Cell Genomics

CellPress
OPEN ACCESS

Perspective

Not all exons are protein coding: Addressing a common misconception

Julie L. Aspden,[1,2,3] Edward W.J. Wallace,[4] and Nicola Whiffin[5,6,*]
[1]School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK
[2]LeedsOmics, University of Leeds, Leeds LS2 9JT, UK
[3]Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds LS2 9JT, UK
[4]Institute for Cell Biology and Centre for Engineering Biology, School of Biological Sciences, The University of Edinburgh, Edinburgh EH9 3BF, UK
[5]Big Data Institute, University of Oxford, Oxford OX3 7LF, UK
[6]Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK
*Correspondence: nwhiffin@well.ox.ac.uk

https://doi.org/10.1016/j.xgen.2023.100296

SUMMARY

Exons are regions of DNA that are transcribed to RNA and retained after introns are spliced out. However, the term “exon” is often misused as synonymous to “protein coding,” including in some literature and textbook definitions. In contrast, only a fraction of exonic sequences are protein coding (<30% in humans). Both exons and introns are also present in untranslated regions (UTRs) and non-coding RNAs. Misuse of the term exon is problematic, for example, “whole-exome sequencing” technology targets <25% of the human exome, primarily regions that are protein coding. Here, we argue for the importance of the original definition of an exon for making functional distinctions in genetics and genomics. Further, we recommend the use of clearer language referring to coding exonic regions and non-coding exonic regions. We propose the use of coding exome sequencing, or CES, to more appropriately describe sequencing approaches that target primarily protein-coding regions rather than all transcribed regions.

WHAT IS AN EXON?

A region of eukaryotic DNA that is transcribed into RNA may contain any number of exons and introns. The sections that are retained in the mature RNA molecule after RNA splicing are termed exons, with the regions that are removed referred to as introns (Figure 1A).[1] The term exon refers to these regions in both DNA and RNA sequences.

Exons and introns were named by Walter Gilbert in 1978.[2]

> The notion of the cistron, the genetic unit of function that one thought corresponded to a polypeptide chain, now must be replaced by that of a transcription unit containing regions which will be lost from the mature messenger - which I suggest we call introns (for intragenic regions) - alternating with regions which will be expressed - exons.

It is important to note that annotation of an exon is transcript specific. A region that is exonic in one transcript may be intronic in another, as a result of alternative splicing.[3] For transcripts where splicing does not occur, the entire transcribed sequence comprises a single exon.

Regions of a transcript that encode protein are always within exons, but protein-coding transcripts also contain exonic regions that do not form part of the final coding sequence, termed untranslated regions (UTRs; Figure 1B). Minimally, the first and last protein-coding exons of a transcript also contain sections of UTRs. Interestingly, the discovery of untranslated RNA regions (by 1970[4,5]) predated the discovery of introns (1977[6,7]). Since 1978,[8] the term exon has been used for both protein-coding and non-coding regions to distinguish those that are retained after splicing and is still used among molecular biologists in this way.[1,9] Additionally, an exon (or part of an exon) may be protein coding in one transcript but non-coding in another, for example, due to alternative start codon usage.

Non-coding RNAs (>25,000 of which are known to exist in humans) are often subject to splicing but are not translated, i.e., they are composed of non-coding exons (Figure 1C).

A common misconception

Across the field of genetics, there is ambiguity in the use of the term exon, which is often used to refer to regions of the genome that code directly for protein, while introns are often classed as the non-coding parts of genes. Several textbooks refer to exons as simply protein coding,[10,11] which is incorrect, and fail to mention UTRs and non-coding RNAs when defining exons and introns. Many public resources also perpetuate this common misuse: for example, the Oxford English Dictionary’s second edition defines an exon as “a section of a DNA or RNA molecule that codes for a protein, in cases where such sections are separated by non-coding ones.”[12] The definition offered by Google is similar (Figure S1). These definitions are incorrect.[13]

