תרגיל בית 3 – מבוא לביואינפורמטיקה

מגישים: אלון הרשקוביץ 203854443, אפרת לבקוביץ 204536270

R markdown and RNA-Seq differential expression analysis

a. sample name, cell line, treatment:

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

e .**How many different cell types? How many samples per cell type?**

```
 EBC1 H1573 H1993  H596
    9     6     6    11
```

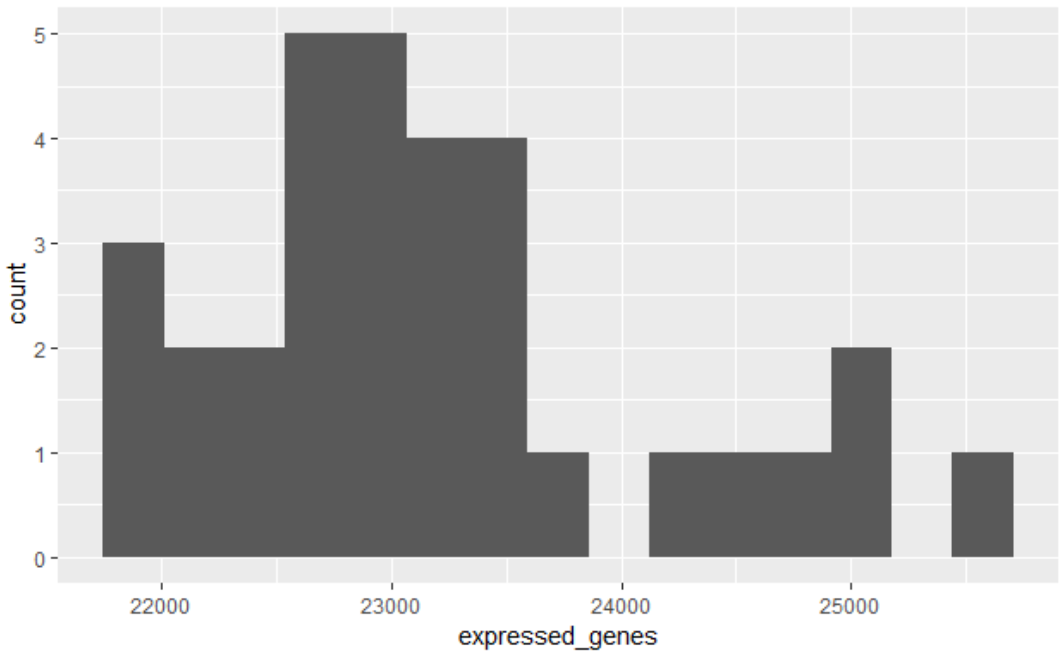
How many treatments? How many samples per treatment type?

```
                                   Crizotinib
                                            3
 Crizotinib + Hepatocyte growth factor (HGF)
                                            1
                              Crizotinib +HGF
                                            2
              Hepatocyte growth factor (HGF)
                                            3
                                 Interferon γ
                                           11
                                 No treatment
                                           12
```

