##### **קמפוס טל**

**החוג לביואינפורמטיקה**

**פרויקט גמר**

**שם הפרויקט: מציאת והצגת מידע על רמת ביטוי גנים בתאים בודדים באנליזת single cell RNA-seq שנעשית בעזרת הכלי Seurat**

**מגישות:**

**הדר חלאווה 323809798**

**מרים חנה בלה 209265230**

**מנחה:**

**ד"ר אוריה ורדי-יעקב**

**5/8/2023**

תקציר:

הפרויקט עוסק בניתוח נתונים עם החבילה Seurat. חבילה לניתוח של נתוני single cell RNA-seq. השימוש בכלי זה נותן מידע על האחוז הכללי של התאים המבטאים גן מסויים בקלסטר נבחר, אך אינו נותן מידע על אחוז ביטוי גנים בכל תא ותא בנפרד בקלסטרים השונים ולכן פיתחנו כלי בשביל להנגיש את המידע המתקבל מהאנליזות שבוצעו באתר. ראינו שלא קיימים כלים נוספים שמנגישים ומסדרים את המידע החסר הנ"ל, ולכן עבדנו על הנגשת המידע למשתמש בצורה מאירת עיניים.

את הפרויקט כתבנו בשפת R ובעזרת חבילת Seurat המיועדת ל-QC ,ניתוח וחקר נתוני תא בודד. יצרנו סקריפט שמעביר את הנתונים תהליך של קיבוץ וחלוקה למשתנים עד לקבלת אחוז ביטוי הגנים שמעל אחוזי הסף, בכל תא ותא מהקלסטר המבוקש. לאחר מכן, בחנו את יעילות הכלי על נתונים ממאמר העוסק בתחום אנהנסורים בקשר לסרטן לוקמיה עם דגש על גן BCL11B. חיפשנו גנים הקשורים לגן המבוקש וחלק מהגנים נמצאו בנתונים,אך הם לא היו סיגניפיקנטים ולכן לא יכולנו להוכיח קשר בין הגנים.

הקדמה:

הפרויקט עוסק בהבנת וניתוח נתונים עם חבילת Seurat . חבילת Seurat נותנת שני סוגי ניתוחים של נתונים. הראשון הוא ניתוח של מערכי נתונים מרחביים מבוססי רצף והדמיה, והשני זה ניתוח רב-מודאלי אינטגרטיבי. הניתוח הראשון של מערכי נתונים שנפתרו במרחב, מגדירים מחדש את ההבנה שלנו לגבי אינטראקציות סלולריות וארגון רקמות אנושיות. גם לטכנולוגיות מבוססות רצף (כלומר Visium, SLIDE-seq וכו'), וגם טכנולוגיות מבוססות הדמיה (MERFISH/Vizgen, Xenium, CosMX וכו') יש יתרונות ייחודיים, ודורשות שיטות אנליטיות מותאמות ותשתית תוכנה. באתר מציגים תמיכה גמישה ומגוונת עבור מגוון רחב של סוגי נתונים עם פתרון מרחבי, ותמיכה בטכניקות אנליטיות לאינטגרציה של scRNA-seq, דה-קונבולוציה וזיהוי נישה.

הניתוח השני, הוא ניתוח רב-מודאלי אינטגרטיבי. התעתיק הסלולרי הוא רק היבט אחד של הזהות התאית, וטכנולוגיות עדכניות מאפשרות פרופיל שגרתי של נגישות כרומטין, שינויים בהיסטונים ורמות חלבון מתאי בודד.

תפקיד האנליזות הוא למצוא אפיון לתאים בקלאסטרים השונים,למצוא גנים שמתבטאים בצורה דיפרנציאלית בין התאים וכו

למרות שהמטרה העיקרית היא קלסטור של סוגי תאים משותפים בין מערכי נתונים, זה יכול לגרום להרבה בעיות. משתמשים רבים עשויים להיות מודאגים לגבי איזו שיטה להשתמש, או שהאינטגרציה עלולה לגרום לאובדן הרזולוציה הביולוגית.

כל השיטות מדגישות הדמיות ברורות מושכות וניתנות לפרשנות, ותוכננו לשימוש [קל הן](https://satijalab.org/seurat/articles/get_started.html) על ידי חוקרי מעבדה יבשה והן על ידי חוקרי מעבדה רטובה.

החבילה של Seurat נותנת אפשרות לבצע כל מיני אנליזות על הנתונים. יש אפשרות לקבל את רמת הביטוי של הגנים בתאים אבל ראינו שקשה להסיק מסקנות מהמספרים היבשים של רמות ביטוי ולכן רצינו שנוכל לקבל גם כל מיני אחוזים ומידע נוסף כדי להבין יותר את הנתונים. אנחנו פיתחנו כלי שלוקח את האנליזות שבוצעו בעזרת החבילה של Seurat ומפשיט ומנגיש את המידע בצורה פשוטה וקלה להבנה. ראינו שלא קיימות שיטות נוספות שנותנות מידע על אחוז ביטוי גנים בכל תא בנפרד, אלא קיים רק מידע כללי לגבי אחוז התאים המבטאים גן מסוים בקלסטר נבחר ולכן עבדנו על להנגיש את המידע למשמש בצורה מאירת עיניים. הנגשת המידע נעשתה בעזרת קבלת ממוצע התאים המבטאים גן מסוים, ומה רמת הביטוי הממוצעת לאותו גן וכו'. (דוגמה בנספחים, תמונה ו'). הנגשת המידע למשתמש או לחוקר יכולה לעזור בשימוש במידע בצורה נכונה ומיטבית. מידע מבולגן ולא מונגש עלול להסתיר בתוכו קורלציות חשובות שיכולות לתרום רבות לחוקרים בניסויים. הנגשת המידע תורמת למניעת בלבול שעלול להיווצר עקב אחוזי הנתונים המוצגים. נירמלנו וסידרנו את המידע כך שמשתמש לא רק יוכל לקבל את המידע הגולמי, אלא גם כמה הוא מתבטא ביחס לנתונים אחרים ובכך לתת עוד פרופורציה למשתמש על טיב הנתונים.

לסיכום, מטרתנו בפרויקט היא למצוא ולהציג מידע על רמת ביטוי גנים בכל תא ותא בקלסטר, בנוסף למידע הקיים, שהוא האחוז הכללי של הגנים המתבטאים בכל הקלסטר.

מתוך המאמר המוזכר עבדנו על נתונים של שני מטופלים, מטופל SJMPAL068275\_D1 (שנקרא לו מספר 75), ומטופל SJAML068287\_D1 (שנקרא לו מספר 87), ניסינו למצוא קשר סיגניפיקנטי בין סוגי תאים מסוימים,כגון אנהנסורים פרומוטורים וכו', את המידע על גנים הקשורים לגן BCL11B השגנו בעזרת אתר geneCard(יפורט בחלק תוצאות ודיון), מובא תקציר מידע אודות אנהנסורים

Enhancer **-** הוא מקטע [DNA](https://he.wikipedia.org/wiki/DNA) קצר,המשמש להגברת [תיעתוק](https://he.wikipedia.org/wiki/%D7%AA%D7%A2%D7%AA%D7%95%D7%A7_(%D7%91%D7%99%D7%95%D7%9C%D7%95%D7%92%D7%99%D7%94)) של [גנים](https://he.wikipedia.org/wiki/%D7%92%D7%9F_(%D7%91%D7%99%D7%95%D7%9C%D7%95%D7%92%D7%99%D7%94)) באופן ספציפי לסוג התא ולזמן ההתפתחותי. המעצם בדרך כלל אינו ממוקם בסמוך לגן המטרה שלו, ולכן על מנת שיוכל לפעול, [הכרומטין](https://he.wikipedia.org/wiki/%D7%9B%D7%A8%D7%95%D7%9E%D7%98%D7%99%D7%9F) (ה-DNA ו[החלבונים](https://he.wikipedia.org/wiki/%D7%97%D7%9C%D7%91%D7%95%D7%9F) העוטפים אותו) מתקפל כך שהאנהנסור יגיע לסביבת [הקדם](https://he.wikipedia.org/wiki/%D7%A4%D7%A8%D7%95%D7%9E%D7%95%D7%98%D7%95%D7%A8) (פרומוטור) של גן המטרה. המעצם אינו מתאפיין ברצף [נוקלאוטידים](https://he.wikipedia.org/wiki/%D7%A0%D7%95%D7%A7%D7%9C%D7%90%D7%95%D7%98%D7%99%D7%93) ספציפי המבדיל בינו לבין אלמנטים אחרים בגנום, אלא על ידי שינויים [אפיגנטיים](https://he.wikipedia.org/wiki/%D7%90%D7%A4%D7%99%D7%92%D7%A0%D7%98%D7%99%D7%A7%D7%94)

שיטות:

את הפרויקט כתבנו בשפת R ובעזרת חבילת Seurat,המיועדת ל-QC ,ניתוח וחקר נתוני single cell RNA-seq.

ראשית עברנו על ההדרכה של האנליזה שיש באתר של Seurat והרצנו אותו כדי לקבל תמונה ברורה של התהליך שעברו הנתונים. ההדרכה באתר כוללת כמה שלבים.

קראנו את הנתונים ויצרנו אובייקט Seurat. האובייקט מכיל גם מטריצת נתונים וגם ניתוח שלהם, כמו PCA וקליסטור עבור מערך נתונים של תא בודד. מערך הנתונים עצמו בנוי כמטריצה דלילה משום שלרוב, רוב הערכים מכילים 0. בצורה כזו, אחסון כזה של הנתונים עוזר לחסוך בזיכרון ובמהירות.

לאחר מכן אנו מסננים תאים מהדטה ע"פ כל מיני קריטריונים, לדוגמה: מספר הגנים הייחודיים שזוהו בכל תא (מספר גבוה או נמוך באופן חריג מראה על טיפות ריקות או תאים כפולים בטיפה), המספר הכולל של מולקולות שזוהו בתוך תא (מתאם חזק עם גנים ייחודיים), ואחוז הקריאה שממפה לגנום המיטוכונדריאלי (תאים עם ספירות מיטוכונדריאליות נמוכה מראה על כך שהם גוססים ולא מתפקדים כראוי ולכן הם ככל הניראה יתנהגו בצורה שונה מתאים רגילים בריאים). לאחר כל הסינונים נירמלנו את הדטה. [גרף ויזואלי של מטופל מספר 75 בתמונה א' בנספחים].

