

Applied bioinformatics: Homework 2

Applied Bioinformatics Homework 2: Suggested solution (one of many possible max-scoring solutions)

Solution text in red, common mistakes in green

1) (4p)

Your lab group wants to create a tiling array investigating the RNA expression, with strand information, from chromosome X in the human genome (hg18). Part of this is to design the probes. Here are the locations of four suggested probes. Your role in this is to decide why these are good or bad probes.

A max-scoring answer says whether the probe is worth making or not (so yes or no), and why that is (an argumentation). No points are rewarded without argumentation. Pictures are not necessary in this answer.

The key issue when making probes is that they have to be uniquely map-able on the genome, otherwise it is not clear what their signal is indicative of. Three of the probes below have such problems

a) chrX:5,818,951-5,818,986

[There is nothing strange with this region – no repeats, and it is unique on the genome. Yes, make the probe]

b) chrX:119,095,715-119,095,747

[This region looks OK, but it is not unique in the genome. A simple BLAT search shows that it maps to four places, where some are overlapping the indicated region but on the opposite strand. Clearly non-ideal, so we cannot approve this probe]

c) chrX:38,989,820-38,989,844

[This region sites in a large repetitive region. This will likely hybridize to anything expressed from any such region over the whole genome, which is not wanted. Do not make this probe]

d) chrX:58,598,710-58,598,742

[This region has repeats, but even worse, a lot of Ns instead of proper nucleotides, since it is bordering the centromeres. Totally meaningless as a probe – not approved.]

Comment: some common mistakes:

Many students thought that tiling array probes are better if they overlap mRNAs (and, conversely, that probes are bad if they are not overlapping mRNAs). This is not true, as the whole point with a tiling array is to be unbiased in terms of annotation, so that you can see any mRNAs, including unknown ones. In the same line of thinking, the mRNA tracks are not the whole truth, just the mRNAs we know about at present. If we wanted to just detect known mRNAs we could have settled for one probe per gene (= an old-school microarray)

2) (4p)

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: AK002007.

What are the first five nucleotides from the first inferred exon of this mRNA (so, what are the first 5 nucleotides that are read by polymerase II according to the browser)?

[This question can actually be read in two ways: As referring to the DNA sequence or the mRNA sequence. The important thing is to remember that the gene is located on the reverse strand. So either 'AAAGG' or 'TTTCC' is correct, and can be found by zooming in the genome browser]

Then look at the raw mRNA sequence of AK002007. What are the first five nucleotides?

[Clicking at the mRNA ID will bring us to an information page, where we can find the sequence. It starts with 'GAAGC' -the first 7 nt are not mapped to the genome, and so don't show up in the browser. It is also possible to see this by looking at the actual alignment]

Applied bioinformatics: Homework 2

How do you explain the discrepancy?

[Remember that the mRNA track is created by mapping the actual cDNA sequences back to the genome, so there may be differences between the sequences. The ends of cDNAs often contain errors, which could be due to low sequence quality, experimental artefacts such as fragments from cloning vectors, or truncated cDNAs. (In this case the extra sequence actually appears to belong to the previous exon (present in other cDNAs from this gene), so presumably a truncated mRNA)].

Common mistakes:

Not realizing that the transcript is one the minus strand (and therefore taking the 3-prime end instead of the 5-prime)

Mistaking the first protein CODING part of the mRNA as the start of the mRNA (confusing transcription and translation). The question clearly says “the first 5 bp read by polymerase II according to the browser”, not “the first 5 bp read by the ribosome”. Again: the definition of exon is that it is part of a mature (usually spliced) mRNA. Protein coding potential has nothing to do with this definition: untranslated parts of mRNAs are also exons.

