Homework 2

Group 5: Mikkel Corfitzen, Katja Johansen, Stefano Pellegrini, Rikke Stausholm

5/13/2020

Homework 2

Names: Mikkel Corfitzen, Katja Johansen, Stefano Pellegrini, Rikke Stausholm
Group: 5

Question 1: Dicer dissected

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?

5' - TTT CC

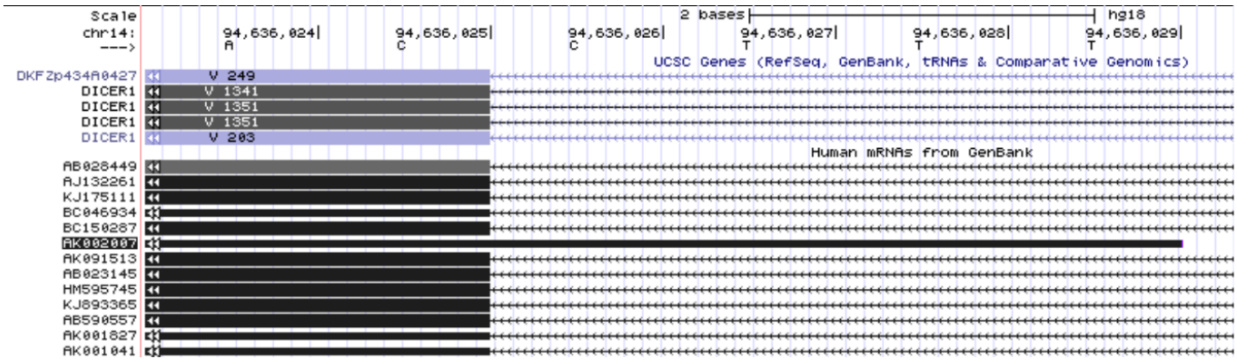


Figure 1: *The 5' UTR of the AK002007 mRNA alignment starts with 5'-TTTCC-3*

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

5' - GAA GC

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

In Figures 2 and 3 we see that the alignment of the mRNA to the genome starts 7 bases further downstream, with AAAGG, which is complementary to TTTCC of the genome. After the first 7 bases we have a 100% alignment of 1753 nucleotides. The mRNA has been going through a cap-targeted selection method called oligo-capping [1]. It is possible that the first 7 bases could be leftovers from the 5' oligo-cap or its adapter. In either case the 7 bases sequence is an artifact from the cDNA library, where the mRNA sequence was

added to the library without being trimmed. During the trimming step the adapters and other contaminant sequences are removed from the reads.

```
gaagcaaAAA GGTCAGCAAC TGTAATCTGT ATCGCCTTGG AAAAAAGAAG   50
GGACTACCCA GCCGCATGGT GGTGTCAATA TTTGATCCCC CTGTGAATTG  100
```

Figure 2: *First 100 nucleotides of the AK002007 cDNA sequence, aligned to the hg18 assembly. Capitalized blue is the coding region, capitalized red is the UTR, light blue marks the intron/exon boundaries in the coding region, orange marks the intron/exon boundaries in the UTR region, and uncapitalized black is a unaligned sequence.*

mRNA/Genomic Alignments

BROWSER	SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
browser	1753	100.0%	14	-	94626558	94636029	AK002007	8	1760	1760

Figure 3: *mRNA aligment of the AK002007 to the hg18 assemly.*

Question 2: ERA and ERB

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: one for estrogen receptor 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?

```
# Sort the files
sort -k1,1V -k2,2n ERa_hg18.bed > ERa_sorted.bed
sort -k1,1V -k2,2n ERb_hg18.bed > ERb_sorted.bed

# Compute the genome coverage
nice bedtools genomecov -i ERa_sorted.bed -g hg18_chrom_sizes.txt -> ERa_coverage.bedreport
nice bedtools genomecov -i ERb_sorted.bed -g hg18_chrom_sizes.txt -> ERb_coverage.bedreport

# Sort the coverage report files using natural sorting
sort -k1,1V ERa_coverage.bedreport > ERa_coverage_sorted.bedreport
sort -k1,1V ERb_coverage.bedreport > ERb_coverage_sorted.bedreport
```

