Hw2

Home work 2 - to be done as groups

Names: Group:

For deadlines etc, see Absalon.

You have to supply both the answer (numbers, a table, plots or combinations thereof) as well as the R or Linux code you used to make the plots. This should be done using this R markdown template: we want both the R markdown file and a resulting PDF. For PDF output, you may have to install some extra programs - RStudio will tell you.

Note that:

1. If the R code gives different results than your results, you will get severe point reductions or even 0 points for the exercise
2. Some questions may request you to use R options we have not covered explicitly in the course: this is part of the challenge
3. While this is a group work, we expect that everyone in the group will have understood the group solution: similar or harder question might show up in the individual homework. So, if something is hard, it means you need to spend more time on it
4. The results should be presented on a level of detail that someone else could replicate the analysis.

For statistical tests, you have to:

1) Motivate the choice of test
2) State exactly what the null hypothesis is (depends on test!)
3) Comment the outcome: do you reject the null hypothesis or not, and what does this mean for the actual question we wanted to answer (interpretation)?

A question marked * means that is more challenging, and likely requires skills from the whole group.

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

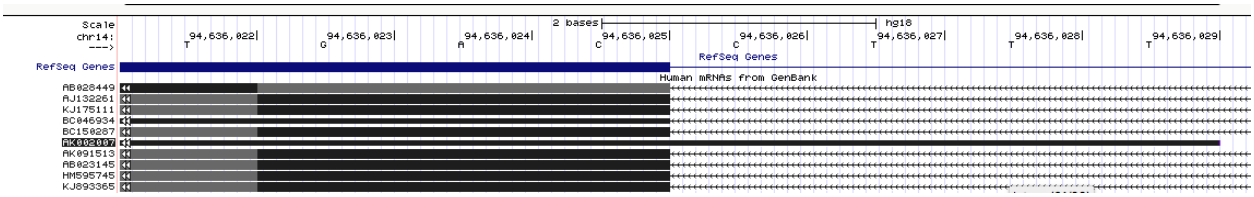

Figure 1: UCSC screenshot, showing the five first bases of transcription of AK002007

```
                     /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 tccttttttca 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.

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.

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: