***[CANDISP3 | Specific Aims]***

**Word counts:**

**specific aim : 744**

**significance: 545**

**Aim1 : 1132**

**Aim2 : 3530**

**Aim3: 1117**

**Aim4: 1006**

**\*\*\*\*\*Updated break down\*\*\*\*\*\***

**specific aim : 856**

**significance: 508**

**innovation : 212**

**Aim1 : 1065**

**Aim2 : 3375**

**C2a -> 137**

**C2b -> 939**

**C2c -> 1252**

**C2d -> 936**

**C2e-> 111**

**Aim3: 1117**

**Aim4: 905**

**Rest : 373**

**Total : 8411**

**- 856 = 7555**

**SPECIFIC AIMS**

Each year, an estimated 14,000 people in the United States die as a consequence of renal cell carcinoma (RCC). RCC is the urologic malignancy with greatest associated mortality. The racial disparities in this disease have been well documented: African-Americans with RCC have significantly poorer clinical outcomes. Various lines of evidence point to genomic differences as major contributors to these disparities. First, the incidence of RCC is 30% greater in African-Americans than in Caucasians. Second, recent studies have shown that African-Americans present with RCC at significantly younger ages, which may suggest a hereditary predisposition. Third, racial differences in RCC histologic subtypes are significant. Clear cell RCC (ccRCC) is the most common histologic type in all races, yet papillary RCC (pRCC) is three-fold more common in African-Americans, accounting for 35-40% of all cases. In this proposal, we aim to systematically investigate whether genomic differences can explain these large racial disparities while controlling for confounding factors. In particular, we will focus on the genes *VHL* and *MET* as the primary drivers of ccRCC and pRCC, respectively.

**Hypothesis:** We hypothesize that specific germline and somatic alterations in coding and non-coding regions are associated with racial disparities in kidney cancer and that these mutations are associated with the known RCC driver genes *VHL* and *MET.*

**Aim 1: To perform whole-genome sequencing (WGS) of African-Americans with ccRCC to complete a missing aspect of The Cancer Genome Atlas (TCGA).** WGS will be performed on a cohort of African-American patients treated for ccRCC at Yale. African-American patients are underrepresented in TCGA ccRCC cohorts. Thus, the proposed Yale sequencing and the TCGA cohorts are complementary. African-American patients will be carefully matched against the TCGA cohort. This results in availability of a combined Yale-TCGA study population, suitable for comparing genomic alterations according race (Caucasian vs. African-American) and histologic subtype (ccRCC vs. pRCC).

**Aim 2: To discover a set of variants that are both associated with kidney cancer and also racially disparate.** We will first find all genomic annotations associated with the canonical drivers of kidney cancer, MET and VHL. We term these assembled annotations the ‘METome’ and ‘VHLome’. These include annotations of binding sites, related transcription factors, enhancers, and other regulatory elements. Then we will find all the variants in these regions and rank them in terms of the impact, according to both functional impact and recurrence. Next, we will re-rank the variants in terms of their racial disparities. We will perform the same functional and disparity rankings in small sub-regions, such as binding sites, that are particularly burdened. Finally, we will pass to the next aim the variants that score highly both in terms of association with kidney cancer and racial disparities.

**Aim 3: To validate specific variants and mutated regions suspected of contributing to kidney cancer racial disparities in a large, independent cohort.** In order to validate our discovered variants (Aim 2), we will form an independent cohort of African-American and Caucasian patients treated for ccRCC and pRCC. We will use the Connecticut State Registry, which has a substantial number of individuals (n = 11,000). Because of the size of this cohort, we will be able to balance it with respect to known kidney cancer risk factors (e.g. obesity) and assure statistically rigorous results. Tumor and normal DNA from candidate genomic regions identified in Aim 3 will be assessed using PCR-Mass spectroscopy.

**Aim 4: To perform functional characterization of a prioritized, high-confidence list of genetic variants.** Immortalized cell lines will be genotyped and gene-edited using a CRISPR/Cas system. This will allow the creation of matched cell lines to study the activity of the VHL/HIF or HGF/MET pathways. In addition to HIF reporter assays, gene expression and protein abundance will be evaluated to discern changes in activity. Finally, we will evaluate the effects of these alterations on cell proliferation, invasion, migration, and anchorage-independent growth.

**Deliverables:** We will produce a number of concrete data and analytic deliverables from this project. First, we will complete whole genome sequencing to cover a missing aspect of TCGA and produce a set of sequencing files. Then we will create annotation files for the VHLome and METome relating non-coding regions to these important driver genes. Then we'll produce a list of variants affecting the VHLome and METome and a ranking of these variants, both in terms of functional impact and racial disparity. We will also perform rankings in terms of small sub-regions such as binding sites. Finally we will report our validation results as an additional annotation to this table of impactful variants. We will make all of this available through a project website by producing standard data files. We'll also make available all our code through standard repositories such as GitHub.

**A. Significance**

Renal cell carcinoma (RCC) constitutes over 90% of kidney cancers and is the most lethal genitourinary malignancy\cite{25559415}. Known RCC risk factors include male gender, age, hypertension, obesity, chronic kidney disease, and smoking\cite{17704406; 18343443; 14665346}. However, these factors are likely contributory in less than half of RCC cases\cite{21486465; 9737554}, 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 African-Americans\cite{25559415; 10235157}. Genetic risk factors have been postulated to play an important role in RCC racial disparities\cite{18343443; 21486465; 21741761; 20448658}, but to date, no comprehensive study has specifically focused on deciphering genetic mechanisms associated with racial disparity. Given that the main driver genes of RCC have been well characterized -- *VHL in* ccRCC and *MET* in pRCC -- much of our analyses will focus on these genes.

The MET protein (encoded by the *c-Met* gene) is a transmembrane receptor-linked tyrosine kinase. 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\cite{26505625; 19496715; 28373408}, which we propose to study in the work detailed here.

*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. VHL is categorized as a tumor suppressor gene in part because loss of VHL function promotes angiogenesis which may increase solid tumor growth. Inherited mutations within this gene have been linked to a number of cancer types, including those in the kidney, pancreas, and brain\cite{19216840; 19657325; 22022277}.

With a focus on *MET* and *VHL*, we will analyze the frequency, functional impact, and genomic burdening of mutations that lead the resulting proteins to lose their native functions or gain new ones, namely, loss-of-function (LoF) and gain-of-function (GoF) mutations across samples. We will study the patterns of LoF and GoF variants in TSGs and oncogenes in patients from different races 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 \cite{5279523}, it is an aggregate effect of several mutations that 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.

[[HM significance+contribution [...]]

**B. Innovation**

In contrast to other cancer types such as prostate cancer\cite{27814645; 23878300}, racial disparities in kidney cancer have not been well studied. In this work, we are interested in identifying key genomic alterations which contribute to the greater incidence and distinct histologic distribution of kidney cancer in African-Americans relative to Caucasians. This study will be the first comprehensive assessment of somatic and germline alterations in kidney cancer by race. We will expand upon prior work from 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. Using our novel bioinformatics tools to analyze WGS data, we will define driver alterations and study the interaction patterns between somatic and germline variations important in kidney cancer risk and progression. Whole genome analysis, including non-coding genome variation analysis in relation to canonical coding driver mutations, will expand our search beyond the small "genic" landscape currently known to be strongly associated with kidney cancer (e.g. MET and VHL). This study will be a first step in addressing the genetic causes of racial disparities in kidney cancer.

**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 has broadened our understanding of the genomic basis of various forms of kidney cancer. TCGA has led to the understanding that different cell types within the kidney may give rise to distinct forms of cancer. Somatic alterations (driver mutations and copy number variants) are also important for determining a cancer’s molecular profile. In TCGA, kidney cancer cases were submitted by 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 surprisingly, only a limited number of African-Americans with clear-cell kidney cancer were included in TCGA analysis. Despite the fact that African-Americans account 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%) (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 the genetic basis of racial disparities.

Despite available data, there has not been a thorough analysis of somatic driver alterations or germline risk variants in kidney cancer among African-Americans. 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, 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, using a patient cohort with a different genetic background, whole genome sequencing might illustrate novel, ethnicity-specific driver events, as recently seen in an African-American prostate cancer study\cite{28515055}.

