POLITECNICO DI TORINO

Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering

XXIX cycle

Advisor

Enrico MACII

Software tools for NGS data analysis

PhD Candidate:

Giulia PACIELLO

1. Introduction

The advent of Next Generation Sequencing (NGS) dramatically reshaped genomics, allowing to obtain huge amounts of data with both reduced per base costs and high accuracy. Data from sequencing technologies comes in the form of sequences of letters named reads, where each letter reports on a DNA/RNA base. Read analysis accounted in the past decade for the identification of several genomic aberrations such as Single Nucleotide Polymorphisms (i.e., single nucleotide variations) or gene fusions (i.e., hybrid genes originating from the joining of two separate genes). These discoveries led to a better understanding of different pathologies as cancer, bridging to a new era of molecular pathology and personalized medicine. The retrieval of information from sequencing experiments is a big data problem that benefit of ad-hoc bioinformatics approaches.

2. Objectives

Three kinds of chromosomal aberrations associated to cancer onset have been investigated by designing ad-hoc bioinformatics software tools:

- **VDJSeq-Solver** *[1]*: Paired-end RNA-Sequencing (RNA-Seq) reads (i.e., reads coming from the sequencing of both the ends of a RNA fragment) are analysed to provide users with a detailed picture of the immunological B cell (i.e., white blood cells) repertoire characterizing both reactive and neoplastic samples.
- **isomiR-SEA** *[2]*: RNA-Seq reads are processed to identify and quantify the amount of miRNAs (i.e., short non coding RNAs), isomiRs (i.e., miRNA variants) and conserved miRNA:messengerRNA (mRNA) interaction sites in sequenced samples.
- **FuGePrior** *[3]*: Results from state-of-the-art tools for gene fusion detection are analysed, filtered and prioritized to identify the most significant and reliable gene fusions.

4. Results

- **VDJSeq-Solver** performance has been assessed on 5 Mantle Cell Lymphoma (MCL) and 3 Reactive Follicular Hyperplasia (RFH) samples from the University Hospital of Verona, and 12 Diffuse Large B Cell Lymphoma (DLBCL) datasets from The Cancer Genome Atlas (TCGA). *VDJSeq-Solver* correctly identified a prevalent clone in both MCL and DLBCL data. The sequence of the main clone has been confirmed by wet lab experiments. Furthermore, according to biological insights, *VDJSeq-Solver* pointed out a polyclonal distribution in RFH samples.
- **isomiR-SEA** run on 4 large public datasets from different tissues, species and cancers. Results have been compared to those provided by widely adopted algorithms for miRNA analysis proving the importance of miRNA-specific alignment procedures to carefully evaluate miRNA regulative functions.
- **FuGePrior** analysis has been conducted on several Acute Myeloid Leukemia (AML) samples in the context of NGS-PTL European Project, where it accounted for the identification of several real fusions. The correct prioritization of real gene fusions has been also confirmed on two public datasets respectively from prostate and breast cancers.

3. Methods

VDJSeq-Solver

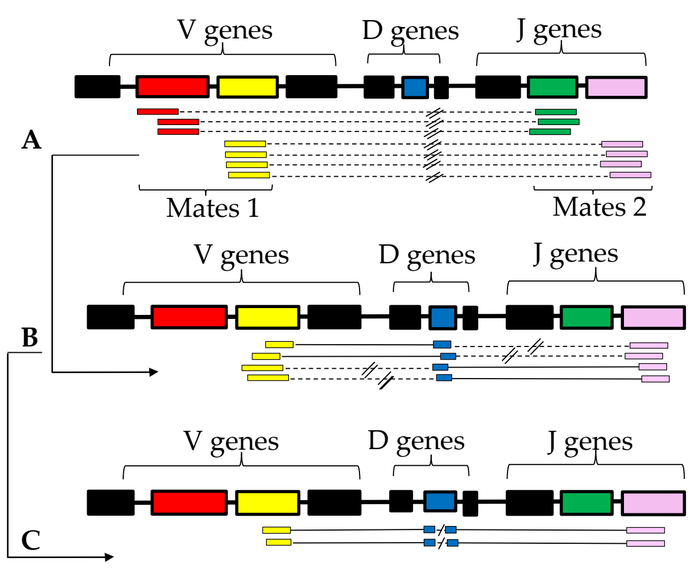

Figure 1: *VDJSeq-Solver* algorithm

The V(D)J Recombination process is responsible for the generation of a huge repertoire of different B cell populations capable to contrast heterogeneous pathogens. These populations, known as B cell clones, are characterized by unique rearrangements of their V, D and J gene segments. *VDJSeq-Solver* exploits paired-end RNA-Seq reads to recognize these rearrangements, identifying different clones. Firstly, reads having the two mates (i.e., the two sequenced ends of a DNA/RNA fragment) aligned to a V and a J gene respectively, define the VJ clones and are labelled as VJ encompassing reads *[A]*. Later, those mates belonging to a VJ encompassing read that are partially mapped onto a D gene identify the D recombined gene *[B]*. Finally, VJ encompassing reads with both the mates partially aligned to a D gene are retrieved to reconstruct the clone sequences *[C]*.

