mRNA Sequencing

Clear Cell Sarcoma of the Kidney (CCSK)

*Protocols were performed at NCI Center for Cancer Research through the laboratory of Dr. Javed Khan.

RNA-seq Library construction and sequencing by Illumina HiSeq2000:

PolyA+ RNA was purified using the MACS mRNA isolation kit (Miltenyi Biotec, Bergisch Gladbach, Germany), from 5-10ug of DNaseI-treated total RNA as per the manufacturer's instructions. Double-stranded cDNA was synthesized from the purified polyA+ RNA using the Superscript Double-Stranded cDNA Synthesis kit (Invitrogen, Carlsbad, CA, USA) and random hexamer primers (Invitrogen) at a concentration of 5µM. The cDNA was fragmented by sonication and a paired-end sequencing library prepared following the Illumina paired-end library preparation protocol (Illumina, Hayward, CA, USA). RNA samples were prepared by Illumina TruSeqRNA Sample Preparation V2 kits according to the manufacturer's protocol. Poly-A containing mRNA was purified using poly-T oligo-attached magnetic beads and then fragmented. RNA fragments of ~200bp were reverse-transcribed and ligated with adaptors for sequencing. RNA libraries were sequenced on Illumina HiSeq2000 using 100bp paired-end sequencing according to the manufacturer's protocol.

Reads Alignment:

Align reads to reference genome (GRCh37) using Tophat version 2.0.8b with default options, expect for options specifying number of processor threads and fusion search. An example code for alignment with fastq files is shown below.

```
-o tophat.out –p 6 --fusion-search –fusion-min-dist 100000 GRCh37 read_1.fq read_2.fq
```

RNA sequencing reads were aligned to GRCh37-lite genome-plus-junctions reference using BWA version 0.5.7. This reference combined genomic sequences in the GRCh37-lite assembly and exon-exon junction sequences whose corresponding coordinates were defined based on annotations of any transcripts in Ensembl (v59), Refseq and known genes from the UCSC genome browser, which was downloaded on August 19 2010, August 8 2010, and August 19 2010, respectively. Reads that mapped to junction regions were then repositioned back to the genome, and were marked with 'ZJ:Z' tags. BWA is run using default parameters, except that the option (-s) is included to disable Smith-Waterman alignment. Finally, reads failing the Illumina chastity filter are flagged with a custom script, and duplicated reads were flagged with Picard Tools.

Gene Coverage Analysis Protocol:

www.bcgsc.ca/downloads/genomes/Homo_sapiens/hg19/1000genomes/bwa_ind/genome/(link is external)
Data Level: 3 Data File: *.gene.quantification.txt

The gene coverage analysis was performed with our internal analysis pipeline version 1.1 using "composite" gene annotations from the hg19 (GRCh37-lite) version of the TCGA GAF v3.0. These composite gene models were created in June 2011 by UNC (with assistance from UCSC) based on the annotations in the "UCSC genes" database. Each composite gene annotation was generated by collapsing all transcripts of that gene into a single model such that exonic bases in a composite gene model were the union of exonic bases from all known transcripts of the gene. Thus, the locations of the exonic boundaries used for the gene coverage analysis were not based on a single canonical transcript for each gene. Consequently, the exonic boundaries in a composite gene model may not correspond to the actual boundaries of the expressed transcripts. For simplicity, throughout this document and in the gene coverage results files, a composite gene model is simply referred to as a gene, and it is associated with the id of the gene whose transcripts contributed to that composite model. To generate the raw read counts, we first counted the number of bases of each read that were inside exonic regions in a gene, and then divided this total base count by the read length. Thus our values for the raw number of reads were not whole numbers (i.e. if the entire 50bp read mapped to an exon, we would add 50 to the total base count, which would ultimately contribute 1 to the raw read count. However, if only 25 bases of the read's alignment fell within an exon's boundaries, the total base count would be incremented by 25, which would ultimately contribute 0.5 to the raw read count). In order to comply with the file format specification enforced by the DCC validator, our raw read counts are rounded to the closest whole number. A gene's raw read count is the sum of raw read counts for exons belonging to the gene. Gene coverage is its raw read count divided by the sum of its exon lengths. RPKM is calculated using the formula: (number of reads mapped to all exons in a gene x 1,000,000,000)/(NORM_TOTAL x sum of the lengths of all exons in the gene)

[Note: NORM_TOTAL = the total number of reads that are mapped to all exons from the composite gene models. (i.e. sum of the fractional read count for all exons)]

If a read alignment contained a deletion or a large gap, the read did not contribute coverage inside the region spanned by the deletion/gap. Each of the paired end reads was counted separately. We excluded reads from pairs that failed Illumina's Chastity filter, as well as reads with mapping quality < 10.

