Gene 537 (2014) 29–40

Contents lists available at ScienceDirect

Gene

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

Analysis of changes in transcription start site distribution by a classification approach ☆

Kuo-ching Liang [a], Yutaka Suzuki [c], Yutaro Kumagai [d], Kenta Nakai [a,b,*]

[a] Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[b] Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba-ken 227-8561, Japan
[c] Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba-ken 227-8561, Japan
[d] Laboratory of Host Defense, World Premier International Immunology Frontier Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan

ARTICLE INFO

Article history:
Accepted 16 December 2013
Available online 31 December 2013

Keywords:
Changes of TSS distribution
Statistical approach
Time-course TSS data
Alternative promoters
Mouse dendritic cells

ABSTRACT

Change in transcription start site (TSS) usage is an important mechanism for the control of transcription process, and has a significant effect on the isoforms being transcribed. One of the goals in the study of TSS is the understanding of how and why their usage differs in different tissues or under different conditions. In light of recent efforts in the mapping of transcription start site landscape using high-throughput sequencing approaches, a quantitative and automated method is needed to process all the data that are being produced. In this work we propose a statistical approach that will classify changes in TSS distribution between different samples into several categories of changes that may have biological significance. Genes selected by the classifiers can then be analyzed together with additional supporting data to determine their biological significance. We use a set of time-course TSS data from mouse dendritic cells stimulated with lipopolysaccharide (LPS) to demonstrate the usefulness of our method.

1. Introduction

With recent advances in the understanding of complex mechanisms involved in the regulation of transcription in eukaryotes, our view of gene transcription landscape has changed dramatically. At the completion of Human Genome Project, the number of genes identified (~20,000) was far smaller than what was previously estimated (~50,000 to ~100,000). Subsequent studies have shown that in order to produce the large number of known proteins from the smaller than expected set of genes, a gene will often produce multiple unique isoforms, accomplished through several different mechanisms (Landry et al., 2003). In particular, production of multiple isoforms due to usage of alternative promoters, which was once considered as uncommon, has now been found to be a mechanism involved in the majority of human genes (Davuluri et al., 2008; The ENCODE Project Consortium, 2012). The analysis of alternative promoter has become an important topic in the study of transcriptional machinery, not only to find genes with alternative isoforms, but also to understand the evolutionary history of regulatory and transcriptional mechanism for these genes (Jordan et al., 2003).

The usage of alternative promoters can result from changes in epigenetic modifications such as DNA methylation, histone modifications and chromatin remodeling, or from changes to using different transcription factors that bind to different promoters (Hatchwell and Greally, 2007).

Abbreviations: AP-1, jun proto-oncogene; *BAI3*, brain-specific angiogenesis inhibitor 3; *CADM2*, cell adhesion molecule 2; CAGE, cap analysis of gene expression; cDNA, DNA complementary to RNA; ChIP, chromatin immunoprecipitation; *FGF14*, fibroblast growth factor 14; *GADD45g*, growth arrest and DNA-damage-inducible 45 gamma; GM-CSF, Granulocyte macrophage colony-stimulating factor; *IFIT1*, interferon-induced protein with tetratricopeptide repeats 1; *IFNR*, interferon production regulator; *IKK*, inhibitor of kappa B kinase; *IKZF1*, IKAROS family zinc finger 1; *IL6*, interleukin 6; *IL27*, interleukin 27; *IRAK*, interleukin receptor-associated kinase; *IRF*, Interferon regulator factor; *ISG15*, ubiquitin-like modifier; JAK/STAT, Janus kinase/signal transducers and activators of transcription; *KCNIP4*, K_v channel interacting protein 4; K_v, voltage-gated potassium channel; LPS, lipopolysaccharide; *IRG1*, immunoresponsive gene 1; *LRRTM4*, leucine rich repeat transmembrane neuronal 4; *LSAMP*, limbic system-associated membrane protein; *MAPK*, mitogen-activated protein kinase; *MyD88*, myeloid differentiation primary response gene 88; *NF-κB*, nuclear factor kappa B; *NFKBIZ*, nuclear factor of kappa light polypeptide gene enhancer in B cells inhibitor, zeta; *NRG3*, neuregulin 3; *NRXN1*, neurexin 1; *PCDH9*, protocadherin 9; *RIP1*, receptor (TNFRSF)-interacting serine-threonine kinase 1; *SOCS1*, suppressor of cytokine signaling 1; *STAT5a*, signal transducer and activator of transcription 5A; *TAK1*, *TGF-beta*-activated kinase 1; *TANK*, *TRAF* family member-associated *NF-κB* activator; *TBK1*, *TANK*-binding kinase 1; *TLR4*, toll-like receptor 4; *TNF-α*, Tumor necrosis factor alpha; *TRAF*, *TNF* receptor associated factors; *TRAFD1*, TRAF-type zinc finger domain containing 1; *TRIF*, toll-like receptor adaptor molecule; TSC, transcription start site cluster; TS, tissue specificity score; TSS, transcription start site; *TTR*, transthyretin; *USP18*, ubiquitin specific peptidase 18.

