RNA Journal

Research Article

Expanding and enriching the LncRNA gene landscape using the GeneCaRNA database

Shalini Aggarwal[1,*], Chana Rosenblum[1], Marshall Gould[2], Shahar Ziman[1], Ruth Barshir[3], Ofer Zelig[4], Yaron Guan-Golan [4], Tsippi Iny-Stein [1], Marilyn Safran [1], Shmuel Pietrokovski[1], Doron Lancet [1,*]

[1]Department of Molecular Genetics, Weizmann Institute of Science, Herzl 234, Rehovot 7610010, Israel
[2]Department of Biological Sciences. University College London, Gower Street, London, WC1E 6BT. United Kingdom
[3] TAD Center for AI and Data Science, Tel Aviv University, Tel Aviv 6997801, Israel
[4]LifeMap Sciences Limited, Wilson House 1001-2, 19 Wyndham Street, Central, Hong Kong, and LifeMap Sciences Inc., Alameda, CA 94501, USA.
*Correspondence: Shalini Aggarwal shalini.aggarwal@weizmann.ac.il and Doron Lancet doron.lancet@weizmann.ac.il

Abstract: The GeneCaRNA human gene database is a member of the GeneCards Suite. It presents ~280,000 human non-coding RNA genes, identified algorithmically from ~690,000 RNAcentral's transcripts. This expands by ~tenfold the ncRNA gene count relative to other sources. GeneCaRNA thus contains ~120,000 long non-coding RNAs (LncRNAs, >200 bases long), including ~100,000 novel genes. The latter have sparse functional information, a vast terra incognita for future research. LncRNA genes are uniformly represented on all nuclear chromosomes, with 10 genes in mitochondrial DNA. Data obtained from MalaCards, another GeneCards Suite member, finds 1,547 genes associated with 1 to 50 diseases. ~15% of the associations portray experimental evidence, with cancers tending to be multigenic. Preliminary text mining within GeneCaRNA discovers interactions of LncRNA transcripts with target gene products, with 25% being ncRNAs and 75% proteins. GeneCaRNA has a biological pathways section, which at present shows 131 pathways for 38 LncRNA genes, a basis for future expansion. Finally, our GeneHancer database provides regulatory elements for ~110,000 LncRNA genes, offering pointers for co-regulated genes and genetic linkages from enhancers to diseases. We anticipate that the broad vista provided by GeneCaRNA will serve as an essential guide for further LncRNA research.

Keywords: Non-coding RNA, long non-coding RNA, human noncoding RNA, LncRNA, GeneCaRNA, GeneCards

1. Introduction

As our understanding of the human genome deepens, the spotlight on non-coding RNAs (ncRNAs) has intensified, revealing a diverse landscape of non-coding transcripts. Currently, the major sources (HGNC (Seal et al., 2020), NCBI Gene (Brown et al., 2015), Ensembl (Harrison et al., 2024) and RNAcentral (The RNAcentral Consortium et al., 2019)), portray ~38,000 well-documented ncRNA genes. In our earlier paper (Barshir et al., 2021), we significantly augmented the count of identified ncRNA genes to 220,000, obtained based on algorithms that analyzed all transcripts in RNAcentral (The RNAcentral Consortium et al., 2019). Our ncRNA gene category consists of 24 classes, with the highest counts belonging to LncRNAs, piRNAs and miRNAs.

GeneCaRNA's *raison d'etre* within the GeneCards Suite is to provide a comprehensive gene-centric knowledgebase for human ncRNA genes and their annotations. The present version 5.19 GeneCaRNA exhibits ~280,000 human non-coding RNA genes, stemming from ~690,000 RNAcentral transcripts. A key advancement in GeneCaRNA is the addition of a very large number of potentially functional ncRNA genes to a gene database that enables the use of a broad range of other types of genomic and biological data and offers many bioinformatic analysis tools.

GeneCaRNA is a multi-source database, offering a promising avenue to enrich annotation information and functions on a single platform. GeneCaRNA effectively addresses the challenge of disparate gene names across sources by leveraging aliases, facilitating accurate gene mapping, and providing comprehensive gene details. GeneCaRNA is frequently updated to address novel knowledge and automatically add functional annotation and disease links which also address the novel Transcripts Inferred GeneCaRNA genes (TRIGGs)(Barshir et al., 2021).

LncRNA genes constitute the largest class within the ncRNA gene category, ~46% of the total. This gene class is defined for transcripts longer than 200 nucleotides, which have scant or no protein-coding characteristics (Kapranov et al., 2007; Wang et al., 2019). The large size of the LncRNA class led to further definition of subclasses, based on criteria such as transcriptional machinery (Mattick et al., 2023), structure (Aliperti et al., 2021), expression regulation (Ma et al., 2013), location with respect to protein coding genes (Frankish et al., 2021; Sun et al., 2018), and binding targets (Ma et al., 2013), as well as transcriptional and post-transcriptional regulation (Sun et al., 2018).

This paper describes thorough analyses of the LncRNA class of human genes based on the power of the GeneCards Suite. GeneCaRNA provides textual analyses on annotations of each LncRNA GeneCard in the database and reveals information for items such as revised sub-classification, interactions between genes products (protein and transcripts), gene to disease relationships, and relations of genes to biological pathways. GeneCaRNA is poised to become a highly useful platform for accessing complete profiles of LncRNAs.

2. Results

2.1. Data extraction and text mining

We extracted 120,982 LncRNA genes from GeneCaRNA, as portrayed in **Supplemental Table S1**. Of these, 19,482 genes are available in the major sources, and 101,500 genes are unique GeneCaRNA TRIGGs. Both categories were explored for annotative information based on their GIFtS value, which represents the cumulative information imported from all the relevant sources (Harel et al., 2009). We found that while 34% of the major source genes had GIFtS > 10, none of the TRIGGs had a GIFtS > 10 (**Fig. 1A**). This indicates the fertile land for future research in the realm of TRIGGs, both for LncRNA genes and other ncRNA categories.