Only a fraction of exonic sequences code for protein

To illustrate the impact of distinguishing exonic from protein-coding sequences, we quantified the amount of an exonic sequence that codes for protein across six different organisms

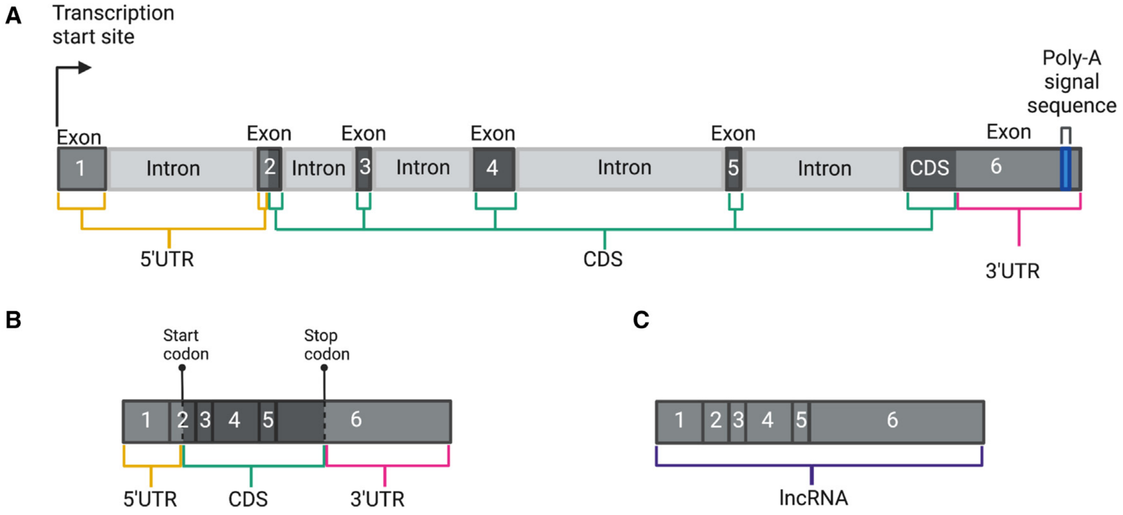

Figure 1. Schematics showing the position of the exons and introns with reference to the coding sequence (CDS) and untranslated regions (UTRs)

(A) The genomic region of a protein-coding transcript with six exons. Exons 3, 4, and 5 are entirely CDSs, exon 1 is entirely 5′ UTR, and exons 2 and 6 contain both CDSs and UTR sequences.

(B) The mature mRNA (after removal of introns by splicing) of the same protein-coding transcript as represented in (A).

(C) The mature RNA of a long non-coding RNA (lncRNA) also with six exons, all of which are entirely non-coding. 5′ UTRs containing exons are indicated in yellow, CDSs containing exons are in green, 3′ UTRs containing exons are in pink, and lncRNA exons are in purple. The poly-A signal is in blue. This figure was made in BioRender.

in all transcripts downloaded from Ensembl (Figure 2A; see Methods S1). While in fission yeast *S. pombe* and roundworm *C. elegans*, the proportion of exonic bases that are annotated as protein coding is high, at 68.4% and 75%, respectively, this proportion generally decreases with increased organism complexity. This decrease coincides with an increase in the length of 3′ UTRs and a greater number of non-coding RNA genes in these more complex organisms (Figure 2A). Only 28.1% of exonic bases in mice and 23.0% in humans are annotated as protein coding, while 30.8% and 32.4% are annotated as UTRs and 29.0% and 37.4% as non-coding RNA, respectively.

One barrier to extending this analysis wider in the Tree of Life is that available genome annotations tend to be incomplete and inconsistent about how they include UTRs and other non-coding exons. Genome annotation focuses on protein-coding regions and is aided by their high levels of sequence conservation. Non-coding regions are less well conserved, hindering automated annotation. Hence, non-coding exon annotation often requires data beyond the DNA sequence, such as RNA sequencing (RNA-seq) and cap analysis gene expression (CAGE), which are unavailable for many species.

“Whole-exome sequencing” captures less than 40% of exonic sequences

Whole-exome sequencing (WES) only covers a fraction of the exons within the genome, and for some of those exons (namely those that also contain UTRs), only a sub-portion is included within the sequencing capture region. In almost all cases, WES only directly captures the regions of exons that are protein coding, although variants in the directly adjacent sequence, or buffer region, may also be detected, some of which may be functionally important.[14] We assessed the overlap of the human WES capture regions used to sequence the UK Biobank dataset with different categories of exonic sequences from Ensembl (Figure 2B; Table S1). The vast majority of captured bases are within the coding sequence (89.1%). In total, WES-captured regions only cover 24.2% of exonic bases, indicating that the name WES is misleading. These captured regions cover 93.7% of bases annotated as protein coding but only 11.3% of 5′ UTR, 5.3% of 3′ UTR, and 1.3% of non-coding RNA exonic bases. When accounting for a 50 bp buffer on either side of each captured region, the proportion of exonic bases covered increases only marginally, to 27.3% (including 21.0% and 8.4% of 5′ UTR and 3′ UTR bases, respectively). Recent comparison of WES and WES data from the UK Biobank calculated that WES missed 72.2% and 89.4% of 5′ UTR and 3′ UTR variants, respectively.[15] There is no doubt that the naming of WES technology has increased the confusion surrounding the definition of the exon, promoting its misuse as synonymous to protein coding.

The non-coding exome plays important roles in regulation, disease, and biotechnology

The non-coding exome plays important roles throughout eukaryotes, with deletion or disruption of these elements linked to dysfunction and many different diseases.[16,17]

UTRs, the non-coding exonic regions of protein-coding genes, are important regulators of RNA stability, localization, and translation, controlling the amount of protein that is produced in the cell.[18–20] This regulation is mediated via interaction with RNA-binding proteins and microRNAs (miRNAs)[18,21] and through regulatory sequence elements including upstream open reading frames (uORFs)[22] and secondary structures.[19] Genetic variants that disrupt these regulatory elements can cause disease; for

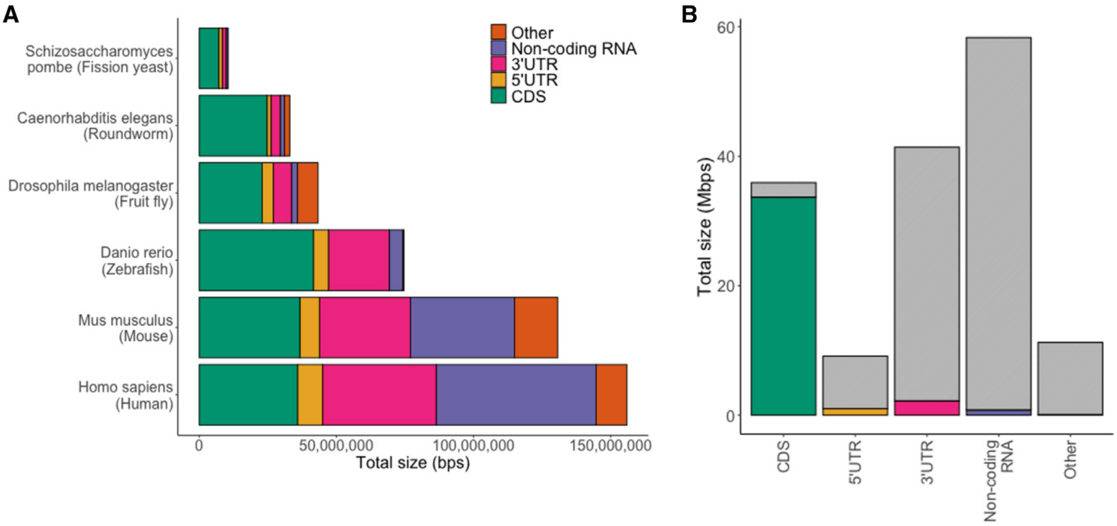

Figure 2. Proportion of exonic sequences and representation in whole-exome sequencing

(A) Comparison of the proportion of exonic bases with annotations of protein coding, 5′ UTR, 3′ UTR, non-coding RNA, and other (including transposable element gene or pseudogene exons not annotated as protein coding) across six different organisms.

