1 **Mutational signatures in upper tract urothelial carcinoma define etiologically distinct**

2 **subtypes with prognostic relevance**

3 Xuesong Li[1†], Huan Lu[2,3†], Bao Guan [1,2,4,5†], Zhengzheng Xu[2,3], Yue Shi[2], Juan Li[2,3], Wenwen

4 Kong[2,3], Jin Liu[6], Dong Fang[1,5], Libo Liu[1,4,5], Qi Shen[1,4,5], Yanqing Gong[1,4,5], Shiming He[1,4,5], Qun

5 He[1,4,5], Liqun Zhou [1,4,5*] and Weimin Ci [2,3,7*]

6 **Affiliations:**

7 [1]Department of Urology, Peking University First Hospital, Beijing 100034, China.

8 [2]Key Laboratory of Genomics and Precision Medicine, Beijing Institute of Genomics, Chinese

9 Academy of Sciences, Beijing 100101, China;

10 [3]University of Chinese Academy of Sciences, Beijing 100049, China.

11 [4]Institute of Urology, Peking University, Beijing 100034, China.

12 [5]National Urological Cancer Center, Beijing, 100034, China.

13 [6]Department of Urology, The Third Hospital of Hebei Medical University, Hebei 050051, China.

14 [7]Institute of Stem cell and Regeneration, Chinese Academy of Sciences, Beijing 100101, China.

15

16 *Correspondence to: zhoulqmail@sina.com; and ciwm@big.ac.cn.

17 † These authors contributed equally to this work.

18

22

23

Abstract

Parts of East Asia have a very high upper tract urothelial carcinoma (UTUC) prevalence, and an etiological link between the medicinal use of herbs containing aristolochic acid (AA) and UTUC has been established. The mutational signature of AA, which is characterized by a particular pattern of A:T to T:A transversions, can be detected by genome sequencing. Thus, integrating mutational signatures analysis with clinicopathological data may be a crucial step toward personalized treatment strategies for the disease. Therefore, we performed whole-genome sequencing (WGS) of 90 UTUC patients from China. Mutational signature analysis via nonnegative matrix factorization method defined three etiologically distinct subtypes with prognostic relevance: (i) AA, the typical AA mutational signature characterized by signature 22, had the highest tumor mutation burden, the best clinical outcomes; (ii) Age, an age-related group featured by signatures 1 and 5, had the lowest weighted genome instability index score, the worst clinical outcomes; and (iii) DSB, signature with deficiencies in DNA double strand break-repair featured by signatures 3, the intermediate clinical outcomes. Additionally, the distinct AA subtype was associated with AA exposure, the highest number of predicted neoantigens and heavier lymphocytes infiltrating. Thus, it may be good candidate for immune checkpoint blockade therapy. Notably, we showed AA-mutational signature was also identified in histologically "normal" urothelial cells. Thus, non-invasive urine test based on the AA-mutational signature could take advantage of this "field effect" to screen individuals at increased risk of recurrence due to exposure to herbal remedies containing the carcinogen AA. Collectively, the findings here may accelerate the development of novel prognostic markers and personalized therapeutic approaches for UTUC patients in China.

Introduction

The epidemiology and clinicopathological characteristics of UTUC is largely influenced by its geographic distribution. Parts of East Asia such as Taiwan have a much higher UTUC prevalence, accounting for more than 30% of urothelial cancers (UCs)(1-3) compared to only 5-10% of UCs in Western populations(4). Chinese patients are more frequently female, of higher rate for kidney failure(5). Conceivably, geographic differences in risk factors for UTUC, such as AA-containing herb drugs consumption may account for some of these observations(1, 6-8). It has been shown that AA-induced mutational signature was characterized by A:T to T:A transversions at the sequence motif A[C│T]AGG(8, 9). The unusual genome-wide AA signature, termed signature 22 in COSMIC, holds great potential as "molecular fingerprints" for AA exposure in multiple cancer types(7, 9). Similarly, other endogenous or exogenous DNA damaging agents will also create mutational signatures detectable by DNA sequencing(10). Thus, analyses of mutational signatures, informative in other tumors(11, 12), have not been comprehensively reported in UTUC especially Chinese patients. The most comprehensive genomic studies of UTUC in Western populations used exome sequencing or targeted panel sequencing of cancer-associated genes(13-15). However, deriving signatures from whole-exome data may be problematic(16). Herein, we reported the first comprehensive genomic analysis of UTUC based on mutational signatures using WGS in 90 Chinese UTUC patients. Mutational signature analysis defined three UTUC subtypes with distinct etiologies and clinical outcomes. We further showed that geographically distinct AA subtype with high tumor mutation burden and higher level of tumor-infiltrating lymphocytes held great potential for immunotherapy. These results provided deeper insights into the biology of this cancer.

70   **Results**

71   **Copy number variations (CNVs) dominate the UTUC landscape**

72   The flowchart of selecting patients was shown in fig. S1**.** Clinicopathologic characteristics of 90

73   UTUC patients were listed in table S1. The samples were sequenced by WGS to an average 30X

74   depth which allowed us to comprehensively catalog both single nucleotide variations (SNVs),

75   small insertions and deletions (indels) and copy number variations (CNVs). We identified a

76   median of 19,639 (interquartile range [IQR], 16,578 to 32,937) SNVs and a median of 2,197 (IQR,

77   1,615 to 2,650) indels. A median of 437 (IQR, 355 to 584) coding mutations was noted (fig. 1A).

78   The increased mutation load was consistent with the increased prevalence of exposure to the potent

79   mutagen AAs (fig. 1B). Next, we examined the candidate driver genes with MutSigCV (Q<0.05).

80   Only two genes, TP53 and FRG2C, were identified (fig. 1C). However, many genes listed in the

81   Cancer Gene Census as known driver genes were affected by non-silent mutations including genes

82   which frequently mutated in Western UTUC patients, such as KMT2D (18%) and ARID1A (14%)

83   (fig. 1C). Additionally, the overall genomic landscape confirmed the dominance of recurrent

84   CNVs compared with SNVs and indels in the cohort (fig. 1D).

85   **Mutational signatures define three UTUC subtypes with distinct etiology**

86   Next, we explored the dynamic interplay of risk factors and cellular processes using mutational

87   signature analysis. We identified 23 mutational signatures defined by COSMIC in our cohort,

88   which were merged by shared aetiologies into 9 signatures by MutationalPatterns(17) (Materials

89   and Methods). Hierarchical clustering based on the number of SNVs attributable to each signature

90   confirmed three major subtypes (fig. 2A): (1) AA, there is significantly higher number of signature

91   22(1, 8) mutations (a median of 1,871) compared to the other two subtypes (fig. 2B, Wilcoxon

92   rank test); (2) Age, there is significantly higher number of signatures 1 and 5, attributed to clocklike

4

93   mutational processes accumulated over cell divisions(18) (fig. 2C, Wilcoxon rank test); and (3)

94   DSB, there is significantly higher number of signature 3 mutations, associated with failure of DNA

95   double-strand break-repair by homologous recombination (fig. 2D, Wilcoxon rank test). We

96   further verified whether the subtypes defined by mutational signatures associated with their

97   attributed etiological and clinicopathologic features. Consistent with previous epidemiological

98   studies in Asian patients(6, 19-21), we found that the AA subtype was significantly associated

99   with AA-containing herb drugs intake, poor renal function, female sex, multifocality and lower T

100   stage (fig. 2E, Kruskal-Wallis test). Thus, the AA subtype defined by mutational signatures did

101   capture the etiological subgroup of UTUC patients who had AAs exposure. Furthermore,

102   consistent with the scenario that DNA damage repair deficiency may increase genomic instability,

103   the DSB subtype exhibited both higher level weighted genome integrity index (wGII) and higher

104   level microsatellite instability (MSI) compared to the Age subtype (fig. 2F and 2G). Additionally,

105   lymphovascular invasion, an independent predictor of clinical outcomes for UTUC(22), was

106   higher in the Age subtype compared to the DSB subtype. Although statistically significance was

107   not achieved which may be due to limited number of cases (table S2).

