Whole-genome analysis of papillary kidney cancer finds significant non-coding alterations

Authors: Shantao Li[1], Brian M. Shuch[2,*], Mark B. Gerstein[1,3,4,*]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA.

[2]Department of Urology, Yale School of Medicine, New Haven, CT, 06520, USA.

[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA.

[4]Department of Computer Science, Yale University, New Haven, CT 06520, USA.

*To whom correspondence should be addressed: brian.shuch@yale.edu, pi@gersteinlab.org

Short title: Whole-genome analysis of papillary kidney cancer

Abstract: To date, studies on papillary renal-cell carcinoma (pRCC) have largely focused on coding alterations in traditional drivers, particularly *MET*. However, for a significant fraction of tumors, researchers have been unable to determine clear molecular etiologies. To address this, we perform the first whole-genome analysis of pRCC. Elaborating on previous results on *MET*, in the coding regions of this gene we find more somatic alternations and a germline SNP predicting prognosis (rs11762213). We identify activation of promoter of retrotransposons in MET due to methylation dysregulation as a driver event. Next, we scrutinize non-coding mutations, discovering potentially impactful ones in regions associated with *MET* and a long non-coding RNA (*NEAT1*). Moreover, *NEAT1* is implicated in other cancer and its mutations in pRCC are associated with increased expression and unfavorable outcome. Finally, we investigate genome-wide mutational patterns, finding they are governed mostly by methylation-associated C-to-T transitions. Also, we observe significantly more mutations in open chromatin and early

Shantao 2/18/2017 7:04 PM
Comment [1]: New results:
1. MET retrotransposons
2. SV: SDHB del.?
3.Evolution tree topology
4.CR defects associated with RT

Shantao 2/17/2017 6:45 PM
Deleted: Papillary renal-cell carcinoma (pRCC) constitutes 10-15% of kidney tumors.

Shantao 2/17/2017 6:45 PM
Deleted: it

Shantao 2/18/2017 4:31 AM
Deleted: Interestingly, we find no enrichment for small structural variants associated with *MET*.

Shantao 2/18/2017 7:04 PM
Comment [2]: Too Strong?

Shantao 2/18/2017 7:04 PM
Formatted: Highlight

replicated regions in tumors with chromatin-modifier alterations. We build evolution trees for individual tumor and find their topologies are associated with tumor subtypes.

Currently word count:177, need to be <=150

Author Summary

Renal cell carcinoma accounts for more than 90% of kidney cancers. Papillary renal cell carcinoma (pRCC) is the second most common subtype of renal cell carcinoma. Previous studies, focusing mostly on the protein-coding regions, have identified several key genomic alterations that are key to cancer initiation and development. However, researchers cannot find any key mutation in a significant portion of pRCC. Therefore, we carry out the first whole-genome study of pRCC to discover triggering DNA changes explaining these cases. By looking at the entire genome, we find additional potentially impactful alterations in and out of the protein-coding regions. These newly identified critical mutations from scrutinizing the entire genome help complete our understanding of pRCC genomes. Two alterations we found are associated with prognosis, which could aid clinical decisions. We are also able to recognize mutation patterns, signatures and tumor evolution structures, which reflect the mutagenesis processes and give hints on how cancer develops. Our study provides valuable additional information to facilitate better tumor subtyping, risk stratification and potentially clinical management.

Introduction

Renal cell carcinoma (RCC) makes up over 90% of kidney cancers and currently is the most lethal genitourinary malignancy (1). Papillary RCC (pRCC) accounts for 10%-15% of the total RCC cases (2). Unfortunately pRCC has been understudied and there are no current forms of effective systemic therapy for this disease. pRCC are further subtyped into two major groups: type 1 and type 2 based on histopathological features. For many years, the only prominent oncogene in pRCC (specifically, type 1) that physicians were able to identify was *MET*, a tyrosine kinase receptor for hepatic growth factor. An amino acid substitution that leads to constitutive activation and/or overexpression are two mechanisms of dysfunction of *MET* in tumorigenesis. Recently, the Cancer Genome Atlas (TCGA) published its first result on pRCC (3), which greatly improves our understanding of the genomic basis of this disease. Several more genes and specific sub-clusters were identified to be significantly mutated in pRCC. Nevertheless, a significant portion of pRCC cases still remains without any known driver. Therefore we think it is time to explore the rest 98% non-coding regions of the genome using whole genome sequencing (WGS). This is sensible because non-coding regions, previously overlooked in cancer, have been showed to be actively involved in tumorigenesis (4-6). Mutations in non-coding regions may cause disruptive changes in both cis- and trans-regulatory elements, affecting gene expression. Understanding non-coding mutations helps fill the missing "dark matter" in cancer research.

Multiple endogenous and environmental mutation processes shape the somatic mutational landscape observed in cancers (7). Analyses of the genomic alterations associated with these processes give information on cancer development, shed light on mutational disparity between cancer subtypes and even indicate potential new treatment strategies (8). Additionally, genomic features such as replication time and chromatin environment govern mutation rate along the genome, contributing to spatial mutational heterogeneity. While identifying mutation signatures is possible using data from whole exome sequencing (WXS), whole genome sequencing (WGS) gives richer information on mutation landscape and minimizes the potential confounding effects of exome capture process and driver selection.

In this study, we comprehensively analyzed 35 pRCC cases that were whole genome sequenced along with an extensive set of WXS data on multiple levels. We went from microscopic examination of driver genes to analyses of whole genome sequencing variants, and finally, to investigation of high-order mutational features. First, we focused on *MET*, an oncogene which plays a central role in pRCC, especially in type 1. We found rs11762213, a germline exonic single nucleotide polymorphism inside *MET*, predicts cancer-specific survival (CSS) in type 2 pRCC. We also discovered several potentially impactful non-coding mutation hotspots in *MET* promoter and its first two exons. The previous TCGA study identifies a *MET* alternative transcription event as a driver event but without illustrating the etiology (3). We found that a cryptic promoter from a long interspersed nuclear element-1 (L1) triggers the alternative isoform expression. Surprisingly, we did not find a significant amount of structural variations affecting *MET* besides polysomy 7. Then we went onto cases not as easily explained as those with *MET* alterations. We analyzed about 160,000 non-coding mutations throughout the entire genomes and found several potentially high-impact mutations in non-coding regions. Further zooming out, we discovered pRCC exhibits mutational heterogeneity in both nucleotide context and genome location, indicating underlying vibrant mutational processes interplay. We found methylation is the leading factor influencing mutation landscape. Methylation status drives the intra-sample mutation variation by promoting more C-to-T mutations in the CpG context. APOBEC activity, although infrequently observed, leaves an unequivocal mutation signature in a pRCC genome but not in ccRCC. Also, we discovered samples with chromatin remodeler alternations accumulate more mutations in open chromatin and early-replicated regions. Last, we

Shantao 1/24/2017 12:04 AM
Deleted: 32

Shantao 1/23/2017 5:32 PM
Formatted: Font:Italic

Shantao 1/24/2017 2:29 AM
Deleted: 3

Shantao 2/9/2017 6:38 PM
Deleted: nearly

Shantao 2/9/2017 6:38 PM
Deleted: 5

Shantao 2/18/2017 7:08 PM
Deleted: Last

inferred evolution tree for each individual samples and found tree structures correlate with tumor subtypes.

