Supplementary Methods:

Title: **A biopsy and blood based molecular biomarker of inflammation in inflammatory bowel disease**

C. Argmann[1*], R. Hou[2*], R. Ungaro[3], H. Irizar[1], R. Huang[2], R. Kosoy[1], S. Venkat[4], W.M. Song[1], AF. Di Narzo [1], B. Losic[1], K. Hao[1], L. Peters[1], P.H. Comella[1], G. Wei[1], A. Atreja[3], M. Mahajan[1], A. Iuga[5],P. T. Desai[4], P. Branigan[4], A. Stojmirovic[4], J. Perrigoue[4], C. Brodmerkel[4], M. Curran[4], J. Friedman[4], A. Hart[4], E. Lamousé-Smith[4], J. Wehkamp[4], S. Mehandru[3], E.E. Schadt[1,6], B. Sands[3], MC. Dubinsky*[3], J.-F. Colombel*[3], A. Kasarskis*[1,2,6], M. Suárez-Fariñas*[1,2]

Analysis of Mount Sinai Crohn's and Colitis Registry (MSCCR) Cohort:

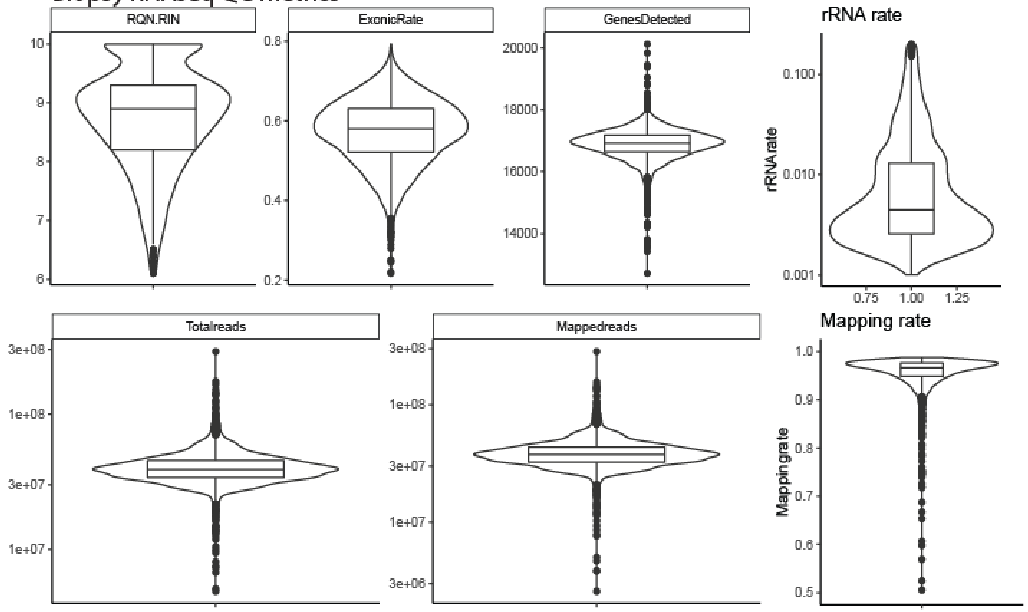
RNA extraction, alignment and QC:

RNA-seq quality metrics: Biopsy and Blood RNA was extracted and processed in randomly allocated batches as previously described [1, 2]. Briefly, biopsy samples were randomized for extraction and sequencing within each batch for biopsy sampling region (small and large intestine), disease type (UC, CD and control) and inflammation status (inflamed vs uninflamed), with considerations for age, gender and ethnicity. Similarly, blood samples were randomized for extraction and sequencing within each batch for disease type with considerations for age, gender and ethnicity. Genetic principal components (PC's) were computed using eigenstrat, v6.0.1[3] with available genotype information described previously[4]. Single nucleotide polymorphisms (SNPs) in the HLA region were removed first, then LD-redundant SNPs were pruned using plink [5] with the option '--indep-pairwise 50 5 0.8'. Demographic information associated with this cohort is summarized in **Supplementary Table 1**. RNA was isolated from frozen tissue (in RNAlater) using Qiagen QIAsymphony RNA Kit (cat.# 931636) on the QIAsymphony. RNA from whole blood collected in PAXgene tubes was isolated using QIAsymphony Blood PAXgene RNA kit (cat.# 762635). In general, one microgram of total RNA was used for the preparation of the sequencing libraries using the RNA Tru Seq Kit (Illumina (Cat # RS-122-2001-48). Ribosomal RNA from biopsy tissue was depleted from total RNA using the Ribozero kit (Illumina Cat # MRZG12324), and globin RNA along with ribosomal RNA was depleted from total blood RNA using Globin zero gold rRNA removal kit (Illumina cat.# GZG1224) to enrich poly-adenylated coding RNA as well as non-coding RNA. The ribozero and globin zero RNA-Seq libraries were sequenced on the Illumina HiSeq 2500 platform using 100 bp single end protocol following manufacturer's procedure. The RIN score had a mean of 8.7 (range: 6.1 to 10) and 8.7 (range: 6.5 to 10) for biopsy RNA-seq and blood RNA-seq respectively. The rRNA rate had a mean of 0.0145 (range: 0.001 to 0.2) and 0.005 (range: 0.0009 to 0.05) for biopsy and blood RNA-seq respectively.

Genomic alignment to GRCh37 of single-end RNA-seq reads was performed using 2-pass STAR[6, 7]. Default parameters for STAR were used, as were those for the quantification of aligned reads to GRCh37.75 gene features via featureCounts[7]. Multimapping reads were flagged and discarded. Raw count data was pre-filtered to keep genes with CPM>0.5 in at least a third of the samples. After filtering, count data was normalized via the weighted trimmed mean of M-values[8]. The quality metrics summary statistics and plots are shown in supplementary methods Figure 1.

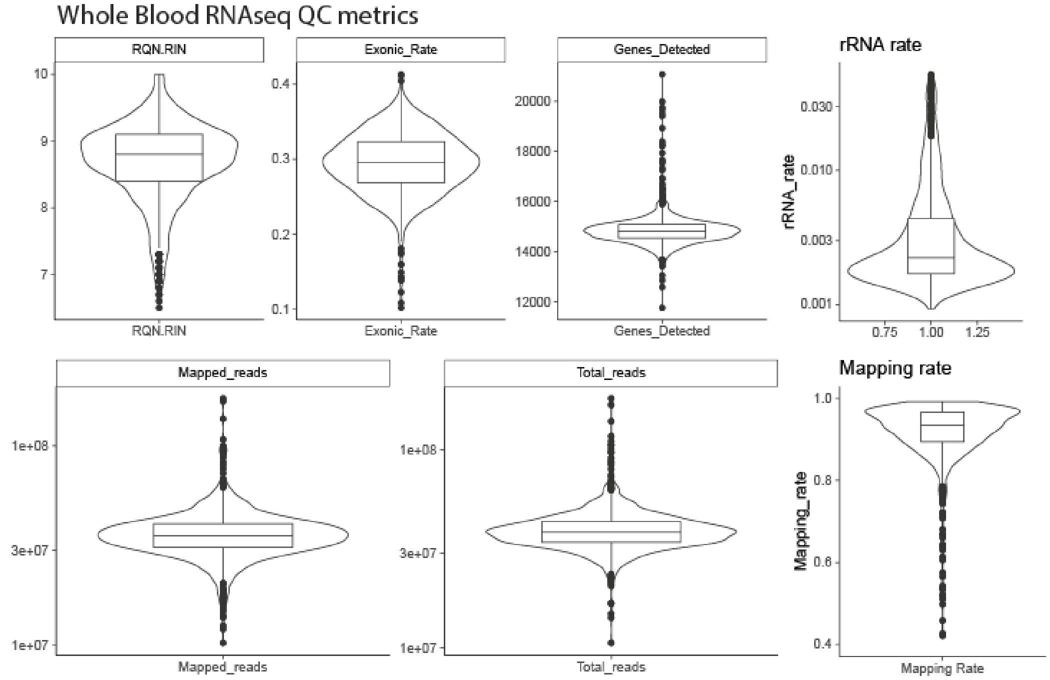
Supplementary Methods Figure 1

Biopsy RNAseq QC metrics



Biopsy Transcriptome	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	IQR
RQN.RIN	6.1000	8.2000	8.9000	8.7417	9.3000	10.0000	1.1000
ExonicRate	0.2174	0.5210	0.5795	0.5721	0.6306	0.7927	0.1096
rRNArate	0.0010	0.0026	0.0045	0.0145	0.0131	0.2003	0.0105
GenesDetected	12725	16640	16933	16882	17189	20119	549
Totalreads	4912436	33981396	39158771	41324909	45459309	290323139	11477913
Mappedreads	2580916	32489385	37491815	39463326	43533558	283887629	11044173
Mappingrate	0.2869	0.9484	0.9665	0.9553	0.9760	0.9884	0.0276