Whole-genome sequencing offers several advantages over traditional chip-based methods. It allows analysis of poorly-tagged or rare SNPs, INDELs and structural variants (SVs). Moreover, whole-genome sequencing has single nucleotide resolution which helps to isolate disease-causing variants rather than large 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 a team of pathologists. Whole blood is procured to serve as a genomic control. In the past 2 years, over 300 patients 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 completing the TCGA dataset, we will study 15 African-American subjects with ccRCC from 2013-2016. TCGA kidney cancer projects have captured patient age, sex, race, smoking history, and have 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 that 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 be performed using Illumina’s HiSeq 2000 technology. In brief, DNA fragments from each sample will be hybridized using HiSeq Paired-End Cluster Kits and 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 our active participation in the 1000 Genomes Consortium - especially from our participation in the analysis working group and the structural variant (SV) and functional interpretation (FIG) subgroups, in which the majority of the variant calling tools were developed, deployed and interpreted\cite{21293372; 20981092; 23128226}.

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 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\cite{21478889} to generate initial raw variant call sets using the GATK haplotype caller. We will use the parameters consistent with those used in TCGA\cite{26536169}.

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\cite{23396013} and Strelka\cite{22581179} 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 a number of diseases, including cancer. We have developed a number of SV calling algorithms, including BreakSeq\cite{20037582}, CNVnator\cite{21324876}, AGE\cite{21233167}, and PEMer\cite{19236709}. Furthermore, we have studied the SVs that originate from different mechanisms and may have potentially divergent functional impacts\cite{24092746; 26028266}. We will run CNVnator to identify CNVs in each cancer sample. We will apply CREST\cite{21666668} to identify large structural variations. Furthermore, we will run our BreakSeq tool to decipher the underlying mechanism.

Along with our new sequenced samples, we will reprocess all the 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 for both African-American ccRCC patients at Yale and TCGA kidney cancer patients. This will be done consistent with methodology already used by TCGA. This catalog 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 catalog of variants will serve as a basis for identifying genomic aberrations associated with racial disparity observed in kidney cancer. We plan to make our sequencing data available via dbGAP (see data dissemination plan).

**C-1-e Problems and Solutions:** We do not anticipate difficulty acquiring samples and performing whole-genome sequencing.

**Aim 2: To discover 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 the canonical driver genes in kidney cancer (MET and VHL). This METome and VHLome will include all regions which we find directly or indirectly impact the function of MET and VHL in RCC. We will utilize this extensive catalog to prioritize coding and non-coding variants associated with kidney cancer. This prioritization step will be performed with the help of an extensive suite of software tools which have been applied in prior studies. In addition to prioritizing germline variants and somatic variants in MET and VHL, we will also study the interplay between somatic and germline mutations (i.e., their combined effects), with a focus on searching for racial disparities. The key mutations identified in these analyses will be evaluated in a larger validation cohort.

**[[[ PDM stops & hM takes]]**

**C-2-b** **Relevant Preliminary Results:** Here, we outline some of our published and publicly available tools and methods that have been devised to prioritize variants in large-scale sequencing studies. These pipelines may 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 We have developed tools for somatic and germline burden tests:** 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 terms of loss-of-function events, as well as in the contexts of both allosteric regulation and local mechanistic perturbations.

We have 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 cancer driver genes.

With respect to allosteric effects, we have developed the STRESS software tool\cite{27066750}. STRESS (STRucturally-identified ESSential residues) employs models of large-scale protein conformational change 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 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 within the cores of TSGs.

In addition to coding variants, we have also developed FunSeq, a tool to prioritize non-coding variants in cancer. 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 TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitivity sites). In our published work using FunSeq\cite{24092746}, we integrated large-scale data from various resources, including ENCODE and 1000 Genomes Project, with cancer genomics data. By comparing patterns of inherited polymorphisms from 1092 humans with somatic variants from cancer patients, FunSeq identifies candidate non-coding driver mutations.

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 noncoding elements by comparing observed mutation counts with expected counts under a whole-genome background mutation model\cite{26304545}. 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\cite{26304545}. Furthermore, we have 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*). Both annotation-based and variant-based permutation are supported.

**C-2-b-2 We have extensive experience in analyzing whole-genome datasets from cancer cohorts including kidney cancer:** We are active participants in TCGA and Pan-cancer Analysis of Whole Genomes (PCAWG) consortium projects. Specifically, we have played key roles in the TCGA investigations into prostate\cite{26544944} and kidney\cite{26536169} cancers. We participated in the TCGA KICH (chromophobe RCC) project\cite{25155756} and a following pan-subtype kidney analysis\cite{4794376}. Our team analyzed the whole-genome sequencing data for the TCGA KIRP (pRCC), now published in *The New England Journal of Medicine*\cite{26536169}. In recent work, we leveraged our expertise in 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. 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 (COCA) immunologic profile from gene and miRNA expression datasets.

Together with other published results on RCC\cite{25790038; 25401301; 23792563; 23797736; 25826081}, 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 have 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 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. In addition to identifying canonical driver mutations, we identified many high-impact mutations which can potentially influence cancer progression.

**C-2-c Research plan**:

**C-2-c-1 [[LS, PDM] Constructing the METome & VHLome: 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.

For example, 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.

This integration will produce an 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 noncoding regions. Our results have biologically interpretability because genetic modules will be linked with genes.

**C-2-c-2 Build a comprehensive mutation catalog for the METome and VHLome**: We aim to build an inclusive, comprehensive mutation catalog with variants assembled from both our newly sequenced dataset and public 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. We will also include regions that are associated with disparities between Caucasian and African-Americans in other forms of cancer, such as prostate cancer\cite{23770567; 27535533}. Prior work has shown that RCC is uniquely characterized by CNVs as an early and major driver event\cite{25790038}. 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. Last, we will also gather both germline and somatic mutations from TCGA and PCAWG. To estimate background mutation rates in the general population, we will leverage both gnomAD\cite{27535533} and ExAC\cite{27535533}.

Currently, ExAC and gnomAD report 677 variants in ~31,000 whole-genome sequenced 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 7 somatic MET mutations. In 161 TCGA WXS pRCC samples, we discovered 15 nonsynonymous 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 within secondary association). Therefore, we estimate that there are ~20,000 germline SNVs and ~1,500 somatic SNVs from public dataset and our newly-sequenced samples.

## **C-2-c-3 Run our coding variant prioritization pipeline on all variants in both coding and non-coding region**: Once we establish our comprehensive set of variants (in both coding as well as non-coding regions) for both the METome and VHLome, we will prioritize them both by their inferred impact as well as 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 has been 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 the MET and VHL genes have resulted in detailed 3D models of their structures. Using high-resolution crystallographic models of these two proteins, we will run our tools to annotate coding variants associated with MET and VHL genes within our sample cohort, followed by running the remainder of our coding variant prioritization pipeline.  
 In addition to coding variants, many changes in noncoding regions regulating the METome and VHLome may play critical roles in renal cell cancer 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 mutation hot spots within the genome. With the addition of many more samples, we will perform comprehensive prioritization to identify many more non-coding variants that may play key roles in kidney cancer.

As mentioned, 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 on the comprehensive kidney cancer variant catalog. 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 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 BMI \cite{25673413}, hypertension \cite{28951623}, smoking\cite{20418890} 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 Inter-relate (co-occurrence) somatic & germline Degeneralize - coocc{{

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

[[HM -- shortened version]] We will investigate mutation patterns in African-American and Caucasian cohorts in the combined Yale-TCGA RCC data repository. Differences in major and minor allele frequency distributions across patient samples constitute an important factor that we will consider. Somatic variants can complement cancer-related germline variants and lead to the development of different cancer types.}}

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

[[HM/PDM -- Two paragraphs -- all the prioritized SNPs -- go into ExAC 1000 Genomes, calculate some differential AF, (FST) is correct -- we rank on that]]

[[HM/PDM -- Put on a line for differential burdening -- SKAT -- (mentioned in approaches)]]

Following prioritization of MET- and VHL-related variants across all TCGA-Yale samples, we will study differences in the frequency and functional impacts between variant sets. We will also 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 SKAT\cite{21737059}, we will find regions that are significantly burdened by germline mutations in the kidney cancer cases relative to healthy controls.

To perform these analyses, we will leverage a number of popular tools. Along with genomic samples in the combined Yale-TCGA cohort, we will mine several other genomic repositories, including the 1000 Genomes (1KG) whole-genome samples\cite{26432245; 21293372; 20981092; 24092746; 21596777}, the ExAC meta-cohort (with TCGA samples excluded)\cite{27535533}, and gnomAD\cite{27535533}. This will allow us to find racially disparate genetic elements based on allele frequency distributions using a Fisher exact test, fixation index (FST), and unified sequence-based tests in He et al\cite{28844485}.

