non-coding RNA

Review

A Survey of Current Resources to Study lncRNA-Protein Interactions

Melcy Philip [1], Tyrone Chen [1] and Sonika Tyagi [1,2,3,*]

[1] School of Biological Sciences, Monash University, 25 Rainforest Walk, Clayton, VIC 3800, Australia; melcy.p.j@gmail.com (M.P.); Tyrone.Chen@monash.edu (T.C.)

[2] Monash eResearch Centre, Monash University, Clayton, VIC 3800, Australia

[3] Department of Infectious Disease, Monash University (Alfred Campus), 85 Commercial Road, Melbourne, VIC 3004, Australia

[*] Correspondence: sonika.tyagi@monash.edu

Abstract: Phenotypes are driven by regulated gene expression, which in turn are mediated by complex interactions between diverse biological molecules. Protein–DNA interactions such as histone and transcription factor binding are well studied, along with RNA–RNA interactions in short RNA silencing of genes. In contrast, lncRNA-protein interaction (LPI) mechanisms are comparatively unknown, likely directed by the difficulties in studying LPI. However, LPI are emerging as key interactions in epigenetic mechanisms, playing a role in development and disease. Their importance is further highlighted by their conservation across kingdoms. Hence, interest in LPI research is increasing. We therefore review the current state of the art in lncRNA-protein interactions. We specifically surveyed recent computational methods and databases which researchers can exploit for LPI investigation. We discovered that algorithm development is heavily reliant on a few generic databases containing curated LPI information. Additionally, these databases house information at gene-level as opposed to transcript-level annotations. We show that early methods predict LPI using molecular docking, have limited scope and are slow, creating a data processing bottleneck. Recently, machine learning has become the strategy of choice in LPI prediction, likely due to the rapid growth in machine learning infrastructure and expertise. While many of these methods have notable limitations, machine learning is expected to be the basis of modern LPI prediction algorithms.

Keywords: LPI; lncRNA; ncRNA; protein; transcriptomics; molecular docking; machine learning; deep learning; databases

check for updates

Citation: Philip, M.; Chen, T.; Tyagi, S. A Survey of Current Resources to Study lncRNA-Protein Interactions. *Non-coding RNA* **2021**, *7*, 33. https://doi.org/10.3390/ncrna7020033

Academic Editor: Y-h. Taguchi

Received: 3 May 2021
Accepted: 7 June 2021
Published: 8 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Transcriptomics is the study of a complete set of RNA transcripts in a cell, measuring variable expression levels of the genome under different conditions. Modern transcriptomics is performed with high-throughput sequencing to investigate the function of genes and biological pathways, commonly with bioinformatics methods applying differential gene expression analyses, splice site identification, transcript variant identification or determining alternative promoter usage for protein-coding transcripts [1]. However, these protein-coding transcripts only represent a small proportion of the transcriptome. A large proportion of the genome generates RNA transcripts which do not directly code for protein products [2]. These non-coding RNA (ncRNA) transcripts have been known to exist, but their properties make them difficult to characterise as compared to the coding transcripts. ncRNA can be divided into multiple categories based on function and length [3]. In this review, we specifically consider the long non-coding RNA (lncRNA) category of ncRNA and their interaction with proteins, an important functional mechanism of lncRNA.

LncRNA are very broadly defined as RNA transcripts exceeding 200 nucleotides (nt) in length without coding potential. Their length varies widely, ranging from hundreds to thousands of nucleotides [4]. LncRNA can act as gene regulators, and like other epigenetic mechanisms are involved in numerous biological processes. They achieve their regulatory function with their ability to interact with a wide range of biological molecules, such as other nucleic acids and proteins [5], as well as with small molecules [4]. Among their more direct modes of action are sequestering and releasing transcripts to modulate gene expression, stabilising transcripts and binding to DNA to sterically hinder transcription initiation [6]. More indirectly, they can recruit proteins and other molecules to form a functional complex, or act as a scaffold for targeted chromatin formation [7].

An important layer of lncRNA-mediated gene regulation is LPI (lncRNA-protein interactions). We illustrate the importance of LPI in developmental and abiotic stress pathways with several examples encompassing multiple distinct species. In *Drosophila melanogaster*, regulatory networks mediated by LPI regulate key eye development [8] and dosage compensation pathways [9] mediated by RNA-binding proteins. In the plant *Arabidopsis thaliana*, LPI control alternative splicing within the nucleus by selectively displacing existing transcripts and subsequently altering root development [10,11]. Response to abiotic stress is also governed by LPI, as shown by an lncRNA recruiting histone methylases to suppress *Arabidopsis thaliana* flowering during cold conditions [12]. *Dario renio* LPI are also observed to interface with transcription factors and other RNA-binding proteins during embryonic development, although their exact mechanism of action is not well known [13]. LPI also act as mediators of other epigenetic mechanisms, for instance as chromatin scaffolds to organise the three-dimensional structure of the genome in *Mus musculus* [14].

Due to the widespread involvement of LPI in epigenetics, dysregulation of certain LPI contributes to disease states, particularly cancers. Severity of a human pancreatic cancer phenotype is driven by an lncRNA-protein complex, which triggers a positive feedback loop of protein overexpression leading to poor patient outcomes [15]. Similarly, formation of an lncRNA-protein complex is associated with poorer prognosis in breast cancer [16], colon cancer [16] and lymphoma [17] by blocking phosphorylation sites, stabilising other epigenetic factors and through an unknown mechanism, respectively. Infectious diseases are also associated with LPI dysregulation, including COVID-19 [18,19]. A more exhaustive list of known LPI–disease associations is available at the LncTarD database [20]. Despite the wealth of information on LPI–disease associations, their precise mechanism of action remains unknown. Therefore, insight into LPI will be valuable in complex disease research, potentially resulting in improved diagnosis and treatment procedures.

Multiple high-throughput laboratory assays were developed to investigate LPI, some of which will be briefly discussed in this review article. However, exhaustively performing an experimental validation for each individual LPI is not practical given their volume and variety. Hence, computational methods are necessary to screen these high-throughput assays for potential LPI which can then be subsequently experimentally validated, similar to transcriptomics workflows for conventional protein-coding RNA [21]. A variety of these computational LPI predictors exist, each applying different strategies to achieve their goals, and are dependent on a few biological databases containing subsets of experimentally validated LPI. In this review, we will discuss recent bioinformatics resources for studying LPI, with an emphasis on software and databases, together with their advantages as well as limitations.

2. LPI Laboratory Assays

Because of the biological importance of LPI, many laboratory assays were developed to identify these interactions. Two general categories of such assays exist, protein-centric assays and RNA-centric assays, which can capture either the cellular environment of a living cell or extracted biological material [22]. Protein-centric assays target the protein component of a LPI, while RNA-centric assays target the lncRNA component. Each method varies in sensitivity and specificity, has different prerequisites and has unique advantages as well as disadvantages. Comprehensively comparing and contrasting these laboratory assays is out of the scope of this review, but we provide a high-level overview only to give the computational methods discussed in this article some biological context. A more detailed overview of these assays can be found in a separate review article [22].

To discover proteins bound to RNA of interest (RNA-centric methods), IVT (in vitro-transcribed) RNA can be tagged with biotin, and selectively bound to streptavidin for purification [23]. RaPID (RNA–protein interaction detection) [24] operates in a conceptually similar way to the previous method. IVT RNA can also be tagged with dyes and bound to protein microarrays, with fluorescence providing a quantitative output [25]. In vivo, cross-linking RNA with protein, either through formaldehyde or UV light, is used to identify LPI by purifying and extracting the RNA-bound proteins. CHART (capture hybridisation analysis of RNA targets) [26], ChIRP (chromatin isolation by RNA purification and capture hybridisation analysis of RNA targets) [27], MS2-BioTRAP (MS2 in vivo biotin-tagged RAP) [28], PAIR (peptide nucleic acid-assisted identification of RBPs) [29], RAP (RNA affinity purification) [30] and TRIP (tandem RNA isolation procedure) [31] all use either of these cross-linking strategies.

To discover RNA bound to proteins of interest (protein-centric methods), exploiting cross-linking is also common. The largest group of protein-centric methods are CLIP (cross-linking immunoprecipitation)-based methods [32]. Many variants of CLIP methods exist [33], and when paired with high-throughput sequencing are capable of generating libraries of data for further analysis. RIP-seq (RNA immunoprecipitation) [34] and TRIBE (targets of RNA-binding proteins identified by editing) [35] also belong to this category of protein-centric methods.

