[CANDISP3 | Main Text]

A. Significance

[[HM - update the signific
w/ DC add a para on VHL parallel to the one MET -- [[done]]
Add a para on the interesting Qs
Som-germline
Tsg-onco - LOF v GOF]]

LOGIC

The MET protein (encoded by the c-Met gene) is a tyrosine kinase that functions as a membrane receptor. It plays key roles in both organism development as well as tissue growth. Given that it may function as an oncogene, hyper-activation of MET may result in rapid tumorigenesis and poor patient prognosis. MET plays especially prominent roles in cancers of the liver, brain, kidney, stomach, and breast. Consistent with its well-characterized roles in growth and development, it is normally only expressed in stem cells and progenitor cells. A number of specific mutations in MET have been shown to be especially important in oncogenesis, which we propose to study in the work detailed here.

[[dc2HM -- below is a paragraph on VHL -- I'd be happy to expand on this if you like]]
[HM2DC - I reckon it should be fine esp. in comparison with the MET paragraph above]
VHL (von Hippel–Lindau) encodes a subunit of a complex that is responsible for downregulating other proteins through ubiquitin ligation. In particular, one target of this complex is HIF1a, which promotes angiogenesis (a characteristic hallmark of solid tumors). As such, VHL is a tumor suppressor gene (TSG). Inherited mutations within this gene have been linked to a number of cancer types, including those in the kidney, pancreas, and brain. Cancer-related variants in VHL exhibit properties of the 'two-hit' hypothesis of oncogenesis: an individual born with a variant in one copy of VHL confers predisposition to cancer, as random mutations to the only healthy copy over the course of an individual's lifetime can result to total loss of this tumor suppressor, thereby promoting oncogenic initiation.

[[
dc2HM: I think we just need to figure out the best place to insert this reference to M. Rubin's paper -- also is the below ref correct (ie, is this the paper to which MG was alluding?
Recurrent somatic SNVs identified in other cancer types have been shown to explain racial disparities in patient outcomes, some of which exhibit racial disparities \cite{28515055}.''
]]

[[HM2DC: I inserted a reference to M. Rubin et al. in section on somatic-germline co-occurrence paragrah - currently under "*Inter-relate (co-occurrence) somatic & germline.*" The one I referred to is this new paper on prostate cancer [here] - STL may confirm if that's exactly the one.]]

The X-ray crystal structure of the c-Met kinase domain, resolved to a resolution of 1.8 angstroms (PDB ID: 1R1W).

[HM - para on interesting Qs: som-germline interactions and TSGs vs OncoGs + co-variates]

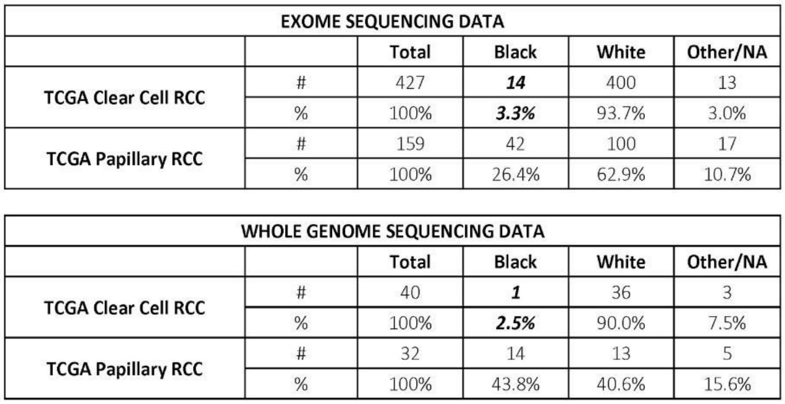
Understanding the underlying genetics of racial disparities in RCC is a crucial, broad question. In this context, we will address a number of related questions. As per the Knudson hypothesis \cite{two-hit hypothesis foundational papers}, where an accumulation of genetic mutations lead to the incidence of cancer, we will investigate how somatic and germline variations complement each other leading to the incidence of RCC. Analyzing the patterns of somatic-germline interactions in African American and Caucasian patients can shed the light on possible differences associated with the etiology of the disease. We will also will analyze the frequency, functional impact, and genomic burdening of loss-of-function (LoF) and gain-of-function (GoF) variants across samples. With a focus on *MET* and *VLF*, an oncogene and a tumor-suppressor gene, respectively, we will study the patterns of LoF and GoF variation in tumor suppressor and oncogenes in patients from different races. In addition, and to gain a more comprehensive insight into the reasons behind existing racial disparities, we will utilize electronic health records in an attempt to identify clinical and environmental factors in relationship with RCC incidence.

B. Innovation

C. Approach

Aim 1: To perform whole genome sequencing (WGS) of African-Americans with ccRCC to complete a missing aspect of the cancer genome atlas (TCGA).

