The value of primary transcripts to the clinical and non-clinical genomics community: survey results and roadmap for improvements.

Authors:
Joannella Morales, Aoife C. McMahon, Jane Loveland, Emily Perry, Adam Frankish, Sarah Hunt, Irina M. Armean, Paul Flicek, Fiona Cunningham.

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

Grant numbers
Ensembl receives majority funding from Wellcome Trust (grant number WT108749/Z/15/Z) with additional funding for specific project components. Research reported in this publication was supported by Wellcome Trust [WT200990/Z/16/Z, WT200990/A/16/Z], EMBL and by National Human Genome Research Institute of the National Institutes of Health under award number 2U41HG007234. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Abstract
Variant interpretation is dependent on transcript annotation and remains time consuming and challenging. There are major obstacles for historical data reuse and for interpretation of new variants. First, both RefSeq and Ensembl/GENCODE produce transcript sets in common use, but there is currently no easy way to translate between the two. Second, the resources often used for variant interpretation (e.g., ClinVar, gnomAD, UniProt) do not use the same transcript set, nor default transcript or protein sequence. Ensembl ran a survey in 2018 to sample attitudes to choosing one default transcript per locus, and to gather data on reference sequences used by the scientific community. This was publicised on the Ensembl and UCSC genome browsers, by email and on social media. We had 788 respondents. Here we report our results and roadmap to create an effective default set of transcripts for resources, and for reporting interpretation of clinical variants.

Keywords: transcript annotation, variant interpretation, survey, default

INTRODUCTION

Many advances in biological understanding and genomic medicine are dependent on variant interpretation and the ability to describe a sequence change with respect to a specific annotated transcript. However, in older publications the transcript version is rarely recorded, hampering the ability to reuse historical data. Occasionally no transcript is specified (e.g., CFTR del-508, BRAF V600E), or the analysis may have used one historic transcript only. Moreover, interpretation of novel data is hampered by the variety of reference sequences used to gather evidence for variant analysis, and lack of coordination across the resources. There are two commonly used transcript sets for annotation: NCBI's RefSeq (Pruitt et al., 2014) and EMBL-EBI's Ensembl/GENCODE (Frankish et al., 2021). Many highly-accessed genomics resources supporting variant interpretation use transcripts from only one set, or default to a single transcript (e.g. ExAC/gnomAD (Karczewski et al., 2020; Lek et al., 2016), Human Cell Atlas (Andersson et al., 2014), GTEx (GTEx Consortium et al., 2015), ClinVar (Landrum et al., 2014), HGMD (Stenson et al., 2012)). None of these are coordinated with UniProt's principal isoform (Bateman et al., 2017) and comparison of annotation across sets is non-trivial. Additionally, some transcript sequences do not perfectly match the reference genome used for variant calling.

With this in mind, we started to explore how to choose one default transcript for each protein-coding locus, and the merits of such a set. In 2018, we surveyed the community to understand the priorities and attitudes surrounding transcript choice and reporting. The survey results supported RefSeq and Ensembl/GENCODE agreeing on an identical transcript for each locus to be used as a common default across resources. Below we detail our other conclusions.

METHODS

To gather input from the scientific community on transcript usage, and attitudes to transcript change, we developed a survey. As well as ascertaining background information about our survey respondents, our questions broadly covered:

- What the demand was for a single transcript per locus, a minimal set of transcripts or a complete set of all known transcripts. For the minimal set, whether that should cover all exons with clinical significance, or all abundant protein-coding exons, or all abundant exons.
- How to choose one primary transcript per locus, raising awareness of the complexities and compromises when selecting one transcript. We had a series of questions where the respondent had to trade off: low abundance and longer coding sequence with higher abundance and a shorter coding sequence; abundance, coding sequence length and coverage of clinically relevant variants.
- The relative importance of transcripts remaining stable, or matching the reference assembly, or avoiding pathogenic alleles or including globally frequent alleles.
- Opinions on updating a transcript for changes in coding sequence, UTR length, transcript splicing or never updating.
- The reference sequences currently used, including for interpreting and reporting variants.

