# **SPECIFIC AIMS**

# There are an estimated 14,000 annual deaths due to renal cell carcinoma (RCC) in the United States, making this disease the urologic malignancy with the highest mortality. The incidence of RCC has risen at an alarming rate in African-Americans and now is 30% greater than Caucasians. African-Americans also have a worse RCC outcome than any other race for unclear reasons. While various studies have found that African-Americans less frequently receive standard kidney cancer systemic and surgical management, even when controlling for treatment, survival disparities still exist. Beyond access to care barriers, similar to other urologic malignancies such as prostate cancer, biologic differences in disease characteristics may account for racial disparities in kidney cancer. Recent studies have shown that African-Americans present at a significantly younger age than Caucasians. Early disease onset is something in RCC considered a possible indicator of a hereditary predisposition. Another key disparity between races is that the histologic distribution is drastically different in African-Americans. While clear cell RCC (ccRCC) is the most common histologic type in all races, papillary RCC (pRCC) is three-fold more common in African-Americans, accounting for 35-40% of cases. Unfortunately pRCC has been understudied and there are no current forms of effective systemic therapy for this disease. With significant racial disparities present in kidney cancer, we set out to identify possible genomic alterations explaining these cancer disparities. To do so, we will analyze available genomic data contained in the Cancer Genome Atlas (TCGA) kidney cancer cohorts. Since there was a scarcity of African-American subjects with ccRCC included in the TCGA, additional whole genomic sequencing on Yale African-American subjects will be needed to fill out the dataset and make appropriate comparisons across race and the major histologic subtypes. Both coding and non-coding as well as germline and somatic variations associated with RCC will be identified and compared between race and subtype to help identify the genomic basis of kidney cancer racial disparity.

**Hypothesis:**We hypothesize that 1) recurrent non-coding as well as coding germline and somatic alterations are associated with kidney cancer and 2) specific germline and somatic alterations are more common in African-Americans with kidney cancer. We will test our central hypothesis with the following aims:

**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 consecutive series of Yale African-Americans patients with ccRCC, using sequencing parameters similar to those in TCGA, to provide a cohort useful for comparing genomic alterations between race (Caucasian vs. African-American) and histologic subtype (ccRCC vs pRCC).

**Aim 2: To assemble a set of coding and non-coding regions associated with kidney cancer, both in terms of somatic and germline alterations.** We will analyze the African-Americans with ccRCC sequenced in Aim 1 as well as existing TCGA RCC cohorts to identify key genomic regions (and collections of regions, notated as modules) associated with kidney cancer. Novel pipelines for evaluating functional impact and germline and somatic mutational burden will consistently prioritize coding and noncoding regions.

**Aim 3: To identify genomic regions differing most between African-Americans and Caucasians with kidney cancer.** We will analyze genomic regions and modules developed in Aim 2 to identify racial differences in the occurrence of common SNPs or in terms of differential burdening with rare, germline or somatic alterations. We will prioritize and rank many (~500) genomic regions with the highest scores, indicating a possible a hereditary predisposition to kidney cancer.

**Aim 4: To validate specific regions with either germline or somatic mutations suspected of contributing to kidney cancer racial disparity.** We will utilize an independent cohort of African-American and Caucasians patients treated at Yale with ccRCC and pRCC. Tumor and normal DNA will be assessed with a PCR-Mass spectroscopy based system to evaluate candidate genomic regions identified in Aim 3.

**Deliverable:** We plan on developing a coherent data resource, mostly available from the project website, for racial disparities in Kidney cancer, which may be useful for subsequent researchers in this area. This will include consistent variant calls to TCGA for Yale cohort of ccRCC and a tabulation of coding and non-coding regions associated with Kidney, a ranking of them in terms of their racial disparity and an indication of the validation results for the most disparate regions .

**Summary:** Significant racial disparities exist in kidney cancer with African-Americans having increased cancer incidence, earlier age of onset, a different subtype distribution, and worse survival. While multiple studies have characterized epidemiologic differences, to date, there has been no comprehensive investigation into a biologic or genetic cause of this cancer health disparity. *Our proposed research plan will improve our understanding of the genomic cause of racial disparity in RCC and may have implications beyond the scope of this project. We anticipate the findings will provide insight into inherited cancer predisposition and acquired driver alterations in African-Americans with RCC.*

**SIGNIFICANCE[[CHECK NUMBERING W JANINE + EXACT SINGLE SPACING]]**

Renal cell carcinoma (RCC) makes up over 90% of kidney cancers and currently is the most lethal genitourinary malignancy [1]. The incidence of RCC 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, 2]. According to the NCI Surveillance and Epidemiology End Results (SEER) Cancer Program [3], the age-adjusted incidence of kidney cancers compared to Caucasians is 30% greater (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 [4]. Various hypotheses have been proposed implicating both genetic risk variants and a greater prevalence of RCC risk factors in African-Americans including obesity, chronic kidney disease, and hypertension [5, 6, 7, 8].

Besides the higher incidence of RCC among African-Americans, several other racial disparities have been described. Some studies have demonstrated that African-Americans have a younger median age of RCC presentation, between 3-8 years earlier than Caucasians (Table 2). In RCC, age of onset is a major criterion for consideration for genetic testing as many hereditary cancers develop at a younger age than observed in the generation population [9]. While over a dozen known RCC syndromes exist, inherited risk and early disease onset may be more frequently related 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 RCC is the large racial difference in the distribution of histologic subtypes. RCC is a group of cancers arising from the nephron with the two most common subtypes, clear cell RCC (ccRCC) and papillary RCC (pRCC) accounting for 85% of all cases. Although pRCC is considered to comprise 10-15% of renal tumors in general (Figure 2), several published and ongoing studies demonstrate this subtype is three-fold more common in African-Americans, accounting for 35-40% of cases (Table 2) [10, 11, 12]. The reason for the increased pRCC frequency in African-Americans is currently unknown. Unfortunately, when metastatic pRCC has an abysmal prognosis with limited therapeutic options.

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 [4, 13, 14]. Similar to prostate cancer where African-Americans patients have a more aggressive disease biology,[15] it has been proposed that differences in molecular biology are involved racial disparities in kidney cancer [4, 6, 16, 17].

While significant racial differences exist in the incidence, mortality, age of onset, and subtype distribution of kidney cancer, no study has addressed genetic mechanisms associated with RCC racial disparity. While the Cancer Genome Atlas (TCGA) and other sequencing efforts have analyzed hundreds of kidney cancer specimens, the majority of tumors are from Caucasian patients. Without tumors from a diverse cohort of subjects it is difficult to explore to the reasons for these racial disparities including whether clinical differences are based on varying genomic backgrounds or specific driver alterations. With significant racial disparities present, we set out to identify possible coding and non-coding alterations explaining the genomic basis of kidney cancer racial disparity.