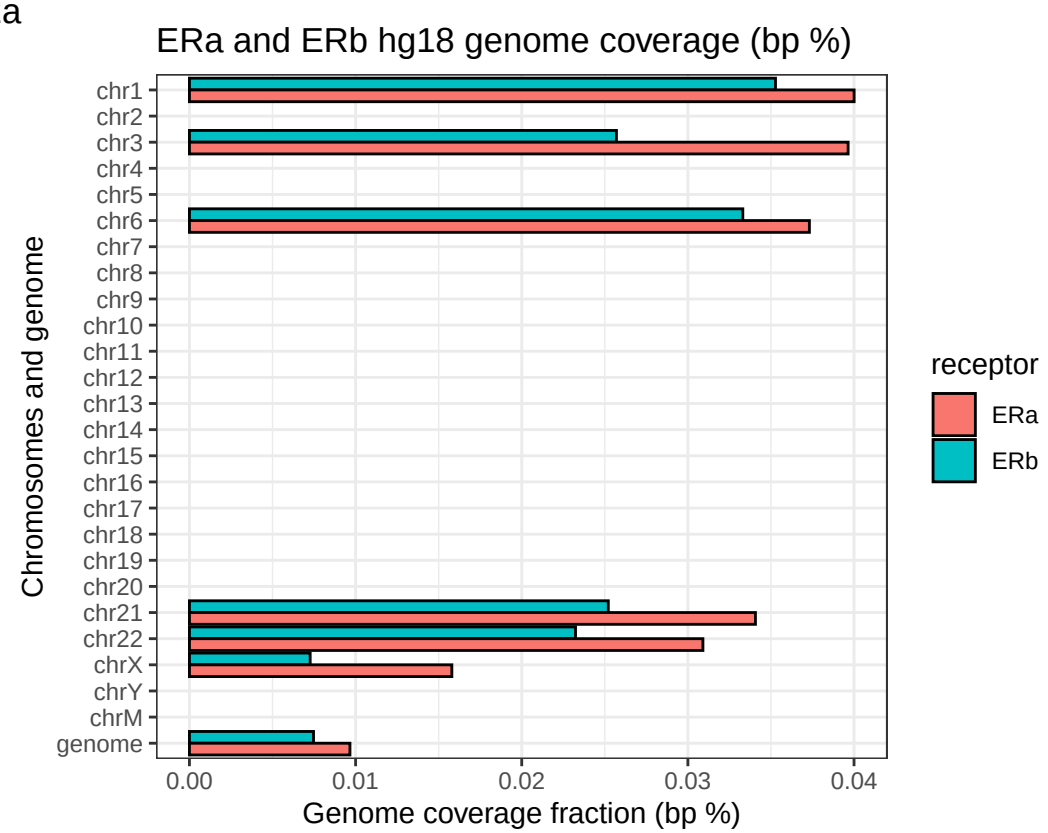
Plot the fractions for all chromosomes as a single barplot in R. Briefly comment the results. Is there anything particularly surprising? Try to explain the outcome (biological and/or experimental setup explanations)?

```
ERa_coverage <- read_tsv("ERa_coverage_sorted.bedreport",
                         col_names = c("chr",
                                       "depth",
                                       "bp",
```

```
                                        "total_chr_bp",
                                        "coverage_fraction"))
ERb_coverage <- read_tsv("ERb_coverage_sorted.bedreport",
                         col_names = c("chr",
                                        "depth",
                                        "bp",
                                        "total_chr_bp",
                                        "coverage_fraction"))
```

```
# Plot the fractions (bp %) for all chromosomes as a single barplot
ERb_coverage %>%
  transmute(chr, depth,                                              # Trasmute coverage to bp %
            'ERb' = coverage_fraction * 100,
            'ERa' = ERa_coverage$coverage_fraction * 100) %>%
  gather(key = "receptor", val = "coverage_fraction", ERb, ERa) %>% # Use long format
  mutate(chr = factor(chr,                                           # Convert chr to factors so
                      levels = c("genome",                           # that we can keep unused levels
                                 paste("chr", c("M","Y", "X", 22:1),# (specify the levels so that
                                 sep = "")))) %>%                    #  chr are sorted in the plot)
  filter(depth == 1) %>%                                             # Plot histograms of covered
  ggplot(aes(x = chr, y = coverage_fraction, fill = receptor)) +     # chromosomes
  geom_bar(stat = "identity", position = "dodge", colour="black") +
  scale_x_discrete(drop = FALSE) +                                   # Keep unused levels
  coord_flip() +                                                     # Flip to horizontal
  labs(title = "ERa and ERb hg18 genome coverage (bp %)",
       tag = "Question 2a") +
  ylab("Genome coverage fraction (bp %)") + xlab("Chromosomes and genome") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.tag = element_text(size = 10)) +
  theme_bw()
```

