Research Letter: Therapeutic targets for haemorrhoidal disease: proteome-wide Mendelian randomisation and colocalization analyses

Shifang Li[#]*, Meijiao Gong[#]

Laboratory of Immunology and Vaccinology, FARAH, ULiège, Liège 4000, Belgium.

[#]Shifang Li and Meijiao Gong contributed equally to this work

*Correspondence:

Shifang Li, fruceslee@gmail.com

Laboratory of Immunology and Vaccinology, FARAH, ULiège, Liège 4000, Belgium

Abstract

Human haemorrhoidal disease (HEM) is a common anorectal pathology. However, the etiology of HEM, as well as its molecular mechanism, remains largely unclear. In this study, we applied a two-sample bi-direction Mendelian randomisation (MR) analysis to estimate the causal effects of 4907 plasma proteins on HEM outcomes and investigated the mediating impacts of plasma proteins on HEM risk factors to uncover potential HEM treatment targets by integrating GWASs statistics of HEM and plasma protein levels. Following MR analysis, our study identified 5 probable causal proteins associated with HEM. ERLEC1 and ASPN levels were genetically predicted to be positively and inversely associated with HEM risk, respectively, with strong evidence of colocalization (H4>0.9). The findings of an independent cohort corroborate the causal relationship between these two proteins and HEM. Furthermore, gene expression analysis of haemorrhoidal tissue and normal specimens revealed that ERLEC1 but not ASPN were differentially expressed. By analyzing single-cell ERLEC1 expression in human rectum tissues, ERLEC1 was found to be highly expressed in transient-amplifying cells. Interestingly, a genetically greater risk of myxoedema was linked to an elevated risk of HEM. However, there was no evidence that dorsalgia, hernia, diverticular disease, and ankylosing spondylitis were causally associated with HEM. Furthermore, no association was found between myxoedema and the genetically predicted ERLEC1 and ASPN levels. Overall, this study identified some causal associations of circulating proteins and risk factors with HEM by integrating the largest-to-date plasma proteome and GWASs of

HEM. The findings could provide further insight into understanding biological mechanisms for HEM.

Keywords

Haemorrhoidal disease, Mendelian randomisation, ERLEC1, myxoedema

Human haemorrhoidal disease (HEM) is a common anorectal disorder. Recently, Zhang *et al.* reported the first and largest genome-wide association study (GWAS) with haemorrhoidal disease (HEM), and these data offered us a resource for understanding the genetic risk factors for HEM.[1] However, the etiology of HEM, as well as its molecular mechanism, remains primarily unclear.[2] In addition, the identification of genes with therapeutic effects needs to be conducted. In recent years, by incorporating protein quantitative trait loci (pQTLs) into MR analysis, such an approach has been successfully used to prioritize therapy targets.[3,4] Here, using a two-sample bidirectional Mendelian randomisation (MR) analysis, we estimated the causal effects of 4907 plasma proteins on HEM outcomes, and investigated the effects of plasma proteins that may mediate the impact of risk factors on HEM in order to identify potential therapeutic targets for HEM.

As stated in the **supplementary methods**, 4907 proteins (*cis*-pQTLs) were used as instrumental variables for exposure and HEM as the outcome to estimate the causal effect of plasma protein levels on HEM in a proteome-wide context using MR analysis.[5-8] Our study revealed 5 potential causative proteins at the Bonferroni-corrected threshold of $p<1.01\times10^{-5}$, including 3 negative and 2 positive

