Exemple de projet : Classification des cancers de primitif inconnu par IA

Julien VIBERT

MD Medical Oncology - PhD Bioinformatics

DITEP - Gustave Roussy

U830 Institut Curie

IFSBM Module 11

Gustave Roussy

17 décembre 2024

Un patient adressé à Curie (patient n°1)

- Patient de 30 ans
- AEG rapide, masse abdominale, lésions osseuses diffuses
- Adressé à Curie pour suspicion de sarcome osseux
- Carcinome inclassé par l'anatomo-pathologie…

- Traitement prévu : Folfirinox (cancer du pancréas ?

Un patient adressé à Curie (patient n°1)

- Patient de 30 ans
- AEG rapide, masse abdominale, lésions osseuses diffuses
- Adressé à Curie pour suspicion de sarcome osseux
- Carcinome inclassé par l'anatomo-pathologie…

- Traitement prévu : Folfirinox (cancer du pancréas ?

RNA-seq disponible à Curie pour le diagnostic des sarcomes

RNA-seq

- **RNA-seq** : séquençage haut débit du transcriptome entier

- Plusieurs types de sarcomes sont caractérisés par des **fusions de gènes :**
  - sarcome d'Ewing : **EWSR1-FLI1**
- Le RNA-seq permet la détection de **toutes les fusions de gènes exprimées**

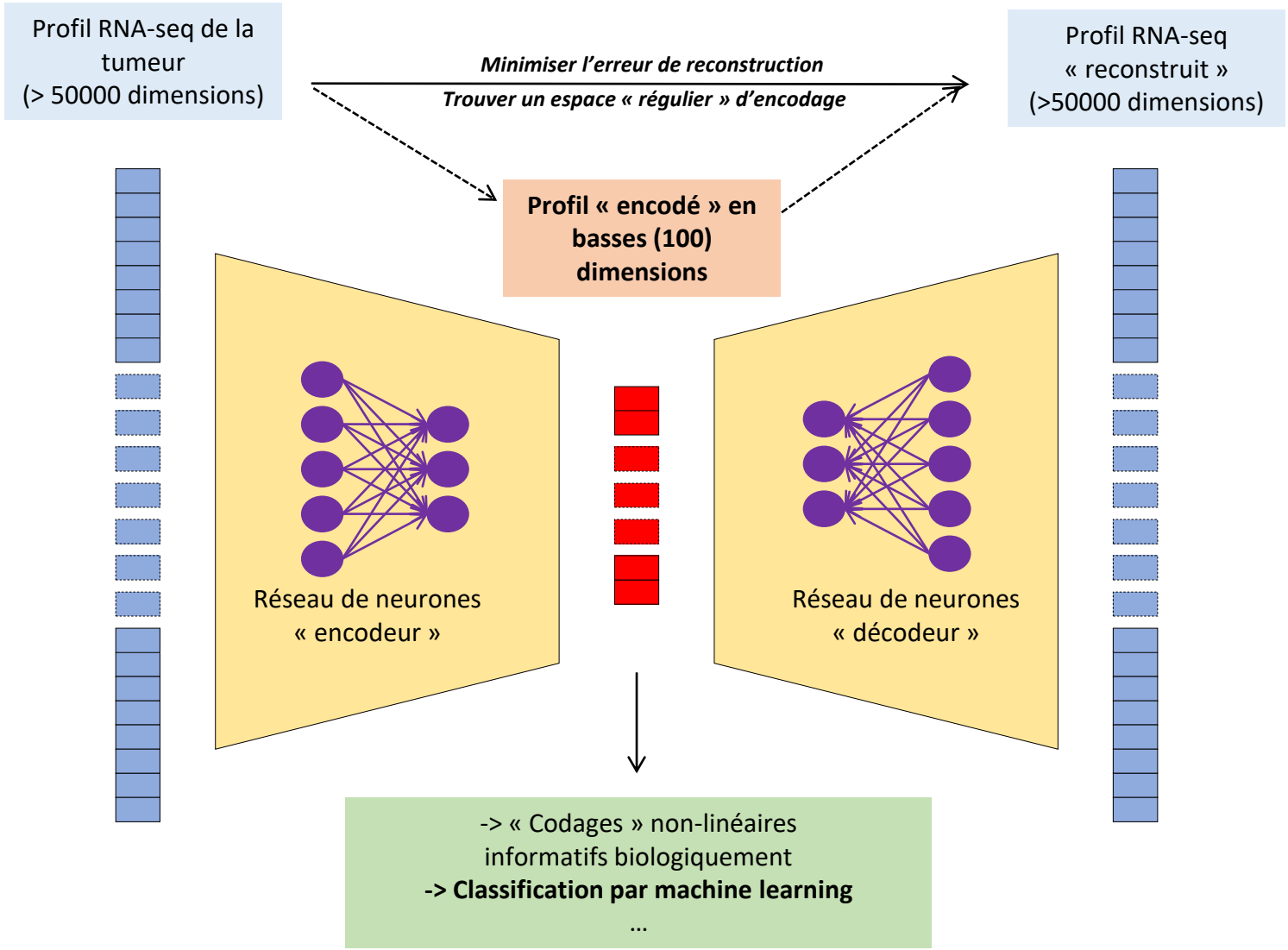
- **Le RNA-seq donne pour un échantillon tumoral donné l'expression de > 50.000 transcrits**

  -> « Big data »

Cancers de primitif inconnu

- **Cancers de primitif inconnu (CUP)** : 2-3 % des cancers métastatiques
- Aucun primitif retrouvé malgré toutes les explorations notamment anatomopathologiques et radiologiques
  -> Peut-on identifier le tissu d'origine à partir du RNA-seq (> 50000 dimensions) ?

Un outil d'intelligence artificielle : L'auto-encodeur variationnel (VAE)



Utilisation du VAE pour l'analyse d'images

Le VAE « apprend » à réduire à <u>des dimensions interprétables par l'humain</u>

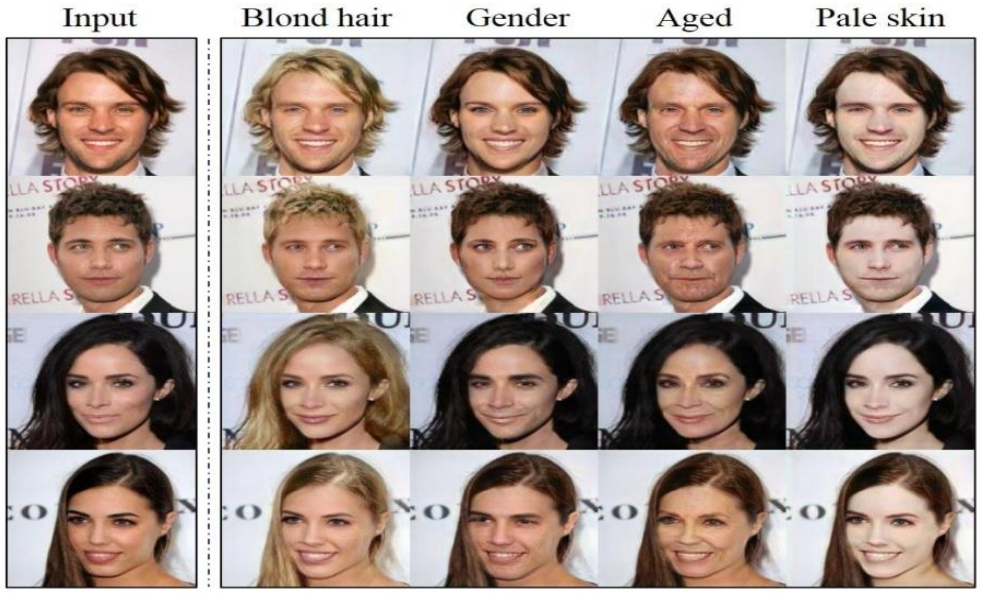
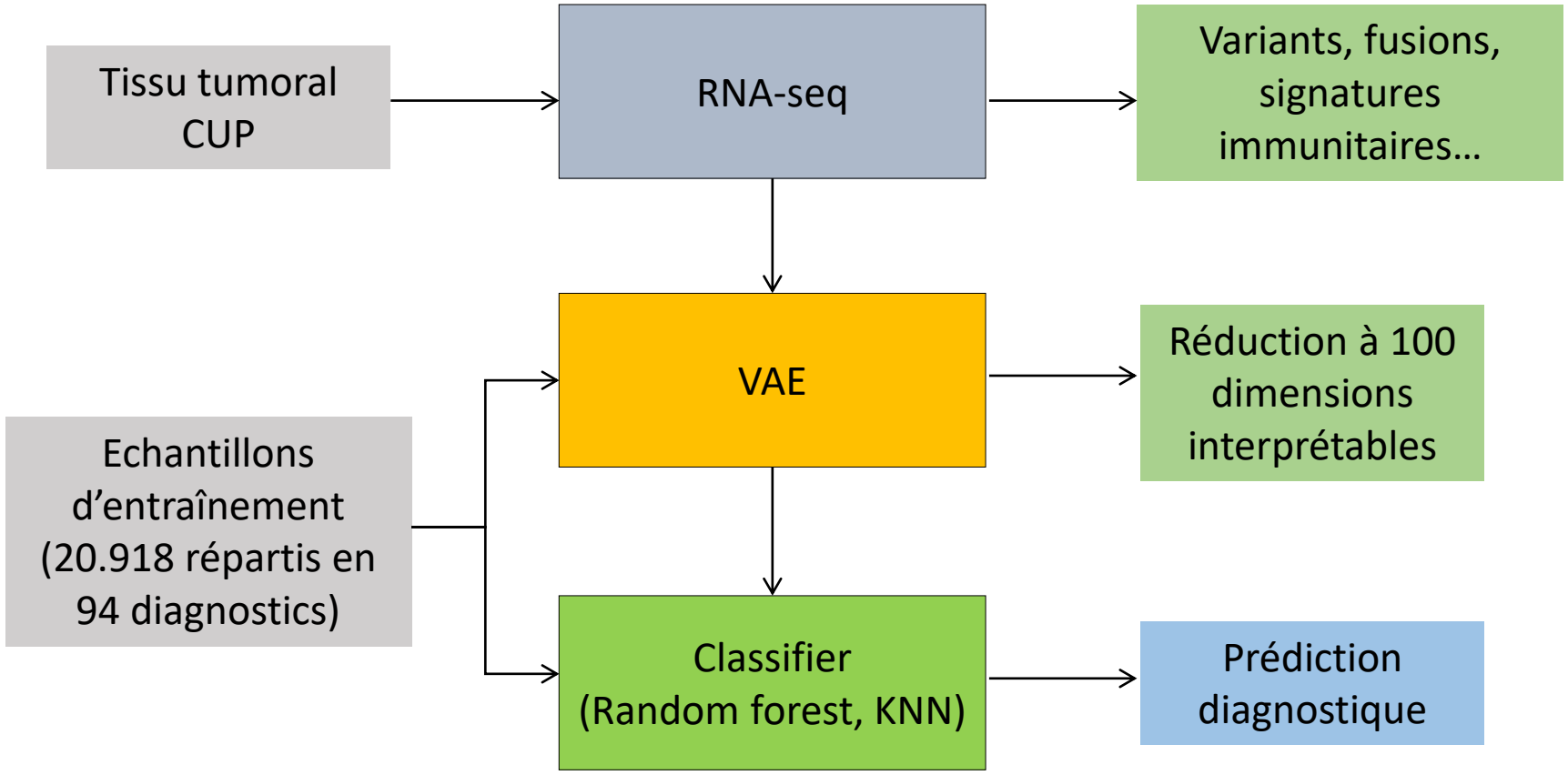