Whole Blood RNAseq QC metrics

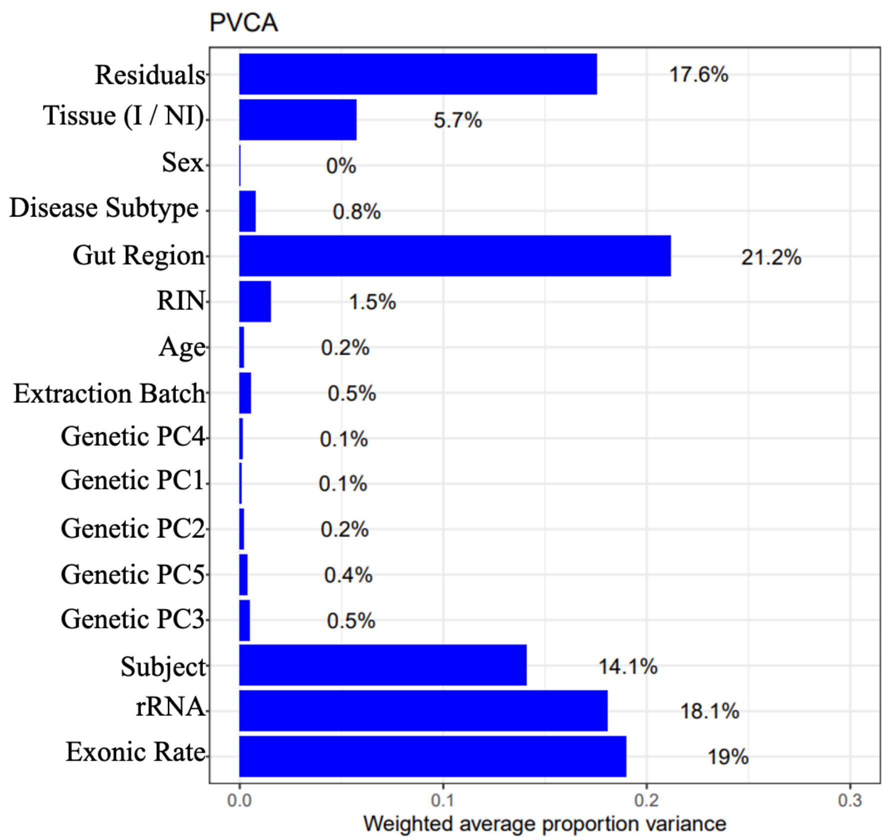


Blood Transcriptome	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	IQR
RQN.RIN	6.500	8.400	8.800	8.687	9.100	10.000	0.700
rRNA_rate	0.001	0.002	0.002	0.005	0.004	0.052	0.003
Exonic_Rate	0.102	0.269	0.295	0.294	0.323	0.413	0.054
Genes_Detected	11753	14533	14810	14872	15078	21074	545
Mapped_reads	10264145	30872857	35455324	37068906	40681240	172996421	9808384
Total_reads	10553848	34082853	38424727	40221012	43408197	182160962	9325344
Mapping_rate	0.421	0.894	0.934	0.917	0.967	0.992	0.072

Supplementary Methods Figure 1: RNA sequencing associated QC metrics for the MSCCR biopsy (top figure) and blood (bottom figure) data.