Not understanding the difference between a sequenced mRNA, and the mapping of that mRNA, and buying the mRNA sequence as “truth”

Not actually explaining the difference, just saying that it is there

In general a lot of these errors can be attributed to not thinking about what the data is (it is not magic truths from heaven, it comes from experiments which then are mapped to the genome)

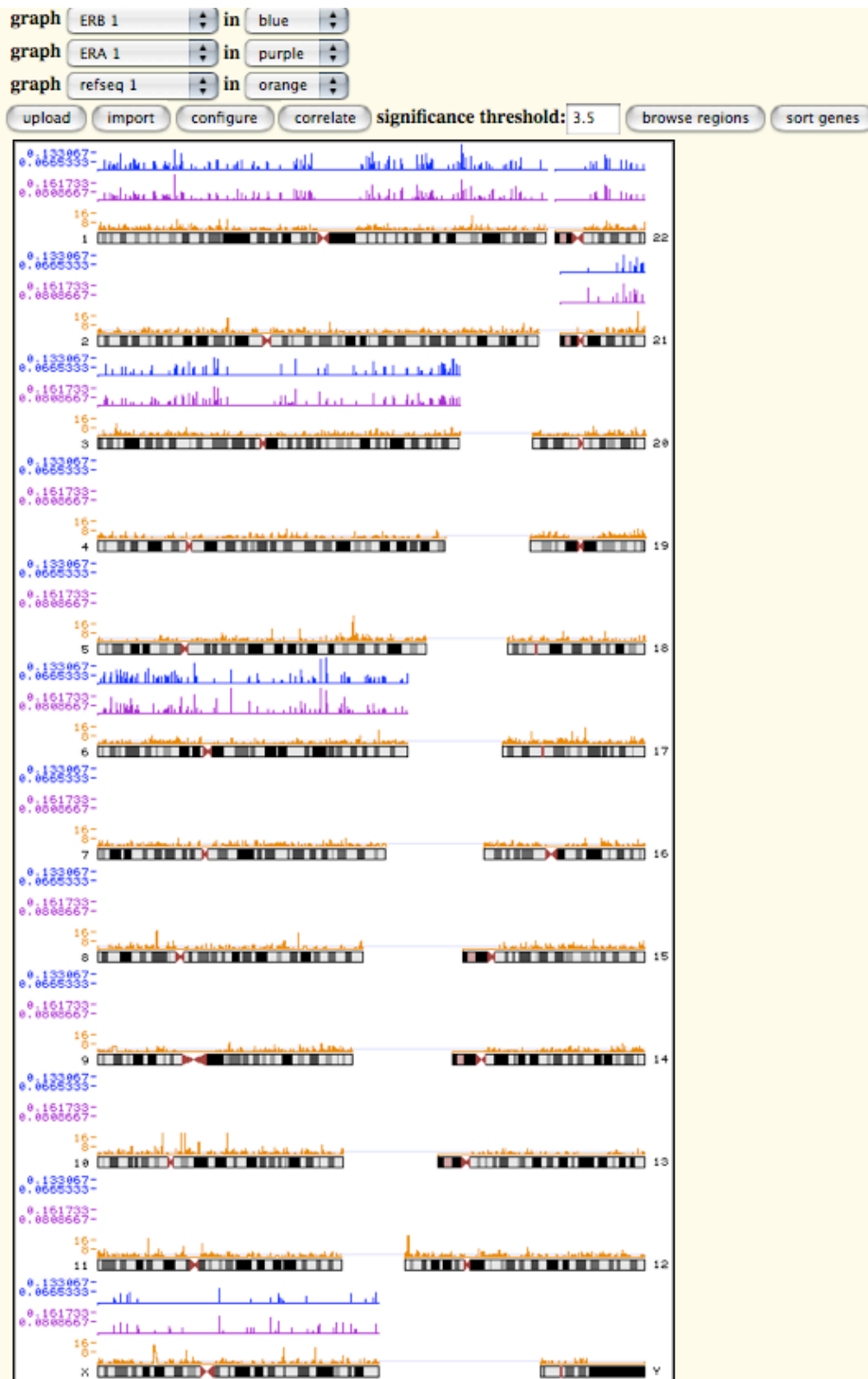
In the two remaining questions we will return to the estrogen receptor ChIP-chip sets we worked on in the last homework, and characterize some of it on a genome-wide scale. The start data correspond to two bed files, one for estrogen receptor alpha (ERA), one for estrogen receptor beta (ERB). These are available in the homework directory.

3) (4p)

Upload these to the browser and then make a genome plot (you can access the tool by clicking “home” in the browser – it is in the left column – “Genome Graphs”) with all ERA sites, all ERB sites, and all Rfseq genes (same plot but distinct barplots for each dataset and different colors). Briefly comment the pictures. Is there anything particularly surprising? Can try to explain the outcome (biological and/or experimental setup explanations)?

We upload these files to the browser using the correct assembly, hg18 (found in the file names) to the browser. This is straightforward, using Add Custom Tracks, and then upload the files, one by one. We then import these into the genome Graphs tool, using import. We choose the custom track to import the ERA and ERB tracks, filling in ERa and ERB as names, respectively, and the Genes and gene prediction group to import the refseq track (Refseq Genes, refGene table). We can show everything in the same picture by setting number of rows to 3 in “configure”

Applied bioinformatics: Homework 2



Click on a chromosome to open Genome Browser at that position.

There are a few things to note with this image:

ERA and ERB signals are overlapping often – they are very similar but not identical. It is possible to correlate them using the correlate button, which gives a pretty nice correlation (not necessary though). This

Applied bioinformatics: Homework 2

might either be due to that they bind very similar target, or that the antibodies used are not very specific (it turned out later that the antibodies are fine, so it is real biology, but there is no way to see this from this particular data)

ERA and ERB signals are only present in some chromosomes, and have no signal whatsoever in the remaining chromosomes

The gene density varies along chromosomes, but is never totally zero. There seems to be no real correlation with the ERA tracks by eye.

It is unrealistic to think that there are no signals whatsoever on those chromosomes lacking one, especially since we have TWO experiments saying the same thing (also, when we looked at human ERA, it covered all human chromosomes). We also have genes all over all the chromosomes.

Either something went horribly wrong with the experiment OR there are simply not any probes covering these regions in the experiment in question. Even if something went wrong, it is unlikely that specific chromosomes are omitted – real experimental errors tend to be much more random.

In fact, this is a “half-genome” tiling array that the experimentalists used to save money. So the probes are only for some of the chromosomes.

Common mistakes:

Over-plotting tracks – this is not a major issue, but I ask for distinct plots. It is hard to see anything if all the plots are on top of each other

Not realizing (or arguing why) the ERA/ERB tracks are totally missing from half the chromosomes.

4) (8p)

We want to know if ERA sites in general are more conserved over evolution than ERB sites.

One way of measuring conservation is to take the mean PhastCons conservation over all the nucleotides in a given binding site. This will then give one value for each site in the data set.

A potential problem in this analysis is that the functions of the two transcription factors overlap – some binding sites are shared. This is an obvious violation to the assumption that the sets are independent.

Therefore, we have to remove these sites from the analysis and focus on independent sites.

Show the distribution of conservation (using mean phastcons scores for individual sites) of both sets of sites after removing those sites that are shared. Are these distributions significantly different in terms of mean or medians?

SOLUTION

We upload the ERA track into galaxy by (in galaxy) saying

Get Data->UCSC Main table browser:

Assembly hg18, group custom tracks, track ERA, table ct_ERA, output format=bed

We do the same for ERB.

We then want to overlap these sets to take away sites that are overlapping between alpha and beta.

It is possible to do this in many ways, but this might be the simplest – it is just four operations.

We do Operate on genomic intervals->subtract where ERA is the first query and ERB is the second. We set “return” to intervals with no overlap, as we want to have full regions with no overlap between the sets. This will give all the ERB regions with no ERA overlaps. It is easy to be confused by which set is which – I looked a few sample regions in the genome browser to see what set they came from to be sure. Sanity checks are always good!

For these “only” ERB sites, we want to fetch PhastCons scores. We do this using Get aggregate scores, to get Phastcons 17 species alignments (cached).

