Title: DNA sequencing and gene-expression profiling assists in making a tissue of origin diagnosis in cancer of unknown primary

Authors
Atara Posner¹, Owen W.J. Prall², Tharani Sivakumaran³, Dariush Etemadamoghadam⁵, Niko Thio⁵, Andrew Pattison¹, Shiva Balachander¹, Krista Fisher⁵, Samantha Webb⁵, Colin Wood⁵, Anna DeFazio⁶,⁷,⁸, Nicholas Wilcken⁹, Bo Gao⁹, Christos S Karapetis¹⁰, Madhu Singh¹¹, Ian M Collins¹², Gary Richardson¹³, Christopher Steer¹⁴, Mark Warren¹⁵, Narayan Karanth¹⁶, Gavin Wright¹⁷, Scott Williams¹⁸, Joshy George¹⁹, Rodney J Hicks¹⁰, Alex Boussioutas²¹, Anthony J Gill²², Benjamin J. Solomon³, Huiling Xu², Andrew Fellowes², Stephen B Fox²,⁴, Penelope Schofield⁴,²³,²⁴, David Bowtell⁴,⁵, Linda Mileshkin*³,⁴ and Richard W. Tothill *¹,⁴

1. Department of Clinical Pathology and Centre for Cancer Research, University of Melbourne, Melbourne, Australia
2. Department of Pathology, Peter MacCallum Cancer Centre, Melbourne, Australia
3. Department of Medical Oncology, Peter MacCallum Cancer Centre, Melbourne, Australia
4. Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia
5. Peter MacCallum Cancer Centre, Melbourne, Australia.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
6. The Westmead Institute for Medical Research, Sydney, Australia
7. Department of Gynaecological Oncology, Westmead Hospital, Sydney, Australia
8. The Daffodil Centre, The University of Sydney, a joint venture with Cancer Council NSW, Sydney, Australia.
9. Department of Medical Oncology, Crown Princess Mary Cancer Centre, Westmead Hospital, Sydney, Australia
10. Department of Medical Oncology, Flinders University and Flinders Medical Centre, Adelaide, Australia
11. Department of Medical Oncology, Barwon Health Cancer Services, Geelong, Australia
12. Department of Medical Oncology, SouthWest HealthCare, Warrnambool and Deakin University, Geelong, Australia
13. Medical Oncology, Cabrini Health, Melbourne, Australia
14. Border Medical Oncology, Albury Wodonga Regional Cancer Centre, Albury, Australia
15. Medical Oncology, Bendigo Health, Bendigo, Australia
16. Division of Medicine, Alan Walker Cancer Centre, Darwin, Australia
17. Department of Cardiothoracic Surgery, St Vincent's Hospital, Melbourne, Australia
18. Department of Radiation Oncology, Peter MacCallum Cancer Centre, Melbourne, Australia
19. Department of Computational Sciences, The Jackson Laboratory, Connecticut, United States
20. The St Vincent's Hospital Department of Medicine, University of Melbourne, Melbourne, Australia

21. Department of Medicine, Royal Melbourne Hospital, University of Melbourne, Melbourne, Australia

22. Cancer Diagnosis and Pathology Group, Kolling Institute of Medical Research and Sydney Medical School, University of Sydney, Sydney, Australia

23. Department of Psychology, and Iverson Health Innovation Research Institute, Swinburne University, Melbourne, Australia

24. Behavioural Sciences Unit, Health Services Research and Implementation Sciences, Peter MacCallum Cancer Centre, Melbourne, Australia

* These authors contributed equally to this work

Corresponding authors:

Richard Tothill, 305 Grattan St, Melbourne VIC 3000, rtothill@unimelb.edu.au

Linda Mileshkin, 305 Grattan St, Melbourne VIC 3000, Linda.Mileshkin@petermac.org

Competing interest statement: The authors declare no competing financial interests
Abstract

Cancer of unknown primary (CUP) constitutes a group of metastatic cancers in which standardized clinical investigations fail to identify a tissue of origin (TOO). Gene-expression profiling (GEP) has been used to resolve TOO, and DNA sequencing to identify potential targeted treatments; however, these methods have not been widely applied together in CUP patients. To assess the diagnostic utility of DNA and RNA tests for TOO classification, we applied GEP classification and/or gene-panel DNA sequencing in 215 CUP patients. Based on a retrospective review of pathology reports and clinical data, 77% of the cohort were deemed True-CUPs (T-CUP). Among the remaining cases, a latent primary diagnosis (10%) (LP-CUP) or TOO was highly suspected based on combined clinicopathological data (13%) (histology-resolved CUP, HR-CUP). High-medium confidence GEP classifications were made for 80% of LP/HR-CUPs, and these classifications were consistent with a pathologist-assigned diagnosis in 94% of cases, while only 56% of T-CUPs had high-medium confidence predictions. The frequency of somatic mutations in cancer genes was similar to 2,785 CUPs from AACR GENIE Project. DNA features, GEP classification, and oncovirus detection assisted making a TOO diagnosis in 37% of T-CUPs. Gene mutations and mutational signatures of diagnostic utility were found in 31% T-CUPs. GEP classification was useful in 13% of cases and viral detection in 4%. Among resolved T-CUPs, lung and biliary were the most frequently identified cancer types, while kidney cancer represented another minor subset. Multivariate survival analysis showed that unresolved T-CUPs had poorer overall survival than LP/HR-CUPs (Hazard ratio=1.9, 95% CI 1.1 – 3.2, p=0.016), while the risk of death was similar in genomically-resolved T-CUPs and LP/HR-CUPs. In conclusion, combining DNA and RNA profiling with clinicopathological data supported a putative TOO diagnosis in over a third of T-CUPs. DNA sequencing benefited T-CUP tumors with atypical transcriptional patterns that hindered reliable GEP classification.
87 **Keywords**

88 cancer of unknown primary; next-generation sequencing; cancer diagnostic; mutation profiling; targeted therapy; tissue of origin classification
Introduction

Cancer of unknown primary (CUP) represents 1-3% of all cancer diagnoses. It has a notoriously poor outcome, and in Australia, despite being the 14th most common diagnosis is the 6th most common cause of cancer related death. In the absence of a known tissue of origin (TOO), most CUP patients have historically been treated with empirical chemotherapy that is non-durable in most cases. Molecularly targeted treatments and immunotherapies may improve the survival outcome for some CUP patients, although drug access can still be challenging without confirming a TOO diagnosis. Furthermore, a CUP diagnosis brings a heavy psychosocial burden, as patients struggle with the absence of a specific diagnosis leading to feelings of high uncertainty. Therefore, a combination of improved diagnostic and treatment options for people with CUP is urgently needed.