Results

Shantao 2/18/2017 7:09 PM
Deleted:

1. An exonic SNP in *MET*, rs11762213, predicts prognosis in type 2 pRCC.

We begin with coding variants in the long known driver *MET*. The TCGA study of 161 pRCC patients found 15 samples carrying somatic, nonsynonymous single nucleotide variant (SNV) in *MET*. By analyzing 117 extra WXS samples (see Methods), we found six more nonsynonymous somatic mutations in six samples (Table S1). V1110I and M1268T were two recurrent mutations in this extra set. Both of them were observed in the TCGA study as well. Additionally, we found two samples carrying H112Y and Y1248C respectively. H1112Y has been observed in two patients the original TCGA study cohort and H1118R is a long-known germline mutation associated with hereditary papillary renal carcinoma (HPRC, 13). Y1248C has been observed in type 1 pRCC before (rs121913246) and the TCGA cohort has a case carrying Y1248H. All mutations occur in the hypermutated tyrosine kinase catalytic domain of *MET*. Two out of these six samples were identified as type 1 pRCC while the subtypes of the rest four were unknown.

Although many MET somatic mutations are believed to play a central role in pRCC, some germline *MET* mutations have also been associated with the disease. In particular, a germline SNP, rs11762213, has been discovered to predict recurrence and survival in a mixed RCC cohort (14). ccRCC predominated the initial discovery RCC cohort. This conclusion was later validated in a ccRCC cohort but never in pRCC (9). We wondered whether this SNP has a prognostic effect in pRCC. Using an extensive WXS set of 277 patients (see Methods; Figure S1 and Table S1;), we found 14 patients carry one risk allele of rs11762213 (G/A, Table 1, minor allele frequency (MAF) = 2.53%). No homozygous A/A was observed. Cancer specific deceases are concentrated in type 2 pRCC. Among 96 type 2 pRCC cases, seven patients carry the minor A allele (MAF = 3.65%, Table 1). Survival is significantly worse in type 2 patients carrying the risk allele of rs11762213 (p = 0.034, Figure 1B). But we did not find significant association of this germline SNP with survival in type 1 patients. We did not find statistically significant association of rs11762213 with *MET* RNA expression in either tumor samples or normal controls (p > 0.1, two-sided rank-sum test). *Met* pY1235 levels in tumor samples, as measured by Reverse phase protein array (RPPA), were not significantly different in patients carrying the minor G allele compared to patients with A/A genotype (p > 0.1, two-sided rank-sum test).

Characteristic	**G/A (n = 7)**	**A/A (n = 89)**
Sex, No. (%)		
Male (%)	4 (57)	25 (28)
Female (%)	3 (43)	64 (72)
Age, median (IQR), y	54 (47-61)	65 (57-73)
Race, No. (%)		
White	6 (86)	65 (73)
Black	1 (14)	16 (18)
Asian	0	4 (4)
NA	0	4 (4)
T stage, No. (%)		
T1	4 (57)	47 (53)
T2	1 (14)	10 (11)
T3	2 (29)	31 (35)
T4	0	1 (1)
N stage, No. (%)		
N0	3 (43)	20 (22)
N1	0	15 (17)
N2	1 (14)	2 (2)
NX	3 (43)	52 (58)
M stage, No. (%)		
M0	3 (43)	54 (61)

M1	1 (14)	4 (4)
MX/NA	3 (43)	31 (35)
AJCC stage, No. (%)		
I	4 (57)	43 (48)
II	0	7 (8)
III	1 (14)	29 (33)
IV	2 (29)	6 (7)
NA	0	4 (4)
Median follow-up for surviving patients, days (IQR)	243 (132-354)	579 (219-1247)

Table 1. Patient clinical profiles of the type 2 pRCC cohort in rs11762213 survival analysis. AJCC: American Joint Committee on Cancer; IQR: interquartile range; NA: not available. Percentages may not add up to 100% because of rounding.

2. Epigenetic alterations and mutation hotspots in non-coding regions

Shantao 1/23/2017 5:41 PM
Formatted: Font:Bold, Not Italic

Shantao 1/23/2017 5:40 PM
Deleted: M

The TCGA study has identified a *MET* alternative translation isoform as a driver event (3). However, the etiology of this new isoform is unknown. We identified this isoform results from the usage of a cryptic promoter from an L1 element (Figure 1A), likely due to a local loss of methylation (REF). This event was reported in several other cancer types (REF). To test its relationship with methylation, we found a closet probe (cg06985664, ~3kb downstream) on the Methylation array show marginally statistically significant (p=0.055, one-side rank-sum test). Additionally, as expected, this event is associated with methylation group 1 (odds ration (OR)= 4.54, p<0.041), indicating genome-wide methylation dysfunction. This association is stronger in type 2 pRCC and it shows a significant association with the C2b cluster (OR= 17.5, p<0.007).

Despite the fact *MET* is the most common driver alteration, about 20% presumably *MET*-driven yet *MET* wild-type pRCC samples were still left unexplained (3). Therefore, we scanned the *MET* non-coding regions. We observed one mutation in *MET* promoter region in a type 1 pRCC sample (Figure 1A and Table S2). This sample shows no evidence of a nonsynonymous

Shantao 2/18/2017 7:12 PM
Deleted: 2

mutation in *MET* gene but it has copy number gain of *MET*. Additionally, we observed 6/35 (17.1%) samples carry mutations in the intronic regions between exon 1-3 of *MET* (Figure 1A and Table S2). Previously it is been established that alternative splicing of these exons is a driver event (3). Therefore we speculated that these non-coding variants might correlate with the alternative splicing. However, likely being hindered by a small size, we were not able to find statistically significant association between the alternative splicing event and these intronic mutations.

We further expanded our scope and ran FunSeq (4-5) to identify potentially high-impact, non-coding variants in pRCC. First, we identified a high-impact mutation hotspot on chromosome 1. 6/35 (17.1%) samples have mutations within this 6.5kb region (Figure 2A and Table S2). This hotspot locates at the upstream of *ERRFI1* (ERBB Receptor Feedback Inhibitor 1) and overlaps with the predicted promoter region. ERRFI1 is a negative regulator of EGFR family members, including EGFR, HER2 and HER3, all have been implicated in cancer. Due to a very limited sample size here, our test power was inevitably low. We didn't observe statistically significant changes among mutated samples in mRNA expression level, protein level and phosphorylation level of EGFR, HER2 and HER3.