**C-2-d Differential burdening between Caucasian and African-Americans within METome and VHLome germline genomic regions**

**C-2-d-1 Variant level analysis:** To analyse coding regions, we will employ the full 467 samples with whole-exome data from TCGA. To analyze common variants 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 and their association with elements in the VHLome and METome networks. 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 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 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-d-2 Region based analysis:** Beyond investigating the associations between single common variants and race, we will focus on evaluating 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, affect phenotype in the same direction (e.g., LoFs disabling a tumor suppressor). We assume that, in total, there are *n* patients with available whole-exome sequencing data. For a target region, for example, consider a gene that harbors *m* variants. Let  denote the a race-based indicator variable for the patient. for African-Americans and 0 otherwise. Let represent the genotype of patient . A logistic regression model can be used to evaluate associations (equation 1). Suppose that describes the mean of the population status. Then

 (1)

For the burden test, we could treat the coefficient for each patient as a weighted coefficient like . Then equation (1) can be rewritten as:

 (2)

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

 (3)

The allele frequency can be used to assign the weight for each variant. For example,, where  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 is treated as an independent random variable with 0 mean and variance . The null hypothesis can be then be changed to H0: , and the test statistic in equation (1) can be written to:

 (4)

In (4), is the kernel matrix, and is the genotype information vector. 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 to

(5)

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:

 (6)

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

 (7)

**C-3-d-3 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, 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-2-d-4 Power analysis using SKAT for per region based analysis:** In the 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 (Figure **XXX**). In our proposed study, we will focus on genomic modules linked with kidney cancer such as MET- and VHLome, and therefore expect a large number of effective mutations. Typically, the MET genomic region consists of 126,027 nucleotides while VHL of 12,035 respectively. We expect these numbers to increase significantly (likely up to ten-fold) after creating the genomic modules of METome and VHLome.

**C-2-e Deliverables:** We aim to generate a list of regions in genome that have the greatest potential to impact RCC development. We will construct a list of genetic modules that are assembled from high-impact regions. We will make these regions available from our project website and as tables in published papers (see data dissemination plan). We will also make all the software tools available in public repositories (e.g. GitHub) to ensure reproducibility. We will then pass lists of ranked genes, regions and variants for validation in Aim 3. 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).

**C-2-f 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\cite{5391127} to search for significant alterations in kidney cancer. We believe that we can readily find additional alterations, some of which will show racial disparities. Thus, we see no difficulties with finding a small number of significant variants within the thousands that we will survey.  
 In particular, by analyzing more than 20,000 SNVs (germline and somatic), we are confident that we will identify more than 55 regions with racial disparity to further process in aim 3. In the event of fewer mutations, we will complete our validation list with our most prioritized regions from aim 2 and reexamine these for racial disparities using an independent cohort.

[[HM- Aim3-- validate the racially disparate variants (as in old aim 4) + include genotyping + covariates + power analysis of validation cohort]]

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

**C-3-a Rationale**: Aims I and II together represent a discovery phase, where we identify and prioritize genomic regions for racial disparities associated with cancer. In Aim 3, 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. This cohort will include both African-Americans and Caucasians with clear cell and papillary RCC, to allow comparisons across histologic type and race. In addition, we will study environmental and clinical factors that might be associated with racial disparity in RCC by utilizing the electronic health records (EHRs) associated with all samples.

**C-3-b Genotyping for specimen acquisition, and DNA extraction:** For our validation cohort, 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 55 highly prioritized regions (100bp each) for 384 individuals from aim 2. As mentioned above, specific kidney cancer risk factors may influence risk of RCC. To control for these differences, cases will be matched 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 cases are selected, 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 Tumor Registry IRB approved protocol. All tumors will have been centrally reviewed by our genitourinary pathologists and classified according to recent International Society of Urologic Pathology (ISUP) criteria\cite{24025519}. For both fresh and FFPE tissue, DNA will be extracted from the tumor and adjacent normal kidney using an automated Maxwell 16® System (Promega, Madison, WI).

**C-3-c Genotyping for sample matching:** The recent 1000 Genomes phase 3 study showed African Americans (ASW) carry a significant amount of European ancestry\cite{26432245}, 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\cite{19407144; 25529636}. 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\cite{24509691}, and ii) by Principal Component Analysis\cite{16862161}. We will include reference genotypes from HAPMAP and 1000 Genomes in this analysis. This will allow 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\cite{26040208; 23161897}.

**C-3-d Studying environmental and clinical covariates using EHR records of the validation cohort**: Racial disparity 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 accommodate for additional genomics and electronic health record data to be collected during or after the study.

**C-3-e 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 2. 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\cite{92795038}, 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 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 (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.

**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 **xxx** candidates, we will find at least one or two to study further.   
Moreover, in our recent publication, we found a germline SNP, rs11762213, predicts type 2 pRCC patients prognosis and might play a role in pRCC incidence in African Americans\cite{5391127}. It has been shown rs11762213 is also associated with prognosis in ccRCC\cite{26505625, 23219378}. The underlying mechanism remains to be elucidated. Therefore, we do functional studies (Aim 4) on rs11762213, along with other variants that we found **or prioritized from aims 2 and 3**.

Aim 4: Functional Characterization of Key Mutations

1500 - 2100 words + 1-2 figs.

**Aim 4: Functional characterization of candidate gene alterations in MET and VHL pathway.**

**C-4-a Rationale:** We focus on the coding and non-coding mutations of *VHL* and *MET*, both important drivers of tumorigenesis in clear cell and papillary RCC. We will interrogate the functional impact of our top racially disparate, candidate gene alterations in these key kidney cancer signaling pathways.

**C-4-b Research Plan:**

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

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 as they have been successfully altered with this technology**.** The urologic oncology laboratory recently has collaborated on projects (Letter of support Dr. Ranjit Bindra) where CRISPR/Cas-based gene targeting protocols have been optimized specifically to interrogate matched immortalized primary human cell lines. This highlight our ability to (a) design gRNAs targeting our candidate regions of *VHL* or *MET*, (b) to transfect plasmids encoding gRNAs with Cas9 specifically in primary human cells, (c) to induce synonymous or non-synonymous SNPs, non-coding mutations, or splice site mutations, and (d) to isolate single-cell clones via FACS for expansion. For both *VHL* and *MET,* sequencing will confirm successful cell line generation. DNA will be isolated from clonally expanded populations. A 200bp segment containing the target sites will be PCR amplified and gel purified for sequencing. For validation of the success of *VHL* models, we will perform western blots to see if we can detect the truncated VHL protein products by western blot analysis using antibodies specific to the N-terminus. To assess off-target CRISPR alterations, the highest scoring off-target coding sites will be interrogated using PCR amplification/sequencing to those regions.

**C-4-b-2 Interrogation of downstream HGF/MET and VHL/HIF/VEGF Pathway Activation.   
C-4-b-2-a MET and VEGFR2 protein/phospho-protein quantification:** Mesoscale 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. Whole cell lysates of our matched cell lines are prepared using Triton X-100 containing buffer with protease and phosphatase inhibitors. Human KDR Base Kit, Phospho-VEGFR-2 Whole Cell Lysate Kit, and Phospho/Total Met Whole Cell Lysate Kit (Meso Scale Discovery, Rockville, MD) can be used for total and phosphoprotein quantification and calculated by the signal intensity from total and phospho-MET levels and total and phospho-KDR.

**C-4-b-2-b 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). Digital gene expression profiling will be performed 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. Gene expression analysis was will be performed using the Nanostring nCounter system for digital RT-PCR and then analyzed with nSolverAnalysis Software. (NanoString Technologies, Seattle, WA). We have used this system for classification of clear cell RCC tumor regions into profiles “A” and “B” based on their angiogenic signature largely due to expression differences in HIF2/EPAS1.

**C-4-b-2-c 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, which will be measured using a tetrazolium (WST-1) colorimetric assay where the cleavage of the tetrazolium salt WST-1 to formazan will be quantified by a plate reader at 420-480 nm; (2) apoptosis, which will be evaluated using the TUNEL assay where the TUNEL-positive cells will be recorded; (3) migration, which a confluent monolayer of cells will be scratched and the distance the cells migrated will be logged from 3 to 24 h; (4) invasion, which will be analyzed using a transwell assay where the cells that invade a Matrigel chamber will be fixed, stained, and visualized; and (5) anchorage-independent growth, which will be evaluated by colony formation in soft agar over three weeks before the colonies are stained, visualized, and quantified. We have successfully performed such assays in 96 well plates with clear cell kidney cancer lines previously to interrogate the role of XL184 (VEGFR2/MET inhibition) with and without HGF stimulation in 96 well plates using absorbance to quantify visible colonies.