108   **Subtypes defined by mutational signatures can predict clinical outcomes**

109   Next, we explored whether clinical outcomes of the three subtypes differed significantly. Notably,

110   a Kaplan-Meier plot revealed that the subtypes can predict both cancer-specific survival (CSS)

111   (fig. 3A, log-rank, p=0.006) and metastasis-free survival (MFS) (fig. 3B, log-rank, p=0.007). The

112   AA and DSB subtypes exhibited favourable outcomes compared with the Age subtype. Since an

113   improved prognostic stratification could offer more tailored therapeutic decisions, we further

114   stratified the patients into subgroups, muscle invasion and non-muscle invasion subgroups.

115   Consistently, the AA and DSB subtypes also exhibited favourable outcomes in muscle-invasive

116   UTUC patients (log-rank, *p*=0.001 for CSS, log-rank, *p*=0.002 for MFS) (fig. 3C and 3D). In

117   multivariate Cox regression proportional analyses adjusted for the effects of standard

118   clinicopathologic variables, mutational subtypes were still an independent risk factor for CSS

119   (Hazard Ratio [HR]=4.05, 95% CI: 1.58~10.42, *p*=0.004) and MFS (HR=3.43, 95% CI: 1.46~8.06,

120   p=0.006) (table S3). We also used the recently described mutational signature deconvolution

121   method, Mutalisk, for comparison(23). This method was based on different statistical frameworks,

122   and therefore some differences were to be expected. Nevertheless, we observed the same key

123   signature patterns with similar-sized patient subgroups expressing the dominant signature types

124   (fig. S2A). Both Mutalisk and MutationalPatterns identified the same group of patients with AA

125   signature. MutationalPatterns identified 9 putative Age subtype of UTUC patients that Mutalisk

126   did not identify (T002, T005, T013, T040, T047, T048, T087, T091 and T094) (fig. S2B). Since

127   the Age subtypes by either MutationalPatterns and Mutalisk showed the worst CSS and MFS (fig.

128   3, fig. S2C and S2D), we would propose that MutationalPatterns was preferable to err on the side

129   of overcalling rather than undercalling the presence of age-related signature.

Integration of mutational signature subtypes with copy number alterations

131   One practical question was how this subtyping could be implemented clinically. Previous study

132   had showed that 10X coverage data can confidently retrieve dominant signatures(11). Therefore,

133   lower-coverage WGS could provide a cost-effective alternative for signature-based stratification

134   especially for the AA subtype (more than 35% mutations contributed by signature 22 in our cohort).

135   Although for the Age and DSB subtype which did not have the dominant signatures, we next

136   explored whether CNV features can distinguish these two subtypes. First, we examined the

137   landscape of CNVs in UTUC by ichorCNA(24) which can predict CNVs in tumor-only samples.

138   Recurrent CNVs in UTUC patients of BCM/MDACC cohort were also identified in our study(13),

139   such as gain in 1q, 8q and loss in 2q, 5q, 10q (fig. S3A). However, the defined mutational subtypes

140   presented similarities in terms of CNV profiles (fig. S4A and S4B). Of note, several focal CNVs

141   were significantly associated with the mutational subtypes, such as deletion of 10q26.2-26.3 was

142   more prevalent in the DSB subtype compared to the Age subtype, and amplification of 1q31.1-

143   1q32.1 was more prevalent in the AA subtype compared to the DSB subtype (fig. S4C). Moreover,

144   we performed unsupervised hierarchical clustering which separated UTUCs into three groups with

145   distinct CNV patterns. However, there was not significantly association between mutational

146   signature subtypes and CNV subtypes (fig. S5A). Meanwhile, the overall survival of the CNV

147   subtypes did not differ significantly (fig. S5B and S5C). Among all the recurrent CNV regions,

148   only amplification of 17q25.1 was significantly associated with better overall survival (fig. S5D).

149   Hence, the CNV profiles seem to capture a different type of clinicopathological information

150   compared with the mutational signature subtypes.

151   **Field cancerization may contribute to malignant transformation especially for the AA**
152   **subtype**

153   Clinical interest in the AA subtype is increasing especially in East Asia since AA-associated

154   UTUCs are prevalent(1, 25). Consistent with our previous epidemiological study(21), we found

155   an increased rate of multifocality and high bladder recurrence in the AA subtype (fig. 2B). Field

156   cancerization(26), which is the development of a field with genetically altered cells, has been

157   proposed to explain the development of multiple primary tumours and local recurrence. Therefore,

158   we further sequenced three AA subtype patients (table S4) including a multifocal patient (fig. 4A

159   and fig. S6). We did find that significant amount of SNVs and indels in the morphologically normal

160   urothelium specimens in all three patients. Strikingly, the AA mutational signature was

161   consistently identified in urothelium specimens and tumor tissues in all 3 patients (fig. 4B), which

162   indicated that AA exposure may contribute to the field cancerization. Probably damaging

163   mutations in genes that were documented by the COSMIC and the KEGG pathway in cancer were

164   shown in table S5. Moreover, in the multifocal patient, the urothelial tumour at renal pelvic from

165   2007 shared no genetic alterations with a renal pelvic tumour from 2015 or a bladder tumour from

166   2015. However, the two tumours from 2015 were genetically related (fig. 4C). Therefore, field

167   cancerization and intraluminal seeding could co-contribute to the multifocality and increased

168   bladder recurrence in the AA subtype patients.

169   **Potential for immunotherapy in the AA subtype of UTUC patients**

170   The AA subtype bears high mutation burdens thus may be good candidate for immune checkpoint

171   blockade therapy(27). To address this possibility, we predicted neoantigens binding to patient-

172   specific human leukocyte antigen (HLA) types. The AA subtype had the highest number of

173   predicted neoantigens (fig. 5A). Moreover, it has been reported that lymphocytes infiltration

174   especially CD3$^+$ lymphocytes in tumour region is associated with improved survival in a range of

175   cancers including urothelial cancer(28-30). Previous study has shown that the number of tumor-

176   infiltrating lymphocyte independently correlates progress-free survival in NSCLC patients treated

177   with nivolumab immunotherapy(31). Pathological assessment of tumor-infiltrating lymphocytes

178   allows an easy evaluation of immune infiltration. Therefore, we further evaluated the extent of

179   tumor-infiltrating mononuclear cells (TIMCs) and CD3$^+$ lymphocytes in 76 available samples

180   (table S6). We found the number of CD3$^+$ lymphocytes was positively associated with the number

181   of stromal TIMCs (R^2=0.72; $p<0.001$) (fig. 5B). The AA subtype had the highest number of both

182   stromal TIMCs (Wilcoxon rank test, $p<0.001$) and CD3$^+$ lymphocytes (Wilcoxon rank, $p<0.001$)

183   (fig. 5C and 5D). Representative images from a patient of the AA subtype and the Age subtype

184   were shown in the fig. 5E and fig. 5F, respectively.