The guidelines for a CUP diagnosis are currently based on standardized gender-appropriate histopathological and clinical investigations, as described by the European Society of Medical Oncology (ESMO). Genomic tests, including gene expression profiling (GEP) and DNA methylation analysis, have also been described for making a TOO diagnosis, and some of these tests are commercially available. These classifiers have an accuracy of 83 to 94% when tested on known primary and metastatic tumors and are superior to immunohistochemistry. The diagnostic utility of molecular classifiers has been validated when a latent primary can be found (LP-CUP), in which a
primary tumor becomes known in time or through alignment with IHC and other clinicopathological information. However, despite the potential diagnostic value of these molecular tests, their clinical utility is questioned. While retrospective and non-randomized studies applying site-directed therapies based on predicted TOO have shown survival benefits, two randomized clinical trials showed no improvements in patient outcomes. As such, the level of recommendation for these tests is low under current guidelines.

DNA mutational profiling has also been explored in CUP. The primary goal of these studies has been to identify actionable mutations to direct targeted therapies with clinically actionable mutations identified in 30-85% of CUP cases. However, mutational profiling also provides insight into the TOO, given that certain mutational processes and the prevalence of cancer driver mutations can be cancer type-dependent. The availability of large amounts of mutation data in public data repositories enables the interpretation of mutations and mutational signatures found in CUP tumors by a priori knowledge of the frequency and specificity across known cancer types. Not surprisingly, the integration of genomics with histology to resolve CUP diagnosis is making its way into clinical practice.

While GEP or mutation profiling of CUP has been described independently in previous studies, few studies have considered the diagnostic utility of these assays benchmarked against clinicopathological review. Here we describe the application of GEP and
targeted DNA sequencing in a CUP series recruited through a national study (Solving Unknown Primary Cancer: SUPER). Starting from a retrospective review of the available clinicopathological data, we firstly classified CUP cases based on a degree of certainty of a single site TOO diagnosis. We then assessed the potential additive diagnostic value of GEP and DNA sequencing in these patient groups. We also compared overall survival differences and the frequency of potential targeted treatments in the CUP groups based on whether genomics and/or clinicopathological review had assisted in resolving a TOO diagnosis. We used the observations from this study to support the use of genomics data in the diagnostic workup of CUP patients and to further characterize CUP disease characteristics.

Methods

Patient cohort

CUP patients were recruited to the SUPER study from 11 Australian sites with informed patient consent under an approved protocol of the Peter MacCallum Cancer Centre (PMCC) human research ethics committee (HREC protocol: 13/62).

Eligibility criteria for patient inclusion in the study were: 1. presenting with carcinoma of no confirmed primary site and who had a preliminary diagnostic workup, including, but not limited to; a detailed clinical assessment; a CT scan of the chest, abdomen, and pelvis; pathological review of tumor tissue; and, other gender-appropriate diagnostic tests; 2. were yet to commence treatment, or had commenced treatment no more than six months prior to recruitment; and 3. could read and write in English, and provide
written informed consent. Exclusion criteria included patients under 18 years of age, who had a poor ECOG performance status (ECOG>=3) or uncontrolled medical or psychological conditions that may have prevented completion of study requirements. Patients were identified and referred by their treating clinician at 11 participating hospitals (Supplementary Methods). Histopathology, clinical characteristics, diagnostic investigations, treatment history, and survival data were collected at baseline, six- and twelve-months post recruitment unless deceased or withdrawn earlier.

Latent primary status was recorded by the treating clinician if a primary tumor was detected with both imaging and histopathology during the study. For study purposes, CUP patients were classified by a medical oncologist (TS) into the ESMO defined clinicopathological subsets defined as favorable and unfavorable prognosis groups, as previously described.

CUP clinicopathological review

A retrospective review of all histopathology reports and clinical data was performed by a single pathologist (OP) (Supplementary Table 1-2). A histopathology review of slides was undertaken for a subset of cases from the PMCC (n=59), including additional IHC staining if necessary and tissue was available (Supplementary Tables 1-2). Further review of the clinical data was done by a medical oncologist (LM) to concur with the pathologist. Cases were assigned a likely diagnosis where possible, or alternatively, designated as a True-CUP (T-CUP) and then subclassified using a modified version of the Memorial Sloan Kettering Cancer Center OncoTree classification system for CUPs.
OncoTree classification included the following: undifferentiated malignant neoplasms (UDMN), poorly differentiated carcinoma (PDC), adenocarcinoma, not otherwise specified (ADNOS), neuroendocrine tumors, not otherwise specified (NETNOS), neuroendocrine carcinomas, not otherwise specified (NECNOS) and squamous cell carcinomas, not otherwise specified (SCCNOS). ADNOS were further subdivided based on cytokeratin 7 (CK7) and cytokeratin 20 (CK20) IHC staining, and in the case of CK7 negative and CK20 positive staining, caudal-type homeobox 2 (CDX2) was annotated. All ADNOS CK7-CK20+ tumors in the cohort had positive IHC staining of CDX2. SCCNOS were further subclassified to "SCC p16+" based on p16INK4A (p16) IHC staining positivity, as this increases the likelihood of high-risk human papillomavirus (HPV) infection and raises the possibility of oral, uterine, cervical, or anal mucosal origin. The diagnosis for all cases was re-assessed following the NGS findings.

