Article

Geographic variation of mutagenic exposures in kidney cancer genomes

https://doi.org/10.1038/s41586-024-07368-2

Received: 4 May 2023

Accepted: 28 March 2024

Published online: 1 May 2024

Open access

Check for updates

Sergey Senkin[1,58], Sarah Moody[2,58], 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[3,5,6], 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[12,13], Surasak Sangkhathat[14], Worapat Attawettayanon[15], Beata Świątkowska[16], Sonata Jarmalaite[17,18], Rasa Sabaliauskaite[17], Tatsuhiro Shibata[19,20], Akihiko Fukagawa[20,21], Dana Mates[22], Viorel Jinga[23], Stefan Rascu[23], Mirjana Mijuskovic[24], Slavisa Savic[25], Sasa Milosavljevic[26], John M. S. Bartlett[27], Monique Albert[28,29], Larry Phouthavongsy[29], Patricia Ashton-Prolla[30,31], Mariana R. Botton[32], Brasil Silva Neto[33,34], Stephania Martins Bezerra[35], Maria Paula Curado[36], Stênio de Cássio Zequi[37,38,39,40], Rui Manuel Reis[41,42], Eliney Ferreira Faria[43,44], Nei Soares de Menezes[45], Renata Spagnoli Ferrari[44], Rosamonde E. Banks[46], Naveen S. Vasudev[46], David Zaridze[47], Anush Mukeriya[47], Oxana Shangina[47], Vsevolod Matveev[48], Lenka Foretova[49], Marie Navratilova[49], Ivana Holcatova[50,51], Anna Hornakova[52], Vladimir Janout[53], Mark P. Purdue[54], Nathaniel Rothman[54], Stephen J. Chanock[54], Per Magne Ueland[55], Mattias Johansson[1], James McKay[1], Ghislaine Scelo[56], Estelle Chanudet[57], Laura Humphreys[2], Ana Carolina de Carvalho[1], Sandra Perdomo[1], Ludmil B. Alexandrov[3,4,5], Michael R. Stratton[2] & Paul Brennan[1] ✉

International differences in the incidence of many cancer types indicate the existence of carcinogen exposures that have not yet been identified by conventional epidemiology make a substantial contribution to cancer burden[1]. In clear cell renal cell carcinoma, obesity, hypertension and tobacco smoking are risk factors, but they do not explain the geographical variation in its incidence[2]. Underlying causes can be inferred by sequencing the genomes of cancers from populations with different incidence rates and detecting differences in patterns of somatic mutations. Here we sequenced 962 clear cell renal cell carcinomas from 11 countries with varying incidence. The somatic mutation profiles differed between countries. In Romania, Serbia and Thailand, mutational signatures characteristic of aristolochic acid compounds were present in most cases, but these were rare elsewhere. In Japan, a mutational signature of unknown cause was found in more than 70% of cases but in less than 2% elsewhere. A further mutational signature of unknown cause was ubiquitous but exhibited higher mutation loads in countries with higher incidence rates of kidney cancer. Known signatures of tobacco smoking correlated with tobacco consumption, but no signature was associated with obesity or hypertension, suggesting that non-mutagenic mechanisms of action underlie these risk factors. The results of this study indicate the existence of multiple, geographically variable, mutagenic exposures that potentially affect tens of millions of people and illustrate the opportunities for new insights into cancer causation through large-scale global cancer genomics.

The incidence rates of most adult cancers vary substantially between geographical regions and many such differences are not explained by known risk factors[1]. Together with unexplained trends in incidence over time, this indicates the probable presence of unknown environmental or lifestyle causes for many cancer types[1]. Traditional epidemiological studies have identified many important lifestyle, environmental and infectious risk factors for cancer. However, they have had limited success in recent decades, suggesting that alternative study designs are required if further risk factors are to be identified.

Characterization of mutational signatures within cancer genomes[3] is an approach that complements conventional epidemiology for investigating unknown causes of cancer. Most cancers contain thousands of somatic mutations that have occurred over the lifetime of the individual. These can be caused by endogenous cellular processes such as

A list of affiliations appears at the end of the paper.

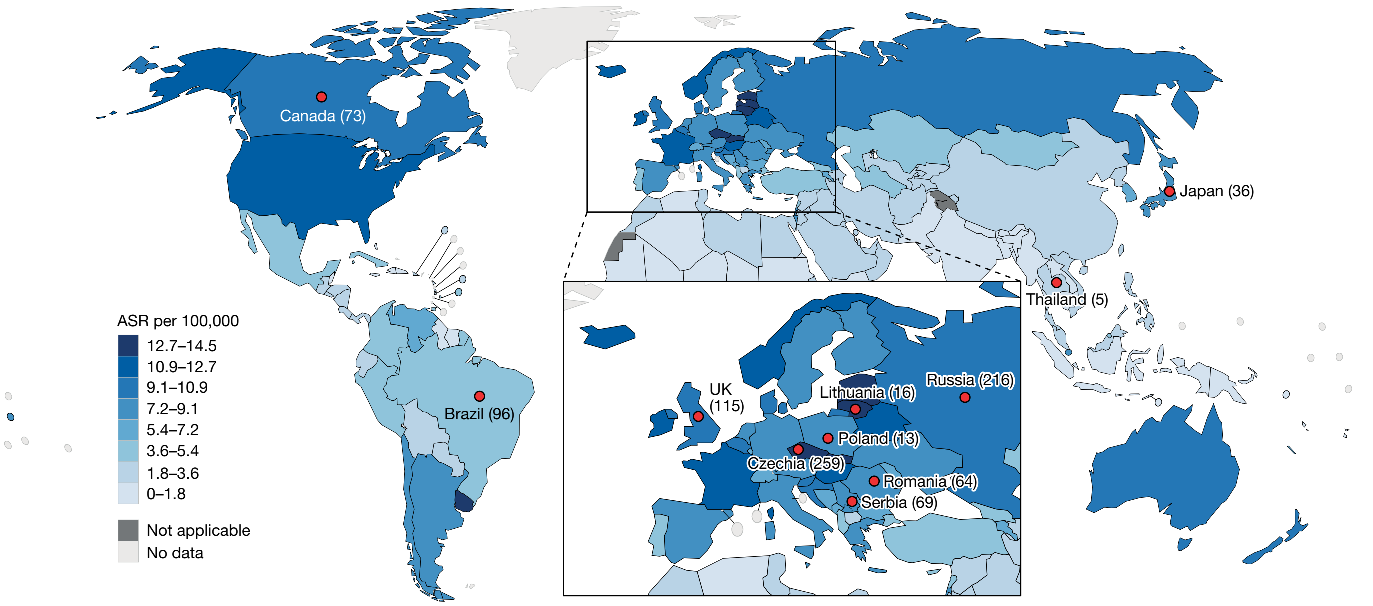


Fig. 1 | Eleven participating countries and estimated ASRs of ccRCCs. Incidence of ccRCC for men and women combined (ASR per 100,000). Data from GLOBOCAN 2020. Markers indicate countries included in this study (number of participants with ccRCC per country).

imperfect DNA replication and repair, or by exposure to exogenous environmental or lifestyle mutagens such as ultraviolet radiation in sunlight and compounds in cigarette smoke. Mutational signatures are the patterns of somatic mutation imprinted on genomes by individual mutational processes. Analysis of thousands of cancer genome sequences from most cancer types has established a set of reference mutational signatures including 71 single base substitution (SBS) or doublet base substitution (DBS) signatures, and 18 small insertion and deletion (indel or ID) signatures[4]. A possible aetiology has been suggested for 47 SBS and DBS signatures and 9 indel signatures.

Kidney cancer has particularly high incidence rates in Central and Northern Europe, notably in the Czech Republic and Lithuania, and has shown increasing incidence in high income countries in recent decades[2] (Fig. 1). Most kidney cancers are clear cell renal cell carcinomas[3] (ccRCCs), for which obesity, hypertension and tobacco smoking are known risk factors[2]. However, these account for less than 50% of the global ccRCC burden and do not explain geographical or temporal incidence trends. Previous ccRCC genome sequencing studies have included relatively small numbers of individuals from a small number of countries with limited variation in ccRCC incidence[5–9] and have not comprehensively examined associations between ccRCC risk factors and mutational signatures. To detect the activity of unknown carcinogens involved in ccRCC development and to investigate the mechanisms of action of known risk factors, we generated and analysed epidemiological and whole-genome sequencing data from a large international ccRCC dataset[10].

A total of 962 cases of ccRCC from 11 countries in 4 continents were studied, including from the Czech Republic (also known as Czechia) ($n = 259$), Russia ($n = 216$), UK ($n = 115$), Brazil ($n = 96$), Canada ($n = 73$), Serbia ($n = 69$), Romania ($n = 64$), Japan ($n = 36$), Lithuania ($n = 16$), Poland ($n = 13$) and Thailand ($n = 5$) (Fig. 1, Table 1 and Methods). These encompass a broad range of ccRCC incidence, from the highest global age-standardized rates (ASRs) of Lithuania and Czech Republic (ASRs of 14.5 and 14.4 per 100,000 respectively) to the relatively low rates of Brazil and Thailand[11] (ASRs of 4.5 and 1.8 per 100,000 respectively). Epidemiological questionnaire data were available on sex, age at diagnosis and important risk factors, including body mass index (BMI), hypertension and tobacco smoking (Table 1). DNA from ccRCCs and blood from the same individuals were extracted and whole-genome sequenced to average coverage of 54-fold and 31-fold, respectively.

Somatic mutation burdens in the 962 ccRCC genomes ranged from 803 to 45,376 (median 5,093) for SBS, 2 to 240 (median 53) for DBS, and 10 to 14,770 (median 695) for indels (Supplementary Table 1). The average burden of all three mutation types differed between the 11 countries (P value < 2×10^{-23}, P value < 2×10^{-14} and P value < 6×10^{-14}, for SBSs, DBSs and indels, respectively). In particular, the burden of all mutation types was higher in Romania compared with other countries (Extended Data Fig. 1). Principal component analysis (PCA) performed on the proportions of the six primary SBS mutation classes (C>A, C>G, C>T, T>A, T>C and T>G) in each sample identified a distinct cluster of mainly Romanian and Serbian cases and a further cluster of mainly Japanese cases (Extended Data Fig. 2). The results, therefore, clearly demonstrate geographical variation of somatic mutation loads and patterns in ccRCC.

To investigate the mutational processes contributing to the geographical variation in mutation burdens, we extracted mutational signatures and estimated the contribution of each signature to each ccRCC genome (Supplementary Tables 2–6). Ten signatures with strong similarity to a reference signature in the Catalogue of Somatic Mutations in Cancer (COSMIC) database were extracted: SBS1, due to deamination of 5-methylcytosine[12]; SBS2 and SBS13, due to cytosine deamination by apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like (APOBEC) DNA-editing enzymes[12]; SBS4, due to tobacco smoke mutagens[13]; SBS5, due to an endogenous mutational process in which mutations accumulate with age[13]; SBS12, of unknown cause; SBS18, due to DNA damage by reactive oxygen species[13]; SBS21 and SBS44, due to defective DNA mismatch repair[13,14]; and SBS22, due to aristolochic acid exposure[15,16].

Five further SBS signatures were identified that were not well described by the COSMIC v3.3 catalogue (Fig. 2 and Supplementary Table 7). SBS_B, SBS_A and SBS_F were present in most ccRCCs, accounting for, on average, around 30%, 20% and 3% of mutations, respectively (Fig. 2b). Combined, they closely resemble the previously reported SBS40 (0.96 cosine similarity), suggesting that the large number of ccRCC whole genomes analysed here provides the power to separate the constituent component signatures of SBS40. This hypothesis was tested by performing a series of extractions using different conditions

Table 1 | Summary of ccRCC risk factors included in this study

Country (ASR per 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)
Total 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	0	0	93	0	0	13	14	18	0	0	0	138
	2005–2010	0	0	111	0	0	0	19	70	1	0	31	232
	2010–2015	0	9	55	28	0	0	31	116	68	0	41	348
	2015–2020	96	64	0	8	16	0	0	12	0	5	43	244
Stage	I	28	3	123	24	6	0	33	94	32	0	53	396
	II	2	0	42	1	0	6	12	24	4	0	8	99
	III	16	23	46	6	5	5	18	65	26	0	38	248
	IV	7	10	38	5	2	2	1	33	7	0	16	121
	Missing	43	37	10	0	3	0	0	0	0	5	0	98
BMI	<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	0	0	0	0	0	0	1	0	0	0	0	1
High blood pressure	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	0	1	0	0	1	0	1	0	0	0	1	4
Diabetes	No	76	55	130	29	9	0	45	186	61	3	95	689
	Yes	20	16	36	7	7	0	4	12	8	2	20	132
	Missing	0	2	93	0	0	13	15	18	0	0	0	141
Family history of RCC	No	90	42	165	35	16	0	54	192	67	5	102	726
	Yes	5	4	22	1	0	0	1	6	2	0	3	43
	Missing	1	27	72	0	0	13	9	18	0	0	10	193
Tobacco smoking status	Current	23	21	66	9	4	6	11	52	18	1	28	239
	Ex	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	0	0	0	1	0	0	1	0	0	0	0	2
PFOA (ng ml^{-1})	Mean (s.d.)	0.7 (0.5)	1.6 (1.1)	3.4 (2.1)		1.3 (0.6)	5.4 (4.1)	1.3 (0.9)	1.5 (1.4)	1.3 (0.6)	2.2 (2.2)	3.3 (1.7)	2.2 (1.9)

PFOA, perfluorooctanoic acid; RCC, renal cell carcinoma.

and subsets of the data (Supplementary Note) which showed that the three extracted signatures were highly reproducible. SBS_B, SBS_A and SBS_F were therefore named SBS40a, SBS40b and SBS40c respectively. SBS40 was previously reported frequently, and at high levels, in kidney cancer, but also in other cancers, and is of unknown aetiology. Similar to the composite SBS40, SBS40a is present in multiple cancer types. However, SBS40b and SBS40c are largely restricted to ccRCC (Extended Data Fig. 3 and Supplementary Note). The two remaining signatures that were not explained were SBS_H and SBS_I, both of which had strong support from individual mutational spectra (Supplementary Note). Analysis of all other types of mutational signatures, including doublet base substitutions, small insertion and deletions, copy number variants and structural variants, is presented in Supplementary Note.

The mutation burdens of multiple SBS mutational signatures varied between the 11 countries. SBS22 is thought to be caused by aristolochic acids, mutagenic derivatives of plants of the *Aristolochia* genus, which are carcinogenic and also cause Balkan endemic nephropathy (BEN), a kidney disease that is prevalent in areas adjacent to the Danube in southeastern Europe[17]. SBS22 has previously been found in ccRCC, other urothelial tract cancers and hepatocellular carcinomas from Romania[5,18] and various countries in East and Southeast Asia[15,16,19]. In this study, SBS22 was present in high proportions of ccRCC from Romania (45 out of 64 (70%)), Serbia (16 out of 69 (23%)) and Thailand (3 out of 5 (60%)), often with very high mutation burdens. Of note, given the limited number of cases in Thailand, they may not be representative of ccRCC in that region. The presence of SBS22 was strongly correlated with that of new signatures SBS_I, DBS_D and ID_C (Extended Data

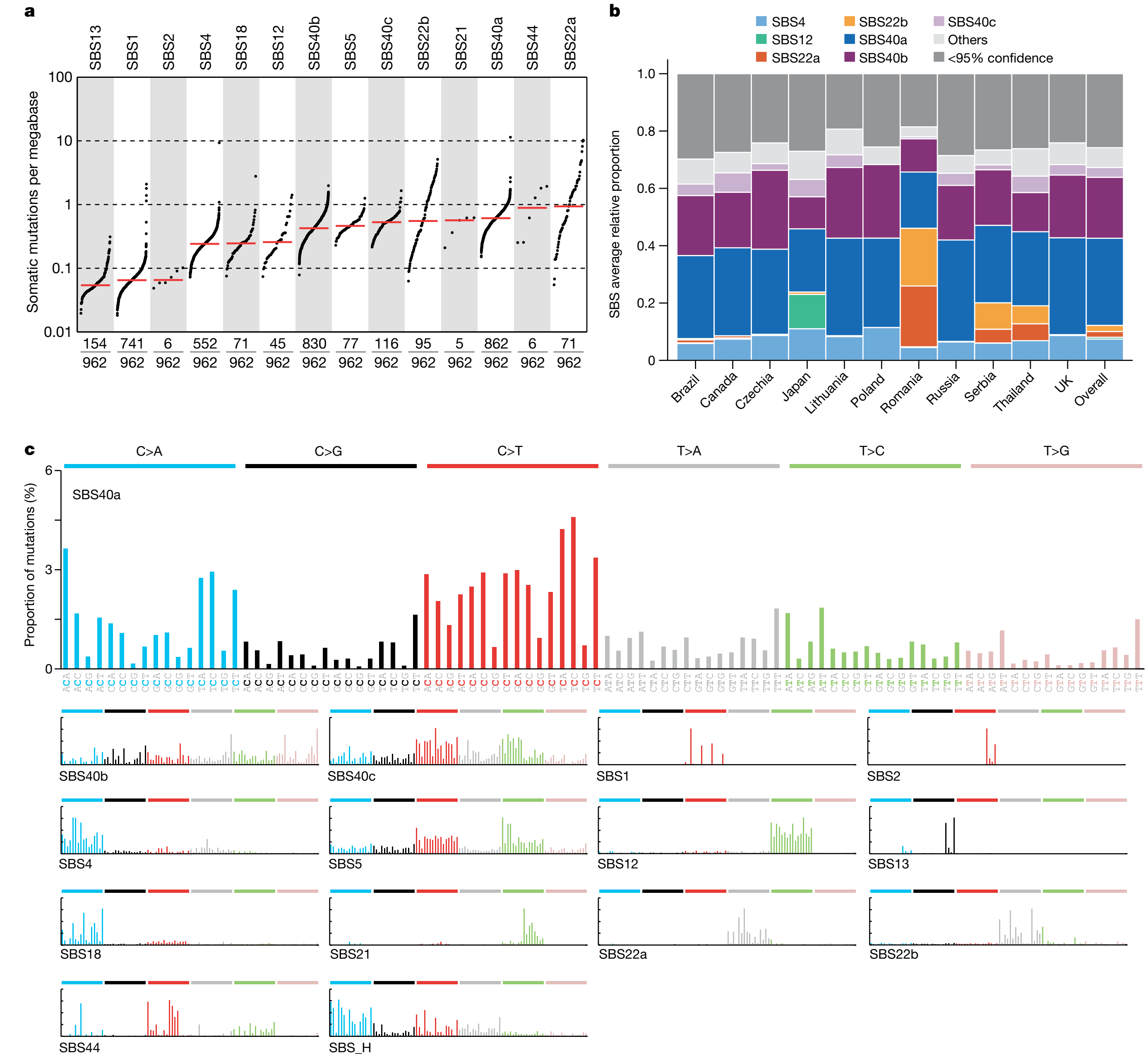


Fig. 2 | SBS signature operative in ccRCCs. a, TMB plot showing the frequency and number of mutations per megabase for each of the decomposed SBS signatures. Data include only samples with more than zero mutations. **b**, Average relative attribution for SBS signatures across countries. Signatures that contribute less than 5% on average are grouped in the 'others' category, except for SBS12 and the aristolochic acid-related signatures SBS22a and SBS22b. The '<95% confidence' category accounts for the proportion of mutation burden that could not be assigned to any signature with confidence level of at least 95%. **c**, Decomposed signatures, including reference COSMIC signatures as well as de novo signatures that are not decomposed into COSMIC reference signatures.

