GeneCaRNA: A Comprehensive Gene-centric Database of Human Non-coding RNAs in the GeneCards Suite

Ruth Barshir 1*, Simon Fishilevich 1, Tsippi Iny-Stein 1, Ofer Zelig 2, Yaron Mazor 2, Yaron Guan-Golan 2, Marilyn Safran 1 and Doron Lancet 1*

1 - Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 7610001, Israel
2 - LifeMap Sciences Inc., Alameda, CA 94501, USA

Correspondence to Ruth Barshir and Doron Lancet: ruth.barshir@weizmann.ac.il (R. Barshir), simon.fishilevich@weizmann.ac.il (S. Fishilevich), tsippi.iny-stein@weizmann.ac.il (T. Iny-Stein), oferz@lifemapsc.com (O. Zelig), ym@lifemapsc.com (Y. Mazor), yg@lifemapsc.com (Y. Guan-Golan), marilyn.safran@weizmann.ac.il (M. Safran), doron.lancet@weizmann.ac.il (D. Lancet)
https://doi.org/10.1016/j.jmb.2021.166913
Edited by Michael Sternberg

Abstract

Non-coding RNA (ncRNA) genes assume increasing biological importance, with growing associations with diseases. Many ncRNA sources are transcript-centric, but for non-coding variant analysis and disease decipherment it is essential to transform this information into a comprehensive set of genome-mapped ncRNA genes. We present GeneCaRNA, a new all-inclusive gene-centric ncRNA database within the GeneCards Suite. GeneCaRNA information is integrated from four community-backed data structures: the major transcript database RNAcentral with its 20 encompassed databases, and the ncRNA entries of three major gene resources HGNC, Ensembl and NCBI Gene. GeneCaRNA presents 219,587 ncRNA gene pages, a 7-fold increase from those available in our three gene mining sources. Each ncRNA gene has wide-ranging annotation, mined from >100 worldwide sources, providing a powerful GeneCards-leveraged search. The latter empowers VarElect, our disease-gene interpretation tool, allowing one to systematically decipher ncRNA variants. The combined power of GeneCaRNA with GeneHancer, our regulatory elements database, facilitates wide-ranging scrutiny of the non-coding terra incognita of gene networks and whole genome analyses.

© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Protein-coding genes constitute a minor part (1–2%) of the human genome. The development of massive parallel sequencing, which enabled one to analyze transcriptomes at a deeper level, revealed that a much larger genome fraction is transcribed, encoding non-coding RNAs (ncRNAs), a diverse group of biomolecules that lack clear potential to encode proteins.1,2 This includes the long-known ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), small regulatory ncRNAs exemplified by microRNAs (miRNAs) that regulate gene expression in various cell types, as well as piwi-interacting RNAs (piRNAs) involved in the regulation of transposon expression in germ cells.3,4 Also included are long non-coding RNAs (lncRNAs), defined arbitrarily as transcripts of more than 200 nucleotides that do not belong to any other well-defined type. Such genes have been implicated in a wide array of cellular processes, including transcriptional regulation, differentiation, and cellular reprogramming.5

There is growing evidence that ncRNA genes significantly contribute to the development of a large number of different human disorders. This
encompasses many ncRNA types, including miRNAs and their processing machinery, snoRNAs, piRNAs and lncRNAs\textsuperscript{5-7}. Such genes are coveted targets for drug discovery\textsuperscript{2,10}, significantly enhancing the need to carry out whole genome sequencing (WGS) analyses in order to fathom the extensive non-coding genomic territories that encode ncRNAs\textsuperscript{11}. This trend necessitates parallel development of databases and analytical tools for the disease interpretation of ncRNAs.

The gene-centric ncRNA world seems to lag behind the protein-coding equivalent, as indicated by there being only \(\sim 8,000\) ncRNA genes with an HGNC approved symbol\textsuperscript{12}. There are major community-acknowledged sources for defining genes in the human genome \(\sim 13-15\), but their repertoire definition rests on ample functional characterization. While currently satisfactory for protein-coding genes, the rapidly emerging vast universe of ncRNA genes requires a different strategy, namely the generation of a comprehensive list of genomically-defined gene candidates, providing infrastructure for further biomedical scrutiny. Significantly, the genome-wide characterization of non-coding transcripts is much more advanced, stemming from extensive worldwide RNA sequencing efforts.

