HW3 – Bioinformatics – 236523

Anna Romanov 321340580 annarom@campus.technion.ac.il
Maxim Kolchinsky 320983216 kolchinsky@campus.technion.ac.il

Question 1

Part a

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

Part e

Cell types: (4 cell types)

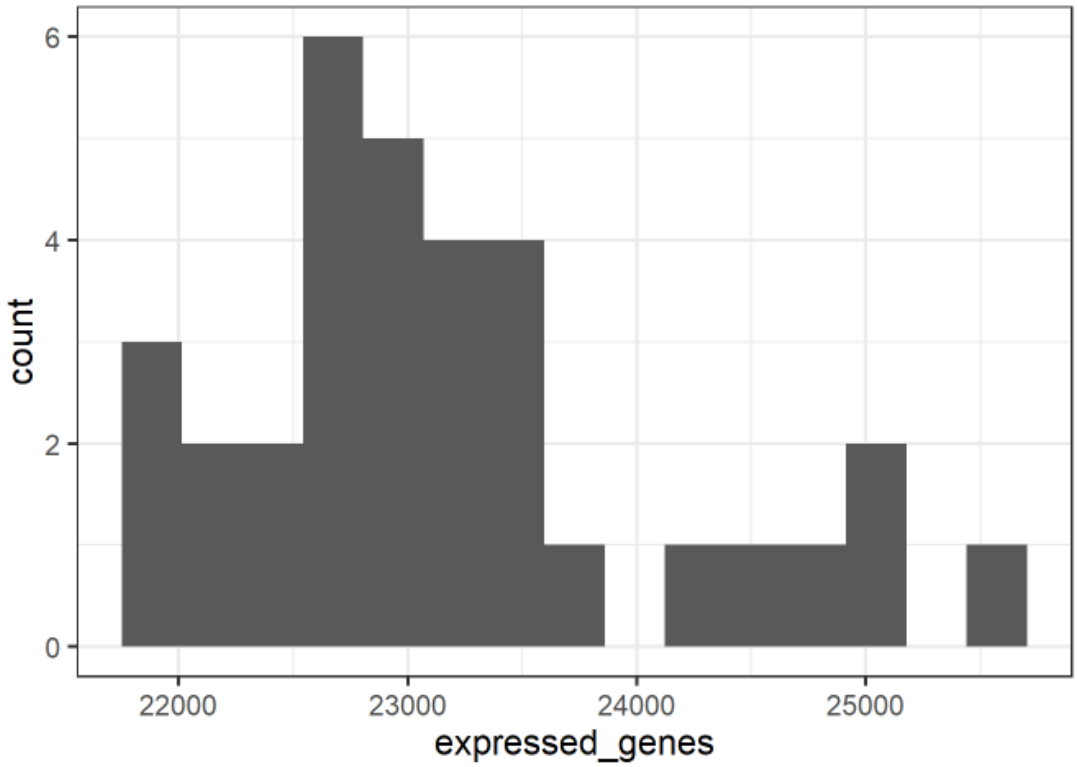
1) EBC1 – 9 samples
2) H1573 – 6 samples
3) H1993 – 6 samples
4) H596 – 12 samples

Treatments: (6 types of treatments)

