nature COMMUNICATIONS

ARTICLE

Received 30 May 2014 | Accepted 3 Sep 2014 | Published 29 Oct 2014

DOI: 10.1038/ncomms6135

Variation in genomic landscape of clear cell renal cell carcinoma across Europe

Ghislaine Scelo[1,*], Yasser Riazalhosseini[2,3,*], Liliana Greger[4,*], Louis Letourneau[3,*], Mar Gonzàlez-Porta[4,*], Magdalena B. Wozniak[1], Mathieu Bourgey[3], Patricia Harnden[5], Lars Egevad[6], Sharon M. Jackson[5], Mehran Karimzadeh[2,3], Madeleine Arseneault[2,3], Pierre Lepage[3], Alexandre How-Kit[7], Antoine Daunay[7], Victor Renault[7], Hélène Blanché[7], Emmanuel Tubacher[7], Jeremy Sehmoun[7], Juris Viksna[8], Edgars Celms[8], Martins Opmanis[8], Andris Zarins[8], Naveen S. Vasudev[5], Morag Seywright[9], Behnoush Abedi-Ardekani[1], Christine Carreira[1], Peter J. Selby[5], Jon J. Cartledge[10], Graham Byrnes[1], Jiri Zavadil[1], Jing Su[4], Ivana Holcatova[11], Antonin Brisuda[12], David Zaridze[13], Anush Moukeria[13], Lenka Foretova[14], Marie Navratilova[14], Dana Mates[15], Viorel Jinga[16], Artem Artemov[17], Artem Nedoluzhko[18], Alexander Mazur[17], Sergey Rastorguev[18], Eugenia Boulygina[18], Simon Heath[19], Marta Gut[19], Marie-Therese Bihoreau[20], Doris Lechner[20], Mario Foglio[20], Ivo G. Gut[19], Konstantin Skryabin[17,18], Egor Prokhortchouk[17,18], Anne Cambon-Thomsen[21], Johan Rung[4], Guillaume Bourque[2,3], Paul Brennan[1], Jörg Tost[20], Rosamonde E. Banks[5], Alvis Brazma[4] & G. Mark Lathrop[2,3,7,20,†]

The incidence of renal cell carcinoma (RCC) is increasing worldwide, and its prevalence is particularly high in some parts of Central Europe. Here we undertake whole-genome and transcriptome sequencing of clear cell RCC (ccRCC), the most common form of the disease, in patients from four different European countries with contrasting disease incidence to explore the underlying genomic architecture of RCC. Our findings support previous reports on frequent aberrations in the epigenetic machinery and PI3K/mTOR signalling, and uncover novel pathways and genes affected by recurrent mutations and abnormal transcriptome patterns including focal adhesion, components of extracellular matrix (ECM) and genes encoding FAT cadherins. Furthermore, a large majority of patients from Romania have an unexpected high frequency of A:T>T:A transversions, consistent with exposure to aristolochic acid (AA). These results show that the processes underlying ccRCC tumorigenesis may vary in different populations and suggest that AA may be an important ccRCC carcinogen in Romania, a finding with major public health implications.

[1] International Agency for Research on Cancer (IARC), 150 cours Albert Thomas, 69008 Lyon, France. [2] Department of Human Genetics, McGill University, 1205 Dr Penfield Avenue, Montreal, Quebec, Canada H3A 1B1. [3] McGill University and Genome Quebec Innovation Centre, 740 Doctor Penfield Avenue, Montreal, Quebec, Canada H3A 0G1. [4] European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK. [5] Leeds Institute of Cancer and Pathology, University of Leeds, Cancer Research Building, St James's University Hospital, Leeds LS9 7TF, UK. [6] Department of Pathology, Karolinska Institutet, SE-171 77 Stockholm, Sweden. [7] Fondation Jean Dausset - Centre d'Etude du Polymorphisme Humain, 27 rue Juliette Dodu, 75010 Paris, France. [8] Institute of Mathematics and Computer Science, University of Latvia, 29 Rainis Boulevard, Riga LV-1459, Latvia. [9] Department of Pathology, The Beatson Institute for Cancer Research, Switchback Road, Bearsden, Glasgow G61 1BD, UK. [10] Leeds Teaching Hospitals NHS Trust, Pyrah Department of Urology, Lincoln Wing, St James's University Hospital, Leeds LS9 7TF, UK. [11] First Faculty of Medicine, Institute of Hygiene and Epidemiology, Charles University in Prague, Studničkova 7, Praha 2, 128 00 Prague, Czech Republic. [12] University Hospital Motol, V Úvalu 84, 150 06 Prague, Czech Republic. [13] Russian N.N. Blokhin Cancer Research Centre, Kashirskoye shosse 24, Moscow 115478, Russian Federation. [14] Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute and MF MU, Zluty Kopec 7, 656 53 Brno, Czech Republic. [15] National Institute of Public Health, Dr Leonte Anastasievici 1-3, sector 5, Bucuresti 050463, Romania. [16] Carol Davila University of Medicine and Pharmacy, Th. Burghele Hospital, 20 Panduri Street, 050659 Bucharest, Romania. [17] Centre 'Bioengineering', The Russian Academy of Sciences, Moscow 117312, Russian Federation. [18] National Research Centre 'Kurchatov Institute', 1 Akademika Kurchatova pl., Moscow 123182, Russia. [19] Centro Nacional de Análisis Genómico, Baldiri Reixac, 4, Barcleona Science Park - Tower I, 08028 Barcelona, Spain. [20] Centre National de Génotypage, CEA - Institute de Génomique, 2 rue Gaston Crémieux, 91000 Evry, France. [21] Faculty of Medicine, Institut National de la Santé et de la Recherche Medicale (INSERM) and University Toulouse III-Paul Sabatier, UMR 1027, 37 allées Jules Guesde, 31000 Toulouse, France. * These authors contributed equally to the study. † Present address: McGill University and Genome Quebec Innovation Centre, 740 Doctor Penfield Avenue, Montreal, Quebec, Canada H3A 0G1. Correspondence and requests for materials should be addressed to G.M.L. (email: tania.abouyounes@mail.mcgill.ca).

Renal cancer is diagnosed in more than 330,000 people each year worldwide, and accounts for 2.4% of all adult cancers and over 140,000 deaths annually[1]. Incidence rates have been increasing sharply with unexplained variation in different countries. The highest rates that are observed worldwide occur in Central Europe[2], and in particular in the Czech Republic, reaching 24.1/100,000 in men and 10.5/100,000 in women (world population age-standardized rates), while the equivalent figures for the United Kingdom are 10.9/100,000 and 5.8/100,000. Approximately 90% of renal cancers are renal cell carcinomas (RCCs) that develop in the renal parenchyma[2], with conventional (clear cell) RCC (ccRCC) being the most common (70–80%) histological type. Somatic mutations or epigenetic alterations of the von Hippel–Lindau tumour suppressor gene (*VHL*) are observed in >80% of ccRCC[3,4]. A modest proportion (2–4%) of RCC is associated with VHL syndrome caused by germline mutations in *VHL*[5]. Results from genome-wide association studies have identified common germline variants associated with increased susceptibility for developing ccRCC[6,7], and recent sequencing efforts have revealed recurrent somatic mutations in a number of genes including *PBRM1*, *SETD2* and *BAP1* (ref. 8). Recognized environmental and lifestyle risk factors for RCC include tobacco smoking, excess body weight and hypertension, as well as a history of chronic kidney diseases[2,9].

Previously somatic mutation patterns in patients from different European populations have not been systematically examined. Here we present the results of a whole-genome sequencing (WGS) study of ccRCC patients from four European countries (Czech Republic, United Kingdom, Romania and Russia), and we identify a specific mutational pattern that is predominant in the patients from Romania but not elsewhere. In addition to the WGS analysis of somatic variations, we also present results from a genome-wide transcriptome analysis in a subset of the study samples.

Results

Molecular profiling of the clinical cohort. Table 1 and Supplementary Data 1 provide summary and individual-level data on patients. We undertook a comprehensive molecular evaluation of the samples using WGS, single-nucleotide polymorphism (SNP) arrays (Illumina Human660-Quad BeadChip) and transcriptomics (RNA-seq). Matched tumour and blood DNA samples were available on 94 of the study participants, and WGS was made to an average depth of 54× coverage in all of these samples (Supplementary Data 2). We observed >99% concordance between SNP genotypes and sequence-based single-nucleotide variation (SNV) at the same sites in all 86 pairs with both data, attesting to the accuracy of the sequence calls. RNA from tumours was available for 92 patients (63 with WGS data), and from matched normal adjacent tissue for 45 of these (36 with WGS). We obtained an average of 81 million reads per sample from the RNA-Seq data, of which 90% were retained for further analysis based on the mapping results, the vast majority (94.0%) of which localized to protein coding genes.

Genome sequencing results. We detected 4,904 somatic mutations on average per sample pair corresponding to a mean of 1.79 somatic mutations per Mb after correcting for regions with low coverage (see Methods). Intergenic regions were seen to have higher mutation rates (2.02 Mb^{-1}) than other genome regions (Supplementary Fig. 1). Notably, the overall somatic mutation rate in coding regions was significantly less than this (1.54 Mb^{-1}, rate ratio = 0.76, 95% confidence interval 0.74–0.79) corroborating previous literature[10]. Similarly, other regions associated with genes (for example, 5′ untranslated region (UTR), 3′UTR and introns) had lower rates than intergenic regions as shown in Supplementary Fig. 1. Regions corresponding to DNaseI hypersensitive (DHS) sites (see Methods for definition) also had a lower overall somatic mutation rate (1.66 Mb^{-1}) when compared with intergenic regions (rate ratio = 0.82, 95% confidence interval 0.81–0.83) in concordance with a recent report in other cancers[11]. We did not detect significant patterns of recurrent mutations in non-coding regions of the genome (see below for coding regions). We identified loss-of-heterozygozity and other copy number variants (CNVs) at 1,008 sites (with sizes ranging from 450 kb to 197.7 Mb and between 0 and 48 CNV

Table 1 | Characteristics of the patients in the study.