3. lncRNA-Protein Resource Databases

Starbase [36], POSTAR [37], RAIN [38], RNAInter [39], NPInter [40], ATtRACT [41] and oRNAment [42] are examples of databases that contain information associated with lncRNA-protein interactions obtained by the previously discussed laboratory assays, computational analysis and literature mining. Two broad classes exist: databases containing curated lncRNA-protein interactions and databases containing RNA-binding motifs.

Starbase, RNAInter, POSTAR, NPInter and RAIN all contain details of curated lncRNA-protein interactions, and many additional attributes (including functional annotation) associated with the interactions, derived from a combination of the laboratory assays discussed in the previous section (Table S1). These are not limited exclusively to lncRNA, and contain various other pieces of interaction information, including interactions with other ncRNA, other nucleic acids and proteins [43–45]. Some contrasts between these databases are also observable from a species, usability and scope perspective, which will be discussed here. Starbase, POSTAR and RAIN contain LPI information from a small number (two to four) of species, while RNAInter and NPInter host a wide range of species. To improve usability, Starbase, RNAInter and RAIN feature third party tool integration to streamline bioinformatics workflows. In terms of scope, POSTAR and NPInter appear to be focused on disease phenotypes, providing disease association information, while Starbase, RNAInter and RAIN have a more generic focus.

ATtRACT and oRNAment databases contain details of RBP (RNA-binding protein) motifs. While not directly containing LPI, these can be applied to predict putative LPI and are a useful starting point or supplementary tool in screening for LPI.

To provide a guide for the community on database selection, we generated a recommendation matrix (Table S2). We considered five lncRNAs, namely NEAT1, MALAT1 and Hotair (well studied) versus Lassie and MaTAR25 (less explored). We discovered that Starbase is an exclusive database which provides MALAT1–protein interactions with the CLIP-seq evidence, whereas POSTAR2 provides RNA- and RBP-centric interactome information for the well-examined lncRNAs. Similarly, RAIN provides RNA–protein interaction details and networks using STRING for NEAT1, MALAT1 and HOTAIR. RNAInter provides information associated with interacting molecules, RNA editing, RNA structure, RNA localisation, RNA modification, evidence support (experimental evidence) and references, interaction networks (the top 100 interactions) and dynamic expression for the major lncRNAs. NPInter integrates NONCODE and ENSEMBL data to document and annotate the available information for NEAT1, MALAT1 and HOTAIR while ATtRACT is an RBP-centric database (keyword should be a RBP) that provides the RBP details and associated motifs. oRNAment consists of detailed information on transcripts and RBP along with numerous downloadable graphical representations of the noted lncRNAs with multiple visualisation options. However, none of these databases include any information on emerging lncRNAs such as Lassie and MaTAR25, further highlighting the reliance of the community on these databases.

All databases feature at least mouse and human datasets, likely due to their status as model organisms relevant to human disease, although some incorporate other model organisms as well. It is interesting to note that all databases feature advanced querying and search functions, likely reflecting the volume and complexity of LPI data. We have reviewed and compared them in Table S1. In summary, we discovered that there is a surprising lack of specialised LPI databases, with most databases featuring combinations of other nucleic acid and protein combinations. The biggest limitation of the current databases is that the LPI data are available only at a gene level and not a transcript level, lowering the resolution of LPI discovery methods which use these data. In a separate (unpublished) study we demonstrated that different isoforms of a lncRNA genes can have different interactomes, and hence functions. We are also developing machine learning methods to annotate lncRNAs at the transcript level (https://bioinformaticslab.erc.monash.edu/linc2function accessed on 27 May 2021).

4. LPI Prediction Algorithms

Most LPI prediction algorithms exploit these curated databases of prior LPI knowledge to tune their predictions. Computational strategies for LPI prediction can be divided into two high-level categories, molecular docking and machine learning. Lower-level subdivisions among the methods we surveyed are visualised in Figure 1, and include deep learning, tree-based methods, graph-based methods, similarity networks, image segmentation, matrix factorisation and variants of the Fourier transform. Conventional molecular docking methods operate by finding the optimal configuration of an lncRNA-protein complex, and ranking the highest scoring configurations for further evaluation. Within the past decade, a large number of prediction algorithms based on machine learning have emerged. Most machine learning methods do not involve molecular docking simulations. Instead, they exploit known interactions between lncRNA and protein and/or biomolecular sequence information directly, although many also leverage known secondary structures to improve their performance table [1,2]. As with the LPI databases, it is worth noting that none of these methods are tuned specifically for LPI prediction, and represent broader scopes of identifying combinations of nucleic acid–protein interaction.

4.1. Molecular Docking Approaches

Before the current ecosystem of machine learning algorithms was established, molecular docking was the dominant strategy used to predict and investigate LPI or RNA–protein interactions in general. By developing custom equations, which account for conformation and other steric properties, the likelihood of lncRNA-protein complex formation is scored. Low-level methodology does not vary significantly, with most methods applying a variant of the FFT (fast Fourier transform) to extract features from three-dimensional molecule representations, template or optimising for a minimal energy state. Key factors considered include docking pose, distance and area of interracial sites, energy-based criteria and selection of the most structurally conserved docked complex [46]. Several methods also account for sequence homology or electrical charge between biological molecules [47]. Hierarchical clustering to group complexes of interest is not uncommon. However, at a high level these strategies are applied in different ways, and on different steric features. In many cases, a set of parameters must be specified by the user.

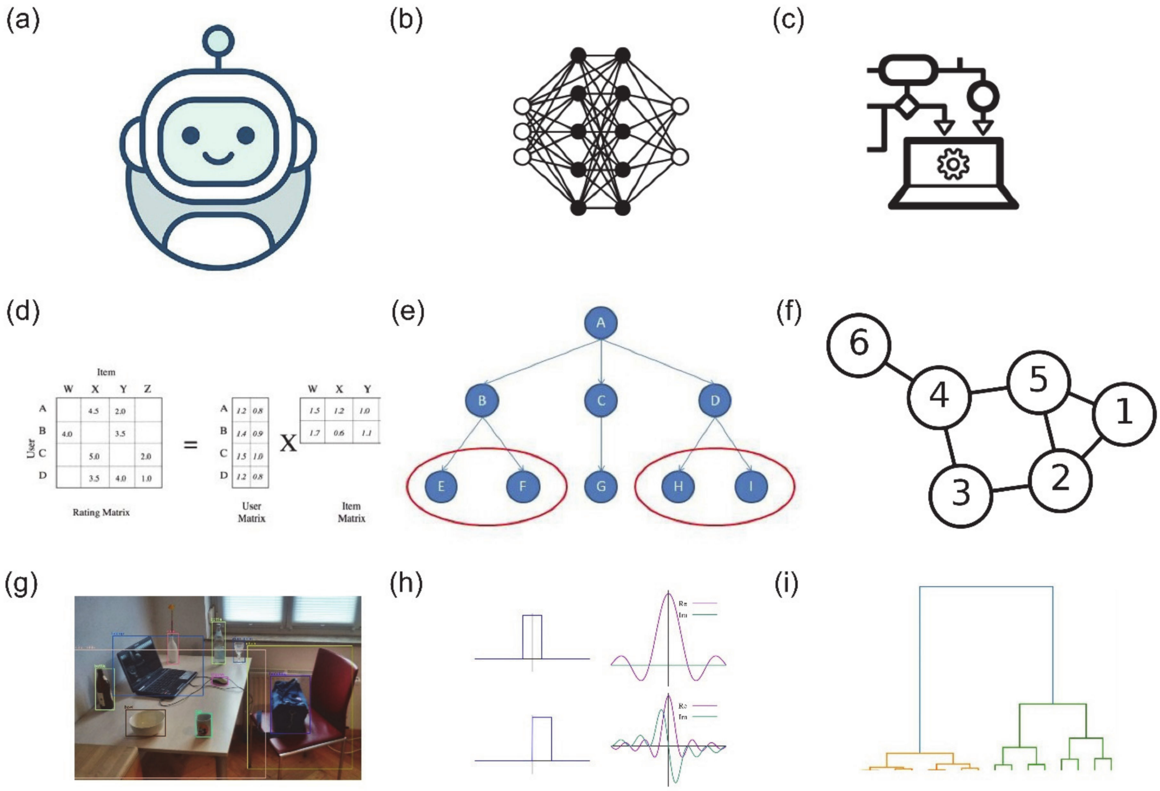

Figure 1. Visualisation of the broad categories of strategies used for predicting lncRNA-protein interactions. (**a**) Machine learning, (**b**) deep learning, (**c**) ensemble learning, (**d**) matrix factorisation, (**e**) similarity network analysis, (**f**) graph theory, (**g**) segmentation, (**h**) Fourier transform (in lncRNA-protein molecular docking simulations) and (**i**) hierarchical clustering. Training data are commonly higher-level features (e.g., structure, orientation) of lncRNA and proteins as well as the sequences recoded into tensors of varying dimensions.