RNA transcriptome modeling:

PCA (main Figure 1b) revealed that region of biopsy was the largest factor contributing to variation in gene expression, followed by inflammation status, with disease sub-type (UC vs CD) showing very little separation. Variance partition analysis was also carried out as part of this analysis (Supplementary method Figure 2). It shows that region of the gut biopsied (variable = RegionRe in figure) had the largest variance, whereas tissue type (Inflamed/Non-Inflamed, variable = TypeRe) and disease subtype (variable= IBD_disease) were much smaller. For bMIS generation, biopsy data for patients with indeterminate IBD



Supplementary Method Figure 2: PVCA analysis of various potential technical and biological sources of variation associated with the biopsy RNA seq dataset. Variables: resid = residual; typeRe (Inflamed vs uninflamed); Demographics_gender (male vs female); IBD_disease (UC, CD or control); RegionRe (small versus large intestine regions); RQN.RIN (RNA quality number); Study_Eligibility_age… (Age at MSCCR study); Extraction… (Technical covariate for disease were removed (n=13) and biopsies identified as inflamed in the healthy control group were also removed (n=7). For subsequent analysis, biopsies from pouch patients (n=18 unique) were also removed. The data are available on GEO (GEO accession: GSE186507 for blood and GSE193677 for biopsy).

Generation of molecular inflammation scores (MIS) for biopsy (bMIS) and peripheral blood (cirMIS):

bMIS (biopsy molecular inflammation score):

Gene expression matrices from biopsy were generated from the count matrices using the voom transformation[9] on the count matrix using the *limma* framework. Voom-transformed gene expression data was modelled using a mixed-effect models with 'tissue type' (i.e. endoscopically inflamed or non-inflamed), 'intestine biopsy region' (ileum, colon, rectum etc) and 'disease sub-type' (Control, UC, CD) and its interactions as factors and a random factor for each patient, with technical (batch, RIN, rRNA rate, exonic rate) and relevant variables (age, gender, and genetic PC's 1-5) as covariates. In the *limma* model, control samples were accounted for as a covariate as "IBD vs. Control" in the development of overall bMIS (i.e. bMIS IBD) and as "CD vs. Control" or "UC vs. Control" for subtype bMIS's (bMIS CD or bMIS UC respectively). In this model differences between endoscopically inflamed and non-inflamed tissue were assessed for each intestinal region (7 possible including: rectum, sigmoid, left colon, transverse, right colon, cecum, ileum) and disease subtype, thus defining intestinal region- and disease subtype- specific inflammation signatures (**Figure 1b-c**). However, as we observed a strong correlation across the inflammation signatures, we generated a general IBD inflammation signature by fitting a model with tissue type, disease sub-type and intestine biopsy region (no interactions) and an inflammation signature for each disease subtype by including only an interaction term for tissue type by disease sub-type.

From the IBD, or CD and UC subtype-specific inflammation gene signatures, we defined the markers of biopsy inflammation as genes differentially expressed (up-regulated genes only) between endoscopically inflamed and non-inflamed biopsies, at FDR<0.05 and fold change (FCH)>2 and the bMIS score was derived by using a gene-set variation analysis (GSVA[10]). The inflammation score was built as the average z-score derived from the expression (adjusted for technical covariates) of the differentially expressed genes (DEGs) normalized by the square root of the number of genes[11]. As a result, each biopsy sample for the MSCCR cohort had a bMIS_IBD score as well as either a UC or CD disease sub-type specific score (bMIS_UC or bMIS_CD), depending on the patient's disease sub-type diagnosis. This score is based on all the genes differentially expressed, as we aimed to quantify the overall level of molecular inflammation (ie continuous score) and not to develop a predictor of endoscopic inflammation status (yes vs no). This rationale was also based on our experience in psoriasis where we developed a similar transcriptome scoring system[12,13, 14]. Alternative to this could be 1. to use a regularized regression, a popular machine learning algorithm to train a model that predicts the probability of a biopsy being inflamed or not or 2. use a regularized regression to identify a subset of predictive genes and then do GSVA. Neither of these approaches provided good validation results (data not shown), however, and our goal was a continuous measure applicable across different IBD cohorts.

cirMIS: Blood gene expression data from 1030 patients for which gut biopsy transcriptome data was available, was used to identify genes whose expression in blood associated with the level of intestinal inflammation. To obtain a patient-level gut inflammation score, we took advantage of the fact that multiple gut regions per individual were sampled and summarized the patient's individual bMIS scores into an intestinal-level (ileum-to-rectum) inflammation score (iMIS). We fit a mixed-effect model with tissue type and gut biopsy region as fixed effects and a random intercept and tissue type coefficient for each participant, with technical variables and covariates (age, gender, genetic PC's #1-5) adjusted data:

$$bMIS_{ij} = a_o+a_1T_{ij}+ a_2R_{ij}+b_i+b_{1i}T_{ij} +e_{ij}.$$

were Tij and Rij define the tissue type and gut biopsy region from the j biopsy of patient i. From this model the overall, region and tissue type independent scores in inflamed tissue were obtained as $Score=b_i+(a_1+b_{1i})*XI_i$ where XI_i is 1 if a patient does have an inflamed biopsy or zero otherwise.

