Bullet points, major findings:

- 32 WGS + extensive set of WXS; in depth analysis
  - Finely scrutinizing local high-impact events as well as giving a macro overlook of the mutation landscape
- rs117652213 predicts cancer-specific survival, first time validated in pRCC.
- Examples of high-impact non-coding mutations
- Mutational heterogeneity
  - Methylation
  - APOBEC (unique in pRCC)
  - Chromatin remodeling genes

Title

Abstract

Papillary renal cell carcinoma (pRCC) constitutes 10-15% cases of renal cell carcinoma, which is the most common kidney cancer. Recent advancing of DNA sequencing unprecedentedly deepened our understanding of pRCC. However, previous research scope was limited to coding alterations in traditional driver genes. A significant proportion of samples still lack clear molecular etiology. In this first pRCC whole genome sequencing analysis study, we took a comprehensive approach to examine pRCC genomes, trying to explain the cases lacking classic driver mutations. Starting from established drivers, we expanded to the entire genome and finally to high-order mutational patterns. We validated a germline MET exonic SNP, rs11762213, predicts prognosis in pRCC. Then in non-coding region, we found several potentially impactful mutation hotspots. Recurrent mutations inside a long non-coding RNA, NEAT1, associated with cancer specific survival. Last, by mining the mutational heterogeneity of both nucleotide context and genomic locations, we revealed key factors that dictates/intervene mutation processes.

Shantao 6/28/2016 12:56 AM
Comment [1]: Or should we say something like “complete the mutation spectrum”

Shantao 6/27/2016 11:47 AM
Formatted: Highlight

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. Traditionally, the only prominent oncogene in pRCC (specifically type 1) that physicians were able to identify for long is MET, a tyrosine kinase receptor for hepatic growth factor. Amino acid substitution that locks MET in “on-state” and overexpression are two mechanisms of dysfunction of MET in tumorgenesis. Accordingly, missense mutations, amplification and other coding region alterations have been reported among many Type 1 pRCC patients. Recently, the Cancer Genome Atlas (TCGA) published its first result on pRCC(REF), which improves our understanding of the genomic basis of this disease. Several more genes were identified to be significantly mutated in pRCC. Nevertheless, a significant portion of pRCC cases still remains “driver-unknown”. [[but good time for us - why?]]

pRCC I

Shantao 6/25/2016 5:53 PM
Deleted:

Shantao 6/25/2016 5:53 PM
Deleted: [[is this a tcga discovery?]]

Shantao 6/25/2016 5:54 PM
Deleted: MET, especially, is the leading driver for Type 1 pRCC. Missense mutations, amplification and other coding region alterations have been reported in these patients. [More intro about MET?][[YES what does it do?]]

Mark Gerstein 6/19/2016 11:44 AM
Formatted: Highlight

Non-coding regions, previously overlooked in cancer, has been showed to be involved in tumorigenesis [REF:Funseq, TERT promoter]. Mutations in non-coding regions may cause disruptive changes in both cis- and trans-regulatory elements. 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 (REF Alexanderov). Analyses of the associated genomic alterations give information of cancer development, shed 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) gives richer information on mutation landscape and minimizes the potential effect of clone selection.

In this study, we comprehensively analyzed 32 pRCC WGS data along with an extensive set of WXS data in multiple levels. We went from microscopic examination of driver genes to analysis of whole genome sequencing variants and finally, to investigation of high-order mutational features. First, We focused on MET, a proto-oncogene which play a central role in pRCC, especially in Type 1. For the first time we validated rs11762213, a germline exonic single nucleotide polymorphism inside MET, as a predictive SNP for cancer-specific survival (CSS) in pRCC. We also found several potentially impactful non-coding alternations around MET promoter and first two exons.[[which expl some cases w/o coding?]] Then we went onto cases not as easily explained as those with MET alterations. We analyzed nearly 150,000 non-coding mutations 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 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 infrequently observed, leaves unequivocal mutation signatures in a pRCC genome but not in ccRCC.

Shantao 6/26/2016 4:40 PM
Deleted: W

Mark Gerstein 6/19/2016 11:38 AM
Formatted: Highlight

Shantao 6/26/2016 4:40 PM
Deleted: for the first time in pRCC.

Mark Gerstein 6/19/2016 11:38 AM
Formatted: Highlight

Mark Gerstein 6/19/2016 11:41 AM
Formatted: Highlight

Results

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

WHAT ABT SOM.

We begin with MET coding variants. Although many MET somatic mutations are believed to play a central role in pRCC, a germline SNP, rs11762213, has [[not been linked?]] been discovered to predict recurrence and survival in an RCC cohort. ccRCC predominated the initial discovery RCC cohort[REF]. This conclusion was later validated in ccRCC cohort but never in pRCC [REF]. We evaluated whether this SNP has a prognostic effect in pRCC. Using an extensive WXS set of 207 patients (see Methods), 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

Mark Gerstein 6/19/2016 11:42 AM
Formatted: Highlight

allele (p < 0.037, Peto & Peto modification of the Gehan-Wilcoxon test; p < 0.044, log-rank test, Fig 1). 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%.