Schéma de l'étude et de l'outil « TransCUPtomics »



Training dataset

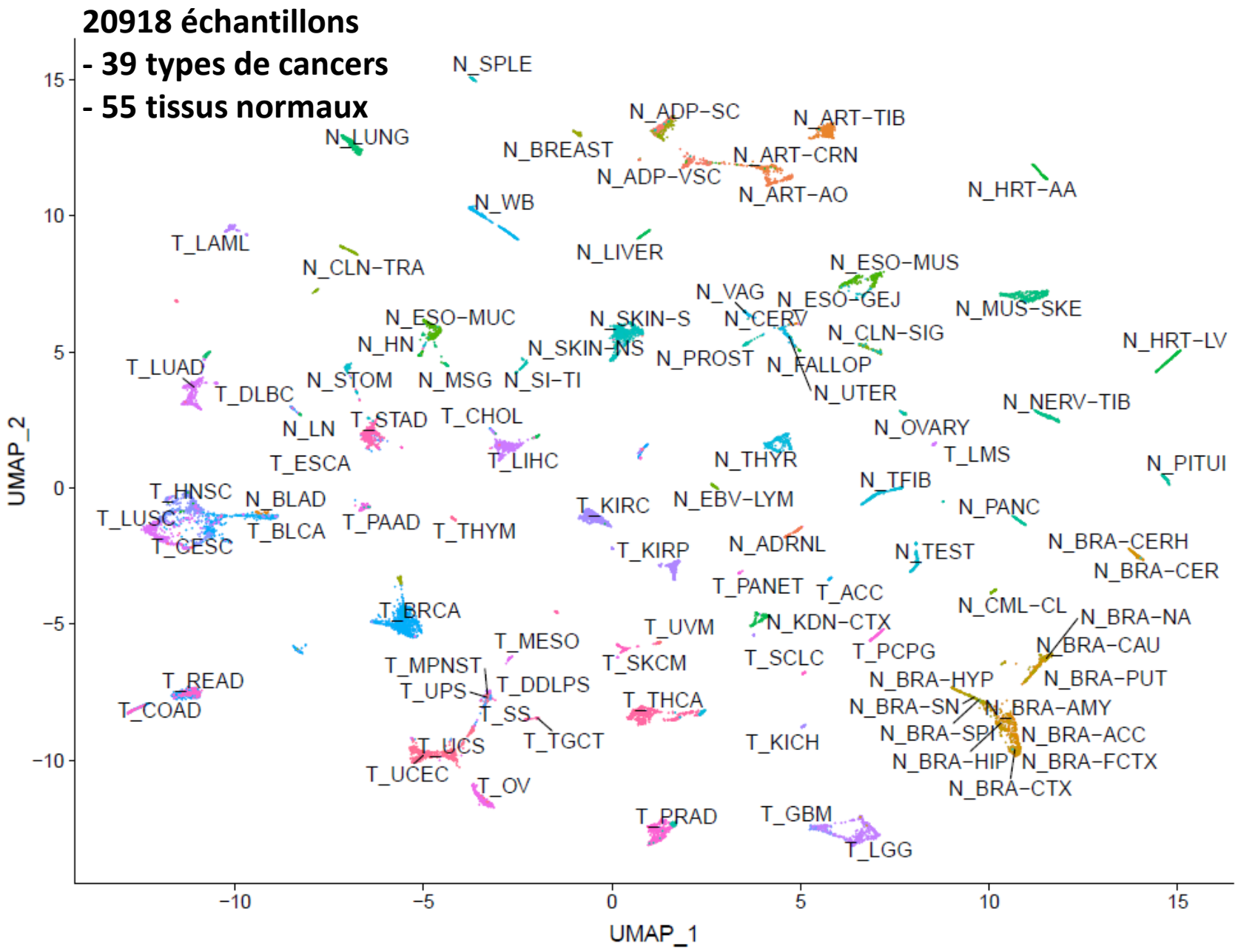
Databases : UGS, TCGA, GTEx, HPA

Tumors	Database	Number of samples	Number of subtypes
	TCGA	10,222	47
	UGS_cleaned	862	49
	PanNET	33	1
	Total	**11,117**	**97**

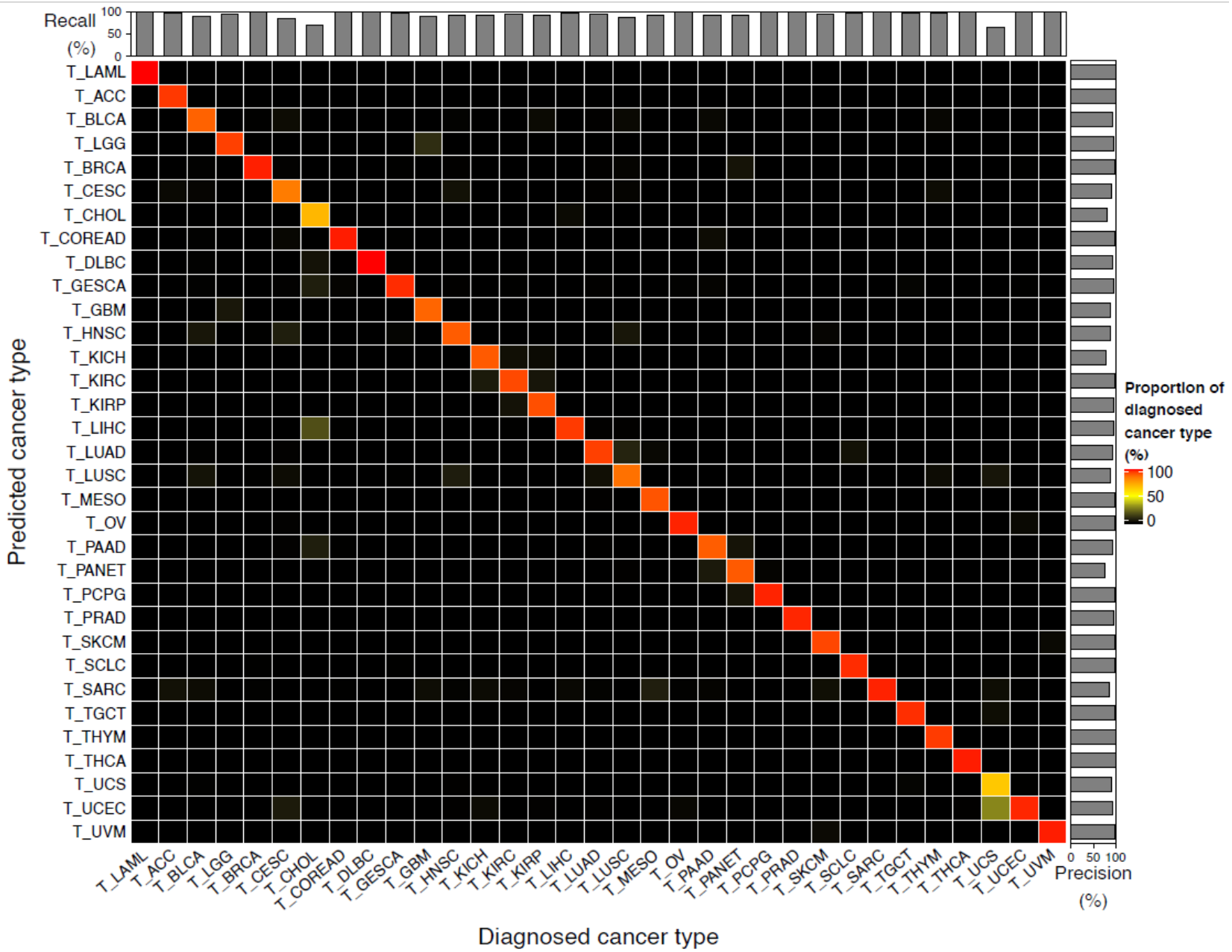
Normal Tissues	Database	Number of samples	Number of subtypes
	TCGA	742	24
	GTEx	9,778	54
	Human Protein Atlas	200	32
	Total	**10,720**	**110**

