Homework 2, group 5

Ann C. Printz
Jesus V. Torresano Lominchar
Peter C. Holm
Sesilie B. Weiss
William P. Bullock

Part I: 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 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)?

The end of the cDNA often contain errors which could be due to artifacts or truncated cDNA. When mapping this to the annotated genes, the algorithm that maps the raw sequence to the genome has trouble with these discrepancies. Looking at the exon prior to the transcription start site of AK002007, some the discrepancy sequence actually corresponds to the last part of that exon. This indicates that the cDNA might be truncated.

Part II: ERA and ERB

Our collaborators designed a ChIP study using so-called tilling arrays: 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.

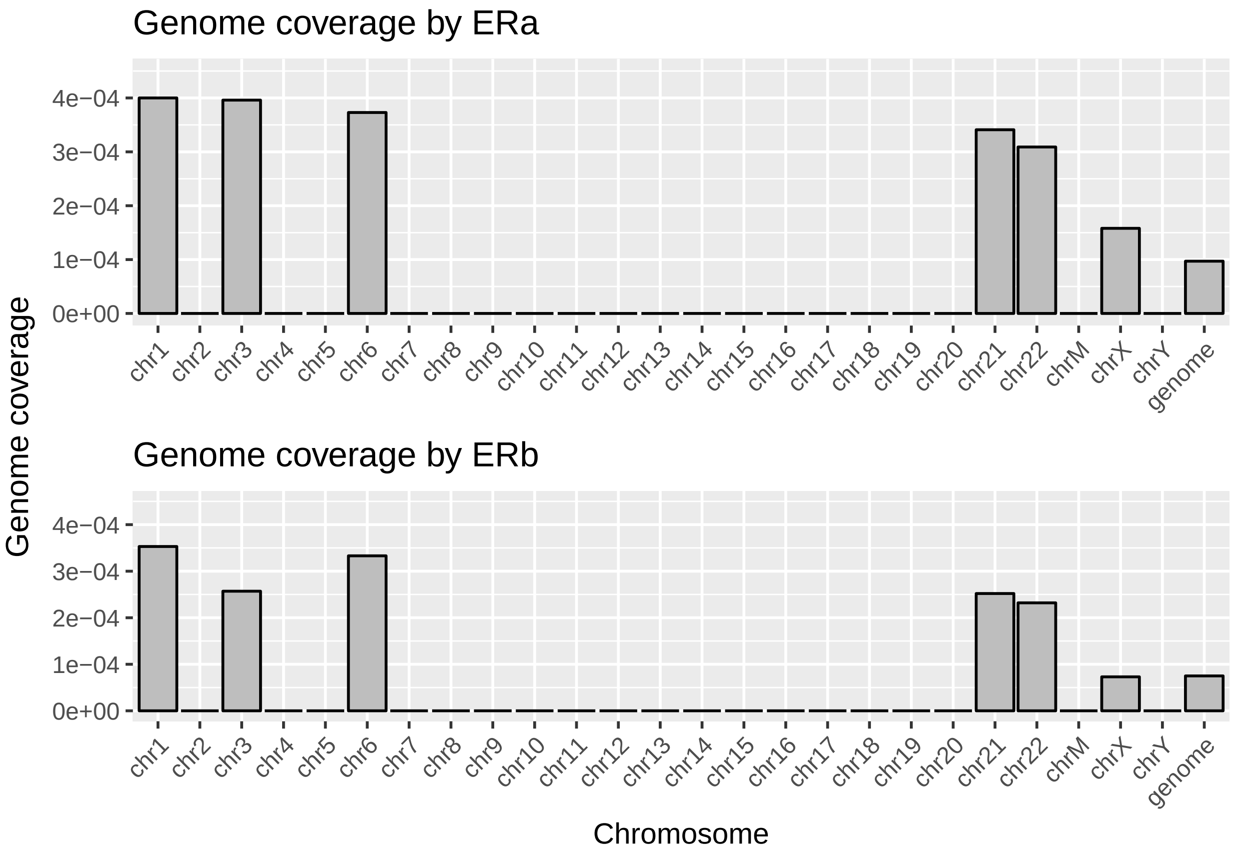
a) What is the genome coverage (% of base pair covered at each chromosome) for ERB and ERA sites? Plot the fractions for all chromosomes as a single barplot in R. Briefly comment the results. Is there anything particularly surprising? Try to explain the outcome (biological and/or experimental setup explanations)?

```
# Connect to ricco-server
ssh xxx@ricco.popgen.dk

# First we want to make sure that ERA and ERB bedfiles are sorted
sort -k1,1V -k2,2n ERa_hg18.bed > ERa_hg18_sorted.bed
sort -k1,1V -k2,2n ERb_hg18.bed > ERb_hg18_sorted.bed

# And also that the hg18_chrom_sizes.txt are correctly sorted
sort -k1,1V hg18_chrom_sizes.txt > hg18_chrom_sizes_sorted.txt

# Getting the genome-coverages for each
nice bedtools genomecov -i ERa_hg18_sorted.bed -g hg18_chrom_sizes_sorted.txt -max 1 >
genomecov_ERa.txt
nice bedtools genomecov -i ERb_hg18_sorted.bed -g hg18_chrom_sizes_sorted.txt -max 1 >
genomecov_ERb.txt
```

