Neuro-Genomics

Final project - Sequencing Data Analysis

מגישים:

דניאל ברוקר 315015594

אור שחר 206582017

PART 1 - analysis of bulk sequencing

- בחלק זה של הפרויקט נעזור לחוקר שמנסה להבין את המנגנון המולקולרי של מחלה נוירולוגית.
  החולים במחלה הזו סובלים מחוסרים פיזיולוגיים ייחודיים המתבטאים בתסמינים נוירולוגיים חמורים. לפני מספר שנים התגלה כי אצל החולים במחלה זו יש מוטציה שמונעת את תרגום של חלבון מסוים. גן זה אינו מוטנטי באוכלוסייה הכללית, אך הקשר בין הגן לבין ההיבטים הפיזיולוגיים של המחלה נותר בלתי מובן.
- כדי להבין יותר על המחלה מבחינה מולקולרית, החוקר יצר מודל עכבר שבו יש "נוקאאוט" (כלומר, החלבון אינו יכול להיווצר) של הגן המסוים הזה. לאחר מכן הוא ביצע ניסוי ריצוף (bulk sequencing) באמצעות רקמות קורטקס.
  נרשמו שלוש דגימות של רקמת קורטקס משלושה עכברים נורמליים (שמסומנים "C"), ושלוש דגימות מרקמת קורטקס משלושה עכברים שבהם נעשה נוקאאוט לגן המסוים (שמסומנים "KO").
- אנחנו ננתח את נתוני הריצוף הגולמיים שהחוקר יצר, ונבין מהם החוסרים המולקולריים הקשורים לנוקאאוט של הגן. נחפש מידע על המנגנון המולקולרי של המחלה, ונחפש כיוון לטיפול אפשרי.

pre processing:

בחלק זה ביצענו עיבוד ראשוני בפייתון והשתמשנו ב-Kallisto ליצירת אינדקסים ולכימות נתוני ה-RNA. לאחר מכן, ביצענו ייבוא של הנתונים ב-R בעזרת tximport, והשתמשנו ב-DESeq2 לצורך ניתוח דיפרנציאלי של ביטוי גנים בין קבוצות ה-control וה-knockout.

נפרט על השלבים

1. **עיבוד ראשוני בפייתון:**
   - נשתמש בפייתון על מנת לבצע הכנה של הנתונים הגולמיים לריצוף RNA שנמצאים בקבצי FASTQ. המטרה היא ליישר את הקריאות שנוצרו מהריצוף ל-transcript או לגנים באופן שיאפשר ניתוח יעיל של ביטוי גנים.

2. **שימוש ב-Kallisto -יצירת אינדקסים וסיכום נתונים:**
   - Kallisto הוא כלי מתקדם ומהיר לכימות של נתוני RNA-seq, שבאמצעותו ניתן לנתח את הביטוי של גנים בצורה יעילה.
   - תהליך היישור הפסאודו-אליינמנט (Pseudo-alignment): בשלב זה, Kallisto משתמש בקבצי ה-FASTA שמכילים את הרצפים הגנטיים של ה transcripts , על מנת ליצור אינדקסים (מבנה נתונים) שמאפשרים ביצוע חיפוש מהיר אחר קריאות ה-RNA בנתונים הגולמיים (FASTQ) בתוך מאגרי הטרנסקריפטים שהתקבלו מריצוף ה-RNA.

- כימות ביטוי גנים: מתבצע חישוב ערכים כמו Transcripts Per Million(TPM), שמודדים את רמת הביטוי של כל גן, וכן ערכים נוספים.

3. **ייבוא נתונים ב-tximport ב-R**

- לאחר השימוש ב-Kallisto ליצירת אינדקסים וכימות ביטוי הגנים, נייבא את הנתונים ב-R באמצעות חבילת tximport.
- המטרה היא לבצע הכנה לניתוח הנתונים מהקליסטו באמצעות -DESeq2.
- מה עשינו בקוד: בוצע ייבוא של תוצאות ה-Kallisto תוך שימוש במיפוי tx2gene, ולאחר מכן בוצע עיגול ערכי ה-counts כדי להתאים את הנתונים ל-DESeq2.

4. **ניתוח דיפרנציאלי ב-DESeq2**

- לאחר ייבוא הנתונים בוצע ניתוח דיפרנציאלי באמצעות DESeq2 על מנת לזהות גנים שמתבטאים בצורה שונה בין קבוצות הניסוי השונות (control מול knockout).
- מה נעשה ב-DESeq2:
  נוצר אובייקט DESeqDataSet מתוך הנתונים המיובאים מ-tximport.
  בוצע עיבוד של הנתונים ב-DESeq2 לזיהוי ביטוי גנים דיפרנציאלי (Differential Gene Expression) בין קבוצות ה-control לבין קבוצות ה-knockout.
  תוצאות הניתוח כללו ערכים של log 2 fold change (הבדלים בביטוי בין הקבוצות) ו-p-value, שמסייעים להגדיר אילו גנים הראו ביטוי דיפרנציאלי משמעותי.

<u>לאחר כל השלבים הללו נקבל את הטבלה הבאה:</u> (זה רק השורות הראשונות שלה)

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>
ENSMUSG00000000001	4389.0940	0.0736535	0.0936202	0.786727	0.431442	0.999899
ENSMUSG00000000003	0.0000	NA	NA	NA	NA	NA
ENSMUSG00000000028	235.3305	-0.0399980	0.2367427	-0.168951	0.865835	0.999899
ENSMUSG00000000037	65.9520	0.3237494	0.3804228	0.851025	0.394755	0.999899
ENSMUSG00000000049	22.2449	0.3589602	0.6510935	0.551319	0.581415	0.999899
ENSMUSG00000000056	7983.2542	-0.0575610	0.0683286	-0.842414	0.399556	0.999899

<u>פירוש הטבלה:</u>

הטבלה מציגה את השוואת השינוי בביטוי הגנים בין תנאי knockout לתנאי control על פי log2 fold change.

1. baseMean: הממוצע של כמות כל גן בין כל 6 הדגימות.
2. log2FoldChange: שינוי בביטוי הגנים (log2 fold change) בין KO ל-control. ערכים חיוביים מציינים גנים שהתבטאו יותר ב-control לעומת KO, וערכים שליליים מציינים גנים שהתבטאו יותר ב-KO לעומת control.

3. lfcSE: סטיית התקן של השינוי בביטוי הגנים (log2 fold change standard error).
זהו מדד לחוסר הוודאות בהערכת השינוי בביטוי בין הקבוצות.
ערך lfcSE גבוה מצביע על כך שהשינוי בביטוי אינו מובהק מאוד, כלומר יש חוסר וודאות רב יותר בקשר לגודל השינוי האמיתי בין הקבוצות.

4. stat: סטטיסטיקת המבחן עבור כל גן (Wald test).

$$stat = \frac{log2FoldChange}{lfcSe}$$

ערך גבוה של סטטיסטיקת Wald מעיד על כך שהשינוי בביטוי בין הקבוצות גדול יחסית לשגיאה הסטנדרטית, כלומר השינוי הוא מובהק יותר.

5. P-value: מציין את הסבירות שהשינוי בביטוי הגן הוא מקרי.

6. P-adj : תיקון סטטיסטי לp-value בעקבות השגיאות שקורות בהשוואות מרובות, עוזר להעריך את השינוי בביטוי הגנים. (תיקון מסוג Benjamini-Hochberg).

Describe the molecular deficiencies in the disease:

To describe the molecular deficiencies in the knockout conditions, first detect genes that are differentially expressed between the control and the knockout.
A common p-value cutoff to detect differentially expressed genes is p-adj (i.e. p-value adjusted for multiple testing correction) which is less than 0.01-0.1, for example, p-adj<0.01, or p-adj<0.05, or p-adj<0.1.

Manually examine the genes detected as differentially expressed with the highest statistical significance (i.e. lowest p-value).

Study separately:
(a) the group of genes that were lower in expression in the treated conditions(i.e. genes that had higher expression in the control conditions compared to the knockout)

(b) the group of genes that had higher expression in the knockout conditions compared to the control.

Can you detect something common among the functions of these genes?

Perform functional analysis on all the genes (for groups a and b)

<u>לאחר שביצענו את ההכנה וקיבלנו את הטבלה הנ"ל, נבצע את השלבים הבאים:</u>

1. נסמן את הגנים עם ערכי p-adj נמוכים מ0.05.
2. נחלק את הגנים שנבחרו לשתי קבוצות:

1. גנים עם log2 fold change חיובי- כלומר גנים שהתבטאו יותר ב- control
2. גנים עם log2 fold change חיובי- כלומר גנים שהתבטאו יותר ב- knockout
3. נבצע המרה של שמות הגנים באמצעות ספריית biomaRt שנועדה להתאים בין שמות שונים של גנים המופיעים בריצופים או בנתונים הגנומים לבין שמות גנים סטנדרטיים (Symbols), כלומר **"Gene Symbols"**.

הסבר: גן עם Log2FoldChange =1 ו- p-adj = 0.05 אומר שהגן מתבטא פי 2 יותר בקבוצת הcontrol בהשוואה ל knockout-והסיכוי שההבדל הזה אקראי הוא קטן מ-5%.

נקבל את הקבוצות הבאות:

קבוצה א'- קבוצת הגנים גנים בעלי ביטוי גבוה יותר בתנאי הבקרה בהשוואה לנוקאאוט.

	ensembl_gene_id	external_gene_name
1	ENSMUSG00000004565	Pnpla6
2	ENSMUSG00000004933	Matk
3	ENSMUSG00000005373	Mlxipl
4	ENSMUSG00000007594	Hapln4
5	ENSMUSG00000013766	Ly6g6e
6	ENSMUSG00000020317	Spmap2
7	ENSMUSG00000024232	Bambi
8	ENSMUSG00000029032	Arhgef16
9	ENSMUSG00000029716	Tfr2
10	ENSMUSG00000031374	Zfp92
11	ENSMUSG00000035041	Creb3l3
12	ENSMUSG00000037095	Lrg1
13	ENSMUSG00000042102	Dmgdh
14	ENSMUSG00000045915	Ccdc42
15	ENSMUSG00000050022	Amz1
16	ENSMUSG00000051067	Lingo3
17	ENSMUSG00000051255	Gm6563
18	ENSMUSG00000056656	Apol8
19	ENSMUSG00000072720	Myo18b
20	ENSMUSG00000073102	Drc1
21	ENSMUSG00000085666	Tdg-ps2
22	ENSMUSG00000121724	Ppp1ccb

נחפש פונקציות משותפות ומכנה משותף בין הגנים:

1. גנים הקשורים למטבוליזם של שומנים וחומצות אמינו:

- Myo18b: משתתף בתהליכים הקשורים לחילוף חומרים של שומנים.
- Apol8: משתתף בתהליכי חילוף חומרים של שומנים ובהובלת ליפידים בדם.
- Dmgdh: משתתף במטבוליזם של חומצת אמינו.
- Ppp1ccb: משפיע על תהליכי מטבוליזם דרך פעילות פוספטאזית.
- Mlxipl: משתתף בבקרה על חילוף חומרים של סוכרים ושומנים.
- Creb3l3: מעורב במטבוליזם של שומנים.

2. גנים הקשורים למערכת העצבים והתפתחות עצבית:

- Lingo3: קשור לתהליכים עצביים, התפתחות ושימור מבנה מערכת העצבים.
- Pnpla6: קשור לחילוף חומרים עצבי והתפתחות מערכת העצבים.

- Creb3l3: משתתף במטבוליזם ושומנים, עם השפעה אפשרית על תהליכים עצביים.
- Drc1: משפיע על ניידות של אברונים בתוך התא.

3. <u>גנים הקשורים למבנה ותפקוד תאי:</u>

- Spmap2 :מעורב בארגון הציטוסקלטון(שלד התא).
- Ccdc42: משתתף בארגון הציטוסקלטון ובבקרה על מבנה התא.
- Arhgef16: משפיע על רה-ארגון הציטוסקלטון.
- Myo18b: משתתף בתנועתיות של תאים וארגון סיבים תאיים.
- Drc1: משפיע על ניידות פנימית של אברונים ותפקוד של מבנה תוך-תא.

