Noncoding RNAs improve the predictive power of network medicine

Deisy Morselli Gysin and Albert-László Barabási

Network medicine has improved the mechanistic understanding of disease, offering quantitative insights into disease mechanisms, comorbidities, and novel diagnostic tools and therapeutic treatments. Yet, most network-based approaches rely on a comprehensive map of protein–protein interactions (PPI), ignoring interactions mediated by noncoding RNAs (ncRNAs). Here, we systematically combine experimentally confirmed binding interactions mediated by ncRNA with PPI, constructing a comprehensive network of all physical interactions in the human cell. We find that the inclusion of ncRNA expands the number of genes in the interactome by 46% and the number of interactions by 107%, significantly enhancing our ability to identify disease modules. Indeed, we find that 132 diseases lacked a statistically significant disease module in the protein-based interactome but have a statistically significant disease module after inclusion of ncRNA-mediated interactions, making these diseases accessible to the tools of network medicine. We show that the inclusion of ncRNAs helps unveil disease–disease relationships that were not detectable before and expands our ability to predict comorbidity patterns between diseases. Taken together, we find that including noncoding interactions improves both the breadth and the predictive accuracy of network medicine.

A key goal of post-genomic medicine is to translate the detailed inventory of cellular components and their disease-related mutations, offered by genomics, into mechanistic insights into disease causation and progression, ultimately resulting in novel treatments. To achieve this, we must catalog the physical interactions between all cellular components, arriving at an accurate and predictive map of the human subcellular network. Network medicine, a discipline whose goal is to exploit the predictive power of subcellular networks (1), has already improved our understanding of disease classification (2) and progression (3), disease–disease comorbidities (4), similarities (4), and treatments (5) and offered tools to identify drug repurposing opportunities (6, 7) and drug combinations (8). Some of these tools have already entered the clinical practice, resulting in network-based diagnostic tools currently used by doctors to improve treatment outcomes for rheumatoid arthritis (RA) patients (9) and the drug repurposing opportunities identified during the COVID-19 pandemic (6). These advances relied on maps of experimentally confirmed protein–protein interactions (PPI) and supported multiple foundational discoveries, like the tendency of proteins associated with the same disease to be co-localized in the same neighborhood of the PPI network (4), the finding that diseases with similar phenotypes lie in similar regions of the interactome (4) and the discovery that the network-based distance of drugs to a disease module affects drug efficacy (5, 6).

However, the current interactome maps ignore an important component of subcellular dependency, the regulatory interactions induced and mediated by noncoding RNAs (ncRNAs). In-depth transcriptome sequencing estimates that while approximately about 74% of the human genome is transcribed (10), only 2 to 3% of the human genome encodes for proteins (10–13), meaning that the remaining transcripts represent different classes of noncoding elements (13). These ncRNAs (Fig. 1A) contribute to multiple biological functions, from the maintenance and regulation of gene expression to pre- and post-processing of mRNAs, splicing, decoding mRNAs into amino acids, and the control of gene expression (14–19), ultimately contributing to multiple diseases (20).

Significance

Network medicine has been used to quantify disease mechanisms, comorbidities, and treatments, but most approaches have ignored interactions mediated by noncoding RNAs (ncRNAs). This study systematically combines experimentally confirmed ncRNA and protein–protein interactions to construct a comprehensive network of all physical interactions in the human cell. The inclusion of ncRNA increases the number of genes and interactions in the interactome and enhances the ability to identify disease modules and predict comorbidity patterns between diseases. Ultimately, this study shows that including noncoding interactions improves the breadth and accuracy of network medicine.
the PPI network, resulting in a noncoding interactome (NCI). We find that the inclusion of ncRNAs increases the number of nodes by 46% and the number of links by 107% compared to the currently used PPI-based interactome. Most importantly, we find that this expansion allows us to identify disease modules for 132 diseases that lacked a statistically significant module before; hence, it could not be previously explored with the tools of network medicine. Finally, we show that the expanded interactome improves the prediction of disease–disease relationships, offering more accurate comorbidity predictions and advances that ultimately will lead to better prevention and personalized medicine.

Results

PPI Network. The human protein–protein interactome (Fig. 1C) was derived from 21 public databases containing different types of experimentally derived PPI data (6): 1) binary PPIs, derived from high-throughput yeast-two hybrid (Y2H) experiments (HI-Union), three-dimensional (3D) protein structures (Interactome3D, Instruct, Insider) or literature curation (PINA, MINT, LitBM17, Interactome3D, Instruct, Insider, BioGrid, HINT, HIPPIE, APID, and InWeb); 2) PPIs identified by affinity purification followed by mass spectrometry present in BioPlex2, QUBIC, CoFrac, HINT, HIPPIE, APID, LitBM17, and InWeb; 3) kinase–substrate interactions from KinomeNetworkX, and PhosphoSitePlus; 4) signaling interactions from SignaLink and InnateDB; and 5) regulatory interactions derived by the ENCODE consortium. We used the curated list of PSI-MI IDs provided by Alonso-López et al. (21) to differentiate binary interactions among the several experimental methods present in the literature-curation databases. Specifically, for InWeb, interactions with curation score <0.175 (75th percentile) were not considered. All proteins were mapped

---

**Fig. 1.** The role of ncRNA in gene regulation and connection to the human interactome. (A) The modern central dogma of biology. A DNA strand showing the transcription process: miRNAs, lncRNAs, and mRNAs are all transcribed from the DNA; however, only processed mRNAs are translated into proteins. (B) Interaction between ncRNAs. miRNAs can bind to lncRNAs, mRNAs, and proteins. When miRNAs interact with mRNAs and lncRNAs they regulate (by activating or repressing) the gene expression process. lncRNAs can also bind to miRNAs, mRNAs, and proteins. (C) A network of protein interactions. Proteins interact with one another, forming a protein–protein interaction network. Some proteins act as TFs, which regulate gene expression. The PPI only accounts for binding interaction among protein-coding genes. (D) A network of all interactions. ncRNAs and protein-coding RNAs interact with each other, forming a densely connected network, the PPI & NCI, which contains multiple types of physical interactions from different genomic elements. (E) PPI & NCI. Each edge on the network represents the relative frequency of all respective interactions across different element types. The PPI is a subgraph of the PPI & NCI, which only accounts for protein-coding genes and their interactions; TFs and Proteins interact with each other, responsible for 33% of the interactions in the PPI & NCI network, showing that even though protein interactions play a big role on the network, their interaction with other groups is also important. The majority of interactions occur between miRNAs and protein-coding genes and TFs. While IncRNAs interact with protein-coding genes and other TF, they interact with few other elements.
to their corresponding Gene Symbol (NCBI) and proteins that could not be mapped were removed. As each database contributes with interactions between a different set of proteins (SI Appendix, Table S1), the resulting PPI network contains 536,965 interactions between 18,217 proteins (Fig. 1E). Interactions containing at least one ncRNA transcript were included in the NCI.

**NCI Network.** The most studied ncRNAs are microRNAs (miRNAs), that contain ~22 nucleotides (18) and mainly act at the post-transcriptional level (19), involved in mRNA cleavage, activating or repressing translation (17, 18). The human genome accounts for approximately 2,300 miRNAs (22), each with hundreds of targets, which together regulate 10 to 30% of all genes (23). miRNA recognizes its mRNA targets by base-pairing the miRNA seed region (containing 2 to 8 nucleotides) to the complementary region on the targeted mRNA (24) (Fig. 1B). Playing a similar role as Transcription Factors (TFs), miRNAs form an interwoven network (25) that affects disease development (26) as documented in asthma (27) or schizophrenia (28), and mutated or dysregulated miRNAs are associated with the lack of function in neurogenesis (29).

Long ncRNAs (lncRNAs), another family of nonprotein-coding RNAs, that exceed 200 nucleotides (30), present a poli-A tail and can be spliced. Even though only a small number of IncRNAs are well characterized (31), they are involved in a wide range of biological functions, from X-chromosome inactivation (32, 33), to imprinting (34, 35), and often interact with proteins (36–38)—acting as a TF, and sponges for miRNA (39), bind to chromatin (40) and enhancers (41). IncRNAs can bind to both RNAs and proteins and hence are classified into protein-focused and RNA-focused (31) (Fig. 1B). Moreover, IncRNAs are also associated with multiple diseases (42), cancers (43), autoimmune neuropathies (25), and neurodegenerative diseases (44), and the IncRNA CRNDE has been identified as a target for therapeutic treatment of prostate cancer (45).

To construct the human NCI (Fig. 1D), we combine nine publicly available databases that collect and curate experimentally derived noncoding interactions: 1) miRNA-targets, derived from reporter assay, western blot, microarray, and next-generation sequencing experiments from mirTarBase; 2) miRNAs and IncRNAs, that exceed 200 nucleotides, that present a poli-A tail and can be spliced. Even though only a small number of IncRNAs are well characterized, they are involved in a wide range of biological functions, from X-chromosome inactivation (32, 33), to imprinting (34, 35), and often interact with proteins (36–38)—acting as a TF, and sponges for miRNA (39), bind to chromatin (40) and enhancers (41). IncRNAs can bind to both RNAs and proteins and hence are classified into protein-focused and RNA-focused (31) (Fig. 1B).

We begin by constructing two networks: i) the protein–protein interaction (PPI) network, which contains 536,965 interactions between 18,217 protein-coding genes, and ii) the joint PPI & NCI network, which has 26,575 genes (18,358 coding, 2,134 IncRNA, 1,650 miRNAs, 3,429 pseudogenes and 1004 other ncRNAs such as piRNA, siRNA (small interfering RNA), tRNAs] connected by 1,114,777 binding interactions (Fig. 1E).

The inclusion of ncRNAs increases the number of nodes by 46% and the number of links by 107%, compared to the PPI. The final interactome is fairly complete, containing 94.5% of all human proteins, 99.6% of all TFs, 86.3% of all miRNAs, and 38.5% of all IncRNAs transcripts (SI Appendix, Fig. S5). The inclusion of ncRNAs in the network increases the diameter of the PPI from 7 to 9, mainly because of the larger number of genes, also inducing an increase in the average shortest path length from 2.66 (PPI) to 2.79 (PPI & NCI; SI Appendix, Table S2). We have also analyzed the degree distribution of the PPI and the PPI & NCI networks, finding that while the inclusion of the NCI does not affect the scale-free nature of the network (56), it does alter the coefficients of 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 2.54 less than the degree exponent for the PPI network, indicating that the NCI increases the role of the hubs, leading to a more heterogeneous network.

While in the PPI proteins interact with 30 [12; 64] median value (interquartile range) other proteins, in the PPI & NCI network, the median degree increases to 54 [20; 107]. A TF in the PPI interacts with 48 proteins [20; 102] and its degree increases to 90 [41; 168] after the inclusion of the ncRNAs. miRNAs and IncRNAs, absent in the PPI, are connected to 52 [16; 224.75] and 3 [1; 6] elements (miRNAs, IncRNAs, proteins, TFs, or other
ncRNAs) respectively (SI Appendix, Table S3). In the PPI & NCI network, 34.8% of the interactions are between miRNAs and protein-coding genes and 16% between miRNAs and TFs, and lncRNAs are responsible for less than 1% of the combined interactions (Fig. 1E). In summary, we find that miRNAs are responsible for the majority of the noncoding interactions with protein-coding genes, TFs, and pseudogenes, as well as 40% of the lncRNAs interactions (Fig. 1E), confirming the central role miRNAs play in the regulation of the underlying cellular network.

### Disease–Disease Associations Under the Light of ncRNAs

Relating genes and their mutations to diseases is the central question of post-genomic medicine, drug-based therapeutics, drug discovery, and drug repurposing. Traditionally, genes were associated with traits via positional cloning (57), but recently most gene–disease associations come from genome-wide association studies (GWAS) followed by functional analysis to evaluate the effect of a single nucleotide polymorphism (SNP) or gene on the trait. Genes can be also associated with diseases using differential gene expression, epigenetic, susceptibility and other association studies, and literature-based curation (followed or not by an expert review). We assembled a Gene Disease Association (GDA) database by retrieving evidence of disease associations from 15 well-curated databases (ClinGen, ClinVar, CTD, Disease Enhancer, DisGeNET, GWAS Catalog, HMDD (58), lncBook, LncRNA disease, LOVD, Monarch, OMIM, Orphanet, PheGenI, and PsyGenNet; see Material and Methods and SI Appendix, section 2.2 and Table S2).

We limit our exploration to diseases with at least 10 gene associations, arriving at a database with 861 diseases and 13,216 disease-associated genes, of which 10,764 are protein-coding and 2,452 encode ncRNAs. The diseases with most miRNAs associations are carcinoma hepatocellular (380 miRNAs), breast neoplasms (372 miRNAs), and colorectal neoplasms (347 miRNAs). Diseases with high lncRNAs involvement include astrocytoma (311 lncRNAs), breast neoplasms (114 IncRNAs), and stomach neoplasms (112 IncRNAs). We find that 250 of the 861 diseases are enriched for ncRNAs (proportions test; P adj < 0.05, FDR corrected), i.e., they have more associated ncRNAs than we would expect by chance. Previous results show that proteins associated with a specific trait, phenotype, disease, or biological process tend to cluster in the same topological neighborhood of the PPI network (59), forming a sub-network known as the disease module. The phenomena that proteins linked to the same phenotype have a strong tendency to interact with each other (and to cluster in the same network neighborhood) has been documented multiple times (1, 60–63) and the occurrence of such phenomena is the basis of disease module, representing a connected sub-network of the disease-associated genes with a potential mechanistic link to a particular phenotype. The size of the largest connected component (LCC) of this sub-network measures the magnitude of the disease module, and its statistical significance is obtained by deriving the LCC distribution based on the random selection of the disease-associated genes. It is important to note that the size of the LCC must be accompanied by its statistical significance, as even a large LCC could emerge by chance. As a second measure, here we introduce the relative LCC (rLCC), defined as the ratio between the size of LCC and the number of genes associated with the disease. The rLCC captures the completeness of the disease module, allowing comparison across diseases with different numbers of genes. For example, schizophrenia has 1,458 associated genes (1,292 coding, 148 ncRNAs) of which 1,364 form an LCC in the PPI & NCI (P adj < 0.05, FDR corrected), resulting in an rLCC of 93.55%. In contrast, even with the inclusion of ncRNAs, obsessive-compulsive disorder, with 92 genes (81 protein-coding and 11 noncoding) has an LCC of 8 in the PPI & NCI, resulting in an rLCC of only 8.69%, indicating that the corresponding disease module is highly incomplete.