*.gene.quantification.txt: A tab-delimited text file containing the following fields: - gene = Gene ID from GAF (version 3.0). The ID follows the nomenclature '<HUGO gene symbol>|<Entrez ID>'. If the combination of the HUGO symbol and the Entrez ID is not unique, an additional 'NofM' descriptor is added. An ID with '?' indicates that the HUGO gene symbol or Entrez ID is not available. e.g. U80769|?; TRNA_Pseudo|?|8of100 - raw_counts = Sum of fraction of reads (rounded off to nearest integer - restricted by the RNA-seq validator) that mapped to collapsed transcripts representing a specific gene. Reads from pairs that did not pass Illumina’s Chastity filter or with mapping quality less than 10, i.e. reads that did not map uniquely, were excluded from calculation. - median_length_normalized = Average coverage over all exons in the collapsed transcripts i.e. sum of the coverage depth at each base in all exons divided by the sum

of the exon lengths - RPKM = Reads per kilobase of exon per million. Calculation described in detail below.

Exon Coverage Analysis:

Data Level: 3 Data File: **.exon.quantification.txt

The exon coverage analysis was performed with our internal analysis pipeline version 1.1 using "composite" gene annotations from the hg19 (GRCh37-lite) version of the TCGA GAF v3.0. These composite gene models were created in June 2011 by UNC (with assistance from UCSC) based on the annotations in the "UCSC genes" database. Similar to the gene coverage analysis, all transcripts of a given gene were collapsed into a single model such that exonic bases in a composite gene model were the union of exonic bases from all known transcripts of the gene. For simplicity, throughout this document and in the exon coverage results files, the collapsed exons are simply referred to as an exon. To generate the raw read counts, we first counted the number of bases of each read that were inside an exonic region, and then divided this total base count by the read length. Thus our values for the raw number of reads were not whole numbers (i.e. if the entire 50bp read mapped to an exon, we would add 50 to the total base count, which would ultimately contribute 1 to the raw read count. However, if only 25 bases of the read's alignment fell within an exon's boundaries, the total base count would be incremented by 25, which would ultimately contribute 0.5 to the raw read count). In order to comply with the file format specification enforced by the DCC validator, our raw read counts are rounded to the closest whole number. Exon coverage is the raw read count of an exon divided by its length. RPKM is calculated using the formula (number of reads (fractional) mapped to an exon x 1,000,000,000)/(NORM_TOTAL x length of an exon) [Note: NORM_TOTAL = the total number of reads (fractional) that mapped to exons, excluding those in the mitochondrial chromosome] If a read alignment contained a deletion or a large gap, the read did not contribute coverage inside the region spanned by the deletion/gap. Each of the paired end reads was counted separately. We excluded reads from pairs that failed Illumina's Chastity filter, as well as reads with mapping quality < 10.

**.exon.quantification.txt A tab-delimited text file containing the following fields: - exon = Exon coordinates according to GAF (version 3.0) with the nomenclature, chr<chromosome number>:<start coordinate>-<end coordinate>:<strand>. '.' in the <strand> indicates that there was no strand information available. e.g. chr10:120810487-120810613:. - raw_counts = Sum of fraction of reads (rounded off to nearest integer - restricted by the RNA-seq validator) that mapped to an exon. Reads from pairs that did not pass Illumina’s Chastity filter or with mapping quality less than 10 were excluded from calculation. - median_length_normalized = Average coverage over the exon i.e. the sum of the coverage depth at each base in an exon divided by the length of the exon. - RPKM = Reads per kilobase of exon per million.

Gene and isoform expression:

Gene and isoform expression from RNA-seq data was generated using Cufflinks version 2.1.1. with default options and supplied reference annotation (Homo_sapiens.GRCh37.71.gtf) for estimation of expression. Cufflinks will not assemble novel transcripts, and it will ignore alignments not structurally compatible with any reference transcript.

Exon expression:

Exon expression file was generated using dexseq_count.py included in R package DEXseq 1.12.1 with annotation (Homo_sapiens.GRCh37.71.gff) and default parameters except for –p yes (indicates the data is paired end) and –s no (indicates the data is not from a strand-specific assay).

Gene fusion:

Gene fusion file was generated using defuse version 0.6.1 with default parameters and with reference annotation Homo_sapiens.GRCh37.69.