2. Mutation hotspots in non-coding region

?

Despite the fact MET is the most important driver; some presumably MET-driven yet MET wildtype pRCC samples were left unexplained. 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 has no evidence ~~of a 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 2A). Notably, RNA splicing variants of exon 1-3 were found in several pRCC samples and thought to be a cancer-driving event. However, we were not able to find statistically significant association between slicing events and intronic mutations.

W/CODING

We further expanded our scope and ran FunSeq2 [REF] 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. Noticeably, 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 terms of mRNA expression level, protein level and phosphorylation level of EGFR, HER2 and HER3 (Supplements X).

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] (FIG 2C). **It can also alter the epigenetic

Brian Shuch 6/7/2016 4:31 PM
Comment [2]: No nonsy.....thats a double negative.

Brian Shuch 6/7/2016 4:31 PM
Comment [3]: Should expand on significance

Shantao 6/26/2016 4:41 PM
Deleted: [[do these have no normal met muts]]

Shantao 6/11/2016 10:24 PM
Comment [4]: 2/32 have splicing event and none in these 6.
But potentially these mutations give similar effects?

Can we just report something without significance discussion.?

Shantao 6/26/2016 4:41 PM
Deleted: did not detect any association of the two.[[need to discuss..]]

Mark Gerstein 6/19/2016 12:01 PM
Formatted: Highlight

Mark Gerstein 6/19/2016 12:03 PM
Formatted: Highlight

Brian Shuch 6/11/2016 10:25 PM
Comment [5]: EGFR Splice variant missing exons 2-7 Ji 2006 PNAS

Shantao 6/26/2016 4:42 PM
Deleted: was greatly compromised. [[say more positively]]

Mark Gerstein 6/19/2016 12:03 PM
Formatted: Highlight

Mark Gerstein 6/19/2016 12:03 PM
Formatted: Highlight

Mark Gerstein 6/19/2016 12:03 PM
Formatted: Highlight

Shantao 6/16/2016 11:08 PM
Comment [6]: Flip the order of NEAT1 and ERRFI1?

?

landscape and promote transcription of target genes (**Dimple Chakravarty Nature comm 2014).**

Samples carry NEAT1 mutations have higher NEAT1 expression (Fig 2D?, $p < 0.044$, two-sided rank sum test). We also found NEAT1 mutations are associated with worse prognosis in patients not carrying rs11762213 minor allele (Fig 2E, $p < 0.022$, log-rank test).

**would see if the NEAT1 or MET intronic alterations are in a specific subgroups of papillary RCC (cluster, type I vs II etc).

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>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 PC1 (which explains 12.5% of the variation) demonstrated C>T in CpGs contributes the most to inter-sample variation (Fig 3B). C>T in CpGs is highly associated with methylation. It reflects the spontaneous deamination of cytosines in CpGs, which is much more frequent in 5-methylcytosines. So we further explored the association between C>T in CpGs and tumor methylation status. We confirmed this by showing samples from methylation cluster 1 (hypermethylated group, Supplement X) had higher PC1 scores as well as higher C>T mutation counts and rates in CpGs (Fig 3C). This trend is further confirmed using WXS as well (Supplement X). Especially the most hypermethylated group, CpG island methylation phenotype (CIMP), showed the greatest C>T in CpGs. Therefore, methylation status was the most prominent factor that shapes the mutation spectra across patients.

Shantao 6/26/2016 4:44 PM
Moved down [2]: Also these mutations [[the neat ones or other ones??]] are ~10-20kb of a downstream microRNA, MIR-612. Higher MIR-612 expression has been proved to negatively regulates colorectal and hepatocellular cancer growth (http://www.ncbi.nlm.nih.gov/gene/693197) . [[move this neg. regulate to a sep. para]]

Mark Gerstein 6/19/2016 12:04 PM
Formatted: Highlight

Mark Gerstein 6/19/2016 12:07 PM
Formatted: Highlight

Shantao 6/26/2016 4:44 PM
Deleted: and higher percentage of non-zero MIR-612 read counts (3/5, 60% versus 3/27, 9%; $p < 0.038$, two-tailed Fisher exact test)

Shantao 6/26/2016 4:44 PM
Deleted: . However,

Shantao 6/26/2016 4:44 PM
Deleted: w

Shantao 6/26/2016 4:44 PM
Deleted: still

Shantao 6/26/2016 4:45 PM
Deleted: ?

Shantao 6/28/2016 12:57 AM
Comment [7]: See figure 5.

Shantao 6/26/2016 4:44 PM
Deleted: ... [1]

Shantao 6/26/2016 4:44 PM
Moved (insertion) [2]

Shantao 6/26/2016 4:46 PM
Deleted: especially

Shantao 6/26/2016 4:47 PM
Deleted: [[need to expl linkage better]]

Mark Gerstein 6/19/2016 12:10 PM
Formatted: Highlight

Shantao 6/7/2016 4:31 PM
Comment [8]: Volcano plots

[[Working on some methylation analyses here: want to demonstrate these mutations indeed happen at hypermethylation sites]]

correlation of SIG5? AGE?