Most of the molecular docking methods we reviewed use methods which incorporate at least two of the previously discussed low-level methodologies (Table 1). To provide some context for the building blocks of these more complex methods, we first present examples of methods that use an individual strategy together with a brief discussion of their advantages and disadvantages, which include 3dRPC [48], HexServer [49], FireDOCK [50], HADDOCK [51] and PatchDOCK [52]. HexServer and 3dRPC are FFT-based methods, and 3dRPC is effective on well-characterised molecules only. By exploiting the fact that LPI complexes have looser packing, 3dRPC implements FFT on geometric complementarity as well as electrostatics with a custom scoring function. HexServer uses an FFT-based algorithm to exploit shape complementarity as a feature for optimisation. Its key advantage is its reformulation of the conventional 3D search space to greatly boost the speed of the FFT, achieving results in seconds. Meanwhile, FireDOCK and HADDOCK optimise the minimum free energy of the lncRNA-protein complex. While FireDOCK focuses on exploiting side chain information, HADDOCK leverages ambiguous interaction restraints, and is one of the few methods which has the advantage of being applicable to multi-body problems as well as other biomolecular interactions. Among molecular docking tools, PatchDOCK takes a more unconventional strategy by summarising low-level geometric features into higher-level features, and has some conceptual similarities to image segmentation. It is interesting to note that FireDOCK and PatchDOCK both complement each other, where PatchDOCK can feed output directly into FireDOCK.

Table 1. A comparison of molecular docking tools used to predict lncRNA-protein interactions. Important attributes of these molecular docking tools, including their effectiveness and a link to their corresponding server, are listed (all weblinks are accessed on 27 May 2021).

Sl:No	Resource	Resource Type	Advantages and Disadvantages	Weblink	Reference Paper
1	P3DOCK	lncRNA-protein docking server (adapted from conventional docking servers)	Free docking and template-based docking strategies in a hybrid approach, results in an accurate classification	http://www.rnabinding.com/P3DOCK/P3DOCK.html	[55]
2	HDOCK	lncRNA-protein docking server (adapted from conventional docking servers)	Integrates template-based modelling as well as ab initio free docking, with a scope that extends to both proteins and nucleic acids	http://hdock.phys.hust.edu.cn	[53]
3	PATCHDOCK	lncRNA-protein docking server (adapted from conventional docking servers)	Low-level geometric features into higher-level features, FireDOCK and PatchDOCK both complement each other, where PatchDOCK can feed output directly into FireDOCK.	https://bioinfo3d.cs.tau.ac.il/PatchDock//	[52]
4	FIREDOCK	lncRNA-protein docking server (adapted from conventional docking servers)	Focuses on exploiting side chain information, optimises the minimum free energy of the lncRNA-protein complex	http://bioinfo3d.cs.tau.ac.il/FireDock/ /	[50]
5	NPDOCK	Exclusively lncRNA-protein docking server, developed for nucleic acid docking only	Chains multiple methods into a pipeline of tools, which implement mostly FFT-based methods.	http://genesilico.pl/NPDock /	[56]
6	HADDOCK	lncRNA-protein docking server (adapted from conventional docking servers)	It averages ambiguous interaction restraints, and it can generalise to multi-body problems as well as other biomolecular interactions, optimises the minimum free energy of the lncRNA-protein complex	https://wenmr.science.uu.nl/haddock2.4/	[51]
7	MPRDOCK	lncRNA-protein docking server (adapted from conventional docking servers)	Implies protein flexibility by applying FFT and considering sequence homology of the target of interest to generate a repertoire of structures for "ensemble docking"	http://huanglab.phys.hust.edu.cn/mprdock/	[54]
8	Hexserver	lncRNA-protein docking server (adapted from conventional docking servers)	FFT-based algorithm to exploit shape complementarity as a feature for optimisation	http://hexserver.loria.fr/	[49]

Methods implementing a mixture of these strategies include HDOCK [53], MPRDOCK [54], P3DOCK [55] and NPDOCK [56]. HDOCK integrates template-based modelling as well as ab initio docking, with a scope that extends to both proteins and nucleic acids. In addition, the user may specify binding sites of interest directly. MPRDOCK exploits protein flexibility by applying FFT and considering sequence homology of the target of interest to generate a repertoire of structures for "ensemble docking". We note that in this specific context of

MPRDOCK, "ensemble docking" refers to the library of proteins generated by MPRDOCK, and is distinct from "ensemble learning" in the machine learning approaches section where the outputs of multiple algorithms are aggregated to obtain a result. P3DOCK integrates the previously discussed 3dRPC, as PRIME that leverages sequence as well as structural homology in addition to the features used by 3dRPC. P3DOCK's authors claim that by complementing free docking and template-based docking strategies in a hybrid approach, a more accurate classification is possible. Finally, NPDOCK does not use a hybrid or ensemble strategy, but chains multiple methods into a pipeline of tools, which implement mostly FFT-based methods. The main advantage of using such ensemble methods is a generally improved performance over single-strategy methods as the limitations of each individual method are complemented.

With the exception of one or two methods such as HexServer, many of these algorithms are computationally expensive and time-consuming (hours to days of real time) to run. Some methods, such as HexServer, require advanced hardware such as GPUs and specialised software engineering tools. Biological molecules are complex and dynamic, with their wide range of possible conformations as well as orientations greatly increasing the search space for algorithms. The molecular docking community is mindful of this, and provides their software on publicly accessible and user-friendly web servers for users to run these programs remotely, although time remains a bottleneck for these workflows.

4.2. Machine Learning Approaches

Most modern lncRNA-protein interaction (LPI) prediction algorithms use machine learning, where large datasets with attributes of interest are passed to an algorithm (Table 2). The algorithm then "learns" from the data, discovering patterns in the data with minimal human intervention such as user-defined equations, a process known as "training". In the case of LPI, known LPI and their corresponding sequences as well as structures are used for training the prediction models. Their strategies can be divided into several broad categories, including graph methods, ensemble learning, matrix factorisation and deep learning. Of these strategies, matrix factorisation appears to be the most popular and is integrated into many other higher-level strategies. LPI are commonly formulated as similarity matrices, which can then be easily formulated as a matrix factorisation problem. Broader strategies incorporating matrix factorisation, such as ensemble learning and methods which leverage multimodal data, appear to have consistently robust performance [57]. Few deep learning models exist, but they both perform and generalise well in comparison to other methods, and are likely to become more popular as they have become in other areas of biology.

Table 2. A comparison of machine learning algorithms used to predict lncRNA-protein interactions. Important attributes of these machine learning algorithms, including their scope, strategies, training data, effectiveness and reproducibility are listed. More than half of these methods are not reproducible as their source code is proprietary or not available. A few methods provide web interfaces for users to enter their own data (all weblinks are accessed on 27 May 2021).

