# 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. בונוס של 5 נק' למי שיעשה את תרגיל כיתה 4 ☺
* פתרון מלא: תשובה ספיציפית, דרך הפתרון והסבר לפתרון.

הנחיות לקוד:

* כל שאלה וכל סעיף יוגשו בנפרד – ניתן לעלות את כל הקבצים כקובץ מכווץ
* כל קובץ יהיה בפורמט: <ID>\_<question><part>.sh, לדוגמה עבור הסטודנט עם ת.ז 123456789, שאלה 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': בדיקת איכות (40 נק')

## שאלה 1 – רמת איכות – FASTQC (20 נק')

במהלך ביצוע NGS אחד השלבים החשובים ביותר זוהי בדיקת האיכות של הקבצים המתקבלים. אנו לרוב נשתמש בתכנה FASTQC שעוברת על קובץ הFASTQ ובודקת אותו. על בסיס התוצאות שלה, נוכל לקבוע מה השלב הבא אותו אנו נבצע.

לפניכם מספר מקרים שהתקבלו בהרצה. לא כל המקרים קרו באותו קובץ. אנא אמרו האם קיימת בעיה ברידים (רמז: כנראה שכן) והציעו מה ניתן לעשות בהתאם. בשאלה זו אין תמיד פתרון חד משמעי – יתקבלו מספר פתרונות.

הערה: הפלטים נוצרו בעזרת התכנה MuliQC למספר קבצי FASTQC. מאוד ממליץ לעבור על המצגת/תרגול שוב, ולבדוק מה הצירים אומרים, מה צפוי להיות כאן לעומת מה שקיבלנו.

|  |  |
| --- | --- |
| א | ב |
| ציר x: המיקום בread (Position in bp).  ציר y: phread score. | ציר x: המיקום בread (Position in bp).  ציר y: אחוז הנוקלאוטידים בעמדה |
| ג | ד |
| ציר x: המיקום בread (Position in bp).  ציר y: אחוז הנוקלאוטידים בעמדה | ציר x: מספר הדופליקטים  ציר y: אחוז הרידים |
| ה |  |
| ציר x: המיקום בread (Position in bp).  ציר y: אחוז הרצפים |  |

## שאלה 2: רצפי CIGAR (10 נק')

פורמט ה-CIGAR משמש כפורמט לתיאור התאמת רצפים. בבסיס קובץ SAM, עומד הרצף, האיכות (שמקורם ב-FASTQ), הכרומוזום, נקודת ההתחלה וה-CIGAR. לפניכם מספר רצפים שונים: אנא כתבו אותם (אין צורך לצייר, רק לכתוב את המשמעות).

בחלק הזה כתבו בבקשה מהי משמעות רצף הCIGAR.

הערה: משמעות אורך הרצף הינה מהו אורך הרצף המיוצג מתחילתו ועד סופו. כך למשל: 4M מייצג אורך של 4 בגנום וברצף. 4M5D2M מייצג אורך של 6 בריד (4+2) ואילו בגנום ייצג אורך של 11 (4+5+2), היות שאנו אומרים שהיתה מחיקה של 5 מהגנום רפרנס.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **רצף הCIGAR** | **אורך הרצף בגנום** | **אורך הרצף בריד המותאם** | **משמעות הCIGAR** |
| **א** | 3S97M |  |  |  |
| **ב** | 3S30I40M10D30M |  |  |  |
| **ג** | 5H20M30I15M29D100N30M |  |  |  |

בחלק הזה כתבו בבקשה מהו הרצף הCIGAR המתקבל כאן:

|  |  |  |
| --- | --- | --- |
| א |  |  |
|  |  |  |
| ב |  |  |

## שאלה 3: קידוד נתונים (10 נק')

בתרגול דיברנו על פעמיים בהם אנו נוטים לקודד נתונים לדגל: פעם אחת במתן הרשאות לקבצים (chmod 777) ופעם שניה הדגל בתוך קובץ sam. העקרון של שניהם מבוסס על עקרון בינארי של חיבור חזקות של 2 שכל אחת בעלות מידע ייחודי. לפניכם מספר דוגמאות של קידודים – אנא תרגמו את משמעות הקידוד.