**Gene expression profiling**

GEP was performed using a previously described microarray-based test (CUPGuide) or a custom NanoString nCounter assay (NanoString Technologies Inc., Seattle, WA, USA). A detailed description of nucleic acid extraction, the NanoString assay and TOO classifier is described in Supplementary Methods. Briefly, the NanoString panel included probe sets targeting 225 genes that were differentially expressed across 18 tumor classes, endogenous control genes, and viral transcripts encoding capsid proteins for HPV16 L1, HPV18 L1, and Merkel cell polyomavirus (VP2) (Supplementary Table 3). The NanoString classifier was trained on harmonized TCGA RNA-seq data previously
described 10, representing 8,454 samples consolidated into 18 tumor classes (Supplementary Table 4). Tumor classes SCC and Neuroendocrine represented tumors sharing the respective histological features but originating from multiple tissues in training. The RNA-seq/NanoString k-nearest neighbor cross-platform classifier was validated on an independent test set of 188 metastatic tumors profiled by NanoString (see Supplementary Tables 5-6 for further details on the prediction performance and overall accuracy/sensitivity/specificity). A probability score was generated for predictions and heuristic thresholds set for classification confidence level (unclassified <0.5, low ≥ 0.5 and ≤ 0.7, medium confidence >0.7 and < 0.9, high confidence ≥9 probability).

**Targeted DNA sequencing**

Targeted enrichment and DNA sequencing were performed on matched blood and tumor DNA, capturing coding regions and exon/intron splice sites of 386 cancer-related genes (listed in Supplementary Table 7) using previously described methods 30. Alignment and variant calling were performed using the ensemble variant caller bcbio-nextgen cancer somatic variant calling pipelines and R tools used for analysis. A detailed description of bioinformatics is described in Supplementary Methods.

**Reference mutation data and identification of putative diagnostic DNA features**

The AACR Project GENIE mutation data for 77,058 tumors (version 3.7.9) 23 was downloaded from the cBioPortal webpage (https://genie.cbiopower.org/) 26,27. The frequency of gene-specific mutations was assessed in tumors annotated as cancer of
unknown primary or other cancer types. Genes investigated were restricted to those included in the targeted capture panel in the current study (Supplementary Table 7). For assessment of cancer gene driver mutations, GENIE cancer classes with less than 50 samples per class with genes of interest were removed from consideration, except for assessing gene fusions, where cancer classes with less than 10 samples per cancer class were removed from analysis (Supplementary Table 8).

For identifying DNA features of potential diagnostic utility, the frequency of gene-wise driver mutations was calculated in 22 pre-defined cancer classes (Supplementary Table 8). Oncogenes and tumor suppressor genes (TSGs) were annotated using OncoKB. For assessing cancer class mutation frequency in the reference data, only truncating mutations in TSGs or hotspot mutations in both oncogenes and TSGs were used (annotated using the web portal http://www.cancerhotspots.org/) Oncogenic gene-fusions were restricted to those involving the genes FGFR2, FGFR3, ERG, FUS, TMPRSS2, ALK, RET, ROS1, NRG1, NTRK1, NTRK2, NTRK3, and EWSR1. Copy-number alterations were restricted to a curated set of frequently amplified cancer genes (Supplementary Table 7). Gene alteration frequencies was calculated within individual cancer classes (Supplementary Table 9). A Fisher’s exact test was then performed using the R package stats to identify genes statistically enriched for DNA alterations in individual cancer classes versus all other cancers. A post hoc adjusted test was performed using the Holm–Bonferroni method. Significance was defined as anything with an adjusted p-value <0.05 and an odds ratio (OR) greater than 1. These
significant cancers type diagnostic DNA alterations are summarized in Supplementary Tables 9-10.

Survival analysis

Overall survival (OS) was measured from the date of CUP diagnosis (histologically confirmed) to the date of death from any cause. Thirty-five patients had death recorded after completion of study follow-up. Patients without a recorded death were censored based on the date of the last follow-up (12 months). Survival analysis was performed using the R package `survival` (v3.1-12). Kaplan-Meier estimates of OS were presented along with log-rank tests for comparison between resolved and true CUP groups. A cox proportional-hazard model was used for multivariable analysis and adjusted for ECOG (0,1,2,3), age (>60 or <60), gender, and ESMO assigned favorable and unfavorable prognosis groups. Kaplan–Meier and forest plots were produced with the `survminer` package (v0.4.8.999).

Results

Study cohort and subclassification of CUPs

A total of 215 patients were recruited to the SUPER study, and a summary of baseline characteristics is shown in Table 1. Eighty-nine percent (191/215) received GEP, 93% (201/215) of patients had DNA sequencing, and 82% (177/215) received both assays (Table 1). A latent primary (LP-CUP) cancer was reported by the treating clinician during clinical follow-up in 10% (22/215) of cases based on histopathology, clinical
presentation, and/or cancer imaging. In another 13% (27/215) of cases, a likely TOO was assigned after a retrospective review of described morphology, clinical picture, and IHC staining; these cases were termed histology resolved CUP (HR-CUP). Notably, among the LP/HR-CUP cases, there was an enrichment of patients with a prior history of cancer (35%, 17/49), eight of whom likely had a recurrence of their previous disease (Supplementary Table 1).

