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.

Training on in silico host-tissue mixtures reduces host attraction in RNA-seq-based tissue-of-origin classification of metastatic biopsies

Juwon Kang
Yonsei University
Junjeong Choi
junjeong@yonsei.ac.kr

Yonsei University

Research Article

Keywords: Cancer of unknown primary, Tissue of origin, Metastasis, Host attraction, Data augmentation, Feature standardization, RNA sequencing, Gene expression classifier, Preregistration

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11293314/v1

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

Additional Declarations: No competing interests reported.

Training on in silico host-tissue mixtures reduces host attraction in RNA-seq-based tissue-of-origin classification of metastatic biopsies

Title page

Authors: Juwon Kang (ORCID 0009-0009-2186-0038) and Junjeong Choi (ORCID 0000-0003-1339-593X)

Affiliations: College of Pharmacy, Yonsei Institute of Pharmaceutical Sciences, Yonsei University, Incheon, Republic of Korea (21983); both authors.

Corresponding author: Junjeong Choi

Email: Juwon Kang, juwonk@yonsei.ac.kr; Junjeong Choi, junjeong@yonsei.ac.kr

Abstract

Background

Transcriptome classifiers that infer the primary site of metastases are trained mostly on primary tumors, but metastatic biopsies also contain normal host-organ cells. Such classifiers can name the biopsied organ instead of the primary, an error we call host attraction. A previous method reduced errors in liver metastases by training a microRNA classifier on in silico tumor–liver mixtures. Using internally preregistered analyses, we tested whether mixtures with several host tissues reduce host attraction in a whole-transcriptome classifier.

Results

We trained logistic regression classifiers for 32 tumor types with or without in silico tumor–normal mixtures from ten host tissues (HostMix-TOO). In POG570, HostMix-TOO reduced host attraction from 22.8% to 2.4% of at-risk biopsies (n = 378; −20.4 percentage points, 95% CI −24.3 to −16.4) and increased top-1 accuracy from 68.2% to 75.6% (n = 512; 7.4 points, 95% CI 3.9 to 10.7). Non-inferiority of accuracy for biopsies from the organ of origin was not

shown (86.8% versus 80.2%, n = 91; one-sided 95% lower bound −12.1 points, margin −10). In six auxiliary RNA-seq cohorts, all three primary hypotheses were confirmed: host attraction fell from 20.4% to 5.6% (n = 427), top-1 accuracy rose from 52.7% to 69.8% (n = 729), mostly in one bladder cancer cohort, and accuracy for biopsies from the organ of origin was non-inferior (n = 315). Published classifiers SCOPE and CUP-AI-Dx showed host attraction in 12.2–30.4% of at-risk biopsies (HostMix-TOO: 2.4–5.6%), although CUP-AI-Dx was more accurate in auxiliary cohorts. In five microarray cohorts, HostMix-TOO reduced host attraction but lowered top-1 accuracy from 72.4% to 45.5% (n = 536). A post hoc ablation attributed the host-attraction reduction to the mixtures and the microarray loss to feature standardization: without standardization, mixtures reduced microarray host attraction from 25.6% to 9.2% (n = 207) with 77.6% top-1 accuracy.

Conclusions

For RNA-seq biopsies, training on in silico host-tissue mixtures reduced host attraction and improved accuracy in two preregistered evaluations. Post hoc analyses traced the microarray loss to feature standardization, which requires confirmation. Benchmarks of tissue-of-origin classifiers on metastatic biopsies should report host attraction, with its denominator, alongside top-1 accuracy.

Keywords

Cancer of unknown primary; Tissue of origin; Metastasis; Host attraction; Data augmentation; Feature standardization; RNA sequencing; Gene expression classifier; Preregistration

Background

Cancers of unknown primary present as metastases without an identifiable primary tumor, and their management depends in part on the presumed tissue of origin [1]. Molecular classifiers have therefore been developed to suggest the tissue of origin from gene expression [2, 3, 4, 5, 6], somatic mutations [7, 8] or histology images [9]. Expression classifiers are usually trained

on large collections of primary tumors, such as The Cancer Genome Atlas (TCGA), and then applied to biopsies of metastases.

A biopsy of a metastasis is not a pure sample of the primary tumor. It contains normal cells of the organ in which the metastasis grows, such as hepatocytes in a liver biopsy or alveolar cells in a lung biopsy, and the tumor cells themselves may take on expression features of the host tissue [10, 11]. A classifier that has learned from primary tumors that hepatocyte transcripts indicate a liver cancer can therefore call a colorectal cancer metastasis in the liver a liver cancer. We call this error host attraction: the classifier predicts an organ that is native to the biopsy site, although the true primary lies elsewhere. Host attraction is hidden in overall accuracy, which pools biopsies at risk of this error with biopsies from lymph nodes, from the organ of origin, or from sites where the tumor types in question rarely arise. To our knowledge, evaluations of tissue-of-origin classifiers have not reported it as a separate quantity with an explicit denominator.

Several responses to host signal have been described. SCOPE adds normal-tissue classes to its tumor classes, so that a profile dominated by normal tissue can be assigned to a normal class [2]. Developmental deconvolution represents each tumor by the estimated contributions of developmental cell types and classifies the origin from that representation [5]. MetGen generates synthetic metastatic profiles from primary-tumor and normal-tissue profiles with a generative model trained with metastatic samples [12]. Most directly, Vincent et al. modeled tissue contamination by training a microRNA classifier on in silico mixtures of primary-tumor and benign-liver profiles; this lowered the test error on liver-metastasis biopsies from 77% to 45%, and to 34% when such biopsies were also used for training [13].

The mixture idea has not, to our knowledge, been evaluated for whole-transcriptome classifiers, for several host tissues at once, or with host attraction as a pre-specified endpoint. Here we define host attraction and its denominator for any biopsy site; train a whole-

transcriptome classifier on in silico mixtures of TCGA primary tumors with normal profiles from ten host tissues, a procedure we call HostMix-TOO; compare this procedure with alternatives on simulated mixtures and on a development cohort; test it against an otherwise similar classifier in two internally preregistered confirmations, POG570 and six further RNA-seq cohorts; describe host attraction in two published classifiers; and examine, after the confirmations, which part of the training change is responsible for the reduction and for the losses, and where the remaining errors go. We present mixture augmentation as an extension of the contamination model of Vincent et al., not as a new idea.

Methods

Study design and internal preregistration

The study had four stages (Fig. 1c). First, we assembled the data, fixed the gene universe, trained a baseline classifier and measured host attraction. Second, we compared training procedures that address host signal on simulated tumor–host mixtures and on MET500, which served as the development cohort, and locked one procedure, HostMix-TOO, by a pre-specified rule. Third, HostMix-TOO was compared with the baseline in two confirmatory evaluations: POG570, and a family of six RNA-seq cohorts collected after the POG570 confirmation. Fourth, both classifiers were compared with two published classifiers, and post hoc analyses were carried out. Supplementary methods, tables and figures are given in Additional file 1, and the source data for all figures and tables in Additional file 2.

Before each confirmatory evaluation, the hypotheses, analysis sets, test statistics, decision rules and random seeds were written to a plan file and committed to a git repository, together with SHA-256 hashes of the model files, label-mapping files and analysis code. The confirmatory code joined labels to predictions only after a separate unlock commit. Before unlocking, the confirmatory code of both confirmations was run on synthetic labels. For the auxiliary confirmation it was also run on planted predictions with known expected results, and

an independent script that did not import the analysis code recomputed every test statistic (274 comparisons without a mismatch); no independent recomputation was recorded for the POG570 confirmation. We call this procedure internal preregistration. It was not deposited in a public registry. Commit times were recorded by git on the analysis workstation and were not certified by a third party; the commit hashes fix the order of commits but do not independently prove their times. The commit history is listed in Additional file 3. The git history is not public because some committed files list POG570 patient identifiers with their tumor labels or per-patient results; a copy of the history without the contents of these files, in which every commit hash can still be verified, is available to editors and reviewers on request (Availability of data and materials).

Each analysis is labeled as preregistered (a confirmatory hypothesis test), pre-specified descriptive (named in a plan committed before its results existed, without a hypothesis test), exploratory (run during development or between the two confirmations) or post hoc (run after both confirmations).

Three deviations from the plans occurred. The first unlocked run of the auxiliary confirmation stopped before writing any hypothesis table because empty biopsy-site fields were not handled; after a fix that changed only the handling of empty fields, the run was repeated, and the second run is reported. Eighteen GSE50760 liver-metastasis samples that belong to the auxiliary at-risk set had been omitted from the input matrices hashed before the published classifiers were run; they were scored afterwards so that the at-risk denominator did not change. One classifier named in the comparison plan [6] was not run because no trained model or gene list was publicly available. Minor deviations in file paths are listed in Additional file 3.

Data

Table 1 lists all data sets and their roles.

Training data. Gene-level expression of TCGA tumors and Genotype-Tissue Expression (GTEx) normal tissues, uniformly processed by the Toil pipeline [14], was downloaded from the UCSC Xena Toil hub [15, 16]. The matrix is distributed as log2(TPM + 0.001), and we converted it to linear transcripts per million (TPM) as $\max(2^x - 0.001, 0)$. We kept one primary tumor per patient (TCGA sample type 01; type 03 for acute myeloid leukemia), 9358 tumors from 33 projects. Colon and rectal adenocarcinoma were merged, which gave 32 tumor labels. The tumors were split 80:20 into TCGA-train (n = 7486) and TCGA-test (n = 1872), stratified by label, so that each patient was in one part only. TCGA metastatic samples (sample type 06) from patients not in TCGA-train (n = 373) were used only descriptively. GTEx samples [17] from tissue donors (n = 7792 after removal of 70 cell-line samples without a donor) were split by donor within each tissue into GTEx-ref (n = 3881), used for training mixtures and reference profiles, and GTEx-sim (n = 3911), used only to simulate mixtures for evaluation.

Development cohort. MET500 [18] expression (fragments per kilobase per million, FPKM) was downloaded from the UCSC Xena public hub [19]. Libraries were collapsed to one per biopsy specimen, preferring poly(A) over exome-capture libraries. Specimens whose cancer type could not be mapped to a TCGA label (n = 59) were excluded, leaving 437 biopsies.

Confirmation cohort 1. POG570 [20, 21] comprises TPM expression profiles of 570 advanced cancers, one per patient, released by the BC Cancer Genome Sciences Centre. The expression file was hashed at download. The analysis code returns that matrix only when called with an explicit unlock, and that call was made only after the POG570 analysis plan had been committed. Fifty-eight samples whose diagnosis could not be mapped to an organ group were excluded, leaving 512.

Confirmation cohort 2 and descriptive cohorts. After the POG570 confirmation, we searched cBioPortal [22, 23, 24] and the Gene Expression Omnibus (GEO) [25, 26] for public

expression data of metastatic biopsies with an annotated biopsy site (search strategy in Additional file 1). Six RNA-seq cohorts formed the auxiliary confirmatory family: IMvigor210 urothelial carcinoma [27, 28], a metastatic triple-negative breast cancer trial cohort [29, 30], DFCI melanoma [31, 32], the PRINCE metastatic pancreatic cancer trial [33, 34], primary colorectal cancers with matched liver metastases (GSE50760) [35, 36] and SU2C/PCF metastatic castration-resistant prostate cancer [37, 38, 39]. A seventh RNA-seq cohort, AURORA US metastatic breast cancer (GSE209998) [40, 41], was given a descriptive role before the auxiliary plan was committed, because its sample identifiers could not be reconciled with the published number of patients (53 identifier prefixes versus 55 patients). Five microarray cohorts — colorectal cancer (GSE41258) [42, 43], breast cancer (GSE14018) [44, 45], pancreatic cancer (GSE71729) [46, 47] and two prostate cancer cohorts (GSE74685 [48, 49] and FHCRC [50, 51]) — formed a pre-specified descriptive layer.

Processing. RNA-seq values were mapped to gene symbols and converted to a linear scale summing to 10^6 per sample: FPKM values were rescaled, counts were divided by GENCODE v23 [52] exon-union gene length before rescaling, and log-scale values were back-transformed. For PRINCE, which provides log2-transformed normalized values, we used $2^v - 1$ without length normalization. The SCOPE input for PRINCE was built separately and was divided by GENCODE v23 exon-union length (Additional file 1). Microarray probes were mapped to genes with the platform annotation; probes matching more than one gene were removed, and the probe with the highest mean signal was kept for each gene. For the two-channel arrays GSE71729 and GSE74685, the sample channel was used and the reference channel was not. Within each analysis set, one sample per patient was used (for SU2C, poly(A) before capture libraries). To exclude samples that might duplicate MET500 or POG570 samples, every RNA-seq sample was compared with all MET500 and POG570 samples by the Spearman correlation of log2(TPM + 1) values; a sample was excluded when

its maximum external correlation was at least 0.98, or at least 0.95 and higher than its maximum correlation with other patients of its own cohort. This rule removed 76 SU2C samples and no sample from the other cohorts.

Tumor labels, organ groups and biopsy sites

Classifiers predicted the 32 tumor labels. Labels were mapped to 26 organ groups (for example, lung adenocarcinoma and lung squamous cell carcinoma to lung, and the three kidney cancer types to kidney; Additional file 1: Table S1), and the probability of an organ was the sum of the probabilities of its labels. All accuracies refer to the organ with the largest summed probability (top-1). The true organ of each evaluation sample was obtained from the cohort's diagnosis annotation with the same mapping.

