Geographic variation of mutagenic exposures in kidney cancer genomes

Sergey Senkin[1], Sarah Moody[2], Marcos Díaz-Gay[3,4,5], Behnoush Abedi-Ardekani[1], Thomas Cattiaux[1], Aida Ferreiro-Iglesias[1], Jingwei Wang[2], Stephen Fitzgerald[2], Mariya Kazachkova[3,6,5], Raviteja Vangara[3,4,5], Anh Phuong Le[2], Erik N. Bergstrom[3,4,5], Azhar Khandekar[3,4,5], Burçak Otlu[3,4,5,7], Saamin Cheema[2], Calli Latimer[2], Emily Thomas[2], Joshua Ronald Atkins[8], Karl Smith-Byrne[8], Ricardo Cortez Cardoso Penha[1], Christine Carreira[9], Priscilia Chopard[1], Valérie Gaborieau[1], Pekka Keski-Rahkonen[10], David Jones[2], Jon W. Teague[2], Sophie Ferlicot[11], Mojgan Asgari[12], Surasak Sangkhathat[13], Worapat Attawettayanon[14], Beata Świątkowska[15], Sonata Jarmalaite[16,17], Rasa Sabaliauskaite[16], Tatsuhiro Shibata[18,19], Akihiko Fukagawa[19,20], Dana Mates[21], Viorel Jinga[22], Stefan Rascu[22], Mirjana Mijuskovic[23], Slavisa Savic[24], Sasa Milosavljevic[25], John M.S. Bartlett[26], Monique Albert[27,28], Larry Phouthavongsy[28], Patricia Ashton-Prolla[29,30], Mariana R. Botton[31], Brasil Silva Neto[32,33], Stephania Martins Bezerra[34], Maria Paula Curado[35], Stênio de Cássio Zequi[36,37,38,39], Rui Manuel Reis[40,41], Eliney Faria[42], Nei Soares Menezes[43], Renata Spagnoli Ferrari[42], Rosamonde E. Banks[44], Naveen S. Vasudev[44], David Zaridze[45], Anush Mukeriya[45], Oxana Shangina[45], Vsevolod Matveev[46], Lenka Foretova[47], Marie Navratilova[47], Ivana Holcatova[48,49], Anna Hornakova[50], Vladimir Janout[51], Mark Purdue[52], Nathaniel Rothman[52], Stephen J. Chanock[52], Per Magne Ueland[53], Mattias Johansson[1], James McKay[1], Ghislaine Scelo[54], Estelle Chanudet[55], Laura Humphreys[2], Ana Carolina de Carvalho[1], Sandra Perdomo[1], Ludmil B. Alexandrov[3,4,5], Michael R. Stratton[2], Paul Brennan[1]*

[1]Genomic Epidemiology Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France, [2]Cancer, Ageing and Somatic Mutation, Wellcome Sanger Institute, Cambridge, UK, [3]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, USA, [4]Department of Bioengineering, University of California San Diego, La Jolla, USA, [5]Moores Cancer Center, University of California San Diego, La Jolla,

28 USA, [6]Biomedical Sciences Graduate Program, University of California San Diego, La Jolla,
29 USA, [7]Department of Health Informatics, Graduate School of Informatics, Middle East
30 Technical University, Ankara, Turkey, [8]Cancer Epidemiology Unit, The Nuffield Department
31 of Population Health, University of Oxford, Oxford, UK, [9]Evidence Synthesis and
32 Classification Branch, International Agency for Research on Cancer (IARC/WHO), Lyon,
33 France, [10]Nutrition and Metabolism Branch, International Agency for Research on Cancer
34 (IARC/WHO), Lyon, France, [11]Service d'Anatomie Pathologique, Assistance Publique-
35 Hôpitaux de Paris, Univeristé Paris-Saclay, Le Kremlin-Bicêtre, France, [12]Oncopathology
36 Research Center, Iran University of Medical Sciences, Tehran, Iran, [13]Translational Medicine
37 Research Center, Faculty of Medicine, Prince of Songkla University, Hat Yai,
38 Thailand, [14]Department of Surgery, Urology, Faculty of Medicine, Prince of Songkla
39 University, Hat Yai, Thailand, [15]Department of Environmental Epidemiology, Nofer Institute of
40 Occupational Medicine, Łódź, Poland, [16]Laboratory of Genetic Diagnostic, National Cancer
41 Institute, Vilnius, Lithuania, [17]Department of Botany and Genetics, Institute of Biosciences,
42 Vilnius University, Vilnius, Lithuania, [18]Laboratory of Molecular Medicine, The Institute of
43 Medical Science, The University of Tokyo, Minato-ku, Japan, [19]Division of Cancer Genomics,
44 National Cancer Center Research Institute, Chuo-ku, Japan, [20]Department of Pathology,
45 Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Japan, [21]Occupational
46 Health and Toxicology, National Center for Environmental Risk Monitoring, National Institute
47 of Public Health, Bucharest, Romania, [22]Urology Department, "Carol Davila" University of
48 Medicine and Pharmacy - "Prof. Dr. Th. Burghele" Clinical Hospital, Bucharest,
49 Romania, [23]Clinic of Nefrology, Faculty of Medicine, Military Medical Academy, Belgrade,
50 Serbia, [24]Department of Urology, University Hospital “Dr D. Misovic” Clinical Center,
51 Belgrade, Serbia, [25]International Organization for Cancer Prevention and Research,
52 Belgrade, Serbia, [26]Cancer Research UK Edinburgh Centre, Institute of Genetics and
53 Cancer, University of Edinburgh, Edinburgh, Scotland, [27]Centre for Biodiversity Genomics,
54 University of Guelph, Guelph, Canada, [28]Ontario Tumour Bank, Ontario Institute for Cancer
55 Research, Toronto, Canada, [29]Experimental Research Center, Genomic Medicine

56 Laboratory, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil, [30]Post-Graduate
57 Program in Genetics and Molecular Biology, Universidade Federal do Rio Grande do Sul,
58 Porto Alegre, Brazil, [31]Diagnostic Laboratory Service, Personalized Medicine, Hospital de
59 Clínicas de Porto Alegre, Porto Alegre, Brazil, [32]Service of Urology, Hospital de Clínicas de
60 Porto Alegre, Porto Alegre, Brazil, [33]Post-Graduate Program in Medicine: Surgical Sciences,
61 Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, [34]Department of Anatomic
62 Pathology, A.C. Camargo Cancer Center, São Paulo, Brazil, [35]Department of Epidemiology,
63 A.C. Camargo Cancer Center, São Paulo, Brazil, [36]Department of Urology, A.C. Camargo
64 Cancer Center, São Paulo, Brazil, [37]National Institute for Science and Technology in
65 Oncogenomics and Therapeutic Innovation, A.C. Camargo Cancer Center, São Paulo,
66 Brazil, [38]Latin American Renal Cancer Group – LARCG, São Paulo, Brazil, [39]Department of
67 Surgery, Division of Urology, Sao Paulo Federal University - UNIFESP, São Paulo,
68 Brazil, [40]Molecular Oncology Research Center, Barretos Cancer Hospital, Brazil, [41]Life and
69 Health Sciences Research Institute (ICVS), School of Medicine, Minho University, Braga,
70 Portugal, [42]Department of Urology, Barretos Cancer Hospital, Brazil, [43]Department of
71 Pathology, Barretos Cancer Hospital, Brazil, [44]Leeds Institute of Medical Research at St
72 James's, University of Leeds, Leeds, UK, [45]Clinical Epidemiology, N.N.Blokhin National
73 Medical Research Centre of Oncology, Moscow, Russia, [46]Department of Urology,
74 N.N.Blokhin National Medical Research Centre of Oncology, Moscow, Russia, [47]Department
75 of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech
76 Republic, [48]Institute of Public Health & Preventive Medicine, 2nd Faculty of Medicine,
77 Charles University, Prague, Czech Republic, [49]Department of Oncology, 2nd Faculty of
78 Medicine, Charles University and Motol University Hospital, Prague, Czech
79 Republic, [50]Institute of Hygiene & Epidemiology, 1nd Faculty of Medicine, Charles University,
80 Prague, Czech Republic, [51]Faculty of Health Sciences, Palacky University, Olomouc, Czech
81 Republic, [52]Division of Cancer Epidemiology and Genetics, National Cancer Institute,
82 Rockville, USA, [53]Bevital AS, Bergen, Norway, [54]Observational & Pragmatic Research

Institute Pte Ltd, Singapore, Singapore, [55]Department of Pathology, Radboud University Medical Centre, Nijmegen, Netherlands

These authors contributed equally: Sergey Senkin, Sarah Moody

* Corresponding author: Paul Brennan

ABSTRACT

International differences in the incidence of many cancer types indicate the existence of carcinogen exposures that have not been identified by conventional epidemiology yet potentially make a substantial contribution to cancer burden[1]. This pertains to clear cell renal cell carcinoma (ccRCC), for which obesity, hypertension, and tobacco smoking are risk factors but do not explain its geographical variation in incidence[2]. Some carcinogens generate somatic mutations and a complementary strategy for detecting past exposures is to sequence the genomes of cancers from populations with different incidence rates and infer underlying causes from differences in patterns of somatic mutations. Here, we sequenced 962 ccRCC from 11 countries of varying incidence. Somatic mutation profiles differed between countries. In Romania, Serbia and Thailand, mutational signatures likely caused by extracts of Aristolochia plants were present in most cases and rare elsewhere. In Japan, a mutational signature of unknown cause was found in >70% cases and <2% elsewhere. A further mutational signature of unknown cause was ubiquitous but exhibited higher mutation loads in countries with higher kidney cancer incidence rates (p-value $<6 \times 10^{-18}$). Known signatures of tobacco smoking correlated with tobacco consumption, but no signature was associated with obesity or hypertension suggesting non-mutagenic mechanisms of action underlying these risk factors. The results indicate the existence of multiple, geographically variable, mutagenic exposures potentially affecting 10s of millions of people and illustrate the opportunities for new insights into cancer causation through large-scale global cancer genomics.

110 **INTRODUCTION**

111 The incidence rates of most adult cancers vary substantially between geographical regions
112 and many such differences are unexplained by known risk factors[1]. Together with unexplained
113 trends in incidence over time, this indicates the likely presence of unknown environmental or
114 lifestyle causes for many cancer types[1]. Traditional epidemiological studies have identified
115 many important lifestyle, environmental and infectious risk factors for cancer. However, they
116 have had limited success in recent decades suggesting that alternative study designs are
117 required if further risk factors are to be identified.

118

119 Characterization of mutational signatures within cancer genomes[3] is an approach
120 complementary to conventional epidemiology for investigating unknown causes of cancer.
121 Most cancers contain thousands of somatic mutations that have occurred over the lifetime of
122 the individual. These can be caused by endogenous cellular processes, such as imperfect
123 DNA replication and repair, or by exposure to exogenous environmental or lifestyle mutagens
124 such as ultraviolet radiation in sunlight and compounds in cigarette smoke. Mutational
125 signatures are the patterns of somatic mutation imprinted on genomes by individual mutational
126 processes. Analysis of thousands of cancer genome sequences from most cancer types has
127 established a set of reference mutational signatures including 71 single base substitution
128 (SBS) or doublet base substitution (DBS) signatures, and 18 small insertion and deletion (ID)
129 signatures[4]. A possible etiology has been suggested for 47 SBS/DBS signatures and nine ID
130 signatures.

131

132 Kidney cancer has particularly high incidence rates in Central and Northern Europe, notably in
133 the Czech Republic and Lithuania, and has shown increasing incidence in high income
134 countries in recent decades (**Fig. 1**)[2]. Most kidney cancers are clear cell renal cell carcinomas
135 (ccRCC)[3] for which obesity, hypertension and tobacco smoking are known risk factors[2].
136 However, these account for <50% of the global ccRCC burden and do not explain geographical
137 or temporal incidence trends. Previous ccRCC genome sequencing studies have included

138 relatively modest numbers of individuals from a small number of countries with limited variation
139 in ccRCC incidence[5–9] and have not comprehensively examined associations between ccRCC
140 risk factors and mutational signatures. To detect the activity of unknown carcinogens involved
141 in ccRCC development and to investigate the mechanisms of action of known risk factors, we
142 generated and analyzed epidemiological and whole genome sequencing data from a large
143 international series of ccRCC[10].
144

145 **RESULTS**

146 A total of 962 ccRCC cases from 11 countries in four continents were studied, including Czech
147 Republic (*n*=259), Russia (*n*=216), United Kingdom (*n*=115), Brazil (*n*=96), Canada (*n*=73),
148 Serbia (*n*=69), Romania (*n*=64), Japan (*n*=36), Lithuania (*n*=16), Poland (*n*=13), and Thailand
149 (*n*=5; **Fig. 1**; **Table 1**; **Methods**). These encompass a broad range of ccRCC incidence, from
150 the highest global age-standardized rates (ASRs) of Lithuania and Czech Republic (ASRs of
151 14.5 and 14.4/100,000 respectively) to the relatively low rates of Brazil and Thailand (ASRs of
152 4.5 and 1.8/100,000 respectively)[11]. Epidemiological questionnaire data were available on sex,
153 age at diagnosis, and important risk factors including body mass index (BMI), hypertension,
154 and tobacco smoking (**Table 1, Supplementary Table 1**). DNAs from ccRCCs and blood from
155 the same individuals were extracted and whole genome sequenced to average coverage of
156 54-fold and 31-fold, respectively.
157

158 Somatic mutation burdens in the 962 ccRCC genomes ranged from 803 to 45,376 (median
159 5,093) for single base substitutions (SBS), 2 to 240 (median 53) for doublet base substitutions
160 (DBS), and 10 to 14,770 (median 695) for small insertions and deletions (**Supplementary**
161 **Table 2**). The average burden of all three mutation types differed between the 11 countries (p-
162 value $<2 \times 10^{-23}$, p-value $<2 \times 10^{-14}$, p-value $<6 \times 10^{-14}$, for SBSs, DBSs, and IDs, respectively).
163 In particular, the burden of all mutation types was elevated in Romania compared to other
164 countries (**Extended Data Fig. 1**). Principal Component Analysis (PCA) performed on the

165 proportions of the six primary SBS mutation classes (C>A, C>G, C>T, T>A, T>C, T>G) in each
166 sample identified a distinct cluster of mainly Romanian and Serbian cases and a further cluster
167 of mainly Japanese cases (**Extended Data Fig. 2**). The results, therefore, clearly demonstrate
168 geographical variation of somatic mutation loads and patterns in ccRCC.
169
170 To investigate the mutational processes contributing to the geographical variation in mutation
171 burdens we extracted mutational signatures and estimated the contribution of each signature
172 to each ccRCC genome (**Supplementary Tables 3-7**). Ten signatures with strong similarity to
173 a reference signature in the Catalogue of Somatic Mutations in Cancer (COSMIC) database
174 were extracted: SBS1, due to deamination of 5-methylcytosine[12]; SBS2 and SBS13, due to
175 cytosine deamination by Apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like
176 (APOBEC) DNA editing enzymes[12]; SBS4, due to tobacco smoke mutagens[13]; SBS5, due to
177 an endogenous mutational process in which mutations accumulate with age[13]; SBS12, of
178 unknown cause; SBS18, due to DNA damage by reactive oxygen species[13]; SBS21 and
179 SBS44, due to defective DNA mismatch repair[13,14]; and SBS22, due to Aristolochic acid
180 exposure[15,16].
181
182 Five further SBS signatures were identified which were not well described by the COSMIC
183 catalogue (**Fig. 2**; **Supplementary Table 8**). SBS40a, SBS40b and SBS40c were present in
184 most ccRCC, accounting for, on average, ~30%, ~20%, and ~3% of mutations respectively
185 (**Fig. 2b**). Combined, they closely resemble the previously reported SBS40 (0.96 cosine
186 similarity), suggesting that the large number of ccRCC whole genomes analyzed here provides
187 the power to separate the constituent component signatures of SBS40. SBS40 was previously
188 reported frequently, and at high levels, in kidney cancer, but also in other cancers, and is of
189 unknown etiology. Like the composite SBS40, SBS40a is present in multiple cancer types.
190 However, SBS40b and SBS40c are largely restricted to ccRCC (**Supplementary Note,**
191 **Extended Data Fig. 3**). SBS_H was found in a single case and SBS_I is related to Aristolochic
192 acid exposure (see below; SBS_I has been renamed as SBS22b). Analysis of all other types

193 of mutational signatures, including doublet bases substitutions, small insertion and deletions,
194 copy number variants and structural variants, is presented in Supplementary results.
195
196 The mutation burdens of multiple SBS mutational signatures varied between the 11 countries.
197 SBS22 is thought to be caused by Aristolochic acids, mutagenic derivatives of plants of the
198 Aristolochia genus which are carcinogenic and also cause Balkan endemic nephropathy
199 (BEN), a kidney disease prevalent in areas adjacent to the Danube in Southeastern Europe[17].
200 SBS22 has previously been found in ccRCC, other urothelial tract cancers, and hepatocellular
201 carcinomas from Romania[5,18] and various countries in East and South-East Asia[15,16,19]. In this
202 study, SBS22 was present in high proportions of ccRCC from Romania (45/64, 70%), Serbia
203 (16/69, 23%), and Thailand (3/5, 60%), often with very high mutation burdens. Of note, given
204 the limited number of cases in Thailand, they may not be representative of ccRCC in that
205 region. The presence of SBS22 was strongly correlated with that of new signatures SBS_I,
206 DBS_D, and ID_C (**Extended Data Fig. 4-6**) which are, therefore, also probably due to
207 Aristolochic acid exposure. SBS_I, like SBS22, is composed predominantly of T>A mutations.
208 The signature identified previously as SBS22, has therefore been renamed SBS22a, and
209 SBS_I has been named SBS22b. The mutation burden of both SBS22a and SBS22b differed
210 between Serbia and Romania, with higher levels being detected in Romania, and away from
211 recognized BEN zones (**Fig 3a-c**). The two signatures may be due to different subsets of
212 Aristolochic acids, and/or to different metabolites, which induce slightly different mutational
213 patterns. Only five ccRCC cases were known to reside within recognized BEN zones,
214 suggesting no clear link between the two diseases. While the source of this exposure is
215 uncertain, these results indicate that a substantial proportion of the population over a wide
216 geographical area of Eastern Europe, possibly numbering in the 10s of millions, has been
217 exposed to Aristolochic acid containing compounds, the public health consequences of which
218 are uncertain.
219

220 SBS12 was present in 72% of Japanese and 2% of non-Japanese ccRCC (p-
221 value=4.7×10^{-78}) (**Extended Data Fig. 7h**). Compared to the mutation burdens imposed by
222 Aristolochic acid in ccRCC, SBS12 contributed modest mutation loads. SBS12 is composed
223 predominantly of T>C substitutions and exhibits strong transcriptional strand bias with more
224 T>C mutations on the transcribed than untranscribed strands of protein coding genes.
225 Transcriptional strand bias is typically caused by activity of transcription-coupled nucleotide
226 excision repair acting on bulky DNA adducts due to exogenous mutagenic exposures such as
227 tobacco smoke chemicals[13], ultraviolet light[13], Aristolochic acids[15], and aflatoxins[20]. Assuming
228 that transcription-coupled repair of DNA adducts is responsible for the SBS12 strand bias, the
229 adducts are likely on adenine. Alternatively, transcriptional strand bias can also be caused by
230 transcription-coupled damage[21,22]. The presence of SBS12 was replicated in two further series
231 of whole genome sequenced ccRCC from Japan including 14 cases from an independent
232 study group who undertook a broad genomic analysis of ccRCC but without detailed mutational
233 signature analysis[23], and a more recent unpublished series of 61 cases from an additional
234 cohort of ccRCC sequenced by the same center as the initial cohort (**Supplementary Note**).
235 SBS12 was present in 12/14 cases (85%) cases and 46/61 (75%) of cases, respectively.
236 SBS12 was previously reported in hepatocellular carcinomas[4,13] and additional analysis of
237 existing datasets revealed strong SBS12 enrichment in hepatocellular carcinomas from Japan
238 compared to other countries (p-value=3.8×10^{-15}; **Supplementary Note**). These results,
239 therefore, indicate that exposure to an agent contributing SBS12 mutations to kidney and liver
240 cancer is common in Japan and rare in the other 10 countries included in this study. The agent
241 responsible for SBS12 is unknown although the precedents provided by other mutational
242 signatures with strong transcriptional strand bias suggest that it is likely of exogenous
243 origin[21,22]. A polymorphism in aldehyde dehydrogenase 2 known to impair metabolism of
244 alcohol to aldehydes and common in Japan did not associate with levels of SBS12, and neither
245 did any other common germline variants (**Supplementary Note**).
246

247 SBS40a, SBS40b, and SBS40c were present in ccRCC from all 11 countries. The country-
248 specific average mutation burdens of SBS40a and SBS40b positively associated with country-
249 specific ASRs of kidney cancer incidence (p-value=0.0022 and p-value=5.1×10^{-18},
250 respectively; **Extended Data Fig. 8a**; **Fig. 4a, Supplementary Table 9**), with the highest
251 mutation loads in the Czech Republic and Lithuania. Kidney cancer incidence rates also vary
252 between the regions of the Czech Republic and SBS40b mutation burdens differed significantly
253 between these (p-value=0.011; **Fig 4b,c, Supplementary Table 10**), with the highest
254 attribution in the highest risk region. SBS40b exhibits modest transcriptional strand bias and,
255 assuming that transcription-coupled repair of DNA adducts is responsible, the adducts
256 underlying SBS40b are likely on pyrimidines. Insertion and deletion (indel) signatures ID5 and
257 ID8, which together contributed ~60% of the indel mutation burden on average, were also
258 strongly associated with country-specific kidney cancer ASR (p-value=1.3×10^{-10} and p-
259 value=7.1×10^{-5}, respectively, **Extended Data Fig. 8b,c**). Signatures ID5 and ID8 correlated
260 with each other (Spearman's r=0.78), as well as with SBS40b (r=0.79 and r=0.74, respectively)
261 indicating that they likely all constitute products of the same underlying mutational process.
262 Thus, the burdens of the full complement of mutation types generated by this mutational
263 process correlate with age-adjusted kidney cancer incidence rates. The overall mutational
264 burden did not, however, associate significantly with kidney cancer incidence rates (**Extended**
265 **Data Fig. 9**).
266
267 To investigate potential mutagenic agents underlying these geographically variable signatures,
268 an untargeted metabolomics screen of plasma was conducted on 901 individuals in the study,
269 from all countries except Japan (**Methods**). 2,392 metabolite features were obtained, including
270 944 independent peaks (r<0.85). Three features were associated with SBS4 (**Supplementary**
271 **Table 11**), with two identified as hydroxycotinine (p-value=2.9×10^{-9}) and cotinine (p-value=1.9
272 $\times 10^{-5}$), two major metabolites of nicotine[24]. Eight features were associated with SBS40b
273 (**Supplementary Table 11**). One feature was identified as N,N,N-trimethyl-L-alanyl-L-proline
274 betaine (TMAP; p-value=1.2×10^{-5}, **Supplementary Table 12**), increased levels of which

275 correlate strongly with reduced kidney function[25]. Other established measures of kidney
276 function, including cystatin C and creatinine, were correlated with TMAP (p-value = 2.5×10^{-30}
277 and 1.7×10^{-69}, respectively) and also showed evidence of positive association with SBS40b
278 (p-value=0.023 and 0.058, respectively). Thus, exposure to the mutagenic agent responsible
279 for SBS40b is associated with reduced kidney function. No recognized metabolome features
280 were significantly associated with any other signatures.
281
282 A total of 1962 "driver" mutations were found in 136 genes including *VHL*, *PBRM1*, *SETD2*
283 and *BAP1*, the known frequently mutated cancer genes in ccRCC (Methods) (**Fig. 5a,**
284 **Supplementary Table 13**)[9,26]. The frequencies of mutations in these genes were consistent
285 across countries (**Fig. 5b**). The spectrum of all driver mutations in ccRCC with Aristolochic
286 acid exposure (**Methods**) was enriched in T>A mutations compared to non-exposed cases
287 (25% vs 13%, p-value=0.0062, **Fig. 5c,d**) with similar enrichment specifically in VHL mutations
288 (30% vs 16%; **Fig. 5e,f**), and in the whole exomes (27% in exposed compared to 12% in
289 unexposed cases). Thus genome-wide Aristolochic acid mutagenesis has contributed in a
290 proportionate fashion to generation of driver mutations in Aristolochic acid-exposed ccRCC.
291 The driver mutation spectrum did not show statistically significant enrichment of T>C mutations
292 in SBS12 exposed cases (20% vs 12%, p = 0.069), but was consistent with the level of
293 enrichment in the exome (21% in exposed compared to 15% in unexposed cases). SBS40b
294 also did not show statistically significant enrichment possibly due to the ubiquitous exposure
295 and its relatively flat and featureless mutation profile.
296
297 Exogenous mutagenic exposures that ultimately cause cancer may be present during the early
298 stages of evolution of cancer clones. To time mutagenic exposures, the contribution of each
299 mutational signature to mutations in the primary clone (relatively early) and to mutations in
300 subclones (relatively late) were estimated[27,28] (**Methods**). All signatures of the putative
301 exogenous mutagenic exposures observed in ccRCC were present at relatively early stages
302 of cancer development, consistent with exposures to normal cells. SBS12, SBS22b, and

303 SBS40b showed higher activities in main clones compared to subclones (q-value=0.04, q-
304 value=0.02, q-value=2.3×10^{-5}, respectively) (**Extended Data Fig. 10**) and SBS22a showed
305 no significant difference[15,16]. By contrast, signatures due to endogenous mutational processes
306 including APOBEC DNA editing (SBS13) and oxidative damage (SBS18), were enriched in
307 subclones (q-value=1.6×10^{-4}, q-value=3.2×10^{-7}, respectively).
308
309 Established or suspected risk factors for ccRCC include age, tobacco smoking, obesity,
310 hypertension, diabetes, and environmental exposure to PFAS compounds[29]. Total SBS, DBS,
311 and ID mutation burdens associated with age, as did SBS1, SBS4, SBS5, SBS40a, SBS40b,
312 SBS22a, SBS22b, DBS2, ID1, ID5, and ID8. Total SBS (p-value=3.1×10^{-5}), DBS (p-
313 value=3.7×10^{-3}) and ID (p-value=1.3×10^{-4}) mutation burdens also associated with sex, with
314 males having higher mutation burdens than females, and with SBS40b showing a similar
315 association (p-value=9.3×10^{-5}). Associations with tobacco smoking were observed for SBS4
316 (p-value=5.3×10^{-6}) and DBS2 (p-value=2.4×10^{-7}), both known to be caused by tobacco
317 carcinogens[30,31]. These results suggest that the known increased risk of ccRCC with tobacco
318 smoking is due to direct exposure of the kidney to tobacco related mutagens (**Supplementary**
319 **Note**). Associations of particular mutational signatures with other ccRCC risk factors were not
320 observed (**Supplementary Tables 14, 15**). To complement this analysis of observational data,
321 associations between polygenic risk scores for known ccRCC risk factors and mutational
322 signatures[32,33] were examined (**Methods**). Consistent with the observational data, no
323 associations were found between genetically inferred risk factors and mutational signatures
324 except for tobacco smoking and DBS2 (p-value=0.01; **Supplementary Table 16**).
325
326
327 **DISCUSSION**
328 Somatic mutations in the genomes of 962 ccRCC patients from 11 countries indicate the
329 existence of multiple, widespread mutational processes exhibiting substantial geographical

330 variation in their contributions to ccRCC mutation loads. The results contrast with those from
331 552 esophageal squamous carcinomas from eight countries with widely different esophageal
332 carcinoma incidence rates in which geographical differences in mutation burdens or signatures
333 were not observed[34]. Together the studies implicate both geographically variable mutagenic
334 and non-mutagenic carcinogenic exposures contributing to global cancer incidence. Indeed,
335 the presence of mutational signatures associated with tobacco smoking but absence of
336 signatures associated with other known ccRCC risk factors, such as obesity and hypertension,
337 suggests that the latter may be mediated by non-mutagenic processes and, therefore, that
338 both classes of carcinogen contribute to the development of ccRCC.
339
340 The existence, identity, and carcinogenic effect of some of the agents underlying these
341 mutational processes are known. Aristolochic acids are believed to cause SBS22a/b, and its
342 associated signatures, but this study suggests that the geographical extent and proportion of
343 the population acquiring mutations in South-Eastern Europe is far greater than previously
344 anticipated, possibly affecting 10s of millions of individuals. The sources of the Aristolochic
345 acid exposure, the manner by which it is ingested and whether the exposure continues today
346 are uncertain, and further definition of the source and extent of this exposure is required in
347 order to provide a foundation for public health action.
348
349 The existence of the mutagenic exposures underlying SBS12 and SBS40b were not previously
350 suspected, and their causative agents are unknown. Based on current information, the
351 exposure causing SBS12 is restricted to Japan. However, larger studies are now indicated to
352 explore the geographical extent of exposure in Japan and neighboring countries, and the
353 proportions of their populations that have been exposed. Studies of Japanese migrants to other
354 countries are also likely to be informative regarding the potential source of exposure. In the
355 first instance this will be achievable by further sequencing of kidney and hepatocellular cancer
356 genomes. However, studies of normal tissues, using recently reported sequencing methods
357 allowing detection of somatic mutations in normal cells[35], and particularly relatively accessible

358 ones such as cells in urine that can be prospectively collected, may enable large population-
359 based studies providing better characterization of the exposure and its consequences. As with
360 exposure to Aristolochic acid in South-Eastern Europe, it is possible that 10s of millions of
361 individuals in East Asia are exposed to a potent mutagen, and identification of the source and
362 extent of exposure would seem to be a public health priority.
363
364 In contrast to Aristolochic acid and the agent causing SBS12, the exposure underlying SBS40b
365 appears to be globally ubiquitous. It predominantly causes mutations in ccRCC, with much
366 lower burdens in other cancer types, and generates mutation loads correlating strongly with
367 age and sex. There are few clues as to its origin or nature.
368
369 The incidence rates of ccRCC vary ~eightfold across the eleven countries from which ccRCCs
370 were sequenced. A strong positive correlation (p-value=5.5×10^{-18}) was found between the
371 average mutation loads attributable to SBS40b in each country (and also those of ID5 and ID8
372 which are correlated with SBS40b) and incidence of kidney cancer within each country. This
373 correlation reflects approximately a tripling of average country-specific SBS40b mutation loads
374 (a difference of ~1000 mutations) in parallel with the eightfold increase of country-specific ASR.
375
376 SBS40b mutation burdens also positively correlated with biomarkers of impaired kidney
377 function, reminiscent of the nephrotoxic effects of Aristolochic acids in Balkan endemic
378 nephropathy. It is possible that the increased SBS40b somatic mutation load itself engenders
379 this reduction in renal function. However, studies of other normal tissues suggest that they are
380 generally tolerant of elevated mutation burdens, except for manifesting a higher incidence of
381 neoplasia[36,37]. It is also possible that the agent underlying SBS40b is directly nephrotoxic, for
382 example by engendering DNA damage and a response to it, and that the mutations it
383 generates are immaterial to kidney function. It is also conceivable, however, that impaired renal
384 function, potentially due to many different causes, results in a metabolic state which itself
385 causes the elevated SBS40b mutation load. Whatever the mutational process underlying

386 SBS40b, it is plausible that it contributes to the geographical variation in the age standardized
387 rates of kidney cancer incidence rates. It is of public health interest to determine the cause of
388 SBS40b and, hence, to consider whether the exposure can be mitigated, potentially with
389 concomitant reduction in global ccRCC incidence rates.
390
391 The absence of any association between several known risk factors for ccRCC and mutation
392 burden, in particular for obesity and hypertension, supports a model of cancer development
393 where mutations are essential but additional factors affect the expansion of a mutated clone
394 and thus the chance of it progressing into cancer[38]. Further efforts at defining how lifestyle and
395 environmental exposures contribute to cancer development will therefore require a greater
396 understanding of both the causes of the mutations in cell clones in normal tissue, and the
397 further promotion of such mutant clones by non-mutagenic processes.
398
399 Finally, the substantial geographical variability of SBS12 and SBS22a/b, with most countries
400 not showing evidence of exposure, raises the possibility that additional mutational signature
401 studies of ccRCC involving more countries may reveal further mutagenic exposures.
402 Furthermore, the results relating to SBS40b highlight the prospect that a significant proportion
403 of global cancer burden may be caused by relatively ubiquitous exposures that are not readily
404 detectable by classical cancer epidemiology studies. The conduct of large scale whole genome
405 sequencing for other cancer sites across high and low risk populations around the world would
406 seem to be an appropriate strategy for detecting such novel cancer causing agents.
407
408
409 **REFERENCES**

410 1. Brennan, P. & Davey-Smith, G. Identifying Novel Causes of Cancers to Enhance
411 Cancer Prevention: New Strategies Are Needed. *JNCI: Journal of the National Cancer*
412 *Institute* **114**, 353–360 (2022).
413 2. Hsieh, J. J. *et al.* Renal cell carcinoma. *Nat Rev Dis Primers* **3**, 17009 (2017).

414 3. Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures:
415 emerging concepts, caveats and clinical applications. *Nat Rev Cancer* **21**, 619–637
416 (2021).
417 4. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer.
418 *Nature* **578**, 94–101 (2020).
419 5. Scelo, G. *et al.* Variation in genomic landscape of clear cell renal cell carcinoma
420 across Europe. *Nat Commun* **5**, 5135 (2014).
421 6. Mitchell, T. J. *et al.* Timing the Landmark Events in the Evolution of Clear Cell Renal
422 Cell Cancer: TRACERx Renal. *Cell* **173**, 611-623.e17 (2018).
423 7. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93
424 (2020).
425 8. Degasperi, A. *et al.* A practical framework and online tool for mutational signature
426 analyses show intertissue variation and driver dependencies. *Nat Cancer* **1**, 249–263
427 (2020).
428 9. The Cancer Genome Atlas Research Network. Comprehensive molecular
429 characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).
430 10. Mutographs Cancer Grand Challenge. *https://cancergrandchallenges.org/teams*.
431 11. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence
432 and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209–
433 249 (2021).
434 12. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast
435 Cancers. *Cell* **149**, 979–993 (2012).
436 13. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature*
437 **500**, 415–421 (2013).
438 14. Drost, J. *et al.* Use of CRISPR-modified human stem cell organoids to study the origin
439 of mutational signatures in cancer. *Science (1979)* **358**, 234–238 (2017).
440 15. Hoang, M. L. *et al.* Mutational Signature of Aristolochic Acid Exposure as Revealed by
441 Whole-Exome Sequencing. *Sci Transl Med* **5**, (2013).