In this paper we describe the establishment of GeneCaRNA, a comprehensive non-redundant gene-centric integrative view of human ncRNA genes and their annotations, covering the entire terra incognita of non-coding RNA genes. We complement the major but limited sources of ncRNA genes by a transcript-to-gene transformation of all RNAcentral\textsuperscript{16} transcripts, followed by gene candidate integration based on genomic coordinates. We leverage the comprehensive collection of ncRNA transcripts from RNAcentral\textsuperscript{16}, in turn obtained from 20 different transcript resources, with every ncRNA sequence mapped to the genome. An early GeneCards effort to create a non-redundant compendium of ncRNA genes was published in 2013\textsuperscript{17}. GeneCaRNA represents a major improvement relative to this previous version, featuring numerous novel data sources, an improved transcript clustering algorithm, significantly enhanced functional annotations, and comprehensive visualization via a hub at the UCSC genome browser\textsuperscript{18}.

Results

Comprehensive set of ncRNA genes

To create GeneCaRNA (https://www.genecards.org/genecarna), a comprehensive non-redundant gene-centric resource of human non-coding RNAs, we integrated records from three gene sources, HGNC\textsuperscript{18}, NCBI Gene\textsuperscript{13} and Ensembl (ENSG)\textsuperscript{14}, and a comprehensive transcript source, RNAcentral\textsuperscript{16} which mines, quality assures and provides information regarding 20 individual transcript sources (Figure 1a). Sources beyond the above four directly-mined sources are mined indirectly via RNAcentral exclusively, leading to the inclusion in GeneCaRNA of only transcripts judged to be sufficiently reliable by RNAcentral. First, we mined all 250,701 unique transcript sequences from RNAcentral with their provided mapped genome coordinates. This information allowed us to define 209,874 Candidate Genes. In parallel, the genomic coordinates of all ncRNA genes in the major gene sources (HGNC, NCBI Gene and Ensembl) were recorded. The pool of all genomic coordinates thus obtained was subjected to an integration process that takes into account a gene’s functional category and exon structure, avoiding clustering transcripts with overlapping genomic territory but no exon overlap (See Methods). This resulted in a comprehensive list of 219,587 GeneCaRNA genes, amalgamated from the four sources. Significantly, only 22,285 of the Candidate Genes existed in any of the major gene sources; the remaining 187,589 records are defined as novel Transcripts Inferred GeneCaRNA genes (TRIGGs). We also note that 9,622 entries from NCBI Gene are related to model transcripts unique to this source\textsuperscript{19} (Figure 1b).

Our previous non-redundant compendium of ncRNAs (in GeneCards version 4, v4) included \(\sim 100,000\) ncRNAs. In GeneCaRNA, embedded within GeneCards version 5 (v5), the ncRNA gene count is more than twice as large (\(\sim 220,000\), Figure 1c). Further, while in v4 the total number of ncRNA genes was three times larger than that in HGNC, NCBI Gene and Ensembl combined, this comparative ratio went up seven-fold in GeneCaRNA. The total coverage of ncRNA gene territories (including exons and introns) increased in v5 by a factor of 1.5, from 26% to 40%. Most of this enhanced genome coverage stems from the addition of RNAcentral transcripts. Relatedly, the exonic genome coverage of GeneCaRNA is 5%, of which half is contributed by our TRIGGs. The exon genome coverage contributed by RNAcentral’s transcripts, both enriching major gene sources genes and in TRIGGs, is three times larger than that of the major gene sources (1.6%). We also note that ncRNA exon coverage in GeneCaRNA is two and a half times larger than that of protein-encoding exons (1.9%)\textsuperscript{20} (Figure 1c).

Within the arsenal of \(\sim 220,000\) GeneCaRNA genes entries, \(\sim 110,000\) genes are found by our algorithm (see Methods) to have a single genomic copy, while almost half (\(\sim 109,000\)) belong to groups containing 2–768 duplicate genomic copies (Figure 1d). The propensity of such gene duplication is highly varied among different types of ncRNA genes, with up to 83% of the genes belonging to multi-loci groups in the piRNA type (Figure 1d, Figure S1). In a group of gene duplicates as defined here, all members are transcribed into ncRNA transcripts with the same
sequence, hence collectively representing a paralog-like functional entity. We then can compute the count of unique ncRNAs to be 115,749.

