DATABASE
The Journal of Biological Databases and Curation

Database, 2018, 1–6
doi: 10.1093/database/bay089
Perspective/Opinion

Perspective/Opinion

Improved ontology for eukaryotic single-exon coding sequences in biological databases

Roddy Jorquera[1,2], Carolina González[1], Philip Clausen[3], Bent Petersen[3,4] and David S. Holmes[1,5,*]

[1]Center for Bioinformatics and Genome Biology, Fundacion Ciencia & Vida, Avenida Zañartu 1482, Ñuñoa, Santiago, Chile, [2]Facultad de Ciencias Biologicas, Universidad Andres Bello, Santiago, Chile, [3]Department of Bio and Health Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark, [4]Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia and [5]Centro de Genómica y Bioinformática Facultad de Ciencias, Universidad Mayor, Santiago, Chile

*Corresponding author: Tel: +56 2 22398969; Fax: +56 2 22372259; Email: dsholmes2000@yahoo.com

Citation details: Jorquera,R., González,C., Clausen,P. *et al*. Improved ontology for eukaryotic single-exon coding sequences in biological databases. *Database* (2018) Vol. 2018: article ID bay089; doi:10.1093/database/bay089

Received 3 April 2018; Revised 27 July 2018; Accepted 30 July 2018

Abstract

Efficient extraction of knowledge from biological data requires the development of structured vocabularies to unambiguously define biological terms. This paper proposes descriptions and definitions to disambiguate the term 'single-exon gene'. Eukaryotic Single-Exon Genes (SEGs) have been defined as genes that do not have introns in their protein coding sequences. They have been studied not only to determine their origin and evolution but also because their expression has been linked to several types of human cancer and neurological/developmental disorders and many exhibit tissue-specific transcription. Unfortunately, the term 'SEGs' is rife with ambiguity, leading to biological misinterpretations. In the classic definition, no distinction is made between SEGs that harbor introns in their untranslated regions (UTRs) versus those without. This distinction is important to make because the presence of introns in UTRs affects transcriptional regulation and post-transcriptional processing of the mRNA. In addition, recent whole-transcriptome shotgun sequencing has led to the discovery of many examples of single-exon mRNAs that arise from alternative splicing of multi-exon genes, these single-exon isoforms are being confused with SEGs despite their clearly different origin. The increasing expansion of RNA-seq datasets makes it imperative to distinguish the different SEG types before annotation errors become indelibly propagated in biological databases. This paper develops a structured vocabulary for their disambiguation, allowing a major reassessment of their evolutionary trajectories, regulation, RNA

processing and transport, and provides the opportunity to improve the detection of gene associations with disorders including cancers, neurological and developmental diseases.

Database URL: http://www.sinex.cl

Introduction

Next-Generation Sequencing (NGS) and other high-throughput technologies are generating vast amount of biological data that are a challenge for downstream data mining. To help address this problem, progress has been made in the development of structured vocabularies and ontologies that facilitate computational data annotation, retrieval and interpretation (1–4). However, no such structured vocabulary exists for describing the different types of eukaryotic single-exon coding sequences (CDSs), causing confusion and leading to misinterpretation of their evolutionary origins as well as their regulation and function within eukaryotic genomes.

This situation is being further exacerbated by the discovery that Single-Exon Isoforms (SEIs) are being misannotated as arising from Single-Exon Genes (SEGs) rather than from Multi-Exon Genes (MEGs) by alternative splicing as is the case. It is urgent to draw the attention of the scientific community to such problems before such errors become indelibly propagated in biological databases. 'This work attempts' to address these concerns by proposing a workflow for developing a structured vocabulary (see also the glossary box) and ontology to describe and distinguish the various types of eukaryotic single-exon CDSs; one that is resilient and inclusive but flexible enough to accommodate new advances in gene interpretation. The workflow is presented as a directed acyclic graph with transitive rules for deriving ontological descriptors (Figure 1).

