Bullet points, major findings:

- 32 WGS + extensive set of WXS; in depth analysis
- Mutational heterogeneity
    - Methylation
    - APOBEC (unique in pRCC)
    - Chromatin remodeling genes
- Examples of high-impact non-coding mutations
- rs117652213 predicts cancer-specific survival, first time validated in pRCC.

Title

Abstract

Introduction

Renal cell carcinoma (RCC) makes up over 90% of kidney cancers and currently is the most lethal genitourinary malignancy \cite{25559415}. Papillary RCC (pRCC) accounts for 10%-15% of the total RCC cases (REF). Unfortunately pRCC has been understudied and there are no current forms of effective systemic therapy for this disease. Recently, the Cancer Genome Atlas (TCGA) published its first result on pRCC(REF), which improves our understanding of the disease in a genomic aspect.

Shantao 5/22/2016 11:11 PM
Comment [1]: [[First three sentences from candisp grant]]

Multiple endogenous and environmental mutation processes shape the somatic mutation spectra observed in cancers (REF Alexanderov). Mutation processes decomposition gives information of cancer development, sheds light on mutational disparity between cancer subtypes and even indicates potential new treatment strategies (REF Alexanderov Gasteric CA). 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), by probing the entire genome, gives richer information on mutation landscape and minimizes the potential effect of clone selection.

Non-coding region, previously often overlooked, has been showed to play an active role in cancer development [REF:Funseq, TERT promoter]. Mutations in non-coding region are able to cause disruptive change in both cis- and trans regulatory elements. Understanding non-coding mutations helps fill the missing “dark matters” in cancer research.

DISC W/ BRIAN

In this study, we comprehensively analyzed 32 pRCC WGS data along with an extensive set of WXS data. We discovered pRCC exhibits mutational heterogeneity in both nucleotide context and genome location, indicating underlying vibrant mutational processes interplay. Methylation and APOBEC activity are two leading factors influencing the mutation landscape. Methylation status drives the intra-sample mutation variation by giving rise to more C>T mutations in the CpG context. APOBEC activity, although sparely occurred, leaves unequivocal mutation signatures in some pRCC genomes but not in ccRCC. Empowered by whole genome sequencing, we scrutinized about 150,000 non-coding mutations and found several potentially high-impact mutations in non-coding regions. Last, we validated rs11762213, a germline exonic single nucleotide polymorphism inside proto-oncogene MET, as a cancer-specific survival (CSS) predictive SNP for the first time in pRCC.

+Jim UNIFY

1

Results

1. **Mutation spectra of pRCC**

   We summarized the mutation spectra of 32 whole genome sequenced pRCC samples (Fig 1A). C>T in CpGs shows the highest mutation rates, which are roughly ten to twenty folds 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 PC1 (explains

12.5% of the variation) demonstrate C>T in CpGs contributes the most to inter-sample variation (Fig 1B). C>T in CpGs reflects the spontaneous deamination of cytosines in CpGs, especially 5-methylcytosine. We confirmed this by showing samples from methylation cluster 1 (hypermethylated group) have higher PC1 scores as well as higher C>T mutation counts and rates in CpGs (Fig 1C). Therefore, methylation status is the most prominent factor that shapes the mutation spectra across patients.

[[Working on some methylation analyses here]]

Using an in-house LASSO-based tool (see Methods) to identify mutation signatures in both WGS and WXS samples, we found 4/161 (2.5%) samples in the WXS data exhibited APOBEC-associated signature 2 and 13. APOBEC mutation pattern enrichment analysis (see Method) further confirms the presence of APOBEC activity in pRCC (Fig 4D). The corresponding four samples are statistically enriched of APOBEC mutations (all p-value < 0.0003). Noticeably, these four samples show significantly higher APOBEC3A and APOBEC3B mRNA expression level ($p < 0.0022$ and $p < 0.0039$ respectively, one-side rank sum test). Both APOBEC3A and APOBEC3B expression levels also correlates well with the APOBEC mutation fraction among the four samples (Spearman correlation, 0.8 in both).

Shantao 5/17/2016 12:54 AM

Comment [2]: These four samples don't have WGS. However, enrichment scores between WXS and WGS in 31 samples show very poor correlation (cor = ~0 at best). Also there seems to be an systematic, baseline elevated mutation rate in TCW motifs (median consistently > 1). Therefore, Fisher-exact test doesn't work in WGS: almost every sample is tested significant. No particularly high enrichment seen in WGS. The range is 1.03-1.59.