Raw biopsy-site annotations were mapped to standard sites with a dictionary fixed before each confirmation. For each standard site we defined a native organ set: the organ groups whose primary tumors arise at that site (Additional file 1: Table S2). For example, the native set of liver is {liver, biliary tract}, of lung {lung, mesothelioma}, of lymph node {lymphoid neoplasms} and of bone marrow {myeloid neoplasms, lymphoid neoplasms}. Biopsies from sites without a native set were not part of any at-risk set.

Genes and features

The gene universe G consisted of the 17552 HGNC protein-coding symbols [53] present in the TCGA/GTEx Toil matrix, the MET500 matrix and the POG570 matrix after Ensembl identifiers were mapped to symbols with GENCODE v23 annotation. Within each sample, the expression values of the genes in G were converted to rank-normal scores, $z_g = \Phi^{-1}((r_g - 0.5)/|G|)$, where r_g is the rank of gene g in that sample, tied values receive their average rank, and Φ^{-1} is the standard normal quantile function. Because these scores depend only on ranks within a sample, they are unaffected by library size and by monotone transformations of the expression scale. The classifiers used the 5000 genes with the highest variance of these rank-

normal scores across TCGA-train, selected once before any model was fitted. For microarray profiles, ranks were computed over the genes measured on the platform, and classifier genes absent from the platform were set to 0, the center of the rank-normal scale.

Baseline classifier and HostMix-TOO

Both classifiers are multinomial logistic regressions with an L2 penalty, class weights inversely proportional to class frequency and the L-BFGS solver (at most 5000 iterations), as implemented in scikit-learn [54]. The baseline classifier was fitted to the 7486 TCGA-train tumors on unstandardized scores. Its inverse regularization strength, C = 0.1, was chosen by five-fold cross-validation on TCGA-train, stratified by label with one sample per patient (macro-averaged F1; candidates 0.01, 0.1, 1 and 10; the smaller value on ties). HostMix-TOO was fitted to the same tumors plus four mixtures per tumor (below), with each feature standardized to zero mean and unit variance on the augmented training matrix. Its C = 0.03 was chosen by three-fold cross-validation in which the pure and mixed profiles of a tumor were kept in the same fold (macro-averaged F1 on held-out folds, mixtures included; candidates 0.03, 0.1 and 0.3; the smallest C within 10^{-12} of the best mean). The baseline had been fixed before mixture augmentation was developed, so the two classifiers differ in feature standardization and C as well as in their training data; a post hoc ablation separated these factors.

Host pool and mixtures

The host pool consisted of ten GTEx tissues standing for common metastatic sites (Additional file 1: Table S3): liver, lung, brain cortex, adrenal gland, skin (not sun-exposed), skeletal muscle, subcutaneous adipose tissue and visceral adipose tissue (omentum), plus spleen as a proxy for lymph node and whole blood as a proxy for bone marrow; together 1103 GTEx-ref samples. GTEx does not include lymph node or bone marrow. Each proxy was the candidate GTEx tissue (spleen, whole blood, EBV-transformed lymphocytes or terminal ileum) whose

mean log2(TPM + 1) profile had the highest Spearman correlation with the Human Protein Atlas consensus profile of the target tissue [55] (0.881 for spleen and lymph node; 0.883 for whole blood and bone marrow).

For each TCGA-train tumor, four mixtures were generated. The tumor profile t and a host profile h were restricted to G, converted to linear TPM and rescaled to sum to 10^6, and the mixture was $x = \rho t + (1 - \rho)h$, with ρ drawn from a uniform distribution on [0.15, 1.0) separately for each mixture. The host profile was one GTEx-ref sample, chosen by drawing one of the ten host-pool tissues with equal probability and then one sample of that tissue with equal probability. Rank-normal scores were computed after mixing, and each mixture received the label of its tumor. The training matrix thus had 37430 rows (7486 pure tumors and 29944 mixtures). The mixing weight ρ is the tumor fraction of RNA, not of cells. Random seeds are listed in Additional file 1.

Alternatives compared during development

During development we compared HostMix-TOO with other ways of handling host signal (Additional file 1: Table S4). (i) Normal classes: the 1103 host-pool samples were added to the training data as ten normal-tissue classes, whose probabilities were discarded at prediction. (ii) Site-specific mixtures: nine models, each trained with mixtures drawn only from the host tissue or tissues of one standard site and applied according to the biopsy site, with HostMix-TOO for sites without a model. (iii) Linear deconvolution: the host component was estimated by weighted non-negative least squares against the mean profile of the site's host tissue and the 32 tumor-label centroids, subtracted, and the baseline was applied to the remainder; the estimated host fraction β was recorded. (iv) Native-organ masking: the probabilities of the native organs of the biopsy site were set to zero. (v) Removal of the 20% of genes whose scores changed most under host mixing, alone or combined with site-specific mixtures (20% or 40% removed). (vi) Gene-set scores: Kolmogorov–Smirnov enrichment

scores of 8267 MSigDB gene sets [56] from the ONCOfind resource [57] in place of single-gene scores, without mixtures, with mixtures, or with a fixed-fraction host correction. The pre-specified rule locked the new training procedure with the highest MET500 top-1 accuracy, breaking ties by the lower host-attraction rate; normal classes, deconvolution and masking served as comparators and were not eligible.

Between the two confirmations we examined exploratory variants of HostMix-TOO (Additional file 1: Table S5): training on pure tumors only with HostMix-TOO's standardization, one or eight mixtures per tumor, lower bounds of ρ of 0.3 and 0.5, C of 0.01 and 0.1, host pools of three tissues (liver, lung and spleen) and of 22 tissues (the ten pool tissues and twelve tissues outside the pool), ten leave-one-host-out pools, and a multilayer perceptron in place of logistic regression. Two of these defined the secondary auxiliary hypotheses: the 22-tissue model, and a gated model that used the baseline when the deconvolution host fraction β was at most 0.02 or undefined and HostMix-TOO otherwise, with the threshold chosen on MET500.

Host attraction and analysis sets

For each evaluation cohort we defined four analysis sets. The evaluation set contains biopsies with a mapped true organ and a standard site. The at-risk set contains biopsies whose site has a native organ set and whose true organ lies outside it. The native-truth set contains biopsies whose true organ lies in the native set of the site, that is, biopsies taken from the organ of origin. The pool-out at-risk set contains at-risk biopsies from sites whose host tissue is not in the host pool. A prediction is a host-attraction error when the predicted organ is in the native set of the biopsy site and differs from the true organ. The host-attraction rate is the number of host-attraction errors in the at-risk set divided by the size of the at-risk set. Host-attraction errors can also occur in the native-truth set when a site has more than one native organ (for

example, a cholangiocarcinoma biopsied in the liver and called liver cancer), but these do not enter the rate.

Simulation benchmark

From TCGA-test, 1000 tumors were drawn, stratified by label. Each tumor was mixed as described above with a randomly drawn GTEx-sim sample of each of 22 tissues at $\rho = 1.0$, 0.8, 0.6, 0.4 and 0.2: the ten host-pool tissues (in-pool) and twelve tissues outside the pool (pool-out: stomach, transverse colon, pancreas, kidney cortex, breast, ovary, prostate, thyroid, esophagus mucosa, minor salivary gland, ectocervix and bladder). Each tissue was assigned the standard site that it represents, and its host-attraction rate was computed over the tumors whose organ is outside the native set of that site. All models were scored on the same mixtures. Within each tissue, the GTEx-sim donors had not been used for training.

Confirmatory hypotheses and tests

POG570 family (one sample per patient; fixed-sequence testing; each test one-sided at $\alpha = 0.05$). H1: in the at-risk set, the host-attraction rate of HostMix-TOO is lower than that of the baseline. H2, tested only if H1 was rejected: in the evaluation set, the top-1 accuracy of HostMix-TOO is higher. H3, judged only if H2 was rejected: in the native-truth set, the top-1 accuracy of HostMix-TOO is not lower than that of the baseline by more than 0.10 (non-inferiority). H1 and H2 used exact one-sided McNemar tests on the discordant pairs. H3 was met if the one-sided 95% lower bound of the difference, the 5th percentile of its bootstrap distribution, exceeded −0.10. Differences are HostMix-TOO minus baseline, reported with two-sided 95% percentile intervals from 2000 bootstrap resamples of patients (seed 20261001).

Auxiliary families (six RNA-seq cohorts pooled; one sample per patient). The primary family, AH1–AH3, repeated H1–H3 with the same tests, margin and sequence. The secondary family was tested regardless of the primary family, with Holm correction at a one-sided family-wise

α of 0.05. AS1: the gated model has higher top-1 accuracy than HostMix-TOO in the native-truth set. AS2: the 22-tissue model has a lower host-attraction rate than HostMix-TOO in the pool-out at-risk set. AS3: the gated model has higher top-1 accuracy than HostMix-TOO in the evaluation set. Design power was computed for each hypothesis before unlocking (Additional file 1).

Pre-specified descriptive analyses

The plans also named descriptive outputs without hypothesis tests: in POG570, results by biopsy-site group and by tertile of the tumor content recorded in the POG570 metadata; for the auxiliary cohorts, leave-one-cohort-out estimates, a DerSimonian–Laird random-effects meta-analysis of cohort-level differences [58], results for GSE209998, and the comparisons of the primary family repeated in the microarray layer as effect sizes with intervals; and conformal prediction sets calibrated on MET500 (Additional file 2).

Comparison with published classifiers

SCOPE [2, 59] and CUP-AI-Dx [3, 60] were chosen because their trained models and prediction code are public. The comparison plan, including input rules, the mapping of each tool's classes to our organ groups and the treatment of SCOPE's normal-tissue classes, was committed before either tool was run on MET500, POG570 or the auxiliary cohorts, and the input matrices were hashed before the runs. Inputs were built from all genes in each source file following each tool's documentation (Additional file 1). Before scoring our cohorts, we reproduced the CUP-AI-Dx authors' example of metastatic samples (363 of 394 correct, 92.1%). SCOPE top-1 was the first-ranked class of its five-model ensemble; a call of a normal-tissue class was counted as incorrect, and a sensitivity analysis counted it as the corresponding organ. The comparison is descriptive: paired differences with bootstrap intervals are given in Additional file 2 but are not interpreted as tests. The original SCOPE report evaluated metastatic samples from the Personalized OncoGenomics program; because

the public repository does not list those identifiers, overlap with POG570 could not be checked, and no sample was removed. Any POG570 sample used during SCOPE development would bias the comparison in favor of SCOPE. Samples listed in the CUP-AI-Dx external metadata did not overlap our evaluation samples.

Post hoc analyses

Ablation. Because the baseline and HostMix-TOO differ in training data, feature standardization and C, we refitted three logistic models on the same features: pure tumors with standardization and C = 0.03; pure tumors with standardization and C = 0.15, which matches the total loss weight of the five-times larger augmented training set; and the HostMix-TOO mixtures without standardization and with C = 0.1 (hereafter the unstandardized mixture model). The plan, committed before any fit, defined for each cohort the gap g between the host-attraction rates of the baseline and HostMix-TOO and the share of g attributable to the mixtures: (standardized pure model − HostMix-TOO)/g for the two pure-tumor models and (baseline − unstandardized mixture model)/g for the mixture model. It stated that the mixtures would be considered the main source of the gap if all nine shares (three models in three cohorts) were at least 0.5. Before fitting the mixture model, we regenerated the HostMix-TOO training mixtures from the mixture records and refitted HostMix-TOO; the coefficients matched the locked model to within 1.05×10^{-14}, and its MET500 predictions were identical. Under a second plan, committed before any of these models was applied to microarray or gene-restricted profiles, we also scored them in the microarray layer and in the missing-gene simulation (below). For the loss of top-1 accuracy in the microarray evaluation set and for the loss under random gene removal, the plan defined the share of HostMix-TOO's loss reproduced by each model and attributed a loss to the mixtures or to standardization when the models with that factor reproduced at least half of it and the models without it reproduced less than half (Additional file 1).

Error analyses. Using the confirmatory predictions and the locked models only, we decomposed the results by cohort and analysis set, recorded how at-risk biopsies whose baseline host-attraction error was removed by HostMix-TOO were then classified, and computed leave-one-cohort-out intervals. To simulate missing genes, we restricted POG570 and MET500 profiles to the genes of each microarray platform (GPL96 [61], GPL20769 [62], GPL15659 [63] and the FHCRC Agilent array [51]), recomputed the scores with the microarray rule, and compared the change in top-1 accuracy with ten random removals of the same number of classifier genes (seed 20261023). For IMvigor210 bladder tumors in the native-truth set that were called esophagus, we computed the mean contribution of each gene to the difference between the esophagus and bladder logits, that is, the coefficient difference multiplied by the model input (the rank-normal score for the baseline and the standardized score for HostMix-TOO). To ask which part of the esophageal carcinoma class attracted calls, the TCGA-train esophageal carcinomas (n = 145) were divided by their Genomic Data Commons primary diagnosis [64] into adenocarcinoma (74), squamous cell carcinoma (65) and other (6), and each evaluation sample called esophagus was assigned to the adenocarcinoma or squamous centroid by Spearman correlation.

Software

Analyses used Python 3.12.4 with NumPy 1.26.4 [65], SciPy 1.13.1 [66], pandas 2.2.2 [67], scikit-learn 1.4.2 [54] and matplotlib 3.8.4 [68]. SCOPE and CUP-AI-Dx were run in separate Python 3.8.20 environments with TensorFlow 2.10.1. Additional file 1 maps every reported result to the script and output file that produced it.

