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.

An Outcome-Independent Analytical Framework Reveals Systemic Discordance Between a Continuous Neural-Mesenchymal Programme and Consensus Molecular Subtyping in Colorectal Cancer: A Multi-Cohort Transcriptomic Study

Anirudhha Roy

anirudhha.roy.kol@hotmail.com

JIS University https://orcid.org/0009-0002-6194-5545

Research Article

Keywords: colorectal cancer, epithelial–mesenchymal transition, consensus molecular subtype, neural-mesenchymal programme, biomarker, transcriptomics, multi-cohort validation

Posted Date: September 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11096455/v1

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

Additional Declarations: The authors declare no competing interests.

An Outcome-Independent Analytical Framework Reveals Systemic Discordance Between a Continuous Neural-Mesenchymal Programme and Consensus Molecular Subtyping in Colorectal Cancer: A Multi-Cohort Transcriptomic Study

Anirudhha Roy

JIS School of Medical Science and Research, JIS University, Kolkata, West Bengal, India

Corresponding author: Anirudhha Roy, JIS School of Medical Science and Research, JIS University, Kolkata, West Bengal, India.

Keywords

colorectal cancer; epithelial–mesenchymal transition; consensus molecular subtype; neural-mesenchymal programme; biomarker; transcriptomics; multi-cohort validation

Highlights

• A four-gene GAP43/VIM/ZEB1/TWIST1 module was fixed by co-expression alone, before CMS or outcome data were consulted.

• The resulting NEM score captures 96.3% of CMS4 but also flags EMT activity outside CMS4.

• This asymmetric discordance reproduces across nine independent GEO cohorts (n≈3,742).

• A Human Protein Atlas cross-check shows no module gene is consistently, independently prognostic across cohorts (one discordant exception).

• GAP43 protein localises to peritumoral nerve, not tumour epithelium, which informs module interpretation.

Abstract

An outcome-independent strategy is presented for deriving a continuous programme score from co-expression structure alone, then testing its agreement with an established classifier across independent cohorts. Consensus Molecular Subtyping (CMS) groups colorectal cancer (CRC) transcriptomes into discrete categories; CMS4 denotes mesenchymal, stromal-rich disease. The question addressed is whether a continuous neural-mesenchymal (NEM) programme, built from a curated 14-gene panel under a pre-specified correlation threshold, tracks CMS4 assignment without being reducible to it. In a TCGA-COAD discovery cohort (n=499), a four-gene module (GAP43, VIM, ZEB1, TWIST1) emerged as the only combination for which every pairwise Spearman correlation exceeded r=0.50 (r=0.51–0.74); the module was fixed before external validation and never trained against CMS or outcome data. The resulting NEM score correlated strongly with Hallmark EMT ssGSEA activity (r=0.93) and identified 96.3% of CMS4 samples correctly. NEM-High status was not confined to CMS4, however: 35.3% of NEM-High tumours in the discovery cohort fell within CMS1–CMS3, a pattern that held across nine external GEO cohorts (24.0–46.8%, mean 38.0%, n≈3,742). NEM-High status carried a numerically elevated but non-significant hazard of death (HR=1.44–1.53, p>0.05). A median-split sensitivity check and a leave-VIM-out re-scoring each ruled out dichotomisation and gene-set overlap as explanations. An orthogonal Human Protein Atlas cross-check found no single module gene independently prognostic in COAD or READ, favouring a module-level interpretation. Continuous NEM activity and categorical CMS status describe overlapping but distinct aspects of CRC biology: a compact, correlation-derived module reconstructs most of a subtype's membership while still detecting programme activity the label leaves out.

1. Introduction

Colorectal cancer (CRC) shows substantial transcriptomic heterogeneity, formalised by the Consensus Molecular Subtypes (CMS1–CMS4) [1]. CMS4, termed “mesenchymal,” is characterised by stromal infiltration and epithelial-mesenchymal transition (EMT) and carries the poorest clinical outcome among the four subtypes. CMS4 assignment, as implemented by reference-template classifiers such as CMScaller [2], produces a categorical output derived from a fixed gene-expression template. Categorical subtype assignment and continuous biological programme activity need not, however, describe identical dimensions of tumour biology. Metastatic CRC transcriptomes show CMS4-like

signal that is attributable in large part to non-tumour stromal composition [3], and single-cell analyses indicate that EMT-associated signal can occur outside the CMS4 label [4]. A separate body of work has approached CMS4 from the opposite direction, building classifiers or markers designed to recover the categorical label itself rather than to test its boundaries. A forty-gene classifier trained across three measurement platforms reproduces CMS1–4 assignment with reported accuracies of 0.88–0.91 [18]; a transcription-factor signature derived from CMS4-associated genes stratifies prognostic risk within the subtype [19]; and fibroblast activation protein has been proposed as a single-marker surrogate for CMS4, detectable by PET imaging [20]. Cross-species and proteomic extensions of the CMS framework have likewise treated categorical assignment as the object of study rather than as a boundary to be tested against a continuous alternative [21,22]. None of this work asks whether a programme, fixed independently of CMS and never trained to match it, diverges from the categorical label in a reproducible, quantifiable way across independent cohorts; that is the question addressed here.

Most recent transcriptomics-only biomarker studies in CRC follow a different design: candidate genes are chosen because they are differentially expressed, then combined into a risk score through outcome-supervised methods such as LASSO-Cox regression, with the resulting signature validated against survival in TCGA and one or more GEO cohorts [5–8]. Such an approach suits the construction of a prognostic classifier well. Because gene selection is guided by the very outcome it optimises, though, it cannot on its own establish whether the resulting score measures something distinct from an existing categorical subtype. A complementary approach was taken here: candidate genes were selected purely from a biologically curated panel and their pairwise co-expression structure, with no use of survival or CMS labels at any stage of module discovery. Any subsequent relationship to CMS4 or to survival is, as a result, a genuinely independent observation rather than a consequence of how the module was built.

Whether a neural-mesenchymal transcriptional module, identified empirically from a curated candidate gene panel and fixed prior to any external validation, would show reproducible, quantifiable discordance from categorical CMS4 assignment across independent cohorts, was the central question addressed. This study reports the derivation, validation, and biological characterisation of the resulting NEM score, followed by an orthogonal protein-level cross-check of the module genes against the Human Protein Atlas, and is informed throughout by REMARK [9] reporting principles for prognostic marker studies.

2. Materials and Methods

2.1 Study design and cohorts

This retrospective, multi-cohort transcriptomic study examined the relationship between a neural-mesenchymal transcriptional programme and Consensus Molecular Subtype (CMS) classification in colorectal cancer. The TCGA-COAD cohort served as the discovery dataset. RNA-sequencing data were retrieved from the Genomic Data Commons using TCGAbiolinks::GDCquery() (data category: Transcriptome Profiling; workflow: STAR-Counts).

External validation drew on nine independent colorectal cancer cohorts obtained from the Gene Expression Omnibus (GEO): GSE13067, GSE13294, GSE14333, GSE17536, GSE20916, GSE2109, GSE33113, GSE35896, and GSE39582 (Table 1). GSE42284 was excluded before any outcome-related analysis, because dataset retrieval failed during acquisition. Each accession sits in a public repository (GEO: https://www.ncbi.nlm.nih.gov/geo) and is cited by accession number in Table 1 and the Data Availability Statement.

The analytical workflow followed a one-directional discovery-and-validation design. Identification of the neural-mesenchymal module and derivation of the NEM score took place exclusively in the TCGA-COAD discovery cohort. The resulting four-gene composition and scoring formula were then fixed and applied to the external cohorts without modification.

2.2 Data processing and quality control

GDC count matrices underwent sample-level quality control, including checks on total library size and duplicate sample records. Samples with total counts below 100,000 were excluded as low-depth observations. Low-expression genes were removed with edgeR::filterByExpr() [10], using group-aware filtering and a manual CPM>1 fallback where needed; all 14 candidate-panel genes were kept regardless of filtering outcome. Quality control left 499 evaluable samples: 458 primary tumours and 41 matched solid-tissue normal samples (Supplementary Table S4, S5).

Retained counts were normalised by the trimmed mean of M-values (TMM) method with edgeR::calcNormFactors() [10] and converted to log2 counts per million using edgeR::cpm() [10] for downstream analyses.

2.3 Candidate gene panel and empirical derivation of the NEM module

A biologically curated 14-gene panel, representing neural-associated, EMT/mesenchymal, and pluripotency-associated programmes, was evaluated: neural-associated genes NES, GAP43, GFAP, MSI1, SOX2, DCX, SYP, and TUBB3; EMT/mesenchymal genes VIM, ZEB1, TWIST1, and SNAI1; and pluripotency-associated genes POU5F1 and NANOG. No gene was added to or removed from this panel once analysis began.

Module membership followed a pre-specified, uniformly applied pairwise Spearman correlation threshold ($r \geq 0.50$), computed across all 91 pairwise combinations of the 14 panel genes in discovery-cohort tumour samples (n=458; Table 3). Four genes — GAP43, VIM, ZEB1, and TWIST1 — were the only combination for which every pairwise correlation among themselves exceeded this threshold (r=0.51–0.74); no other subset of the panel met the criterion for even two genes tested against each other. This module was fixed before external validation, before CMS classification was consulted, and before any survival data were examined. The NEM score was calculated as the mean of the within-cohort z-scores of these four genes:

$$NEM_i = [Z(GAP43_i) + Z(VIM_i) + Z(ZEB1_i) + Z(TWIST1_i)] / 4$$

Because z-scoring was performed independently within each cohort before pooling, within-cohort scale differences between RNA-sequencing and microarray platforms were reduced without requiring pooled cross-platform batch correction. This within-cohort standardisation was a deliberate choice, made to limit platform-driven variation across the multi-cohort validation; it reduces rather than eliminates platform-related differences, and does not by itself establish full cross-platform comparability of absolute score values. Equal weighting of the four z-scored genes, rather than a regression- or PCA-derived weighting scheme, was likewise a deliberate choice: the aim was an interpretable, portable score computable identically across platforms and cohorts without a fitted model, not an optimised predictive signature.

2.4 NEM stratification, CMS assignment, and pathway analyses

Samples were split into NEM-High and NEM-Low groups using k-means clustering of the continuous NEM score (k=2, nstart=20) with a fixed random seed. The four-gene composition and scoring formula were fixed a priori and applied without modification across cohorts; NEM-High/NEM-Low assignment, in contrast, was performed independently within each cohort using this same prespecified k=2 procedure on that cohort's own NEM score distribution, since scores were z-scored within cohort. A two-group split was chosen a priori for interpretability rather than derived from a data-driven cluster-number criterion; this dichotomisation simplifies a continuous score, a point addressed further in the Discussion. Cluster reproducibility was checked computationally at the point of use, and a median-split sensitivity analysis (Section 3.5) was performed to confirm that the NEM-CMS asymmetry does not depend on the k-means cut point.

