

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

3
4 **Authors**

5 Atara Posner¹, Owen W.J. Prall², Tharani Sivakumaran^{3,4}, Dariush
6 Etemadamoghadam⁵, Niko Thio⁵, Andrew Pattison¹, Shiva Balachander¹, Krista
7 Fisher⁵, Samantha Webb⁵, Colin Wood⁵, Anna DeFazio^{6,7,8}, Nicholas Wilcken⁹, Bo
8 Gao⁹, Christos S Karapetis¹⁰, Madhu Singh¹¹, Ian M Collins¹², Gary Richardson¹³,
9 Christopher Steer¹⁴, Mark Warren¹⁵, Narayan Karanth¹⁶, Gavin Wright¹⁷, Scott
10 Williams¹⁸, Joshy George¹⁹, Rodney J Hicks²⁰, Alex Boussioutas²¹, Anthony J Gill²²,
11 Benjamin J. Solomon³, Huiling Xu², Andrew Fellowes², Stephen B Fox^{2,4}, Penelope
12 Schofield^{4,23,24}, David Bowtell^{4,5}, Linda Miles^{3,4} and Richard W. Tothill ^{*1,4}

13

14

15 1. Department of Clinical Pathology and Centre for Cancer Research, University of
16 Melbourne, Melbourne, Australia

17 2. Department of Pathology, Peter MacCallum Cancer Centre, Melbourne, Australia

18 3. Department of Medical Oncology, Peter MacCallum Cancer Centre, Melbourne,
19 Australia

20 4. Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne,
21 Australia

22 5. Peter MacCallum Cancer Centre, Melbourne, Australia.

- 23 6. The Westmead Institute for Medical Research, Sydney, Australia
- 24 7. Department of Gynaecological Oncology, Westmead Hospital, Sydney, Australia
- 25 8. The Daffodil Centre, The University of Sydney, a joint venture with Cancer Council
26 NSW. Sydney, Australia.
- 27 9. Department of Medical Oncology, Crown Princess Mary Cancer Centre, Westmead
28 Hospital, Sydney, Australia
- 29 10. Department of Medical Oncology, Flinders University and Flinders Medical Centre,
30 Adelaide, Australia
- 31 11. Department of Medical Oncology, Barwon Health Cancer Services, Geelong,
32 Australia
- 33 12. Department of Medical Oncology, SouthWest HealthCare, Warrnambool and Deakin
34 University, Geelong, Australia
- 35 13. Medical Oncology, Cabrini Health, Melbourne, Australia
- 36 14. Border Medical Oncology, Albury Wodonga Regional Cancer Centre, Albury,
37 Australia
- 38 15. Medical Oncology, Bendigo Health, Bendigo, Australia
- 39 16. Division of Medicine, Alan Walker Cancer Centre, Darwin, Australia
- 40 17. Department of Cardiothoracic Surgery, St Vincent's Hospital, Melbourne, Australia
- 41 18. Department of Radiation Oncology, Peter MacCallum Cancer Centre, Melbourne,
42 Australia
- 43 19. Department of Computational Sciences, The Jackson Laboratory, Connecticut,
44 United States

45 20. The St Vincent's Hospital Department of Medicine, University of Melbourne,
46 Melbourne, Australia

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

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

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

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

55

56 * These authors contributed equally to this work

57 Corresponding authors:

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

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

60

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

62

63 **Abstract**

64 Cancer of unknown primary (CUP) constitutes a group of metastatic cancers in which
65 standardized clinical investigations fail to identify a tissue of origin (TOO). Gene-expression
66 profiling (GEP) has been used to resolve TOO, and DNA sequencing to identify potential
67 targeted treatments; however, these methods have not been widely applied together in CUP
68 patients. To assess the diagnostic utility of DNA and RNA tests for TOO classification, we
69 applied GEP classification and/or gene-panel DNA sequencing in 215 CUP patients. Based on a
70 retrospective review of pathology reports and clinical data, 77% of the cohort were deemed
71 True-CUPs (T-CUP). Among the remaining cases, a latent primary diagnosis (10%) (LP-CUP)
72 or TOO was highly suspected based on combined clinicopathological data (13%) (histology-
73 resolved CUP, HR-CUP). High-medium confidence GEP classifications were made for 80% of
74 LP/HR-CUPs, and these classifications were consistent with a pathologist-assigned diagnosis in
75 94% of cases, while only 56% of T-CUPs had high-medium confidence predictions. The
76 frequency of somatic mutations in cancer genes was similar to 2,785 CUPs from AACR GENIE
77 Project. DNA features, GEP classification, and oncovirus detection assisted making a TOO
78 diagnosis in 37% of T-CUPs. Gene mutations and mutational signatures of diagnostic utility
79 were found in 31% T-CUPs. GEP classification was useful in 13% of cases and viral detection in
80 4%. Among resolved T-CUPs, lung and biliary were the most frequently identified cancer types,
81 while kidney cancer represented another minor subset. Multivariate survival analysis showed
82 that unresolved T-CUPs had poorer overall survival than LP/HR-CUPs (Hazard ratio=1.9, 95%
83 CI 1.1 – 3.2, p=0.016), while the risk of death was similar in genomically-resolved T-CUPs and
84 LP/HR-CUPs. In conclusion, combining DNA and RNA profiling with clinicopathological data
85 supported a putative TOO diagnosis in over a third of T-CUPs. DNA sequencing benefited T-
86 CUP tumors with atypical transcriptional patterns that hindered reliable GEP classification.

87 **Keywords**

88 cancer of unknown primary; next-generation sequencing; cancer diagnostic; mutation

89 profiling; targeted therapy; tissue of origin classification

90

91 **Introduction**

92 Cancer of unknown primary (CUP) represents 1-3% of all cancer diagnoses ¹. It has a
93 notoriously poor outcome, and in Australia, despite being the 14th most common
94 diagnosis is the 6th most common cause of cancer related death ². In the absence of a
95 known tissue of origin (TOO), most CUP patients have historically been treated with
96 empirical chemotherapy that is non-durable in most cases ³. Molecularly targeted
97 treatments and immunotherapies may improve the survival outcome for some CUP
98 patients, although drug access can still be challenging without confirming a TOO
99 diagnosis. Furthermore, a CUP diagnosis brings a heavy psychosocial burden, as
100 patients struggle with the absence of a specific diagnosis leading to feelings of high
101 uncertainty ⁴. Therefore, a combination of improved diagnostic and treatment options for
102 people with CUP is urgently needed.

103
104 The guidelines for a CUP diagnosis are currently based on standardized gender-
105 appropriate histopathological and clinical investigations, as described by the European
106 Society of Medical Oncology (ESMO) ⁵. Genomic tests, including gene expression
107 profiling (GEP) and DNA methylation analysis, have also been described for making a
108 TOO diagnosis, and some of these tests are commercially available. These classifiers
109 have an accuracy of 83 to 94% when tested on known primary and metastatic tumors ⁶⁻
110 ¹⁰ and are superior to immunohistochemistry ^{11,12}. The diagnostic utility of molecular
111 classifiers has been validated when a latent primary can be found (LP-CUP), in which a

112 primary tumor becomes known in time or through alignment with IHC and other
113 clinicopathological information ^{6,9,11-13}. However, despite the potential diagnostic value
114 of these molecular tests, their clinical utility is questioned. While retrospective and non-
115 randomized studies applying site-directed therapies based on predicted TOO have
116 shown survival benefits ^{14,15}, two randomized clinical trials showed no improvements in
117 patient outcomes ^{16,17}. As such, the level of recommendation for these tests is low under
118 current guidelines ⁵.

119
120 DNA mutational profiling has also been explored in CUP ¹⁸⁻²². The primary goal of these
121 studies has been to identify actionable mutations to direct targeted therapies with
122 clinically actionable mutations identified in 30-85% of CUP cases ¹⁸⁻²². However,
123 mutational profiling also provides insight into the TOO, given that certain mutational
124 processes and the prevalence of cancer driver mutations can be cancer type-dependent
125 ¹⁸. The availability of large amounts of mutation data in public data repositories enables
126 the interpretation of mutations and mutational signatures found in CUP tumors by a
127 priori knowledge of the frequency and specificity across known cancer types ²³⁻²⁷. Not
128 surprisingly, the integration of genomics with histology to resolve CUP diagnosis is
129 making its way into clinical practice ²⁸.

130
131 While GEP or mutation profiling of CUP has been described independently in previous
132 studies, few studies have considered the diagnostic utility of these assays benchmarked
133 against clinicopathological review. Here we describe the application of GEP and

134 targeted DNA sequencing in a CUP series recruited through a national study (Solving
135 Unknown Primary Cancer: SUPER). Starting from a retrospective review of the
136 available clinicopathological data, we firstly classified CUP cases based on a degree of
137 certainty of a single site TOO diagnosis. We then assessed the potential additive
138 diagnostic value of GEP and DNA sequencing in these patient groups. We also
139 compared overall survival differences and the frequency of potential targeted treatments
140 in the CUP groups based on whether genomics and/or clinicopathological review had
141 assisted in resolving a TOO diagnosis. We used the observations from this study to
142 support the use of genomics data in the diagnostic workup of CUP patients and to
143 further characterize CUP disease characteristics.

