

יישום מערכות לומדות (Machine Learning) לניבוי ערכי הטיפוח באמצעות סמנים של נוקלאוטיד יחיד (SNPs) באוכלוסיית בקר לחלב בישראל.

איל סרוסי, דרור שגיא. מכון לחקר בע"ח, המנהל למחקר חקלאי.

תקציר: לניבוי ערכי הטיפוח חשיבות רבה במשק הבקר לחלב. הערכות גנומיות (GEBVs) חיוניות לתוכנית ההשבחה של בקר לחלב. בעבר פותחו שיטות יעילות לברירה גנומית על בסיס קביעה נרחבת של גנוטיפים עם טכנולוגית השבב הגנטי (Illumina BovineSNP50™). עם זאת, יישום אפקטיבי דרש סדרת אימון גדולה, שאינה זמינה עבור העדר הישראלי. בעת ביצוע תחזיות גנומיות באמצעות רגרסיה של פנוטיפים לשבבים גנטיים עם מערכי SNPs בצפיפות גבוהה, מספר הסמנים (p) עולה בהרבה על מספר הדגימות הזמינות (n). רגרסיה נרחבת כזאת, כאשר $n > p$ הופכת לבעיה סטטיסטית, כך שרגרסיה זאת דורשת שיטות לברירה או לייצוג מופחת של מספר הסמנים. עקב אובדן מידע פוטנציאלי באמצעות שיטות אלה, שיטות מבוססות על זיהוי תבניות (pattern recognition) הפכו לחלופות מבטיחות לחיזוי על בסיס נתונים גנומיים. שיטות מבוססות זיהוי תבניות הן חלק ממגוון רחב של מערכות לומדות (Machine Learning), המסוגלות לאתר ולאפיין אינטראקציות גנים המתבטאות כדפוסי גנוטיפ. בהקשר זה, רשתות עצביות מלאכותיות (neural nets) הן כלים חזקים לניבוי גנומי, תוך ניצול של הממד המלא של הנתונים. רשתות אלה נחשבות כקירוב אוניברסלי של פונקציות מורכבות ומייצגות של פונקציות כל שהן, ובכך מסוגלות לקרב יחסים לא ליניאריים בין הסמנים הגנטיים והפנוטיפים הנצפים, באופן אינטואיטיבי. תכונה זאת מושגת באמצעות יכולת למידה אוניברסאלית הניתנת על ידי אלגוריתמים מסוימים, שמיושמים לאימון הפונקציה. בקשר לניבוי הגנומי, אופן פעולה זה מוביל ליתרון נוסף, למשל לא נדרש לקבוע מראש מודל הורשה גנטי, כך שקימת מראש התאמה לכל ארכיטקטורה גנטית אפשרית של תכונות המטרה. המחקר השווה ביצועי מערכות אלה בניבוי ערכי הטיפוח על בסיס האוכלוסייה המקומית בלבד, לביצועים שמתקבלים בשיטה הנוכחית שמפעילה אלגוריתם שמרני על מדגם נרחב יותר. המחקר כלל בקרת הרשומות הגנטיות בספר העדר והתאמתן לנדרש עבור ניבוי ערכי הטיפוח מדויקים (תוצאות אלה פורסמו בכתב העת GENES); זיהוי תבניות בבסיס הנתונים כאמצעי להפחתת מורכבותו; ובדיקה ראשונית של אלגוריתמים לניבוי ערכי הטיפוח מדויקים בעזרת בסיס הנתונים המשופר. בהמשך, נבחן אלגוריתמים לחיזוי ערכי טיפוח על בסיס למידת מכונה הניסויים הראו שהאלגוריתם לניבוי ערכי חמ"מ של פרים צעירים בעזרת רשת עצבית מלאכותית וייצוג הגנום כתמונה מגלה תוצאות דומות לאלה שמוסגות במערכת הקיימת. כדי להשיג שיפור הוחלט לחקור בתכנית המשך (362-0774), שילוב של נתונים גנומיים קיימים עבור פרות; והוצע (בפרסום נוסף בכתב העת GENES) שילוב באינדקס הטיפוח של תכונה אלטרנטיבית להערכה גנטית של דלקת עטין על בסיס תפוקת החיים.

