**Introduction**

###Cut out from Result 1###

Multiple endogenous and environmental mutation processes shape the somatic mutation spectrum observed in cancers. Whole genome sequencing, by probing the entire genome, gives richer information on mutation landscape and minimizes the potential effect of clone selection.

**Results**

1. Mutation spectrum of PRCC

We summarized the mutation spectrum 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 the factors that explain the most inter-sample variation. The first principle component (PC1) explains 12.5% of the variation. The loadings on PC1 demonstrate C>T in CpGs contributes the most to inter-sample variation (Fig 1B). C>T in CpGs reflects the spontaneous deamination of 5-methylcytosine in CpGs. We confirmed this by showing samples from methylation cluster 1 (hypermethylated group) have higher scores on PC1 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 spectrums across patients.

Using an in-house tool (more details to come) to identify mutation signatures in both WGS and WXS samples, we found 5/154 (3.25%) [[so low? more on majority]] samples that have at least one signature identified using WXS data showed either signature 2 or 13. Three samples show both signature 2 and 13. These two signatures are linked with ApoBEC activities. ApoBEC mutation pattern enrichment analysis (See Method) further confirms the presence of ApoBEC activity in PRCC (Fig 4D). The ApoBEC pattern enrichment fold correlates well with the identified ApoBEC signatures percentage (Spearman’s correlation = 0.9). The enrichment fold plot indicates potentially more samples influenced by ApoBEC activities. Due to low signature detecting power in samples with low mutation count, we were not able to confidentially warrant ApoBEC signatures presence in all of them.

Interestingly, although being considered to have similar cellular origin with PRCC, we could not detect ApoBEC activities in ccRCC. This is in agreement with previous studies (REF). DISCUSSION?

1. 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 2A). This hotspot locates at the upstream of ERRFI1 (ERBB Receptor Feedback Inhibitor 1) and it 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 caused by mutations in ERRFI1 on mRNA expression level, protein level and phosphorylation level of EGFR, HER2 and HER3 (Supplements X).[[b/c too few samples]]

We also observed one mutation in MET promoter region in a type 1 PRCC sample (Fig 2B). This sample has no nonsynonymous mutation in MET gene but copy number gain of MET. (The MET mRNA level is LOW)

1. Probing rs11762213 in PRCC prognosis

A germline SNP, rs11762213, has been discovered to be able to predict recurrence and survival in RCC, mostly constituted by ccRCCs.[REF]. It was later validated in ccRCC but not in PRCC [REF]. We would like to know whether this SNP has a prognostic effect in PRCC. Using whole exome sequencing data from 202 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 having the risk allele (p < 0.037, Peto & Peto modification of the Gehan-Wilcoxon test).

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%. This 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.

1. Defective chromatin remodeler affects mutation landscape[[move earlier & integrate]]

Chromatin remodeling genes are frequently mutated in PRCC and many other cancers. We postulate defects in chromatin remodeling cause dysregulation of chromatin status, which further changes the mutation landscape. To test this hypothesis, we calculated the number of mutations in DNase I hypersensitive site (DHS) as defined by HEK293 (human embryonic kidney). Samples with non-silent mutations in eleven chromatin remodeling genes show higher mutation counts in DHS region (p < 0.005) but not in the total mutation counts (p > 0.05).

**Methods**

**Testing rs11762213 on prognosis**

We downloaded PRCC whole exome sequencing, germline SNV calling from TCGA Data Portal (<https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp>). Excluding criteria: “Follow-up days” not available or 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 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” was “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.