144

145 **Methods**

146 ***Patient cohort***

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

150 Eligibility criteria for patient inclusion in the study were: 1. presenting with carcinoma of
151 no confirmed primary site and who had a preliminary diagnostic workup, including, but
152 not limited to; a detailed clinical assessment; a CT scan of the chest, abdomen, and
153 pelvis; pathological review of tumor tissue; and, other gender-appropriate diagnostic
154 tests; 2. were yet to commence treatment, or had commenced treatment no more than
155 six months prior to recruitment; and 3. could read and write in English, and provide

156 written informed consent. Exclusion criteria included patients under 18 years of age,
157 who had a poor ECOG performance status (ECOG \geq 3) or uncontrolled medical or
158 psychological conditions that may have prevented completion of study requirements.
159 Patients were identified and referred by their treating clinician at 11 participating
160 hospitals (Supplementary Methods). Histopathology, clinical characteristics, diagnostic
161 investigations, treatment history, and survival data were collected at baseline, six- and
162 twelve-months post recruitment unless deceased or withdrawn earlier.

163 Latent primary status was recorded by the treating clinician if a primary tumor was
164 detected with both imaging and histopathology during the study. For study purposes,
165 CUP patients were classified by a medical oncologist (TS) into the ESMO
166 defined clinicopathological subsets defined as favorable and unfavorable prognosis
167 groups, as previously described ⁵.

168

169 ***CUP clinicopathological review***

170 A retrospective review of all histopathology reports and clinical data was performed by a
171 single pathologist (OP) (Supplementary Table 1-2). A histopathology review of slides
172 was undertaken for a subset of cases from the PMCC (n=59), including additional IHC
173 staining if necessary and tissue was available (Supplementary Tables 1-2). Further
174 review of the clinical data was done by a medical oncologist (LM) to concur with the
175 pathologist. Cases were assigned a likely diagnosis where possible, or alternatively,
176 designated as a True-CUP (T-CUP) and then subclassified using a modified version of
177 the Memorial Sloan Kettering Cancer Center OncoTree classification system for CUPs

178 ²⁹. OncoTree classification included the following: undifferentiated malignant neoplasms
179 (UDMN), poorly differentiated carcinoma (PDC), adenocarcinoma, not otherwise
180 specified (ADNOS), neuroendocrine tumors, not otherwise specified (NETNOS),
181 neuroendocrine carcinomas, not otherwise specified (NECNOS) and squamous cell
182 carcinomas, not otherwise specified (SCCNOS). ADNOS were further subdivided based
183 on cytokeratin 7 (CK7) and cytokeratin 20 (CK20) IHC staining, and in the case of CK7
184 negative and CK20 positive staining, caudal-type homeobox 2 (CDX2) was annotated.
185 All ADNOS CK7-CK20+ tumors in the cohort had positive IHC staining of CDX2.
186 SCCNOS were further subclassified to "SCC p16+" based on p16INK4A (p16) IHC
187 staining positivity, as this increases the likelihood of high-risk human papillomavirus
188 (HPV) infection and raises the possibility of oral, uterine, cervical, or anal mucosal
189 origin. The diagnosis for all cases was re-assessed following the NGS findings.

190

191 ***Gene expression profiling***

192 GEP was performed using a previously described microarray-based test (CUPGuide) ¹⁸
193 or a custom NanoString nCounter assay (NanoString Technologies Inc., Seattle, WA,
194 USA). A detailed description of nucleic acid extraction, the NanoString assay and TOO
195 classifier is described in Supplementary Methods. Briefly, the NanoString panel included
196 probe sets targeting 225 genes that were differentially expressed across 18 tumor
197 classes, endogenous control genes, and viral transcripts encoding capsid proteins for
198 HPV16 L1, HPV18 L1, and Merkel cell polyomavirus (VP2) (Supplementary Table 3).
199 The NanoString classifier was trained on harmonized TCGA RNA-seq data previously

200 described ¹⁰, representing 8,454 samples consolidated into 18 tumor classes
201 (Supplementary Table 4). Tumor classes SCC and Neuroendocrine represented tumors
202 sharing the respective histological features but originating from multiple tissues in
203 training. The RNA-seq/NanoString k-nearest neighbor cross-platform classifier was
204 validated on an independent test set of 188 metastatic tumors profiled by NanoString
205 (see Supplementary Tables 5-6 for further details on the prediction performance and
206 overall accuracy/sensitivity/specificity). A probability score was generated for predictions
207 and heuristic thresholds set for classification confidence level (unclassified <0.5, low ≥
208 0.5 and ≤ 0.7, medium confidence >0.7 and < 0.9, high confidence ≥9 probability).

209

210 ***Targeted DNA sequencing***

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

217

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

219 The AACR Project GENIE mutation data for 77,058 tumors (version 3.7.9) ²³ was
220 downloaded from the cBioPortal webpage (<https://genie.cbioportal.org/>) ^{26,27}. The
221 frequency of gene-specific mutations was assessed in tumors annotated as cancer of

222 unknown primary or other cancer types. Genes investigated were restricted to those
223 included in the targeted capture panel in the current study (Supplementary Table 7). For
224 assessment of cancer gene driver mutations, GENIE cancer classes with less than 50
225 samples per class with genes of interest were removed from consideration, except for
226 assessing gene fusions, where cancer classes with less than 10 samples per cancer
227 class were removed from analysis (Supplementary Table 8).

228 For identifying DNA features of potential diagnostic utility, the frequency of gene-wise
229 driver mutations was calculated in 22 pre-defined cancer classes (Supplementary Table
230 8). Oncogenes and tumor suppressor genes (TSGs) were annotated using OncoKB ²⁴.
231 For assessing cancer class mutation frequency in the reference data, only truncating
232 mutations in TSGs or hotspot mutations in both oncogenes and TSGs were used
233 (annotated using the web portal <http://www.cancerhotspots.org/>) ³¹. Oncogenic gene-
234 fusions were restricted to those involving the genes *FGFR2*, *FGFR3*, *ERG*,
235 *FUS*, *TMPRSS2*, *ALK*, *RET*, *ROS1*, *NRG1*, *NTRK1*, *NTRK2*, *NTRK3*, and *EWSR1*.
236 Copy-number alterations were restricted to a curated set of frequently amplified cancer
237 genes (Supplementary Table 7). Gene alteration frequencies was calculated within
238 individual cancer classes (Supplementary Table 9). A Fisher's exact test was then
239 performed using the R package *stats* to identify genes statistically enriched for DNA
240 alterations in individual cancer classes versus all other cancers. A post hoc adjusted
241 test was performed using the Holm–Bonferroni method. Significance was defined as
242 anything with an adjusted p-value <0.05 and an odds ratio (OR) greater than 1. These

243 significant cancers type diagnostic DNA alterations are summarized in Supplementary
244 Tables 9-10.

245

246 ***Survival analysis***

247 Overall survival (OS) was measured from the date of CUP diagnosis (histologically
248 confirmed) to the date of death from any cause. Thirty-five patients had death recorded
249 after completion of study follow-up. Patients without a recorded death were censored
250 based on the date of the last follow-up (12 months). Survival analysis was performed
251 using the R package *survival* (v3.1-12). Kaplan-Meier estimates of OS were presented
252 along with log-rank tests for comparison between resolved and true CUP groups. A cox
253 proportional-hazard model was used for multivariable analysis and adjusted for ECOG
254 (0,1,2,3), age (>60 or <60), gender, and ESMO assigned favorable and unfavorable
255 prognosis groups. Kaplan–Meier and forest plots were produced with
256 the *survminer* package (v0.4.8.999).

257

258 **Results**

259 ***Study cohort and subclassification of CUPs***

260 A total of 215 patients were recruited to the SUPER study, and a summary of baseline
261 characteristics is shown in Table 1. Eighty-nine percent (191/215) received GEP, 93%
262 (201/215) of patients had DNA sequencing, and 82% (177/215) received both assays
263 (Table 1). A latent primary (LP-CUP) cancer was reported by the treating clinician
264 during clinical follow-up in 10% (22/215) of cases based on histopathology, clinical

265 presentation, and/or cancer imaging. In another 13% (27/215) of cases, a likely TOO
266 was assigned after a retrospective review of described morphology, clinical picture, and
267 IHC staining; these cases were termed histology resolved CUP (HR-CUP). Notably,
268 among the LP/HR-CUP cases, there was an enrichment of patients with a prior history
269 of cancer (35%, 17/49), eight of whom likely had a recurrence of their previous disease
270 (Supplementary Table 1).

271
272 Most of the patients (166/215, 77%) were deemed to be True-CUP (T-CUP) and had
273 sufficient IHC workup according to current ESMO guidelines ⁵. Additional IHC stains
274 may have been informative in a small subset (7/166, 4%) but could not be done owing
275 to tissue availability. The T-CUPs were classified into histomorphological subtypes using
276 a modified version of the Memorial Sloan Kettering Cancer Center OncoTree
277 classification (see Methods) (Supplementary Table 2). The majority of T-CUPs were
278 adenocarcinomas (51%) or poorly differentiated carcinomas (25%) with minor subsets
279 of SCCs (13%), undifferentiated neoplasms (7%), and neuroendocrine neoplasms (2%).
280 The most frequently assigned cancer types for LP/HR-CUPs or CUP OncoTree
281 classifications are summarized in Supplementary Table 11.

282
283 ***GEP classification confidence is lower for True-CUPs than known metastatic***
284 ***cancers***

285 We used two GEP methods for TOO classification of 191/215 CUP tumors, where
286 sufficient RNA was available. A previously described 18-class microarray-based

287 classifier (CUPGuide) was used in 20 cases ¹⁸, while a novel NanoString classifier was
288 used in the remaining cases (Supplementary Table 1-2). The NanoString classifier was
289 validated using an independent cohort of 188 metastatic tumors of known origin
290 (Supplementary Table 6), achieving an overall prediction accuracy of 82.9%, increasing
291 to 91.5% considering only high-medium confidence classifications (n=154) (Figure 1A,
292 Supplementary Table 5.