CMS classification proceeded independently of NEM group assignment, using CMScaller [2] with the expression-processing workflow appropriate to each dataset.

Hallmark EMT activity was calculated independently for individual tumour samples through single-sample gene-set enrichment analysis (GSVA::gsva [11]) with the MSigDB Hallmark gene-set collection, then compared with the continuous NEM score by Spearman correlation.

To characterise the broader transcriptional context tied to NEM activity, genome-wide Spearman correlations between the continuous NEM score and individual genes were calculated in the discovery cohort, excluding the four genes used to build the score. The top 200 genes by absolute correlation magnitude were then subjected to Hallmark gene-set enrichment analysis using hypergeometric testing, with Benjamini-Hochberg false-discovery-rate correction.

2.5 Differential expression and NEM-CMS cross-classification

Tumour-versus-normal differential expression for all 14 candidate genes was assessed by the Mann-Whitney U test on log2-CPM expression — a distribution-free test suited to a small, targeted panel, rather than a genome-wide

generalised linear model — comparing primary tumour and matched solid-tissue normal samples, with Benjamini-Hochberg correction across the 14-gene panel.

For each cohort, samples were cross-classified by NEM group and CMS subtype. The proportion of CMS4 tumours classified as NEM-High, and the proportion of NEM-High tumours not assigned to CMS4, were each calculated in the discovery cohort and independently in all nine external GEO cohorts. This cross-classification pattern is read descriptively as NEM-CMS discordance; it does not imply that either NEM or CMS assignment represents an absolute biological ground truth.

2.6 Survival analysis

Overall survival was analysed by Kaplan-Meier estimation with log-rank testing and Cox proportional-hazards regression [12]. Univariable models assessed the association between NEM group and survival; multivariable models evaluated the association after adjustment for CMS4 status and, separately, for age and stage. The proportional-hazards assumption was checked using Schoenfeld residuals (global p=0.23), which showed no evidence of violation. Survival analyses were treated as secondary and exploratory relative to the primary molecular and cross-cohort replication analyses.

2.7 Statistical software and reproducibility

All analyses were run in R using edgeR [10], GSVA [11], msigdbr, CMScaller [2], and survival [12], among related packages. A global random seed was fixed a priori, to support computational reproducibility of stochastic analytical steps, including k-means clustering. Two-sided statistical testing was used throughout, with Benjamini-Hochberg correction applied where appropriate.

2.8 Orthogonal protein-level cross-check (Human Protein Atlas)

The Human Protein Atlas cross-check was used to assess whether the transcriptomic NEM module corresponded to detectable protein-level and tissue-compartment evidence, rather than to validate the prognostic performance of its individual genes. Publicly available data for GAP43, VIM, ZEB1, and TWIST1 were retrieved from the Human Protein Atlas (proteinatlas.org; version accessed 17 August 2026) [13]. This resource aggregates RNA-sequencing-based normal-tissue and single-cell expression, TCGA- and CPTAC-derived cancer transcript/protein data, and immunohistochemistry (IHC)-based protein staining across validated antibodies. For each gene the following were recorded: (i) normal colon and rectum tissue- and single-cell-level RNA and protein expression, where retrieved; (ii) the HPA-reported prognostic verdict for colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ), in both the TCGA cohort and an independent HPA validation cohort; and (iii) IHC-based protein detection in colorectal adenocarcinoma tissue microarrays, across all antibodies listed for that gene. Human Protein Atlas content is distributed under a CC BY-SA 4.0 licence; this cross-check used publicly released summary data only and did not involve access to primary patient-level images beyond those already displayed on the platform. The Human Protein Atlas applies its own pre-specified prognostic threshold and antibody-validation criteria, independent of the analytical pipeline used here, so this comparison functions as an external, protein-oriented check on the NEM module rather than a replication of the transcriptomic NEM/CMS analysis itself.

3. Results

3.1 Cohort characteristics and analytical framework

The discovery cohort comprised 499 evaluable TCGA-COAD samples after quality control: 458 primary tumours and 41 matched solid-tissue normal samples (Table 1). The nine external GEO validation cohorts contributed approximately 3,742 samples in aggregate, spanning RNA-sequencing and microarray platforms.

Table 1. Summary of cohorts analysed

Cohort	Platform	n (role)	Role
TCGA-COAD (GDC)	RNA-seq	499 (458 tumour / 41 normal)	Discovery
GSE13067	Microarray	74	External validation
GSE13294	Microarray	155	External validation

GSE14333	Microarray	290	External validation
GSE17536	Microarray	177	External validation
GSE20916	Microarray	145	External validation
GSE2109	Microarray	2158	External validation
GSE33113	Microarray	96	External validation
GSE35896	Microarray	62	External validation
GSE39582	Microarray	585	External validation

3.2 Development of the neural-mesenchymal transcriptional module

Pairwise Spearman correlation analysis among the 14 candidate genes in discovery-cohort tumour samples (Table 3, Figure 2) identified GAP43, VIM, ZEB1, and TWIST1 as the unique combination for which all pairwise correlations exceeded the pre-specified $r \geq 0.50$ threshold ($r = 0.51–0.74$). This four-gene block stood apart clearly from the pluripotency-associated markers POU5F1 and NANOG ($r \approx 0–0.07$) and from the remaining neural-associated genes, none of which formed a comparable, fully inter-correlated block at this threshold (Figure 2). This empirically identified module defined the composite NEM score and was fixed, without modification, before external validation.

3.3 Association of NEM activity with tumour biology and EMT

Of the 14 panel genes, 11 were significantly differentially expressed between primary tumour (n=458) and matched normal tissue (n=41) at FDR<0.05 (Mann-Whitney U test with Benjamini-Hochberg correction; Table 2, Figure 1). TUBB3 (log2FC=+3.00), SNAI1 (+1.58), and TWIST1 (+1.39) showed the strongest upregulation in tumour tissue; GAP43 (log2FC=−4.02), SYP (−2.27), and DCX (−1.61) showed the strongest downregulation. ZEB1 (log2FC=−1.58) and VIM (log2FC=−0.82) were also significantly downregulated overall in bulk tumour tissue relative to normal tissue. NES, NANOG, and POU5F1 did not reach statistical significance, and MSI1 narrowly missed it (FDR=0.064). Principal-component analysis separated tumour from normal tissue along the first principal component, which explained 16.2% of total variance (Supplementary Figure S2).

The continuous NEM score correlated strongly with independently calculated Hallmark EMT activity in the discovery cohort (Spearman r=0.93; Figure 5). Of the four NEM module genes, only VIM is itself a member of the 200-gene Hallmark EMT set; GAP43, ZEB1, and TWIST1 are not. To test whether this single shared gene drove the association, Hallmark EMT ssGSEA was recomputed with VIM removed from the gene set (199 genes). The correlation with the NEM score was essentially unchanged (Spearman r=0.933 versus r=0.934 for the full 200-gene set), which indicates that the NEM–Hallmark EMT concordance is not attributable to gene-set overlap. Within the CMS4 subset, GAP43 remained strongly correlated with ZEB1 (Figure 6). A positive relationship between GAP43 and VIM was also observed in the independent GSE39582 cohort (Spearman r=0.256, n=585; Figure 7), which supports partial preservation of the underlying transcriptional relationship in an external, platform-different dataset.

3.4 Distribution of NEM activity across CMS subtypes

K-means clustering of the continuous NEM score split the discovery cohort into NEM-High (n=201) and NEM-Low (n=257) groups (Figure 3).

NEM-High status was strongly enriched within CMS4: of the 135 CMS4 tumours with an available NEM group assignment, 130 (96.3%) were classified as NEM-High, a near-complete capture of the categorical CMS4 label by the four-gene continuous score. The relationship was markedly asymmetric in the reverse direction, though. Of the 201 NEM-High tumours, 130 (64.7%) were CMS4, while 71 (35.3%) were assigned to CMS1–CMS3 (Figure 4). Substantial neural-mesenchymal transcriptional activity was therefore reproducibly present outside the CMS4 label, even as CMS4 itself remained overwhelmingly composed of NEM-High tumours.

3.5 Reproducible NEM-CMS asymmetry across external cohorts

The asymmetric enrichment pattern observed in the discovery cohort — near-complete NEM-High capture of CMS4, alongside substantial NEM-High representation outside CMS4 — reproduced across all nine external GEO cohorts (Table 4, Figure 10). The proportion of NEM-High tumours assigned outside CMS4 ranged from 24.0% (GSE2109)

to 46.8% (GSE13294), with a mean of 38.0% across the nine cohorts; the discovery-cohort value of 35.3% fell within this external range.

NEM-High tumours were thus reproducibly enriched within CMS4 across every cohort tested, while elevated NEM activity was also reproducibly observed outside the CMS4 category, at a consistent order of magnitude across platforms and cohort compositions.

3.5b Sensitivity of the NEM-CMS asymmetry to dichotomisation method

Because the central NEM-CMS asymmetry finding rests on a dichotomised NEM-High/NEM-Low classification, a median-split sensitivity analysis was performed as an alternative to k-means, computed independently within each cohort on the same continuous NEM score. In the discovery cohort, median-split classification captured 97.0% of CMS4 tumours as NEM-High and placed 37.0% of NEM-High tumours outside CMS4, closely matching the k-means-based values (96.3% and 35.3%, respectively). Across the nine external GEO cohorts, median-split classification gave a mean of 37.3% of NEM-High tumours outside CMS4 (range 24.1–47.5%), closely matching the k-means-based mean of 38.0% (range 24.0–46.8%). The qualitative pattern of asymmetric enrichment was therefore unchanged by the choice of dichotomisation method.

3.6 Functional characterisation of the NEM-associated transcriptome

The top 200 genes associated with NEM activity were subjected to Hallmark gene-set enrichment analysis (Table 5). Hallmark Epithelial-Mesenchymal Transition dominated the enrichment, with 147 overlapping genes and an FDR of approximately 2.6×10^{-55}. Additional enrichment appeared for myogenesis, allograft rejection, inflammatory response, and KRAS signalling. Together with the strong sample-level correlation between NEM and Hallmark EMT activity, these findings tie the four-gene NEM score to a broader EMT-associated transcriptional context.

