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,*,#]

Abstract: Papillary renal cell carcinoma (pRCC) constitutes 10-15% of kidney tumors. Recent advances in DNA sequencing have significantly deepened our understanding of the molecular genetics of this disease. However, much of this work has been limited to coding alterations in traditional cancer driver genes, especially *MET*. Moreover, despite identifying specific sub-groups of alterations, researchers cannot find a clear molecular etiology for a significant proportion of the tumors. To address this, we carry out the first whole genome analysis of pRCC. We take a comprehensive approach, trying to explain the cases lacking classic drivers. First, we elaborate on these results on *MET* and we find more somatic mutations in this gene and we find a *MET* germline SNP (rs11762213) that predicts prognosis in pRCC. Interestingly, we find no enrichment for structural variants to be associated with *MET*, perhaps consistent with it having an oncogenic as opposed to tumor suppressor role. Next, we analyze mutations in the non-coding regions throughout the entire genome. We discover several potentially impactful hotspots, including one in non-coding regions near *MET*. Another hotspot is inside a long non-coding RNA that has been implicated in cancer, *NEAT1*. The *NEAT1* mutations, moreover, are associated with increased expression and worse cancer-specific survival. Finally, we analyze genome-wide mutational patterns. We notice that these are dictated mostly by the prevalence of methylation-associated C-to-T transitions. Also, we observe significantly more mutations in open chromatin regions in tumors with chromatin modifier alterations.

Shantao 8/14/2016 6:56 PM
Comment [1]: To be precise, it's introns and promoters
Use "of" instead of "near"?

[[150 words version (Cell Report), no "Significance"]]

Summary: Papillary renal cell carcinoma (pRCC) constitutes 10-15% of kidney tumors. However, previous pRCC genomes studies have been largely limited to coding alterations in traditional cancer drivers, especially *MET*. Moreover, despite identifying sub-groups of alterations, researchers cannot find clear molecular etiology for a significant proportion of the tumors. Therefore, we perform the first whole genome analysis of pRCC. First, we elaborate on previous results on *MET* and we find a germline SNP (rs11762213) predicts prognosis. Next, we scrutinize non-coding mutations throughout the entire genome. We discover potentially impactful hotspots around functional elements, including non-coding regions of *MET* and a long non-coding RNA that has been implicated in cancer, *NEAT1*. The *NEAT1* mutations are associated with increased expression and unfavorable outcome. Finally, we investigate genome-wide mutational patterns. We notice they are dictated mostly by methylation-associated C-to-T transitions. Also, we observe significantly more mutations in open chromatin in tumors with chromatin modifier alterations. (150 words)

Keywords: Papillary renal cell carcinoma, whole genome, non-coding alternations

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

Highlights [[up to four bullet points, max. 85 char. each, core findings]]

- A germline SNP in *MET*, rs11762213, predicts prognosis in pRCC
- Mutations in NEAT1 correlate with high expression and worse outcome in pRCC
- Chromatin modifier alterations correlate with more mutation in open chromatin
- Non-coding alterations analyses helps reveal molecular etiology in pRCC

eTOC Blurb [[max 50 words, third person]]

Li et al. preformed the first comprehensive pRCC whole genome analysis. Besides finding a germline SNP predicts prognosis, they discover a significant amount of potential impactful alternations and several mutation patterns in non-coding regions. These findings help explain molecular etiology of the disease and further complete pRCC genomic alteration landscape. (50 words)

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. 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 tumorgenesis. 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 "driver-unknown". 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, has 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.

Looking at the mutations at a higher level, multiple endogenous and environmental mutation processes shape the somatic mutational landscape observed in cancers (*7*). Analyses of the associated genomic alterations give information of 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 effect of exome capture process and driver selection.

