Supplementary material

Machine learning hyperparameters and cross-validation schemes

Detailed hyperparameter settings and cross-validation strategies for all five machine learning algorithms are as follows:

LASSO regression: Performed using the glmnet R package with $\alpha = 1$ (L1 regularization). The optimal penalty parameter λ was determined by 10-fold cross-validation, with λ_min (corresponding to the minimum binomial deviance) selected as the feature selection threshold.

Random Forest: Constructed using the randomForest R package with 2,000 decision trees (ntree = 2,000). Feature importance was evaluated by the Mean Decrease Gini index, and the top 15 features were retained for downstream analysis.

XGBoost: Implemented using the xgboost R package with the following hyperparameters: maximum tree depth (max_depth) = 6, learning rate (eta) = 0.5, number of boosting iterations (nround) = 25, and objective function set to binary logistic regression (binary:logistic).

Boruta algorithm: Applied as a Random Forest–based all-relevant feature selection method with a maximum of 500 iterations (maxRuns = 500). Features were classified by comparing their importance scores against shadow variables, with "Confirmed" features retained and "Tentative" features further resolved using the TentativeRoughFix function.

SVM-RFE: Performed using the caret R package with a radial basis function kernel (svmRadial). The optimal feature subset was determined by 5-fold cross-validation, evaluating classification accuracy and error rate across varying feature numbers.

RNA isolation and library preparation

Fresh target tissues of suitable dimensions were promptly excised and briefly rinsed with ice-cold, enzyme-free water or phosphate-buffered saline (PBS) to eliminate blood and debris. The specimens were immediately immersed in liquid nitrogen for rapid freezing, then transferred into pre-cooled cryovials following complete freezing, and subsequently stored in liquid nitrogen prior to long-term preservation at −80 °C. Total RNA was extracted using the TRIzol reagent (Invitrogen, CA, USA) according to the manufacturer 's protocol. RNA purity and quantification were evaluated using the NanoDrop 2000 spectrophotometer (Thermo Scientific, USA). RNA integrity was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Then the libraries were constructed using VAHTS Universal V6 RNA-seq Library Prep Kit according to the manufacturer 's instructions. The transcriptome sequencing was conducted by OE Biotech Co., Ltd. (Shanghai, China).

Isolation of B cells from tissues

Fresh anal fistula tissues were initially washed with phosphate-buffered saline (PBS) and subsequently minced into pieces measuring approximately 1 mm^3. The minced tissues underwent enzymatic digestion using Collagenase IV (1 mg/mL) and DNase I (100 μg/mL) in RPMI 1640 medium at 37°C for 45 minutes with gentle agitation. The resultant cell suspension was passed through a 70-μm strainer and washed. Cells were then resuspended in fluorescence-activated cell sorting (FACS) buffer and treated with an Fc receptor blocking solution. Following this, cells were stained with anti-CD3 , anti-CD19, and a viability dye to exclude non-viable cells. CD19+ CD3- B cells were isolated using a cell sorter. Dead cells and doublets were excluded from the analysis. The sorted cells were immediately lysed for subsequent RNA extraction.

RNA extraction

Total RNA was extracted from from sorted $CD3^{-}CD19^{+}$ B cells using RNeasy Micro Kit (Qiagen) according to the manufacturer's specifications. The yield of RNA was determined using a NanoDrop 2000 spectrophotometer (Thermo Scientific, USA), and the integrity was evaluated using agarose gel electrophoresis stained with ethidium bromide.

Real-time quantitative RT-PCR

mRNA quantification was performed with a two-step reaction process: reverse transcription (RT) and PCR. The RT reaction was performed in a GeneAmp® PCR System 9700 (Applied Biosystems, USA) as follows: 50 ng RNA, 2μl of 5×TransScript All-in-one SuperMix for qPCR, 0.5μl of gDNA Removerin a total volume of 10 μl was incubated at 42˚C for 15 min, followed by heat inactivation of RT for 5 s at 85˚C. The 10 μl RT reaction mixture was then held at -20˚C. Real-time PCR was performed using LightCycler® 480 II Real-time PCR Instrument (Roche, Swiss) with 10 μl PCR reaction mixture that included 1 μl of cDNA, 5μl of AceQ U+ Probe Master Mix、0.5μl of TaqMan® Gene Expression Assays（20×）, and 3.5 μl of nuclease-free water. Reactions were incubated in a 384-well optical plate (Roche, Swiss) at 37℃ for 2 min, 95˚C for 5 min, followed by 45 cycles of 95˚C for 10 s, 60˚C for 30 s. Each sample was run in triplicate for analysis.