Single-exon genes

Eukaryotic genes are usually interrupted by intragenic, non-protein coding regions termed introns that are removed by RNA splicing during maturation of the final RNA product. However, more than 2000 protein-coding genes in human genome have been shown to lack introns and have been termed SEGs, defined as a nuclear, protein-coding gene that lack introns in their CDSs (5). This definition excludes genes that generate functional RNAs such as tRNA, rRNA and long non-coding RNAs. A large proportion of genes encoding G-protein-coupled receptors (GPCRs), especially the olfactory receptors, the major subfamily of class A GPCRs (6), and genes encoding canonical histones (7) are known to be SEGs. It has been proposed that the expression of many human SEGs is linked to several types of cancer and neurological and developmental disorders (8). In addition, the expression of some SEGs is testis and neuro-specific (8, 9). These discoveries highlight the importance of studying SEGs to uncover properties and evolutionary trajectories that underlie their relationships with both pathologies and normal phenotypes.

As shown in Figure 1 (left side), SEGs can be divided into the following two main groups: (i) SEGs having introns in their untranslated region (UTR), so-called "UTR intron-containing SEGs" (uiSEGs) and (ii) SEGs lacking introns in the entire gene, termed 'intronless genes' (IGs) (10) (see also the glossary box for definitions).

Examples of uiSEGs with experimentally validated phenotypes of clinical relevance are as follows: ERAS, embryonic stem cell expressed Ras; NEUROD2, neuronal differentiation 2; and NFIL3, nuclear factor, interleukin-3-regulated protein (5, 11, 12).

Examples of SEGs that can be classified as IGs are as follows: Reprimo (RPRM), a TP53 dependent G2 arrest mediator (10, 12, 13); CDR1, cerebellar degeneration-related protein; and NPBWR2, neuropeptides B/W receptor type 2 (10, 11).

Ontological ambiguity

There is a significant operational problem in distinguishing between uiSEGs and IGs. Today, most genes are predicted based mainly on bioinformatics analyses. Inparticular, SEGs are identified based on CDS (protein coding) gene identifiers in annotated genomes (5, 14–16) and these identifiers do not include information from the UTR of genes, resulting in the identification of all SEGs as IGs, regardless of the presence or absence of introns in their UTRs (14–18) with the result that the terms 'SEG' and 'IG' are rife with ambiguity.

The distinction between uiSEGs versus IGs is important to make because the presence or absence of introns in the UTR, and whether the intron is in the 5′ UTR or 3′ UTR, can impact transcriptional regulation and post-transcriptional processing (19–21). For example, RNA transcripts derived from 5′ UTR intron-containing genes (35% of all human transcripts) (22) are exported from the nucleus by a splicing-dependent mechanism involving the TRanscription and EXport (TREX) complex (20). TREX is a conserved multi-subunit complex that is