442 16. Poon, S. L. *et al.* Genome-wide mutational signatures of aristolochic acid and its
443 application as a screening tool. *Sci Transl Med* **5**, 197ra101 (2013).
444 17. Grollman, A. P. Aristolochic acid nephropathy: Harbinger of a global iatrogenic
445 disease. *Environ Mol Mutagen* **54**, 1–7 (2013).
446 18. Turesky, R. J. *et al.* Aristolochic acid exposure in Romania and implications for renal
447 cell carcinoma. *Br J Cancer* **114**, 76–80 (2016).
448 19. Wang, X.-M. *et al.* Integrative genomic study of Chinese clear cell renal cell carcinoma
449 reveals features associated with thrombus. *Nat Commun* **11**, 739 (2020).
450 20. Huang, M. N. *et al.* Genome-scale mutational signatures of aflatoxin in cells, mice, and
451 human tumors. *Genome Res* **27**, 1475–1486 (2017).
452 21. Haradhvala, N. J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal
453 Mechanisms of DNA Damage and Repair. *Cell* **164**, 538–549 (2016).
454 22. Nik-Zainal, S. *et al.* The genome as a record of environmental exposure. *Mutagenesis*
455 gev073 (2015) doi:10.1093/mutage/gev073.
456 23. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat*
457 *Genet* **45**, 860–867 (2013).
458 24. Dempsey, D. *et al.* Nicotine metabolite ratio as an index of cytochrome P450 2A6
459 metabolic activity. *Clin Pharmacol Ther* **76**, 64–72 (2004).
460 25. Velenosi, T. J. *et al.* Untargeted metabolomics reveals N, N, N-trimethyl-L-alanyl-L-
461 proline betaine (TMAP) as a novel biomarker of kidney function. *Sci Rep* **9**, 6831
462 (2019).
463 26. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat*
464 *Genet* **45**, 860–867 (2013).
465 27. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
466 28. Dentro, S. C., Wedge, D. C. & van Loo, P. Principles of Reconstructing the Subclonal
467 Architecture of Cancers. *Cold Spring Harb Perspect Med* **7**, (2017).
468 29. Shearer, J. J. *et al.* Serum Concentrations of Per- and Polyfluoroalkyl Substances and
469 Risk of Renal Cell Carcinoma. *J Natl Cancer Inst* **113**, 580–587 (2021).

470 30. Nik-Zainal, S. *et al.* The genome as a record of environmental exposure. *Mutagenesis*
471 gev073 (2015) doi:10.1093/mutage/gev073.
472 31. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents.
473 *Cell* **177**, 821-836.e16 (2019).
474 32. Gabriel, A. A. G. *et al.* Genetic Analysis of Lung Cancer and the Germline Impact on
475 Somatic Mutation Burden. *JNCI: Journal of the National Cancer Institute* **114**, 1159–
476 1166 (2022).
477 33. Liu, Y., Gusev, A., Heng, Y. J., Alexandrov, L. B. & Kraft, P. Somatic mutational
478 profiles and germline polygenic risk scores in human cancer. *Genome Med* **14**, 14
479 (2022).
480 34. Moody, S. *et al.* Mutational signatures in esophageal squamous cell carcinoma from
481 eight countries with varying incidence. *Nat Genet* **53**, 1553–1563 (2021).
482 35. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature*
483 **593**, 405–410 (2021).
484 36. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic
485 mutations in normal human skin. *Science (1979)* **348**, 880–886 (2015).
486 37. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age.
487 *Science (1979)* **362**, 911–917 (2018).
488 38. Fowler, J. C. & Jones, P. H. Somatic Mutation: What Shapes the Mutational
489 Landscape of Normal Epithelia? *Cancer Discov* **12**, 1642–1655 (2022).
490

491 **FIGURE AND TABLE LEGENDS**

492 **Fig. 1: Eleven participating countries and estimated age-standardized incidence rates**
493 **of clear cell renal cell carcinomas.**

494 Incidence of clear cell renal cell carcinomas (ccRCC), men and women combined, age-
495 standardized incidence rates (ASR) per 100,000, data from GLOBOCAN 2020. Markers
496 indicate countries included in this study (number of participating ccRCC patients per
497 country).

498

499 **Table 1. Summary of clear cell renal cell carcinomas risk factors included in this**

500 **study.**

501

502 **Fig. 2: Single base substitution signature operative in clear cell renal cell carcinomas.**

503 **(a)** TMB plot showing the frequency and mutations per Mb for each of the decomposed SBS

504 signatures. **(b)** Average relative attribution for single base substitution (SBS) signatures

505 across countries. Signatures contributing less than 5% on average are grouped in the

506 'Others' category, apart from SBS12 and AA-related signatures SBS22a and SBS22b. '<95%

507 confidence' category accounts for the proportion of mutation burden which could not be

508 assigned to any signature with confidence level of at least 95%. **(c)** Decomposed signatures,

509 including reference COSMIC signatures as well as *de novo* signatures not decomposed into

510 COSMIC reference signatures.

511

512 **Fig. 3: Geospatial analysis of Aristolochic acid-related SBS signatures.**

513 Distribution of Romanian and Serbian cases with known residential history, along with the

514 summed levels of SBS22a and SBS22b attributions (per-case and regional estimate), with

515 respect to the Balkan endemic nephropathy (BEN) areas. White circles represented cases

516 with no detected activity of SBS22a and SBS22b.

517

518 **Fig. 4: Association of SBS40b signature attribution with incidence of kidney cancer.**

519 **(a)** Number of mutations attributed to signature SBS40b against age-standardized incidence

520 rates (ASR) of kidney cancer in each of the eleven countries represented in the cohort. Error

521 bars represent standard errors of the mean. **(b)** Number of mutations attributed to signature

522 SBS40b in four regions of Czech Republic against ASR of kidney cancer in each region.

523 Error bars represent standard errors of the mean. In **(a)** and **(b)**, the p-values are shown for

524 the ASR variable in linear regressions across all cases, adjusted for sex and age of

525 diagnosis. **(c)** Levels of attribution of SBS40b signature within Czech Republic, with bar plots

526 showing the number of cases for each quartile of SBS40b attribution across Prague,
527 Olomouc, Ceske Budejovice, and Brno regions.
528
529 **Fig. 5: Driver mutation analysis in clear cell renal cell carcinomas.**
530 **(a)** Frequency of driver genes in the cohort. Only genes mutated in at least 10 cases are
531 shown. **(b)** Frequency of driver genes across countries. Thailand, Poland and Lithuania are
532 not shown due to low sample numbers. **(c)** SBS-96 mutational spectra of all driver mutations
533 in ccRCC for Aristolochic acid (AA)-exposed and unexposed cases. **(d)** Percentage of T>A
534 driver mutations in AA-exposed and unexposed cases. **(e)** SBS-96 mutational spectra of VHL
535 mutations in ccRCC for AA-exposed and unexposed cases. **(f)** Percentage of T>A VHL
536 mutations in AA-exposed and unexposed cases.
537

538 **ONLINE METHODS**

539 **Recruitment of cases and informed consent**

540 The International Agency for Research on Cancer (IARC/WHO) coordinated case recruitment
541 through an international network of over 40 collaborators from the 11 participating countries
542 (**Table1**; **Supplementary Table 17**). The inclusion criteria for patients were >=18 years of age
543 (ranging from 23 to 87, with a mean of 60 and a standard deviation of 12), confirmed diagnosis
544 of primary ccRCC and no prior cancer treatment. Informed consent was obtained for all
545 participants. Patients were excluded if they had any condition that could interfere with their
546 ability to provide informed consent or if there were no means of obtaining adequate tissues or
547 associated data as per the protocol requirements. Ethical approvals were first obtained from
548 each Local Research Ethics Committee and Federal Ethics Committee when applicable, as
549 well as from the IARC Ethics Committee.

550

551 **Bio-samples, data collection, and expert pathology review**

552 Dedicated standard operating procedures, following guidelines from the International Cancer
553 Genome Consortium (ICGC), were designed by IARC/WHO to select appropriate case series
554 with complete biological samples and exposure information as described previously[1]
555 (**Supplementary Table 17**). In brief, for all case series included, anthropometric measures
556 were taken, together with relevant information regarding medical and familial history.
557 Comparable smoking and alcohol history was available from all centers. Detailed
558 epidemiological information on residential history was collected in Czech Republic, Romania,
559 and Serbia. Potential limitations of using retrospective clinical data collected using different
560 protocols from different populations were addressed by a central data harmonization to ensure
561 a comparable group of exposure variables (**Supplementary Table 17**). All patient related data
562 as well as clinical, demographical, lifestyle, pathological and outcome data were
563 pseudonymized locally through the use a dedicated alpha-numerical identifier system before
564 being transferred to IARC/WHO central database.

565 Original diagnostic pathology departments provided diagnostic histological details of
566 contributing cases through standard abstract forms. IARC/WHO centralized the entire
567 pathology workflow and coordinated a centralized digital pathology examination of the frozen
568 tumor tissues collected for the study as well as formalin-fixed, paraffin-embedded (FFPE)
569 sections when available, via a web-based report approach and dedicated expert panel
570 following standardized procedures as described previously[1]. A minimum of 50% viable tumor
571 cells was required for eligibility to whole genome sequencing.
572 In summary, frozen tumor tissues were first examined to confirm the morphological type and
573 the percentage of viable tumor cells. A random selection of tumor tissues was independently
574 evaluated by a second pathologist. Enrichment of tumor component was performed by
575 dissection of non-tumoral part, if necessary. 90 cases overlapped with a previously published
576 cohort recruited under the Cancer Genomics of the Kidney (CAGEKID) project[2], which were
577 also part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project[3].
578
579 **DNA extraction**
580 Extraction of DNA from fresh frozen tumor and matched blood samples was centrally
581 conducted at IARC/WHO except for Japan, which performed DNA extractions at the local
582 center following a similarly standardized DNA extraction procedure. Of the cases which
583 proceeded to the final analysis (*n*=962), germline DNA was extracted from either buffy-coat,
584 whole blood, or from adjacent normal tissue (*viz.*, samples from Japan) using previously
585 described protocols and methods[1].
586
587 **Whole genome sequencing**
588 In total, 1583 renal cell carcinoma cases were evaluated, with 1267 confirmed as ccRCC
589 cases. 116 (9%) were excluded due to insufficient viable tumor cells (pathology level), or
590 inadequate DNA (tumor or germline). DNA from 1151 cases was received at the Wellcome
591 Sanger Institute for whole genome sequencing. Fluidigm SNP genotyping with a custom panel
592 was performed to ensure that each pair of tumor and matched normal samples originated from

593 the same individual. Whole genome sequencing (150bp paired end) was performed on the
594 Illumina NovaSeq 6000 platform with target coverage of 40X for tumors and 20X for matched
595 normal tissues. All sequencing reads were aligned to the GRCh38 human reference genome
596 using Burrows-Wheeler-MEM (v0.7.16a and v0.7.17). Post-sequencing QC metrics were
597 applied for total coverage, evenness of coverage and contamination. Cases were excluded if
598 coverage was below 30X for tumor or 15X for normal tissue. For evenness of coverage, the
599 median over mean coverage (MoM) score was calculated. Tumors with MoM scores outside
600 the range of values determined by previous studies to be appropriate for whole genome
601 sequencing (0.92 – 1.09) were excluded. Conpair[4] (https://github.com/nygenome/Conpair)
602 was used to detect contamination, cases were excluded if the result was greater than 3%[5]. A
603 total of 962 cases passed all criteria and were included in subsequent analysis.
604
605 **Somatic variant calling**
606 Variant calling was performed using the standard Sanger bioinformatics analysis pipeline
607 (https://github.com/cancerit). Copy number profiles were determined first using the algorithms
608 ASCAT[6] and BATTENBERG[7], where tumor purity allowed. SNV were called with
609 cgpCaVEMan[8], indels were called with cgpPINDEL[9], and structural rearrangements were
610 called using BRASS. CaVEMan and BRASS were run using the copy number profile and purity
611 values determined from ASCAT where possible (complete pipeline, n=857). Where tumor
612 purity was insufficient to determine an accurate copy number profile (partial pipeline, n=105),
613 CaVEMan and BRASS were run using copy number defaults and an estimate of purity
614 obtained from ASCAT/BATTENBERG. For SNV additional filters (ASRD >= 140 and CLPM
615 ==0) were applied to remove potential false positive calls. A second variant caller, Strelka2,
616 was run for SNVs and indels as consensus variant calling was previously shown to eliminate
617 algorithm specific artefacts and to generate highly reproducible mutational spectra compared
618 to using a single variant calling algorithm[1,10]. Only variants called by both the Sanger variant
619 calling pipeline and Strelka2 were included in subsequent analysis.
620

621 **Validation of Japanese sequencing**

622 The matched normal tissue sequenced was blood for all countries with the exception of Japan,
623 where adjacent normal kidney was used. As Japan displayed an enrichment of SBS12,
624 matched blood was obtained from 28 of the 36 patients to confirm that the source of the
625 matched normal tissue was not influencing the result. In all cases, the mutational spectra of
626 Japanese ccRCC generated using either blood or adjacent normal kidney matched each other
627 with a cosine similarity of >0.99.

628

629 **Generation of mutational matrices**

630 Mutational matrices for single base substitutions (SBS), doublet base substitutions (DBS) and
631 small insertions and deletions (ID) were generated using SigProfilerMatrixGenerator
632 (https://github.com/AlexandrovLab/SigProfilerMatrixGenerator) with default options
633 (v1.2.12)[11].

634

635 **Mutational signature analysis**

636 Mutational signatures were extracted using two algorithms, SigProfilerExtractor
637 (https://github.com/AlexandrovLab/SigProfilerExtractor), based on nonnegative matrix
638 factorization, and mSigHdp[12] (https://github.com/steverozen/mSigHdp), based on the
639 Bayesian hierarchical Dirichlet process. For SigProfilerExtractor, *de novo* mutational
640 signatures were extracted from each mutational matrix using SigProfilerExtractor with
641 nndsvd_min initialization (NMF_init="nndsvd_min") and default parameters (v1.1.9)[13]. Briefly,
642 SigProfilerExtractor deciphers mutational signatures by first performing Poisson resampling of
643 the original matrix with additional renormalization (based on a generalized mixture model
644 approach) of hypermutators to reduce their effect on the overall factorization[13]. Nonnegative
645 matrix factorization (NMF) was performed using initialization with nonnegative singular value
646 decomposition and by applying the multiplicative update algorithm using the Kullback–Leibler
647 divergence as an objective function[13]. NMF was applied with factorizations between k=1 and
648 k=20 signatures; each factorization was repeated 500 times[13]. *De novo* single base

649 substitution mutational signatures were extracted with SigProfilerExtractor for both SBS-288
650 and SBS-1536 contexts[11]. The results were largely concordant with the SBS-1536 *de novo*
651 signatures allowing additional separation of mutational processes, therefore the SBS-1536 *de*
652 *novo* signatures were taken forward for further analysis (**Supplementary Table 3**). Mutational
653 signatures for DBS and ID were extracted in DBS-78 and ID-83 contexts respectively
654 (**Supplementary Tables 4, 5**). Where possible, SigProfilerExtractor matched each *de novo*
655 extracted mutational signature to a set of previously identified COSMIC signatures[14], for SBS-
656 1536 signatures this requires collapsing the 1536 classification into the standard 96
657 substitution type classification with six mutation classes having single 3' and 5' sequence
658 contexts (**Supplementary Table 8**). This step makes it possible to distinguish between *de*
659 *novo* signatures which can be explained by a combination of the known catalog of mutational
660 process (which have not been completely separated during the extraction), and those which
661 have not been previously identified. mSigHdp extraction of SBS-96 and ID-83 signatures was
662 performed using the suggested parameters and using the country of origin to construct the
663 hierarchy. SigProfilerExtractor's decomposition module was subsequently used to match
664 mSigHdp *de novo* signatures to previously identified COSMIC signatures[14]. Further details on
665 the comparison of results between SigProfilerExtractor and mSigHdp and decomposition of *de*
666 *novo* signatures into COSMIC reference signatures can be found in the **Supplementary Note.**
667

668 **Attribution of activities of mutational signatures**

669 The *de novo* (SigProfiler) and COSMIC signature (SigProfiler and mSigHdp) activities were
670 attributed for each sample using the MSA signature attribution tool (v2.0,
671 https://gitlab.com/s.senkin/MSA)[15]. For COSMIC attributions, only COSMIC reference
672 signatures, which were identified in the decomposition of *de novo* signatures, were included in
673 the panel for attribution, in addition to *de novo* signatures which could not be decomposed into
674 COSMIC reference. At its core, the tool utilizes the nonnegative least squares (NNLS)
675 approach minimizing the L2 distance between the input sample and the one reconstructed
676 using available signatures. To limit false positive attributions, an automated optimization

677 procedure was applied by repeated removal of all signatures that do not increase the L2
678 similarity of a sample by >0.008 for SBS, >0.014 for DBS, and >0.03 for ID mutation types, as
679 suggested by simulations. These optimal penalties were derived using an optional parameter
680 (params.no_CI_for_penalties = false) utilizing a conservative approach in calculation of
681 penalties. Finally, a parametric bootstrap approach was applied to extract 95% confidence
682 intervals for each attributed mutational signature activity.
683
684 **Driver mutations**
685 A dNdS approach was used to identify genes under positive selection in ccRCC[16]. The analysis
686 was performed both for the whole genome (q-value<0.01), and with restricted hypothesis
687 testing (RHT) for a panel of 369 known cancer genes[16]. Variants in any gene identified as
688 under positive selection in global dNdS or in the 369-cancer gene panel were assessed as
689 potential drivers[16]. Candidate driver mutations were annotated with the mode of action using
690 the Cancer Gene Census (https://cancer.sanger.ac.uk/census) and the Cancer Genome
691 Interpreter tool (https://www.cancergenomeinterpreter.org). Missense mutations were
692 assessed using the MutationMapper tool (http://www.cbioportal.org/mutation_mapper).
693 Variants were considered likely drivers if they met any of the following criteria: *(i)* Truncating
694 mutations in genes annotated as tumor suppressors; *(ii)* mutations annotated as likely or
695 known oncogenic in MutationMapper; (iii) truncating variants in genes with selection (q-
696 value<0.05) for truncating mutations assumed to be tumor suppressors and thus likely drivers;
697 *(iv)* missense variants in all genes under positive selection and with dN/dS ratios for missense
698 mutations above 5 (assuming 4 of every 5 missense mutations are drivers) labelled as likely
699 drivers; or *(v)* in-frame indels in genes under significant positive selection for in-frame indels.
700
701 **Evolutionary analysis**
702 Subclonal architecture reconstruction was performed using the DPClust R package v2.2.8[7,17],
703 after obtaining cancer cell fraction (CCF) estimates by dpclust3p v1.0.8
704 (https://github.com/Wedge-lab/dpclust3p) based on the variant allele frequency provided by

