol intake frequency
35 GWAS hits omitted
Set1
ADAP2
MIR4733
QMG
RAB11FIP4
RNF135
NF1
DPRXP4
EVI2B
EVI2A
Position on chr17 (Mb)
Recombination rate (cM/Mb)


rs4423457
Height
height
6 GWAS hits omitted
Set1
APOH
PRKCA
Position on chr17 (Mb)
Recombination rate (cM/Mb)

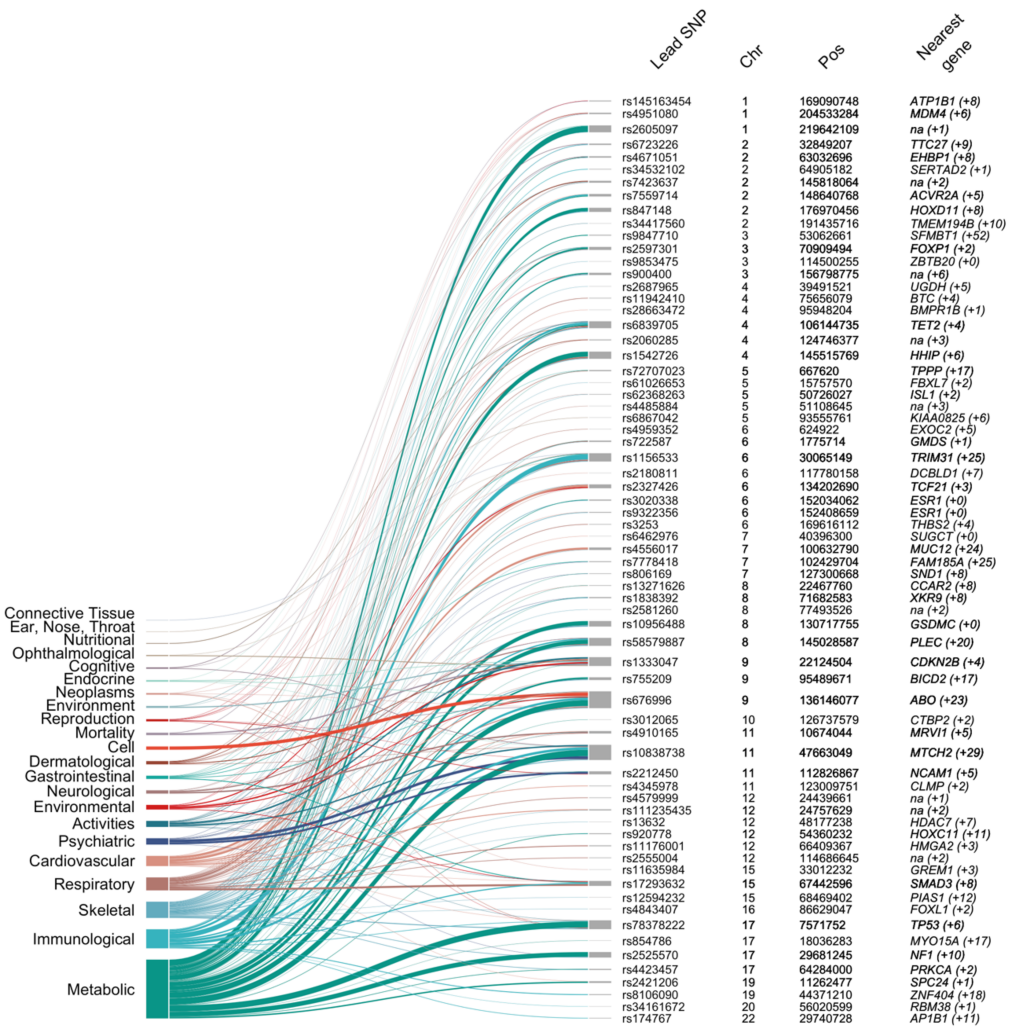

rs2421206
Comparative height size at age 10
Free cholesterol in large LDL
Height
8 GWAS hits omitted
Set1
LDLR
SPC24
KANK2
MIR5886
Position on chr19 (Mb)
Recombination rate (cM/Mb)


rs8106090
breast cancer
Granulocyte percentage of myeloid white cells
Granulocyte percentage of myeloid white cells
4 GWAS hits omitted
Set1
KCNN4
ZNF283
ZNF404
ZNF45
ZNF221
ZNF155
ZNF222
ZNF223
LYPD5
LOC100505715
LOC101928063
ZNF284
ZNF230
Position on chr19 (Mb)
Recombination rate (cM/Mb)



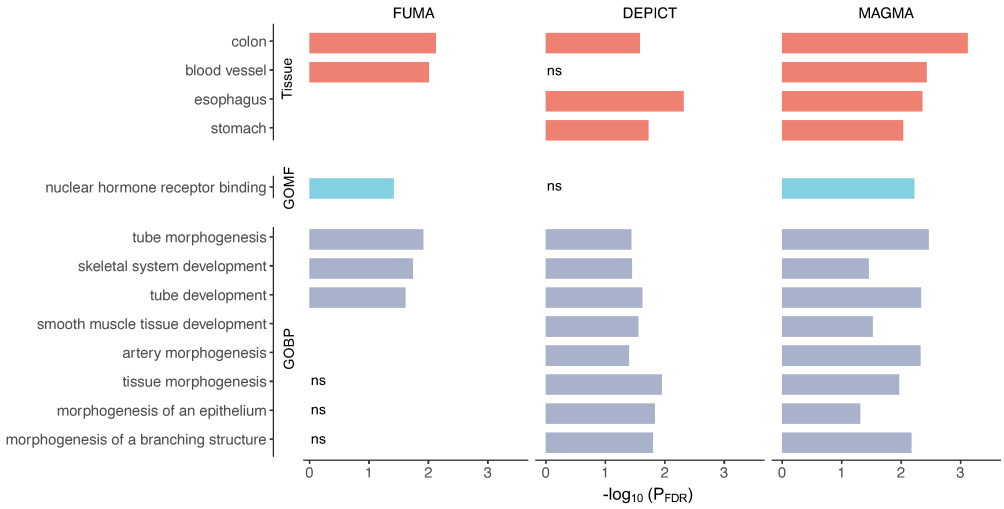
Online Supplementary Figure S5. ***Regional association plots of HEM GWAS risk loci.***

Shown are the -log$_{10}$ *P*-values from meta-analysis with regard to the physical location of markers and the degree of linkage disequilibrium (r^2). Purple circle: lead SNP; line: recombination intensity (cM/Mb). Positions and gene annotations are according to NCBI's build 37 (hg19). Plots were generate using LocusZoom[42], also reporting the 95%-fine mapped credible sets at each locus.

