FP7-HEALTH-2009 – 241669

CAGEKID

Cancer Genomics of the Kidney

Funding Scheme: **Large-scale Integrating Project**
Topic: **FP7-HEALTH-2009-2.1.1-2**

D9.6 Final report (M54)

Due date of the deliverable: 31th August 2014
Date of latest version of Annex I: 30th of July 2014
Lead organisation short name: CEPH
Author(s): Mark Lathrop

Table of Contents

FINAL PUBLISHABLE SUMMARY REPORT

1. EXECUTIVE SUMMARY

The principal objective of the CAGEKID project was to obtain a comprehensive understanding of genetic and epigenetic changes and resultant downstream proteomic changes in the most common form of renal cancer, conventional clear cell renal cell carcinoma (ccRCC) and, as feasible, non-conventional forms of RCC. Ultimately, the results will be applied to obtain new biological markers that will give better understanding of the disease aetiology, provide new diagnostic and prognostic tools, and lead to new therapies. The project was motivated by the advent of new technologies and associated methodologies to analyse DNA variation and gene expression on a genome-wide basis. CAGEKID was undertaken as part of the International Cancer Genome Consortium (ICGC), which has the goal of obtaining genomic analyses of the 50 most common cancer types.

CAGEKID brought together clinical and epidemiological resources together with the necessary genetics and genomics expertise across Europe to contribute to international efforts to decipher the cancer genome. During the project period (March 1 2010-August 30 2014), the consortium completed sample collections, provided integrated genomic assessments of ccRCC and non-conventional (non-clear cell) forms of renal cancer, and conducted follow-up genomic and proteomic studies. An enduring legacy of the project will be the strong collaborative network across Europe. Adopting a European-wide approach furnished us with ability to obtain the clinical collections and biological resources for these studies, which are not available within any single country. The results from the genomic analyses address needs for molecular markers in renal cancer, which is one of the few cancer sites in which such markers are not yet available for clinical use, while also providing a unique opportunity to investigate the variable incidence of the disease in Europe. The clinical and biological resources collected by CAGEKID will be available to address future scientific and public health questions. CAGEKID responded to the specific requirements of the FP7 call by structuring EU participation in the ICGC. CAGEKID has served as a test case for addressing ELSI (Ethical, Legal and Social Implications) of cancer genomics and other large-scale genomics research. Another important contribution of the project has been in the development and diffusion of methodologies and bioinformatics tools for studies involving whole-genome resequencing, and related techniques.

The most striking finding from CAGEKID is that ccRCC is characterised by a specific mutational signature in large proportion of patients from central Europe. These patients' tumours have a very high number of mutations (markedly higher than other patients) with a predominance of A:T>T:A transversions. Experimental and other data show that this pattern is consistent with exposure to aristolochic acid (AA), a previously unknown risk factor for common cancer. This shows that the processes underlying ccRCC tumourigenesis vary in different European populations, and suggest that AA may be an important ccRCC carcinogen in some regions, a finding with major public health implications to be pursued beyond the project. Many additional results on the genomic landscape of RCC in Europe are described in the final project report.

2. SUMMARY DESCRIPTION OF PROJECT CONTEXT AND OBJECTIVES

The CAGEKID consortium has undertaken a comprehensive investigation of genetic and epigenetic changes and resultant downstream proteomic changes in the most common form of renal cancer, renal cell carcinoma (RCC). This is in order to obtain new biological markers that will give better understanding of the disease aetiology, provide new diagnostic and prognostic tools, and lead to new therapies. RCC is a significant public health problem particularly in Europe, where in some regions the incidence is highest worldwide and increasing. Our investigations will meet a major need to identify novel biological markers for RCC, which is one of very few tumour types for which there are currently no biological markers that are in routine clinical use.

The approaches used by CAGEKID were motivated by the advent of new technologies and associated methodologies to analyse DNA variation and gene expression on a genome-wide basis. These advances provide new opportunities for studying the molecular basis of cancer, and obtaining novel biomarkers of disease for diagnostic applications and the development of new treatments. Comparative analysis of transcript patterns in tumour and normal tissue with expression arrays covering the genome has already proven to be a powerful diagnostic tool in several cancer types. High-density single-nucleotide polymorphism (SNP) genotyping has uncovered many constitutional DNA markers that are associated with increased risk of developing specific cancers, and SNP and other array-based approaches are important tools for identifying chromosomal rearrangements in cancer. Furthermore, a new generation of sequencing technologies has made possible the analysis of DNA and expression patterns in an unbiased manner (i.e. without a priori selection of candidate genes) at very high resolution, providing much additional power to detect novel gene modifications in cancer. Prior to the start of the project, several preliminary studies had shown the application of these technologies leading to the identification of novel genetic markers and affected pathways for several cancer sites. Such markers can be examined with respect to disease progression and treatment responses, and potentially used for the development of targeted therapies. Such results strongly suggest that the application of systematic sequencing approaches to obtain comprehensive catalogues of genetic modifications in different cancers will have a profound impact on the understanding and treatment of these diseases in the near future. Achieving this impact requires mobilisation of research on an international scale.

The above considerations led to the formation of the International Cancer Genome Consortium (ICGC) with the ultimate goal of full characterisation of the constitutional and tumour genomes for 25,000 patients (50 cancers, 500 patients per cancer), the largest genome-based project to date. CAGEKID consortium scientists participate in the ICGC, and are members of the scientific steering committee, biobanking, technology, data analysis, ethics and other ICGC working groups.

In this context, CAGEKID provided a unique opportunity for involvement of a European-wide consortium in the ICGC through a study of renal cancer. RCC represents a significant public health issue within Europe, and it is a cancer that can only be effectively studied on a European-wide basis and through an integrated consortium with clinical and genomic expertise

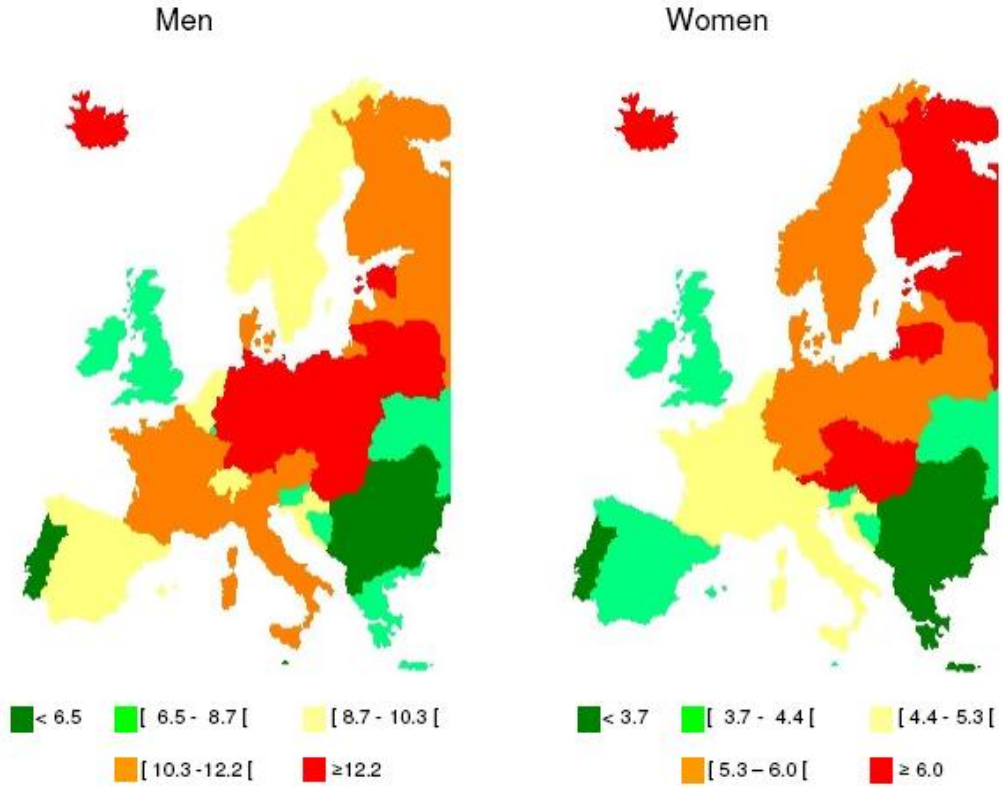
Renal cancer Our aim was to carry out comprehensive detection of DNA markers for conventional (clear cell) renal carcinoma, a significant public health issue within Europe and an important tumour type for inclusion in the ICGC. Because surgical resection of the kidney is a standard intervention in RCC, extensive amounts of fresh tumour and non-tumour tissue are available. This, together with several compelling clinical needs as described below, have led to the International Cancer Genome Consortium (ICGC) recognising that RCC is a particularly appropriate tumour for genomic investigations.

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 (Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr, accessed on 5 February 2014). Approximately 90% of renal cancers are renal cell carcinomas (RCC) that develop in the renal parenchyma, with conventional (clear cell) RCC (ccRCC) being the most common (70-80%) histological type. Somatic mutations or epigenetic alterations of the von Hippel-Lindau tumor suppressor gene (*VHL*) are observed in >80% of ccRCC. A modest proportion (2-4%) of RCC shows Mendelian inheritance and most of these cases are caused by germ-line mutations in *VHL.* Results from genome-wide association studies have identified common germ-line variants associated with increased susceptibility for developing ccRCC, and recent sequencing efforts including ours have revealed recurrent somatic mutations in a number of genes including *PBRM1*, *SETD2* and *BAP1.* 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.

Renal cancer is of particular significance within Europe, as the incidence rates are very high in areas of Central Europe and some other European regions compared to elsewhere in the world (Fig. 1). Incidence rates have been increasing sharply with unexplained variation in different countries. The highest rates that are observed worldwide occur in Central Europe, 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 UK are 10.9/100,000 and 5.8/100,000. Outside of central Europe, similar high kidney cancer rates are also reported within the US among African American populations, whereas the lowest incidence rates in both men and women are found in Africa and Asia. A sharp recent increase in the incidence has also been observed, most notably in the Czech Republic and among the black population in the USA. The epidemiology of RCC is only moderately understood, with known lifestyle risk factors including tobacco smoking, obesity, hypertension and a history of diabetes, and these alone do not account for the marked ethnic and geographical variations in incidence. Approximately 50% of patients diagnosed with kidney cancer are alive after 5 years. There are also important geographical disparities for survival within Europe, being around 70% in Germany and Austria, and less than 50% in the UK. Survival is influenced not only by stage at diagnosis, but also by age, being on average 70% for cancers diagnosed before the age of 45, to 40% for cancers diagnosed at age 75 or over.