C-1-a **Rationale:** In recent years, TCGA efforts have furthered our understanding of the genomic basis of various forms of kidney cancer. TCGA studies have led to the understanding that different cell types within the kidney may give rise to distinct forms of kidney cancer. Somatic alterations (driver mutations and copy number variants) are also important in determining a cancer's molecular profile. In TCGA, kidney cancer cases were submitted from various high volume tertiary centers to the Bio-specimen Core Resource (BCR) for accessioning and specimen processing. However, specimens were not submitted in a coordinated fashion to ensure a study population of similar profile to that encountered nationally. Not surprising, there were a limited number of African-Americans with clear-cell kidney cancer included in TCGA analysis. Despite African-Americans accounting for approximately 1 in 7 cases of ccRCC, only a cursory analysis was performed in this population, including 14/427 (3.3%) samples that underwent whole exome sequencing and 1/40 (2.5%) ~~[[hello?]]~~ (Table 2) that underwent whole genome sequencing. A failure to include a larger population of African-Americans with clear cell RCC limits our ability to explore genomic bases for racial disparities. This contrasts with higher incidence of pRCC in African-Americans, the pRCC TCGA cohort was able to include a larger number of African-Americans. However, despite available data, there has not been a thorough analysis of somatic driver alterations or germline risk variants more prevalent in African-American kidney cancer. We propose to complete TCGA analysis of the top two subtypes of kidney cancer, papillary and clear cell, by analyzing an additional cohort of African-Americans with ccRCC. By performing whole genome sequencing on this additional cohort of samples, we will have an adequate number of cases to allow balanced comparisons between African-American and Caucasian clear cell and papillary kidney cancers.



EXOME SEQUENCING DATA					
		Total	Black	White	Other/NA
TCGA Clear Cell RCC	#	427	***14***	400	13
	%	100%	***3.3%***	93.7%	3.0%
TCGA Papillary RCC	#	159	42	100	17
	%	100%	26.4%	62.9%	10.7%

WHOLE GENOME SEQUENCING DATA					
		Total	Black	White	Other/NA
TCGA Clear Cell RCC	#	40	***1***	36	3
	%	100%	***2.5%***	90.0%	7.5%
TCGA Papillary RCC	#	32	14	13	5
	%	100%	43.8%	40.6%	15.6%

Table 2: Racial and histologic distribution of available whole exome and whole genome data available from TCGA datasets.

Furthermore, using a patient cohort with a different genetic background, sequencing might illustrate novel, ethnic-specific driver events as recently seen in an African American prostate cancer study \cite{28515055}.

Whole genome sequencing offers several advantages over chip-based methods. It allows analysis of poorly-tagged or rare SNPs, INDELs and structural variants (SVs). Moreover, whole genome sequencing has nucleotide resolution which helps pin down the disease causing variants rather than big DNA blocks in linkage disequilibrium.

C-1-b Sample acquisition, comorbidity/demographics matching, and DNA extraction: All patients undergoing scheduled kidney cancer surgery at Yale New Haven Hospital are offered enrollment into an IRB-approved Genitourinary Biospecimen repository (P.I. Shuch, HIC# 0805003787). Within 30 minutes of removal, fresh tumor tissue is snap frozen in liquid nitrogen by the pathology team. Additionally whole blood is procured to serve as a genomic control. In the past 2 years, over 300 subjects with kidney cancer have been prospectively enrolled. All fresh bio-specimens are stored at -80˚C and are available for immediate analysis. For the purpose of completion of the TCGA dataset, we will utilize a 15 African-American subjects with ccRCC from 2013-2016. TCGA kidney cancer projects have captured patient age, sex, race, smoking history, and has limited information from a secondary analysis on obesity status. Self-reported racial identity may be imprecise, yet is necessary to account for patient demographics and the influence of RCC comorbidities. We therefore intend to prospectively genotype candidate individuals for WGS, to ensure they follow the same racial distribution as in TCGA. To determine the ideal candidates for WGS we will employ both phylogenetic and data mining clustering methods (See section C-4-d).

C-1-c WGS and variant calling: Sequencing of normal and tumor samples will performed using Illumina's Hiseq 2000 technology. In brief, DNA fragments from each sample will be hybridized using HiSeq Paired-End Cluster Kits and will be further amplified using the Illumina cBOT.

Paired–end libraries will be generated by utilizing HiSeq (2x101) cycle and imaging will be performed by TruSeq kits.

We have extensive experience in large-scale variant calling and interpretation through active membership in the 1000 Genomes Consortium, particularly from our participation in the analysis working group and the structural variant (SV) and functional interpretation (FIG) subgroups of the consortium, where the majority of the variant calling tools were developed, deployed and interpreted [23]–[25]. We have already set up a prototype pipeline for calling germline and somatic variants. We will use the Genome Analysis Toolkit (GATK) [26] to call germline SNPs and INDELS. We use parameters consistent with those used in TCGA [27]. We will map raw FASTQ files of each sample to the hg19 reference genome using the bwa-mem algorithm with default parameters to generate BAM files. These bam files will be further processed to sort and mark duplicates reads before calling variants.

UPDATE PCAWG ?

We will follow GATK best practices [26] to generate initial raw variant call sets using GATK haplotype caller. We will filter these initial call sets by running GATK variant recalibration tool. This filtering strategy based on a variant recalibration method uses a continuous adaptive error model.

The adaptive error model takes into account variant annotations including quality score, mapping quality, strandedness and allele information. Using this information, it classifies variant calls as true positives or sequencing artifacts. We will exclude any filtered variant, which falls in a low mappability region of the genome. MuTect [28] and Strelka [29] will call somatic SNVs and INDELs, respectively.

Structural variations (SVs) are important contributors to human polymorphism, have great functional impact and are often implicated in diseases including cancer. We have developed a number of SV calling algorithms, including BreakSeq which compares raw reads with a breakpoint library (junction mapping) [30], CNVnator which measures read depth [31], AGE which refines local alignment [32], and PEMer which uses paired ends [33]. We have also developed array-based approaches [34] and a sequencing-based Bayesian model [35]. Furthermore, we have studied the distinct features of SVs that originate from different mechanisms, and showed how creation processes may have potentially divergent functional impacts [36], [37]. We will perform extensive molecular characterization of germline and somatic SVs in these cancer samples. We will run CNVnator to identify germline and somatic copy number variations in each cancer sample. We will apply CREST [38] to identify germline and somatic large structural variations including large deletions, insertion, inversion, intra and inter-chromosomal translocations. Furthermore, we will run our BreakSeq tool to decipher the underlying mechanism of somatic and germline SV formation.

