# תרגיל בית 3 – Unix shell commands, CIGAR, QC data

הנחיות כלליות:

* בבקשה נסו לפתור את כל השאלות המופיעות בתרגיל.
* אתם מתבקשים לרשום פתרון מלא ולא רק פתרון סופי. המשמעות: שניתן לשחזר את הפתרון שלכם. ניתן לפרט בקווים כלליים באופן דומה לתרשים ההנחיות.
* בשלבים הדורשים להגיע לתצוגה ויזואלית (ה-UCSC), מאוד ממליץ לבצע צילומי מסך (ניתן להשתמש ב-windows snipping tool, <windows key>+<shift>+<S>).
* ניתן כמובן להיעזר אחד בשני, אך הכתיבה וצילומי המסך עצמאיים לכל זוג
* מאוד עדיף לכתוב בוורד אך במידה ואתם בוחרים לכתוב בכתב יד: אין בעיה (כל עוד הוא קריא). תצרפו שני קבצים: אחד צילום כתב היד והשני צילומי המסך.
* בונוס: לציון התרגיל הנוכחי במידה וטעיתם במקומות אחרים :) לא חייבים לבצע אותו.
* פתרון מלא: תשובה ספיציפית, דרך הפתרון והסבר לפתרון.

הנחיות לקוד:

* כל הפקודות בתרגיל יהיו רשומות יחד בקובץ sh אחד. אם השתמשתם בפונקציות או סקריפטי עזר – אנא כללו אותם גם כן.
* כל קובץ יהיה בפורמט: <ID>\_part>.sh, לדוגמה עבור הסטודנט עם ת.ז 123456789, במידה והפרדתם לחלקים 1.א תהיה: 123456789\_1a.sh.
* וודאו מראש שהקוד שלכם ניתן להרצה! נסו להריץ אותו אצלכם בשרת, האם הוא עובד?
* שימו לב: הבדיקה תתבצע במערכת ubunto18.04 או במערכת CentoOS7. לא תהיה בדיקה בmac! סטודנטים המעוניינים לפתור בmac יצטרכו לבדוק שהקוד אכן רץ במערכות לעיל.
* הציון מורכב משני חלקים: בדיקה אוטומטית וידנית. בבדיקה האוטומטית הקוד שלכם יורץ וייבדק האם הוא מחזיר את התוצאה הסופית הדרושה. בבדיקה הידנית הקוד ייבדק מבחינה אלגוריתמית ותקינות כללית.
* אינכם נדרשים ליעילות, אך הקוד צריך לעבוד בפרק זמן סביר.
* פורמט הקוד המתקבל הינו פסאודו קוד, סקריפטי unix bash השונים שנלמדו. אין לכתוב סקריפט python לפתרון התרגילים =]
* שימו לב מה הפלט המדויק שצריך לצאת בכל שאלה!
* המלצה: התחילו מפסאודו קוד ובנו את הקוד משם
* במידה ואתם מסתבכים: זכרו שיש לכם אפשרות לעשות man command ב-unix (לדוגמה man cat עבור הפקודה cat) ולקבל את כל הפרמטרים.

בהצלחה!

## חלק א – אנליזה בסיסית

בתרגיל זה אנו נעבוד עם dataset הומני (אנושי) שהופק באחד מהניסויים הבאים. בחרו דאטסט שמעניין אותכם ובחרו 2-3 דוגמאות מייצגות. מייצגות הכוונה חולה ובריא, או שני דברים אחרים. ניתן לחפש עצמאית דאטסט. כך למשל: מהמאגר הראשון ניתן לבחור דגימת חולה קורונה ודגימת אדם בריא. אנחנו נרצה לבצע עימוד (Alignment) לכרומוזום ספיציפי, לבצע אנליזה בסיסית ולנסות לזהות מי הדגימה החולה ומי הבריאה.

