CARTtools

- Documentation -

1 Introduction

CARTtools is a collection of Python tools for selecting and processing **CARTs (Clinical Annotation Reference Templates)**. The included tools can be run individually (see *Section 7*) or together as an integrated end-to-end pipeline (*Section 3*). The pipeline can be easily customised with a configuration file and run with a single command.

In this documentation we discuss the installation, configuration and execution of the pipeline and individual tools and we also describe the generated output files.

The CARTtools package consists of the following 8 tools:

- **RefSeqDB**
- **RefSeqCheck**
- **SelectNMs**
- **MapNMs**
- **EnsemblDB**
- **SelectENSTs**
- **CompareENSTs**
- **FormatCARTs**

2 Installation

CARTtools can be downloaded from GitHub (https://github.com/RahmanTeamDevelopment/CARTtools/releases) in either .zip or .tar.gz format. To unpack these run one of the following commands:

```
unzip CARTtools-x.x.x.zip
```

or:

```
tar -xvzf CARTtools-x.x.x.tar.gz
```

and then you can install CARTtools with the following commands:

```
cd CARTtools-x.x.x
```

```
./install.sh
```

CARTtools uses virtualenv and pip to manage all its extra dependencies, which means that it will not clutter up your system by installing things globally. Everything it installs will go into a sub-directory in the CARTtools-x.x.x directory. If you delete CARTtools then everything it has installed will also be deleted. Once the installation script has finished successfully, CARTtools is ready for use.

3 Running the pipeline

Once correctly installed, the CARTs pipeline can be run with the following command:

CARTtools-x.x.x/cart_pipeline --config config.txt

where config.txt is the configuration file discussed in *Section 4*.

The main input file of the pipeline is the list of HGNC IDs of the genes for which CARTs are selected (the input file name is set in the configuration file). The pipeline follows the CART selection process described in our paper and outputs the selected CARTs in various file formats (see *Section 5*).

4 Configuration file

The configuration file has to follow the INI format (https://en.wikipedia.org/wiki/INI_file) and may contain the following options:

- **reference_37**: GRCh37 reference genome fasta file *(mandatory option)*
- **reference_38**: GRCh38 reference genome fasta file *(mandatory option)*
- **output_prefix**: output file names prefix *(mandatory option)*

[SelectNMs] section:

- **input_genes:** txt file containing the HGNC IDs of the genes of interest (each HGNC ID in a separate line) *(mandatory option)*
- **genes_dict**: Gene ID dictionary file providing translation from HGNC ID to NCBI gene ID; a txt file with two columns, first column providing the HGNC ID and second column providing the NCBI gene ID. (This file can be downloaded from the HGNC BioMart website: https://biomart.genenames.org/)
- **appris**: APPRIS Principal Isoforms data file (downloaded from the APPRIS website: http://appris.bioinfo.cnio.es/#/downloads)

[MapNMs] section:

- **hgncid_to_symbol**: gene dictionary file providing translation from HGNC ID to gene symbol; a txt file with two columns, first column providing the HGNC ID and second column providing the gene symbol (downloaded from the HGNC BioMart website: https://biomart.genenames.org/)

- **more_symbols**: txt file supplying gene symbols missing from the above HGNC ID to gene symbol dictionary file

[EnsemblDB] section:

- **release_37**: Ensembl release version for build GRCh37
- **release_38**: Ensembl release version for build GRCh38

[SelectENSTs] section:

- **gene_synonyms**: txt file containing gene symbol synonyms, first column providing the gene symbol and second column providing the comma-separated list of synonyms (data downloaded from the HGNC BioMart website: https://biomart.genenames.org/)

[FormatCARTs] section:

- **series_37**: GRCh37 series code for CART IDs (e.g. CART37A) *(mandatory option)*

- **series_38**: GRCh38 series code for CART IDs (e.g. CART38A) *(mandatory option)*
- **canonical_37***:* txt file containing Ensembl canonical transcript for build GRCh37. First column is gene symbol, second column is the canonical ENST ID.
- **canonical_38***:* txt file containing Ensembl canonical transcript for build GRCh38. First column is gene symbol, second column is the canonical ENST ID.

Note that a configuration file template is provided in the CARTtools-x.x.x/config subfolder.

