Whole-exome sequencing identifies somatic mutations associated with mortality in metastatic clear cell kidney cancer carcinoma

Alejandro Mendoza-Alvarez[1], Beatriz Guillen-Guio[1], Adrian Baez-Ortega, Carolina Hernandez-Perez[2], Sita Lakhwani-Lakhwani[1], Maria-del-Carmen Maeso, Jose M. Lorenzo-Salazar[3], Manuel Morales[2], Carlos Flores[1, 3, 4*]

[1]Hospital Universitario Nuestra Señora de Candelaria, Spain, [2]Service d'Oncologie Médicale, Hospital Universitario Nuestra Señora de Candelaria, Spain, [3]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Spain, [4]Centro de Investigación Biomédica en Red (CIBER), Spain

Submitted to Journal:
Frontiers in Genetics

Specialty Section:
Cancer Genetics

Article type:
Original Research Article

Manuscript ID:
429604

Received on:
08 Oct 2018

Frontiers website link:
www.frontiersin.org

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Author contribution statement

ABO, AMA, BGG, and CF wrote the manuscript and the supplementary files. BGG and SLL performed DNA extractions and quantifications, as well as the preparation of samples for sequencing. CHP, MDCM, and MM performed data collection and the record of patient and sample features. ABO and JMLS conceived and implemented software procedures. AMA, CF and JMLS performed the statistical tests on detected somatic variants. CF provided a general supervision of the project, giving guidelines for each step. All the authors provided insights, corrections and approved the final version of the manuscript.

Keywords

ccRCC, Whole-exome sequencing, Kidney cancer, somatic mutation, Mortality

Abstract

Word count: 298

Clear cell renal cell carcinoma (ccRCC) is among the most aggressive histologic subtypes of kidney cancer, representing about 3% of all human cancers. Patients at stage IV have nearly 60% of mortality in 2–3 years after diagnosis. To date, most ccRCC studies have used DNA microarrays and targeted sequencing of a small set of well-established, commonly altered genes. Nonetheless, whole exome-sequencing (WES) has presently become the methodology of choice for the effective analysis of pathogenic coding genetic variation while maintaining clinical utility. Applying WES to simultaneously interrogate virtually all exons in the human genome for somatic variation, here we analyzed the burden of coding somatic mutations in metastatic ccRCC primary tumors, and its association with patient mortality from cancer, in patients who received VEGF receptor-targeting drugs as the first-line therapy. To this end, we sequenced the exomes of ten tumor–normal pairs of ccRCC patient tissues from primary biopsies at >100× mean depth and called somatic coding variation. Mutation burden analysis prioritized 138 genes displaying nominal associations with patient mortality. A gene set enrichment analysis evidenced strong statistical support for the abundance of genes involved in the development of kidney cancer (p=2.31×10–9) and carcinoma (p=1.22×10–5), with 49 genes having direct links with kidney cancer according to the published records. Three mutational signatures were found to be operative in the tumor exomes, one of which (COSMIC signature 12) has not been previously reported in ccRCC. Selection analysis yielded no detectable evidence of overall positive or negative selection, with the exome-wide number of nonsynonymous substitutions per synonymous site reflecting largely neutral tumor evolution. Taken together, our results provide evidence for a set of candidate genes in which somatic mutation burden is tentatively associated with patient mortality in metastatic ccRCC, offering new potential pharmacological targets and a basis for further validation studies.

Funding statement

This study was funded by Fundación CajaCanarias (SALUCAN11). AMA was supported by a fellowship from CajaSiete-ULL. BGG was supported by a fellowship from the Canarian Agency for Research, Innovation and Information Society (ACIISI, grant number TESIS2015010057) co-funded by the European Social Fund (ESF).

Ethics statements

(Authors are required to state the ethical considerations of their study in the manuscript, including for cases where the study was exempt from ethical approval procedures)

Does the study presented in the manuscript involve human or animal subjects: Yes

Please provide the complete ethics statement for your manuscript. Note that the statement will be directly added to the manuscript file for peer-review, and should include the following information:

- Full name of the ethics committee that approved the study
- Consent procedure used for human participants or for animal owners
- Any additional considerations of the study in cases where vulnerable populations were involved, for example minors, persons with disabilities or endangered animal species

As per the Frontiers authors guidelines, you are required to use the following format for statements involving human subjects:
This study was carried out in accordance with the recommendations of [name of guidelines], [name of committee]. The protocol was approved by the [name of committee]. All subjects gave written informed consent in accordance with the Declaration of Helsinki.
For statements involving animal subjects, please use:
This study was carried out in accordance with the recommendations of 'name of guidelines, name of committee'. The protocol was approved by the 'name of committee'.

If the study was exempt from one or more of the above requirements, please provide a statement with the reason for the exemption(s).
Ensure that your statement is phrased in a complete way, with clear and concise sentences.

This study was carried out in accordance with the recommendations of the Ethics Committee for Clinical Research from the Hospital Universitario Nuestra Señora de Candelaria with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics Committee for Clinical Research from the Hospital Universitario Nuestra Señora de Candelaria.

Data availability statement

Generated Statement: The datasets generated for this study are available on request to the corresponding author.

Whole-exome sequencing identifies somatic mutations associated with mortality in metastatic clear cell kidney cancer carcinoma

Alejandro Mendoza-Alvarez[1], Beatriz Guillen-Guio[1], Adrian Baez-Ortega[2], Carolina Hernandez-Perez[3], Sita Lakhwani-Lakhwani[1], Maria-del-Carmen Maeso[4], Jose M. Lorenzo-Salazar[5], Manuel Morales[3], Carlos Flores[1,5,6*]

[1]Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain.

[2]Transmissible Cancer Group, Department of Veterinary Medicine, University of Cambridge, UK.

[3]Service of Medical Oncology, Hospital Universitario Nuestra Señora de Candelaria, Santa Cruz de Tenerife, Spain.

[4]Deparment of Pathology, Hospital Universitario Nuestra Señora de Candelaria, Santa Cruz de Tenerife, Spain.

[5]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain.

[6]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain.

*** Correspondence:**
Carlos Flores, PhD.
Unidad de Investigación
Hospital Universitario N.S. de Candelaria
Carretera del Rosario s/n
38010 Santa Cruz de Tenerife
Phone: (+34) 922-602938
Fax: (+34) 922-600562
e-mail: cflores@ull.edu.es

***Number of words:** 5851

***Number of figures:** 3 (+ 1 in the supplementary material)

***Number of tables:** 2 (both in the supplementary material)

Keywords: ccRCC, whole-exome sequencing, kidney cancer, somatic mutation, mortality.