**GeneCaRNA user interfaces**

GeneCaRNA (URL https://www.genecards.org/genecarna) is a searchable, integrative database that provides comprehensive, user-friendly information on all ncRNA human genes. It inherits its capabilities from the GeneCards framework within which it is embedded. Thus, each GeneCaRNA entry has a dedicated Web page (GeneCard), presenting a comprehensive view of gene annotations integrated from multiple sources. The GeneCards infrastructure encompasses - 178 automatically mined data sources, and a large number of them is available for ncRNA annotations,
Figure 2. a. GeneCaRNA UCSC hub view. Shown is a 26 kb genomic locus (chr1:29,400–55,400, hg38) encompassing 5 ncRNA genes. Two tracks are shown: the combined gene track (top) and the combined transcript track (bottom). ncRNA types are color coded, with transcripts in a lighter hue. Shown in this view are IncRNA (red), miRNA (blue) and piRNA (green). b. GeneCards page views statistics. Rank plot of GeneCards’ top-visited Cards, with protein coding in blue, and ncRNAs in orange. Naturally, many more protein-coding genes get significant attention, but some ncRNAs, the top ones being MALAT1, TERC and NEAT1, ranked 88, 164 and 315 respectively, are also quite popular. Such performance is expected to be greatly augmented by GeneCaRNA. These statistics represent GeneCards page visits between 1-Jan-2020 and 28-Jul-2020.

including links to the external sources for more detailed scrutiny (Figure S2).

The powerful GeneCards search functionality and index\textsuperscript{22} produce a list of gene symbol “hits”, sorted by an evidence-based relevance score, in response to user-defined Boolean queries. Detailed instructions are provided at https://www.genecards.org/Guide/Search. When a search is performed from the GeneCaRNA home page, results are filtered to show only ncRNA gene hits. Notably, when query terms contain information shared with protein coding genes, e.g. participation in the same pathway\textsuperscript{22}, sharing a Gene Ontology term\textsuperscript{23} or having relation to the same disease\textsuperscript{24}, the evidence shown reflects such connections. The advanced search mode supports focused querying by searching for specific terms in specific sections of the Card. Hit contexts, i.e. the complete evidence associating each hit with query term(s), can be explored by opening the mini-cards, which show the specific fields in relevant Card sections in which the term(s) were found.

GeneCaRNA-based searches are leveraged by VarElect\textsuperscript{25} the Suite’s NGS phenotyper, which helps effectively and rapidly identify and prioritize direct and indirect associations between genes and user-supplied disease terms, together with extensive evidence for such associations. VarElect infers indirect (“guilt by association”) relationships between genes and phenotype keywords by exploiting diverse gene-to-gene relationships generated using the GeneCards search engine to search for gene symbols in selected sections.

Further, GeneCaRNA search results can be exported to (1) GeneALacart, the Suite’s batch query facility\textsuperscript{51}, to extract GeneCards annotations of interest for all of the genes in the set; and (2) GeneAnalytics, the Suite’s gene set enrichment tool\textsuperscript{76}.

To enable exploring GeneCaRNA records within a genome browser, we developed the GeneCaRNA UCSC\textsuperscript{18} hub. The hub displays and their associated transcripts in the UCSC genome browser (Figure 2a). The GeneCaRNA hub is a track set containing nine gene tracks and nine transcript tracks, each in a different color. It includes one combined track for all ncRNA genes and a separate gene track for each ncRNA type. Each gene is represented by its exon structure, created by merging overlapping exons of all of its clustered transcripts (See Methods). Similarly, the hub includes a combined track for all ncRNA transcripts, and a separate transcript track for each type. The hub can be launched using dedicated links in GeneCaRNA’s Genomics and Transcripts sections, as well as directly from the UCSC genome browser (by specifying https://genecards.weizmann.ac.il/geneloc/ncRNA_hub/hub.txt in the UCSC hub upload page).