<u>נסכם את התהליכים הפונקציונליים העיקריים בקבוצה א':</u>

- **הראשונה**- גנים המשפיעים על חילוף חומרים של שומנים וחומצות אמינו, מה שיכול להיות קשור לניהול האנרגיה של הגוף.
- **השנייה**- כוללת גנים שמשפיעים על תהליכים עצביים, שמירה על תפקוד נוירולוגי והתפתחות מערכת העצבים.
- **השלישית**- גנים שמשפיעים על מבנה התא ושלד התא, מה שחשוב לתנועתיות התאים והמבנה הכללי שלהם.

<u>קבוצה ב'-</u> קבוצת הגנים בעלי ביטוי גבוה יותר בתנאי הנוקאאוט בהשוואה לבקרה

	ensembl_gene_id	external_gene_name
1	ENSMUSG00000006403	Adamts4
2	ENSMUSG00000006782	Cnp
3	ENSMUSG00000009075	Cabp7
4	ENSMUSG00000021567	Nkd2
5	ENSMUSG00000022382	Wnt7b
6	ENSMUSG00000026841	Fibcd1
7	ENSMUSG00000026879	Gsn
8	ENSMUSG00000027375	Mal
9	ENSMUSG00000027858	Tspan2
10	ENSMUSG00000029798	Herc6
11	ENSMUSG00000031425	Plp1
12	ENSMUSG00000032517	Mobp
13	ENSMUSG00000032554	Trf
14	ENSMUSG00000033595	Lgi3
15	ENSMUSG00000036634	Mag
16	ENSMUSG00000037185	Krt80
17	ENSMUSG00000037625	Cldn11
18	ENSMUSG00000037674	Rfx7
19	ENSMUSG00000037984	Neurod6
20	ENSMUSG00000038173	Enpp6
21	ENSMUSG00000041607	Mbp
22	ENSMUSG00000043162	Pyurf
23	ENSMUSG00000043448	Gjc2
24	ENSMUSG00000047904	Sstr2
25	ENSMUSG00000049414	Gm5417
26	ENSMUSG00000060261	Gtf2i
27	ENSMUSG00000061086	Myl4
28	ENSMUSG00000062190	Lancl2
29	ENSMUSG00000074978	Actg-ps1
30	ENSMUSG00000076439	Mog
31	ENSMUSG00000079113	Defa43

<u>נחפש פונקציות משותפות ומכנה משותף בין הגנים:</u>

1. <u>גנים שקשורים למיאלין:</u>

- Plp1: חיוני ליצירת מיאלין במערכת העצבים המרכזית.
- Cldn11: משתתף ביצירת מחסום הדם-מוח ומיאלין בתאי עצב.
- Mog: מעורב במבנה ובשמירה על יציבות המיאלין.

- Cnp: קריטי למבנה ולתפקוד המיאלין.
- Mag: חיוני לתקשורת בין אוליגודנדרוציטים (תאי גליה שיוצרים את תאי המיילין מסביב לסיבים) ותאי עצב ולשמירה על מבנה המיאלין.
- Mal: משתתף בתהליכים של יצירת מיאלין.
- Mbp: חיוני ליצירת מיאלין במערכת העצבים המרכזית.
- Mobp: מעורב בשמירה על מבנה המיאלין.

2. <u>גנים הקשורים לתהליכים נוירולוגיים והתפתחות עצבית:</u>

- Lgi3: משתתף בהעברת אותות בתאי עצב ובהתפתחות מערכת העצבים.
- Wnt7b: תפקיד חשוב בהתפתחות עצבית ובפלסטיות תאית.
- Sstr2 : בקרה על שחרור הורמונים במערכת העצבים.
- Neurod6: קשור להתמיינות ולתפקוד נוירונים.
- Cabp7: משתתף בוויסות רמות סידן, שמשפיע על העברת אותות עצביים.
- Gjc2 : משתתף ביצירת צמתים בין תאים עצביים.

3. <u>גנים הקשורים לתהליכים של תיקון, רגולציה ושליטה על תפקוד תאי:</u>

- Adamts4: מעורב בתהליכים של תיקון רקמות ובקרה על דלקת.
- Gsn : משפיע על תנועתיות תאים, ארגון ציטוסקלטון ומוות תאי מתוכנן (אפופטוזיס).
- Enpp6 : קשור למטבוליזם של שומנים ובניית מיאלין.
- Fibcd1 : מעורב במסלולי תיקון רקמות.
- Pyurf: משפיע על מסלולי העברת אותות הקשורים למערכת העצבים ולמערכת החיסון.
- Rfx7: קשור לרגולציה של מערכת החיסון ותהליכי שעתוק של גנים חיסוניים.

<u>סיכום התהליכים העיקריים הקשורים לקבוצה ב':</u>

- יצירת מיאלין ותחזוקתו- מרכזי לתפקוד מערכת העצבים, מאפשר העברת אותות עצביים מהירה ויעילה
- התפתחות וחיווט של תאי עצב- כולל דיפרנציאציה של נוירונים ויצירת קשרים סינפטיים, חיוני ליצירת רשתות עצביות מורכבות

- גנים המעורבים בתהליכי תיקון רקמות, דלקת, תנועתיות תאית ורגולציה של תפקוד תא.

Describe the molecular functions most affected by the knockout:

קבוצה ב' (ביטוי גבוה בנוקאאוט): הפונקציות המולקולריות שהושפעו כוללות:

- פונקציות של ייצור מיילין ושל אוליגודנדרוציטים, אשר חיוניים לתפקוד המערכת העצבית (למשל *Plp1, Mog, Mbp*).
- התפתחות עצבית ופעילות נוירונים (*Neurod6, Gjc2*).

Given the molecular functions affected, what is the main deficiency that you detect in this disease?

בהתבסס על הניתוח של הקבוצות, ניתן להסיק שהחוסר המרכזי הוא בפונקציות הקשורות לתהליך היצירה של המיילין על תאי עצב ואוליגודנדרוציטים. נראה שהנוקאאוט גורם לפגיעה ביכולת של המערכת העצבית לשמור על מיילין תקין, שחשוב לתקשורת עצבית תקינה.

Can you suggest a possible treatment direction?

כיוון טיפול אפשרי יכול לכלול תמיכה בתהליכי יצירת מיילין או שיקום פעילותם של תאי אוליגודנדרוציטים (תאי גליה במערכת העצבים המרכזית שאחראיים על ייצור מיילין).

<u>כמה המלצות אפשריות שמצאנו עבור טיפול בחוסר מיאליין:</u>

- טיפולים תרופתיים- שיכולות לעזור להאט את קצב אובדן המיאלין במחלות כמו טרשת נפוצה, ושימוש בסטרואידים במינון גבוה.
- תוספי תזונה-לקיחת ויטמין D שנמצא כי רמות נמוכות שלו קשורות לסיכון מוגבר לחוסר במיאלין.
  וחומצות שומן כמו אומגה-3, שעשויות לתמוך בבריאות המיאלין.
- תזונה- דיאטה עשירה בשומנים בריאים כמו אבוקדו, אגוזים וזרעים, שיכולים לתמוך ביצירת מיאלין.
  הפחתת צריכת סוכר ופחמימות מעובדות, שעלולות לגרום לדלקת.
- פעילות גופנית- נמצא כי תרגילים אירוביים קבועים, משפרים את בריאות המוח ויכולים לתמוך בתהליכי תיקון המיאלין.
- טיפולים משלימים- דיקור סיני, מדיטציה וטכניקות הפחתת מתח, נמצאו כמסייעים בהפחתת תסמינים במחלות שפוגעות במיילין .
- טיפולים ניסיוניים- תאי גזע מזנכימליים (תאי גזע בוגרים רב-תכליתיים), נחקרים כאפשרות לעידוד תיקון מיאלין.
  תרפיות גנטיות המכוונות לשיפור יצירת המיאלין, שנמצאות בשלבי מחקר מוקדמים.

Part 2- analysis of single cell sequencing in situ

בחלק זה ננתח ביופסיה של חולת סרטן שד.

התאים מהביופסיה של הרקמה רוצפו באמצעות טכנולוגיית single cell שמשמרת את מיקומם של התאים בתוך הרקמה (situ sequencing).

המטרה שלנו היא לבדוק האם טיפול בעזרת תרופה אימונותרפיית תעזור למטופל מסוים.

Analyze the single cell data and determine the cell type of each cell. The list of 297 genes include known cell type marker genes, and this can help you in identifying the cell type of each cell:

T cells (immune cells) marker genes- CD3G, FOXP3, CD8A, CD3D, CD3E; Macrophage (immune cells) marker genes- HLA-DRA, CD68, CD4; B cells (immune cells) marker genes- IGHG1, IGHG4, IGKC, IGHM; Tumor marker genes- EGFR, GRB7, ERBB2, PGR, CD44, CD24, ALDH1A3, EPCAM, KRT19, KRT18, CDH1; Fibroblast (non-tumor and non-immune cells) marker gene- HSPG2, SULF1.

Pre-processing:

השתמשנו בחבילת seurat כמו שלמדנו בתרגול.

נעבור על השלבים שביצענו עבור עיבוד המידע הגולמי:

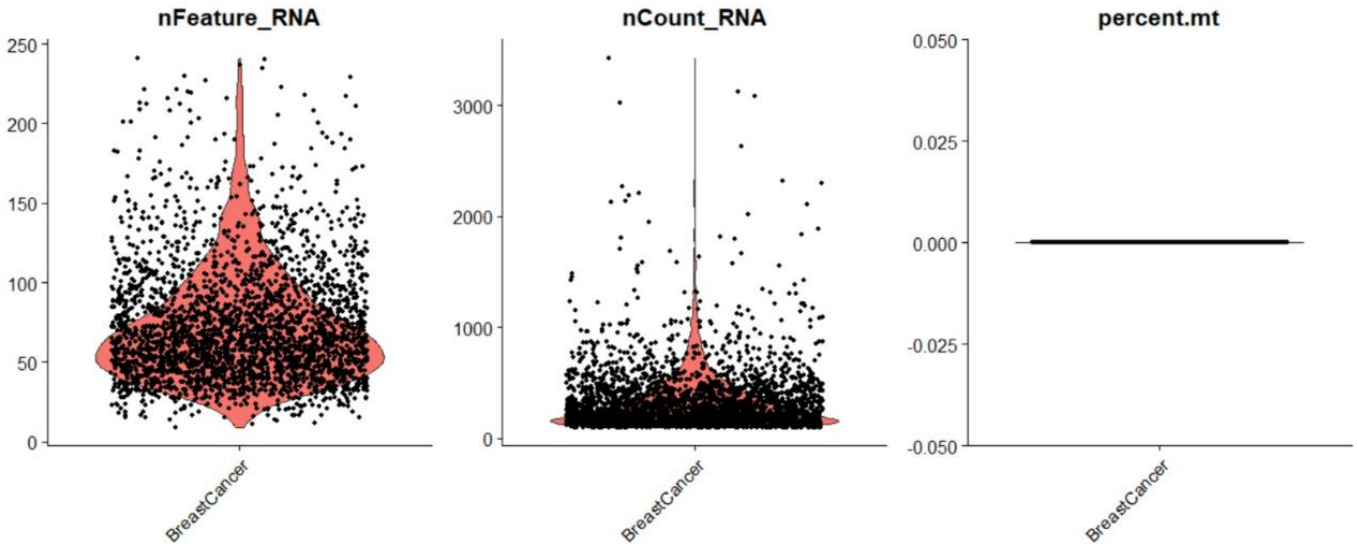
1. **ביצוע transpose** לטבלאות האקסל, כך שהמידע עבור הגנים יהיה בשורות עבור seurat.
2. **יצירת Seurat Object**

- החבילה מאפשרת לנו יצירת מבנה נתונים מאורגן לניתוח. כך שנוצר ניהול יעיל של הנתונים והדאטא הקשורים לכל תא.