Category	Group	No. in category (%) or median (range)	
		All samples ($n=121$)	Sequenced samples ($n=94$)
Country of residence	Czech Republic	38 (31.4%)	28 (29.8%)
	Romania	14 (11.6%)	14 (14.9%)
	Russia	38 (31.4%)	23 (24.5%)
	UK	31 (25.6%)	29 (30.8%)
Sex	Female	53 (43.8%)	42 (44.7%)
	Male	68 (56.2%)	52 (55.3%)
Smoking status	Never	60 (49.6%)	41 (43.6%)
	Former	34 (28.1%)	31 (33.0%)
	Current	27 (22.3%)	22 (23.4%)
Body mass index		27.6 (20.7–49.5)	27.6 (20.9–49.5)
History of hypertension	Yes	54 (44.6%)	46 (48.9%)
	No	67 (55.4%)	48 (51.1%)
Age at surgery		60 (35–83)	60 (38–83)
Stage	I	67 (55.4%)	51 (54.3%)
	II	12 (9.9%)	9 (9.6%)
	III	29 (24.0%)	21 (22.3%)
	IV	13 (10.7%)	13 (13.8%)
Tumour grade	1	3 (2.5%)	2 (2.1%)
	2	74 (61.2%)	54 (57.4%)
	3	23 (19.0%)	18 (19.1%)
	4	21 (17.4%)	20 (21.3%)

differences per pair; Supplementary Data 3) and we found 1,418 additional putative structural rearrangements (Supplementary Data 4) with further analyses as described (Methods). Among recurrently affected loci, loss of chromosome 3p, the most frequent CNV, was observed in 90% of samples, whereas loss of all of chromosome X was observed in 7.4% of samples (Supplementary Fig. 2), which is a novel recurrent aberration that has not previously been reported in ccRCC.

A geographic specific mutational signature. We observed different somatic mutation frequencies associated with patient country-of-residence (Kruskal–Wallis $P = 0.003$; see Fig. 1). This was principally due to A:T > T:A transversions, which were found to be strikingly elevated in some tumours (Fig. 1a,b). We analysed the SNV patterns to identify such outlier samples as described in the Methods. Twelve sample pairs, all from Romania, accounted for the significantly increased A:T > T:A transversion rate (Fig. 1c and see also Supplementary Data 2 and Supplementary Fig. 3a). One additional Romanian sample pair (RO6) was a potential A:T > T:A outlier, while only one Romanian sample pair (RO12) exhibited no evidence of such a deviation. In the 12 outliers from Romania, the overall rates of somatic mutations were substantially higher than elsewhere (4.01 Mb^{-1} versus 1.48 Mb^{-1}) and A:T > T:A transversions comprised 21–74% of all the somatic mutations detected (Fig. 1a). After removal of A:T > T:A, the pattern of mutations in other SNV classes remained slightly higher in the outliers compared with other samples

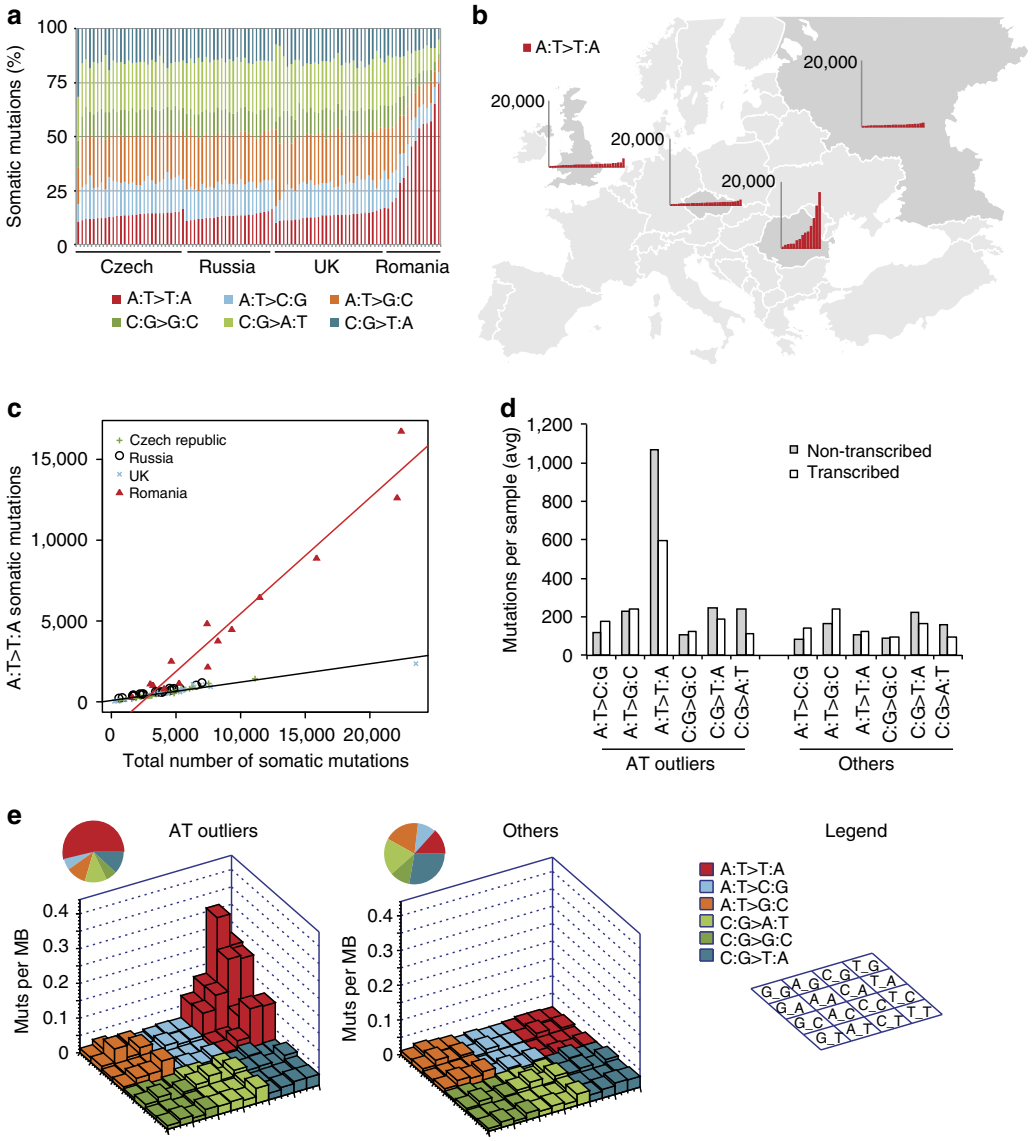

Figure 1 | Single-base substitution patterns in 94 ccRCC samples. (**a**) Proportion of observations within different base substitution classes for each ccRCC sample pair. (**b**) Number of A:T > T:A transversions in samples from the four countries included in the study. (**c**) The number of A:T > T:A transversions plotted against the total number of mutations in each sample. The graph also shows the linear regression lines for Romanian samples (red) and other samples (black). (**d**) Frequencies of six base substitution classes in Romanian outliers and other samples for somatic mutations within genes categorized into non-transcribed and transcribed strands. A:T > T:A transversions occur preferentially on the non-transcribed strand in the Romanian outliers. (**e**) Lego plot showing the number of somatic mutations with the surrounding sequence context for the Romanian outliers and other sample pairs. The plot illustrates the preference for A:T > T:A transversions within the context C/T[A:T]A/G, but also the overall increased frequency of A:T > T:A and increases of other mutational classes for the Romanian outliers.

(Supplementary Fig. 3b, Kruskal–Wallis $P=0.046$). Country-of-origin was not significantly correlated to mutation counts when the Romanian sample pairs were excluded from the analysis (Kruskal–Wallis, $P=0.10$).

In the Romanian outliers, we observed a preference for A:T>T:A to occur within the specific sequence context of C/T[A:T]A/G (Fig. 1e and Supplementary Fig. 3c,d). This pattern accounted for 50% of A:T>T:A transversions in outliers compared with 26% in other samples. However, A:T>T:A was more frequent in the outliers irrespective of the context (Supplementary Fig. 3d). Finally, we observed a strong bias for a location on the non-transcribed strand for A:T>T:A transversions within genes in the Romanian outliers (Fig. 1d). Exposure to aristolochic acid (AA) is known to produce an increased frequency of A>T mutations with such strand bias[12] and has previously been shown to lead to AA nephropathy (AAN) characterized by chronic renal disease and to rare transitional cell carcinomas of the upper urinary tract[13,14], but has not been reported as an environmental risk factor in ccRCC.

Other mutational patterns. One UK sample pair (UK24) with a very high somatic mutation rate ($7.61\,Mb^{-1}$) exhibited a relative predominance of A:T>G:C and C:G>A:T mutations (Supplementary Fig. 4). This was the only obvious outlier with a potential unique mutational pattern outside of the Romanian sample pairs. We eliminated UK24 and all the Romanian sample pairs to conduct further analyses of the somatic mutation patterns. We found that the total number of somatic mutations increased significantly with the patients' age (Pearson $r^2=0.3$; $P=2\times10^{-5}$; Supplementary Fig. 5a). This observation included C>T mutations in the context of CpG dinucleotides, as reported previously in ccRCC and other cancers[15,16], but it held equally for all other SNV classes (Supplementary Fig. 5b).

Somatic mutations in genes. We identified genes that harboured non-silent somatic mutations with a frequency significantly greater than background rates ($P<0.05$ from MuSiC[17]), or with non-silent somatic mutations seen in at least two patients. These mutations were further filtered by *in silico* analyses and subject to verification by an orthogonal sequencing method resulting in 1,239 validated mutations out of 1,317 that could be tested (94.1%; see Methods for details). Among tested mutations, 200 were indels from which 187 (93.5%) were validated and 1,117 were nucleotide substitutions from which 1052 (94.2%) were validated. The filtering and validation procedures led to a list of 583 genes (Supplementary Data 5). Non-silent A:T>T:A substitutions were more frequent in the Romanian outliers compared with other patients (38% versus 11%; $P<10^{-33}$); however, only 9 of the 583 genes were in the list uniquely because of A:T>T:A substitutions in these patients (see Supplementary Data 6 for further information on non-silent A:T>T:A mutations). After removal of the Romanian sample pairs, the number of genes harbouring non-silent mutations did not significantly vary with country-of-residence (Kruskal–Wallis, $P=0.08$). We found a significant association between number of non-silent mutations and the patients' age (Pearson $r^2=0.46$; $P=5\times10^{-7}$ with Romanian pairs excluded; Supplementary Fig. 5c) that was stronger than that in the WGS data overall.

The most frequently mutated genes are shown in Table 2 and Fig. 2. The prevalence of *VHL* mutations in our patients (73%) is similar to that reported in our previous studies (70–80%)[3,4] but substantially higher than those found in recent next-generation sequencing studies of ccRCC, which have reported *VHL* mutations in 27% (ref. 18), 40% (ref. 19), 52% (ref. 20) and 55% (ref. 21) of samples. Although some of the variation may be due to ethnic origin, patient selection, pathology review criteria and/or sample tumour cell content, false-negative next-generation sequencing results have been evoked as a factor affecting previous studies[20]. *VHL* was mutated in 8 of the 12 Romanian outlier; however, none of these mutations were A:T>T:A transversion. In one patient, we identified a germline mutation (missense) in VHL of unknown functional significance.