Then we read the genomecov outputs into R

```
# Load packages
library(tidyverse)
library(magrittr)
library(gridExtra)

# Import genome coverage data
genomecov_ERa <- read_tsv(file = "genomecov_ERa.txt", col_names = F)
genomecov_ERb <- read_tsv(file = "genomecov_ERb.txt", col_names = F)

# Assign factor levels for chromosomes in the correct order
chr_order <- genomecov_ERa$X1 %>% unique() %>% gtools::mixedsort()
genomecov_ERa$X1 <- genomecov_ERa$X1 %>% factor(levels = chr_order)
genomecov_ERb$X1 <- genomecov_ERb$X1 %>% factor(levels = chr_order)

# Create plot 1
p1 <- genomecov_ERa %>%
  subset(X2 == 0) %>%
  ggplot(aes(x = X1, y = 1-X5)) +
  geom_bar(stat = "identity", color = "black", fill = "grey") +
  ylim(0, 4.5e-4) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = NULL, y = NULL,
       title = "Genome coverage by ERa")
```

```
# Create plot 2
p2 <- genomecov_ERb %>%
  subset(X2 == 0) %>%
  ggplot(aes(x = X1, y = 1-X5)) +
  geom_bar(stat = "identity", color = "black", fill = "grey") +
  ylim(0, 4.5e-4) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Chromosome", y = NULL,
       title = "Genome coverage by ERb")

# Puts the two plots on top of another
grid.arrange(p1, p2, left = "Genome coverage")
```


Genome coverage by ERa

Genome coverage by ERb

From the plot it can be seen that ERa and ERb-binding are positively correlated, albeit ERa-binding is more enriched on the chromosomes included. If one can determine that the antibodies used to pulldown are specific to either isoform, this could suggest that ERa and ERb have similar biological binding profiles.

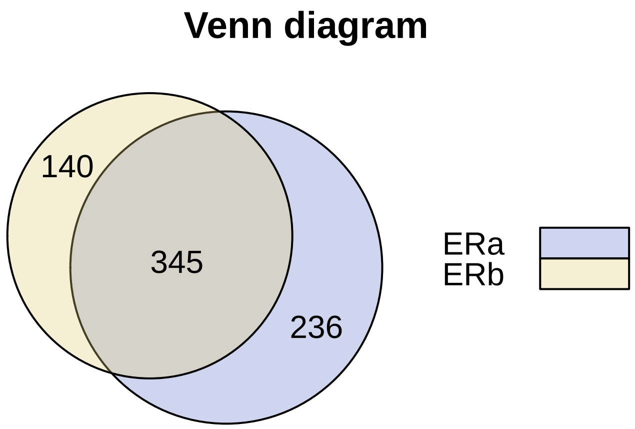
It's worth noting that there is absolutely no signal from several chromosomes, and that the same chromosomes missing in both experiments. Assuming this is raw-data without any threshold to compensate for noise, this suggests that something went wrong experimentally, or that there isn't any probes directed at these 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 the Venn package, but you can also make it in any drawing program.

```
bedtools intersect -a ERa_hg18_sorted.bed -b ERb_hg18_sorted.bed -wa | wc -l
345

bedtools intersect -a ERa_hg18_sorted.bed -b ERb_hg18_sorted.bed -v | wc -l
236

bedtools intersect -a ERb_hg18_sorted.bed -b ERa_hg18_sorted.bed  -wa | wc -l
345

bedtools intersect -a ERb_hg18_sorted.bed -b ERa_hg18_sorted.bed  -v | wc -l
140
```

And using R we can create a Venn-diagram.

```
# Load the eulerr package for making venn-diagrams
library(eulerr)

fit <- euler(c(ERa = 236, ERb = 140, "ERa&ERb" = 345))

# Plot the diagram
plot(fit, fill_opacity = 0.3, counts = T,
     key = list(space = "right"),
     main = "Venn diagram"
     )
```



Part III: 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) 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 mRNA 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)

Answer:

The argument of the supervisor is sound, but there are a lot of things that he has not considered;

1) Zooming a bit out from the mentioned area, the conserved area can be readily seen on the conservation track as seen in figure 1. From this figure, we can also see that it is only a minor part of the conserved region where ribosomal genes from other species map. The gene region thought to be a ribosomal protein is very short – only 81 bp. If we compare this to the drosophila melanogaster rpl41, this has an mRNA length of 320 bp (4x longer than the aligned area). This can be seen in figure 2.