3. **שלב בקרת האיכות:**

- נרצה לסנן את ה"תאים" שאינם מכילים מידע (אמיצלות ריקות או עם מעט RNA ללא תא) או לחילופין קיים בהם כפילויות (אמיצלות עם יותר מתא אחד) כך שיש להם ספירת גנים גבוהה יחסית. שלב זה בעצם מזהה ומסנן את התאים באיכות נמוכה או שלא אמינים ומבטיח שהמשך הניתוח יהיה מבוסס על נתונים אמינים, ללא רעש והטיות.

תחילה נציג את המידע המקורי לפני הסינון:



ניתן לראות 3 גרפים, כאשר כל נקודה שחורה מייצגת תא.

- הגרף הימני מייצג את אחוז הגנים המיטוכונדריים, ניתן לראות שנקבל 0% של גנים מיטוכונדריים.
  התבטאות מוגברת של גנים מיטוכונדריאליים משמשת כאינדיקציה לכך שהתא נמצא במצב לחץ או נזק.
  אם הביטוי של גנים מיטוכונדריאליים בתא מסוים גבוה מאוד, הוא עשוי להיחשב לתא שנפגע או תא "חולה" כלומר במידה והיינו מקבלים בדאטא גנים מסוג זה, היינו מבצעים סינון שלהם בשלב בקרת האיכות.

- הגרף האמצעי מייצג את הnCounet_RNA- הכמות הכוללת של רמות הביטוי של כל הגנים, עבור כל תא בנפרד.

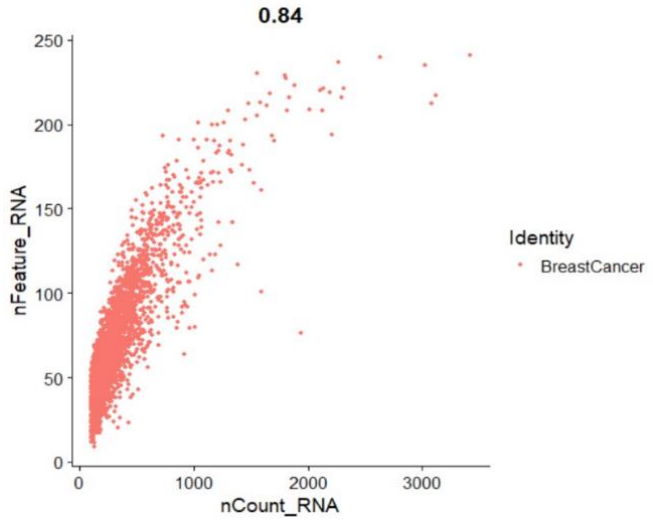
- הגרף השמאלי מייצג את הnFeature_RNA- מספר הגנים המתבטאים בכל תא.

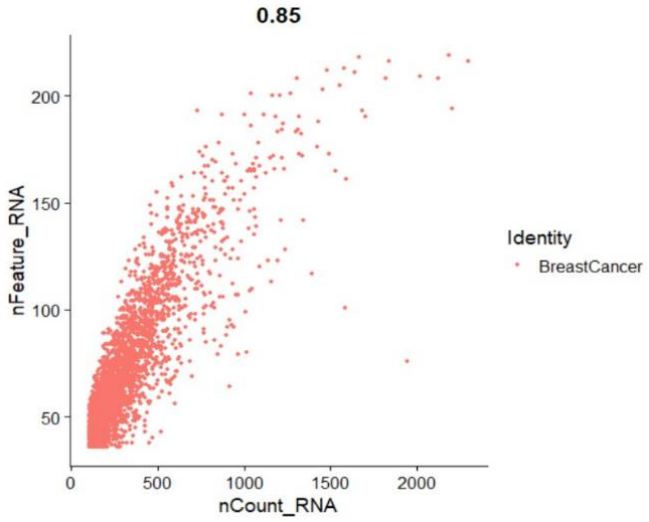
נבצע סינון כך שישארו תאים בעלי המאפיינים הבאים:

$$35 < nFeature_{RNA} < 220$$

$$nCount < 2500$$

לאחר הסינון נקבל את התוצאות הבאות:





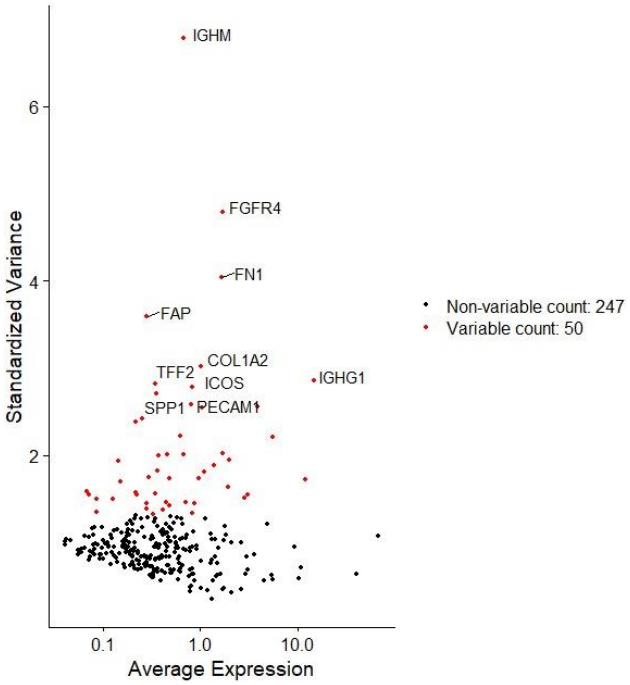
- הגרף הימני הוא לאחר הסינון, ניתן לראות שהוא יותר לינארי.
הגרפים מציגים את הקורלציה בין הnFeature_rna לבין nCoute_rna . המספר מציג את הקורלציה, תוצאה אידיאלית תהיה 1.

4. **ביצוע נורמליזציה:**

- נרצה להתאים את הנתונים לסקלה אחידה, נבצע נרמול לוקטורים כך שיהיה שווין בין תהליכי הריאקציות השונות, כלומר נסיר את ההבדלים הטכניים המפרידים ביניהם שנובעים למשל מהשוני בין כל אמיצלה או חלקיק (מספר זרועות).

5. **זיהוי גנים בעלי שונות גבוהה (סטיית תקן):**

- נסמן את 50 הגנים בעלי השונות הגבוהה ביותר, ונציין את השמות עבור 10 הגנים בעלי השונות הכי גבוהה. גנים אלו מעניינים אותנו יותר בגלל שהם משתנים יותר בין תא לתא.
שלב זה מפחית רעש ומשפר יעילות חישובית.



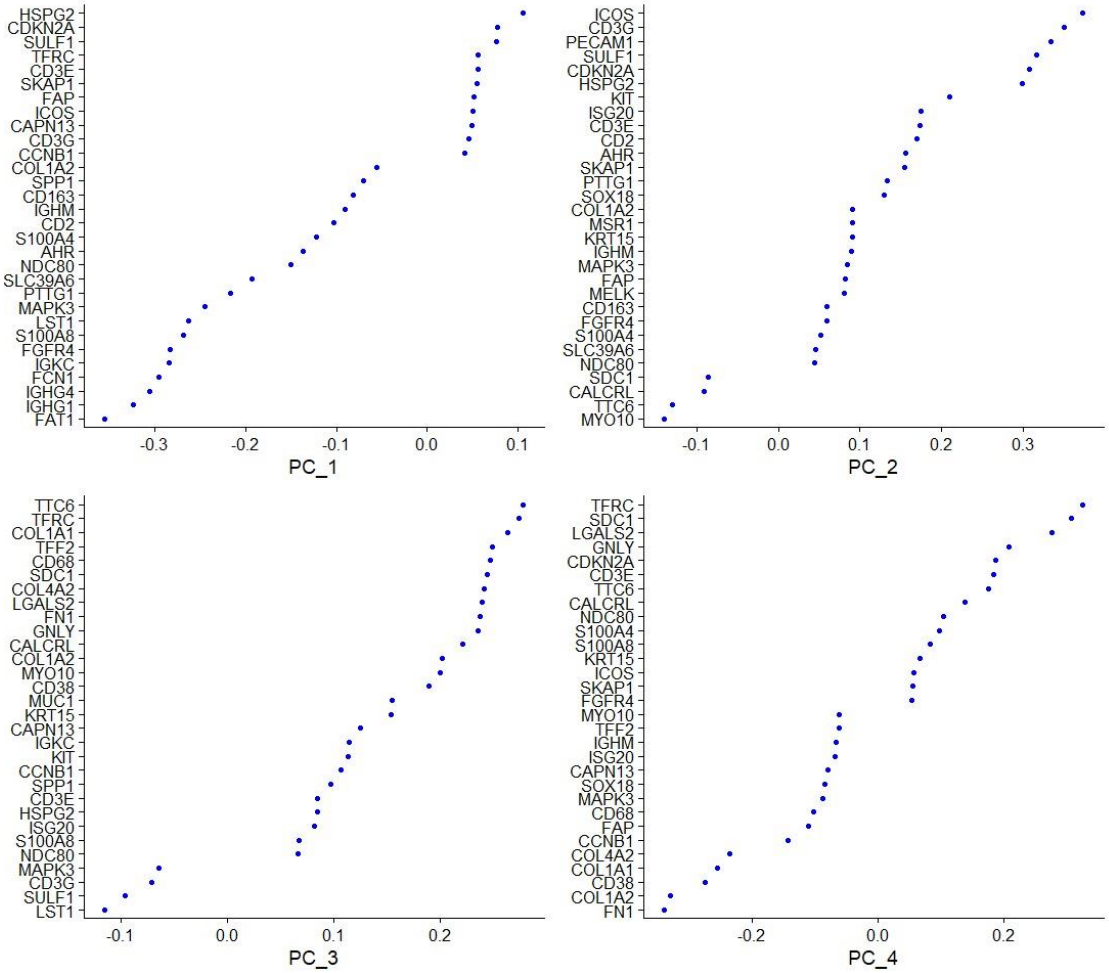
6. **<u>ביצוע scaling לנתונים:</u>**

- נבצע סטנדרטיזציה של הנתונים, התהליך משנה את הביטוי של כל גן כדי שלא נקבל החלטה לסיווג התא רק על סמך הגנים החזקים ביותר בתא.
לאחר הנרמול נקבל שהממוצע של כל גן על פני כל תאים הוא 0 ,הרמות ביטוי יהיו לרוב בין -1 ל 1 והשונות בין התאים היא 1.

7. **<u>תהליך PCA:</u>**

- תהליך הPCA נועד להפחית מימדים תוך שמירה על מידע מרבי. התהליך מאפשר ויזואליזציה וניתוח של נתונים מורכבים במרחב פשוט יותר (לדוגמה דרך הצגת הקלאסטרים בדו-מימד).
המטרה היא לזהות את הרכיבים (PC) בנתונים שמסבירים את מרב השונות. בחירת מספר רכיבים ותהליך הפחתת המימדים מפחית רעשים.
בסופו של דבר התהליך הוא שילוב לינארי של המימדים המקוריים (הגנים שלנו) וקבלת מימדים חדשים שמסבירים את ההבדל בצורה טובה ופשוטה יותר להבנה בין התאים.