Sl:no	Resource	Scope	Advantages and Disadvantages	Strategy	Problem Formulation	Model Training Data	Weblink/Source Code	Reference Paper
1	LPI-FKLKRR (lncRNA-protein interaction kernel ridge regression, based on fast kernel learning)	Prediction	Effective in datasets with imbalanced classes.	Kernel ridge regression	Similarity matrices formulated as kernels	lncRNA-protein interactions, lncRNA expression, protein ontology, lncRNA sequence, protein sequence	https://github.com/6gbluewind/LPI_FKLKRR	[58]
2	LPI-KTASLP (prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information)	Prediction, discovery	Effective in datasets with imbalanced classes.	Multiple kernel learning	Similarity matrices formulated as kernels	lncRNA-protein interactions, lncRNA expression, lncRNA sequence	https://github.com/6gbluewind/LPI_KTASLP	[59]
3	LPI-NRLMF (lncRNA-protein interaction prediction by neighbourhood regularised logistic matrix factorisation)	Prediction, discovery	Prediction bias is expected due to the sparsity of the training dataset.	Matrix factorisation	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, protein sequence	NA	[60]
4	LPI-INBRA (long non-coding RNA–protein interaction prediction based on improved bipartite network recommender algorithm)	Prediction	Robust against false positives.	Matrix factorisation	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, protein sequence	NA	[61]
5	LPI-BNPRA (long non-coding RNA–protein interaction bipartite network projection recommended algorithm)	Prediction	Effective in humans and closely related species.	Bipartite network recommendation	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, protein sequence	NA	[62]

Table 2. *Cont.*

Sl:no	Resource	Scope	Advantages and Disadvantages	Strategy	Problem Formulation	Model Training Data	Weblink/Source Code	Reference Paper
6	PBLPI (path-based lncRNA-protein interaction)	Prediction, discovery	Prediction accuracy limited due to technical limitations.	Graph	Similarity matrices	lncRNA-protein interactions, protein semantic similarity, lncRNA functional similarity, Gaussian interaction profile kernel similarity, integrated similarity for lncRNAs and proteins	NA	[63]
7	PLPIHS (predicting lncRNA-protein interactions using HeteSim scores)	Prediction, discovery	Performance is improved by preserving information regarding the biological network, taking into account lncRNA-protein interactions similar to the target.	Graph	Similarity matrices	Co-expression data of lncRNA-protein pairs, lncRNA-protein interaction data	NA	[64]
8	IRWNRLPI (integrating random walk and neighbourhood regularised logistic matrix factorisation for lncRNA-protein interaction prediction)	Prediction	Robust due to hybrid approach, but known to be unstable.	Hybrid: random walk, neighbourhood regularised logistic matrix factorisation algorithm	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, protein sequence	NA	[65]

Table 2. *Cont.*

Sl:no	Resource	Scope	Advantages and Disadvantages	Strategy	Problem Formulation	Model Training Data	Weblink/Source Code	Reference Paper
9	SFPEL-LPI (sequence-based feature projection ensemble learning method)	Prediction, discovery	Multimodal approach boosts prediction accuracy.	Ensemble: graph Laplacian regularisation	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, protein sequence	http://www.bioinfotech.cn/SFPEL-LPI/	[66]
10	HLPI-Ensemble (human lncRNA-protein interactions ensemble)	Prediction	Scope restricted to humans.	Ensemble: support vector machines (SVM), random forests (RF) and extreme gradient boosting (XGB)	Recoded feature vectors	lncRNA-protein interactions, lncRNA sequence, lncRNA features, protein sequence, protein features	NA	[63]
11	GPLPI (graph predict lncRNA-protein interaction)	Prediction	Scope restricted to plants.	Deep learning, ensemble learning, graph attention LSTM autoencoder	Recoded sequence and structure vectors	lncRNA sequences, protein sequences, structural features from predicted secondary structures from lncRNA and protein sequences.	https://github.com/Mjwl/GPLPI	[67]
12	LPI-BLS (predicting lncRNA-protein interactions with a broad learning system-based stacked ensemble classifier)	Prediction	Flat network architecture boosts speed and accuracy. Effective in several model organisms.	Ensemble: broad learning system (flat neural network)	Recoded feature vectors	lncRNA-protein interactions, lncRNA sequence, lncRNA features, protein sequence, protein features	https://github.com/NWPU-903PR/LPI_BLS	[69]

Table 2. *Cont.*

Sl:no	Resource	Scope	Advantages and Disadvantages	Strategy	Problem Formulation	Model Training Data	Weblink/Source Code	Reference Paper
13	LPI-CNNCP (lncRNA-protein interactions convolutional neural network copy-padding trick)	Prediction	Can be extended to predict other biomolecular interactions, effective across different species.	Deep learning (convolutional neural network)	Recoded feature vectors	lncRNA-protein interactions, lncRNA sequence, protein sequence	https://github.com/NWPU-903PR/LPI-CNNCP	[71]
14	DeepLPI (deep lncRNA-protein interactions)	Prediction, discovery	Can be extended to other biomolecular interactions, unique capability to predict lncRNA interaction with different protein isoforms.	Deep learning (embedding, convolution, LSTM)	Recoded feature tensors	lncRNA-protein interactions, lncRNA sequence, lncRNA structure, protein sequence, protein structure	https://github.com/dls03/DeepLPI	[72]
15	LPI-SKF (lncRNA-protein interaction similarity kernel fusion)	Prediction, discovery	Aggregating multiple similarities increases robustness against noise.	Similarity kernel fusion, manifold learning	Similarity matrices	lncRNA-protein interactions, pairwise similarities for lncRNAs, pairwise similarities for proteins	https://github.com/zyk2118216069/LPI-SKF	[75]
16	PMKDN (projection-based neighbourhood non-negative matrix decomposition model)	Prediction	Strategy avoids overfitting and sparsity issues, allowing more generalisability to different datasets.	Neighbourhood regularised matrix factorisation algorithm	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, lncRNA expression, protein sequence, protein annotation	NA	[73]

Table 2. *Cont.*

Sl:no	Resource	Scope	Advantages and Disadvantages	Strategy	Problem Formulation	Model Training Data	Weblink/Source Code	Reference Paper
17	LPI-miRNA	Prediction, discovery	Can operate on datasets without prior knowledge of lncRNA interactions but relies on known miRNA–lncRNA and miRNA–protein interactions.	Heterogeneous network model	Similarity matrices	lncRNA–miRNA interactions, protein–miRNAs interactions	https://github.com/zyk2118216069/LncRNA-protein-interactions-prediction	[74]
18	lncPro	Prediction	Training dataset limited, effective on short sequences.	Fourier transform, matrix factorisation	Recoded feature tensors	lncRNA-protein interactions, lncRNA sequence, lncRNA features, protein sequence, protein features	http://cmbi.bjmu.edu.cn/lncpro/	[76]
19	catRAPID	Prediction	Visualisation is available, prediction accuracy may be limited by reliance on very old lncRNA annotations.	Discrete Fourier transform	lncRNA and protein secondary structure, hydrogen bonding, van der Waals forces	NA	http://s.tartaglialab.com/page/catrapid_group	[77]

Table 2. *Cont.*

Sl:no	Resource	Scope	Advantages and Disadvantages	Strategy	Problem Formulation	Model Training Data	Weblink/Source Code	Reference Paper
20	3dRPC	Prediction	Effective on well-characterised molecules, may have lower accuracy if this is not the case.	Fast Fourier transform, root mean square deviation	Conformations of nucleotide-amino-acid pairs	NA	http://biophy.hust.edu.cn/3dRPC.html	[48]
21	DeepBind	Prediction	Effective, generalisable across species, but more effective at predicting protein–DNA binding than protein–RNA binding.	Deep learning (convolutional neural network)	Recoded feature tensors	lncRNA-protein interactions, lncRNA sequence, protein sequence	http://tools.genes.toronto.edu/deepbind/	[70]
22	LPLNP	Prediction, discovery	Effective and robust in humans, capable of discovering novel interactions.	Ensemble: linear neighbourhood similarity	Similarity matrices	lncRNA expression, lncRNA features lncRNA-protein interactions, lncRNA sequence, protein features, protein sequence	https://github.com/BioMedicalBigDataMiningLabWhu/lncRNA-protein-interaction-prediction	[68]

Matrix factorisation is the most common way to formulate LPI for prediction algorithms, including LPI-FKLKRR (lncRNA-protein interaction kernel ridge regression, based on fast kernel learning) [58], LPI-KTASLP (prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information) [59], LPI-NRLMF (lncRNA-protein interaction prediction by neighbourhood regularised logistic matrix factorisation) [60], LPI-INBRA (long non-coding RNA–protein interaction prediction based on improved bipartite network recommender algorithm) [61] and LPI-BNPRA (long non-coding RNA–protein interaction bipartite network projection recommended algorithm) [62]. These methods share a common theme of formulating lncRNA-protein interactions as a matrix factorisation problem and using them in broader strategies, such as multiple kernel learning or recommender algorithms. Known structural features are often used together with sequence features. In the special case of LPI-FKLKRR, matrices are reformulated into kernels for direct optimisation with kernel ridge regression, increasing performance in the common scenario of class imbalance. Further comparing and contrasting the advantages as well as disadvantages of these methods shows that LPI-FKLKRR and LPI-KTASLP are expected to be effective in the case of imbalanced classes. In LPI-NRLMF, the authors note a slight prediction bias may occur due to the sparsity of their training data. LPI-INBRA is robust against false positives, and LPI-BNPRA is effective on closely related species other than humans.

