A. SIGNIFICANCE

Renal cell carcinoma (RCC) constitutes over 90% of kidney cancers and is the most lethal genitourinary malignancy [1]. Known RCC risk factors include male gender, age, hypertension, obesity, chronic kidney disease, and smoking [2, 3, 4]. However, these factors are likely contributory in less than half of RCC cases [5, 6], and measured associations between known risk factors and RCC are modest. Among all races, the incidence of RCC has tripled in recent years; however, the most dramatic increase has been observed in the African-American population (Fig. 1) [1, 7]. Genetic risk factors have been postulated to play an important role in RCC racial disparities [3, 5, 8, 9], but to date, no comprehensive study has specifically focused on deciphering genetic mechanisms associated with racial disparity.

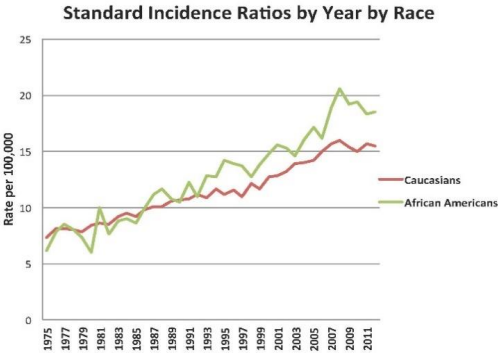

Figure 1: Standardized incidence ratios of cancer of the kidney and renal pelvis for Caucasians (Green) and African Americans (Red). Data from the Surveillance Epidemiology and End Result program from 1975-2011 [83]

There are two main subtypes of RCC (Fig. 2), clear cell RCC (ccRCC) and papillary RCC (pRCC). Clear cell, the most common histologic type, accounts for 75% of cases and is linked to alterations in the VHL (von Hippel–Lindau) gene. VHL encodes a subunit of a complex that is responsible for downregulating other proteins through ubiquitin ligation (Fig. 3). In particular, one target of this complex is HIF1a, which promotes angiogenesis. VHL is categorized as a tumor suppressor gene in part because loss of VHL function promotes angiogenesis which may increase solid tumor growth. A germline mutation within this gene have been linked to VHL syndrome, characterized by a number of cancer types, including clear cell RCC [10, 11, 12]. However, over 80% of sporadic cases have alterations in this gene.

There is a markedly higher incidence of pRCC in African-Americans relative to Caucasians. The MET protein (encoded by the c-Met gene) is the most prominent driver in pRCC, which represents roughly 16% of the RCC cases.

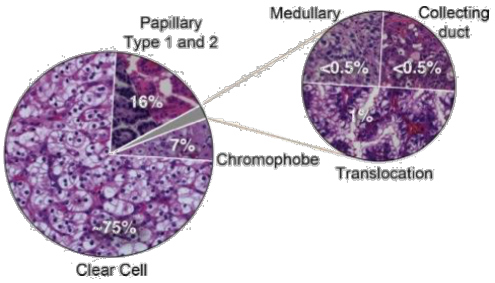

Figure 2: Two histologic types of RCCs; clear cell (ccRCC) and papillary type (pRCC) [81].

MET is a transmembrane receptor-linked tyrosine kinase (Fig. 3). 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 in a variety of cancers including liver, brain, kidney, stomach, and breast cancers. Consistent with its well-characterized roles in growth and development, it is normally only expressed in stem cells and progenitor cells. Similar to the clear cell, there are both hereditary and sporadic cases of MET mutations in papillary RCC important to oncogenesis [13, 14, 15]. Multiple mechanisms of MET alterations have been described, all believed to contribute to tumorigenesis.

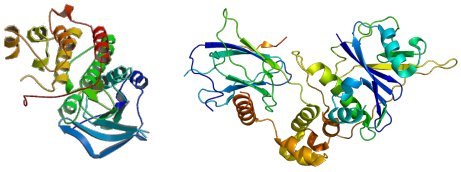

Figure 3: X-ray crystal structure of MET protein (left, pdb ID: 1R0P) and HIF-VHL complex (right, pdb ID: 1lm8)

Given that the main driver genes of RCC have been well characterized -- VHL in ccRCC and MET in pRCC – we hypothesize that racial disparity may be linked to genomic alterations in these coding and non-coding regions of these genes. We will analyze the frequency, functional impact, and genomic burdening of mutations that cause loss-of-function (LoF) and gain-of-function (GoF) mutations across samples. We will study the patterns of LoF and GoF variants in tumor-suppressor genes (TSGs) and oncogenes among Caucasians and African Americans to provide a new perspective on the underlying biology of RCC.

We will also investigate how somatic and germline variants complement each other to promote RCC. Per Knudson's 'two-hit' hypothesis [16], an aggregate effect of several mutations is what often leads to cancer. 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. By analyzing the patterns of somatic-germline co-occurrence in African-American and Caucasian patients in *VHL* and other genes, we hope to shed light on possible differences associated with the etiology of this disease.

B. INNOVATION

This study will be the first comprehensive assessment of somatic and germline alterations in kidney cancer by race. In contrast to other cancer types such as prostate cancer [17, 18], racial disparities in kidney cancer, particularly genomic aspects, have not been well studied. We are interested in identifying key genomic alterations that contribute to the greater incidence and distinct histological distribution of kidney cancer in African-Americans relative to Caucasians. In particular, they have higher incidence of pRCC (Fig. 4). We will expand upon prior whole-genome analysis in The Cancer Genome Atlas (TCGA) by including an additional large, independent validation cohort of African-American patients with ccRCC, matched by kidney cancer risk factors. By including these samples and performing secondary data analysis of the existing ccRCC and pRCC datasets, we can compare differences in risk variants, driver mutations, and driver copy number alterations by race.

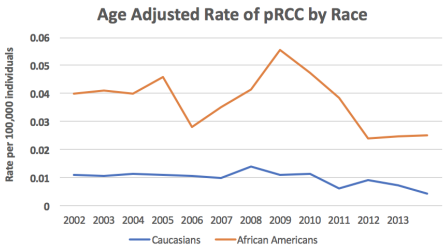

Figure 4: African Americans have much high age adjusted incident rate for pRCC compared with Caucasians.

