# Hw2

## Home work 2 - to be done as groups

Names: Group:

For deadlines etc, see Absalon.

### 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?

**Answer:** The first 5 genomic nucleotides seen from the UCSC-genome browser is: *AAAGG*

This is seen on the following screenshot (Fig 1.) with the first exon starting on the right and running to the left. The sequence CCTTT is reserve complemented and gives AAAGG.

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

The first 5 nucleotides from the GenBank sequence is gaagcaa. This is seen in the screenshot on figure 2:

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

The discrepancy is hard to explain, but we have a couple of theories. Looking at the GenBank entry we can see that the sequences are found by oligo-capping. In this method a cDNA library is constructed by removal of the 5'-Cap and insertion of a small synthetic oligo. This sequence could also show up in the sequencing and shown in the genbank, but removed when aligned to the genome

*Source: 1. Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A. & Sugano, S. Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. Gene 200, 149–156 (1997).*

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Our collaborators designed a ChIP study using so-called tilling arrays (an outdated technique these days, but the top of the pop at the time: see http://en.wikipedia.org/wiki/Tiling_array): one for estrogen receptor alpha (ERA), one for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

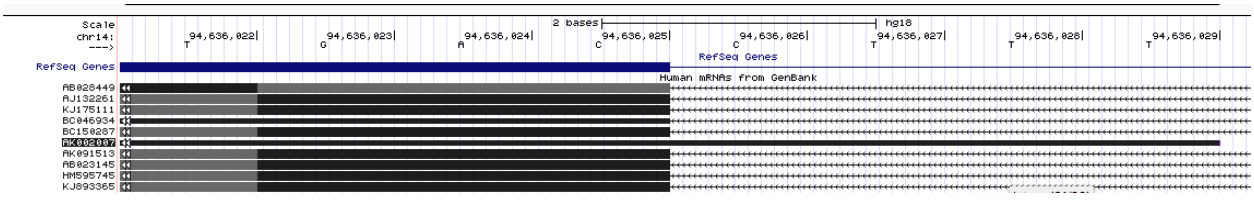


Figure 1: UCSC screenshot, showing the five first bases of transcription of AK002007

```
                     Endoribonuclease Dicer (EC 3.1.26.-)
                     /codon_start=1
                     /protein_id="BAG51002.1"
                     /translation="MVVSIFDPPVNWLPPGYVVNQDKSNTDKWEKDEMTKDCMLANGK
                     LDEDYEEEDEEEESLMWRAPKEEADYEDDFLEYDQEHIRFIDNMLMGSGAFVKKISLS
                     PFSTTDSAYEWKMPKKSSLGSMPFSSDFEDFDYSSWDAMCYLDPSKAVEEDDFVVGFW
                     NPSEENCGVDTGKQSISYDLHTEQCIADKSIADCVEALLGCYLTSCGERAAQLFLCSL
                     GLKVLPVIKRTDREKALCPTRENFNSQQKNLSVSCAAASVASSRSSVLKDSEYGCLKI
                     PPRCMFDHPDADKTLNHLISGFENFEKKINYRFKNKAYLLQAFTHASYHYNTITDCYQ
                     RLEFLGDAILDYLITKHLYEDPRQHSPGVLTDLRSALVNNTIFASLAVKYDYHKYFKA
                     VSPELFHVIDDFVQFQLEKNEMQGMDSELRRSEEDEEKEEDIEVPKAMGDIFESLAGA
                     IYMDSGMSLETVWQVYYPMMRPLIEKFSANVPRSPVRELLEMEPETAKFSPAERTYDG
                     KVRVTVEVVGKGKFKGVGRSYRIAKSAAARRALRSLKANQPQVPNS"
ORIGIN
        1 gaagcaaaaa ggtcagcaac tgtaatctgt atcgccttgg aaaaaagaag ggactaccca
       61 gccgcatggt ggtgtcaata tttgatcccc ctgtgaattg gcttcctcct ggttatgtag
      121 taaatcaaga caaaagcaac acagataaat gggaaaaaga tgaaatgaca aaagactgca
      181 tgctggcgaa tggcaaactg gatgaggatt acgaggagga ggatgaggag gaggagagcc
      241 tgatgtggag ggctccgaag gaagaggctg actatgaaga tgatttcctg gagtatgatc
      301 aggaacacat cagatttata gataatatgt taatggggtc aggagctttt gtaaagaaaa
      361 tctctctttc tccttttca accactgatt ctgcatatga atggaaaatg cccaaaaaat
      421 cctccttagg tagtatgcca ttttcatcag attttgagga ttttgactac agctcttggg
      481 atgcaatgtg ctatctggat cctagcaaag ctgttgaaga agatgacttt gtggtggggt
      541 tctggaatcc atcagaagaa aactgtggtg ttgacacggg aaagcagtcc atttcttacg
      601 acttgcacac tgagcagtgt attgctgaca aaagcatagc ggactgtgtg gaagccctgc
      661 tgggctgcta tttaaccagc tgtggggaga gggctgctca gcttttcctc tgttcactgg
      721 ggctgaaggt gctcccggta attaaaagga ctgatcggga aaaggccctg tgccctactc
      781 gggagaattt caacagccaa caaaagaacc tttcagtgag ctgtgctgct gcttctgtgg
      841 ccagttcacg ctcttctgta ttgaaagact cggaatatgg ttgtttgaag attccaccaa
      901 gatgtatgtt tgatcatcca gatgcagata aaacactgaa tcaccttata tcggggtttg
      961 aaaattttga aaagaaaatc aactacagat tcaagaataa ggcttacctt ctccaggctt
     1021 ttacacatgc ctcctaccac tacaatacta tcactgattg ttaccagcgc ttagaattcc
     1081 tgggagatgc gattttggac tacctcataa ccaagcacct ttatgaagac ccgcggcagc
     1141 actccccggg ggtcctgaca gacctgcggt ctgccctggt caacaacacc atctttgcat
     1201 cgctggctgt aaagtacgac taccacaagt acttcaaagc tgtctctcct gagctcttcc
     1261 atgtcattga tgactttgtg cagtttcagc ttgagaagaa tgaaatgcaa ggaatggatt
     1321 ctgagcttag gagatctgag gaggatgaag agaaagaaga ggatattgaa gttccaaagg
     1381 ccatgggggа tatttttgag tcgcttgctg gtgccattta catggatagt gggatgtcac
     1441 tggagacagt ctggcaggtg tactatccca tgatgcggcc actaatagaa aagttttctg
     1501 caaatgtacc ccgttcccct gtgcgagaat tgcttgaaat ggaaccagaa actgccaaat
     1561 ttagcccggc tgagagaact tacgacggga aggtcagagt cactgtggaa gtagtaggaa
     1621 aggggaaatt taaaggtgtt ggtcgaagtt acaggattgc caaatctgca gcagcaagaa
     1681 gagccctccg aagcctcaaa gctaatcaac ctcaggttcc caatagctga aaccgctttt
     1741 taaaattcaa aacaagaaac
//
```

Figure 2: Screenshot from the GenBank databse of AK002007. The 7 nucleotides which are different between the UCSC browser and the GenBank entry are highlighted in blue.

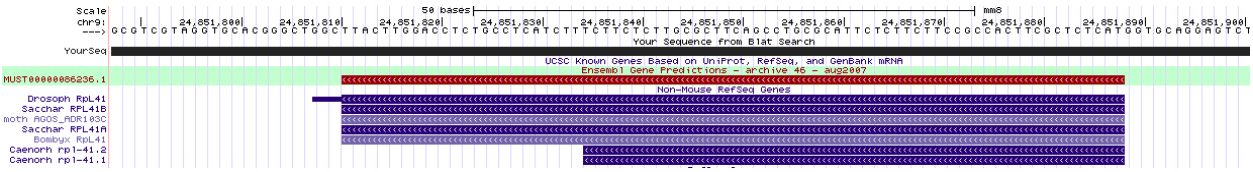


Figure 3: USCS screenshot from the mouse genomic region.

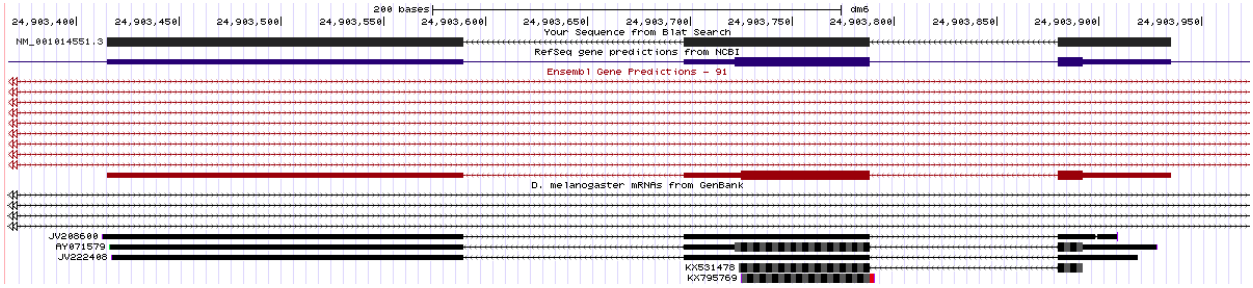


Figure 4: USCS screenshot from the fly BLAT result.