Another potentially impactful mutation hotspot is in *NEAT1*. We saw mutations inside this nuclear long non-coding RNA in 6/35(17.1%) samples (Figure 2B and Table S2). Several studies indicated *NEAT1* is associated in many other cancers (15-16). It promotes cell proliferation in hypoxia (17) and alters the epigenetic landscape, increasing transcription of target genes (18).

All the mutations we found fell into a putative promoter region of *NEAT1*. We noticed *NEAT1* mutations were associated with higher *NEAT1* expression (Figure 2C, p < 0.032, two-

Shantao 1/24/2017 12:09 AM
Deleted: 32

Shantao 1/24/2017 12:09 AM
Deleted: 18.8

Shantao 2/18/2017 7:11 PM
Deleted: 2

Shantao 1/24/2017 12:05 AM
Deleted: 2

Shantao 1/24/2017 12:05 AM
Deleted: 8.8

Shantao 2/18/2017 7:12 PM
Deleted: 2B

Shantao 1/24/2017 12:04 AM
Deleted: 5

Shantao 1/24/2017 12:05 AM
Deleted: 2

Shantao 1/24/2017 12:05 AM
Deleted: 15.6

Shantao 2/18/2017 7:12 PM
Deleted: 2C

Shantao 2/18/2017 7:12 PM
Deleted: 2D

Shantao 2/18/2017 7:13 PM
Deleted: 044

sided rank sum test). We also found *NEAT1* mutations were associated with worse prognosis (Figure 2D, $p < 0.041$, log-rank test). However, without mutation status, *NEAT1* expression level is not significantly linked with pRCC survival. Nonetheless, *NEAT1* is overexpressed in 5% ccRCC samples from the TCGA cohort. *NEAT1* overexpression is significantly associated with shorted overall survival (Fig SXX). *MALAT1*, another noticeable lncRNA in cancer, is tightly co-expressed with *NEAT1* in both pRCC and ccRCC (Spearman's correlation: 0.79 and 0.87 respectively). Catalogue of Somatic Mutations in Cancer (COSMIC) (REF) annotates MALAT1 as cancer consensus gene, associating it with pediatric RCCs and lung cancer. Overexpression of *MALAT1* is reported to be associated with cancer progression (REF).

We used DELLY (10) to perform structural variants (SVs) discovery from WGS reads information (see Methods and Table S3). The SV discovery approach has higher sensitivity and resolution than array-based methods, which were employed in the TCGA analysis. In the end we found 424 somatic SV events, includes deletions, duplications, inversions and translocations (Figure SXX). Based on the SV event number, samples clearly split into two types: genome unstable (>40 events) and genome stable (<10 events).

First, by overlapping SVs with known pRCC related genes, we found two cases with deletion in SDHB. The median SDHB expression is only ~50% compared to cases without alternation (Figure SXX). We confirmed three cases carrying deletions affecting *CDKN2A* called by TCGA array-based methods but not the other two cases, possibly due to large-scale events (aneuploidy). Notably, three confirmed cases have significantly lower *CDKN2A* expression but not in the unconfirmed two cases. This suggests SV calling from WGS is accurate and predicts expression better. One sample, TCGA-B9-4116, which has extensive amplification of *MET*, showed multiple SVs of various classes hitting *MET* regions. However,

Shantao 2/18/2017 7:12 PM
Deleted: 2E

Shantao 2/18/2017 7:14 PM
Deleted: 022

Shantao 2/17/2017 7:21 PM
Formatted: Highlight

Shantao 2/17/2017 7:21 PM
Formatted: Font:Italic, Highlight

Shantao 2/17/2017 7:21 PM
Formatted: Highlight

Shantao 1/24/2017 12:17 AM
Formatted: Font:Italic

Shantao 2/18/2017 4:43 AM
Formatted: Font:Italic

Shantao 1/24/2017 12:18 AM
Formatted: Font:Italic

Shantao 1/24/2017 12:19 AM
Formatted: Font:Not Italic

Shantao 1/24/2017 12:18 AM
Formatted: Font:Italic

Shantao 1/24/2017 12:20 AM
Formatted: Font:Italic

Shantao 2/17/2017 7:13 PM
Deleted: 343

Shantao 2/17/2017 7:12 PM
Formatted: Not Highlight

Shantao 2/17/2017 7:12 PM
Formatted: Font:Italic

surprisingly, we did not find SVs affecting *MET* except this one example. We postulate trisomy/polysomy 7 is the main mechanism of *MET* structural alteration rather than duplication in a smaller scale. Besides duplication, we did not expect to find deletion, inversion or translocation disrupting oncogene *MET*. These SVs are likely to cause loss-of-function rather than gain-of-function mutations. This is consistent with the putative role of *MET* as an oncogene, rather than a tumor suppressor. Last, we observed several interesting sporadic events, including duplications in *EGFR* and *HIF1A* duplication and deletions in *DNMT3A* and *STAG2* (see SXX).

3. Mutation spectra and mutation processes of pRCC

To further get a high-order overview of the mutation landscape, we summarized the mutation spectra of 35 whole genome sequenced pRCC samples (Figure 3A). C-to-T in CpGs showed the highest mutation rates, which were roughly ten to twenty-fold higher than mutation rates in other nucleotide contexts.

We used principle components analysis (PCA) to reveal factors that explain the most inter-sample variation. The loadings on the first principle component (which explained 12.5% of the variation) demonstrated C-to-T in CpGs contributed the most to inter-sample variation (Figure 3B). C-to-T in CpGs is highly associated with methylation. It reflects the spontaneous deamination of cytosines in CpGs, which is much more frequent in 5-methyl-cytosines (REF). So we further explored the association between C-to-T in CpGs and tumor methylation status. First we validated the TCGA identified methylation cluster 1 showed higher methylation lever than cluster 2 in all annotation regions (Figure S2, see Methods), prominently in CpG Islands (OR of sites being differentially hypermethylated: 1.29, 95%CI: 1.20-1.39, p<0.0001). We confirmed this association by showing samples from methylation cluster 1 had higher PC1 scores

Shantao 2/17/2017 7:12 PM
Formatted: Not Highlight

Shantao 2/17/2017 7:11 PM
Deleted: .