- The value of having different transcripts sets versus having increased agreement between RefSeq and Ensembl/GENCODE.

The examples we chose for picking transcripts were cartoon versions of real loci. We advertised the survey by email, on the Ensembl (Cunningham et al., 2018) and UCSC (Tyner et al., 2017) genome browsers, via social media, and through contacts to ClinGen and NCBI's Genetic Testing Registry participants.

RESULTS

The survey generated 788 responses (see questions and results here: https://tinyurl.com/embl-ebi-transcript-survey) from 32 different countries: the largest contributors were the USA, UK and Germany (40%, 19% and 5% respectively). We analysed our results into two categories based on the response to the multiple-choice question 'Where do you work?'. Those who selected 'clinical diagnostics' or 'clinical research' were labelled 'clinical' (N=285; 36%) and those who selected from (University/college/academia/non-profit /research; commercial/industry; government; other) were 'non-clinical' (N=503; 64%). The results and requirements from these categories were different. We assayed how transcripts were used across the scientific community (question 14). The most common words in the answers included: variants, analysis, expression, RNA-seq, clinical, reporting, gene and annotation.

When presented with two choices for a primary transcript, the more abundant or the longest coding sequence, the non-clinical group showed a clear preference for choosing the more abundant transcript (question 2a, 2b). In contrast, no clear preference emerged in the clinical group (see Figure 1). In question 3a, the choice was between the transcript that covers the most clinically relevant variants, that is most abundant, that is longest, or that is used historically. The clinical group preferred the transcript that covered the most clinically relevant variants (see Figure 2); (see also question 3b). In contrast, there was no obvious preference between these choices in question 3 for the non-clinical group. There was lower preference for historical transcripts (12%; 14% of respondents - question 3a; 3b).

We received >800 additional comments across questions 1-3. Themes that emerged from these: rejected the value of a primary transcript, stated that all transcripts should be used, or proposed an artificial transcript be created to cover all exons. Many comments called for ranking and filtering methods in genome browsers and resources, supported by specific data on transcript abundance, tissue-specificity/expressivity, cell-specificity, background conditions, environmental, developmental stage and transcript quality metrics. More data was requested on flagging transcripts that were computationally determined, predicted, fully functional, validated, chosen by expert consensus as clinically relevant, or rare. The importance of cell/tissue-specificity and the difficulty of assessing abundance or relative expression was often mentioned.

For transcript sequences, respondents were asked to prioritise either that a transcript sequence matches the reference assembly, does not contain pathogenic alleles, matches the global major allele or never changes. Here, the transcript that matches the reference was the priority choice (48%) across all respondents (question 4) (Figure 3). There was only a minority to whom transcript sequences never changing was important (<10%, questions 4 and 5).

For transcript usage for reporting and interpretation, there was a preference captured by the respondent comment “I wouldn’t use just one transcript for INTERPRETATION unless it was the only one known” over only using one transcript (question 6). The preferred option for clinical respondents was to report on the primary transcript and the affected transcript (39%) rather than across all transcripts (14%). The opposite was true for the ‘non-clinical’ group (18% vs 40% respectively) (question 7).

We surveyed the reference sequences used for reporting in question 8 (Figure 4). In general, ‘clinical’ respondents used RefSeq, Locus Reference Genomic (LRG) (Dalgleish et al., 2010; MacArthur et al., 2014) and GRCh37, rather than Ensembl/GENCODE or GRCh38. Whereas the ‘non-clinical’ community replies were more equally spread across using GRCh38 and GRCh37, RefSeq or Ensembl/ GENCODE but not LRG.

Results from the survey indicated that having RefSeq and Ensembl/GENCODE agree on one primary transcript per gene would be welcome (54% overall; 67% of ‘clinical’ respondents, question 10). We revisited the question ‘Do you want us to provide one primary transcript’ at the end of the survey requiring a Yes, No or ‘Not sure’ answer. Here 60% of the ‘clinical’ respondents were in favour, compared with 48% of ‘non-clinical’ ones.