Most of the patients (166/215, 77%) were deemed to be True-CUP (T-CUP) and had sufficient IHC workup according to current ESMO guidelines. Additional IHC stains may have been informative in a small subset (7/166, 4%) but could not be done owing to tissue availability. The T-CUPs were classified into histomorphological subtypes using a modified version of the Memorial Sloan Kettering Cancer Center OncoTree classification (see Methods) (Supplementary Table 2). The majority of T-CUPs were adenocarcinomas (51%) or poorly differentiated carcinomas (25%) with minor subsets of SCCs (13%), undifferentiated neoplasms (7%), and neuroendocrine neoplasms (2%). The most frequently assigned cancer types for LP/HR-CUPs or CUP OncoTree classifications are summarized in Supplementary Table 11.

**GEP classification confidence is lower for True-CUPs than known metastatic cancers**

We used two GEP methods for TOO classification of 191/215 CUP tumors, where sufficient RNA was available. A previously described 18-class microarray-based
classifier (CUPGuide) was used in 20 cases, while a novel NanoString classifier was used in the remaining cases (Supplementary Table 1-2). The NanoString classifier was validated using an independent cohort of 188 metastatic tumors of known origin (Supplementary Table 6), achieving an overall prediction accuracy of 82.9%, increasing to 91.5% considering only high-medium confidence classifications (n=154) (Figure 1A, Supplementary Table 5).

GEP TOO classification was possible for 45 LP/HR-CUPs (Figure 1B), of which 80% (36/45) had a high-medium confidence classification (Figure 1C). Three LP/HR-CUPs were considered to be outside the 18-class GEP TOO differential, including one rare ampullary tumor and two uterine tumors. Among high-medium confidence predictions within the GEP diagnostic differential, 94% (31/33) were concordant with their assigned diagnosis (Supplementary Table 1). Common high-medium confidence misclassifications observed among LP/HR-CUPs as well as the 188 known metastatic cancers included cholangiocarcinoma (1/1 and 5/11, respectively) and pancreatic adenocarcinomas (3/10 in the known metastatic group), illustrating classification difficulty among the pancreatobiliary group (Figure 1A, B).

GEP TOO classification was performed on 146 T-CUPs. High-medium confidence classifications were made for only 56% (82/146) of T-CUPs (Figure 1C). The most frequent high-medium confidence classifications included SCC (32%), liver (13%), colorectal (12%), breast (10%) and lung (8.5%) (Figure 1D). A reduced proportion of
high-medium confidence classifications for T-CUPs compared to known metastatic
 cancers suggested T-CUPs either have an atypical transcriptional profile or potentially
 are cancer types outside of the GEP classifier differential.

*The mutation profile of the SUPER CUP cohort is consistent with other CUP
cohorts*

DNA panel sequencing was performed for 201/215 CUP tumors and their matching
germline controls. We detected mutational features including single nucleotide variants
(SNVs), gene-fusions, copy-number amplifications (CNAs), SNV 96 trinucleotide
mutational signatures (COSMIC v2)\(^2\), tumor mutation burden (TMB), and off-target viral
DNA sequences (HPV, EBV) (Figure 2). Viral RNA transcripts detected by NanoString
also supported viral status for HPV-positive tumors.

At least one protein-coding mutation was found in 98.5% (198/201) of all CUPs, with a
median TMB of 4.4 mutations/Mb (range 0.5-149 mutations/Mb). The most frequently
mutated genes were *TP53* (55%), *LRP1B* (20%), *PIK3CA* (17%), *KMT2D* (15%), *KRAS*
(12%), *ARID1A* (11%) and *SMARCA4* (11%) (Figure 2). The variant allele frequency of
these mutations ranged between 16-40.5%, consistent with clonal cancer driver events
when tumor purity was considered. Additionally, 8% (n=16) of the cohort had dominant
Signature 4 (tobacco smoking), 5% (n=11) Signature 7 (ultra-violet light, UV) and 2%
(n=4) Signature 6 (microsatellite instability, MSI). HPV16 (DNA and RNA) was detected
in five cases and EBV (DNA only) in one case.
The gene-wise mutation frequency in the CUP cohort was also compared to 2,785 CUPs with panel sequencing in the AACR Project GENIE database (Figure 2). The mutation profile between the study CUP and GENIE CUP cohorts was highly similar with some minor differences, including a higher frequency of KRAS mutations (12% vs. 22%) among the GENIE CUPs and higher LRP1B (18% vs. 2%) mutations in the SUPER cohort; the latter explained by LRP1B not being included in MSK-IMPACT panel 33.

Actionable mutations were also investigated in reference to the CUPISCO trial criteria of actionability described previously 22. Eighty-six CUP patients (40%) were matched to either a targeted therapy or immunotherapy arm of CUPISCO (Figure 2) (Supplementary Table 12).

**Mutation profiling and GEP can augment histopathology review**

We next considered the diagnostic value of DNA and RNA features, including GEP classification in combination with the histopathological review. Here we considered driver gene mutations, gene fusions, mutational signatures, oncoviruses, and high-medium confidence GEP classifications only. To identify gene features with significant cancer type associations, we referenced the GENIE database (77,058 tumor samples) involving 22 solid cancer types. The potential diagnostic utility of a gene feature was determined by comparing one cancer type versus all others (Fisher exact test adjusted
p-value < 0.05 and odds-ratio (OR) >1) (Supplementary Table 10). A total of 171 genes were significantly enriched in one or more cancer types, and 90 genes were enriched in only one cancer type (Figure 3A, Supplementary Figure 1). Mutational signatures also provided cumulative evidence to support likely TOO, for example, a Signature 7 (UV) associated with skin cancer or a Signature 4 (tobacco) found in cancers of the airways, although potentially not excluding liver cancer.

We first assessed LP/HR-CUP cases that had been assigned a single TOO prior to considering the genomics data. Sixty-nine percent of cases (33/49) had one or more features (RNA and/or DNA) consistent with the TOO diagnosis. Both GEP classification and mutation profiling was informative in 20/33 cases, DNA features alone in 7/33, GEP alone in 5/33, and viral detection alone in 1/33 cases. The seven cases where GEP was not informative included two where GEP was not possible, four with low confidence classification, and one was a rare cancer part of the classifier differential. High-medium confidence GEP classifications were the most useful diagnostic feature (n=25) followed by gene mutations (n=24), mutational signatures (n=4), copy number amplifications (n=2), gene-fusions (n=1), and oncoviruses (n=1) (Figure 3B, Supplementary Figure 2).

