Computational Biology and Chemistry xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Characterization and distribution of repetitive elements in association with genes in the human genome

Kai-Chiang Liang [a], Joseph T. Tseng [a], Shaw-Jenq Tsai [b], H. Sunny Sun [a,c,*]

[a] Institute of Bioinformatics and Biosignal Transduction, College of Bioscience and Biotechnology, National Cheng Kung University, Tainan, Taiwan
[b] Department of Physiology, College of Medicine, National Cheng Kung University,Tainan, Taiwan
[c] Institute of Molecular Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan

ARTICLE INFO

Article history:
Received 31 December 2014
Accepted 3 February 2015
Available online xxx

Keywords:
Tandem repeats
Interspersed repeats
Genome evolution
Gene regulation

ABSTRACT

Repetitive elements constitute more than 50% of the human genome. Recent studies implied that the complexity of living organisms is not just a direct outcome of a number of coding sequences; the repetitive elements, which do not encode proteins, may also play a significant role. Though scattered studies showed that repetitive elements in the regulatory regions of a gene control gene expression, no systematic survey has been done to report the characterization and distribution of various types of these repetitive elements in the human genome. Sequences from 5′ and 3′ untranslated regions and upstream and downstream of a gene were downloaded from the Ensembl database. The repetitive elements in the neighboring of each gene were identified and classified using cross-matching implemented in the RepeatMasker. The annotation and distribution of distinct classes of repetitive elements associated with individual gene were collected to characterize genes in association with different types of repetitive elements using systems biology program. We identified a total of 1,068,400 repetitive elements which belong to 37-class families and 1235 subclasses that are associated with 33,761 genes and 57,365 transcripts. In addition, we found that the tandem repeats preferentially locate proximal to the transcription start site (TSS) of genes and the major function of these genes are involved in developmental processes. On the other hand, interspersed repetitive elements showed a tendency to be accumulated at distal region from the TSS and the function of interspersed repeat-containing genes took part in the catabolic/metabolic processes. Results from the distribution analysis were collected and used to construct a gene-based repetitive element database (GBRED; http://www.binfo.ncku.edu.tw/GBRED/index.html). A user-friendly web interface was designed to provide the information of repetitive elements associated with any particular gene(s).

This is the first study focusing on the gene-associated repetitive elements in the human genome. Our data showed distinct genes associated with different kinds of repetitive element and implied such combination may shape the function of these genes. Aside from the conventional view of these elements in genome evolution, results from this study offer a systemic review to facilitate exploitation of these elements in genome function.

1. Introduction

The completion of the human genome revealed that coding sequences of the nearly 25,000 human genes comprise only about 1% of the genome (Lander, 2001) and more than 55% of the human genome is constituted by repetitive elements (Rollins et al., 2006). Two major groups of repetitive elements in eukaryotic genomes are tandem repetitive elements and interspersed repetitive elements. The tandem repetitive elements are two or more adjacent, approximate copies of a pattern of nucleotides and include satellite DNA, simple repeats, and low complexity sequences (Merkel and Gemmell, 2008). While satellite DNA comprises about 4.17% of the human genome (Rollins et al., 2006), simple repeats (also known as microsatellite) and low complexity sequences compose 1.7% and 1.3% in the human genome, respectively (Rollins et al., 2006). The interspersed repetitive elements are mainly transposable elements

* Corresponding author at: Institute of Molecular Medicine, College of Medicine, National Cheng Kung University, Tainan 701, Taiwan. Tel.: +886 6 2353535; fax: +886 6 2095845.
E-mail addresses: z2696103@mail.ncku.edu.tw (K.-C. Liang), tctseng@mail.ncku.edu.tw (J.T. Tseng), seantsai@mail.ncku.edu.tw (S.-J. Tsai), hssun@mail.ncku.edu.tw (H. S. Sun).

http://dx.doi.org/10.1016/j.compbiolchem.2015.02.007
1476-9271/

(TE) and are divided into two groups based on their transposition mechanism and sequence organization (Smit, 1996). The DNA transposon group predominantly move through a DNA-mediated mechanism and constitute approximately 3% of the human genome. The retroelement group, including long terminal repeat-containing elements (LTRs), long interspersed elements (LINEs or L1-element), and short interspersed elements (SINEs) accounts for 9.3%, 22.3% and 16.1% of the human genome, respectively (Rollins et al., 2006).