In the following, we focus on three diseases, RA, Chron’s disease (CD), and pre-eclampsia (PE), to illustrate the value of adding the NCI to the interactome. RA is a multisystemic, chronic inflammatory disease characterized by destructive synovitis, erosive arthritis, progressive articular damage, and systemic organ involvement (64–66), which can lead to decreased functional capacity and quality of life, increased morbidity and mortality. RA is associated with 391 genes, (302 protein-coding, 66 miRNAs, 14 lncRNAs, and 9 other ncRNAs). While there is a significant disease module (P adj < 0.05, FDR corrected) in both the PPI and the PPI & NCI, the PPI’s disease module accounts for only 189 disease-associated genes (rLCCPPI: 62%), whereas the PPI & NCI includes 345 disease-associated genes (Fig. 2A), increasing the rLCCPPI & NCI to 88%.

CD is an autoimmune disease, that causes inflammation of the digestive tract, which can lead to abdominal pain, severe diarrhea, fatigue, weight loss, and malnutrition. CD is associated with 279 genes (204 protein-coding, 47 miRNAs, and 14 IncRNAs). It has a significant disease module in both networks; however, the size of the LCC more than doubles with the inclusion of the ncRNA (LCCPPI: 93; LCCPPI & NCI: 217), and the proportion of disease-associated genes also increases greatly (from rLCCPPI: 45% to rLCCPPI & NCI: 77%) (Fig. 2B), collecting many more known disease genes.

Finally, PE is characterized by persistent high blood pressure during pregnancy or postpartum, and in rare cases, it can progress to severe PE very quickly, which can lead to the death of the mother and the baby. PE can also lead to premature birth, malnutrition, and lack of oxygen in the womb; adults whose mothers had PE have higher chances of developing diabetes, congestive heart failure, and hypertension (67). PE has no statistically significant disease module in the PPI network, hence, previously we could not apply network medicine tools to explore this disease. Indeed, PE is associated with 136 genes; the majority of which (95) encode miRNAs. It does, however, have a significant module in the combined PPI & NCI network (Fig. 2C). The PPI network accounts for only 33 of the 136 genes associated with the disease, of which only 2 are part of the LCC, while the PPI & NCI contains 136 genes of which 111 are in the LCC. Similarly, the ratio of disease genes found in the disease module jumps from 6% (rLCCPPI) to 81% (rLCCPPI & NCI), confirming the key role miRNAs play in regulation of PE. In other words, the PPI, historically the starting point of network medicine studies, provides a highly incomplete map of the PE disease module. However, the inclusion of ncRNAs allows us to now detect the disease module, opening up the possibility to explore the disease using the tools of network medicine.

To generalize beyond RA, CD, and PE, we calculated the LCC, rLCC, and the significance of the disease module for all 861 diseases in the PPI and in the PPI & NCI. Of the 861 diseases, 505 have a statistically significant LCC in the PPI network and 602 have a statistically significant LCC in the PPI & NCI (permutation test, N = 1,000, P adj < 0.05, FDR corrected; Fig. 3A). Taken together, we find that for 132 diseases the identification of a statistically significant disease module cannot be achieved without the inclusion of ncRNA-based interactions. We also find that the inclusion of ncRNAs decreases the LCC P-value (FDR corrected) for 522 diseases, hence increasing our confidence over the observed disease module (Fig. 3 B, C, and E) and also increases the median size of
the significant LCCs (Wilcoxon test, $P < 0.0001$; Fig. 3D). Finally, we find that the rLCC in the PPI & NCI increases for 430 diseases by 26.6% on average, indicating that the NCI considerably reduces the number of unconnected disease genes by linking many previously isolated components to the LCC.

Taken together, we find that including ncRNAs in the PPI increases the size, significance, and disease-gene retrieval of the disease module, expanding the reach of network medicine to a large number of diseases that previously could not be explored using network-based tools. Digestive problems, pathologies, and neoplasm have the highest ratio of ncRNAs (Fig. 3F), indicating that those disease classes benefit most from the inclusion of ncRNAs.

**Fig. 2.** Disease modules for RA, CD, and PE. (A) RA disease module. In RA, both PPI and the PPI & NCI networks do form a significant LCC; however, the inclusion of ncRNAs into the network allows a much better retrieval of disease genes, increases the rLCC from rLCC_{PPI}: 62% to rLCC_{PPI & NCI}: 88%. Gray nodes are not in the LCC of the complete interactome, while the colored nodes are present. (B) CD disease module. Both PPI and the PPI & NCI identify significant disease modules. However, by including ncRNAs into the disease module of CD, the rLCC increases from rLCC_{PPI}: 45% to rLCC_{PPI & NCI}: 77%, increasing the disease gene retrieval. (C) PE disease module. PE shows that the inclusion of ncRNAs can change our ability to identify significant disease modules. In the PPI, we are unable to define and identify a disease module; however, when in the PPI & NCI, a disease module emerges (rLCC_{PPI}: 6%, and rLCC_{PPI & NCI}: 81%).

**Direct and Indirect Bindings Are Supported by Co-Expression Networks.** Gene co-expression networks are often used to shed light on the molecular mechanisms that underlie biological processes (68). As gene co-expression is driven by the regulatory network, it leads to the hypothesis, supported by previous evidence (69), that interacting proteins are products of genes with higher co-expression, compared to proteins that do not physically interact. Here we extend this hypothesis to noncoding elements, asking if genes connected by noncoding interactions show higher co-expression than expected by chance. In other words, we use gene co-expression to probe the relative strength of the interactions induced by coding and noncoding elements. As bulk RNA-seq is not appropriate for measuring miRNA, due to the lack of poli-A tail in miRNAs, here we derive an indirect physical network that includes co-regulatory interactions (Fig. 4A), driven by the hypothesis that two proteins are co-expressed if they are co-regulated by the same ncRNAs. We distinguish three interactions that could modulate co-expression networks: i) Direct PPI (Fig. 4B), constructed using direct physical interaction between two proteins; ii) Indirect NCI (iNCI; Fig. 4C), connecting two proteins if proteins “A” and “B” are co-regulated by the same ncRNA; iii) Direct & iNCI (Fig. 4D), representing the combination of the PPI and the iNCI network; or iv) For control, we measure co-expression between nodes that have no known physical interaction between them.

We compare the measured co-expression for the three different interaction types (PPI, iNCI, and PPI & iNCI), relying on gene
co-expression derived from whole blood samples by GTEx (70). We use both Pearson correlation ($\rho$) and wTO (71, 72) ($\omega$) to ensure that the results are not biased by the methodology. We then compare the absolute co-expression weights ($\rho$ and $\omega$) for the three binding interaction types (PPI, iNCI, and PPI & iNCI, and no interaction). We find that the absolute co-expression weights ($\rho$ and $\omega$) are, on
Disease Comorbidity. Diseases with similar symptoms tend to share their disease network neighborhood (69), a feature captured by the network-based separation of two diseases, \( a \) and \( b \), defined as (4)

\[
S_{a,b} = \frac{<d_{a,b}^- > - <d_{a,b}^+ >}{2},
\]

where \(<d_{i,j}^- >\) is the average shortest distance from disease \( i \) to \( j \). A negative \( S_{a,b} \) indicates that two diseases are in overlapping network neighborhoods, while \( S_{a,b} \geq 0 \) implies that components associated with diseases \( a \) and \( b \) are in distinct network neighborhoods.

To illustrate how NCI can help improve our understanding of disease relationships, we focus first on RA and its comorbidities. If we limit the network to the PPI, RA has only one statistically overlapping disease, CD (Fig. 5 A and B—PPI). In contrast, in the extended PPI & NCI network, \( S_{a,b} \) uncovers statistically significant network-based overlap with CD, ulcerative colitis (UC), inflammatory bowel disease, inflammation, systemic lupus erythematosus (SLE), multiple sclerosis (MS), diabetes mellitus types 1 and 2, asthma, heart failure, stroke, atherosclerosis, and multiple neoplasms (Fig. 5 A and B—PPI & NCI). Most of these diseases predicted to be in the same network neighborhood as RA have known comorbidities to RA (76), similarly to inflammatory bowel diseases (77) [e.g., UC and CD (78)]. Also, patients with RA have a higher risk of developing diabetes type I (79) due to

**Fig. 4.** Physically interacting genes are co-expressed. (A) Schematics of the complete PPI and NCI network. A noncoding RNA “X” interacts with three proteins “A,” “B,” and “C,” and protein “B” binds to protein “D.” (B) The PPI network contain only interactions among proteins, and all interactions containing NC interactions are absent. (C) The induced NCI (NCI) contains indirect interactions. ncRNA “X” interacts with genes “A” and “B”; therefore, they are co-regulated by the same ncRNA, leading to an indirect interaction. In the same fashion, proteins “A” and “C” and “B” and “C” are also co-regulated by the same ncRNA “X,” inducing a triangle among them. (D) PPI & NCI includes direct and indirect interactions. Combining the interactions identified in (B and C). (E) 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. The boxplot indicates that the absolute Pearson correlation is higher when there is physical interaction, compared to then nonexistent links, indicating an association between physical binding and strength of co-expression. (F) Co-expression networks can predict physical interactions. We use the correlation 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 & NCI.
Fig. 5. Disease similarity on the two networks. (A) RA disease similarities and possible comorbidities. Two diseases are connected if they have a negative separation (meaning that they co-exist in the same network neighborhood) and if their overlap is significant. RA Disease–Disease network (LCC $P < 0.05$, $S_{a,b} < 0$, $S_{a,b} P < 0.05$, Jaccard index $> 0.05$, hypergeometric test $P < 0.05$; all $P$ values are FDR corrected) is shown for both the PPI and the PPI & NCI. Diseases are colored according to literature references for comorbidity. We find in the PPI that only CD is possible comorbidity, while the PPI & NCI predicts several other comorbid diseases, and most of them have a reported RR greater than 1, indicating comorbidity, few have RR smaller than 1, indicating a protective effect from those diseases. (B) RA comorbidities are also connected. Diseases that have high comorbidity with RA are also connected. Similar to A, we find an expansion of disease associations in the PPI & NCI. The PPI alone identifies two distinct clusters: A neoplasm and an inflammation module. The inclusion of ncRNA into the PPI helps us identify how those two clusters (neoplasm and inflammation) are also interconnected with each other. (C) Complete map of disease–disease relationships. We show the complete disease–disease network, unveiling the comorbidity map between 466 diseases. Each disease with a significant module (full dots) has its node size representing the size of the disease module, and the link width is relative to the normalized absolute $S_{a,b}$ value. Note that the PPI & NCI network forms a connected component with all the 213 diseases that have a significant LCC, moreover, we see that neoplasm (represented in blue) are close and form a module. The PPI, in its turn, forms a connected component with only 100 diseases, and the combination of both gives us a component that includes 249 diseases.
insulin resistance, and patients with RA often develop diabetes mellitus (80); additionally, patients with asthma have higher risk of developing RA (81, 82). Moreover, patients with RA have reported higher rates of heart failure (83, 84), such as myocardial infarction (83) and stroke (84) in addition to atherosclerosis (85). Higher risk of multiple neoplasms has also been reported in patients with RA treated with anti-TNF drugs (86, 87), interestingly, even though multiple myeloma and colonic neoplasms (88) share the same network neighborhood, they have reported decreased risk in patients with RA (89), suggesting that two diseases in the same network vicinity might also grant protection from each other. Taken together, we find that while the PPI can identify only one comorbidity for RA, most of the clinically documented comorbitides can be detected in the joint PPI & NCI network, indicating that the inclusion of ncRNA interactions is necessary to reveal disease comorbidities.