The CARTs pipeline includes a default APPRIS data file, default HGNC ID to NCBI gene ID / Gene Symbol dictionary files and gene synonym file (located in the CARTtools-x.x.x/default folder). If any of these options are not specified in the configuration file, the default file is used.

If Ensembl releases are not specified for any of GRCh37 or GRCh38 builds, Ensembl v75 and v92 are used by default.

5 Output files

The CARTs pipeline creates the following output files:

- **Genomic coordinates / sequences of the CARTs saved into various file formats (GenBank, genePred, GFF2, GFF3, FASTA) and transcript database file for CAVA:**

  - GFF2 file: <prefix>.gff2.gz

  - GFF3 file: <prefix>.gff3.gz

  - GenePred file: <prefix>.gp

  - GenBank files: <prefix>_gbk.zip (zipped folder containing the .gbk files for each CART)

  - FASTA file: <prefix>.fa (+.fai)

  - CAVA database: <prefix>_cava.gz (+.tbi)

(See *Section 6* for more details of the output file formats.)

- **Summary table of results (<prefix>_summary.txt):**

Each line of the table represents a gene described by the following 24 columns.

  - **GeneSymbol**

  - **HGNC_ID**

  - **NCBI.Gene:** NCBI Gene ID

  - **GDM_Colour**: gene 'colour' as defined by the Gene Disease Map (GDM) (https://osf.io/s4pva/)

  - **Algorithmic_NM**: NM selected by algorithmic pipeline

  - **Algorithmic_NM_Version**: version of NM selected by algorithmic pipeline

  - **UTRDifferenceType**: difference category/categories in the UTR selection step of algorithmic NM selection

- **UTRDecisiveCriteria**: decisive criteria applied in UTR selection step of algorithmic NM selection

- **GenomeDifference**: the value of the third column in the output file created by RefSeqCheck (see *Section 7.2*)

- **AlignmentDifference**: differences between the associated NM and GRCh37 derived from the CIGAR string provided by the NCBI RefSeq interim alignment file

- **NMs.in.RefSeq**: List of NMs in RefSeq for this gene

- **RefSeqGene_NMs**: List of RefSeqGene NMs for this gene

- **ClinVar_NMs:** List of ClinVar NMs for this gene

- **Community_NM:** NM selected as community transcript

- **CART associated NM:** Selected associated RefSeq transcript (NM)

- **CART_ASSOCIATED_ENST_37:** Selected associated ENST for build GRCh37

- **UTR_DIFF_37:** UTR difference between the mapped NM and the ENST selected for build GRCh37 (possible values: ".", "UTR3", "UTR5" or "UTR5,UTR3")

- **UTR_EXON_NUM_DIFF_37:** Any difference in the number of UTR exons between the mapped NM and the ENST selected for build GRCh37. The difference in the number of exons (relative to the number of exons in the NM) is also given: e.g. "UTR5:-1" means the ENST contains one exon less in its UTR5 than the NM, while "UTR5:+1,UTR3:+1" means both UTR5 and UTR3 have one more exon in the ENST than in the NM

- **CART_ASSOCIATED_ENST_38:** Selected associated ENST for build GRCh38

- **UTR_DIFF_38:** UTR difference between the mapped NM and the ENST selected for build GRCh38 (same values as for UTR_DIFF_37)

- **UTR_EXON_NUM_DIFF_38:** Any difference in the number of UTR exons between the mapped NM and the ENST selected for build GRCh38 (same values as for UTR_EXON_NUM_DIFF_37)

- **CART37_and_CART38_DIFF:** Difference of the Ensembl transcripts selected for GRCh37 and GRCh38 (i.e. semicolon separated flags describing differences in the CDS, UTR5 and UTR3 sequences of the two ENSTs. See *Section 7.7* for more details.)

- **CART_ID_37**: CART ID for GRCh37 (or ENST ID of selected Ensembl transcript)

- **CART_ID_38**: CART ID for GRCh38 (or ENST ID of selected Ensembl transcript)

- CART Pipeline Report:

Summary of most important information about the finished pipeline run and summary of results.

All additional intermediate files created by the pipeline (i.e. outputs of the different tools) are provided in the **cart_pipeline_files/** folder.