* Corresponding author.
E-mail address: knakai@ims.u-tokyo.ac.jp (K. Nakai).

0378-1119/$ – see front matter © 2013 The Authors. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.12.038

These two mechanisms allow genes to utilize different promoters through different means: one by blocking access to promoters, forcing transcription factors to find different binding targets, the other by using different transcription factors to bind to different targets. Genes with possible alternative promoter usage under different conditions can be found by analyzing promoter binding or transcription start site data. In this work we focus our efforts on the analysis of transcription start sites.

Many computational methods (Bajic et al., 2002; Down and Hubbard, 2002; Knudsen, 1999; Lu and Luo, 2008; Zhang, 1998) and experimental approaches such as Cap Analysis of Gene Expression (CAGE), massively parallel Paired End Tag (PET)-tagging, and TSS-Seq have been proposed to identify TSS and the corresponding promoters (Birney et al., 2007; Carninci et al., 2005; Suzuki et al., 2001; Tsuchihara et al., 2009). Recent cDNA sequencing projects such as FLJ (Ota et al., 2004) and FANTOM (Okazaki et al., 2002) have revealed that instead of utilizing only a single TSS, a promoter can be associated with a number of TSS that are distributed around its immediate neighborhood. Databases such as DBTSS (Yamashita et al., 2012) have made public up to 418 million TSS tags generated using oligo-capping and TSS-Seq techniques, providing a comprehensive overview of TSS landscapes and allowing for their comparisons in tissues under different conditions. The understanding of how the distribution of TSS changes under different conditions can help to shed further insight into the mechanism for transcribing different isoforms, and possibly their differences in functions.

Researchers have already begun to explore the relationship between TSS and transcription mechanisms. Some take the integrative approach where RNA-Seq and ChIP-Seq data are utilized in the analysis of TSS data (Yamashita et al., 2011). Others have taken the approach to analyze the significance of differences in TSS distributions. In Carninci et al. (2006), distributions of TSS are classified into four groups: 1) Single dominant peak, 2) Broad, 3) Bi- or multi-modal, and 4) Broad with dominant peak; and shapes of TSS distributions are correlated to nucleotide sequences and expression levels in human and mouse. In particular, TSS distributions with a single dominant peak are often associated with promoters with TATA-box motif, whereas broad distributions are typically found in promoter regions that have high CG content or are enriched with CpG islands (Gustincich et al., 2006). Other similar classification systems based on shapes of TSS distributions have also been proposed (Ni et al., 2010). However, while characterizing TSS distribution based on shapes of distributions has revealed some correlation with gene expression, the heuristic-based approach in determining the type of distribution shape may be a limiting factor in the uncovering of more complex relationships. In Yamashita et al. (2011), genes with TSS distribution changes in different tissues are grouped into categories based on the pattern of distribution differences, and functional overrepresentations are identified from gene ontology analysis for each category. These findings highlight the utility of not simply looking at whether differences in tag distributions exist between samples, but also taking one step further in identifying genes with specific kinds of distribution change patterns that are of interest for the given study. Furthermore, with advances in sequencing technology that allow researchers to generate TSS data in an unprecedented quantity and speed, a need has arisen for statistical methods that can automatically compare TSS distributions between different samples to identify such unique patterns.

