Home work 2

Names: Abdullah Faqih Al Mubarok,August Nygaard Bodilsen, Natalie Christiansen, Maria Madrazo I Montoya

Group: 21

Question 1

The human DICER1 gene encodes an important ribonuclease, involved in miRNA and siRNA processing. Several mRNAs representing this gene have been mapped to the human genome (March 2006 assembly). We will look closer at one of them with the accession number AK002007.

a) What are the first five genomic nucleotides that are read by RNA polymerae II from this transcript?

The first five nucleotides (template strand) that the RNA polymerase II will read are: TTTCC. The first five nucleotides (coding strand) in the transcript are thus: AAAGG.

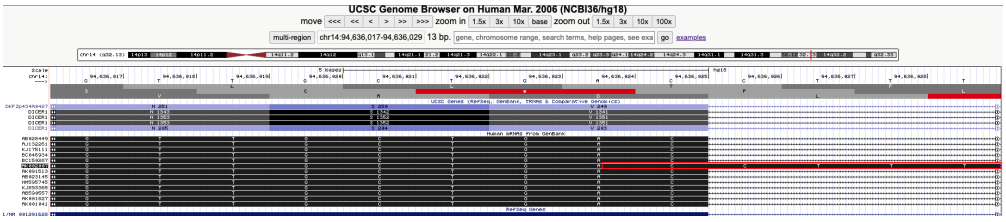


Figure 1: *AK002007 first five nucleotides from UCSC*

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

From GenBank we have GAAGC.

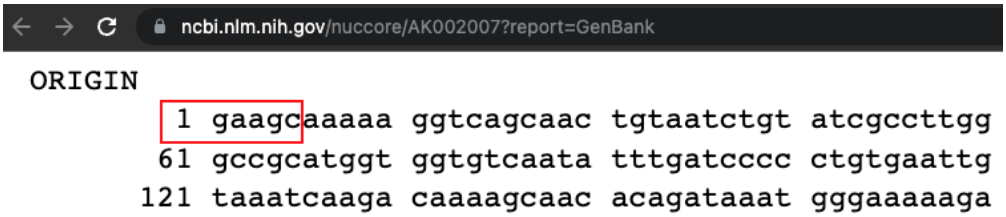


Figure 2: *AK002007 first five nucleotides from GenBank*

c) How do you explain the discrepancy (maximum 5 lines)?

The mismatch can be due to the usage of oligo-capping from the library preparation of the sequencing process. Despite mismatch for the first seven nucleotides, the downstream are match (see Figure 3).

Question 2

Our collaborators designed a ChIP study using so-called tilling arrays (an outdated technique these days, but the top of the pop at the time: see http://en.wikipedia.org/wiki/Tiling_array): one for estrogen receptor

