Report

Contents

Background

COVID-19 is a single-stranded positive RNA (+ss-RNA) virus from the betacoronavirus family, which also includes MERS-CoV, and SARS-CoV. Coronaviruses possess the largest genome (30 kb) of the RNA viruses, contributing to their complex genome regulation strategies. Once COVID-19 infects a human host cell, it makes use of the proteins and host machinery including human RNA binding proteins (RBPs) that bind to RNA for processing such as post transcriptional modifications, regulation of stability and degradation (Tay et al. 2020). Several human RBPs have been previously shown to directly interact with coronaviruses to regulate their transcription, replication and translation. The full range of RBPs that may interact with COVID-19 remain to be described. In order to make strides towards the development of therapeutic strategies for COVID-19 and to understand the pathology of the virus, it is important to study the interactions between the viral genome and RBPs. If an RBP is involved in COVID-19 replication, targeting it for inhibition may reduce the rate at which the virus reproduces and may yield better clinical results in patients. Previously, as part of BioHackathon 2020, Ferraini et al. (2020) identified a set of proteins predicted to bind to the SARS-CoV-2 genome in order to characterize host-virus interactions post-infection and suggest pharmaceutical interventions for COVID-19 patients. Their workflow used position-weight matrices to locate potential binding sites for human RBPs on the genomes of SARS-CoV-2, SARS-CoV-1, and RaTG13, as well as simulated genomes that shared the nucleotide compositions of the aforementioned viruses. The difference in number of binding sites for each RBP between real and simulated genomes was tested for statistical significance to identify enriched RBP binding sites, and these binding sites were compared between viruses. This analysis identified 38 human RBPs that bind with SARS-CoV-2's genome, a handful of which do so exclusively.

Introduction

To build upon the work by Ferraini et al., we will first reproduce the pipeline presented in the hackathon to obtain baseline results. This includes differential expression analysis to identify human mRNAs that are significantly up or down-regulated in COVID-19 infected cells. Since the release of the original hackathon results, new publications have revealed datasets that may be useful for modeling COVID-19 and human RBP interactions. For example, a study from December 2020 (Schmidt et al. 2020) published a list of 104 human RBPs they found to be experimentally interacting with COVID-19. Similarly, another study from 2021 mapped COVID-19 RBP interactions (Kamel et al. 2021). Thus, we plan on building upon previous methods by integrating known interactions for training a classifier that could identify COVID-19 interacting RBPs. Further, once we obtain a high confidence candidate list of RBPs, we can further evaluate their druggability to suggest human RBPs for further study (Gordon et al. 2020). We will divide the list of published COVID-19 interacting RBPs into training and testing sets to train a classifier using viral sequence information and differential expression results for RBP mRNAs. This model is likely to yield additional interacting candidates that have not yet been described

Analyse

Project composition and functions

They first created a file (rbp_functions) with all the functions they needed: - to read the different type of file (readGFF, readFasta, readPWMsFromFasta), - to calculate PWM information content (GetPWMEntropy) - to scan sequence(s) with multiple PWMs (ScanSeqWithPWMs) - to simulate a DNA sequence (simulateSeq) - to do enrichment test for RBP (enrich_rbps_real)