# **B. INNOVATION**

In this work, we are interested in identifying key genomic alterations, which primarily contribute to the greater incidence, earlier age of onset, and different histologic distribution of kidney cancer in African-Americans compared to Caucasians. This study will be the first to comprehensively assess genomic alterations in kidney cancer by race. We expand upon prior work from the TCGA by including an additional cohort of African-American’s with ccRCC. 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 bioinformatic tools to analyze whole genome data, we will define and then validate non-coding driver alterations important in kidney cancer risk and progression. This study will be the initial 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.

# **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).[[harmonize]]**

C**-1-a Rationale:** In recent years various TCGA efforts have characterized the genomic basis of the various forms of kidney cancer. These studies have led to the understanding that some of the diversity within kidney cancer results from different cells of origin giving 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 Bio-specimen 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 ccRCC, only a cursory analysis was performed in this population including 14/427 (3.3%) samples that underwent whole exome sequencing (Table 1) and 1/40 (2.5%) (Table 2) that underwent whole genome sequencing. Failing to include a larger population of African-Americans with clear cell RCC limits our ability to explore the genomic basis for racial disparities. With a higher incidence of pRCC in African-Americans, the papillary kidney cancer TCGA cohort was able to include a larger number of African-Americans. However, despite the available data, there has not been a thorough analysis of somatic driver alterations or germline risk variants more prevalent in African-American kidney cancer. We propose to complete the TCGA analysis of the top two subtypes of kidney cancer, papillary and clear cell, by analyzing an additional cohort of African-Americans with ccRCC. By performing whole genome sequencing on this additional cohort of samples, we will have an adequate number of cases to allow balenced comparisons between African-American and Caucasian clear cell and papillary kidney cancers.

**C-1-b 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 bio-specimens are stored at -80˚C and are available for immediate analysis. For the purpose of completion of the TCGA dataset, we will utilize a consecutive series of 15 African-American subjects with ccRCC from 2013-2015. DNA will be extracted from fresh tumor tissue and whole blood using an automated Maxwell 16® System (Promega, Madison, WI).

**C-1-c WGS[[try to get consistent use of abbrev]] and variant calling**: Sequencing of the normal and tumor sample will performed using Illumina’s Hiseq 2000 technology. In brief, DNA fragments from each sample will be hybridized using HiSeq Paired-End cluster Kits and will be further amplified using the Illumina cBOT. Paired–end libraries will be generated by utilizing HiSeq (2x101) cycle and imaging will be performed by TruSeq kits.

We have extensive experience in large-scale variant calling and interpretation through 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 were developed, deployed and interpreted [18, 19, 20]. We have already developed a prototype pipeline for calling germline and somatic variants. We will use the Genome Analysis Toolkit (GATK)[21] to call germline SNPs and INDELS. We use parmaters consistent with those used in TCGA.[[add ref]] 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 [21] 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 and allele information) and the probability of it being a true positive instead of a sequencing artifact. Furthermore, we will exclude any filtered variant, which falls in a low mappability region of the genome. In addition, we will utilize MuTect [22] and Strelka [23] to call somatic SNVs and INDELs, respectively.

Structural variations (SVs) are important contributors to human polymorphism, have great functional impact and are often implicated in various diseases including cancer. We have developed a number of SV calling algorithms, including BreakSeq [24], which compares raw reads with a breakpoint library (junction mapping) , CNVnator, which measures read depth[25], AGE, which refines local alignment [26], and PEMer, which uses paired ends [27]. We have also developed array-based approaches [28] and a sequencing-based Bayesian model [29]. Furthermore, we have intensively studied the distinct features of SVs originated from different mechanisms. This indicates specific creation processes and potentially divergent functional impacts [30, 31]. 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 [32] 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.

**C-1-d Deliverables:** In this aim, we will generate an extensive catalogue of germline and somatic variants including SNPs, INDELs and large SVs for African-American ccRCC cases. This will be done consistently with the methodology already used in the TCGA, so this catalogue can be used conjunction with TCGA kidney cancer genomic variant datasets to serve as an excellent comparison for the identification of genomic aberrations, associated with racial disparity observed in the emergence of kidney cancer. We plan to make our sequencing data available via dbGAP (see data dissemination plan).

**Aim 2: To assemble a set of coding and non-coding regions associated with kidney cancer, both in terms of somatic and germline alterations.**

**C-2-a Rationale**: In this study, we aim to discover underlying genetic regions that explain racial disparity in RCC. However, due to the limited size of sequenced samples, it is not feasible to test every single region in the genome. In fact, as we discuss later, we have to limit our search space to achieve sufficient statistical power. Therefore, we will first 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 RCC and cancer genomics into this study. This allows us to decrease the number of tests, avoiding losing statistical power.

**C-2-b Relevant Preliminary Results**

D-2-b-1 We have developed ways of prioritizing high-functional impact variants: We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets [33, 34, 35]. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations [33]. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region[36]. Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery) [37, 38]. In recent studies [30, 39], we have integrated and extended these methods to develop a prioritization pipeline called FunSeq. 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). 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 specific disruptiveness to TF binding sites, annotating 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 [30]. We have also applied our method to investigate non-coding mutation patterns in subtypes of gastric cancer [40]. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group. [[did we def'n pcawg & icgc???]]

We have also used allelic variability to prioritize regions of the genome. That is we prioritize regions that differ in functional genomic response, for example, allele-specific expression and binding, between the maternal and paternal alleles. Our variant analysis work includes AlleleSeq [41], a computational pipeline to identify allele-specific events, and AlleleDB, our database connecting single nucleotide variants with allele-specific binding and expression [42].

C-2-b-2 We have developed tools for somatic and germline burden tests: We have worked on statistical methods for analysis of non-coding regulatory regions. LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichments in noncoding elements by comparing observed mutation counts with expected counts under a whole genome background mutation model. LARVA includes corrections for biases in mutation rate owing to DNA replication timing. LARVA can also be used in a mode exclusively on coding regions to prioritize genes. 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 [43].

C-2-b-3 We have already identified some regions associated with kidney cancer through our involvement in the papillary TCGA team: Due to Yale’s expertise in the clinical management and genetics of kidney cancer, we were invited to participate in the various TCGA kidney cancer projects. Our role in the TCGA KICH (chromophobe RCC) included being the manuscript coordinator for the Cancer Cell manuscript. Next for the TCGA KIRP (pRCC) our team performed the analysis of the whole genome sequencing, now published in New England Journal of Medicine [44], providing us with further experience with the available RCC genomic datasets. Finally our team has participated in two ongoing pan-RCC manuscripts serving a central role assessing evaluating the cluster of clusters’ (Cluster of cluster assignments -- COCA) immunologic profile from gene and miRNA expression datasets. Together with other published results on RCC [45, 46, 47, 48, 49], we already have the ability assembled an extensive 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.