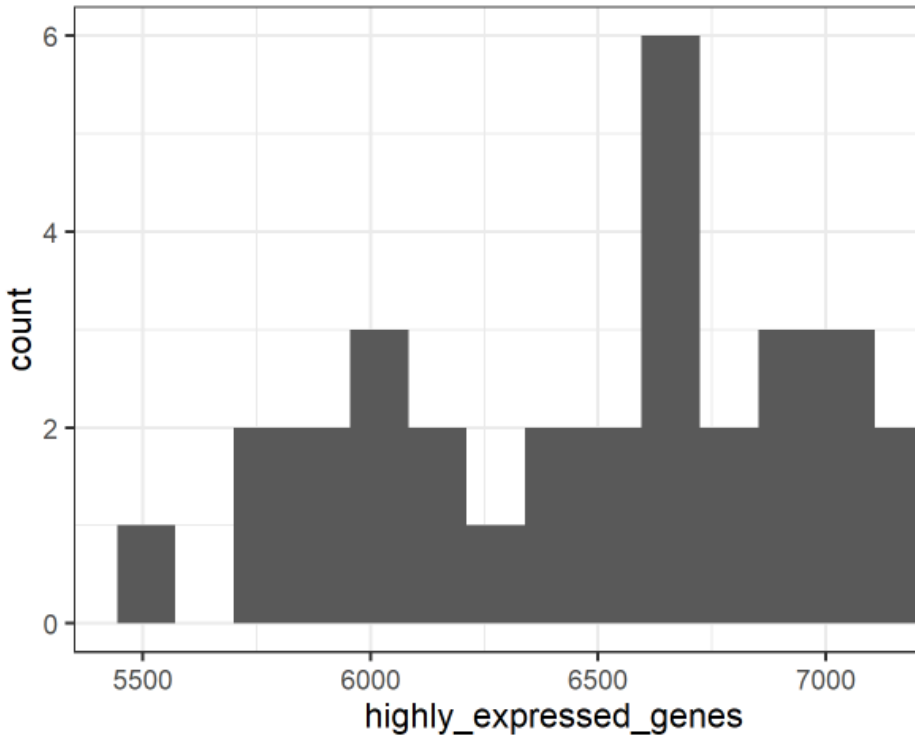
1) No treatment – 12 samples
2) Crizotinib – 3 samples
3) Interferon Î³ - 12 samples
4) Crizotinib + Hepatocyte growth factor (HGF) – 1 sample
5) Crizotinib +HGF – 2 samples
6) Hepatocyte growth factor (HGF) – 3 samples

Expressed Genes:

The histogram:



Highly expressed Genes:
The histogram:



Filtered data.frame:
For the following 2 treatments:

```
treatment1<- 'No treatment'
treatment2<- 'Crizotinib'
```

And celltype EBC1:
There are a total of 6 samples as can be observed here:

```
filtered_metadata
```

id <chr>	celltype <chr>	dex <chr>	geo_id <chr>
AE1160	EBC1	No treatment	GSM2949380
AE1163	EBC1	No treatment	GSM2949381
AE1166	EBC1	No treatment	GSM2949382
AE1162	EBC1	Crizotinib	GSM2949383
AE1165	EBC1	Crizotinib	GSM2949384
AE1168	EBC1	Crizotinib	GSM2949385