185
186   **Discussion**

187   Herein, we performed the first comprehensive genomic analysis of Chinese UTUC patients.

188   Three subtypes were defined by mutational signatures. We found that the AA subtype was

189   significantly associated with the usage of AA-containing herb drugs. However, a substantial

190   percentage of patients with the AA-mutational signature did not report usage of AA-containing

191   herb drugs. This finding may be due to the difficulty to track the dosage of AAs from various

192   herbal remedies. Some AA-containing herbal remedies have been officially prohibited in China

193   since 2003, but Aristolochia contains over 500 species, many of which have medicinal

194   properties(32). In the current study, we only took 70 AA-containing single products, such as

195   *Guang Fangchi*, *Qing Mu Xiang*, *Ma Dou Ling*, *Tian Xian Teng*, *Xun Gu Feng*, and *Zhu Sha Lian*

196   in Mandarin Chinese and mixed herbal formulas, such as, *Long Dan Xie Gan*, into account. And

197   the detail list of AA-containing herb drugs surveyed by current study was shown in table S7. Some

198   herb plants containing AA were not banned such as *Xi Xin(33)*. Exposure to AAs and their

199   derivatives may still be widespread in China. Notably, we showed AA-mutational signature was

200   also identified in histologically "normal" urothelial cells. In a published study(34), similar "field

201   effect" was also identified in an Chinese UTUC patient. Thus, non-invasive urine test based on the

202   AA-mutational signature could take advantage of this "field effect" to screen individuals at

203   increased risk of recurrence due to exposing to herbal remedies containing the carcinogen AA.

204   One limitation of current study was lack of a matched normal for 85 of 90 UTUC patients due

205   to lack of funds. Therefore, we relied on the public mutation database and in-house genomic data

206   of 1000 healthy Chinese individuals for the filtering germline variances. But in recent years, these

207   catalogs of human variation have grown exponentially, questioning the necessity of sequencing a

208   matched normal for every tumor sample(35). Additionally, a matched normal will not be available

209   in many pathological conditions, such as cancer cell lines and liquid biopsy. Although we can not

9

210   make definite conclusions that the SNVs identified in current study were somatic mutations. We

211   may also overcall the Age subtypes without a matched normal since germline variants contributed

212   to signature 5 in COSMIC which combined with signature 1 as age-related signature in current

213   study. However, with this caveat, the mutational signature subtypes did capture distinct clinical

214   and genomic features, such as better overall survival, higher level of wGII and higher level of MSI

215   in DSB subtype compared to Age subtype of patients.

216   Additionally, our study supports a rationale for management of UTUC, most notably the AA

217   subtype. Patients with AA mutational signature may predict better checkpoint inhibitors response.

218   Strikingly, the DSB subtype was featured by signature 3 in COSMIC. In pancreatic cancer,

219   responders to platinum therapy usually exhibit signature 3 mutations(36). These results suggested

220   that platinum-based therapy may also benefit UTUC patients of DSB subtype. More recently,

221   protein components of the DNA damage repair (DDR) systems have been identified as promising

222   avenues for targeted cancer therapeutics(37). We also further assessed the mutation rates across

223   345 genes associated with DDR, as previously described in a pan-cancer analysis(37) (fig. S7).

224   There was a 1.6 fold enrichment (*P*=0.007, 95% CI:1.12-2.20) and a 1.4 fold enrichment (*P*=0.01,

225   95% CI:1.07-1.86) of DDR genes alterations in DSB and Age subtype, respectively. Further

226   mechanistic insights of the DSB and Age subtypes may provide the basis for therapeutic

227   opportunities within the DNA damage response. In summary, our study provides the most

228   comprehensive genomic profile of Chinese UTUC patients to date. The findings may accelerate

229   the development of novel prognostic markers and personalized therapeutic approaches for UTUC

230   patients especially those from geographical regions with exposure to AA-containing herb drugs.

231

232   **Materials and Methods**

10

Patients cohort

All the fresh samples in this study were obtained from Peking University First Hospital between January 2016 to December 2017 (Grant No.2015[977]). These fresh samples were stored in liquid nitrogen after surgery immediately. Formalin-fixed, paraffin-embedded (FFPE) samples were provided by the Institute of Urology after pathologic diagnosis. The main endpoint events consisted of overall survival (OS), cancer-specific survival (CSS), metastasis-free survival (MFS), progression-free survival (PFS) and bladder-recurrence free survival (BRFS). All patients were not treated with neoadjuvant chemotherapy. AA exposure assessment was performed according to self-reported data on 70 AA and its derivatives-containing herb drugs (collectively, AA) intake(21, 38). These herbs were taken as single products (*Guan Mu Tong* (*Aristolochia manshuriensis Kom*), *Guang Fangchi* (*Aristolochia fangchi*), *Qing Mu Xiang* (*Radix Aristolochiae*), *Ma Dou Ling* (*Fructus Aristolochiae*), *Tian Xian Teng* (*Caulis Aristolochiae*), *Xun Gu Feng* (*herba Aristolochiae mollissimae*), and *Zhu Sha Lian* (*Aristolochia cinnabarina*)) or were components of mixed herbal formulas (eg, *Guan Mu Tong* in the *Long Dan Xie Gan* mixture). And the detail list of AA-containing herb drugs surveyed by current study was shown in table S7. The accumulated self-reported usage for the above drugs more than a year was termed as AA exposure patients. Clinical and demographic information was obtained from a prospectively maintained institutional database. The study was approved by the Ethics Committee of Peking University First Hospital.

Whole-genome sequencing

The sheared DNA was repaired and 3´dA-tailed using the NEBNext Ultra II End Repair/dA-Tailing Module unit and then ligated to paired-end (PE) adaptors using the NEBNext Ultra II Ligation Module unit. After purification by AMPure XP beads, the DNA fragments were amplified

255   by PCR for 6-8 cycles. Finally, the libraries were sequenced with an Illumina X-ten instrument,

256   thus generating 2 x 150-bp PE reads.

Single nucleotide variations (SNVs) and indels calling

258   SNVs and indels were called using VarScan2(39)and Vardict(40) in the form of single sample

259   mode, then Rtg-tools(41) was used to remove variants called in a set of 1000 healthy, unrelated

260   Chinese individuals (unpublished data from CAS Precision Medicine Initiative) and got the

261   common variants called by the two software. Further filtering criteria were carried out according

262   to the reference(11, 42). Finally, Annovar(43) was applied to remove variants whose mutation

263   frequency was no less than 0.001 in 1000 Genomes project, latest Exome Aggregation Consortium

264   dataset, NHLBI-ESP project with 6500 exomes, latest Haplotype Reference Consortium database

265   and latest Kaviar database.

Mutational signature analysis

267   R package MutationalPatterns(17) was carried out to assign each sample' signature. We discovered

268   23 COSMIC signatures which were merged by shared etiologies into 9 signatures in our cohort.

269   We named signature 22 which has been found in cancer samples with known exposures to

270   aristolochic acid(1, 8) as AA, DNA double strand break-repair deficiency featured by signature 3

271   as DSB. We merged signature 1 and 5 as Age(18), signature 2, 13 as APOBEC, combined

272   signatures including signature 6, 15, 20, 21, 26, 27 as MSI, combined signatures including

273   signature 8, 12, 16, 17, 18, 19, 23, 25, 28, 30 as unknown. Mutational signatures was extracted

274   from the mutation count matrix, with non-negative matrix factorization (NMF). Hierarchical

275   clustering was performed by the number of single nucleotide variants (SNVs) attributable to each

276   signature.

Copy number variations (CNVs) analysis

278   To determine copy number states, we used a recently published method(44). Briefly, the genome

279   was divided into bins of variable length with an equal amount of potential uniquely mapping reads.

