# **Significance [1.5 pg, #659]**

Renal cell carcinoma makes up over 90% of kidney cancers and currently is the most lethal genitourinary malignancy.[1] The incidence of kidney cancer has nearly tripled in recent years in all races however the most dramatic increase is seen in African Americans relative to other populations in the United States.[1-3]According to the NCI Surveillance and Epidemiology End Results (SEER) Cancer Program,[4] the age-adjusted incidence of kidney cancers compared to Caucasians(18.5 and 15.5 cases per 100,000 persons, respectively (Figure 1). To date, research has not fully explained increased susceptibility to RCC among African Americans.[5] Various hypotheses have been proposed implicating both genetic risk variants and a greater prevalence of kidney cancer risk factors in African Americans including obesity, chronic kidney disease (CKD), and hypertension.[6-10]

Besides the higher incidence of kidney cancer among African Americans several other racial disparities have been described. Several studies have demonstrated that African American with kidney cancer have a much younger of presentation with a median age of onset is 3-8 years earlier than Caucasians (Table 2). In kidney cancer, age of onset is important in the accepted criteria for consideration for genetic testing as many hereditary cancers develop at a younger age than generally observed in the generation population.(8) While over a dozen known monogenic cancer syndromes exist, more often inherited risk and early disease onset is believed due to a complex inheritance pattern. Specific risk alleles may contribute to the racial disparity in kidney cancer perhaps, predisposing to an earlier age of onset.

An additional disparity in kidney cancer is the large racial differences in the types of kidney cancers found between African Americans and Caucasians. RCC is a collection of tumors with the two most common subtypes, clear cell and papillary accounting 85% of all cases. Although papillary RCC is considered to comprise 10-15% of renal tumors in general (Figure 2), several published and ongoing studies have shown that this subtype is approximately three-fold more common in African Americans, accounting for 35-40% of cases (Table 2).3, 4, 20, 21 The reasons for the increased incidence of papillary RCC in AA are currently unknown. Unfortunately when metastatic, papillary RCC has an abysmal prognosis, possibly related to the lack of therapeutic options for this disease**.[1]**

Another major aspect of racial disparity in kidney cancer is that survival is also significantly worse among African Americans. One explanation is that various studies have found that African Americans less frequently receive standard treatments in the United States. However even controlling for treatment and tumor characteristics including, stage, grade, and subtype, survival is still significantly worse.[2-4] Similar to prostate cancer where AA patients have a more aggressive disease biology,[5] it has been proposed that differences in molecular biology are involved ethnic disparities in kidney cancer.[2, 6, 7] While various studies have investigated genomic differences in prostate cancer among specific races,[8, 9] to date, no research has investigated the biologic differences between African American and Caucasian forms of kidney cancer.

Clear racial differences exist in the incidence, mortality, age of onset, and subtype distribution of kidney cancer. However no study has addressed genetic mechanisms associated with racial disparities in African American kidney cancer. While the Cancer Genome Atlas (TCGA) has analyzed hundreds of kidney cancer specimens, due to the referral patterns from the centers involved, many cohorts are predominantly from Caucasian patients. Therefore we intend expand upon TCGA analyses by including an additional cohort of African American’s with kidney cancer. Including these samples and performing secondary data analysis of the existing dataset we will be able to compare differences in possible risk variants, driver mutation, and driver copy number alterations. Additionally using our novel sequencing tools to analyze whole genome data, we will define and then validate non-coding, intronic driver alterations important in kidney cancer risk and progression. This study will be the first step in addressing the biological/genetic causes of cancer health disparities in kidney cancer and the findings have implications far beyond the scope of this current proposal.