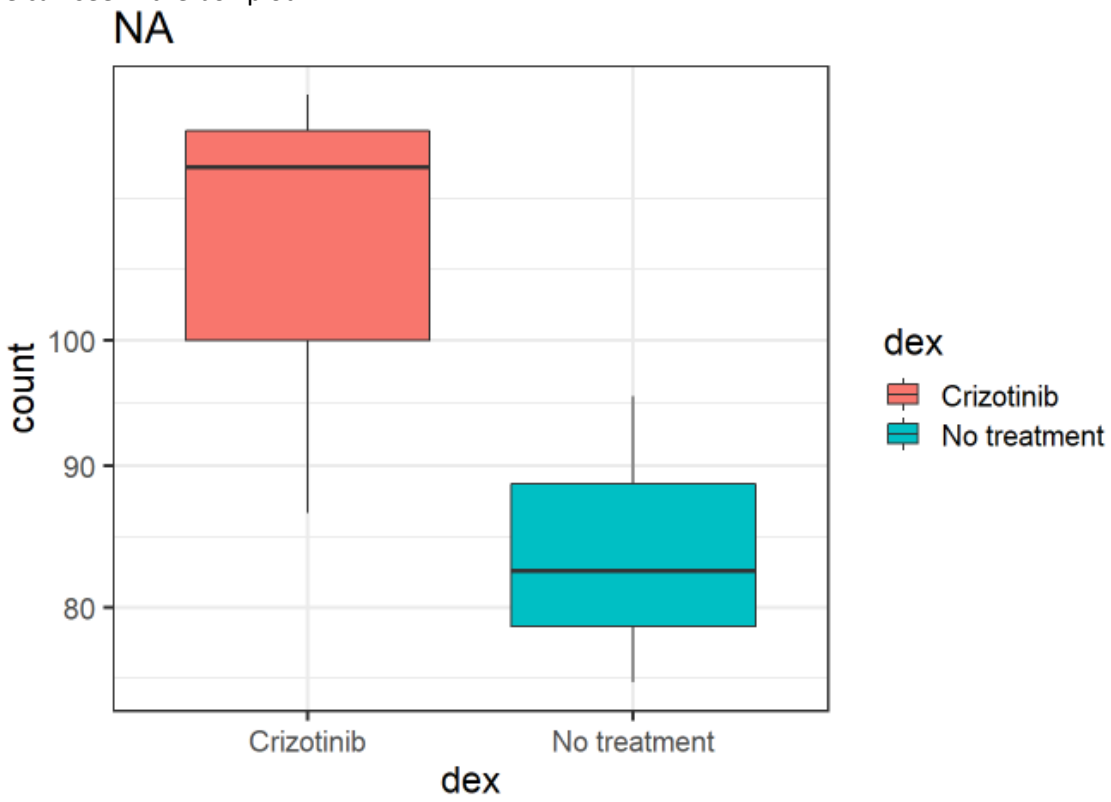
6 rows

Genes differentially expressed with threshold < 0.05:
There 5264 are genes with p-value < 0.05
Genes differentially expressed with threshold < 0.01:
There 4001 are genes with p-value < 0.01
Log2fold change:

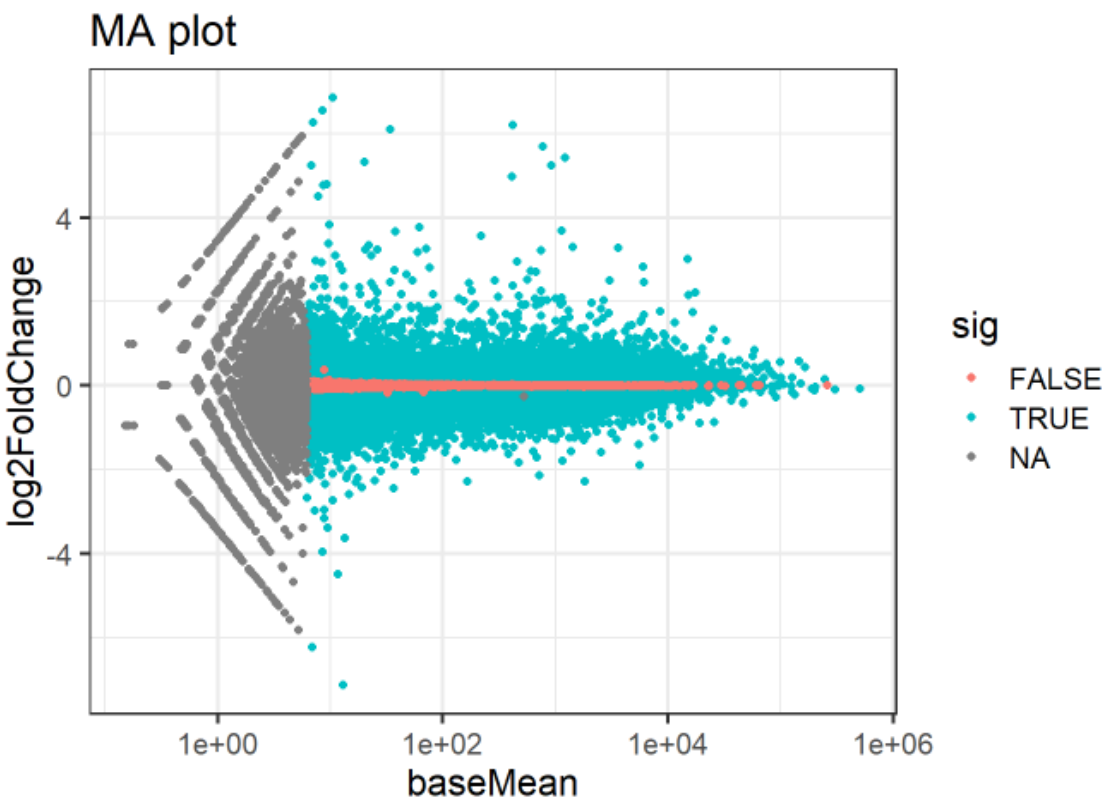
There are 573 with log fold change above 2