Additional known ccRCC genes mutated in our cohort include *PBRM1* (39%), *SETD2* (19%), *BAP1* (12%) and *KDM5C* (7%). No A:T>T:A mutations were found in *SETD2*, *BAP1* or *KDM5C* in

Table 2 | Frequency of events for principal genes and pathways harbouring non-silent somatic variants.

Gene or pathway	Focal adhesion or PI3K	% with events (n = number of sample pairs evaluated)			mRNA expression in tumour compared with normal	*P*-value
		Non-silent mutations ($n=94$) (%)	CNVs ($n=85$) (%)	Switch events ($n=36$) (%)		
VHL		73.40	73.40	0.00		1.3×10^{-56}
Focal adhesion/PI3K	Yes	58.50	82.97	34.00	Up	2.7×10^{-4}
PBRM1		39.36	72.34	5.32		5.2×10^{-44}
SETD2		19.15	74.47	0.00		1.1×10^{-15}
BAP1		11.70	71.28	20.21		1.9×10^{-9}
ZFHX4		9.57	7.45	0.00		3.8×10^{-7}
CSMD3		8.51	5.32	0.00		8.1×10^{-5}
MTOR	Yes	8.51	6.38	9.57		1.7×10^{-5}
FAT3		7.45	4.26	0.00		7.1×10^{-4}
KDM5C		7.45	9.57	0.00		1.4×10^{-7}
ZNF469		7.45	5.32	0.00		2.6×10^{-3}
ANPEP		6.38	0.00	0.00	Down	6.1×10^{-9}
COL11A1	Yes	6.38	2.13	0.00		1.3×10^{-5}
MLL3		6.38	8.51	12.77		1.2×10^{-3}
NRXN1		6.38	1.06	0.00		6.1×10^{-5}
PIEZO2		6.38	5.32	0.00	Up	1.6×10^{-2}
TRRAP		6.38	9.57	0.00		2.2×10^{-4}
WDFY3		6.38	3.19	0.00		3.3×10^{-4}

CNV, copy number variant; PI3K, phosphatidylinositide 3-kinase.
Genes are listed when non-silent mutations were detected in >6% of the samples. Two genes, *TTN* and *MUC16*, were excluded from this table because the mutation patterns are generally observed in genome/exome sequencing experiments. *P*-values are calculated using the convolution test from Genome MuSiC[17] except for focal adhesion/PI3K as described in methods.

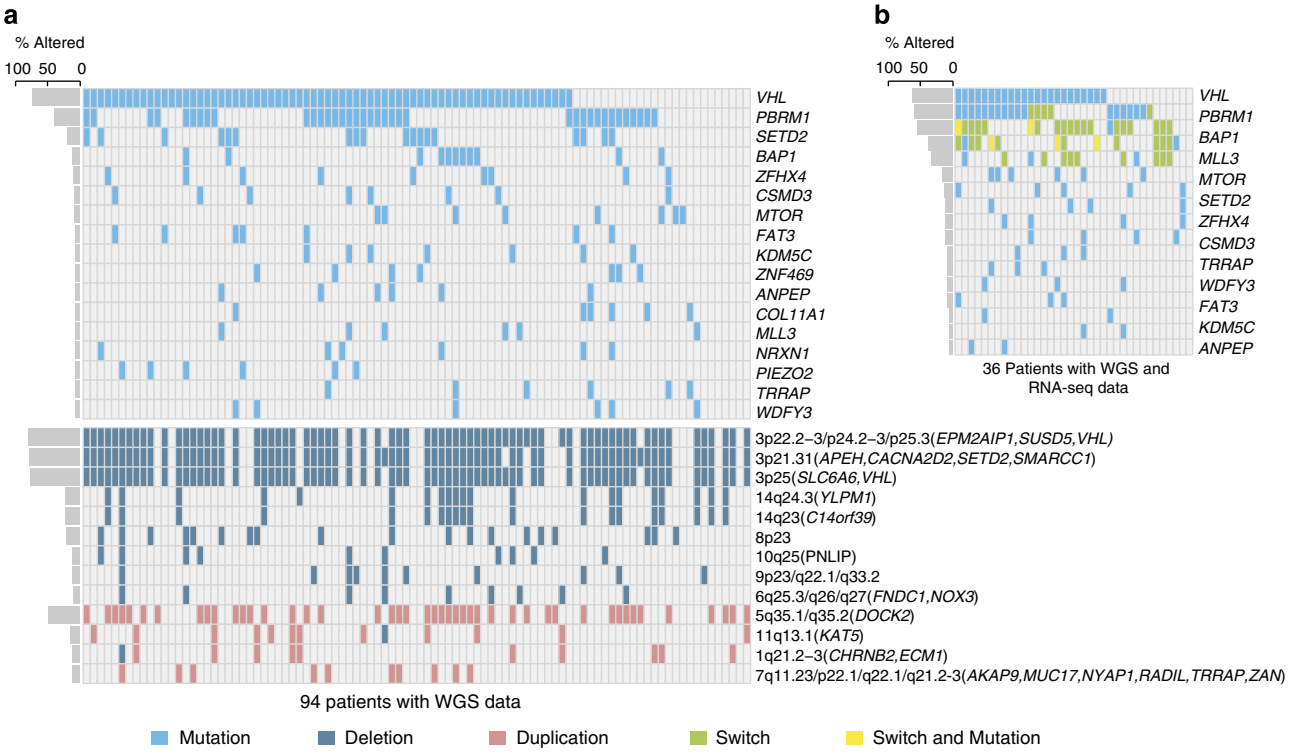

Figure 2 | Patterns of somatic variations in ccRCC samples. Left histograms show percentages of affected samples. (**a**) Upper heat map, somatic mutations; bottom heat map, copy number losses (blue) and gains (red) in 94 samples with sequencing data. Affected genes in each locus that harbour somatic SNVs are listed in the parentheses. (**b**) Patient-specific mutation and switch events affecting most frequently mutated genes (shown in **a**) detected in the 36 samples paired samples with both genomic and RNA analysis.

Romanian outliers (see below for *PBRM1*). In addition to *PBRM1*, genes encoding members of the SWI/SNF complex were mutated in 55% of ccRCCs studied here. Taking histone modifier genes into account, we found that 80% of the series was affected by mutations in epigenetic regulator pathways (Supplementary Fig. 6). Novel genes with recurrent mutations in our data include *FAT3* (7%), *WDFY3* (6%) and *ANPEP* (6%) as shown in Table 2. In addition to *FAT3*, somatic mutations were also identified in other genes encoding Fat cadherins including *FAT1*, *FAT2* and *FAT4* accounting for 20% of the subjects included in this study. We performed additional analysis across all genes but did not find any recurrent mutation in the promoters, UTRs or introns, nor in other regions (for example, DHS regions).

Interestingly, *PBRM1* was mutated in 10/12 (83%) of Romanian outliers compared with 27/82 (33%) in other sample pairs ($P=0.001$), but only three of the mutations in the outlier group have the characteristic transversion pattern discussed above. This observation led us to further examine the overall relationship of somatic mutation frequencies with *PBRM1* mutations. Because of the predominance of A:T > T:A mutations in *PBRM1* mutated patients in Romania, we initially removed these transversions from consideration. We found that the presence of non-silent *PBRM1* mutations was associated with higher overall mutation rates in the other mutation classes (Kruskal–Wallis, $P=0.002$). When we excluded the Romanian and UK outlier samples, but this time included A:T > T:A in the analysis, we again found a significant association (Kruskal–Wallis, $P=0.009$). However, we saw similar but nonsignificant trends for the other frequently mutated genes, where there may be lower power to detect an association because of the relative numbers of mutated and non-mutated sample pairs; thus, the *PBRM1* observation may simply be due to association of higher somatic mutation rates genome wide with increased mutation frequencies in all the principal driver genes.

Analysis of the relationship between clinical variables and the presence or absence of somatic mutations in genes with a mutation frequency of >10% identified significant associations only of both *PBRM1* mutations ($P=0.043$) and *SETD2* mutations ($P=0.014$) with higher-stage tumours (Supplementary Fig. 7). As tumour stage is defined by local invasiveness (T), lymphatic infiltration and growth in lymph node (N) and metastasis (M), we also examined these variables. *PBRM1* mutations only were associated with T ($P=0.05$) but other variables and genes showed no significant relationships.

Differential gene expression and differential splicing. We found that 12,849 protein-coding genes (60% of genes annotated as such in Ensembl 66) were expressed on average at 1 fragments per kilo bases of exons for per million mapped reads (FPKM) or more in either the panel of 91 tumour samples, or in the 45 normal samples. In a paired analysis using only 45 samples with RNA-Seq data from matched tumour and adjacent normal tissue, we detected 3,272 protein-coding genes that were differentially expressed with more than twofold change between tumour and normal (false discovery rate (FDR) < 0.01; Supplementary Data 7 and 8; see also Methods). Hierarchical clustering did not reveal any subgroups with correlations with clinical variables (results not shown). Expression data in paired samples were available for only two of the Romanian patients and in tumour only for three additional patients, which precluded an analysis of the subset with frequent A:T > T:A mutations.

Individual genes were then examined for statistically significant differences in the levels of exon usage in the 45 matched tumour

and adjacent normal samples, a metric that can be used as an indicator of differential splicing (see Methods). We detected 7,842 genes (7,182 protein coding) that manifested significant differences for at least one exon (with FDR < 0.01; Supplementary Data 9). On the other hand, from transcript-centric analysis, we found that the average number of expressed transcripts per gene was higher in tumours compared with normal samples (Supplementary Fig. 8). Transcript expression levels were also more variable in tumours compared with normal samples (Supplementary Fig. 8), and such variability was correlated with the number of annotated transcripts (Pearson $r^2 = 0.75$; $P < 0.001$).

Consistent with previous observations[22], we observed gene expression to be dominated by one transcript in most cases, both in tumours and normal samples (Supplementary Fig. 9). However, the number of genes with a dominant transcript was significantly different between tumour and normal samples ($P < 0.001$; Methods). When focusing on the most abundant transcript within each gene (major transcript), 50% of the genes with altered splicing were predicted to change major transcripts between the normal and tumour in at least one sample pair (referred to as switch events; Fig. 3). However, when applying additional filtering to require at least twofold or fivefold difference between the first and second most abundant transcripts, the number of such switch events was reduced to 25% and 2% of the genes, respectively, most of which were observed in a small number of samples (Fig. 3b). Figure 3c,d provide examples of switch events in *PP2R4* and *SRSF6*, respectively, and show that a possible loss of function can occur through changes in the dominant transcript isoform even when no significant differential expression is observed on a gene level (see Supplementary Notes for details). Taken together, these findings indicate that extreme splicing changes are mostly non-recurrent, and suggest that splicing variation in ccRCC principally affects lower abundance transcripts.