With input from this survey results, our conclusions and recommendations are that:

1. RefSeq and Ensembl/GENCODE collaborate to agree on:
    - one identical primary transcript per locus that matches the reference assembly. This is to ensure the community, browsers and resources use a good, consensus choice of transcript for analyses or situations that require only one (e.g., default display per gene).
    - minimal additional identical transcripts that match the reference assembly required for clinical reporting.
2. Transcripts are updated from historical exemplars, using modern datasets to choose a representative transcript:
    - evaluated on predicted functional significance and abundance rather than due to longest length, or being defined first (i.e., the historical transcript).
    - whose sequence is an exact reference genome sequence match.
3. All resources adopt this primary agreed transcript for the most effective benefit of the workings of the scientific community.
4. Genome browsers and resources consider improvements to their methods of filtering and ranking transcripts to facilitate choosing the appropriate transcript(s). Often, using only the one primary transcript per locus may not be right.

DISCUSSION

Across the survey results as a whole, there is no agreed method for designating a primary transcript. However, the value of consensus between Ensembl/GENCODE and RefSeq was highlighted as important. There is a history of collaboration between the two groups, for example on the Consensus CDS (CCDS) project (Pujar et al., 2018) and LRG. For many transcripts, the CCDS project has achieved consensus for the exon/intron structure over the protein-coding region, but there remain coding sequence discrepancies and structure differences in the untranslated regions (UTRs). The LRG project focuses on recording historical sequences for variant reporting that should never change, and therefore many of these do not perfectly match the reference assembly. However, the survey demonstrated a tolerance for change (only 6% selected 'Never update' in question 5).

Interestingly, many suggested the ideal primary transcript should contain all exons. This 'meta transcript' approach has been used for a few LRGs (e.g., LRG_391 for TTN; and LRG_202 for NEB) that represent an inferred transcript model containing all identifiable in-frame coding exons. However, it leads to the creation of primary transcripts that do not reflect biological reality and which are not guaranteed to be comprehensive: they may contain exons that show huge differences in their inclusion rates generally, and in specific tissues; they may include mutually exclusive exons; they cannot include exons in different frames; and they will need to be updated if novel coding exons are subsequently discovered.

The survey reported many, especially clinical groups, are still using GRCh37, released in 2009. GRCh38, released in 2013, offers a more complete genome that is being continuously improved by the Genome Reference Consortium (GRC) (Schneider et al., 2017) through a supplemental release model. Ensembl/GENCODE gene annotation is only being updated on GRCh38. Therefore, it is only the annotation on GRCh38 that will benefit from all the improvements supported by the incorporation of new data sets (such as long transcriptomic data generated using methods developed by Oxford Nanopore Technologies and Pacific Biosciences), and of tools (such as the PhyloCSF method for identifying regions of the genome with conserved protein-coding potential). Major resources such as gnomAD and DECIPHER are also now using GRCh38.

Worth noting is that many survey comments expressed resistance to the very idea of a default transcript. They rightly pointed out that biology cannot be simplified in this manner, however appealing the concept. We agree completely that genome analysis requires considering multiple transcripts per gene and Ensembl remains absolutely committed to annotating all evidence-based transcripts at every locus. Analysis, including the interpretation of variants identified from clinical sequencing, should always be in relation to the most relevant and abundant isoform(s) for the tissue of interest at the developmental stage of interest and in the correct cell type. In general, we do not yet have the data to determine this. Although projects such as GTEx and Human Cell Atlas have and will change the landscape of transcriptomic data available, currently for the majority of developmental stages, there is a lack of this critical information. As a result, in the absence of tissue-specific data, any analysis should consider all transcripts or proteins at the locus. We urge more cooperation between clinical diagnostics and research to use a broader transcript set and thereby remove the bias in reported transcripts.

