BIG DATA in Biomarker discovery: Prostate cancer

The deluge in sequence data in the last 10 years from spectacular advances in sequencing technologies has made Big Biology indispensable in addressing human diseases. During my tenure as a graduate student, I have made efforts to comprehend both the breadth of NGS data pertaining to a given biological context and the level of integration that can be achieved by simply mining the data. The choice of prostate cancer to demonstrate the use of Big Data in biology was accidental. One of the first transcriptome dataset that was available for download was for prostate cancer. Subsequently many other groups had put their prostate cancer data online triggering the curiosity to compare the findings from various transcriptome datasets. Needless to say that these were diverse data in terms of both ethnicity and protocols, providing significance to findings common to diverse dataset.

While working on this non-coding RNAs took important stage in prostate cancer. A large number of non-coding RNAs were reported to be differentially regulated in prostate cancer, with a few shown to be implicated in cancer progression. Parallel to this, tens of thousands of non-coding RNAs were discovered in transcriptome data from normal tissues. I was curious to know the regulation of these normal non-coding RNAs in cancer. This quest resulted in majority of the work that makes my thesis.

Chapter 1 and 2 reviews the status of big biology and biomarkers in prostate cancer respectively. Chapter 3 describes my efforts in the identification of expression status of lncRNAs in the various transcriptome datasets from which we had already extracted cancer-specific coding gene. This chapter also describes our efforts to prioritize a few candidate lncRNA for translation based on their proximity to coding genes on the chromosomes. Chapter 4 provides details of translational effort towards one of the most promising candidate non-coding RNAs. Chapter 5 describes our integrated approach to prioritize additional candidate lncRNAs using methylation status from a bisulfite data from another independent group.

Integration of data from diverse applications required development and validation of various pipelines. Also, our variant calling efforts exposed i) a lacuna in variant calling using hg19 as reference because of large number of minor alleles and ii) lack of representation of Indian population-specific variants in public databases for use in identifying somatic mutations. This prompted us to create an alternate reference genome called hg19K and initiate a sequencing effort to create both population-specific and prostate cancer-specific variants in Indian diaspora. The last chapter highlights some of the immediate future goals and inferences.

This thesis is my humble effort where a large-scale integration of bioinformatics and molecular biology lab work has resulted in identification, prioritization and validation of one of the novel genes. Our efforts have shown that bioinformatics should not be considered as just a service to aid research similar to IT services but it should be an integral part of any research framework to
understand biology better. We have demonstrated that in the era of data driven biology, one can start addressing a biological query without even entering the lab by mining relevant data from public repositories. The work here highlights the power of collaboration between bioinformatics and molecular biology in taking biomarkers from data to bench and initiate efforts to traverse from bench to bedside.