שימוש ב-Chip-Seq על DNA חופשי הקשור להיסטונים עם מודיפיקציית H3K4me3 כדי לסווג חולים במחלות כבד

שם	מייל	ת.ז	משתמש CS
עמית אלפר	amit.alper@mail.huji.ac.il	214981789	amit_alper
הלל דרויש	darvish.hillel@mail.huji.ac.il	325938835	hillel_darvish
דוד גרשליס	david.gershelis@mail.huji.ac.il	214907305	davidgersh2004
גסאן גבר	ghasanjb12@gmail.com	212319685	ghasanjb
אילי גוסרסקי	ilay.gussarsky@mail.huji.ac.il	214713091	ilay_gussarsky

רקע

דנ"א חופשי הוא דנ"א המשתחרר לזרם הדם מתאים שמתו (מסיבות שונות ומגוונות). בגרעין, הדנ"א ארוז מרחבית, סביב קומפלקסים חלבוניים הנקראים נוקלאוזומים. כל נוקלאוזום מורכב מתת יחידות הנקראות היסטונים. היסטונים אלו עוברים מודיפיקציות הנקראות Post translation modifications, ומכונות הקוד ההיסטוני. מודיפיקציות היסטוניות שונות, קשורות לפונקציות ביולוגיות שונות [1]. חלקן מייצגות אזורים של אנהנסרים, חלקן מייצגות אזורי פרומוטור לגנים מושתקים, חלקן אזורי פרומוטור לגנים שעוברים ביטוי, וקיימות עוד מגוון פונקציות נוספות. המודיפיקציה עליה נסתכל היא H3K4me3, הקשורה באזורים של פרומוטורים של גנים אשר יכולים לעבור ביטוי בתא [2]. במסגרת שחרור הדנ"א למחזור הדם על ידי התאים המתים, משתחרר גם דנ"א הקשור לנוקלאוזומים, אשר חלקם מכילים היסטון מסומן במודיפיקציה זו. כל ריד שנקלט בשיטה משוייך לגן מסוים. שיוך כל הרידים בדגימה יוצרים פרופיל של הגנים המבוטאים בקרב התאים שמתו ושחררו את הדנ"א שלהם למחזור הדם בדגימה זו. מתוך הגנים הללו יכולים להיות כאלו שמסמנים את קיומה של המחלה.

שאלת המחקר ומטרות

השאלה הראשונה שנרצה לשאול היא - האם נוכל בהינתן סיגנל הביטוי הגנטי מהדנ"א החופשי בדם לבנות מסווג שידע להבחין בין אדם חולה, ובין אדם בריא, בדגש על מחלות כבד. המטרה שלנו להקאתון היא ראשית להגיע למסווג שידע לסווג חולים מול בריאים. בהמשך נרצה לשאול האם נוכל בעזרת פרופיל הביטוי להבדיל בין פתולוגיות שונות של המחלה, בדגש על הפתולוגיות הנפוצות בדאטה שלנו (למשל MASH, AIH).

הדאטה

הדאטה נאסף על ידי תרומת דם, שממנה הופקה פלסמה. הפלסמה הוכנסה למערכת Chip-seq, שתופסת היסטונים על ידי נוגדנים מתאימים היודעים לקשור את זנב ההיסטונים עם המודיפיקציה 3H3K4me. לאחר מכן מבוצעת שטיפה של מה שלא נקשר, ומתבצע ריצוף של ה-DNA להיסטונים שנשארו קשורים לנוגדנים. הרידים עוברים השוואה מרובת רצפים לגנום מייצג, וכל ריד מסווג לגן שהכי מתאים לו [3]. הדאטה מורכב משלוש טבלאות:_טבלה על האנשים הבריאים, כל עמודה הגיע מניסוי Chip-seq על החתימה ההיסטונית H3K4me3 אצל האדם, וכל שורה היא כמות ה-read-ים המנורמלים (הנרמול מתואר ב-supplementary של [3]) עבור גן מסוים. לכל אדם (כותרת עמודה בטבלה) יש קוד מאפיין, ושמות השורות הם קודים של שמות הגנים. כמות הדגימות בטבלה זו היא 715, וכמות הגנים היא 18,182. טבלה עם האנשים החולים, מכילה את אותן שורות כמו הטבלה של האנשים הבריאים, כאשר לה יש 207 דגימות של חולים. טבלה עם מידע קליני על החולים, בה יש מידע על 188 חולים מתוך ה-207,כאשר לכל אחד יש מידע על דיאגנוסטיקת הביופסיה שאומרת איזה סוג של פתולוגיה יש להם, יש מידע על תאריך הדיאגנוסטיקה, מידע נוסף על הביופסיה ומידע האם הכבד מושתל והאם יש דחייה של השתל.