6 CART file formats

6.1 FASTA file

Detailed FASTA format specification:
https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp

A FASTA file can contain multiple sequences, with the description of each sequence beginning with a single-line description, followed by multiple lines of sequence data. The description line begins with a greater-than (">") symbol.

The CARTs pipeline outputs the concatenated exonic sequences of the CARTs in a single FASTA file.

The sequence description line contains the CART ID for each sequence.

For reverse-stranded CARTs, the reverse complemented sequence is included.

The FASTA file outputted by the CARTs pipeline is automatically indexed by Samtools faidx (.fai index file is provided).

6.2 GenePred (Gene Predictions) file

Detailed GenePred format specification:
http://genome.ucsc.edu/FAQ/FAQformat#format9

Note: The CARTs pipeline outputs the Extended GenePred (genePredExt) format

A 15-column, tab-delimited file in which each CART is described by a single line

Specific to the CARTs pipeline output:
- Column 1 (name): CART ID
- Column 12 (alternative name): HGNC ID

6.3 GFF3 (Generic Feature Format 3) file

Detailed GFF3 format specification:
https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

A 9-column, tab-delimited file in which each CART is described by multiple lines (referred to as features)

The CARTs pipeline outputs three types of features for each CART:
- transcript (a single line for each CART)
- exon (for each exon of the CART)
- CDS (for each CDS exon of the CART)

For the "transcript" feature line, the following attributes are included in column 9:
- ID: CART ID
- hgnc_id: HGNC ID of the gene, without the "HGNC:" prefix
- gene_symbol: gene symbol
- biotype: its value is set to "protein_coding" for all CARTs
- assoc_nm: associated NM
- assoc_enst: associated ENST

For "exon" and "CDS" features line, the following attributes are included in column 9:
- ID: automatically generated exon or CDS ID, derived from the CART ID
- Parent: reference to the ID of the corresponding transcript feature (the Parent ID is used to indicate a partof relationship)

Note: the CARTs pipeline outputs a chromosome/position-sorted, bgzipped, Tabix-indexed GFF3 file (.tbi index file is provided).

6.4 GBK (GenBank) file

Detailed GBK format specification:
https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

Each CART is described by a separate GBK file

Specific to the CARTs pipeline output:
- Locus name and ACCESSION are set to the CART ID
- Two features provided in the FEATURES section: mRNA and CDS
- Entire reference sequence spanned by the CART (introns included) is provided in the ORIGIN section
- If CART is reverse-stranded, the sequence provided in ORIGIN section is reverse complemented

6.5 CAVA database file

Tab-delimited file where each CART is described by a single line. Each line contains 11+2*N columns where N is the number of exons in the given CART.

Description of columns:
- CART ID
- Gene symbol
- HGNC ID (without the "HGNC:" prefix)
- TRINFO flag used in CAVA
- Chromosome
- Strand (1 or -1)
- Genomic start coordinate
- Genomic end coordinate
- cDNA coordinate of CDS start
- Genomic coordinate of CDS start
- Genomic coordinate of CDS end
- Following column pairs: Genomic start and end coordinates of each exon

Note: the CAVA transcript database is chromosome/position-sorted, bgzipped and Tabix-indexed (the .tbi index file is also provided).

7 Individual tools

In this section we describe each tool included in the CARTtools package so users can also run them individually.

7.1 RefSeqDB

This tool creates a RefSeq transcript database file that contains the mapped genomic coordinates of each NM transcript. The user can select from downloading the mapping from the NCBI interim release (13/01/2017) or UCSC (latest version of RefSeq). The database can be generated for either GRCh37 or GRCh38 genome build. RefSeqDB outputs a position-sorted, bgzipped, Tabix-indexed transcript data file and its Tabix index file (tbi). In addition, the tool creates two txt files containing the list of transcripts included and exlcuded from the database.