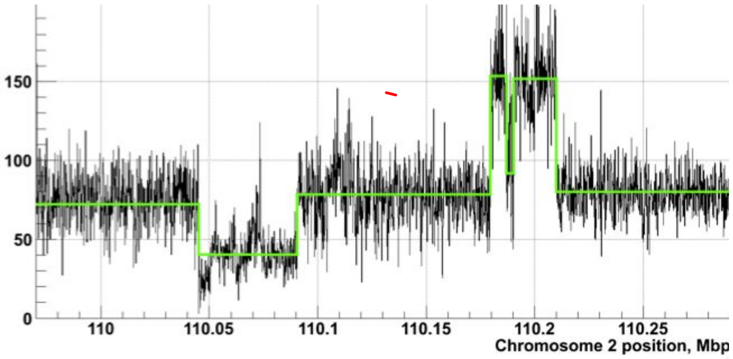


Figure 3: Read depth based identification of copy number variation by CNVnator.

Along with our new sequenced samples, we will reprocess all the TCGA data using our own calling pipeline, to mitigate any potential processing or batch effects.

C-1-d **Deliverables:** In this aim, we will generate an extensive catalogue variants, for both African-American ccRCC patients at Yale, and TCGA kidney cancer patients. This will be done consistent with methodology already used in TCGA. This catalogue will include both germline and somatic variants, including SNPs, INDELs and large SVs. We will cover both coding and non-coding regions of the genome. Our catalogue of variants, will serve as an excellent basis for identification of genomic aberrations associated with racial disparity observed in kidney cancer. We plan to make our sequencing data available via dbGAP (see data dissemination plan).

SAME CONSENT

Aim 2: To find MET-related key mutation patterns, regions and residues associated with kidney cancer [[find key racially disparte snvs]]

C-2-b Relevant Preliminary Results:

C-2-b-2 We have developed tools for somatic and germline burden tests:
[[DC & SK: stress & frustration & alfot]\]] [[done]]

2ND SITE

DESC LOGIC

We have developed a number of software tools that have been designed to annotate and understand the effects of variants within the coding regions of the human genome. Coding variants are first annotated (for example, determined to be synonymous, non-synonymous, premature stop codons, splice-site change, etc) using our VAT software \cite{22743228}. Once mapped to 3D structures from the PDB, the effects of annotated variants may be studied in detail by measuring their associated effects in the contexts of both allosteric regulation and local mechanistic perturbations.

With respect to allosteric effects, we have developed the STRESS software tool \cite{27066750}. STRESS employs models of large-scale protein conformational change in order to predict key allosteric residues on both the protein surface (by finding essential pockets) as well as the interior (by identifying information-flow bottlenecks). Our reported results demonstrate that this software selects residues that are highly conserved over both long- and short evolutionary timescales, and it has also been used to help rationalize otherwise poorly-understood ("cryptic") disease-associated SNVs.

With respect to localized perturbations, we have reported a separate study \cite{27915290} to demonstrate how localized changes in biomolecular frustration may be used to better understand the differential effects of variants in oncogenes and TSGs. Specifically, these results shed light on potential gain-of-function variants on the surfaces of oncogenes, and loss-of-function variants in TSGs.

We have also developed ALoFT, a tool specifically tailored to annotate and predict the disease-causing potential of loss-of-function events \cite{28851873}. Short for "annotation of loss-of-function transcripts", ALoFT has been used to successfully discriminate between LoF mutations that are deleterious in heterozygous states from those that may cause disease in the homozygous state. We analyzed somatic variants in more than 6500 cancer exomes, and demonstrated that variants predicted to be deleterious by ALoFT are enriched in canonical driver genes.

We have worked on statistical methods for analysis of non-coding regulatory regions. LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichments in noncoding elements, by comparing observed mutation counts with expected counts under a whole genome background mutation model [53]. LARVA includes corrections for biases in mutation rate owing to DNA replication timing. LARVA can be targeted to coding regions to prioritize genes. We used this tool in a pan-cancer analysis of variants in 760 cancer whole genomes, spanning a number of cancer data portals and published datasets. Our analyses demonstrated that LARVA can recapitulate previously established coding and noncoding cancer drivers, including the TERT and TP53 promoters [53]. Furthermore, we have developed MOAT (Mutations Overburdening Annotations Tool), an alternative empirical mutation burden approach that evaluates mutation enrichments based upon permutations of the input data (submitted). Both annotation-based and variant-based permutation is supported.

