Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Exploratory cross-cohort assessment of a post-hoc FAP-associated epithelial transcriptomic module in colorectal datasets

Yidi Zhao
Xizang Minzu University
Yunfei Jia
Xizang Minzu University
Hehe Liao
Xizang Minzu University
Yujie Ma
Xizang Minzu University
Zifei Zhang
Xizang Minzu University
Yao Zheng
Xizang Minzu University
Jianan Zhang
Xizang Minzu University
Haidie Xie
Xizang Minzu University
Pingyi Wang
wangpingyi@xzmu.edu.cn
Xizang Minzu University

Research Article

Keywords: colorectal adenoma, familial adenomatous polyposis, single-cell RNA sequencing, single-nucleus RNA sequencing, pseudobulk, random gene-set resampling, transcriptomic data reuse, reproducibility, negative results

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10669004/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Public transcriptomic reuse can test whether exploratory signals remain coherent across datasets, but post-hoc selection can create apparent reproducibility. We evaluated the evidence boundary of a locked 36-gene epithelial module generated after a negative donor-level screen in colorectal single-nucleus data containing paired familial adenomatous polyposis (FAP) samples.

Methods

We processed datasets independently, used donor-level pseudobulk where identifiers were recoverable and retained negative and non-estimable branches. We compared M02 with 10,000 expression- and variability-matched random 36-gene sets from its 1,000-gene source background, plus an unmatched reference. Resampling summaries were not treated as post-selection P values.

Results

The project-recorded normal-to-adenoma gene screen identified no genes at FDR < 0.05 in four epithelial states. In the four FAP donors used during filtering, mean M02 differences were 1.276 in all epithelium and 0.778 in Stem/progenitor profiles; both exceeded all matched random modules, and all leave-one-donor-out means remained positive. This was a selection-conditioned descriptive association. Three matched GSE161277 donors, unused during M02 generation, showed positive differences (mean 1.645, 95% CI 0.592–2.698; FDR = 0.128). Three bulk datasets showed positive associations, but two were sample scoped and heterogeneity was high (I^2 = 87.6%). Stool models were negative and did not directly test M02; LST-G and patient-level spatial endpoints were not estimable.

Conclusions

M02 remains a post-hoc FAP/adenoma-associated candidate, not an independently supported sporadic-adenoma signature. Random-module resampling addressed arbitrariness but did not remove design-time selection bias or limited external patient-level evidence.

Background

Colorectal tumorigenesis is commonly described as progression from normal mucosa through adenoma to invasive cancer, but hereditary background, anatomical site and lesion morphology create multiple biological routes rather than a uniform sequence [1]. Familial adenomatous polyposis (FAP), caused predominantly by germline APC alterations, is biologically distinct from most sporadic adenomas [2]. Single-cell, single-nucleus and bulk transcriptomic datasets permit comparison of epithelial states across these contexts [3–5], but they also create analytical hazards. Thousands of cells or nuclei from one donor do not constitute independent biological replicates, and methods that ignore donor structure can generate false discoveries [6–8]. Cross-platform agreement can also be exaggerated when matrices are concatenated, sample-level arrays are relabeled as patients or failed feature mapping is hidden.

The project initially asked whether epithelial genes changed at the normal-to-adenoma transition and continued in the same direction toward cancer. GSE201348 supplied this donor-level single-nucleus screen [3]. The project-recorded gene-level result was negative. A subsequent post-hoc amendment formed stage-label-blind covariance modules and then filtered them using stage effects, paired FAP direction and leave-one-donor-out behavior. Six modules met those exploratory criteria, and a 36-gene Stem/progenitor module, M02, was selected for cross-dataset assessment.

This sequence creates an essential dependency: the four paired FAP donors used to assess direction also influenced which module was retained. Recalculated summaries in those donors are therefore conditional on selection and likely optimistic. They cannot independently support the module, estimate a FAP population effect or justify generalization to sporadic adenoma. The term “stage-label-blind” applies only to covariance clustering, not to the subsequent stage-aware and FAP-aware filtering step.

We therefore asked a narrower question: what are the estimability, direction and evidence boundaries of a post-hoc FAP/adenoma-associated candidate when its membership is fixed across public colorectal transcriptomic datasets? The analysis retained the original single-nucleus screen and module-generating cohort, one three-donor matched single-cell cohort and three separately processed bulk-tissue datasets. Cancer-only, stool-RNA, laterally spreading tumor and spatial branches provided supplementary context or technical feasibility assessments; they were not counted as accumulating votes for a biological claim.

Our objective was to evaluate the locked score across datasets while preserving each dataset’s biological unit, platform and eligibility constraints. We retained negative and non-estimable branches, did not tally the number of positive sensitivities as evidence and kept causal, diagnostic and sporadic-disease claims outside the recorded claim ceiling.

Methods

Study design and dataset roles

All analyses used public, de-identified data. Dataset roles, biological units and independence constraints are summarized in Fig. 1 and Table 1. Records from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) were GSE201348 [3], GSE161277 [4] and GSE132465 [5] for single-nucleus or single-cell analyses; GSE41657 (repository record), GSE100179 [9] and GSE8671 [10] for separately processed bulk comparisons; GSE99573 [11] for stool-RNA analysis; and GSE200997 [16] and GSE226997 for cancer-context or secondary technical roles. Figshare record 29925404 and its source report [12, 13] supplied the LST-G data. The DNA Data Bank of Japan (DDBJ) Genomic Expression Archive (GEA) record E-GEAD-622 [14] supplied the spatial feasibility branch; E-GEAD-619 and E-GEAD-579 [17] had secondary technical roles. TCGA-COAD [15] was accessed through the National Cancer Institute (NCI) Genomic Data Commons (GDC). Repository URLs and complete accession details are provided in the Availability of data and materials section and Supplementary Information 4.

Table 1
Dataset roles, biological units and principal constraints

Dataset	Role in manuscript	Lesion/context	Biological/statistical unit	Unit count	Pairing	M02 feature coverage	Principal constraint
GSE201348	Primary negative single-nucleus screen; post-hoc candidate generation	normal mucosa; polyp/adenoma; cancer	Donor pseudobulk	14 total donors; four paired FAP donors used during candidate filtering	Partially paired; four paired FAP donors entered candidate filtering	36/36 for the module branch; not applicable to the primary gene screen	Internal M02 statistics are selection affected; sporadic adenoma and stage-by-FAP effects not estimable
GSE161277	Small matched single-cell description	normal mucosa; adenoma; carcinoma	Donor paired difference	3 matched donors	Yes; three matched donors	36/36	Underpowered; FDR 0.128; hereditary stratification unavailable
GSE132465	Supplementary cancer-only context	normal mucosa to cancer	Donor paired difference	8 evaluable matched donors	Yes; eight cancer–normal donor pairs	33/36	No adenoma endpoint
GSE41657	Bulk adenoma association	normal; low-grade adenoma; high-grade adenoma; cancer	Array/sample-scoped unit	63 normal/adenoma arrays; 88 arrays overall	No verified pairing	29/36	Patient identity unresolved; bulk tissue
GSE100179	Bulk adenoma association	normal; adenoma; cancer	Array/sample-scoped unit	40 normal/adenoma arrays; 60 arrays overall	No verified pairing	32/36	Patient identity unresolved; bulk tissue
GSE8671	Paired bulk adenoma association	normal mucosa to adenoma	Patient pair	32 verified pairs	Yes; 32 verified pairs	29/36	Bulk tissue; hereditary status unresolved
TCGA-COAD	Supplementary cancer-only context	primary cancer to normal; stage; MSI; CMS; survival	Patient	41 paired cancer-normal patients	YES for cancer-normal subset	34/36	No adenoma endpoint
GSE99573	Supplementary model-transport audit	stool RNA normal; adenoma; cancer	Participant	65 independent test participants	NO	31/36 mapped; 27/36 detectable	Models selected 0/0/1 features; test set retired
Figshare 29925404	Supplementary mapping-limited LST-G sensitivity	LST-G; protruded adenoma; normal	Patient	7 patients; 2 matched donors	Yes for P1 and P5 only	33/36 mapped; canonical endpoint not estimable	36-gene endpoint not estimable; hereditary status unknown
E-GEAD-622	Supplementary spatial technical feasibility	pTis sections with frozen normal, adenoma, carcinoma and other regions	Patient required; spots and pseudo-spots technical only	1/4 patients passed region gate	Planned within-section pairing; only one eligible patient	30/36	No independent pathology gold standard; patient-level analysis not estimable

Numerical summary with effect definitions, confidence intervals, P values, FDR values and mandatory interpretation. FAP summaries are marked selection conditioned and non-independent. NA indicates not available or not estimable, never zero.