Examples of LP/HR-CUP diagnoses supported by genomic data included 8/10 (80%) ovarian cases with high-confidence GEP classification as well as a TP53 mutation, the latter occurring in >96% of high-grade serous ovarian cancers. A dominant mutation Signature 7 (UV) (n=3) and either oncogenic mutations in NRAS (melanoma OR=20.7)
or truncating \textit{NF1} mutations (melanoma OR=5.2) was consistent with cutaneous melanoma. High-confidence GEP prediction of gastric cancer combined with \textit{ERBB2} amplification (concordant with positive HER2 IHC staining) (esophagus/stomach OR=3) and a frameshift deletion in \textit{CDH1} (esophagus/stomach OR=3) supported a diagnosis of gastric adenocarcinoma. High-confidence colorectal GEP prediction and \textit{APC} mutations (colorectal OR= 77.5) were present in two cases diagnosed as colorectal adenocarcinomas (Supplementary Figure 2, Supplementary Table 1).

A single HR-CUP (1097) had molecular features resulted in a change in classification from colorectal to T-CUP (ADNOS CK7-CK20+CDX2+). No mutations supported a colorectal origin (e.g., \textit{APC}, \textit{RAS/RAF} mutations), and a high-confidence GEP prediction of kidney with mutations in \textit{NF2} and \textit{SMARCA4} was identified. Confirmatory IHC staining (e.g., for PAX8) may have supported a kidney cancer diagnosis, but no tissue was available.

Molecular features supported a likely single TOO diagnosis consistent with clinicopathological features in 37% (61/166) of T-CUPs (Figure 3C). DNA features supported a diagnosis in 31% (51/166) of T-CUP cases, GEP TOO classification in 13% (21/166), and viral detection in 4% (6/166). Combined GEP classification and mutation profiling features were informative in 10% 16/166 cases (Figure 3D). Types of genomic feature supporting a diagnosis included driver gene mutations (n=36), mutational signatures (n=23), HPV16 (n=5) and EBV (n=1) viral nucleic acids, gene amplification (n=4), gene-fusions (n=3) and high-medium confidence GEP prediction (n=21) (Figure
DNA features could narrow the differential diagnosis in a further 11/166 T-CUPs (7%), although a TOO assignment of a single site could not be confidently made (Figure 3C). Considering the genomics data, the most frequently suspected cancer types among T-CUPs were lung (n=18, including a single pleomorphic carcinoma of the lung (LUPC)), biliary tract (n=8), breast (n=5), colorectal (n=5), HPV+ SCC (n=5) and kidney (n=4) (Figure 3E).

Lung-CUP was the single largest single group among T-CUPs. Dominant mutational Signature 4 (tobacco) was found in 14 cases. Driver gene mutations associated with non-small cell lung cancer (NSCLC) included \( \text{KEAP1} \) (Lung OR=9.8), \( \text{STK11} \) (Lung OR=14.1), \( \text{SMARCA4} \) (Lung OR=2.8), and \( \text{KRAS} \) (Lung OR=2.1) (Figure 3E, Supplementary Figure 1 and Supplementary Table 10). Notably, all but one putative Lung-CUP lacked TTF1 IHC staining. Among the Lung-CUPs where GEP was possible, a high-medium confidence classification of Lung was made in 5/16 cases. Two additional T-CUPs that were unresolved had high-medium lung classification, but mutation profiling was unsuccessful, and there was no other evidence to support a lung diagnosis. Three Lung-CUPs were CDX2-positive by IHC but had mutational features consistent with lung cancer, including a Signature 4 (tobacco) in all three cases. GEP was uninformative in these cases as one was classified as colorectal (0.9-confidence probability), and two were predicted SCC, a non-specific classification but potentially in keeping with lung-squamous cancer (Figure 3E, Supplementary Table 2). These CDX2+ Lung-CUPs were favored to be enteric-like lung adenocarcinomas.  

37
Eight ADNOS tumors were putatively diagnosed as intrahepatic cholangiocarcinomas supported by mutations in $BAP1$ (cholangiocarcinoma OR=7.2) and $IDH1$ (cholangiocarcinoma OR=32) (Figure 3E and Supplementary Table 10). Seven of these tumors presented with liver masses. These tumors lacked $KRAS$ mutations, making a pancreatic origin less likely, given that $KRAS$ mutations occur in approximately 90% of pancreatic adenocarcinomas (Supplementary Table 9-10) $^{38}$. Hotspot $IDH1$ (R132C) mutations have the highest frequency in cholangiocarcinoma (cholangiocarcinoma OR=33) but can be detected at a low frequency in melanoma (2%, melanoma OR=3.6). Similarly, $FGFR2$-fusions are frequently detected in intrahepatic cholangiocarcinoma (cholangiocarcinoma OR>100) $^{39-41}$, a feature that supported a cholangiocarcinoma diagnosis in two T-CUP cases (Supplementary Table 2 and 10).

In four T-CUPs subsequently assigned a kidney TOO, two cases expressed PAX8 by IHC, both confirmed by high PAX8 mRNA expression (z-score > 2), and an additional case did not have PAX8 IHC performed but had high PAX8 mRNA expression. The fourth case did not have PAX8 staining or high PAX8 mRNA expression. Only two cases were classified as kidney cancer by GEP (Figure 3E). Driver mutations significantly associated with kidney cancer, and detected among kidney-CUPs, included $BAP1$ (kidney OR=7.6) and $NF2$ (kidney OR=7.2). Another case had a truncating $FH$ mutation consistent with FH-deficient kidney cancer (kidney OR=8.1) (Supplementary Figures 1). Notably, no $VHL$ mutations were detected, representing the most common driver gene in renal cell carcinoma (RCC) (42.5% RCC in GENIE, OR=763).
assigned kidney case was confirmed as a recurrence of late-onset adult Wilm's tumor by detecting a somatic \textit{FGFR1} (p.K656D) mutation in both the original primary tumor and a recurrent tumor that presented over twenty years after first diagnosis (Supplementary Table 2 and 10).

