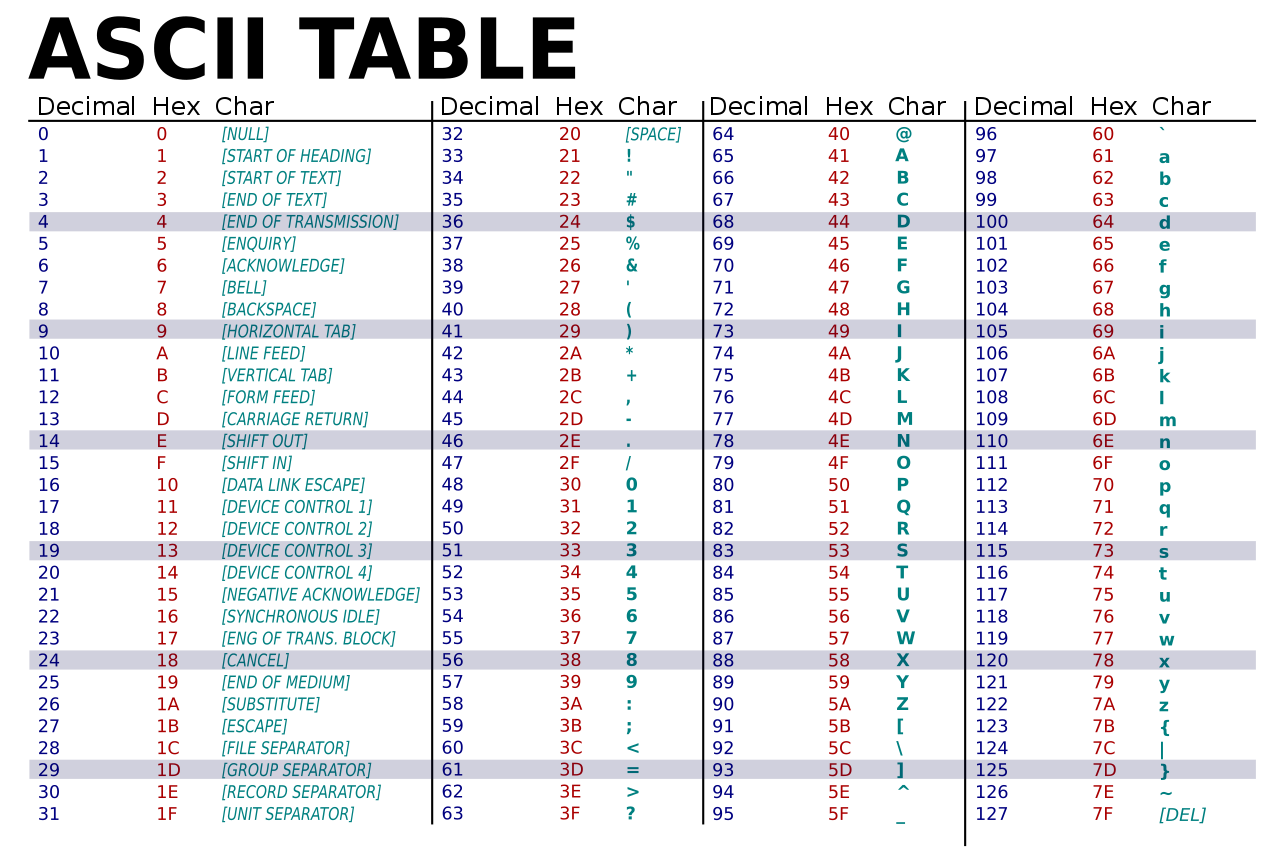
# תרגיל בית 2 – Table browser & NGS

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 3 ☺
* פתרון מלא: תשובה ספיציפית, דרך הפתרון והסבר לפתרון.

בהצלחה!



## שאלה 1 – רמת איכות – FASTQ

במהלך ביצוע NGS אנו רוצים לבצע בדיקת איכות לכל בסיס בפורמט FASTQ. בהנחה וההסתברות לטעות הינה הבאות מטה, חשבו מה האיכות של הרצף, ואיזה אות אסקי אנו ניתן לו? שימו לב: הנכם מתבקשים לכתוב את הסיכוי לטעות, או את ציון הPhred score.

הערה: ניתן ורצוי להיעזר בטבלת ascii המצורפת כנספח ובמשוואות:|Qphred=-10\*log10(mistake), (Q האיכות, P הסיכוי לטעות) ובנוסחה charASCII = Qphred +33. רמז: אם אתם מקבלים מספרים עשרוניים: עגלו לשלם.

דוגמה: הסיכוי לטעות הוא 79% = 0.79. נציב במשוואה: Qphred=-10\*log10(mistake):

Qphred=-10\*log10(0.79) 🡪 Qphred=1.023 🡪 Qphred= 1

כעת נציב במשוואה השניה:

charASCII = 1 +33 🡪 34 🡪 ‘"’

לפיכך, עבור הסיכוי 0.79, קיבלנו את הציון 34 ואת התיו ".

