DATA NOTE

Clinical Annotation Reference Templates: a resource for consistent variant annotation [version 1; referees: 1 approved with reservations]

Shawn Yost1, Márton Münz1, Shazia Mahamdallie1, Anthony Renwick1, Elise Ruark1, Nazneen Rahman1,2

1Division of Genetics and Epidemiology, Institute of Cancer Research, UK, 15 Cotsold Road, London, SM2 5NG, UK
2Cancer Genetics Unit, Royal Marsden NHS Foundation Trust, London, SM2 5PT, UK

Abstract
Annotating the impact of a variant on a gene is a vital component of genetic medicine and genetic research. Different gene annotations for the same genomic variant are possible, because different structures and sequences for the same gene are available. The clinical community typically use RefSeq NMs to annotate gene variation, which do not always match the reference genome. The scientific community typically use Ensembl ENSTs to annotate gene variation. These match the reference genome, but often do not match the equivalent NM. Often the transcripts used to annotate gene variation are not provided, impeding interoperability and consistency.

Here we introduce the concept of the Clinical Annotation Reference Template (CART). CARTs are analogous to the reference genome; they provide a universal standard template so reference genomic coordinates are consistently annotated at the protein level. Naturally, there are many situations where annotations using a specific transcript, or multiple transcripts are useful. The aim of the CARTs is not to impede this practice. Rather, the CART annotation serves as an anchor to ensure interoperability between different annotation systems and variant frequency accuracy. Annotations using other explicitly-named transcripts should also be provided, wherever useful.

We have integrated transcript data to generate CARTs for over 18,000 genes, for both GRCh37 and GRCh38, based on the associated NM and ENST identified through the CART selection process. Each CART has a unique ID and can be used individually or as a stable set of templates; CART37A for GRCh37 and CART38A for GRCh38.

We have made the CARTs available on the UCSC browser and in different file formats on the Open Science Framework: https://osf.io/tcvbq/. We have also made the CARTtools software we used to generate the CARTs available on GitHub.

We hope the CARTs will be useful in helping to drive transparent, stable, consistent, interoperable variant annotation.

Keywords
Variant annotation, transcript, template, CART, next generation sequencing, RefSeq NM, Ensembl ENST
This article is included in the Transforming Genetic Medicine Initiative (TGMI) gateway.

Corresponding author: Nazneen Rahman (rahmanlab@icr.ac.uk)

Author roles: Yost S: Data Curation, Formal Analysis, Methodology, Resources, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; Münz M: Data Curation, Formal Analysis, Methodology, Resources, Software, Validation, Writing – Review & Editing; Mahamdallie S: Methodology, Writing – Review & Editing; Renwick A: Methodology, Writing – Review & Editing; Ruark E: Data Curation, Formal Analysis, Methodology, Writing – Review & Editing; Rahman N: Conceptualization, Data Curation, Funding Acquisition, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: NR is a Non-Executive Director of AstraZeneca. ER is Product Manager for Foresite Capital

Grant information: This work was supported by the Wellcome Trust (200990).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Yost S et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Yost S, Münz M, Mahamdallie S et al. Clinical Annotation Reference Templates: a resource for consistent variant annotation [version 1; referees: 1 approved with reservations] Wellcome Open Research 2018, 3:146 (https://doi.org/10.12688/wellcomeopenres.14924.1)

First published: 14 Nov 2018, 3:146 (https://doi.org/10.12688/wellcomeopenres.14924.1)
Introduction

An integral component of next generation sequencing (NGS) gene analysis methods is the annotation of variation using the human reference genome as a baseline. By contrast, historical gene analysis methods, such as Sanger sequencing, can choose which sequences to use for variant annotation. The majority of the clinical community, and much of the clinical research community, use RefSeq NM transcripts as baseline sequences for variant annotation. This can result in annotation inconsistencies for several reasons. Firstly, there are often several different NMs available for a particular gene and laboratories choose to use different NMs to annotate the same variant in different ways. Secondly, NMs are curated mRNA derived sequences and do not always match the reference genome, leading to annotation inconsistencies with NGS-based gene analyses. Thirdly, gene structure information, such as intron-exon boundary positions, is not included in the NMs. Instead this information is overlaid inconsistently, which can lead to different annotations of exon deletion/duplication variants, which are an important cause of disease.

