# Supplemental Information

# Non-Coding RNAs Improve the Predictive Power of Network Medicine

Deisy Morselli Gysi and Albert-László Barabási

## Table of Contents

# 1 NON-CODING RNAS: A SHORT REVIEW

The non-coding RNAs (ncRNAs) are responsible for different biological functions, such as the maintenance of gene expression. In brief, gene expression relies on DNA as a template for RNAs which then migrate to the cytoplasm, where they are translated into proteins. Yet, not all RNAs will be translated into proteins, and many ncRNAs are needed for diverse cell functions[1].

The **small nuclear RNAs (snRNAs)** are mainly involved with the mRNA splicing, Transfer RNAs (tRNAs) are the decoders from mRNAs into peptides or proteins, they recognize three nucleotide sequences in the mRNA (codons) and recruit amino acids in the same sequence to ribosomes; Ribosomal RNAs (rRNAs), the most abundant RNA molecules in the cell, are the building blocks for ribosomes, essential for protein translation (housekeeping RNAs). Some housekeeping RNAs can carry modifications, added by small nucleolar RNAs (snoRNAs)[2]. **Small double-stranded RNAs (dsRNAs)** are known to mediate post-transcriptional gene silencing of mRNAs, by a process known as RNA interference (RNAi)[3]. Additionally, some ncRNAs were also discovered to be involved in regulatory processes such as micro RNAs (miRNAs), piwi-associated RNAs (piRNAs), and small interfering RNAs (siRNAs).

**MicroRNAs (miRNAs)** are a family of small nonprotein coding RNAs, that contain ~ 22 nucleotides[4], and function as important regulators of gene expression (activating or repressing their translation[4,5], and mainly act at the post-transcriptional level[6]A mature miRNA can recognize their mRNA targets by base-pairing the seed region (2-8 of the miRNA nucleotides) to the complementary region on the targeted mRNA[7]. Each miRNA can have hundreds of

targets and has been estimated that the human genome has more than 1000 miRNAs and about 10 to 30% of all human genes may be regulated by miRNAs[8]. One can use Argonaute (AGO) immunoprecipitation followed by high-throughput sequencing, termed AGO CLIP–seq, to examine the association between miRNA and mRNAs in a very specific manner[9,10]. However, this technique does not provide a direct linkage between miRNA and its mRNAs targets, but it uses as an intermediate the argonaute protein (AGO), which binds to miRNAs. To overcome this issue, a new technology called CLASH (crosslinking, ligation, and sequencing of hybrids) has been proposed[11], which ligates the 5′ end of an AGO-bound miRNA to the 3′ end of an AGO-bound mRNA, and as consequence, it provides direct linkage information of the two RNAs[12,13]. Because miRNAs function in a similar manner as Transcription Factors (TF), they have an intertwined and well-regulated network[14] that can contribute to disease development[15] such as asthma[16] and schizophrenia[17], when mutated or dysregulated miRNAs are highly associated with lack of function, such as neurogenesis[18].

**PIWI-interacting RNAs (piRNAs),** were discovered from germlines, and are a little bit longer than miRNAs (24-32 nucleotides), piRNAs originate from single-strand RNAs, as opposed to miRNAs that originate from double-stranded RNAs and require post-processing. piRNAs act by directly cleaving mRNAs[19]. Until 2018 more than 8 million piRNAs have been discovered in humans[20]. Similarly, **small interfering RNA (siRNA),** like miRNAs, are part of the RNAi class. Those are double-stranded ncRNAs with ~ 20-25 nucleotides and they are involved with the cell defense against undesired transcripts[21] and external invasion[22] and probably being our first biological mechanism that acted as the immune system[22]. siRNAs are highly sequence-specific, and in theory, can silence any disease-related genes, by mediating targeted

mRNA. Given that, three drugs using siRNAs have been already approved to treat hereditary amyloidogenic transthyretin and acute hepatic porphyria diseases[22,23] and several others are in clinical trials for hemophilia, acute kidney injury, and others are being studied for IBD[24] and Ocular Diseases[25]; for a review see [26].

**Long non-coding RNAs (lncRNA)** are a family of nonprotein coding RNAs that exceed 200 nucleotides[27], present a poly-A tail, similar to mRNAs, and have potential to be spliced. Even though only a small number has been well-characterized[28], it is known that lncRNAs are involved in a wide range of biological functions, such as X-chromosome inactivation[29,30], imprinting[31,32], they can act as Gene Regulatory Factors (GRFs), and interact with one or more proteins[33–35], they also can bind to chromatin[36], enhancers[37] and act as sponges for miRNA[38]. Moreover, lncRNAs are associated with multiple diseases[39], including different types of cancers[40], autoimmune neuropathies[14], and neurodegenerative diseases[41]. Due to their function as GRF, lncRNAs can physically interact with proteins, and the measurement of the physical binding started by the low-throughput assays such as RNA electrophoretic mobility shift assay[42], RNA pull-down assay[43], oligonucleotide-targeted RNase H protection assay[44], and FISH co-localization[45]. However, these methods offer only limited information, mainly due to the many-to-many bindings that occur between proteins and RNAs. Nowadays, several different high-throughput methods exist, that can be classified into protein-focused and RNA-focused[28]. The protein-focused focus is on the binding of RNAs to a protein of interest and can be further classified into *in vitro* or *in vivo*. In the *in vitro* approaches, RNA libraries are tested against a protein, and high-affinity RNAs are isolated after rounds of selection. In the *in vivo* methods, RNAs bound to the protein of interest in a sample are pulled down using variants of

immunoprecipitation techniques. In RNA-focused approaches, the goal is to identify all proteins bound to an RNA of interest. For a review with a complete description and comparison of the methods, see [28]. Of note, the lncRNA CRNDE has been identified as a promising target for the therapeutic treatment of prostate cancer by targeting miR-146a-5p[46].

# 2 DATABASE CONSTRUCTION

## 2.1 NON-CODING RNA DATABASES

We combine data from nine publicly available datasets that compile experimentally validated non-coding interactions. We describe each dataset, its data, and its normalization procedure. Table S 1 details their gene composition and interactions.

**DIANA Tools**[47] ranges from target prediction algorithms to databases of experimentally verified miRNA targets on coding and non-coding RNAs. In total it provides 91,249 interactions between 12,254 genes (10,087 proteins, 2,167 ncRNAs).

**lncBook**[48] carries information on lncRNAs, from functions, and associations to diseases and interactions from lncRNAs to other elements (proteins and miRNAs). The interactions are experimentally validated or predicted. In total, the database provides 21 interactions between 30 ncRNAs.

**lncRNome**[49] is a knowledge graph on human lncRNAs and provides predicted and experimentally validated interactions from lncRNAs and other RNAs. We only retrieved the database of experimentally validated targets, which offers information about 978 interactions between 521 genes (130 proteins, 391 ncRNAs).

**mirTARbase**[50,51] is a collection of experimentally validated miRNAs and their targets, validated by reporter assay, western blot, microarray, and next-generation sequencing experiments. After gene name conversion, it provides 6,667 interactions between 3,022 genes (2,511 proteins, 511 ncRNAs).

**miRecords**[52] is a manually curated database of experimentally validated miRNA-target interactions. It provides 975 validated interactions between 818 genes (651 proteins, 167 ncRNAs).

**miRNet**[53] aggregates information from miRTarBase v8.0, TarBase v8.0, and miRecords and allows for the selection of experimentally validated miRNA-targets. In total miRNet provides 2,546 interactions between 1,264 genes (917 proteins, 347 ncRNAs).

**NPinter4**[54] contains only experimentally validated interactions from ncRNA to DNA, TF, proteins, and other RNAs using CLIP-seq, AGO CLIP-seq, ChIRP-seq, and literature-mined interactions. For humans, it provides binding information for 1,153 interactions between 1,023 genes (538 proteins, 485 ncRNAs)

**RAIN**[55] contains experimentally validated and predicted interactions. Here we considered only experimentally validated interactions with a confidence score higher than 0.15 (as suggested by the authors), resulting in 391,209 interactions between 16,244 genes (13,271 proteins, 2,973 ncRNAs)

**RISE**[56] focuses on RNA-RNA interactions, which come from transcriptome-wide sequencing-based experiments such as PARIS, SPLASH, LIGRseq, and MARIO, and targeted studies like RIAseq, RAP-RNA, and CLASH. RISE also aggregates data from other databases

with experimental validation, 64,084 validated interactions between 20,660 genes (15,559 proteins, 5,101 ncRNAs).

Note that, during the construction of the PPI some databases reported protein and ncRNA binds, we included those interactions only in the NCI. The Table S 1 provides a complete overview of how many genes and interactions exist in each final database, along with the number of experimentally validated interactions present in each network.

## 2.2 GENE DISEASE DATABASES

To link genes to diseases, we rely on several databases. We start by normalizing gene and disease names. A description of the gene-disease association in each database is presented in Table S 4.

**ClinGen**[57] is a worldwide effort to associate genes with diseases along with data curation from a panel of experts. Although it provides different levels of evidence, for each entry we focused only on the ones with strong or weak evidence. Overall, it accounts for 523 associations from 446 genes and 138 diseases.

**ClinVar**[58] reports relationships among human variations and phenotypes, with supporting evidence. It accounts for 6,769 associations between 3,847 genes and 917 diseases.

