WSI-Inferred Spatial Transcriptomics for Colorectal Cancer

Duxiuju et al.

2026-02-21

目录

1 Abstract

Spatial transcriptomics (ST) provides an interpretable molecular readout of tissue architecture, yet its clinical scalability is constrained by cost and tissue requirements. We developed and benchmarked a whole-slide-image (WSI)-to-ST inference framework across internal leave-one-patient-out cohorts and external datasets in colorectal cancer. Across 418 target genes, the framework achieved robust spot-level concordance with measured ST profiles and preserved biologically meaningful spatial gradients. Comparative evaluation against multiple state-of-the-art baselines showed consistently stronger correlation distributions and a higher fraction of genes above practical concordance thresholds. These findings indicate that histology-driven virtual ST can recover substantial transcriptomic structure from routine pathology images and may support hypothesis generation in settings where direct ST is unavailable.

2 Introduction

Spatial context is central to colorectal cancer biology, where epithelial programs, stromal remodeling, and immune exclusion co-exist within heterogeneous tissue niches. Although ST can resolve this architecture, widespread deployment remains limited in retrospective cohorts and routine pathology workflows. Computational inference of ST from H&E WSIs offers a pragmatic alternative, but the field still requires rigorous cross-cohort validation and transparent reporting of per-gene and per-sample behavior.

Here, we evaluate a WSI-to-ST pipeline under internal and external settings, emphasizing clinically relevant robustness. We focus on gene-wise correlation structure, sample-level reproducibility, and interpretable summaries of model behavior. The study is designed to test not only aggregate performance but also whether inferred expression retains tissue-context fidelity across heterogeneous specimens.

3 Results

3.1 Robust cross-cohort concordance of WSI-inferred transcriptomes

We first evaluated gene-wise concordance between inferred and measured spatial transcriptomics in internal and external settings. In the external cohort, the global correlation distribution remained shifted toward positive agreement, indicating that the model generalizes beyond the training-like internal samples.

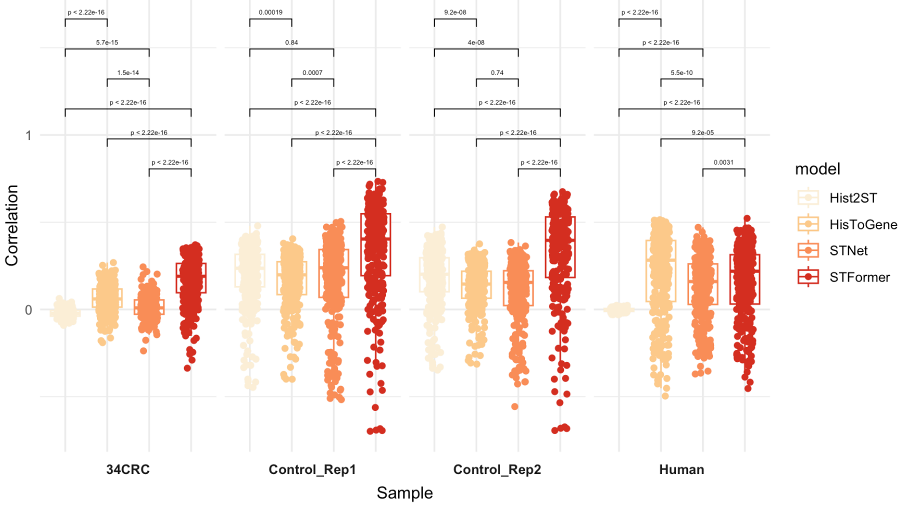


图 1: External cohort gene-wise correlation distribution.

Within the internal validation setting, per-sample correlation profiles showed consistent performance across patients, with expected heterogeneity in difficulty across tissue contexts.

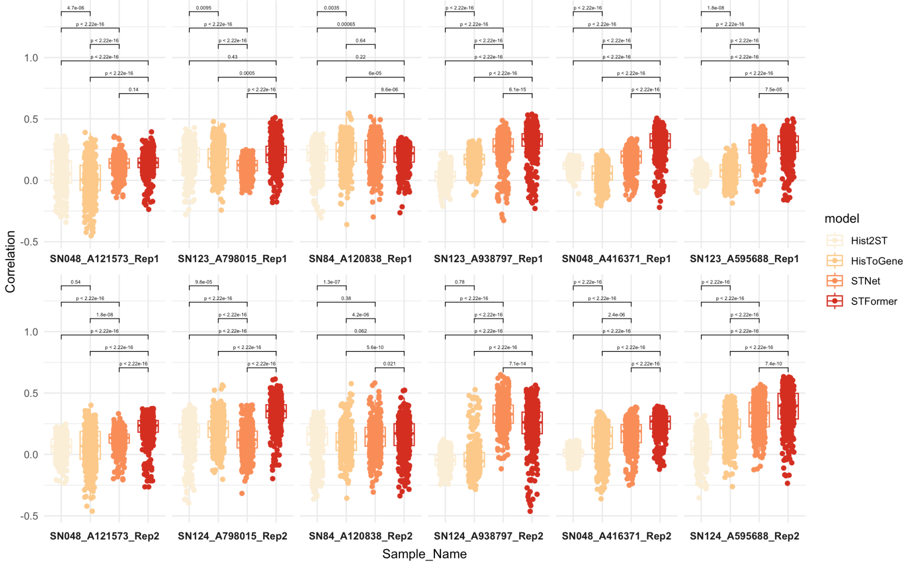
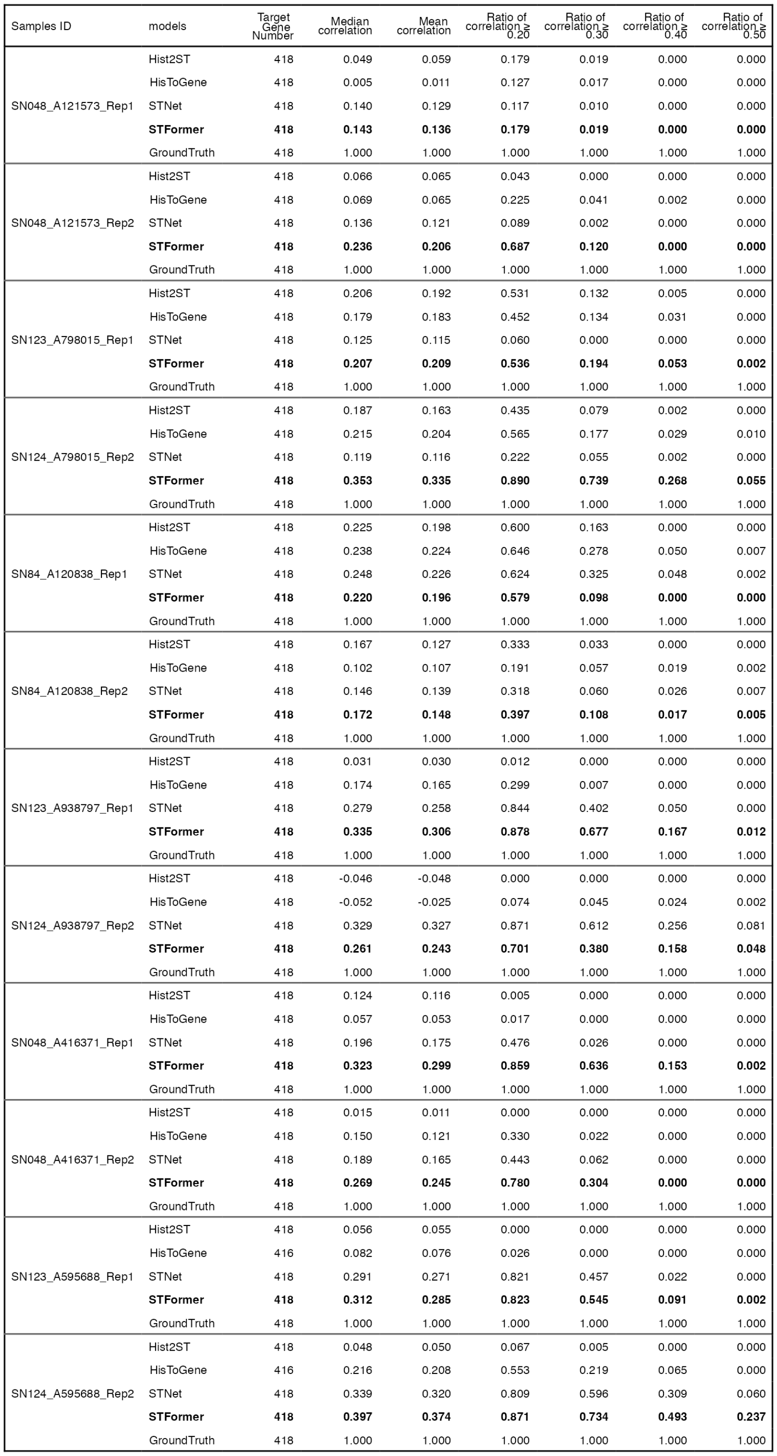