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157103>

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164873>

[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183701](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183701https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148822)

[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148822](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183701https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148822)

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131705>

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164196>

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158952>

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160690>

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92472>

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE166530>

### שלב מקדים

|  |  |
| --- | --- |
| **שם המאגר שלקחתם (GSE)** | **שם הדוגמאות שלקחתם (SRR)** |
| 157103 | SRR12544526 |
| 157103 | SRR12544527 |

### שאלה 1: הורדת הקבצים לשרת

בשלב זה אנחנו נצטרך שלושה דברים:

תשובה – הורדנו את שני קבצי הFASTQ בעזרת הסקריפט הבא:

Graphical user interface

Description automatically generated with low confidence

השתמשנו בסקריפט הנתון מהתרגול על מנת ליצור את התיקיות והמבנה. לאחר מכן השתמשנו בprefetch על מנת לייבא את קובץ ההגדרות של שני הקבצים הנבחרים – ואחר כך הורדנו מהם 500,000 רידים.

#### סעיף ב': הורדת הגנום

### על מנת להוריד את הכרומוזום הנבחר (כרומוזום X) השתמשנו בסקריפט הבא:



### בשאלה 2: בדיקת ה-FASTQ

אנא הקפידו לפרט את שורות ההרצה והפלט:

1. כמה reads (קריאות) יש בכל אחד מהקבצים? (הכוונה לפיזית כמה שורות ובהם מידע גנטי).

השתמשנו בשורת הקוד:

echo $(zcat SRR12544526\_1.fastq.gz|wc -l)/4|bc

מאחר שמבנה FASTQ הוא:

Table

Description automatically generated

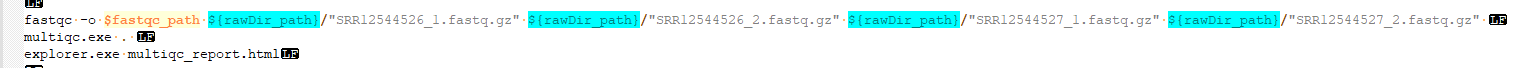
לכל קובץ, מספר הרידים בקובץ הוא מספר השורות חלקי 4. מההורדה עם הפקודה קיבלנו 2 קבצים לכל דוגמא שנבחרה – ולכל אחד מהם K500 רידים, ויש K500 שורות של רצפים כאשר לכל שורה יש רצף (ריד).

A picture containing text

Description automatically generated

1. חלצו את הקבצים המכווצים במידת הצורך.
2. הריצו fastqc על הקבצים. מה איכות הקבצים? אנא צלמו מסך

הרצנו את הפקודות הבאות:



וקיבלנו את התוצאות הבאות :

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

איכות:

Graphical user interface, application, table

Description automatically generated

A picture containing table

Description automatically generated

התפלגות בסיסים עבור כל קובץ:

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated

רמות דופליקטים:

Chart, line chart

Description automatically generated

אורך הרידים:

Graphical user interface, text, application, chat or text message

Description automatically generated

אדפטורים:

Graphical user interface, text, application

Description automatically generated

מהתוצאות נובע כי רמות הדופליקטים נמוכות, איכות הריצוף תקינה, כל הרידים באורך זהה של 51 בסיסים ורמות האדפטורים זניחות. מכאן שניתן להמשיך לריצוף של הדוגמאות ללא עיבוד לקבצים.

הדרכה: זכרו איך קוראים קובץ מכווץ ומחלצים אותו (היינו אך). אין חובה כלל להריץ multiqc – אם כי זה אכן יקל מעט עליכם ☺

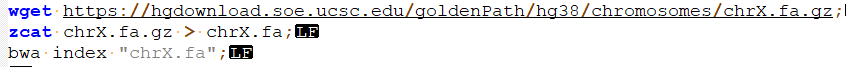
### שאלה 3: ריצוף בסיסי

בשלב זה אנו נרצה לעמד (Align) את הטרנסקריפט שלנו בחזרה לגנום.

אנו נשתמש ב-HISAT2 או BWA, שהן תכנת Alignment טובה ויעילה עבור NGS. שימו לב שהתקנתם את שניהם בשרת, אך לבחירתכם במה להשתמש. שניהם יביאו לאותה תוצאה (ניתן כמובן להשתמש גם ב-STAR). בצעו את השלבים הבאים:

1. יצרו אינדקס מקובץ ה-Fasta.

