New Insights into Polycistronic Transcripts in Eukaryotes

Haiwei Pi, PhD; Li-Wei Lee, PhD; Szecheng J. Lo, PhD

In bacteria and archaea, many functionally related genes are organized into operons in order to be transcribed and translated simultaneously. Operons are rarely seen in eukaryotes except for the *Trypanosome* and nematode, in which they are first transcribed into polycistronic transcripts but then processed into individual mature mRNAs. Recently, several researchers described the findings of polycistronic transcripts also in insects, which revised the previous thoughts that polycistronic genes were absent or few in eukaryotes. Similar to prokaryotic operons, the encoded peptides or proteins are translated simultaneously from a single polycistronic mRNA, providing new insights into the evolution of polycistronic genes. More interestingly, one type of the newly identified polycistronic genes encodes biologically important peptides composed of as few as 11 amino acids. These new findings will spur scientists to identify more small peptides in genome-solved organisms, and change the definition of coding sequences in genomic annotation. *(Chang Gung Med J 2009;32:494-8)*

Key words: evolution, genome, open reading frame (ORF), operon, polycistronic mRNA, small peptide

Polycistronic transcripts are defined as single mRNAs encoding two or more peptides or proteins (Fig. 1). The entire DNA sequence required for polycistronic gene expression is called an operon. The first and best-known operon is the *lac* operon in *Escherichia coli*, which encodes three proteins, β-galactosidase, permease, and transacetylase involved in lactose metabolism (Fig. 1A). Operons are thought to encode functionally related proteins, and commonly are found in bacteria and archaea. In contrast, most of functionally related genes of eukaryotes, i.e., α and β-hemoglobulin genes of vertebrates are not clustered and are transcribed independently. It is thought that if operons are present in eukaryotes, they are only found in limited phyla.[1,2] This simple concept may face new challenges since more forms of polycistronic mRNA are being found in various groups of eukaryotic organisms.

Operons in nematodes

An example of *Caenorhabditis elegans* operon (CEOP5428) is shown in Fig. 1B. Two functionally related genes, *fib-1* and *rps-16*, which encode the major nucleolus protein, fibrillarin, and the ribosomal small subunit protein 16, respectively, are transcribed as a single polycistronic mRNA. This mRNA is then processed into two mature monocistronic mRNAs by *trans*- and *cis*-splicing. After they are exported out of the nucleus, they can be translated as many other single protein mRNAs.

Trans-splicing is a special form of RNA splicing. It requires a short 22-nucleotide sequence, called splice leader (SL) sequences, as the donor to the 5'-end of transcripts. Two types of SL sequences, SL1 and SL2, are found in *C. elegans* and it is thought that SL2 evolved from SL1. The presence of SL sequences provides a chance for organisms to evolve a specific gene arrangement, called a *trans*-splicing

From the Department of Life Sciences, College of Medicine, Chang Gung University, Taoyuan, Taiwan.
Received: Oct. 9, 2008; Accepted: Dec. 23, 2008
Correspondence to: Prof. Szecheng J. Lo, Department of Life Sciences, College of Medicine, Chang Gung University. 259, Wunhua 1st Rd., Gueishan Township, Taoyuan County 333, Taiwan (R.O.C.). Tel: 886-3-2118800 ext. 3295; Fax: 886-3-2118392; Email: losj@mail.cgu.edu.tw

based operon. *C. elegans* and *Trypanosomes* are the first organisms identified to contain *trans*-splicing based operons.

Genome-wide screening has identified more than 1000 operons in *C. elegans*. These operons encode at least 2600 genes, comprising more than 15% of the total number of genes of *C. elegans*. The gene number within a single *C. elegans* operon ranges from 2 to 8, with an average of 2.6 genes per operon. Most operons are dicistronic. It is common that operons encode mitochondrial proteins or proteins involved in basic machinery for gene expression, such as transcription, splicing, and translation. Nevertheless, a few operons may encode proteins that are not functionally related. Interestingly, the number of operons and the total gene number encoded by operons are similar in the second genome-solved nematode, *Caenorhabditis briggsae*. *C. briggsae* evolutionally separated from *C. elegans* about 50 to 100 million years ago, indicating that the *trans*-splicing based operon is highly conserved in nematodes.