# **C-2-c Research plan**

C-2-c-1 Collected database of what is known for kidney cancer: First, we will mine the literature and condense results from previous studies, alluded to above. We will gather genetic changes that include but not limited to: single nucleotide variation (SNP), structural variation/copy number variation (SV/CNV), and mutation process signature. Our study will take a comprehensive approach of the entire genome, sweeping a larger pool to unearth genomic regions for racial disparity in RCC. Last, we will also curate known regions that have discovered in studies on Caucasian and African-American disparity in other types of cancer, for example, prostate cancer [50, 51]. That is, we will add to our collection of kidney-cancer prioritized regions, regions known to be associated with racial disparities in other types of cancer.[[keep??]] Studies have pointed out that RCC is uniquely characterized by copy number variations (RCC) as early and major driver event [45]. Repeats are triggering factors for many structural variation events. Therefore in the germline analysis below, we will also pay attention to repeats polymorphisms around pre-eminent cancer associated genes and recurrent CNV regions in RCC. Particularly, we assume excessive repeats put certain RCC related genes predisposed to CNV events.

C-2-c-2 We will extend FunSeq to find connected modules of elements: FunSeq has already had a limited way of connecting non-coding elements to target genes. It exploits locality of promoters and correlates epigenetic markers on distal regulatory elements with gene expression. Here will extend this capability to develop modules of elements. Genetic modules extend high impact regions by linking them with other genomic elements according to physical interaction, epigenetic marker and expression correlation, molecule pathway/network and other evidence. Elements in the same genetic module are expected to play similar roles in ccRCC and pRCC initialization and development. In the end, we will integrate this new feature into FunSeq and use to assemble genetic modules. Genetic modules group potentially impactful elements that share similar or collaborative biological functions, increasing statistical power in our study. Last, genetic modules offer annotation to less known noncoding regions. Our results will be more biologically interpretable because these regions will be linked with genes.

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

C-2-c-3 We will extend LARVA to include additional covariates: It is known that various genomic features affect background mutation rate in most cancer types, which results in numerous false positives in somatic mutation recurrence analysis [52]. Hence, we have been working on an update to LARVA, which incorporates corrections for additional covariates that influence the somatic mutation rate in different genomic regions, including sequence content, replication timing, expression level, histone modification marks, and chromatin status. 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.

C-2-c-4 We will run our updated and extended FunSeq & LARVA on WGS sequences from TCGA and aim 1: We expect many changes in noncoding regions play a critical role in renal cell cancer. In order to find high impact mutations in noncoding regions, we will run our updated and extended FunSeq and LARVA on variation calls from TCGA whole genome sequenced samples as well as our newly sequenced samples, both cancer and normal. In a first pass run, we have already run FunSeq on 32 whole genome sequenced samples from the TCGA KIRP group. We have found several disruptive mutation hot spots in the genome. We also have found excessive somatic mutations in MET intronic and promoter regions, along with several other recurrent mutated regions that merit further investigation.[[STL ok??]]

C-2-c-5 We will identify critical regions burdened by germline mutations: First, we will try to find regions that are significantly burdened by germline mutations in the kidney cancer cases versus healthy people. As a non-cancer control, we will use both the 1000 Genome Project (2504 people) for the whole genome as controls as well as the Exome Aggregation Consortium (ExAC, non cancer) for the exome[[refs]]. We will look for regions and genes that are burdened significantly in RCC, compared to the control. Given the size of the datasets, we will be well powered in our tests (comparing numbers discussed later in aims 3 and 4). We will also prioritize regions that are less significantly associated with RCC that are known to have racial disparities in healthy people.

**C-2-d Deliverables**: We aim to generate a list of regions in genome that, to our best knowledge, potentially have the highest impacts on the development of RCC. We will also 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). In Aim 3, we will directly test those elements on our samples.

**Aim 3: To compare genomic alterations that differs between African-Americans and Caucasians with kidney cancer**

## C-3-a Rationale: In this aim, we are going to take the genomic regions and modules that we have developed in Aim 2 and test for any evidence of racial disparity. Our overall intention is to investigate differences in the occurrence of common SNPs or differential burdening in terms of rare germline or somatic mutations between African-American and Caucasians. Our specific goal for this aim is to score and prioritize these genomic regions in order to select 550 regions with the highest score to be validated in Aim 4 using a larger cohort.

**C-3-b Compare the germline mutations in coding regions between Caucasian and African-Americans in the prioritized regions using WES Data**

C-3-b-1 Variant level analysis: For the coding region analysis, we will utilize the full category of 556 samples with whole exome data analysis from TCGA. For the common variants analysis at a single locus, Fisher’s exact test can be used to evaluate the racial disparity between Caucasian and African-American subjects with RCC. Here, we prioritize common variants according to their associations with RCC disparity in race. For a common SNP identified in African Americans and Caucasians with RCC, we record the minor allele frequencies and major allele frequencies in African Americans and Caucasians with RCC, respectively. For these counts of the focal SNP, Fisher-exact test is used to determine whether the SNP tends to be associated with the African Americans with RCC. The p-values of the tests for all the common variants are used to prioritize them for further study and validation. Moreover, the power of the Fisher[[cap on Fisher?]] exact test can readily be estimated in this context. For instance, for an ordinary SNP with allele frequency 7% in the total samples, when its frequency in the African American subjects is 12%, the power of the test can reach 0.4 with a p-value < 5e-5. This indicates that the such a SNP can be detected with statistical significance from 1000 candidates, even when a Bonferroni correction is used. Therefore, It is likely possible to detect common SNPs and rare SNPs with high effect sizes from SNP candidates.[[why rare is last sentence??, check this section]]

C-3-b-2 Region based analysis: Beyond investigating the association between the single common variant and race, we will focus on the evaluating the cumulative effects of a set of rare variants in certain genomic regions, such as genes, using both burden and non burden test. Burden test are often applied on regions where most of the variants in the same region are causal and their effect on the phenotype are on the same direction.  We assume that in total there are n patients with whole exome sequencing data available. Also for a target region, for example, a gene, there are m variants. Let denote the population information of the patient. for African-Americans and 0 otherwise. Let represent the genotype of patient . Then a logistic regression model can be set up to evaluate the association as in (1). Suppose that describes the mean of the population status, then



  (1). [[can inline this]]



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



               (2).



Then under the null hypothesis that there is no association of variants in this region with race, the coefficient should be zero. So the test statistic for H0: should be [[H sub 0?]]



  (3).[[explain??]]



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 quite heterogenous, burden tests, such as equation (3), may lose statistical power. Here, sequence kernel association test (SKAT) can be used. Instead of assuming a weighted coefficient effect in the burden test, s are treated as independent random variables with 0 mean and variance . Then the null hypothesis can be changed to H0: . Then the test statistic under equation (1) can be written into:



(4). [[can inline this]]



In (4), is the kernal matrix, and is the genotype information vector. is the weight matrix which can choose based on allele frequency or external information, such as conservation score. The test statistic in (4) can be rewritten into



(5).