In addition, we propose to integrate coding and non-coding alterations to characterize and prioritize key variants that drive racial disparities in kidney cancer. Whole genome analysis, including non-coding genome variation analysis in relation to canonical coding driver mutations, will expand our search beyond the small gene-centered landscape currently known to be strongly associated with kidney cancer (e.g., MET and VHL). Specifically, one novel aspect of our approach will entail focusing our analyses on the METome and VHLome, which will include alterations falling within coding and regulatory regions associated with these genes. We will prioritize coding and non-coding mutations falling within the METome and VHLome of our kidney cancer cohort. We will subsequently carry out comprehensive experimental assays to validate key variants from this prioritization scheme. This will be the first comprehensive analysis of the METome and VHLome, and it will also be a first step in addressing genetic aspects driving kidney cancer racial disparity. Given the limited number of African-American patients included in TCGA, the current TCGA cohort is underpowered for investigating racial disparities. As such, the work proposed here will greatly bridge these gaps in TCGA.

C. APPROACH

Aim 1: To perform WGS of African-Americans with ccRCC to complete a missing aspect of TCGA

C-1-a Rationale: In recent years, TCGA has broadened our understanding of the genomic basis of various forms of kidney cancer. For example, TCGA helped demonstrate that different cell types in the kidney may give rise to distinct forms of cancer and that somatic alterations (driver mutations and copy number variants) are important for determining a cancer's molecular profile. As part of TCGA, various high-volume tertiary centers submitted kidney cancer samples to the Bio-specimen Core Resource for accessioning and specimen processing. However, specimens were not submitted in a coordinated fashion to ensure a study population with a similar profile to that encountered nationally (stage, race, etc.).

Despite the fact that African-Americans account for approximately 1 in 7 cases of ccRCC, only a limited number of African-Americans with clear-cell kidney cancer were included in TCGA analysis [14/427 (3.3%) samples underwent whole-exome sequencing and 1/40 (2.5%) underwent WGS; Table 1]. The failure to include a large population of African-Americans with ccRCC limits our ability to explore the genetic basis of racial disparities. With limited available data, a preliminary analysis of somatic driver alterations or germline risk variants in kidney cancer among African-Americans with ccRCC showed that African Americans have fewer *VHL* alterations than Caucasians [19].

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

We propose a complete, whole-genome analysis of the top two subtypes of kidney cancer -- papillary and clear cell -- by analyzing a cohort of African-Americans with ccRCC in addition to TCGA cohorts. By including this additional cohort, we will have an adequate number

of cases to allow balanced comparisons of clear cell and papillary kidney cancers between African-American and Caucasian patients. Furthermore, by using a patient cohort of a different genetic background, WGS might reveal novel, ethnicity-specific driver events, was recently observed in an African-American prostate cancer study [20].

WGS offers several advantages over traditional chip-based methods. It allows analysis of poorly tagged or rare single nucleotide polymorphisms (SNPs), insertions and deletions (INDELs), and structural variants (SVs). Moreover, WGS offers single-nucleotide resolution, helping to isolate disease-causing variants rather than large DNA blocks in linkage disequilibrium.

C-1-b Sample acquisition, comorbidity/demographic matching, and DNA extraction: All patients undergoing scheduled kidney cancer surgery at Yale-New Haven Hospital are offered enrollment into an Institutional Review Board (IRB)-approved Genitourinary Biospecimen repository (P.I. Shuch, HIC# 0805003787). Within 30 minutes of removal, a team of pathologists snap freeze fresh tumor tissue in liquid nitrogen. The pathologists procure whole blood as well to serve as a genomic control. In the past 4 years, over 500 patients with kidney cancer have been prospectively enrolled, including a large proportion of African Americans from New Haven's diverse population. All fresh bio-specimens are stored at -80˚C and are available for immediate analysis. For the purpose of completing TCGA dataset, we will study 15 African-American subjects with ccRCC from 2013-2016. TCGA kidney cancer projects have captured patient age, sex, race, and smoking history, and have limited information from a secondary analysis on obesity status. Self-reported racial identity may be imprecise, yet it 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 that their genetic ancestry matches their reported ancestry and matches our target population. 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: We will perform sequencing of normal and tumor samples using Illumina's HiSeq 2000 technology. In brief, we will hybridize DNA fragments from each sample using HiSeq Paired-End Cluster kits and further amplify them using the Illumina cBOT. To generate paired-end libraries, we will utilize HiSeq (2x101) cycle sequencing and perform imaging using TruSeq kits.

We have extensive experience in large-scale variant calling and interpretation through our active participation in the 1000 Genomes Consortium. In particular, we were involved in the analysis working group and the SV and functional interpretation subgroups, in which the majority of the variant calling tools were developed, deployed, and interpreted [21, 22, 23].

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. We will further process these BAM files to sort and mark duplicate reads before calling variants. We have already set up a prototype pipeline for calling germline and somatic variants. We will follow the GATK best practices [24] to generate initial raw variant call sets using the GATK haplotype caller. We will use parameters consistent with those used in TCGA [25]. We will filter these initial call sets by running the 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 [26] and Strelka [27] will call somatic single nucleotide variants (SNVs) and INDELs, respectively.

SVs are important contributors to human polymorphisms, have great functional impact, and are implicated in a number of diseases such as cancer. We have developed a number of SV-calling algorithms, including BreakSeq [28], CNVnator [29], AGE [30], and PEMer [31]. Furthermore, we have studied the SVs that originate from different mechanisms and may have potentially divergent functional impacts [32, 33]. We will run CNVnator to identify copy number variations (CNVs) in each cancer sample. We will apply CREST [34] to identify large structural variations. Finally, we will run our BreakSeq tool to decipher the underlying mechanism. Along with our new sequenced samples, we will reprocess all TCGA data using our own calling pipeline, thereby mitigating any potential processing or batch effects.