GeneCaRNA annotations

Each GeneCaRNA gene is identified by a unique gene symbol, derived from multiple data sources, using a hierarchical symbol assignment algorithm (Methods, Table S1). We provide first priority to the ~8,000 ncRNA genes already assigned a symbol by the Human Genome Nomenclature committee (HGNC\textsuperscript{15}). Additional ~24,000 symbols are derived from NCBI Gene and Ensembl. All remaining ncRNA genes are assigned symbols based on RNAcentral and its sources. In addition, for each gene, we portray a comprehensive list of aliases based on the aforementioned sources (Figures 1 and S2).

A great advantage of being part of the GeneCards knowledgebase is that GeneCaRNA genes are automatically annotated with a plethora of integrated data in the sections defined in GeneCards. To extend the characterization of ncRNAs, we developed additional ncRNA-specific annotations, summarized below.

GeneCaRNA genes are cataloged into 17 types, based on similar partitions in 4 major data sources
The largest type (109,820 genes) is piRNA, amounting to ~50% of all GeneCaRNA genes; the second largest type is IncRNA, with 75,839 genes, ~35% of the compendium; the rest encompass the remaining 15 types (Figure 1d). In parallel, the GeneCards/GeneCaRNA Summaries section includes mapping to RNA families by sequence alignment and secondary structure from Rfam.

A table in the GeneCards/GeneCaRNA Transcripts section summarizes integrated transcript records assigned to an ncRNA gene in the clustering procedure, with external links to all relevant transcript sources for the gene. Visualization of the transcript cluster is triggered by clicking the customized link to the GeneCaRNA UCSC track hub.

GeneCaRNA genes are associated with regulatory elements by GeneHancer, the GeneCards Suite database of human regulatory elements and their inferred target genes. Because enhancer–gene mapping depends on several methods (e.g. eQTL, capture HiC) for which the needed data are scarce for newly defined ncRNAs, TRIGGS receive regulatory annotations based only on genomic proximity.

As previously reported, some RNA genes may encode a peptide product. In the current GeneCaRNA version 390 genes have evidence of peptide product from Uniprot, and these are now shown in the Card’s Proteins section. We note that our 390 cases include all those shown in.

Gene-disease associations are a central focus of the GeneCards Suite, which includes the affiliated disease database MalaCards. Whereas 73% of protein-coding genes have MalaCards disease affiliation at different evidence levels, a much lower fraction (1%) of ncRNA genes is similarly disease-
associated (Figure 3b). To demonstrate an additional path for augmenting this low ncRNA annotation coverage regarding diseases and phenotypes, we integrated EBI’s GWAS Catalog in the Function section. While this augmented the “saturated” annotation landscape of protein coding genes by a factor (X1.07), there was an impressive annotation depth increase (X7) for ncRNA genes (Figure 3c). Significantly, the improved annotations by variants residing within exons is mostly attributed to our novel TRIGGs. Finally, the GeneCaRNA Variants section shows variants from Clinvar and dbSNP with different molecular consequences and with varying levels of evidence as to their clinical significance. Such variants are selected by the sources as related to clinical genetics linkage analyses, hence may reside upstream or downstream to the nominal gene territory in the genome maps.

Discussion

The major gene-centric sources are highly curated, and mostly focus on well-studied genes. Transcript data are accumulated by other sources at a much faster rate, but gene definition for such ncRNAs is often lacking. To enhance the universe of ncRNA gene data, we created GeneCaRNA within the powerful user interface of the GeneCards Suite. GeneCaRNA obtains transcript data from the transcript-centric RNACentral database, indirectly mining its 20 source databases having human data, defining transcript-inferred GeneCaRNA Genes (TRIGGs). We then integrate such data with ncRNA genes from the major gene sources, to define genomic coordinates for every ncRNA gene in the human genome. Thus, GeneCaRNA increases the availability of ncRNA gene entries by a large factor (7 fold) compared to the HGNC, NCBI Gene and Ensembl repertoires.

GeneCaRNA provides users with an orderly compendium of >200,000 ncRNA gene entries, each with a gene symbol and ncRNA type affiliation. The latter entities are also integrated from multiple sources, in ways that portray relationships to the original sources.