Abstract

Clear cell renal cell carcinoma (ccRCC) is among the most aggressive histologic subtypes of kidney cancer, representing about 3% of all human cancers. Patients at stage IV have nearly 60% of mortality in 2–3 years after diagnosis. To date, most ccRCC studies have used DNA microarrays and targeted sequencing of a small set of well-established, commonly altered genes. Nonetheless, whole exome-sequencing (WES) has presently become the methodology of choice for the effective analysis of pathogenic coding genetic variation while maintaining clinical utility. Applying WES to simultaneously interrogate virtually all exons in the human genome for somatic variation, here we analyzed the burden of coding somatic mutations in metastatic ccRCC primary tumors, and its association with patient mortality from cancer, in patients who received VEGF receptor-targeting drugs as the first-line therapy. To this end, we sequenced the exomes of ten tumor–normal pairs of ccRCC patient tissues from primary biopsies at >100× mean depth and called somatic coding variation. Mutation burden analysis prioritized 138 genes displaying nominal associations with patient mortality. A gene set enrichment analysis evidenced strong statistical support for the abundance of genes involved in the development of kidney cancer (p=2.31×10^{-9}) and carcinoma (p=1.22×10^{-5}), with 49 genes having direct links with kidney cancer according to the published records. Three mutational signatures were found to be operative in the tumor exomes, one of which (COSMIC signature 12) has not been previously reported in ccRCC. Selection analysis yielded no detectable evidence of overall positive or negative selection, with the exome-wide number of nonsynonymous substitutions per synonymous site reflecting largely neutral tumor evolution. Taken together, our results provide evidence for a set of candidate genes in which somatic mutation burden is tentatively associated with patient mortality in metastatic ccRCC, offering new potential pharmacological targets and a basis for further validation studies.

1 Introduction

Clear cell renal cell carcinoma (ccRCC) represents only 2–3% of all human cancers (Manley et al., 2017). Notwithstanding, over 30% of ccRCC patients have metastases at the time of diagnosis, and 60% die in the first 2–3 years after diagnosis (Casuscelli et al., 2017). ccRCC is characterized by the resistance to radiation, cytotoxic and hormone therapies. Current treatments for ccRCC include diverse chemotherapeutic agents targeting the vascular endothelial growth factor (VEGF) pathway (Sternberg et al., 2010).

Roughly a decade ago, genetic approaches to disease diagnosis were postulated as a costly new way to progress towards the paradigm shift aimed by precision medicine. In ccRCC, the vast majority of studies have been directed at assessing genes that are known to be directly involved in pathogenesis, most of them using DNA arrays for genetic screening. The drawbacks and advantages of holistic vs. targeted gene studies have been extensively discussed in the literature (Iglesias et al., 2014; Kong et al., 2018). Nowadays, high-throughput next generation sequencing (NGS) technologies have made genetic testing affordable and cost-effective, hence consolidating as a central instrument for the progress towards the implementation of precision medicine. Furthermore, the reduction in per-base sequencing cost has popularized the use of whole-exome sequencing (WES) for the investigation of the pathogenic impact of genetic variation in coding regions (Damiati et al., 2016; Fay et al., 2016; Lata et al., 2018). This is reflected by the sheer number of WES studies being published, including a multitude of analyses of cancer exomes (Samarakoon et al., 2014; Lata et al., 2018).

To our knowledge, research in ccRCC using WES has previously focused on the treatment response or toxicity variables in relation to chemotherapeutic treatment. Moreover, kidney cancer studies were often limited to genes which are frequently altered in this condition, most commonly focusing on a gene panel conformed by *VHL*, *PBRM1*, *BAP1*, *SETD2*, *TP53*, *PTEN*, *KDM5C* and *TERT* genes (Casuscelli et al., 2017; Manley et al., 2017; Tennenbaum et al., 2017). Here, for the first time, we apply high-depth WES to assess the association between somatic mutation burden in metastatic ccRCC primary tumors and patient survival.

2 Material and methods

2.1 Patient population and setting

A total of 13 metastatic ccRCC patients (stage IV) from the two tertiary hospitals of Tenerife (Spain), Hospital Universitario Nuestra Señora de Candelaria (HUNSC) and Hospital Universitario de Canarias (HUC), were included in the study. The patients were aged 31–80 years old (mean age of 56 years), with a male percentage of 61.5%. Seven (54%) of these patients died of cancer-related causes during the course of the study. The study was approved by the HUNSC Ethics Committee and written informed consent was obtained from all patients.

Nephrectomies were performed with curative intentions in 6 patients. For the rest of individuals, surgery was performed with cytoreductive purposes (Flanigan et al., 2001; Mickisch et al., 2001). Patients were classified into prognosis groups according to the Heng scoring system (Heng et al., 2013). At the moment of the diagnosis of metastasis, 5 patients showed good prognosis, while 6 had an intermediate prognosis and 2 a bad prognosis. All patients received tyrosine kinase inhibitors of the VEGF pathway, namely pazopanib (Sternberg et al., 2010) or sunitinib (Motzer et al., 2013), as the first-line treatment, except for one patient with bad prognosis who received temsirolimus (Hudes et al., 2007) as first-line treatment and pazopanib as second-line treatment.

Formalin-fixed paraffin-embedded (FFPE) biopsies from the primary tumors were obtained in blocks for subsequent DNA extraction. After evaluation by a pathologist, hematoxylin-eosin stained tissues were used to determine the limits of tumoral tissues. Whenever possible, non-tumoral (thereafter referred to as normal) and tumoral tissues for DNA isolation were obtained from independent tissue slices. The GeneRead DNA FFPE Kit (QIAGEN, Hilden, Germany) was used for DNA isolation according to manufacturer's instructions. The integrity and concentration of DNA was evaluated with the Qubit® 3.0 Fluorometer, using the dsDNA BR Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA), and the TaqMan™ RNase P Detection Reagents Kit (Thermo Fisher Scientific).

2.2 Whole-exome sequencing

Genomic DNA was enriched for exome regions using the Ion AmpliSeq™ Exome RDY Kit (Thermo Fisher Scientific) and Ion PI™ Chip Kit v3 (Thermo Fisher Scientific). Exome-enriched DNA was sequenced on the Ion Proton™ platform (Thermo Fisher Scientific), with two exomes per run to attain a theoretical depth of 100× per sample. Sequence data were aligned to the hg19/GRCh37 human reference genome using the Torrent Mapping Alignment Program v.5.0.13 included in the Torrent Suite Software for Sequencing Data Analysis v.5.0.4 (Thermo Fisher Scientific).

