Collaborations

Analysis of various coming-into the core data (single Cell and Bulk RNASeq, ATAC, miRNA, etc)

Merkle Lab

Pilot experiment for measuring selective pressures experienced during human pluripotent stem cell (hPSC) culture
Variations in human cells using exome and whole genome data.

Julian Griffin Lab

Finding variation causing **Creatine deficiency syndrome**

Main Projects

Max Planck Institute - Germany

Ribosomal Profiling Drop-off rate

Structure variations in **Burkitt lymphoma cancer**

Ribofilio: A tool to estimate the drop-off rate of ribosomes from Ribo-seq data

Are there any factors that affect the drop-off rate?

Sherine Awad
Davide Chiarugi
Angel Valleriani

Ribosomal profiling

- ➔ This technique begins with drug-mediated interruption of the cellular translation process followed by hydrolysis of the mRNA regions that are not covered (protected) by the ribosomes.
- ➔ The residual mRNA oligomers known as ribosome protected fragments (RPF) that were protected by the ribosome are deep sequenced.
- ➔ The positions of the ribosomes are determined by mapping the sequence to the reference genome

Sin, Celine, Davide Chiarugi, and Angelo Valleriani. "Quantitative assessment of ribosome drop-off in E. coli." *Nucleic acids research* 44.6 (2016): 2528-2537.

- ➔ The abundance of the RPFs that map to different parts of the single genes usually evaluated in terms of **Ribosome Density (RD)** and measured in number of RPF per codon is typically used for protein synthesis rate for each gene.
- ➔ The distribution of RPFs along the genes also provide information about the possible presence of **ribosome drop-off**; an average decrease of the RD from the 5’ end to the 3’ end of each open reading frame (ORF) reflects that a significant number of ribosomes fail to reach the 3’ end.

Sin, Celine, Davide Chiarugi, and Angelo Valleriani. "Quantitative assessment of ribosome drop-off in E. coli." *Nucleic acids research* 44.6 (2016): 2528-2537.

- The goal is to have a quantitative estimate for ribosome drop -off
- Are there any factors that affect the ribosome drop-off

Sin, Celine, Davide Chiarugi, and Angelo Valleriani. "Quantitative assessment of ribosome drop-off in E. coli." *Nucleic acids research* 44.6 (2016): 2528-2537.

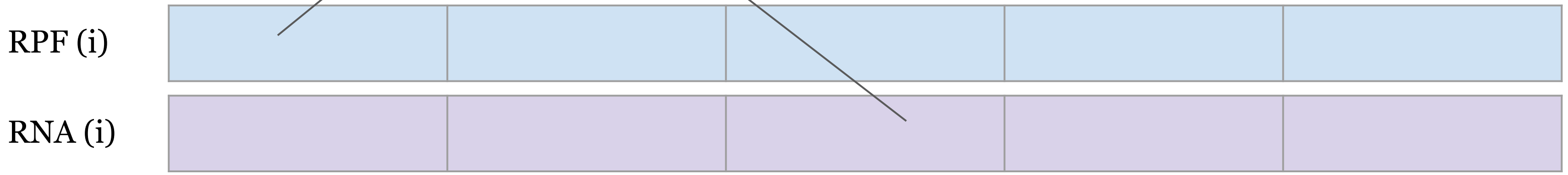
Binning Strategy: Computing the average number of RPFs per ORF

We normalize the amount of RPFs with the abundance of the corresponding RNA-seq reads

Average number of RPF

$$\mathrm{NRPF}\ (i) = \frac{\mathrm{RPF}\ (i)}{\mathrm{RNA}\ (i)}$$

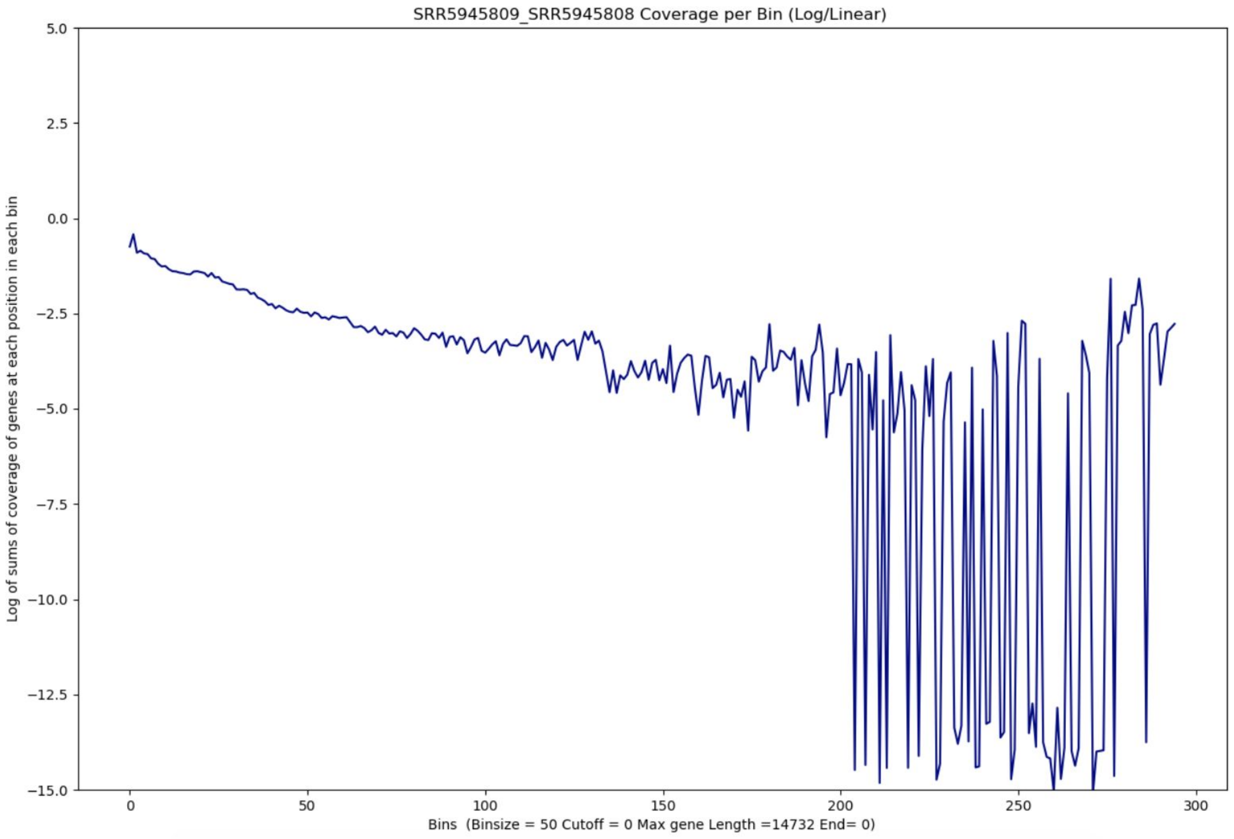
$$Y = Ae^{-RX}$$

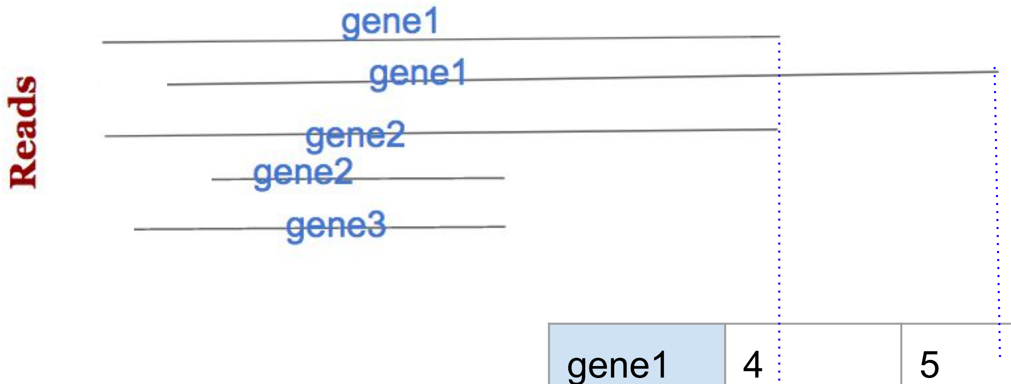


The relationship between the average number of RPFs per bin Y and the bin number X. This study the dependence of Y on X based on exponential decay:

$$Y = A\mathrm{e}^{-RX}$$

X is the bin number and A is the intercept (no interest) and the value R is the drop-off rate per bin





coverage

gene1	4	5
gene2	4	2
gene3	2	….

```
coverage[str(gname)].append(int(x[2]))
pos[int(x[2])] =0
```