293
294 GEP TOO classification was possible for 45 LP/HR-CUPs (Figure 1B), of which 80%
295 (36/45) had a high-medium confidence classification (Figure 1C). Three LP/HR-CUPs
296 were considered to be outside the 18-class GEP TOO differential, including one rare
297 ampullary tumor and two uterine tumors. Among high-medium confidence predictions
298 within the GEP diagnostic differential, 94% (31/33) were concordant with their assigned
299 diagnosis (Supplementary Table 1). Common high-medium confidence
300 misclassifications observed among LP/HR-CUPs as well as the 188 known metastatic
301 cancers included cholangiocarcinoma (1/1 and 5/11, respectively) and pancreatic
302 adenocarcinomas (3/10 in the known metastatic group), illustrating classification
303 difficulty among the pancreatobiliary group (Figure 1A, B).

304
305 GEP TOO classification was performed on 146 T-CUPs. High-medium confidence
306 classifications were made for only 56% (82/146) of T-CUPs (Figure 1C). The most
307 frequent high-medium confidence classifications included SCC (32%), liver (13%),
308 colorectal (12%), breast (10%) and lung (8.5%) (Figure 1D). A reduced proportion of

309 high-medium confidence classifications for T-CUPs compared to known metastatic
310 cancers suggested T-CUPs either have an atypical transcriptional profile or potentially
311 are cancer types outside of the GEP classifier differential.

312

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

315 DNA panel sequencing was performed for 201/215 CUP tumors and their matching
316 germline controls. We detected mutational features including single nucleotide variants
317 (SNVs), gene-fusions, copy-number amplifications (CNAs), SNV 96 trinucleotide
318 mutational signatures (COSMIC v2)³², tumor mutation burden (TMB), and off-target viral
319 DNA sequences (HPV, EBV) (Figure 2). Viral RNA transcripts detected by NanoString
320 also supported viral status for HPV-positive tumors.

321

322 At least one protein-coding mutation was found in 98.5% (198/201) of all CUPs, with a
323 median TMB of 4.4 mutations/Mb (range 0.5-149 mutations/Mb). The most frequently
324 mutated genes were *TP53* (55%), *LRP1B* (20%), *PIK3CA* (17%), *KMT2D* (15%), *KRAS*
325 (12%), *ARID1A* (11%) and *SMARCA4* (11%) (Figure 2). The variant allele frequency of
326 these mutations ranged between 16-40.5%, consistent with clonal cancer driver events
327 when tumor purity was considered. Additionally, 8% (n=16) of the cohort had dominant
328 Signature 4 (tobacco smoking), 5% (n=11) Signature 7 (ultra-violet light, UV) and 2%
329 (n=4) Signature 6 (microsatellite instability, MSI). HPV16 (DNA and RNA) was detected
330 in five cases and EBV (DNA only) in one case.

331
332 The gene-wise mutation frequency in the CUP cohort was also compared to 2,785
333 CUPs with panel sequencing in the AACR Project GENIE database (Figure 2). The
334 mutation profile between the study CUP and GENIE CUP cohorts was highly similar
335 with some minor differences, including a higher frequency of *KRAS* mutations (12% vs.
336 22%) among the GENIE CUPs and higher *LRP1B* (18% vs. 2%) mutations in the
337 SUPER cohort; the latter explained by *LRP1B* not being included in MSK-IMPACT
338 panel³³.

339
340 Actionable mutations were also investigated in reference to the CUPISCO trial criteria of
341 actionability described previously²². Eighty-six CUP patients (40%) were matched to
342 either a targeted therapy or immunotherapy arm of CUPISCO (Figure 2)
343 (Supplementary Table 12).

344

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

346 We next considered the diagnostic value of DNA and RNA features, including GEP
347 classification in combination with the histopathological review. Here we considered
348 driver gene mutations, gene fusions, mutational signatures, oncoviruses, and high-
349 medium confidence GEP classifications only. To identify gene features with significant
350 cancer type associations, we referenced the GENIE database (77,058 tumor samples)
351 involving 22 solid cancer types. The potential diagnostic utility of a gene feature was
352 determined by comparing one cancer type versus all others (Fisher exact test adjusted

353 p-value < 0.05 and odds-ratio (OR) >1) (Supplementary Table 10). A total of 171 genes
354 were significantly enriched in one or more cancer types, and 90 genes were enriched in
355 only one cancer type (Figure 3A, Supplementary Figure 1). Mutational signatures also
356 provided cumulative evidence to support likely TOO, for example, a Signature 7 (UV)
357 associated with skin cancer or a Signature 4 (tobacco) found in cancers of the airways,
358 although potentially not excluding liver cancer ³².

359
360 We first assessed LP/HR-CUP cases that had been assigned a single TOO prior to
361 considering the genomics data. Sixty-nine percent of cases (33/49) had one or more
362 features (RNA and/or DNA) consistent with the TOO diagnosis. Both GEP classification
363 and mutation profiling was informative in 20/33 cases, DNA features alone in 7/33, GEP
364 alone in 5/33, and viral detection alone in 1/33 cases. The seven cases where GEP was
365 not informative included two where GEP was not possible, four with low confidence
366 classification, and one was a rare cancer part of the classifier differential. High-medium
367 confidence GEP classifications were the most useful diagnostic feature (n=25) followed
368 by gene mutations (n=24), mutational signatures (n=4), copy number amplifications
369 (n=2), gene-fusions (n=1), and oncoviruses (n=1) (Figure 3B, Supplementary Figure 2).

370
371 Examples of LP/HR-CUP diagnoses supported by genomic data included 8/10 (80%)
372 ovarian cases with high-confidence GEP classification as well as a *TP53* mutation, the
373 latter occurring in >96% of high-grade serous ovarian cancers ³⁴. A dominant mutation
374 Signature 7 (UV)(n=3) and either oncogenic mutations in *NRAS* (melanoma OR=20.7)

375 or truncating *NF1* mutations (melanoma OR=5.2) was consistent with cutaneous
376 melanoma³⁵. High-confidence GEP prediction of gastric cancer combined with *ERBB2*
377 amplification (concordant with positive HER2 IHC staining) (esophagus/stomach OR=3)
378 and a frameshift deletion in *CDH1* (esophagus/stomach OR=3) supported a diagnosis of
379 gastric adenocarcinoma. High-confidence colorectal GEP prediction and *APC* mutations
380 (colorectal OR= 77.5) were present in two cases diagnosed as colorectal
381 adenocarcinomas³⁶(Supplementary Figure 2, Supplementary Table 1).

382

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

389 Molecular features supported a likely single TOO diagnosis consistent with
390 clinicopathological features in 37% (61/166) of T-CUPs (Figure 3C). DNA features
391 supported a diagnosis in 31% (51/166) of T-CUP cases, GEP TOO classification in 13%
392 (21/166), and viral detection in 4% (6/166). Combined GEP classification and mutation
393 profiling features were informative in 10% 16/166 cases (Figure 3D). Types of genomic
394 feature supporting a diagnosis included driver gene mutations (n=36), mutational
395 signatures (n=23), HPV16 (n=5) and EBV (n=1) viral nucleic acids, gene amplification
396 (n=4), gene-fusions (n=3) and high-medium confidence GEP prediction (n=21) (Figure

397 3B). DNA features could narrow the differential diagnosis in a further 11/166 T-CUPs
398 (7%), although a TOO assignment of a single site could not be confidently made (Figure
399 3C). Considering the genomics data, the most frequently suspected cancer types
400 among T-CUPs were lung (n=18, including a single pleomorphic carcinoma of the lung
401 (LUPC)), biliary tract (n=8), breast (n=5), colorectal (n=5), HPV+ SCC (n=5) and kidney
402 (n=4) (Figure 3E).

403 Lung-CUP was the single largest single group among T-CUPs. Dominant mutational
404 Signature 4 (tobacco) was found in 14 cases. Driver gene mutations associated with
405 non-small cell lung cancer (NSCLC) included *KEAP1* (Lung OR=9.8), *STK11* (Lung
406 OR=14.1), *SMARCA4* (Lung OR=2.8), and *KRAS* (Lung OR=2.1) (Figure 3E,
407 Supplementary Figure 1 and Supplementary Table 10). Notably, all but one putative
408 Lung-CUP lacked TTF1 IHC staining. Among the Lung-CUPs where GEP was possible,
409 a high-medium confidence classification of Lung was made in 5/16 cases. Two
410 additional T-CUPs that were unresolved had high-medium lung classification, but
411 mutation profiling was unsuccessful, and there was no other evidence to support a lung
412 diagnosis. Three Lung-CUPs were CDX2-positive by IHC but had mutational features
413 consistent with lung cancer, including a Signature 4 (tobacco) in all three cases. GEP
414 was uninformative in these cases as one was classified as colorectal (0.9-confidence
415 probability), and two were predicted SCC, a non-specific classification but potentially in
416 keeping with lung-squamous cancer (Figure 3E, Supplementary Table 2). These CDX2+
417 Lung-CUPs were favored to be enteric-like lung adenocarcinomas ³⁷.