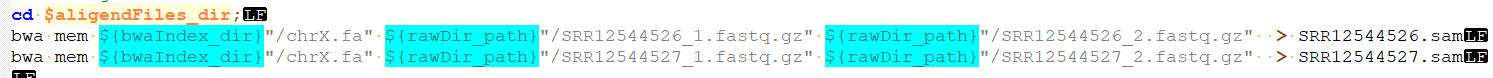
יצרנו אינדקס בעזרת הפקודות הבאה :



(חלק מהפקודות שימשו להורדה של הכרומוזום שבו השתמשנו)

1. מפו את כל אחד מהקבצים לגנום – העזרו באינדקס שיצרתם בסעיף הקודם. (הפלט צריך להיות בפורמט SAM).

עשינו את המיפוי בעזרת BWA בעזרת הפקודות הבאות:



1. כמה קריאות (Reads) התמפו באופן ייחודי לגנום? (Uniquely mapped reads)?

עבור הדוגמא SRR12544526 קיבלנו:

Graphical user interface

Description automatically generated with low confidence

נתון שיש לנו 0 רידים דופליקטים,מכאן שכל הרידים שהתמפו לגנום הינם ייחודיים ומכאן שיש 113755

עבור הדוגמא SRR12544527 קיבלנו:

Text

Description automatically generated

בדומה למקרה הקודם יש לנו 0 דופליקטים ולכן כל 129463 הרידים שמופו אצלנו הם ייחודיים.

1. מה היתה איכות המיפוי הכללית? (Overall alignment rate)

מ2 התמונות מסעיף 3 נובע כי עבור דוגמא SRR12544527 מופו 12.95% מהרידים ועבור הדוגמא SRR12544526 מופו 11.38% מהרידים.

הדרכה: לסעיף 1 ו-2 העזרו ב-HISAT2 או BWA וב-manual שלו. לסעיפים 3-4 העזרו samtools flagstat. בבקשה הקפידו לפרט כל שלב!

### שאלה 4: שימוש ב-Samtools

1. המירו את קבצי ה-SAM לקבצי BAM

השתמשנו בפקודות:

Graphical user interface

Description automatically generated with low confidence

1. למה אנחנו נעדיף ככלל לעבוד עם קבצי BAM ולא עם SAM? אנא תנו סיבה אחת עיקרית.

קבצי BAM מכילים את אותו המידע של SAM אך תופסים משמעותית פחות מקום ולכן קל ונוח יותר לעבוד איתם.

1. כמה קריאות יש בקובץ ה-BAM שיצרתם? האם הכמות זהה לכמות שקיבלתם בשאלה 2?

כמות הרידים בקבצי הBAM היא כמיליון לכל קובץ (תמונות מצורפות). בשאלה 2 קיבלנו שיש לנו לכל דוגמא 2 קבצים ולכל אחד מהם K500 רידים, כלומר סהכ מיליון רידים לכל דוגמא. יצר מצב שמספר הרידים של קבצי הBAM זהה למספר הרידים הכולל של כל דוגמא בהתאמה.

Text

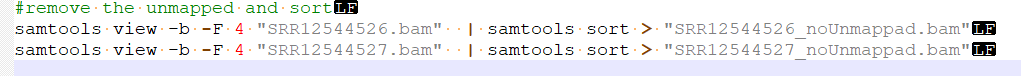
Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

1. סננו אל קובץ חדש רק את הקריאות שהתמפו (Mapped reads).
2. מיינו את הקובץ BAM שיצרתם בסעיף 4.

את סעיפים 4 ו5 עשינו ביחד בעזאת הפקודה הזאת



* הסימון 4 F- מוריד את כל אלו שמסומנים בביט ה4 שמסמן שהריד לא מופה.

הדרכה: לסעיף 1 ו-3 העזרו ב-Samtools view. לסעיף 5 העזרו ב-samtools sort. סעיף 4 הוא מעט מאתגר, חשבו כיצד אתם יכולים לגשת בעזרת Samtools רק אל הקריאות שהתמפו? רמז, יש דרך מובנית.

### שאלה 5: Variant calling

בשלב זה נרצה להמיר את הקובץ BAM שלנו ל-mpileup ואת הקובץ BCF להמיר לפורמט VCF. ישנם מספר דרכים לביצוע השלב הבא

1. הריצו Variant calling על קבצי הBAM: השתמשו בsamtools mpileup ו-BCF call על מנת להמיר את הקובץ מ-BCF ל-VCF (מומלץ לבצע זאת בעזרת פקודה אחת).

התשמשנו בפקודות הבאות עבור 2 הקבצים שלנו:



A picture containing text

Description automatically generated

1. הביטו בשני קבצי ה-VCF המתקבלים, תארו שני שינויים משמעותיים אשר אתם רואים בין הקבצים. הסבירו איך בחרתם אותם

השתמשנו בפקודות:

A screenshot of a computer

Description automatically generated with medium confidence

וקיבלנו עבור SRR12544526.vcf:

Text

Description automatically generated

ועבור SRR12544527.vcf:

Text

Description automatically generated

שני ההבדלים הראשונים שאנחנו רואים בין הקבצים הם ראשית מספר הרשומות של כל קובץ (num of records) כאשר בדוגמא הראשונה יש לנו 14,758 שורות ובשניה יש 12,945. הבדל נוסף שאנחנו רואים הוא היות מספר הSNPים בכל דוגמא כאשר בדוגמא הראשונה יש 14,269 ובשניה יש 12,529. ניתן לראות הבדלים גם במספר האתרים בעלי מספר אללים שונים ומספר המוטציות מסוג indel.

הדרכה: בסעיף 1 מאוד מומלץ לבצע זאת בפקודה אחת. ניתן להשתמש בbcf mpileup במקום ב-samtools mpileup. בסעיף 2 יש הרבה דרכים כיצד לבצע את ההשוואה – ניתן לחפש הבדלים פיזיים בין הדוגמאות, ניתן להיעזר ב-bcf stats ועוד. הסתפקו רק בשני הבדלים. אמנם, כמובן: כל המרבה הרי זה משובח.

## חלק ב – אנליזה מתקדמת ☺

בשלב זה אנו נבצע שני אנליזות אפשריות על הנתונים (6-7, ניתן להיות מקוריים ולעשות דברים אחרים, מי שיבצע שלושה זכאי לבונוס ☺).

### שלב מקדים משותף: סינון מוטציות A-G ו-C-U

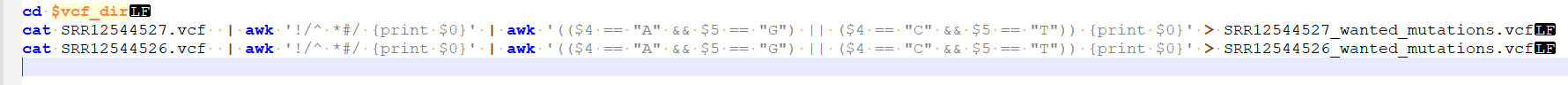
לא בחרנו מוטציות אלו סתם. מוטציות A-G מקוטלזות ע"י ADAR ו-C-U (C-T) מקוטלזות ע"י משפחת Apobec. נניח שנרצה להתמקד רק במוטציות האלו. ישנם מספר אפשרויות לרשותכם. אנו נעשה את זה על קבצי ה-VCF של שתי הדוגמאות:

* **סינון ב-Excel** – פשטני אבל אפשרי. שימו לב שעמודת ה-ref יש A (או C) ובעמודת ה-Alt יש G (או T).
* **סינון ב-AWK** – לדעתי הרבה יותר מהיר. שימו לב שאתם בוחרים את העמודות הנכונות. . לנוחותכם שורת AWK שעוזרת ;) :

cat file.vcf | awk '!/^ \*#/ {print $0}' | awk '($4 == "A" && $5 == "G") {print $0}'

בסוף השלב תקבלו את נקודות המוטציה.

על מנת לסנן את המוטציות הרלוונטיות השתמשנו בפקודות הבאות עבור שני הקבצים:



### שאלה 6: webANNOVAR

1. הפעילו את הכלי של ה-ANNOVAR – סננו את האזורים שהינם מגיעים מ-SNP . שימו לב ש-wAnnovar מכיל מגבלת גודל אז צריך לסנן ☺ (חושב שעד 100Mb).
2. האם קיבלתם תוצאות? הציגו דוגמה (ייתכן ואין וזה בסדר לחלוטין).

הכנסנו את הקבצים שקיבלנו בשאלה הקודמת ל wAnnovar ושמרנו על ההגדרות הבאות:

Graphical user interface, application

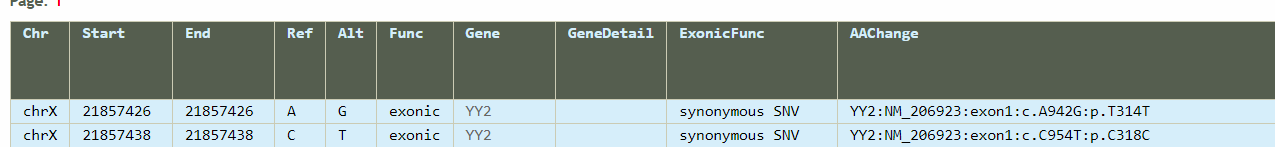
Description automatically generated

לשתי הדוגמאות בחרנו באופציה של **exome summary results** והורדנו את התוצאות לCSV.

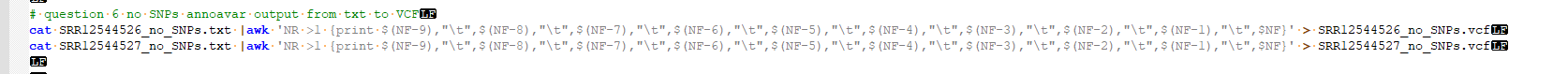
Graphical user interface, application

Description automatically generated

דוגמאות לתוצאות משותפות בין שתי הדוגמאות:



2 הדוגמאות אינן SNP, מהאקסום שקיבלנו לדוגמא SRR12544527 יש לנו 63 תוצאות בעוד שמהאקסום של הדוגמא SRR12544526 יש לנו 68 תוצאות, כאשר משני האקסומים 15 מהתוצאות קיימות במאגר dbsnp – כלומר הרוב המוחלט של התוצאות אינו SNPים. לבסוף כאשר הורדנו את הקובץ בפורמט אקסל, מחקנו ידנית את כל התוצאות שהן SNPs ושמרנו את התוצאות בתור קובץ טקסט. לבסוף בעזרת הפקודות הבאות המרנו את הקבצים לVCF :



קובץ הVCF שקיבלנו הוא

SRR12544526

A picture containing text

Description automatically generated

SRR12544527:

A picture containing diagram

Description automatically generated

הקפידו להציג את ההרצה. המשיכו מכאן לחלק ד' ☺

### שאלה 7: נקודות וגנים ידועים

סביר להניח שקיים מידע עבר על גנים שנמצאים באסוציאציה עם המחלה שאותה אתם חוקרים, נוכל לשלוף את הנקודות ולחפש אותם.

1. הכנסו אל מאגר הgenecards ושלפו משם גנים הנמצאים באסוציאציה עם המחלה (לפחות 5)

הדאטאסט שבחרנו מתמקד במחלת הקורונה. הלכנו למאגר ה <https://www.genecards.org/> , חיפשנו בשורת החיפוש את הביטוי covid 19 ביחד עם הכרומוזום שבחרנו מוקדם יותר (chrX) ובחרנו את הגנים הבאים:

* ACE2
* TLR7
* DDX3X
* G6PD
* GRIPAP1

שורת החיפוש הייתה



הסיבה שבחרנו את הגנים שבחרנו היא שראינו אותם בטבלה שקיבלנו בannovar, על מנת שנקבל תוצאות לחיתוך בהמשך האנליזה.

1. הכניסו את הגנים שמצאתם אל ה-table browser והורידו את טבלת הגנים הללו בפורמט BED.

האזורים של הגנים האלה בגנום הם:

ACE2 - chrX:15,557,039-15,607,236

TLR7 - chrX:12,867,072-12,890,361

DDX3X - chrX:41,333,348-41,364,472  
G6PD - chrX:154,532,000-154,547,572

GRIPAP1 - chrX:48,973,723-49,002,264

נכנסנו לtable browser והזננו את האזורים בגנום תחת הסימון של (defined regions) באופן הבא :

Graphical user interface, text, application, email

Description automatically generated

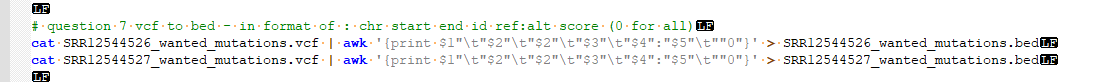
Graphical user interface, application, Word

Description automatically generated

כמו כן השתמשנו בטבלת ה refGene להורדה של קובץ הBED המכיל את כל הגנים המדוברים והורדנו את הקובץ בפורמט BED.

