Identification of Tissue of Origin and Guided Therapeutic Applications in Cancers of Unknown Primary Using Deep Learning and RNA Sequencing (TransCUPtomics)

Julien Vibert, Gäelle L. Pierron, Camille Benoist, Nadège Gruel, Delphine Guillemot, Anne Vincent-Salomon, Christophe Le Tourneau, Alain Livartowski, Odette Mariani, Sylvain Baulande, et al.

Identification of tissue of origin and guided therapeutic applications in cancers of unknown primary using deep learning and RNA sequencing (TransCUPtomics)

Running Title: RNA-seq and AI for CUP

Julien Vibert[1], Gaëlle Pierron[2], Camille Benoist[3], Nadège Gruel[1,4], Delphine Guillemot[2], Anne Vincent-Salomon[5], Christophe Le Tourneau[6], Alain Livartowski[7], Odette Mariani[5], Sylvain Baulande[8], François-Clément Bidard[7,9], Olivier Delattre[1,2], Joshua J. Waterfall[4,10], Sarah Watson[1,7]

1 INSERM U830, Équipe Labellisée Ligue Nationale Contre le Cancer, Diversity and Plasticity of Childhood Tumors Lab, PSL Research University, Institut Curie Research Center, Paris, France
2 Somatic Genetics Unit, Department of Genetics, Institut Curie Hospital, Paris, France
3 Clinical Bioinformatic Unit, Department of Diagnostic and Theranostic Medecine, Institut Curie Hospital, Paris, France
4 Department of Translational Research, PSL Research University, Institut Curie Research Center, Paris, France
5 Department of Diagnostic and Theranostic Medecine, Institut Curie Hospital, Paris, France
6 Department of Drug Development and Innovation, INSERM U900, Paris-Saclay University, Institut Curie Hospital and Research Center, Paris and Saint-Cloud, France
7 Department of Medical Oncology, Institut Curie Hospital, Paris, France
8 Institut Curie Genomics of Excellence (ICGex) Platform, PSL Research University, Institut Curie Research Center, Paris, France
9 INSERM CIC-BT 1428, UVSQ, Paris-Saclay University, Saint-Cloud, France
10 INSERM U830, PSL Research University, Institut Curie Research Center, Paris, France

Corresponding author:
Dr Sarah Watson, MD, PhD
INSERM U830, Équipe Labellisée Ligue Nationale Contre le Cancer, Diversity and Plasticity of Childhood Tumors Lab, PSL Research University, Institut Curie Research Center, Paris, France

Department of Medical Oncology, Institut Curie Hospital, Paris, France

26 rue d'Ulm 75005 Paris, France

e-mail: sarah.watson@curie.fr

Funding

This work was supported by ANR-10-EQPX-03, Institut Curie Génomique d'Excellence (ICGex) (S.B, O.D, S.W) and by INCa-DGOS- 4654 (J.J.W). S.W. was supported by a grant from the Institut National de la Santé et de la Recherche Medicale (INSERM) and the Foundation Bettencourt-Schueller. J. V. was supported by a grant from la Ligue Contre le Cancer and by Institut Curie. J.J.W acknowledges support from SIRIC INCa-DGOS-INSERM_12554. The ICGex NGS platform of the Institut Curie was supported by the grants ANR-10-EQPX-03 (Equipex) and ANR-10-INBS-09-08 (France Génomique Consortium) from the Agence Nationale de la Recherche ("Investissements d'Avenir" program), by the ITMO-Cancer Aviesan (Plan Cancer III) and by the SiRIC-Curie program (SiRIC Grant INCa-DGOS- 4654).

Disclosures: None declared.

Abstract

Cancers of unknown primary (CUP) are metastatic cancers for which the primary tumor is not found despite thorough diagnostic investigations. Multiple molecular assays have been proposed to identify the tissue of origin (TOO) and inform clinical care, however none has been able to combine accuracy, interpretability and easy access for routine use.

We developed a classifier tool based on the training of a variational autoencoder (VAE) to predict tissue of origin based on RNA-seq data. We used as training data 20,918 samples corresponding to 94 different categories, including 39 cancer types and 55 normal tissues. The TransCUPtomics classifier was applied to a retrospective cohort of 37 CUP patients, and to 11 prospective patients.

TransCUPtomics showed an overall accuracy of 96% on reference data for TOO prediction. The TOO could be identified in 38/48 CUP patients (79%). 8/11 prospective CUP patients (73%) could receive first-line therapy guided by TransCUPtomics prediction with responses observed in most patients. The VAE added further utility by enabling prediction interpretability, and diagnostic predictions could be matched to detection of gene fusions and expressed variants.

TransCUPtomics confidently predicted TOO for CUP and enabled tailored treatments leading to significant clinical responses. The interpretability of our approach is a powerful addition to improve the management of CUP patients.

Introduction

Cancers of unknown primary (CUP) are heterogeneous metastatic cancers for which the primary tumor cannot be identified despite a thorough diagnostic workup. CUP represent 1-2% of metastatic cancers and remain a diagnostic and therapeutic challenge. Usual first-line therapeutic strategy consists of unspecific platinum-based chemotherapy, with the response rate and median overall survival remaining below 30% and 9 to 12 months, respectively (1).

Over the last decades, multiple attempts have been made to characterize the genomic and transcriptomic landscapes of CUP in order to identify relevant molecular alterations and gene expression signatures that could orientate toward a specific tissue of origin and guide therapeutic strategies. Among them, data from microarrays, targeted DNA and RNA sequencing, DNA methylation, and whole-genome sequencing have been used with varying success (2). However, those techniques have not yet been widely included in the current diagnostic workup of CUP, due to difficulty of access, cost, and lack of standardization. Thus, current guidelines for CUP management still rely on extensive clinical and immunohistochemical characterization for tissue of origin determination (1), despite which most cases remain unclassified and are treated with unspecific systemic drugs.

The utility of gene expression data to investigate tissue of origin has already been studied extensively, with whole-transcriptomic approaches (RNA-seq) being more robust than microarrays to identify tumor characteristics and to improve diagnostic accuracy (3, 4). However, the amount of data generated by whole transcriptomic sequencing renders their analysis difficult within standard diagnostic procedures. Recently, artificial intelligence and machine learning approaches have been successfully applied to the analysis of large, high-dimensional molecular datasets (5-8).

We hypothesized that such approaches could be applied to analyze high-dimensional RNA-seq data and trained to identify tissues of origin on datasets of tumors and non-malignant tissues. We present here a classifier based on the training of a variational autoencoder (VAE), a neural network used in the field of deep learning for dimensionality reduction (9), to predict tissue of origin based on RNA-seq data. The TransCUPtomic classifier was trained on an unprecedented reference dataset of more than 20,000 tumors and normal tissues. We report our experience on 48 CUP patients on which we applied our classifier and evaluated its clinical utility by assessing the efficacy of matched therapy.

Methods

Patients and tumors

The retrospective cohort included patients over 18-years-old treated in Institut Curie over the last decade for a clinicopathological diagnosis of CUP, as assessed by a multidisciplinary tumor board after standard diagnostic workup, and for whom fresh-frozen tissue was available. This included patients treated at Institut Curie in the SHIVA 01 clinical trial (10).

Diagnostic workup included standard biological and radiological procedures to search for the primary tumor, as well as appropriate extensive pathological examination and immunohistochemistry (IHC) testing.

The prospective cohort included all new untreated CUP patients referred to Institut Curie from June 2019 to August 2020.

Response rate to first-line therapy was evaluated according to RECIST 1.1 criteria.

The study was approved by the institutional review board of Institut Curie. All living patients provided written informed consent for molecular analysis.

RNA preparation and sequencing

Total RNA was isolated from fresh-frozen tumor tissue samples using Trizol reagent. Library construction was performed following the TruSeq Stranded mRNA LS protocol (Illumina, San Diego, CA, USA). Sequencing was performed on Illumina sequencing machines: NextSeq 500 (150 nt paired-end) and NovaSeq (100 nt paired-end). Atropos (v1.1.21) was used to trim adapters from FASTQ files. Quality controls including RNA concentration, RNA integrity number, and standard sequencing criteria of raw read data were performed prior to analysis. Reads were aligned to the human reference genome (hg19) with GENCODE version 19 as the reference gene annotation with the use of STAR, version 2.7.0e and were quantified with the GeneCounts algorithm. Raw counts were normalized to transcripts per million (TPM).

Reference samples used for training

To train our classifier, we used public RNA-seq data from fresh-frozen samples of all primary tumors in The Cancer Genome Atlas (TCGA) (n=10,201). To allow for contamination of samples by normal cells and prevent overfitting in the case of low-tumor content, we also included in the training dataset all samples from juxta-tumor normal tissues in TCGA (n=746), Genotype-Tissue Expression (GTEx) (n=9,659) and Human Protein Atlas (HPA)

(n=200). We excluded categories with fewer than 10 samples available to avoid classes with too few samples for training. We divided the TCGA-SARC category into the different subtypes of sarcomas, and fused normal tissues from different platforms under the same annotation (some tissues such as liver clustered separately on the Uniform Manifold Approximation and Projection plot between different platforms, either reflecting true biological variability and/or residual batch effect). We also used data from small cell lung cancers (SCLC) (n=79, Formalin-Fixed Paraffin-Embedded (FFPE), GEO https://www.ncbi.nlm.nih.gov/geo/, accession number GSE60052) (11) and pancreatic neuroendocrine tumors (PanNET) (n=33, GEO accession GSE118014) (12). Raw data FASTQ files were downloaded from GDC archive (TCGA), SRA (GTEx, SCLC and PanNET) and HPA (HPA). In total, our reference dataset contains 20918 samples divided into 94 “diagnoses”. Supplemental Table S1 lists all diagnoses and sample numbers.