C-1-d Deliverables: In this aim, we will generate an extensive catalog of variants in kidney cancer for both African-American ccRCC patients at Yale and TCGA patients. We will achieve this using methodology

consistent with that used by TCGA. This catalog will encompass both germline and somatic variants, including SNPs, INDELs, and large SVs. We will cover both coding and non-coding regions of the genome. Our catalog of variants will serve as a basis for identifying racially disparate genomic variants in kidney cancer. We plan to make our sequencing data available via the Database of Genotypes and Phenotypes (dbGAP; see data dissemination plan).

C-1-e Problems and solutions: We do not anticipate major difficulties acquiring samples or performing WGS. However, we note that it may not be possible to precisely match all variables between TCGA and Yale samples. As such, we may need to expand comparisons or work closely with various outside collaborators to ensure matches that are as close as possible.

Aim 2: To identify key genomic variants associated with kidney cancer that exhibit racial disparities

C-2-a Rationale: We aim to develop a fully curated catalog of somatic and germline variants associated with canonical driver genes in kidney cancer (*MET* and *VHL*). First, we will identify all genomic regions associated with these genes. We term these networks of association the METome and VHLome. This genome-wide survey will include *all regions* (introns, promoters, exons, etc.) that we find directly or indirectly impact the function of *MET* and *VHL* in RCC. Then we will use this extensive catalog to prioritize coding and non-coding variants associated with kidney cancer. For this prioritization step, we will leverage an extensive suite of software tools that we have applied in prior studies. In addition to prioritizing germline and somatic variants in *MET* and *VHL*, we will study the interplay between somatic and germline mutations (i.e., their combined effects). After prioritizing for association with cancer, we will evaluate the racial disparity of the variants. In addition to analyzing rare variants directly, we will also analyze the association with cancer and racially disparate variant-burdening of smaller genomic subregions.

C-2-b Relevant preliminary results: Here, we outline some of our published and publicly available tools and methods devised to prioritize variants in large-scale sequencing studies. These pipelines can be readily combined to provide multiple lines of evidence for prioritizing variants, and they have already been successfully applied to a number of disease variant datasets. The corresponding software code for each tool is computationally efficient, thereby enabling us to scale them to large patient cohorts.

C-2-b-1 Tools for somatic and germline burden tests: We have developed a number of software tools to annotate and understand the effects of variants within the coding regions of the human genome. We developed **VAT** to annotate coding variants; for example, VAT can determine them to be synonymous, non-synonymous, premature stop codons, or splice-site changes [35]. Once mapping the annotated variants to three-dimensional (3D) structures from the Protein Data Bank, we can study the effects of them in detail by measuring events associated with their LoF or in the contexts of allosteric regulation and local mechanistic perturbations.

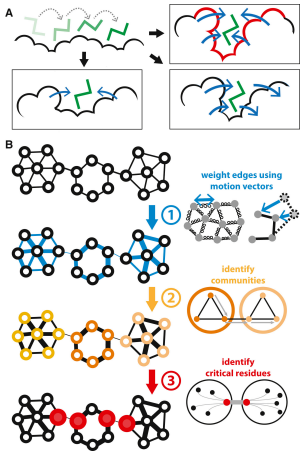

Figure 6: Prioritizing the effects of SNVs based on predicted allosteric residues at the surface (A) and within the interior (B).

In addition, we have developed **ALoFT**, a tool specifically tailored to annotate and predict the disease-causing potential of LoF events [36]. Short for "annotation of loss-of-function transcripts", we have used ALoFT 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 6,500 cancer exomes and demonstrated that variants predicted to be deleterious by ALoFT are enriched in canonical cancer driver genes [36].

With respect to allosteric effects, we have developed the **STRESS** software tool [37]. STRESS (STRucturally-identified ESSential residues) employs models of large-scale protein conformational changes in order to predict key allosteric residues from 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

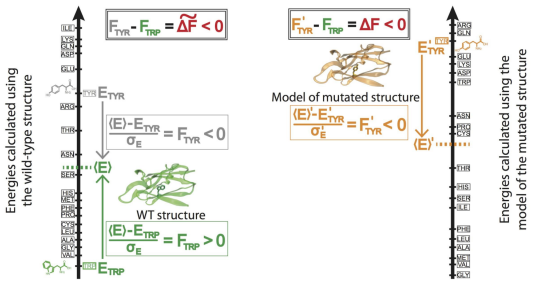

Figure 5: Prioritizing the effect of SNVs based on changes in localized perturbations (as measured by frustration).

residues that are highly conserved over both long and short evolutionary timescales [37]. This software has also been used to help rationalize otherwise poorly understood ("cryptic") disease-associated SNVs.

With respect to localized perturbations, we performed a separate study [38] to demonstrate how localized changes in biomolecular frustration can be used to better understand the differential effects of variants in oncogenes and TSGs (Fig. 6). Specifically, these results shed light on potential GoF variants on the surfaces of oncogenes, and LoF variants within the cores of TSGs.

In addition to coding variants, we have developed a tool to prioritize non-coding variants in cancer called **FunSeq** (Fig. 7). In brief, FunSeq prioritizes variants based on network connectivity and their disruptiveness (e.g., by finding motif breakers), by identifying deleterious variants in many non-coding functional elements (including transcription factor binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitivity sites). In our published work using FunSeq [32], we integrated large-scale data from various resources, including ENCODE and the 1000 Genomes Project, with cancer genomics data. By comparing patterns of inherited polymorphisms from 1,092 humans with somatic variants, FunSeq identified candidate non-coding driver mutations.

