Homework 2

Marie Holm Abildgaard, Natthawut Adulyanukosol, Mathias Isager Fessler, Enrique Goñi Echeverria, Adrian Otamendi Laspiur, Katrine Harpelunde Poulsen

Part I: Dicer dissected

a) What are the first five genomic nucleotides from the first exon of this transcript?
5'-AAAGG-3'

b) Look at the raw mRNA sequence of AK002007, from the database it actually comes from. What are the first five nucleotides?
5'-GAAGC-3'

c) How do you explain the discrepancy (maximum 5 lines)?
There are seven additional nucleotides (gaagcaa) in the raw sequence of AK002007, which do not align to the genome. These nucleotides matches the end of the preceding exon from the complete isoforms of DICER1 and the AK002007 sequence could be a product of fragmentation of the mRNA during sample preparation, library construction or a sequencing error.

Part II: ERa and ERb

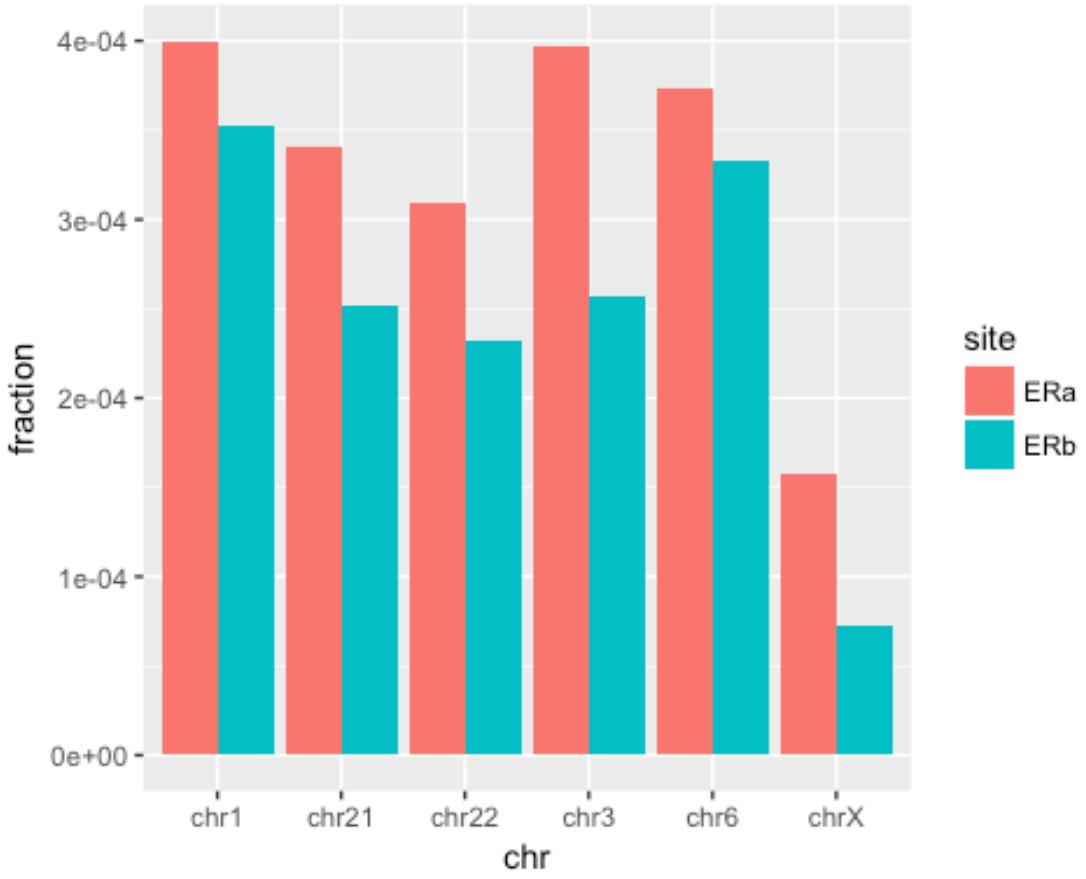
a) Plot the fractions for all chromosomes as a single barplot in R. Briefly comment the results. Is there anything particularly surprising?