Variational autoencoder (VAE) and Machine learning classifier

The VAE encoded high-dimensional transcriptomic profiles into low-dimensional representations in a latent space composed of 100 features with interpretable gene weights (Supplemental Figure S1A). To design the VAE, we profited from the seminal work of Way and Greene (13) who designed a VAE model named "Tybalt" to encode RNA-seq samples from TCGA. As this model was already optimized in their work, we elaborated our model based on a similar architecture: as input to the VAE, we selected the 5000 most variable features after a variance-stabilizing transformation on log-tranformed TPM values (SelectVariable Features in R package Seurat v3.1.4). Our encoder neural network was fully connected and one layer deep, with an encoding intermediate layer of 100 neurons, and decoder network also fully connected and one layer deep. The latent space is therefore 100-dimensional. Input features were scaled between 0 and 1 before training (divided by maximum value of the corresponding feature). The VAE was implemented and trained with Keras version 2.2.4 (TensorFlow version 1.14.0), optimized with Adam, batch-normalized. Activation was relu (rectified linear unit) for the encoding layer and sigmoid for the decoding layer. Learning rate was 0.0005 and we trained the model for 50 epochs with no evidence of overfitting. The latent space was visualized with the use of Uniform Manifold Approximation and Projection (UMAP) in two dimensions (14). The VAE was trained with all reference samples as input, resulting in 100-dimensional encoded representations for each sample, then two different machine learning classifiers were trained on these 100 features: one random forest (RF) classifier, and another using k-nearest neighbors (KNN). The RF classifier was

trained using the RandomForest (v4.6-14) package in R, with 5000 trees, mtry= 10. The KNN classifier was trained using the kknn (v1.3.1) package in R, we used weighted 10-nearest neighbors with a gaussian kernel.

Cross-validation of the classifier

To measure the performance of the classifiers on the training dataset, we performed the following 3-fold cross-validation procedure: the reference dataset was randomly divided into 3 equally sized parts, and 3 classifiers were independently trained with 2 out of the 3 parts, the third being reserved from the entire training procedure for use as a validation dataset. In this way, the cross-validation was entirely "blind" to the training of the classifier, including in the feature selection step from the VAE. As each of the reference sample was included in one of the 3 validation datasets we could calculate a confusion matrix for the classifier on the entire reference dataset, as well as precision and recall values for each of the diagnoses in the classifier.

To account for diagnoses which arise from similar tissues and, as expected, share similar transcriptome profiles (as depicted in the UMAP plot), different labels corresponding to subtypes of the same normal tissue were grouped together for this procedure, namely: 13 subtypes of brain tissue, 4 subtypes of gynecologic tissue (cervix, uterus, vagina, fallopian tube), 3 subtypes each of artery and esophageal tissues, 2 subtypes each of adipose, colon, skin and cardiac tissues. Moreover, tumors with similar clinical management and related transcriptome profiles were regrouped as well, namely: 5 subtypes of soft tissue sarcomas from the TCGA-SARC project (soft tissue sarcoma), colon and rectum adenocarcinomas (colorectal adenocarcinoma), stomach and esophageal carcinomas (gastroesophageal carcinoma).

Classification of test samples

CUP transcriptomic profiles were encoded in the 100-dimensional latent space with the VAE encoder neural network, and the 100-feature vector was input into RF and KNN classifiers to give a prediction of the most probable diagnosis with corresponding scores (Supplemental Figure S1B). Each test sample was projected on the original reference UMAP for visualization.

We used two criteria for confidence of the prediction: 1) The same diagnosis is predicted by both classifiers; 2) At least one of the diagnoses is predicted with a large score (> 50%).

Thus we could define confidence of prediction as: 1) high: both criteria present; 2) moderate: one of the two criteria present; 3) low: both criteria absent, samples are left "unclassified". Considering that some diagnoses may overlap and orient towards similar clinical management, criteria number one was considered to be fulfilled when a pair of diagnoses of the same family were predicted. Specifically, we fused the following pairs of diagnoses in our samples:

- "Kidney renal clear cell carcinoma" and "Kidney renal papillary cell carcinoma" were grouped into "Kidney carcinoma";
- "Liver hepatocellular carcinoma" and "Cholangiocarcinoma" were grouped into "Liver HCC/Cholangiocarcinoma";
- "Ovarian serous cystadenocarcinoma" and "Uterine corpus endometrial carcinoma" were grouped into "Gynecological carcinoma";
- subtypes of soft tissue sarcoma were grouped into "Soft tissue sarcoma";
- upper tract gastro-intestinal cancers were grouped into "GI cancer".

Exploration of VAE features for interpretability

The VAE encodes all samples inside a 100-dimensional latent space with 100 features that can be interpreted. For each feature, we calculated mean values for samples of each diagnosis to infer diagnoses with high values for this feature. Conversely, an "average" profile could also be calculated for each diagnosis by taking the mean value for each of the 100 features. Each feature is the result of a non-linear combination of the initial transcriptomic features, and the associated weights in the decoder network of the VAE give us an idea of the genes most contributing to this feature. We performed GO analysis on the 100 highest-weight genes in each feature with the package gprofiler2 v0.7.0 in R. Results of these analyses are in Supplemental Table S2.

Analysis of expressed variants and fusion detection

We used RNA-seq to detect gene fusions and infer expressed variants as described below.

<u>Fusion gene detection:</u>

Fusion gene detection was performed by two complementary approaches:

1/ a targeted analysis using a curated list of known fusion gene sequences to detect well-documented fusions.

2/ an exploratory analysis with 5 fusion-detection tools:

- Defuse v0.6.0,

- StarFusion v2.5.3,
- Fusion Catcher v1.00,
- FusionMap Oshell toolkit v10.0.1.50,
- ARRIBA v1.2.0

Interpretation combined the results of the targeted fusion analysis and those of the exploratory analysis.

Variant Calling:

Read alignment was performed with STAR on hg19 and read cleaning was done as described by GATK good practice recommandations (v3.5). Variant calling was performed on a list of 499 genes (Cancer Gene Census, COSMIC 24.05.2016) using haplotype Caller (GATK v.3.5) and Mutect2 (GATK v.4). Reads with mapping quality lower than 6 and sequenced bases quality lower than 20 were not considered for variant calling. Variants were annotated with ANNOVAR (v2018Apr16). Modelization of SNVs belonging to a list of 499 genes (Cancer Gene Census, COSMIC 24.05.2016) was afterwards validated using Alamut Visual 2.9.0 (Interactive BiosoftWare) and annotated for pathogenicity in 5 classes using Varsome Educational use v9.3.4 (https://varsome.com) following ACMG recommandations.

Results

Training, performance and interpretation of the classifier

The TransCUPtomics classifier was trained to predict tissue of origin based on RNA-seq data. In total, the reference dataset contained 20,918 samples corresponding to 94 diagnostic categories including 39 tumor types and 55 normal tissue types (Supplemental Table S1).

UMAP visualization of the transcriptomic landscape of the reference dataset captured by the VAE showed strong separation of most tumor types according to their clinical and pathological diagnosis (Figure 1A). For example, adenocarcinoma from lung (T_LUAD), prostate (T_PRAD), and pancreatic (T_PAAD) origins clearly formed separated groups of tumors. Moreover, different histological subtypes of tumors developing from the same primary site such as kidney tumors (T-KIRC and T-KIRP) could be distinguished, and similar tumors of distinct grades such as glioma (T-LGG) and glioblastoma (T-GBM) showed a continuous distribution. On the contrary, squamous cell carcinoma of various tissue origins including head and neck (T-HNSC), cervix (T-CESC) and lung (T-LUSC) showed partial overlap (Figure 1B), in accordance with previous reports (15) . Normal tissues clustered

according to their tissue of origin and apart from malignant tumors originating from the same organs. Of note, a transcriptomic continuum could be identified between different organs of the same embryonal origin (for example between vagina (N-VAG), cervix (N-CER), uterus (N-UTER) and fallopian tubes (N-FALLOP)), or across different structures of the same organ such as brain (N-BRA) (Figure 1C).

The 100 VAE features were used to train two machine learning classifiers based on random forest (RF) and K-nearest neighbors (KNN) to predict the most probable diagnoses. Cross-validation showed robust overall accuracy, with predictions matching the true diagnosis in 96.26% of cases with the RF classifier and in 96.03% of cases with KNN (94.99% for RF and 94.53% for KNN when restricting the analysis to tumor samples) (Figure 2 and Supplemental Figure S2A-C).

Each of the 100 VAE features was a weighted combination of genes (Table S2), allowing biological interpretation of the classification. Gene Ontology (GO) performed on the high-weight genes of each feature enabled identification of biological processes. This included features associated to neural development (VAE_52, highly expressed in all brain samples), immune infiltration (VAE_2, 9, 39, 64, highly expressed in all blood samples), or keratinization (VAE_10, 21, 23, 65, 78, 85, 95, 98, 99) (Supplemental Table S3). VAE_2 was associated with GO terms related to the adaptive immune system, with its highest-weight genes including numerous immunoglobulin and T-cell receptor genes. Notably, diagnoses associated to VAE_2 were related to T cell-hosting tissues (normal lymph node, diffuse large B-cell lymphoma and thymoma), but also included tumor types with frequent T cell infiltration and benefit from immunotherapy (kidney carcinoma and skin melanoma).