associations (**figure 1A-1B**). MR analysis, for example, revealed that genetically predicted ERLEC1 levels were linked to an increased risk of HEM (p=5.18e-07). To determine whether the identified relationships of the circulating protein with HEM shared causative variations, colocalization analysis was carried out and a high level of support for colocalization evidence was discovered between two proteins (ERLEC1 and ASPN) and HEM (H4>0.9) (**figure 1C**). The findings of the INTERVAL cohort corroborated the causal relationship between these two proteins and HEM (**figure 1D**).[9] Interestingly, the deCODE study's lead cis-pQTL for the ERLEC1 (rs2542580) but not ASPN (rs10992273) were not found to be associated with all available secondary traits (**supplementary table1**). Gene expression analysis of haemorrhoidal tissue and normal specimens revealed that ERLEC1 but not ASPN were differentially expressed after controlling for gender and BMI (**figure 1E**), further supporting that a high ERLEC1 expression level was associated with an increased risk of HEM. Following that, we investigated the tissues in which ERLEC1 is expressed in bulk tissues using GTEx v8 (https://gtexportal.org/), and found that ERLEC1 was considerably expressed in multiple tissues, including the small intestine and colon, as compared to the whole blood (p<0.001) (**figure 1F**). To further understand the origin of ERLEC1, single-cell ERLEC1 expression was assessed in human rectum tissues, and ERLEC1 was found to be highly expressed in transient-amplifying (TA) cells (p< 0.05) (**figure 1G**).[10]

In order to investigate whether the causal protein mediates the effect of risk factors on HEM, the causal risk factors for HEM were first identified. 5 clinical traits

that genetically correlated with HEM were selected (**supplementary methods**), with instrumental variables generated from GWASs confined to European populations. It was discovered that a genetically greater risk of myxoedema was linked to an elevated risk of HEM (p<0.05) (**figure 1H**). Although genetic correlations with HEM were reported,[1] there was no evidence that dorsalgia, hernia, diverticular disease, and ankylosing spondylitis were causally associated (p>0.05). In order to identify the protein related to HEM risk factors, we conducted MR analysis again on 2 plasma proteins impacting HEM with myxoedema. After filtering, there was a lack of evidence that myxoedema had a causal relationship with these two plasma proteins (**figure 1I**).

Overall, by integrating the largest-to-date plasma proteome and GWAS of HEM, we discovered that ERLEC1 could serve as prospective protein therapeutic targets for HEM. In-depth research is needed to investigate the mechanisms by which putative risk factors affect HEM (**figure 1J**).

Competing interests

None declared.

Contributors

SF was involved in conceptualization. SF and MJ were involved in the formal analysis. SF was involved in writing, reviewing, and editing.

Acknowledgments

The authors would like to thank all of the researchers who contributed to the GWAS datasets used in this study for making them available for research purposes.

References

1 Zheng T, Ellinghaus D, Juzenas S, *et al*. Genome-wide analysis of 944 133 individuals provides insights into the etiology of haemorrhoidal disease. *Gut* 2021;70:1538-49.

2 EAM Festen & RK Weersma. Large-scale genetic analyses in an understudied disease: haemorrhoidal disease. *Gut 2021*;70:1429-1430.

3 Bovijn J, Lindgren CM & Holmes MV. Genetic variants mimicking therapeutic inhibition of IL-6 receptor signaling and risk of COVID-19. *The Lancet Rheumatology* 2020;2:e658-9.

4 Dewey, F. E. et al. Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. *N Engl J Med* 2017;377:211-21.

5 Ferkingstad E, Sulem P, Atlason BA, *et al*. Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* 2021;53:1712-21.

6 Zheng J, Haberland V, Baird D, *et al*. Phenome-wide Mendelian randomisation mapping the influence of the plasma proteome on complex diseases. *Nat Genet* 2020;52:1122-31.

7 Chen L, Peters JE, Prins B, *et al*. Systematic Mendelian randomisation using the human plasma proteome to discover potential therapeutic targets for stroke. *Nat Commun* 2022;13:1-14.

8 Yoshiji S, Butler-Laporte G, Lu T, *et al*. Proteome-wide Mendelian randomisation implicates nephronectin as an actionable mediator of the effect of obesity on COVID-19 severity. *Nat Metab* 2023;5:248-64.

9 Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, *et al*. Genomic atlas of the human plasma proteome. *Nature* 2018;558, 73-79.

10 Wang Y, Song W, Wang J, *et al*. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. *J Exp Med* 2020;217(2):e20191130.