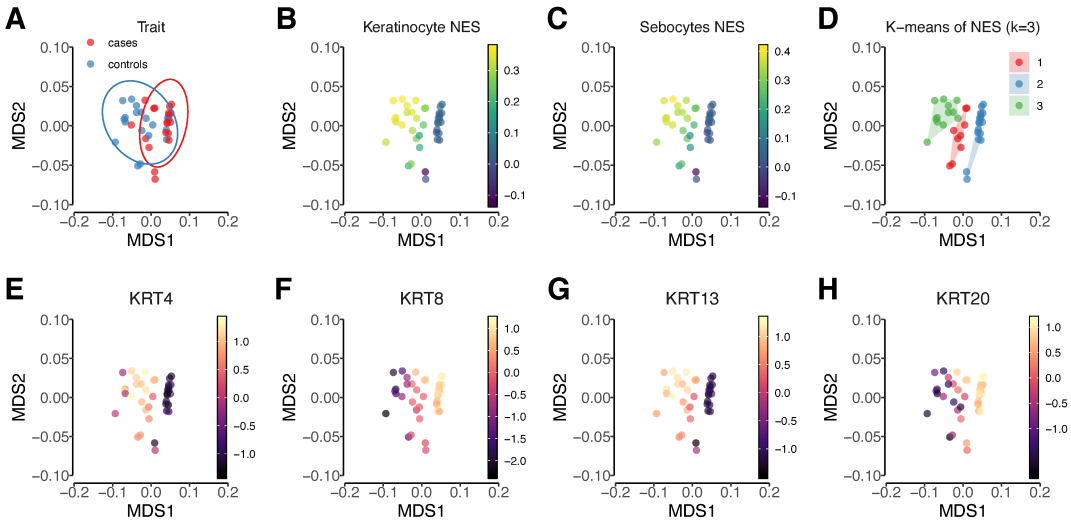


Online Supplementary Figure S6: ***Previously reported associations of HEM risk loci with other traits and diseases, clustered by biological areas.***

The plot shows associations with other traits, extracted from the GWAS ATLAS for the 102 lead SNPs (and/or their r^2>0.8 LD proxies) ordered by chromosome and chromosomal position. Associations are grouped by domain and represented with different colors. The ribbon size is proportional to the number of traits associated at the genome-wide significance level ($P_{Meta}<5\times10^{-8}$). Columns from left to right: Lead SNP – marker showing strongest association signal from each locus; Chr – chromosome; Pos – SNP position on chromosome (genome build hg19); Nearest gene (#genes within locus boundaries) – gene closest to the lead SNP (if within 100 kb distance, otherwise "na") (**Methods**).