21,837 samples representing 207 tissue/ tumor subtypes

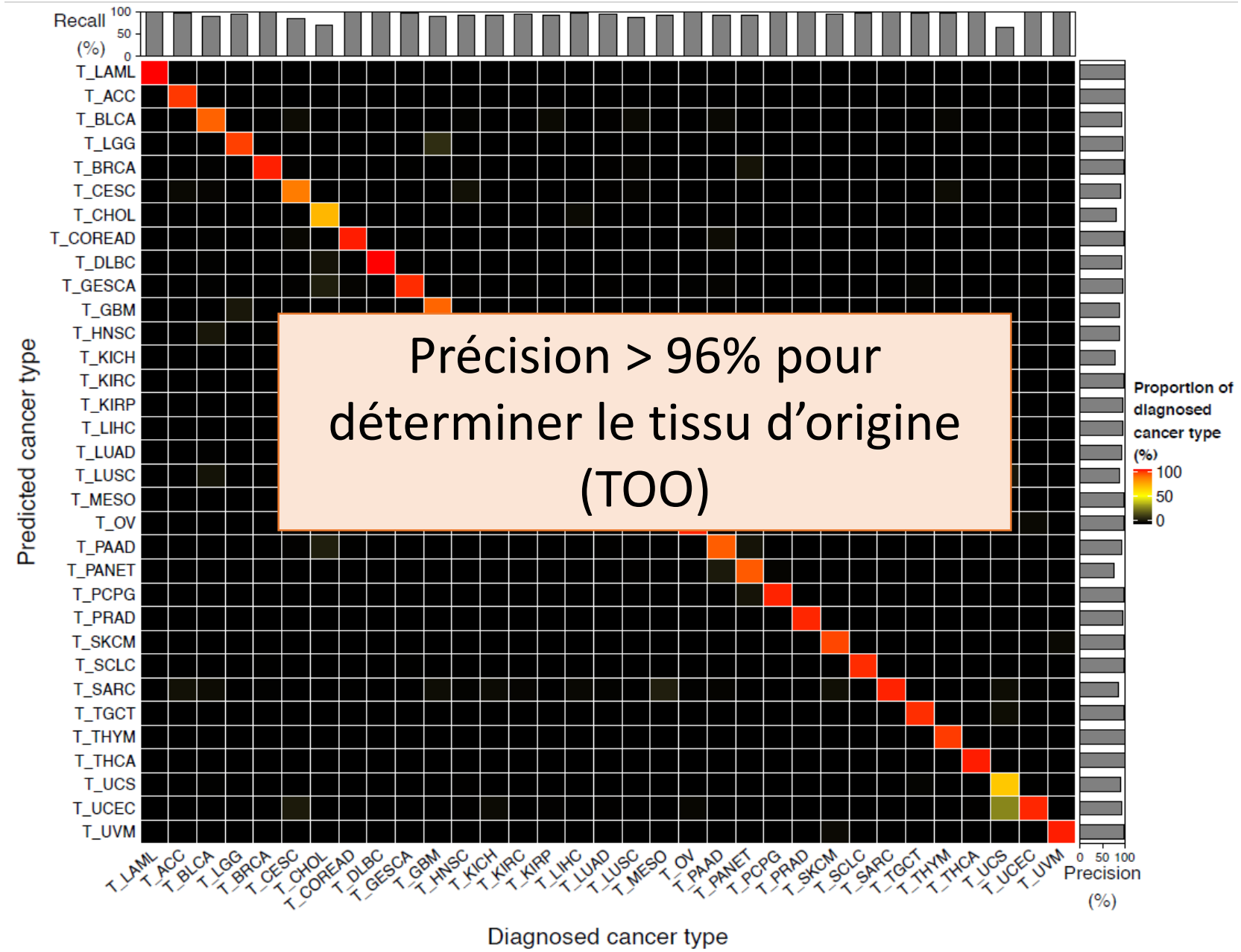
UMAP 2D des échantillons connus (TCGA, GTEx, HPA…)



Performance diagnostique sur les échantillons connus

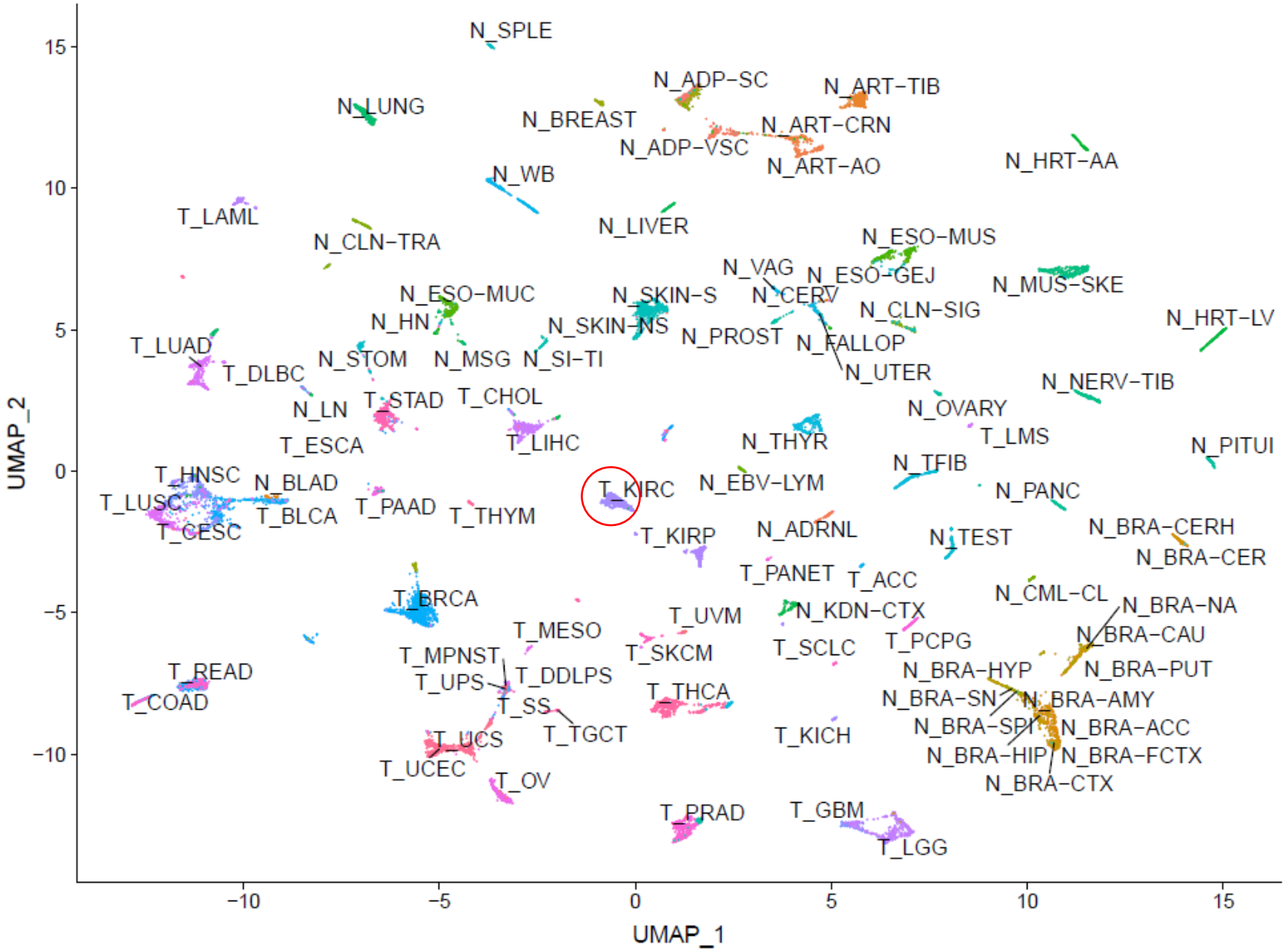


Performance diagnostique sur les échantillons connus



Précision > 96% pour déterminer le tissu d'origine (TOO)

Patient n°1


N_SPLE
N_ADP-SC
N_ART-TIB
N_LUNG
N_BREAST
N_ART-CRN
N_ADP-VSC
N_ART-AO
N_HRT-AA
N_WB
T_LAML
N_LIVER
N_CLN-TRA
N_ESO-MUS
N_VAG
N_ESO-GEJ
N_MUS-SKE
N_ESO-MUC
N_SKIN-S
N_CERV
N_CLN-SIG
N_HRT-LV
N_HN
N_SKIN-NS
N_PROST
N_FALLOP
T_LUAD
T_DLBC
N_STOM
N_MSG
N_SI-TI
N_UTER
N_NERV-TIB
N_LN
T_STAD
T_CHOL
N_OVARY
T_LMS
T_ESCA
T_LIHC
N_THYR
N_PITUI
N_TFIB
T_HNSC
N_BLAD
T_KIRC
N_EBV-LYM
T_LUSC
T_BLCA
T_PAAD
T_THYM
N_PANC
T_CESC
T_KIRP
N_ADRNL
N_TEST
N_BRA-CERH
N_BRA-CER
T_PANET
T_ACC
T_BRCA
N_CML-CL
N_BRA-NA
T_UVM
N_KDN-CTX
T_MESO
N_BRA-CAU
T_MPNST
T_SKCM
T_SCLC
T_PCPG
T_READ
T_DDLPS
N_BRA-HYP
N_BRA-PUT
T_UPS
T_THCA
T_COAD
T_SS
N_BRA-SN
N_BRA-AMY
T_TGCT
N_BRA-SPI
N_BRA-ACC
T_UCS
T_KICH
N_BRA-HIP
N_BRA-FCTX
T_UCEC
T_OV
N_BRA-CTX
T_PRAD
T_GBM
T_LGG
UMAP_2
UMAP_1
15
10
5
0
−5
−10
−10
−5
0
5
10
15