3.7 Exploratory survival analysis

NEM-High status was associated with a numerically increased hazard of death in the discovery cohort (HR=1.44; log-rank p=0.064; Figure 8, Figure 9, Table 6). After adjustment for CMS4 status, the estimated hazard ratio stayed elevated (HR=1.53), though the association did not reach statistical significance (p=0.14). An age- and stage-adjusted analysis yielded an HR of 1.34 (p=0.24). The proportional-hazards assumption was not violated (Schoenfeld global p=0.23; Supplementary Figure S1). These findings should be read as exploratory, not as evidence of independent prognostic significance.

3.8 Orthogonal protein-level evidence for the NEM module genes (Human Protein Atlas)

To assess whether the four NEM module genes carried independent, protein-level support consistent with their inclusion in the module, GAP43, VIM, ZEB1, and TWIST1 were cross-checked against the Human Protein Atlas (Table 7) [13].

GAP43 protein was not detected in glandular epithelial cells of normal colon or rectum by any of three independent antibodies (HPA013392, HPA015600, CAB004417); strong staining was confined to peripheral nerve/ganglion elements. Single-cell RNA data from normal colon showed a similarly near-zero expression across colonocyte, goblet, tuft, and enteric stem/transit-amplifying cell populations. GAP43 was not reported as prognostic in COAD or READ in either the TCGA or HPA validation cohort, and IHC staining was negative for all three antibodies in colorectal adenocarcinoma tissue microarrays.

VIM and ZEB1 protein were each detected by IHC in colorectal adenocarcinoma tissue (VIM: low-category staining, with visible DAB positivity for antibody CAB000080; ZEB1: medium/low staining for antibody CAB079943), consistent with their established roles as an EMT effector protein and EMT transcription factor, respectively [14,16]. Neither gene reached the Human Protein Atlas's own significance threshold for prognostic value in COAD or READ, in either the TCGA or validation cohorts.

TWIST1 was likewise non-prognostic in COAD (TCGA and validation) and in READ TCGA. In the READ validation cohort alone, TWIST1 was flagged by the Human Protein Atlas as potentially prognostic, with high expression favourable — the opposite direction from the EMT-harm expectation. This discordant, single-cohort result is reported

explicitly rather than left out. TWIST1 IHC data for colorectal cancer were listed as pending cancer-tissue analysis at the time of access.

Taken together, none of the four constituent genes individually reached the Human Protein Atlas's own prognostic-significance threshold in COAD or READ, with the single exception of TWIST1's counterintuitive, favourable-direction association in one validation cohort. Protein-level detectability in colorectal adenocarcinoma tissue followed the order VIM (clearly detected), then ZEB1 (partially detected), then TWIST1 (protein data incomplete), then GAP43 — a transcript-level module member that went IHC-undetected in tumour epithelium, with protein confined to peripheral nerve elements in normal tissue. This pattern matches the discovery-cohort finding that GAP43 was the most weakly and most platform-dependently correlated module member (Section 3.3), and it supports reading the NEM score as a module-level signal rather than the arithmetic sum of four independently prognostic single-gene biomarkers.

Table 7. Orthogonal protein-level cross-check of the NEM module genes (Human Protein Atlas)

Gene	Normal colon/rectum protein	COAD/READ prognostic verdict (HPA)	IHC in CRC tumour tissue	CPTAC / cancer protein data
GAP43	Not detected in glandular cells; high in peripheral nerve/ganglion (all 3 antibodies)	Not prognostic (COAD TCGA n=254, COAD validation n=486, READ TCGA n=88, READ validation n=207)	Not detected (0/3 antibodies; HPA013392, HPA015600, CAB004417)	Low nRPX (~0.2–2.1); no tumour enrichment
VIM	Not independently retrieved in this cross-check	Not prognostic (COAD TCGA, COAD validation, READ TCGA, READ validation)	Detected — Low category, visible DAB positivity (CAB000080, CAB058687)	nRPX ~0.4–0.8, mixed tumour/normal
ZEB1	Not independently retrieved in this cross-check	Not prognostic (COAD TCGA, COAD validation, READ TCGA, READ validation)	Detected — Medium/Low category (CAB079943); 2/3 antibodies unmarked	197 samples; quantification N/A
TWIST1	Not independently retrieved in this cross-check	Not prognostic in COAD (TCGA, validation) and READ TCGA; READ validation only: potential prognostic, high expression favourable (discordant direction)	Pending cancer-tissue analysis at time of access	197 samples; quantification N/A

4. Discussion

Colorectal cancer shows substantial transcriptional heterogeneity, and the Consensus Molecular Subtype framework offers a widely used categorical classification of it. In this multi-cohort analysis, an empirically derived four-gene neural-mesenchymal module (GAP43, VIM, ZEB1, TWIST1) was identified from a biologically curated 14-gene candidate panel, using a single, pre-specified pairwise correlation threshold applied uniformly across the panel, without reference to CMS labels or outcome data. The resulting NEM score was strongly associated with independently calculated Hallmark EMT activity and captured the great majority of CMS4 tumours (96.3%). High NEM activity was, at the same time, reproducibly observed outside CMS4: 35.3% of NEM-High tumours in the discovery cohort were assigned to CMS1–CMS3, a pattern that reproduced across nine external cohorts (24.0–46.8%, mean 38.0%).

This asymmetry forms the central conceptual finding of the study. A four-gene continuous score can recover a categorical subtype almost completely in one direction, while at the same time detecting biologically coherent signal that the categorical label misses in the other. Such a pattern fits a model in which CMS assignment identifies a dominant, template-matched molecular subtype, while a continuous score measures the relative activity of a specific transcriptional programme that need not be confined to any single template.

Most existing transcriptomics-only CRC biomarker studies build their gene sets in the opposite order: candidates are filtered for differential expression or a biological theme, then combined into a score through an outcome-supervised procedure such as LASSO-Cox regression, with prognostic performance validated in TCGA and additional GEO cohorts [5–8]. A five-gene EMT-related signature (COMP, MYL9, PCOLCE2, SCG2, TIMP1), for example, was derived by LASSO-Cox regression in TCGA-COAD and validated for survival prediction in two GEO cohorts [5]. Comparable outcome-supervised pipelines have produced six-gene hormone-secretion [6], six-gene lysosomal-pathway [7], and complement-pathway [8] risk signatures in colon adenocarcinoma, each validated against survival across multiple GEO datasets. These studies suit the construction of prognostic classifiers well. Because gene selection is guided by the outcome being predicted, however, they cannot by construction address whether the resulting score measures something distinct from an existing categorical subtype such as CMS4. The present study fixed module membership from co-expression structure alone instead, before any outcome or subtype label was consulted, so the subsequent relationship to CMS4 and to survival forms an independent test rather than a built-in consequence of the selection procedure.

The composition of the NEM module is biologically notable, and independent literature on each constituent gene supports its selection. VIM, ZEB1, and TWIST1 are core, well-characterised components of the canonical EMT-transcription-factor and intermediate-filament programme. ZEB1 and TWIST1 belong to the small set of master EMT-activating transcription factors — alongside the Snail and ZEB families — that repress epithelial genes and induce mesenchymal ones, while vimentin acts both as a downstream effector and an upstream regulator of this same programme and is widely used as the prototypical mesenchymal marker [1,4,15,16]. GAP43 (also termed neuromodulin) is, by contrast, conventionally characterised as a neuronal growth-cone and axonal-plasticity protein rather than an EMT gene. An EMT-facilitating role for GAP43 outside the nervous system has independent experimental support, however: in papillary thyroid carcinoma, GAP43 knockdown raises E-cadherin and lowers N-cadherin and vimentin, and reduces proliferation, migration, and invasion, which indicates that GAP43 can causally modulate the EMT programme in an epithelial tumour context rather than merely correlate with it [16]. Colorectal and other gastrointestinal tumours are, separately, among the most densely innervated solid-tumour types, and perineural invasion — tumour-cell association with, and invasion of, nerve fibres — is an established, independent prognostic feature of CRC. It arises from bidirectional molecular crosstalk between tumour cells and the peritumoral nervous system, and neuromodulin/GAP43 is reported among the proteins induced in nerve fibres contacted by invading tumour cells in other neurotropic malignancies [17]. Within this framework, the empirical co-expression observed between GAP43 and the VIM/ZEB1/TWIST1 EMT core in the present data plausibly reflects a shared transcriptional context linking tumour mesenchymal/invasive activity to a reactive peritumoral neural microenvironment, rather than direct expression of GAP43 within mesenchymal-state tumour epithelial cells themselves. The Human Protein Atlas cross-check reported in Section 3.8 fits this interpretation: GAP43 protein was confined to peripheral nerve/ganglion elements in normal colorectal tissue and stayed undetectable by immunohistochemistry in tumour epithelium, despite the gene's retention in the co-expression-derived module, whereas VIM and ZEB1 protein were each directly detectable in tumour tissue. Within the discovery cohort, GAP43 was nonetheless the only neural-associated gene to meet the $r \geq 0.50$ threshold against all three mesenchymal genes, and this relationship partially persisted in an independent, platform-different cohort (GSE39582, GAP43-VIM $r=0.256$).

An additional observation concerns the heterogeneous tumour-versus-normal behaviour of the module genes. GAP43, ZEB1, and VIM were significantly downregulated on average in bulk tumour tissue relative to normal tissue, while TWIST1 was upregulated. A gene may, therefore, be downregulated on average in bulk tumour tissue while still remaining positively associated with a particular transcriptional state within the tumour population. Cellular heterogeneity may contribute to this pattern, but the present bulk-transcriptomic analysis cannot establish whether the NEM signal arises predominantly from tumour cells, stromal/neural cells, or their interaction. Single-cell and spatial transcriptomic studies will be needed to resolve this directly.

The Human Protein Atlas cross-check also offers an important calibration for the module-versus-single-gene reading of the NEM score. None of the four constituent genes individually reached the Human Protein Atlas's own significance threshold for prognostic value in COAD or READ, with the sole exception of a counterintuitive, favourable-direction result for TWIST1 in one validation cohort. This pattern does not undermine the present findings; instead, it reinforces the module-based framing adopted throughout this study. Biological and prognostic signal appears to reside in the composite four-gene programme rather than in any single constituent transcript or protein considered alone.

The survival findings call for cautious interpretation. NEM-High status was associated with a numerically increased hazard of death, but neither the univariable nor the CMS4-adjusted association reached conventional statistical significance, and the present data do not establish independent prognostic significance for the NEM score. The proportional-hazards assumption was formally verified and not violated, which supports the validity of the Cox estimates obtained, even though their precision remained limited. These analyses should be regarded as exploratory and hypothesis-generating for larger, harmonised cohorts.