**C-4-c Expected Outcome and Alternative Approaches.** These experiments are important to functionally validate the activity of the candidate racially disparate alterations. *VHL* and *MET* in our primary kidney cell lines. Although unlikely, We we may encounter various difficulties in discovering including finding a small number of high impact candidate gene alterations. Nevertheless, we We already have identified several novel variants in *MET* that have not been characterized and, including the *MET* SNP rs11762213 which we will have the capability to interrogate regardless of aims 1-3. As *VHL* is a tumor suppressor gene, 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.

[[STL: paste the end of doc last year below]]

**D. Potential Pitfalls and Alternative Strategies**

A potential issue with our proposal is that, despite our design, we may not find any regions in the validation cohort that are significantly different between races. 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: (1) 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. (2) 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. (3) The validation cohort can be expanded 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 with a diagnosed with kidney cancer from 1998 to present.

**E. Expected Outcomes and Future Directions**

At the conclusion of this research, our analyses will determine candidate coding and non-coding regions associated with papillary and clear cell kidney cancer. We will identify and validate specific germline and somatic alterations that are disparate between African-Americans and Caucasians with kidney cancer. These findings will be an initial step towards understanding the genomic cause of kidney cancer racial disparities, and have implications beyond the scope of this project. Understanding inherited predisposition to kidney cancer may have important screening implications in high-risk individuals, such as African-Americans. Additionally, racial disparities in candidate driver alterations has the potential to impact how we view treatment in the age of precision-based cancer therapeutics. The findings from this project have far reaching implications, justifying further research beyond the scope of this proposal.

References

1. Siegel, R. L., Miller, K. D., Jemal, A. (2015) Cancer statistics, 2015. *CA Cancer J Clin* 65(1):5-29.

2. Berndt, S. I., Carter, H. B., Schoenberg, M. P., Newschaffer, C. J. (2007) Disparities in treatment and outcome for renal cell cancer among older black and white patients. *J Clin Oncol* 25(24):3589-95.

3. Stafford, H. S., Saltzstein, S. L., Shimasaki, S., Sanders, C., Downs, T. M., Sadler, G. R. (2008) Racial/ethnic and gender disparities in renal cell carcinoma incidence and survival. *J Urol* 179(5):1704-8.

4. Vaishampayan, U. N., Do, H., Hussain, M., Schwartz, K. (2003) Racial disparity in incidence patterns and outcome of kidney cancer. *Urology* 62(6):1012-7.

5. Lipworth, L., Tarone, R. E., McLaughlin, J. K. (2011) Renal cell cancer among African Americans: an epidemiologic review. *BMC Cancer* 11():133.

6. Benichou, J., Chow, W. H., McLaughlin, J. K., Mandel, J. S., Fraumeni, Jr, J. F. (1998) Population attributable risk of renal cell cancer in Minnesota. *Am J Epidemiol* 148(5):424-30.

7. Chow, W. H., Devesa, S. S., Warren, J. L., Fraumeni, Jr, J. F. (1999) Rising incidence of renal cell cancer in the United States. *JAMA* 281(17):1628-31.

8. Ljungberg, B., Campbell, S. C., Choi, H. Y., Cho, H. Y., Jacqmin, D., Lee, J. E., Weikert, S., Kiemeney, L. A. (2011) The epidemiology of renal cell carcinoma. *Eur Urol* 60(4):615-21.

9. Chow, W.-H., Dong, L. M., Devesa, S. S. (2010) Epidemiology and risk factors for kidney cancer. *Nat Rev Urol* 7(5):245-57.

10. Hakimi, A. A., Ostrovnaya, I., Jacobsen, A., Susztak, K., Coleman, J. A., Russo, P., Winer, A. G., Mano, R., Sankin, A. I., Motzer, R. J., Voss, M. H., Offit, K., Purdue, M., Pomerantz, M., Freedman, M., Choueiri, T. K., Hsieh, J. J., Klein, R. J. (2016) Validation and genomic interrogation of the MET variant rs11762213 as a predictor of adverse outcomes in clear cell renal cell carcinoma. *Cancer* 122(3):402-10.

11. Giubellino, A., Linehan, W. M., Bottaro, D. P. (2009) Targeting the Met signaling pathway in renal cancer. *Expert Rev Anticancer Ther* 9(6):785-93.

12. Salgia, R. (2017) MET in Lung Cancer: Biomarker Selection Based on Scientific Rationale. *Mol Cancer Ther* 16(4):555-565.

13. Cowey, C. L., Rathmell, W. K. (2009) VHL gene mutations in renal cell carcinoma: role as a biomarker of disease outcome and drug efficacy. *Curr Oncol Rep* 11(2):94-101.

14. Clark, P. E. (2009) The role of VHL in clear-cell renal cell carcinoma and its relation to targeted therapy. *Kidney Int* 76(9):939-45.

15. Moore, L. E., Nickerson, M. L., Brennan, P., Toro, J. R., Jaeger, E., Rinsky, J., Han, S. S., Zaridze, D., Matveev, V., Janout, V., Kollarova, H., Bencko, V., Navratilova, M., Szeszenia-Dabrowska, N., Mates, D., Schmidt, L. S., Lenz, P., Karami, S., Linehan, W. M., Merino, M., Chanock, S., Boffetta, P., Chow, W.-H., Waldman, F. M., Rothman, N. (2011) Von Hippel-Lindau (VHL) inactivation in sporadic clear cell renal cancer: associations with germline VHL polymorphisms and etiologic risk factors. *PLoS Genet* 7(10):e1002312.

16. Knudson, Jr, A. G. (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* 68(4):820-3.

17. Bhardwaj, A., Srivastava, S. K., Khan, M. A., Prajapati, V. K., Singh, S., Carter, J. E., Singh, A. P. (2017) Racial disparities in prostate cancer: a molecular perspective. *Front Biosci (Landmark Ed)* 22():772-782.

18. Cooperberg, M. R. (2013) Re-examining racial disparities in prostate cancer outcomes. *J Clin Oncol* 31(24):2979-80.

19. Huang, F. W., Mosquera, J. M., Garofalo, A., Oh, C., Baco, M., Amin-Mansour, A., Rabasha, B., Bahl, S., Mullane, S. A., Robinson, B. D., Aldubayan, S., Khani, F., Karir, B., Kim, E., Chimene-Weiss, J., Hofree, M., Romanel, A., Osborne, J. R., Kim, J. W., Azabdaftari, G., Woloszynska-Read, A., Sfanos, K., De Marzo, A. M., Demichelis, F., Gabriel, S., Van Allen, E. M., Mesirov, J., Tamayo, P., Rubin, M. A., Powell, I. J., Garraway, L. A. (2017) Exome Sequencing of African-American Prostate Cancer Reveals Loss-of-Function ERF Mutations. *Cancer Discov* 7(9):973-983.

20. Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stütz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. A., Korbel, J. O., 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470(7332):59-65.

21. 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., McVean, G. A. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061-73.

22. 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., McVean, G. A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56-65.

23. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., Daly, M. J. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491-8.