הערה: כמובן שיש לאמר ממה הוא מורכב.

דגלי SAM: בונוס: האם לדעתכם התוצאה תקינה? (1 נק' על כל סעיף)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | קידוד | היה aligment (התאמה לגנום?) | אם כן: מי הותאם? איפה (אם ידוע)? מי מהpair? | בונוס: התוצאה תקינה? |
| א | 67 |  |  |  |
| ב | 99 |  |  |  |
| ג | 4 |  |  |  |
| ד | 147 |  |  |  |
| ה | 50 |  |  |  |

הרשאות: בונוס נק' על כל סעיף אם יש רעיון מקורי למה שבכלל נחשוב להשתמש בזה =]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | קידוד | הרשאות משתמש | הרשאות קבוצה | הרשאות אחרים | שימוש? (בונוס) |
| א | 744 |  |  |  |  |
| ב | 444 |  |  |  |  |
| ג | 421 |  |  |  |  |
| ד | 404 |  |  |  |  |
| ה | 763 |  |  |  |  |

## חלק ב' – Shell (60 נק')

### שאלה 4: קוד בסיסי בShell

בחלק זה המטרה לכתוב מספר סקריפטים בסיסיים ב-shell. בכל חלק ניתן להשתמש בכל פונקציית shell שהיא (או שניתן לזמן דרך shell).

### סעיף 1 (5 נק')

כתבו קוד המקבל רדיוס מהמשתמש ומחשב שטח מעגל (כערך pi ניתן להשתמש בקירוב 3.14). (מותר להניח שיוזן מספר).

|  |  |  |
| --- | --- | --- |
| **קלט** | **פלט** | **הנחות** |
| רדיוס (מספר) | שטח מעגל (מספר, יכול להיות עשרוני!) | יוזן מספר תקין  ניתן להשתמש בקירוב pi=3.14 |

### סעיף 2 (5 נק')

כתבו קוד המקבל נתיב לקובץ מהמשתמש (מותר להניח שהקובץ קיים) מפריד (separator) ומפריד אחר. המטרה שלכם זה לעבור על הקובץ ולהחליף כל מפריד במפריד החדש. אינכם צריכים לבדוק תקינות קלט. הקובץ ייבדק על קובץ csv סטנדרטי (קובץ מופרד ב’,’ ויומר ל-‘\t’) ולהפך. הבהרה: עליכם להחליף כל אירוע בשורה.

|  |  |  |
| --- | --- | --- |
| **קלט** | **פלט** | **הנחות** |
| נתיב לקובץ  מפריד א'  מפריד ב' | הדפסת כלל השורות בקובץ כאשר: מפריד א' מוחלף במפריד ב' | קלט תקין  הקובץ קיים |

### סעיף 3 (5 נק')

בעקבות עבודה אינטנסיבית של קבוצת המחקר בתקופת הקורונה, השרתים והכוננים של הקבוצה בקריסה. אתם מוניתם על ידי האדמין למצוא את המכשלה – קבצים בעלי סיומת ספיציפית הגדולים מגודל מסויים.

כתבו קוד המוצא את כל הקבצים (files) בעלי סיומת x ומעל גודל y שנמצאים בתיקיה.

|  |  |  |
| --- | --- | --- |
| **קלט** | **פלט** | **הנחות** |
| נתיב לתיקיה  גודל (בkb)  סיומת ספיציפית | כל הקבצים שעוברים את הגודל הנתון ובעלי הסיומת הנתונה | קלט תקין |

### בונוס (5 נק'): סעיף 4 (המשך לסעיף 3)

על מנת לחסוך שיחת נזיפה מכלל הקבוצה, האדמין מעוניין לזמן רק את האדם הספיציפי שאשם בכך. לצורך כך הוא ביקש מכם שתחשבו כמה כל חוקר צורך מהזכרון. הנחתו הינה "אין חף מפשע, אך יש שחפים יותר". הרחיבו את סעיף 3 כך שתוכלו לסכום כמה כל משתמש צורך מהזכרון.

|  |  |  |
| --- | --- | --- |
| **קלט** | **פלט** | **הנחות** |
| נתיב לתיקיה  גודל (בkb)  סיומת ספיציפית | כמה כל משתמש צורך מהזכרון | קלט תקין |

## שאלה 5: AWK (25 נק')

הכלי AWK הינו מאפשר לבצע הרצות מאוד מהירות על טבלאות ומסדי נתונים. הוא עובר שורה שורה בקובץ ומפרסר אותו לפי ההוראות שלנו. היתרון המשמעותי של הכלי הינו השילוב של מהירות הפעולה של Unix (בייחוד לקבצים גדולים) והיכולת להשתלב לגמרי בשורת פקודה. בכל סעיף תקבלו טבלה לדוגמה, ותתצטרכו לכתוב סקריפט קצר ב-AWK שיבצע את המשימה. לנוחותכם, הטבלה המתקבלת תמוספר מבחינת עמודות ושורות. שימו לב שהמספור הוא רק מטעמי נוחות ואינו מופיע הלכה למעשה בקובץ. ראו נספח כדוגמה.

בכל שאלה אנו נעבוד עם טבלה שונה (והמידע ייבדק עליה). אנא שימו לב שהורדתם את הטבלה מה-Table browser.

### סעיף א' (5 נק')

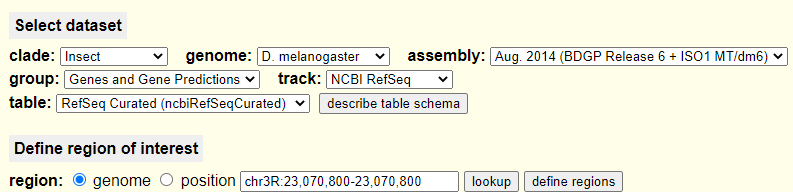
להלן דוגמה לטבלת BED:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **1** | **2** | **3** | **4** | **5** | **6** |
| **1** | chr3R | 613277 | 613278 | Myo81F | intronic | + |
| **2** | chr3L | 27422103 | 27422104 | Dbp80 | nonsyn | + |
| **3** | chr3R | 610517 | 610518 | Myo81F | intronic | + |

הדפיסו את כל השורות בקובץ (כמו שהם, $0) אשר מקיימים:

* אורך הגן הוא לפחות 1000
* בגדיל החיובי
* בכרומוזום “chr3R”.

בשאלה זו נעבוד עם טבלת כל הגנים בדרוזופילה – אלו הפרמטרים לטבלה.



|  |  |  |
| --- | --- | --- |
| **קלט** | **פלט** | **הנחות** |
| נתיב לקובץ גנים בפורמט BED | שורות המקיימות את התנאים לעיל | קלט תקין של דרוזופילה |

### סעיף ב' (10 נק')

כתבו סקריפט הממיר פורמט GTF (ראו דוגמה מטה) לפורמט BED. הפורמט צריך להיות פורמט BED תקני (6 עמודות), כאשר:

* עמודה מספר 4: שם הגן (מה שרשום ב-Gene ID).
* עמודה מספר 5: שם החלק בגן (המידע בעמודה 3).