C-2-b-3 We have identified regions associated with kidney cancer through our involvement in the papillary TCGA team: Given that Yale has expertise in the clinical management and genetics of kidney cancer, we were invited to participate in TCGA kidney cancer projects. Our role in the TCGA KICH (chromophobe RCC) included coordination of the Cancer Cell manuscript [[need citation]]. Our team analyzed the whole genome sequencing data for the TCGA KIRP (pRCC), now published in New England Journal of Medicine [27]. In recent work, we leveraged our expertise of studying non-coding regions in the first whole genome analysis of pRCC samples \cite{5391127}. Our work finds significant genomic alterations beyond traditional known drivers of pRCC. We hypothesize that these alterations may have non-canonical effects on known tumorigenic pathways (for example, MET in type 1 pRCC). We discovered genomic markers in MET and NEAT1 that predict prognosis. We investigated mutational signatures and the mutational landscape and evolutionary trees in pRCC and identified several meaningful etiological factors explaining inter-patient genomic variation in pRCC. This experience provided further practical knowledge of working with available RCC genomic datasets. Finally, our team has participated in two ongoing pan-RCC manuscripts serving a central role assessing evaluating the cluster of cluster assignments (COCA) immunologic profile from gene and miRNA expression datasets. Together with other published results on RCC [54]–[58], we have assembled an extensive list of impactful and statistically significant regions of RCC genomes.

RS PHAT

[[SK to put a para on funseq + pcawg]] [[done]]
C-2-b-4 We have extensive experience in analyzing whole-genome datasets from cancer cohorts: We are active participants in The Cancer Genome Atlas (TCGA) and Pancancer Analysis of Whole Genomes (PCAWG) consortium projects \cite{}. Specifically, we have played key roles in the TCGA investigations into prostate \cite{26544944} and kidney \cite{26536169} cancers. More recently, we have conducted a detailed investigation into the noncoding variants of kidney papillary cancer samples in TCGA \cite{28358873}. As part of the driver discovery subgroup in PCAWG, we have participated in a comprehensive variant prioritization exercise to generate a catalogue of driver elements in many cancer cohorts. Furthermore, we are currently leading the PCAWG group investigating the aggregated impact of mutations on cancer development, progression, and prognosis. As part of this effort, we ran FunSeq on each variant(~30 million total somatic mutations among 39 cancer subtypes) in PCAWG. We identified many high-impact mutations, in addition to canonical driver mutations, which can potentially influence cancer progression.

C-2-c Research plan:

C-2-a Assessing the functional impact of coding mutations in MET

C-2-a-0 [[LS, PDM] Constructing the MET-ome & VHL-ome: integrating MET- And VHL-associated elements across annotations.

WE WILL LINK

LOGIC

High impact regions associated with MET and VHL are linked to other genomic regions through functional relationships that exist across networks of biomolecules. Examples of these connections include physical interactions among the molecular binding partners of MET and VHL, or the gene regulatory networks that influence MET and VHL expression.

We plan to link genomic regions associated with MET and VHL across functional annotations. For example, we'll link transcription factors to enhancer elements, and enhancers to their target genes. We'll seek to clarify the influence of distal epigenetic regulatory markers, like methylation and chromatin-state, on MET and VHL expression. We'll use protein interaction networks to better understand broader consequence of variation as transmitted through a molecular interaction network. We'll build maps of the molecular pathways influencing MET and VHL function.

This integration of will produce an extended MET and VHL annotation. Genetic modules will group potentially impactful elements that share similar or collaborative biological functions. These groupings will increase the statistical power in our study for resolving contributory genetic variation. Genetic modules also offer annotation of lesser known noncoding regions. Our results have biologically interpretability because genetic modules will be linked with genes.

ANNOTES ?

C-2-a-0 building a mut catalog for the METome [[STL]]

Est. the number of mutations
Get al muts from somatic,pcawg, exact, gnome
Est the size

1 para.

WHERE SOM TCGA + US

C-2-a-2 Build a comprehensive mutation catalog for the METome and VHLome. We aim to build an all-inclusive, comprehensive mutation catalog with variants assembled from both our dataset and public available data. First we will perform a literature search, identifying previous work documenting association between genomic alteration and RCC. We will gather genetic changes that include single nucleotide variation (SNV), structural variation/copy number variation (SV/CNV), and mutation process signature. We will also annotate regions that are associated with disparity between Caucasian and African-Americans in other forms of cancer, such as prostate cancer [59], [60]. Prior study has shown that RCC is uniquely characterized by copy number variations (CNV) as an early and major driver event [54]. Because repeats are triggering factors for many structural variation events, we will pay particular attention to repeats polymorphisms around known cancer associated genes, and recurrent CNV regions in RCC. Repeats may put certain RCC related genes at predisposition to CNV events. Last, we also gather somatic RCC mutations from TCGA and PCAWG project. For background mutation landscape in general population, we will leverage on gnomAD and ExAC.
ExAC and gnomAD reports 677 variants in ~31k alleles in MET and additional 1218 variants in 250k exome-sequenced alleles. In VHL, the numbers are 448 and 225 respectively. In 35 pRCC whole genomes, we found 12 somatic MET mutations. In TCGA….PCWAG...WXS…
By functional elements linking, we expect the regions will grow exponentially with the association degree. We expect the mutation number will grow by one order. … and estimate…around germline 20,000 SNVs from public dataset and somatic 5000 (??) SNVs.

?