24. Cancer Genome Atlas Research Network, Linehan, W. M., Spellman, P. T., Ricketts, C. J., Creighton, C. J., Fei, S. S., Davis, C., Wheeler, D. A., Murray, B. A., Schmidt, L., Vocke, C. D., Peto, M., Al Mamun, A. A. M., Shinbrot, E., Sethi, A., Brooks, S., Rathmell, W. K., Brooks, A. N., Hoadley, K. A., Robertson, A. G., Brooks, D., Bowlby, R., Sadeghi, S., Shen, H., Weisenberger, D. J., Bootwalla, M., Baylin, S. B., Laird, P. W., Cherniack, A. D., Saksena, G., Haake, S., Li, J., Liang, H., Lu, Y., Mills, G. B., Akbani, R., Leiserson, M. D. M., Raphael, B. J., Anur, P., Bottaro, D., Albiges, L., Barnabas, N., Choueiri, T. K., Czerniak, B., Godwin, A. K., Hakimi, A. A., Ho, T. H., Hsieh, J., Ittmann, M., Kim, W. Y., Krishnan, B., Merino, M. J., Mills Shaw, K. R., Reuter, V. E., Reznik, E., Shelley, C. S., Shuch, B., Signoretti, S., Srinivasan, R., Tamboli, P., Thomas, G., Tickoo, S., Burnett, K., Crain, D., Gardner, J., Lau, K., Mallery, D., Morris, S., Paulauskis, J. D., Penny, R. J., Shelton, C., Shelton, W. T., Sherman, M., Thompson, E., Yena, P., Avedon, M. T., Bowen, J., Gastier-Foster, J. M., Gerken, M., Leraas, K. M., Lichtenberg, T. M., Ramirez, N. C., Santos, T., Wise, L., Zmuda, E., Demchok, J. A., Felau, I., Hutter, C. M., Sheth, M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J. C., Zhang, J., Ayala, B., Baboud, J., Chudamani, S., Liu, J., Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Ally, A., Balasundaram, M., Balu, S., Beroukhim, R., Bodenheimer, T., Buhay, C., Butterfield, Y. S. N., Carlsen, R., Carter, S. L., Chao, H., Chuah, E., Clarke, A., Covington, K. R., Dahdouli, M., Dewal, N., Dhalla, N., Doddapaneni, H. V., Drummond, J. A., Gabriel, S. B., Gibbs, R. A., Guin, R., Hale, W., Hawes, A., Hayes, D. N., Holt, R. A., Hoyle, A. P., Jefferys, S. R., Jones, S. J. M., Jones, C. D., Kalra, D., Kovar, C., Lewis, L., Li, J., Ma, Y., Marra, M. A., Mayo, M., Meng, S., Meyerson, M., Mieczkowski, P. A., Moore, R. A., Morton, D., Mose, L. E., Mungall, A. J., Muzny, D., Parker, J. S., Perou, C. M., Roach, J., Schein, J. E., Schumacher, S. E., Shi, Y., Simons, J. V., Sipahimalani, P., Skelly, T., Soloway, M. G., Sougnez, C., Tam, A., Tan, D., Thiessen, N., Veluvolu, U., Wang, M., Wilkerson, M. D., Wong, T., Wu, J., Xi, L., Zhou, J., Bedford, J., Chen, F., Fu, Y., Gerstein, M., Haussler, D., Kasaian, K., Lai, P., Ling, S., Radenbaugh, A., Van Den Berg, D., Weinstein, J. N., Zhu, J., Albert, M., Alexopoulou, I., Andersen, J. J., Auman, J. T., Bartlett, J., Bastacky, S., Bergsten, J., Blute, M. L., Boice, L., Bollag, R. J., Boyd, J., Castle, E., Chen, Y.-B., Cheville, J. C., Curley, E., Davies, B., DeVolk, A., Dhir, R., Dike, L., Eckman, J., Engel, J., Harr, J., Hrebinko, R., Huang, M., Huelsenbeck-Dill, L., Iacocca, M., Jacobs, B., Lobis, M., Maranchie, J. K., McMeekin, S., Myers, J., Nelson, J., Parfitt, J., Parwani, A., Petrelli, N., Rabeno, B., Roy, S., Salner, A. L., Slaton, J., Stanton, M., Thompson, R. H., Thorne, L., Tucker, K., Weinberger, P. M., Winemiller, C., Zach, L. A., Zuna, R. (2016) Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N Engl J Med* 374(2):135-45.

25. Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., Getz, G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 31(3):213-9.

26. Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., Cheetham, R. K. (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 28(14):1811-7.

27. Lam, H. Y. K., Mu, X. J., Stütz, A. M., Tanzer, A., Cayting, P. D., Snyder, M., Kim, P. M., Korbel, J. O., Gerstein, M. B. (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* 28(1):47-55.

28. Abyzov, A., Urban, A. E., Snyder, M., Gerstein, M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 21(6):974-84.

29. Abyzov, A., Gerstein, M. (2011) AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* 27(5):595-603.

30. Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M., Gerstein, M. B. (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 10(2):R23.

31. Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., Das, J., Abyzov, A., Balasubramanian, S., Beal, K., Chakravarty, D., Challis, D., Chen, Y., Clarke, D., Clarke, L., Cunningham, F., Evani, U. S., Flicek, P., Fragoza, R., Garrison, E., Gibbs, R., Gümüş, Z. H., Herrero, J., Kitabayashi, N., Kong, Y., Lage, K., Liluashvili, V., Lipkin, S. M., MacArthur, D. G., Marth, G., Muzny, D., Pers, T. H., Ritchie, G. R. S., Rosenfeld, J. A., Sisu, C., Wei, X., Wilson, M., Xue, Y., Yu, F., 1000 Genomes Project Consortium, Dermitzakis, E. T., Yu, H., Rubin, M. A., Tyler-Smith, C., Gerstein, M. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342(6154):1235587.

32. Abyzov, A., Li, S., Kim, D. R., Mohiyuddin, M., Stütz, A. M., Parrish, N. F., Mu, X. J., Clark, W., Chen, K., Hurles, M., Korbel, J. O., Lam, H. Y. K., Lee, C., Gerstein, M. B. (2015) Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun* 6():7256.

33. Wang, J., Mullighan, C. G., Easton, J., Roberts, S., Heatley, S. L., Ma, J., Rusch, M. C., Chen, K., Harris, C. C., Ding, L., Holmfeldt, L., Payne-Turner, D., Fan, X., Wei, L., Zhao, D., Obenauer, J. C., Naeve, C., Mardis, E. R., Wilson, R. K., Downing, J. R., Zhang, J. (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* 8(8):652-4.

34. Habegger, L., Balasubramanian, S., Chen, D. Z., Khurana, E., Sboner, A., Harmanci, A., Rozowsky, J., Clarke, D., Snyder, M., Gerstein, M. (2012) VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* 28(17):2267-9.

35. Clarke, D., Sethi, A., Li, S., Kumar, S., Chang, R. W. F., Chen, J., Gerstein, M. (2016) Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. *Structure* 24(5):826-837.

36. Kumar, S., Clarke, D., Gerstein, M. (2016) Localized structural frustration for evaluating the impact of sequence variants. *Nucleic Acids Res* 44(21):10062-10073.

37. Lochovsky, L., Zhang, J., Fu, Y., Khurana, E., Gerstein, M. (2015) LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res* 43(17):8123-34.

38. Cancer Genome Atlas Research Network (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell* 163(4):1011-25.

39. Davis, C. F., Ricketts, C. J., Wang, M., Yang, L., Cherniack, A. D., Shen, H., Buhay, C., Kang, H., Kim, S. C., Fahey, C. C., Hacker, K. E., Bhanot, G., Gordenin, D. A., Chu, A., Gunaratne, P. H., Biehl, M., Seth, S., Kaipparettu, B. A., Bristow, C. A., Donehower, L. A., Wallen, E. M., Smith, A. B., Tickoo, S. K., Tamboli, P., Reuter, V., Schmidt, L. S., Hsieh, J. J., Choueiri, T. K., Hakimi, A. A., The Cancer Genome Atlas Research Network, Chin, L., Meyerson, M., Kucherlapati, R., Park, W.-Y., Robertson, A. G., Laird, P. W., Henske, E. P., Kwiatkowski, D. J., Park, P. J., Morgan, M., Shuch, B., Muzny, D., Wheeler, D. A., Linehan, W. M., Gibbs, R. A., Rathmell, W. K., Creighton, C. J. (2014) The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* 26(3):319-330.

40. Leont'ev, A. P., Minchenkova, B. I., Osipov, V. T. (1973) [Emitters for the apparatus for ultrasonic therapy]. *Med Tekh* 5(0):23-5.

41. Bütler, R., Brunner, E. (1969) Inability of Ag-Anti-Ag complexes to fix complement and its significance. *Vox Sang* 17(5):462-4.

42. Kovac, M., Navas, C., Horswell, S., Salm, M., Bardella, C., Rowan, A., Stares, M., Castro-Giner, F., Fisher, R., de Bruin, E. C., Kovacova, M., Gorman, M., Makino, S., Williams, J., Jaeger, E., Jones, A., Howarth, K., Larkin, J., Pickering, L., Gore, M., Nicol, D. L., Hazell, S., Stamp, G., O'Brien, T., Challacombe, B., Matthews, N., Phillimore, B., Begum, S., Rabinowitz, A., Varela, I., Chandra, A., Horsfield, C., Polson, A., Tran, M., Bhatt, R., Terracciano, L., Eppenberger-Castori, S., Protheroe, A., Maher, E., El Bahrawy, M., Fleming, S., Ratcliffe, P., Heinimann, K., Swanton, C., Tomlinson, I. (2015) Recurrent chromosomal gains and heterogeneous driver mutations characterise papillary renal cancer evolution. *Nat Commun* 6():6336.