1. חתכו את הנקודות שקיבלתם (VCF שהמרתם לBED) כנגד הקובץ שהורדתם. האם התקבלה תוצאה?

המרנו את הקובץ שקיבלנו בסינון המקדים לקובץ BED בעזרת הפקודות הבאות:

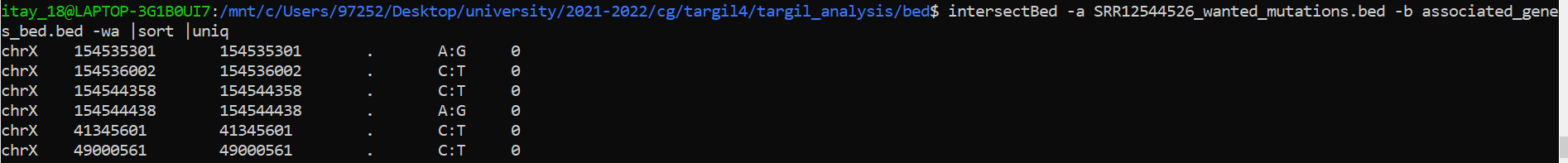


חתכנו בעזרת הפקודות הבאות:

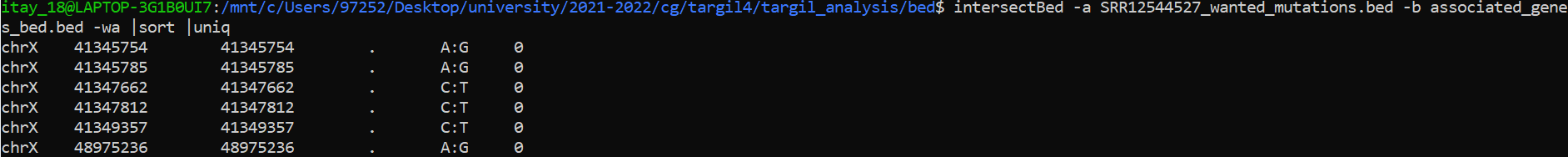
A screenshot of a computer

Description automatically generated with medium confidence

התוצאות עבור החיתוך של SRR12544526 עם הטבלה שהורדנו:



התוצאות עבור החיתוך של SRR12544527 עם הטבלה שהורדנו:



כמו שניתן לראות קיבלנו את אותו מספר תוצאות, אך הן שונות בין הדוגמאות.

1. הורידו מה-table browser את טבלת ה-SNP של הclinVar וחזרו על התהליך בסעיף 2 ו-3.

הורדנו את טבלת הclinvar לפי המיקומים של הגנים הנבחרים שלנו (תמונה של הtable browser למטה) וחתכנו בין הטבלה שהורדנו לבין קבצי הBED שקיבלנו לאחר הסינון שביצענו קודם לדוגמאות איתן אנחנו עובדים .

הפקודות בהן השתמשנו:

A picture containing text

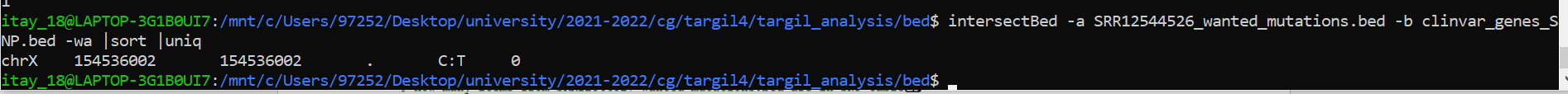
Description automatically generated

Graphical user interface, text, application

Description automatically generated

התוצאות:

עבור הדוגמא SRR12544526 קיבלנו רק את התוצאה הבאה:



עבור הדוגמא SRR12544527 לא קיבלנו אף תוצאה.

הקפידו להציג את ההרצה. המשיכו מכאן לחלק ד' ☺

## חלק ג' – הבונוס (תרגיל base editor)

מי שלא מעוניין בביצוע הבונוס – אין דבר. המשיכו לחלק ד'. שימו לב שההגשה היא ביחידים – כל אחד מגיש מוטציה ☺.