The patient or donor was the biological replicate whenever a reliable identifier was available. Nuclei, cells, libraries, tissue fragments, arrays, capture areas, regions and spatial spots were nested or technical observations. GSE41657 and GSE100179 remained sample scoped because patient identities could not be verified; GSE8671 contained 32 verified normal–adenoma patient pairs. Expression matrices from different cohorts or platforms were never concatenated. The available study size was determined by repository eligibility and recoverable biological identifiers, not by observed M02 direction; no post-hoc power calculation was used to justify the included cohorts.

Bias-control measures included donor/patient-level inference, independent dataset preprocessing, fixed M02 membership, separation of candidate generation from later assessment, a selection-conditioned random-gene-set audit and complete retention of negative and non-estimable branches. Missing clinical or technical metadata were not imputed: unresolved fields remained NA, analyses were omitted when a design was not estimable and arrays without recoverable donor identifiers remained sample scoped. These safeguards do not remove the selection bias inherent in the post-hoc M02 construction.

GSE201348 preprocessing, quality control and annotation

GSE201348 comprised single-nucleus RNA-sequencing libraries. Each library was processed independently before integration. The final primary gate retained nuclei with > 400 and < 4,000 detected genes, < 10,000 UMI and < 5% mitochondrial RNA, following the deposited analysis design. The previously approved 100-UMI/100-gene setting for GSM6061701 was retained only as a pilot sensitivity setting and was not mixed into the primary object. After primary hard QC, that library contributed 172 nuclei and remained flagged as low depth/high risk. Ribosomal proportion was diagnostic only. Doublets were assessed independently per eligible library using scDblFinder [18]. Three of 72 libraries had < 100 nuclei after hard QC and were excluded. The final objects contained 201,649 nuclei from 69 libraries, 67 biological tissues and all 14 donors.

Seurat was used for normalization, variable-feature selection and principal components [19]. Harmony was applied only to the low-dimensional representation; raw counts and unintegrated objects were retained [20]. Major lineages were assigned from combinations of positive and exclusionary markers. Epithelial profiles were reclustered into broad states including Absorptive, Adenoma-like, Cycling, Stem/progenitor, BEST4-positive, enteroendocrine, goblet/secretory and tuft states. CopyKAT supplied auxiliary copy-number evidence [21]; tissue origin or EPCAM expression alone did not define malignancy.

The annotation audit reports marker-expression dot plots, cluster-to-state assignments, donor composition and stage/FAP composition (Fig. S1; Table S2). Three M02 genes occurred in the recorded epithelial-state marker evidence: ASCL2 in the broad Stem/progenitor panel and NKD1/APCDD1 in cluster-level evidence. This overlap is disclosed because the module was generated within the annotated Stem/progenitor compartment; it limits independence between state definition and module characterization. No M02 gene is mitochondrial, ribosomal or a canonical cell-cycle QC gene. INPP5D was flagged before the contamination sensitivity analysis as an immune-lineage concern.

Donor-level primary gene screen

Raw epithelial counts were summed by donor, lesion stage and epithelial state. Pseudobulks required at least 20 nuclei and 10,000 total raw UMI. Four states had at least three eligible donors in each of the normal, adenoma and cancer stages: Absorptive (6, 8 and 3 donors), Adenoma-like (6, 8 and 4), Cycling (6, 9 and 4) and Stem/progenitor (6, 9 and 4).

Counts were filtered and normalized by trimmed mean of M-values (TMM) in edgeR [22, 23] and modeled with limma-voom [24]. Repeated stages from the same donor were handled with duplicateCorrelation; FAP status was included only when estimable. Anatomical-location covariates were audited but not forced into underidentified models. Missing hereditary, location or patient-identity fields were not imputed. The primary normal-to-adenoma screen used a Benjamini–Hochberg FDR threshold of 0.05 [25]. Effect estimates, 95% confidence intervals (CIs), two-sided P values, FDR values and donor counts were retained for every tested gene.

Post-hoc module construction and candidate selection

Module construction and candidate filtering were distinct steps. First, within each epithelial state, up to 1,000 highly variable genes were selected without using lesion-stage labels. Spearman correlation matrices were clustered into at most 20 modules. Modules required at least 10 genes and a median within-module correlation of at least 0.20.

Second, stage labels and paired FAP data were used to filter these covariance modules. A module required a normal-to-adenoma FDR of at most 0.05, an absolute standardized effect of at least 0.50, a cancer-to-normal effect in the same direction and at least 80% of the early effect, a compatible cancer-to-adenoma direction or plateau, concordance with the paired FAP direction, and leave-one-donor-out sign stability of at least 0.75. Six modules passed. M02 was selected from these six for downstream cross-cohort assessment. This was post-hoc candidate generation; it was not part of the original gene-level discovery question and was not publicly preregistered.

M02 contained 36 genes with equal weights of 1/36. Its canonical membership was documented before specified downstream evaluations with SHA256 d4e34472243b0259650aca3123a5df7e767e9a408a409cef08861937f2360a30. That timestamp demonstrates file order within the project but is not claimed as independent preregistration. No gene was added or replaced after downstream results were viewed.

Module composition and selection-conditioned audits

The 36 genes were grouped into broad, non-exclusive descriptive categories, and gene descriptions were mapped with g:Profiler [26]. Overrepresentation analysis used g:Profiler version e114_eg62_p19_27110d83, g:SCS correction and a custom background of the 1,000 unique genes entering Stem/progenitor covariance modules. Functional characterization was performed after module selection and did not modify membership.

For the four paired FAP donors, TMM-normalized log2 counts per million (log2CPM; prior count 2) were calculated for the 36 genes in all epithelium and the Stem/progenitor compartment. Patient-specific adenoma-minus-normal differences, gene contributions and leave-one-gene-out score differences were summarized descriptively. Because paired FAP direction was part of candidate selection, no P value from these four donors is interpreted as an independent test. Their role is a selection-conditioned exploratory association, not independent inferential support.

Selection-conditioned random-gene-set resampling audit

Before inspecting resampling output, we froze the source universe, matching variables, iteration counts, seeds, scoring method and reporting limits. The universe comprised the 1,000 genes used for Stem/progenitor covariance clustering; the 36 M02 genes were excluded from random sampling. For the primary reference, each source-universe gene was classified by quintiles of mean and standard deviation across eligible Stem/progenitor TMM-log2CPM pseudobulks. We generated 10,000 unique 36-gene modules that matched the number of M02 genes in every two-dimensional expression-variability stratum. A second 10,000-module reference sampled uniformly from the same background. The same sampled membership was scored in all epithelium and Stem/progenitor profiles using the locked equal-weight TMM-log2CPM method.

We compared the observed mean paired FAP difference with each random-module distribution and reported its percentile and plus-one upper-tail fraction. These quantities are descriptive random-gene-set references, not post-selection P values: the same four FAP donors contributed to M02 filtering, so random resampling cannot restore independence or type-I error control. We also summarized donor-specific effects and all four leave-one-donor-out means. No resampling result was used to change M02, its weights, the gene universe, either compartment or any external dataset.