图 2: Internal per-sample correlation landscape (repository file: `Correlations_internalCV_persample.png`; corresponds to the requested internal per-sample correlation panel).

Together, these results support stable transfer of histology-derived molecular signal and suggest that performance differences are driven more by sample complexity than by systematic model collapse.

3.2 Quantitative benchmarking supports model-level separation

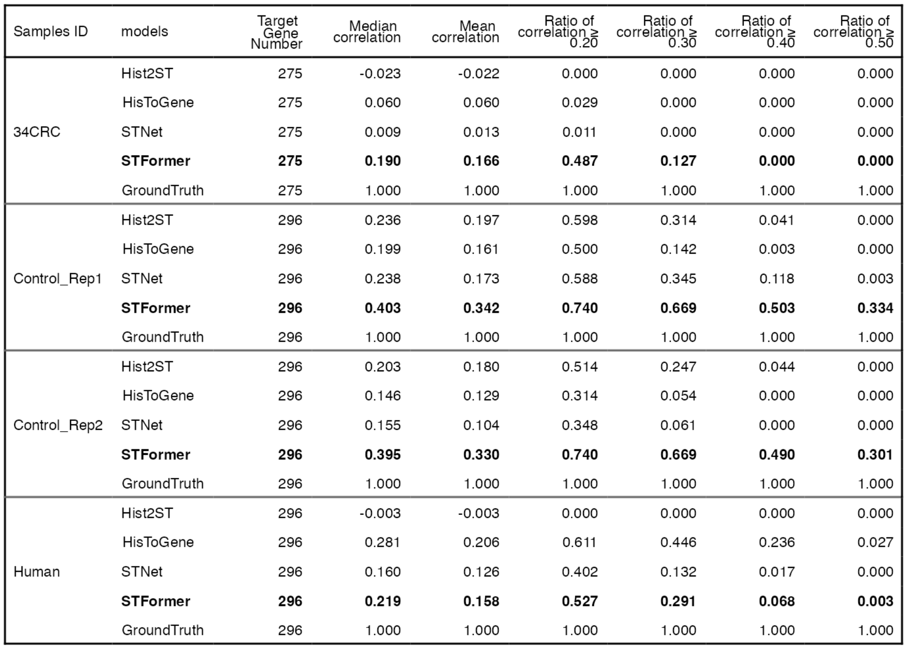
To quantify practical utility, we summarized target-gene coverage and threshold-based concordance metrics. In the internal 14-CRC setting, the leading method shows higher central tendency and a larger fraction of genes exceeding moderate-to-high correlation cutoffs, consistent with distributional trends in Fig. 图 2.