Some graph-based methods for LPI prediction are PBLPI (path-based lncRNA-protein interaction) [63] and PLPIHS (predicting lncRNA-protein interactions using HeteSim scores) [64]. PBLPI takes into account both functional and semantic similarity between proteins, while PLPIHS uses a custom distance metric to unify co-expression, lncRNA-protein interactions and protein–protein interaction scores to construct a network which is then provided to a SVM classifier. In the case of PBLPI, a disadvantage is that prediction accuracy may be reduced due to technical limitations, while in PLPIHS performance is improved by preserving information regarding the biological network, taking into account lncRNA-protein interactions similar to the target.

Examples of hybrid and ensemble learning approaches are IRWNRLPI (integrating random walk and neighbourhood regularised logistic matrix factorisation for lncRNA-protein interaction prediction) [65], SFPEL-LPI (sequence-based feature projection ensemble learning method) [66], HLPI-Ensemble (human lncRNA-protein interactions ensemble) [63], GPLPI (graph predict lncRNA-protein interaction) [67], LPLNP (linear neighbourhood propagation method) [68] and LPI-BLS (predicting lncRNA-protein interactions with a broad learning system-based stacked ensemble classifier) [69]. IRWNRPLI uses lncRNA-protein interactions and lncRNA/protein sequence similarity as the input into a hybrid approach of random walk and neighbourhood regularised logistic matrix factorisation. Being an integrative model, it appears to be robust, although its accuracy varies on different biological systems. Ensemble approaches PMKDN, SFPEL-LPI, HLPI-Ensemble, LPI-BLS and LPLNP all have the advantage of being robust against noise due to their ensemble strategy, incorporating multiple approaches, and are capable of discovering new LPI. LPLNP and LPI-BLS in particular stand out: LPI-BLS for its unconventional flat network architecture and aggregation strategy, as well as its effectiveness in multiple species, and LPLNP for its unique application of neighbourhood similarity to LPI. However, we note that HLPI-Ensemble is specifically intended for human LPI only. GPLPI uses both sequence features and known secondary structures to train a graph-based neural network. In addition, by using an ensemble of features including evolutionary information, GPLPI's effectiveness was increased. An important distinction between these two methods is that GPLPI is trained on known plant lncRNA, and plant non-coding RNA have different properties (some ncRNA lose function even with 1–2 nucleotide changes) to those of animal non-coding RNA [70]. For this model to be effective on non-plant organisms, retraining is likely necessary but viable due to the relatively higher volume of data associated with animals, in particular humans [63].

Only a few deep learning approaches exist: DeepBind [70], LPI-CNNCP (lncRNA-protein interactions convolutional neural network copy-padding trick) [71] and DeepLPI (deep lncRNA-protein interactions) [72]. DeepBind was one of the first applications of deep learning to predict nucleic acid–protein binding, and is applicable to LPI. By reformulating the classical position weight matrix [73] as a convolutional kernel, it operates on raw sequence data to provide a simple prediction score for a nucleic acid–protein interaction [74]. LPI-CNNCP uses only lncRNA and protein sequence data recorded as k-mers as input into a CNN but achieves good results. It is also interesting to note that it appears to be one of the few models that are effective across different species, which is a less common advantage. Meanwhile, DeepLPI feeds co-expression, sequence and structural data to a neural network optimised by a conditional random field. Using isoform data makes DeepLPI the only method to date with the ability to predict lncRNA interaction with different protein isoforms. Furthermore, its flexibility allows it to be extended to other biomolecular interactions, such as miRNA.

Other methods used to predict LPI that do not fall into a specific category include LPI-SKF (lncRNA-protein interaction similarity kernel fusion) [75], PMKDN (projection-based neighbourhood non-negative matrix decomposition model) [73] and LPI-MiRNA [74]. LPI-SKF uses an integrative approach where verified lncRNA-protein interactions are used to build a network, and similarity kernel fusion is used to integrate protein and lncRNA similarity scores before applying manifold learning. PMKDN uses multiple features from lncRNA (nucleotide composition, expression levels) and protein (amino acid subcategories) to build a similarity matrix for similarity network fusion with a nearest neighbour′s approach. Both these methods have the advantages of being robust against noise and capable of interaction discovery, but like most methods that express LPI as similarity matrices, they make a strong assumption that sequence homology correlates with interactivity, which may not hold in all cases. LPI-MiRNA takes a unique approach, exploiting miRNA as an intermediate unit of lncRNA-protein binding, and uses this in a network-based approach. While this gives LPI-MiRNA the ability to operate on datasets without prior knowledge of lncRNA interactions, a different limitation is introduced of relying on known miRNA–lncRNA and miRNA–protein interactions. An assumption is also made that miRNAs which interact with both lncRNA and a protein would also form LPI, which may not always hold. Nevertheless, this method was shown to be effective.

Although lncPro [76] and catRAPID [77] are older methods, these are featured in this manuscript because of their historical significance. lncPro was one of the first published machine learning LPI prediction algorithms, and many LPI algorithms resemble it. Higher-level features are extracted from lncRNA and protein sequence, which are then recorded as vectors as input into their model. Although the authors noted limitations associated with data availability and computational complexity at the time, this method became a template for many other machine learning methods, including those discussed in this manuscript. catRAPID does not apply machine learning, but instead constructs an interaction matrix from known secondary structure and other molecular features. A major limitation of this approach is its reliance on obsolete genomic data, which is expected to reduce prediction accuracy.

However, it is important to note that the scope of most LPI prediction algorithms are limited. Not all methods can predict interactions for novel lncRNA or proteins, and few methods generalise across species [62,69,71]. This is partly due to the limited availability of curated training data, with a small number of samples mostly from human or mouse present in a few databases [63,66,69]. LPI prediction for different protein isoforms is also not an active area of prediction algorithm development, with only one method having this functionality. Another limitation observed is that some methods exploit sequence similarity as an intermediate metric for LPI prediction, particularly methods which formulate LPI as similarity matrices. While this appears to be effective within the specific training datasets used by each study, this implicit assumption of similar sequence homology correlating to interactivity may not always hold, especially across different species [78,79]. At the same time, we consider that small nucleotide changes in biological molecules can cause major functional changes, which can potentially cause improperly trained prediction algorithms to produce misleading results [80].

We also note the limited accessibility of many of these machine learning methods. Among the methods reviewed that were published within the last five years, many do not make their source code publicly available and/or are written in proprietary programming languages such as MATLAB [81]. This restricts reproducibility and prevents usage of more than half of the methods we reviewed (Table 2). At least partly because of the computational complexity required, machine learning methods which are well suited to resolving non-linear variables in high dimensional data have recently become a focus of the LPI field. Computational methods that both identify and functionally annotate LPI are limited, leaving a gap in the field.

In contrast to published molecular docking algorithms, only a few machine learning methods provide active web servers for convenient use by the community, further raising the barrier for usability by biologists.

5. Future Directions

Computational surveying is not a substitute for experimental validation. However, as the intention of computational modelling is to generate a subset of the most likely testable hypotheses for laboratory biologists, we believe that developments in both the laboratory and computational fields will complement each other. With computational modelling reducing the quantity of experiments required, and with the experimentally validated data generated as a result, more efficient algorithms can be developed which further reinforces the developmental cycle. As a result, biologists interested in LPI will gain access to more refined tools, allowing them to streamline their experiments.