GSE161277 descriptive assessment

GSE161277 was processed independently and was not used during M02 generation, gene ranking, module filtering or module selection. After per-capture quality control and doublet assessment, 38,239 tissue cells remained, including 13,734 epithelial cells. Multiple adenoma tissues from the same donor were aggregated within donor and stage before scoring. The assessment used three fully matched donors. Raw counts were aggregated to donor-stage pseudobulks, TMM-normalized and transformed to log2CPM. The 36-gene equal-weight score was calculated without redefining the module. The mean paired effect, t confidence interval, nominal two-sided P value and BH FDR across the six previously selected modules were reported. Given n = 3, patient-specific directions are the primary display; inferential estimates are descriptive and underpowered.

Independent bulk-tissue analyses and scale sensitivity

GSE41657, GSE100179 [9] and GSE8671 [10] were normalized and analyzed separately. Platform-specific probe mappings were audited before scoring. GSE41657 compared normal, low-grade adenoma, high-grade adenoma and cancer arrays; GSE100179 compared normal, adenoma and cancer arrays; GSE8671 used 32 verified normal–adenoma pairs. GSE41657 and GSE100179 used heteroskedasticity-consistent regression at the array/sample level because patient identifiers were unresolved. GSE8671 used paired patient differences.

The original analysis combined accession-level adenoma-versus-normal effects using a random-effects model with restricted maximum likelihood and Knapp-Hartung inference [27, 28]. These effects were on platform-specific module-score scales. As an additional non-selective analysis, we evaluated scale sensitivity: each accession-specific effect and standard error was divided by the standard deviation of M02 scores among its normal and adenoma samples. The three standardized accession effects were synthesized with REML/Knapp-Hartung and repeated with Paule-Mandel heterogeneity estimation. Leave-one-accession-out estimates were reported. This sensitivity does not recover patient independence in the two sample-scoped cohorts and was not used to redefine the module.

Supplementary transportability and technical-feasibility audits

Cancer-only analyses in GSE132465 and TCGA-COAD [15] were kept separate from adenoma-onset evidence. The GSE99573 stool-RNA models [11] were trained and locked before one use of the held-out test set. LASSO-logistic regression used only the training set for feature and penalty selection [29], and AUCs were computed with pROC [30]. After testing, the split was retired. We report the frozen coefficients, number of unique predicted probabilities and test metrics; no model was refit, and the 36-gene M02 score was not directly tested in this branch.

For Figshare 29925404 [12, 13], 33 of 36 canonical genes were present in the integer-count feature space; AL022068.1, AP001636.3 and LINC01748 were absent. The historical name M02_29OF36_COVERAGE_SENS denoted a minimum coverage threshold. The implemented score averaged all 33 mapped genes; it was not a fixed 29-gene subset. Predefined reduced-coverage, gene-z and UCell [31] scores were sensitivity descriptions only. Because the canonical endpoint was unavailable and only two same-site matched LST-G donors were present, this branch remained non-inferential.

Pseudo-spots were used solely to evaluate score recoverability under simulated depth, dropout, epithelial proportion and contamination; they were not biological replicates. For E-GEAD-622 [14], two image-processing pipelines generated region labels without using M02 values, but no independent pathologist gold standard was available. Only case4 met the frozen agreement and paired-region criteria; cases1–3 failed. The patient-level comparison required at least three eligible patients and was therefore not estimable. In case4, the primary 30/36 score and most sensitivities were negative, whereas the two-region gene-z sensitivity was effectively zero; the frozen decision was DESCRIPTIVE_METHOD_DEPENDENT, with no population inference.

Statistical reporting and reproducibility

In this manuscript, 'prespecified' means recorded in the project workflow before the relevant downstream endpoint was inspected; it does not mean that the analysis was publicly preregistered.

All hypothesis tests were two sided. For gene-level analyses, Benjamini–Hochberg FDR was applied within each epithelial-state-by-contrast family; for GSE161277, the FDR family comprised the six previously selected modules. Other families are identified in the source tables. NA denotes unavailable or not computed, never zero. Non-estimable outcomes were not converted to negative outcomes. Nuclei, cells, arrays and spots did not increase the biological sample size. Small-sample results are accompanied by the number and identity scope of biological units.

Analytical scripts, configuration files, small source-data tables, environment records and checksums are provided in the supplementary source package, consistent with reproducible-computation principles [32]. The versioned baseline package is available on GitHub and Zenodo [33]; it supports provenance checks and reconstruction of key small source data, not a complete automated rerun of large single-cell or single-nucleus objects.

Results

The donor-level normal-to-adenoma gene screen was negative

Of 72 deposited GSE201348 single-nucleus libraries, three FAP adenoma/polyp libraries had < 100 nuclei after hard QC and were excluded; the analysis retained 201,649 nuclei from 69 libraries, 67 tissues and all 14 donors (Table S1). No gene reached FDR < 0.05 in any of the four epithelial states eligible for all three stages (Fig. 2a; Table 2). Minimum normal-to-adenoma P values were 7.17×10^{-6}, 1.44×10^{-5}, 1.10×10^{-5} and 9.49×10^{-6} for Absorptive, Adenoma-like, Cycling and Stem/progenitor states; the corresponding minimum FDR values were 0.0866, 0.0519, 0.0625 and 0.0534. The project-recorded screen therefore yielded zero genes and zero primary candidate programs. Cancer-associated changes did not satisfy the required adenoma-stage endpoint and were not relabeled as early-transition findings.

Table 2
Primary, candidate-generation and cross-cohort results

Dataset	Analysis	Biological/statistical unit	Effect (95% CI)	P	FDR	Result status	Principal limitation
GSE201348	Absorptive	Donor-level pseudobulk (n=normal = 6;adenoma = 8;cancer = 3)	No gene passed FDR < 0.05	7.17e-06	0.0866	Primary gene-level screen: zero genes at FDR < 0.05	State-specific eligible donor counts; no threshold relaxation
GSE201348	Adenoma_like	Donor-level pseudobulk (n=normal = 6;adenoma = 8;cancer = 4)	No gene passed FDR < 0.05	1.44e-05	0.0519	Primary gene-level screen: zero genes at FDR < 0.05	State-specific eligible donor counts; no threshold relaxation
GSE201348	Cycling	Donor-level pseudobulk (n=normal = 6;adenoma = 9;cancer = 4)	No gene passed FDR < 0.05	1.10e-05	0.0625	Primary gene-level screen: zero genes at FDR < 0.05	State-specific eligible donor counts; no threshold relaxation
GSE201348	Stem_progenitor	Donor-level pseudobulk (n=normal = 6;adenoma = 9;cancer = 4)	No gene passed FDR < 0.05	9.49e-06	0.0534	Primary gene-level screen: zero genes at FDR < 0.05	State-specific eligible donor counts; no threshold relaxation
GSE201348	All_epithelial	donor (n = 4)	1.276 (0.243 to 2.310)	0.0293	0.0586	Selection-conditioned candidate-generation description; not an independent test	Same-donor selection-conditioned description; n = 4; no independent-evidence credit
GSE201348	Stem_progenitor	donor (n = 4)	0.778 (0.066 to 1.491)	0.0402	0.0482	Selection-conditioned candidate-generation description; not an independent test	Same-donor selection-conditioned description; n = 4; no independent-evidence credit
GSE161277	Normal to adenoma	Donor paired difference (n = 3)	1.645 (0.592 to 2.698)	0.0214	0.1285	Three-donor concordant description; underpowered; multiplicity-adjusted result not significant	One independent cohort; n = 3; FDR = 0.128; counted once
GSE41657	Adenoma versus normal	Array/sample-scoped unit (n = 63)	2.420 (2.215 to 2.626)	2.48e-32	1.49e-31	Bulk-tissue association	Patient identity not verifiable; sample-scoped arrays; bulk tissue
GSE100179	Adenoma versus normal	Array/sample-scoped unit (n = 40)	2.080 (1.832 to 2.327)	2.31e-19	1.38e-18	Bulk-tissue association	Patient identity not verifiable; sample-scoped arrays; bulk tissue
GSE8671	Adenoma versus normal	Patient pair (n = 32)	1.846 (1.666 to 2.026)	7.65e-20	4.59e-19	Bulk-tissue association	32 verified matched patients; bulk tissue