705 the somatic variant callers and the copy number profiles determined by the BATTENBERG
706 algorithm. Only tumors with at least 40% purity according to BATTENBERG were considered
707 for further evolutionary analysis. For each tumor with at least one subclone, the respective
708 somatic mutations were split into clonal and subclonal mutations using the most probable
709 cluster assignment for each mutation as per the DPClust output. Mutations not assigned to a
710 cluster by DPClust were removed from further analysis. Clusters centered at a CCF>1.5 and
711 ones where chromosome X contributed the highest number of mutations were deemed
712 artifactual, and the respective mutations were removed. Samples with a total number of clonal
713 or subclonal mutations below 256 were also removed. Additionally, samples with poor
714 separation between the clonal and subclonal distributions (*e.g.*, subclone centered at a
715 CCF>0.80) were removed. Finally, only samples that had both a clone and at least one
716 subclone post-filtering were retained for further analysis. This yielded a total of 223 samples,
717 each with clonal and subclonal mutations. SigProfilerAssignment (v0.0.13)
718 (https://github.com/AlexandrovLab/SigProfilerAssignment) was used to identify the activity of
719 each mutational signature in each clone/subclone, and these activities were then normalized
720 by the total number of mutations belonging to the clone/subclone (*i.e.*, clonal mutations were
721 not included in the subclone). A two-sided Wilcoxon Signed-Rank Test[18] was used to assess
722 the differences in the relative activity of each mutational signature between the clones and
723 their respective subclones. P-values were corrected using the Benjamini-Hochberg
724 procedure[19] and reported as q-values in the manuscript.
725
726 **Regressions**
727 Signature attributions were dichotomized into presence and absence using confidence
728 intervals, with presence defined as both lower and upper limits being positive, and absence as
729 the lower limit being zero. If a signature was present in at least 75% of cases (SBS1, SBS40a,
730 SBS40b, ID1, and ID5), it was dichotomized into above and below the median of attributed
731 mutation counts. The binary attributions served as dependent variables in logistic regressions,
732 and relevant risk factors were used as factorized independent variables. To adjust for

733 confounding factors, sex, age of diagnosis, country, and tobacco status were added as
734 covariates in regressions. The Bonferroni method was used to test for significant p-values (*i.e.,*
735 a total of 224 comparisons for regressions with signatures, and a total of 24 comparisons for
736 regressions with mutation burden). P-values reported are raw (not corrected). Regressions
737 with incidence of renal cancer were performed as linear regressions with mutation burdens or
738 signature attributions (those present in at least 75% of cases) with confidence intervals not
739 consistent with zero as a dependent variable, and age-standardized rates (ASR) of renal
740 cancer obtained from Global Cancer Observatory (GLOBOCAN)[20], sex and age of diagnosis
741 as independent variables. ASR of renal cancer for regions of Czech Republic were obtained
742 from SVOD web portal[21]. Signatures present in less than 75% of cases were dichotomized into
743 presence and absence as previously mentioned and analyzed using the logistic regressions.
744 All regressions were performed on a sample basis.
745
746 **Polygenic risk score (PRS) analysis of lifestyle risk factors**
747 In this analysis, we used the genome-wide association studies (GWAS) summary statistics
748 estimated in European populations for well-established risk factors for ccRCC. For tobacco
749 smoking status, we used results from the GSCAN consortium meta-analysis of smoking
750 initiation (ever vs never status)[22]. For body mass index (BMI), the results of UK biobank (UKBB)
751 and GIANT consortium meta-analysis of continuous BMI were used[23]. GWAS summary
752 statistics related to hypertension, namely systolic blood pressure and diastolic blood pressure,
753 as well as the ones related to diabetes[24], such as fasting glucose and fasting insulin were also
754 obtained using UKBB studies[25].
755
756 Since all the GWAS summary statistics used in the current work were based on European
757 populations, we used ADMIXTURE tool (v1.3.0)[26] and principal component analysis (PCA) to
758 infer the unsupervised cluster of individuals with European genetic background within ccRCC
759 cases. Hapmap SNPs (n=1,176,821 variants) were extracted from the ccRCC whole-genome
760 sequence genotype data. After basic quality control using PLINK (v1.9b, www.cog-

761 genomics.org/plink/1.9/), 333 variants were removed due to missing genotype rate > 5%, 1,236
762 variants failed Hardy-Weinberg equilibrium test (p-values<10^{-8}), and 18,702 variants had
763 MAF<1% in our cohort. Additionally, 3 ambiguous variants and 21,358 variants within regions
764 of long-range, high linkage disequilibrium (LD) in the human genome (hg38) were excluded.
765 After pruning for linkage disequilibrium, 143,727 variants remained in ccRCC genotype data.
766 The 1000 genome reference population genotype data (phase 3) for Europeans (N=489),
767 Africans (YRI, N=108) and East Asians (N=103 from China and 104 from Japan)
768 (https://www.internationalgenome.org/data/) were filtered and merged with ccRCC genotype
769 data based on the pruned set of variants present in both datasets. ADMIXTURE was run on
770 the merged genotype data with *k*=3, which would correspond to the three ancestral continental
771 population groups that likely reflect the participants of our study. The ccRCC cases with
772 European genetic fraction greater than 80% by the ADMIXTURE analysis were selected for
773 the polygenic risk scores (PRS) analyses. To complement the ADMIXTURE analysis, PCA
774 was run on the same samples.
775
776 The initial genotype data based on whole-genome sequence from 849 ccRCC cases with
777 European genetic background consisted of biallelic SNPs with MAF >0.01% (to exclude ultra-
778 rare variants; N ~ 30 million variants). After basic quality control, variants with missing
779 genotype rate of greater than 5% (N=7,519,196 variants) with strong deviation from Hardy-
780 Weinberg equilibrium (p-values<10^{-8}, N=220,862) were excluded. For each GWAS trait, we
781 restricted our analyses to the biallelic SNPs with minor allele frequency (MAF) greater than 1%
782 in the 1000 genomes reference for European populations. For the selection of the independent
783 genome-wide significant hits (p-values<5×10^{-8}) of each GWAS summary statistic used to
784 generate the PRS, SNPs were clumped (r^2=0.1 within a LD window of 10 MB) using PLINK
785 (v1.9b, www.cog-genomics.org/plink/1.9/) based on the 1000 genomes European reference
786 population genotype data (N=489; ~ 10 million variants). Where a selected GWAS hit was not
787 found in ccRCC genotype data, we extracted proxies (r^2>0.8 in 1000 genomes) also present
788 in ccRCC dataset where possible (**Supplementary Table 18**). The variance of each genetic

789 trait explained by the genetic variants were calculated as previously suggested[27]. PRS was
790 subsequently calculated as the sum of the individual's beta-weighted genotypes using PRSice-
791 2 software[28]. Associations were estimated per standard deviation increase in the PRS, which
792 was normalized to have a mean of zero across ccRCC cases of European genetic ancestry.
793
794 **Untargeted metabolomics association with signatures**
795 Of the 962 subjects from the main analysis, 901 subjects were included in this sub-study – all
796 Japanese samples (*n*=36) as well as few cases from Czech Republic (*n*=13), Romania (*n*=5)
797 and Russia (*n*=3) were not included due to lack of available plasma samples. Samples were
798 randomized and analyzed as two independent analytical batches. Analysis was performed with
799 a UHPLC-QTOF-MS system that consisted of a 1,290 Binary LC and a 6,550 QTOF mass
800 spectrometer equipped with Jet Stream electrospray ionization source (Agilent Technologies),
801 using previously described methods[29]. Pre-processing was performed using Profinder
802 10.0.2.162 and Mass Profiler Professional B.14.9.1 software (Agilent Technologies). A "Batch
803 recursive feature extraction (small molecules)" process was employed for samples and blanks
804 to find $[M+H]^+$ ions. The two batches were processes separately and the resulting features
805 were aligned in Mass Profiler Professional. Chromatographic peak areas were used as a
806 measurement of intensity. No normalization or transformation of raw data was performed prior
807 to the downstream data analysis.
808
809 A total of 2,392 features were detectable in at least one of the 901 samples. Features present
810 in only one of the two batches were filtered out. Recursive filtering elimination was applied to
811 decrease redundancy from highly correlated variables ($r \geq 0.85$, Pearson's r calculated before
812 any transformation/imputation) by selecting the features with least missing data within clusters
813 of features. A total of 944 features were included in the statistical analysis. Features were pre-
814 processed: missing values were replaced with 1/5 of the minimal value of the feature before
815 applying mean centering and Pareto scaling. Each feature was regressed against both *de novo*
816 and COSMIC signatures, adjusting for sex and age of diagnosis, as well as body mass index

817 (BMI) and technical factors (batch, acquisition order) that could impact chromatographic peak
818 area. Models for SBS22a and SBS22b were restricted to Romanian and Serbian samples to
819 find potential pathways of Aristolochic acid exposure in the Balkan region. Logistic models
820 were used for zero-inflated signatures (≥30% zeros) while quasi-Poisson regressions were
821 used for the least zero-inflated signatures (SBS1, SBS40a, and SBS40b). To derive specific
822 false detection rates, random variables were created from permutations of the initial features
823 and regressed against signatures in the same fashion as true features. Maximum p-value
824 thresholds from regressions with random features were compared to adjusted p-value
825 thresholds according to Bonferroni's procedure. The more conservative approach was used in
826 selecting features of interest. Random forest models were also used as cross-checking
827 multivariate models to assess the relative importance of each feature in explaining the
828 signature attribution. As with univariate models, regression models were used for the least
829 zero-inflated signatures (<30% of zeros) while classification models were used for all other
830 signatures, with restriction to Romanian and Serbian samples for SBS22a and SBS22b.
831 Importance was estimated from the total decrease in node impurities from splitting on the
832 variable, averaged over all trees. Node impurity was measured by the Gini index for
833 classification, and by residual sum of squares for regression. The significance of importance
834 metrics for Random Forest models were estimated by permuting the response variable
835 (https://github.com/EricArcher/rfPermute).
836
837 Features considered for identification, along with their highly correlated counterparts, were
838 searched in Human Metabolome Database (HMDB), LipidMaps, Metlin, and Kegg. Compound
839 identity was confirmed by comparison of retention times and MS/MS fragmentation against
840 chemical standards when available, or otherwise against reference MS/MS spectra. Since the
841 feature 240.1468@0.8929933 was strongly correlated with several features identified as
842 TMAP (**Supplementary Table 11**), the integration of these features was inspected and
843 corrected manually, and regressed against SBS40b using the same model applied to features
844 selected for analysis. Creatinine was identified among the features by matching its retention

845 time and MS/MS spectra against a reference standard and also regressed against SBS40b in
846 the same fashion as other metabolites. Estimation of correlation between metabolic features
847 was done using linear regression adjusting for batch and acquisition order.
848
849 **Targeted metabolomics analyses**
850 Circulating levels of PFAS (Per- and Polyfluorinated Substances) and cystatin C compounds
851 were investigated using targeted mass spectrometry-based methods as described
852 previously[30,31].
853
854 Out of the 962 subjects from the main analysis, plasma samples from 909 subjects (from all
855 countries except Japan) were randomized and sent frozen in dry ice to each respective
856 laboratory for analyses. Measurement of cystatin C from 906 subjects included its native form
857 and isoforms (3Pro-OH cystatin C, cystatin C-desS, 3Pro-OH cystatin C-desS and cystatin C-
858 desSSP) that were modeled individually and for the total concentration of cystatin C isoforms.
859 Measurements of PFAS compounds included PFOA (Perfluorooctanoic Acid; total, branch,
860 linear), PFOS (Perfluorooctanoic Acid; total, branch, linear), PFHxS (Perfluorohexane
861 sulfonate), PFNA (Perfluorononanoic acid), PFDA (Perfluorodecanoic acid), MePFOSAA (n-
862 methylperfluoro-1 octanesulfonamido acetic acid), EtPFOSAA (2-(N-Ethyl-perfluorooctane.
863 sulfonamido) acetic acid).
864
865 Multivariable quasi-Poisson (for the least sparse signatures SBS1, SBS40a and SBS40b) and
866 logistic regression were used to estimate the association between plasma concentrations of
867 the aforementioned substances and mutational signatures. All compounds were modeled
868 continuously (log2-transformed) and categorically, with adjustments made by sex, age, date
869 of recruitment, country, BMI, tobacco and alcohol status in the case of PFAS molecules and
870 by sex, age and BMI, in the case of cystatin C.
871
872

873 **Geospatial analyses**

874 Geospatial analyses were performed to estimate the regional effect for signature attribution,
875 particularly for signatures thought to be from exogenous exposure (SBS40b – unknown – and
876 SBS22a/SB22b - Aristolochic acid). Residential history information was available for a large
877 proportion of cases from the countries of interest: Czech Republic for SBS40b and
878 Romania/Serbia for SBS22a/SBS22b. The 259 cases from Czech Republic within this study
879 were recruited from 4 separate regions including Prague, České Budějovice (in Southern
880 Bohemia), as well as Brno and Olomouc in the east of the country. Each individual residence
881 was geocoded to its administrative region. All locations outside the country of recruitment were
882 labeled as "Abroad". A multi-membership mixed model was used to account for the full list of
883 regions in which each subject resided, as well as the proportion of life spent in that region
884 before diagnosis. As dependent variable, signatures were inverse-normal transformed. Models
885 were adjusted for sex and age of diagnosis (fixed effects). The regional effect was treated as
886 random effect.

887

888 **Data availability**

889 Whole genome sequencing data and patient metadata are deposited in the European
890 Genome-phenome Archive (EGA) associated with study EGAS00001003542. Aligned BAM
891 files for all ccRCC cases included in the final analysis were deposited in dataset
892 EGAD00001012102, variant calling files in dataset EGAD00001012222 and patient metadata
893 in dataset EGAD00001012223. The metabolomics data are available upon request. All other
894 data are provided in the accompanying **Supplementary Tables**.

895

896 **Code availability**

897 All algorithms used for data analysis are publicly available with repositories noted within the
898 respective method sections and in the accompanying reporting summary. Code used for
899 regression analysis and figures is available at:
900 https://gitlab.com/Mutographs/Mutographs_RCC.

901

902

903 **Methods references**

904 1. Moody, S. *et al.* Mutational signatures in esophageal squamous cell carcinoma from
905 eight countries with varying incidence. *Nat Genet* **53**, 1553–1563 (2021).
906 2. Scelo, G. *et al.* Variation in genomic landscape of clear cell renal cell carcinoma
907 across Europe. *Nat Commun* **5**, 5135 (2014).
908 3. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93
909 (2020).
910 4. Whalley, J. P. *et al.* Framework for quality assessment of whole genome cancer
911 sequences. *Nat Commun* **11**, 5040 (2020).
912 5. Bergmann, E. A., Chen, B.-J., Arora, K., Vacic, V. & Zody, M. C. Conpair:
913 concordance and contamination estimator for matched tumor–normal pairs.
914 *Bioinformatics* **32**, 3196–3198 (2016).
915 6. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the*
916 *National Academy of Sciences* **107**, 16910–16915 (2010).
917 7. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
918 8. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to
919 Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics*
920 **56**, (2016).
921 9. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion
922 Events from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**, (2015).
923 10. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat*
924 *Methods* **15**, 591–594 (2018).
925 11. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring
926 patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).

927 12. Liu, M., Wu, Y., Jiang, N., Boot, A. & Rozen, S. G. mSigHdp: hierarchical Dirichlet
928 process mixture modeling for mutational signature discovery. *bioRxiv*
929 2022.01.31.478587 (2022) doi:10.1101/2022.01.31.478587.
930 13. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction
931 with SigProfilerExtractor. *Cell genomics* **2**, None (2022).
932 14. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer.
933 *Nature* **578**, 94–101 (2020).
934 15. Senkin, S. MSA: reproducible mutational signature attribution with confidence based
935 on simulations. *BMC Bioinformatics* **22**, 540 (2021).
936 16. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues.
937 *Cell* **171**, 1029-1041.e21 (2017).
938 17. Dentro, S. C., Wedge, D. C. & van Loo, P. Principles of Reconstructing the Subclonal
939 Architecture of Cancers. *Cold Spring Harb Perspect Med* **7**, (2017).
940 18. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**, 80
941 (1945).
942 19. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
943 Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series*
944 *B (Methodological)* **57**, 289–300 (1995).
945 20. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence
946 and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209–
947 249 (2021).
948 21. Dušek, L. *et al.* Epidemiology of Malignant Tumours in the Czech Republic
949 [online]. Masaryk University, Czech Republic, [2005]. *http://www.svod.cz/. Version 7.0*
950 *[2007], ISSN 1802 – 8861*.
951 22. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into
952 the genetic etiology of tobacco and alcohol use. *Nat Genet* **51**, 237–244 (2019).

