HW3 – Bioinformatics – 236523

Anna Romanov 321340580 [annarom@campus.technion.ac.il](mailto:annarom@campus.technion.ac.il)

Maxim Kolchinsky 320983216 [kolchinsky@campus.technion.ac.il](mailto:kolchinsky@campus.technion.ac.il)

# Question 1

Part a

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | sample\_name | cell line | dex | | AE1148 | EBC1 | No treatment | | AE1149 | EBC1 | No treatment | | AE1150 | EBC1 | No treatment | | AE1151 | EBC1 | Crizotinib | | AE1152 | EBC1 | Crizotinib | | AE1153 | EBC1 | Crizotinib | | AE1154 | EBC1 | Interferon Î³ | | AE1155 | EBC1 | Interferon Î³ | | AE1156 | EBC1 | Interferon Î³ | | AE1157 | H1573 | No treatment | | AE1158 | H1573 | No treatment | | AE1159 | H1573 | No treatment | | AE1160 | H1573 | Interferon Î³ | | AE1161 | H1573 | Interferon Î³ | | AE1162 | H1573 | Interferon Î³ | | AE1163 | H1993 | No treatment | | AE1164 | H1993 | No treatment | | AE1165 | H1993 | No treatment | | AE1166 | H1993 | Interferon Î³ | | AE1167 | H1993 | Interferon Î³ | | AE1168 | H1993 | Interferon Î³ | | AE1169 | H596 | No treatment | | AE1170 | H596 | No treatment | | AE1171 | H596 | No treatment | | AE1172 | H596 | Crizotinib + Hepatocyte growth factor (HGF) | | AE1173 | H596 | Crizotinib +HGF | | AE1174 | H596 | Crizotinib +HGF | | AE1175 | H596 | Hepatocyte growth factor (HGF) | | AE1176 | H596 | Hepatocyte growth factor (HGF) | | AE1177 | H596 | Hepatocyte growth factor (HGF) | | AE1178 | H596 | Interferon Î³ | | AE1179 | H596 | Interferon Î³ | | AE1180 | H596 | Interferon Î³ | |  |  |

Part e

Cell types: (4 cell types)

1. EBC1 – 9 samples
2. H1573 – 6 samples
3. H1993 – 6 samples
4. H596 – 12 samples

Treatments: (6 types of treatments)

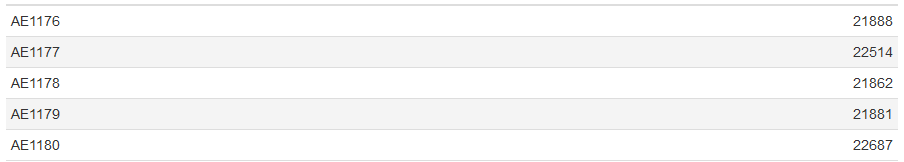
1. No treatment – 12 samples
2. Crizotinib – 3 samples
3. Interferon Î³ - 12 samples
4. Crizotinib + Hepatocyte growth factor (HGF) – 1 sample
5. Crizotinib +HGF – 2 samples
6. Hepatocyte growth factor (HGF) – 3 samples

Expressed Genes:









Highly expressed Genes:

I used excel command in counts file for each row: COUNTIF and got the following result

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | AE1148 | 6601 | | AE1149 | 7346 | | AE1150 | 6082 | | AE1151 | 6955 | | AE1152 | 6935 | | AE1153 | 7270 | | AE1154 | 5759 | | AE1155 | 6826 | | AE1156 | 6407 | | AE1157 | 6938 | | AE1158 | 6109 | | AE1159 | 6675 | | AE1160 | 7066 | | AE1161 | 5855 | | AE1162 | 7193 | | AE1163 | 6717 | | AE1164 | 6092 | | AE1165 | 6595 | | AE1166 | 7059 | | AE1167 | 6227 | | AE1168 | 6981 | | AE1169 | 7173 | | AE1170 | 6082 | | AE1171 | 6627 | | AE1172 | 5986 | | AE1173 | 6769 | | AE1174 | 6523 | | AE1175 | 6643 | | AE1176 | 5917 | | AE1177 | 6441 | | AE1178 | 5553 | | AE1179 | 5823 | | AE1180 | 6678 | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Filtered data.frame:

For the following 2 treatments:



I have a total of 15 samples

Genes differentially expressed with threshold < 0.05:

There are 353 genes with adjusted p-value < 0.05

Genes differentially expressed with threshold < 0.01:

There are 182 genes with adjusted p-value < 0.01

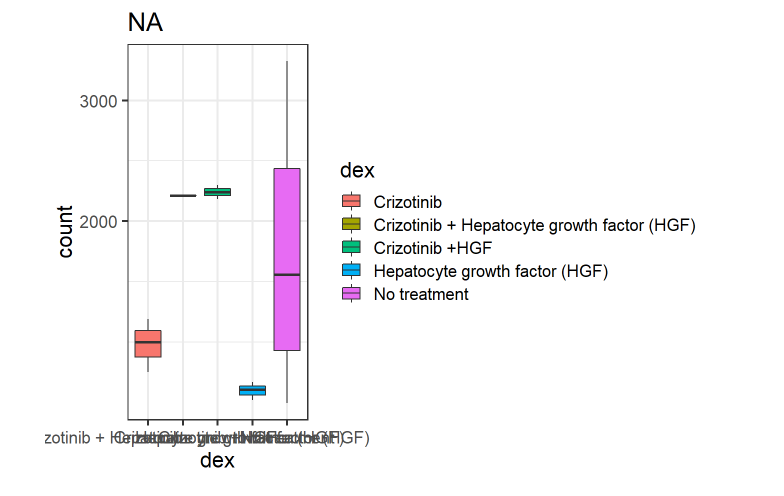
Log2fold change:

There are 3220 with log fold change above 2

There are 233 with log fold change below -2

How significant is the effect of the treatment on gene expression levels?