1. מבוא

לחיזוי על בסיס נתונים גנומיים וניבוי ערכי הטיפוח חשיבות הולכת וגוברת בגידול בעלי חיים ואף בגנטיקה האנושית (Nayeri וחבריו, 2019). בשנים האחרונות, ההתפתחות המהירה של טכנולוגיות לקביעת רצף נוקלאוטידים, בעלות תפוקה גבוהה, אפשרה קביעת גנוטיפים בהיקף נרחב של אלפי סמנים גנטיים, למשל סמנים של נוקלאוטיד יחיד (SNPs) בגנום של האדם ובעלי חיים. מערכים צפופים כאלה של סמנים מולקולריים מאפשרים מיפוי של הגנום לאתרים לתכונות כמותיות (QTLs) של פרטים ואוכלוסיות באמצעות קישור חוסר שיווי משקל בין סמנים בודדים לבין שונות גנומית. למרות שניתן ליישם מבחנים סטטיסטיים לאיתור גנים כמותיים עיקריים (Seroussi, 2009), ההנחה היא שאלה נדירים ושרוב התכונות נקבעות על ידי מספר גדול של גנים עם השפעות קטנות (Falconer & Mackay, 1996), שמצטברות באופן אדטיבי לכדי ביטוי של הפנוטיפ של הפרט. בשנת 2001, על בסיס הנחה זאת, Meuwissen וחבריו, הציעו גישה פורצת דרך לחיזוי של תכונות מורכבות על בסיס נתונים גנומיים. הכותבים יישמו מודל רגרסיה של גנום שלם כאשר הפנוטיפים מנותחים באופן לינארי על אלפי סמנים בו זמנית. בניגוד לתחזיות מבוססות אילן יוחסין, גישה זו מאפשרת חיזוי מדויק של ערכים גנטיים של בעלי חיים, בשלב מוקדם של החיים. מספר שנים אחר כך, Schaeffer (2006) הציב את היסודות להצלחה רבה בפועל של גישה זאת של טבע את המושג Genomic-Estimated Breeding Values (GEBV). בשנים האחרונות יושם חיזוי וטיפוח על בסיס גנומי בעלי חיים במספר משקים ותוכניות טיפוח (VanRaden וחבריו, 2009). במקביל, המשיכה התופעה של "חוסר תורשה של תכונות מורכבות" למשך עניין עולמי בקהילה המדעית של חוקרי בעלי חיים ונושאי מחקר קשורים כמו רפואה אנושית (Manolio וחבריו, 2009). למרות שבעת ההיא, מחקרים כבר סיפקו תובנות יקרות ערך על בסיס גנטי, הם הסבירו חלק קטן יחסית של תורשת התכונות המורכבות. הסברים רבים לחסר זה הוצעו וכמה קבוצות עבודה ציינו כי הבעיה המהותית עשויה להיות חוסר התאמה בין הארכיטקטורה הגנטית של התכונות המורכבות והטכניקות הסטטיסטיות המיושמות לאפיון (Yang וחבריו, 2010). לפיכך, גישה מסבירה חשובה היא קיומן של השפעות לא תוספתיות ואי-ליניאריות גבוהה. מנגנון גנטי עשוי לכלול אינטראקציות מורכבות בין גנים, בין גנים לתנאים סביבתיים, או אפקטים אפיגנטיים שלא מיוצגים במלואם על ידי מודלים ליניאריים תוספתיים. בהקשר זה גברה ההתעניינות בשימוש בשיטות לא-פרמטריות או פרמטריות-למחצה עבור חיזוי גנומי של תכונות כמותיות כדי להסביר את השפעות הגנים הלא תוספתיים, את האינטראקציות הלא-ליניאריות המורכבות יותר כמו גם אינטראקציות גנוטיפיות-סביבתיות (Gianola & van Kaam, 2008; de los Campos וחבריו, 2010). עם זאת, בעת ביצוע תחזיות גנומיות באמצעות רגרסיה של פנוטיפים לשבבים גנטיים עם מערכי SNPs בצפיפות גבוהה, מספר הסמנים (p) עולה בהרבה על מספר הדגימות הזמינות (n). רגרסיה נרחבת כזאת, כאשר $n > p$ הופכת לבעיה סטטיסטית, כך שרגרסיה זאת דורשת שיטות לברירה או לייצוג מופחת של מספר הסמנים (de los Campos וחבריו, 2013). עקב אובדן מידע פוטנציאלי באמצעות שיטות כאלה, שיטות מבוססות על זיהוי תבניות (recognition pattern) הפכו לחלופות מבטיחות לחיזוי על בסיס נתונים גנומיים. שיטות מבוססות זיהוי תבניות הן חלק ממגוון רחב של מערכות לומדות (Machine Learning), המסוגלות לאתר ולאפיין אינטראקציות גנים המתבטאות כדפוסי גנוטיפ (Musani, 2007). בהקשר זה במיוחד, רשתות עצביות מלאכותיות (neural nets) הן כלים מעניינים לניבוי גנומי, תוך ניצול של הממד המלא של נתונים. רשתות אלה נחשבות כקירוב אוניברסלי של פונקציות מורכבות (Hornik וחבריו, 1989). על פי משפט קולמוגורוב (Kolmogorov, 1957), רשתות אלה מוכחות כמייצגות של פונקציות כל שהן, ובכך מסוגלות לקרב יחסים לא ליניאריים בין הסמנים הגנטיים והפנוטיפים הנצפים, באופן אינטואיטיבי. תכונה זאת מושגת באמצעות יכולת למידה אוניברסאלית הניתנת על ידי אלגוריתמים מסוימים,

