RBPs mutations impact in lncRNAs

November 29, 2017

1 Project Details

Title: Mutation Analysis of RNA binding proteins, impact on lncRNAs.

Student: Alejandro Athie

Host: Institute of Biomedicine, Goteborg University,

Time period: May - July 2016

2 Background

During the last decade lncRNAs have emerged as new players in cancer biology. There are many examples of lncRNAs whose expression is altered and also multiple mechanisms have been proposed to explain the role they play in the disease. Most of the molecular mechanisms proposed so far relay on protein partners. Protein-RNA interactions are key to explain how lncRNAs exert their functions.

RNA-binding proteins alterations have been linked to several pathologies including cancer. It has been reported that expression of RBPs is altered in different tumor types. Recent studies showed that mutations at RBPs alter the splicing of protein coding genes. However, not much work has been done to analyze the impact of these mutations in the lncRNAs.

Novel experimental techniques, sucha as eCLIP, add valuable data to characterize RNA-protein interactions by mapping the RBPs binding sites to their RNA targets. Integration of all these new data will help us to better understand if mutations at the RNA binding proteins alter protein-lncRNAs interactions?

3 Goals

-Learn how to work with mutation data from the TCGA.

-Analyze RNA-seq data for the shRNAs.

-Analyze eCLIP data.

Answer the following question: Do mutations at the RNA binding proteins alter protein-lncRNAs interactions?

4 Working Plan

5 First month

1) Get a list of experimentally validated RNA binding proteins, by merging the two lists of the following publications:

   Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell , Volume 149 , Issue 6 , 1393 - 1406

   The mRNA-bound proteome and its global occupancy profile on proteincoding transcripts. Mol Cell. 2012;46(5):674–90.

2) Define which of them can act as drivers by intersecting with the following list of mutational cancer drivers:

   Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci Rep 3:2650.

   And compare with previously published list of RBPs altered in cancer:

   Dissecting the expression landscape of RNA-binding proteins in human cancers. Genome Biol. 15, R14 (2014). The list of Strongly upregulated RNA binding proteins SUR RBPs.(31)

3) Analyze the mutations present in this set of proteins, with particular interest in those present at the RNA binding domains.

4) Correlate the presence of the mutations in these proteins to the expression of lncRNAs, (the same as Arghavan's paper by adding these proteins to your list).

6 Second month

1) Analyze the eCLIP data for those RBPs released by ENCODE (K562, HepG2) data. The goal is to define a set of lncRNA targets bound by this protein and the presence of a binding motif, if it exist.

2) Check in the mutation data, the presence of mutations in the lncRNAs at the eCLIP binding motifs. These mutations could be altering the binding site.

7 Third month

1) Analyze the RNA-seq data for the shRNAs experiments of RBPs (K562, HepG2) from the ENCODE, to check for lncRNAs whose expression is altered upon the knockdown of these proteins.

2) Combine previous analysis to check for interesting lncRNA candidates.

3) Mutation status of binding site in the targets. Presence of a mutation that alters the binding site.

8 Progress Report Week ONE:

9 First goal

1) Get a list of experimentally validated RNA binding proteins, by merging the two lists of the following publications:

   Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell , Volume 149 , Issue 6 , 1393 - 1406.

   The mRNA-bound proteome and its global occupancy profile on proteincoding transcripts. Mol Cell. 2012;46(5):674–90.

```
In [5]: setwd("/Users/felipeathie/Desktop/RBPs/")
        install.packages("venneuler", repos = "http://ftp.acc.umu.se/mirror/CRAN/")
        library(venneuler)
```

```
also installing the dependency ‘rJava’




The downloaded source packages are in
        ‘/private/var/folders/mt/p9qx2w495c77xmh1hv9lmmkr0000gn/T/RtmpHhm9ZI/downloaded_packages


Updating HTML index of packages in '.Library'
Making 'packages.html' ... done
Loading required package: rJava
```

The first data set comes from 'Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell'. Contains 860 proteins that were identified in Hela cells to bind RNAs.

```
In [8]: RBPs_list_Cell<-read.table(file="RBPs_Cell.txt", sep="\t", header=TRUE)
        head(RBPs_list_Cell,1)
        tail(RBPs_list_Cell,1)
        dim(RBPs_list_Cell)
```

Out[8]:

	ENSEMBLID	Symbol	mRNAbinding	IdentifiedCL	IdentifiedControl	p.adj.DESeq
1	ENSG00000001497	LAS1L	mRNA-interactome	identified	not identified	0,000257809

Out[8]:

	ENSEMBLID	Symbol	mRNAbinding	IdentifiedCL	IdentifiedControl	p.adj.DESeq
860	ENSG00000249714	NA	mRNA-interactome	identified	not identified	0,002275245

Out[8]:

1. 860 2. 8

The second data set comes from 'The mRNA-bound proteome and its global occupancy profile on protein coding transcripts. Mol Cell' Sci Rep. 2013 Oct 2. Contains 799 proteins identified in HEK293 cells.

```
In [9]: RBPs_list_MolCell<-read.table(file="RBPs_MolCell.txt", sep="\t", header=TRUE)
        head(RBPs_list_MolCell,1)
        tail(RBPs_list_MolCell,1)
        dim(RBPs_list_MolCell)
```

Out[9]:

	id	Class	Protein.ID	RefSeq	Offical.gene.symbol	Uniprot	log2FC.exp
1	155	I	IPI00021570	NP$_0$03783	EDF1	O60869-1;O60869;O60869-2	-7,97290711

Out[9]:

	id	Class	Protein.ID	RefSeq	Offical.gene.symbol	Uniprot	log2FC.experiment.L1	log2FC.e
800	NA							

Out[9]:

1. 800 2. 20

By intersecting both lists using the Gene.symbol, 528 are shared in the two studies and can be classified as true RNA-binding proteins. While 1129 are the union of both lists. I will continue working with the two lists separately.

```
In [21]: RBPs_shared<-intersect(RBPs_list_Cell$Symbol, RBPs_list_MolCell$Offical.gene.symbol)
         length(RBPs_shared)
         head(RBPs_shared)
         RBPs_both<-union(RBPs_list_Cell$Symbol, RBPs_list_MolCell$Offical.gene.symbol)
         length(RBPs_both)
         head(RBPs_shared)
         v1<-venneuler(c(Cell_paper=860, Mol_Cell=799, "Cell_paper&Mol_Cell"=528))
         v1$labels<-c("")
         plot(v1, main="RBPs proteome")
         text(.2, .5, "Mol_Cell.799")
         text(.5, .5, "528")
         text(.8, .5, "Cell.860")
```

Out[21]:

528

Out[21]:

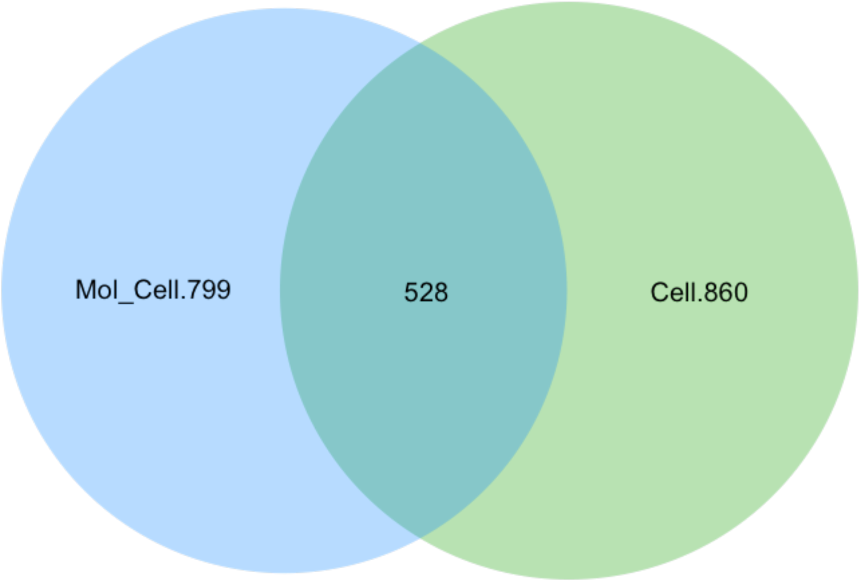
1. 'LAS1L' 2. 'RBM6' 3. 'UPF1' 4. 'ELAC2' 5. 'RPS20' 6. 'CSDE1'

Out[21]:

1129

Out[21]:

1. 'LAS1L' 2. 'RBM6' 3. 'UPF1' 4. 'ELAC2' 5. 'RPS20' 6. 'CSDE1'



10 Second Goal

To classify which of these proteins can act as a cancer drivers, the two lists of RBPs_(528, 1129) were intersected with a list of known mutational driver genes (435) coming from Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci Rep 3:2650.

```
In [23]: cancer_drivers<-read.table(file="srep02650-s3.csv", sep=",", header=TRUE)
         head(cancer_drivers,1)
         tail(cancer_drivers,1)
         dim(cancer_drivers)
```

```
   Out[23]:
```

	Gene.Symbol	Cancer.Gene.Census	MuSIC	OncodriveFM	OncodriveCLUST	ActiveDriver	Mu
1	AARS2						Sel

Out[23]:

	Gene.Symbol	Cancer.Gene.Census	MuSIC	OncodriveFM	OncodriveCLUST	ActiveDriver
435	ZNF703		Selected			

Out[23]:

1. 435 2. 8

In [30]:

```
#Intersection with list RBPs_shared
RBPs_shared_df<-data.frame(RBPs_shared)
RBPs_both_df<-data.frame(RBPs_both)

RBPs_shared_drivers<-merge(RBPs_shared_df, cancer_drivers, by.y="Gene.Symbol", by.x="RB
RBPs_both_drivers<-merge(RBPs_both_df, cancer_drivers, by.y="Gene.Symbol", by.x="RBPs_b

#36 in the shared list
head(RBPs_shared_drivers)
dim(RBPs_shared_drivers)

#64 in the both list
head(RBPs_both_drivers)
dim(RBPs_both_drivers)

#shared
v2<-venneuler(c(Drivers=435, shared=528, "Drivers&shared"=36))
v2$labels<-c("")
plot(v2, main="RBPs.shared Mutational cancer drivers")
text(.2, .5, "Driver.435")
text(.5, .5, "36")
text(.8, .5, "RBPs.Shared.528")


#both
v3<-venneuler(c(Drivers=435, both=1129, "Drivers&both"=64))
v3$labels<-c("")
plot(v3, main="RBPs.both Mutational cancer drivers")
text(.2, .5, "Driver.435")
text(.55, .5, "64")
text(.8, .5, "RBPs.Both.1129")
```

Out[30]:

	RBPs$_{s}hared$	Cancer.Gene.Census	MuSIC	OncodriveFM	OncodriveCLUST	ActiveDriver	Mu
1	BCLAF1		Selected			Selected	
2	CCAR1			Selected			
3	CNOT1			Selected		Selected	
4	CSDE1		Selected				
5	DDX3X			Selected			
6	DDX5	CGC					Sele

Out[30]:

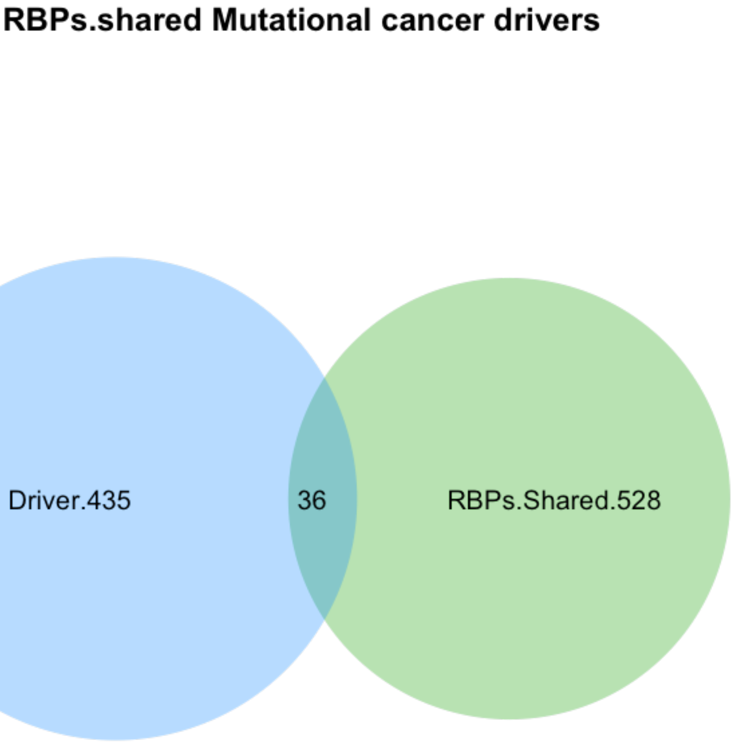
1. 36 2. 8

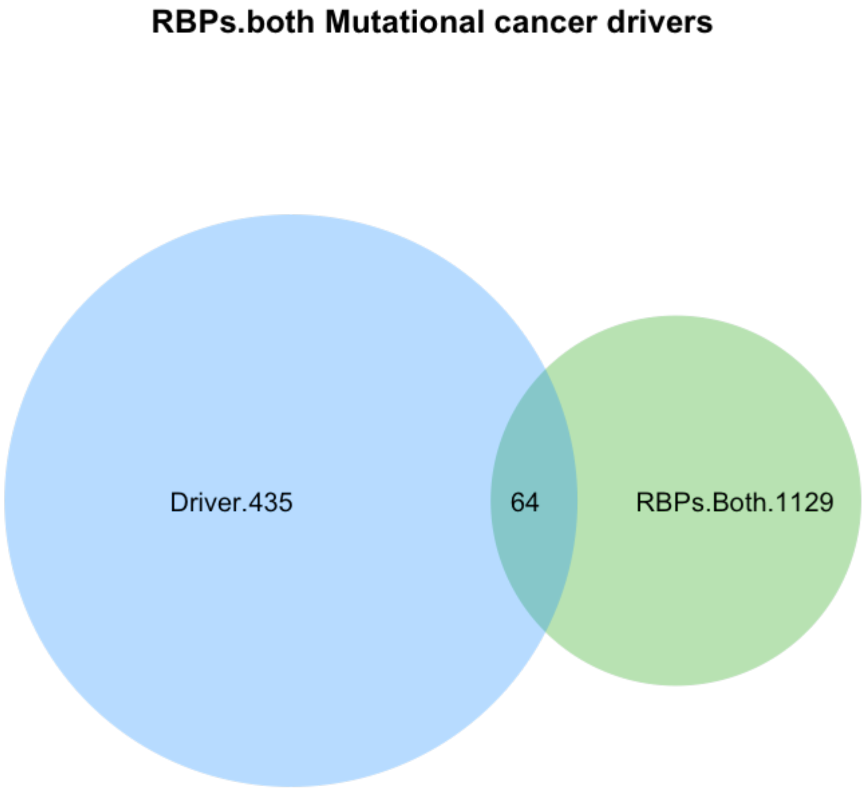
Out[30]:

	RBPs$_both$	Cancer.Gene.Census	MuSIC	OncodriveFM	OncodriveCLUST	ActiveDriver	MutS
1	AHNAK					Selected	
2	BCLAF1		Selected			Selected	
3	CAST			Selected		Selected	
4	CCAR1			Selected			
5	CDKN2A	CGC	Selected	Selected	Selected		Selec
6	CHD3			Selected			

Out[30]:

1. 64 2. 8





Compare with this other publication where they analyze the expression of RBPs altered in cancer, This will add not just the mutational status that make these proteins drivers but also expression data and prognosis value for some of them.

Dissecting the expression landscape of RNA-binding proteins in human cancers. Genome Biol. 15, R14 (2014). The list of Strongly upregulated RNA binding proteins SUR RBPs.(31)

```
In [31]: SUR_RBPs<-read.table(file="SUR_RBPs.txt", sep="\t", header=TRUE)
         head(SUR_RBPs,1)
         dim(SUR_RBPs)
         RBPs_shared_drivers_SUR<-merge(RBPs_shared_drivers, SUR_RBPs, by.y="Gene.Symbol", by.x=
         RBPs_both_drivers_SUR<-merge(RBPs_both_drivers, SUR_RBPs, by.y="Gene.Symbol", by.x="RBP
         dim(RBPs_shared_drivers_SUR)
         dim(RBPs_both_drivers_SUR)
```

Out[31]:

	Gene.Symbol
1	CCDC124

Out[31]:

1. 31 2. 1

Out[31]:

1. 6 2. 8

Out[31]:

1. 10 2. 8

Not so many less than 20% of the analyzed also appear to have altered expression. This are the strongest RBPs that can be classified as drivers based on its mutations and also have altered expression (upregulated).

```
In [32]: RBPs_both_drivers_SUR
```

Out[32]:

	RBPs$_{b}oth$	Cancer.Gene.Census	MuSIC	OncodriveFM	OncodriveCLUST	ActiveDriver	Mut
1	DDX3X			Selected			
2	FLNA					Selected	
3	GNL3L				Selected		
4	HUWE1					Selected	
5	NONO	CGC					Selec
6	PHF6	CGC	Selected	Selected			Selec
7	RBM10			Selected			
8	RBMX						Selec
9	RPGR						Selec
10	UTP14A					Selected	

Continue working just with the lists of RBPs intersected with the cancer drivers 36 shared and 64 both. Get the mutations for these RBPs

Take a look at the ENCODE data For downloading go to https://www.encodeproject.org/matrix/?type=Experiment&assay_title=eCLIP select the eCLIP experiments and download the file.txt contains the list of urls to download all the data, in total it contains 270 different samples (K562 cells 152 samples, HepG2 112 samples, 2 adrenal tissue samples)

I only downloaded the .bam and .bigBed files, (not the fastq.gz file) grep -v .*.fastq.gz eCLIP_files.txt > eCLIP_files_bigBed_bam.txt xargs -n 1 curl -O -L < eCLIP_files_bigBed_bam.txt

For downloading de shRNAs data go to https://www.encodeproject.org/matrix/?type=Experiment&assay_t seq select the shRNAs experiments and download the file.txt contains the list of urls to download all the data, in total it contains 445 different samples (K562 cells 227 samples, HepG2 218 samples)

I only downloaded the .bam and .bigBed files, (not the fastq.gz file) grep -v .*.fastq.gz shRNA_files.txt > shRNA_files_bigBed_bam.txt xargs -n 1 curl -O -L < shRNA_files_bigBed_bam.txt

Downloading..... takes a while.