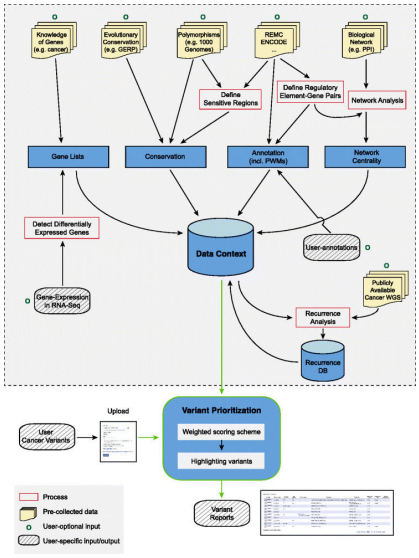

Figure 7: The workflow of FunSeq

We have developed statistical methods for the analysis of non-coding regulatory regions. **LARVA** (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichment in non-coding elements by comparing observed mutation counts with expected counts under a whole-genome background mutation model [39]. 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 non-coding cancer drivers, including the TERT and TP53 promoters [39]. Finally, we developed **MOAT** (Mutations Overburdening Annotations Tool), an alternative empirical mutation burden approach that evaluates mutation enrichment based upon permutations of the input data (in press). This tool supports both annotation-based and variant-based permutation.

C-2-b-2 Analyzing whole-genome datasets from cancer cohorts including kidney cancer: We have played key roles in TCGA investigations into prostate [40] and kidney [25] cancers. We participated in TCGA KICH (chromophobe RCC) project [41] and a following pan-subtype kidney analysis [42]. Our team analyzed the WGS data for the TCGA KIRP (pRCC), now published in The New England Journal of Medicine [25]. In recent work, we leveraged our expertise in non-coding regions in the first whole-genome analysis of pRCC samples [43]. Our work found significant genomic alterations beyond traditional known drivers of pRCC located within coding exons (Fig. 8). 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 influence prognosis. Moreover, we constructed evolutionary trees using the abundant mutation information from WGS. The tree structure implies tumor evolution path and correlates with tumor subtypes (Fig. 9).This experience provided further practical knowledge of working with available RCC genomic datasets. Finally, our team has participated in two ongoing pan-RCC manuscripts by playing a central role in assessing the cluster-of-cluster assignments immunologic profile from gene and microRNA expression datasets.

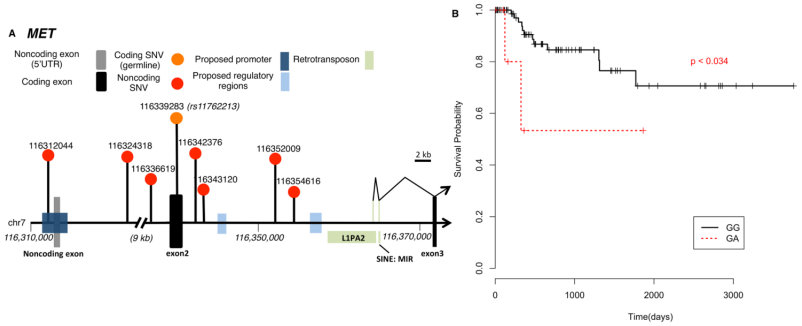

Figure 8: A. Whole genome analysis of 35 pRCC samples finds significant non-coding mutations in MET. B. A germline SNP (rs11762213) predicts survival in type 2 pRCC patients.

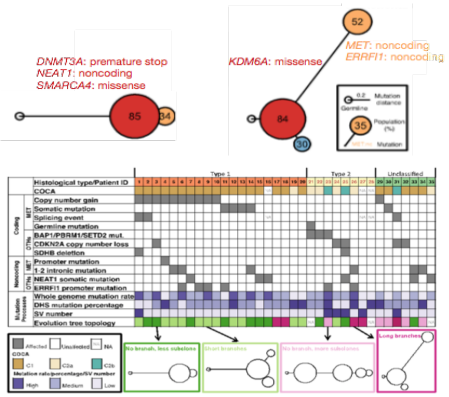

Figure 9: Evolutionary trees help elucidate pRCC tumor development and complete molecular subtyping.

Together with other published results on RCC [44, 45, 46, 47, 48], we have assembled an extensive list of impactful and statistically significant regions of RCC genomes. Similarly, as part of the driver discovery subgroup in PCAWG, we participated in a comprehensive variant prioritization exercise to generate a catalog of driver elements in many cancer cohorts. Furthermore, we are currently leading the PCAWG group to investigate the impact of non-coding mutations on cancer development, progression, and prognosis. As part of this effort, we ran our FunSeq pipeline on each variant (~30 million total somatic mutations among 39 cancer subtypes) in PCAWG. In addition to identifying canonical driver mutations, we identified many high-impact mutations that can potentially influence cancer progression.

C-2-c Research plan:

C-2-c-1 Construct the METome and VHLome by integrating *MET*-And *VHL*-associated elements across annotations: We will link high-impact regions associated with *MET* and *VHL* 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 will build a METome and VHLome that includes regions that are associated with the function of these two genes in RCC. Mapping these relationships among coding and non-coding elements is important because apparently incidental or unimportant cancer mutations may significantly affect cancer biology through indirect network mechanisms [32, 49].

We will link transcription factors to enhancer elements, and enhancers to their target genes. We will seek to clarify the influence of distal epigenetic regulatory markers, like methylation and chromatin-state, on *MET* and *VHL* expression. We will use protein interaction networks to better understand the broader consequences of variation as transmitted through a molecular interaction network. We will build maps of the molecular pathways influencing MET and VHL function.

In order to systematically integrate evidence from various sources, which can often be represented in graph form, we will use a random walk on multiple graph layers. At each step, we will update the state on a graph. The walk will stop at a certain distance from its starting point (boundary condition). Starting with a gene of interest, and simulating this random walk multiple times, we will finally tally the number of visits to each node and pick out “hot” nodes that are frequently covered in walks. Those nodes represent the nodes linked with our starting gene. Since random walk will give an empirical distribution of the number of visits to the nodes, we will be able to set up our cut-off value for linked nodes in a rigorous manner.

This integration will produce extended *MET* and *VHL* annotations. 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 non-coding regions.