cDNA AK002007

```
gaagcaaAAA GGTCAGCAAC TGTAATCTGT ATCGCCTTGG AAAAAAGAAG  50
GGACTACCCA GCCGCATGGT GGTGTCAATA TTTGATCCCC CTGTGAATTG  100
GCTTCCTCCT GGTTATGTAG TAAATCAAGA CAAAAGCAAC ACAGATAAAT  150
GGGAAAAAGA TGAAATGACA AAAGACTGCA TGCTGGCGAA TGGCAAACTG  200
GATGAGGATT ACGAGGAGGA GGATGAGGAG GAGGAGAGCC TGATGTGGAG  250
GGCTCCGAAG GAAGAGGCTG ACTATGAAGA TGATTTCCTG GAGTATGATC  300
AGGAACAcAT CAGATTTATA GATAATATGT TAATGGGGTC AGGAGCTTTT  350
GTAAAGAAAA TCTCTCTTTC TCCTTTTTCA ACCACTGATT CTGCATATGA  400
ATGGAAAATG CCCAAAAAAT CCTCCTTAGG TAGTATGCCA TTTTCATCAG  450
ATTTTGAGGA TTTTGACTAC AGCTCTTGGG ATGCAATGTG CTATCTGGAT  500
CCTAGCAAAG CTGTTGAAGA AGATGACTTT GTGGTGGGGT TCTGGAATCC  550
ATCAGAAGAA AACTGTGGTG TTGACACGGG AAAGCAGTCC ATTTCTTACG  600
ACTTGCACAC TGAGCAGTGT ATTGCTGACA AAAGCATAGC GGACTGTGTG  650
GAAGCCCTGC TGGGCTGCTA TTTAACCAGC TGTGGGGAGA GGGCTGCTCA  700
GCTTTTCCTC TGTTCACTGG GGCTGAAGGT GCTCCCGGTA ATTAAAAGGA  750
CTGATCGGGA AAAGGCCCTG TGCCCTACTC GGGAGAATTT CAACAGCCAA  800
CAAAAGAACC TTTCAGTGAG CTGTGCTGCT GCTTCTGTGG CCAGTTCACG  850
CTCTTCTGTA TTGAAAGACT CGGAATATGG TTGTTTGAAG ATTCCACCAA  900
GATGTATGTT TGATCATCCA GATGCAGATA AAACACTGAA TCACCTTATA  950
TCGGGGTTTG AAAATTTTGA AAAGAAAATC AACTACAGAT TCAAGAATAA  1000
GGCTTACCTT CTCCAGGCTT TTACACATGC CTCCTACCAC TACAATACTA  1050
TCACTGATTG TTACCAGCGC TTAGAATTCC TGGGAGATGC GATTTTGGAC  1100
TACCTCATAA CCAAGCACCT TTATGAAGAC CCGCGGCAGC ACTCCCCGGG  1150
GGTCCTGACA GACCTGCGGT CTGCCCTGGT CAACAACACC ATCTTTGCAT  1200
CGCTGGCTGT AAAGTACGAC TACCACAAGT ACTTCAAAGC TGTCTCTCCT  1250
GAGCTCTTCC ATGTCATTGA TGACTTTGTG CAGTTTCAGC TTGAGAAGAA  1300
TGAAATGCAA GGAATGGATT CTGAGCTTAG GAGATCTGAG GAGGATGAAG  1350
AGAAAGAAGA GGATATTGAA GTTCCAAAGG CCATGGGGGA TATTTTTGAG  1400
TCGCTTGCTG GTGCCATTTA CATGGATAGT GGGATGTCAC TGGAGACAGT  1450
CTGGCAGGTG TACTATCCCA TGATGCGGCC ACTAATAGAA AAGTTTTCTG  1500
CAAATGTACC CCGTTCCCCT GTGCGAGAAT TGCTTGAAAT GGAACCAGAA  1550
ACTGCCAAAT TTAGCCCGGC TGAGAGAACT TACGACGGGA AGGTCAGAGT  1600
CACTGTGGAA GTAGTAGGAA AGGGGAAATT TAAAGGTGTT GGTCGAAGTT  1650
ACAGGATTGC CAAATCTGCA GCAGCAAGAA GAGCCCTCCG AAGCCTCAAA  1700
GCTAATCAAC CTCAGGTTCC CAATAGCTGA AACCGCTTTT TAAAATTCAA  1750
AACAAGAAAC
```

Figure 3: *Alignment of AK002007*

alpha (ERA), one for estrogen receptor beta (ERB). All the sites are stored in BED files respectively for two ERs. These are now available in the homework directory, and are both mapped on hg18 genome. The current situation is that we know to some degree what ERA does, but not what ERB does (there are some evidence that they share some functions, but not all). So, we need bigger experiments and better statistics.

a) Using BEDtools within Linux: What is the genome coverage (% of base pair covered at each chromosome) for ERB and ERA sites? If you need a file with chromosome sizes for hg18, it included in the assignment: hg18_chrom_sizes.txt. Plot the fractions for all chromosomes as a single barplot in R. Briefly comment the results. Is there anything particularly surprising? Try to explain the outcome (biological and/or experimental setup explanations)?

First of all, we need to sort each of BED file before calculating the coverage

```
sort -k1,1 -k2,2n ERa_hg18.bed -o sorted_ERa_hg18.bed
sort -k1,1 -k2,2n ERb_hg18.bed -o sorted_ERb_hg18.bed
```

Then, we can do the coverage calculation

```
nice bedtools genomecov -i sorted_ERa_hg18.bed -g hg18_chrom_sizes.txt -max 1 > ERa_coverage
nice bedtools genomecov -i sorted_ERb_hg18.bed -g hg18_chrom_sizes.txt -max 1 > ERb_coverage
```

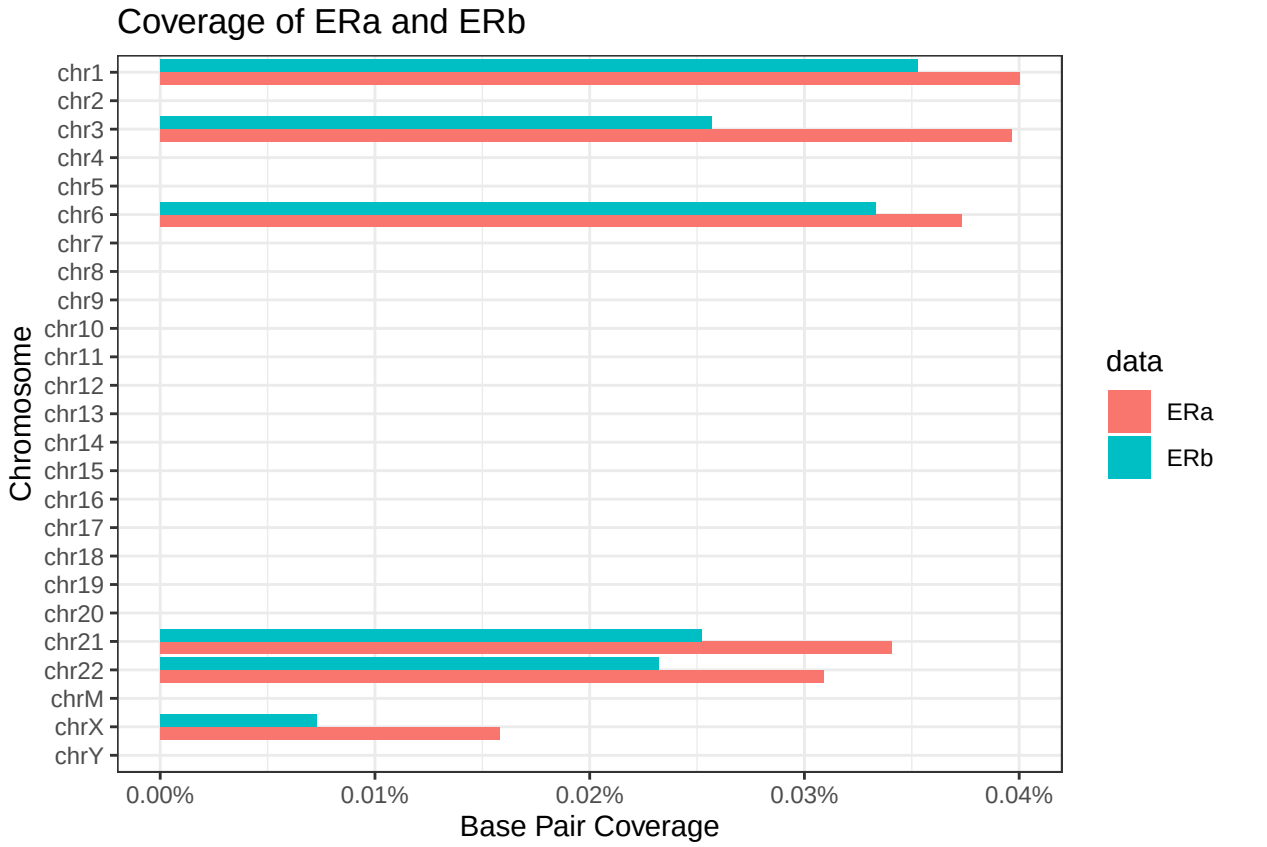
After that, we load the data into R for plotting the fractions

```
ERa_coverage <- read_tsv("ERa_coverage", col_names = FALSE, show_col_types = FALSE)
colnames(ERa_coverage) <- c("chrom", "depth", "n_bases", "chrom_size",
    "pct")
ERa_coverage <- ERa_coverage |>
    mutate(data = "ERa")
```

```
ERb_coverage <- read_tsv("ERb_coverage", col_names = FALSE, show_col_types = FALSE)
colnames(ERb_coverage) <- c("chrom", "depth", "n_bases", "chrom_size",
    "pct")
ERb_coverage <- ERb_coverage |>
    mutate(data = "ERb")

combined_coverage <- bind_rows(ERa_coverage, ERb_coverage)
combined_coverage <- combined_coverage |>
    filter(chrom != "genome") |>
    mutate(chrom = factor(chrom, levels = c(paste("chr", c("Y",
        "X", "M", 22:1), sep = ""))))
rm(ERa_coverage, ERb_coverage)

# Plot the fraction
ggplot(combined_coverage |>
    filter(depth > 0)) + geom_bar(mapping = aes(x = chrom, y = pct,
    fill = data), stat = "identity", position = "dodge") + scale_x_discrete(drop = FALSE) +
    coord_flip() + scale_y_continuous(labels = scales::percent) +
    labs(x = "Chromosome", y = "Base Pair Coverage", title = "Coverage of ERa and ERb") +
    theme_bw()
```



