Gene 660 (2018) 92–101

Contents lists available at ScienceDirect

Gene

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

Research paper

The matrices and constraints of GT/AG splice sites of more than 1000 species/lineages

Hai Nguyen[a,b], Urmi Das[a], Benjamin Wang[a,c], Jiuyong Xie[a,*]

[a] Department of Physiology & Pathophysiology, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB R3E 0J9, Canada
[b] University of Winnipeg, Winnipeg, MB R3B 2E9, Canada
[c] University of Illinois Urbana-Champaign, IL, USA

ARTICLE INFO

Keywords:
Splice sites
Intron
Diversity
Alternative splicing

ABSTRACT

To provide a resource for the splice sites (SS) of different species, we calculated the matrices of nucleotide compositions of about 38 million splice sites from > 1000 species/lineages. The matrices are enriched of a GGTAAGT (5′SS) or $(Y)_6$N(C/t)AG(g/a)t (3′SS) overall; however, they are quite diverse among hundreds of species. The diverse matrices remain prominent even under sequence selection pressures, suggesting the existence of diverse constraints as well as U snRNAs and other spliceosomal factors and/or their interactions with the splice sites. Using an algorithm to measure and compare the splice site constraints across all species, we demonstrate their distinct differences quantitatively. As an example of the resource's application to answering specific questions, we confirm that high constraints of particular positions are significantly associated with transcriptome-wide, increased occurrences of alternative splicing when uncommon nucleotides are present. More interestingly, the abundance of alternative splicing in 16 species correlates with the average constraint index of splice sites in a bell curve. This resource will allow users to assess specific sequences/splice sites against the consensus of every Ensembl-annotated species, and to explore the evolutionary changes or relationship to alternative splicing and transcriptome diversity. Web-search or update features are also included.

1. Introduction

Splice sites (SS) demarcate exons and introns allowing the proper joining of exons during the expression of most eukaryotic genes. Their selective usage during alternative splicing produces more than one transcript from a single gene thereby contributing to transcriptomic and proteomic diversity (Black, 2003; Nilsen and Graveley, 2010). Their importance has been clearly demonstrated by splice site mutations that cause diseases (Tazi et al., 2009; Scotti and Swanson, 2016; Daguenet et al., 2015; Feng and Xie, 2013). Genomic analyses of different individual species/groups have given a glimpse of the consensus and diversity of both constitutive and alternative splice sites (Dou et al., 2006; Thanaraj and Stamm, 2003; Rogozin and Milanesi, 1997; Sibley et al., 2016; Szczesniak et al., 2013; Abril et al., 2005; Garg and Green, 2007; Burset et al., 2001). However, a centralized source for an overview of the annotated, millions of splice sites among all the currently sequenced eukaryotes remains to be created.

In biological or biomedical research, one often needs to assess the strength of particular sequences as splice sites or compare them between species, in fields such as genetics, cell biology, biochemistry or physiology. A resource with quantitative, comparable measurements of the splice site consensus and constraints of different species would be a very helpful reference. We thus compiled this resource for the consensus and diversity of the splice sites of the GT/AG introns of the Ensembl-annotated eukaryotic species as a reference for simple search or further exploration.

The GT/AG splice sites present in the majority of eukaryotic introns, characterized in humans with a consensus AGGTRAGT at the 5′ splice site and **A**$(Y)_n$NYAGG (underlined: intron start/end GT/AG, **A**: branch point, Yn: polypyrimidine tract Py, N: A,C,G or T, n:6–35) at the 3′ splice site, with variations (except the GT/AG) in other species (Sibley et al., 2016; Moore, 2000; Burge et al., 1998; Spingola et al., 1999; Mount et al., 1992; Lorkovic et al., 2000). These sequences are recognized through direct base-pairing by the snRNAs of snRNP splicing factors (U1, U2, U5 and U6, with the participation of U4) or through contact by accessory proteins such as U2AFs during the dynamic assembly of spliceosomes (Will and Luhrmann, 2011; Shi, 2017). In this report, we used the Ensembl-annotated databases to compile a complete list of the matrices and constraints of the splice sites. Since the branch point is not as easy to assess accurately as the other motifs, it is not

Abbreviations: RNA, ribonucleic acid; SS, splice site

* Corresponding author.
E-mail address: xiej@umanitoba.ca (J. Xie).

https://doi.org/10.1016/j.gene.2018.03.031
Received 17 October 2017; Received in revised form 8 March 2018; Accepted 12 March 2018
Available online 26 March 2018
0378-1119/ © 2018 Elsevier B.V. All rights reserved.

