Supplemental material

IBDome: An integrated molecular, histopathological, and clinical atlas of inflammatory bowel diseases

Christina Plattner, Gregor Sturm, Anja A. Kühl, Raja Atreya, Sandro Carollo, Dietmar Rieder, Raphael Gronauer, Michael Günther, Steffen Ormanns, Claudia Manzl, Stefan Wirtz, Asier Rabasco Meneghetti, Ahmed N. Hegazy, Jay V. Patankar, Zunamys I. Carrero, Felix Grabherr, Moritz Meyer, TRR241 IBDome Consortium, Timon E. Adolph, Herbert Tilg, Markus F. Neurath, Jakob Nikolas Kather, Christoph Becker, Britta Siegmund, and Zlatko Trajanoski

Supplementary Methods

Ethics approval and consent to participate

The IBDome was approved by the institutional ethics boards in both Erlangen and Berlin (project identifiers 332-17B and EA1/200/17, respectively). The IBDome is granted permission to collect and share patient samples, clinical and molecular data. All included participants are 18 years or older and have provided informed consent before inclusion into the study.

The non-IBD control subjects consisted of patients who underwent surveillance colonoscopy and did not have signs of active intestinal inflammation upon the performed colonoscopy procedure, and also did not have any history of diagnosed chronic intestinal inflammation. If the samples of the non-IBD control group were taken from surgical gut specimens, it was ensured that they were taken from patients who had to undergo surgery due to the diagnosis of colorectal neoplasia and that the selected site consisted of non-neoplastic and non-inflammatory tissue.

H&E-stained tissue sections from patients in Innsbruck were retrieved from the existing biobank which was approved by the institutional ethics board with the identifiers AN4994 323/4.4 (3806a).

Data management

We distinguish between clinical databases at the study center and a centralized research database. The former was implemented by the IT departments of the study centers in accordance with data protection laws, while the latter only contains non-identifiable information that may be shared publicly according to the ethics approval. In regular intervals, data are transferred from the clinical centers to the central research database located in Innsbruck (Biocenter, Institute of Bioinformatics at the Medical University of Innsbruck).

Study participants were assigned a randomly generated pseudonym when entering the study, which was used to label specimens and samples in the research database. The data related to biomaterials are stored in pseudonymized form in the Starlims biobank management software. Access to the systems (clinical databases and Starlims) was restricted and regulated by an authorization concept.

To ensure data security, all systems are hosted in a secured environment of the university hospital IT infra-structure of Erlangen and Berlin with an information security management

system (ISMS) based on guidelines from the German Federal Office for Information Security. The ISMS specifies procedures and rules within the hospital to define, manage, control, maintain, and continuously improve data security. The documented standard operating procedures for data security and data safety were followed and were checked on a regular basis. The data management fulfills all requirements of the EU General Data Protection Regulation and good scientific practice.

Biomaterial collection, processing and storage

The following specimen are collected from patients in the study:

- Whole blood, collected in heparinized tubes (Vacuette® Greiner Bio-One plasma tube with heparin, Thermo Fisher Scientific) for peripheral blood mononuclear cell isolation as well as K3EDTA tubes (Vacuette® Greiner Bio-One, Thermo Fisher Scientific) for DNA isolation.
- Serum, collected in (Vacuette® Greiner Bio-One Z Serum Sep Clot Activator tubes, Thermo Fisher Scientific).
- Mucosal biopsies collected during endoscopy or after surgery from surgical specimen, stored in test tubes containing RNA protect reagent (RNAprotect Tissue Reagent, Qiagen) for RNA isolation and neutral buffered, 10 % formalin solution (Sigma-Aldrich) for histopathology.
- Surgical resections, including ileocecal resection, hemicolectomy, colectomy, and normal tissue during cancer surgery, where we collected the unaffected tissue at the resection margin for IBDome.
- Stool samples, by providing patients with a stool sample tube containing RNA protect reagent (RNAprotect Tissue Reagent, Qiagen) and a questionnaire to sample stool 3-5 days after endoscopy or surgery.

In brief, samples were processed as follows. Peripheral blood mononuclear leukocytes (PBMC) are isolated from whole blood employing the SepMate™-50 (IVD) tube for density gradient centrifugation (StemCell Technologies). PBMCs are stimulated with PMA/Ionomycin and LPS or left unstimulated for 4 hours. Naïve PBMC (directly after isolation), stimulated PBMC and unstimulated PBMC (with or without brefeldin A) are fixed in Proteomic Stabilizer PROT1 (SMART TUBE Inc.) and stored at -80°C for CyTOF analysis. The supernatants of LPS-stimulated PBMC are stored at -80°C for cytokine analysis. Whole blood from EDTA tubes is stored in 1 mL aliquots at -80°C for DNA isolation. Serum is stored in 1 mL aliquots at -80°C for proteomics (Olink). After incubation of biopsies in RNA protect reagent (RNAprotect Tissue Reagent, Qiagen) overnight at 4°C, biopsies are stored individually at

-80°C until RNA isolation. Formalin-fixed biopsies or resected tissue is processed by and stored at iPATH.Berlin, the core unit of Charité-Universitätsmedizin Berlin for histopathology. Stool samples in RNA protect reagent (RNAprotect Tissue Reagent, Qiagen) are stored in pea-sized aliquots or 1 mL aliquots when liquid at -80°C until analysis.

Histopathological assessment

Formalin-fixed tissues were embedded overnight and paraffin blocks were prepared. Paraffin sections (1-2 µm) were dewaxed and hydrated in a descending alcohol series. Sections were stained with hematoxylin (Merck) and eosin (Sigma-Aldrich). Sections were dehydrated in an ascending alcohol series with xylene (Carl Roth) as intermediate and coverslipped with corbit balsam (Hecht). Histomorphology of the ileum and colon was evaluated according to modified scores based on Naini and Cortina[1] for CD and Riley[2] for UC. The main modification of both scores include the evaluation of resected tissue with scores for submucosal and transmural inflammation, fissures and increased lymphatic follicles. Minor modifications to the Nini and Cortina scoring system add villous atrophy and fibrosis. Also, for the Riley scoring scheme, the modifications include the scores for resected tissue as well as the scoring for ileal involvement (evaluation of infiltration with acute and chronic inflammatory cells, architectural distortion and epithelial integrity).

Stool score assessment

Stool samples were taken by the patients and shipped in RNAprotect reagent accompanied by a questionnaire. In order to classify various types of feces the Bristol stool chart was used[3].

Whole exome sequencing analysis

Germline mutations were called using a custom-built nextflow pipeline. Briefly: Whole exome sequencing raw reads were cleaned from residual adapter sequences and low-quality sequences using fastp v0.12.4[4]. The reads were then aligned to the reference genome (hg38) using BWA v0.7.17[5]. Duplicate reads were marked with sambamba v0.8.0[6]. Base-call quality score recalibration was performed with GATK4 v4.2.3[7]. Germline variants are called using the haplotypecaller program from GATK4 and Strelka2 v2.9.10[8]. Variants that were called from both algorithms were used as high-confidence variants and annotated using the Ensembl variant effect prediction (VEP v104.3) tool[9].

To investigate *NOD2*, all mutations were filtered to retain only coding variants associated with protein-coding transcripts. Exon regions were extracted from the Gencode v33 primary

assembly annotation GTF file using the R-package GenomicFeatures (v.1.56.0). A transcript database (TxDb) was created with the *makeTxDbFromGFF* function. Transcript names were retrieved using the *transcripts* function and filtered to match *NOD2* transcript IDs present in our dataset. The distribution of *NOD2* mutations was visualized using the trackViewer R-package (v.1.40.0). A lollipop plot was generated, highlighting the most frequent mutations in red.

