Getting closer to the whole picture of the LINE-1 protein complex

Valentina Chernova, Olga Khasina, Kat Poliakova-Georgantas, Daria Repkina, Ksenia Meteleva, Marc Masramon, and Dmitry Alexeev, Ilya Altukhov, John LaCava

Abstract

LINE-1 (Long Interspersed Nuclear Element-1; L1) is a mobile genetic element in the human genome that replicates by a copy and paste mechanism known as retrotransposition. ORF1p and ORF2p proteins, encoded by the L1 gene, form protein complexes with host proteins, and these complexes, which remain poorly characterized, participate in the L1 lifecycle. We studied the protein-protein interactions formed with L1 using the String and BioPlex databases.

We have developed an algorithm that allowed us to determine which proteins interact with one another and where they were localised relative to the complex centre. We also characterised putative physical connections that are not currently annotated in the above databases.

Introduction

Approximately half of the human genome consists of repetitive DNA sequences attributable to retro-elements. ~20% of the genome is made up of L1 sequences, specifically. Notably, L1 has been implicated in the development of the nervous system, and L1 is observed to be expressed in over half of all cancers. It is estimated that in one in twenty human births contains a novel L1 mobilised insertion.

A full-length L1 transcript is ~6 knt long and functions as a bicistronic mRNA, and codes two proteins - ORF1 and ORF2, which in turn are RNA-binding proteins with chaperone activity and a multifunctional endonuclease functionality.

We have used the recent work of Molloy et al. [1], to evaluate L1 protein complexes that may occur in living cells. Molloy et al. define a set of putative protein complexes based on protein co-behaviors across a number of affinity proteomic experiments.

Materials and methods

LINE-1 protein clusters

During our work, we used the protein clusters of the LINE-1 protein complex outlined in the KR Molloy et al. manuscript [2]. Collectively the clusters contained 31 proteins.

I-DIRT experiments

24 I-DIRT experiments at different conditions were collected from the unpublished data of J. LaCava as a table of affinity values for 886 proteins; 21 were found in the clusters.

PPI databases

During our work, we used two databases to identify and describe PPIs within our clusters: String DB [3] and BioPlex V.2.0 [4].

In the String database, we considered only experimentally supported protein interactions or ones in other curated databases at medium confidence (0.4).

Network visualisation

The networks formed from I-DIRT data were visualised using Cytoscape [5]. The correlation values and number of approaching edges were mapped to edge width and node size respectively.

Identification of new PPI algorithm

RStudio Version 1.0.153 was used to build a new cluster identification and directional interaction algorithms. Raw data (37 proteins) from J. La Cava's experiments provided the affinity of proteins present in variable conditions. The presence or absence of proteins were defined as 1 and 0. Correlation was calculated for each possible pair of proteins based on their '1' values.

Protein-protein reliance was determined by comparing all possible protein pairs and proteins were labeled as 'source' nodes (more peripheral) and then compared based on their '0' and '1' values. If the 'source' protein was present for every instance of the 'target' (more central) protein's presence, then that protein pair was saved in a data frame. That data frame along with the correlation data was used to create a network displaying directed protein interaction by correlation strength.

Results and discussion

Clusters' Functional Characterisation (see fig. 1)

Cluster 1: They are involved in the cell growth cycle. In the String database, the ORF2 protein is absent, while all the other proteins are interconnected. BioPlex does not have data on ORF2, HSPAA1, and TUBB4B.

Cluster 2: They are associated with RNA binding, mRNA degradation and translational control. In String DB, only ORF1 is absent from the database, however few interactions are recorded - only UPF1 with PABPC4 (0.4 confidence) and UPF1 with MOV10 (0.125 confidence). BioPlex does not have data on UPF1, PABPC1, PABPC4, and ORF1.

Cluster 3: HSPA8 and HSPA1A are protein-folding chaperones and transcriptional repressors. MEPCE is a capping enzyme that stabilises snRNA and a negative regulator of RNA Pol II promoter binding. According to the String database, MEPCE does not interact with the HSPA8 and HSPA1A proteins, while both of them interact closely with one another. BioPlex does not have data on HSPA1A.