Operons in insects

Co-expression of functionally related proteins through a dicistronic transcript has been found in tomatoes, *Drosophila* and mammals.[2] For example, genes encoding γ-glutamyl kinase (*GK*) and γ-glutamyl phosphate reductase *(GPR)* in tomatoes as well as genes encoding alcohol dehydrogenase *(Adh)* and alcohol dehydrogenase related protein *(Adhr)* in the fly are transcribed into dicistronic transcripts. In contrast to the *trans*-splicing mechanism which processes the dicistronic transcript into two monocistronic mRNAs in nematodes, it has been proposed that the second gene may be translated from the same transcript using internal ribosome entry sites, ribosomal leaky scanning mechanism, or re-initiation after translation termination at the first gene. These operons probably originated many millions of years ago from prokaryotes and evolved differently from the *trans*-splicing based operons. At present time, around 100 dicistronic genes have been predicted in *Drosophila melanogaster,* which are based on the comparison of genomes of 12 species of flies using a new program of gene annotation.[3]

Apart from dicistronic transcripts, two new examples of polycistronic genes have been found in flies and other insects.[4-7] The first group was an operon with six sugar receptor genes in *Drosophila*. RT-PCR analyses suggested that it must be transcribed as a polycistronic transcript. These receptor genes span 11 kb of genomic DNA and encode gustatory receptors of sucrose, maltose, glucose, arabinose, trehalose, and glycerol.[4] In addition to the results of RT-PCR, the fact that the adjacent receptor genes are separated only by a short distance of less than 200 nucleotides, and no transcription termination signals (AAUAAA) were found in the intergenic regions supports the hypothesis that they are transcribed as a polycistronic transcript. Since dimeric or multimeric proteins are required for gustatory receptors of sugars, co-expression of six genes could result in various combinational receptors to discriminate between the different sugars. Moreover, a similar cluster arrangement was found in other odor receptors in the fly, suggesting that the polycistronic gene structure might be more common in eukaryotes than previously assumed.

The second group of polycistronic genes was originally classified as an mRNA-like ncRNA (non-coding RNA) because their open reading frames (ORFs) only encode small polypeptides of less than 50 amino acids, which were not expected to be translated. Now it turns out that they encode peptides of 11 to 32 amino acids (aa) with important functions during early embryonic patterning of insects.[5-7] One feature of this type of polycistronic operons is that they contain multiple copies of small ORF; therefore, they are also named as polycistronic peptide coding RNAs (ppcRNAs).[5] Fig. 1C shows one example of ppcRNA: the *pri (polished rice)/ tal (tarsal-less)* transcript of *Drosophila*.[6,7] It contains five predicted small ORFs, termed ORF1 to ORF5. ORF1 to ORF3 encode 11-aa peptides, and these three peptides share a core sequence of LDPTGXY, while the ORF4 encodes a 32-aa peptide containing two copies of this core sequence. Rescue experiments showed that the four ORFs were functionally redundant. ORF5 encoding a 49-aa peptide which lacks the core sequence, in contrast, was most likely not translated. Sequence comparison among different species reveals that these small ORFs are conserved among insects and crustaceans. The homologous genes in *Daphnia* and primitive insects contain two copies of the sORF, while the *Bombyx* insect has two copies of homologous genes.[6] Although it is not known whether such small ORFs are conserved across all