953 23. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body
954 mass index in ~700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641–
955 3649 (2018).
956 24. Lagou, V. *et al.* Sex-dimorphic genetic effects and novel loci for fasting glucose and
957 insulin variability. *Nat Commun* **12**, 24 (2021).
958 25. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
959 *Nature* **562**, 203–209 (2018).
960 26. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry
961 in unrelated individuals. *Genome Res* **19**, 1655–1664 (2009).
962 27. Shim, H. *et al.* A Multivariate Genome-Wide Association Analysis of 10 LDL
963 Subfractions, and Their Response to Statin Treatment, in 1868 Caucasians. *PLoS*
964 *One* **10**, e0120758 (2015).
965 28. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-
966 scale data. *Gigascience* **8**, (2019).
967 29. Loftfield, E. *et al.* Novel Biomarkers of Habitual Alcohol Intake and Associations With
968 Risk of Pancreatic and Liver Cancers and Liver Disease Mortality. *JNCI: Journal of the*
969 *National Cancer Institute* **113**, 1542–1550 (2021).
970 30. Shearer, J. J. *et al.* Serum Concentrations of Per- and Polyfluoroalkyl Substances and
971 Risk of Renal Cell Carcinoma. *J Natl Cancer Inst* **113**, 580–587 (2021).
972 31. Gao, J., Meyer, K., Borucki, K. & Ueland, P. M. Multiplex Immuno-MALDI-TOF MS for
973 Targeted Quantification of Protein Biomarkers and Their Proteoforms Related to
974 Inflammation and Renal Dysfunction. *Anal Chem* **90**, 3366–3373 (2018).
975

976 **Acknowledgements**

977 The authors would like to thank Laura O'Neill, Kirsty Roberts, Katie Smith, Maisie Farenden,
978 Siobhan Austin-Guest and the staff of DNA Pipelines at the Wellcome Sanger Institute for their
979 contribution. We are grateful for the support provided by Maja Milojevic, Christophe Lallemand,
980 Helene Renard, Aude Bardot, Andreea Spanu and Nivonirina Robinot as well as IARC General

Services, including the Laboratory Services and Biobank team led by Zisis Kozlakidis, the Section of Support to Research overseen by Tamas Landesz and the Evidence Synthesis and Classification Section led by Ian Cree, under IARC regular budget funding. The authors would like to thank Gislaine Bergo, Riley Cox and Juliana Oliveira for help with data/sample preparation and processing. The authors would also like to acknowledge the contributions of the Leeds Biobanking and Sample Processing Lab, the Leeds Multidisciplinary RTB and the Leeds NIHR BioRTB for provision of samples. The authors would like to thank Peter Campbell, Inigo Martincorena, Tim Butler, Daniela Mariosa, Laura Torrens Fontanals, Wellington Oliveira Dos Santos, Hana Zahed, Marc Gunter, Maggie Blanks and Mimi McCord for useful discussions. The authors would also like to thank all the patients and their families involved in this study.

Funding

This work was delivered as part of the Mutographs team supported by the Cancer Grand Challenges partnership funded by Cancer Research UK (C98/A24032). This work was supported by the Wellcome Trust grants 206194 and 220540/Z/20/A. The work was also partly funded by Barretos Cancer Hospital, the Public Ministry of Labor of Campinas (Research, Prevention, and Education of Occupational Cancer, 2015 to R.M.R.), and by Hospital de Clínicas de Porto Alegre (180330 to P.A.-P., M.B., B.S.N.). The work was also partly supported by the Practical Research Project for Innovative Cancer Control from the Japan Agency for Medical Research and Development (AMED) (JP20ck0106547h0001 to T.S.), and by the National Cancer Center Japan Research and Development Fund (2020-A-7 to A.F.). The work was also partly funded by the 1st and 2nd Faculties of Medicine, Charles University, Prague (CAGEKID to I.H.; Occupation, Environment and Kidney Cancer in Central and Eastern Europe to A.H.). The work was also partly supported by the Ministry of Health of the Czech Republic (MH CZ – DRO (MMCI, 00209805) to L.F. and M.N.). Measurement of PFAS compounds was funded by Division of Cancer Epidemiology and Genetics of the National

Cancer Institute (USA). Measurement of cystatin C was funded by Cancer Research UK (C18281/A29019).

Contributions

The study was conceived, designed and supervised by M.R.S., P.B. and L.B.A. Analysis of data was performed by S.Senkin, S.Moody, M.D.-G., T.C., A.F.-I., J.W., S.F., M.K., R.V., A.P.L., E.N.B., A.K., B.O., S.C., E.T., J.A., K.S.-B., R.C.C.P., V.G., D.J., J.W.T. and J.M. Pathology review was carried out by B.A.-A., S.F. and M.A. Sample manipulation was carried out by C.L., C.C. and P.C. Patient and sample recruitment was led or facilitated by S.Sangkhathat, W.A., B.S., S.J., R.S., D.M., V.Jinga, S.R., S.Milosavljevic, M.M., S.Savic, J.M.S.B, M.A., L.P., P.A.-P., M.B., B.S.N., S.M.B., M.P.C., S.C.Z., R.M.R., E.F., N.S.M., R.S.F., R.B., N.V., D.Z., A.M., O.S., V.M., L.F., M.N., I.H., A.H., V.Janout, S.C. and C.L., M.P. P.K.-R., S.C., M.P., P.M.U. and M.J. contributed to data generation. Patient and sample recruitment for Japanese cases was led by T.S. and A.F. Scientific project management was carried out by L.H., E.C., G.S., A.C.D.C., A.F.-I. and S.P. S.Moody and S.Senkin jointly contributed and were responsible for overall scientific coordination. The manuscript was written by S.Senkin, S.Moody, M.R.S. and P.B. with contributions from all other authors.

Competing interests

LBA is a compensated consultant and has equity interest in io9, LLC and Genome Insight. His spouse is an employee of Biotheranostics, Inc. LBA is also an inventor of a US Patent 10,776,718 for source identification by non-negative matrix factorization. ENB and LBA declare U.S. provisional applications with serial numbers: 63/289,601; 63/269,033; and 63/483,237. LBA also declares U.S. provisional applications with serial numbers: 63/366,392; 63/367,846; 63/412,835; and 63/492,348. VM received honoraria from Ipsen, Bayer, AstraZeneca, Janssen, Astellas Pharm and MSD, and provided expert testimony to BMS, Bayer, MSD and Janssen. No other authors declare any competing interests.

1036 **Disclaimer**

1037 Where authors are identified as personnel of the International Agency for Research on Cancer
1038 / World Health Organization, the authors alone are responsible for the views expressed in this
1039 article and they do not necessarily represent the decisions, policy or views of the International
1040 Agency for Research on Cancer / World Health Organization.

1041

1042 **Corresponding author**

1043 Correspondence to Paul Brennan.

1044

1045 **EXTENDED DATA FIGURE AND TABLE LEGENDS**

1046 **Extended Data Fig. 1: Mutation burdens in clear cell renal cell carcinomas across**
1047 **countries.**

1048 Mutation burdens for single base substitutions (SBS) **(a)**, doublet base substitutions (DBS)
1049 **(b)** and small insertions and deletions (ID) **(c)** show significant differences between countries
1050 using the Kruskal-Wallis (two-sided) test (n=961 biologically independent samples). Four
1051 SBS hypermutators and four ID hypermutators above mutation burden of 30000 and 3000,
1052 respectively, were removed for clarity. Box and whiskers plots are in the style of Tukey. The
1053 line within the box is plotted at the median while the upper and lower ends are indicated 25th
1054 and 75th percentiles. Whiskers show 1.5*IQR (interquartile range) and values outside it are
1055 shown as individual data points.

1056

1057 **Extended Data Fig. 2: Principal component analysis of relative mutation counts.**

1058 PCA performed on relative mutation counts of all ccRCC tumors incorporating the six
1059 mutation classes (C>A, C>G, C>T, T>A, T>C, T>G). Principal component 1 (PC1) clearly
1060 separates the cluster of mostly Romanian cases that are enriched with AA signatures, often
1061 at high mutation burdens. Principal component 3 (PC3) identifies a cluster of mostly
1062 Japanese cases, enriched with signature SBS12.

1063

1064 **Extended Data Fig 3: Attribution of signatures SBS40a, SBS40b, and SBS40c in a pan-**
1065 **cancer cohort.**
1066 Attribution of signatures SBS40a, SBS40b, and SBS40c in a pan-cancer cohort, showing a
1067 widespread distribution for SBS40a whilst SBS40b and SBS40c are only seen consistently in
1068 clear cell renal cell carcinomas (ccRCC). The size of each dot represents the proportion of
1069 samples of each tumor type where the signature is present. The color of each dot represents
1070 the average mutation burden.
1071
1072 **Extended Data Fig. 4: Doublet-base substitution signatures operative in clear cell**
1073 **renal cell carcinomas.**
1074 **(a)** Tumour mutation burden (TMB) plot showing the frequency and mutations per Mb for
1075 each of the decomposed DBS signatures. **(b)** Average relative attribution for doublet-base
1076 substitution (DBS) signatures across countries. Signatures contributing less than 5% on
1077 average are grouped in the 'Other' category, apart from signature DBS_D. Category named
1078 '<95% confidence' accounts for the proportion of mutation burden which could not be
1079 assigned to any signature with confidence level of at least 95%. **(c)** Decomposed DBS
1080 signatures, including reference COSMIC signatures as well as *de novo* signatures not
1081 decomposed into COSMIC reference signatures.
1082
1083 **Extended Data Fig. 5: Small insertions and deletion signatures operative in clear cell**
1084 **renal cell carcinomas.**
1085 **(a)** Tumour mutation burden (TMB) plot showing the frequency and mutations per Mb for
1086 each of the decomposed ID signatures. **(b)** Average relative attribution for small insertion and
1087 deletion (ID) signatures across countries. Signatures contributing less than 5% on average
1088 are grouped in the 'Others' category, apart from signature ID_C. Category named '<95%
1089 confidence' accounts for the proportion of mutation burden which could not be assigned to
1090 any signature with confidence level of at least 95%. **(c)** Decomposed ID signatures, including

reference COSMIC signatures as well as *de novo* signatures not decomposed into COSMIC reference signatures.

Extended Data Fig. 6: Correlation amongst signatures SBS22a, SBS22b, DBS_D, ID_C.

Extended Data Table 1: Presence of signatures SBS22a, SBS22b, DBS_D, ID_C across countries.

Extended Data Fig. 7: Single base substitution signatures showing significant differences in attributed mutation burden between countries.

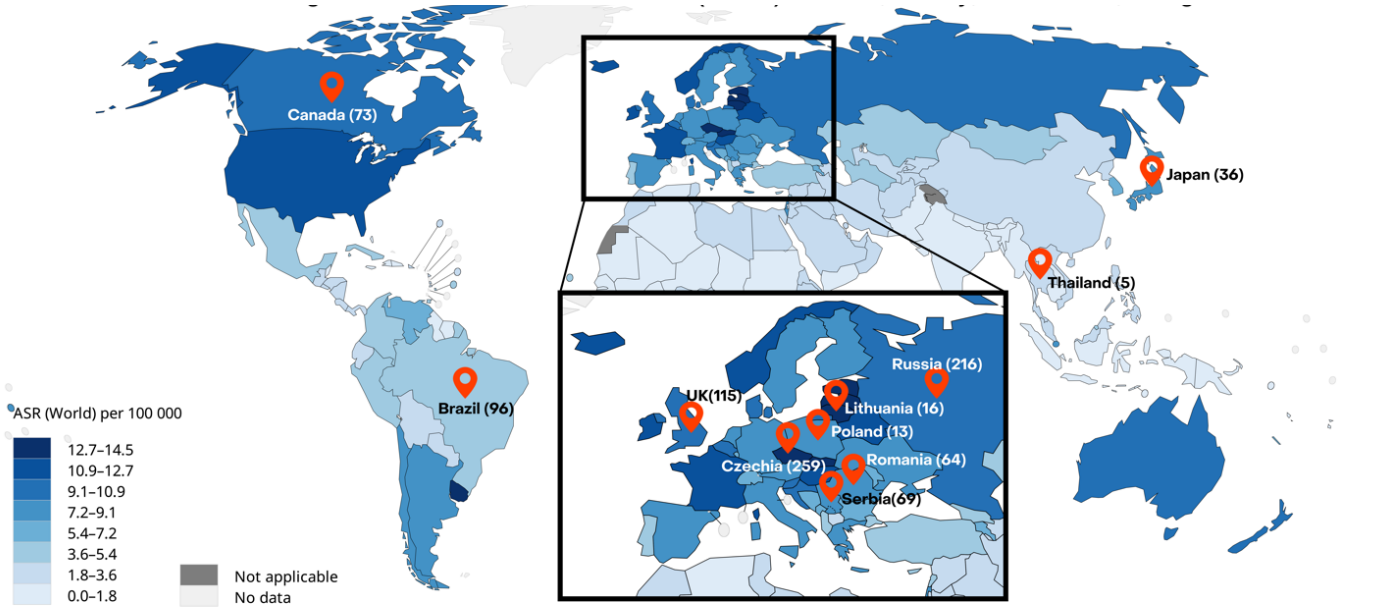
Signatures SBS40a **(a)** and SBS40b **(b)** were more prevalent in high-incidence regions of Czech Republic and Lithuania. Signatures SBS22a **(c)** and SBS22b **(d)** were enriched in Romania and Serbia. SBS1 **(e)**, SBS5 **(f)** and SBS4 **(g)** showed moderate differences across countries. Signature SBS12 **(h)** is highly prevalent in Japan. Five SBS1 hypermutators above mutation burden of 1000 were removed for clarity. Box and whiskers plots are in the style of Tukey. The line within the box is plotted at the median while the upper and lower ends are indicated 25th and 75th percentiles. Whiskers show 1.5*IQR (interquartile range) and values outside it are shown as individual data points.

Extended Data Fig. 8: Association of mutational signatures with incidence of renal cancer.

Number of mutations attributed to signatures **(a)** SBS40a, **(b)** ID5 and **(c)** ID8 against age-standardized incidence rate (ASR) of kidney cancer in each of the eleven countries represented in the cohort. Error bars represent standard errors of the mean. The p-values shown are for the ASR variable in linear regressions across all cases, adjusted for sex and age of diagnosis.

1119 **Extended Data Fig. 9: Association of mutation burden with incidence of renal cancer.**
1120 Association of age-standardized rates (ASR) of kidney cancer incidence with SBS (**a)**, DBS
1121 (**b)** and ID (**c)** mutation burdens across countries. Error bars represent standard errors of the
1122 mean. The p-values shown are for the ASR variable in linear regressions across all cases,
1123 adjusted for sex and age of diagnosis.
1124
1125 **Extended Data Fig. 10: Evolutionary analysis of mutational signatures in ccRCC.**
1126 Comparison of mutational signatures between clonal and subclonal mutations. Lines show
1127 the change in relative activity between the clonal mutations (main) and subclonal mutations
1128 (sub) within a sample. Blue and red lines represent an activity change of more than 6% (blue
1129 indicates higher in the clonal mutations; red indicates higher in the subclonal mutations). Bar
1130 plots show the distribution of activities in samples where the signature was present in the
1131 clonal and/or subclonal mutations; this number is represented in the title of each plot as
1132 X/223 for each signature. Black bars indicate one standard deviation away from the mean.
1133 Significance was assessed using a two-sided Wilcoxon signed-rank test, and q-values were
1134 generated using the Benjamini-Hochberg Procedure.

Fig.1



All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.

Data source: GLOBOCAN 2020
Map production: IARC
(http://gco.iarc.fr/today)
World Health Organization

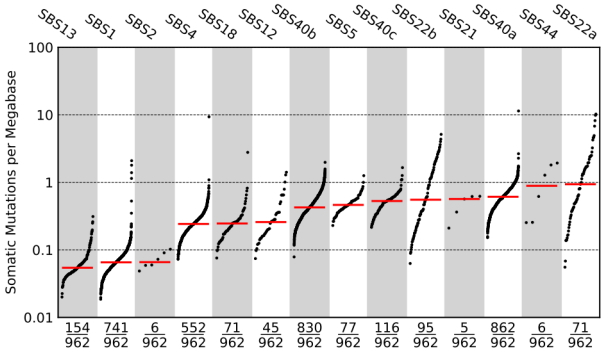
World Health Organization
© International Agency for Research on Cancer 2020
All rights reserved

Table 1

Country (ASR/100,000)		**Brazil (4.5)**	**Canada (10.4)**	**Czechia (14.4)**	**Japan (7.6)**	**Lithuania (14.5)**	**Poland (8.1)**	**Romania (7.7)**	**Russia (10.3)**	**Serbia (7.4)**	**Thailand (1.8)**	**UK (10.3)**	**Total (4.6)**
Number of cases		96	73	259	36	16	13	64	216	69	5	115	962
Sex	Female	44	22	93	8	9	5	25	98	30	4	42	380
	Male	52	51	166	28	7	8	39	118	39	1	73	582
Age at diagnosis (years)	0-45	15	6	27	3	1	2	6	43	16	0	6	125
	45-55	20	17	51	5	0	6	10	44	11	0	22	186
	55-65	30	17	77	8	9	1	20	91	27	2	41	323
	65-75	24	27	72	13	4	4	20	32	9	2	31	238
	75+	7	6	32	7	2	0	8	6	6	1	15	90
Year of recruitment	1999-2005			93			13	14	18				138
	2005-2010			111				19	70	1		31	232
	2010-2015		9	55	28			31	116	68		41	348
	2015-2020	96	64		8	16			12		5	43	244
Stage	I	28	3	123	24	6	0	33	94	32		53	396
	II	2	0	42	1	0	6	12	24	4		8	99
	III	16	23	46	6	5	5	18	65	26		38	248
	IV	7	10	38	5	2	2	1	33	7		16	121
	Missing	43	37	10		3					5		98
Body mass index	<20	3	2	5	2	0	2	2	9	8	0	6	39
	20-25	21	10	100	25	2	3	17	84	28	3	23	316
	25-30	35	24	85	7	6	6	30	40	20	1	45	299
	>30	37	37	69	2	8	2	14	83	13	1	41	307
	Missing							1					1
Hypertension	No	45	28	129	16	5	9	39	125	28	2	58	484
	Yes	51	44	130	20	10	4	24	91	41	3	56	474
	Missing		1			1		1				1	4
Diabetes	No	76	55	130	29	9		45	186	61	3	95	689
	Yes	20	16	36	7	7		4	12	8	2	20	132
	Missing		2	93			13	15	18				141
Family history of ccRCC	No	90	42	165	35	16		54	192	67	5	102	726
	Yes	5	4	22	1	0		1	6	2	0	3	43
	Missing	1	27	72			13	9	18			10	193
Tobacco status	Current smoker	23	21	66	9	4	6	11	52	18	1	28	239
	Ex-smoker	21	30	62	15	3	3	15	27	15	0	44	235
	Never	52	22	131	11	9	4	37	137	36	4	43	486
	Missing				1			1					2
PFOA (ng/mL)	Mean	0.7	1.6	3.4		1.3	5.4	1.3	1.5	1.3	2.2	3.3	2.2
	(st. dev.)	(0.5)	(1.1)	(2.1)		(0.6)	(4.1)	(0.9)	(1.4)	(0.6)	(2.2)	(1.7)	(1.9)

