USC Department Of Translational Genomics
Keck Medicine of USC

TRGN 599: Applied Data Science and Bioinformatics

UNIT VI. Enrichment Analysis, Linear Regression

Week 13 - Lecture 2

Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S. | **Assistant Professor**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

David W. Craig, Ph.D. | **Professor and Vice Chair**

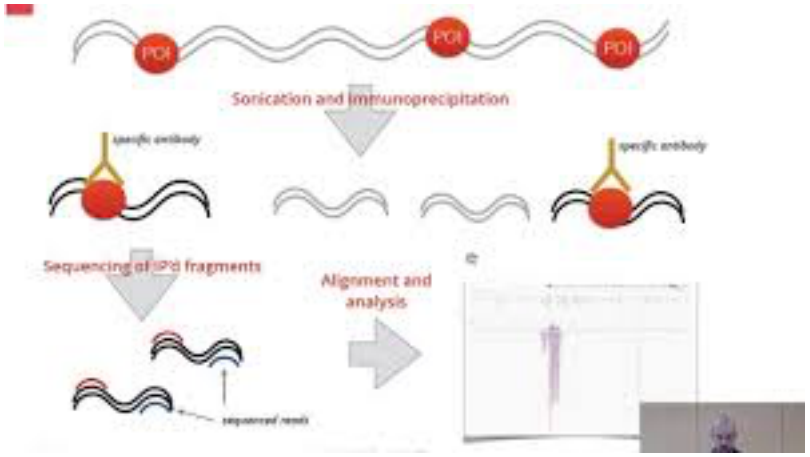
Dept. of Translational Genomics

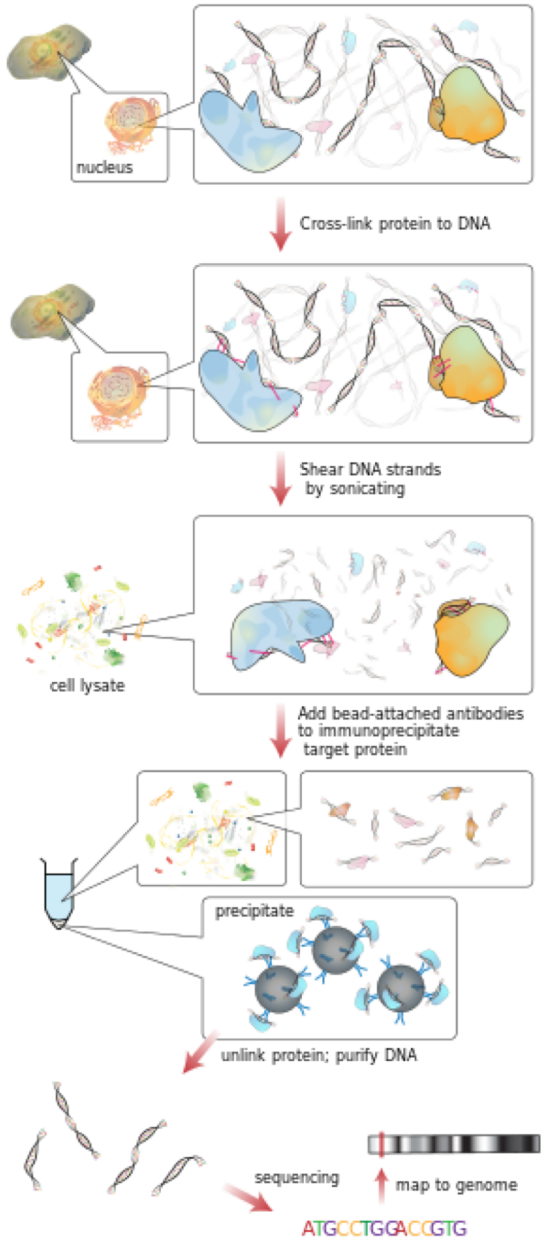
USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Co-Director, Institute of Translational Genomics

Topics

- Chromatin Immunoprecipitation (ChIP) - ChIPseq.





High-throughput sequencing of ChIP fragments

- Last class we generated a list of 26 peaks.

```
RangedData with 26 rows and 3 value columns across 1 space
       space        ranges   |       max    maxpos       sum
    <factor>     <IRanges>   | <integer> <integer> <integer>
1          1         1-261   |       285        96     60429
2          1     6452-8295   |      1983      7172   1454805
3          1   61278-61891   |       562     61522    230848
4          1   89394-90052   |       719     89834    248001
5          1 169521-169928   |       363    169711    108088
6          1 196049-196160   |       192    196104     20107
7          1 279578-280348   |       511    280092    232479
8          1 300972-301130   |       163    301073     25026
9          1 321577-322158   |       453    321920    174458
...      ...           ... ...       ...       ...       ...
18         1 545923-546424   |       327    546121    125455
19         1 560747-560868   |       188    560808     21008
20         1 561231-561792   |       313    561452    121825
21         1 635495-635721   |       285    635645     55198
22         1 635769-635976   |       229    635935     39015
23         1 691546-691970   |       387    691715    121786
24         1 737874-739387   |       857    738740    691885
25         1 763761-764625   |       378    764123    206480
26         1 782526-784469   |      2231    784190   1666690
```

- And we converted to a Granges object for subsequent analysis.

```
321 ```{R}
322 bind.sites <- GRanges(seqnames='chrXIV',ranges=unlist(ranges(chip.sel.peaks)),strand='*')
323 ```
```