Figure Legends

Figure 1 Mendelian randomisation results. (**A**) The effect of plasma protein levels on HEM. Volcano plot indicating the effect of plasma protein on HEM using MR analysis. (**B**) Forest plots shows the effect of plasma ERLEC1 and ASPN levels on HEM. (**C**) Colocalization analysis of ERLEC1 levels (Up) and ASPN (Down). (**D**) Forest plots shows the effect of plasma ERLEC1 and ASPN levels on HEM using INTERVAL cohort. (**E**) Boxplot shows differentially expressed genes in HEM patients when compared to healthy individuals. *p*-values were corrected the effect of gender and BMI using linear model. (**F**) The violin plot depicts ERLEC1 gene expression across multiple bulk tissues. (**G**) Data visualization of cell populations in human rectum tissues using UMAP (left) and gene expression of ERLEC1 in different cell types (right). (**H**) Forest plots showing the causal effect of chosen risk factors on HEM. (**I**) Forest plots for the effect of myxoedema on plasma ERLEC1 and ASPN levels. (**J**) Schematic illustration of the proposed model in the study. HEM, haemorrhoidal disease.

Supplementary Methods The statistics method used in the study.

Supplementary Tables1 The significant MR summary statistics obtained in this study.

Supplementary Methods

GWASs of haemorrhoidal disease and risk factors

We used recently published large-scale genome-wide associations (GWASs) for haemorrhoidal disease (HEM).[1] This GWAS summary statistics were derived from 944,133 European ancestry individuals (Ncase = 218,920 and Ncontrol = 725,213) from 5 cohorts and downloaded from the GWAS Catalog (https://www.ebi.ac.uk/gwas/, access ID: EFO_0009552). Diverticular disease of the intestine, ankylosing spondylitis (AS), dorsalgia, hernia, and myxoedema were evaluated as potential causal risk factors associated with HEM in order to determine the probable causal risk factors. All GWASs for the five risk factors were obtained from the ieu open gwas project (https://gwas.mrcieu.ac.uk/datasets/). The summary statistics of the large GWAS (14,357 cases and 182,423 controls) were used for diverticular disease of the intestine (access ID: finn-b-K11_DIVERTIC). The GWAS for ankylosing spondylitis (access ID: finn-b-M13_ANKYLOSPON) have a sample size of 1,462 cases and 164,682 controls. The GWAS for myxoedema (access ID: ieu-b-4877) has a sample size of 311,629 cases and 321,173 controls. The GWAS for dorsalgia (access ID: finn-b-M13_DORSALGIA) included 193467 individuals, with 28,785 cases and 164,682 controls. A total of 218792 individuals were reported with GWAS of hernia (access ID: finn-b-K11_HERNIA), including 28,235 cases and 190,557 controls.

Plasma protein quantitative trait loci (pQTL) data

To conduct proteome-wide Mendelian randomisation (MR), we first obtained

199 genetic instrumental variables using the protein quantitative trait loci (pQTL) data
200 generated by Ferkingstad *et al*.[2] The largest-to-date pQTL analysis on plasma
201 proteome (a total of 4907 proteins) in 35,559 Icelanders was performed in their study,
202 and an amount of 18,084 pQTL associations between genetic variation and protein
203 levels in plasma were identified. A total of 4907 pQTLs were successfully
204 downloaded from the deCODE study using aria2c.[3] To minimize the risk of horizontal
205 pleiotropy, instrumental variables to *cis*-pQTLs (SNPs located within a 500 kb
206 window from the target gene body) of protein were selected for the following analysis.
207 In order to validate the MR results for ASPN and ERLEC1 using independent study,
208 the *cis*-pQTLs of ASPN and ERLEC1 were obtained from the INTERVAL cohort and
209 used for the following analysis.[4]