\textbf{Unresolved CUPs have a reduced survival outcome compared to resolved CUPs}

Of the 215 patients, 177 patients had recorded survival data. Among all cases recruited to the SUPER study, the majority were classified as unfavorable according to the ESMO guidelines (137/159= 86% of T-CUP, and 39/48=81% of LP/HR-CUP, where data was available) \(^5\) (Figure 4A). Overall, 85% of the cohort was unfavorable CUP and, as expected, these had significantly poorer overall survival (OS)- by log-rank test (\(p < 0.001\)), with a median OS of 11 months compared to a median OS of 15 months for favorable CUP. Additionally, 46% of unfavorable CUP achieved 12 months of survival compared with 92% of favorable CUP (Figure 4B). Furthermore, log-rank testing showed longer survival in LP/HR-CUPs, and T-CUPs assigned a single TOO after interpretation of the genomics data, compared with diagnostically unresolved T-CUPs (\(p=0.04\), Figure 4C). Multivariate Cox-regression analysis, including gender, age, ECOG performance, and ESMO prognosis groups (Figure 4D), showed that T-CUPs had worse survival outcomes compared to LP/HR-CUPs (HR=1.89, 95\% CI 1.13-3.2, \(p=0.016\)) while there was no significant difference between LP/HR-CUPs and resolved T-CUPs (HR=1.18, 95\% CI 0.65-2.1, \(p=0.595\)).
Discussion

The reported incidence of CUP has decreased in recent years. This is most likely due to increased awareness and standardization of investigations, including access to more specific IHC stains, improved diagnostic imaging, and the use of molecular profiling. Consistent with other large retrospective CUP studies, we found that approximately one-quarter of CUP may be assigned a likely TOO based on a centralized histopathology review. This is similar to the recent experience of the international CUP clinical trial CUPISCO, where ~20% of patients had a single primary site diagnosis supported by available evidence or was TOO was strongly suspected. In the current study, we have shown that incorporating DNA and RNA tests may help to resolve more than a third of T-CUP cases not otherwise resolved using conventional testing. Importantly, despite GEP being the most commonly explored molecular test for resolving CUP, we found DNA sequencing was of potentially greater diagnostic value among CUP tumors due to many CUPS having an atypical transcriptional profile.

Despite positive evidence that GEP can be diagnostically informative for CUP, cases with poorly differentiated histopathology and atypical disease presentation remain challenging. This is perhaps not surprising given that GEP classification relies upon the expression of cellular differentiation markers that are often lost or equivocal in CUP tumors. Our observation that fewer CUP pass a high-medium confidence GEP classification compared to known metastatic cancers, hence have a reduced classifier performance, is supported by other studies. For instance, the CancerTYPE ID
(Biotheranostics) 92-gene test, which had extensive multisite validation, showed an overall accuracy of 85% for known cancer metastases. But in CUP, the concordance of GEP with IHC and clinicopathological evidence was lower at 75% for LP-CUP and 70% compared to the clinical picture only 7. Additionally, rare malignancies are infrequently included in the training set of TOO classifiers 6, which poses a limitation for rare entities among a cohort of CUP. More specifically, subsets of lung and biliary cancers can be problematic for GEP classifiers to resolve. For example, among lung CUPs, GEP classification was found to be concordant with a latent primary tumor in only 50% of cases 13. GEP classification accuracy can also be low for cholangiocarcinomas, as the transcriptional profile is similar to pancreatic and upper gastrointestinal neoplasms 10. Notably, some previously described GEP and DNA methylation tests have also excluded cholangiocarcinoma in their models or combined them into a pancreaticobiliary class 9,48. We found that DNA sequencing may be particularly useful in pancreaticobiliary CUPs given that gene mutations characteristic of cholangiocarcinomas can have both diagnostic and occasionally therapeutic significance, including alterations in IDH1, FGFR2, and BAP1 39-41.

Integrated molecular profiling of CUP can help identify rare disease subtypes and our data supports their being some recurrent CUP entities. We identified examples of pulmonary enteric adenocarcinomas that lack TTF1 expression but can express the gastrointestinal marker CDX2 37. GEP or IHC alone could not resolve such cases, given their atypical profile; however, they still retained DNA features highly suggestive of
NSCLC, including a tobacco mutational signature and somatic mutations in *KRAS*, *STK11*, and *SMARCA4*. *SMARCA4*-deficient lung cancers are known to lack TTF1 expression \(^{49}\) and are likely to be frequent in the CUP population \(^{19,20}\). *SMARCA4*-deficient lung cancer models similarly lose expression of lung differentiation markers and have a pro-metastatic behavior consistent with aggressive clinical course and poor survival outcomes in patients \(^{50}\). Kidney cancers are another emerging CUP entity representing ~4-6% of CUPs \(^{44,51,52}\). Four CUPs in the SUPER cohort were consistent with a primary kidney tumor. Interestingly, we found somatic mutations in *NF2*, *FH*, and *BAP1* in some RCC cases. Somatic *NF2* mutations are characteristic of advanced papillary renal cell tumors and those with biphasic hyalinizing psammomatous features \(^{53}\). Papillary carcinomas have also been found to be enriched among other Kidney-CUPs \(^{51,52}\) with a mutation profile similar to the Kidney-CUPs described in the SUPER cohort \(^{54}\). Identifying CUP entities and recurrent therapeutic targets in these groups may help guide future CUP clinical trials. For instance, while empirical chemotherapy is ineffective in RCC, targeted therapies and ICIs are likely more efficacious \(^{51,55}\). Furthermore, detection of *NF2* mutations among RCC and mesothelioma could direct targeted treatment of the Hippo pathway using inhibitors of TEAD auto-palmitoylation \(^{56}\), which have now entered clinical trials (NCT04665206).