Dataset	Analysis	Biological/statistical unit	Effect (95% CI)	P	FDR	Result status	Principal limitation
GSE41657;GSE100179;GSE8671	Random-effects adenoma versus normal	Cohort-level random-effects meta-analysis (n = 3)	2.113 (1.389 to 2.837)	0.0063	0.0377	Three-accession random-effects summary; not pooled patients	I2 = 87.6%; prediction interval 0.725 to 3.500; not a pooled-patient analysis
GSE99573	adenoma_vs_normal	Participant (n = 42)	0.500 (0.500 to 0.500)	NA	NA	Frozen model-transport result	Immutable test set used once and retired
GSE99573	cancer_vs_normal	Participant (n = 45)	0.500 (0.500 to 0.500)	NA	NA	Frozen model-transport result	Immutable test set used once and retired
GSE99573	neoplasia_vs_normal	Participant (n = 65)	0.530 (0.380 to 0.679)	NA	NA	Frozen model-transport result	Immutable test set used once and retired
Figshare 29925404	Primary 36/36 LST-G matched-pair score	Patient matched descriptive effect (n = 2 matched donors)	Not estimable	NA	NA	Canonical 36-gene endpoint not estimable	Primary status precedes sensitivities; hereditary status unknown; Stage 11B-specific Table 2 inclusion
E-GEAD-622 Stage 10F	Patient-level spatial endpoint	Patient gate (n = 1 eligible patient of minimum 3)	Not estimable	NA	NA	Patient-level spatial analysis not estimable	Not-estimable due to ROI eligibility shortfall; not a negative biological effect
GSE41657; GSE100179; GSE8671	Additional within-accession SD scaling sensitivity	Accession-level meta-analysis (n = 3 accession effects)	1.951 (1.542 to 2.361)	0.0024	NA	Additional scale sensitivity; not a new primary result	k = 3; two cohorts have unresolved patient identity; pooled estimate is sensitivity-only and not a pooled-patient analysis; I2 = 67.2%

M02 showed a selection-conditioned descriptive association in the donors used to select it

Stage-label-blind covariance clustering produced multiple modules; subsequent stage-aware, paired-FAP and leave-one-donor-out filters reduced them to six exploratory candidates (Fig. 2b). M02 contained 36 genes and was selected for cross-dataset assessment. The module included six genes in each of four broad categories—WNT-linked/epithelial-state, receptor/signal-transduction, extracellular-matrix/adhesion and long-noncoding/poorly characterized genes—plus smaller transcriptional, metabolic, membrane and other groups (Fig. 2c; Table S3). No functional term reached adjusted $P < 0.05$ against the 1,000-gene Stem/progenitor module background. INPP5D was the only explicitly immune-linked member.

All four paired FAP donors had positive 36-gene adenoma-minus-normal differences in all epithelium (0.413–1.917) and Stem/progenitor profiles (0.219–1.286) (Fig. 3a). Omitting any single gene left all four Stem/progenitor donor differences positive; mean paired differences ranged from 0.746 to 0.810 across leave-one-gene-out scores (Fig. 3b). These values describe a selection-conditioned association in the donors that helped select the module; they are not independent evidence. Historical paired summaries—1.276 (95% CI 0.243–2.310; P = 0.0293; FDR = 0.0586) in all epithelium and 0.778 (95% CI 0.066–1.491; P = 0.0402; FDR = 0.0482) in Stem/progenitor profiles—are retained in Table S4 and explicitly marked selection conditioned.

Random-gene-set resampling contextualized but did not remove selection dependence

In the expression- and variability-matched reference, the observed M02 mean exceeded all 10,000 random-module effects in both all epithelium and Stem/progenitor profiles. For all epithelium, the observed effect was 1.276, whereas the random-module median was 0.209 and its 2.5th–97.5th percentile interval was − 0.143 to 0.562. For Stem/progenitor profiles, the observed effect was 0.778 versus a random-module median of 0.185 and interval of − 0.139 to 0.502. The plus-one upper-tail fraction was 1/10,001 (1.0×10 − 4) in each compartment (Fig. S10; Table S11).

The unmatched reference gave the same qualitative ordering: M02 exceeded all 10,000 all-epithelial modules and 9,999 of 10,000 Stem/progenitor modules; the corresponding upper-tail fractions were 1.0×10 − 4 and 2.0×10 − 4. Omitting each FAP donor in turn left mean effects positive, ranging from 1.063 to 1.564 in all epithelium and from 0.609 to 0.965 in Stem/progenitor profiles. These observations show that the selected score was not typical of random 36-gene sets under the stated references. They do not correct selection bias, because paired FAP direction and leave-one-donor-out behavior helped determine which module was retained.

Three matched GSE161277 donors showed concordant but underpowered differences

The three GSE161277 donors had M02 adenoma-minus-normal differences of 2.012, 1.181 and 1.743 (Fig. 4a). The mean was 1.645 (95% CI 0.592–2.698; nominal P = 0.0214), but the FDR across six candidate modules was 0.128. This single three-donor dataset, which was not used during M02 generation, was directional and underpowered, not evidence of a general sporadic effect: hereditary status was insufficient for FAP-specific or sporadic-specific estimation, and three donors cannot characterize population heterogeneity.

Separately processed bulk-tissue datasets showed associations limited by unit resolution and heterogeneity

Original-scale adenoma-versus-normal module-score differences were positive in GSE41657 (2.420, 95% CI 2.215–2.626; 63 sample-scoped arrays), GSE100179 (2.080, 95% CI 1.832–2.327; 40 sample-scoped arrays) and GSE8671 (1.846, 95% CI 1.666–2.026; 32 verified pairs) (Fig. 4b). The accession-level random-effects summary was 2.113 (95% CI 1.389–2.837), with I^2 = 87.6% and a prediction interval of 0.725–3.500. Because score scales differed by platform and two datasets lacked patient identifiers, this synthesis is not a pooled patient effect.

After division by each accession's normal/adenoma score standard deviation, the standardized effects were 2.136, 1.862 and 1.846. The REML/Knapp–Hartung summary was 1.951 SD units (95% CI 1.542–2.361), with I^2 = 67.2% and a prediction interval of 1.234–2.668; Paule–Mandel results were nearly identical (Fig. 4c; Table S6). Leave-one-accession-out estimates remained positive but were highly uncertain with only two accessions. These analyses show a cross-platform bulk-tissue association, not epithelial specificity or independent patient-level reproducibility in all three cohorts.

Transport and spatial branches exposed failure modes rather than additional support

The locked stool-RNA models degenerated. Adenoma-versus-normal and cancer-versus-normal LASSO models selected no non-intercept feature and assigned one unique test probability, producing AUC = 0.500 for both endpoints. The composite neoplasia-versus-normal model selected B4GALT7 alone, produced 61 unique probabilities among 65 test samples and had AUC = 0.530 (95% CI 0.380–0.679) (Fig. S7; Table S7). This was a negative assessment of the project-locked feature-selection pipeline, not a direct test of the 36-gene M02 score.

In Figshare 29925404, 33 canonical genes mapped, whereas the 36/36 endpoint required by that branch was unavailable (Fig. S8). The reduced-coverage scores in two matched LST-G donors remained descriptive because the canonical endpoint was not estimable, FAP/APC status was unknown and LST-G is not equivalent to ordinary sporadic adenoma.

Spatial pseudo-spot analyses established only simulation-based score recoverability. In E-GEAD-622, the absence of an independent pathologist reference and failure of three of four cases at the frozen region-registration gate prevented patient-level analysis (Fig. S9). In the sole eligible case, the primary score was negative, most sensitivities were also negative and gene-z was effectively zero; the result was therefore method dependent. No spatial population, biological or causal analysis was performed.