418

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

430
431 In four T-CUPs subsequently assigned a kidney TOO, two cases expressed PAX8 by
432 IHC, both confirmed by high *PAX8* mRNA expression (z-score > 2), and an additional
433 case did not have PAX8 IHC performed but had high *PAX8* mRNA expression. The
434 fourth case did not have PAX8 staining or high *PAX8* mRNA expression. Only two
435 cases were classified as kidney cancer by GEP (Figure 3E). Driver mutations
436 significantly associated with kidney cancer, and detected among kidney-CUPs, included
437 *BAP1* (kidney OR=7.6) and *NF2* (kidney OR=7.2). Another case had a truncating *FH*
438 mutation consistent with *FH*-deficient kidney cancer (kidney OR=8.1) (Supplementary
439 Figures 1). Notably, no *VHL* mutations were detected, representing the most common
440 driver gene in renal cell carcinoma (RCC) (42.5% RCC in GENIE, OR=763). Another

441 assigned kidney case was confirmed as a recurrence of late-onset adult Wilm's tumor
442 by detecting a somatic *FGFR1* (p.K656D) mutation in both the original primary tumor
443 and a recurrent tumor that presented over twenty years after first diagnosis
444 (Supplementary Table 2 and 10).

445

446 ***Unresolved CUPs have a reduced survival outcome compared to resolved CUPs***

447 Of the 215 patients, 177 patients had recorded survival data. Among all cases recruited
448 to the SUPER study, the majority were classified as unfavorable according to the ESMO
449 guidelines (137/159= 86% of T-CUP, and 39/48=81% of LP/HR-CUP, where data was
450 available) ⁵ (Figure 4A). Overall, 85% of the cohort was unfavorable CUP and, as
451 expected, these had significantly poorer overall survival (OS)- by log-rank test ($p <$
452 0.001), with a median OS of 11 months compared to a median OS of 15 months for
453 favorable CUP. Additionally, 46% of unfavorable CUP achieved 12 months of survival
454 compared with 92% of favorable CUP (Figure 4B). Furthermore, log-rank testing
455 showed longer survival in LP/HR-CUPs, and T-CUPs assigned a single TOO after
456 interpretation of the genomics data, compared with diagnostically unresolved T-CUPs
457 ($p=0.04$, Figure 4C). Multivariate Cox-regression analysis, including gender, age, ECOG
458 performance, and ESMO prognosis groups (Figure 4D), showed that T-CUPs had
459 worse survival outcomes compared to LP/HR-CUPs (HR=1.89, 95% CI 1.13-3.2,
460 $p=0.016$) while there was no significant difference between LP/HR-CUPs and resolved
461 T-CUPs (HR=1.18, 95% CI 0.65-2.1, $p=0.595$).

462

463 **Discussion**

464 The reported incidence of CUP has decreased in recent years ¹. This is most likely due
465 to increased awareness and standardization of investigations, including access to more
466 specific IHC stains, improved diagnostic imaging, and the use of molecular profiling¹.
467 Consistent with other large retrospective CUP studies, we found that approximately
468 one-quarter of CUP may be assigned a likely TOO based on a centralized
469 histopathology review ^{42,43}. This is similar to the recent experience of the international
470 CUP clinical trial CUPISCO, where ~20% of patients had a single primary site diagnosis
471 supported by available evidence or was TOO was strongly suspected ⁴⁴. In the current
472 study, we have shown that incorporating DNA and RNA tests may help to resolve more
473 than a third of T-CUP cases not otherwise resolved using conventional testing.
474 Importantly, despite GEP being the most commonly explored molecular test for
475 resolving CUP, we found DNA sequencing was of potentially greater diagnostic value
476 among CUP tumors due to many CUPS having an atypical transcriptional profile.

477
478 Despite positive evidence that GEP can be diagnostically informative for CUP, cases
479 with poorly differentiated histopathology and atypical disease presentation remain
480 challenging ^{45,46}. This is perhaps not surprising given that GEP classification relies upon
481 the expression of cellular differentiation markers that are often lost or equivocal in CUP
482 tumors ⁴⁷. Our observation that fewer CUP pass a high-medium confidence GEP
483 classification compared to known metastatic cancers, hence have a reduced classifier
484 performance, is supported by other studies. For instance, the CancerTYPE ID

485 (Biotheranostics) 92-gene test, which had extensive multisite validation, showed an
486 overall accuracy of 85% for known cancer metastases. But in CUP, the concordance of
487 GEP with IHC and clinicopathological evidence was lower at 75% for LP-CUP and 70%
488 compared to the clinical picture only ⁷. Additionally, rare malignancies are infrequently
489 included in the training set of TOO classifiers ⁶, which poses a limitation for rare entities
490 among a cohort of CUP. More specifically, subsets of lung and biliary cancers can be
491 problematic for GEP classifiers to resolve. For example, among lung CUPs, GEP
492 classification was found to be concordant with a latent primary tumor in only 50% of
493 cases ¹³. GEP classification accuracy can also be low for cholangiocarcinomas, as the
494 transcriptional profile is similar to pancreatic and upper gastrointestinal neoplasms ¹⁰.
495 Notably, some previously described GEP and DNA methylation tests have also
496 excluded cholangiocarcinoma in their models or combined them into a
497 pancreaticobiliary class ^{9,48}. We found that DNA sequencing may be particularly useful
498 in pancreaticobiliary CUPs given that gene mutations characteristic of
499 cholangiocarcinomas can have both diagnostic and occasionally therapeutic
500 significance, including alterations in *IDH1*, *FGFR2*, and *BAP1* ³⁹⁻⁴¹.

501
502 Integrated molecular profiling of CUP can help identify rare disease subtypes and our
503 data supports their being some recurrent CUP entities. We identified examples of
504 pulmonary enteric adenocarcinomas that lack TTF1 expression but can express the
505 gastrointestinal marker CDX2 ³⁷. GEP or IHC alone could not resolve such cases, given
506 their atypical profile; however, they still retained DNA features highly suggestive of

507 NSCLC, including a tobacco mutational signature and somatic mutations in *KRAS*,
508 *STK11*, and *SMARCA4*. *SMARCA4*-deficient lung cancers are known to lack TTF1
509 expression⁴⁹ and are likely to be frequent in the CUP population^{19,20}. *SMARCA4*-
510 deficient lung cancer models similarly lose expression of lung differentiation markers
511 and have a pro-metastatic behavior consistent with aggressive clinical course and poor
512 survival outcomes in patients⁵⁰. Kidney cancers are another emerging CUP entity
513 representing ~4-6% of CUPs^{44,51,52}. Four CUPs in the SUPER cohort were consistent
514 with a primary kidney tumor. Interestingly, we found somatic mutations in *NF2*, *FH*, and
515 *BAP1* in some RCC cases. Somatic *NF2* mutations are characteristic of advanced
516 papillary renal cell tumors and those with biphasic hyalinizing psammomatous features
517⁵³. Papillary carcinomas have also been found to be enriched among other Kidney-
518 CUPs^{51,52} with a mutation profile similar to the Kidney-CUPs described in the SUPER
519 cohort⁵⁴. Identifying CUP entities and recurrent therapeutic targets in these groups may
520 help guide future CUP clinical trials. For instance, while empirical chemotherapy is
521 ineffective in RCC, targeted therapies and ICIs are likely more efficacious^{51,55}.
522 Furthermore, detection of *NF2* mutations among RCC and mesothelioma could direct
523 targeted treatment of the Hippo pathway using inhibitors of TEAD auto-palmitoylation⁵⁶,
524 which have now entered clinical trials (NCT04665206).
525
526 CUP can be classified into favorable and unfavorable prognosis groups. Given that
527 ~85% of CUPs belong to the unfavorable group, research and clinical trials directed at
528 identifying other treatments for these patients or therapeutic responsive subsets should

529 be sought. Approximately 40% of T-CUPs in the SUPER cohort had potentially
530 actionable mutations and genomic alterations, broadly consistent with other CUP
531 profiling studies ^{19,20,22}. As curated findings from genomic tests were returned to
532 clinicians during the study, it is possible that the genomic data allowed site-specific
533 treatments and access to targeted treatments, potentially explaining the improved
534 survival outcome among the resolved compared to the unresolved T-CUP group.
535 Alternatively, genomically resolved CUPs may be enriched for cancer types where
536 targeted or immunotherapy treatments are available and effective.

537
538 In conclusion, we have shown that DNA and RNA tests can be incorporated into a
539 pathology assessment to improve cancer type diagnosis, identify CUP subtypes, and
540 direct treatments. Rather than replacing traditional histopathological analysis, molecular
541 testing can augment conventional testing to either confirm a suspicion of primary tissue
542 of origin or provide robust diagnostic leads that are not otherwise evident using other
543 modalities. In practice, in cases where tissue is limited, it may be more informative to
544 prioritize genomic testing to guide additional investigations before consuming tissue on
545 extended IHC panels. With steady improvements in technology and reduction in the
546 sequencing costs, more comprehensive whole-genome and transcriptome analysis will
547 likely increase the sensitivity to detect features such as structural variants and
548 mutational signatures that are not reliably detected by panel sequencing ⁵⁷. Recent
549 studies have also utilized machine-learning for classification based on genome-wide
550 patterns of somatic mutations that are not otherwise interpretable by human curation

551 ^{58,59} and combined DNA and RNA features ⁶⁰ to improve the TOO prediction accuracy
552 further (60-91% and 94%, respectively). Although molecular testing is not currently
553 recommended in most CUP guidelines, careful prospective evaluation of these tests
554 may justify their cost, helping to resolve cancer type in a significant fraction of CUP
555 cases as well as direct treatments that will ultimately translate to better outcomes for
556 patients.