Samples ID	models	Target Gene Number	Median correlation	Mean correlation	Ratio of correlation ≥ 0.20	Ratio of correlation ≥ 0.30	Ratio of correlation ≥ 0.40	Ratio of correlation ≥ 0.50
SN048_A121573_Rep1	Hist2ST	418	0.049	0.059	0.179	0.019	0.000	0.000
	HisToGene	418	0.005	0.011	0.127	0.017	0.000	0.000
	STNet	418	0.140	0.129	0.117	0.010	0.000	0.000
	STFormer	**418**	**0.143**	**0.136**	**0.179**	**0.019**	**0.000**	**0.000**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN048_A121573_Rep2	Hist2ST	418	0.066	0.065	0.043	0.000	0.000	0.000
	HisToGene	418	0.069	0.065	0.225	0.041	0.002	0.000
	STNet	418	0.136	0.121	0.089	0.002	0.000	0.000
	STFormer	**418**	**0.236**	**0.206**	**0.687**	**0.120**	**0.000**	**0.000**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN123_A798015_Rep1	Hist2ST	418	0.206	0.192	0.531	0.132	0.005	0.000
	HisToGene	418	0.179	0.183	0.452	0.134	0.031	0.000
	STNet	418	0.125	0.115	0.060	0.000	0.000	0.000
	STFormer	**418**	**0.207**	**0.209**	**0.536**	**0.194**	**0.053**	**0.002**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN124_A798015_Rep2	Hist2ST	418	0.187	0.163	0.435	0.079	0.002	0.000
	HisToGene	418	0.215	0.204	0.565	0.177	0.029	0.010
	STNet	418	0.119	0.116	0.222	0.055	0.002	0.000
	STFormer	**418**	**0.353**	**0.335**	**0.890**	**0.739**	**0.268**	**0.055**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN84_A120838_Rep1	Hist2ST	418	0.225	0.198	0.600	0.163	0.000	0.000
	HisToGene	418	0.238	0.224	0.646	0.278	0.050	0.007
	STNet	418	0.248	0.226	0.624	0.325	0.048	0.002
	STFormer	**418**	**0.220**	**0.196**	**0.579**	**0.098**	**0.000**	**0.000**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN84_A120838_Rep2	Hist2ST	418	0.167	0.127	0.333	0.033	0.000	0.000
	HisToGene	418	0.102	0.107	0.191	0.057	0.019	0.002
	STNet	418	0.146	0.139	0.318	0.060	0.026	0.007
	STFormer	**418**	**0.172**	**0.148**	**0.397**	**0.108**	**0.017**	**0.005**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN123_A938797_Rep1	Hist2ST	418	0.031	0.030	0.012	0.000	0.000	0.000
	HisToGene	418	0.174	0.165	0.299	0.007	0.000	0.000
	STNet	418	0.279	0.258	0.844	0.402	0.050	0.000
	STFormer	**418**	**0.335**	**0.306**	**0.878**	**0.677**	**0.167**	**0.012**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN124_A938797_Rep2	Hist2ST	418	-0.046	-0.048	0.000	0.000	0.000	0.000
	HisToGene	418	-0.052	-0.025	0.074	0.045	0.024	0.002
	STNet	418	0.329	0.327	0.871	0.612	0.256	0.081
	STFormer	**418**	**0.261**	**0.243**	**0.701**	**0.380**	**0.158**	**0.048**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN048_A416371_Rep1	Hist2ST	418	0.124	0.116	0.005	0.000	0.000	0.000
	HisToGene	418	0.057	0.053	0.017	0.000	0.000	0.000
	STNet	418	0.196	0.175	0.476	0.026	0.000	0.000
	STFormer	**418**	**0.323**	**0.299**	**0.859**	**0.636**	**0.153**	**0.002**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN048_A416371_Rep2	Hist2ST	418	0.015	0.011	0.000	0.000	0.000	0.000
	HisToGene	418	0.150	0.121	0.330	0.022	0.000	0.000
	STNet	418	0.189	0.165	0.443	0.062	0.000	0.000
	STFormer	**418**	**0.269**	**0.245**	**0.780**	**0.304**	**0.000**	**0.000**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN123_A595688_Rep1	Hist2ST	418	0.056	0.055	0.000	0.000	0.000	0.000
	HisToGene	416	0.082	0.076	0.026	0.000	0.000	0.000
	STNet	418	0.291	0.271	0.821	0.457	0.022	0.000
	STFormer	**418**	**0.312**	**0.285**	**0.823**	**0.545**	**0.091**	**0.002**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000
SN124_A595688_Rep2	Hist2ST	418	0.048	0.050	0.067	0.005	0.000	0.000
	HisToGene	416	0.216	0.208	0.553	0.219	0.065	0.000
	STNet	418	0.339	0.320	0.809	0.596	0.309	0.060
	STFormer	**418**	**0.397**	**0.374**	**0.871**	**0.734**	**0.493**	**0.237**
	GroundTruth	418	1.000	1.000	1.000	1.000	1.000	1.000

图 3: Internal 14-CRC summary table of model performance.

External summaries recapitulated this ranking, supporting model robustness under distribution shift and independent sample characteristics.

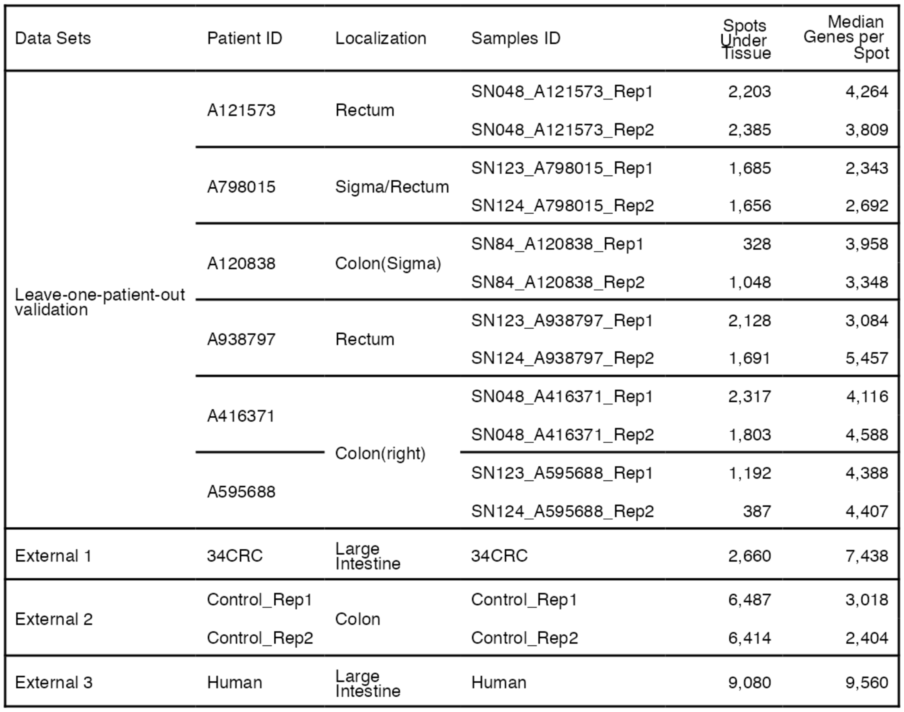


Samples ID	models	Target Gene Number	Median correlation	Mean correlation	Ratio of correlation ≥ 0.20	Ratio of correlation ≥ 0.30	Ratio of correlation ≥ 0.40	Ratio of correlation ≥ 0.50
34CRC	Hist2ST	275	-0.023	-0.022	0.000	0.000	0.000	0.000
	HisToGene	275	0.060	0.060	0.029	0.000	0.000	0.000
	STNet	275	0.009	0.013	0.011	0.000	0.000	0.000
	STFormer	**275**	**0.190**	**0.166**	**0.487**	**0.127**	**0.000**	**0.000**
	GroundTruth	275	1.000	1.000	1.000	1.000	1.000	1.000
Control_Rep1	Hist2ST	296	0.236	0.197	0.598	0.314	0.041	0.000
	HisToGene	296	0.199	0.161	0.500	0.142	0.003	0.000
	STNet	296	0.238	0.173	0.588	0.345	0.118	0.003
	STFormer	**296**	**0.403**	**0.342**	**0.740**	**0.669**	**0.503**	**0.334**
	GroundTruth	296	1.000	1.000	1.000	1.000	1.000	1.000
Control_Rep2	Hist2ST	296	0.203	0.180	0.514	0.247	0.044	0.000
	HisToGene	296	0.146	0.129	0.314	0.054	0.000	0.000
	STNet	296	0.155	0.104	0.348	0.061	0.000	0.000
	STFormer	**296**	**0.395**	**0.330**	**0.740**	**0.669**	**0.490**	**0.301**
	GroundTruth	296	1.000	1.000	1.000	1.000	1.000	1.000
Human	Hist2ST	296	-0.003	-0.003	0.000	0.000	0.000	0.000
	HisToGene	296	0.281	0.206	0.611	0.446	0.236	0.027
	STNet	296	0.160	0.126	0.402	0.132	0.017	0.000
	STFormer	**296**	**0.219**	**0.158**	**0.527**	**0.291**	**0.068**	**0.003**
	GroundTruth	296	1.000	1.000	1.000	1.000	1.000	1.000

图 4: External cohort summary table of model performance.

