

# 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 polymerase II from this transcript?
- b) Look at the raw mRNA sequence of AK002007, from the database it actually comes from. What are the first five nucleotides?
- c) How do you explain the discrepancy (maximum 5 lines)?

### Question 2: ERA and ERB

Our collaborators designed a ChIP study using so-called tiling arrays (an outdated technique these days, but the top of the pop at the time: see [http://en.wikipedia.org/wiki/Tiling\\_array](http://en.wikipedia.org/wiki/Tiling_array)): one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)