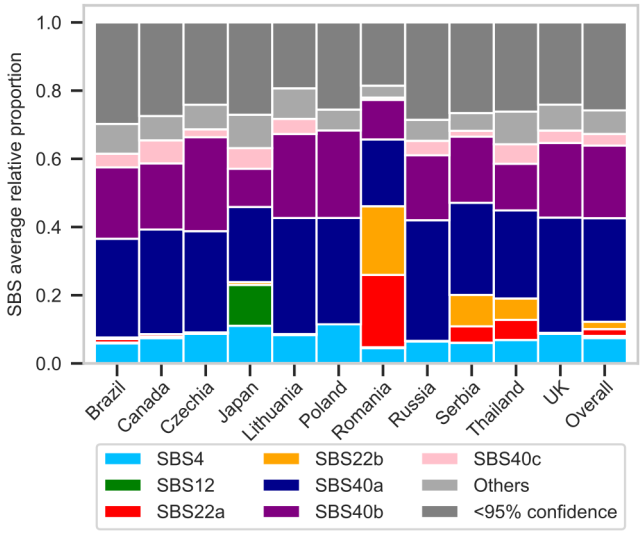
Fig. 2

a



*Showing samples with counts more than 0

b



c

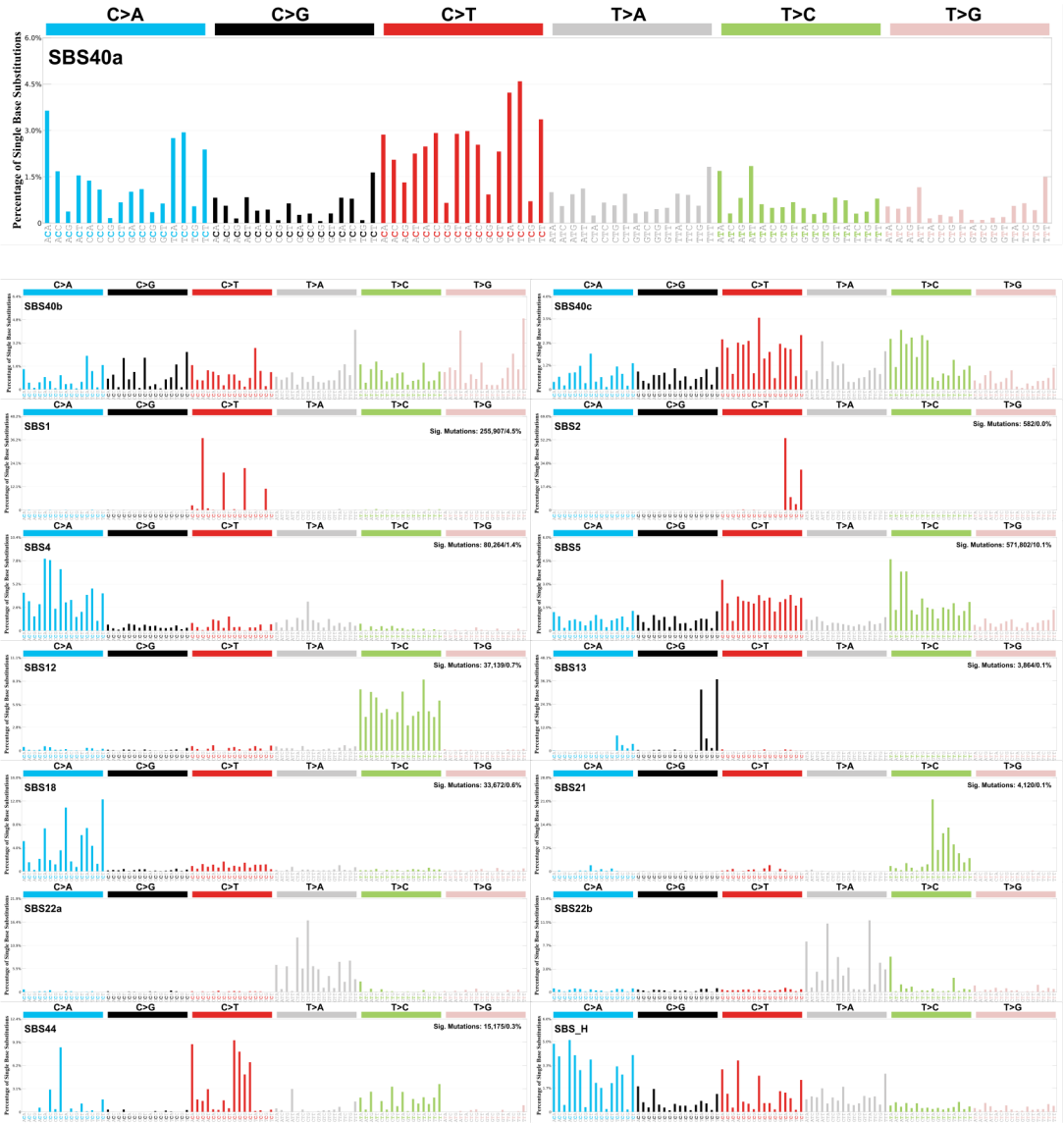


Fig. 3

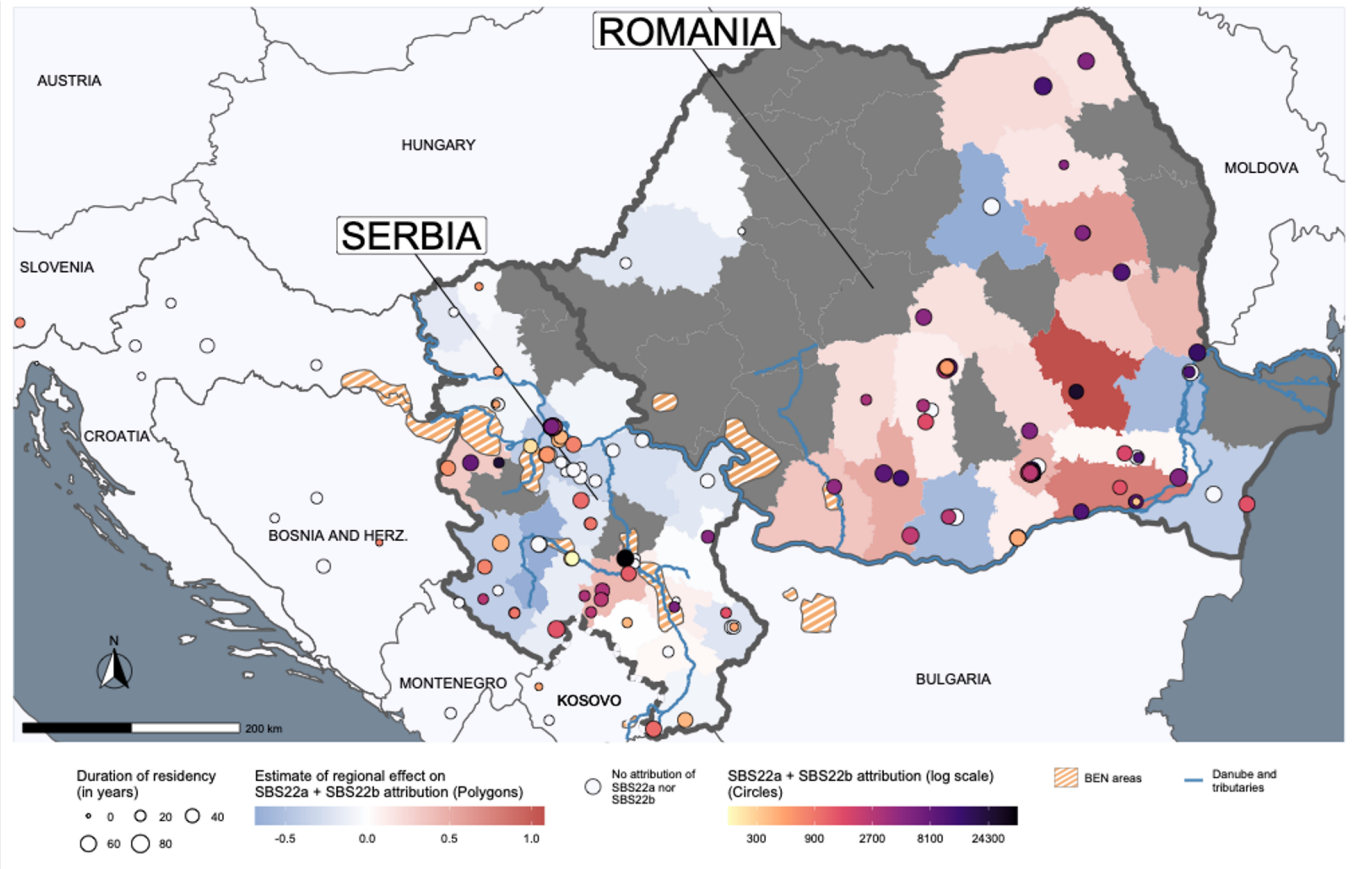

ROMANIA
SERBIA
AUSTRIA
HUNGARY
MOLDOVA
SLOVENIA
CROATIA
BOSNIA AND HERZ.
MONTENEGRO
KOSOVO
BULGARIA
N
200 km
Duration of residency (in years)
0
20
40
60
80
Estimate of regional effect on SBS22a + SBS22b attribution (Polygons)
-0.5
0.0
0.5
1.0
No attribution of SBS22a nor SBS22b
SBS22a + SBS22b attribution (log scale) (Circles)
300
900
2700
8100
24300
BEN areas
Danube and tributaries

Fig. 4

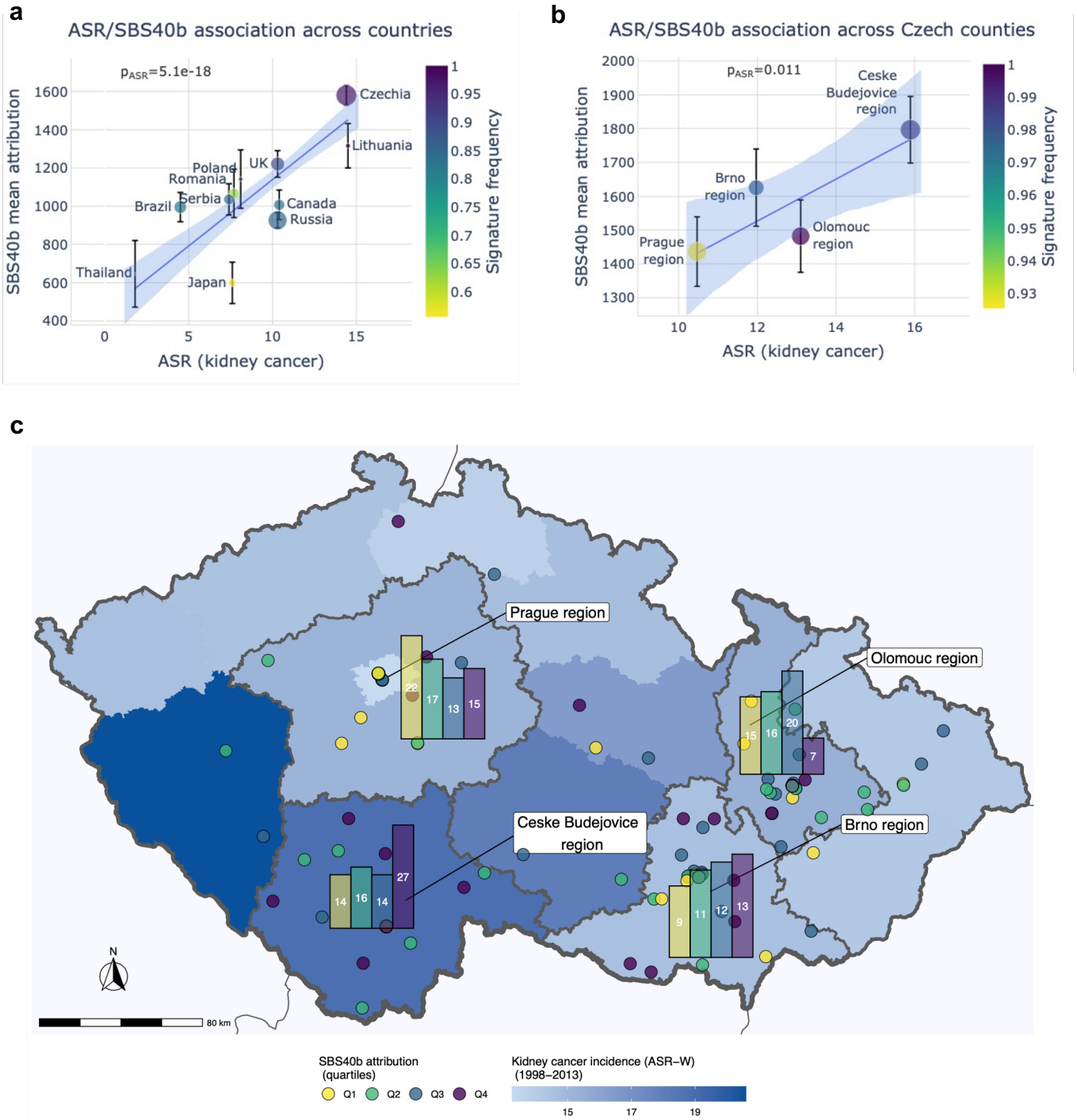
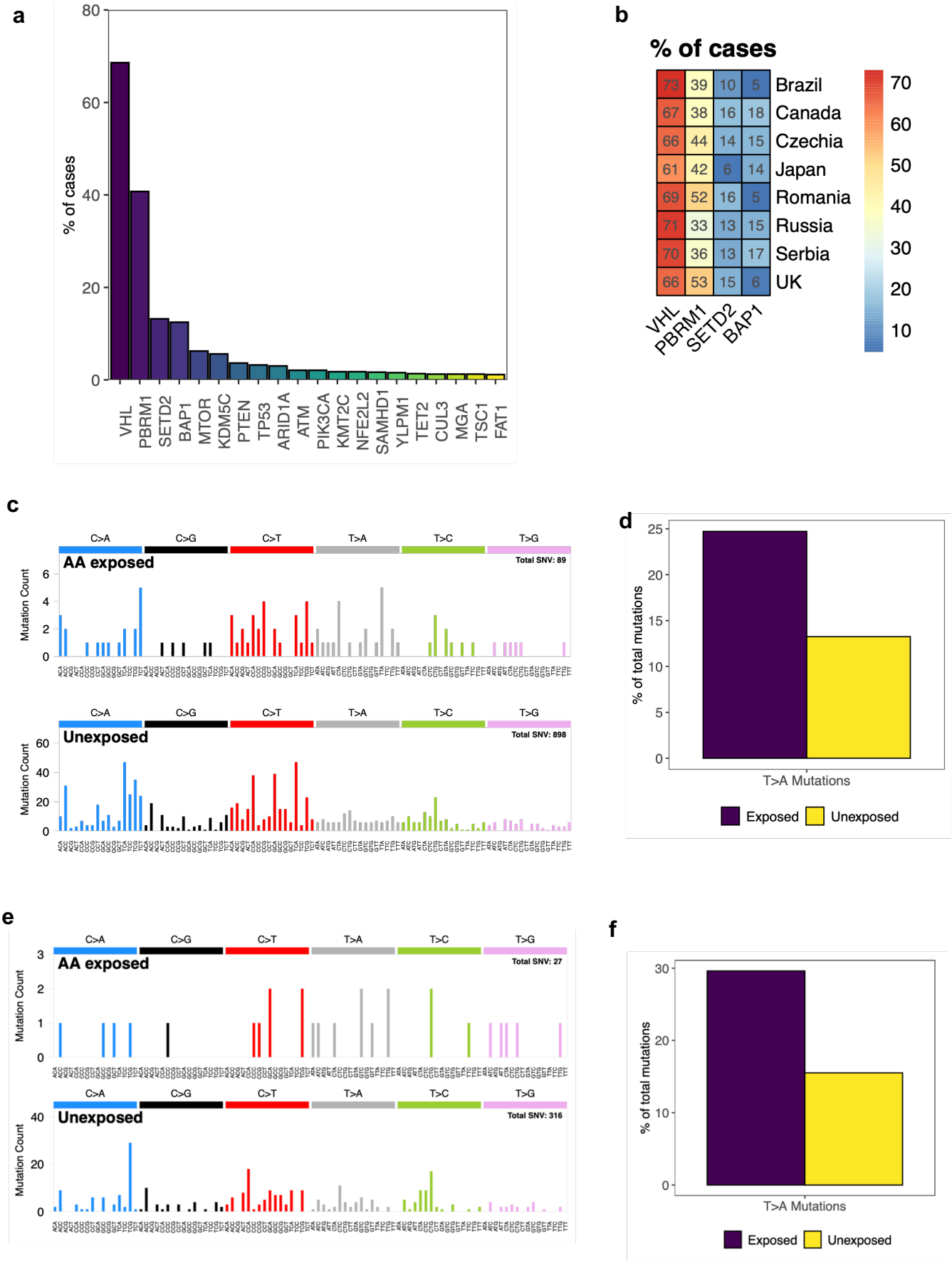
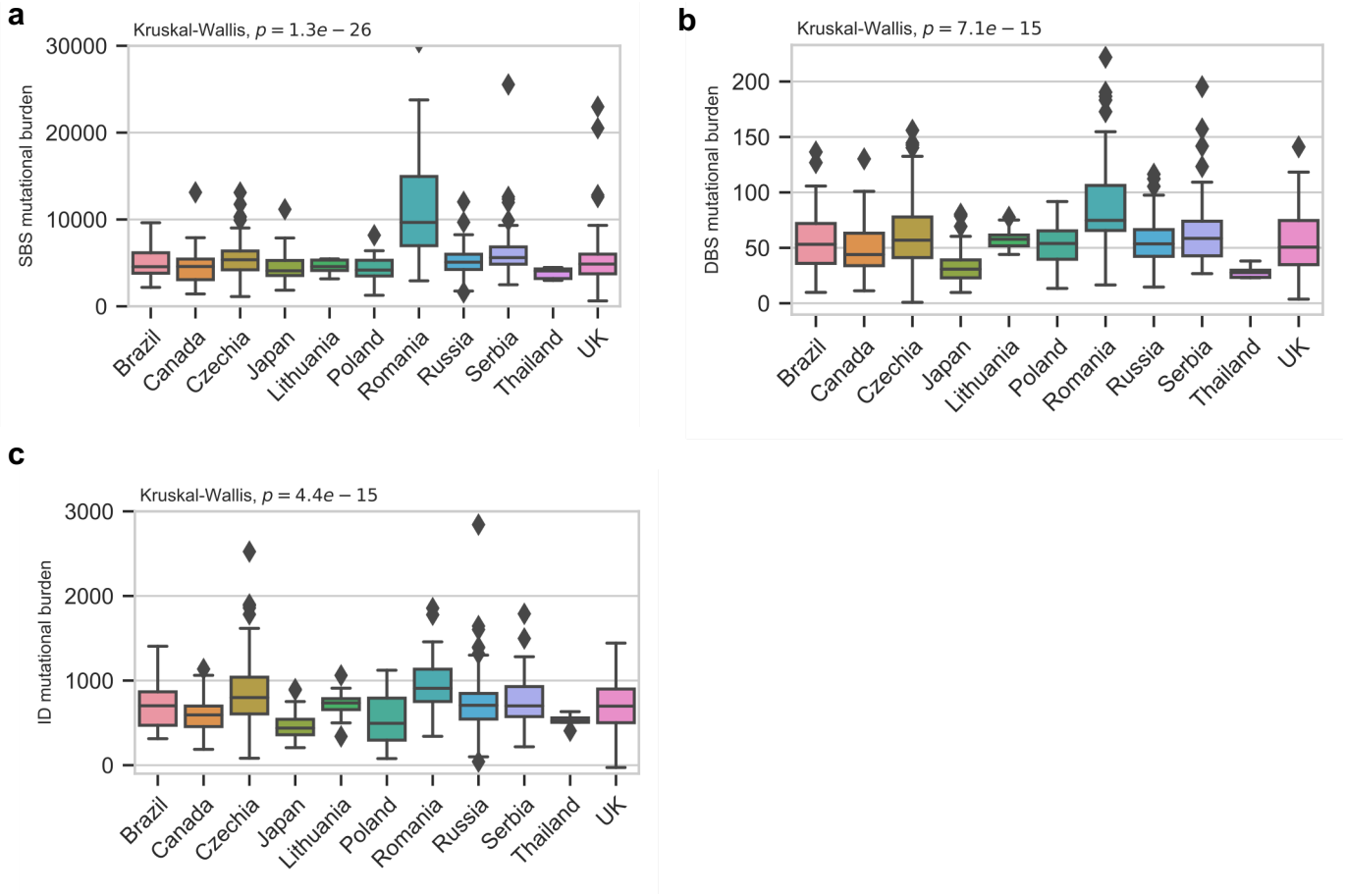