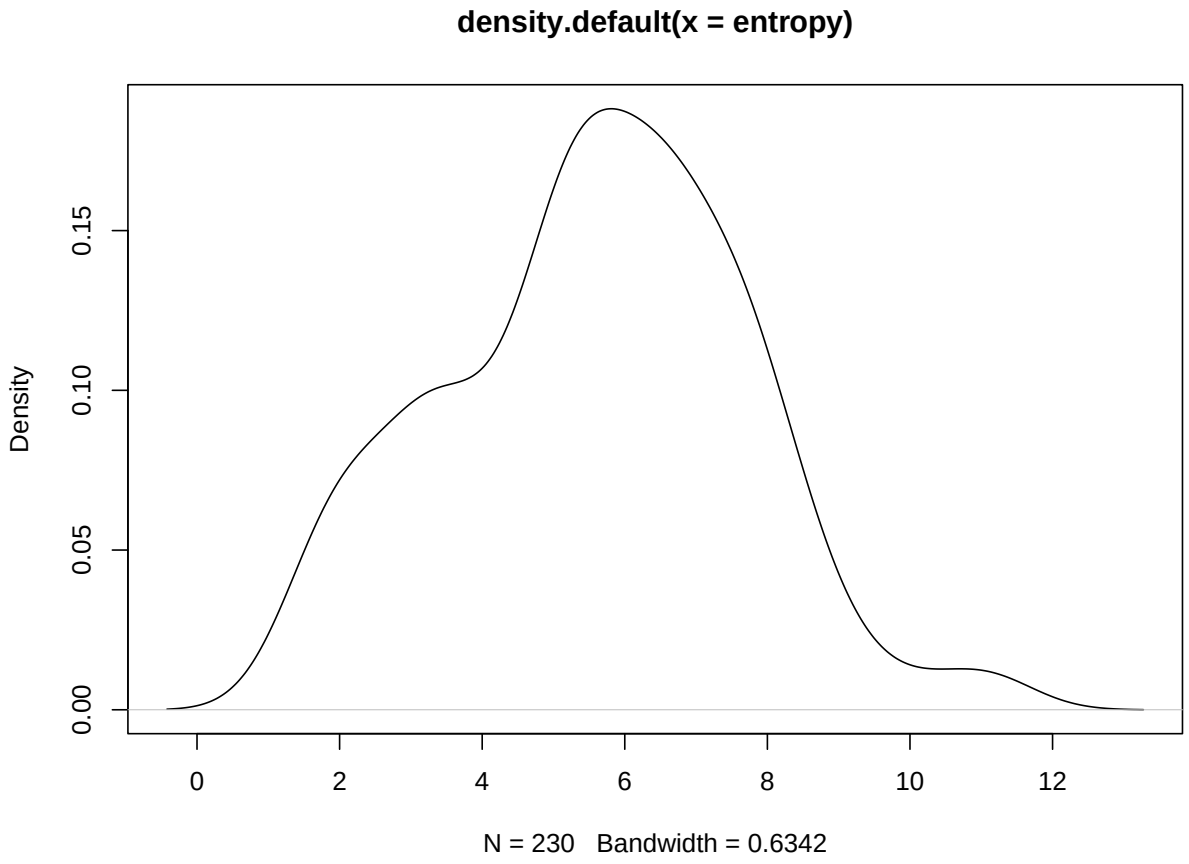
Then they created 7 seven files to analyse and to find binding sites enriched in the SARS-CoV-2 genome: - filter_ATtRACT - find_binding_sites - annotate_sites - simulate_genomes - find_sim_binding_sites - enrichment_analysis - filter_hits

1) RNA-binding proteins (filter_ATtRACT)

We start by filtering the RNA-binding proteins (RBP) coming from the daTabase of RNA binding proteins and AssoCiated moTifs (ATtRACT). Before filtration, there were 327 RBPs. We keep only human RBP (159 RBP) and we remove the ones consisting of only a single motif (106 RBP).

Then we filter the position weight matrices (PWM) which are used to represente motifs (patterns) in biological sequences by matching then with RBPs and by removing high entropy PWM. We reduced the number of PWMs from 1583 to 230 and the number of RBPs in the search from 106 to 102.

We can visualize the plot of the density vs the entropy.



2) Genome

We upload the SARS2 genome and we scan the genome with the PWMs. After the selection of unique RBP-PWM mappings and finding duplicate binding sites for the same protein and choose the one with the highest score, we reduced number of sites from 30597 to 29290.

3) Annotated sites

We load the gff file. We can visualize the head of the file.

	seqnames	start	end	width	strand	source	type	score	phase	ID	Dbxref	collection.date
1	NC_045512.2	1	265	265	+	RefSeq	five_prime_UTR	NA	NA	id-NC_045512.2:1..265		NA
2	NC_045512.2	266	21555	21290	+	RefSeq	gene	NA	NA	gene-GU280_gp01	GeneID:4....	NA
3	NC_045512.2	266	13468	13203	+	RefSeq	CDS	NA	0	cds-YP_009724389.1	Genbank:....	NA
4	NC_045512.2	13468	21555	8088	+	RefSeq	CDS	NA	0	cds-YP_009724389.1	Genbank:....	NA
5	NC_045512.2	266	805	540	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:1..180		NA
6	NC_045512.2	806	2719	1914	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:181..818		NA
7	NC_045512.2	2720	8554	5835	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:819..2763		NA
8	NC_045512.2	8555	10054	1500	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:2764..3263		NA
9	NC_045512.2	10055	10972	918	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:3264..3569		NA
10	NC_045512.2	10973	11842	870	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:3570..3859		NA
11	NC_045512.2	11843	12091	249	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:3860..3942		NA
12	NC_045512.2	12092	12685	594	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:3943..4140		NA
13	NC_045512.2	12686	13024	339	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:4141..4253		NA
14	NC_045512.2	13025	13441	417	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:4254..4392		NA
15	NC_045512.2	13442	13468	27	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:4393..5324		NA
16	NC_045512.2	13468	16236	2769	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:4393..5324		NA
17	NC_045512.2	16237	18039	1803	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:5325..5925		NA
18	NC_045512.2	18040	19620	1581	+	RefSeq	mature_protein_region_of_CDS	NA	NA	id-YP_009724389.1:5926..6452		NA