 **Aim 1 - fill out TCGA kidney cancer dataset for racial disparity analysis by doing select WGS on ~15 black cases [2pg #856]**

In recent years various TCGA efforts have characterized the genomic basis of clear cell, papillary, and chromophobe kidney cancer. These studies have led to the understanding that some of the diversity within kidney cancer results from different cells of origin give rise to distinct types of cancer within the same organ. Additionally differences in somatic alterations (driver mutations and copy number variations) are important in determining a cancer’s molecular profile. In the TCGA, cases were submitted from various high volume tertiary centers to the Biospecimen Core Resource (BCR) for accessioning and specimen processing. Specimens however were not submitted in a coordinated fashion to ensure the study population has a similar profile of that encountered nationally. Not surprising, there was clearly a limited number of African Americans with clear cell kidney cancer included in the TCGA analysis. Despite African Americans accounting for approximately 1 in 7 cases of kidney cancer, only a cursory analysis was performed in this population including 14/427 (3.3%) samples that underwent whole exome sequencing (WES) (Table X) and 1/40 (2.5%) (Table Y) that underwent whole genome sequencing (WGS). Failing to include a larger population of African Americans with clear cell RCC limits our ability to explore the genomic rationale for racial disparities. As African Americans do have a much more frequency greater number of African Americans were included in the papillary kidney cancer TCGA cohort. However, despite the available data, there has not been a thorough analysis of somatic driver alterations and germline risk variants that may be more common in African American kidney cancer. We propose to complete the TCGA analysis of the top two subtypes of kidney cancer, papillary and clear cell, but analyzing an additional cohort of African Americans with clear cell kidney cancer. By performing whole genome sequencing on this additional cohort of samples, we will have an adequate number of cases to allow comparisons between African American and Caucasian clear cell and papillary kidney cancers.

**Sample Acquisition and DNA Extraction:**

All patients undergoing scheduled kidney cancer surgery at Yale New Haven Hospital are offered enrollment into an IRB-approved Genitourinary Biospecimen repository (P.I. Shuch, HIC# 0805003787). Within 30 minutes of removal, fresh tumor tissue is snap frozen in liquid nitrogen by the pathology team. Additionally whole blood is procured to serve as a genomic control. In the past 2 years, over 300 subjects with kidney cancer have been prospectively enrolled. All fresh biospecimens are stored at -80˚C and are available for immediate analysis. For the purpose of completion of the TCGA dataset, we will utilize a consecutive series of 15 African American subjects with clear cell kidney cancer from 2013-2015. DNA will be extracted from fresh tumor tissue and whole blood using an automated Maxwell 16® System (Promega, Madison, WI).

**Whole Genome Sequencing:** Describe techniques

We have extensive experience in large-scale variant calling through being active members of the 1000 Genomes Consortium, especially in the analysis working group and the structural variant (SV) and functional interpretation (FIG) subgroups of the consortium where the majority of the variant calling tools are developed and used \cite{12952525,19015660,12952525}. We have already developed a prototype pipeline for calling germline and somatic variants. We will use the Broad’s Genome Analysis Toolkit (GATK)\cite{21478889} to call germline SNPs and INDELS.We will map raw FASTQ files of each sample to the hg19 reference genome using bwa-mem algorithm with default parameters to generate BAM files. These bam files will be further processed to sort and mark duplicates reads before calling variants.

We will follow GATK best practices to generate initial raw variant call sets using GATK haplotype caller. We will filter these initial call sets by running GATK variant recalibration tool. The filtering strategy based on variant recalibration method uses a continuous adaptive error model. The adaptive error model takes into account the relationship between annotation of each variant (Quality score, mapping quality, strandedness & allele info) and the probability of it being a true positive instead of a sequencing artifact. Furthermore, we will exclude any filtered variant, which falls in low mapability region of the genome. In addition, we will utilize MuTect \cite{23396013} and Strelka \cite{ 22581179} to call somatic SNVs and INDELs, respectively.