Over 80% of RCC are histologically of conventional (clear cell) subtype. VHL gene mutations have been reported to occur in the majority of sporadic conventional RCC. Previous reports have indicated a mutation level of between 50% and 60%, although two recent large and independent analyses from CAGEKID PIs indicate that approximately 80% of tumours are likely to harbour VHL mutations or methylation changes. Identification of additional biological markers of the disease is likely to be important for further understanding of disease heterogeneity, and may help to explain differences in disease incidence and survival patterns across Europe and elsewhere.

Fig. 1 Variation in the incidence of renal cell cancer There are marked worldwide geographical variations in incidence rates of kidney cancers in men and women with the highest rates found for both sexes in Czech Republic (24.1/100,000 in men and 10.5/100,000 in women). In men, intermediate high incidence rates are found in Estonia (17.3/100,000), Iceland (16.5/100,000), Lithuania (14.7/100,000), Hungary (14.7/100,000), Slovakia (13.7/100,000), and Poland (13.5/100,000). Among females, the intermediate high incidence rates are found in Lithuania (8.4/100,000), Iceland (8.1/100,000), Estonia (7.1/100,000), Austria (6.8/100,000), Finland (6.7/100,000), Slovakia (6.6/100,000), Hungary (6.6/100,000), and Australia (6.5/100,000). In both sexes, the lowest rates are found in Africa and Asia.



In the absence of further biological markers of disease, histological diagnosis, detection of relapse and the monitoring of response to therapy require either invasive procedures or the use of cross-sectional or other radiological imaging approaches. Invasive procedures carry risks and morbidity whilst cross-sectional imaging is unpleasant for patients, expensive, can involve anxiety and certainly involves a delay before the result is communicated to the patient. Patients undergoing surgery are usually followed clinically and with intermittent imaging tests. In terms of prognosis, pathological factors alone or combined into algorithms or nomograms are most often used, although estimates of risk are relatively wide for individual patients. The tumour, node, metastasis (TNM) staging system (TNM staging atlas, Rubin P and Hansen JT, Lippencott, Williams and Wilkins, 2007) is currently the most extensively used tool for providing prognostic information for RCC while performance status and tumour grade have also been shown to be independent prognostic factors using multivariate analysis. Pathological stage and nuclear grade are the two factors consistently included in these prognostic models, but the major problem from a pathological perspective is that of reproducibility with variations in assignation of stage depending on the experience and meticulousness of the pathologist performing the dissection and only moderate agreement on grade between pathologists or by the same pathologist on different days.

Radical nephrectomy is standard treatment for localised RCC but approximately ~30% of patients of these patients will go on to develop metastatic disease. Due to the relatively asymptomatic nature of the illness, approximately 40% of patients already have locally advanced or metastatic disease at the time of diagnosis for which therapeutic options are more limited due to inherent resistance to chemotherapy and radiotherapy. The prognosis of patients with metastatic disease is poor with a 5-year survival of less than 10%. As indicated above, the VHL tumour suppressor gene has been implicated in conventional/clear cell RCC and this represents a prime example of how knowledge of underlying genetic changes can be used for rational treatment design. A series of recently introduced with anti-angiogenic properties such as sorafenib, sunitinib and bevacizumab are based on inhibiting downstream pathways affected by the VHL gene, particularly VEGF-mediated changes, and have increased the options for patients with renal cancer with improved response rates and relapse-free survival. This illustrates that knowledge of genomic changes in RCC can lead to the development of new therapies. However, these agents are expensive, carry significant toxicity and not all patients will respond underlining the interest in identifying additional genomic markers of disease.

3. PRINCIPAL RESULTS

The following describes the principal achievements of the project from the start date to the end of the final reporting period.

Sample collections & preparation The CAGEKID collaborators have established the largest set of well-characterised RCC samples available to date, with representation from Eastern, Central and Western Europe (n=1,837), all meeting stringent ICGC study entry criteria for the discovery and initial validation sets.

The criteria as agreed at the start of the study were: ≥ 18 years of age, diagnosis of conventional (clear cell) renal cell carcinoma (ccRCC) or non-conventional RCC for the non-conventional sub-study, 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 hemodialysis.

A total of 3,253 potential cases in Leeds (926) and at IARC (2,327) entered the study at the screening level. All were screened for potential eligibility in terms of meeting the defined clinical criteria and also having suitable tissue (including diagnostic FFPE sections), blood samples and associated clinical data available for use. From these, 2,184 entered the study for pathology review, with a final total of 1,837 being accepted for study (35 with no consensus).

The breakdown of cases for the different phases is as follows:

Conventional RCC discovery set (n=121; 94 for whole genome resequencing with 64 of these also used for RNA sequencing plus an additional 27 for RNA sequencing alone). These include samples from the Czech Republic (38 patients), Romania (14 patients), Russia (38 patients) and the UK (31 patients).

Initial validation set conventional RCC (n=394). These include samples from the Czech Republic (113 patients), Romania (28 patients), Serbia (7 patients), Russia (111 patients) and the UK (135 patients).

Non-conventional RCC discovery set (n=73). These include samples from the Czech Republic (24 patients), Romania (8 patients), Russia (19 patients) and the UK (22 patients).

Combined RCC final validation set (n=1,249). These include 897 conventional RCC cases, 317 non-conventional cases and 35 with no consensus currently. Broken down by country samples are included from Czech Republic (289 patients), Serbia (75 patients), Romania (148 patients), Russia (175 patients) and the UK (562 patients).

In all cases, 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 (PH, LE, MS). Tumours were subtyped histopathologically using the WHO classification and graded using the Fuhrmann grading system (the caveats of basing this on a single section only are realised).

For those samples included in the discovery or initial validation sets where diagnosis of ccRCC or non-conventional 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 (PH, LE, MS) to confirm diagnosis and achieve a consensus assessment of grade, 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. Additionally 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 absence of tumour cells and assessed for viability and the degree of inflammation.

For the final validation set, 1,424 cases were entered (1,249 accepted) in total. Of these 621 cases entered (490 accepted) followed the process above but with a threshold of at least 50% viable tumour cells being used, whereas 803 entered cases (759 accepted) were entered as FFPE material only. In these latter cases, pathology review of H&E sections only was undertaken as above and a threshold of at least 50% viable tumour cells being used. Of these, 581 were prepared in parallel with the TMA (described in Task 1.7) with a further 178 prepared in Leeds recruited through additional centres in the UK. In all cases one pathologist selected areas for coring and a minimum of 3 x 1 mm cores were taken from each block for DNA extraction.

Standard operating procedures for recruitment Protocols for case recruitment were harmonized across all recruiting centres to ensure a high-quality biosample collection. SOP have been developed and implemented following the original plan, and distributed through the CAGEKID website for use by other groups.

The SOPs cover the following activities:

- Patient recruitment and clinical, demographical, lifestyle and outcome data collection
- Collection and processing of blood samples
- Collection and processing of renal tissue biospecimens (tumour and normal tissues)
- Collection of initial pathological diagnosis data and review of diagnosis
- Process and review of renal tissue biospecimens (tumoural and normal tissues)
- DNA purification from blood samples
- DNA purification from renal tissue biospecimens (tumour and normal tissues)
- RNA purification from renal tissue biospecimens (tumour and normal tissues)
- Sample shipping and transportation

- Compiling “dry” data (clinical, demographical, lifestyle, outcome, pathological data) and tracking biosamples and aliquots
- Overall workflow for case selection
- Biobank quality control and assurance

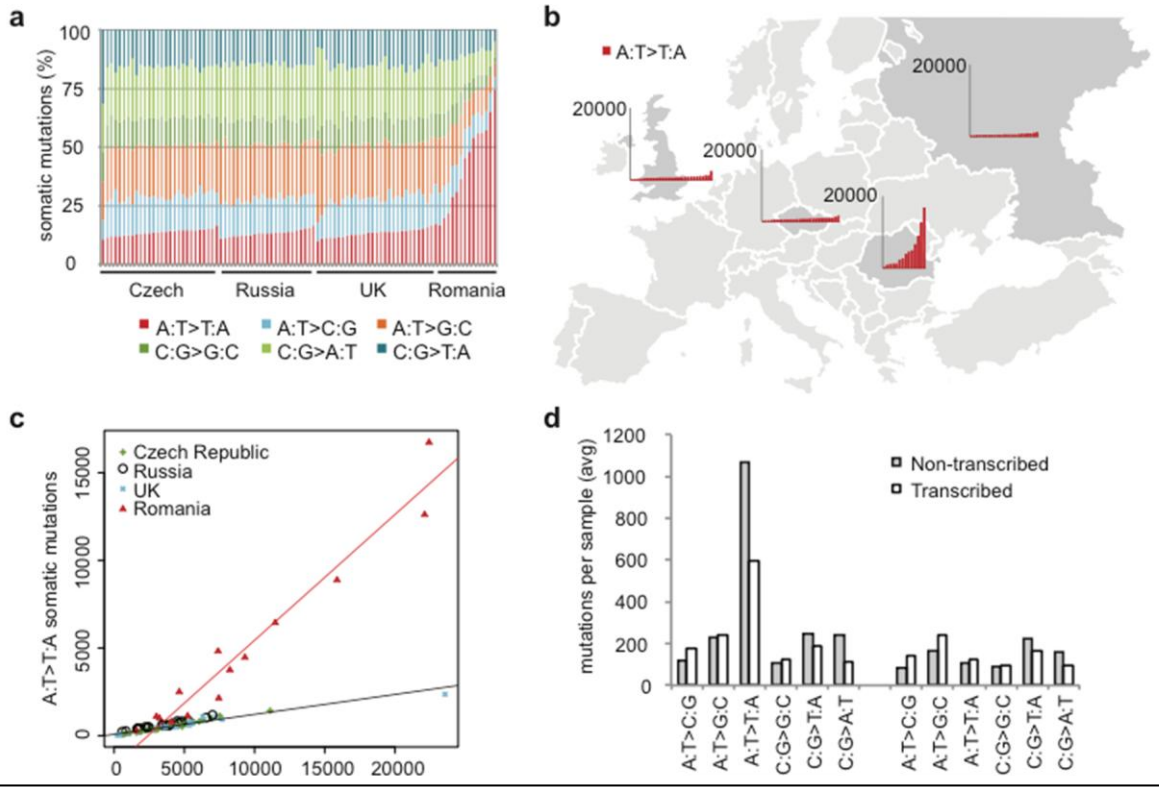
Construction of tissue microarrays (TMAs) TMAs are important resources for downstream characterisation of disease markers. The CAGEKID consortium provided an important opportunity for obtaining such a resource. From the cases with only formalin-fixed paraffin-embedded tissue available, 610 FFPE blocks of tumour tissue were sent to Karolinska Instituted, Stockholm, Sweden from Leeds, UK (392 blocks) and IARC, France (218 blocks). Additionally 21 blocks with normal tissue were sent from Leeds. Two adjacent sections were cut from each block and H & E stained with one set being simultaneously reviewed and areas of tumour cells marked for coring by a pathologist (LE) in Sweden and one set scanned in Leeds for second pathologist review.