Patient n°1

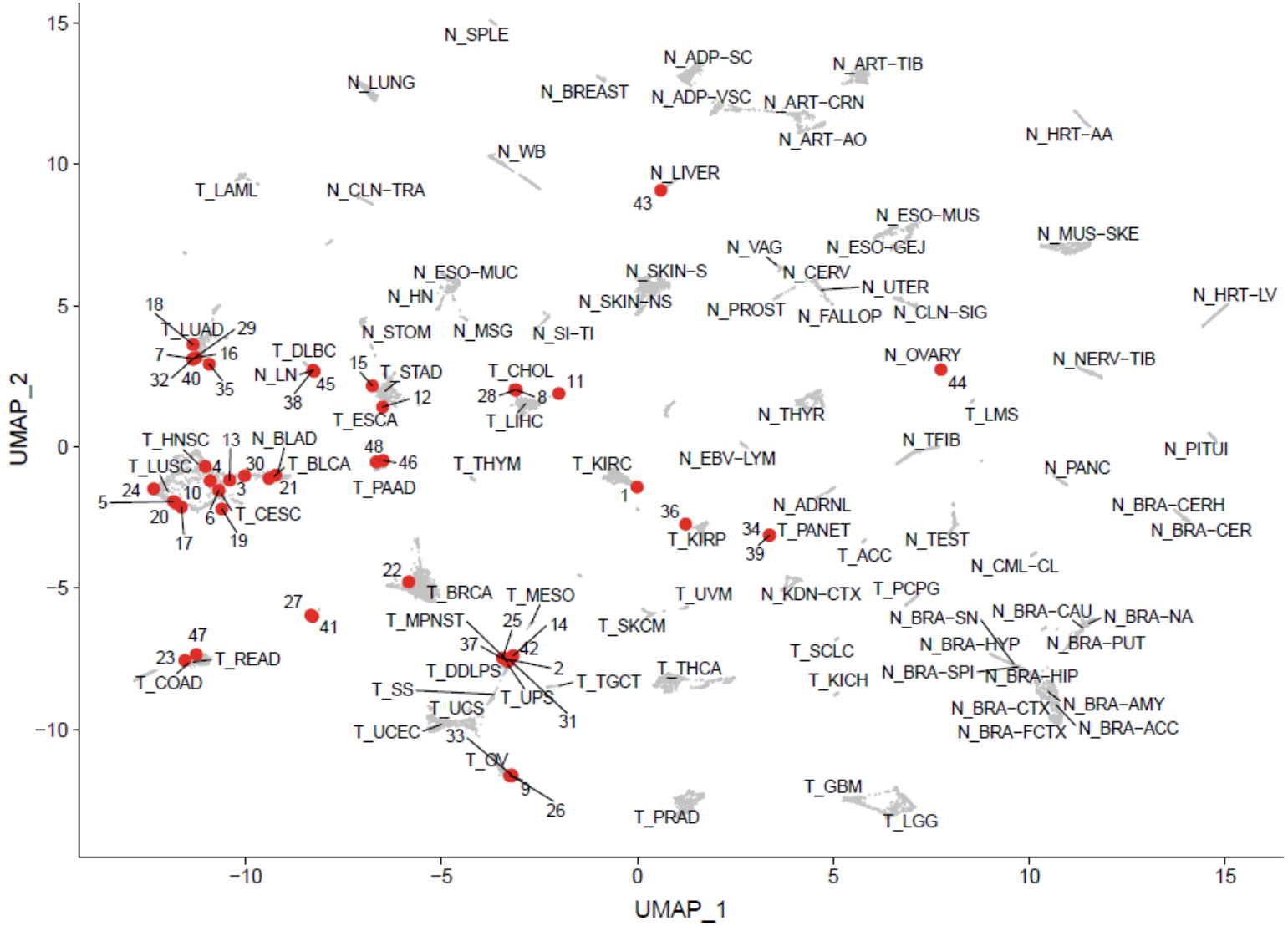
- **Carcinome rénal** extrarénal
- Patient traité par immunothérapie + anti-angiogénique

Patient n°1

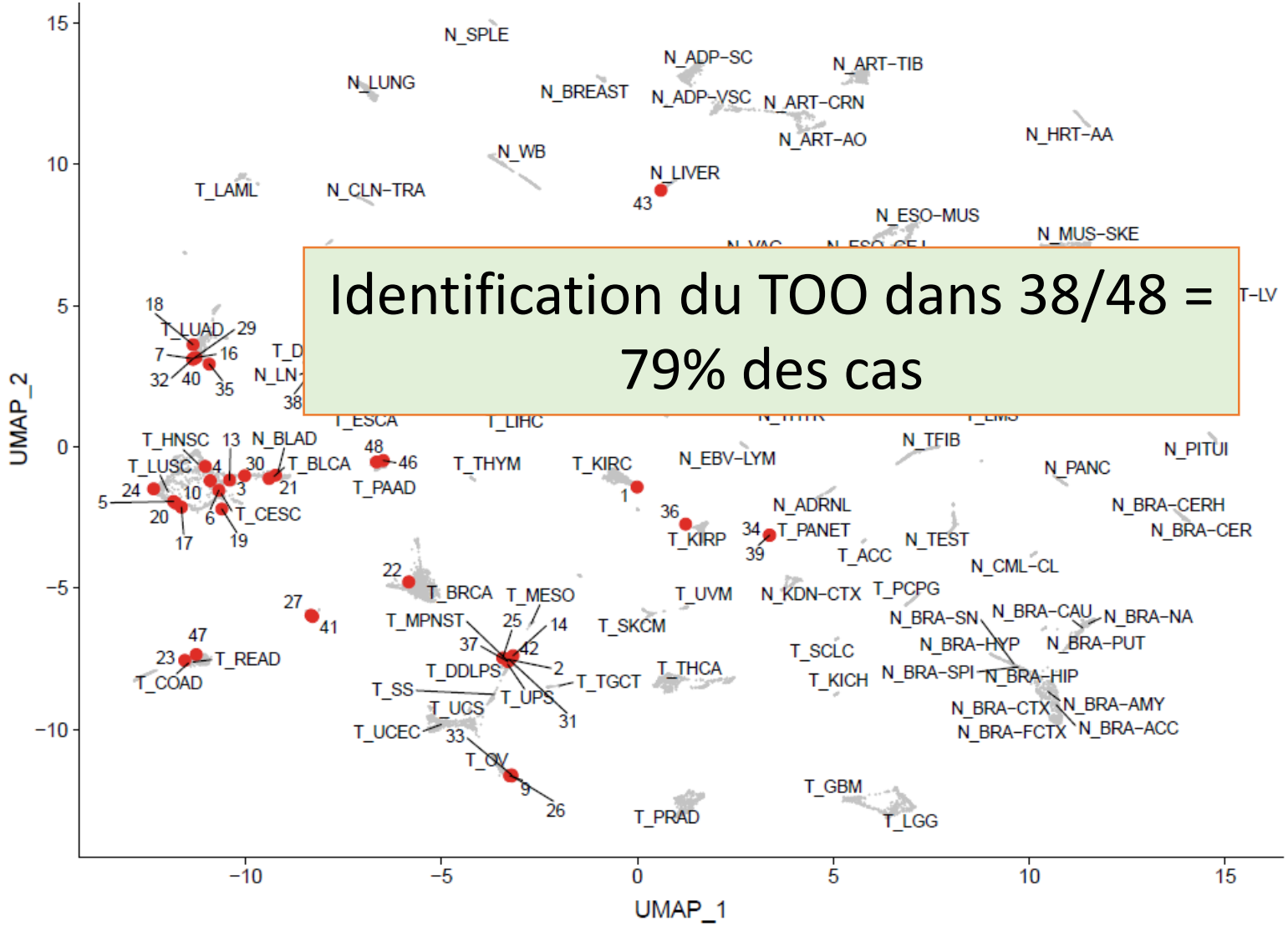
- **Carcinome rénal** extrarénal
- Patient traité par immunothérapie + anti-angiogénique

-> *Réponse complète*

Prédiction diagnostique sur une série de CUPs



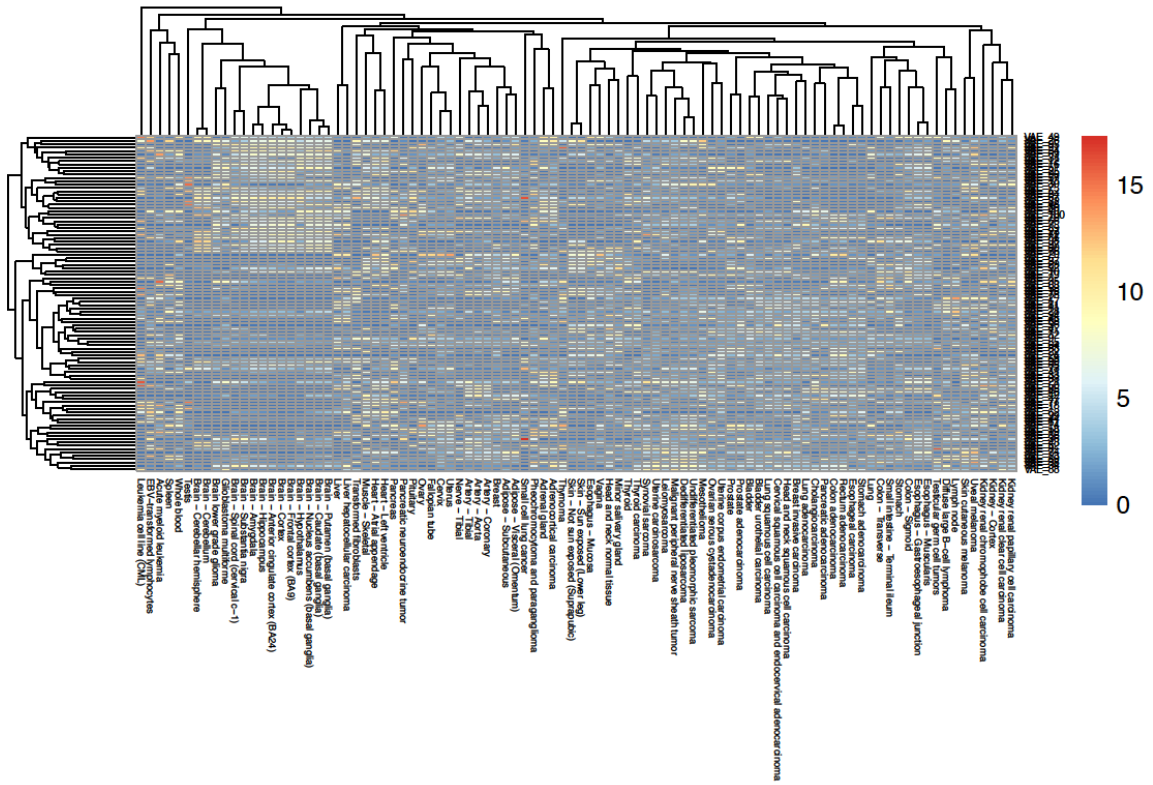
Prédiction diagnostique sur une série de CUPs



