Gene 436 (2009) 12–22

Contents lists available at ScienceDirect

Gene

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

Repetitive DNA elements, nucleosome binding and human gene expression

Ahsan Huda [a], Leonardo Mariño-Ramírez [b,c], David Landsman [b], I. King Jordan [a,*]

[a] School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332, USA
[b] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[c] Computational Biology and Bioinformatics Unit, Biotechnology and Bioindustry Center, Corporacion Colombiana de Investigacion Agropecuaria - CORPOICA, Km. 14 Via a Mosquera, Bogota, Colombia

ARTICLE INFO

Article history:
Received 22 January 2009
Accepted 23 January 2009
Available online 5 February 2009

Received by M. Batzer

Keywords:
Transposable elements
Nucleosome binding
Epigenetics
Simple sequence repeats
Gene regulation
Promoter architecture
Human genome

ABSTRACT

We evaluated the epigenetic contributions of repetitive DNA elements to human gene regulation. Human proximal promoter sequences show distinct distributions of transposable elements (TEs) and simple sequence repeats (SSRs). TEs are enriched distal from transcriptional start sites (TSSs) and their frequency decreases closer to TSSs, being largely absent from the core promoter region. SSRs, on the other hand, are found at low frequency distal to the TSS and then increase in frequency starting ~150 bp upstream of the TSS. The peak of SSR density is centered around the −35 bp position where the basal transcriptional machinery assembles. These trends in repetitive sequence distribution are strongly correlated, positively for TEs and negatively for SSRs, with relative nucleosome binding affinities along the promoters. Nucleosomes bind with highest probability distal from the TSS and the nucleosome binding affinity steadily decreases reaching its nadir just upstream of the TSS at the same point where SSR frequency is at its highest. Promoters that are enriched for TEs are more highly and broadly expressed, on average, than promoters that are devoid of TEs. In addition, promoters that have similar repetitive DNA profiles regulate genes that have more similar expression patterns and encode proteins with more similar functions than promoters that differ with respect to their repetitive DNA. Furthermore, distinct repetitive DNA promoter profiles are correlated with tissue-specific patterns of expression. These observations indicate that repetitive DNA elements mediate chromatin accessibility in proximal promoter regions and the repeat content of promoters is relevant to both gene expression and function.

1. Introduction

The prevalence of repetitive DNA sequences in mammalian genomes has been appreciated since the classic re-association kinetic (COT-curve) experiments of the late nineteen-sixties (Britten and Kohne, 1968). The completion of the human genome projects at the turn of the millennium further underscored the extent to which the human genome sequence is made up of repetitive DNA elements (Lander et al., 2001; Venter et al., 2001). There are several distinct categories of repetitive sequence elements in the human genome. Interspersed repeat sequences, also known as transposable elements (TEs), make up at least 45% of the euchromatic genome sequence, and novel human TE families continue to be discovered and characterized (Wang et al., 2005; Nishihara et al., 2006). Simple sequence repeats (SSRs) consist of tandem repeats of exact or nearly exact units of length k (k-mers), with $k = 1$–13 corresponding to microsatellites and $k = 1$–500 for minisatellites. Analysis of the human genome sequence showed that ~3% of the euchromatic sequence was made up of SSRs, and both SSRs and TEs are thought to be far more abundant in heterochromatin. Segmental duplications of 1–200 kb were initially shown to account for ~3% of the human genome sequence (Lander et al., 2001), and more recent results reveal that copy number variants populate the genome to an even greater extent (Kidd et al., 2008).

