הגנה על מאגר הגנום הנחקר בעזרת פרטיות דיפרנציאלית

האוניברסיטה הפתוחה 2024ב'
סמינר בביו-אינפורמטיקה 20552

בן שוור
בהנחיית ד״ר מיה הרמן

תוכן עניינים

1. מבוא

ההתקדמות המהירה במחקר הגנומי מחזיקה פוטנציאל אדיר והוא כולל לא רק את המבנה והפעולה של הגנומים אלא גם את ההשלכות שלהם על תחומים הנמצאים בין רפואה מותאמת אישית ועד לחקלאות ולמדע פורנזי. עם זאת, הנתונים הגנומיים מכילים מידע רגיש שמעורר דאגות משמעותיות בנושא פרטיות. הגנה על פרטיותם של אנשים התורמים את המידע הגנומי שלהם למאגרי מחקר היא חיונית לשמירה על האמון ולעידוד השתתפות במחקרים אלו. אחת הגישות המבטיחות ביותר להגנת מידע רגיש זה היא באמצעות פרטיות דיפרנציאלית.

פרטיות דיפרנציאלית היא פתרון מתמטי חזק שנועד להבטיח כי הכללת או החרגת נתונים של אדם יחיד לא תשפיע באופן משמעותי על תוצאות ניתוח נתונים מסוים. המשמעות היא שחוקרים יכולים להפיק תובנות חשובות מהמאגר הכולל מבלי לחשוף נקודות נתונים בודדות. פרטיות דיפרנציאלית פותחה במקור כדי להתמודד עם סוגיות פרטיות בתחומים שונים במדעי הנתונים, וכיום היא מהווה כלי חיוני בהגנת נתונים גנומיים.

הצורך באמצעי פרטיות מחמירים במחקר גנומי הוא הכרחי. נתונים גנומיים הם מזהים מטבעם, מכיוון שהם מכילים סמנים ייחודיים שניתן להתחקות אחריהם אל פרטים וקרוביהם. גישה בלתי מורשית או פרצות אבטחה יכולות להוביל לתוצאות חמורות, כולל אפליה בתעסוקה ובביטוח והפרת פרטיות אישית ומשפחתית. לכן, יש ליישם שיטות המעניקות הבטחות חזקות נגד סיכונים אלו. יישום פרטיות דיפרנציאלית במאגרי נתונים גנומיים כרוך באתגרים טכניים רבים. אלו כוללים את הצורך לאזן בין פרטיות לשימושיות הנתונים, מכיוון שאמצעי פרטיות מוגזמים יכולים להפוך את הנתונים לפחות מועילים למטרות מחקר. בנוסף, הקשרים הנוצרים במאגרי נתונים גנומיים בין היתר על ידי קשרים משפחתיים מוסיפים קושי נוסף הרי הנחת הפתרון של פרטיות דיפרנציאלית היא אי-תלות בין הנתונים.

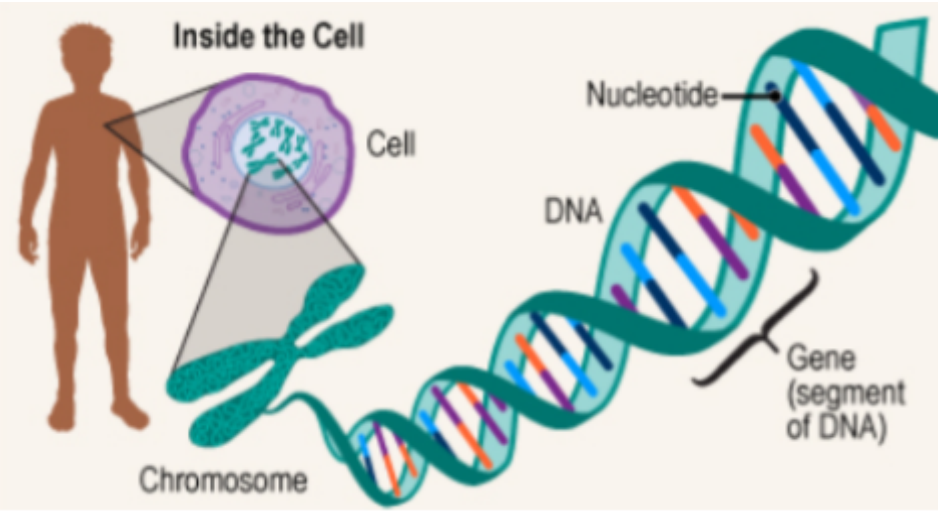
בסמינר זה נבין לעומק את עולם הגנומיקה ונחקור את עקרונות הפרטיות הדיפרנציאלית ואת יישומה בהגנת מאגרי מחקר גנומיים. נדון כיצד ניתן לשלב פרטיות דיפרנציאלית באופן אפקטיבי בתהליכי ניתוח נתונים גנומיים, והטכניקות הספציפיות שפותחו להשגת מטרה זו.

2. גנומיקה

2.1 רקע

גנומיקה היא תחום ביולוגי העוסק בחקר והלמידה של הגנום (הסט השלם של ה-DNA) של אורגניזמים.
הגנום מכיל את כל המידע הגנטי של האורגניזם, כולל הקוד הגנטי שלו, המציין את המאפיינים הפיזיים והביוכימיים שלו.
גנומיקה משלבת טכנולוגיות חדישות כמו ריצוף DNA, תיאור מבנה הגנים, וניתוחים ביולוגיים וביואינפורמטיים כדי להבין את תפקודם של הגנים ואת השפעתם על האורגניזם.

אם נסתכל על גוף האדם למשל, בממוצע עבור אדם בוגר ממין זכר יש כ37.2 טריליון תאים אשר יוצרים את מבנה הגוף ובפרט רקמות ואיברים [6].
כל אחד מהתאים האלו מלבד סוגי תאים בודדים כמו למשל תאי דם אדומים מכילים העתק שלם של הגנום של אותו אדם כלומר הם מכילים את הDNA השלם שלו.



איור 2.1 - ה-DNA בתא האדם.
https://www.studocu.com/in/document/pondicherry-university/general-microbio
[ogy/genetic-mutation-germinal-and-somatic-mutations/33159017

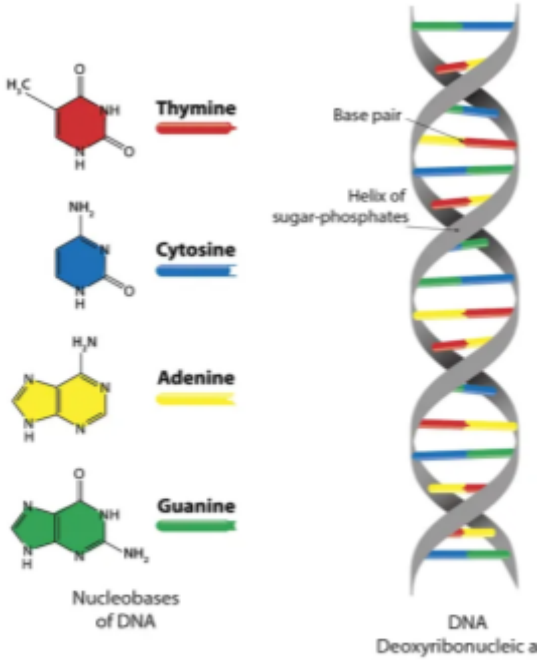
2.2 גנום

2.2.1 מבנה הגנום

הגנום הוא המאגר המלא של כל ה-DNA שבתא של אורגניזם, הוא מורכב מה-DNA של המיטוכונדריה (בתנאי שהתאים של אותו אורגניזם מכילים מיטוכונדריה) וכן מה-DNA בגרעין התא.
הגנום בתוך גרעין התא מאורגן ביחידות הנקראות כרומוזומים וכל אחד מהם מכיל רצפי DNA מסוימים.

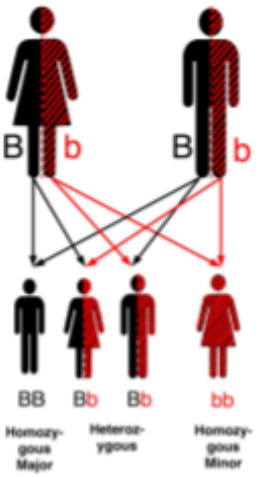
ה-DNA עצמו מורכב משתי שרשראות מקושרות שבסופן הן מסודרות כך שהן יוצרות מבנה של סליל כפול (double-helix), לכל אחת מהשרשאות האלו נהוג לקרוא גדיל.

כל גדיל כזה מורכב משורות של נוקלאוטידים.
הנוקלאיטידים מורכבים משלושה חלקים, קבוצת פוספט, בסיס חנקני ומולקולת סוכר המחברת אותן.
החלק העיקרי הוא הבסיס החנקני כאשר יש ארבעה בסיסים חנקניים אפשריים:
אדנין (A), טימין (T), גואנין (G) וציטוזין (C).
בין הבסיסים החנקניים נוצרים קשרי מימן המתקיימים רק בין שני זוגות:
אדנין וטימין (AT), גואנין וציטוזין (GC).
קשרי המימן מתרחשים בין הבסיסים הנגדיים במבנה ה-DNA כלומר בין כל זוג בסיסים (Base pair) הנמצאים זה מול זה ב2 הגדילים וממשיכה זו נוצר מבנה ה-DNA.



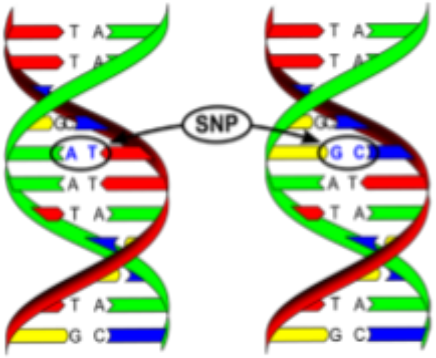
איור 2.2 - מבנה ה-DNA
[https://hevdel.co.il/🧬-מה-ההבדל-בין-dna-ל-rna/]

הגנום האנושי מורכב מכ-3.2 מיליארד זוגות בסיסים אלו ובין שני אנשים הרוב מהן זהה לחלוטין כך שעבור שני אנשים זרים רק כ0.5% מהם שונה בין אותם אנשים, ועבור 2 אנשים עם קרבה משפחתית הערך קטן עוד יותר וזאת בגלל ירושת האללים.
אללים הם גרסאות שונות של גנים שהם מספר זוגות בסיסים וכל אדם יורש עבור כל גן אלל מכל הורה, זהו חלק מחוק ההפרדה של מנדל.



איור 2.3 - תורשה מנדלית לילד [מאמר 2].

השוני העיקרי בגנום האנושי בין אנשים נקרא Single Nucleotide Polymorphism או בקצרה SNP והוא מכונה גם סניפ, סניפ הוא וריאציה של רצף ה-DNA המשפיעה רק על נוקלאוטיד יחיד ובפרט על הבסיס של אותו נוקלאוטיד שהתרחשה ליותר מאחוז 1 מהאוכלוסיה [2].



איור 2.4 - החלפה של נוקלאוטיד אדנין (A) בגואנין (G). לפיכך, במיקום הספציפי הזה, ה-SNP נצפה, ושתי וריאציות הנוקלאוטידים האפשריות - A או G, נקראות אללים עבור מיקום הייחוס הזה.
[https://www.openpr.com/news/406308/global-single-nucleotide-polymorphism-snp-market-2017-beckman-coulter-qiagen-luminex-corporation-enzo-life-sciences-bio-rad-sequenom-genscript.html]

2.3 נתונים גנומיים

כמו שנאמר גנומיקה היא תחום ביולוגי העוסק בחקר והלמידה של הגנום, לכן נרצה להבין מאיפה המידע הזה מגיע, איך נוכל להשיג אותו ומה הוא המידע עצמו. נתונים גנומיים יכולים להגיע ממקורות רבים, ובאמצעות הטכנולוגיות המתקדמות שיש לנו היום ניתן להפיק אותם במהירות ובכמויות עצומות.

2.3.1 ריצוף DNA

המידע עצמו שאנחנו מעוניינים בו הוא הסדר של הנוקלואטידים בגנום, התהליך הקשור להבנה של אותו הסדר נקרא ריצוף (Sequencing). ריצוף DNA מאפשר לחוקרים לקרוא את הקוד הגנטי הנמצא בתוך גנום האורגניזם והוא מלמד אותנו על מבנה, פעולות והיסטוריית האבולוציה של אותו אורגניזם.

יש מספר טכניקות לביצוע ריצוף ה-DNA, למשל:

- Sanger Sequencing
- Next-Generation Sequencing (NGS)
    - Sequencing By Synthesis
    - Sequencing By Ligation
    - Nanopore Sequencing

תהליך הריצוף מתחיל מלקיחת דגימה ביולוגית, למשל דם, רקמה או רוק מהאורגניזם הרצוי. לאחר מכן בעזרת טכניקות מסוימות מחלצים את ה-DNA מהדגימה ומכינים אותו לריצוף. נתייחס לשיטות שציונו מעל.

2.3.1.1 ריצוף בשיטת סנגר (Sanger Sequencing)

שיטת ריצוף זו הייתה השיטה הראשונה לריצוף DNA והיא פותחה על ידי פרדריק סנגר ועמיתיו בשנת 1977. בשיטה זו מרצפים שברים של DNA, עד כ900 זוגות בסיסים.
ריצוף בשיטה זו כוללת שכפול גדילי DNA במבחנות, כלומר הוא כולל:

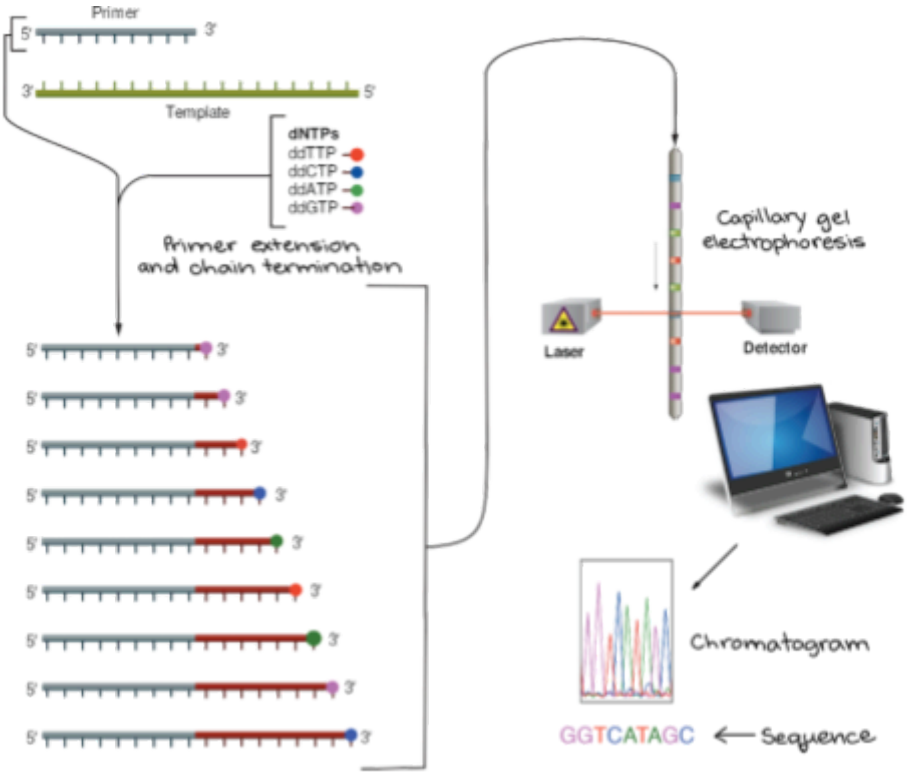
- DNA פולימראז שהוא אנזים שתפקידו להשלים גדיל DNA הנמצא במצב של חד-גדיל לדו-גדיל.
- primer, זהו חלק קצר של DNA חד-גדילי המתחבר לשבר ה-DNA אותו מרצפים והוא מתנהג כמעין מתנע לפולימראז וגורם לו להתחיל להשלים את ה-DNA החד-גדילי.
- ארבעת הנוקלאוטידים המרכיבים את ה-DNA.
- שבר ה-DNA אותו נרצה לרצף.

בנוסף לארבעת הרכיבים המצוינים מעל, ריצוף בשיטה זו כולל רכיב ייחודי, רכיב זה הוא גרסה נוספת של אותם ארבעת הנוקלאוטידים אך בשונה מהם, נוקלאוטידים חדשים לא יכולים להתווסף לגדיל לאחר הוספה של נוקלאוטידים ייחודיים אלו. כל אחד מארבעת הסוגים של הנוקלאוטידים המיוחדים מתוייגים בעזרת צבע פלורסנטי (צבע שונה עבור כל אחד מארבעת הבסיסים) ושמם של נוקלאוטידים אלו הוא דידאוקסינוקלאוטיד - dideoxynucleotide.