3.3 Clinical and biological context of inferred expression

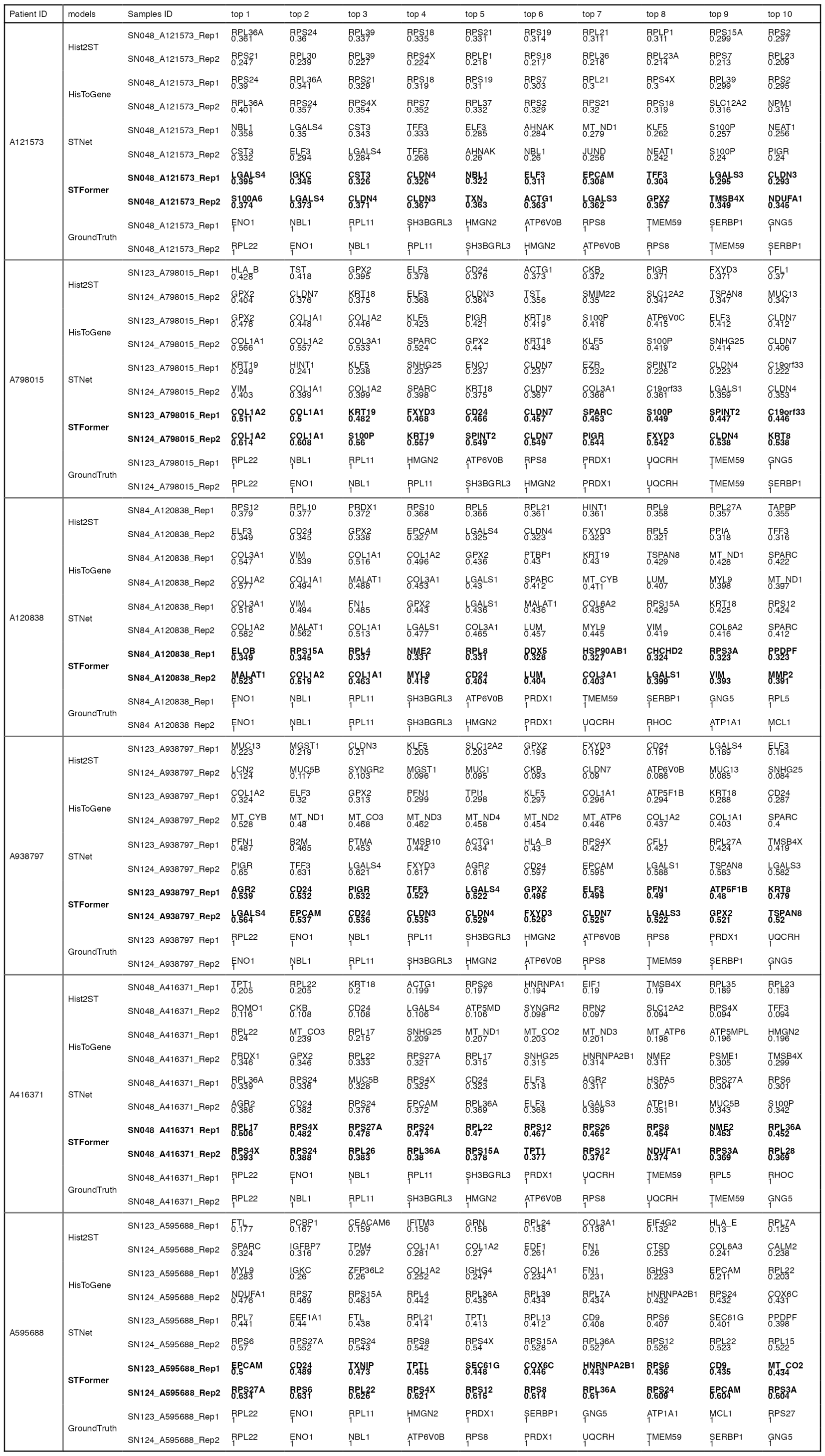
Clinical composition across cohorts (dataset source, localization, and spot-level sequencing depth surrogates) provides context for the observed variation in model behavior.



Data Sets	Patient ID	Localization	Samples ID	Spots Under Tissue	Median Genes per Spot
Leave-one-patient-out validation	A121573	Rectum	SN048_A121573_Rep1	2,203	4,264
			SN048_A121573_Rep2	2,385	3,809
	A798015	Sigma/Rectum	SN123_A798015_Rep1	1,685	2,343
			SN124_A798015_Rep2	1,656	2,692
	A120838	Colon(Sigma)	SN84_A120838_Rep1	328	3,958
			SN84_A120838_Rep2	1,048	3,348
	A938797	Rectum	SN123_A938797_Rep1	2,128	3,084
			SN124_A938797_Rep2	1,691	5,457
	A416371	Colon(right)	SN048_A416371_Rep1	2,317	4,116
			SN048_A416371_Rep2	1,803	4,588
	A595688		SN123_A595688_Rep1	1,192	4,388
			SN124_A595688_Rep2	387	4,407
External 1	34CRC	Large Intestine	34CRC	2,660	7,438
External 2	Control_Rep1	Colon	Control_Rep1	6,487	3,018
	Control_Rep2		Control_Rep2	6,414	2,404
External 3	Human	Large Intestine	Human	9,080	9,560