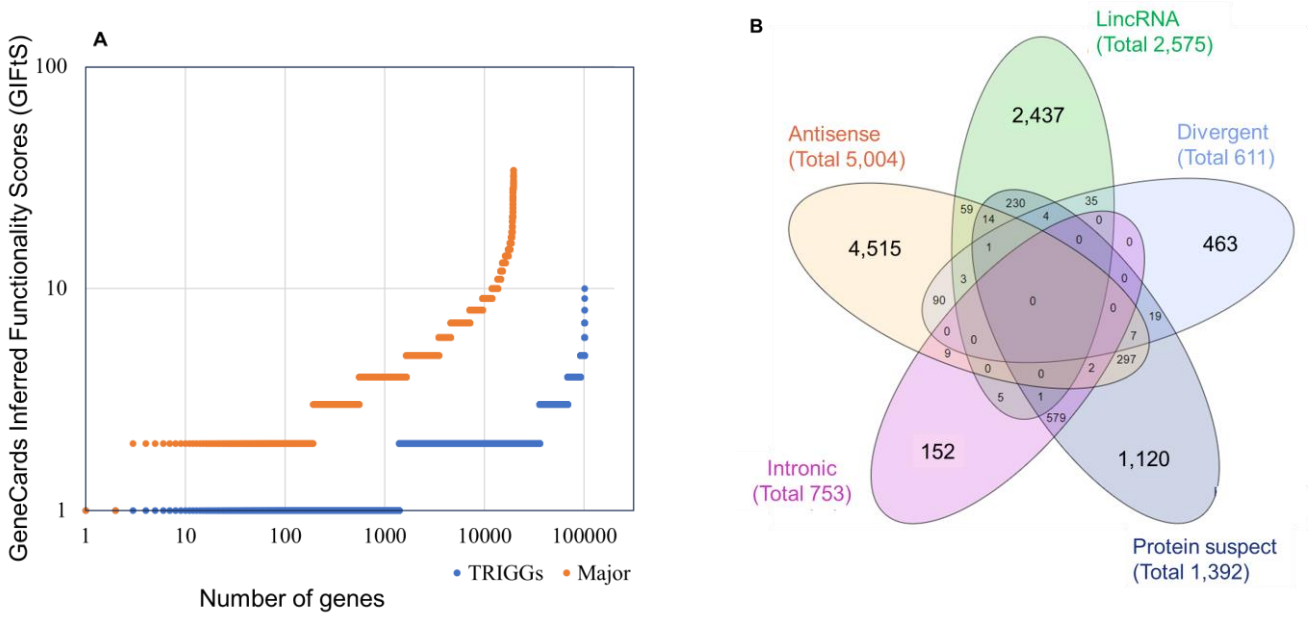

Figure 1. (A) Rank graph for annotative information, separately illustrated for two groups: Genes from the major gene sources (NCBI, HGNC and ENSEMBL) (orange), and genes inferred from RNAcentral transcripts without any annotation from the major gene sources, TRIGGS (Blue). **(B)** A GeneCaRNA-based suggested subclassification of 10,042 LncRNA subclass-definable genes (8.3% of the total). LincRNA: LncRNAs that are transcribed from the DNA stretch between the two protein coding genes. Divergent: LncRNAs that are transcribed by a promoter shared with a protein coding gene. Protein suspect: LncRNAs that have the potential to encode a peptide/protein. Intronic: LncRNAs transcribed purely from the intron(s) of a coding gene. Antisense: LncRNAs that are transcribed in antisense to a protein coding DNA strand.

2.2. LncRNA Subclass definition

We developed a pipeline for defining sub-classes for LncRNA genes (see Methods). (Frankish et al., 2021; Seal et al., 2023). We performed sub-family term unification (see Methods). As a result, five subclasses were created: "Antisense", "LincRNA", "Divergent", "Protein Suspect", and "Intronic" **(Supplemental Table S2)**. These sub-classes cover only 8.3% of the relevant genes; the rest were classified as "no subclass" (**Fig. 1B**).

2.3. Gene-disease and Gene-Pathway associations

We systematically explored GeneCaRNA for relationships of LncRNA genes to diseases, as portrayed in MalaCards, the disease database of the GeneCards suite (Rappaport et al., 2017). A total of 5,100 gene-disease associations were found, spanning 1,554 LncRNA genes and 2,021 diseases. We identified 715 elite associations, involving 618 contributing diseases, and 340 contributing genes (**Supplemental Table S3)**. Among the 1,554 genes, 793 were associated with only 1 disease, and the rest had 2-111 associated diseases (**Fig. 2A**). The top five disease-contributing genes are H19 (111 diseases), MALAT1 (87 diseases), MEG3 (76 diseases), HOTAIR (69 diseases), and CDKN2B-AS1 (59 diseases).