In bioinformatics, the general feature format (gene-finding format, generic feature format, GFF) is a file format used for describing genes and other features of DNA, RNA and protein sequences.

We can also visualize the head of the sites file.

Matrix_id	seqname	source	feature	start	end	score	strand	frame	seq	Gene_name	len
75	NC_045512.2	TFBS	TFBS	1	6	2.999063	+	.	ATTAAA	PPIE	6
75	NC_045512.2	TFBS	TFBS	1	6	2.999063	-	.	TTTAAT	PPIE	6
s63	NC_045512.2	TFBS	TFBS	1	4	3.450980	+	.	ATTA	TIAL1	4
s1	NC_045512.2	TFBS	TFBS	3	7	3.700284	-	.	CTTTA	ELAVL2	5
s6	NC_045512.2	TFBS	TFBS	3	7	3.700284	-	.	CTTTA	TIAL1	5
s9	NC_045512.2	TFBS	TFBS	3	7	3.737562	-	.	CTTTA	TIA1	5
143	NC_045512.2	TFBS	TFBS	4	8	2.997815	-	.	CCTTT	PTBP1	5
155	NC_045512.2	TFBS	TFBS	4	9	3.664262	-	.	ACCTTT	ZFP36	6
165	NC_045512.2	TFBS	TFBS	4	7	2.496259	-	.	CTTT	PTBP1	4
s21	NC_045512.2	TFBS	TFBS	6	11	4.189243	+	.	AGGTTT	ZRANB2	6
M033_0.6	NC_045512.2	TFBS	TFBS	7	14	5.822878	-	.	TATAAACC	KHDRBS3	8
s38	NC_045512.2	TFBS	TFBS	7	10	1.884956	-	.	AACC	HNRNPL	4
s1	NC_045512.2	TFBS	TFBS	8	12	3.700284	+	.	GTTTA	ELAVL2	5
s6	NC_045512.2	TFBS	TFBS	8	12	3.558239	+	.	GTTTA	TIAL1	5
s9	NC_045512.2	TFBS	TFBS	8	12	3.737562	+	.	GTTTA	TIA1	5
75	NC_045512.2	TFBS	TFBS	9	14	2.999063	+	.	TTTATA	PPIE	6
75	NC_045512.2	TFBS	TFBS	9	14	2.999063	-	.	TATAAA	PPIE	6
s21	NC_045512.2	TFBS	TFBS	12	17	4.488048	-	.	AGGTAT	ZRANB2	6

The new file (*annoted_sites*) is more complete.