The evolutionary significance and the functional role that repetitive genomic elements, TEs in particular, play has long been a matter of speculation and inquiry. Once regarded as selfish, or parasitic, genomic elements with little or no phenotypic relevance (Doolittle and Sapienza, 1980; Orgel and Crick, 1980), it has since become apparent that TEs make substantial contributions to the structure, function and evolution of their host genomes (Kidwell and Lisch, 2001). Perhaps the most significant functional effect that TEs have had on their host genomes is manifest through the donation of regulatory sequences that control the expression of nearby genes (Feschotte, 2008). Studies of TE regulatory effects have focused, for the most part, on discrete well characterized regulatory elements such as transcription factor binding sites (Jordan et al., 2003; van de Lagemaat et al., 2003; Wang et al., 2007), enhancers (Bejerano et al., 2006) and alternative promoters (Dunn et al., 2003; Conley et al., 2008). A number of recent studies have also outlined the contributions of TEs to regulatory RNA genes (Smalheiser and Torvik, 2005; Borchert et al., 2006; Piriyapongsa and Jordan, 2007; Piriyapongsa et al., 2007). For this study, we sought to analyze the contribution of

Abbreviations: TE, transposable element; SSR, simple sequence repeat; TSS, transcriptional start site; GNF2, Novartis mammalian gene expresssion atlas 2.

* Corresponding author. Tel.: +1 404 385 2224; fax: +1 404 894 0519.
E-mail address: king.jordan@biology.gatech.edu (I.K. Jordan).

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.01.013

repetitive DNA to epigenetic aspects of gene regulation, specifically the relationship between repetitive DNA elements and the chromatin environment of human promoter sequences.

Genomic DNA in eukaryotes is wrapped around histone proteins and packaged into repeating subunits of chromatin called nucleosomes (Kornberg and Lorch, 1999). The importance of specific genomic sequences in determining the binding locations of nucleosomes has recently been confirmed (Segal et al., 2006). A number of factors point to a relationship between repetitive DNA elements, the local chromatin environment and epigenetic gene regulation. Densely compact heterochromatin is enriched for both TEs and SSRs in a number eukaryotic organisms (Dimitri and Junakovic, 1999). Heterochromatin functions to mitigate potentially deleterious effects associated with TEs by repressing both element transcription and ectopic recombination between dispersed element sequences (Grewal and Jia, 2007). In fact, it has been proposed that heterochromatin originally evolved to serve as a genome defense mechanism by silencing TEs (Henikoff and Matzke, 1997; Henikoff, 2000). In the plant Arabidopsis, *de novo* heterochromatin formation can be caused by insertions of TEs into euchromatin, and TEs are able to epigenetically silence genes when they are inserted nearby or inside them (Lippman et al., 2004). In other words, TEs have been shown to cause specific *in situ* changes in the chromatin environment that can spread locally and regulate gene expression in a way that is region-specific but sequence-independent (*i.e.* epigenetic).

The previously established connections between genome repeats, chromatin environment and gene regulation for model organisms, taken together with the repeat-rich nature of the human genome, suggest that repetitive sequence elements may play a role in regulating human gene expression by modulating the local chromatin environment. Specifically, we hypothesized that gene regulatory related differences in nucleosome binding at human promoter sequences are mediated in part by repetitive genomic elements. We evaluated the relationship between nucleosome binding, repetitive element promoter distributions and human gene expression to test this idea. Human proximal promoter sequences were characterized with respect to both their repetitive DNA architectures and predicted nucleosome binding affinities, and the repetitive DNA environment of the promoters was considered with respect to patterns of gene expression.

2. Materials and methods

2.1. Promoter sequence analysis

Our analysis focused on proximal promoter sequence regions, which we define for a gene as ranging from −1 kb at the 5′ end to the transcription start (TSS) at the 3′ end. We relied on the Database of Transcriptional Start Sites (DBTSS) to identify experimentally characterized TSS, based on aligned full-length cDNA sequences, in the human genome (Suzuki et al., 2002). These TSS were mapped to the March 2006 human genome reference sequence (NCBI Build 36.1) and used to extract 1 kb proximal promoter sequences as described previously (Marino-Ramirez et al., 2004; Tharakaraman et al., 2005). This procedure was used to ensure analysis of the most accurate set of human proximal promoter sequences possible. For the additional three mammalian species analyzed – chimpanzee (*Pan troglodytes*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*) – the locations of proximal promoter sequences were determined based on the 5′ most position of NCBI Refseq gene models (Pruitt et al., 2007). These positions were used to download 1 kb proximal promoter sequences from the latest respective genome builds for each organism from the UCSC Genome Browser (Karolchik et al., 2003): chimpanzee $n=24{,}170$, mouse $n=20{,}589$ and rat $n=8737$.