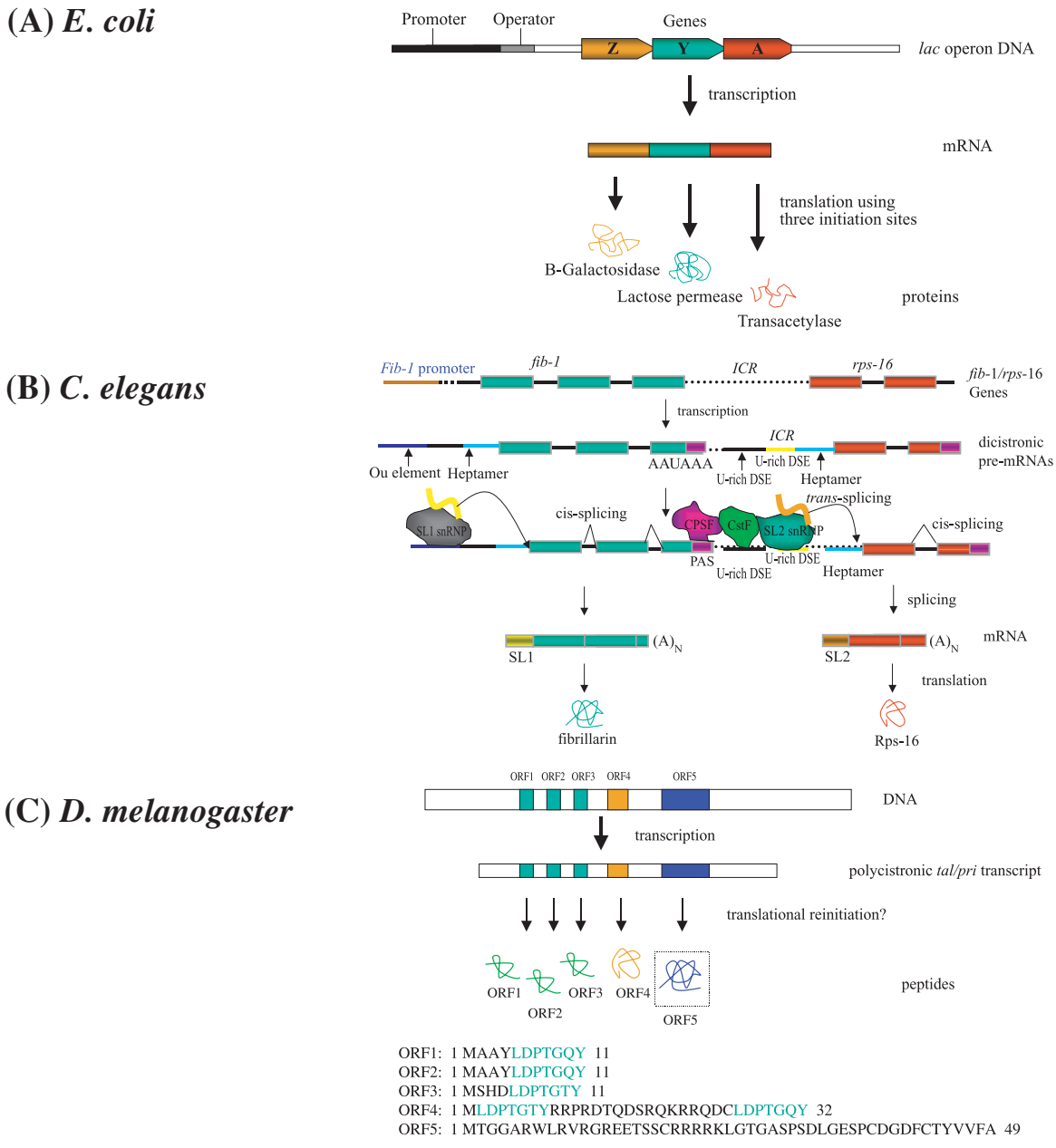

Fig. 1 Structure and arrangement of polycistronic genes in *E. coli*, *C. elegans*, and *Drosophila*. (A) The *lac* operon of *E. coli* is under the control of a single promoter and is transcribed as a polycistronic mRNA which is then translated into three proteins, β-galactosidase, permease, and transacetylase that are all involved in lactose metabolism. (B) The CEOP5428 of *C. elegans* is transcribed into a single dicistronic transcript. After *trans*- and *cis*-splicing, the SL1 is added to the 5'-end of *fib*-1 while the SL2 is added to *rps*-16 to become two individual mRNAs. The SL1 snRNP (small nuclear ribonucleoproteins) is shown at the 5'-end of polycistronic transcript while the SL2 snRNP and two other RNA binding proteins, CPSF and CstF, are shown binding to the intercistronic region of the transcript. (C) The *tal*/*pri* gene of *Drosophila* is a transcript that contains five small ORFs as indicated by their relative positions. The peptides encoded by the five ORFs are shown below as sequences of one-letter amino acid abbreviations, in which the core sequence (LDPTGXY) shared by ORF1 to ORF4 is indicated in green. It has been speculated that the translation of ORF2-ORF4 is carried out through a translational reinitiation mechanism. The longest ORF 5 encoding a 49-aa peptide, and is probably not translated (indicated by a dash-line square).

metazoans, nevertheless, it presents a great challenge to design a new program to annotate ppcRNA genes more broadly in genome-solved organisms.