Shantao 2/17/2017 7:19 PM
Formatted: Font:Italic

Shantao 2/17/2017 7:19 PM
Formatted: Font:Italic

Shantao 2/17/2017 7:19 PM
Formatted: Font:Italic

Shantao 2/17/2017 7:20 PM
Formatted: Font:Italic

Shantao 1/24/2017 2:27 AM
Deleted: 32

Shantao 1/23/2017 6:07 PM
Deleted: (hypermethylated group, Figure S2)

as well as higher C-to-T mutation counts and mutation percentages in CpGs (Figure 3C). This trend was further validated using a larger WXS dataset as well. Especially, the most hypermethylated group, CpG island methylation phenotype (CIMP), showed the greatest C-to-T in CpGs (Figure S2). As expected, C-to-T mutations in CpGs in group 1 showed higher but not statistically significant percentage overlapping with CpG islands compared with group 2 (1.8% versus 1.4%, p=0.14). Therefore, methylation status is the most prominent factor shaping the mutation spectra across patients. We further explored the functional impact of the excessive mutations driven by methylation. C-to-T mutations in CpGs were more likely to be in the coding region (OR=1.54, 95%CI: 1.27-1.85, p<0.0001) and nonsynonymous (OR=1.47, 95%CI: 1.17-1.84, p<0.001). Yet, C-to-T mutations in CpGs did not show functional bias between two methylation groups in non-coding regions.

Shantao 2/18/2017 7:20 PM
Comment [3]: Do we want this sentence?

Recently, several somatic mutation signatures were identified. Many have putative etiology, revealing the underlying mutation processes (7). We used a LASSO-based approach (see Methods) to decompose mutations into a linear combination of these canonical mutation signatures in both WGS and WXS samples (Figure S3). The leading signature was signature 5, which is consistent with previous studies (7). Interestingly, we found one type 2 pRCC case out of 155 somatic WXS sequenced samples exhibited APOBEC-associated mutation signature 2 and 13. APOBEC mutation pattern enrichment analysis (see Method) further confirmed the presence of APOBEC activity (Figure 3D). This sample was statistically enriched of APOBEC mutations (adjusted p-value < 0.0003).

Prominent APOBEC activities were also incidentally detected in three upper track urothelial cancer (UC) samples sequenced and processed in the same pipeline with pRCC samples. UC often carries APOBEC mutation signatures and our result is consistent with TCGA bladder urothelial cancer study (19).

The APOBEC-signature carrying pRCC case was centrally reviewed by six pathologists in the original study and confirmed to be type 2 pRCC (3). Thus this tumor is likely a special case of type 2 with genomic alterations share some similarities with UC. It has non-silent mutations in *ARID1A* and *MLL2* and a synonymous mutation in *RXRA*, all are identified as significantly mutated genes in UC but not in pRCC. Potential pRCC driver events, for example low expression of *CDKN2A* and nonsynonymous alternations in significantly mutated genes of pRCC, are absent in this sample.

Noticeably, all four samples with APOBEC activities showed significantly higher *APOBEC3A* and *APOBEC3B* mRNA expression level ($p < 0.0022$ and $p < 0.0039$ respectively, one-side rank sum test, Figure S4). This is in concordance with previous studies of APOBEC mutagenesis in various types of cancer (12).

Consistent with previous studies (12), we failed to detect statistically significant APOBEC activities in an extensive WXS dataset consisting of 418 clear cell RCC (ccRCC) samples, even after resampling to avoid p-value adjustment eroding the power. Very low levels of APOBEC signatures (<15%) was found in less than 1%(4/418) samples. With a much larger sample size, this result was unlikely to be confounded by detecting power.

4. Defects in chromatin remodeling affects mutation landscape

Chromatin remodeling genes are frequently mutated in pRCC and many other cancers including ccRCC (20). Defects in chromatin remodeling cause dysregulation of chromatin environment. Open chromatin regions show lower mutation rate, presumably due to more effective DNA repair (21). Thus chromatin remodeler alternations could possibly alter the mutation landscape, specifically increase mutation rate in previously open chromatin regions. To test this hypothesis, we tallied the number of mutations inside DNase I hypersensitive sites (DHS) in eleven normal fetal kidney cortex samples (The NIH Roadmap Epigenomics Mapping Consortium, REF), which represent the normal, physiological condition. 9/35 samples with disruptive mutations in ten chromatin remodeling, cancer associated genes show higher genome-wide mutation counts ($p < 0.021$, one-side rank-sum test), partially driven by higher mutation counts in DHS region ($p < 0.0023$, one-side rank-sum test). The median number of mutations in DHS region considerably increases by 60% (67.5 versus 108) in samples carrying chromatin remodeling defects. The effect is significant after normalizing against the total mutation counts ($p < 0.019$, one-side rank-sum test, Figure 3E).

Replication time is known to correlate greatly with mutation rate. Early replicating regions have lower mutation rate compared to late replicating ones. Researchers reason replication errors are more likely to be corrected by DNA repair system in early replicating regions. With defects in mutated chromatin remodeling, we observed this trend became less pronounced ($p<0.031$, one-side rank-sum test, Figure S5). This is likely because dysregulation of the chromatin environment hinders replication error repair by changing the accessibility of newly synthesized DNA chains.

5. Evolution tree analysis

Shantao 2/18/2017 7:22 PM
Deleted:

Shantao 2/18/2017 7:22 PM
Deleted: HEK293, a cell line derived from human embryonic kidney cells, the closest match we could find in the ENCODE DHS database.

Shantao 2/18/2017 7:24 PM
Deleted: 12

Shantao 2/18/2017 7:24 PM
Deleted: 2

Shantao 2/18/2017 7:24 PM
Deleted: non-silent

Shantao 2/18/2017 7:24 PM
Deleted: eleven

Shantao 2/18/2017 7:25 PM
Deleted: 32

Shantao 2/18/2017 7:25 PM
Deleted: 0

Shantao 2/18/2017 7:38 PM
Deleted: about 50

Shantao 2/18/2017 7:38 PM
Deleted: 75

Shantao 2/18/2017 7:38 PM
Deleted: 12

Shantao 2/18/2017 7:25 PM
Deleted: 5

Shantao 2/18/2017 9:16 PM
Formatted: Font:Bold

Shantao 2/18/2017 9:16 PM
Formatted: Font:Bold

With the richness of SNVs in WGS samples, we inferred 35 individual evolution trees (Figure SXX). Three trees have a largest population faction <0.5 (likely due to low mutation number, high sequence error and/or high heterogeneity) and thus excluded from downstream analysis. We could further classify the trees into four types based on topology (Figure 4A, 4B): no branch, less subclones (10, 32.3%), short branches (12, 37.5%), no branch, more subclones (5, 15.6%) and long branches (5, 15.6%).

Short branch type is significantly enriched in Type I pRCC (p<0.011, two-tailed fisher exact test) while the more heterogeneous types: long branches and no branch, more subclones type are significantly depleted in Type I (p < 0.0034, two-tailed fisher exact test). This indicates type I tumors are more homogenous and show less complex evolution features compared to type II and unclassified samples.