In the coding variant analysis, because for most genes, we do not know which of the two cases each gene falls into, a unified test can be used as the following

(6)



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

(7)



C-3-b-3 Power analysis using SKAT for per region based analysis: In the above, we are planning to use aggregated burden tests (i.e. SKAT) to look for differential burdening between populations and use this to rank the regions. While we are not striving for absolute statistical significance in the differential burdening, we do think that the sample size is enough to get an appreciable signal for ranking. Here, we discuss the power aspects of the burden tests in detail and substantiate. SKAT analysis has been developed for rare genomic mutations and remains robust even for common variants. We will utilize SKAT to identify genomic regions with significant variant disparity in kidney cancer between Caucasian and African-American populations. To estimate the sample size that we need to use in order to obtain statistical power, we used the SKAT package from R project, running several population models with different parameters (FIG XXX2). In the proposed study, we will focus on genomic modules linked with kidney cancer; therefore we expect a greater number of effective mutations. In Table XXX2, we present the effective sample size based on different values of causal and negative percentage.

**C-3-c Compare the germline mutations in noncoding regions between Caucasian and African-Americans in the prioritized regions using WGS Data**

C-3-c-1 Pooled variant test for limited target regions: For the noncoding region analysis, since we only have limited power with 32 WGS samples in both populations, target regions instead of the whole genome wide analysis will be carried out on only a small set of regions. From our experience with TCGA KIRP, we already prioritized MET intronic and promoter regions, along with several other recurrent mutated regions that merit further investigation. We will only focus on these selected regions to use the unified test mentioned above (in section D-3-b).[[check numbering]]

D-3-c-2 Non-parametric test for FunSeq score distribution difference: We suspect that the casual regions may not only be under differential mutational burden between races, but may also be overly affected by high-impact mutations. Thus, for the prioritized regions given above, we plan to calculate all the FunSeq scores on both African-American and Caucasian populations. Subsequently, by ranking and pairing the scores between the two population groups we intend to use the Wilcoxon signed-rank test to evaluate the significance of the 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 the population size increases, a Z-score can then be calculated.

**D-3-d Compare the somatic mutations between Caucasian and African-Americans in the prioritized regions:** Previously, we developed an integrative framework LARVA to discover the highly recurrent regions in cancer genomes as candidates for drivers [43]. We will further develop our method by correcting many other genomic features for more accurate background mutation rate calculation. Specifically, in a region with length , suppose the mutation rate is known as , then the number of mutations y within given  should follow a Poisson distribution as the following:



   (8)



However, we discovered in our previous analysis that there is great cancer type, sample, and regional heterogeneity in the mutation count data [43]. Such mutational heterogeneity violates the constant mutation rate assumption and results in over-dispersion. Hence, instead of supposing  is constant, we set up the following model

  (9)



Then the marginal distribution of Y [[gamma??]] could be expressed as the type I negative binomial distribution

   (10),



where,.



Let , are the genomic covariates to be corrected, such as replication timing, GC content, and chromatin status, we could use the following negative binomial regression to estimate the local mutation rate under the covariant set.



(11)



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

We will apply our new method on the 16 African-American and then the 16 Caucasian WGS samples separately. Highly recurrent regions in each will be reported and compared. Those regions that are unique to either population will be prioritized for detailed validation.

**C-3-e Deliverables**: This aim will create a ranking on the list of genes, non-coding regions and variants from Aim 2 to pass to the validation in Aim 4. We will combine the rankings from the sections above by comparing their corresponding p-values. However, we will keep a minimum number of validations of each category. Also, we plan to make our racial disparity rankings of genes and non-coding regions publicly available from the project web server (see data dissemination plan) . [[see deliverable available]]

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

**C-4-a Rationale**: Typically, traditional GWAS studies require thousands of sequenced genomes to associate genetic variants and disease with confidence. Therefore, aim 2 and 3[[consistent way to refer to aims, funseq, &c]] may not render the necessary statistical support to associate kidney cancer variants with racial disparities. However, in aim II and III, our main intention is to obtain, prioritize and rank ~550 genomic regions using our FUNSEQ2 algorithm [39]. These regions will be further processed and validated in aim 4, by using sequences from Yale’s Genitourinary Biospecimen Repository. We intend to validate 550[[exact?? Mention to BRIAN]] regions (100bp each) for 384 individuals.

In particular, we will assemble a independent validation cohort from Yale’s Repository. This will contain both African-Americans and Caucasian with clear cell and papillary RCC to allow comparisons across both histologic type and race. Besides confirming an association with kidney cancer, a large cohort will help us better understanding of how frequently these alterations occur.

**C-4-b Power analysis for the validation cohort:** In the most extreme case, suppose all 550 regions are from common SNPs described in Aim 3. With 384 samples (case-control ratio = 1:1)[[Is this right??]], we can safely detect a SNP, which is of >0.1 frequency and 4(heterozygous)/6(homozygous) risk size, with power >0.8134 and alfa < 9e-5. Then after Bonferroni correction, the corrected P value are considerably less than 0.05. In the other extreme, when the all the prioritized regions are genes after pooled rare SNP tests, we suppose 25 genes with 2kb length. We could still detect regions with risk effect >\*\*\* with this number of samples. [[we need to go over this para]]

[[JX:

In the 550 genomic regions from 384 individuals, here we focus on the SNPs prioritized by the simple method in Aim 3. The parameter space of a hypothetical SNP in the 384 individuals is assumed to test the method’s ability to identify the common SNPs associated with race disparity. The parameters of a SNP are its frequencies in the total 384 individuals (f) and in the 192 African American (fa) and Caucasian (fc) individuals. The parameter f is assumed to vary uniformly from 0.05 to 0.1, instead of the theoretical upper bound 0.5, to be conservative. The fa to fc ratio is assumed to be between 1 and 5. For each hypothetical SNP, the method is used to calculate the significance level (p-value) of the SNP associated with race disparity. Because in reality we need to test the top ranked SNPs in the 550 genomic regions, Bonferroni correction is used to deal with this multiple-test issue. Here, series sets containing 100 to 5000 hypothetical SNPs are generated to estimate the performance of the method with multiple test correction (alpha < 0.05). As a result, the maximum number of SNPs with statistical significance is 51 when the sample size of SNPs is 1000. This analysis roughly estimated the number of top ranked SNPs chosen in real validation.

I am still working on coming up reasonable parameters and settings for the power analysis software used by LS

]]

**C-4-c Sample acquisition and DNA extraction:** As mentioned above, fresh kidney cancer tissue is procured on our IRB-approved Genitourinary Bio-specimen protocol within 30 minutes of removal. Additionally our protocol 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) criteria [53]. 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).