In this study, we comprehensively analyzed 32 pRCC cases that were whole genome sequenced along with an extensive set of WXS data in 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 play a central role in pRCC, especially in Type 1. For the first time we found rs11762213, a germline exonic single nucleotide polymorphism inside *MET*, predicts cancer-specific survival (CSS) in pRCC. We also ~~discovered several potentially impactful non-coding mutation hotspots around *MET* promoter and its first two exons.~~ Surprisingly, we did not find a significant amount of structural variations affecting *MET* besides polysomy 3. Then we went onto cases not as easily explained as those with MET alterations. We analyzed nearly 150,000 non-coding mutations through out 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. Methylation is the leading factor influencing mutation landscape. Methylation status drives the intra-sample mutation variation by giving rise to 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. Last, we found samples with chromatin remodeler alternations accumulate more mutations in open chromatin regions.

Results

1. Probing an exonic SNP in MET, rs11762213, in pRCC prognosis.

We begin with *MET* coding variants. The TCGA study of 161 pRCC patients found 15 samples carrying somatic, nonsynonymous single nucleotide variant (SNV) in MET. By analyzing 45 extra WXS samples (see Methods), we found two more nonsynonymous somatic mutations, H1112Y and Y1248C. Both mutations occur in the hypermutated tyrosine kinase catalytic domain of *MET*. H1112Y has been observed in two patients the original TCGA study cohort. Y1248C has been observed in Type 1 pRCC before (REF) and TCGA cohort has a case carrying Y1248H. The sample with Y1248C is a Type 1 pRCC case while the subtype for the H111Y sample is unknown.

Although many MET somatic mutations are believed to play a central role in pRCC, a germline SNP, rs11762213, has been discovered to predict recurrence and survival in a RCC cohort (*9*). ccRCC predominated the initial discovery RCC cohort. This conclusion was later validated in a ccRCC cohort but never in pRCC (*10*). We evaluated whether this SNP has a prognostic effect in pRCC. Using an extensive WXS set of 207 patients, we found 12 patients carry one risk allele of rs11762213 (G/A, Table 1). No homozygous A/A was observed. The cancer-specific survival is statistically significantly worse in patients with the risk allele ($p < 0.037$, Peto & Peto modification of the Gehan-Wilcoxon test; $p < 0.044$, log-rank test, Fig 1).

Genotype (rs11762213)	G/A (n = 12)	G/G (n = 195)
Sex, No. (%)		
Male (%)	7 (58)	142 (73)
Female (%)	5 (42)	53 (27)
Age, median (IQR), y	53 (46-59.5)	60 (53-69)
Race, No. (%)		
White	9 (75)	130 (67)
Black or African American	3 (25)	47 (24)

Shantao 8/14/2016 9:26 PM
Deleted: validated

Shantao 8/14/2016 9:27 PM
Deleted: as a predictive SNP for

Shantao 8/14/2016 9:27 PM
Deleted: found

Shantao 8/14/2016 9:27 PM
Deleted: alternations

Shantao 8/14/2016 9:30 PM
Formatted: Font:Italic

Shantao 8/14/2016 9:29 PM
Deleted: >

Shantao 8/14/2016 9:34 PM
Deleted: pRCC

Shantao 8/14/2016 9:35 PM
Deleted: including

Shantao 8/14/2016 9:36 PM
Deleted: Y1248C and

Shantao 8/14/2016 9:36 PM
Formatted: Font:Italic

Shantao 8/14/2016 9:36 PM
Deleted: two recurrences

Shantao 8/14/2016 9:37 PM
Deleted: found

Shantao 8/14/2016 9:37 PM
Deleted: of

Shantao 8/14/2016 9:38 PM
Deleted: carries

Shantao 8/7/2016 10:00 PM
Comment [2]: We mostly want to show an overview table for 32 WGS…An overview for WXS (161 samples) was in the NEJM paper.

But I couldn't find a perfect place to insert an overview for WGS. Also it means adding an extra table/figure/subfigure

Shantao 8/15/2016 12:47 AM
Deleted: (*9*).