As a second case study, we focus on PE, which lacks a disease module in the PPI; hence, we could not predict comorbidities based on the PPI alone. By using the PPI & NCI combination, the top 10 closest diseases with significant topological overlap with PE are atherosclerosis, uterine cervical neoplasms, osteosarcoma, pancreatic neoplasms, glioma, cholangiocarcinoma, multiple myeloma, triple-negative breast neoplasms, heart failure, and carcinoma pancreatic ductal. Indeed, there is clinical evidence that women with PE are at increased risk for atherosclerosis (90) and other cardiovascular diseases such as heart failure (91), women with HPV infection, the main cause for uterine cervical neoplasms, also have increased risks for PE (92), and PE has been associated with an increased risk for several types of cancer (93). In other words, the joint PPI & NCI can accurately predict the known comorbidities of PE.

Finally, we expand our investigation by mapping out the disease–disease relationships that capture the network proximity of all disease pairs. The PPI predicts 543 comorbidity links between 350 diseases, revealing distinct clusters for neoplasm, cardiovascular, and gastrointestinal disease (Fig. 5C—PPI). In the combined PPI & NCI, we find 2,659 pairwise disease links between 466 diseases (Fig. 5C—PPI & NCI). We find that by including ncRNAs into the PPI, we retrieve more complete and biologically more meaningful list of disease–disease interactions, offering a better quantitative understanding of disease similarity and comorbidity, ultimately helping us understand disease progression in patients (3).

The NCI Predicts Relative Risk (RR). The PPI-based disease–disease separation is a known predictor for disease comorbidity (4), prompting us to ask whether the inclusion of ncRNAs improves not only our ability to detect clinically documented comorbidity in patients but can also help quantify comorbidity by predicting the RR between diseases. We measured the pairwise RR between all disease-pairs, relying on the disease history extracted from 13,039,018 elderly Americans enrolled in Medicare (94). RR estimates the strength of the association between two diseases, so that RR > 1 represents a risk factor, while RR < 1 indicates a protective factor. For example, UC affects 26,432 patients and Crohn’s disease (CD), 24,936. If the two diseases were independent of each other, we would expect only 45 individuals with both diseases. In contrast, we find 1,462 individuals with both UC and CD in our database. In other words, the chance of a patient developing CD is 30.55 times higher (29.0, 32.16; Wald interval, 95% confidence) in a patient with UC compared to a patient who does not have a history of UC, meaning that UC is a risk factor for CD.

Next, we investigate if diseases with a network overlap ($S_{ab} < 0$) have a higher RR, meaning that diseases that are located in the same network neighborhood have a higher risk of being comorbid. We find that negative $S_{ab}$ disease pairs have a statistically higher RR than the ones with positive $S_{ab}$ in both the PPI and the PPI & NCI (Material and Methods and SI Appendix, Fig. S9A). The RR in the PPI is on average 8.6 (se 2.98) for diseases with an $S_{ab} < 0$ ($P < 0.05$), and 6.74 (se 0.41) for diseases with an $S_{ab} > 0$ ($P < 0.05$). For the PPI & NCI, we observe an increase in the RR average to 9.5 for negative $S_{ab}$ (se 2.93), and a decrease to 6.5 (se 0.41) for positive $S_{ab}$ (SI Appendix, Fig. S9A). The increase of the RR for closer diseases in the PPI & NCI networks indicates, once again, that the inclusion of ncRNAs enhances our ability to quantify comorbidity by predicting the RR for patients, offering a better understanding of the network-based roots of disease progression.