The **Comparative Toxicogenomics Database (CTD**[59]**)** is a well-established database of chemicals, genes, phenotypes, exposures, and diseases. All its 24,857 entries are annotated with publication information, allowing for transparency and for tracing any of its 7,327 genes and 7,711 diseases.

**Disease Enhancer**[60] is a manually curated database, based on literature, for disease-associated enhancers. It provides 518 associations between 303 genes and 121 diseases.

**DisGeNet**[61,62] integrates expert-curated databases with text-mined data and covers information on Mendelian and complex diseases. Here, we focus only on the curated databases based on genes or variants. The gene-based data source contains 36,317 associations, from 8,912 genes to 2,099 diseases.

**GWAS catalogue**[63] combines information across multiple GWAS studies extracted from literature and goes through a double curation process, where only associations with sufficient evidence are kept. We removed inferred and imputed SNPs and retrieved 8,957 associations, between 4,254 genes and 434 diseases.

**HMDD**[64]**, Human microRNA Disease Database**, focuses only on the association of miRNAs and diseases. Its data is experimentally validated through miRNA circulation, tissue differential expression, genetics, epigenetics, or targeted analysis. HMDD accounts for 13,662 associations between 916 miRNAs and 622 diseases.

**LncBook**[48]**,** used in the construction of the NCI, also provides information on lncRNAs' associations with diseases. It contains 1,478 associations between 669 genes and 246 diseases.

**lncRNADisease**[65,66] is a knowledge base that focuses on associations between diseases and lncRNA, it provides different levels of evidence for its associations. In our study, we focus on 3,146 associations with experimental evidence between 1,533 genes and 286 diseases.

**Leiden Open Variation Database (LOVD)**[67] is an integrative project connecting expert-curated variants and genes with multiple phenotypes. LOVD includes 2,931 associations from 2,113 genes to 664 diseases.

**Monarch** is an integrative project that connects phenotypes and genotypes in several species. It does provide experimental evidence for part of its results and includes text-mining information. In total, provides 19,542 associations between 8,059 genes and 1,058 diseases.

**Online Mendelian Inheritance in Man (OMIM[68])** is a continuous catalog of human genes and traits, with a focus on the molecular relationship between genetic variation and phenotype. OMIM includes 3,517 associations based on peer-reviewed biomedical literature from 2,578 genes to 751 diseases.

**Orphanet** is a manually-curated database that focuses on experimentally validated associations from rare diseases to genes. It contains 3,787 associations from 2,555 genes to 721 rare diseases.

**PheGenI**[69] is the Phenotype-Genotype Integrator from NCBI and combines multiple of its hosted databases (such as Gene, dbGaP, OMIM, eQTL, and dbSNP). From this integrator, we retrieved 11,232 associations from 5,929 genes and 485 diseases.

**Psygenet**[70] is a database that focuses on psychiatric disorders. A gene is associated with a disease if it plays a role in the disease pathogenesis or is a marker for the disease. For that, the associations are identified using a text-mining tool, and later manually curated. It provides 2,924 associations between 1,324 genes in 40 mental disorders.

## 2.3 DEGREE DISTRIBUTION

We analyzed the degree distribution of the PPI and the PPI & NCI networks, following the methodology defined in [71]. We find that while the inclusion of the NCI does not affect the power law nature of the network, it does alter the degree exponent, which is crucial for determining the properties of the network. Specifically, we find that the degree exponent for the PPI & NCI network is $\gamma_{PPI\,\&\,NCI} = 2.54$, less than the degree exponent for the PPI network, $\gamma_{PPI} = 2.71$. We also calculate γ for each element type and present its results on Table S 6. We find that Protein Coding Genes behave differently than TFs in the PPI & NCI network. This difference is important because a $\gamma < 3$ indicates a scale-free property with a diverging second moment, which has significant implications for the network robustness.

### 2.4 DISEASE MODULE SIGNIFICANCE

Disease modules significances were calculated using both a degree preserving and a non-degree preserving randomization. Results referring to the degree preserving randomization indicate that 41 diseases are statistically significant only using the PPI, while the PPI & NCI alone finds 65 diseases with statistical significance, and 109 diseases are identified in both networks.

## 3 GENE CO-EXPRESSION NETWORKS

Gene co-expression networks are often used to shed light on the molecular mechanisms that underlie biological processes and how changes in those interactions can lead

to a disorder[72]. These networks are inferred using the association between pairwise gene expression, measured using RNAseq or microarray. The association is often derived from a correlation, such as Pearson correlation, or a transformation, such as WGCNA[73] or Weighted Topological Overlap (wTO)[74,75], which calculates the co-expression between two genes based on their correlation normalized by their commonalities, and removes false positives. The former focuses on positive correlations, while the latter allows for the inclusion of positive and negative correlations. Here, we construct a gene co-expression network from whole blood samples provided by GTEx[76], using Pearson Correlation ($\rho$) and the wTO[74] ($\omega$).

Using the physical networks we derived from the PPI and the PPI & NCI (iNCI), we compared the co-expression values for both the Pearson Correlation and the wTO. We observed higher co-expression values on the physical networks (PPI, iNCI, and PPI & iNCI), leading us to ask whether the co-expression weights are predictive of physical binding. We calculated the Area Under the ROC (AUROC), for both co-expression values in the three physical networks, and we find that the AUCs increased from 0.59 in the PPI to 0.63 in the iNCI and PPI & iNCI networks for $\rho$ (Figure 4 F) and from 0.58 in the PPI to 0.63 in the iNCI and PPI & iNCI networks for $\omega$ (Figure S 7 B).

# 4 THE EFFECT ON NON-CODING RNAS IN DRUGS

Traditionally, drug-targets are proteins, currently, there is no ncRNA in DrugBank[77] listed as a drug-target. Yet, some ncRNAs are used as drugs – such as Bevasiranib[78], a small interfering RNA (siRNA) that targets the VEGF-A gene, currently in clinical trial for treating

241 macular degeneration and diabetic retinopathy, or Inclisiran, an LDL cholesterol lowering
242 siRNA that targets PCSK9, the first approved ncRNA-based drug [26,79].

243 Here, we explore the potential effect of ncRNAs on the drug-target network, by
244 exploring the neighborhood of drug-targets. For that, we explore the differences in gene
245 classes (protein-coding, TF, miRNA, lncRNA, etc) each drug-target interacts with, compared to
246 non-drug-targets. We find that drug target proteins interact with more miRNAs, Protein-
247 coding, and TF than proteins not targeted by drugs (Mann Whitney Test, p-adj < 0.05; FDR
248 corrected, Table S 7), meaning that targets are, on median, more connected to miRNAs,
249 protein-coding and TFs than not drugged proteins.

250 We begin by evaluating the effect of ncRNAs on drug proximity. For that, we calculate
251 the average proximity of drug targets to the RA disease module in both networks. We first
252 focus on three drugs used to treat RA: Adalimumab, Abatacept, and Penicillamine.
253 Adalimumab targets the TNF gene which is also located in the RA disease module. In other
254 words, both the PPI or the PPI & NCI disease modules already embed this gene in its network,
255 meaning its average proximity is already zero. Abatacept, in turn, targets CD80 and CD86 in
256 the PPI. The average proximity of those genes to the RA disease module is 2.34 in the PPI and
257 2.11 in the PPI & NCI, indicating that by adding non-coding interactions, we observe a
258 reduction of the proximity values. Penicillamine targets only the gene SLCO1B1, having
259 average proximity of two in the PPI, while the distance decreases to one in the PPI & NCI
260 (Figure S 10 A). We next test the change in distance for all drugs indicated for RA in Drug Bank,
261 finding that the average proximity reduces its proximity values from PPI to PPI & NCI. In other

262 words, the inclusion of ncRNAs into the interactome reduces the distance between drug-
263 targets and disease modules.

264 Even though the correlation between the proximity in the PPI and the PPI & NCI
265 networks is high ($\rho = 0.97$), we still find a significant reduction of the average proximity in
266 drugs indicated for RA the PPI & NCI ($p < 0.01$; Mann-Whitney's test). We next compare the
267 average proximity between drugs indicated for RA compared to drugs without official
268 indication for RA, finding that the proximity is statistically smaller for RA-indicated drugs than
269 the ones not indicated to treat RA (Wilcoxon test; $p < 0.01$, (Figure S 10 B), both in the PPI and
270 the PPI & NCI. Meaning that drugs used to treat RA are, on median, closer to the disease
271 module than drugs not indicated for the RA treatment, in both PPI and PPI & NCI.
272

# 5 TABLES

Table S 1 **Descriptive of Databases Genes and Reported Binding Experimental Interactions.** Gene names were normalized to Gene Symbols and if no official name was found the gene was removed from the final database along with all its interactions. We kept only experimentally validated interactions and classified the interactions according to gene type (protein-coding - PC, non-coding – NC).