280   We simulated approximately 37X hg19 reads mapped by BWA and divided the genome into

281   50,000 bins of the same mapping ability. Loess normalization was used to correct for GC bias.

282   **Candidate driver gene analysis**

283   Candidate driver genes were explored by MutSigCV(45) to analyse SNVs and indels. Chromatin

284   state, DNA replication time, expression level and the first 20 principal components of 169

285   chromatin marks(46) from the RoadMap Epigenomics Project(47) of the mutated gene were used

286   as covariates. We applied the cut off 0.05 for Q value. Cancer census genes from COSMIC(48)

287   release v87 were calculated for their mutation rates (mutations/gene length) and genes with more

288   than 10 samples were selected with high mutation rates.

289   **The weighted genome integrity index score (wGII) analysis**

290   We calculated the wGII score as reported(49). Briefly, the percentage of gained and lost genomic

291   material was calculated relative to the ploidy of the sample. The use of percentages eliminates the

292   bias induced by differing chromosome sizes. The wGII score of a sample was defined as the

293   average of this percentage value over the 22 autosomal chromosomes.

294   **Microsatellite instability (MSI) analysis**

295   mSINGS(50) was used to calculate the ratio of unstable loci in the genome of each sample. First,

296   the microsatellites at the reference genome were calculated by MSIsensor(51). Sixteen

297   microsatellite stable esophageal squamous cell carcinomas(52) were used to calculate the baseline

298   for the absence of MSI signature (signature 6, 15, 20, 21, 26, 27) as defined by COSMIC(48).

299   **Neoantigen prediction**

300 HLAscan(53) was used to genotype the HLA region with HLA-A, HLA-B and HLA-C taken into

301 consideration. NetMHC4.0(54) was used for predictions of peptide-MHC class I interaction, SNVs

302 annotated by Annovar(43) as nonsynonymous were used to do this analysis. An in-house script

303 was carried out to obtain possible 9–amino acid sequences covering the mutated amino acids

304 according to the manual. Rank in the output that was no more than two were considered as binding

305 in the light of BindLevel suggestion.

306 **Assessing tumor-infiltrating lymphocytes**

307 The tumor infiltrating mononuclear lymphocytes were measured according to a standardized

308 method from the International Immuno-Oncology Biomarkers Working Group(55). And the CD3

309 antibody (ab5690, 1:10000; Abcam) was used to evaluate the $CD3^+$ lymphocytes in tumor section.

320 **Disclosure:** Weimin Ci certifies that there are no conflicts of interest.

321 **Data availability**

322   The raw sequence data reported in this paper have been deposited in the Genome Sequence

323   Archive(56) in BIG Data Center(57), Beijing Institute of Genomics (BIG), Chinese Academy of

324   Sciences (accession numbers HRA000029). That can be accessed at http://bigd.big.ac.cn/gsa-

325   human/s/HiObV4f3.

326
327
328   1.   Chen CH, Dickman KG, Moriya M, Zavadil J, Sidorenko VS, Edwards KL, et al. Aristolochic acid-
329   associated urothelial cancer in Taiwan. Proc Natl Acad Sci U S A. 2012;109(21):8241-6.
330   2.   Yang MH, Chen KK, Yen CC, Wang WS, Chang YH, Huang WJ, et al. Unusually high incidence of
331   upper urinary tract urothelial carcinoma in Taiwan. Urology. 2002;59(5):681-7.
332   3.   Hsiao PJ, Hsieh PF, Chang CH, Wu HC, Yang CR, Huang CP. Higher risk of urothelial carcinoma in the
333   upper urinary tract than in the urinary bladder in hemodialysis patients. Ren Fail. 2016;38(5):663-70.
334   4.   Roupret M, Babjuk M, Comperat E, Zigeuner R, Sylvester RJ, Burger M, et al. European Association of
335   Urology Guidelines on Upper Urinary Tract Urothelial Carcinoma: 2017 Update. European urology.
336   2018;73(1):111-22.
337   5.   Singla N, Fang D, Su X, Bao Z, Cao Z, Jafri SM, et al. A Multi-Institutional Comparison of
338   Clinicopathological Characteristics and Oncologic Outcomes of Upper Tract Urothelial Carcinoma in China
339   and the United States. J Urol. 2017;197(5):1208-13.
340   6.   Grollman AP. Aristolochic acid nephropathy: Harbinger of a global iatrogenic disease. Environmental
341   and molecular mutagenesis. 2013;54(1):1-7.
342   7.   Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P, et al. Genome-wide mutational
343   signatures of aristolochic acid and its application as a screening tool. Sci Transl Med. 2013;5(197):197ra01.
344   8.   Hoang ML, Chen CH, Sidorenko VS, He J, Dickman KG, Yun BH, et al. Mutational signature of
345   aristolochic acid exposure as revealed by whole-exome sequencing. Sci Transl Med. 2013;5(197):197ra02.
346   9.   Ng AWT, Poon SL, Huang MN, Lim JQ, Boot A, Yu W, et al. Aristolochic acids and their derivatives
347   are widely implicated in liver cancers in Taiwan and throughout Asia. Science translational medicine.
348   2017;9(412).
349   10.   Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers.
350   Nature reviews Genetics. 2014;15(9):585-98.
351   11.   Secrier M, Li X, de Silva N, Eldridge MD, Contino G, Bornschein J, et al. Mutational signatures in
352   esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. Nat Genet.
353   2016;48(10):1131-41.
354   12.   Schulze K, Imbeaud S, Letouze E, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing
355   of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat
356   Genet. 2015;47(5):505-11.
357   13.   Moss TJ, Qi Y, Xi L, Peng B, Kim TB, Ezzedine NE, et al. Comprehensive Genomic Characterization of
358   Upper Tract Urothelial Carcinoma. European urology. 2017.
359   14.   Audenet F, Isharwal S, Cha EK, Donoghue MTA, Drill EN, Ostrovnaya I, et al. Clonal Relatedness and
360   Mutational Differences between Upper Tract and Bladder Urothelial Carcinoma. Clinical cancer research : an
361   official journal of the American Association for Cancer Research. 2018.
362   15.   Moss TJ, Qi Y, Xi L, Peng B, Kim TB, Ezzedine NE, et al. Comprehensive Genomic Characterization of
363   Upper Tract Urothelial Carcinoma. European urology. 2017;72(4):641-9.
364   16.   Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of

365   mutational processes in human cancer. Nature. 2013;500(7463):415-21.

366   17.   Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide
367   analysis of mutational processes. Genome Med. 2018;10(1):33.

368   18.   Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational
369   processes in human somatic cells. Nat Genet. 2015;47(12):1402-7.

370   19.   Grollman AP, Shibutani S, Moriya M, Miller F, Wu L, Moll U, et al. Aristolochic acid and the etiology of
371   endemic (Balkan) nephropathy. Proc Natl Acad Sci U S A. 2007;104(29):12129-34.

372   20.   Yang HY, Wang JD, Lo TC, Chen PC. Increased risks of upper tract urothelial carcinoma in male and
373   female chinese herbalists. Journal of the Formosan Medical Association = Taiwan yi zhi. 2011;110(3):161-8.

374   21.   Zhong W, Zhang L, Ma J, Shao S, Lin R, Li X, et al. Impact of aristolochic acid exposure on oncologic
375   outcomes of upper tract urothelial carcinoma after radical nephroureterectomy. OncoTargets and therapy.
376   2017;10:5775-82.

377   22.   Kikuchi E, Margulis V, Karakiewicz PI, Roscigno M, Mikami S, Lotan Y, et al. Lymphovascular invasion
378   predicts clinical outcomes in patients with node-negative upper tract urothelial carcinoma. Journal of clinical
379   oncology : official journal of the American Society of Clinical Oncology. 2009;27(4):612-8.