שמישמים לאימון הפונקציה. בקשר לניבוי גנומי, אופן פעולה זה מוביל ליתרון נוסף, למשל לא נדרש לקבוע מראש מודל הורשה גנטי, כך שקימת מראש התאמה לכל ארכיטקטורה גנטית אפשרית של תכונות המטרה. הערכות גנומיות (GEBVs) חיוניות לתוכנית ההשבחה של בקר לחלב. בארה"ב, שימוש בהערכות האלה הביא להאצת קצב ההתקדמות הגנטית. למשל בזן הולשטיין, מוערך שעבור תכונות הייצור העיקריות של כמויות חלב, חלבון ושומן, בחמש השנים האחרונות, נצברה בממוצע 58% יותר התקדמות מאשר בתקופה של חמש שנים קודמות (Norman וחבריו, 2020). בישראל הגנוטיפים לסמני SNPs נקבעים באופן שיגרתי לכל הפרטים שעשויים להיות מעורבים בתוכנית ההשבחה (ראה תוצאות הקדמיות). עדר הבקר לחלב בישראל קטן יחסית לזה של מדינות כארה"ב, שם הסטטיסטיקאים נהנים מיחס משופר של $p < n$, והכוח הסטטיסטי לחישוב הערכות הגנומיות נשען על אנליזה של עשרות אלפי פריים. בעבר ניסינו להתמודד מקומית עם בעיה זאת על ידי בחינת יכולת חיזוי של קבוצות משנה נבחרות של SNPs באוכלוסיית בקר החלב המקומית (Weller וחבריו, 2014). או שימוש הפלוטיפים, שמהווה למעשה אמצעי לזיהוי תבניות שמוביל לייצוג מופחת של מספר הסמנים (Baruch וחבריו, 2006). נעשה אף ניסיון פרלימינרי ליישם מערכות לומדות להתמודדות עם הבעיה (Seroussi & Seroussi, 2013; תוצאות הקדמיות). ניסיונות אלה הדגימו הצלחה חלקית בלבד וכפתרון לבעיית המדגם הקטן העדיפו הגורמים המקצועיים בענף לחבור לחברה הולנדית שמבצעת את ניתוח הנתונים על בסיס משותף של אוכלוסיית הבקר לחלב בשתי המדינות. בשנים האחרונות, בצד גידול טבעי באוכלוסיית המדגם הישראלית, שופרו הרשומות ואומתו הדגימות של אוכלוסיית פרי הזרעה הישראלית והתברר שבמדגם הישן מספר ניכר של רשומות (~5%) היה לקוי. בעקבות התפתחויות אלה וההכרה הבין לאומית הגוברת בעליונות של יישום מערכות לומדות בתחום הגנטי (Nayeri וחבריו, 2019), המחקר המוצע ישווה ביצועי מערכות אלה בניבוי ערכי הטיפוח על בסיס האוכלוסייה המקומית בלבד, לביצועים שמתקבלים בשיטה הנוכחית שמפעילה אלגוריתם שמרני על מדגם נרחב יותר.

2. מטרת המחקר:

מטרת העל של המחקר המוצע היא להאיץ את קצב ההתקדמות הגנטית בתוכנית ההשבחה של בקר לחלב בישראל. מטרה זאת נקשרת למטרות הספציפיות הבאות:

2.1 בקרת הרשומות הגנטיות בספר העדר והתאמתן לנדרש עבור ניבוי ערכי הטיפוח מדויקים: הקביעה המדויקת של גנוטיפ הסמן ושל מקומו הכרומוזומלי היא בסיס חשוב לזיהוי תבניות בגנום. בהתאם לשיפורים האחרונים בבניית גנום הבקר יש להטמיע בספר עדר את המידע העדכני ביותר ולוודא שלמות הנתונים ונכונותם על ידי הצלבה לקרובי משפחה בכפוף לחוקי ההורשה המנדליים.

2.2 זיהוי תבניות בבסיס הנתונים כאמצעי להפחתת מורכבות: המצב בבקר לחלב הוא שמספר פרי הזרעה בודדים מהווים מקור להורשת הגנים לכלל האוכלוסייה. הגודל האפקטיבי של אוכלוסיית בקר הולשטיין בעולם מוארך בכמאה פרטים. במצב זה סמנים גנטיים סמוכים תורמים מעט למידע ורק מגדילים את מורכבות המדגם. מניסיונו באזור של מיליון בסיסים (כעשרה סמני BovineSNP50) די בכעשרים הפלוטיפים להסביר את רוב המוחלט של השונות הגנטית (>95%). כלומר ניתוחים סטטיסטיים על בסיס הפלוטיפים מאפשרים הפחתת מורכבות המדגם במספר סדרי גודל. בעבר פתחנו תוכנה מותאמת לזיהוי הפלוטיפים בעדר הבקר (Baruch וחבריו, 2006). התאמת תוכנה זאת ואחרות לניתוח נתוני העדר כיום היא מטרה חשובה לזיהוי תבניות בבסיס הנתונים כאמצעי להפחתת מורכבות.

2.3 בחירת האלגוריתם היעיל ביותר לניבוי ערכי הטיפוח מדויקים בעזרת בסיס הנתונים המשופר: על ידי חלוקת המדגם לסדרות אימון וקבוצות מבחן נשווה ניבוי ערכי הטיפוח באמצעות שיטות ניתוח קלסיות וחדשניות ונבחר את המוצלחת שביניהן.