A total of 581 cases met the criteria of >50% viable tumour cells and were entered into the study (described above for FFPE only cases). In each case 3 x 1 mm cores were obtained using biopsy punches (43 cases required additional cores) and were shipped for DNA extraction.

Six TMA blocks were constructed in triplicate. A total of 518 of the 581 tumour samples described above were used for construction of the six TMA blocks, comprising 406 conventional, 42 papillary, 24 chromophobe RCC and 46 oncocytomas. In addition to the tumour cores each TMA block contained six normal kidney cores (three from medulla, three from cortex).

Genomic analyses To explore the underlying genomic architecture of RCC and variability in Europe, we undertook whole-genome and transcriptome sequencing of clear cell RCC (ccRCC), the most common form of the disease, in more than 90 patients from four different European countries in CAGEKID collection with contrasting disease incidence. We undertook a comprehensive molecular evaluation of the samples using WGS, 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 54X coverage in all of these samples. 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 wide methylation patterns were measured quantitatively for 86 matched ccRCC sample pairs (all with WGS) using Illumina 450K methylation arrays.

Fig. 2 Variation in the incidence of renal cell cancer 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 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.



Somatic mutations and structural variation We detected 4,904 somatic mutations on average per sample pair corresponding to a mean of 1.79 somatic mutations/Mb after correcting for regions with low coverag. Intergenic regions were seen to have higher mutation rates (2.02/Mb) than other genome regions. Notably, the overall somatic mutation rate in coding regions was significantly less than this (1.54/Mb, rate ratio = 0.76, 95%CI 0.74-0.79) corroborating previous literature. Similarly other regions associated with genes (e.g. 5'UTR, 3'UTR, introns) had lower rates than intergenic regions. Regions corresponding to DNaseI hypersensitive (DHS) sites also

had a lower overall somatic mutation rate (1.66/Mb) when compared to intergenic regions (rate ratio = 0.82, 95%CI 0.81-0.83) in concordance with a recent report in other cancers. We identified loss-of-heterozygozity and other copy number variants (CNVs) at 1,019 sites (with sizes ranging from 450kb-197.7Mb and between 0-48 CNV differences per pair), and we found 1,421 additional putative structural rearrangements with further analyses as described. Among recurrently affected loci, loss of chromosome 3p, the most frequent CNV, was observed in 90% of samples,

Our findings support previous reports on frequent aberrations in the epigenetic machinery and PI3K/mTOR signaling, 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. Strikingly, 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 tumourigenesis may vary in different populations, and they strongly suggest that AA may be an important ccRCC carcinogen in Romania, a finding with major public health implications. Evidence of AA exposure has never previously been implicated in a common cancer.

In CAGEKID data, 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 (UTUC) in patients as well as in cultured primary cells 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. We have generated additional data, as yet unpublished, with CAGEKID that confirms the very high frequency of this signature in other Romanian patients, and we are performing whole-genome sequencing in other sample from Croatia, Serbia and Bosnia-Herzegovina collected as part of the follow-up of this result.

Unlike in UTUC, 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 higher frequency of somatic mutations of PBRM1, which is the second most common mutated genes in ccRCC, among the Romanian outliers as compared to 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. Furthermore, it has been shown that PBRM1 is required for TP53-driven replicative senescence. 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 ccRCC, it is relevant to investigate the extent to which PBRM1 status influences somatic mutational patterns. Likewise, the high rate of somatic mutations in epigenetic regulators supports the importance of chromatin remodeling/histone modification pathways in ccRCC as suggested recently.

While 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 this may be due to the increased number of somatic mutations by age observed in kidney epithelial cells, involvement of a cancer-associated deficiency in related cellular processes such as DNA repair cannot be excluded. Further studies in which patient-matched non-tumour kidney tissue samples are analysed in addition to the tumours are being undertaken to further understanding of these processes.

The prevalence of *VHL* mutations in our patients (73%) is similar to that reported in our previous studies (70-80%) but substantially higher than those found in recent next-generation sequencing (NGS) studies of ccRCC, which have reported *VHL* mutations in between 27%-55% of samples. While some of the variation may be due to ethnic origin, patient selection, pathology review criteria and/or sample tumour cell content, false-negative NGS results have been evoked as a factor affecting previous studies.

Additional known ccRCC genes mutated in our cohort include *PBRM1* (39%), *SETD2* (19%), *BAP1* (12%), and *KDM5C* (7%). Interestingly, *PBRM1* was mutated in 10 of the 12 Romanian outliers, with three of these mutations having the characteristic transversion pattern discussed above. 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. Novel genes with recurrent mutations in our data include *FAT3* (7%), *WDFY3* (6%) and *ANPEP* (6%). 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.

Analysis of the relationship between clinical variables and presence or absence of somatic mutations in genes with a mutation frequency of >10% identified significant associations between both *PBRM1* mutations (p=0.043) and *SETD2* mutations (p=0.014) with higher stage tumours and these plus additional clinical associations will be explored further after the project end.

In line with recent reports, the PI3K/AKT/mTOR pathway was recurrently affected by somatic mutations and abnormal gene expression patterns (see below) 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 (RTK) 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. COL5A3 encodes alpha 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 progression39 either by

affecting downstream growth-promoting pathways such as ERK and PI3K signaling40 or by contributing to angiogenesis and metastasis by influencing tumour microenvironment and tumour-stroma communication. 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 signaling pathways such as Hippo and Wnt signaling. Emerging evidence has shown that FAT genes are mutated in different cancers, and has suggested tumour suppressor activity for these genes. Together with our data, this motivates further research to dissect the mechanisms by which FAT abnormalities can contribute to cancer.

Another novel finding from our data is widespread loss of chromosome Y occurring in the majority of male patients in our study. This finding will be pursued further beyond the end of the CAGEKID project.

Transcriptome Analysis of transcriptome patterns highlighted significant alterations in metabolic pathways consistent with the Warburg effect. This phenomenon is a hallmark of ccRCC22 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, an approach limited to the study of a subset of splicing events. By analyzing differences in exon usage and major transcript expression patterns, we showed that while major recurrent changes in splicing are rare, splicing patterns are broadly altered in ccRCC, consistent with the observation that the mRNA processing pathway is commonly disrupted in ccRCC.

We found that 12,849 protein-coding genes (60% of genes annotated as such in Ensembl 66) were expressed on average at 1 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 two-fold change between tumour and normal (FDR<0.01). Hierarchical clustering did not reveal any subgroups with correlations with clinical variables.

Following *in silico* screening and validation by RT-PCR, we identified six tumour specific fusion events. Two fusion partner genes (*MED15* and *TFE3*) shown in Figure 4 participate in the TGFβ/SMAD pathway, known to play a role in renal cancer development, while nine out of the twelve fusion partners (*CWC25*, *CGNL1, SH2D3C*, *RAB31, LRSAM1, MED15, SLC12A4, TFE3, TCF12*) code for phosphoproteins, which are known with important roles in signaling pathways. All the confirmed chimeras appeared to be associated with inversion, since the fusion partner genes were located on opposite strands.

Using the KEGG datasets, we performed pathway analysis for genes affected by somatic mutations or transcriptome alterations. Pathway analysis of the 583 genes harboring somatic mutations showed significant enrichment for focal adhesion and PI3K pathways (FDR= $5x10^{-07}$ and FDR=0.003, respectively). Overall, 59% of tumours in our study showed non-silent somatic mutations in one or more of 32 genes from these two pathways. Recurrent mutations in constituents of PI3K-mTOR signaling have recently been reported in ccRCC. We observed non-silent somatic mutations or CNVs at each of *PIK3R1*, *PTEN* and *MTOR* in more than 5% of tumours, and somatic variations at lower frequencies in other genes involved in PI3K-mTOR signaling. 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.

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, carbon metabolism, the citrate cycle, and amino acid metabolism. Among key pathways enriched for upregulated genes were cytokine-cytokine receptor interaction, cell adhesion molecules, and the chemokine signaling. Focal adhesion and PI3K pathways were also enriched for upregulated genes (FDR=$2x10^{-5}$ and FDR=$2x10^{-8}$, respectively). “Metabolic pathways” was the only KEGG pathway that showed evidence of enrichment for genes involved in switch events (using the criteria of a two-fold expression difference seen in at least two patients). However, the protein processing in endoplasmic reticulum and mTOR signaling pathways also showed some evidence of enrichment (p=0.001 without multiple testing correction, FDR=0.14). 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.

Methylation 897 genes exhibited significantly different methylation values at promoter or transcription start sites in at least 50% of the pairs (553 hypermethylated in the tumour and 344 hypomethylated in the tumour). VEGFA, which was overexpressed in tumours and affected by somatic mutations (see below), was hypomethylated in 83 (97%) of the pairs, and the methylation values had a significant inverse correlation with VEGFA expression levels (ρ=-0.61, p=$5x10^{-5}$). We found five tumours to be hypermethylated at VHL, of which three do not harbour VHL mutations.

In contrast to our findings on non-conventional forms of RCC described below, we identified no methylation signals that appeared to identify relevant subsets of ccRCC, we nevertheless analysed both 450K Illumina arrays, and undertook targeted follow-up of specific loci. The targeted follow-up was performed on a total of 300 samples tumours, including 200 tumours, 100 with their paired peri-tumoural tissues, more than originally planned In total 28 regions were analysed; exceeding slightly the number of 25 targets initially planned. The analyses showed excellent correlation between the high resolution quantitative data obtained with the Pyrosequencing technology and the genome-wide analysis using the 450K BeadChip data. The analysis of a large number of CpGs surrounding the CpGs analyzed on the 450K BeadChip

showed that the DNA methylation alterations are stable over the analysed region and might represent changes of potential biological importance.

An objective in CAGEKID was to set up an assay based on the detection of DNA methylation patterns of circulating DNA in the plasma of ccRCC patients for the diagnosis of ccRCC (or other cancer). As found no correlation between clinical parameters and DNA methylation patterns could be defined, the assay development successfully focused on methodology for the detection and potentially early detection of ccRCC. Eleven different protocols for the extraction and quantification of DNA from plasma were evaluated and we identified protocols giving a high and reproducible yield. Nineteen DNA methylation positions that were highly methylated in the tumour, but showed little methylation in the peritumoral and corresponding PBMC samples were identified from the 450K data and methylation specific QPCRs were successfully implemented for six of them, with some of them being analysable in a multiplex format.