Interprétabilité

Dimensions du VAE associées à des processus biologiques :

- Kératinisation
- Neurogenèse
- Activation du système immunitaire…



The Journal of Molecular Diagnostics, Vol. 23, No. 10, October 2021

ELSEVIER

the Journal of Molecular Diagnostics

jmdjournal.org

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

Check for updates

Julien Vibert,[*] Gaëlle Pierron,[†] Camille Benoist,[‡] Nadège Gruel,[*§] Delphine Guillemot,[†] Anne Vincent-Salomon,[¶] Christophe Le Tourneau,[‖] Alain Livartowski,[**] Odette Mariani,[¶] Sylvain Baulande,[††] François-Clément Bidard,[**‡‡] Olivier Delattre,[*†] Joshua J. Waterfall,[§,§§] and Sarah Watson[*,**]

From the INSERM U830,[*] Équipe Labellisée Ligue Nationale Contre le Cancer, Diversity and Plasticity of Childhood Tumors Lab, PSL Research University, the Department of Translational Research,[§] PSL Research University, the Institut Curie Genomics of Excellence (ICGex) Platform,[††] PSL Research University, and the INSERM U830,[§§] PSL Research University, Institut Curie Research Center, Paris; the Somatic Genetics Unit,[†] Department of Genetics, the Clinical Bioinformatic Unit,[‡] Department of Diagnostic and Theranostic Medecine, the Department of Diagnostic and Theranostic Medecine,[¶] and the Department of Medical Oncology,[**] Institut Curie Hospital, Paris; the Department of Drug Development and Innovation,[‖] INSERM U900, Paris-Saclay University, Institut Curie Hospital and Research Center, Paris and Saint-Cloud; and the INSERM CIC-BT 1428,[‡‡] UVSQ, Paris-Saclay University, Saint-Cloud, France

Accepted for publication July 14, 2021.

Address correspondence to Sarah Watson, M.D., Ph.D., 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.

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 to predict tissue of origin based on RNA-sequencing 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 11 prospective patients. TransCUPtomics exhibited an overall accuracy of 96% on reference data for TOO prediction. The TOO could be identified in 38 (79%) of 48 CUP patients. Eight of 11 prospective CUP patients (73%) could receive first-line therapy guided by TransCUPtomics prediction, with responses observed in most patients. The variational autoencoder 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. *(J Mol Diagn 2021, 23: 1380–1392; https://doi.org/10.1016/j.jmoldx.2021.07.009)*

Huit des 11 patients prospectifs atteints de CUP ***(73 %) ont pu recevoir une thérapie de première ligne guidée par la prédiction de TransCUPtomics,*** *avec des réponses observées chez la plupart des patients*

The Journal of Molecular Diagnostics, Vol. 23, No. 10, October 2021

ELSEVIER

the Journal of Molecular Diagnostics

jmdjournal.org

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

Julien Vibert,[*] Gaëlle Pierron,[†] Camille Benoist,[‡] Nadège Gruel,[*§] Delphine Guillemot,[†] Anne Vincent-Salomon,[¶] Christophe Le Tourneau,[‖] Alain Livartowski,[**] Odette Mariani,[¶] Sylvain Baulande,[††] François-Clément Bidard,[**‡‡] Olivier Delattre,[*†] Joshua J. Waterfall,[§,§§] and Sarah Watson[*,**]

From the INSERM U830,[*] Équipe Labellisée Ligue Nationale Contre le Cancer, Diversity and Plasticity of Childhood Tumors Lab, PSL Research University, the Department of Translational Research,[§] PSL Research University, the Institut Curie Genomics of Excellence (ICGex) Platform,[††] PSL Research University, and the INSERM U830,[§§] PSL Research University, Institut Curie Research Center, Paris; the Somatic Genetics Unit,[†] Department of Genetics, the Clinical Bioinformatic Unit,[‡] Department of Diagnostic and Theranostic Medecine, the Department of Diagnostic and Theranostic Medecine,[¶] and the Department of Medical Oncology,[**] Institut Curie Hospital, Paris; the Department of Drug Development and Innovation,[‖] INSERM U900, Paris-Saclay University, Institut Curie Hospital and Research Center, Paris and Saint-Cloud; and the INSERM CIC-BT 1428,[‡‡] UVSQ, Paris-Saclay University, Saint-Cloud, France

Accepted for publication July 14, 2021.

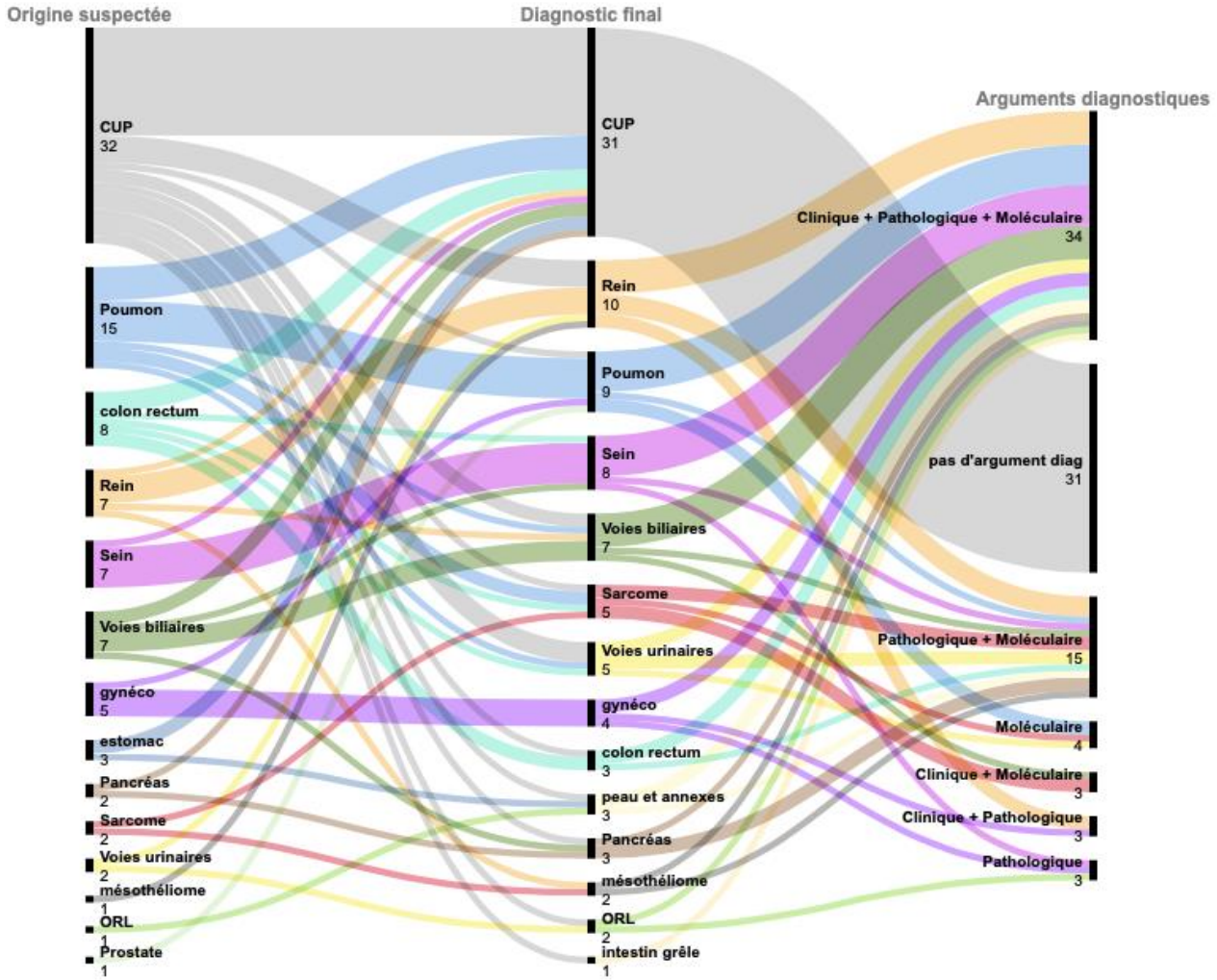
Address correspondence to Sarah Watson, M.D., Ph.D., INSERM U830, Équipe Labellisée Ligue Nationale Contre le Cancer, Diversity and Plasticity of Childhood Tumors Lab, PSL Research University, Institut Curie Research Center, Paris, France, Department of Medical Oncology, Institut Curie Hospital, Paris, France, 26 rue d'Ulm 75005 Paris, France. E-mail: sarah.watson@curie.fr.