2.3 Variant calling and annotation

Aligned sequence data were analyzed to identify somatic and germline single-nucleotide variants (SNVs) and small insertions and deletions (indels). We called genetic variation using a bespoke computational pipeline (**Supplementary Figure 1**) built on the variant caller Platypus v0.8.1 (Rimmer et al., 2014). As part of the pipeline, Platypus was run twice on each BAM file with two different settings: (i) default mode with additional options --minReads=3 and --minPosterior=0, (ii) default mode with options --minReads=3, --minPosterior=0, --minFlank=0 and --trimReadFlank=10. Variants (SNVs and indels) flagged with Platypus quality flags 'badReads', 'MQ', 'strandBias', 'SC' and 'QD' were subsequently discarded, and the remaining variants were merged into a single file and genotyped across each sample. Variants that continued to be flagged with 'badReads', 'MQ', 'strandBias', 'SC' and 'QD' during this genotyping were discarded. We then filtered germline variation and retained somatic variants for subsequent analyses. To that aim, we filtered out the variants that were present in any of the normal tissues, as well as the variants that were supported by less than 3 sequence reads. Remaining variants were considered somatic and annotated using the Ensembl Variant Effect Predictor (VEP) v91.0 (McLaren et al., 2016).

2.3.1 Mutation burden and selection analysis

We analyzed the annotated somatic variants in each gene using bespoke analysis routines coded in the R programming language (R Development Core Team, 2008). To test for associations between the mutation burden and patient mortality, a Fisher's exact test on the mutation count data was performed in R. Results were evaluated for inflation with a quantile-quantile (QQ) plot, using the qqplot v3.4.2 R package (Becker, Chambers, & Wilks, 1988), and by estimating lambda with GenABEL v1.8-0 (Aulchenko et al., 2007). To assess evidence of positive or negative selection on somatic substitutions and detect any potential germline contamination in the somatic variant set, the dNdScv v0.0.0.9 R package (Martincorena et al., 2017) was employed to estimate exome-wide and per-gene number ratios of nonsynonymous substitutions per synonymous site (dN/dS).

2.3.2 Mutational signature analysis

The sigfit v1.1.0 R package (Gori & Baez-Ortega, 2018) was used to identify mutational processes (Alexandrov & Stratton, 2014), by fitting the mutational signatures published in the COSMIC catalogue (https://cancer.sanger.ac.uk/cosmic/signatures) to the mutational profiles of the somatic SNVs in each tumor. The latter were obtained by classifying SNVs into 96 categories according to substitution type (interpreting the pyrimidine base in the Watson–Crick pair as the reference base) and the bases immediately 5' and 3' to the mutated base in the reference genome (Alexandrov & Stratton, 2014). Fitting of mutational signatures to somatic variants was initially performed using all 30 COSMIC signatures; subsequently, those signatures displaying significant activity and biological coherence were fitted again to obtain more-accurate signature activity estimates.

2.4 Gene set enrichment analysis

The mutational landscape of ccRCC was explored through gene set enrichment analysis (GSEA), which was performed on those genes with $p<0.05$ in the Fisher's exact test of mutation burden (described above). This was performed via the EnrichR tool (Chen et al., 2013; Kuleshov et al., 2016) focusing on disease links through the Jensen Diseases database, which compiles evidence of gene–disease associations through the analysis of existing literature on genetic studies.

3 Results

3.1 DNA extraction and sequencing

We extracted and quantified genetic material from the original 13 patient FFPE samples for further evaluation via qPCR amplification with TaqMan probes of the housekeeping gene RNAsaP. Three of the samples were discarded from the study due to insufficient amount of extracted DNA and high fragmentation levels, caused by the formalin fixation process. We subsequently sequenced 20 paired DNA samples, extracted from normal and tumor tissues from the remaining 10 patients. The average age of the sequenced individuals was 55 years (range 31–80 years), where 70% were male and 40% died during the course of the study. Amplicon size ranged between 157 and 182 base pairs (bp), with a mean insert length of 172 bp. The Ion AmpliSeq™ Exome RDY Kit yielded a median of 91.17% reads covering the on-target region with at least 20× depth. Sequencing metric summaries are shown in **Supplementary Table 1**.

3.2 Variant calling and annotation

A total of 122,019 SNVs and 31,646 indels were initially called by the variant calling pipeline. The elevated number of indels was likely due to characteristic sequencing errors at polynucleotide tracts, associated with the Ion Torrent sequencing chemistry (Fujita et al., 2017; Lata et al., 2018). A further categorization of all these variants into germline and somatic sets resulted in a total of 23,157 SNVs (18.98%) and 9 indels (0.28%) of somatic origin. Further filtering of somatic SNVs and indels, according to whether the alternate allele presented sufficient support across the tumor samples, resulted in a refined set of 9,220 (40%) high-confidence somatic SNVs, which were considered for subsequent analyses; all indels were filtered out at this stage. This figure agrees with previous results (Miao et al., 2018), confirming that ccRCC is among the cancer types with lowest somatic mutation prevalence. We then predicted the functional consequences of the somatic variants using the Ensembl Variant Effect Predictor (VEP) software (McLaren et al., 2016). The predictions indicated that, of the 9,189 SNVs categorized, 65% were missense variants, 31% were synonymous variants, and 4% were nonsense variants.

Finally, to evaluate the evidence for selection on somatic substitutions and identify any potential contamination from germline polymorphisms, exome-wide and per-gene estimates of the ratio of nonsynonymous substitutions per synonymous site (dN/dS) were obtained for the set of somatic variants using a dN/dS model optimized for the analysis of selection in cancer (Martincorena et al., 2017). Somatic variants identified in more than one tumor (n=464) were excluded from the analysis in order to avoid spurious inflation of dN/dS estimates. The analysis yielded an exome-wide dN/dS≈1, which is indicative of largely neutral evolution, in agreement with previous studies of selection in cancer (Martincorena et al., 2017). No genes were found to display detectable evidence of selection on missense or truncating substitutions.