Use of large language models

Large language model tools assisted this study. Claude (Anthropic; Claude Opus 5.5) was used to draft the analysis plans, the stage-wise work instructions and the text of this manuscript. An agent in the Cursor editor (xAI; Grok 4.7) was used to write and run the

analysis code, to produce the figures and tables, and to check the manuscript's references and numbers against Crossref records and result files with independent scripts. Juwon Kang decided the research questions, approved each plan before it was committed, and reviewed the code outputs and the manuscript. Confirmatory analyses followed the internal preregistration described above; deviations are listed in Additional file 3. The authors take full responsibility for the content. The tools are not authors.

Results

Host attraction in simulated mixtures

In simulated mixtures of held-out TCGA tumors with GTEx-sim tissue, the baseline classifier named the host organ more often as the tumor RNA fraction fell (Fig. 2). Across the ten host-pool tissues, its mean host-attraction rate was 0.111 at $\rho = 0.6$ and 0.431 at $\rho = 0.2$; the corresponding rates for HostMix-TOO were 0.002 and 0.007 (Fig. 2a). The effect was largest for liver: at $\rho = 0.6$, top-1 accuracy was 0.394 for the baseline and 0.976 for HostMix-TOO (n = 1000). For the twelve tissues outside the host pool, HostMix-TOO reduced host attraction less ($\rho = 0.6$: 0.120 for the baseline and 0.084 for HostMix-TOO), whereas the exploratory 22-tissue model reduced it to 0.002 (Fig. 2b). Models trained without one pool tissue were less protected against that tissue: at $\rho = 0.6$, the median host-attraction rate on the left-out tissue was 0.018, compared with 0.039 for the baseline and 0.001 for HostMix-TOO on the same tissues, and the model trained without liver had a rate of 0.549 on liver mixtures, compared with 0.625 for the baseline and 0.001 for HostMix-TOO; the model trained without brain cortex had a rate of 0.758 on brain cortex mixtures, higher than the baseline's 0.068 and HostMix-TOO's 0.000 (Fig. 2c). Augmentation did not lower accuracy on pure tumors: top-1 accuracy on TCGA-test was 0.982 for the baseline and 0.979 for HostMix-TOO (difference −0.003, 95% CI −0.006 to 0.000; n = 1872), and on TCGA metastatic samples it was 0.933 for the baseline and 0.949 for HostMix-TOO (n = 373).

Development on MET500

In MET500, the baseline classifier had a host-attraction rate of 0.136 (49 of 361 at-risk biopsies) and a top-1 accuracy of 0.689 (n = 437). Among the candidate training procedures, HostMix-TOO had the highest top-1 accuracy, 0.728, with a host-attraction rate of 0.044, and was locked by the pre-specified rule; site-specific mixture models performed similarly (0.714 and 0.030) (Additional file 1: Table S4, Fig. S1). Normal classes and gene removal left most host attraction in place (0.119 for both), linear deconvolution removed part of it (0.089), and native-organ masking removed it by construction but lowered top-1 accuracy to 0.600, because a masked classifier cannot call a tumor that arose in the biopsied organ. Gene-set scores were less accurate than single-gene scores, without and with mixtures (top-1 0.568 and 0.670). Because MET500 informed the choice of HostMix-TOO, its MET500 results are not independent evidence.

Confirmation 1: POG570

In POG570, HostMix-TOO reduced the host-attraction rate from 0.228 to 0.024 (H1; n = 378; difference −0.204, 95% CI −0.243 to −0.164; 77 discordant pairs in favor of HostMix-TOO and 0 against; one-sided exact McNemar $p = 6.6 \times 10^{-24}$) and increased top-1 accuracy from 0.682 to 0.756 (H2; n = 512; difference 0.074, 95% CI 0.039 to 0.107; 61 versus 23 discordant pairs; $p = 2.0 \times 10^{-5}$) (Fig. 3, Table 2). In the native-truth set, top-1 accuracy was 0.868 for the baseline and 0.802 for HostMix-TOO (H3; n = 91; difference −0.066, 95% CI −0.132 to 0.000). The one-sided 95% lower bound, −0.121, was below the margin of −0.10, so non-inferiority was not shown.

The pre-specified descriptive strata show where the reduction occurred (Fig. 4b). In liver biopsies, the baseline named liver or biliary tract for 0.451 of at-risk biopsies (n = 175) and HostMix-TOO for 0.046, and top-1 accuracy rose from 0.478 to 0.726 (n = 186). Under the baseline, the host-attraction rate was 0.273 and 0.274 in the lower two tertiles of tumor

content and 0.135 in the highest; under HostMix-TOO it was 0.055, 0.008 and 0.008. In the native-truth set, the net loss of six correct calls under HostMix-TOO came from sarcomas biopsied from soft tissue (10 versus 8 correct of 10) and from other sites (41 versus 37 of 45); accuracy in native-truth biopsies from liver, lung and lymph node was unchanged. Among the alternatives locked with HostMix-TOO, site-specific mixture models gave almost identical results (top-1 0.758; host attraction 0.024), whereas normal classes (0.709; 0.180) and linear deconvolution (0.723; 0.143) retained more host attraction (Additional file 1: Table S6).

Confirmation 2: six RNA-seq cohorts

In the six auxiliary RNA-seq cohorts, all three primary hypotheses were confirmed (Fig. 3, Table 2). HostMix-TOO reduced the host-attraction rate from 0.204 to 0.056 (AH1; n = 427; difference −0.148, 95% CI −0.183 to −0.112; 65 versus 2 discordant pairs; $p = 1.5 \times 10^{-17}$), increased top-1 accuracy from 0.527 to 0.698 (AH2; n = 729; difference 0.171, 95% CI 0.145 to 0.200; 127 versus 2; $p = 1.2 \times 10^{-35}$), and met the non-inferiority criterion in the native-truth set (AH3; n = 315; 0.473 versus 0.730; difference 0.257, 95% CI 0.210 to 0.305; one-sided lower bound 0.216). No secondary hypothesis was supported after Holm correction (Table 2). The gated model was less accurate than HostMix-TOO in the native-truth set (AS1; 0.473 versus 0.730) and in the evaluation set (AS3; 0.550 versus 0.698), because the gate sent most biopsies to the baseline (post hoc count of the recorded gate decisions: the baseline was chosen for 552 of 729 evaluation samples and 303 of 315 native-truth samples), and the 22-tissue model did not detectably lower host attraction in the pool-out at-risk set (AS2; n = 71; 0.000 versus 0.028; 2 discordant pairs in favor and 0 against; p = 0.25).

The cohorts contributed differently (Fig. 4a, c). Most of the host-attraction reduction came from two cohorts: PRINCE pancreatic cancer biopsies (37 of the 65 favorable discordant pairs; 50 of its 56 at-risk biopsies were taken from the liver) and SU2C prostate cancer biopsies (13 pairs). In the pre-specified leave-one-cohort analyses, the reduction remained

whichever cohort was removed (differences −0.070 to −0.240). Most of the accuracy gain came instead from IMvigor210, in which the baseline classified 145 of 194 bladder tumors biopsied from the bladder as esophageal cancer, compared with 65 for HostMix-TOO. Without IMvigor210, the difference in top-1 accuracy was 0.038 (pre-specified leave-one-cohort estimate; n = 453), and the native-truth difference was 0.025 (95% CI −0.008 to 0.066; n = 121; post hoc), which still met the non-inferiority criterion (one-sided lower bound 0.000).

Microarray cohorts

In the five microarray cohorts, analyzed descriptively as pre-specified, HostMix-TOO also reduced host attraction (difference −0.155, 95% CI −0.217 to −0.097; n = 207) but lowered top-1 accuracy in the evaluation set (−0.269, 95% CI −0.312 to −0.231; n = 536) and in the native-truth set (−0.317, 95% CI −0.367 to −0.267; n = 341) (Additional file 1: Fig. S1). Top-1 accuracy was 0.724 for the baseline and 0.455 for HostMix-TOO. Most of the additional errors were calls of esophagus (see below). Among the alternatives locked with HostMix-TOO, linear deconvolution had the highest top-1 accuracy in this layer (0.812; host-attraction rate 0.010; post hoc aggregation; Additional file 1: Fig. S1). A post hoc ablation traced this loss to feature standardization (see below).

The mixtures accounted for the reduction and standardization for the microarray loss

In the post hoc ablation (Fig. 5a, b), the two standardized models trained on pure tumors stayed close to the baseline (host-attraction rates 0.122 and 0.139 in MET500, 0.196 and 0.222 in POG570, and 0.176 and 0.201 in the auxiliary cohorts), whereas the unstandardized model trained with mixtures stayed close to HostMix-TOO (0.044, 0.034 and 0.082, compared with 0.044, 0.024 and 0.056 for HostMix-TOO). The share of the gap between the baseline and HostMix-TOO attributable to the mixtures ranged from 0.810 to 1.030 across the nine comparisons, so the pre-specified reading rule (every share at least 0.5) was met. Top-1

accuracy followed the same division: the unstandardized mixture model was at least as accurate as HostMix-TOO (0.728 and 0.728 in MET500, 0.777 and 0.756 in POG570, and 0.706 and 0.698 in the auxiliary cohorts), whereas the standardized pure-tumor models stayed near the baseline (0.680 and 0.666 in MET500, 0.705 and 0.670 in POG570, and 0.494 and 0.491 in the auxiliary cohorts, compared with 0.689, 0.682 and 0.527 for the baseline). In the POG570 native-truth set, the unstandardized mixture model was correct for 0.846 of biopsies, compared with 0.868 for the baseline and 0.802 for HostMix-TOO (n = 91; Additional file 1: Table S8).

The same five models located the loss on microarray profiles in feature standardization (Fig. 5; Additional file 1: Table S9). In the microarray evaluation set (n = 536), the unstandardized mixture model had a top-1 accuracy of 0.776, compared with 0.724 for the baseline and 0.455 for HostMix-TOO, whereas the two standardized pure-tumor models had 0.360 and 0.332. The unstandardized mixture model also had the lowest host-attraction rate in the microarray at-risk set (0.092, compared with 0.256 for the baseline and 0.101 for HostMix-TOO; n = 207), was correct for 0.891 of native-truth biopsies (baseline 0.871, HostMix-TOO 0.554; n = 341) and called esophagus for 11 biopsies, compared with 30 for the baseline and 147 for HostMix-TOO. In the missing-gene simulation, removing as many randomly chosen classifier genes as each platform lacks lowered POG570 and MET500 top-1 accuracy by 0.006 on average for both the baseline and the unstandardized mixture model, compared with 0.049 for each standardized pure-tumor model and 0.031 for HostMix-TOO (Fig. 5c). The shares of HostMix-TOO's microarray accuracy loss reproduced by the unstandardized mixture model and by the two standardized pure-tumor models were −0.194, 1.354 and 1.458, so by the reading rule committed before these models were applied, standardization, not the mixtures, accounted for the loss; the same held for the loss under random gene removal (shares −0.028, 1.711 and 1.704).

Robustness to augmentation settings

In exploratory variants run between the two confirmations (Additional file 1: Table S5), the POG570 host-attraction rate stayed between 0.019 and 0.034 with one to eight mixtures per tumor, a lower bound of ρ between 0.15 and 0.5, C between 0.01 and 0.1, or a multilayer perceptron in place of logistic regression, and was 0.037 with a pool of only liver, lung and spleen. It was 0.222 when the standardized model was trained on pure tumors only, and 0.183 when liver was left out of the host pool.

Published classifiers

SCOPE and CUP-AI-Dx showed host attraction in all three cohorts (Fig. 6, Table 3): SCOPE in 0.122, 0.159 and 0.157 of at-risk biopsies in MET500, POG570 and the auxiliary cohorts, and CUP-AI-Dx in 0.180, 0.304 and 0.272, compared with 0.044, 0.024 and 0.056 for HostMix-TOO. HostMix-TOO had the highest top-1 accuracy in MET500 (0.728) and POG570 (0.756). CUP-AI-Dx had the highest in the auxiliary cohorts (0.820 versus 0.698 for HostMix-TOO), an advantage concentrated in IMvigor210 (0.899 versus 0.605; n = 276) and the breast cancer cohort (0.933 versus 0.400; n = 30) (Fig. 4c). Conversely, CUP-AI-Dx named the host organ for 0.875 of at-risk PRINCE biopsies (HostMix-TOO 0.179; n = 56) and for 0.778 of GSE50760 liver metastases (HostMix-TOO 0.056; n = 18). SCOPE assigned a normal-tissue class to 8 auxiliary biopsies; counting these as the corresponding organ changed its auxiliary top-1 accuracy from 0.418 to 0.424.

Remaining errors

Removing a host-attraction error did not always produce a correct call (Fig. 4d). Of the at-risk biopsies whose baseline host-attraction error was removed by HostMix-TOO, 53 of 77 in POG570, 19 of 34 in MET500 and 20 of 65 in the auxiliary cohorts were then classified correctly; the others received another incorrect organ.