Structural variations (SVs) are important contributors to human polymorphism, have functional impact and are often implicated in various diseases including cancer. We have developed a number of SV calling algorithms, including BreakSeq \cite{20037582, which compares raw reads with a breakpoint library (junction mapping) , CNVnator, which measures read depth\cite{21324876}, AGE, which refines local alignment \cite{21233167}, and PEMer, which uses paired ends\cite{19236709}. We have also developed array-based approaches \cite{19037015} and a sequencing-based bayesian model\cite{21034510}. Furthermore, we have intensively studied the distinct features of SVs originated from different mechanisms. This indicates specific creation processes and potentially divergent functional impacts \cite{24092746,26028266}. We will perform extensive molecular characterization of germline and somatic SVs in these cancer samples. We will run CNVnator to identify germline and somatic copy number variations in each cancer samples. We will apply CREST \cite{21666668} to generate germline and somatic large structural variations including large deletions, insertion, inversion, intra & inter-chromosomal translocations. Furthermore, we will run our BreakSeq tool to decipher the underlying mechanism of somatic and germline SV formation.

# **Aim 2 - Assemble a catalog of somatic & germline mutations relevant to kidney cancer & prioritize regions with greatest impact [STL - 3.5pg #1250]**

**2.1 Rational**: In this study, we aim to discover underlying genetic regions that are responsible for race disparity in renal cell carcinoma. However, due to the limited size of sequenced samples, it is not feasible to test every region in the genome. In fact, we have to carefully limit our searching space to get enough sufficient statistical power. Therefore, we will assemble a catalog of mutations that are relevant to renal cell carcinoma and prioritize regions with greatest impact. In this way, we will incorporate our best prior knowledge of renal cell carcinoma and cancer genomics in general into this study and bring down the number of tests significantly to avoid losing statistical power. In the end, we will generate a list of potentially high influential regions in both coding and noncoding associated with renal cell carcinoma.

**2.2 Preliminary result**

**2.2.a We've developed ways of Prioritizing high-functional impact variants.** We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets\cite{21596777,22950945,22955619}. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations\cite{21596777}. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region\cite{22955616}. Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery\cite{18077332,23505346}). In recent studies\cite{24092746,25273974}, we have integrated and extended these methods to develop a prioritization pipeline called FunSeq (Fig 2). It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). FunSeq links each non-coding mutation to target genes, and prioritizes such variants based on scaled network connectivity. It identifies deleterious variants in many non-coding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. It also detects their disruptiveness in TF binding sites (both loss-of and gain-of function events). Integrating large-scale data from various resources (including ENCODE and The 1000 Genomes Project) with cancer genomics data, our method is able to prioritize the known TERT promoter driver mutations, and it scores somatic recurrent mutations higher than those that are non-recurrent. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples \cite{24092746}. We have also applied our method to investigate non-coding mutation patterns in subtypes of gastric cancer\cite{submitted}. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group. A new, developing feature of FunSeq will link genomic elements, especially noncoding regions, together with genes or other well-annotated regions in the genome and assemble them into modules. Those modules help better understand functions and impacts of noncoding variations.

**2.2.b Tools for identifying enrichment of variations in coding and noncoding regions.** We have worked on statistical methods for analysis of nc regulatory regions. LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichments in nc elements by comparing observed mutation counts with expected counts under a whole genome background mutation model. LARVA also includes corrections for biases in mutation rate owing to DNA replication timing. For coding region analysis, we developed MuSiC[44] to analyze genetic changes using standardized sequence-based inputs, along with multiple types of clinical data, to establish correlations among variants, affected genes and pathways, and to ultimately separate commonly abundant passenger events from truly significant events.

LARVA was proposed in our initial noncoding variant characterization plans in earlier grants. Since that time, we have developed a highly efficient software tool that implements the proposed mutation burden statistical calculations. We used this tool in a pan-cancer analysis of 760 cancer whole genomes’ variants spanning a number of cancer data portals and some published datasets. Our analyses demonstrated that LARVA can recapitulate previously established coding and noncoding cancer drivers, including the TERT and TP53 promoters. We published results of this analysis and the source code in *Nucleic Acids Research* in Sept 2015 \cite{26304545).