3.3 Gene-based mutation burden and mortality by ccRCC

We conducted comparative analyses between surviving and non-surviving ccRCC patients, testing for differences per gene in the somatic mutation burden. We found a total of 5,267 genes with evidence of somatic variation among the 10 patients, where the most altered gene in terms of the number of mutations was *CDC27*, which harbored a total of 89 somatic variants. We then applied Fisher's exact test to evidence significant differences in the somatic burden and found 138 genes showing nominal significance (lowest $p=2.0\times10^{-6}$; **Supplementary Table 2**). A QQ-plot of the distribution of gene-based *p*-values nearly followed the null (**Figure 1**) suggesting a minimal lambda factor (1.071) and a minimal inflation of results. Interestingly, among the genes with strongest

statistical significance, we found a number of genes expressed in kidney tissues and previously associated with a variety of human malignancies of neoplastic and non-neoplastic origin, such as *GPR155* (ranked 1st), *INPP5K* (ranked 3rd), and *KRT7* (ranked 4th) (**Supplementary Table 2**). Another notable result was the presence in the list of various mucin-encoding genes (*MUC5B*, *MUC12*, and *MUC16*), which have been previously linked to colorectal, ovarian and hepatological cancers (Yin et al., 2013; Felder et al., 2014; Wang et al., 2018), as well as to severe fibrotic lung disorders (Seibold et al., 2011). In agreement with a previous targeted sequencing study (Tennenbaum et al., 2017), the mutation burden of *VHL*, which is the main hallmark of ccRCC, showed no differences between survivors and non-survivors, supporting its role only during early stages of tumorigenesis (Mandriota et al., 2002; Mitchell et al., 2018).

3.4 Gene set enrichment analysis and mutational signatures

An enrichment analysis focused on the set of 138 genes having nominally significant differences for somatic mutation burden between survivors and non-survivors was performed to reveal disease links according to the Jensen Diseases database. Those genes most likely to be driving such relationships were prioritized. In agreement with a visual inspection of the prioritized gene list, this analysis showed a strong association between these genes and both kidney cancer development (adjusted $p=2.32\times10^{-9}$) and carcinoma development (adjusted $p=1.22\times10^{-5}$). A clustergram of the 49 genes that were directly associated with kidney cancer development is shown in **Figure 2**.

Finally, analysis of mutational signatures revealed three signatures (COSMIC signatures 1, 5 and 12) with significant activity in the tumors (**Figure 3**). Signatures 1 and 5 correspond to endogenous mutational processes that are consistently operative in nearly all human cells (Alexandrov et al., 2015). On the other hand, signature 12, whose etiology is presently unknown, has been previously described only in liver cancer, and thus its presence in ccRCC tumors is unprecedented (Alexandrov et al., 2013). Strikingly, although with borderline significance, the mutational contribution of signature 12 in the tumors tends to associate with the age at diagnosis (rho=0.71, $p=0.02$). Notwithstanding this result, the overall somatic mutation burden was not correlated with the age at diagnosis (rho=0.32, $p=0.36$).

4 Discussion

This study constitutes the first exome-wide approach for revealing genes with differential accumulation of somatic mutations in relation to cancer-associated mortality in ccRCC patients. Previous studies have either used targeted approaches directed to a limited set of genes that commonly accumulate mutations (Tennenbaum et al., 2017) or to evidence associations with treatment responses to the therapies (Fay et al., 2016; Miao et al., 2018). At most, these studies revealed that recurrent mutations in *PBRM1*, one of the well-known ccRCC genes, might have implications in the treatment responses. In contrast, our study enabled prioritizing 138 genes based on a refined set of high-confidence somatic SNVs, where 49 of those genes had previously been related to kidney cancer according to the literature. Our study has also yielded unprecedented evidence of the activity of COSMIC signature 12 in kidney cancer, in addition to two well-established endogenous mutational processes.

Among the genes displaying the strongest associations between somatic mutation and mortality from ccRCC, a few genes are strong candidates for further study. Of these, *GPR155*, *INPP5K*, *KRT7*, and *CYP4B1* are worth highlighting. *GPR155* encodes the G protein-coupled receptor 155, a

transmembrane transporter involved in the entry of growth factors and chemotherapeutic agents in tumor cells, and has been previously related to hepatocellular carcinoma on the basis of gene expression, methylation and copy number analyses (Umeda et al., 2017). The inositol polyphosphate-5-phosphatase K (also known as Skeletal muscle and kidney-enriched inositol phosphatase), encoded by *INPP5K*, is suggested to be involved in oncogenesis through participation in the PI3K/Akt pathway, which has an established role in cancer cell growth and survival. *INPP5K* is a tumor suppressor residing in 17p, which is commonly altered in the genome of a variety of human malignancies (Hedberg Oldfors et al., 2015). *KRT7* encodes the keratin 7 protein that is expressed *de novo* in pre-neoplastic lesions and associated tumors in chronic kidney disease (CKD) patients, being highly suggestive of a role in tissue remodeling and tumorigenesis (Sarlos et al., 2018). In fact, this protein is used in the clinic as a biomarker for tumor classification (Giunchi et al., 2016; Renshaw et al., 2018; Sarlos et al., 2018). Finally, *CYP4B1*, encoding cytochrome P450 family 4 subfamily B member 1, is frequently downregulated through hypermethylation in carcinomas and is associated with malignancy in renal tissues (Imaoka et al., 2000).

Previously, three mucin-encoding genes (*MUC2*, *MUC4* and *MUC12*) were found repeatedly altered in colorectal tumor tissues using RNA-seq data (Yin et al., 2013). In our ccRCC patients, *MUC12*, *MUC16* and *MUC5B* showed evidence of somatic mutation recurrence, associated with mortality. There is much evidence supporting the link between *MUC16* and progression and metastasis of ovarian cancer, promotion of cancer cell proliferation and inhibition of anti-cancer immune responses. In fact, one of its epitopes (CA125) is routinely used in serum assays for patient monitoring. Consistently, The Cancer Genome Atlas (TCGA) patient data show that carriers of *MUC16* mutations have slightly lower survival rates than non-carriers, although such difference is not statistically significant, likely due to low statistical power (Felder et al., 2014). Similarly, *MUC12* has lower expression in cancer tissues than in the adjacent normal tissues, and has been put forward as a candidate biomarker for disease-free survival in colorectal cancer (Matsuyama et al., 2010). Recently, *MUC12* was also identified through a WES approach as one of the highly mutated genes in a particular type of liver cancer (Wang et al., 2018). The results for *MUC5B* are less clear, since there is no direct evidence in the literature of a link with oncogenic processes, but only with susceptibility and survival in pulmonary fibrosis through germline regulatory variants (Seibold et al., 2011; Noth et al., 2013; Fingerlin et al. 2013; Peljto et al., 2013; Allen et al., 2017). Curiously, a recent WES study by Lata et al (2018), aimed at providing diagnosis of adult probands with CKD of unknown cause, identified causal germline mutations in *PARN* (poly(A)-specific ribonuclease), another pulmonary fibrosis susceptibility gene (Stuart et al., 2015). In agreement with these results, some studies have argued in support of pathogenic similarities between pulmonary fibrosis and cancer (Vancheri, 2015). Under such scenario, it could be speculated that coding mutations in *MUC5B* and *PARN* may play a role in oncogenesis in lung and kidney tissues.