Many of the remaining errors were calls of esophagus. Of the 114 auxiliary biopsies that CUP-AI-Dx classified correctly and HostMix-TOO did not, 82 were IMvigor210 bladder tumors and 16 were breast tumors, and HostMix-TOO called esophagus for 102. In the microarray cohort GSE41258, HostMix-TOO called esophagus for 102 of 183 primary colorectal tumors (Additional file 1: Fig. S3). The single esophageal carcinoma class of TCGA-train combined 74 adenocarcinomas, 65 squamous cell carcinomas and 6 other tumors. Of the esophagus calls by HostMix-TOO, 145 of 147 in GSE41258 and 45 of 47 in PRINCE were closer to the adenocarcinoma centroid, whereas 69 of 105 in IMvigor210 were closer to the squamous centroid (Additional file 1: Fig. S5). In IMvigor210, whose RNA was extracted from formalin-fixed, paraffin-embedded tissue and profiled with an exome-capture protocol [27], the genes with the largest positive contributions to the esophagus calls were NPIPB5, NPIPB4, FN1, ITGAV and NKX6-3 for the baseline and NPM3, SAMD9, BOLA2, ITGAV and MIA2 for HostMix-TOO (Additional file 1: Fig. S4); this analysis does not identify a mechanism.

Missing genes explain part of the microarray loss. Restricting POG570 profiles to the genes measured on GPL96 left 1443 of the 5000 classifier genes missing and lowered top-1 accuracy by 0.035 for the baseline and 0.104 for HostMix-TOO; removing the same number of randomly chosen classifier genes lowered it by 0.014 and 0.079 on average (Additional file 1: Fig. S3). HostMix-TOO was thus more sensitive than the baseline to the number of missing genes, whereas the additional loss specific to the GPL96 gene set was similar for the two classifiers (0.021 for the baseline and 0.024 for HostMix-TOO). In the ablation, this sensitivity followed feature standardization rather than the mixtures (Fig. 5c).

Finally, pancreatic cancer biopsies in PRINCE were rarely classified correctly by any method (top-1 accuracy 0.017 to 0.086; n = 58).

Discussion

This study asked whether a simple change to the training data — mixing primary-tumor profiles with normal profiles of the organs in which metastases grow — reduces a specific error of tissue-of-origin classifiers on metastatic biopsies. For RNA-seq data the answer was consistent across simulated mixtures, a development cohort and two internally preregistered confirmations: host attraction fell to between one-tenth and one-third of the baseline rate, and overall accuracy rose. A post hoc ablation attributed the reduction to the mixtures rather than to the accompanying changes in feature standardization and regularization. Two results limit the claim for HostMix-TOO as locked: in POG570, accuracy for tumors biopsied from their organ of origin decreased by 6.6 percentage points and non-inferiority was not shown, and on microarray profiles HostMix-TOO lost accuracy although it reduced host attraction. Post hoc, most of both losses disappeared when the same mixtures were used without feature standardization.

The ablation separated two components of the training change that the preregistered comparison could not. The mixtures carried the reduction in host attraction, with or without feature standardization and on RNA-seq as well as microarray profiles. Standardization, which had been adopted together with the mixtures during development, carried the loss on microarray profiles: models with standardized features lost accuracy there and under random gene removal whether or not they were trained with mixtures, whereas the unstandardized mixture model did not. The smaller loss in POG570 for tumors biopsied from their organ of origin arose mainly when the mixtures and standardization were combined (see below). A plausible mechanism is that standardization divides each feature by its standard deviation in the training data, so that, under the same penalty, genes whose rank-normal scores vary little across training tumors can receive large weights on the original scale; missing or differently measured values of such genes on another platform then shift the logits strongly. Because

rank-normal scores already put genes on a common scale within each sample, standardization was not needed for the reduction in host attraction. These results favor training mixture-augmented classifiers on unstandardized rank-normal scores, but they are post hoc: the unstandardized mixture model was specified before it was fitted, yet it was evaluated on cohorts that had already been used, its regularization strength was not tuned, and on RNA-seq its host-attraction rates were slightly higher than those of HostMix-TOO in POG570 and in the auxiliary cohorts (0.034 and 0.082, compared with 0.024 and 0.056).

The results extend the contamination model of Vincent et al. [13] from a microRNA classifier and one host organ to whole-transcriptome RNA-seq and ten host tissues in a single model that does not need the biopsy site at prediction. Site-specific mixture models were not better than one pooled model, so site information was not needed in our data. Two designs that add normal tissue in other ways removed much less host attraction. Normal-tissue classes, the design used by SCOPE [2], let a classifier recognize a profile that is mostly normal tissue, but they do not show it what a tumor diluted by normal tissue looks like; our normal-class model and SCOPE itself retained host attraction in 12–18% of at-risk biopsies. Linear deconvolution removed part of the host signal, consistent with the view that subtracting an estimated host component is harder than learning to ignore it. Generative approaches such as MetGen [12] can model the metastatic state more richly but use metastatic samples for training, whereas mixture augmentation needs only primary tumors and normal tissues, which are abundant in public resources.

The comparison with CUP-AI-Dx shows that host attraction and top-1 accuracy can diverge. CUP-AI-Dx was the most accurate classifier in the auxiliary cohorts, mainly because it classified the IMvigor210 and breast cancer biopsies well, yet it named the host organ for most at-risk biopsies of pancreatic cancer and for most liver metastases of colorectal cancer; HostMix-TOO showed the opposite pattern. We interpret these as two different failure modes.

Host attraction arises from normal cells of the biopsied organ, which the mixtures address. The esophagus calls in IMvigor210 and in the microarray cohorts more likely reflect a shift in the measurement process — formalin-fixed tissue with exome capture, or a platform that measures fewer genes — which the mixtures were not designed to address: they reduced the esophagus calls in IMvigor210 only partly (from 145 to 65 of 194), and on microarray profiles the esophagus calls followed standardization (147 for HostMix-TOO and 183 and 191 for the standardized pure-tumor models, compared with 30 for the baseline and 11 for the unstandardized mixture model). The single TCGA esophageal carcinoma class, which combines adenocarcinomas and squamous cell carcinomas, may act as a broad attractor for such shifted profiles; consistent with this, though not proof of it, esophagus calls of colorectal and pancreatic tumors were mostly closer to its adenocarcinoma component and those of urothelial tumors to its squamous component. A benchmark that reports only top-1 accuracy mixes these failure modes, whereas reporting host attraction isolates the first.

The loss of accuracy for tumors biopsied from their organ of origin in POG570 has two candidate explanations. The mixtures may teach the classifier to discount the expression of host tissues, which a tumor arising from such a tissue partly shares; two of the six net losses were sarcomas biopsied from soft tissue, whose host tissues are in the pool, whereas native-truth biopsies from liver, lung and lymph node were unaffected. The ablation suggests, however, that the loss arose mainly from combining the mixtures with standardization: neither the unstandardized mixture model (0.846) nor the standardized pure-tumor models (0.857 and 0.868) lost as much as HostMix-TOO (0.802; baseline 0.868; n = 91), although these differences rest on a few calls and were not tested. In the auxiliary cohorts the same comparison favored HostMix-TOO, and non-inferiority held even without IMvigor210. For HostMix-TOO as locked, the trade-off is therefore real but cohort-dependent, and it should be

weighed against the larger number of at-risk biopsies corrected (77 discordant pairs in favor and none against in POG570).

Four practical points follow. First, benchmarks of tissue-of-origin classifiers on metastatic biopsies should report the host-attraction rate with its at-risk denominator alongside top-1 accuracy, because the two can disagree. Second, mixture augmentation is inexpensive: it changes only the training data, one mixture per tumor was about as effective as four in exploratory analyses, and it can in principle be added to other model families, although we tested only logistic regression and one multilayer perceptron. Third, the host pool should cover the sites from which metastases are biopsied: a model trained without liver lost most of its protection in POG570, in which about a third of the biopsies came from the liver, tissues outside the pool were protected less in simulation, and a model trained without brain cortex was more attracted to brain cortex than the baseline (0.758 versus 0.068 at $\rho = 0.6$), so an incomplete pool can be worse than none for the missing tissue. Fourth, feature standardization deserves scrutiny whenever a classifier will meet profiles from another platform: in our post hoc analyses it caused the loss on microarray profiles and the sensitivity to missing genes, and gene coverage should be checked before a model trained on full RNA-seq gene sets is applied to such profiles.

This study has limitations. The preregistration was internal: commit times were recorded on the analysis workstation and not certified externally, and the same team wrote the plans and ran the analyses. The evaluation cohorts consist of metastases of known primary tumors rather than cancers of unknown primary, which may differ in biology and sampling. Truth labels come from cohort annotations; the site dictionary and native organ sets required judgment; and lymph node and bone marrow are represented by proxy tissues. The host-attraction definition counts a call of a native organ as an error even when part of the signal may reflect adaptation of tumor cells to their environment rather than admixed normal cells [10, 11], and

our data do not separate the two. The mixtures model RNA admixture linearly with tissue from post-mortem GTEx donors, which differs from tumor-adjacent tissue and lacks metastasis-specific stroma and immune infiltrates. Sites outside the host pool were protected less, and the secondary hypothesis that a 22-tissue pool would help there was not supported, although that test rested on 71 at-risk biopsies with few host-attraction errors. MET500 informed the selection of HostMix-TOO. The unstandardized mixture model, which avoided the loss on microarray profiles, was evaluated only post hoc on cohorts already used in this study. The cohort search was not exhaustive: 142 GEO series that passed title screening were not reviewed further (Additional file 1). We prepared the inputs of the published classifiers, which may not have been optimal for them, and possible overlap between POG570 and the SCOPE development data could not be checked. Finally, most of the gain in top-1 accuracy in the auxiliary family came from one cohort with an assay shift that HostMix-TOO only partly corrected.

The most direct next step is a preregistered evaluation of the unstandardized mixture model in independent RNA-seq and microarray cohorts. Further extensions can each be evaluated with the host-attraction endpoint defined here: splitting the esophageal carcinoma class by histology, augmenting with random gene dropout, building host pools from tumor-adjacent normal tissue and from additional metastatic sites, and adding the mixture step to other architectures, including published classifiers. Evaluation in cancers of unknown primary whose origin was later established would test clinical relevance.

Conclusions

Tissue-of-origin classifiers trained on primary tumors can name the organ from which a metastasis was biopsied instead of its primary site. Training on in silico mixtures of primary-tumor and normal host-tissue profiles reduced this host attraction and improved overall accuracy on RNA-seq profiles in two internally preregistered evaluations, and an ablation

attributed the reduction to the mixtures. The method needs only primary tumors and normal tissues for training and no biopsy-site information at prediction. As locked, it did not transfer to microarray profiles, protected tissues outside the training pool less, and lowered accuracy for tumors biopsied from their organ of origin in one cohort. Post hoc, the microarray loss was traced to feature standardization, and the same mixtures without standardization kept the reduction in host attraction while losing less accuracy for tumors biopsied from their organ of origin; this remains to be confirmed. Because host attraction and top-1 accuracy can diverge, as the comparison with two published classifiers showed, benchmarks of tissue-of-origin classifiers on metastatic biopsies should report host attraction with its denominator.

List of abbreviations

CI: confidence interval; FFPE: formalin-fixed, paraffin-embedded; FPKM: fragments per kilobase of transcript per million mapped fragments; GEO: Gene Expression Omnibus; GTEx: Genotype-Tissue Expression; HGNC: HUGO Gene Nomenclature Committee; MSigDB: Molecular Signatures Database; POG: Personalized OncoGenomics; RNA-seq: RNA sequencing; SU2C/PCF: Stand Up To Cancer/Prostate Cancer Foundation; TCGA: The Cancer Genome Atlas; TPM: transcripts per million.

Declarations

Ethics approval and consent to participate

This study is a secondary analysis of de-identified data and did not involve recruitment of participants. TCGA, GTEx, MET500 and the auxiliary cohorts were obtained from public repositories (UCSC Xena, cBioPortal and GEO). POG570 data were obtained from the BC Cancer Genome Sciences Centre under its terms of use; the Personalized OncoGenomics program was approved by the University of British Columbia – BC Cancer Research Ethics Board (H12-00137, H14-00681; ClinicalTrials.gov NCT02155621), and all patients gave written informed consent [20]. No attempt was made to identify individual participants.

Consent for publication

Not applicable.

Availability of data and materials

The code for data processing, training, evaluation and figures; the trained baseline and HostMix-TOO models and the unstandardized mixture model of the post hoc ablation, in scikit-learn format and as portable NumPy archives with a loader that does not require pickle; the host-pool profiles; the site dictionary and native organ sets; and sample-level predictions for the public cohorts are available at https://github.com/juwonk-yonsei/hostmix-too and archived at Zenodo (https://doi.org/10.5281/zenodo.23205336) [69]. The public repository is a snapshot without the git history and without the files that list POG570 patient identifiers with their tumor labels or per-patient results, which are listed there with their SHA-256 hashes. A copy of the git history in which these files are present only as their git object hashes, so that every commit hash, including those of the preregistration commits, can be verified, is available to editors and reviewers on request.

Project name: HostMix-TOO

Project home page: https://github.com/juwonk-yonsei/hostmix-too

Archived version: https://doi.org/10.5281/zenodo.23205336

Operating system(s): Linux; analyses were run on Ubuntu 24.04 LTS

Programming language: Python

Other requirements: Python 3.12.4, NumPy 1.26.4, SciPy 1.13.1, pandas 2.2.2, scikit-learn 1.4.2, matplotlib 3.8.4

License: MIT (code); CC BY 4.0 (trained models, host-pool profiles and predictions for the public cohorts other than cBioPortal studies); ODbL 1.0 (predictions derived from cBioPortal studies)

Any restrictions to use by non-academics: None

TCGA and GTEx expression data (Toil recomputation) are available from the UCSC Xena Toil hub [16], and MET500 expression from the UCSC Xena public hub [19]. The auxiliary cohorts are available from the cBioPortal datahub [28, 30, 32, 34, 38, 39, 51] and from GEO under accessions GSE50760, GSE209998, GSE41258, GSE14018, GSE71729 and GSE74685. cBioPortal datahub files are distributed under the Open Database License, and predictions derived from them are released under the same license. POG570 data [21] are available from the BC Cancer Genome Sciences Centre under its terms of use, which do not permit redistribution; we therefore do not release POG570 data or sample-level POG570 predictions, and report aggregate results only. SCOPE and CUP-AI-Dx were run from their public repositories at commits fd9db4cd20122c8c7b675400c0df2a4c5bebb257 and 79f9aa210a5b023d2d7525c48a7288f7423b15f6, respectively.

