

Supplementary methods and supplementary item legends

Article title: Tissue composition quantitatively reproduces bulk transcriptomic differences between paired primary gastric tumours and solid peritoneal metastases

Journal: Functional & Integrative Genomics

Authors: Chao Huang¹, Qin Yu², Zheng-Ming Zhu¹, Yinfei Kong³

Affiliations: ¹ Department of Gastrointestinal Surgery, The Second Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi Province, China; ² School of Management, University of Science and Technology of China, Hefei 230026, China; ³ Department of Information Systems and Decision Sciences, College of Business and Economics, California State University Fullerton, 800 N State College Blvd, Fullerton, CA 92831, USA

Corresponding authors: Zheng-Ming Zhu (zzm8654@163.com) and Yinfei Kong (yikong@fullerton.edu)

S1. Specimen and patient definitions

One row per biological specimen was built for every source from GEO SOFT records, the supplementary tables of the corresponding publications, and two small curated maps stored with the analysis code, each carrying a source locator. Solid peritoneal metastasis was recorded separately from omental metastasis, malignant ascites, peritoneal lavage and normal peritoneum. Ascites and lavage were never merged with solid tissue at any stage.

The analysis unit was the patient. Each patient contributed one primary value, one metastasis value and one difference. In the primary cohort, one patient provided two pairs — a pre-treatment pair and a pair obtained at recurrence one year after treatment — and the protocol assigned the earliest eligible pair to the primary analysis. The 18 pair units described in the source publication therefore correspond to 17 independent patients, and this is the denominator used throughout.

GSE234129 was examined for overlap with GSE239676 by comparing 16-nucleotide cell barcodes per sample. All 17 libraries matched their expected counterpart with a shared fraction between 0.774 and 0.956, against a next-best match of 0.014 to 0.031, identifying them as the same libraries. GSE234129 therefore contributes no additional patients and was counted once.

S2. Expression matrices and their actual content

The GSE314812 supplementary file was identified as gene-level RSEM expected counts rather than TPM: values are non-negative with a non-integer fraction consistent with multi-mapping allocation, column sums of 1.9 to 3.2×10^7 rather than 10^6 , and identifiers are unversioned Ensembl GRCh37 genes. The authors' TPM matrix is not public and was not reconstructed. GSE237876 provides transcript-level Salmon output per specimen; transcript estimates were summed to gene level using an Ensembl GRCh37 release 75 transcript-to-gene map, and gene-level TPM was obtained by summing transcript TPM. All 57,773 GSE314812 gene identifiers and all 196,501 GSE237876 transcript identifiers mapped to that annotation.

S3. Single-cell processing

Each sample was processed independently, with no cross-sample integration used for any quantitative comparison. Symbols were harmonised to approved HGNC symbols through Ensembl identifiers, approved symbols, unique previous symbols and unique aliases, in that order, retaining the original identifier and the route taken. Quality control was per sample: a minimum of 200 detected genes, a mitochondrial fraction threshold derived from the sample's own robust distribution with an absolute ceiling of 25 per cent, and robust outlier removal on log counts and log genes. Doublets were identified per sample and excluded. Clustering used a graph-based method on highly variable genes; each cluster was assigned the lineage with the highest mean marker score when it exceeded the runner-up by a fixed margin, and was otherwise labelled uncertain. The lineage marker sets are fixed in the analysis code and share no genes with any scored programme.

Malignant epithelium was identified by expression-based copy-number scoring within each sample, using that sample's own non-epithelial cells as reference, with a signal threshold set at the 95th percentile of reference cells and a requirement for positive correlation with the sample's inferred profile. Epithelial cells in normal-tissue samples were never labelled malignant, and in the calibration source no cell in a normal gastric sample met the malignant criteria, which provides an internal specificity check.

S4. Programme definitions and the coverage rule

Definitions were derived from primary source files rather than transcribed. The primary programme is the core high-relapse-cell programme of the cited colorectal study, taken from its supplementary table. Control programmes came from the Molecular Signatures Database release 2026.1.Hs, the ESTIMATE supplementary data, and the pan-cancer malignant meta-programme collection. The negative control comprises consensus YAP/TAZ-TEAD target genes appearing in at least two independent published signatures and restricted to genes that are not epithelium-dominant in the Human Protein Atlas.

Host-compartment sets required a compartment maximum of at least 10 nCPM and at least threefold the maximum of every other compartment, taking the 40 highest-expressing genes, after excluding every gene belonging to any scored programme. This exclusion prevents circularity between the endpoint and its composition covariates.

Scores are within-sample equal-weight mean gene ranks on a common gene universe defined by expression alone: a gene enters when it reaches 1 count per million in at least 25 per cent of the samples on that platform, a rule that uses no site or group information. A programme is not evaluable on a platform when fewer than 80 per cent of its defined genes, or fewer than ten genes, are in that universe. In the primary cohort the universe contained 15,888 genes and the primary programme had 99 of 100 genes measured. The mesothelium compartment reached 31 of 40 and was declared not evaluable; the rule was not relaxed after the primary result was known.

S5. Measurement-model calibration (post hoc)

This calibration was added after the primary result was known and is labelled post hoc throughout. Using calibration donors only, pseudo-bulk mixtures of 300 cells were assembled at known fractions of 0.05 to 0.60 along two axes — malignant fraction and total epithelial fraction — with ten replicates per fraction per donor. Frozen programmes were scored on each mixture with the same rule used for real specimens. The slope of score on known fraction was estimated with a donor term. The observed tissue-level difference was divided by that slope to express it as the cell-content difference that would reproduce it. The relationship saturates at high epithelial content, so these conversions are approximations; both axes are reported.