The blood gene expression data was then modeled using a linear model with the continuous variables iMIS and technical covariates including imputed genetic PC's (#1-5), age at endoscopy, sex, and IBD disease sub-type. iMIS-associated blood genes were selected if the iMIS slope was significantly different than zero with FDR<0.05 and $|slope| \geq log_2(1.3)/\Delta$, where Δ represents 1/3 of the range of iMIS in all participants, that induced an absolute FCH>1.3 in gene expression. Those genes were then used as the input to generate a circulating molecular score that reflects intestinal inflammation (cirMIS) as the GSVA score of the gene expression (adjusted by technical covariates) matrix $\Theta = \begin{bmatrix} X_{Ui} \\ -X_{Di} \end{bmatrix}$ with U and D the set of genes positively and negatively associated respectively.

Statistical modeling:

Statistical analysis was carried out using R language version R v4.0.5[15] and its available packages. Each MIS was modeled using linear models with relevant factors depending on the comparison. When data was paired, ie several biopsies were available for the same patient, or different time points, mixed-effect models were fitted including fixed factors and random intercept for each subject using the *nlme* package in R. Marginal means and hypothesis of interests were tested using the *emmeans* package capabilities.

Logistic regression was used to evaluate the performance of iMIS, cirMIS, CRP and fecal calprotectin in classifying patients in endoscopic (SESCD<3 in CD patients or Mayo Score=0 in UC patients) or histological remission (GHAS score=0 in CD patients or Nancy score =0 in UC patients). AUC was calculated for each model from logistic regression and the AUC performance was compared by Delong's method.

Correlation of the endoscopic, histological, and clinical disease activity measures with the molecular scores was assessed using Spearman correlations, and Fisher's Z test was used to compare the correlations.

References:

1. Suarez-Farinas M, Tokuyama M, Wei G, et al. Intestinal inflammation modulates the expression of ACE2 and TMPRSS2 and potentially overlaps with the pathogenesis of SARS-CoV-2 related disease. Gastroenterology 2020.
2. Uzzan M, Martin J, Kenigsberg E. Mapping of B cell landscape in Ulcerative Colitis lesions reveals a pathogenic response that associates with treatment resistance and disease complications. Nature Medicine 2020:Under revision.
3. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904-9.
4. Di'Narzo AF, Houten SM, Kosoy R, et al. Integrative Analysis of the Inflammatory Bowel Disease Serum Metabolome Improves Our Understanding of Genetic Etiology and Points to Novel Putative Therapeutic Targets. Gastroenterology 2022;162:828-843 e11.
5. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.
6. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) 2013;29:15-21.
7. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2013;30:923-930.
8. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 2010;11:R25.
9. Law CW, Chen Y, Shi W, et al. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014;15:R29.

10. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 2013;14:7.
11. Lee E, Chuang HY, Kim JW, et al. Inferring pathway activity toward precise disease classification. PLoS Comput Biol 2008;4:e1000217.
12. Hamilton JD, Suarez-Farinas M, Dhingra N, et al. Dupilumab improves the molecular signature in skin of patients with moderate-to-severe atopic dermatitis. J Allergy Clin Immunol 2014;134:1293-1300.
13. Guttman-Yassky E, Ungar B, Malik K, et al. Molecular signatures order the potency of topically applied anti-inflammatory drugs in patients with atopic dermatitis. J Allergy Clin Immunol 2017;140:1032-1042 e13.
14. Beck LA, Thaci D, Hamilton JD, et al. Dupilumab treatment in adults with moderate-to-severe atopic dermatitis. N Engl J Med 2014;371:130-9.
15. R Development Core Team. R: A language and environment for statistical computing. v3.0.3 ed: R foundation for statistical computing, 2020.