seqnames	start	end	width	strand	Matrix_id	source	feature	score	frame	seq	Gene_name	len	type	gene
NC_045512.2	1259	1265	7	+	s50	TFBS	TFBS	4.073349	.	CAGACGG	SRSF1	7	gene	ORF1ab
NC_045512.2	1259	1265	7	+	s50	TFBS	TFBS	4.073349	.	CAGACGG	SRSF1	7	CDS	ORF1ab
NC_045512.2	1259	1265	7	+	s50	TFBS	TFBS	4.073349	.	CAGACGG	SRSF1	7	CDS	ORF1ab
NC_045512.2	1259	1265	7	+	s50	TFBS	TFBS	4.073349	.	CAGACGG	SRSF1	7	mature_protein_region_of_CDS	NA
NC_045512.2	1259	1265	7	+	s50	TFBS	TFBS	4.073349	.	CAGACGG	SRSF1	7	mature_protein_region_of_CDS	NA
NC_045512.2	1260	1266	7	+	M020_0.6	TFBS	TFBS	4.757508	.	AGACGGG	FXR2	7	gene	ORF1ab
NC_045512.2	1260	1266	7	+	M020_0.6	TFBS	TFBS	4.757508	.	AGACGGG	FXR2	7	CDS	ORF1ab
NC_045512.2	1260	1266	7	+	M020_0.6	TFBS	TFBS	4.757508	.	AGACGGG	FXR2	7	CDS	ORF1ab
NC_045512.2	1260	1266	7	+	M020_0.6	TFBS	TFBS	4.757508	.	AGACGGG	FXR2	7	mature_protein_region_of_CDS	NA
NC_045512.2	1260	1266	7	+	M020_0.6	TFBS	TFBS	4.757508	.	AGACGGG	FXR2	7	mature_protein_region_of_CDS	NA
NC_045512.2	1262	1266	5	+	s19	TFBS	TFBS	3.113263	.	ACGGG	SRSF5	5	gene	ORF1ab
NC_045512.2	1262	1266	5	+	s19	TFBS	TFBS	3.113263	.	ACGGG	SRSF5	5	CDS	ORF1ab
NC_045512.2	1262	1266	5	+	s19	TFBS	TFBS	3.113263	.	ACGGG	SRSF5	5	CDS	ORF1ab
NC_045512.2	1262	1266	5	+	s19	TFBS	TFBS	3.113263	.	ACGGG	SRSF5	5	mature_protein_region_of_CDS	NA
NC_045512.2	1262	1266	5	+	s19	TFBS	TFBS	3.113263	.	ACGGG	SRSF5	5	mature_protein_region_of_CDS	NA
NC_045512.2	1263	1269	7	+	167	TFBS	TFBS	3.749375	.	CGGGCGA	SRSF1	7	gene	ORF1ab
NC_045512.2	1263	1269	7	+	167	TFBS	TFBS	3.749375	.	CGGGCGA	SRSF1	7	CDS	ORF1ab
NC_045512.2	1263	1269	7	+	167	TFBS	TFBS	3.749375	.	CGGGCGA	SRSF1	7	CDS	ORF1ab
NC_045512.2	1263	1269	7	+	167	TFBS	TFBS	3.749375	.	CGGGCGA	SRSF1	7	mature_protein_region_of_CDS	NA

4) Simulated genomes

To simulate the genome, we used the reference genome with the number of occurrence of each nucleotide. They did 5000 simulations which made our computers crash. So we decided to compare the results with 50, 500 and 5000 simulations to see if it was necessary to have this number of simulation.

We also simulated untranslated regions (or UTR): 5' side (sim_f_utrs) and 3' side (sim_t_utrs).

We have 3 outputs for this file: sim_t_utrs.RData, sim_f_utrs.RData and sim_genomes.RData.

```
##
## sim1 CTAGAGCTGATACGTGTTCCAGACGCGTAGTAAACCGGTAGCTGCTGTATACGGATTGGGGAAGATATTACCGGCAGCATTTACCTTAGAACACCC
## sim2 TAGTGACTTTACGTTGTTCTATGTTCATCATGCGTTTTATTATGAAAACTGATGCTCAAGCTCATAAGGAATCCACTTCGTTAGCGGTAACCCTAT
## sim3 AAAACATGATCGTCAAAATGATGGACTTTTGAGATCGTTACATTTAACTTTTTTCGACACTTTTTTGTCCATTTATATGGCTAGTTTTTAAGAGCA
## sim4 TTTGTAAATGCACTACCACGAAGTCGTTTTTGGTGACCCTATTTAACTGATGGTAGAAGAGGTTTTCCTTCGAAACAGTGAAAGACCCACACATTT
## sim5 TCAGCATATTTTGTTTTTAGTGTGGTGAAAGGATAAGTTATCCTGATGCAAGAAGATTTTTGAACAAGTCGATACTTATGACTCAGTGGTTGCTCT
## sim6 CTCTACTCCCTTGGGGCAACCAACTATCAGCGAAGGATAGGAAGGAATGGGCTGAGAATATTCAATTACTACCGGGATTTTACGAACCCGGTTGAA
```

5) Find the binding sites of the simulated genomes

We load all the data we have: the genomes, the binding sites, the matrices and the proteins. We start by listing all PWMs for RBPs that were found in the reference genome and then we look at the simulated genomes with these PWMs. For 50 simulations, we obtain 1520059 binding sites from simulated genomes.

6) Enrichment analysis

We match binding sites with protein name Find duplicate binding sites for the same protein and choose the one with the highest score *Enrichment*: is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted groups of genes.

For 50 simulations, we have 195 rows in the *site_count* file but each gene is represented two times (*why?*) so we have 98 genes.