GenePy score calculation

To estimate the pathogenic burden of genomic variants we adopted the GenePy[10] score. Briefly, the score was calculated from variant effect predictor (VEP) annotations by integrating functional impact, allele frequency, and zygosity for each variant. First, the raw CADD scores[11] (v1.6) for all variants were rescaled between the minimum and maximum values observed across the dataset. To account for population frequency, each variant's gnomAD allele frequency was adjusted to avoid zero or one values, and converted to a probability tuple representing the reference and alternate allele frequencies. The variant-level score was then computed as the scaled CADD score multiplied by the negative log10 of the product of these allele frequencies; for homozygous variants, this value was doubled. Only variants with predicted coding consequences were retained. Finally, variant scores were summed per gene for each individual, producing a gene-level "GenePy" score that reflects the cumulative predicted deleterious burden of coding variants in that gene.

eQTL analysis

To identify cis-expression quantitative trait loci (cis-eQTLs), genotyping was performed using GATK HaplotypeCaller[12] to generate a genomic VCF (GVCF), which was subsequently lifted from GRCh38 to GRCh19 using BCFtools/liftover[13]. Variants were imputed using a local Nextflow pipeline running the Michigan Imputation Server 2[14] in combination with the 1000 Genomes reference panel. Post-imputation quality control was conducted using bcftools[15] (v1.23) and PLINK2[16] (v2.0.0-a.6.9LM); variants were initially filtered for imputation quality ($R^2 > 0.85$) and minor allele frequency (MAF > 0.01). For the eQTL analysis, a stricter subset of variants was generated with thresholds for MAF (> 0.1), genotype missingness (< 0.1), individual missingness (< 0.05), and Hardy-Weinberg equilibrium (hwe $1x10^{-6}$). Gene expression data was processed by filtering low-expression genes (TPM > 0.1 in ≥ 30% of samples) and applying gene-level inverse normal transformation. Hidden confounders were estimated using the PEER[17] (v.2.0) package, with the number of factors (*K*) determined by sample size (*K*=15 for *N*<150, *K*=30 for *N*=150–250, *K*=60 for *N*>250), and combined with optimal genotype principal components (PCs) selected via the elbow method. Cis-eQTL

mapping was performed using MatrixEQTL[18] (v2.3) with a linear model within a 1 Mb window. To identify a single lead SNP per gene, significant associations (FDR < 0.05) were prioritized using a hierarchical tie-breaking strategy: variants were sorted first by lowest FDR, followed by highest empirical imputation quality (ER^2), highest imputation quality (R^2), and finally highest MAF. Fine-mapping was conducted using susieR[19] (v0.14.2) on expression residuals after regressing out covariates to identify 95% credible sets (**Supplementary Table 2**). Significant gene-SNP pairs (FDR < 0.05) were compared with all significant cis-eQTL results from the healthy intestinal datasets of the GTEx[20] project (GTEx eQTL analysis v8: colon transverse, colon sigmoid, small intestine terminal ileum). Colocalization with IBD GWAS summary statistics from the IEU OpenGWAS[21] database (dataset: IBD ebi-a-GCST004131) was assessed using the coloc[22] (v5.2.3) package by requiring a posterior probability (PP4) greater than 0.5 for a model with a single common causal variant and leveraging the effect sizes (β) and their variance (var-β) for shared variants. Within the colocalization model, eQTLs were defined as quantitative traits (type = "quant"), whereas GWAS summary statistics were defined as case-control traits (type = "cc") utilizing the respective sample sizes and case proportions. To assess inflammation-dependent effects, the cohort was stratified into low (≤ 33rd percentile) and high (≥ 66th percentile) inflammation groups based on the normalized histological score. To prevent overcorrection and preserve the stratification signal, this score was excluded from the final covariate matrix during the group-specific eQTL mapping. Finally, context-specific interactions were identified by comparing the resulting group-specific β coefficients using a Z-score test with $p_{interaction} < 0.05$.

Transcriptomics analysis

RNA-sequencing samples from four different batches were processed with the nf-core RNA-seq pipeline version 3.4[23]. In brief, sequencing reads were aligned to the hg38/GRCh38 reference genome with Gencode v33 annotations using STAR v2.7.7a[24]. Read counts and transcripts per million (TPM) were quantified using Salmon[25].

Differential expression analysis was performed in R v.4.4.1 with DESeq2 (v.1.44.0) using raw counts and the covariate formula ~ *group* + *batch* + *sex* + *scaled age*. For comparisons between IBD inflamed and non-IBD samples *tissue_coarse* was added as an additional covariate to account for the different tissues involved. Genes were considered differentially expressed if they met an adjusted p-value threshold of < 0.05 and a |log2FoldChange| threshold of >1. For visualization of the results we used the EnhancedVolcano (v.1.22.0), ggplot2 (v.3.5.1), ComplexHeatmap (v.2.20.0), and ggvenn (v.0.1.10) R-packages.

Cytokine signaling activities for bulk gene expression data were inferred using CytoSig[26] in Python v.3.8.20, leveraging the cytosig.v0.1 implementation available on GitHub (https://github.com/data2intelligence/CytoSig). TPM values were log-transformed as $\log_2(\text{TPM} + 1)$ prior to analysis and used as input. CytoSig calculates the z-score by dividing the regression coefficient by the standard error. The p-values are obtained using a permutation test when the random count is > 0 or using a Student's t-test if the random count is 0.

Biopsy and circulating molecular inflammation signatures were obtained from Argmann et al.[27]. To calculate the biopsy molecular inflammation scores (bMIS) for our samples, we applied gene-set variation analysis (GSVA)[28] using the GSVA R-package (v.1.52.3).

Cell fractions were inferred from TPM-normalized expression values with quanTIseq[29,30] as implemented in the quantiseqr R-package (v.1.12.0). We applied the built-in "TIL10" signature matrix and performed deconvolution with constrained least squares regression. For visualization cell types with no detected fractions (dendritic cells and monocytes) as well as the "Other" category were excluded. Differences between conditions were assessed using the Wilcoxon rank-sum test with significance thresholds set at: $p < 0.05$ (*), $p < 0.01$ (**), $p<0.001$ (***), and $p<0.0001$ (****).

Serum protein analysis

Data tables containing normalized protein expression (NPX) values, Olink Proteomics' arbitrary unit on $\log_2$ scale, were loaded into R v.4.4.1 and further processed with the OlinkAnalyze (v.4.0.1) R-package. Differential protein analysis was conducted using the *olink_ttest* function. Only proteins detected in at least 90% of the measured samples were included in the analysis. Statistical differences were assessed using the Welch two-sample t-test with Benjamini-Hochberg correction applied to adjust for multiple testing. Proteins were considered differentially abundant if they met a FDR threshold of < 0.05. Results were visualized using the EnhancedVolcano (v.1.22.0) R-package. Intersections were retrieved and plotted with the ggVennDiagram (v.1.5.2) or the UpSetR (v.1.4.0) R-package.

We developed an IBD Inflammatory Protein Severity Signature (IBD-IPSS) using a method consistent with the approach outlined by Argmann et al.[27]. In brief, differential protein abundance between inflamed and non-inflamed IBD samples was analyzed using OlinkAnalyze as described above, identifying significantly upregulated proteins for inclusion in the IBD-IPSS. Similarly, entity-specific signatures were generated: the UC-IPSS and CD-IPSS, derived by analyzing protein abundance separately in ulcerative colitis and

Crohn's disease samples. Correlation analysis with the various inflammatory scores available within IBDome including endoscopic scores (SES-CD and UCEIS), clinical scores (HBI and PMS), histopathology scores (modified Riley and modified Naini Cortina score) and the computed bMIS scores (bMIS-CD and bMIS-UC) was conducted using Pearson correlation with pairwise complete observations.

Functional analysis and clustering of the IBD-IPSS proteins was performed using the STRING database[31]. Evidence for protein interactions was considered only from curated databases and experimentally validated interactions. Clustering was performed using MCL (Markov Cluster Algorithm)[32] with an inflation parameter set to 3. Clusters were annotated using the default settings of the STRING database web application. This annotation process prioritized general terms or pathways that summarize multiple specific terms and pathways, derived from various databases integrated within STRING.

Multi-omics factor analysis

For the multi-omics factor analysis, we used MOFA[33] as implemented in the MOFA2 R-package (v.1.14.0). As input for the model, we included four data modalities: genomic, transcriptomic, proteomic, and histological features. The genomic layer consisted of the top 1000 most variable mutations assessed by median absolute deviation of the GenePy-scores. For transcriptomic data, we used gene expression values in log(TPM+1), retaining the top 5000 most variable genes across samples. The proteomic layer included normalized Olink protein expression values (NPX), filtered to retain only proteins detected in at least 10 samples. Finally, the histological features were derived from H&E-stained tissue sections as described in "*Generation of classification-derived feature matrix for MOFA integration*". Samples from the different modalities were merged by subject ID, sampling date, and tissue (total N=1485).

We trained the MOFA2 model using default parameters, except that the *scale_groups* data option was set to *TRUE* and the convergence mode was set to "slow". The model identified 20 latent factors that captured major sources of variation. To visualize the latent structure of the learned factors, we applied Uniform Manifold Approximation and Projection (UMAP) for non-linear dimensionality reduction with the *run_umap* and *plot_dimred* functions as implemented in the MOFA2 R-package.