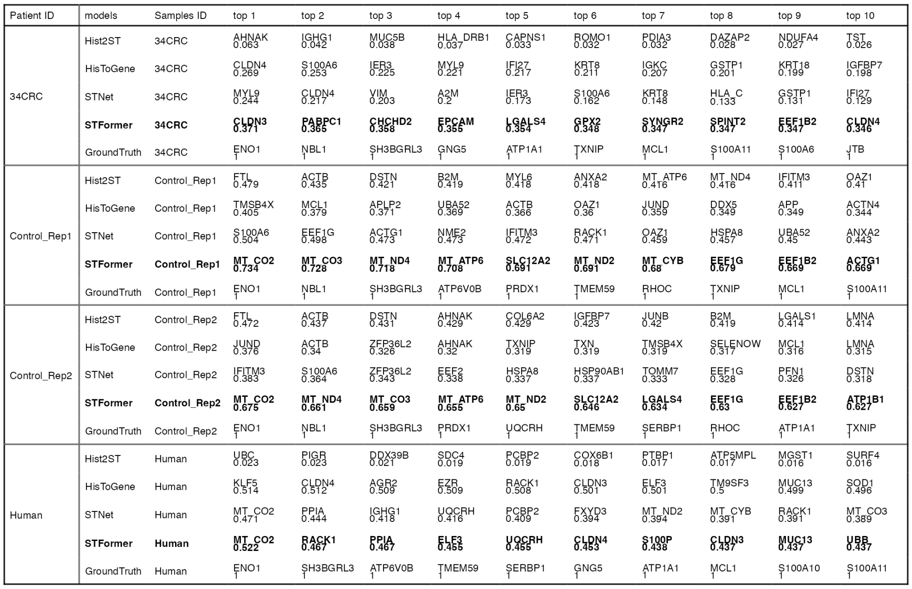
图 5: Clinical characteristics across internal and external cohorts.

At the gene level, top-ranked concordant genes remained biologically coherent across internal and external sets, indicating that recovered signals are not dominated by idiosyncratic sample artifacts.



