Sequencing Platform Modeling and Analysis

- Li-Xuan Qin
Associate Member of Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

- Yen-Tsung Huang
Assistant Professor of Epidemiology and Biostatistics, Brown University, Providence, RI, USA.

Supplement Aims and Scope

Cancer informatics represents a hybrid discipline encompassing the fields of oncology, computer science, bioinformatics, statistics, computational biology, genomics, proteomics, metabolomics, pharmacology, and quantitative epidemiology. The common bond or challenge that unifies the various disciplines is the need to bring order to the massive amounts of data generated by researchers and clinicians attempting to find the underlying causes and effective means of treating cancer.

The future cancer informatician will need to be well-versed in each of these fields and have the appropriate background to leverage the computational, clinical, and basic science resources necessary to understand their data and separate signal from noise. Knowledge of and the communication among these specialty disciplines, acting in unison, will be the key to success as we strive to find answers underlying the complex and often puzzling diseases known as cancer.

This supplement is focused on sequencing platform modeling and analysis, and article topics may include:

- RNA Sequencing
- DNA Sequencing
- ChIP-Seq
- RNA Expression Level
- RNA Slicing Variant
- DNA Copy Number
- DNA Mobile Element
- DNA Transcription Factor Binding
- DNA Histone Modification Marks
- Splicing Quantitative Trait Loci
- Experimental Design
- Normalization Methodology
- Differential Expression Analysis
- Prognostic Biomarker Selection
- Classification and Prognostication
- Functional and Pathways Analysis
- Multi-Dimensional Association Studies
- Multi-Platform Integrative Analysis

As the cost rapidly declines, the next-generation sequencing (NGS) technologies are becoming the main work horse of modern molecular biology to measure the abundance of genomic markers. NGS offers a number of technical advantages such as a greater dynamic range for measuring abundance and a higher resolution for low abundance molecules; moreover, it allows studying innovative biological problems such as the detection of novel molecular features and performing a more seamless integrative analysis of multiple types of molecular data for the same set of samples. Their utilization in cancer genomic studies such as The Cancer Genome Atlas (TCGA) presents unprecedented opportunities to advance our understanding of cancer.

While certain aspects of NGS data analysis fall under the general framework of high-throughput data analysis and can be addressed by existing statistical methodologies, it also presents unique analytic challenges that require development of new statistical approaches. We take the opportunity of this special issue to highlight several recent methodological developments on NGS data analysis by leading researchers in the field. Some of the articles in this special issue are summarized below.

- ***RNA Expression Level***: Identification of molecular signatures is an important aspect of the analysis of RNA expression data. Ha et al developed a novel method in a causal structure learning framework to discover prognostic gene signatures. Their method represented the causal structure by directed acyclic graphs and constructed gene-specific network modules to constitute a gene and its corresponding regulators; the method then correlated each module with a survival outcome to allow for a network oriented approach to select prognostic genes. They applied the new method to a clear cell renal cell carcinoma study from TCGA and found several novel prognostic genes.
- ***RNA Splicing Variant***: Alternative splicing is a post-transcriptional process that allows a single gene to produce multiple mRNA isoforms (namely, mRNA slicing variants). This process can be regulated by genetic elements such as single nucleotide polymorphisms (SNPs), which are called splicing quantitative trait loci (sQTL). The main analytic challenge of identifying sQTL is the estimation of isoform-specific mRNA expression based on RNA sequencing data. Jia et al evaluated three statistical

methods for the analysis of sQTL using both simulated and real datasets. They observed favorable results for one of the methods and discussed possible directions to further improve it.
- ***DNA Histone Modification***: Chromatin immunoprecipitation sequencing (ChIP-seq) is a powerful approach to examining DNA-protein interactions, such as binding of transcription factors (TFs) and marks of histone modifications (HMs). In contrast to ChIP-seq for TFs that is characterized by strong signals for distinctive peaks, ChIP-seq data for HMs typically contain weak signals for multiple local peaks. Wu et al developed a novel statistical framework to effectively model ChIP-seq data for HMs and proposed a novel test statistic to identify differentially histone modified regions between two experimental conditions. Their method was based on nonparametric hypothesis testing and kernel smoothing, and was illustrated using data from an adipogenesis study and the ENCODE study.
- ***Mobile Elements:*** Mobile elements constitute nearly half of the human genome with most elements fixed and inactive within the human population. However, some younger elements such as ALU and LINE are still actively duplicating, which causes structural alterations of the genome and may result in human diseases such as cancer. The NGS data provide a unique resource for identification of such important genomic variations. Lee et al. developed a new bioinformatics algorithm to detect mobile elements insertion in the 1000 Genome Projects, and further extended it to incorporate alignment mapped by any short-read mapper. They applied the new algorithm to TCGA data and demonstrated its advantage in efficiency and accuracy.

NGS is a young technology of high potential. Its full use requires extensive developments of novel statistical methodologies in many areas of data analysis. One particularly important area that needs further development is in the critical preprocessing step of converting raw reads to an accurate measure of the underlying abundance. Highly innovative analytic methods are needed to properly deal with the bias due to mis-alignment, the bias due to differences in gene length and GC content, the dependence on genome annotation, and the requirement of read depth. With continued advances in statistical methodologies, we are optimistic that they will help NGS bear more and better fruits in the understanding of the molecular basis of organisms and diseases such as cancer in the years to come.

Lead Guest Editor **Dr Li-Xuan Qin**

Dr. Li-Xuan Qin is an Associate Member of Biostatistics at Memorial Sloan Kettering Cancer Center. She completed her PhD at the University of Washington. Her current work focuses on the statistical analysis of high-dimensional data for translational cancer research. Dr Qin is the author or co-author of many published papers, and holds several NIH grants as PI or co-investigator. She has been invited to give presentations in national/international research conferences and academic departments.

qinl@mskcc.org
http://www.mskcc.org/research/epidemiology-biostatistics/biostatistics/staff/li-xuan-qin

Guest Editor

DR YEN-TSUNG HUANG

Dr. Yen-Tsung Huang is an Assistant Professor of Epidemiology and Biostatistics at Brown University. He completed his ScD at Harvard University. His research focuses on the incorporation of new biological discoveries into statistical methodologies for a better understanding of cancer genomics. Dr Huang is the author or co-author of 24 published papers and has presented at 12 conferences.

yen-tsung_huang@brown.edu
https://vivo.brown.edu/display/yh70

SUPPLEMENT TITLE: Sequencing Platform Modeling and Analysis

CITATION: Qin and Huang. Sequencing Platform Modeling and Analysis. *Cancer Informatics* 2015:14(S1) 55–56 doi: 10.4137/CIN.S24698.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Editorial

CORRESPONDENCE: qinl@mskcc.org

All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines.