Homework2

Name: Nuttapong Mekvipad, Ryan William Moreau, Rani Nielsen, Silvija Pupsaite, Liuqing Zheng

Group: 7

Question 1

a) The first five nucleotides of this transcript on the genome browser were 'AAAGG'.

b) The first five nucleotides of the raw AK002007 from GenBank were 'GAAGC'.

c) The discrepancy between the first five nucleotide in the genome browser and in the raw sequence might be due to the sequence found in genome browser was the result of mapping the raw AK002007 to human genome. Some part of raw mRNA (in this case first 7 nucleotides) might not be mapped to the genome due to various reasons. One of which, could be that the first 7 bp's might belong to other exon and this AK002007 is a truncated sequence without the rest of that exon. Additionally, it could be that there were sequencing error at the end of sequence so the similarity at the end of sequence might not be significant enough for mapping.

Question 2

a) Finding the coverage of ERa na ERb sites on chromosome

We sorted the ERa_hg18.bed and ERb_hg18.bed by chromosome number and starting position of ERa or ERb sites. Then caluculated the coverage of ERa and ERb site on each chromosome.

```
sort -k1,1 -k2,2n ERa_hg18.bed > Sorted_ERa_hg18.bed

sort -k1,1 -k2,2n ERb_hg18.bed > Sorted_ERb_hg18.bed

nice bedtools genomecov -i Sorted_ERa_hg18.bed -g hg18_chrom_sizes.txt -max 1 > ERa_coverage.txt

nice bedtools genomecov -i Sorted_ERb_hg18.bed -g hg18_chrom_sizes.txt -max 1 > ERb_coverage.txt
```

We then loaded results into R for plotting.

```
era_coverage <- read_tsv('ERa_coverage.txt', col_names = FALSE)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_double()
## )
```

```
erb_coverage <- read_tsv('ERb_coverage.txt', col_names = FALSE)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_double()
```

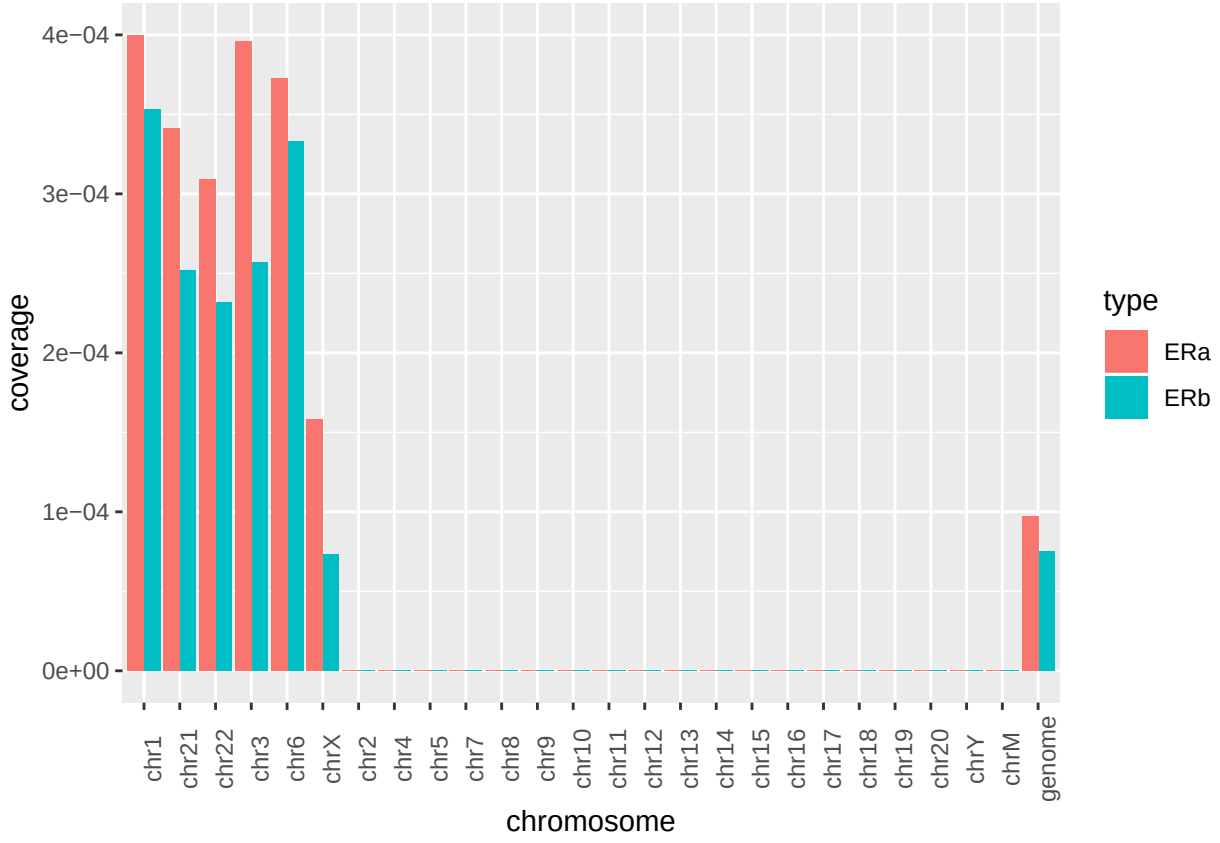
```
## )
```

```
era_coverage_prop <- filter(era_coverage, X2==0) %>%
  transmute(chromosome = X1, ERa= 1 - X5)
erb_coverage_prop <- filter(erb_coverage, X2==0) %>%
  transmute(chromosome = X1, ERb = 1 - X5)

# convert chromosome to factor to prevent ggplot from sorting values
era_coverage_prop$chromosome <- factor(era_coverage_prop$chromosome,
                                       levels = era_coverage_prop$chromosome)

erb_coverage_prop$chromosome <- factor(erb_coverage_prop$chromosome,
                                       levels = erb_coverage_prop$chromosome)

full_join(era_coverage_prop, erb_coverage_prop, by = 'chromosome') %>%
  gather(key = "type", value = "coverage", -chromosome) %>%
  ggplot(aes(x=chromosome, y=coverage, fill=type)) + geom_bar(stat = "identity",
                                                              position = "dodge") +
  theme(axis.text.x = element_text(angle = 90))
```



From the results, we can see that ERa and ERb site cover only the chromosome 1, 3, 6, 21, 22 and X. There are two possible reasons why we don't find the ERa and ERb sites on other chromosomes. One is that there actually is no ERa and ERb sites on the other chromosomes. However, it is unlikely for both sites to not be found on the rest of the 17 chromosomes. Furthermore, when we looked up the distribution of ERa and ERb sites, we found that they actually could be found on other chromosomes. For example, the ERa site could be found in the wnt11 and CTSD genes which are on chromosome 11 (Lin et al. 2007). Therefore, the reason might be that people who performed the experiment only searched for the ERa and ERb sites on just

chromosome 6. This experimental limitation could be found in paper doing similar experiment such as Liu et al. 2008.

b) Finding the number of overlapping ERa and ERb sites

Next we wanted to see how much ERa and ERb sites overlap in data set. We used bedtools with following command to find the number of non-overlap sites.

```
bedtools intersect -a ERa_hg18.bed -b ERb_hg18.bed -v | wc -l  # 236
bedtools intersect -a ERb_hg18.bed -b ERa_hg18.bed -v | wc -l  # 140
```

Then we plotted Venn diagram and found that there are a total of 582 ERa sites and 486 ERb sites with 346 overlapping sites between ERa and ERb.

```
library(VennDiagram)
```

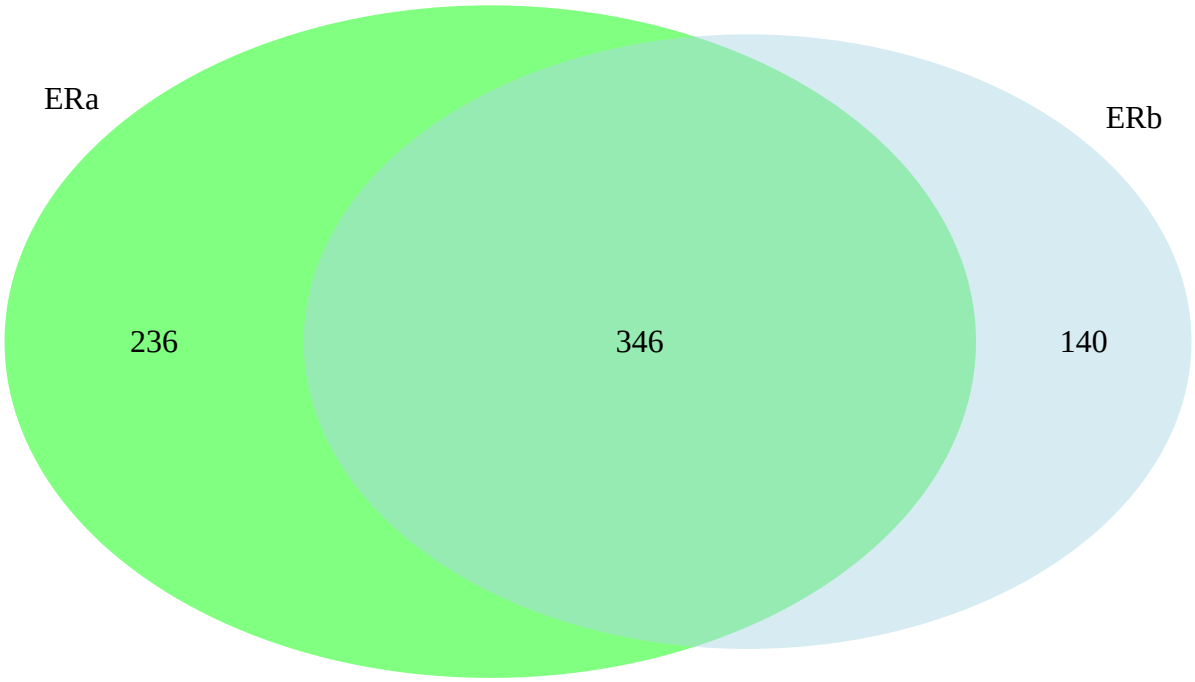
```
## Warning: package 'VennDiagram' was built under R version 3.4.4
```

```
## Loading required package: grid
```

```
## Loading required package: futile.logger
```

```
## Warning: package 'futile.logger' was built under R version 3.4.4
```

```
grid.newpage()
draw.pairwise.venn(582, 486, 346, category = c("ERa", "ERb"),
                   lty = rep("blank", 2), fill = c("green", "light blue"))
```




```
## (polygon[GRID.polygon.72], polygon[GRID.polygon.73], polygon[GRID.polygon.74], polygon[GRID.polygon.
```

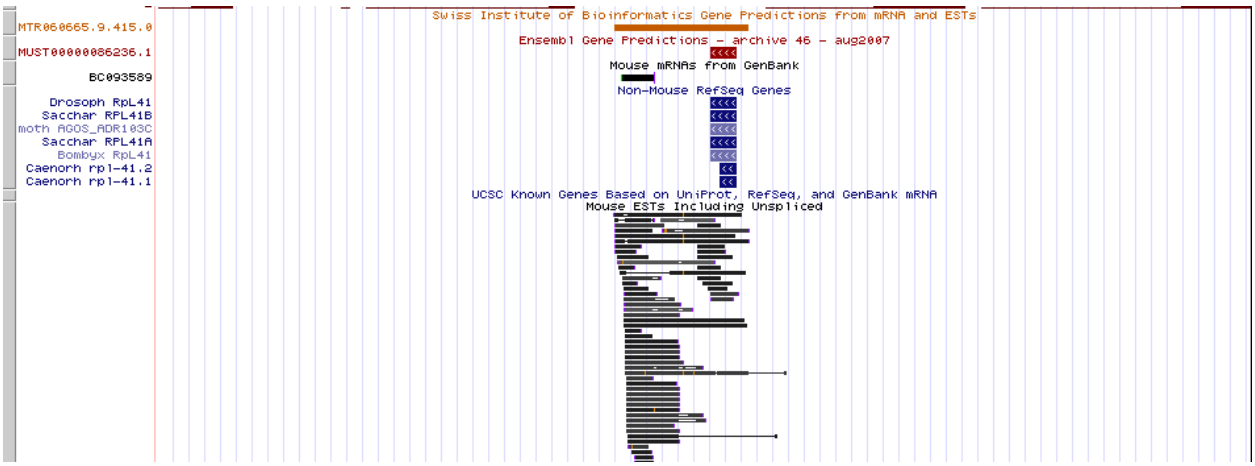


Figure 1: Predicted mouse gene structure

Question 3

To examine the argument of professor, we first looked at the possible structure of the gene that this 80 bp sequence belongs. Either, while looking at the tracks under gene and gene prediction e.g. RefSeq, Ensembl or SIB Genes, or looking at mRNA and ESTs tracks. From the gene and gene prediction tracks, there were 2 possible structures of the gene. These include a gene with one long exon (SIB Genes) and a gene with one short exon (Ensembl). The ESTs information also supported both structure (Figure 1).

Next we looked at the BLAT result of *D. melanogaster rpl*41, we can see from the figure (Figure 2) that the sturcture of *D. melanogaster rpl*41 was very different from the potential gene in the mouse. *D. melanogaster rpl*41 had 3 exons, and the conserved 80 bp sequence belongs to th 2nd exon.

We also looked at the conservation along this potential gene in the mouse (Figure 3) and compared it with the conservation along the gene in *D. melanogaster* (Figure 2). We can see that in *D. melanogaster* the conservation was very high in all 3 exon regions. While, in the mouse the conservation of the gene was high, it was only at the region that the 80 bp sequence comes from. This different in structure might be due to the mouse gene being a truncated version of *D. melanogaster rpl*41, and the actual functional region is the conserved 80 bp region.

These discrepancies between the structure of *D. melanogaster rpl*41 and the mouse gene led us to question whether this sequence actually belongs to active ribosomal protein gene equivalent to *D. melanogaster rpl*41, or is part of rpl41 or not.

To check the validity of the gene, we next looked at Broad H3 ChIPseq track to see the distribution of histone mark around the mouse potential gene. If this gene is actually an active functional gene, we should see some activation mark like H3K4Me1 to H3K4Me3. However, we can see from the figure (Figure 3) that in the mouse neuroprogenitor cells the mark found is mostly around the upstream of the gene where repressive marks such as H3K27Me3 or H3K9Me3 and very low level of H3K4Me1 to H3K4Me3 could be seen.

Interestingly, when we turned on RepatMasker track, we could see that in mouse there were many transposon e.g. LINE or SINE directly upstream of the potential gene, while the region around *D. melanogaster rpl*41 were free of transposon. This may explain why there were many repressive histone marks up-and downstream.

This information implies that this 80 bp sequence might belong to the truncated mouse rpl41 gene that may not be active due to the transposon upstream of this truncated gene.

\ \ \

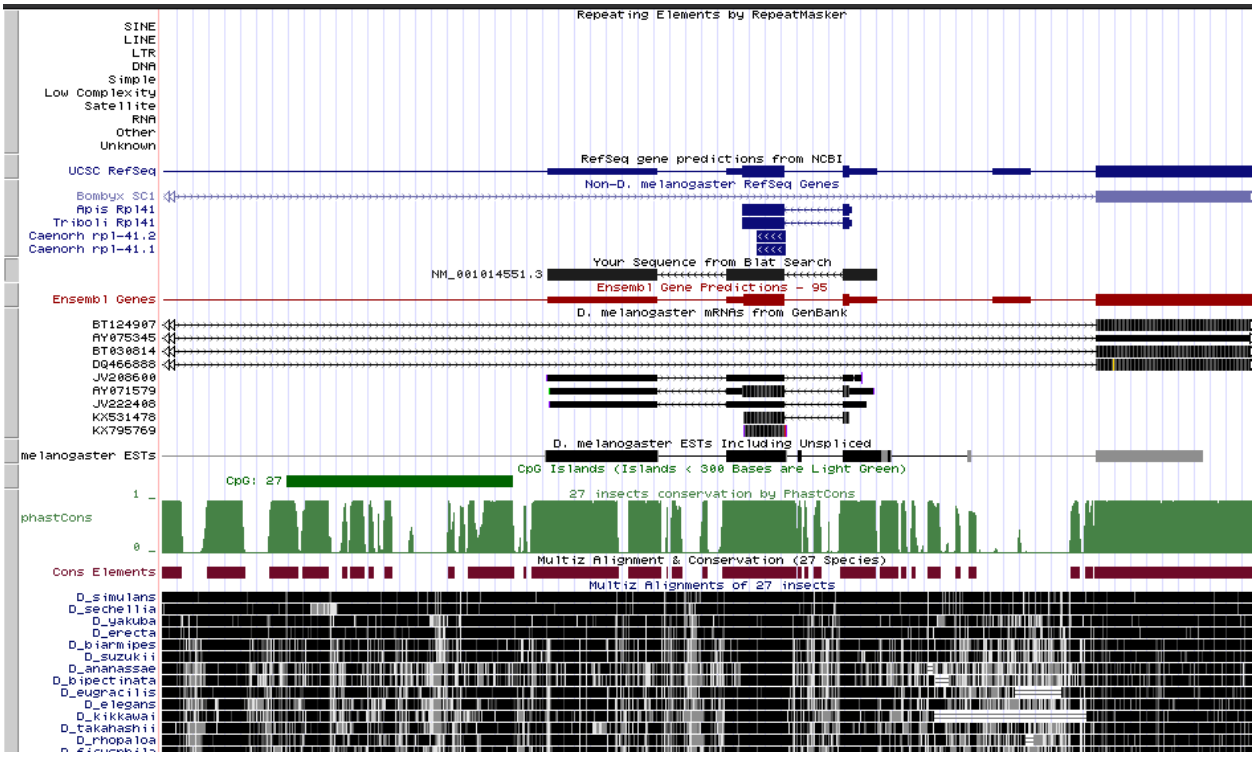


Figure 2: Overall genome browser tracks of *D. melanogaster rpl*41

Figure 3: Overall genome browser tracks of mouse gene

References

Lin, Z., Reierstad, S., Huang, C.C. and Bulun, S.E., 2007. Novel estrogen receptor-alpha binding sites and estradiol target genes identified by chromatin immunoprecipitation cloning in breast cancer. Cancer research, 67(10), pp.5017-5024.

Liu, Y., Gao, H., Marstrand, T.T., Str?m, A., Valen, E., Sandelin, A., Gustafsson, J.?. and Dahlman-Wright, K., 2008. The genome landscape of ER-alpha-and ER-beta-binding DNA regions. Proceedings of the National Academy of Sciences, 105(7), pp.2604-2609.