| DATABASE | GENES | | INTERACTIONS | | |
|---|---|---|---|---|---|
| | NC | PC | NC <-> NC | PC <-> NC | PC <-> PC |
| APID | 411 | 15930 | 32 | 3151 | 169730 |
| BIOGRID | 616 | 15308 | 31 | 9883 | 194221 |
| BIOPLEX | 386 | 13547 | 177 | 3912 | 104600 |
| COFRAC | 19 | 2959 | 3 | 205 | 13689 |
| DIANA | 2167 | 10087 | 15189 | 76060 | 0 |
| DIP | 2 | 1389 | 1 | 1 | 1285 |
| ENCODE | 1229 | 13808 | 1412 | 101569 | 23473 |
| HI-UNION | 165 | 8882 | 17 | 1703 | 61969 |
| HINT | 320 | 14576 | 66 | 2528 | 116789 |
| HIPPIE | 100 | 8520 | 17 | 1437 | 121691 |
| INATEDB | 153 | 5548 | 1 | 251 | 13378 |
| INSIDER | 17 | 3362 | 18 | 12 | 3939 |
| INSTRUCT | 206 | 11234 | 45 | 774 | 50456 |
| INTACT | 56 | 4567 | 10 | 131 | 9102 |
| INTERACTOME3D | 68 | 7327 | 108 | 135 | 15071 |
| INWEB | 67 | 10530 | 43 | 94 | 58507 |
| KINOMENETWORKX | 7 | 2340 | 0 | 25 | 7367 |
| LITBM17 | 20 | 6020 | 1 | 46 | 13320 |
| LNC_BOOK | 30 | 0 | 21 | 0 | 0 |
| LNCRNOME | 391 | 130 | 0 | 733 | 245 |
| MINT | 42 | 4871 | 6 | 121 | 10616 |
| MIRECORDS | 167 | 651 | 7 | 967 | 1 |
| MIRNET | 347 | 917 | 14 | 2532 | 0 |
| MIRTARBASE | 511 | 2511 | 51 | 6616 | 0 |
| NPINTERV4 | 485 | 538 | 210 | 860 | 83 |
| PHOSPHOSITEPLUS | 7 | 2840 | 0 | 7 | 7014 |
| PINA | 282 | 14895 | 12 | 958 | 163068 |
| QUBIC | 34 | 5271 | 7 | 241 | 27453 |
| RAIN | 2973 | 13271 | 7840 | 383213 | 156 |
| RISE | 5101 | 15559 | 2333 | 26660 | 35091 |

Table S 2 **Network proprieties comparison between PPI and PPI & NCI.** The PPI has smaller network diameter and average shortest path length. The PPI & NCI, on its turn has smaller network density and transitivity.

| CHARACTERISTIC | PPI | PPI & NCI |
|---|---|---|
| DIAMETER | 7 | 9 |
| DENSITY | 0.00324 | 0.00316 |
| COMPONENTS | 13 | 52 |
| TRANSITIVITY | 0.04785 | 0.02594 |
| LARGEST CLIQUES | 57 | 57 |
| LCC | 18204 | 26483 |
| AV. DEGREE | 58.13 | 83.32 |
| MEDIAN DEGREE | 32 | 34 |
| AV. SHORTEST PATH | 2.66131 | 2.79465 |

Table S 3 **Median and interquartile ranges for each gene category in all two networks.** TFs and protein-coding genes increase their degree median when we include the NCI in the PPI.

| GENE CATEGORY | PPI | PPI & NCI |
|---|---|---|
| **PROTEIN-CODING** | 30 [12; 64] | 54 [20; 107] |
| **TF** | 48 [20; 102] | 90 [41; 168] |
| **LNCRNA** | | 3 [1; 6] |
| **MIRNA** | | 52 [16; 224.75] |
| **OTHER** | | 11 [3; 25] |
| **PSEUDOGENE** | | 2 [1; 4] |

Table S 4 **Descriptive of Databases Genes and Reported Disease Associations.** Gene names were normalized to Gene Symbols and disease names were matched to MESH terms.

| | FINAL | | | ORIGINAL | | |
|---|---|---|---|---|---|---|
| **DATABASE** | **Associations** | **Genes** | **Diseases** | **Associations** | **Genes** | **Diseases** |
| **CLINGEN** | 523 | 446 | 138 | 896 | 695 | 352 |
| **CLINVAR** | 6,769 | 3,847 | 917 | 57,941 | 9,230 | 30,925 |
| **CTD** | 24,857 | 7,327 | 1,912 | 35,997 | 8,745 | 5,806 |
| **DISEASE ENHANCER** | 518 | 303 | 121 | 650 | 349 | 165 |
| **DISGENET (GDA)** | 36,317 | 8,912 | 2,099 | 95,959 | 10,147 | 11,181 |
| **GWAS CATALOG** | 8,957 | 4,254 | 434 | 70,192 | 12,786 | 2,339 |
| **HMDD** | 13,662 | 916 | 622 | 15,968 | 945 | 862 |
| **LNCBOOK** | 1,478 | 669 | 246 | 1,952 | 863 | 369 |
| **LNCRNA DISEASE** | 3,146 | 1,533 | 286 | 4,546 | 2,251 | 402 |
| **LOVD** | 2,931 | 2,113 | 664 | 6,611 | 3,910 | 4,860 |
| **MONARCH** | 19,542 | 8,059 | 1,058 | 33,549 | 11,531 | 8,593 |
| **OMIM** | 3,517 | 2,578 | 751 | 12,333 | 7,132 | 7,349 |
| **ORPHANET** | 3,787 | 2,555 | 721 | 8,252 | 4,324 | 3,784 |
| **PHEGENI** | 11,232 | 5,929 | 485 | 24,926 | 9,635 | 1,001 |
| **PSYGENET** | 2,924 | 1,354 | 40 | 3,621 | 1,493 | 108 |

Table S 5 **Genes with physical binding have higher co-expression in the physical networks.** Genes with a direct or indirect physical binding (PPI, PPI & NCI, or co-regulated by a ncRNA) have, in median, higher co-expression values than genes that do not physically interact. Kruskal-Wallis Test, Dunn's Post hoc test.

| | CORRELATION | | | WTO | | |
|---|---|---|---|---|---|---|
| **COMPARISON** | **P** | **P adj** | **Kruskal-Wallis** | **P** | **P adj** | **Kruskal-Wallis** |
| **DIRECT - NONE** | p < 0.001 | p < 0.001 | | p < 0.001 | p < 0.001 | |
| **INDIRECT - NONE** | p < 0.001 | p < 0.001 | p < 0.001 | p < 0.001 | p < 0.001 | p < 0.001 |
| **INDIRECT + DIRECT - NONE** | p < 0.001 | p < 0.001 | | p < 0.001 | p < 0.001 | |

Table S 6 **Power Law fit for each element in their respective networks.** The inclusion of the NCI does not change the PL nature of the network, but does change the coefficients of the degree exponent for $\gamma_{PPI} = 2.71$ to $\gamma_{PPI\ \&\ NCI} = 2.54$.

| NETWORK | TYPE | $K_{CUT}$ | $K_{SAT}$ | $\gamma$ |
|---|---|---|---|---|
| **PPI** | **PPI** | 2427 | 90 | 2.71 |
| | **PPI** | 2427 | 90 | 2.71 |
| | **TF** | 1564 | 100 | 2.42 |
| | **Protein Coding** | 1688 | 83 | 2.84 |
| **PPI & NCI** | **PPI & NCI** | 3918 | 100 | 2.54 |
| | **Protein Coding** | 1801 | 100 | 3.00 |
| | **TF** | 2632 | 100 | 2.43 |
| | **lncRNA** | 1618 | 1 | 1.59 |
| | **miRNA** | 4413 | 77 | 1.62 |
| | **Pseudogenes** | 123 | 1 | 1.76 |

Table S 7 **First Neighbors of drug-targets distribution.** Drug-Targets are more connected to miRNAs, TFs and Protein-coding Genes than non-drugged genes.

| | DRUGGED | NON-DRUGGED | P | P-ADJ |
|---|---|---|---|---|
| **PROTEIN-CODING** | 35 [15, 76] | 15 [4, 41] | <0.0001 | <0.0001 |
| **TF** | 9 [4, 21] | 6 [2, 16] | <0.0001 | <0.0001 |
| **LNCRNA** | 1 [1, 2] | 1 [1, 2] | 0.3875 | 0.3975 |
| **MIRNA** | 19 [5, 50] | 16 [4, 43] | <0.0001 | <0.0001 |
| **PSEUDOGENE** | 1 [1, 2] | 1 [1, 2] | 0.3237 | 0.2590 |

# 6 FIGURES

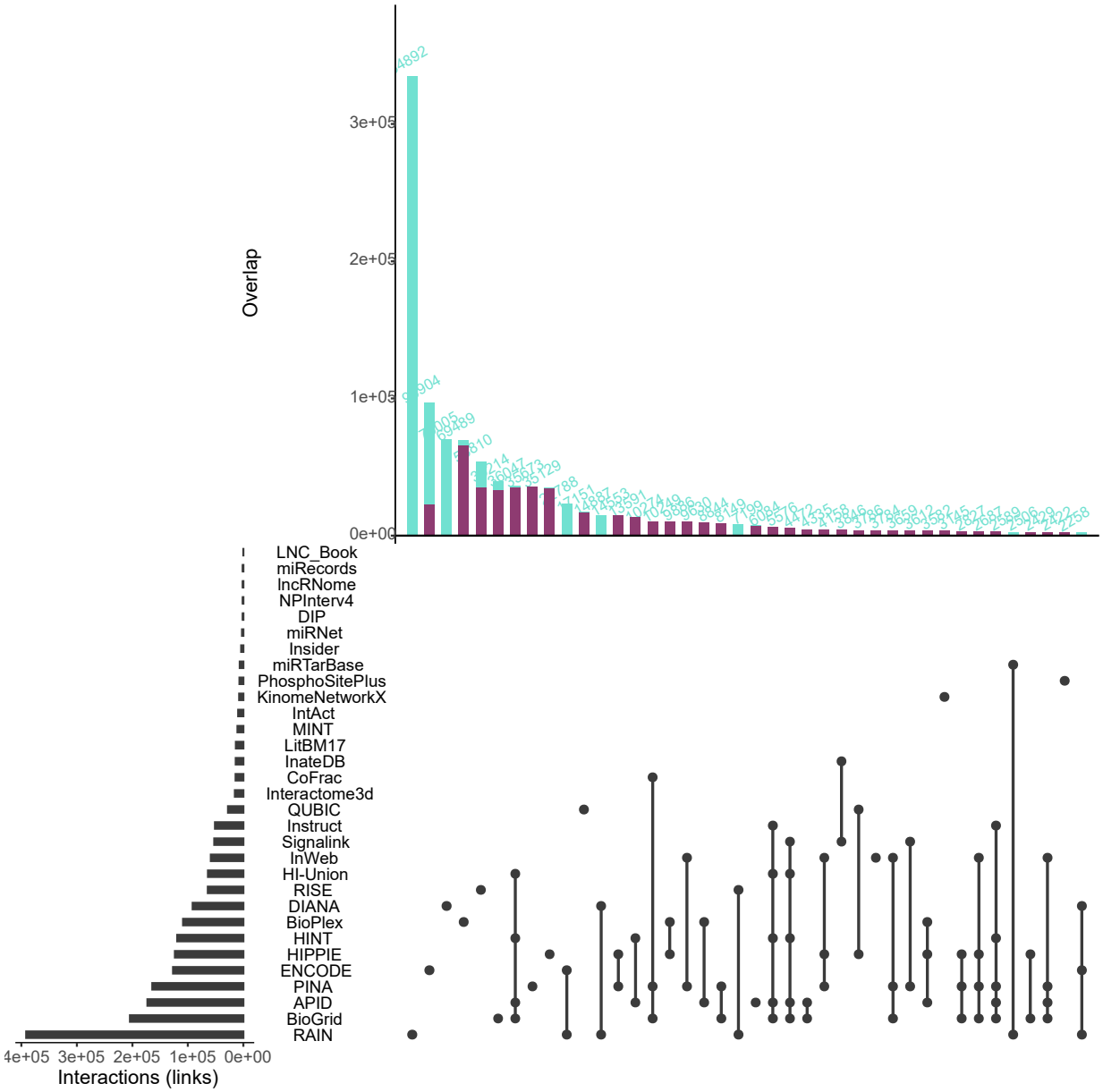


Figure S 1 **Network Validation: Gene-Interaction Overlap**. The upset plot depicts the number of binding interactions in each dataset, and how it overlaps across all PPI and NCI databases. Most of the interactions are database specific and do not occur in multiple sources, which provide us with complementary information. Interactions among proteins are represented in purple, while interactions involving at least one ncRNA are represented in turquoise.

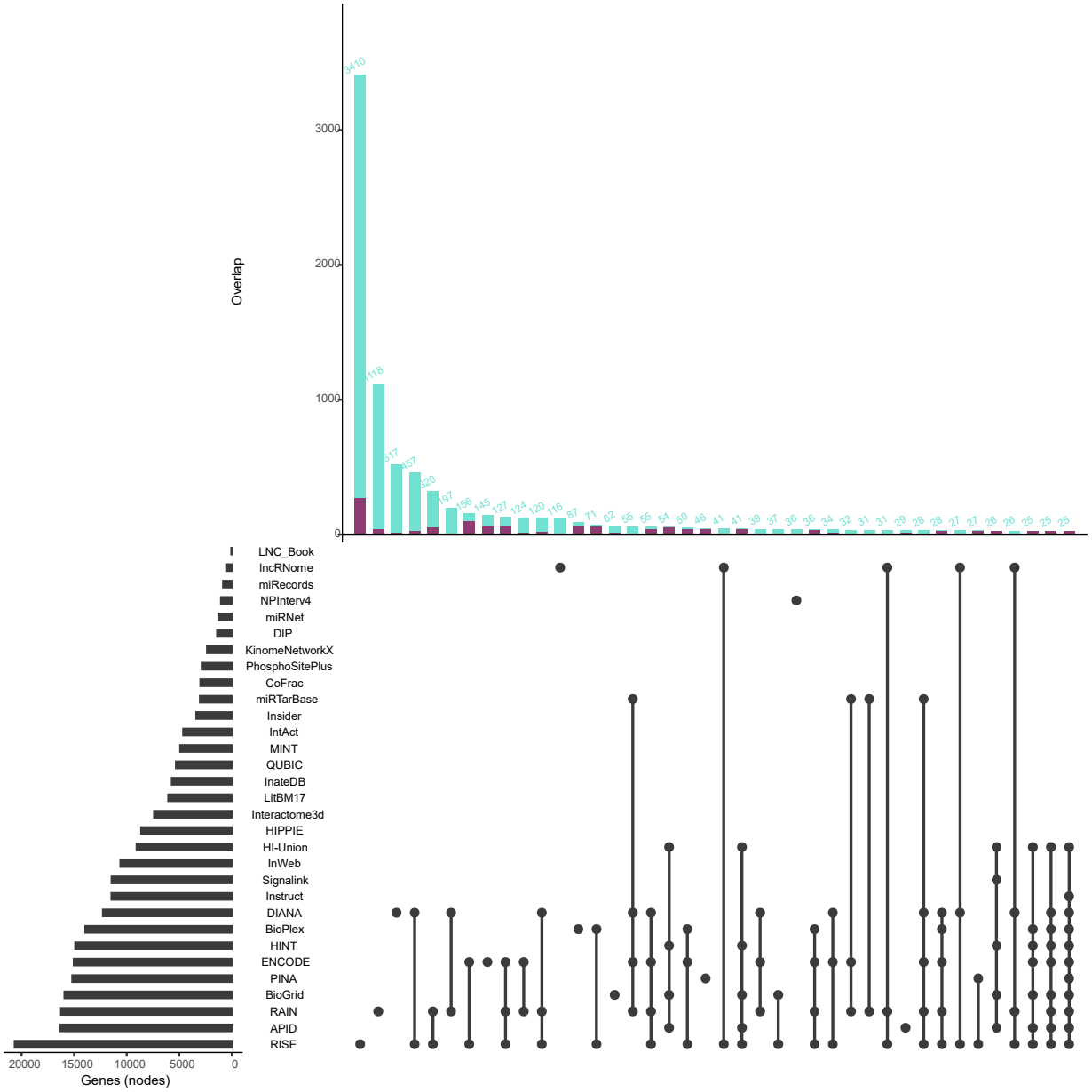


Figure S 2 **Network Validation: Gene Overlap**. The upset plot depicts the number of genes in each dataset, and how it overlaps across all PPI and NCI databases. Most of the genes are reported in multiple datasets. Proteins are represented in purple, while ncRNAs are represented in turquoise.

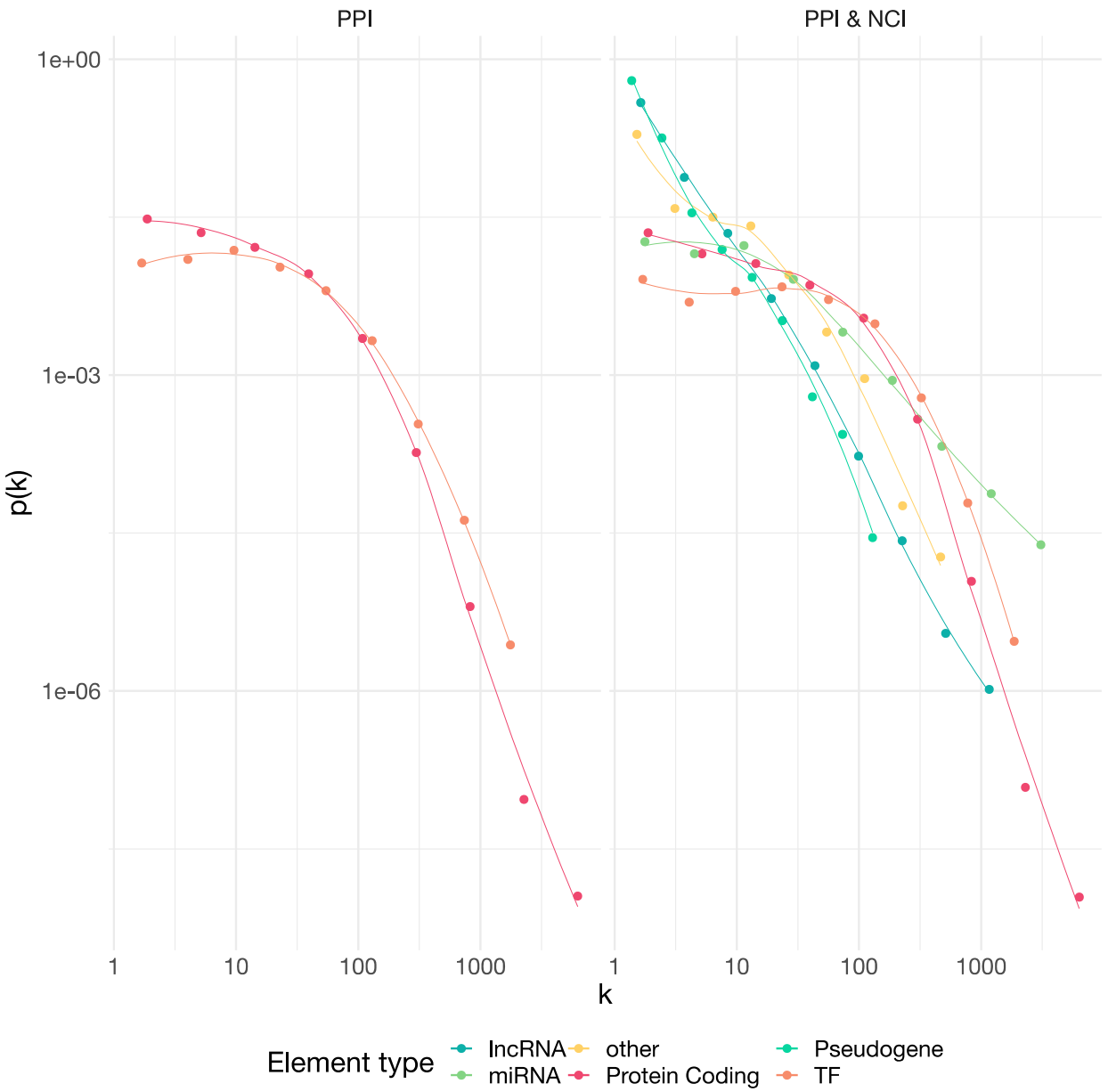


Figure S 3 **Degree distribution of different genomic elements.** The degree distribution on the three networks changes when we include non-coding elements into the PPI, the inclusion of miRNAs shows that they tend to have a higher degree, acting as regulatory factors and possibly as master regulators. Both PPI and PPI & NCI follow assintotically, for high (k), a power-law distribution. The diameter of the PPI is 7, while the diameter of the PPI & NCI is 9. Therefore, the diameter of the network increases with the NCI inclusion. Even though it might be counter intuitive, the PPI is more densely connected than the PPI & NCI. Meaning that, the inclusion of ncRNA mediated interactions more than double the number of interactions, and include more genes, creating a less dense network, where nodes are more dispersed and less closely connected.

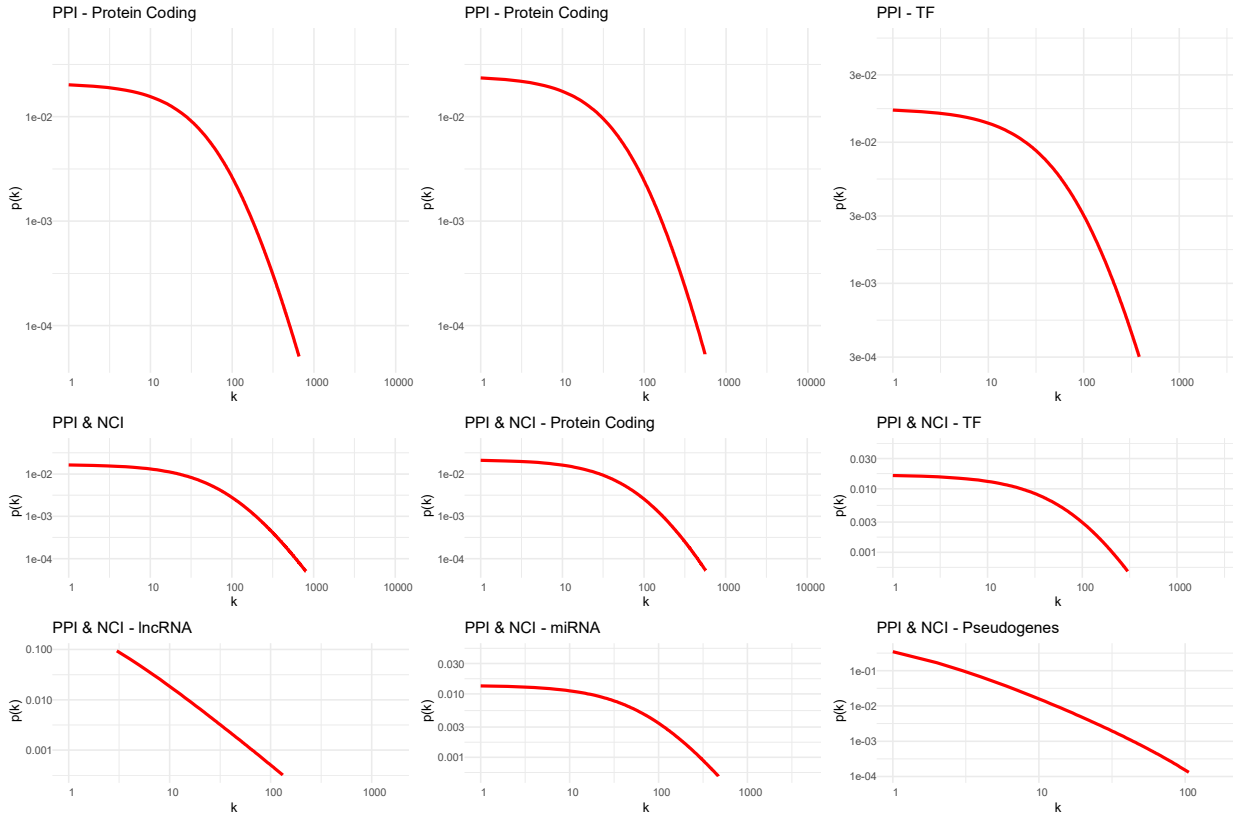


Figure S 4 **Degree distribution and power law fit (in red) for the network-based genomic elements.** The degree distribution and the power law for each genomic element was fitted independently. We fitted a Power-Law for real data as $(k + k_{sat})^{-\gamma} e^{\frac{-k}{k_{cut}}}$, where k is the degree, $k_{cut}$ is defined by the low-k saturation and is and $k_{sat}$ is the large-k cutoff. The identified $k_{cut}$ and $k_{sat}$ can be found on Table S 6.

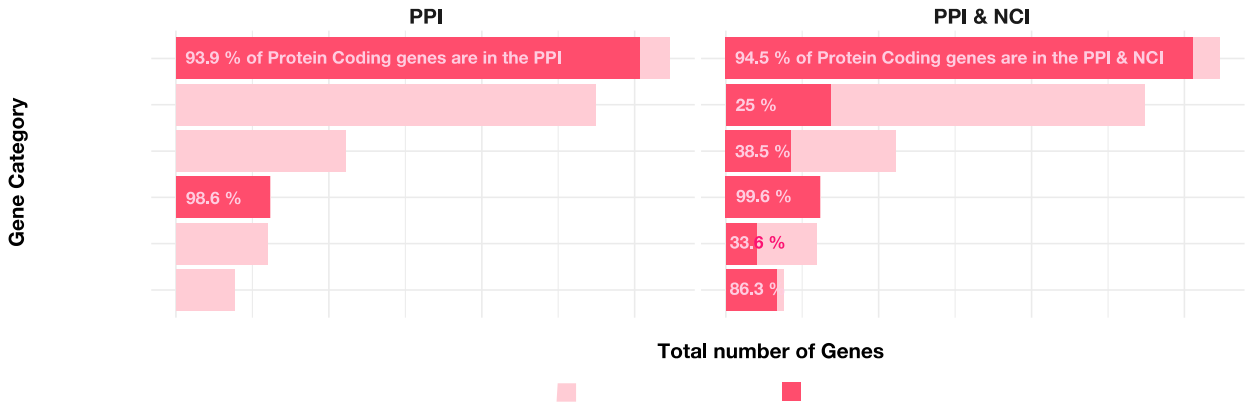


Figure S 5 **Gene completion mapped in the PPI and the combined network.** In the combined PPI & NCI we retrieve 86.3% of miRNAs, 99.6% of transcription factors and 38.5% of lncRNAs, increasing the interactions and coverage of the human transcriptomic.

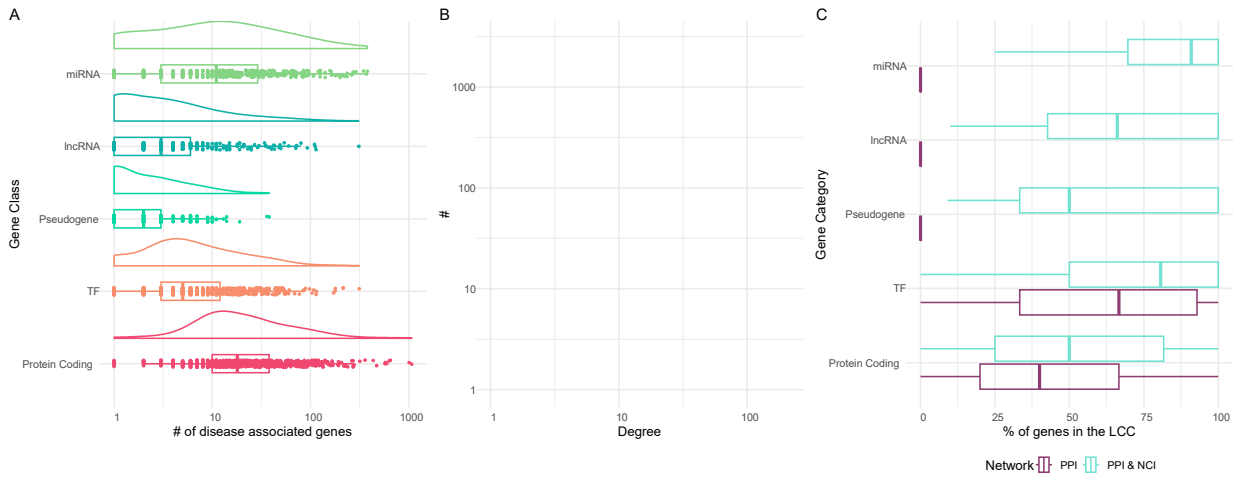


Figure S 6 **Disease Associated genes and their classification. A)** Rainplot of Number of disease-associated genes classified by gene category. We calculate the number of disease-associated genes in each disease, for each gene category. We find that Protein-Coding genes are the most associated with diseases, followed by miRNAs, TFs and lncRNAS. **B)** Degree distribution of gene-disease associations. miRNAs have a fat tail, indicating that few miRNAs can be associated with multiple diseases. **C)** The % of genes in each category found in the LCC. The inclusion of ncRNAs in the PPI allows us to increase the percentage of protein-coding genes retrieved in the disease modules from 40 to 50%. On median, 90% of miRNAs associated with a disease are found in the disease module.

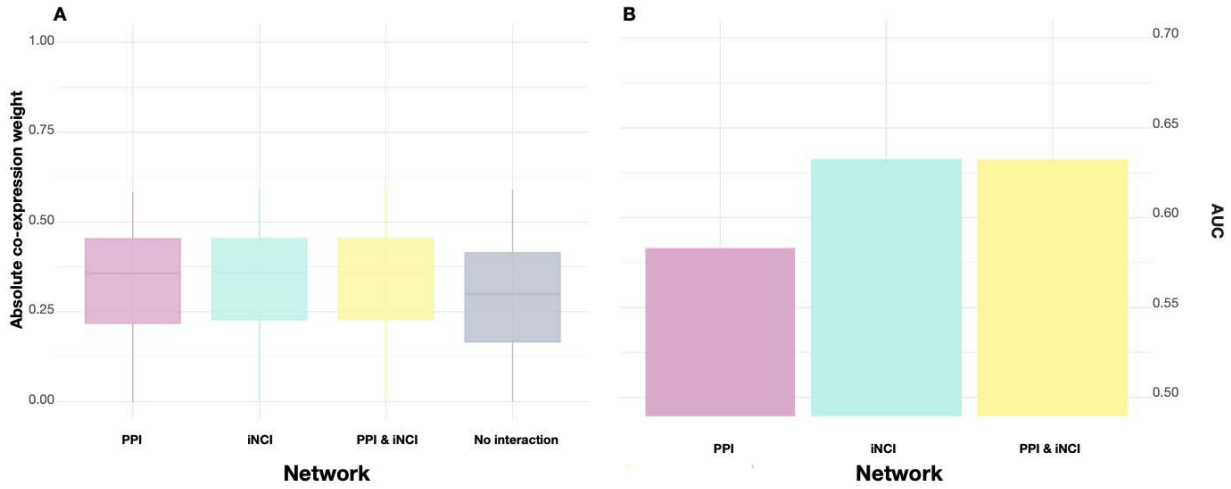


Figure S 7 **Absolute co-expression values (wTO) are higher in physical networks. A)** Genes with direct or indirect physical binding (PPI, PPI & NCI, or co-regulated by an ncRNA) have higher co-expression values than genes that do not physically interact in the wTO. The boxplot indicates that the absolute weighted Topological Overlap values are higher when there is physical interaction, compared to then non-existing links, indicating an association between physical binding and strength of co-expression. **B)** Co-Expression Networks Can Predict Physical Interactions. We use the wTO values between two transcripts to predict a direct or indirect binding, finding that the inclusion of ncRNAs increases the AUC in the iNCI and the PPI & iNCI.

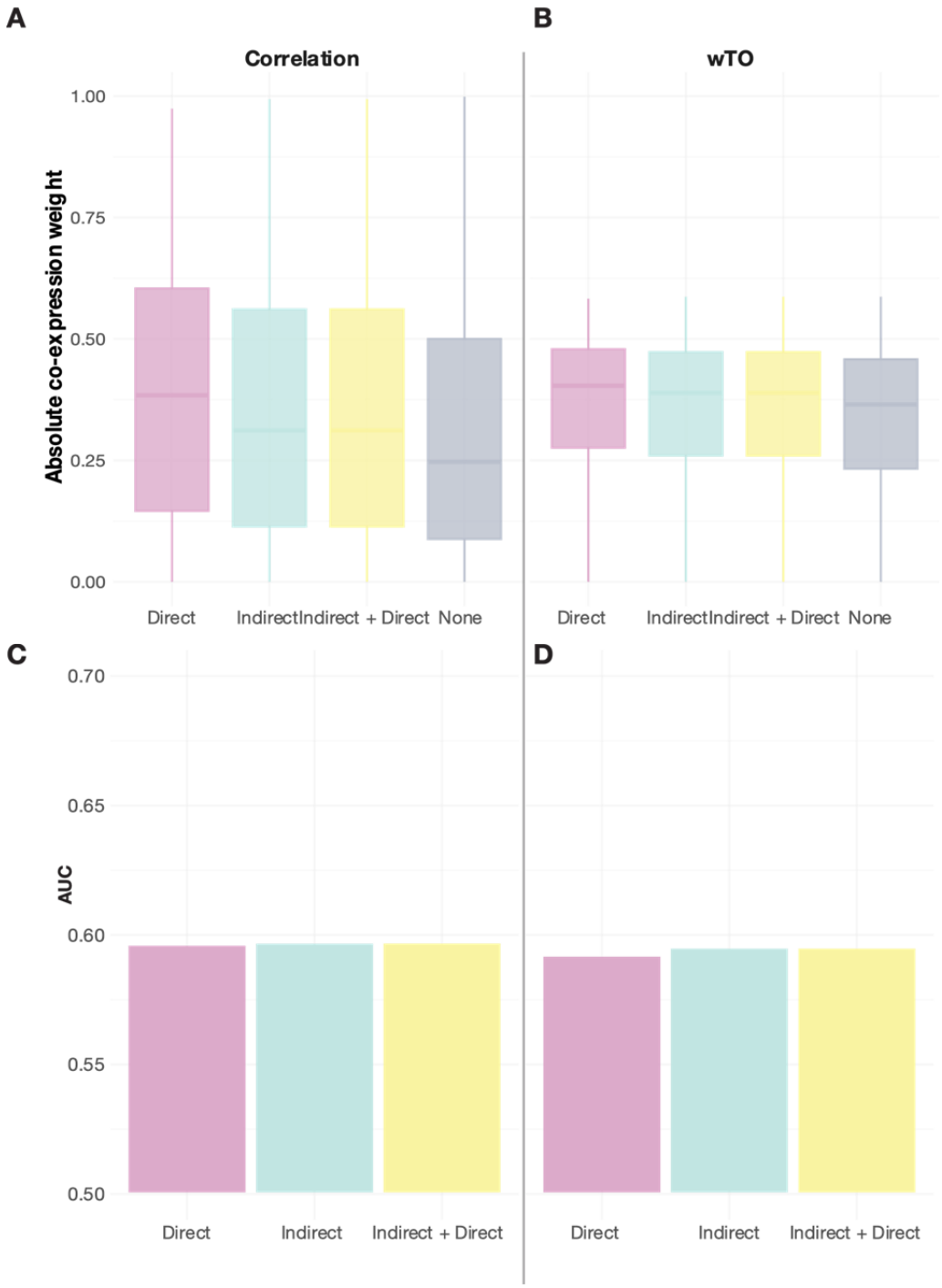


Figure S 8 Absolute co-expression values (correlation and wTO) are higher in TF-TF physical network and Co-Expression Networks Can Predict TF-TF Physical Interactions. A) TFs with direct or indirect physical binding (PPI, PPI & NCI, or co-regulated by an ncRNA) have higher co-expression values than genes that do not physically interact in the wTO. The boxplot indicates that the absolute correlation and B) weighted Topological Overlap values are higher when there is physical interaction, compared to then non-existing links, indicating an association between physical binding and strength of co-expression ($p < 0.01$; Kruskal-Wallis Test, Dunn's Post hoc test). C) Co-Expression Networks Can Predict Physical Interactions. We use the corrletion and D) wTO values between two TFs to predict a direct or indirect binding, finding that the inclusion of ncRNAs slighlty increases the AUC in the iNCI and the PPI & iNCI.

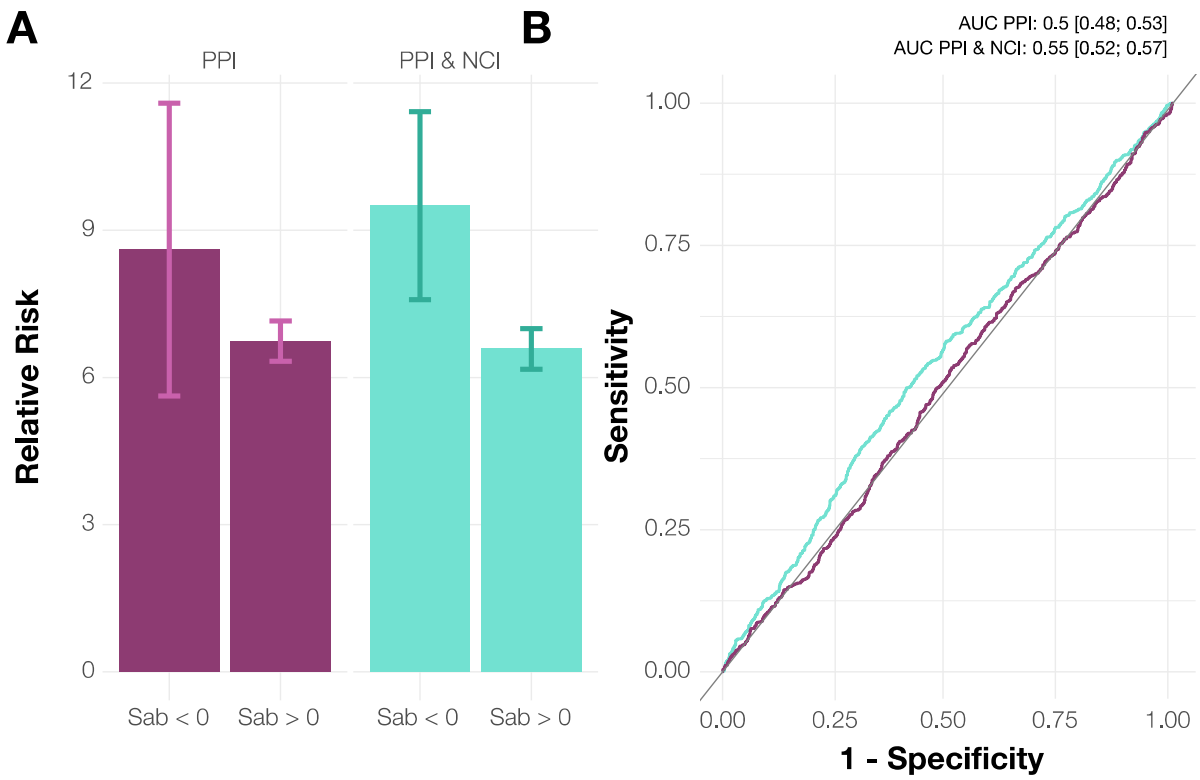


Figure S 9 **Disease Comorbidity and their separation. A)** Relative Risk is significatively higher for diseases closer to each other. Diseases with a negative $S_{ab}$ have significantly higher relative risks when compared to diseases with positive $S_{ab}$ for both the PPI and the PPI & NCI (Wilcoxon Test, p < 0.05). We observe for the PPI & NCI, on average, an increase on the RR of 9.5 for negative $S_{a,b}$ (se 2.93), and a decrease to 6.5 (se 0.41) for positive $S_{a,b}$ (Wilcoxon Test, p > 0.05). **B)** Network Separation is predictive of comorbidity. We use the network separation as a predictor of a significant Relative Risk > 1, finding that the PPI alone is as good as random (AUC = 0.5), while the PPI & NCI slightly increases the AUC to 0.55, even though the predictive power is still close to random, it suggests that ncRNAs might hold the key for improving disease comorbidity and progression identification.

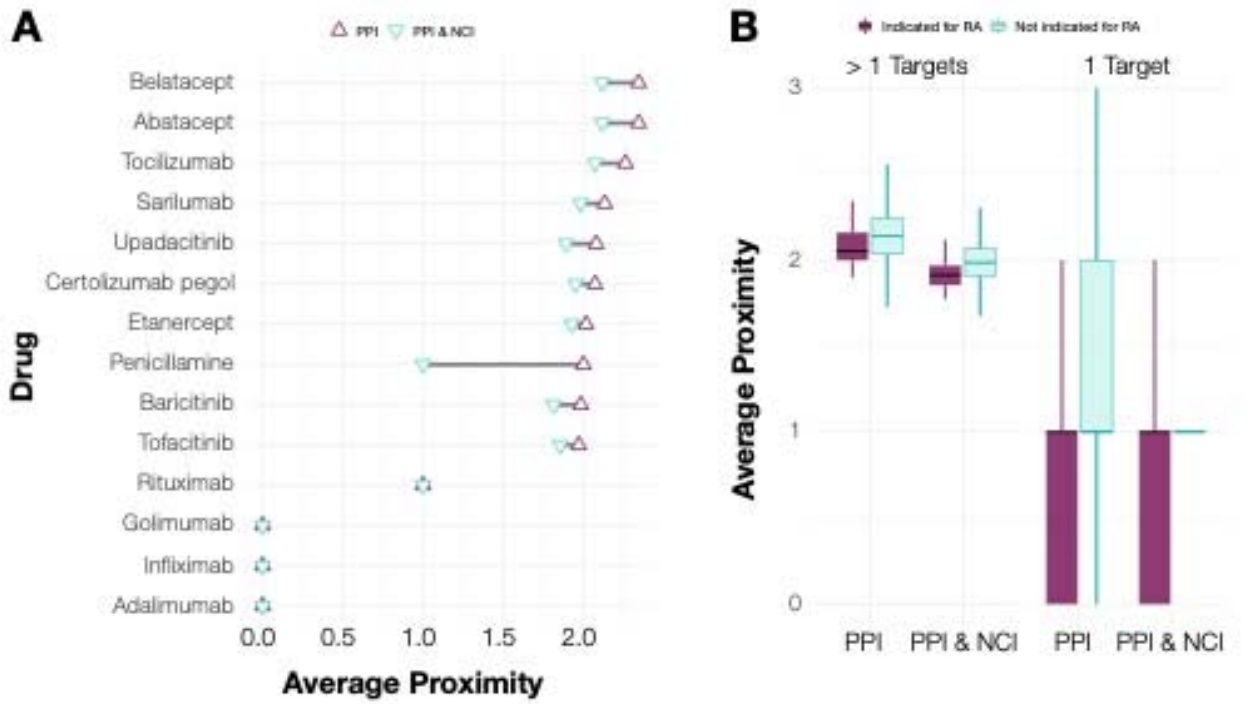


Figure S 10 **Average Proximity decreases for Rheumatoid Arthritis indicated drugs. A)** The dumbbell plot shows in the x-axis the average proximity for 14 drugs indicated to treat RA. In purple, the PPI values for each disease and in turquoise the PPI & NCI values. For all drugs with proximity higher than 1 we find that the PPI & NCI network decreases the distance of the drug-targets to the disease module**. B)** Average proximity for drug-targets with more than 1 target is statistically significant smaller for the PPI & NCI when compared to the PPI. The boxplots indicate a reduction on the proximity for drugs indicated to treat RA.

# 7 REFERENCES


1. Hombach, S. & Kretz, M. Non-coding RNAs: Classification, biology and functioning. *Advances in Experimental Medicine and Biology* **937**, 3–17 (2016).
2. Matera, A. G., Terns, R. M. & Terns, M. P. Non-coding RNAs: Lessons from the small nuclear and small nucleolar RNAs. *Nature Reviews Molecular Cell Biology* **8**, 209–220 (2007).
3. Hammond, S. M., Caudy, A. A. & Hannon, G. J. Post-transcriptional gene silencing by double-stranded RNA. *Nature Reviews Genetics 2001 2:2* **2**, 110–119 (2001).
4. John, B. *et al.* Human microRNA targets. *PLoS Biology* **2**, 661–663 (2004).
5. Pillai, R. S., Bhattacharyya, S. N. & Filipowicz, W. *Repression of protein synthesis by miRNAs: how many mechanisms? Trends in Cell Biology* **17**, 118–126 (Elsevier Current Trends, 2007).
6. Cui, C. & Cui, Q. The relationship of human tissue microRNAs with those from body fluids. *Scientific Reports* **10**, 1–7 (2020).
7. Zhuo, Y., Gao, G., Shi, J. an, Zhou, X. & Wang, X. MiRNAs: Biogenesis, origin and evolution, functions on virus-host interaction. *Cellular Physiology and Biochemistry* **32**, 499–510 (2013).
8. Lewis, B. P., Burge, C. B. & Bartel, D. P. *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets*. *Cell* **120**, 15–20 (2005).
9. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of Mammalian MicroRNA Targets. *Cell* **115**, 787–798 (2003).
10. Grimson, A. *et al.* MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell* **27**, 91–105 (2007).
11. Kudla, G., Granneman, S., Hahn, D., Beggs, J. D. & Tollervey, D. Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10010–10015 (2011).
12. Martin, M. M. *et al.* The human angiotensin II type 1 receptor +1166 A/C polymorphism attenuates microRNA-155 binding. *Journal of Biological Chemistry* **282**, 24262–24269 (2007).
13. Zhao, Y., Samal, E. & Srivastava, D. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature* **436**, 214–220 (2005).
14. Martinez, N. J. & Walhout, A. J. M. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *BioEssays* **31**, 435–445 (2009).
15. Ghafouri-Fard, S. *et al.* The interaction between miRNAs/lncRNAs and nuclear factor-κB (NF-κB) in human disorders. *Biomedicine and Pharmacotherapy* **138**, 111519 (2021).
16. Mattes, J., Collison, A., Plank, M., Phipps, S. & Foster, P. S. Antagonism of microRNA-126 suppresses the effector function of T H2 cells and the development of allergic airways disease. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 18704–18709 (2009).
17. Topol, A. *et al.* Dysregulation of miRNA-9 in a Subset of Schizophrenia Patient-Derived Neural Progenitor Cells. *Cell Reports* **15**, 1024–1036 (2016).
18. Kutsche, L. K. *et al.* Combined Experimental and System-Level Analyses Reveal the Complex Regulatory Network of miR-124 during Human Neurogenesis. *Cell Systems* **7**, 1–15 (2018).
19. Huang, X. & Wong, G. *An old weapon with a new function: PIWI-interacting RNAs in neurodegenerative diseases*. *Translational Neurodegeneration* **10**, (BioMed Central, 2021).
20. Wang, J. *et al.* PiRBase: A comprehensive database of piRNA sequences. *Nucleic Acids Research* **47**, D175–D180 (2019).
21. Kutter, C. & Svoboda, P. miRNA, siRNA, piRNA: Knowns of the unknown. *RNA Biology* **5**, 181–188 (2008).
22. Hu, B. *et al.* Therapeutic siRNA: state of the art. *Signal Transduction and Targeted Therapy* **5**, (2020).
23. Zhang, M. M., Bahal, R., Rasmussen, T. P., Manautou, J. E. & Zhong, X. bo. The growth of siRNA-based therapeutics: Updated clinical studies. *Biochemical Pharmacology* **189**, (2021).
24. D, B.-S. *et al.* siRNA-based identification of IBD-related targets in human monocyte-derived dendritic cells. *Journal of immunological methods* **494**, (2021).
25. Moreno-Montañés, J. *et al.* siRNA therapeutics in ocular diseases. in *Methods in Molecular Biology* **2282**, 417–442 (Methods Mol Biol, 2021).
26. MM, Z., R, B., TP, R., JE, M. & XB, Z. The growth of siRNA-based therapeutics: Updated clinical studies. *Biochemical pharmacology* **189**, (2021).
27. Dahariya, S. *et al.* Long non-coding RNA: Classification, biogenesis and functions in blood cells. *Molecular Immunology* **112**, 82–92 (2019).
28. Ferrè, F., Colantoni, A. & Helmer-Citterich, M. Revealing protein–lncRNA interaction. *Briefings in Bioinformatics* **17**, 106–116 (2016).

29. Tian, D., Sun, S. & Lee, J. T. The long noncoding RNA, Jpx, Is a molecular switch for X chromosome inactivation. *Cell* **143**, 390–403 (2010).
30. Lee, J. T. Lessons from X-chromosome inactivation: Long ncRNA as guides and tethers to the epigenome. *Genes and Development* **23**, 1831–1842 (2009).
31. Lyle, R. *et al.* The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. *Nature Genetics* **25**, 19–21 (2000).
32. Williamson, C. M. *et al.* Uncoupling antisense-mediated silencing and DNA methylation in the imprinted Gnas cluster. *PLoS Genetics* **7**, (2011).
33. Zhu, J. J., Fu, H. J., Wu, Y. G. & Zheng, X. F. Function of lncRNAs and approaches to lncRNA-protein interactions. *Science China Life Sciences* **56**, 876–885 (2013).
34. Wang, X. *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* **454**, 126–130 (2008).
35. Geisler, S. & Coller, J. RNA in unexpected places: Long non-coding RNA functions in diverse cellular contexts. *Nature Reviews Molecular Cell Biology* **14**, 699–712 (2013).
36. Rinn, J. L. *et al.* Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
37. Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120–126 (2011).
38. Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147**, 358–369 (2011).
39. Chen, Y., Li, Z., Chen, X. & Zhang, S. *Long non-coding RNAs: From disease code to drug role*. *Acta Pharmaceutica Sinica B* **11**, 340–354 (Chinese Academy of Medical Sciences, 2021).
40. Li, C. H. & Chen, Y. Targeting long non-coding RNAs in cancers: Progress and prospects. *International Journal of Biochemistry and Cell Biology* **45**, 1895–1910 (2013).
41. Wu, P. *et al.* Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. *Brain Research Bulletin* **97**, 69–80 (2013).
42. Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature Protocols* **2**, 1849–1861 (2007).
43. Wang, W., Caldwell, M. C., Lin, S., Furneaux, H. & Gorospe, M. HuR regulates cyclin A and cyclin B1 mRNA stability during cell proliferation. *EMBO Journal* **19**, 2340–2350 (2000).
44. Günzl, A., Palfi, Z. & Bindereif, A. Analysis of RNA-protein complexes by oligonucleotide-targeted RNase H digestion. *Methods* **26**, 162–169 (2002).
45. Shih, J. D., Waks, Z., Kedersha, N. & Silver, P. A. Visualization of single mRNAs reveals temporal association of proteins with microRNA-regulated mRNA. *Nucleic Acids Research* **39**, 7740–7749 (2011).
46. Fu, D. *et al.* Long non-coding RNA CRNDE regulates the growth and migration of prostate cancer cells by targeting microRNA-146a-5p. *Bioengineered* **12**, 2469–2479 (2021).
47. Karagkouni, D. *et al.* DIANA-LncBase v3: Indexing experimentally supported miRNA targets on non-coding transcripts. *Nucleic Acids Research* **48**, D101–D110 (2020).
48. Ma, L. *et al.* Lncbook: A curated knowledgebase of human long non-coding rnas. *Nucleic Acids Research* **47**, D128–D134 (2019).
49. Bhartiya, D. *et al.* LncRNome: A comprehensive knowledgebase of human long noncoding RNAs. *Database* **2013**, (2013).
50. Hsu, S. Da *et al.* MiRTarBase: A database curates experimentally validated microRNA-target interactions. *Nucleic Acids Research* (2011). doi:10.1093/nar/gkq1107
51. Huang, H. D. Y. *et al.* MiRTarBase 2020: Updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Research* **48**, D148–D154 (2020).
52. Xiao, F. *et al.* miRecords: An integrated resource for microRNA-target interactions. *Nucleic Acids Research* **37**, (2009).
53. Chang, L., Zhou, G., Soufan, O. & Xia, J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. *Nucleic acids research* **48**, W244–W251 (2020).
54. Teng, X. *et al.* NPInter v4.0: An integrated database of ncRNA interactions. *Nucleic Acids Research* **48**, D160–D165 (2020).
55. Junge, A. *et al.* RAIN: RNA-protein association and interaction networks. *Database* **2017**, (2017).
56. Gong, J. *et al.* RISE: A database of RNA interactome from sequencing experiments. *Nucleic Acids Research* **46**, D194–D201 (2018).
57. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *New England Journal of Medicine* (2015). doi:10.1056/nejmsr1406261
58. Landrum, M. J. *et al.* ClinVar: Improvements to accessing data. *Nucleic Acids Research* (2020). doi:10.1093/nar/gkz972
59. Davis, A. P. *et al.* Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research* (2020).

doi:10.1093/nar/gkaa891
60. Zhang, G. *et al.* DiseaseEnhancer: A resource of human disease-associated enhancer catalog. *Nucleic Acids Research* **46**, D78–D84 (2018).
61. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* (2020). doi:10.1093/nar/gkz1021
62. Piñero, J. *et al.* DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* **45**, D833–D839 (2017).
63. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* (2019). doi:10.1093/nar/gky1120
64. Huang, Z. *et al.* HMDD v3.0: A database for experimentally supported human microRNA-disease associations. *Nucleic Acids Research* (2019). doi:10.1093/nar/gky1010
65. Bao, Z. *et al.* LncRNADisease 2.0: An updated database of long non-coding RNA-associated diseases. *Nucleic Acids Research* (2019). doi:10.1093/nar/gky905
66. Chen, G. *et al.* LncRNADisease: A database for long-non-coding RNA-associated diseases. *Nucleic Acids Research* (2013). doi:10.1093/nar/gks1099
67. Fokkema, I. F. A. C. *et al.* LOVD v.2.0: The next generation in gene variant databases. *Human Mutation* **32**, 557–563 (2011).
68. McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. *American Journal of Human Genetics* (2007). doi:10.1086/514346
69. Ramos, E. M. *et al.* Phenotype-genotype integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources. *European Journal of Human Genetics* **22**, 144–147 (2014).
70. Gutierrez-Sacristan, A. *et al.* PsyGeNET: a knowledge platform on psychiatric disorders and their genes. *Bioinformatics* **31**, 3075 (2015).
71. Barabási, A.-L. *Network Science*. (Cambridge University Press, 2016).
72. Gysi, D. M. & Nowick, K. Construction, comparison and evolution of networks in life sciences and other disciplines. *Journal of the Royal Society Interface* **17**, 20190610 (2020).
73. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
74. Gysi, D. M., Voigt, A., Fragoso, T. de M., Almaas, E. & Nowick, K. wTO: an R package for computing weighted topological overlap and a consensus network with integrated visualization tool. *BMC Bioinformatics* **19**, 392 (2018).
75. Nowick, K., Gernat, T., Almaas, E. & Stubbs, L. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proceedings of the National Academy of Sciences* **106**, 22358–22363 (2009).
76. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
77. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research* (2018). doi:10.1093/nar/gkx1037
78. Singerman, L. Combination therapy using the small interfering RNA bevasiranib. *Retina* **29**, (2009).
79. Ray, K. K. *et al.* Two Phase 3 Trials of Inclisiran in Patients with Elevated LDL Cholesterol. *New England Journal of Medicine* **382**, 1507–1519 (2020).