The repetitive elements were thought to be 'selfish' or 'junk' DNA (Orgel and Crick, 1980) and may originally from parasites (Hickey, 1982; Schmid, 2003) or error(s) during DNA replication, repair, or recombination (Levinson and Gutman, 1987). However, a growing body of evidences implied that enormous biological diversity is not caused by coding sequences alone (John, 2003; Gaillard and Strauss, 2006), instead, it is the other part of the genome where constituted by repetitive elements that may play a significant role in the regulation (Jasinska and Krzyzosiak, 2004; Wang et al., 2007 Wang et al., 2007). Thus, it is of importance to recognize the distribution and characterization of these repetitive elements in order to thoroughly understand the regulation of human genome. Such fundamental studies have been performed in many different species including *Escherichia Coli* (Kawano et al., 2002), fish (Cioffi and Bertollo, 2012), and bovine (Adelson et al., 2009) genomes. Since a thorough study of repetitive elements will provide pivotal insights into genome functions and evolutionary process, there is no surprise that many computational (Jurka, 1998; Sharma et al., 2007; Wang and Xu, 2009) as well as experimental (Hormozdiari et al., 2011; Cao et al., 2014) approaches had been developed to identify novel repetitive elements using genomic sequences. For example, the L1Base provides the functional annotation and prediction of active LINE1 elements (Penzkofer et al., 2005) and the TRbase emphasizes tandem repeats related to disease genes in the human genome (Boby et al., 2005). Though the web server/database for repetitive elements identification (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org) or visualization (UCSC Genome Browser; https://genome.ucsc.edu/) are also available, these web tools provide only the general information while the data on gene and repetitive element association is missing. Thus, users may have difficulties to propose the functional roles of these repetitive elements.

Growing evidences support that repetitive elements have special functions to influence gene expression. Previous studies suggested that repetitive elements can influence genome stability and gene expression (Charlesworth et al., 1994; Rodriguez et al., 2008). For example, the mutational event of interspersed repeats may cause genome rearrangement and illegitimate recombination, leading to various human disorders (Jurka, 1990; Kreahling and Graveley, 2004). Recent studies also demonstrated that the insertion of interspersed repeats into a new genomic position will introduce sequence motifs for transcription of nearby genes (Eller et al., 2007), such as promoter, enhancer motifs (Sverdlov, 1998), alternative splicing sites or polyadenylation sites (Makalowski, 2000) and results in a change of overall level of gene expression. Further studies demonstrated that a large fraction of transcription factor binding sites are embedded in distinctive families of interspersed repetitive elements, the so-called repeat-associated binding sites (RABS) (Bourque et al., 2008). In addition to transposable elements, many studies have demonstrated that microsatellites DNA in the non-coding region may function in gene regulation, presumably by forming specific DNA structures like Z-DNA (Rothenburg et al., 2001) or H-DNA (Fabregat et al., 2001). Besides, the expansion of tandem repeats are known to affect neighboring gene expression and cause nearly 30 human diseases (Cummings and Zoghbi, 2000; Liquori et al., 2001; Mirkin, 2007). These results implied that a sizable fraction of gene-associated elements are functionally important in eukaryote; therefore, the distribution of these gene-associated repetitive elements is worth further investigation. This study aimed to thoroughly examine the repetitive elements in association with genes in the human genome using computational approaches. Especially, our efforts focused on the regions that are known to be involved in gene regulation such as upstream flanking region and untranslated regions.

2. Materials and method

2.1. Data sources and tools

All sequences and annotation of genes were downloaded from Ensembl genome via BioMart program (Spudich et al., 2007). To investigate repetitive elements nearby an annotated gene, we downloaded the sequences flanking all human genes, including 5′ untranslated region, 3′ untranslated region and the 5000 base pairs of upstream and downstream flanking regions. All the annotations of repetitive elements, including their reference sequences and the definition of repetitive element families, were downloaded from Repbase (Jurka et al., 2005). The most commonly used library-based tool for repetitive element identification, RepeatMasker program (version 3.3, A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org), was applied in this study. The cutoff Smith–Waterman scores are 225 for most repeats, 300 for the low-complexity and LINE1 search, and 180 for the very old MIR, LINE2 and MER5 sequences.

2.2. Establish gene-based repetitive element database (GBRED)

The database was written using MySQL (structured query language) on the Windows XP operating system. In addition, web interface was written using Hypertext Preprocessor (PHP) scripting language. The database mainly consists of four related tables: "Original information", "Original sequence", "RepeatMasker information", and "Repeat length". The "Original information" table summarizes attributes for each gene in the human genome, and "Original sequence" stores the sequences flanking all the human genes, and the "RepeatMasker information" table and "Repeat length" table provide detail information for repetitive elements. The detailed schema diagram of GBRED is given in Supplementary Fig. 1. A manual that explains all information downloaded from Ensembl was provided online for internet users (http://multi-course.binfo.ncku.edu.tw/GBRED/manual.htm).

2.3. Functional annotation of repetitive elements

An integrated "knowledge-based" platform for pathway analysis of experimental data, Metacore (GeneGo, Inc., St. Joseph, MI), was used to further illustrate functional relationship among repetitive element-containing genes (Bugrim et al., 2004). This tool is based on a proprietary manually curated database of human protein–protein, protein–DNA and protein compound interactions, metabolic and signaling pathways and the effects of bioactive molecules in gene expression. The significance was evaluated and ranked based on the size of the intersection between user's dataset and set of genes corresponding to a network module/pathway/process in question (Bessarabova et al., 2012). In addition, we randomly selected similar number of genes in the human genome that do not contain repetitive elements in a given region using MySQL script. These sets of genes were used to run the same annotation analyses and results were provided as controls.

Please cite this article in press as: K.-C. Liang, et al., Characterization and distribution of repetitive elements in association with genes in the human genome, Comput. Biol. Chem. (2015), http://dx.doi.org/10.1016/j.compbiolchem.2015.02.007