Add a column to the 36 and 64 table if eCLIP is available and/or shRNA is available by merging with the gene.symbol

```
In [33]: eCLIP_protein_list<-read.table(file="eCLIP_proteins_list.tsv", sep="\t", header=FALSE)
         shRNA_protein_list<-read.table(file="shRNA_proteins_list.tsv", sep="\t", header=FALSE)
         names(eCLIP_protein_list)<-c("Gene.Symbol","ENCODE_eCLIP")
         names(shRNA_protein_list)<-c("Gene.Symbol","ENCODE_shRNA")
```

```
RBPs_shared_driver_eCLIP<-merge(RBPs_shared_drivers, eCLIP_protein_list, by.x="RBPs_sha
RBPs_both_driver_eCLIP<-merge(RBPs_both_drivers, eCLIP_protein_list, by.x="RBPs_both",

RBPs_shared_driver_eCLIP_shRNA<-merge(RBPs_shared_driver_eCLIP, shRNA_protein_list, by.
RBPs_both_driver_eCLIP_shRNA<-merge(RBPs_both_driver_eCLIP, shRNA_protein_list, by.x="R

index_shared<-(which(RBPs_shared_driver_eCLIP_shRNA$ENCODE_shRNA=='shRNA' & RBPs_shared
index_both<-(which(RBPs_both_driver_eCLIP_shRNA$ENCODE_shRNA=='shRNA' & RBPs_both_drive

RBPs_shared_driver_eCLIP_shRNA[index_shared,]

RBPs_both_driver_eCLIP_shRNA[index_both,]
```

Out[33]:

	RBPs$_shared$	Cancer.Gene.Census	MuSIC	OncodriveFM	OncodriveCLUST	ActiveDriver	Mu
5	DDX3X			Selected			
10	EWSR1	CGC					Sel
18	NONO	CGC					Sel
19	NPM1	CGC	Selected	Selected			Sel
24	PRPF8			Selected			
25	RBFOX2			Selected			
35	U2AF1	CGC	Selected	Selected	Selected		Sel

Out[33]:

	RBPs$_both$	Cancer.Gene.Census	MuSIC	OncodriveFM	OncodriveCLUST	ActiveDriver	MutS
10	DDX3X			Selected			
17	EWSR1	CGC					Selec
35	NONO	CGC					Selec
36	NPM1	CGC	Selected	Selected			Selec
43	PRPF8			Selected			
45	QKI			Selected			
46	RBFOX2			Selected			
61	U2AF1	CGC	Selected	Selected	Selected		Selec

By merging all the available RBPs/Drivers/eCLIP/shRNA: -->shared 7 DDX3X, EWSR1, NONO, NPM1, PRPF8, RBFOX2, U2AF1 -->both 8 DDX3X, EWSR1, NONO, NPM1, PRPF8, QKI, RBFOX2, U2AF1 The only difference between both lists is 'QK1'

RBFOX2, has already been deeply analyzed in the paper: Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nature Methods.

Check the numbers without filtering by the drivers the initial lists RBPs_shared_df RBPs_both_df They go up ten times more without the driver filtering RBPs_shared_df 65 RBPs_both_df 80

```
In [34]: RBPs_shared_df_eCLIP<-merge(RBPs_shared_df, eCLIP_protein_list, by.x="RBPs_shared", by.
         RBPs_both_df_eCLIP<-merge(RBPs_both_df, eCLIP_protein_list, by.x="RBPs_both", by.y="Gen

         RBPs_shared_df_eCLIP_shRNA<-merge(RBPs_shared_df_eCLIP, shRNA_protein_list, by.x="RBPs_
         RBPs_both_df_eCLIP_shRNA<-merge(RBPs_both_df_eCLIP, shRNA_protein_list, by.x="RBPs_both

         index_shared<-(which(RBPs_shared_df_eCLIP_shRNA$ENCODE_shRNA=='shRNA' & RBPs_shared_df_
         index_both<-(which(RBPs_both_df_eCLIP_shRNA$ENCODE_shRNA=='shRNA' & RBPs_both_df_eCLIP_
         length(index_shared)
         length(index_both)
```

Out[34]:

```
65
```

Out[34]:

```
80
```

Better to take the list of 65 RBPs (shared in Hela & HEK293T) that seem to be true RNA binder, and that have both shRNA and eCLIP data available keep in mind those 8 that were classified as mutational drivers (DDX3X, EWSR1, NONO, NPM1, PRPF8, QKI, RBFOX2, U2AF1)

!!Note: To save the code in pdf , first save it as a *ipynb and then convert it in the command line to* .tex and then to *.pdf jupyter nbconvert* .ipynb --to latex pdflatex *.tex