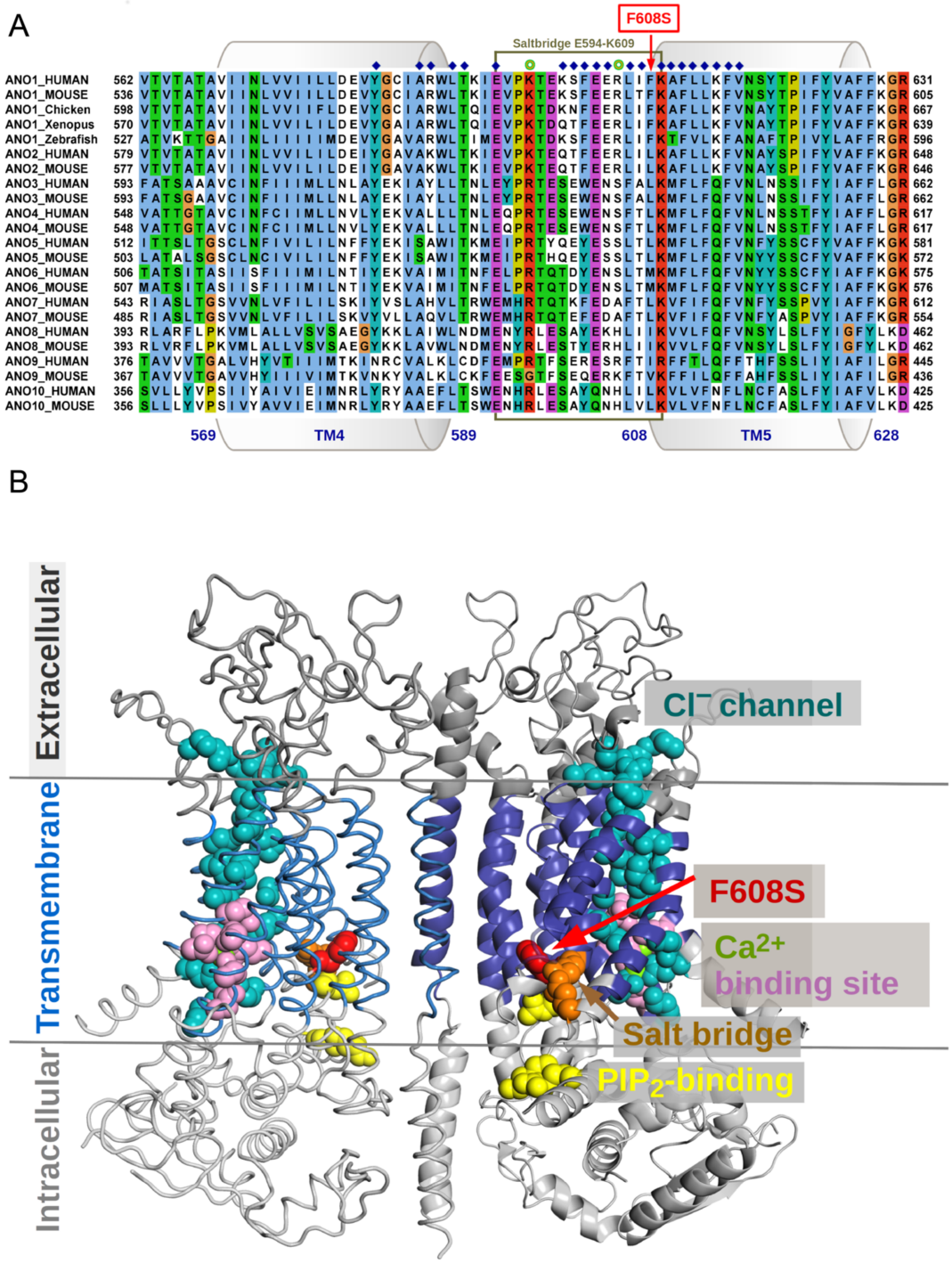


Online Supplementary Figure S7. ***Gene set enrichment analyses of HEM genes.*** Tissues and pathways are shown, which resulted significantly enriched in at least 2/3 analyses (using FUMA, MAGMA or DEPICT generate HEM gene lists; **Methods**). Gene Ontology Biological Processes (GOBP) and Molecular Function (GOMF) categories are reported; ns=non-significant; some tissues/pathways were not available in all analyses (missing bars).



Online Supplementary Figure S8. ***Gene signature-based determination of anal canal zones.***

Multidimensional scaling (MDS) analysis of the transcriptome data using Spearman's correlation distance (1 – correlation coefficient); **(A)** Colored by trait status, where the cases are enlarged hemorrhoidal tissue samples and controls are healthy hemorrhoidal tissue; **(B)** Colored by normalized enrichment score (NES) of keratinocyte cells; **(C)** Colored by NES of sebocytes; **(D)** Colored by clusters obtained by applying the k-means algorithm; **(E-H)** Colored by normalized expression values of anal canal marker genes, including *KRT4*, *KRT8*, *KRT13* and *KRT20.*



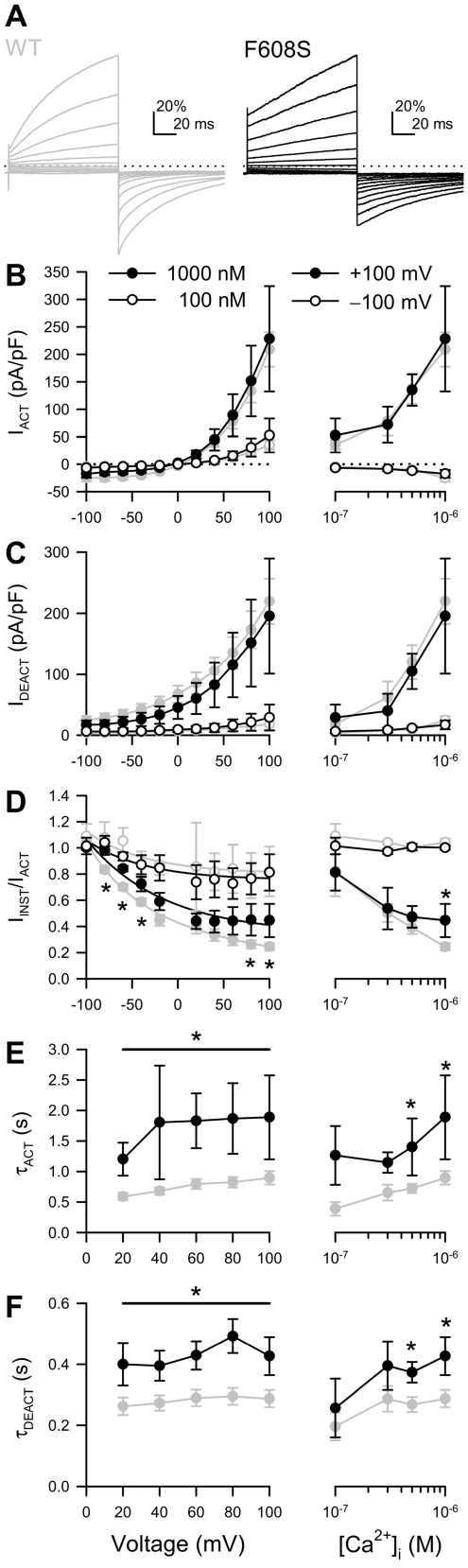
Online Supplementary Figure S9. ***ANO1 Alignment of TM4-5 and ANO1 structure.***

(A) Protein sequence alignment of ANO1 transmembrane helices TM4 and TM5 including the intracellular linker connected to TM5 via a salt bridge. Blue diamonds

mark conserved amino acid positions next to and in close proximity to the F608S variant. Green spheres mark PIP2-binding positions K597 and R605, whose mutation has been shown to lead to rapid channel inactivation through increased desensitization to Ca^{2+}. The same effect was observed with the mutation of E594 or K609 which form a stabilizing salt bridge[69]. Only hydrophobic amino acids (blue) are conserved at the site of the F608S variant. It is therefore predicted that the mutation to the polar serine destabilizes the local protein structure and affects the integrity of this salt bridge. Accelerated desensitization of the anion channel may result from conformational changes of the putative PIP2 binding site due to a disruption of the salt bridge[69]**. Consequently, F608S may be able to down-regulate ANO1 activity.**

(B) Structural model of the ANO1 dimer and localization of the F608S variant. The F608S variant (red spheres) is located at the beginning of transmembrane helix 5 and thus at the membrane-cytosolic interface and a predicted PIiP2 interaction site[69] (yellow spheres). The exchange of the hydrophobic sidechain of phenylalanine (F) to a polar serine (S) within a conserved hydrophobic region is expected to destabilize the structure by disrupting the stability conducted by the salt bridge of K609 and E594 (orange spheres), which could accelerate the down-regulation of ANO1 by a faster channel inactivation by desensitization to Ca^{2+}. This effect was shown by an alanine mutation of the salt bridge[69]. The ANO1 structural model is based on cryo-electron microscopy of the murine homolog (PDB ID 5oyb[70]) The two monomers are distinguished by representation as ribbons and cartoons, respectively.