The IBDome research database

A relational database was designed and implemented in the Python package sqlalchemy using SQLite as database engine. Data integrity is ensured through check constraints and foreign key validation. SQLite was chosen over other database systems, because it makes the database easy to share as a single file, does not require a server, and offers good performance for a use-case without concurrent writes. Inconsistencies in clinical data were resolved manually, and implausible entries were removed. Both clinical and molecular data were processed and imported into the database in a set of Jupyter notebooks and a custom helper library written in Python. All data loading steps are integrated into a Nextflow pipeline[34], which allows rebuilding the database from scratch in a single command.

Web application

The IBDome web application is implemented in R Shiny and directly interacts with the IBDome SQLite database. Dependencies are packaged in a Docker container and a docker-compose file is provided which allows executing the app locally. Plots were generated in R using the ggplot2, ggpubr, plotly, and ggbeeswarm packages. For visualization of gene expression data, transcripts per million (TPM) values were $\log_{10}$(TPM+1) transformed. P-values were computed using a two-tailed Wilcoxon test on the transformed values.

Foundation Model Embedding Extraction

H&E WSI were tessellated into patches with dimensions of 224×224 pixels, representing a 256 µm edge length. To ensure consistent color distribution across cohorts, patches from each cohort underwent color normalization using the Macenko spectral matching technique[35], which maps images to a standardized color space. For performance comparison purposes and to ensure the robustness of our findings, we extracted patch-level feature embeddings using four distinct Foundation models—CHIEF[36], UNI2[37], Virchow2[38,39] and H-optimus-1[40]. These models produced an embedding matrix of dimensions n × 768, n × 1536, n × 2560, and n × 1536, respectively, for each patient's pre-processed patches, where n is the number of (224 ×224 pixels) pre-processed image patches obtained per whole slide image. All preprocessing steps followed the STAMP protocol[41], which harmonizes patch extraction and feature computation across cohorts. The resulting slide-level embedding matrices were subsequently used for both the inflammation score prediction (regression task) and IBD subtype classification (classification task).

Deep Learning Inflammation score prediction

The extracted slide-level embeddings were processed in an attention-based multiple instance learning (attMIL) framework[42] designed for weakly supervised regression. For each foundation model, a separate attMIL model was trained using 5-fold cross-validation on the Berlin cohort to predict the normalized modified Naini Cortina score and the normalized modified Riley score. The cross-validation employed score-based stratification to maintain consistent data distributions across all folds, resulting in five models trained and tested on distinct and balanced splits. To externally validate the model's prognostic performance, all five attMIL models from the cross-validation folds were independently deployed to the Erlangen and Innsbruck cohorts to mitigate fold-specific variability. For the Innsbruck cohort, histologic disease activity scores were assigned independently by averaging the scores of two gastrointestinal histology experts, including one board-certified gastrointestinal pathologist, both blinded to clinical and molecular data, and aggregated by arithmetic averaging to generate a consensus reference score for model evaluation. Slide-level predictions were generated by each model and then aggregated through arithmetic averaging to produce the final prognostic scores. These steps were performed using the open-source Deep Learning pipeline “marugoto”[43], with the default hyperparameters (learning rate = 0.0001, weight decay = 0.01, batch size = 1) using weighted L1 loss.

Explainability of the Deep Learning model

To interpret the decision-making process of the regression models, we leveraged the attention mechanism of the attMIL architecture. High-resolution attention heatmaps were created by loading the attMIL model architectures for regression into a fully convolutional equivalent[44] with their respective weights from the training procedure. By running inference on each patient’s WSI, we extracted the attention layer associated with the score prediction and overlaid it on the WSI, highlighting the regions of focus for the model’s predictions of the scores. For visualization, we selected the Berlin cohort to observe the model performance in predicting disease activity scores. For a more detailed evaluation, we selected the top 10 attention heatmaps for each scoring system based on prediction accuracy. These heatmaps were then reviewed by an expert pathologist, who assessed the highlighted regions for correspondence with areas of known clinical relevance.

IBD subtype classification

In order to evaluate the ability of foundation model embeddings to support subtype discrimination within IBD, we trained a weakly supervised transformer classifier aimed at

distinguishing CD from UC. The classification pipeline was implemented using the STAMP framework[41], utilizing the same slide-level embeddings previously extracted from each foundation model (CHIEF[36], UNI2[37], Virchow2[38,39], and H-optimus-1[40]). For each model, a 5-fold cross-validation was conducted on the Berlin cohort using stratified splitting by disease subtype. The cohort had a near-equal distribution of CD and UC cases, minimizing class imbalance during training. Following cross-validation, a final classifier model was retrained on the entire Berlin cohort and then deployed on four distinct subsets of the Erlangen cohort to assess generalization: all tissue samples (full Erlangen set), all inflamed tissues, colon-only samples, and inflamed colon samples. The classifier was trained using cross-entropy loss, while predictions were obtained by applying a softmax activation to the model outputs, with class assignment based on the highest predicted probability. Model performance metrics were computed for each subset, and 1,000 bootstrap resampling were performed to derive 95% confidence intervals for all performance estimates.

Generation of classification-derived feature matrix for MOFA integration

To derive histopathology features for multi-modal factor analysis, we leveraged the Virchow2 model[39], which showed the strongest classification performance in our benchmarking. A 5-fold cross-validation was performed on the combined Berlin and Erlangen cohorts, and for each image, we used the specific fold model responsible for generating its classification prediction. From the slide-level h5 embedding feature files, we selected the top 10 tiles per slide based on the softmax score for the prediction of the class. A weighted average of the 1280-dimensional embeddings of the top 10 tiles, using the normalized softmax scores as weights, was then computed to obtain a single feature vector per slide. This resulted in a patient-level feature matrix of size 1250 × 1280, which was used as the histology modality input for MOFA[33] integration.

Supplementary Results

Genetic regulation of tissue gene expression

We performed expression quantitative trail locus (eQTL) analysis to examine how genetic variation associates with gene expression in the tissue. cis-eQTL mapping identified 3,600 distinct eGenes. Comparison of significant eSNP-eGene associations with GTEx healthy intestinal eQTL data yielded 3,138 overlapping pairs, including 224 lead cis-eQTLs. Integration of these cis-eQTLs with IBD GWAS summary statistics (EBI study GCST004131) identified 60 shared causal variants with high colocalization probability. Furthermore, our analysis highlighted 396 inflammation-dependent lead eQTLs affecting 388 distinct eGenes (**Supplementary Figure 1B-C, Supplementary Table 1-5)**.

Multi-omic factor analysis reveals tissue- and inflammation-driven variation

To integrate the diverse datasets in our cohort, we performed multi-omic factor analysis using MOFA2[33], incorporating genomics, transcriptomics, serum proteomics, and image-derived features from H&E-stained tissue sections. After merging samples by patient, date, and tissue (**Supplementary Figure 2A**), MOFA identified 20 latent factors explaining variation across modalities (**Supplementary Figure 2B**). The majority of variance was captured by transcriptomics and histology-derived features, followed by proteomics.

The leading factors predominantly captured tissue type and inflammation, consistent with known drivers of molecular heterogeneity in IBD, while disease subtype (CD vs. UC) accounted for only minor differences (**Supplementary Figure 2C**). UMAP embedding showed continuous gradients without clear CD/UC separation (**Supplementary Figure 2D**), highlighting the molecular similarity of inflamed tissues across IBD subtypes and the challenge of distinguishing disease type from bulk tissue data alone.

Supplementary References

1. Naini BV, Cortina G. A histopathologic scoring system as a tool for standardized reporting of chronic (ileo)colitis and independent risk assessment for inflammatory bowel disease. Human Pathology 2012;43:2187–2196.

2. Riley SA, Mani V, Goodman MJ, et al. Microscopic activity in ulcerative colitis: what does it mean? Gut 1991;32:174–178.

3. Lewis SJ, Heaton KW. Stool Form Scale as a Useful Guide to Intestinal Transit Time. Scandinavian Journal of Gastroenterology 1997;32:920–924.

4. Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–i890.

5. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–1760.

6. Tarasov A, Vilella AJ, Cuppen E, et al. Sambamba: fast processing of NGS alignment formats. Bioinformatics 2015;31:2032–2034.

7. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;43:11.10.1-11.10.33.

8. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods 2018;15:591–594.

9. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:122.

10. Mossotto E, Ashton JJ, O'Gorman L, et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. BMC Bioinformatics 2019;20:254.

11. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310–315.

12. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. 2017. Available at: http://biorxiv.org/lookup/doi/10.1101/201178 [Accessed February 18, 2026].

13. Genovese G, Rockweiler NB, Gorman BR, et al. BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies Kendziorski C, ed. Bioinformatics 2024;40:btae038.