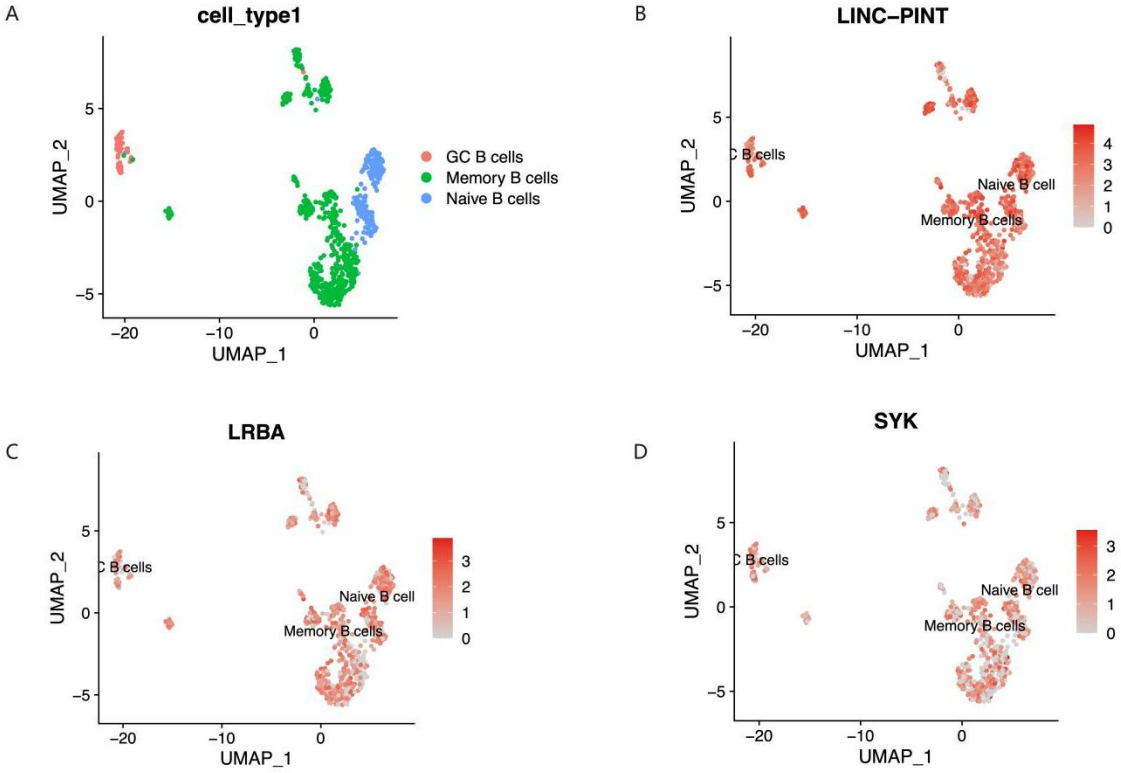
The primer sequences used were as follows:

- SYK forward primer: 5′-CGTATGAGCCAGAACTTGCACC-3′, reverse primer: 5′-CTTTCGGTCCAGGTAAACCTCC-3′;
- LINC-PINT forward primer: 5′-CCACAGGGCTCACTTCGTTT-3′, reverse primer: 5′-GGTGCAGGGAGACATTGGTA-3 ′;
- LRBA forward primer: 5′-GCCAGTGTTTCAGCCATTTG-3′, reverse

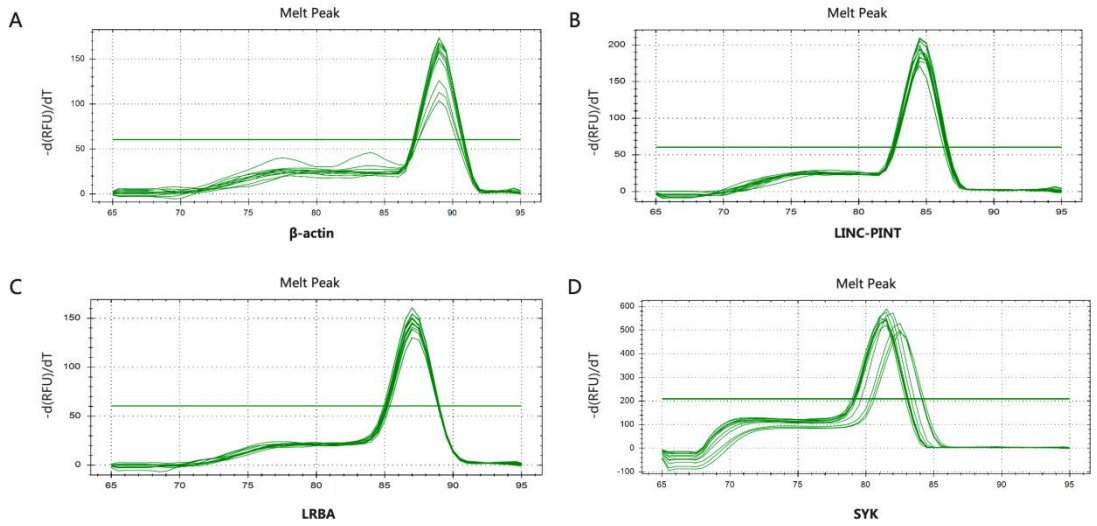
primer: 5′-CTCCATTGTCCTGGTCACTTC-3′;

- β-actin (ACTB) forward primer: 5′-CATGTACGTTGCTATCCAGGC-3′,reverse primer: 5′-CTCCTTAATGTCACGCACGAT-3 ′.