One of the most notable strengths of this study is that it focuses on a homogeneous patient population, with all patients being included at stage IV and similarly treated. Besides, the combination of high-depth WES of matched tumor–normal sample pairs from each patient, and the multiple filtering routines performed after variant calling, enabled the derivation of a high-confidence set of somatic variants. Notwithstanding, we also acknowledge some limitations in the study. First, the study evaluated a small patient series and focused on the high-mortality risk spectrum of ccRCC cases. Therefore, the results may not be representative of the full disease spectrum. Second, we only sequenced a single specimen at the point of patient diagnosis. As a consequence, our capacity to identify candidate genes linked to ccRCC survival was limited to those at the pre-treatment stage, hindering the possibility of identifying additional genes as the tumors responded to therapy. Third,

because of the capture design of WES, we were unable to assess the association of non-coding variants with patient mortality, as has been previously suggested for regulatory variants in the telomerase reverse transcriptase encoding gene *TERT* (Casuscelli et al., 2017). Fourth, structural variants are common in ccRCC (Thiesen et al., 2018) and have been associated with poorer prognosis (Moore et al., 2012; Chen et al., 2009). However, we did not explore the implications of these on patient mortality, as our analyses focused on SNVs and indels. Finally, we did not adjust for multiple testing because our analyses were exploratory in nature. As such, our results should be regarded as hypothesis-generating findings.

5 Conclusions

In this study, we identify 138 genes that are recurrently altered in ccRCC tumors and that associate with patient mortality. Many of these have been previously suggested as biomarkers of cancer prognosis, and participate in molecular pathways linked to tumor development and progression. Additionally, we provide unprecedented evidence of the activity of COSMIC mutational signature 12 in kidney cancer, suggesting that the understanding of the mutational processes involved in this kind of malignancy remains incomplete. Independent validation studies achieving larger statistical power are needed to better evaluate the impact on ccRCC patient mortality of somatic mutations in our list of prioritized genes.

6 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7 Author Contributions

ABO, AMA, BGG, and CF wrote the manuscript and the supplementary files. BGG and SLL performed DNA extractions and quantifications, as well as the preparation of samples for sequencing. CHP, MDCM, and MM performed data collection and the record of patient and sample features. ABO and JMLS conceived and implemented software procedures. AMA, CF and JMLS performed the statistical tests on detected somatic variants. CF provided a general supervision of the project, giving guidelines for each step. All the authors provided insights, corrections and approved the final version of the manuscript.

8 Funding

This study was funded by Fundación CajaCanarias (SALUCAN11). AMA was supported by a fellowship from CajaSiete-ULL. BGG was supported by a fellowship from the Canarian Agency for Research, Innovation and Information Society (ACIISI, grant number TESIS2015010057) co-funded by the European Social Fund (ESF).

9 Acknowledgments

We would like to thank Joaquín Martínez-Muñoz for the helpful technical support at the early stages of this project.

10 References

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., et al. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. 47, 1402–1407. doi:10.1038/ng.3441.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. Nature 500, 415–421. doi:10.1038/nature12477

Alexandrov, L. B., and Stratton, M. R. (2014). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr. Opin. Genet. Dev. 24, 52–60. doi:10.1016/j.gde.2013.11.014.

Allen, R. J., Porte, J., Braybrooke, R., Flores, C., Fingerlin, T. E., Oldham, J. M., et al. (2017). Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. Lancet Respir Med 5, 869–880. doi.org/10.1016/S2213-2600(17)30387-9.

Aulchenko, Y. S., Ripke, S., Isaacs, A., and van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics 23, 1294–1296. doi:10.1093/bioinformatics/btm108.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). The new S language. A programming environment for data analysis and graphics. Wadsworth & Brooks.

Casuscelli, J., Becerra, M. F., Manley, B. J., Zabor, E. C., Reznik, E., Redzematovic, A., et al. (2017). Characterization and Impact of TERT Promoter Region Mutations on Clinical Outcome in Renal Cell Carcinoma. Eur Urol Focus. doi:10.1016/j.euf.2017.09.008.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14, 128. doi:10.1186/1471-2105-14-128.

Chen, M., Ye, Y., Yang, H., Tamboli, P., Matin, S., Tannir, N. M., et al. (2009). Genome-wide profiling of chromosomal alterations in renal cell carcinoma using high-density single nucleotide polymorphism arrays. Int. J. Cancer 125, 2342–2348. doi.org/10.1002/ijc.24642 Cited

Chilamakuri, C. S. R., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., et al. (2014). Performance comparison of four exome capture systems for deep sequencing. BMC Genomics 15, 449. doi:10.1186/1471-2164-15-449.

Damiati, E., Borsani, G., and Giacopuzzi, E. (2016). Amplicon-based semiconductor sequencing of human exomes: performance evaluation and optimization strategies. Hum. Genet. 135, 499–511. doi:10.1007/s00439-016-1656-8.

Farwell, K. D., Shahmirzadi, L., El-Khechen, D., Powis, Z., Chao, E. C., Tippin Davis, B., et al. (2015). Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. Genet. Med. 17, 578–586.

Fay, A. P., de Velasco, G., Ho, T. H., Van Allen, E. M., Murray, B., Albiges, L., et al. (2016). Whole-Exome Sequencing in Two Extreme Phenotypes of Response to VEGF-Targeted Therapies in Patients With Metastatic Clear Cell Renal Cell Carcinoma. J. Natl. Compr. Canc. Netw. 14, 820–824.