Classification of CUP

We then evaluated the performance of the TransCUPtomics classifier to predict the tissue of origin in a series of 48 CUP patients, including 37 retrospective cases and 11 prospective patients (Table 1). All CUP diagnoses were confirmed by a multidisciplinary tumor board and had gone through extensive diagnostic workup as recommended. The median age at diagnosis was 57 years (range 30-80) and 60.4% of patients were female. The most frequent metastatic sites were the lymph nodes (70.8%), liver (29.1%) and bone (20.8%). The most frequent pathological subtypes were adenocarcinoma (47.9%) and undifferentiated carcinoma (29.2%). The extensive immunohistochemical (IHC) profile of each sample is described in Supplemental Table S4.

RNA-seq was performed on fresh-frozen tissue from the diagnostic workup. All samples met standard quality controls (Supplemental Table S5). Transcriptomic profiles were encoded in the latent space of the VAE trained on the reference dataset. When plotted on the reference UMAP representation, every CUP localized within or near a specific diagnosis, with 45 cases fitting into a specific tumor group and 3 cases within a normal tissue cluster (Figure 3). The most probable tissue of origin was rigorously predicted with both RF and KNN algorithms to evaluate robustness of classification. For each sample, a highly confident prediction was defined by: 1) a similar diagnosis given by both algorithms, and 2) at least one score of prediction over the 50% threshold. Moderate confidence diagnoses referred to cases for which only one criteria was present. The remaining cases were considered as unclassified.

Overall, a predicted diagnosis could be established in 38/48 cases (79%), and matched clinical and pathological presentation (Table 2 and Supplemental Table S6). This included 32/48 (67%) high confidence predictions and 6/48 (12%) moderate confidence predictions. The most frequent diagnoses established with high confidence scores were lung adenocarcinoma (N=6), bladder urothelial carcinoma (N=3) and breast invasive carcinoma (N=3). In three cases, a high confidence prediction was made toward a non-malignant tissue of origin (liver: CUP11, CUP43; ovary: CUP44), and was in agreement with pathological review showing tumor cellularity below 10% in all three samples.

10/48 (21%) samples remained unclassified (CUP10, 12 13, 14, 15, 19, 20, 25, 31, 45). These samples came from lymph node biopsies or lymphadenectomy specimens in 7/10 cases. 3/10 of these samples were characterized by an unusually high distance to the nearest neighbor in the 100-dimensional latent space (> 99% quantile of all samples in the cohort), suggesting that their tumor type of origin may not be represented in the reference dataset (Supplemental Table S6).

Molecular alterations

Gene fusion and variant detection algorithms were applied to all CUP samples. CUP3 showed an in-frame *HELB-HMGA2* fusion, whereas no relevant gene fusion was detected in the other samples.

Variants in genes of interest in oncogenesis were detected in 46/48 samples (Supplemental Tables S7 and S8). The most frequently mutated genes were *TP53* (*N*=19/48) and *KRAS* (*N*=9/48), and other actionable alterations were mostly detected in genes involved in DNA repair and RAS/MAPK pathways, as previously described in CUP (16). Of note, oncogenic mutations in *KRAS* and *BRCA1* (CUP43), and *TP53* (CUP11) were detected with a low depth

of coverage in two samples predicted as normal tissues by TransCUPtomics, in line with their poor tumor cellularity (Supplemental Table S8).

Clinical impact of TransCUPtomics classification

We next evaluated the potential application of TransCUPtomics classification for tailored treatment guidance. Among the 37 retrospective cases of CUP patients, 29 had received first-line unspecific platinum-based chemotherapy, 7 had received a treatment oriented towards a putative primary tumor determined by clinical and immunohistochemical characteristics, and 1 had not received any systemic treatment. The overall response rate to first-line therapy was 36%, including 4/36 complete responses and 9/36 partial responses. The diagnosis prediction given by the TransCUPtomics algorithm could have given therapeutic alternatives to platinum-based chemotherapy in 24/37 (64.8%) patients (Table 3).

Out of 11 prospective cases, eight could receive first-line systemic treatment according to the TransCUPtomics predicted tissue of origin. The remaining three patients included one who was not treated due to altered performance status and two who were treated according to clinical and pathological characteristics due to low tumor cellularity of the sample analyzed by RNAseq. Among the 8 patients that could receive TransCUPtomics-tailored first-line treatment, there were 2 complete responses and 5 partial responses (Table 3). This included a 30-year old male (CUP1) with diffuse bone and sub-diaphragmatic lymph nodes metastases, whose bone biopsy showed undifferentiated adenocarcinoma with an IHC profile (CKAE1/AE3+, CK7-, CK20-, CDX2-, TTF1-, PSA-, CD20-, PAX8+, CD10+, Vimentin+, PDL1 3+) compatible with kidney or biliopancreatic primary. TransCUPtomics showed a highly confident prediction for kidney carcinoma, further supported by the detection of a truncating *SMARCA4* rearrangement (17). The patient was included in a clinical trial evaluating an anti-PD1 immune checkpoint inhibitor in combination with an anti-angiogenic tyrosine kinase inhibitor. First evaluation at 3 months showed a complete response, and at the time of this report the patient is still progression-free after 12 months of follow-up. Other TransCUPtomics-tailored therapeutic strategies included notably frontline surgery for predicted soft tissue sarcoma, oxaliplatin and 5FU-based chemotherapy for predicted colorectal carcinoma, and paclitaxel-trastuzumab-pertuzumab for predicted HER2-amplified breast carcinoma.

Discussion

CUP consist of a heterogeneous group of metastatic tumors, for which the primary cannot be identified despite extensive radiological and pathological investigations. This raises critical clinical issues, since therapeutic strategies in oncology are primarily based on the determination of tissue of origin and treatments tailored to primary site are more effective than unspecific chemotherapy (18).

Pathological analyses are the gold standard approaches for tissue of origin determination, by enabling the detection of tissue-specific antigens by IHC. However, IHC faces several limits, including unequal access to up-to-date panels, lack of specificity of markers and absence of expression of any informative antigen in poorly differentiated tumors. As a result, no precise hypothesis of putative tissue of origin can be made after extensive IHC profiling in approximately 75% of CUP (19).

In this study, we used the largest collection of primary cancers and normal tissues assembled so far to design a deep learning algorithm to identify tissue of origin based on features derived from whole-transcriptomic data. Our classifier reached an overall accuracy of more than 95% for cancer type prediction. When applied to a series of CUP patients, TransCUPtomics could predict the likely tissue of origin in 79% of cases, which was in line with clinical and pathological tumor characteristics.

Over the last decades, multiple techniques have been developed to identify tumor tissue of origin based on molecular data. Transcriptomic profiling has been widely studied, and several commercial systems are available to predict primary tumor type using microarrays or RT-qPCR for targeted mRNA or miRNA quantification (20-22). DNA methylation patterns have also been shown to be strongly correlated with tissue of origin and enable the successful identification of likely primary tumors in CUP (23). More recently, targeted DNA profiling (24) and whole genome sequencing (25) have also been used for primary tumor type identification.

TransCUPtomics combines whole transcriptomic data analysis and deep learning for primary tumor prediction and shows several advantages compared to previous methods in addition of its high accuracy and proportion of high-confidence predictions for CUP. First, major efforts have been made over the last decade to provide collections of RNA-seq data of most tumor types, enabling the establishment of an unprecedented reference dataset of 39 different cancer types including rare diagnoses. Second, the inclusion of normal tissues in our reference dataset minimizes the risk of over-classification of samples based on expression of non-

malignant cells. Third, RNA-seq contains functional information giving insights into the biological mechanisms underlying classification and allowing identification of genetic and immune signatures for potential therapeutic applications, as opposed to DNA sequencing, targeted RNA sequencing and DNA methylation. Combined with the VAE, a deep learning technique with high potential for biomedical big data analysis, it also allows interpretability and biological insights into the transcriptomic landscape of cancers. Last, RNA-seq enables identification of fusion transcripts and expressed variants useful for primary tumor identification and precision medicine.

We show here that the TransCUPtomics classifier results in a major clinical impact for CUP patients, with an estimated 65% (24/37) of therapeutic alternatives to platinum-based chemotherapy in our retrospective cohort, a significant proportion since more than 75% of CUP patients are not offered any second-line systemic therapy (19). This included notably two cases of soft tissue sarcoma, whose diagnosis may sometimes mimick undifferentiated carcinoma due to the expression of cytokeratins by rare tumor cells. Moreover, in 8 prospective patients that could receive TransCUPtomics-tailored first-line chemotherapy, no patient experienced tumor progression and 7/8 showed tumor response at 3 months.

Prospective clinical trials investigating the efficacy of tissue-specific systemic treatments determined by molecular profiling have so far failed to show a survival benefit for CUP patients (26, 27), probably due to the heterogeneity and overall poor prognosis of tumor types enrolled in these trials. However, initial results on our prospective cohort suggest that using TransCUPtomics may improve the prognosis of individual patients, which should be confirmed in larger prospective studies. Of note, our results are in agreement with the EPICUP study, showing an improved overall survival in CUP patients receiving tumor-type specific therapy compared to patients treated with empiric approaches (23).

Our classifier faces several limits. Despite the attempt to design a reference dataset as exhaustive as possible, many rare tumor types are missing which may result in incorrect classification or absence of classification of samples as seen in 10 cases. The observation that some of these samples were characterized by an unusually high distance to the nearest neighbor in the 100-dimensional latent space indeed supports the hypothesis that their tumor type of origin may not be represented in the reference dataset. Also, some CUPs may have lost most of their differentiation characteristics rendering prediction of tissue of origin intrinsically impossible. However, TransCUPtomics algorithms still give a diagnostic orientation useful for treatment determination in those cases, and genomic features can help refine diagnostic hypotheses.