6. Conclusions

LPI forms a unique layer of gene regulation across many species, and a growing interest in the field has resulted in the creation and expansion of curated databases as well as LPI prediction algorithms. Here, we are reviewing some of the established (older than five years) and recent (within the last five years) LPI prediction approaches as well as databases. We note four important points. First, there has been a recent shift from conventional molecular docking algorithms to machine learning methods, which attempt the direct prediction of LPI from biomolecular sequence identity and higher-level features. This shift to machine learning is observable across different fields of biology and is likely to continue with the rising availability of computational infrastructure as well as machine learning expertise. Secondly, these methods are heavily dependent on a set of curated data across several databases. Across these databases, a lack of universal standardisation complicates data merging [82], preventing the community from unlocking the full potential of LPI data, in contrast to conventional transcriptomics databases such as SRA [83], EBI [84] and DDBJ [85]. This is in part due to the diversity of assays used to capture the LPI information, as well as the scope of the databases, which may subsequently bias the machine learning algorithms developed on these data. Third, there is a distinct lack of methods and databases which are specifically designed for LPIs' unique properties, with most having a generic scope despite LPIs' biological significance. Finally, it is concerning that more than half of the recent machine learning methods we surveyed are not reproducible or usable due to the absence of or restrictions on their source code. However, LPI act as an important but less-studied regulatory layer and understanding them will provide key context to deepen our understanding of biological systems.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ncrna7020033/s1. Table S1: lncRNA-protein data repositories. Seven databases, four with LPI information and three with RNA motif information, are surveyed. Each database holds information on at least one combination of nucleic acid and protein interaction. The number of species each database contains varies widely, from 4–154. Every database contains at least human and mouse data, and has been updated within the past five years, Table S2: LPI database recommendation matrix. Seven databases are analysed with respect to lncRNAs, namely, NEAT1, MALAT1 and Hotair (well studied) versus Lassie and MaTAR25 (less explored). Each database includes information on NEAT1, MALAT1 and Hotair and there are no data available regarding Lassie and MaTAR25.

Author Contributions: Conceptualisation, S.T.; methodology, M.P., T.C. and S.T.; formal analysis, M.P., T.C. and S.T.; resources, S.T.; data curation, M.P. and T.C.; writing—original draft preparation, M.P. and T.C.; writing—review and editing, M.P., T.C. and S.T.; visualisation, T.C. and M.P.; supervision, S.T.; project administration, S.T.; funding acquisition, S.T. All authors have read and agreed to the published version of the manuscript.

Funding: S.T. acknowledges the AISRF EMCR Fellowship by the Australian Academy of Science and Australian Women Research Success Grant at Monash University. T.C. received funding from the Australian Government Research Training Program Scholarship and Monash Faculty of Science Dean's Postgraduate Research Scholarship.

Acknowledgments: We thank Yashpal Ramakrishnaiah for his proofreading and feedback on this article. The authors thank the HPC team at Monash eResearch Centre for their continuous personnel support. This work was supported by the www.massive.org.au (accessed on 27 May 2021) MASSIVE HPC facility. We acknowledge and pay respects to the Elders and Traditional Owners of the land on which our four Australian campuses stand. https://www.monash.edu/indigenous-australians/about-us/recognising-traditional-owners (accessed on 27 May 2021).

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

References

1. Lowe, R.; Shirley, N.; Bleackley, M.; Dolan, S.; Shafee, T. Transcriptomics technologies. *PLoS Comput. Biol.* **2017**, *13*, e1005457. [CrossRef] [PubMed]
2. Guo, J.-C.; Fang, S.-S.; Wu, Y.; Zhang, J.-H.; Chen, Y.; Liu, J.; Wu, B.; Wu, J.-R.; Li, E.-M.; Xu, L.-Y.; et al. CNIT: A fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. *Nucleic Acids Res.* **2019**, *47*, W516–W522. [CrossRef] [PubMed]
3. Wang, C.; Wang, L.; Ding, Y.; Lu, X.; Zhang, G.; Yang, J.; Zheng, H.; Wang, H.; Jiang, Y.; Xu, L. LncRNA Structural Characteristics in Epigenetic Regulation. *Int. J. Mol. Sci.* **2017**, *18*, 2659. [CrossRef] [PubMed]
4. Kazimierczyk, M.; Kasprowicz, M.K.; Kasprzyk, M.E.; Wrzesinski, J. Human Long Noncoding RNA Interactome: Detection, Characterization and Function. *Int. J. Mol. Sci.* **2020**, *21*, 1027. [CrossRef] [PubMed]
5. Jalali, S.; Bhartiya, D.; Lalwani, M.K.; Sivasubbu, S.; Scaria, V. Systematic Transcriptome Wide Analysis of lncRNA-miRNA Interactions. *PLoS ONE* **2013**, *8*, e53823. [CrossRef] [PubMed]
6. Li, J.; Chen, Y.; Xu, X.; Jones, J.; Tiwari, M.; Ling, J.; Wang, Y.; Harismendy, O.; Sen, G.L. HNRNPK maintains epidermal progenitor function through transcription of proliferation genes and degrading differentiation promoting mRNAs. *Nat. Commun.* **2019**, *10*, 1–14. [CrossRef] [PubMed]
7. Fang, Y.; Fullwood, M.J. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genom. Proteom. Bioinform.* **2016**, *14*, 42–54. [CrossRef] [PubMed]
8. Piccolo, L.L.; Mochizuki, H.; Nagai, Y. The lncRNA hsrω regulates arginine dimethylation of FUS to cause its proteasomal degradation in Drosophila. *J. Cell Sci.* **2019**, *132*, jcs.236836. [CrossRef]
9. Militti, C.; Maenner, S.; Becker, P.; Gebauer, F. UNR facilitates the interaction of MLE with the lncRNA roX2 during Drosophila dosage compensation. *Nat. Commun.* **2014**, *5*, 4762. [CrossRef]
10. Bardou, F.; Ariel, F.; Simpson, C.G.; Romero-Barrios, N.; Laporte, P.; Balzergue, S.; Brown, J.W.; Crespi, M. Long Noncoding RNA Modulates Alternative Splicing Regulators in Arabidopsis. *Dev. Cell* **2014**, *30*, 166–176. [CrossRef]
11. Rigo, R.; Bazin, J.; Romero-Barrios, N.; Moison, M.; Lucero, L.; Christ, A.; Benhamed, M.; Blein, T.; Huguet, S.; Charon, C.; et al. The Arabidopsis lnc RNA ASCO modulates the transcriptome through interaction with splicing factors. *EMBO Rep.* **2020**, *21*, e48977. [CrossRef]
12. Zhao, X.; Li, J.; Lian, B.; Gu, H.; Li, Y.; Qi, Y. Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. *Nat. Commun.* **2018**, *9*, 1–12. [CrossRef]
13. Huang, C.; Zhu, B.; Leng, D.; Ge, W.; Zhang, X.D. Long noncoding RNAs implicated in embryonic development in Ybx1 knockout zebrafish. *FEBS Open Bio* **2021**, *11*, 1259–1276. [CrossRef]
14. Zhao, T.; Cai, M.; Liu, M.; Su, G.; An, D.; Moon, B.; Lyu, G.; Si, Y.; Chen, L.; Lu, W. lncRNA 5430416N02Rik Promotes the Proliferation of Mouse Embryonic Stem Cells by Activating Mid1 Expression through 3D Chromatin Architecture. *Stem Cell Rep.* **2020**, *14*, 493–505. [CrossRef]
15. Li, N.; Yang, G.; Luo, L.; Ling, L.; Wang, X.; Shi, L.; Lan, J.; Jia, X.; Zhang, Q.; Long, Z.; et al. lncRNA THAP9-AS1 Promotes Pancreatic Ductal Adenocarcinoma Growth and Leads to a Poor Clinical Outcome via Sponging miR-484 and Interacting with YAP. *Clin. Cancer Res.* **2020**, *26*, 1736–1748. [CrossRef]