This gives three extra columns to the file – mean score, max score and min score. We save this file for later R analysis, with the name only_erb.txt. I open this file and add headers –

```
chr      start      end      mean_val      min_val max_val
```

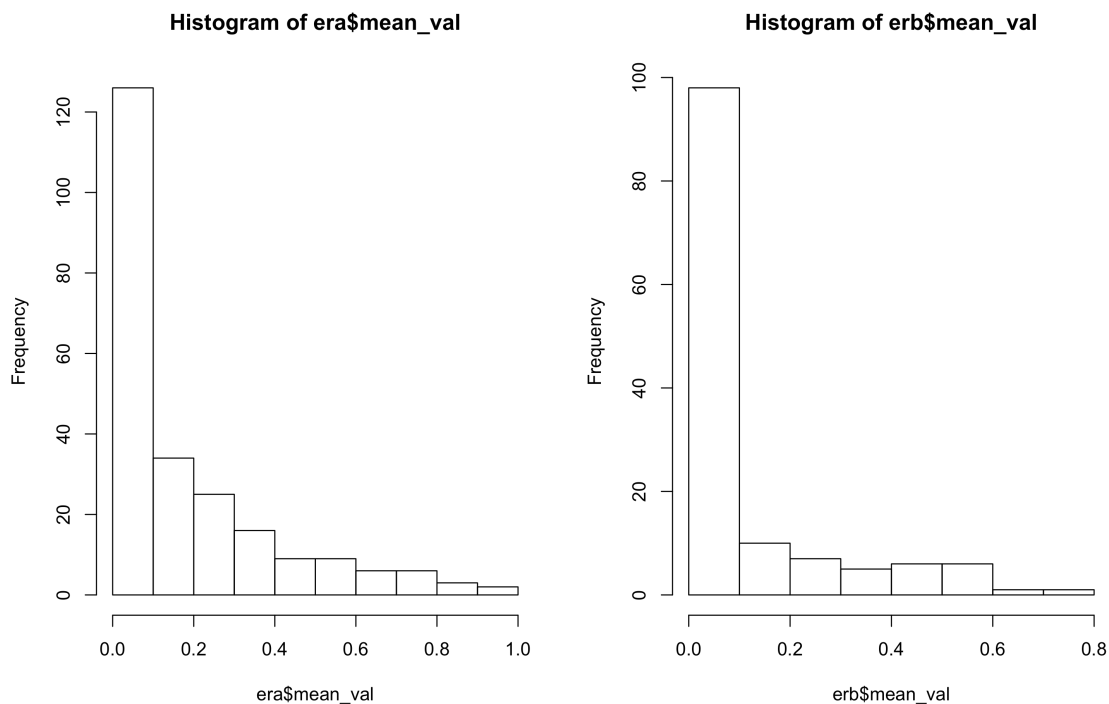
Applied bioinformatics: Homework 2

It is a real bad idea to call the columns things like “mean”, since there is also a function called mean() in R

We repeat the above with ERA (so, we get all ERA regions that are NOT overlapped by ERB), and repeat the aggregate score analysis on these, save them as only_era.txt, and add headers.

Then we open R and make the analysis. We want to have the distributions and to see if one distribution is significantly more “conserved” than the other. This is a two-sample tests, so either t-test or Wilcox test is applicable, depending on if the data is normal-shaped or not.

```
era<-read.table("era_only.txt", h=T)
erb<-read.table("erb_only.txt", h=T)
par(mfrow=c(1,2))
hist(era$mean_val); hist(erb$mean_val)
```



Clearly not normal-distributed! So, a Wilcox-test is in order, since it is a two-sample test.

Just to get our head straight – do the distributions have different observed medians?

```
median( era$mean_val)
[1] 0.08296669
> median( erb$mean_val)
[1] 0.02194220
```

Yes they do. It seems ERB is less conserved. The null hypothesis is that they have the same median (or, very technically, that they have a location shift of >0, as Wilcox test is not exactly looking at medians)

```
wilcox.test(era$mean_val, erb$mean_val)

Wilcoxon rank sum test with continuity correction
```

data: era\$mean_val and erb\$mean_val

W = 19885, p-value = 3.811e-05

alternative hypothesis: true location shift is not equal to 0

The P-values is much lower than 0.05 and we can reject the null hypothesis. ERB sites are less conserved than ERA sites (with the data that we look at) .

Applied bioinformatics: Homework 2

Common mistakes:

The most common mistake is to give too few details for someone else to replicate the test, in particular the overlap.

The second most common mistake is to make an error when making the overlaps. The solution to this is to do sanity checks. There is a high correlation between these two errors, which sort of shows that it helps to explain things to another person