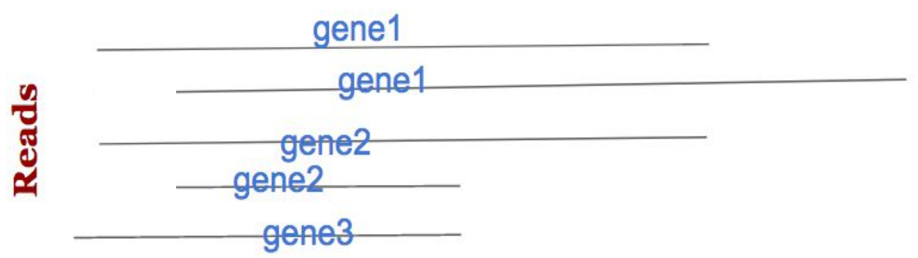
(i) is read end from coverage

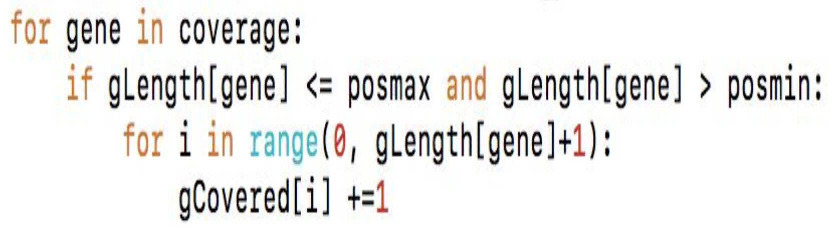
posmax is the longest gene in whole set or longest gene in subset if **--subset** is chosen

pos

1	2	3	4	5
0	2	0	2	1

```
for gene in coverage:
    for i in coverage[gene]:
        if gLength[gene] <= posmax and gLength[gene] > posmin:
            pos[int(i)] +=1
print("Filling pos is done")
```





```
for gene in coverage:
    if gLength[gene] <= posmax and gLength[gene] > posmin:
        for i in range(0, gLength[gene]+1):
            gCovered[i] +=1
```

gCovered

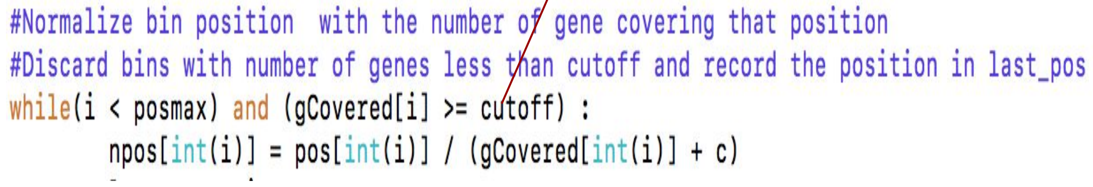
How many genes cover each position based on length

1	2	3	4	5
3	3	2	2	1

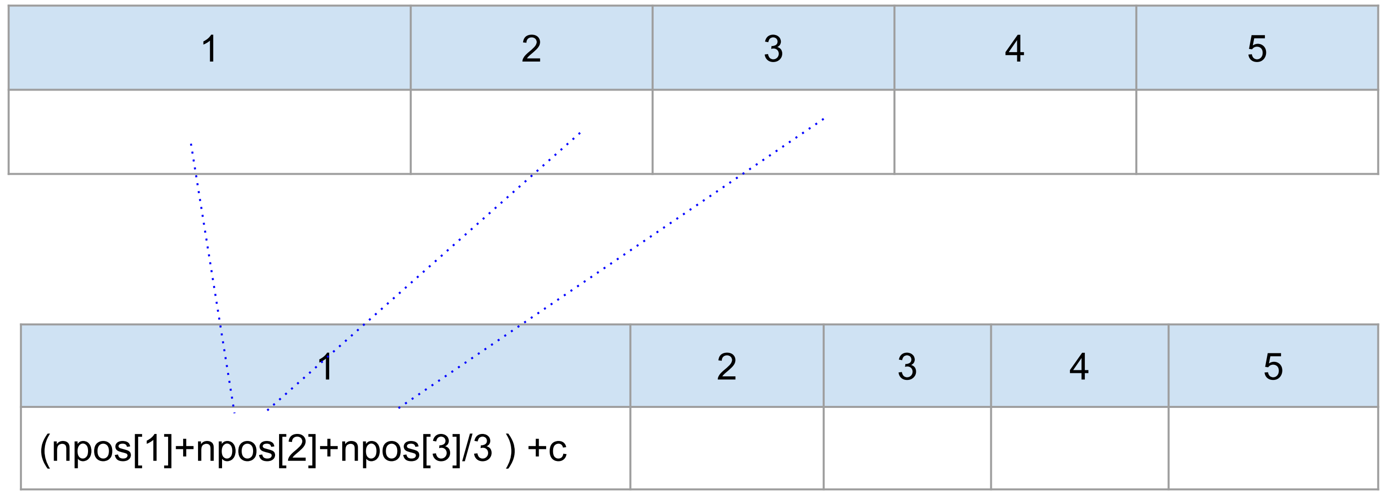
Cutoff is usually 0 unless set otherwise for cutoff experiment

npos

1	2
pos[1]/gcovered[1] +c	..



```
#Normalize bin position  with the number of gene covering that position
#Discard bins with number of genes less than cutoff and record the position in last_pos
while(i < posmax) and (gCovered[i] >= cutoff) :
        npos[int(i)] = pos[int(i)] / (gCovered[int(i)] + c)
```



npos

1	2	3	4	5

BINSIZE =3

Bins

RPF (i)

1	2	3	4	5
(npos[1]+npos[2]+npos[3]/3) +c				

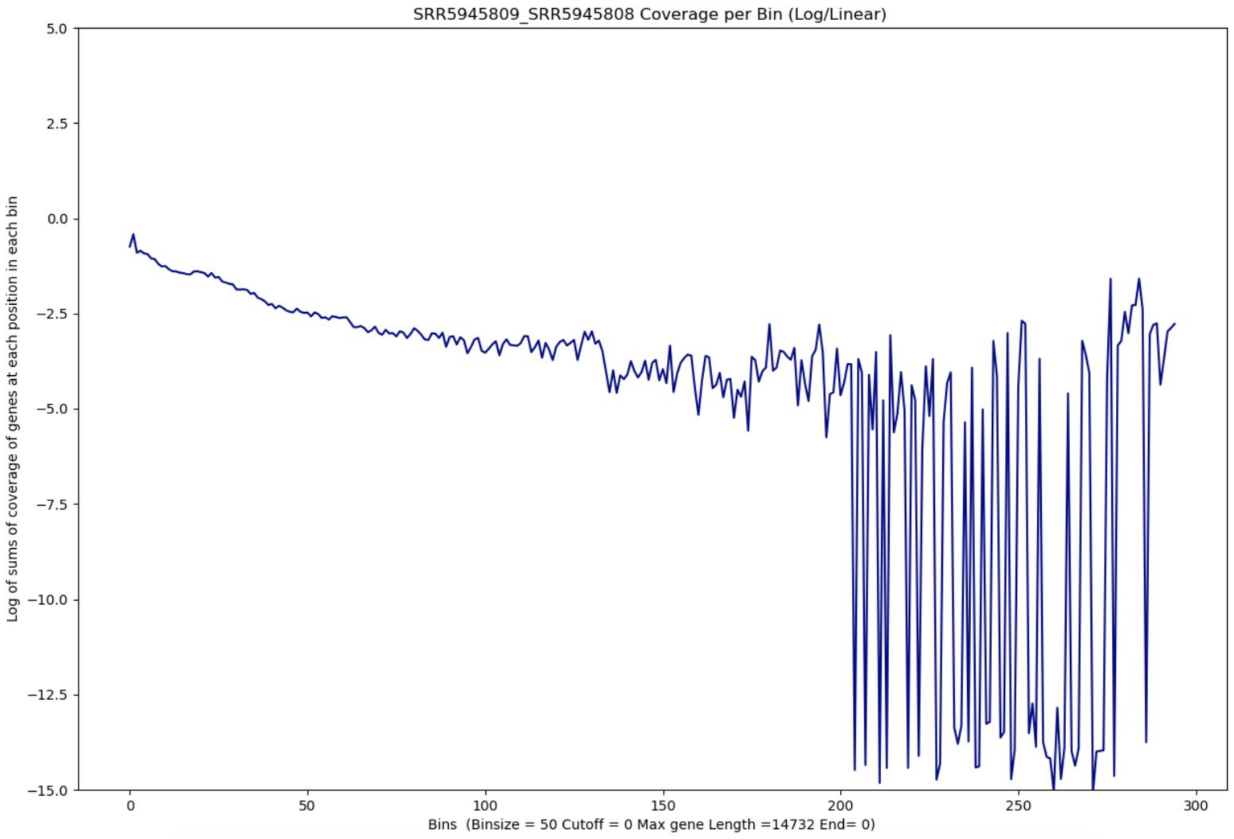
```
while a< posmax:
    for i in range (a,b+1):
        if i >  (len(npos) - 1):
            break
        if i > last_pos:
            break
        posum = float(npos[i])
        gbins[index] +=posum
    gbins[index] = (float(c +  (gbins[index]/BINSIZE) ) )
    index +=1
    a = b+1
    b = b +BINSIZE
```