C-2-c-2 Build a comprehensive mutation catalog for the METome and VHLome: We aim to build an inclusive, comprehensive mutation catalog with all identified variants assembled from both our newly sequenced dataset and publicly available data. We will first search the literature for previously documented RCC-associated genomic alterations. We will collect genetic changes that include SNVs, indels, SVs/CNVs, and mutation process signatures [50, 51]. Prior work has shown that papillary and clear cell RCC subtypes are uniquely characterized by CNVs as an early and major driver event [44]. Because repeats are triggering factors for many SV events, we will pay particular attention to repeat polymorphisms around known cancer-associated genes and recurrent CNV regions in RCC. Finally, we will gather both germline and somatic mutations from TCGA and PCAWG. To estimate background mutation rates in the general population, we will leverage both gnomAD and ExAC [51].

Currently, ExAC and gnomAD report 677 variants in ~31,000 WGS alleles in *MET* and an additional 1,218 variants in 250,000 exome-sequenced alleles. In *VHL*, these numbers are 448 and 225, respectively. In 35 pRCC whole genomes, we found seven somatic *MET* mutations. In 161 TCGA whole-exome pRCC samples, we discovered 15 non-synonymous somatic mutations in *MET*. *VHL* is mutated in 234/418 TCGA WGS ccRCC samples. In PCAWG, we identified 35 *MET* and 46 *VHL* mutations in 144 WGS samples.

By linking functional elements, the number of regions grows exponentially with the degree of association, which is the linkage distance between the target region and the core gene *MET* or *VHL*. We expect the number of mutations to grow by roughly one order of magnitude, assuming a branch factor of 3-to-5 and including all associations among secondary associations. Therefore, we estimate that there are ~20,000 germline SNVs and ~1,500 somatic SNVs from the public dataset and our newly sequenced samples.

C-2-c-3 Run our variant prioritization pipeline on all variants in both coding and non-coding regions: Once we establish our comprehensive set of variants for both the METome and VHLome, we will prioritize them both by their inferred impact and by using recurrence-based approaches. Together, the tools we developed for these methods (detailed in section C-2-b-2 above) constitute a comprehensive pipeline that we designed to readily process and evaluate large datasets of coding and non-coding variants.

Intense research efforts to gain mechanistic insights into the functioning of MET and VHL have resulted in detailed 3D structural models. Using high-resolution crystallographic models of these two proteins, we will run our tools to annotate coding variants associated with these genes within our cohort, followed by running the remainder of our coding variant prioritization pipeline.

In addition to coding variants, many changes in non-coding regions regulating the METome and VHLome may play critical roles in RCC initiation and progression. In order to comprehensively characterize key non-coding alterations influencing these genomic subsystems, we will run our updated and extended FunSeq pipeline on the METome and VHLome variant catalog. As part of our initial analysis, we ran FunSeq and carefully curated the results. We found several disruptive non-coding mutation hot spots within the genome. With the addition of many more samples, we will perform comprehensive prioritization to identify additional non-coding variants that may play key roles in kidney cancer.

As mentioned above, in addition to evaluating functional impact we will also evaluate variant recurrence to identify key mutations associated with the METome and VHLome in kidney cancer. We will apply our LARVA and MOAT tools to the comprehensive kidney cancer variant catalog. Our prior analysis of TCGA WGS samples indicated 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 variant catalog.

We will run our tools to identify critical regions burdened by germline variants. The statistics for germline variants are distinct from those of somatic variants and thus demand distinct analytical approaches. We will run our pipeline to identify *MET*-associated regions that are significantly burdened by germline mutations in kidney cancer relative to healthy controls. We will mask known SNPs and flanking regions associated with high body mass index [52], hypertension [53], smoking [54], and other known risk factors in previous association studies, thereby reducing the possibility of misattribution of these known RCC comorbidities to direct genetic effects.

C-2-c-4 Study germline-somatic interplay: Following prioritization of *MET*- and *VHL*-related variants across all Yale-TCGA samples, we will study differences in the frequency and functional impact of variants between variant sets. In addition, we will analyze patterns of co-occurrence between somatic and germline variations and identify differentially burdened regions and variants 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 individuals to renal cancer. These analyses provide an opportunity to identify genetic signatures and impactful and recurrent mutations that partially explain racial disparities in RCC. In addition, using a sequence kernel association test (SKAT) [55] we will find regions that are significantly burdened by germline mutations in kidney cancer cases relative to healthy controls. To perform these analyses, we will leverage a number of popular tools including normal reference genomes from racially diverse cohorts. Along with genomic samples in the combined Yale-TCGA cohort, we will mine several other genomic repositories, including the 1000 Genomes whole-genome samples [21, 22, 32, 56, 57], the ExAC meta-cohort (with TCGA samples excluded), and gnomAD [51].

C-2-c-5 Determine differential burdening between Caucasians and African-Americans within METome and VHLome germline genomic regions

C-2-c-5-i Overall approach on disparities: We will re-prioritize the variants and regions from C-2-c-3 to find racially disparate genetic elements based on allele frequency distributions using a Fisher's exact test, fixation index differences (F_{ST}), and the unified sequence-based tests in He et al. [58].

C-2-c-5-ii Specific Variant level analysis: To analyze coding regions, we will employ the full 467 samples with whole-exome data from TCGA. To analyze common variants at a single locus, we will use Fisher's exact test to evaluate the racial disparity between Caucasian and African-American subjects with RCC. Here, we will

prioritize common variants according to their associations with RCC disparity in race and their association with elements in the VHLome and METome networks. For a common SNP identified in African-Americans and Caucasians with RCC, we will record minor allele frequencies and major allele frequencies in African-Americans and Caucasians with RCC. For these counts of a focal SNP, Fisher's exact test is used to determine whether the SNP tends to be associated with African-Americans with RCC. The statistical significance will allow prioritization for further study and validation.

The power of the Fisher's exact test can readily be estimated in this context. For instance, for an ordinary SNP with an allele frequency of 7% among all 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. In addition, our focus on mutations falling within the VHLome and METome will mitigate the multiple hypothesis testing burden, thereby increasing the power to detect variants in these networks.

