Updated on 05/07/2019

RNA-seq (Cancer) – gene expression and fusion transcripts

Sequencing data will be first checked for quality control and then aligned to the human "indexed" reference genome (hg38) (or corresponding organism) using STAR (Spliced Transcripts Alignment to a Reference) [REF 1]. Gene quantification will be done using HTSeq-count [REF 2]. If the client is specifically looking for long noncoding RNAs (lncRNAs), annotations from GENCODE lncRNAs in alignment (STAR) and gene quantification (HTSeq-count) steps will be used [REF 3]. Gene count data from HTSeq-count will be filtered before analysis to remove lowly expressed genes. For example, a gene is retained if it has at least 10 counts for at least 30% samples. CPM (counts per million) data will be generated after normalization using trimmed method of M-values (TMM), as implemented in the Bioconductor package edgeR [REF 4].

DIFFERENTIAL EXPRESSION: To determine genes important for a case-control (e.g., disease vs. normal) study, we will perform a t-test (or edgeR test) followed by a FDR (BH-method) adjustment on normalized log2-transformed CPM. EdgeR-tests for differential expression between two groups using a method conceptually similar ~~in idea~~ to the Fisher's Exact Test. EdgeR is for a two-group comparison, but if there are additional covariates, a generalized linear model (GLM) framework is utilized. A gene is defined as a differentially expressed gene (DEG) if its FDR is ≤ 0.05.

PATHWAY ENRICHMENT: To characterize expressed genes, a pre-ranked permutation based gene set enrichment analysis (GSEA) will be performed. A gene set from the Molecular Signatures Database (current version) will be used as input data for this analysis [REF 5]. GSEA generates enriched pathways ordered by enrichment score (ES) of the leading-edge genes based on a weighted Kolmogorov–Smirnov (K–S) statistical test, which measures the difference between the number of genes in a given dataset and the number of occurrences if the genes are observed in a pathway [REF 6]. Alternatively, the identified differentially expressed genes will be used in pathway analysis. In this case, the hypergeometric distribution will be used to find the exact probabilities to compute enrichment likelihoods [REF 7].

FUSION GENE/TRANSCRIPT DISCOVERY: Fusion transcripts are characteristic of cancer tumors. STAR-Fusion uses chimeric-reads collected during STAR-alignment for fusion RNA prediction [REF 8]. In order to reduce the number of false-positive fusion genes: fusion events with fusion fragments per million total reads<0.1 and putative fusions between homologous genes will be discarded. We also remove the fusions between mitochondria and autosomes. Fusion transcript plot(s) with protein domain annotations will be created using chimeraviz [REF 9]

Required from client:

1. Raw data files (fastq)
2. Metadata spreadsheet

3. Organism (Human/Mouse)

Deliverables of RNA-seq data analysis service:

1. Hierarchical clustering analysis and box plots of selected genes
2. Differential expression analysis with multiple hypothesis testing
3. Heatmaps (selected genes or a pathway)
4. Enrichment analysis of gene annotations/pathways
5. List of fusion transcripts and fusion plot(s)
6. Analysis and methodology descriptions for manuscript

References:

1. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
2. Anders S, Pyl PT, Huber W. HTSeq - a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31:166–9.
3. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22: 1760–1774.
4. Robinson MD, Smyth GK: Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007, 23: 2881-2887.
5. http://software.broadinstitute.org/gsea/msigdb
6. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550 (2005).
7. Dinasarapu AR, Gupta S, Ram Maurya M, et al. A combined omics study on activated macrophages-enhanced role of STATs in apoptosis, immunity and lipid metabolism. Bioinformatics. 2013; 29:2735–2743.
8. https://github.com/STAR-Fusion/STAR-Fusion/wiki
9. https://github.com/stianlagstad/chimeraviz

Note: For sequencing data acquisition please contact Emory Integrated Genomics Core (EIGC@emory.edu).

Questions? Comments?

Please email us at EICC@emory.edu