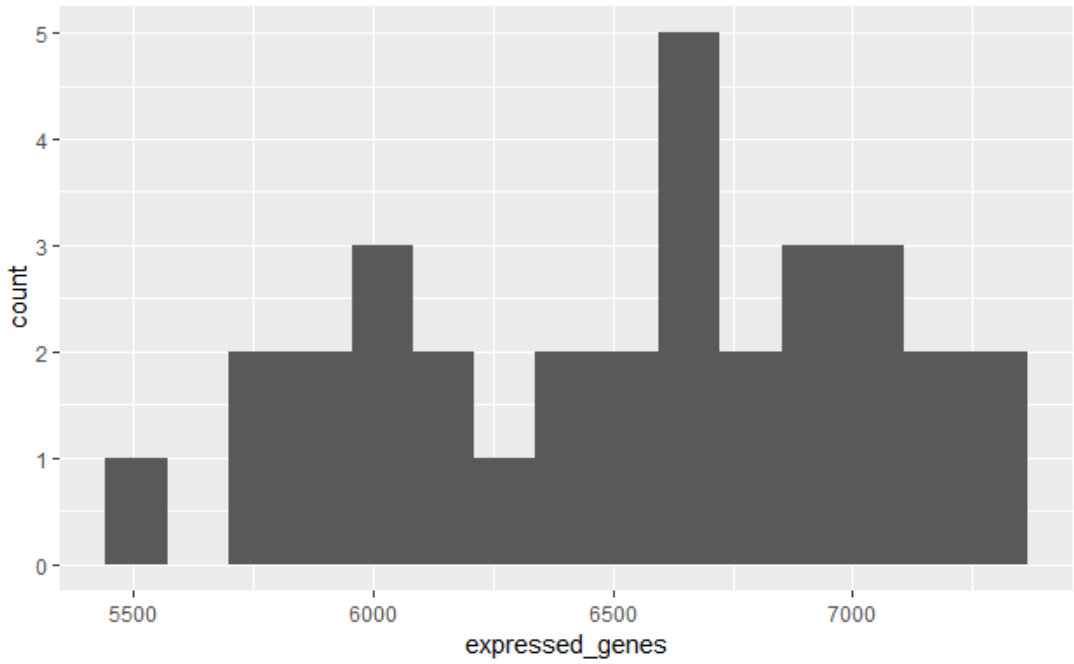
the number of samples for every combination of cell type and treatment.

```
                                            EBC1 H1573 H1993 H596
 Crizotinib                                    3     0     0    0
 Crizotinib + Hepatocyte growth factor (HGF)   0     0     0    1
 Crizotinib +HGF                               0     0     0    2
 Hepatocyte growth factor (HGF)                0     0     0    3
 Interferon γ                                  3     3     3    2
 No treatment                                  3     3     3    3
```

For each sample, how many genes are expressed (expression above zero)? Present a histogram of the number of expressed genes.



For each sample, how many genes are highly expressed (expression above 1000)? Present a histogram of the number of expressed genes.



A filtered data.frame that contains only samples of one cell type and two treatment options.

```
treatment1<- 'No treatment'
treatment2<- 'Crizotinib'
celltype0 <- 'EBC1'
```

6 samples. `"AE1160" "AE1163" "AE1166" "AE1162" "AE1165" "AE1168"`

How many genes are differentially expressed using a threshold of 0.05 on the adjusted p-value?

There are 5265 genes with adjusted p-value < 0.05

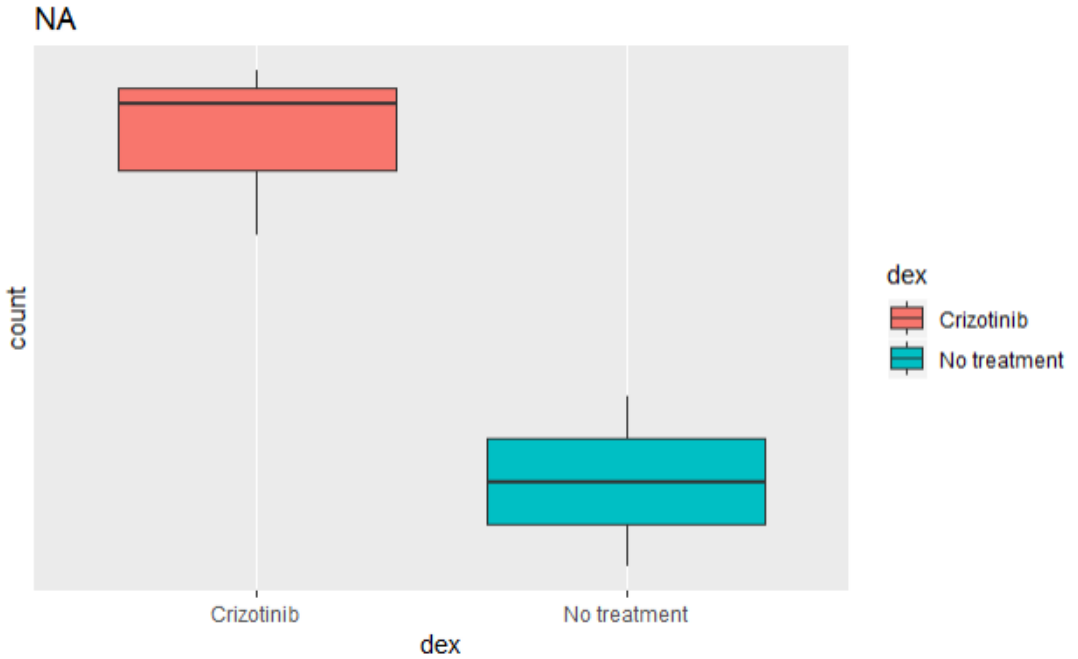
How many with a threshold of 0.01?

There are 4002 genes with adjusted p-value < 0.01

How many gene have a log fold change of above 2? How many below -2?

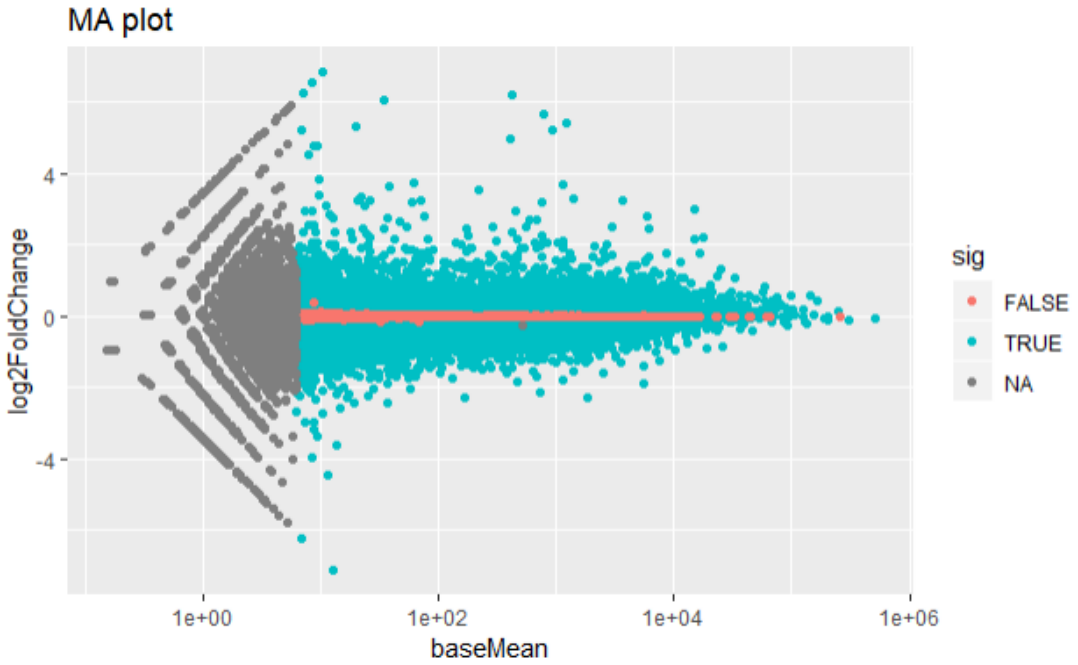
There are 574 genes with log fold change above 2.
There are 655 genes with log fold change below -2.