Competing interests

The authors declare that they have no competing interests.

Funding

This study received no specific funding.

Authors' contributions

JK: conceptualization, methodology, software, formal analysis, writing – original draft. JC: conceptualization, supervision, writing – review and editing. All authors read and approved the final manuscript.

Acknowledgements

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. This work would not be possible without the participation of our patients and families, the POG team, and the

746 generous support of the BC Cancer Foundation and Genome British Columbia (project
747 B20POG). We also acknowledge contributions towards equipment and infrastructure from
748 Genome Canada and Genome BC (projects 202SEQ, 212SEQ, 12002), Canada Foundation
749 for Innovation (projects 20070, 30981, 30198, 33408) and the BC Knowledge Development
750 Fund. All POG data were generated and are maintained by the BC Cancer Genome Sciences
751 Centre, a part of the Provincial Health Services Authority.

752 **References**

753 1. Spurgeon L, Mitchell C, Cook N, Conway AM. Cancer of unknown primary: the hunt for
754 its elusive tissue-of-origin – is it time to call off the search? Br J Cancer. 2025;133(6):733–
755 742. doi:10.1038/s41416-025-03073-7.
756 2. Grewal JK, Tessier-Cloutier B, Jones M, Gakkhar S, Ma Y, Moore R, et al. Application of
757 a Neural Network Whole Transcriptome–Based Pan-Cancer Method for Diagnosis of Primary
758 and Metastatic Cancers. JAMA Netw Open. 2019;2(4):e192597.
759 doi:10.1001/jamanetworkopen.2019.2597.
760 3. Zhao Y, Pan Z, Namburi S, Pattison A, Posner A, Balachander S, et al. CUP-AI-Dx: A tool
761 for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data
762 and artificial intelligence. EBioMedicine. 2020;61:103030. doi:10.1016/j.ebiom.2020.103030.
763 4. Vibert J, Pierron G, Benoist C, Gruel N, Guillemot D, Vincent-Salomon A, et al.
764 Identification of Tissue of Origin and Guided Therapeutic Applications in Cancers of
765 Unknown Primary Using Deep Learning and RNA Sequencing (TransCUPtomics). J Mol
766 Diagn. 2021;23(10):1380–1392. doi:10.1016/j.jmoldx.2021.07.009.
767 5. Moiso E, Farahani A, Marble HD, Hendricks A, Mildrum S, Levine S, et al. Developmental
768 Deconvolution for Classification of Cancer Origin. Cancer Discov. 2022;12(11):2566–2585.
769 doi:10.1158/2159-8290.cd-21-1443.
770 6. He B, Sun H, Bao M, Li H, He J, Tian G, et al. A cross-cohort computational framework to

771 trace tumor tissue-of-origin based on RNA sequencing. Sci Rep. 2023;13(1):15356.
772 doi:10.1038/s41598-023-42465-8.
773 7. Nguyen L, Van Hoeck A, Cuppen E. Machine learning-based tissue of origin classification
774 for cancer of unknown primary diagnostics using genome-wide mutation features. Nat
775 Commun. 2022;13(1):4013. doi:10.1038/s41467-022-31666-w.
776 8. Moon I, LoPiccolo J, Baca SC, Sholl LM, Kehl KL, Hassett MJ, et al. Machine learning for
777 genetics-based classification and treatment response prediction in cancer of unknown primary.
778 Nat Med. 2023;29(8):2057–2067. doi:10.1038/s41591-023-02482-6.
779 9. Lu MY, Chen TY, Williamson DFK, Zhao M, Shady M, Lipkova J, et al. AI-based
780 pathology predicts origins for cancers of unknown primary. Nature. 2021;594(7861):106–110.
781 doi:10.1038/s41586-021-03512-4.
782 10. Sanghvi N, Calvo-Alcañiz C, Rajagopal PS, Scalera S, Canu V, Sinha S, et al. Charting
783 the transcriptomic landscape of primary and metastatic cancers in relation to their origin and
784 target normal tissues. Sci Adv. 2024;10(49):eadn0220. doi:10.1126/sciadv.adn0220.
785 11. Nagel L, Wenzel M, Hoppe S, Karimpour M, Plum PS, Hamoudi A, et al. Gene
786 expression adaptation of metastases to their host tissue. iScience. 2026;29(8):116744.
787 doi:10.1016/j.isci.2026.116744.
788 12. Liu Z, Chiu YC, Chen Y, Huang Y. A Metastatic Cancer Expression Generator (MetGen):
789 A Generative Contrastive Learning Framework for Metastatic Cancer Generation. Cancers
790 (Basel). 2024;16(9):1653. doi:10.3390/cancers16091653.
791 13. Vincent M, Perell K, Nielsen FC, Daugaard G, Hansen NR. Modeling tissue
792 contamination to improve molecular identification of the primary tumor site of metastases.
793 Bioinformatics. 2014;30(10):1417–1423. doi:10.1093/bioinformatics/btu044.
794 14. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil enables
795 reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017;35(4):314–

316. doi:10.1038/nbt.3772.

15. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675–678. doi:10.1038/s41587-020-0546-8.

16. TCGA TARGET GTEx Toil gene expression, UCSC Xena. https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/TcgaTargetGtex_rsem_gene_tpm.gz. Accessed 1 October 2026.

17. The GTEx Consortium, Aguet F, Anand S, Ardlie KG, Gabriel S, Getz GA, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318–1330. doi:10.1126/science.aaz1776.

18. Robinson DR, Wu YM, Lonigro RJ, Vats P, Cobain E, Everett J, et al. Integrative clinical genomics of metastatic cancer. Nature. 2017;548(7667):297–303. doi:10.1038/nature23306.

19. MET500 gene expression, UCSC Xena public hub. https://ucscpublic.xenahubs.net/download/MET500/geneExpression/M.mx.txt.gz. Accessed 1 October 2026.

20. Pleasance E, Titmuss E, Williamson L, Kwan H, Culibrk L, Zhao EY, et al. Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. Nat Cancer. 2020;1(4):452–468. doi:10.1038/s43018-020-0050-6.

21. POG570 data, BC Cancer Genome Sciences Centre. https://www.bcgsc.ca/downloads/POG570/POG570_TPM_expression.txt.gz. Accessed 1 October 2026.

22. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov. 2012;2(5):401–404. doi:10.1158/2159-8290.cd-12-0095.

23. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative

821 Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci
822 Signal. 2013;6(269):pl1. doi:10.1126/scisignal.2004088.
823 24. de Bruijn I, Kundra R, Mastrogiacomo B, Tran TN, Sikina L, Mazor T, et al. Analysis and
824 Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE
825 Biopharma Collaborative in cBioPortal. Cancer Res. 2023;83(23):3861–3867.
826 doi:10.1158/0008-5472.can-23-0816.
827 25. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and
828 hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–210.
829 doi:10.1093/nar/30.1.207.
830 26. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI
831 GEO: archive for functional genomics data sets—update. Nucleic Acids Res.
832 2013;41(D1):D991–D995. doi:10.1093/nar/gks1193.
833 27. Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y, et al. TGFβ
834 attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature.
835 2018;554(7693):544–548. doi:10.1038/nature25501.
836 28. cBioPortal datahub blca_iatlas_imvigor210_2017.
837 https://raw.githubusercontent.com/cBioPortal/datahub/master/public/blca_iatlas_imvigor210_
838 2017/data_mrna_seq_tpm.txt. Accessed 6 October 2026.
839 29. Anders CK, Woodcock MG, Van Swearingen AED, Moore DT, Sambade MJ, Laurie S, et
840 al. Evaluating the efficacy of a priming dose of cyclophosphamide prior to pembrolizumab to
841 treat metastatic triple negative breast cancer. J Immunother Cancer. 2022;10(2):e003427.
842 doi:10.1136/jitc-2021-003427.
843 30. cBioPortal datahub brca_iatlas_anders_2022.
844 https://raw.githubusercontent.com/cBioPortal/datahub/master/public/brca_iatlas_anders_2022
845 /data_mrna_seq_tpm.txt. Accessed 6 October 2026.

31. Liu D, Schilling B, Liu D, Sucker A, Livingstone E, Jerby-Arnon L, et al. Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. Nat Med. 2019;25(12):1916–1927. doi:10.1038/s41591-019-0654-5.

32. cBioPortal datahub mel_dfci_2019. https://raw.githubusercontent.com/cBioPortal/datahub/master/public/mel_dfci_2019/data_mrna_seq_tpm.txt. Accessed 6 October 2026.

33. Padrón LJ, Maurer DM, O'Hara MH, O'Reilly EM, Wolff RA, Wainberg ZA, et al. Sotigalimab and/or nivolumab with chemotherapy in first-line metastatic pancreatic cancer: clinical and immunologic analyses from the randomized phase 2 PRINCE trial. Nat Med. 2022;28(6):1167–1177. doi:10.1038/s41591-022-01829-9.

34. cBioPortal datahub paad_iatlas_prince_2022. https://raw.githubusercontent.com/cBioPortal/datahub/master/public/paad_iatlas_prince_2022/data_mrna_seq_expression.txt. Accessed 6 October 2026.

35. Kim SK, Kim SY, Kim JH, Roh SA, Cho DH, Kim YS, et al. A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. Mol Oncol. 2014;8(8):1653–1666. doi:10.1016/j.molonc.2014.06.016.

36. GEO series GSE50760. https://ftp.ncbi.nlm.nih.gov/geo/series/GSE50nnn/GSE50760/suppl/GSE50760_RAW.tar. Accessed 6 October 2026.

37. Abida W, Cyrta J, Heller G, Prandi D, Armenia J, Coleman I, et al. Genomic correlates of clinical outcome in advanced prostate cancer. Proc Natl Acad Sci U S A. 2019;116(23):11428–11436. doi:10.1073/pnas.1902651116.

38. cBioPortal datahub prad_su2c_2019 polyA. https://raw.githubusercontent.com/cBioPortal/datahub/master/public/prad_su2c_2019/data_mrna_seq_fpkm_polya.txt. Accessed 6 October 2026.

871 39. cBioPortal datahub prad_su2c_2019 capture.
872 https://raw.githubusercontent.com/cBioPortal/datahub/master/public/prad_su2c_2019/data_m
873 rna_seq_fpkm_capture.txt. Accessed 6 October 2026.
874 40. Garcia-Recio S, Hinoue T, Wheeler GL, Kelly BJ, Garrido-Castro AC, Pascual T, et al.
875 Multiomics in primary and metastatic breast tumors from the AURORA US network finds
876 microenvironment and epigenetic drivers of metastasis. Nat Cancer. 2023;4(1):128–147.
877 doi:10.1038/s43018-022-00491-x.
878 41. GEO series GSE209998.
879 https://ftp.ncbi.nlm.nih.gov/geo/series/GSE209nnn/GSE209998/suppl/GSE209998_AUR_129
880 _raw_counts.txt.gz. Accessed 6 October 2026.
881 42. Sheffer M, Bacolod MD, Zuk O, Giardina SF, Pincas H, Barany F, et al. Association of
882 survival and disease progression with chromosomal instability: A genomic exploration of
883 colorectal cancer. Proc Natl Acad Sci U S A. 2009;106(17):7131–7136.
884 doi:10.1073/pnas.0902232106.
885 43. GEO series GSE41258.
886 https://ftp.ncbi.nlm.nih.gov/geo/series/GSE41nnn/GSE41258/matrix/GSE41258_series_matri
887 x.txt.gz. Accessed 6 October 2026.
888 44. Zhang XHF, Wang Q, Gerald W, Hudis CA, Norton L, Smid M, et al. Latent Bone
889 Metastasis in Breast Cancer Tied to Src-Dependent Survival Signals. Cancer Cell.
890 2009;16(1):67–78. doi:10.1016/j.ccr.2009.05.017.
891 45. GEO series GSE14018.
892 https://ftp.ncbi.nlm.nih.gov/geo/series/GSE14nnn/GSE14018/matrix/GSE14018_series_matri
893 x.txt.gz. Accessed 6 October 2026.
894 46. Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SGH, Hoadley KA, et al. Virtual
895 microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal

adenocarcinoma. Nat Genet. 2015;47(10):1168–1178. doi:10.1038/ng.3398.

47. GEO series GSE71729.
https://ftp.ncbi.nlm.nih.gov/geo/series/GSE71nnn/GSE71729/suppl/GSE71729_RAW.tar.
Accessed 6 October 2026.

48. Haider M, Zhang X, Coleman I, Ericson N, True LD, Lam HM, et al. Epithelial mesenchymal-like transition occurs in a subset of cells in castration resistant prostate cancer bone metastases. Clin Exp Metastasis. 2016;33(3):239–248. doi:10.1007/s10585-015-9773-7.

49. GEO series GSE74685.
https://ftp.ncbi.nlm.nih.gov/geo/series/GSE74nnn/GSE74685/suppl/GSE74685_RAW.tar.
Accessed 6 October 2026.