We have been working on an update to LARVA, called LARVA 2, which incorporates corrections for additional covariates that influence the somatic mutation rate in different genomic regions, including GC content, transcription level, and chromatin marks. Our intention is to iteratively refine the underlying whole genome background mutation model to reflect all factors that influence the accumulation rate of background mutations. We are also collecting new variant datasets to expand our analyses to new cancers, and identify drivers for individual cancer types.

Beyond LARVA, our variant analysis work includes AlleleDB, our database connecting single nucleotide variants with allele-specific binding and expression, which was recently submitted for publication. We have also participated in PCAWG8, which has coordinated the production of a number of new variant datasets to which we will apply our methods.

**2.2.c We have already identified regions with significant somatic mutations in kidney cancer that we could focus on.**

As an active participant of the TCGA KIRP (Papillary renal cell carcinoma) group, we have gained a significant amount of experience in analyzing renal cell carcinoma sequence results. The manuscript from KIRP group is submitted to New England Journal of Medicine and currently under review. Together with other published results on renal cell carcinoma (cite …), we already have assembled a list of impactful regions on the genome that have shown statistical significance in previous studies. However, most of these studies focused on coding regions only.

In order to examine noncoding regions in the genome, we have run FunSeq on 32 whole genome sequenced samples from the TCGA KIRP group. We have found several harmful mutation hot spots in the genome. Namely, we have found excessive mutations in MET intronic and promoter regions. We expect changes in noncoding regions play a critical role in renal cell cancer.

**2.3 Approach** First, we will mine the literature and condense results from previous studies. We will gather genetic changes that including but not limit to: single nucleotide variation (SNP), structural variation/copy number variation (SV/CNV), and mutation process signature. However, most of the published studies on renal cell carcinoma focus on coding region only. In order to find high impact mutations in noncoding regions, we will run FunSeq and LARVA on variation calls from TCGA whole genome sequenced samples, both cancer and normal. Using existing mutation process decomposition tools, we will also conduct an inspection on all available renal cell carcinoma sequenced samples, but only on tumor samples.

In the end, we will use the developing module of FunSeq to assemble modules. Modules extend high impact regions by linking them with other genomic elements according to physical interaction, molecule pathway and network and other evidence. We anticipate most new elements assembled are noncoding regions with little annotation. But they may play an important role in renal cell carcinoma by interacting and affecting cancer driver genes. Thus we feel it is important to assemble modules to include those elements in our study. Last, modules offer annotation to less known noncoding regions. Our results will be more biologically interpretable.

**2.4 Deliverables - a list of prioritized regions & modules.** We aim to generate a list of regions on the genome that, to our best knowledge, potentially have highest impacts on renal cell carcinoma. We will also construct a list of modules that are assembled from high impact regions. Modules further extend regions in the genome by linking genomic elements using evidence from physical interactions, molecular pathways and network etc. Elements in the same module are expected to play similar roles in renal cell carcinoma initialization and development. We expect those regions and modules have high possibility of involvement in race disparity in renal cell carcinoma. In Aim 3, we will directly test those elements on our samples.

# **Aim 3 - Look for ancestry differences in high impact mutations [LS #1657]**

**Specific AIM 3: Ancestry Differences in High Impact Mutations**

Kidney cancer has almost tripled in recent years, where clear cell and papillary cancers account for 85% of these cases. However, papillary RCC, which comprise 10-15% of renal tumors, shows large racial differences resulting in a threefold increased prevalence for African-Americans, an ethnic group often under-represented in cancer studies and datasets. This enrichment involves different clinical values including the onset of the disease, as well as its prevalence amongst kidney cancer cases overall.