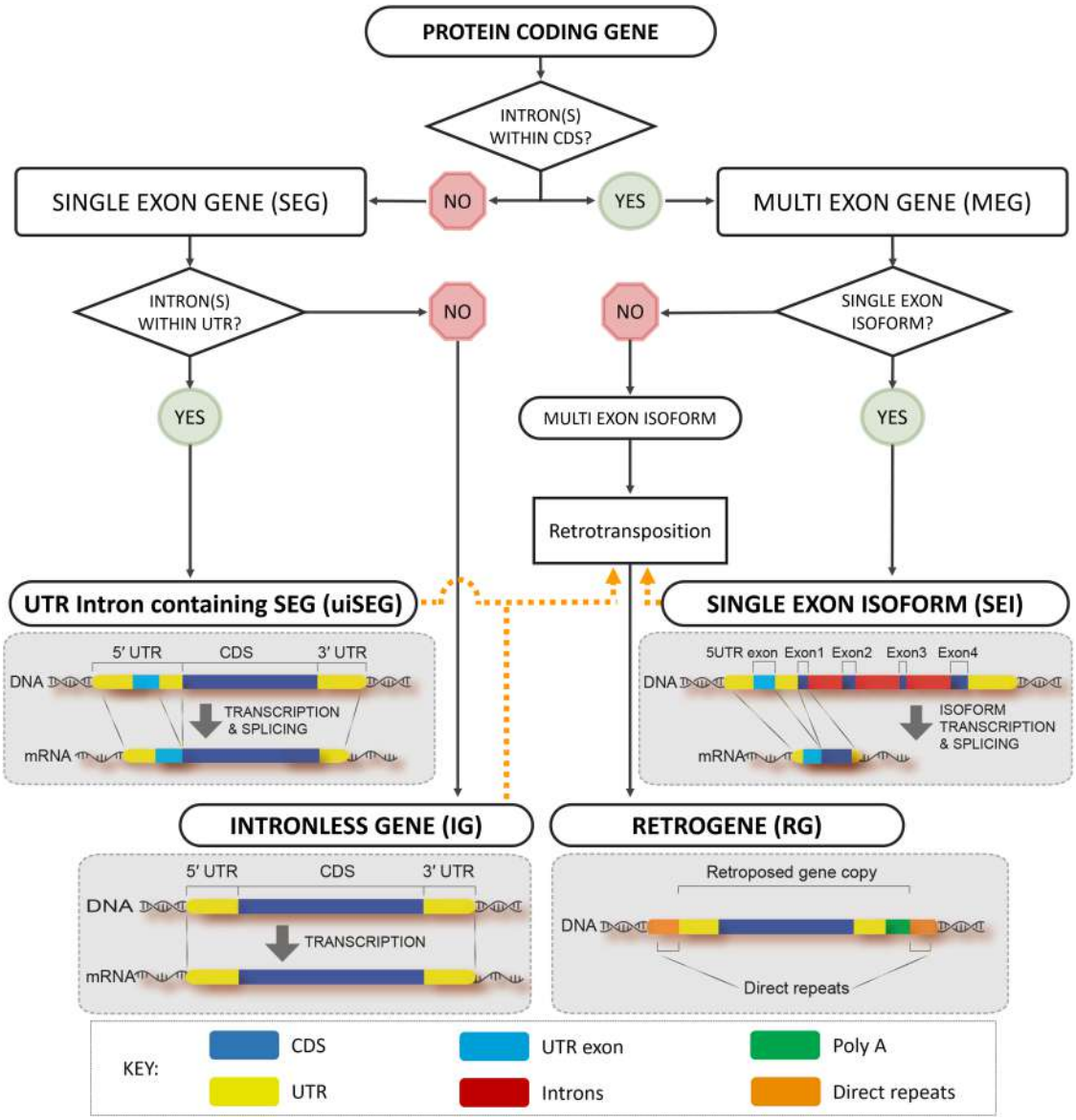

Figure 1. Workflow for developing a structured vocabulary (ontology) to distinguish different types of single exon CDSs. UTR intron containing genes (uiSEG); Intronless Gene (IG); Single Exon Isoform (SEI) and Retrogene (RG). Orange dotted lines connect genes to retrogenes via retrotransposition processes. CDS = protein coding region; UTR = untranslated region. The orange dotted lines connect RGs with potential parental genes. RGs are synonymous with IGs only when the retrotransposition origin of these sequences can be implied.

recruited to the 5′ end of mRNA transcripts by capping and splicing events (23). The TREX export pathway has been implicated in several diseases (23). On the other hand, RNA transcripts lacking 5′ UTR introns, such as IGs, can harbor specific sequences in their early coding regions (24) that promote an alternative mRNA nuclear export pathway (20). The majority of these mRNAs encode secreted, membrane-bound or mitochondrial proteins (25). The presence or absence of introns in the 5′ UTR of genes has also been shown to affect transcriptional activity, protein accumulation and determination of tissue-specific transcription (26, 27), providing another example of the need for disambiguation of uiSEGs from IGs.

Less is known about the function of 3′ UTR intron-containing genes, although some have been shown to target mRNA for degradation by the nonsense-mediated decay pathway (19, 27). It has also been observed that 3′ UTR introns can modulate gene expression at multiple levels and has been associated with miRNA targets (28) and specific mRNA localization in neurons (29, 30).

These data together suggest that in order to clearly distinguish between uiSEGs and IGs, an experimental validation of UTR introns should be considered.

SEGs are not always synonymous with retrogenes

Retrogenes (RGs) (Figure 1, bottom right-hand side) arise by retrotranscription of mRNA followed by insertion of the resulting DNA copy into the genome (31, 32). Most RGs are believed to have originated from multi-exon (intron-containing) parental genes (33) (Figure 1) although theoretically they could also arise from mRNA derived from uiSEGs and IG transcripts (Figure 1, dotted arrows). RGs are generally thought to be intronless but recently, intron-containing RGs derived from retrotransposition of parental isoforms with retained introns (31, 34) and MEGs including RG-derived exons (31) has been discovered.