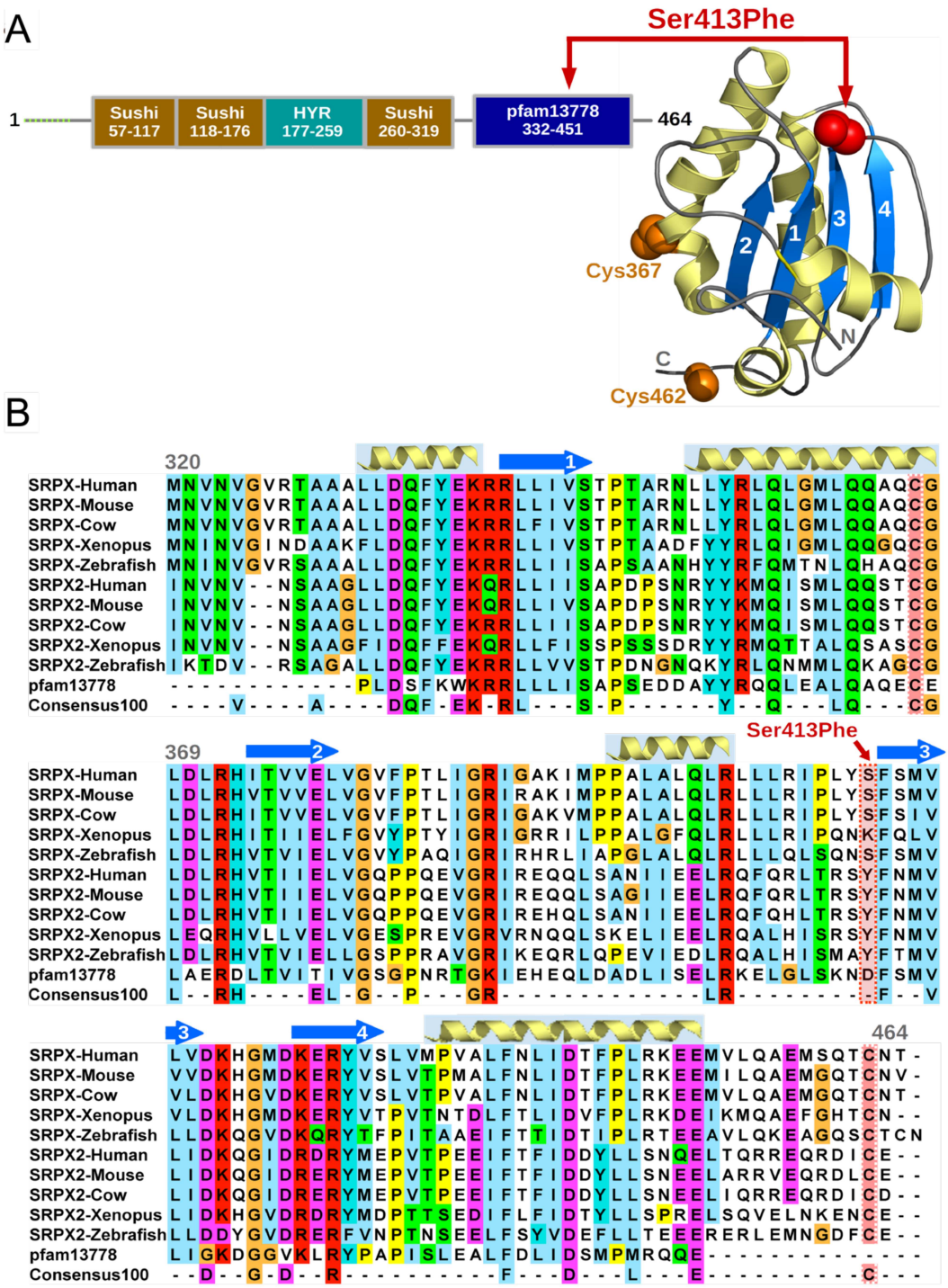
Extracellular and intracellular domains are colored dark and light grey, the transmembrane domain is blue. The Cl anion channel is highlighted as teal spheres. The Ca^{2+}-binding site is colored pink, the calcium atoms are shown as yellow-green spheres. Predicted PIP2-interacting residues (R481, K597 and R605[69]) are depicted as yellow spheres. Protein sequences were derived from the UniProt sequence database and visualized using JalView[95] with the Clustal X color scheme.


A
WT
F608S
20%
20 ms
20%
20 ms
B
1000 nM
100 nM
+100 mV
−100 mV
I_{ACT} (pA/pF)
C
I_{DEACT} (pA/pF)
D
I_{INST}/I_{ACT}
E
τ_{ACT} (s)
F
τ_{DEACT} (s)
Voltage (mV)
$[Ca^{2+}]_i$ (M)

Online Supplementary Figure S10. F608S mutant of ANO1 has high instantaneous current but slow voltage-dependent activation and deactivation kinetics *in vitro*.

(A) Representative Cl^- currents recorded from HEK293 cells transfected with wild-type ANO1 (left) or F608S-ANO1 (right), elicited by stepping for 1 s from -100 mV holding voltage to 100 through +100 mV**. (B-F)** Left, voltage-dependence at 1000 (●) or 100 nM $[Ca^{2+}]_i$ (○); or right, $[Ca^{2+}]_i$-dependence at +100 (●) or −100 mV (○); of Cl^- current parameters from HEK293 cells expressing WT- (gray) or F608S-ANO1 (black): Cl^- current densities at the 1-s plateau (**B**, IACT), tail currents immediately upon deactivation (**C**, I_{DEACT}), ratios of the instantaneous Cl^- current at 20 ms versus the plateau current at 1 s (**D**, I_{INST}/I_{ACT}), time constants of Cl^- current during activation (**E**, τ_{ACT}) or deactivation (**F**, τ_{DEACT}) (*P < 0.05, F608S vs. WT, by unpaired two-tailed t-test; n = 5-27 cells per $[Ca^{2+}]_i$).