Patient ID	models	Samples ID	top 1	top 2	top 3	top 4	top 5	top 6	top 7	top 8	top 9	top 10
A121573	Hist2ST	SN048_A121573_Rep1	RPL36A 0.361	RPS24 0.36	RPL39 0.337	RPS18 0.335	RPS21 0.331	RPS19 0.314	RPL21 0.311	RPLP1 0.311	RPS15A 0.299	RPS2 0.297
		SN048_A121573_Rep2	RPS21 0.247	RPL30 0.239	RPL39 0.227	RPS4X 0.224	RPLP1 0.218	RPS18 0.217	RPL36 0.216	RPL23A 0.214	RPS7 0.213	RPL23 0.209
	HisToGene	SN048_A121573_Rep1	RPS24 0.39	RPL36A 0.341	RPS21 0.329	RPS18 0.319	RPS19 0.31	RPS7 0.303	RPL21 0.3	RPS4X 0.3	RPL39 0.299	RPS2 0.295
		SN048_A121573_Rep2	RPL36A 0.401	RPS24 0.357	RPS4X 0.354	RPS7 0.352	RPL37 0.332	RPS2 0.329	RPS21 0.32	RPS18 0.319	SLC12A2 0.316	NPM1 0.315
	STNet	SN048_A121573_Rep1	NBL1 0.358	LGALS4 0.35	CST3 0.343	TFF3 0.333	ELF3 0.285	AHNAK 0.284	MT_ND1 0.279	KLF5 0.262	S100P 0.257	NEAT1 0.256
		SN048_A121573_Rep2	CST3 0.332	ELF3 0.294	LGALS4 0.284	TFF3 0.266	AHNAK 0.26	NBL1 0.26	JUND 0.256	NEAT1 0.242	S100P 0.24	PIGR 0.24
	STFormer	**SN048_A121573_Rep1**	**LGALS4 0.395**	**IGKC 0.345**	**CST3 0.326**	**CLDN4 0.326**	**NBL1 0.322**	**ELF3 0.311**	**EPCAM 0.308**	**TFF3 0.304**	**LGALS3 0.295**	**CLDN3 0.293**
		SN048_A121573_Rep2	**S100A6 0.374**	**LGALS4 0.373**	**CLDN4 0.371**	**CLDN3 0.367**	**TXN 0.363**	**ACTG1 0.363**	**LGALS3 0.362**	**GPX2 0.357**	**TMSB4X 0.349**	**NDUFA1 0.345**
	GroundTruth	SN048_A121573_Rep1	ENO1 1	NBL1 1	RPL11 1	SH3BGRL3 1	HMGN2 1	ATP6V0B 1	RPS8 1	TMEM59 1	SERBP1 1	GNG5 1
		SN048_A121573_Rep2	RPL22 1	ENO1 1	NBL1 1	RPL11 1	SH3BGRL3 1	HMGN2 1	ATP6V0B 1	RPS8 1	TMEM59 1	SERBP1 1
A798015	Hist2ST	SN123_A798015_Rep1	HLA_B 0.428	TST 0.418	GPX2 0.395	ELF3 0.378	CD24 0.376	ACTG1 0.373	CKB 0.372	PIGR 0.371	FXYD3 0.371	CFL1 0.37
		SN124_A798015_Rep2	GPX2 0.404	CLDN7 0.376	KRT18 0.375	ELF3 0.368	CLDN3 0.364	TST 0.356	SMIM22 0.35	SLC12A2 0.347	TSPAN8 0.347	MUC13 0.347
	HisToGene	SN123_A798015_Rep1	GPX2 0.478	COL1A1 0.448	COL1A2 0.446	KLF5 0.423	PIGR 0.421	KRT18 0.419	S100P 0.416	ATP6V0C 0.415	ELF3 0.412	CLDN7 0.412
		SN124_A798015_Rep2	COL1A1 0.566	COL1A2 0.557	COL3A1 0.533	SPARC 0.524	GPX2 0.44	KRT18 0.434	KLF5 0.43	S100P 0.419	SNHG25 0.414	CLDN7 0.406
	STNet	SN123_A798015_Rep1	KRT19 0.249	HINT1 0.241	KLF5 0.238	SNHG25 0.237	ENO1 0.237	CLDN7 0.237	EZR 0.232	SPINT2 0.226	CLDN4 0.223	C19orf33 0.222
		SN124_A798015_Rep2	VIM 0.403	COL1A1 0.399	COL1A2 0.399	SPARC 0.398	KRT18 0.375	CLDN7 0.367	COL3A1 0.366	C19orf33 0.361	LGALS1 0.359	CLDN4 0.353
	STFormer	**SN123_A798015_Rep1**	**COL1A2 0.511**	**COL1A1 0.5**	**KRT19 0.482**	**FXYD3 0.468**	**CD24 0.466**	**CLDN7 0.457**	**SPARC 0.453**	**S100P 0.449**	**SPINT2 0.447**	**C19orf33 0.446**
		SN124_A798015_Rep2	**COL1A2 0.614**	**COL1A1 0.608**	**S100P 0.56**	**KRT19 0.557**	**SPINT2 0.549**	**CLDN7 0.549**	**PIGR 0.544**	**FXYD3 0.542**	**CLDN4 0.538**	**KRT8 0.538**
	GroundTruth	SN123_A798015_Rep1	RPL22 1	NBL1 1	RPL11 1	HMGN2 1	ATP6V0B 1	RPS8 1	PRDX1 1	UQCRH 1	TMEM59 1	GNG5 1
		SN124_A798015_Rep2	RPL22 1	ENO1 1	NBL1 1	RPL11 1	SH3BGRL3 1	HMGN2 1	PRDX1 1	UQCRH 1	TMEM59 1	SERBP1 1
A120838	Hist2ST	SN84_A120838_Rep1	RPS12 0.379	RPL10 0.377	PRDX1 0.372	RPS10 0.368	RPL5 0.366	RPL21 0.361	HINT1 0.361	RPL9 0.358	RPL27A 0.357	TAPBP 0.355
		SN84_A120838_Rep2	ELF3 0.349	CD24 0.345	GPX2 0.338	EPCAM 0.327	LGALS4 0.325	CLDN4 0.323	FXYD3 0.323	RPL5 0.321	PPIA 0.318	TFF3 0.316
	HisToGene	SN84_A120838_Rep1	COL3A1 0.547	VIM 0.539	COL1A1 0.516	COL1A2 0.496	GPX2 0.436	PTBP1 0.43	KRT19 0.43	TSPAN8 0.429	MT_ND1 0.428	SPARC 0.422
		SN84_A120838_Rep2	COL1A2 0.577	COL1A1 0.494	MALAT1 0.488	COL3A1 0.453	LGALS1 0.43	SPARC 0.412	MT_CYB 0.411	LUM 0.407	MYL9 0.398	MT_ND1 0.397
	STNet	SN84_A120838_Rep1	COL3A1 0.518	VIM 0.494	FN1 0.485	GPX2 0.443	LGALS1 0.436	MALAT1 0.436	COL6A2 0.435	RPS15A 0.429	KRT18 0.425	RPS12 0.424
		SN84_A120838_Rep2	COL1A2 0.582	MALAT1 0.562	COL1A1 0.513	LGALS1 0.477	COL3A1 0.465	LUM 0.457	MYL9 0.445	VIM 0.419	COL6A2 0.416	SPARC 0.412
	STFormer	**SN84_A120838_Rep1**	**ELOB 0.349**	**RPS15A 0.345**	**RPL4 0.337**	**NME2 0.331**	**RPL8 0.331**	**DDX5 0.328**	**HSP90AB1 0.327**	**CHCHD2 0.324**	**RPS3A 0.323**	**PPDPF 0.323**
		SN84_A120838_Rep2	**MALAT1 0.523**	**COL1A2 0.519**	**COL1A1 0.463**	**MYL9 0.415**	**CD24 0.404**	**LUM 0.404**	**COL3A1 0.403**	**LGALS1 0.399**	**VIM 0.393**	**MMP2 0.391**
	GroundTruth	SN84_A120838_Rep1	ENO1 1	NBL1 1	RPL11 1	SH3BGRL3 1	ATP6V0B 1	PRDX1 1	TMEM59 1	SERBP1 1	GNG5 1	RPL5 1
		SN84_A120838_Rep2	ENO1 1	NBL1 1	RPL11 1	SH3BGRL3 1	HMGN2 1	PRDX1 1	UQCRH 1	RHOC 1	ATP1A1 1	MCL1 1