557

558

559 ***Figure legends***

560 **Figure 1:** Gene expression profiling (GEP) tissue of origin classification of known
561 metastatic cancers and SUPER cancer of unknown primary (CUP) tumors. A)
562 NanoString GEP classifier was tested on 188 known origin metastatic tumors with
563 confusion matrix showing concordance of tissue of origin prediction and known cancer
564 type. B) GEP classifier tested on latent primary/histology resolved (LP/HR)-CUPs
565 showing concordance between the likely tissue of origin and the predicted cancer type.
566 LP/HR-CUPs representing cancer types that were not represented in the classifier
567 model were removed from analysis C) Fraction of cases within confidence probability
568 score grouping contrasting classification of true-CUPs (T-CUP) and LP/HR-CUPs
569 combined with known metastatic tumors (unclassified <0.5, low ≥ 0.5 and ≤ 0.7 medium
570 confidence ≥ 0.8 and ≤ 0.9 , high confidence = 1). D) GEP cancer class predictions of all
571 T-CUPs with high-medium confidence predictions.

572

573 **Figure 2:** DNA mutational profiling of the SUPER cancer of unknown primary (CUP)
574 cohort. Oncoplot shows somatic mutations in CUPs in descending order of frequency.
575 Genes and mutational features colored in red are actionable and targeted in the
576 CUPISCO trial. The proportion of mutations in the SUPER CUPs was compared with
577 the AACR GENIE CUP cohort (right-hand bar plot). The left-hand plot of the variant
578 allele frequency (VAF) distribution per gene. The top bar plot shows the number of
579 coding mutations per sample. Annotations include OncoTree class for true-CUPs (T-
580 CUP) or the assigned tumor class for latent primary/histology resolved (LP/HR)-CUPs,
581 detection of COSMIC mutational signatures (V2): smoking signature, ultra-violet (UV)
582 signature, DNA mismatch repair signature, oncoviruses: Human papillomavirus 16
583 (HPV16) and Epstein-Barr virus (EBV), and tumor mutation burden (TMB) status: High
584 >10 mutations/Mb or Low <10 mutations/Mb.

585
586 **Figure 3:** Identification of diagnostically useful DNA features using AACR Project
587 GENIE mutation data and summary of evidence used to support tissue of origin (TOO)
588 diagnosis among SUPER cancer of unknown primary (CUP) cases. A) Volcano plot
589 showing significantly mutated features by cancer types using GENIE pan-cancer cohort.
590 All genes are plotted for all cancers using a Fisher's exact test to calculate adjusted p-
591 values and odds ratio (OR) (significance threshold adjusted p-value < 0.05 and OR >1).
592 B) Proportion and number of cases where each genomic feature supported a putative
593 TOO for true-CUPs (T-CUP) and latent primary/histology resolved (LP/HR)-CUPs.
594 Genomic features included single nucleotide variants (SNV), Gene expression profiling

595 (GEP), mutational signatures, oncoviral sequences, and copy number amplifications
596 (CNA). C) OncoTree cancer classification of T-CUPs before and after genomic analysis.
597 CUPs were either resolved to a putative TOO, had reduced occult diagnosis, or
598 remained T-CUP. D) Proportion of cases where DNA and/or GEP classification
599 supported TOO diagnosis among T-CUPs. E) Detailed summary of supportive genomic
600 features and IHC staining used to assign a putative TOO for all genomically resolved T-
601 CUPs. NGS= Next generation sequencing.

602
603 **Figure 4:** A comparison of overall survival outcome among SUPER CUP patients based
604 on clinical and diagnostic groups A) Proportion and number of true-CUPs (T-CUP) and
605 latent primary/histology resolved (LP/HR)-CUPs classified by European Society of
606 Medical Oncology (ESMO) favorable and unfavorable prognosis groups. B) Kaplan-
607 Meier curve of overall survival (OS) in unfavorable and favorable CUP types. Censored
608 at 12 months and significance using log-rank p-values. C-D) Log-rank test and
609 multivariable Cox-proportional hazard model of SUPER CUP clinical and diagnostic
610 group. Data censored at 12 months and significance using log-rank p values for Kaplan-
611 Meier analysis. Cox-proportional hazard model including covariates of ECOG, age (>60
612 or <60), ESMO favorable and unfavorable outcome categories, and patient sex. GR =
613 genomically resolved T-CUPs, U= unfavorable type, F= favorable type, HR= hazard
614 ratio, CI= confidence interval.

615
616

617

Table 1. Study cohort characteristics

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

*DNA sequencing and GEP performed successfully

618

619 **Reference list**

- 620 1 Rassy, E. & Pavlidis, N. The currently declining incidence of cancer of unknown primary.
621 *Cancer Epidemiol* **61**, 139-141 (2019).
- 622 2 Australian Institute of Health and Welfare (AIHW) 2018 Cancer Data in Australia;
623 Australian Cancer Incidence and Mortality (ACIM) books: cancer of unknown primary
624 site Canberra: AIHW.
- 625 3 Massard, C., Lorient, Y. & Fizazi, K. Carcinomas of an unknown primary origin--diagnosis
626 and treatment. *Nat Rev Clin Oncol* **8**, 701-710 (2011).
- 627 4 Hyphantis, T., Papadimitriou, I., Petrakis, D., Fountzilias, G., Repana, D.,
628 Assimakopoulos, K. et al. Psychiatric manifestations, personality traits and health-related
629 quality of life in cancer of unknown primary site. *Psycho-oncology* **22**, 2009-2015 (2013).
- 630 5 Fizazi, K., Greco, F. A., Pavlidis, N., Daugaard, G., Oien, K., Pentheroudakis, G. et al.
631 Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis,
632 treatment and follow-up. *Ann Oncol* **26**, v133-138 (2015).
- 633 6 Erlander, M. G., Ma, X. J., Kesty, N. C., Bao, L., Salunga, R. & Schnabel, C. A.
634 Performance and clinical evaluation of the 92-gene real-time PCR assay for tumor
635 classification. *J Mol Diagn* **13**, 493-503 (2011).
- 636 7 Kerr, S. E., Schnabel, C. A., Sullivan, P. S., Zhang, Y., Singh, V., Carey, B. et al.
637 Multisite validation study to determine performance characteristics of a 92-gene
638 molecular cancer classifier. *Clin Cancer Res* **18**, 3952–3960 (2012).

- 639 8 Hainsworth, J. D. & Greco, F. A. Gene expression profiling in patients with carcinoma of
640 unknown primary site: from translational research to standard of care. *Virchows Arch*
641 **464**, 393-402 (2014).
- 642 9 Moran, S., Martínez-Cardús, A., Sayols, S., Musulén, E., Balañá, C., Estival-Gonzalez,
643 A. et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre,
644 retrospective analysis. *Lancet Oncol* **17**, 1386-1395 (2016).
- 645 10 Zhao Y, P. Z., Namburi S, Pattison A, Posner A, Balachander S, Paisie C.A., Reddi H.V.,
646 Rueter J, Gill AJ, Fox S, Raghav KPS., Flynn WF, Tothill RW, Li S, Karuturi RKM,
647 George J. CUP-AI-Dx: a tool for inferring cancer tissue of origin and molecular subtype
648 using RNA gene-expression data and artificial intelligence. *EBioMedicine* (2020).
- 649 11 Handorf, C. R., Kulkarni, A., Grenert, J. P., Weiss, L. M., Rogers, W. M., Kim, O. S. et al.
650 A multicenter study directly comparing the diagnostic accuracy of gene expression
651 profiling and immunohistochemistry for primary site identification in metastatic tumors.
652 *Am J Surg Pathol* **37**, 1067-1075 (2013).
- 653 12 Tothill, R. W., Shi, F., Paiman, L., Bedo, J., Kowalczyk, A., Mileskin, L. et al.
654 Development and validation of a gene expression tumour classifier for cancer of
655 unknown primary. *Pathology* **47**, 7-12 (2015).
- 656 13 Greco, F. A., Lenington, W. J., Spigel, D. R. & Hainsworth, J. D. Molecular profiling
657 diagnosis in unknown primary cancer: accuracy and ability to complement standard
658 pathology. *J Natl Cancer Inst* **105**, 782-790 (2013).

- 659 14 Hainsworth, J. D., Rubin, M. S., Spigel, D. R., Boccia, R. V., Raby, S., Quinn, R. et al.
660 Molecular gene expression profiling to predict the tissue of origin and direct site-specific
661 therapy in patients with carcinoma of unknown primary site: a prospective trial of the
662 Sarah Cannon research institute. *J Clin Oncol* **31**, 217-223 (2013).
- 663 15 Yoon, H. H., Foster, N. R., Meyers, J. P., Steen, P. D., Visscher, D. W., Pillai, R. et al.
664 Gene expression profiling identifies responsive patients with cancer of unknown primary
665 treated with carboplatin, paclitaxel, and everolimus: NCCTG N0871 (alliance). *Ann*
666 *Oncol* **27**, 339-344 (2016).
- 667 16 Haratani, K., Hayashi, H., Takahama, T., Nakamura, Y., Tomida, S., Yoshida, T. et al.
668 Clinical and immune profiling for cancer of unknown primary site. *J Immunother Cancer*
669 **7**, 251 (2019).
- 670 17 Fizazi, K., Maillard, A., Penel, N., Baciarello, G., Allouache, D., Daugaard, G. et al.
671 LBA15_PRA phase III trial of empiric chemotherapy with cisplatin and gemcitabine or
672 systemic treatment tailored by molecular gene expression analysis in patients with
673 carcinomas of an unknown primary (CUP) site (GEFCAPI 04). *Annals of Oncology* **30**
674 (2019).
- 675 18 Tothill, R. W., Li, J., Mileskin, L., Doig, K., Siganakis, T., Cowin, P. et al. Massively-
676 parallel sequencing assists the diagnosis and guided treatment of cancers of unknown
677 primary. *J Pathol* **231**, 413-423 (2013).
- 678 19 Ross, J. S., Wang, K., Gay, L., Otto, G. A., White, E., Iwanik, K. et al. Comprehensive
679 Genomic Profiling of Carcinoma of Unknown Primary Site: New Routes to Targeted
680 Therapies. *JAMA Oncol* **1**, 40-49 (2015).