2.4 הטמעת השימוש במערכות לומדות לצורך ניהול משק החי.

מעבר לבעיה הספציפית של ניבוי ערכי הטיפוח, יישום מערכות לומדות למתן תשובות לחקלאי בתחומי ניהול וממשק במשק החי הוכיח את יעילותו בשנים האחרונות (Nayeri וחבריו, 2019). למשל נדרש מודל למידה להזרעה במועד מיטבי. מחקר שיטמיע שימוש במערכות החדשניות האמורות בישראל יאפשר קידום החקלאות בישראל והתחרותיות שלה בעולם.

3. חומרים, תהליכי ושיטות עבודה.

נרכשו שתי תחנות עבודה שפועלות כשרתים עצמאיים והוסבו למערכת הפעלה לינוקס Ubuntu ייעודית לביואינפורמטיקה, הכוללת בתוכה תוכנות ביואינפורמטיקה פופולריות כגון Blast לחיפוש והשוואה של רצפים ביולוגיים, ו-Samtools, לניתוח של רצפי נוקלאוטידים וסמנים שהתקבלו משבבים גנטיים ומריצוף עמוק של הגנום. הותקנו תוכנות קוד פתוח נוספות שמפעילות מערכות לומדות ביניהן *Scikit-learn* (<https://scikit-learn.org/stable>) ו-*TensorFlow* (<https://www.tensorflow.org>). שתי התחנות הזוהו הוצבו בשתי המעבדות של החוקרים המעורבים, כך שניתן להשוות ביצועי אלגוריתמים שונים ביניהן. נקלטו נתוני ספר העדר, ובוצעו אנליזות כדי להשלים את הרשומות החסרות באמצעות, נתוני קביעת הרצף הגנומי של 17 פרים ולתקן שגיאות על ידי הצלבה לקרובי משפחה בהתאם למטרה 2.1. בהמשך, נבדקו אלגוריתמים יעילים לשחזור הפלוטיפים ונבחרה תוכנת *FINDHAP* (<https://aipl.arsusda.gov/software/findhap>) שמאפשרת צמצום מורכבות בסיס הנתונים על ידי השלמה חישובית (imputation) של גנוטיפים חסרים בבסיס הנתונים בכפוף למטרה 2.2.

במסגרת ניסיון לחזות ערכי טיפוח עבור פרים בלמידה עמוקה (רשתות עצביות) נעשתה המרה של גנוטיפים של פרים שהתקבלו מציפ גנומי לתמונות. התחום של תחזית ע"י תמונות מפותח ביותר ומכאן הרציונל לנסות את הגישה הזאת. בעזרת הספרייה *Fastai* (Howard וחבריו, 2020), שפועלת תחת *PyTorch* (<https://pytorch.org>) בשפת התכנות *Python* נבחנו רשתות עצביות שונות. ביניהן ספרייה בשם *Timm* (<https://pypi.org/project/timm/>), שמכילה מגוון רחב של רשתות עצביות מוכנות לשימוש, מהמתקדמות ביותר ללימוד תמונה, וניתן ליישם אותם דרך *Fastai*. כך ניתן לנסות בקלות ארכיטקטורות שונות של *Convolutional Neural Networks (CNN)* כגון: *He Resnet* (חבריו, 2016) ו-*EfficientNet* (Tan וחבריו, 2019). המנגנון של *Fastai* מאפשר שליטה קלה בפרמטרים שונים לצורך ניסיונות לשיפור הדיוק של התחזיות. גישה חדשה לשימוש ב-*CNN* במידע טבלאי (*Spikelab*), שזכתה במקום השני בתחרות *Kaggle* היא אפשרות נוספת שאנו בודקים לתחזית ערכי חמ"מ על בסיס הנתונים הגנומיים.

4. תוצאות.

על ידי אנליזה של הגנומים של 17 פרי הולשטיין ישראלים, השווינו את דיוק הגנוטיפים בין רצף גנום שלם (WGS) וטכניקות מבוססות שבבים גנטיים. באמצעות פרוטוקול עיבוד הנתונים הסטנדרטי (*GATK*), וריאנטים ברצף הנוקלאוטידים שמופו לגנום הייחוס (*ARS-UCD1.2*) הושוו לגנוטיפים מהשבב הגנטי (*Illumina BovineSNP50*) (*BeadChip*) ולגנוטיפים שאופיינו בשיטה חלופית, שפיתחנו. שיטה זאת מחקה מבחינה חישובית את הליך ההיברידיזציה שמתבצעת בטכנולוגית השבב הגנטי, על ידי מיפוי קטעי הרצף ל-50 הבסיסים המאפיינים כל סמן גנטי לרצפי המקור של השבב. מספר חוסר ההתאמות בין הגנוטיפים של השבב ושל הריצוף הגנומי היה נמוך (0.2%). עם זאת, ל-17,197 (40% מהסמנים האינפורמטיביים) הייתה וריאציה נוספת בטווח של 50 בסיסים מאתר השונות המקורית. וריאציה זאת מפריעה לקביעת גנוטיפים המבוססת על היברידיזציה. כתוצאה מכך, בהתייחס לשגיאות גנוטיפ שנצפו, קביעת הגנוטיפים שמבוססת על היברידיזציה הייתה שונה באופן משמעותי ושיטתי מהגנוטיפים שהתקבלו מהרצף הגנומי, והציגה אפקטים דמויי אלל חסר (*null allele*) ושגיאות מנדליאניות (<0.5%), בעוד שהאלגוריתם ה-*GATK* של זיהוי מקומי של הפלוטיפים קבע בהצלחה את הגנוטיפים בסמנים בהם סמוכה וריאציה נוספת לאתר המקור. ממצאים אלה מצביעים על כך שתכנון השבב צריך להימנע מאזורים גנומיים