Conclusion

In addition to the polycistronic transcript which encodes two or more different proteins or peptides on a single transcript, several other structures are known where multiple proteins can be made from a single transcript. For instance, translational frame-shift in a retrovirus produces the gag and gag-pol proteins. Alternative splicing is another example of multiple proteins encoded from the same segment of DNA. These phenomena make it more difficult and complicated to annotate the exact number of protein-coding genes in organisms. The recent discoveries of biologically active small peptides encoded by polycistronic RNA further revise the original concept of functional ORFs and provide a new avenue for the annotation and functional analyses of genome-solved organisms.

Acknowledgements

We would like to thank Simon Silver (a visiting professor of Department of Life Sciences, Chang Gung University) for critically reviewing the manuscript. Grants for *C. elegans* and *Drosophila* research were supported by CMRP 33031and CMRP32037 from the Chang Gung Memorial Hospital to HP and SJL, respectively.

REFERENCES

1. Nimmo R, Woollard A. Widespread organisation of *C. elegans* genes into operons: fact or function? BioEssays 2002;24:983-7.
2. Blumenthal T. Operons in eukaryotes. Brief Funct Genomic Proteomic 2004;3:199-211.
3. Lin MF, Carlson JW, Crosby MA, Matthews BB, et al. Revisiting the protein-coding gene catalog of *Drosophila melanogaste*r using 12 fly genomes. Genome Res 2007;17:1823-36.
4. Slone J, Daniels J, Amrein H. Sugar receptors in *Drosophila*. Curr Biol 2007;17:1809-16.
5. Savard J, Maqques-Souza H, Aranda M, Tautz D. A segmentation gene in *Tribolium* produces a polycistronic mRNA that codes for multiple conserved peptides. Cell 2006;126:559-69.
6. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLos Biol 2007;5:1052-62.
7. Kondo T, Hashomoto Y, Kato K, Inagaki S, Hayashi S, Kageyama Y. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nature Cell Biol 2007;9:660-5.

眞核生物多基因轉錄物之新見

皮海薇 李立緯 羅時成

在細菌及古菌中，許多功能相關之基因連在一起組成操作組，以便同時進行轉錄及轉譯。相對地，除了錐蟲和線蟲，眞核生物較少有操作組基因結構。但它們的操作組基因在轉錄後進行同位及異位剪裁形成單獨成熟 mRNA 再進行轉譯，有別於原核生物轉錄與轉譯同時進行。近來，一些有關昆蟲多基因轉錄物的發現，改變了過去認爲在眞核生物缺乏或很少操作組基因的概念。這些昆蟲的操作組基因有如原核生物一樣，轉錄後同時進行轉譯，提供了操作組基因演化的新觀念。更重要的是，一些多基因轉錄物只合成 11 個胺基酸並具生物功能的短胜肽。這些新發現將加速科學家在基因解碼的生物中找尋更多作小胜肽的基因，同時也改變了對基因開放編閱框合成胜肽至少要超過 50 個胺基酸的定義。(長庚醫誌 2009;32:494-8)

關鍵詞：演化，基因體，開放編閱框，操作組基因，多基因 mRNA，小胜肽

長庚大學 醫學院 生命科學系
受文日期：民國97年10月9日；接受刊載：民國97年12月23日
通訊作者：羅時成教授，長庚大學 醫學院 生命科學系。桃園縣333龜山鄉文化一路259號。Tel.: (03)2118800轉3295; Fax: (03)2118392; E-mail: losj@mail.cgu.edu.tw