Modified expression for somatically mutated genes. Of the 583 genes with non-silent somatic mutations, 104 (18%) also exhibited differential expression (Supplementary Data 7 and 8) and 91 (15%) were observed to have switch events (using a twofold difference criterion and required observation in at least two patients; Supplementary Data 10). We evaluated the possible effect that mutation status might have on transcript expression levels by considering differences in transcript abundance for the 32 most frequently mutated genes. The results are shown in Supplementary Data 11. We categorized tumours according to mutation status (nonsense/frame-shift, missense and no mutation) and also compared with transcript abundance in normal tissue. *SETD2* exhibited nominally significant differences between mutation classes with tumours in the class of nonsense/frame-shift mutations had lower average expression (Supplementary Fig. 10). We also found marginally significant results for *ZFHX4*. However, neither result remained significant after adjustment for multiple testing.

Interestingly, *VHL* exhibited no switch events, while *PBRM1*, the second most frequently mutated gene, was affected by switch events in 5 of the 36 patients tested, none of whom harboured

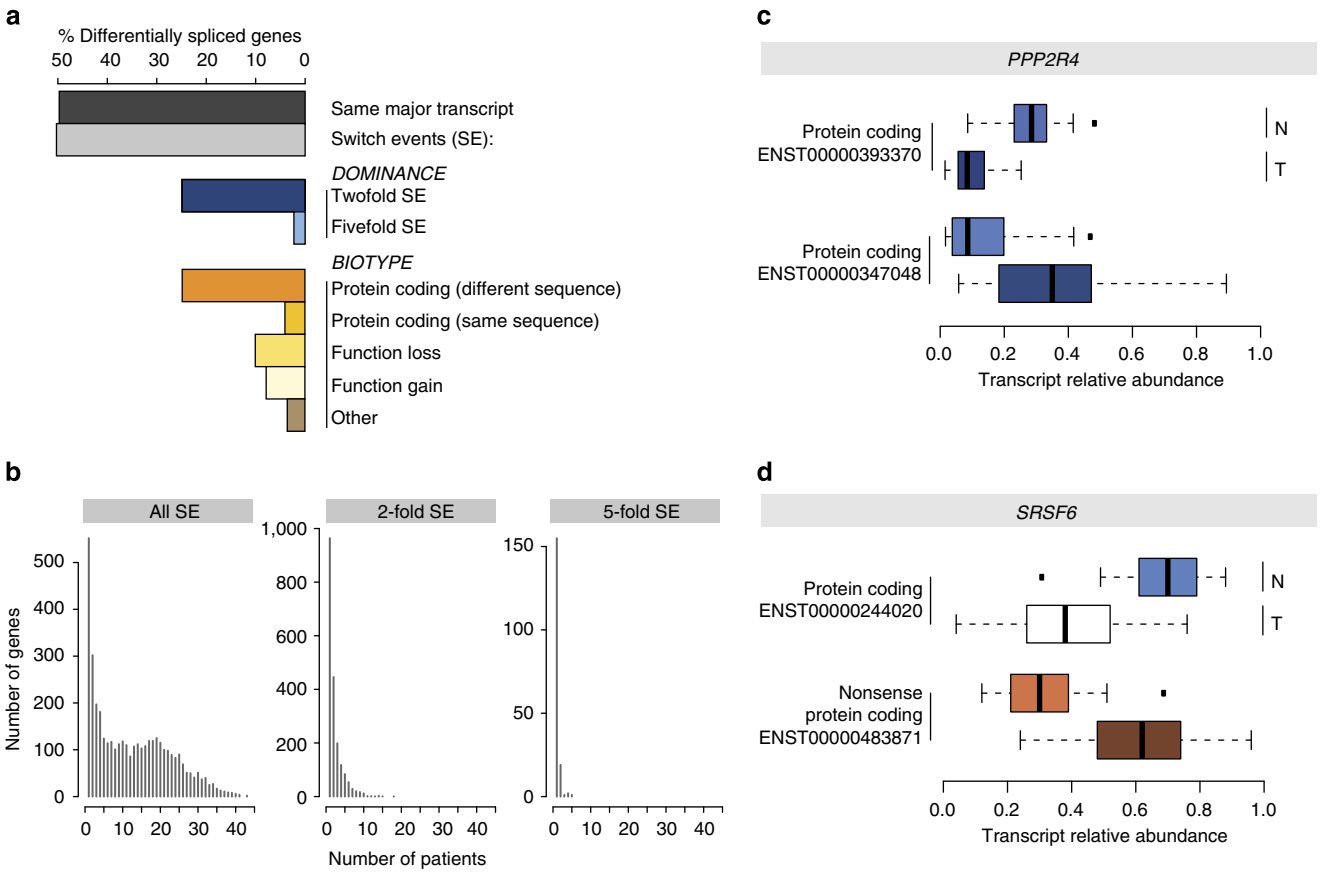

Figure 3 | Switch events in 45 matched normal and tumour samples. (**a**) Frequencies of different categories of switch events. (**b**) Number of patient sample pairs affected by switch events in different categories. (**c**,**d**) Box plots showing median, upper and lower quartiles of transcript relative abundance for *PP2R4* and *SRSF6*.

somatic mutations in this gene. Other mutated genes recurrently affected by switch events included *BAP1*, *PTEN* and *MTOR* affected by switch events in 20.2%, 6.4% and 9.6% of the studied samples, respectively (Fig. 2b and Supplementary Fig. 11). *MLL3*, a histone methyltransferase involved in transcriptional activation, also emerged as an interesting example, as it was mutated in 6% of tumours in our data, and was affected by switch events in an additional 9.5%. Similarly, the transcription factor *ETS1* was mutated in only one patient but was widely affected by switch events and expression changes (Supplementary Fig. 12). This is consistent with previous reports that point to a disruption of expression levels for this gene in ccRCC[23]. Switch events affecting other recurrently mutated genes and genes involved in cellular processes associated with renal cancer are depicted in Supplementary Figs 11 and 12, and discussed in the Supplementary Notes. The occurrence of switch events and somatic mutations in the same gene and patient were not mutually exclusive (Supplementary Data 12).

Identification of fusion genes. Following *in silico* screening (see Methods) and validation by reverse transcription–PCR, we identified six tumour-specific fusion events (Table 3, Fig. 4 and Supplementary Fig. 13). Two fusion partner genes (*MED15* and *TFE3*) shown in Fig. 4 participate in the TGFβ/SMAD pathway, known to play a role in renal cancer development[24], while 9 out of the 12 fusion partners (*CWC25*, *CGNL1*, *SH2D3C*, *RAB31*, *LRSAM1*, *MED15*, *SLC12A4*, *TFE3* and *TCF12*) code for phosphoproteins, which are known with important roles in signalling pathways[25]. All the confirmed chimeras appeared to be associated with inversion, as the fusion partner genes were located on opposite strands. Further details about fusion events are provided in the Supplementary Notes.

Pathway analysis. Using the Kyoto Encyclopedia of Genes and Genomes (KEGG) data sets, we performed pathway analysis for genes affected by somatic mutations or transcriptome alterations (see Methods). Pathway analysis of the 583 genes harbouring somatic mutations showed significant enrichment for focal adhesion and phosphatidylinositide 3-kinase (PI3K) pathways (FDR $= 5 \times 10^{-07}$ and FDR $= 0.003$, respectively; Supplementary Data 13). Overall, 59% of tumours in our study showed non-silent somatic mutations in one or more of 32 genes from these two pathways (Table 2). Recurrent mutations in constituents of PI3K–mammalian target of rapamycin (mTOR) signalling have recently been reported in ccRCC[19,20]. We observed non-silent somatic mutations or CNVs at each of *PIK3R1*, *PTEN* and *MTOR* in $>5\%$ of tumours and somatic variations at lower frequencies in other genes involved in PI3K–mTOR signalling. The Focal adhesion pathway contains genes that act upstream of PI3K (and of the FAK and Src pathways). We observed frequent somatic variations in genes encoding extracellular matrix (ECM) molecules, integrin receptors, members of the collagen family and genes coding for laminin chains in addition to genes coding for receptor tyrosine kinases (Supplementary Data 14 and Fig. 5).

Among downregulated genes, we observed highly significant (FDR $< 10^{-9}$) enrichment in pathways involved in energy metabolism such as oxidative phosphorylation, known to be frequently impaired in ccRCC cells as part of a general metabolic shift[26], carbon metabolism, the citrate cycle and amino acid metabolism (Supplementary Data 15). Among key pathways enriched for upregulated genes were cytokine–cytokine receptor interaction, cell adhesion molecules and the chemokine signalling. Focal adhesion and PI3K pathways were also enriched for upregulated genes (FDR $= 2 \times 10^{-5}$ and FDR $= 2 \times 10^{-8}$, respectively; Supplementary Data 16). 'Metabolic pathways' was the only KEGG pathway that showed evidence of enrichment for genes involved in switch events (using the criteria of a twofold expression difference seen in at least two patients). However, the protein processing in endoplasmic reticulum and mTOR signalling pathways also showed some evidence of enrichment ($P = 0.001$ without multiple testing correction, FDR $= 0.14$; Supplementary Data 17). The latter is relevant in the context of PI3K-Akt and focal adhesion.

Collectively, these results show that focal adhesion–PI3K–mTOR molecular axis is recurrently affected by somatic mutations and/or abnormal gene expression patterns in ccRCC (Supplementary Data 14 and Fig. 5).

Discussion

A high rate of A:T > T:A transversions, as seen in the large majority of our Romanian samples, has not previously been reported in ccRCC. The predominance of A > T transversion on the non-transcribed strand of DNA along with a preference for deoxyadenosine in the C/T[A:T]A/G motif supports exposure to AA as the underlying factor. Similar patterns have recently been observed in urothelial carcinoma of the upper urinary tract in patients[12,27] as well as in cultured primary cells[27,28] exposed to AA, with C/T[A:T]G being the preferential sequence context. Our WGS data show that a less marked increase in A > T transversions also occurs in other sequence contexts. According to GLOBOCAN data, Romania is in the lower range of annual kidney cancer incidence rates, with 8.19 new cases per 100,000 in men and 3.71 per 100,000 in women (world population age-standardized rates). However, GLOBOCAN estimates for Romania are largely based on the regional registry that is located in the northwest part of the country (far from the area where cases were recruited for this study) as well as registries of neighbouring countries (Bulgaria and Slovakia)[29]. In 2007, The Romanian Ministry of Health initiated additional regional cancer registries but these are not yet fully functional and several more years will be needed to accurately map kidney cancer incidence across Romania to evaluate whether there are regional disparities that could be due to lifestyle habits.

Table 3 | Tumour-specific fusion events.