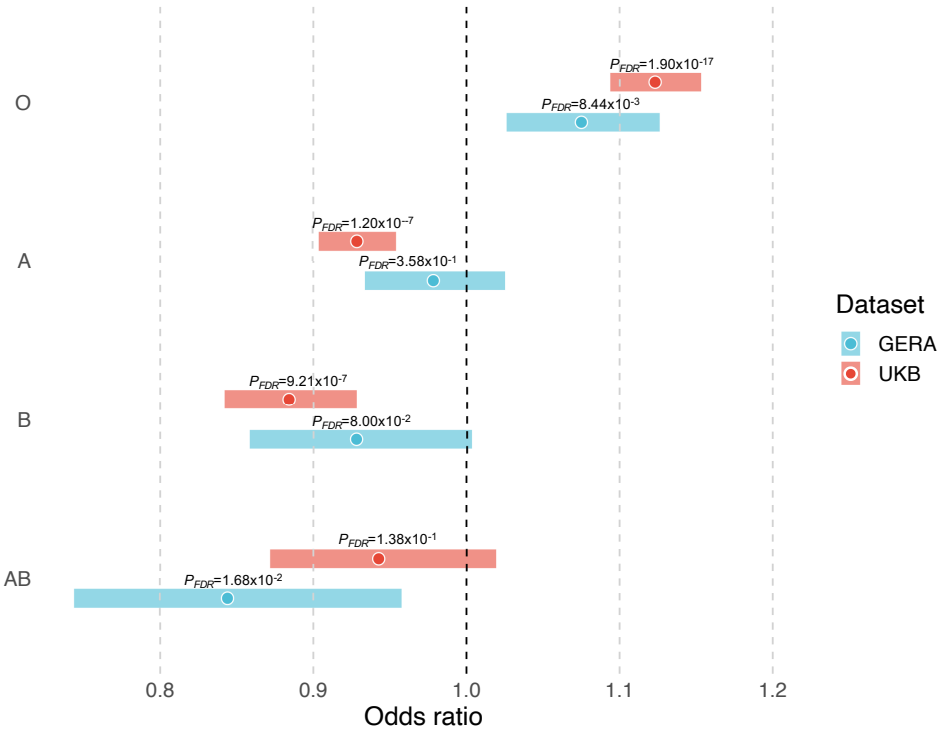
To determine the functional impact of F608S on human ANO1, we recorded whole-cell voltage-dependent Ca^{2+}-activated Cl^- currents from HEK293 cells expressing wild-type or F608S-ANO1 at 100-1000 nM intracellular Ca^{2+} ($[Ca^{2+}]_i$). Ca^{2+}-activated Cl^- current densities of F608S-ANO1 were similar to WT for both activation **(B)** and deactivation **(C)** at all tested voltages and $[Ca^{2+}]_i$ concentrations. However, the kinetics of the two constructs were different. F608S-ANO1 Cl^- currents had a larger ratio of instantaneous-to-plateau current (I_{INST}/I_{ACT}) at high $[Ca^{2+}]_i$ **(D)**. Moreover, F608S-ANO1 activated and deactivated slower than WT, as reflected in an increase in the time constants of activation (τ_{ACT}, **E**) and deactivation (τ_{DEACT}, **F**) at positive voltages (+20 to +100 mV) and high $[Ca^{2+}]_i$ (500-1000 nM).


A
Ser413Phe
1
Sushi 57-117
Sushi 118-176
HYR 177-259
Sushi 260-319
pfam13778 332-451
464
Cys367
Cys462
C
N
1
2
3
4
B
320
369
464
Ser413Phe
SRPX-Human
SRPX-Mouse
SRPX-Cow
SRPX-Xenopus
SRPX-Zebrafish
SRPX2-Human
SRPX2-Mouse
SRPX2-Cow
SRPX2-Xenopus
SRPX2-Zebrafish
pfam13778
Consensus100

Online Supplementary Figure S11. ***Sushi repeat-containing protein (SRPX) structure und alignment.***

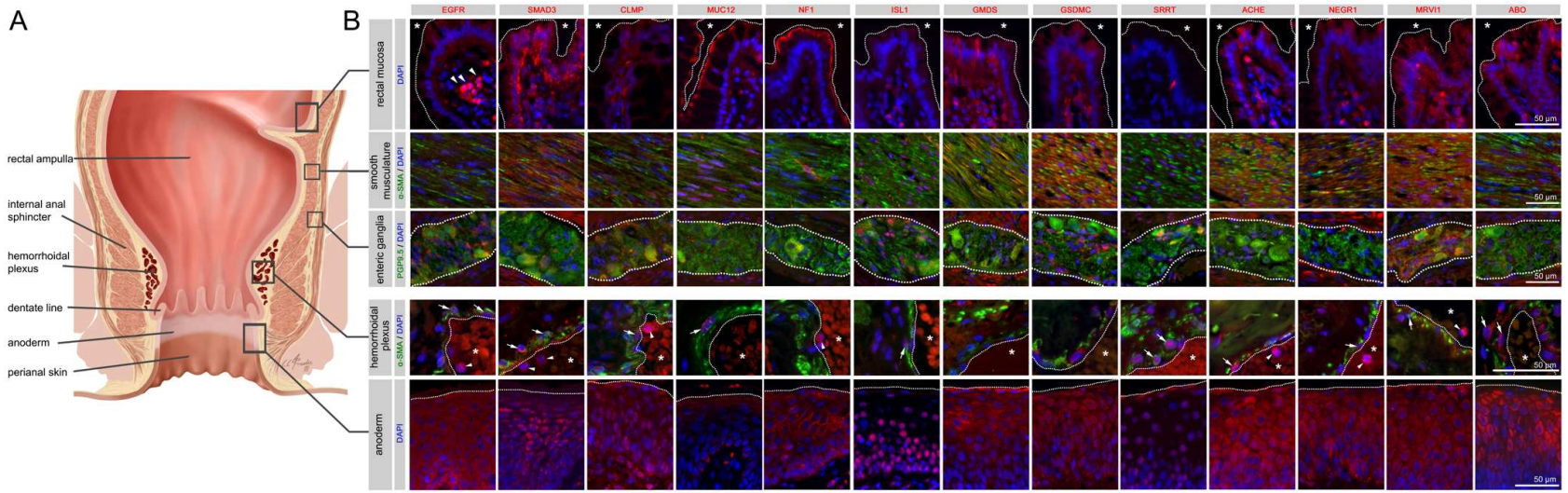
(A) SRPX domain structure and the predicted protein fold of the C-terminal domain. The N-terminal signal peptide is shown as a green dashed line. Predicting the 3D location of the Ser413Phe variant is based on a model with lower confidence, with loop and helical structures being less reliable than the central beta sheet. In this model it is predicted that the polar Ser413 stabilizes loops originating from strands 1, 3 and 4, and a mutation to a hydrophobic phenylalanine could interfere with this function **(B)** Multiple sequence alignment with predicted secondary structures. Conserved sequence positions are largely consistent with the pfam13778 family, in particular with the central beta sheet, which enhances the confidence of the core regions in the above structural model. Ser413Phe is located adjacent to the conserved beta strand 3 and the invariant Phe414 which supports an important structural role of the variant. For further details see **Methods**, section ***In silico* variant protein analysis**.