התהליך של ריצוף DNA בשיטה זו מתרחש באופן הבא:

1. מחממים את תערובת השברים של ה-DNA הדו-גדילי כדי להפריד בין שני הגדילים ובכך מקבילים DNA חד-גדילי.
2. מקררים את התערובת החדשה שקיבלנו מהשלב הקודם כדי שה-primer יתחבר לשבר DNA החד-גדילי אותו נרצה לרצף.
3. לאחר שה-primer נקשר לשבר הרצוי, מעלים שוב את הטמפרטורה כדי לאפשר ל-DNA פולימראז להשלים את ה-DNA בכך שהוא ימשיך להוסיף לשרשרת נוקלאוטידים נוקלאוטידים חדשים עד שמתווסף אותו נוקלאוטיד מיוחד הדידאוקסינוקלאוטידים באופן אקראי במקום נוקלאוטיד רגיל, בשלב זה אף נוקלאוטיד נוסף לא יוכל להתווסף כך שהגדיל שה-DNA פולימראז יצר מסתיים עם הדידאוקסינוקלאוטידים. בגלל שהנוקלאוטידים הרגילים נמצאים בריכוז פי מאה מאשר הדידאוקסינוקלאוטידים נקבל עותקים של גדיל ה-DNA המקורי באורכים שונים ואקראיים.

תהליך זה חוזר על עצמו מספר פעמים כך שתיאורטית מובטח שבכל מיקום בשבר DNA הרצוי יופיע דידאוקסינוקלאוטידים, ולבסוף כל אותם עותקים באורכים שונים של ה-DNA מועברים דרך צינור ארוך המכיל ג'ל אלקטרופורזה.
בג'ל האלקטרופורזה, גדילים קצרים ינועו יותר מהר מאשר גדילים ארוכים ונוכל לנצל זאת כדי ליצור סדר, ובעזרת לייזר בקצה הצינור ניתן לזהות בדיוק מהו הבסיס במיקום המדויק על גבי אותו שבר ה-DNA בעזרת הדידאוקסינוקלאוטידים והתיוגים שלהם.



איור 2.5 - תהליך ריצוף בעזרת שיטת סנגר
[https://favpng.com/png_view/sanger-sequencing-dna-sequencing-dideoxynucleotide-png/7uyLAB07]

שיטת הריצוף של סנגר הביאה להצלחתו של פרויקט הגנום האנושי שמטרתו היית לזהות את רצף זוגות הבסיסים המרכיבים את ה-DNA של האדם ולמפות את כל הגנים בגנום האדם.

2.3.1.2 ריצוף מהדור החדש Next-Generation Sequencing

ריצוף מהדור החדש (Next-Generation Sequencing) מתייחס לקבוצה של מספר טכניקות שונות לריצוף המשתמשות בטכנולוגיות שונות, אך רובן חולקות במשותף מספר תכונות המייחדות אותן משיטת הריצוף של סנגר:

- **מקביליות -** פעולות הריצוף מתבצעות בו זמנית וכך מתאפשר ריצוף מקבילי של מספר שברים של ה-DNA.
- **מהירות -** בגלל היתרון של המקביליות, נוצר לטכניקות אלו גם יתרון של מהירות על גבי הריצוף של סנגר.
- **מחיר -** המחיר לריצוף בסיסים בשיטות מהדור החדש משמעותית יותר זול.
- **ריצוף על ידי סינתזה -** בשונה מריצוף בשיטה של סנגר שבסופה אנחנו מקבלים אורכים שונים של עותקים של שבר ה-DNA אותו אנחנו מרצפים ועל יד ג'ל האלקטרופורזה אנחנו יכולים לבצע סדר בין אותם עותקים, בטכניקת הריצוף על ידי סינתזה, כבר עם פעילות ה-DNA פולימראז אשר משלים את ה-DNA החד-גדילי, בעזרת הוספת נוקלאוטידים

מיוחדים אשר פולטים חלקיק כלשהו למשל אור או פרוטון ברגע שהם מצליחים להתחבר לבסיס המשלים שלהם בשבר DNA אותו מרצפים, ובעזרת כלים שיודעים לנטר ולזהות זאת, ניתן לפענח את רצף ה-DNA.

השיטות מהדור החדש אפשרו גישה יותר רחבה ליכולות הריצוף והן קידמו מחקרים רבים בגנומיקה בזכות המהירות והיחס של המחיר ליעילות. שיטות אלו הפכו כיום לסטנדרט של פעולות הריצוף.

2.3.2 יישומי ריצוף נפוצים

נרצה להבין כעת מה הם היישומים אליהם אנחנו צריכים את כל היכולות של הריצוף.

2.3.2.1 ריצוף גנום מלא (Whole Genome Sequencing - WGS)

רצפי גנום מלא מכילים את כל המידע הגנטי של האורגניזם, ובעזרתם ניתן לזהות באופן מפורט ומדויק תהליכי מוטציה, סניפים (SNP), ושינויים במבנה הגנום [22]. ריצוף גנום מלא מהווה כלי עוצמתי וחשוב במחקרים בתחום הגנומיקה. לריצוף גנום מלא יש המון שימושים, אל חלקם נתייחס בהמשך.

2.3.2.2 ריצוף אקסום מלא (Whole-Exome Sequencing - WES)

ריצוף אקסום מלא מתמקד בריצוף של האזורים המקודדים לחלבונים של גנים בגנום. יתרונו לעומת ריצוף גנום מלא הוא שכלל האקסום מהווה כ-1% מהגנום האנושי אך בו זמנית ניתן ללמוד ממנו הרבה מכיוון שוריאציות גנטיות המשמשות את רצפי החלבון יכולים לגרום למחלות רבות כגון מחלת האלצהיימר [22].
היכולת לרצף רק את האקסום מתוך כלל הגנום מאפשרת לחוקרים לחקור את אותו אקסום בעזרת הריצוף היעיל יותר מאשר ריצוף כלל הגנום.

2.3.2.3 ריצוף ממוקד (Targeted Sequencing)

בריצוף אקסום התמקדנו באזורים המקודדים לחלבונים של גנים אך יש מחקרים רבים המתמקדים באזורים ספציפיים עוד יותר הממוקדים למחקר עצמו, לכן ריצוף ממוקד הוא פתרון אידיאלי למחקרים אלו למשל ריצוף גנים המקושרים למחלה מסוימת. במקום לרצף את הגנום המלא ניתן לרצף אזור ממוקד ולבצע זאת ביחס עלות תועלת גבוה מאוד [22].

2.4 התרומה למדע

הגנומיקה זכתה להיות אבן הבניין של המחקר המדעי המודרני, והיא מהווה מקור למהפכות בתחומים רבים ומגוונים, החל מרפואה ועד לחקלאות. נתייחס למספר דוגמאות לתרומה של הגנומיקה למדע.

2.4.1 רפואה מותאמת אישית

הגנומיקה המהפכנית תרמה לרפואה מותאמת אישית, שבה אסטרטגיות טיפול מותאמות לפרופיל הגנטי האישי של כל אדם. עם היכולות של חשיפת נטיות מוקדמות (Genetic predisposition) למחלות ותגובות למחלות (Pharmacogenetics ו-Pharmacogenomics), הגנומיקה מקלה על פיתוח טיפולים ממוקדים עם יעילות מוגברת ותופעות לוואי מוזלות.

2.4.1.1 נטיות מוקדמות (Genetic predisposition)

נטיה מוקדמת היא הנטייה לחלות במחלות מסוימות ולגנומיקה יש משמעות רבה כדי לחשוף נטיות אלו. לרוב זיהוי נטיה מוקדמת כולל זיהוי של סניפים (SNPs) המשויכים לסיכון גבוה של תכונה או מחלה כלשהי. בעזרת המחקרים של Genome-wide association studies (GWAS) אשר מנתחים וריאציות גנטיות רבות על גבי הגנום כדי למצוא קשר בין וריאציות מסוימות כמו למשל סניפ (SNP) כלשהו לבין תכונות או מחלות על ידי השוואה של הגנום של אנשים עם ואנשים בלי תכונות ומחלות מסוימות אפשר לאתר את הוריאציות הגנטיות שיותר שכיחות בקבוצת אנשים מסוימת בהשוואה לקבוצה השנייה [29].

2.4.1.2 פרמקוגנטיקה ופרמקוגנומיה (Pharmacogenetics ו-Pharmacogenomics)

פרמקוגנטיקה הוא ענף מדעי שעושה שימוש בגנטיקה כדי לגלות ולשפר תרופות. ענף זה בוחן כיצד שונות גנטית באנשים יכולה להשפיע על התגובה שלהם לתרופות. בדומה לכך גם פרמקוגנומיה הוא ענף העושה שימוש בגנטיקה כדי לגלות ולשפר תרופות. ההבדל העיקרי בין השניים הוא שבפרמקוגנטיקה הפוקוס העיקרי הוא על ההשפעה של וריאציה בגן בודד למשל סניפ (SNP) או מוטציה, וכיצד היא משפיעה על התגובה של בן אדם לתרופה לעומת הפרמקוגנומיה אשר עובדת בגישה יותר מקיפה שבה מנסים להבין איך הגנום השלם של בן אדם משפיע על התגובה לתרופה מסוימת [7].

2.4.2 הבנה של מחלות מורכבות ופיתוח טיפולים

הלמידה של הגנומיקה העמיקה את ההבנה של מחלות מורכבות כגון סרטן, פגמים גנטיים למשל תסמונת דאון ומחלות מדבקות כגון מחלת ה-Covid19.
בעזרת תהליך הריצוף היה ניתן לעקוב ולנטר אחר וירוס האחראי על מחלת ה-Covid19 ובנוסף לזהות מוטציות חדשות בניגוד לבדיקת ה-PCR עוד לפני שהן מתפשטות ובכך למנוע את ההתפשטות של הזן החדש [26].

2.4.3 הנדסה גנטית וביוטכנולוגית

הגנומיקה אפשרה קידמה משמעותית בהנדסה גנטית ובתחום הביוטכנולוגיה על ידי מתן הבנה עמוקה יותר של מבנה, פעולות ובקרת הגנים השונים והגנום עצמו.
הגנומיקה הובילה לפיתוחים של כלים לעריכה מדויקת של ה-DNA למשל CRISPR-Cas9.

2.4.3.1 מערכות CRISPR-Cas

מערכות ה-CRISPR-Cas מאפשרות לחוקרים לערוך גנים ב-DNA וחלקם מאפשרים לערוך את ה-RNA (המשמשת להעברת מידע גנטי מה-DNA לחלבונים).
את המערכות האלו ניתן לנצל ליישומים טיפוליים כמו חיסול נגיף ה-SARS-CoV-2 אשר גורם למחלת הCovid19 למשל [1].

2.4.4 ביולוגיה אבולוציונית (Evolutionary biology) ושימור המגוון הביולוגי (Conservation biology, Genetic diversity)

ביולוגיה אבולוציונית היא תחום שחוקר את מוצא המינים והשינויים לאורך זמן שהפיקו את מגוון החיים על פני כדור הארץ. בעזרת הגנומיקה והיכולות שהיא מביאה, חוקרים יכולים ללמוד על השינויים הגנטיים בין המינים ובין אותו המין. בעזרת השוואה של הגנום של אורגניזמים שונים, חוקרים יכולים למצוא דמיון ושוני ברצפי ה-DNA שלהם ובמבנה הגנום. אחת המסקנות הגדולות

בנושא זה הוא שככל הנראה לבני אדם ולשימפנזות יש היסטוריה אבולוציונית משותפת ויש לנו אב קדמון משותף. מסקנה זאת מגיעה לאחר חקר הגנום וההשוואה בין האדם לשימפנזה ומתברר כי לבני אדם ולשפנזות יש דמיון ב-DNA בכ98.8 אחוז [13].

2.4.5 זיהוי פלילי (Forensic science)

זיהוי פלילי הוא תחום מאוד חשוב והוא עוזר לחקירת פשיעות. התחום כולל איסוף מידע מזירת הפשע וניתוחו כדי לעזור להגיע למסקנות. הגנומיקה באה לידי ביטוי בתחום זה כאשר נאספים דגימות DNA של הפושע בזירת הפשע ועל ידי הריצוף שלהן ניתן להשוואת אותן למאגר רצפי DNA הזמין לחוקרים ובכך לגלות את זהות הפושע. דגימת ה-DNA מאפשרת לנו להגיע לפושע גם אם דגימת ה-DNA שלו לא חשופה ישירות לבלשים על ידי הקשר הגנטי של אותו פושע למשפחתו ובאופן עקיף יהיה ניתן לזהות את הפושע [14]. חשיפת הפושע המכונה The Golden State Killer צלחה בעזרת הגנומיקה ובעזרת ההשוואה של דגימת ה-DNA של אותו פושע לבני משפחתו שהשתמשו באתר המכיל דגימות DNA כדי ללמוד על העץ המשפחתי שלהם [23][21].

2.4.6 גנום חקלאי (Agriculture genome)

המונח "גנום חקלאי" מתייחס למידע גנטי או רצף גנום של אורגניזמים הקשורים לחקלאות. זה כולל את הגנום של היבול, של הצמחים, בעלי חיים ועוד. הגנום החקלאי כולל את הלמידה של אותם הגנומים כדי להאיץ את מאמצי שיפור היבול על ידי זיהוי גנים העומדים בבסיס התכונות הרצויות, כגון עמידות למחלות ותכולה תזונתית [28].

2.4.7 סיכום ומסקנות

השפעתה העמוקה של הגנומיקה על ההתקדמות המדעית אינה ניתנת להכחשה, היא חוצה את הגבולות הדיסציפלינריים של עצמה ומזינה חדשנות בתחומים מגוונים ומאפשרת הסתכלות ממבט שונה על בעיות מגוונות.
החל מרפואה מותאמת אישית ועד לשימוש בגנומיקה כדי לשפר את החקלאות, הגנומיקה ממשיכה לעצב את התפיסה שלנו על מורכבות החיים ולהניב שינויים מהפכניים בחברה.
בעוד שאנו מעמיקים את המחקר הגנומי, הפוטנציאל לגילויים ויישומים נוספים נותר חסר גבולות.

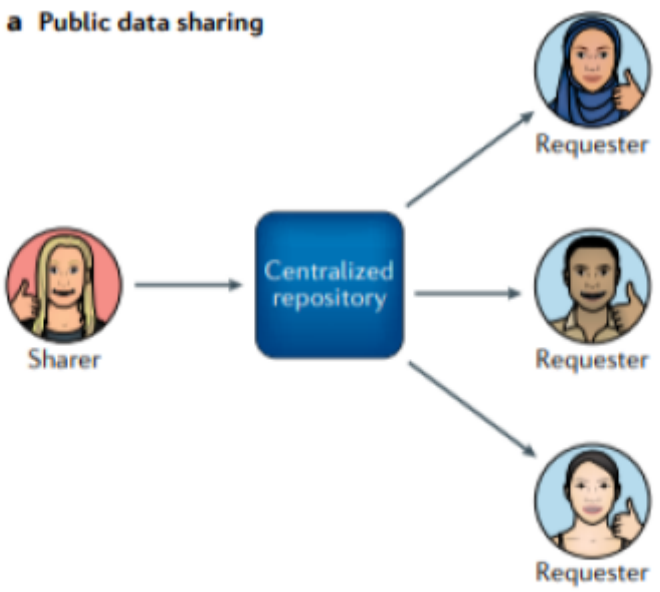
2.5 שיתוף מאגרי גנום

שיתוף מידע גנומי הוא הכרחי כדי לאפשר התקדמות מהירה ויעילה יותר במחקרים מדעיים וכדי לשפר ולמצוא פתרונות לתחומים רבים בעזרת גישת הגנומיקה כמו שהזכרנו כבר מעל.
לשיתוף של מידע גנומי יש מספר גישות, אנו נתייחס לשיטות הבאות:

- גישה חופשית/פומבית - Public
- גישה מבוקרת - Controlled access

2.5.1 גישה חופשית/פומבית - Public

השיטה הכי פשוטה ונוחה למשתמש היא השיטה שמאפשרת גישה לכולם ללא הגבלות אלא חוץ מאשר חוקים ושיקולים אתיים שאותם מצופה שהמשתמש יכיר. בגישה זו, חוקרים שמבצעים בדיקות ומשיגים גישה למידע גנומי כלשהו כמו למשל רצפי גנום מלאים (WGS), משתפים את המידע שבידם לבסיסי נתונים בעלי גישה חופשית כך שכל אחד יוכל להיחשף לאותו מידע, ללמוד ממנו ולבצע מחקרים בעצמו [9].