43. Durinck, S., Stawiski, E. W., Pavía-Jiménez, A., Modrusan, Z., Kapur, P., Jaiswal, B. S., Zhang, N., Toffessi-Tcheuyap, V., Nguyen, T. T., Pahuja, K. B., Chen, Y.-J., Saleem, S., Chaudhuri, S., Heldens, S., Jackson, M., Peña-Llopis, S., Guillory, J., Toy, K., Ha, C., Harris, C. J., Holloman, E., Hill, H. M., Stinson, J., Rivers, C. S., Janakiraman, V., Wang, W., Kinch, L. N., Grishin, N. V., Haverty, P. M., Chow, B., Gehring, J. S., Reeder, J., Pau, G., Wu, T. D., Margulis, V., Lotan, Y., Sagalowsky, A., Pedrosa, I., de Sauvage, F. J., Brugarolas, J., Seshagiri, S. (2015) Spectrum of diverse genomic alterations define non-clear cell renal carcinoma subtypes. *Nat Genet* 47(1):13-21.

44. Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499(7456):43-9.

45. Sato, Y., Yoshizato, T., Shiraishi, Y., Maekawa, S., Okuno, Y., Kamura, T., Shimamura, T., Sato-Otsubo, A., Nagae, G., Suzuki, H., Nagata, Y., Yoshida, K., Kon, A., Suzuki, Y., Chiba, K., Tanaka, H., Niida, A., Fujimoto, A., Tsunoda, T., Morikawa, T., Maeda, D., Kume, H., Sugano, S., Fukayama, M., Aburatani, H., Sanada, M., Miyano, S., Homma, Y., Ogawa, S. (2013) Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet* 45(8):860-7.

46. Christinat, Y., Krek, W. (2015) Integrated genomic analysis identifies subclasses and prognosis signatures of kidney cancer. *Oncotarget* 6(12):10521-31.

47. Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., Getz, G. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457):214-218.

48. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G., Exome Aggregation Consortium (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285-91.

49. Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., Croteau-Chonka, D. C., Esko, T., Fall, T., Ferreira, T., Gustafsson, S., Kutalik, Z., Luan, J., Mägi, R., Randall, J. C., Winkler, T. W., Wood, A. R., Workalemahu, T., Faul, J. D., Smith, J. A., Zhao, J. H., Zhao, W., Chen, J., Fehrmann, R., Hedman, Å. K., Karjalainen, J., Schmidt, E. M., Absher, D., Amin, N., Anderson, D., Beekman, M., Bolton, J. L., Bragg-Gresham, J. L., Buyske, S., Demirkan, A., Deng, G., Ehret, G. B., Feenstra, B., Feitosa, M. F., Fischer, K., Goel, A., Gong, J., Jackson, A. U., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Mangino, M., Leach, I. M., Medina-Gomez, C., Medland, S. E., Nalls, M. A., Palmer, C. D., Pasko, D., Pechlivanis, S., Peters, M. J., Prokopenko, I., Shungin, D., Stančáková, A., Strawbridge, R. J., Sung, Y. J., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S. W., van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L., Zhang, W., Isaacs, A., Albrecht, E., Ärnlöv, J., Arscott, G. M., Attwood, A. P., Bandinelli, S., Barrett, A., Bas, I. N., Bellis, C., Bennett, A. J., Berne, C., Blagieva, R., Blüher, M., Böhringer, S., Bonnycastle, L. L., Böttcher, Y., Boyd, H. A., Bruinenberg, M., Caspersen, I. H., Chen, Y.-D. I., Clarke, R., Daw, E. W., de Craen, A. J. M., Delgado, G., Dimitriou, M., Doney, A. S. F., Eklund, N., Estrada, K., Eury, E., Folkersen, L., Fraser, R. M., Garcia, M. E., Geller, F., Giedraitis, V., Gigante, B., Go, A. S., Golay, A., Goodall, A. H., Gordon, S. D., Gorski, M., Grabe, H.-J., Grallert, H., Grammer, T. B., Gräßler, J., Grönberg, H., Groves, C. J., Gusto, G., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hartman, C. A., Hassinen, M., Hayward, C., Heard-Costa, N. L., Helmer, Q., Hengstenberg, C., Holmen, O., Hottenga, J.-J., James, A. L., Jeff, J. M., Johansson, Å., Jolley, J., Juliusdottir, T., Kinnunen, L., Koenig, W., Koskenvuo, M., Kratzer, W., Laitinen, J., Lamina, C., Leander, K., Lee, N. R., Lichtner, P., Lind, L., Lindström, J., Lo, K. S., Lobbens, S., Lorbeer, R., Lu, Y., Mach, F., Magnusson, P. K. E., Mahajan, A., McArdle, W. L., McLachlan, S., Menni, C., Merger, S., Mihailov, E., Milani, L., Moayyeri, A., Monda, K. L., Morken, M. A., Mulas, A., Müller, G., Müller-Nurasyid, M., Musk, A. W., Nagaraja, R., Nöthen, M. M., Nolte, I. M., Pilz, S., Rayner, N. W., Renstrom, F., Rettig, R., Ried, J. S., Ripke, S., Robertson, N. R., Rose, L. M., Sanna, S., Scharnagl, H., Scholtens, S., Schumacher, F. R., Scott, W. R., Seufferlein, T., Shi, J., Smith, A. V., Smolonska, J., Stanton, A. V., Steinthorsdottir, V., Stirrups, K., Stringham, H. M., Sundström, J., Swertz, M. A., Swift, A. J., Syvänen, A.-C., Tan, S.-T., Tayo, B. O., Thorand, B., Thorleifsson, G., Tyrer, J. P., Uh, H.-W., Vandenput, L., Verhulst, F. C., Vermeulen, S. H., Verweij, N., Vonk, J. M., Waite, L. L., Warren, H. R., Waterworth, D., Weedon, M. N., Wilkens, L. R., Willenborg, C., Wilsgaard, T., Wojczynski, M. K., Wong, A., Wright, A. F., Zhang, Q., LifeLines Cohort Study, Brennan, E. P., Choi, M., Dastani, Z., Drong, A. W., Eriksson, P., Franco-Cereceda, A., Gådin, J. R., Gharavi, A. G., Goddard, M. E., Handsaker, R. E., Huang, J., Karpe, F., Kathiresan, S., Keildson, S., Kiryluk, K., Kubo, M., Lee, J.-Y., Liang, L., Lifton, R. P., Ma, B., McCarroll, S. A., McKnight, A. J., Min, J. L., Moffatt, M. F., Montgomery, G. W., Murabito, J. M., Nicholson, G., Nyholt, D. R., Okada, Y., Perry, J. R. B., Dorajoo, R., Reinmaa, E., Salem, R. M., Sandholm, N., Scott, R. A., Stolk, L., Takahashi, A., Tanaka, T., van 't Hooft, F. M., Vinkhuyzen, A. A. E., Westra, H.-J., Zheng, W., Zondervan, K. T., ADIPOGen Consortium, AGEN-BMI Working Group, CARDIOGRAMplusC4D Consortium, CKDGen Consortium, GLGC, ICBP, MAGIC Investigators, MuTHER Consortium, MIGen Consortium, PAGE Consortium, ReproGen Consortium, GENIE Consortium, International Endogene Consortium, Heath, A. C., Arveiler, D., Bakker, S. J. L., Beilby, J., Bergman, R. N., Blangero, J., Bovet, P., Campbell, H., Caulfield, M. J., Cesana, G., Chakravarti, A., Chasman, D. I., Chines, P. S., Collins, F. S., Crawford, D. C., Cupples, L. A., Cusi, D., Danesh, J., de Faire, U., den Ruijter, H. M., Dominiczak, A. F., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Felix, S. B., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N. G., Forrester, T., Franco, O. H., Gansevoort, R. T., Gejman, P. V., Gieger, C., Gottesman, O., Gudnason, V., Gyllensten, U., Hall, A. S., Harris, T. B., Hattersley, A. T., Hicks, A. A., Hindorff, L. A., Hingorani, A. D., Hofman, A., Homuth, G., Hovingh, G. K., Humphries, S. E., Hunt, S. C., Hyppönen, E., Illig, T., Jacobs, K. B., Jarvelin, M.-R., Jöckel, K.-H., Johansen, B., Jousilahti, P., Jukema, J. W., Jula, A. M., Kaprio, J., Kastelein, J. J. P., Keinanen-Kiukaanniemi, S. M., Kiemeney, L. A., Knekt, P., Kooner, J. S., Kooperberg, C., Kovacs, P., Kraja, A. T., Kumari, M., Kuusisto, J., Lakka, T. A., Langenberg, C., Marchand, L. L., Lehtimäki, T., Lyssenko, V., Männistö, S., Marette, A., Matise, T. C., McKenzie, C. A., McKnight, B., Moll, F. L., Morris, A. D., Morris, A. P., Murray, J. C., Nelis, M., Ohlsson, C., Oldehinkel, A. J., Ong, K. K., Madden, P. A. F., Pasterkamp, G., Peden, J. F., Peters, A., Postma, D. S., Pramstaller, P. P., Price, J. F., Qi, L., Raitakari, O. T., Rankinen, T., Rao, D. C., Rice, T. K., Ridker, P. M., Rioux, J. D., Ritchie, M. D., Rudan, I., Salomaa, V., Samani, N. J., Saramies, J., Sarzynski, M. A., Schunkert, H., Schwarz, P. E. H., Sever, P., Shuldiner, A. R., Sinisalo, J., Stolk, R. P., Strauch, K., Tönjes, A., Trégouët, D.-A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M.-C., Völker, U., Waeber, G., Willemsen, G., Witteman, J. C., Zillikens, M. C., Adair, L. S., Amouyel, P., Asselbergs, F. W., Assimes, T. L., Bochud, M., Boehm, B. O., Boerwinkle, E., Bornstein, S. R., Bottinger, E. P., Bouchard, C., Cauchi, S., Chambers, J. C., Chanock, S. J., Cooper, R. S., de Bakker, P. I. W., Dedoussis, G., Ferrucci, L., Franks, P. W., Froguel, P., Groop, L. C., Haiman, C. A., Hamsten, A., Hui, J., Hunter, D. J., Hveem, K., Kaplan, R. C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N. G., März, W., Melbye, M., Metspalu, A., Moebus, S., Munroe, P. B., Njølstad, I., Oostra, B. A., Palmer, C. N. A., Pedersen, N. L., Perola, M., Pérusse, L., Peters, U., Power, C., Quertermous, T., Rauramaa, R., Rivadeneira, F., Saaristo, T. E., Saleheen, D., Sattar, N., Schadt, E. E., Schlessinger, D., Slagboom, P. E., Snieder, H., Spector, T. D., Thorsteinsdottir, U., Stumvoll, M., Tuomilehto, J., Uitterlinden, A. G., Uusitupa, M., van der Harst, P., Walker, M., Wallaschofski, H., Wareham, N. J., Watkins, H., Weir, D. R., Wichmann, H.-E., Wilson, J. F., Zanen, P., Borecki, I. B., Deloukas, P., Fox, C. S., Heid, I. M., O'Connell, J. R., Strachan, D. P., Stefansson, K., van Duijn, C. M., Abecasis, G. R., Franke, L., Frayling, T. M., McCarthy, M. I., Visscher, P. M., Scherag, A., Willer, C. J., Boehnke, M., Mohlke, K. L., Lindgren, C. M., Beckmann, J. S., Barroso, I., North, K. E., Ingelsson, E., Hirschhorn, J. N., Loos, R. J. F., Speliotes, E. K. (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518(7538):197-206.