<u>נסתכל על ניתוח הרכיבים העיקריים:</u>



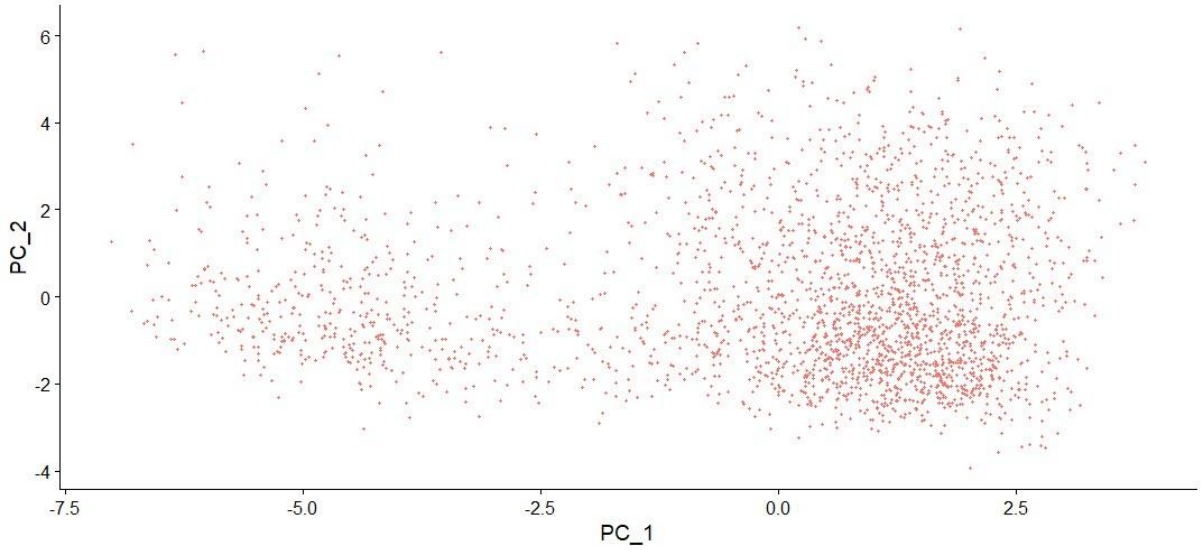
- הגרפים מציגים את הגנים המשפיעים ביותר על כל אחד מהרכיבים העיקריים, נציג את ארבעת הPC הראשונים.
כל נקודה בגרף מייצגת גן, המיקום שלה מייצג את התרומה של הגן לשונות בכל רכיב.

נדפיס את הגנים בעלי המשקל החיובי והשלילי הגבוהים ביותר מהגרפים הנ"ל:

```
PC_ 1
Positive:  HSPG2, CDKN2A, SULF1, TFRC, CD3E
Negative:  FAT1, IGHG1, IGHG4, FCN1, IGKC
PC_ 2
Positive:  ICOS, CD3G, PECAM1, SULF1, CDKN2A
Negative:  MYO10, TTC6, CALCRL, SDC1, S100A8
PC_ 3
Positive:  TTC6, TFRC, COL1A1, TFF2, CD68
Negative:  LST1, SULF1, CD3G, MAPK3, FAT1
PC_ 4
Positive:  TFRC, SDC1, LGALS2, GNLY, CDKN2A
Negative:  FN1, COL1A2, CD38, COL1A1, COL4A2
```

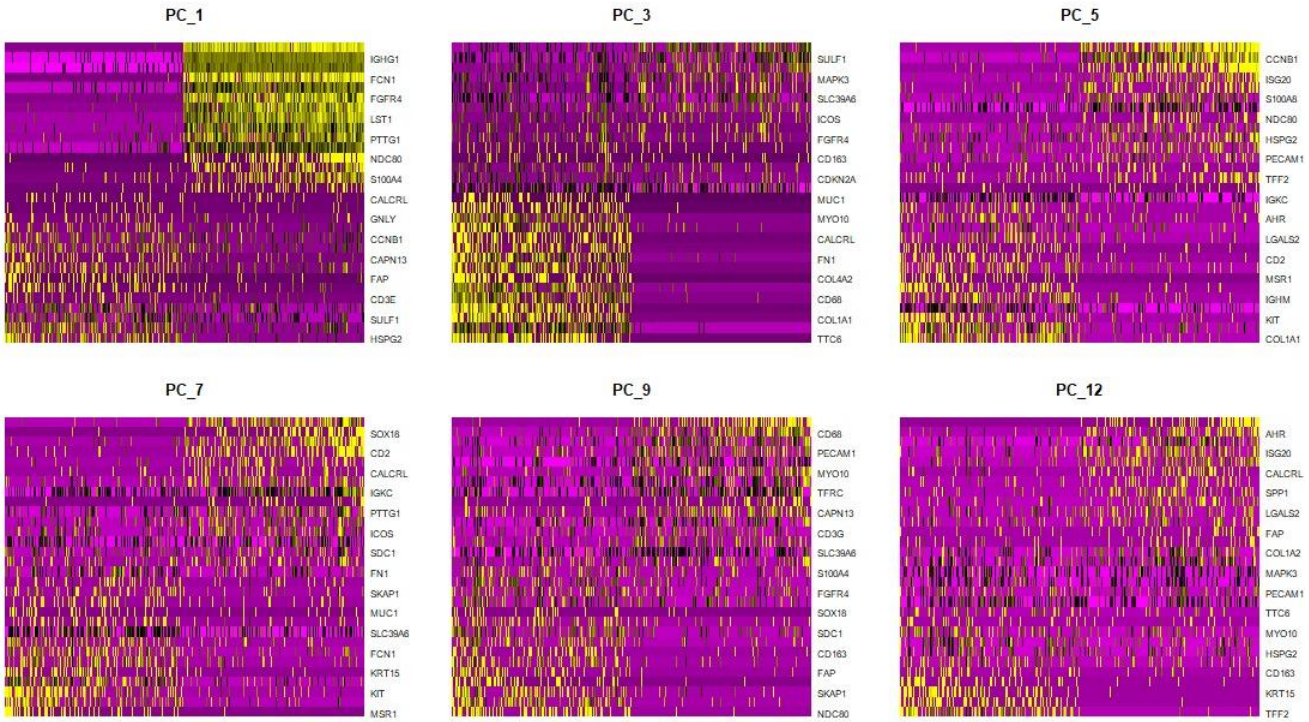
חלק זה מסייע לנו להבין מהן הגנים בעלי התרומה המשמעותית לשונות, מאפשר הבנה ראשונית לגבי הקשרים בין הגנים.

<u>פיזור התאים במרחב הPCA:</u>



- הגרף הבא מציג פיזור תאים במרחב ה-PCA על בסיס שני רכיבים עיקריים בלבד, PC1 ו-PC2.
הפיזור מראה התפלגות של התאים לפי הרכיבים הראשיים.
ניתן לזהות בגרף אשכולות של תאים שמתפצלים בהתאם לשונות בנתונים. כלומר, ניתן לזהות קבוצות שונות של תאים שעשויות להוות סוגי תאים שונים או תת-קבוצות.

נציג את התהליך של כמות רכיבי הPCA על השונות הגנים בתאים:



- כל עמודה מייצגת תא, וכל שורה מייצגת גן.

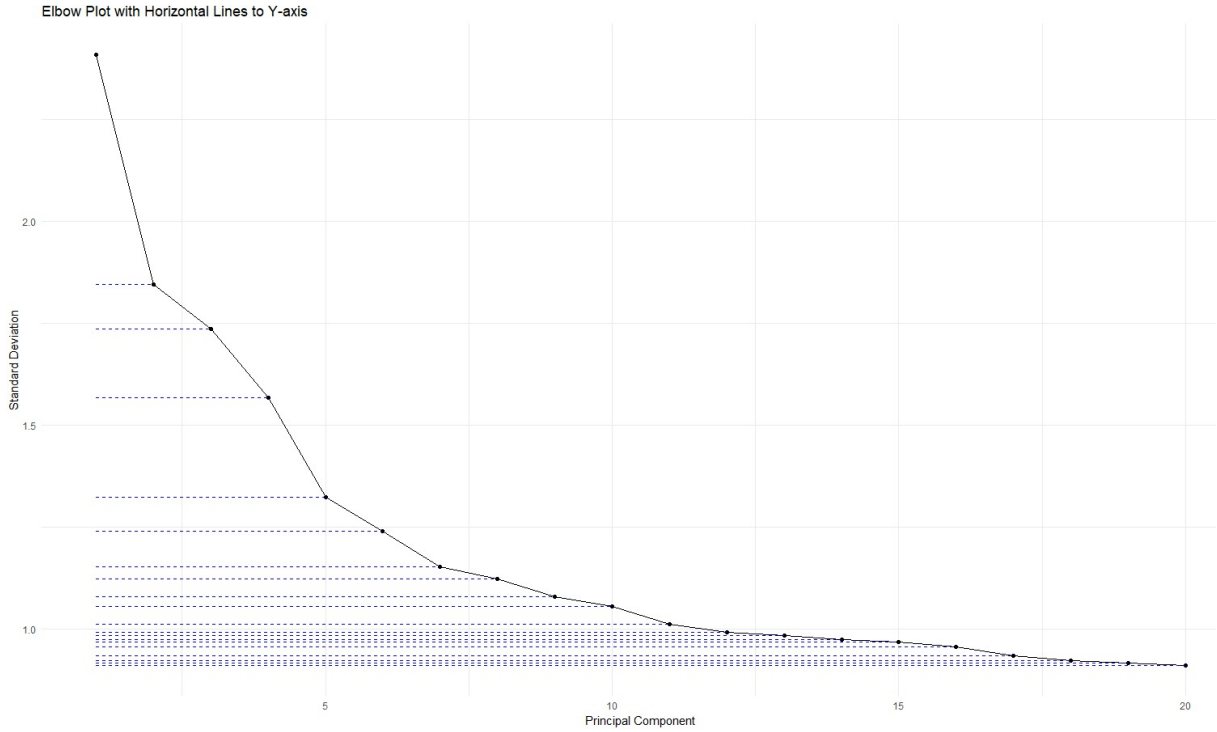
ריבועים בצהוב מייצגים רמת ביטוי- גבוהה, סגול- נמוכה, שחור-0.

ניתן לראות שככל שניקח PC יותר גבוה התוצאה פחות ברורה, כלומר קשה יותר להפריד בין גנים עם ביטוי גבוה לנמוך (לבצע ניתוח של התאים אל מול הביטוי של הגן בתא).

הגרף מייצג 400 תאים בלבד אך ממחיש טוב את ההבדל.

8. **<u>קביעת מספר המימדים:</u>**

- נבצע דירוג של רכיבים עיקריים המבוסס על אחוז השונות שכל רכיב מציג.
  נשתמש בפונקציית ElbowPlot.



- בגרף ניתן לראות שבכל שנשתמש ביותר רכיבים עיקריים (יותר מימדים) סטיית התקן יורדת.
  נרצה לבחור את מספר הרכיבים הקטן ביותר אך שסטיית התקן תהיה מספיק נמוכה.
  נשים לב שהרכיב שאנחנו בוחרים מקיים הבדל מהותי בין סטיית התקן שלו לסטיית התקן של הנקודה הקודמת אליו, כלומר לא נרצה לבחור את PC=20 כיוון שאין תרומה משמעותית בין נקודה 19 לנקודה 20.
  במילים אחרות, סטיית התקן מפסיקה לרדת בשלב כלשהו, ושם נרצה לעצור את מספר הרכיבים שאנחנו לוקחים בחשבון.
  במידה וניקח יותר מידי רכיבים יהיו רעשים מיותרים ותרומה מועטה.

- לפי הגרף נבחר PC=12

שלב עיבוד הנתונים:

1. **חלוקה לקלאסטרים:**

- נזהה קבוצות תאים דומות.
- נשתמש באלגוריתם KNN למציאת K השכנים הקרובים ביותר, התהליך מתבסס על בניית גרף שמייצג את המרחק האוקלידי בין התאים במרחב הPCA שבחרנו, לאחר מכן נקבץ לפי גנים שדומים. שלב זה מבוצע באמצעות הפונקציה FindNeighbors.
- נשתמש באלגוריתם Louvian כדי לסווג את התאים לקלאסטרים.

לחלק זה נשתמש בפונקציה FindClusters שמכילה פרמטר רזולוציה שקובע את ה'פירוט' של האשכול , ככך שהרזולוציה גבוהה יותר נקבל יותר קבוצות.