ALMS1 provides an example of the variant annotation inconsistencies that can occur. Pathogenic variants in ALMS1 cause Alstrom syndrome (MIM 606844). There is only one RefSeq transcript for ALMS1, NM_015120.4. This transcript has two different 3 bp insertions in exons 1 and 8, compared to reference genome build GRCh37. This means variants with the same genomic coordinates can have different annotations. For example variant chr2:73717247C>T (GRCh37) is annotated as c.8164C>T; p.Arg2721Ter in ClinVar, OMIM and the medical literature, but as c.8158C>T; p.Arg2720Ter in resources that use reference genome based transcripts for annotation such as ExAC, VEP or CAVA. Adding further complexity, NM_015120.4 is different to build GRCh38 by only one 3 bp insertion, in exon 1. So the same variant annotated on GRCh38 would have genomic coordinates of chr2:73490120C>T, and would be annotated as c.8164C>T; p.Arg2722Ter using NM_015120.4 but c.8161C>T;p.Arg2721Ter in resources using reference genome based transcripts for annotation.

NGS-based gene analyses often use ENST transcripts as the baseline sequences for variant annotation. ENSTs always match the reference genome. However, similar to NMs, multiple ENSTs are available for many genes and laboratories may choose to use different ENSTs to annotate variation in a given gene. This can result in variant annotation differences between laboratories using different ENSTs and between laboratories using NMs for annotation and those using ENSTs. A further issue is that ENSTs are frequently updated, potentially compromising the stability of annotations, particularly if an ENST is retired. For example, compared with Ensembl release 91, Ensembl release 92 retired 314 transcripts, included 3,226 new transcripts, and 1,336 new version numbers for existing transcripts, predominantly due to changes in the untranslated region (UTR) and/or coding sequence (CDS). BMPR1A exemplifies the problems updates can inadvertently cause. NM_004329.2 is the only RefSeq transcript available for BMPR1A. Historically, this RefSeq NM was associated with ENST00000224764, which has been used to annotate BMPR1A variants in many publications, reports and databases. However, this ENST00000224764 is no longer available, links to it now state ‘this identifier is not in the current EnsEMBL database’, compromising integration of historical and current BMPR1A variant annotations.

Although it is usually possible to work out how different ENSTs and NMs relate to each other, it is difficult, time-consuming and rarely done. Instead people often assume annotations are consistent. If this is a misassumption, it can lead to downstream scientific and clinical errors, particularly in relation to clinical interpretations about variant pathogenicity. A common error of this type is the assumption that if a variant (annotated using an NM) is not present in the default presentation of ExAC (annotated using an ENST) it must be exceptionally rare, and hence more likely to be pathogenic. However, it is possible the relevant variant is present in ExAC but has a different annotation, because the ENST selected differs from the NM selected. BDNF is an example of this problem. BDNF is associated with 17 different NMs. In 2002, a BDNF variant, c.29C>T;p.Thr2Ile, annotated using NM_001709.4, was proposed to cause a severe condition called congenital central hypoventilation syndrome (CCHS). In ClinVar NM_170731.4 is used and the same BDNF variant is called p.Thr10Ile. Neither p.Thr2Ile nor p.Thr10Ile appear in the default annotations in ExAC, which use ENST00000438929. This results in the variant, g.11:27680107G>A, being called p.Thr84Ile, which is present in 132 individuals. At this allele frequency (0.001) it would be a major cause of CCHS if it was a disease-causing variant. However, to our knowledge no one with CCHS and this variant has been reported since the original publication, and it is highly unlikely to be a pathogenic variant. OMIM have downgraded the variant from pathogenic to uncertain significance since we brought this issue to their attention (MIM 113505).

Given the intrinsic differences in the widely used variant annotation systems it is essential that the transcripts used for variant calling are transparently provided and stably available. However this often does not occur. Moreover, it is becoming increasingly challenging to provide this information on a gene-by-gene basis, because many analyses now generate variant calls from thousands of genes.

To address this important issue we here introduce the concept of the Clinical Annotation Reference Template (CART) and provide CARTs for GRCh37 and GRCh38. The CARTs aim to provide standard, interoperable, stable gene templates for variant annotation that are based on the reference genome sequence, include the required structural information, and can be used either individually or as set.

CARTs can be considered analogous to the reference genome; they provide a universal standard template so the reference genomic coordinates of a variant are consistently annotated at the protein level. Of course, there are many situations where annotations using a specific transcript, or all available transcripts are useful. The aim of the CARTs is not to impede or curb this practice. Rather, we propose that the CART annotation is
always provided, as an anchor to ensure interoperability between different annotation systems and variant frequency accuracy. Additionally, annotations using other explicitly-named transcripts should also be provided where necessary or useful. To facilitate transparent, consistent variant annotations of panel/exome/genome tests the CARTs can be used as a set. For example at the bottom of a clinical exome report, or in a publication it could be stated that variants were called using the CART37A series, except where otherwise stated.