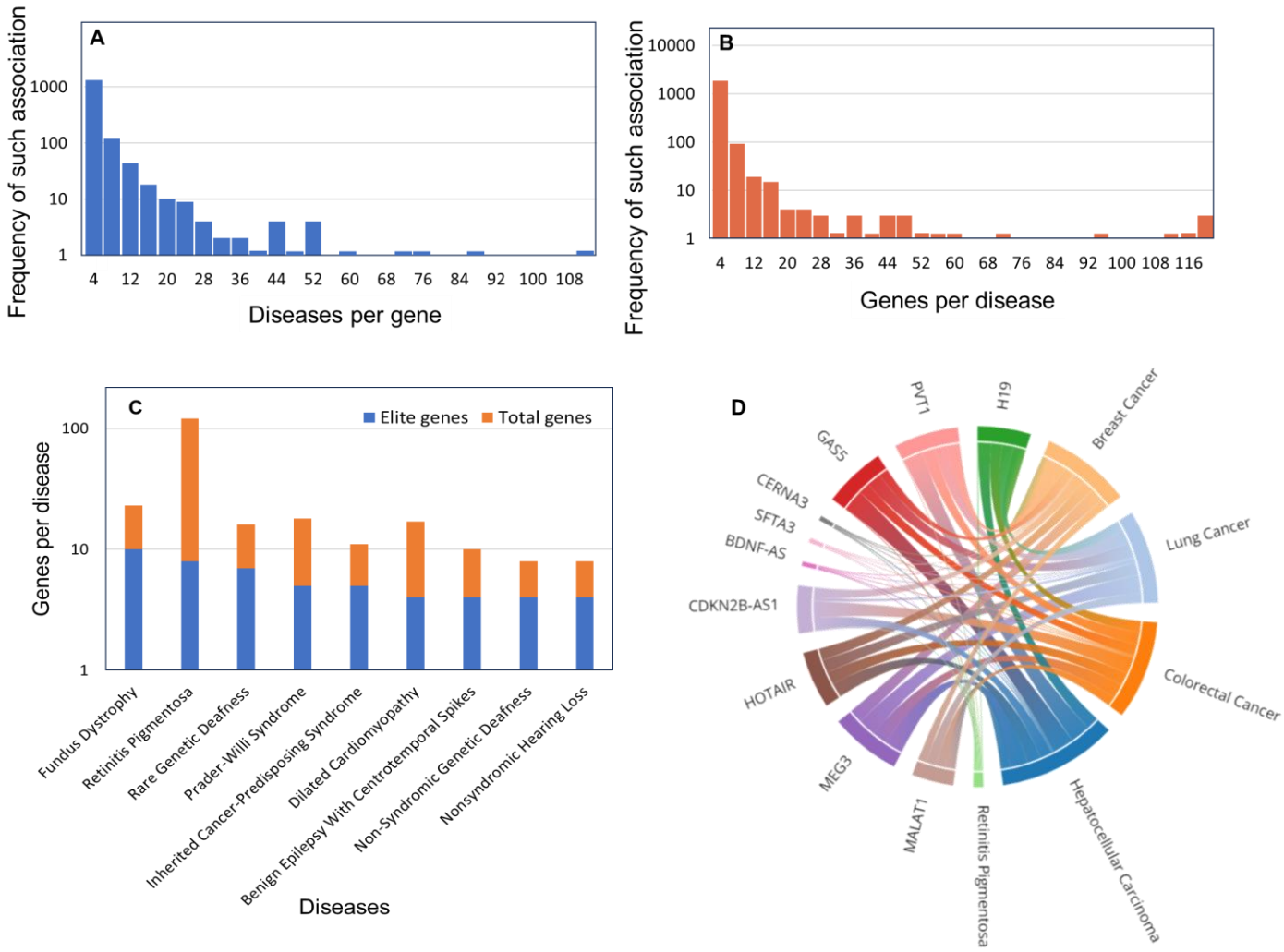

Figure 2. Gene-disease associations: Detailed scrutiny of the gene-disease associations among the 1,554 LncRNA genes and 2,021 diseases. **(A)** The distribution of the count of diseases per gene values; **(B)** The distribution of the count of gene per disease values; **(C)** The number of Genes per disease for a sample of 9 diseases with 4 or more elite associations. **(D)** Chord diagram of the network of the top 10 most-recurring genes, and the top 5 most-recurring diseases. Each chord represents the gene/disease association, where the width of the line corelates with the association score. The association score follows the MalaCards gene to disease scoring system, where score values depend on the level of manual curation of the information source, and on the significance assigned by the source itself to its different annotation classes (Rappaport et al., 2017).

Similarly, out of 2,021 diseases, 1,395 diseases are associated with only 1 gene (**Fig. 2B**). The top five diseases with the most genes are: Gastric cancer (120 genes), Retinitis pigmentosa (118 genes), Hepatocellular carcinoma (117 genes), Colorectal cancer (114 genes) and Lung cancer (111 genes). Only 9 diseases were found to have 4 or more elite disease-gene associations; featured in **Fig. 2C**. Around 50% of the genes are associated with diseases in many-to-many relationships, and the connectivity of selected cases of gene to disease associations are shown in **Fig. 2D**.

Involvement of multiple genes in a disease may indicate a gene-gene network, i.e. a biological pathway. Hence, all LncRNA genes associated with one or more disease were analyzed with respect to their pathway affiliation, as portrayed in the Pathways section of GeneCaRNA, which is based on the PathCards database in the GeneCards suite (Belinky et al., 2015). 131 pathways were found to be associated with 38 LncRNA genes (**Fig. 3A** and **B and Supplemental Table S4**). In an example, the role of four LncRNAs genes in a pathway with 28 genes is shown (**Fig. 3C**).