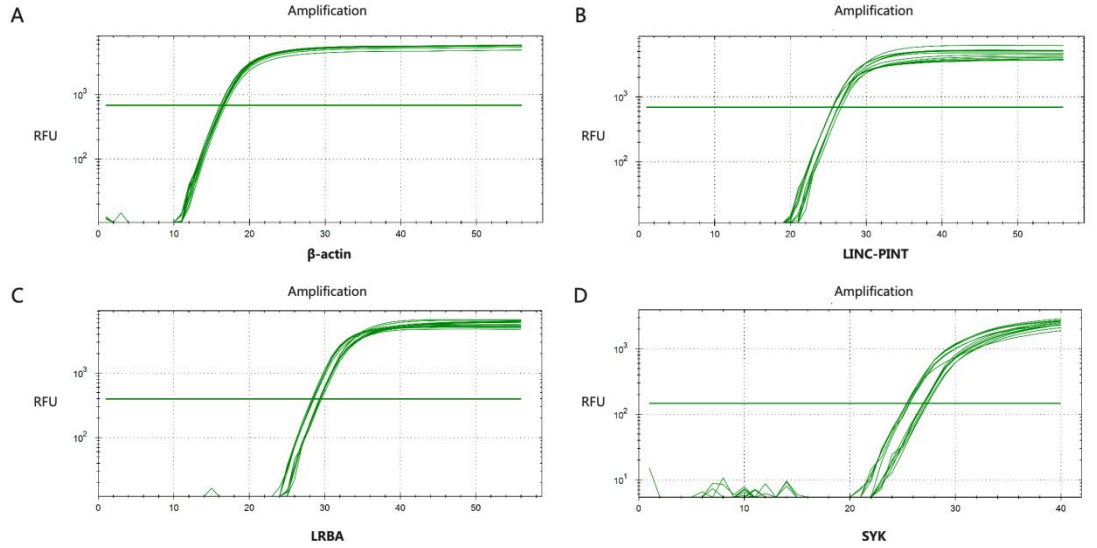
The expression levels of mRNAs were normalized to ACTB and were calculated using the $2^{-\Delta\Delta Ct}$ method. The qRT-PCR were conducted by OE Biotech Co., Ltd. (Shanghai, China).



Supplementary Figure 1 uniform manifold approximation and projection clustering and gene expression mapping of B-cell subsets. A: UMAP plot showing the sub-clustering of B cells into three distinct subsets: GC B cells (red), Memory B cells (green), and Naive B cells (blue); B-D: Feature plots displaying the expression distribution of LINC-PINT (B), LRBA (C), and SYK (D) across the identified B-cell subsets. The color scale (grey to red) indicates the expression level, showing a broad distribution of these genes across all subsets without restriction to a specific cluster.



Supplementary Figure 2 Melting curve analysis for specificity validation. A-D: Representative melting curves of the qPCR products are shown for the reference gene β-actin (A), LINC-PINT (B), LRBA (C), SYK (D). The presence of a single, sharp dissociation peak in each plot confirms the specific amplification of the target genes and the absence of primer dimers or non-specific products. The X-axis represents the temperature, and the Y-axis represents the derivative reporter signal.



Supplementary Figure 3 Real-time PCR amplification plots. A-D: The reference gene β-actin (A), LINC-PINT (B), LRBA (C), SYK (D). The curves display a typical exponential growth phase, indicating efficient PCR amplification across the samples.

Supplementary Table 1 Clinical characteristics of the individuals included in the single cell sequence analysis

Sample ID	Age (years)	Sex	BMI (kg/m2)	Smoking History	Disease Duration (months)	Fistula Classification	Etiology	Medication History
Anal fistula 1	45	Male	24.5	No	2	simple	Cryptoglandular	No
Anal fistula 2	52	Female	22.1	Yes	8	complex	Cryptoglandular	No
Anal fistula 3	48	Male	25.8	No	15	complex	Cryptoglandular	No
Control 1	46	Male	23.4	Yes	N/A	N/A	N/A	No
Control 2	50	Female	25.6	No	N/A	N/A	N/A	No
Control 3	47	Male	24.9	No	N/A	N/A	N/A	No

Supplementary Table 2 Comparison of baseline clinical characteristics included in the RT-qPCR analysis between two groups

Characteristics	Anal Fistula Group (n=18)	Control Group (n=20)	P-value
Demographics			
Age, years (Mean ± SD)	36.1 ± 12.7	38.6 ± 11.5	0.53
Sex (Male/Female)	16 /2	16 / 4	0.45
BMI, kg/m²	24.1 ± 2.4	23.6 ± 2.2	0.48
Smoking History (Yes/No)	13 /5	16 / 4	0.57
Clinical Features			
Disease Duration (months)	14.8 ± 12.6	N/A	-
Fistula Classification			
Complex	7 (39%)	N/A	-
Simple	11(61%)	N/A	-
Etiology			
Cryptoglandular	18 (100%)	N/A	-
Crohn's Disease	0 (0%)	N/A	-
Medication History			
Antibiotics (last 2 weeks)(Yes/No)	2(11%)	0 (0%)	0.13
Immunomodulators	0 (0%)	0 (0%)	-