14. **Das S, Forer L, Schönherr S**, et al. Next-generation genotype imputation service and methods. Nat Genet 2016;48:1284–1287.

15. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10:giab008.

16. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:s13742-015-0047–8.

17. Stegle O, Parts L, Piipari M, et al. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc 2012;7:500–507.

18. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 2012;28:1353–1358.

19. Cui R, Elzur RA, Kanai M, et al. Improving fine-mapping by modeling infinitesimal effects. Nat Genet 2024;56:162–169.

20. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 2020;369:1318–1330.

21. Lyon M, Andrews SJ, Elsworth B, et al. The variant call format provides efficient and robust storage of GWAS summary statistics. 2020. Available at: http://biorxiv.org/lookup/doi/10.1101/2020.05.29.115824 [Accessed February 23, 2026].

22. Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants Cordell HJ, ed. PLoS Genet 2021;17:e1009440.

23. **Ewels PA, Peltzer A**, Fillinger S, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol 2020;38:276–278.

24. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.

25. Patro R, Duggal G, Love MI, et al. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 2017;14:417–419.

26. **Jiang P, Zhang Y, Ru B**, et al. Systematic investigation of cytokine signaling activity at the tissue and single-cell levels. Nat Methods 2021;18:1181–1191.

27. Argmann C, Hou R, Ungaro RC, et al. Biopsy and blood-based molecular biomarker of inflammation in IBD. Gut 2023;72:1271–1287.

28. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 2013;14:7.

29. Finotello F, Mayer C, Plattner C, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med 2019;11:34.

30. Plattner C, Finotello F, Rieder D. Chapter Ten - Deconvoluting tumor-infiltrating immune cells from RNA-seq data using quanTIseq. In: Galluzzi L, Rudqvist N-P, eds. *Methods in Enzymology*.Vol 636. Tumor Immunology and Immunotherapy – Integrated Methods Part B. Academic Press; 2020:261–285. Available at: https://www.sciencedirect.com/science/article/pii/S007668791930237X [Accessed August 11, 2025].

31. Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res 2023;51:D638–D646.

32. Van Dongen S. Graph Clustering Via a Discrete Uncoupling Process. SIAM J Matrix Anal Appl 2008;30:121–141.

33. **Argelaguet R, Velten B**, Arnol D, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology 2018;14:e8124.

34. Di Tommaso P, Chatzou M, Floden EW, et al. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017;35:316–319.

35. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*.; 2009:1107–1110. Available at: https://ieeexplore.ieee.org/document/5193250 [Accessed October 30, 2024].

36. **Wang X, Zhao J**, Marostica E, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. Nature 2024;634:970–978.

37. **Chen RJ, Ding T, Lu MY,** et al. Towards a general-purpose foundation model for computational pathology. Nat Med 2024;30:850–862.

38. **Vorontsov E, Bozkurt A, Casson A**, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat Med 2024;30:2924–2935.

39. Zimmermann E, Vorontsov E, Viret J, et al. Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology. arXiv e-prints 2024:arXiv:2408.00738.

40. Saillard C, Jenatton R, Llinares-Lopez F, et al. H-optimus-1. 2024. Available at: https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

41. El Nahhas OSM, Treeck M van, Wölflein G, et al. From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. Nat Protoc 2024:1–24.

42. Leiby JS, Hao J, Kang GH, et al. Attention-based multiple instance learning with self-supervision to predict microsatellite instability in colorectal cancer from histology whole-slide images. In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*.; 2022:3068–3071. Available at: https://ieeexplore.ieee.org/document/9871553 [Accessed October 30, 2024].

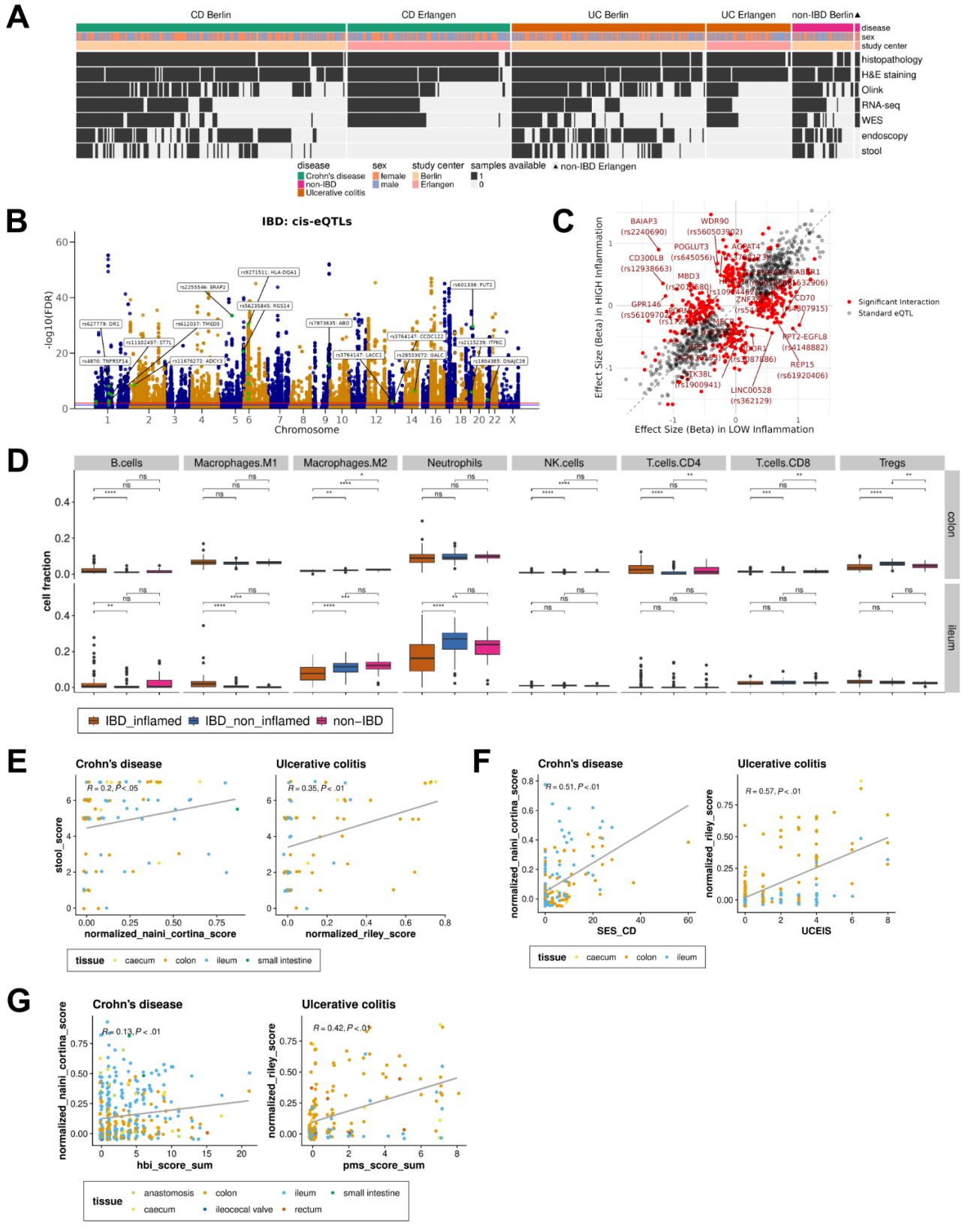
43. Ilse M, Tomczak J, Welling M. Attention-based Deep Multiple Instance Learning. In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR; 2018:2127–2136. Available at: https://proceedings.mlr.press/v80/ilse18a.html [Accessed October 30, 2024].

44. Pathak D, Shelhamer E, Long J, et al. Fully Convolutional Multi-Class Multiple Instance Learning. 2015. Available at: http://arxiv.org/abs/1412.7144 [Accessed October 30, 2024].

Author names in bold designate shared co-first authorship.