50. Kumar A, Coleman I, Morrissey C, Zhang X, True LD, Gulati R, et al. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. Nat Med. 2016;22(4):369–378. doi:10.1038/nm.4053.

51. cBioPortal datahub prad_fhcrc Agilent.
https://raw.githubusercontent.com/cBioPortal/datahub/master/public/prad_fhcrc/data_mrna_agilent_microarray.txt. Accessed 6 October 2026.

52. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766–D773. doi:10.1093/nar/gky955.

53. Seal RL, Braschi B, Gray K, Jones TEM, Tweedie S, Haim-Vilmovsky L, et al. Genenames.org: the HGNC resources in 2023. Nucleic Acids Res. 2023;51(D1):D1003–D1009. doi:10.1093/nar/gkac888.

54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–2830.
http://jmlr.org/papers/v12/pedregosa11a.html.

921 55. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al.
922 Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.
923 doi:10.1126/science.1260419.
924 56. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene
925 set enrichment analysis: A knowledge-based approach for interpreting genome-wide
926 expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–15550.
927 doi:10.1073/pnas.0506580102.
928 57. ONCOfind repository, commit bc4a55e63569bd743412530353a35c3609fd17c8.
929 https://github.com/yeonuk-Jeong/ONCOfind. Accessed 6 October 2026.
930 58. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials.
931 1986;7(3):177–188. doi:10.1016/0197-2456(86)90046-2.
932 59. CancerScope repository, commit fd9db4cd20122c8c7b675400c0df2a4c5bebb257.
933 https://github.com/jasgrewal/cancerscope. Accessed 6 October 2026.
934 60. CUP-AI-Dx repository, commit 79f9aa210a5b023d2d7525c48a7288f7423b15f6.
935 https://github.com/TheJacksonLaboratory/CUP-AI-Dx. Accessed 6 October 2026.
936 61. GEO platform GPL96.
937 https://ftp.ncbi.nlm.nih.gov/geo/platforms/GPLnnn/GPL96/annot/GPL96.annot.gz. Accessed
938 6 October 2026.
939 62. GEO series GSE71729, source of the GPL20769 symbols used here.
940 https://ftp.ncbi.nlm.nih.gov/geo/series/GSE71nnn/GSE71729/suppl/GSE71729_RAW.tar.
941 Accessed 6 October 2026.
942 63. GEO platform GPL15659.
943 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL15659&targ=self&view=data&for
944 m=text. Accessed 6 October 2026.
945 64. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a

946 Shared Vision for Cancer Genomic Data. N Engl J Med. 2016;375(12):1109–1112.
947 doi:10.1056/nejmp1607591.
948 65. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al.
949 Array programming with NumPy. Nature. 2020;585(7825):357–362. doi:10.1038/s41586-
950 020-2649-2.
951 66. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy
952 1.0: fundamental algorithms for scientific computing in Python. Nat Methods.
953 2020;17(3):261–272. doi:10.1038/s41592-019-0686-2.
954 67. McKinney W. Data Structures for Statistical Computing in Python. In: Proceedings of the
955 9th Python in Science Conference. 2010. p. 56–61. doi:10.25080/majora-92bf1922-00a.
956 68. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9(3):90–95.
957 doi:10.1109/mcse.2007.55.
958 69. Kang J, Choi J. HostMix-TOO: code, trained models and public-cohort predictions for
959 host attraction in tissue-of-origin classification. Zenodo. 2026.
960 https://doi.org/10.5281/zenodo.23205336.

961 ## Figure titles and legends

962 **Figure 1. Host attraction, mixture-augmented training and study design**

963 **a** A metastatic biopsy contains tumor cells and normal cells of the host organ. A classifier
964 trained on primary tumors can name the host organ, here liver for a colorectal cancer
965 metastasis in the liver; we call this error host attraction. A biopsy is at risk when its true organ
966 lies outside the organs native to the biopsy site, and the host-attraction rate is the fraction of
967 at-risk biopsies predicted as a native organ. **b** HostMix-TOO training. Each TCGA-train
968 tumor profile is used once unchanged and in four in silico mixtures (orange border) with a
969 normal profile of one of ten GTEx host tissues, mixed in linear TPM with tumor RNA
970 fraction ρ. Within-sample rank-normal scores of 5000 genes are standardized and used as the

input of a multinomial logistic regression (C = 0.03) with 32 tumor labels, whose probabilities are summed to 26 organ groups. The baseline classifier uses the same tumors and features without mixtures or standardization (C = 0.1). **c** Study stages. Simulated mixtures and MET500 were used for development (thin border). The baseline and HostMix-TOO were then locked (shaded box) and compared in two internally preregistered confirmations (thick borders). The comparison with published classifiers was specified in advance and is descriptive (dashed border); the ablation and the error analyses were post hoc (dotted border).

Figure 2. Host attraction in simulated tumor–host mixtures

Held-out TCGA tumors (n = 1000) were mixed with GTEx-sim samples at tumor RNA fraction ρ; within each tissue, these donors had not been used for training. **a** Mean host-attraction rate across the ten tissues of the training host pool for the baseline and HostMix-TOO. **b** Mean host-attraction rate across twelve tissues outside the host pool for the baseline, HostMix-TOO and an exploratory model trained with all 22 tissues. Rates are computed over tumors whose organ is not native to the site that the tissue represents and are averaged over tissues without weighting; values per tissue are in Additional file 2. ρ = 1 corresponds to pure tumors. **c** Host-attraction rate at ρ = 0.6 on each tissue of the host pool for the baseline, HostMix-TOO and the leave-one-host-out model trained without that tissue; rows are ordered by the baseline rate, markers within a row are offset vertically for legibility, and the number of at-risk tumors for each tissue is given on the right. Spleen and whole blood represent lymph node and bone marrow sites. The simulations were part of development and are descriptive.

Figure 3. Preregistered comparisons of HostMix-TOO with the baseline classifier

Differences, HostMix-TOO minus baseline, with two-sided 95% patient-bootstrap percentile intervals in POG570 (H1–H3) and in the six auxiliary RNA-seq cohorts (AH1–AH3). **a** Host-attraction rate in the at-risk set. **b** Top-1 organ accuracy in the evaluation set. **c** Top-1 accuracy in the native-truth set, that is, biopsies taken from the organ of origin; the dashed

line marks the non-inferiority margin of −0.10, and the open diamond shows the one-sided 95% lower bound. The columns on the right give the baseline and HostMix-TOO estimates and the size of each analysis set. H1, H2, AH1 and AH2 were tested with exact one-sided McNemar tests. The secondary hypotheses are reported in Table 2.

Figure 4. Host attraction and accuracy by cohort and stratum

a Host-attraction rate of the baseline and HostMix-TOO in each evaluation cohort, grouped by the role of the cohort; AURORA US (other RNA-seq) and the microarray cohorts are descriptive. Each line joins the two classifiers within one cohort, and the at-risk n in parentheses matches Table 1. **b** POG570 host-attraction rate by biopsy-site group and by tertile of tumor content (pre-specified descriptive); n is the number of at-risk biopsies in each group, and 0 marks a rate of zero. **c** Top-1 accuracy of the baseline, HostMix-TOO, SCOPE and CUP-AI-Dx in each auxiliary RNA-seq cohort, with the evaluation n (post hoc decomposition of Table 3). **d** At-risk biopsies whose baseline host-attraction error was removed by HostMix-TOO, divided into correct calls and other errors, in MET500 (n = 34), POG570 (n = 77) and the auxiliary cohorts (n = 65) (post hoc); the numbers in the bars are counts.

Figure 5. Post hoc ablation separating mixtures from feature standardization and regularization

a Host-attraction rate in the at-risk sets and **b** top-1 accuracy in the evaluation sets of MET500, POG570, the six auxiliary RNA-seq cohorts and the five microarray cohorts (n under each cohort) for five logistic models: the baseline (pure tumors, unstandardized, C = 0.1); two standardized models trained on pure tumors (C = 0.03 and C = 0.15); an unstandardized model trained with the HostMix-TOO mixtures (C = 0.1); and HostMix-TOO (mixtures, standardized, C = 0.03). **c** Mean loss of top-1 accuracy in MET500 and POG570 when as many randomly chosen classifier genes were removed as each of four microarray

platforms lacks (four platforms, ten draws each). Separate plans for the RNA-seq cohorts and for the microarray and missing-gene extensions, each with a reading rule, were committed before the models were applied; all results are post hoc. Shares of the differences attributable to the mixtures are given in Additional file 1: Tables S8 and S9.

Figure 6. Host attraction and accuracy of two published classifiers

a Top-1 organ accuracy in the evaluation sets and **b** host-attraction rate in the at-risk sets of the baseline, HostMix-TOO, SCOPE and CUP-AI-Dx in MET500 (n = 437; at risk, 361), POG570 (512; 378) and the six auxiliary RNA-seq cohorts (729; 427); the n under each cohort is the denominator of that panel, and the numbers above the bars in b are the rates. SCOPE and CUP-AI-Dx were run with the authors' code and models on locked input matrices. The comparison was specified before the runs and is descriptive; paired differences with bootstrap intervals are in Additional file 2. SCOPE calls of normal-tissue classes were counted as incorrect. MET500 was the development cohort for HostMix-TOO.

Tables

Table 1. Data sets and their roles

Data set	Role	Assay (unit)	Tumor type	n: evaluation / at risk / native truth	Reference
TCGA-train	Training	RNA-seq, Toil (TPM)	32 tumor labels	7486 primary tumors	[14]
TCGA-test	Held-out test; simulation tumors	RNA-seq, Toil (TPM)	32 tumor labels	1872 primary tumors	[14]
GTEx-ref	Host pool (10 tissues), reference profiles	RNA-seq, Toil (TPM)	Normal tissue	1103 of 3881 samples	[17]
GTEx-sim	Simulation hosts (22 tissues)	RNA-seq, Toil (TPM)	Normal tissue	1874 of 3911 samples	[17]
MET500	Development	RNA-seq (FPKM)	Mixed	437 / 361 / 76	[18]
POG570	Confirmation 1 (H1–H3)	RNA-seq (TPM)	Mixed advanced cancers	512 / 378 / 91	[20]
IMvigor210	Confirmation 2 (AH, AS)	RNA-seq, FFPE exome capture (TPM)	Urothelial carcinoma	276 / 82 / 194	[27]
Anders et al.	Confirmation 2	RNA-seq (TPM)	Triple-negative breast cancer	30 / 17 / 13	[29]
DFCI melanoma	Confirmation 2	RNA-seq (TPM)	Melanoma	121 / 35 / 83	[31]
PRINCE	Confirmation 2	RNA-seq (log2 normalized)	Pancreatic adenocarcinoma	58 / 56 / 0	[33]
GSE50760	Confirmation 2	RNA-seq	Colorectal cancer	18 / 18 / 18	[35]

		(FPKM)			
SU2C/PCF	Confirmation 2	RNA-seq, poly(A) or capture (FPKM)	Prostate cancer	226 / 219 / 7	[37]
GSE209998 (AURORA US)	Descriptive	RNA-seq (counts)	Breast cancer	123 / 64 / 44	[40]
GSE41258	Descriptive, microarray	GPL96	Colorectal cancer	236 / 58 / 183	[42]
GSE14018	Descriptive, microarray	GPL96	Breast cancer	36 / 36 / 0	[44]
GSE71729	Descriptive, microarray	GPL20769	Pancreatic cancer	206 / 56 / 145	[46]
GSE74685	Descriptive, microarray	GPL15659	Prostate cancer	38 / 49 / 0	[48]
FHCRC	Descriptive, microarray	Agilent	Prostate cancer	20 / 8 / 13	[50]

Legend. Evaluation, at-risk and native-truth sets are defined in Methods. Evaluation, at-risk and native-truth sets were selected separately with one sample per patient; for patients profiled in both GSE74685 and FHCRC, the retained sample can differ between sets, so the GSE74685 at-risk count exceeds its evaluation count. Auxiliary RNA-seq totals: 729 / 427 / 315; pool-out at-risk set, 71 (67 from IMvigor210). Microarray totals as used in the descriptive layer: 536 / 207 / 341. TCGA-test also supplied the 1000 tumors of the simulation benchmark. FFPE, formalin-fixed, paraffin-embedded.