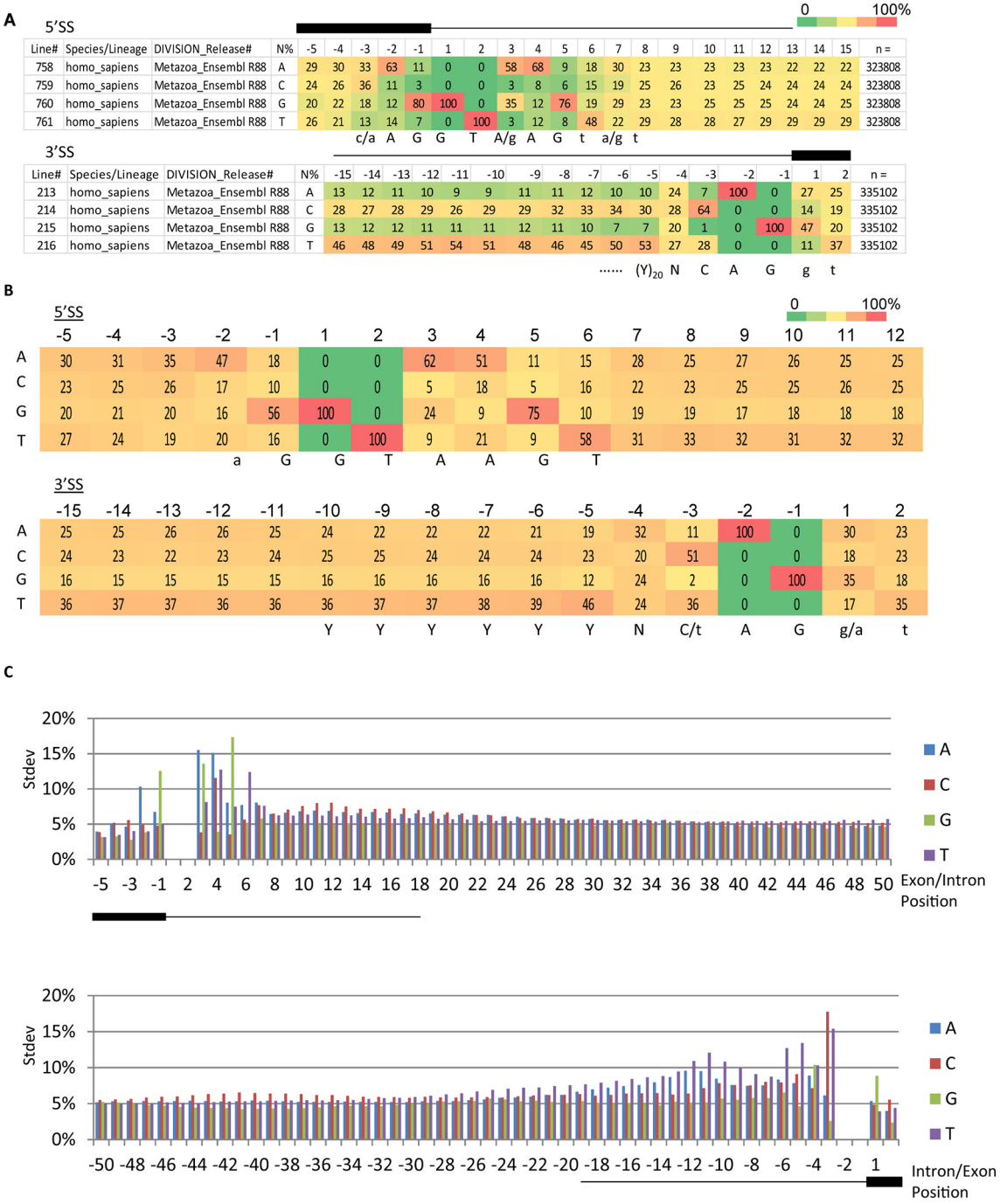

Fig. 1. Matrices and diversity of the splice sites. A. An example of the format of splice site matrices, with the human 5′ and 3′ splices sites in heat maps of the percentages of nucleotides at each position. Heavy black bar: exons, black line: introns. Below the matrices, the lowercased nucleotides are enriched well above background but < 50%, and the uppercased are above 50%. B. Average percentages of the nucleotide compositions of 30,995,943 5′ splice sites of 680 species (upper) or of 31,473,266 3′ splice sites of 682 species (lower). C. Standard deviations of the average percentages of the nucleotides in B.

included here. Also not included are the minor AT/AC introns (< 0.5%) (Burge et al., 1998; Verma et al., 2017; Wu and Krainer, 1999; Hall and Padgett, 1994; Levine and Durbin, 2001).

2. Results

2.1. The matrices of GT/AG splice sites of > 1000 eukaryotic species/lineages

We calculated the percent nucleotide compositions of the 5′ and 3′ GT/AG splice sites of 1074(5′)/1076(3′) species or their lineages/strains (hereafter 'lineage', to represent all in the same species, S_Tables I–II). An example of the resulting matrix format is shown in Fig. 1A, with the average percentages of > 300 thousands of human splice sites. The matrices are enriched of (c/a)AGGT(A/g)AGt (5′SS) or $(Y)_{20}$NCAGgt (3′SS, upstream beyond the $(Y)_{20}$ is T/A-rich), similar to those based on about 3000 human splice sites in total (Zhang, 1998). There is also a slight increase of A_7T_8, which could also participate in U1 snRNA base-pairing and splicing (Freund et al., 2005; Roca et al., 2012). However, the A_4 and G_5 of the 5′SS and the G_1 of the 3′SS are 3.5%, 4.5% and 7% less than those in the previous one on average. There is also substantial enrichment of 5′SS G_3 (35% vs 3% of C_3 or T_3),

Download English Version:

https://daneshyari.com/en/article/8645171

Download Persian Version:

https://daneshyari.com/article/8645171

Daneshyari.com