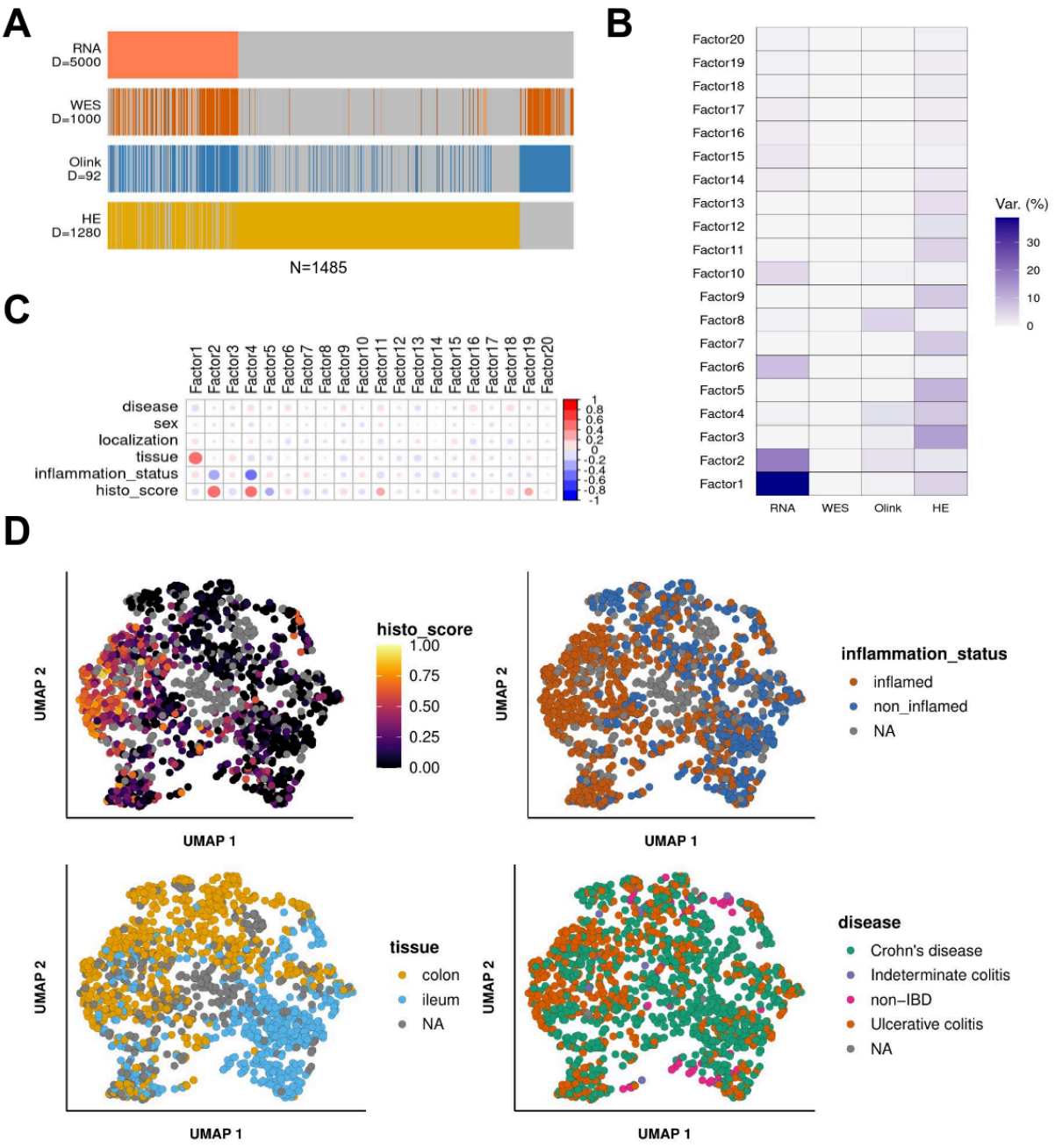
Supplementary Figures

Supplementary Figure 1



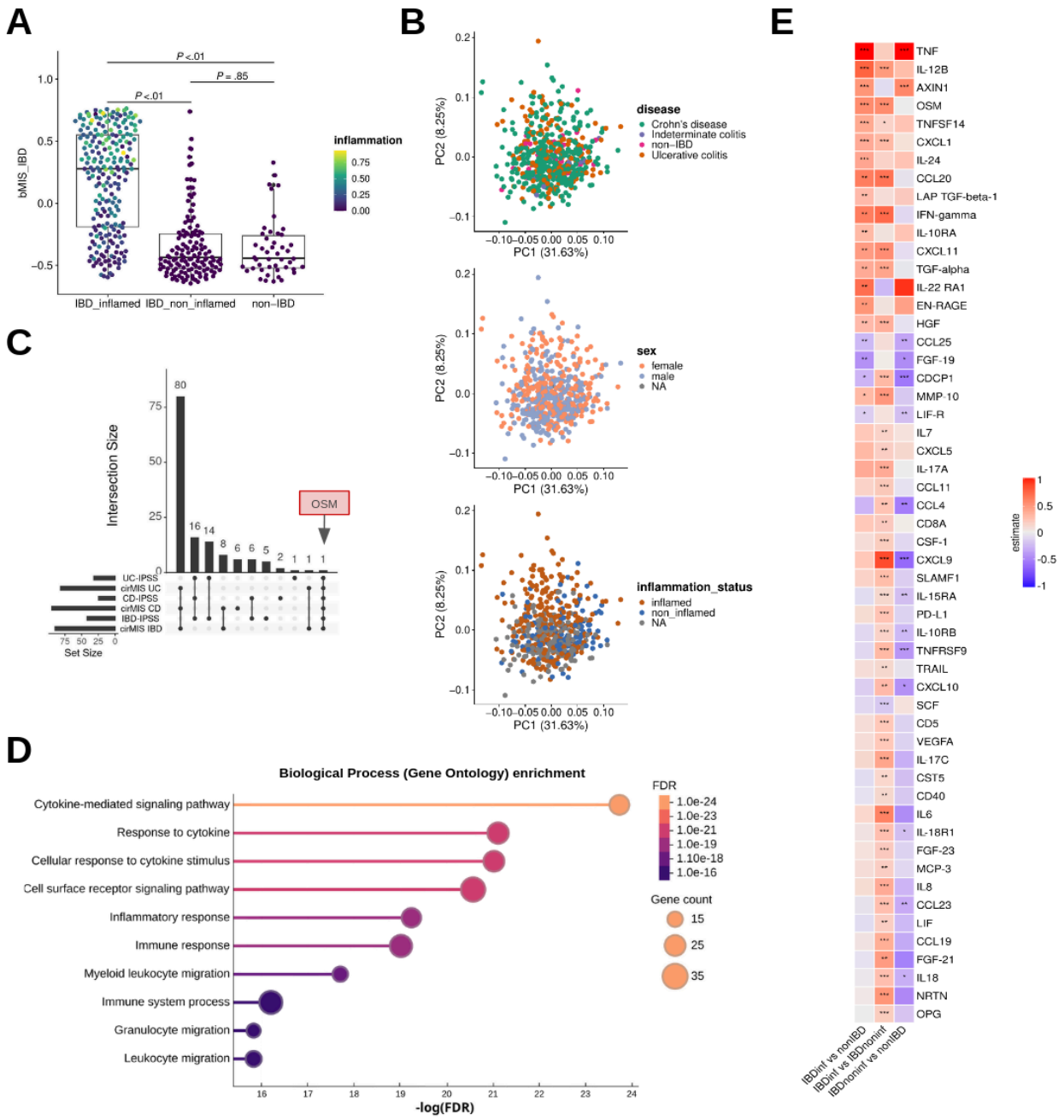
Supplementary Figure 1. IBDome overview. (A) Cohort distribution illustrated in a heatmap showing the available samples per patient. The heatmap is split by disease and study center and ordered according to the amount of different sample types available per patient. **(B)** Manhattan plot of cis-eQTL associations in the IBDome cohort. The x-axis displays the genomic position of tested variants across chromosomes, and the y-axis represents the significance of the eQTL associations expressed as -log10(FDR). The solid horizontal lines denote False Discovery Rate (FDR) significance thresholds at 0.05 (blue line) and 0.01 (red line). Highlighted data points (green) indicate the cis-eQTLs identified as shared causal variants with IBD susceptibility loci, determined through colocalization analysis with IBD GWAS data (ebi-a-GCST004131). The labels specify the top 15 lead eSNP identifiers and their corresponding eGene. **(C)** Comparison of genetic effect sizes β for inflammation-dependent cis-eQTLs. Low inflammation group (x-axis) versus the high inflammation group (y-axis). Red data points indicate cis-eQTLs with a significant inflammation-dependent interaction ($p_{interaction} < 0.05$). Grey points represent standard eQTLs that do not demonstrate a significant context-dependent interaction. **(D)** Immune cell fractions estimated with quanTIseq with p-values assessed by Wilcoxon test with * $P < .05$, ** $P < .01$, *** $P < .001$, **** $P < .0001$. **(E)** Correlation between histopathology (normalized Riley score and normalized Naini-Cortina score) and the Bristol stool score. **(F)** Correlation between endoscopic (UCEIS = Ulcerative Colitis Endoscopic Index of Severity, SES-CD = Simple Endoscopic Score for Crohn's Disease) and histopathology scores for UC and CD, respectively. **(G)** Correlation between histopathology (normalized Riley score and normalized Naini-Cortina score) and clinical activity scores (PMS= Partial Mayo Score, HBI=Harvey-Bradshaw Index) for UC and CD, respectively.

