מעבדה לסטטיסטיקה 2024, מטלה 2
הצגה 28.5

במטלה זו נחקור קבצי מיפוי, ונבדוק את התפלגות הפרגמנטים על פני הכרומוזום הראשון.

קראו את קבצי הפרגמנטים כפי שהראינו בכיתה, ע״פ החלוקה הבאה:

1. קבוצות A עובדות על קובץ `TCGA-13-0723-01A_lib2_all_chr1.forward`
2. קבוצות B עובדות על קובץ `TCGA-13-0723-01A_lib1_all_chr1.forward`
3. קבוצות C עובדות על קובץ `TCGA-13-0723-10B_lib1_all_chr1.forward`

שימו לב שהקבצים מכווצים בעזרת gzip, ויש לפתוח את הכיווץ .

א. כתבו פונקציה שלוקחת נקודת התחלה ונקודת סיום, ומייצרת וקטור שעבור כל בסיס רושם את מספר הפרגמנטים המתחילים בבסיס (השלימו את השלד בקוד המעבדה). הריצו את הפונקציה על אזור של 20 מיליון בסיסים בכרומוזום לבחירתכם. הציגו את הקוד שכתבתם, וכן כמה זמן לוקח לפונקציה שלכם לרוץ עבור 20 מיליון הבסיסים (ציינו גם כמה פרגמנטים סה״כ היו באזור זה). הקפידו שהקוד יהיה קריא; בדקו שאכן אין לכם טעויות בחישוב.
רשות – נסו לכתוב פונקציה שרצה יותר מהר, והשוו מהירות.

ב. הציגו היסטוגרמה עבור הכיסוי השולי בנקודות בודדות (כפי שעשינו בכיתה).

ג. בחרו גודל תא (בכפולות 10,000) וסכמו את מספר הפרגמנטים בכל תא מתוך Loc. (גם פה אולי כדאי לכתוב כפונקציה).
1. הציגו גרף אחד המתאר את ההתפלגות על פני הכרומוזום (כלומר מספר רידים מול מיקום התא) .
2. הציגו גרף שני המתאר את ההתפלגות בחלון של 20 מיליון בסיסים, כלומר zoom-in של הראשון.

דונו במה שאתם רואים. האם אתם מזהים איזור שפוטנציאלית יש בו שינוי במספר העותקים (משמעותית יותר או פחות רידים מהרקע)?

ד. הסתכלו על ההתפלגות השולית של מספר reads בתא. התאימו לה את ההתפלגות הנורמלית הקרובה ביותר, ותארו באופן ויזואלי את ההתאמה. האם הם מתאימים?

אתם מוזמנים לחקור עוד את התפלגות הפרגמנטים על פני הגנום .