Shantao 8/15/2016 12:47 AM
Deleted:

Shantao 8/15/2016 12:48 AM
Deleted: (see Methods)

Shantao 8/15/2016 12:49 AM
Comment [3]: Combining table 1 with figure 1?

American Indian or Alaska native	0	2 (1)
Asian	0	5 (3)
NA	0	11 (6)
Tumor type, No. (%)		
Type 1	4 (33)	86 (44)
Type2	7 (58)	66 (34)
Unclassified	1 (8)	25 (13)
Not centrally reviewed	0	18 (9)
T stage, No. (%)		
T1	7 (58)	132 (68)
T2	2 (17)	18 (9)
T3	3 (25)	43 (22)
T4	0	2 (1)
N stage, No. (%)		
N0	5 (54)	32 (16)
N1	0	18 (9)
N2	1 (8)	2 (1)
NX	6 (50)	143 (74)
M stage, No. (%)		
M0	5 (42)	74 (38)
M1	1 (8)	6 (3)
MX/NA	6 (50)	115 (59)
AJCC stage, No. (%)		
I	7 (58)	125 (64)
II	1 (8)	11 (6)
III	2 (17)	39 (20)
IV	2 (17)	10 (5)
NA	0	10 (5)
Median follow-up for surviving patients, days (IQR)	266 (118-643)	493 (167-884)

AJCC: American Joint Committee on Cancer; IQR: interquartile range
Due to rounding, the percentages may not add up to one

Table 1. Clinical characteristics of subjects in pRCC The Cancer Genome Atlas Cohort

2. Mutation hotspots in non-coding region

Despite the fact *MET* is the most common driver alteration, about 20% presumably *MET*-driven yet *MET* wild-type pRCC samples were 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 (Fig 2A). This sample shows no evidence of a nonsynonymous mutation in *MET* gene but it has copy number gain of *MET*. Additionally, we observed 6/32 (18.8%) samples carry mutations in the intronic regions between exon 1-3 of *MET* (Fig 2A). Previously it is been established that alternative splicing of these exons is a driver event. Therefore we speculate that these non-coding variance might correlate with the alternative splicing. However, likely being hindered by a small size, we were not able to find statistically significant association between alternative splicing event and these intronic mutations.

We further expanded our scope and ran FunSeq (5) to identify potentially high-impact, non-coding variants in pRCC. First, we identified a high-impact mutation hotspot on chromosome 1. 6/32 (18.8%) samples have mutations within this 6.5kb region (Fig 2B). 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 associated with cancer. Due to a very limited sample size here, our test power was inevitably low. We didn't observe statistically significant

Shantao 8/15/2016 12:50 AM
Deleted: ha

Shantao 8/15/2016 12:51 AM
Deleted: t

Shantao 8/15/2016 12:52 AM
Deleted: be

Shantao 8/15/2016 12:52 AM
Deleted: the source

Shantao 8/15/2016 12:52 AM
Deleted: of

Shantao 8/15/2016 12:52 AM
Deleted: s

changes among mutated samples in mRNA expression level, protein level and phosphorylation level of EGFR, HER2 and HER3 (S1-S3).

Another potentially impactful mutation hotspot is in *NEAT1*. We saw mutations inside this nuclear long non-coding RNA in 5/32(15.6%) samples (FIG 2C). Several studies indicated *NEAT1* is associated in many other cancers (*11, 12*). It promotes cell proliferation in hypoxia (13) and alters the epigenetic landscape, increasing transcription of target genes (*14*).

All the mutations we found fell into a putative promoter region of *NEAT1*. We noticed *NEAT1* mutations were associated with higher *NEAT1* expression (Fig 2D, $p < 0.044$, two-sided rank sum test). We also found *NEAT1* mutations were associated with worse prognosis (Fig 2E, $p < 0.022$, log-rank test).