Running the tool from the command line:
CARTtools/refseqdb <options>

Options:
--build : Genome build (GRCh37 or GRCh38)
--mapping : Mapping source (ncbi or ucsc)
--output : Output file name prefix

The following output files are generated:
- <prefix>.gz (the actual transcript database including coordinate data)
- <prefix>.gz.tbi (Tabix index file for the database)
- <prefix>_included.txt (list of transcripts included in the database)
- <prefix>_excluded.txt (list of transcripts excluded from the database with reason for each; see below)

The second column of the <prefix>_excluded.txt file describes the reason the listed transcripts are excluded from the transcript database. Possible values:
- “joined_cds" (CDS coordinates could not be extracted, they were given in a 'joined_cds' format)
- “multiple_mapping” (Multiple mapping to GRCh37 are available for this transcript)
- “no_mapping” (No mapping available for this transcript on chromosomes 1-22, X, Y, MT)
- “missing_hgncid” (HGNC ID was missing from RefSeq data)
- “incorrect_cds_length" (CDS length provided by mapped coordinates not divisible by 3)

7.2 RefSeqCheck

The input of RefSeqCheck are the transcript database created by the RefSeqDB tool and the reference fasta file. RefSeqCheck generates a single output file containing all NM transcripts included in the RefSeq transcript database with an added flag describing any difference between the transcript sequence and reference genome sequence.

Running the tool from the command line:
CARTtools/refseqcheck <options>

Options:
--input : Input RefSeq database file (output of RefSeqDB)
--reference : Reference genome fasta file
--output : Output file name

The third column of the output file created by RefSeqCheck have the following possible values:
- “." (RefSeq sequence and reference sequence are identical)
- comma-separated c. annotations (RefSeq sequence and reference differ only in the listed substitutions)
- “lengthChange" (RefSeq sequence and reference sequence have different lengths)

7.3 SelectNMs

This tool selects an NM transcript for each gene on the input gene list. The transcript is selected from the NM chosen by an automatic transcript selection pipeline and gene transcripts used by the clinical genetics community. (Details of the selection process is discussed in our paper.) SelectNMs requires the APPRIS data file downloaded from the APPRIS website, the RefSeq transcript database (output of RefSeqDB), the output file of RefSeqCheck, and the Gene ID dictionary file.

Running the tool from the command line:
CARTtools/selectnms <options>

Options:
--input_genes : Input file containing HGNC IDs
--appris : APPRIS data file
--refsdb : RefSeq transcript database file (output of RefSeqDB)
--refsdbinc : List of transcripts included in the RefSeq DB (output of RefSeqDB)
--refschk : RefSeqCheck output file
--genes_dict : Gene ID dictionary file
--build : Genome build (GRCh37 or GRCh38)
--out_auto : Output file name prefix for automatic selection
--out : Output file name prefix for final selection

The output of SelectNMs is a table containing all genes (identified by HGNC ID) and the selected NM transcript. In addition, SelectNMs outputs the result of the automatic NM selection pipeline, the list of excluded genes with reason of exclusion, and a log file including the steps of NM selection for each gene.

7.4 MapNMs

This tool creates a file containing the mapped genomic coordinates of the NM transcript selected for each gene by the SelectNMs tool. The mapped coordinates are read from the RefSeqDB output files (NCBI/UCSC). If an NM is included in the NCBI coordinate file, the NCBI mapping is used, otherwise the UCSC mapping is outputted. If the NM is not included in any of the NCBI or UCSC coordinate files, the gene is excluded.

Running the tool from the command line:
CARTtools/mapnms <options>

Options:
--input : Input file (output of SelectNMs)
--ncbi : RefSeqDB output file with NCBI interim mapping data
--ucsc : RefSeqDB output file with UCSC mapping data
--hgncid_to_symbol : HGNC ID to Gene Symbol dictionary file
--output : Output file name prefix
--more_symbols: txt file for specifying gene symbols for missing HGNC IDs

Note that MapNMs checks if any HGNC IDs are missing from the HGNC ID to Gene Symbol dictionary file. If yes, the script terminates and prints out the list of these HGNC IDs asking the user to supply these in an additional text file with command line option --more_symbols.

MapNMs outputs a position-sorted, bgzipped, Tabix-indexed data file containing the mapped genomic coordinates of the NMs. It also creates a txt file reporting for each NM if it is represented by NCBI or UCSC mapping. Finally, it creates a txt files listing all excluded genes with reason of exlcusion for both the NCBI and UCSC coordinate files.