RNA-seq is becoming cost-effective and increasingly used to guide diagnosis and therapeutic choices in cancer patients. This technique is widely available, standardized, and rapid: sequencing results could be delivered in a few days, and prediction with the trained algorithm within minutes. It is currently used in routine in our national reference center. Thus, our classifier could be easily applied to prospective cohorts of patients and enriched with diverse diagnoses. We emphasize that such a tool will not replace clinical and pathological diagnoses but is designed to be an additional element to help in the diagnostic workup and therapeutic decision makings, albeit frozen tissue availability is currently restricted to specialized cancer centers. However, as frozen tissue allows higher-quality transcriptomic profiling and our reference data, including all tumor samples profiled by TCGA, are also from frozen tissue, we expected lower performance for classification of FFPE samples due to batch effect and did not evaluate them in our study. Developments are therefore needed for applying TransCUPtomics to RNA-seq data from FFPE samples.

Artificial intelligence, including classical machine learning and deep learning techniques, is increasingly being used with success for prediction tasks involving high-throughput biomedical data. However, interpretability of the vast majority of these approaches is often hampered by the "black-box" nature of these algorithms. The VAE used in this study is a promising technique to address this shortcoming of artificial intelligence, since it not only allows state-of-the-art predictive performance, but also easier interpretation and visualization of its decision, as demonstrated in this study for CUP, in comparison to other tools that primarily make a prediction without interpretation. Moreover, the VAE enables potential discovery of previously unknown biology by extracting relevant non-linear features from high-dimensional biological data. It can also be used as a generative model to create realistic synthetic data for further purposes. Altogether, the VAE is a powerful technique from artificial intelligence that exhibits high potential in multiple biomedical contexts and is increasingly being used in tasks as diverse as imaging and pharmacology (28, 29).

In summary, we present a powerful and interpretable deep learning-based classifier trained on RNA-seq data to identify tissue of origin in CUP. We propose to integrate RNA-seq and TransCUPtomics in the standard management of CUP as a cost-effective aid to pathologists and oncologists, as this widely available and standardized technique may lead to a meaningful improvement of their clinical management.

Acknowledgements

We thank the patients and their family members, and clinicians involved in their care. We also acknowledge support from Institut Curie for sample collection, banking, and processing: the Biological Resource Center and its members, the Unité de Génétique Somatique and its members and the Department of Pathology and its members. We thank Maud Kamal for access to SHIVA01 CUP samples.

References

1. Fizazi K, Greco FA, Pavlidis N, Daugaard G, Oien K, Pentheroudakis G, Committee EG. Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2015;26 Suppl 5:v133-8.
2. Rassy E, Pavlidis N. Progress in refining the clinical management of cancer of unknown primary in the molecular era. Nat Rev Clin Oncol. 2020;17(9):541-54.
3. Groschel S, Bommer M, Hutter B, Budczies J, Bonekamp D, Heining C, Horak P, Frohlich M, Uhrig S, Hubschmann D, Georg C, Richter D, Pfarr N, Pfutze K, Wolf S, Schirmacher P, Jager D, von Kalle C, Brors B, Glimm H, Weichert W, Stenzinger A, Frohling S. Integration of genomics and histology revises diagnosis and enables effective therapy of refractory cancer of unknown primary with PDL1 amplification. Cold Spring Harb Mol Case Stud. 2016;2(6):a001180.
4. Wei IH, Shi Y, Jiang H, Kumar-Sinha C, Chinnaiyan AM. RNA-Seq accurately identifies cancer biomarker signatures to distinguish tissue of origin. Neoplasia. 2014;16(11):918-27.
5. Grewal JK, Tessier-Cloutier B, Jones M, Gakkhar S, Ma Y, Moore R, Mungall AJ, Zhao Y, Taylor MD, Gelmon K, Lim H, Renouf D, Laskin J, Marra M, Yip S, Jones SJM. Application of a Neural Network Whole Transcriptome-Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers. JAMA Netw Open. 2019;2(4):e192597.
6. Xu-Monette ZY, Zhang H, Zhu F, Tzankov A, Bhagat G, Visco C, Dybkaer K, Chiu A, Tam W, Zu Y, Hsi ED, You H, Huh J, Ponzoni M, Ferreri AJM, Moller MB, Parsons BM, van Krieken JH, Piris MA, Winter JN, Hagemeister FB, Shahbaba B, De Dios I, Zhang H, Li Y, Xu B, Albitar M, Young KH. A refined cell-of-origin classifier with targeted NGS and artificial intelligence shows robust predictive value in DLBCL. Blood Adv. 2020;4(14):3391-404.
7. Menden K, Marouf M, Oller S, Dalmia A, Magruder DS, Kloiber K, Heutink P, Bonn S. Deep learning-based cell composition analysis from tissue expression profiles. Sci Adv. 2020;6(30):eaba2619.
8. Zhao Y, Pan Z, Namburi S, Pattison A, Posner A, Balachander S, Paisie CA, Reddi HV, Rueter J, Gill AJ, Fox S, Raghav KPS, Flynn WF, Tothill RW, Li S, Karuturi RKM, George J. CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. EBioMedicine. 2020;61:103030.
9. Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv:13126114v10. 2013.
10. Le Tourneau C, Delord JP, Goncalves A, Gavoille C, Dubot C, Isambert N, Campone M, Tredan O, Massiani MA, Mauborgne C, Armanet S, Servant N, Bieche I, Bernard V, Gentien D, Jezequel P, Attignon V, Boyault S, Vincent-Salomon A, Servois V, Sablin MP, Kamal M, Paoletti X, investigators S. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol. 2015;16(13):1324-34.
11. Jiang L, Huang J, Higgs BW, Hu Z, Xiao Z, Yao X, et al. Genomic Landscape Survey Identifies SRSF1 as a Key Oncodriver in Small Cell Lung Cancer. PLoS Genet. 2016;12(4):e1005895.
12. Chan CS, Laddha SV, Lewis PW, Koletsky MS, Robzyk K, Da Silva E, Torres PJ, Untch BR, Li J, Bose P, Chan TA, Klimstra DS, Allis CD, Tang LH. ATRX, DAXX or MEN1 mutant pancreatic neuroendocrine tumors are a distinct alpha-cell signature subgroup. Nat Commun. 2018;9(1):4158.
13. Way GP, Greene CS. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pac Symp Biocomput. 2018;23:80-91.
14. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:180203426 2018.
15. Campbell JD, Yau C, Bowlby R, Liu Y, Brennan K, Fan H, et al. Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas. Cell Rep. 2018;23(1):194-212 e6.
16. Ross JS, Wang K, Gay L, Otto GA, White E, Iwanik K, Palmer G, Yelensky R, Lipson DM, Chmielecki J, Erlich RL, Rankin AN, Ali SM, Elvin JA, Morosini D, Miller VA, Stephens PJ.

Comprehensive Genomic Profiling of Carcinoma of Unknown Primary Site: New Routes to Targeted Therapies. JAMA Oncol. 2015;1(1):40-9.
17. Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499(7456):43-9.
18. Greco FA. Molecular diagnosis of the tissue of origin in cancer of unknown primary site: useful in patient management. Curr Treat Options Oncol. 2013;14(4):634-42.
19. Varadhachary GR, Raber MN. Cancer of unknown primary site. N Engl J Med. 2014;371(8):757-65.
20. Ferracin M, Pedriali M, Veronese A, Zagatti B, Gafa R, Magri E, Lunardi M, Munerato G, Querzoli G, Maestri I, Ulazzi L, Nenci I, Croce CM, Lanza G, Querzoli P, Negrini M. MicroRNA profiling for the identification of cancers with unknown primary tissue-of-origin. J Pathol. 2011;225(1):43-53.
21. Bridgewater J, van Laar R, Floore A, Van TVL. Gene expression profiling may improve diagnosis in patients with carcinoma of unknown primary. Br J Cancer. 2008;98(8):1425-30.
22. Santos MTD, Souza BF, Carcano FM, Vidal RO, Scapulatempo-Neto C, Viana CR, Carvalho AL. An integrated tool for determining the primary origin site of metastatic tumours. J Clin Pathol. 2018;71(7):584-93.
23. Moran S, Martinez-Cardus A, Sayols S, Musulen E, Balana C, Estival-Gonzalez A, Moutinho C, Heyn H, Diaz-Lagares A, de Moura MC, Stella GM, Comoglio PM, Ruiz-Miro M, Matias-Guiu X, Pazo-Cid R, Anton A, Lopez-Lopez R, Soler G, Longo F, Guerra I, Fernandez S, Assenov Y, Plass C, Morales R, Carles J, Bowtell D, Mileshkin L, Sia D, Tothill R, Tabernero J, Llovet JM, Esteller M. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet Oncol. 2016;17(10):1386-95.
24. Penson A, Camacho N, Zheng Y, Varghese AM, Al-Ahmadie H, Razavi P, Chandarlapaty S, Vallejo CE, Vakiani E, Gilewski T, Rosenberg JE, Shady M, Tsui DWY, Reales DN, Abeshouse A, Syed A, Zehir A, Schultz N, Ladanyi M, Solit DB, Klimstra DS, Hyman DM, Taylor BS, Berger MF. Development of Genome-Derived Tumor Type Prediction to Inform Clinical Cancer Care. JAMA Oncol. 2019.
25. Jiao W, Atwal G, Polak P, Karlic R, Cuppen E, Subtypes PT, Clinical Translation Working G, Danyi A, de Ridder J, van Herpen C, Lolkema MP, Steeghs N, Getz G, Morris Q, Stein LD, Consortium P. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat Commun. 2020;11(1):728.
26. Hayashi H, Kurata T, Takiguchi Y, Arai M, Takeda K, Akiyoshi K, Matsumoto K, Onoe T, Mukai H, Matsubara N, Minami H, Toyoda M, Onozawa Y, Ono A, Fujita Y, Sakai K, Koh Y, Takeuchi A, Ohashi Y, Nishio K, Nakagawa K. Randomized Phase II Trial Comparing Site-Specific Treatment Based on Gene Expression Profiling With Carboplatin and Paclitaxel for Patients With Cancer of Unknown Primary Site. J Clin Oncol. 2019;37(7):570-9.
27. Hainsworth JD, Rubin MS, Spigel DR, Boccia RV, Raby S, Quinn R, Greco FA. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon research institute. J Clin Oncol. 2013;31(2):217-23.
28. Jorgensen PB, Schmidt MN, Winther O. Deep Generative Models for Molecular Science. Mol Inform. 2018;37(1-2).
29. Kell DB, Samanta S, Swainston N. Deep learning and generative methods in cheminformatics and chemical biology: navigating small molecule space intelligently. Biochem J. 2020;477(23):4559-80.