Discussion

This secondary analysis illustrates a common trajectory in public-data research: a valid donor-level screen can be negative while a later post-hoc module appears coherent in selected contexts. The relevant question is not whether the module can be made to look positive, but whether its provenance, estimability and cross-dataset limits are explicit enough for independent testing.

Limitations and boundary conditions

The strongest limitation is post-hoc selection dependence. M02 was not an outcome of the primary gene-level screen. Its covariance structure was derived without stage labels, but the candidate filter used stage effects, paired FAP direction and leave-one-donor-out behavior. The same four FAP donors were then summarized again. Consequently, the internal effect sizes, confidence intervals and FDR values are conditional on selection and likely optimistic. Their appropriate role is to document a selection-conditioned exploratory association, not provide independent inferential support.

The random-gene-set audit addresses arbitrariness, not independence. M02 occupied the extreme upper tail of both matched and unmatched 36-gene references, and its leave-one-donor-out means remained positive. However, this pattern is partly expected because paired FAP direction and donor-omission behavior entered the candidate filter. The empirical tail fractions therefore have no valid post-selection P-value interpretation and do not raise the evidence level.

FAP-stage confounding further restricts interpretation. GSE201348 did not contain a pathology-defined sporadic-adenoma design cell adequate for a sporadic normal-to-adenoma effect, and a stage-by-FAP interaction was not estimable. M02 should therefore be described as a FAP/adenoma-associated exploratory epithelial stem/progenitor module, not a general colorectal initiation program. GSE161277 adds a concordant observation in three matched donors, but its FDR of 0.128, very small n and insufficient hereditary stratification prevent broad generalization.

The module audit also tempers biological interpretation. M02 spans WNT-linked genes, extracellular/adhesion genes, signaling genes, transcriptional regulators and six noncoding or poorly characterized loci, but no functional category was significantly enriched against the relevant module-construction background. Three module genes overlapped state-annotation evidence. This does not invalidate the score, but it means that Stem/progenitor localization is not fully independent of the labels used to define that compartment. INPP5D creates a plausible immune-contamination sensitivity, although removing it did not reverse the four FAP donor descriptions.

Bulk-tissue associations were the most numerically consistent external observations. They remain limited in three ways. First, bulk scores are not epithelial-specific. Second, patient identity was unresolved in two accessions, so arrays cannot be assumed to be independent patients. Third, only three accession effects were available. Knapp-Hartung and alternative heterogeneity estimates address some small-k uncertainty but cannot manufacture cohort diversity or patient metadata [28]. The reduced heterogeneity after within-cohort scaling shows that score scale contributed to the original I2, yet substantial residual heterogeneity remained.

The technically failed and non-estimable branches are informative. The stool models collapsed to intercept-only or one-feature solutions, but this branch did not directly test M02. Because the test set was used once and retired, a direct M02 assessment would require a new independent dataset. In the LST-G branch, the 36-gene endpoint was unavailable, and the historical “29/36” label obscured that all 33 mapped genes were averaged after meeting a 29-gene minimum threshold. The spatial branch lacked independent pathology adjudication and retained one eligible case; its method-dependent description cannot be converted into population evidence.

Additional limitations arise from uneven public metadata, platform-specific preprocessing and incomplete anatomical or hereditary information. Candidate selection and evaluation were not registered in a public immutable protocol before analysis; repository timestamps establish file order only. Automated annotation was auxiliary, but state labels still depend on marker choices and several clusters were donor dominated. Within-accession bulk standardization was a later sensitivity analysis introduced after the original-scale synthesis; it did not alter module membership. No wet-laboratory experiment was performed, so causal or clinical-utility claims remain outside the evidence.

Implications for independent evaluation

A decisive next study would require pathology-defined sporadic adenomas and matched normal mucosa, recorded hereditary status, raw counts, epithelial states defined without M02 genes and enough independent donors for patient-level estimation. The 36-gene score, mapping rule and analysis code should be fixed publicly before outcome analysis. Until such data exist, M02 should remain a candidate for prospective assessment rather than an independently supported signature.

Conclusions

The project-recorded donor-level normal-to-adenoma screen was negative. M02 emerged only after a post-hoc workflow that used paired FAP direction, making its four-donor summaries selection conditioned and unsuitable for independent inference. Three matched GSE161277 donors and three bulk-tissue accessions showed concordant module-score differences, but small n, unresolved patient identities, tissue composition and heterogeneity prevent broad inference. Stool, LST-G and spatial branches did not provide independent support. The reusable output is a fixed exploratory candidate and an auditable account of where cross-dataset evaluation was feasible, technically failed or was not estimable.

Abbreviations

APC
adenomatous polyposis coli
AUC
area under the receiver-operating-characteristic curve
CI
confidence interval
FAP
familial adenomatous polyposis
FDR
false-discovery rate
LST-G
granular-type laterally spreading tumor
M02
Stem_progenitor_SB_M02
ROI
region of interest

TMM
trimmed mean of M-values
UMI
unique molecular identifier.

Declarations

Ethics approval and consent to participate

Ethics approval: Not applicable. The present study comprised only secondary analyses of publicly available, de-identified repository datasets. It involved no prospective recruitment, intervention, new specimen collection, animal experiments or access to identifiable private information; formal ethics committee approval was therefore not required for the present analysis. Ethical approval and consent for the original studies were the responsibility of, and were reported by, the respective data generators.

Consent to participate:

Not applicable. No individuals participated directly in the present study, and no additional informed consent was required for secondary analysis of the publicly available, de-identified data.

Consent for publication

Not applicable. The manuscript contains no directly identifying participant information.

Competing interests

The authors declare that they have no competing interests.

Use of AI-assisted tools

During manuscript preparation, the authors used OpenAI ChatGPT for English-language refinement. These tools did not generate or modify the frozen analytical results. All outputs were reviewed and verified by the authors, who take full responsibility for the manuscript.

Authors' information

Yidi Zhao, Yunfei Jia and Hehe Liao contributed equally. Pingyi Wang is the corresponding author (ORCID 0009-0008-3831-9013).

Funding

This work was supported by the Xizang Natural Science Foundation (XZ202501ZR0034, XZ202601ZR0076, XZ202601ZR0008 and XZ202601ZR0019); the Shaanxi Provincial Department of Education General Special Scientific Research Program (24JK0687 and 25JK0702); and the Scientific Research Project of Xizang Minzu University (25MD10, 26MDQ06 and 26MDQ09). The funders had no role in study design, data analysis, interpretation, manuscript preparation or the decision to submit.

Author Contribution

Yidi Zhao, Yunfei Jia, Hehe Liao and Pingyi Wang conceptualized the study. Yidi Zhao, Yunfei Jia and Hehe Liao performed data curation, formal analysis, software implementation, visualization and initial drafting. Yujie Ma, Zifei Zhang, Yao Zheng, Jianan Zhang and Haidie Xie contributed metadata curation, literature verification, quality-control review and interpretation. Pingyi Wang contributed methodology, supervision, project administration and funding acquisition. Yidi Zhao drafted the manuscript; Yunfei Jia, Hehe Liao and Pingyi Wang critically reviewed and edited it. All authors read and approved the final manuscript. Yidi Zhao, Yunfei Jia and Hehe Liao contributed equally.

Acknowledgement

The authors thank the investigators who generated and publicly shared the datasets used in this study.

Availability of data and materials

The datasets analyzed or formally audited in this study were obtained from public repositories. National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO): GSE201348 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE201348); GSE161277 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161277); GSE132465 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132465);

GSE41657 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41657); GSE100179 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100179); GSE8671 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8671); GSE99573 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99573); GSE200997 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE200997); GSE226997 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE226997).