Cluster 4: They have assorted functions: nucleosome assembly, protein transport, mRNA degradation/stabilisation, apoptosis cascade mediation and mitochondrial protease activity. In the String database, only NAP1L1 and NAPL4 have recorded interactions. BioPlex does not have data on HAX1 and YME1L1.

Cluster 5: All proteins from this cluster enable the binding of proteins to DNA; some of them are also involved in DNA repair and replication.

In the String database, all of these proteins are present, however interactions are only shown between two pair of proteins - PURB and PURA, PCNA and PARP1.

As can be seen from the above data, the String and BioPlex databases are incomplete. Furthermore, the databases provide no directionality within the PPIs and little context as to the mechanisms of these complexes.

New PPI Identification and Protein Localisation

We developed an algorithm that allows the analysis of the associated I-DIRT datasets to be used to predict new PPIs and hence wider complexes. This allows us to better understand not only whether these relationships exist, but their localisation in the complex and the correlation with which they co-occur. This correlation frequency as shown in a histogram format can be seen in Figure.

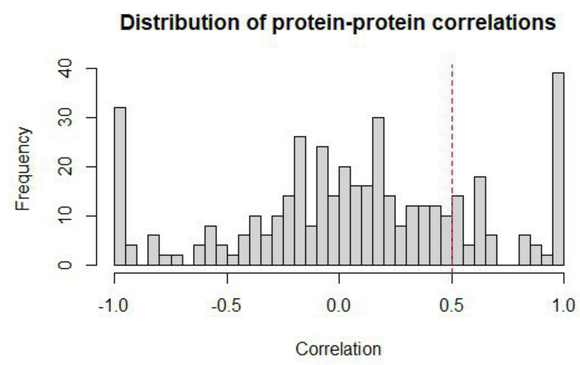

Figure 4: Histogram of correlation (x-axis) to its frequency (y-axis) in the PPIs generated by algorithm.

The algorithm generated 21 nodes (proteins) and 138 directional interactions of all correlation levels, with 44 of those at a correlation greater than 0.5. Analysing the generated network in Cytoscape provided us not only with an approximation of their distance from the centre of the complex, but also whether they were upstream or downstream from other members within the complex. This provides a measure of context and a more complete understanding of the nature of these complexes, which is simply not provided in conventional databases.

Not all proteins from the fifth cluster are present in the table, our algorithm is generated from. According our results, it is clear that the PARP1 protein is located at the centre, while LARP7 and PURA are on the periphery. We discovered two high-correlation interactions missing from the PPI databases - LARP7 and PARP1 (0.95), and PURA and PARP1 (0.8).

Whole Complex Description

The predicted relative localisation of proteins in the complex, as defined by their number of incoming edges, can be seen in Table 1. The fact that HAX1 is the most central protein according to this model is quite surprising, as the ORF1 protein would generally be anticipated to be so. Nevertheless, as the baiting was done with ORF2 proteins, it is entirely possible that ORF1 could have been removed into the solution at a faster rate than HAX1. The tendency to group by cluster in many cases is very interesting though and may indicate that localisation in the complex is a complementary influence on experimental behaviour along with functional relationships.

There are few high-correlation (>0.8) interactions between different clusters modelled by the network, bar the links to HIST1H2BO. These include HAX1 (Cluster 4) being upstream of ORF1 (Cluster 2) with 0.87 correlation; DDX6 (Cluster 4) being upstream of HSPA1A (Cluster 3) with 0.82 correlation; MOV10 (Cluster 2) being downstream of LARP7 (Cluster 5) with 0.99 correlation. Below that cut-off point, Cluster 1 proteins are very highly connected to PABPC4 at 0.65 correlation.

Conclusions and Perspectives

In light of the transition from "garbage in-garbage out" to big data philosophies, there comes a need to develop new algorithms for data analysis. The algorithm we have developed is one such example, however its potential use is not limited to our work on the LINE-1 retrotransposon, as the importance of interactomics research is becoming ever clearer.

The predictions and models we have generated have provided new avenues for research and proven their efficacy in this limited case, however without appropriate result validation with both "wet" and "dry" lab approaches, there is no way forward. Confirmatory experiments will have to conducted with much larger data-sets on both already-known and unknown protein complexes to ensure the algorithm's validity.

Nevertheless, further steps must be taken, for investigations into protein-protein interactions are a significant opportunity for medical developments, including the discovery of biomarkers for disease progression and potential drug targets. Interactomics is just at its outset and it is our hope that such algorithms can assist it in its growth.

Number of incoming edges	1	2	3	4	5	6	7	8	9	10
Proteins	PCNA			PURA	HSPA 1A	HSPA 8	LARP 7	DDX6	IPO7	HAX1
				PARP 1	HSPA 90AA 1	MOV1 0	TUBB 4B	UPF1	L1RE 1	
					HIST1 H2BO	ORF2		PABP C1		
						HSPA 90AB 1		PABP C4		
								TUBB		
								RPS2 7A		

Table 1: The higher the number of incoming edges, the closer the protein is predicted to be to the complex centre. The colours indicate different clusters.

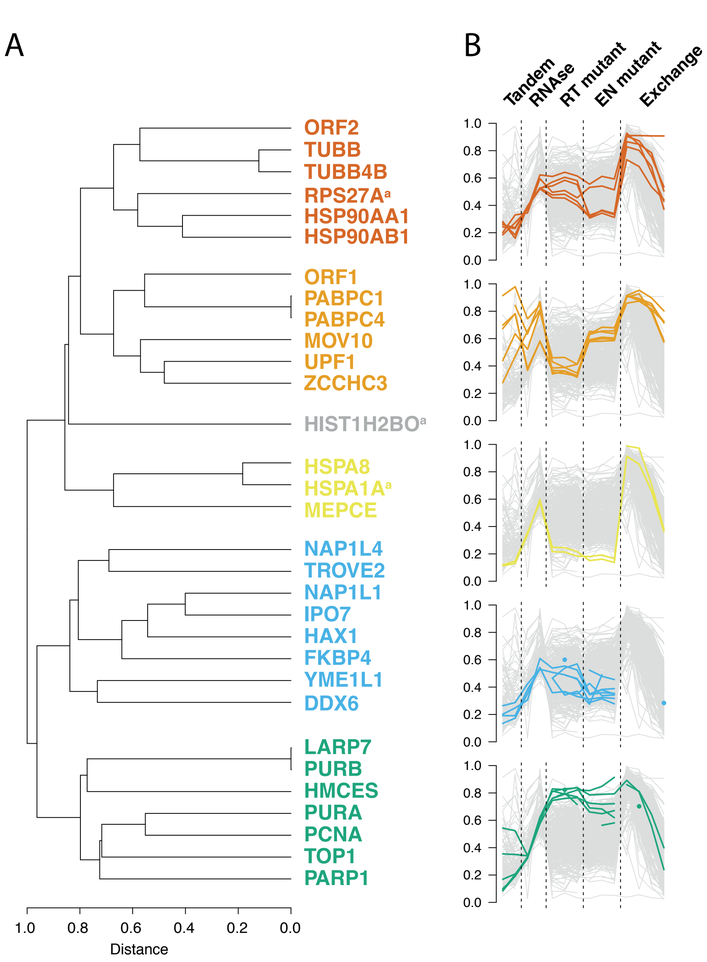

Figure 1: X-axis - different experimental set-ups; y-axis - affinity of the identified proteins during Western Blotting.

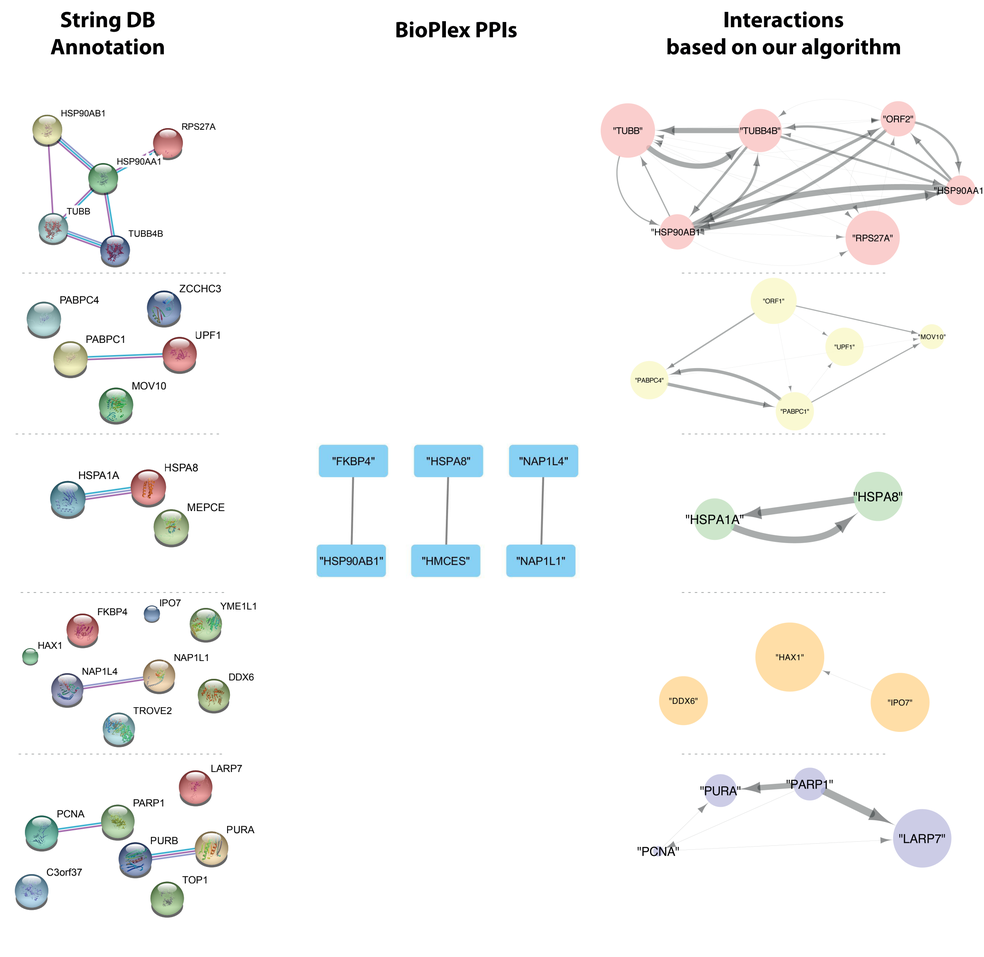

Figure 2: The interaction networks from String DB and BioPlex; medium confidence (0.4)

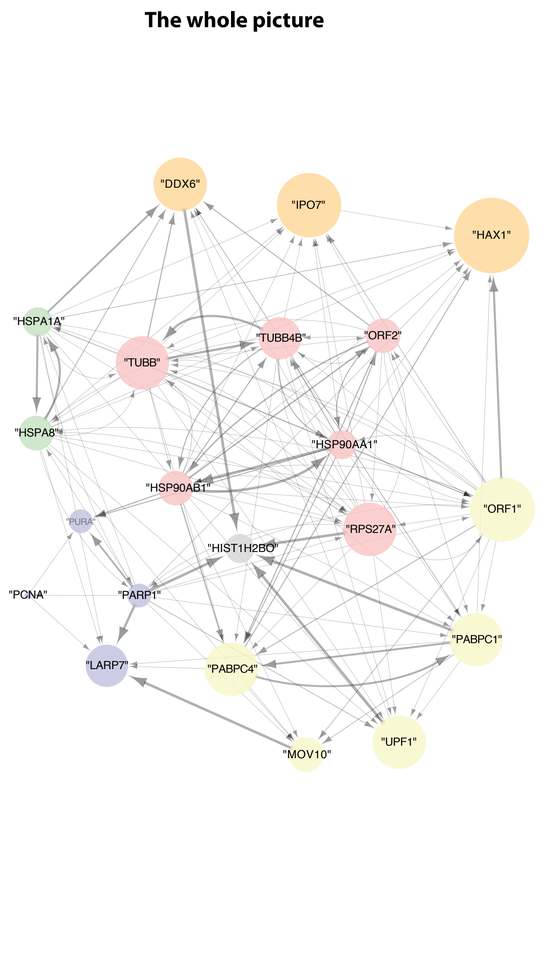

Figure 3: The networks generated from the created algorithm; initially per cluster and between clusters.
Thick to thin edge width detrmined by 0.5 correlation cut-off point. Node size determined by 'Outdegree' - number of incoming nodes.