**C-4-d Genotyping kidney cancer non-coding genomic variants:** Frequently, next generation sequencing results find variants that require validation to confirm significance. We will employ similar methods to various other studies involving novel variants found on either exome or whole genome sequencing [54, 55, 56, 57] These studies utilize the MassArray system (Agena Biosciences, San Diego, CA), a mass spectrometry platform that measures PCR-derived amplicons. The system can be multiplexed to analyze a large number of alterations with high sensitivity but a low cost. DNA from FFPE and fresh tissue can be analyzed on the same chip without difficulty. The MassArray system has been utilized to study genomes in the study of cancers, benign conditions, and even ecologic research. This approach has been used in the screening of large case-control series and independent validation cohorts of affected individuals. For the analysis, 20 ng of DNA from genomic control will be used in Yale’s Validation cohort. The 384-subject cohort will have their germline DNA assessed for non-coding variants identified. The MassArray Assay Design Suite will be used for designing custom PCR primers to detect potential germline variants using the genomic coordinates of interest. For genotyping, up to 40 genomic variants can be multiplexed per well per silicon chip. 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.

**D-4-e Tumor profiling for somatic non-coding mutations:** The MassArray system is frequently used for the rapid detection of known or suspected somatic alterations important in cancer [58, 59]. Panels exist to detect common alterations in specific cancers and are employed for testing at various clinical laboratories. We will perform somatic mutational profiling using the MassArray System that allows multiplexing for up to 15 somatic variants per 384-well chip. The technology can detect variants with as low as 1% mutant allele frequency using a small DNA quantity. 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 assessed in the validation cohort. For somatic variants, the MassArray Assay Design Suite will be used for designing custom PCR primers. Similar to above, mutation calls will be assessed using the MassArray Typer 3.4 Analyzer.

**XXX Deliverables**: We plan to make the results from the validation available in conjunction with the ranked list of regions on the project website (see data dissemination plan).

[[thoughts???]]

**We didn’t put any Potential Pitfalls and Alternative Stategies In: This is something the reviewers like…..**

**Perhaps a final conclusion?**

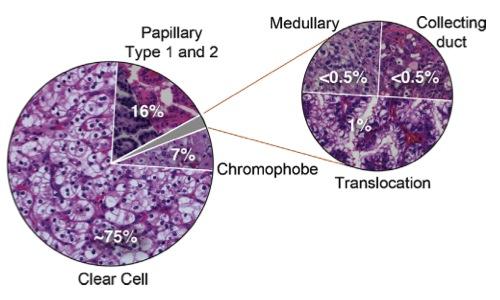


Fig1 : Histologic distribution of kidney cancer ([60])

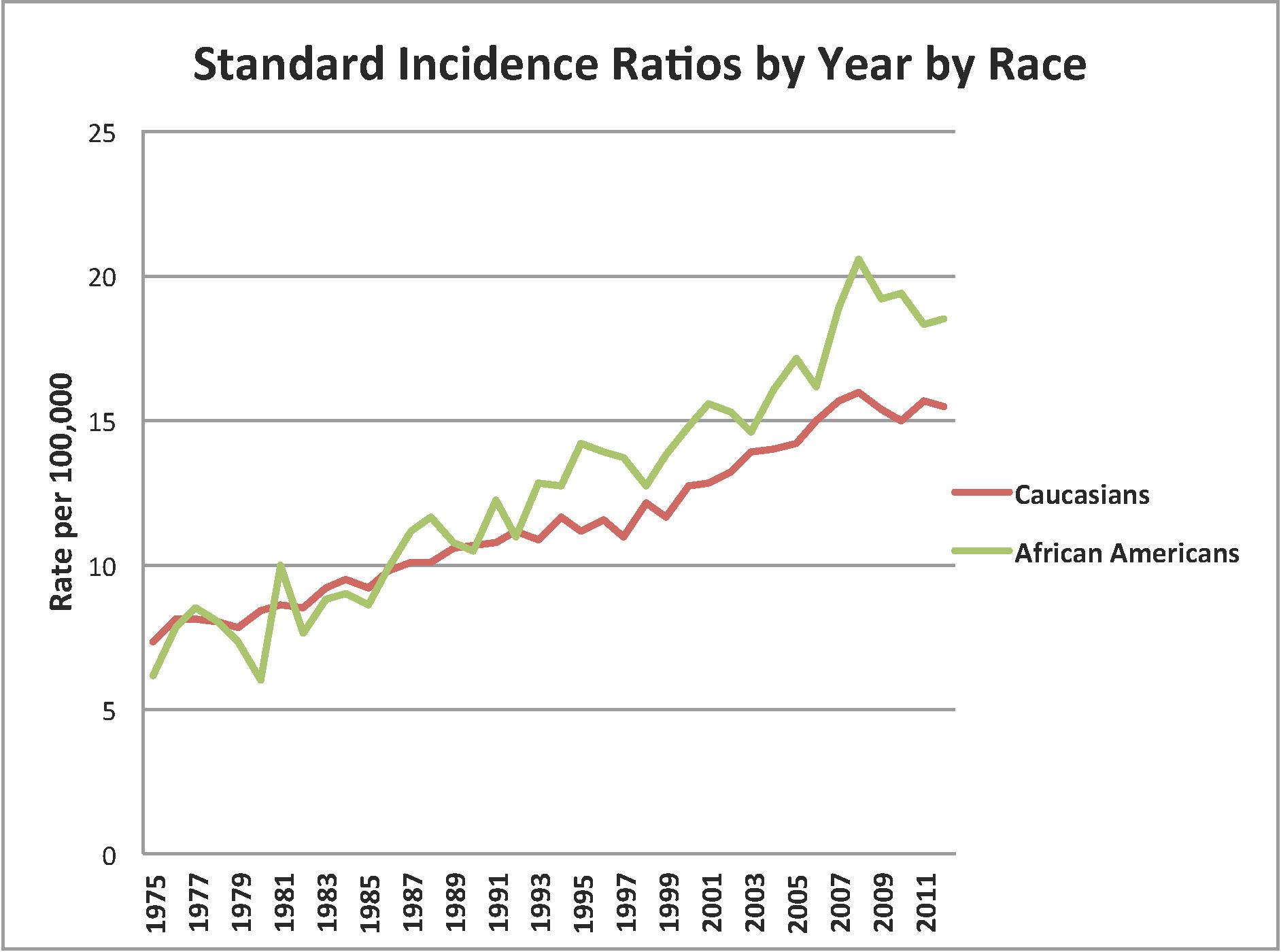


Fig2: Standardized incidence ratios of cancer of the kidney and renal pelvis for Caucasians (Green) and African Americans (Red). Data from the Surveillance Epidemiology and End Result program from 1975-2011, http://seer.cancer.gov/data/)