Table 2. Preregistered hypotheses and results

Hypothesis	**Analysis set (n)**	**Reference model**	**Tested model**	**Difference, tested − reference (95% CI)**	**Decision rule**	**Result**
H1 Lower host-attraction rate	POG570 at risk (378)	Baseline 0.228	HostMix-TOO 0.024	−0.204 (−0.243 to −0.164)	One-sided exact McNemar test, α = 0.05	Met; $p = 6.6 \times 10^{-24}$
H2 Higher top-1 accuracy	POG570 evaluation (512)	Baseline 0.682	HostMix-TOO 0.756	0.074 (0.039 to 0.107)	As H1; tested only if H1 met	Met; $p = 2.0 \times 10^{-5}$
H3 Top-1 accuracy not lower by more than 0.10	POG570 native truth (91)	Baseline 0.868	HostMix-TOO 0.802	−0.066 (−0.132 to 0.000)	One-sided 95% lower bound > −0.10; judged only if H2 met	Not shown; bound −0.121
AH1 Lower host-attraction rate	Auxiliary at risk (427)	Baseline 0.204	HostMix-TOO 0.056	−0.148 (−0.183 to −0.112)	As H1	Met; $p = 1.5 \times 10^{-17}$
AH2 Higher top-1 accuracy	Auxiliary evaluation (729)	Baseline 0.527	HostMix-TOO 0.698	0.171 (0.145 to 0.200)	As H2	Met; $p = 1.2 \times 10^{-35}$
AH3 Top-1 accuracy not lower by more than 0.10	Auxiliary native truth (315)	Baseline 0.473	HostMix-TOO 0.730	0.257 (0.210 to 0.305)	As H3	Shown; bound 0.216
AS1 Higher top-1 accuracy	Auxiliary native truth (315)	HostMix-TOO 0.730	Gated model 0.473	−0.257 (−0.308 to −0.210)	One-sided exact McNemar test;	Not supported; p = 1.0

					Holm, family-wise α = 0.05	
AS2 Lower host-attraction rate	Auxiliary pool-out at risk (71)	HostMix-TOO 0.028	22-tissue model 0.000	−0.028 (−0.070 to 0.000)	As AS1	Not supported; p = 0.25
AS3 Higher top-1 accuracy	Auxiliary evaluation (729)	HostMix-TOO 0.698	Gated model 0.550	−0.148 (−0.174 to −0.123)	As AS1	Not supported; p = 1.0

Legend. For H1–H3 and AH1–AH3, the tested model is HostMix-TOO and the reference is the baseline; for AS1–AS3, the reference is HostMix-TOO. Intervals are two-sided 95% percentile intervals from 2000 patient-level bootstrap resamples. p values are one-sided and unadjusted; for AS1–AS3, Holm's procedure was applied to these p values. "Bound" is the one-sided 95% lower bound used for non-inferiority.

Table 3. Pre-specified descriptive comparison with published classifiers

Cohort	Classifier	Top-1 accuracy (n)	Host-attraction rate (n at risk)
MET500	Baseline	0.689 (437)	0.136 (361)
MET500	HostMix-TOO	0.728 (437)	0.044 (361)
MET500	SCOPE	0.627 (437)	0.122 (361)
MET500	CUP-AI-Dx	0.668 (437)	0.180 (361)
POG570	Baseline	0.682 (512)	0.228 (378)
POG570	HostMix-TOO	0.756 (512)	0.024 (378)
POG570	SCOPE	0.658 (512)	0.159 (378)
POG570	CUP-AI-Dx	0.611 (512)	0.304 (378)
Auxiliary RNA-seq	Baseline	0.527 (729)	0.204 (427)
Auxiliary RNA-seq	HostMix-TOO	0.698 (729)	0.056 (427)
Auxiliary RNA-seq	SCOPE	0.418 (729)	0.157 (427)
Auxiliary RNA-seq	CUP-AI-Dx	0.820 (729)	0.272 (427)

Legend. The comparison was specified before SCOPE and CUP-AI-Dx were run and is not a hypothesis test. All truth organs could be represented by both tools, so the common-label set equals the full set. MET500 was the development cohort for HostMix-TOO. The auxiliary at-risk set includes 18 GSE50760 liver metastases that are not in the evaluation set.

Additional files

Additional file 1

File name: Additional file 1

File format: .docx

Title of data: Supplementary methods, tables and figures

Description of data: Data processing, cohort search, duplicate screening and other exclusions, label and site mapping (Tables S1–S2), host pool and proxy selection (Table S3),

development comparison (Table S4), exploratory augmentation variants (Table S5), POG570 alternatives and strata (Table S6), the post hoc ablation (Tables S8 and S9), input rules for the published classifiers, design power and pipeline checks, Figures S1–S6, and a map from every manuscript result to the script and file that produced it.

Additional file 2

File name: Additional file 2

File format: .xlsx

Title of data: Source values for the figures and tables

Description of data: The values drawn in the figures and the confirmatory, pre-specified descriptive and post hoc tables, including Tables S8 and S9. The first row of each sheet states its content and status. POG570 sample identifiers are not included.