Question 2a




```
# Just to better specify the coverage fraction of the covered chromosomes,
# (omitted chromosomes have no coverage for ERb or ERa sites)
ERb_coverage %>%
  transmute(chr, depth,
            'ERb' = coverage_fraction * 100,
            'ERa' = ERa_coverage$coverage_fraction * 100) %>%
  filter(depth==1) %>%
  transmute(chr, ERb, ERa) %>%
  rename(ERb_bp_percentage = ERb,
         ERa_bp_percentage = ERa) -> ERab_covered_chr_percentage
ERab_covered_chr_percentage
```

```
## # A tibble: 7 x 3
##   chr    ERb_bp_percentage ERa_bp_percentage
##   <chr>              <dbl>             <dbl>
## 1 chr1             0.0353            0.0400
## 2 chr3             0.0257            0.0396
## 3 chr6             0.0333            0.0373
## 4 chr21            0.0252            0.0341
## 5 chr22            0.0232            0.0309
## 6 chrX             0.00727           0.0158
## 7 genome           0.00747           0.00966
```

From our results the smallest binding area for ERb is on chrX where it covers 0.007% of the chromosome compared to the largest binding area on chr1 that is covered 0.035%. The smallest binding area for ERa is also on chrX where it covers 0.016% of the chromosome. And the largest binding area is also on chr 1 that is covered 0.040%. ERa and ERb are both activated when binding the ligand Estrogen. Estrogen is a steroid

hormone with fundamental functions in growth and cardiovascular system but most importantly serves as the primary female sex hormone responsible for the development and regulation of the female reproductive system. Because of this very important role in general development and regulation, we would have expected a high presence of ERa and ERb binding sites random across the 23 chromosomes with exception from chromosome Y. We can see that ERa and ERb binding sites are only present on 6 chromosomes (chr1 ,3 ,6 ,21, 22, x). This observation can be explained by the technique used. The collaborators used a tiling array which is almost the same as a microarray with the probe design as the main difference. Tiling arrays probe for known contiguous sequences meaning that the collaborators only probe for a known sequence and not for the whole genome. We would expect to have contiguous sequences on multiple chromosomes but for some reason we do not get that. The tiling array is probably expensive and time consuming and for this reason our collaborators do not probe for all the chromosomes. Also, there are no ERa or ERb binding sites on chromosome Y, which makes sense with the fact that the receptors primarily are involved in the development and regulation of the female reproductive system. We also observe an extreme difference between the sites of ERb and ERa on chromosome X, in fact, ERb sites are approximately half of the ERa sites. This may indicate that the two receptors have different roles even though they are activated by the same ligand.

b) Again, using BEDtools in Linux: How many ERA sites do/do not overlap ERB sites, and vice versa?