However, for practical reasons it is sometimes helpful to have only one transcript for sharing and comparing results across experiments, datasets and collaborations. Indeed, many browsers,

bioinformatics tools and variant interpretation pipelines have chosen a default transcript, independently from each other. For example, Ensembl and UniProt have had their own 'canonical' (available only through the Ensembl API) and 'principal isoform' choices, respectively, for default transcripts and proteins for over a decade while RefSeq has a 'select' transcript and HGMD has a default RefSeq. Often these have been based on the longest transcript (https://www.ensembl.org/Help/Glossary), or the first sequences published, or most prevalent (https://www.uniprot.org/help/canonical_and_isoforms) but are not necessarily consistent or coordinated with other resources.

It is clear, therefore, that the concept of a default transcript already exists across resources but is uncoordinated. The survey results demonstrated a desire for a default transcript, but in the absence of a consensus choice so far, we see that each genomics resource, scientist and experiment choose a different transcript. Selecting one particular transcript per locus comes with a risk of biasing the scientific community towards ignoring the full transcriptome. However, a collaboration between RefSeq and Ensembl/GENCODE would provide the leadership necessary to unite the community and provide a consensus choice for a set of results and opinions that lack a clear consensus from the survey. This would be a practical and coordinated effort to define one default transcript per locus. There is no overall 'correct' choice but the most important and valuable property of a default transcript is that it is consistent, for reporting and to ease use of different resources and tools that require a default transcript. Equally important would be to work with all major browsers and resources (e.g., NCBI, Ensembl, Ensembl's Variant Effect Predictor, UCSC Genome Browser, gnomAD, DECIPHER, UniProt, Panel App, COSMIC etc.) to ensure adoption of the common default transcript.

ACKNOWLEDGEMENTS:

We would like to thank the 788 individuals who completed our survey and everyone who helped advertise it. Thank you also to Caroline Wright for useful analysis discussions, and the following for their feedback on the survey design: Deanna Church, Mark Diekhans, Terence Murphy, Heidi Rehm, Magali Ruffier, Andrew Yates.

References

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., … Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461. doi: 10.1038/nature12787

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., … Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. doi: 10.1093/nar/gkw1099

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., … Flicek, P. (2018). Ensembl 2019. *Nucleic Acids Research*. doi: 10.1093/nar/gky1113

Dalgleish, R., Flicek, P., Cunningham, F., Astashyn, A., Tully, R. E., Proctor, G., … Maglott, D. R. (2010). Locus Reference Genomic sequences: An improved basis for describing human DNA variants. *Genome Medicine*, *2*(4), 24. doi: 10.1186/gm145

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., … Flicek, P. (2021). GENCODE 2021. *Nucleic Acids Research*, *49*(D1), D916–D923. doi: 10.1093/nar/gkaa1087

GTEx Consortium, T., Ardlie, K. G., Deluca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., … Dermitzakis, E. T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. doi: 10.1126/science.1262110

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. doi: 10.1038/s41586-020-2308-7

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(D1), D980–D985. doi: 10.1093/nar/gkt1113

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. doi: 10.1038/nature19057

MacArthur, J. A. L., Morales, J., Tully, R. E., Astashyn, A., Gil, L., Bruford, E. A., … Cunningham, F. (2014). Locus Reference Genomic: Reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Research*, *42*(D1), D873–D878. doi: 10.1093/nar/gkt1198

Pujar, S., O'Leary, N. A., Farrell, C. M., Loveland, J. E., Mudge, J. M., Wallin, C., … Pruitt, K. D. (2018). Consensus coding sequence (CCDS) database: A standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Research*, *46*(Database issue), D221–D228. doi: 10.1093/nar/gkx1031

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *27*(5), 849–864. doi: 10.1101/gr.213611.116

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., & Cooper, D. N. (2012). The Human Gene Mutation Database (HGMD) and its exploitation in the fields of

personalized genomics and molecular evolution. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *Chapter 1*, Unit1.13. doi: 10.1002/0471250953.bi0113s39