From a translational perspective, the observed asymmetry may matter when molecular analyses rely exclusively on categorical CMS assignment to study mesenchymal biology. Restricting analyses to CMS4 alone would still capture the great majority of NEM-High tumours, but it would miss roughly one-third to two-fifths of NEM-High tumours classified elsewhere. Several explanations for this NEM-High, non-CMS4 group remain open: transitional tumours positioned between subtypes, stromal or neural contamination of bulk tissue, mixed cellular composition within a sample, a continuous mesenchymal gradient that the categorical template does not fully capture, template-classifier misassignment, or a genuinely distinct biology. Bulk transcriptomic data cannot adjudicate between these possibilities. The safest interpretation, and the one adopted here, is that NEM activity identifies a transcriptional dimension not confined to the CMS4 category, rather than a newly discovered molecular subtype; whether these tumours form a distinct clinical entity, or differ in therapeutic response, remains a hypothesis for future work rather than a conclusion supported by the present data. The practical consequence is narrower than the phrase "molecular subtyping" might suggest. Studies that condition analysis, cohort selection, or biomarker validation on CMS4 status alone will, on the cohorts examined here, still capture most tumours with elevated neural-mesenchymal activity, but they will systematically exclude a substantial minority, roughly a third to two-fifths, that a continuous score identifies and a categorical label does not. Whether that excluded minority differs in treatment response or outcome was not tested in this study. The finding reported here is limited to showing that the minority exists and recurs across nine independent cohorts, which is a precondition for asking that question rather than an answer to it.

Several limitations should be noted. First, this was a retrospective and observational analysis, and it cannot establish causal roles for the genes comprising the NEM module. Second, module discovery was constrained to a biologically curated 14-gene candidate panel and a single, pre-specified correlation threshold, applied once in the TCGA-COAD discovery cohort; the module was not independently re-derived in an external cohort, only fixed and then tested there, so its composition reflects discovery-cohort correlation structure rather than a structure confirmed to arise independently elsewhere. Alternative candidate panels or thresholds may identify additional or different transcriptional relationships. The NEM-High/NEM-Low dichotomisation used for cross-tabulation and survival analysis was checked against a median-split alternative (Section 3.5b) and gave closely matching results, which limits, though does not eliminate, concern about threshold-dependence of the categorical framework. Third, only one NEM module gene (VIM) overlaps with the Hallmark EMT gene set; a leave-VIM-out re-scoring (Section 3.3) gave a correlation with the NEM score ($r=0.933$) nearly identical to the full-set result ($r=0.934$), which addresses gene-set-overlap circularity directly rather than as an open concern. Fourth, the external datasets were generated on heterogeneous transcriptomic platforms; within-cohort z-scoring reduces but cannot remove all platform-related variation. Fifth, the biological source of the NEM signal cannot be resolved using bulk transcriptomic data alone. Sixth, the Human Protein Atlas cross-check in Section 3.8 is itself a bulk, unstratified analysis of publicly deposited TCGA- and antibody-based data, applying HPA's own pre-specified thresholds rather than the present study's methodology; it found no single NEM gene individually prognostic in COAD or READ, and this null single-gene result should be read as support for the module-level interpretation of the NEM score, not as an independent replication of the present cohort-level findings. Seventh, survival analysis was performed only in the TCGA-COAD discovery cohort; per-sample clinical outcome data required for Cox and Kaplan-Meier analysis were not available for the nine external GEO cohorts, so the reported hazard ratios (Section 3.6, Figures 8-9) are not independently replicated externally and should be read as discovery-cohort findings, in contrast to the NEM-CMS discordance metric, which was replicated across all nine external cohorts. Finally, clinical and treatment annotation was not uniformly available across cohorts, which limits conclusions about prognosis and therapeutic response.

5. Conclusions

This multi-cohort analysis identified a four-gene neural-mesenchymal transcriptional module (GAP43, VIM, ZEB1, TWIST1), selected solely on a pre-specified pairwise co-expression threshold and fixed before any exposure to CMS labels or outcome data. The resulting continuous NEM score was strongly associated with EMT-associated transcription, captured 96.3% of CMS4 tumours, and reproducibly identified substantial neural-mesenchymal activity

outside the CMS4 category across nine independent colorectal cancer cohorts (mean 38.0%, range 24.0–46.8%). An orthogonal protein-level cross-check against the Human Protein Atlas found no single module gene independently prognostic, which supports a module-level rather than single-gene interpretation of the underlying biology. These findings support a model in which continuous neural-mesenchymal programme activity and categorical molecular subtype assignment represent overlapping but non-identical dimensions of colorectal cancer biology.

CRediT Author Statement

A.R.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing — original draft, Writing — review & editing.

Data Availability Statement

TCGA-COAD data are publicly available via the Genomic Data Commons (https://portal.gdc.cancer.gov). All nine GEO datasets are publicly available under accessions GSE13067, GSE13294, GSE14333, GSE17536, GSE20916, GSE2109, GSE33113, GSE35896, and GSE39582 (https://www.ncbi.nlm.nih.gov/geo). Human Protein Atlas summary data used for the orthogonal cross-check in Section 2.8/3.8 are publicly available at https://www.proteinatlas.org under a CC BY-SA 4.0 licence. Derived summary statistics (including per-sample continuous NEM scores, NEM-group assignment, and CMS labels, provided as Supplementary Table S8), analysis code, and figures/tables are provided as a submission-ready supplementary package and are available from the corresponding author upon reasonable request.

Declaration of Competing Interests

The author declares no conflict of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics Statement

Not applicable; this study used publicly available, de-identified data only.

Competing Interests

The author declares no competing financial or non-financial interests.

Data Availability

All data used in this study are publicly available from The Cancer Genome Atlas (TCGA-COAD, via GDC) and the NCBI Gene Expression Omnibus (accession numbers GSE13067, GSE13294, GSE14333, GSE17536, GSE20916, GSE2109, GSE33113, GSE35896, and GSE39582).

Author Contributions

A.R. conceived the study, performed all data acquisition, analysis, and interpretation, and wrote the manuscript.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used Claude (Anthropic) to assist with cross-referencing and tabulating publicly available Human Protein Atlas summary data, targeted literature retrieval in support of gene-panel rationale, and language editing of the manuscript text. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Figure Legends

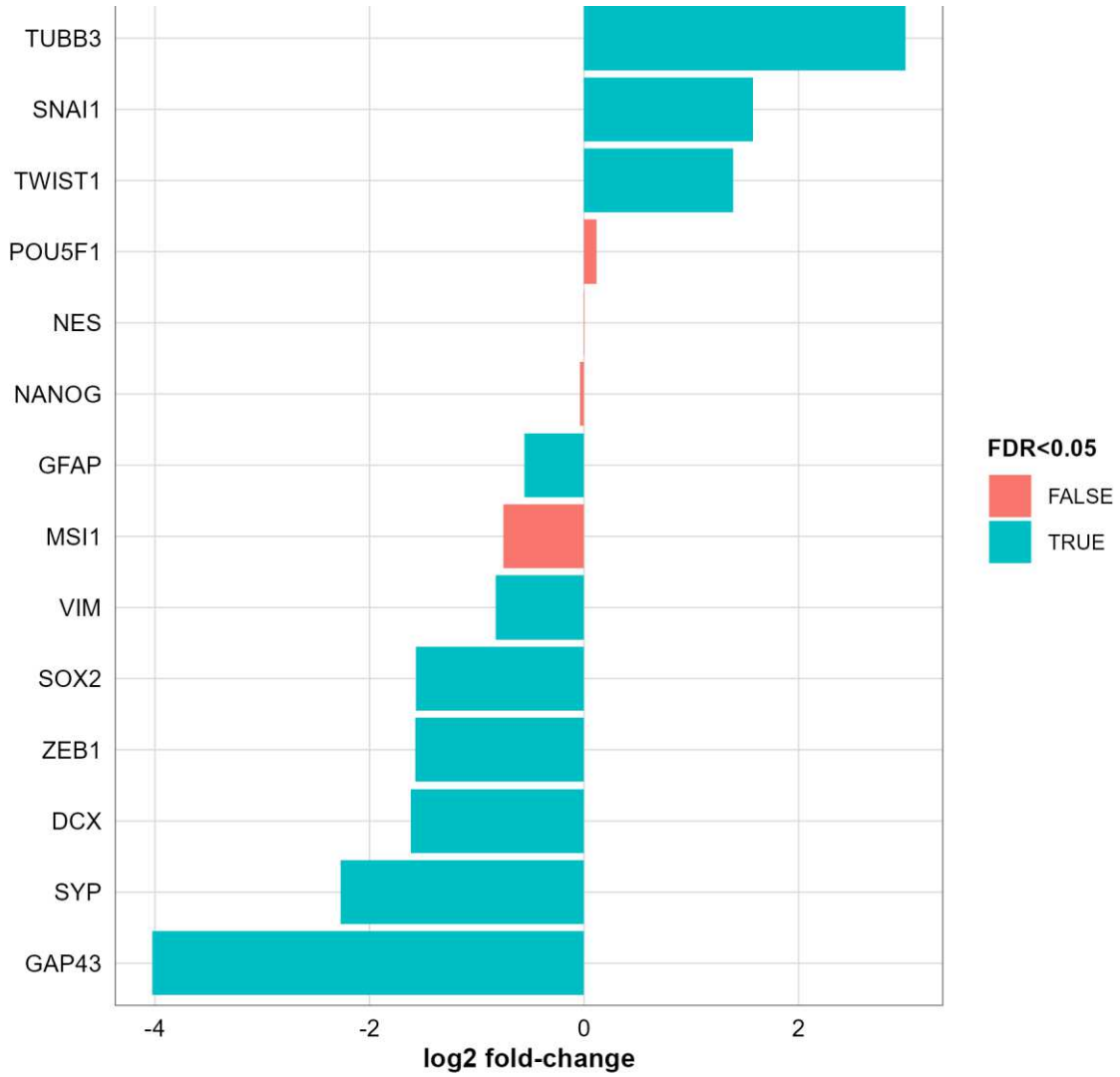


Figure 1. Differential expression of the 14 candidate panel genes between primary tumour (n=458) and matched solid-tissue normal (n=41) samples in the TCGA-COAD discovery cohort. Bars show log2 fold-change (tumour vs normal); genes reaching FDR<0.05 (Mann-Whitney U test with Benjamini-Hochberg correction) are indicated.