{bash}

```
# Pre-process the bed files by sorting instances
bedtools sort -i ERa_hg18.bed > out/ERa_hg18_sorted.bed
bedtools sort -i ERb_hg18.bed > out/ERb_hg18_sorted.bed

# Calculate genome coverage
bedtools genomecov -i out/ERa_hg18_sorted.bed -g hg18_chrom_sizes.txt > out/E
Ra_hg18_sorted.bed

bedtools genomecov -i out/ERb_hg18_sorted.bed -g hg18_chrom_sizes.txt > out/E
Rb_hg18_sorted.bed
```

{r}

```r
# Load packages
library("ggplot2")
library("VennDiagram")
# Read the coverage files
df_ERa <- read.table("out/ERa_hg18_coverage.txt")
df_ERb <- read.table("out/ERb_hg18_coverage.txt")
# Add a column with the site name
df_ERa$site <- "ERa"
df_ERb$site <- "ERb"
# Combine 2 dataframes
df_ER <- rbind(df_ERa, df_ERb)
# Edit column names
colnames(df_ER) <- c("chr", "depth", "coverage_size", "chr_size", "fraction",
"site")
# Remove rows of 'genome'
df_ER <- df_ER[df_ER$chr != 'genome', ]
# Keep only rows with coverage
df_ER_hits <- df_ER[df_ER$depth == 1, ]
# Plot
p <- ggplot(df_ER_hits, aes(chr, fraction))
p + geom_bar(stat = "identity", aes(fill = site), position = "dodge")
```



The ChIP study only identified ERa sites and ERb sites at 6 chromosomes (chromosome 1, 3, 6, 21, 22 and X). It seems unlikely that there are only ERa and ERb sites on the 6 chromosomes and not a single sites on the remaining chromosomes. A potential explanation of the restricted distribution of the ERa and ERb sites to the 6 chromosomes could be that the tiling array is not genome-wide but limited to the 6 chromosomes.

On all 6 chromosomes the fraction of ERa sites are larger than the fraction of ERb sites which might suggest that a larger number of genes that are controlled by ERa compared to ERb.The fractions of ERa and ERb binding sites are similar across the autosomes. However, the fractions on X chromosomes is less than a half of those in the autosomes. If the gene densities are similar across all chromosomes, this finding could suggest there are less genes controlled by ERa and ERb on X chromosomes compared to the autosomes.

b) How many ERA sites do/do not overlap ERB sites, and vice versa?

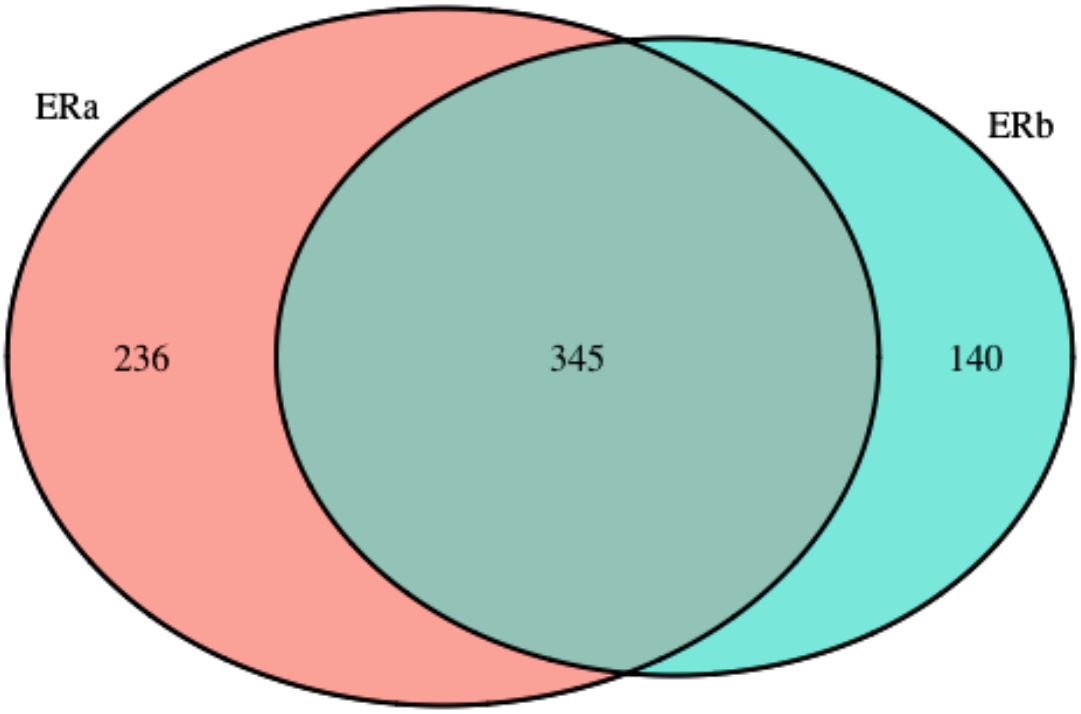
{bash}

```
wc -l out/ERa_hg18_sorted.bed
# A = 581

wc -l out/ERb_hg18_sorted.bed
# B = 485

bedtools intersect -a out/ERa_hg18_sorted.bed -b out/ERb_hg18_sorted.bed -c |
awk '$4 == 1' | wc -l
# A and B = 345
```