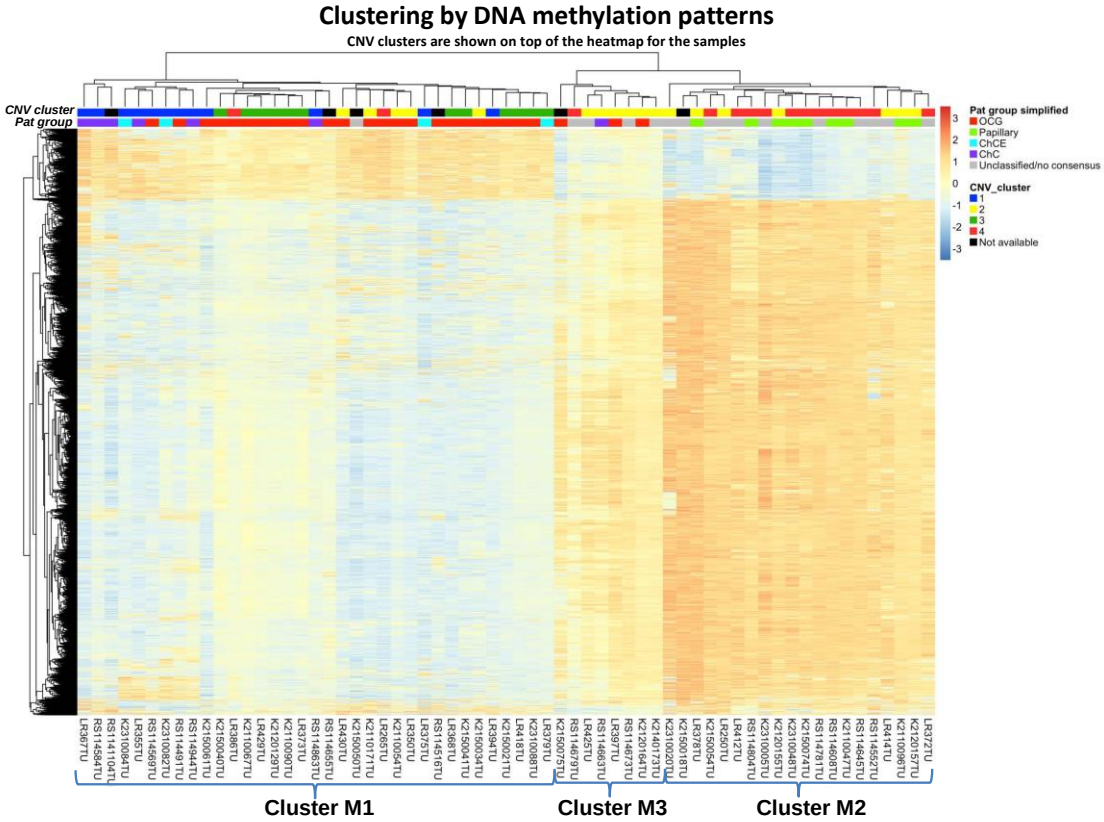
AT mutational signature follow-up As indicated above, our initial study of samples from patients affected by ccRCCs revealed that a majority of tumours procured from Romanian patients (12 out of 14) exhibited a mutational signature consistent with exposure to aristolochic acid (AA). This pattern was limited to Romanian patients. In a follow-up study we performed whole-exome sequencing (WES) on tumour, peri-tumour and and blood DNA from 30 additional Romanian patients which confirmed the existence of the same mutational pattern in these samples. On the other hand, Balkan endemic nephropathy (BEN) which has been associated with ingestion of seeds of Aristolochia clematitis affects people of alluvial plains along the tributaries of the Danube River in Croatia, Bosnia and Herzegovina, Serbia, Bulgaria and Romania. Accordingly, we have extended our investigation on possible involvement of AA in the aetiology of ccRCC in these regions by performing whole-genome sequencing (WGS) of additional samples from Croatia, Serbia and Bosnia and Herzegovina in an on going research project. The use of WGS vs. WES has been chosen based on predominance of AA mutational signature in non-coding regions allowing improved detection of the signature.

Non-conventional RCC We have also performed whole-exome sequencing and DNA methylation profiling of non-conventional renal cell carcinomas collected through CAGEKID program. The initial analysis included samples from 76 patients diagnosed with different pathological subtypes of RCC or appearances within this spectrum including 11 chromophobe (classical: ChC; eosinophilic/classical: ChCE), 28 within the distal nephron tumour spectrum oncocytic/chromophobe group (OCG), 9 papillary (type I: Pap I; mixed type I and II: Pap M), and 17 unclassified tumours where consensus wasn't achieved, several of which had elements of papillary architecture (Uncl Pap).. DNA methylation profiling highlights widespread differences in tumour-associated DNA methylome between papillary tumours and the other non-conventional RCC cases suggesting a distinct epigenetic signature for papillary tumours. Within the definite chromophobe and oncocytic/chromophobe grouping, similar DNA methylation profiles were seen across these cases , but recurrent somatic copy number variations (CNVs) differed. Similarly, CNV patterns of papillary tumours are distinct from those of other tumours type. Combination of somatic CNVs and DNA methylation profiles provides distinct epi-genomic signature groupings within non-conventional RCCs. Using these signatures, several, but not all, of the unclassified tumours grouped with specific pathological subtypes. . In addition, we noted

that a minority of unclassified tumours harbour somatic mutations in driver genes of different subtypes such as in VHL (driver for ccRCC) and in Met (driver for papillary RCC) simultaneously. This observation reflects a complex heterogeneity that exists at genetic level in these tumours, which may comprise a distinct subtype of RCC.

The analysis of DNA methylation data revealed global variations across sample pairs. Subsequently, unsupervised clustering of samples based on differences in methylation levels between paired tumour and normal samples (using methylation data of 3098 loci with the most variable cancer-associated patterns) resulted in three clusters of samples, which are hereafter referred as clusters M1 to M3 (see Figure 3). We observed that cluster M1 was composed of ChC, ChCE and OCG tumours whereas none of these tumour types were found in subjects of cluster M2. Cluster M2 was composed of nineteen cases including all 9 patients with papillary cancers and 10 cases with unclassified tumours, of which 8 had “papillary in part” based on pathological information. Cluster M3 was composed of samples who exhibited methylation patterns similar to those observed in cluster M2 subjects, but the methylation differences between sample pairs were weaker in M3 subjects. Therefore, it seems that two main clusters of patients are identified by DNA methylation patterns that can distinguish papillary tumours (cluster M2) from other tumour types (cluster M1). Notably, the grouping of samples in cluster M1 indicates the existence of subgroups within this cluster, which are seemingly following pathological diagnosis types. For example, although it is not a perfect grouping, a co-clustering pattern is observed for diagnosed ChC and ChCE cases within cluster M1.

Fig. 3 Methylation patterns distinguish clusters of non-conventional RCC



Taken together, these data indicate that papillary tumours represent a distinct DNA methylome, which distinguishes them from other non-conventional RCCs. In comparison to papillary tumours, ChC and OCG tumours exhibit a similar global DNA methylation patterns; however, they look different in the extent of abnormal DNA methylation (hyper- or hype-methylation).

Additional follow-up From our discovery sequencing set results, a panel of 56 genes of interest was developed and tested, and applied to follow-up samples (394 pairs or 788 samples sequenced). The 56 genes were chosen based on high frequency of mutated ccRCCs (>5%) in Cagekid or other major studies, predominance in tumours with non-mutated VHL, association to clinical features (outcome) or recurrence in non-conventional RCCs (see Table below). For the follow-up sequencing experiments, sequencing libraries are generated using Nextera technology of Illumina, which requires 50ng of DNA as starting material. According to the manufacturer protocols, libraries generated from 12 samples are pooled together and target regions are captured following Illumina TruSeq protocols. Subsequently, captured regions from 96 samples are sequenced on on lane of HiSeq instrument resulting in at least 150X in depth of coverage. The sequence data have been generated but because of the delays in sample collection, the analysis will be finished after the project end date, together with further validation on our extended validation sets.

Public deposit of data The analysis of the data from CAGEKID has resulted in novel discoveries, however to make these available for further research they need to be archived and disseminated to the scientific community. The archival resources and related infrastructure available through the European Bioinformatics Institute, the CAGEKID partner, were used to achieve this goal. The European Bioinformatics Institute is the host of the European Genome-phenome Archive (EGA), which was the most appropriate archive for archiving and dissemination of the identifiable data. Moreover, EGA has also been the designed archive of data generated in all ICGC projects. The fact that access to ICGC project data deposited in EGA is controlled by the ICGC Data Access Committee provides a solution the legal and ethical issues related to identifiability of the sequencing data. Therefore, the decision was made to use EGA as the main archive for the sample/phonotype and raw sequence data generated in the project. Additional sample and phenotype data are archived in the ICGC portal. Archiving raw data from DNA and RNA sequencing of the tumour and normal samples from CAGEKID patients along side with the patient medical data and all necessary sample information gathered in the project, and dissemination of these data with the help of the ICGC Data Access Committee.

The analysis of these data have resulted in novel discoveries and their publication, however to make these data available for further research they need to be archived and disseminated to the scientific community. The archival resources and related infrastructure at the European Bioinformatics Institute were used to achieve this goal.

From the archiving perspective the data can be divided in two major types – data the can be disseminated publicly and data that requires controlled access. Generally the data that may potentially lead to the donor indentifiability fall into the second category, while the rest of the

data into the first. The potentially identifiable data includes all sequence data (DNA and RNA) and may include some of the (anonymised) phenotypic data, such as the date of birth.

The main non-identifiable data type generated by the project was gene and transcript level expression data, which is obtained by processing the raw RNA sequencing data. Researchers are often interested in these processed data – gene and transcript expression levels, rather than the raw sequences. These data can be disseminated publicly, which lower their usage barriers. Therefore it is important also to achieve maximum dissemination of processed gene expression data. The permanent public archive of the functional genomics data at the EBI – ArrayExpress was chosen for this task.

It was decided that the most appropriate data to disseminate via EGA was the raw sequencing reads aligned to the genome – the so-called BAM files. To achieve this goal, whole genome sequencing data from 95 matched blood normal and tumour patients mapped with BWA aligner with an average depth 54 x coverage was uploaded into the EGA archive via the EGA uploading system (https://www.ebi.ac.uk/ega/submission /tools/EGA_webin_data_uploader). Similarly, RNA sequencing data from 45 matched normal/tumour and 46 non-matched tumour patients were mapped with tophat version 2 splice aware aligner and the generated bam files uploaded securely into the EGA archive. Briefly, all mapped data files were encrypted prior submission to ensure the secured access to the data and post encryption md5sum values for each file were generated to preserve the integrity of the data. Subsequently, the encrypted files were submitted alongside with the md5sum values via the EGA Webin Data uploader tool.

The uploaded data is under study ID EGAS00001000083 consisting of 5 datasets : EGAD00001000709 describes 95 normal DNA samples isolated from blood ; EGAD00001000717 consists of 95 tumour DNA samples ; EGAD00001000718 describes 91 tumour RNA samples ; EGAD00001000719 consists of 45 matched normal RNA samples and EGAD00001000720 describes 90 matched normal/tumour RNA samples.