380   23.   Lee J, Lee AJ, Lee JK, Park J, Kwon Y, Park S, et al. Mutalisk: a web-based somatic MUTation AnaLysIS
381   toolKit for genomic, transcriptional and epigenomic signatures. Nucleic Acids Res. 2018;46(W1):W102-W8.

382   24.   Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-
383   exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun.
384   2017;8(1):1324.

385   25.   Colin P, Koenig P, Ouzzane A, Berthon N, Villers A, Biserte J, et al. Environmental factors involved in
386   carcinogenesis of urothelial cell carcinomas of the upper urinary tract. BJU Int. 2009;104(10):1436-40.

387   26.   Vanharanta S, Massague J. Field cancerization: something new under the sun. Cell.
388   2012;149(6):1179-81.

389   27.   Yarchoan M, Hopkins A, Jaffee EM. Tumor Mutational Burden and Response Rate to PD-1 Inhibition.
390   The New England journal of medicine. 2017;377(25):2500-1.

391   28.   Huh JW, Lee JH, Kim HR. Prognostic significance of tumor-infiltrating lymphocytes for patients with
392   colorectal cancer. Archives of surgery. 2012;147(4):366-72.

393   29.   Thomas NE, Busam KJ, From L, Kricker A, Armstrong BK, Anton-Culver H, et al. Tumor-infiltrating
394   lymphocyte grade in primary melanomas is independently associated with melanoma-specific survival in the
395   population-based genes, environment and melanoma study. Journal of clinical oncology : official journal of
396   the American Society of Clinical Oncology. 2013;31(33):4252-9.

397   30.   Lipponen PK, Eskelinen MJ, Jauhiainen K, Harju E, Terho R. Tumour infiltrating lymphocytes as an
398   independent prognostic factor in transitional cell bladder cancer. European journal of cancer.
399   1992;29A(1):69-75.

400   31.   Gataa I ML, Auclin E, Le Moulec S, Alemany P, Kossai M, Massé J , CARAMELLA C, Remon Masip J,
401   Lahmar J, Ferrara R, Gazzah A, Soria JC, Planchard D, Besse B, Adam J. . 112P Pathological evaluation of
402   tumor infiltrating lymphocytes and the benefit of nivolumab in advanced non-small cell lung cancer (NSCLC).
403   Annals of Oncology. 2017;28(suppl_5).

404   32.   KreLL D SJ. Aristolochia: the malignant truth. Lancet Oncology. 2013;14(1):25-6.

405   33.   Hsieh SC, Lin IH, Tseng WL, Lee CH, Wang JD. Prescription profile of potentially aristolochic acid
406   containing Chinese herbal products: an analysis of National Health Insurance data in Taiwan between 1997
407   and 2003. Chinese medicine. 2008;3:13.

408   34.   Du Y, Li R, Chen Z, Wang X, Xu T, Bai F. Mutagenic Factors and Complex Clonal Relationship of
409   Multifocal Urothelial Cell Carcinoma. European urology. 2016.

410   35.   Kumar A, White TA, MacKenzie AP, Clegg N, Lee C, Dumpit RF, et al. Exome sequencing identifies a
411   spectrum of mutation frequencies in advanced and lethal prostate cancers. Proceedings of the National

412   Academy of Sciences of the United States of America. 2011;108(41):17087-92.

413   36.   Greer JB, Whitcomb DC. Role of BRCA1 and BRCA2 mutations in pancreatic cancer. Gut.
414   2007;56(5):601-5.

415   37.   Pearl LH, Schierz AC, Ward SE, Al-Lazikani B, Pearl FM. Therapeutic opportunities within the DNA
416   damage response. Nat Rev Cancer. 2015;15(3):166-80.

417   38.   Li WH, Yang L, Su T, Song Y, Li XM. [Influence of taking aristolochic acid-containing Chinese drugs
418   on occurrence of urinary transitional cell cancer in uremic uremic patients undergoing dialysis]. Zhonghua yi
419   xue za zhi. 2005;85(35):2487-91.

420   39.   Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and
421   copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-76.

422   40.   Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and
423   versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res.
424   2016;44(11):e108.

425   41.   Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, et al. 2015.

426   42.   Li H. Toward better understanding of artifacts in variant calling from high-coverage samples.
427   Bioinformatics. 2014;30(20):2843-51.

428   43.   Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-
429   throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

430   44.   Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, et al. Genome-wide copy number
431   analysis of single cells. Nat Protoc. 2012;7(6):1024-41.

432   45.   Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational
433   heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-8.

434   46.   Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of
435   Selection in Cancer and Somatic Tissues. Cell. 2017;171(5):1029-41.e21.

436   47.   Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of
437   111 reference human epigenomes. Nature. 2015;518(7539):317-30.

438   48.   Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of
439   Somatic Mutations In Cancer. Nucleic acids research. 2018.

440   49.   Endesfelder D, Burrell RA, Kanu N, McGranahan N, Howell M, Parker PJ, et al. Chromosomal
441   Instability Selects Gene Copy-Number Variants Encoding Core Regulators of Proliferation in ER+Breast
442   Cancer. Cancer Research. 2014;74(17):4853-63.

443   50.   Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC. Microsatellite instability detection
444   by next generation sequencing. Clin Chem. 2014;60(9):1192-9.

445   51.   Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, et al. MSIsensor: microsatellite instability detection
446   using paired tumor-normal sequence data. Bioinformatics. 2014;30(7):1015-6.

447   52.   Song Y, Li L, Ou Y, Gao Z, Li E, Li X, et al. Identification of genomic alterations in oesophageal
448   squamous cell cancer. Nature. 2014;509(7498):91-5.

449   53.   Ka S, Lee S, Hong J, Cho Y, Sung J, Kim HN, et al. HLAscan: genotyping of the HLA region using
450   next-generation sequencing data. BMC Bioinformatics. 2017;18(1):258.

451   54.   Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-
452   MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol.
453   2017;199(9):3360-8.

454   55.   Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, et al. Assessing Tumor-Infiltrating
455   Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method
456   from the International Immuno-Oncology Biomarkers Working Group: Part 2: TILs in Melanoma,
457   Gastrointestinal Tract Carcinomas, Non-Small Cell Lung Carcinoma and Mesothelioma, Endometrial and
458   Ovarian Carcinomas, Squamous Cell Carcinoma of the Head and Neck, Genitourinary Carcinomas, and

459   Primary Brain Tumors. Advances in anatomic pathology. 2017;24(6):311-35.

460   56.   Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: Genome Sequence Archive<sup/>.

461   Genomics, proteomics & bioinformatics. 2017;15(1):14-8.

462   57.   Members BIGDC. Database Resources of the BIG Data Center in 2018. Nucleic Acids Res.

463   2018;46(D1):D14-D20.