{r}

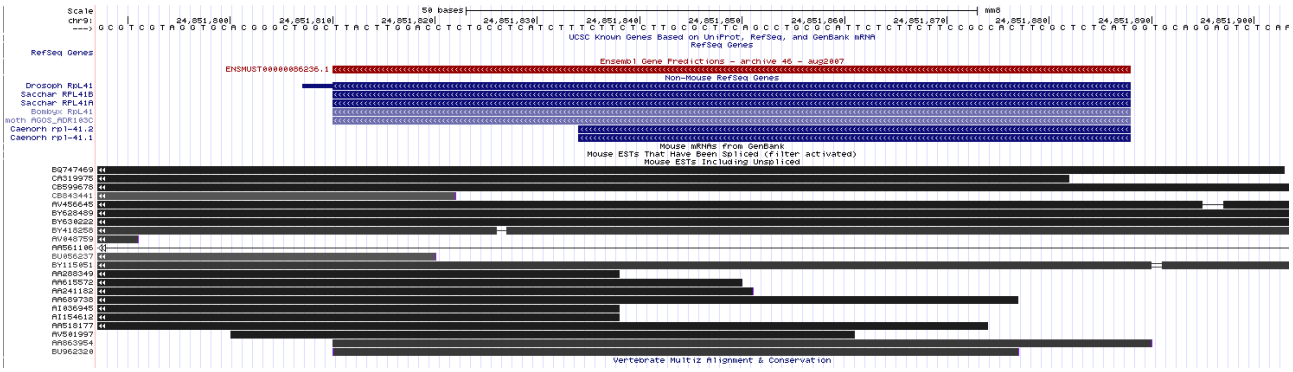
```
venn.plot <- draw.pairwise.venn(581, 485, 345,
                                c("ERa", "ERb"), fill = c("salmon", "turquois
e"));
grid.draw(venn.plot);
grid.newpage();
```

345 binding sites can be bound by both ERa and ERb. In the literature (Cowley et al., 1997), it is evident that ERa and ERb can form heterodimers in addition to homodimers. There are 236 sites specific to ERa and 140 sites specific to ERb which might account for the homodimers.

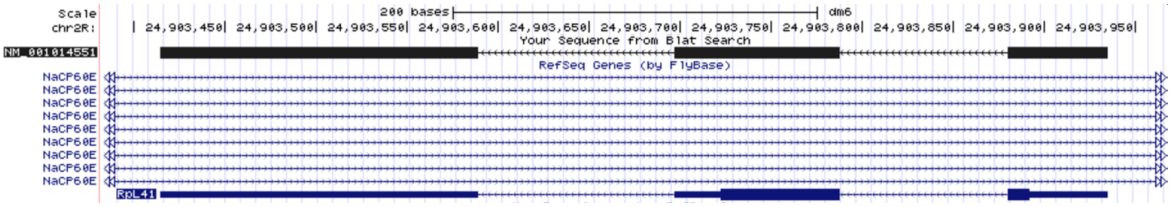


Part III: Ribosomal Gene

First of all, when looking at the genome browser the hypothesized new ribosomal protein (sequence chr9:24,851,809-24,851,889) does not contain any introns (red), and therefore it is very unlikely that it is a functional gene. Besides, there is no evidence for the transcription of the mouse gene as there are no mRNAs from GenBank and ESTs matching the gene:



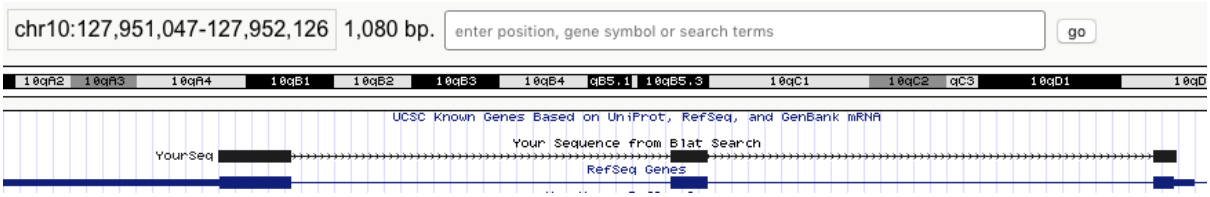
When performing a BLAT of the same sequence to the fly genome, you get this instead:



Here, we can see that the fly gene is actually three exons long, which makes it likely that it is protein coding here. When doing a BLAT of the sequence chr9:24,851,809-24,851,889 to the mouse's own genome (assembly mm8) a lot of results show up with high identity on many different chromosomes:

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	YourSeq	81	1	81	81	100.0%	14	-	104567546	104567626	81
browser details	YourSeq	81	1	81	81	100.0%	9	+	24851810	24851890	81
browser details	YourSeq	81	1	81	81	100.0%	13	+	112714411	112714491	81
browser details	YourSeq	81	1	81	81	100.0%	10	+	43144838	43144918	81
browser details	YourSeq	79	1	81	81	98.8%	16	+	3932207	3932287	81
browser details	YourSeq	79	1	81	81	100.0%	10	+	127951335	127951911	577
browser details	YourSeq	77	1	81	81	97.6%	11	-	12548092	12548172	81
browser details	YourSeq	77	1	81	81	97.6%	17	+	12680974	12681054	81
browser details	YourSeq	77	1	80	81	98.8%	1	+	51407765	51407885	121
browser details	YourSeq	75	1	81	81	92.4%	15	+	28003396	28003473	78
browser details	YourSeq	73	1	81	81	95.1%	18	+	10274593	10274673	81
browser details	YourSeq	72	2	81	81	95.0%	2	+	113896864	113896943	80
browser details	YourSeq	71	1	81	81	93.9%	2	-	150618772	150618852	81
browser details	YourSeq	70	2	81	81	95.0%	13	+	55192053	55192147	95
browser details	YourSeq	68	2	81	81	92.5%	16	+	38488929	38489008	80
browser details	YourSeq	67	1	70	81	98.6%	17	+	6779920	6779993	74
browser details	YourSeq	67	1	69	81	98.6%	12	+	81699244	81699312	69
browser details	YourSeq	66	2	81	81	91.3%	16	-	96214017	96214096	80
browser details	YourSeq	66	2	81	81	91.3%	1	+	147698430	147698509	80
browser details	YourSeq	65	1	70	81	97.2%	17	-	7797103	7797176	74
browser details	YourSeq	65	1	81	81	92.3%	16	+	20242235	20242320	86
browser details	YourSeq	63	2	81	81	94.5%	6	-	70898421	70898500	80
browser details	YourSeq	63	2	81	81	89.1%	4	-	131819711	131819788	78
browser details	YourSeq	62	1	81	81	83.2%	11	+	97117711	97117787	77
browser details	YourSeq	60	2	81	81	83.4%	11	+	20174315	20174392	78
browser details	YourSeq	59	16	81	81	95.4%	4	-	134649232	134649743	512
browser details	YourSeq	57	3	81	81	86.1%	11	-	43750932	43751010	79
browser details	YourSeq	49	4	62	81	91.6%	11	+	6176859	6176917	59
browser details	YourSeq	42	5	52	81	93.8%	8	+	10717096	10717143	48
browser details	YourSeq	42	2	61	81	85.0%	6	+	107061992	107062051	60
browser details	YourSeq	38	13	54	81	95.3%	7	-	57411777	57411818	42
browser details	YourSeq	37	43	81	81	97.5%	18	+	79300084	79300122	39
browser details	YourSeq	27	2	28	81	100.0%	6	+	28096491	28096517	27
browser details	YourSeq	27	5	35	81	86.7%	4	+	116541238	116541267	30
browser details	YourSeq	25	55	81	81	96.3%	7	-	34131782	34131808	27
browser details	YourSeq	24	13	36	81	100.0%	7	+	75772459	75772482	24
browser details	YourSeq	21	13	33	81	100.0%	12	-	32829422	32829442	21

Most of them look similar to that of chr9:24,851,809-24,851,889 (no introns). However, one of the hits on chromosome 10 looks like the fly Rpl41 gene (with multiple exons and introns, see figure on the next page) and it also matches very nicely with a mouse Rpl41 RefSeq gene:



Taking all this into consideration the most likely explanation is that the hypothesized new ribosomal protein is a processed pseudogene generated from the spliced Rpl41 mRNA, which has been reversed transcribed and re-inserted in the genome at the site the supervisor found. Ribosomal genes are reported to be the most common group of processed pseudogenes (Zhang et al., 2004). This explains why the sequence at chr9 is the same as on chr10 (picture above) but without introns. As the sequence is found at several places in the mouse genome, it has not only been inserted into the position the supervisor found.

Reference:

Cowley, S.M., Hoare, S., Mosselman, S., and Parker, M.G. (1997). Estrogen receptors alpha and beta form heterodimers on DNA. J. Biol. Chem. *272*, 19858–19862.

Zhang, Z., Carriero, N., and Gerstein, M. (2004). Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. *20*, 62–67.