7) Filter hits

Number	Name	Num_sites	P_adj
1	HNRNPL	632	2.7e-32
2	FUS	140	5.4e-24
3	MBNL1	682	2.0e-19
4	SRSF1	335	1.5e-13
5	RBMY1A1	107	1.3e-11
6	ZFP36	609	8.3e-11
7	SRSF10	88	8.7e-10
8	PTBP1	3151	1.5e-08
9	SRSF3	74	3.1e-07
10	YBX2	51	3.5e-07
11	PABPC1	118	1.9e-06
12	PABPN1	50	1.3e-05
13	SART3	49	3.6e-05
14	PABPC4	28	3.1e-04
15	ZNF638	27	3.1e-04
16	PABPC5	33	7.4e-04
17	CELF2	698	1.6e-03
18	PABPC3	34	1.8e-03
19	YBX1	204	3.3e-03

In the paper, they found 19 proteins whose binding sites are enriched in the SARS-CoV-2 genome.

Filter significant hits by keeping genes with qval<0.01, strand==“+” and N>2.

a) 50 simulations

```
## Warning: le package 'kableExtra' a été compilé avec la version R 4.1.2
```

\begin{table}[!h]

\caption{genome_hits}

Gene_name	strand	N	mean_count	sd_count	z	pval	qval
HNRNPL	+	632	389.66	19.658939	12.327217	0.0000000	0.0000000
MBNL1	+	682	402.76	28.330815	9.856405	0.0000000	0.0000000
FUS	+	140	61.68	8.049439	9.729871	0.0000000	0.0000000
ZFP36	+	609	470.62	16.919847	8.178561	0.0000000	0.0000000
SRSF1	+	335	223.16	14.060162	7.954389	0.0000000	0.0000000
RBMY1A1	+	107	56.14	7.461493	6.816330	0.0000000	0.0000000
PTBP1	+	3151	2834.22	48.527413	6.527857	0.0000000	0.0000000
SRSF10	+	88	43.34	7.029413	6.353304	0.0000000	0.0000000
YBX2	+	51	22.72	4.611609	6.132350	0.0000000	0.0000000
SRSF3	+	74	37.82	6.548687	5.524771	0.0000000	0.0000002
PABPN1	+	50	25.70	4.752014	5.113622	0.0000002	0.0000014
PABPC1	+	118	73.04	9.682638	4.643363	0.0000017	0.0000141
SART3	+	49	23.98	5.615176	4.455782	0.0000042	0.0000318
PABPC4	+	28	10.62	4.471634	3.886722	0.0000508	0.0003593
ZNF638	+	27	13.32	3.716702	3.680682	0.0001163	0.0007676
PABPC5	+	33	18.56	3.985484	3.623149	0.0001455	0.0009004
CELF2	+	698	586.92	36.765023	3.021350	0.0012583	0.0073275
NOVA2	+	35	23.22	3.950149	2.982166	0.0014311	0.0078710
ENOX1	+	16	8.94	2.393955	2.949095	0.0015935	0.0083031
YBX1	+	204	165.28	13.295757	2.912207	0.0017944	0.0088824

\end{table}

Table 1: 5' UTR sequences

Gene_name	strand	N	mean_count	sd_count	z	pval	qval
ZRANB2	+	4	0.76	0.8703741	3.722537	9.86e-05	0.0009368

Table 2: 3' UTR sequences

Gene_name	strand	N	mean_count	sd_count	z	pval	qval
PABPC4	+	27	0.26	0.8283251	32.282012	0.00e+00	0.0000000
SART3	+	28	0.88	1.1542291	23.496201	0.00e+00	0.0000000
SRSF10	+	27	1.34	1.1885508	21.589317	0.00e+00	0.0000000
PABPC1	+	28	1.44	1.2803061	20.745039	0.00e+00	0.0000000
KHDRBS3	+	28	1.98	1.8679903	13.929408	0.00e+00	0.0000000
PPIE	+	42	13.94	4.4511406	6.304002	0.00e+00	0.0000000
HNRNPA1	+	4	0.60	0.8081220	4.207285	1.29e-05	0.0000531
LIN28A	+	3	0.32	0.6833292	3.921975	4.39e-05	0.0001625

Genome hits

For 50 simulations, we find 20 genes. With only 50 simulations, we can observe some differences between the paper and our results. PABPC3 gene appears in the paper but not in our results and NOVA2 and ENOX1 genes appear in our results but not in the paper. However, the genes with very small p-value are in both tables.