- 681 20 Varghese, A. M., Arora, A., Capanu, M., Camacho, N., Won, H. H., Zehir, A. et al.
682 Clinical and molecular characterization of patients with cancer of unknown primary in the
683 modern era. *Ann Oncol* **28**, 3015-3021 (2017).
- 684 21 Löffler, H., Pfarr, N., Kriegsmann, M., Endris, V., Hielscher, T., Lohneis, P. et al.
685 Molecular driver alterations and their clinical relevance in cancer of unknown primary
686 site. *Oncotarget* **7**, 44322-44329 (2016).
- 687 22 Ross, J. S., Sokol, E. S., Moch, H., Mileskin, L., Baciarello, G., Losa, F. et al.
688 Comprehensive Genomic Profiling of Carcinoma of Unknown Primary Origin:
689 Retrospective Molecular Classification Considering the CUPISCO Study Design.
690 *Oncologist* (2020).
- 691 23 Consortium, A. P. G. AACR Project GENIE: Powering Precision Medicine through an
692 International Consortium. *Cancer Discov* **7**, 818-831 (2017).
- 693 24 Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J. et al. OncoKB: A
694 Precision Oncology Knowledge Base. *JCO Prec Oncol* 1-16 (2017).
- 695 25 Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J. et al.
696 COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* **45**, D777–D783
697 (2017).
- 698 26 Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A. et al. The
699 cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer
700 Genomics Data. *Cancer Discov* **2**, 401 (2012).

- 701 27 Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O. et al.
702 Integrative analysis of complex cancer genomics and clinical profiles using the
703 cBioPortal. *Sci Signal* **6**, pl1 (2013).
- 704 28 Groschel, S., Bommer, M., Hutter, B., Budczies, J., Bonekamp, D., Heining, C. et al.
705 Integration of genomics and histology revises diagnosis and enables effective therapy of
706 refractory cancer of unknown primary with PDL1 amplification. *Cold Spring Harb Mol*
707 *Case Stud* **2**, a001180 (2016).
- 708 29 Kundra, R., Zhang, H., Sheridan, R., Sirintrapun, S. J., Wang, A., Ochoa, A. et al.
709 OncoTree: A Cancer Classification System for Precision Oncology. *JCO Clin Cancer*
710 *Inform* **5**, 221-230 (2021).
- 711 30 McEvoy, C. R., Semple, T., Yellapu, B., Choong, D. Y., Xu, H., Mir Arnau, G. et al.
712 Improved next-generation sequencing pre-capture library yields and sequencing
713 parameters using on-bead PCR. *Biotechniques* **68**, 48–51 (2020).
- 714 31 Chang, M. T., Bhattarai, T. S., Schram, A. M., Bielski, C. M., Donoghue, M. T. A.,
715 Jonsson, P. et al. Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer*
716 *Discov* **8**, 174-183 (2017).
- 717 32 Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A.
718 V. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415-421
719 (2013).
- 720 33 Cheng, D. T., Mitchell, T. N., Zehir, A., Shah, R. H., Benayed, R., Syed, A. et al.
721 Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets

- 722 (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical
723 Assay for Solid Tumor Molecular Oncology. *J Mol Diagn* **17**, 251-264 (2015).
- 724 34 Ahmed, A. A., Etemadmoghadam, D., Temple, J., Lynch, A. G., Riad, M., Sharma, R. et
725 al. Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary.
726 *J Pathol* **221**, 49–56 (2010).
- 727 35 Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**,
728 1681-1696 (2015).
- 729 36 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon
730 and rectal cancer. *Nature* **487**, 330-337 (2012).
- 731 37 Li, H. & Cao, W. Pulmonary enteric adenocarcinoma: a literature review. *J Thorac Dis*
732 **12**, 3217-3226 (2020).
- 733 38 Network, C. G. A. R. Integrated Genomic Characterization of Pancreatic Ductal
734 Adenocarcinoma. *Cancer cell* **32**, 185–203.e113 (2017).
- 735 39 Arai, Y., Totoki, Y., Hosoda, F., Shirota, T., Hama, N., Nakamura, H. et al. Fibroblast
736 growth factor receptor 2 tyrosine kinase fusions define a unique molecular subtype of
737 cholangiocarcinoma. *Hepatology* **59(4)**, 1427–1434 (2014).
- 738 40 Javle, M. M., Murugesan, K., Shroff, R. T., Borad, M. J., Abdel-Wahab, R., Schrock, A.
739 B. et al. Profiling of 3,634 cholangiocarcinomas (CCA) to identify genomic alterations
740 (GA), tumor mutational burden (TMB), and genomic loss of heterozygosity (gLOH). *J*
741 *Clin Oncol* **37**, 4087-4087 (2019).

- 742 41 Silverman, I. M., Murugesan, K., Lihou, C. F., Féliz, L., Frampton, G. M., Newton, R. C.
743 et al. Comprehensive genomic profiling in FIGHT-202 reveals the landscape of
744 actionable alterations in advanced cholangiocarcinoma. *J Clin Oncol* **37**, 4080-4080
745 (2019).
- 746 42 Greco, F. A., Oien, K., Erlander, M., Osborne, R., Varadhachary, G., Bridgewater, J. et
747 al. Cancer of unknown primary: progress in the search for improved and rapid diagnosis
748 leading toward superior patient outcomes. *Ann Oncol* **23**, 298-304 (2012).
- 749 43 Horlings, H. M., van Laar, R. K., Kerst, J. M., Helgason, H. H., Wesseling, J., van der
750 Hoeven, J. J. et al. Gene expression profiling to identify the histogenetic origin of
751 metastatic adenocarcinomas of unknown primary. *J Clin Oncol* **26(27)**, 4435–4441
752 (2008).
- 753 44 Pauli, C., Bochtler, T., Mileskin, L., Baciarello, G., Losa, F., Ross, J. S. et al. A
754 Challenging Task: Identifying Patients with Cancer of Unknown Primary (CUP)
755 According to ESMO Guidelines: The CUPISCO Trial Experience. *Oncologist* **26**, e769-
756 e779 (2021).
- 757 45 Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M. et al.
758 Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci*
759 *U S A* **98**, 15149-15154 (2001).
- 760 46 Palmeri, S., Lorusso, V., Palmeri, L., Vaglica, M., Porta, C., Nortilli, R. et al. Cisplatin and
761 gemcitabine with either vinorelbine or paclitaxel in the treatment of carcinomas of
762 unknown primary site : results of an Italian multicenter, randomized, phase II study.
763 *Cancer* **107**, 2898-2905 (2006).

- 764 47 Pavlidis, N. Optimal therapeutic management of patients with distinct clinicopathological
765 cancer of unknown primary subsets. *Ann Oncol* **23 Suppl 10**, x282-285 (2012).
- 766 48 Monzon, F. A., Lyons-Weiler, M., Buturovic, L. J., Rigl, C. T., Henner, W. D., Sciulli, C. et
767 al. Multicenter validation of a 1,550-gene expression profile for identification of tumor
768 tissue of origin. *J Clin Oncol* **27**, 2503-2508 (2009).
- 769 49 Herpel, E., Rieker, R. J., Dienemann, H., Muley, T., Meister, M., Hartmann, A. et al.
770 SMARCA4 and SMARCA2 deficiency in non-small cell lung cancer:
771 immunohistochemical survey of 316 consecutive specimens. *Ann Diagn Pathol* **26**, 47–
772 51 (2017).
- 773 50 Concepcion, C. P., Ma, S., LaFave, L. M., Bhutkar, A., Liu, M., DeAngelo, L. P. et al.
774 Smarca4 Inactivation Promotes Lineage-Specific Transformation and Early Metastatic
775 Features in the Lung. *Cancer Discov* **12**, 562-585 (2022).
- 776 51 Greco, F. A. & Hainsworth, J. D. Renal Cell Carcinoma Presenting as Carcinoma of
777 Unknown Primary Site: Recognition of a Treatable Patient Subset. *Clin Genitourin*
778 *Cancer* **16**, e893-e898 (2018).
- 779 52 Overby, A., Duval, L., Ladekarl, M., Laursen, B. E. & Donskov, F. Carcinoma of
780 Unknown Primary Site (CUP) With Metastatic Renal-Cell Carcinoma (mRCC) Histologic
781 and Immunohistochemical Characteristics (CUP-mRCC): Results From Consecutive
782 Patients Treated With Targeted Therapy and Review of Literature. *Clin Genitourin*
783 *Cancer* **17**, e32-e37 (2019).

- 784 53 Yakirevich, E., Perrino, C., Necchi, A., Grivas, P., Bratslavsky, G., Shapiro, O. et al. NF2
785 mutation-driven renal cell carcinomas (RCC): A comprehensive genomic profiling (CGP)
786 study. *J Clin Oncol* **38**, 726-726 (2020).
- 787 54 Wei, E. Y., Chen, Y. B. & Hsieh, J. J. Genomic characterisation of two cancers of
788 unknown primary cases supports a kidney cancer origin. *BMJ Case Rep* **2015** (2015).
- 789 55 Escudier, B. Combination Therapy as First-Line Treatment in Metastatic Renal-Cell
790 Carcinoma. *N Eng J Med* **380**, 1176-1178 (2019).
- 791 56 Tang, T. T., Konradi, A. W., Feng, Y., Peng, X., Ma, M., Li, J. et al. Small Molecule
792 Inhibitors of TEAD Auto-palmitoylation Selectively Inhibit Proliferation and Tumor Growth
793 of NF2-deficient Mesothelioma. *Mol Cancer Ther* **20**, 986–998 (2021).
- 794 57 Priestley, P., Baber, J., Lolkema, M. P., Steeghs, N., de Bruijn, E., Shale, C. et al. Pan-
795 cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210-216
796 (2019).
- 797 58 Yuan, Y., Shi, Y., Li, C., Kim, J., Cai, W., Han, Z. et al. DeepGene: an advanced cancer
798 type classifier based on deep learning and somatic point mutations. *BMC Bioinformatics*
799 **17**, 476 (2016).
- 800 59 Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Subtypes, P. T. et al. A deep
801 learning system accurately classifies primary and metastatic cancers using passenger
802 mutation patterns. *Nat Commun* **11**, 728 (2020).