50. Tobacco and Genetics Consortium (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 42(5):441-7.

51. Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89(1):82-93.

52. 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., Abecasis, G. R. (2015) A global reference for human genetic variation. *Nature* 526(7571):68-74.

53. Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. K., Gerstein, M. B. (2011) Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* 39(16):7058-76.

54. He, Z., Xu, B., Lee, S., Ionita-Laza, I. (2017) Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in Metabochip Data. *Am J Hum Genet* 101(3):340-352.

55. Srigley, J. R., Delahunt, B., Eble, J. N., Egevad, L., Epstein, J. I., Grignon, D., Hes, O., Moch, H., Montironi, R., Tickoo, S. K., Zhou, M., Argani, P., ISUP Renal Tumor Panel (2013) The International Society of Urological Pathology (ISUP) Vancouver Classification of Renal Neoplasia. *Am J Surg Pathol* 37(10):1469-89.

56. Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., Williams, S. M. (2009) The genetic structure and history of Africans and African Americans. *Science* 324(5930):1035-44.

57. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., Mountain, J. L. (2015) The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am J Hum Genet* 96(1):37-53.

58. Salichos, L., Stamatakis, A., Rokas, A. (2014) Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol Biol Evol* 31(5):1261-71.

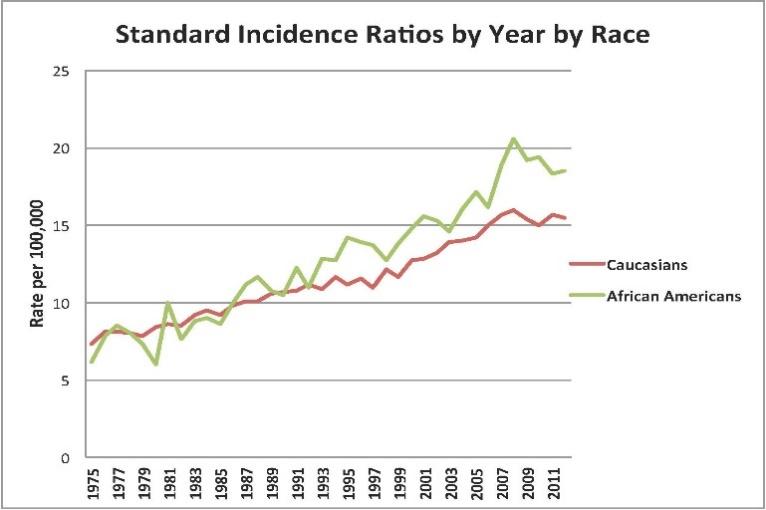
59. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904-9.

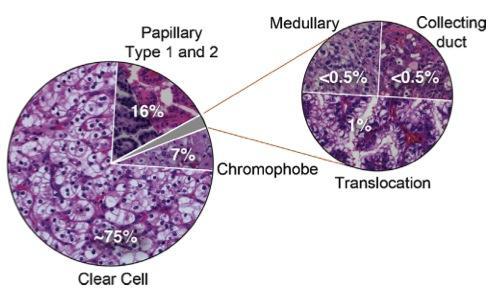
60. Zhang, K., Wiener, H., Aissani, B. (2015) Admixture mapping of genetic variants for uterine fibroids. *J Hum Genet* 60(9):533-8.

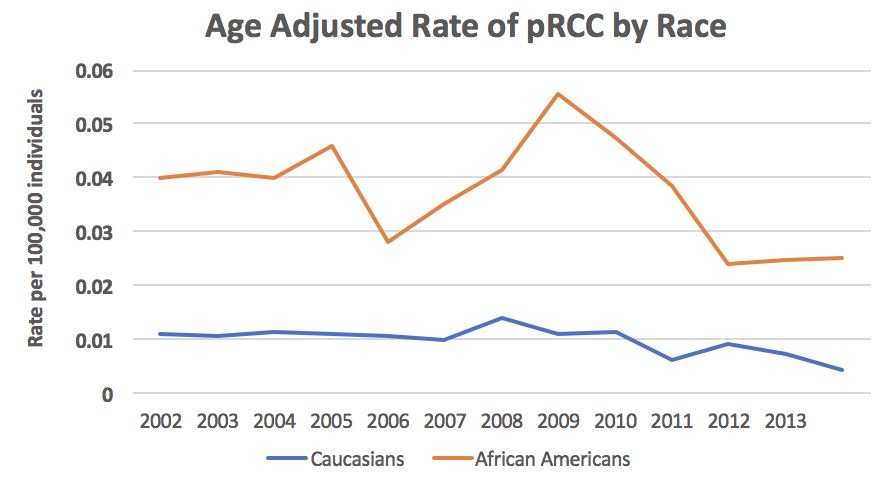
61. Wise, L. A., Ruiz-Narvaez, E. A., Palmer, J. R., Cozier, Y. C., Tandon, A., Patterson, N., Radin, R. G., Rosenberg, L., Reich, D. (2012) African ancestry and genetic risk for uterine leiomyomata. *Am J Epidemiol* 176(12):1159-68.

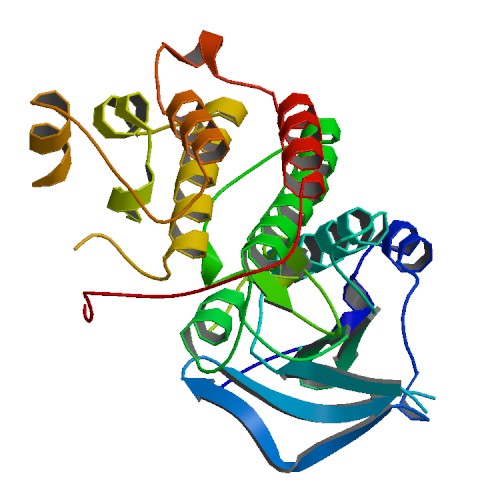
62. Dunn, P. K., Smyth, G. K. (1996) Randomized Quantile Residuals. *J. Comput. Graph. Stat* 5(3):236--244.