Although many SEGs are thought to be RGs (14), molecular mechanisms distinct to retrotransposition have been proposed for the origin of SEGs, such as *de novo* origin (35), DNA-based duplication from intron-containing genes (36) and intron loss, among others (37, 38). Clusters of genes encoding canonical replication-dependent histones that evolved to possess a specialized 3′ processing pathway that is coupled to DNA replication (39, 40) are remarkable examples of SEGs that are not RGs (31, 41).

Initially, RGs may exhibit sequence signatures of their mRNA origin and genome insertion such as poly-A tails and direct repeats. But these molecular signals may become blurred over time that could impact the annotation of RGs. Furthermore, many RGs contain mutations that may render them inactive and are termed ‘processed pseudogenes’ (31, 42).

These data together suggest that despite their similar molecular structure, RGs are synonymous with SEGs only when the retrotransposition origin of the sequence can be implied.

Single-exon isoforms

A previously undefined class of mRNA transcripts, which we term SEIs, is similar to mRNAs derived from SEGs (9). However, unlike mRNA from SEGs, SEIs originate by alternate splicing of RNA transcribed from MEGs (Figure 1, top right-hand side and Supplementary Figure S1) in which only one protein coding exon is retained in the mature mRNA (SEI). The gene structure, transcriptional regulation and evolutionary origin of SEIs differ radically from SEGs.

The emerging problem is that SEGs and SEIs are being confused because bioinformatics techniques for gene identification based on CDS annotation do not take into account the underlying gene structure. To aid in the resolution of this problem, we suggest that a closer examination of the underlying gene structure at the DNA level would facilitate the distinction between a single-exon transcript that arose from a SEG (uiSEG or IG) and one that arose from a MEG by alternate splicing (SEI).

Analysis of the human genome (GRCh38.p9 Refseq assembly GCF_000001405.35) using a genomics protocol described in Supplementary Material, predicted 2783 putative SEGs of which 621 (22.3%) correspond to predicted SEIs and 38 sequences (1.4%) were identified as potential annotation errors. Using manually curated methods, 687 sequences have been previously predicted to be IGs (10), these data together let us to estimate that an approximate of 1437 (51.6%) predicted SEGs—using our protocol—correspond to uiSEGs (Supplementary Figure S2).

The estimated proportion of uiSEGs, IGs and SEIs within the human genome is about 2:1:1 and the latter group could have caused bias in earlier studies of SEG function and evolution in the human genome (5, 8, 14–17, 43). The unambiguous identification of SEIs also provides opportunities to study novel aspects of alternative splicing and intron evolution, gene evolution, RNA editing, nuclear export pathways and transcriptional regulation.

Examples of SEIs with experimental validation and clinical relevance that have previously been classified as SEGs include the following: BDNF, brain-derived neurotrophic factor isoform a preproprotein; HIC1, hypermethylated in cancer 1 protein isoform 1; and GDNF, glial cell line-derived neurotrophic factor isoform (11, 12).

Implementation

The proposed structured vocabulary has been submitted to a publicly accessible ontology project: Sequence Ontology (SO, http://www.sequenceontology.org/).

Conclusions

- Eukaryotic SEGs can be divided into the following two main groups: IGs and uiSEGs. This distinction is important to make because the presence or absence of introns in the UTR, can impact transcriptional regulation and post-transcriptional processing of the mRNA. In order to understand their evolution and biology, it is important to clearly distinguish between the different types of genes.
- Despite similar molecular architecture between some RGs and SEGs, these terms are synonymous only when the retrotransposition origin of the sequence can be implied.
- A previously undefined class of mRNA transcripts, which we term SEIs, is being incorrectly annotated as SEGs,

exacerbating the operational problem in distinguishing the different types of SEGs.

- With the increase of RNA sequencing approaches, this confusion is likely to become aggravated if it is not solved. This task is urgent so as to reduce the propagation of erroneous gene annotations in biological databases.
- A structured vocabulary with unambiguous definitions of SEGs, IGs and SEIs provides opportunities to study novel aspects of alternative splicing and intron evolution, gene evolution, RNA editing, nuclear export pathways and transcriptional regulation. It also provides the opportunity to improve the detection of gene associations with disorders including cancers, neurological and developmental diseases.

Supplementary data

Supplementary data are available at *Database* Online.

Funding

Fondecyt (1090451, 1130683, 1181717) and Basal (AFB 170004).

Conflict of interest. None declared.

References

1. Hoehndorf,R., Schofield,P.N. and Gkoutos,G.V. (2015) The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinform.*, **16**, 1069–1080.
2. Spasic,I., Schober,D., Sansone,S.A. *et al.* (2008) Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC Bioinformatics*, **9** Suppl 5, S5.
3. Smith,B., Ashburner,M., Rosse,C. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
4. He,Y., Xiang,Z., Zheng,J. *et al.* (2018) The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. *J. Biomed. Semantics*, **9**, 3.
5. Jorquera,R., Ortiz,R., Ossandon,F. *et al.* (2016) SinEx DB: a database for single-exon coding sequences in mammalian genomes. *Database (Oxford)*, **2016**, baw095.
6. Tine,M., Kuhl,H., Beck,A. *et al.* (2011) Comparative analysis of intronless genes in teleost fish genomes: insights into their evolution and molecular function. *Mar. Genomics*, **4**, 109–119.
7. Draizen,E.J., Shaytan,A.K., Marino-Ramirez,L. *et al.* (2016) HistoneDB 2.0: a histone database with variants–an integrated resource to explore histones and their variants. *Database (Oxford)*, **2016**, baw014.
8. Grzybowska,E.A. (2012) Human intronless genes: functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochem. Biophys. Res. Commun.*, **424**, 1–6.
9. Shabalina,S.A., Ogurtsov,A.Y., Spiridonov,A.N. *et al.* (2010) Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. *Mol. Biol. Evol.*, **27**, 1745–1749.
10. Louhichi,A., Fourati,A. and Rebai,A. (2011) IGD: a resource for intronless genes in the human genome. *Gene*, **488**, 35–40.
11. Yates,A., Akanni,W., Amode,M.R. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
12. Yang,I.S., Son,H., Kim,S. *et al.* (2016) ISOexpresso: a web-based platform for isoform-level expression analysis in human cancer. *BMC Genomics*, **17**, 631.
13. Bernal,C., Aguayo,F., Villarroel,C. *et al.* (2008) Reprimo as a potential biomarker for early detection in gastric cancer. *Clin. Cancer Res.*, **14**, 6264–6269.
14. Sakharkar,M.K., Chow,V.T., Ghosh,K. *et al.* (2005) Computational prediction of SEG (single-exon gene) function in humans. *Front Biosci.*, **10**, 1382–1395.
15. Yan,H., Jiang,C., Li,X. *et al.* (2014) PIGD: a database for intronless genes in the Poaceae. *BMC Genomics*, **15**, 832.
16. Zou,M., Guo,B. and He,S. (2011) The roles and evolutionary patterns of intronless genes in deuterostomes. *Comp. Funct. Genomics*, **2011**, 680673.
17. Agarwal,S.M. and Gupta,J. (2005) Comparative analysis of human intronless proteins. *Biochem. Biophys. Res. Commun.*, **331**, 512–519.
18. Sakharkar,K.R., Sakharkar,M.K., Culiat,C.T. *et al.* (2006) Functional and evolutionary analyses on expressed intronless genes in the mouse genome. *FEBS Lett.*, **580**, 1472–1478.
19. Bicknell,A.A., Cenik,C., Chua,H.N. *et al.* (2012) Introns in UTRs: why we should stop ignoring them. *Bioessays*, **34**, 1025–1034.
20. Palazzo,A.F., Mahadevan,K. and Tarnawsky,S.P. (2013) ALREX-elements and introns: two identity elements that promote mRNA nuclear export. *Wiley Interdiscip. Rev. RNA*, **4**, 523–533.
21. Palazzo,A.F. and Akef,A. (2012) Nuclear export as a key arbiter of "mRNA identity" in eukaryotes. *Biochim. Biophys. Acta*, **1819**, 566–577.
22. Cenik,C., Derti,A., Mellor,J.C. *et al.* (2010) Genome-wide functional analysis of human 5′ untranslated region introns. *Genome Biol.*, **11**, R29.
23. Heath,C.G., Viphakone,N. and Wilson,S.A. (2016) The role of TREX in gene expression and disease. *Biochem. J.*, **473**, 2911–2935.
24. Cenik,C., Chua,H.N., Singh,G. *et al.* (2017) A common class of transcripts with 5′-intron depletion, distinct early coding sequence features, and N1-methyladenosine modification. *RNA*, **23**, 270–283.
25. Cenik,C., Chua,H.N., Zhang,H. *et al.* (2011) Genome analysis reveals interplay between 5′ UTR introns and nuclear mRNA export for secretory and mitochondrial genes. *PLoS Genet.*, **7**, e1001366.
26. Moabbi,A.M., Agarwal,N., El Kaderi,B. *et al.* (2012) Role for gene looping in intron-mediated enhancement of transcription. *Proc. Natl. Acad. Sci. USA*, **109**, 8505–8510.
27. Barrett,L.W., Fletcher,S. and Wilton,S.D. (2012) Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell Mol. Life Sci.*, **69**, 3613–3634.
28. Tan,S., Guo,J., Huang,Q. *et al.* (2007) Retained introns increase putative microRNA targets within 3′ UTRs of human mRNA. *FEBS Lett.*, **581**, 1081–1086.

29. Mayr,C. (2016) Evolution and biological roles of alternative 3′ UTRs. *Trends Cell Biol*., **26**, 227–237.
30. Sharangdhar,T., Sugimoto,Y., Heraud-Farlow,J. *et al*. (2017) A retained intron in the 3′ UTR of Calm3 mRNA mediates its Staufen2- and activity-dependent localization to neuronal dendrites. *EMBO Rep*.**18**, 1762–1774.
31. Kubiak,M.R. and Makalowska,I. (2017) Protein-coding genes' retrocopies and their functions. *Viruses*, **9**, 80.
32. Navarro,F.C. and Galante,P.A. (2015) A genome-wide landscape of retrocopies in primate genomes. *Genome Biol. Evol*., **7**, 2265–2275.
33. Navarro,F.C. and Galante,P.A. (2013) RCPedia: a database of retrocopied genes. *Bioinformatics*, **29**, 1235–1237.
34. Zhang,C., Gschwend,A.R., Ouyang,Y. and Long,M. (2014) Evolution of gene structural complexity: an alternative-splicing-based model accounts for intron-containing retrogenes. *Plant Physiol*., **165**, 412–423.
35. Knowles,D.G. and McLysaght,A. (2009) Recent de novo origin of human protein-coding genes. *Genome Res*., **19**, 1752–1759.
36. Zhang,Y.E., Vibranovski,M.D., Krinsky,B.H. *et al*. (2011) A cautionary note for retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single-exon genes. *Bioinformatics*, **27**, 1749–1753.
37. Roy,S.W. and Gilbert,W. (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet*., **7**, 211–221.
38. Chen,S., Krinsky,B.H. and Long,M. (2013) New genes as drivers of phenotypic evolution. *Nat. Rev. Genet*., **14**, 645–660.
39. Rattray,A.M.J. and Müller,B. (2012) The control of histone gene expression. *Biochem. Soc. Trans*., **40**, 880–885.
40. Ederveen,T.H., Mandemaker,I.K. and Logie,C. (2011) The human histone H3 complement anno 2011. *Biochim. Biophys. Acta*, **1809**, 577–586.
41. Erives,A.J. (2017) Phylogenetic analysis of the core histone doublet and DNA topo II genes of Marseilleviridae: evidence of proto-eukaryotic provenance. *Epigenetics Chromatin*, **10**, 55.
42. Pink,R.C., Wicks,K., Caley,D.P. *et al*. (2011) Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA*, **17**, 792–798.
43. Sakharkar,M.K., Chow,V.T., Chaturvedi,I. *et al*. (2004) A report on single-exon genes (SEG) in eukaryotes. *Front Biosci*., **9**, 3262–3267.