The raw and normalized counts per gene and per sample alongside with the metadata describing some of the clinical data was loaded into Array express with an accession number E-MTAB-3033. Metadata about the experiment aims, submitter contact details, samples, and sample and data processing protocols were entered into MAGE-TAB format files (the Investigation Design Format file and Sample and Data Relationship Format file) by a curator for loading into the ArrayExpress database. Sample information was broken down into specific sample attributes such as sex, country of origin, and tumour stage. Ages were binned into 5year ranges. Experimental Factor Ontology (EFO, www.ebi.ac.uk/efo) sample attribute types and values were used where possible. This limited phenotype was chosen for public dissemination, the full phenotype is available via the ICGC data portal. The study was also annotated with EFO experiment type terms and a link to the read data in EGA added. The processed data was included in the study data for ArrayExpress. The study was then inserted into the ArrayExpress submission processing system, checked for MAGE-TAB validity and then loaded into the ArrayExpress database. MAGE-TAB files and processed data were made available for download on the EBI web site and FTP server (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3033/files/).

As described below, we developed a software tool (KIDREP) designed specifically to handle the integration of clinical, histological and clinical data associated with the project. The integration of data acquired during the CAGEKID project with ICGC involved two major subtasks: 1) development and regular updating of mapping between KIDREP data fields and data submission fields as specified in ICGC submission manual; 2) development of KIDREP software module implementing data export according to specifications of mapping between KIDREP and ICGC data fields. A number of related subtasks also included providing linking of KIDREP data with in-house sample barcodes used by other project partners. This was needed to correctly link sample and assays data sets that were submitted to ICGC and EGA and has to be done in somewhat implicit way to guarantee the preserving of anonymity of sample and patient data.

The development and updating of mapping between KIDREP data fields and data submission fields as specified in ICGC data dictionaries was done in close collaboration with project Partners 1, 2, 5, 8 and 12. The first version of data mapping has been completed already in April 2011, however in response to the changes in specifications in newer releases of ICGC data dictionaries the mappings had to be regularly (and sometimes considerably) redeveloped. The last changes in mapping were introduced in response to release of ICGC data dictionary v.0.10a as recently as September 2014.

Whilst regarding semantic content there is a good compatibility of data fields stored in KIDREP and these required by ICGC specifications, the structuring of these data is considerably different, making generic KIDREP data export tools not suitable for providing data exports in ICGC compatible format and special external KIDREP plug-in module to provide such functionality has to be developed. Initial prototype of such module was provided already in 2011 together with the availability of data mapping specifications and later (by 2012) was redeveloped and more closely integrated into KIDREP by adapting it newly developed API that provides KIDREP plug-in mechanism. Regular updates have been made in this module afterwards to keep it up to date with the most recent specifications of KIDREP data mapping to ICGC data dictionaries.

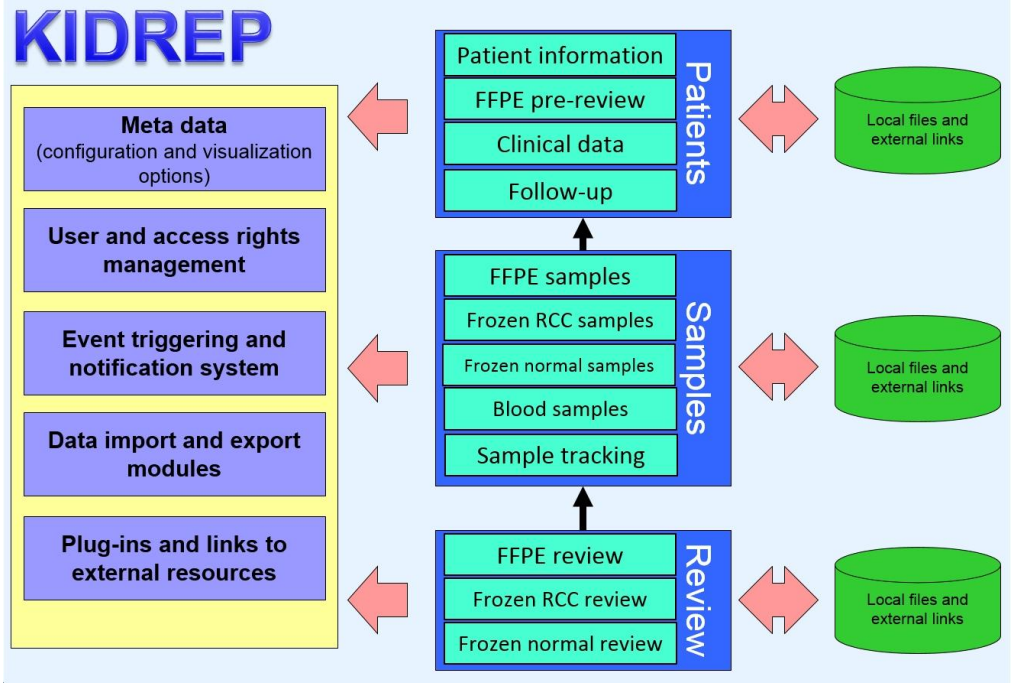
The practical approbation of data export to ICGC module started in 2013 when sequenced assay data for submission to ICGC become available. The first data set was successfully submitted to ICGC Release 15 on January 2014. Data about 122 samples were uploaded to ICGC, this included samples linked to approximately 90 DNA and 90 RNA sequencing datasets. Later two more submission updates have been made due to changes in ICGC requirements, with the latest update complying to requirements of the currently newest ICGC Release 18.

Bioinformatics tool development In addition to maintaining and updating of public and internal software for the CAGKID NGS and statistical analysis pipelines, we have undertaken development of novel software with applications beyond the project. Details of the basic pipelines and procedures that we are used are similar to other genomic analysis packages. Here, we highlight three examples of novel tool development arising from CAGEKID:

KIDREP The online web based software system – KIDREP was developed and was the basis of achieving all the work package objectives. The system successfully provided all the above

outlined functionality. As the overall project was progressing, the detailed requirements to the system were constantly updated (specifically import/export functionality with external software systems and ICGC portal), leading to regular updating of specifications and component redevelopment. At the end of the project KIDREP repository hosted data about 2079 patients, 2153 samples and 7873 pathology review forms. The system was developed as open source software and a generic version of KIDREP software, usable for many projects similar to CAGEKID, will be released to the community in 2015.

Fig. 4 Representation of the KIDREP web based software system



From the functionality perspective the main features of the KIDREP software system include:

- Online management of clinical and epidemiological information and follow-up data about treatments and disease progress of the patients;
- Integrated access to digital microscopy images of the samples and specimens;
- Online review forms for pathology review forms for FFPE and frozen samples;
- Automatic notification tools facilitating the monitoring of the review process and enabling rapid decision making (specifically regarding the suitability of samples for sequencing);
- Tracking tools to monitor sample location and conditions;
- Data import and export tools from in-house LIMS used by project partners;
- Tools for linking and integration with external software systems and data repositories.

The main functionality of KIDREP including links of the system with external environment is depicted in the Figure above.

The KIDREP system has the following features:

1. The prototype version of KIDREP repository developed on the basis of pre-existing software component SIMS (Sample Information Management System). Due to significantly different and broader functionality requirements the existing SIMS architecture was redesigned and the available functionality significantly extended. Notably more complex hierarchical structure of data entries was introduced, envisaging availability of different types of entries at each hierarchical level and extending the previous simple hierarchical structure to structure supporting hierarchies represented by arbitrary three level undirected graphs. Navigation facilities between different hierarchy levels have been expanded and tools and macro language for configuration of navigation behaviour have been developed. Also data import/export functionality and level of its configurability has been extended.

2. A more advanced user access rights management system was introduced, providing different access levels to each of the entries that can be configured either on the level of individual users or user groups.

3. The data filtering mechanisms support queries representing expressions in CNFs as well as provide an intuitive user-friendly interface for configuring them. Also management and storing of predefined filters and queries by individual users is supported.

4. Application Programmatic Interface (API) for communicating with external software systems (e.g. in-house LIMS, SlidePath digital microscopy image system) and external modules (data export to ICGC, email notification system has been developed) are incorporated.

5. A module of data querying from the Slidepath digital image repository was developed. The module connects Slidepath via web services interface and is used for automated import of links to digital images and snapshots from digital images created by pathologists during the review process. A number of modifications and updates were regularly made for this module to maintain compatibility with upgrades in SlidePath software.

6. An external module for data export to ICGC according to ICGC data dictionary specifications was developed. A number of regular updates have also been introduced to this module in response to frequent changes in ICGC specifications.

7. An external module for providing an email notifications to system users was developed and integrated within KIDREP software platform. The purpose of this module is to inform the designated system users about the availability of new samples and their microscopy images, sample assignments to reviewers, completion of sample reviews and changes in sample status.

Methylation analysis software A new methodology was developed for pre-processing and normalization of data from the Illumina Infinium Human Methylation 450K BeadChip that we used as part of CAGKID. This pipeline performs standard quality controls, elimination of unrelated signal variations and filtering of technical noise as well as a correction of the shift between Infinium I and II signals using an original subset quantile normalization approach. This has been widely used by groups outside of CAGEKID.

As the 450K Infinium technology was relatively new when adopted by CAGEKID, it posed a number of issues that need to be resolved at the level of analysis. The array consists of two subsets of markers that use different chemical assays to measure methylation. The first are Infinium I probes which are the bead types used in the previous 27K methylation array, and the second are Infinium II probes, which us a dual channel single-nucleotide primer extension with labeled dideoxynucleotides on the methylation variable position of a CpG. Issues of normalization and shift correction between the two assay formats have not yet been optimally solved. We also identified additional points that may interfere with the quality of the methylation data produced with the Illumina Infinium Human Methylation 450K BeadChip The crucial steps are 1) data quality control to estimate the quality of a dataset after data extraction, 2) probe filtering to eliminate signal variation unrelated to DNA methylation differences or unrelated to the biological context of the study, 3) signal correction for the adjustment of the color balance and background level correction as well as the 4) InfI/InfII shift correction and between sample normalization.

The Infinium II (InfII) probes are attached to a single type of beads and the methylation information is obtained through dual channel single-nucleotide primer extension with labeled dideoxynucleotides on the methylation variable position of a CpG. A shift in the density curves between InfI and InfII probes and a different dynamic behaviour of the two assays have previously been shown. Although a correction strategy has been devised this approach is based on a strong assumption about the bimodal shape of the methylation density profiles and when implementing the algorithm we found its efficiency to be sensitive to variations in the shape of the methylation density curves. We also identified additional points that may interfere with the quality of the methylation data produced with the Illumina Infinium Human Methylation 450K BeadChip The crucial steps are 1) data quality control to estimate the quality of a dataset after data extraction, 2) probe filtering and remapping to eliminate signal variation unrelated to DNA methylation differences or unrelated to the biological context of the study, 3) signal correction for the adjustment of the colour balance and background level correction as well as the 4) InfI/InfII shift correction and between sample normalization.