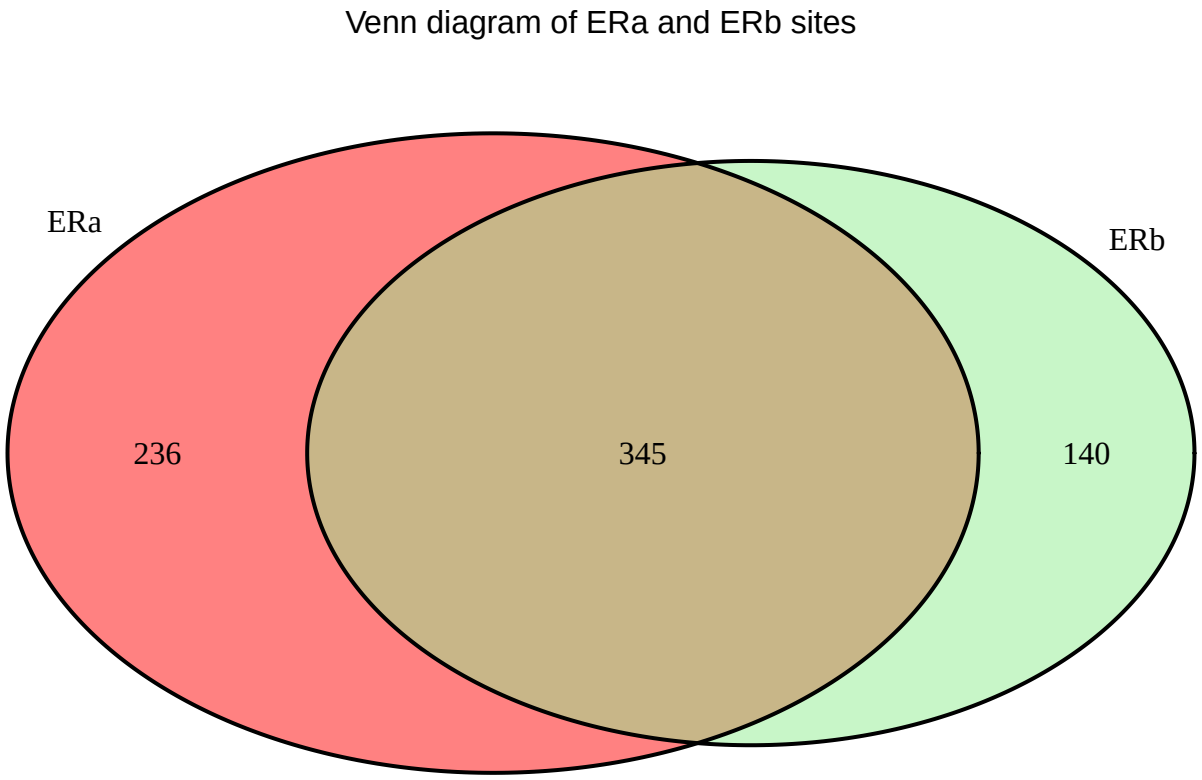
345 ERA sites do overlap ERB sites and 236 don't overlap them.
345 ERB sites do overlap ERA sites and 140 don't overlap them.

Show the Linux commands and then a Venn diagram summarizing the results.

```
## Linux commands

# Output ERA sites overlapping ERB sites and vice versa, count rows
bedtools intersect -a ERa_hg18.bed -b ERb_hg18.bed  | wc -l
bedtools intersect -a ERb_hg18.bed -b ERa_hg18.bed  | wc -l

# Output ERA sites not overlapping ERB sites and vice versa, count rows
bedtools intersect -a ERa_hg18.bed -b ERb_hg18.bed -v | wc -l
bedtools intersect -a ERb_hg18.bed -b ERa_hg18.bed -v | wc -l
```

```
## Venn diagram

library(VennDiagram)
VennDiagram = draw.pairwise.venn(581, 485, 345,
                                 category = c("ERa", "ERb"),
                                 alpha = rep(0.5, 2),
                                 fill = c("red", "lightgreen"))
```

```
# Add title
require(gridExtra)
grid.arrange(gTree(children = VennDiagram),
             top = "Venn diagram of ERa and ERb sites")
```



Question 3: Ribosomal Gene

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) 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)

As we can see in Figure 4, the ribosomal protein gene that the supervisor claims to have found, is a very conserved sequence across different species, and in these other species is (part of) a gene called Rpl41.