Felder, M., Kapur, A., Gonzalez-Bosquet, J., Horibata, S., Heintz, J., Albrecht, R., et al. (2014). MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. Mol. Cancer 13, 129. doi.org/10.1186/1476-4598-13-129

Fingerlin, T. E., Murphy, E., Zhang, W., Peljto, A. L., Brown, K. K., Steele, M. P., et al. (2013). Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat. Genet. 45, 613–620. doi.org/10.1038/ng.2609

Flanigan, R. C., Salmon, S. E., Blumenstein, B. A., Bearman, S. I., Roy, V., McGrath, P. C., et al. (2001). Nephrectomy followed by interferon alfa-2b compared with interferon alfa-2b alone for metastatic renal-cell cancer. N. Engl. J. Med. 345, 1655–1659. doi:10.1056/NEJMoa003013

Fujita, S., Masago, K., Okuda, C., Hata, A., Kaji, R., Katakami, N., et al. (2017). Single nucleotide variant sequencing errors in whole exome sequencing using the Ion Proton System. Biomed Rep 7, 17–20. doi:10.3892/br.2017.911.

Giunchi, F., Fiorentino, M., Vagnoni, V., Capizzi, E., Bertolo, R., Porpiglia, F., et al. (2016). Renal oncocytosis: a clinicopathological and cytogenetic study of 42 tumours occurring in 11 patients. Pathology 48, 41–46. doi.org/10.1016/j.pathol.2015.11.009

Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. bioRxiv 372896 (2018). doi:10.1101/372896

Gu, Y., Zou, Y. M., Lei, D., Huang, Y., Li, W., Mo, Z., et al. (2017). Promoter DNA methylation analysis reveals a novel diagnostic CpG-based biomarker and RAB25 hypermethylation in clear cell renel cell carcinoma. Sci. Rep. 7, 14200. doi.org/10.1038/s41598-017-14314-y

Hedberg Oldfors, C., Dios, D. G., Linder, A., Visuttijai, K., Samuelson, E., Karlsson, S., et al. (2015). Analysis of an independent tumor suppressor locus telomeric to Tp53 suggested Inpp5k and Myo1c as novel tumor suppressor gene candidates in this region. BMC Genet. 16, 80. doi.org/10.1186/s12863-015-0238-4

Heng, D. Y. C., Xie, W., Regan, M. M., Harshman, L. C., Bjarnason, G. A., Vaishampayan, U. N., et al. (2013). External validation and comparison with other models of the International Metastatic Renal-Cell Carcinoma Database Consortium prognostic model: a population-based study. Lancet Oncol. 14, 141–148. doi:10.1016/S1470-2045(12)70559-4.

Hudes, G., Carducci, M., Tomczak, P., Dutcher, J., Figlin, R., Kapoor, A., et al. (2007). Temsirolimus, interferon alfa, or both for advanced renal-cell carcinoma. N. Engl. J. Med. 356, 2271–2281. doi:10.1056/NEJMoa066838.

Iglesias, A., Anyane-Yeboa, K., Wynn, J., Wilson, A., Truitt Cho, M., Guzman, E., et al. (2014). The usefulness of whole-exome sequencing in routine clinical practice. Genet. Med. 16, 922–931. doi:10.1038/gim.2014.58.

Imaoka, S., Yoneda, Y., Sugimoto, T., Hiroi, T., Yamamoto, K., Nakatani, T., et al. (2000). CYP4B1 is a possible risk factor for bladder cancer in humans. Biochem. Biophys. Res. Commun. 277, 776–780. doi.org/10.1006/bbrc.2000.3740

Kong, S. W., Lee, I.-H., Liu, X., Hirschhorn, J. N., and Mandl, K. D. (2018). Measuring coverage and accuracy of whole-exome sequencing in clinical context. Genet. Med. doi:10.1038/gim.2018.51

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 44, W90–7. doi:10.1093/nar/gkw377

Lata, S., Marasa, M., Li, Y., Fasel, D. A., Groopman, E., Jobanputra, V., et al. (2018). Whole-Exome Sequencing in Adults With Chronic Kidney Disease: A Pilot Study. Ann. Intern. Med. 168, 100–109. doi:10.7326/M17-1319

Manley, B. J., Zabor, E. C., Casuscelli, J., Tennenbaum, D. M., Redzematovic, A., Becerra, M. F., et al. (2017). Integration of Recurrent Somatic Mutations with Clinical Outcomes: A Pooled Analysis of 1049 Patients with Clear Cell Renal Cell Carcinoma. Eur Urol Focus 3, 421–427. doi:10.1016/j.euf.2016.06.015

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., et al. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 171, 1029–1041.e21. doi:10.1016/j.cell.2017.09.042

Matsuyama, T., Ishikawa, T., Mogushi, K., Yoshida, T., Iida, S., Uetake, H., et al. (2010). MUC12 mRNA expression is an independent marker of prognosis in stage II and stage III colorectal cancer. Int. J. Cancer 127, 2292–2299. doi.org/10.1002/ijc.25256

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The Ensembl Variant Effect Predictor. Genome Biol. 17, 122. doi:10.1186/s13059-016-0974-4

Mandriota, S. J., Turner, K. J., Davies, D. R., Murray, P. G., Morgan, N. V., Sowter, H. M., et al. (2002). HIF activation identifies early lesions in VHL kidneys: evidence for site-specific tumor suppressor function in the nephron. Cancer Cell 1, 459–468. doi.org/10.1016/S1535-6108(02)00071-5

Miao, D., Margolis, C. A., Gao, W., Voss, M. H., Li, W., Martini, D. J., et al. (2018). Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. Science 359, 801–806. doi:10.1126/science.aan5951

Mickisch, G. H., Garin, A., van Poppel, H., de Prijck, L., Sylvester, R., and European Organisation for Research and Treatment of Cancer (EORTC) Genitourinary Group (2001). Radical nephrectomy plus interferon-alfa-based immunotherapy compared with interferon alfa alone in metastatic renal-cell carcinoma: a randomised trial. Lancet 358, 966–970. doi.org/10.1016/S0140-6736(01)06103-7

Mitchell, T. J., Turajlic, S., Rowan, A., Nicol, D., Farmery, J. H. R., O'Brien, T., et al. (2018). Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. Cell. doi:10.1016/j.cell.2018.02.020.

Moore, L. E., Jaeger, E., Nickerson, M. L., Brennan, P., De Vries, S., Roy, R., et al. (2012). Genomic copy number alterations in clear cell renal carcinoma: associations with case characteristics and mechanisms of VHL gene inactivation. Oncogenesis 1, e14. doi.org/10.1038/oncsis.2012.14

Motzer, R. J., Hutson, T. E., Cella, D., Reeves, J., Hawkins, R., Guo, J., et al. (2013). Pazopanib versus sunitinib in metastatic renal-cell carcinoma. N. Engl. J. Med. 369, 722–731. doi:10.1056/NEJMoa1303989

Noth, I., Zhang, Y., Ma, S.-F., Flores, C., Barber, M., Huang, Y., et al. (2013). Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. Lancet Respir Med 1, 309–317. doi.org/10.1016/S2213-2600(13)70045-6

Peljto, A. L., Zhang, Y., Fingerlin, T. E., Ma, S.-F., Garcia, J. G. N., Richards, T. J., et al. (2013). Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. JAMA 309, 2232–2239. doi:10.1001/jama.2013.5827

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2008). ISBN 3-900051-07-0, URL http://www.R-project.org.