The Cancer Genome Atlas Colon Adenocarcinoma project (TCGA-COAD) was accessed through the National Cancer Institute (NCI) Genomic Data Commons (GDC; https://portal.gdc.cancer.gov/projects/TCGA-COAD). Spatial records were obtained from the DNA Data Bank of Japan (DDBJ) Genomic Expression Archive (GEA): E-GEAD-622 (https://ddbj.nig.ac.jp/public/ddbj_database/gea/experiment/E-GEAD-000/E-GEAD-622/), E-GEAD-619 (https://ddbj.nig.ac.jp/public/ddbj_database/gea/experiment/E-GEAD-000/E-GEAD-619/) and E-GEAD-579 (https://ddbj.nig.ac.jp/public/ddbj_database/gea/experiment/E-GEAD-000/E-GEAD-579/). The LST-G count matrices and metadata are available from Figshare (record 29925404; https://doi.org/10.6084/m9.figshare.29925404).

The analytical roles of these datasets are stated in the Methods, Table 1 and Supplementary Table S1. Accession-specific links, download dates and checksums are provided in Supplementary Information 4.

The versioned baseline reproducibility package is available at https://github.com/Wang-Pingyi/CRC-M02-reproducibility/tree/v1.0.1 and archived as Zenodo v1.0.1 at https://doi.org/10.5281/zenodo.21743195 (concept DOI https://doi.org/10.5281/zenodo.21743194). Software is licensed under MIT; original small source data and documentation are licensed under CC BY 4.0. Manuscript-specific scripts and source data are included in Supplementary Information 4. Raw FASTQ, CEL, H5/H5AD, RDS and large count matrices are not redistributed. The release supports provenance checks and reconstruction of key small source data; it is not a complete automated rerun of all large bioinformatic analyses.

References

1. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990;61:759–67. https://doi.org/10.1016/0092-8674(90)90186-I.
2. Half E, Bercovich D, Rozen P. Familial adenomatous polyposis. Orphanet J Rare Dis. 2009;4:22. https://doi.org/10.1186/1750-1172-4-22.
3. Becker WR, Nevins SA, Chen DC, et al. Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. Nat Genet. 2022;54:985–95. https://doi.org/10.1038/s41588-022-01088-x.
4. Zheng X, Song J, Yu C, et al. Single-cell transcriptomic profiling unravels the adenoma-initiation role of protein tyrosine kinases during colorectal tumorigenesis. Signal Transduct Target Ther. 2022;7:60. https://doi.org/10.1038/s41392-022-00881-8.
5. Lee HO, Hong Y, Etlioglu HE, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet. 2020;52:594–603. https://doi.org/10.1038/s41588-020-0636-z.
6. Squair JW, Gautier M, Kathe C, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12:5692. https://doi.org/10.1038/s41467-021-25960-2.
7. Zimmerman KD, Espeland MA, Langefeld CD. A practical solution to pseudoreplication bias in single-cell studies. Nat Commun. 2021;12:738. https://doi.org/10.1038/s41467-021-21038-1.
8. Crowell HL, Soneson C, Germain PL, et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nat Commun. 2020;11:6077. https://doi.org/10.1038/s41467-020-19894-4.
9. Kalmar A, Nagy ZB, Galamb O, et al. Genome-wide expression profiling in colorectal cancer focusing on lncRNAs in the adenoma-carcinoma transition. BMC Cancer. 2019;19:1059. https://doi.org/10.1186/s12885-019-6180-5.
10. Sabates-Bellver J, Van der Flier LG, de Palo M, et al. Transcriptome profile of human colorectal adenomas. Mol Cancer Res. 2007;5:1263–75. https://doi.org/10.1158/1541-7786.MCR-07-0267.
11. Barnell EK, Kang Y, Wurtzler EM, et al. Noninvasive detection of high-risk adenomas using stool-derived eukaryotic RNA sequences as biomarkers. Gastroenterology. 2019;157:884. https://doi.org/10.1053/j.gastro.2019.05.058. 7.e3.
12. Gong Y, Zhang Y, Liu X, et al. Single-cell RNA profiling of colorectal granular-type laterally spreading tumor uncovers progression trajectory toward carcinoma and transcriptional signatures favoring lateral morphogenesis. Front Oncol. 2025;15:1552841. https://doi.org/10.3389/fonc.2025.1552841.
13. KC F. Expression matrices of colorectal laterally spreading tumours. Figshare. 2025. https://doi.org/10.6084/m9.figshare.29925404
14. Hashimoto M, Kojima Y, Sakamoto T, et al. Spatial and single-cell colocalisation analysis reveals MDK-mediated immunosuppressive environment with regulatory T cells in colorectal carcinogenesis. EBioMedicine. 2024;103:105102. https://doi.org/10.1016/j.ebiom.2024.105102.
15. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7. https://doi.org/10.1038/nature11252.
16. Khaliq AM, Erdogan C, Kurt Z, et al. Refining colorectal cancer classification and clinical stratification through a single-cell atlas. Genome Biol. 2022;23:113. https://doi.org/10.1186/s13059-022-02677-z.
17. Park SS, Lee YK, Choi YW, et al. Cellular senescence is associated with the spatial evolution toward a higher metastatic phenotype in colorectal cancer. Cell Rep. 2024;43:113912. https://doi.org/10.1016/j.celrep.2024.113912.
18. Germain PL, Lun A, Garcia Meixide C, Macnair W, Robinson MD. Doublet identification in single-cell sequencing data using scDblFinder. F1000Res. 2022;10:979. https://doi.org/10.12688/f1000research.73600.2.

19. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573–87. https://doi.org/10.1016/j.cell.2021.04.048. .e29.
20. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96. https://doi.org/10.1038/s41592-019-0619-0.
21. Gao R, Bai S, Henderson YC, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021;39:599–608. https://doi.org/10.1038/s41587-020-00795-2.
22. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40. https://doi.org/10.1093/bioinformatics/btp616.
23. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25. https://doi.org/10.1186/gb-2010-11-3-r25.
24. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47. https://doi.org/10.1093/nar/gkv007.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
26. Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. 2023;51:W207–12. https://doi.org/10.1093/nar/gkad347.
27. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36:1–48. https://doi.org/10.18637/jss.v036.i03.
28. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol. 2015;15:99. https://doi.org/10.1186/s12874-015-0091-1.
29. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1–22. https://doi.org/10.18637/jss.v033.i01.
30. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S + to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77. https://doi.org/10.1186/1471-2105-12-77.
31. Andreatta M, Carmona SJ. UCell: robust and scalable single-cell gene signature scoring. Comput Struct Biotechnol J. 2021;19:3796–8. https://doi.org/10.1016/j.csbj.2021.06.043.
32. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013;9:e1003285. https://doi.org/10.1371/journal.pcbi.1003285.
33. Zhao Y, Jia Y, Liao H, Ma Y, Zhang Z, Zheng Y, Zhang J, Xie H, Wang P. CRC M02 cross-dataset transcriptomic reproducibility package. Zenodo. 2026. https://doi.org/10.5281/zenodo.21743195.