Zheng T, *et al. Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868



Online Supplementary Figure S12. ***ABO blood groups and HEM risk in UKBB and GERA.***

The plot shows odds ratios (and 95% confidence intervals) form testing ABO blood groups vs HEM risk in UKBB and GERA (**Methods**). An association test based on logistic regression is used to test for a significant HEM association for each of the four blood groups, taking into account sex, age, BMI and the top 10 PCs from PCA. FDR correction was applied to correct for multiple testing. FDR: false discovery rate.



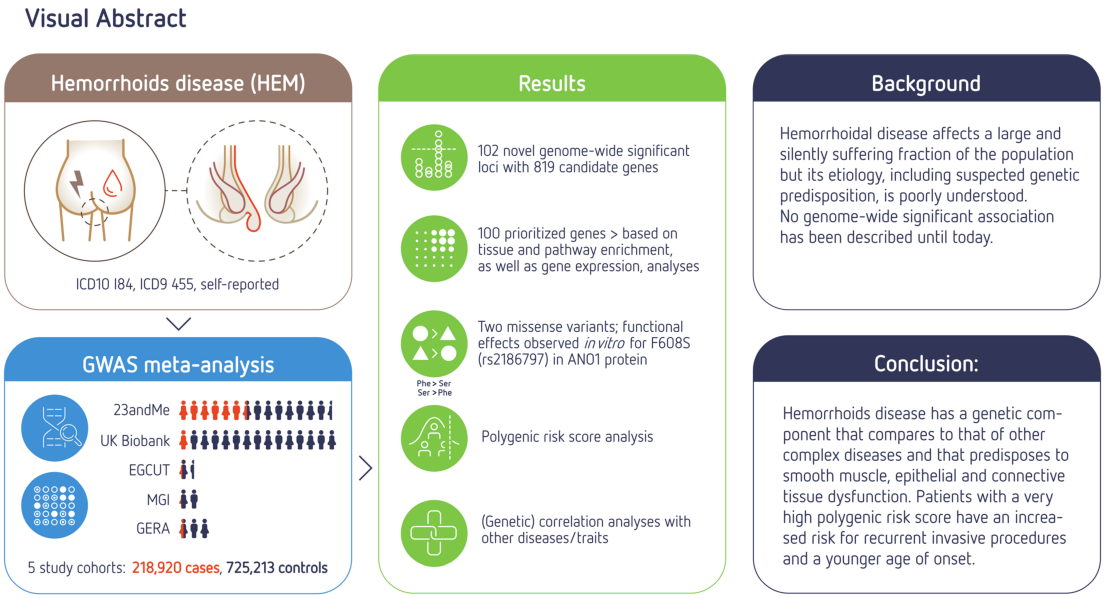
Online Supplementary Figure S13. ***Immunohistochemistry for selected HEM candidate proteins.***

Illustration of the rectum and anal canal **(A)** with indication the site-specific localization of the immunohistochemical panels analyzed in **(B).** Fluorescence immunohistochemistry (B) for selected HEM candidate proteins (see also **online supplementary table S11**), encoded by candidate genes within our 102 identified genome-wide significant loci, are shown. *SRPX* (rs35318931), *ANO1* (rs2186797) and *MYH11* (rs6498573) were determined as prioritized HEM genes in our study. *ANO1* and *SRPX* are interesting HEM candidate genes since the lead SNPs at these loci are (missense) coding variants. *MYH11* is also a main hub gene within the M1 co-expression module of our transcriptome analysis. Given the ABO blood group association observed in our study in HEM patients (**online supplementary figure S12**), we have included ABO as further target for immunohistochemistry.

Antibody staining was performed on colorectal FFPE tissue specimens from control individuals. The rows correspond to the rectal mucosa (top row, epithelial surface delimited by dashed line,*: intestinal lumen), smooth musculature (second row), enteric ganglia

(third row, ganglionic boundaries delimited by dashed line), hemorrhoidal plexus (fourth row, endothelial surface delimited by dashed line, *: vascular lumen), and the anoderm (bottom row, border of the anoderm delimited by dashed line). Blue: DAPI; green: α-SMA (anti-alpha smooth muscle actin antibody) for row 2 and 4 (smooth musculature/hemorrhoidal plexus) and PGP9.5 (member of the ubiquitin hydrolase family of proteins, neuronal marker) for row 3 (enteric ganglia); red: antibody for the respective candidate protein. Arrows point to corresponding candidate-positive cells within the vascular wall. Arrowheads point to corresponding candidate-positive nucleated immune cells.

Zheng T, *et al*. *Gut* 2021; 70:1538–1549. doi: 10.1136/gutjnl-2020-323868



Online Supplementary Figure S14. ***Graphical abstract of the study.***

SUPPLEMENTARY REFERENCES

1 Tung JY, Do CB, Hinds DA, Kiefer AK, Macpherson JM, Chowdry AB*, et al.* Efficient replication of over 180 genetic associations with self-reported medical data. PLoS One 2011;**6**:e23473.
2 Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K*, et al.* The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;**562**:203-9.
3 Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H*, et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol 2015;**44**:1137-47.
4 Fritsche LG, Gruber SB, Wu Z, Schmidt EM, Zawistowski M, Moser SE*, et al.* Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. Am J Hum Genet 2018;**102**:1048-61.
5 Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE, Schreiber S. PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. Community Genet 2006;**9**:55-61.
6 Krokstad S, Langhammer A, Hveem K, Holmen TL, Midthjell K, Stene TR*, et al.* Cohort Profile: the HUNT Study, Norway. Int J Epidemiol 2013;**42**:968-77.
7 Hansen TF, Banasik K, Erikstrup C, Pedersen OB, Westergaard D, Chmura PJ*, et al.* DBDS Genomic Cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors. BMJ Open 2019;**9**:e028401.
8 Burgdorf KS, Simonsen J, Sundby A, Rostgaard K, Pedersen OB, Sorensen E*, et al.* Socio-demographic characteristics of Danish blood donors. PLoS One 2017;**12**:e0169112.
9 Jensen TB, Jimenez-Solem E, Cortes R, Betzer C, Boge Breinholt S, Meidahl Petersen K*, et al.* Content and validation of the Electronic Patient Medication module (EPM)-the administrative in-hospital drug use database in the Capital Region of Denmark. Scand J Public Health 2020;**48**:43-8.
10 Durand EY, Do CB, Mountain JL, Macpherson JM. Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution. bioRxiv 2014:010512.
11 Henn BM, Hon L, Macpherson JM, Eriksson N, Saxonov S, Pe'er I*, et al.* Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One 2012;**7**:e34267.
12 Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM*, et al.* A map of human genome variation from population-scale sequencing. Nature 2010;**467**:1061-73.
13 Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;**81**:1084-97.
14 Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics 2015;**31**:782-4.
15 Wain LV, Shrine N, Miller S, Jackson VE, Ntalla I, Soler Artigas M*, et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. Lancet Respir Med 2015;**3**:769-81.
16 Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. PLoS One 2014;**9**:e93766.