Pyrosequencing can be considered as one of the “gold standard” technologies for DNA methylation analysis due to its high quantitative precision and its ability to provide data with single nucleotide resolution. We therefore compared the CpG methylation values derived from the 450K arrays to the quantitative methylation values of the very same CpGs provided by the pyrosequencing technology. Among all tested normalization variants, the subset quantile normalization using the “relation to CpG” annotations to identify category related “anchors” provided the greatest number of closest methylation values (n=7) to those obtained by pyrosequencing for the very same CpG. This approach, together with the peak-based correction approach, provided also the smallest absolute differences in the methylation values when compared to the pyrosequencing-based methylation values.

A tool for integrated genomic/proteomic analyses Use of protein interaction network (referred to as interactome here) in combination with genomic data has emerged as a promising approach for integrative analysis, and for the identification of novel factors that are not captured by traditional pathway analysis methods. As part of CAGEKID we introduced a new approach, that

we call Cancer Genomics Network enrichment Analysis (CGNA), in which the whole interactome is divided into all of its constituting subnetworks that are treated independently in all further analyses. All proteins included in these subnetworks are labeled with abnormal patterns detected for their genes (e.g. mutations or DNA methylation) and/or mRNA (differential expression) and those surrounded by significantly high number of deregulated interacting partners are identified. This method is based on the assumption that proteins with significantly high number of deregulated interacting partners are likely important factors in the studied disease.

Briefly, to identify proteins whose interacting neighborhoods are deregulated in a certain group of samples, we applied hypergeometric modeling supplemented with 2-fold cross validation. Implementation of permutation-based statistics such as 2-fold cross validation or bootstrapping (available in the R package of CGNA) in network-based analysis is necessary in order to avoid low-confidence results. In our approach we used 2-fold cross validation because, in comparison to bootstrapping, in the context of network hypergeometric modeling both training and test sets change each time, and therefore the method is less prone to the biases that can originate from hypermutated or outlier samples.

We tested the performance of the method using data of The Cancer Genome Atlas breast cancer project, and used it to identify candidate driver aberrations in our renal cancer data. For example, only one protein (CDK1) was common in the list of factors identified through the same analysis applied to breast cancer and renal carcinoma datasets indicating that CGNA findings are not biased towards proteins with large number of interacting partners. Interestingly, CDK1 is shown to be an inhibitor of FOXO1, which is a known as common tumour suppressor in different types of human cancer. Consequently several proteins that were identified by analysis of the breast cancer datasets were supported by previous literature confirming the performance of CGNA. Accordingly, the additional factors identified in this study may present novel relevant players in breast cancer.

We then applied CGNA to study ccRCC, which is a malignancy that has not been investigated as extensively as breast cancer, with CAGEKID data. Our findings revealed an oncogenic core signaling which is activated in patients. Molecules involved in this network were members of VEGF signaling, the main driver of angiogenesis, which is a known target for development of therapies against ccRCC. However, resistance to anti-VEGF factors often develops in ccRCC patients, and interferes with the control of the disease. Our findings uncover novel factors of this family, which can serve as alternative targets for future drug development strategies.

Proteomic analyses Based on results of genomic analyses, we established a list of genes with mutations involved in epigenetic regulation, genes with mutations undergoing switch events (e.g. gain or loss of function), genes with mutations and expression differences in patients. A successful pilot study has been undertaken with 32 of the most promising targets. This work will be pursued by CAGEKID partners beyond the project end date.

Ethical and societal issues Compliance with ethical and legal rules was essential element of the project. Project documentation (information, consent, ethical approvals) was assessed and revised as appropriate to take into account changing international guidelines. Consent was

revised according to the current ICGC policies, therefore harmonizing the activities. We clarified the procedures and responsibilities for the use of CAGEKID samples for other researches than CAGEKID and possible transfer in 3rd countries outside CAGEKID. The authorization of the General Assembly of CAGEKID was obtained. A demonstration of the practical impact of our work in this respect is the recent evolutions in ethical issues management with respect to IARC samples collected in Moscow, Brno, Prague and Bucharest. The consent forms used to initially collect these samples were amended so as to include additional information about whole genome sequencing. This required the coordinated action of IARC and local ethics committees. Moreover, a waiver was obtained from local ethics committees when it was judged impractical or unnecessary to send additional information to patients following the guidelines that we had established. For samples coming from the University of Leeds, a new Material Transfer Agreement was prepared. We also prepared a document providing the legal and ethical rules to be respected for the transfer of samples and data from France to Canada, as this was a country considered and a data transfer agreement for the transfer of data from France to Canada.

A report on "Guidelines and ethical recommendations for a governance model, coherent with existing ethical frameworks and governance models applying to cohorts of patients included in CAGEKID", was circulated to partners, and presented as a poster at the ICGC meeting. An analysis of the proposed new regulation (January 25, 2012) for replacing the present Data Protection Directive was done and a comparative Table for aspects relevant to research and to genetic data prepared, and an update bibliography on ethics of sequencing in research and transfer to clinics.

Another important goal of our activities was coordination of the ethics component of CAGEKID not only with ICGC, as mentioned above, but also with other relevant activities. Anne Cambon-Thomsen from Partner 11 (INSERM) is member of the Ethics and policy committee (EPC) of ICGC and of the IDAC (International Data Access Committee) and participates regularly in the teleconferences and annual meetings of these committees and to the work performed. This allowed cross-fertilisation between ICGC policies and CAGEKID research. At the 2012 ICGC consortium a work on surveying actions in the various ICGC countries on how patients views are expressed and taken into consideration has been proposed by A. Cambon-Thomsen and is especially relevant for 3rd year objectives of CAGEKID. This close involvement in ICGC allows timely information of the CAGEKID consortium. Coordination has also been undertaken with two other initiatives connected to CAGEKID: GEUVADIS, a coordination action on sequencing technologies in healthy and diseased individuals and the new infrastructure of sequencing/genotyping (ESGI).

We also undertook assessment of the needs for education in the domain on ELSI aspects. This has started from the questions posed and the discussions in the consortium as well as in other forums. For example, a summer school on health law and bioethics has been organized by Partner 11 (INSERM) July 4-7 2011 in Toulouse and young investigators of CAGEKID were invited to attend it as the theme of this summer school was genetics and especially a full session is devoted to high throughput techniques and especially sequencing. This summer school was supported by the EU projects GEUVADIS and TECHGENE.

Training CAGEKID provided training to the wide community of researchers, clinicians and other people interested in cancer genomic studies by engaging in workshops, in several instances with other European projects, with multiple formats. Some of the training materials established through CAGEKID are now used for training courses at EBI and elsewhere.

4. POTENTIAL IMPACT

CAGEKID has brought together clinical and epidemiological resources together with the necessary genetics and genomics expertise across Europe to make a major contribution to international efforts to decipher the cancer genome. We have assembled the most important resources existing worldwide for genomic studies of renal cell carcinoma, which is of particular interest as a public health issue within Europe. Our programme provides the most complete and systematic analysis of this tumour site, incorporating both conventional and non-conventional histological subtypes. These data provides new insights into disease aetiology with applications for diagnosis and treatment. The results address presently unmet needs for biological markers in renal cancer, which is one of the few cancer sites in which such markers are not yet available for clinical use. Adopting a European-wide approach furnished us with ability to obtain exceptional clinical collections and biological resources for these studies, which are not available within any single country. This provided us with a valuable opportunity to examine differences in disease patterns across Europe, and the relationship of the variable incidence of the disease. CAGEKID addressed the specific issue of the FP7 call under which it was funded by structuring EU participation in the ICGC, where we will contribute the principal data on this cancer site and make the primary data available to the scientific community, and by establishing norms for the manipulation and storage of biological samples.

Ours is the first systematic exploration of human tumours from different backgrounds for specific DNA mutation profiles to identify previously unsuspected exposures associated with cancer. Exposure to aristolochic acid (AA) has previously been shown to lead to aristolochic acid nephropathy (AAN) characterized by chronic renal disease. AAN occurs in parts of the Balkans through ingestion of wheat flour contaminated with seeds of *Aristolochia clematitis*, as well as in parts of Asia through widespread use of herbal remedies from plants of the same family. Some very rare transitional cell carcinomas of the upper urinary tract are also known to be associated with AA exposure, and P53 mutations exhibit similar patterns to those we describe for the much more frequent ccRCC. Collaborators at IARC established experimentally that mouse cultured fibroblasts exposed to AA also exhibit a global pattern of A:T>T:A mutations with similar specific sequence context and strand bias. Taken together, our results demonstrate that AA is a likely carcinogen in ccRCC tumourigenesis, and that impact of exposure to AA may be far broader than previously thought, extending beyond known regions of endemic AAN. Indeed, in Romania and perhaps neighbouring central European countries (based on unpublished data), it seems that most cases of renal cancer could be related to AA exposure. This will have important consequences for public health and potential for disease prevention.

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). In 2007, by order of The Romanian Ministry of Health, eight regional cancer registries were initiated to cover the following regions: North-West, North-East, South-East, South Muntenia, South-West Oltenia, West, Central, and Bucharest. These registries 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 (Dana Mates, personal communication). In parts of Asia, AAN occurs through widespread use of AA-containing herbal remedies. 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 – appeared to be unintentional. The so-called Balkan endemic nephropathy (BEN) is thought instead to be due to the ingestion of wheat flour contaminated with seeds of Aristolochia clematitis. 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. It is remarkable that the kidney cancer series from Romania, out of four European countries (including three Eastern European countries) in our study, strongly showed the AA signature, while there was no evidence of any specific signature in the other countries included in the primary data from CAGEKID (we are now investigating others from the BEN affected region). Yet, the hospital in Bucharest where cases were recruited does cover the population of the BEN area. BEN patients are usually hospitalized in Timisoara or Craiova, two cities located in the Western and South Western part of Romania (Dana Mates and Viorel Jinga, personal communication). Additional studies are urgently needed to investigate the potential routes of exposure to AA in Romania and neighbouring countries because of the possibility of intervention.