עבור התיו ‘+’, נציב במשוואה charASCII = Qphred +33: (נעזר בטבלת ASCII)

+ = Qphred +33 🡪 43 = Qphred +33 🡪 = Qphred =10

כעת נציב במשוואה Qphred=-10\*log10(mistake):

10=-10\*log10(mistake) 🡪 -1 = log10(mistake) 🡪 mistake =0.1 = 10%

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **טעות ב(%)** | **הסיכוי לטעות** | **האיכות של בסיס** | **התיו** | **חישוב** |
| 1 | 100% |  |  |  |  |
| 2 | 42% |  |  |  |  |
| 3 | 7% |  |  |  |  |
| 4 | 0.001% |  |  |  |  |
| 5 | 0.025% |  |  |  |  |
| 6 |  |  |  | < |  |
| 7 |  |  |  | H |  |
| 8 |  |  |  | ) |  |
| 9 |  |  |  | / |  |
| 10 |  |  |  | ^ |  |

## USCS table browse

הנחיות:

* עיקר העבודה היא ב-UCSC - <https://genome.ucsc.edu/index.html>
* אנו נעבוד הפעם עם Human genome Dec. 2013 (GRCh38/hg38).
* אנא הקפידו לרשום את השלבים שביצעתם על מנת להשיג את מטרתכם. מומלץ לצרף צילום מסך של התוצאה (במידת הצורך).
* אנא הקפידו על כתיבת דרך מלאה והסבר לבחירה שלכם. לרוב תשובה נכונה תהיה גם תשובה קצרה יחסית ☺, אז אם אתם רואים שאתם מתארכים: חשבו האם התשובה אכן נכונה.

שימו לב שאתם עובדים עם הגנום הנכון (hg19) ואתם רושמים פתרון, דרך והסבר!

### שאלה 2: פלט בסיסי

השתמשו בטבלת ה-'refGene table' (‘Gene and Gene Prediction Tracks’ group, ‘NCBI RefSeq’ track), על מנת לקבל את המידע מטה על הגן SCT (refseq gene NM\_021920.4, לא בכרומוזום אלטרנטיבי).

1. הביטו ורשמו את המיקום של הגן (refseq) בפורמט BED מלא (שש עמודות, הקפידו על פורמט תקני).
2. מה מספר האקסונים שקיבלתם?
3. רשמו את האזור המקודד של הגן
4. רשמו את אורכי האקסון הראשון והאחרון. הסבירו איך מצאתם אותם (המלצה: בסעיף הזה כדאי לוודא את עצמכם עם הGenome browser).

הדרכה: אם אתם רואים שאתם מסתבכים עם השאלה: אתם לא בכיוון הנכון =] מטרת השאלה היא לראות פלט בצורה נאיבית. כל הפרטים שאתם צריכים נמצאים כבר בטבלה, אך איך תציגו את כל המידע הנדרש? חשבו איך אנחנו מקבלים פורמט BED.

בונוס (1 נק'): מה המשמעות של כרומוזום אלטרנטיבי/שברי כרומוזומים? כיצד הדבר מתבטא? (רמז: קריוטיפ)

### שאלה 3: סינון בסיסי

השתמשו בטבלת ה-'refGene table' (‘Gene and Gene Prediction Tracks’ group, ‘NCBI RefSeq’ track), על מנת למצוא את כל הגנים בכל הגנום אשר:

* בגדיל השלילי
* בכרומוזומים 12-23 או בכרומוזום M.
* בהכרח יש 3’UTR.
* גודל כלל הגן מתחילתו ועד סופו הוא לכל היותר 1000
* ישנו יותר מאקסון אחד.

הדרכה: מאוד מומלץ לצייר את זה מולכם ולחשוב מהו 3’UTR ומהו 5’UTR בכל אחד מהגדילים. הקפידו לרשום את הפרמטרים איתם סיננתם, ואת שורת הquery במידה והשתמשתם בה! השורה תצא מעט ארוכה, אך הדבר בסדר. ממליץ לבנות את השורה בשלבים ואז לבדוק את עצמכם.

ענו בבקשה על הסעיפים הבאים:

1. כמה גנים מצאתם?
2. בחרו את הגן בעל מספר האקסונים הרב ביותר: רשמו את השם שלו (NM\_xxxxx) ואת המיקום שלו בפורמט BED.

הדרכה: חשבו כיצד תקבלו רק את הפלט הרלוונטי לכם?

את השלב הבא בצעו עבור הגן אותו מצאתם בסעיף ב'. שימו לב שעליכם לשנות את הטבלה והמיקום!!!

1. כמה SNP קיימים באזור זה? השתמשו ב-All SNP’s(151) מתוך הקבוצה Variation
2. כמה מתוך ה-SNP הללו יוצרים מוטציית missense בגן שלנו? הקפידו להסביר איך הגעתם לתוצאה. (שימו לב לגדיל!)

