International Journal of Biochemistry and Cell Biology 91 (2017) 116–123

Contents lists available at ScienceDirect

International Journal of Biochemistry and Cell Biology

journal homepage: www.elsevier.com/locate/biocel

A computational integrative approach based on alternative splicing analysis to compare immortalized and primary cancer cells

Kumar Parijat Tripathi[1], Ilaria Granata[1,*], Mario Rosario Guarracino

High Performance Computing and Networking Institute, National Research Council, Italy

ARTICLE INFO

Keywords:
Splicing
Single nucleotide polymorphism
Integrative analysis

ABSTRACT

Immortalized cell lines are widely used to study the effectiveness and toxicity of anti cancer drugs as well as to assess the phenotypic characteristics of cancer cells, such as proliferation and migration ability. Unfortunately, cell lines often show extremely different properties than tumor tissues. Also the primary cells, that are deprived of the *in vivo* environment, might adapt to artificial conditions, and differ from the tissue they should represent. Despite these considerations, cell lines are still one of the most used cancer models due to their availability and capability to expand without limitation, but the clinical relevance of their use is still a big issue in cancer research. Many studies tried to overcome this task, comparing cell lines and tumor samples through the definition of the genomic and transcriptomic differences. To this aim, most of them used nucleotide variation or gene expression data. Here we introduce a different strategy based on alternative splicing detection and integration of DNA and RNA sequencing data, to explore the differences between immortalized and tissue-derived cells at isoforms level. Furthermore, in order to better investigate the heterogeneity of both cell populations, we took advantage of a public available dataset obtained with a new simultaneous omics single cell sequencing methodology. The proposed pipeline allowed us to identify, through a computational and prediction approach, putative mutated and alternative spliced transcripts responsible for the dissimilarity between immortalized and primary hepato carcinoma cells.

1. Introduction

Human-derived cancer cell lines have been for decades the elective model to study the cancer biology and to test new anti cancer therapies (Goodspeed et al., 2016). Although the rapid scientific progress, cell line-based assays still represent an important resource for pharmaceutical, chemical, medical and cosmetic industries. The lower costs, the culture methods easy to handle and the high reproducibility ensure their extensive use. Primary tumors represent a reliable but more expensive and less available resource. Furthermore, they are constituted by a highly heterogeneous cell population containing also non-cancerous cells that might affect the results of the performed experiments. However, the relevance of cell lines as tumor models strongly depends on the type of experimental approach and on how close their properties are to those of tumor tissue (Gillet et al., 2013). Thus, the investigation and the definition of this closeness is a very important issue for biologists and might lead to the development of new *in vitro* pre-clinical models. Many studies have been carried out with the aim to determine the differences in terms of functions between primary cells and cell lines (Ertel et al., 2006; Pan et al., 2009). The used approaches are commonly based on the comparison of gene expression (Vincent et al., 2015; Tyakht et al., 2014; Chen et al., 2015) or genomic (Domcke et al., 2013) profiles. The advent and the incessant development of high throughput technologies have provided a huge amount of omics data aimed at understanding the biological processes and functions of living organisms and the tight regulation existing among their constituent molecules. In particular, their application has led to the molecular classification of many diseases and to the identification of biomarkers and therapeutic targets involved in the mechanisms responsible of rise, progression and outcome of the pathology under study. Recent successes in research have finally defined the ”one gene, one protein, one function” dogma, postulated by Beadle and Tatum (Beadle and Tatum, 1941), as outdated. The latest GENCODE release (version 25) (Harrow et al., 2012) reports almost 80,000 transcript variants encoded by about 20,000 protein-coding genes in humans, suggesting an average of four transcripts per gene, although the number of transcripts per gene varies accordingly to the different databases, highlighting the current limitations in fully characterizing the transcriptome. More than 90% of

[*] Corresponding author.
E-mail address: ilaria.granata@icar.cnr.it (I. Granata).
[1] Equal contributors.

http://dx.doi.org/10.1016/j.biocel.2017.07.010
Received 15 March 2017; Received in revised form 3 July 2017; Accepted 15 July 2017
Available online 27 July 2017

human genes are alternatively spliced, with a role in many physiological functions. The dys-regulation of the splicing machinery has been associated with a wide variety of human diseases (Hsu and Hertel, 2009; Scotti and Swanson, 2016). The proteins translated from the alternatively spliced transcripts can have similar, different or even opposing functions. Five major alternative splicing events are distinguished: exon skipping (SE), also called *cassette exon*, use of alternative acceptor and/or donor sites (A5SS, A3SS), mutually exclusive exons (MXE) and intron retention (RI). How the spliceosome recognizes alternative exons and decides how to splice the mRNA still remains not fully understood. The aberrant use of alternative mRNA isoforms has been found linked to cancer formation. It is well known that several oncogenes and tumor-suppressor genes (for example, LEF1, TP63, TP73, HNF4A, RASSF1, and BCL2L1) have multiple promoters and alternative splice variants (Zhang et al., 2013; Hovanes et al., 2001; Wilhelm et al., 2010; Nekulova et al., 2011; Tomasini et al., 2008). These findings highlight the importance of focusing on the isoform level expression profiles and on the understanding of the tightly regulated splicing machinery to better define the signature of cancer cells. The huge amount of omics data, which range from DNA to RNA-sequencing and to proteomic data, allows to develop analysis pipelines based on integration approaches. One of the challenges faced by these approaches is the combination of single nucleotide polymorphisms (SNPs) and splicing events in order to identify the genetic variants affecting the splicing machinery. A large fraction of DNA variants takes place within splice site sequences at the intron-exon junction, or within enhancer and silencer sequences. As a consequence, they may alter the splicing machinery activity and its tight regulation. The dysfunctionality introduced by these nucleotide variations in pre-mRNA splicing could lead both to novel transcripts and to an abnormal ratio of alternative splicing patterns. Based on these considerations, we designed and use a bioinformatics pipeline to compare the molecular properties of immortalized and primary cell lines through an integrative approach based on the study of the alternative splicing. Furthermore, in order to take into account the high heterogeneity of primary cells, we used, as case study, a public available single cell sequencing dataset containing both hepatocellular cell line samples (HepG2) and related cancer tissue cells. DNA and RNA were sequenced contemporary through a novel technique developed by the authors called scTrio-seq (Hou et al., 2016). This work highlights the importance and the possible involvement of differentially spliced isoforms in determining the differences between immortalized and tissue-derived cell lines.