המפריד בין השורות צריך להיות tab (‘\t’).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| # | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **1** | chr11 | ncbiRefSeq | stop\_codon | 113410727 | 113410729 | 0 | - | . | gene\_id"NM\_016574.3";transcript\_id "NM\_016574.3"; |
| **2** | chr11 | ncbiRefSeq | CDS | 113410730 | 113410920 | 0 | - | 2 | gene\_id"NM\_016574.3";transcript\_id "NM\_016574.3"; |
| **3** | chr11 | ncbiRefSeq | exon | 113409595 | 113410920 | 0 | - | . | gene\_id"NM\_016574.3";transcript\_id "NM\_016574.3"; |
| **4** | chr11 | ncbiRefSeq | CDS | 113412556 | 113412883 | 0 | - | 0 | gene\_id"NM\_016574.3";transcript\_id "NM\_016574.3"; |

הדרכה: שימו לב בבקשה שבעוד רוב ההמרות הינן "פשוטות" ורק דורשות החלפת מיקום, על מנת לחלץ את שם הגן (gene\_id מבחינתנו) נאלץ להשתמש ב-split. שימו לב שאתם זוכרים מה ההבדל בין GTF לBED.

את טבלת הגנים ניתן להוריד מה-table browser. הבדיקה תהיה על טבלת כל הגנים של דרוזופילה (בפורמט GTF).

|  |  |  |
| --- | --- | --- |
| **קלט** | **פלט** | **הנחות** |
| קובץ בפורמט GTF | קובץ בפורמט BED | קלט תקין |

### סעיף ג' (10 נק')

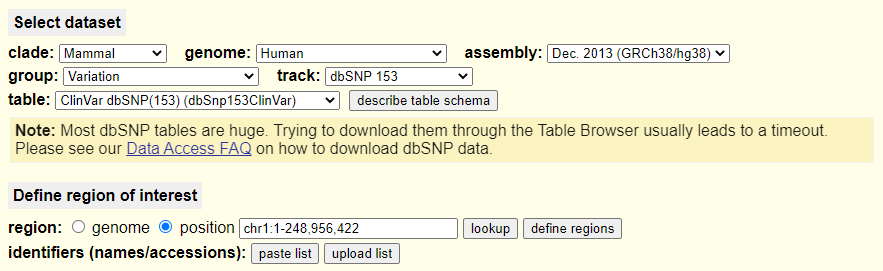
צרו קטע קוד המתרגם פורמט VCF לפורמט BED. הפורמט צריך להיות בפורמט BED תקני (6 עמודות).

בשאלה זו נעבוד עם הוריאנטים של ה-clinvar.

שימו לב:

* עמודה 4: שם הSNP ב-clinvar.
* עמודה 5: עמודת הref ועמודת הalt בפורמט הבא: <ref>:<alt>, כדוגמאת: A:G, או C,T:A.

שימו לב: דלגו על שורות המתחילות בסימן ‘#’.



|  |  |  |
| --- | --- | --- |
| **קלט** | **פלט** | **הנחות** |
| קובץ בפורמט VCF | קובץ בפורמט BED (להתעלם משורות הערה) | קלט תקין |

## שאלה 6 – יצירת מילון (20 נק')

לעיתים קרובות אנו מורידים טבלאות מה-table browser והן מכילות שם קוד שמאפשר גישה ייחודית אליהן (קוד refseq שמתחיל ב-NM\_, גנים של gene bank המתחילים ב-GB ועוד). לחילופין: אנו עושים שימוש בשמות הללו על מנת לקבל תוצאות ייחודיות, אך לא תמיד אנו מבינים מיד מהו הגן או האזור המדובר.

בתרגיל זה אנו נוריד שני טבלאות מה-UCSC:

* גנים בפורמט BED
* מילון (נסביר את שלבי היצירה שלו).

ידוע לנו בהגדרה ששם הגן מצוי בעמודה הרביעית. מטרתנו היא לקחת כל שורה מקובץ הגנים, לחפש אם היא מצויה בקובץ המילון ואם כן: להוסיף את זה בעמודה מספר 5 (עמודת ה-Score). אחרת: לרשום "NA".

|  |  |  |
| --- | --- | --- |
| **קלט** | **פלט** | **הנחות** |
| קובץ בפורמט BED  מילון | עבור כל השורות, להחליף את עמודה 5 בשם הגן מהמילון במידה וקיים. | קלט תקין |

הקבצים (מודגמים עבור מספר שורות, אבל ייבדקו על כרומוזום שלם).

|  |  |
| --- | --- |
| **קובץ הגנים** | **המילון** |
|  |  |
|  |  |
|  |  |

הדרכה: אינכם נדרשים לפתרון יעיל בשאלה – חשבו כיצד ניתן לממש את זה בפתרון נאיבי והסתפקו בו.