Using a LASSO-based approach (see Methods) to identify mutation signatures in both WGS and WXS samples, 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). It was statistically enriched of APOBEC mutations (adjusted p-value < 0.0003).

[[para]]

This Type II pRCC case with APOBEC activities had non-silent mutations in ARID1A and MLL2 and a synonymous mutation in RXRA, all are identified as significantly mutated genes in UC. 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.

Mark Gerstein 6/19/2016 12:21 PM
Moved (insertion) [1]

Prominent APOBEC activities were also incidentally detected in three upper track urothelial cancer samples sequenced and processed in the same pipeline with pRCC samples. This result is consistent with TCGA bladder urothelial cancer study [REF]. Noticeably, 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).

Mark Gerstein 6/19/2016 12:18 PM
Formatted: Highlight

Consistent with previous studies (REF), 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 level of APOBEC signatures (<15%) was found in only four samples. With a much larger sample size, this result was unlikely to be confounded by detecting power.

Mark Gerstein 6/19/2016 12:21 PM
Moved up [1]: This Type II pRCC case with APOBEC activities had non-silent mutations in ARID1A and MLL2 and a synonymous mutation in RXRA, all are identified as significantly mutated genes in UC. 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.

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 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. HEK293 cell line is derived from human embryonic kidney cells, which is the closest match we found 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 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, Fig 4A).

Shantao 6/26/2016 4:50 PM
Deleted: (human embryonic kidney).

Shantao 6/26/2016 4:50 PM
Deleted:

2

Shantao 6/7/2016 4:31 PM
Comment [9]: 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 dissipates 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).

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 5). Beyond traditionally driver events, we suggested several novel noncoding alterations that could potentially drive tumorgenesis.

First, we validated an exonic SNP in MET, 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. Nevertheless, this finding is potentially very meaningful in clinical management of pRCC patients. rs11762213 genotyping could be a reliable low-cost risk stratification method for patients.

Interestingly, MAF of rs11762213 among African American patients is 3.95%, higher than MAFs 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 is worth further investigation.

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 mutations relevant to tumorigenesis. In our pRCC cohort, 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

Shantao 6/26/2016 4:51 PM
Deleted:

Brian Shuch 6/11/2016 10:30 PM
Comment [10]: What are the breakdown of the Type I and II between groups?

Shantao 6/26/2016 4:54 PM
Moved (insertion) [3]

Shantao 6/26/2016 4:55 PM
Deleted: W

observed. 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. Another hotpot is in NEAT1, a long non-coding RNA that has been speculated to involved in cancer. Patients carrying mutations in NEAT1 have higher NEAT1 expression and worse prognosis.

Last, focusing on the high-level land scape of mutations in pRCC, 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 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>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 (REF). 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 APOBEC activities in ccRCC. This is in agreement with previous studies (REF). APOBEC activities have been linked with genetic predisposition and viral infection (REF). 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 meaningful clinical impact.

Brian Shuch 6/11/2016 10:32 PM
Comment [11]: Phase II Study of Erlotinib in Patients With Locally Advanced or Metastatic Papillary Histology Renal Cell Cancer: SWOG S0317
Michael S. Gordon, Michael Hussey, Raymond B. Nagle, Primo N. Lara Jr, Philip C. Mack, Janice Dutcher, Wolfram Samlowski, Joseph I. Clark, David I. Quinn, Chong-Xian Pan, and David Crawford

Shantao 6/26/2016 4:54 PM
Moved up [3]: We also discovered mutations associated with MET promoter and first two introns.

Brian Shuch 6/11/2016 10:32 PM
Comment [12]: Discuss implications of both

The MET SNP may be important for determining management as it may be an important predictive/prognostic biomarker of disease behavior

Shantao 6/26/2016 4:51 PM
Deleted: [STL: I will work on this part]

Shantao 6/27/2016 11:40 AM
Deleted: Interestingly

Shantao 6/27/2016 11:40 AM
Deleted: although

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.

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

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 or equals zero 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 49 samples from single-center, automated calls to form an extensive patients set.

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.

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.

Methylation association analysis

In total, we collected HumanMethylation450 BeadChip array data for 139 samples that are either methylation cluster 1 or 2 (REF). We used an R package

“IMA” to facilitate analysis [REF]. 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 must 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.

Methylated CpG mutation rates are calculated by first obtain +/- 100bp context of each hypermethylation site. Then we counted the number of N[C>T]G and C[G>A]N divided by all NCG/CGN motifs in the context. To get empirical p-value, we randomly permutated the labels of hypermethylation sites for 1,000 times to establish the distribution of methylated CpG mutation rate.

APOBEC enrichment analysis

We used the method described by XXX [REF]. For every C>{T,G} and ./oG>{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.