A Genome-wide Study of Dual Coding Regions in Human Alternatively Spliced Genes

Han Liang[1] and Laura F. Landweber[2]
Department of Chemistry[1] and Ecology& Evolutionary Biology[2]
Princeton University, Princeton, NJ 08544, USA

Supplementary File 2

Fig. S1

Characterization of alternatively spliced genes with multiple coding regions.

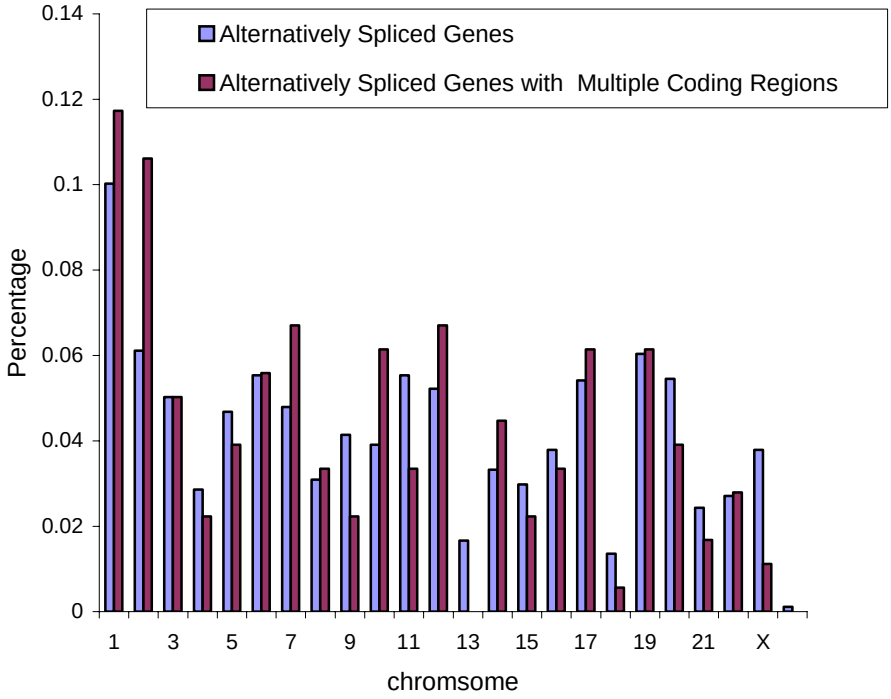
(a) The distribution over different chromosomes. The blue bars represent the percentage of alternatively spliced genes in each chromosome; the red bars represent the percentage of alternatively spliced genes with multiple coding regions in each chromosome.

(b) The distribution in different biological processes. The blue bars represent the percentage of alternatively spliced genes in each process; the red bars represent the percentage of alternatively spliced genes with multiple coding regions in each process.

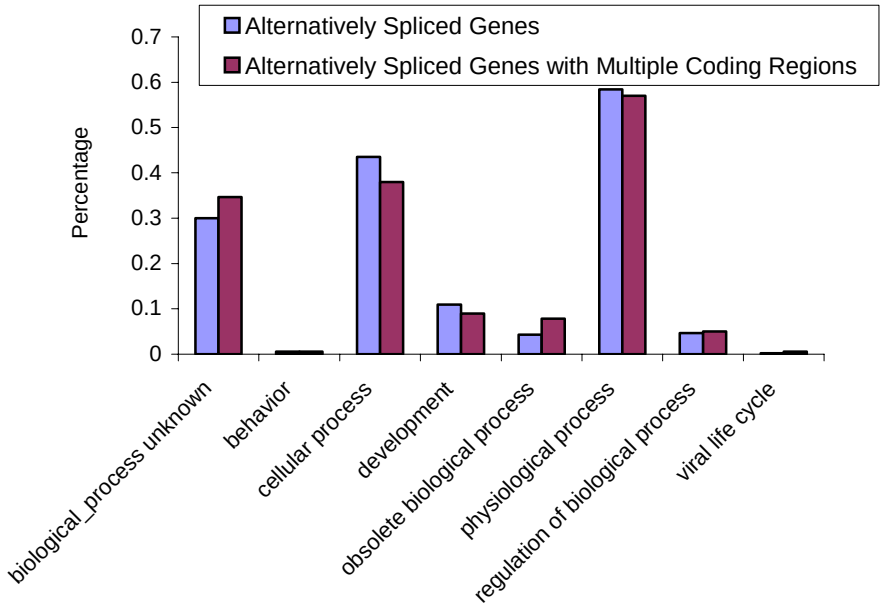
(c) The distribution in different biological function categories. The blue bars represent the percentage of alternatively spliced genes in each category; the red bars represent the percentage of alternatively spliced genes with multiple coding regions in each category.