Figs. 4–6 and Extended Data Table 1), which are therefore also probably due to aristolochic acid exposure. SBS_I, similar to SBS22, is composed predominantly of T>A mutations. The signature identified previously as SBS22 was therefore renamed SBS22a, and the three newly identified signatures were named SBS22b, DBS20 and ID23, respectively. The mutation burden of SBS22a and SBS22b differed between Serbia and Romania, with higher levels being detected in Romania and away from recognized BEN zones[20] (Fig. 3 and Extended Data Fig. 7c,d). The two signatures may be due to different subsets of aristolochic acids and/or to different metabolites, which induce slightly different mutational patterns. Only five ccRCC cases were known to reside within recognized BEN zones, suggesting no clear link between the two diseases. Although the source of this exposure is uncertain, these results indicate that a substantial proportion of the population over a wide geographical area of eastern Europe, possibly tens of millions of people, has been exposed to aristolochic acid-containing compounds, the public health consequences of which are uncertain.

SBS12 was present in 72% of Japanese and 2% of non-Japanese ccRCC (P value = 4.7×10^{-78}) (Extended Data Fig. 7h). Compared with the mutation burdens imposed by aristolochic acid in ccRCC, SBS12 contributed modest mutation loads. SBS12 is composed predominantly of T>C substitutions and exhibits strong transcriptional strand bias with more T>C mutations on the transcribed strand than on the untranscribed strand of protein-coding genes. Transcriptional strand bias is typically

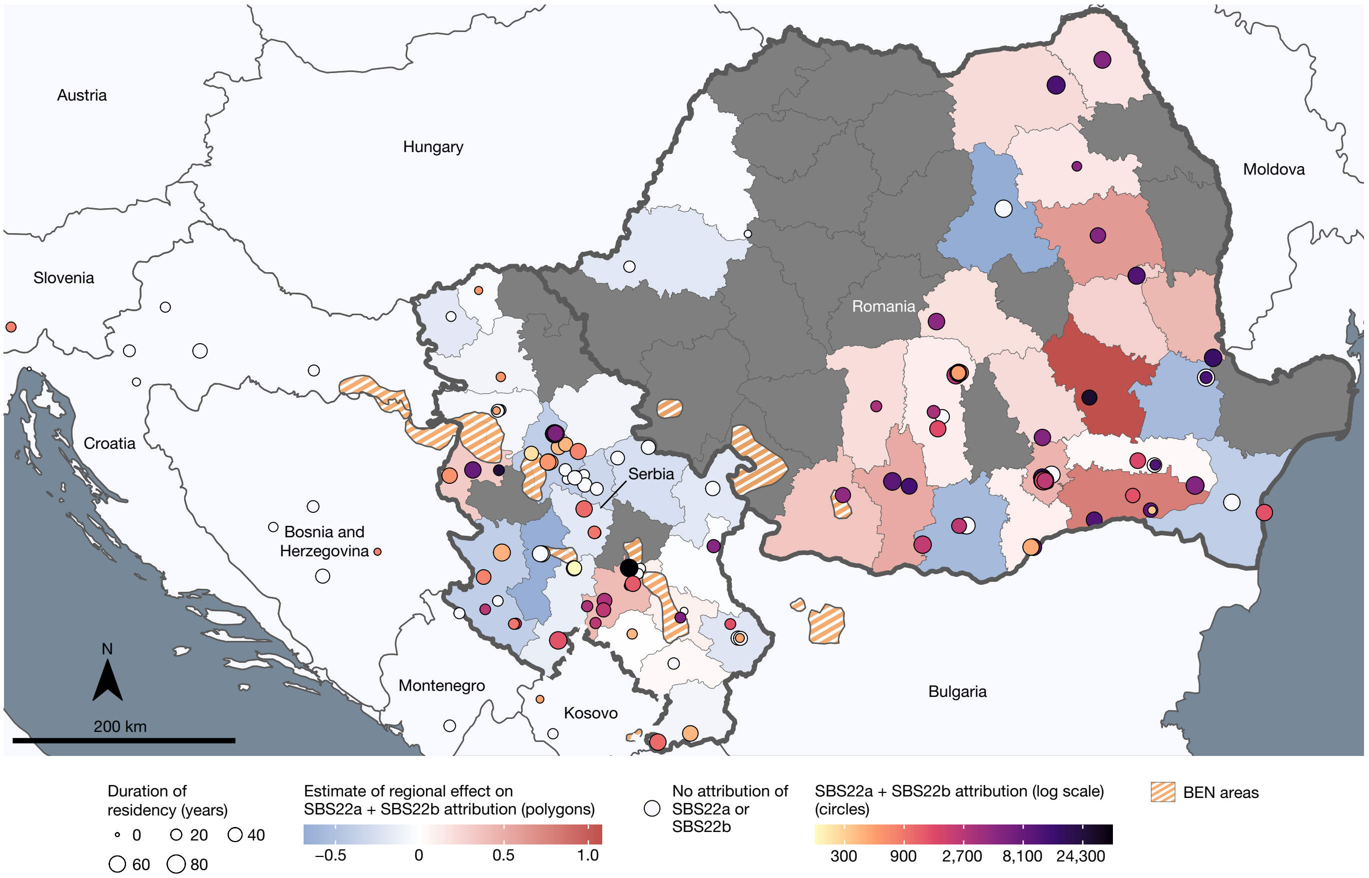

Austria
Hungary
Moldova
Slovenia
Romania
Croatia
Serbia
Bosnia and Herzegovina
Montenegro
Kosovo
Bulgaria
N
200 km
Duration of residency (years)
0
20
40
60
80
Estimate of regional effect on SBS22a + SBS22b attribution (polygons)
−0.5
0
0.5
1.0
No attribution of SBS22a or SBS22b
SBS22a + SBS22b attribution (log scale) (circles)
300
900
2,700
8,100
24,300
BEN areas

Fig. 3 | Geospatial analysis of aristolochic acid-related SBS signatures. Distribution of cases of ccRCC from Romania and Serbia with known residential history, along with the summed levels of SBS22a and SBS22b attributions (per case and regional estimate), with respect to BEN areas. White circles represented cases with no detected activity of SBS22a and SBS22b.

caused by activity of transcription-coupled nucleotide excision repair acting on bulky DNA adducts owing to exogenous mutagenic exposures such as tobacco smoke chemicals[13], ultraviolet light[13], aristolochic acids[15] and aflatoxins[21]. Assuming that transcription-coupled repair of DNA adducts is responsible for the SBS12 strand bias, the adducts are probably on adenine. Alternatively, transcriptional strand bias can also be caused by transcription-coupled damage, which results in an increase of mutations in genic regions compared with intergenic regions, however, prior topography analysis of mutational signatures has shown that SBS12 does not exhibit such an enrichment[22–24]. The presence of SBS12 was replicated in two further series of whole-genome sequenced ccRCCs from Japan, including 14 cases from an independent study group who undertook a broad genomic analysis of ccRCC but without detailed mutational signature analysis[25] and a more recent unpublished series of 61 cases from an additional cohort of ccRCC sequenced by the same centre as the initial cohort (Supplementary Note). SBS12 was present in 12 out of 14 (85%) and 46 out of 61 (75%) cases, respectively. SBS12 was previously reported in hepatocellular carcinomas[4,13] and additional analysis of existing datasets revealed strong SBS12 enrichment in hepatocellular carcinomas from Japan compared with other countries (P value = 3.8×10^{-15}; Supplementary Note). These results, therefore, indicate that exposure to an agent that contributes SBS12 mutations to kidney and liver cancer is common in Japan and rare in the other ten countries included in this study. The agent responsible for SBS12 is unknown, although the precedents provided by other mutational signatures with strong transcriptional strand bias suggest that it is probably of exogenous origin[22,24]. A polymorphism in aldehyde dehydrogenase 2 that is known to impair metabolism of alcohol to aldehydes and is common in Japan did not associate with levels of SBS12, and neither did any other common germline variants (Supplementary Note).

SBS40a, SBS40b and SBS40c were present in ccRCCs from all 11 countries. The country-specific average mutation burdens of SBS40a and SBS40b were positively associated with country-specific ASRs of kidney cancer incidence (P value = 0.0022 and P value = 5.1×10^{-18}, respectively; Fig. 4a, Extended Data Fig. 8a and Supplementary Table 8), with the highest mutation loads in the Czech Republic and Lithuania. Kidney cancer incidence rates also vary between the regions of the Czech Republic and SBS40b mutation burdens differed significantly between these (P value = 0.011; Fig 4b,c and Supplementary Table 9), with the highest attribution in the highest-risk region. SBS40b exhibits modest transcriptional strand bias and—assuming that transcription-coupled repair of DNA adducts is responsible—the adducts underlying SBS40b are probably on pyrimidines. Indel signatures ID5 and ID8, which together contributed around 60% of the indel mutation burden on average, were also strongly associated with country-specific kidney cancer ASR (P value = 1.3×10^{-10} and P value = 7.1×10^{-5}, respectively; Extended Data Fig. 8b,c). Signatures ID5 and ID8 correlated with each other (Spearman's r = 0.78), as well as with SBS40b (r = 0.79 and r = 0.74, respectively) indicating that they probably all constitute products of the same underlying mutational process. Thus, the burdens of the full complement of mutation types generated by this mutational process correlate with age-adjusted kidney cancer incidence rates. The overall mutational burden did not, however, associate significantly with kidney cancer incidence rates (Extended Data Fig. 9).

To investigate potential mutagenic agents underlying these geographically variable signatures, we conducted an untargeted metabolomics screen of plasma on 901 individuals in the study, from all

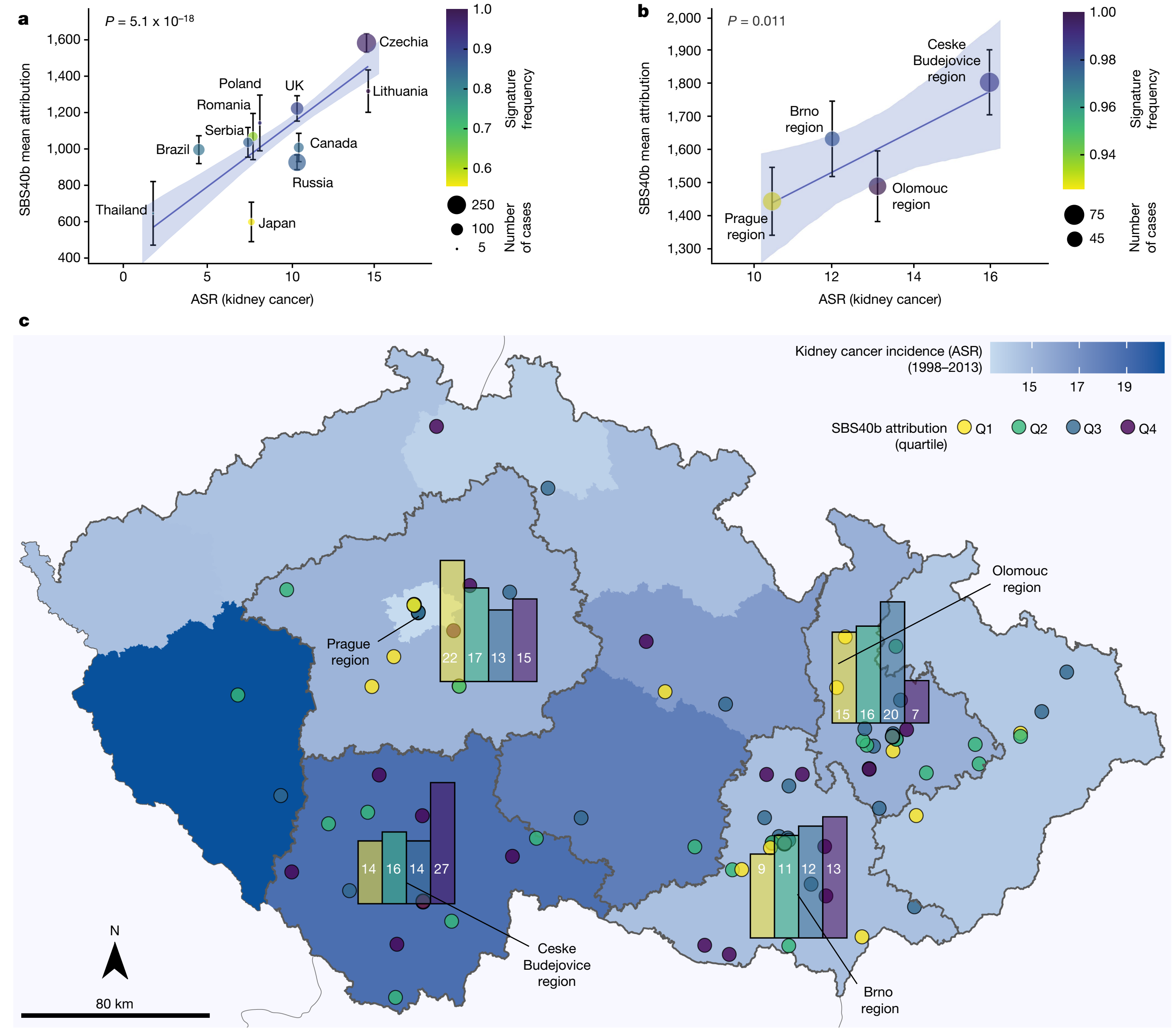


Fig. 4 | Association of SBS40b signature attribution with incidence of kidney cancer. a, Number of mutations attributed to signature SBS40b against ASR of kidney cancer in each of the 11 countries represented in the cohort. Data are mean ± s.e.m. ($n = 961$ biologically independent samples examined over 1 independent experiment). **b**, Number of mutations attributed to signature SBS40b in four regions of Czech Republic against ASR of kidney cancer in each region. Data are mean ± s.e.m. ($n = 961$ biologically independent samples examined over 1 independent experiment). **a**,**b**, *P* values are shown for the ASR variable in linear regressions across all cases, adjusted for sex and age of diagnosis. **c**, Attribution of SBS40b signature within the Czech Republic, with bar plots showing the number of cases for each quartile of SBS40b attribution across Prague, Olomouc, Ceske Budejovice and Brno regions.

countries except Japan (Methods). A total of 2,392 metabolite features was obtained, including 944 independent peaks ($r < 0.85$). Three features were associated with SBS4 (Supplementary Table 10), with two identified as hydroxycotinine (*P* value = 2.9 × 10^{-9}) and cotinine (*P* value = 1.9 × 10^{-5}), two major metabolites of nicotine[26]. Eight features were associated with SBS40b (Supplementary Table 10). One feature was identified as *N*,*N*,*N*-trimethyl-L-alanyl-L-proline betaine (TMAP) (*P* value = 1.2 × 10^{-5}; Supplementary Table 11), increased levels of which correlate strongly with reduced kidney function[27]. Other established measures of kidney function, including cystatin C and creatinine, were correlated with TMAP (*P* value = 2.5 × 10^{-30} and 1.7 × 10^{-69}, respectively) and also showed evidence of positive association with SBS40b (*P* value = 0.023 and 0.058, respectively). Thus, exposure to the mutagenic agent responsible for SBS40b is associated with reduced kidney function. No recognized metabolome features were significantly associated with any other signatures.