464

465

466
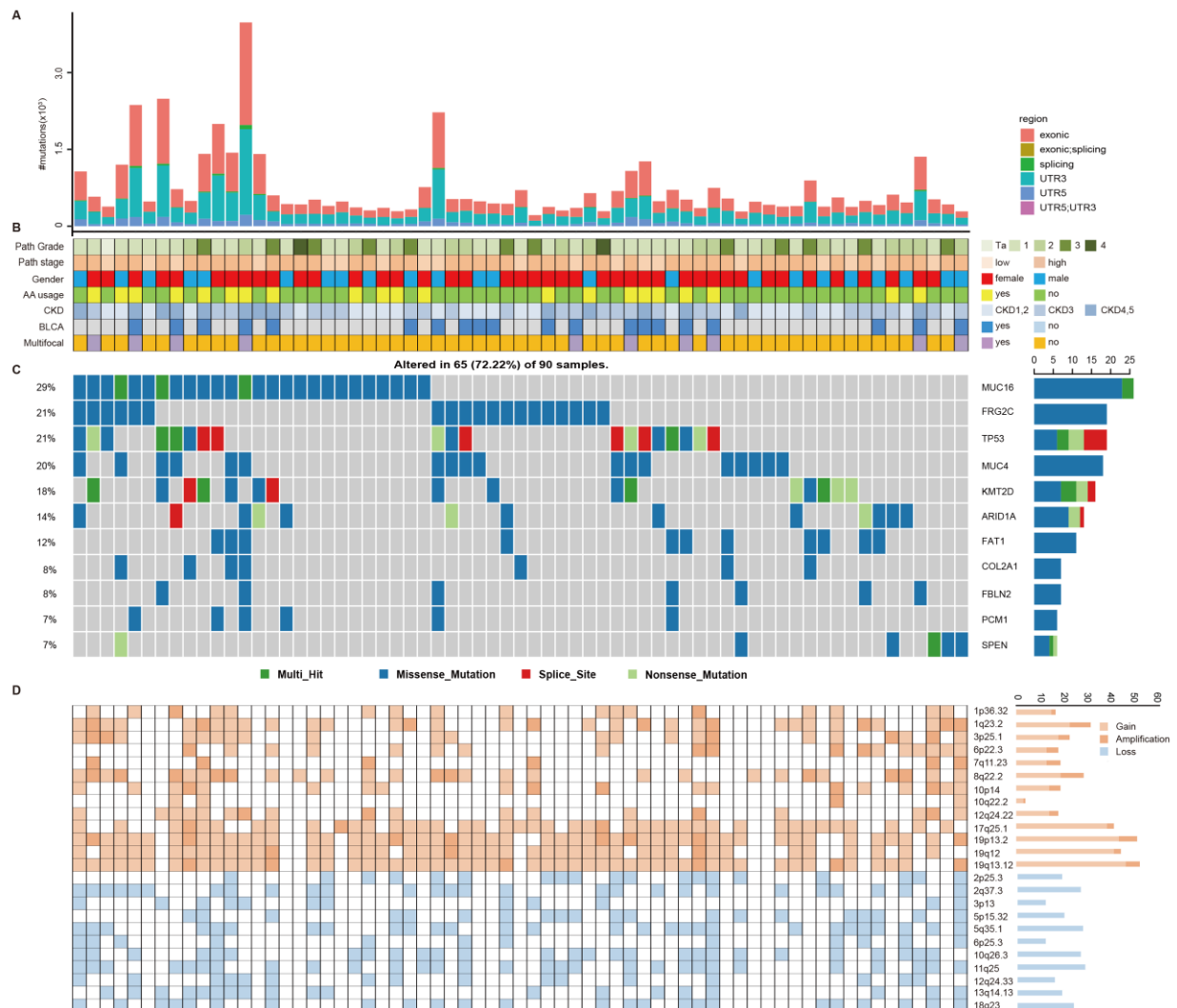
467

Fig. 1. Oncoplot of genomic alterations in UTUC. (A) The number of coding mutations were represented by barplots. **(B)** Selected patient clinical features were included. **(C)** Cancer consensus frequently mutated genes and mutation types were indicated in the legend. **(D)** Selected loci with high recurrence rate among patients were shown. Copy number gain/loss and amplification/deletion were inferred as genes with GISTIC scores of 1 and 2, respectively. BLCA: synchronous bladder cancer, bladder recurrence and bladder cancer history. AA: aristolochic acid. CKD: chronic kidney disease.

475

476
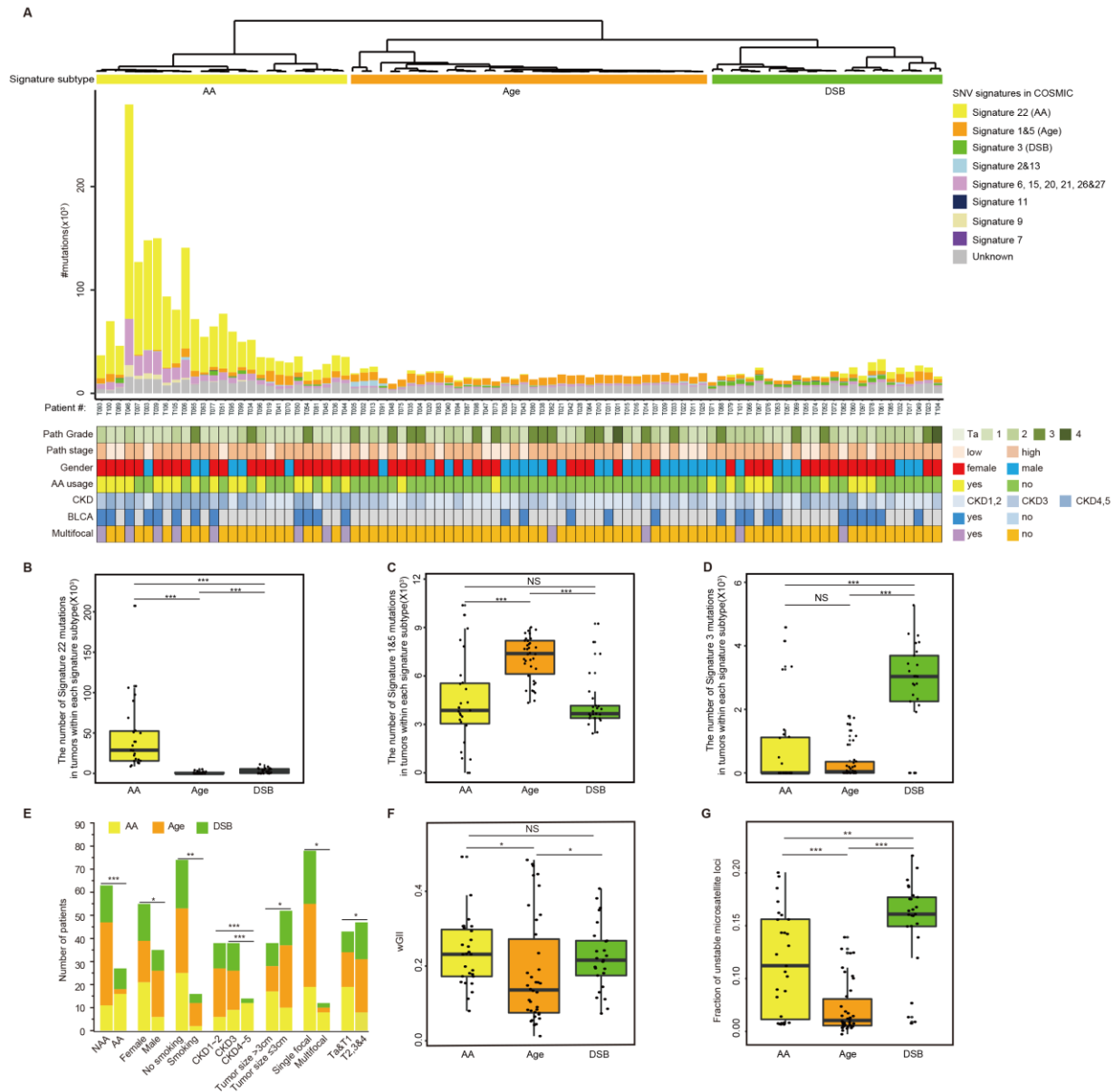
477
Fig. 2. Genomic subtypes defined by mutational signatures. (A) Bar plot of the number of
SNVs attributable to 9 merged signatures in each of the 90 tumors, sorted by hierarchical clustering
(dendrogram at bottom), revealing AA-related (AA, yellow), Age-related (Age, orange), and DNA
damage repair deficient (DSB, green). Selected clinical features were represented in the bottom
tracks. **(B)** The box plot showed the number of signature 22 mutations in tumors within each
signature subtype. $*P<0.05$; $**P<0.01$; $***P< 0.001$ (Wilcoxon rank test). **(C)** The box plot
showed the number of signatures 1&5 mutations in tumors within each signature subtype. $*P<0.05$;
$**P<0.01$; $***P< 0.001$ (Wilcoxon rank test). **(D)** The box plot showed the number of signature 3
mutations in tumors within each signature subtype. $*P<0.05$; $**P<0.01$; $***P< 0.001$ (Wilcoxon
rank test). **(E)** The bar graph showed the association between three subtypes and clinicopathologic
features. $*P<0.05$; $**P<0.01$; $***P< 0.001$ (Kruskal-Wallis test). **(F)** Comparison of the weighted
Genome Instability Index (wGII) scores among three subtypes. **(G)** Microsatellite instability (MSI)
status was evaluated by fraction of unstable microsatellite loci identified by WGS data. P-values