איור 2.6 - אופן הפעולה של שיתוף מידע בגישה פומבית [מאמר 9]

כמה דוגמאות למאגרי מידע הדוגלים בגישה זו הן:

1. National Center for Biotechnology Information (NCBI) GenBank
2. European Bioinformatics Institute (EMBL-EBI) European Nucleotide Archive (ENA)
3. DNA Data Bank of Japan (DDBJ)

בסיסי נתונים אלו מאפשרים לחוקרים בכל העולם לגשת למידע ולנתח אותו באופן חופשי.
בנוסף, אותם בסיסי נתונים מרכיבים את שיתוף הפעולה הבינלאומי של רצפי נוקלאוטידים - https://www.insdc.org, שמטרתו העיקרית היא לאפשר גישה לרצפי נוקלאוטידים באופן עולמי לקהילת החוקרים, וגיבוש סטנדרליזציה לפורמט של המידע עצמו ולמטא-דאטא (האופן בו התבצעה הדגימה).

שיתוף מידע באופן חופשי בשילוב של דגימות מפורטות והמטא-דאטא מאפשר לחוקרים לענות על שאלות רבות. אחת הדוגמאות הגדולות והמוצלחות למאגר נתונים פומבי הוא הפרויקט The Cancer Genome Atlas - TCGA או בעברית "אטלס הגנום הסרטני", שמטרתו היא לזהות את השינויים ב-DNA הגורמים לסוגים שונים של סרטן למשל בעזרת זיהוי מוטציה בגן מסוים וזאת על ידי ריצוף גנומים וניתוחם בשיטות ביואינפורמטיות ובכך לעזור לחוקרים להבין כיצד נוצרים סוגים שונים של סרטן ולהוביל לדרכים חדשות למניעה, אבחון וטיפול בסרטן.

2.5.2 גישה מבוקרת - Controlled access

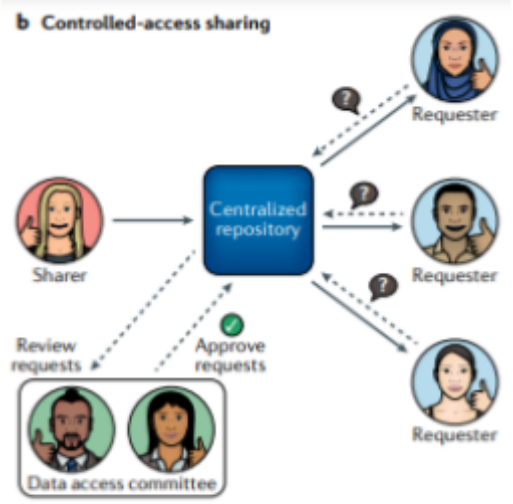
למרות הרצון לפרסם מידע בין חוקרים ומדענים באופן פומבי ולאפשר בכך שימוש באותו מידע כדי לחקור אותו ולהגיע למסקנות רבות, יש מידע גנטי שעבורו גישה זאת בעייתית בעיקר אם מדובר במידע של אדם פרטי לדוגמא רצף הגנום שלו, הרי הגנום מאפשר לנו ללמוד המון על הבן אדם, החל מהנראות (צבע עיניים, צבע שיער, סיכוי להתקרחות) ועד סיכויים לחלות במחלות מסוימות.

מהסיבה הזאת יש צורך גם במאגרי נתונים עם גישה מבוקרת [9]. הגישה למאגרי נתונים אלו מתאפשרת כל עוד חוקרים עומדים בקריטריונים מסוימים. קריטריונים אלו יכולים לכלול:

1. ביקורת על פרוטוקולים, כלומר, כאשר מתבצעת בקשה לגישה לנתונים בגישה מבוקרת, החוקרים צריכים להגיש פרוטוקול התואם את המחקר יחד עם מידע רלוונטי נוסף כמו המוסד/ות אליהם שייכים, מקורות מימון ואמצעי אבטחת המידע לועדת גישה לנתונים DAC Data access committee או ארגון ניהולי דומה להערכת הפרוטוקולים ושאר הנתונים המוגשים. הפרוטוקול עצמו הוא תכנית מפורטת או סט הפעולות המתארות כיצד המחקר יבוצע, והוא כולל את היעדים, השיטות, הטכניקות לניתוח הנתונים, שיקולים אתיים, סיכונים ויתרונות אפשריים.

2. התחייבות לשימוש במידע אך ורק למחקרים הקשורים לבריאות.

3. דרישות המחייבות את כלל החוקרים אשר מבקשים גישה למידה לעמוד בהן לדוגמא, רמת ושיטת אמצעי אבטחה מינימליים, הסכם לתנאים הכוללים למשל הגבלות לשיתוף המידע ועוד.



איור 2.7 - אופן הפעולה של שיתוף מידע בגישה המבוקרת [מאמר 9]

גישה מבוקרת היא הגישה המועדפת עבור נתונים גנומים אשר עברו התממה (de-identification) ועדיין יש חשש משמעותי מבניית זיהוי מחדש (re-identification) של אדם פרטי על ידי הנתונים הגנומים שלו ושילוב של אותם נתונים עם נתונים פומביים אחרים על אותו אדם.
בהמשך יופיע הסבר מפורט על התממה והיכולות לזיהוי מחדש.

מספר דוגמאות למאגרי נתונים גנומים הדוגלים בגישה המבוקרת הן למשל - The UK Biobank https://www.ukbiobank.ac.uk , אפשר לראות את הקרטריונים הנדרשים בקישור הבא - https://www.ukbiobank.ac.uk/enable-your-research/register.
עוד דוגמאות הן:

1. The database of Genotypes and Phenotypes (dbGaP)
2. The European Genome-Phenome Archive (EGA)

3 הגנת הפרט

3.1 הקדמה

בתקופה שבה אנו חיים המוכרת בתרבותה הדיגיטלית והפצת המידע, עניין הגנת הפרט מקבל חשיבות רבה. החל מפרטים אישיים שאנחנו מפרסמים במדיות חברתיות ועד למידע רגיש המאוחסן בבסיסי נתונים רפואיים, העקבות הדיגיטליות שלנו שהפכו להיות נפוצות יותר ויותר, מעלות שאלות עמוקות בנוגע לאוטונומיה, אבטחה ואתיקה.

הגנת הפרט בבסיסה מתמקדת בזכותם של אנשים לשלוט בגישה, השימוש והפצת המידע האישי שלהם. היא משמשת הגנה מפני חדירה בלתי מוצדקת לפרטיותו של האדם, ניצול ואפליה, ושומרת לא רק על כבודנו אלא גם על הפעולות שלנו בנוף דיגיטלי הולך ומתרחב.

התפתחות הטכנולוגיה, בעודה פותחת הזדמנויות מדהימות לחדשנות ושיתוף פעולה, הציבה את האנשים לפני סיכונים רבים. פריצות נתונים למשל מתקפת סייבר על ספק שירותי בריאות גדול, מעקבים סמויים ועיוותי אלגוריתמים מדגישים את שבריריות הפרטיות שלנו בעולם המובל בנתונים.

בתחום הגנומיקה, עם כל ההתקדמות במחקר הגנומיקה והטכנולוגיות שאפשרו להוזיל וליעל את כל תהליך הריצוף, הגיע עידן חדש של שירותי בריאותי וגילויים מדעיים, המבטיחים טיפולים מותאמים אישית מותאמים למבנה הגנטי של הפרט ותובנות לגבי יחסי הגומלין המורכבים בין גנטיקה למחלות. עם זאת, בתוך התקדמות זו, טמונה דאגה דחופה: הגנה על הפרט [8].

בלב העניין נמצאים התכונות של נתונים גנומיים. כל רצף גנטי הוא מתווה ייחודי שלא רק חושף רגישות למחלות, אלא הוא חושף גם פרטים אינטימיים כמו מוצאו של הפרט, הקשרים המשפחתיים, ואפילו תכונות התנהגותיות. למידע רגיש שכזה, אם הוא מטופל בצורה לא נכונה וייפול לידיים לא נכונות, עלולות להיות השלכות עמוקות כלפי הפרט, החל מאפליה בין אם בקבלה לעבודות או הסדרי ביטוחי בריאות ועד לגניבת זהות וסחיטות לא חוקיות ופוגעניות רק על ידי הסיכון לחלות במחלות מסויימות המתגלה על ידיי אותו מידע גנומי על הפרט.
יתרה מזאת, כאשר בסיסי נתונים גנומיים גדלים בגודלם ובהיקפם, הסיכון לגישה בלתי מורשית ושימוש לרעה עולה באופן אקספוננציאלי.

מידע גנומי של פרט מספק מידע רב עד כדי היכולת לזהות את אותו האדם, ובשונה מנתונים רפואיים רבים, מידע גנומי נוטה להישאר סטטי לאורך חייו של אדם ולספק מידע ביומטרי ייחודי לאותו אדם. ואף מתברר כי על ידי סך הכל כ-75 סניפים (SNPים) של פרט מסויים, ניתן לסרוק בסיסי נתונים בעלי גישה חופשית/פומבית ועל ידי השוואה אפשר למצוא בסיכוי גבוהה ולזהות את אותו אדם בין האוכלוסיה כולה ברחבי העולם ומכאן אפשר להמשיך ולהתחיל לנסות להשיג נתונים רבים נוספים [18].

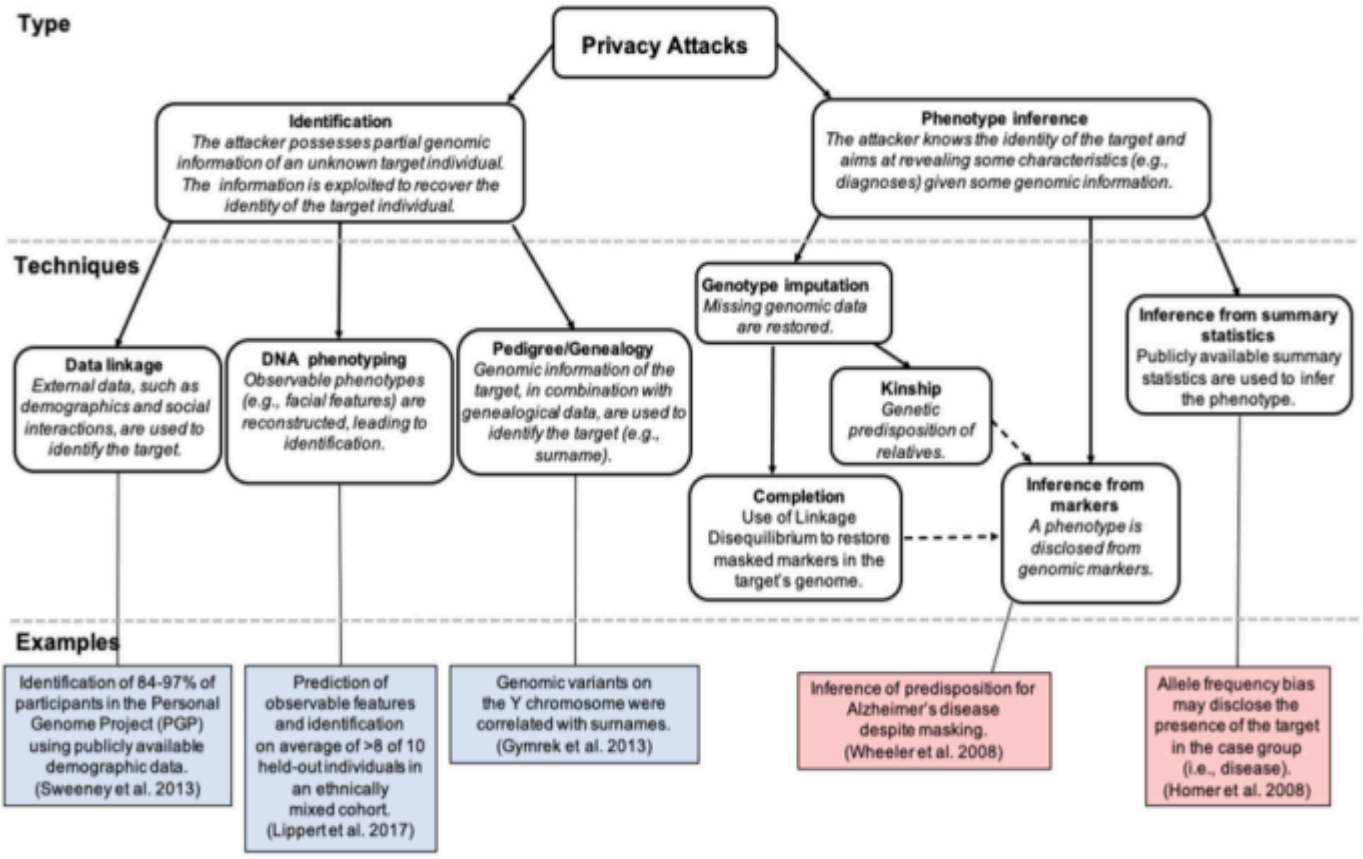
חשוב לזכור ולהדגיש את הדמיון בגנום בין אנשים קרובים משפחתית (ביולוגית) וזאת מפני שאם אדם מסויים מאשר לקבל על עצמו את כלל הסיכונים והוא רוצה לתרום את המידע הגנטי שלו למדע, הוא פוגע בפרטיותם של קטינים שלא יכולים חוקית לתת אישור בעצמם ואף בקרובי משפחתו שטרם נולדו.

3.2 סיכוני פרטיות בשיתוף מידע גנומי של אנשים

ישנם התקפות מסוגים שונים שניתן לבצע, הפוגעות בפרטיותו של אדם בעזרת ניצול הנתונים הזמינים לציבור ובפרט לתוקף [8]. נסווג התקפות אלו לשתי קטגוריות:

- זיהוי (Identification)
- הסקת פנוטיפים (Phenotype Inference)

נוכל לראות את הטכניקות השונות למתקפות אלו והתוצאות באיור 3.1.



איור 3.1 - סיווג של התקפות פרטיות ידועות בשיתוף מידע גנומי. אנחנו מבדילים בין שתי קטגוריות עיקריות של התקפות פרטיות: זיהוי והסקת פנוטיפ. לכל סוג של התקפה, אנו מדגישים את הטכניקות הידועות העיקריות ומדווחים על דוגמאות רלוונטיות שפורסמו [מאמר 8].

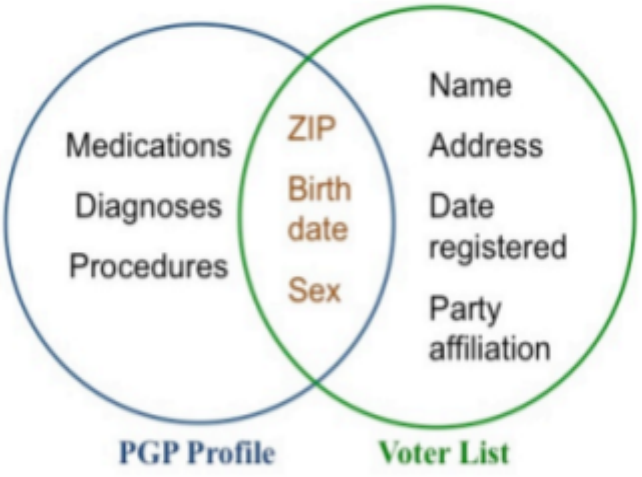
מספר דוגמאות למידע עזר כלומר מידע שאותו ניתן להשיג מבסיסי נתונים החשופים לציבור ועל ידו ניתן לבצע מתקפות זיהוי והסקת פנוטיפ מופיעות באיור 3.2.

Auxiliary information	Identification	Phenotype inference	Examples of data sources
Demographics, Surnames	X		Census Data (https://www.census.gov/data.html)
Pedigree, Family Tree	X	X	PGP (https://pgp.med.harvard.edu) CEPH (http://www.cephb.fr)
Gene Expression	X	X	GTEx Project (https://gtexportal.org/home/)
Genotype Data	X	X	OpenSNP (https://www.opensnp.org) 1000 Genomes Project (https://www.internationalgenome.org) dbGaP (https://www.ncbi.nlm.nih.gov/gap/)
Social Relationships	X	X	Population Registry, Social Networks
Observable Phenotypes		X	Social Networks
Clinical Data	X	X	Clinical Data Research Networks
Summary of Statistics		X	UK BioBank (https://www.ukbiobank.ac.uk)

איור 3.2 - דוגמה למידע עזר. מספר דוגמאות למידע עזר ונתונים זמינים
והמקורות שיכולים להיות מנוצלים על ידי תוקפים לביצוע התקפות זיהוי והסקת פנוטיפ [מאמר 8].

3.2.1 זיהוי (Identification)

כשמדובר על מתקפות זיהוי מדובר על היכולת לשחזר את הזהות של תורמים (אנשים אשר תרמו וביצעו בדיקה כלשהי בין אם על ידי לקיחת רוק, רקמה או דם לסיבות רפואיות או מחקריות וחשפו נתונים גנומיים) בעזרת ניצול גישה לנתונים גנומים של בן אדם שעברו התממה (de-identification) כלומר אנונימיזציה למידע הגורמת להפחתת הסיכון של זיהוי התורם על ידי הסרת פרטים אישיים כגון שם פרטי, שם משפחה, תעודת זהות או כל תעודה מזהה ולפעמים אף הסרה של פרטים למשל מיקוד ואזור מגורים.
פעולה כזאת של התממה לא תמיד מספיקה, נתייחס לPGP כדוגמא, PGP או פרויקט הגנום האנושי שם לו למטרה לרצף ולפרסם את רצף הגנום השלם של 100,000 מתנדבים לצד הרשומה הרפואית המלאה שלהם. חזון הפרויקט, שהחל בשנת 2005, הוא לאפשר התקדמות במחקר הגנומיקה האישית ופיתוח הרפואה המותאמת אישית [27].
פרויקט ה-PGP פעל לפי פרוטוקול פרטיות הנקרא "open consent" ואיפשר למתנדבים לבחור באופן חופשי לחשוף נתונים אישיים רבים ככל שהם רוצים, לעתים קרובות כולל נתונים דמוגרפים, כגון תאריך לידה, מין ומיקוד אך למרות זאת המידע המוצג ברשת עבר התממה ולכן לא ניתן לזהות את המתנדבים באופן ישיר לכאורה.
אך מתברר שעל ידי הצלבה בין נתונים מה-PGP לנתוני הצבעה בבחירות הצליחו לזהות כ-84%-97% מהאנשים שהשתתפו עד לאותו הזמן בפרויקט הגנום האנושי [31][27].



איור 3.3 - הצלבה בין פרופיל PGP לרשימת בוחרים באמצעות ערכים דמוגרפיים כדי להתאים שמות לתוכן המופיע בפרופיל PGP. [מאמר 27]

בנוסף מתברר כי ניתן לזהות יותר מ-99% מאזרחיה של ארצות הברית עם כל בסיס נתונים המכיל רק 15 נתונים דמוגרפיים. מכך נובע שתהליך ההתממה שהוא חלק הכרחי בכל הנוגע להגנה על הפרט אינו מספיק בעידן של היום כאשר מידע רב זמין לציבור הרחב המאפשר לבצע הצלבות בין נתונים ובכך להרכיב תוצאות [25].

3.2.2 הסקת פנוטיפים (Phenotype Inference)

בטכניקות להסקת פנוטיפים של אדם מדובר על היכולת לשער "תכונות" של אותו אדם לפי המידע הגנטי שלו. בגנטיקה, פנוטיפ מתייחס לתכונות או למאפיינים שניתן לראותם, אשר נקבעים על ידי המבנה הגנטי שלו (גנוטיפ) באינטראקציה עם גורמים סביבתיים. תכונות אלו יכולות לכלול תכונות פיזיות כמו גובה, צבע עיניים או רגישות למחלה.

מתקפת הסקת פנוטיפים עובדת בצורה הבאה:

1. <u>ניתוח נתוני הגנום -</u> התוקפים מתחילים על ידי ניתוח הנתונים הגנומיים של האדם. הנתונים הללו כוללים מידע על וריאציות בסדרת הדנ"א, הידועות גם כסמנים גנטיים.
   סמן גנטי הוא רצף DNA שמיקומו על גבי הכרומוזום ידוע, ויכול לשמש למשל לזיהוי ולחקר מחלות תורשתיות.
2. <u>הצלבה עם מידע פנוטיפי -</u> התקופים בשלב זה מצליבים בין הסמנים הגנטיים של אותו אדם למידע פנוטיפי ידוע.
3. <u>הסקת פנוטיפים -</u> על ידי הצלבה של סמנים גנטיים ופנוטיפים, התוקפים יכולים להסיק או לחזות תכונות ומאפיינים שונים של האדם. לדוגמה, הם עשויים להסיק את סיכון ההתפתחות של מחלות מסוימות, את מוצאם, או אפילו תכונות פיזיות כמו צבע העיניים או גובה.

מתקפות כאלו יכולות לשמש למשל חברות ביטוח ולגרום לאפליה בין אנשים שונים על ידיי הפנוטיפים שלהם (הסיכון לחלות במחלות מסוימות).

3.2.3 מסקנה

כאשר מידע גנטי נחשף ומשותף על ידי מישהו זה הופך להיות בלתי אפשרי לעקוב בפועל אחר היעדים הפוטנציאלים המרובים שלהם. ואם הוא מגיע לידיים הלא נכונות הוא יוכל לגרום לפגיע משמעותית בפרטיות של רבים והפחד הגדול הוא שאנו עוד לא מודעים לכוח שיש לחשיפת מידע גנומי ולאילו דברים ניתן לנצל אותם.

3.3 טכניקות להגנת הפרטיות במאגרי נתונים

הגנת הפרט היא חלק מרכזי בהקשרים של מאגרי נתונים המכילים מידע רגיש ובפרט מידע גנומי, מאחר ומידע כזה יכול להוות מקור ייחודי של אינפורמציה על כל אדם ואדם. לכן השמירה על הפרטיות במידע רגיש על אדם היא לא רק על מנת להגן על פרטיותו וכל הזכויות האישיות שלו, אלא גם למנוע מימוש של סיכונים פוטנציאליים כגון הפרת הפרטיות במקרים של גילוי תופעות גנטיות רגישות. אך מנגד עומד הרצון של חוקרים רבים לשתף מידע רגיש בעיקר אם מדובר על מידע גנומי על מנת לקדם את המחקר הרפואי ולפתח טיפולים רפואיים מתקדמים, ולכן חשוב לאפשר גישה למידע כזה בצורה שיהיה ניתן לשתף אותו אך בו זמנית להבטיח הגנה ושמירה על פרטיות הנתונים. נציין מספר טכניקות המשומשות למקרים אלו.

	Goal	Techniques	Privacy Protection	Pros	Cons	Examples of Relevant Applications
Data Security	Blocking Unauthorized users to access the original data	Access control, Trust-but-verify	Grant access only to authorized users	Easy to implement Allow monitoring of data usage	Vulnerable to internal attacks (e.g., a dishonest user who has access to the data)	Data Repositories such as: dbGaP (https://www.ncbi.nlm.nih.gov/gap/), EGA (https://www.ebi.ac.uk/ega/home), and the All of Us Research Program (https://www.researchallofus.org
		Homomorphic Encryption (HE)	Data are encrypted, generating ciphertext, on which certain operations can be performed and produce the same results as when the original, nonencrypted data are used	Strong and provable security guarantees	Computationally intense	Genomic sequence matching 53,54,56, outsource computation 57–62
		Secure Multiparty Computation (SMC)	Data are encrypted and multiple parties can jointly compute a function without learning anything about each others' private data	Strong and provable security guarantees	High communication cost	Genomic sequence comparison 64–66, secure statistical test evaluation 67,68, and GWAS 69
Data Anonymization	Protect the identity/ presence of the individual in shared data	k-anonymity via generalization and suppression of SNPs	Data are transformed such that, for each record in the output, there are k-1 other records with the same set of quasi identifiers	Intuitive notion of privacy	Vulnerable against an informed adversary May lead to overly generalized data	"Anonymization" of DNA sequences 73,74
		Differential privacy (adversary wants to know if target is in the database) achieved via random perturbation	Data results are perturbed to guarantee that an adversary who observes the outputs cannot determine the presence of any individual record in the data	Strong and provable privacy guarantees	Released data may have limited usability due to the noise injected	GWAS test statistics (e.g.,$\chi 2$) and SNPs highly associated with diseases of interest 77–79,93,94

איור 3.4 - סקירה כללית של טכניקות מרכזיות המשמשות לשמירה על פרטיות נתונים גנומיים [מאמר 8].

3.3.1 בקרת גישה (Access Control)

שיטת בקרת הגישה מגבילה את חשיפת המידע בכך שהיא מאפשרת גישה לאותם נתונים רגישים או לחלקו אך ורק לאנשים בעלי הרשאה לכך. אותם אנשים בעלי גישה יקבלו את המידע בצורה המקורית שלו כך שעליהם לדאוג שברגע שהמידע מגיע אליהם הוא נשמר בצורה מיטבית ולא ישמש לרעה. בגישה זו ניתן להימנע מלחשוף את המידע לכלל הציבור ולאפשר גישה אך ורק לחוקרים שקיבלו הרשאה לכך, אך נזכור שברגע שהמידע משותף עם לפחות בן אדם אחד לא נוכל להמשיך לעקוב מכאן והלאה לאן המידע יגיע כלומר בגישה זאת אנחנו מסתמכים רבות על עניין האמון.

3.3.2 הצפנה (Encryption)

בעזרת הצפנה ניתן להפוך את המידע המקורי למידע שלא ניתן לזהות אותו אלא רק על ידי המפתח המתאים. בכך אם המידע נחשף ומגיע ליידיים הלא נכונות, קשה מאוד להסיק מסקנות מהמידע המוצפן ללא מפתח פענוח. יש מספר רב של שיטות למימוש הצפנה עבור נתונים רבים אך החיסרון הוא שכדי שחוקרים שונים יוכלו לבצע ניתוחים על מידע זה, יש צורך תחילה לפענח את אותו מידע. כדי להימנע ממצב זה, ניתן להשתמש בהצפנה הומומורפיסטית אשר מאפשרת לבצע חישובים על המידע המוצפן ללא פגיע במידע עצמו ובכך חוקרים רבים יכולים לנתח את המידע המוצפן ולהגיע למסקנות על המידע המוצפן בלי להכיר בכלל את המידע שעומד מאחורי ההצפנה, ולבסוף רק החוקרים אשר יש להם את המפתח פענוח, יוכלו לפענח את המסקנות שחוקרים רבים אחרים הסיקו ולשתף אותם ובכך ניתן למנוע גישה מלאה למידע המקורי.

3.3.3 K-אנונימיות (K-anonymity)

k-אנונימיות היא טכניקה נוספת לשמירת פרטיות המשמשת להגנה על נתונים רגישים במאגרי מידע על ידי הבטחה שכל נתון של אדם מסוים אינו ניתן להבחנה מנתוניהם של לפחות k-1 אנשים אחרים בהתבסס על מאפיינים מזהים מסוימים אשר הם פרטי מידע אישיים של האדם למשל מיקוד או מין. כלומר, אנחנו דואגים שכל נתון של אדם יהיה דומה לנתונים של לפחות עוד k-1 אנשים אחרים.
המשמעות היא שאם מישהו ינסה לזהות את האדם באמצעות מידע מסוים כמו מיקוד או מין, הוא לא יוכל להבדיל בין אדם זה לבין עוד k-1 אנשים אחרים במאגר הנתונים.
הדבר מושג בעזרת שתי פעולות, הכללה - בפעולה זו אנחנו הופכים את הנתונים לפחות ספציפיים, כגון הגדרת טווחי גילאים לעומת הגיל עצמו של אותו אדם, והפעולה השנייה, הסתרה - בפעולה זו אנחנו מסתירים חלק מהנתונים וכך אנחנו יוצרים קבוצות או מחלקות שקילות של לפחות k רשומות.

3.3.4 פרטיות דיפרנציאלית (Differential Privacy)

פרטיות דיפרנציאלית היא טכניקה מתקדמת לשמירת פרטיות המבטיחה את פרטיותם של יחידים במאגר נתונים תוך שמירה על היכולת להפיק מידע שימושי מהנתונים. שיטה זו פועלת על ידי הוספת רעש מבוקר לשאילתות על הנתונים או על הנתונים עצמם, כך שהכללתו או אי הכללתו של נתון של אדם יחיד תשפיע באופן מינימלי על התוצאה. בדרך זו, קשה מאוד לתוקפים להסיק נתונים על אדם מסוים מהתוצאות שפורסמו. פרטיות דיפרנציאלית מספקת ערבויות מתמטיות חזקות לפרטיות, מה שהופך אותה ליעילה במיוחד לשימוש בניתוח סטטיסטי, למידת מכונה ויישומים נוספים המבוססים על נתונים וכל זאת תוך שמירה על סודיות המידע האישי.

3.3.5 שילוב טכניקות

לכל אחד מהטכניקות יש יתרונות וחסרונות משלה ולכן הדבר הנכון לעשות הוא לדעת לשלב בצורה חכמה מספר טכניקות כדי להגביר את הפרטיות. אחת הדוגמאות היא פתרון המשלב הצפנה הומומורפיסטית ביחד עם פרטיות דיפרנציאלית [24].

3.4 פרטיות דיפרנציאלית

3.4.1 הקדמה

פרטיות דיפרנציאלית (Differential Privacy) היא הגדרה מתמטית מתקדמת שנועדה להבטיח את פרטיותם של יחידים כאשר הנתונים שלהם נכללים במאגר הנבדק. המושג הוצג לראשונה על ידי סינתיה דבורק ועמיתיה בשנת 2006, והוא מבוסס על הרעיון שהתוצאות של כל ניתוח על מאגר הנתונים לא ישתנו באופן משמעותי אם נתוני פרט מסוים נכללים או לא נכללים במאגר. כדי להשיג זאת, מוסיפים רעש אקראי מבוקר לתוצאות השאילתות או לנתונים על ידי אלגוריתם פרטי דיפרנציאלי, מה שמטשטש את ההשפעה של רשומה בודדת [12].

ישנן שתי טכניקות מרכזיות בפרטיות דיפרנציאלית:

1. **פרטיות דיפרנציאלית מרכזית**: הרעש מתווסף לתוצאות השאילתות ולא למידע עצמו במאגר הנתונים.
2. **פרטיות דיפרנציאלית לוקאלית**: הרעש מתווסף למידע עצמו לפני שהוא נשמר או מנותח.

בהגדרת הפרטיות הדיפרנציאלית קיים הפרמטר ε (אפסילון) המשמש לכימות רמת הפרטיות. ערך נמוך של ε מציין פרטיות חזקה יותר אך פחות דיוק, בעוד שערך גבוה מאפשר דיוק רב יותר עם הגנה חלשה יותר על הפרטיות. בנוסף, קיים הפרמטר δ (דלתא), המספק גמישות נוספת למקרי קצה ולמעבר מתיאוריה למציאות.

אחד ההיבטים המרכזיים של פרטיות דיפרנציאלית הוא **תקציב הפרטיות**. תקציב הפרטיות מתאר את הסכום הכולל של ערכי ε (אפסילון) עבור כל השאילתות שניתן לבצע לפני שהפרטיות נפגעת במידה בלתי קבילה. לכל שאילתא יש ערך ε קבוע משלה, וכל שאילתא על מאגר הנתונים צורכת חלק מהתקציב הזה. חשוב לנהל את תקציב הפרטיות בזהירות, שכן שימוש יתר בתקציב עלול לגרום לחשיפת מידע פרטי על יחידים. התקציב מאפשר שליטה על האיזון בין רמת הפרטיות לרמת הדיוק של התוצאות, ומשמש ככלי חשוב למעקב וניהול פרטיות הנתונים.

פרטיות דיפרנציאלית מיושמת בתחומים רבים כמו מפקדי אוכלוסין, טכנולוגיה ובריאות. היא מאפשרת לנתח ולשתף מידע סטטיסטי בצורה בטוחה, המגנה על פרטיות המשתמשים ומונעת חשיפת מידע אישי רגיש. המטרה העיקרית היא ללמוד מידע חשוב על אוכלוסייה מבלי לחשוף פרטים על פרט בודד.

3.4.2 הגדרה פורמלית

פרטיות דיפרנציאלית מבטיחה שכל מי שיבחן תוצאות של ניתוח פרטי דיפרנציאלי, בסופו של דבר יגיע לאותה הנחה עבור פרטים אישיים של כל האנשים, אף אם אדם מסוים לא נכלל בניתוח הסופי. מה שאנחנו יכולים ללמוד על אדם מהפרטים האישיים שלו, מוגבל על ידי הפרמטרים ε ו-δ.

