**Whole Genome Sequencing and variant calling**

Sequencing of the normal and tumor sample will performed using Illumina’s Hiseq 2000 or 2500?X 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 \cite{21293372,20981092,23128226}. We have already developed a prototype pipeline for calling germline and somatic variants. We will use the 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 \cite{21478889}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 mappability 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.

**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\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). 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 to 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.

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 (Allelic activity). Our variant analysis work includes AlleleSeq \cite{21811232}, a computational pipeline to identify allele-specific events, and AlleleDB, our database connecting single nucleotide variants with allele-specific binding and expression.

**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 also 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. \cite{26304545).

**We will run FunSeq & LARVA on WGS sequences from TCGA and aim 1**

In order to examine noncoding regions in the RCC genome, we have run FunSeq on 32 whole genome sequenced samples from the TCGA KIRP group. We have found several disruptive mutation hot spots in the genome. In initial runs, we have found excessive mutations in MET intronic and promoter regions, along with several other recurrent mutated regions that merit further investigation. 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 FunSeq and LARVA on variation calls from TCGA whole genome sequenced samples as well as our newly sequenced samples, both cancer and normal. Using existing somatic mutation process decomposition tools, we will also conduct an inspection on all available renal cell carcinoma sequenced tumor samples.