210 **Mendelian randomisation analysis**

211 MR analysis is an analytical method that uses genetic variation as an
212 instrumental variable (IV) to estimate causal effects. It overcomes the limitations of
213 measurement error and confounding factors that are common in observational studies
214 and is widely used to assess causal relationships.[5] In this study, the TwoSampleMR
215 package (v0.5.6, https://mrcieu.github.io/TwoSampleMR/) was used for MR
216 analysis.[6] The instrumental variables that determined the exposure in each MR study
217 were specified as genome-wide significant ($p \leqslant$ 5e-08) SNPs. SNPs in the human
218 major histocompatibility complex (MHC) region at chromosome 6:
219 28,477,797-33,448,354 (GRCh37) were excluded from the analysis due to its complex
220 linkage disequilibrium (LD) structure. Using the 1000 Genomes Project European

reference panel and an LD threshold of r^2 <0.001 with a clumping window of 10,000 kb, PLINK v.1.9 (http://pngu.mgh.harvard.edu/purcell/plink/) was employed to derive instrumental variables.[7-8] F-statistics were used to determine the strength of each SNP's association with exposure, and F-statistics of more than 10 were considered strong. For the main MR analysis, the inverse variance weighted approach for proteins with two or more instrumental variables and the wald ratio method for proteins with a single instrumental variable was used for evaluating the causal influence of exposure on outcome. In addition, in the case of more instrumental variables used in MR analysis, four additional MR methods (weighted median, simple mode, weighted mode, and MR-Egger method) were used to assess the reliability of the primary results. For exposures with multiple IVs, we additionally investigated heterogeneity across variant-level MR estimations with the "mr_heterogeneity()" function in the TwoSampleMR package (Cochrane's Q test). In addition, a pleiotropy test was performed using MR Egger analysis to determine whether there is horizontal pleiotropy among IVs. Meanwhile, "phenoscanner" (https://github.com/phenoscanner/phenoscanner) was to be utilized to determine any pleiotropy of SNPs used in the MR analysis. An SNP was regarded to be pleiotropic when the reported SNP-traits association was genome-wide significant ($p \leqslant$5e-08) in the European population.

Finally, in the event there were more than two IVs in exposure, a leave-one-out analysis was performed, and the MR findings of the remaining IVs were calculated by deleting the IVs one by one to ensure the robustness of the MR data. To acquire robust

evidence for the casual estimation, MR findings that meet all of the following criteria were chosen as described by Yoshiji and others: (1) no pleiotropy was found using MR-Egger regression (p>0.05); (2) results with an $I^2 < 50\%$ (no substantial heterogeneity); (3) leave-one-out analysis MR p<0.05 after removing outliers; and (4) reverse MR p>0.05.[9] The same procedure as mentioned above was utilized to explore the causal effect of the given exposure and associated outcome in the reverse MR analysis. p-values less than a Bonferroni adjusting ($p=1.01\times10^{-5}$ (0.05/4,907)) are deemed significant for multiple testing.

Colocalization analysis

The coloc R package was employed to investigate whether the reported relationships between proteins and HEM were driven by linkage disequilibrium.[10] The analysis offers posterior probability for each hypothesis tested: no association in either group (H0), pQTL only (H1), the GWAS of HEM only (H2), associations with both GWAS but by separate causal signals (H3), and associations with both GWAS but by the same signals (H4).[11] A higher H4 (H4>0.8) was considered as strong evidence for colocalization, implying a shared variation between the two phenotypes.[10,11]

Differentially expressed genes analysis in bulk tissues

The GSE154650 dataset was downloaded from NCBI Gene Expression Omnibus (GEO) and analyzed using the R program.[12] The RPM value of ERLEC1 and ASPN were further subjected to linear model analysis to investigate the differential gene expression in HEM and healthy individuals after correcting for the effects of gender

and BMI. The expression data of ERLEC1 from 39 tissues across 838 individuals were obtained from the GTEx v8 (https://gtexportal.org/).[13] Mann-Whitney U test was performed to determine the significance of ERLEC1 expression differences between the two groups, and $p<0.01$ was declared significant.