ההגדרה הפורמלית של פרטיות דיפרנציאלית כוללת שני בסיסי נתונים שכנים D ו-'D, בסיסי נתונים שכנים הם בסיסי נתונים זהים לחלוטין חוץ מהבדל ברשומה אחת (בין אם הרשומה קיימת בשני בסיסי הנתונים אך הנתונים שונים או שהרשומה נמצאת רק באחד מבסיסי הנתונים) ואלגוריתם A שמחזיר את התוצאה לשאילתא.
A נחשב ל-(ϵ,δ)-פרטי דיפרנציאלי אם לכל התוצאות האפשריות S של האלגוריתם A מתקיים:

$$Pr[A(D) \in S] \leq e^{\epsilon} * Pr[A(D') \in S] + \delta$$

בעצם, הנוסחה משווה את ההסתברות שהאלגוריתם A יוצר תוצאה מסוימת עבור הבסיס נתונים D אל מול ההסתברות שהאלגוריתם A יוצר אל מול הבסיס נתונים 'D, וכדי ש-A יספק אותנו, התוצאות צריכות להיות קרובות עד כדי e^{ϵ} והתוספת של δ.
אם אכן הנוסחה מתקיימת עבור אפסילון נמוך, זה אומר שגם אם מידע על בן אדם מסוים נמצא בבסיס נתונים וגם אם לא, התוצאות של ניתוח בסיס הנתונים המקורי לא ישתנה הרבה, כלומר קשה להניח האם הנתונים של אדם מסוים נכללים במאגר מידע וכך אנחנו שומרים על פרטיותו [12].

3.4.3 למה צריך פרטיות דיפרנציאלית

פרטיות דיפרנציאלית היא תחום חדש ולפעמים לא ברור למה אנחנו צריכים אותו, בעיקר כאשר מדובר בהשגת נתונים סטטיסטיים ממאגרי נתונים הנקראים "Aggregated data" שהם נתונים כוללים/קבוצתיים ולא אישיים של אדם מסוים, למשל כמות האנשים שיש להם סניפ (SNP) מסוים או אפילו ממוצע גילאים. נשאלת השאלה האם זה בכלל אפשרי לזהות אדם מסוים על ידי נתונים סטטיסטיים אלו? אז מתברר שכן, יש מספר טכניקות לעשות זאת.

טכניקה ראשונה מנצלת את האופי של העידן שלנו, שבו יש המון מידע פומבי החשוף לכולם ובעזרתו ניתן להגיע למסקנות, ראינו דוגמה לכך בהסבר על מתקפות זיהוי (Identification) תחת 3.2.1.
וטכניקה נוספת שטוענת כי מתברר שעל ידי ביצוע כמות מספקת של שאילתות למאגר נתונים סטטיסטיים אשר מכילים רעש ניתן לחשוף מידע פרטי. את זאת הראו אירית דינור וקובי ניסים שעל מאמרם עוד משנת 2003, זכו ב-Test Of Time Award ובעזרתו, החלה התקדמות בפתרון לפרטיות דיפרנציאלית ומימוש תקציב הפרטיות למניעת פגיע זו [10][12].

כעת כדי להבין קצת יותר טוב איך ניתן להשיג מידע על אדם פרטי אף על פי שמדובר רק על Aggregated Data, אציג דוגמא יחסית פשוטה [30]:

אליס ובוב הם שני פרופסורים ב"אוניברסיטה הפרטית". לשניהם יש גישה לבסיס נתונים המכיל מידע אישי על הסטודנטים באותה אוניברסיטה. המידע כולל מידע הקשור לעזרה הכלכלית שכל סטודנט מקבל, למשל המלגות שלו או ההלוואות שהוא לוקח. בגלל שבסיס נתונים זה מכיל מידע אישי, הגישה אליו מוגבלת. כדי לקבל אליו גישה, אליס ובוב נדרשו להפגין שהם מתכננים לעקוב אחר הפרוטוקולים של האוניברסיטה לטיפול בנתונים אישיים על ידי ביצוע הדרכה בנושא סודיות וחתימה על הסכמי שימוש בנתונים אישיים האוסרים את השימוש בהם וחשיפת מידע אישי המתקבל מהמאגר.

במרץ, אליס פרסמה מאמר המתבסס על המידע במאגר. אליס ציינה: "נכון לעכשיו יש כ-3,005 סטודנטים בשנה הראשונה שלהם באוניברסיטה הפרטית, 202 מהם מגיעים ממשפחות המרוויחות מעל ל-350 אלף דולר בשנה". אליס מנמקת כי היא פרסמה סטטיסטיקה מצטברת (Aggregated) שנלקחה מלמעלה מ-3,005 אנשים, ואף מידע אישי של אדם מסוים לא ייחשף.

חודש לאחר מכן, באפריל, בוב גם כן פרסם מאמר המכיל את הסטטיסטיקה הבאה: "ל-201 סטודנטים באוניברסיטה הפרטית מתוך 3,004 סטודנטים אשר נמצאים בשנה הראשונה שלהם ללימודים באוניברסיטה יש הכנסת משק בית מעל ל-350 אלף דולר בשנה". אליס ובוב לא מודעים לכך שהם פרסמו מאמרים המכילים מידע דומה.

כעת נשים לב שאם נצליח להבין איזה סטודנט לשנה הראשונה עזב את האוניברסיטה בין חודש מרץ לאפריל, נלמד עליו מידע אישי והוא שמשפחתו מכניסה מעל ל-350 אלף דולר בשנה. זיהוי הסטודנט (נקרא לו ג'ון) כעת יכול להסתכם בתשאול מספר סטודנטים בודדים מהשנה הראשונה: "מי הוא הסטודנט שלמד איתם ועזב בין מרץ לאפריל?".

היופי בהסקת מסקנה זו הוא שאליס ובוב לעולם לא חשפו מידע אישי הקשור לג'ון. כל מה שנעשה זה פרסום מידע סטטיסטי המציין כמות, ולא הייתה כאן שום הפרה של פרוטוקול האוניברסיטה, ובכל זאת מידע פרטי על ג'ון נחשף יחסית בקלות.

3.4.5 יישום

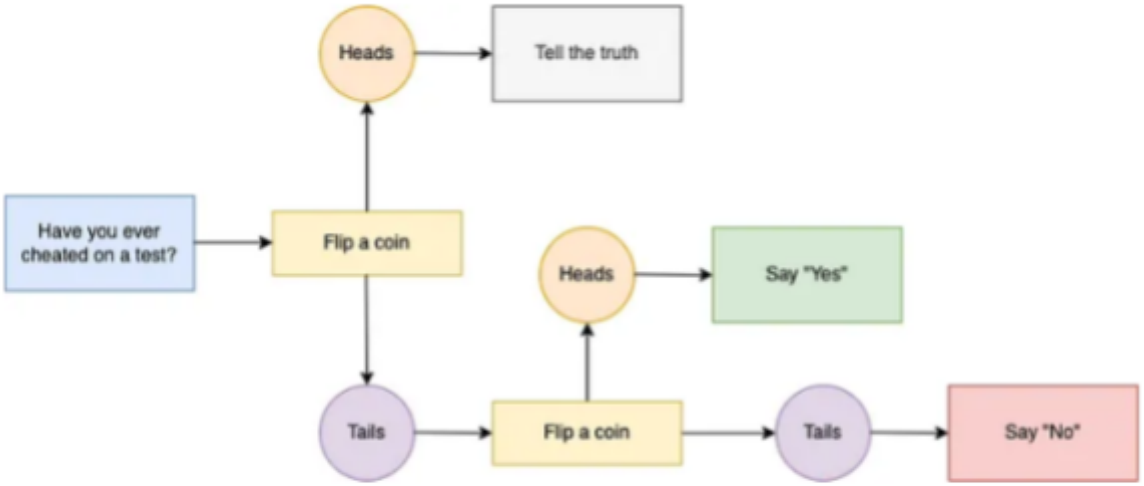
פרטיות דיפרנציאלית מתבססת על הוספת רעש מבוקר לתוצאות או לנתונים בעזרת אלגוריתמים פרטיים דיפרנציאלים. לשם הוספת הרעש יש מספר טכניקות: הוספת רעש באופן אקראי (למשל הטלת מטבע), התפלגות מעריכית, התפלגות נורמלית (גאוסית) או התפלגות לפלס.
עניין חשוב נוסף הוא הבחירה של הפרמטרים ϵ ו-δ, בעזרת הפרמטרים האלה נוכל לשלוט על כיול הפרטיות אל מול דיוק הנתונים.

3.4.5.1 דוגמא בסיסית

נציג דוגמא הממחישה את הרעיון של פרטיות דיפרנציאלית [20], ניקח למשל את הסיטואציה שיש בבית ספר 300 תלמידים, ומנהל בית הספר מעוניין להבין מה הוא אחוז הרמאות במבחנים בקרב התלמידים בבית ספרו. לשם כך התפרסם שאלון אינטרנטי לכלל התלמידים ובו הם נשאלו "האם רימית אי פעם במבחן בבית הספר?", אך ברור כי לתלמידים יהיה חשש מלענות "כן" לשאלה זו מפני שכל מי שיש לו גישה לבסיס נתונים הסופי יכול לבצע התאמה בין התשובה לשם התלמיד ולכן הם קיבלו את ההנחייה הבאה שכל תלמיד אמור לעקוב אחריה לפני מענה שאלות זה:

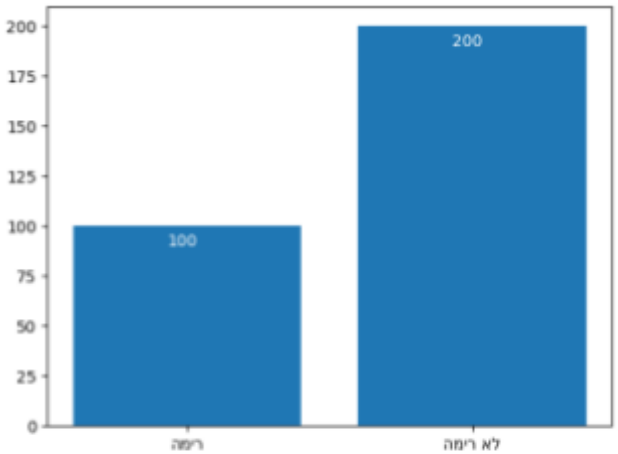
1. הטילו מטבע.
2. אם המטבע נופל על "עץ", ענו את האמת.
3. אם המטבע נופל על "פלי":
    a. הטילו את המטבע פעם נוספת.
    b. אם המטבע נופל על "עץ", ענו על השאלה ב"כן".
    c. אחרת, ענו על השאלה ב"לא".

בעזרת איור 3.5 ניתן לראות ויזואלית את זרימת התשובה לשאלון זה.



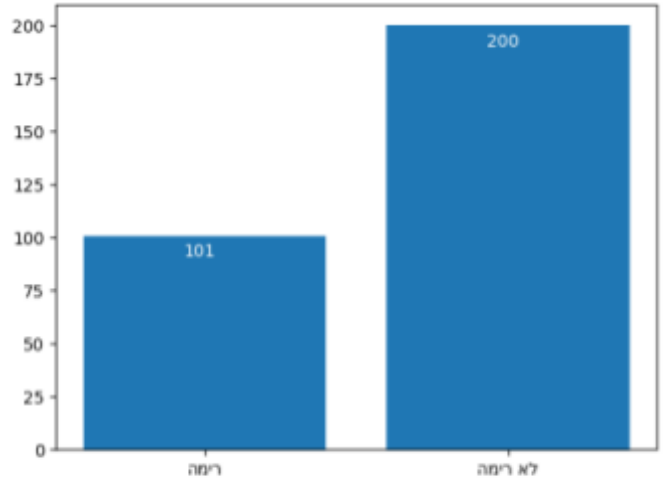
איור 3.5 - דיאגרמה המדגימה את שיטת הטלת המטבע [מאמר 20]

כעת ברור, שלא כל מי שענה "כן" אכן אומר את האמת כלומר לא ניתן באמת להפיק מידע אמיתי אך בכל זאת אם נסתכל על כל התלמידים ככלל נגיע לתוצאות מספקות.
ננתח את התוצאות על ידי סימולציית הניסוי בפייתון עבור הנתונים הבאים - 200 תלמידים לא רימו ו-100 מהם רימו.
נבחן את הנתונים ללא פרטיות דיפרנציאלית כמו שנראה באיור 3.6.



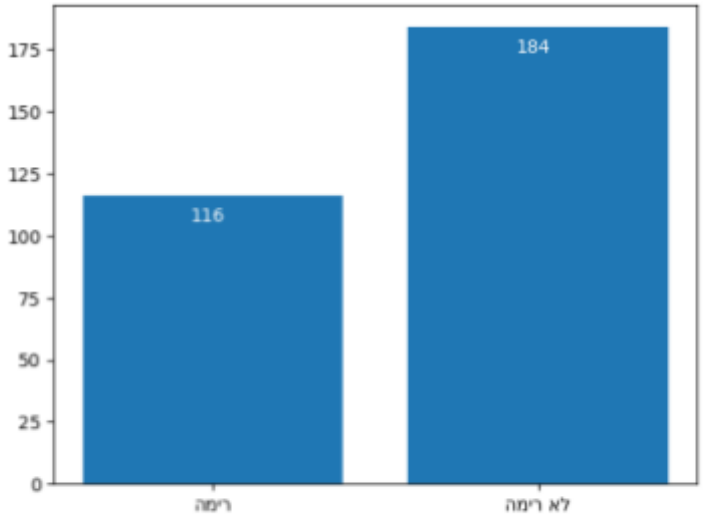
איור 3.6 - סטטיסטיקה של 300 תלמידים שרימו אל מול כאלו שלא, ללא פרטיות דיפרנציאלית.

כעת ברגע שהתלמיד הנוסף עונה על הסקר נוכל לזהות כי הוא רימה הרי נקבל את הנתונים הבאים - 200 תלמידים לא רימו ו-101 מהם רימו (הבט באיור 3.7).



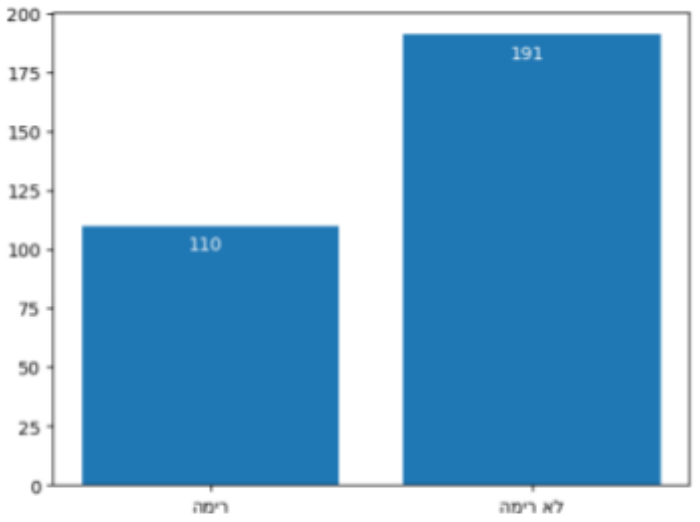
איור 3.7 - סטטיסטיקה של 301 תלמידים שרימו אל מול כאלו שלא,
ללא פרטיות דיפרנציאלית.

עכשיו נחזור על אותה סימולציה ונפעל לפי ההנחיות המדמות לנו פרטיות דיפרנציאלית בעזרת הטלת המטבע.
עבור הנתונים הבאים - 200 תלמידים שלא רימו ו-100 אשר רימו, נראה לפי איור 3.8 כי קיבלנו תוצאה מבלבלת שטוענת כי 116 תלמידים רימו ו-184 לא רימו.



איור 3.8 - סטטיסטיקה של 300 תלמידים שרימו אל מול כאלו שלא, עם
הוספת רעש בעזרת פרטיות דיפרנציאלית.

נוכל לראות כי המידע עצמו לא מאוד מדויק אך זה נובע מההנחיה אותה קיבלנו, השאלה האמיתית היא האם הגענו למצב בו לא נוכל לזהות לאיזו קבוצה שייך התלמיד ה-301.

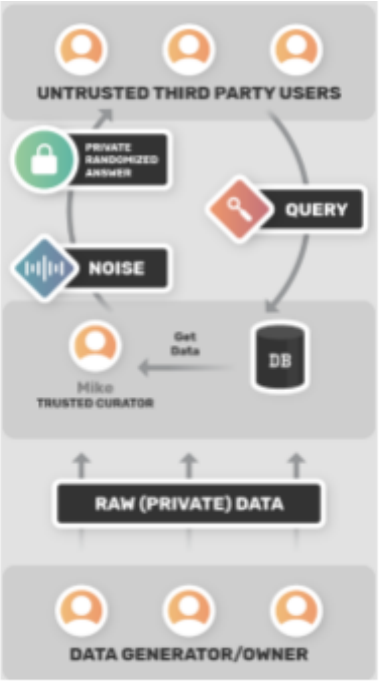


איור 3.9 - סטטיסטיקה של 301 תלמידים שרימו אל מול כאלו שלא, עם הוספת רעש בעזרת פרטיות דיפרנציאלית.