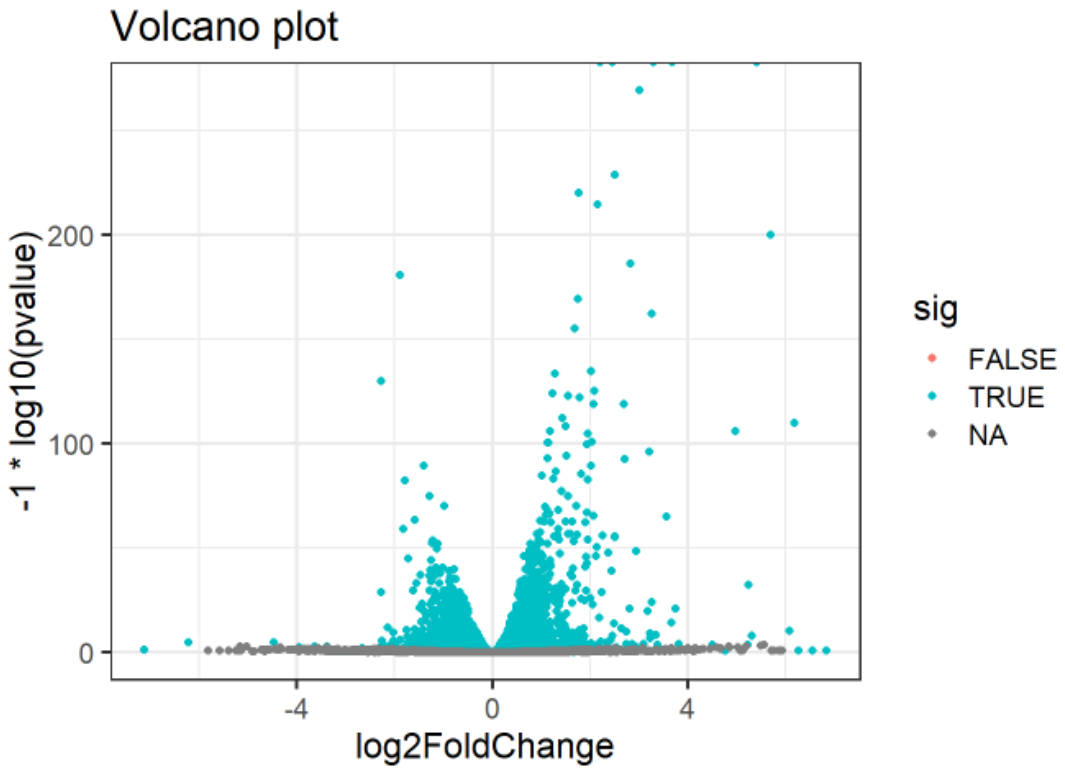
There are 654 with log fold change below -2

<u>How significant is the effect of the treatment on gene expression levels</u>?
As we can see in the box plot:



It seems like Crizotinib is has significant gene expression compared to no treatment.
MA and volcano plots:


MA plot
log2FoldChange
4
0
-4
1e+00
1e+02
1e+04
1e+06
baseMean
sig
FALSE
TRUE
NA


Volcano plot
-1 * log10(pvalue)
200
100
0
-4
0
4
log2FoldChange
sig
FALSE
TRUE
NA

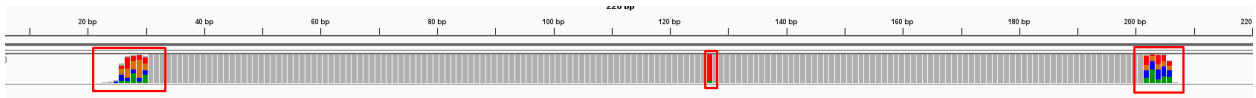
Question 2

b. Gene name is : TMC1 in Mus musculus

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

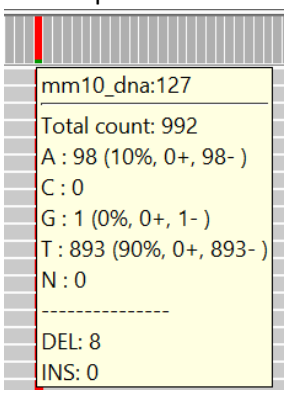
c.