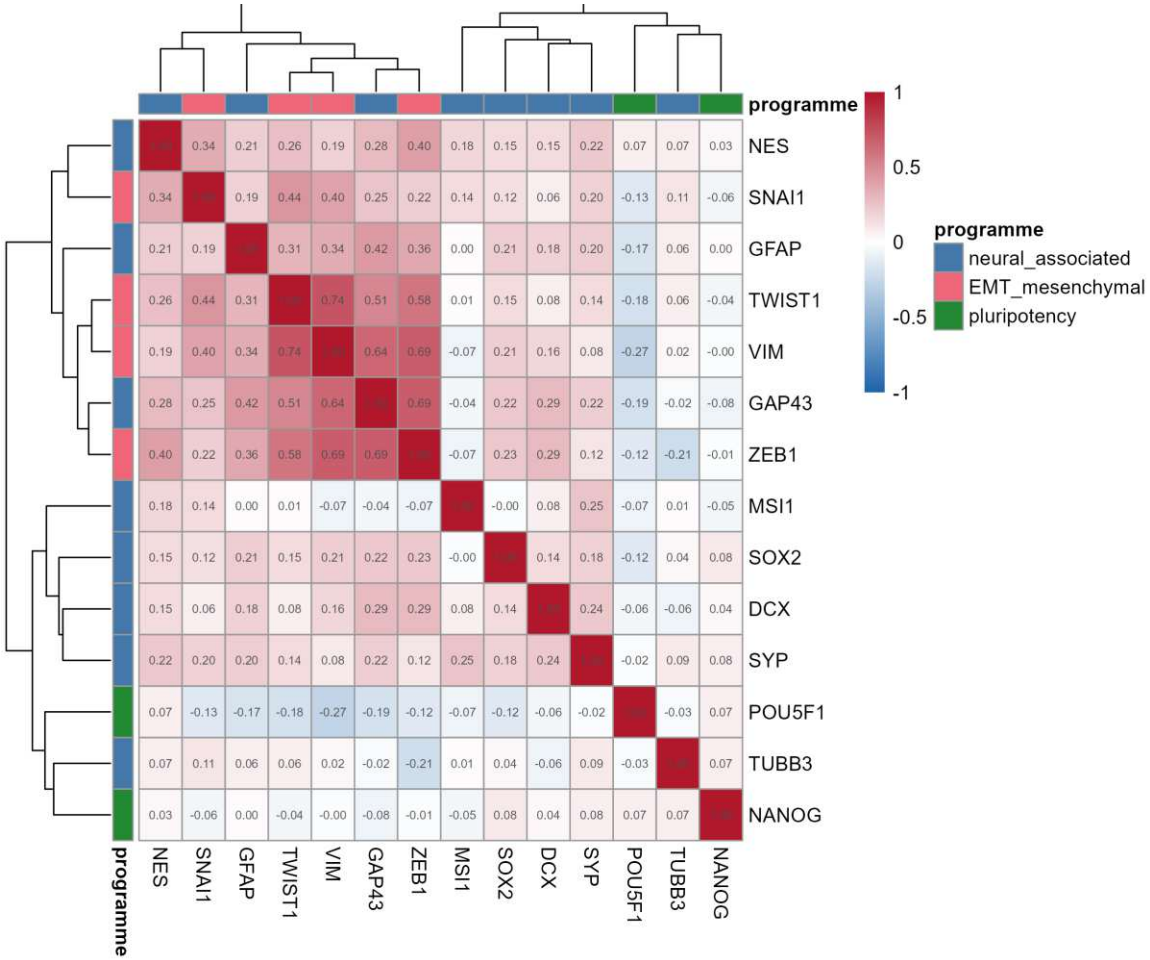


Figure 2. Pairwise Spearman correlation heatmap of the 14 candidate panel genes in TCGA-COAD discovery-cohort tumour samples, showing the fully inter-correlated four-gene block (GAP43, VIM, ZEB1, TWIST1; r=0.51–0.74) that defined the NEM module, distinguished from the pluripotency markers (POU5F1, NANOG) and remaining neural-associated genes.

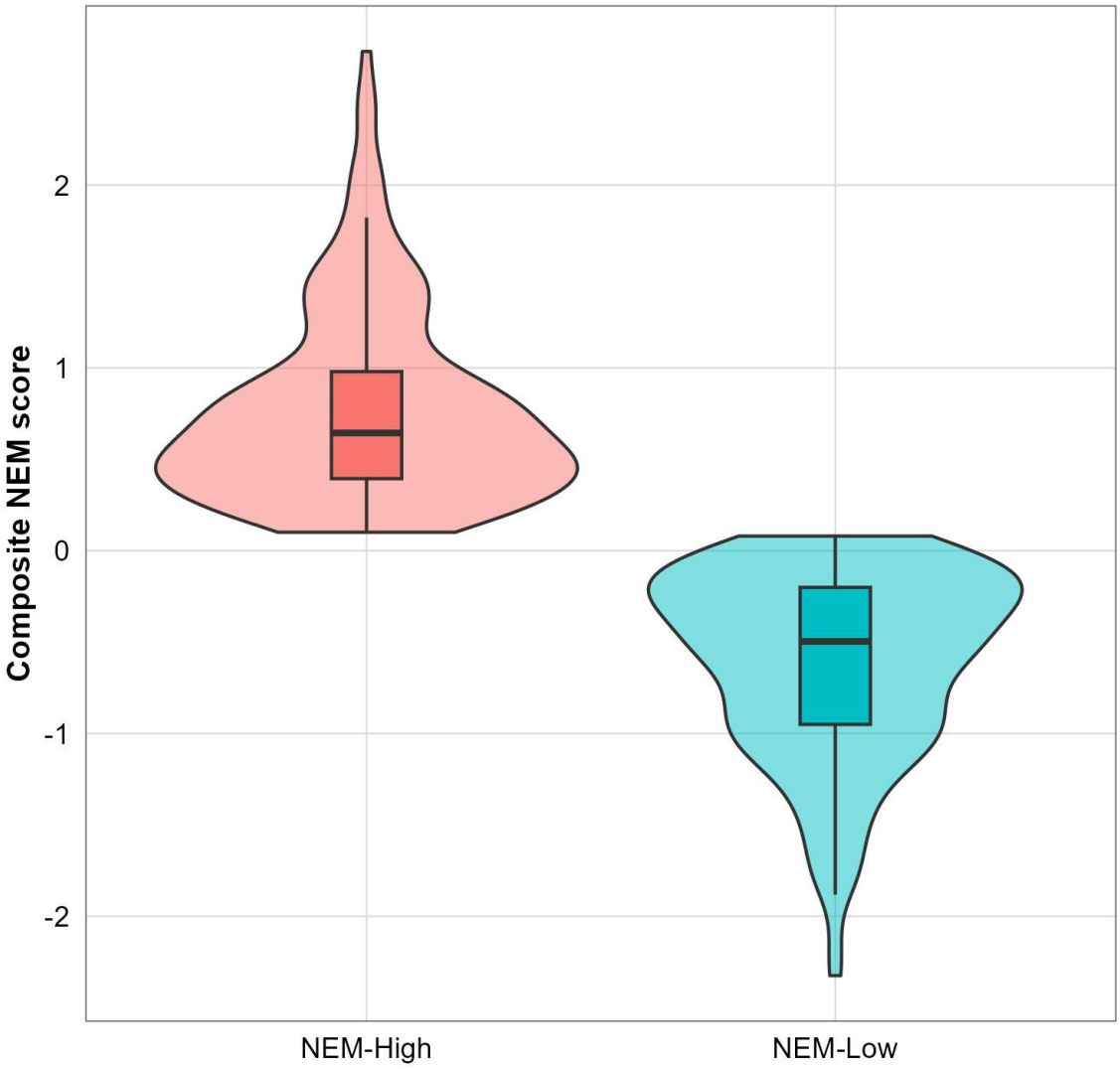


Figure 3. K-means partition of the continuous NEM score into NEM-High (n=201) and NEM-Low (n=257) groups in the TCGA-COAD discovery cohort.

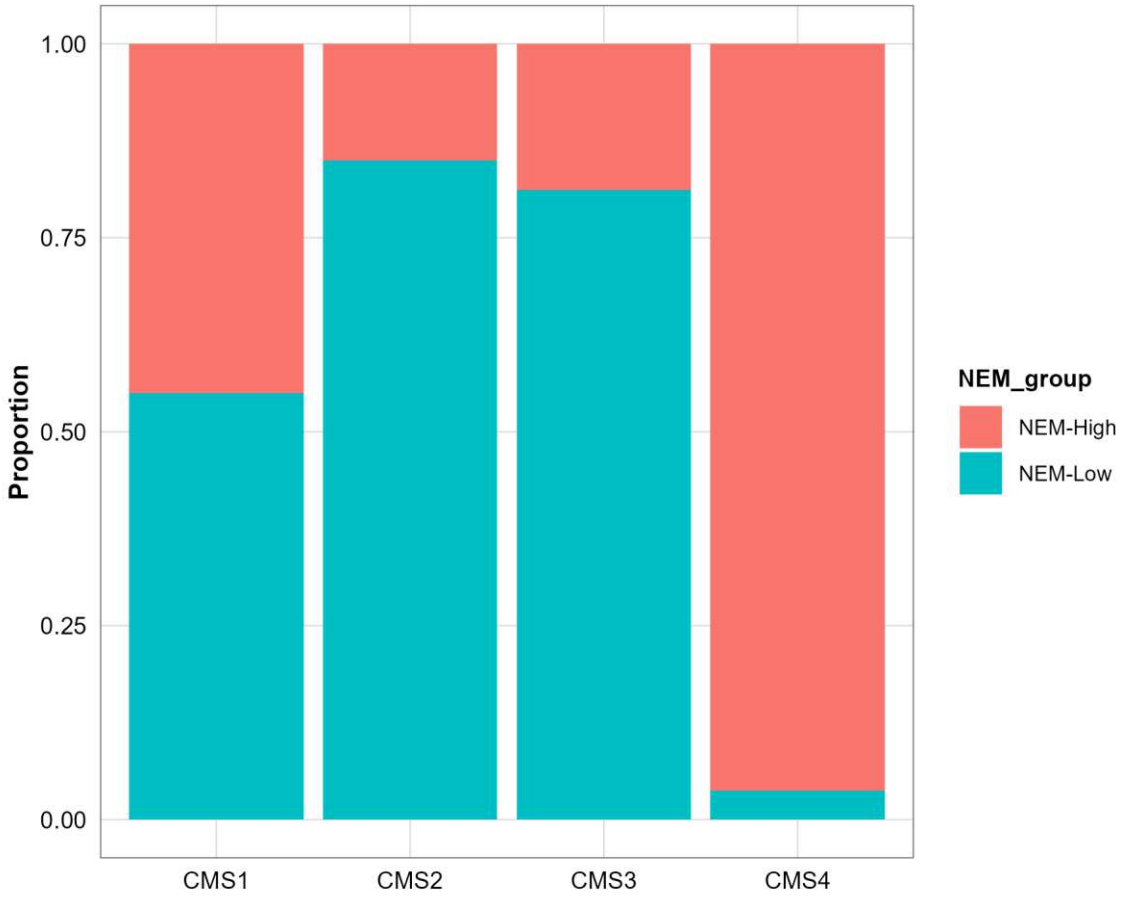


Figure 4. Cross-tabulation of NEM group (NEM-High/NEM-Low) against CMS subtype assignment in the discovery cohort, illustrating near-complete NEM-High capture of CMS4 (96.3%) alongside substantial NEM-High representation among CMS1–CMS3 tumours (35.3% of NEM-High cases).