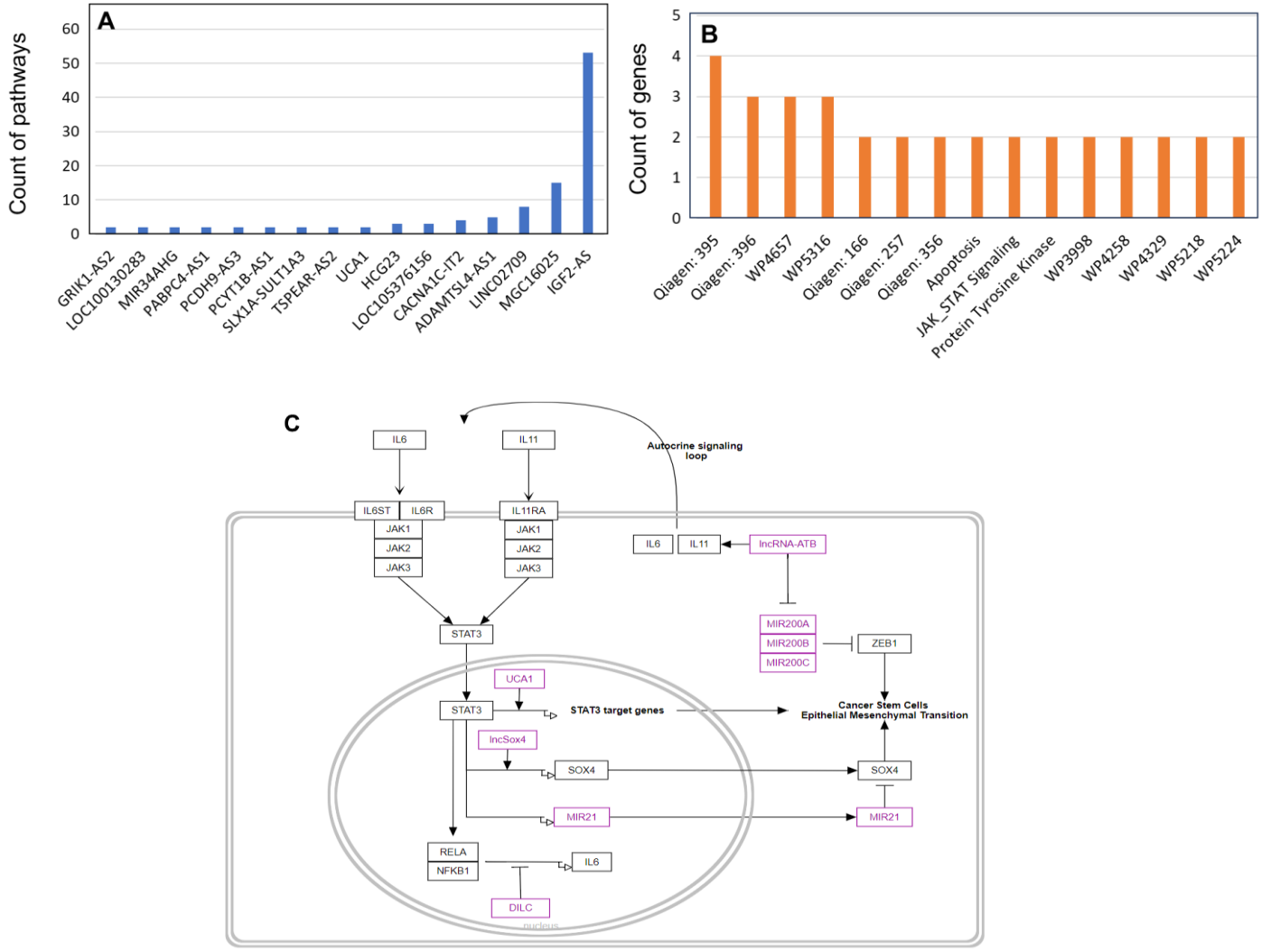

Figure 3. LncRNA-related pathways. (A) The distribution of counts of pathways (≥ 2) mapped to LncRNA genes; **(B)** The distribution of count of genes per pathway. 96 pathways include 1 gene (not shown) and at maximum a pathway had 4 genes; **(C)** A map of the "STAT3 signaling in hepatocellular carcinoma" obtained from WikiPathways (ID: WP4337). The LncRNA genes involved represent the cruciality of LncRNA genes in regulating protein coding genes, joining five MirRNA genes in the ncRNA category.

2.4. *Gene product Interactions*

Our multiple gene text analysis tool (MuSe) was applied to obtain a partial image of interactions of LncRNA transcripts with transcripts and encoded proteins of other genes. This was done with 3,000 adequately annotated LncRNA genes using 5 interaction terms. This facilitated the identification of 199 interacting partners mapped to 118 genes**.** These partners included 28 ncRNA genes and 171 protein/protein-coding genes **(Fig. 4)**. A sample of these interactions is shown in more detail in **Table 1,** showing the broad range of ncRNA interactions.

Table 1. List of LncRNA genes and their associated target gene products.

S.no	Input lncRNA genes		Action (Keyword)	Target genes			
	Symbol	Description		Symbol	Description	Catagory	Class
1	BCAR4	Breast Cancer Anti-Estrogen Resistance 4	Activates	GLI2	GLI Family Zinc Finger 2	Protein Coding	transcription regulator
2	HOTAIR	HOX Transcript Antisense RNA	Activates	DNMT1	DNA Methyltransferase 1	Protein Coding	methyltransferase
3	HOXA11-AS	HOXA11 Antisense RNA	Activates	ITGB3	Integrin Subunit Beta 3	Protein Coding	receptor
4	UCA1	Urothelial Cancer Associated 1	Activates	MTOR	Mechanistic Target Of Rapamycin Kinase	Protein Coding	kinase

5	SNHG1	Small Nucleolar RNA Host Gene 1	Activates	HOXA1	Homeobox A1	Protein Coding	transcription factor
6	LINC00261	Long Intergenic Non-Protein Coding RNA 261	Binds	MIR8485	MicroRNA 8485	RNA Gene	miRNA
7	ZFAS1	ZNFX1 Antisense RNA 1	Binds	MIR548E	MicroRNA 548e	RNA Gene	miRNA
8	H19	H19 Imprinted Maternally Expressed Transcript	Binds	EZH2	Enhancer Of Zeste 2 Polycomb Repressive Complex 2 Subunit	Protein Coding	enzyme
9	HOTAIR	HOX Transcript Antisense RNA	Binds	AR	Androgen Receptor	Protein Coding	receptor
10	LINC01413	Long Intergenic Non-Protein Coding RNA 1413	Binds	HNRNPK	Heterogeneous Nuclear Ribonucleoprotein K	Protein Coding	Nuclear Ribonucleoprotein
11	NORAD	Non-Coding RNA Activated By DNA Damage	Binds	WEE1	WEE1 G2 Checkpoint Kinase	Protein Coding	kinase
12	MIR222HG	MiR222/221 Cluster Host Gene	Interacts	DNM3OS	DNM3 Opposite Strand/Antisense RNA	RNA Gene	lncRNA
13	SNHG16	Small Nucleolar RNA Host Gene 16	Interacts	MIR195	MicroRNA 195	RNA Gene	miRNA
14	FOXD2-AS1	FOXD2 Adjacent Opposite Strand RNA 1	Interacts	EZH2	Enhancer Of Zeste 2 Polycomb Repressive Complex 2 Subunit	Protein Coding	enzyme
15	CASC11	Cancer Susceptibility 11	Interacts	HNRNPK	Heterogeneous Nuclear Ribonucleoprotein K	Protein Coding	Nuclear Ribonucleoprotein
16	DSCAM-AS1	DSCAM Antisense RNA 1	Interacts	YBX1	Y-Box Binding Protein 1	Protein Coding	RNA-binding protein
17	HOTAIR	HOX Transcript Antisense RNA	Interacts	AR	Androgen Receptor	Protein Coding	receptor
18	CRNDE	Colorectal Neoplasia Differentially Expressed	Interacts	TLR3	Toll Like Receptor 3	Protein Coding	receptor
19	PCAT1	Prostate Cancer Associated Transcript 1	Interacts	AR	Androgen Receptor	Protein Coding	receptor
20	RMST	Rhabdomyosarcoma 2 Associated Transcript	Interacts	SOX2	SRY-Box Transcription Factor 2	Protein Coding	transcription factor
21	H19	H19 Imprinted Maternally Expressed Transcript	Stimulates	STAR	Steroidogenic Acute Regulatory Protein	Protein Coding	regulator of steroid hormone synthesis
22	HULC	Hepatocellular Carcinoma Up-Regulated Long Non-Coding RNA	Stimulates	EMT	EMT ITK (IL2 Inducible T Cell Kinase)	Protein Coding	kinase
23	SNHG8	Small Nucleolar RNA Host Gene 8	Stimulates	NPC	NPC (Intracellular Cholesterol Transporter)	Protein Coding	tansporter
24	MALAT1	Metastasis Associated Lung Adenocarcinoma Transcript 1	Suppresses	MIR211	MicroRNA 211	RNA Gene	miRNA
25	SNHG1	Small Nucleolar RNA Host Gene 1	Suppresses	SOX9	SRY-Box Transcription Factor 9	Protein Coding	transcription factor