Shantao 2/17/2017 6:50 PM

Deleted: However, a non-parametric permutation Kolmogorov–Smirnov test (see Methods) failed to detect a statistical significance (p > 0.05), likely because of the small number of samples and the prudence of our conserved test.

Discussion

We comprehensively analyzed both WGS and an extensive set of WXS of pRCC, scrutinizing local high-impact events as well as giving a macro overlook of the mutation landscape. Our work further completed the genomic alteration landscape of pRCC (Figure 4B). Beyond traditionally driver events, we suggested several novel noncoding alterations potentially drive tumorigenesis.

First, we elaborated on previous results of the long known driver *MET*. In an extended 117 WXS dataset, we found six additional nonsynonymous somatic mutations in the hypermutated tyrosine kinase catalytic domain. These somatic mutations are highly recurrent, concentrated on a few critical amino acids. This is in line with *MET* being an oncogene and supports the central role of *MET* in pRCC. Then we found an exonic SNP in *MET*, rs11762213, to be a prognostic germline variance in type 2 pRCC. Previously, rs11762213 was found to predict outcome in a mixed RCC samples, predominated by ccRCC (14). Later, the result is confirmed in a large ccRCC cohort (9). However, it is never clear whether rs11762213 only predicts the outcome in ccRCC or other histological types as well. In this study, we concluded that the minor alternative allele of rs11762213 also forecasts unfavorable outcome in type 2 pRCC patients. The mechanism of this exonic germline SNP remains unsettled. A previous study proposes it disrupt a putative enhancer and thus affect *MET* expression. However, researchers cannot find significant difference in *MET* expression in either tumor or normal tissues. We noticed there is other gene within 100 kb of this SNP. Given the significant role of *MET* in pRCC, we also think rs11762213 is affecting survival through *MET*, although the mechanism unknown.

Remarkably, similar to ccRCC, type 2 pRCC is not primarily driven by *MET*. Not as significantly mutated in ccRCC and type 2 pRCC, *MET* nonetheless seems to play a role in cancer development. ccRCC responses to *MET* inhibitors (REF). This finding is potentially meaningful in clinical management of patients with the more aggressive type 2 pRCC. rs11762213 genotyping could become a reliable, low-cost risk stratification tool for these patients. Also, rs11762213 might become a biomarker for predicting patient response to MET inhibitors.

Interestingly, rs11762213 is prevalent mostly in European and American populations but not in African populations and rare in populations in Asia. MAF of rs11762213 among African American patients in our cohort is 2.73%, higher than MAFs in general African populations observed in 1000 Genome phase 3 dataset (0.2%, 0% in Americans with African ancestry

Shantao 2/18/2017 9:07 PM
Formatted: Font:Italic

Shantao 2/18/2017 9:08 PM
Formatted: Font:Italic

Shantao 2/18/2017 9:11 PM
Formatted: Font:Italic

Shantao 2/18/2017 9:12 PM
Formatted: Font:Italic

Shantao 1/24/2017 1:54 AM
Deleted: central

Shantao 2/18/2017 9:09 PM
Formatted: Font:Italic

Shantao 1/24/2017 1:55 AM
Deleted: of

Shantao 1/24/2017 1:55 AM
Deleted: Potentially

Shantao 2/18/2017 7:45 PM
Deleted: , patients with

Shantao 1/24/2017 2:32 AM
Deleted: also benefit from

(ASW))) and the ExAC dataset (1.1%, excluding TCGA cohorts). This implies a possible effect of rs11762213 on pRCC incidence among African Americans that is worth further investigation.

Shantao 1/24/2017 2:11 AM
Deleted: Perhaps this variant could play a role in the significant racial disparities are known to exist in the overall incidence, histologic distribution, and survival of African Americans with kidney cancer.

Besides, in *MET* non-coding regions, we first find a cryptic promoter from a retrotransposon in the second intron initiates the alternative splicing event, which is classified as a driver event by the TCGA study (3). Methylation is a major source of silencing retrotransposon activities in human genome (REF). Indeed, we observed evidence for a local loss of methylation and global methylation dysregulation in samples expressing alternative isoforms. Therefore, we showed methylation change might drive pRCC growth through MET pathway.

We also discovered mutations associated with *MET* promoter and first two introns. Although the implication is unknown, our analysis suggests there is a mutation hotspot in *MET* that calls for further research.

Shantao 2/18/2017 7:50 PM
Deleted: w

Expanding our scope from coding to non-coding and use FunSeq to group SNVs by functional elements, we found several potentially significant non-coding mutation hotspots relevant to tumorigenesis throughout the entire genome. A mutation hotspot was found upstream of *ERRFI1*, an important regulator of the EGFR pathway, which may serve as a potential tumor suppressor. EGFR inhibitors have been used in papillary kidney cancer with an 11% response rate observed (22). These mutations potentially disrupt regulatory elements of *ERRFI1* and thus play a role in tumorigenesis. However, likely limited by a small sample size, we were not able to detect statistically significant functional changes in ERRFI1 and related pathways. Another non-coding hotpot is in *NEAT1*, a long non-coding RNA that has been speculated to involve in cancer. All mutations locate in a putative regulatory region of the gene. Patients carrying mutations in *NEAT1* have significantly higher *NEAT1* expression and worse prognosis. High expression of *NEAT1* predicts significantly worse survival in ccRCC as well. *NEAT1* has been shown to be hypermutated in other cancers and some studies also linked high *NEAT1* association with unfavorable prognosis in several other tumors (23-24). Last, a downstream lncRNA, *MALAT1*, shows tight co-expression pattern with *NEAT1* in both pRCC and ccRCC. *MALAT1* is in COSMIC consensus cancer gene list and annotated as related with pediatric RCCs.

WGS provides many times more SNVs compared to WXS, and noncoding SNSs are less constrains by selection pressure. Thus it gives us a great opportunity to look into the high-level landscape of mutations in pRCC. We identified mutation rate dispersion of C-to-T in CpG motif contributes the most to the inter-sample mutation spectra variations. We further pinned down the cause of dispersion by showing the hypermethylated cluster, identified in the previous TCGA study (3), has higher C-to-T rate in CpGs. This hypermethylated cluster is associated with later stage, type 2 pRCC, *SETD2* mutation and worse prognosis (3). Although increased C-to-T in CpG is likely the result of hypermethylation, we cannot rule out the possibility the change of mutation landscape plays a role in cancer development. For example, C-to-T in methylated CpG causes loss of methylation, which could have effects on local chromatin environment, trans-elements recruitment and gene expression regulation. In our study, we observed C-to-Ts in CpG are enriched in coding regions, which indicates they have higher functional impacts in cancer genome.