19

491   were calculated by Wilcoxon rank test (*P<0.05, **P< 0.01, ***P< 0.001).
492
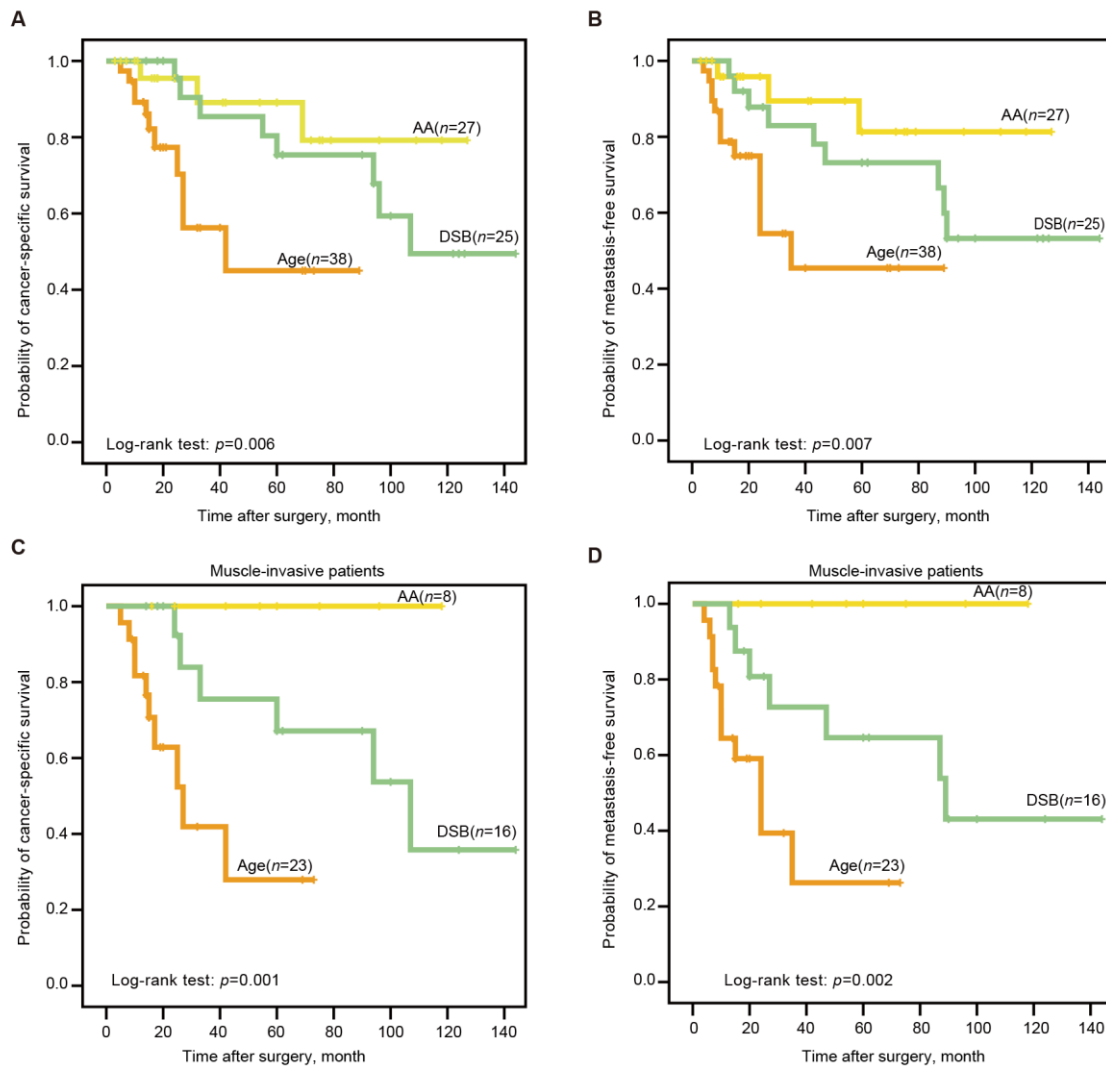
493
494   **Fig. 3. Subtypes defined by mutational signatures predict clinical outcome. (A)-(D)**Kaplan-
495   Meier survival curves showed the mutational signature subtypes can predict both CSS and MFS
496   for the whole cohort, as well as for muscle-invasive UTUC patients. CSS: cancer-specific survival.
497   MFS: metastasis-free survival. *P*-values were calculated by the log-rank test. *n*, the number of
498   cases.
499
500