C-2-c-5-iii Specific region-based analysis: Beyond investigating the associations between single common variants and race, we will evaluate the cumulative effects of a set of rare variants in genomic regions (such as VHL/METome genes, promoters, and enhancers, as well as each network as a whole) by using both burden and non-burden tests.

Burden tests are often applied on regions where most of the variants in the same neighborhood are causal and affect phenotype in the same direction (e.g., LoFs disabling a TSG). We assume that, in total, we have available whole-exome sequencing data for *n* patients. For a target region, for example, consider a gene that harbors *m* variants. Let y_i denote the a race-based indicator variable for the i^{th} patient. $y_i = 1$ for African-Americans and 0 otherwise. Let $\mathbf{G_i} = (g_{i1},...,g_{im})'$ represent the genotype of patient i. We can use a logistic regression model can be used to evaluate associations (equation 1). Suppose that π_i describes the mean of the population status. Then

$$\text{logit}(\pi_i) = \gamma_0 + \mathbf{G_i'b} \quad (1)$$

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

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

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:

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

We can use the allele frequency 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, we can use a sequence kernel association test(SKAT).Instead of assuming a weighted coefficient effect in the burden test, we can treat each b_j as an independent random variable with 0 mean and variance $w_j^2 \tau$. We can then change the null hypothesis to H0: $\tau = 0$, and write the test statistic in equation (1) as:

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

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

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

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

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

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

$$Q_{opt} = \min\left(Q_{\rho 1}, \cdots, Q_{\rho_k} \right) \quad (7)$$

C-2-c-5-iv 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 the 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; we will use it when we cannot assume that the populations follow a normal distribution. As population size increases, we can calculate a Z-score.

C-2-c-5-v Power analysis using SKAT for per region based analysis: As described above, we plan to use aggregated burden tests to look for differential burdening between populations and to use this to rank genomic regions. While we are not striving for absolute statistical significance in differential burdening, our sample size provides an appreciable signal for ranking. Here, we discuss the power aspects of burden tests applied to our sample populations. To estimate the sample size needed to obtain statistical power, we ran the SKAT package (available from the R project) on several population models for genomic regions of 5000 nucleotides (Fig. 10). In our proposed study, we will focus on genomic modules linked to kidney cancer (i.e., the METome and VHLome); therefore, we expect a large number of effective mutations. Typically, the *MET* genomic region consists of 126,027 nucleotides and the *VHL* region consists of 12,035. We expect these numbers to increase significantly (likely up to ten-fold) after creating the genomic modules of the METome and VHLome.

C-2-d Deliverables: We aim to generate a list of genomic regions that have the greatest potential to impact RCC development and progression. In particular, we will construct a list of genetic modules that are assembled from high-impact regions, with an emphasis on those genomic annotations that are associated with the canonical drivers of pRCC and ccRCC, *MET* and *VHL* (i.e., our assembled METome and VHLome). Furthermore, after evaluating METome and VHLome variants with our suite of software tools (detailed in section C-2-b-1), we will rank and prioritize the variants on the basis of both their inferred functional impact and the extent to which they exhibit racial disparities. We will format our annotations and rankings for genes and non-coding regions into structured online data that are designed for easy accessibility. We will make this data publicly available on our project web server, in tables in published papers, and/or through dbGaP (see data dissemination plan). We will also make the software tools available in public repositories (e.g., GitHub) to ensure reproducibility.

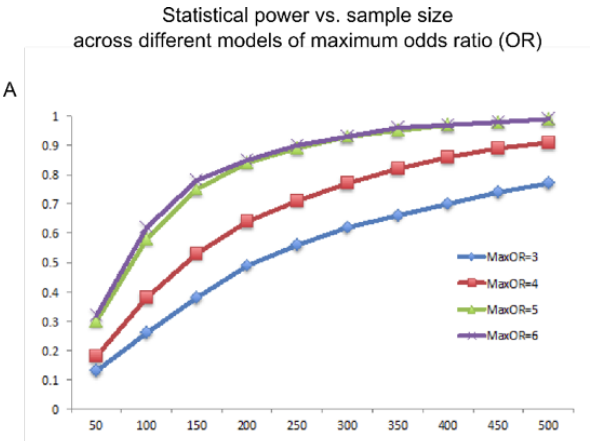

Figure 10: Using the default haplotype information in the SKAT haplotypes dataset, we randomly selected subregions of size = 5k and ran 100 simutations. In A, we show the statistical power obtained across the different models of maximum Odds Ratio. In B, we show the required sample size for each of these models in order to obtain significant statistical power (alpha = 0.01, beta = 0.2)

C-2-e Problems and solutions: We anticipate no major difficulties identifying a set of variants from the genome sequencing data in the first aim and from the large amount of data in the databases. We have already completed a study [43] to search for significant alterations in kidney cancer. We note that identifying impactful variants exhibiting clear racial disparities is not a guarantee. However, with additional cancer-associated variants, we anticipate that we may find a number of significant variants among the thousands that will be surveyed. In particular, by analyzing more than 20,000

SNVs (germline and somatic), we would expect to identify more than 55 regions with racial disparities to further process in Aim 3.

Aim 3: To validate specific variants and mutated regions and study clinical and environmental covariates suspected of contributing to kidney cancer racial disparities

C-3-a Rationale: Aims 1 and 2 together represent a discovery phase, where we will identify and prioritize genomic regions for racial disparities associated with cancer. They are low-risk in the sense that we are confident that if we do the sequencing we will produce a list of variants to test. In Aim 3, we will validate our findings from our discovery phase in an independent patient cohort. This independent validation cohort will include patients with RCC from Yale's Genitourinary Biospecimen Repository, as well as patients with RCC from a statewide sampling through the Connecticut Tumor Registry (P.I. Shuch, HIC# 0805003787). This cohort will include both African-Americans and Caucasians with ccRCC and pRCC to allow comparisons across histological type and race. In addition, we will match for environmental and clinical factors that may be associated with racial disparity in RCC by examining the electronic health records (EHRs) associated with all samples.