A938797	Hist2ST	SN123_A938797_Rep1	MUC13 0.223	MGST1 0.219	CLDN3 0.21	KLF5 0.205	SLC12A2 0.203	GPX2 0.198	FXYD3 0.192	CD24 0.191	LGALS4 0.189	ELF3 0.184
		SN124_A938797_Rep2	LCN2 0.124	MUC5B 0.117	SYNGR2 0.103	MGST1 0.096	MUC1 0.095	CKB 0.093	CLDN7 0.09	ATP6V0B 0.086	MUC13 0.085	SNHG25 0.084
	HisToGene	SN123_A938797_Rep1	COL1A2 0.324	ELF3 0.32	GPX2 0.313	PFN1 0.299	TPI1 0.298	KLF5 0.297	COL1A1 0.296	ATP5F1B 0.294	KRT18 0.288	CD24 0.287
		SN124_A938797_Rep2	MT_CYB 0.528	MT_ND1 0.48	MT_CO3 0.468	MT_ND3 0.462	MT_ND4 0.458	MT_ND2 0.454	MT_ATP6 0.446	COL1A2 0.437	COL1A1 0.403	SPARC 0.4
	STNet	SN123_A938797_Rep1	PFN1 0.487	B2M 0.465	PTMA 0.453	TMSB10 0.442	ACTG1 0.434	HLA_B 0.43	RPS4X 0.427	CFL1 0.427	RPL27A 0.424	TMSB4X 0.419
		SN124_A938797_Rep2	PIGR 0.65	TFF3 0.631	LGALS4 0.621	FXYD3 0.617	AGR2 0.616	CD24 0.597	EPCAM 0.595	LGALS1 0.588	TSPAN8 0.583	LGALS3 0.582
	STFormer	**SN123_A938797_Rep1**	**AGR2 0.539**	**CD24 0.532**	**PIGR 0.532**	**TFF3 0.527**	**LGALS4 0.522**	**GPX2 0.495**	**ELF3 0.495**	**PFN1 0.49**	**ATP5F1B 0.48**	**KRT8 0.479**
		SN124_A938797_Rep2	**LGALS4 0.564**	**EPCAM 0.537**	**CD24 0.536**	**CLDN3 0.535**	**CLDN4 0.529**	**FXYD3 0.526**	**CLDN7 0.525**	**LGALS3 0.522**	**GPX2 0.521**	**TSPAN8 0.52**
	GroundTruth	SN123_A938797_Rep1	RPL22 1	ENO1 1	NBL1 1	RPL11 1	SH3BGRL3 1	HMGN2 1	ATP6V0B 1	RPS8 1	PRDX1 1	UQCRH 1
		SN124_A938797_Rep2	ENO1 1	NBL1 1	RPL11 1	SH3BGRL3 1	HMGN2 1	ATP6V0B 1	RPS8 1	TMEM59 1	SERBP1 1	GNG5 1
A416371	Hist2ST	SN048_A416371_Rep1	TPT1 0.205	RPL22 0.205	KRT18 0.2	ACTG1 0.199	RPS26 0.197	HNRNPA1 0.194	EIF1 0.19	TMSB4X 0.19	RPL35 0.189	RPL23 0.189
		SN048_A416371_Rep2	ROMO1 0.116	CKB 0.108	CD24 0.108	LGALS4 0.106	ATP5MD 0.106	SYNGR2 0.098	RPN2 0.097	SLC12A2 0.094	RPS4X 0.094	TFF3 0.094
	HisToGene	SN048_A416371_Rep1	RPL22 0.24	MT_CO3 0.239	RPL17 0.215	SNHG25 0.209	MT_ND1 0.207	MT_CO2 0.203	MT_ND3 0.201	MT_ATP6 0.198	ATP5MPL 0.196	HMGN2 0.196
		SN048_A416371_Rep2	PRDX1 0.346	GPX2 0.346	RPL22 0.333	RPS27A 0.321	RPL17 0.315	SNHG25 0.315	HNRNPA2B1 0.314	NME2 0.311	PSME1 0.305	TMSB4X 0.299
	STNet	SN048_A416371_Rep1	RPL36A 0.339	RPS24 0.336	MUC5B 0.328	RPS4X 0.325	CD24 0.323	ELF3 0.318	AGR2 0.311	HSPA5 0.307	RPS27A 0.304	RPS6 0.301
		SN048_A416371_Rep2	AGR2 0.386	CD24 0.382	RPS24 0.376	EPCAM 0.372	RPL36A 0.369	ELF3 0.368	LGALS3 0.359	ATP1B1 0.351	MUC5B 0.343	S100P 0.342
	STFormer	**SN048_A416371_Rep1**	**RPL17 0.506**	**RPS4X 0.482**	**RPS27A 0.478**	**RPS24 0.474**	**RPL22 0.47**	**RPS12 0.467**	**RPS26 0.465**	**RPS8 0.454**	**NME2 0.453**	**RPL36A 0.452**
		SN048_A416371_Rep2	**RPS4X 0.393**	**RPS24 0.388**	**RPL26 0.383**	**RPL36A 0.38**	**RPS15A 0.378**	**TPT1 0.377**	**RPS12 0.376**	**NDUFA1 0.374**	**RPS3A 0.369**	**RPL28 0.369**
	GroundTruth	SN048_A416371_Rep1	RPL22 1	ENO1 1	NBL1 1	RPL11 1	SH3BGRL3 1	PRDX1 1	UQCRH 1	TMEM59 1	RPL5 1	RHOC 1
		SN048_A416371_Rep2	RPL22 1	NBL1 1	RPL11 1	SH3BGRL3 1	HMGN2 1	ATP6V0B 1	RPS8 1	UQCRH 1	TMEM59 1	GNG5 1
A595688	Hist2ST	SN123_A595688_Rep1	FTL 0.177	PCBP1 0.167	CEACAM6 0.159	IFITM3 0.156	GRN 0.156	RPL24 0.138	COL3A1 0.136	EIF4G2 0.132	HLA_E 0.13	RPL7A 0.125
		SN124_A595688_Rep2	SPARC 0.324	IGFBP7 0.316	TPM4 0.297	COL1A1 0.281	COL1A2 0.27	EDF1 0.261	FN1 0.26	CTSD 0.253	COL6A3 0.241	CALM2 0.238
	HisToGene	SN123_A595688_Rep1	MYL9 0.283	IGKC 0.26	ZFP36L2 0.26	COL1A2 0.252	IGHG4 0.247	COL1A1 0.234	FN1 0.231	IGHG3 0.223	EPCAM 0.211	RPL22 0.203
		SN124_A595688_Rep2	NDUFA1 0.476	RPS7 0.469	RPS15A 0.463	RPL4 0.442	RPL36A 0.435	RPL39 0.434	RPL7A 0.434	HNRNPA2B1 0.432	RPS24 0.432	COX6C 0.431
	STNet	SN123_A595688_Rep1	RPL7 0.441	EEF1A1 0.44	FTL 0.438	RPL21 0.414	TPT1 0.413	RPL13 0.412	CD9 0.408	RPS6 0.407	SEC61G 0.401	PPDPF 0.398
		SN124_A595688_Rep2	RPS6 0.57	RPS27A 0.552	RPS24 0.543	RPS8 0.542	RPS4X 0.54	RPS15A 0.528	RPL36A 0.527	RPS12 0.526	RPL22 0.523	RPL15 0.522
	STFormer	**SN123_A595688_Rep1**	**EPCAM 0.5**	**CD24 0.489**	**TXNIP 0.473**	**TPT1 0.455**	**SEC61G 0.448**	**COX6C 0.446**	**HNRNPA2B1 0.443**	**RPS6 0.436**	**CD9 0.435**	**MT_CO2 0.434**
		SN124_A595688_Rep2	**RPS27A 0.634**	**RPS6 0.631**	**RPL22 0.626**	**RPS4X 0.621**	**RPS12 0.615**	**RPS8 0.614**	**RPL36A 0.61**	**RPS24 0.609**	**EPCAM 0.604**	**RPS3A 0.604**
	GroundTruth	SN123_A595688_Rep1	RPL22 1	ENO1 1	RPL11 1	HMGN2 1	PRDX1 1	SERBP1 1	GNG5 1	ATP1A1 1	MCL1 1	RPS27 1
		SN124_A595688_Rep2	RPL22 1	ENO1 1	NBL1 1	ATP6V0B 1	RPS8 1	PRDX1 1	UQCRH 1	TMEM59 1	SERBP1 1	GNG5 1