a
ASR/SBS40b association across countries
pASR=5.1e-18
SBS40b mean attribution
ASR (kidney cancer)
Signature frequency
Czechia
Lithuania
UK
Poland
Romania
Serbia
Brazil
Canada
Russia
Thailand
Japan
b
ASR/SBS40b association across Czech counties
pASR=0.011
Ceske Budejovice region
Brno region
Olomouc region
Prague region
c
Prague region
Olomouc region
Ceske Budejovice region
Brno region
N
80 km
SBS40b attribution (quartiles)
Q1
Q2
Q3
Q4
Kidney cancer incidence (ASR–W) (1998–2013)

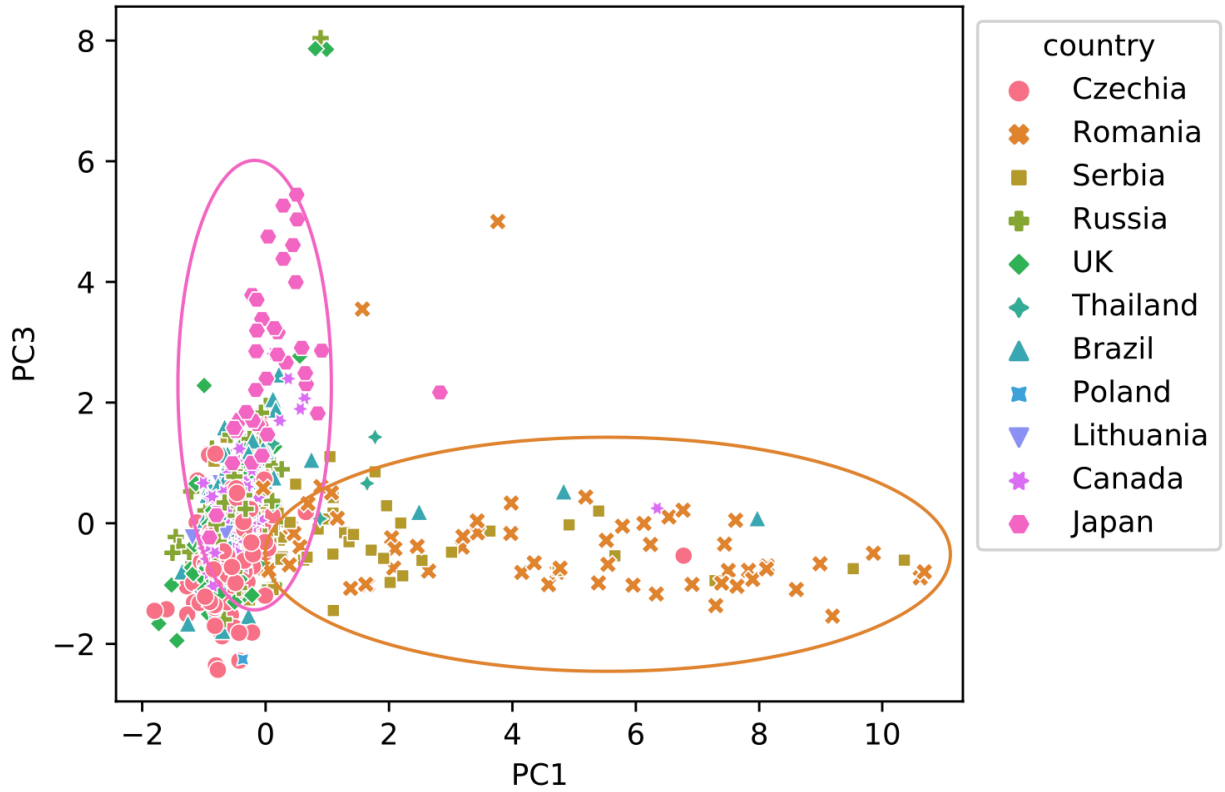
Fig. 5


a
% of cases
VHL
PBRM1
SETD2
BAP1
MTOR
KDM5C
PTEN
TP53
ARID1A
ATM
PIK3CA
KMT2C
NFE2L2
SAMHD1
YLPM1
TET2
CUL3
MGA
TSC1
FAT1
b
% of cases
Brazil
Canada
Czechia
Japan
Romania
Russia
Serbia
UK
VHL
PBRM1
SETD2
BAP1
c
C>A
C>G
C>T
T>A
T>C
T>G
AA exposed
Total SNV: 89
Unexposed
Total SNV: 898
Mutation Count
d
% of total mutations
T>A Mutations
Exposed
Unexposed
e
AA exposed
Total SNV: 27
Unexposed
Total SNV: 316
Mutation Count
f
% of total mutations
T>A Mutations
Exposed
Unexposed

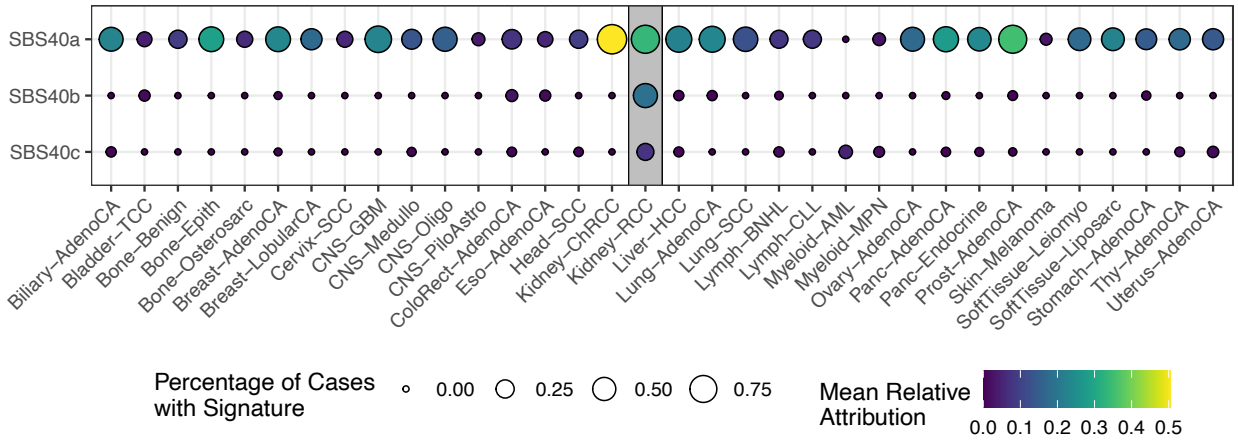
Extended Data Fig. 1.



Extended Data Fig. 2.

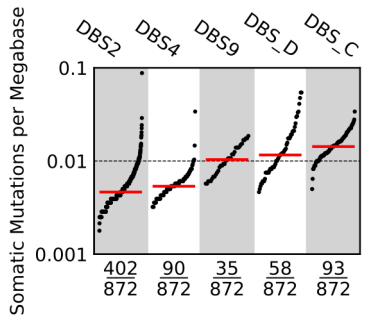


Extended Data Fig. 3.



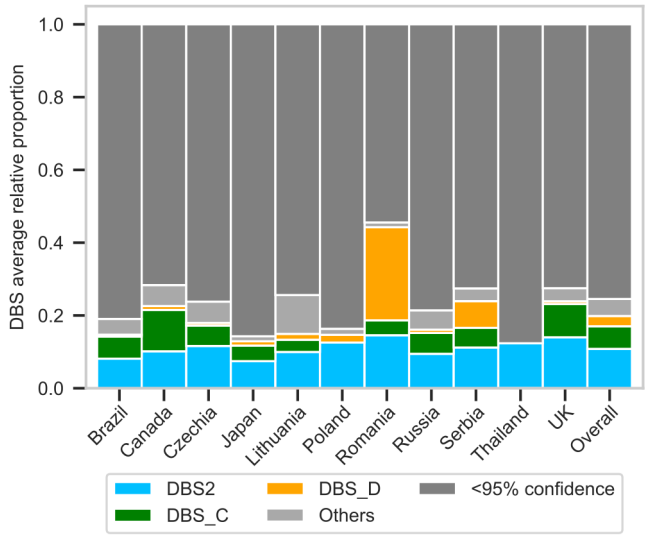
Extended Data Fig. 4.

a

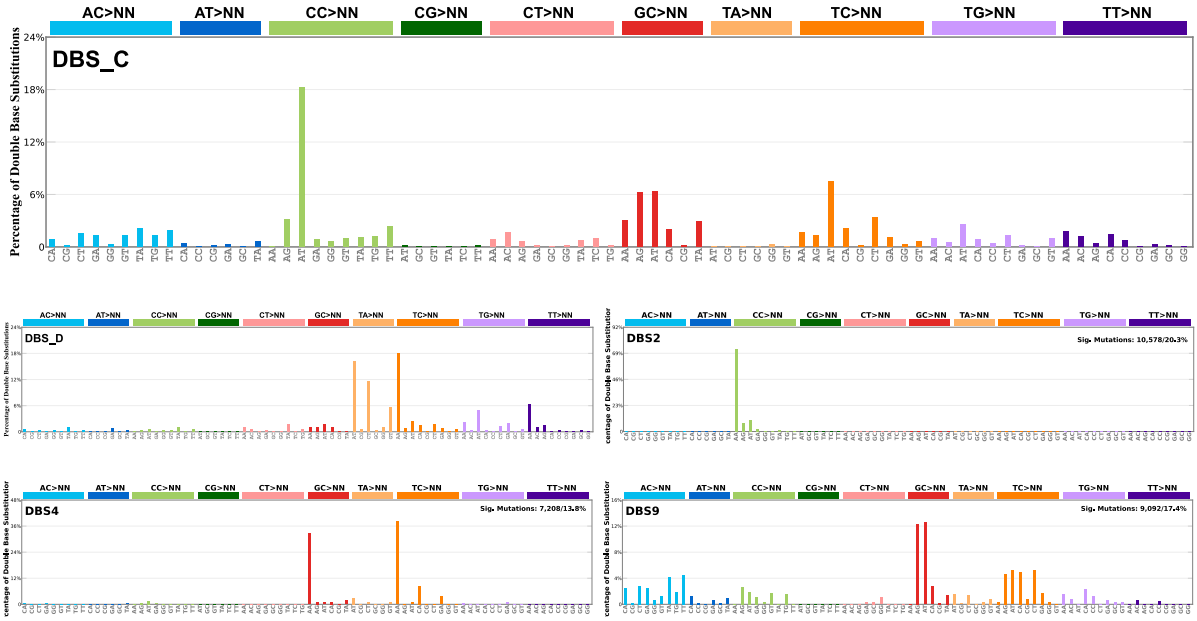


*Showing samples with counts more than 0

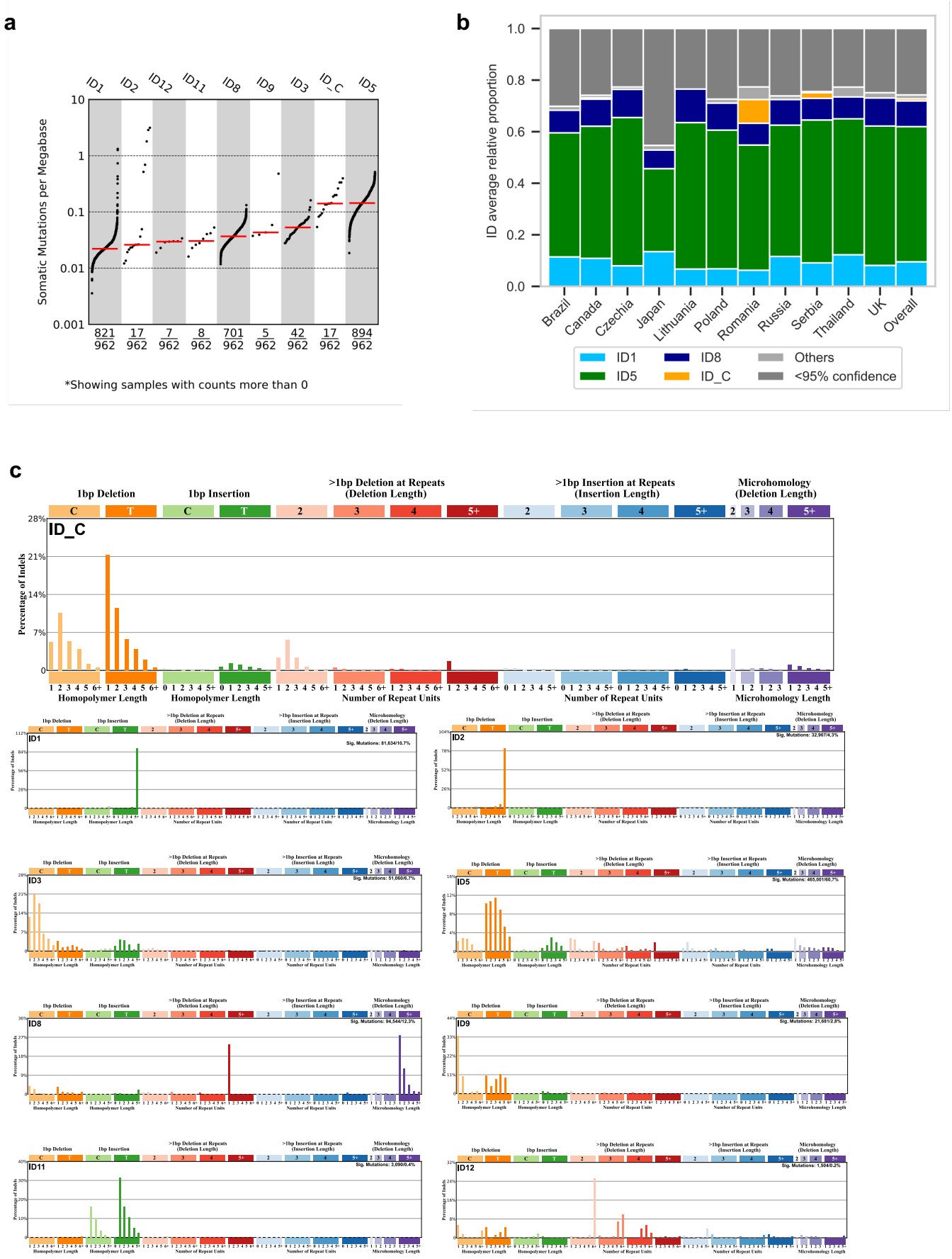
b



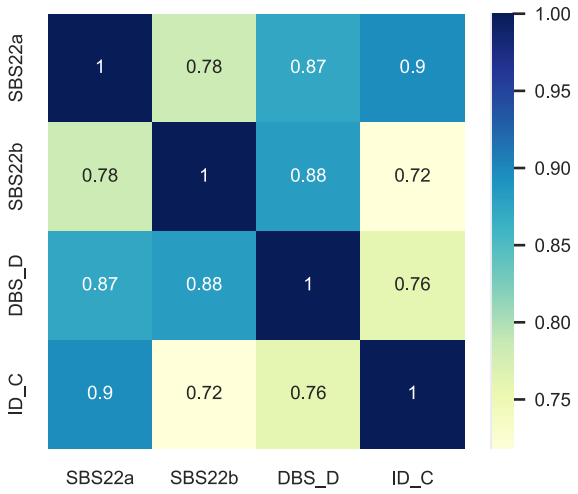
c



Extended Data Fig. 5.



