Published online 20 November 2017

Nucleic Acids Research, 2018, Vol. 46, No. 2 ***917–928***
doi: 10.1093/nar/gkx1169

Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs

Diogo M. Ribeiro[1], Andreas Zanzoni[1], Andrea Cipriano[2], Riccardo Delli Ponti[3,4], Lionel Spinelli[1], Monica Ballarino[2], Irene Bozzoni[2], Gian Gaetano Tartaglia[3,4,5,*] and Christine Brun[1,6,*]

[1]Aix-Marseille Université, Inserm, TAGC UMR_S1090, Marseille, France, [2]Dept. of Biology and Biotechnology Charles Darwin, Sapienza University, Rome, Italy, [3]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, 08003 Barcelona, Spain, [4]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain, [5]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluıs Companys, 08010 Barcelona, Spain and [6]CNRS, Marseille, France

Received October 06, 2017; Revised November 03, 2017; Editorial Decision November 07, 2017; Accepted November 07, 2017

ABSTRACT

The human transcriptome contains thousands of long non-coding RNAs (lncRNAs). Characterizing their function is a current challenge. An emerging concept is that lncRNAs serve as protein scaffolds, forming ribonucleoproteins and bringing proteins in proximity. However, only few scaffolding lncRNAs have been characterized and the prevalence of this function is unknown. Here, we propose the first computational approach aimed at predicting scaffolding lncRNAs at large scale. We predicted the largest human lncRNA–protein interaction network to date using the *cat*RAPID *omics* algorithm. In combination with tissue expression and statistical approaches, we identified 847 lncRNAs (∼5% of the long non-coding transcriptome) predicted to scaffold half of the known protein complexes and network modules. Lastly, we show that the association of certain lncRNAs to disease may involve their scaffolding ability. Overall, our results suggest for the first time that RNA-mediated scaffolding of protein complexes and modules may be a common mechanism in human cells.

INTRODUCTION

More than 60% of the human genome is transcribed into tens of thousands of RNAs with low coding potential (1). Long non-coding RNAs (lncRNAs) are a subset of those transcripts longer than 200 nt, transcribed by RNA polymerase II, often capped, spliced and polyadenylated (2). The possible function of most of the > 26 000 GENCODE annotated lncRNAs is yet to be addressed (3), and many are thought to be transcription errors or noise. However, thousands of lncRNAs have been found to be differentially expressed in distinct cell types, with dozens shown to be implicated in transcription regulation (4), stress responses (5) and disease (6). Indeed, lncRNAs are versatile molecules able to perform numerous tasks in the cell through binding of proteins, DNA or other RNA molecules (2).

All cellular functions are performed by interactions between molecules, such as interaction between proteins and RNAs. These interactions can be stable, leading to ribonucleoprotein (RNP) complexes such as the ribosome, the spliceosome or the telomerase complex, or transient such as those involved in transport and degradation of nuclear transcripts. Similarly, components of complexes or pathways need to be physically close to each other (either transiently or permanently) in order to perform their function. One way to achieve this, while attaining selectivity in a crowded cell, is to employ platform or scaffold molecules that piece together components of a complex or a pathway (7). Although proteins can and do serve as scaffolds for other proteins (8), the use of RNA scaffolds would present several advantages, since 'one protein comprising 100 amino acids can capture only one or two proteins, whereas one RNA molecule comprising 100 nt can capture around 5–20 proteins', simultaneously (9). Moreover, lncRNAs can act immediately after transcription, while protein scaffolds require at least the step of translation before being functional (2).

Several ncRNAs have been found to function as scaffolds for RNP complexes such as TERC (Telomerase RNA Component), SRP (Signal Recognition Particle RNA) and LINP1 (LncRNA In Nonhomologous End Joining Pathway 1) (2,10,11) or found to transiently assemble groups of proteins as in the case of XIST (X-inactive specific transcript) and both the granule-forming NEAT1 (Nuclear Paraspeckle Assembly Transcript 1) and MALAT1

*To whom correspondence should be addressed. Tel: +33 491828712; Email: christine-g.brun@inserm.fr
Correspondence may also be addressed to Gian Gaetano Tartaglia. Tel: +34 933160116; Email: gian.tartaglia@crg.eu

(Metastasis Associated Lung Adenocarcinoma Transcript 1) (5,12). Although known scaffolding lncRNAs carry out important cellular functions, only a few dozen cases have been uncovered so far (7), many while studying the protein complexes rather than the lncRNAs. We therefore hypothesize that other yet uncharacterized lncRNAs may act as scaffolds.

Recently, with the development of RNA interactome capture methodologies, the repertoire of RNA-binding proteins (RBPs) has greatly expanded (13), leading to the discovery of hundreds of novel RNA-interacting proteins, many of which contain no known RNA-binding domain (RBD). In addition, studies using high-throughput methods to detect RNAs bound by RBPs including iCLIP, PAR-CLIP and recently eCLIP (14), demonstrate that most RBPs bind thousands of different RNA molecules depending on the cell line. However, these investigations have been limited to a set of ∼140 RBPs containing known RBDs (14,15) and do not cover the full extent of the protein–RNA interaction space. Furthermore, only one fraction of the RNAs targeted by the RBPs are found in common by independent replicate experiments, suggesting that the interaction maps of the studied RBPs are far from complete (14). Computational prediction of protein–RNA interactions can therefore help fill the gap in our knowledge of protein–RNA interactions and be applied to large-scale analyses.

In this paper, we study for the first time the prevalence of protein complex scaffolding as a function of lncRNAs. By exploiting a computed protein–RNA interaction network, we developed and applied an original large-scale approach to identify candidate lncRNAs possibly acting as scaffolding molecules for protein complexes and network functional modules. We discovered hundreds of scaffolding lncRNA candidates, suggesting that RNA scaffolding is a prevalent and widespread mechanism in the cell. In addition, we found that more than half of the protein complexes and network modules in the cell may be scaffolded by lncRNAs, reinforcing the widespread nature of their action.

MATERIALS AND METHODS

LncRNA–protein interaction predictions

The *cat*RAPID *omics* protein–RNA interaction predictor (16) was used to predict interactions between the human long non-coding RNA transcriptome (Ensembl v82) and the human canonical proteome, leading to ∼243 million predictions. Predictions with interaction propensity score ≥50 were kept for further analyses (∼30.8 million interactions). See Supplementary Material for details.

Tissue expression filtering

To create a set of high confidence protein–RNA interaction predictions, we restricted the analysis to pairs of lncRNA–proteins that are likely to be found together in at least one tissue. Human tissue expression data from the GTEx v6.0 project (17) was used. We downloaded RPKM (Reads Per Kilobase of transcript per Million mapped reads) information from 8555 samples across 53 tissues, already mapped to human transcripts (GENCODE v19). RPKM values of samples coming from the same tissue were averaged after a step of removing outlier values (below or above 1.5-times the interquartile range). Protein expression was derived from their coding mRNA expression, by selecting the highest RPKM value among the protein's mRNAs for each tissue. Only protein–RNA interactions where both the RNA and the protein have a minimum RPKM value of 1.58 in at least one of the 53 tissues, were retained. This cutoff was determined as the optimal expression cutoff (maximizing the sum of specificity and sensitivity) in a ROC curve experiment between the pre-filtering lncRNA–protein interaction prediction dataset (∼243 million interaction predictions) and a set of 2438 experimentally detected CLIP interactions taken from StarBase v2.0 (18) with at least 100 mapped reads (area under the ROC = 0.71). The expression metric used ('paired expression') was calculated for each protein–RNA pair as the lowest RPKM expression between the protein and RNA for each tissue, to which the maximum RPKM value among tissues for that protein–RNA pair is then withdrawn, i.e.

$$E\ (\text{Protein},\ \text{RNA}) = \max_{t\ \in\ \text{tissues}} \left(\min\left(E_t\,(\text{Protein}),\ E_t\,(\text{RNA})\right)\right)$$

where E(Protein, RNA) denotes the 'paired expression' for each protein–RNA pair and E_t denotes the RPKM expression in tissue t (RPKM values were $\log_{10}$-transformed).

Protein complex and network module datasets

We collected protein complex information from the (i) BioPlex publication (19) Supplementary Table S3, which includes 354 complexes; (ii) list of conserved protein complexes from Wan *et al.* (20), Supplementary Table S4, which includes 981 complexes; (iii) list of non-redundant CORUM (21) complexes from Havugimana *et al.* (22), Supplementary Table S3, which includes 324 complexes, referred to as 'non-redundant CORUM complexes'. Protein network modules were extracted from a human interactome as described in (23). See Supplementary Material for details.