Another method for finding functional relations is offered by GeneHancer, yet another facility of the GeneCards Suite (Fishilevich et al., 2017). This resource portrays regulatory elements (enhancers and promoters) for every gene in GeneCards, including all ncRNA genes. GeneHancer also provides a list of genes sharing enhancer(s) with the gene in question. It also exhibits potential phenotypes and diseases based on genome-wide association studies (GWAS) as shown in **Supplemental Fig. S1**. We extracted enhancer information for each LncRNA gene and found ~108,000 genes having at least one enhancer mapped to it due to gene/enhancer proximity (**Supplemental Fig. S2**). 19,105 of the genes were contributed via major gene sources, with ~11% having one-to-one relationship, and 88,974 were TRIGGs, with ~17% of them in one-to-one relationship. Importantly, the very large group of novel TRIGG gene definitions often have no annotations except those inferred via GeneHancer, opening a way for a massive body of functional leading information.

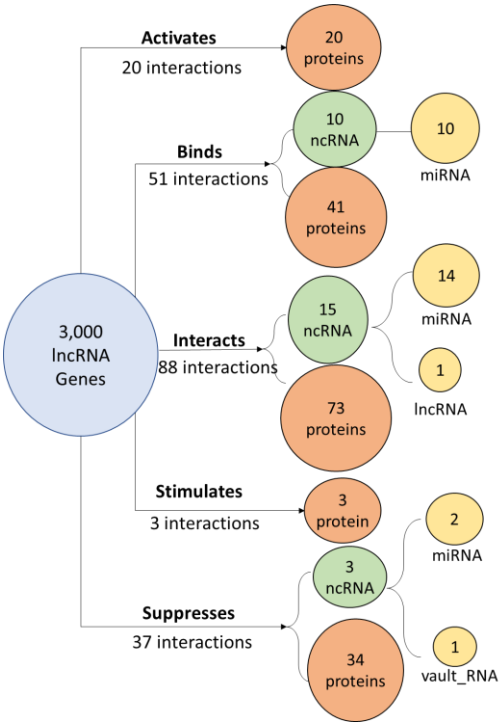

Figure 4. A diagram showing a sample of 199 target gene products showing one-to-one interactions with 118 LncRNA transcripts. The LncRNAs targets include 171 proteins, and a range of ncRNAs spanning three classes.

3. Discussion

GeneCaRNA was established 3 years ago to provide a compendium of all definable genes stemming from a full list of human transcripts. Its annotation system benefits from GeneCaRNA being part of the GeneCards Suite, which incorporates data-mined from 194 sources (Safran et al., 2021). Our database succeeds in portraying as many as ~280,000 ncRNA genes, ~7-times more than several other resources. This significant improvement was attained by the TRIGG algorithm (Safran et al., 2021). A great majority of the transcripts leading to the novel ~240,000 genes are weakly annotated, but the richness of their hosting GeneCards Suite often provides new information for such genes. In this paper, we use this power to scrutinize the largest class of ncRNA genes, namely LncRNAs, a functionally diverse group of genes, which currently embodies an avalanche of new research. GeneCaRNA, encompassing ~120,000 LncRNA genes, 43% of all ncRNA genes, provides substantial enrichment compared to other sources. Our GeneCaRNA-based enhancing pipeline for data extraction and analyses furnishes a fruitful approach to reporting and augment the relevant information published.

The immense size of the LncRNA class calls for sub-classification. The HUGO gene nomenclature committee (HGNC) defines six subclasses based on 6,000 LncRNAs (Seal et al., 2020). On the other hand, GENCODE define nine sub-classes, based on 18,000 genes (Frankish et al., 2021). When analyzing all 120,000 GenCaRNA's genes, we ascribed five subclasses to ~10,000 genes, of which ~600 were TRIGGS. We are now in the process of integrating GENCODE's sub-class assignments for 8,000 genes not yet covered by GeneCaRNA. All our five sub-classes appear in other subclassification systems, whereby our sub-class "Protein suspect", a seeming contradiction to the definition of LncRNA (Mattick & Rinn, 2015), is a merger of "3' overlapping" and "Sense overlapping" in GENCODE. We elected not to include HGNC's "contain microRNA or snoRNA" and GENCODE's "macro lncRNA", "non_coding" and "processed_transcript" LncRNA.