Figure legends:

Figure 1: Transcriptomic landscape of the reference dataset captured by the VAE

A. In total, the reference dataset contains 20,918 samples corresponding to 94 diagnostic categories including 39 different tumor types and 55 normal tissue types. The VAE was trained to encode all reference samples in a 100-dimensional space of latent features and later encoded in two dimensions with UMAP. B. Enlarged view of the transcriptomic profiles of lung, head and neck and cervical squamous cell carcinoma showing partial overlap. C. Enlarged view of non-malignant brain structures showing a transcriptomic continuum between cortex and basal ganglia structures. Abbreviations: N correspond to normal tissues, T to tumor types. N_ADP-SC: Adipose - Subcutaneous; N_ADP-VSC: Adipose - Visceral (Omentum); N_ADRNL: Adrenal gland; N_ART-AO: Artery - Aorta; N_ART-CRN: Artery - Coronary; N_ART-TIB: Artery - Tibial; N_BLAD: Bladder; N_BRA-ACC: Brain - Anterior cingulate cortex (BA24); N_BRA-AMY: Brain - Amygdala; N_BRA-CAU: Brain - Caudate (basal ganglia); N_BRA-CER: Brain - Cerebellum; N_BRA-CERH: Brain - Cerebellar hemisphere; N_BRA-CTX: Brain - Cortex; N_BRA-FCTX: Brain - Frontal cortex (BA9); N_BRA-HIP: Brain - Hippocampus; N_BRA-HYP: Brain - Hypothalamus; N_BRA-NA: Brain - Nucleus accumbens (basal ganglia); N_BRA-PUT: Brain - Putamen (basal ganglia); N_BRA-SN: Brain - Substantia nigra; N_BRA-SPI: Brain - Spinal cord (cervical c-1); N_BREAST: Breast; N_CERV: Cervix; N_CLN-SIG: Colon - Sigmoid; N_CLN-TRA: Colon - Transverse; N_CML-CL: Leukemia cell line (CML); N_EBV-LYM: EBV-transformed lymphocytes; N_ESO-GEJ: Esophagus - Gastroesophageal junction; N_ESO-MUC: Esophagus - Mucosa; N_ESO-MUS: Esophagus - Muscularis; N_FALLOP: Fallopian tube; N_HN: Head and neck normal tissue; N_HRT-AA: Heart - Atrial appendage; N_HRT-LV: Heart - Left ventricle; N_KDN-CTX: Kidney - Cortex; N_LIVER: Liver; N_LN: Lymph node; N_LUNG: Lung; N_MSG: Minor salivary gland; N_MUS-SKE: Muscle - Skeletal; N_NERV-TIB: Nerve - Tibial; N_OVARY: Ovary; N_PANC: Pancreas; N_PITUI: Pituitary; N_PROST: Prostate; N_SI-TI: Small intestine - Terminal ileum; N_SKIN-NS: Skin - Not sun exposed (Suprapubic); N_SKIN-S: Skin - Sun exposed (Lower leg); N_SPLE: Spleen; N_STOM: Stomach; N_TEST: Testis; N_TFIB: Transformed fibroblasts; N_THYR: Thyroid; N_UTER: Uterus; N_VAG: Vagina; N_WB: Whole blood; T_ACC: Adrenocortical carcinoma; T_BLCA: Bladder urothelial carcinoma; T_BRCA: Breast invasive carcinoma; T_CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma; T_CHOL: Cholangiocarcinoma; T_COAD: Colon adenocarcinoma; T_DDLPS: Dedifferentiated liposarcoma; T_DLBC: Diffuse large B-cell lymphoma; T_ESCA: Esophageal carcinoma; T_GBM: Glioblastoma multiforme; T_HNSC: Head and neck squamous cell carcinoma; T_KICH: Kidney renal chromophobe cell carcinoma; T_KIRC: Kidney renal clear cell carcinoma; T_KIRP: Kidney renal papillary cell carcinoma; T_LAML: Acute myeloid leukemia; T_LGG: Brain lower grade glioma; T_LIHC: Liver hepatocellular carcinoma; T_LMS: Leiomyosarcoma; T_LUAD: Lung adenocarcinoma; T_LUSC: Lung squamous cell carcinoma; T_MESO: Mesothelioma; T_MPNST: Malignant peripheral nerve sheath tumor; T_OV: Ovarian serous cystadenocarcinoma; T_PAAD: Pancreatic adenocarcinoma; T_PANET: Pancreatic neuroendocrine tumor; T_PCPG: Pheochromocytoma and paraganglioma; T_PRAD: Prostate adenocarcinoma; T_READ: Rectum adenocarcinoma; T_SCLC: Small cell lung cancer; T_SKCM: Skin cutaneous melanoma; T_SS: Synovial sarcoma; T_STAD: Stomach adenocarcinoma; T_TGCT: Testicular germ cell tumors; T_THCA: Thyroid carcinoma; T_THYM: Thymoma; T_UCEC: Uterine corpus endometrial carcinoma; T_UCS: Uterine carcinosarcoma; T_UPS: Undifferentiated pleomorphic sarcoma; T_UVM: Uveal melanoma.

Figure 2: Performance of the TransCUPtomics classifier for tissue of origin detection

Confusion matrix showing the accuracy of the Random Forest tool of classification for tumor type prediction and evaluated according to a three-fold cross-validation procedure. The reference dataset was randomly divided in three equally sized parts, and three classifiers were independently trained with two out of the three parts, the third part being left aside from the entire training procedure as a validation dataset. Rows correspond to the predicted diagnoses and columns to the true diagnoses. Recall and precision are shown at the top and right sides of the matrix. Abbreviations: T_COREAD: Colorectal adenocarcinoma; T_GESCA: Gastroesophageal carcinoma; T_SARC : Soft tissue sarcoma. The rest are similar to Figure 1.

Figure 3: Detection of tissue of origin in CUP patients

RNA-seq was performed on fresh-frozen biopsies from the diagnostic workup of each CUP patient (CUP1 to CUP48). Transcriptomic profiles were encoded in the 100-dimensional latent space of the VAE trained on the reference dataset, and then plotted on the reference UMAP representation. Each CUP sample is highlighted by a red dot with its corresponding identity number.

Tables

Characteristic	**CUP cohort (*N*= 48)**
Median age - yr (range)	57 (30-80)
Female sex - *N*. (%)	29 (60.4)
Prospective - *N*. (%)	11 (22.9)
Site of metastases - *N*. (%)	
lymph node	34 (70.8)
liver	14 (29.1)
bone	10 (20.8)
lung	6 (12.5)
peritoneum	7 (14.6)
brain	3 (6.2)
other	15 (31.2)
Histology - *N*. (%)	
adenocarcinoma	23 (47.9)
squamous cell carcinoma	5 (10.4)
undifferentiated carcinoma	14 (29.2)
other	6 (12.5)
IHC - *N*. (%)	
CK7+ CK20-	32 (66.7)
CK7+ CK20+	5 (10.5)
CK7- CK20+	1 (2)
CK7- CK20-	10 (20.8)
Suspected clinicopathological diagnosis - *N*. (%)	
Unknown primary	30 (62.5)
GI cancer	8 (16.6)
Breast cancer	4 (8.3)
Lung cancer	3 (6.3)
Gynecological cancer	3 (6.3)

Table 1: Characteristics of CUP patients

The clinicopathological diagnosis refers to the suspicion of tissue of origin that could be made based on clinical presentation and pathological analysis. Abbreviations: IHC: immunohistochemical CK7 and CK20 profiles; GI: gastrointestinal; yr: years; N: number.