17 International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL*, et al.* A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;**449**:851-61.
18 Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods 2011;**9**:179-81.
19 Mitt M, Kals M, Parn K, Gabriel SB, Lander ES, Palotie A*, et al.* Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet 2017;**25**:869-76.
20 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**:e1000529.
21 Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR*, et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 2012;**91**:839-48.
22 Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics 2010;**26**:2867-73.
23 Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S*, et al.* Worldwide human relationships inferred from genome-wide patterns of variation. Science 2008;**319**:1100-4.
24 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;**4**:7.
25 McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A*, et al.* A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016;**48**:1279-83.
26 Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;**2**:e190.
27 Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 2012;**44**:955-9.
28 Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV*, et al.* Long-range LD can confound genome scans in admixed populations. Am J Hum Genet 2008;**83**:132-5; author reply 5-9.
29 Jun TH, Rouf Mian MA, Michel AP. Genetic mapping revealed two loci for soybean aphid resistance in PI 567301B. Theor Appl Genet 2012;**124**:13-22.
30 Guo Y, He J, Zhao S, Wu H, Zhong X, Sheng Q*, et al.* Illumina human exome genotyping array clustering and quality control. Nat Protoc 2014;**9**:2643-62.
31 Wang C, Zhan X, Liang L, Abecasis GR, Lin X. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. Am J Hum Genet 2015;**96**:926-37.
32 Loh PR, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet 2016;**48**:811-6.
33 Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A*, et al.* Next-generation genotype imputation service and methods. Nat Genet 2016;**48**:1284-7.
34 Konig IR, Loley C, Erdmann J, Ziegler A. How to include chromosome X in your genome-wide association study. Genet Epidemiol 2014;**38**:97-103.
35 Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM*, et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 2015;**47**:284-90.
36 Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN*, et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet 2018;**50**:1335-41.

37 Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Magi R*, et al.* Quality control and conduct of genome-wide association meta-analyses. Nat Protoc 2014;**9**:1192-212.
38 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;**26**:2190-1.
39 Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun 2017;**8**:1826.
40 Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J*, et al.* Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun 2015;**6**:5890.
41 de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 2015;**11**:e1004219.
42 Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP*, et al.* LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010;**26**:2336-7.
43 Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 2016;**32**:1493-501.
44 Cuéllar-Partida G, Lundberg M, Kho PF, D'Urso S, Gutiérrez-Mondragón LF, Ngo TT*, et al.* Complex-Traits Genetics Virtual Lab: A community-driven web platform for post-GWAS analyses. bioRxiv 2019:518027.
45 Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P*, et al.* A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. Am J Hum Genet 2012;**90**:821-35.
46 Schafmayer C, Harrison JW, Buch S, Lange C, Reichert MC, Hofer P*, et al.* Genome-wide association analysis of diverticular disease points towards neuromuscular, connective tissue and epithelial pathomechanisms. Gut 2019;**68**:854-65.
47 Bonfiglio F, Zheng T, Garcia-Etxebarria K, Hadizadeh F, Bujanda L, Bresso F*, et al.* Female-Specific Association Between Variants on Chromosome 9 and Self-Reported Diagnosis of Irritable Bowel Syndrome. Gastroenterology 2018;**155**:168-79.
48 Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience 2019;**8**.
49 Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J*, et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics 2019;**35**:4851-3.
50 Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C*, et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019;**47**:D1005-D12.
51 Watanabe K, Stringer S, Frei O, Umicevic Mirkov M, de Leeuw C, Polderman TJC*, et al.* A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet 2019;**51**:1339-48.
52 Jensen AB, Moseley PL, Oprea TI, Ellesoe SG, Eriksson R, Schmock H*, et al.* Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun 2014;**5**:4022.
53 Rozowsky J, Kitchen RR, Park JJ, Galeev TR, Diao J, Warrell J*, et al.* exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling. Cell Syst 2019;**8**:352-7 e3.
54 Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;**30**:207-10.
55 Fenger C. The anal transitional zone. Acta Pathol Microbiol Immunol Scand Suppl 1987;**289**:1-42.