The wide dissemination of the genomic project results will provide an important contribution to future research in renal cancer, and will represent one of the major cancer sites contributing to the international effort in cancer genomics. The archiving and diffusion of these data through the ICGC and the European Bioinformatic Institute will assure wide diffusion of the datasets and their integration with similar data generated for other cancer sites. Comparisons of the results between cancer sites will aid in obtaining an understanding of the similarities and differences between biological mechanisms underlying these diseases. The clinical and biological data collected on the large number of RCC samples included in this study has the power to greatly impact renal cancer in particular through the identification new disease markers and their correlation with clinical and other parameters of disease. CAGEKID is also preparing for a next step in the systematic development of resources for cancer studies by integrating a pilot programme to obtain antibodies and characterise antibodies for genomically defined targets emerging from our studies. In keeping with goal of resource development, this activity will call upon the platforms and expertise established at the Human Protein Atlas in Sweden and the reagents will be distributed to the scientific community through the infrastructure that is already financed there.

An additional and important contribution of the project has been in the development and diffusion of methodologies and bioinformatics tools for studies involving whole-genome

resequencing, and related techniques. Moreover, CAGEKID will contribute to developing, implementing and testing archiving and analysis tools for managing large datasets based around many applications of 2nd generation sequencing at EBI.

In addition to their diffusion through EBI, ICGC and Human Protein Atlas, results from the CAGEKID have been disseminated in the form of publications in international peer-reviewed journals, contributions to symposia, annual reports to the Commission, and uploaded to the information and communication platform external portal. This has been configured and launched at the start of the project and maintained from the Management Office. The portal has served to communicate research progress, consortium events (meetings, workshops) and publications to the scientific community. The portal will also mirror other research activities in genomics technologies in order to increase the interactivity between the various scientific and applied fields.

Members of the project have organised workshops, symposia and training courses for external dissemination and discussion of results. These actions have been designed to provide training in the methodologies that will be developed and applied within CAGEKID, which have wide application for research in cancer on other disease. Communication with potential stake-holders (such as patient associations) and dialogue with the public is important for informing them of developments in this area, particularly important in view of the high degree of public concern about ethical and societal matters related to developments in the human genome. As well as training aspects of this project, we have therefore engaged outreach and educational efforts to intensify the dialogue with a wider public.

Protection of intellectual property associated validated biological markers of renal cell carcinoma has been considered but no action taken as the principal objective of the project has been to contribute data under the rules governing the ICGC. All primary data have been made available to the scientific community with strict respect to ICGC policy.

5. PROJECT WEBSITE

To December 15, 2014: http://www.cng.fr/cagekid

From December 15, 2014: http://www.cagekid.org/

USE AND DISSEMENTATION OF FOREGROUND

SECTION A DISSEMINATION MEASURES

The principal publication summarising results CAGEKID appeared in October 2014 (Scelo et al. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. Nat Commun. 2014, 5:5135). Other publications and communications are listed in Templates A1 and A2.

The plan for future use and dissemination of the results has the following components:

1. The primary data from CAGEKID form part of the ICGC project, and are available through the ICGC portal. The data have been contributed to a pan-cancer genome analysis that is being conducted under the auspices of ICGC.

2. Following the publication of the primary CAGEKID paper, several other manuscripts will now be submitted describing additional results and other aspects of the project. Manuscripts that are presently in preparation include new results on the prevalence and distribution of AA exposure patterns in central Europe, results follow-up sequencing studies in the total CAGEKID collection, new methodology for integrated genomic/proteomic analysis, results from screening of non-conventional ccRCC, and a paper describing the KIDREP software.

3. Distribution of the KIDREP software package is planned for 2015.

SECTION B EXPLOITABLE FOREGROUND AND PLANS FOR EXPLOITATION

1. The genomic analysis has identified an environmental exposure to aristolochic acid (AA) that may have a large impact on ccRCC in regions of central Europe. New epidemiological studies will be designed and implemented to determine the exact origin and frequency of exposure and to determine if the exposure that causes the mutation pattern is also causative of the cancer. These additional studies are urgently needed because of the possibility of intervention.
2. Additional biological investigations are planned to understand the mechanisms underlying RCC in general, and the effect of AA exposure in different mutational backgrounds in particular.
3. These investigations include new proteomic studies issuing from the efforts to generate antibodies and other investigative tools in the CAGEKID project.
4. Testing of clinical applications of sequencing for classification of patients based on patterns of genomic variation are underway. Blood-based methylation assays will be tested for effectiveness in diagnosis based on results from the non-conventional RCC.
5. To achieve a yet wider characterisation of RCC, discussions have progressed to extend the project by inclusion of additional samples and clinical groups from east and central European countries, the UK, Canada, and elsewhere, and to involve other potential partners in the studies.

ANNEX

SECTION A

This section includes two templates

- Template A1: List of all scientific (peer reviewed) publications relating to the foreground of the project.
- Template A2: List of all dissemination activities (publications, conferences, workshops, web sites/applications, press releases, flyers, articles published in the popular press, videos, media briefings, presentations, exhibitions, thesis, interviews, films, TV clips, posters).

These tables are cumulative, which means that they should always show all publications and activities from the beginning until after the end of the project. Updates are possible at any time.

Template A1: List of Scientific (Peer Reviewed) Publications

TEMPLATE A1: LIST OF SCIENTIFIC (PEER REVIEWED) PUBLICATIONS, STARTING WITH THE MOST IMPORTANT ONES									
NO.	Title	Main author	Title of the periodical or the series	Number, date or frequency	Publisher	Place of Publication	Relevant pages	Permanent identifiers (if available)	Is/Will open access be provided to this publication
1	Variation in genomic landscape of clear cell renal cell carcinoma across Europe	Scelo G et al.	Nat Commun	5, 2014	Nature	London	5135	PMID: 25351205	No
2	Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation.	Touleimat N and Tost J	Epigenomics	4, 2012	FSG	London	325-41		Yes
3	ELSI 2.0 for Genomics and Society	Kaye J et al.	Science	336, 2012	AAAS	New York	673-4		Yes
4	Disclosing Results to Genomic Research Participants: Differences That Matter	Balsimme A et al.	The American Journal of Bioethics	12, 2012	Taylor & Francis Group	London	20-2		Yes
5	Renal cancer biomarkers: the promise of personalized care.	Vasudev N et al.	BMC Medicine	10, 2012	Biomed Central	London	112 onwards	PMID: 23016578	Yes

Template A2: List of Dissemination Activities

TEMPLATE A2: LIST OF DISSEMINATION ACTIVITIES								
NO.	Type of activities	Main leader	Title	Date/ Period	Place	Type of audience	Size of audience	Countries addressed
1	Poster	IMCS	KIDREP: a system for cancer sample information management in collaborative research projects.	20-22.3.2012	Cannes, France	6th Scientific Workshop of International Cancer Genome Consortium		Worldwide
2	Poster	INSERM	Coping with diversity in consent forms: example of the CAGEKID project.	20-22.3.2012	Cannes, France	6th Scientific Workshop of International Cancer Genome Consortium		Worldwide
3	Poster	INSERM	Coping with ethical novelty in CAGEKID.	20-22.3.2012	Cannes, France	6th Scientific Workshop of International Cancer Genome Consortium		Worldwide
4	Oral presentation	INSERM	Professional and family attitudes regarding large scale genetic information generated through next generation sequencing in research.	20.6.2012	London, UK	ESRC Genomics Network Conference 2012 - Genomics in Society: Facts, Fictions and Cultures		Worldwide
5	Invited lecture	INSERM	Public health genomics: Bringing genomics to society	23-24.4.2012	Hinxton, UK	Translational genomics pipeline: from populations to individuals – summer school held at the Wellcome Trust Genome Campus		Worldwide
6	Invited lecture	INSERM	Informed consent and sharing system for biobank	23-26.6.2012	Shangai, China	2nd annual conference on biobanking		Worldwide
7	Poster	INSERM	From phenotype driven medicine to data driven medicine: ethical challenges in genetics	20.6.2012	Rotterdam, Netherlands	11th world congress of bioethics. Rotterdam		Worldwide
8	Invited lecture	INSERM	The network of influences between ethics, law and practices in regulation of genomic information: a 10 years overview	27-28.5.2010	Amsterdam	Ten years after: mapping the societal landscape of genomics		Worldwide
9	Abstract	INSERM	Professional and publics attitudes regarding large scale genetic information generated in research	6-8.7.2012	Rotterdam, Netherlands	11th world congress of bioethics. Rotterdam		Worldwide
10	Oral presentation	INSERM	Post-genomics cancer research: trajectories of convergence and innovation	26-29.6.2012	Copenhagen, Denmark	Society for the Social Study of Science annual meeting		Worldwide
11	Invited lecture	INSERM	Ethical aspects of personalized medicine: any new issue?	26-29.6.2012	Tampere, Finland	8th FinBioNet Symposium, revolutionary biosciences: from advanced technologies to personalized medicine		Worldwide
12	Workshop	CNG	4th Paris Workshop on Genomic Epidemiology	18.10.2012	Paris	Scientific	200	Worldwide
13	Workshop	M. Lathrop	5th Paris Workshop on Genomic Epidemiology	2-3.10.2012	Paris	Scientific	200	Worldwide
14	Summit meeting	M. Lathrop	P3G Privacy Summit: "Data Sharing, Cloud Computing and Privacy"	2-4.5.2013	Paris	Policy makers & scientists	200	Worldwide
15	Training workshop	EBI	CAGEKID Cancer Genomics Workshop	3.5.2013	Hixton	PhD students and post-doctoral researchers	40	European

TEMPLATE A2: LIST OF DISSEMINATION ACTIVITIES – Cont'd

NO.	Type of activities	Main leader	Title	Date/ Period	Place	Type of audience	Size of audience	Countries addressed
16	Workshop	Leeds	Renal Cancer Workshop	18-22.3.2013	York	Scientists & clinicians	25	European
17	Workshop	Leeds	Renal Cancer Workshop	16-17.6.2014	York	Scientists & clinicians	25	European
18	Training workshop	EBI	RNA-seq workshop	4-6.12.2012	Hixton	PhD students and post-doctoral researchers	40	European
19	Poster	CEA	Identification of non-coding mutations on regulatory elements in clear cell renal carcinoma		Cambridge, UK	Epigenomics of Common Diseases, a Wellcome Trust conference	300	Worldwide
20	Invited lecture	Uni of Leeds	Biomarkers of kidney cancer		Manchester, UK	Renal Cancer North Regional Meeting		Northern England
21	Invited lecture	Uni of Leeds	Genetic changes in kidney cancer: initial results from the EU CAGEKID study		Leeds, UK	St James's Institute of Oncology-departmental meeting		Leeds, UK
22	Invited update	Uni of Leeds	Update on current renal cancer biomarker studies		London, UK	National Cancer Research Institute Renal Clinical Studies Group Meeting		UK
23	Invited lecture	Uni of Leeds	Biomarker discovery: new emerging science		Leeds, UK	ECMC UK Nurse Educational Session		UK
24	Invited lecture	INSERM	The world of sharing: Feedback of what, to whom and how?	20-23.9-2010	Oxford, UK	The international data sharing conference		Worldwide
25	Presentation	INSERM	Implications of next generation sequencing for clinical practice - A debate & Legal regulation for genetic testing	Jun-10	Göteborg, Sweden	ESHG Conference		Worldwide
26	Invited lecture	INSERM	Translating proteomic biomarkers into clinic: new ethical issues or different perspectives on classical ethical dilemma?	18-19.3.2010	Hinxton, UK	Translating clinical proteomics into clinical practice, Welcome Trust Genome Campus		Worldwide
27	Invited lecture	INSERM	Informed consent and its actors: puzzle, maze or Babel tower?	14-16.6.2010	Uppsala, Sweden	Is Medical Ethics Really in the Best Interest of the Patient?, Medical Ethics Conference		Worldwide
28	Presentation	INSERM	Privacy in the context of high throughput technologies in genetics	4.3.2011	Paris, France	UNESCO Tenth Meeting of the UN Interagency Committee on Bioethics "Genetic privacy and non-discrimination"		Worldwide

SECTION B (Confidential or public: confidential information to be marked clearly)

Part B1

The applications for patents, trademarks, registered designs, etc. shall be listed according to the template B1 provided hereafter.

The list should, specify at least one unique identifier e.g. European Patent application reference. For patent applications, only if applicable, contributions to standards should be specified. This table is cumulative, which means that it should always show all applications from the beginning until after the end of the project.

TEMPLATE B1: LIST OF APPLICATIONS FOR PATENTS, TRADEMARKS, REGISTERED DESIGNS, ETC.					
Type of IP Rights[18]:	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	Application reference(s) (e.g. EP123456)	Subject or title of application	Applicant (s) (as on the application)
Not Applicable					

Part B2

Please complete the table hereafter:

Type of Exploitable Foreground[19]	Description of exploitable foreground	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application[20]	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
	Ex: New supercond uctive Nb-Ti alloy			*MRI equipment*	*1. Medical 2. Industrial inspection*	*2008 2010*	*A materials patent is planned for 2006*	*Beneficiary X (owner) Beneficiary Y, Beneficiary Z, Poss. licensing to equipment manuf. ABC*
NOT APPLICABLE								

In addition to the table, please provide a text to explain the exploitable foreground, in particular:

- Its purpose
- How the foreground might be exploited, when and by whom
- IPR exploitable measures taken or intended
- Further research necessary, if any
- Potential/expected impact (quantify where possible)

REPORT ON SOCIETAL IMPLICATIONS

Replies to the following questions will assist the Commission to obtain statistics and indicators on societal and socio-economic issues addressed by projects. The questions are arranged in a number of key themes. As well as producing certain statistics, the replies will also help identify those projects that have shown a real engagement with wider societal issues, and thereby identify interesting approaches to these issues and best practices. The replies for individual projects will not be made public.

A. General Information *(completed automatically when **Grant Agreement number** is entered).*

Grant Agreement Number:	241669
Title of Project:	Cancer Genomics of the Kidney
Name and Title of Coordinator:	Mark Lathrop – Scientific Director

B. Ethics

1. Did your project undergo an Ethics Review (and/or Screening)? • If Yes: have you described the progress of compliance with the relevant Ethics Review/Screening Requirements in the frame of the periodic/final project reports? Special Reminder: the progress of compliance with the Ethics Review/Screening Requirements should be described in the Period/Final Project Reports under the Section 3.2.2 *'Work Progress and Achievements'*	⊗**Yes** O **No**
2. Please indicate whether your project involved any of the following issues (tick box) :	***YES***
RESEARCH ON HUMANS	
• Did the project involve children?	NO
• Did the project involve patients?	YES
• Did the project involve persons not able to give consent?	NO
• Did the project involve adult healthy volunteers?	NO
• Did the project involve Human genetic material?	YES
• Did the project involve Human biological samples?	YES
• Did the project involve Human data collection?	YES
RESEARCH ON HUMAN EMBRYO/FOETUS	
• Did the project involve Human Embryos?	NO
• Did the project involve Human Foetal Tissue / Cells?	NO
• Did the project involve Human Embryonic Stem Cells (hESCs)?	NO
• Did the project on human Embryonic Stem Cells involve cells in culture?	NO
• Did the project on human Embryonic Stem Cells involve the derivation of cells from Embryos?	NO
PRIVACY	

• Did the project involve processing of genetic information or personal data (eg. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?	YES
• Did the project involve tracking the location or observation of people?	NO
RESEARCH ON ANIMALS	
• Did the project involve research on animals?	NO
• Were those animals transgenic small laboratory animals?	NO
• Were those animals transgenic farm animals?	NO
• Were those animals cloned farm animals?	NO
• Were those animals non-human primates?	NO
RESEARCH INVOLVING DEVELOPING COUNTRIES	
• Did the project involve the use of local resources (genetic, animal, plant etc)?	NO
• Was the project of benefit to local community (capacity building, access to healthcare, education etc)?	NO
DUAL USE	
• Research having direct military use	O Yes ⊗ No
• Research having the potential for terrorist abuse	NO

C. Workforce Statistics

3. Workforce statistics for the project: Please indicate in the table below the number of people who worked on the project (on a headcount basis).

Type of Position	Number of Women	Number of Men
Scientific Coordinator		1
Work package leaders	4	3
Experienced researchers (i.e. PhD holders)		
PhD Students		
Other		

4. How many additional researchers (in companies and universities) were recruited specifically for this project?	
Of which, indicate the number of men:	

D. Gender Aspects

5. Did you carry out specific Gender Equality Actions under the project?	o	Yes
	⊗	No

6. Which of the following actions did you carry out and how effective were they?

	Not at all effective				Very effective
□ Design and implement an equal opportunity policy	O	O	O	O	O
□ Set targets to achieve a gender balance in the workforce	O	O	O	O	O
□ Organise conferences and workshops on gender	O	O	O	O	O
□ Actions to improve work-life balance	O	O	O	O	O

○ Other:

7. Was there a gender dimension associated with the research content – i.e. wherever people were the focus of the research as, for example, consumers, users, patients or in trials, was the issue of gender considered and addressed?

⊗ Yes- please specify: Gender specific genetic profiles examined

○ No

E. Synergies with Science Education

8. Did your project involve working with students and/or school pupils (e.g. open days, participation in science festivals and events, prizes/competitions or joint projects)?

○ Yes- please specify

⊗ No

9. Did the project generate any science education material (e.g. kits, websites, explanatory booklets, DVDs)?

○ Yes- please specify

⊗ No

F. Interdisciplinarity

10. Which disciplines (see list below) are involved in your project?

⊗ Main discipline[21] : **3.1**

⊗ Associated discipline[21] **:5.4** | o | Associated discipline[21] :

G. Engaging with Civil society and policy makers

11a. Did your project engage with societal actors beyond the research community? *(if 'No', go to Question 14)*	⊗	Yes
	o	No

11b. If yes, did you engage with citizens (citizens' panels / juries) or organised civil society (NGOs, patients' groups etc.)?

- ○ No
- ○ Yes- in determining what research should be performed
- ○ Yes - in implementing the research
- ⊗ Yes, in communicating /disseminating / using the results of the project

11c. In doing so, did your project involve actors whose role is mainly to organise the dialogue with citizens and organised civil society (e.g. professional mediator; communication company, science museums)?	O ⊗	Yes No

12. Did you engage with government / public bodies or policy makers (including international organisations)

- O No
- O Yes- in framing the research agenda
- O Yes - in implementing the research agenda
- ⊗ Yes, in communicating /disseminating / using the results of the project

13a. Will the project generate outputs (expertise or scientific advice) which could be used by policy makers?

- ⊗ Yes – as a primary objective (please indicate areas below- multiple answers possible)
- O Yes – as a secondary objective (please indicate areas below - multiple answer possible)
- O No

13b. If Yes, in which fields?

Agriculture		Energy		Human rights	
Audiovisual and Media		Enlargement		Information Society	
Budget		Enterprise		Institutional affairs	
Competition		Environment		Internal Market	
Consumers		External Relations		Justice, freedom and security	
Culture		External Trade		Public Health	✓
Customs		Fisheries and Maritime Affairs		Regional Policy	
Development Economic and Monetary Affairs		Food Safety		Research and Innovation	
Education, Training, Youth		Foreign and Security Policy		Space	
Employment and Social Affairs		Fraud		Taxation	
		Humanitarian aid		Transport	

13c. If Yes, at which level?

- ○ Local / regional levels
- ○ National level
- ○ European level
- ⊗ International level

H. Use and dissemination

14. How many Articles were published/accepted for publication in peer-reviewed journals?	5
To how many of these is open access[22] provided?	4
How many of these are published in open access journals?	
How many of these are published in open repositories?	
To how many of these is open access not provided?	1
Please check all applicable reasons for not providing open access:	
☑ publisher's licensing agreement would not permit publishing in a repository ☐ no suitable repository available ☐ no suitable open access journal available ☐ no funds available to publish in an open access journal ☐ lack of time and resources ☐ lack of information on open access ☐ other[23]: ……	

15. How many new patent applications ('priority filings') have been made? *("Technologically unique": multiple applications for the same invention in different jurisdictions should be counted as just one application of grant).*		0
16. Indicate how many of the following Intellectual Property Rights were applied for (give number in each box).	Trademark	0
	Registered design	0
	Other	0
17. How many spin-off companies were created / are planned as a direct result of the project?		0
Indicate the approximate number of additional jobs in these companies:		

18. Please indicate whether your project has a potential impact on employment, in comparison with the situation before your project:

☐ Increase in employment, or	☐ In small & medium-sized enterprises
☐ Safeguard employment, or	☐ In large companies
☐ Decrease in employment,	☐ None of the above / not relevant to the project
☑ Difficult to estimate / not possible to quantify	

19. For your project partnership please estimate the employment effect resulting directly from your participation in Full Time Equivalent (*FTE = one person working fulltime for a year*) jobs:	*Indicate figure:*
Difficult to estimate / not possible to quantify	☑

I. Media and Communication to the general public

20. As part of the project, were any of the beneficiaries professionals in communication or media relations?

○ Yes ⊗ No

21. As part of the project, have any beneficiaries received professional media / communication training / advice to improve communication with the general public?

○ Yes ⊗ No

22. Which of the following have been used to communicate information about your project to the general public, or have resulted from your project?

☑ Press Release	☑ Coverage in specialist press
☑ Media briefing	☐ Coverage in general (non-specialist) press
☐ TV coverage / report	☐ Coverage in national press
☐ Radio coverage / report	☑ Coverage in international press
☐ Brochures /posters / flyers	☐ Website for the general public / internet
☐ DVD /Film /Multimedia	☐ Event targeting general public (festival, conference, exhibition, science café)

23. In which languages are the information products for the general public produced?

☐ Language of the coordinator	☑ English
☐ Other language(s)	