פולימורפיים המועדים לשונות נוספת ושניתן להשתמש בנתוני הריצוף הגנומי כדי לתקן גנוטיפים שגויים. תוצאות אלה סוכמו במאמר שהתפרסם לאחרונה בכתב העת GENES (Gershoni וחבריו, 2022).

הגנוטיפים הבודקים של הפרים, שרוצפו גנומית, שולבו במדגם של 1750 פרי הזרעה עם שבב גנטי. מדגם משופר זה מבוסס על נתוני ספר העדר שנצברו עד ובשנת 2018, כאשר לפרים הצעירים ביניהם עדיין חסרים רשומות אמת של ביצועי הבנות. עם התקדמות הזמן, צפוי שבשנה הבאה לרוב הפרים הצעירים במדגם זה יצברו רשומות שיאפשרו קביעת ערכים גנומיים מהימנים לתכונות האינדקס. בעזרת התוכנה *FINDHAP*, בוצעה השלמה חישובית (imputation) של גנוטיפים חסרים בבסיס הנתונים המשופר לסמנים נבחרים. נבחרו 50,392 סמנים גנטיים כבסיס להמשך המחקר (Gershoni וחבריו, 2022). על בסיס סמנים אלה, בוצעו הרצות הקדמיות לבחינת יכולת ניבוי החמ"מ על ידי למידת מכונה של עשרת הפרים הצעירים ביותר בלוח הפרים, שלאחרונה רשומות ביצועי הבנות שלהם הגיעו להישנות סבירה (>75%). בין המודלים שהורצו *LASSO* (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html), *Random forest regressor* (<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>), ו-*GPNET* (<https://github.com/gwyns6287/GpNet>). עבור עשרת פרים אלה, הקורלציות שהתקבלו בין הערכי החמ"מ האמתיים והחזויים מפורטות בטבלאות 1 ו-2. תוצאות אלה הוצגו בכנס בקר וחלקן סוכמו בתקציר (שגיא וחבריו, 2022).

בהמשך, נבחנו ארכיטקטורות שונות של CNN כולל Resnet ו-EfficientNet שרובן לא סיפקו תחזיות מוצלחות (תוצאות לא מוצגות, למעט המוצג בדיון).

בדיעבד הסתבר שהניסוי הראשון (טבלה 1) כלל פרים שהביצועים שלהם היו רחוקים מאוד מהצפוי לפי ממוצע ההורים (מתאם שלילי של -0.47). ההטיה הממוצעת (440 יחידות חמ"מ) הייתה גבוהה. הטיה זאת חושבה לפי ממוצע של הערכים האבסולוטיים של הפרשים בין ערכי החמ"מ בפועל וערכים הצפויים לפי ההורים. תוצאה זאת נובעת בעיקר משרידות נמוכה של בנות שלא צלחו את התחלובה הראשונה, תופעה שהנזק שלה לחקלאי גדול, ושיתכן שמצדיקה שינוי באינדקס הטיפוח כדי לצמצמה. דלקת עטין היא הסיבה הישירה הנפוצה ביותר שמובילה לקריסת פרות בתחלובה הראשונה. תכנית ההשבחה השיגה התקדמות מועטה בהפחתת שכיחות דלקת העטין. תיעוד של דלקת העטין הוא בעייתי עקב היעדר רגישות של הנתונים והסתמכות על אבחון קליני, לעיתים קרובות חסרים מקרים תת-קליניים ו/או על מדידות חודשיות של ספירת תאים סומטיים (SCC). המדד הנוכחי לדלקת בעטין הוא ממוצע תחלובה של מדד התאים הסומטיים (LSCS). בדקנו שני מערכי נתונים: (1) 148 עגלות מחולקות לקבוצות שאינן נגועות, נגועות תת-קליניים וקבוצות דלקת עטין קלינית; (2) נתונים מ-89,601 עגלות הולשטיין ישראלי, במהלך אותה תקופה מחולקים ל"עטין בריא" (UH) ו-"לא בריא" (UNH) לפי סף של 120,000 SCC תאים/מ"ל בכל תשע רישומי החלב החודשיים. בניסוי 1, לעגלות הלא נגועות היו באופן מובהק ($p < 0.05$) יותר המלטות, ימי ייצור ותפוקת חיים, בהשוואה לקליניות ותת-קליניות. בניסוי 2, לעגלות UH (20.3%) היה חלב גבוה יותר באופן מובהק ($p < 0.01$) לכל החיים, ימי ייצור ותחלובה. באנליזה של הנתונים לפי אבות, אותם ניתוחים גילו הבדלים באחוזים של בנות UH בין קבוצות האבות. ייצור חלב לכל החיים נמצא בקורלציה ($r = +0.83, p < 0.001$) עם מצב בריאות העטין. סף SCC של פחות מ-120,000 תאים/מ"ל במהלך כל מדידות התחלובה הראשונות הצביע על עטין בריא, מה שסיפק תובנה רבת ערך שתכונה דיכטומית זאת היא יתרון לחישוב אינדקס טיפוח של תפוקת החיים על פני LSCS. תוצאות אלה סוכמו ופורסמו בכתב העת GENES (Leitner וחבריו, 2024).