Significant APOBEC activities and consequential mutation signatures were observed in one type 2 pRCC case. APOBEC activities were known to be prevalent in UCs (12, 19). We also successfully detected prominent APOBEC signatures in all three UC samples processed in the same pipeline as pRCCs. Intriguingly, despite being considered to have the same cellular origin with pRCC, we were not able to detect significant APOBEC activities in ccRCC. This is in

Shantao 2/18/2017 8:02 PM
Formatted: Font:Italic

Shantao 2/18/2017 8:03 PM
Formatted: Font:Italic

Shantao 2/18/2017 8:03 PM
Formatted: Font:Italic

Shantao 2/18/2017 8:04 PM
Formatted: Font:Italic

Shantao 2/18/2017 8:20 PM
Deleted: Last, focusing on the high-level landscape of mutations in pRCC, w

agreement with previous studies (12). APOBEC mutation signature was also found in a small percentage of chromophobe renal cell carcinoma (25), although they are believed to have a different cellular origin. APOBEC activities have been linked with genetic predisposition and viral infection (26). Given a statistically robust signal in our conservative algorithm, it is plausible that a small fraction of otherwise driver mutation absent type 2 pRCCs might share some etiologically and gnomically similarity with UC. Standard treatment for UC involves cytotoxic chemotherapy and radiation while RCC shows low response rate to cytotoxic therapy. Pending further research, this finding might lead to actionably clinical implications. (still too strong?)

Chromatin remodeling pathway is highly mutated in pRCC (3). Several chromatin remodelers, for example *SETD2* and *PBRM1*, have been identified as cancer drivers in pRCC. We investigate the relationship between samples with mutated chromatin remodelers and those without such mutations in terms of overall mutational spectrum. We demonstrated pRCC with defects in chromatin remodeling genes shows higher mutation rate in general, driven by an even stronger mutation rate increase in putative open chromatin regions in normal kidney tissues. This is likely because chromatin remodeling defects affect normal open chromatin environment and impede DNA repairing in these regions.

It is known that replication time strongly governs local mutation rate. Early replication regions have fewer mutations. But the difference dissipates when DNA mismatch repair becomes defective (21). In our study, we found this correlation weakened in chromatin remodeling genes mutated samples, presumably caused by failure of replication error repair in an abnormal chromatin environment. By adapting defects in chromatin remodeling genes, tumor alters its mutation rate and landscape, which might further provide advantage in cancer evolution. Yet,

Shantao 1/24/2017 2:18 AM
Deleted: b

Shantao 1/24/2017 2:18 AM
Deleted: e

Shantao 1/24/2017 2:18 AM
Deleted: genomically

Shantao 1/24/2017 2:19 AM
Deleted: to

Shantao 1/24/2017 2:19 AM
Deleted: ince s

Shantao 1/24/2017 2:22 AM
Formatted: Highlight

Shantao 1/24/2017 2:19 AM
Deleted: ,

Shantao 1/24/2017 2:19 AM
Deleted: this

Shantao 1/24/2017 2:20 AM
Deleted: could have

Shantao 1/24/2017 2:20 AM
Deleted: a very

Shantao 1/24/2017 2:20 AM
Deleted: meaningful clinical impact.

Shantao 2/18/2017 8:05 PM
Deleted: , *BAP1*

Shantao 1/24/2017 2:24 AM
Deleted: due to

high mutation burden in functional important open chromatin regions also raises the chance that tumor antigens activate host immune system. Researchers found tumors with DNA mismatch repair deficiency response better to PD-1 blockage (27), while these tumors also accumulates more mutations in early replicated regions (21). Thus chromatin remodeler alterations might as well correlate with higher response rate of immunotherapy,

In this first whole genome study of pRCC, we found several novel non-coding alterations that might have meaningful clinical impacts. However, due to a limited sample size, our statistical tests were underpowered. As the cost of sequencing keeps dropping, we expect to have more pRCC whole genome sequenced in the near future (28). With a larger cohort, we hope to gain enough power to test the hypotheses we formed as well as further explore the noncoding regions of pRCC.

Materials and Methods

Data acquisition

We downloaded pRCC and ccRCC WXS and pRCC WGS variation calls from TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp) and TCGA Jamboree respectively. pRCC RNAseq, RPPA and methylation data were downloaded from TCGA Data Portal as well. Repli-seq and DHS data were obtained from ENCODE (https://www.encodeproject.org/).

Testing rs11762213 on prognosis and exploring somatic mutations in *MET*

We downloaded pRCC clinical outcomes from TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp). pRCC samples that failed the histopathological review were excluded (3). In total, we included 277 patients in our analyses (Figure S1, Table S1). For germline calls, the majority of samples, 163 out of 277, were supported by SNV callings from at least two centers (102 from three centers). 100% genotype concordance rate was observed. Also, 162 curated rs11762213 genotypes were in agreement with automated callsets. With proved high confidence in accuracy of genotyping rs11762213 in germline, we recruited additional 114 samples from single-center (BCM), automated calls to form an extensive patients set (Figure S1). For somatic SNVs in *MET*, after excluding cases that were recruited in the TCGA study, we formed an additional set encompassing 117 patients. Five callings were supported by two centers. The rest were supported by single-center (BCM) automated calls.

Cancer-specific survival was defined using the same criteria as described in a ccRCC study (9). Deaths were considered as cancer-specific if the “Personal Neoplasm Cancer Status” is “With Tumor”. If “Tumor Status” is not available, then the deceased patients were classified as cancer-specific death if they had metastasis (M1) or lymp node involvement (>= N1) or died within two years of diagnosis. An R package, “survival”, was used for the survival analysis.

SV calling procedure

We remapped the reads using bwa 0.7.12, which support split read mapping. Then we used DELLY (10) with default parameters for somatic SV calling. To avoid sample contamination or germline SVs, we filtered our callsets against the entire TCGA pRCC WGS dataset, regardless of sample match or pathological reviews. We discharge all callings that were marked “LowQual” (PE/SR support below 3 or mapping quality below 20). Last, to further

Shantao 2/17/2017 7:05 PM
Deleted: **precedure**

Shantao 2/17/2017 7:07 PM
Formatted: Indent: First line: 0"

Shantao 2/17/2017 7:07 PM
Deleted: ... [1]

Shantao 2/17/2017 7:07 PM
Deleted: 2

Shantao 2/17/2017 7:08 PM
Deleted: Lastly, we

eliminate germline contamination, we filtered out SVs that show at least 0.8 reciprocally overlapping with 1000 Genome Phase III SV callsets (only 1/425 found).

Shantao 2/17/2017 7:07 PM
Deleted: .

Mutation spectra study

WGS Mutations were extracted from flanking 5’ and 3’ nucleotide context. The raw mutation counts were normalized by trinucleotide frequencies in the whole genome.