Fusion	Breakpoint gene 1	Interpro domains in gene 1	Breakpoint gene 2	Interpro domains in gene 2	ORF
CWC25-PLXDC1	36962550	NA	37265644 and 37283259	NA	Out-of-frame
LRSAM1-SH2D3C	130214363	NA	130505243	NA	Out-of-frame
MED15-TFE3	20922918	IPR019087 retained	48891766	IPR011598, IPR021802 retained	In-frame
RAB31-PIEZO2	9792232	NA	10871413	NA	Out-of-frame
SLC12A4-DPEP2	67995478	IPR018491, IPR004841 lost	68025087 and 68024900	IPR008257 partially lost	Out-of-frame/ in-frame
CGNL1-TCF12	57754090	IPR002928 lost	57484356	IPR011598 retained	In-frame

NA, not available; ORF, open reading frame.

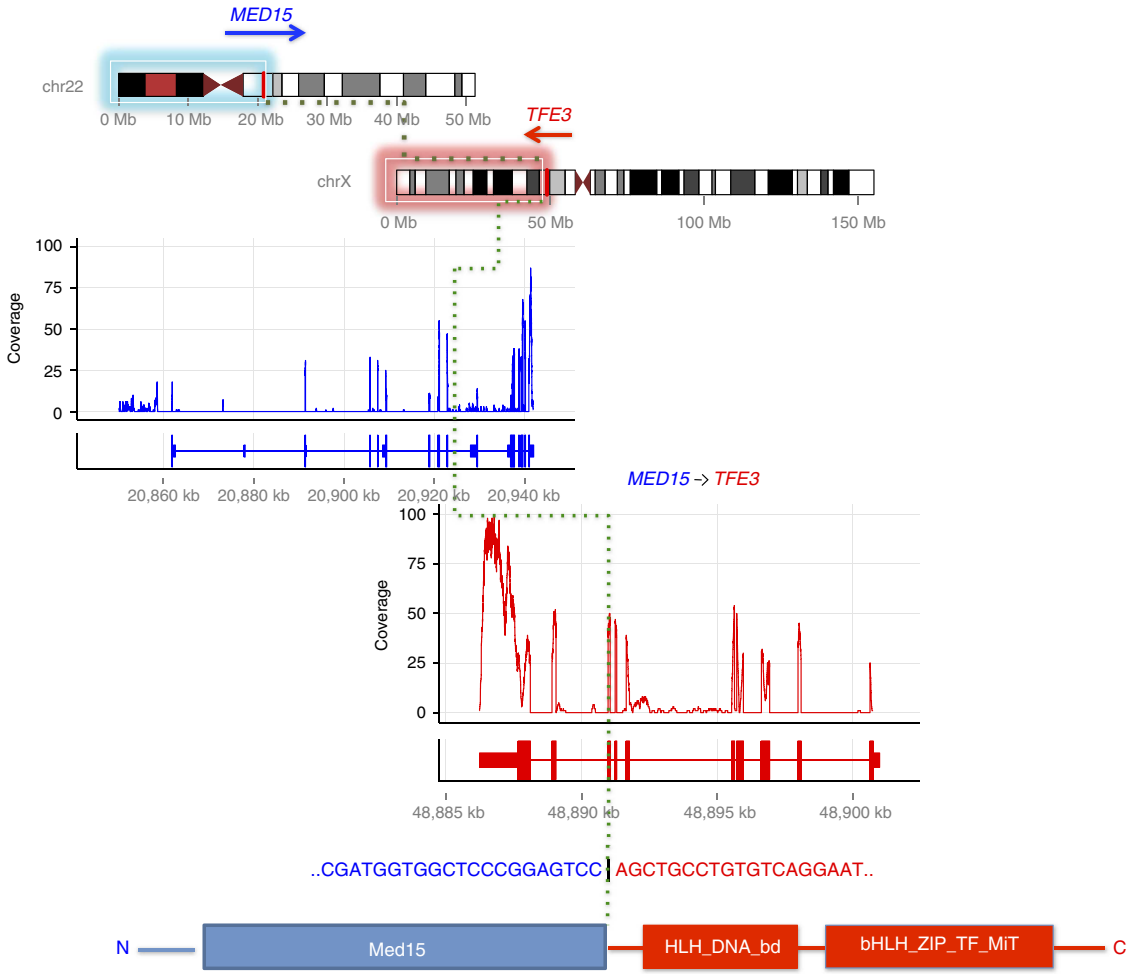

Figure 4 | Schematic representation of the *MED15->TFE3* fusion event. The figure shows the chromosome ideograms, transcript structures and number of reads mapping to the two partner genes. The dashed green line indicates the fusion breakpoint. The retained protein domains are shown at the bottom of the figure.

In parts of Asia, AAN occurs through widespread use of AA-containing herbal remedies[30]. In Europe, this practice has not been commonly reported and the use of *Aristolochia fangchi* in slimming regimen—which first triggered the attention to the associated risk of urothelial cancer in Belgium women[14]—appeared to be unintentional. Balkan endemic nephropathy (BEN) is thought to be due to the ingestion of wheat flour contaminated with seeds of *Aristolochia clematitis*[31]. BEN has been described as affecting people of the alluvial plains along the tributaries of the Danube River in Croatia, Bosnia and Herzegovina, Serbia, Bulgaria and Romania[31]. Interestingly, the catchment area of the Bucharest hospital where cases were recruited does not cover the Romanian population of the BEN area, who are usually hospitalized in Timisoara or Craiova, in the western and southwestern part of Romania. Our results provide strong motivation for further studies to investigate the potential routes of exposure to AA in Romanian ccRCC patients.

Although we did not find evidence for other mutational patterns that differentiate patients from higher and lower regions of incidence across our total cohort, we established a strong correlation between the number of somatic mutations in ccRCC and the age of patients at surgery. We observed that all SNV classes represent this age-associated pattern, suggesting that a general underlying process is involved. Although our observations may be due to the increased number of somatic mutations by age observed in kidney epithelial cells[32], involvement of a cancer-associated deficiency in related cellular processes such as DNA repair cannot be excluded. Tomasetti *et al.*[33] have detected similar patterns in other cancers, and they argue that most somatic mutations in self-renewing tissue occur in normal cells before tumour initiation and accumulate with patient age, and most do not play a causal role in neoplasia. Further studies in which patient-matched non-tumour kidney tissue samples are analysed in addition to the tumours might contribute further understanding of these processes.

Unlike in urothelial carcinoma of the upper urinary tract, the AA signature observed in Romanian ccRCC patients from our series was not associated with an increased rate of *TP53* somatic mutations, a gene that is not frequently mutated in ccRCC. We observed a high frequency of somatic mutations of *PBRM1*, which is the second most common mutated gene in ccRCC, among the Romanian outliers as compared with other sample pairs. However, the majority of *PBRM1* mutations were not A:T > T:A transversions, implying that the higher rate of *PBRM1* mutations in Romanians are not due to the AA exposure. Silencing of *PBRM1* in renal cancer cell lines has shown that this protein regulates pathways involved in chromosomal instability and cell proliferation[34]. Furthermore, it has been shown that *PBRM1* is required for *TP53*-driven replicative senescence[35].

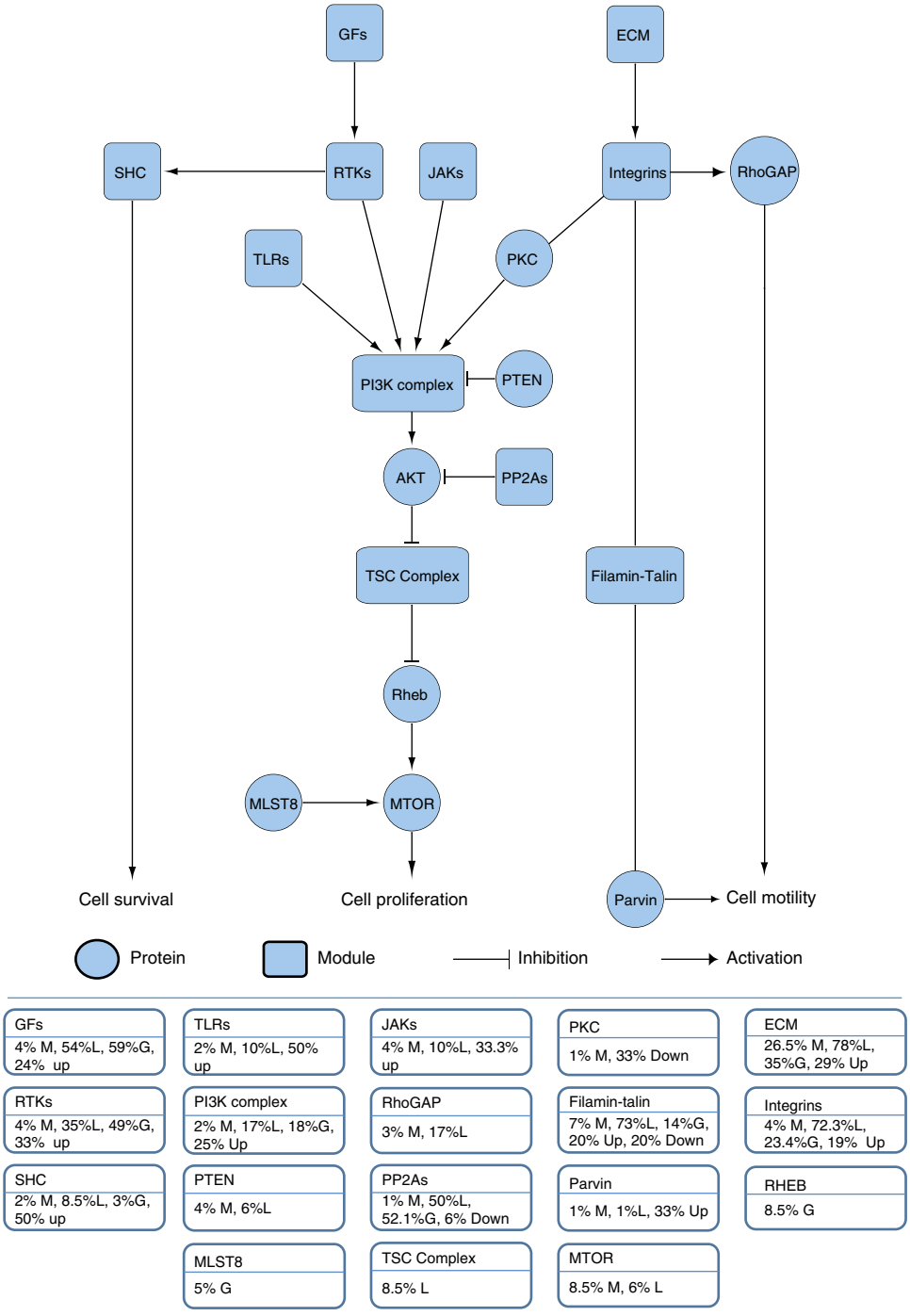

Figure 5 | Somatic variants in the focal adhesion and PI3K pathways identified in the ccRCC patients from our study. This figure shows interactions of the focal adhesion and PI3K pathway modules (adapted from KEGG). GFs, growth factors; ECM, extracellular matrix; RTKs, receptor tyrosine kinases; TLRs, Toll-like receptors; Jaks, janus kinases; PKC, protein kinase C; PP2As, Ser/Thr protein phosphotase 2A subunits; M, mutated; L, copy number loss; G, copy number gain; Up, upregulated; Down, downregulated; %, percentage of affected patients for M, L and G, and percentage of affected genes within each module for Up and Down from Supplementary Data 14.