The study of LncRNAs is at the forefront of RNA biology due to the rise in reports claiming the role of LncRNAs in human disorders (Wapinski & Chang, 2011). A key strength of GeneCaRNA is its inclusion in a framework that has profound information about diseases via the inter-database relations with MalaCards (Rappaport et al., 2017). In parallel, there are mechanisms to discover indirect relationships between LncRNA genes and diseases via GeneHancer (**Supplemental Fig. S2**). Regarding the distinction between elite and inferred gene-disease associations, while the former are based on solid experimental support, the latter often provide a hint that opens a gateway for more research that may improve our understanding of a disease.

A key capacity of GeneCaRNA is text analyses on gene annotations, revealing information on interactions between ncRNA transcripts and other genes products (protein and transcripts). Using this platform, we were able to portray a

partial view of such interaction network (**Supplemental Table S5**). Further database development will allow us to obtain a holistic picture. Some examples of currently missing data include published functional roles of LncRNAs in the context of inflammatory diseases and cancers (Lin et al., 2021; Salmena et al., 2011). This happens by binding the LncRNAs to miRNAs, thus blocking their regulation of mRNAs in translation of downstream genes. Broadening the LncRNA interactome in GeneCaRNA will strengthen its capacity to support future research in the field.

The GeneCards suite includes VarElect (Safran et al., 2021; Stelzer et al., 2016), which can connect DNA variants to diseases. Initially, it was developed for exome sequencing, linking variants in protein coding genes to diseases. The establishment of GeneCaRNA now enables VarElect to function in the realm of disease decipherment of variants in ncRNAs genes obtained from whole genome sequencing. The indirect mode of VarElect helps discover association of variants in ncRNA genes with a disease known to be caused by mutations in a protein coding gene. This is achieved based on GeneCards information pointing to functional relationships between this ncRNA and the protein gene. This is important for solving undeciphered disease cases. This is an essential instrument for expanding the knowledge on the mechanisms by which LncRNA gene mutations lead to disease. In an example, somatic copy number mutations in lncRNA genes are shown to be associated with cancer clinical prognosis (Du et al., 2013), where, at present, little is known about the molecular pathways involved.

GeneCaRNA provides a comprehensive unified repository of LncRNA and their annotative information. Exploited the forte of the GeneCards Suite, it is slated to become a powerful go-to place for researchers to explore these genes. The proposed text analysis pipeline for revealing gene-gene interactions to be integrated into the GeneCards platform, will likely improve the understanding of the role of LncRNAs in biological processes. As knowledge of LncRNAs continues to evolve, the frequently updating GeneCaRNA will serve as an essential platform, enabling researchers to navigate the expansive landscape of non-coding RNA, including LncRNAs, their targets, and their associated regulatory routes.

4. Methods

4.1. Data extraction and text mining

Data extraction from the GeneCaRNA database (GC V5.19, based on RNAcentral V23) employed Structured Query Language (SQL) or our in-house tool GeneAlaCart (Safran et al., 2021). For routine analysis, we generated an interim LncRNA table containing the gene symbol, gene description, mined data sources, GeneCards Inferred Functionality Scores (GIFtS) (Harel et al., 2009), and chromosomal location of each gene (**Supplemental Table S1**).

The extracted table contained 120,982 LncRNAs genes, whose categorization was based on transcript categorization in our sources. Some genes are assigned symbols based on the information in our major sources, and if absent there, are assigned our own GeneCards symbols as done before (Safran et al., 2021). Most genes have additional (non-symbol) alias names featured in various external sources, which all are included in GeneCaRNA. GeneCaRNA was also used to extract all of the aliases and protein existence (PE) (Lane et al., 2014), using GeneAlaCart.

4.2. LncRNA Subclass definition

We extracted keywords from the Aliases section of GeneCaRNA, which contains multi-word gene descriptions. In the first stage, a list of the 10 most frequent words appearing in the entire collection of LncRNA genes was identified, using in-house python code (program code shown in S1 in **Supplementary file**). These terms were "Intronic", "Intergenic", "LINC", "Overlapping", "Divergent", "Pseudogene", "Antisense", "Sense", "Open Reading Frame", and "ORF". In a second stage, 5 candidate sub-classification terms were chosen, consulting other sub-classification approaches (Frankish et al., 2021; Harrison et al., 2024; Seal et al., 2023). We performed sub-family term unification, e.g. "Overlapping", "Pseudogene", "Sense", and "ORF" were clustered in our proposed system as "Protein Suspect". Under the same sub-class, we also joined genes with the highest degree (1,2,3) of UniProt's PE Levels (Lane et al., 2014), and the genes with lower levels (4,5) were also included provided that "protein" was a keyword appearing in our keyword search in GeneCaRNA's Aliases section. Finally, each gene was assigned one or more sub-classes, based on the appearance of these keywords in the original gene/alias-specific text in the Aliases section of GeneCaRNA.

4.3. Gene-disease and Gene-Pathway associations

A list of diseases associated with each gene was obtained by an SQL query. Each gene-disease pairing was assigned an association score (Safran et al., 2021). The score value depends on the level of manual curation of the information source, and on the significance assigned by the source itself to its different annotation classes. In addition, an "elite" grade is assigned for associations with experimental evidence support, and an "inferred" grade for the rest, where evidence is derived only from text-mining. GeneCards policy is to portray inferred association to catalyze future experimentation.

For gene-pathways relationships, we included only LncRNA genes with at least one associated disease. These were extracted using GeneAlaCart applied to the Pathways section of GeneCaRNA. The latter is mined from WikiPathways, SIGNOR, Reactome, PharmGKB, and Sino Biological pathways. In GeneCaRNA, we also show amalgamations to our SuperPathways facility (Belinky et al., 2015).