LncRNA–protein complex enrichment analysis

Using the set of predicted interactions between lncRNA and proteins filtered by (i) interaction propensity and (ii) minimum RPKM expression, we performed the following enrichment analysis: for each lncRNA and protein group, we assessed the enrichment of the lncRNA's interacting-proteins among the proteins in the group using a hypergeometric test (one-tailed test; FDR = 5%, multiple test corrected with the Benjamini–Hochberg procedure; Figure 1B), using as background the set of proteins in complexes or modules retaining at least one interaction after interaction filters. We considered only enrichments where: (i) the lncRNA is interacting with at least two proteins of the protein group and (ii) all the proteins in the complex or the module are expressed in a same tissue as the lncRNA with at least 1.58 RPKM. To exclude lncRNAs with high background levels of enrichments, we built a null hypothesis distribution by performing 10 000 hypergeometric tests for each lncRNA, each time randomly shuffling the proteins labels between the protein groups. We excluded lncRNAs

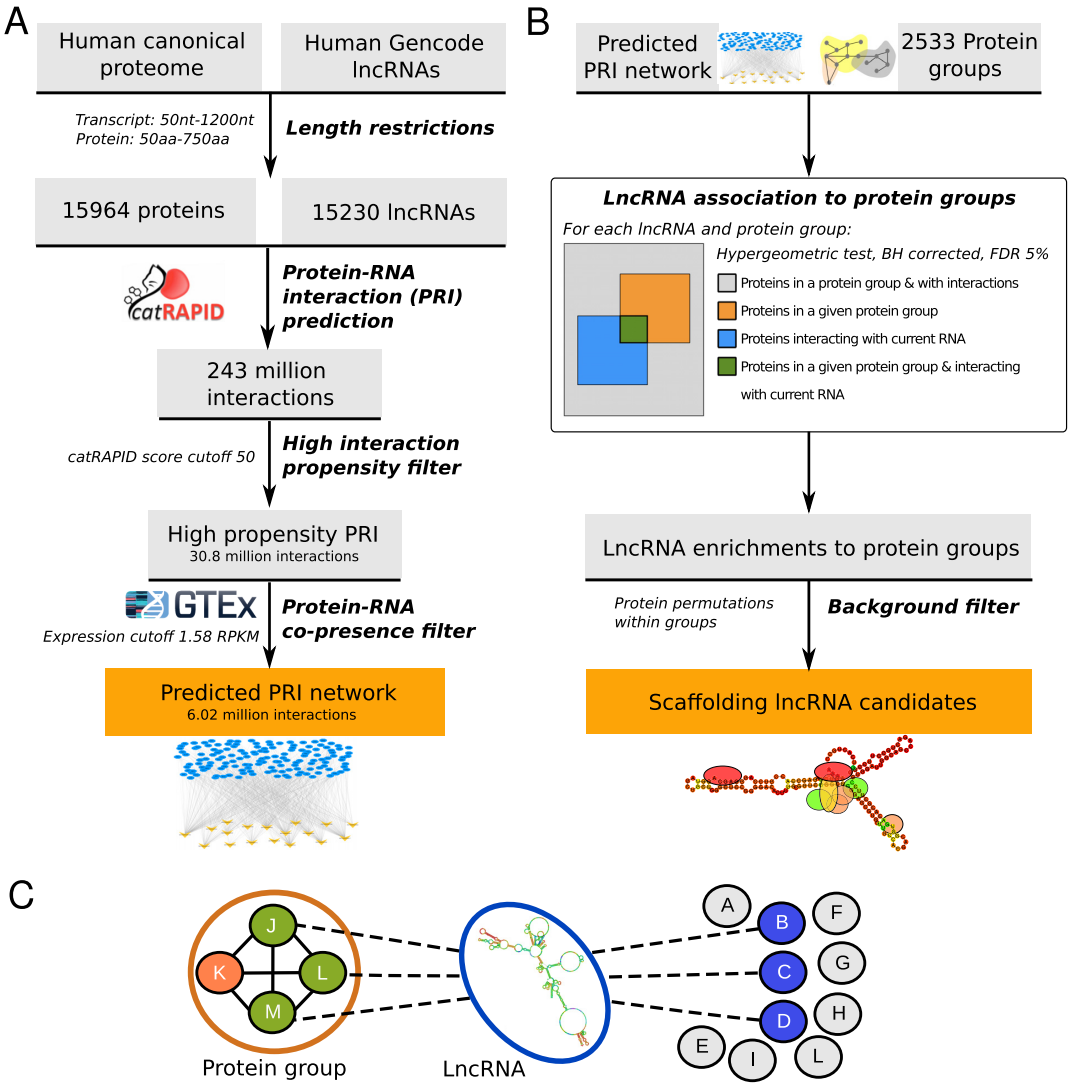

Figure 1. Data production and analysis workflows. (**A**) Predictions of protein-lncRNA interactions (PRI) using *cat*RAPID *omics* for the human proteome and long non-coding transcriptome. Interactions are further filtered by co-presence in the same GTEx tissue. The produced PRI network contains 6.02 million interactions. (**B**) Protein groups and lncRNAs are tested for enrichment in lncRNA protein's targets among groups of proteins. After noise filtering, a final list of scaffolding lncRNA candidates is produced. (**C**) Principle of the enrichment in lncRNA protein's targets among groups of proteins. Colors of nodes correspond to the ones used on the lncRNA association to protein groups box on (B).

with (i) enrichments not significant in respect to the null hypothesis (empirical P-value > 0.01); (ii) an enrichment ratio lower than 2-fold.

RESULTS

A predicted human interaction network between the non-coding transcriptome and the proteome

Aiming to extensively identify lncRNA molecules interacting with protein complexes and potentially acting as protein scaffolds, we first computed the protein–RNA interaction potential between most of the human proteome and the long non-coding transcriptome (79% and 81%, respectively; Supplementary Material) using the *cat*RAPID *omics* algorithm (16) (Figure 1A). The *cat*RAPID algorithm is a protein–RNA interaction predictor based on the physicochemical features of the molecules that has been extensively used and tested on lncRNAs with good performances (16,24,25). With this method we produced 243 million predicted interactions, of which 30.8 million display high interaction propensity scores (*cat*RAPID score $\geq$ 50). Since many lncRNAs have only been found to be expressed at very low levels and often in a tissue-specific manner (26), we only retained 6.02 million protein–lncRNA interactions between molecules co-present in at least one out of the 53 human tissues from the GTEx RNA-seq dataset (17) (see Materials and Methods). Globally, the 6.02 million predicted interactions occur between 12629 proteins and 2799 lncRNAs (Figure 2), i.e. between 80% of the tested proteins and 18% of our initial set of lncRNAs. Individual proteins are predicted to interact with up to 2.5% of the lncRNAs on average (Supplementary Figure S1). When considering only RBPs (Supplementary Material), we predict them to interact with 4.14% of the lncRNAs on average, in the same

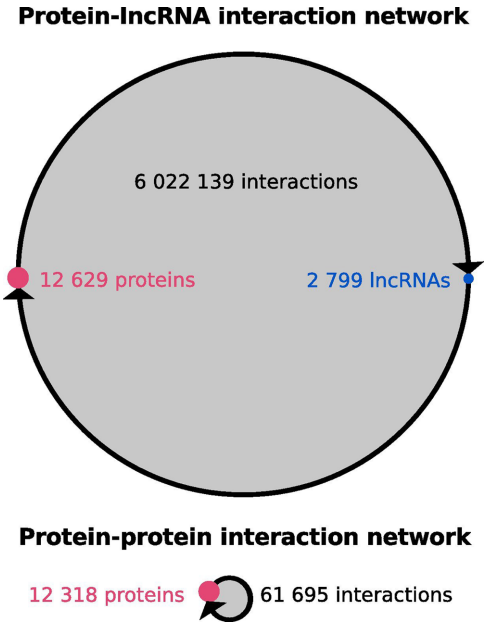

Figure 2. A global lncRNA–protein interaction network. Predicted protein-lncRNA interaction network composed by more than 6 millions interactions (grey circle) between 12629 proteins (pink circle) and 2799 lncRNAs (blue circle). The size of the network is compared to the human binary protein-protein interaction network (see Supplementary Methods). All circles are proportional to their components.

range as eCLIP results on 82 RBPs (14), which interact with 7.98% of the lncRNAs from the same dataset. On the lncRNA side, their median number of protein interactions is 1267 (Supplementary Figure S1), a higher number than suggested by current RNA pull-down studies that report between 126 and 852 interacting proteins per lncRNA (27,28).

As evident in recent high-throughput screenings, the complexity of biological systems challenges interpretation of experimental results due to the specific interactions occurring in different contexts (12,14). Yet our predictions, based on molecular physicochemical properties, represent a set of possible interactions between co-expressed proteins and RNA, independent of the cellular sub-localization and the cellular states. Our predictions therefore cover a larger spectrum of conditions in which protein–RNA interactions may occur, compared to the ones assessed in specific *in vivo* studies. This allows us to detect, for example, lncRNAs acting exclusively upon DNA damage and other stress conditions, or interactions restricted to a few cell types. Despite all this, 9414 of our predicted interactions are found in the relatively small set of eCLIP experiments, a highly significant overlap considering the 82 proteins and 7381 transcripts present in both eCLIP and *cat*RAPID datasets (P-value < 2.2e-271, OR = 1.85, two-tailed Fisher's exact test), therefore increasing our confidence in the predicted network.

Overall, to the best of our knowledge, we have predicted the largest human lncRNA–protein interaction network to date.

Interactions between lncRNAs and protein complexes or network modules

To assess our capacity to computationally predict lncRNA interactions with protein complexes, we studied the possible association between a recently discovered evolutionarily-conserved and muscle-restricted lncRNA, *lnc-405* (29), and the Puro–Purβ–YBX1 protein complex, implicated in gene regulation of muscle cells (30). The *cat*RAPID *omics* algorithm predicts the interaction of human *lnc-405* with Puro, Purβ and YBX1 with moderate to high scores (38.56, 44.05, 67.84, respectively).

To determine if *cat*RAPID correctly predicted the interactions of the lncRNA to the protein complex in a cellular context, we performed endogenous *lnc-405* RNA pull-down from nuclear extracts of C2C12 mouse myotubes followed by a mass spectrometry (MS) analysis. Murine cells were used since *lnc-405* is highly conserved in mouse and very abundant in differentiated C2C12 cells, allowing the easy production of the large amounts of nuclear extracts which are required for the pull-down. Efficient enrichment of *lnc-405* was detected in both odd and even RNA pull-down samples, while no recovery was observed with lacZ control (Supplementary Figure S2A).

Notably, MS analysis applied on the odd, even and lacZ (control) samples allowed the identification of 19 *lnc-405* interactors, including two components of the Puro-Purβ-YBX1 complex (Supplementary Table S1; Supplementary Material). RIP assays performed in mouse and human myotubes, using an antibody against Purβ, allowed to validate the specificity of the interaction with *lnc-405* and to confirm the evolutionary conservation of such interaction (Supplementary Figure S2B and C). Moreover, a GSEA experiment shows that the top interactors of *lnc-405* predicted by *cat*RAPID are enriched in proteins identified in the MS experiment (Figure 3, P-value = 0.017). These results remarkably show that *cat*RAPID is able to correctly predict interactions between lncRNAs and proteins (whether in a complex or not), in line with good *cat*RAPID performances observed for other ncRNAs and reported in previous articles (24,25,31).

We thus proceeded with the exploration of our *cat*RAPID predicted lncRNA–protein interaction network, aiming to test the hypothesis that lncRNAs frequently scaffold known protein complexes through protein–RNA interaction. For this, we investigated three public datasets of human macromolecular complexes. Briefly, we used the (i) non-redundant dataset of 326 CORUM complexes (21) collected by Havugimana *et al.* (22) (hereafter referred to as 'non-redundant CORUM'), (ii) a set of 981 metazoan-conserved complexes produced by Wan *et al.* (20) through biochemical fractionation with quantitative MS (hereafter referred to as 'Wan 2015'), as well as (iii) the BioPlex dataset (19) of 354 complexes detected through affinity purification, MS experiments and interaction network analysis. Moreover, the human cell contains groups of functionally-related proteins that interact more transiently but may nevertheless be assembled or gathered together by lncRNA scaffolds to participate in metabolic or signaling pathways. For these reasons, we also used a dataset of 874 functional modules identified

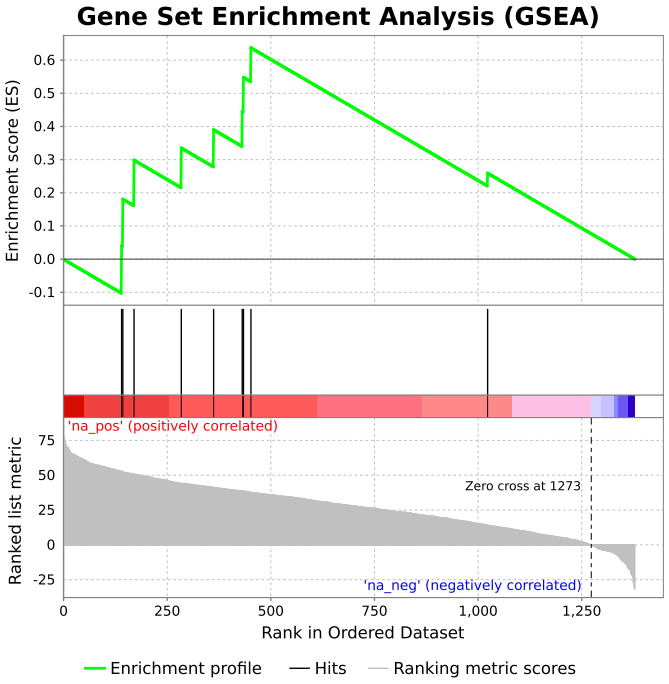

Figure 3. Experimentally-determined *lnc-405*-interacting proteins are enriched as top *cat*RAPID predictions. Gene Set Enrichment Analysis (GSEA) (69) of *cat*RAPID predictions between *lnc-405* lncRNA and 1459 human RBPs (Supplementary Material), using the RBPs identified as interactors in the MS experiment as a gene set. Note that only RBPs with *cat*RAPID predictions (within size restrictions) were considered. *P*-value = 0.017 (10 000 simulations), normalized enrichment score = 1.59).

in the human interactome using OCG, an algorithm that decomposes a network into overlapping modules, based on modularity optimization (32). These modules (hereafter referred to as 'Network modules') are groups of highly interacting proteins, which tend to be involved in the same cellular processes, metabolic or signalling pathways (33) (Supplementary Table S2).

Using these datasets of protein groups and our protein-lncRNA interaction predictions, we identified lncRNAs that may scaffold complexes or modules by assessing first, for each lncRNA, the enrichment of the lncRNA's interacting proteins among those proteins composing each complex or network module (hypergeometric test, Benjamini-Hochberg corrected FDR 5%) (Figure 1B). Second, because some lncRNAs are predicted to bind a large number of proteins, we estimated the number of protein groups we would expect to find enriched by chance for each lncRNA, as a control, by shuffling the protein labels between protein groups (10 000 times). Only lncRNAs predicted to bind significantly more (empirical *P*-value < 0.01), and at least twice as many, protein groups than expected by chance were considered candidates for scaffolding function.

After filtering using the randomised control, we obtained a total of 27 090 statistically significant enrichments between 1517 protein groups and 847 distinct lncRNA transcripts, encoded by 820 lncRNA genes (Supplementary Table S3). These 847 lncRNAs, ∼5% of our 15 230 tested transcripts, are hereafter referred to as 'scaffolding lncRNA candidates' and constitute a set of lncRNAs predicted to be involved in a scaffolding function (Supplementary Table S4). Remarkably, we also predict that ∼56% of the known protein complexes and 66% of the network modules are scaffolded by at least one lncRNA (Supplementary Table S3). These results suggest that lncRNAs scaffolding complexes and modules are highly prevalent. Moreover, as the set of predicted complexes and modules found to be scaffolded by lncRNAs are involved in most cellular biological processes (Supplementary Figure S3), the scaffolding function of lncRNAs appears therefore to be a general feature and not restricted to specific cellular processes.

Although current experimental protein-lncRNA interaction datasets are largely incomplete and limited to 148 RBPs (14,15,18), we find that 832 out of 6186 lncRNA–protein-group interactions including at least one of the 148 RBPs contain one or more known experimental interactions (Supplementary Table S4). Importantly, as a control, when restricting our scaffolding lncRNA candidate detection method to protein–RNA interactions involving only RNA-binding proteins (1459 RBPs; Supplementary Material), instead of the whole proteome, we identify 788 scaffolding lncRNA candidates among which 572 (72.5%) were also found by our proteome-wide approach. This highly significant overlap (*P*-value $< 2.2e{-}16$, OR = 158, Fisher's exact test; Supplementary Figure S4) reinforces the confidence of our predictions.

Overall, our large-scale approach predicted tens of thousands of lncRNA–protein-group interactions between hundreds of lncRNAs and protein groups, many of which containing experimentally determined interactions, suggesting an abundant presence of lncRNA scaffolds.

Global analysis of scaffolded complexes and modules

In order to analyse the patterns of predicted interactions between lncRNAs and protein groups, we represent them as a clustered matrix (Figure 4). Clusters of protein groups with similar enrichment profiles often share proteins, while clusters of lncRNAs with similar enrichment profiles are largely composed of transcript isoforms from the same or paralog genes. While some protein groups and lncRNAs interact specifically, others—protein groups as well as lncRNAs—do so more promiscuously, and this occurs for each of the four protein group datasets used. Indeed, we observe that some lncRNAs are predicted to interact with 1 to 98 protein groups, according to the dataset, i.e. at most 54 (16.7% of total) non-redundant CORUM complexes, 35 (9.9%) in BioPlex, 98 (10.0%) in Wan 2015, 68 (7.8%) in network modules (Figure 4; Supplementary Figure S5A). Likewise, protein complexes are predicted to interact with 1 to 401 lncRNAs i.e. at most 401 lncRNAs (2.6% of total tested) in non-redundant CORUM, 17 (0.1%) in BioPlex, 248 (1.6%) in Wan 2015, 115 (0.7%) in network modules (Figure 4; Supplementary Figure S5B).

Interestingly, some of our predictions corroborate and further extend the current knowledge of protein–RNA complexes. For instance, the polycomb repressive complex 2 (PRC2 complex), previously found associated with lncRNAs (4), is predicted to be scaffolded by 101 different lncRNAs in our analysis. Indeed, the PRC2 complex and some of its constituent proteins have previously been found to bind

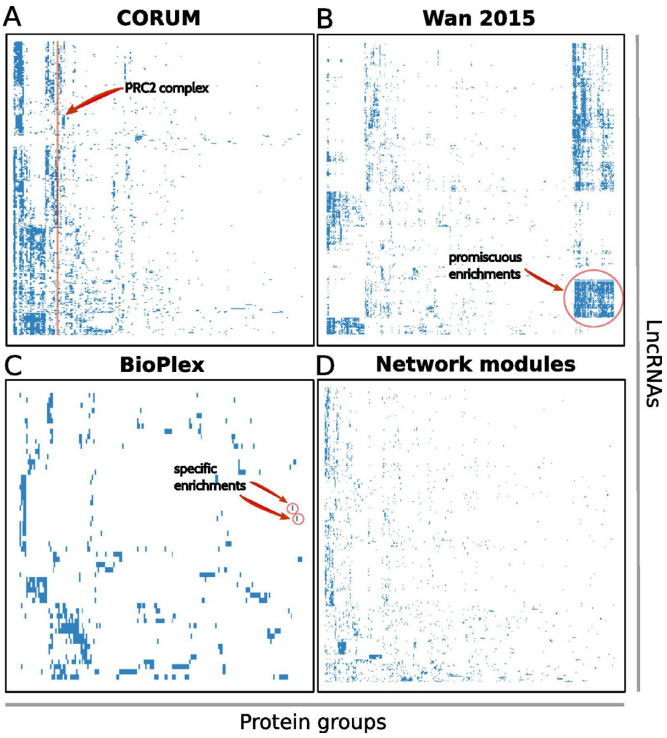

Figure 4. Interactions between lncRNAs and protein groups. Boolean matrix representing enrichment between lncRNAs and protein groups on (**A**) non-redundant CORUM, (**B**) Wan 2015, (**C**) BioPlex and (**D**) Network modules. Blue color represents significant enrichments, white color represents non-significant enrichments. Only lncRNAs/protein groups with at least one significant enrichment are displayed. Matrix was clustered by hierarchical clustering with euclidean distance, dendrograms are not displayed due to the very high number of rows and columns. The PRC2 complex, as well as examples of promiscuous and specific enrichments are highlighted.

hundreds of lncRNAs, presumably as a part of its targeted gene repression mechanism or its regulation by decoy lncRNAs (4,34).

Overall, we find that some lncRNA candidates may act as general scaffolds for several protein groups, while others are specific to one or a few protein groups. Likewise, some protein groups are predicted to interact with many different lncRNAs, perhaps reflecting their function, exemplified by the PRC2 complex.

Scaffolding lncRNA candidates display functional features

To determine if our scaffolding lncRNA candidates are likely to be functional, we gathered several orthogonal datasets of lncRNAs displaying functional features. Together these include lncRNAs (i) displaying a metabolism profile characteristic of functional transcripts (35), (ii) overlapping eQTLs (36); (iii) that alter cell-growth when subjected to inactivation by CRISPRi (37); (iv) involved in disease (38,39), as well as lncRNAs (v) conserved in tetrapods (40) or (vi) possessing structurally conserved elements (41). Strikingly, even though these functional lncRNAs have been found to act not only through protein-binding but also RNA- and DNA-binding, many were successfully identified by our protein–RNA interaction-based approach (Figure 5A). Indeed, we observe a significant (P-value < 0.05, one-tailed Fisher's exact test) and often strong overlap (OR $>$ 2) between our scaffolding lncRNA candidate dataset and every functional or conserved lncRNA dataset analysed except therian-conserved lncRNAs. This latter result suggests that most human scaffolding lncRNAs may have appeared later in evolution or may be highly species-specific.

Additionally, when considering the different sets of scaffolding lncRNA candidates identified using our four different protein group datasets separately, they are all found significantly enriched in functional or conserved lncRNAs from all tested orthogonal datasets (P-value < 0.05, OR from 1.73 to 1.96, one-tailed Fisher's exact test; Supplementary Figure S6A). Different pertinent lncRNA candidates can therefore be detected from each protein group dataset, consistent with the relatively low overlap observed between lncRNAs candidates found from each dataset (Supplementary Figure S6B).

In agreement with our findings, we observe that mutations in exons of scaffolding long non-coding intergenic RNA (lincRNAs) candidates have a higher predicted consequence on fitness than mutations in other lincRNAs, by measuring their fitCons scores (Figure 5B; Supplementary Material), a metric that takes into account sequence polymorphisms in human and sequence divergence in primates (42).

Altogether, these results suggest that our candidates generally possess the features of functional transcripts, therefore lending further weight to our predictions.

LncRNA-associated disease mechanisms could involve lncRNA scaffolding function

Hundreds of lncRNA genes have been associated with several human diseases and conditions including cancer, diabetes and neurodegenerative diseases. As most of these associations were identified through the analysis of lncRNA differential expression in disease states (38,39), knowledge on the molecular role of these lncRNAs in disease is lacking.

We have found 30 scaffolding lncRNA candidate genes associated with disease in lnc2cancer (38) and LncRNADisease (39) databases (Figure 5A; Supplementary Material). We then assessed whether these lncRNA-disease associations could occur through the predicted protein group scaffolding functions of the lncRNAs. For this, we mapped proteins involved in disease from the OMIM database (43) to protein groups and found that 15 out of 30 scaffolding lncRNA candidate genes associated with disease are possibly interacting with a protein group that includes at least one protein associated with the same or similar disease (Figure 6; Supplementary Table S5).

In several cases (e.g. lncRNA genes SNHG1, SOX2-CT and RP11–356I2.4), lncRNAs and diseases are linked through different protein complexes, and involving different proteins, which provides further evidence of the association.

For instance, the SNHG15 lncRNA gene has been associated to Hereditary Haemorrhagic Telangiectasia (HHT) (44), a disease known to be caused by mutations in genes that modulate the TGF-β superfamily (45). Here, we find that two of its transcripts possibly interact with a com-

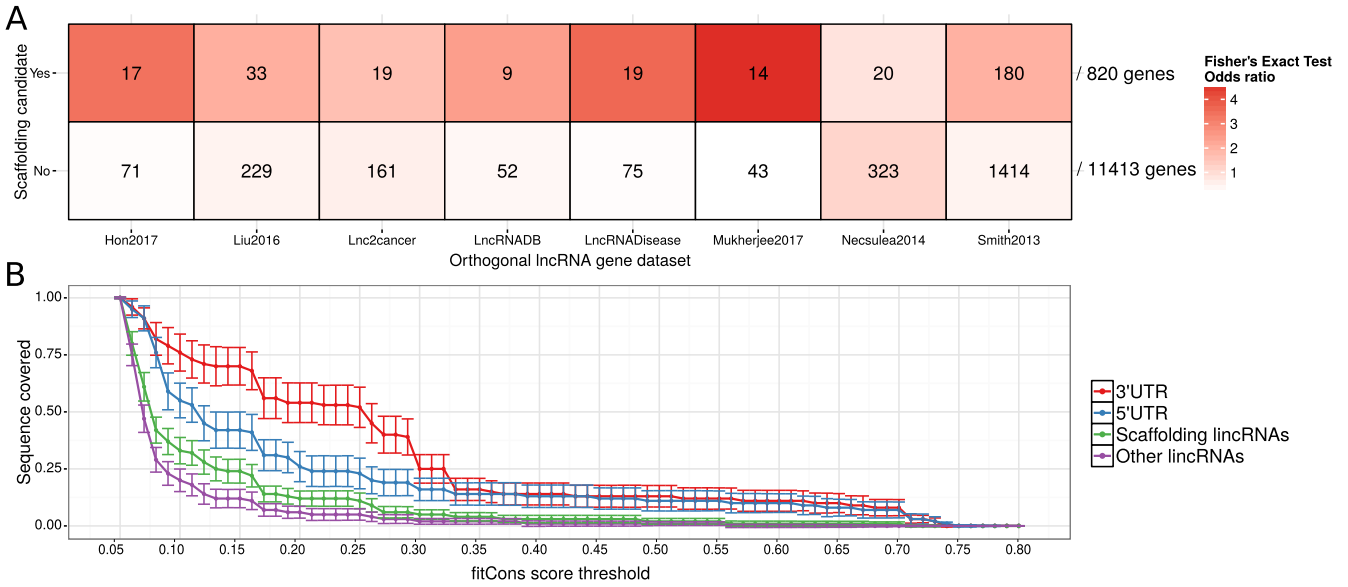

Figure 5. Scaffolding lncRNA candidates display functional features. (**A**) Overlap between scaffolding lncRNA candidate gene (820 genes, 847 transcripts) and the following groups of functional or conserved lncRNA genes characterized in other studies: Hon2017 (36): lncRNAs displaying four features of functionality; Liu2016 (37): lncRNAs affecting cell growth according to CRISPRi experiments; Lnc2cancer (38): lncRNAs involved in cancer; LncRNADB (70): compendium of known functional lncRNAs; LncRNADisease (39): lncRNAs involved in human diseases; Mukherjee2017 (35): lncRNAs with a metabolic profile characteristic of functional transcripts; Necsulea2014 (40): lncRNAs conserved in therians; Smith2013 (41): lncRNAs containing at least one exonic conserved structural element (see Supplementary Material). Enrichment was tested with one-tailed Fisher's exact tests, background included all genes (12233 lncRNA genes, 15230 transcripts) analysed in this study. All *P*-values for the 'Yes' category are significant (*P*-value < 0.05), except for Necsulea2014. (**B**) Proportion of sequence covered with fitCons score above the threshold (x-axis), for different gene features (3′UTR, 5′UTR), lincRNA exons on scaffolding lincRNAs candidates and all other lincRNAs accessed in this study. Error bars: standard deviation of 100 subsampling experiments (with replacement) of 50 genes per category. 'Scaffolding lincRNAs' have a higher proportion of sequence covered above the threshold than 'Other lincRNAs' (one-tailed Kolmogorov–Smirnov test *P*-value = 0.008). As observed in other studies (35,42), lincRNA fitCons scores are lower than UTR regions of protein-coding genes.

plex containing 11 components and regulators of the TGF-β pathway out of 23 proteins (ENST00000585030, non-redundant CORUM complex 81), and with a module composed of signalling proteins and transcription factors (ENST00000578968, network module 686, Supplementary Table S5). Notably, whereas these SNHG15-interacting protein groups are largely composed of different sets of proteins, both contain the SMAD4 protein, a TGF-β pathway component mutated in HHT (46). Overall, further credibility is given to an involvement of SNHG15 in this disease through its predicted scaffolding function.

Moreover, the MEG3 lncRNA gene has been linked to colorectal cancer (47), and has been shown to bind chromatin-remodeling complexes (4). Interestingly, we detected a short MEG3 lncRNA isoform (ENST00000524131, 721 nucleotides) possibly interacting with a complex containing DNA polymerase epsilon subunits as well as chromatin-remodeling proteins (Wan 2015 complex 79), including POLE1, also associated to colorectal cancer.

Finally, the SNHG1 gene is associated to hepatocellular carcinoma (HCC) (48) and non-small cell lung cancer (49). Here we find one of its transcripts (ENST00000539975) interacting with 18 different protein groups associated with one or both of those diseases. Moreover, the interaction of SNHG1 lncRNA with 6 of those protein groups is corroborated by experimental interactions (14,15,18) through five distinct RBPs. Several pathway components of the TNFα/NF-κB signaling pathway have been associated with both lung cancer and HCC, as well as other cancers (50,51). The SNHG1 lncRNA is predicted to interact with the TNFα/NF-κB signaling complex (non-redundant CORUM complex 10) through PAPOLA (poly(A) polymerase α) and CHUK (inhibitor of nuclear factor κ-B kinase subunit α). Notably, the lncRNA interaction with the protein complex is further corroborated by two experimental interactions with two RBPs of the complex, DDX3X and AKAP8L (Supplementary Table S5). Additionally, the SNHG1 lncRNA has been associated with HCC through suppression of miR-195 (52), a microRNA known to target the TNFα/NF-κB pathway by repressing the CHUK protein, and thus suppressing HCC (53) (Figure 7). Given our predictions, we can thus propose that beyond its known effect through miR-195, SNHG1 may regulate elements of the TNFα/NF-κB pathway and therefore directly affect HCC through its possible protein group scaffolding function.

Globally, we propose that the association of 15 lncRNA genes to 22 diseases is due to protein-lncRNA interaction-based mechanisms, notably through the scaffolding of protein complexes and modules by lncRNAs.

DISCUSSION

The current scarcity of experimentally determined lncRNA–protein interaction data hinders the investigation of lncRNA function at large-scale. We thus

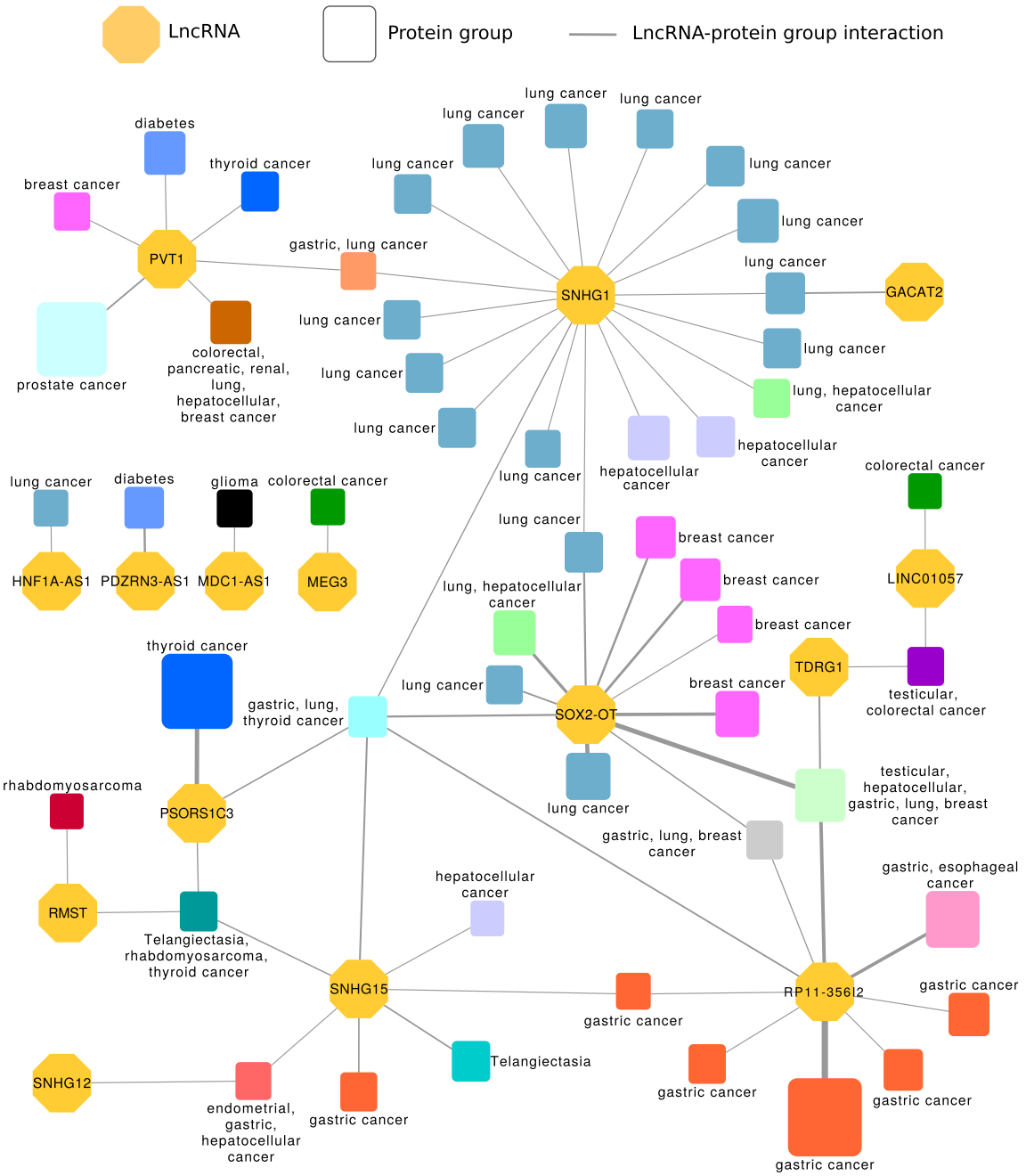

Figure 6. Disease-associated lncRNA network. Network representation of disease-associated lncRNAs (hexagonal nodes in yellow) potentially scaffolding protein groups (square colored nodes) containing at least one protein known to be involved in the same disease. Colors correspond to different diseases. Node size reflects the number of proteins in the group. Edges represent lncRNA–protein-group interactions. Edge width reflects the number of proteins interacting with the lncRNA. LncRNA transcripts were mapped to genes. Some disease names have been abbreviated for simplicity.

computationally predicted a comprehensive lncRNA–protein interaction network in order to better cover the lncRNA–protein interaction space. For this, we used *cat*RAPID, a protein–RNA interaction predictor based on the physicochemical features of the molecules, which can be used large-scale and has been initially validated on a large collection of protein associations with lncRNA (24). Indeed, *cat*RAPID performed well against the NPInter database (area under the receiver operating characteristic (ROC) of 0.88), as well as on the non-nucleic-acid-binding database (area under the ROC curve of 0.92) (31). In addition, we showed herein that *cat*RAPID predictions provide relevant information about lncRNA–protein-complex interactions by experimentally validating that part of the Purα-Purβ-YBX1 complex — predicted here to interact with the *lnc-405* lncRNA — effectively binds the lncRNA *in vivo*.

Noticeably, as the *cat*RAPID predicted interaction network contains the set of biophysically possible interactions between co-expressed molecules, which may differ from in-

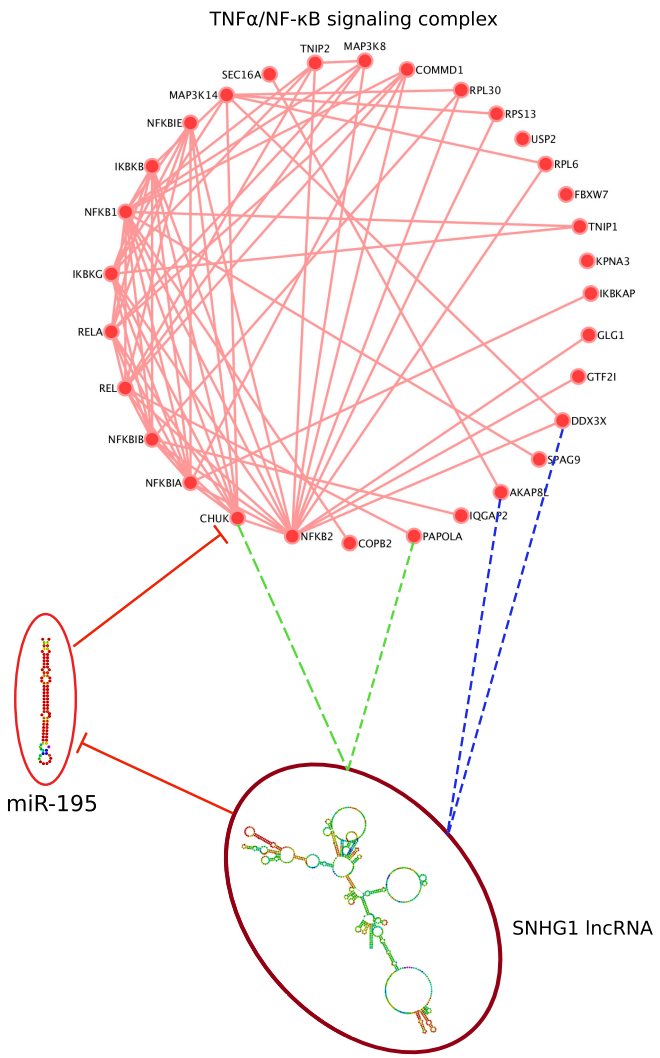

Figure 7. SNHG1 lncRNA gene association to hepatocellular carcinoma through interaction with the TNFα/NF-κB signaling complex. Red nodes represent protein components of the TNFα/NF-κB signaling complex (non-redundant CORUM complex 10). Pink edges correspond to the identified protein-protein interactions between those proteins, downloaded from IntAct (71) on 22 May 2017. Interactions predicted in this study are represented by green dashed edges, experimentally determined ones (see Supplementary Material) by blue dashed edges. Negative regulatory interactions are shown in red and are taken from (53) and (52). ViennaRNA web services were used to predict the secondary structure of SNHG1 and miR-195 (72).

teractions occurring in particular biological contexts, experimentally assessing the quality of the predicted interactions in our network is a desirable goal. However, issues relative to the fraction of interactions to be tested, the sensitivity of the chosen experimental assay, the fraction of interactions identifiable by the chosen assay, and its precision have to be solved beforehand as proposed in the case of the assessment of large-scale binary protein interactomes (54). Moreover, validating the predicted protein complex scaffolding function of lncRNAs is yet another challenge that should involve a wealth of experimental work — e.g., knocking-down of the lncRNA, determination of the localization of the predicted associated complex, its effect on the cell, as well as analysis of binding sites involved in the binding of each protein by the lncRNA (55) — which is beyond the scope of our analysis. Overall, these reasons justify our integration of several orthologous functional datasets to validate our interaction predictions and the possibility of the lncRNA to be indeed functional in the cell.

A growing body of evidence suggests that a significant fraction of lncRNAs has a function (36,37). Large-scale efforts to determine or predict lncRNA function have used their metabolic properties (35,36), sequence or structural conservation (40,41), differential expression in disease (56), lncRNA and protein-coding gene co-expression profiles (57), variant analysis (58), as well as combinations thereof (59). Methods to understand the function of individual lncRNAs through direct interaction with proteins have been exploited to a lesser extent, and are generally restricted to the limited number of known RBPs assessed to date. Hence, there is a clear need for novel large-scale methods to investigate the functions of ncRNAs acting through protein–RNA interactions, such as their ability to scaffold protein groups.

Although protein–RNA interactions are usually perceived as a protein-centric mechanism, they are now also envisioned as a RNA-centric question, where the interactions are driven by the RNA (13). However, even for RNA-centric experiments where the RNA is precipitated and its interacting proteins are identified with MS, each experiment seems to underestimate the number of proteins interacting with lncRNAs. This was observed for the XIST lncRNA, where five independent studies found >600 proteins in total associated with XIST, of which only one is in common between the five studies (12). Hence, we used a method based on proteome-wide and transcriptome-wide interaction predictions combined with tissue-expression information, and predict the presence of millions of protein–RNA interactions in human cells.

As our knowledge of proteins with RNA-binding capabilities is still incomplete (13), we produced proteome-wide protein–RNA interaction predictions to explore the action of lncRNAs at a wider level, going beyond the current knowledge. Indeed, using the *cat*RAPID algorithm, we find that many proteins not yet identified as RBPs have a high propensity to interact with several lncRNAs, as RBPs do. However, with increasingly stringent interaction-propensity cutoffs, we observe a significant increase in the proportion of proteins that are annotated as RBPs (e.g. Spearman's rank correlation coefficient = 0.985, P-value < 2.2e–16, for proteins with at least five interaction partners; Supplementary Figure S7), even though many RBPs display milder binding propensities (e.g. we retain only 79.3% of RBPs with at least 10 interactions above score = 100; 6.6% for score = 200). As RBPs are predicted to interact with lncRNAs with different interaction propensities, we selected an interaction-propensity score cutoff (≥50) that would ensure that we capture biological information, as applied in previous studies (60), while allowing for a large number of possible interactions to be detected.

Due to computational constraints, we have restricted our analysis to lncRNAs of up to 1200 nucleotides, thus excluding well characterized moderately long or very long scaffolding molecules such as MALAT1, NEAT1 and XIST,

that are known to bind dozens to hundreds of proteins (61,62). In addition, lncRNA identification studies have found from tens- to hundreds-of-thousands of novel lncRNAs (63) that are shown to vary according to the methodology used and experimental conditions. This suggests the identification of human lncRNAs is far from complete, but also that lncRNA identification methods are not yet convergent (64). In our study, we therefore restricted our analysis to lncRNAs from the curated dataset of GENCODE, widely considered as the human gene annotation reference standard. However, this also means that several recently found lncRNA scaffolds such as the LUNAR-1 (65), linc-RAM (66) and PARTICLE (67) lncRNAs are not yet present in the GENCODE dataset.

Importantly, we identified for the first time 847 lncRNAs, accounting for ∼5% of the human long non-coding transcriptome, that potentially act as RNA scaffolding molecules for a total of 1517 protein complexes or modules, roughly half of the human protein complexes known to date. As for protein–RNA interactions, knowledge of the human protein complexome is not yet comprehensive. Therefore, we used several datasets of protein complexes to better cover the protein complex space. Indeed, these datasets are largely non-redundant, with 0 to 12.4% of complexes sharing $> = 50\%$ of their constituent proteins with another complex of the same dataset (Supplementary Table S6). In addition, the three datasets are largely complementary, with at the most 20.4% of complexes sharing ≥50% of their proteins between datasets (Supplementary Table S7), and none of the complexes being entirely shared between datasets. A slightly higher inter-dataset overlap (26.2%) is found for network modules, mostly due to the higher module size compared to the protein complexes. As expected, we found that each protein group dataset used allows identifying a different set of scaffolding lncRNA candidates and the majority of the candidates (57%) are detected exclusively with one dataset of protein groups (Supplementary Figure S6B). Overall, this reveals the necessity of considering several datasets for a global analysis of human cellular complexes, as performed in this study.

Notably, our study indicates that RNA scaffolding may be an important regulatory mechanism, not limited to the few well-known cases. We indeed greatly expand the current knowledge on RNA-mediated scaffolding, by proposing that scaffolding occurs with a high prevalence and for most cellular processes. Even though major cellular functions such as telomere repair, signal peptide recognition and translation are known to closely involve RNA components, usual methods to identify cellular macromolecular complexes routinely use an RNA nuclease step before protein purification (2), thereby hindering the possible detection of RNA components in protein complexes. It is therefore likely that many ribonucleoprotein (RNP) complexes have previously been overlooked. These can possibly be retraced with a computational approach, as suggested by our results. Moreover, cellular functions are not only performed via stable macromolecular complexes, but also through stepwise reactions performed by molecules whose temporal and spatial proximity may be mediated by other molecules, as exemplified by the MAYA lncRNA, which links two pathways related to cancer metastasis through protein interaction (68). Such situations are also taken into account by our analyses when investigating interaction enrichment of lncRNAs to functional network modules. Indeed, our data revealed hundreds of modules which may be organized by RNA scaffolding.

Several lncRNAs have been shown to bind protein complexes by interacting with a single protein of the complex. Examples include HOTAIR, MEG3 and Linc-RAM which have been shown to regulate gene expression through their binding to only one component of chromatin-remodeling complexes (PRC2 (4), LSD1 (10), and MyoD–Baf60c–Brg1 complexes (66)). As our enrichment-based approach only allows identification of lncRNAs that bind at least two proteins of the same complex or module, single-protein-binding lncRNAs are beyond the scope of our approach. However, we report a short isoform of the MEG3 gene predicted to interact with several proteins of a chromatin-remodeling-related complex, suggesting that here again, some functional protein–RNA interactions may have been missed by experimental approaches, therefore emphasizing the power of predictive computational analyses.

Overall, our findings suggest the widespread prevalence of scaffolding function for lncRNAs. By proposing that lncRNAs perform such a scaffolding function for a large fraction of protein complexes and functional modules, we further characterize their function and open new questions regarding the importance and essential nature of RNA-mediated scaffolding in the cell.

AVAILABILITY

The filtered protein-lncRNA interaction network produced and analysed in this study is provided at: http://tagc.univ-mrs.fr/MoonDB/protein_lncrna_interaction_network.tsv.gz [38 Mbytes]. Source code used for data processing and analyses can be found at: https://github.com/TAGC-Brun/RAINET-RNA.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are indebted to Charles E. Chapple for helpful suggestions and careful editing of the manuscript. We thank Anaïs Baudot, Alberto Valdeolivas, Nieves Lorenzo, Davide Cirillo and Alexandros Armaos for fruitful discussions.

FUNDING

Work in IB's lab was partially supported by grants from ERC-2013 [AdG 340172–MUNCODD]; Telethon [GGP16213]; Human Frontiers Science Program Award [RGP0009/2014]; Parent Project Italia, AFM-Telethon [17835]; Epigen-Epigenomics Flagship Project and AriSLA full grant 2014 'ARCI'. Work in GGT's lab was supported by the European Research Council [RIBOMYLOME_309545]; Spanish Ministry of Economy and Competitiveness [BFU2014-55054-P]. The RAINET project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University—A*MIDEX,

a French 'Investissements d'Avenir' programme (to C.B.). Funding for open access charge: Excellence Initiative of Aix-Marseille University—A*MIDEX [RAINET].
Conflict of interest statement. None declared.

REFERENCES

1. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
2. Geisler,S. and Coller,J. (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.*, **14**, 699–712.
3. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.
4. Mondal,T., Subhash,S., Vaid,R., Enroth,S., Uday,S., Reinius,B., Mitra,S., Mohammed,A., James,A.R., Hoberg,E. *et al.* (2015) MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures. *Nat. Commun.*, **6**, 7743.
5. Clemson,C.M., Hutchinson,J.N., Sara,S.A., Ensminger,A.W., Fox,A.H., Chess,A. and Lawrence,J.B. (2009) An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol. Cell*, **33**, 717–726.
6. Wapinski,O. and Chang,H.Y. (2011) Long noncoding RNAs and human disease. *Trends Cell Biol.*, **21**, 354–361.
7. Spitale,R.C., Tsai,M.C. and Chang,H.Y. (2011) RNA templating the epigenome: Long noncoding RNAs as molecular scaffolds. *Epigenetics*, **6**, 539–543.
8. Good,M.C., Zalatan,J.G. and Lim,W.A. (2011) Scaffold proteins: hubs for controlling the flow of cellular information. *Science*, **332**, 680–686.
9. Chujo,T., Yamazaki,T. and Hirose,T. (2015) Architectural RNAs (arcRNAs): A class of long noncoding RNAs that function as the scaffold of nuclear bodies. *Biochim. Biophys. Acta*, **1859**, 139–146.
10. Tsai,M.-C., Manor,O., Wan,Y., Mosammaparast,N., Wang,J.K., Lan,F., Shi,Y., Segal,E. and Chang,H.Y. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329**, 689–693.
11. Zhang,Y., He,Q., Hu,Z., Feng,Y., Fan,L., Tang,Z., Yuan,J., Shan,W., Li,C., Hu,X. *et al.* (2016) Long noncoding RNA LINP1 regulates repair of DNA double-strand breaks in triple-negative breast cancer. *Nat. Struct. Mol. Biol.*, **23**, 1–12.
12. Cirillo,D., Blanco,M., Armaos,A., Buness,A., Avner,P., Guttman,M., Cerase,A. and Tartaglia,G.G. (2016) Quantitative predictions of protein interactions with long noncoding RNAs. *Nat. methods*, **14**, 5–6.
13. Beckmann,B.M., Castello,A. and Medenbach,J. (2016) The expanding universe of ribonucleoproteins: of novel RNA-binding proteins and unconventional interactions. *Pflügers Arch. - Eur. J. Physiol.*, **468**, 1029–1040.
14. Van Nostrand,E.L., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., Surka,C., Elkins,K. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. methods*, **13**, 1–9.
15. Hao,Y., Wu,W., Li,H., Yuan,J., Luo,J., Zhao,Y. and Chen,R. (2016) NPInter v3.0: an upgraded database of noncoding RNA-associated interactions. *Database: J. Biol. databases Curation*, **2016**, baw057.
16. Agostini,F., Zanzoni,A., Klus,P., Marchese,D., Cirillo,D. and Tartaglia,G.G. (2013) CatRAPID omics: a web server for large-scale prediction of protein–RNA interactions. *Bioinformatics*, **29**, 2928–2930.
17. GTEx Consortium (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
18. Li,J.H., Liu,S., Zhou,H., Qu,L.H. and Yang,J.H. (2014) StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.*, **42**, 92–97.
19. Huttlin,E.L., Ting,L., Bruckner,R.J., Gebreab,F., Gygi,M.P., Szpyt,J., Tam,S., Zarraga,G., Colby,G., Baltier,K. *et al.* (2015) The BioPlex Network: a systematic exploration of the human interactome. *Cell*, **162**, 425–440.
20. Wan,C., Borgeson,B., Phanse,S., Tu,F., Drew,K., Clark,G., Xiong,X., Kagan,O., Kwan,J., Bezginov,A. *et al.* (2015) Panorama of ancient metazoan macromolecular complexes. *Nature*, **525**, 339–344.
21. Ruepp,A., Waegele,B., Lechner,M., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Mewes,H.W. (2009) CORUM: the comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Res.*, **38**, 497–501.
22. Havugimana,P.C., Hart,G.T., Nepusz,T., Yang,H., Turinsky,A.L., Li,Z., Wang,P.I., Boutz,D.R., Fong,V., Phanse,S. *et al.* (2012) A census of human soluble protein complexes. *Cell*, **150**, 1068–1081.
23. Chapple,C.E., Robisson,B., Spinelli,L., Guien,C., Becker,E. and Brun,C. (2015) Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.*, **6**, 7412.
24. Bellucci,M., Agostini,F., Masin,M. and Tartaglia,G.G. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.
25. Agostini,F., Cirillo,D., Bolognesi,B. and Tartaglia,G.G. (2013) X-inactivation: quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Res.*, **41**, 1–9.
26. Kornienko,A.E., Dotter,C.P., Guenzl,P.M., Gisslinger,H., Gisslinger,B., Cleary,C., Kralovics,R., Pauler,F.M. and Barlow,D.P. (2016) Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol.*, **17**, 14.
27. Lee,S., Kopp,F., Chang,T.C., Sataluri,A., Chen,B., Sivakumar,S., Yu,H., Xie,Y. and Mendell,J.T. (2015) Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell*, **164**, 69–80.
28. Minajigi,A., Froberg,J.E., Wei,C., Sunwoo,H., Kesner,B., Colognori,D., Lessing,D., Payer,B., Boukhali,M., Haas,W. *et al.* (2015) Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science*, **349**, aab2276.
29. Ballarino,M., Cazzella,V., D'Andrea,D., Grassi,L., Bisceglie,L., Cipriano,A., Santini,T., Pinnarò,C., Morlando,M., Tramontano,A. *et al.* (2015) Novel long noncoding RNAs (lncRNAs) in myogenesis: a miR-31 overlapping lncRNA transcript controls myoblast differentiation. *Mol. Cell. Biol.*, **35**, 728–736.
30. Kelm,R.J., Cogan,J.G., Elder,P.K., Strauch,A.R. and Getz,M.J. (1999) Molecular interactions between single-stranded DNA-binding proteins associated with an essential MCAT element in the mouse smooth muscle alpha-actin promoter. *J. Biol. Chem.*, **274**, 14238–14245.
31. Cirillo,D., Agostini,F. and Tartaglia,G.G. (2013) Predictions of protein–RNA interactions. *WIRES Comput. Mol. Sci.*, **3**, 161-175.
32. Becker,E., Robisson,B., Chapple,C.E., Guénoche,A. and Brun,C. (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinforma.*, **28**, 84–90.
33. Brun,C., Chevenet,F., Martin,D., Wojcik,J., Guénoche,A. and Jacq,B. (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol.*, **5**, R6.
34. Davidovich,C., Zheng,L., Goodrich,K.J. and Cech,T.R. (2013) Promiscuous RNA binding by Polycomb repressive complex 2. *Nat. Struct. Mol. Biol.*, **20**, 1250–1257.
35. Mukherjee,N., Calviello,L., Hirsekorn,A., de Pretis,S., Pelizzola,M. and Ohler,U. (2017) Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat. Struct. Mol. Biol.*, **24**, 86–96.
36. Hon,C., Ramilowski,J., Harshbarger,J., Bertin,N., Rackham,O., Gough,J., Denisenko,E., Schmeier,S., Poulsen,T., Severin,J. *et al.* (2017) An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*, **543**, 199-204.
37. Liu,S.J., Liu,S.J., Horlbeck,M.A., Cho,S.W., Birk,H.S., Malatesta,M., Attenello,F.J., Villalta,J.E., Cho,M.Y., Chen,Y. *et al.* (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, **355**, aah7111.
38. Ning,S., Zhang,J., Wang,P., Zhi,H., Wang,J., Liu,Y., Gao,Y., Guo,M., Yue,M., Wang,L. *et al.* (2016) Lnc2Cancer: a manually curated