### שאלה 4 – חיתוך טבלאות ו-tracks

מגפת הקורונה השפיעה עלינו רבות. כחלק מהמאמץ למלחמה בקורונה, גוייסתם למערך הביואינפורמטי כנגד הנגיף. פתחו את הקבוצה COVID-19, COVID GWAS v4 והטבלה: Severe COVID vars.

1. כמה אתרים מקודדים יש בכרומוזום 10?
2. כמה מן האתרים מצויים באזורים של גנים? (רמז: חיתוך)
3. כמה מן האתרים שמצאת בסעיף ב' הם כן בתוך אלמנטים חוזרניים? (רמז: צרו custom track מהתוצאות בסעיף ב').
4. כמה מן האתרים שמצאתם בסעיף ב' מצויים בתוך אלמנטים חוזרניים ממשפחת Alu? (רמז: אתם צריכים ליצור את זה).
5. כמה מתוך הוריאנטים שמצאתם בסעיף ד' מצויים גם ב-dbSNP153 של clinvar? (יש טבלה בשם זה).
6. חוקר צעיר שמע שקיימים אנזימים המסוגלים לערוך בסיסי אדנוזין וציטוזין. חזרו על סעיף ה' אבל הפעם סננו מראש את ה-clinvar כך שיכיל רק “A,” או “C,” (רמז: filter alts). צרו track פעם אחת עבור כל בסיס בנפרד. כעת:
   1. כמה קיבלתם מכל סוג? כלומר בחיתוך עם טבלאות ל-A או ל-C?
   2. במידה וקיבלתם עד 10 מכל סוג – צרו טבלה של שם הSNP ושם הגן. אתם יכולים לחפש את שם הגן בGenome browser

הדרכה: כל סעיף דורש מכם לשחק עם הפרמטרים שאתם מגדירים, שימו לב שהטבלאות שלכם נכונות בכל סעיף!! השתמשו בטבלת הגנים של GENCODE V38 ובערוץ RepeatMasker. סעיף ד' דורש מכם קודם לייצר את הטבלה. שימו לב: repClass הינו SINE, repFamily הינו Alu. שמרו את זה בtrack.

### שאלה 5: עבודה בסיסית עם BEDtools

שאלה זו מטרתה להכין לקראת שני התרגילים הבאים, בהם אנו נעבוד בשורת פקודה באופן מקיף יותר. הורידו אליכם לשרת את הטבלאות הבאות בפורמט BED. ממליץ להוריד אותם מכווצות בgzip. אנו עובדים רק על אזור ספיציפי הפעם –כרומוזום 21. שימו לב שאין filters או intersections!

* ה-Gencode V38 של הכרומוזום. בפלט הורידו טבלה לכל אחד מהבאים:
  + כל הגן
  + אקסונים
  + אינטרונים
  + CDS
* טבלת ה-Svere COVID vars (COVID GWAS v4) (בכל הגן)
* טבלת הרפיטים באזור

אחרי שכל הטבלאות בשרת, נבצע חיתוכים עם הטבלאות. היעזרו בפקודה intersectBed. שימו לב להנחיות:

1. רשמו את הפקודה שהרצתם בוורד ואת התוצאה. אין צורך לפרט מעבר.
2. רשמו את כלל הפקודות בקובץ טקסט עם סיומת (sh) אשר נמצא בפורמט Unix (כלומר EOL הוא “\n”. היעזרו בnotepad++. נא לרשום הערה (השורה תתחיל ב#) עם מטרת הפקודות. העלו את הקובץ יחד עם התרגיל.
3. שימו לב שהסינטקס שלכם ייבדק!

והשאלות:

1. כמה שורות יש בתוך Covid vars? כמה מתוך הרשומות נמצאים ב-Whole genes?
2. כמה אקסונים יש להם אינטרון שחופף להם (רמז: שימו לב! עליכם להחזיר את טבלת האקסונים, לא את החיתוך בין אינטרון לאקסון. חשבו איזה פרמטר נותן לכם את זה.
   1. כמה אקסונים קיבלתם?
   2. סננו רשומות כפולות בעזרת sort וuniq). כמה תוצאות אתם מקבלים?
   3. שמרו את הרשימה שקיבלתם.
3. חתכו את הרשימה שלכם מסעיף 3 עם ה-CDS, סננו רשומות כפולות כמו מקודם. כמה תוצאות קיבלתם? שמרו את תוצאת החיתוך.
4. . סננו רשומות כפולות.
   1. כמה תוצאות קיבלתם?
   2. מזגו את התוצאות שקיבלתם (bedtools merge) (שימו לב שצריך למיין לפני ☺ ). כמה רשומות קיבלתם הפעם? (התוצאה אמורה להיות שונה)
   3. שמרו את התוצאה שלכם
5. חתכו את ה-COVID vars יחד עם התוצאה שקיבלתם בסעיף הקודם.
   1. כמה תוצאות קיבלתם?
   2. הדפיסו למסך את 20 השורות הראשונות וצרפו צילום מסך.

בונוס (3 נק'): תוכלו להסביר מה למעשה אנו מקבלים? מה חיפשנו ומה קיבלנו (בשאיפה)?