[STL: I have shaky HPC connection, will update the number estimate]

DONE

*** C-2-c-2 Run our coding variant prioritization pipeline on all variants (func impact coding): [[sK, dc]] [[done]]

WE HAVE MUT THEN PRIORITIZE

The MET and VHL genes are known to play important roles in many cancer types, including those of the kidney. As mentioned in our significance statement, MET and VHL function as TSGs and oncogenes, respectively. Intense research efforts to gain mechanistic insights into the functioning of these genes have resulted in detailed 3D models of their structures. Using high-resolution crystallographic models of these two proteins, we will employ a number of our tools to evaluate the functional impacts of their point mutations. We will first apply VAT \cite{22743228} to annotate coding variants of the MET and VHL genes within our sample cohort. Furthermore, we will map their coding variants onto their respective crystal structures and then apply our STRESS \cite{27066750} tool to identify cryptic sites influenced by non-synonymous coding mutation. The cryptic sites identified through STRESS play important roles in allosteric regulation. Non-synonymous mutations at allosteric sites are likely to affect proper functioning. We will also apply our Frustration tool to prioritize non-synonymous mutations by identifying those variants that disrupt the local stability of MET and VHL proteins. Such detailed mechanistic annotations of the roles of individual variants on MET and VHL protein functionality may provide essential resources for targeted drug development. We also note that the differential and highly heterogeneous effects of SNVs (as elucidated through STRESS and frustration profiles) may also provide prognostic value. In addition, we will apply our ALoFT tool to identify loss-of-function (LoF) coding mutations potentially inactivating copies of MET and VHL genes in our study cohort.

MET/VHL

*** C-2-c-3 Run our variant prioritization pipeline on all variants (noncoding func impat): [[sK, dc]] [[done]]

In addition to coding variants, many changes in noncoding regions may play critical roles in renal cell cancer initiation and progression. In order to identify high-impact mutations in noncoding regions for both cancer and normal samples, we will run our updated and extended FunSeq pipeline on the kidney cancer variant catalogue. As part of our initial analysis, we ran FunSeq on and carefully curated the results. We found several disruptive mutation hot spots in within the genome. With the addition of many more samples, we will perform comprehensive prioritization analyses to identify many more non-coding variants that may play key roles in kidney cancer.

*** C-2-c-4 Run our variant prioritization pipeline on all variants (recurrence):

Prioritizing variants within non-coding regions of the genome is especially challenging. Thus, we will apply the alternative approach of evaluating variant recurrence to identify key mutations in kidney cancer. We will apply our LARVA and MOAT tools on the comprehensive

CODING TOO

kidney cancer variant catalogue. Our prior analysis of TCGA whole-genome sequenced samples indicate the presence of excessive somatic mutations in the MET intronic and promoter regions, along with several other recurrent mutated regions that merit further investigation. We expect to further identify other important variants in kidney cancer with large-fold increases in our kidney cancer variant catalogue.

INTEG

C-2-c-5 Identify critical regions burdened by germline mutations: Above, we have explained our approach to key somatic regions associated with kidney cancer. In this section we explain our approach to germline variation. Statistics for germline variants are different than for somatic ones, and thus require a different approach to analysis. First, using SKAT [61] we will find MET associated regions that are significantly burdened by germline mutations in the kidney cancer cases versus healthy controls. As non-cancer controls, we will use both the 1000 Genomes Project (2504 individuals) whole genome samples, as well as the Exome Aggregation Consortium (ExAC, TCGA samples excluded) meta-cohort [60]. To mitigate genetic background as a confounding factor, we will match our patient samples with normal controls using both self-reported ethnicities and racial SNP markers. We will look for regions and genes that are burdened significantly in RCC compared to control. Given the size of these datasets, we will be well powered in our testing (see also cancer population sampling discussed in aims 3 and 4).

We will mask known SNPs and flanking regions associated with high BMI [62], hypertension [63], cigarette smoking [64] and other known risk factors in previous association studies, to reduce the possibility of misattribution of these known RCC comorbidities to direct genetic effect.

1

}}}

Inter-relate (co-occurrence) somatic & germline Degeneralize - coocc{{

SHORT
EN
&
INT
SG

[HM, STL] **C-2-c:**

We will comparatively investigate mutation patterns in African American and Caucasian cohorts in the combined Yale-TCGA RCC data. Differences in major and minor allele frequency distributions across patient samples is an important factor we will take into account. Somatic variants can complement cancer-related germline ones and lead to the development of different types of cancer. We consider this co-occurrence a mode of somatic-germline variant interaction, and we will study the interaction patterns in patients across races at the levels of coding and noncoding regions associated with all human genes and known COSMIC genes. In addition, along the lines of other cancer studies that focus on specific genes \cite{mrubin_etal}, we will focus on *MET* and *VLH*. Variations in interaction patterns of somatic and germline variants associated with genes can also help us locate regions of interest and unravel part of the intricate underlying genetics that lead to RCC across races. To perform the aforementioned tasks, we will develop new methods and leverage an amalgam of tools and pipelines we have developed to prioritize variants. Alongside with genomic samples in the combined Yale-TCGA cohort, we will mine several genomic repositories including ExAC \cite{exac}, 1000 Genomes (1KG) \cite{1KG paper(s)}, and gnomAD \cite{gnomAD}.
}}

**** Find the impactful & recurrent mutations that are racially disparate {{[[HM, PDM]]

Following prioritization of MET- and VHL-related variants across all TCGA-Yale samples, we will study differences in the population frequency, functional impact, and genomic burdening between variant sets to identify racially disparate genetic elements. The variety of tools we have developed in the lab allow us to flexibly identify impactful variants and mutational signatures in novel data sets.

We will also analyze patterns of somatic and germline variation in samples from both African American and Caucasian patients. By identifying relationships between recurrent somatic and germline mutations, we may identify novel germline mutations that predispose to renal cancer. This analysis provides an opportunity to identify genetic signatures and impactful and recurrent mutations that partially explain racial disparities in RCC.

[[PDM2HM: The above 1-2 paragraphs are a bit non-specific (e.g. no mention of methods, specific tools, etc.) Could be a good thing or a bad thing -- provides flexibility in later approach, but reviewer may question how exactly we accomplish this. Of course, some methods are listed in other sections of grant app.]]