isomiR-SEA

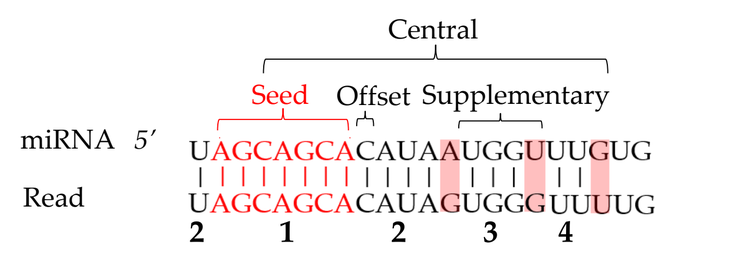

Figure 2: *isomiR-SEA* algorithm

miRNAs act as post-transcriptional regulators by binding target mRNA molecules. The effectiveness of the binding mainly depends on four regions on miRNA sequence named seed, offset, supplementary and central sites.
Mutations on miRNA sequences give rise to miRNA variants known as isomiRs that have been proven to target a different set of mRNA with respect to the miRNA from which they origin.
isomiR-SEA exploits a miRNA tailored alignment procedure that begins with the identification of the seed sequence on the read *[A]*. Follows the ungapped extension of the alignment in both 5′ and 3′ directions until the first encountered mismatch *[B]*. Two consecutive extensions in 3′ direction are then performed according to user specified configuration parameters *[C, D]*. By keeping track of the positions of the encountered mismatches (red boxes in *Fig. 2*), *isomiR-SEA* is able to distinguish among miRNAs and isomiRs, and to evaluate the conservation of miRNA:mRNA interaction sites.

FuGePrior

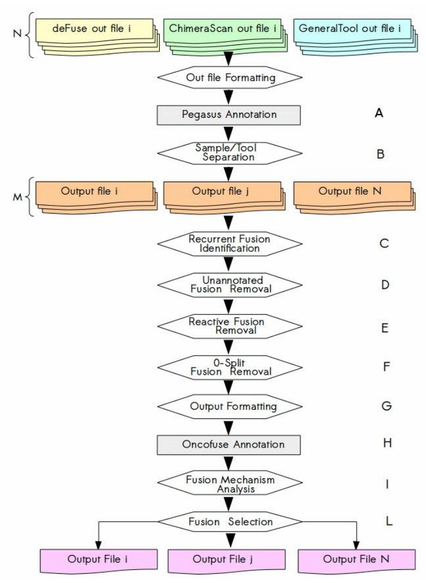

Figure 3: *FuGePrior* algorithm

The 90% of known gene fusions have been identified by NGS experiments, specifically RNA-Seq. However, gene fusion detection tools exploiting sequencing data, suffer of different drawbacks as poorly overlapping results and huge amounts of false positives. *FuGePrior* prioritizes gene fusion candidates by implementing a multi-step processing and filtering approach designed by considering gene fusion features. Information as the occurrence of the fusion in reactive samples *[E]*, the number of supporting reads *[F]* or the biological mechanism responsible for its generation *[I]* are evaluated by *FuGePrior* to focus on the most reliable fusions with a potential driver role in the pathology.

5. Conclusions

The advent of NGS technologies opened the way to a novel era of genomic studies. The analysis of data coming from sequencing experiments allowed to identify several chromosomal aberrations involved in complex diseases such as cancer. These findings benefited from ad-hoc bioinformatics pipelines and tools capable to extract information from the huge amount of data from sequencing experiments. In this context, *VDJSeq-Solver*, *isomiR-SEA* and *FuGePrior* have been respectively implemented for B cell clonal population, miRNA/isomiR and gene fusion analyses. The good performance achieved on both public and private RNA-Seq datasets makes their adoption a viable solution to gain novel biological and clinical insights.

6. References

1. Paciello, G., Acquaviva, A., Pighi, C., Ferrarini, A., Macii, E., & Ficarra, E. (2015). VDJSeq-Solver: in silico V(D)J recombination detection tool. PloS one, 10(3), e0118192.
2. Urgese, G., Paciello, G., Acquaviva, A., & Ficarra, E. (2016). isomiR-SEA: an RNA-Seq analysis tool for miRNAs/isomiRs expression level profiling and miRNA-mRNA interaction sites evaluation. BMC bioinformatics, 17(1), 1.
3. Paciello G., & Ficarra, E. (2016). FuGePrior: A novel gene fusion prioritization algorithm based on accurate fusion structure analysis in cancer RNA-seq samples. BMC bioinformatics, [Under Review].