Discussion

Network medicine, with its promise to offer a better mechanistic understanding of diseases (4), their progression (3), comorbidities (4), similarities (4), and treatments, such as drug repurposing (6, 7) and drug combinations (8), traditionally relied on PPI, capturing binding interactions between proteins. Two decades after the Human Genome Project, there is overwhelming evidence that noncoding genes and ncRNAs regulate multiple biological processes and functions, playing important roles in multiple diseases, and hence must be incorporated into the network medicine framework. Here, we show that the inclusion of ncRNAs into the PPI significantly improves the breath and the predictive power of network medicine. Protein-coding and noncoding genes are intertwined into a densely connected network, hence the inclusion of ncRNAs improves disease module identification and our ability to uncover disease–disease relationships, more accurately predicting the RR for patients. We also show that the rLCC increases when we incorporate the ncRNAs, helping us retrieve more complete disease modules, and offering biologically more interpretable identification of the molecular components contributing to a disease. We find that several neoplasms share the same noncoding neighborhoods, confirming the role ncRNAs play in neoplasm regulation. We also find improved comorbidities for many other diseases after the inclusion of ncRNAs, suggesting that noncoding elements contribute to most disease mechanisms. Disease modules are tissue-specific and are only expressed if the respective disease-genes are expressed as a connected component (95). In our study, we do not explore the tissue specificity of the specific diseases, but rather, consider the completeness of the disease modules in a tissue-independent network. Disease-focused studies should filter the PPI and the PPI & NCI to contain only genes expressed in a particular tissue, cell line, or even disease.

The more accurate disease module detection enabled by the inclusion of ncRNAs can lead to the development and identification of novel drug–targets, that hit closer to the disease module. They also raise the possibility that for some diseases targeting ncRNAs in the disease module may have a better therapeutic potential than targeting proteins. The clinical relevance of such intervention is illustrated by Bevasiranib (96), a siRNA that targets the VEGF-A gene, currently in clinical trial for treating macular degeneration and diabetic retinopathy, or Inclisiran, an LDL cholesterol lowering siRNA that targets PCSK9, the first approved ncRNA-based drug (97, 98). We also found an improved comorbidity prediction when considering ncRNAs, suggesting that the systematic inclusion of noncoding elements can offer a better understanding of disease progression, potentially opening a path toward precision medicine.
Materials and Methods

GDAs. We surveyed around 130 databases with GDA and selected those that i) were not compiled from other data sources and ii) provided at least one kind of evidence type classified as: Strong (functional evidence using an experimental essay); Weak (GWAS evidence but no experimental validation); Inferred (relying on bioinformatics or SNPs from imputation in GWAS); not compatible ([lncRNA, miRNA and other transcripts with or without experimental validation]). For each database we kept the disease name, gene converted to HGNC names (HUGO Gene Nomenclature Committee), and evidence level. At the end, we combined the following data sources: GWAS from ClinGen, ClinVar, CTD, Disease Enhancer, DisGeNET, GWAS Catalog, HMDD (58), IncBook, LncRNA disease, LOVD, Monarch, OMIM, Orphanet, PheGeni, and PsyGeNet (SI Appendix, section 2.2).

We searched for datasets that provide noncoding interactions derived from experimental evidence (ncRNA vs. proteins or ncRNA vs. ncRNA). We kept only databases that provided experimental evidence for their binding interactions and removed all interactions without strong evidence (SI Appendix, section 2). To validate the collected interactions, we assessed each database’s interactions and gene’s overlap.

Combining Data. All disease names were converted into MeSH terms after a word2vec embedding and all gene IDs into Gene Symbols. Gene names were normalized to HGNC symbol using biomart(99), Gene Cards (100), and gene2en-sembi from NCBI. Genes were classified into coding, and noncoding (miRNA, ncRNA, etc.) based on their classification from Gene Cards. Coding genes were classified as TF according to Perdomo-Sabogal 2019 (101).

To normalize disease names, we first converted all strings to lowercase and kept only alphanumeric characters. We next removed diseases with keywords that represented measures (such as “body mass”, “volume”, “count”, “susceptibility”, etc.). The renamed diseases were combined into clusters based on their similarity to MeSH C or F terms and synonyms based on a word2vec embedding trained on PUBMED (102).

Disease names with a cosine distance lower than 0.8 to a MeSH term or a MeSH synonym were removed, and the term with higher similarity was selected. Diseases are classified into different classes based on the MeSH’s first level of classification. As some diseases may have multiple first-level classifications, we define their first level as a level with more single diseases in it. Let us take RA as an example, which is defined as “Immune System Diseases”, “Musculoskeletal Diseases” and “Skin and Connective Tissue Diseases”, to avoid counting RA three times, we classify it as one of the three classes. We do so by selecting the class with more diseases, meaning that, 69 diseases are classified as “Immune System Diseases”, 84 are Musculoskeletal Diseases and 78 are “Skin and Connective Tissue Diseases”, in the RA case, the class is therefore classified as Musculoskeletal Diseases. After disease and gene names normalization, we filtered for diseases that have at least five associated genes, with at least one strong, weak, or incompatible experimental evidence.