C-3-b Genotyping for specimen acquisition, and DNA extraction: For our validation cohort, we will select an equal number (n=96) of Caucasian and African-American clear cell and papillary tumors (total n=384) will be selected as a Validation Cohort. We intend to validate the top 55 highly prioritized regions (100bp each) for 384 individuals from Aim 2. As mentioned above, specific kidney cancer risk factors may influence the risk of RCC. To control for these differences, we will match cases using a similar optimal match algorithm to that described above. This matching algorithm will include RCC comorbidities such as age, sex, smoking status, and obesity. Once we select cases, we will access archival fresh or formalin-fixed, paraffin-embedded (FFPE) tissue blocks to retrieve tumor and the adjacent normal kidney tissue for a genomic control. Our IRB-approved Genitourinary Bio-specimen protocol provides access to tissue from 1988-2016. If further cases are needed, we have access to specimens and clinical data over the past two decades in Connecticut State, through a Connecticut State Tumor Registry IRB approved protocol (n = ~11,000 index cases). Our genitourinary pathologists (Dr. Adeniran) will have already centrally reviewed all of the tumors and classified them according to recent International Society of Urologic Pathology criteria [59]. For both fresh and FFPE tissue, we will extract DNA from the tumor and adjacent normal kidney using an automated Maxwell 16® system (Promega, Madison, WI).

C-3-c Genotyping for sample matching: A recent 1000 Genomes Phase 3 study showed that African Americans (African Ancestry in SW USA) carry a significant amount of European ancestry [56], revealing an admixed population structure. Methods like admixture mapping are used to assign a degree of correlation between the genetic background of admixture populations, according to ancestry composition and differences in phenotypes associated with genetic background [60, 61]. We will genotype markers of ancestry (~30-50 markers are needed) from samples and controls to construct and evaluate genetic clusters in two ways: i) by using maximum likelihood phylogenetic algorithms to infer clusters of ethnic individuals [62] and ii) by performing principal component analysis [63]. We will include reference genotypes from HAPMAP and 1000 Genomes in this analysis, allowing us to assess the genetic topology of different genotypes and confirm the genetic background of the African-American and Caucasian populations. Equally important, this will enable us to correlate disease phenotypes with the genetic background in the case of an admixture population [64, 65].

C-3-d Including environmental and clinical covariates: Racial disparity in cancer is likely the result of multiple factors. While genetics can provide a valuable insight into kidney cancer etiology, we will also take into account other perspectives. In this context, we plan to (i) find significant correlations between clinical and environmental conditions and the disease incidence across races and (ii) rigorously correct for non-genetic biases and stratification errors in collected samples. We will build an automated pipeline to consider environmental and clinical factors along with genetic ones. Both our pipeline and results will be available on our project website. We will use the pipeline to provide an optimal match between the individuals in our validation cohort.

C-3-e Power analysis for the validation cohort:

We have calculated our statistical power for detecting both common and rare SNPs associated with racial disparity in the *MET* and *VHL* genomic modules. For the common SNP arm of the power analysis, we will focus on 550 common SNPs prioritized by the Fisher's exact test proposed in Aim 2. We will use Fisher's exact test 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 surveyed 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 a p-value < 1.0e-4 are considered to be associated with race disparity in RCC. Using the STATMOD R package [66], we found 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 ten regions 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 equal to four (power > 0.8).

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

C-3-f Problems and solutions: By the time we get to this aim, we will have clear hypotheses to test. We cannot anticipate how the validation will work out. However, we are optimistic that with about 55 candidates, we will find at least one or two to study further. We have designed a study that we believe will have adequate power to detect racial differences. If after performing the first third of the validations (~185) we do not find any regions that are significantly different, we have a number of courses of action: i) We can remove somatic variants from the validation. Validating somatic variants is more expensive than germline ones. Removing them will allow us to validate a larger number of regions. ii) We can focus only on disparities in coding genes as opposed to non-coding regions. There are many more kidney cancer exome sequences than WGS (by more than an order of magnitude). Coupling a larger sample population with a much smaller genomic space being queried should substantially increase the power of our analysis. iii) We can expand the validation cohort to increase power. Currently, the Yale Biospecimen Repository is adding 150 new kidney cancer subjects each year. Additionally, our close collaborator in the US Kidney Cancer Study has access to a large cohort of genomic DNA in individuals with kidney cancer (843 Caucasians and 358 African Americans). Finally, the Yale Kidney Cancer Program recently was granted approval from the Connecticut State Tumor Registry to access records and/or tissue from individuals diagnosed with kidney cancer from 1998 to present. Moreover, in our recent publication, we found that a germline SNP, rs11762213, predicts type 2 pRCC patients prognosis and might play a role in pRCC incidence in African Americans [43]; rs11762213 has also been associated with prognosis in ccRCC [13, 67]. As the underlying mechanism remains to be elucidated, we can perform functional studies (as outlined in Aim 4) on rs11762213, along with our other prioritized variants.

Aim 4: To perform functional characterization of a prioritized, high-confidence list of genetic variants

C-4-a Rationale: We will interrogate the functional impact of our top racially disparate candidate gene alterations on key kidney cancer signaling pathways associated with *MET* and *VHL*. Our objective is to determine the mechanistic basis for cancer progression.

C-4-b Research Plan:

C-4-b-1 Creation/validation of matched cell lines with genomic variants of interest

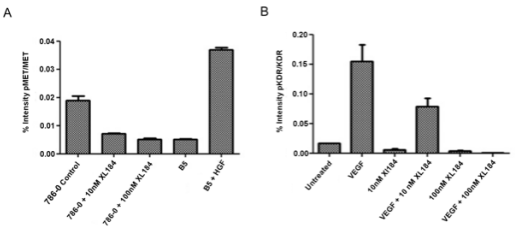

Figure 11: Using MSD for pathway analysis A. for pMET/MET with Cabozantinib B. for VEGFR2 with VEGF simulation using Cabozantinib.