המטרה בתרגיל היא לבחור מוטציה (בעדיפות רבה זאת המצויה בגן ב-CDS) ולנתח אותה. בסופה עליכם לכתוב בקצרה בפורמט דוח מעבדה מקוצר (לכל היותר חצי עמוד-עמוד). הכתיבה והסיכום של כלל הדברים היא חלק ד' בכל מקרה.

הנחיה כללית: הקפידו לבחור נקודה שקוראת בתוך גן או שהיתה בתוך clinvar.

### שלב מקדים: התבוננו במוטציה

הריצו ב-IGV בדיקה על המוטציה, או השתמשו ב-samtools tview (יש דוגמה במצגת, מעט יותר מהיר). כיצד המוטציה נראית?

### שאלה 9 (בונוס): חקרו את הגן ב-UCSC

מצאו פרטים על הנקודה שבחרתם. השתמשו ב-UCSC genome browser (או ה-table browser) פרטים כמו:

* האם היא בגן?
* איזה גן? אם לא, ליד איזה גן? באיזה מרחק מהגן הכי קרוב?
* איפה בגן?
* כמה אקסונים/אינטרונים יש לו?
* מה אורכו?
* מה תפקידו הביולוגי?
* האם הנקודה שמורה באבולוציה?
* אם הנקודה באלמנטים חוזרניים? אם כן איזה? (חשוב, במיוחד אם היא ב-Alu)
* האם הנקודה היא ב-SNP (פחות אידיאלי).
* מה התפקידו המולקולרי של הגן?
* באיזה רקמה הגן מתבטא?

אין צורך להתייחס לכל השאלות הללו, אך עליכם להראות שבדקתם את הגן באמת. שימו לב שאתם יכולים להיעזר בתרגילים 1 ו-2 על מנת לקבל רעיונות לניתוח.

### שאלה 10 (בונוס): עריכת בסיס

כעת נניח ואנו קופצים אל העתיד (הקרוב). בידנו הדרך לשלוט בביטוי של האות או כ-A או כ-G (או בהתאמה C ו-T). אנו ננסה כעת למצוא שינויים:

* אם הנקודה באקסון: בדקו מה הח.א המקורית (והעמדה שלה) ומה הח.א האלטרנטיבית אם נשנה את הקודון. בדקו האם יש מסגרות קריאה אלטרנטיביות.
* מה סוג המוטציה שתקרה? הכוונה synonymous, non-synonymous ועוד.
* האם יש איזופורמים שונים? כיצד כל אחד מהם יושפע? הכוונה בח.א או ברצף

נסחו במילים שלכם את השינוי שעלול להיות ואת השפעתו. לא בהכרח לכל שינוי קיימת סיבה נראית, אך אם אתם מוצאים שינוי נראה וברור, זה מחזק את הטענה שלכם. שלב 6 בא כאן לשרותכם.

### שאלה 11 (הבונוס): מחקר טרנסקריפטומי עתידי

במהלך העבודה הסתכלתם על דאטה של חולים במחלה ספיציפית, ובדקתם מוטציה ספיציפית בשאלה 9. נניח והייתם רוצים לבוא לבדוק את המוטציה לעומק האם היא מקושרת למחלה אותה חקרתם.הציעו כיצד נוכל לוודא זאת.

נניח: אתם רואים מוטציות בגןMDM2 בדאטסט על הקורונה אנו רואים שהגן מתבטא בעיקר בזרם הדם. על כן: אם נאסוף דגימת דם ודגימת מטוש מנבדקים בריאים וחולים נוכל לוודא את הטענה.

## חלק ד': סיכום

בלי קשר לכמות השלבים שביצעתם, אנחנו נרצה לסכם את העבודה שביצעתם. פרטו:

* מה הדאטה שעבדתם עליו: ניתן להשתמש בהנחיה של תרגיל 4.