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., … Kent, W. J. (2017). The UCSC Genome Browser database: 2017 update. *Nucleic Acids Research*, *45*(D1), D626–D634. doi: 10.1093/nar/gkw1134

FIGURE LEGENDS

Figure 1:

An example of a cartoon version of a locus we used in the survey to understand opinions across the scientific community on different options for choosing one transcript. These are the transcript scenarios presented for questions 2a (top panel) and 2b (bottom panel). For question 2a, and for question 2b, we asked respondents to choose either the first longer coding transcript, or the second more abundant (but shorter) one as a primary transcript. For both questions, the more abundant one (indicated by the blue arrow) was the most popular transcript choice for the non-clinical community (75%; 68%). However, there was no clear preference for this one (indicated by the blue arrow) from the clinical respondents (54%; 46%).

Figure 2:

Top panel: question 3a from the survey.

Bottom panel: bar chart of answers across 503 'non-clinical' respondents and 285 'clinical' ones. Respondents chose between the transcript that covers the most clinically relevant variants (D), that is most abundant (E), that has the longest coding sequence (C) or that is used historically.

The results favoured (D), the transcript that covers the most clinically relevant variants, or (E) the most abundant overall. However, for the clinical group, there was a strong preference for the transcript that covers the most clinically relevant variants (D) (64%) despite having lower abundance overall. In contrast, there was no obvious preference between these choices for the non-clinical group. Here neither the longest coding transcript (C), nor the historical transcript were popular preferences.

Figure 3:

Bar chart of results from question 4 which asked 'Considering the sequence of a transcript, which is the most important to you (choose one):

- That the sequence matches the reference assembly sequence (e.g. GRCh37/ hg19), even if it contains minor alleles
- That the sequence does not contain any pathogenic alleles
- That the sequence matches the global major allele
- That the sequence does not change
- It doesn't matter to me

Both the clinical (N=285) and non-clinical (N=503) respondents had "that the sequence matches the reference.." as most important (44%; 50%). For many in the clinical group, however, it was also important that a transcript did not contain any pathogenic alleles (7% of 'non-clinical' respondents but 23% 'clinical' ones). Only a minority prioritised that a transcript sequence never changes (<10%).

Figure 4:

Answers across respondents (503 ‘non-clinical’ and 285 ‘clinical’) to question 8: “Which reference sequences do you use for reporting variants (select all that apply)”:

- RefSeq transcripts or proteins
- GRCh37/hg19 genome
- LRG transcripts or LRG proteins
- Ensembl/GENCODE transcripts or proteins
- GRCh38/hg38 genome
- Use both RefSeq and Ensembl
- Use both 37 and 38 genome references

In general, the ‘clinical’ respondents used:

- RefSeq transcripts or proteins rather than Ensembl/GENCODE (73% vs 24%),
- GRCh37/hg19 (71% vs 19% for GRCh38) and
- LRG transcripts or proteins (27%).

Whereas the ‘non-clinical’ community replies were more equally spread across using:

- GRCh38 and GRCh37 (46% vs 42%), and
- RefSeq or Ensembl/GENCODE (46% vs 52%) and
- little usage of LRG (4%).

Declarations

Ethics approval and consent to participate n/a

Consent for publication - n/a

Availability of data and materials

The datasets generated and/or analysed during the current study are available here.
https://tinyurl.com/embl-ebi-transcript-survey

Conflicts of interest statement

PF is a member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd.

Author's contributions

- Joannella: survey design; survey analysis; survey promotion
- Aoife: survey design; survey design feedback, survey figures
- Jane: survey design, survey design feedback
- Adam: survey design, survey design feedback
- Emily P: survey review; survey
- Sarah H: bioinformatics analysis
- Irina M. A: bioinformatics analysis
- Paul F: survey dissemination - twitter; survey design feedback, manuscript input
- Fiona: wrote this manuscript; survey design; survey analysis; publication text; survey promotion