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 to predict tissue of origin based on RNA-sequencing 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 11 prospective patients. TransCUPtomics exhibited an overall accuracy of 96% on reference data for TOO prediction. The TOO could be identified in 38 (79%) of 48 CUP patients. Eight of 11 prospective CUP patients (73%) could receive first-line therapy guided by TransCUPtomics prediction, with responses observed in most patients. The variational autoencoder 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. *(J Mol Diagn 2021, 23: 1380–1392; https://doi.org/10.1016/j.jmoldx.2021.07.009)*

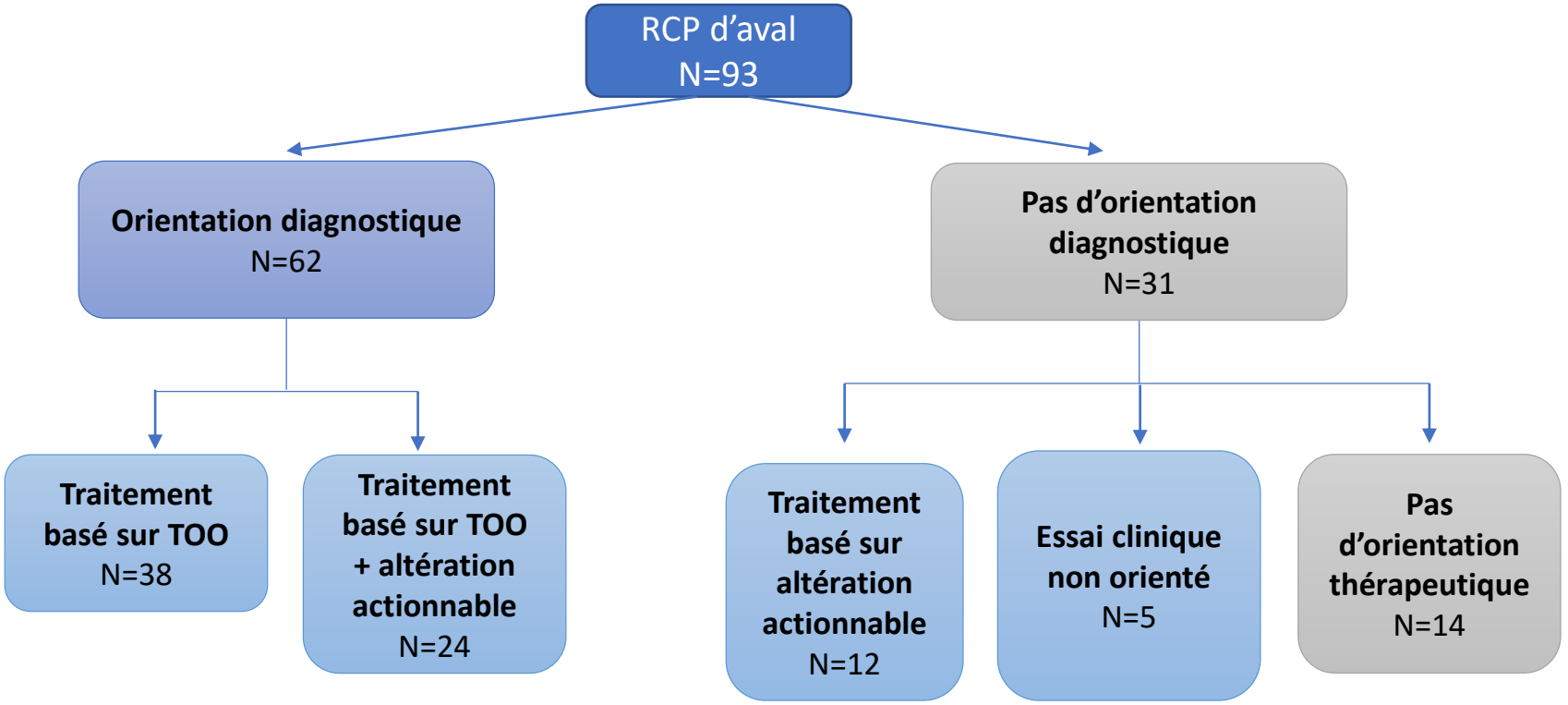
Huit des 11 patients prospectifs atteints de CUP ***(73 %) ont pu recevoir une thérapie de première ligne guidée par la prédiction de TransCUPtomics,*** *avec des réponses observées chez la plupart des patients*

TransCUPtomics est actuellement déployé sur le Plan **France Médecine Génomique 2025**, un programme national de séquençage des patients CUP

RCP nationale CUP / Plan France Médecine Génomique 2025
(Sarah Watson)



Bilan de la RCP nationale CUP (Sarah Watson)



Au total: 79 patients (85%) ont reçu une orientation thérapeutique

Inscription: RCPmolCurieSeqoia@curie.fr

Le Monde

Consulter le journal

ACTUALITÉS ÉCONOMIE VIDÉOS DÉBATS CULTURE LE GOÛT DU M

SCIENCES • MÉDECINE

Quand l'intelligence artificielle permet d'identifier l'origine inconnue d'un cancer métastasé

Un jeune homme de 30 ans présentant un cancer métastasé d'origine inconnue a été le premier à tester un outil d'intelligence artificielle développé à l'Institut Curie. Le crible a permis d'identifier le rein comme l'organe présentant la tumeur d'origine et le traitement spécifique qui lui a permis de guérir.

Le Parisien

Décryptage Société, Santé

Comment l'intelligence artificielle a identifié le cancer introuvable de Wilfrid

Les médecins ne trouvaient pas l'origine du cancer métastasé de ce quinquagénaire. Une IA, développée par l'Institut Curie, a estimé qu'« à 90 % » la tumeur initiale se trouvait dans les reins. L'outil est mis à l'honneur mardi dans un congrès d'oncologie aux États-Unis.

LE FIGARO Lire le journal

france inter Grille des programmes Podcasts Info Culture Humour

Une intelligence artificielle devine les cancers masqués, le Parisien.

Lundi 17 avril 2023

ÉCOUTER (5 MIN)

Tech & web

Actualités tech Crypto Start-up Tests Pratique Jeux vidéos

DOSSIER

Intelligence artificielle: tout ce qu'il faut savoir sur cette nouvelle révolution

Réservé aux abonnés

L'intelligence artificielle, un formidable outil pour sauver des vies

Par Pascal Grandmaison

Publié le 17/02/2023 à 06:01

Le Dr Sarah Watson, de l'Institut Curie. *Christophe LEPETIT pour Le Figaro Magazine*

DÉCRYPTAGE - En révélant des anomalies inaccessibles aux facultés humaines, l'intelligence artificielle s'impose comme l'assistant numéro 1 du médecin.

D’autres algorithmes d’IA pour la classification des CUPs

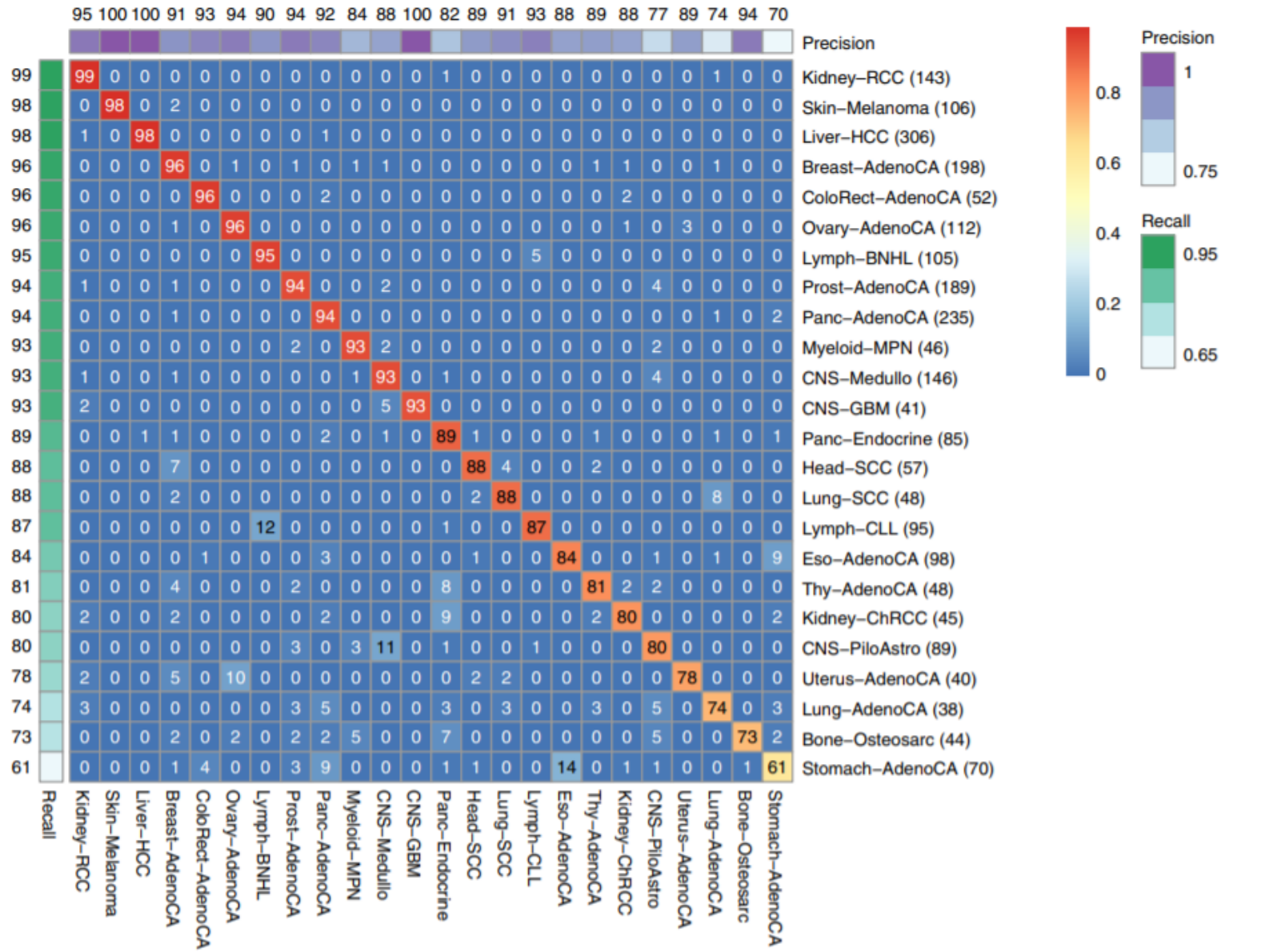
Deep learning sur des données de séquençage du génome