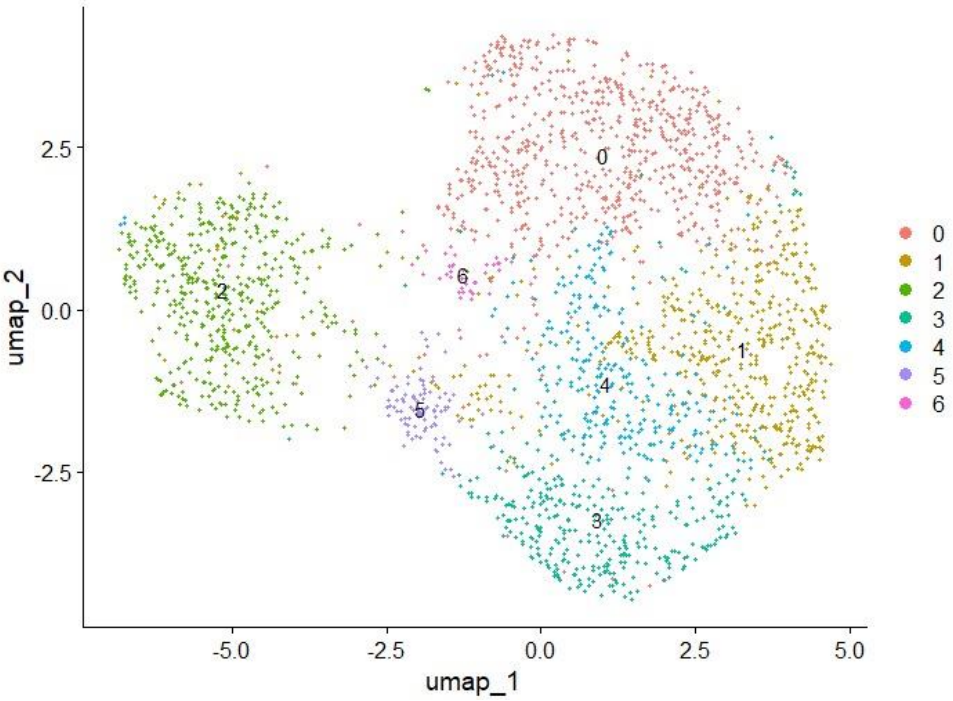
נבחר רזולוציה=0.6.

2. **ביצוע ויזואליזציה של הנתונים:**

- נשתמש בטכניקת UMAP - הפחתת מימדים בצורה לא לינארית.

הגרף שנקבל מציג את תוצאות החלוקה לקלסטרים בתצוגה דו מימדית, כלומר הצגה של קבוצות התאים שיש ביניהם דפוסי ביטוי דומים.

שלב זה מאפשר הבנה של הנתונים והסקת מסקנות להמשך.



3. **<u>סיווג הקלסטרים:</u>**

- זיהוי גנים המבוטאים באופן דיפרנציאלי בין הקלאסטרים, בעזרת בפונקציית FindAllMarkers.

  Seurat משתמש בניתוח ביטוי דיפרנציאלי (DE) כדי למצוא גנים אלו.

  ברירת המחדל היא להשוות קלאסטר אחד לכל שאר התאים.

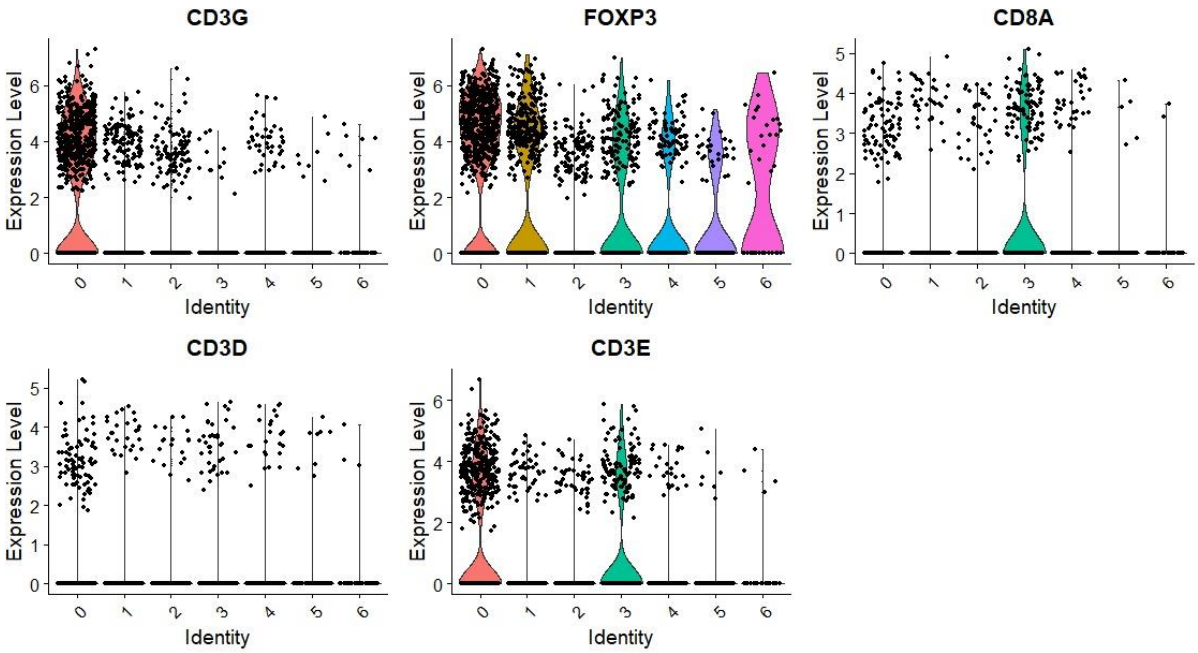
<u>נקבל את הטבלה הבאה עבור כל הגנים:</u>

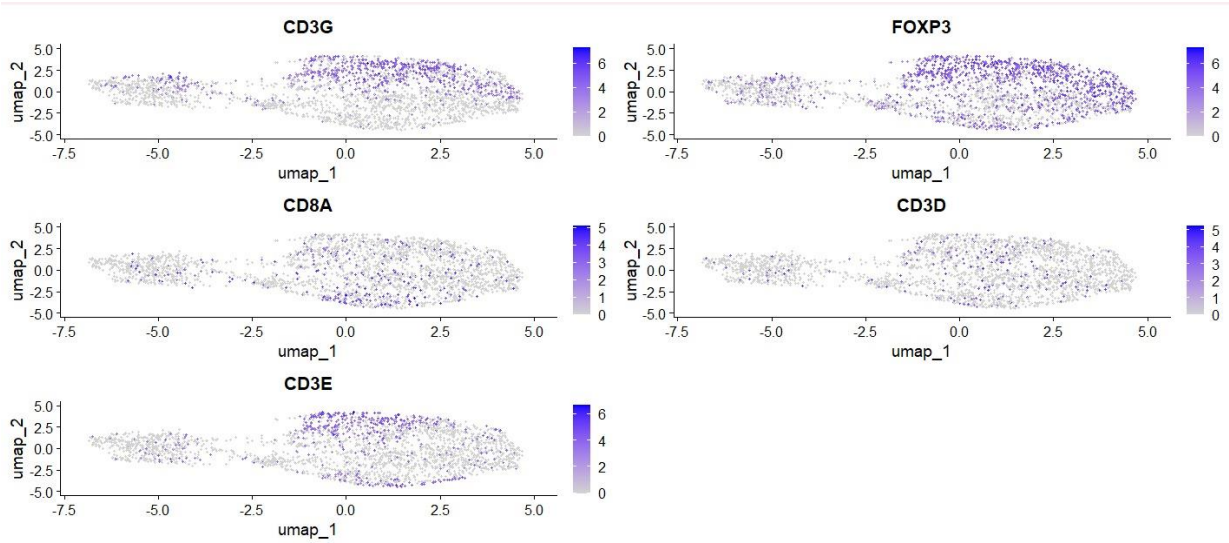
	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
CLDN4	2.522249e-09	0.4118989	0.266	0.153	7.491080e-07	0	CLDN4
STMN1	6.320880e-08	0.3523712	0.345	0.220	1.877301e-05	0	STMN1
XCL1	1.646728e-07	0.4878938	0.356	0.236	4.890781e-05	0	XCL1
MMP12	1.991228e-07	0.3540837	0.365	0.245	5.913948e-05	0	MMP12
AURKA	2.931414e-07	0.4180843	0.332	0.218	8.706299e-05	0	AURKA
TTYH1	6.192690e-07	0.3520248	0.310	0.201	1.839229e-04	0	TTYH1
TFEC	6.632718e-07	0.3918287	0.358	0.240	1.969917e-04	0	TFEC
HLA-DRB5	7.429813e-07	0.3131738	0.265	0.166	2.206654e-04	0	HLA-DRB5
CD24	9.781073e-72	0.7651046	0.995	0.979	2.904979e-69	1	CD24

- p value- מודד מובהקות סטטיסטית של הגן באשכול, ערכים נמוכים מצביעים על כך שהגן מבוטא באופן משמעותי באשכול.
- Avg_log2FC- ערך הlog2 fold change שמבטא את ההבדל בביטוי הגן בין האשכול הנוכחי לשאר האשכולות. ערכים חיובים מצביעים על ביטוי גבוה של הגן באשכול הספציפי.
- Pct.1- אחוז התאים באשכול שמבטאים את הגן.
- Pct.2- אחוז התאים בשאר האשכולות שמבטאים את הגן.
- P_val_adj- ערך הp value מתוקן להשוואות מרובות. התיקון נעשה בעזרת שיטת FDR (שיטה סטטיסטית לתיקון בעיית ההשוואות המרובות).
- Cluster – מספר הקלאסטר שהגן משויך אליו.

כעת כדי לבצע את הסיווג נשתמש ב violin plot שמציג את רמת הביטוי עבור גן מסויים בכל אשכול, ו-FeaturePlot שמציג את פיזור הגן הנבדק בתצוגת דו מימד.