We used DELLY2 (REF) to perform structural variants (SVs) calling from WGS reads information (Supplements). 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 about 500 somatic SV events, includes deletions, duplications, inversions and translocations. We confirmed three cases carrying deletions affecting *CKDN2A*. One sample, TCGA-B9-4116, which had extensive amplification of *MET*, showed multiple SVs of various classes hitting *MET* regions. However, surprisingly, we did not find SVs affecting *MET* except this one example. We postulate trisomy or polysomy 3 might be 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 of this oncogene.

MORE

3. Mutation spectra of pRCC

To further get a high-order overview of the mutation landscape, we summarized the mutation spectra of 32 whole genome sequenced pRCC samples (Fig 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 context.

We used principle components analysis (PCA) to reveal factors that explain the most inter-sample variation. The loadings on the first principle component (which explains 12.5% of the variation) demonstrated C-to-T in CpGs contributes the most to inter-sample variation (Fig 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. So we further explored the association between C-to-T in CpGs and tumor methylation status. We confirmed this association by showing samples from methylation cluster 1 (hypermethylated group, Supplement 4-5) had higher PC1 scores as well as higher C-to-T mutation counts and mutation percentage in CpGs (Fig 3C). This trend was further confirmed using a larger WXS dataset as well (S6). Especially the most hypermethylated group, CpG island methylation phenotype (CIMP), showed the greatest C-to-T in CpGs. Therefore, methylation status was the most prominent factor that shapes the mutation spectra across patients.

Recently, several somatic mutation signatures were identified. Many have putative etiology, revealing the underlying mutation processes (REF). 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. The leading signature was signature 5,

Shantao 8/14/2016 9:38 PM
Deleted: was

Shantao 8/14/2016 9:39 PM
Deleted: didn't

Shantao 8/15/2016 12:58 AM
Formatted: Font:Italic

Shantao 8/15/2016 12:57 AM
Deleted: .

Shantao 8/15/2016 1:04 AM
Deleted: >

Shantao 8/15/2016 1:04 AM
Deleted: >

Shantao 8/15/2016 1:04 AM
Deleted: >

Shantao 8/15/2016 1:04 AM
Deleted: >

Shantao 8/15/2016 1:04 AM
Deleted: >

Shantao 8/15/2016 1:07 AM
Deleted: rates

Shantao 8/15/2016 1:04 AM
Deleted: >

Shantao 8/15/2016 1:09 AM
Deleted: most prominent

Shantao 8/15/2016 1:09 AM
Deleted: five

which is consistent with previous studies (Supplement X). Interestingly, we found one Type II pRCC case out of 155 somatic WXS sequenced samples exhibited APOBEC-associated signature 2 and 13. APOBEC mutation pattern enrichment analysis (see Method) further confirmed the presence of APOBEC activity in pRCC (Fig 3D). This sample was statistically enriched of APOBEC mutations (adjusted p-value < 0.0003).

This case was centrally reviewed by six pathologists in the original study and confirmed to be Type 2 pRCC. Thus the tumor is likely a special case of Type 2 with genomic alterations share some similarities with urothelial cancer (UC), which often carries APOBEC mutation signatures. Indeed, it had 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* or non-synonymous alternations in significantly mutated genes of pRCC, were absent in this sample.

Prominent APOBEC activities were also incidentally detected in three upper track UC samples sequenced and processed in the same pipeline with pRCC samples. This result is consistent with TCGA bladder urothelial cancer study (*15*). Noticeably, along with the Type II pRCC case, all four samples showed significantly higher *APOBEC3A* and *APOBEC3B* mRNA expression level ($p < 0.0022$ and $p < 0.0039$ respectively, one-side rank sum test, S7). This is in line with previous studies of APOBEC mutagenesis in various types of cancer (16).