The GeneCaRNA gene-centric view is crucial for interpretation of clinical genomic sequencing for several reasons. It is useful for functional annotation of ncRNAs with data from many existing repositories that increasingly add ncRNA data and use gene rather than transcript as their core entity. Examples of such resources are OMIM, Gene Ontology and KEGG. It is also crucial for generating a genomic map facilitating the clear visualization of the relative positions of different genes in the genome, as well as their relationships to regulatory elements, as exemplified by the GeneCards UCSC tracks. Importantly our strategy allows GeneCaRNA to be an integral part of the GeneCards framework, inheriting an array of capabilities and tools that empower comprehensive and facile navigation of the ncRNA gene universe. The GeneCaRNA search has the potency of the GeneCards search, and users can choose between a mode that returns only ncRNA entries, and another that allows a comprehensive display of all gene category results.

Genes belonging to the ncRNA category are generally poorly annotated relative to protein-coding genes, and the TRIGG entries we have generated from transcripts have even lower annotation levels (Figure S3). A key goal of GeneCaRNA is to serve as a platform that reflects the gradually increasing annotation levels for ncRNAs as research progresses. Towards that goal, GeneCaRNA does not exclude genes based on their current annotation level. By being part of the GeneCards Suite, GeneCaRNA genes are automatically annotated by relevant sources among the 178 sources integrated in the GeneCards knowledgebase, including 17 sources introduced specifically for the development of GeneCaRNA. An example of immediate annotation enrichment for ncRNA genes, made possible by a GeneCards affiliation, are ncRNA genes’ associations with regulatory elements via GeneHancer (Figure S4). A second annotative enhancement relates to the GeneCards gene-gene and disease-gene matrices, which enable one to generate networks that connect ncRNA genes to genes of other categories, as well as to diseases cataloged in MalaCards (Figure 3a).

Our results point to 40% of the genome being covered by ncRNA gene territories. This highlights the significance of ncRNA scrutiny for disease interpretation. Indeed, a growing number of non-coding genes has become associated with diseases and phenotypes, and GeneCards ncRNA gene searches attest to this emerging interest (Figure 2b). The functional entities that mediate disease are ncRNA transcripts, as documented in many of GeneCaRNA’s data sources. However, efficient clinical variant analyses necessitate a comprehensive genomic coordinate map of all ncRNA genes across the genome, as provided by GeneCaRNA. This is crucial for turning what at present constitutes mostly clinical terra incognita into fertile clinical terrain. It is noteworthy that half of the ncRNA exonic territories reside in novel TRIGG-classified genes that are not found in any of the currently available variant-to-genome mapping data. Further, about 100,000 of all GeneCaRNA genes (nearly 50%, mostly TRIGGs) constitute duplicated genomic loci with identical transcripts. While such multiplicity appears to have minimal functional ramifications, full documentation of all of them is crucial for variant analyses, e.g. in cases of dominant negative inheritance.
A growing challenge in the decipherment of the genetic underpinnings of human diseases is a full capacity to tackle variations in ncRNA genes, when performing medically oriented whole genome sequencing. Two complementary tools within the GeneCards Suite are geared to address this challenge. The first is GeneALaCarte, a multigene batch search facility that extracts user-selected annotations for all genes submitted as a set\textsuperscript{21}. The second is VarElect\textsuperscript{23}, an NGS interpretation tool, which rapidly identifies and prioritizes direct and indirect associations between genes and user-supplied disease terms, together with extensive evidence for such associations. The indirect mode exploits diverse gene-to-gene relationships generated by the GeneCards/GeneCaRNA search engine.

The establishment of GeneCaRNA offers vistas for applying all of these tools to a large number of genes with a variable level of annotation. This is made possible by the fact that GeneCaRNA, like GeneCards, is all-inclusive and provides a mode of scientific communication based on unique gene identifiers and current annotation status. This serves the community by gradually enhancing the annotation of “unverified” genes to eventually gain the status of “verified”, and in doing so supports the progress of scientific scrutiny.

Materials and methods

Clustering RNAcentral transcripts

ncRNA transcripts data were obtained from version 15 of RNAcentral\textsuperscript{16} at the European Bioinformatics Institute (EBI)\textsuperscript{16}, including information from its source databases with human data (Table S1). We included only human transcript records having genomic mapping, exon definition (genome version hg38) and annotations such as identifiers and RNA type.