ARTICLE

https://doi.org/10.1038/s41467-019-13825-8 OPEN

A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns

Wei Jiao[1,63], Gurnit Atwal[1,2,3,63], Paz Polak[4,63], Rosa Karlic[5], Edwin Cuppen[6,7], PCAWG Tumor Subtypes and Clinical Translation Working Group, Alexandra Danyi[8], Jeroen de Ridder[8], Carla van Herpen[9], Martijn P. Lolkema[10], Neeltje Steeghs[11], Gad Getz[12], Quaid Morris[1,2,3,13,14,64], Lincoln D. Stein[1,2,64*] & PCAWG Consortium


95 100 100 91 93 94 90 94 92 84 88 100 82 89 91 93 88 89 88 77 89 74 94 70
Precision
99 Kidney–RCC (143)
98 Skin–Melanoma (106)
98 Liver–HCC (306)
96 Breast–AdenoCA (198)
96 ColoRect–AdenoCA (52)
96 Ovary–AdenoCA (112)
95 Lymph–BNHL (105)
94 Prost–AdenoCA (189)
94 Panc–AdenoCA (235)
93 Myeloid–MPN (46)
93 CNS–Medullo (146)
93 CNS–GBM (41)
89 Panc–Endocrine (85)
88 Head–SCC (57)
88 Lung–SCC (48)
87 Lymph–CLL (95)
84 Eso–AdenoCA (98)
81 Thy–AdenoCA (48)
80 Kidney–ChRCC (45)
80 CNS–PiloAstro (89)
78 Uterus–AdenoCA (40)
74 Lung–AdenoCA (38)
73 Bone–Osteosarc (44)
61 Stomach–AdenoCA (70)
Recall
Kidney–RCC
Skin–Melanoma
Liver–HCC
Breast–AdenoCA
ColoRect–AdenoCA
Ovary–AdenoCA
Lymph–BNHL
Prost–AdenoCA
Panc–AdenoCA
Myeloid–MPN
CNS–Medullo
CNS–GBM
Panc–Endocrine
Head–SCC
Lung–SCC
Lymph–CLL
Eso–AdenoCA
Thy–AdenoCA
Kidney–ChRCC
CNS–PiloAstro
Uterus–AdenoCA
Lung–AdenoCA
Bone–Osteosarc
Stomach–AdenoCA
0
0.2
0.4
0.6
0.8
Precision
1
0.75
Recall
0.95
0.65

Deep learning sur de l'anatomopathologie

Article

AI-based pathology predicts origins for cancers of unknown primary

https://doi.org/10.1038/s41586-021-03512-4

Received: 27 June 2020

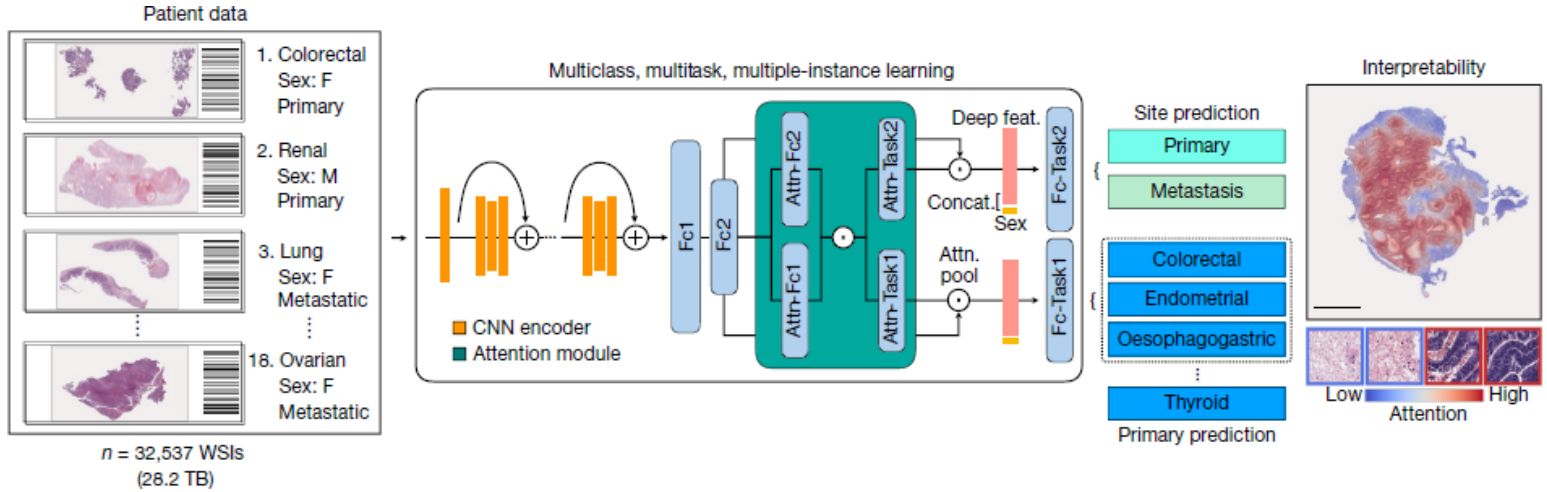
Accepted: 1 April 2021

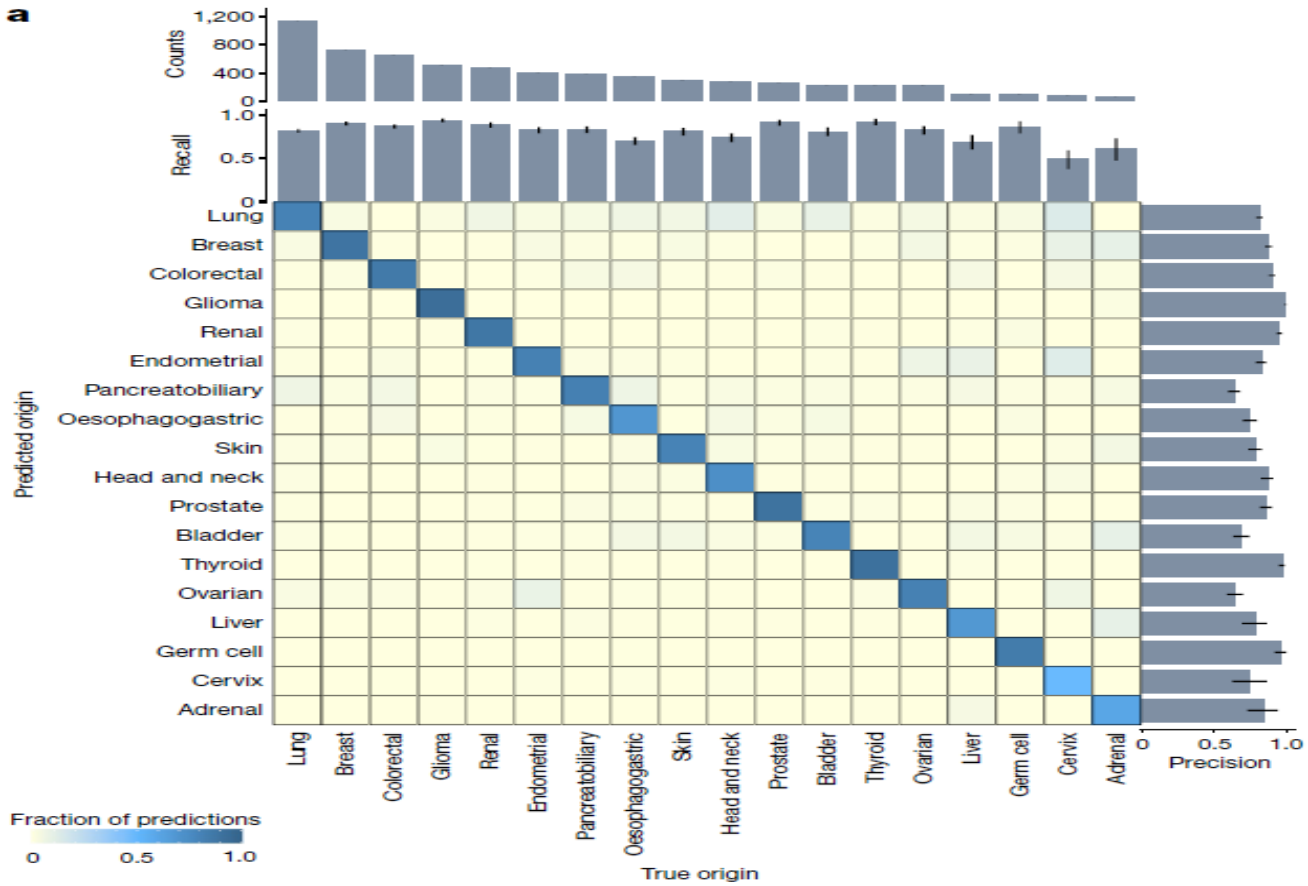
Published online: 5 May 2021

Check for updates

Ming Y. Lu[1,2,3], Tiffany Y. Chen[1,2,5], Drew F. K. Williamson[1,2,5], Melissa Zhao[1], Maha Shady[1,2,3,4], Jana Lipkova[1,2,3] & Faisal Mahmood[1,2,3]✉