Supplementary Figure 2



Supplementary Figure 2. Multi-omics factor analysis. (A) Overview of the cohort merged by subject ID, date, and tissue across the four input modalities: RNA-seq (log(TPM+1), top 5000 most variable genes), whole-exome sequencing (WES, GenePy-score, top 1000 most variable mutations), serum proteomics (Olink), and image-derived features from H&E stained tissue sections (HE). **(B)** Variance explained (%) by each MOFA factor per modality. **(C)** Pearson correlation of MOFA factors with clinical covariates. **(D)** UMAP embedding of MOFA factors colored by the normalized histopathology score, inflammation status, tissue type, and disease subtype.

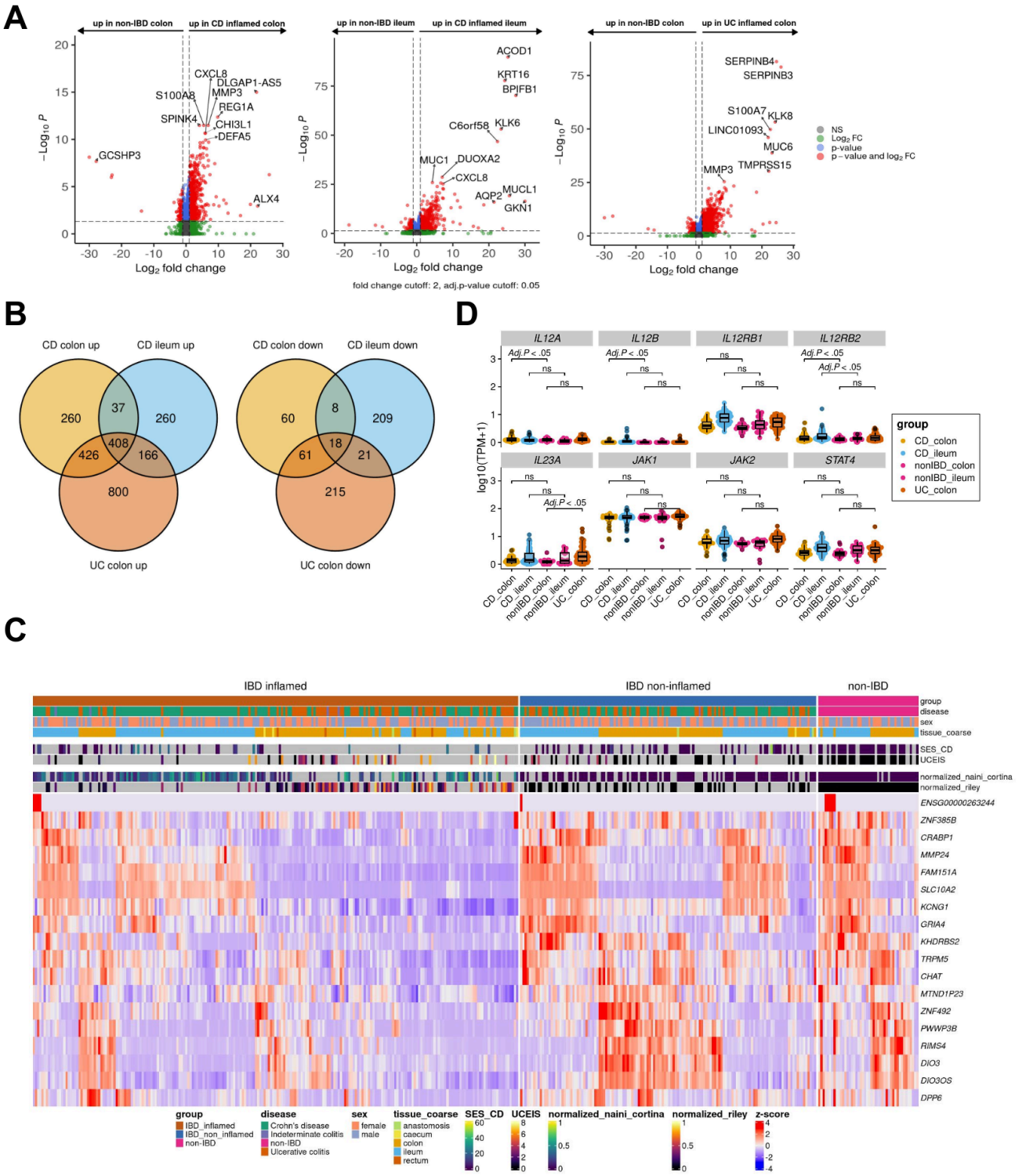
Supplementary Figure 3



Supplementary Figure 3. Protein abundance in the serum. (A) Biopsy-based molecular inflammation scores (bMIS-IBD) across different groups, calculated using Gene Set Variation Analysis (GSVA); p-values were assessed using the Wilcoxon test. The color depicts the normalized inflammation score assessed by histopathology (modified Riley score for UC and modified Naini Cortina score for CD). **(B)** Principal Component Analysis (PCA) of serum protein abundances, colored by disease type, sex, and inflammation status **(C)** UpSet plot showing the intersection of different blood-based scores (cirMIS = circulating molecular inflammation score from RNA-seq from the blood, IPSS = inflammatory protein severity

signature from serum proteins) highlighting OSM as the only protein shared across all signatures. **(D)** Significantly enriched Gene Ontology - Biological Processes (GO-BP) terms (FDR < 0.05) from over representation analysis (ORA) in the IBD-IPSS. **(E)** Heatmap of the differentially abundant proteins with significance levels determined by t-test: *** = adj.p<0.01, ** = adj.p < 0.05, * = adj.p < 0.1; “estimate” represents the numeric difference in mean NPX (normalized protein expression; Olink’s arbitrary unit) between groups.