803 60 Abraham, J., Heimberger, A. B., Marshall, J., Heath, E., Drabick, J., Helmstetter, A. et al.
804 Machine learning analysis using 77,044 genomic and transcriptomic profiles to
805 accurately predict tumor type. *Transl Oncol* **14**, 101016 (2021).

806

807 **Supplementary Information**

808 **Supplementary Methods**

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

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

813

814 **Supplementary Tables 1-12**

815

816 ***Acknowledgments***

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

828 ***Ethics Approval / Consent to Participate***

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

832 ***Author Contribution Statement***

833 RT and LM conceived the study. AP performed the analysis, drafted the figures and
834 tables. OWJP undertook the histopathology review and data analysis. TS and LM
835 reviewed the clinical data. CW, KF, SW collected the data for the study. ADF, NW, CSK,

836 BG, CS, MS, IMC, GR, MW, NK screened patients for eligibility. DE and NT developed
837 the NanoString classifier, and JG and SB validated the classifier. GW, SW, AB, AJG,
838 BJS, DB, SF, RJH contributed samples for training the classifier. AF, HX and SF
839 performed the mutation profiling. APa performed the bioinformatic analysis. AP and RT
840 co-wrote the manuscript. All authors critically reviewed the manuscript. PS, DB, LM and
841 RT are the principal investigators and obtained research funding to support the study.

842 ***Funding Statement***

843 The study was supported by funding from Cancer Australia (APP1048193,
844 APP1082604) and the Victorian Cancer Agency (TRP13062). RWT was supported by
845 funding from the Victorian Cancer Agency (TP828750). The Westmead, Blacktown and
846 Nepean study sites were supported by the Cancer Institute NSW 11/TRC/1-06,
847 15/TRC/1-01 and 15/RIG/1-16.

848 ***Data Availability Statement***

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

851

852

853

Figure 1

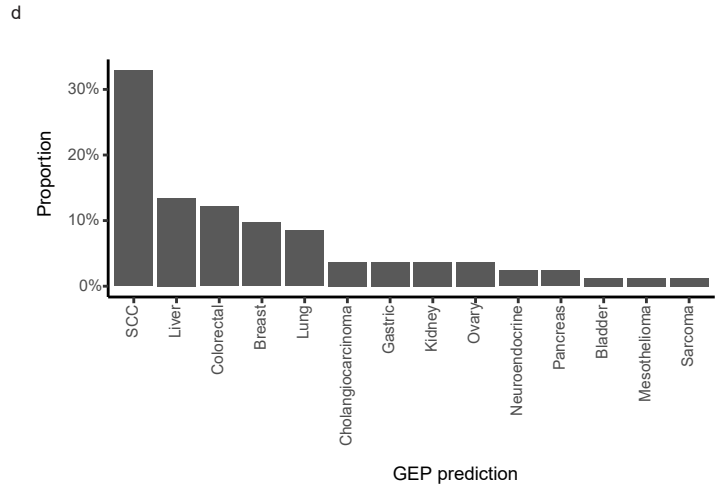
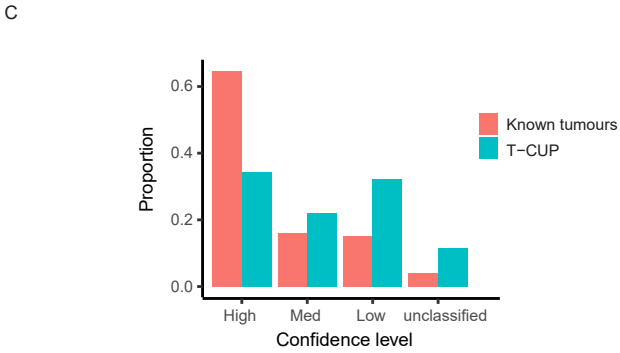
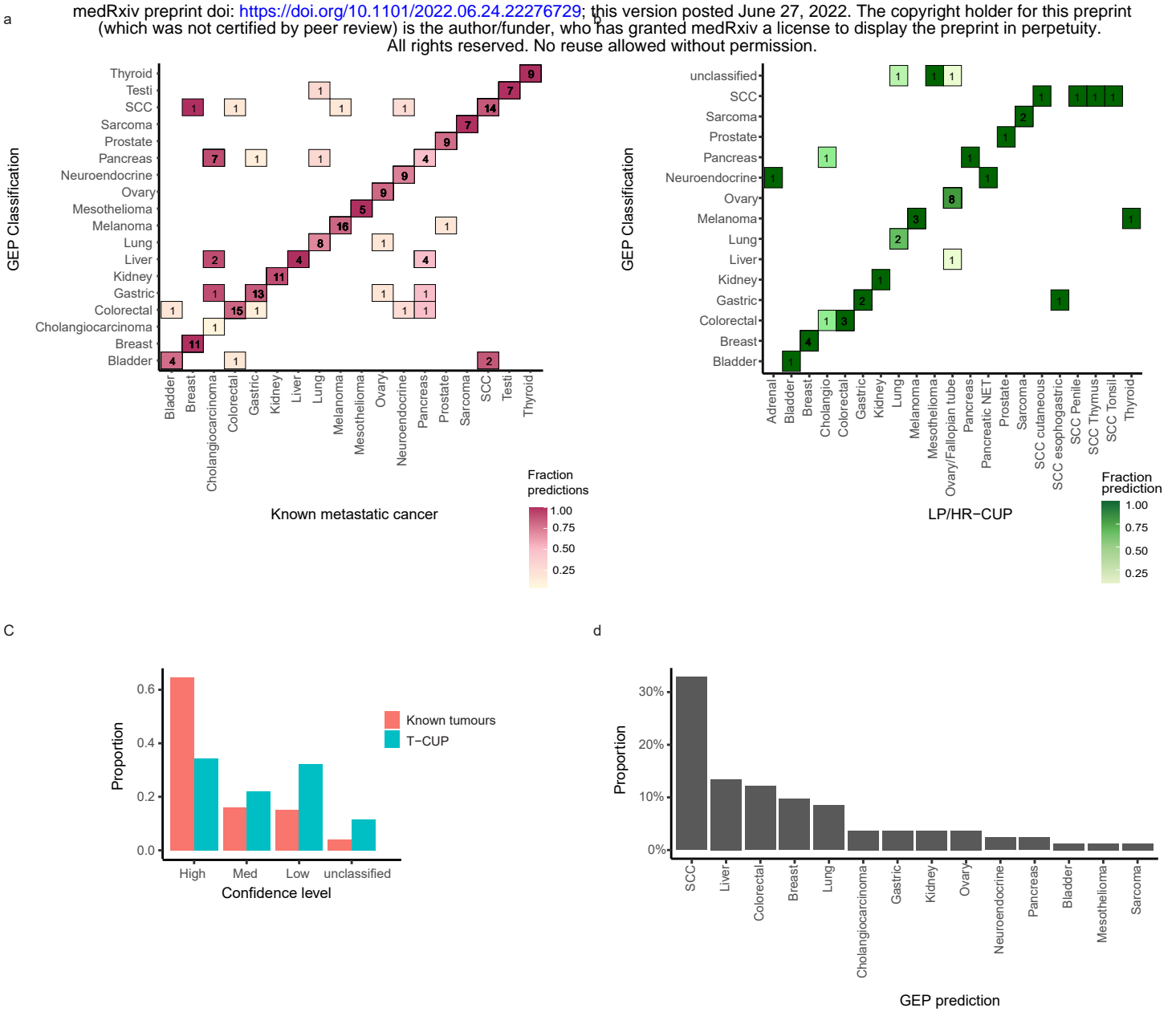
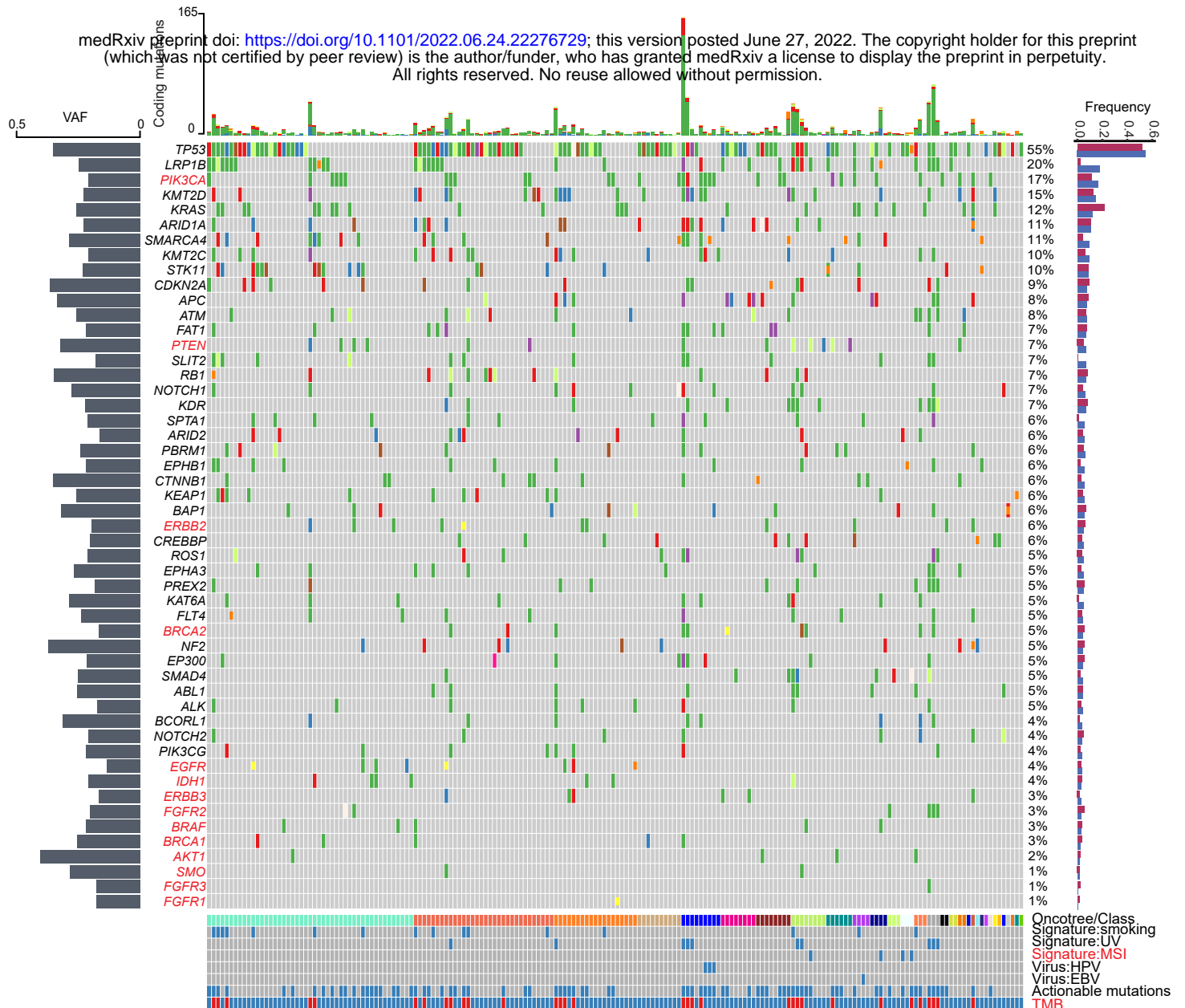