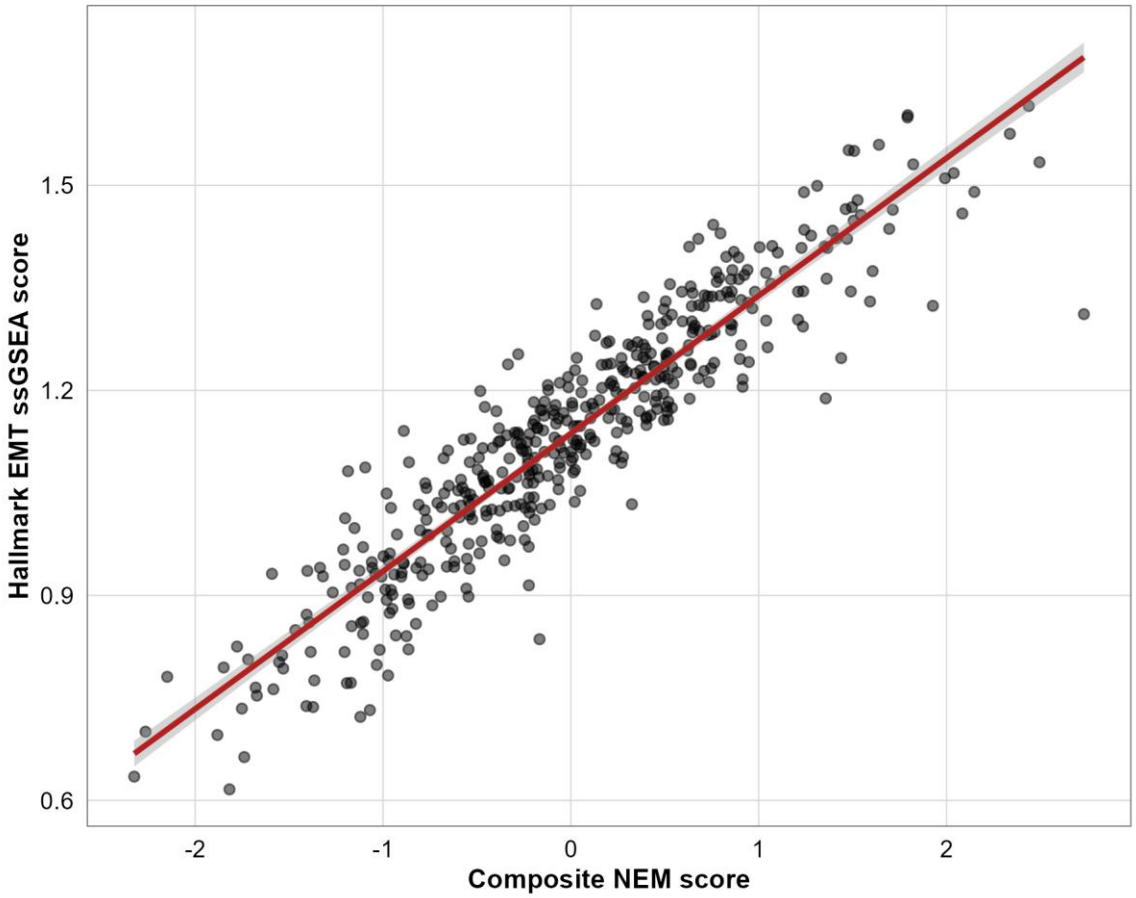


Figure 5. Correlation between the continuous NEM score and independently computed Hallmark Epithelial-Mesenchymal Transition single-sample GSEA activity in the discovery cohort (Spearman r=0.93).

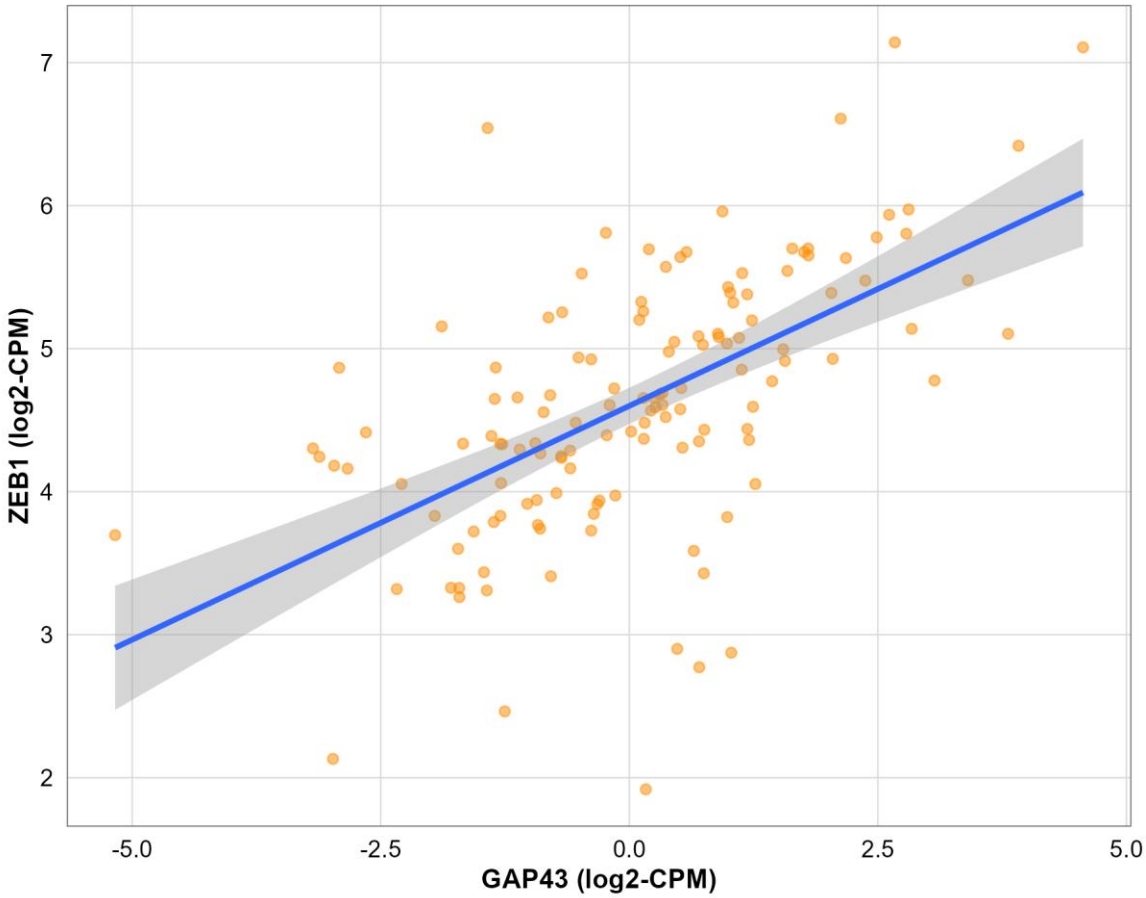


Figure 6. Correlation between GAP43 and ZEB1 expression within the CMS4 tumour subset of the discovery cohort.

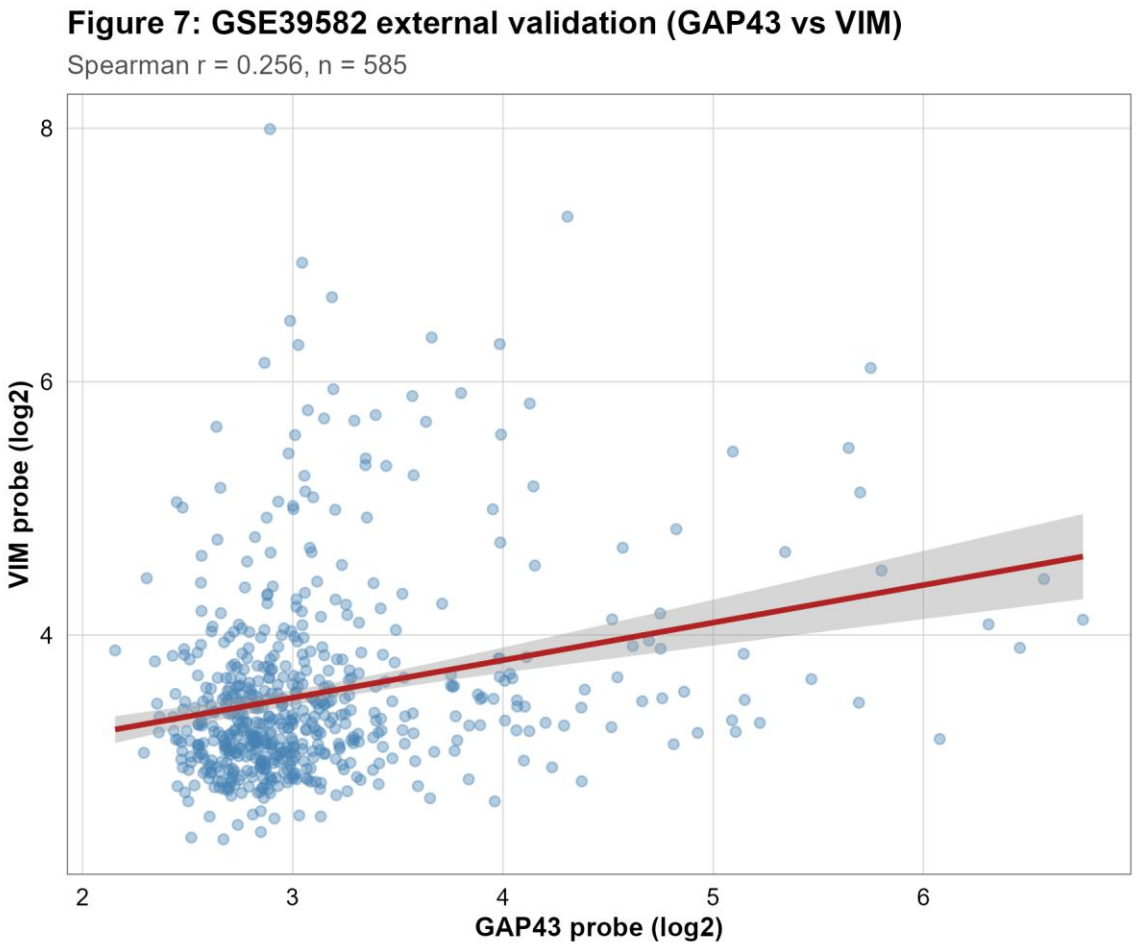


Figure 7. Correlation between GAP43 and VIM expression in the independent, platform-different GSE39582 cohort (Spearman r=0.256, n=585), supporting partial preservation of the discovery-cohort relationship externally.