We hope the CARTs will be useful in helping to drive transparent, stable, consistent, interoperable variant annotations.

**Methods**

**Gene selection**

We downloaded the approved HGNC IDs\(^{12}\) from the HGNC BioMart portal page (https://biomart.genenames.org/martform/#!/default/HGNC?datasets=hgnc_gene_mart) for every gene on chromosome 1-22, X, Y and mitochondria with a ‘Locus type’ equal to ‘gene with protein product’ on 08/01/2018. This gave a set of 19,171 protein-coding genes (Extended Data File 1\(^{13}\)).

**Datasets**

We used the following datasets in the CART selection process.

**NCBI Gene ID:** For each HGNC ID we downloaded the corresponding NCBI ID from the HGNC BioMart portal page on 08/01/2018.

**APPRIS:** We downloaded the APPRIS principal isoforms data file from the APPRIS website (http://appris.bioinfo.cnio.es/#/downloads) corresponding to RefSeq release 107 from APPRIS version 20 (rs107v20) on 13/01/2017. For each gene APPRIS identifies a single principle isoform, if possible, and identifies every NM associated with the gene in RefSeq that matches the principal isoform\(^{11}\).

**RefSeq NM genomic alignments:** We downloaded the RefSeq genomic mapping of NMs for GRCh37.13.1_interim_annotation and GRCh38.10 from the RefSeq FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/) on 13/01/2017.

**UCSC NM genomic alignments:** We downloaded the UCSC mapping\(^{14}\) of NMs for GRCh37 and GRCh38 from the UCSC website (ftp://hgdownload.cse.ucsc.edu/goldenPath/) on 15/10/2018.

**Ensembl ENSTs:** We downloaded the Ensembl\(^{15}\) transcript database files from their FTP site (ftp://ftp.ensembl.org/pub/) for the given release. For GRCh37 and GRCh38 we used Ensembl release 75 and 92, respectively.

**RefSeqGene:** We downloaded the set of RefSeqGenes from the FTP site (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/RefSeqGene/) on 02/05/2017. For each reference standard gene (RefSeqGene) we extracted the NCBI gene ID, gene symbol, and NMs.

**ClinVar:** We downloaded the ClinVar\(^{4}\) variant summary file from the FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/) on 02/05/2017. For each gene in ClinVar we extracted the NM, gene symbol and NCBI ID.

**CART generation**

CART is the term used to describe the below information:

1. Human reference genome build (e.g.GRCh37)
2. Strand
3. The total number of exons, both CDS and UTR
4. All exon boundary positions
5. Translation start c.1 and translation stop position
6. The reference genome sequence of the above

The CART selection and generation process is described below. The process uses eight scripts which are described in detail in Extended Data File 2\(^{2}\). We have made the scripts available as CARTtools (see CART availability)\(^{16}\). They can be run in one command or each script can be used separately.

**Algorithmic NM selection**

For each gene we used both the APPRIS and RefSeq genomic alignment data to identify a single NM on which to base the CART. We call this the ‘Algorithmic NM’ (Figure 1). If a gene had an APPRIS principal isoform associated with only one correctly aligned NM it was selected as the Algorithmic NM. If a gene had multiple NMs associated with the APPRIS principal isoform we used a UTR selection process to select a single Algorithmic NM (Figure 1). The goal of the UTR selection process is to reduce the number of transcripts through a sequential selection process until only one NM remains. The UTR selection process includes three major criteria (A, B, C) and 10 minor criteria (A1-A3, B1-B4, and C1-C3) as described in Figure 1. The criteria are applied sequentially to all the available NMs associated with the APPRIS principal isoform until one NM is removed. The UTR selection process then restarts at A1 using the remaining NMs, until a single NM remains, which becomes the Algorithmic NM.

We did not select an Algorithmic NM if a) there were no NMs to select from, b) there were multiple NMs with the same genomic coordinates, c) one or more NMs assigned as principal by APPRIS had different CDS or d) we could not match the NCBI Gene ID to an HGNC ID. The Algorithmic NMs are given in Extended Data File 1\(^{11}\).

**Community NM selection**

We used RefSeqGene and ClinVar data to identify NMs used in the clinical diagnostic and clinical research communities. If the gene was in the RefSeqGene database we used the RefSeqGene NM(s) as the Community NM(s). If the gene was not in the RefSeqGene database we used the ClinVar NM(s) as the Community NM(s). If a gene was in neither database we did not...
select a Community NM. The Community NMs are given in Extended data file 11.