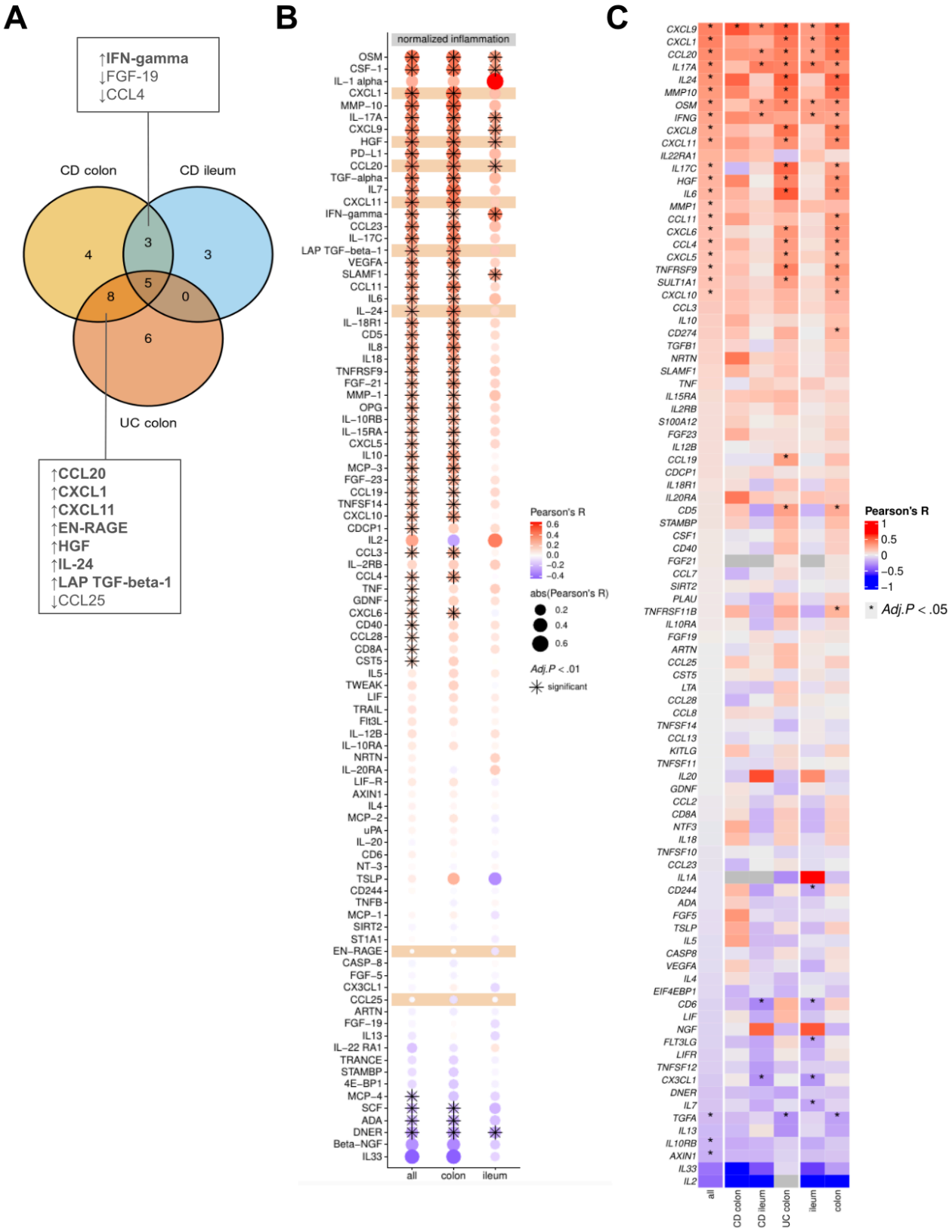
Supplementary Figure 4



Supplementary Figure 4. Transcriptional characterization. (A) Volcano plots displaying the DE analysis results of inflamed CD vs. non-IBD colon samples, inflamed CD vs. non-IBD ileum samples and inflamed UC colon vs. non-IBD colon samples with thresholds: |log2FC|>1 and adjusted *P*<.05. **(B)** Overlap of significantly upregulated and downregulated

genes in the different comparisons. **(C)** Heatmap displaying the gene expression of the commonly downregulated genes (n=18) across the different groups (inflamed IBD, non-inflamed IBD and non-IBD) clustered by euclidean distance and complete linkage. **(D)** Gene expression of genes involved in the IL12 signaling pathway with adjusted p-values retrieved from the differential expression analysis with DESeq2.

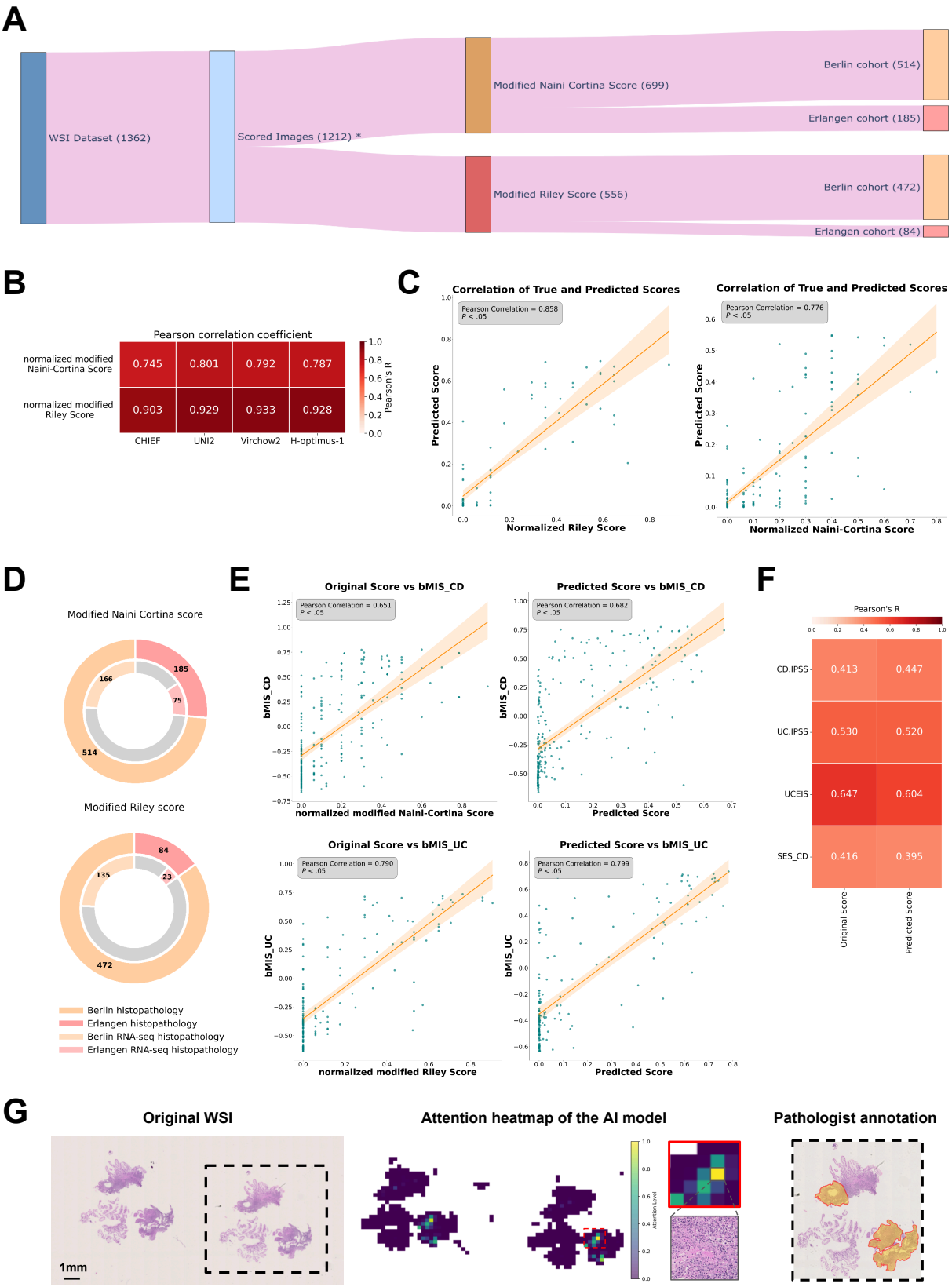
Supplementary Figure 5



Supplementary Figure 5. Correlation of serum proteins. (A) Venn diagram depicting the intersections of differential abundant proteins in the different comparisons. **(B)** Dot plot of

Pearson correlation coefficients (R) between serum protein abundance and histopathology scores (normalized inflammation = normalized modified Riley or normalized modified Naini Cortina score) across all samples and the tissue groups; adjusted p-value cutoff: 0.01; **(C)** Heatmap of Pearson correlation coefficients between serum protein abundance and tissue gene expression across all samples (n=335) and splitted by group and tissue; * adjusted p-value <0.05.