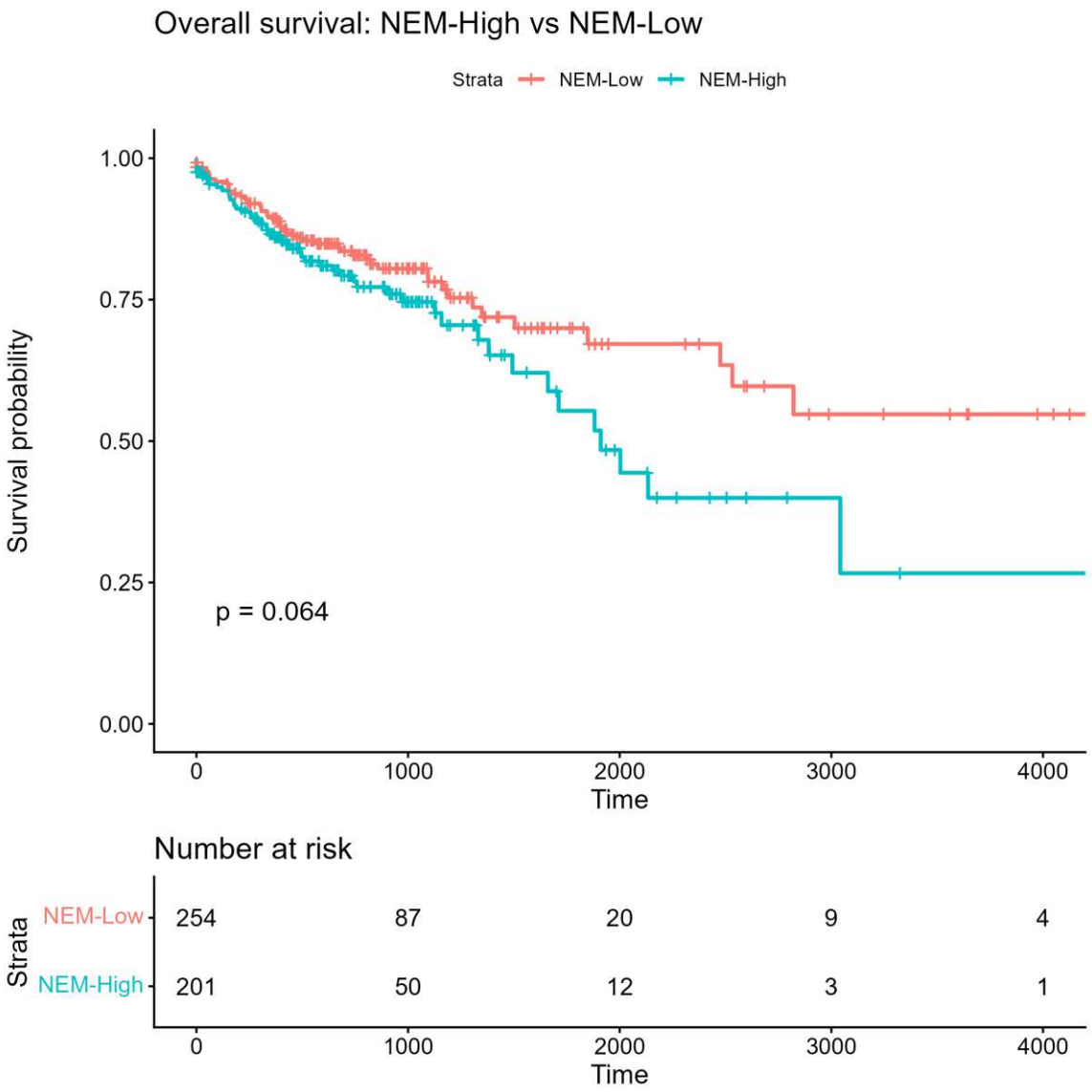


Figure 8. Kaplan-Meier overall survival curves stratified by NEM group (NEM-High vs NEM-Low) in the discovery cohort.

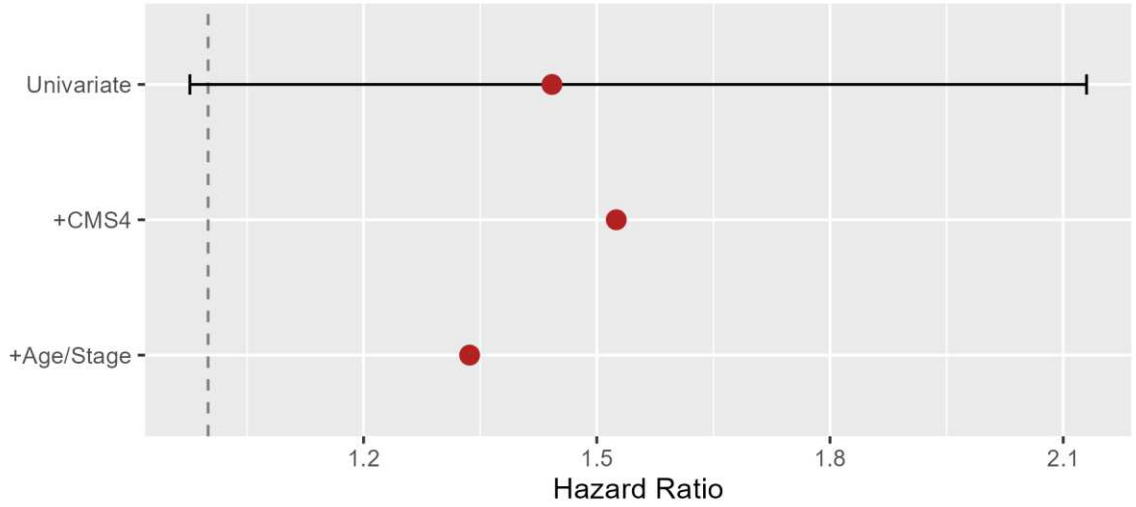


Figure 9. Forest plot of hazard ratios for NEM-High status from unadjusted, CMS4-adjusted, and age/stage-adjusted Cox proportional-hazards models in the discovery cohort.

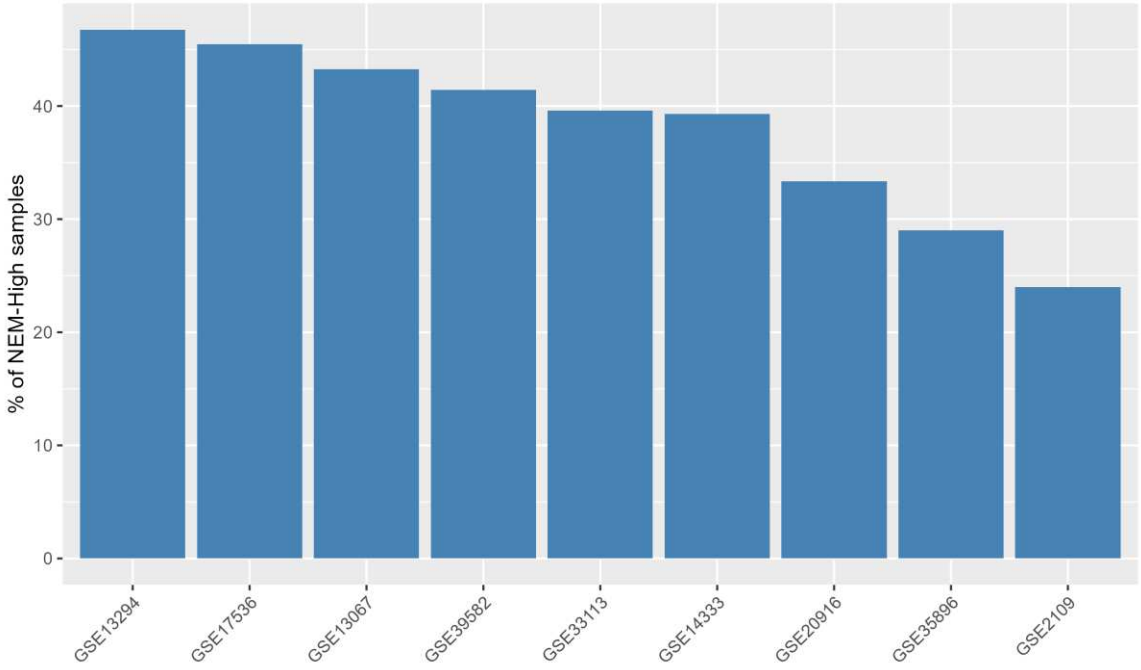


Figure 10. Proportion of NEM-High tumours assigned outside CMS4 across the TCGA-COAD discovery cohort and all nine external GEO validation cohorts, illustrating reproducibility of the NEM-CMS asymmetry (range 24.0–46.8%; mean 38.0%).

Table Captions

Table 2. Differential expression statistics (log2 fold-change, Mann-Whitney U p-value, Benjamini-Hochberg-adjusted FDR) for each of the 14 candidate panel genes, tumour vs matched normal tissue, TCGA-COAD discovery cohort.

gene_symbol	n_tumour	n_normal	mean_log2cpm_tumour	mean_log2cpm_normal	log2FC	p_value	FDR	significant_FDR05
TUBB3	458	41	-0.758	-3.756	2.998	0	0	TRUE
SNAI1	458	41	2.583	1.006	1.577	0	0	TRUE
SYP	458	41	1.027	3.296	-2.268	4.85×10^-23	2.26×10^-22	TRUE

GAP43	458	41	-1.666	2.357	-4.023	1.11×10^-22	3.88×10^-22	TRUE
DCX	458	41	-3.911	-2.297	-1.615	1.23×10^-12	2.86×10^-12	TRUE
ZEB1	458	41	3.536	5.111	-1.576	1.12×10^-12	2.86×10^-12	TRUE
TWIST1	458	41	1.034	-0.356	1.390	9.26×10^-11	1.85×10^-10	TRUE
VIM	458	41	7.085	7.910	-0.825	1.29×10^-9	2.27×10^-9	TRUE
SOX2	458	41	-1.777	-0.212	-1.565	3.19×10^-7	4.97×10^-7	TRUE
GFAP	458	41	-2.137	-1.581	-0.555	0.006	0.009	TRUE
MSI1	458	41	1.041	1.795	-0.755	0.050	0.064	FALSE
NANOG	458	41	-4.569	-4.531	-0.038	0.159	0.173	FALSE
POU5F1	458	41	1.222	1.105	0.117	0.160	0.173	FALSE
NES	458	41	5.097	5.093	0.004	0.969	0.969	FALSE

Table 3. Pairwise Spearman correlation coefficients among the 14 candidate panel genes in TCGA-COAD discovery-cohort tumour samples (n=458), underlying the derivation of the four-gene NEM module.