Currently, there are many well-established methods that can be used to detect differential expression in RNA-Seq analyses. For example, in *edgeR* (Robinson et al., 2010) and *DESeq* (Anders and Huber, 2010), read count of a gene, transcript, or exon is modeled as a negative binomial distribution. In both methods the mean and variance of a negative binomial distribution are modeled as functions of the true relative abundance, due to the often lack of samples to estimate variance separately. Thus, differential expression is detected by testing the null hypothesis that the true relative abundances are the same in different samples. However, such methods pool all the reads into a single read count, and provide no information regarding how the reads are mapped to different parts of the gene/transcript/exon, and whether the distribution of these mappings are different between the samples being compared. In Kawaji et al. (2006), differences in the distribution of CAGE tags for TSS are categorized into positional bias and regional bias. For positional bias, Kruskal–Wallis one-way analysis of variance is used to test the null hypothesis that a gene's TSS distributions in different samples have the same median. For regional bias, a tissue specificity score (TS) is computed for each 21-bp window. High TS indicates that the tissue has a tissue-specific preference for TSS usage in this 21-bp region compared to other tissues. However, the Kruskal–Wallis test does not actually test for equal median or mean, and may give inaccurate results when the distribution have different shapes [Handbook of Biological Statistics]. Furthermore, while TS can locate regional differences in tag distributions, it is unclear how TS from various regions can be combined to give a single score to represent how well the overall distribution change matches the change pattern of interest. In (Zhao et al., 2011), Minimum Difference of Pair Assignments, which is similar to Earth Mover's Distance (Rubner et al., 1998), is proposed to compare the similarity between TSS distributions. However, this is again a global measure of difference between distributions, and does not contain any information on the pattern of the difference between the distributions. In (Balwierz et al., 2009), TSS loci are grouped into TSS clusters (TSC), and the likelihood was derived for two neighboring TSCs under the assumption that they have fixed relative expression. While this approach provides a comparison of the proportionality of adjacent TSCs, its computation may become overly complex when we want to make a gene-level comparison where many TSCs may be involved. Furthermore, in a multi-sample comparison, the approach cannot distinguish in which sample the change in TSC expression has occurred, and in a two-sample comparison, the likelihood function may not be accurately estimated.

In this work, we propose a classifier that can be reconfigured to test for specific patterns of TSS distribution change between tissues. We will use this approach to construct classifiers to identify genes that show differential expression in two different samples while utilizing the same TSS, and genes that exhibit TSS shift between two different samples, which we name Class 1 and Class 2 genes, respectively. The pattern of distribution change of Class 2 genes is of particular interest in our analysis of TSS, due to the possible link to alternative promoter usage, and the unavailability of such information in traditional transcriptome analysis such as microarrays and RNA-Seq. The proposed classifier analyzes TSS distributions in different samples by directly comparing their distributions in high resolution, using only a user-defined window size to merge TSS loci that might be using the same promoter. To test its usefulness, we will apply the proposed classifier to a set of TSS-Seq data for a time-course experiment on mouse dendritic cells to discover genes with possible alternative promoter usage after stimulation. It should be noted here that the classifier proposed in this paper is for single sample experiment only. While in recent years many works have argued that noise found in biological replicates is significant enough to put doubt in findings from single sample experiments as to whether statistical significant findings are due to biological phenomenon or within sample variations, when used with caution, single sample experiments can still be informative in a preliminary manner, providing candidates for more in-depth follow-up studies. In particular, many databases, including DBTSS, which is one of the largest repositories of sequencing data for TSS, contain many single sample experiments, and analyses of these data can still provide valuable knowledge about the mechanism for transcription.

1.1. TLR signaling pathways

An important motivating application for the TSS distribution change classifier is for the understanding of potential changes in TSS usage when a dendritic cell is being stimulated by lipopolysaccharide (LPS). Dendritic cells act as intermediaries between external environment