A total of 1,962 'driver' mutations were found in 136 genes including *VHL*, *PBRM1*, *SETD2* and *BAP1*, the known frequently mutated cancer genes in ccRCC[9,25] (Fig. 5a, Supplementary Table 12 and Methods). The frequencies of mutations in these genes were consistent across countries (Fig. 5b). The spectrum of all driver mutations in ccRCC with aristolochic acid exposure (Methods) was enriched in T>A mutations compared with non-exposed cases (25% versus 13%, *P* value = 0.0062; Fig. 5c,d) with similar enrichment specifically in *VHL* mutations (30% versus 16%; Fig. 5e,f), and in the whole exomes (27% in exposed compared to 12% in unexposed cases). Thus genome-wide aristolochic acid mutagenesis has contributed in a proportionate fashion to generation of driver mutations in aristolochic acid-exposed ccRCC. The driver

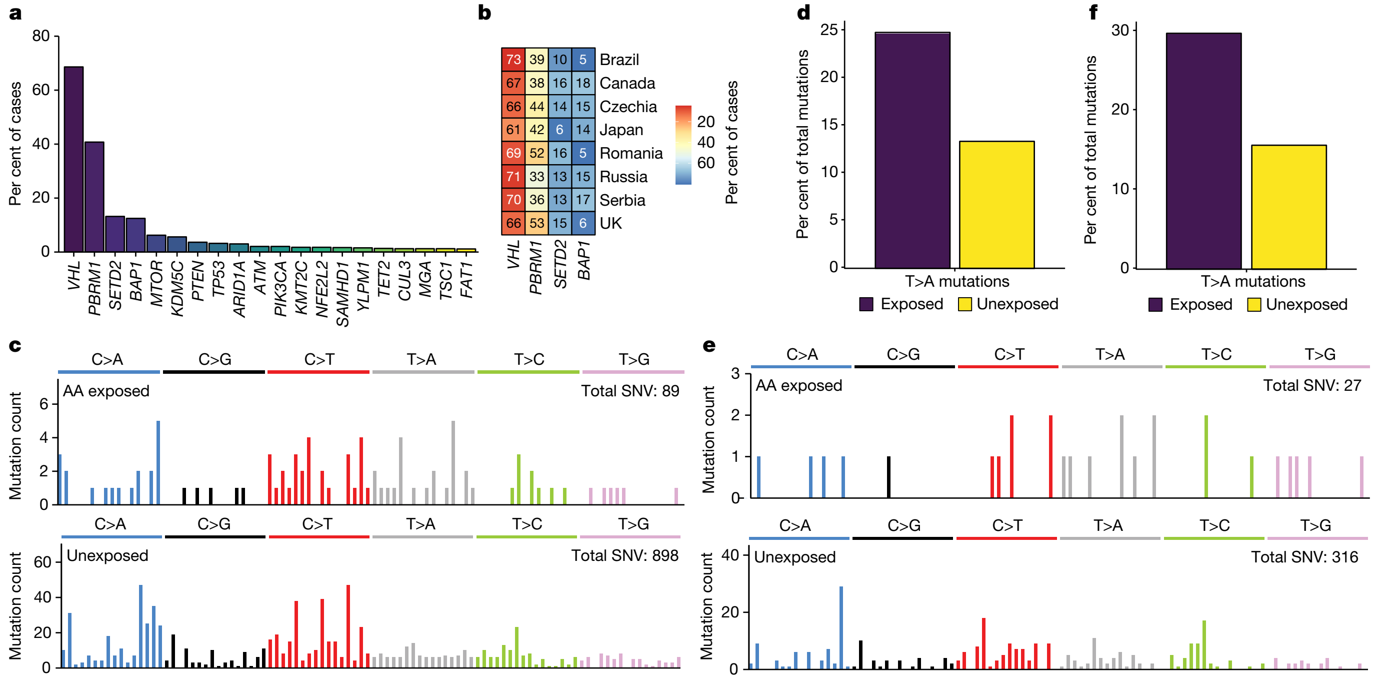


Fig. 5 | Driver mutation analysis in ccRCCs. a, Frequency of driver genes in the cohort. Only genes mutated in at least ten cases are shown. **b**, Frequency of driver genes across countries. Thailand, Poland and Lithuania are not shown owing to low numbers of samples. **c**, SBS-96 mutational spectra of all driver mutations in ccRCC for aristolochic acid-exposed and unexposed cases. **d**, Percentage of T>A driver mutations in aristolochic acid-exposed and unexposed cases. **e**, SBS-96 mutational spectra of *VHL* mutations in ccRCC for aristolochic acid-exposed and unexposed cases. **f**, Percentage of T>A *VHL* mutations in AA-exposed and unexposed cases.

mutation spectrum did not show statistically significant enrichment of T>C mutations in SBS12 exposed cases (20% versus 12%, $P = 0.069$), but was consistent with the level of enrichment in the exome (21% in exposed cases compared with 15% in unexposed cases). SBS40b also did not show statistically significant enrichment, possibly owing to the ubiquitous exposure and its relatively flat and featureless mutation profile.

Exogenous mutagenic exposures that ultimately cause cancer may be present during the early stages of evolution of cancer clones. To time mutagenic exposures, we estimated the contribution of each mutational signature to mutations in the primary clone (relatively early) and to mutations in subclones[28,29] (relatively late) (Methods). All signatures of the putative exogenous mutagenic exposures observed in ccRCC were present at relatively early stages of cancer development, consistent with exposures to normal cells. SBS12, SBS22b and SBS40b showed higher activities in main clones compared with subclones (q value = 0.04, q value = 0.02 and q value = 2.3×10^{-5}, respectively) (Extended Data Fig. 10) and SBS22a showed no significant difference[15,16]. By contrast, signatures due to endogenous mutational processes including APOBEC DNA editing (SBS13) and oxidative damage (SBS18), were enriched in subclones (q value = 1.6×10^{-4} and q value = 3.2×10^{-7}, respectively).

Established or suspected risk factors for ccRCC include age, tobacco smoking, obesity, hypertension, diabetes and environmental exposure to PFAS compounds[30]. Total SBS, DBS, and indel mutation burdens associated with age, as did SBS1, SBS4, SBS5, SBS40a, SBS40b, SBS22a, SBS22b, DBS2, ID1, ID5 and ID8. Total SBS (P value = 3.1×10^{-5}), DBS (P value = 3.7×10^{-3}) and indel ($P = 1.3 \times 10^{-4}$) mutation burdens also associated with sex, with males having higher mutation burdens than females, and with SBS40b showing a similar association ($P = 9.3 \times 10^{-5}$). Associations with tobacco smoking were observed for SBS4 ($P = 5.3 \times 10^{-6}$) and DBS2 ($P = 2.4 \times 10^{-7}$), both of which are known to be caused by tobacco carcinogens[24,31]. These results suggest that the known increased risk of ccRCC with tobacco smoking is owing to direct exposure of the kidney to tobacco-related mutagens (Supplementary Note). Associations of particular mutational signatures with other ccRCC risk factors were not observed (Supplementary Tables 13 and 14). To complement this analysis of observational data, associations between polygenic risk scores for known ccRCC risk factors and mutational signatures[32,33] were examined (Methods). Consistent with the observational data, no associations were found between genetically inferred risk factors and mutational signatures except for tobacco smoking and DBS2 (P value = 0.01; Supplementary Table 15).

Discussion

Somatic mutations in the genomes of 962 individuals with ccRCC from 11 countries indicate the existence of multiple, widespread mutational processes exhibiting substantial geographical variation in their contributions to ccRCC mutation loads. The results contrast with those from 552 oesophageal squamous carcinomas from 8 countries with widely different oesophageal carcinoma incidence rates in which geographical differences in mutation burdens or signatures were not observed[34]. Together the studies implicate both geographically variable mutagenic and non-mutagenic carcinogenic exposures contributing to global cancer incidence. Indeed, the presence of mutational signatures associated with tobacco smoking but absence of signatures associated with other known ccRCC risk factors, such as obesity and hypertension, suggests that the latter may be mediated by non-mutagenic processes and, therefore, that both classes of carcinogen contribute to the development of ccRCC.

The existence, identity and carcinogenic effect of some of the agents underlying these mutational processes are known. Aristolochic acids are believed to cause SBS22a and SBS22b and its associated signatures, but this study suggests that the geographical extent and proportion of the population acquiring mutations in southeastern Europe is far greater than previously anticipated, possibly affecting tens of millions of individuals. The sources of the aristolochic acid exposure, the manner by which it is ingested and whether the exposure continues today are uncertain, and further definition of the source and extent of this exposure is required in order to provide a foundation for public health action.

The existence of the mutagenic exposures underlying SBS12 and SBS40b were not previously suspected, and their causative agents are unknown. Based on current information, the exposure causing SBS12 is restricted to Japan. However, larger studies are now indicated to explore the geographical extent of exposure in Japan and neighbouring countries, and the proportions of their populations that have been exposed. Studies of Japanese migrants to other countries are also likely to be informative regarding the potential source of exposure. In the first instance, this will be achievable by further sequencing of kidney and hepatocellular cancer genomes. However, studies of normal tissues using recently reported sequencing methods that enable detection of somatic mutations in normal cells[35], and particularly relatively accessible ones such as cells in urine that can be prospectively collected, may enable large population-based studies providing better characterization of the exposure and its consequences. As with exposure to aristolochic acid in southeastern Europe, it is possible that tens of millions of individuals in East Asia are exposed to a potent mutagen, and identification of the source and extent of exposure must be a public health priority.

In contrast to aristolochic acid and the agent that causes SBS12, the exposure underlying SBS40b appears to be globally ubiquitous. It causes mutations predominantly in ccRCC, with much lower burdens in other cancer types, and generates mutation loads that correlate strongly with age and sex. There are few clues as to its origin or nature.

The incidence rates of ccRCC vary approximately eightfold across the 11 countries from which ccRCCs were sequenced. A strong positive correlation (P value = 5.5×10^{-18}) was found between the average mutation loads attributable to SBS40b in each country (and also those of ID5 and ID8, which are correlated with SBS40b) and incidence of kidney cancer within each country. This correlation reflects approximately a tripling of average country-specific SBS40b mutation loads (a difference of around 1,000 mutations) in parallel with the eightfold increase of country-specific ASR.

SBS40b mutation burdens also positively correlated with biomarkers of impaired kidney function, reminiscent of the nephrotoxic effects of aristolochic acids in BEN. It is possible that the increased SBS40b somatic mutation load itself engenders this reduction in renal function. However, studies of other normal tissues suggest that they are generally tolerant of elevated mutation burdens, except for manifesting a higher incidence of neoplasia[36,37]. It is also possible that the agent underlying SBS40b is directly nephrotoxic—for example, by engendering DNA damage and a response to it—and that the mutations it generates are immaterial to kidney function. It is also conceivable, however, that impaired renal function, potentially owing to many different causes, results in a metabolic state which itself causes the elevated SBS40b mutation load. Whatever the mutational process underlying SBS40b, it is plausible that it contributes to the geographical variation in the ASRs for kidney cancer. It is of public health interest to determine the cause of SBS40b and thus to consider whether the exposure can be mitigated, potentially with concomitant reduction in global ccRCC incidence rates.

The absence of any association between several known risk factors for ccRCC and mutation burden—in particular for obesity and hypertension—supports a model of cancer development in which mutations are essential but additional factors affect the expansion of a mutated clone and thus the chance of it progressing into cancer[38]. Further efforts at defining how lifestyle and environmental exposures contribute to cancer development will therefore require a greater understanding of both the causes of the mutations in cell clones in normal tissue and the further promotion of such mutant clones by non-mutagenic processes.

Finally, the substantial geographical variability of SBS12, SBS22a and SBS22b, with most countries not showing evidence of exposure, raises the possibility that additional mutational signature studies of ccRCC involving more countries may reveal further mutagenic exposures. Furthermore, the results relating to SBS40b highlight the prospect that a significant proportion of global cancer burden may be caused by relatively ubiquitous exposures that are not readily detectable by classical cancer epidemiology studies. The conduct of large-scale whole-genome sequencing for other cancer sites across high- and low-risk populations around the world would seem to be an appropriate strategy for detecting such novel cancer-causing agents.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07368-2.

1. Brennan, P. & Davey-Smith, G. Identifying novel causes of cancers to enhance cancer prevention: new strategies are needed. *J. Natl Cancer Inst.* **114**, 353–360 (2022).
2. Hsieh, J. J. et al. Renal cell carcinoma. *Nat. Rev. Dis. Primers* **3**, 17009 (2017).
3. Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat. Rev. Cancer* **21**, 619–637 (2021).
4. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
5. Scelo, G. et al. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat. Commun.* **5**, 5135 (2014).
6. Mitchell, T. J. et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx renal. *Cell* **173**, 611–623.e17 (2018).
7. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
8. Degasperi, A. et al. A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nat. Cancer* **1**, 249–263 (2020).
9. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).
10. Mutographs. *Cancer Grand Challenges* https://cancergrandchallenges.org/teams (2023).
11. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA* **71**, 209–249 (2021).
12. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
13. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
14. Drost, J. et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234–238 (2017).
15. Hoang, M. L. et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci. Transl. Med.* **5**, 197ra102 (2013).
16. Poon, S. L. et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
17. Grollman, A. P. Aristolochic acid nephropathy: harbinger of a global iatrogenic disease. *Environ. Mol. Mutagen.* **54**, 1–7 (2013).
18. Turesky, R. J. et al. Aristolochic acid exposure in Romania and implications for renal cell carcinoma. *Br. J. Cancer* **114**, 76–80 (2016).
19. Wang, X.-M. et al. Integrative genomic study of Chinese clear cell renal cell carcinoma reveals features associated with thrombus. *Nat. Commun.* **11**, 739 (2020).
20. Stefanovic, V. & Radovanovic, Z. Balkan endemic nephropathy and associated urothelial cancer. *Nat. Clin. Pract. Urol.* **5**, 105–112 (2008).
21. Huang, M. N. et al. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res.* **27**, 1475–1486 (2017).
22. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
23. Otlu, B. et al. Topography of mutational signatures in human cancer. *Cell Rep.* **42**, 112930 (2023).
24. Nik-Zainal, S. The genome as a record of environmental exposure. *Mutagenesis* **30**, 763–770 (2015).
25. Sato, Y. et al. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* **45**, 860–867 (2013).
26. Dempsey, D. et al. Nicotine metabolite ratio as an index of cytochrome P450 2A6 metabolic activity. *Clin. Pharmacol. Ther.* **76**, 64–72 (2004).
27. Velenosi, T. J. et al. Untargeted metabolomics reveals *N*,*N*,*N*-trimethyl-L-alanyl-L-proline betaine (TMAP) as a novel biomarker of kidney function. *Sci. Rep.* **9**, 6831 (2019).
28. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
29. Dentro, S. C., Wedge, D. C. & van Loo, P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harb. Perspect. Med.* **7**, a026625 (2017).
30. Shearer, J. J. et al. Serum concentrations of per- and polyfluoroalkyl substances and risk of renal cell carcinoma. *J. Natl Cancer Inst.* **113**, 580–587 (2021).

31. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836.e16 (2019).
32. Gabriel, A. A. G. et al. Genetic analysis of lung cancer and the germline impact on somatic mutation burden. *J. Natl Cancer Inst.* **114**, 1159–1166 (2022).
33. Liu, Y., Gusev, A., Heng, Y. J., Alexandrov, L. B. & Kraft, P. Somatic mutational profiles and germline polygenic risk scores in human cancer. *Genome Med.* **14**, 14 (2022).
34. Moody, S. et al. Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. *Nat. Genet.* **53**, 1553–1563 (2021).
35. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
36. Robinson, P. S. et al. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat. Genet.* **53**, 1434–1442 (2021).
37. Robinson, P. S. et al. Inherited *MUTYH* mutations cause elevated somatic mutation rates and distinctive mutational signatures in normal human cells. *Nat. Commun.* **13**, 3949 (2022).
38. Fowler, J. C. & Jones, P. H. Somatic mutation: what shapes the mutational landscape of normal epithelia? *Cancer Discov.* **12**, 1642–1655 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[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, CA, USA. [4]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA. [5]Moores Cancer Center, University of California San Diego, La Jolla, CA, USA. [6]Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA, USA. [7]Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey. [8]Cancer Epidemiology Unit, The Nuffield Department of Population Health, University of Oxford, Oxford, UK. [9]Evidence Synthesis and Classification Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France. [10]Nutrition and Metabolism Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France. [11]Service d'Anatomie Pathologique, Assistance Publique-Hôpitaux de Paris, Univeristé Paris-Saclay, Le Kremlin-Bicêtre, France. [12]Oncopathology Research Center, Iran University of Medical Sciences, Tehran, Iran. [13]Hasheminejad Kidney Center, Iran University of Medical Sciences, Tehran, Iran. [14]Translational Medicine Research Center, Faculty of Medicine, Prince of Songkla University, Hat Yai, Thailand. [15]Division of Urology, Department of Surgery, Faculty of Medicine, Prince of Songkla University, Hat Yai, Thailand. [16]Department of Environmental Epidemiology, Nofer Institute of Occupational Medicine, Łódź, Poland. [17]Laboratory of Genetic Diagnostic, National Cancer Institute, Vilnius, Lithuania. [18]Department of Botany and Genetics, Institute of Biosciences, Vilnius University, Vilnius, Lithuania. [19]Laboratory of Molecular Medicine, The Institute of Medical Science, The University of Tokyo, Minato-ku, Japan. [20]Division of Cancer Genomics, National Cancer Center Research Institute, Chuo-ku, Japan. [21]Department of Pathology, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Japan. [22]Occupational Health and Toxicology Department, National Center for Environmental Risk Monitoring, National Institute of Public Health, Bucharest, Romania. [23]Urology Department, Carol Davila University of Medicine and Pharmacy, Prof. Dr. Th. Burghele Clinical Hospital, Bucharest, Romania. [24]Clinic of Nephrology, Faculty of Medicine, Military Medical Academy, Belgrade, Serbia. [25]Department of Urology, University Hospital Dr D. Misovic Clinical Center, Belgrade, Serbia. [26]International Organization for Cancer Prevention and Research, Belgrade, Serbia. [27]Cancer Research UK Edinburgh Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. [28]Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada. [29]Ontario Tumour Bank, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [30]Experimental Research Center, Genomic Medicine Laboratory, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil. [31]Post-Graduate Program in Genetics and Molecular Biology, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. [32]Transplant Immunology and Personalized Medicine Unit, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil. [33]Service of Urology, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil. [34]Post-Graduate Program in Medicine: Surgical Sciences, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. [35]Department of Anatomic Pathology, A. C. Camargo Cancer Center, São Paulo, Brazil. [36]Department of Epidemiology, A. C. Camargo Cancer Center, São Paulo, Brazil. [37]Department of Urology, A. C. Camargo Cancer Center, São Paulo, Brazil. [38]National Institute for Science and Technology in Oncogenomics and Therapeutic Innovation, A.C. Camargo Cancer Center, São Paulo, Brazil. [39]Latin American Renal Cancer Group (LARCG), São Paulo, Brazil. [40]Department of Surgery, Division of Urology, Sao Paulo Federal University (UNIFESP), São Paulo, Brazil. [41]Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil. [42]Life and Health Sciences Research Institute (ICVS), School of Medicine, Minho University, Braga, Portugal. [43]Faculdade Ciências Médicas de Minas Gerais, Belo Horizonte, Brazil. [44]Department of Urology, Barretos Cancer Hospital, Barretos, Brazil. [45]Department of Pathology, Barretos Cancer Hospital, Barretos, Brazil. [46]Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK. [47]Department of Clinical Epidemiology, N. N. Blokhin National Medical Research Centre of Oncology, Moscow, Russia. [48]Department of Urology, N. N. Blokhin National Medical Research Centre of Oncology, Moscow, Russia. [49]Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic. [50]Institute of Public Health and Preventive Medicine, 2nd Faculty of Medicine, Charles University, Prague, Czech Republic. [51]Department of Oncology, 2nd Faculty of Medicine, Charles University and Motol University Hospital, Prague, Czech Republic. [52]Institute of Hygiene and Epidemiology, 1st Faculty of Medicine, Charles University, Prague, Czech Republic. [53]Faculty of Health Sciences, Palacky University, Olomouc, Czech Republic. [54]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA. [55]Bevital AS, Bergen, Norway. [56]Observational and Pragmatic Research Institute Pte Ltd, Singapore, Singapore. [57]Department of Pathology, Radboud University Medical Centre, Nijmegen, Netherlands. [58]These authors contributed equally: Sergey Senkin, Sarah Moody. ✉e-mail: brennanp@iarc.who.int

Methods

Recruitment of cases and informed consent

The International Agency for Research on Cancer (IARC/WHO) coordinated case recruitment through an international network of more than 40 collaborators from the 11 participating countries (Table 1 and Supplementary Table 16). The inclusion criteria for patients were ≥18 years of age (range from 23 to 87, with a mean of 60 and a s.d. of 12), confirmed diagnosis of primary ccRCC and no prior cancer treatment. Informed consent was obtained for all participants. Patients were excluded if they had any condition that could interfere with their ability to provide informed consent or if there were no means of obtaining adequate tissues or associated data as per the protocol requirements. Ethical approvals were first obtained from each Local Research Ethics Committee and Federal Ethics Committee when applicable, as well as from the IARC Ethics Committee.

Bio-samples, data collection and expert pathology review

Dedicated standard operating procedures, following guidelines from the International Cancer Genome Consortium (ICGC), were designed by IARC/WHO to select appropriate case series with complete biological samples and exposure information as described previously[34] (Supplementary Table 16). In brief, for all case series included, anthropometric measures were taken, together with relevant information regarding medical and familial history. Comparable smoking and alcohol history was available from all centres. Detailed epidemiological information on residential history was collected in Czech Republic, Romania, and Serbia. Potential limitations of using retrospective clinical data collected using different protocols from different populations were addressed by a central data harmonization to ensure a comparable group of exposure variables (Supplementary Table 16). All patient-related data as well as clinical, demographical, lifestyle, pathological and outcome data were pseudonymized locally through the use a dedicated alpha-numerical identifier system before being transferred to IARC/WHO central database.

Original diagnostic pathology departments provided diagnostic histological details of contributing cases through standard abstract forms. IARC/WHO centralized the entire pathology workflow and coordinated a centralized digital pathology examination of the frozen tumour tissues collected for the study as well as formalin-fixed, paraffin-embedded sections when available, via a web-based report approach and dedicated expert panel following standardized procedures as described previously[34]. A minimum of 50% viable tumour cells was required for eligibility to whole-genome sequencing.

In summary, frozen tumour tissues were first examined to confirm the morphological type and the percentage of viable tumour cells. A random selection of tumour tissues was independently evaluated by a second pathologist. Enrichment of tumour component was performed by dissection of non-tumoral part, if necessary. 90 cases overlapped with a previously published cohort recruited under the Cancer Genomics of the Kidney (CAGEKID) project[5], which were also part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project[7].

DNA extraction

Extraction of DNA from fresh frozen tumour and matched blood samples was centrally conducted at IARC/WHO except for Japan, which performed DNA extractions at the local centre following a similarly standardized DNA extraction procedure. Of the cases which proceeded to the final analysis ($n = 962$), germline DNA was extracted from either buffy-coat, whole blood, or from adjacent normal tissue (samples from Japan) using previously described protocols and methods[34].

Whole-genome sequencing

In total, 1,583 renal cell carcinoma cases were evaluated, with 1,267 confirmed as ccRCC cases. One hundred and sixteen cases (9%) were excluded due to insufficient viable tumour cells (pathology level), or inadequate DNA (tumour or germline). DNA from 1,151 cases was received at the Wellcome Sanger Institute for whole-genome sequencing. Fluidigm single nucleotide polymorphism (SNP) genotyping with a custom panel was performed to ensure that each pair of tumour and matched normal samples originated from the same individual. Whole-genome sequencing (150 bp paired end) was performed on the Illumina NovaSeq 6000 platform with target coverage of 40X for tumours and 20X for matched normal tissues. All sequencing reads were aligned to the GRCh38 human reference genome using Burrows-Wheeler-MEM (v0.7.16a and v0.7.17). Post-sequencing quality control metrics were applied for total coverage, evenness of coverage and contamination. Cases were excluded if coverage was below 30X for tumour or 15X for normal tissue. For evenness of coverage, the median over mean coverage score was calculated. Tumours with median over mean coverage scores outside the range of values determined by previous studies to be appropriate for whole-genome sequencing (0.92–1.09) were excluded. Conpair[39] (https://github.com/nygenome/Conpair) was used to detect contamination, cases were excluded if the result was greater than 3%[40]. A total of 962 cases passed all criteria and were included in subsequent analysis.

Somatic variant calling

Variant calling was performed using the standard Sanger bioinformatics analysis pipeline (https://github.com/cancerit). Copy number profiles were determined first using the algorithms ASCAT[41] and BATTENBERG[28], where tumour purity allowed. Single nucleotide variants (SNVs) were called with cgpCaVEMan[42], indels were called with cgpPINDEL[43], and structural rearrangements were called using BRASS. CaVEMan and BRASS were run using the copy number profile and purity values determined from ASCAT where possible (complete pipeline, $n = 857$). Where tumour purity was insufficient to determine an accurate copy number profile (partial pipeline, $n = 105$), CaVEMan and BRASS were run using copy number defaults and an estimate of purity obtained from ASCAT/BATTENBERG. For SNV additional filters on ASRD (read length-adjusted alignment score of reads showing the variant allele) and CLPM (median number of soft-clipped bases in variant supporting reads) (ASRD ≥ 140 and CLPM = 0) were applied to remove potential false positive calls. A second variant caller, Strelka2, was run for SNVs and indels as consensus variant calling was previously shown to eliminate algorithm specific artefacts and to generate highly reproducible mutational spectra compared to using a single variant calling algorithm[34,44]. Only variants called by both the Sanger variant calling pipeline and Strelka2 were included in subsequent analysis.

Validation of sequencing for Japanese cases

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

Generation of mutational matrices

Mutational matrices for SBS, DBS and indels were generated using SigProfilerMatrixGenerator (https://github.com/AlexandrovLab/SigProfilerMatrixGenerator) with default options (v1.2.12)[45].

Mutational signature analysis

Mutational signatures were extracted using two algorithms, SigProfilerExtractor (https://github.com/AlexandrovLab/SigProfilerExtractor), based on non-negative matrix factorization, and mSigHdp[46] (https://github.com/steverozen/mSigHdp), based on the Bayesian hierarchical Dirichlet process. For SigProfilerExtractor, de novo mutational signatures were extracted from each mutational matrix using SigProfilerExtractor with nndsvd_min initialization (NMF_init = “nndsvd_min”) and default parameters (v1.1.9)[47]. Briefly, SigProfilerExtractor deciphers mutational signatures by first performing Poisson resampling of the original matrix with additional renormalization (based on a generalized mixture model approach) of hypermutators to reduce their effect on the overall factorization[47]. Non-negative matrix factorization (NMF) was performed using initialization with non-negative singular value decomposition and by applying the multiplicative update algorithm using the Kullback–Leibler divergence as an objective function[47]. NMF was applied with factorizations between $k = 1$ and $k = 20$ signatures; each factorization was repeated 500 times[47]. De novo SBS mutational signatures were extracted with SigProfilerExtractor for both SBS-288 and SBS-1536 contexts[45]. The results were largely concordant with the SBS-1536 de novo signatures allowing additional separation of mutational processes, therefore the SBS-1536 de novo signatures were taken forward for further analysis (Supplementary Table 2). Mutational signatures for DBS and indels were extracted in DBS-78 and ID-83 contexts respectively (Supplementary Tables 3 and 4). Where possible, SigProfilerExtractor matched each de novo extracted mutational signature to a set of previously identified COSMIC signatures[4], for SBS-1536 signatures this requires collapsing the 1536 classification into the standard 96 substitution type classification with six mutation classes having single 3′ and 5′ sequence contexts (Supplementary Table 7). This step makes it possible to distinguish between de novo signatures which can be explained by a combination of the known catalogue of mutational process (which have not been completely separated during the extraction), and those which have not been previously identified. mSigHdp extraction of SBS-96 and ID-83 signatures was performed using the suggested parameters and using the country of origin to construct the hierarchy. SigProfilerExtractor’s decomposition module was subsequently used to match mSigHdp de novo signatures to previously identified COSMIC signatures[4]. Further details on the comparison of results between SigProfilerExtractor and mSigHdp, decomposition of de novo signatures into COSMIC reference signatures and support for the splitting of SBS40 components can be found in the Supplementary Note.

Attribution of activities of mutational signatures

The de novo (SigProfiler) and COSMIC signature (SigProfiler and mSigHdp) activities were attributed for each sample using the MSA signature attribution tool (v2.0, https://gitlab.com/s.senkin/MSA)[48]. For COSMIC attributions, only COSMIC reference signatures, which were identified in the decomposition of de novo signatures, were included in the panel for attribution, in addition to de novo signatures which could not be decomposed into COSMIC reference. At its core, the tool utilizes the non-negative least squares (NNLS) approach minimizing the L2 distance between the input sample and the one reconstructed using available signatures. To limit false positive attributions, an automated optimization procedure was applied by repeated removal of all signatures that do not increase the L2 similarity of a sample by >0.008 for SBS, >0.014 for DBS, and >0.03 for ID mutation types, as suggested by simulations. These optimal penalties were derived using an optional parameter (params.no_CI_for_penalties = false) utilizing a conservative approach in calculation of penalties. Finally, a parametric bootstrap approach was applied to extract 95% confidence intervals for each attributed mutational signature activity.

Driver mutations

A dNdS approach was used to identify genes under positive selection in ccRCC[49]. The analysis was performed both for the whole genome (q value < 0.01), and with restricted hypothesis testing for a panel of 369 known cancer genes[49]. Variants in any gene identified as under positive selection in global dNdS or in the 369-cancer gene panel were assessed as potential drivers[49]. Candidate driver mutations were annotated with the mode of action using the Cancer Gene Census (https://cancer.sanger.ac.uk/census) and the Cancer Genome Interpreter tool[50] (https://www.cancergenomeinterpreter.org). Missense mutations were assessed using the MutationMapper tool[51] (http://www.cbioportal.org/mutation_mapper). Variants were considered likely drivers if they met any of the following criteria: (1) Truncating mutations in genes annotated as tumour suppressors; (2) mutations annotated as probably or known oncogenic in MutationMapper; (3) truncating variants in genes with selection (q value < 0.05) for truncating mutations assumed to be tumour suppressors and thus likely drivers; (4) missense variants in all genes under positive selection and with dN/dS ratios for missense mutations above 5 (assuming 4 of every 5 missense mutations are drivers) labelled as likely drivers; or (5) in-frame indels in genes under significant positive selection for in-frame indels.

Evolutionary analysis

Subclonal architecture reconstruction was performed using the DPClust R package v2.2.8 (refs. 28,29), after obtaining cancer cell fraction (CCF) estimates by dpclust3p v1.0.8 (https://github.com/Wedge-lab/dpclust3p) based on the variant allele frequency provided by the somatic variant callers and the copy number profiles determined by the BATTENBERG algorithm. Only tumours with at least 40% purity according to BATTENBERG were considered for further evolutionary analysis. For each tumour with at least one subclone, the respective somatic mutations were split into clonal and subclonal mutations using the most probable cluster assignment for each mutation as per the DPClust output. Mutations not assigned to a cluster by DPClust were removed from further analysis. Clusters centred at a CCF > 1.5 and ones where chromosome X contributed the highest number of mutations were deemed artifactual, and the respective mutations were removed. Samples with a total number of clonal or subclonal mutations below 256 were also removed. Additionally, samples with poor separation between the clonal and subclonal distributions (*e.g.*, subclone centred at a CCF > 0.80) were removed. Finally, only samples that had both a clone and at least one subclone post-filtering were retained for further analysis. This yielded a total of 223 samples, each with clonal and subclonal mutations. SigProfilerAssignment (v0.0.13)[52] (https://github.com/AlexandrovLab/SigProfilerAssignment) was used to identify the activity of each mutational signature in each clone/subclone, and these activities were then normalized by the total number of mutations belonging to the clone or subclone (that is, clonal mutations were not included in the subclone). A two-sided Wilcoxon signed-rank test[53] was used to assess the differences in the relative activity of each mutational signature between the clones and their respective subclones. P values were corrected using the Benjamini–Hochberg procedure[54] and reported as q values in this Article.

Regressions

Signature attributions were dichotomized into presence and absence using confidence intervals, with presence defined as both lower and upper limits being positive, and absence as the lower limit being zero. If a signature was present in at least 75% of cases (SBS1, SBS40a, SBS40b, ID1 and ID5), it was dichotomized into above and below the median of attributed mutation counts. The binary attributions served as dependent variables in logistic regressions, and relevant risk factors were used as factorized independent variables. To adjust for confounding factors, sex, age of diagnosis, country, and tobacco status were added as covariates in regressions. The Bonferroni method was used to test for significant P values (that is, a total of 224 comparisons for regressions with signatures, and a total of 24 comparisons for regressions with mutation burden). P values reported are raw (not corrected). Regressions with incidence of renal cancer were performed as linear regressions with mutation burdens or signature attributions (those present in at least 75% of cases) with confidence intervals not consistent with zero as a dependent variable, and ASR of renal cancer obtained from

Global Cancer Observatory (GLOBOCAN)[11], sex and age of diagnosis as independent variables. ASR of renal cancer for regions of Czech Republic were obtained from SVOD web portal[55]. Signatures present in less than 75% of cases were dichotomized into presence and absence as previously mentioned and analysed using the logistic regressions. All regressions were performed on a sample basis.

Polygenic risk score analysis of lifestyle risk factors

In this analysis, we used the genome-wide association studies (GWAS) summary statistics estimated in European populations for well-established risk factors for ccRCC. For tobacco smoking status, we used results from the GSCAN consortium meta-analysis of smoking initiation (ever vs never status)[56]. For BMI, the results of UK Biobank and GIANT consortium meta-analysis of continuous BMI were used[57]. GWAS summary statistics related to hypertension, namely systolic blood pressure and diastolic blood pressure, as well as the ones related to diabetes[58], such as fasting glucose and fasting insulin were also obtained using UK Biobank studies[59].

Since all the GWAS summary statistics used in the current work were based on European populations, we used ADMIXTURE tool (v1.3.0)[60] and PCA to infer the unsupervised cluster of individuals with European genetic background within ccRCC cases. Hapmap SNPs ($n = 1,176,821$ variants) were extracted from the ccRCC whole-genome sequence genotype data. After basic quality control using PLINK (v1.9b, www.cog-genomics.org/plink/1.9/), 333 variants were removed due to missing genotype rate > 5%, 1,236 variants failed Hardy–Weinberg equilibrium test (P values $< 10^{-8}$), and 18,702 variants had MAF < 1% in our cohort. Additionally, 3 ambiguous variants and 21,358 variants within regions of long-range, high linkage disequilibrium in the human genome (hg38) were excluded. After pruning for linkage disequilibrium, 143,727 variants remained in ccRCC genotype data. The 1000 genome reference population genotype data (phase 3) for Europeans ($N = 489$), Africans (Yoruba in Ibadan, Nigeria, $N = 108$) and East Asians ($N = 103$ from China and 104 from Japan) (https://www.internationalgenome.org/data/) were filtered and merged with ccRCC genotype data based on the pruned set of variants present in both datasets. ADMIXTURE was run on the merged genotype data with $k = 3$, which would correspond to the 3 ancestral continental population groups that probably reflect the participants of our study. The ccRCC cases with European genetic fraction greater than 80% by the ADMIXTURE analysis were selected for the PRS analyses. To complement the ADMIXTURE analysis, PCA was run on the same samples.

The initial genotype data based on whole-genome sequence from 849 ccRCC cases with European genetic background consisted of biallelic SNPs with MAF > 0.01% (to exclude ultra-rare variants; $N \approx 30$ million variants). After basic quality control, variants with missing genotype rate of greater than 5% ($N = 7,519,196$ variants) with strong deviation from Hardy–Weinberg equilibrium (P values $< 10^{-8}$, $N = 220,862$) were excluded. For each GWAS trait, we restricted our analyses to the biallelic SNPs with minor allele frequency (MAF) greater than 1% in the 1000 Genomes reference for European populations. For the selection of the independent genome-wide significant hits (P values $< 5 \times 10^{-8}$) of each GWAS summary statistic used to generate the PRS, SNPs were clumped ($r^2 = 0.1$ within a linkage disequilibrium window of 10 MB) using PLINK v1.9b[61] (www.cog-genomics.org/plink/1.9/) based on the 1000 Genomes European reference population genotype data ($N = 489$; ~10 million variants). Where a selected GWAS hit was not found in ccRCC genotype data, we extracted proxies ($r^2 > 0.8$ in 1000 Genomes) also present in ccRCC dataset where possible (Supplementary Table 17). The variance of each genetic trait explained by the genetic variants were calculated as previously suggested[62]. PRS was subsequently calculated as the sum of the individual's beta-weighted genotypes using PRSice-2 software[63]. Associations were estimated per standard deviation increase in the PRS, which was normalized to have a mean of zero across ccRCC cases of European genetic ancestry.

Untargeted metabolomics association with signatures

Of the 962 subjects from the main analysis, 901 subjects were included in this sub-study—all Japanese samples ($n = 36$) as well as few cases from Czech Republic ($n = 13$), Romania ($n = 5$) and Russia ($n = 3$) were not included due to lack of available plasma samples. Samples were randomized and analysed as two independent analytical batches. Analysis was performed with a UHPLC-QTOF-MS system that consisted of a 1,290 Binary LC and a 6,550 QTOF mass spectrometer equipped with Jet Stream electrospray ionization source (Agilent Technologies), using previously described methods[64]. Pre-processing was performed using Profinder 10.0.2.162 and Mass Profiler Professional B.14.9.1 software (Agilent Technologies, https://www.agilent.com/). A 'batch recursive feature extraction (small molecules)' process was employed for samples and blanks to find $[M + H]^+$ ions. The two batches were processes separately and the resulting features were aligned in Mass Profiler Professional. Chromatographic peak areas were used as a measurement of intensity. No normalization or transformation of raw data was performed prior to the downstream data analysis.

A total of 2,392 features were detectable in at least one of the 901 samples. Features present in only one of the two batches were filtered out. Recursive filtering elimination was applied to decrease redundancy from highly correlated variables ($r \geq 0.85$, Pearson's r calculated before any transformation/imputation) by selecting the features with least missing data within clusters of features. A total of 944 features were included in the statistical analysis. Features were pre-processed: missing values were replaced with one-fifth of the minimal value of the feature before applying mean centering and Pareto scaling. Each feature was regressed against both de novo and COSMIC signatures, adjusting for sex and age of diagnosis, as well as BMI and technical factors (batch, acquisition order) that could impact chromatographic peak area. Models for SBS22a and SBS22b were restricted to Romanian and Serbian samples to find potential pathways of aristolochic acid exposure in the Balkan region. Logistic models were used for zero-inflated signatures (≥30% zeros) while quasi-Poisson regressions were used for the least zero-inflated signatures (SBS1, SBS40a, and SBS40b). To derive specific false detection rates, random variables were created from permutations of the initial features and regressed against signatures in the same fashion as true features. Maximum P value thresholds from regressions with random features were compared to adjusted P value thresholds according to Bonferroni's procedure. The more conservative approach was used in selecting features of interest. Random forest models were also used as cross-checking multivariate models to assess the relative importance of each feature in explaining the signature attribution. As with univariate models, regression models were used for the least zero-inflated signatures (<30% of zeros) while classification models were used for all other signatures, with restriction to Romanian and Serbian samples for SBS22a and SBS22b. Importance was estimated from the total decrease in node impurities from splitting on the variable, averaged over all trees. Node impurity was measured by the Gini index for classification, and by residual sum of squares for regression. The significance of importance metrics for Random forest models were estimated by permuting the response variable (https://github.com/EricArcher/rfPermute).

Features considered for identification, along with their highly correlated counterparts, were searched in Human Metabolome Database (HMDB), LipidMaps, Metlin and KEGG. Compound identity was confirmed by comparison of retention times and MS/MS fragmentation against chemical standards when available, or otherwise against reference MS/MS spectra. Since the feature 240.1468@0.8929933 was strongly correlated with several features identified as TMAP (Supplementary Table 10), the integration of these features was inspected and corrected manually, and regressed against SBS40b using the same model applied to features selected for analysis. Creatinine was identified among the features by matching its retention time and MS/MS spectra against a reference standard and also regressed against SBS40b in the same fashion as other metabolites. Estimation of correlation between metabolic features was done using linear regression adjusting for batch and acquisition order.

Targeted metabolomics analyses

Circulating levels of per- and polyfluorinated substances (PFAS) and cystatin C compounds were investigated using targeted mass spectrometry-based methods as described previously[30,65].

Out of the 962 subjects from the main analysis, plasma samples from 909 subjects (from all countries except Japan) were randomized and sent frozen in dry ice to each respective laboratory for analyses. Measurement of cystatin C from 906 subjects included its native form and isoforms (3Pro-OH cystatin C, cystatin C-desS, 3Pro-OH cystatin C-desS and cystatin C-desSSP) that were modelled individually and for the total concentration of cystatin C isoforms. Measurements of PFAS compounds included PFOA (total, branch, linear), PFOS (perfluorooctanoic acid; total, branch, linear), PFHxS (perfluorohexane sulfonate), PFNA (perfluorononanoic acid), PFDA (perfluorodecanoic acid), MePFOSAA (*n*-methylperfluoro-1 octanesulfonamido acetic acid) and EtPFOSAA (2-(*N*-ethyl-perfluorooctane. sulfonamido) acetic acid).

Multivariable quasi-Poisson (for the least sparse signatures SBS1, SBS40a and SBS40b) and logistic regression were used to estimate the association between plasma concentrations of the aforementioned substances and mutational signatures. All compounds were modelled continuously ($\log_2$-transformed) and categorically, with adjustments made by sex, age, date of recruitment, country, BMI, tobacco and alcohol status in the case of PFAS molecules and by sex, age and BMI, in the case of cystatin C.

Geospatial analyses

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

Statistics and reproducibility

Analyses were conducted using R version 4.1 (ref. 66) and python version 3.9.13 (ref. 67). Handling of geospatial and other data was conducted using the R packages lme4, matrixStats, Matrix, geojsonio, raster, rgeos, sf, sp, tmaptools, patchwork, leaflet, data.table, dplyr, haven, Hmisc, openxlsx, rgdal, scales, stringr, tidyr, tibble, xlsx, rfPermute, randomForest, forcats, and in python using the packages pandas, numpy, scipy, statsmodels, firthlogist, patsy and jupyter[68–97]. Figures were created using ggplot, ggnewscale, ggpattern, ggrepel, ggsflabel, ggspatial, ggpubr, cowplot, matplotlib, plotly (https://plot.ly), seaborn and TMB_plotter[98–108]. Open-source maps of Czech Republic, Romania and Serbia were obtained from the Global Administrative Areas project[109] (https://gadm.org), with the surrounding borders obtained from the Natural Earth project[110] (https://www.naturalearthdata.com/). Signature extraction was replicated two times independently at both Wellcome Sanger Institute and UCSD, with similar results. Signature attribution was replicated two times independently at both Wellcome Sanger Institute and IARC, with similar results. All attempts at replication were successful. No other experiments other than those mentioned here were replicated independently due to limited resources. Additional details relating to the methods used in this study can be found in Supplementary Figs. 1–27 and Supplementary Note Tables 1–10.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Whole-genome sequencing data and patient metadata are deposited in the European Genome–Phenome Archive (EGA) associated with study EGAS00001003542. Aligned BAM files for all ccRCC cases included in the final analysis are deposited in dataset EGAD00001012102, consensus SNV and indel variant calling files are in dataset EGAD00001012222, patient metadata are in dataset EGAD00001012223, structural rearrangement variant calling files are in dataset EGAD00001013726 and copy number variant calling are in dataset EGAD00001013727. Mutational catalogues for the PCAWG dataset can be accessed at https://dcc.icgc.org/releases/PCAWG. Data used for validation of SBS12 in additional cohorts can be retrieved from the original publication[25] (validation cohort 1) and EGA dataset EGAD00001009866 (validation cohort 2). The metabolomics data have been uploaded to the MetaboLights repository as study MTBLS9394. The human reference genome used for alignment is available at ftp://ftp.sanger.ac.uk/pub/cancer/support-files/reference/GRCh38_full_analysis_set_plus_decoy_hla.fa. All other data are provided in the accompanying Supplementary Tables.

Code availability

All algorithms used for data analysis are publicly available with repositories noted within the respective method sections and in the accompanying reporting summary. Code used for regression, geospatial, metabolomics and germline analyses as well as figures is available at https://gitlab.com/Mutographs/Mutographs_RCC.

39. Whalley, J. P. et al. Framework for quality assessment of whole genome cancer sequences. *Nat. Commun.* **11**, 5040 (2020).
40. Bergmann, E. A., Chen, B.-J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and contamination estimator for matched tumor–normal pairs. *Bioinformatics* **32**, 3196–3198 (2016).
41. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
42. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinform.* **56**, 15.10.1–15.10.18 (2016).
43. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinform.* **52**, 15.7.1–15.7.12 (2015).
44. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
45. Bergstrom, E. N. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genom.* **20**, 685 (2019).
46. Liu, M., Wu, Y., Jiang, N., Boot, A. & Rozen, S. G. mSigHdp: hierarchical Dirichlet process mixture modeling for mutational signature discovery. *NAR Genom. Bioinform.* **5**, lqad005 (2023).
47. Islam, S. M. A. et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom.* **2**, 100179 (2022).
48. Senkin, S. MSA: reproducible mutational signature attribution with confidence based on simulations. *BMC Bioinform.* **22**, 540 (2021).
49. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
50. Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
51. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
52. Díaz-Gay, M. et al. Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. *Bioinformatics* **39**, btad756 (2023).
53. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bull.* **1**, 80 (1945).
54. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

55. Dušek, L. et al. Epidemiology of Malignant Tumours in the Czech Republic, Version 7.0. *Masaryk University* http://www.svod.cz (2007).
56. Liu, M. et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
57. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
58. Lagou, V. et al. Sex-dimorphic genetic effects and novel loci for fasting glucose and insulin variability. *Nat. Commun.* **12**, 24 (2021).
59. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
60. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
61. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
62. Shim, H. et al. A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PLoS ONE* **10**, e0120758 (2015).
63. Choi, S. W. & O'Reilly, P. F. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).
64. Loftfield, E. et al. Novel biomarkers of habitual alcohol intake and associations with risk of pancreatic and liver cancers and liver disease mortality. *J. Natl Cancer Inst.* **113**, 1542–1550 (2021).
65. Gao, J., Meyer, K., Borucki, K. & Ueland, P. M. Multiplex Immuno-MALDI-TOF MS for targeted quantification of protein biomarkers and their proteoforms related to inflammation and renal dysfunction. *Anal. Chem.* **90**, 3366–3373 (2018).
66. R Core Team. *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/ (R Foundation for Statistical Computing, 2022).
67. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, 2009).
68. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, https://doi.org/10.18637/jss.v067.i01 (2015).
69. Bengtsson H. matrixStats: Functions that apply to rows and columns of matrices (and to vectors). https://cran.r-project.org/web/packages/matrixStats/index.html (2023).
70. Bates, D. et al. Matrix: Sparse and dense matrix classes and methods. https://matrix.r-forge.r-project.org/ (2023).
71. Chamberlain, S., Teucher, A. & Mahoney, M. geojsonio. https://github.com/ropensci/geojsonio (2023).
72. Hijmans, R. J. raster. https://github.com/rspatial/raster (2024).
73. Bivand, R. & Rundel, C. rgeos. https://github.com/cran/rgeos/ (2023).
74. Pebesma, E. & Bivand, R. *Spatial Data Science: With Applications in R* (Chapman and Hall/CRC, 2023).
75. Tennekes, M. tmaptools: Thematic map tools. https://github.com/r-tmap/tmaptools (2021).
76. Lin Pedersen, T. patchwork: The composer of plots. https://github.com/thomasp85/patchwork (2024).
77. Cheng, J. Leaflet. https://github.com/rstudio/leaflet (2023).
78. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A grammar of data manipulation. https://dplyr.tidyverse.org (2023).
79. Wickham, H., Miller, E. & Smith, D. haven: Import and export 'SPSS', 'Stata' and 'SAS' files. https://haven.tidyverse.org (2023).
80. Harrell Jr F. Hmisc: Harrell Miscellaneous. R package version 5.1-1. https://hbiostat.org/r/hmisc/ (2023).
81. Schauberger, P. & Walker, A. openxlsx: Read, write and edit xlsx files. https://github.com/ycphs/openxlsx (2022).
82. Bivand, R., Keitt, T. & Rowlingson, B. rgdal: Bindings for the 'Geospatial' data abstraction library. https://rgdal.r-forge.r-project.org (2023).
83. Wickham, H., Pedersen, T. L. & Seidel, D. scales: Scale functions for visualization. https://scales.r-lib.org (2023).
84. Wickham, H. stringr: Simple, consistent wrappers for common string operations. https://stringr.tidyverse.org (2023).
85. Wickham, H., Vaughan, D. & Girlich, M. tidyr: Tidy messy data. https://tidyr.tidyverse.org (2024).
86. Müller, K. & Wickham, H. tibble: Simple data frames. https://tibble.tidyverse.org/ (2023).
87. Dragulescu, A. & Arendt, C. xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. https://github.com/colearendt/xlsx (2022).
88. Archer, E. rfPermute. https://github.com/EricArcher/rfPermute (2023).
89. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
90. Wickham, H. forcats: Tools for working with categorical variables (factors). https://forcats.tidyverse.org/ (2023).
91. The pandas development team. pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.7093122 (2022).
92. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
93. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
94. Seabold, S. & Perktold, J. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference* https://doi.org/10.25080/Majora-92bf1922-011 (2010).
95. Luo, J. firthlogist. https://github.com/jzluo/firthlogist (2022).
96. Smith, N. J. et al. pydata/patsy. https://doi.org/10.5281/zenodo.5529350 (2021).
97. Kluyver, T. et al. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (eds Loizides, F. & Schmidt, B.) 87–90 (2016).
98. Wickham, H. ggplot2: Elegant graphics for data analysis. https://ggplot2.tidyverse.org (2016).
99. Campitelli, E. ggnewscale: Multiple fill and color scales in ggplot2. https://doi.org/10.5281/zenodo.7971612 (2023).
100. Mike, F. C., Davis, T. L. & ggplot2 authors. ggpattern: 'ggplot2' pattern geoms. https://github.com/trevorld/ggpattern (2022).
101. Slowikowski, K. ggrepel: Automatically position non-overlapping text labels with 'ggplot2'. https://github.com/slowkow/ggrepel (2024).
102. Yutani, H. ggsflabel. https://yutannihilation.github.io/ggsflabel/ (2023).
103. Dunnington, D. ggspatial: Spatial data framework for ggplot2. https://paleolimbot.github.io/ggspatial/ (2023).
104. Kassambara, A. ggpubr: 'ggplot2' based publication ready plots. https://rpkgs.datanovia.com/ggpubr/ (2023).
105. Wilke, C. O. cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. https://wilkelab.org/cowplot/ (2024).
106. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
107. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
108. He, Y. TMB_plotter. https://github.com/AlexandrovLab/TMB_plotter (2020).
109. Global Administrative Areas. GADM v4.1. https://gadm.org (2022).
110. Patterson, T. & Nathaniel, V. World Countries, 1:10 million. Natural Earth v5.1.1. https://www.naturalearthdata.com (2022).