$$nRPF(i) = RPF\,(i)\;/RNA\,(i)$$

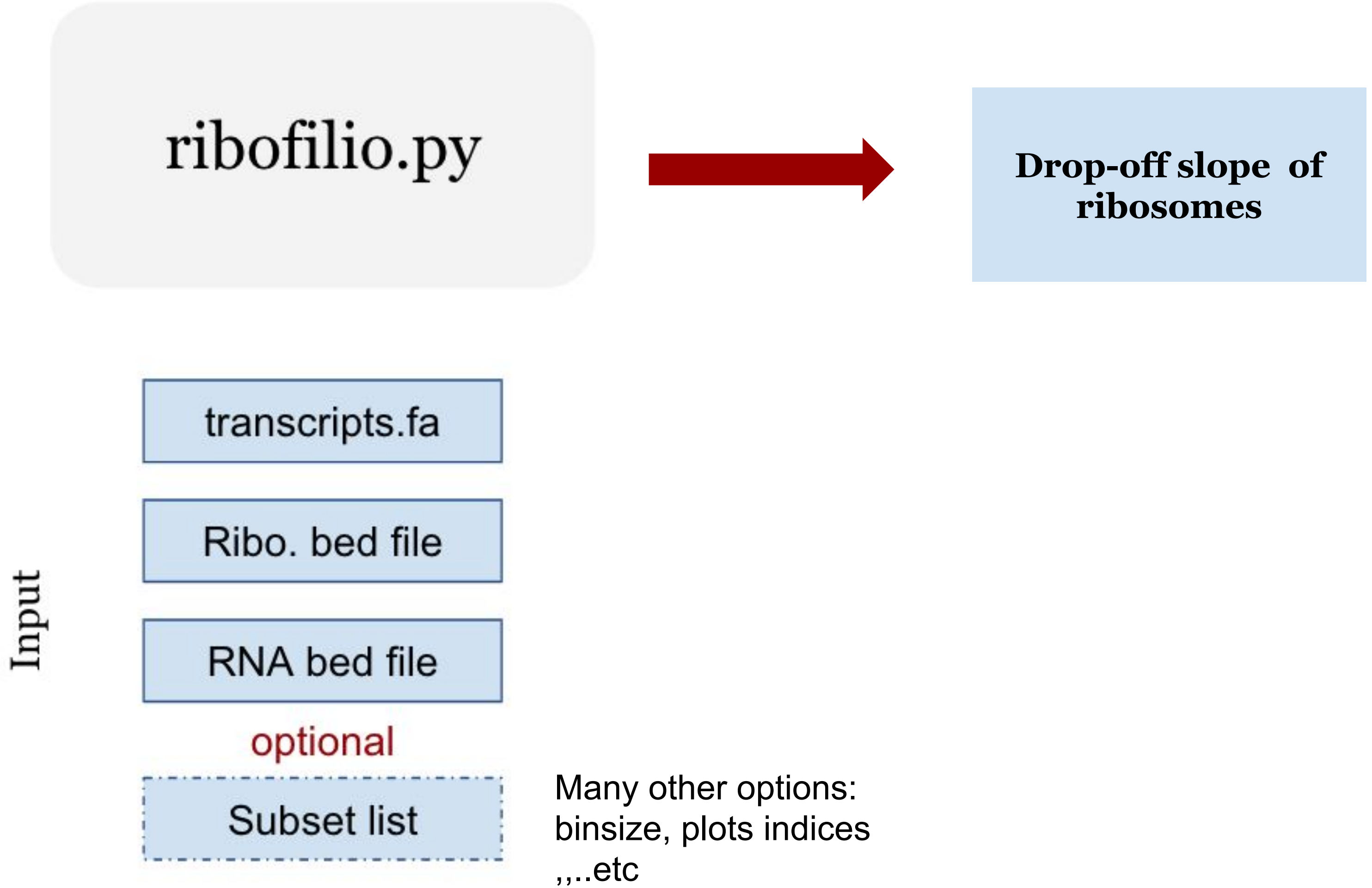
The relationship between the average number of RPFs per bin Y and the bin number X. This study the dependence of Y on X based on exponential decay:

$$Y = A\mathrm{e}^{-RX}$$

X is the bin number and A is the intercept (no interest) and the value R is the drop-off rate per bin