(d) The distribution in different biological components. The blue bars represent the percentage of alternatively spliced genes in each component; the red bars represent the percentage of alternatively spliced genes with multiple coding regions in each component.

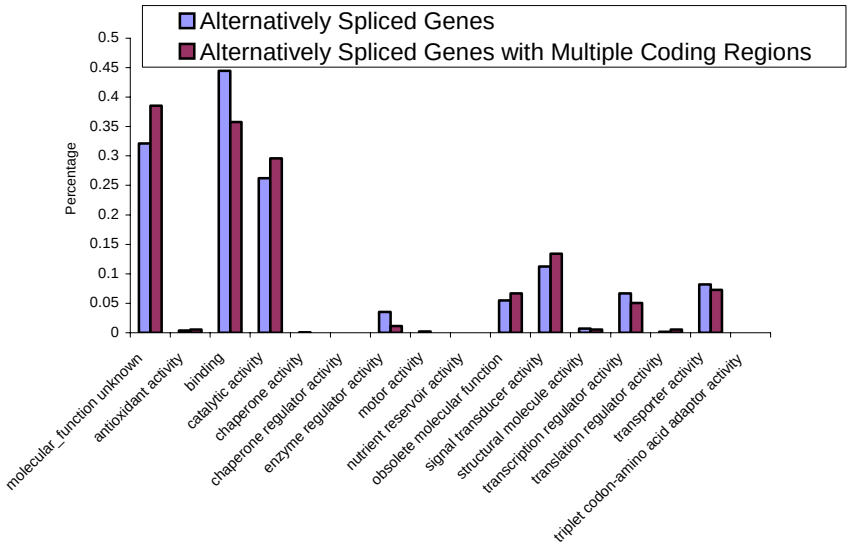
(a)


Alternatively Spliced Genes
Alternatively Spliced Genes with Multiple Coding Regions
Percentage
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1
3
5
7
9
11
13
15
17
19
21
X
chromsome

(b)


Alternatively Spliced Genes
Alternatively Spliced Genes with Multiple Coding Regions
Percentage
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
biological_process unknown
behavior
cellular process
development
obsolete biological process
physiological process
regulation of biological process
viral life cycle

(c)


Alternatively Spliced Genes
Alternatively Spliced Genes with Multiple Coding Regions
Percentage
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
molecular_function unknown
antioxidant activity
binding
catalytic activity
chaperone activity
chaperone regulator activity
enzyme regulator activity
motor activity
nutrient reservoir activity
obsolete molecular function
signal transducer activity
structural molecule activity
transcription regulator activity
translation regulator activity
transporter activity
triplet codon-amino acid adaptor activity

(d)

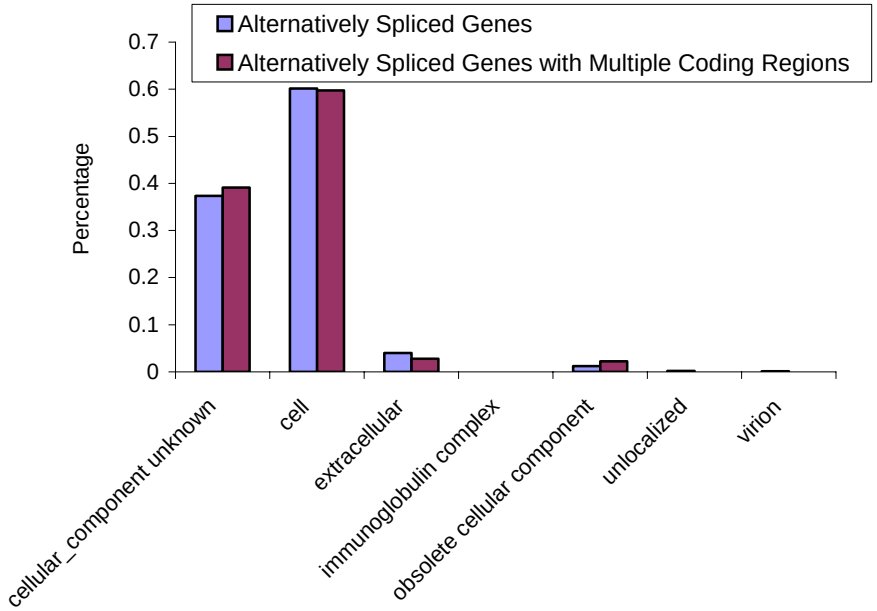

Alternatively Spliced Genes
Alternatively Spliced Genes with Multiple Coding Regions
Percentage
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
cellular_component unknown
cell
extracellular
immunoglobulin complex
obsolete cellular component
unlocalized
virion