Consistent with previous studies (16), we could not 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. Accordingly, 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. We postulate defects in chromatin remodeling cause dysregulation of chromatin environment. Open chromatin regions show lower mutation rate, presumably due to more effective DNA repair (REF, 17). Thus chromatin remodeler alternations further alter the mutation landscape, specifically increase mutation rate in previously open chromatin. To test this hypothesis, we tallied the number of mutations inside DNase I hypersensitive sites (DHS) in HEK293, a cell line derived from human embryonic kidney cells, the closest match we could find in ENCODE DHS database. 12/32 samples with non-silent mutations in eleven chromatin remodeling, cancer associated genes show higher genome-wide mutation counts ($p < 0.032$, one-side rank-sum test), partially driven by higher mutation counts in DHS region ($p < 0.003$, one-side rank-sum test). The median number of mutations in DHS region considerably increases by about 50% (75.5 versus 112). The effect is still significant after normalizing against the total mutation counts ($p < 0.015$, one-side rank-sum test, Fig 3E).

Replication time is known to correlate greatly with mutation rate (REF). Early replicated regions have lower mutation rate compared to late replicated ones. Researchers reasoned replication errors are more likely to be corrected by DNA repair system in early replicated regions. With defects in mutated chromatin remodeling, we observed this trend became less accentuated (S8). This is likely because dysregulation of the chromatin environment hinders

PRONOUNCED

BECOME

Shantao 8/15/2016 1:10 AM
Deleted: It

Shantao 8/15/2016 1:11 AM
Deleted: Type II pRCC case with APOBEC activities

Shantao 8/15/2016 1:11 AM
Deleted: .

Shantao 8/15/2016 1:12 AM
Deleted: is

Shantao 8/15/2016 1:14 AM
Deleted: one

Shantao 8/15/2016 1:14 AM
Deleted:

Shantao 8/15/2016 1:57 AM
Deleted: This

Shantao 8/14/2016 9:39 PM
Deleted: regions

Shantao 8/15/2016 2:04 AM
Deleted: disrupts

replication error repair by changing the accessibility of newly synthesized DNA chains 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.

Discussion

We comprehensively analyzed both WGS and an extensive set of WXS of pRCC, finely 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 (Fig 4). Beyond traditionally driver events, we suggested several novel noncoding alterations that could potentially drive tumorgenesis

First, we elaborated on previous results of the long known driver *MET*. In an extended 45 WXS dataset, we found two additional nonsynonymous somatic mutations in the hypermutated tyrosine kinase catalytic domain. This further supports the central role of *MET* in pRCC. Then we found an exonic SNP in *MET*, rs11762213, to be a prognostic germline variance in pRCC. Previously, rs11762213 was found to predict outcome in a mixed RCC samples, predominated by ccRCC (9). Later, this is confirmed in a large ccRCC cohort (10). 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 alternative allele of rs11762213 also forecasts unfavorable outcome in pRCC patients. The mechanism of this exonic germline SNP remains unsettled. Remarkably, pRCC has two subtypes. We noticed cancer-specific deaths in our cohort concentrate in type 2 patients. Thus we hypothesized rs11762213 potentially has different prognostic power in two subtypes. A larger pRCC dataset is required to test our hypothesis. Nevertheless, this finding is potentially very meaningful in clinical management of pRCC patients. rs11762213 genotyping could become a reliable, low-cost risk stratification tool for patients.

Interestingly, MAF of rs11762213 among African American patients is 3.0%, higher than MAFs observed in general African populations in both 1000 Genome phase 3 dataset (0.2%, 0% in Americans with African ancestry (ASW)) and the ExAC dataset (1.27%). This implies a possible effect of rs11762213 on pRCC incidence among African Americans that is worth further investigation. 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 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.

Expand our scope from coding to non-coding, 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 (REF). 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 involved in cancer. Patients carrying mutations in *NEAT1* have higher

Shantao 8/15/2016 2:12 AM
Deleted: validated

Shantao 8/15/2016 2:13 AM
Deleted: for the first time

Shantao 8/15/2016 2:13 AM
Formatted: Font:Italic

Shantao 8/15/2016 2:12 AM
Deleted: as

Shantao 8/15/2016 2:13 AM
Deleted: The original discovery

Shantao 8/15/2016 2:13 AM
Deleted: made

Shantao 8/15/2016 2:14 AM
Deleted: , and

Shantao 8/15/2016 2:14 AM
Deleted: l

Shantao 8/15/2016 2:14 AM
Deleted:

Shantao 8/15/2016 2:15 AM
Deleted: was un

NEAT1 expression and worse prognosis. *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 (*18, 19*).