We applied a clustering algorithm to all transcripts as follows. Transcripts were clustered based on three criteria: 1. Belonging to the same ncRNA type, based on a unification table (Table S2); 2. Mapped to the same DNA strand; 3. Having overlap of at least one exon pair amounting to 70% of the shorter of the two. Transcripts of RNAcentral type ‘miscRNA’ were clustered with transcripts of other types, but only when all of the transcripts overlapping with the miscRNA record belong to a single additional type. Such clusters are considered as Candidate Genes for the next step. We further define gene multiplicity by identifying all genomic loci sharing the same transcript combination and having the same length of genome mapping. Each such group of loci constitutes unique ncRNAs genes (cf. Figure 1d).

Defining GeneCaRNA genes

A GeneCaRNA gene belongs to one of two different divisions: the first constitutes those mined from the major gene sources HGNC, Ensembl and NCBI Gene, as implemented in the standard GeneCards gene integration process\textsuperscript{21}. The second division constitutes the Transcripts Inferred GeneCaRNA Genes (TRIGGs) as described below.

To avoid defining new genes where major gene sources genes exist, we execute a merger of these genes with Candidate Genes where relevant. Merger to a specific gene occurs if one or more of the Candidate Gene’s transcripts have an RNAcentral source annotation of HGNC\textsuperscript{15}, Ensembl\textsuperscript{14}, or RefSeq\textsuperscript{17}, and the merger is directed by the identity of the gene name annotation of the transcript with the gene name in the major gene source. Merger also requires that the Candidate Gene and the merge target gene have a genomic territory overlap. The genomic territory of such a merged gene inherits the full extent of this overlap and is adjusted accordingly. If a Candidate Gene matches two or more merge target genes, the Candidate Gene is recorded as affiliated with all matching genes, without adjusting the coordinates.

Once merger is executed, the merged Candidate Gene is taken off the TRIGG list. A minority set of 576 ncRNA transcripts merged into 77 protein-coding genes, and are currently not included in GeneCaRNA; in future versions we will explore defining genes with more than one category.

Unmerged Candidate Gene transcript clusters define novel TRIGGs. Each TRIGG is associated with all clustered transcripts and its coordinates span the overall territory of all of its transcripts. Each GeneCaRNA gene, either TRIGG or major gene sources-based, inherits all annotations of its associated transcripts. Consequently, the genomic coordinates of 96% of the GeneCaRNA genes contain RNAcentral transcripts. The minority 4% are genes defined by major gene sources, not including any RNAcentral transcripts, in which case the coordinates are derived from the relevant major gene sources.

Defining gene types

GeneCaRNA genes are classified into types inherited from their annotating sources using a conversion table (Table S2). Types are based on those derived from the major gene sources, where applicable, employing the following hierarchy: NCBI Gene, HGNC, Ensembl. The type of a TRIGG is assigned from the shared type of all members of the transcript cluster (see criterion 1 in section Clustering RNAcentral transcripts above). TRIGG genes that include miscRNA transcripts in addition to transcripts of a different type inherit their type from the non-miscRNA transcripts.
Gene symbols

GeneCaRNA genes are assigned gene symbols by a hierarchical algorithm. For genes from major gene sources, the legacy GeneCards symbol-assigning algorithm is used, which chooses the HGNC symbol if it exists, followed by the NCBI symbol if it exists, and the Ensembl symbol (ENSEG00...) otherwise. For TRIGGs, since they do not have symbols from the major gene sources, we use identifiers from other sources, with a source-rank-based election hierarchy as defined in Table S1. If there are multiple candidate symbols from the highest ranked source, the first in alphabetical order is chosen. If genes with multiple genomic copies are assigned an identical symbol, a numerical suffix (e.g., 001, 002, ...) is used, ordered by genomic location. All contributing source identifiers not selected as gene symbols are provided as gene aliases. Other modifications of symbols relative to those provided by a source are: for Ensembl, the version identified (0.1, 0.2 etc.) is removed; for ENA, all text beginning with a first dot is removed; and for SNOMDB, the prefix ‘SNOMDB’ is added to the numeric identifier.