Consistent with previous studies (REF), we could not detect APOBEC activities in an extensive WXS dataset consisting of 418 clear-cell RCC (ccRCC) samples. Only very low level of APOBEC signatures (<15%) was found in four samples. Because of a much larger sample size, this is unlikely to be confounded by detecting power.

DEFEATING

2. Defects in chromatin remodeler affects mutation landscape

Chromatin remodeling genes are frequently mutated in pRCC and many other cancers. We postulate defects in chromatin remodeling cause dysregulation of chromatin status. This further alters the mutation landscape, specifically increases mutation rate in open chromatin. To test this hypothesis, we tallied the number of mutations inside DNase I hypersensitive sites (DHS) in HEK293 (human embryonic kidney). 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 an even 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).

Shantao 5/22/2016 11:30 PM
Comment [3]: Is our set of chromatin remodelers arbitrary?
Using the gene set of chromatin remodeler + SWI/SNF modifier from NEJM pathway analysis didn't work. Further applying a second PolyPhen filter didn't help.

Shantao 5/22/2016 11:30 PM
Comment [4]: Test statistics changed. I pulled the mutations from MAF directly this time rather than relying on the NEJM supplement tables. Somewhat there is minor disagreement.
One sample is missing in MAF. I looked into VCFs from two centers.

Replication time is known to correlate greatly with mutation rate. Early replicated regions have lower mutation rate but the difference dissolves when DNA mismatch repair becomes defective (REF). We discovered the distribution of replication time at each non-coding mutation correlated with percentage of mutations inside DHS (Spearman's correction: 0.69). We found a trend of shifting to earlier replication in the mutated group. The AUC of replication time distribution is significantly different between two groups ($p<0.05$, one-side rank-sum test). However, this shift is not statistically significant (empirical p-value < 0.17).

3. Mutations in non-coding region

Mutations in non-coding region have been demonstrated to play a critical role in cancer. We ran FunSeq2 to identify potentially high-impact non-coding variants in pRCC. First, we identified a mutation hotspot on chromosome 1. 6/32 (18.8%) samples have mutations within this 6.5kb region (Fig 3A). This hotspot locates at the upstream of ERRFI1 (ERBB

Receptor Feedback Inhibitor 1) and overlaps with the predicted promoter region. ERRFI1 is the negative regulator of EGFR family members including EGFR, HER2 and HER3. However, we didn't observe statistically significant changes among mutated samples in terms of mRNA expression level, protein level and phosphorylation level of EGFR, HER2 and HER3 (Supplements X). Noticeably, due to a very limited sample size here, our test power was greatly compromised.

We also observed one mutation in MET promoter region in a type 1 pRCC sample (Fig 3B). This sample has no nonsynonymous mutation in MET gene but copy number gain of MET. Additionally, we have observed 6/32 (18.8%) samples carry mutations in the first or the second introns of MET (Fig 3C).

Shantao 5/15/2016 3:07 PM
Comment [5]: The MET mRNA level in this sample is LOW

Shantao 5/15/2016 8:30 PM
Comment [6]: No association with the MET splicing event. Only 2/32 have this event and none of them has mutation here.

Worth meaning at all?

Shantao 5/17/2016 1:00 AM
Comment [7]: Should we mention NEAT1?

Another potentially impactful mutation hotspot is NEAT1. We saw mutations inside this nuclear long non-coding RNA in 5/32(15.6%) samples. Several studies indicated NEAT1 is associated in lung and prostate cancer [REF]. It promotes cell proliferation in hypoxia [REF].

4. Probing rs11762213 in pRCC prognosis

A germline SNP, rs11762213, has been discovered to predict recurrence and survival in a RCC cohort, predominated by ccRCCs [REF]. This conclusion was later validated in ccRCC but never in pRCC [REF]. We would like to know 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). No homozygous A/A was observed. The cancer-specific survival is significantly worse in patients with the risk allele ($p < 0.037$, Peto & Peto modification of the Gehan-Wilcoxon test, FIG 4).

The minor allele (A) frequency in our dataset is 2.90%, slightly lower than the previous studies. However, among patients with African ancestry, the MAF is 3.95%. It is higher than MAFs previously observed in general African populations in both 1000 Genome phase 3 dataset (0.2%) and the ExAC dataset (1.27%). This implies a possible effect of rs11762213 on pRCC incidence among African Americans that worth further investigation.

SUMMARIZE

Shantao 5/22/2016 11:53 PM
Comment [8]: Sample size too small. Move to discussion?