501

502
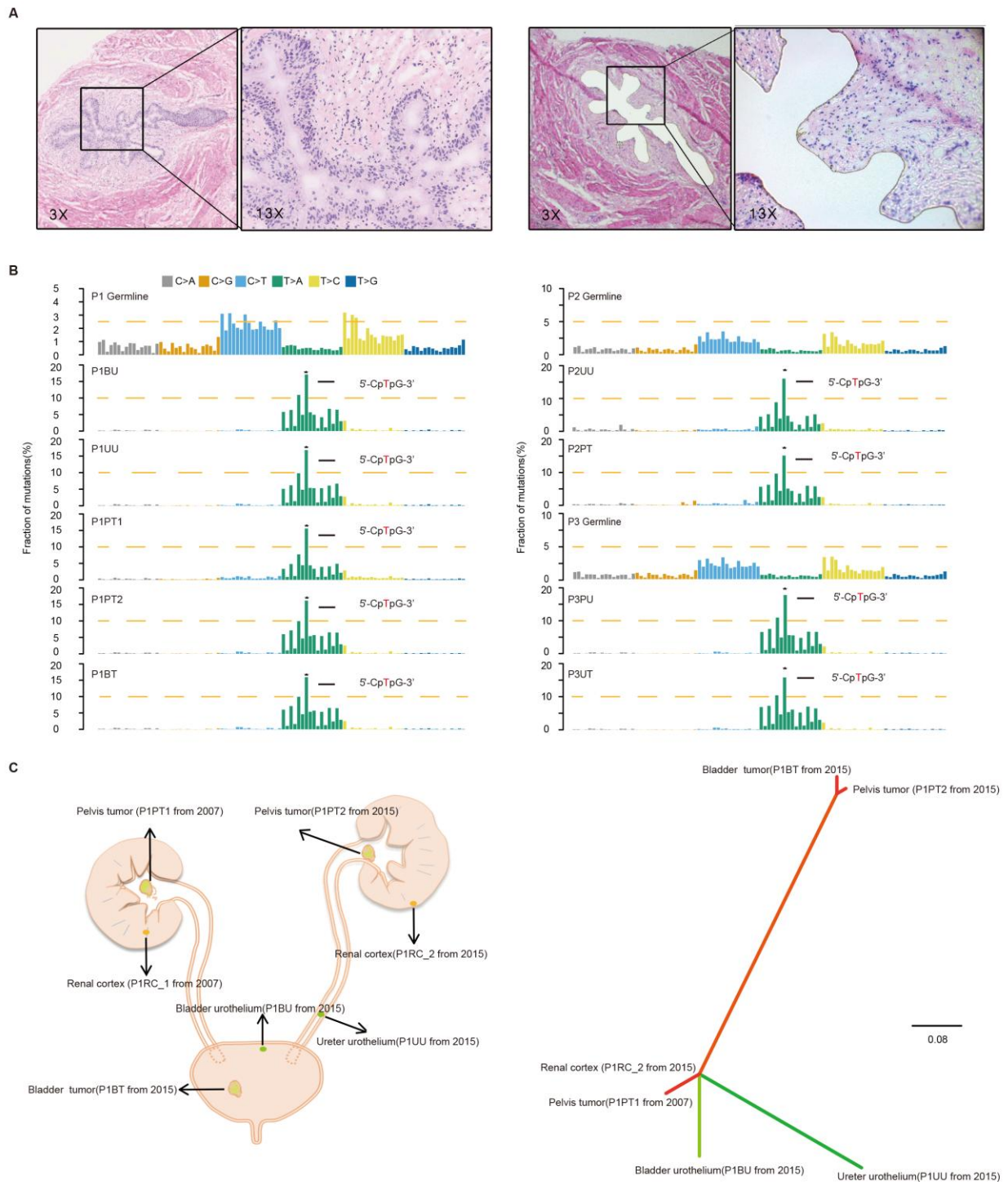
Fig. 4. Field cancerization may contribute to malignant transformation especially for the AA subtype. (A) Laser capture microdissection of urothelium from a representative patient. Magnified insets (13x) of the black frame area showed ureteral urothelium before laser capture and the remaining cells after laser capture. **(B)** Trinucleotide contexts for somatic mutations in biopsies from three patients of the AA subtype. The height of each bar (the *y*-axis) represented the proportion of all observed mutations that fall into a particular trinucleotide mutational class. AA:

509 aristolochic acid; UU: ureter urothelium; BU: bladder urothelium; RC: renal cortex; BT: bladder
510 tumor; UT: urothelium tumor; and PT: pelvis tumour. **(C)** Phylogenetic relationships of the six
511 samples from the multifocal patient were deciphered using mrbayes_3.2.2. Spatial locations of
512 core biopsies of the patient were presented in the left panel.

513
514
515
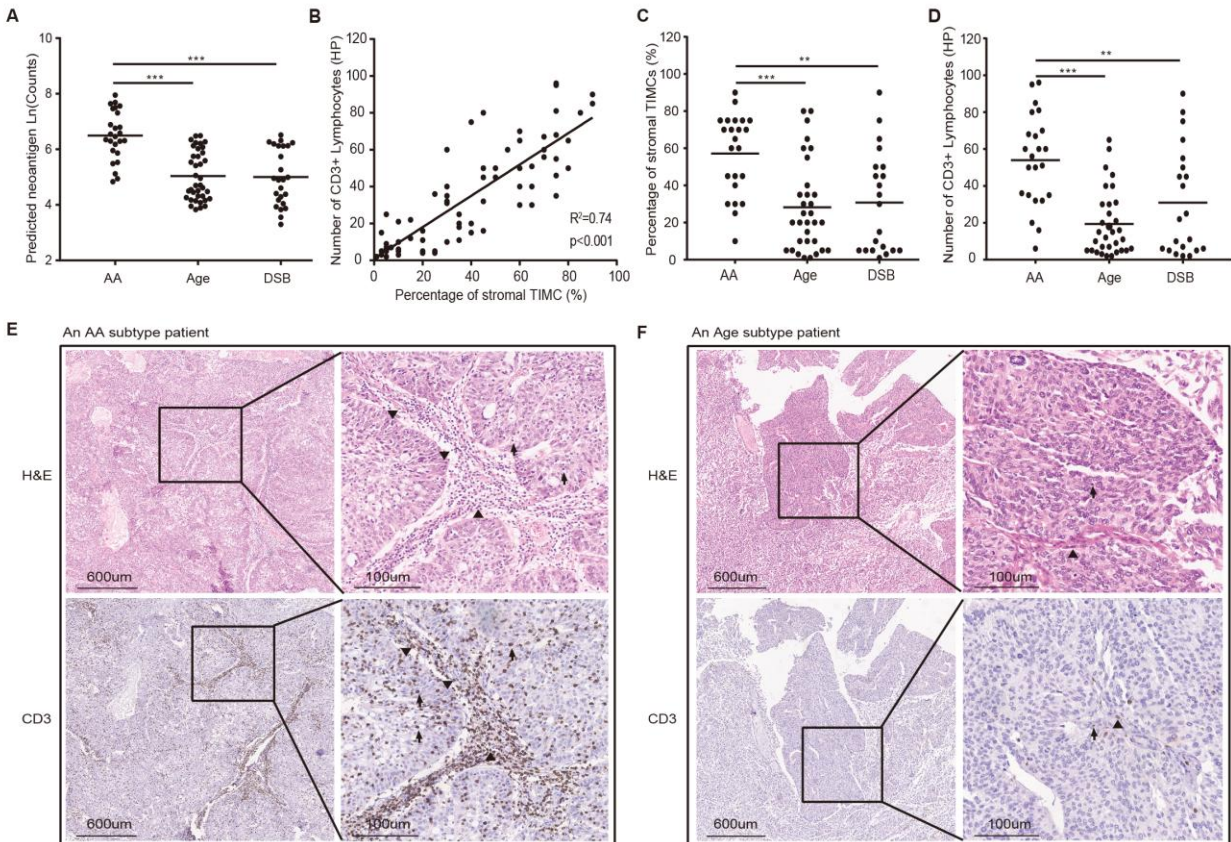
516
517 **Fig. 5. Predicted neoantigens and tumor-infiltrating lymphocytes in each mutational subtype.**
518 **(A)** Neoantigen burden was significantly higher in the AA subtype. Statistical significance was
519 determined by Wilcoxon rank test (***$P<0.001$). **(B)** Positive correlation of the percentage of
520 stromal tumor-infiltrating mononuclear cells (TIMCs) and the number of CD3$^+$ lymphocytes in 76
521 UTUC patients of our cohort (n_{AA}=23; n_{Age}=32; n_{DSB}=21). The percentage of stromal TIMCs **(C)**
522 and CD3$^+$ lymphocytes **(D)** were shown in each subtype of patients. Statistical significance was
523 determined by the Wilcoxon rank test (**$P<0.01$, ***$P<0.001$). HP represented high-power field.
524 Images of TIMCs and CD3$^+$ lymphocytes of a representative patient from the AA subtype **(E)** and
525 the Age subtype **(F)** Triangle highlighted the TIMCs or CD3$^+$ lymphocytes in stromal tumour
526 region. Arrow highlighted the TIMCs or CD3$^+$ lymphocytes in intra-tumour region.

527
528