היפותזות

הפרדה בין חולים ובריאים: ההערכה שלנו היא שגנים שמתבטאים רק/בעיקר בכבד (כמו 2Albumin, APOA וכו') יופיעו בחלק יחסי גדול יותר בקרב חולי כבד מאשר בקרב נבדקים בריאים. מכאן אנו מעריכים שבעיקר באמצעות גנים אלו נוכל להבדיל בין נבדקים בריאים וחולים [4]. הפרדה בין פתולוגיות שונות: בהסתכלות ראשונית על הדאטה אנו מבחינים בשתי נקודות חשובות: הראשונה, יש bias מאוד חזק לפתולוגיות מסוימות (MASH, AIH הן הפתולוגיות של 40% מהחולים); השנייה, שמעל 60% מהפתולוגיות שזוהו מופיעות בדגימה אחת בלבד. על כן, אנחנו משוכנעים שבהינתן הדאטה הקיים לא נוכל להפריד בין כל הפתולוגיות השונות, אך מעריכים שנוכל להבדיל בין שלוש הפתולוגיות הנפוצות ביותר (AIH, MASH ואולי PBC).

מודל חישובי, אלגוריתם, ומבחנים סטטיסטיים

- ויזואליזציות בעזרת הורדת מימד - PCA, tSNE.
- מבחנים סטטיסטיים כמו anova, t-test, כדי להבדיל בין קבוצות (חולים מול בריאים, וקבוצות חולים שונים) עבור כל גן.
- הפקת התפלגויות בעזרת GMM ו-CDF לכל גן, כדי להבדיל בין קבוצות.
- ביצוע NMF למציאת גנים מבדילים בין הקבוצות.
- ביצוע סינונים ו-preprocessing, והפקת מסווג בעזרת למידת מכונה (רגרסיה, random forest, ואולי רשתות).

שלבים לביצוע

שלב 1: להבין את הדאטה

- הבנת מקור הדאטה, איך הוא נאסף, ואיזו אינפורמציה סוג הדאטה מכיל.
- ספירה והסתכלות על כמויות הסיגנלים של הגנים השונים אצל חולים מול בריאים.
- ספירת כמות הדיאגנוזות השונות של החולים.
- ביצוע ויזואליזציה עם הורדת מימד לפי עמודות (אנשים) ולפי שורות (גנים), לפי סטטוס מחלה.

שלב 2: אנליזה של דוגמאות תורמים בריאים אל מול דוגמאות חולים

- ביצוע מבחנים סטטיסטים על כל גן, חולים מול בריאים, כדי לזהות biomarkers (גנים מבדילים) בין שתי הקבוצות.
- הפקת GMM ו-CDF לכל גן, חולים מול בריאים, כדי לזהות גנים עם שוני בהתפלגות בין שתי הקבוצות.
- ביצוע NMF על חולים מול בריאים לפי ביטוי הגנים של כל אדם.
- יצירת מסווג לפי הגנים שנמצאו משמעותיים בשלבים הקודמים.

שלב 3: אנליזה של דוגמות החולים לפי פתולוגיות

- ביצוע אותם פעולות כמו בשלב 2, רק שבמקום שתי קבוצות של בריאים וחולים, יש מספר קבוצות של חולים סוגים, לפי תוצאת הביופסיה שלהם.

בונוס

הקבוצה שלנו מכילה: שלושה עתודאים לביולוגיה חישובית בתכנית BIO, צוער תלפיות וסטודנט לתואר ראשון במדעי המחשב. בקבוצה אישה וארבעה גברים. בזכות הגיוון הנ"ל נטען לבונוס של בין 3 ל-4 נקודות.

ביבליוגרפיה

1. Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research*, *21*(3), 381-395. https://doi.org/10.1038/cr.2011.22

2. Wang, H., & Helin, K. (2024). Roles of H3K4 methylation in biology and disease. *Trends in Cell Biology*. https://doi.org/10.1016/j.tcb.2024.06.001

3. Sadeh, R., Sharkia, I., Fialkoff, G., Rahat, A., Gutin, J., Chappleboim, A., ... & Friedman, N. (2021). ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. *Nature biotechnology*, *39*(5), 586-598. https://doi.org/10.1038/s41587-020-00775-6

4. Ogawa, A., Yano, M., Tsujinaka, T., Ebisui, C., Morimoto, T., Kishibuchi, M., ... & Monden, M. (1997). Gene expression of albumin and liver-specific nuclear transcription factors in liver of protein-deprived rats. The Journal of nutrition, 127(7), 1328-1332. https://doi.org/10.1093/jn/127.7.1328