CUP can be classified into favorable and unfavorable prognosis groups. Given that ~85% of CUPs belong to the unfavorable group, research and clinical trials directed at identifying other treatments for these patients or therapeutic responsive subsets should
be sought. Approximately 40% of T-CUPs in the SUPER cohort had potentially actionable mutations and genomic alterations, broadly consistent with other CUP profiling studies. As curated findings from genomic tests were returned to clinicians during the study, it is possible that the genomic data allowed site-specific treatments and access to targeted treatments, potentially explaining the improved survival outcome among the resolved compared to the unresolved T-CUP group. Alternatively, genomically resolved CUPs may be enriched for cancer types where targeted or immunotherapy treatments are available and effective.

In conclusion, we have shown that DNA and RNA tests can be incorporated into a pathology assessment to improve cancer type diagnosis, identify CUP subtypes, and direct treatments. Rather than replacing traditional histopathological analysis, molecular testing can augment conventional testing to either confirm a suspicion of primary tissue of origin or provide robust diagnostic leads that are not otherwise evident using other modalities. In practice, in cases where tissue is limited, it may be more informative to prioritize genomic testing to guide additional investigations before consuming tissue on extended IHC panels. With steady improvements in technology and reduction in the sequencing costs, more comprehensive whole-genome and transcriptome analysis will likely increase the sensitivity to detect features such as structural variants and mutational signatures that are not reliably detected by panel sequencing. Recent studies have also utilized machine-learning for classification based on genome-wide patterns of somatic mutations that are not otherwise interpretable by human curation.
and combined DNA and RNA features to improve the TOO prediction accuracy further (60-91% and 94%, respectively). Although molecular testing is not currently recommended in most CUP guidelines, careful prospective evaluation of these tests may justify their cost, helping to resolve cancer type in a significant fraction of CUP cases as well as direct treatments that will ultimately translate to better outcomes for patients.

**Figure legends**

**Figure 1**: Gene expression profiling (GEP) tissue of origin classification of known metastatic cancers and SUPER cancer of unknown primary (CUP) tumors. A) NanoString GEP classifier was tested on 188 known origin metastatic tumors with confusion matrix showing concordance of tissue of origin prediction and known cancer type. B) GEP classifier tested on latent primary/histology resolved (LP/HR)-CUPs showing concordance between the likely tissue of origin and the predicted cancer type. LP/HR-CUPs representing cancer types that were not represented in the classifier model were removed from analysis C) Fraction of cases within confidence probability score grouping contrasting classification of true-CUPs (T-CUP) and LP/HR-CUPs combined with known metastatic tumors (unclassified <0.5, low ≥ 0.5 and ≤ 0.7 medium confidence ≥ 0.8 and ≤ 0.9, high confidence = 1). D) GEP cancer class predictions of all T-CUPs with high-medium confidence predictions.
**Figure 2:** DNA mutational profiling of the SUPER cancer of unknown primary (CUP) cohort. Oncoplot shows somatic mutations in CUPs in descending order of frequency. Genes and mutational features colored in red are actionable and targeted in the CUPISCO trial. The proportion of mutations in the SUPER CUPs was compared with the AACR GENIE CUP cohort (right-hand bar plot). The left-hand plot of the variant allele frequency (VAF) distribution per gene. The top bar plot shows the number of coding mutations per sample. Annotations include OncoTree class for true-CUPs (T-CUP) or the assigned tumor class for latent primary/histology resolved (LP/HR)-CUPs, detection of COSMIC mutational signatures (V2): smoking signature, ultra-violet (UV) signature, DNA mismatch repair signature, oncoviruses: Human papillomavirus 16 (HPV16) and Epstein-Barr virus (EBV), and tumor mutation burden (TMB) status: High >10 mutations/Mb or Low <10 mutations/Mb.

**Figure 3:** Identification of diagnostically useful DNA features using AACR Project GENIE mutation data and summary of evidence used to support tissue of origin (TOO) diagnosis among SUPER cancer of unknown primary (CUP) cases. A) Volcano plot showing significantly mutated features by cancer types using GENIE pan-caner cohort. All genes are plotted for all cancers using a Fisher's exact test to calculate adjusted p-values and odds ratio (OR) (significance threshold adjusted p-value < 0.05 and OR >1). B) Proportion and number of cases where each genomic feature supported a putative TOO for true-CUPs (T-CUP) and latent primary/histology resolved (LP/HR)-CUPs. Genomic features included single nucleotide variants (SNV), Gene expression profiling
(GEP), mutational signatures, oncoviral sequences, and copy number amplifications (CNA). C) OncoTree cancer classification of T-CUPs before and after genomic analysis. CUPs were either resolved to a putative TOO, had reduced occult diagnosis, or remained T-CUP. D) Proportion of cases where DNA and/or GEP classification supported TOO diagnosis among T-CUPs. E) Detailed summary of supportive genomic features and IHC staining used to assign a putative TOO for all genomically resolved T-CUPs. NGS= Next generation sequencing.