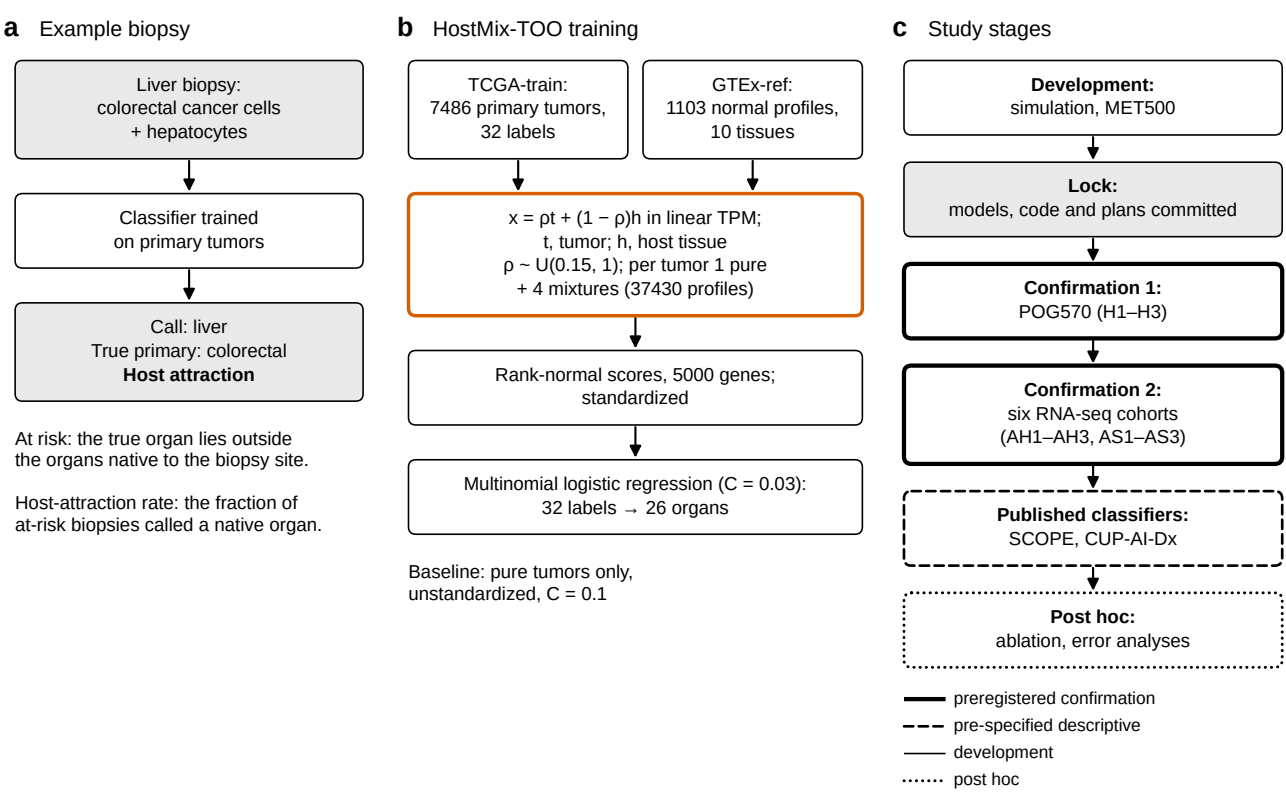
Additional file 3

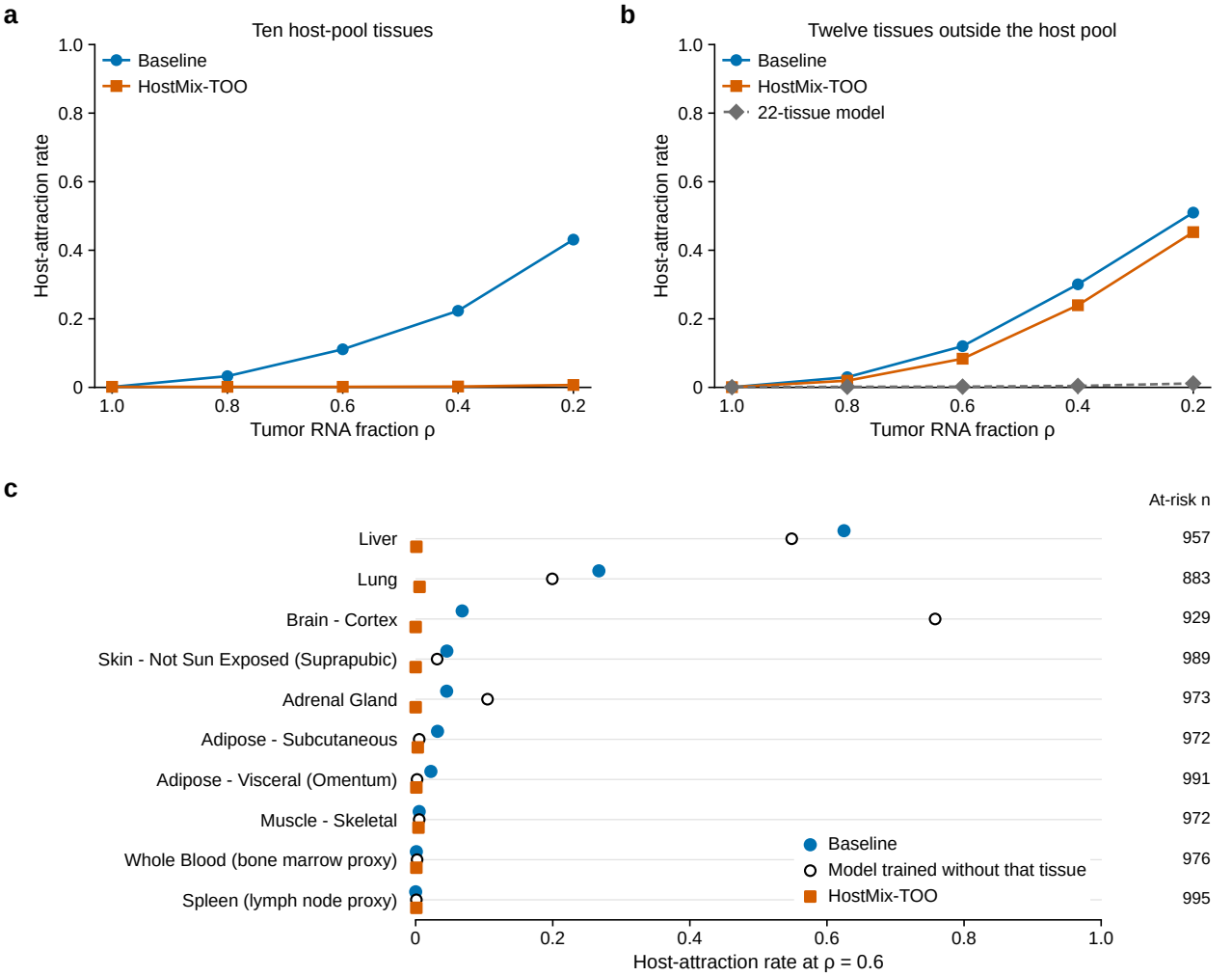
File name: Additional file 3

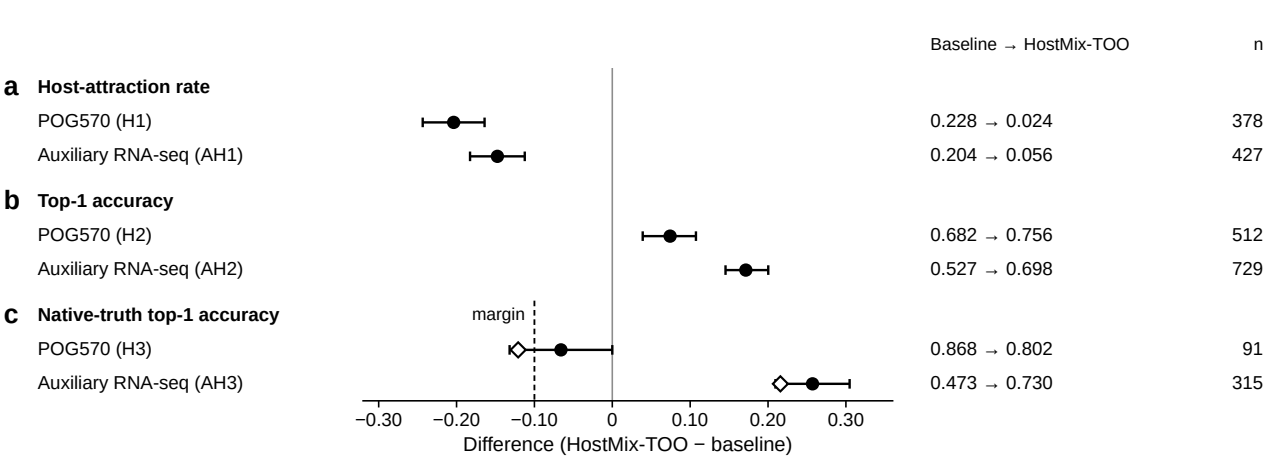
File format: .xlsx

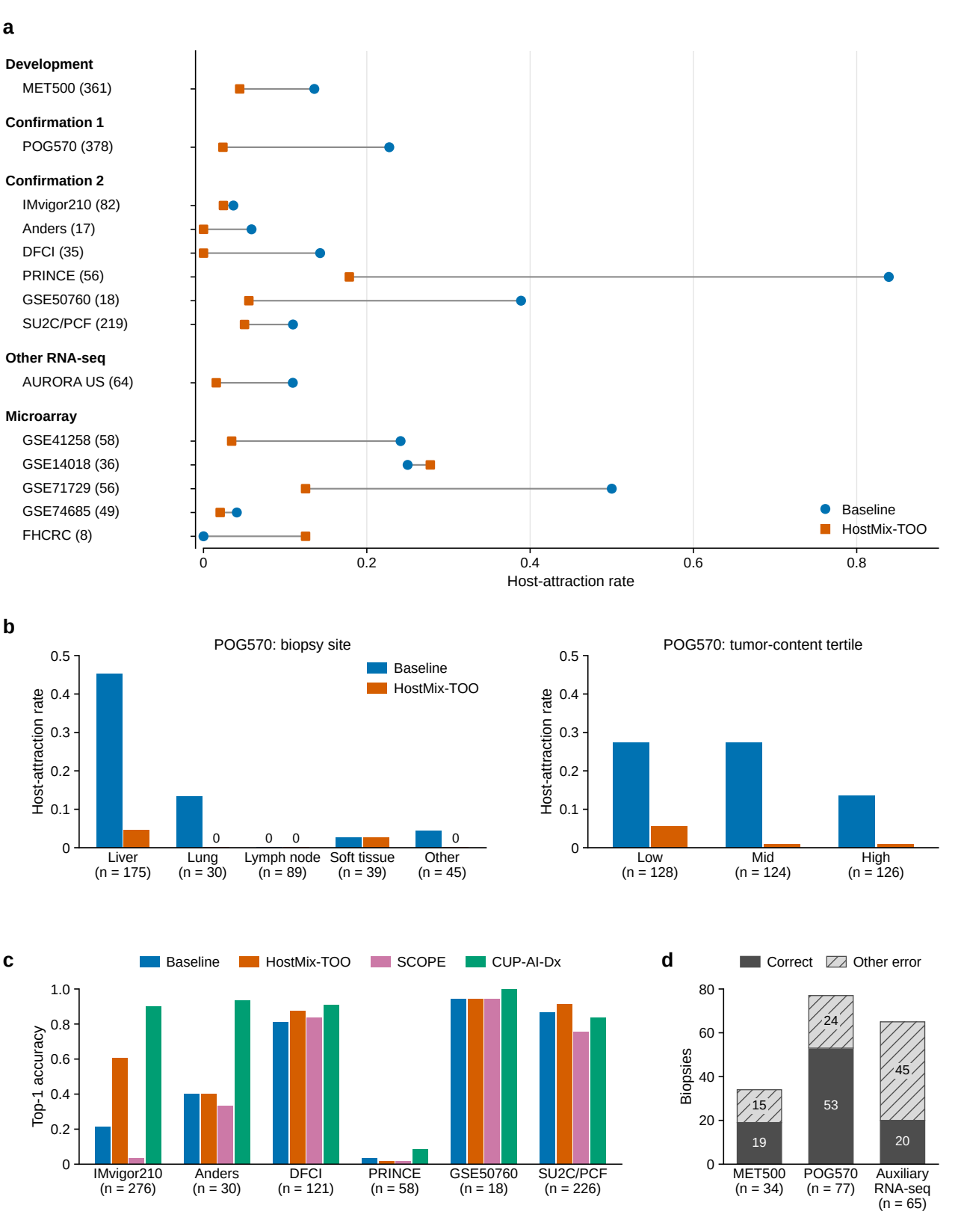
Title of data: Internal preregistration history and deviations

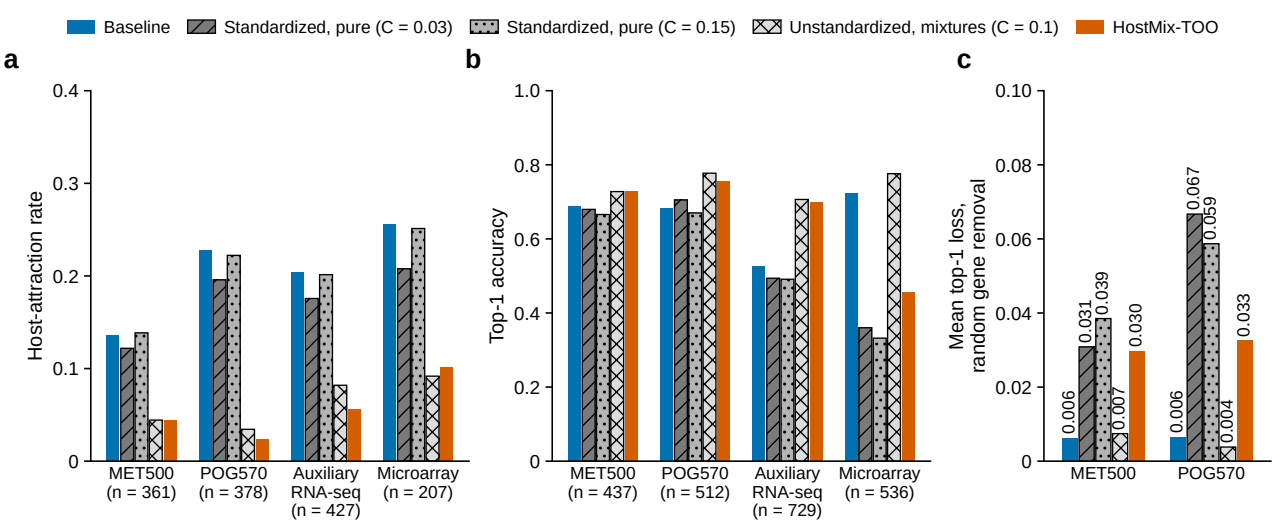
Description of data: Commit hash, time, role and recorded deviations for every lock, freeze, unlock and confirmatory output and for the two post hoc ablation plans, plus an English list of deviations from the committed plans.

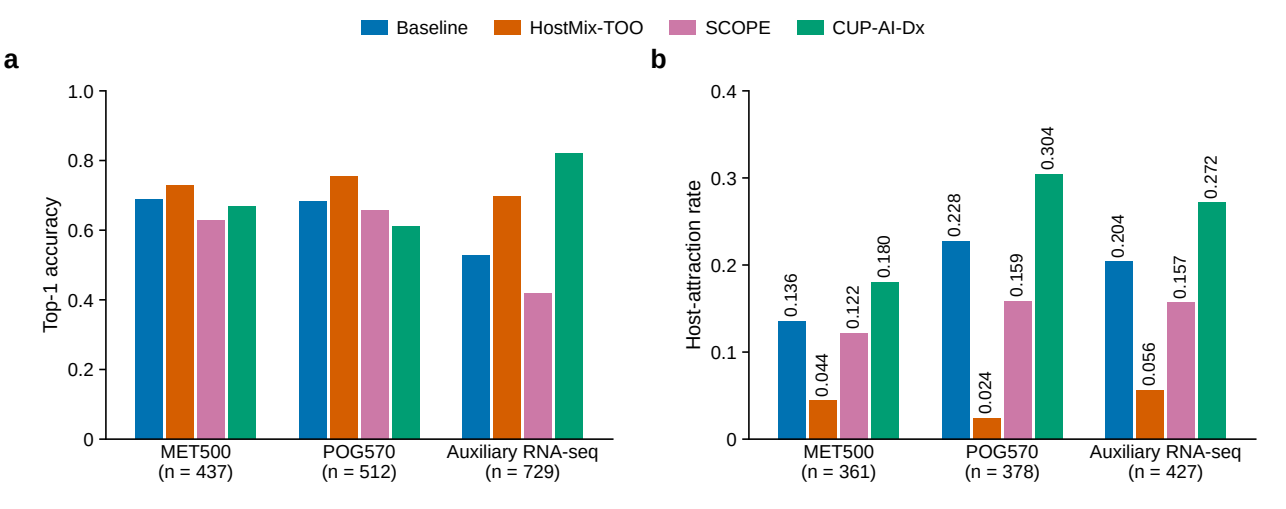

a
Example biopsy
Liver biopsy:
colorectal cancer cells
+ hepatocytes
Classifier trained
on primary tumors
Call: liver
True primary: colorectal
Host attraction
At risk: the true organ lies outside
the organs native to the biopsy site.
Host-attraction rate: the fraction of
at-risk biopsies called a native organ.
b
HostMix-TOO training
TCGA-train:
7486 primary tumors,
32 labels
GTEx-ref:
1103 normal profiles,
10 tissues
x = ρt + (1 − ρ)h in linear TPM;
t, tumor; h, host tissue
ρ ~ U(0.15, 1); per tumor 1 pure
+ 4 mixtures (37430 profiles)
Rank-normal scores, 5000 genes;
standardized
Multinomial logistic regression (C = 0.03):
32 labels → 26 organs
Baseline: pure tumors only,
unstandardized, C = 0.1
c
Study stages
Development:
simulation, MET500
Lock:
models, code and plans committed
Confirmation 1:
POG570 (H1–H3)
Confirmation 2:
six RNA-seq cohorts
(AH1–AH3, AS1–AS3)
Published classifiers:
SCOPE, CUP-AI-Dx
Post hoc:
ablation, error analyses
preregistered confirmation
pre-specified descriptive
development
post hoc


a
Ten host-pool tissues
Baseline
HostMix-TOO
Host-attraction rate
1.0
0.8
0.6
0.4
0.2
0
1.0
0.8
0.6
0.4
0.2
Tumor RNA fraction ρ
b
Twelve tissues outside the host pool
Baseline
HostMix-TOO
22-tissue model
Host-attraction rate
Tumor RNA fraction ρ
c
At-risk n
Liver
Lung
Brain - Cortex
Skin - Not Sun Exposed (Suprapubic)
Adrenal Gland
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Muscle - Skeletal
Whole Blood (bone marrow proxy)
Spleen (lymph node proxy)
957
883
929
989
973
972
991
972
976
995
Baseline
Model trained without that tissue
HostMix-TOO
0
0.2
0.4
0.6
0.8
1.0
Host-attraction rate at ρ = 0.6


Baseline → HostMix-TOO
n
a
Host-attraction rate
POG570 (H1)
0.228 → 0.024
378
Auxiliary RNA-seq (AH1)
0.204 → 0.056
427
b
Top-1 accuracy
POG570 (H2)
0.682 → 0.756
512
Auxiliary RNA-seq (AH2)
0.527 → 0.698
729
c
Native-truth top-1 accuracy
margin
POG570 (H3)
0.868 → 0.802
91
Auxiliary RNA-seq (AH3)
0.473 → 0.730
315
−0.30
−0.20
−0.10
0
0.10
0.20
0.30
Difference (HostMix-TOO − baseline)


a
Development
MET500 (361)
Confirmation 1
POG570 (378)
Confirmation 2
IMvigor210 (82)
Anders (17)
DFCI (35)
PRINCE (56)
GSE50760 (18)
SU2C/PCF (219)
Other RNA-seq
AURORA US (64)
Microarray
GSE41258 (58)
GSE14018 (36)
GSE71729 (56)
GSE74685 (49)
FHCRC (8)
Baseline
HostMix-TOO
0
0.2
0.4
0.6
0.8
Host-attraction rate
b
POG570: biopsy site
Baseline
HostMix-TOO
Host-attraction rate
0.5
0.4
0.3
0.2
0.1
0
0
0
0
0
Liver (n = 175)
Lung (n = 30)
Lymph node (n = 89)
Soft tissue (n = 39)
Other (n = 45)
POG570: tumor-content tertile
Low (n = 128)
Mid (n = 124)
High (n = 126)
c
Baseline
HostMix-TOO
SCOPE
CUP-AI-Dx
Top-1 accuracy
1.0
0.8
0.6
0.4
0.2
0
IMvigor210 (n = 276)
Anders (n = 30)
DFCI (n = 121)
PRINCE (n = 58)
GSE50760 (n = 18)
SU2C/PCF (n = 226)
d
Correct
Other error
Biopsies
80
60
40
20
0
15
19
24
53
45
20
MET500 (n = 34)
POG570 (n = 77)
Auxiliary RNA-seq (n = 65)


Baseline
Standardized, pure (C = 0.03)
Standardized, pure (C = 0.15)
Unstandardized, mixtures (C = 0.1)
HostMix-TOO
a
Host-attraction rate
0.4
0.3
0.2
0.1
0
MET500 (n = 361)
POG570 (n = 378)
Auxiliary RNA-seq (n = 427)
Microarray (n = 207)
b
Top-1 accuracy
1.0
0.8
0.6
0.4
0.2
0
MET500 (n = 437)
POG570 (n = 512)
Auxiliary RNA-seq (n = 729)
Microarray (n = 536)
c
Mean top-1 loss, random gene removal
0.10
0.08
0.06
0.04
0.02
0
0.006
0.031
0.039
0.007
0.030
0.006
0.067
0.059
0.004
0.033
MET500
POG570


Baseline
HostMix-TOO
SCOPE
CUP-AI-Dx
a
Top-1 accuracy
1.0
0.8
0.6
0.4
0.2
0
MET500
(n = 437)
POG570
(n = 512)
Auxiliary RNA-seq
(n = 729)
b
Host-attraction rate
0.4
0.3
0.2
0.1
0
0.136
0.044
0.122
0.180
0.228
0.024
0.159
0.304
0.204
0.056
0.157
0.272
MET500
(n = 361)
POG570
(n = 378)
Auxiliary RNA-seq
(n = 427)

Supplementary Files

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

- Additionalfile3.xlsx
- FigureS6.pdf
- FigureS1.pdf
- FigureS2.pdf
- FigureS3.pdf
- FigureS5.pdf
- FigureS4.pdf
- Additionalfile2.xlsx
- Additionalfile1.docx