CART associated NM selection
We used the Algorithmic and Community NMs to select the final CART associated NM. We used the Algorithmic NM as the CART associated NM if there was no Community NM or if the Algorithmic NM was a Community NM. We used the Community NM as the CART associated NM if there was no Algorithmic NM or if there was a single Community NM that differed from the Algorithmic NM. The CART associated NMs are given in Extended Data File 11.

CART associated ENST selection
For each CART associated NM we next selected the closest matching ENST using the coordinates of the mapped NM (from RefSeq or UCSC if coordinates were not available from RefSeq) and Ensembl’s ENST coordinates (Extended Data File 11). To be selected, the CDS genomic coordinates of the ENST had to be identical to the CART associated NM. If there was only one ENST with identical CDS to the CART associated NM, it was selected as the associated ENST. If the CDS matched but there were UTR differences between the CART associated NM and available ENSTs we used the following selection process to select a single ENST. We prioritised ENSTs with the same number of 5’ UTRs as the CART associated NM. If none were available we prioritised ENSTs in which the 5’ UTR genomic location encompassed the 5’ UTR genomic location in the CART associated NM. If more than one ENST was available that matched these prioritisation criteria, or no ENST was available that matched the prioritisation criteria we used the UTR selection process shown in Figure 1, to select a single CART associated ENST.

Figure 1. Algorithmic NM selection process. Diagram showing the Algorithmic NM selection process implemented in CARTtools using the APPRIS and RefSeq datasets.
The CARTs
The genomic coordinates of the associated ENST were used as the genomic coordinates of the CART. Each CART has a unique identifier (CART ID) defined as: CART<genomeBuild><series><CARTNumber> (Figure 2). The genomeBuild is the human reference genome build the CARTs are aligned to, for example 37 for GRCh37. The CART series represents the full set of stable templates that the template belongs to, for example A for series A. The CARTNumber is a unique template number starting at 10,001. We used the same CARTNumber if the genomic sequence of the CART template for the gene did not change between builds. Thus for KCNC3 the CART IDs are CART37A25530 and CART38A25530 because the sequence and structures of the UTR and CDS are identical on GRCh37 and GRCh38. If the genomic sequence of the CART changed between genome builds the CARTNumber also changes, with the new CARTNumber always being the next available CARTNumber. For example, the CARTs for UMPS are CART37A11618 and CART38A28332 because the UMPS 3’ UTR is longer in GRCh38 than in GRCh37.

Using the above process, we were able to generate CARTs for 94% (18,000/19,171) of genes on GRCh37 and 96% (18,330/19,171) of genes on GRCh38. With respect to the differences between the CART associated ENST and the CART associated NM, all have identical CDS (by definition) and 16% (3,110) in GRCh37 and 17% (3,325) in GRCh38 also have identical UTRs (Figure 3A). The CARTs for GRCh37 and GRCh38 have identical CDS and UTR for 75% (14,350/19,171) of genes and identical CDS for 91% of genes (17,514/19,171).

![Figure 2. CART identifier. Each CART has a unique identifier with the format CART<genome build><series><CARTNumber>.](image)

![Figure 3. Comparisons of CART associated NM and CART associated ENST and CARTs on different genome builds. (A) Comparison of the CART associated NM and CART associated ENSTs for 19,171 genes. The CDS is identical for all genes, but UTR differences are common. B) Comparison of CARTs on GRCh37 (CART37A) and GRCh38 (CART38A) for 19,171 genes. The majority of CARTs are identical.](image)
The CARTtools output provides further details about the CARTs as shown in Extended Data File 1.

Data availability

Underlying data

We have made the CARTs available on the UCSC browser. They can be searched for as ‘CART37A’ or CART37B’ or directly through the following links. For CART37A, the CARTs for GRCh37: https://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=Rahman.team&hgS_otherUserSessionName=CART37A.

For CART38A, the CARTs for GRCh38: https://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=Rahman.team&hgS_otherUserSessionName=CART38A.

We have also made the CARTs available in the following annotation file formats GFF2, GFF3, GenePred, GenBank, FASTA and CAVA database, so the CARTs can be easily integrated into popular variant annotation or analysis tools such as VEP, Snpeff, ANNOVAR, Mutation Surveyor and CAVA.

If a gene does not have a CART we provide Ensembl’s ‘canonical’ ENST for that gene in the output files. Further information is available in the CARTtools documentation (Extended Data File 2)

The data files for CART37A and CART38A are available on the Open Science Framework (OSF): http://doi.org/10.17605/OSF.IO/TCVBQ. Data are available under the terms of a CC0 1.0 Universal licence.