Connecting Annotation to Peaks

- Once peaks are produced in GRanges format we can connect genomic annotation to these regions.
  - Similarly than the RNA-seq data.
- Transcript annotation of the yeast can be found in the TxDb.Scerevisiae.UCSC.sacCer3.sgdGene package.

````
345 library(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene)
346 txdb <- TxDb.Scerevisiae.UCSC.sacCer3.sgdGene
347
348 seqlevels(txdb)
349 #seqlevels(txdb, force=TRUE) <- c("chrXIV")
350
351 ```
````

```
 [1] "chrI"    "chrII"   "chrIII"  "chrIV"   "chrV"    "chrVI"   "chrVII"  "chrVIII" "chrIX"   "chrX"    "chrXI"   "chrXII"  "chrXIII" "chrXIV"  "chrXV"   "chrXVI"  "chrM"
```

````
352
353 ```{R}
354
355 genelist <- genes(txdb)
356
357 head(genelist)
358
359 ```
````

```
GRanges object with 6 ranges and 1 metadata column:
       seqnames      ranges strand |     gene_id
          <Rle>   <IRanges>  <Rle> | <character>
  Q0010     chrM   3952-4415      + |       Q0010
  Q0032     chrM 11667-11957      + |       Q0032
  Q0055     chrM 13818-26701      + |       Q0055
  Q0075     chrM 24156-25255      + |       Q0075
  Q0080     chrM 27666-27812      + |       Q0080
  Q0085     chrM 28487-29266      + |       Q0085
  -------
  seqinfo: 17 sequences (1 circular) from sacCer3 genome
```

Connecting Annotation to Peaks

- Limiting the annotation to Chromosome 14.

```
353 ```{R}
354 seqlevels(txdb) <- c("chrXIV")
355
356 seqlevels(txdb)
357 ```
```

```
[1] "chrXIV"
```

```
358
359 ```{R}
360
361 genelist <- genes(txdb)
362
363 head(genelist)
364
365 ```
```

```
GRanges object with 6 ranges and 1 metadata column:
          seqnames        ranges strand |     gene_id
             <Rle>     <IRanges>  <Rle> | <character>
  YNL001W   chrXIV 627456-628616      + |      YNL001W
  YNL002C   chrXIV 626174-627142      - |      YNL002C
  YNL003C   chrXIV 624975-625829      - |      YNL003C
  YNL004W   chrXIV 622915-624621      + |      YNL004W
  YNL005C   chrXIV 621313-622428      - |      YNL005C
  YNL006W   chrXIV 620067-620978      + |      YNL006W
  -------
  seqinfo: 1 sequence from sacCer3 genome
```

Connecting Annotation to Peaks

- As the genelist is now in an appropriate format, we can directly apply the findOverlaps() function again to overlap our generated **bind.sites** object and **genelist** object

````
```{R}

covers <- findOverlaps(bind.sites,genelist,ignore.strand=T)

affected.genes <- as.data.frame(genelist[subjectHits(covers)])

head(affected.genes)

```
````

	seqnames	start	end	width	strand	gene_id
YNL337W	chrXIV	7165	7419	255	+	YNL337W
YNL338W	chrXIV	6561	6719	159	+	YNL338W
YNL304W	chrXIV	60297	61857	1561	+	YNL304W
YNL242W	chrXIV	191324	196102	4779	+	YNL242W
YNL192W	chrXIV	276502	279897	3396	+	YNL192W
YNL179C	chrXIV	300666	301103	438	-	YNL179C

Connecting Annotation to Peaks

- The binding of the **target proteins** in the origin recognition complex (ORC) strongly affects the expression of these genes.
- The org.Sc.sgd.db package uncovers the details of these genes.

```{R}

#if (!requireNamespace("BiocManager", quietly = TRUE))
#    install.packages("BiocManager")
#BiocManager::install("org.Sc.sgd.db", version = "3.8")

library(org.Sc.sgd.db)

select(org.Sc.sgd.db, as.data.frame(affected.genes)$gene_id, keytype="ORF",columns=c("GENENAME","ENTREZID"))

```

```
'select()' returned 1:1 mapping between keys and columns
         ORF        SGD GENENAME ENTREZID
1    YNL337W S000005281     <NA>     <NA>
2    YNL338W S000005282     <NA>     <NA>
3    YNL304W S000005248    YPT11   855412
4    YNL242W S000005186     ATG2   855479
5    YNL192W S000005136     CHS1   855529
6    YNL179C S000005123     <NA>     <NA>
7    YNL143C S000005087     <NA>   855579
8    YNL114C S000005058     <NA>     <NA>
9    YNL112W S000005056     DBP2   855611
10   YNL093W S000005037    YPT53   855631
11   YNL094W S000005038     APP1   855630
12 YNL067W-A S000007623     <NA>     <NA>
13 YNL067W-B S000028810     <NA>  1466514
14   YNL067W S000005011    RPL9B   855658
15   YNR004W S000005287     SWM2   855738
16   YNR059W S000005342     MNT4   855796
17   YNR077C S000005360     <NA>   855814
```

Analysis of binding site motifs

- Another goal of ChIP studies is to identify sequence motifs to which different TFs (or another chromatin binding proteins) attach.

- To perform this analysis you should become familiar with the Bioconductor package rGADEM.

- We will not review this specific analysis but in the provided Rmarkdown for this lecture there is an example (including data) about using this specific analysis.

Chip-chip and ChIPseq Analysis

- Thank you!