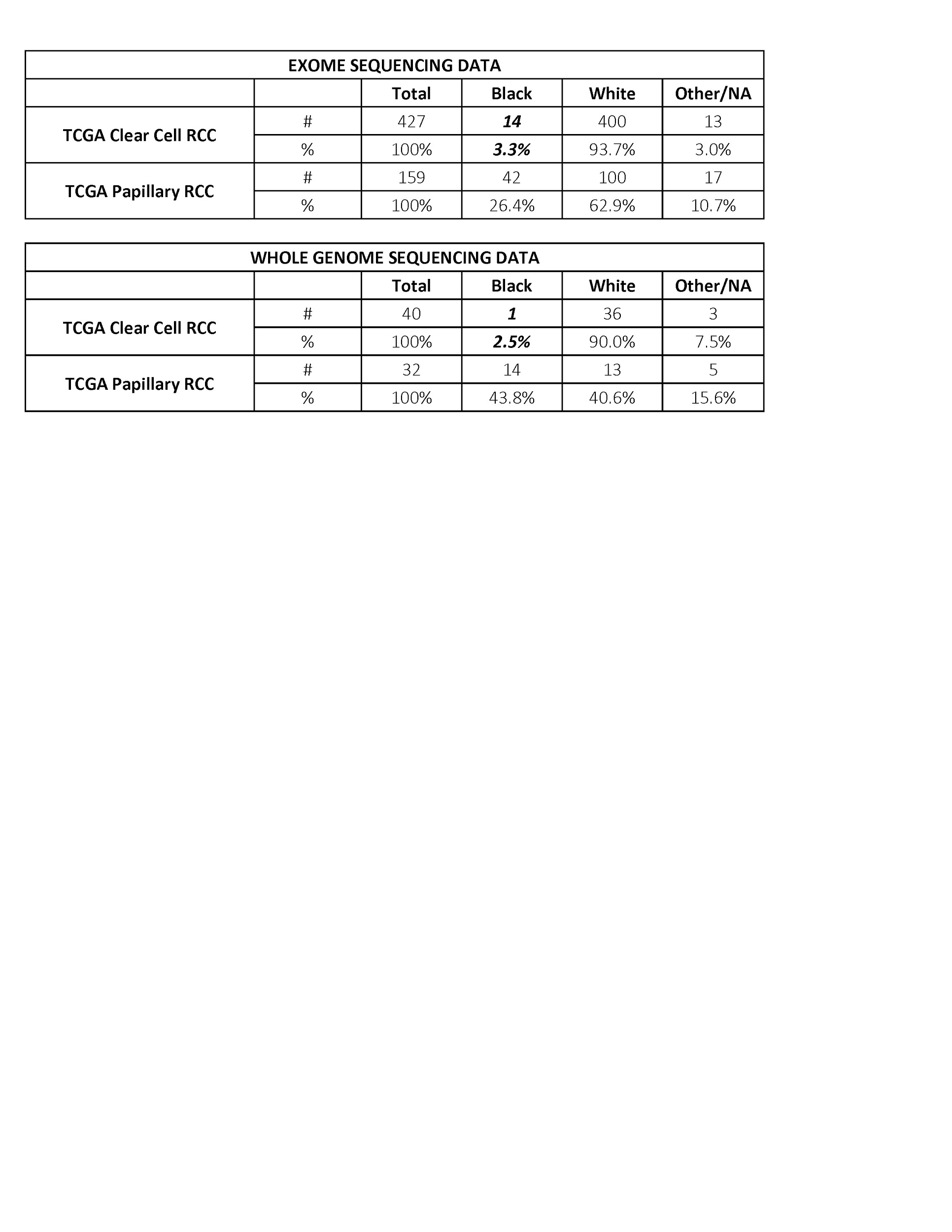
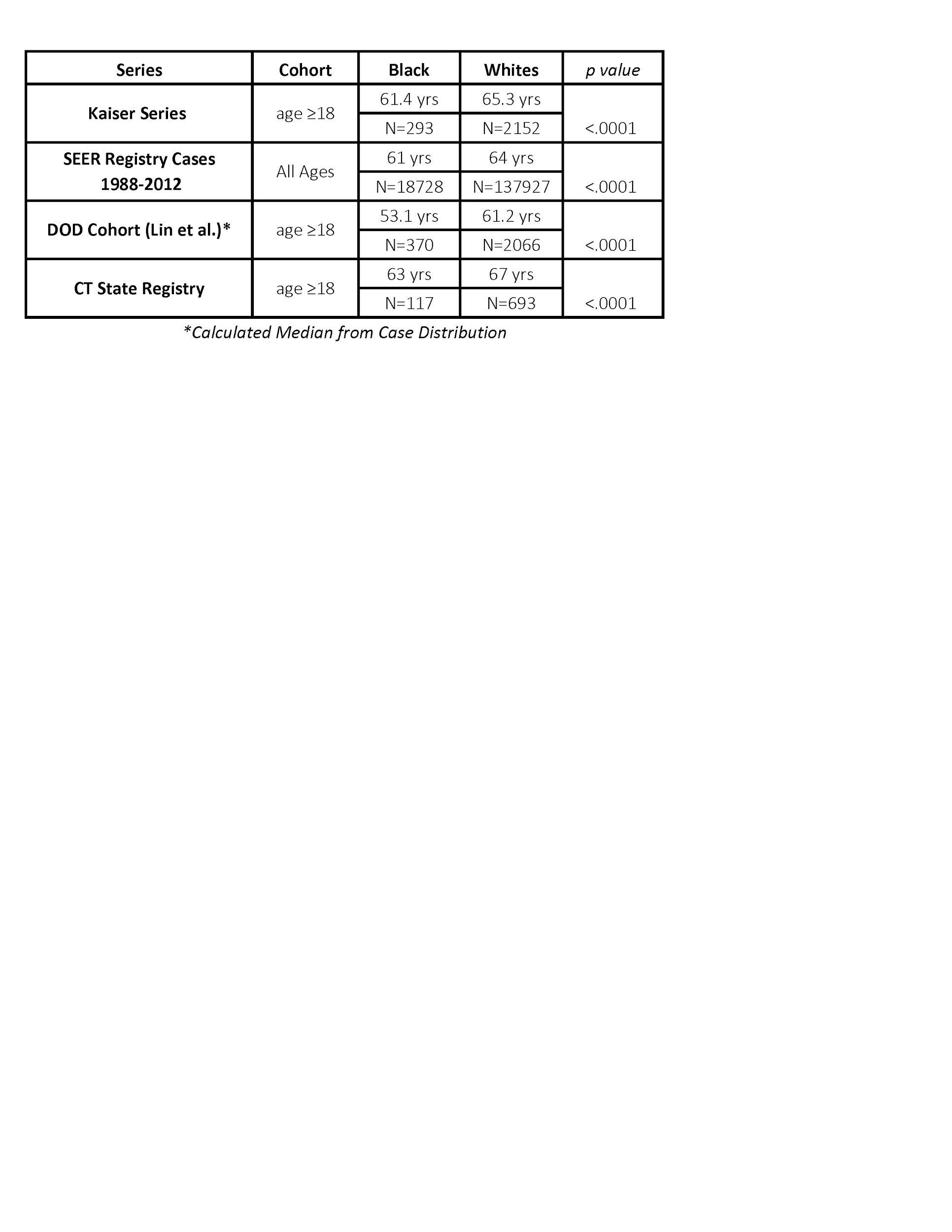