图 6: Top 10 concordant genes per sample in internal 14-CRC cohort.



Patient ID	models	Samples ID	top 1	top 2	top 3	top 4	top 5	top 6	top 7	top 8	top 9	top 10
34CRC	Hist2ST	34CRC	AHNAK 0.063	IGHG1 0.042	MUC5B 0.038	HLA_DRB1 0.037	CAPNS1 0.033	ROMO1 0.032	PDIA3 0.032	DAZAP2 0.028	NDUFA4 0.027	TST 0.026
	HisToGene	34CRC	CLDN4 0.269	S100A6 0.253	IER3 0.225	MYL9 0.221	IFI27 0.217	KRT8 0.211	IGKC 0.207	GSTP1 0.201	KRT18 0.199	IGFBP7 0.198
	STNet	34CRC	MYL9 0.244	CLDN4 0.217	VIM 0.203	A2M 0.2	IER3 0.173	S100A6 0.162	KRT8 0.148	HLA_C 0.133	GSTP1 0.131	IFI27 0.129
	STFormer	**34CRC**	**CLDN3** **0.371**	**PABPC1** **0.365**	**CHCHD2** **0.358**	**EPCAM** **0.355**	**LGALS4** **0.354**	**GPX2** **0.348**	**SYNGR2** **0.347**	**SPINT2** **0.347**	**EEF1B2** **0.347**	**CLDN4** **0.346**
	GroundTruth	34CRC	ENO1 1	NBL1 1	SH3BGRL3 1	GNG5 1	ATP1A1 1	TXNIP 1	MCL1 1	S100A11 1	S100A6 1	JTB 1
Control_Rep1	Hist2ST	Control_Rep1	FTL 0.479	ACTB 0.435	DSTN 0.421	B2M 0.419	MYL6 0.418	ANXA2 0.418	MT_ATP6 0.416	MT_ND4 0.416	IFITM3 0.411	OAZ1 0.41
	HisToGene	Control_Rep1	TMSB4X 0.405	MCL1 0.379	APLP2 0.371	UBA52 0.369	ACTB 0.366	OAZ1 0.36	JUND 0.359	DDX5 0.349	APP 0.349	ACTN4 0.344
	STNet	Control_Rep1	S100A6 0.504	EEF1G 0.498	ACTG1 0.473	NME2 0.473	IFITM3 0.472	RACK1 0.471	OAZ1 0.459	HSPA8 0.457	UBA52 0.45	ANXA2 0.443
	STFormer	**Control_Rep1**	**MT_CO2** **0.734**	**MT_CO3** **0.728**	**MT_ND4** **0.718**	**MT_ATP6** **0.708**	**SLC12A2** **0.691**	**MT_ND2** **0.691**	**MT_CYB** **0.68**	**EEF1G** **0.679**	**EEF1B2** **0.669**	**ACTG1** **0.669**
	GroundTruth	Control_Rep1	ENO1 1	NBL1 1	SH3BGRL3 1	ATP6V0B 1	PRDX1 1	TMEM59 1	RHOC 1	TXNIP 1	MCL1 1	S100A11 1
Control_Rep2	Hist2ST	Control_Rep2	FTL 0.472	ACTB 0.437	DSTN 0.431	AHNAK 0.429	COL6A2 0.429	IGFBP7 0.423	JUNB 0.42	B2M 0.419	LGALS1 0.414	LMNA 0.414
	HisToGene	Control_Rep2	JUND 0.376	ACTB 0.34	ZFP36L2 0.326	AHNAK 0.32	TXNIP 0.319	TXN 0.319	TMSB4X 0.319	SELENOW 0.317	MCL1 0.316	LMNA 0.315
	STNet	Control_Rep2	IFITM3 0.383	S100A6 0.364	ZFP36L2 0.343	EEF2 0.338	HSPA8 0.337	HSP90AB1 0.337	TOMM7 0.333	EEF1G 0.328	PFN1 0.326	DSTN 0.318
	STFormer	**Control_Rep2**	**MT_CO2** **0.675**	**MT_ND4** **0.661**	**MT_CO3** **0.659**	**MT_ATP6** **0.655**	**MT_ND2** **0.65**	**SLC12A2** **0.646**	**LGALS4** **0.634**	**EEF1G** **0.63**	**EEF1B2** **0.627**	**ATP1B1** **0.627**
	GroundTruth	Control_Rep2	ENO1 1	NBL1 1	SH3BGRL3 1	PRDX1 1	UQCRH 1	TMEM59 1	SERBP1 1	RHOC 1	ATP1A1 1	TXNIP 1
Human	Hist2ST	Human	UBC 0.023	PIGR 0.023	DDX39B 0.021	SDC4 0.019	PCBP2 0.019	COX6B1 0.018	PTBP1 0.017	ATP5MPL 0.017	MGST1 0.016	SURF4 0.016
	HisToGene	Human	KLF5 0.514	CLDN4 0.512	AGR2 0.509	EZR 0.509	RACK1 0.508	CLDN3 0.501	ELF3 0.501	TM9SF3 0.5	MUC13 0.499	SOD1 0.496
	STNet	Human	MT_CO2 0.471	PPIA 0.444	IGHG1 0.418	UQCRH 0.416	PCBP2 0.409	FXYD3 0.394	MT_ND2 0.394	MT_CYB 0.391	RACK1 0.391	MT_CO3 0.389
	STFormer	**Human**	**MT_CO2** **0.522**	**RACK1** **0.467**	**PPIA** **0.467**	**ELF3** **0.455**	**UQCRH** **0.455**	**CLDN4** **0.453**	**S100P** **0.438**	**CLDN3** **0.437**	**MUC13** **0.437**	**UBB** **0.437**
	GroundTruth	Human	ENO1 1	SH3BGRL3 1	ATP6V0B 1	TMEM59 1	SERBP1 1	GNG5 1	ATP1A1 1	MCL1 1	S100A10 1	S100A11 1

图 7: Top 10 concordant genes per sample in external cohorts.

Collectively, these analyses indicate that WSI-inferred ST captures reproducible transcriptomic structure across datasets while preserving biologically interpretable gene-level patterns.

4 Discussion

The present analysis demonstrates that virtual ST from WSIs can reach reproducible concordance across both internal and external colorectal datasets, with a clear advantage for the top-performing model in this benchmark. Importantly, improvements were not restricted to a few marker genes but extended to distribution-level shifts in per-gene correlations and threshold-based quality metrics.

Several limitations remain. First, performance still varies by sample and tissue context, suggesting unresolved domain shifts in staining, cellular composition, or section quality. Second, correlation-based metrics do not fully capture pathway-level conservation or downstream clinical utility. Third, retrospective evaluation cannot substitute prospective deployment constraints.

Future work should include prospective multi-center validation, uncertainty-aware calibration at spot level, and integration with pathology annotation priors to improve robustness in low-signal regions.

5 Methods

5.1 Study Design

We analyzed internal leave-one-patient-out data and external colorectal datasets, using measured ST as reference and model-inferred ST as prediction.

5.2 Computational Workflow

The repository pipeline follows ordered scripts: 1. `1_Correlation_0705_Parallel.R`: computes correlation outputs. 2. `2-Prepare_gt_pre_csv_for_newh5_Parallel.R`: prepares matched GT/prediction matrices. 3. `3_newh5_from_csv_0707_Parallel.R`: builds intermediate `newh5` assets. 4. `4_spe_from_newh5.R`: constructs `SpatialExperiment` objects and visualization outputs. 5. `5_three_line_Table.R`: exports summary tables and top-gene reports.

5.3 Metrics and Reporting

Primary endpoint: gene-wise spot-level correlation between inferred and measured ST. We report median/mean correlation and the proportion of genes above thresholds (0.20, 0.30, 0.40, 0.50).

5.4 Statistical Notes

Results are descriptive and benchmark-oriented; figures summarize distributional trends across genes and samples.

6 Supplements

6.1 Supplementary Figures and Tables

- Correlation summaries and per-sample views are provided in `Figures/Correlations/`.
- Cohort-level and external summary tables are provided in `ThreeLineTable/`.

6.2 Reproducibility

All outputs reported in this manuscript are generated from scripts in the repository root and can be reproduced via `Rscript` execution in sequence.