16. Liu, B.; Sun, L.; Liu, Q.; Gong, C.; Yao, Y.; Lv, X.; Lin, L.; Yao, H.; Su, F.; Li, D.; et al. A Cytoplasmic NF-κB Interacting Long Noncoding RNA Blocks IκB Phosphorylation and Suppresses Breast Cancer Metastasis. *Cancer Cell* **2015**, *27*, 370–381. [CrossRef]
17. Kim, S.H.; Kim, S.H.; Yang, W.I.; Yoon, S.O.; Kim, S.J. Association of the long non-coding RNA MALAT1 with the polycomb repressive complex pathway in T and NK cell lymphoma. *Oncotarget* **2017**, *8*, 31305–31317. [CrossRef]
18. Turjya, R.R.; Khan, A.-A.-K.; Islam, A.B.M.M.K. Perversely expressed long noncoding RNAs can alter host response and viral proliferation in SARS-CoV-2 infection. *Futur. Virol.* **2020**, *15*, 577–593. [CrossRef]
19. Laha, S.; Saha, C.; Dutta, S.; Basu, M.; Chatterjee, R.; Ghosh, S.; Bhattacharyya, N.P. In silico analysis of altered expression of long non-coding RNA in SARS-CoV-2 infected cells and their possible regulation by STAT1, STAT3 and interferon regulatory factors. *Heliyon* **2021**, *7*, e06395. [CrossRef]
20. Zhao, H.; Shi, J.; Zhang, Y.; Xie, A.; Yu, L.; Zhang, C.; Lei, J.; Xu, H.; Leng, Z.; Li, T.; et al. LncTarD: A manually-curated database of experimentally-supported functional lncRNA–target regulations in human diseases. *Nucleic Acids Res.* **2019**, *48*, D118–D126. [CrossRef]
21. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef] [PubMed]
22. Ramanathan, M.; Porter, D.F.; Khavari, P.A. Methods to study RNA–protein interactions. *Nat. Methods* **2019**, *16*, 225–234. [CrossRef] [PubMed]
23. Faoro, C.; Ataide, S.F. Ribonomic approaches to study the RNA-binding proteome. *FEBS Lett.* **2014**, *588*, 3649–3664. [CrossRef] [PubMed]
24. Ramanathan, M.; Majzoub, K.; Rao, D.; Neela, P.H.; Zarnegar, B.J.; Mondal, S.; Roth, J.; Gai, H.; Kovalski, J.R.; Siprashvili, Z.; et al. RNA–protein interaction detection in living cells. *Nat. Methods* **2018**, *15*, 207–212. [CrossRef] [PubMed]
25. Kretz, M.; Siprashvili, Z.; Chu, C.; Webster, D.; Zehnder, A.; Qu, K.; Lee, C.S.; Flockhart, R.J.; Groff, A.F.; Chow, J.; et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nat. Cell Biol.* **2012**, *493*, 231–235. [CrossRef]
26. Simon, M.D.; Wang, C.I.; Kharchenko, P.V.; West, J.A.; Chapman, B.A.; Alekseyenko, A.A.; Borowsky, M.L.; Kuroda, M.I.; Kingston, R.E. Te genomic binding sites of a noncoding RNA. *Proc. Natl Acad. Sci. USA* **2011**, *108*, 20497–20502. [CrossRef]
27. Chu, C.; Qu, K.; Zhong, F.; Artandi, S.E.; Chang, H.Y. Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. *Mol. Cell* **2011**, *44*, 667–678. [CrossRef]
28. Tsai, B.P.; Wang, X.; Huang, L.; Waterman, M.L. Quantitative profiling of in vivo–assembled RNA-protein complexes using a novel integrated proteomic approach. *Mol. Cell. Proteom.* **2011**, *10*, M110.007385. [CrossRef]
29. Zeng, F.; Peritz, T.; Kannanayakal, T.J.; Kilk, K.; Eiríksdóttir, E.; Langel, Ü.; Eberwine, J. A protocol for PAIR: PNA-assisted identification of RNA binding proteins in living cells. *Nat. Protoc.* **2006**, *1*, 920–927. [CrossRef]
30. McHugh, C.A.; Guttman, M. RAP-MS: A Method to Identify Proteins that Interact Directly with a Specific RNA Molecule in Cells. *Methods Mol. Biol.* **2018**, *1649*, 473–488. [CrossRef]
31. Matia-González, A.M.; Iadevaia, V.; Gerber, A.P. A versatile tandem RNA isolation procedure to capture in vivo formed mRNA-protein complexes. *Methods* **2017**, *118–119*, 93–100. [CrossRef]
32. Ule, J.; Jensen, K.B.; Ruggiu, M.; Mele, A.; Ule, A.; Darnell, R. CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science* **2003**, *302*, 1212–1215. [CrossRef]
33. Kim, B.; Kim, V.N. fCLIP-seq for transcriptomic footprinting of dsRNA-binding proteins: Lessons from DROSHA. *Methods* **2019**, *152*, 3–11. [CrossRef]
34. Nicholson, C.O.; Friedersdorf, M.B.; Keene, J.D. Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq. *RNA* **2016**, *23*, 32–46. [CrossRef]
35. McMahon, A.; Rahman, R.; Jin, H.; Shen, J.L.; Fieldsend, A.; Luo, W.; Rosbash, M. TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell* **2016**, *165*, 742–753. [CrossRef]
36. Quinodoz, S.; Guttman, M. Long noncoding RNAs: An emerging link between gene regulation and nuclear organization. *Trends Cell Biol.* **2014**, *24*, 651–663. [CrossRef]
37. Ulitsky, I. Interactions between short and long noncoding RNAs. *FEBS Lett.* **2018**, *592*, 2874–2883. [CrossRef]
38. Ramakrishnaiah, Y.; Kuhlmann, L.; Tyagi, S. Towards a comprehensive pipeline to identify and functionally annotate long noncoding RNA (lncRNA). *Comput. Biol. Med.* **2020**, *127*, 104028. [CrossRef]
39. Li, J.-H.; Liu, S.; Zhou, H.; Qu, L.-H.; Yang, J.-H. starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **2014**, *42*, D92–D97. [CrossRef]
40. Hu, B.; Yang, Y.; Huang, Y.; Zhu, Y.; Lu, Z.J. POSTAR: A platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.* **2017**, *45*, D104–D114. [CrossRef]
41. Junge, A.; Refsgaard, J.C.; Garde, C.; Pan, X.; Santos, A.; Alkan, F.; Anthon, C.; von Mering, C.; Workman, C.T.; Jensen, L.J.; et al. RAIN: RNA-protein Association and Interaction Networks. *Database* **2017**, *2017*. [CrossRef]
42. Lin, Y.; Liu, T.; Cui, T.; Wang, Z.; Zhang, Y.; Tan, P.; Huang, Y.; Yu, J.; Wang, D. RNAInter in 2020: RNA interactome repository with increased coverage and annotation. *Nucleic Acids Res.* **2020**, *48*, D189–D197. [CrossRef]
43. Teng, X.; Chen, X.; Xue, H.; Tang, Y.; Zhang, P.; Kang, Q.; Hao, Y.; Chen, R.; Zhao, Y.; He, S. NPInter v4.0: An integrated database of ncRNA interactions. *Nucleic Acids Res.* **2019**, *48*, D160–D165. [CrossRef]
44. Giudice, G.; Sánchez-Cabo, F.; Torroja, C.; Lara-Pezzi, E. ATtRACT—A database of RNA-binding proteins and associated motifs. *Database* **2016**, *2016*. [CrossRef]

