Journal of Glycomics & Lipidomics

Extended Abstract

Analysis of short-read data from NCBI SRA and CGHub/TCGA using HIVE-XLD: Evaluation of the effects of non-synonymous variation in the Human proteome

Raja Mazumder

Abstract

Current sequencing technologies are generating petabytes of data that are inaccessible to majority of the research community because of the costs and expertise required to analyze big data. Another roadblock to analyzing such data is the lack of curated information in NGS data repositories such as NCBI SRA. To address the above challenges, we have implemented a low-cost Highperformance Integrated Virtual Environment of eXtra-Large Data (HIVE-XLD) private cloud at GWU and US FDA. Effects of variation on active sites and glycosylation sites will be presented to illustrate the power of integration of big-data with functional objects such as active sites, binding sites and pathways.

DNA sequencing technology advances have enabled genetic investigation of more samples in a shorter time than has previously been possible. Furthermore, the ability to analyze and understand large sequencing datasets has improved due to concurrent advances in sequence data analysis methods and software tools. Constant improvements to both technology and analytic approaches in this fast moving field are evidenced by many recent publications of computational methods, as well as biological results linking genetic events to human disease. Cancer in particular has been the subject of intense investigation, owing to the genetic underpinnings of this complex collection of diseases. New massively-parallel sequencing (MPS) technologies have enabled the investigation of thousands of samples, divided across tens of different tumor types, resulting in new driver gene identification, mutagenic pattern characterization, and other newly uncovered features of tumor biology. This review will focus both on methods and recent results: current analytical approaches to DNA and RNA sequencing will be presented followed by a review of recent pan-cancer sequencing studies. This overview of methods and results will not only highlight the recent advances in cancer genomics, but also the methods and tools used to accomplish these advancements in a constantly and rapidly improving field.

Cancer is not one disease, but a collection of different diseases that share common features: origination from patients' own cells and a disrupted regulatory program resulting in uncontrolled growth. Cancers also generally have a genetic origin; changes in the biological blueprint initiate the signaling and regulatory alterations that lead to tumorigenesis. Early work at the beginning of the 20th century suggested that certain chromosomal aberrations could cause unregulated growth in sea urchin eggs, leading to a variety of hypotheses regarding the role of chromosomes in cancer1. The role of genetic alterations in human cancer was confirmed with the discovery and subsequent classification of the Philadelphia chromosome2, 3. This and other genetic discoveries led to theories of the requirement of multiple genetic events driving clonal selection of tumor cells4-6. Advances in molecular manipulation played a role in the cloning and identification of the first oncogenes, followed by the discovery of the exact nucleotide change responsible for the oncogenic phenotype (in this case, the genetic changes resulting in the RAS G12V amino acid substitution)7-9. This body of work has cemented the role of genetic alterations, large and small, in the biology of cancer.

Cancer genes had been discovered prior to the public release of the human genome using techniques such as positional cloning, biological screening assays, and candidate gene studies (for review see10). However, the Human Genome Project has greatly increased our understanding of the structure and contents of our chromosomes11, 12, allowing for the design and execution of experiments that were not previously realistic. In the continued study of cancer genetics, this led to numerous large-scale investigations of many genes across different cancer types, resulting in the identification of a number of new cancer genes. Initial studies focused on smaller groups of genes, including tyrosine kinases13, a more comprehensive set of kinases14-16, and other genes17. Using this approach, BRAF was discovered to contain a common mutation in several cancer types, with a high prevalence in melanoma18. This discovery eventually resulted in a targeted therapy (vemurafenib) that is commonly used today. Sequencing and sample handling improvements allowed the subsequent investigation of almost all protein coding genes by specific PCR targeting and capillary-based sequencing in 22 breast and colorectal cancers19, 24 pancreatic cancers20, and 22 glioblastoma multiforme tumors21. These improvements were also used to extend kinase sequencing to 210 tumor and matched normal samples, allowing for precise definition and characterization of somatic mutations22. These studies resulted in better understanding of the specific base change classes that were different across cancer types, passenger vs. driver mutations, and new genes important for cancer biology.

Although previous large-scale genetic studies yielded many new insights into the genetic underpinnings of cancer, there were several limitations. These limitations stemmed from the cost of targeting and sequencing many genes with PCR and capillary sequencing based methods.

Few laboratories in the world had the resources to perform such studies, and even those labs had to limit either the number of genes targeted, or the number of tumor samples investigated. And even though sequencing costs had fallen rapidly during the Human Genome Project, costs had not falling rapidly enough to consider sequencing many more samples across all genes, or even whole genomes.

This work is partly presented at 2nd International Conference on Big Data Analysis and Data Mining 30-December 01, 2015 San Antonio, USA

Raja Mazumder
George Washington University, USA E-mail: mazumder@gwu.edu