56 Iacobuzio-Donahue CA. Inflammatory and Neoplastic Disorders of the Anal Canal. In: Robert Odze JG, ed. Surgical Pathology of the GI Tract, Liver, Biliary Tract, and Pancreas: Elsevier Inc., 2009:733-61.
57 Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 2017;**18**:220.
58 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;**15**:550.
59 Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF*, et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009;**462**:108-12.
60 Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 2013;**14**:7.
61 Williams GR, Talbot IC, Northover JM, Leigh IM. Keratin expression in the normal anal canal. Histopathology 1995;**26**:39-44.
62 McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 2012;**40**:4288-97.
63 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;**9**:559.
64 Russo PST, Ferreira GR, Cardozo LE, Burger MC, Arias-Carrasco R, Maruyama SR*, et al.* CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinformatics 2018;**19**:56.
65 Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;**16**:284-7.
66 Severe Covid GG, Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A*, et al.* Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med 2020;**383**:1522-34.
67 Cossais F, Lange C, Barrenschee M, Moding M, Ebsen M, Vogel I*, et al.* Altered enteric expression of the homeobox transcription factor Phox2b in patients with diverticular disease. United European Gastroenterol J 2019;**7**:349-57.
68 De Jesus-Perez JJ, Cruz-Rangel S, Espino-Saldana AE, Martinez-Torres A, Qu Z, Hartzell HC*, et al.* Phosphatidylinositol 4,5-bisphosphate, cholesterol, and fatty acids modulate the calcium-activated chloride channel TMEM16A (ANO1). Biochim Biophys Acta Mol Cell Biol Lipids 2018;**1863**:299-312.
69 Le SC, Jia Z, Chen J, Yang H. Molecular basis of PIP2-dependent regulation of the Ca(2+)-activated chloride channel TMEM16A. Nat Commun 2019;**10**:3769.
70 Paulino C, Kalienkova V, Lam AKM, Neldner Y, Dutzler R. Activation mechanism of the calcium-activated chloride channel TMEM16A revealed by cryo-EM. Nature 2017;**552**:421-5.
71 Dang S, Feng S, Tien J, Peters CJ, Bulkley D, Lolicato M*, et al.* Cryo-EM structures of the TMEM16A calcium-activated chloride channel. Nature 2017;**552**:426-9.
72 Kirkitadze MD, Barlow PN. Structure and flexibility of the multiple domain proteins that regulate complement activation. Immunol Rev 2001;**180**:146-61.
73 Callebaut I, Gilges D, Vigon I, Mornon JP. HYR, an extracellular module involved in cellular adhesion and related to the immunoglobulin-like fold. Protein Sci 2000;**9**:1382-90.
74 Bommer GT, Jager C, Durr EM, Baehs S, Eichhorst ST, Brabletz T*, et al.* DRO1, a gene down-regulated by oncogenes, mediates growth inhibition in colon and pancreatic cancer cells. J Biol Chem 2005;**280**:7962-75.

75 Pawlowski K, Muszewska A, Lenart A, Szczepinska T, Godzik A, Grynberg M. A widespread peroxiredoxin-like domain present in tumor suppression- and progression-implicated proteins. BMC Genomics 2010;**11**:590.
76 Inoue H, Pan J, Hakura A. Suppression of v-src transformation by the drs gene. J Virol 1998;**72**:2532-7.
77 Tambe Y, Yoshioka-Yamashita A, Mukaisho K, Haraguchi S, Chano T, Isono T*, et al.* Tumor prone phenotype of mice deficient in a novel apoptosis-inducing gene, drs. Carcinogenesis 2007;**28**:777-84.
78 Tambe Y, Hasebe M, Kim CJ, Yamamoto A, Inoue H. The drs tumor suppressor regulates glucose metabolism via lactate dehydrogenase-B. Mol Carcinog 2016;**55**:52-63.
79 Burgstaller G, Oehrle B, Gerckens M, White ES, Schiller HB, Eickelberg O. The instructive extracellular matrix of the lung: basic composition and alterations in chronic lung disease. Eur Respir J 2017;**50**.
80 Wilson R, Norris EL, Brachvogel B, Angelucci C, Zivkovic S, Gordon L*, et al.* Changes in the chondrocyte and extracellular matrix proteome during post-natal mouse cartilage development. Mol Cell Proteomics 2012;**11**:M111 014159.
81 Naba A, Clauser KR, Whittaker CA, Carr SA, Tanabe KK, Hynes RO. Extracellular matrix signatures of human primary metastatic colon cancers and their metastases to liver. BMC Cancer 2014;**14**:518.
82 Perea-Gil I, Uriarte JJ, Prat-Vidal C, Galvez-Monton C, Roura S, Llucia-Valldeperas A*, et al.* In vitro comparative study of two decellularization protocols in search of an optimal myocardial scaffold for recellularization. Am J Transl Res 2015;**7**:558-73.
83 Burnicka-Turek O, Kata A, Buyandelger B, Ebermann L, Kramann N, Burfeind P*, et al.* Pelota interacts with HAX1, EIF3G and SRPX and the resulting protein complexes are associated with the actin cytoskeleton. BMC Cell Biol 2010;**11**:28.
84 Royer-Zemmour B, Ponsole-Lenfant M, Gara H, Roll P, Leveque C, Massacrier A*, et al.* Epileptic and developmental disorders of the speech cortex: ligand/receptor interaction of wild-type and mutant SRPX2 with the plasminogen activator receptor uPAR. Hum Mol Genet 2008;**17**:3617-30.
85 Song X, Tanaka H, Ohta K. Multiple roles of Equarin during lens development. Dev Growth Differ 2014;**56**:199-205.
86 O'Leary EE, Mazurkiewicz-Munoz AM, Argetsinger LS, Maures TJ, Huynh HT, Carter-Su C. Identification of steroid-sensitive gene-1/Ccdc80 as a JAK2-binding protein. Mol Endocrinol 2013;**27**:619-34.
87 Nasseri YY, Krott E, Van Groningen KM, Berho M, Osborne MC, Wollman S*, et al.* Abnormalities in collagen composition may contribute to the pathogenesis of hemorrhoids: morphometric analysis. Tech Coloproctol 2015;**19**:83-7.
88 Mazzone A, Gibbons SJ, Bernard CE, Nowsheen S, Middha S, Almada LL*, et al.* Identification and characterization of a novel promoter for the human ANO1 gene regulated by the transcription factor signal transducer and activator of transcription 6 (STAT6). FASEB J 2015;**29**:152-63.
89 Strege PR, Bernard CE, Mazzone A, Linden DR, Beyder A, Gibbons SJ*, et al.* A novel exon in the human Ca2+-activated Cl- channel Ano1 imparts greater sensitivity to intracellular Ca2. Am J Physiol Gastrointest Liver Physiol 2015;**309**:G743-9.
90 Ferrera L, Caputo A, Ubby I, Bussani E, Zegarra-Moran O, Ravazzolo R*, et al.* Regulation of TMEM16A chloride channel properties by alternative splicing. J Biol Chem 2009;**284**:33360-8.
91 Strege PR, Gibbons SJ, Mazzone A, Bernard CE, Beyder A, Farrugia G. EAVK segment "c" sequence confers Ca(2+)-dependent changes to the kinetics of full-length human Ano1. Am J Physiol Gastrointest Liver Physiol 2017;**312**:G572-G9.

92 Margetis N. Pathophysiology of internal hemorrhoids. Ann Gastroenterol 2019;**32**:264-72.
93 Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;**55**:997-1004.
94 Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR*, et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 2015;**47**:1228-35.
95 Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 2009;**25**:1189-91.