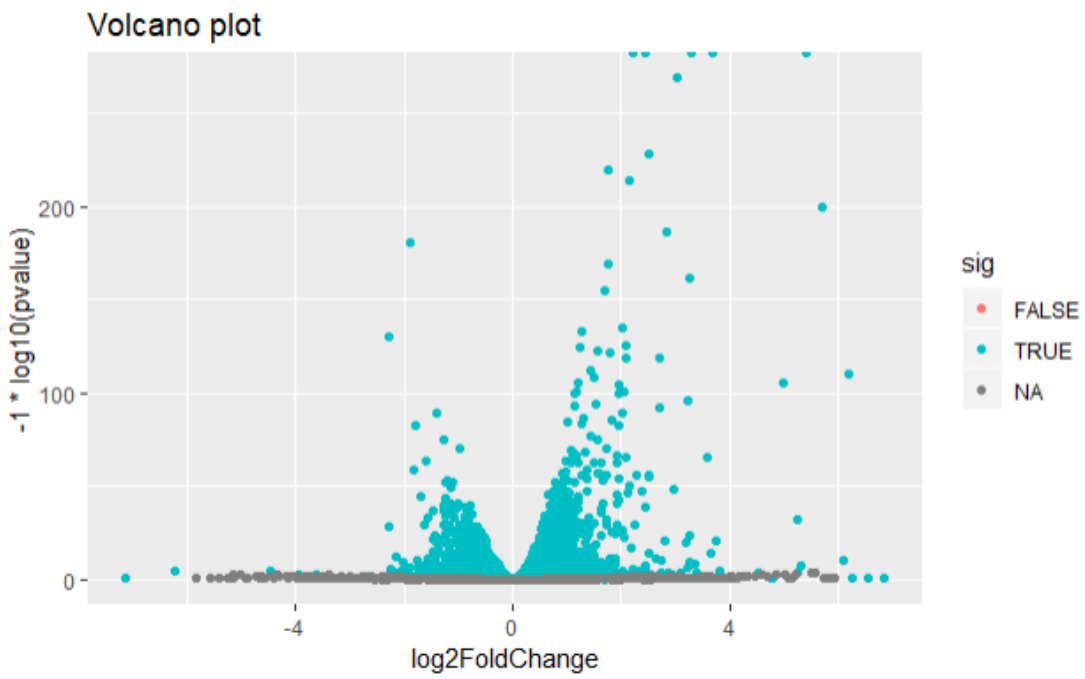
Plot a boxplot of a differentially expressed gene. In which treatment group is it expressed higher?



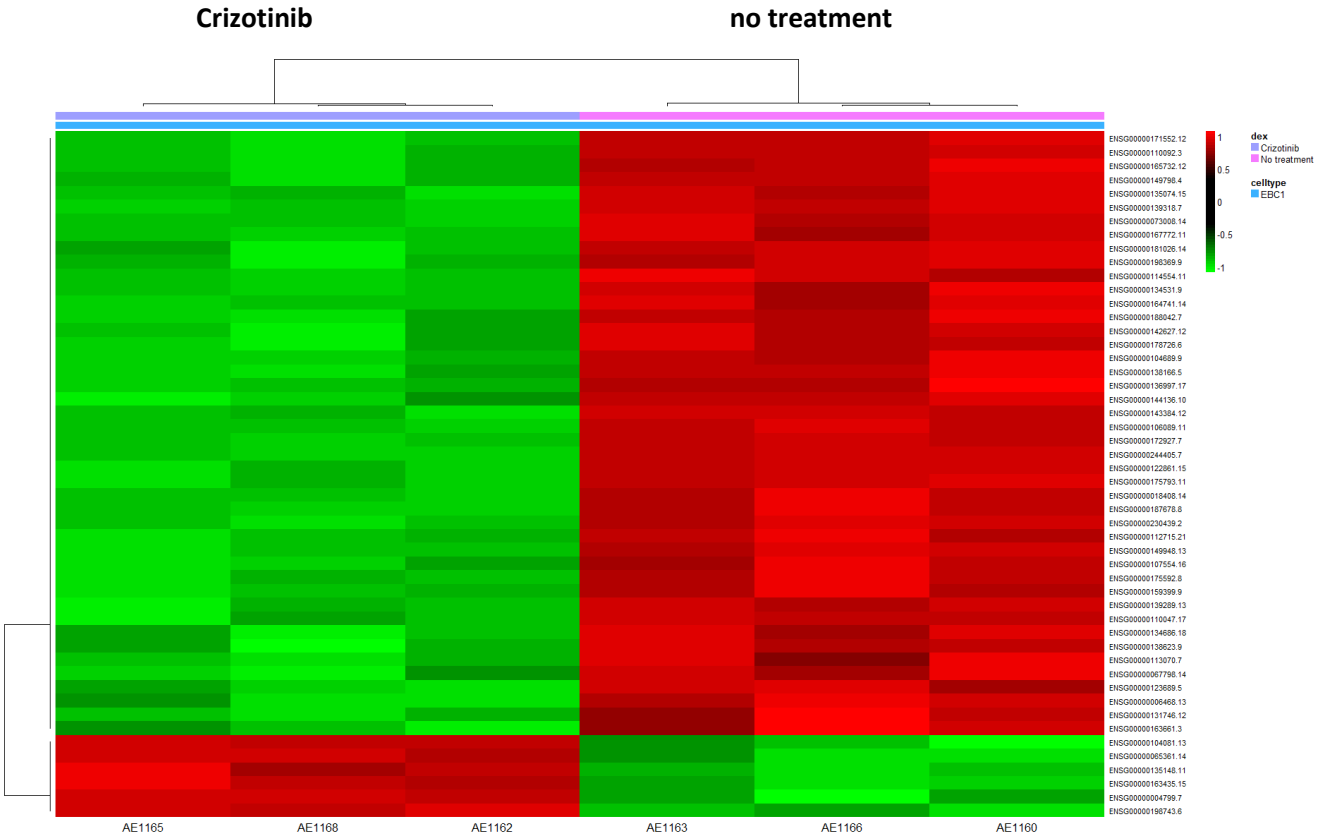
In the Crizotinib treatment group it is expressed higher.