Figures

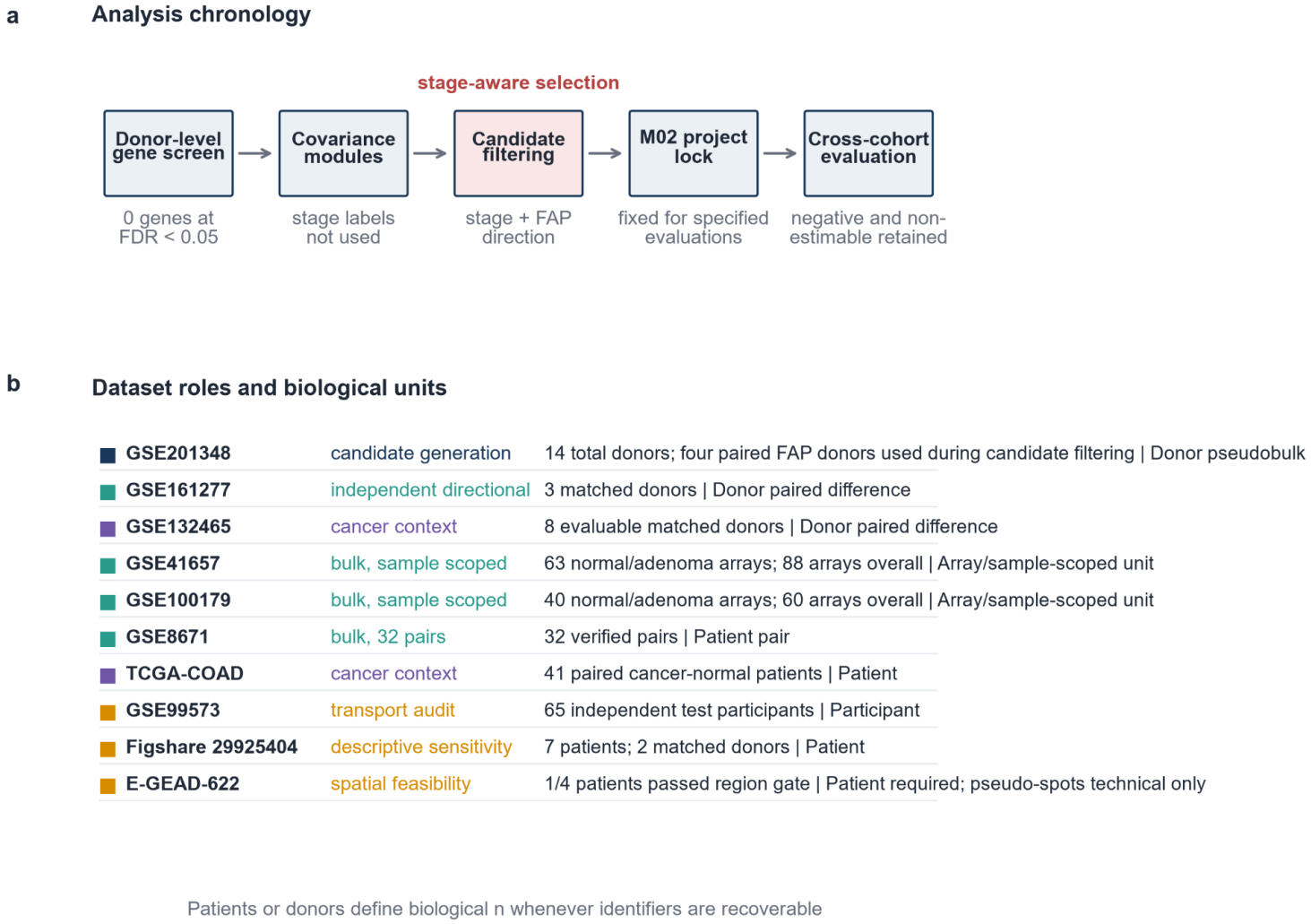


Figure 1

Study design, dataset roles and biological units

Chronology and dataset-role matrix for the negative donor-level screen, post-hoc module generation and subsequent or supplementary assessments. The module-candidate filter used paired FAP direction after stage-label-blind covariance clustering. Patient or donor is the biological unit where identifiers are recoverable; GSE41657 and GSE100179 remain sample-scoped

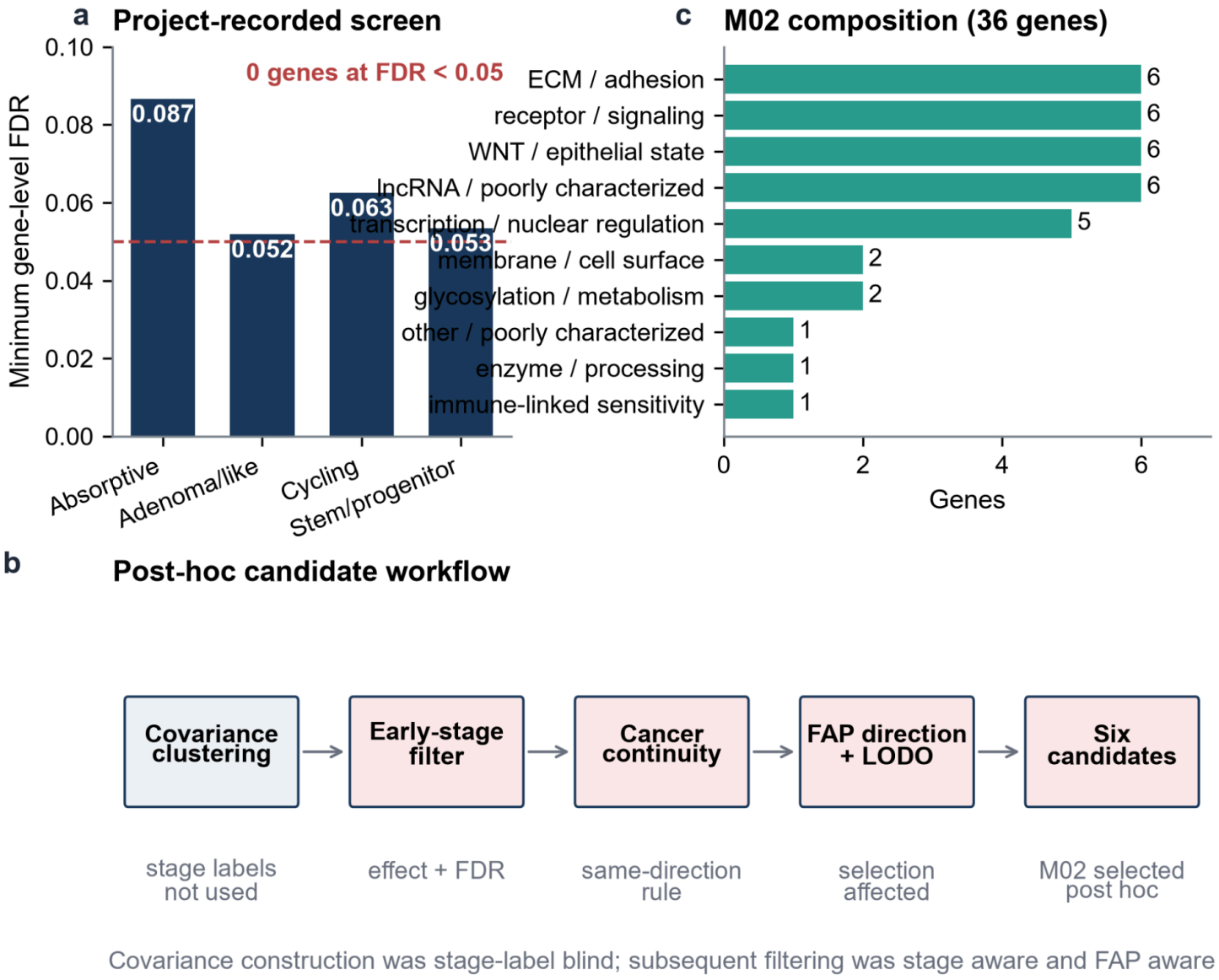


Figure 2

Negative primary screen and post-hoc M02 construction

(a) Minimum normal-to-adenoma FDR in the four eligible epithelial states; no gene reached FDR<0.05. (b) Separation of stage-label-blind covariance clustering from stage-aware and FAP-direction candidate filtering. (c) Descriptive composition of the 36-gene M02 module. Functional enrichment against the 1,000-gene Stem/progenitor module background yielded no adjusted term below 0.05