There seems to be a relation between the covered chromosomes that could be explained biologically by the fact that these two forms, in several cell types, are co-expressed and in occasions can form heterodimers by binding to two half sites of a response element. Therefore, we would expect a relation between the binding sites for ERalpha and ERbeta. As we saw, there is a lower coverage for ERb in comparison to ERa, thus ERa binding sites are more common than ERb binding sites.

Additionally, it would seem that there are chromosomes for which ERa and ERb do not map to at all. This is likely explained by the way the data was generated (tiling arrays). Tiling arrays differ from traditional microarrays in the way probes are designed. Therefore, the tiling array used in this experiment might only contain probes for a subset of chromosomes.

b) Again, using BEDtools in Linux: How many ERA sites do/do not overlap ERB sites, and vice versa? Show the Linux commands and then a Venn diagram summarizing the results. The Venn diagram can be made in R using one of many venn diagram packages, but you can also make it in any drawing program.

First of wall, we count for the overlapped and non overlapped data of A and B

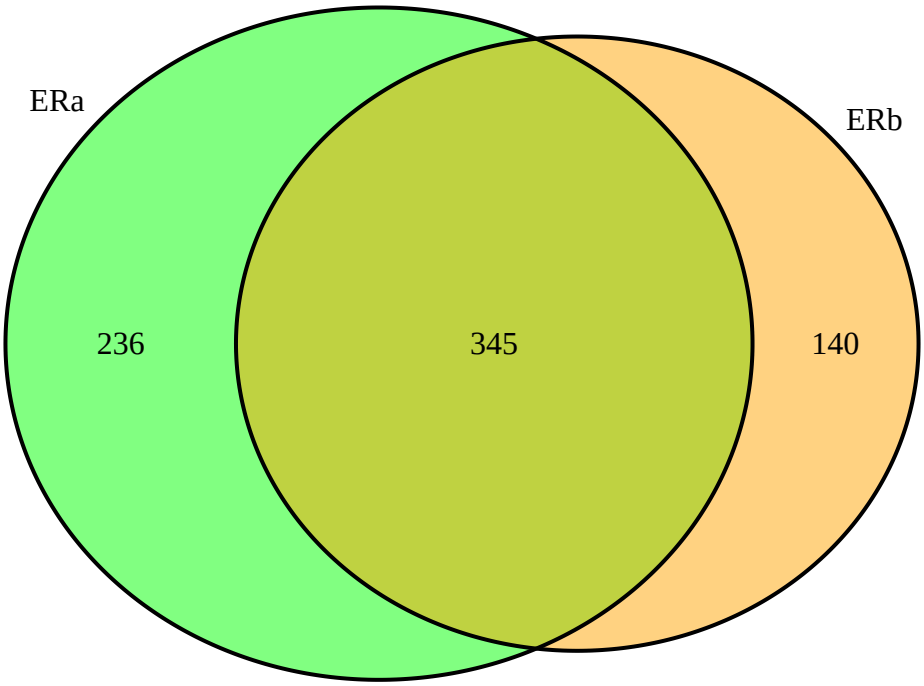
```
# Removing the first line of the files
tail -n +2 ERa_hg18.bed > cleaned_ERa_hg18.bed
tail -n +2 ERb_hg18.bed > cleaned_ERb_hg18.bed
```

```
(bedtools intersect -a cleaned_ERa_hg18.bed -b cleaned_ERb_hg18.bed -wa | wc -l;
  bedtools intersect -a cleaned_ERa_hg18.bed -b cleaned_ERb_hg18.bed -v | wc -l;
  bedtools intersect -a cleaned_ERb_hg18.bed -b cleaned_ERa_hg18.bed -v | wc -l) > ER_venn_data
```