database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.*, **44**, D980–D985.
39. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, 983–986.
40. Necsulea,A., Soumillon,M., Warnefors,M., Liechti,A., Daish,T., Zeller,U., Baker,J.C., Grützner,F. and Kaessmann,H. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, **505**, 635–640.
41. Smith,M.A., Gesell,T., Stadler,P.F. and Mattick,J.S. (2013) Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.*, **41**, 8220–8236.
42. Gulko,B., Hubisz,M.J., Gronau,I. and Siepel,A. (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Publ. Group*, **47**, 276–283.
43. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
44. Tørring,P.M., Larsen,M.J., Kjeldsen,A.D., Ousager,L.B., Tan,Q. and Brusgaard,K. (2014) Long non-coding RNA expression profiles in hereditary haemorrhagic telangiectasia. *PLoS One*, **9**, e90272.
45. Dupuis-Girod,S., Bailly,S. and Plauchu,H. (2010) Hereditary hemorrhagic telangiectasia: from molecular biology to patient care. *J. Thromb. Haemostasis: JTH*, **8**, 1447–1456.
46. Gallione,C.J., Repetto,G.M., Legius,E., Rustgi,A.K., Schelley,S.L., Tejpar,S., Mitchell,G., Drouin,E., Westermann,C.J.J. and Marchuk,D.A. (2004) A combined syndrome of juvenile polyposis and hereditary haemorrhagic telangiectasia associated with mutations in MADH4 (SMAD4). *Lancet*, **363**, 852–859.
47. Yin,D.-D., Liu,Z.-J., Zhang,E., Kong,R., Zhang,Z.-H. and Guo,R.-H. (2015) Decreased expression of long noncoding RNA MEG3 affects cell proliferation and predicts a poor prognosis in patients with colorectal cancer. *Tumour Biol. J. Int. Soc. Oncodev. Biol. Med.*, **36**, 4851–4859.
48. Zhang,M., Wang,W., Li,T., Yu,X., Zhu,Y., Ding,F., Li,D. and Yang,T. (2016) Long noncoding RNA SNHG1 predicts a poor prognosis and promotes hepatocellular carcinoma tumorigenesis. *Biomed. Pharmacother.*, **80**, 73–79.
49. You,J., Fang,N., Gu,J., Zhang,Y., Li,X., Zu,L. and Zhou,Q. (2014) Noncoding RNA small nucleolar RNA host gene 1 promote cell proliferation in nonsmall cell lung cancer. *Indian J. Cancer*, **51**, e99–e102.
50. Luedde,T. and Schwabe,R.F. (2011) NF-κB in the liver–linking injury, fibrosis and hepatocellular carcinoma. *Nat. Rev. Gastroenterol. Hepatol.*, **8**, 108–118.
51. Wu,Y. and Zhou,B.P. (2010) TNF-alpha/NF-kappaB/Snail pathway in cancer cell migration and invasion. *Br. J. Cancer*, **102**, 639–644.
52. Zhang,H., Zhou,D., Ying,M., Chen,M., Chen,P., Chen,Z. and Zhang,F. (2016) Expression of long non-coding RNA (lncRNA) small nucleolar RNA host gene 1 (SNHG1) exacerbates hepatocellular carcinoma through suppressing miR-195. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.*, **22**, 4820–4829.
53. Ding,J., Huang,S., Wang,Y., Tian,Q., Zha,R., Shi,H., Wang,Q., Ge,C., Chen,T., Zhao,Y. *et al.* (2013) Genome-wide screening reveals that miR-195 targets the TNF-α/NF-κB pathway by down-regulating IκB kinase alpha and TAB3 in hepatocellular carcinoma. *Hepatol.*, **58**, 654–666.
54. Venkatesan,K., Rual,J.-F., Vazquez,A., Stelzl,U., Lemmens,I., Hirozane-Kishikawa,T., Hao,T., Zenkner,M., Xin,X., Goh,K.-I. *et al.* (2009) An empirical framework for binary interactome mapping. *Nat. Methods*, **6**, 83–90.
55. Wang,K.C. and Chang,H.Y. (2012) Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, **43**, 904–914.
56. Li,J., Han,L., Roebuck,P., Diao,L., Liu,L., Yuan,Y., Weinstein,J.N. and Liang,H. (2015) TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.*, **75**, 3728–3737.
57. Zhao,Z., Bai,J., Wu,A., Wang,Y., Zhang,J., Wang,Z., Li,Y., Xu,J. and Li,X. (2015) Co-LncRNA: investigating the lncRNA combinatorial effects in GO annotations and KEGG pathways based on human RNA-Seq data. *Database: J. Biol. databases Curation*, **2015**, 1–7.
58. Chen,X., Hao,Y., Cui,Y., Fan,Z., He,S., Luo,J. and Chen,R. (2017) LncVar: a database of genetic variation associated with long non-coding genes. *Bioinformatics*, **33**, 112–118.
59. Park,C., Yu,N., Choi,I., Kim,W. and Lee,S. (2014) LncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs. *Bioinformatics*, **30**, 2480–2485.
60. Zanzoni,A., Marchese,D., Agostini,F., Bolognesi,B., Cirillo,D., Botta-Orfila,M., Livi,C.M., Rodriguez-Mulero,S. and Tartaglia,G.G. (2013) Principles of self-organization in biological pathways: a hypothesis on the autogenous association of alpha-synuclein. *Nucleic Acids Res.*, **41**, 9987–9998.
61. West,J., Davis,C., Sunwoo,H., Simon,M., Sadreyev,R., Wang,P., Tolstorukov,M. and Kingston,R. (2014) The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell*, **55**, 791–802.
62. Cerase,A., Pintacuda,G., Tattermusch,A. and Avner,P. (2015) Xist localization and function: new insights from multiple levels. *Genome Biol.*, **16**, 166.
63. Zhao,Y., Li,H., Fang,S., Kang,Y., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. and Chen,R. (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. **44**, D203–D208.
64. Kashi,K., Henderson,L., Bonetti,A. and Carninci,P. (2015) Discovery and functional analysis of lncRNAs: methodologies to investigate an uncharacterized transcriptome. *Biochim. Biophys. Acta*, **1859**, 3–15.
65. Trimarchi,T., Bilal,E., Ntziachristos,P., Fabbri,G., Dalla-Favera,R., Tsirigos,A. and Aifantis,I. (2014) Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. *Cell*, **158**, 593–606.
66. Yu,X., Zhang,Y., Li,T., Ma,Z., Jia,H., Chen,Q., Zhao,Y., Zhai,L., Zhong,R., Li,C. *et al.* (2017) Long non-coding RNA Linc-RAM enhances myogenic differentiation by interacting with MyoD. *Nat. Commun.*, **8**, 14016.
67. O'Leary,V.B., Hain,S., Maugg,D., Smida,J., Azimzadeh,O., Tapio,S., Ovsepian,S.V. and Atkinson,M.J. (2017) Long non-coding RNA PARTICLE bridges histone and DNA methylation. *Sci. Rep.*, **7**, 1790.
68. Li,C., Wang,S., Xing,Z., Lin,A., Liang,K., Song,J., Hu,Q., Yao,J., Chen,Z., Park,P.K. *et al.* (2017) A ROR1-HER3-lncRNA signalling axis modulates the Hippo – YAP pathway to regulate bone metastasis. *Nat. Cell Biol.*, **19**, 106–119.
69. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
70. Quek,X.C., Thomson,D.W., Maag,J.L.V., Bartonicek,N., Signal,B., Clark,M.B., Gloss,B.S. and Dinger,M.E. (2015) lncRNAdb v2.0: Expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.*, **43**, D168–D173.
71. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids Res.*, **42**, D358–D363.
72. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol. AMB*, **6**, 26.