Currently, the reasons for these racial disparities and whether clinical differences are based on varying genomic backgrounds -driver or absolute number of mutations- remain largely unknown. Moreover, in the case that the increased prevalence of papillary RCC amongst African-American is associated with the genetic background, it is unclear whether this discordance is due to an increased number of germline or somatic mutations, in the coding or non-coding region. In this aim, we intend to apply various statistical tests and methods in order to evaluate and model the association between the load of mutations, their location and racial disparity. More specifically, we will use the gene modules from aim II to compare 1) absolute numbers of germline mutations, 2) rare germline mutations, 3) somatic mutations

**3.1 Compare Germline Mutations Between Caucasian and African-American.** Initially, we are planning to compare the total numbers of germline mutations between Caucasian and African-American cancer genotypes (with respect to control groups) to identify whether the latter appear to have a higher number of mutations in either coding or non-coding regions. Moreover, we will test whether there is a particular enrichment of variants in specific gene modules that have already been linked with kidney cancer. In this way, we intend to establish whether African-American individuals bear a higher number of mutations, especially in regions that have been previously associated with cancer. The statistical significance between case-control, as well as between Caucasian and African-American subgroups will be assessed using Fisher exact test.

**3.2 Compare Rare –High Impact- Germline Variants Between Caucasian and African-American.** In aim II, we intend to map and identify rare germline variants by testing each variant against a database of known recurrent mutations. Subsequently, using LARVA burden test for rare variants, we intend to assess whether one ethnic group is more enriched . Due to a potential lack of statistical power, as a confirmatory analysis, we will also apply a SKAT test to assign statistical confidence. Finally, in aim IV, we will proceed with the validation of the most significant rare variants.

**3.3 Compare Somatic mutations Between Caucasian and African-American**. In part three of our third aim we intend to compare the number of somatic mutations between ethnic groups in control cases, cancer genome-scale and cancer-related gene modules. This will enable us to test whether there is different enrinchment in somatic variants between Caucasian and African-American and compare the effect to germline variants from part 3.1.

**Power size**

In our study, power analysis depends upon various parameters including the effect and population size. *Population size* reflects the number of SNPs or SVs that we expect to identify after genome sequencing, while *effect size* has to be indirectly suggested based on reasonable estimations. In figure XXX we provide the number of mutations needed for various effect sizes in order to obtain a 0.05 statistical level of significance. Previous studies have suggested that we should expect about 8-9 mutations per 10Kb (SNPs per 10 kb was 8.33, 8.44, and 8.09 in the human genome, in intergenic regions, and in genic regions, respectively), although in GWAS studies it is often suggested that 1 mutation per 10k is expected to be meaningful (Gtex 2013). However, in our proposed study we have constructed and focus our interest on modules already suggested to be associated with kidney cancer, therefore expecting more than one mutations to be relatively effective. Using Fisher’s exact test for small sizes, in a realistic but optimistic scenario, in order to obtain significance under an effect size of 1.2, which is reasonable provided the threefold prevalence of cancer incidents among African-Americans, we need about 750 mutations to obtain significance. This means that, even though we should have no problem obtaining significance for an overall assessment when comparing ethnic populations, an individual module should have 750:15 which equals about 50 mutations. This may result in not obtaining statistical significance for many individual gene modules, in which case we will provide a careful manual assessment with further validation in the next aim.

**Tools for identifying and comparing rare variants in coding and noncoding regions**

LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichments in nc elements by comparing observed mutation counts with expected counts under a whole genome background mutation model. LARVA also includes corrections for biases in mutation rate owing to DNA replication timing. For coding region analysis, we developed MuSiC[44] to analyze genetic changes using standardized sequence-based inputs, along with multiple types of clinical data, to establish correlations among variants, affected genes and pathways, and to ultimately separate commonly abundant passenger events from truly significant events. We have already applied these tools in a pan-cancer analysis of 760 cancer whole genomes’ variants spanning a number of cancer data portals and some published datasets. Our analyses demonstrated that LARVA can recapitulate previously established coding and noncoding cancer drivers, including the TERT and TP53 promoters. We published results of this analysis and the source code in *Nucleic Acids Research* in Sept 2015 \cite{26304545).