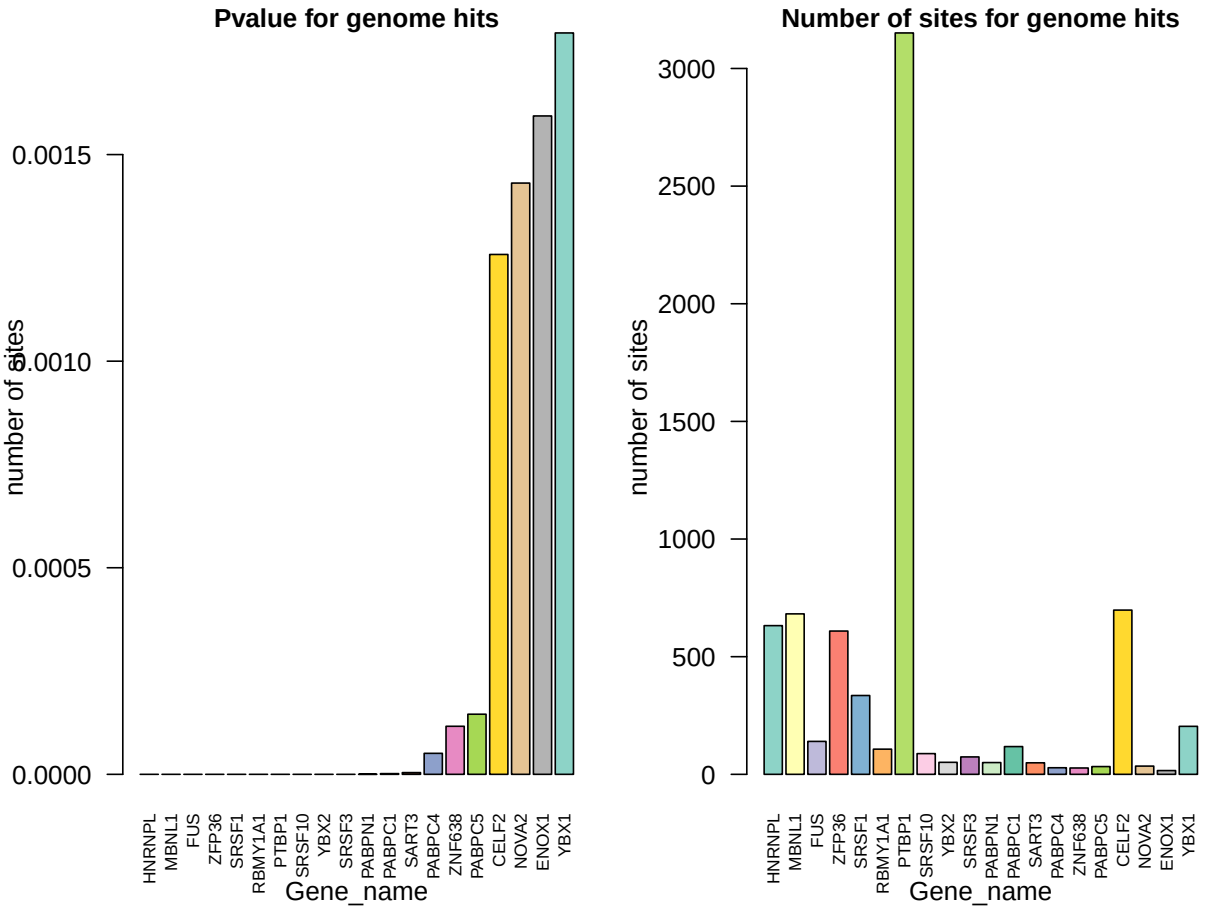
Negative genome hits

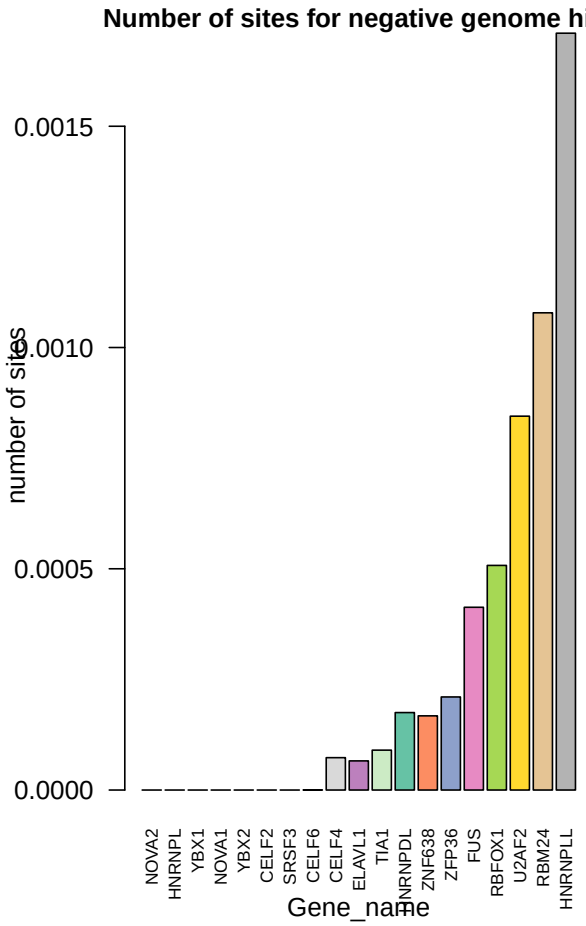
5' and 3' UTR sequences

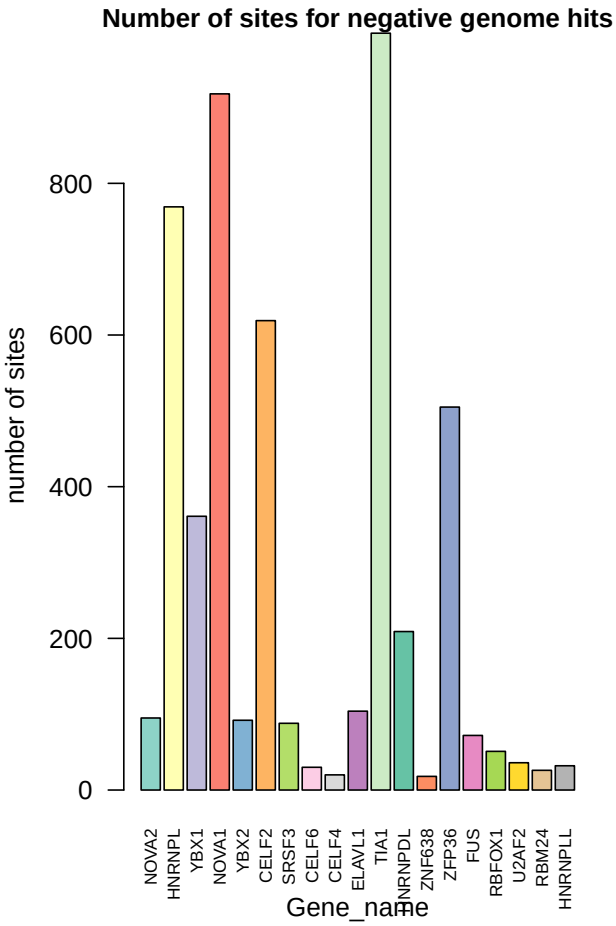
```
## Warning in brewer.pal(n = 9, name = "Set2"): n too large, allowed maximum for palette Set2 is 8
## Returning the palette you asked for with that many colors
```