Last, focusing on 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 poorer prognosis. Although increased C-to-T in CpG is likely the results 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 trans-elements recruitment.

Significant APOBEC activities and consequential mutation signatures were observed in one Type II pRCC case. APOBEC activities were known to be prevalent in UCs (15, 16). 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 agreement with previous studies (*16*). Interestingly, APOBEC mutation signature was also found in a small percentage of chromophobe renal cell carcinoma (*20*), which is believed to have a different cellular origin. APOBEC activities have been linked with genetic predisposition and viral infection (*21*). Although we could not rule out sample processing contamination, given a statistically robust signal in our conservative algorithm, it is plausible that a small fraction of otherwise driver mutation absent Type II pRCCs might be etiologically and genomically similar to UC. Since standard treatment for UC involves cytotoxic chemotherapy and radiation, this finding could have a meaningful clinical impact.

Chromatin remodeling pathway is highly mutated in pRCC (*3*). Several chromatin remodelers, for example *SETD2*, *BAP1* and *PBRM1*, have been identified as cancer drivers in pRCC. We investigate the relationship between samples with mutated chromatin remodelers and those without in terms of overall mutational spectrum. We demonstrated pRCC with defects in chromatin remodeling genes show higher mutation rate in general, driving by an even higher mutation rate rise in putative open chromatin regions. This is because chromatin remodeling defects affect the open chromatin environment and impede DNA repairing in these regions.

It has been known replication time strongly dictates local mutation rate. Early replication regions tend to have less mutation. But the difference dissipates when DNA mismatch repair becomes defective (*17*). In our study, we found this correlation weakened in chromatin remodeling genes mutated samples, presumably caused by failure of replication error repair due to an abnormal chromatin environment. By adapting a defective chromatin remodeling pathway, tumor alters its mutation rate and landscape, which could further provide advantage in cancer evolution. Yet, high mutation burden in functional important open chromatin regions raises the chance that tumor antigens activate the immune system. Thus chromatin remodeler alterations might 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 small sample size, some of 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. 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.

Shantao 8/15/2016 2:48 AM
Deleted: will

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. pRCC samples that failed the histopathological review were excluded (*3*). 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 on MET

We downloaded pRCC clinical outcomes from TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp). In total, we included 207 patients in our analyses. The majority of samples, 162 out of 207, were supported by high-quality, curated SNV callings from multiple centers. 100% genotype concordance rate was observed in samples harbor the minor allele (A) in germline as well as samples with homozygous reference allele (G/G). Also, these curated rs11762213 genotypes were in agreement with automated callsets. With proved high confidence in accuracy of genotyping rs11762213 in germline, we recruited additional 45 samples from single-center (BCM), automated calls to form an extensive patients set (Supplement). We also use this set to find additional somatic mutations in *MET*.

Cancer-specific survival was defined using similar method as described in a ccRCC study (10). 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 precedure

We use DELLY2 (REF) 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. Last, we discharge all callings that were marked "LowQual" and retain the ones marked as "PASS".

Mutation spectra study

WGS Mutations were extracted from with flaking 5' and 3' nucleotide context. Then the raw mutation counts were normalized based on trinucleotide frequency 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.

Shantao 8/15/2016 2:55 AM
Comment [4]: Write down the L1 regression math formula here?

The penalty parameter lambda was determined empirically using 10-fold cross-validation individually for every sample. 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 (*22*). 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.

Shantao 8/15/2016 2:57 AM
Deleted: must

APOBEC enrichment analysis

We used the method described by Roberts et al. (*16*). 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.