45. Bouvrette, L.P.B.; Bovaird, S.; Blanchette, M.; Lécuyer, E. oRNAment: A database of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic Acids Res.* **2019**, *48*, D166–D173. [CrossRef]
46. Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Drug Des.* **2011**, *7*, 146–157. [CrossRef]
47. Suravajhala, R.; Gupta, S.; Kumar, N.; Suravajhala, P. Deciphering lncRNA-protein interactions using docking complexes. *J. Biomol. Struct. Dyn.* **2020**, 1–8. [CrossRef]
48. Huang, Y.; Li, H.; Xiao, Y. 3dRPC: A web server for 3D RNA–protein structure prediction. *Bioinformatics* **2017**, *34*, 1238–1240. [CrossRef]
49. Ghoorah, A.W.; Devignes, M.-D.; Smaïl-Tabbone, M.; Ritchie, D.W. Protein docking using case-based reasoning. *Proteins Struct. Funct. Bioinform.* **2013**, *81*, 2150–2158. [CrossRef]
50. Andrusier, N.; Nussinov, R.; Wolfson, H.J. FireDock: Fast interaction refinement in molecular docking. *Proteins Struct. Funct. Bioinform.* **2007**, *69*, 139–159. [CrossRef]
51. van Zundert, G.; Rodrigues, J.; Trellet, M.; Schmitz, C.; Kastritis, P.; Karaca, E.; Melquiond, A.; van Dijk, M.; de Vries, S.; Bonvin, A.M. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428*, 720–725. [CrossRef] [PubMed]
52. Duhovny, D.; Nussinov, R.; Wolfson, H.J. Efficient Unbound Docking of Rigid Molecules. In *Algorithms in Bioinformatics, Proceedings of Second International Workshop, WABI 2002, Rome, Italy, 17-21 September 2002*; Guigo, R., Gusfield, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2002; Volume 2452, pp. 185–200.
53. Yan, Y.; Tao, H.; He, J.; Huang, S.-Y. The HDOCK server for integrated protein–protein docking. *Nat. Protoc.* **2020**, *15*, 1829–1852. [CrossRef] [PubMed]
54. He, J.; Tao, H.; Huang, S.-Y. Protein-ensemble–RNA docking by efficient consideration of protein flexibility through homology models. *Bioinformatics* **2019**, *35*, 4994–5002. [CrossRef] [PubMed]
55. Zheng, J.; Hong, X.; Xie, J.; Tong, X.; Liu, S. P3DOCK: A protein–RNA docking webserver based on template-based and template-free docking. *Bioinformatics* **2019**, *36*, 96–103. [CrossRef]
56. Tuszynska, I.; Magnus, M.; Jonak, K.; Dawson, W.; Bujnicki, J.M. NPDock: A web server for protein–nucleic acid docking. *Nucleic Acids Res.* **2015**, *43*, W425–W430. [CrossRef]
57. Chen, T.; Tyagi, S. Integrative computational epigenomics to build data-driven gene regulation hypotheses. *GigaScience* **2020**, *9*. [CrossRef]
58. Shen, C.; Ding, Y.; Tang, J.; Guo, F. Multivariate information fusion with fast kernel learning to kernel ridge regression in predicting lncrna-protein interactions. *Front. Genet.* **2019**, *9*, 716. [CrossRef]
59. Shen, C.; Ding, Y.; Tang, J.; Jiang, L.; Guo, F. LPI-KTASLP: Prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information. *IEEE Access* **2019**, *7*, 13486–13496. [CrossRef]
60. Liu, H.; Ren, G.; Hu, H.; Zhang, L.; Ai, H.; Zhang, W.; Zhao, Q. LPI-NRLMF: lncRNA-protein interaction prediction by neighborhood regularized logistic matrix factorization. *Oncotarget* **2017**, *8*, 103975–103984. [CrossRef]
61. Xie, G.; Wu, C.; Sun, Y.; Fan, Z.; Liu, J. LPI-IBNRA: Long Non-coding RNA-Protein Interaction Prediction Based on Improved Bipartite Network Recommender Algorithm. *Front. Genet.* **2019**, *10*, 343. [CrossRef]
62. Zhao, Q.; Yu, H.; Ming, Z.; Hu, H.; Ren, G.; Liu, H. The Bipartite Network Projection-Recommended Algorithm for Predicting Long Non-coding RNA-Protein Interactions. *Mol. Ther. Nucleic Acids* **2018**, *13*, 464–471. [CrossRef]
63. Hu, H.; Zhang, L.; Ai, H.; Zhang, H.; Fan, Y.; Zhao, Q.; Liu, H. HLPI-Ensemble: Prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol.* **2018**, *15*, 797–806. [CrossRef]
64. Xiao, Y.; Zhang, J.; Deng, L. Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks. *Sci. Rep.* **2017**, *7*, 3664. [CrossRef]
65. Zhao, Q.; Zhang, Y.; Hu, H.; Ren, G.; Zhang, W.; Liu, H. IRWNRLPI: Integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. *Front. Genet.* **2018**, *9*, 239. [CrossRef]
66. Zhang, W.; Yue, X.; Tang, G.; Wu, W.; Huang, F.; Zhang, X. SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PLoS Comput. Biol.* **2018**, *14*, e1006616. [CrossRef]
67. Wekesa, J.S.; Meng, J.; Luan, Y. A deep learning model for plant lncRNA-protein interaction prediction with graph attention. *Mol. Genet. Genom.* **2020**, *295*, 1091–1102. [CrossRef]
68. Zhang, W.; Qu, Q.; Zhang, Y.; Wang, W. The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* **2018**, *273*, 526–534. [CrossRef]
69. Fan, X.-N.; Zhang, S.-W. LPI-BLS: Predicting lncRNA-protein interactions with a broad learning system-based stacked ensemble classifier. *Neurocomputing* **2019**, *370*, 88–93. [CrossRef]
70. Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **2015**, *33*, 831–838. [CrossRef]
71. Zhang, S.W.; Zhang, X.X.; Fan, X.N.; Li, W.N. LPI-CNNCP: Prediction of lncRNA-protein interactions by using convolutional neural network with the copy-padding trick. *Anal. Biochem.* **2020**, *601*, 113767. [CrossRef]
72. Shaw, D.; Chen, H.; Xie, M.; Jiang, T. DeepLPI: A multimodal deep learning method for predicting the interactions between lncRNAs and protein isoforms. *BMC Bioinform.* **2021**, *22*, 1–22. [CrossRef] [PubMed]

73. Ma, Y.; He, T.; Jiang, X. Projection-Based Neighborhood Non-Negative Matrix Factorization for lncRNA-Protein Interaction Prediction. *Front. Genet.* **2019**, *10*, 1148. [CrossRef] [PubMed]
74. Zhou, Y.-K.; Shen, Z.-A.; Yu, H.; Luo, T.; Gao, Y.; Du, P.-F. Predicting lncRNA-protein Interactions With miRNAs as Mediators in a Heterogeneous Network Model. *Front. Genet.* **2020**, *10*, 1341. [CrossRef] [PubMed]
75. Zhou, Y.-K.; Hu, J.; Shen, Z.-A.; Zhang, W.-Y.; Du, P.-F. LPI-SKF: Predicting lncRNA-Protein Interactions Using Similarity Kernel Fusions. *Front. Genet.* **2020**, *11*. [CrossRef]
76. Lu, Q.; Ren, S.; Lu, M.; Zhang, Y.; Zhu, D.; Zhang, X.; Li, T. Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genom.* **2013**, *14*, 651. [CrossRef]
77. Agostini, F.; Zanzoni, A.; Klus, P.; Marchese, D.; Cirillo, D.; Tartaglia, G.G. catRAPIDomics: A web server for large-scale prediction of protein–RNA interactions. *Bioinformatics* **2013**, *29*, 2928–2930. [CrossRef]
78. Jacq, C.; Miller, J.; Brownlee, G. A pseudogene structure in 5S DNA of Xenopus laevis. *Cell* **1977**, *12*, 109–120. [CrossRef]
79. Lou, W.; Ding, B.; Fu, P. Pseudogene-Derived lncRNAs and Their miRNA Sponging Mechanism in Human Cancer. *Front. Cell Dev. Biol.* **2020**, *8*, 85. [CrossRef]
80. Denning, G.M.; Anderson, M.P.; Amara, J.F.; Marshall, J.; Smith, A.E.; Welsh, M. Processing of mutant cystic fibrosis transmembrane conductance regulator is temperature-sensitive. *Nat. Cell Biol.* **1992**, *358*, 761–764. [CrossRef]
81. *MATLAB.version 7.10.0 (R2010a)*; The MathWorks Inc.: Natick, MA, USA, 2010. Available online: https://www.mathworks.com/products/matlab.html (accessed on 27 May 2021).
82. Ramakrishnaiah, Y.; Kuhlmann, L.; Tyagi, S. Linc2function: A deep learning model to identify and assign function to long noncoding RNA (lncRNA). *bioRxiv* **2021**. Available online: https://www.biorxiv.org/content/10.1101/2021.01.29.428785v1.abstract (accessed on 27 May 2021). [CrossRef]
83. Leinonen, R.; Sugawara, H.; Shumway, M. On behalf of the International Nucleotide Sequence Database Collaboration the Sequence Read Archive. *Nucleic Acids Res.* **2010**, *39*, D19–D21. [CrossRef]
84. RNAcentral Consortium; Sweeney, B.A.; Petrov, A.I.; Ribas, C.E.; Finn, R.D.; Bateman, A.; Szymanski, M.; Karlowski, W.M.; Seemann, S.E.; Gorodkin, J.; et al. RNAcentral 2021: Secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* **2021**, *49*, D212–D220. [CrossRef]
85. Ogasawara, O.; Kodama, Y.; Mashima, J.; Kosuge, T.; Fujisawa, T. DDBJ Database updates and computational infrastructure enhancement. *Nucleic Acids Res.* **2019**, *48*, D45–D50. [CrossRef]