Given that *PBRM1* is the only gene other than *VHL* whose mutations have been identified at the root of tumour evolution in a subgroup of ccRCCs[36], it will be relevant to investigate the extent to which *PBRM1* status is related to the AA-mutational signature or potentially other mutational patterns with additional samples in future studies. Likewise, the high rate of somatic mutations in epigenetic regulators supports the importance of chromatin remodelling/histone modification pathways in ccRCC as suggested recently[20].

In line with recent reports[19,20], the PI3K/AKT/mTOR pathway was recurrently affected by somatic mutations and abnormal gene expression patterns in our series, highlighting the relevance of this signalling cascade as a therapeutic target for ccRCC. In addition, we found many components of the focal adhesion pathway among the frequently mutated genes. Besides receptor tyrosine kinase proteins that are common to focal adhesion and PI3K pathways, other members of focal adhesion were also recurrently mutated, among which *COL5A3* and *ARHGAP35*

have recently been identified as novel ccRCC genes in a pan-cancer analysis of TCGA data[37]. *COL5A3* encodes α-chain of collagen type V involved in ECM. Genes coding for other constitutes of ECM including additional collagen proteins, integrins and laminins were frequently mutated in our series. Abnormalities of ECM dynamics are common features of tumour microenvironment and play key roles in tumour formation and progression[38], either by affecting downstream growth-promoting pathways such as ERK and PI3K signalling[39] or by contributing to angiogenesis and metastasis by influencing tumour microenvironment and tumour–stroma communication[40]. Thus, our data revealing frequent mutations in ECM components points to the likelihood of an important role of the tumour microenvironment in ccRCC. This is further supported by deregulation of cell adhesion and focal adhesion pathways observed in the gene expression profiles.

Fat cadherins constitute a novel gene family identified mutated in our cohort. Fat cadherins are surface proteins that are involved in cell adhesion and modulation of signalling pathways such as Hippo and Wnt signalling[41]. Emerging evidence has shown that *FAT* genes are mutated in different cancers and has suggested tumour suppressor activity for these genes[42,43]. Together with our data, this motivates further research to dissect the mechanisms by which FAT abnormalities can contribute to cancer.

Analysis of transcriptome patterns highlighted significant alterations in metabolic pathways consistent with the Warburg effect. This phenomenon is a hallmark of ccRCC[20] and emphasizes the important underlying metabolic shift in cancer cells. More generally, our study constitutes the first genome-wide characterization of the splicing alterations that are associated with renal cancer. Previous studies relied on the identification of differentially expressed exons to assess exon skipping events[44,45], an approach limited to the study of a subset of splicing events. By analysing differences in exon usage and major transcript expression patterns, we showed that although major recurrent changes in splicing are rare, splicing patterns are broadly altered in ccRCC, consistent with the observation that the messenger RNA processing pathway is commonly disrupted in ccRCC[46].

Our study presents a comprehensive genomic and transcriptome characterization of ccRCC in patients from several European countries. Genome-level mutational patterns and in-depth analysis of transcriptome abnormalities reported here advance our knowledge about the molecular background of the disease, with potential clinical applications. In particular, combination of aberrant genomic and transcriptome patterns underlines focal adhesion and PI3K/mTOR pathways as recurrent molecular targets and highlights the importance of ECM and tumour microenvironment in ccRCC.

Moreover, we observe strikingly different somatic mutation patterns that distinguish most of the Romanian patients from patients from other European countries included in our study. If due to AA exposure, these findings imply that the impact might be far broader than previously thought, both geographically, as none of the patients originated from known regions of endemic AAN, and for risk of developing a common cancer. Although we cannot ignore that other unknown substances could lead to similar mutational patterns, our results provide strong motivation for additional studies designed to investigate the epidemiology, biology and clinical features of ccRCCs with distinct mutational patterns in Romania, and potentially in other populations, which may lead to intervention strategies. This result also highlights the importance of including cases from multiple and diverse populations when conducting large-scale sequencing studies to determine the mutational profiles of a particular cancer.

Methods

Patients and samples. Patients undergoing nephrectomy for suspected renal cancer during the period December 2008 to March 2011 at St James's University Hospital in Leeds, UK; University Hospital Motol, Prague, Czech Republic; Masaryk Memorial Cancer Institute, Brno, Czech Republic; Th. Burghele Hospital, Bucharest, Romania; and N. N. Blokhin Cancer Research Centre, Moscow, Russia, were recruited to the study after informed consent was obtained. Recruitment in Central and Eastern Europe was coordinated by the International Agency for Research on Cancer. Ethical approvals were obtained from the Leeds (East) Local Research Ethics Committee, the International Agency for Research on Cancer Ethics Committee, as well as from local ethics committee for recruiting centres in Czech Republic, Romania and Russia. All sampling and clinical data collection was undertaken according to predefined standard operating procedures following guidelines from the International Cancer Genome Consortium.

The inclusion criteria for patients were ≥18 years of age, diagnosis of conventional ccRCC and no prior treatment. Exclusion criteria were a known family history of renal cancer or a defined genetic predisposition to kidney disease, such as von Hippel–Lindau disease or polycystic kidney disease, and those on haemodialysis. For patients entering the study, venous blood was obtained and the buffy coat stored for subsequent extraction of DNA. Samples of tumour and distant non-tumour cortical renal tissue were snap-frozen (with or without prior embedment in optical coherence tomography) or preserved in RNAlater as rapidly as possible following surgery (usually within 90 min). A representative haematoxylin and eosin (H&E)-stained section of formalin-fixed paraffin-embedded (FFPE) tissue was obtained for each case from the original diagnostic pathology departments and diagnostic histological details collected through standard abstract forms at each centre with grading according to the Fuhrman grading system[47] and staging using the TNM 2010 criteria[48].

Samples from 121 patients diagnosed with ccRCC were accepted into this phase of the study for analysis having passed pathological review as described below. These include samples from the Czech Republic (38 patients), Romania (14 patients), Russia (38 patients) and the UK (31 patients). WGS was undertaken for 94 cases and RNA sequencing for 91, with 36 cases analysed in both. For each case, original diagnostic H&E-stained FFPE sections were scanned using a Leica digital scanner and reviewed remotely using Slidepath software by at least two pathologists of the pathology review panel (P.H., L.E. and M.S.). For those where diagnosis of ccRCC was confirmed, frozen tissue samples (tumour and non-tumour) were processed as follows: two 5-μm sections placed on glass slides, thirty 20-μm sections placed in a tube for subsequent RNA extraction, forty 20-μm sections placed in a tube for DNA extraction and again two 5-μm sections placed on glass slides. The flanking sections at both ends of processed tumour samples were stained with H&E and also CD45, scanned and the digital images reviewed remotely by at least two pathologists of the pathology review panel (P.H., L.E. and M.S.) to confirm diagnosis and achieve a consensus assessment of grade, the presence or absence of necrosis and percentage of viable tumour cells. With one exception (due to a block selection error), all samples contained at least 70% viable tumour cells on average across the two flanking sections (Supplementary Data 1). In addition, a consensus of at least 70% viable tumour cells at both ends of the blocks was achieved for all except four cases. Frozen sections from non-tumour cortical renal tissue from each patient were reviewed to confirm the absence of tumour cells and assessed for viability and the degree of inflammation. Considering the cases submitted to the study for potential entry overall (that is, to produce the 121 cases fully accepted), ~10% were rejected at the level of FFPE review as not meeting the diagnostic criteria, with a further 25% of those cases entering at frozen tissue level being rejected, largely due to insufficient viable tumour cells, that is, approximately one-third of samples considered for study entry were unsuitable.

Preparation of DNA and RNA. DNA samples were extracted with an Autopure (Qiagen) instrument using an automated protocol based on the salting-out method. Briefly, kidney tissue sections were manually lysed by Proteinase K digestion at a concentration of 100 μg ml^{-1} in the Qiagen 'Autopure Cell Lysis solution' overnight at 55 °C. DNA was then purified on the Autopure extractor using the manufacturer's protocol 'Cell Lysate Increased Spin'. DNA quality was evaluated by visualization on agarose gel of the genomic DNA or of the PCR amplification reactions of two microsatellites (D19S879 and D7S2473). For RNA extraction, tissue sections were lysed using a TissueLyzer (Qiagen) in QIAzol Lysis Reagent. After chloroform separation, the RNA was purified on a QIAcube using the miRNeasy Mini Kit following the manufacturer's (Qiagen) instructions. The RNA purity and concentration was assessed using an Agilent 2100 Bioanalyzer and a Nanodrop spectrophotometer, respectively.

Whole-genome sequencing. Genomic DNA (1.5 μg) extracted from blood and tumour tissues were sheared to 300–600 bp using a Covaris E210 (Covaris, Woburn, Massachusetts, USA). Fragment libraries for a 100-bp paired-end sequencing were robotically prepared on a SPRI-TE Nucleic Acid Extractor workstation (Beckman Coulter, Inc., Fullerton, CA, USA) according to the Illumina protocol. Samples were sequenced on GAIIx and HiSeq2000 instruments (Illumina Inc., San Diego, CA, USA).

RNA sequencing. Indexed complementary DNA libraries were prepared from 1.5 μg of total RNA following the Illumina TRUSEQ protocol. Average size of the AMPure XP beads (Beckman Coulter, Inc.) purified PCR products was 278 ± 9 bp. The paired-end 100 bp reads sequencing of the transcriptome was performed on pools of four cDNA libraries on a HiSeq 2000.

SNP genotyping. Genotyping was performed on 86 of the study blood/tumour sample pairs for which sufficient tumour DNA was available with the Human 660W-Quad v1 array (Illumina, Inc.), according to the standard manufacturer's protocol using 200 ng genomic DNA input. Genotypes were retained on the basis of standard quality control criteria, including minor allele frequency >1%, genotype call rate >98% and lack of deviation from Hardy–Weinberg equilibrium.