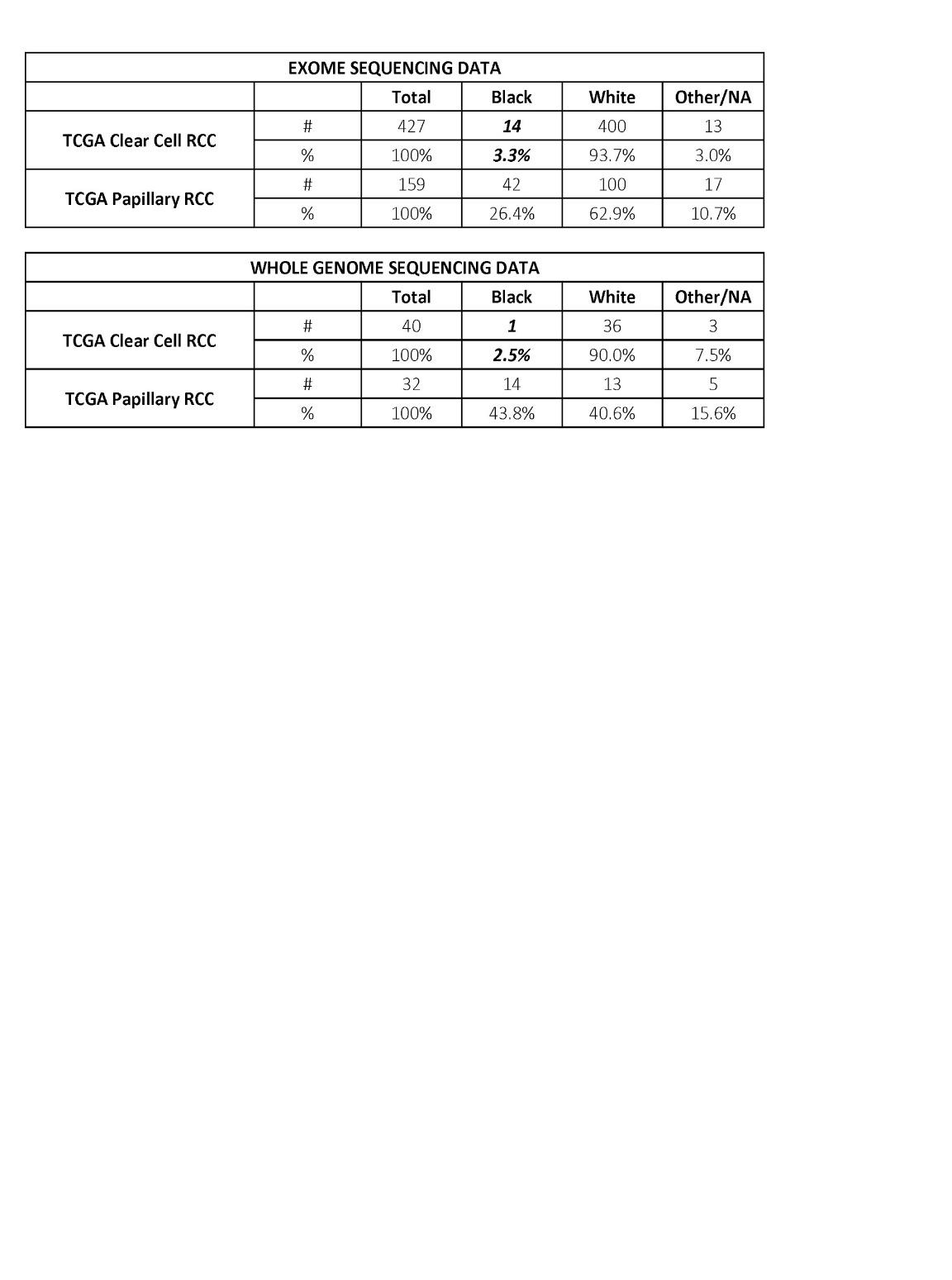
63. Schutz, F. A. B., Pomerantz, M. M., Gray, K. P., Atkins, M. B., Rosenberg, J. E., Hirsch, M. S., McDermott, D. F., Lampron, M. E., Lee, G.-S. M., Signoretti, S., Kantoff, P. W., Freedman, M. L., Choueiri, T. K. (2013) Single nucleotide polymorphisms and risk of recurrence of renal-cell carcinoma: a cohort study. *Lancet Oncol* 14(1):81-7.



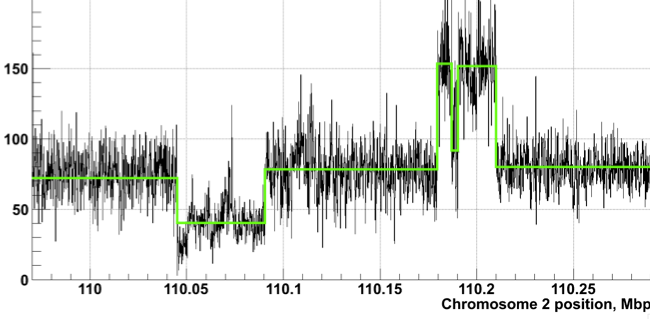




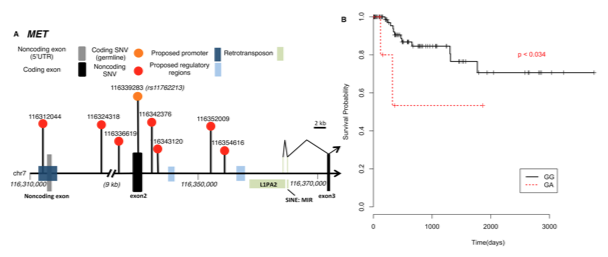












## 

## 





Summary

There exist significant racial disparities in kidney cancer including early disease onset for African-Americans, different histologic distribution, and worst disease outcome, even when controlling for treatment. No study has yet evaluated a biologic or cause of racial disparities in kidney cancer. In our study, we intend to investigate whether these disparities have a genetic basis in clear cell (ccRCC) and papillary (pRCC) kidney cancer focusing on two major kidney cancer genes MET and VHL. Our first aim is to perform whole genome sequencing of 15 African-Americans with ccRCC in order to increase the number of African-Americans with kidney cancer in TCGA and complement existing TCGA kidney cancer cohorts. For these individuals and for the TCGA ccRCC and pRCC cohorts, we intend to perform high-quality mutation calls for structural and genomic variation including SNPs, deletions, insertions, inversions and copy number variations (CNV). Our second aim is to assemble a comprehensive list of both somatic & germline mutations associated directly or indirectly with MET and VHL (what we term the ‘METome’ and ‘VHLome’) and prioritize regions according to greatest impact and racial disparity. We will begin constructing the METome and VHLome by finding all annotations of molecular interaction and regulatory relationships for these genes. Then we will use a novel pipeline that includes i) FunSeq, a sophisticated algorithm that prioritizes high functional impact variants, and ii) LARVA, a burden test algorithm that identifies significant mutation enrichment in non coding elements, by also considering covariates such as mutation rate, replication time etc. Then, we will test a list of prioritized regions and modules for racial genomic differences. More specifically, we intend to identify racial disparities across i) germline mutations in coding regions using whole-genome sequencing (WGS) and whole-exome sequencing (WES) data, ii) genomic regions with higher mutational burdens, iii) germline mutations in non-coding prioritized regions using WGS data and iv) somatic mutations in prioritized regions. In our third objective, we intend to validate our prioritized racially disparate regions using a large, independent Yale validation cohort consisting of total 384[1] individuals, balanced to including clear cell and papillary tumors with 96 Caucasian and African-American patients for each tumor type. Patients in this large validation cohort will be genotyped using PCR-Mass spectroscopy. Finally, in our fourth objective, we will perform a CRISPR/Cas validation on a selected subset of validated and prioritized mutations in genotyped immortalized cell lines. We will determine the effect of gene and protein expression alterations on proliferation, invasion, migration, and anchorage-independent growth.

Is this still the number we’re aiming for?

Narrative

We investigate the genomic cause of racial disparities in renal cell carcinoma. We will evaluate the genetic burdening of MET and VHL genes and their respective interactive regions in clear cell and papillary renal tumors from African American and Caucasians. We will identify and prioritize both common and rare, germline and somatic variants, in coding and non-coding regions. Utilizing a range of cancer genome data and novel bioinformatics tools, our goal is to prioritize and experimentally validate variants contributing to racial disparity.