To identify signatures in the mutation spectra, we used a robust, objective LASSO-based method. First, 30 known signatures were downloaded from COSMIC (http://cancer.sanger.ac.uk/cosmic/signatures). Then we solve a positive, zero-intercept linear regression problem with L1 regularizer to obtain signatures and corresponding weights for each genome. Specifically, we solve the problem:

$$\min_{W}(\|SW - M\|_2 + \lambda\|W\|)$$

Where M is the mutation matrix, containing the mutations of each sample in 96 nucleotide contexts. S is the 96×30 signature matrix, representing the mutation probability in 96 nucleotide contexts of the 30 signatures. W is the weighting matrix, representing the contribution of 30 signatures to each sample.

The penalty parameter lambda (λ) was determined empirically using 10-fold cross-validation individually for every sample. λ was chosen to maximize sparsity and constrained to keep mean-square error (MSE) within one standard error of its minimum. Last, we discharged signatures that composite less than 5% of the total detectable signatures.

Methylation association analysis

In total, we collected HumanMethylation450 BeadChip array data for 139 samples that are either methylation cluster 1 or 2. We used an R package "IMA" to facilitate analysis (11). After discharging sites with missing values or on sex chromosomes, we obtained beta-values on 366,158 CpG sites in total. Then we test beta-values of each site by Wilcoxon rank sum test between two methylation clusters. After adjusting p-value using Benjamini-Hochberg procedure, we called 9,324(2.55%) hypermethylation sites. These sites have an adjusted p-value of less than 0.05 and mean beta-values in methylation cluster 1 are 0.2 or higher than the ones in methylation cluster 2.

APOBEC enrichment analysis

We used the method described by Roberts et al. (12). For every C>{T,G} and G>{A,C} mutation we obtained 20bp sequence both upstream and downstream. Then enrichment fold was defined as:

$$Enrichment\ Fold = \frac{Mutation_{\mathrm{TCW/WGA}} \times Context_{C/G}}{Mutation_{C/G} \times Context_{TCW/WGA}}$$

Here TCW/WGA stands for T[C>{T,G}]W and W[G>{A,C}]A. W stands for A or T. p-value for enrichment were calculated using one-side Fisher-exact test. To adjust for multiple hypothesis testing, p-values were corrected using Benjamini-Hochberg procedure.

WXS data for APOBEC enrichment and signature analysis was obtained from a high quality somatic callset: hgsc.bcm.edu_KIRP.IlluminaGA_DNASeq.1.protected.maf. This dataset includes 155 pRCC samples and three UC samples. We use hgsc.bcm.edu_KIRC.Mixed_DNASeq.1.protected.maf for ccRCC analyses.

Chromatin remodeling genes and replication time association

We identified chromatin remodeling genes based on its significance in pRCC and function. Our gene list is the intersection of gene lists in the original TCGA pRCC study molecular feature table (supplementary table 3) with the chromatin remodeling and SNI/SWF pathway gene lists (supplementary table 4). Our gene set include ten genes: *SETD2, KDM6A, PBRM1, SMARCB1, ARID1A, ARID2, MLL2 (KMT2D), MLL3(KMT2C), MLL4(KMT2B), EP300.* We found adding BAP1 into the list won't change the significance of our tests. We defined chromatin remodeling defect as nonsynonymous mutations in these genes. For missense mutations, we filtered out mutations with polyphen score less then 0.8 (benign).

In order to avoid cell type redundancy, we only kept GM12878 as the representative of all lymphoblastoid cell lines. Eleven cell types were included in our analysis: BG02ES, BJ, GM12878, HeLaS3, HEPG2, HUVEC, IMR90, K562, MCF7, NHEK, SK-NSH. Wave smoothed replication time signal was averaged in a ±10kb region from every mutation. To avoid potential selection effects, we removed mutations in exome and flanking 2bp. Regions overlap with reference genome gaps and DAC blacklist (https://genome.ucsc.edu/) were removed as well. Last, we picked the median number from 11 cell types at each mutation position for further analysis.

To test the significance of replication time of non-coding mutations between two groups, we assigned all the mutation with its local replication time and then defined the ones stand above

Shantao 2/17/2017 7:03 PM
Deleted:

Shantao 2/17/2017 7:02 PM
Formatted: Indent: First line: 0.5"

Shantao 2/17/2017 7:05 PM
Formatted: Font:Italic

Shantao 2/17/2017 7:03 PM
Formatted: Highlight

Shantao 2/17/2017 7:01 PM
Deleted: included eleven genes. They are *ARID1A*, *ARID2*, *BAP1*, *DNMT3A*, *KDM6A*, *MLL2*, *MLL3*, *MLL4*, *PBRM1*, *SETD2*, *SMARCB1*.

Shantao 2/17/2017 7:03 PM
Formatted: Highlight

Shantao 2/17/2017 6:51 PM
Deleted: we adapted a conservative non-parametric Kolmogorov–Smirnov test (K-S test) using empirical p-value.

Shantao 2/17/2017 6:51 PM
Deleted: W

90 percentile in all pooled mutations as “mutations in early replicated regions”. Then we calculate the percentage of “mutations in early replicated regions” in total mutations for each sample and compare between two groups using rank-sum test.

Evolution tree inference:

We use PhyloWGS (REF) to infer the evolution trees for each individual tumor. To mitigate the effects on copy number change, we removed all the SNVs inside the copy number change regions as defined by assay-based method in the original TCGA study (REF). To be prudent, we defined any region with an absolute log tumor copy number to normal ratio larger than 0.3. Last, we removed all SNVs with allele frequency higher than 0.6 as they are likely affected by copy number loss.

Author contributions: SL, BMS and MG conceived and designed the study. SL carried out the computation and data analysis, SL, BMS and MG interpreted the results. SL wrote the manuscript. BMS and MG co-directed this work. All authors have read and approved the final manuscript. **Competing interests**: The authors declare no competing interests.

Acknowledgments: This work was supported by the National Institutes of Health, AL Williams Professorship, and in part by the facilities and staffs of the Yale University Faculty of Arts and Sciences High Performance Computing Center. We thank Patrick Mcgillivray for his help in manuscript preparation.

References

Shantao 2/17/2017 6:52 PM
Deleted: its percentile among all mutations replication time shifted ±100kb from the origin (represents the background replication time)

Shantao 2/17/2017 6:52 PM
Deleted: K-S test

Shantao 2/17/2017 6:54 PM
Deleted: statistics

Shantao 2/17/2017 6:54 PM
Deleted: in

Shantao 2/17/2017 6:54 PM
Deleted: and compare

Shantao 2/17/2017 6:52 PM
Deleted: To obtain the empirical p-value, we randomly permutated the chromatin remodeling genes mutation labels for 1,000 times to estimate the test statistics distribution under null hypothesis.