Figure 2

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.



Mutation

- Splice_Site
- Missense_Mutation
- Nonsense_Mutation
- Frame_Shift_Del
- Frame_Shift_Ins
- In_Frame_Ins
- Del
- Amp
- Multi_Hit
- Complex_Event

Oncotree / Class

- SCC p16+
- AD CK7-CK20-
- AD CK7+CK20-
- PDC
- Breast
- Colorectal
- Lung
- Sarcoma
- Gastric
- Prostate
- Cholangio
- NECNOS
- Uterus
- NETNOS
- ADNOS
- Mesothelioma
- AD CK7+CK20+
- Melanoma
- UDMN
- AD CK7-CK20+CDX2+
- Ovary/Fallopian tube
- SCCNOS
- Pancreas
- SCC cutaneous
- Adrenal
- Thyroid
- SCC Penile
- Ampulla
- SCC Thymus
- SCC esophogastric
- Kidney

Mutational feature detected

- NA
- Yes
- TMB
 - Low TMB
 - HighTMB
- Gene proportion
 - Genie CUP
 - T-CUP

Figure 3

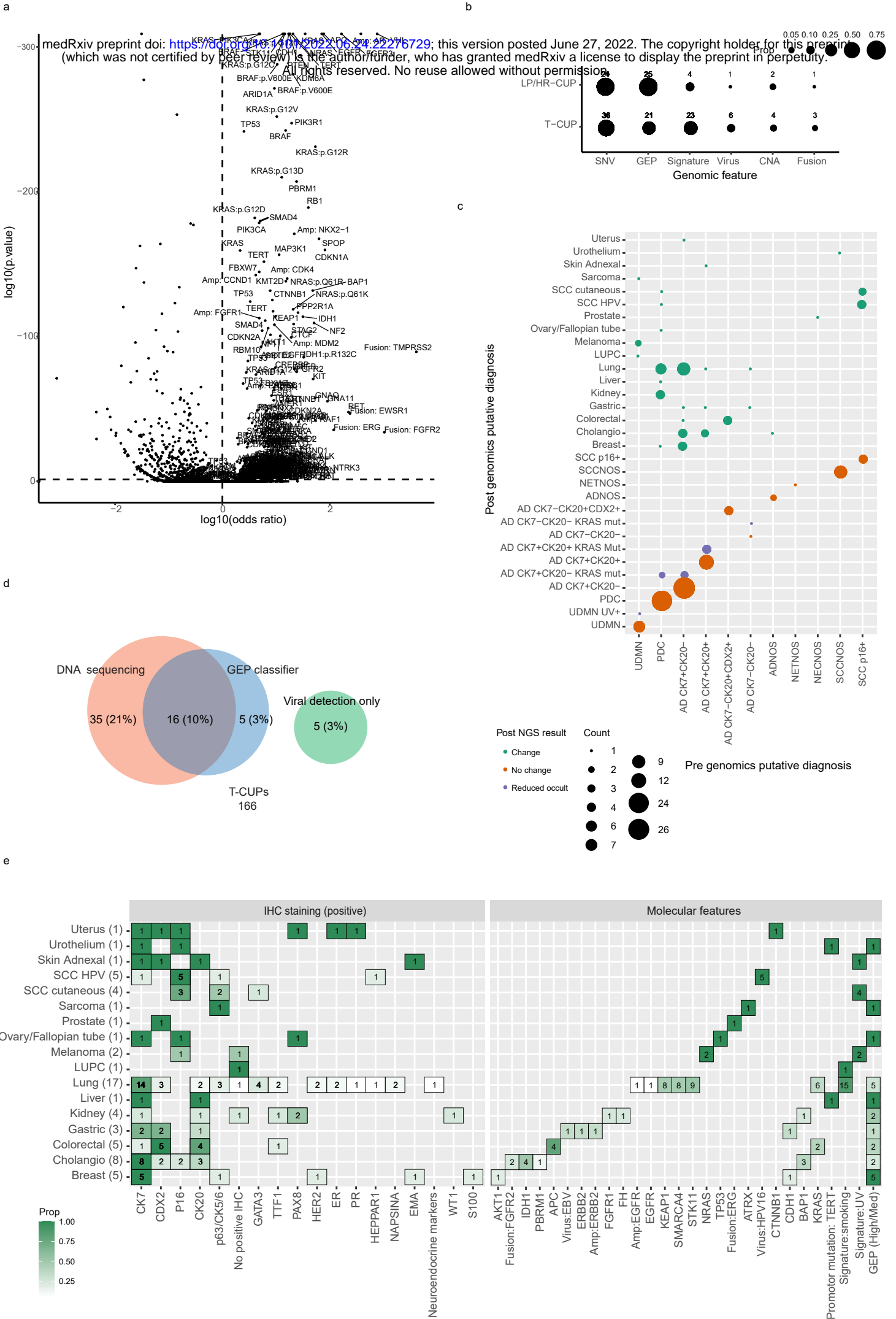
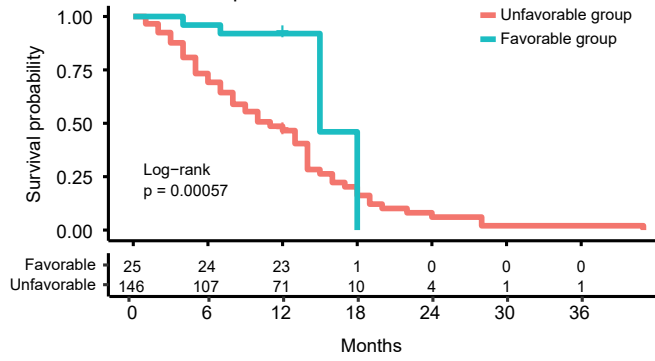
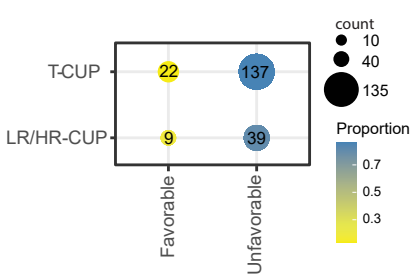


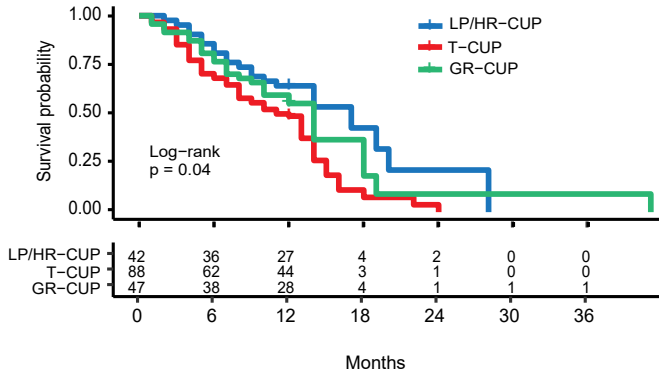
Figure 4

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.

a



c



d