Chromatin remodeling genes and replication time association

We identified chromatin remodeling genes based on its significance in pRCC (appears in the TCGA study summary table (3)) and function. Our gene list included eleven genes. They are *ARID1A*, *ARID2*, *BAP1*, *DNMT3A*, *KDM6A*, *MLL2*, *MLL3*, *MLL4*, *PBRM1*, *SETD2*, *SMARCB1*.

Shantao 8/15/2016 2:58 AM
Deleted: its

In order to avoid cell type redundancy, we only kept Gm12878 as the representative of all lymphoblastoid cell lines. Wave smoothed replication time signal is 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. 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 adapted a conservative non-parametric Kolmogorov–Smirnov test (K-S test) using empirical p-value. We assigned all the mutation with its percentile among all mutations replication time

shifted +/- 100kb from the origin (represents the background replication time). Then we calculate the K-S test statistics of mutation counts of 100 bins in two groups and compare. 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.

Supplementary Materials

Supplement figure 1. Mutations on *ERRFI1* promoter region has no effect on *ERRFI1* RNA expression.

Supplement figure 2. Mutations on *ERRFI1* promoter region has no effect on *ERR* family protein levels.

Supplement figure 3. Mutations on *ERRFI1* promoter region has no effect on *ERR* family Phosphorylation.

Supplement figure 4. Volcano plot of rank sum test of all CpG probe sites between methylation cluster 1 and 2.

Supplement figure 5. Volcano plot of rank sum test of all CpG probe sites between methylation cluster 1 and 2 after grouped by functional regions.

Supplement figure 6. Comparison of C-to-T in CpGs mutation counts and fractions in pRCC WXS set among three different methylation clusters.

Supplement figure 7. The expression levels of *APOBEC3A* and *APOBEC3B* are significantly higher in samples carrying APOBEC signatures

Supplement figure 8. Replication time distribution shifts among samples having chromatin remodeling genes mutations.

References and Notes:

1. Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. "Cancer statistics, 2015." CA: a cancer journal for clinicians 65.1 (2015): 5-29.

2. Shuch, Brian, Ali Amin, Andrew J. Armstrong, John N. Eble, Vincenzo Ficarra, Antonio Lopez-Beltran, Guido Martignoni, Brian I. Rini, and Alexander Kutikov. "Understanding pathologic variants of renal cell carcinoma: distilling therapeutic opportunities from biologic complexity." European urology 67, no. 1 (2015): 85-97.

3. Cancer Genome Atlas Research Network. "Comprehensive molecular characterization of papillary renal-cell carcinoma." N Engl J Med 2016, no. 374 (2016): 135-145.

4. Khurana, Ekta, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner et al. "Integrative annotation of variants from 1092 humans: application to cancer genomics." Science 342, no. 6154 (2013): 1235587.

5. Fu, Yao, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Jasmine Mu, Kevin Y. Yip, Ekta Khurana, and Mark Gerstein. "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer." Genome biology 15, no. 10 (2014): 1.

6. Huang, Franklin W., Eran Hodis, Mary Jue Xu, Gregory V. Kryukov, Lynda Chin, and Levi A. Garraway. "Highly recurrent TERT promoter mutations in human melanoma." Science 339, no. 6122 (2013): 957-959.

7. Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Peter J. Campbell, and Michael R. Stratton. "Deciphering signatures of mutational processes operative in human cancer." Cell reports 3, no. 1 (2013): 246-259.

8. Alexandrov, Ludmil B., Serena Nik-Zainal, Hoi Cheong Siu, Suet Yi Leung, and Michael R. Stratton. "A mutational signature in gastric cancer suggests therapeutic strategies." Nature communications 6 (2015).