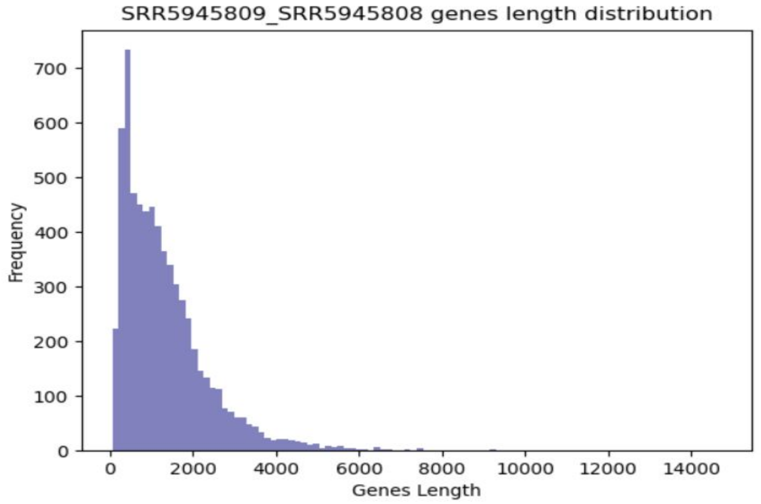
Conda
Docker
..etc



Many other options:
binsize, plots indices
,,..etc

Results on Yeast SRR5945809/SRR5945808

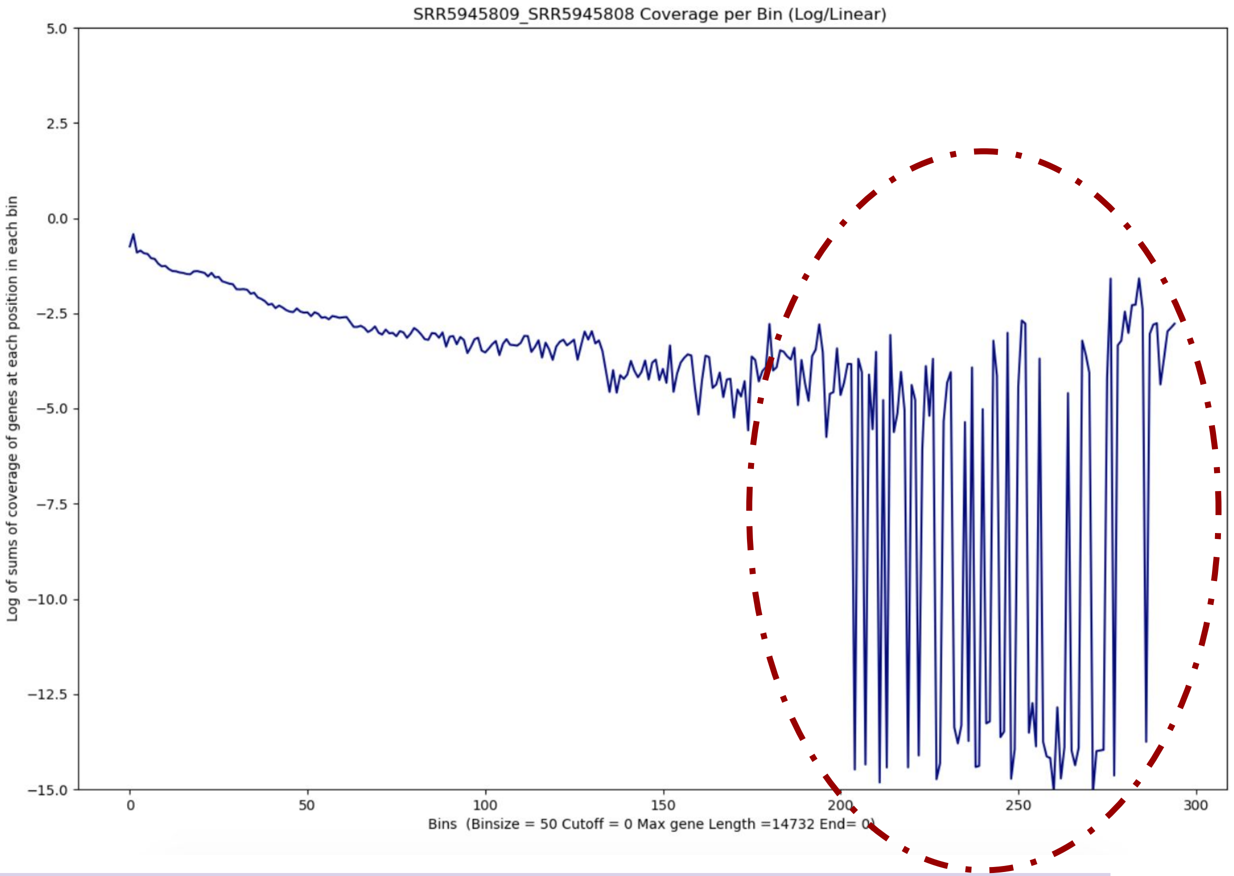
mRNA/FP	Unique Alignment
SRR5945808 (mRNA)	26.83%
SRR5945809 (FP)	28.51%



The relationship between the average number of RPFs per bin Y and the bin number X. This study the dependence of Y on X based on exponential decay:

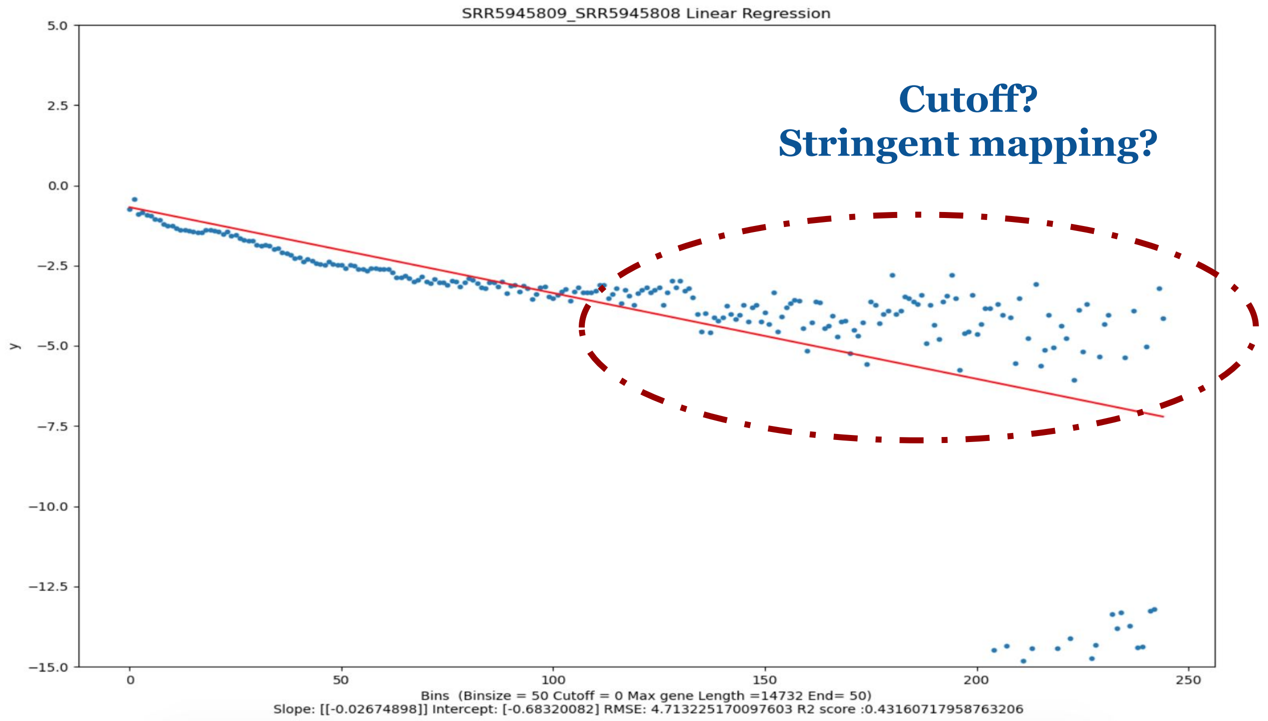
$$Y = Ae^{-RX}$$

X is the bin number and A is the intercept (no interest) and the value R is the drop-off rate per bin



```
python ribofilio.py -t yeast.fa -f SRR5945809.bed -r SRR5945808.bed -b 50
```

Linear Regression



Weighted -Linear Regression

- ➔ Better RMSE and R2
- ➔ Better fit and handle this noise?

?

Factors that could affect the drop-off rate

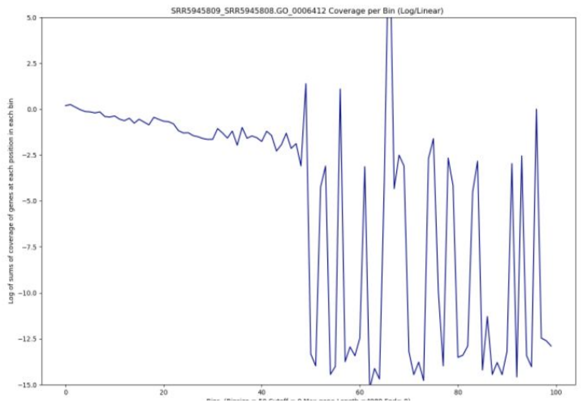
Ribofilio parameter

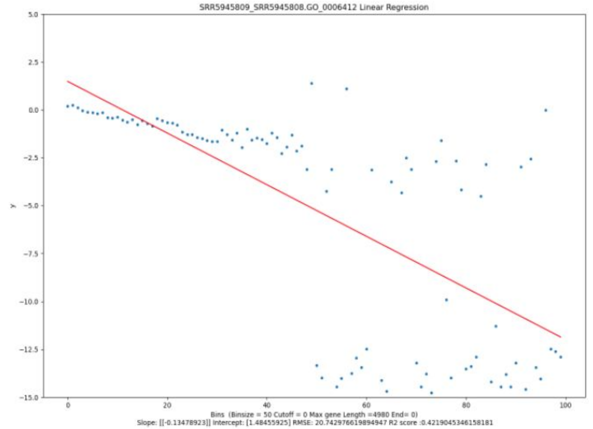
```
--subset  or -s
```

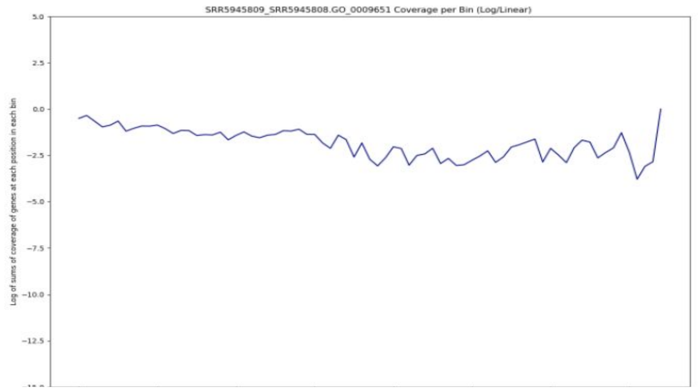
Factor
Gene Ontology (GO) of the gene
Gene Length
TPM (Over or Under expressed transcripts)
SNPs
Distance (By chromosome for example)
Orthogonality (Orthologous)