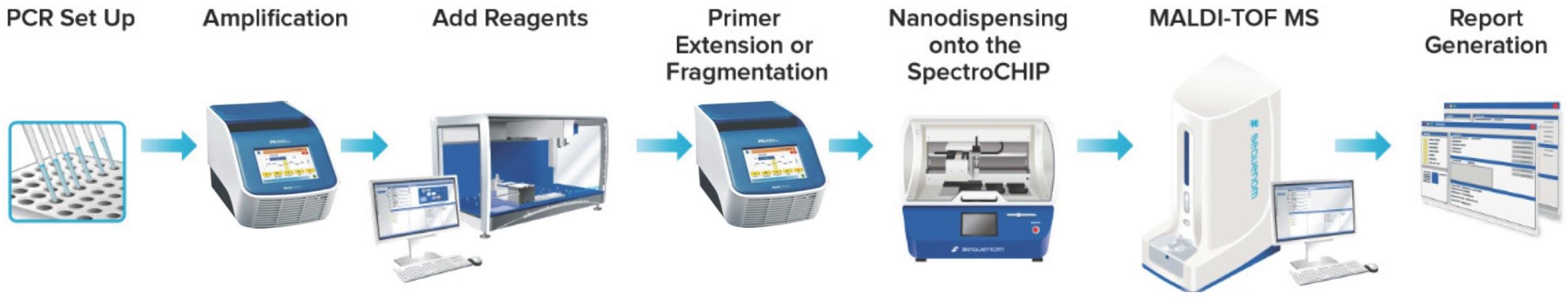
Table 1: Racial and histologic distribution of available whole exome and whole genome data available from TCGA datasets



# **Table 2: Age distribution of kidney cancer by race from a prior series [61] and ongoing work from the Yale Kidney Cancer group.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Service** | **Units** | **$/Unit** | **Total** |
| **DNA QC and Sample Shipment** | **10** | **$16** | **$160** |
| **Human Whole Genome Seq, 135Gb (45X) ('Normal' samples)** | **5** | **$2,100** | **$10,500** |
| **Human Whole Genome Seq, 135Gb (60X) ('Tumor' samples)** | **5** | **$2,800** | **$14,000** |
| **Data handling fee (fastq only) 45X** | **5** | **$24** | **$120** |
| **Data handling fee (fastq only) 60X** | **5** | **$30** | **$150** |
| **Total** |  |  | **$24,930** |

Table3: Sequencing cost breakdown for 5 tumor/normal pair



# **Figure 3: Workflow for validation of whole genome sequencing findings using Yale cohort of tissue. DNA from formalin-fixed, paraffin embedded tumor tissue and also genomic controls will be amplified, have primer extension, mass spectroscopy detection, and analysis. Germline and somatic, coding and non-coding variants will be validated with a large Yale patient cohort.**