Shantao 2/18/2017 8:10 PM
Formatted: Font:Bold

1. Siegel, R, Naishadham, D, Jemal, A. Cancer statistics, 2015. CA: a cancer journal for clinicians. 2015; 65(1*)*, 5-29.
2. Shuch B, Amin A, Armstrong AJ, Eble JN, Ficarra V, Lopez-Beltran A, et al. Understanding pathologic variants of renal cell carcinoma: distilling therapeutic opportunities from biologic complexity. European urology. 2015;67(1):85-97.
3. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of papillary renal-cell carcinoma. N Engl J Med. 2016;2016(374):135-45.
4. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013;342(6154):1235587.
5. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome biology. 2014;15(10):1.
6. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. Science. 2013;339(6122):957-9.
7. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell reports. 2013;3(1):246-59.
8. Alexandrov LB, Nik-Zainal S, Siu HC, Leung SY, Stratton MR. A mutational signature in gastric cancer suggests therapeutic strategies. Nature communications. 2015;6.
9. Hakimi AA, Ostrovnaya I, Jacobsen A, Susztak K, Coleman JA, Russo P, et al. Validation and genomic interrogation of the MET variant rs11762213 as a predictor of adverse outcomes in clear cell renal cell carcinoma. Cancer. 2016;122(3):402-10.

10. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333-9.

11. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, et al. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics. 2012;28(5):729-30.

12. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nature genetics. 2013;45(9):970-6.

13. Schmidt L, Junker K, Weirich G, Glenn G, Choyke P, Lubensky I, et al. Two North American families with hereditary papillary renal carcinoma and identical novel mutations in the MET proto-oncogene. Cancer research. 1998;58(8):1719-22.

14. Schutz FA, Pomerantz MM, Gray KP, Atkins MB, Rosenberg JE, Hirsch MS, et al. Single nucleotide polymorphisms and risk of recurrence of renal-cell carcinoma: a cohort study. The lancet oncology. 2013;14(1):81-7.

15. Guo S, Chen W, Luo Y, Ren F, Zhong T, Rong M, et al. Clinical implication of long non-coding RNA NEAT1 expression in hepatocellular carcinoma patients. International journal of clinical and experimental pathology. 2015;8(5):5395.

16. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016;534(7605):47-54.

17. Choudhry H, Albukhari A, Morotti M, Haider S, Moralli D, Smythies J, et al. Tumor hypoxia induces nuclear paraspeckle formation through HIF-2α dependent transcriptional activation of NEAT1 leading to cancer cell survival. Oncogene. 2015;34(34):4482-90.

18. Chakravarty D, Sboner A, Nair SS, Giannopoulou E, Li R, Hennig S, et al. The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. Nature communications. 2014;5.

19. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507(7492):315-22.

20. Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, et al. Integrated molecular analysis of clear-cell renal cell carcinoma. Nature genetics. 2013;45(8):860-7.

21. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015;521(7550):81-4.

22. Gordon MS, Hussey M, Nagle RB, Lara PN, Mack PC, Dutcher J, et al. Phase II study of erlotinib in patients with locally advanced or metastatic papillary histology renal cell cancer: SWOG S0317. Journal of Clinical Oncology. 2009;27(34):5788-93.

23. Li Y, Li Y, Chen W, He F, Tan Z, Zheng J, et al. NEAT expression is associated with tumor recurrence and unfavorable prognosis in colorectal cancer. Oncotarget. 2015;6(29):27641.

24. He C, Jiang B, Ma J, Li Q. Aberrant NEAT1 expression is associated with clinical outcome in high grade glioma patients. Apmis. 2016;124(3):169-74.

25. Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer cell. 2014;26(3):319-30.

26. Henderson S, Chakravarthy A, Su X, Boshoff C, Fenton TR. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. Cell reports. 2014;7(6):1833-41.

27. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 blockade in tumors with mismatch-repair deficiency. New England Journal of Medicine. 2015;372(26):2509-20.

28. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome biology. 2016;17(1):1.

Figure 1. MET noncoding alterations and Survival analysis of rs11762213 in pRCC patients.

(A) A schematics diagram of non-coding mutations on *MET*. The germline SNP, rs11762213, is also shown. Thin black lines indicate retrotransponson initiated alternative isoform.

(B) Genotypes are shown in the legend. Peto & Peto modification of the Gehan-Wilcoxon test.

Figure 2. Noncoding alterations in pRCC.

(A) A schematics diagram of non-coding mutations on *ERRFI1*. (B) A schematics diagram of non-coding mutations on *NEAT1*. One tumor carries two mutations on *NEAT1*. (C) Tumors with mutations on *NEAT1* show higher *NEAT1* expression. (D) Survival analysis shows mutations in *NEAT1* are associated with worse prognosis. To avoid potential confounding effects, we removed one subject who carries rs11762213 but not *NEAT1* mutation. Log-rank test.

Figure 3. **Mutation spectra and mutation processes in pRCC.**

(A) The mutation spectrum of all pRCC WGS samples. Mutations are ordered in alphabetical order of the reference trinucleotides (with the mutated nucleotide in the middle, from A[C>A]A to T[T>G]T) from left to right. Then we use PCA to maximize inter-sample variation. The loadings on the first principle component is strongly dominated by C>T in CpGs. (B) PC1, along with C>T in CpGs mutation counts and the fractions of such mutations among total mutations are significantly different between two methylation groups. (C) APOBEC mutation signatures are shown for both pRCC (along with three UC sampels, which have blue outer circles) and ccRCC TCGA cohorts. Red dashed line represents the median APOBEC enrichment. (D) Comparison of total mutation counts, mutations counts in open chromatin regions and percentages of mutations in open chromatin regions of total mutations between tumors with chromatin remodeling genes alterations and the ones without.

Figure 4. **Evolution trees and genomic alteration landscape of 35 whole genome sequenced pRCC samples.**

(A) Two individual evolutions trees. Cancer related mutations firstly appear in each population is marked by corresponding colors (B) Index: patient index, see Table S2

Shantao 2/18/2017 9:59 PM
Deleted: (A) A schematics diagram of non-coding mutations on *MET*. The germline SNP, rs11762213, is also shown.

Shantao 2/18/2017 10:00 PM
Deleted: B

Shantao 2/18/2017 10:00 PM
Deleted: C

Shantao 2/18/2017 10:00 PM
Deleted: D

Shantao 2/18/2017 10:00 PM
Deleted: E

Shantao 2/18/2017 10:00 PM
Deleted: (B)

Shantao 2/18/2017 10:00 PM
Deleted: W

Shantao 2/18/2017 10:00 PM
Deleted: C

Shantao 2/18/2017 10:00 PM
Deleted: D

Shantao 2/18/2017 10:00 PM
Deleted: E

Shantao 2/18/2017 10:01 PM
Deleted: **The**

Shantao 2/18/2017 10:01 PM
Deleted: **32**

Shantao 2/18/2017 10:01 PM
Deleted: Grey cells represent genomic alterations. CN: copy number.