בתחילת המחקר ניצלנו את הזמינות של נתוני הרצף העמוק של 17 גנומים של פרי הולשטיין ישראליים ששימשו להזרעה מלאכותית של פרות חלב. נתונים אלה אפשרו השוואה של דיוק קביעת הגנוטיפים בין טכניקות השבב הגנומי והריצוף הגנומי. הריצוף סיפק מידע מפורט שאפשר מיפוי מדויק לגנום הייחוס וקביעת גנוטיפ שמתגברת על שיבושים שנגרמים מווריאציה גנטית נפוצה, נוספת ובלתי צפויה באזורים שליד הסמן. שיבושים בקביעת גנוטיפ עשויים גם להסביר חלק מהתורשה החסרה שנתקלה במחקרים מבוססי שבבים גנטיים (Gershoni וחבריו, 2022). כדי לקבוע גנוטיפ, טכנולוגית השבב עושה שימוש רק באחד משתי הכיוונים האפשריים. לפיכך, במדגם שלנו, הווריאציה הנוספת עשויה להסתכם ב- 25% מהסמנים האינפורמטיביים. כך שהגנוטיפ מבוסס ריצוף עמוק משקף טוב יותר את הגנוטיפים האמתיים מאשר גנוטיפ מבוסס היברידיזציה. אכן, בהתחשב בסוגיית התורשה החסרה, הוצע בשלב מוקדם שכאשר מחיר הריצוף יורד, יהיה זה הגיוני להפסיק להשתמש ב-SNPs ולהתחיל לבצע ריצוף של גנומים שלמים. לחלופין, הוכח כי בבני אדם, השלמה חישובית (imputation) של נתוני הגנוטיפ יכולה להוריד את התורשתיות המחושבת החסרה לתכונות הגובה ומסת הגוף לרמות זניחות (Gershoni וחבריו, 2022). עם זאת, ראוי לציין שרוב תוכניות ההשבחה בימינו משתמשות בגרסאות מותאמות של שבבים שאומצו לצרכיהם ויתכן ששבבים המתוכננים בקפידה נמנעים מאזורים גנומיים הפולימורפיים ביתר (Gershoni וחבריו, 2022). על בסיס מסקנות אלה בוצעה השלמה חישובית (imputation) של גנוטיפים חסרים בבסיס הנתונים של ספר העדר משנת 2018, לאחר שהוטמעו בו הגנוטיפים שנקבעו לפי הריצוף הגנומי העמוק. לבדיקה ראשונית, נתונים אלה הוזנו לתוכנות למידת מכונה שהניבו תחזיות לערכי חמ"מ בדיוק גבוה (מתאם >0.7) עבור פרים ששמשו בעבר ובדיוק נמוך (<0.3) עבור פרים יותר צעירים (למשל מדגם 10 הפרים האחרונים במבחן פרים 2021-10). כמוצג בטבלה 1, בעיה זאת חוזרת גם בשיטות הניבוי הנוכחיות (ממוצע חמ"מ הורים, תחזית חמ"מ הולנד, ותחזית חמ"מ אינטרבול).

טבלה 1: חמ"מ מדוד לעומת צפוי במדגם עשרת הפרים האחרונים במבחן פרים 2021-10

GPNET	תחזית אינטרבול	תחזית הולנד	ממוצע הורים	אב האם	אב	שם	מס. פר	חמ"מ בפועל
281	-67	773	724	ג'קי	פומיקי	פוג'י	9225	588
403	412	742	663	סימי	איסר	אסמון	9210	517
470	543	754	427	ג'רום	ברולו	באגי	9194	470
322	75	788	626	ג'קי	פומיקי	פירקלו	9224	418
262	140	774	665	ג'רום	אדלוויס	אינייסטה	9191	358
233	494	824	837	סופון	סופרשוט	סגאן	9222	209
-154	-78	660	615	ווינס	בוסני	בטומי	9192	173
63	158	854	666	מניפולד	ביוואס	בימיני	9193	12
-44	63	738	855	ג'מצי'	סיטבון	סוגיטה	9203	-22
128	198	739	753	ווינס	ג'מצי'	ג'פסי	9216	-208
137	257	513	440	הטיה ממוצעת				
0.70	0.14	0.02	-0.47	מתאם עם חמ"מ בפועל				