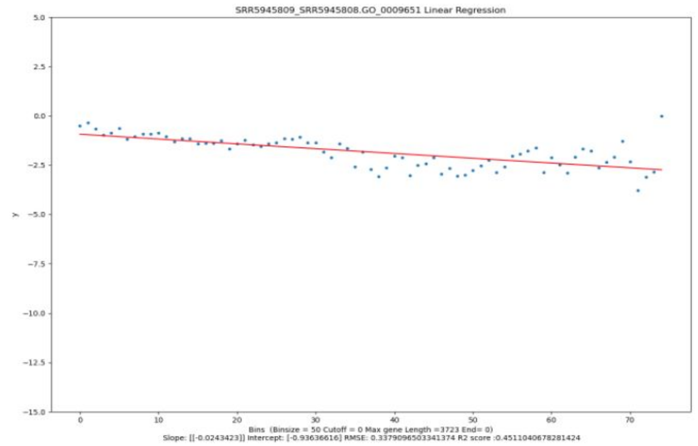
GO Translation

GO Response to Salt Stress



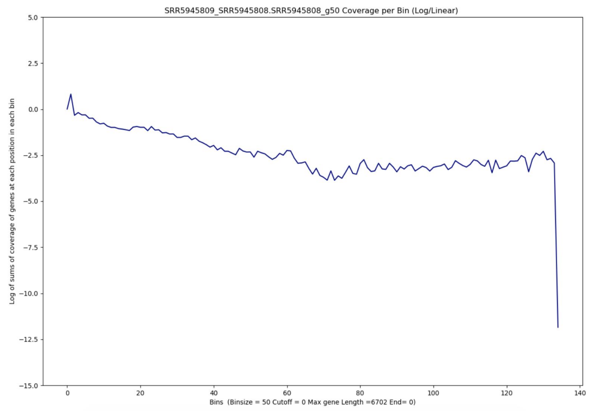


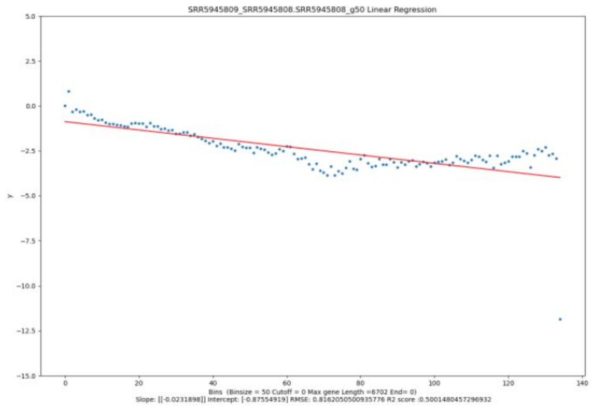




```
python ribofilio.py -t yeast.fa -f SRR5945809.bed -r SRR5945808.bed -b 50 -s  GO0006412.txt
```

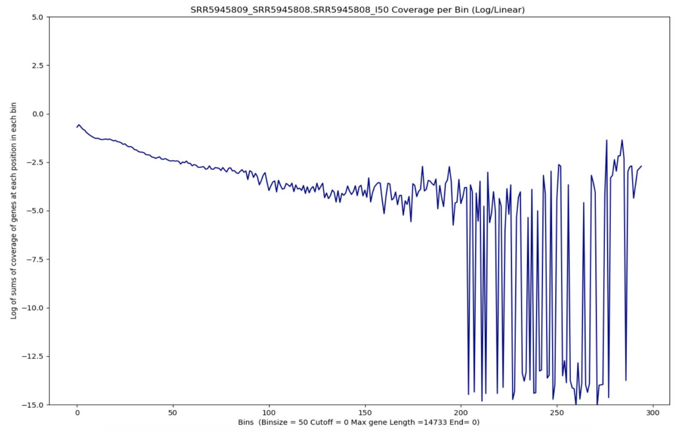
TPM >=50

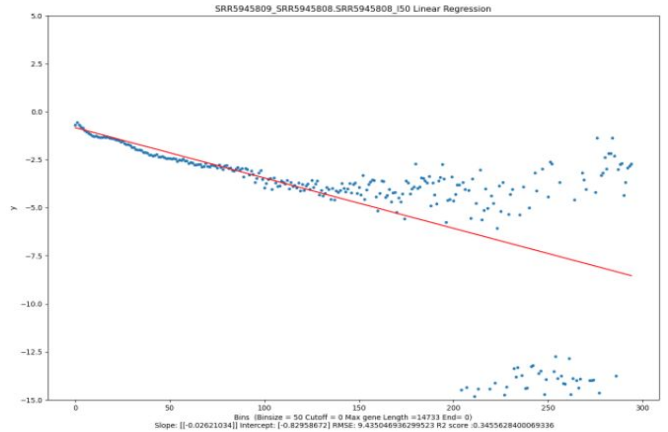




```
Slope: [[-0.0231898]] Intercept: [-0.87554919] RMSE: 0.8162050500935776 R2 score :0.5001480457296932
```

TPM < 50

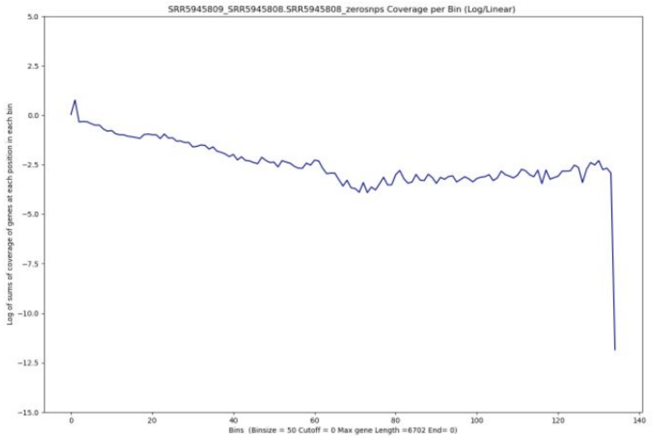


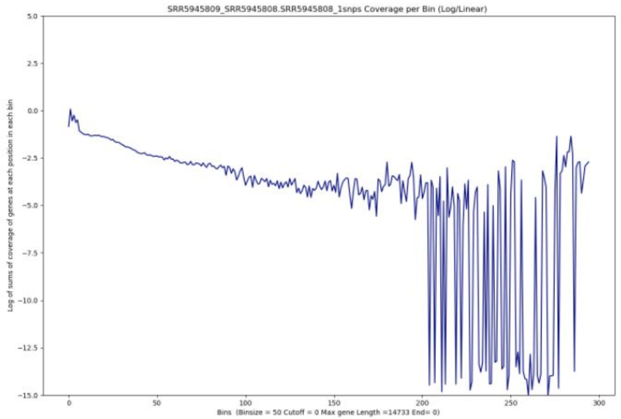


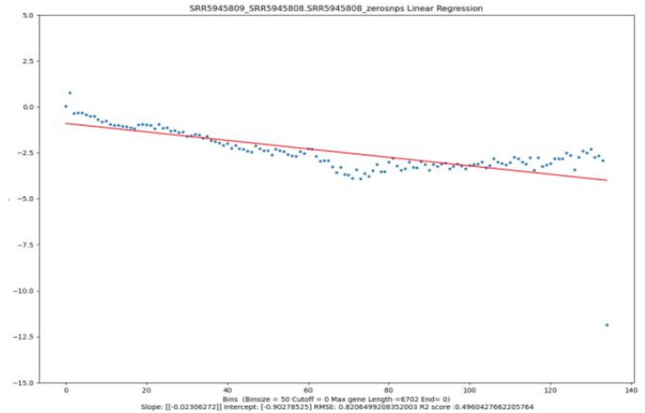
```
Slope: [[-0.02621034]] Intercept: [-0.82958672] RMSE: 9.435046936299523 R2 score :0.3455628400069336
```