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)?

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.

### 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) 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)

**Answer:**

This is the genomic region in mouse in which we can see that there is a conserved gene in fly and some other species, all of them ribosomal proteins (RpL41).

We can see that the mRNA from the gene we are looking for does not seem to be spliced, but when we take the sequence from the fly and BLAT it against its own genome, we get that the mRNA from gene we find (RpL41) is spliced here, while was not in the possible gene from mouse.

In order to find out a reason for that, we took the mouse genomic sequence and BLAT against the mouse genome (mm8). These are the hits we got:

| QUERY | SCORE | START | END | QSIZE | IDENTITY | CHRO | STRAND | START | END | SPAN |
|---|---|---|---|---|---|---|---|---|---|---|
| YourSeq | 78 | 1 | 78 | 78 | 100.0% | 9 | - | 24851811 | 24851888 | 78 |
| YourSeq | 78 | 1 | 78 | 78 | 100.0% | 13 | - | 112714412 | 112714489 | 78 |
| YourSeq | 78 | 1 | 78 | 78 | 100.0% | 10 | - | 43144839 | 43144916 | 78 |
| YourSeq | 78 | 1 | 78 | 78 | 100.0% | 14 | + | 104567548 | 104567625 | 78 |
| YourSeq | 76 | 1 | 78 | 78 | 98.8% | 16 | - | 3932208 | 3932285 | 78 |
| YourSeq | 76 | 1 | 78 | 78 | 100.0% | 10 | - | 127951336 | 127951909 | 574 |
| YourSeq | 74 | 1 | 78 | 78 | 97.5% | 17 | - | 12680975 | 12681052 | 78 |
| YourSeq | 74 | 1 | 78 | 78 | 97.5% | 1 | - | 51407807 | 51407884 | 78 |
| YourSeq | 74 | 1 | 78 | 78 | 97.5% | 11 | + | 12548094 | 12548171 | 78 |
| YourSeq | 72 | 1 | 78 | 78 | 96.2% | 2 | - | 113896864 | 113896941 | 78 |
| YourSeq | 72 | 1 | 78 | 78 | 92.0% | 15 | - | 28003397 | 28003471 | 75 |
| YourSeq | 70 | 1 | 78 | 78 | 94.9% | 18 | - | 10274594 | 10274671 | 78 |
| YourSeq | 68 | 1 | 78 | 78 | 94.9% | 13 | - | 55192053 | 55192145 | 93 |
| YourSeq | 68 | 1 | 78 | 78 | 93.6% | 2 | + | 150618774 | 150618851 | 78 |
| YourSeq | 66 | 10 | 78 | 78 | 98.6% | 17 | - | 6779921 | 6779993 | 73 |
| YourSeq | 66 | 1 | 78 | 78 | 92.4% | 16 | - | 38488929 | 38489006 | 78 |
| YourSeq | 66 | 11 | 78 | 78 | 98.6% | 12 | - | 81699245 | 81699312 | 68 |
| YourSeq | 64 | 1 | 78 | 78 | 91.1% | 1 | - | 147698430 | 147698507 | 78 |
| YourSeq | 64 | 10 | 78 | 78 | 97.2% | 17 | + | 7797103 | 7797175 | 73 |
| YourSeq | 64 | 1 | 78 | 78 | 91.1% | 16 | + | 96214019 | 96214096 | 78 |
| YourSeq | 61 | 1 | 78 | 78 | 88.8% | 4 | + | 131819713 | 131819788 | 76 |
| YourSeq | 59 | 1 | 78 | 78 | 82.5% | 11 | - | 97117712 | 97117785 | 74 |
| YourSeq | 58 | 1 | 78 | 78 | 82.9% | 11 | - | 20174315 | 20174390 | 76 |
| YourSeq | 55 | 1 | 63 | 78 | 93.7% | 4 | + | 134649234 | 134649296 | 63 |
| YourSeq | 49 | 18 | 76 | 78 | 91.6% | 11 | - | 6176859 | 6176917 | 59 |
| YourSeq | 45 | 30 | 78 | 78 | 96.0% | 6 | + | 70898452 | 70898500 | 49 |
| YourSeq | 42 | 28 | 75 | 78 | 93.8% | 8 | - | 10717096 | 10717143 | 48 |
| YourSeq | 42 | 19 | 78 | 78 | 85.0% | 6 | - | 107061992 | 107062051 | 60 |
| YourSeq | 42 | 26 | 75 | 78 | 87.8% | 2 | + | 42065104 | 42065152 | 49 |
| YourSeq | 38 | 26 | 67 | 78 | 95.3% | 7 | + | 57411777 | 57411818 | 42 |
| YourSeq | 35 | 1 | 37 | 78 | 97.3% | 18 | - | 79300084 | 79300120 | 37 |
| YourSeq | 35 | 39 | 77 | 78 | 94.9% | 11 | + | 43750972 | 43751010 | 39 |
| YourSeq | 28 | 45 | 78 | 78 | 84.9% | 4 | - | 116541235 | 116541267 | 33 |
| YourSeq | 27 | 52 | 78 | 78 | 100.0% | 6 | - | 28096491 | 28096517 | 27 |
| YourSeq | 25 | 1 | 25 | 78 | 100.0% | 7 | + | 34131784 | 34131808 | 25 |
| YourSeq | 24 | 44 | 67 | 78 | 100.0% | 7 | - | 75772459 | 75772482 | 24 |
| YourSeq | 21 | 47 | 67 | 78 | 100.0% | 12 | + | 32829422 | 32829442 | 21 |
| YourSeq | 20 | 59 | 78 | 78 | 100.0% | 6 | - | 100360374 | 100360393 | 20 |