ולפי איור 3.9 ניתן לראות שההוספה של תלמיד 301 לא גרמה לשינוי בתוצאות החושף מידע אודות התלמיד ה-301 וקשה מאוד להבין לאיזו קבוצה הוא שייך. הגענו לתוצאות אלו בזכות האקראיות שאנו מכניסים לנתונים. בעזרת דוגמא זאת אפשר להבין קצת יותר טוב את המטרה והצורך בפרטיות דיפרנציאלית אך חשוב לציין מספר דברים, דבר ראשון השיטה של הטלת המטבע היא שיטה מאוד בסיסית ופחות מומלצת לשימוש אמיתי של פרטיות דיפרנציאלית, ובנוסף לגודל מאגר הנתונים יש משמעות רבה כאשר עובדים עם פרטיות דיפרנציאלית, ככל שמאגר הנתונים גדול יותר כך תועלת המידע גדולה יותר למרות הרעש המתווסף. ולבסוף חשוב להבין שהדוגמא הזאת רק ממחישה את הרעיון מאחורי פרטיות דיפרנציאלית אך במציאות המימוש יכול לעבוד אחרת למשל בכך שהמידע עצמו במאגר הנתונים לא משתנה ולמעשה רק כאשר מתשאלים את בסיס הנתונים, התוצאות עוברות דרך אלגוריתם פרטי דיפרנציאלי שהוא זה שמשנה את התוצאות על ידי הוספת רעש לפני שהן מגיעות אל היעד.

3.4.5.2 פרטיות דיפרנציאלית מרכזית

פרטיות דיפרנציאלית מרכזית היא אחת מהטכניקות המרכזיות במימוש פרטיות דיפרנציאלית. בפרטיות דיפרנציאלית מרכזית בעלי הנתונים משתפים את המידע הפרטי עם אוצר מידע (curator) שסומכים עליו וכל המידע החשוף נשמר אצלו והוא אחראי על אבטחתו כך שהמידע כולו נמצא כעת אצלו והלקוחות המעוניינים במידע מתשאלים אותו. כמו שנאמר המידע עצמו נמצא אצל ה-curator (לפעמים אף יעבור התממה), אך כעת נרצה לאפשר שיתוף של המידע ללקוחות ללא חשיפה של מידע אישי על אנשים הנמצאים במאגר נתונים עצמו [16]. כדי לאפשר זאת, נרצה להשתמש בפרטיות דיפרנציאלית. בפועל, יהיה פורטל או API backend אליו הלקוחות יפנו, ואותו פורטל או API backend יפנה למאגר הנתונים להשגת הנתונים המקוריים, יוסיף לתוצאה רעש בעזרת אלגוריתם פרטי דיפרנציאלי, ואת התוצאה המעודכנת יחזיר ללקוח. נשים לב שרק התוצאות הן אלו שעברו שינוי ולא המידע עצמו במאגר הנתונים.
נוכל לראות איך כל התהליך נראה בעזרת איור 3.10.



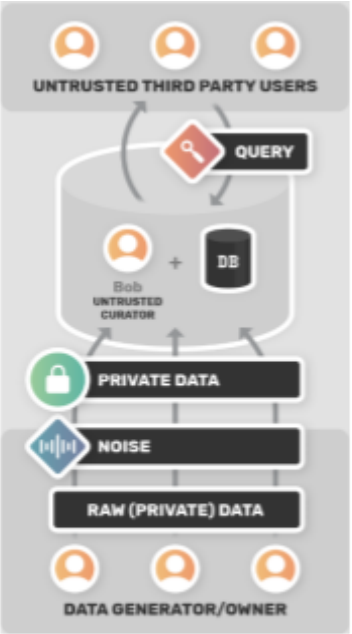
איור 3.10 - אופן הפעולה של פרטיות דיפרנציאלית מרכזית [מאמר 16].

למימוש זה יש מספר יתרונות וחסרונות אשר נרצה להבין ולכן נסקור אותם. אחד היתרונות הגדולים בשיטה זו היא הפשטות, מפני שכל מה שבעלי המידע נדרשים לבצע הוא העברה של המידע לאוצר מידע ורק הוא דואג ואחראי לאבטחתו ומימוש אלגוריתם פרטי דיפרנציאלי המטשטש את התוצאות. אופן פעולה זה מאפשר לבצע אופטימיזציה לאלגוריתם מפני שהוא עובר על המידע כולו (ולא על חלקים של המידע בדומה למימוש של פרטיות דיפרנציאלית לוקאלית שעליה נרחיב ב-3.4.5.3). אך מנגד שיטה זו מניחה כי ניתן לסמוך על אוצר המידע וחשיפתו אל המידע הרגיש חסרת סיכון, אך בפועל הרבה מאוד מהפעמים זה אינו המצב. חסרון בולט נוסף הוא נקודת הכשל הבודדת (SPOF) שאותו אוצר מידע יוצר, כלומר אם הוא נפגע, כל המידע הרגיש עלול להיחשף.

אחד השימושים הגדולים ביותר של פרטיות דיפרנציאלית מרכזית הוא על ידי לשכת מפקד האוכלוסין האמריקאית, לשכת מפקד האוכלוסין של ארה"ב עורכת מפקד אוכלוסין בן עשור, ואוספת מידע דמוגרפי מפורט מכל משק בית בארצות הברית. נתונים אלה חיוניים למטרות רבות, כולל ייצוג פוליטי, הקצאת כספים פדרליים ומחקר חברתי וכלכלי. עם זאת, היא מחויבת לשמור על פרטיותם של כלל האנשים, ולכן במפקד האוכלוסין שהתרחש ב-2020 נעשה שימוש בפרטיות דיפרנציאלית [17].

3.4.5.3 פרטיות דיפרנציאלית לוקאלית

פרטיות דיפרנציאלית לוקאלית היא טכניקות מרכזית נוספת במימוש פרטיות דיפרנציאלית. בפרטיות דיפרנציאלית לוקאלית כל אחד מבעלי המידע משתפים את האוצר מידע (curator) עם המידע הפרטי אך בשונה מפרטיות דיפרנציאלית מרכזית, כאן התווסף רעש למידע עוד לפני שה-curator מקבל את המידע עצמו, המשמעות היא שבטכניקה זו גם ה-curator לא חשוף למידע הרגיש עצמו. למעשה זה מתבצע באופן דומה לדוגמא הבסיסית שהצגנו עם שיטת הטלת המטבעות בחלק 3.4.5.1, ונוכל לראות זאת באיור 3.11 [16].

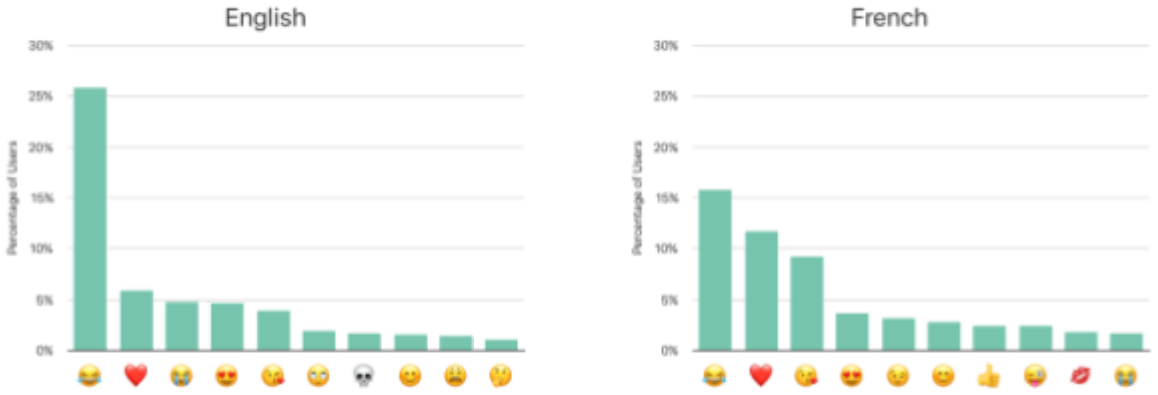


איור 3.11 - אופן הפעולה של פרטיות דיפרנציאלית לוקאלית [מאמר 16].

נבחן גם כאן את היתרונות והחסרונות, ואכן בשונה מפרטיות דיפרנציאלית מרכזית, אין דרישת אמון בין בעלי המידע אל מול אוצר המידע מפני שהמידע עובר "טשטוש" עוד לפני שהוא מגיע אליו אך מכאן גם נובע חסרון משמעותי, בגלל שהמידע עובר טשטוש עוד בהתחלה, כדי להגיע לאותה רמת פרטיות שהיינו יכולים להגיע בעזרת פרטיות דיפרנציאלית מרכזית יש צורך בהרבה יותר רעש מה שיכול להפחית משמעותית את התועלת של הנתונים.

השימושים של פרטיות דיפרנציאלית לוקאלית למעשה רבים יותר משום שהיא מעבירה את שלב הוספת הרעש קרוב ככל האפשר ללקוח והדוגמאות הפופולריות ביותר הן:

1. Google's RAPPOR - פיתוח זה משמש את גוגל לאיסוף מידע על משתמשי ה-Google Chrome בעזרת שימוש בפרטיות דיפרנציאלית, ומכך ללמוד מה הם האתרים או החיפושים הנפוצים ביותר בלי לדעת איזה חיפושים שייכים לאיזה אדם [15].
2. Apple - לאפל יש שימוש נרחב בפרטיות דיפרנציאלית לוקאלית, התהליך מתבצע ממש על הטלפונים (iPhone) או המחשבים (MacBook) ולאחר מכן המידע נשלח לאפל עצמם. למעשה באופן דומה לגוגל, אפל אוספת נתונים על המשתמשים שלה למשל סטטיסטיקות השימוש של אמוג'ים בין שפות שונות של המקלדת (נשים לב לתוצאות באיור 3.12) [5].



איור 3.12 - שימוש אימוג'ים עבור מקלדות בשפה האנגלית והצרפתית [מאמר 5].

נתונים כאלו מאפשרים לאפל להציע את האמוג'ים המתאימים ביותר ב-iOS QuickType בהתאם לשפת המקלדת. וכמובן לאפל יש שימושים רבים נוספים לפרטיות דיפרנציאלית לוקאלית כמו זיהוי חריגות זיכרון באתרים מסויימים ב- Safari ואף למידה של מילים חדשות כמו סלנג מסוים כדי לדעת להתאים את עצמם להשלמה אוטומטית.

3.4.6 מה פרטיות דיפרנציאלית לא עושה

פרטיות דיפרנציאלית אינה פתרון לכל הבעיות, לכן בנוסף להבנה של מהי פרטיות דיפרנציאלי ובמה היא עוזרת, חשוב להבין מנגד במה פרטיות דיפרנציאלית לא עוזרת [11].
פרטיות דיפרנציאלית לא מתאימה לבעיות הבאות:

3.4.6.1 ניתוח נתוני קצה

נתוני קצה הם נתונים ששונים משמעותית מרוב הנתונים במאגר נתונים, ונזכור כי המטרה של פרטיות דיפרנציאלית היא להקשות על הזיהוי של נתונים עבור פרט מסוים בין אם הם נמצאים במאגר הנתונים או לא, גם אם מדובר על נתוני קצה.
בפועל, במקרים אלו, האלגוריתמים יסתירו את אותם נקודות קיצון אם הם נמצאים במאגר נתונים או שהם יצרו שגיאות מסוג ראשון (false positive) וזאת בגלל כמות הרעש הנדרשת במקרים כאלו.

3.4.6.2 ניתוח בסיסי נתונים קטנים

הוספה או החסרה של נתונים עבור פרט מסוים בבסיס נתונים קטן משפיע באופן משמעותי יותר על תוצאות הניתוח בהשוואה לבסיס נתונים גדול.

3.4.6.3 חוסר הגבלה באופן בו נעשה שימוש במידע המופק על האוכלוסייה

על אף שפרטיות דיפרנציאלית מקשה על היכולת להבין האם פרט מסוים נמצא במאגר נתונים, אין לה איך להגביל את השימוש במידע המופק על האוכלוסייה כולה. ניקח למשל את הדוגמה שבה מחקר מסוים שהשתמש במאגר נתונים הפועל לפי פרטיות דיפרנציאלית מרכזית הגיעה למסקנה שעבור אנשים אשר נוהגים לשתות יין אדום באופן מרובה, יש סבירות יותר גבוהה לפתח סרטן מסוג מסוים. נשים לב שבמסקנה זו מדובר על האוכלוסייה כולה ולא על פרט מסוים, אך מנגד, ברגע

שמסקנה זו מפורסמת, במידה ומכירים מישהו העומד בתנאים של המחקר (במקרה הזה, אדם אשר שותה יין אדום באופן מרובה), ניתן להניח על אותו אדם שיש לו סבירות גבוהה יותר לפתח סרטן, וזהו מידע רגיש שהצלחנו בכל זאת לגלות על הפרט אף על פי השימוש בפרטיות דיפרנציאלית. מנגד כמובן שמחקרים כאלה לרוב עוזרים ומקדמים את החברה בכך שהם גורמים לשינוי וקידום בין אם חברתי או טכנולוגי ובנוסף הם יוצרים ומאפשרים שקיפות [30].

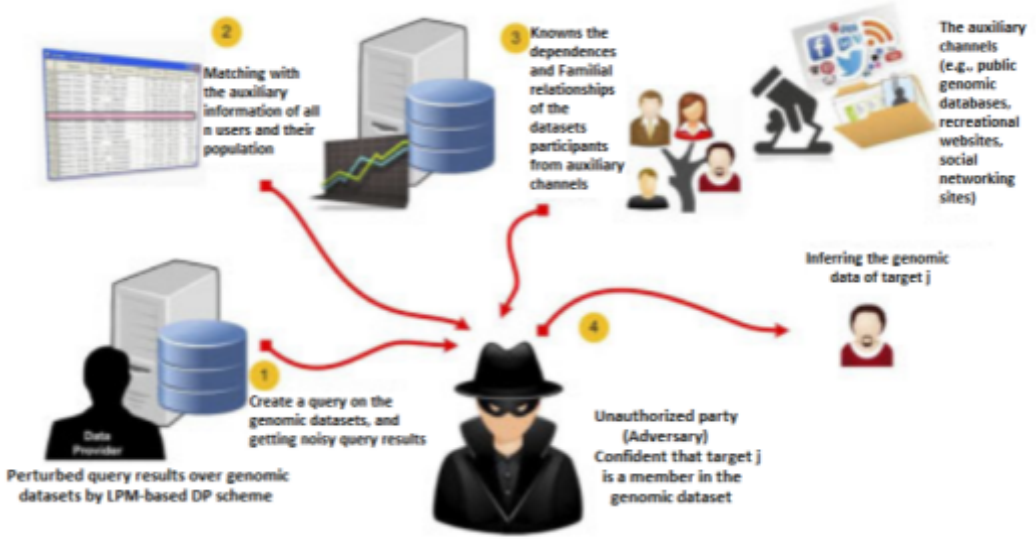
3.4.6.4 הנחת האי-תלות בין רשומות בבסיסי נתונים

אחת ההנחות ברעיון של פרטיות דיפרנציאלית היא שלא קיימת תלות בין רשומות, הנחה זאת הכרחית למחויבות התיאורטיות של פרטיות דיפרנציאלית. נתייחס לצורך העניין לפרטיות דיפרנציאלית מרכזית, בה הרעש המתווסף בתוצאות לשאילתות מבוסס על רגישות השאילתא. הרגישות מודדת כמה תוצאת השאילתא יכולה להשתנות כאשר נתונים של אדם בודד כלשהו מתווספים או מוחסרים מהבסיס נתונים. אותו חישוב מניח כי הפרטים של אנשים לא תלויים זה בזה. כאשר מדובר על נתונים תלויים, הרגישות של אותה שאילתא יכולה להיות גבוהה יותר מאשר עבור נתונים לא תלויים ולכן נוצרת בעיה בכך שחישוב תוספת הרעש לא מתחשבת ברגישות משתנה זו אלא היא קבועה עבור אותה השאילתא [19].

4 הגנת הפרטיות עבור נתונים גנומיים בעזרת פרטיות דיפרנציאלית

4.1 הקושי במימוש פרטיות דיפרנציאלית עבור נתונים גנומיים

כבר הזכרנו שמידע גנומי הוא מידע מאוד רגיש וניתן ללמוד על ידו המון על בן אדם מסויים. לכן, עבור מאגרי מידע רבים המכילים מידע רגיש שכזה, נרצה לפעול על ידי פרטיות דיפרנציאלית. הבעיה בגישה זאת היא החסרון של פרטיות דיפרנציאלית, שהרי היא מניחה כי קיימת אי-תלות בין הרשומות בבסיסי הנתונים אך בסיסי נתונים המכילים מידע גנומי על אנשים לא עומדים בתנאי זה, מפני שיש תלות בין הגנום של אנשים מאותה משפחה ולכן אנו עומדים כאן בקושי חדש בהתמודדות עם הגנה על נתונים גנומיים.



איור 4.1 - מודל התקיפה. תוקף מנצל ידע קודם שיש לו על אנשים שלקחו חלק במחקר כדי לממש מתקפות ולפגוע בפרטיותם [מאמר 2].

ננסה תחילה להבין את הבעייתיות קצת יותר לעומק על ידיי דוגמא בסיסית.
נגדיר בסיס נתונים המכיל מידע גנטי של 6 אנשים. בהנחה שאנחנו יודעים מי הם אותם 6 אנשים, ניתן יהיה להשתמש במקורות חיצוניים כדי ללמוד על הקשר בין אנשים אלו. בעידן שלנו שבו כל כך הרבה מידע נמצא ברשתות חברתיות, ניתן לזהות קשרים משפחתיים בצורה יחסית קלה, בין אם מדובר בעוקבים או תיוגים בתמונות באינסטגרם שיכולים להצביע על קשר כלשהו או אפילו פוסטים תמימים בפייסבוק. בנוסף בהקשרי הגנומיקה יש אתרים רבים העוסקים בחקר היוחסין (גנאלוגיה) כמו myheritage, ancestry.com ו-23andme שהם אתרים העוזרים לבנות עץ משפחתי ולחקור קשרים גנטיים אך למרות כל התנאים וההגנות שאותם אתרים מבצעים מהצד שלהם, לא מעט מהמשתמשים של אותם האתרים בוחרים לחשוף באופן פומבי את העץ משפחה שלהם וכך גם ניתן לנצל אתרים כאלו באופן יחסית תמים ופשוט כדי ללמוד על קשרים משפחתיים.
כעת נניח שאנו יודעים כי לאנשים 2, 4, 5 ו-6 יש קשר משפחתי מתוך כל ששת האנשים במאגר הנתונים, ונניח כי את המידע הזה השגנו באופן שהוסבר מעל, לכן ברור כי בסבירות גבוהה יחסית יהיו להם את אותם התכונות הגנטיות הרי הם יותר קרובים ביולוגית זה לזה מאשר המשתתפים שאינם בני משפחה.

כדי להגן על המידע הגנטי של אותם שישה אנשים, מתבצע שימוש בפרטיות דיפרנציאלית, ורק לשם הדוגמא נהיה יותר ספציפיים ונגיד כי מדובר על פרטיות דיפרנציאלית מרכזית, כלומר הוספת רעש לתוצאות השאילתות על ידי התפלגות לפלס למשל.

כמו שאמרנו בסיס נתונים זה מכיל מידע גנטי, כלומר שאילתא לדוגמא שניתן לבצע על בסיס נתונים זה, היא "לכמה מהפרטים בבסיס הנתונים יש סמן גנטי ספציפי (למשל וריאציה הקשורה למחלה תורשתית)?", שאלה זאת סך הכל מביאה לנו תוצאה לכמות הפעמים שהמאורע בו לבן אדם יש מחלה תורשתית חוזר על עצמו עבור כל האנשים בבסיס נתונים זה וכביכול לא נוכל ללמוד על אף פרט מתוך השישה שום אינפורמציה אישית בזכות הוספת הרעש של הפרטיות הדיפרנציאלית.
נראה איך בכל זאת נוכל לנצל את הקשר המשפחתי. נתשאל את הבסיס נתונים באותה השאילתא בגרסאות שונות ונבדוק האם עכשיו אנחנו מצליחים להשיג מידע פרטי על בן אדם בעזרת הידע שיש לנו על הקשר בין ששת האנשים המופיעים בו.

נתשאל את הבסיס נתונים בשתי השאלות הבאות:

1. "לכמה מהאנשים בבסיס נתונים יש סמן גנטי בוריאציה הקשורה למחלה תורשתית מתוך החמישה האנשים הראשונים?"
2. "לכמה מהאנשים בבסיס נתונים יש סמן גנטי בוריאציה הקשורה למחלה תורשתית מתוך כל ששת האנשים?"

כעת נניח שהתשובות שנקבל לאחר הוספת הרעש הן אלו בהתאמה:

1. ל-3 מהאנשים בבסיס נתונים יש סמן גנטי בוריאציה הקשורה למחלה תורשתית מתוך חמשת האנשים הראשונים.
2. ל- 4 מהאנשים בבסיס נתונים יש סמן גנטי בוריאציה הקשורה למחלה תורשתית מתוך כל ששת האנשים.

לתוצאות שקיבלנו מתווסף רעש כלשהו הרי אנחנו פועלים לפי פרטיות דיפרנציאלית מרכזית, כלומר למשל לשאלה הראשונה, סביר להניח שהתשובה האמיתית היא בכלל 2, 3 או 4, והתשובה האמיתית לשאלה השנייה יכולה להיות למשל 3, 4 או 5 כמובן אלו השערות ולא יותר מכך.
מהתוצאות המתקבלות נוכל להסיק שאם מתוך ה-5 אנשים הראשונים (אשר כוללים שלושה בני משפחה 2, 4 ו-5) יש בערך 3 אנשים העומדים בתנאים אז ככל הנראה שלושת בני המשפחה הם אלו עם המחלה התורשתית ומהתוצאה הטוענת כי מתוך 6 אנשים יש בערך 4 אנשים כאלה אז עבור הבן אדם השישי שהוא בן משפחה של 2, 4 ו-5 בסבירות גבוהה יש את אותה מחלה תורשתית שיש לבני משפחתו, ולמרות שהתוצאה המתקבלת עברה הוספת רעש אנו יכולים לטעון זאת בסבירות יותר גובהה מפני שאנו מכירים בקשר המשפחתי בין אותם אנשים. לעומת זאת אם לבן אדם ה-6 לא היה קשר משפחתי כלל לשלושת האנשים האחרים, התוצאה האומרת כי יש 4 אנשים מתוך 6 אנשים עם אותה מחלה תורשתית כלל לא אומרת לנו כלום הרי מדובר בתוצאה שעברה רעש ולכן לא נוכל לדעת האם מדובר בתוצאת אמת או פשוט אקראיות שנוצרה כחלק מהאלגוריתם הפרטי דיפרנציאלי.

נזכור שכמובן המטרה של פרטיות דיפרנציאלית היא לערפל את הנוכחות או העדרות של פרט כלשהו אך בעזרת הקשר המשפחתי הידוע מראש, ניתן לבצע ניחושים מושכלים ולהפחית את חוסר הוודאות הנוצר מהרעש של האלגוריתם הפרטי דיפרנציאלי.

כדי להבין לעומק את המשמעויות של הקושי שאנו מציגים כאן במימוש פרטיות דיפרנציאלית עבור נתונים גנומיים, נציג דוגמא דומה אך פשוטה יותר וקלה להבנה על ידיי ויזואליזציה ונתונים ממשיים [2].

נניח כי מדובר על מאגר נתונים המכיל נתונים דמוגרפיים על אנשים ופרטים על סניפים, ובאותו מאגר נתונים קיימים נתונים עבור משפחה הכוללת את האב, האם והבן.
בפרטי הסניפים נשתמש במספרים 0, 1 או 2, מספרים אלו מייצגים את תכונות האללים:
הומוזיגוט לאלל ראשי - 0.
הטרוזיגוט - 1.
הומוזיגוט לאלל מינורי - 2.

להגדרה של 3 האופציות מעל לא נתייחס בהרחבה, אך נוכל להיעזר באיור 2.3 כדי להבין אותם.
המטרה שלנו היא ללמוד על המידע הגנטי של הבן ללא גישה ישירה לרשומה שלו (וכמובן גם לא של ההורים). לשם כך נתשאל את בעל המידע בשאילתא הבאה: "מה הוא הסכום של SNP1 עבור

אנשים שגרים בפתח תקווה ברחוב עין גנים?'", ונניח לשם הפשטות כי על ידיי הפרטים הדמוגרפיים בשאילתא צימצמנו את התוצאה רק לאותה משפחה (האב, האם, והבן) הכוללת 3 בני משפחה.
אנו יודעים לעשות זאת הרי ההנחה היא שאנחנו יכולים ללמוד על הקשר המשפחתי ופרטים נוספים על המשפחה על ידיי רשתות חברתיות ובדרכים שונות ומגוונות.
תוצאה לסכום זה כמובן עוברת דרך אלגוריתם פרטי דיפרנציאלי, אך נדגים איך הקשר המשפחתי בכל זאת מקל עלינו לבצע ניחוש מושכל כדי להגיע לפרטים הגנטיים של הבן.
נניח כי הסכום המתקבלת הוא 4 עבור 3 אנשים אלו, ננסה להבין כעת מה הן כלל התמורות שלנו בהתחשב בכך שהתוצאה יכולה להיות לא מדויקת הרי התווסף אליה רעש כלשהו.

פרמוטציות לייצוג ה-SNP עבור k אנשים וסכום i				
2	1	0	מספר האנשים	הסכום שהתקבל
1	2	0	3	4
1	1	1	3	4
1	1	2	3	4

איור 4.2 - דוגמא למספר פרמוטציות לקבלת סכום 4 של ערכי ה-SNP של 3 אנשים.
(1) שורה ראשונה מציגה פרמוטציה תקינה, ובה יש אדם אחד מתוך השלושה בעל וריאציה 2 של אללים ל-SNP1 ושני אנשים בעלי וריאציה 1 עבור אללים לאותו SNP1, ונראה כי קיבלנו במדויק כי התוצאה הכוללת היא 4.
(2) שורה שנייה מציגה פרמוטציה תקינה נוספת, ובה כל אדם עם וריאציה שונה של אללים ל-SNP1, נשים לב שדווקא במקרה זה הסכום הכולל הוא 3 ולא 4, אך כיוון שהתשובה 4 התקבלה על ידיי פרטיות דיפרנציאלית, התוצאה האמיתית יכולה להיות גם 3 ולכן לא נשלול אופציה זו.
(3) שורה שלישית מציגה דוגמא לא תקינה, הרי קיימת הצבה של 4 וריאציות (2 + 1 + 1) כלומר 4 אנשים שונים למרות שאנו טוענים כי מדובר ב-3 אנשים סך הכל, לכן נשלול פרמוטציה זאת.

יש המון פרמוטציות אפשריות שכמובן תלויות בכמות מספר האנשים הנכללים בשאילתא, אך נתייחס במקרה זה רק לשתי השורות הראשונות מאיור 4.2 המציגות לנו שתי אפשרויות לפיזור הוריאציות בין 3 אנשים.
כעת ננסה "להציב" את הוריאציות של האללים עבור SNP1 למשפחה (אב, אם ובן) על ידיי כל אחת מהאופציות.
למשל עבור השורה הראשונה, קיבלנו כי יש אדם אחד בעל הוריאציה 2, ושני אנשים בעלי הוריאציה 1, לכן ננסה להציב את האופציות ולשייך אותם לבני משפחה.

אבא	אמא	בן
2	1	1
1	2	0

איור 4.3 - דוגמה להצבה של הוריאציות לבני משפחה.
(1) שורה ראשונה מתבססת על הפרמוטציה הראשונה מאיור 4.2.
(2) שורה שנייה מתבססת על הפרמוטציה השנייה מאיור 4.2.

יש מספר אופציות להצבה במקרה זה, אך נתייחס כעת להצבות באיור 4.3.
בהצבות הראשונה והשנייה אכן מתקיימת הצבה תקינה לפי הפרמוטציות מאיור 4.2 אך נשים לב שבעוד היותן תקינות, לפי חוק ההפרדה של מנדל שהוזכר בחלק 2.2.1, ההצבה השנייה אינה

הגיונית, נוכל לראות זאת גם לפי איור 2.3, הרי אם לאב יש את וריאציה 1 ולאם יש את וריאציה 2, לבן לא יכולה להיות וריאציה 0 ומכך אנו יודעים לשלול אופציה זאת.

כמובן שיש לא מעט אופציות תקינות נוספות אך יש גם המון תוצאות לא תקינות ולכן נשים לב ליופי שבדבר, בעזרת הקשר המשפחתי שקיים במאגר נתונים זה, למרות הפרטיות הדיפרנציאלית, הצלחנו לצמצם המון תוצאות שלא היינו יכולים לבצע ללא התלות.

דוגמאות אלו מציגות חולשה של פרטיות דיפרנציאלית למרות ההוספה של הרעש להגנת הפרט, וזאת בעזרת ההנחה שאנו מכירים את הקשר המשפחתי בין המשתתפים מראש.

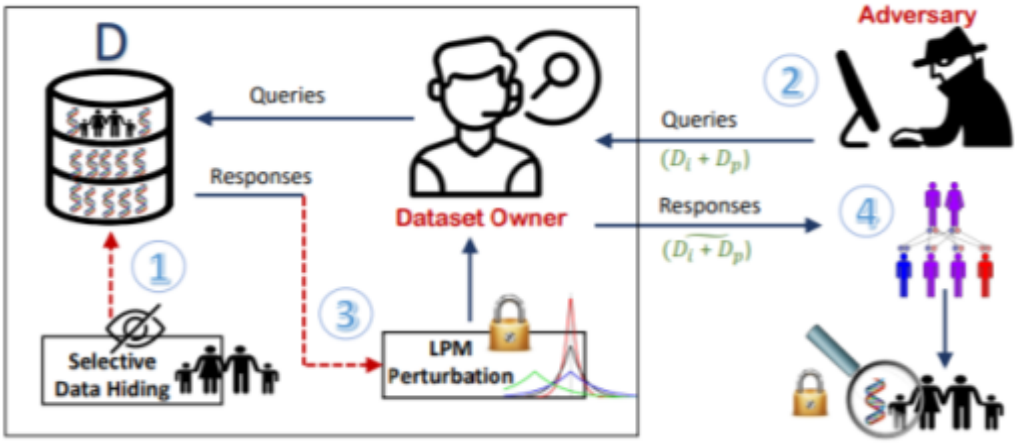
ברור כעת כי למרות הפרטיות הדיפרנציאלית לא הצלחנו למנוע לגמרי את היכולת להסיק מידע גנטי רגיש כאשר קיימת תלות בין הנתונים, וכמובן עבור בסיסי נתונים במחקרים העוסקים בגנומיקה חולשות מסוג זה מהוות בעיה והן מאפשרות פגיעה בפרטיות של אנשים.

4.2 פתרון ואפחות החולשות

אחרי שהבנו איך פרטיות דיפרנציאלית עובדת, אפשר להגיד שהפתרון לבעיה שהצגנו יכול להיות פשוט מאוד והוא כולל סך הכל משחק של הפרמטרים ε ו-δ וליתר דיוק הקטנת הערך של אותם פרמטרים כדי להגדיל את כמות הרעש שנוסף לתוצאות השאילתות. ככל שערך הפרמטרים קטן יותר כך הפרטיות חזקה יותר, מפני שזה מחייב שהתוצאה של שאילתה עבור בסיסי נתונים שכנים תהיה יותר "זהה" (מקטינים את החסם העליון) אך מנגד ככל שמוסיפים יותר רעש כך התוצאות יהיו משמעותית פחות מדוייקות ולא תמיד נרצה בכך בעיקר במחקרים העוסקים בגנומיקה. לשם כך נבחן מספר של פתרונות שנועדו לעזור לנו להתמודד עם ההנחה לאי-תלות עבור מחקרים בתחום זה. בשתי הפתרונות שנסקור, נתייחס לכך שחישוב סכום הערכים של ה-SNPים זהה לחישובו ב4.1, כאשר הסניפים מייוצגים על ידיי תכונות האללים.

4.2.1 הסתרה סלקטיבית של סניפים (SNP)

הסתרה סלקטיבית של סניפים (SNPים) עולה כאמצעי בניתוח נתוני גנומי, ומטרתה לשמור על הפרטיות האישית בלא פגיעה בשימושיות הנתונים לצורך ניתוח. השיטה מסתירה באופן אסטרטגי ערכי SNP מסוימים בתוך מאגר הנתונים, ומפריעה ליכולת לזהות קשרי משפחה או לנצל קשרים ביולוגיים לחשיפת מידע רגיש [3].



איור 4.4 - תהליך הפתרון של הסתרה סלקטיבית של סניפים [מאמר 3].

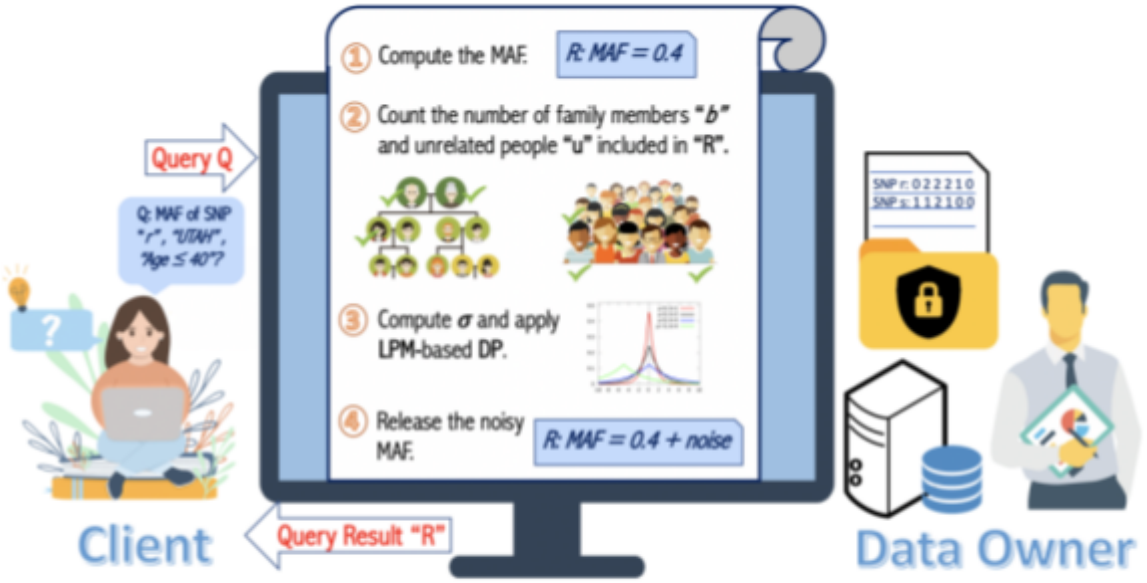
בשיטה זאת, תחילה אנחנו מוחקים או עורכים על ידי הוספת רעש את ערכי ה-SNP של פרטים כדי להפחית את הקשר המשפחתי שקיים במידע גנומי בין בני משפחה, את טכניקת ההסתרה אנו עושים אך ורק בין בני אותה משפחה ובכך אנחנו מצליחים לעשות זאת על ידי פגיעה מינימלית בתועלת המידע במאגר הנתונים.
כעת בזמן תשאול בעל המידע, למשל "מהו סכום ערכי ה-SNP עבור אנשים בטווח גילאים 30-35?", מתווסף רעש לתוצאה של שאילתא זו בעזרת מימוש פרטיות דיפרנציאלית מרכזית, ובנוסף אליה מחזירים תוצאה נוספת והיא כמות האנשים שהיו חלק מהשאילתא אך בחישוב זה נתחשב גם באנשים שערך ה-SNP שלהם מוסתר, כלומר במקרה הזה, מחזירים בנוסף את כמות האנשים שהטווח גילאים שלהם הוא בין 30 ל-35 וכמובן לתוצאה זו גם כן מתווסף רעש על ידי פרטיות דיפרנציאלית מרכזית.
אחד היתרונות המיידיים בפתרון זה הוא שאם המידע כולו נחשף, לא נוכל למצוא קשרים משפחתיים והמון סניפים שתוקף היה יכול לנצל כדי ללמוד מידע על פרטים מוסתרים או שהתווסף להם רעש שזה גם אחד מהחסרונות של פרטיות דיפרנציאלית מרכזית (SPOF).
בעזרת הסתרת ה-SNPים המקשרים בני משפחה שיכולים גם להשתייך לפנוטיפ של מחלה מסוימת אנו מקשים על תוקף שהיה רוצה לנסות לנצל את הקשר המשפחתי שאנחנו מניחים שהוא יודע מראש מפני שהתוצאות שהוא יקבל אינן בהכרח יכללו את כל בני המשפחה.

הצלחנו למצוא פתרון הנמנע מחשיפת פרטיות ובו זמנית פתרון המאפשר לאבד את תועלת המידע באופן מינימלי, אך כמובן עבור מחקרים החוקרים ספציפית את אותם סניפים אותם החלטנו להסתיר, יכולה להיווצר בעייתיות והמידע לא יהיה מספיק מדויק בשבילם. חיסרון נוסף של טכניקה זאת היא הבחירה של הסניפים שאותם צריכים להסתיר, את הבחירה הזאת צריך לבצע עבור כל משתתף במאגר נתונים וכך תהליך השיתוף של המידע עצמו נהיה מאתגר בעיקר ברגע שמתווספים נתונים ורשומות חדשות לבסיס נתונים המקיימים קשר משפחתי כלשהו עם משתתף הנכלל כבר במאגר הנתונים.

GenShare 4.2.2

השיטה GenShare משתמשת במבנה של הנתונים הגנומיים כדי לשפר את הדיוק של התוצאות עבור שאילתות למאגרי נתונים תוך שמירה על פרטיותם של המשתתפים. היא משיגה זאת על ידי

הגברת הרעש המתווסף לתוצאות השאילתות באופן דינמי בהתאם לקשרים המשפחתיים של המשתתפים הקשורים לשאילתא [4].
לעומת אופן הפעולה הרגיל של פרטיות דיפרנציאלית בו אנו קובעים את ערך האפסילון באופן קבוע עבור כל סוג שאילתא, במקרה זה אנחנו מחשבים את ערכו כל פעם מחדש בהתאם לכמות המשתנים התלויים שלוקחים חלק בשאילתא עצמה. למשל עבור השאילתא הקודמת - "מהו סכום ערכי ה-SNP עבור אנשים בטווח גילאים 30-35?", תחילה נחשב את כמות האנשים שיש ביניהם קשר משפחתי וכמות האנשים שאין להם קשר משפחתי הנמצאים בטווח הגילאים 30-35 ובהתאם לתוצאות שנקבל נתאים את ערך האפסילון כך שהפרטיות תגדל ויהיה יותר רעש ככל שיש יותר קשרים משפחתיים.



איור 4.4 - תהליך הפתרון של GenShare להגנה על נתונים גנומיים [מאמר 4].

לסיכום, GenShare מתאים את העקרונות של פרטיות דיפרנציאלית כדי לקחת בחשבון את הקשר בין הנתונים במאגרי נתונים גנומיים, ומספק מנגנון לשימור פרטיות לשיתוף נתונים סטטיסטיים הנגזרים מנתונים כאלה. אך כמובן גם לו יש חסרונות.
חישוב האפסילון כמו שאמרנו מחושב כל פעם מחדש בהתאם לשאילתא עצמה ובהתאם לקשרים משפחתיים הקיימים במאגר הנתונים, מצב זה גורם ל"עומס" חישובי על אוצר המידע.

5. סיכום

הסמינר עוסק בשילוב בין נתונים גנומיים לפרטיות והגנה על הפרט, עם דגש מיוחד על פרטיות דיפרנציאלית כאמצעי להגנה על מידע גנטי רגיש. ההתקדמות המהירה בטכנולוגיות מאפשרת הבנה חסרת תקדים של הגנטיקה האנושית, ומניעה את ההתקדמות ברפואה מותאמת אישית, הבנת מחלות ופיתוחים ביוטכנולוגיים. עם זאת, התקדמות זו מביאה גם חששות משמעותיים לפרטיות, שכן נתונים גנומיים מזוהים מטבעם ויכולים לחשוף מידע רגיש רב על הפרטים ומשפחותיהם.

הסמינר מתחיל במבוא לגנומיקה, ומספק סקירה מקיפה של מבנה הגנום והטכנולוגיות השונות לריצוף, כמו ריצוף סנגר וריצוף מהדור החדש. הוא דן ביישומים של טכנולוגיות אלה, כולל ריצוף גנום מלא (WGS), ריצוף אקסום מלא (WES) וריצוף ממוקד, תוך הדגשת תרומותיהם לרפואה מותאמת אישית, נטיות גנטיות, פרמקוגנטיקה ועוד.

חלק חשוב עוסק בשיתוף מאגרי גנום, ומציג את היתרונות והסיכונים הכרוכים בגישה חופשית ופומבית לעומת גישה מבוקרת לנתונים גנומיים. הדיון עובר לאחר מכן לחשיבות בהגנת הפרטיות, ומפרט את הסיכונים הפוטנציאליים בשיתוף מידע גנומי, כגון זיהוי והסקת פנוטיפים.

כדי להפחית סיכונים אלו, נבחנות טכניקות שונות לשמירת פרטיות. אלה כוללות בקרת גישה, הצפנה, k-אנונימיות ופרטיות דיפרנציאלית. פרטיות דיפרנציאלית שהיא טכניקה הנחשבת חדשה, נבחנת לעומק, כולל הגדרה פורמלית, דיון בצורך בה ויישומים שונים שלה ומימושים שונים בין אם מימוש מרכזי או מימוש מקומי. כמו כן, נדונות המגבלות של פרטיות דיפרנציאלית, כגון הקשיים בעבודה עם מערכות נתונים קטנות וההנחות לגבי אי-תלות בין רשומות.

הסמינר מתמקד בקשיים הספציפיים של יישום פרטיות דיפרנציאלית לנתונים גנומיים, ומציג פתרונות אפשריים, כמו הסתרה סלקטיבית של SNPים ופתרון ה-GenShare, אשר שואפים לאזן בין השימושיות של נתונים גנומיים לצורך בשמירה על הפרטיות.

לסיכום, הסמינר מדגיש את המורכבות שבהגנת נתונים גנומיים ואת תפקיד הפרטיות הדיפרנציאלית. בעוד שנותרו אתגרים משמעותיים, גישות חדשניות ממשיכות להתפתח, במטרה להבטיח שהיתרונות של המחקר הגנומי יתממשו עם פגיעה מינימלית אם בכלל בפרטיות האישית.

6. מקורות

1. Abbott TR, Dhamdhere G, Liu Y, Lin X, Goudy L, Zeng L, Chemparathy A, Chmura S, Heaton NS, Debs R, Pande T, Endy D, La Russa MF, Lewis DB, Qi LS. (2020). Development of CRISPR as an Antiviral Strategy to Combat SARS-CoV-2 and Influenza. Cell. 2020 May 14;181(4):865-876.e12. https://doi.org/10.1016/j.cell.2020.04.020
2. Nour Almadhoun, Erman Ayday, Özgür Ulusoy. (2020). Differential privacy under dependent tuples—the case of genomic privacy, Bioinformatics, Volume 36, Issue 6, March 2020, Pages 1696–1703, https://doi.org/10.1093/bioinformatics/btz837
3. Almadhoun Alserr, Nour & Kale, Gulce & Mutlu, Onur & Tastan, Oznur & Ayday, Erman. (2021). Near-Optimal Privacy-Utility Tradeoff in Genomic Studies Using Selective SNP Hiding. https://doi.org/10.48550/arXiv.2106.05211
4. Almadhoun Alserr, Nour & Ulusoy, Ozgur & Ayday, Erman & Mutlu, Onur. (2021). GenShare: Sharing Accurate Differentially-Private Statistics for Genomic Datasets with Dependent Tuples. https://doi.org/10.48550/arXiv.2112.15109
5. Apple Differential Privacy Team. (2017). Learning with Privacy at Scale. https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf
6. Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, Vitale L, Pelleri MC, Tassani S, Piva F, Perez-Amodio S, Strippoli P, Canaider S. (2013). An estimation of the number of cells in the human body. Ann Hum Biol. 2013 Nov-Dec;40(6):463-71. https://doi.org/10.3109/03014460.2013.807878
7. Jeffrey R. Bishop. (2018). Chapter 6 - Pharmacogenetics, Handbook of Clinical Neurology, Elsevier, Volume 147, 2018, Pages 59-73, ISSN 0072-9752, ISBN 9780444632333, https://doi.org/10.1016/B978-0-444-63233-3.00006-3
8. Bonomi, L., Huang, Y. & Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. Nat Genet 52, 646–654 (2020). https://doi.org/10.1038/s41588-020-0651-0
9. Byrd, J.B., Greene, A.C., Prasad, D.V. et al. (2020). Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet 21, 615–629 (2020). https://doi.org/10.1038/s41576-020-0257-5
10. Irit Dinur and Kobbi Nissim. (2003). Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03). Association for Computing Machinery, New York, NY, USA, 202–210. https://doi.org/10.1145/773153.773173
11. Dwork, Cynthia, Nitin Kohli, and Deirdre Mulligan. (2019). Differential Privacy in Practice: Expose Your Epsilons!. Journal of Privacy and Confidentiality 9 (2). https://doi.org/10.29012/jpc.689

12. Cynthia Dwork and Aaron Roth. (2014), The Algorithmic Foundations of Differential Privacy, Foundations and Trends® in Theoretical Computer Science: Vol. 9: No. 3–4, pp 211-407.
http://dx.doi.org/10.1561/0400000042
13. Ebersberger, I., Metzler, D., Schwarz, C., & Pääbo, S. (2002). Genomewide comparison of DNA sequences between humans and chimpanzees. American journal of human genetics, 70 6, 1490-7.
https://doi.org/10.1086/340787
14. Erlich, Y., Shor, T., Pe'er, I., & Carmi, S. (2018). Identity inference of genomic data using long-range familial searches. Science, 362, 690 - 694.
https://doi.org/10.1126/science.aau4832
15. Fanti, G.C., Pihur, V., & Erlingsson, Ú. (2015). Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries. Proceedings on Privacy Enhancing Technologies, 2016, 41 - 61.
https://doi.org/10.48550/arXiv.1503.01214
16. Haroutunian, M. E., & Mastoyan, K. A. (2021). The Role of Information Theory in the Field of Big Data Privacy. Mathematical Problems of Computer Science, 55, 35–43.
https://doi.org/10.51408/1963-0071
17. Kenny, C.T., Kuriwaki, S., McCartan, C., Rosenman, E.T., Simko, T., & Imai, K. (2021). The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census. Science Advances, 7.
https://doi.org/10.1126/sciadv.abk3283
18. Lin, Z., Owen, A.B., & Altman, R.B. (2004). Genomic Research and Human Subject Privacy. Science, 305, 183 - 183.
https://doi.org/10.1126/science.1095019
19. Liu, C., Chakraborty, S., & Mittal, P. (2016). Dependence Makes You Vulnerable: Differential Privacy Under Dependent Tuples. Network and Distributed System Security Symposium.
https://doi.org/10.14722/ndss.2016.23279
20. MIT ETI. (2021). What is Differential Privacy?. MIT Ethical Technology Initiative.
https://eti.mit.edu/what-is-differential-privacy
21. Derek Muller. (2021). How They Caught The Golden State Killer.
https://youtu.be/KT18KJouHWg?si=Nz-bH79C2towH5Ss
22. Pei, Xiao Meng, Martin Ho Yin Yeung, Alex Ngai Nick Wong, Hin Fung Tsang, Allen Chi Shing Yu, Aldrin Kay Yuen Yim, and Sze Chuen Cesar Wong. (2023). Targeted Sequencing Approach and Its Clinical Applications for the Molecular Diagnosis of Human Diseases. Cells 12, no. 3: 493.
https://doi.org/10.3390/cells12030493
23. Phillips, C. (2018). The Golden State Killer investigation and the nascent field of forensic genealogy. Forensic science international. Genetics, 36, 186-188 .
https://doi.org/10.1016/j.fsigen.2018.07.010
24. Raisaro, J.L., Choi, G., Pradervand, S., Colsenet, R., Jacquemont, N., Rosat, N., Mooser, V., & Hubaux, J. (2018). Protecting Privacy and Security of Genomic Data in i2b2 with Homomorphic Encryption and Differential Privacy. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15, 1413-1426.
https://doi.org/10.1109/TCBB.2018.2854782

25. Rocher, L., Hendrickx, J.M. & de Montjoye, YA. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 10, 3069 (2019). https://doi.org/10.1038/s41467-019-10933-3
26. Stockdale, J.E., Liu, P. & Colijn, C. (2022). The potential of genomics for infectious disease forecasting. Nat Microbiol 7, 1736–1743 (2022). https://doi.org/10.1038/s41564-022-01233-6
27. Sweeney, L., Abu, A., & Winn, J. (2013). Identifying Participants in the Personal Genome Project by Name. Innovation Law & Policy eJournal. https://doi.org/10.2139/SSRN.2257732
28. Thottathil, G.P., Jayasekaran, K., & Othman, A.S. (2016). Sequencing Crop Genomes: A Gateway to Improve Tropical Agriculture. Tropical life sciences research, 27 1, 93-114 . https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4807965/
29. Uffelmann, E., Huang, Q.Q., Munung, N.S. et al. (2021). Genome-wide association studies. Nat Rev Methods Primers 1, 59 (2021). https://doi.org/10.1038/s43586-021-00056-9
30. Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D., Steinke, T., & Vadhan, S.P. (2018). Differential Privacy: A Primer for a Non-Technical Audience. ChemRN: Computational Materials Science (Topic). https://doi.org/10.2139/ssrn.3338027
31. Zarate, O., Brody, J.G., Brown, P., Ramírez-Andreotta, M.D., Perovich, L.J., & Matz, J. (2016). Balancing Benefits and Risks of Immortal Data: Participants' Views of Open Consent in the Personal Genome Project. The Hastings Center report, 46 1, 36-45 . https://doi.org/10.1002/hast.523