טבלה 2: חמ"מ מדוד לעומת צפוי במדגם עשרת הפרים האחרונים במבחן פרים 2022-3

GPNET	תחזית אינטרבול	תחזית הולנד	ממוצע הורים	אב האם	אב	שם	מס. פר	חמ"מ בפועל
278	413	986	879	ארגמן	סגריר	סרג'	9244	1077
473	508	1036	817	ארגמן	איסר	אכילס	9248	970
165	260	808	659	מקרו	סיגרה	סימקו	9234	552
-30	206	769	649	דנקול	וידל	ויראל	9233	545

-147	178	767	773	זקא	ראג'ר	רגנאר	9246	463
189	129	459	357	ארגמן	סימונה	סהרה	9235	439
63	270	950	812	ארגמן	ג'מצי	ג'ון	9229	211
245	284	946	802	פטרשה	בוז'י	בופה	9227	161
94	229	959	812	ארגמן	ג'מצי	ג'ורג'	9230	100
-86	153	777	805	ג'וניור	גרסון	ג'אג'ו	9256	33
348	278	409	368	הטיה ממוצעת				
0.58	0.71	0.21	0.06	מתאם עם חמ"מ בפועל				

למעשה, במדגם המוצג בטבלה 1 (הממוינת לפי חמ"מ), נכשלו כל התחזיות משום שהפר שדורג ראשון הוא בעל תחזיות חמ"מ נמוכות מאלה שנחזו עבור הפר שדורג אחרון בפועל. יתר על כן, יתכן שהמתאם השלילי שנצפה עם ממוצע הורים מעיד שיתכן ועדיף להסתמך על הניבוי הגנומי בלבד ולא לשלב את הערכים של ממוצע ההורים בתחזית הגנומית, כפי שמתבצע כיום. החמ"מ של הפרים שדורגו נמוך במדגם טבלה 1, נפגע בעיקר על ידי שרידות נמוכה של בנות הפר כלומר יציאה של פרות אלה במהלך התחלובה הראשונה. ניכר שיש לשפר את יכולת הניבוי של תכונת השרידות בתחלובה הראשונה שיש לה ערך כלכלי קריטי. בהתאם למסקנות אלה ולעובדה שנוספו בדיעבד נתוני שבבים גנטיים לפרים הרלוונטיים למדגם 2018 (בעיקר פרי חו"ל שנשלחו לבדיקה באיחור) התבצעה בניה של בסיס נתונים גנומיים עדכני ובשנת המחקר השנייה שולבו הנתונים החדשים עם נתוני הריצוף העמוק כדי לבצע השלמה חישובית (imputation) משופרת של גנוטיפים חסרים ובחינה מעמיקה של אלגוריתמים לחיזוי על בסיס למידת מכונה. חישוב חוזר על מדגם חדש הראה שהמערכת הקיימת השתפרה כנראה משום שבמדגם השני (טבלה 2), ממוצע ההורים של הפר הטוב ביותר היה אכן הגבוה משאר הנדגמים. ניסינו גם אלגוריתמים מבוססי CNN אולם הניסיונות שנעשו לא התקבלה תחזית מוצלחת עבור פרים חדשים, מתן משקל יתר לפרטים חדשים גם לא שיפר את התחזיות, מה שמראה שהשימוש הפשטני ב-CNN עבור תחזיות אלו איננו מתאים. אכן (Pook וחבריו 2020) טענו ששימוש פשטני ב-CNN אינו מתאים לנתונים גנומיים. תוצאות ראשוניות של שימוש בגישה החדשה ל-CNN במידע טבלאי Spikelab הניבו תוצאות מרשימות. לצורך הבדיקה נלקחו נתוני פרים עם אמינות גבוהה ממבחן הפרים האחרון (הישנות <0.75). בסיס הנתונים כלל 1663 פרים. אחוז אחד מבסיס הנתונים (17 פרים הצעירים ביותר) שימשו כקבוצת תיקוף ושאר הפרים שימשו לבניית המודלים לכל התכונות שמרכיבות את האינדקס. לאחר שקלול תוצאות מודלים אלה, התקבלה תחזית חמ"מ עם מתאם גבוה לנתוני מבחן הפרים ($r=0.75$) עבור 17 פרי התיקוף. כך ש-Spikelab מהווה אפשרות מעניינת לתחזית על בסיס נתונים גנומיים.

למרות שאלגוריתמים לחיזוי ערכי טיפוח על בסיס למידת מכונה הניסויים הראו שהאלגוריתם לניבוי ערכי חמ"מ של פרים צעירים בעזרת רשת עצבית מלאכותית וייצוג הגנום כתמונה מגלה תוצאות דומות לאלה שמוסגות במערכת הקיימת, הוחלט שכדי להשיג שיפור רצוי להגדיל את אוכלוסיית הייחוס, שמשמשת להם כסדרת אימון. הואיל שנתוני ספר העדר הנוכחיים כוללים אלפי שבבים גנטיים של פרות, הוחלט לחקור בתכנית המשך (הנהלת ענף בקר, 362-0774), שילוב של נתונים גנומיים קיימים עבור פרות באוכלוסיית הייחוס.