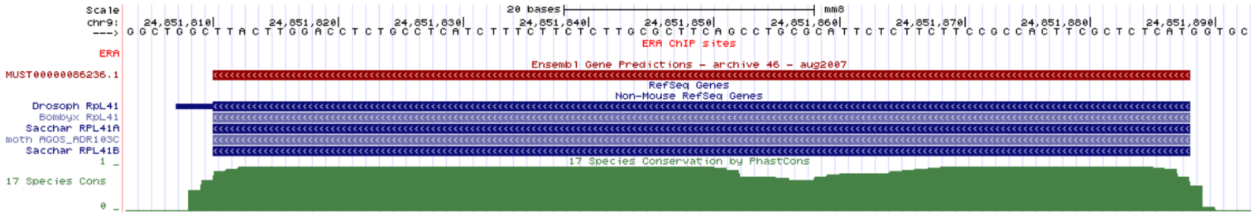


Figure 4: *Ensembl mouse gene prediction and known protein-coding and non-protein-coding genes for other organisms.*

Figure-5 confirms the student's statement, in fact the Rpl41 gene from D. Melanogaster is very different from the ribosomal protein gene that the supervisor claims to have found in the mouse genome.

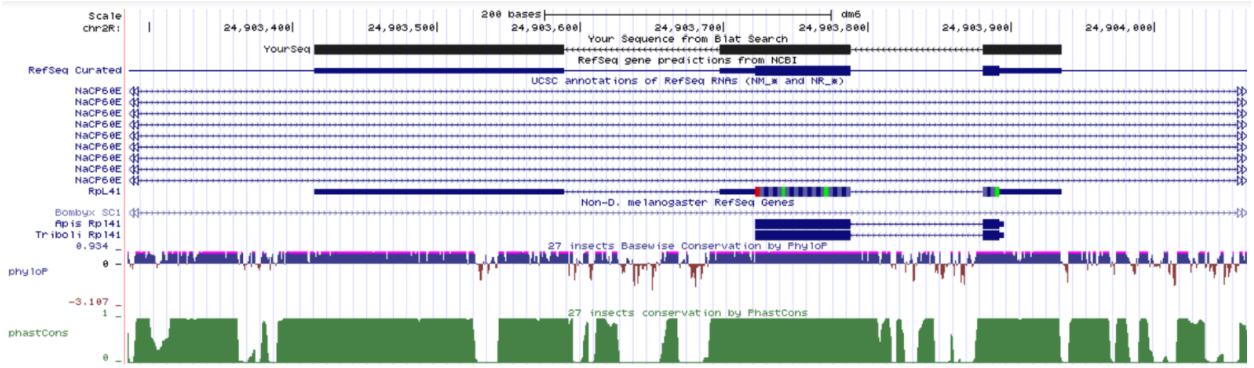


Figure 5: *Rpl41 gene from D. Melanogaster, showing the conservation track and known protein-coding and non-protein-coding genes for organisms.*

The fly refseq sequence from the mouse genome is 81 bp (Figure 6) long and contains only a single exon (Figure 4). When we BLAT the sequence to the fly genome we get a 320 bp long sequence containing three exons and two introns. Our sequence from the BLAT search corresponds very well to the mRNA of the RpL41 gene in the fly genome. Since it is in the refseq database, we are not in doubt that the RpL41 is a real active gene in the fly genome, and since it is very conserved we believe that it has an important role in ribosomal activity.

mRNA/Genomic Alignments

The alignment you clicked on is first in the table below.

BROWSER	SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
browser	81	81.5%	9	+-	24851808	24851888	NM 001014551	44	124	320

Figure 6: *mRNA aligment of the predicted mouse gene to the D. Melanogaster genome, 81 bases aligned with 81.5 % identity.*

During evolution a lot of mutations occur and some of them will by chance be inhibitory for the given gene. The mouse gene is most likely a pseudogene that shares some DNA sequence with the real active fly gene but has become inactive due to inhibitory mutations. The possible biological process explaining the generation of pseudogene is called "gene duplication". A pseudogene often lacks introns and other functional DNA sequences that are critical for the gene activity which is in agreement with our observations between the mouse and fly genome. Therefore we do not believe the mouse gene is a real gene, that does not have any function besides telling us about the evolution of ribosomal protein genes.

Reference

[1] Ota, T., Suzuki, Y., Nishikawa, T. et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36, 40–45 (2004).