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

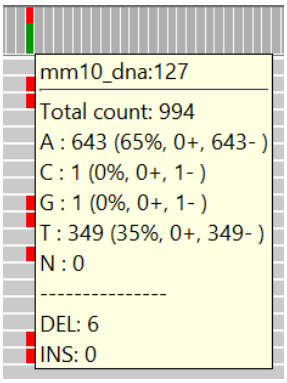


  - The position is mm10 genome reference sequence, at locations 25-30, 127 and 202-207. The one significant variation we will be talking about next is in position 127.
  - According to HW1 a potential phenotype could be hearing loss.
  - The abundance of bases in position 127:

    for sample1 –

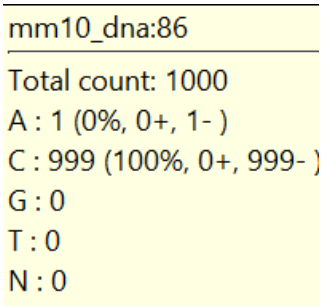


    for sample2 –



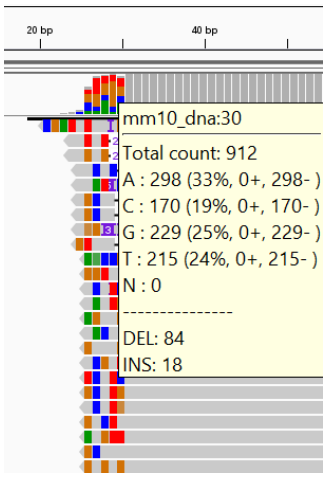
    We see that the abundance of bases is different in the two samples. In both samples, A and T are swapped, however in the first sample there are mostly T's (90%) while in the second sample there are more A's (65%). The reason for this could potentially be due to one sample belonging to a homozygote and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

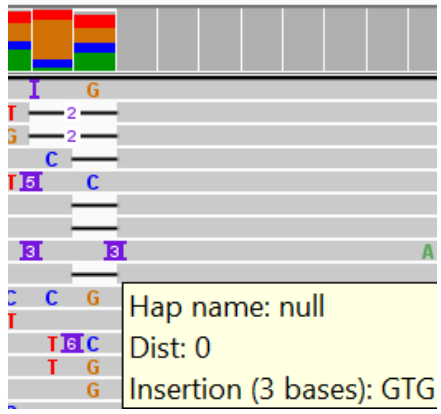
- For example, in sample1 we can see that in position 86 there's a variation in a single read:



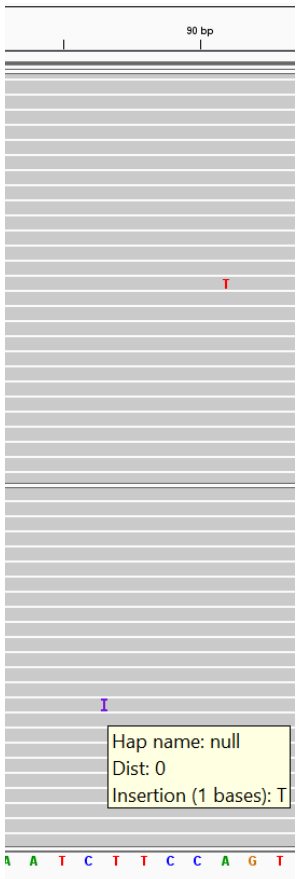
Out of the 1000 reads, in one of them the base C was swapped with base A. This variation is insignificant since we see it in only 0.1% of the reads, therefore we assume it could be caused by an error in the sequencing process.

- 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:



- 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.
As we've seen in the first home assignment, if the indels were indeed caused by CRISPR, the change in phenotype might be an improvement in hearing.

Question 3

a. 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.
b. 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 (which uses alignment-based quantification methods) for RNASeq. For micro arrays, there are solutions such as Gene Meter (which calibrates the probes using a calibration pool of transcripts with known concentrations).