4.4. Gene product Interactions

The interaction targets for each LncRNA transcript (proteins, ncRNAs or DNA segments) were identified by the Multigene Search (MuSe) in-house text mining program. The query gene can be a symbol or any of the aliases. This allows one to receive target genes for a long list of submitted genes, (e.g.) by submitting "<gene name> binds", or "<gene name> interacts with". The output is texts from e.g. the GeneCaRNA Publications or Summaries sections, that include the text query. The output file of the MuSe is further curated using in-house python code **(**see S2 in **Supplemental Methods)**.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Supplemental_Fig_S1.pptx; Supplemental_Fig_S2.pptx; Supplemental_Table_S1.xlsx: List of all LncRNA genes in GeneCaRNA; Supplemental_Table_S2.xlsx: List of genes mapped to five sub-classes; Supplemental_Table_S3.xlsx: List of elite gene-disease associations; Supplemental_Table_S4.xlsx: List of gene-to-pathway associations; Supplemental_Table_S5.xlsx: List of gene-to-gene interactions for five selected keywords; Supplemental_Methods.pdf.

Author Contributions: Conceptualization, D.L. and S.A.; methodology, D.L., S.A. and M.G.; software, S.A, C.R., M.G., O.Z., and Y.G.; validation, T.S. and S.A.; formal analysis, S.A.; investigation, S.A.; resources, C.R., O.Z., Y.G.; data curation, S.A.; writing—original draft preparation, S.A.; writing—review and editing, M.S., S.P., M.G., and D.L.; visualization, S.A., and S.Z.; supervision, D.L.; funding acquisition, D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants from LifeMap Sciences (California, USA and Hong Kong), PIONEER, a European Network of Excellence for Big Data in Prostate Cancer, Horizon 2020 (777492); DC-ren, Horizon 2020 EU grant on Drug combinations for rewriting trajectories of renal pathologies in type II diabetes (SEP-210574920).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Use of GeneCards Suite websites is free for academic non-profit institutions. The suite's extensive knowledgebase is available for research purposes via an academic collaboration agreement (see https://www.genecards.org/Guide/Datasets), Other users must acquire a commercial license from LifeMap Sciences.

Acknowledgments: None

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

- LncRNA: Long non-coding RNA
- ncRNA: non-coding RNA
- HGNC: HUGO Gene Nomenclature Committee
- NCBI: National Center for Biotechnology Information
- TRIGGs: Transcripts Inferred GeneCaRNA genes.
- SQL: Structured Query Language
- GIFtS: GeneCards Inferred Functionality Scores
- LincRNA: Long intergenic non-coding RNA
- MuSe: Multigene Search
- ORF: Open Reading Frame
- PE: Protein Evidence

References

1. Aliperti, V., Skonieczna, J., & Cerase, A. (2021). Long Non-Coding RNA (lncRNA) Roles in Cell Biology, Neurodevelopment and Neurological Disorders. *Non-Coding RNA*, *7*(2), 36. https://doi.org/10.3390/ncrna7020036
2. Barshir, R., Fishilevich, S., Iny-Stein, T., Zelig, O., Mazor, Y., Guan-Golan, Y., Safran, M., & Lancet, D. (2021). GeneCaRNA: A Comprehensive Gene-centric Database of Human Non-coding RNAs in the GeneCards Suite. *Journal of Molecular Biology*, *433*(11), 166913. https://doi.org/10.1016/j.jmb.2021.166913

3. Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., & Lancet, D. (2015). PathCards: Multi-source consolidation of human biological pathways. *Database: The Journal of Biological Databases and Curation*, *2015*, bav006. https://doi.org/10.1093/database/bav006
4. Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K. D., Maglott, D. R., & Murphy, T. D. (2015). Gene: A gene-centered information resource at NCBI. *Nucleic Acids Research*, *43*(Database issue), D36-42. https://doi.org/10.1093/nar/gku1055
5. Du, Z., Fei, T., Verhaak, R. G. W., Su, Z., Zhang, Y., Brown, M., Chen, Y., & Liu, X. S. (2013). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature Structural & Molecular Biology*, *20*(7), Article 7. https://doi.org/10.1038/nsmb.2591
6. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., Lancet, D., & Cohen, D. (2017). GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards. *Database: The Journal of Biological Databases and Curation*, *2017*, bax028. https://doi.org/10.1093/database/bax028
7. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., … Flicek, P. (2021). GENCODE 2021. *Nucleic Acids Research*, *49*(D1), D916–D923. https://doi.org/10.1093/nar/gkaa1087
8. Harel, A., Inger, A., Stelzer, G., Strichman-Almashanu, L., Dalah, I., Safran, M., & Lancet, D. (2009). GIFtS: Annotation landscape analysis with GeneCards. *BMC Bioinformatics*, *10*(1), 348. https://doi.org/10.1186/1471-2105-10-348
9. Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., Chougule, K., … Yates, A. D. (2024). Ensembl 2024. *Nucleic Acids Research*, *52*(D1), D891–D899. https://doi.org/10.1093/nar/gkad1049
10. Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., … Gingeras, T. R. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science (New York, N.Y.)*, *316*(5830), 1484–1488. https://doi.org/10.1126/science.1138341
11. Lane, L., Bairoch, A., Beavis, R. C., Deutsch, E. W., Gaudet, P., Lundberg, E., & Omenn, G. S. (2014). Metrics for the Human Proteome Project—2013-2014 and Strategies for Finding Missing Proteins. *Journal of Proteome Research*, *13*(1), 15–20. https://doi.org/10.1021/pr401144x
12. Lin, Y., Jin, L., Tong, W. M., Leung, Y. Y., Gu, M., & Yang, Y. (2021). Identification and integrated analysis of differentially expressed long non-coding RNAs associated with periodontitis in humans. *Journal of Periodontal Research*, *56*(4), 679–689. https://doi.org/10.1111/jre.12864
13. Ma, L., Bajic, V. B., & Zhang, Z. (2013). On the classification of long non-coding RNAs. *RNA Biology*, *10*(6), 924–933. https://doi.org/10.4161/rna.24604
14. Mattick, J. S., Amaral, P. P., Carninci, P., Carpenter, S., Chang, H. Y., Chen, L.-L., Chen, R., Dean, C., Dinger, M. E., Fitzgerald, K. A., Gingeras, T. R., Guttman, M., Hirose, T., Huarte, M., Johnson, R., Kanduri, C., Kapranov, P., Lawrence, J. B., Lee, J. T., … Wu, M. (2023). Long non-coding RNAs: Definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology*, *24*(6), Article 6. https://doi.org/10.1038/s41580-022-00566-8
15. Mattick, J. S., & Rinn, J. L. (2015). Discovery and annotation of long noncoding RNAs. *Nature Structural & Molecular Biology*, *22*(1), 5–7. https://doi.org/10.1038/nsmb.2942
16. Rappaport, N., Twik, M., Plaschkes, I., Nudel, R., Iny Stein, T., Levitt, J., Gershoni, M., Morrey, C. P., Safran, M., & Lancet, D. (2017). MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Research*, *45*(Database issue), D877–D887. https://doi.org/10.1093/nar/gkw1012
17. The RNAcentral Consortium, Sweeney, B. A., Petrov, A. I., Burkov, B., Finn, R. D., Bateman, A., Szymanski, M., Karlowski, W. M., Gorodkin, J., Seemann, S. E., Cannone, J. J., Gutell, R. R., Fey, P., Basu, S., Kay, S., Cochrane, G., Billis, K., Emmert, D.,