scRNA-sequencing analysis of human rectum tissues

For processing scRNA data (GSE125970), the raw data of the gene expression matrix was first downloaded from NCBI Gene Expression Omnibus (GEO) and converted into a Seurat object using the R Seurat package.[14,15] Low-quality cells were eliminated if they met any of the following requirements: (1) 3000 UMIs; (2) 200 genes; and (3) >50% of UMIs derived from the mitochondrial genome. UMI counts were normalized using the NormalizeData function, and the top 2000 features with the greatest cell-to-cell variation were calculated using the FindVariableFeatures function. To correct the batch effects among samples, the "FindIntegrationAnchors" and "IntegrateData" functions were employed. Following that, the ScaleData function was used to scale and center features in the datasets, and the RunPCA function with default parameters was used to reduce dimensionality. The data were then used for nonlinear dimensional reduction with the RunUMAP function and cluster analysis with the FindNeighbors and FindClusters functions. The FindAllMarkers function was used to identify differentially expressed genes (DEG) for a given cluster. The clusters were labeled in the same way that Wang *et al*. did in their study.[15]

References

1 Zheng T, Ellinghaus D, Juzenas S, *et al*. Genome-wide analysis of 944 133

287 individuals provides insights into the etiology of haemorrhoidal disease. *Gut*
288 2021;70:1538-49.
289 2 Ferkingstad E, Sulem P, Atlason BA, *et al*. Large-scale integration of the plasma
290 proteome with genetics and disease. *Nat Genet* 2021;53:1712-21.
291 3 Aria2c Multi-souorce Download Utilily. Available: http://aria2.sourceforge.net/
292 4 Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S,
293 Jiang T, Paige E, Surendran P, *et al*. Genomic atlas of the human plasma proteome.
294 *Nature* 2018;558, 73-79.
295 5 Skrivankova VW, Richmond RC, Woolf BAR, *et al*. Strengthening the reporting of
296 observational studies in epidemiology using mendelian randomisation
297 (STROBE-MR): Explanation and elaboration. BMJ 2021;375. doi:10.1136/bmj.n2233
298 6 Hemani G, Zheng J, Elsworth B, et al. The MR-base platform supports systematic
299 causal inference across the human phenome. *Elife* 2018;7:1-29.
300 7 Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic
301 variation. *Nature* 2015;526:68-74.
302 8 Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome
303 association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559-75.
304 9 Yoshiji S, Butler-Laporte G, Lu T, *et al*. Proteome-wide Mendelian randomisation
305 implicates nephronectin as an actionable mediator of the effect of obesity on
306 COVID-19 severity. *Nat Metab* 2023;5:248-64.
307 10 Giambartolomei C, Vukcevic D, Schadt EE, *et al*. Bayesian test for colocalisation
308 between pairs of genetic association studies using summary statistics. *PLoS Genet*

2014;10:e1004383.