Graceful cut {{
We will integrate our previously developed tools for variant prioritization (see section C-2-b-1) into a unified software pipeline in order to investigate the impacts of SNVs throughout coding elements of the genome. In particular, we will run our pipeline on all somatic variants that fall within X-ray crystal structures within the PDB. The computational efficiency of each of these software tools will enable somatic variant evaluation within a matter of days or weeks. This workflow will also include an annotation of somatic variants in the context of conserved protein motifs (our proposed term for this new method is "Intensification"), and the entire integrated pipeline will constitute our "Interrogation" workflow. This integrated package will be made available on GitHub (with each component having a dedicated user-friendly web page), and the results of our somatic variant analysis (and in particular, the atlas of high-impact somatic SNVs) will be made available as downloadable data files for downstream analysis to other investigators.
}}

Aim 3: To correct for and study clinical and environmental co-variates using electronic health records of RCC patients

[HM] **C-3-a** (ideas on Preliminary Results/Rationale/Deliverables split?)

Racial disparities in cancer is likely the result of a multiple factors. Genetics might provide a valuable insight into kidney cancer etiology, but because of the breadth of the undertaking, we plan to approach the problem from other perspectives. In this context, we plan to (1) rigorously correct stratification and biases in samples and (2) find significant correlations between clinical and environmental conditions and the disease incidence. We will analyze the electronic health records of all patients in the TCGA-Yale cohort and identify any statistically significant relationships between health and living conditions on one side, and the genotypic and phenotypic aspects of RCC cases present in the cohort. The ultimate goal is to attain the ability to predict RCC incidence based on a combination of genetic and non-genetic factors and to help in crafting recommendations that would help in eliminating existing racial disparities. The analysis pipeline will be automated to accommodated for additional genomics and electronic health record data to be collected during or after the study.

<u>**C-3-c-3**</u> <u>**Power analysis using SKAT for per region based analysis:**</u> In the above, we plan to use aggregated burden tests (e.g. SKAT) to look for differential burdening between populations and use this to rank the regions. While we are not striving for absolute statistical significance in differential burdening, our sample size is provides an appreciable signal for ranking. Here, we discuss the power aspects of burden tests applied to our sample populations. SKAT analysis has been developed for rare genomic mutations, and remains robust for common variants. We will utilize SKAT to test for significant disparity of variants in kidney cancer between Caucasian and African-American populations. To estimate the sample size needed to obtain statistical power, we ran a SKAT package available from the R project, on several population models for genomic regions of 5k nts(Figure XXX). In our proposed study, we will focus on genomic modules linked with kidney cancer such as MET- and VHL-ome, and therefore expect a large number of effective mutations. Typically, the MET genomic region consists of 126,027 nt while VHL of 12,035 respectively. We expect these numbers to increase significantly after creating the modules.

(LATER

[[[***Last year***SEE BELOW**]]]

<u>C-3-c Compare germline mutations in coding regions between Caucasian and African-Americans in prioritized regions using WES Data:</u>

REDUND

<u>C-3-c-1</u> <u>Variant level analysis:</u> For coding region analysis, we will employ the full 556 samples with whole exome data from TCGA. For common variants analysis at a single locus, Fisher's exact test can be used to evaluate the racial disparity between Caucasian and African-American subjects with RCC. Here, we prioritize common variants according to their associations with RCC disparity in race. For a common SNP identified in African Americans and Caucasians with RCC, we record minor allele frequencies and major allele frequencies in African Americans and Caucasians with RCC. For these counts of a focal SNP, the Fisher-exact test is used to determine whether the SNP tends to be associated with the African Americans with RCC. The p-values of tests for all common variants are used to prioritize variants for further study and validation. The power of the Fisher exact test can readily be estimated in this context. For instance, for an ordinary SNP with allele frequency 7% in the total samples, when its frequency in the African American subjects is 12%, the power of the test can reach 0.4 with a p-value < 5e-5. This indicates that these SNPs can be detected with statistical significance from 1000 candidates, even when the most conservative Bonferroni correction is used.

<u>C-3-c-2</u> <u>Region based analysis</u>: Beyond investigating the association between single common variants and race, we will focus on the evaluating the cumulative effects of a set of rare variants in genomic regions, such as genes, using both burden and non-burden test. Burden tests are often applied on regions where most of the variants in the same region are causal, affect

phenotype in the same direction (e.g. LOFs disabling a tumor suppressor). We assume that in total there are *n* patients with whole exome sequencing data available. Also for a target region, for example, a gene, there are *m* variants. Let y_i denote the population information of the i^{th} patient. $y_i = 1$ for African-Americans and 0 otherwise. Let $\mathbf{G}_i = (g_{i1}, \ldots, g_{im})'$ represent the genotype of patient i . Then a logistic regression model can be set up to evaluate the association as in (1). Suppose that π_i describes the mean of the population status, then:

$$\mathrm{logit}(\pi_i) = \gamma_0 + \mathbf{G}_i'\mathbf{b} \quad (1)$$

For the burden test, we could treat the coefficient b_j for each patient as a weighted coefficient like $b_j = w_j \times b_c$. Then equation (1) can be rewritten to:

$$\mathrm{logit}(\pi_i) = \gamma_0 + b_c\left\{\sum_{j=1}^{m} w_j g_{ij}\right\} \quad (2)$$

Under the null hypothesis that there is no association of variants in this region with race, the coefficient b_c should be zero. The test statistic for H0: $b_c = 0$ becomes:

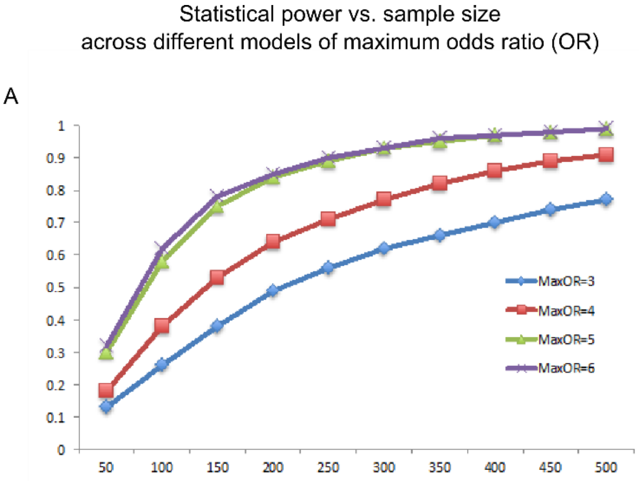
$$Q_B = \left[\sum_{i=1}^{n} (y_i - \hat{\pi}_i)\left(\sum_{j=1}^{m} w_j g_{ij}\right)\right]^2 \quad (3)$$

The allele frequency can be used to assign the weight for each variant. For example, $w_j = 1\Big/\sqrt{\hat{p}_j(1-\hat{p}_j)}$, where $\hat{p}_j$ is the minor allele frequency. However, in some cases, where the target region has many non-causal variants or the effect of such variants is heterogeneous, burden tests, such as equation (3), may lose statistical power. Here, a sequence kernel association test (SKAT) can be used. Instead of assuming a weighted coefficient effect in the burden test, each b_j is treated as an independent random variable with 0 mean and variance $w_j^2\tau$. Then the null hypothesis can be changed to H0: $\tau = 0$, and the test statistic under equation (1) can be written into:

$$Q_S = (\mathbf{y}-\boldsymbol{\pi})'\,\mathbf{K}(\mathbf{y}-\boldsymbol{\pi}) \quad (4)$$

In (4), $\mathbf{K} = \mathbf{GWWG}'$ is the kernel matrix, and $\mathbf{G}$ is the genotype information vector. $\mathbf{W} = diag\{w_1, \ldots, w_m\}$ is the weight matrix which can employ allele frequency or other external information, such as conservation score. The test statistic in (4) can be rewritten into

$$Q_S = \sum_{j=1}^{m} w_j^2 S_j^2 = \sum_{j=1}^{m} w_j^2 \left\{\sum_{i=1}^{n} g_{ij}(y_i - \hat{\pi}_i)\right\}^2 \quad (5)$$



B

Maximum OR	3	4	5	6
effective sample size	763	326	172	160

In coding variant analysis, because we generally do not know which of the two cases each gene falls into, a unified test is the following:

$$Q_\rho = \rho Q_B + (1-\rho)Q_S, 0 \le \rho < 1 \quad (6)$$

Since the best route in (6) is unknown, a best test statistic can be used as follows:

$$Q_{opt} = \min\left(Q_{p1}, \cdots, Q_{p_k}\right) \quad (7)$$

<u>**C-3-c-3 Power analysis using SKAT for per region based analysis**</u>**:** In the above, we plan to use aggregated burden tests (e.g. SKAT) to look for differential burdening between populations and use this to rank the regions. While we are not striving for absolute statistical significance in differential burdening, our

sample size is provides an appreciable signal for ranking. Here, we discuss the power aspects of burden tests applied to our sample populations. SKAT analysis has been developed for rare genomic mutations, and remains robust for common variants. We will utilize SKAT to identify genomic regions with significant disparity of variants in kidney cancer between Caucasian and African-American populations. To estimate the sample size needed to obtain statistical power, we ran a SKAT package available from the R project, on several population models (Figure 6). In our proposed study, we will focus on genomic modules linked with kidney cancer, and therefore expect a large number of effective mutations.

Figure 6: Using the default haplotype information in the SKAT haplotypes dataset, we randomly selected subregions of size=5k and ran 100 simulations. In A, we show the statistical power obtained across the different models of maximum Odds Ratio. In B) we show the required sample size for each of these models in order to obtain significant statistical power (α=0.01, β=0.2)

C-3-d Compare germline mutations in noncoding regions between Caucasian and African-Americans in prioritized regions using WGS Data:

C-3-d-1 Pooled variant test for limited target regions: For our noncoding region analysis, since we have limited power with 32 WGS samples in both populations, targeted analysis will be carried out on a smaller set of regions. From our experience with TCGA KIRP, we have already prioritized MET intronic and promoter regions, along with several other recurrently mutated regions that merit further investigation. We will focus on these selected regions, to enable use of the unified statistical testing mentioned above (in section C-3-c).

C-3-d-2 Non-parametric test for FunSeq score distribution difference: We expect that casual regions may not only be under differential mutational burden between races, but may also be disproportionately affected by high-impact mutations. Thus, for prioritized regions given above, we plan to calculate all FunSeq scores on both African-American and Caucasian populations. By subsequently ranking and pairing scores between the two population groups, we intend to use a Wilcoxon signed-rank test to evaluate the significance of mutational impact on each region. This test is a non-parametric version of the paired t-test, and is used when we cannot assume that the populations follow a normal distribution. As population size increases, a Z-score can be calculated.