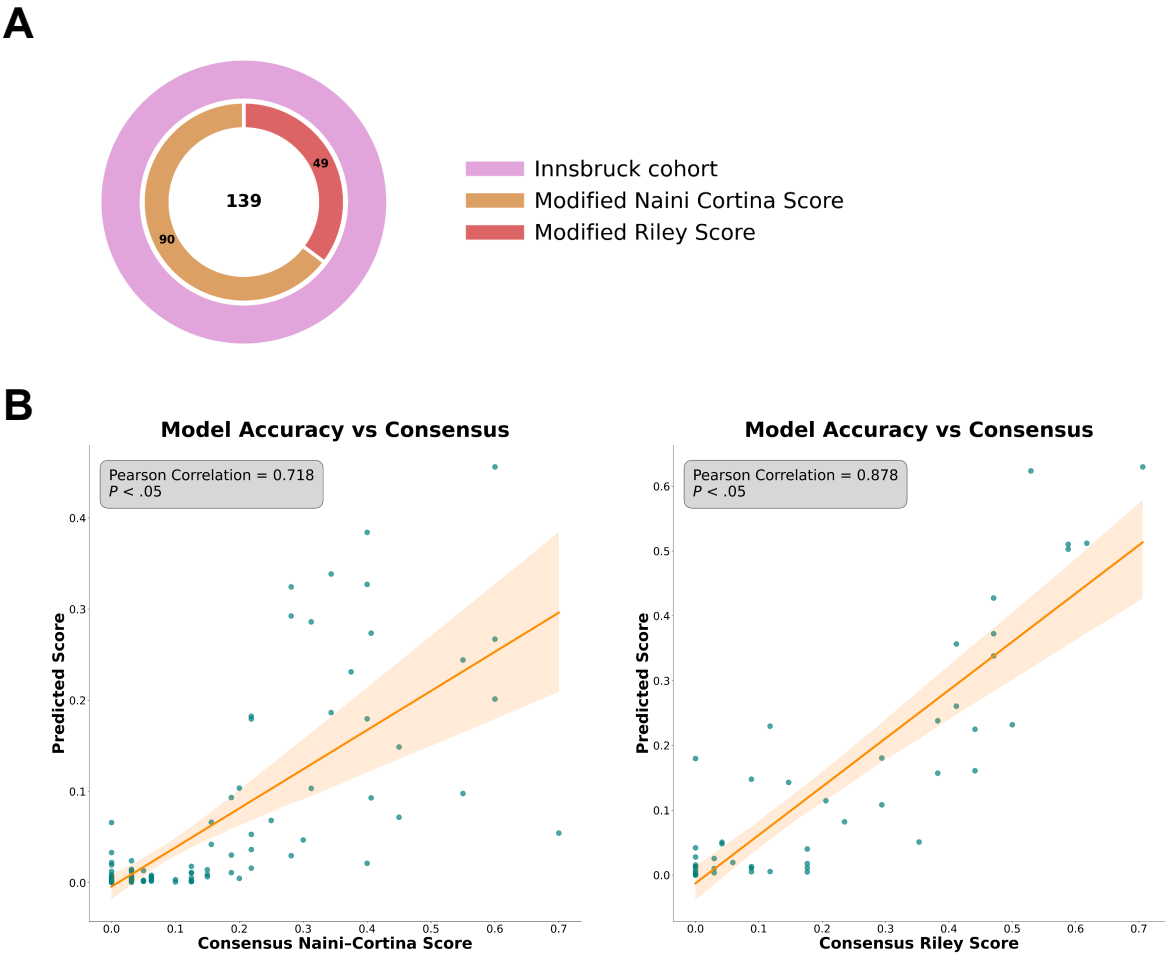
Supplementary Figure 6



Supplementary Figure 6. Prediction of histopathologic scores from H&E images. (A)

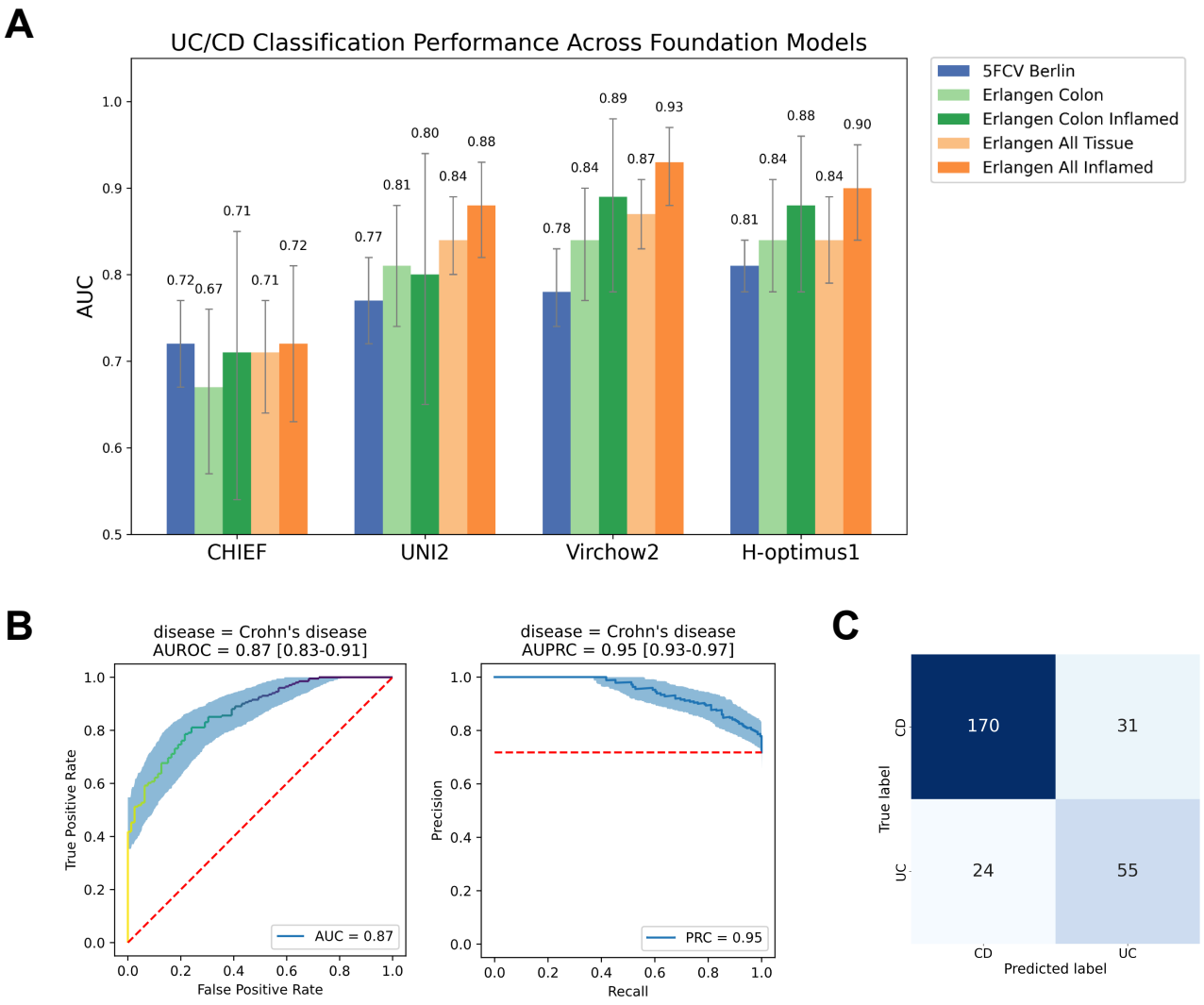
Distribution of all H&E images (n = 1362). After preprocessing, the number of images is reduced to 1212*, which are then categorized with two scoring systems: the modified Naini Cortina score (n = 699) and the modified Riley score (n = 556). Each scoring category is further divided in the two center subsets: Berlin (n = 514 for modified Naini Cortina, n = 472 for modified Riley) and Erlangen (n = 185 for modified Naini Cortina, n = 84 for modified Riley). *Some indeterminate colitis and non-IBD samples were scored with both scoring systems. **(B)** Performance of different Foundation Models (CHIEF, UNI2, Virchow2 and H-optimus-1) on the 5 folds cross-validation regression task on the Berlin cohort, shown by Pearson correlation (R) coefficients. **(C)** Correlation plots between the original histologic disease activity scores (x-axis) and AI-predicted scores (y-axis) for both modified Naini Cortina and modified Riley scoring systems. Predictions are based on the ensemble of 5 cross-validation models trained on the Berlin cohort and evaluated on the Erlangen cohort. **(D)** Cohort distribution: the outer circles represent the Berlin and Erlangen histopathology cohorts, and the inner circles indicate the proportion of available RNA-seq data within each cohort. **(E)** Correlation plots between the bMIS score for UC and CD with the original score (normalized Naini Cortina and Riley score) on the left and the model's predicted scores on the right. **(F)** Comparison of Pearson correlation coefficients (R) between original and predicted histologic disease activity scores against CD-IPSS, UC-IPSS, UCEIS, and SES-CD. **(G)** On the left side, a representative attention heatmap of a biopsy slide image from a CD patient with high histologic disease activity. The heatmap shows the model's attention levels. Higher scores (yellow) mark regions that strongly influence the model's prediction, while lower scores (green) indicate less critical regions. In the middle, a zoomed-in view of the highest-attention region showing the top attention tile and on the right side the pathologist annotation highlighting the important areas to consider for disease activity scoring.

Supplementary Figure 7



Supplementary Figure 7. Histopathologic score prediction from H&E images in the validation cohort. (A) Cohort distribution: The outer circle shows the Innsbruck cohort (n=139), and the inner circle indicates the proportion of modified Naini-Cortina score (n = 90) and modified Riley score (n=49) samples. **(B)** Correlation plots between the consensus histologic disease activity scores (x-axis) and AI-predicted scores (y-axis) for both modified Naini Cortina and modified Riley scoring systems. Predictions are based on the ensemble of 5 cross-validation models trained on the Berlin cohort and evaluated on the Innsbruck cohort.

Supplementary Figure 8



Supplementary Figure 8. Prediction of disease type from pathology images. (A) Classification performance of four Foundation Models (CHIEF, UNI2, Virchow2 and H-optimus-1) across five evaluation settings: 5-fold cross-validation on the Berlin cohort, and training on the same cohort with deployment on the Erlangen cohort using either colon samples, colon-inflamed samples, all tissue types, or all inflamed tissues. Bars represent AUROC values, with error bars denoting 95% confidence intervals. **(B)** AUROC and AUPRC curves for the best-performing model (Virchow2) trained on the Berlin cohort and evaluated on the Erlangen all-tissue subset. **(C)** Confusion matrix summarizing the UC/CD classification results from the model deployment in (B).