2. Materials and methods

2.1. Data

In order to test our approach and develop an appropriate pipeline, we downloaded a publicly available dataset from Gene Expression Omnibus (GEO) portal. This dataset (GSE65364) has been selected since it was generated through a novel triple-omics sequencing protocol developed by the authors (Hou et al., 2016). It is a single cell sequencing technique, called scTrio-seq, able to analyze the genome, DNA methylome, and transcriptome simultaneously. In particular, 6 single human HepG2 cell line and 16 single cells from hepato-cellular carcinoma (HCC) samples were downloaded to perform the integration between nucleotide variants and splicing events.

2.2. Sequencing data pre-processing

Genomic and transcriptomic data used in this study were sequenced through scTrio RRBS protocol. SRA files were downloaded and converted to fastq format using *SRA toolkit* (Leinonen et al., 2010). Quality assessment was performed through *FastQC* tool (Andrews et al., 2010). The low quality read ends, library construction adapters and amplification primers were removed using *Trim Galore* (Krueger and Galore, 2015). In the case of scRRBS-seq data the option "-rrbs" for MspI-digested RRBS libraries was set.

2.3. Gene expression and alternative splicing analysis of RNA-seq data

Good quality reads were aligned to the human genome (assembly hg19) using *Tophat2* (Kim et al., 2013). Samples having concordant pairing rate of forward and reverse reads greater than 60% were selected. The range of concordant alignment were from 45% to 75%, and since we analyzed single cell sequencing data we chose this threshold to get enough statistical support for splicing analysis. In order to quantify and normalize the gene expression levels in each cell sample, *cuffquant*, *cuffnorm*, *cuffdiff* tools from *cufflinks* suite (Trapnell et al., 2012) were used. Before investigating differential splicing events in each sample, we first explored the distance between HepG2 and HCC samples using dimensionality reduction and distance matrix calculation approach. In the case of dimensionality reduction, we employed Multi Dimension Scaling (MDS) and Principal Components Analysis (PCA) methods from *CummeRbund* package (Goff et al., 2013). We also calculated the Jensen–Shannon divergence (Endres and Schindelin, 2003; & sterreicher and Vajda, 2003) between the HepG2 and HCC populations, both using the whole gene expression dataset and selecting the significant differentially spliced transcripts. Furthermore, Multivariate Analysis of Transcript Splicing (MATS) computational tool (Shen et al., 2012) was used to detect alternative splicing events in samples under study.

2.4. Variant calling on scRRBS-seq data

Trimmed sequences were aligned to the reference genome (assembly hg19) using *BSMAP* (Xi and Li, 2009), a short reads mapping software designed ad-hoc for bisulfite sequencing reads. The bam files obtained from the alignment step were used as input to *BS-SNPer*, a package for the exploration of single nucleotide polymorphism (SNP) sites from BS-Seq data (Gao et al., 2015). The output files containing the called SNP of the samples were sorted by Picard (http://picard.sourceforge.net) and merged together by *GATK suite* (McKenna et al., 2010). The sorted files were annotated and filtered using *Var2GO* (Granata et al., 2016), a web tool to annotate variants as well as related genes and to filter VCF (Variant Calling Format) files in an interactive way. SNPs were filtered based on the "FILTER" field of the *BS-SNPer* output. Furthermore, bam files were used to perform Multi Dimensional Scaling (MDS) clustering of samples through *Samtools* (Li et al., 2009) and *Plink* (Chang et al., 2015) tools.

2.5. Integrative analysis

In order to integrate the results obtained by splicing events detection and variant calling, a in-house Python script was developed. RefSeq GTF file was used to define the exonic and intronic boundaries of each gene and to annotate the coordinates of the exons involved in alternative splicing events as well as those of SNPs called in each samples. For each splicing events, the exons which undergo splicing events were associated to the corresponding genes. SNPs were then annotated in terms of associated gene symbol. A schematic view of the pipeline is illustrated in Fig. 1.

2.6. Exonic splicing enhancer regions detection

In order to get information about a possible link between the mapped variants and the mechanism of deterring splicing, we investigated whether detected point mutations could affect exonic splicing enhancer (ESE) regions. From the literature, it is well known that point mutations can inactivate an ESE and hence lead to alternative splicing events (Cartegni et al., 2003). ESEs are common elements in both skipping and constitutive exons, acting as a binding sites for Ser/Arg-rich proteins (SR proteins). SR proteins are conserved splicing