Plot an MA plot and a volcano plot





plot a heatmap of the samples and the 50 most differentially expressed genes.



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

We can see that the Crizotinib treatment is very significant.

DNA sequencing and variant calling

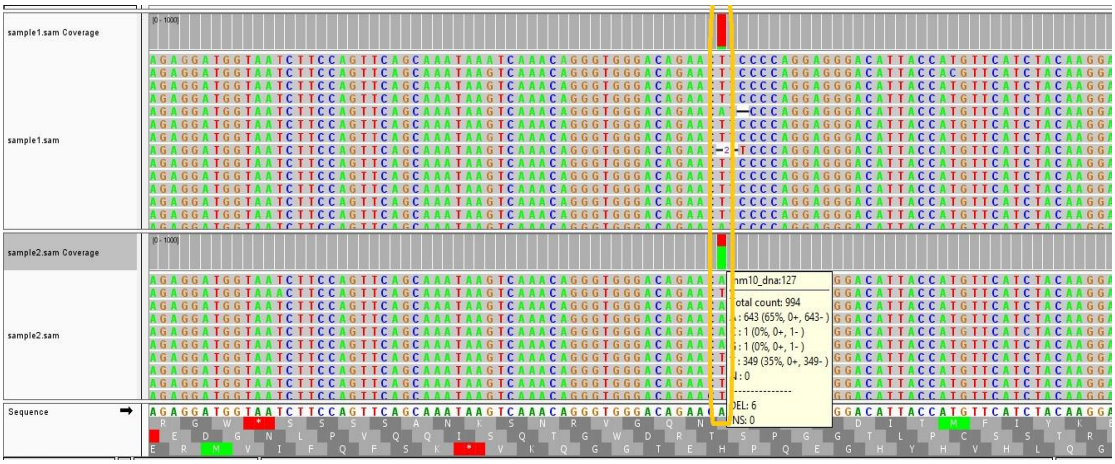
a. --

b. הגן הוא Tmc1 בעכבר. מצאנו אותו ע"י פתיחת קובץ ה-fasta בעורך טקסט, והעתקת הרצף למנוע החיפוש של blastn, יחד עם הרמז ואחוזי ההתאמה הגבוהים אין ספק שמדובר בגן הזה.