The program RepeatMasker (Smit et al., 1996–2004) was used to detect and annotate repetitive elements in the proximal promoter sequences. RepeatMasker was run using 500 bp of flanking sequence on either end of the proximal promoter regions analyzed to avoid edge effects in the detection of repeats. Repetitive elements detected by RepeatMasker were broken down into two main categories: interspersed repeats, also known as transposable elements (TEs), and simple sequence repeats (SSRs). SSRs may be annotated as low complexity sequences and correspond to runs of repeating *k*-mers where $k=1$–13 bp for microsatellites and $k=14$–500 for minisatellites. TEs were further divided into specific classes: LINEs, SINEs, LTR and DNA as well as specific families L1 and Alu.

Proximal promoter sequences, including 500 bp flanks, were analyzed using the Nucleosome Prediction software developed by the Segal lab (Segal et al., 2006). This software was used to calculate the probability of each nucleotide being occupied by a nucleosome in all promoter sequences. These nucleosome occupancy probabilities are based on the periodicity of dinucleotides – AA/TT/TA – that are a characteristic of genomic sequences that have been experimentally isolated as bound to nucleosomes. Predictions for the relative placement of nucleosomes along genomic sequence are further informed by a thermodynamic stability model. The nucleosome prediction model used in our analysis is based on experimentally characterized nucleosome bound sequences reported for chicken (Satchwell et al., 1986). The chicken model has been proven accurate when used on other vertebrate genomes (Segal et al., 2006). For sets of promoter sequences, nucleosome occupancy averages were calculated over each position of the 1 kb proximal promoter regions and these average values were taken as the position-specific nucleosome binding affinities (nba) reported here.

Two sets of promoter sequence randomizations were done and position-specific nucleosome binding affinities were re-calculated on the randomized sequence sets. The first randomization consisted of randomly shuffling entire 1 kb proximal promoter sequences. This has the effect of maintaining overall nucleotide composition of the promoter sequences while changing the dinucleotide composition as well as any regional nucleotide biases along the promoters. The second randomization procedure consisted on randomly shuffling non-overlapping 100 bp windows along the promoter sequences in place. This has the effect of maintaining both overall and local nucleotide compositions of the promoters while changing the dinucleotide composition.

2.2. Repeat-based promoter clustering

Human proximal promoter sequences were clustered solely based on their repetitive DNA architectures. To do this, we generated 1000-unit vectors that represent the position-specific repeat content for each promoter sequence. A discrete value was assigned to each promoter sequence position (nucleotide) in the following manner:

$$X_i = \begin{cases} 1 & \text{if the nucleotide is part of a TE sequence} \\ -1 & \text{if the nucleotide is part of a SSR sequence} \\ 0 & \text{if the nucleotide is part of a non-repetitive sequence} \end{cases}$$

where X_i represents the nucleotide at position *i*.

Promoter sequence repeat vectors were then clustered using a combination of *k*-means clustering ($k=5$, 10, 20) and Self Organized Mapping using the program Genesis (Sturn et al., 2002). We found that using *k*-means clustering with $k=5$ followed by a Self Organized Map generated the most coherent clusters in terms of the repeat content of the vectors.

2.3. Gene expression analysis

We used version 2 of the Novartis mammalian gene expression atlas (GNF2), which provides replicate Affymetrix microarray data for 44,775 probes across 79 human tissues (Su et al., 2004). GNF2