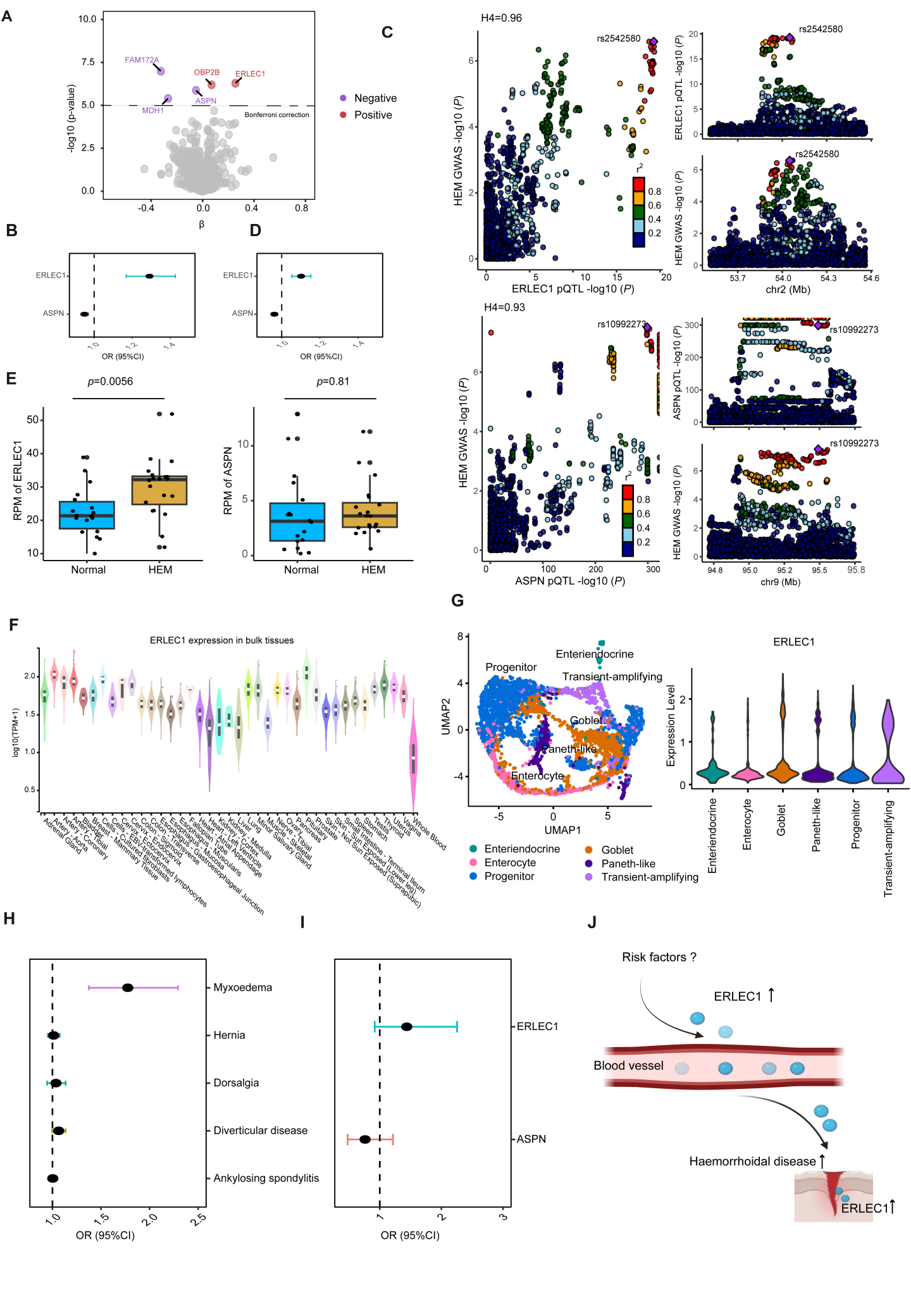
11 Foley CN, Staley JR, Breen PG, *et al*. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat Commun* 2021;12.

12 Zheng T, Ellinghaus D, Juzenas S, et al. Genome-wide analysis of 944 133 individuals provides insights into the etiology of haemorrhoidal disease. Gut 2021;70:1538-49.

13 Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. *Biopreserv Biobank* 2015;13:307-8.

14 Hao Y, Hao S, Andersen-Nissen E & Mauck WM. Integrated analysis of multimodal single-cell data. Preprint at bioRxiv https://doi.org/10.1101/2020.10.12.335331.

15 Wang Y, Song W, Wang J, *et al*. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. *J Exp Med* 2020;217(2):e20191130.


A
FAM172A
OBP2B
ERLEC1
ASPN
MDH1
Bonferroni correction
-log10 (p-value)
β
Negative
Positive
B
ERLEC1
ASPN
OR (95%CI)
C
H4=0.96
rs2542580
HEM GWAS -log10 (P)
ERLEC1 pQTL -log10 (P)
r2
chr2 (Mb)
H4=0.93
rs10992273
ASPN pQTL -log10 (P)
chr9 (Mb)
D
E
p=0.0056
p=0.81
RPM of ERLEC1
RPM of ASPN
Normal
HEM
F
ERLEC1 expression in bulk tissues
log10(TPM+1)
G
Enteriendocrine
Progenitor
Transient-amplifying
Goblet
Paneth-like
Enterocyte
UMAP1
UMAP2
ERLEC1
Expression Level
H
Myxoedema
Hernia
Dorsalgia
Diverticular disease
Ankylosing spondylitis
I
ERLEC1
ASPN
J
Risk factors ?
ERLEC1 ↑
Blood vessel
Haemorrhoidal disease ↑