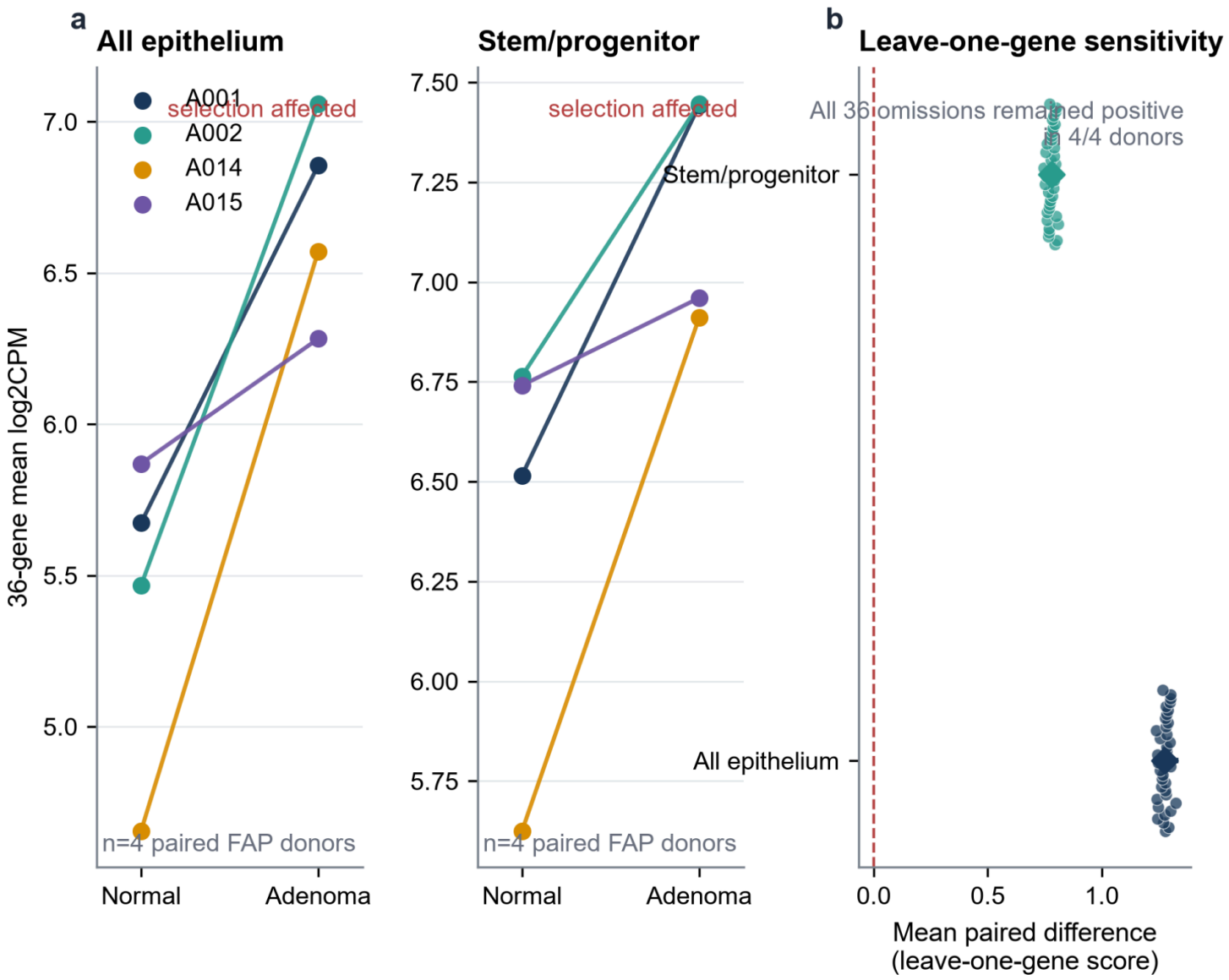


Figure 3

Selection-conditioned descriptive scores in four FAP patients used during module development

(a) Patient-level 36-gene scores in paired normal and adenoma pseudobulks for all epithelium and Stem/progenitor cells. (b) Leave-one-gene-out paired differences. The same four donors contributed to the candidate-selection rule, so no displayed quantity is an independent inferential test

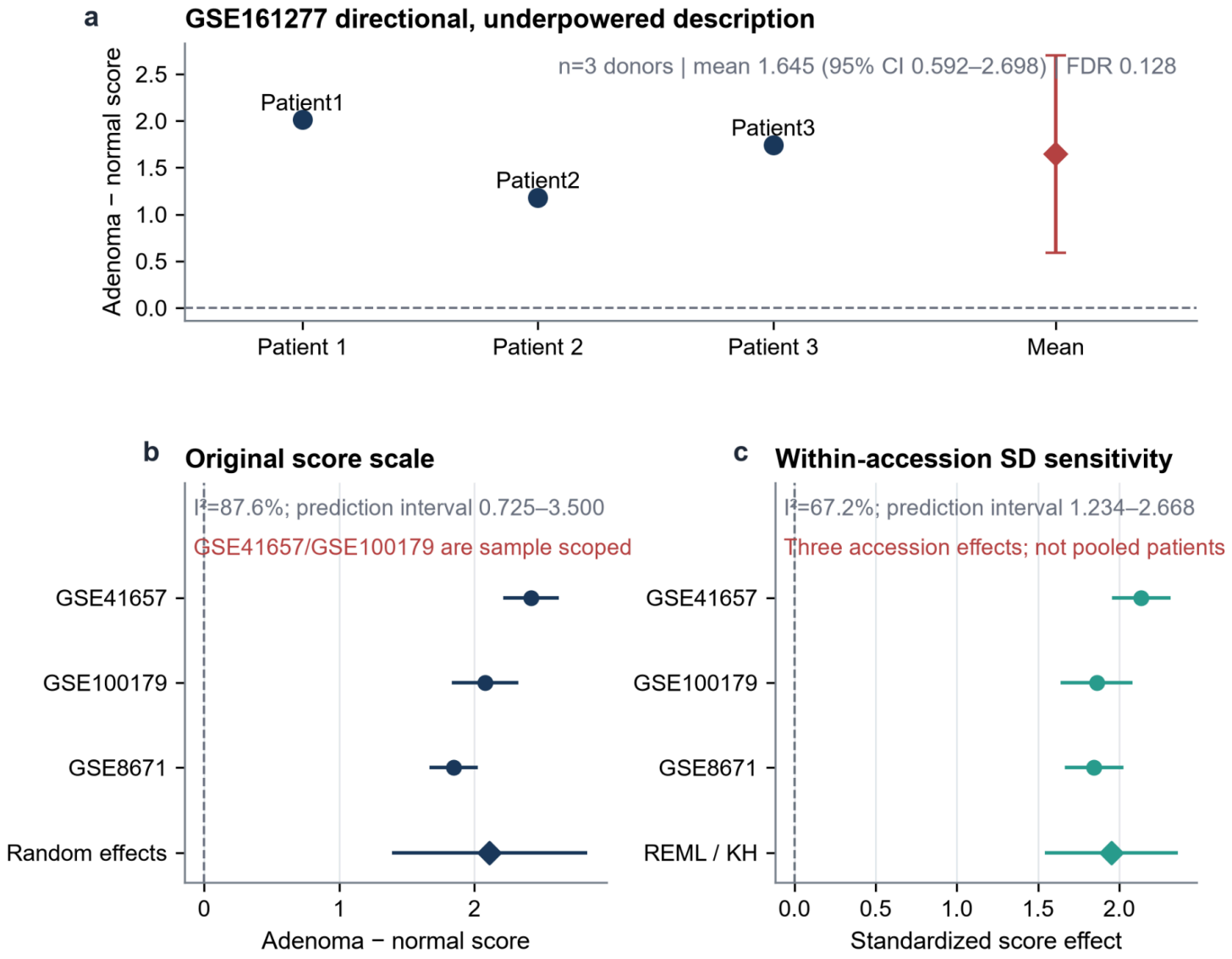


Figure 4

Three-donor single-cell description and bulk-tissue associations

(a) GSE161277 patient-specific normal-to-adenoma differences for three matched donors; the mean effect was 1.645 (95% CI 0.592 to 2.698; FDR 0.128). (b) Original-scale accession-specific bulk effects and random-effects summary. (c) Within-accession standardized sensitivity. GSE41657 and GSE100179 are sample-scoped; GSE8671 contains 32 verified pairs. Meta-analyses combine three accession effects, not pooled patients

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation1Methods.docx
- SupplementaryInformation2Figures.pdf
- SupplementaryInformation3Tables.xlsx
- SupplementaryInformation4SourceData.zip