בשביל לאפשר חקירה קלה של הנתונים אנו משתמשים במפות חום שמראות לנו באופן ויזואלי את הביטוי של הגנים בתאים. מפות חום יכולות לעזור לנו כאשר אנו מנסות להחליט אילו PC לכלול לניתוחים נוספים בהמשך האנליזה. גם התאים וגם התכונות מסודרים לפי ציוני ה-PCA שלהם. הגדרה למספר מתווה את התאים ה'קיצוניים' בשני קצות הספקטרום, מה שמאיץ באופן דרמטי את ההתוויה עבור מערכי נתונים גדולים

**PCs** אלו מימדים שיצאו מPCA, נניח PC1 הוא המימד(תכונה) החשוב ביותר( מסביר את השונות בין התאים באופן הגדול ביותר). וכך, PC2 הוא גם מימד חשוב אך פחות מסביר שונות מאשר PC1( לרוב נשתמש ב-PC1 ו- PC2 כי השאר כבר פחות מועילים).

ממפת החום והגרפים שיצאו מ-PCA נלמד את התכונות החשובות שנרצה להמשיך לעבוד איתן, מפת חום זו מפה שהערכים בה מיוצגים על ידי סקאלה של צבעים, ב-PCA נוריד למספר מימדים מסויים שיתן לנו את אחוזי השונות הגבוהים ביותר. כל מפת חום מתארת מימד שהורידו אליו כך ש- PC1 הוא המימד המשמעותי ביותר וזה הולך ופוחת ככל שהמימדים נוספים. לכן לרוב אנו נעבוד עם PC1 וPC2 שהם המימדים שמסבירים הכי טוב את השונות בין הקלסטרים. [דוגמה בנתונים שלנו בתמונות ד']

להורדת רעש טכני, מבצעים שינוי אקראי של קבוצת משנה של הנתונים (1% כברירת מחדל) ומפעילים חדש את PCA, בונים 'התפלגות אפסית' של ציוני תכונות, וחוזרים על הליך זה.

אחרי כל האנליזות והסידורים של הנתונים אנו רוצים לחלק את הנתונים שלנו לקלסטרים. אנו בונים תחילה גרף KNN המבוסס על המרחק האוקלידי בחלל PCA, ומצמצמים את משקלי הקצוות בין כל שני תאים בהתבסס על החפיפה המשותפת בשכונות המקומיות שלהם (דמיון ג'קארד). שלב זה מבוצע באמצעות הפונקציה ()FindNeighbors ולוקח כקלט את הממדיות שהוגדרה קודם לכן של מערך הנתונים. המטרה של אלגוריתמים אלה, היא ללמוד את המעטפת הבסיסית של הנתונים על מנת למקם תאים דומים יחד במרחב בעל ממדים נמוכים. תאים בתוך האשכולות מבוססי הגרף שנקבעו לעיל צריכים להתמקם במשותף על חלקות הפחתת ממדים אלה. [דוגמה לקלסטרים בתמונה ג']

לאחר שחילקנו את הנתונים לקלסטרים, אנו רוצים למצוא את התכונות שמגדירים כל קלסטר. אנו מוצאים אותם באמצעות ביטוי דיפרנציאלי. כברירת מחדל, הוא מזהה סממנים חיוביים ושליליים של אשכול יחיד בהשוואה לכל התאים האחרים. ישנה גם דרך לראות את רמת הביטוי של גן מסויים בכל התאים עם החלוקה של הקלסטרים. [תמונה ב' בנספחים]

אנו רוצים למצוא גנים דיפרנציאלים המבדילים בין קלסטר מסויים לשאר הקלסטרים. בשביל זה, היה עלינו להבין את החישוב שעושים כדי לחלק לקלסטרים, ולהבין מתוך אחוזי הגנים המתבטאים בכל קלסטר, אלו תאים מסוימים כן מבטאים את הגן הנדרש ואלו לא מבטאים אותו. העברנו את הנתונים נרמול , והכנסנו אותם לתוך מטריצה שיכולה לקבל שם של תא ושם של גן והיא תחזיר את רמת הביטוי של אותו גן באותו תא. לאחר מכן, יצרנו פונקציה שמקבלת שם קלסטר (או כמה שמות של קלסטרים) ומדפיסה את שמות כל התאים שיש באותו קלסטר.

חיפשנו את סף ביטוי הגנים ומצאנו שסף הביטוי הוא 0.1 כברירת מחדל- יכול להשתנות לפי הצורך. כתבנו פונקציה שתקבל שם של קלסטר ושם של גן ותחזיר את התאים באותו קלסטר המבטאים את הגן הנדרש (יותר מהסף) וגם כמה אחוזים זה מכל קלסטר. לאחר מכן יצרנו פונקציה שמדפיסה את רמת ביטוי הגן עבור כל תא ותא בקלסטר וכן ממוצע כולל של רמת ביטוי הגן בכל התאים, ולבסוף יצרנו היסטוגרמה עם רמת ביטוי הגנים עבור כל תא. (דוגמה נמצאת בנספחים בתמונה ה', ו').

מבנה הDB בו השתמשנו לאורך התהליך הוא מערך נתונים של תאים חד-גרעיניים היקפיים בדם (PBMC) הזמינים באופן חופשי מ- 10X Genomics. ישנם 2,700 תאים בודדים שרוצפו ב- Illumina NextSeq 500.(מצורף קישור להורדת הנתונים)

דרך השימוש בפונקציות שיצרנו:

כדי להשתמש בסקריפט שיצרנו ראשית יהיה על המשתמש להעביר על הנתונים את האנליזות הראשונות המופיעות באתר Seurat עד הסוף, שמירת הנתונים לקובץ RDS. לאחר מכן יהיה ניתן להשתמש בכלי שלנו עם הקובץ הנל'.

תוצאות ודיון:

יצרנו סקריפט בשפת R.ראשית קוראים לפונקציות קונסטרקטור שמקבלת כתובת של הנתונים ששמרנו מהאנליזה, ורשימה של כל השמות של הקלסטרים. פונקציית הקונסטרקטור מסדרת את הנתונים עבור כל שאר הפונקציות כדי שיהיה אפשר לעבוד עליהן.

בתוך הסקריפט קיימות כל מיני פונקציות לשימוש ,אם יש צורך אז אפשר לקרוא לפונקציה "InpoScript\_finalProject" שמחזירה טקסט ארוך שמסביר אלו פונקציות קיימות, מה הן עושות, מה הן מקבלות ומה הן מחזירות.

יש את הפונצקיה "Cells\_in\_clusters" שמקבלת שם אחד או יותר קלסטרים ומחזירה את התאים שבאותו קלסטר. יש גם את פונקציית "FindMarkers\_New" שמקבלת את אותם פרמטרים כמו FindMarkers ומחזירה רשימה של 3 אובייקטים. האובייקט הראשון זה הטבלה שמתקבלת משימוש ב-"FindMarkers", ושני האובייקטים האחרים הם וקטורים של התאים שבכל קלסטר שניתן לפונקציה (כי הפונקציה מקבבלת שני שמות של קלסרים).

הפונקצייה האחרונה היא פונקציית "ExpressionGenes" והיא מקבלת שם של קלסטר (אחד או יותר), שם של גן וביטוי מינימלי (ברירת מחדל של 0.1). הפונקציה מחזירה שלוש אובייקטים והם וקטור של כל התאים שבקלסטר/ים שביטאו את הגן הניתן ברמת ביטוי גבוה מרמת הביטוי המינימאלית (בניתנה לפונקציה), ממוצע של מספר התאים שביטאו את הגן הנ"ל בקלסטר/ים, וביטוי ממוצע של התאים.

בדקנו את נכונות ויעילות הסקריפט על נתונים ממאמר "Enhancer Hijacking Drives Oncogenic *BCL11B* Expression in Lineage-Ambiguous Stem Cell Leukemia"

המאמר עוסק בתחום אנהנסורים בקשר לסרטן לוקמיה עם דגש על גן *BCL11B*. ערכנו בירור באתר genecards על הגן המבוקש והבאנו תקציר על פעולותיו העיקריות:

* החלבון המקודד ידוע כמדכא שעתוק.
* מווסת מפתח הן של התמיינות והן של הישרדות של לימפוציטים מסוג T.
* וסת של מקדם IL2.
* משפר את ביטוי IL2 בלימפוציטים.
* מדכא גידולים המדכא שעתוק באמצעות קשירה ישירה ובלתי תלויה ב-TFCOUP2 לאלמנט תגובה עשיר ב-GC.
* עשוי לתפקד גם במסלול איתות P53.

חיפשנו את הגנים המוזכרים וגנים נוספים הקשורים לגן המבוקש. חלק מהגנים נמצאו בנתונים,אך הם לא היו סיגניפיקנטים ולכן לא יכולנו להוכיח קשר בין הגנים. הוספנו בנספח תמונות מהרצת התוכנית על המאמר.

רעיונות להמשך תהליך עם הכלי, ניתן לחשב על כיוון ה-scRNA-seq עבור תאים עם ביטוי משמעותי בקלסטר, כך ניתן ליצור כלי שיוכל לזהות דפוסים שחוזרים על עצמם בתנאים מסוימים וכשנכנס שם של תא ונזין את סוג התא נוכל לקבל מידע ראשני עליו כמו אחוז ביטוי גן ספציפי בזמן מסויים של התא וכמובן שיכולים להיות עוד כיוונים מגוונים לפיתוח הכלי.

ביבליוגרפיה:

קישור לאתר Seurat-

[Tools for Single Cell Genomics • Seurat (satijalab.org)](https://satijalab.org/seurat/)

להורדת הנתונים עליהם עבדנו לחץ [כאן](https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz)

קישור למאמר -

<https://aacrjournals.org/cancerdiscovery/article/11/11/2846/666485/Enhancer-Hijacking-Drives-Oncogenic-BCL11B>

הנתונים - GEO

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE173320>

קישור לנתונים מאתרgenecards אודות *BCL11B*

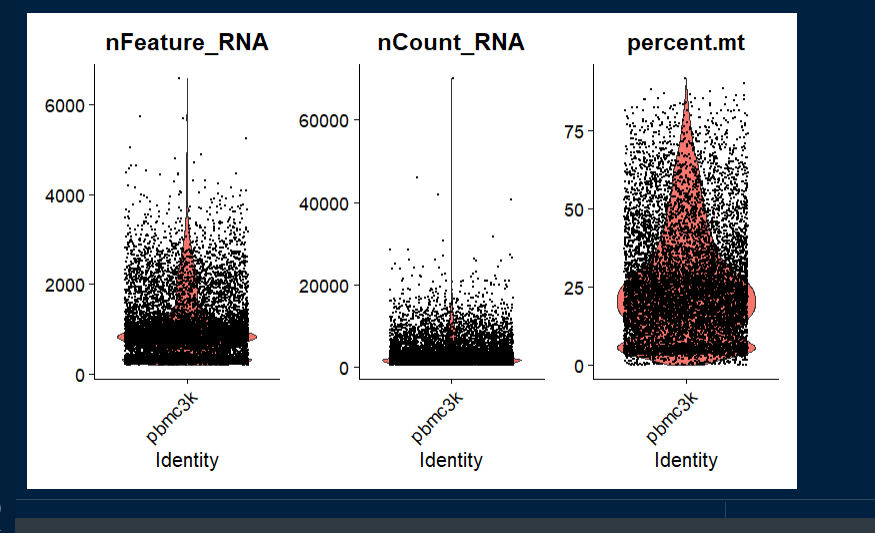
[BCL11B Gene - GeneCards | BC11B Protein | BC11B Antibody](https://www.genecards.org/cgi-bin/carddisp.pl?gene=BCL11B&keywords=BCL11B)

נספחים:

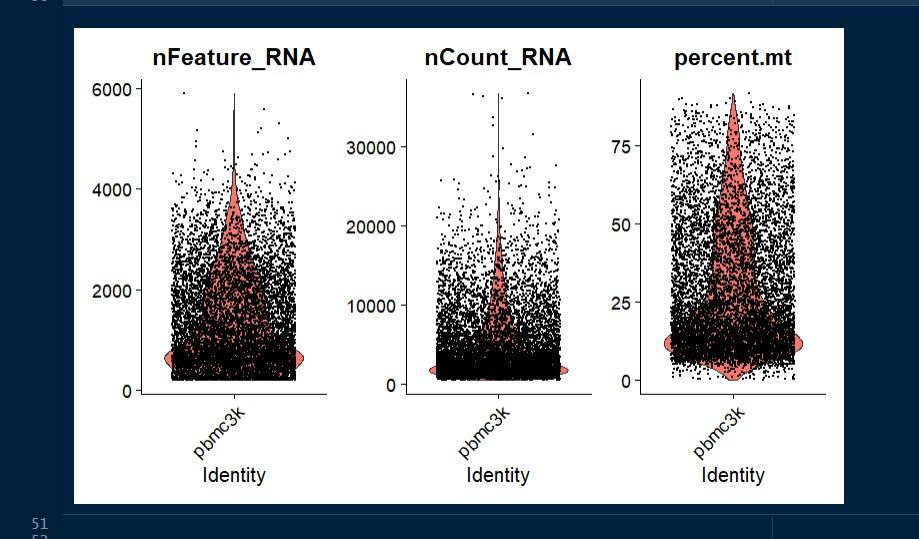
תמונה א': גרף ויזואלי שמראה את התפלגות הנתונים בשביל סינון התאים.

מובאות תמונות של האנליזות שיצאו מהנתונים במאמר המוזכר.

אדם 75:

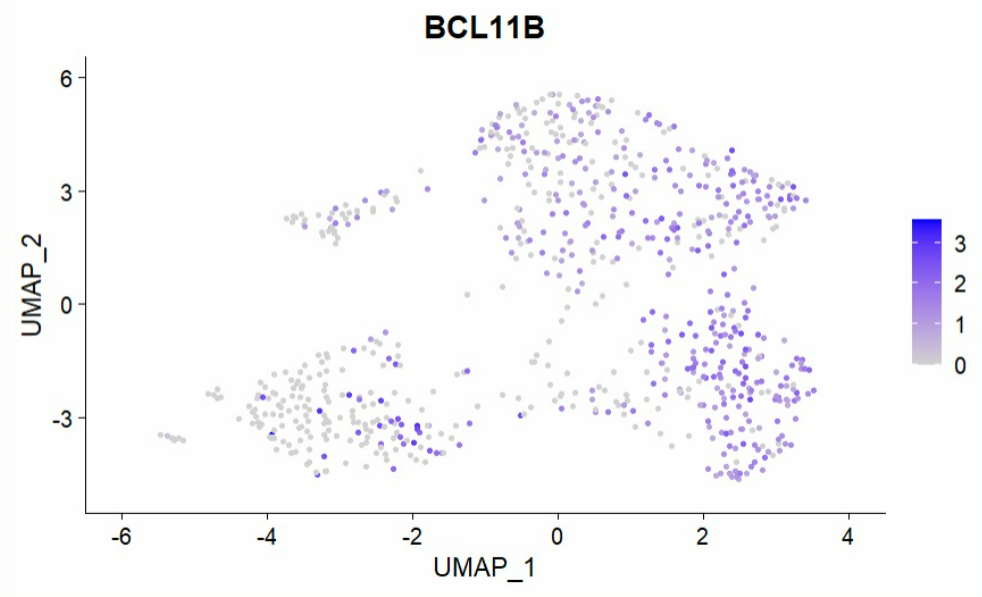


אדם 87:



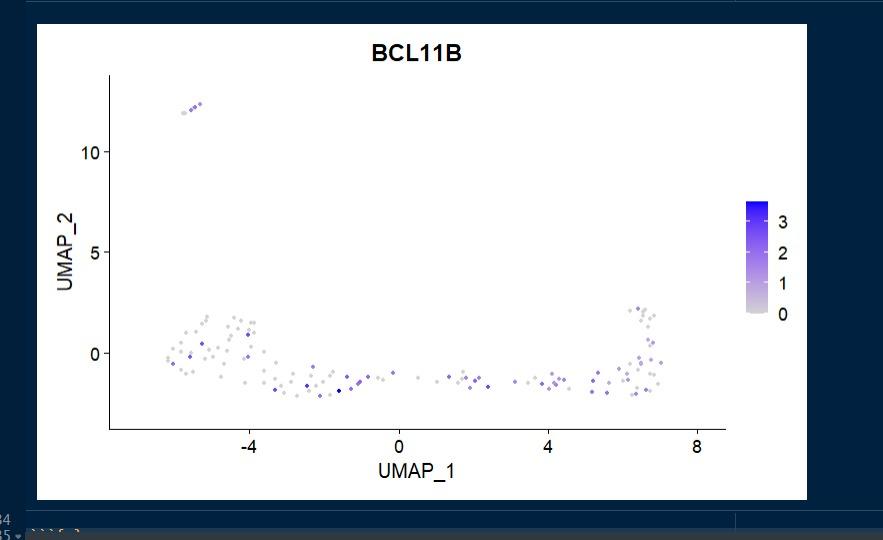
תמונה ב': רמת ביטוי של הגן *BCL11B* בכל התאים בדטה-סט.

אדם 75:



אנו יכולים לראות שבקלסטר B1 יש הכי הרבה תאים המבטאים את הגן וברמה הכי גבוהה.

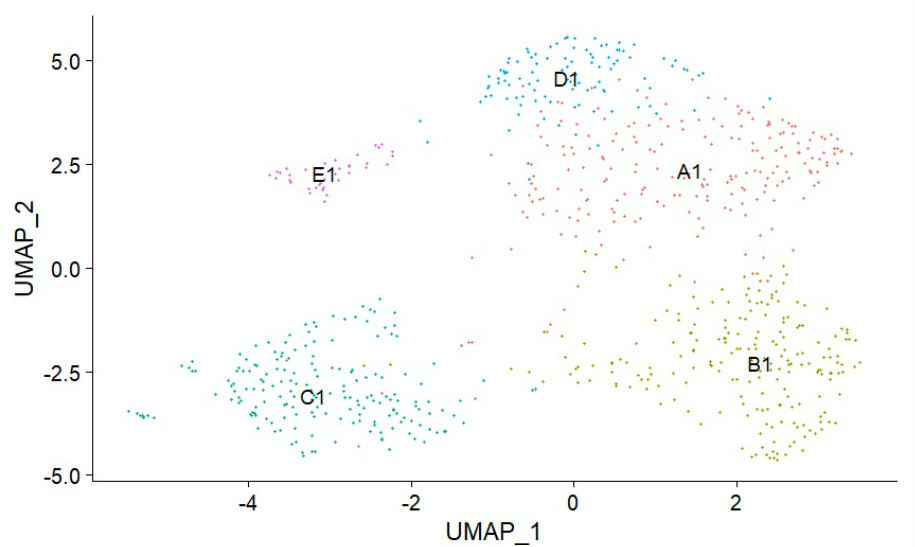
אדם 87:



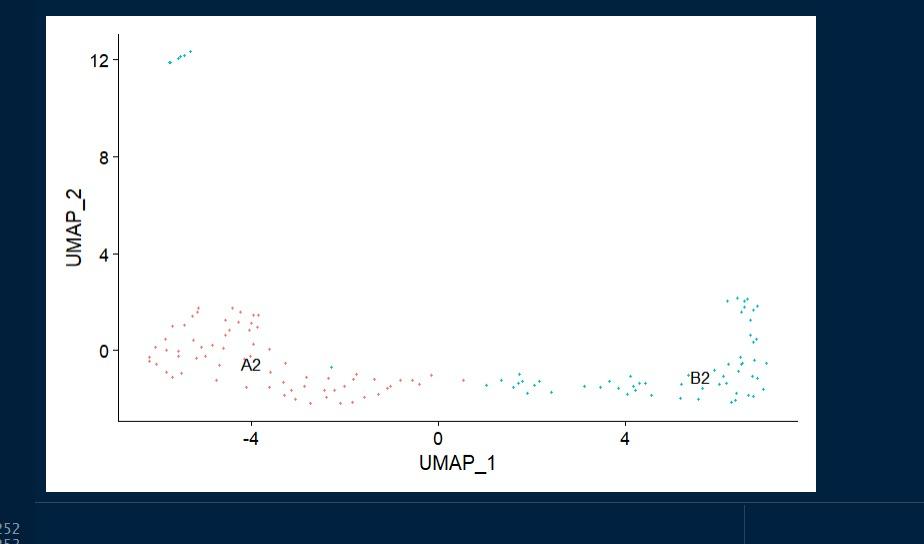
אנו יכולים לראות שבקלסטר B2 יש הכי הרבה תאים המבטאים את הגן וברמה הכי גבוהה.

תמונה ג': הסתדרות לקלסטרים במרחב לפי רמת דמיון בין התאים, מסתדר לפי צבע ומיקום -המיקום משמעותי יותר. הדתה סט התחלק ל5 קלסטים במטופל 75 ונתנו להם שמות A1, B1, C1, D1, E1, והדטה-סט של מטופל 87 התחלק ל-2 קלסטרים ונתנו להם את השמות A2, B2. השמות חסרי משמעות ונועדו עבורנו בשביל להבדיל ביניהם.

אדם 75:

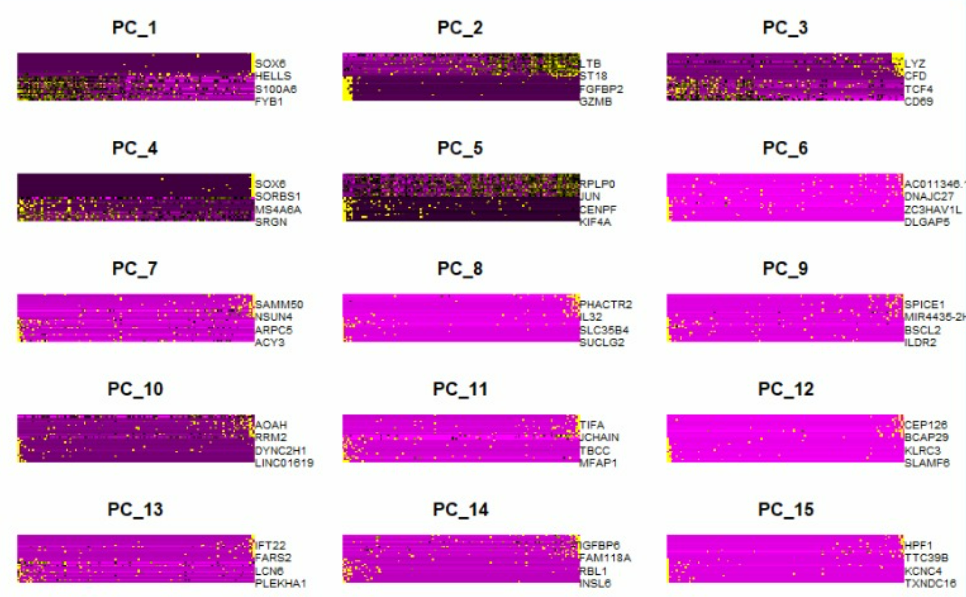


אדם 87:

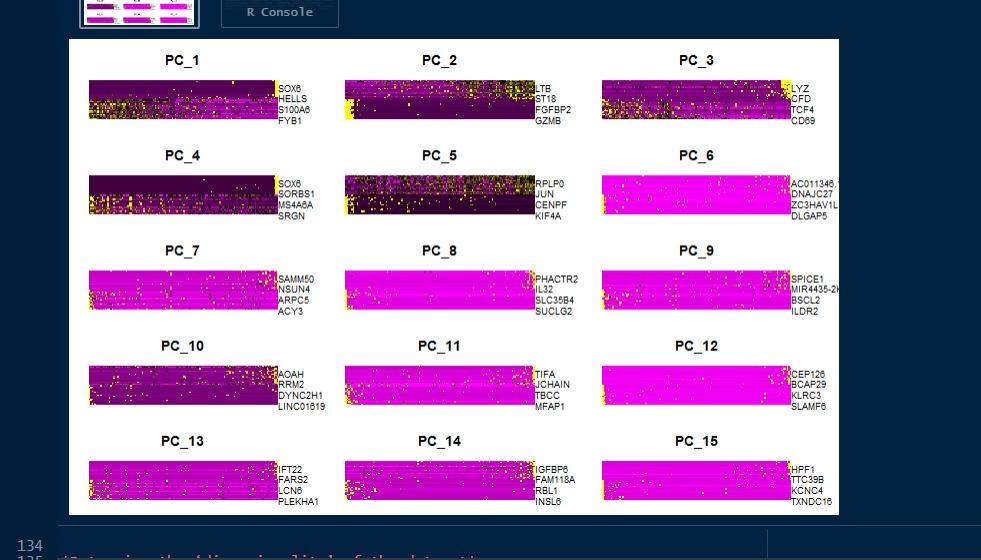


תמונה ד': תמונה של מפות החום לפי קבוצות הגנים, ע"פ חשיבות הקבוצה לקילסטור. כל מפת חום מתארת מימד שהורידו אליו כך ש- PC1 הוא המימד המשמעותי ביותר וזה הולך ופוחת ככל שהמימדים נוספים.

לכן לרוב אנו נעבוד עם PC1 וPC2 שהם המימדים שמסבירים הכי טוב את השונות בין הקלסטרים

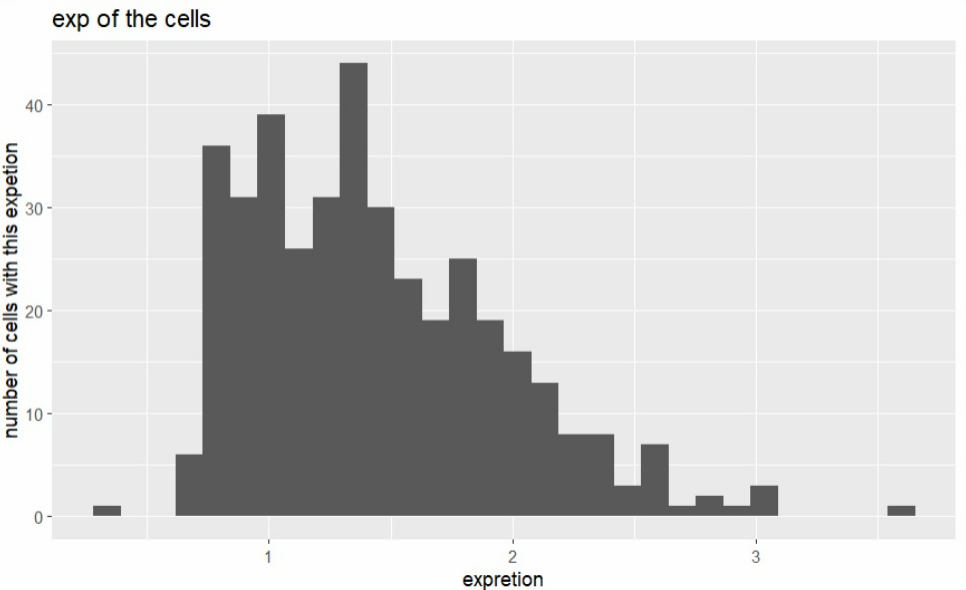
אדם 75:

אדם 87:

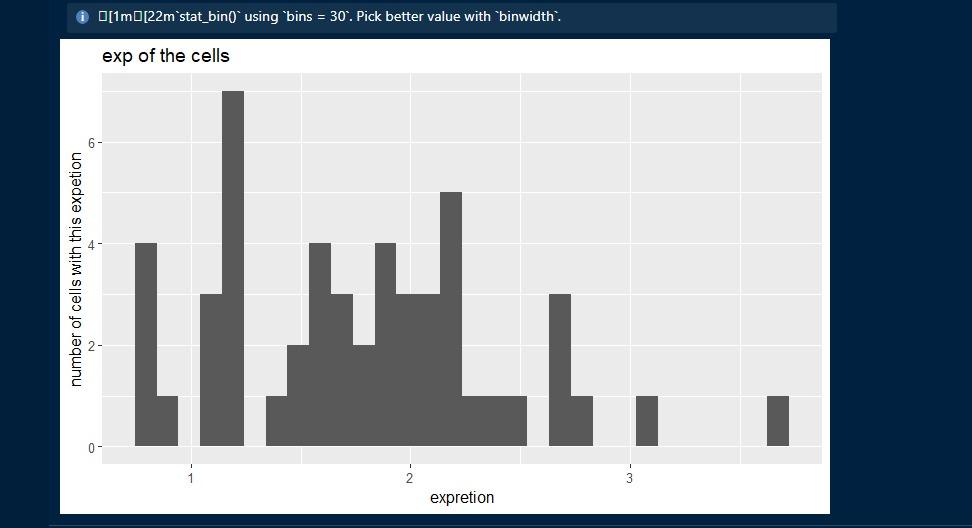


תמונה ה': התמונה הבאה מתקבלת מהרצת הפונקציה שיצרנו בפרויקט, בתמונה יש את רמת הביטוי של הגן *BCL11B* בתאים. אנחנו בדקנו את רמת הביטוי בכל התאים שבכל הקלסטרים ולא הגבלנו לקלסטר מסוים למרות שהפונקציה נותנת אפשרות להגביל את הבדיקה לקלסטרים מסוימים.

אדם 75:



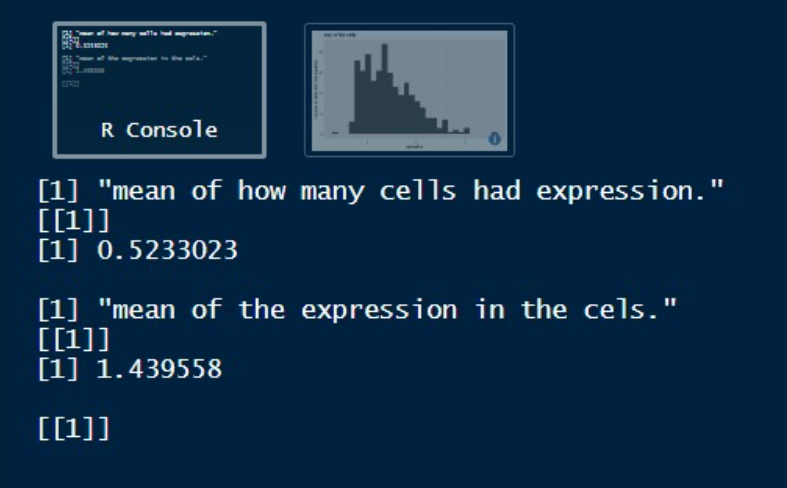
אדם 87:



תמונה ו': התוצאות הבאות מתקבלות מאותה הפונקציה שהביאה את התמונה שלפני כן. מידע זה הוא חדש והוספנו אותו לשם הסברת הנתונים בצורה מפורטת. המספר הראשון זה כמה אחוז מהתאים עברו ביטוי מעל רמת המינימום (ברירת מחדל זה 0.1).

אדם 75:

יש כ 52% תאים בדתה סט שביטאו את הגן. המספר השני, זה הממוצע של רמת הביטוי עבור כל התאים שביטאו. כל תא ביטא כמות שונה של הגן ולכן עשינו ממוצע לכל רמות הביטוי. בדוגמה שלנו, התאים בדתה סט ביטאו את הגן, ברמת ביטוי של 1.4 בממוצע.



אדם 87:

יש כ 38% תאים בדתה סט שביטאו את הגן. המספר השני, זה הממוצע של רמת הביטוי עבור כל התאים שביטאו. כל תא ביטא כמות שונה של הגן ולכן עשינו ממוצע לכל רמות הביטוי. בדוגמה שלנו, התאים בדתה סט ביטאו את הגן, ברמת ביטוי של 1.7 בממוצע.