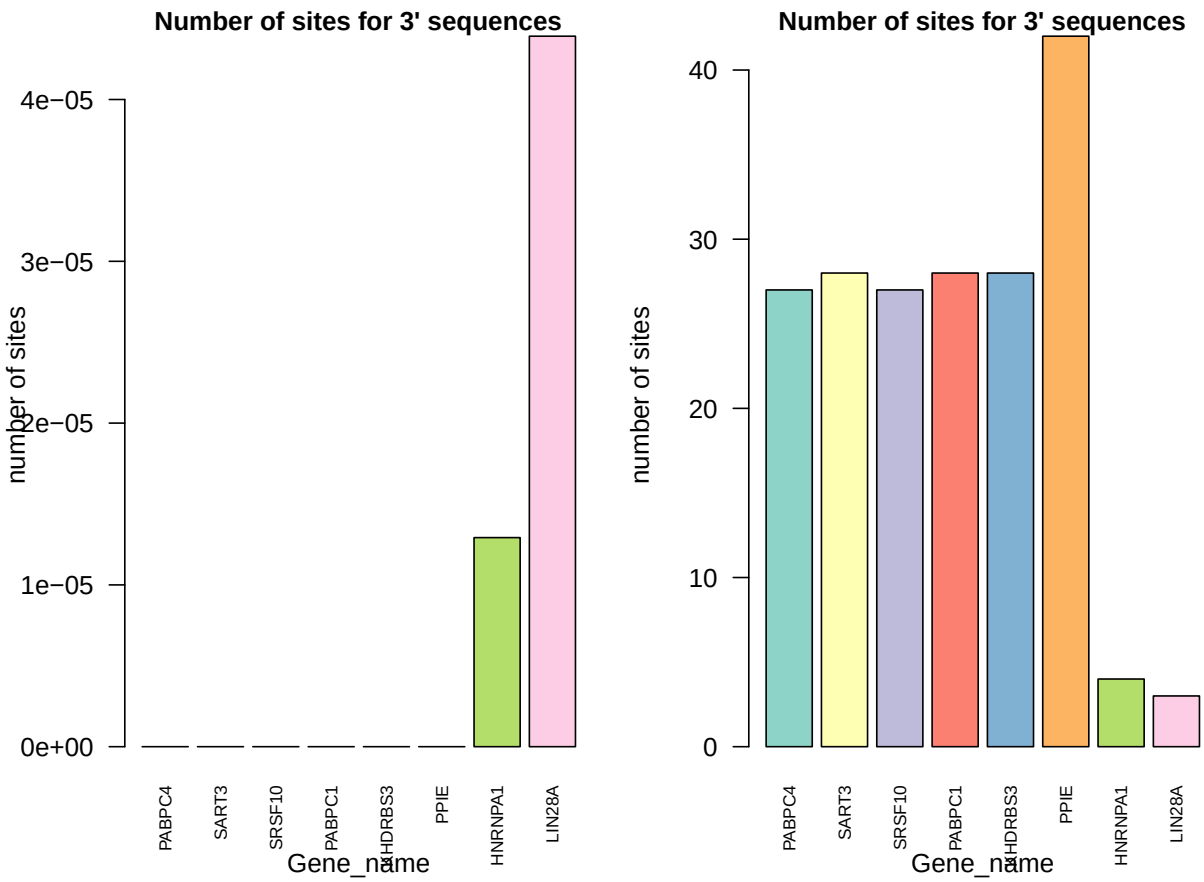
Table 3: Genome negative hits

Gene_name	strand	N	mean_count	sd_count	z	pval	qval
NOVA2	-	95	30.24	5.109255	12.675037	0.0000000	0.0000000
HNRNPL	-	769	510.94	21.920459	11.772564	0.0000000	0.0000000
YBX1	-	361	198.74	13.863489	11.704124	0.0000000	0.0000000
NOVA1	-	918	659.04	22.676033	11.419987	0.0000000	0.0000000
YBX2	-	92	32.68	5.242838	11.314483	0.0000000	0.0000000
CELF2	-	619	412.56	29.671783	6.957452	0.0000000	0.0000000
SRSF3	-	88	46.20	6.770283	6.174040	0.0000000	0.0000000
CELF6	-	30	12.44	3.459149	5.076392	0.0000002	0.0000023
CELF4	-	20	7.76	3.223384	3.797252	0.0000732	0.0007023
ELAVL1	-	104	65.54	10.059436	3.823276	0.0000658	0.0007023
TIA1	-	998	893.78	27.827669	3.745193	0.0000901	0.0007866
HNRNPDL	-	209	171.42	10.511587	3.575102	0.0001750	0.0012926
ZNF638	-	18	8.72	2.587568	3.586379	0.0001677	0.0012926
ZFP36	-	505	438.98	18.719727	3.526761	0.0002103	0.0014423
FUS	-	72	47.78	7.242984	3.343926	0.0004130	0.0026433
RBFOX1	-	51	29.30	6.603184	3.286293	0.0005076	0.0030455
U2AF2	-	36	20.12	5.057385	3.139963	0.0008448	0.0047709
RBM24	-	26	14.26	3.826919	3.067742	0.0010784	0.0057515
HNRNPLL	-	32	18.46	4.625616	2.927177	0.0017103	0.0086414




Number of sites for negative genome hits
number of sites
0.0015
0.0010
0.0005
0.0000
NOVA2
HNRNPL
YBX1
NOVA1
YBX2
CELF2
SRSF3
CELF6
CELF4
ELAVL1
TIA1
HNRNPDL
ZNF638
ZFP36
FUS
RBFOX1
U2AF2
RBM24
HNRNPLL
Gene_name


Number of sites for negative genome hits
number of sites
800
600
400
200
0
NOVA2
HNRNPL
YBX1
NOVA1
YBX2
CELF2
SRSF3
CELF6
CELF4
ELAVL1
TIA1
HNRNPDL
ZNF638
ZFP36
FUS
RBFOX1
U2AF2
RBM24
HNRNPLL
Gene_name



b) 500 simulations

Results

Conclusions

References