Multi-loci gene groups

We define multi-loci gene group as a group of identical copies of a gene in different genomic locations. Genes are considered identical copies if they share RNACentral transcript combination (indicated by transcripts identifiers) and have the same gene-territory length.

UCSC hub viewer

The gene exon-intron structure can be viewed in the GeneCaRNA hub at the UCSC genome browser. In its gene track, gene exons are constructed by merging the exons of the different RNACentral transcripts, and then mapped to the genome. In the transcript track, all transcripts are displayed.

Phenotype association by GWAS

Associations between SNPs, phenotypes and genes were extracted from the current GeneCards table that appears in the Function section of GeneCards/GeneCaRNA Cards, in the “Phenotypes From GWAS Catalog” table, in the “Gene Relation” column. These data are mined from the EBI’s GWAS catalog ‘All associations v1.0.2’ file. For Figure 1c (light purple columns) the gene to phenotype relations appear as mined from the published data. For the dark purple columns we performed an additional mapping of SNPs to all gene exons, whereby we could also get, for the first time, results for the previously unexplored TRIGGs. We note that the same GeneCards/GeneCaRNA table includes mapping of SNPs to regulatory elements as portrayed in GeneHancer, associating such regulatory SNPs to target genes, including all GeneCaRNA’s ncRNA entries.

Comparison with the previous GeneCards ncRNA compendium

All analyses in this manuscript are based on GeneCards Suite version 5.0 (v5). The sources integrated in this version were downloaded during June 2020. Data for the comparison of the new GeneCaRNA database with our earlier ncRNA compendium was taken from the last GeneCards version that included this information, V4.11 (v4), released in June 2019.

Data updates

GeneCaRNA, as part of the GeneCards Suite, is typically updated with each major version release three times a year and several minor versions in-between. In major releases, the entire GeneCards Suite knowledgebase is updated, including downloading the latest available versions of our input databases, rebuilding the genes and updating their annotations; minor versions incorporate localized improvements for particular Suite members.

Research data

GeneCaRNA data are incorporated into the GeneCards database, making it freely available for educational and research purposes by non-profit institutions at https://www.genecards.org/. GeneCards are available through the GeneALaCart batch query tool within the GeneCards Suite. Data dumps are available upon request. All data access options are described at https://www.genecards.org/ Guide/Datasets linked from the main GeneCaRNA and GeneCards menu under the title “Data Access”. GeneCaRNA data can also be visualized at the University of California Santa Cruz (UCSC) Genome Browser, as exemplified for the region shown in Fig. 2A and linked to in the Genomics section of each relevant GeneCard.

CRediT authorship contribution statement

Ruth Barshir: Data curation, Formal analysis, Investigation, Methodology, Software, Writing - original draft, Writing - review & editing. Simon Fishilevich: Data curation, Formal analysis, Investigation, Methodology, Software, Writing - review & editing. Tsippi Iny-Stein: Validation, Writing - review & editing. Ofer Zelig: Software, Validation, Writing - review & editing Yaron Mazor: Software, Validation. Yaron Guan-Golan: Software, Supervision. Marilyn Safran:
Supervision, Validation, Writing - review & editing. **Doron Lancet**: Conceptualization, Funding acquisition, Methodology, Supervision, Writing - review & editing.

Acknowledgements

Grant support is provided by LifeMap Sciences Inc., California, USA; PIONEER, a European Network of Excellence for Big Data in Prostate Cancer; DC-ren, Horizon 2020 EU grant on Drug combinations for rewriting trajectories of renal pathologies in type II diabetes. We thank Anton I. Petrov and Blake Sweeney for their generous cooperation, and for fruitful discussions regarding the best use of their RNACentral database.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2021.166913.

Keywords: comprehensive ncRNA compendium; RNACentral; non-coding universe; disease-gene interpretation; whole genome sequencing

Abbreviations:

TRIGG, Transcripts Inferred GeneCaRNA genes; WGS, Whole Genome Sequencing; eQTL, Expression quantitative trait locus; GWAS, Genome Wide Association Studies; SNP, Single Nucleotide Polymorphism

References