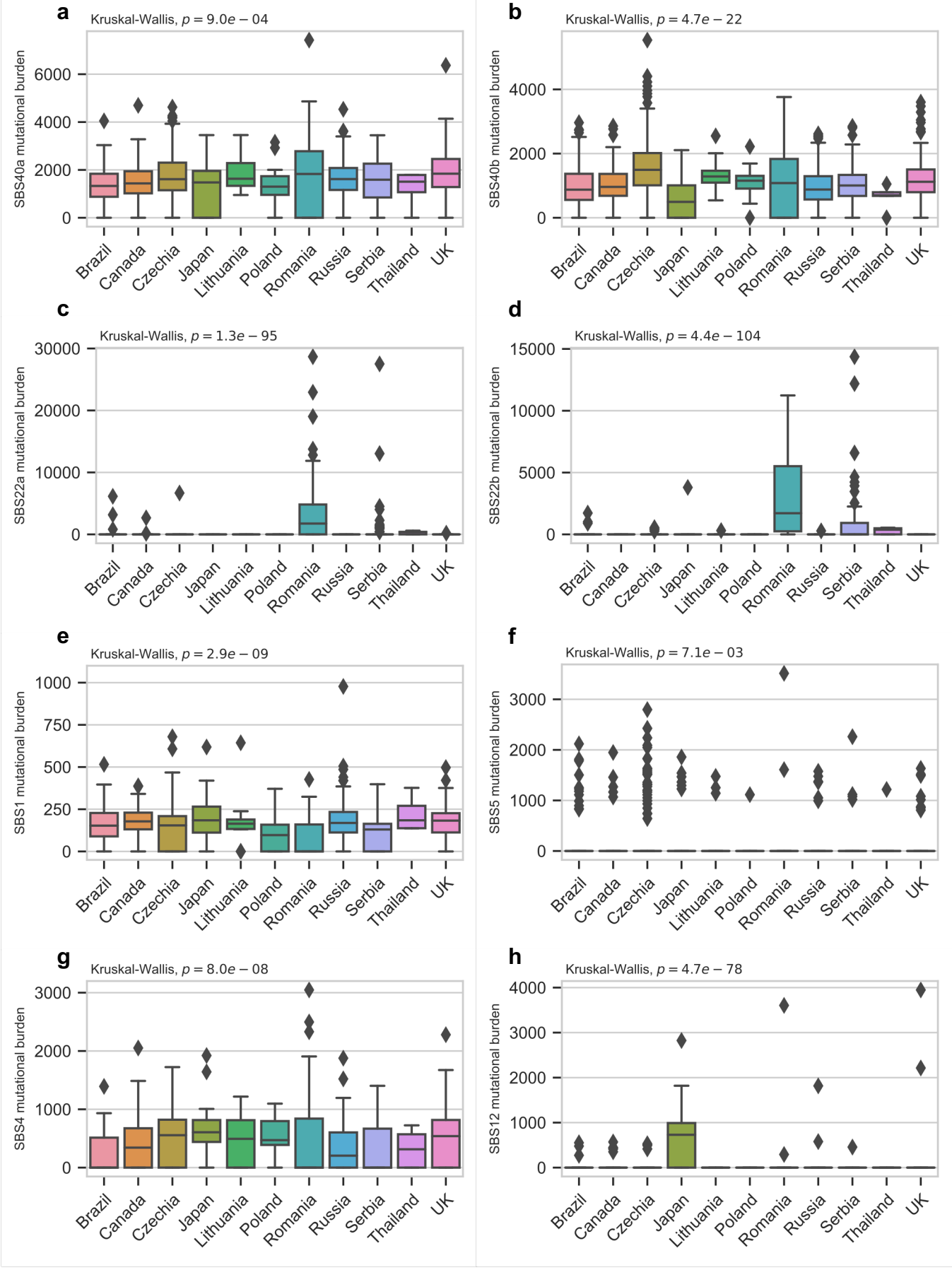
Extended Data Fig. 6.



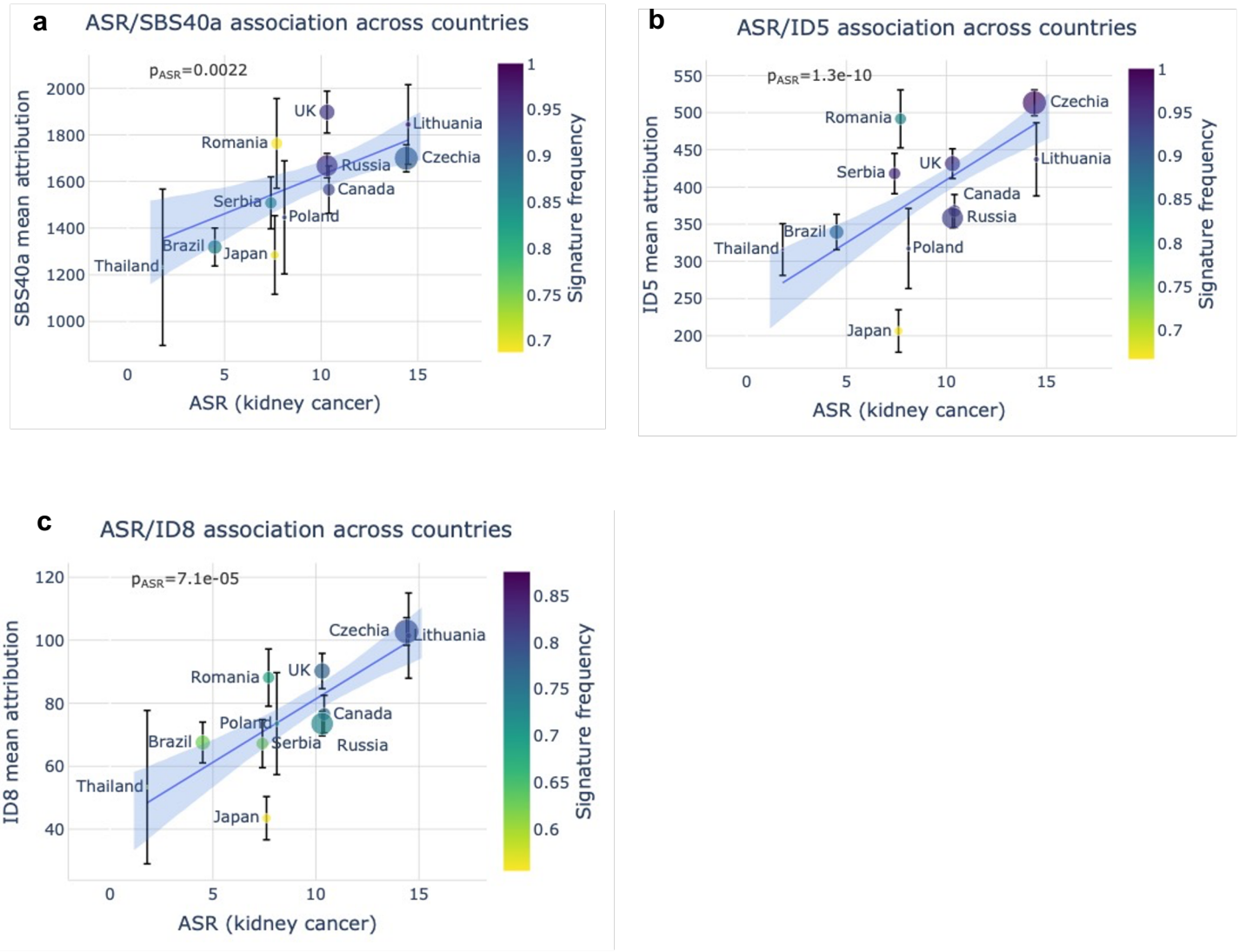
Extended Data Table 1.

Country	N cases	SBS22a (%)	SBS22b (%)	DBS_D (%)	ID_C (%)	SBS22a or SBS22b (%)	Any (%)
Romania	64	45 (70.3)	48 (75.0)	42 (65.6)	13 (20.3)	53 (82.8)	54 (84.4)
Serbia	69	16 (23.2)	33 (47.8)	11 (15.9)	3 (4.3)	35 (50.7)	36 (52.2)
Thailand	5	3 (60.0)	3 (60.0)	0 (0.0)	0 (0.0)	4 (80.0)	4 (80.0)
Brazil	96	3 (3.1)	3 (3.1)	1 (1.0)	0 (0.0)	3 (3.1)	3 (3.1)
Canada	73	2 (2.7)	0 (0.0)	2 (2.7)	1 (1.4)	2 (2.7)	3 (4.1)
Czechia	259	1 (0.4)	5 (1.9)	32 (12.4)	0 (0.0)	6 (2.3)	37 (14.3)
UK	115	1 (0.9)	0 (0.0)	31 (27.0)	0 (0.0)	1 (0.9)	31 (27.0)
Russia	216	0 (0.0)	1 (0.5)	26 (12.0)	0 (0.0)	1 (0.5)	27 (12.5)
Poland	13	0 (0.0)	0 (0.0)	1 (7.7)	0 (0.0)	0 (0.0)	1 (7.7)
Lithuania	16	0 (0.0)	1 (6.2)	1 (6.2)	0 (0.0)	1 (6.2)	2 (12.5)
Japan	36	0 (0.0)	1 (2.8)	1 (2.8)	0 (0.0)	1 (2.8)	1 (2.8)

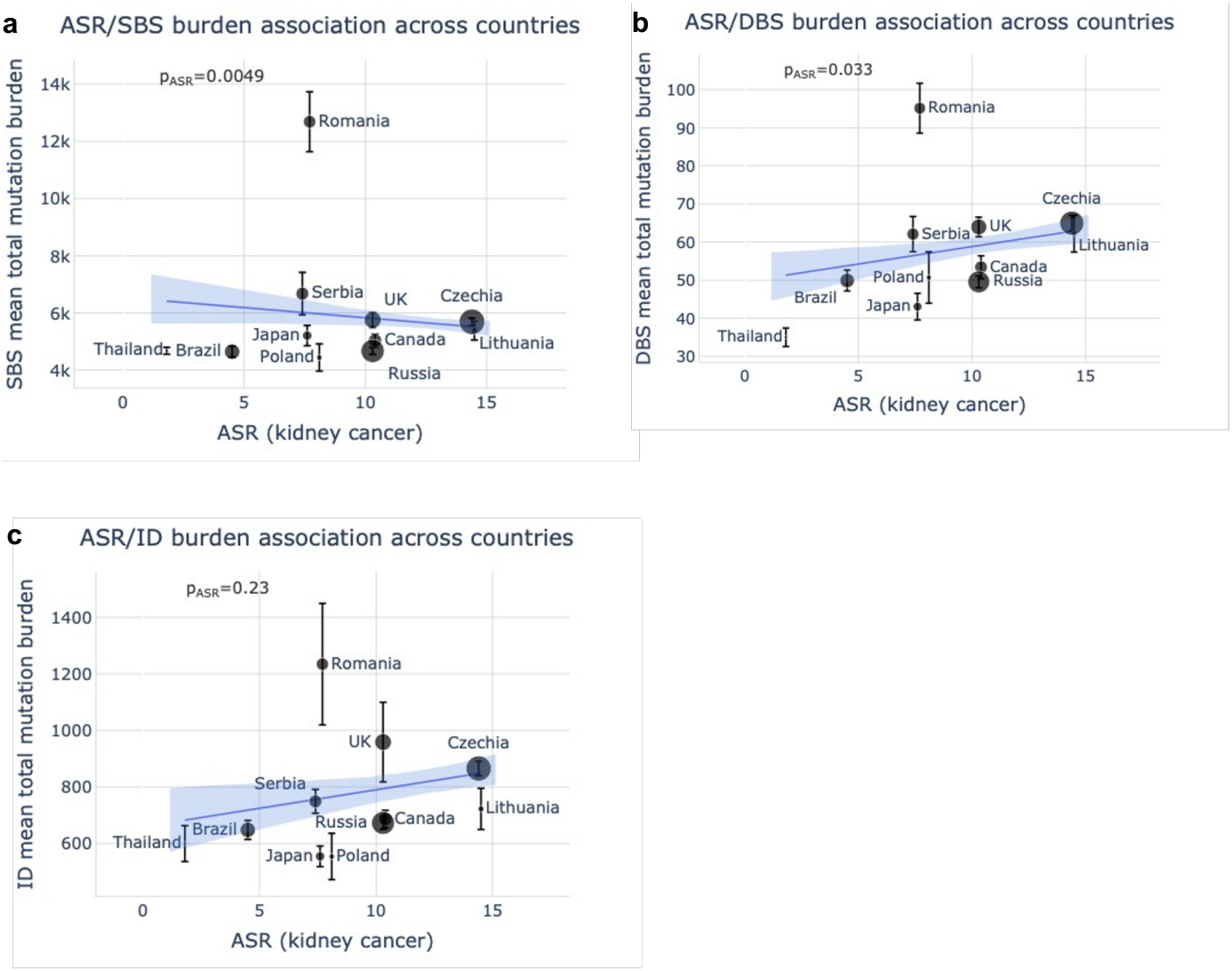
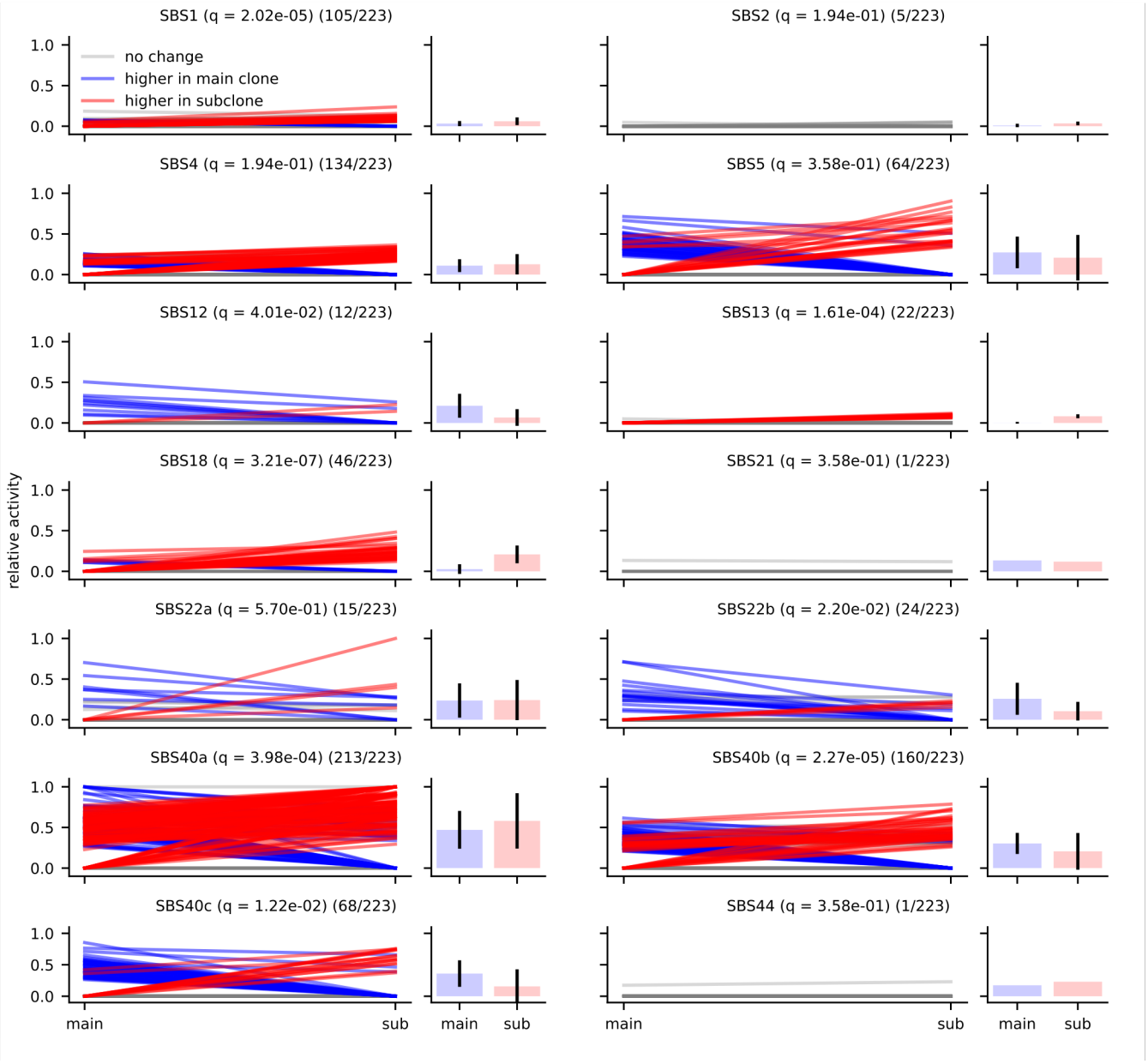
Extended Data Fig. 7.



Extended Data Fig. 8.



Extended Data Fig. 9.



Extended Data Fig. 10.