S6. Post-hoc robustness analyses

Three alternative algorithms scored the same measured genes as the frozen primary programme. The *singscore* implementation first ranked all genes within each sample and computed the centred single-sample score for the programme. The mean-log-expression implementation averaged log2 counts per million plus one. The gene-standardised implementation centred and scaled each programme gene across the full site-blind sample set before averaging. For each method, the same 17 eligible patient pairs were reduced to one metastasis-minus-primary difference per patient and analysed with the paired procedures used for the primary endpoint. Cohen's d_z was calculated as the mean divided by the standard deviation of patient differences; its interval used 5,000 patient bootstrap resamples.

The independent epithelial panel contained EPCAM, KRT8, KRT18 and KRT19. All four genes were measured, none occurred in the primary programme, and no outcome or site contrast was used to select them. Mean-rank and mean-log-expression scores were evaluated in the same 17 pairs.

Published histological purity in GSE237876 was analysed across all 66 specimens from 14 patients. The main model was $\text{HRC score} \sim \text{purity} + \text{patient}$, so the purity coefficient used only within-patient variation. Its interval was estimated by resampling patients with replacement 5,000 times and retaining every specimen from each sampled patient. As a sensitivity analysis, multiple regions were first averaged within patient and lesion role. A separate five-patient calculation compared eligible primary and solid peritoneal samples; its null result is reported because it addresses site-level purity rather than the across-lesion measurement association.

S7. Analyses that were pre-specified but could not be run

Three pre-specified items produced no result and are recorded with their reason rather than omitted.

The direct site contrast, comparing peritoneal with another metastatic site relative to the primary, was not estimable: no patient in the primary cohort has a primary tumour, a solid peritoneal metastasis and another metastatic site with the required pairing, the omental and lymph-node specimens coming from patients without that combination. No peritoneum-specific claim is therefore made anywhere in the paper.

The fluid-context contrast in malignant cells from ascites could not be evaluated because the public matrix for that study retains only 8,630 genes and the primary programme fell below the frozen coverage threshold on that platform.

The clinical peritoneal-recurrence module was switched off by protocol before analysis. The question addressed here concerns established lesions rather than future risk, and the available cohort has roughly forty to fifty first peritoneal-recurrence events with no second time-to-event cohort, which cannot support a prediction claim.

S8. Supplementary table legends

Supplementary Table S1. Sample crosswalk. One row per biological specimen across all sources, with patient, site, specimen context, lesion role, pairing evidence, therapy, preservation, assay, matrix unit, metadata source and locator, identity confidence, eligibility and exclusion reason.

Supplementary Table S2. Overlap register. Source pairs examined for shared patients or libraries, the evidence used, the handling applied and whether the sources can be treated as independent.

Supplementary Table S3. Frozen programme definitions. Every programme with its role, source, version, dataset overlap status, original and mapped gene symbols, mapping route, Human Protein Atlas compartment specificity, direction, weight and scoring method.

Supplementary Table S4. Programme coverage by platform. Defined and measured gene counts per programme per platform, with the evaluability decision under the frozen rule.

Supplementary Table S5. Matrix-type audit. Dimensions, missing values, negative values, zero fraction, non-integer fraction, column-sum distribution, duplicate keys and the resulting judgement of what each expression file actually contains.

Supplementary Table S6. Patient-level differences. Per-patient differences for every evaluable programme and compartment in the primary cohort, with preservation change.

Supplementary Table S7. Hypothesis registry. Every pre-specified hypothesis with its tier, source, test, family, family size, estimability, P and q values, and the reason where it was not executed.

Supplementary Table S8. Claims ledger. Each claim with its evidence tier, estimand, supporting result file, patient numbers, effect, interval, measurement level, contradictory evidence, main alternative explanation, permitted wording and prohibited overclaim.

Supplementary Table S9. Number trace. Every study number appearing in the manuscript with its result file, object key, denominator and unit.

Supplementary Table S10. Reference verification. Every cited work with PMID, DOI, bibliographic fields, publication types, retraction and erratum status, the claim it supports and the verification date.

Supplementary Table S11. Scoring-method robustness. Paired effects for the frozen HRC programme under mean-rank, *singscore*, mean-log-expression and gene-standardised scoring, including patient-bootstrap intervals for the standardised effect.

Supplementary Table S12. Independent epithelial-marker check. Paired effects for EPCAM/KRT8/KRT18/KRT19 under mean-rank and mean-log-expression scoring; the panel shares no genes with HRC.

Supplementary Table S13. External histological-purity association. Raw and within-patient correlations, patient fixed-effects estimates, patient-cluster bootstrap interval, patient-site sensitivity analysis and the separate five-patient site contrast.

Supplementary Table S14. Paired external purity differences. Patient-level histological-purity and frozen-HRC differences for the five eligible external pairs.

Supplementary Figure S1 Adjustment for composition and technical covariates. The model intercept is the estimated HRC difference remaining after adjustment. Each model adds one composition term, except the categorical preservation-change model. Mesothelium was not evaluable under the frozen coverage rule. Error bars are 95% confidence intervals; 17 paired patients. The figure was generated from the cited public data and archived analysis output by the archived code; the limited assistance with selected code refinement is described in the main manuscript's Methods section.