Zero SNPs

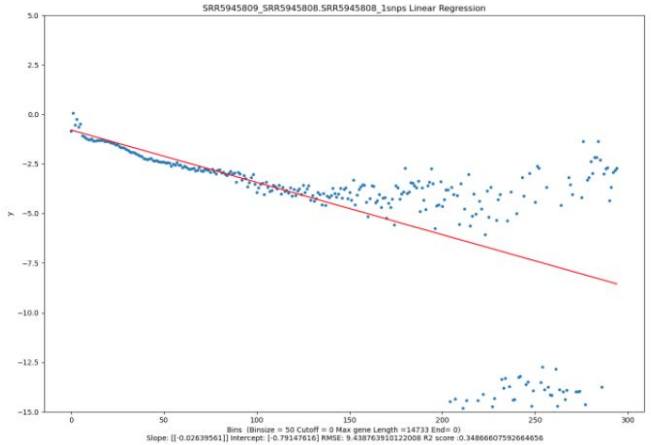
One or more SNPs







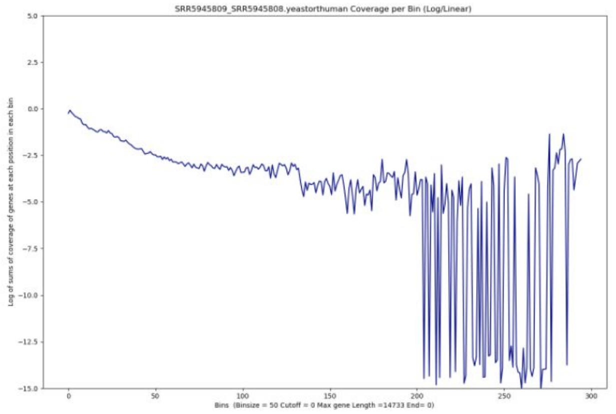
Slope: [[-0.02306272]] Intercept: [-0.90278525] RMSE: 0.8206499208352003 R2 score :0.4960427662205764

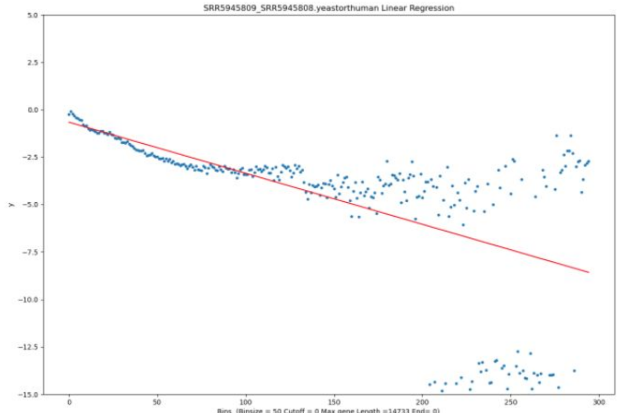


Slope: [[-0.02639561]] Intercept: [-0.79147616] RMSE: 9.438763910122008 R2 score :0.3486660759266456

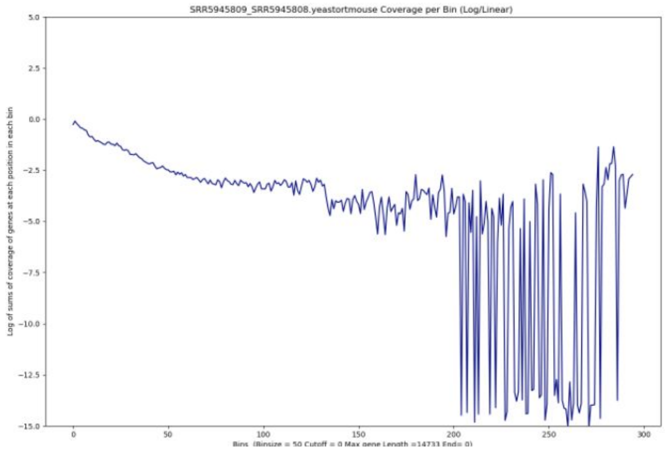
Orthologous to human

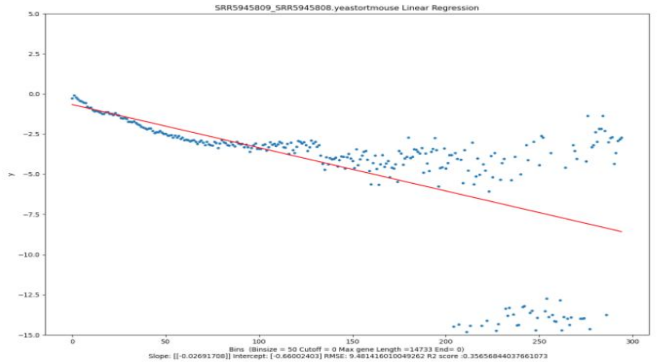
Orthologous to Mouse





Slope: [[-0.02691139]] Intercept: [-0.66189435] RMSE: 9.481267893850392 R2 score :0.35647506084193537

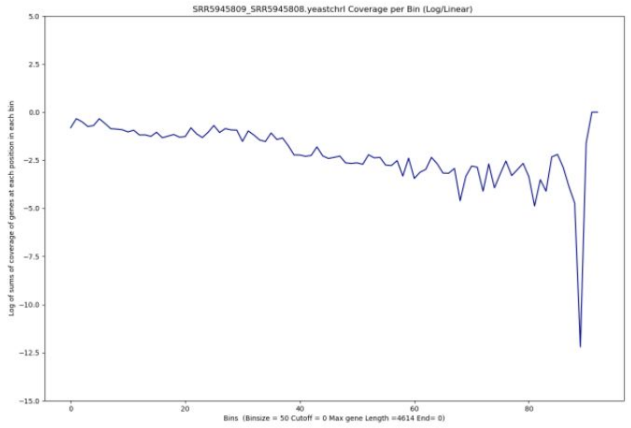


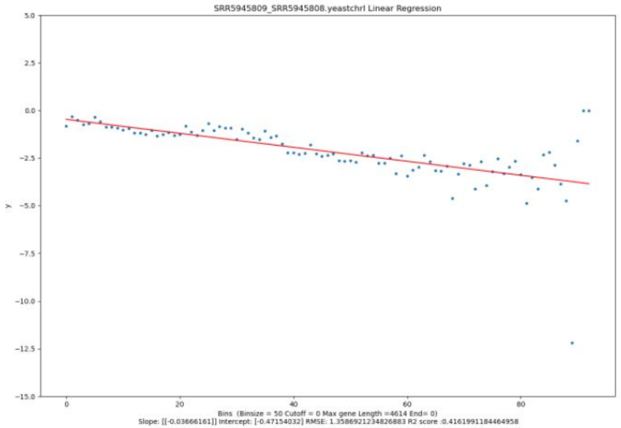


Slope: [[-0.02691708]] Intercept: [-0.66002403] RMSE: 9.481416010049262 R2 score :0.35656844037661073

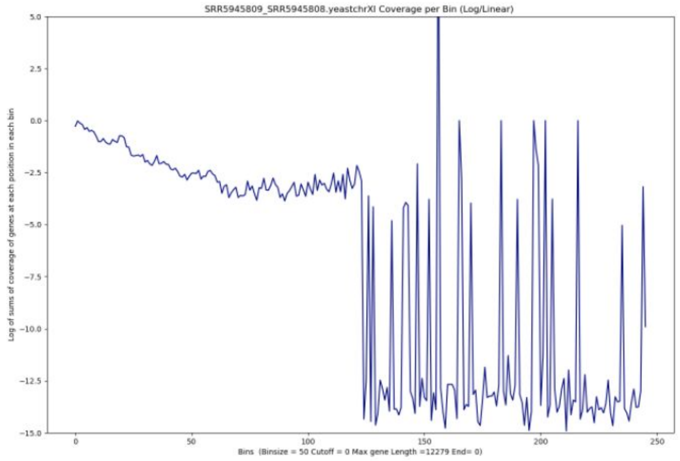
Chromosome I

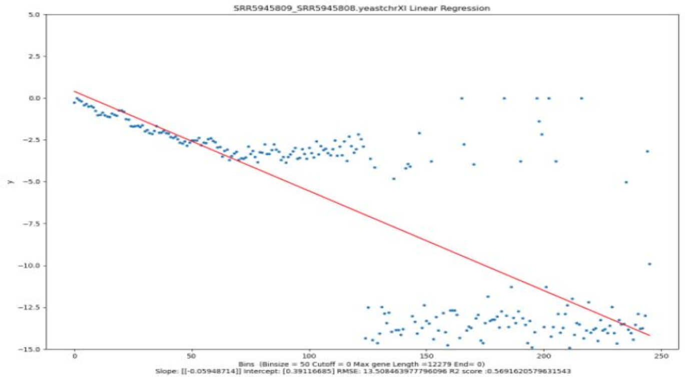
Chromosome XI





Slope: [[-0.03666161]] Intercept: [-0.47154032] RMSE: 1.3586921234826883 R2 score :0.4161991184464958