Patient ID	Cohort	Clinico-pathological diagnosis	Tissue	Predicted diagnosis	Confidence
CUP1	Prospective	CUP	bone biopsy	Kidney carcinoma	High
CUP4	Prospective	CUP	retroperitoneal biopsy	Head and neck squamous cell carcinoma	High
CUP37	Prospective	CUP	muscular biopsy	Soft tissue sarcoma (UPS/LMS)	High
CUP39	Prospective	CUP/NET	liver biopsy	Pancreatic neuroendocrine tumor	High
CUP41	Prospective	CUP/BrCa	liver biopsy	Breast invasive carcinoma	High
CUP42	Prospective	CUP	peritoneal biopsy	Soft tissue sarcoma (UPS/DDLPS)	Moderate
CUP43	Prospective	ACUP/Lca/PDAC	liver biopsy	Liver	High
CUP44	Prospective	ACUP/GaCa	ovarian biopsy	Ovary	High
CUP46	Prospective	ACUP/CRC	peritoneal biopsy	Colon adenocarcinoma	Moderate
CUP47	Prospective	ACUP/CRC/GaCa	peritoneal biopsy	Colon adenocarcinoma	High
CUP48	Prospective	ACUP/CRC/GaCa	peritoneal biopsy	GI cancer	High
CUP2	Retrospective	CUP	muscular biopsy	Undifferentiated pleomorphic sarcoma	Moderate
CUP3	Retrospective	ACUP	subcutaneous biopsy	Bladder urothelial carcinoma	High
CUP5	Retrospective	CUP/Lca	lung biopsy	Lung squamous cell carcinoma	High
CUP6	Retrospective	CUP/LCa	inguinal lymph node biopsy	Cervical squamous cell carcinoma and endocervical adenocarcinoma	High
CUP7	Retrospective	CUP/PDAC	bone biopsy	Lung adenocarcinoma	High
CUP8	Retrospective	ACUP/PDAC	liver biopsy	Liver HCC / Cholangiocarcinoma	High
CUP9	Retrospective	CUP/GyCa/BrCa	cervical lymphadenectomy	Ovarian serous cystadenocarcinoma	High
CUP10	Retrospective	ACUP	cervical lymphadenectomy	Unclassified	Low
CUP11	Retrospective	ACUP	liver biopsy	Liver	High
CUP12	Retrospective	CUP	retroperitoneal lymphadenectomy	Unclassified	Low
CUP13	Retrospective	CUP/GaCa	cavum biopsy	Unclassified	Low
CUP14	Retrospective	CUP	nephrectomy	Unclassified	Low
CUP15	Retrospective	ACUP/PDAC	cervical lymph node biopsy	Unclassified	Low
CUP16	Retrospective	CUP	cervical lymphadenectomy	Lung adenocarcinoma	High
CUP17	Retrospective	CUP	cervical lymphadenectomy	Uterine corpus endometrial carcinoma	Moderate
CUP18	Retrospective	ACUP/Lca	cervical lymphadenectomy	Lung adenocarcinoma	High
CUP19	Retrospective	CUP/NET	axillary lymphadenectomy	Unclassified	Low
CUP20	Retrospective	CUP	sub-clavicular lymphadenectomy	Unclassified	Low
CUP21	Retrospective	ACUP	cervical lymphadenectomy	Bladder urothelial carcinoma	High
CUP22	Retrospective	ACUP/BrCa	axillary lymphadenectomy	Breast invasive carcinoma	High
CUP23	Retrospective	ACUP	lymph node biopsy	Colon adenocarcinoma	High
CUP24	Retrospective	CUP	axillary lymphadenectomy	Lung squamous cell carcinoma	High
CUP25	Retrospective	CUP/HNCa	cervical lymphadenectomy	Unclassified	Low
CUP26	Retrospective	ACUP/GyCa	inguinal lymphadenectomy	Gynecological carcinoma	High
CUP27	Retrospective	CUP/BrCa	lymph node biopsy	Breast invasive carcinoma	High
CUP28	Retrospective	ACUP	liver biopsy	Cholangiocarcinoma	High
CUP29	Retrospective	CUP	lymph node biopsy	Lung adenocarcinoma	High

CUP30	Retrospective	ACUP	kidney biopsy	Bladder urothelial carcinoma	High
CUP31	Retrospective	CUP/OvCa	retroperitoneal biopsy	Unclassified	Low
CUP32	Retrospective	ACUP	cervical lymphadenectomy	Lung adenocarcinoma	High
CUP33	Retrospective	ACUP	lymph node biopsy	Gynecological carcinoma	High
CUP34	Retrospective	CUP/BrCa	cervical lymphadenectomy	Pancreatic neuroendocrine tumor	High
CUP35	Retrospective	ACUP	cervical lymphadenectomy	Lung adenocarcinoma	Moderate
CUP36	Retrospective	ACUP	lymph node biopsy	Kidney carcinoma	High
CUP38	Retrospective	CUP	lymph node biopsy	Skin cutaneous melanoma	Moderate
CUP40	Retrospective	ACUP/KyCa	bone biopsy	Lung adenocarcinoma	High
CUP45	Retrospective	ACUP/BrCA	cervical lymphadenectomy	Unclassified	Low

Table 2: Results of TransCUPtomics prediction for all CUP patients

The most probable tissue of origin was predicted with both RF and KNN machine learning classifiers to evaluate robustness of classification. For each test sample, a highly confident prediction was defined by: 1) a similar diagnosis given by both machine learning algorithms and 2) at least one score of prediction over the 50% threshold. Moderate confident diagnoses referred to cases for which only one criteria was present. The remaining cases were considered as unclassified. Abreviations: CUP: carcinoma of unknown primary; ACUP: adenocarcinoma of unknown primary; NET: neuroendocrine tumor; BrCa: breast carcinoma; LCa: lung carcinoma; PDAC: pancreatic ductal adenocarcinoma; GaCa: gastric carcinoma; CRC: colorectal carcinoma; GyCa: gynecological carcinoma; HNCa: head and neck carcinoma; OvCa: ovarian carcinoma; KyCa: kidney carcinoma; UPS: undifferentiated pleomorphic sarcoma; LMS: leiomyosarcoma; DDLPS: dedifferentiated liposarcoma; GI: gastrointestinal; HCC: hepatocellular carcinoma.

Patient ID	Cohort	Predicted diagnosis	1st line treatment at diagnosis	Tumor response at 1st line (3months)	Potential therapeutic alternative with VAE (retrospective cases)
CUP1	Prospective	Kidney carcinoma	clinical trial anti-PD1 + VEGF inhibitor	CR	
CUP4	Prospective	Head and neck squamous cell carcinoma	Carboplatin	PR	
CUP37	Prospective	Soft tissue sarcoma (UPS/LMS)	surgery	CR	
CUP39	Prospective	Pancreatic neuroendocrine tumor	Carboplatin-Etoposide	PR	
CUP41	Prospective	Breast invasive carcinoma	Paclitaxel-Trastuzumab-Pertuzumab	PR	
CUP42	Prospective	Soft tissue sarcoma (UPS/DDLPS)	palliative care	NA	
CUP43	Prospective	Liver	5FU-folinic acid-oxaliplatin*	SD	
CUP44	Prospective	Ovary	5FU-folinic acid-oxaliplatin*	PD	
CUP46	Prospective	Colon adenocarcinoma	5FU-folinic acid-irinotecan-cetuximab	PR	
CUP47	Prospective	Colon adenocarcinoma	5FU-folinic acid-oxaliplatin-bevacizumab	PR	
CUP48	Prospective	GI cancer	5FU-folinic acid-oxaliplatin	SD	
CUP2	Retrospective	Undifferentiated pleomorphic sarcoma	Pembrolizumab	PD	adriamycin, ifosfamide, VEGFR inhibitors
CUP3	Retrospective	Bladder urothelial carcinoma	Carboplatin-Paclitaxel	PD	immune checkpoint inhibitors
CUP5	Retrospective	Lung squamous cell carcinoma	Vinorelbin	PD	immune checkpoint inhibitors
CUP6	Retrospective	Cervical squamous cell carcinoma and endocervical adenocarcinoma	Cisplatin- Vinorelbin	PD	immune checkpoint inhibitors
CUP7	Retrospective	Lung adenocarcinoma	5FU-folinic acid-oxaliplatin-irinotecan	PD	immune checkpoint inhibitors
CUP8	Retrospective	Liver HCC / Cholangiocarcinoma	Gemcitabin-Oxaliplatin	PD	Cisplatin, 5FU, sunitinib, clinical trials
CUP9	Retrospective	Ovarian serous cystadenocarcinoma	Carboplatin-Paclitaxel	CR	bevacizumab, PARP inhibitors
CUP10	Retrospective	Unclassified	Carboplatin-Paclitaxel	PR	0
CUP11	Retrospective	Liver	Cisplatin-Gemcitabine	PD	NA
CUP12	Retrospective	Unclassified	Cisplatin-Gemcitabine	PD	0
CUP13	Retrospective	Unclassified	Carboplatin-Paclitaxel	PR	0
CUP14	Retrospective	Unclassified	Cisplatin-Gemcitabine	SD	0
CUP15	Retrospective	Unclassified	Cisplatin-Gemcitabine	PR	0
CUP16	Retrospective	Lung adenocarcinoma	Cisplatin-5FU-Epirubicine	PD	immune checkpoint inhibitors
CUP17	Retrospective	Uterine corpus endometrial carcinoma	Cisplatin-Docetaxel	SD	0
CUP18	Retrospective	Lung adenocarcinoma	Cisplatin-Docetaxel	PR	immune checkpoint inhibitors
CUP19	Retrospective	Unclassified	Cisplatin-Etoposide	NA	0
CUP20	Retrospective	Unclassified	Gemcitabin-Oxaliplatin	CR	0
CUP21	Retrospective	Bladder urothelial carcinoma	Cisplatin-5FU-Epirubicine	PR	immune checkpoint inhibitors
CUP22	Retrospective	Breast invasive carcinoma	5FU-Epirubicin-Cyclophosphamide/ Docetaxel	NA	0
CUP23	Retrospective	Colon adenocarcinoma	Carboplatin-Paclitaxel	PD	5FU, oxaliplatin, irinotecan
CUP24	Retrospective	Lung squamous cell carcinoma	Cisplatin-5FU-Cetuximab	PR	immune checkpoint inhibitors
CUP25	Retrospective	Unclassified	Carboplatin-Paclitaxel	PR	0
CUP26	Retrospective	Gynecological carcinoma	Carboplatin-Paclitaxel	CR	bevacizumab, PARP inhibitors
CUP27	Retrospective	Breast invasive carcinoma	Adriamycin-Cyclophosphamide	SD	eribulin, PARP inhibitors, immune checkpoint inhibitors
CUP28	Retrospective	Cholangiocarcinoma	Carboplatin-Paclitaxel	SD	Cisplatin, 5FU, sunitinib, clinical trials