Figure 5: List of hits from BLAT.

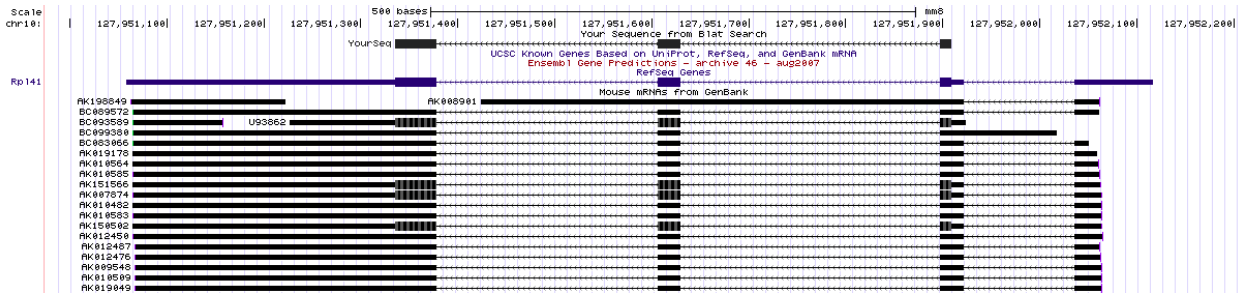


Figure 6: Genome browser showing the sixth hit.

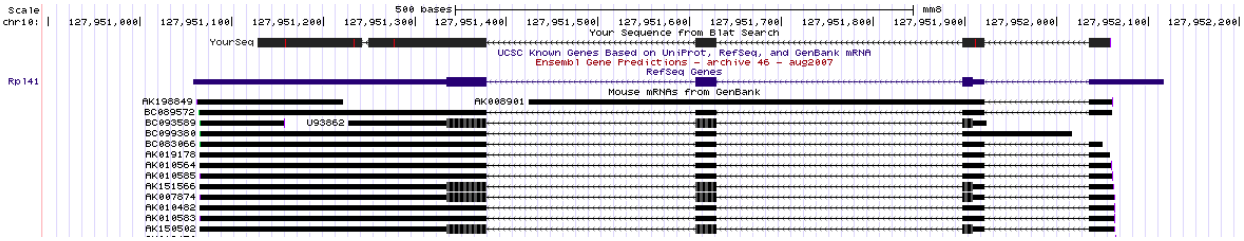


Figure 7: USCS screenshot from the mouse chromosomic region.

We observe that the same phenomenon, is happening in multiple places in different chromosomes. Now focusing on the sixth hit, we see that the span (574) is significantly higher than in the others (78). If we go to the genome browser for that region, we can see that there is an actual RefSeq gene there, a ribosomal protein.

We also took a extended sequence, 200 bp from both sides, and performed the BLAT alignment again, getting a longer alignment for the same ribosomal gene.

The explanation we have for this is that the original sequence in mouse is likely to be a processed retrotransposon. This could have happended due to a retro-transcription of the spliced mRNA of the actual ribosomic protein gene we found in mouse followed by an insertion in the genome. The fact that we found many identical sequences in the genome also supports this explanation.

From searching in literature we see that this is a very common phenomenon for ribosomic protein genes [1]. It has been long thought that this was junk DNA, but now it is being investigated wether these sequences could have some function [2]. In our own research, we observe a high conservation in these sequences and also some ESTs overlapping with the original sequence. Usually pseudogenes are not coding any protein but they might perform some function as RNA altering the expression.

References:

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

2. Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Francisco Carter DR. Pseudogenes: Pseudofunctional or key regulators in health and disease? RNA. 2011;17(5):792-798. doi:10.1261/rna.2658311.