Slope: [[-0.05948714]] Intercept: [0.39116685] RMSE: 13.508463977796096 R2 score :0.5691620579631543

➔ We have more plots on the project documentation site:

https://ribosomesprofiling.readthedocs.io/en/latest/index.html

➔ We can see different patterns by eye, but how we can interpret results

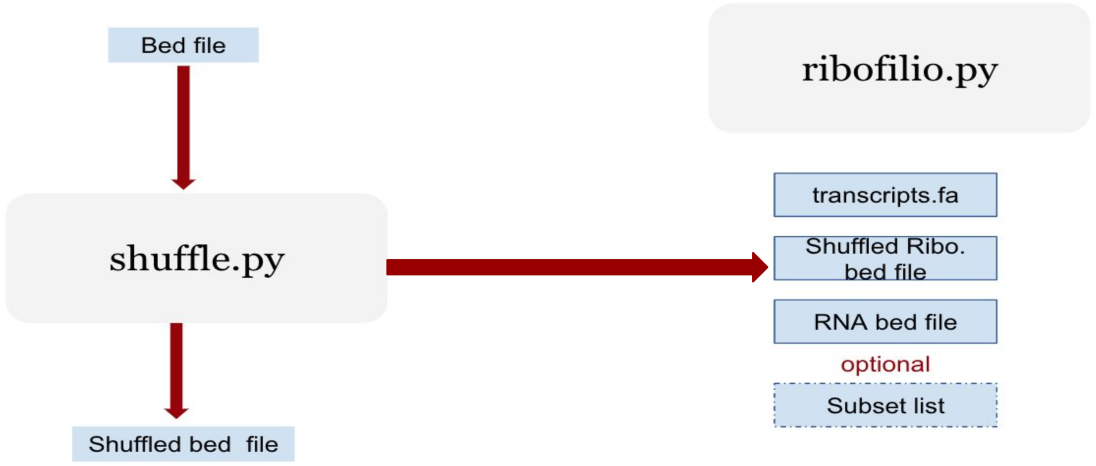
How to Interpret the Results

How far our slopes from random?

Using a bootstrap approach, we simulate footprints and compare ours slopes with the slopes of the simulated shuffled bed files

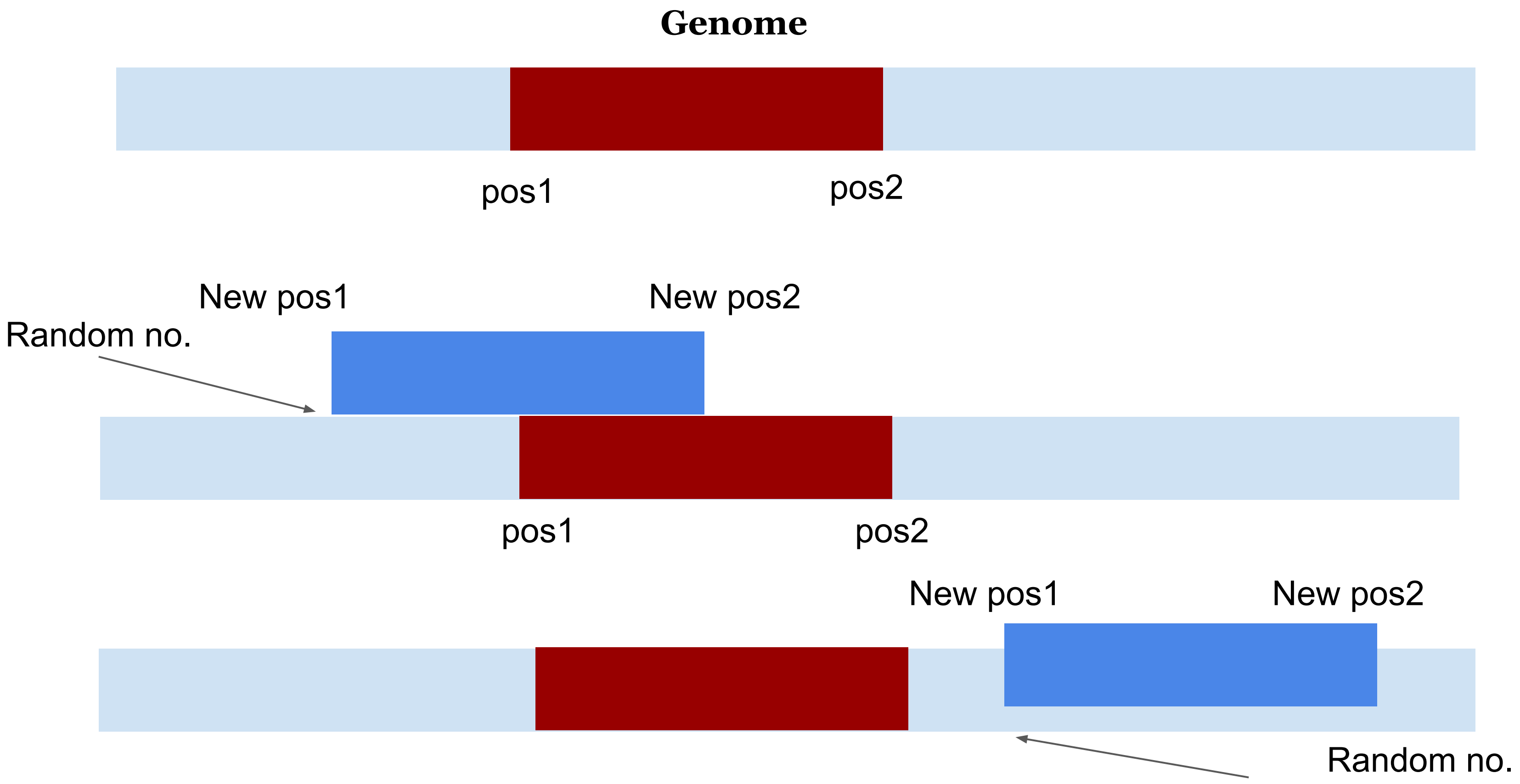
Create a 1000 simulated slopes from shuffled footprint bed files

Calculated the P-value of the of how far our real slope is from random



Shuffling

Generated a random number between [0 and (Gene Length - read Length)]



Main Slope

How far our real slope from the 1000 shuffled slopes?

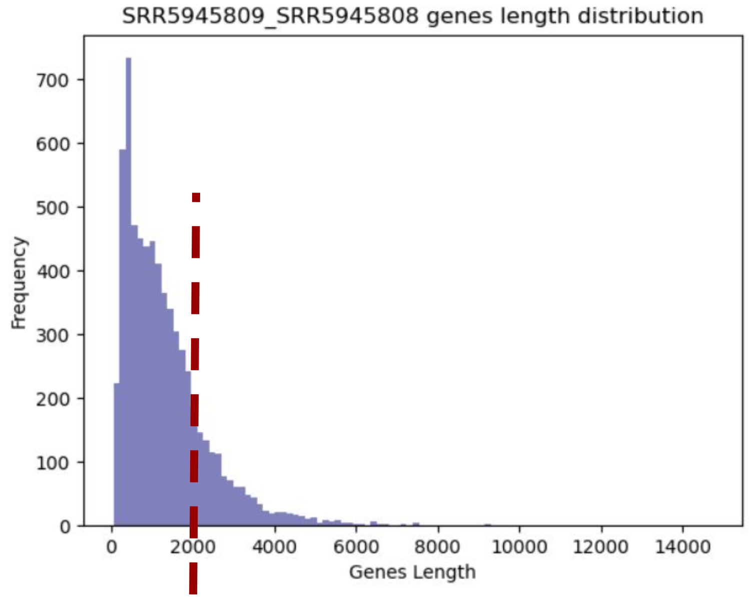
Sample	No. shuffled	P-value
SRR5945809/SRR5945808	1000	0.00

➔ **I ignored in this presentation the details of how we calculated the P-value**

Gene Ontology

GO		No.shuffled	P-Value
GO0006412	Translation	1000	0.00
GO0042254	Ribosome biogenesis	1000	0.00
GO0006950	Response to Stress	1000	1.00
GO0009651	Response Salt Stress	1000	0.139
.....More on the way			

Gene Length



How long is long?