As we can see in the box plot:



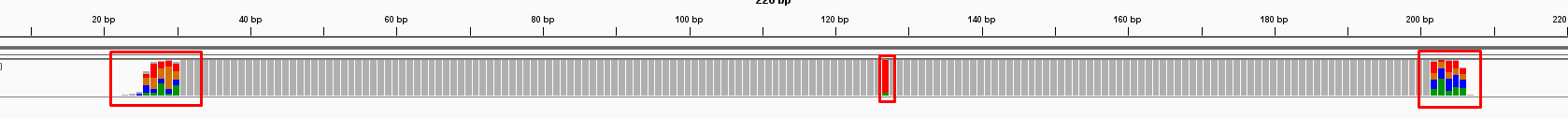
It seems like HGF is the only treatment which is significant on the gene expression.

# Question 2

1. Gene name is : TMC1 in Mus musculus

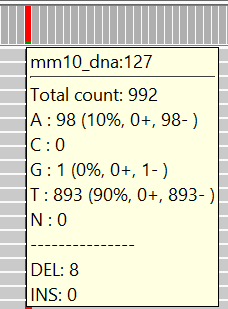
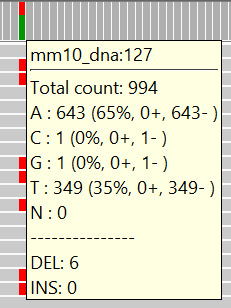
I found it by searching the content of gene.fa in blastn.

* Here’s the significant variation that is present in both samples according to the coverage track:



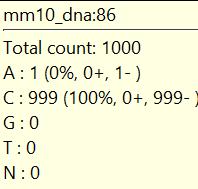
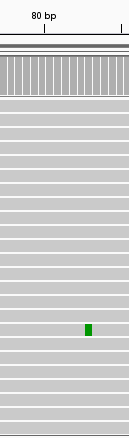
* The position is mm10 genome reference sequence, at locations 25-30, 127 and 202-207. The one significant variation we will be talking about next is in position 127.
* According to HW1 a potential phenotype could be hearing loss.
* The abundance of bases in position 127:

for sample1 – for sample2 –

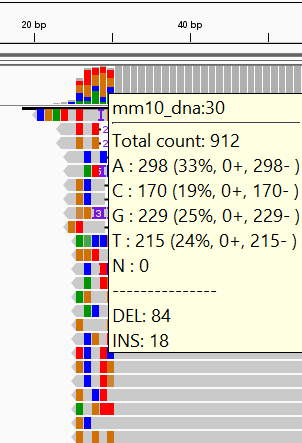
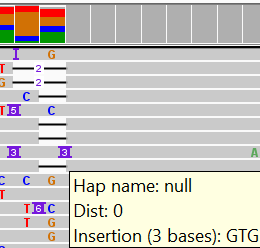
We see that the abundance of bases is different in the two samples. In both samples, A and T are swapped, however in the first sample there are mostly T’s (90%) while in the second sample there are more A’s (65%). The reason for this could potentially be due to one sample belonging to a homozygote and the other to a heterozygote for that gene.

* For example, in sample1 we can see that in position 86 there’s a variation in a single read:

Out of the 1000 reads, in one of them the base C was swapped with base A. This variation is insignificant since we see it in only 0.1% of the reads, therefore we assume it could be caused by an error in the sequencing process.

* Example for indels in sample1:

 close up: 

* We can see many indels near the ends of the sequence, specifically around locations ~30 bp and ~200 bp. However there are also indels in the other locations, for example in sample2 there’s an insertion near 90 bp:



* The cause for deletion events
* The effect of deletion events on he phenotype
* A deletion in TMC1 could either be a random phenomenon (like most mutations are random) or it could be, based on HW1, caused by CRISPR. The abundance of changes at the beginning and end of the sample could be explained most likely by gene editing while the low amount of changes in the middle is most likely random.
* A single base deletion mutation will either cause a missense or a nonsense mutation which replaces a codon in the amino acid. We learned that there are different categories for amino acids with different properties. This kind of mutation can cause creation of a different amino acids from the one expected. A change in a single amino acid could (depending on the properties) completely change the phenotype or it could be a different amino acid with similar properties and the result be no change to the phenotype.

# Question 3

1. Estimating the abundance of different transcripts of a gene is more challenging than estimating the abundance of different genes, since there is more variability in different genes compared to different transcripts, which are more similar to each other. In RNASeq, the used sequences are short and correspond to a portion rather than the full length of an mRNA. This often causes ambiguity in the source of a sequenced RNA fragment, because a read can map to multiple locations in the genome or to a unique location that belongs to multiple isoforms. This is caused by the loss information about the exon in short reads. This way, we are unable to map sequences that come from shared exons. In microarrays information is similarly lost, causing ambiguity.
2. A possible way to overcome the difficulty is to use longer reads if possible. In RNASeq, we need to consider the different exons when looking at the counts, and similarly in microarrays we need probes for different exons.

There are algorithmic solutions such as the [Trinity Transcript Quantification](https://github.com/trinityrnaseq/trinityrnaseq/wiki/Trinity-Transcript-Quantification) for RNASeq. Also there are solutions such as [Gene Meter](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5595416/) for micro arrays.