Analysis of WGS data. We applied the FASTX-Toolkit for 3′ quality trimming using a minimum quality of Q30 and a minimum length of 32 bp. BWA backtrack was used to align each lane to GRCh37 reference genome with MT NC_012920 and the non-chromosomal supercontigs[49]. Picard (http://picard.sourceforge.net) was applied to adjust pair coordinates, flag duplicates and merge lanes, and GATK was used for realignment[50]. SNVs and short indels were called using SAMtools mpileup[51] with paired calling enabled and filtered with a minimum depth of 10×. Variants were annotated using SnpEff[52], which also includes dbNSFP and COSMIC annotations. To identify tumour-specific variants, we used a CLR (Phred log ratio of genotype likelihoods with and without pair constraint) threshold of 90 and a variant quality of 100. Putative variants were further inspected manually using Integrative Genomics Viewer visualization tool[53]. Cross-validation was made with SNP array data by extracting base calls overall and on a per lane basis for each site in the SNP array and comparing with the SNP genotypes. To identify significantly mutated genes, we used the convolution test *P*-value with FDR <0.2 from Genome MuSiC with region of interest built from all of RefSeq exons.

CNVs were detected with DNACRD, a programme that implements an extension of the DNAcopy[54] comparative genomic hybridization array algorithm to WGS data. We only retained calls that were also detected by the control-FREEC method[55]. BreakDancer[56] and Pindel[57] were used in parallel as orthogonal techniques to identify discordant reads from Illumina paired-end sequencing data. BreakDancer was run with the following parameters: (1) minimum mapping quality of 35, (2) read-pairs within ±3 s.d. of insert size for tumour and ±2 s.d. for normal were excluded. Pindel was run according to the following parameters: (1) minimum match around breakpoint of 10 bp, (2) minimum match to reference of 50 bp and (3) minimum read mismatch rate of 0.1. For both BreakDancer and Pindel, somatic events were selected for by requiring read call support of ≥10 reads in the tumour with 0 supporting reads in the normal. Hyper- and hypo-mappability regions and microsatellite regions were discarded, and the resulting outputs were annotated according to the RefSeq (August 2010), repeat masker (open-3.30) and the Genomic Database of Structural Variants (v10) using custom scripts based on the use of the BEDTools suite[58]. BreakDancer insertion events were removed from the list of predicted structural variants, as the small insert-size selection of the library generates numerous read pairs with overlapping mate. SNVs and indels found in <10% in tumour samples or observed in control samples were excluded from further analysis.

SNVs in genomic regions. We classified the genome into six categories, to examine the regional distribution of SNVs. The categories were exons (CDS), 3′ UTRs, 5′ UTRs, introns, and upstream and downstream flanking regions of genes, intergenic and regions corresponding to DHS sites. Regions were extracted in BED format from the snpEff 3.4 database of hg19 for the locations of CDS, UTR and intronic regions. The upstream and downstream regions were defined as 5 kb before 5′ transcription site and 5 kb after the 3′ site. These regions are partially overlapping. DHS clusters were taken from the UCSC 'Digital DNaseI Hypersensitivity Clusters in 74 cell types (2 reps) from ENCODE' track. Any part of the genome not covered by the above was classified as intergenic.

A genome position (base) was included in the regional analysis only if it had at least 10× sequence coverage in all samples (calculated by the SAMtools pileup routine). Somatic mutation rates were calculated for each category and sample pair by dividing each count by the number of bases included in the category. We also calculated the normalized somatic mutation rates for each sample pair and category (the frequency of somatic mutations within the category compared with the total number of somatic mutations meeting the coverage criteria for the sample). For each category, we compared the rate with that for all other regions of the genome over all samples retained for the analysis assuming Poisson distributions and using the 'rateratio.test' procedure implemented in R. The data were trimmed before analysis by removing four samples that had a normalized CDS rate that was >3 s.d. from the category median across all samples.

Analysis of mutational classes for outliers. A statistical procedure was used to identify samples with high somatic mutation rates for one or more of the following six mutational classes: A:T>C:G, A:T>G:C, A:T>T:A, C:G>A:T, C:G>G:C and C:G>T:A. For each patient, we calculated a standardized mutation rate for each class by dividing the number of mutations in the class by the total number of mutations observed for that patient. For each of the standardized rates, we calculated a one-sided *P*-value for obtaining an observation as or more extreme excess using a binominal distribution with parameter θ, where θ is the median of the rate for the same class. Supplementary Fig. 3a shows the base 10 logarithm for the *P*-values corresponding to A:T>T:A after adjustment for multiple tests.

Validation of somatic mutations. We validated experimentally somatic mutations in genes that were either affected by somatic non-synonymous mutations with a frequency significantly greater than background rates ($P<0.05$ from MuSiC), or mutated in at least two patients. Somatic SNVs and insertion/deletion variants were verified by pooled amplicon sequencing on Illumina MiSeq or by Sanger sequencing in the tumours and matched normal samples. PCR primers were designed to capture loci harbouring putative somatic mutations. Target loci were captured through multiplexed amplification using a Fluidigm access array system, and pools of amplicons were sequenced on MiSeq instruments. Alignments, variant extraction and quality metrics were made with BWA and in-house programmes. We considered a somatic mutation to be validated if the targeted re-sequencing confirmed the presence of the mutation in tumour and it was not found in the matched normal sample. Owing to the lack of corresponding DNA samples or unsuccessful PCR assays, we could not test further 374 out of 1,691 putative somatic non-synonymous mutations. Of the remaining 1,317 putative non-synonymous mutations, we confirmed 723 of 771 (93.7%) by Sanger sequencing and 516 of 546 (94.5%) by pooled amplicon sequencing. In total, 1,239 out of 1,317 (94.1%) of the non-synonymous mutations were validated.

Analysis of RNA-seq data. RNA-Seq raw reads were initially trimmed to 95 nucleotides using PRINSEQ v0.19.5 (ref. 59) and mapped to the reference genome using TopHat v2.0.6 (ref. 60). For each gene, expression counts were estimated using HTSeq v0.5.3p7 (http://www-huber.embl.de/users/anders/HTSeq) summarized across all its exons as annotated in Ensembl 66 using the –intersection-nonempty and –stranded = no parameters. Genes that had zero counts for ten or more samples were removed from further analysis. For the 45 matched tumour/normal samples, a paired test for differential expression was performed using edgeR[61] with tag-wise dispersion estimation and Trimmed Mean of M-values (TMM) normalization of the counts. Genes with FDR<0.01 were labelled as 'differentially expressed', and differentially expressed genes with median expression >1 FPKM were selected for further analysis. Gene FPKM levels were obtained by multiplying the read count for each gene with 1,000, and dividing with the total number of reads mapped to proper mate pairs to all genes and the gene length, estimated as the sum of the length of all its exons.

For transcript-level analyses, we applied DEXSeq[62] to identify those genes for which at least one of the exons is differentially used between normal and tumour in the 45 matched samples. Relative abundances for annotated transcripts were obtained using MISO[63]. Transcript FPKMs were then obtained by multiplying those relative abundances with the corresponding gene FPKMs. Based on those quantifications, we identified the most abundant transcript within each gene, referred to as 'major transcript'. The 'major transcript dominance' was calculated as the ratio of transcript FPKMs for the second and first most abundant transcripts for each gene. We then compared for each patient the number of genes with a different dominant transcript in the matched normal and tumour samples using a McNemar's test, and for each major transcript in each gene we compared its dominance values in all normal against all tumour samples with a Wilcoxon test. Controlling for multiple testing was done with the Benjamini–Hochberg FDR procedure. A patient was said to exhibit a 'switch event' for a gene when we observed a different major transcript in the matched tumour and normal sample. Such events were further classified as either twofold or fivefold dominant if the major transcript dominance from both the tumour and normal samples exceeded this threshold. Splicing variability was calculated with the methods provided in Gonzàlez-Porta *et al.*[64]

Identification of fusion genes. Fusion genes were identified by complementary approaches. The first, implemented in deFuse[65], identifies clusters of discordant paired-end alignments, which inform split read alignment, while the other, implemented in FusionMap[34], searches for chimeric transcripts by identifying split reads. Putative fusion events obtained from FusionMap with the number of supporting split reads >3, or obtained with deFuse with number of spanning reads >5 and split reads >3 were retained for further analysis. The list of potential chimeric genes obtained from these putative events was further filtered by removing instances when similar patterns were found in the normal samples from study, normal samples from 462 healthy individuals provided by the Geuvadis consortium[66] and 16 normal tissues samples (Illumina Body Map, ENA: ERP000546). Subsequently, to reduce the number of false-positive chimeras, we removed mitochondrial, ribosomal genes, pseudogenes, homologous fusion partner genes, fusion events overlapping with repetitive regions and naturally occurring read-through transcripts annotated at the AceView database[67]. Finally, we examined the remaining putative fusion events with the Integrative Genomics Viewer browser[53] and pairoscope (http://pairoscope.sourceforge.net/), and those that were supported by visualization were subjected to validation by reverse transcription–PCR.

For validated fusion events, the open reading frame was calculated by using the frame column for each exon in the ensembl GTF file as described in ref. 65. Protein domains were classified by 'retained', 'lost' or 'truncated' by extracting InterPro[68] domains of the partner genes from Ensembl via the Ensembl API (http://www.ensembl.org/) and further assessing the genomic domain locations relative to the breakpoints. Gene list enrichment analysis was performed with ToppGene Suite[69]. The ideograms, coverage plots and the gene models of the fusion partner genes were drawn with the R package ggbio[70].

Validation of fusion events. Fusion mRNA transcript validation experiments were performed for 16 putative fusion transcripts corresponding to 12 fusion genes identified from RNA-Seq data analysis in 17 tumour samples. For each fusion transcript, different PCR primer pairs were designed with Primer 3 (http://frodo.wi.mit.edu/) using the default setting values except for the product size parameter where the minimum accepted length was reduced to 70 bp. Primer sequences are listed in Supplementary Data 18. One microgram of tumoural and of their corresponding peritumoural total RNA, as well as of one commercial kidney sample were reverse transcribed using Superscript III (Invitrogen) and oligo(dT) primers according to the manufacturer's instructions. Five microlitres of the cDNA solution corresponding to 5 ng of total RNA was used for PCR experiments in a final volume of 60 µl. Final concentrations of PCR mix included 1 U of HotStar DNA Taq polymerase (Qiagen), 1 × of HotStar DNA polymerase Buffer, 200 µM of each dNTPs, 200 nM of each primer and 2% of dimethylsulphoxide. The initial denaturation/activation step was performed for 15 min at 95 °C, followed by 45 cycles of 30 s denaturation at 95 °C, 30 s annealing at 60 °C and 30 s elongation at 72 °C, and a final extension step of 5 min at 72 °C in a Mastercyler Pro S (Eppendorf, Le Pecq/France). Five microlitres of PCR product was deposited with 1 × loading dye (Solis BioDyne) on a 2.5% agarose gel with 100 bp Ladder (Invitrogen) and analysed by horizontal electrophoresis in 1 × TBE buffer. For each couple of primer presenting a positive amplification in a tumoural sample, a second PCR on 5 ng of total RNA of the tumour and corresponding normal samples was performed in the same conditions to evaluate possible genomic DNA contamination.