Network Medicine Tools. Disease modules were inferred from the gene–disease association curated database, by calculating the LCC of each disease and deriving its P-value from the density of 1,000 simulations of a permutation test. The P-values are computed based on the empirical distribution derived from selecting random genes in the network and calculating the LCC size. The gene set is of the same size as the original set of genes, following a uniform distribution, i.e., each gene has the same probability of being associated with a disease. We can assume a uniform distribution here because genes are associated with diseases a high throughput setup, such as GWAS, differential gene expression, epigenetic, etc.; in other words, genes are not pre-filtered for the association study. Results using both degree-preserving and nondegree-preserving randomization are presented in SI Appendix.

Diseases separation was calculated using the measure proposed by Menche et al. (4). Separation significance is calculated by resampling 1,000 times the genes in each disease, and calculating the S_g, one-sided P-values are obtained as approximations of the area under the empirical density curve from −∞ to the found S_g.

To assess the disease similarity network, we selected overlapping elements given their disease separation (4) (S_g) < 0, significance of the S_g (permutation test, N = 1,000, P < 0.05), significance of the gene overlap (hypergeometric test, P-adj < 0.05, FDR corrected) and Jaccard index > 0 for each disease pair that forms a significant LCC using the two pre-defined networks (PP, and PPI & NCI). LCCs and S_g were estimated using the NetSoR package.

Drug-Targets and Gene Co-Expression. Drug-target interactions were retrieved from DrugBank (103) (version 5.1.9), keeping all target types (polypeptides, enzymes, carriers, and transporters). We filtered for drugs that have at least one drug-target described.

We analyzed gene expression in whole blood samples from GTEx (70) by selecting genes that are present in PPI & NCI database and have expression levels between the 10th and 90th quantiles. This approach enabled us to remove any outliers. We further filtered the GTEx data to focus on blood samples and retained only those genes that were expressed in at least 80% of the selected samples and had a SD greater than 0.01. Gene co-expression networks were inferred using the association between pairwise gene expression, measured using RNAseq or microarray, and the association is often derived from a correlation, such as Pearson correlation, or a transformation, such as the Weighted Topological Overlap (wTO) (71, 72). Gene expression was accessed using whole blood samples from GTEx (70), and the co-expression was constructed using Pearson Correlation and the wTO, from the wTO R Package (71), which calculates the co-expression between two genes based on their normalized correlation and further removes false positives.

Direct and Indirect Physical Interactions. We created an indirect-physical interaction network, which identifies the co-regulation of any two protein-coding genes by the same noncoding gene. For that, we constructed a bipartite network, based on the PPI & NCI, where one set of nodes represented noncoding genes and the other set represented the protein-coding genes. From this network, we created the projection on the protein–coding network, identifying if two protein-coding genes are targets of the same ncRNAs. We combine the original PPI with the indirect-physical interaction network.

Statistical Analysis and RR. Statistical analyses were performed using the R environment. Tests and CIs have been reported along with their P-values. All tests, unless stated, were FDR-corrected.

To access the RR of two diseases (comorbidities), we used the data from patients enrolled in MediCare (4, 94). To analyze the network of diseases represented by 3-digit codes from the International Classification of Diseases (ICD), we converted these codes into MeSH codes using the procedure outlined in Material and Methods: Combining Data. Since each MeSH code can correspond to multiple ICD codes, we combined patients with different ICD codes that mapped to the same MeSH code into the identified MeSH category. This enabled us to examine the data at a higher level of abstraction and identify patterns and connections between different disease categories. Relative risk and CI were assessed using the RRelRisk function from the DescTools R package, using the Wald test.

To access the difference between the RR and the S_ab, we select RR > 1 with statistical significance (P < 0.05, FDR corrected) and focus on diseases that affect at least 5% of the population, avoiding inflated RR. Differences between groups were assessed using the Mann–Whitney test.

Data, Materials, and Software Availability. The data and code utilized in the preparation of this manuscript are publicly accessible and can be obtained from the following sources: Code and Data Repository: The code and data used for data analysis and generating the results presented in this paper is available on GitHub at the following repository: https://github.com/Barabasi-Lab/NonCoding (104). Online Tool for Disease-Disease Exploration: An interactive online tool for exploring disease-disease relationships using both interactions is accessible at the following web address: https://deisgyii.shinyapps.io/Network_NCI/ (105).

Acknowledgments. We thank Italo F. do Valle and Xiao Gan for fruitful discussions. This research was supported in part by a NIH award 1P01HL132825, a Department of Veterans Affairs award Contract No. 36C2412000027, and by Scipher Inc. Agreement 21-01472. A.-L.B. is supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 810115 – DYNASNET.

Author affiliations: *Network Science Institute, Northeastern University, Boston, MA 02115; Department of Physics, Northeastern University, Boston, MA 02115; *Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA 02115; *US Department of Veteran Affairs, Boston, MA 02130; and *Department of Network and Data Science, Central European University, Budapest 1051, Hungary
43. C. H. Li, Y. Chen, Targeting long noncoding RNAs in brain development, functional diversification and