CUP29	Retrospective	Lung adenocarcinoma	Carboplatin-Paclitaxel	PR	immune checkpoint inhibitors
CUP30	Retrospective	Bladder urothelial carcinoma	palliative care	NA	immune checkpoint inhibitors
CUP31	Retrospective	Unclassified	Carboplatin-Paclitaxel	PD	0
CUP32	Retrospective	Lung adenocarcinoma	Carboplatin-Gemcitabin	PD	immune checkpoint inhibitors
CUP33	Retrospective	Gynecological carcinoma	Cisplatin-Gemcitabin	CR	bevacizumab, PARP inhibitors
CUP34	Retrospective	Pancreatic neuroendocrine tumor	Adriamycin-Cyclophosphamide	PR	etoposide, oxaliplatin, sunitinib
CUP35	Retrospective	Lung adenocarcinoma	Cisplatin-5FU-Etoposide-Adriamycin	PD	immune checkpoint inhibitors, RET inhibitors
CUP36	Retrospective	Kidney carcinoma	Carboplatin-Paclitaxel	PD	immune checkpoint inhibitors
CUP38	Retrospective	Skin cutaneous melanoma	Cisplatin-Etoposide	PD	immune checkpoint inhibitors
CUP40	Retrospective	Lung adenocarcinoma	Carboplatin-Paclitaxel	PD	immune checkpoint inhibitors
CUP45	Retrospective	Unclassified	Epirubicin-Cyclophosphamide	PD	0

Table 3: Therapeutic applications of TransCUPtomics prediction

For retrospective cases, the response to first-line therapy guided by clinical and pathological suspicion is indicated, as well as potential therapeutic alternatives that could have been made based on TransCUPtomics predictions. For prospective case, TransCUPtomics-tailored treatments and responses are indicated. Abreviations: CR: complete response; PR: partial response; SD: stable disease: PD: progression disease; NA: not applicable. * treatment based on clinical and pathological characteristics due to low tumor cellularity of the tumor sample analyzed by RNA-seq.

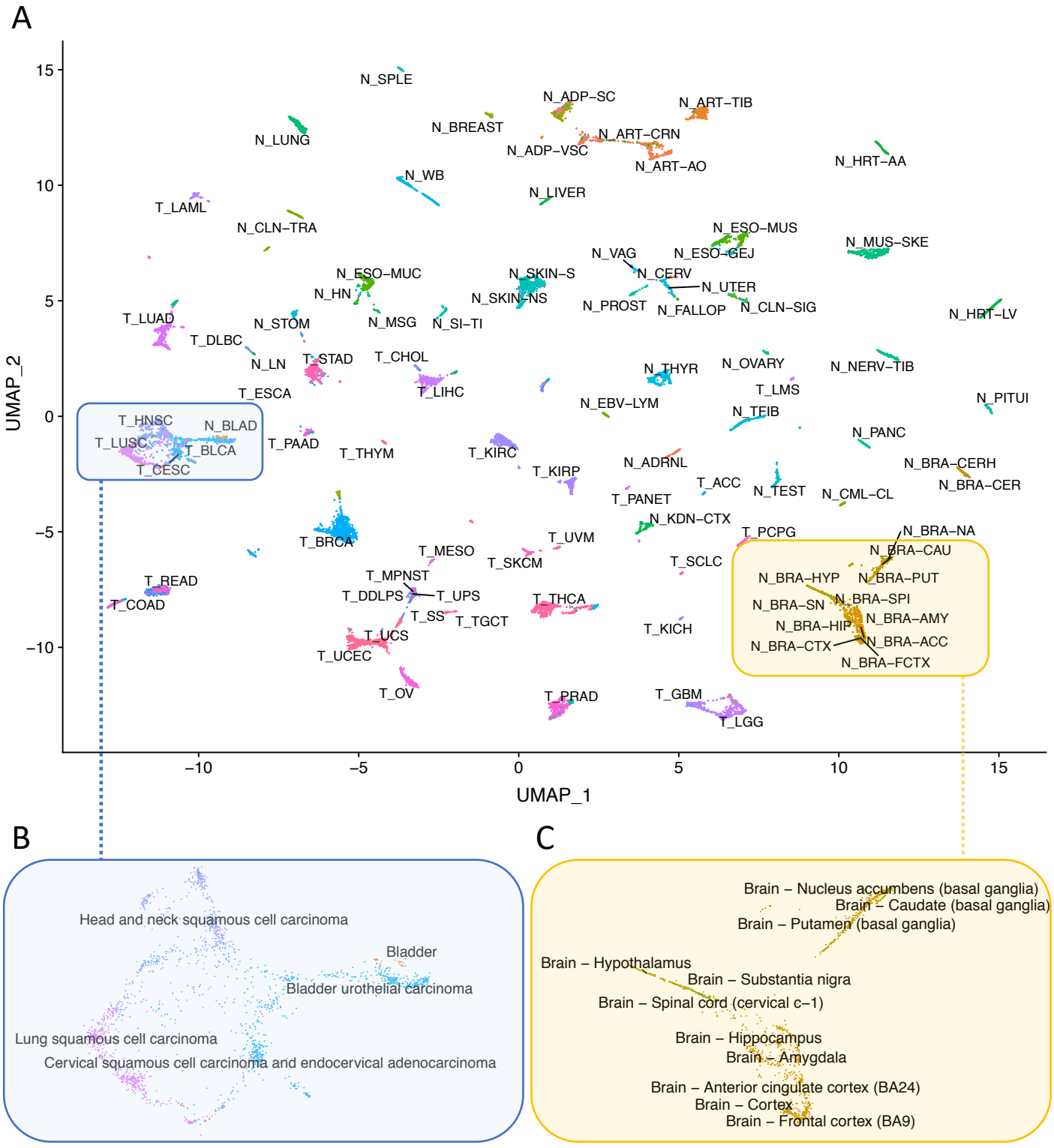

A
UMAP_2
UMAP_1
N_SPLE
N_ADP–SC
N_ART–TIB
N_BREAST
N_LUNG
N_ART–CRN
N_ADP–VSC
N_ART–AO
N_HRT–AA
N_WB
N_LIVER
T_LAML
N_CLN–TRA
N_ESO–MUS
N_MUS–SKE
N_VAG
N_ESO–GEJ
N_ESO–MUC
N_SKIN–S
N_CERV
N_HN
N_UTER
N_SKIN–NS
N_PROST
N_FALLOP
N_CLN–SIG
N_HRT–LV
T_LUAD
N_STOM
N_MSG
N_SI–TI
T_DLBC
N_LN
T_STAD
T_CHOL
N_THYR
N_OVARY
N_NERV–TIB
T_ESCA
T_LIHC
T_LMS
N_PITUI
N_EBV–LYM
N_TEIB
T_HNSC
N_BLAD
T_LUSC
T_BLCA
T_CESC
T_PAAD
N_PANC
T_THYM
T_KIRC
N_ADRNL
N_BRA–CERH
T_KIRP
T_ACC
N_BRA–CER
T_PANET
N_TEST
N_CML–CL
N_KDN–CTX
T_PCPG
N_BRA–NA
T_BRCA
T_UVM
N_BRA–CAU
T_MESO
T_SCLC
T_SKCM
N_BRA–HYP
N_BRA–PUT
T_READ
T_MPNST
N_BRA–SPI
T_COAD
T_DDLPS
T_UPS
N_BRA–SN
T_THCA
N_BRA–AMY
T_SS
T_TGCT
N_BRA–HIP
T_KICH
T_UCS
N_BRA–ACC
N_BRA–CTX
T_UCEC
N_BRA–FCTX
T_OV
T_PRAD
T_GBM
T_LGG
B
Head and neck squamous cell carcinoma
Bladder
Bladder urothelial carcinoma
Lung squamous cell carcinoma
Cervical squamous cell carcinoma and endocervical adenocarcinoma
C
Brain – Nucleus accumbens (basal ganglia)
Brain – Caudate (basal ganglia)
Brain – Putamen (basal ganglia)
Brain – Hypothalamus
Brain – Substantia nigra
Brain – Spinal cord (cervical c–1)
Brain – Hippocampus
Brain – Amygdala
Brain – Anterior cingulate cortex (BA24)
Brain – Cortex
Brain – Frontal cortex (BA9)