Acknowledgements The authors thank L. O'Neill, K. Roberts, K. Smith, M. Farenden, S. Austin-Guest and the staff of DNA Pipelines at the Wellcome Sanger Institute for their contribution; M. Milojevic, C. Lallemand, H. Renard, A. Bardot, A. Spanu and N. Robinot for support; IARC General Services, including the Laboratory Services and Biobank team led by Z. Kozlakidis, the Section of Support to Research overseen by T. Landesz and the Evidence Synthesis and Classification Section led by I. Cree, under IARC regular budget funding; G. Bergo, R. Cox and J. Oliveira for help with data and sample preparation and processing; P. Campbell, I. Martincorena, T. Butler, D. Mariosa, L. T. Fontanals, M. Olanipekun, W. O. Dos Santos, H. Zahed, M. Gunter, M. Blanks and M. McCord for useful discussions; and all the patients involved in this study and their families. The authors acknowledge the contributions of the Leeds Biobanking and Sample Processing Lab, the Leeds Multidisciplinary RTB and the Leeds NIHR BioRTB for provision of samples. 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 partly supported or 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.R.B. and B.S.N.), the Practical Research Project for Innovative Cancer Control from the Japan Agency for Medical Research and Development (AMED) (JP20ck0106547h0001 to T.S.), the National Cancer Center Japan Research and Development Fund (2020-A-7 to A.F.), 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.) and 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. Measurement of cystatin C was funded by Cancer Research UK (C18281/A29019). This publication presents data from the DISCERN study funded by the Horizon Europe 2021–Cancer Mission (grant no. 101096888).

Author 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. Fitzgerald, M.K., R.V., A.P.L., E.N.B., A.K., B.O., S.C., E.T., J.R.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. Ferlicot and M. Asgari. 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. Albert, L.P., P.A.-P., M.R.B., B.S.N., S.M.B., M.P.C., S.d.C.Z., R.M.R., E.F.F., N.S.d.M., R.S.F., R.E.B., N.S.V., D.Z., A.M., O.S., V.M., L.F., M.N., I.H., A.H., V. Janout, S.C. and C.L. P.K.-R., S.J.C., N.R., M.P.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 L.B.A. is a compensated consultant and has equity interest in io9 and Genome Insight. His spouse is an employee of Biotheranostics. L.B.A. is also an inventor of a US Patent 10,776,718 for source identification by non-negative matrix factorization. E.N.B. and L.B.A. declare US provisional patent applications with serial numbers: 63/289,601, 63/269,033 and 63/483,237. L.B.A. also declares US provisional patent applications with serial numbers: 63/366,392, 63/367,846, 63/412,835 and 63/492,348. V.M. received honoraria from Ipsen, Bayer, AstraZeneca, Janssen, Astellas Pharm and MSD, and provided expert testimony to BMS, Bayer, MSD and Janssen. The other authors declare no competing interests.

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

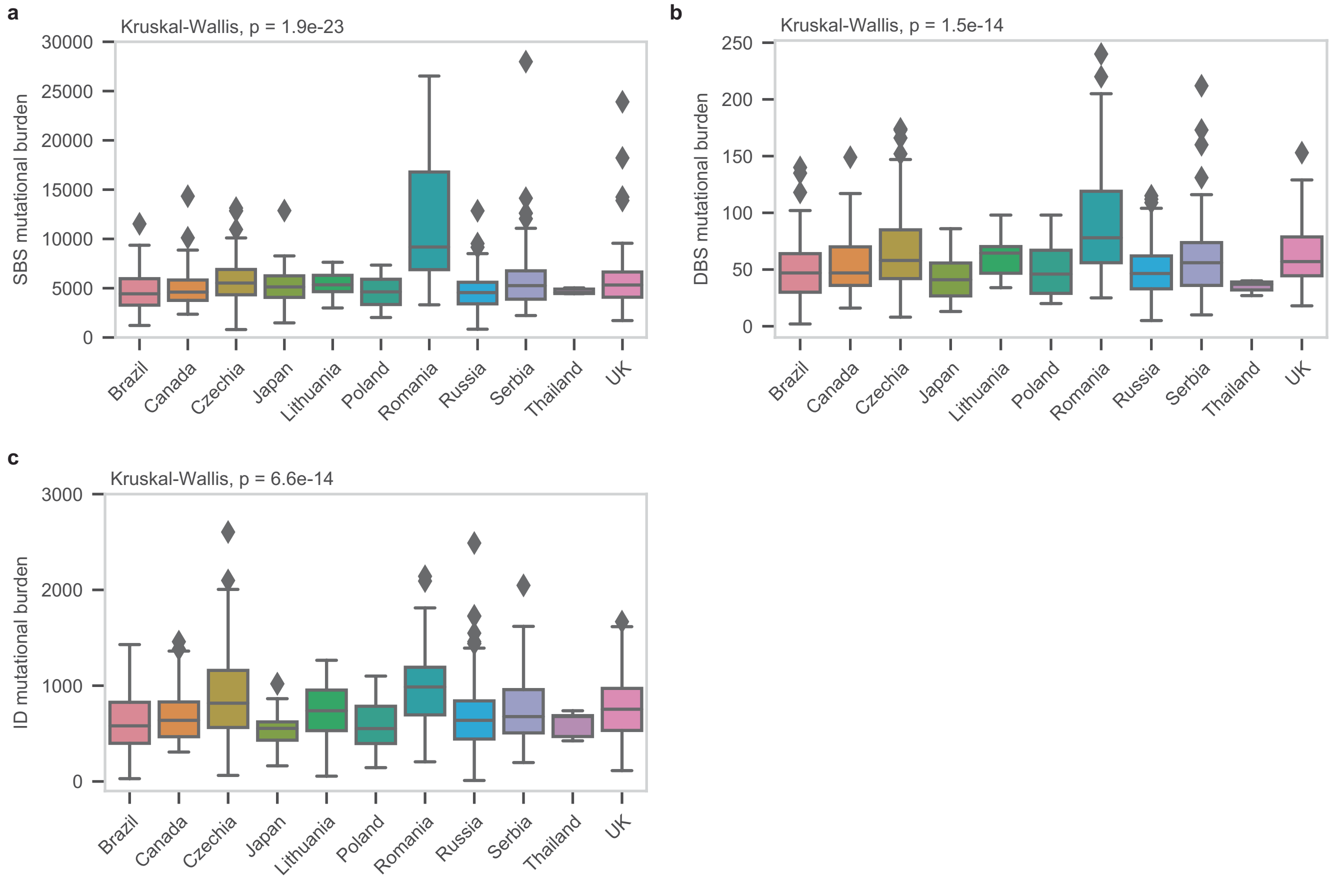
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07368-2.

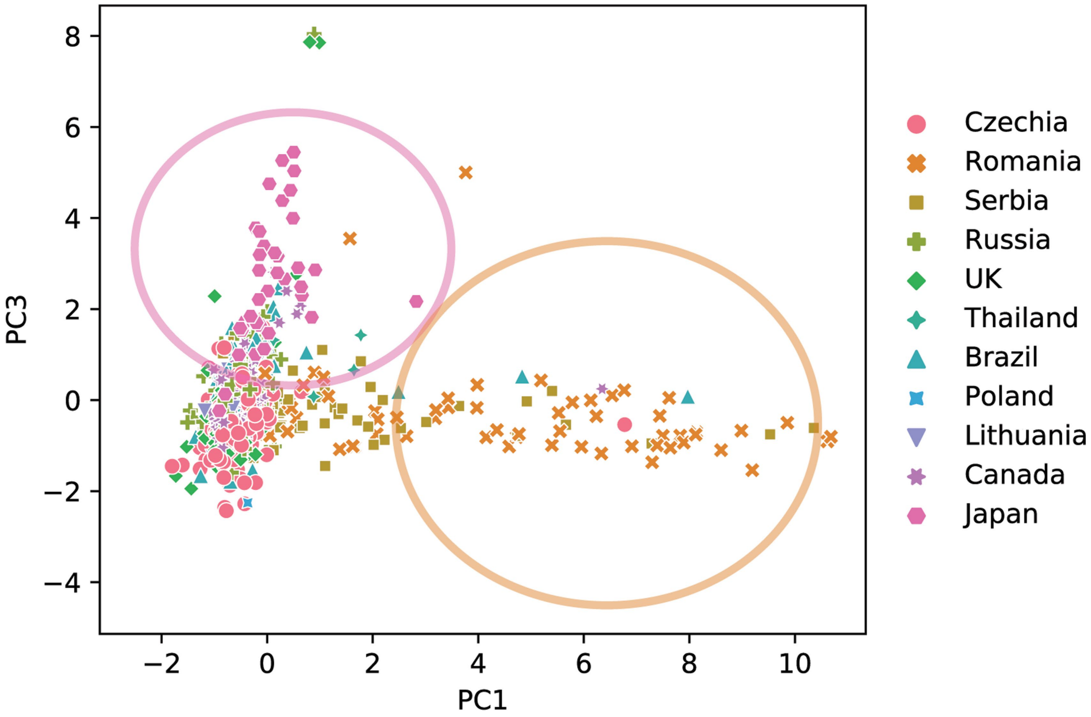
Correspondence and requests for materials should be addressed to Paul Brennan.

Peer review information *Nature* thanks Fran Supek, Jason Wong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

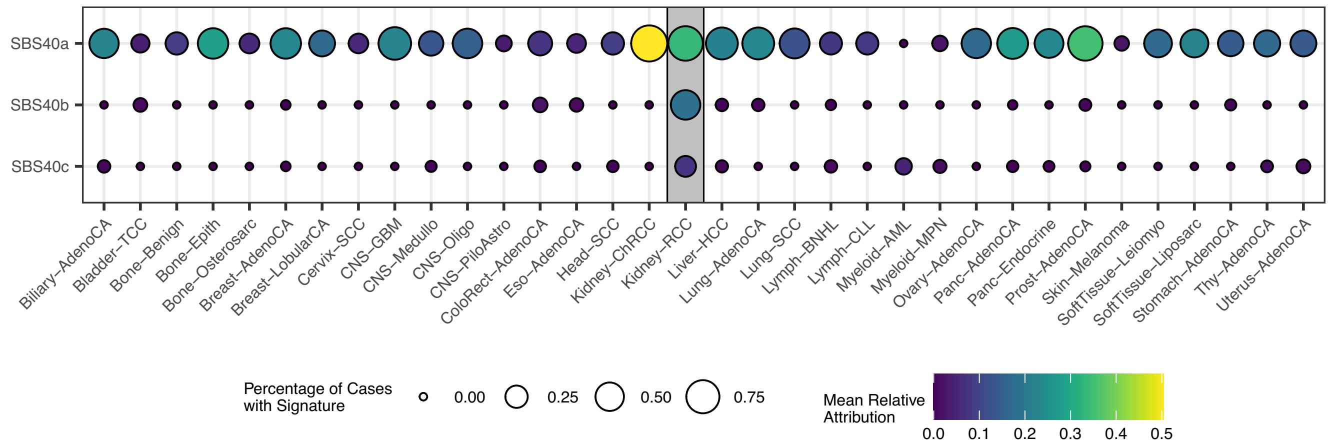
Reprints and permissions information is available at http://www.nature.com/reprints.



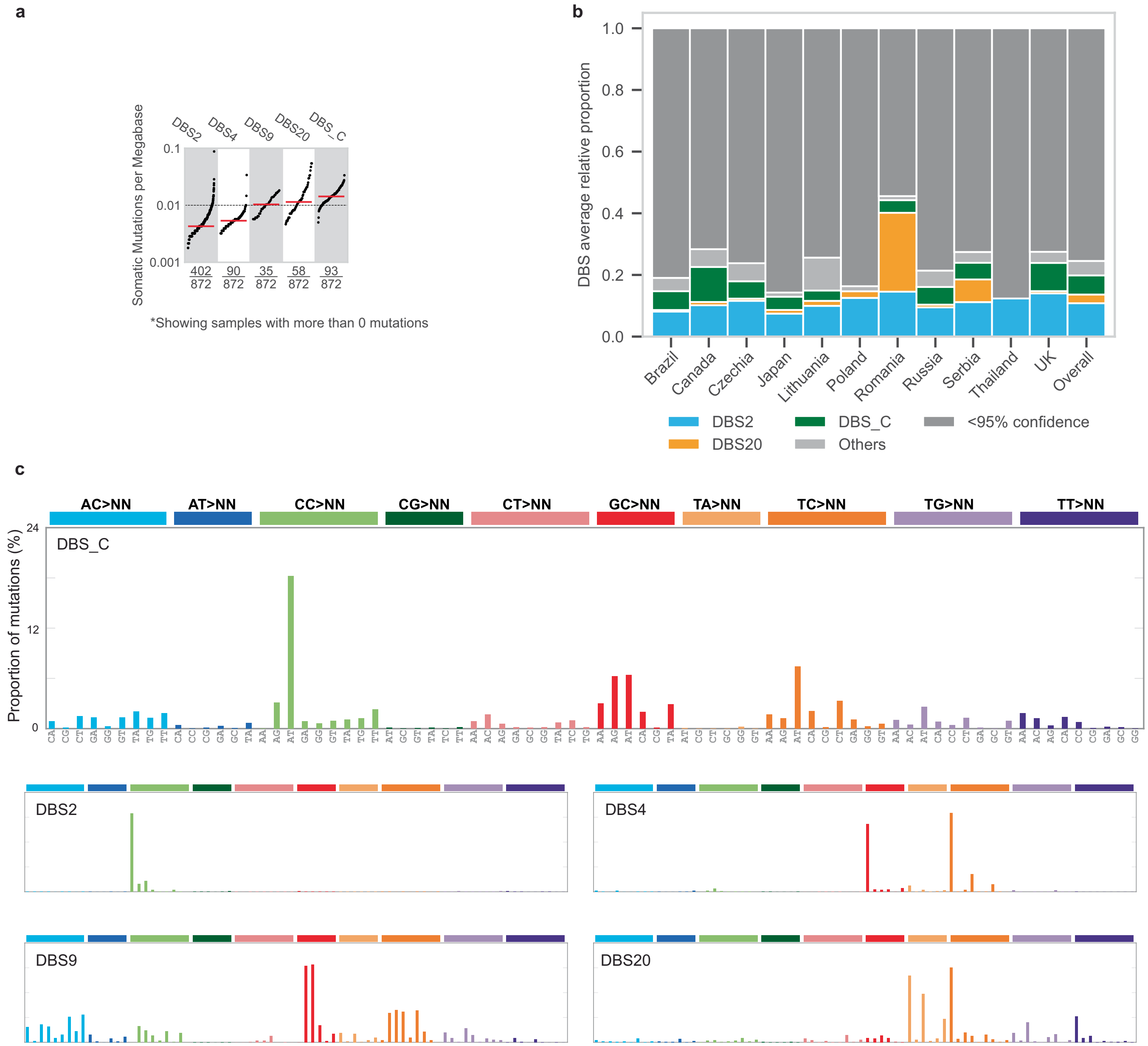
Extended Data Fig. 1 | Mutation burdens in clear cell renal cell carcinomas across countries. Mutation burdens for single base substitutions (SBS) (**a**), doublet base substitutions (DBS) (**b**) and small insertions and deletions (ID) (**c**) show significant differences between countries using the Kruskal-Wallis (two-sided) test (n = 961 biologically independent samples over 1 independent experiment). Four SBS hypermutators and four ID hypermutators above mutation burden of 30000 and 3000, respectively, 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.