Renshaw, A. A., and Gould, E. W. (2018). Ancillary studies in fine needle aspiration of the kidney. Cancer Cytopathol. 126 Suppl 8, 711–723. doi.org/10.1002/cncy.22029

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., WGS500 Consortium, et al. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. 46, 912–918. doi:10.1038/ng.3036

Samarakoon, P. S., Sorte, H. S., Kristiansen, B. E., Skodje, T., Sheng, Y., Tjønnfjord, G. E., et al. (2014). Identification of copy number variants from exome sequence data. BMC Genomics 15, 661. doi:10.1186/1471-2164-15-661

Sarlos, D. P., Peterfi, L., Szanto, A., and Kovacs, G. (2018). Shift of Keratin Expression Profile in End-stage Kidney Increases the Risk of Tumor Development. Anticancer Res. 38, 5217–5222. doi: 10.21873/anticanres.12845

Seibold, M. A., Wise, A. L., Speer, M. C., Steele, M. P., Brown, K. K., Loyd, J. E., et al. (2011). A common MUC5B promoter polymorphism and pulmonary fibrosis. N. Engl. J. Med. 364, 1503–1512. doi: 10.1056/NEJMoa1013660

Sternberg, C. N., Davis, I. D., Mardiak, J., Szczylik, C., Lee, E., Wagstaff, J., et al. (2010). Pazopanib in locally advanced or metastatic renal cell carcinoma: results of a randomized phase III trial. J. Clin. Oncol. 28, 1061–1068. doi:10.1200/JCO.2009.23.9764

Stuart, B. D., Choi, J., Zaidi, S., Xing, C., Holohan, B., Chen, R., et al. (2015). Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. Nat. Genet. 47, 512–517. doi.org/10.1038/ng.3278

Tennenbaum, D. M., Manley, B. J., Zabor, E., Becerra, M. F., Carlo, M. I., Casuscelli, J., et al. (2017). Genomic alterations as predictors of survival among patients within a combined cohort with

clear cell renal cell carcinoma undergoing cytoreductive nephrectomy. Urol. Oncol. 35, 532.e7–532.e13. doi:10.1016/j.urolonc.2017.03.015

Thiesen, H.-J., Steinbeck, F., Maruschke, M., Koczan, D., Ziems, B., and Hakenberg, O. W. (2017). Stratification of clear cell renal cell carcinoma (ccRCC) genomes by gene-directed copy number alteration (CNA) analysis. PLoS One 12, e0176659. doi.org/10.1371/journal.pone.0176659

Umeda, S., Kanda, M., Sugimoto, H., Tanaka, H., Hayashi, M., Yamada, S., et al. (2017). Downregulation of GPR155 as a prognostic factor after curative resection of hepatocellular carcinoma. BMC Cancer 17, 610. doi: 10.1186/s12885-017-3629-2

Valencia, C. A., Husami, A., Holle, J., Johnson, J. A., Qian, Y., Mathur, A., et al. (2015). Clinical Impact and Cost-Effectiveness of Whole Exome Sequencing as a Diagnostic Tool: A Pediatric Center's Experience. Front Pediatr 3, 67. doi:10.3389/fped.2015.00067

Vancheri, C. (2015). Idiopathic pulmonary fibrosis and cancer: do they really look similar? BMC Med. 13, 220. doi.org/10.1186/s12916-015-0478-1

Velmurugan, K. R., Varghese, R. T., Fonville, N. C., and Garner, H. R. (2017). High-depth, high-accuracy microsatellite genotyping enables precision lung cancer risk classification. Oncogene 36, 6383–6390. doi:10.1038/onc.2017.256.

Wang, A., Wu, L., Lin, J., Han, L., Bian, J., Wu, Y., et al. (2018). Whole-exome sequencing reveals the origin and evolution of hepato-cholangiocarcinoma. Nat. Commun. 9, 894. doi:10.1038/s41467-018-03276-y.

Yin, H., Liang, Y., Yan, Z., Liu, B., and Su, Q. (2013). Mutation spectrum in human colorectal cancers and potential functional relevance. BMC Med. Genet. 14, 32. doi.org/10.1186/1471-2350-14-32

11 Figure legends

Figure 1. Quantile-quantile plot of the mutation burden test results from the association with mortality.

Figure 2. Clustergram representing the association of the subset of 49 prioritized genes that have direct links with kidney cancer. Relationships with other cancer types are also shown. Significance values are shown on the top.

Figure 3. Proportion of COSMIC signatures displaying significant activity in each patient tumor. Color code correspondence is: green, signature 1; red, signature 5; and purple, signature 12.