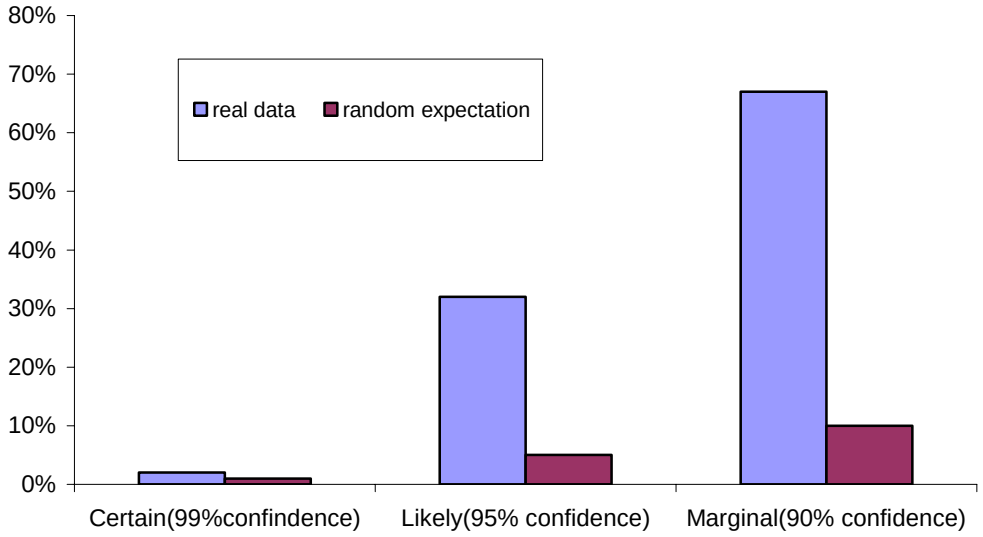
Fig. S2

(a) Statistical significance of sequence similarity search (by BLASTP) in the peptide sequences in DRFs in dual coding regions.

	Hits to PDB	No Hits
Peptide sequences in DRFs	12	90
Randomly shuffled control	4	98

$\chi^2 = 4.3$, P-value < 0.04

(b) Statistical significance of threading prediction (by FUGUE) in the peptide sequences in DRFs in dual coding regions.



$\chi^2 = 70$, P-value < 10^{-14}

Fig. S3

Schematic representation of a dual coding region in the human *UPK3B* gene.

Exons are represented by boxes, and introns are represented by connecting lines. Numbers inside the boxes refer to base pairs. Roman numerals indicate intron phases. The dual coding region is marked by a red arrow. NM_030570 and NM_182684 are RefSeq accession numbers.

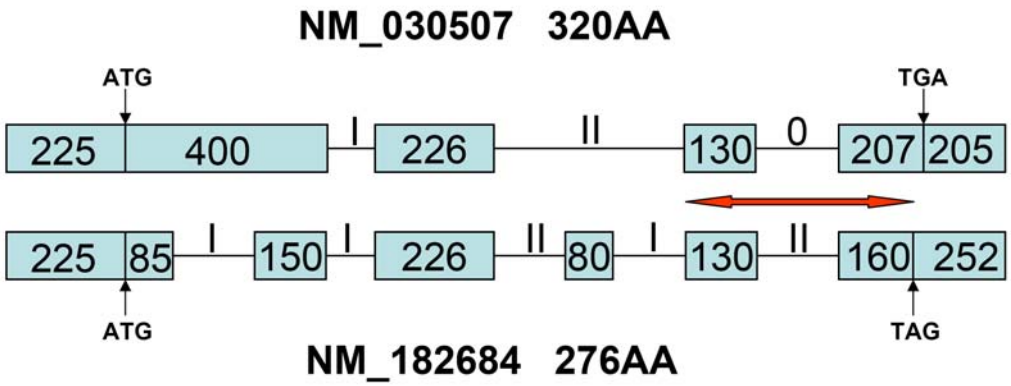


Fig. S4. The distribution of amino acids in Human DRFs corresponding to stop codons in dog DRFs.