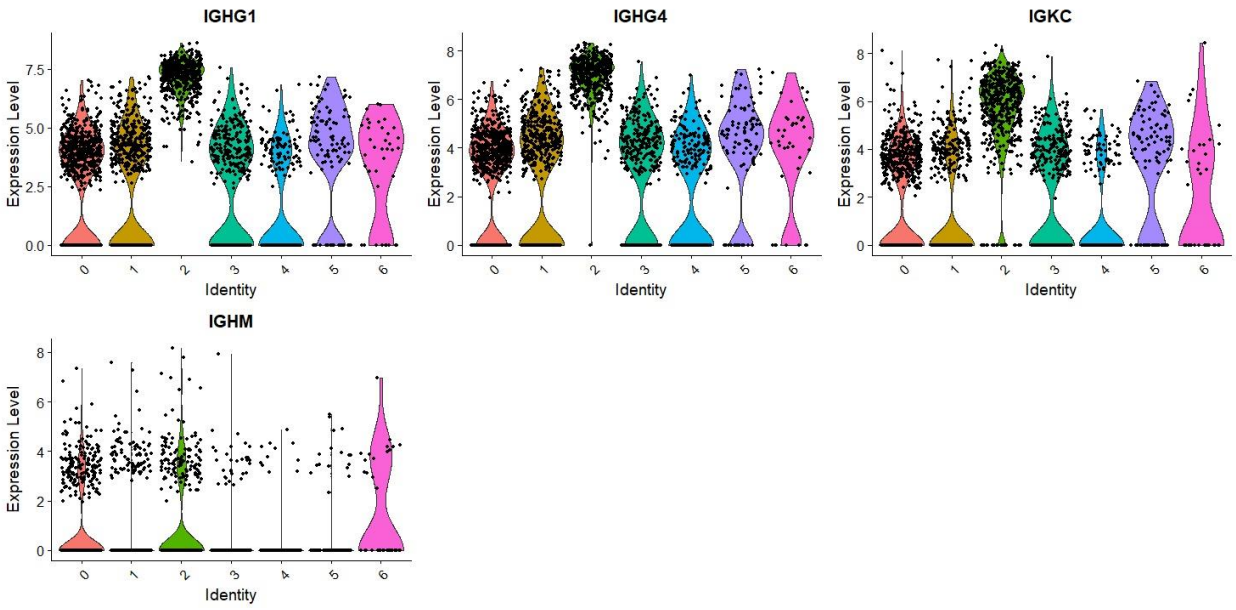
<u>עבור מרקרים של T cell:</u>

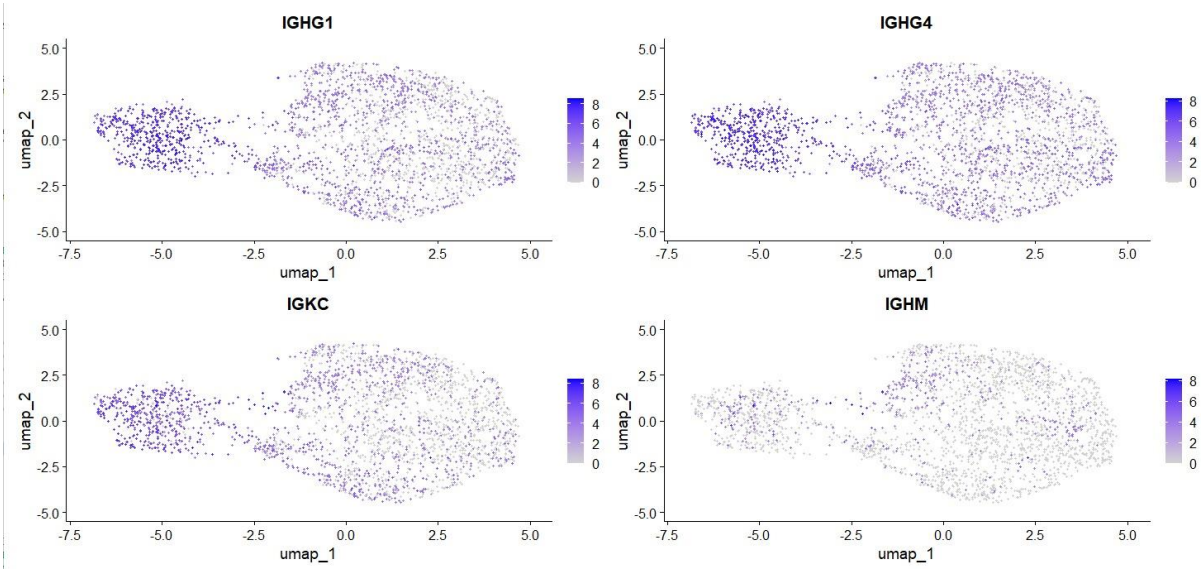




- ניתן לראות שביטוי גנים מסוג T cell נמצאים בעיקר בקלסטר 0.

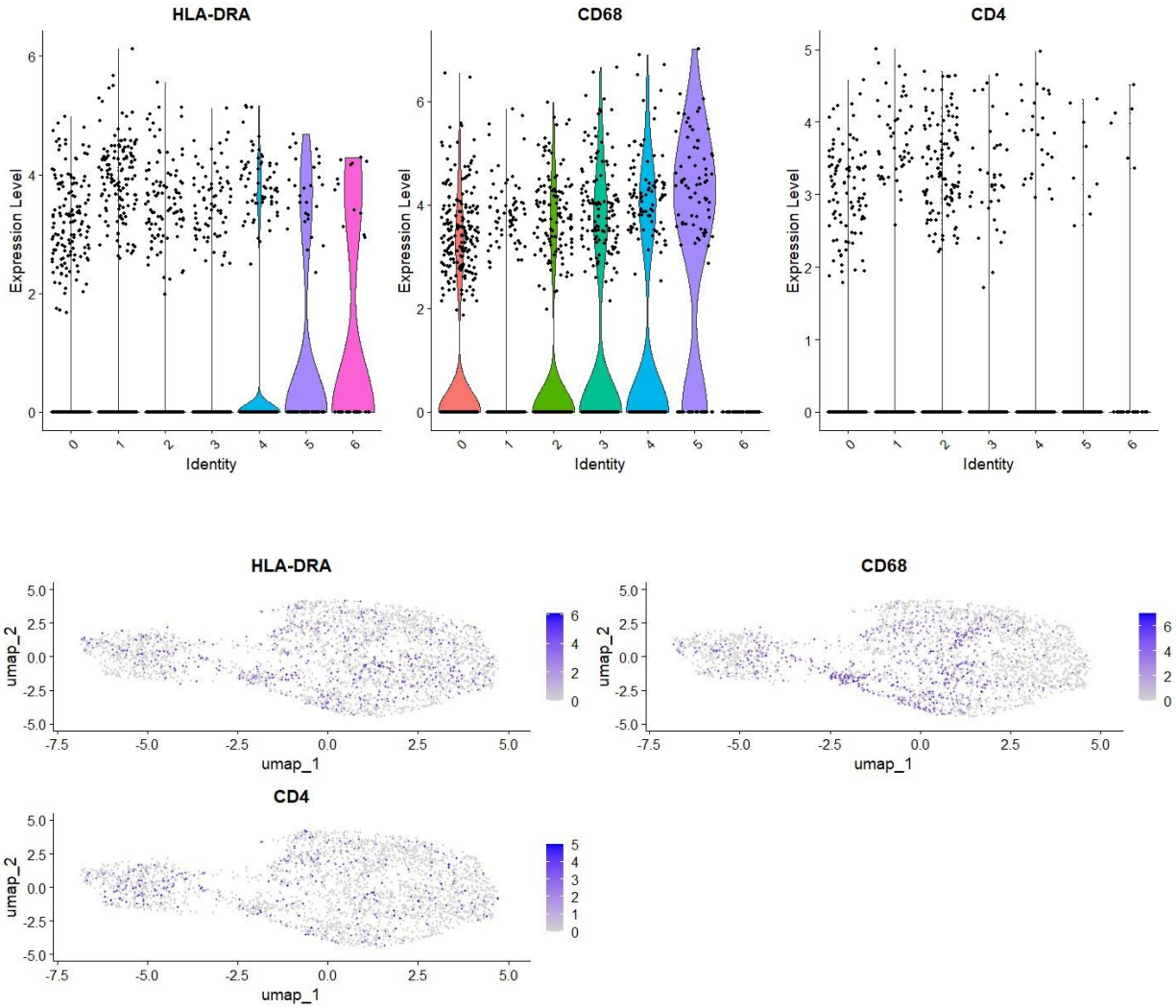
עבור מרקרים של B cell:





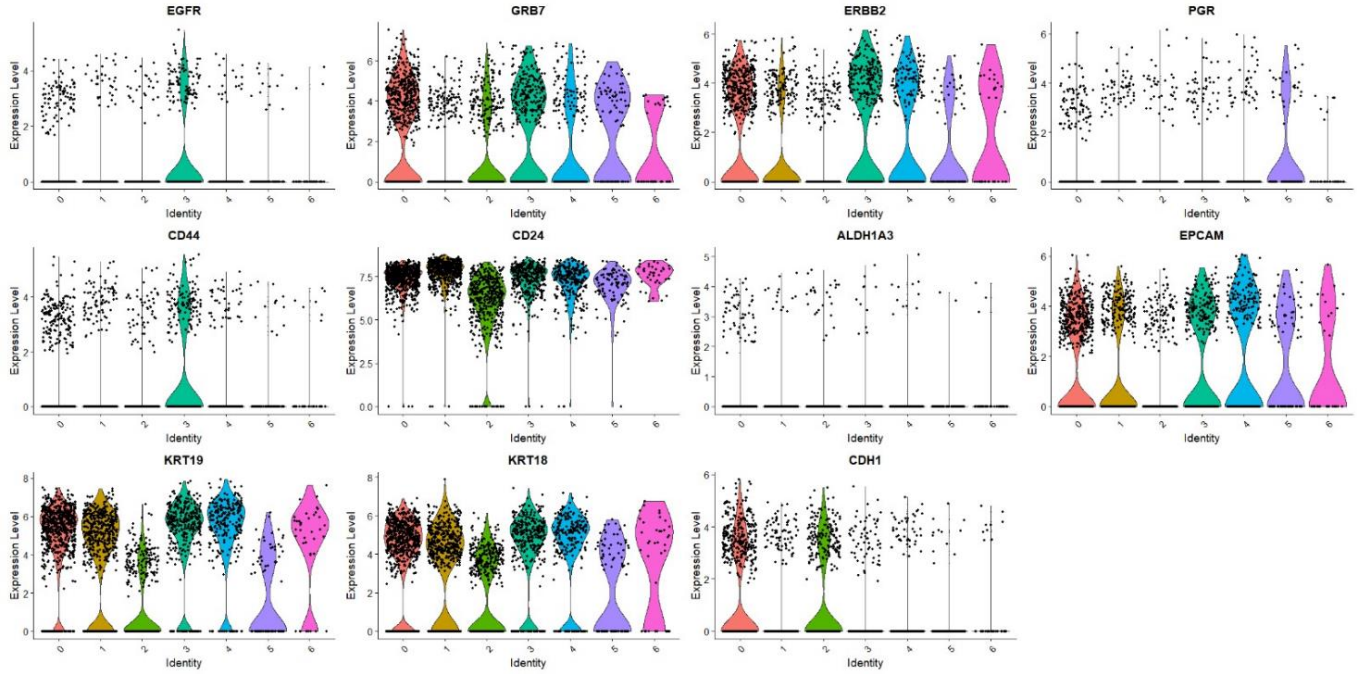
- ניתן לראות שביטוי גנים מסוג B cell נמצאים בעיקר בקלסטר 2.

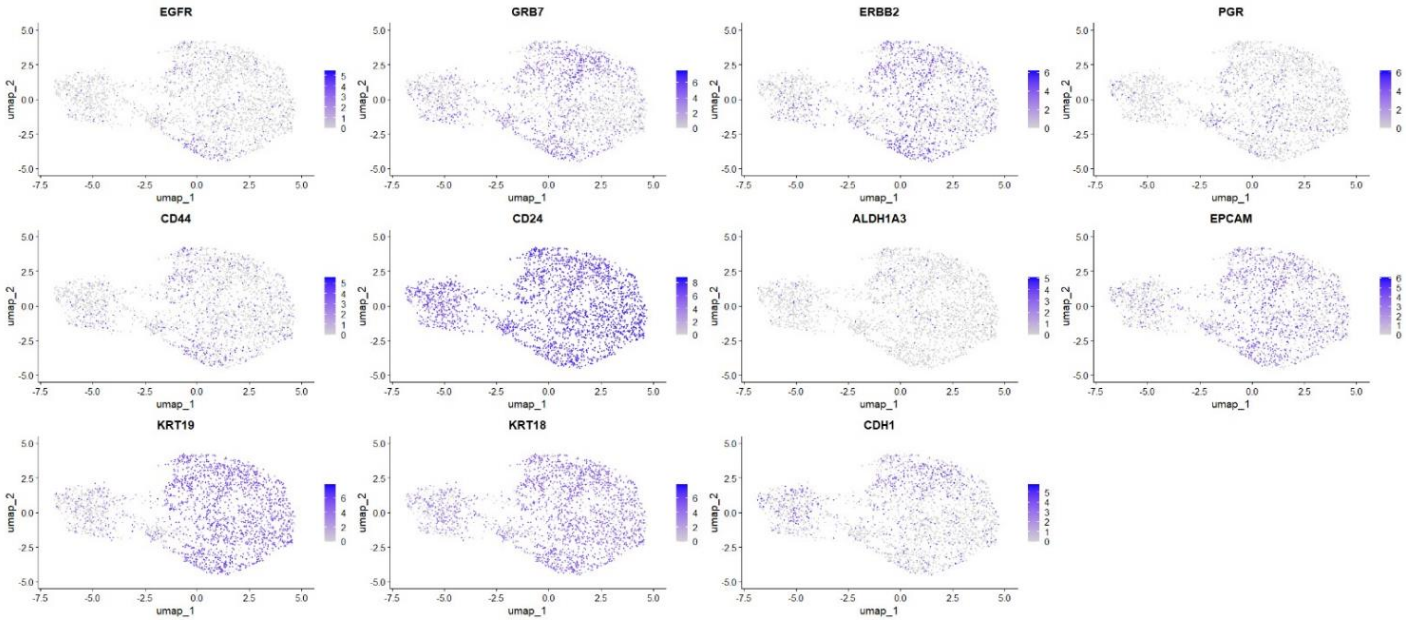
עבור מרקרים מסוג Macrophage:



- ניתן לראות שביטוי גנים מסוג Macrophage נמצאים בעיקר בקלסטר 5.

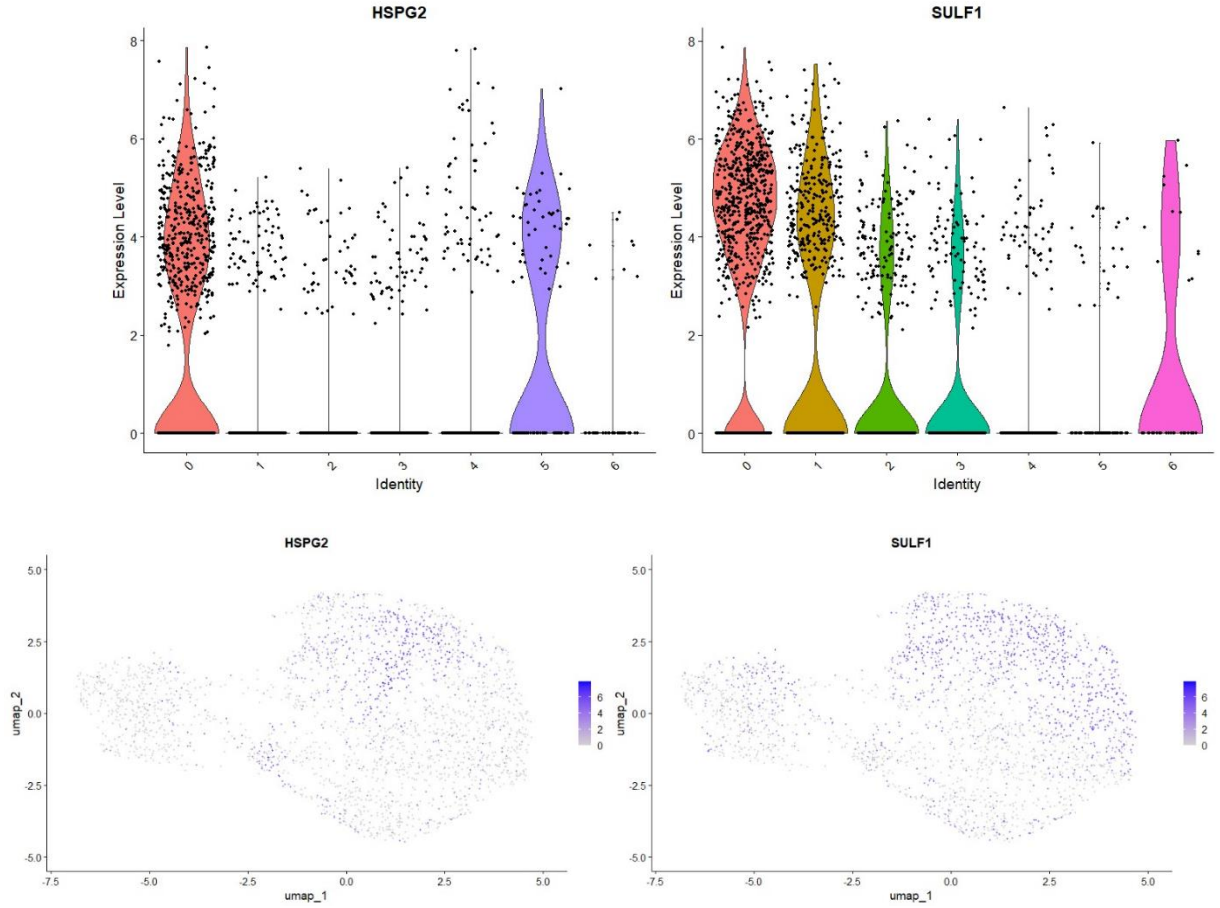
עבור מרקרים מסוג Tumor marker genes:





- ניתן לראות שביטוי גנים מסוג tumor מבוטאים בעיקר בקלסטרים 4,3.

עבור מרקרים מסוג Fibroblast:



- ניתן לראות שביטוי גנים Fibroblast מסוג נמצאים בעיקר בקלסטר 0.

נשים לב שבדרך זאת קשה להסיק במדויק מהו סוג כל קלאסטר, לכן בנוסף לאבחנה הנ"ל נבצע בדיקה של שלושת הגנים (מתוך קבוצת המרקרים הנתונה בשאלה) בעלי ערך average log2 fold change הגבוה ביותר בכל קלאסטר, כלומר הגנים שמתבטאים הכי חזק בכל קלסאטר.

נקבל את התוצאה הבאה:

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>	<chr>
1	4.02e-133	2.76	0.649	0.161	1.19e-130	0	CD3G
2	3.06e-108	2.08	0.612	0.165	9.09e-106	0	HSPG2
3	1.91e- 62	2.08	0.435	0.129	5.66e- 60	0	CD3E
4	9.78e- 72	0.765	0.995	0.979	2.90e- 69	1	CD24
5	4.42e-248	4.42	1	0.616	1.31e-245	2	IGHG1
6	3.83e-239	4.01	0.998	0.686	1.14e-236	2	IGHG4
7	9.53e-197	3.54	0.954	0.46	2.83e-194	2	IGKC
8	2.00e- 23	2.02	0.267	0.089	5.94e- 21	3	EGFR
9	3.70e- 21	1.75	0.267	0.094	1.10e- 18	3	CD8A
10	8.84e- 16	1.32	0.351	0.175	2.63e- 13	3	CD44
11	3.63e- 10	1.19	0.535	0.406	1.08e- 7	4	EPCAM
12	1.13e- 5	1.04	0.367	0.263	3.34e- 3	4	CD68
13	8.27e- 22	0.967	0.895	0.716	2.46e- 19	4	KRT19
14	4.27e- 30	2.39	0.719	0.256	1.27e- 27	5	CD68
15	1.00e- 5	1.48	0.26	0.118	2.98e- 3	5	PGR
16	8.00e- 5	0.414	0.458	0.285	2.38e- 2	5	HSPG2
17	1.29e- 3	1.39	0.385	0.192	3.84e- 1	6	IGHM

מסקנות מהטבלה:

קלסאטר 0 הוא מסוג T cell

קלסאטר 1 הוא מסוג Tumor cell

קלסאטר 2 הוא מסוג B cell

קלסאטר 3 הוא מסוג Tumor cell

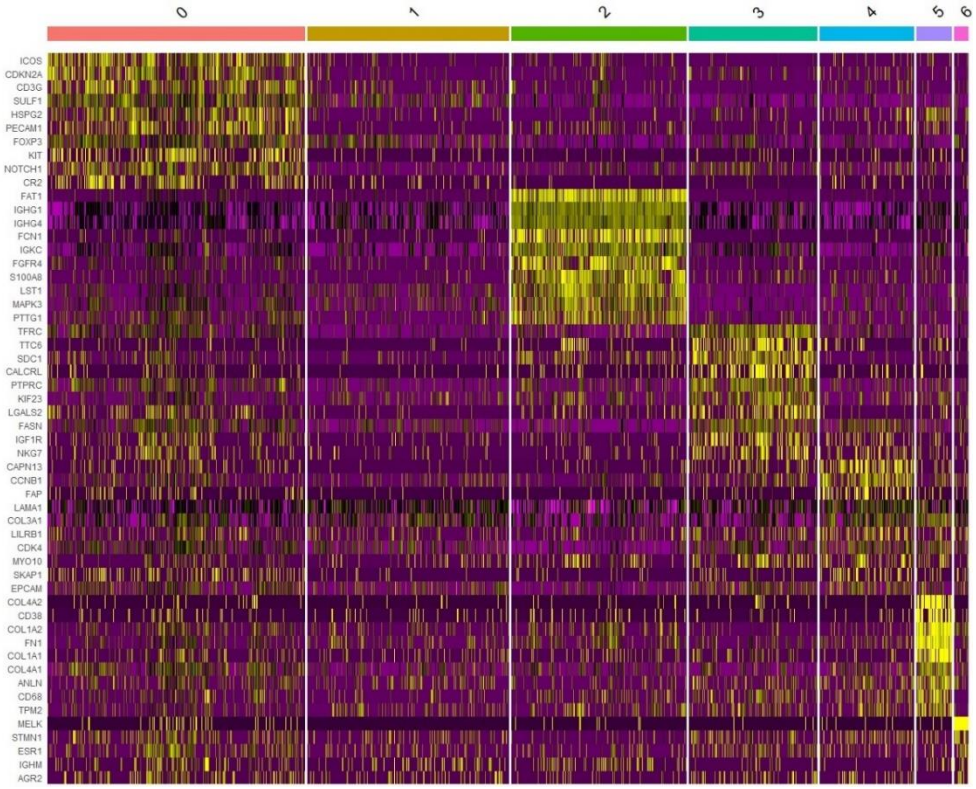
קלסאטר 4 הוא מסוג Tumor cell

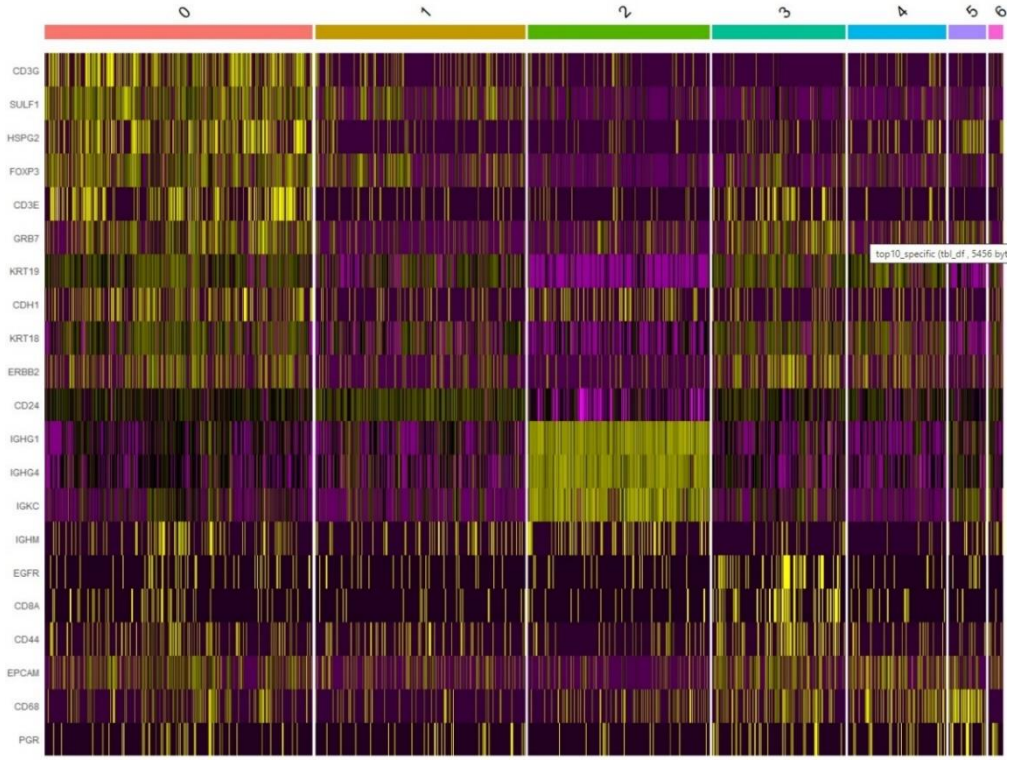
קלסאטר 5 הוא מסוג Macrophage

קלסאטר 6 הוא מסוג B cell

נדפיס גם גרף מסוג heat map כלומר, ביטוי הגן עבור כל קלאסטר:

כל שורה זה גן וכל עמודה מייצג תא, העמודות מחולקות לפי הקלסאטרים.
נקודות צהובות מייצגות רמת ביטוי י גבוהה של גן מסויים בקלסטאר מסויים.

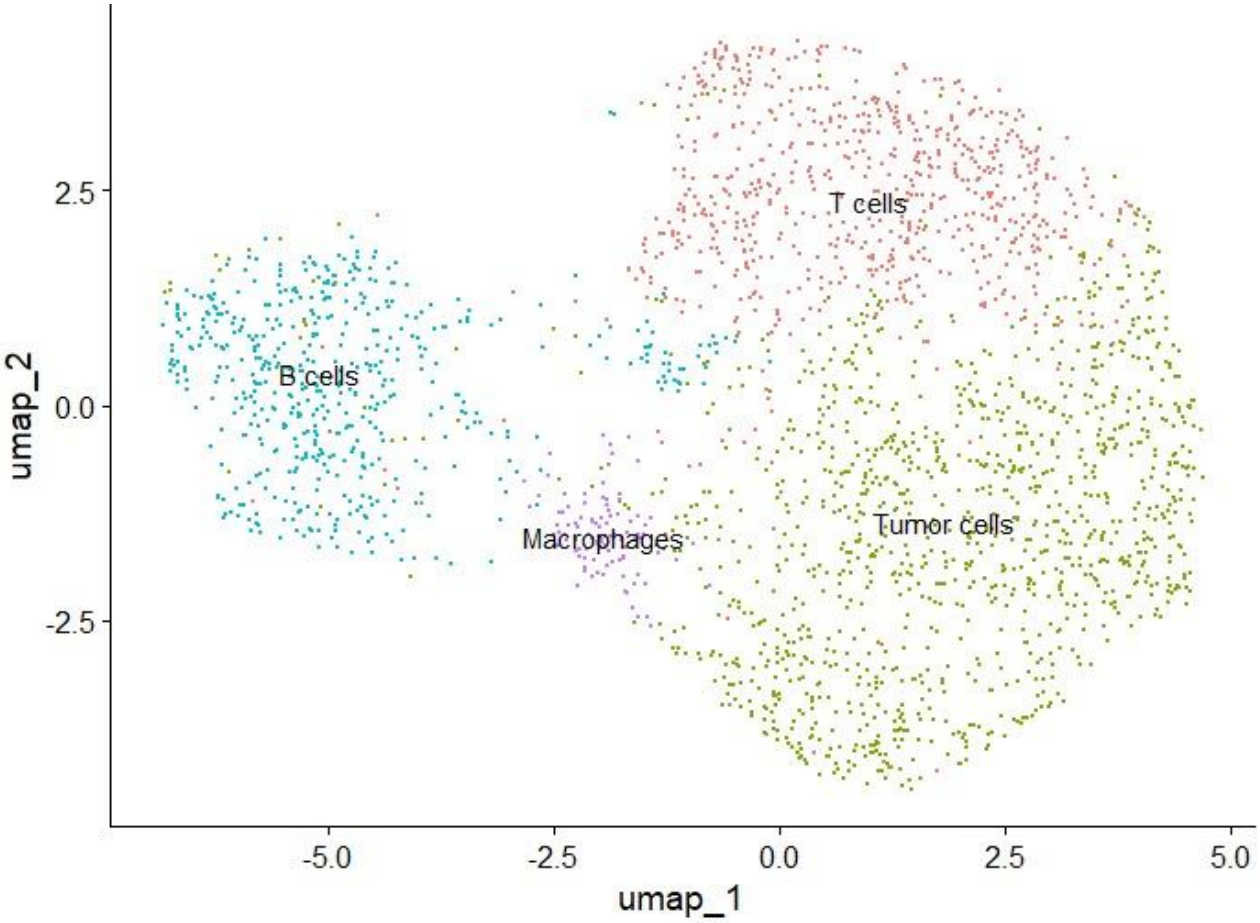




- התמונה הראשונה מציגה את ביטויי כל הגנים(לא ניתן להציג את כולם, אבל הכוונה זה לקבוצת ה297 גנים- בתמונה רואים את הגנים בעלי הביטוי הגבוה יותר) ובתמונה השנייה מציגה רק את הגנים הספצייפים שנקבעו להיות cell type markers.

- ניתן לראות שהתוצאות מתיישבות עם המסקנות שלנו לסיווג הקלאטרים.
למשל ניתן לראות שבקלסטר 2 יש רמות ביטוי גבוהות עבור הגנים : IGHG1,IGHG4,IGKG שהם מרקרים מסוג B cell.
בקסלטר 0 יש רמת ביטוי גבוהה עבור גנים: CD3G ,FOXP3 שהם מרקרים מסוג T cell.

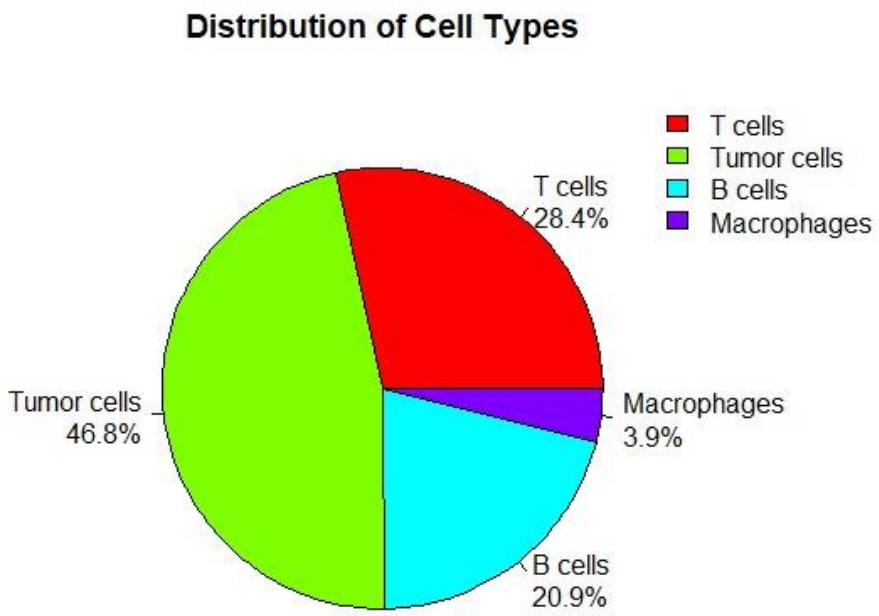
לסיכום- חלוקת הקלאסטרים תהיה בצורה הבאה:



שאלות

1. **Does the sample contain at least 10% immune cells (from the total number of cells studied)?**

נדפיס גרף עוגה שמראה את החלוקה של סוגי התאים:



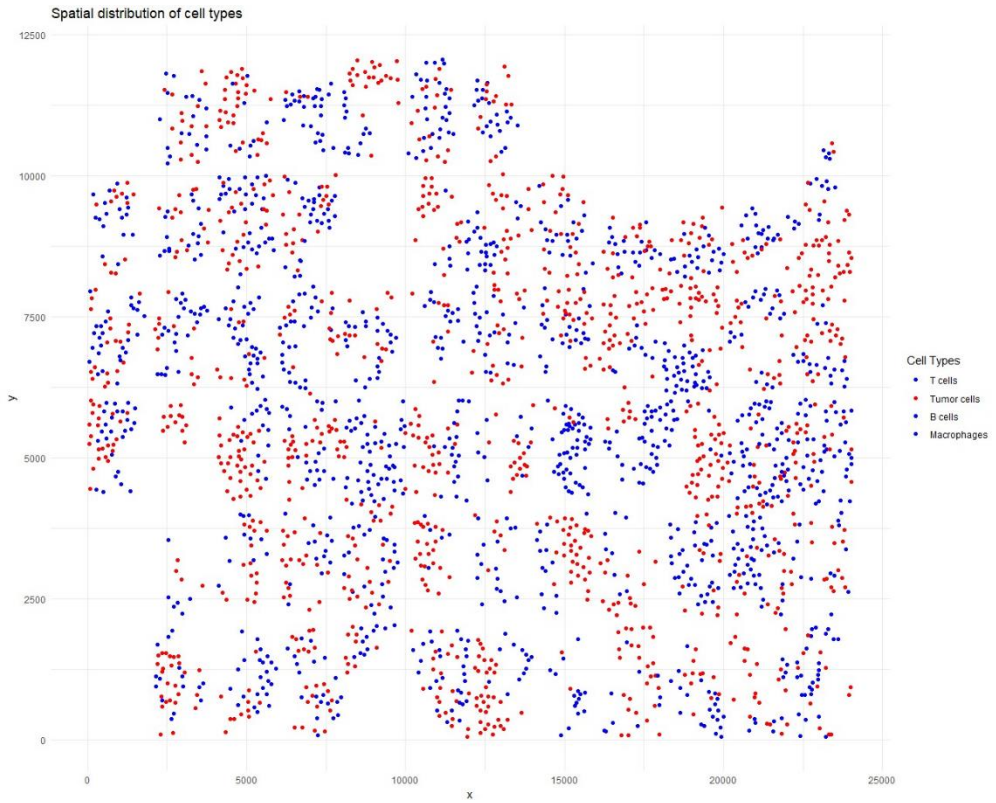

```
    T cells Tumor cells     B cells Macrophages
        699        1152         514          96
> print(cell_percentages)

    T cells Tumor cells     B cells Macrophages
  28.403088   46.810240   20.885819    3.900853
>
> |
```

נקבל שיש 53.2% מסוג immune

2. **Are the immune cells mixed with the tumor cells in the biopsy? The alternative is that the immune cells reside in one location inside the biopsy and the tumor cells reside in a spatially different location.**

בשביל לבדוק האם תאי החיסון ותאי הסרטן מעורבבים, נשתמש בקובץ מיקומי התאים.



בגרף ניתן לראות כי אכן קיים ערבוב בין הimmune cell (הנקודות הכחולות) בין תאי הסרטן (הנקודות האדומות)- כך שהסיכוי של הטיפול להצליח גדל.

3. Are at least 10% of the cells in the biopsy express the gene for PD-L1? Note that PD-L1 is not the official gene symbol name for this gene

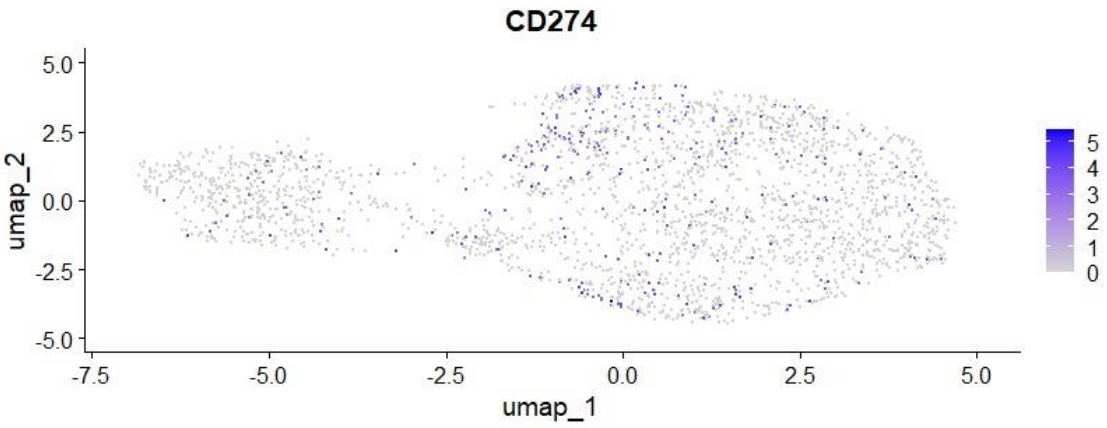
PD-L1 הוא שם מקוצר של הגן Programmed Death-Ligand 1, והשם הרשמי שלו בגנום של עכברים ובני אדם הוא CD274. לכן, נחפש את הביטוי של הגן CD274 בטבלה שלנו.

נקבל כי אחוז התאים שמבטאים את הגן PL-L1 הוא:

```
> print(paste0("Percentage of cells expressing PD-L1: ", round(percent_PD_L1, 2), "%"))
[1] "Percentage of cells expressing PD-L1: 12.11%"
```

כלומר לפחות 10% מהתאים בביופסיה מבטאים את הגן שחיפשנו.

נציג גם את פיזורו ברקמה:



כיוון שקיבלנו שעבור שלושת השאלות התשובה יצאה חיובית, ניתן לומר שקיים סיכוי טוב שהמטופל יגיב לטיפול בתקופה האימונטרפית.