Moreover, we have been working on an update to LARVA, called LARVA 2, which incorporates corrections for additional covariates that influence the somatic mutation rate in different genomic regions, including GC content, transcription level, and chromatin marks. Our intention is to iteratively refine the underlying whole genome background mutation model to reflect all factors that influence the accumulation rate of background mutations. We are also collecting new variant datasets to expand our analyses to new cancers, and identify drivers for individual cancer types.

We have developed LARVA as a highly efficient software tool that implements the proposed mutation burden statistical calculations. Burden tests are generally based on summarizing the rare variants within a collapsed region by a single value (count or weighted value){citation}. In our analysis, we are planning to obtain a single value for each gene module. One potential problem with burden tests arises when the effect of variants tends to be distributed in opposite directions. In other words, if for every deleterious mutation there is an advantageous mutation; the algorithm cannot provide an effect, as there is not any particular direction. Therefore, in this case, as a confirmatory analysis, we will also apply a sequence kernel association test SKAT (for common and rare variants) which uses a multiple regression model to regress the variants into case-control groups{citation}. It should be mentioned though, that a balanced bidirectional effect, should in general contradict the hypothesis that a higher number of mutations is responsible for the higher prevalence of kidney cancer incidents in African-American populations.

# **AIM 4 VALIDATION the key mutations on a larger group [2-3pg #426]**

To confirm somatic and germline variants that are present on the WGS and secondary data analysis a validation cohort will be used

**Sample Acquisition and DNA Extraction:**

As mentioned above, fresh kidney cancer tissue is procured on our IRB-approved Genitourinary Biospecimen repository within 30 minutes of removal. Additionally our protocol also allows access to archival tumor tissue from 1988-2013. Yale pathology archives have available formalin fixed paraffin embedded (FFPE) tissue blocks to retrieve tumor and the adjacent normal kidney tissue for a genomic control. All tumors have recently been centrally reviewed by our genitourinary pathologists and classified according to recent International Society of Urologic Pathology (ISUP) schemes (cite). 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 Yale Validation Cohort. For both fresh and FFPE tissue, DNA will be extracted using an automated Maxwell 16® System (Promega, Madison, WI).

**Tumor Profiling for Somatic Non-Coding Mutations:** A total of 20 ng of total DNA will be obtained from tumor DNA from the Yale Validation Cohort to determine if racial differences exist in somatic non-coding mutations between African Americans and Caucasians with kidney cancer. Small insertions/deletions or single nucleotide alterations found from the WGS and secondary TCGA data analysis will be assess in the validation cohort. Somatic mutational profiling will be evaluated using the MassArray System (Agena Biosciences, San Diego, CA). This platform utilizes mass spectrometry to measure PCR-derived amplicons and allows multiplexing for up to 15 somatic variants per silicon chip. The technology allows screening a large quantity of samples and can detect variants with as low as 1% mutant allele frequency using a small DNA quantity. For somatic variants, the MassArray Assay Design Suite will be used for designing custom PCR primers for use with MassArray.

**Genotyping Kidney Cancer Non-Coding Genomic Variants:** Similar to above, 20 ng of DNA from genomic control will be used in Yale Validation cohort. Using the MassArray system. The 384 subject cohort will have their germline DNA assessed for non-coding variants identified. Similar to somatic variants, the MassArray Assay Design Suite will be used for custom PCR primers design to detect germline variants. For genotyping, up to 40 genomic variants can be multiplexed per well per silicon chip.

**Bioinformatic Analysis of Somatic and Germline Variants in Validation Cohort:** Mutation calls for each tumor and germline sample will be assessed using the MassArray Typer 3.4 Analyzer. Unlike next generation sequencing, the results can be quickly automated and generated into both a graphical or table interactive format.

CITE NATURE GENETICS- Thomas, R. K., Baker, A. C., Debiasi, R. M., Winckler, W., Laframboise, T., Lin, W. M., et al. (2007). High-throughput oncogene mutation profiling in human cancer. *Nat Genet*, *39*(3), 347–351.