Next, we plot the venn diagram

```
ER_venn_data <- read_tsv("ER_venn_data", col_names = FALSE, show_col_types = FALSE)
ER_venn_data <- bind_cols(ER_venn_data, type = c("intersect",
    "ERa_only", "ERb_only"))
ER_venn_data <- ER_venn_data |>
    rename(count = X1) |>
    pivot_wider(names_from = type, values_from = count)


grid.newpage()
venn.plot <- draw.pairwise.venn(area1 = ER_venn_data$ERa_only +
    ER_venn_data$intersect, area2 = ER_venn_data$ERb_only + ER_venn_data$intersect,
    cross.area = ER_venn_data$intersect, fill = c("green", "orange"),
    category = c("ERa", "ERb"))
```

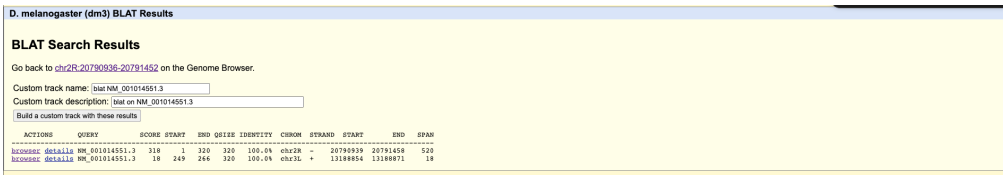


Question 3

Your group just got this email from a frustrated fellow student:

My supervisor has found something he thinks is a new ribosomal protein gene in mouse. It is at chr9:24,851,809-24,851,889, assembly mm8. His arguments for this are a) It has high conservation in other species because ribosomal protein genes from other species map to this mouse region b) And they are all called Rpl41 in the other species (if you turn on the other Refseq you see this clearly in fly and other species).

But, I found out that if you take the fly refseq sequence mentioned above (from Genbank) and BLAT this to the fly genome, you actually get something that looks quite different from the one in the mouse genome. How can this be? Is the mouse gene likely to be real? If not, why? (Maximum 20 lines, plus possibly genome browser pictures)


D. melanogaster (dm3) BLAT Results

BLAT Search Results

Go back to chr2R:20790936-20791452 on the Genome Browser.

Custom track name: blat NM_001014551.3

Custom track description: blat on NM_001014551.3

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	NM_001014551.3	318	1	320	320	100.0%	chr2R	-	20790939	20791458	520
browser details	NM_001014551.3	18	249	266	320	100.0%	chr3L	+	13188854	13188871	18

Figure 4: *BLAT RpL41 to the Fly (md3)*

If we blast the mRNA sequence from the predicted mouse gene against the fly cDNA sequence (Figure 4) and we compare the aligned bases to the Drosophila reference genome, it can be seen that the predicted mouse gene corresponds to the most of exon 2 and last part of exon 1 in the fly Rpl41 gene. Thus, the conserved region in the mouse genome corresponds to a truncated exon structure of the Drosophila.

This could suggest that the predicted gene is a pseudogenic fragment. Pseudogenes are non-functional DNA fragments which structurally resemble actual genes, and they are very common for genes encoding ribosomal proteins. Pseudogenes can arise either through DNA duplication or through reverse transcription of mRNA. Processed pseudogenes are produced by retrotransposition, whereby a processed mRNA is reverse transcribed into cDNA and reintegrated in the genome at a new locus. Because of this mechanism, processed pseudogenes

can be found on different chromosomes from the parental gene". This is in accordance with the mouse Rpl41 gene being located at chr10:127,950,396-127,952,779.

The hypothesis of the predicted gene being a processed pseudogene produced through retrotransposition might be in agreement with the fact that LINE elements are flanking the predicted gene.

References:

Genecards database: https://www.genecards.org/cgi-bin/carddisp.pl?gene=RPL41.

Zhang, Z., Harrison, P., & Gerstein, M. (2002). Identification and Analysis of Over 2000 Ribosomal Protein Pseudogenes in the Human Genome. Genome Research, 12(10), 1466-1482. https://doi.org/10.1101/gr.331902)