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. We identified mutation rate dispersion of C>T in the CpG motif contributes to the largest proportion of inter-sample variations. We further pinned down the cause of dispersion by showing the hypermethylated cluster, identified in the previous TCGA study (REF), has higher C>T rate in CpGs. This hypermethylated cluster is associated with later stage, type 2 pRCC, SETD2 mutation and poorer prognosis. Although increased C>T in CpG is the results of hypermethylation, we cannot rule out the possibility the change of mutation landscape plays a role in cancer development.

DIFF DISEASES

Despite coming with a low prevalence, significant APOBEC activities and consequential mutation signatures are observed in four pRCC cases. Interestingly, although being considered to have the same cellular origin with pRCC, we were not able to detect APOBEC activities in ccRCC. This is in agreement with previous studies (REF). APOBEC activities have been linked with genetic predisposition and viral infection (REF). Thus the divergence of ccRCC and pRCC might be dictated by APOBEC in some patients. With unusual strong APOBEC-related signatures that could make up to more than 70% in total detectable signatures in some pRCC cases, APOBEC activities could greatly shape pRCC spectra.

Chromatin remodeling pathway is highly mutated in pRCC (REF). Several chromatin remodelers, for example SETD2, BAP1 and PBRM1, have been identified as cancer drivers in pRCC. We demonstrated pRCC with defects in chromatin remodeling genes show higher mutation rate in general, driving by an even higher mutation rate in open chromatin regions. By adapting a defective chromatin remodeling pathway, tumor alters its mutation rate and landscape, which could further provide advantage in cancer evolution. However, excessive mutation in functional important open chromatin regions would also lead to disastrous mutational meltdown.

We found several potentially significant non-coding mutations. In our pRCC cohort, a mutation hotspot was found upstream of ERRFI1. Served as a potential tumor suppressor, these mutations potentially disrupt regulatory elements of ERRFI1 and thus play a role in tumorigenesis. However, likely limited by small sample size, we were not able to detect statistically significant functional changes in ERRFI1 and related pathways. We also discovered mutations associated with MET promoter and first two introns. Another hotpot is in NEAT1, a long non-coding RNA that has been speculated to involved in cancer.

Last, we validated rs11762213 as a prognostic germline variance in pRCC for the first time. The original discovery was made in a mixed RCC samples, predominated by ccRCC. Recently, the discovery was confirmed in a ccRCC cohort. It is unclear whether rs11762213 only predicts the outcome in ccRCC. 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 death events in our cohort concentrate in type 2 patients, due to type 2 pRCC inferior prognosis. Thus we further hypothesized rs11762213 potentially has different prognostic power in subtypes, likely to be more powerful in type 2 pRCC. Unlike type 2 pRCC and ccRCC, Type 1 pRCC often carry somatic MET mutations. A larger pRCC dataset is required to test our hypothesis.

Methods

Data acquisition

We downloaded pRCC and ccRCC WXS SNV calls and pRCC WGS variation calls from TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp). pRCC samples that failed the histopathological review were excluded. Patients included in this study were summarized in supplemental table X. pRCC RNAseq, RPPA and methylation data were downloaded from TCGA Data Portal as well.

Shantao 5/22/2016 11:39 PM
Comment [9]: Originally from NCI Jamboree.

Repli-seq and DHS data were obtained from ENCODE (https://www.encodeproject.org/).

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. 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.

SEE. PAPER

APOBEC enrichment analysis

We used the method described by XXX [REF]. 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.

Replication time association

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 plot the cumulative mass function of the mutation replication time in each sample. Area under curve (AUC) is used as a measurement of the distribution. Specifically, a smaller AUC indicated a shift of mutations to the early replicate regions and vice versa.

we adapted a non-parametric test using empirical p-value. We calculated the rank sum of replication time of mutations in every sample and then normalized by its mutation count. Then we sum up the ranks in both group and compare. To obtain the empirical p-value, we randomly sample

10,000 times the tumor samples with equal sizes of these two groups to estimate the rank sum distribution.

Testing rs11762213 on prognosis

We downloaded pRCC clinical outcomes from TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp). Excluding criteria are “Follow-up days” not available and identified as non-pRCC by histopathological review. In total, we included 207 patients in our analyses. The majority of samples, 158 out of 207, were supported by high-quality, curated SNV callings from two centers. 100% genotype concordance rate was observed in samples harbor the minor allele (A, 10 samples) in germline as well as samples with homozygous reference allele (GG, 148 samples). 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 59 samples from single-center, automated calls.

Cancer-specific survival was defined using similar method as described in a ccRCC study (REF). 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. An R package, “survival”, was used for the survival analysis.