9. Schutz, Fabio AB, Mark M. Pomerantz, Kathryn P. Gray, Michael B. Atkins, Jonathan E. Rosenberg, Michelle S. Hirsch, David F. McDermott et al. "Single nucleotide polymorphisms and risk of recurrence of renal-cell carcinoma: a cohort study." The lancet oncology 14, no. 1 (2013): 81-87.

10. Hakimi, A. Ari, Irina Ostrovnaya, Anders Jacobsen, Katalin Susztak, Jonathan A. Coleman, Paul Russo, Andrew G. Winer et al. "Validation and genomic interrogation of the MET variant rs11762213 as a predictor of adverse outcomes in clear cell renal cell carcinoma." Cancer 122, no. 3 (2016): 402-410.

11. Nik-Zainal, Serena, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena et al. "Landscape of somatic mutations in 560 breast cancer whole-genome sequences." Nature534, no. 7605 (2016): 47-54.

12. Guo, Sien, Wenjie Chen, Yihuan Luo, Fanghui Ren, Tengfei Zhong, Minhua Rong, Yiwu Dang, Zhenbo Feng, and Gang Chen. "Clinical implication of long non-coding RNA NEAT1 expression in hepatocellular carcinoma patients."International journal of clinical and experimental pathology 8, no. 5 (2015): 5395.

13. Choudhry, H., A. Albukhari, M. Morotti, S. Haider, D. Moralli, J. Smythies, J. Schödel et al. "Tumor hypoxia induces nuclear paraspeckle formation through HIF-2α dependent

transcriptional activation of NEAT1 leading to cancer cell survival." Oncogene 34, no. 34 (2015): 4482-4490.

14. Chakravarty, Dimple, Andrea Sboner, Sujit S. Nair, Eugenia Giannopoulou, Ruohan Li, Sven Hennig, Juan Miguel Mosquera et al. "The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer." Nature communications 5 (2014).

15. Cancer Genome Atlas Research Network. "Comprehensive molecular characterization of urothelial bladder carcinoma." Nature 507, no. 7492 (2014): 315-322.

16. SA. Roberts, Steven A., Michael S. Lawrence, Leszek J. Klimczak, Sara A. Grimm, David Fargo, Petar Stojanov, Adam Kiezun et al. "An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers."Nature genetics 45, no. 9 (2013): 970-976.

17. Supek, Fran, and Ben Lehner. "Differential DNA mismatch repair underlies mutation rate variation across the human genome." Nature 521, no. 7550 (2015): 81-84.

18. Li, Yunlong, Yaohui Li, Wenping Chen, Fenfei He, Zhaobang Tan, Jianyong Zheng, Weizhong Wang, Qingchuan Zhao, and Jipeng Li. "NEAT expression is associated with tumor recurrence and unfavorable prognosis in colorectal cancer." Oncotarget 6, no. 29 (2015): 27641.

19. He, Chengbiao, Bing Jiang, Jianrong Ma, and Qiaoyu Li. "Aberrant NEAT1 expression is associated with clinical outcome in high grade glioma patients."Apmis 124, no. 3 (2016): 169-174.

20. Davis, Caleb F., Christopher J. Ricketts, Min Wang, Lixing Yang, Andrew D. Cherniack, Hui Shen, Christian Buhay et al. "The somatic genomic landscape of chromophobe renal cell carcinoma." Cancer cell 26, no. 3 (2014): 319-330.

21. Henderson, Stephen, Ankur Chakravarthy, Xiaoping Su, Chris Boshoff, and Tim Robert Fenton. "APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development." Cell reports 7, no. 6 (2014): 1833-1841.

22. Wang, Dan, Li Yan, Qiang Hu, Lara E. Sucheston, Michael J. Higgins, Christine B. Ambrosone, Candace S. Johnson, Dominic J. Smiraglia, and Song Liu. "IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data." Bioinformatics 28, no. 5 (2012): 729-730.

Acknowledgments: Funding: This work was supported by the National Institutes of Health, AL Williams Professorship, and in part by the facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center (Grant Number RR029676-01). **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.