2) The fly rpl41 gene aligns to 4 other regions in the mouse genome besides the one in question, all having a length of 81 bp, covering the same part of the fly rpl41 gene with an 81,5% identity, also seen in figure 2.

3) Taking the seemingly conserved region from chr9:24,851,520-24,851,889 and extracting the DNA-sequence (lower strand) from the "get dna"-option in the genome browser, one can then blat this region to the mm8 genome. The BLAT-results show 52 different places in the mouse genome with similar sequences, some of them are only aligning with 20 nucleotides from the 370bp long query, but several are aligning with a high BLAT-score. This is highly indicative that this could be some sort of transposable element that resides multiple places in the genome.

4) Looking at repetitive regions in the area, one can find; a repetitive region downstream from the suspected gene (this is marked by a red-square in figure 1). Zooming in on this area (figure 3) reveals that the repetitive element is a series of successive a-nucleotides, remnants of a poly-A tail, which indicates that this is indeed a retrotransposed pseudogene and therefore not a novel ribosomal protein gene.

5) Finally, looking at the different reading frames for the reverse strand (seen in figure 1), reveals that all three reading frames have several stop codons - all leading to very small transcripts.

Therefore we conclude that the suspected gene is not a novel ribosomal protein gene, but instead a retrotransposed pseudogene.

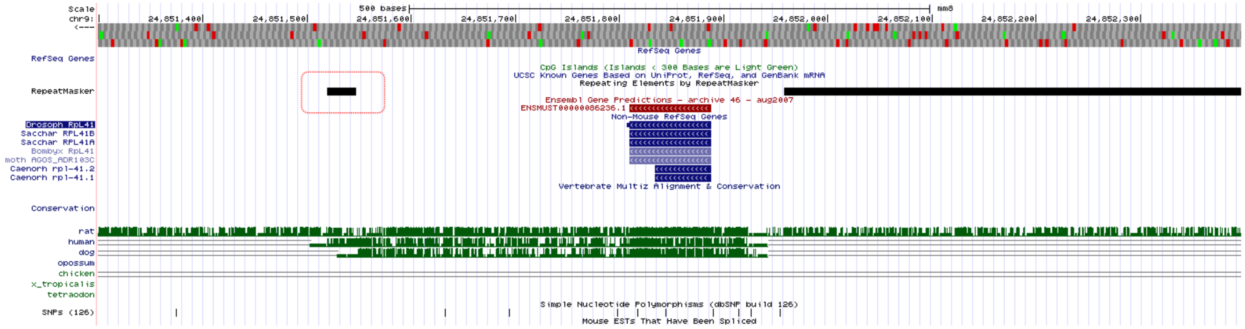
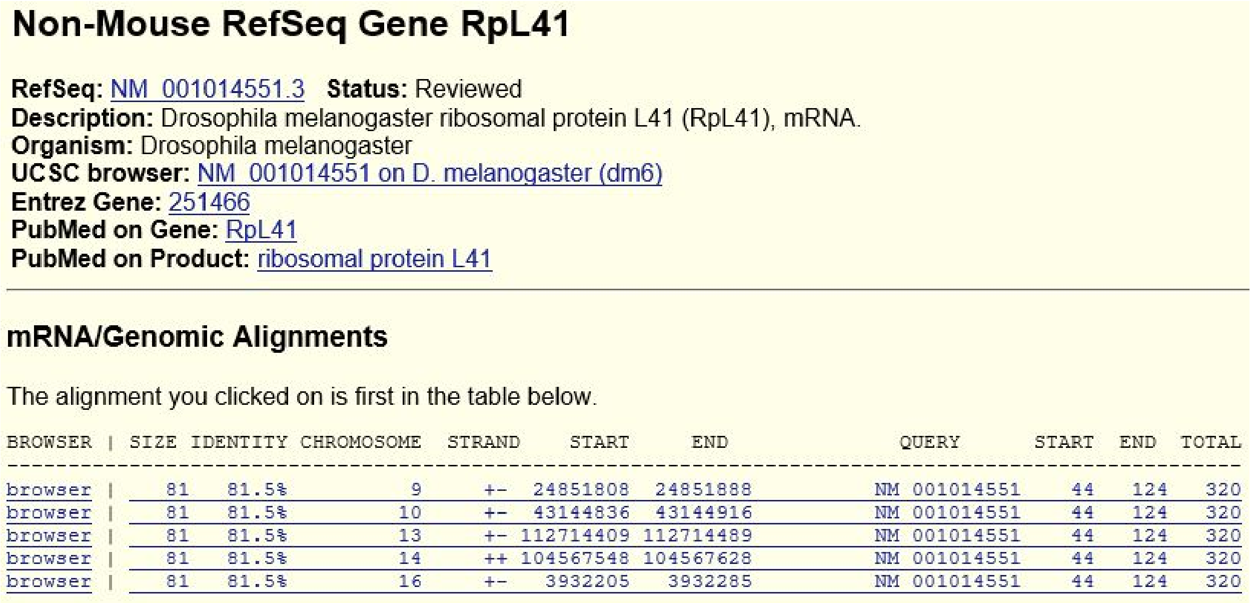