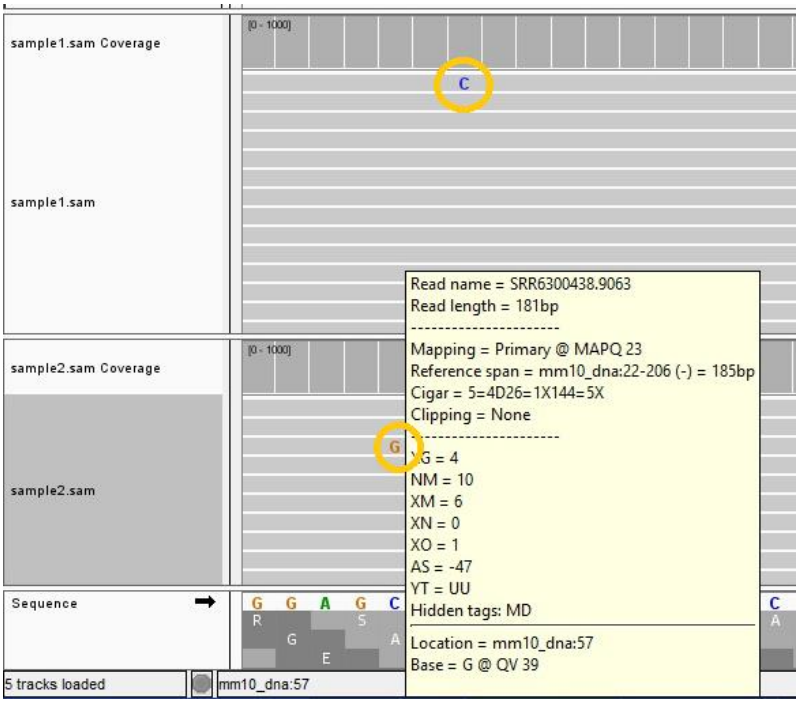
c.

-
  - השוני נמצא בעמדה 127 (כפי שמוצג בתמונה בסוף הסעיפים).
  - כן, חירשות, כפי שראינו בת"ב 1. ניתן למצוא זאת ע"י חיפוש ב-blastn במאגר המידע של גנום עכבר. התוצאה הראשית הינה של הגן המכונה beethoven.
  - בדוגמה הראשונה:
    A – 98, C – 0, G – 1, T – 893
    בדוגמה השניה:
    A – 643, C – 1, G – 1, T – 349
    הריבוי השונה בין הדוגמאות – יתכן שנובע מכך שדוגמה 1 הינה של הומוזיגוט ודוגמה 2 הינה של הטרוזיגוט.
- (תמונה בסוף הסעיפים)
  "מוטציה" כזו יכולה להגרם משגיאה בתהליך הריצוף בה מוכנס בטעות נוקלאוטיד לא נכון ונשאר לאורך הריצוף.
- (תמונה בסוף הסעיפים)
  - לא ניכר שיש מיקומים ספציפיים (מוטציות הן אקראיות), אבל מהסתכלות על כלל הדוגמאות הם לרוב מופיעים ליד המוטציה מתחילת הסעיף, או יותר באיזור הקצוות (או קרוב לקצוות).
  - בקצוות התופעה יכולה פשוט לנבוע מהתהליך עצמו, שפחות יעיל ככל שמתקרבים לקצה/בתחילת הרצף. באיזור השוני ניזכר שזהו אזור מטרה של CRISPR שהינו מערכת עריכת גנים, כלומר ה-deletion יכול להגרם ע"י מערכת עריכת הגנים.
  - עקרונית תופעות מסוג deletion אינן צפויות מאחר ומשנות את מסגרת הקריאה ועלולות לגרום למוטציות נוספות, כמו גם להשתקת המוטציה.

