Gene xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Gene

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

RNA-Seq SSRs and small RNA-Seq SSRs: New approaches in cancer biomarker discovery

Arghavan Alisoltani [a,1], Hossein Fallahi [b,c,1], Behrouz Shiran [a], Anousheh Alisoltani [d], Esmaeil Ebrahimie [e,f,*]

[a] Biotechnology Group, Shahrekord University, Shahrekord, Iran
[b] Department of Biology, School of Sciences, Razi University, Kermanshah, Iran
[c] Medical Biology Research Center, Kermanshah University of Medical Sciences, Kermanshah, Iran
[d] School of veterinary science, Shahrekord University, Shahrekord, Iran
[e] Institute of Biotechnology, Shiraz University, Shiraz, Iran
[f] Discipline of Genetics, School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, Australia

ARTICLE INFO

Article history:
Received 23 September 2014
Received in revised form 2 December 2014
Accepted 13 January 2015
Available online xxxx

Keywords:
Biomarker
EST–SSR
Leukemia
Next Generation Data Sequencing
Pancreatic cancer
RNA-seq
Small RNA-seq

ABSTRACT

The recent exponential increase in the number of next generation sequencing studies provides a new source of data for the discovery of functional genomics based markers. The RNA-seq and small RNA-seq provide a new source for the discovery of differentially expressed SSRs (simple sequence repeats) as biomarkers in various diseases. In the present study, for the first time, we applied RNA-seq SSR to find new biomarkers for pancreatic cancer (PC) diagnosis. Analysis of RNA-seq data revealed a significant alternation in the frequency of SSR motifs during cancer progression. In particular, RNA-seq SSR showed an increase in the frequencies of GCC/GGC and GCG/CGC motifs in PC samples compared to healthy pancreas. These findings were further confirmed using meta-analysis of EST–SSR data in 11 different cancers. Interestingly, the genes containing GCC/GGC and GCG/CGC motifs in their sequences were involved in many cancer-related biological processes, particularly regulation processes. The small RNA-seq data were also mined for the conserved patterns in SSR frequencies (sRNA-seq SSR) during cancer progression. Based on the results, we suggest the potential use of GCC/GGC and GCG/CGC motifs as biomarkers in PC. Based on the findings of this study, it seems that RNA-seq SSR and sRNA-seq SSR could open a new paradigm in the diagnostic and even therapeutic strategies for PC along the other types of cancers.

1. Introduction

Simple sequence repeats (SSRs) or microsatellites are tandem repeats of 1–6 nucleotide motifs in nucleic acid sequences (Ebrahimi et al., 2011). These motifs are located on the non-coding regions as well as the coding regions of genomes (Bakhtiarizadeh et al., 2011). Variation of SSRs that is defined as microsatellite instability (MSI) has been observed in human diseases including different cancers (Gonzalez-Zulueta et al., 1993; de la Chapelle and Hampel, 2010; Bakhtiarizadeh et al., 2011). Our recent investigation demonstrated that SSRs have differential expression pattern between hematopoietic normal and cancer stem cells (Hosseinpour et al., 2014).

SSRs undergo quantitative and qualitative variations due to mutations that add or subtract repeat units (Kashi and King, 2006). Therefore, the influence of SSRs on gene regulation, transcription and protein function can happen through the number of repeats or the sequences of repeats (Kashi and King, 2006). Microsatellites or SSR markers are informative and versatile biomarkers that can be used in many areas of research (Bakhtiarizadeh et al., 2011, 2012; Zalapa et al., 2012; Hajmansoor et al., 2013). One of the most important features of SSR markers is that they can detect multiple alleles per locus (Zalapa et al., 2012). Expressed SSRs have opened a new vista in biomarker discovery as they can be functional and are present in transcriptomic level.