Pathway enrichment analyses. Information about gene annotation to *Homo sapiens (hsa)* pathways was obtained from KEGG, release of December 2013, through KEGG ftp database. Total of 6,695 genes (based on entrez gene IDs) are annotated in 280 pathways according to December 2013 release. We used Fisher's one-tail exact test to identify pathways enriched for genes affected by abnormal patterns. We controlled for multiple testing by using the Benjamini–Hochberg FDR procedure. The one-tail Fisher's exact test for overrepresentation calculates hypergeometric probability for observing k or more genes among total K genes of each pathway based on number of total annotated genes and total queried genes.

Data management and availability. Clinical and sample metadata were collected and annotated using a custom-made information system KIDREP, which is based on open-source software[71]. Clinical and processed data are available from the International Cancer Genome Consortium portal.

References

1. Ferlay, J. *et al.* GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. (International Agency for Research on Cancer; 2013), Available from http://globocan.iarc.fr, accessed on 5 February (2014).
2. Chow, W.-H., Dong, L. M. & Devesa, S. S. Epidemiology and risk factors for kidney cancer. *Nat. Rev. Urol.* **7,** 245–257 (2010).
3. Banks, R. E. *et al.* Genetic and epigenetic analysis of von Hippel-Lindau (VHL) gene alterations and relationship with clinical variables in sporadic renal cancer. *Cancer Res.* **66,** 2000–2011 (2006).
4. Nickerson, M. L. *et al.* Improved identification of von Hippel-Lindau gene alterations in clear cell renal tumors. *Clin. Cancer Res.* **14,** 4726–4734 (2008).
5. Latif, F. *et al.* Identification of the von Hippel-Lindau disease tumor suppressor gene. *Science* **260,** 1317–1320 (1993).
6. Wu, X. *et al.* A genome-wide association study identifies a novel susceptibility locus for renal cell carcinoma on 12p11.23. *Hum. Mol. Genet.* **21,** 456–462 (2012).
7. Purdue, M. P. *et al.* Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3. *Nat. Genet.* **43,** 60–65 (2011).
8. Bellmunt, J., Teh, B. T., Tortora, G. & Rosenberg, J. E. Molecular targets on the horizon for kidney and urothelial cancer. *Nat. Rev. Clin. Oncol.* **10,** 557–570 (2013).
9. Scelo, G. & Brennan, P. The epidemiology of bladder and kidney cancer. *Nat. Clin. Pract. Urol.* **4,** 205–217 (2007).
10. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–507 (2012).
11. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32,** 71–75 (2014).
12. Hoang, M. L. *et al.* Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci. Transl. Med.* **5,** 197ra102 (2013).
13. Grollman, A. P. *et al.* Aristolochic acid and the etiology of endemic (Balkan) nephropathy. *Proc. Natl Acad. Sci. USA* **104,** 12129–12134 (2007).
14. Nortier, J. L. *et al.* Urothelial carcinoma associated with the use of a Chinese herb (*Aristolochia fangchi*). *New Engl. J. Med.* **342,** 1686–1692 (2000).
15. Jones, D. T. *et al.* Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. *Nat. Genet.* **45,** 927–932 (2013).
16. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
17. Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22,** 1589–1598 (2012).
18. Guo, G. *et al.* Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. *Nat. Genet.* **44,** 17–19 (2012).
19. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* **45,** 860–867 (2013).
20. The Cancer Genome Atlas Research, N. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–49 (2013).
21. Dalgliesh, G. L. *et al.* Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463,** 360–363 (2010).
22. Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* **14,** R70 (2013).
23. Mikami, S. *et al.* Expression of Ets-1 in human clear cell renal cell carcinomas: implications for angiogenesis. *Cancer Sci.* **97,** 875–882 (2006).
24. Park, J. H., Lee, C., Suh, J. H., Chae, J. Y. & Moon, K. C. Nuclear expression of Smad proteins and its prognostic significance in clear cell renal cell carcinoma. *Hum. Pathol.* **44,** 2047–2054 (2013).
25. Lim, Y. P. Mining the tumor phosphoproteome for cancer markers. *Clin. Cancer Res.* **11,** 3163–3169 (2005).
26. Linehan, W. M., Srinivasan, R. & Schmidt, L. S. The genetic basis of kidney cancer: a metabolic disease. *Nat. Rev. Urol.* **7,** 277–285 (2010).
27. 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).
28. Olivier, M. *et al.* Modelling mutational landscapes of human cancers in vitro. *Sci. Rep.* **4,** 4482 (2014).
29. Ferlay, J. *et al.* Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur. J. Cancer* **49,** 1374–1403 (2013).
30. Debelle, F. D., Vanherweghem, J. L. & Nortier, J. L. Aristolochic acid nephropathy: a worldwide problem. *Kidney Int.* **74,** 158–169 (2008).
31. Stefanovic, V. & Polenakovic, M. Fifty years of research in Balkan endemic nephropathy: where are we now? *Nephron. Clin. Pract.* **112,** c51–c56 (2009).
32. Martin, G. M. *et al.* Somatic mutations are frequent and increase with age in human kidney epithelial cells. *Hum. Mol. Genet.* **5,** 215–221 (1996).
33. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl Acad. Sci. USA* **110,** 1999–2004 (2013).
34. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469,** 539–542 (2011).
35. Burrows, A. E., Smogorzewska, A. & Elledge, S. J. Polybromo-associated BRG1-associated factor components BRD7 and BAF180 are critical regulators of p53 required for induction of replicative senescence. *Proc. Natl Acad. Sci. USA* **107,** 14280–14285 (2010).
36. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46,** 225–233 (2014).
37. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).
38. Cox, T. R. & Erler, J. T. Remodeling and homeostasis of the extracellular matrix: implications for fibrotic diseases and cancer. *Dis. Model. Mech.* **4,** 165–178 (2011).
39. Levental, K. R. *et al.* Matrix crosslinking forces tumor progression by enhancing integrin signaling. *Cell* **139,** 891–906 (2009).
40. Lu, P., Weaver, V. M. & Werb, Z. The extracellular matrix: a dynamic niche in cancer progression. *J. Cell Biol.* **196,** 395–406 (2012).
41. Sadeqzadeh, E., de Bock, C. E. & Thorne, R. F. Sleeping giants: emerging roles for the fat cadherins in health and disease. *Med. Res. Rev.* **34,** 190–221 (2014).
42. Morris, L. G. *et al.* Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation. *Nat. Genet.* **45,** 253–261 (2013).
43. Zang, Z. J. *et al.* Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat. Genet.* **44,** 570–574 (2012).
44. Valletti, A. *et al.* Genome-wide analysis of differentially expressed genes and splicing isoforms in clear cell renal cell carcinoma. *PLoS ONE* **8,** e78452 (2013).

45. Zhao, Q. *et al.* Tumor-specific isoform switch of the fibroblast growth factor receptor 2 underlies the mesenchymal and malignant phenotypes of clear cell renal cell carcinomas. *Clin. Cancer Res.* **19,** 2460–2472 (2013).
46. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* **45,** 860–867 (2013).
47. Fuhrman, S. A., Lasky, L. C. & Limas, C. Prognostic significance of morphologic parameters in renal cell carcinoma. *Am. J. Surg. Pathol.* **6,** 655–663 (1982).
48. Edge, S. *et al. AJCC Cancer Staging Manual* (Springer, 2010).
49. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
50. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
51. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
52. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).
53. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).
54. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23,** 657–663 (2007).
55. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28,** 423–425 (2012).
56. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–681 (2009).
57. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).
58. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
59. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).
60. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
61. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40,** 4288–4297 (2012).
62. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–2017 (2012).
63. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** 1009–1015 (2010).
64. Gonzalez-Porta, M., Calvo, M., Sammeth, M. & Guigo, R. Estimation of alternative splicing variability in human populations. *Genome Res.* **22,** 528–538 (2012).
65. McPherson, A. *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.* **7,** e1001138 (2011).
66. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
67. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* **7**(Suppl 1), S12 1–S1214 (2006).
68. Apweiler, R. *et al.* InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16,** 1145–1150 (2000).
69. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37,** W305–W311 (2009).
70. Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* **13,** R77 (2012).
71. Viksna, J. *et al.* PASSIM--an open source software system for managing information in biomedical studies. *BMC Bioinformatics* **8,** 52 (2007).

Acknowledgements

This study was supported by the EU FP7 under grant agreement number 241669 (the CAGEKID project, http://www.cng.fr/cagekid) and a grant from Génome Québec, le Ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie (MESRST) Québec and McGill University. We also acknowledge the support of Cancer Research UK Centre and ECMC infrastructure funding in Leeds, and are grateful to the sample processing, urology, pathology and oncology clinical teams at St James's University Hospital, Leeds. This work was also supported in part by MH CZ—DRO (MMCI, 00209805) and RECAMO, CZ.1.05/2.1.00/03.0101, Brno, Czech Republic. We are grateful to the patients for consenting to take part in this study. We thank Eamonn Maher and Joerg Hoheisel for valuable advice on the project. This research was enabled in part by support provided by Compute Canada.

Author contributions

G.M.L., G.S., P.B., R.E.B., I.G.G., J.T., A.C.-T., K.S., E.P. and A.Bra. planned the study; A.C.-T. was responsible for ethical requirements; G.S., M.B.W., P.H., L.E., S.M.J., N.S.V., M.S., B.A.-A., C.C., P.J.S., J.J.C., I.H., A.Bri., D.Z., A.Mo., L.F., M.N., D.M., V.J. and R.E.B. were responsible for patient selection, sample collection, sample preparation and pathological reviews; J.V., E.C., M.O., A.Z. and A.Bra. created the database for clinical data; Y.R., A.D., H.B. and M.A. prepared DNA and RNA; Y.R., P.L., A.H.-K., A.D., S.H., M.G., M.T.B., D.L., A.A., A.N., A.Ma., S.R., E.B., I.G.G., K.S., E.P. and G.M.L. contributed to the generation of genomic data; G.S., Y.R., L.G., L.L., M.G.-P., J.R., M.B., M.K., V.R., E.T., J.Se., G.Bo., G.By., J.Z., J.Su., M.F., M.B.W., J.T., N.S.V., R.E.B., A.Bra. and G.M.L. analysed the data; G.S., Y.R., M.K., A.C.T., G.Bo., G.By., J.Z., P.B., J.T., R.E.B., A.Bra. and G.M.L. drafted the initial versions of the paper.

Additional information

Accession codes: Raw sequence data have been deposited in the European Genome-phenome Archive (EGA), under the accession code EGAS00001000083.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Scelo, G. *et al.* Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat. Commun.* 5:5135 doi: 10.1038/ncomms6135 (2014).