	NES	SOX2	TUBB3	DCX	GFAP	GAP43	MSI1	SYP	ZEB1	TWIST1	SNAI1	VIM	NANOG	POU5F1
NES	1	0.147	0.068	0.147	0.210	0.280	0.175	0.222	0.403	0.260	0.343	0.190	0.027	0.072
SOX2	0.147	1	0.037	0.142	0.212	0.224	-0.001	0.184	0.228	0.154	0.123	0.208	0.084	-0.124
TUBB3	0.068	0.037	1	-0.064	0.062	-0.022	0.015	0.094	-0.207	0.061	0.107	0.023	0.070	-0.026
DCX	0.147	0.142	-0.064	1	0.182	0.285	0.076	0.235	0.294	0.076	0.064	0.160	0.039	-0.056
GFAP	0.210	0.212	0.062	0.182	1	0.424	0.004	0.201	0.364	0.306	0.192	0.341	0.003	-0.169
GAP43	0.280	0.224	-0.022	0.285	0.424	1	-0.042	0.215	0.685	0.506	0.247	0.641	-0.078	-0.191
MSI1	0.175	-0.001	0.015	0.076	0.004	-0.042	1	0.252	-0.073	0.010	0.143	-0.069	-0.054	-0.065
SYP	0.222	0.184	0.094	0.235	0.201	0.215	0.252	1	0.117	0.140	0.196	0.082	0.076	-0.017
ZEB1	0.403	0.228	-0.207	0.294	0.364	0.685	-0.073	0.117	1	0.581	0.216	0.695	-0.014	-0.121
TWIST1	0.260	0.154	0.061	0.076	0.306	0.506	0.010	0.140	0.581	1	0.443	0.736	-0.043	-0.185
SNAI1	0.343	0.123	0.107	0.064	0.192	0.247	0.143	0.196	0.216	0.443	1	0.398	-0.057	-0.134
VIM	0.190	0.208	0.023	0.160	0.341	0.641	-0.069	0.082	0.695	0.736	0.398	1	-0.002	-0.268
NANOG	0.027	0.084	0.070	0.039	0.003	-0.078	-0.054	0.076	-0.014	-0.043	-0.057	-0.002	1	0.069
POU5F1	0.072	-0.124	-0.026	-0.056	-0.169	-0.191	-0.065	-0.017	-0.121	-0.185	-0.134	-0.268	0.069	1

Table 4. NEM-High/NEM-Low group sizes, CMS4 capture rate, and proportion of NEM-High tumours outside CMS4, reported separately for the TCGA-COAD discovery cohort and each of the nine external GEO validation cohorts.

cohort	n	n_nem_high	n_cms4	n_cms4_missed_by_nem	pct_missed
GSE13067	74	37	19	16	43.2
GSE13294	155	77	39	36	46.8
GSE14333	290	145	90	57	39.3
GSE17536	177	88	51	40	45.5
GSE20916	145	72	39	24	33.3
GSE2109	2158	1079	720	259	24.0
GSE33113	96	48	29	19	39.6
GSE35896	62	31	22	9	29.0
GSE39582	585	292	163	121	41.4

Table 5. MSigDB Hallmark gene-set enrichment results for the top 200 genes associated with NEM activity, discovery cohort, ranked by FDR.

gene_set	hits	gene_set_size	input_size	p_value	FDR
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	147	200	4244	5.22×10^-57	2.61×10^-55
HALLMARK_MYOGENESIS	84	179	4244	1.14×10^-14	2.84×10^-13
HALLMARK_ALLOGRAFT_REJECTION	89	196	4244	2.15×10^-14	3.59×10^-13
HALLMARK_INFLAMMATORY_RESPONSE	88	196	4244	6.77×10^-14	8.46×10^-13
HALLMARK_KRAS_SIGNALING_UP	83	198	4244	2.87×10^-11	2.87×10^-10
HALLMARK_IL2_STAT5_SIGNALING	78	198	4244	3.55×10^-9	2.96×10^-8
HALLMARK_COMPLEMENT	77	196	4244	5.15×10^-9	3.68×10^-8
HALLMARK_UV_RESPONSE_DN	60	143	4244	1.45×10^-8	9.08×10^-8
HALLMARK_APICAL_JUNCTION	74	192	4244	2.63×10^-8	1.46×10^-7
HALLMARK_COAGULATION	52	126	4244	2.45×10^-7	1.22×10^-6
HALLMARK_ANGIOGENESIS	20	35	4244	3.56×10^-6	1.62×10^-5
HALLMARK_ADIPOGENESIS	69	197	4244	4.54×10^-6	1.89×10^-5

HALLMARK_INTERFERON_GAMMA_RESPONSE	69	198	4244	5.54×10^-6	2.13×10^-5
HALLMARK_TNFA_SIGNALING_VIA_NFKB	67	199	4244	2.69×10^-5	9.62×10^-5
HALLMARK_HYPOXIA	63	194	4244	0.000	0.001

Top 15 of 50 enriched gene sets shown; the complete ranked list is provided as Supplementary Data.

Table 6. Cox proportional-hazards regression results for NEM-High status (unadjusted, CMS4-adjusted, and age/stage-adjusted models), discovery cohort, including Schoenfeld residual global test statistic.

quantity	computed_value
cox_univariate_nem_hr	1.442
cox_univariate_nem_pvalue	0.066
cox_multivariable_nem_hr_adj_cms4	1.525
cox_multivariable_nem_pvalue_adj_cms4	0.139
logrank_pvalue_nem_high_vs_low	0.064
cox_ph_assumption_pvalue	0.228
cox_age_stage_adjusted_nem_hr	1.336
cox_age_stage_adjusted_nem_pvalue	0.241

References

References are numbered in order of first appearance in the text.

[1] Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21(11):1350-1356.

[2] Eide PW, Bruun J, Lothe RA, Sveen A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci Rep. 2017;7:16618.

[3] Eide PW, Moosavi SH, Eilertsen IA, et al. Metastatic heterogeneity of the consensus molecular subtypes of colorectal cancer. NPJ Genomic Med. 2021;6:59.

[4] Sehgal M, Ramu S, Vaz JM, et al. Characterizing heterogeneity along EMT and metabolic axes in colorectal cancer reveals underlying consensus molecular subtype-specific trends. Transl Oncol. 2023;40:101845.

[5] Zhang Y, Li Y, Zuo Z, Li T, An Y, Zhang W. An epithelial-mesenchymal transition-related mRNA signature associated with the prognosis, immune infiltration and therapeutic response of colon adenocarcinoma. Pathol Oncol Res. 2023;29:1611016.

[6] Jia X, Zhang T, Lv X, Du H, Sun Y, Guan Y. Identification of the six-hormone secretion-related gene signature as a prognostic biomarker for colon adenocarcinoma. Cancer Biomark. 2023;38(4):523-535.

[7] Luo Z, Wang Y, Zeng S, et al. Harnessing lysosomal genetics: development of a risk stratification panel and unveiling of DPP7 as a biomarker for colon adenocarcinoma. J Genet Genomics. 2025;52(12):1549-1562.

[8] Li J, Yu K, Chen Z, Xing D, Zha B, Xie W, Ouyang H, Yu C. A machine-learning prognostic model for colorectal cancer using a complement-related risk signature. Oncol Res. 2025;33(11):3469-3492.

[9] McShane LM, Altman DG, Sauerbrei W, et al. REporting recommendations for tumor MARKer prognostic studies (REMARK). J Natl Cancer Inst. 2005;97(16):1180-1184.

[10] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-140.

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

[12] Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Springer; 2000.

[13] Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.

[14] Zheng Y, Dai M, Dong Y, et al. ZEB2/TWIST1/PRMT5/NuRD multicomplex contributes to the epigenetic regulation of EMT and metastasis in colorectal carcinoma. Cancers. 2022;14:3426.

[15] Usman S, Waseem NH, Nguyen TKN, Mohsin S, Jamal A, Teh MT, Waseem A. Vimentin is at the heart of epithelial mesenchymal transition (EMT) mediated metastasis. Cancers. 2021;13(19):4985.

[16] Zheng C, Quan RD, Wu CY, et al. Growth-associated protein 43 promotes thyroid cancer cell lines progression via epithelial-mesenchymal transition. J Cell Mol Med. 2019;23(12):7974-7984.

[17] Wang H, Huo R, He K, Cheng L, Zhang S, Yu M, Zhao W, Li H, Xue J. Perineural invasion in colorectal cancer: mechanisms of action and clinical relevance. Cell Oncol (Dordr). 2024;47(1):1-17.

[18] Buechler SA, Stephens MT, Hummon AB, et al. ColoType: a forty gene signature for consensus molecular subtyping of colorectal cancer tumors using whole-genome assay or targeted RNA-sequencing. Sci Rep. 2020;10(1):12123.

[19] Zhong ME, Huang ZP, Wang X, et al. A transcription factor signature can identify the CMS4 subtype and stratify the prognostic risk of colorectal cancer. Front Oncol. 2022;12:902974.

[20] Strating E, Wassenaar E, Verhagen M, et al. Fibroblast activation protein identifies Consensus Molecular Subtype 4 in colorectal cancer and allows its detection by 68Ga-FAPI-PET imaging. Br J Cancer. 2022;127(1):145-155.

[21] Amirkhah R, Gilroy K, Malla SB, et al. MmCMS: mouse models' consensus molecular subtypes of colorectal cancer. Br J Cancer. 2023;128(7):1333-1343.

[22] Feliu J, Gámez-Pozo A, Martínez-Pérez D, et al. Functional proteomics of colon cancer Consensus Molecular Subtypes. Br J Cancer. 2024;130(10):1670-1678.

Supplementary Files

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

- SupplementaryFigureS1.jpg.png
- SupplementaryFigureS2.jpg.png
- SupplementaryTableS1CorrelationPvalues.csv
- SupplementaryTableS2GSE39582Validation.csv
- SupplementaryTableS3GEOCohortImportSummary.csv
- SupplementaryTableS4QCGeneFilterReport.csv
- SupplementaryTableS5QCSampleReport.csv
- SupplementaryTableS6PanelGeneSummaryStats.csv
- SupplementaryTableS7PCAVarianceExplained.csv