References

1. Siegel, R. L., Miller, K. D., & Jemal, A. (2015) Cancer statistics, 2015. *CA Cancer J Clin* 65, 5-29, PMCID: 25559415.2. 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, 1628-31, PMCID: 10235157.3. Howlader N, Noone AM, K. M. G. J. M. D. A. S. K. C. Y. M. R. J. T. Z. M. A. L. D. C. H. F. E. C. K. (2014) http://seer.cancer.gov/csr/1975\_2012/. , , PMCID: .4. Chow, W.-H., Shuch, B., Linehan, W. M., & Devesa, S. S. (2013) Racial disparity in renal cell carcinoma patient survival according to demographic and clinical characteristics. *Cancer* 119, 388-94, PMCID: 23147245.5. 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, 3589-95, PMCID: 17704406.6. Olshan, A. F., Kuo, T.-M., Meyer, A.-M., Nielsen, M. E., Purdue, M. P., & Rathmell, W. K. (2013) Racial difference in histologic subtype of renal cell carcinoma. *Cancer Med* 2, 744-9, PMCID: 24403240.7. 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, 1704-8, PMCID: 18343443.8. Vaishampayan, U. N., Do, H., Hussain, M., & Schwartz, K. (2003) Racial disparity in incidence patterns and outcome of kidney cancer. *Urology* 62, 1012-7, PMCID: 14665346.9. Shuch, B., Vourganti, S., Ricketts, C. J., Middleton, L., Peterson, J., Merino, M. J., Metwalli, A. R., Srinivasan, R., & Linehan, W. M. (2014) Defining early-onset kidney cancer: implications for germline and somatic mutation testing and clinical management. *J Clin Oncol* 32, 431-7, PMCID: 24378414.10. Hofmann, J. N., Schwartz, K., Chow, W.-H., Ruterbusch, J. J., Shuch, B. M., Karami, S., Rothman, N., Wacholder, S., Graubard, B. I., Colt, J. S., & Purdue, M. P. (2013) The association between chronic renal failure and renal cell carcinoma may differ between black and white Americans. *Cancer Causes Control* 24, 167-74, PMCID: 23179659.11. Purdue, M. P., Moore, L. E., Merino, M. J., Boffetta, P., Colt, J. S., Schwartz, K. L., Bencko, V., Davis, F. G., Graubard, B. I., Janout, V., Ruterbusch, J. J., Beebe-Dimmer, J., Cote, M. L., Shuch, B., Mates, D., Hofmann, J. N., Foretova, L., Rothman, N., Szeszenia-Dabrowska, N., Matveev, V., Wacholder, S., Zaridze, D., Linehan, W. M., Brennan, P., & Chow, W.-H. (2013) An investigation of risk factors for renal cell carcinoma by histologic subtype in two case-control studies. *Int J Cancer* 132, 2640-7, PMCID: 23150424.12. Sankin, A., Cohen, J., Wang, H., Macchia, R. J., & Karanikolas, N. (2011) Rate of renal cell carcinoma subtypes in different races. *Int Braz J Urol* 37, 29-32;discussion 33-4, PMCID: 21385477.13. Saigal, C. S., Deibert, C. M., Lai, J., & Schonlau, M. (2010) Disparities in the treatment of patients with IL-2 for metastatic renal cell carcinoma. *Urol Oncol* 28, 308-13, PMCID: 19070518.14. Tsao, C.-K., Small, A. C., Moshier, E. L., Gartrell, B. A., Wisnivesky, J. P., Sonpavde, G., Godbold, J. H., Palese, M. A., Hall, S. J., Oh, W. K., & Galsky, M. D. (2012) Trends in the use of cytoreductive nephrectomy in the United States. *Clin Genitourin Cancer* 10, 159-63, PMCID: 22651971.15. Shuch, B., Mikhail, M., Satagopan, J., Lee, P., Yee, H., Chang, C., Cordon-Cardo, C., Taneja, S. S., & Osman, I. (2004) Racial disparity of epidermal growth factor receptor expression in prostate cancer. *J Clin Oncol* 22, 4725-9, PMCID: 15570072.16. Lipworth, L., McLaughlin, J. K., Tarone, R. E., & Blot, W. J. (2011) Renal cancer paradox: higher incidence but not higher mortality among African-Americans. *Eur J Cancer Prev* 20, 331-3, PMCID: 21633203.17. Lipworth, L., Tarone, R. E., & McLaughlin, J. K. (2011) Renal cell cancer among African Americans: an epidemiologic review. *BMC Cancer* 11, 133, PMCID: 21486465.18. 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, 59-65, PMCID: 21293372.19. 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, 1061-73, PMCID: 20981092.20. 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, 56-65, PMCID: 23128226.21. 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, 491-8, PMCID: 21478889.22. 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, 213-9, PMCID: 23396013.23. 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, 1811-7, PMCID: 22581179.24. 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, 47-55, PMCID: 20037582.25. 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, 974-84, PMCID: 21324876.26. Abyzov, A. & Gerstein, M. (2011) AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* 27, 595-603, PMCID: 21233167.27. 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, R23, PMCID: 19236709.28. Wang, L.-Y., Abyzov, A., Korbel, J. O., Snyder, M., & Gerstein, M. (2009) MSB: a mean-shift-based approach for the analysis of structural variation in the genome. *Genome Res* 19, 106-17, PMCID: 19037015.29. Zhang, Z. D. & Gerstein, M. B. (2010) Detection of copy number variation from array intensity and sequencing read depth using a stepwise Bayesian model. *BMC Bioinformatics* 11, 539, PMCID: 21034510.30. 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., Gümüs, 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, 1235587, PMCID: 24092746.31. 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, PMCID: 26028266.32. 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, 652-4, PMCID: 21666668.33. 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, 7058-76, PMCID: 21596777.34. Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., & Gerstein, M. (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* 13, R48, PMCID: 22950945.35. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M., & Snyder, M. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91-100, PMCID: 22955619.36. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74, PMCID: 22955616.37. Kim, P. M., Korbel, J. O., & Gerstein, M. B. (2007) Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* 104, 20274-9, PMCID: 18077332.38. Khurana, E., Fu, Y., Chen, J., & Gerstein, M. (2013) Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* 9, e1002886, PMCID: 23505346.39. Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X. J., Yip, K. Y., Khurana, E., & Gerstein, M. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* 15, 480, PMCID: 25273974.40. () submitted. , , PMCID: .41. Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., Bhardwaj, N., Rubin, M., Snyder, M., & Gerstein, M. (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 7, 522, PMCID: 21811232.42. et. al., J. (2015) Allele-specific binding and expression: a uniform survey over many assays and 1000-Genomes-Project individuals . *Submitted* , , PMCID: .43. 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, 8123-34, PMCID: 26304545.44. 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., Shaw, K. R. M., 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., & Cancer Genome Atlas Research Network (2015) Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N Engl J Med* , , PMCID: 26536169.45. 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, PMCID: 25790038.46. 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, 13-21, PMCID: 25401301.47. Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499, 43-9, PMCID: 23792563.48. 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, 860-7, PMCID: 23797736.49. Christinat, Y. & Krek, W. (2015) Integrated genomic analysis identifies subclasses and prognosis signatures of kidney cancer. *Oncotarget* 6, 10521-31, PMCID: 25826081.50. Freedman, M. L., Haiman, C. A., Patterson, N., McDonald, G. J., Tandon, A., Waliszewska, A., Penney, K., Steen, R. G., Ardlie, K., John, E. M., Oakley-Girvan, I., Whittemore, A. S., Cooney, K. A., Ingles, S. A., Altshuler, D., Henderson, B. E., & Reich, D. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci U S A* 103, 14068-73, PMCID: 16945910.51. Haiman, C. A., Patterson, N., Freedman, M. L., Myers, S. R., Pike, M. C., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., McDonald, G. J., Greenway, S. C., Stram, D. O., Le Marchand, L., Kolonel, L. N., Frasco, M., Wong, D., Pooler, L. C., Ardlie, K., Oakley-Girvan, I., Whittemore, A. S., Cooney, K. A., John, E. M., Ingles, S. A., Altshuler, D., Henderson, B. E., & Reich, D. (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* 39, 638-44, PMCID: 17401364.52. 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., Lee, R. S., 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, 214-8, PMCID: 23770567.53. 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, 1469-89, PMCID: 24025519.54. Buggs, R. J. A., Chamala, S., Wu, W., Gao, L., May, G. D., Schnable, P. S., Soltis, D. E., Soltis, P. S., & Barbazuk, W. B. (2010) Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Mol Ecol* 19 Suppl 1, 132-46, PMCID: 20331776.55. Ren, L. P., Xian, Y. S., Diao, D. M., Chen, Y., Guo, Q., & Dang, C. X. (2013) Further evidence for the contribution of the BRCA1-interacting protein-terminal helicase 1 (BRIP1) gene in breast cancer susceptibility. *Genet Mol Res* 12, 5793-801, PMCID: 24301948.56. Guan, X., Song, Y., Ott, J., Zhang, Y., Li, C., Xin, T., Li, Z., Gan, Y., Li, J., Zhou, S., & Zhou, Y. (2015) The ADAMTS1 Gene Is Associated with Familial Mandibular Prognathism. *J Dent Res* 94, 1196-201, PMCID: 26124221.57. Zhao, L., Zhang, Z., Lin, J., Cao, L., He, B., Han, S., & Zhang, X. (2015) Complement receptor 1 genetic variants contribute to the susceptibility to gastric cancer in chinese population. *J Cancer* 6, 525-30, PMCID: 26000043.58. Quinn, A. M., Hickson, N., Adaway, M., Priest, L., Jaeger, E., Udar, N., Keeling, C., Kamieniorz, M., Dive, C., Wallace, A., Byers, R. J., Newman, W. G., Nonaka, D., & Blackhall, F. H. (2015) Diagnostic Mutation Profiling and Validation of Non-Small-Cell Lung Cancer Small Biopsy Samples using a High Throughput Platform. *J Thorac Oncol* 10, 784-92, PMCID: 25634010.59. Barresi, V., Bonetti, L. R., & Bettelli, S. (2015) KRAS, NRAS, BRAF mutations and high counts of poorly differentiated clusters of neoplastic cells in colorectal cancer: observational analysis of 175 cases. *Pathology* 47, 551-6, PMCID: 26352110.60. Shuch, B., Amin, A., Armstrong, A. J., Eble, J. N., Ficarra, V., Lopez-Beltran, A., Martignoni, G., Rini, B. I., & Kutikov, A. (2015) Understanding pathologic variants of renal cell carcinoma: distilling therapeutic opportunities from biologic complexity. *Eur Urol* 67, 85-97, PMCID: 24857407.61. Lin, J., Zahm, S. H., Shriver, C. D., Purdue, M., McGlynn, K. A., & Zhu, K. (2015) Survival among Black and White patients with renal cell carcinoma in an equal-access health care system. *Cancer Causes Control* 26, 1019-26, PMCID: 25956269.