Extended data

Extended data files have been archived on Open Science Framework: http://doi.org/10.17605/OSF.IO/TCVBQ. Data are available under the terms of a CC0 1.0 Universal licence.

Extended Data File 1. CART summary information. Descriptions of the column headings are given on OSF.

Extended Data File 2. CARTtools documentation.

Software availability

The latest release of CARTtools is available at: https://github.com/RahmanTeamDevelopment/CARTtools/releases.


Software license: MIT.

Grant information

This work was supported by the Wellcome Trust (200990).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We are very grateful to the many different people we had helpful discussions with over the last three years as we were developing the CARTs, in particular within the TGMI, EBI, APPRIS, the RefSeq team, ClinVar, UCSC, GenCC and many others. This work was undertaken as part of the Transforming Genetic Medicine Initiative (https://www.thetgmi.org/).

References


Summary:
Yost et al describe a transcript curation effort they call the Clinical Annotation Reference Template (CART). As there are two different transcript annotation bodies, RefSeq from NCBI and Ensembl from EMBL-EBI, clinical and research annotation often take place on different transcripts, hindering interoperability. The authors have developed CART sets for genome builds 37 and 38 in an effort to unify RefSeq and Ensembl transcripts for the community. These CARTS have an NM Refseq transcript and a closely mapped Ensembl transcript with clearly defined sequence and exon-intron boundaries. NMs were split into two categories: the algorithmic NM and the community NM and one of these was chosen as the NM for the CART. The Ensembl transcript that matched this NM as closely as possible was then chosen for the CART. These CARTS have been loaded into UCSC for ease of use by the research and the clinical community.

General comments:
A curation effort that unites Refseq and Ensembl users was extremely necessary. Mapping between transcripts, particularly while using resources to curate variants, is tricky and time-consuming. This is solidified by the many practical examples included in the introduction. Yost et al's efforts to match the sequences of NMs and ENSTs is an applaudable effort. However, designating a third set of transcripts with their own identifiers is not likely to be easily accepted by the research and clinical community. Clinical pipelines are complex and not particularly responsive to change. The same could perhaps be said of research pipelines used to generate large amounts of sequencing data.

Major points:
-While this effort is necessary and important, other unification efforts do exist that should be commented on in the introduction or the discussion. How is this effort different from the Locus Reference Genome (LRG) effort? Although the Matched Annotation of NCBI and Embl-EBI (MANE) was released fairly recently, perhaps after this article was submitted, it deserves a mention too.

- More flow diagrams need to be added to figure 1. Choosing the Community NM, Choosing the ENST, and Choosing the final NM (Community vs Algorithmic) should also be visualized.

- More information should be provided about the comparison of the Community vs the Algorithmic NMs. How often did they match?

- Only a small set of genes could not have a CART designated. Why was this the case? Did they fall in
particular disease areas or was it just random? Please comment.

-One of the criticisms for LRG transcripts is their inflexibility when transcripts change. How flexible are CARTs? Is each CART stable for the genome build or could a new series of CARTS be generated before the release of a new genome build? Do CARTS always have to be generated in sets or could an update be done on a per-gene basis?

-Why were ENSTs with the same number of 5' UTRs as the CART-associated NM prioritized?

-Because the ENST was chosen second based on the CART-associated NM, how often was the canonical ENST chosen?

-Why were ENSTs with the same number of 5' UTRs as the CART-associated NM prioritized?

-Why were ENSTs with the same number of 5' UTRs as the CART-associated NM prioritized?

-Why were ENSTs with the same number of 5' UTRs as the CART-associated NM prioritized?

-Minor points:
   -More general information about transcript curation efforts would be useful in the introduction

-It would be useful to provide the list of the NM and ENST that make up each CART in case researchers would still like to take advantage of your thorough curation efforts, but would prefer to annotate using those transcripts so as not to change their pipelines.

-In the first paragraph of the introduction “By contrast, historical gene analysis methods, such as Sanger sequencing, can choose which sequences to use for variant annotation.” is awkwardly worded and should be re-written.

-On page 3, second column, 3rd full paragraph, there is a typo: “The CARTs aim to provide standard, interoperable, stable gene templates for variant annotation that are based on the reference genome sequence, include the required structural information, and can be used either individually or as a set.

**Is the rationale for creating the dataset(s) clearly described?**
Yes

**Are the protocols appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and materials provided to allow replication by others?**
Partly

**Are the datasets clearly presented in a useable and accessible format?**
Partly

**Competing Interests:** No competing interests were disclosed.

**Referee Expertise:** Clinical molecular genetics, variant and gene curation

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.