7.5 EnsemblDB

This tool creates an Ensembl transcript database file that contains the genomic coordinates of each ENST transcript in a particular Ensembl release. The user can specify the Ensembl release version and optionally provide an input list of ENST IDs to be retrieved.

Running the tool from the command line:
CARTtools/ensembldb <options>

Options:
--input : Input filename (list of ENST IDs)
--release : Ensembl release version
--output : Output file name prefix

EnsemblDB outputs a position-sorted, bgzipped, Tabix-indexed transcript data file and its Tabix index file (tbi).

7.6 SelectENSTs

This tool identifies the "most similar" Ensembl transcript in an Ensembl release to the selected NMs (considering their mapped genomic coordinates). Details of the ENST selection process are provided in our paper. SelectENSTs requires the outputs of EnsemblDB and MapNMs, and a txt file containing gene symbol synonyms. By default SelectENSTs matches the mapped NMs to ENSTs that either have the same gene symbol or a synonymous gene symbol. Optionally, the mapped NMs can be matched to ENSTs of any gene symbol.

Running the tool from the command line:
CARTtools/selectensts <options>

Options:
--mapped_nms : Mapped NMs data file (output of MapMMs)
--ensembl_data : Ensembl data file (output of EnsemblDB)
--gene_synonyms : Gene synonyms file
--any_gene : Match a mapped NM to ENSTs with any gene symbol
--input : Input file (list of NMs)
--output : Output file name prefix

SelectENSTs outputs a table listing all NMs with its "most similar" ENST. The table also contains additional columns describing the reason of ENST selection.

7.7 CompareENSTs

This tool compares the Ensembl transcripts selected for the the same gene for build GRCh37 and GRCh38. It outputs any differences in the CDS, UTR5 and UTR3 sequences of the two ENSTs. CompareENSTs requires the output of EnsemblDB for both GRCh37 and GRCh38, the table of selected ENSTs for both builds, and the reference genome fasta file for both builds.

Running the tool from the command line:
CARTtools/compareensts <options>

Options:
--ensts37 : Table of selected ENSTs for GRCh37 (output of SelectedENSTs)
--ensts38 : Table of selected ENSTs for GRCh38 (output of SelectedENSTs)
--data37 : Ensembl database for GRCh37 (output of EnsemblDB)
--data38 : Ensembl database for GRCh38 (output of EnsemblDB)
--ref37 : Reference genome fasta file (GRCh37)
--ref38 : Reference genome fasta file (GRCh38)
--output: Output file name
--discr : Output only ENST pairs where there are differences
--simple : Create simplified output

The output file has three columns. The first two columns describe the pair of ENSTs selected for GRCh37 and GRCh38, respectively. The third column may contain the following flags with the listed meaning:
- NF37 - transcript not found in build GRCh37

- NF38 - transcript not found in build GRCh38
- CDS:i - difference in the CDS sequence in exon i
- UTR5:i - difference in the 5' UTR sequence in exon i
- UTR3:i - difference in the 3' UTR sequence in exon i

A transcript pair may have multiple flags separated by semicolons.

If the transcript pair does not have any flag (i.e. '.' value is given in column 3), this means the entire mRNA sequences (UTR5+CDS+UTR3) are identical in the two builds.

7.8 FormatCARTs

This tool outputs the CARTs in various common file formats (GenBank, genePred, GFF2, GFF3, and FASTA) as well as creating a transcript database for CAVA. FormatCARTs requires the outputs of SelectENSTs and EnsemblDB, as well as the reference genome fasta file. Optionally, the CART ID numbering can be continued from a previous FormatCARTs output.

Running the tool from the command line:
CARTtools/formatcarts <options>

Options:
--selected_nms : Final selected NMs file (output of SelectNMs)
--selected_ensts : Selected ENSTs file (output of SelectENSTs)
--canonical : Canonical ENSTs file
--ensembl : Ensembl transcript database (output of EnsemblDB)
--series : CARTs series code (e.g. CART37A)
--gbk : Create GBK output
--ref : Reference genome file
--annovar : Create GenePred and FASTA files for Annovar
--prev_cava_db : CAVA db output of previous run from which CART numbering will be continue
--prev_ref : Reference genome of previous run from which CART numbering will be continued
--output : Output file name prefix

FormatCARTs generates the output files discussed in *Section 5.*