# נספחים

**הושאר ריק על מנת להדגיש: השאר זה נספחים ולא תרגיל נוסף.**

### נספח 1: סקריפט UNIX

כך לדוגמה:

|  |  |  |
| --- | --- | --- |
| **Pseudo-code** | **Code** |  |
| 1. הדפס את כל הקבצים בתיקיה הנוכחית למסך (קבצים ותיקיות) 2. כל עוד יש אברים ברשימה, קרא אותם אחד אחד למשתנה name ובצע:    1. הדפס /(path to current directory)/(file name) | **ls** **|** **while** **read** -r name**;** **do** **echo** **$(pwd)/$name;** **done;** | **One line** |
| **ls** **|** **while** **read** -r name**;**  **do**  **echo** **$(pwd)/$name;**  **done;** | **Script** |
| מטרת הסקריפט: דרך להדפיס את הנתיב האבסולוטי לכל הקבצים בתיקיה. | | **Misson** |

### נספח 2: סקריפט AWK

נדגים סקריפט AWK שיקבל קובץ BED (לרוב הסקריפט יהיה פשוט יותר מזה) ויפרסר את העשר שורות הראשונות בקובץ, להלן הטבלה:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **1** | **2** | **3** | **4** | **5** | **6** |
| **1** | chr3R | 613277 | 613278 | Myo81F | intronic | + |
| **2** | chr3L | 27422103 | 27422104 | Dbp80 | nonsyn | + |
| **3** | chr3R | 610517 | 610518 | Myo81F | intronic | + |
| **4** | chr3L | 23053958 | 23053959 | nrm-RpL10 | intergenic | + |
| **5** | chr2R | 4051448 | 4051449 | Haspin-CG17683 | intergenic | + |
| **6** | chr3R | 649337 | 649338 | Myo81F | intronic | + |
| **7** | chr3R | 3455951 | 3455952 | spok-Parp | intergenic | - |

נכתוב סקריפט להדפסה של חלק מהמידע בקובץ.שימו לב שהצבע הוא מלאכותי ונועד רק כדי להקל על הקריאה ולהבדיל בין פונקציה, טקסט ומשתנה.

|  |  |  |
| --- | --- | --- |
| **Pseudo-code** | **Script** |  |
| 1. הפעל את awk על המשתנה bed\_file    1. הדפס בתחילת הסקריפט את ההודעה:       1. קרא את הקובץ. אם עבור השורה מתקיים התנאי: מספר השורה קטן, שווה ל-4, בצע:       2. הגדר משתנה chr, ושמור בו את עמודה 1.       3. הגדר משתנה start, ושמור בו את עמודה 2.       4. הגדר משתנה end, ושמור בו את עמודה 3.       5. הגדר משתנה name, ושמור בו את עמודה 4.       6. הגדר משתנה score, ושמור בו את עמודה 5.       7. הגדר משתנה strand, ושמור בו את עמודה 6       8. הדפס: הגן name, אשר מצוי בגדיל strand מ start עד end אורכו (תוצאת החישוב של end-start). השורה המקורית היתה (שורת הקלט) | **awk** 'BEGIN {**print** "Analysis for BED file"}  NR <= 4 {  **chr**=**$1**;  **start**=**$2**;  **end**=**$3**;  **name**=**$4**;  **score**=**$5**  **strand**=**$6**;  **print** "The gene " **$4** ", which located at the " **$6** " strand from " **$2** " to " **$3** ", its length: " **start - end** "\nThe original line was" **$0**;'  **}** **$bed\_file** |  |

כמובן שניתן לגשת למשתנים גם דרך מספר העמודה, וגם דרך משתנה כפי שהדגמנו. זה היינו הך.