Ovarian Cancer: Integration of Exome and Transcriptome Data

Introduction

The advancements in NGS technologies and the emergence of Omics field has led to the development of various approaches in studying cancer. The common approaches to identify molecular mechanisms in cancer include scanning the genome for cancer-specific mutations, exploring differential expression of mRNA through transcriptomics or that of protein through proteomics [Chakraborty S et al. (2018)]. However, limiting the analysis to only one type of data is not very efficient because it does not provide a holistic view of the system and the major causal differences between diseased states. Recently, multi-omics approach that looks at integration of omics data using different methods and tools, is believed to give a more comprehensive understanding of biological systems[Ashar Ahmad & Holger Fröhlich (2016), Hasin et al, (2017)].

The incidence of Cancer has been increasing in India, with higher female:male cases being reported in the year 2015[Mathew A et al. (2018)] . Among female malignancies, breast is ranked first, cervical second and ovarian fourth [Masakazu Toi et al, (2010)], although over the years the trend of cervical cancer is decreasing compared to that of ovarian which is emerging at a high pace [Malvia S et al. (2017)]. It was also observed that the onset of breast and ovarian cancer in Indian population occurs at a much earlier age (45-50 years) than in other high-income countries (age>60 years) [Masakazu Toi et al. (2010)]. Several works have focused on identifying genomic alterations, aberrant mRNA and protein expression of breast and ovarian cancers, but there have been minimal efforts towards integration of such data sets for a systemic understanding of disease mechanisms.

This study will be centred towards understanding the molecular mechanisms of ovarian and breast cancer through integration of exome and transcriptome data. The use of exome and transcriptome sequencing for cancer research is rapidly increasing due to its reduced costs. Exome-seq data is used for the identification of variants across the exonic regions of the genome, while RNA-seq is usually used for expression profiling and to study events like splicing and RNA-editing[O'Brien, T. et al. (2015)]. Individually, these two data sets have been applied in various cancer studies, but only a few studies have carried out an integrated analysis. Such an analysis enables one to determine the effects of variants on gene expression, validate common mutations at the gene and transcript level or even explore variants that may no longer be observed at the transcript level due to events like RNA-editing. Moreover, when using only exome seq data, some of the variants may not be functionally relevant and similarly when using only RNA-seq, the underlying cause of differential expression may not be identified. An example would be integrating mutations and expression data to model interactions between groups of different mutated genes and the resulting modifications at the gene expression level.

Literature Review

Data integration in the field of life sciences is not new, many review papers and research

articles have explained the various software tools, methods and workflow of how integration of data has been implemented in their study [Pinu F.R et al. (2019), Brunk E et al. (2016), Yizhak K et al. (2010), Zampieri M et al. (2017), Beal D J et al. (2016), Chen R et al. (2012), Gunther O P et al. (2014)]. A review article in 2014, claimed that the number of papers which had the term data integration in their abstract or title had doubled from the years 2006 to 2013 [Gomez-Cabrero D et al. (2014)]. The review paper published in the 'Briefings in Bioinformatics' journal has explained in detail about the various tools used in analysis of genomic, transcriptomic and proteome data as well as the various initiatives happening in the field of data integration. [Claudia Manzoni et al. (2018)].

A review article published in 2019, has explained all the strengths and challenges involved in all the omics data individually. They have also represented challenges that are specific and those that are shared by different omic fields. Apart from these, a list of all the tools, their platforms and what they can be used for has also been summarized. Some examples of where these integrations have been implemented has been briefed about in their paper [Misra et al. (2019)].

Chakraborty S et al. (2018), a review article has explained the potentials of how multi-omics can be used to interrogate cancer at molecular, cellular, and systems levels. There have been successful integration of transcriptome and epigenome for identifying methylation that affect gene expression; proteome and transcriptome data to identify concordance between mRNA and protein levels in cancer; proteogenomics which is to understand to identify novel genes and updating the annotation of existing genomes; metabolome and transcriptome data for overall insight into altered metabolic networks that are tightly controlled by the transcriptional network.

Davis J. McCarthy et al. (2018), Fleck, J.L., Pavel, A.B. & Cassandras, C.G. (2016), Xiong, Q et al. (2012), Junfei Zhao et al. (2012), Burkhardt et al. (2015) are some cases where integration of mutation and gene expression data has been done for various reasons . Davis J. McCarthy et al. (2018) has identified a novel approach to integrate DNA-seq and RNA-seq data to identify the clonal cell from which the RNA-seq data was obtained. It was also pointed out that the cell cycle pathways were commonly distorted in the mutated clonal cell populations. This study was mainly focused only for inferring the clone of origin of individual cells that have been assayed using single-cell RNA-seq.