Figure 1: Suspected novel ribosomal protein gene



Non-Mouse RefSeq Gene RpL41

RefSeq: NM_001014551.3 **Status:** Reviewed
Description: Drosophila melanogaster ribosomal protein L41 (RpL41), mRNA.
Organism: Drosophila melanogaster
UCSC browser: NM_001014551 on D. melanogaster (dm6)
Entrez Gene: 251466
PubMed on Gene: RpL41
PubMed on Product: ribosomal protein L41

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
browser |    81   81.5%         10     +-   43144836  43144916        NM_001014551     44   124    320
browser |    81   81.5%         13     +-  112714409 112714489        NM_001014551     44   124    320
browser |    81   81.5%         14     ++  104567548 104567628        NM_001014551     44   124    320
browser |    81   81.5%         16     +-    3932205   3932285        NM_001014551     44   124    320
```

Figure 2: Genomic Alignments

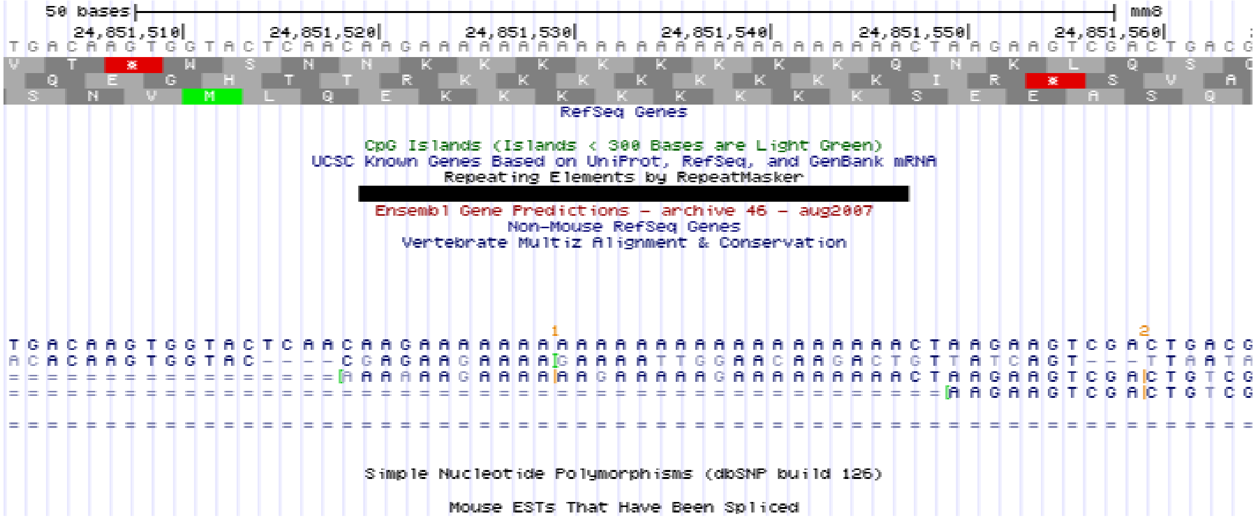

50 bases
mm8
24,851,510
24,851,520
24,851,530
24,851,540
24,851,550
24,851,560
RefSeq Genes
CpG Islands (Islands < 300 Bases are Light Green)
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
Repeating Elements by RepeatMasker
Ensembl Gene Predictions - archive 46 - aug2007
Non-Mouse RefSeq Genes
Vertebrate Multiz Alignment & Conservation
1
2
T G A C A A G T G G T A C T C A A C A A G A A A A A A A A A A A A A A A A A A A C T A A G A A G T C G A C T G A C G
Simple Nucleotide Polymorphisms (dbSNP build 126)
Mouse ESTs That Have Been Spliced

Figure 3: Repetitive element