Figure 1

Figure 2

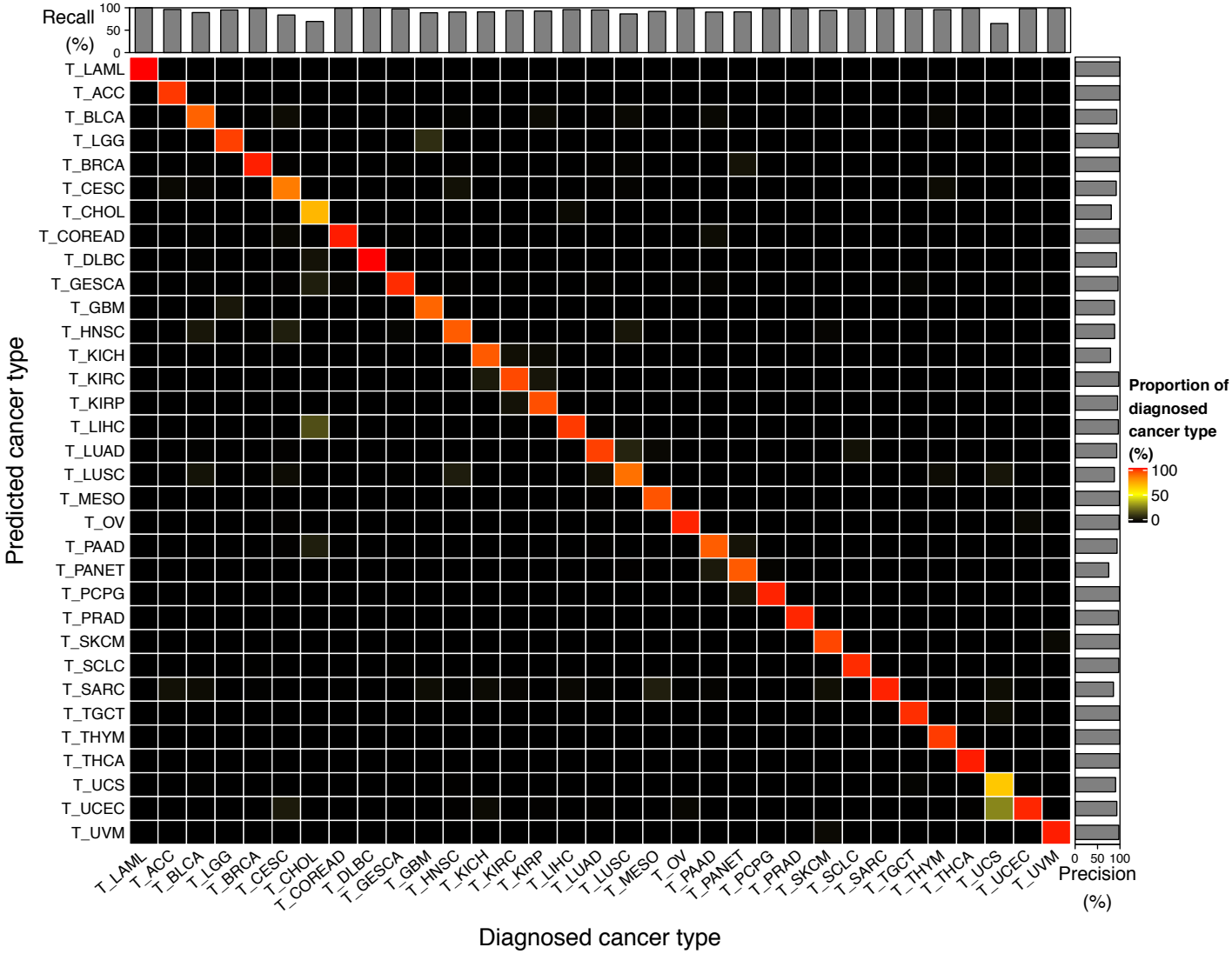
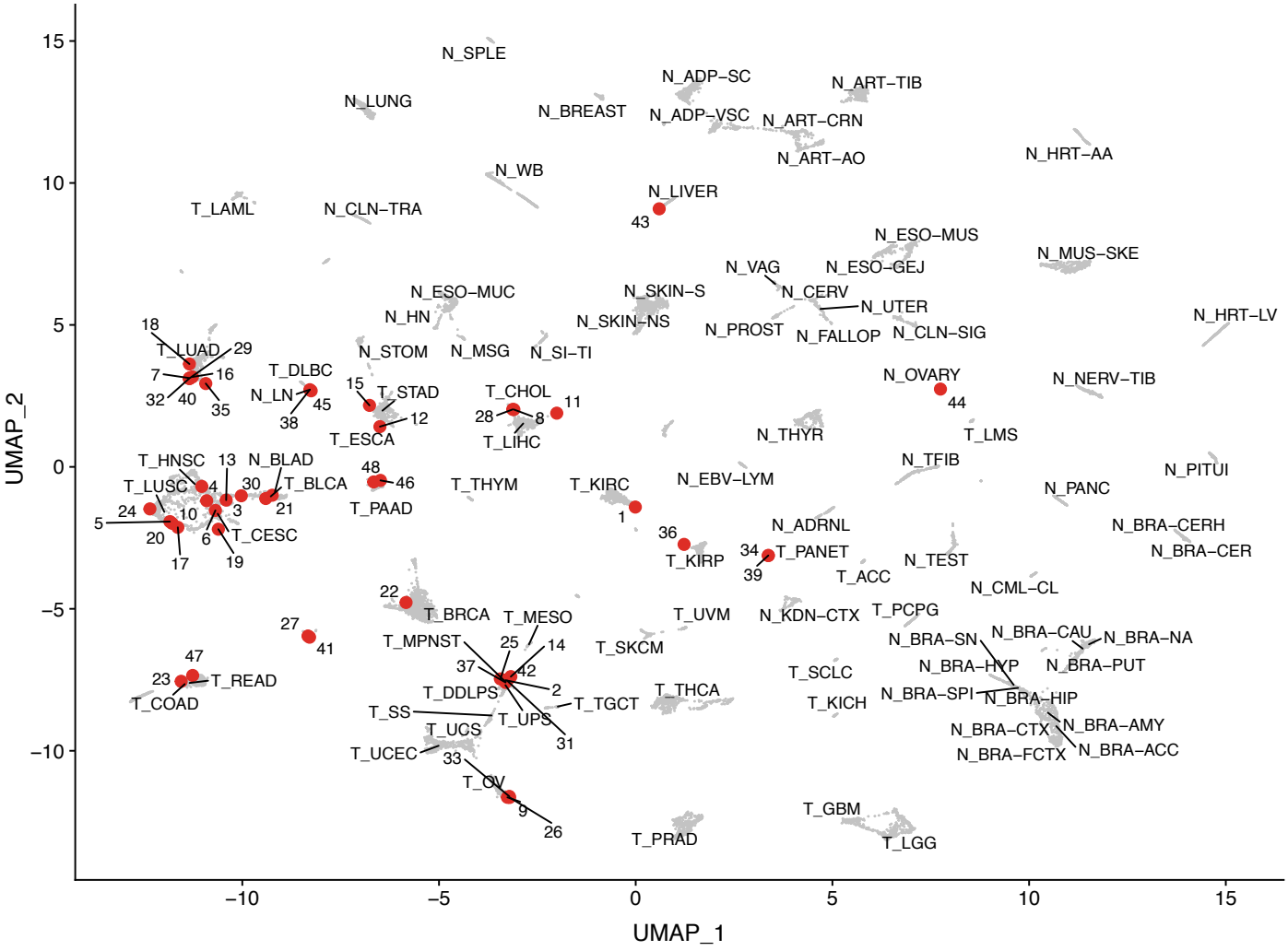

Recall (%)
100
50
0
Predicted cancer type
T_LAML
T_ACC
T_BLCA
T_LGG
T_BRCA
T_CESC
T_CHOL
T_COREAD
T_DLBC
T_GESCA
T_GBM
T_HNSC
T_KICH
T_KIRC
T_KIRP
T_LIHC
T_LUAD
T_LUSC
T_MESO
T_OV
T_PAAD
T_PANET
T_PCPG
T_PRAD
T_SKCM
T_SCLC
T_SARC
T_TGCT
T_THYM
T_THCA
T_UCS
T_UCEC
T_UVM
Diagnosed cancer type
Proportion of diagnosed cancer type (%)
100
50
0
0 50 100
Precision (%)

Figure 3


UMAP_2
UMAP_1
15
10
5
0
−5
−10
−10
−5
0
5
10
15
N_SPLE
N_LUNG
N_BREAST
N_ADP−SC
N_ADP−VSC
N_ART−TIB
N_ART−CRN
N_ART−AO
N_HRT−AA
N_WB
N_LIVER
43
T_LAML
N_CLN−TRA
N_ESO−MUS
N_MUS−SKE
N_VAG
N_ESO−GEJ
N_ESO−MUC
N_SKIN−S
N_CERV
N_UTER
N_HRT−LV
N_HN
N_SKIN−NS
N_PROST
N_FALLOP
N_CLN−SIG
18
T_LUAD
29
7
16
32
40
35
N_STOM
N_MSG
N_SI−TI
T_DLBC
N_LN
45
38
15
T_STAD
12
T_ESCA
T_CHOL
11
28
8
T_LIHC
N_OVARY
44
N_NERV−TIB
N_THYR
T_LMS
N_TFIB
N_PITUI
T_HNSC
13
N_BLAD
T_LUSC
4
30
T_BLCA
48
46
T_THYM
T_KIRC
N_EBV−LYM
N_PANC
24
10
3
21
T_PAAD
1
5
20
6
T_CESC
17
19
36
N_ADRNL
N_BRA−CERH
34
T_PANET
N_BRA−CER
39
T_KIRP
N_TEST
T_ACC
N_CML−CL
22
T_BRCA
T_MESO
T_UVM
N_KDN−CTX
T_PCPG
27
41
T_MPNST
25
14
T_SKCM
N_BRA−SN
N_BRA−CAU
N_BRA−NA
47
37
42
T_SCLC
N_BRA−HYP
N_BRA−PUT
23
T_READ
T_COAD
T_DDLPS
2
T_TGCT
T_THCA
T_KICH
N_BRA−SPI
N_BRA−HIP
T_SS
T_UPS
T_UCS
N_BRA−CTX
N_BRA−AMY
31
T_UCEC
33
N_BRA−FCTX
N_BRA−ACC
T_OV
9
26
T_GBM
T_PRAD
T_LGG