	No.shuffled	**P-Value**
Genes with Length [0-2000]	414	1.0
Genes with Length > 2000	977	0.0

TPM

Run on Amazon AWS

- Mapped mRNA genes to transcript using Salmon
- Clustered genes based on TPM, run ribofilio on FP and mRNA with each cluster as a subset

	No.shuffled	P-Value
TPM > 50	500	1.0
TPM <= 50	83	0.0

Chromosomes

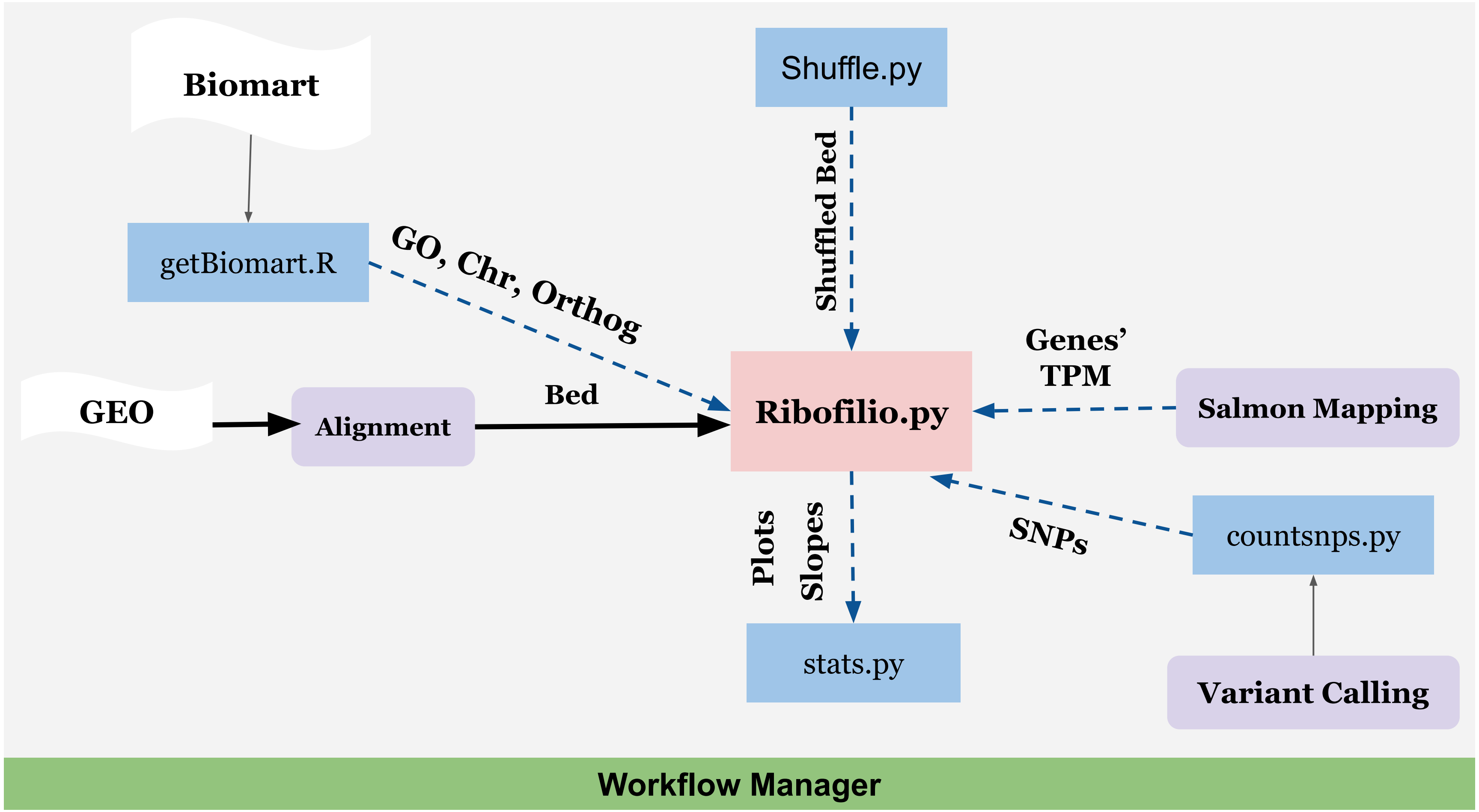
Run on Amazon AWS

Chromosome	Size	No.shuffled	P-Value
chrI	**126**	350	0.9942857142857143
chrVI	**156**	1000	0.003

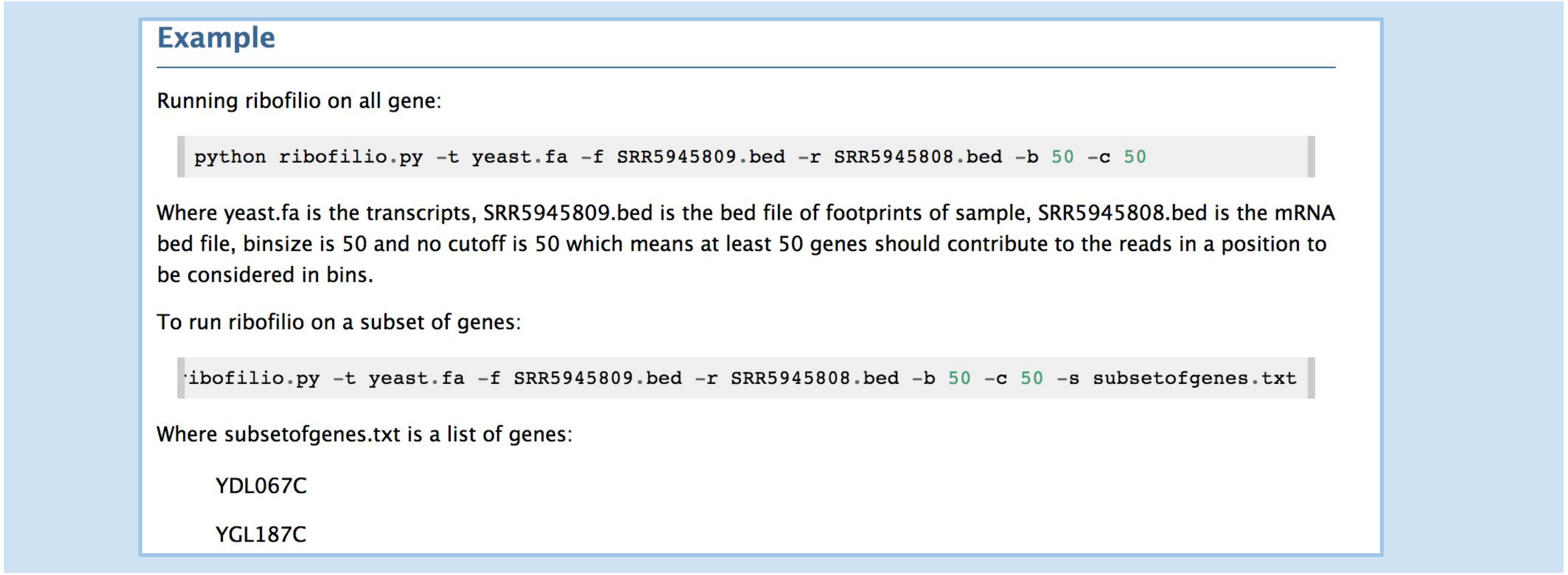
SNPs

SNPs	Size	No.shuffled	P-Value
Zero	2122	176	0.5568181818181818
1 or more SNP	4490	77	0.0

Reproducibility



- More datasets and factors running
- After finalizing the code currently on Github, we will push to conda/docker, binder, Jupyter, etc.
- Documentation is heavily changing as we progress



Example

Running ribofilio on all gene:

```
python ribofilio.py -t yeast.fa -f SRR5945809.bed -r SRR5945808.bed -b 50 -c 50
```

Where yeast.fa is the transcripts, SRR5945809.bed is the bed file of footprints of sample, SRR5945808.bed is the mRNA bed file, binsize is 50 and no cutoff is 50 which means at least 50 genes should contribute to the reads in a position to be considered in bins.

To run ribofilio on a subset of genes:

```
ibofilio.py -t yeast.fa -f SRR5945809.bed -r SRR5945808.bed -b 50 -c 50 -s subsetofgenes.txt
```

Where subsetofgenes.txt is a list of genes:

YDL067C

YGL187C

https://ribosomesprofiling.readthedocs.io/en/latest/index.html

Thank You