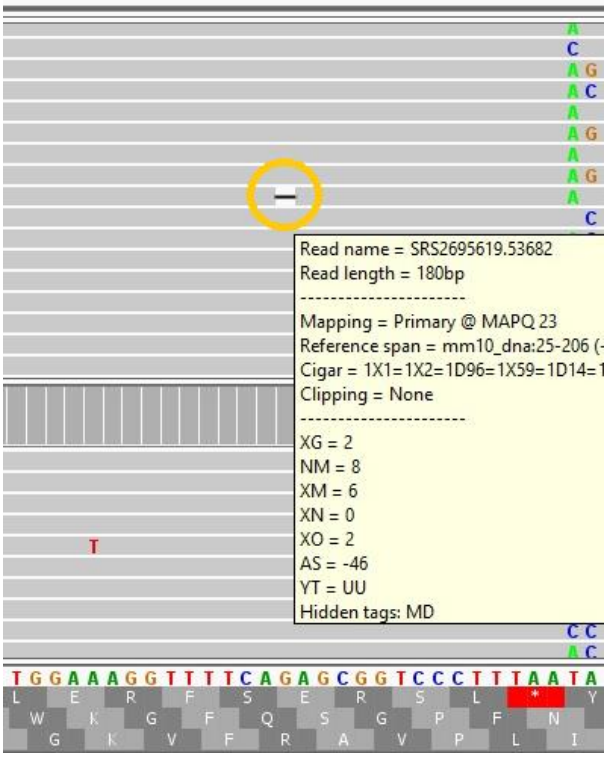
:significant variation



:point variation (shows 2 different ones in both samples)



:indel



Gene expression – theoretical questions

a. הקושי יכול לנבוע ממגוון סיבות – שימוש במקטעים גדולים מדי שאינם מאפשרים זיהוי אקסונים נפרדים, או לחלופין שכיחות נמוכה של איזופורמים כך שלא מגיעים לכמות הדרושה בשביל שתזוהה בשיטה. כמו כן, דמיון גבוה בין הרצפים מאחר ומדובר באותו מקטע דנ"א שמשועתק, כך שזיהוי באמצעות אינדיקטורים על הגן פחות יעיל.
רוב הבעיות הנ"ל מאפיינות בעיקר את שיטת ה-microarray, שדורש כמות גדולה יחסית של RNA וכן מסתמך על הצמדת גלאים וכן על ידע מוקדם של הרצפים.
בשיטת ה-RNA-seq יש הרבה פחות קושי בעניין זה נעשה גם פירוק למקטעים קטנים המאפשרים לזהות חיבורים בין אקסונים, וגם מקטעים גדולים יותר או ביצוע הריצוף בצורה של pair-end שנותנים מידע על חיבור של בין מס' אקסונים, וכן נדרשת כמות קטנה יותר של RNA ואין שימוש בגלאים כי אם בריצוף. אך הקושי העיקרי הוא עיבוד המידע שנוצר בתהליך שכן ב-RNA-seq מתקבלת כמות גדולה מאוד של מידע.

infograph שמסכם את העניין, נלקח מכאן

Microarray	RNA-seq
Advantages • Well-defined protocols for hybridization • Well-defined analysis pipelines • Standardised approaches for data submission • Relatively low cost	**Advantages** • Not reliant on previous sequence information • High dynamic range (no saturation) • Direct sequence alignment, no hybridization • Alternative splicing detected if aligned to genome • Paralogous genes can be defined • Can be used for SNP identification
Disadvantages • Analysis only for pre-defined sequences • Dynamic range limited by scanner • Relies on hybridisation • Hybridisation potentially non-specific • Might not give paralogue information • High variance for low expressed genes • Will generally not identify splice variants	**Disadvantages** • Protocols still not fully optimised • High cost (but continually reducing) • Requires high power computing facilities • High set-up costs if carried out in house • Complex analysis of splice variants • Analysis can be complex if paralogues present

b. ניתן לפתור את הבעיות בשיטת ה-microarray באמצעות שיטה המכונה exon junction array המשתמשת בגלאים הנקשרים לאיזורים הצפויים לשמש כאיזורי חיתוך בקצוות האקסונים, כלומר מאפשרת לסמן את גבולות האקסונים וע"י כך לאפיין ביטוי איזומורפים של הגן. כמו כן בשיטה הנקראת exon array הקושרת גלאים לאקסונים עצמם ניתן לזהות איזופורמים שונים.
עבור שיטת ה-RNA-seq ניתן ליצור אלגוריתמים מהירים למציאה וספירה של איזופורמים, למשל כפי שמיושם ע"י cufflinks המשתמשת באלגוריתמים הסתברותיים למטרה זו.