(B) A bar plot of the total size of exonic bases in humans with different annotations showing the overlap with whole-exome sequencing capture regions. Bases that are within the capture are shown in color with bases that are not in gray. The raw numbers behind this figure are in Table S1.

example, variants that create or disrupt uORFs in the 5′ UTR of *NF1* cause neurofibromatosis type 1.[23] UTRs may also regulate protein function. In humans, CD47 proteins encoded by transcript isoforms that differ only in their 3′ UTR (i.e., the protein sequence is identical) interact with different protein complexes[18] due to different cellular localization of the mRNAs. Furthermore, UTR sequences have been demonstrated to tune protein expression in synthetic biology[24] and in mRNA therapeutics.[25]

Non-coding RNAs are also composed of non-coding exons and have a range of important functions, with many yet to be discovered. For example, in humans, the microRNA miR-204 is essential for normal photoreceptor function. Genetic variants in the miR-204 seed region (the conserved sequence where miRNAs bind to an RNA molecule) that disrupt miR-204 targeting can cause a dominant retinal dystrophy.[26] Furthermore, many long non-coding RNAs (lncRNAs) have important roles in development, such as *Xist*, which is required for X chromosome inactivation in early development.[27]

Recommendations

We recommend clearer use of language that more accurately reflects the regions being described by, for example, referring to “coding exonic regions” and “non-coding exonic regions.” We also suggest that capture technologies that target primarily coding exons should be named in a way that better describes what they are assessing. We propose the use of coding exome sequencing, or CES, as an alternative to WES. Improving the way we refer to these approaches will remove confusion about what they are capturing, as well as improve understanding of the range of exonic regions in the genome with important functions.

We also recommend more thorough measurement and annotation of the non-coding exome across species. Accurate transcript maps are essential to understand all aspects of gene expression regulation and to interpret genetic variation, particularly as our knowledge of the role of non-coding regions and variation within them continues to evolve.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100296.

ACKNOWLEDGMENTS

We thank Jean Beggs, Bill Earnshaw, and two anonymous reviewers for helpful comments on the manuscript. N.W. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (220134/Z/20/Z) and grant funding from the Rosetrees Trust (PGL19-2/10025). J.L.A. is supported by grant funding from BBSRC BB/S007407/1 and BB/W011387/1. E.W.J.W. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (208779/Z/17/Z).

AUTHOR CONTRIBUTIONS

J.L.A., E.W.J.W., and N.W. together conceived of, wrote, and performed analysis for the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Krebs, J.E., Goldstein, E.S., and Kilpatrick, S.T. (2009). Lewin’s GENES X (Jones & Bartlett Publishers).
2. Gilbert, W. (1978). Why genes in pieces? Nature *271*, 501.
3. Black, D.L. (2000). Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. Cell *103*, 367–370.