Marygold, S. J., … Williams, K. P. (2019). RNAcentral: A hub of information for non-coding RNA sequences. *Nucleic Acids Research*, *47*(D1), D221–D229. https://doi.org/10.1093/nar/gky1034

18. Safran, M., Rosen, N., Twik, M., BarShir, R., Stein, T. I., Dahary, D., Fishilevich, S., & Lancet, D. (2021). The GeneCards Suite. In I. Abugessaisa & T. Kasukawa (Eds.), *Practical Guide to Life Science Databases* (pp. 27–56). Springer Nature. https://doi.org/10.1007/978-981-16-5812-9_2
19. Salmena, L., Poliseno, L., Tay, Y., Kats, L., & Pandolfi, P. P. (2011). A ceRNA hypothesis: The Rosetta stone of a hidden RNA language? *Cell*, *146*(3), 353–358. https://doi.org/10.1016/j.cell.2011.07.014
20. Seal, R. L., Chen, L.-L., Griffiths-Jones, S., Lowe, T. M., Mathews, M. B., O'Reilly, D., Pierce, A. J., Stadler, P. F., Ulitsky, I., Wolin, S. L., & Bruford, E. A. (2020). A guide to naming human non-coding RNA genes. *The EMBO Journal*, *39*(6), e103777. https://doi.org/10.15252/embj.2019103777
21. Seal, R. L., Tweedie, S., & Bruford, E. A. (2023). A standardised nomenclature for long non-coding RNAs. *IUBMB Life*, *75*(5), 380–389. https://doi.org/10.1002/iub.2663
22. Stelzer, G., Plaschkes, I., Oz-Levi, D., Alkelai, A., Olender, T., Zimmerman, S., Twik, M., Belinky, F., Fishilevich, S., Nudel, R., Guan-Golan, Y., Warshawsky, D., Dahary, D., Kohn, A., Mazor, Y., Kaplan, S., Iny Stein, T., Baris, H. N., Rappaport, N., … Lancet, D. (2016). VarElect: The phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics*, *17*(Suppl 2), 444. https://doi.org/10.1186/s12864-016-2722-2
23. Sun, Q., Hao, Q., & Prasanth, K. V. (2018). Nuclear long noncoding RNAs: Key regulators of gene expression. *Trends in Genetics : TIG*, *34*(2), 142–157. https://doi.org/10.1016/j.tig.2017.11.005
24. Wang, J., Zhu, S., Meng, N., He, Y., Lu, R., & Yan, G.-R. (2019). ncRNA-Encoded Peptides or Proteins and Cancer. *Molecular Therapy*, *27*(10), 1718. https://doi.org/10.1016/j.ymthe.2019.09.001
25. Wapinski, O., & Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends in Cell Biology*, *21*(6), 354–361. https://doi.org/10.1016/j.tcb.2011.04.001

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Table 1. List of LncRNA genes and their associated target gene products.

Figure 1. (A) Rank graph for annotative information, separately illustrated for two groups: Genes from the major gene sources (NCBI, HGNC and ENSEMBL) (orange), and genes inferred from RNAcentral transcripts without any annotation from the major gene sources, TRIGGS (Blue). **(B)** A GeneCaRNA-based suggested subclassification of 10,042 LncRNA subclass-definable genes (8.3% of the total). LincRNA: LncRNAs that are transcribed from the DNA stretch between the two protein coding genes. Divergent: LncRNAs that are transcribed by a promoter shared with a protein coding gene. Protein suspect: LncRNAs that have the potential to encode a peptide/protein. Intronic: LncRNAs transcribed purely from the intron(s) of a coding gene. Antisense: LncRNAs that are transcribed in antisense to a protein coding DNA strand.

Figure 2. Gene-disease associations: Detailed **s**crutiny of the gene-disease associations among the 1,554 LncRNA genes and 2,021 diseases. **(A)** The distribution of the count of diseases per gene values; (**B)** The distribution of the count of gene per disease values; **(C)** The number of Genes per disease for a sample of 9 diseases with 4 or more elite associations. **(D)** Chord diagram of the network of the top 10 most-recurring genes, and the top 5 most-recurring diseases. Each chord represents the gene/disease association, where the width of the line corelates with the association score. The association score follows the MalaCards gene to disease scoring system, where score values depend on the level of manual curation of the information source, and on the significance assigned by the source itself to its different annotation classes (Rappaport et al., 2017).

Figure 3. LncRNA-related pathways. (A) The distribution of counts of pathways (≥ 2) mapped to LncRNA genes; (**B)** The distribution of count of genes per pathway. 96 pathways include 1 gene (not shown) and at maximum a pathway had 4 genes; **(C)** A map of the "STAT3 signaling in hepatocellular carcinoma" obtained from WikiPathways (ID: WP4337). The LncRNA genes involved represent the cruciality of LncRNA genes in regulating protein coding genes, joining five MirRNA genes in the ncRNA category.

Figure 4. A diagram showing a sample of 199 target gene products showing one-to-one interactions with 118 LncRNA transcripts. The LncRNAs targets include 171 proteins, and a range of ncRNAs spanning three classes.