Cancer of unknown primary (CUP) origin is an enigmatic group of diagnoses in which the primary anatomical site of tumour origin cannot be determined[1,2]. This poses a considerable challenge, as modern therapeutics are predominantly specific to the primary tumour[3]. Recent research has focused on using genomics and transcriptomics to identify the origin of a tumour[4–9]. However, genomic testing is not always performed and lacks clinical penetration in low-resource settings. Here, to overcome these challenges, we present a deep-learning-based algorithm—Tumour Origin Assessment via Deep Learning (TOAD)—that can provide a differential diagnosis for the origin of the primary tumour using routinely acquired histology slides. We used whole-slide images of tumours with known primary origins to train a model that simultaneously identifies the tumour as primary or metastatic and predicts its site of origin. On our held-out test set of tumours with known primary origins, the model achieved a top-1 accuracy of 0.83 and a top-3 accuracy of 0.96, whereas on our external test set it achieved top-1 and top-3 accuracies of 0.80 and 0.93, respectively. We further curated a dataset of 317 cases of CUP for which a differential diagnosis was assigned. Our model predictions resulted in concordance for 61% of cases and a top-3 agreement of 82%. TOAD can be used as an assistive tool to assign a differential diagnosis to complicated cases of metastatic tumours and CUPs and could be used in conjunction with or in lieu of ancillary tests and extensive diagnostic work-ups to reduce the occurrence of CUP.


Patient data
1. Colorectal
Sex: F
Primary
2. Renal
Sex: M
Primary
3. Lung
Sex: F
Metastatic
18. Ovarian
Sex: F
Metastatic
n = 32,537 WSIs
(28.2 TB)
Multiclass, multitask, multiple-instance learning
CNN encoder
Attention module
Fc1
Fc2
Attn-Fc2
Attn-Fc1
Attn-Task2
Attn-Task1
Deep feat.
Concat.
Sex
Attn. pool
Fc-Task2
Fc-Task1
Site prediction
Primary
Metastasis
Colorectal
Endometrial
Oesophagogastric
Thyroid
Primary prediction
Interpretability
Low
High
Attention


a
Counts
1,200
800
400
0
Recall
1.0
0.5
0
Lung
Breast
Colorectal
Glioma
Renal
Endometrial
Pancreatobiliary
Oesophagogastric
Skin
Head and neck
Prostate
Bladder
Thyroid
Ovarian
Liver
Germ cell
Cervix
Adrenal
Predicted origin
True origin
Precision
0
0.5
1.0
Fraction of predictions
0
0.5
1.0

XGBoost sur un panel de NGS ciblé

nature medicine

Article

https://doi.org/10.1038/s41591-023-02482-6

Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary

Received: 6 January 2023

Accepted: 30 June 2023

Published online: 7 August 2023

Check for updates

Intae Moon[1,2], Jaclyn LoPiccolo[3], Sylvan C. Baca[3,4], Lynette M. Sholl[5], Kenneth L. Kehl[2], Michael J. Hassett[2], David Liu[2,3,6], Deborah Schrag[7] & Alexander Gusev[2,6,8]

Cancer of unknown primary (CUP) is a type of cancer that cannot be traced back to its primary site and accounts for 3–5% of all cancers. Established targeted therapies are lacking for CUP, leading to generally poor outcomes. We developed OncoNPC, a machine-learning classifier trained on targeted next-generation sequencing (NGS) data from 36,445 tumors across 22 cancer types from three institutions. Oncology NGS-based primary cancer-type classifier (OncoNPC) achieved a weighted F1 score of 0.942 for high confidence predictions (≥ 0.9) on held-out tumor samples, which made up 65.2% of all the held-out samples. When applied to 971 CUP tumors collected at the Dana-Farber Cancer Institute, OncoNPC predicted primary cancer types with high confidence in 41.2% of the tumors. OncoNPC also identified CUP subgroups with significantly higher polygenic germline risk for the predicted cancer types and with significantly different survival outcomes. Notably, patients with CUP who received first palliative intent treatments concordant with their OncoNPC-predicted cancers had significantly better outcomes (hazard ratio (HR) = 0.348; 95% confidence interval (CI) = 0.210–0.570; $P = 2.32 \times 10^{-5}$). Furthermore, OncoNPC enabled a 2.2-fold increase in patients with CUP who could have received genomically guided therapies. OncoNPC thus provides evidence of distinct CUP subgroups and offers the potential for clinical decision support for managing patients with CUP.

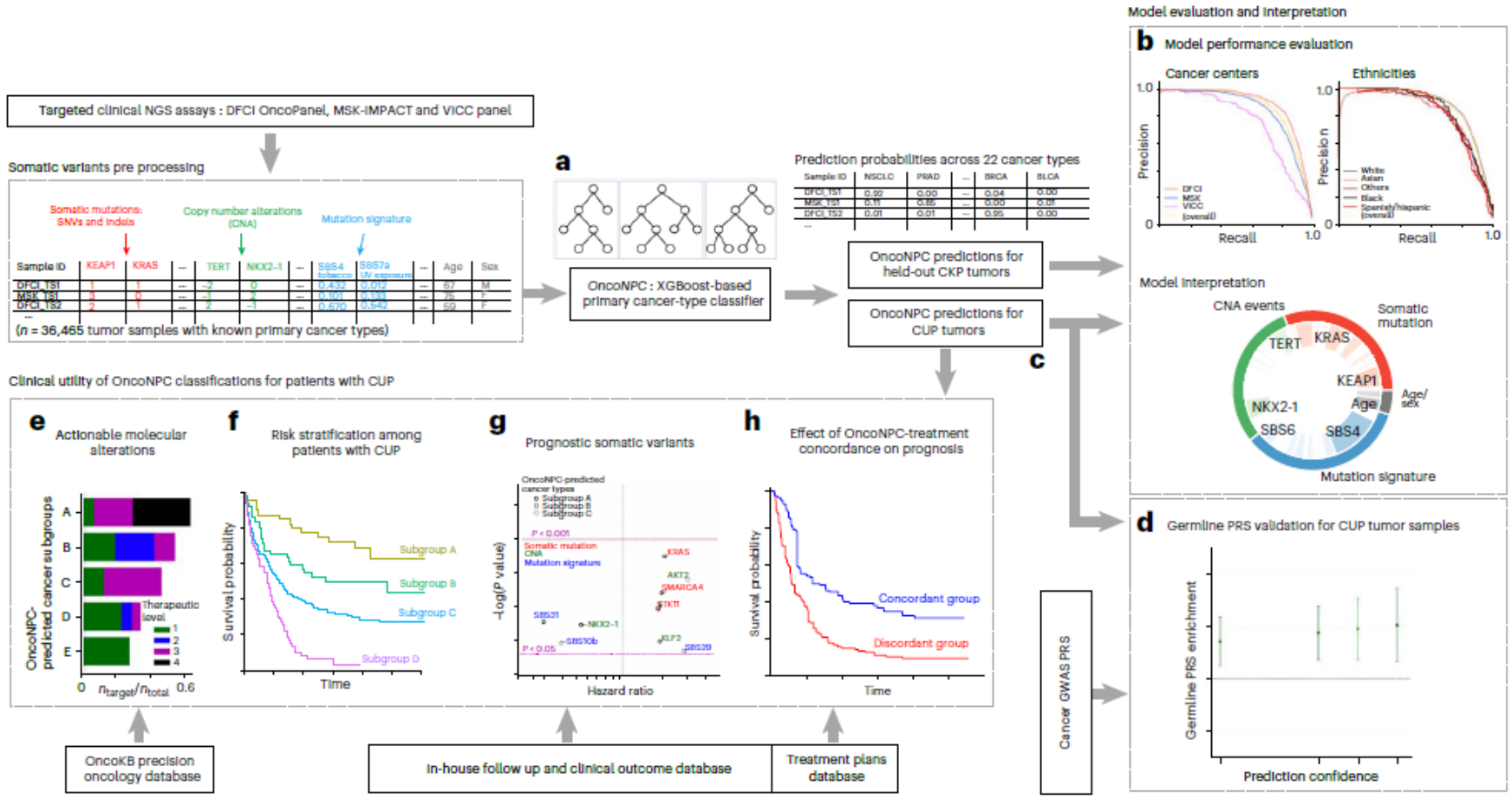


Fig. 1 | Overview of model development and analysis workflow. a, OncoNPC, an XGBoost-based classifier, was trained and evaluated using 36,465 cancer with known primary (CKP) tumor samples across 22 cancer types collected from three different cancer centers. **b**, OncoNPC performance was evaluated on the held-out tumor samples ($n = 7,289$). **c**, OncoNPC was applied to 971 CUP tumor samples at a single institution to predict primary cancer types. **d**–**g**, OncoNPC-predicted CUP subgroups were then investigated for association with elevated germline risk (**d**), actionable molecular alterations (**e**), overall survival (**f**) and prognostic somatic features (**g**). **h**, A subset of CUP patients with detailed treatment data was evaluated for treatment-specific outcomes.

Acknowledgments

- Sarah Watson
- My thesis directors:
  Olivier Delattre and Joshua Waterfall
- U830 Institut Curie
- Unité de Génétique Somatique (UGS)
- Patients and families
- Doctors and other healthcare workers

- **THANK YOU FOR**
- **YOUR ATTENTION!**

Courir pour Mathieu