אשר לבעיה הנקודתית, שתוארה בעבודה זאת, של שרידות נמוכה של עגלות בתחלובה הראשונה בגלל דלקות עטין, הוצע שילוב באינדקס הטיפוח של תכונה אלטרנטיבית להערכה גנטית של דלקת עטין על בסיס תפוקת החיים. הצעה זאת סוכמה בפרסום בכתב העת GENES (Leitner וחבריו, 2024). שרידות נמוכה של בנות שלא צלחו את התחלובה הראשונה היא תופעה שהנזק שלה לחקלאי גדול, שערכי הטיפוח במערכת ההשבחה הנוכחית אינם משקפים. לכן, שינוי אינדקס הטיפוח כמוצע או לחילופין העלאת המשקל של תכונת השרידות באינדקס החמ"מ הם פתרונות שיש לשקול בהמשך תכנית הטיפוח.

6. רשימת ספרות רלבנטית (* כוכביות מסמנות מובאות בהן שותפים הכותבים):

* שגיא ד, רק ר, סרוסי א. (2022) יישום מערכות לומדות לניבוי ערכי חמ"מ באמצעות סמנים גנטיים באוכלוסיית בקר לחלב בישראל. הכנס השנתי ה-33 למדעי הבקר והצאן. מלון רמדה, ירושלים.

<https://www.kenesbakar.co.il/Portals/166/for%20web-final-2022.pdf>

*Baruch E, Weller JI, Cohen-Zinder M, Ron M, **Seroussi E.** (2006) Efficient inference of haplotypes from genotypes on a large animal pedigree. *Genetics* 172:1757-1765.

De los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J. (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* 92:295-308.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193:327-345.

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics, 4th edition. Longman, Harlow, England.

*Gershoni M, Shirak A, Raz R, **Seroussi E.** (2022). Comparing BeadChip and WGS Genotyping: Non-Technical Failed Calling Is Attributable to Additional Variation within the Probe Target Sequence. *Genes* 13:485.

Gianola D, van Kaam JB. (2008) Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178:2289-2303.

Hornik K, Stinchcombe M, White H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks* 2:359-366.

He K, Zhang X, Ren S, Sun J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Howard J, Gugger S. (2020). Fastai: A layered API for deep learning. *Information*, 11, 108.

Kolmogorov AN. (1957) On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition. *Doklady Akademii. Nauk USSR* 114:679-681.

*Leitner, G.; Blum, S.E.; Krifucks, O.; Lavon, Y.; Jacoby, S.; **Seroussi, E.** (2024) Alternative Traits for Genetic Evaluation of Mastitis Based on Lifetime Merit. *Genes* 15:92. <https://doi.org/10.3390/genes15010092>

Meuwissen T, Hayes B, Goddard M. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL,

- Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753.
- Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB. (2007). Detection of gene × gene interactions in genome-wide association studies of human population data. *Human Heredity* 63:67-84.
- Nayeri S, Sargolzaei M, Tulpan D. (2019) A review of traditional and machine learning methods applied to animal breeding. *Anim. Health Res. Rev.* 20:31-46.
- Norman HD, Paul VanRaden P, Wiggans G. (2020) April 2020: Genetic Base Change. Council on Dairy Cattle Breeding, News, https://www.uscdcb.com/wp-content/uploads/2020/02/Norman-et-al-Genetic-Base-Change-April-2020-FINAL_new.pdf.
- Pook T, Freudenthal J, Korte A, Simianer H. (2020). Using local convolutional neural networks for genomic prediction. *Front. Genet.* 11:561497.
- Schaeffer LR. (2006) Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.* 123: 218–223.
- ***Seroussi E.** (2009) The concordance test emerges as a powerful tool for identifying quantitative trait nucleotides: lessons from BTA6 milk yield QTL. *Anim. Genet.* 40:230-234.
- *Seroussi U, **Seroussi E.** (2013) Applying machine learning to the prediction of breeding values using single nucleotide polymorphisms (SNPs) in a small-sized dairy cattle population. <http://ibs13.cs.bgu.ac.il/sites/default/files/IBS13-posters.pdf>. The 15th Israeli Bioinformatics Symposium, Beer-Shiva, Israel.
- ***Seroussi E,** Shirak A, Gershoni M, Ezra E, de Abreu Santos DJ, Ma L, Liu GE. (2019) *Bos taurus*-indicus hybridization correlates with intralocus sexual-conflict effects of PRDM9 on male and female fertility in Holstein cattle. *BMC Genet.* 20:71.
- Tan M, Le Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.
- VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS. (2009) Invited Review: Reliability of genomic predictions for north American Holstein bulls. *J. Dairy Sci.* 92:16-24.
- *Weller JI, Glick G, Shirak A, Ezra E, **Seroussi E,** Shemesh M, Zeron Y, Ron M. (2014) Predictive ability of selected subsets of single nucleotide polymorphisms (SNPs) in a moderately sized dairy cattle population. *Animal* 8:208-216.
- Yang J., Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42:565-569.