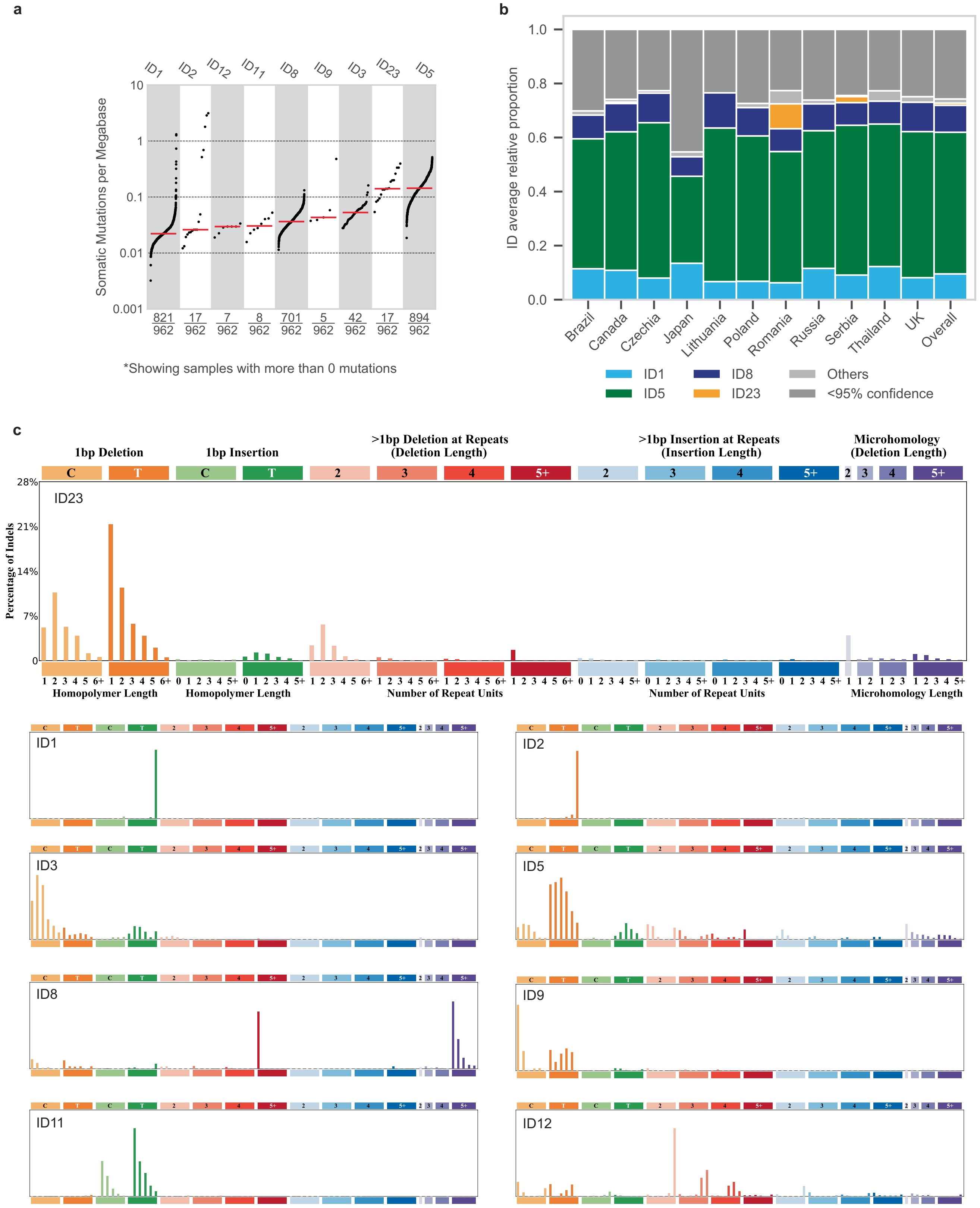
Extended Data Fig. 2 | Principal component analysis of relative mutation counts. PCA performed on relative mutation counts of all ccRCC tumors incorporating the six mutation classes (C > A, C > G, C > T, T > A, T > C, T > G). Principal component 1 (PC1) clearly separates the cluster of mostly Romanian cases that are enriched with AA signatures, often at high mutation burdens. Principal component 3 (PC3) identifies a cluster of mostly Japanese cases, enriched with signature SBS12.



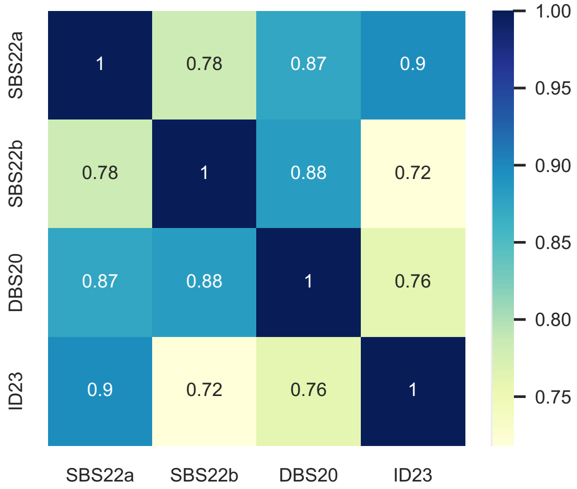
Extended Data Fig. 3 | Attribution of signatures SBS40a, SBS40b, and SBS40c in a pan-cancer cohort. Attribution of signatures SBS40a, SBS40b, and SBS40c in a pan-cancer cohort, showing a widespread distribution for SBS40a whilst SBS40b and SBS40c are only seen consistently in clear cell renal cell carcinomas (ccRCC). The size of each dot represents the proportion of samples of each tumor type where the signature is present. The color of each dot represents the average mutation burden.



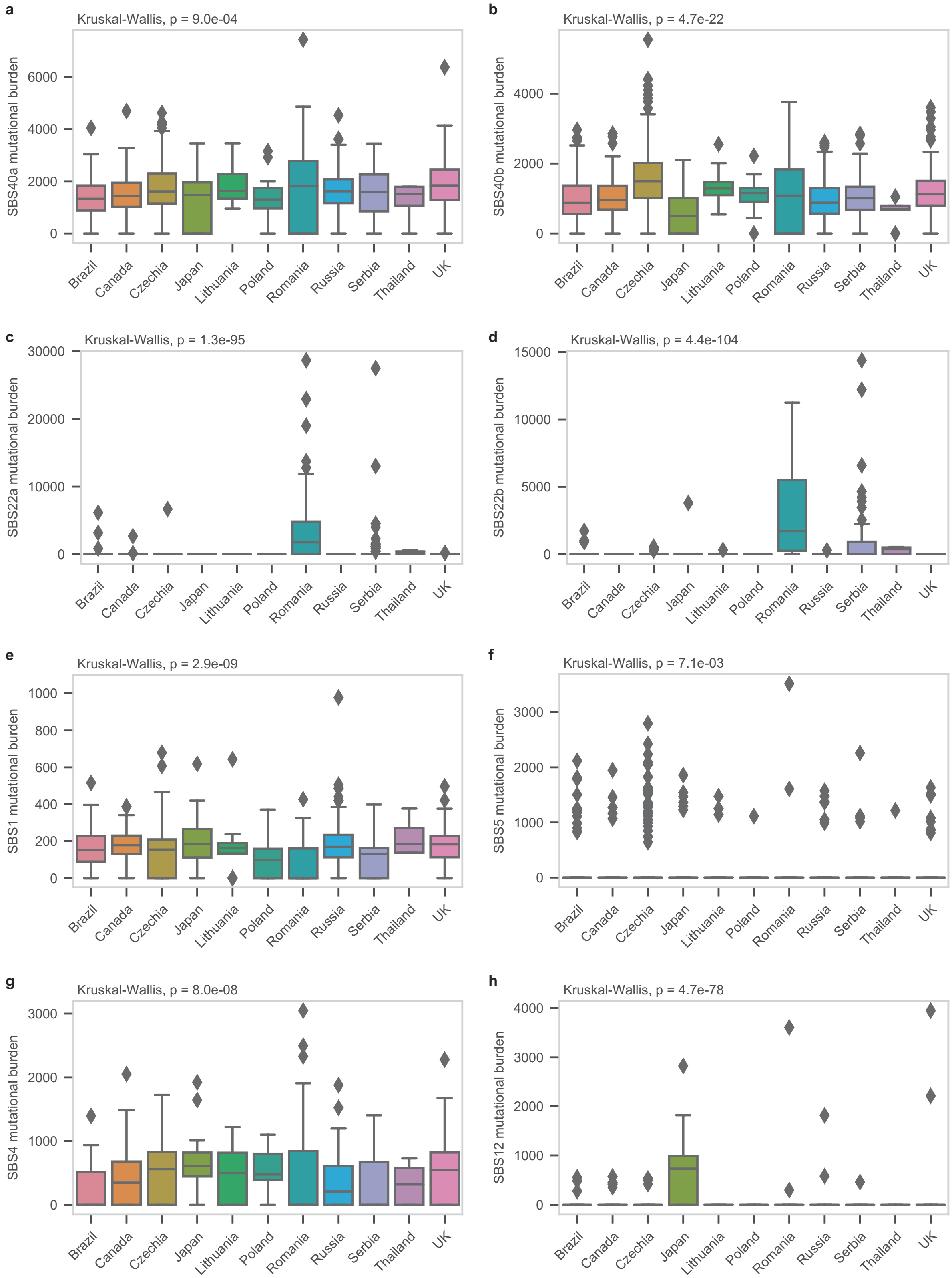
Extended Data Fig. 4 | Doublet-base substitution signatures operative in clear cell renal cell carcinomas. (**a**) Tumour mutation burden (TMB) plot showing the frequency and mutations per Mb for each of the decomposed DBS signatures. (**b**) Average relative attribution for doublet-base substitution (DBS) signatures across countries. Signatures contributing less than 5% on average are grouped in the 'Other' category, apart from signature DBS20. Category named '<95% confidence' accounts for the proportion of mutation burden which could not be assigned to any signature with confidence level of at least 95%. (**c**) Decomposed DBS signatures, including reference COSMIC signatures as well as de novo signatures not decomposed into COSMIC reference signatures.



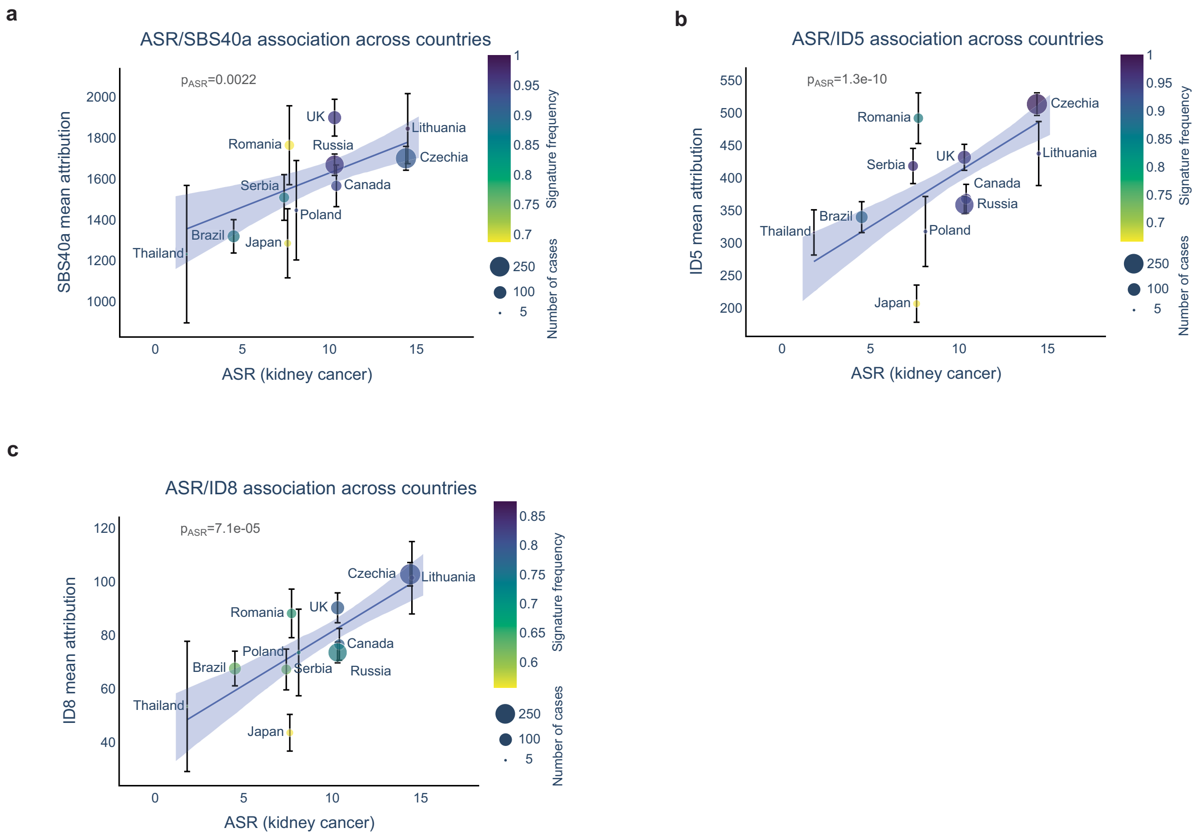
Extended Data Fig. 5 | Small insertions and deletion signatures operative in clear cell renal cell carcinomas. (**a**) Tumour mutation burden (TMB) plot showing the frequency and mutations per Mb for each of the decomposed ID signatures. (**b**) Average relative attribution for small insertion and deletion (ID) signatures across countries. Signatures contributing less than 5% on average are grouped in the 'Others' category, apart from signature ID23. Category named '<95% confidence' accounts for the proportion of mutation burden which could not be assigned to 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.



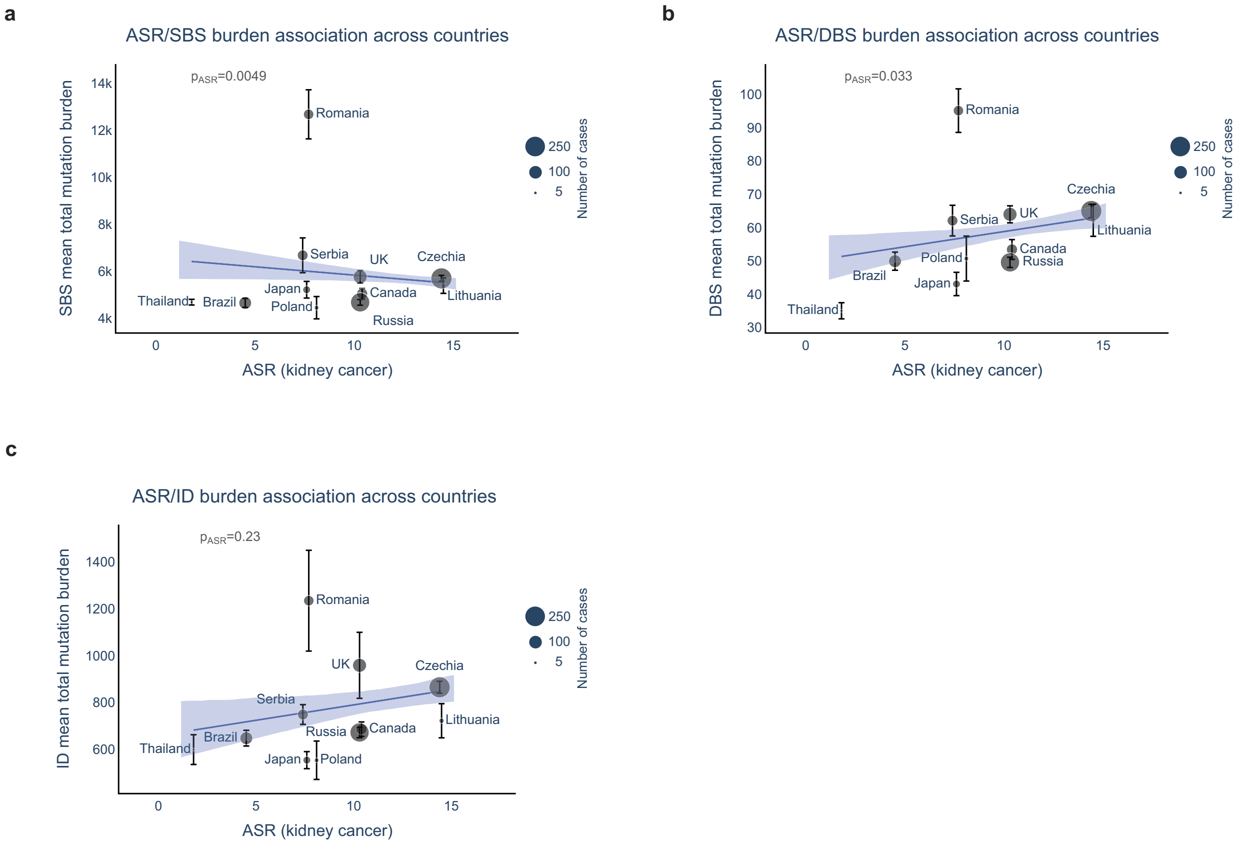
Extended Data Fig. 6 | Correlation between signatures SBS22a, SBS22b, DBS20, ID23. Heatmap of pairwise Pearson correlation between signatures SBS22a, SBS22b, DBS20 and ID23. Numbers and colors indicate correlation coefficient.