High-throughput datasets including expressed sequence tags (ESTs) are valuable resources for SSR discovery, functional genomics and biodiversity studies (Bakhtiarizadeh et al., 2011, 2013; Hosseinpour et al., 2014). Our recent studies highlighted the potential use of EST–SSR approach to find biomarkers in lung-cancer and hematopoietic cancer stem cells (Bakhtiarizadeh et al., 2011; Hosseinpour et al., 2014). EST–SSRs can be used efficiently in cancer studies; however, traditionally development of EST libraries is difficult and costly, which limits its application in clinical diagnosis. Next generation sequencing (NGS) technologies, in contrast, allow the efficient identification of a large number of sequences at a fraction of the cost and effort offered by traditional

Abbreviations: BPC, Benign pancreatic cancer; CNVs, Copy number variations; DE, Differentially expressed; ESTs, Expressed sequence tags; ncRNAs, Non-coding RNAs; NGS, Next generation sequencing; PBMCs, Peripheral blood mononuclear cells; PC, Pancreatic cancer; RNA-seq, RNA-sequencing; SNPs, Single nucleotide polymorphism; sRNA-seq, Sequencing of small RNA; SSRs, Simple sequence repeats.

* Corresponding author at: Institute of Biotechnology, Shiraz University, Shiraz, Iran.
E-mail addresses: a_alisoltani@ut.ac.ir (A. Alisoltani), h.fallahi@razi.ac.ir (H. Fallahi), esmaeil.ebrahimie@adelaide.edu.au (E. Ebrahimie).

[1] These authors equally contributed to this work.

http://dx.doi.org/10.1016/j.gene.2015.01.027
0378-1119/

approaches. Among NGS techniques, high-throughput RNA sequencing (RNA-seq) is rapidly growing as a major quantitative transcriptome profiling approach (Wang et al., 2009). Another type of NGS is high throughput sequencing of small RNA (sRNA-seq) that is recently used as a profiling approach for small non-coding RNAs (ncRNAs) (Fasold et al., 2011).

We suggest that RNA-seq and sRNA-seq data can be used to investigate the relationships between expressed SSRs and cancer, in order to uncover the influence of such alteration on the development and progression of cancer. These insights can then be translated to clinical benefits, including the development of reliable cancer biomarkers and effective strategies for cancer prevention and therapy. Furthermore, the rapid accumulation of transcriptome data offers a unique opportunity to integrate and utilize these sources for cancer studies. Integrating data from several studies, termed as meta-analysis, can also increase the efficiency and reliability of the results.

Pancreatic cancer (PC) is one of the most lethal malignancies, where no reliable modality is available for the early detection of this disease (Yi et al., 2013). The poor prognosis of PC is partly due to the late clinical presentation and the lack of the effective early detection measures (Wang et al., 2013). Due to the fact that most cancers show high degree of instability (de la Chapelle and Hampel, 2010), finding more reliable and stable markers for diagnosis of PC and other types of cancers is highly desirable.

So far, there has been no report on the application of RNA-seq SSR or sRNA-seq SSR, in cancer diagnosis. There is also no study on the integrative analysis of SSRs using ESTs of different types of cancers. These types of integrative analysis can rigorously increase generalizability and reliability of findings. Here, for the first time, we have conducted RNA-seq SSR and sRNA-seq SSR analyses to identify SSRs in peripheral blood mononuclear cells (PBMCs). The results of RNA-seq SSR were further confirmed using the analysis of EST–SSRs in 11 different cancers. One of the main goals in this study was the identification of reliable biomarkers shared across a variety of cancers. Using SSR analysis, we identified potential motifs for PC diagnosis as well as the other types of cancers.