C-3-e Compare somatic mutations between Caucasian and African-Americans in prioritized regions: Previously, we developed an integrative framework, LARVA, to discover highly recurrent regions in cancer genomes as candidate cancer drivers [53]. It is known that various genomic features affect background mutation rate in most cancer types, and this results in numerous false positives in somatic mutation recurrence analysis [65]. Hence, we have extended LARVA into a new system, NIMBus (a Negative Binomial Regression based Integrative Method for Mutation Burden Analysis), which incorporates corrections for additional covariates that influence somatic mutation rate in genomic regions. These covariates include sequence content, replication timing, expression level, histone modification marks, and chromatin status. Specifically, in a region with length l, suppose the mutation rate is known as μ, then the number of mutations y within l given μ should follow a Poisson distribution as follows:

?

$$p_Y(y|\mu) = \frac{e^{-\mu}\mu^y}{y!} \quad (8)$$

However, we discovered in our previous analysis, that there is significant cancer type, sample, and regional heterogeneity in mutation count data [53]. Such mutational heterogeneity violates a constant mutation rate assumption and results in over-dispersion. Hence, instead of supposing μ is constant, we set up the following model:

$$p_Y(y|\mu\gamma) = Poisson(\mu\gamma)$$
$$\gamma \sim Gamma\left(1, \frac{1}{\sigma^2}\right) \qquad (9)$$

?

The marginal distribution of *Y* can be expressed as the type I negative binomial distribution:

$$p_Y(y|\mu,\sigma) = \frac{\Gamma\left(y+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)\Gamma(y+1)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y\left(\frac{1}{1+\sigma\mu}\right)^{\frac{1}{\sigma}} \qquad (10)$$

Where $E(Y)=\mu$, $Var(Y)=(1+\sigma)\mu$. Let $x_1, x_2, \cdots, x_k$ be the genomic covariates to be corrected, such as replication timing, GC content, and chromatin status. We can then use the following negative binomial regression to estimate the local mutation rate under the covariant set:

$$g_1(\mu) = \log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$
$$g_2(\sigma) = \log(\sigma) = \alpha_0 \qquad (11)$$

Consequently, instead of estimating a genome wide mutation rate, we are now estimating a coefficient vector for the mean, and a constant over-dispersion value. For each region to be estimated, a local mutation rate can be reconstructed by equation (11), for accurate background rate and false positive/negative controls.

We will start off with prioritized regions from aim 2 to minimize multiple testing issues. For each region we will use our procedure to identify differential somatic burdening. In particular, we will apply our new method on the 16 African-American and then 16 Caucasian WGS samples separately. Highly recurrent regions in each will be reported and compared. Those regions that are unique to either population will be prioritized for detailed validation.

<u>C-3-f</u> <u>Deliverables</u>: This aim will create ranked lists of genes, non-coding regions and variants from Aim 2, to pass to validation in Aim 4. We will combine rankings from the categories of variants described above, by comparing their corresponding p-values. To obtain results with greatest possible significance, we will aim for a minimum number of validations from each category. Also, we plan to make our racial disparity rankings of genes and non-coding regions publicly available from our project web server (see data dissemination plan).

Aim 4: To validate specific regions with either germline or somatic mutations suspected of contributing to kidney cancer racial disparity.

<u>C-4-a</u> <u>Rationale</u>: Aims II and III together represent a discovery phase, where we identify and prioritize genomic alterations associated with MET and VHL genomic regions for racial disparities associated with cancer. In Aim 4 an independent patient cohort will be used to validate findings from our discovery phase. The independent validation cohort includes patients with RCC from Yale's Genitourinary Biospecimen Repository, in addition to availability of statewide sampling of patients with RCC through the Connecticut Tumor Registry. We intend to validate 55 regions (100bp each) for 384 individuals. This will contain both African-Americans and Caucasians with clear cell and papillary RCC, to allow comparisons across histologic type and race.

Apart from confirming associations between genomic alterations and kidney cancer, this large cohort helps us to better understand how frequently discovered alterations occur.

C-4-b Power analysis for the validation cohort: Here we calculate our statistical power for detecting both common and rare SNPs associated with racial disparity in MET and VHL genomic modules.

For the common SNP arm of the power analysis, we focus on 550 common SNPs prioritized by the Fisher exact test proposed in Aim 3. The Fisher exact test is adopted to detect SNPs associated with racial disparity in RCC, using equal numbers (192) of African American and Caucasian patients with RCC. To determine test power, we survey the parameter space of a candidate SNP, i.e. the frequency of a SNP in all patients (f) and in African American (fa) and Caucasian (fc) patients. According to multiple testing correction with the Bonferroni method, only SNPs with p-value < 1.0e-4 are considered to be associated with race disparity in RCC. Using the STATMOD R package [66] [cite STATMOD package], we find that for detection with a power of 0.8, a candidate SNP requires an f and fa/fc larger than 0.08 and 3.5 respectively. We note that the Bonferroni correction is overly stringent, rendering this power analysis conservative.

For the rare SNP arm of the power analysis, we pool adjacent rare SNPs together. Following testing on all pooled rare SNPs tests, if we assume prioritized regions are genes, we expect approximately 10 genes of 5kb length. Using the SKAT R package, we performed a power analysis of 100 simulated samples. Even at this low number of samples, we were able to detect regions with an Odds Ratio (OR) equal to 4 (power > 0.8).

We expect to be able to match patients in our validation cohort, given scale of the statewide population sampled. Pairing of subjects allows use of paired statistical testing. A paired test has much greater power than a pooled Fisher exact test. Therefore, our power analysis above is conservative, and should serve as a lower bound.