CRISPR/Cas generation of matched human cell lines of candidate gene alterations. To interrogate candidate somatic and germline variants, we will employ the CRISPR/Cas system to introduce matched cell lines with and without variants of interest. HEK293 and YUNK1 (Yale Urology Normal Kidney, immortalized with SV40) will serve as useful controls**.** The urologic oncology laboratory recently collaborated on projects (Letter of support Dr. Ranjit Bindra) where CRISPR/Cas-based gene targeting protocols were optimized specifically to interrogate matched immortalized primary human cell

lines. This highlight our ability to i) design guide RNAs (gRNAs) targeting our candidate regions of *VHL* or *MET*, ii) to transfect plasmids encoding gRNAs with Cas9 specifically in primary human cells, iii) to induce synonymous or non-synonymous SNPs, non-coding mutations, or splice site mutations, and iv) to isolate single-cell clones via fluorescence-activated cell sorting for expansion. For both *VHL* and *MET,* sequencing will confirm successful cell line generation. We will isolate DNA from clonally expanded populations, and then PCR amplify and gel purify a 200bp segment containing the target sites for sequencing. To validate the success of *VHL* models, we will perform Western blots to detect the truncated VHL protein products using antibodies specific to the N-terminus. To assess off-target CRISPR alterations, we will interrogate the highest scoring off-target coding sites with PCR amplification and sequencing.

C-4-b-2 Interrogation of downstream HGF/MET and VHL/HIF/VEGF pathway activation. *C-4-b-2-i MET and VEGFR2 protein/phospho-protein quantification:* Meso scale discovery (MSD; Meso Scale Discovery, Rockville, MD) is novel technology to measure small biomedical markers such as cytokines and intracellular signaling proteins. MSD allows multiplex detection of markers using electrochemiluminescence to provide an extremely sensitive assay allowing measurement across several log-fold differences including validated assays of the pathways of interest. We will prepare whole-cell lysates of our matched cell lines using Triton X-100. We will use the following kits for total and phosphoprotein quantification: the Human KDR Base Kit, Phospho-VEGFR-2 Whole Cell Lysate Kit, and Phospho/Total Met Whole Cell Lysate Kit (Meso Scale Discovery, Rockville, MD). We will calculate phosphoprotein levels by the signal intensity from total and phospho-MET levels and total and phospho-KDR (Fig. 11).

C-4-b-2-ii Upregulation of MET and HIF gene expression: To interrogate changes at the mRNA level, we will extract RNA from the matched cell lines using the Maxwell-16 DNA Purification Kit (Promega, Madison, WI). We will perform digital gene expression profiling on 200ng of extracted RNA using a custom-designed hybridization probe set to evaluate a custom array of genes involved in VHL/HIF/VEGF and HGF/MET signaling pathways. We will perform gene expression analysis using the Nanostring nCounter system for digital RT-PCR and then analyze the data with nSolverAnalysis (NanoString Technologies, Seattle, WA) software. We have used this system for classification of ccRCC tumor regions into profiles “A” and “B” based on their angiogenic signature largely due to expression differences in HIF2/EPAS1 [68, 69].

C-4-b-2-iii Evaluation of hypoxia-reporter assay: To interrogate activation of the hypoxia-inducible pathway, various *in vitro* assays have been developed. In matched cell lines, we will utilize a lentiviral vector for detection of HIF upregulation (Qiagen, Hilden, Germany). This plasmid contains a firefly luciferase reporter assay under the control of a CMV promoter and several hypoxia transcriptional response elements (HRE). Quantification of luminescence will be performed using a luminescence microplate reader and flow cytometry.

C-4-b-3 Assessment of Functional Impact of Variants on *in vitro* Characteristics: Proliferation, apoptosis, Invasion, and Agar Growth: We will assess: (1) cell proliferation, using a tetrazolium colorimetric assay where we will quanitfy the cleavage of the tetrazolium salt WST-1 to formazan using a plate reader at 420-480 nm; (2) apoptosis, using the TUNEL assay; (3) migration, by scratching a confluent monolayer of cells and logging the distance the cells migrate from 3 to 24 h; (4) invasion, using a transwell assay where we will fix, stain, and visualize the cells that invade a Matrigel chamber; and (5) anchorage-independent growth, by evaluating colony formation in soft agar over three weeks prior to staining, visualizing, and quantifying colonies. We have successfully performed these assays in 96-well plates with clear cell kidney cancer lines to interrogate the role of XL184 (VEGFR2/MET inhibition) with and without HGF stimulation using absorbance to quantify visible colonies (Fig. 12).

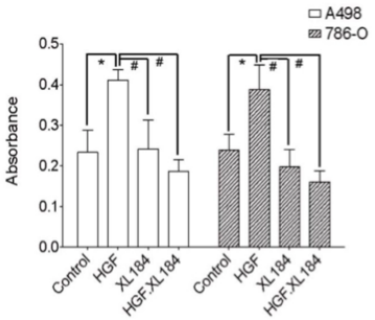

Figure 12: Using MSD for pathway analysis A. for pMET/MET with Cabozantinib B. for VEGFR2 with VEGF simulation using Cabozantinib.

C-4-c Expected Outcome and Alternative Approaches: These experiments are important to functionally validate the activity of candidate racially disparate alterations in our primary kidney cell lines. Although unlikely, we may encounter difficulties such as finding a small number of high-impact candidate gene alterations. However, we already have identified several novel variants in *MET* that have not been characterized, including the *MET* SNP rs11762213; we will have the capability to interrogate this variant regardless of Aims 1-3. As *VHL* is a TSG, losing one copy may be insufficient to alter gene expression and global hypoxia; therefore, it may be necessary to perform CRISPR/Cas on the other *VHL* allele to delete this copy. If this poses a challenge since the *VHL* gene is a small gene (3 exons, <700 nucleotides), we may be able to transfect plasmids with candidate variants in this gene.