2. Material and methods

In this study, we have developed a simple pipeline aiming to detect reliable biomarkers in cancer diagnosis. The overview of our workflow is illustrated in Fig. 1. In brief, the different types of datasets were selected for SSR analysis, including RNA-seq and sRNA-seq data, EST libraries as well as microarray datasets. RNA-seq and sRNA-seq data were checked for quality and trimmed before SSR analysis, while EST libraries were directly used for SSR scanning. Then, to obtain differentially expressed (DE) SSRs, the number of SSRs was compared between cancerous and healthy libraries. To find target genes with DE SSRs, sequences with specific altered SSRs were further extracted and annotated. In the final step, Gene Ontology of target genes was performed. In the case of microarray data, the over- and down-expressed genes in 11 cancers were obtained from our previous study (Alisoltani et al., 2014). After retrieval of sequences for over- and down-expressed genes, sequences were scanned for SSRs. Target genes with specific SSR motif were filtered and finally, the common altered genes across different cancers were selected. The details of each part of analysis are described below.

2.1. Data collection

Two sets of data including RNA-seq and sRNA-seq data of blood sample from pancreatic cancer (PC) patients were obtained from Short Read

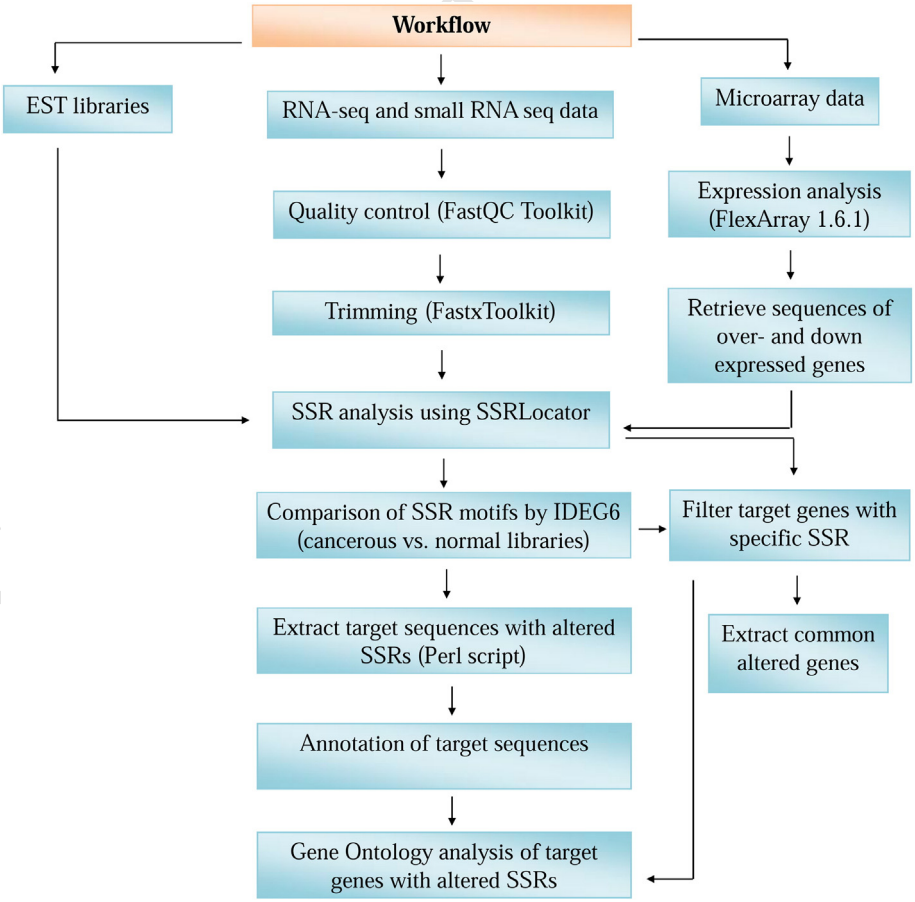

Fig. 1. Workflow of SSR analysis across different cancers using RNA-seq and EST libraries as well as microarray datasets.

Please cite this article as: Alisoltani, A., et al., RNA-Seq SSRs and small RNA-Seq SSRs: New approaches in cancer biomarker discovery, Gene (2015), http://dx.doi.org/10.1016/j.gene.2015.01.027