Fleck, J.L., Pavel, A.B. & Cassandras, C.G. (2016) proposed a model to infer the temporal sequence in which mutations occur and lead to changes at the gene expression level during cancer progression. It used an integer linear program to identify the order in which mutations occured in three different phases of cancer progression and also the corresponding gene expression changes were studied. They selected only the genes which had previously been reported as tumor suppressor genes or oncogenes and checked for mutations and gene expression changes associated with these genes and the model was tested on stimulated as well as TCGA obtained breast carcinoma data.

Xiong Q et al. (2012) developed a single statistical framework GSAA (Gene set association analysis) which measures the genetic variations across the genome and gene expression variations simultaneously to identify sets of genes enriched with gene expression changes and/or trait associated genetic markers. They had concluded that the power to detect real associations between the genes is better in the case of integrated analysis rather than individual genomic data analysis. They performed the analysis on two human diseases (Glioblastoma and Crohn's disease), the study identified abnormalities in pathways that were previously known to

be involved with the disease.

Junfei Zhao et al. developed a method for integration that could differentiate between genes having identical mutational profiles and also identified gene sets with an optimal score. Their method focused on identifying driver genes and pathways by combining the two measures coverage and exclusivity. Here, high coverage meant that a mutation occurred at least once in a pathway for many samples and high exclusivity meant that most samples had no more than one mutation in the pathway.

Burkhardt et al. (2015) studied underlying mechanisms associated with metabolic disorders by integrating SNP, expression and metabolite data. They first identified variants associated with metabolite levels through GWAS, correlated the validated variants with expression data and analyzed the relationship between these genes expression level and metabolites. At the end, they integrated the three associations and inferred causal relationships among them.

Shi, Kai et al. (2016) used the network based integration of mutation, expression and gene network to identify the driver mutations from a series of passage mutations. This method was used for breast cancer, ovarian cancer and glioblastoma. Apart from identifying the high frequency mutations in driver genes, the method was efficient to identify the driver genes with rare mutations too. The data sets were obtained from the DriverNet.

Multi-omics analysis though having many advantages has some constraints with respect to various aspects. Multi-omics research is expensive and requires a lot of funding. Multi-omics data have demonstrated their utility mainly within the field of precision medicine and personalized and these are considered to be moderately useful. Until the scale is increased to population studies their utility will not be truly recognized. Even though there have been improvements in the tools used for integrating and visualization of the integrated data, there still needs to be more easier and user friendly tools that have open access, so that it can reach biologists with limited knowledge in the area of mathematics and statistics. Another major problem involved in the integration of data is dispersed data sets and the noninterpretability of the tools used in bioinformatics. Above all these, nature of the omics data is an important parameter that decides if the data can be integrated; and if it can be done, how good the integrated data is and so on [Pinu, F.R (2019)].

Literature Gap and Rationale of work

As mentioned in the previous section various studies were carried out on the integration of exome and transcriptome, with some focused on ovarian and breast, but these works were not specific to an Indian cohort. No work on the integration of transcriptome and exome of breast and ovarian cancer has been reported till date which is specific to Indian patient samples. This study is directed towards Indian patients, hence the exome-seq and RNA-seq will be specific to an Indian cohort. The research carried out on breast and ovarian cancer has not resulted in an effective therapy till date. Hence, looking specifically into indigenous populations might be advantageous.

Though the methods used for integration have been described in many research works, only a few of these methods have been replicated successfully in other studies. Also, some of them have not been used on breast and ovarian cancer data. In this study we aim to identify these

gaps and evaluate the best tools which are easy to use, efficient in integrating data and provide a new perspective with a deeper understanding of pathways and mechanisms in cancer.

Many studies have focused on mutations that occur in the coding regions which have an affect on the proteins that are translated. In this study we also aim to look at the mutations in UTR regions and how these mutations may affect gene expression. It is known that most miRNA's bind to the 3' UTR regions of mRNA. Identifying mutations in these regions along with the gene expression levels of the respective transcripts can shed some light on miRNA-based post-transcriptional control in cancers.

Objectives

- To identify and compare methods and tools used for integration of exome and transcriptome data
- To integrate SNP and gene expression data in order to understand the underlying molecular mechanisms involved in breast and ovarian cancer
- Comparison of breast and ovarian cancer results

Scope of Work

The integrated analysis of exome and transcriptome can lead to the identification of pathways and mechanisms that would have otherwise been missed when a single data set would have been analysed against databases. Although an integrated data analysis can prove to give more insights, it will surely require appropriate selection of tools and methods based on the purpose of the research. This study will involve the identification and comparison of such methods and tools and their subsequent use in understanding underlying molecular mechanisms of breast and ovarian cancer by integrating mutation and expression data of an Indian cohort. Further, the mutations in the 3' UTR regions could be used to interpret gene regulation by miRNA. The resulting mechanisms could lead to the identification of novel therapeutic targets that would be highly effective for therapies directed towards Indian patients.