Figure 1.TIFF

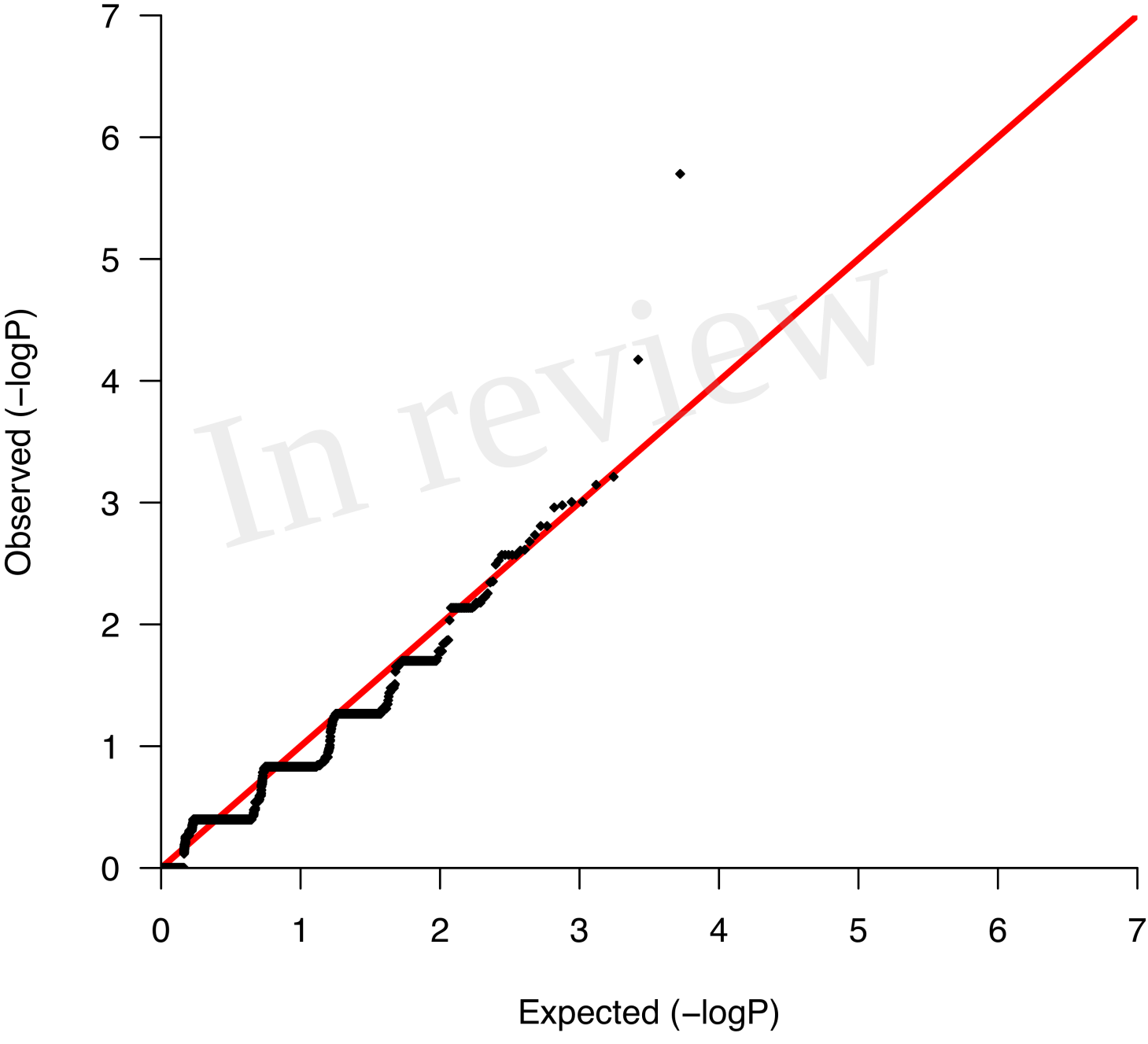

Figure 2.TIFF

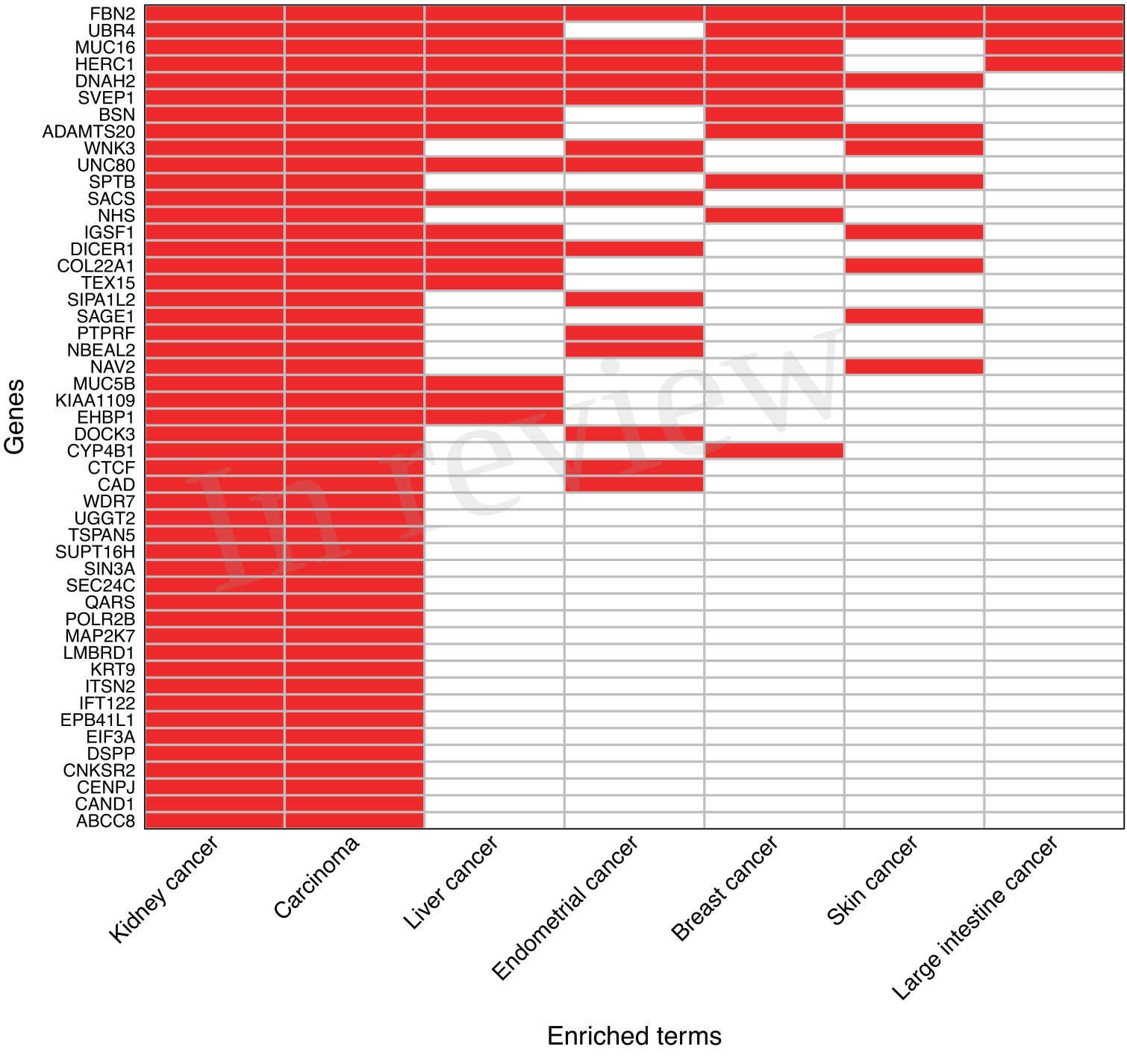

Figure 3.TIFF