פסקה 1:

הדאטה שעבדנו עליו הגיע במקור מהמאגר GSE157103. המאגר עוסק במחקר שנעשה בנושא Large-scale Multi-omic Analysis of COVID-19 Severity. במהלך המחקר נבדקו 128 דגימות של תאי דם לבנים ופלסמה מחולים שאושפזו, כאשר 102 מהדגימות היו מחולים עם קורונה והשאר (26) היו מחולים ללא קורונה. לדגימות נעשו אנליזות בתהליכים של rna-seq ו mass spectrometry. בחרנו 2 דגימות (SRR12544526, SRR12544527) כאשר הראשונה נלקחה מחולה עם קורונה והשניה עם חולה ללא קורונה.

קישור למאגר: (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157103>)

לאחר מכן בהתאם להנחיות בחרנו בכרומוזום X של הגנום האנושי בגרסה hg38 (העדכנית ביותר) על מנת לבצע את האנליזה שלנו איתו – הסיבה לכך היא החשד שנקבל כמות מספיקה של תוצאות במהלך העבודה עם הכרומוזום (מה שהתברר כנכון).

* מה שלבי המחקר שביצעתם? מקבלת הדוגמאות (בקצרה, פסקה אחת. בסגנון: קיבלנו דוגמאות של חולה קורונה, מיפינו בעזרת bwa וזו כמות הרידים שהתמפו (ואחוז המיפוי), הרצנו את x והפקנו קבצי vcf שהם....).

פסקה 2:

תחילה הורדנו את הדוגמאות בעזרת הפקודות מלמעלה וקיבלנו 2 קבצים לכל דוגמא מהפקודה split-files. לאחר מכן בחננו את טיב קבצי הfastq שקיבלנו ושאכן קיבלנו את מה שרצינו להוריד (הורדנו K500 רידים לכל דוגמא). את הבחינה של איכות הקבצים בדקנו בעזרת פקודת fastqc והצגנו את התוצאות בעזרת התוכנה multiqc לשם הנוחות (במקום שיהיו 4 קבצים שונים – קיבלנו קובץ יחיד שמציג את כל התוצאות ביחד). לאחר שראינו שאיכות הקבצים שהורדנו טובה (מפורט למעלה), עשינו לדוגמאות שלנו התאמה (alignment) לכרומוזום שבחרנו בעזרת התוכנה BWA וקיבלנו קובץ SAM עבור כל דוגמא שהתחלנו איתה (סהכ 2 קבצים).עבור דוגמא SRR12544527 מופו 129463 רידים שהם 12.95% מהרידים במיפוי של הדוגמא ועבור הדוגמא SRR12544526 מופו 113755 רידים שהם 11.38% מהרידים במיפוי של הדוגמא. לאחר בחינה קצרה של קבצי הSAM שקיבלנו, המרנו אותם לקבצי BAM ומהם הוצאנו רק את הרידים שהתמפו לכרומוזום שלנו ומיינו אותם. מאותם קבצים שקיבלנו עם הרידים הרצויים, הפקנו קובץ VCF המכיל מידע על הוריאנטים שקיבלנו עד כה בהתאמה שלנו והעיבוד של תוצאותיה. אחר כך סיננו מתוך קבצי הVCF שלנו רק מוטציות A-G ו C-T בעזרת פקודת AWK, ואת הקבצים שקיבלנו מהסינון הזננו ב webANNOVAR. מהפלט של אנובר קיבלנו המון תוצאות לגנום (יותר מ40 עמודים של תוצאות לכל קובץ) ולכן הורדנו את התוצאות של האקסום על מנת לצמצם את מרחב החיפוש שלנו. קיבלנו כמה עשרות תוצאות בכל אקסום, כאשר הרוב המכריע אינו SNPים.המרנו את התוצאות לפורמט VCF בעזרת פקודות AWK. לאחר מכן חזרנו לקבצי הVCF לאחר הסינון של המוטציות הראשוני ועשינו חיתוך בינו לבין 5 גנים נבחרים שקשורים למחלה ונמצאים באותו הכרומוזום אליו עשינו את ההשוואה ובוא נמצאות כל הנקודות בקבצי הVCF. לאחר מכן חתכנו את אותם קבצים עם מאגר של סניפים של clinvar במיקומים של הגנים הנבחרים.

* תוצאות האנליזה שביצעתם
* בונוס:
  + איך בררתם את המוטציה? קצרו את השלבים שביצעתם כאן.
  + מה היא המוטציה? פרטו עליה (השלבים שביצענו כאן).
  + סכמו את דעתכם עליה לאור הממצאים.