4. Adams, J.M., and Cory, S. (1970). Untranslated nucleotide sequence at the 5’-end of R17 bacteriophage RNA. Nature *227*, 570–574.
5. Proudfoot, N.J., and Brownlee, G.G. (1974). Sequence at the 3’ end of globin mRNA shows homology with immunoglobulin light chain mRNA. Nature *252*, 359–362.
6. Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA. Cell *12*, 1–8.
7. Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. USA *74*, 3171–3175.
8. Catterall, J.F., O’Malley, B.W., Robertson, M.A., Staden, R., Tanaka, Y., and Brownlee, G.G. (1978). Nucleotide sequence homology at 12 intron–exon junctions in the chick ovalbumin gene. Nature *275*, 510–513. https://doi.org/10.1038/275510a0.
9. Strachan, T., and Lucassen, A. (2022). Genetics and Genomics in Medicine (CRC Press).
10. Albert, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2008). Molecular Biology of the Cell, 5th edition (Garland Science).
11. Berg, J.M., Stryer, L., Tymoczko, J.L., and Gatto, G.J. (2015). Biochemistry (Macmillan Learning).
12. Simpson, J.A., and Weiner, E.S.C. (1989). The Oxford English Dictionary (Clarendon Press).
13. Wikipedia contributors. Talk:Exon. Wikipedia, the free Encyclopedia https://en.wikipedia.org/w/index.php?title=Talk:Exon&oldid=1098434673.
14. Wright, C.F., Quaife, N.M., Ramos-Hernández, L., Danecek, P., Ferla, M.P., Samocha, K.E., Kaplanis, J., Gardner, E.J., Eberhardt, R.Y., Chao, K.R., et al. (2021). Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. Am. J. Hum. Genet. *108*, 1083–1094.
15. Halldorsson, B.V., Eggertsson, H.P., Moore, K.H.S., Hauswedell, H., Eiriksson, O., Ulfarsson, M.O., Palsson, G., Hardarson, M.T., Oddsson, A., Jensson, B.O., et al. (2022). The sequences of 150,119 genomes in the UK Biobank. Nature *607*, 732–740.
16. Ellingford, J.M., Ahn, J.W., Bagnall, R.D., Baralle, D., Barton, S., Campbell, C., Downes, K., Ellard, S., Duff-Farrier, C., FitzPatrick, D.R., et al. (2022). Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Med. *14*, 73.
17. Esteller, M. (2011). Non-coding RNAs in human disease. Nat. Rev. Genet. *12*, 861–874.
18. Mayr, C. (2019). What are 3’ UTRs doing? Cold Spring Harb. Perspect. Biol. *11*, a034728.
19. Hinnebusch, A.G., Ivanov, I.P., and Sonenberg, N. (2016). Translational control by 5’-untranslated regions of eukaryotic mRNAs. Science *352*, 1413–1416.
20. Srivastava, A.K., Lu, Y., Zinta, G., Lang, Z., and Zhu, J.-K. (2018). UTR-dependent control of gene expression in plants. Trends Plant Sci. *23*, 248–259.
21. Huntzinger, E., and Izaurralde, E. (2011). Gene silencing by microRNAs: Contributions of translational repression and mRNA decay. Nat. Rev. Genet. *12*, 99–110.
22. Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. USA *106*, 7507–7512.
23. Whiffin, N., Karczewski, K.J., Zhang, X., Chothani, S., Smith, M.J., Evans, D.G., Roberts, A.M., Quaife, N.M., Schafer, S., Rackham, O., et al. (2020). Characterising the loss-of-function impact of 5’untranslated region variants in 15,708 individuals. Nat. Commun. *11*, 1–12.
24. De Nijs, Y., De Maeseneire, S.L., and Soetaert, W.K. (2020). 5’ untranslated regions: the next regulatory sequence in yeast synthetic biology. Biol. Rev. Camb. Philos. Soc. *95*, 517–529.
25. Orlandini von Niessen, A.G., Poleganov, M.A., Rechner, C., Plaschke, A., Kranz, L.M., Fesser, S., Diken, M., Löwer, M., Vallazza, B., Beissert, T., et al. (2019). Improving mRNA-based therapeutic gene delivery by expression-augmenting 3′ UTRs identified by cellular library screening. Mol. Ther. *27*, 824–836.
26. Conte, I., Hadfield, K.D., Barbato, S., Carrella, S., Pizzo, M., Bhat, R.S., Carissimo, A., Karali, M., Porter, L.F., Urquhart, J., et al. (2015). MiR-204 is responsible for inherited retinal dystrophy associated with ocular coloboma. Proc. Natl. Acad. Sci. USA *112*, E3236–E3245.
27. Plath, K., Mlynarczyk-Evans, S., Nusinow, D.A., and Panning, B. (2002). Xist RNA and the mechanism of X chromosome inactivation. Annu. Rev. Genet. *36*, 233–278.