**Figure 4:** A comparison of overall survival outcome among SUPER CUP patients based on clinical and diagnostic groups A) Proportion and number of true-CUPs (T-CUP) and latent primary/histology resolved (LP/HR)-CUPs classified by European Society of Medical Oncology (ESMO) favorable and unfavorable prognosis groups. B) Kaplan-Meier curve of overall survival (OS) in unfavorable and favorable CUP types. Censored at 12 months and significance using log-rank p-values. C-D) Log-rank test and multivariable Cox-proportional hazard model of SUPER CUP clinical and diagnostic group. Data censored at 12 months and significance using log-rank p values for Kaplan-Meier analysis. Cox-proportional hazard model including covariates of ECOG, age (>60 or <60), ESMO favorable and unfavorable outcome categories, and patient sex. GR = genomically resolved T-CUPs, U= unfavorable type, F= favorable type, HR= hazard ratio, CI= confidence interval.
## Table 1. Study cohort characteristics

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Count</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total cohort</td>
<td>215</td>
<td>100%</td>
</tr>
<tr>
<td>Age mean (range)</td>
<td>61 (20-86)</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>89</td>
<td>41%</td>
</tr>
<tr>
<td>Female</td>
<td>107</td>
<td>49%</td>
</tr>
<tr>
<td>Unrecorded</td>
<td>19</td>
<td>9%</td>
</tr>
<tr>
<td>ECOG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>67</td>
<td>31%</td>
</tr>
<tr>
<td>1</td>
<td>101</td>
<td>47%</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>11%</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>Not determined</td>
<td>22</td>
<td>10%</td>
</tr>
<tr>
<td>Previous cancer diagnosis</td>
<td>58</td>
<td>27%</td>
</tr>
<tr>
<td>ESMO Outcome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favourable</td>
<td>31</td>
<td>14%</td>
</tr>
<tr>
<td>Unfavourable</td>
<td>176</td>
<td>82%</td>
</tr>
<tr>
<td>Not determined</td>
<td>8</td>
<td>4%</td>
</tr>
<tr>
<td>Gene Expression Profile</td>
<td>191</td>
<td>89%</td>
</tr>
<tr>
<td>NanoString</td>
<td>172</td>
<td>80%</td>
</tr>
<tr>
<td>CUPguide</td>
<td>19</td>
<td>9%</td>
</tr>
<tr>
<td>DNA sequencing</td>
<td>201</td>
<td>93%</td>
</tr>
<tr>
<td>Both molecular tests*</td>
<td>177</td>
<td>82%</td>
</tr>
</tbody>
</table>

*DNA sequencing and GEP performed successfully
Reference list


2. Australian Institute of Health and Welfare (AIHW) 2018 Cancer Data in Australia; Australian Cancer Incidence and Mortality (ACIM) books: cancer of unknown primary site Canberra: AIHW.


**Supplementary Information**

**Supplementary Methods**

**Supplementary Figure 1**: Diagnostically useful mutated features per cancer type using GENIE pan-cancer cohort.

**Supplementary Figure 2**: Supportive genomic features and IHC staining for latent primary and histology resolved CUPs.

**Supplementary Tables 1-12**
**Acknowledgments**

We acknowledge Cameron Patrick of the Melbourne Statistical Consulting Platform for providing statistical support. We wish to thank Jillian Hung and Niklyn Nevins, SUPER study coordinators at Westmead and Blacktown, Lisa Kay at Nepean, Karin Lyon (ethics and governance), and acknowledge the contributions of the Nepean Cancer Biobank and the Westmead GynBiobank for facilitating the study. The authors would like to acknowledge the American Association for Cancer Research and its financial and material support in the development of the AACR Project GENIE registry, as well as members of the consortium for their commitment to data sharing. Interpretations are the responsibility of study authors. We wish to acknowledge the patients who have contributed to this study and the CUP consumer steering committee, Cindy Bryant (chair), Kym Sheehan, Christine Bradford, Clare Brophy, Dale Witton, and Frank Stoss.

**Ethics Approval / Consent to Participate**

All participating patients provided informed consent to partake in the study. This was approved by the Peter MacCallum Cancer Centre (PMCC) human research ethics committee (HREC protocol: 13/62).

**Author Contribution Statement**

RT and LM conceived the study. AP performed the analysis, drafted the figures and tables. OWJP undertook the histopathology review and data analysis. TS and LM reviewed the clinical data. CW, KF, SW collected the data for the study. ADF, NW, CSK,
BG, CS, MS, IMC, GR, MW, NK screened patients for eligibility. DE and NT developed the NanoString classifier, and JG and SB validated the classifier. GW, SW, AB, AJG, BJS, DB, SF, RJH contributed samples for training the classifier. AF, HX and SF performed the mutation profiling. AP performed the bioinformatic analysis. AP and RT co-wrote the manuscript. All authors critically reviewed the manuscript. PS, DB, LM and RT are the principal investigators and obtained research funding to support the study.

**Funding Statement**

The study was supported by funding from Cancer Australia (APP1048193, APP1082604) and the Victorian Cancer Agency (TRP13062). RWT was supported by funding from the Victorian Cancer Agency (TP828750). The Westmead, Blacktown and Nepean study sites were supported by the Cancer Institute NSW 11/TRC/1-06, 15/TRC/1-01 and 15/RIG/1-16.

**Data Availability Statement**

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.
Figure 1

(a) GEP Classification

(b) Known metastatic cancer

(c) Proportion of Known tumours vs. T-CUP

(d) GEP Prediction
Figure 2

All rights reserved. No reuse allowed without permission.
Resolved T-CUP (putative diagnosis)

Ovary/Fallopian tube (1)

Skin Adnexal (1)

Melanoma (2)

Sarcoma (1)

Prostate (1)

Gastric (3)

Kidney (4)

Breast (5)

Lung (17)

Liver (1)

Uterus (1)

Urothelium (1)

Skin Adnexal (1)

SCC cutaneous (4)

Sarcoma (1)

Prostate (1)

Ovary/Fallopian tube (1)

Melanoma (2)

Lung (17)

Liver (1)

Kidney (4)

Gastric (3)

Colorectal (5)

Cholangio (6)

Breast (5)
Figure 4

medRxiv preprint doi: https://doi.org/10.1101/2022.06.24.22276729; this version posted June 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.