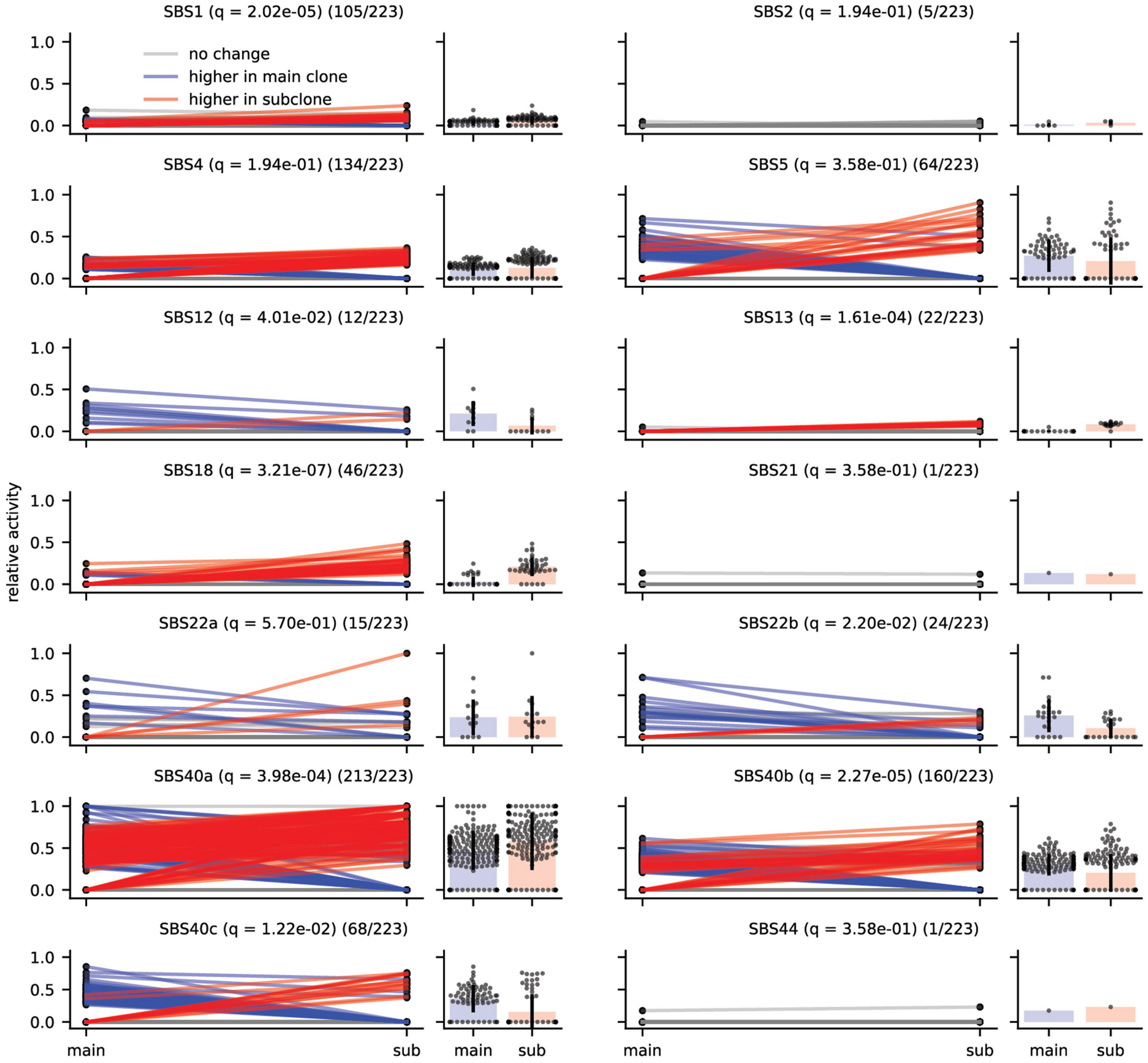
Extended Data Fig. 7 | Single base substitution signatures showing significant differences in attributed mutation burden between countries. Signatures SBS40a (**a**) and SBS40b (**b**) were more prevalent in high-incidence regions of Czech Republic and Lithuania. Signatures SBS22a (**c**) and SBS22b (**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. N = 961 biologically independent samples examined over 1 independent experiment.



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. Data are presented as mean values +/− SEM (n = 961 biologically independent samples examined over 1 independent experiment). The p-values shown are for the ASR variable in linear regressions across all cases, adjusted for sex and age of diagnosis.

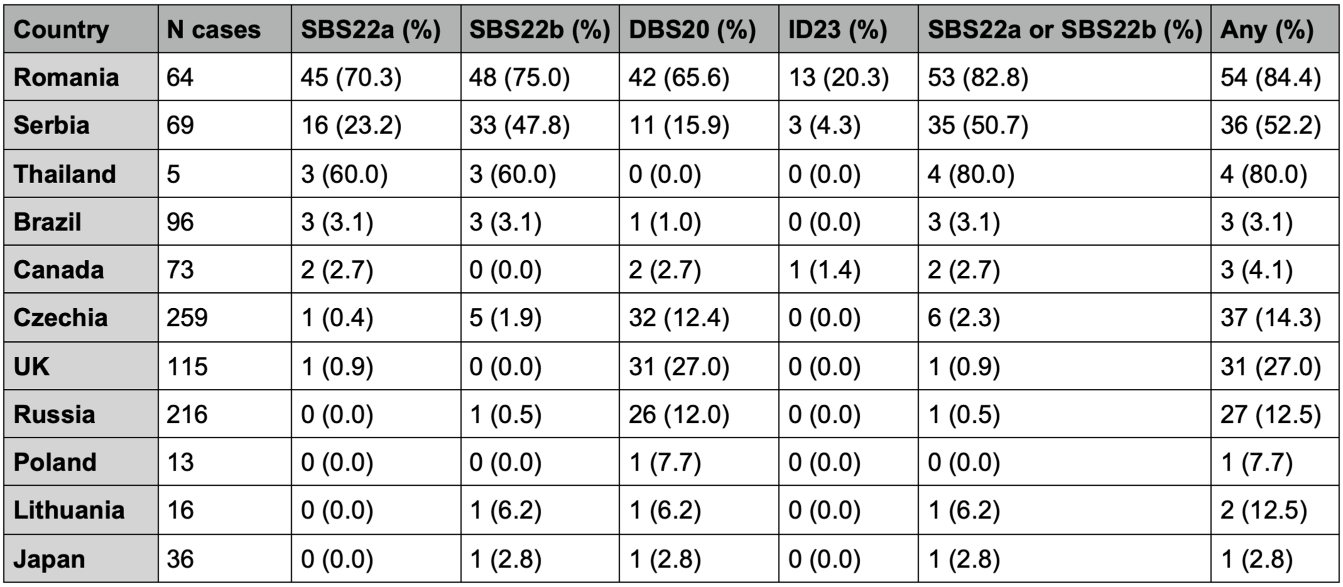


Extended Data Fig. 9 | Association of mutation burden with incidence of renal cancer. Association of age-standardized rates (ASR) of kidney cancer incidence with SBS (**a**), DBS (**b**) and ID (**c**) mutation burdens across countries. Data are presented as mean values +/− SEM (n = 961 biologically independent samples examined over 1 independent experiment). The p-values shown are for the ASR variable in linear regressions across all cases, adjusted for sex and age of diagnosis.



Extended Data Fig. 10 | Evolutionary analysis of mutational signatures in ccRCC. Comparison of mutational signatures between clonal and subclonal mutations. Lines show the change in relative activity between the clonal mutations (main) and subclonal mutations (sub) within a sample. Blue and red lines represent an activity change of more than 6% (blue indicates higher in the clonal mutations; red indicates higher in the subclonal mutations). Bar plots show the distribution of activities in samples where the signature was present in the clonal and/or subclonal mutations; this number is represented in the title of each plot as X/223 for each signature (n = 223 biologically independent samples examined over 1 independent experiment). Black bars indicate one standard deviation away from the mean. Significance was assessed using a two-sided Wilcoxon signed-rank test, and q-values were generated using the Benjamini-Hochberg Procedure.

Extended Data Table 1 | Presence of signatures SBS22a, SBS22b, DBS20, ID23 across countries



Country	N cases	SBS22a (%)	SBS22b (%)	DBS20 (%)	ID23 (%)	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)

Presence of signatures SBS22a, SBS22b, DBS20, ID23, as well as the union of signatures SBS22a/SBS22b and the union of all signatures across different countries in the cohort. Whole numbers indicate the numbers of cases with signatures attributed, whereas percentages indicate the frequencies of signatures in each country.

nature portfolio

Corresponding author(s): Paul Brennan

Last updated by author(s): Feb 16, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Whole genome sequencing (150bp paired end) was performed on the Illumina NovaSeq 6000 platform with target coverage of 40X for tumors and 20X for paired blood. REDCap 13.1.27 was used to collect epidemiological data for cases from Barretos, Porto Alegre, Sao Paulo, Leeds, Vilnius, Hat Yai, Belgrade and Bucharest
Data analysis	Algorithms used: Variant Calling pipelines (available at https://github.com/cancerit): BWA-Mem v0.7.17-r1188 ASCAT v4.3.3 and v4.5.0 BATTENBERG v3.5.3 cgpCaVEMan v1.11.2, v1.14.1 and v1.15.1 cgpPINDEL v2.2.5, v3.3.0 and v3.5.0 BRASS v6.1.2, v6.2.0, v6.3.0 and v6.3.4 Strelka2 v2.9.10 and Manta 1.6.0 Other packages: Conpair v0.2 (https://github.com/nygenome/Conpair) SigProfilerMatrixGenerator v1.2.12 (https://github.com/AlexandrovLab/SigProfilerMatrixGenerator) SigProfilerExtractor v1.1.9: (https://github.com/AlexandrovLab/SigProfilerExtractor) MSA v2.0 (https://gitlab.com/s.senkin/MSA) mSigHdp v2.0.1 (https://github.com/steverozen/mSigHdp)

Cancer Genome Interpreter 2022 (https://www.cancergenomeinterpreter.org)
MutationMapper tool v6.0.0 (http://www.cbioportal.org/mutation_mapper)
DPClust R package v2.2.8 (https://github.com/Wedge-lab/dpclust)
dpclust3p v1.0.8 (https://github.com/Wedge-lab/dpclust3p)
SigProfilerAssignment (v0.0.13)
SnpEff 5.0e (https://pcingola.github.io/SnpEff/)
ADMIXTURE v1.3.0 (https://dalexander.github.io/admixture/)
PLINK v1.9 and v2.00a (www.cog-genomics.org/plink/2.0/)
PRSice 2.3.3 (https://choishingwan.github.io/PRSice/)
Profinder 10.0.2.162 (https://www.agilent.com/)
Mass Profiler Professional B.14.9.1 (https://www.agilent.com/)
lme4 1.1-34 (https://github.com/lme4/lme4/)
matrixStats 1.0.0 (https://github.com/HenrikBengtsson/matrixStats)
Matrix 1.6-1.1 (https://matrix.r-forge.r-project.org/)
geojsonio 0.11.3 (https://github.com/ropensci/geojsonio)
ggnewscale 0.4.9 (https://eliocamp.github.io/ggnewscale/)
ggpattern 1.0.1 (https://github.com/trevorld/ggpattern)
ggrepel 0.9.3 (https://github.com/slowkow/ggrepel)
ggsflabel 0.0.1 (https://yutannihilation.github.io/ggsflabel/)
ggspatial 1.1.9 (https://paleolimbot.github.io/ggspatial/)
ggpubr 0.6.0 (https://github.com/kassambara/ggpubr/)
raster 3.6-23 (https://github.com/rspatial/raster)
rgeos 0.6-4 (https://github.com/cran/rgeos/)
sf 1.0-14 (https://github.com/r-spatial/sf)
sp 2.0-0 (https://github.com/edzer/sp/)
tmaptools 3.1-1 (https://github.com/r-tmap/tmaptools)
patchwork 1.1.3 (https://github.com/thomasp85/patchwork)
leaflet 2.2.0 (https://github.com/rstudio/leaflet)
ggplot2 3.4.3 (https://github.com/tidyverse/ggplot2)
cowplot 1.1.1 (https://wilkelab.org/cowplot/)
data.table 1.14.8 (https://github.com/Rdatatable/data.table)
dplyr 1.1.3 (https://github.com/tidyverse/dplyr)
haven 2.5.3 (https://github.com/tidyverse/haven)
Hmisc 5.1-1 (https://hbiostat.org/r/hmisc/)
openxlsx 4.2.5.2 (https://github.com/ycphs/openxlsx)
rgdal 1.6-7 (https://rgdal.r-forge.r-project.org/)
scales 1.2.1 (https://github.com/r-lib/scales)
stringr 1.5.0 (https://github.com/tidyverse/stringr)
tidyr 1.3.0 (https://github.com/tidyverse/tidyr)
tibble 3.2.1 (https://github.com/tidyverse/tibble)
xlsx 0.6.5 (https://github.com/colearendt/xlsx)
rfPermute 2.5.2 (https://github.com/EricArcher/rfPermute)
randomForest 4.7-1.1 (https://www.stat.berkeley.edu/~breiman/RandomForests/)
forcats 1.0.0 (https://github.com/tidyverse/forcats)
jupyter 1.0.0 (https://jupyter.org/)
pandas 1.5.0 (https://pandas.pydata.org/)
numpy 1.23.3 (https://numpy.org/)
scipy 1.9.1 (https://scipy.org/)
statsmodels 0.13.2 (https://www.statsmodels.org/)
firthlogist 0.5.0 (https://github.com/jzluo/firthlogist)
patsy 0.5.2 (https://github.com/pydata/patsy)
matplotlib 3.3.4 (https://matplotlib.org/)
seaborn 0.12.0 (https://seaborn.pydata.org)
plotly 5.10.0 (https://plot.ly)
TMB_plotter (https://github.com/AlexandrovLab/TMB_plotter)

Statistical analysis was performed in R version 4.1 and Python version 3.9.13

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Whole genome sequencing data and patient metadata are deposited in the European Genome-phenome Archive (EGA) associated with study EGAS00001003542. Aligned BAM files for all ccRCC cases included in the final analysis were deposited in dataset EGAD00001012102, consensus SNV and indel variant calling files in dataset EGAD00001012222, patient metadata in dataset EGAD00001012223, structural rearrangement variant calling files in dataset EGAD00001013726 and copy number variant calling in dataset EGAD00001013727. Mutational catalogs for the PCAWG dataset can be accessed at https://dcc.icgc.org/releases/PCAWG. Data used for validation of SBS12 in additional cohorts can be retrieved from the original publication (validation cohort 1)32 and EGA dataset EGAD00001009866 (validation cohort 2). The metabolomics data have been uploaded to the MetaboLights repository as study MTBLS9394. The human reference genome used for alignment is available at ftp://ftp.sanger.ac.uk/pub/cancer/support-files/reference/GRCh38_full_analysis_set_plus_decoy_hla.fa. All other data are provided in the accompanying Supplementary Tables.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender

Sex information was self-reported and collected using epidemiological questionnaires. Overall numbers are provided in the population characteristics section of the Reporting summary. Consent was obtained for sharing individual-level data. Sex-based epidemiological regressions were performed, as described in the Methods section.

Reporting on race, ethnicity, or other socially relevant groupings

To infer the individuals with European genetic background, ADMIXTURE tool and principal component analysis (PCA) were used as described in the Methods section. This variable was not used as a proxy for any other socially constructed variables. Being an unbiased estimate based on genotyping data, no confounding variables were controlled for in the relevant analyses.

Population characteristics

962 cases (380 women and 582 men) diagnosed with ccRCC were included from the following countries: A total of 962 ccRCC cases from 11 countries in four continents were studied, encompassing: Czech Republic (n=259), Russia (n=216), United Kingdom (n=115), Brazil (n=96), Canada (n=73), Serbia (n=69), Romania (n=64), Japan (n=36), Lithuania (n=16), Poland (n=13), and Thailand (n=5). Age of diagnosis ranging from 23 to 87 y.o., mean (SD): 60 (12).

Recruitment

IARC/WHO coordinated cases recruitment through an international network of collaborators in 11 countries. The inclusion criteria for patients were >=18 years of age, confirmed diagnosis of primary renal cell carcinoma (RCC) and no prior treatment. Patients were excluded if they had any condition that could interfere with their ability to provide informed consent or if there were no means of obtaining adequate tissue/ blood samples as per the protocol requirements. The authors are not aware of any potential self-selection bias or other biases present

Ethics oversight

Ethical approvals were obtained from each Local Research Ethics Committee and Federal Ethics Committee as listed below. The study was submitted and approved by the IARC Ethics Committee. Informed consent was obtained from all participants.
Barretos Cancer Institute, Barretos, Sao Paulo, Brazil
Hospital de Clinicas de Porto Alegre, Porto Alegre, Brazil
A.C.Camargo Cancer Center, Sao Paulo, Brazil
Ontario Tumor Bank, Canada
Charles University in Prague, 2nd faculty of Medicine, Prague, Czech Republic
National Cancer Center, Tokyo, Japan
National Cancer Institute, Vilnius, Lithuania
Nofer Institute of Occupational Medicine, Warsaw, Poland
University of Medicine and Pharmacy “Carol Davila”, Bucharest, Romania
N.N. Blokhin Cancer Research Center, Moscow, Russian Federation
International Organization for Cancer Prevention and Research, Belgrade, Serbia
Faculty of Medicine Prince of Songkla, Hat Yai, Thailand
St James’ University Hospital, Leeds, United Kingdom

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Cases were selected from retrospective and prospective studies from populations which reflect a range of renal cell carcinoma (RCC) incidence rates. Numbers were limited by the number of cases available

Data exclusions

Cases were excluded for any of the following pre-established criteria; 1) Incomplete data on core set variables (age at diagnosis, sex, alcohol and tobacco consumption) 2) Failure to pass pathology review as described in the Methods 3) If matched tumour/normal tissue did not originate from the same individual as determined by Fluidigm SNP genotyping. 4) If sequencing coverage was below 30X for tumour, or 15X for matched normal tissue 5) Evenness of coverage criteria 6) if contamination level was above 3% as determined by Conpair. For evenness of coverage, the median over mean coverage (MoM) score was calculated. Tumors with MoM scores outside the range of values determined by previous studies to be appropriate for whole genome sequencing (0.92 – 1.09) were excluded.

Replication

Signature extraction was replicated two times independently at both Wellcome Sanger Institute and UCSD, with similar results. Signature attribution was replicated two times independently at both Wellcome Sanger Institute and IARC, with similar results. All attempts at replication were successful. No other experiments other than those mentioned here were replicated independently due to limited resources.

Randomization	Randomization is not relevant for this study. Cases did not undergo interventions. All cases were collected based on diagnosis of primary renal cell carcinoma (RCC) and no prior treatment.
Blinding	Blinding is not relevant for this study. Cases were not subject to any interventions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging