Iranian Journal of Animal Science
Vol 49, No 1, Spring 2018 (145-157)
DOI: 10.22059/ijas.2018.253649.653623

علوم دامی ایران
دورهٔ ۴۹، شمارهٔ ۱، بهار ۱۳۹۷ (ص ۱۵۷-۱۴۵)

تأثیر مستندسازی نژادگان و معماری‌های مختلف ژنگانی بر عملکرد روش‌های جنگل تصادفی و بیز آستانه‌ای A در پیش‌بینی ژنگانی

یوسف نادری*
استادیار، دانشگاه آزاد اسلامی، واحد آستارا
(تاریخ دریافت: ۱۳۹٦/۱۲/۱۵ - تاریخ پذیرش: ۱۳۹۷/۲/٤)

چکیده

انتخاب ژنگانی (ژنومی) با بهره‌گیری از مستندسازی (ایمپیوتیشن) می‌تواند نقش مهمی در افزایش بهره‌وری اقتصادی و پیشرفت ژنتیکی صفات آستانه‌ای ایفا کند. هدف این تحقیق بررسی درستی مستندسازی و تأثیر آن در سطح زیر منحنی مشخصهٔ عملکرد (AUROC) ارزیابی ژنگانی روش‌های بیز آستانه‌ای A (TBA) و جنگل تصادفی (RF) در ویژگی‌های آستانه‌ای با معماری‌های مختلف ژنگانی است. داده‌های ژنگانی برای سطوح متفاوت وراثت‌پذیری (۰/۱ و ۰/۳)، سطوح مختلف LD (۰/۱۳۵ و ۰/۲۹۵) و شمار متفاوت جایگاه‌های ویژگی‌های کمی (۱۰۸ و ۱۰۸۰) روی کروموزم ۲۷ کروموزم همانندسازی شدند. برای همانندسازی شرایط واقعی برای هر پیش‌فرض (سناریو)، از بین ٥٤ هزار نشانگر همانندسازی‌شده به‌طور تصادفی اقدام به حذف ۵۰ درصد و ۹۰ درصد نشانگرها کرده و در مرحلهٔ بعد با مستندسازی اقدام به پیش‌بینی نژادگان (ژنوتیپ) نشانگرها کرده و درستی مستندسازی ارزیابی شد. در گام آخر، نژادگان‌های اصلی و مستندشده با استفاده از روش TBA و RF برای ارزیابی AUROC استفاده شدند. با افزایش سطح LD و کاهش میزان حذف نشانگرها، درستی مستندسازی بهبود یافت. میانگین AUROC پیش‌فرض‌های همانندسازی‌شده برای جنگل تصادفی و TBA به‌ترتیب ۰/٦٤ و ۰/٦٦ بود. استفاده از نژادگان‌های مستندشده با میزان حذف ۵۰ درصد و ۹۰ درصد، به‌ترتیب AUROC را به میزان ۰/۰۱۳ و ۰/۰۲ برای RF و ۰/۰۱۸ و ۰/۰۲٦ برای TBA کاهش داد. به‌رغم AUROC بالای روش بیز آستانه‌ای A در پیش‌فرض‌های مختلف، روش جنگل تصادفی عملکرد بهتری در شمار بالای QTL نشان داد. به‌طورکلی استفاده از نژادگان‌های مستندشده (5k) می‌تواند راهکار مهمی برای کاهش هزینه‌های ارزیابی ژنگانی باشد.

واژه‌های کلیدی: درستی مستندسازی، نبود تعادل پیوستگی، میزان نژادگان ازدست‌رفته، همانندسازی، AUROC.

Impact of genotype imputation and different genomic architectures on the performance of random forest and threshold Bayes A methods for genomic prediction

Yousef Naderi*
Assistant Professor of Genetics and Animal Breeding, Islamic Azad University, Astara Branch, Astara, Iran
(Received: Mar. 6, 2018 - Accepted: Apr. 24, 2018)

ABSTRACT

Genomic selection using imputed genotypes can have an important role in increasing economic efficiency and the genetic improvement of the threshold traits. The objective of this study was to: investigate the accuracy of imputation and to evaluate its effect on area under receiver operating characteristic (AUROC) of threshold BayesA (TBA) and random forest (RF) algorithms for discrete traits with different genomic architectures. Genomic data were simulated to reflect variations in heritability (0.30 and 0.10), number of QTL (108 and 1080) and linkage disequilibrium (low and high) for 27 chromosomes. To simulate a condition close to reality, we randomly masked markers with 50% and 90% missing rate for each scenario; afterwards, missing genotypes were imputed and imputation accuracy was estimated. In the last step, to evaluate the AUROC of TBA and RF, original or imputed genotypes were used. The accuracy of imputation was improved with increasing level of LD and decreased missing rate. The total average of AUROC values were 0.64 and 0.66 when using RF and TBA, respectively. Comparing to original genotypes, using imputed genotypes with 50% and 90% missing rate decreased the average AUROC about 0.013 and 0.02 for RF and 0.0018 and 0.026 for TBA, respectively. Despite the higher AUROC of TBA at different scenarios, RF showed a better performance in large number QTL. Generally, genomic prediction based on imputed genotypes (5K) can be implemented to reduce of the cost of a genomic evaluation.

Keywords: AUROC, imputation accuracy, linkage disequilibrium, missing genotype rate, simulation.

* Corresponding author E-mail: yousefnaderi@gmail.com

مقدمه

با وجود موفقیت‌های چشمگیر استفاده از گزینش پدیدگانی (فنوتیپی) و کلاسیک در بهبود ژنتیکی حیوان‌ها (Dekkers, 2002; Hayes, 2007) از سال ۱۹۷۰ تاکنون، این جعبهٔ سیاه با کشف ژن‌های مؤثر بر ویژگی‌های کمی در حال باز شدن است (Montaldo, 2006). مفاهیم اولیهٔ انتخاب ژنگانی (ژنومی) نخستین بار توسط Nejati-Javaremi *et al.* (1997) مطرح شد و در سال ۲۰۰۱ روش‌ها و اصول آن (Meuwissen *et al.*, 2001) ارائه شد، که نقش بسیار مهمی در افزایش درستی ارزیابی ژنتیکی حیوان‌های اهلی داشته است. به‌رغم گسترش روش‌های تعیین نژادگان (ژنوتیپ) و در دسترس بودن داده (پنل)های با تراکم نشانگری بالا، هنوز هم کاربرد این داده‌های نشانگری با تراکم بالا در سطح گسترده مقرون به‌صرفه نیست. از سوی دیگر هنگام تعیین نژادگان دام‌ها به‌طورمعمول داده‌های نژادگانی برخی نشانگرها از دست می‌رود لذا پیش از ارزیابی ژنگانی این داده‌های ازدست‌رفته باید به‌گونه‌ای بازیابی شود چراکه برخی از این نشانگرها ممکن است بزرگ اثر بوده و نبود داده‌های مربوط به آن‌ها منجر به کاهش میزان درستی ارزیابی ژنگانی شود. در این راستا محققان سعی بر آن دارند تا با استفاده از راهکارهایی منطقی هزینه‌های گزاف ارزیابی ژنگانی در کوتاه‌مدت را کاهش دهند. از این‌رو مستندسازی (ایمپیوتیشن) نژادگان (Genotype Imputation) راه‌حل ضروری و پیش‌نیازی برای ارزیابی ژنگانی به شمار می‌آید (Calus *et al.*, 2013). مستندسازی ژنگانی افزون بر برآورد نشانگرهای گم‌شده در یک تراشه (چیپ) می‌تواند با ادغام تراشه‌های پرشمار و همچنین گسترش تراشه‌های با تراکم پایین به تراکم بالا ارزیابی ژنگانی را ارتقاء بخشد. نتایج بررسی‌ها درزمینهٔ مستندسازی نشان داد، داده‌های ژنگانی با درستی مستندسازی بالا می‌توانند سطوح همسان و معقولی از درستی پیش‌بینی ژنگانی در قیاس با داده‌های اصلی ایجاد کرده‌اند (Felipe *et al.*, 2014). همچنین مستندسازی داده‌های کم تراکم به تراکم بالا با درستی بالای مستندسازی افزون بر مقرون به‌صرفه بودن، امکان انتخاب شمار زیادی افراد جوان را برای توالی یابی ژنگان افزایش داده و در جهت پیشبرد انتخاب ژنگانی یک گام روبه‌جلو به‌شمار می‌آید (Chen *et al.*, 2014).

عامل‌های مختلفی می‌توانند درستی ارزش‌های اصلاحی ژنگانی و ارزیابی ژنگانی را تحت تأثیر قرار دهد، این عامل‌ها شامل توزیع اثرهای QTL، مقدار نداشتن تعادل پیوستگی، نوع و تراکم نشانگر (مارکر)ها، وراثت‌پذیری، نحوهٔ رکوردگیری، شمار داده‌های پدیدگانی در جمعیت مرجع، فاصلهٔ زمانی (شمار نسل) بین جمعیت مرجع و جمعیت تأیید، نوع صفت مورد بررسی (پیوسته و گسسته) و مدل آماری مورد استفاده به‌منظور برآورد اثر نشانگرها هستند (Muir, 2007; Villumsen *et al.*, 2009).

استفادهٔ بهتر از انتخاب ژنگانی و کاربرد آن در اصلاح نژاد دام هم گام با توسعهٔ روش‌های مولکولی، افزون بر مقرون به‌صرفه بودن، به فهم بالای روش‌های آماری مورد استفاده در انتخاب ژنگانی وابسته است. روش‌های آماری مختلفی برای برآورد اثر نشانگرها ارائه شده است. تفاوت عمدهٔ این روش‌ها افزون بر معماری ژنتیکی صفت مورد بررسی، به فرضیه‌های در نظر گرفته شده برای مدل ژنتیکی پشت‌صحنهٔ آن‌ها وابسته است (Chen *et al.*, 2014). روش‌های بیزی و یادگیری ماشین از جمله روش‌های آماری برای غلبه بر مشکل $p<n$ در انتخاب ژنگانی هستند. استفاده از روش بیز A در انتخاب ژنگانی نخستین بار توسط Meuwissen *et al.* (2001) ارائه و بعدها توسط González-Recio & Forni (2011) مدل آستانه‌ای آن گسترش یافت. از دیگر روش‌های مورد استفاده در انتخاب ژنگانی، روش جنگل تصادفی است که یکی از الگوریتم‌های یادگیری ماشینی است که نخستین بار توسط Breiman (2001) پیشنهاد شد و بعدها از آن برای تجزیهٔ ژنگانی ویژگی‌های آستانه‌ای (González-Recio & Forni, 2011) و بررسی‌های پویش ژنگانی (Nguyen *et al.*, 2015) استفاده شد. یکی از روش‌های مناسب برای ارزیابی نتایج به‌دست‌آمده از مدل‌های آماری و ارزیابی میزان قابلیت آن‌ها در شناسایی آستانهٔ موردنظر، استفاده از سطح زیر منحنی

مشخصهٔ عملکرد (AUROC) است که در مقایسه با ضریب همبستگی معیار مناسب‌تری برای ارزیابی درستی و کارایی الگوریتم‌های دسته‌بندی است. این گستره بیانگر احتمال این است که یک آستانهٔ انتخاب‌شده به‌طور تصادفی، درست طبقه‌بندی می‌شود و هرچه بیشتر باشد، قابلیت اطمینان بهتر روش یادشده را نشان می‌دهد (Swets, 1988; Hand, 2009).

امروزه تلاش پژوهشگران بر آن است، تا با شناسایی عامل‌های مؤثر بر درستی برآوردهای ارزش‌های اصلاحی ژنگانی، بهترین بهره‌برداری از داده‌های موجود انجام شود. از آنجا که توالی‌یابی کل ژنگان روی شمار بسیار زیادی از افراد، سخت و هزینه‌بر است، انجام بررسی‌ها با استفاده از همانندسازی روشی منطقی برای چنین امری خواهد بود. لذا این پژوهش با هدف بررسی AUROC در برآورد ارزش‌های اصلاحی ژنگانی ویژگی‌های آستانه‌ای دودویی با استفاده از مستندسازی نژادگان و معماری‌های مختلف ژنگانی شامل شمار متفاوت QTL، سطوح متفاوت وراثت‌پذیری و LD با استفاده از روش جنگل تصادفی و بیز آستانه‌ای A در دو مجموعهٔ داده‌های همانندسازی‌شده اصلی و مستندشده انجام شده است.

مواد و روش‌ها

همانندسازی ژنگان

جمعیت‌ها با استفاده از نرم‌افزار QMSim (Sargolzaei & Schenkel, 2009) همانندسازی شدند. در مرحلهٔ اول، برای تولید جمعیتی با LD پایین و با در نظر گرفتن اندازهٔ مؤثر جمعیت در گوسفند، یک جمعیت پایه با ۴،۸۰۰ ماده و ۲۰۰ نر طی ۱۰۰۰ نسل همانندسازی شد. برای تولید جمعیتی با LD بالا، پس از همانندسازی جمعیت با LD پایین، شمار افراد جمعیت با ایجاد یک گلوگاه ژنتیکی (Bottleneck) به ۱۰۰ رأس در نسل ۱۰۵۰ کاهش یافت. آنگاه در آخرین جمعیت پایه، پس از ۵۰ نسل (در نسل ۱۱۰۰) شمار افراد جمعیت به فاز اول خود یعنی ۴۸۰۰ ماده و ۲۰۰ نر برگشت داده شدند. در گام دوم، برای ایجاد جمعیت مرجع و تأیید، همهٔ افراد (۵۰۰۰ رأس) آخرین نسل جمعیت پایه برای تولیدمثل در این جمعیت استفاده شدند که در این بین ۲۰۰ رأس نر و اندازهٔ مؤثر جمعیت ۷۶۸ در نظر گرفته شد. نوع نظام تلاقی تصادفی بود و برای ده نسل دیگر جمعیت افزونش شد. شانس تلاقی در همهٔ حیوانات برابر (در هر دو جنس) و یک فرزند برای هر زایش در نظر گرفته شد. درصد جایگزینی برای نر و ماده به ترتیب ۸۰ و ۲۰ درصد در نظر گرفته شد. انتخاب حیوان‌های برتر برای نسل بعد بر پایهٔ ارزش اصلاحی صورت گرفت. نشانگرها به‌صورت دو آللی و به‌صورت فاصله‌های یکسان در بین ۲۷ جفت کروموزوم به طول ۱۰۰ سانتی مورگان توزیع شدند. به ازای هر کروموزوم ۱۰۰۰ نشانگر همانندسازی شد. در نتیجه ۵۴۰۰۰ نشانگر برای داده‌های ۵۴K همانندسازی شد. دو سطح مختلف QTL (۱۰۸ و ۱۰۸۰) همانندسازی شد که به‌صورت تصادفی در طول کروموزوم‌ها توزیع شدند. میزان جهش برای نشانگرها و QTLها در هر جایگاه و در هر نسل $۲/۵\times۱۰^{-۵}$ فرض شد (Solberg *et al.*, 2008). فراوانی آللی اولیه برای نشانگرها ۰/۵ و توزیع اثر QTLها، گاما فرض شد. در هر نسل و هر جایگاه کل میزان واریانس افزایشی توسط QTL توجیه شد. دو سطح مختلف وراثت‌پذیری (۰/۱ و ۰/۳) برای هر صفت در نظر گرفته شد (جدول ۱). درمجموع ۴ پیش‌فرض یا سناریو (پیش‌فرض اول: وراثت‌پذیری صفت ۰/۳ - شمار QTL ۱۰۸۰- سطح پایین LD؛ پیش‌فرض دوم: وراثت‌پذیری صفت ۰/۳ - شمار QTL ۱۰۸- سطح پایین LD؛ پیش‌فرض سوم: وراثت‌پذیری صفت ۰/۱ - شمار QTL ۱۰۸- سطح پایین LD؛ پیش‌فرض چهارم: وراثت‌پذیری صفت ۰/۱- شمار QTL ۱۰۸- سطح بالای LD) در این تحقیق همانندسازی شد. برای همانندسازی پدیدگان آستانه‌ای دودویی، کد صفر برای حیوان‌های پایین‌تر از میانگین صفت و کد یک برای حیوان‌ها با پدیدگان بالاتر از میانگین صفت در نظر گرفته شد. در طراحی جمعیت، افراد آخرین نسل به‌عنوان گروه تأیید و افراد ۴ نسل پیش از آن در گروه مرجع طبقه‌بندی شدند. نشانگرهای با فراوانی آللی ماینور (MAF) کمتر از ۰/۰۳ و همچنین تعادل عدم هاردی-واینبرگ $p<۱۰^{-۶}$ با استفاده از آزمون کای مربع حذف شدند. مرحله‌های

مختلف فرایند طراحی‌شده در این تحقیق را به‌صورت راهبرد (استراتژی) کلی نشان داده شده است (شکل ۱).

مستندسازی نژادگان

پس از همانندسازی جمعیت با تراکم ۵۴K، با کمک برنامه‌نویسی در نرم‌افزار R، به‌طور تصادفی اقدام به حذف ۹۰ و ۵۰ درصد نشانگرها کرده و در مرحلهٔ بعد با برنامهٔ Flmpute (Sargolzaei *et al*., 2011) اقدام به مستندسازی و پیش‌بینی نژادگان‌های گم‌شده از راه رابطه‌های خویشاوندی و الگوریتم‌های بر پایهٔ جمعیت شد و درنهایت درستی مستندسازی با محاسبهٔ ضریب همبستگی بین نژادگان‌های اصلی و مستندشده برای نشانگرها، ارزیابی خواهد شد.

عدم تعادل پیوستگی

سطح LD برای پیش‌فرض‌های مختلف همانندسازی‌شده با استفاده از محاسبهٔ توان دوم ضریب همبستگی (r^2) بین همهٔ جفت نشانگرهای ارزیابی شد. نرم‌افزار PLINK (Purcell *et al*., 2007) برای برآورد LD بین جفت نشانگرهای مختلف در ژنگان همهٔ حیوان‌های موجود در آخرین نسل استفاده شد.

روش‌های آماری

در این تحقیق از دو روش جنگل تصادفی (یکی از زیرمجموعه‌های روش یادگیری ماشین) و بیز آستانه‌ای A برای ارزیابی AUROC استفاده شد. جنگل تصادفی یک روش غیر فراسنجه‌ای و از نوع روش‌های باز نمونه‌گیری است که توانایی لحاظ کردن ساختارهای متقابل پیچیدهٔ داده‌ها را دارد. الگوریتم‌های جنگل تصادفی برخلاف روش‌های آماری بیزی نیازمند فرضیات کمتری درباره توزیع داده‌ها داشته، و انعطاف‌پذیری بالایی در تجزیهٔ داده‌ها پیشنهاد دارد (Goldstein *et al*., 2010). از برتری‌های دیگر این روش می‌توان به توانایی بالای آن در تجزیهٔ حجم زیاد داده‌های ژنگانی، مشکلات مربوط به نداشتن نمونهٔ کافی، وجود واریانس زیاد در نمونه‌های مورد آزمایش و توانایی بالای آن در پیش‌بینی ژنگانی با در نظر گرفتن اثر غیر افزایشی عنوان کرد. این الگوریتم درختان تصمیم زیادی را در مجموعهٔ نمونه‌های بوت استرپ‌شده (Bootstrap) ایجاد می‌کند که میانگین هر برآورد پیش‌بینی‌های نهایی را تشکیل می‌دهد. این راهبرد که بگینگ (Bagging) نامیده می‌شود (Breiman, 2001) خطای پیش‌بینی را با عامل از شمار درختان، کاهش می‌دهد. بگینگ نمونه‌گیری بوت استرپ را برای کاهش واریانس استفاده می‌کند و می‌تواند قابلیت اعتماد را افزایش و میانگین مربع خطا را کاهش دهد. جنگل تصادفی همچنین از انتخاب ویژگی (Feature selection) که یک روش موفق برای ترکیب یادگیرنده‌های ناپایدار و انتخاب متغیر تصادفی برای ایجاد درخت است استفاده می‌کند. در جنگل تصادفی هر درخت به‌طور کامل رشد بدون هرس می‌کند تا جایی که درختان با اریب پایین به دست می‌آید، در همین زمان، بگینگ و انتخاب متغیر تصادفی منجر به ایجاد درختان با همبستگی پایین می‌شوند و الگوریتم ترکیبی با واریانس و اریب پایین تولید می‌کند. فراسنجه‌های کلیدی برای مدل جنگل تصادفی، شمار درختان و شمار متغیرهای پیشگو هستند (Breiman, 2001). سه فراسنجهٔ مهمی که در جنگل تصادفی در مورد خوشه‌بندی بایستی تنظیم شود عبارت‌اند از mtry، شمار SNP یا کواریت‌هایی نمونه‌برداری شده در هر بار نمونه‌گیری تصادفی، ntree یا شمار بوت استرپ و یا شمار درختانی که بایستی رشد کنند و معیاری برای انتخاب بهترین SNP برای تقسیم شدن هر گره است. node (شمار گره) یا وزن دهی است که نشان‌دهندهٔ شمار مشاهده‌ها در هر خوشهٔ درخت است.

مدل کلی جنگل تصادفی به‌صورت زیر است.

$$\gamma = \mu + \sum_{t=1}^{T} c_t h_t(\mathbf{y}; \boldsymbol{X})$$

در اینجا برای هر مشاهدهٔ γ از طریق میانگین پیش‌بینی‌های هر درخت محاسبه می‌شود. μ میانگین جمعیت و $\mathbf{y}$ بردار $n\times1$ برای مشاهده‌های پدیدگانی گسسته، $\boldsymbol{X}=\{x_i\}$ که در اینجا x_i بردار $p\times1$ نژادگان هر حیوان برای p نشانگر که T درخت تصمیم ساخته شده است. هر درخت ($h_t(\mathbf{y}; \boldsymbol{X})$) یک نمونهٔ تصادفی با جایگزینی از n نمونه است و هر گره یک نمونهٔ

کوچکی تصادفی از SNP و c_t عامل انقباضی میانگین درختان است. برای اجرای جنگل تصادفی نمونهٔ تصادفیp نشانگرها (به شمار جذر نشانگرها) برای ساخت هر درخت استفاده شد. دیگر حیوان‌هایی که جز این نمونه‌گیری نیستند به‌عنوان خارج از مجموعهٔ شناخته‌شده و در اعتبارسنجی هر درخت گزینش می‌شوند. در هر گره داده‌ها در ۲ شاخه بر پایهٔ نژادگان SNPj تقسیم و تا زمان رسیدن به همگرایی ادامه یافت. در این بررسی ۵۰۰ درخت برای تراشهٔ ۵۴K ساخته شد. داده‌های ژنگانی از طریق بستهٔ RanFoG (González-Recio & Forni, 2011) و نرم‌افزار R تجزیه شدند.

مدل کلی بیز آستانه‌ای A به‌صورت زیر است.

$$\boldsymbol{\lambda} = \mu\mathbf{1} + \mathbf{Xb} + \mathbf{e}$$

در اینجا $\boldsymbol{\lambda}$ بردار ستونی با ابعاد n×1 برای متغیر $\mathbf{y}$ است. μ میانگین جمعیت، ۱ بردار ستونی با ابعاد n×1 با درایه‌های ۱ است. b برداری برای برآوردهای ضریب‌های رگرسیونی اثر p نشانگر با فرض توزیع نرمال ($N(0, \sigma_j^2)$) و واریانس متفاوت و مستقل برای هر نشانگر است. σ_j^2 به‌صورت واریانس ناشناخته در ارتباط با نشانگرها مفروض شده است و دارای توزیع کای مربع مقیاس‌دار معکوس ($\sigma_j^2 \sim \upsilon_j s^2_j \chi^{-1}_{\upsilon j}$) با $\upsilon_j=4$ و $s^2_j=0.002$ بود. عنصرهای ماتریس $\mathbf{X}$ به ابعاد $n\times p$ دربرگیرندهٔ اثر افزایشی است. در بیز آستانه‌ای A باقی‌مانده‌ها ($\mathbf{e}$) با فرض میانگین ۰ و واریانس ۱ در نظر گرفته شدند. این روش از طریق نمونه‌گیری گیبس انجام گرفت. بیز آستانه‌ای A از طریق بستهٔ BGLR در نرم‌افزار R تجزیه شد.

سطح زیر منحنی مشخصهٔ عملکرد (AUROC)

برای بررسی درستی پیش‌بینی ژنگانی، نتایج همانندسازی ناشی از نژادگان‌های اصلی و مستندشده با محاسبهٔ آمارهٔ AUROC ارزیابی شدند. برای محاسبهٔ این آماره از بستهٔ (پکیج) pROCR در محیط R استفاده شد.

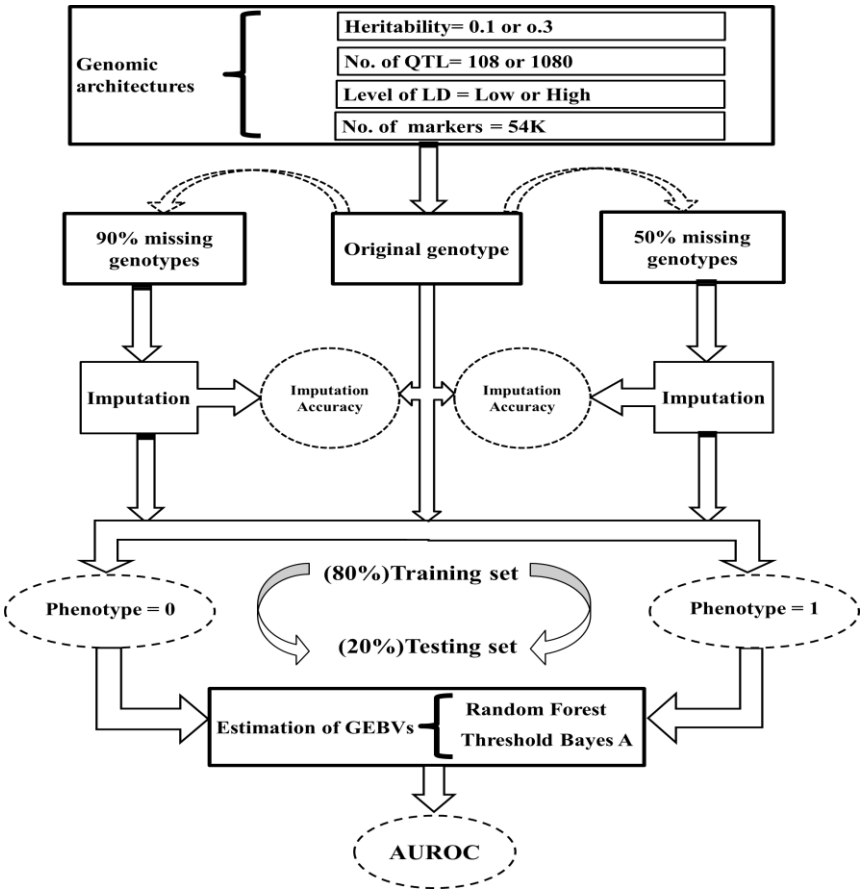
$$\text{AUROC} = \frac{\text{True Positive} - \text{False Positive}}{1 - \text{False Positive}} + \left(1 - \frac{\text{True Positive} - \text{False Positive}}{1 - \text{False Positive}}\right) \times \text{False Positive}$$

ارزیابی‌ها با استفاده از ده تکرار همانندسازی برای هر پیش‌فرض انجام گرفت و میانگین و انحراف معیار برای AUROC هر یک از پیش‌فرض‌ها گزارش شد.

جدول ۱. فراسنجه‌های فرایند همانندسازی

Table 1. Parameters of the simulation process

Parameter	Low linkage disequilibrium	High linkage disequilibrium
Historical population		
No. of generations (population size) in phase 1	1,000 (5,000)	1,000 (5,000)
No. of generations (population size) in phase 2	-	1,050 (100)
No. of generation (population size) in phase 3	-	1,100 (5,000)
Recent population		
No. of founder sires (dams)	200 (4,800)	
No. of generations	10	
No. of offspring per dam	1	
Mating system	Random	
Replacement ratio for males (females)	0.8 (0.2)	
Criteria for selection/culling	EBV/age	
Sex probability for offspring	0.5	
No. of chromosomes	27	
Total length of chromosomes (cM)	2,700	
Marker distribution	Evenly spaced	
No. of QTL alleles	Random (2, 3, or 4)	
Effects of QTL alleles	Gamma (0.4)	
Marker and QTL mutation rate	$2.5\ 10^{-5}$	
Position of marker and QTL	Random	
No. of QTL	108 or 1080	
No. of markers	54,000	
Heritability of the trait	0.1 or 0.3	



شکل ۱. راهبرد کلی فرایند تحقیق

Figure 1. The overall strategy of the research process

نتایج و بحث

درستی مستندسازی

جدول ۲ میانگین عدم تعادل پیوستگی و درستی مستندسازی برای هریک از پیش‌فرض‌های همانندسازی‌شده از طریق همبستگی نژادگان‌های اصلی و مستندشده (با میزان حذف ۵۰ و ۹۰ درصد نشانگرها) را نشان می‌دهد. به‌طورکلی با افزایش میزان حذف نشانگرها از ۵۰ به ۹۰، درستی ایپیوتیشن کاهش یافت، اما تفاوت معنی‌داری مشاهده نشد. در هر دو سری دادهٔ مستندشده (۵۰ و ۹۰ درصد) با افزایش سطح LD تفاوت معنی‌داری در درستی مستندسازی مشاهده شد ($P<۰/۰۵$). این میزان افزایش برای داده‌های با حذف ۹۰ درصد نشانگرها مشهودتر بود. به‌طوری‌که با افزایش میزان LD (پیش‌فرض ۳ نسبت به ۴)، برای نژادگان‌های با میزان حذف ۵۰ و ۹۰ درصد، درستی مستندسازی به ترتیب ۱/۶ و ۲/۳ درصد افزایش یافت. اثر عدم تعادل پیوستگی و میزان حذف نشانگری بر درستی مستندسازی به‌صورت نمودار جعبه‌ای در شکل ۲ ترسیم شده است. میانگین درستی مستندسازی برای داده‌های ۵۰ و ۹۰ درصد مستندشده به ترتیب ۰/۹۷۷ و ۰/۹۶۸ بود اما تفاوت معنی‌داری مشاهده نشد.

نوع معماری ژنگانی و اندازهٔ مؤثر جمعیت از عامل‌های مؤثر بر درستی مستندسازی در گوسفندان نژاد رامنی بود (Ventura *et al.*,2016). نتایج بررسی‌ها نشان دادند، درستی پایین مستندسازی در برخی از مناطق ژنگان (کمتر از ۰/۶) به ماهیت ژنگان و سطح بسیار پایین LD در این مناطق بستگی داشت (VanRaden *et al.*, 2013). با وجود درستی بالای مستندسازی تراشه‌های ۷k به تراشه ۵۰K، بااین‌حال استفاده از تراشه با تراکم ۳K منجر به نتایج شایان پذیرشی از درستی ژنگانی شد (Toghiani *et al.*, 2016). در تحقیقات صورت گرفته در مورد ذرت نتایج نشان داد، میزان حذف نشانگرها و سطح LD از عامل‌های مؤثر بر درستی مستندسازی است. همچنین

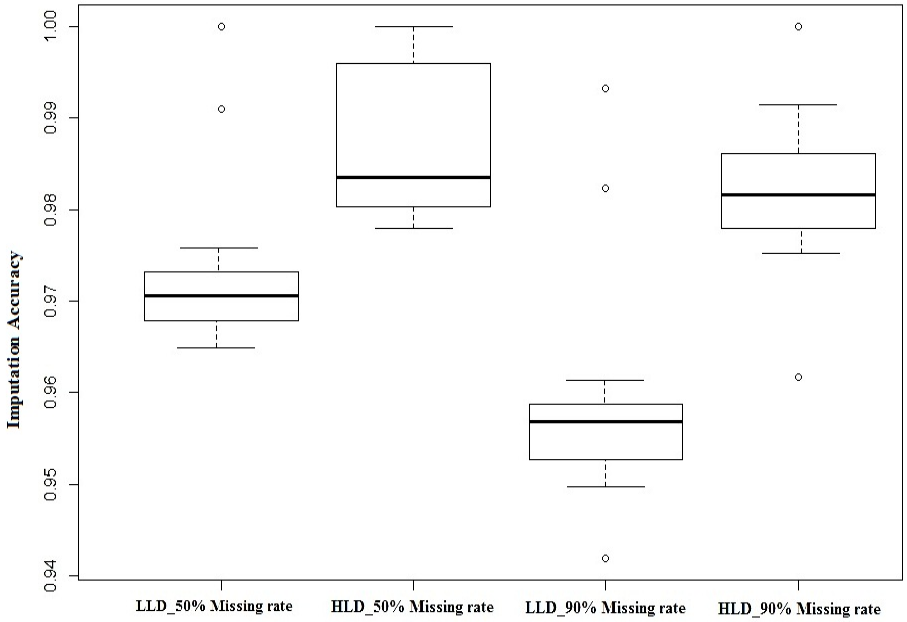
درستی مستندسازی برای داده‌های ژنگانی با میزان حذف درصد۵۰ نشانگرها حدود ۰/۹۰ بود (Hickey *et al*., 2012). نتایج بررسی‌ها در مورد نژادهای مختلف گاو شیری بر نقش مستقیم و مثبت LD بر درستی مستندسازی دلالت دارد (Pauschat al., 2017; Ogawa at al., 2016; Boison at al., 2014; Mulder at al., 2012; Khatkar *et al*., 2012). نتایج تحقیقات در مورد نژاد گاوهای فرانسه نشان داد، LD اثر جزئی و غیر معنی‌داری بر درستی مستندسازی دارد و مستندسازی تراشه‌های ۷K به تراشه‌های HD (با میزان حذف ۹۹/۱ درصد نشانگرها) درستی مستندسازی بالای ۰/۹۲۵ را در پی داشت (Carvalheiro *et al*., 2014).

مقدار AUROC روش جنگل تصادفی و بیز آستانه‌ای A برای هر یک از پیش‌فرض‌های همانندسازی‌شده (اصلی و مستندسازی) ویژگی‌های آستانه‌ای در جدول ۳ نشان داده شده است.

جدول ۲. میانگین عدم تعادل پیوستگی و درستی مستندسازی بین نژادگان‌های اصلی و مستندشده در پیش‌فرض‌های مختلف

Table 2. Mean of linkage disequilibrium and imputation accuracy between imputed and original genotypes in different scenarios

Scenarios	LD mean (in 0.1 cM)	Imputation accuracy (50% missing rate)	Imputation accuracy (90% missing rate)
I (h^2 = 0.3, QTL=1080 and LD=Low)	0.138	0.973±0.013	0.963±0.017
II (h^2 = 0.3, QTL=108 and LD=Low)	0.136	0.975±0.011	0.965±0.015
III (h^2 = 0.1, QTL=108 and LD=Low)	0.135	0.971±0.013	0.960±0.015
IV (h^2 = 0.1, QTL=108 and LD=High)	0.295	0.987±0.009	0.982±0.010



شکل ۲. نمودار جعبه‌ای، درستی مستندسازی برای سطوح مختلف عدم تعادل پیوستگی

Figure 2. The box-plots of imputation accuracy in different levels of LD

جدول ۳. میانگین و انحراف استاندارد AUROC نژادگان‌های اصلی و مستندشده با استفاده از روش جنگل تصادفی (RF) و بیز آستانه‌ای (TBA)A

Table 3. Mean and standard deviation of AUROC by random forest (RF) and threshold Bayesian A (TBA) method in the original and imputed SNP genotypes

Scenarios	RF			TBA		
	Original genotypes	Imputed genotypes (50% missing rate)	Imputed genotypes (90% missing rate)	Original data	Imputed genotypes (50% missing rate)	Imputed genotypes (90% missing rate)
I	0.740±0.01	0.719±0.02	0.708±0.03	0.682±0.04	0.663±0.05	0.658±0.05
II	0.657±0.03	0.655±0.03	0.650±0.03	0.708±0.04	0.679±0.05	0.672±0.05
III	0.579±0.02	0.569±0.02	0.561±0.02	0.588±0.05	0.580±0.05	0.574±0.06
IV	0.591±0.02	0.571±0.02	0.566±0.02	0.660±0.04	0.641±0.04	0.627±0.05
Average	0.641±0.02	0.628±0.02	0.621±0.03	0.659±0.04	0.641±0.05	0.633±0.05

اثر مستندسازی بر AUROC

جدول ۳ اثر میزان حذف نشانگرها بر AUROC را در روش جنگل تصادفی (RF) و بیز آستانه‌ای A (BTA) نشان می‌دهد. به‌طورکلی تفاوت معنی‌داری بین AUROC داده‌های اصلی و مستندشده درون هر روش آماری مشاهده نشد. بااین‌حال حساسیت AUROC در مدل TBA به تغییرات میزان حذف نشانگری نسبت به روش جنگل تصادفی بیشتر بود. به‌طوری‌که کاهشی AUROC داده‌های مستندشده با میزان ۵۰ نسبت به داده‌های اصلی برای بیز آستانه‌ای A و جنگل تصادفی به ترتیب دامنه‌ای بین ۰/۰۲۷ و ۰/۰۲ بود. در داده‌های مستندسازی، AUROC با استفاده از مدل بیز آستانه‌ای A (به‌جز پیش‌فرض I) برآورد بهتری نسبت به روش جنگل تصادفی داشت. با افزایش میزان مستندسازی، مقدار AUROC کاهش یافت و این افت AUROC برای جنگل تصادفی در درستی‌های پایین مستندسازی (پیش‌فرض I و III) محسوس‌تر بود. به‌طورکلی درستی مستندسازی بالا (۰/۹۶۷) و همچنین AUROC نزدیک به داده‌های اصلی (۵۴K) باعث مطلوبیت استفاده از مستندسازی برای داده‌های با میزان ۹۰ درصدی مستندسازی شد.

ارزیابی مدل‌های آماری با استفاده از زیرمجموعه‌های نشانگری در داده‌های با و بدون مستندسازی به این نتیجه رسیده‌اند که توانایی پیش‌بینی ژنگانی هنگام استفاده از مستندسازی نژادگان بهبود بخشید و بسیاری از محققان این راهبرد را برای کاهش هزینه‌ها در برنامه‌های انتخاب ژنگانی توصیه کرده‌اند (Mulder *et al.*, 2012; Berry & Kearney, 2011; Daetwyler *et al.*, 2011). Chen *et al.* (2014) به بررسی اثر درستی مستندسازی بر درستی پیش‌بینی ژنگانی در تراشه‌های ۳k و ۶k در مقایسه با تراشهٔ ۵۰k پرداختند و نشان دادند، درستی پیش‌بینی ژنگانی با استفاده از تراشه‌های ۶k عملکرد بهتری نسبت به تراشه‌های ۳k دارد. به‌طوری‌که در پیش‌فرض‌های مورد بررسی کاهش ناچیزی و غیر معنی‌داری در درستی ژنگانی روش بیزی برای تراشهٔ ۶k نسبت به ۵۰k مشاهده شد. نتایج در مورد گاوهای جرسی نشان داد، مستندسازی یک تراشه با تراکم پایین به تراشهٔ ۵۰k هنگامی‌که درستی مستندسازی بالا باشد منجر به بهبود درستی ژنگانی خواهد شد. در نتیجه توجه ویژه به‌درستی مستندسازی در بررسی‌های ژنگانی بسیار اهمیت خواهد داشت (Weigel *et al.*, 2010). هنگامی‌که خطای مستندسازی بالا و در پی آن درستی مستندسازی پایین باشد در نتیجه استفاده از تراشه‌های با تراکم بالا توصیه شد (Mulder *et al.*, 2012). در بررسی اخیر با توجه به درستی بالای مستندسازی و در پی آن نبود تفاوت معنی‌دار AUROC در گروه‌های مستندشده و اصلی در هر دو روش آماری، استفاده از مستندسازی امری اقتصادی و مقرون به‌صرفه بود. بررسی‌های روی جمعیت موش با میزان مستندسازی ۷۵ و ۵۰ درصد به ترتیب درستی مستندسازی ۰/۹۴ و ۰/۹۸ را به همراه داشت، و استفاده از داده‌های مستندسازی نسبت به داده‌های اصلی در مدل‌های نافراسنجه‌ای، صرفهٔ اقتصادی را به همراه داشت (Felipe *et al.*, 2014).

درستی پیش‌بینی ارزش‌های اصلاحی ژنگانی در داده‌های واقعی و مستندشده (با میزان حذف ۹۲/۸۶ درصد) با استفاده از مدل بیز A متأثر از میزان حذف نشانگری گزارش شد (Toghiani *et al.*, 2016). کاهش درستی ژنگانی نژادگان‌های مستندشده (6K) نسبت به نژادگان‌های اصلی (۵۰K) با استفاده از مدل بیز را حدود ۰/۰۶ گزارش کرد و تفاوت معنی‌داری مشاهده نشد (Chen *et al.*, 2014). Pimentel *et al.* (2015) خطای مستندسازی را از عامل‌های دخیل و مؤثر بر درستی پیش‌بینی ارزش‌های اصلاحی ژنگانی گزارش کردند.

نتایج این تحقیق در پیش‌فرض‌های با شمار کم QTL نشان داد که AUROC به‌دست‌آمده از طریق بیز آستانه‌ای A نسبت به خطای مستندسازی حساسیت بیشتری نشان می‌دهد. در تأیید نتایج اخیر، دیگر محققان (Chen *et al.*, 2014; Zhang *et al.*, 2011) نشان دادند، که درستی پیش‌بینی ژنگانی (با استفاده از روش‌های بیزی) در ویژگی‌های تحت تأثیر شمار کم QTL حساسیت بیشتری به خطای مستندسازی نشان می‌دهند. Felipe *et al.* (2014) گزارش کردند، همیشه مستندسازی منجر به بهبود پیش‌بینی ژنگانی نمی‌شود و شایستگی مستندسازی نژادگان را وابسته به درجهٔ ارتباط

بین جمعیت مورد آزمون و مرجع، شمار نشانگرها، معماری ژنگانی صفت و نوع روش آماری در پیش‌بینی تأثیر نشانگرها دانستند.

نقش معماری‌های مختلف ژنگانی بر AUROC در داده‌های مستندسازی

تأثیر شمار QTL بر AUROC

برای ارزیابی تأثیر شمار QTL بر AUROC، پیش‌فرض ۱ (۱۰۸ QTL) و ۲ (۱۰۸۰ QTL) با هم مقایسه شدند (شکل ۳). در هر سه گروه (نژادگان‌های اصلی و مستندشده با حذف ۵۰ و ۹۰درصد نشانگرها)، با افزایش شمار QTL، مقدار AUROC روش جنگل تصادفی افزایش معنی‌داری نشان داد (P<۰/۰۵). به‌طورکلی به‌کارگیری داده‌های ۵۰K همراه با شمار بالای QTL با نزدیک بودن فاصلهٔ SNPها با QTLها همراه بود در نتیجه شانس نمونه‌گیری در RF افزایش یافت که نتیجهٔ مثبت آن در جنگل تصادفی مشهود بود. افزایش شمار QTL، کاهشAUROC را در روش بیز آستانه‌ای A به همراه داشت، که دلیل این امر را می‌توان به توزیع محدود واریانس ژنتیکی بر شمار زیادی QTL دانست که در نتیجه سهم هر QTL در ارزش ژنتیکی کل کاهش‌یافته، و قدرت مدل‌ها در پیش‌بینی ارزش‌های اصلاحی کاهش خواهد یافت. به خاطر اینکه روش بیزی A از برخی ویژگی‌های مانند انتخاب متغیر سود می‌برند و مفروض‌های آن با شمار QTL کم سازگارتر است در نتیجه در شمار اندک QTL روش بیز آستانه‌ای A بهتر عمل می‌کنند. نتایج این تحقیق در مورد اثربخشی شمار بالای QTL بر AUROC در مورد جنگل تصادفی با نتایج Naderi *et al.* (2016) همخوانی داشت. Hayes *et al.* (2009)، عملکرد مثبت بیز A را در حضور شمار کم QTL گزارش کردند. در پژوهشی با استفاده از شمار QTL برابر با ۹۰ و ۱۰۰۰، مقدار AUROC روش جنگل تصادفی به ترتیب ۰/۷۰ و ۰/۶۹ و برای بیز آستانه‌ای A به‌ترتیب ۰/۶۱ و ۰/۶۶ گزارش شد (González-Recio & Forni, 2011).

تأثیر وراثت‌پذیری بر AUROC

برای ارزیابی تأثیر وراثت‌پذیری بر AUROC، پیش‌فرض ۲ (h^2=۰/۳) و ۳ (h^2=۰/۱) با هم مقایسه شدند (شکل۴). در هر سه مجموعهٔ داده‌ها، با افزایش وراثت‌پذیری افزایش معنی‌داری در AUROC مشاهده شد (P<۰/۰۵). به‌طورکلی AUROC در روش بیز آستانه‌ای A نسبت به روش جنگل تصادفی به افزایش وراثت‌پذیری حساسیت بیشتری نشان داد.

هنگامی‌که وراثت‌پذیری بالا است، مقدار AUROC در روش بیز آستانه‌ای A بیشتر متأثر از درستی مستندسازی قرار گرفت، به‌طوری‌که AUROC در داده‌های مستندشده با میزان حذف درصد ۵۰ و درصد ۹۰، نسبت به داده‌های اصلی به‌ترتیب ۰/۰۲۹ و ۰/۰۳۶ واحد کاهش یافت. بیشتر بودن وراثت‌پذیری به معنای بیشتر بودن نقش ژن‌ها با بیان افزایشی در ایجاد پراکنش در صفت است که باعث برآورد درست‌تر تأثیر نشانگرها می‌شود، و هرچه وراثت‌پذیری صفت بیشتر باشد، پدیدگان فرد به ارزش نژادگانی فرد نزدیک‌تر بوده و در نتیجه آثار نشانگرها و به دنبال آن ارزش‌های اصلاحی ژنگانی افراد به‌طور درست‌تری پیش‌بینی می‌شود (Villumsen *et al*., 2009; Goddard & Hayes, 2009). در بررسی‌های همانندسازی، با افزایش وراثت‌پذیری افزایش شایان‌توجهی در AUROC روش جنگل تصادفی مشاهده شد (Naderi *et al*., 2016).

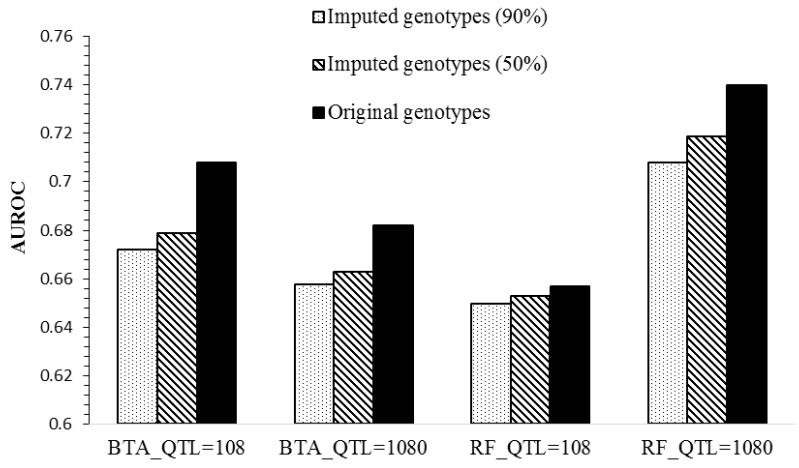
عدم تعادل پیوستگی

میانگین r^2 برای فاصله‌های مختلف تا فاصلهٔ دو مگا جفت باز (۲*Mbp*) در شکل ۵ نشان داده شده است. میانگین r^2 با افزایش فاصلهٔ بین نشانگرها کاهش یافت. به‌طورکلی r^2 محاسبه‌شده برای پیش‌فرض‌های ۴ (HLD) بزرگ‌تر از پیش‌فرض‌های ۳ (LLD) بود. این تفاوت در فواصل کمتر، مشهودتر بود و با افزایش فاصلهٔ بین نشانگرها، تفاوت r^2 محاسبه‌شده بین پیش‌فرض‌های HLD و LLD کمتر بود. روند نمایی کاهش LD با افزایش فاصلهٔ فیزیکی بین دو نشانگر با نتایج به‌دست‌آمده در دیگر بررسی‌ها همخوانی دارد (Naderi *et al*., 2016; Yin *et al*., 2014). به‌طورکلی وجود LD به ساختار ژنتیکی و اندازهٔ مؤثر جمعیت وابسته است (Wang *et al*., 2017). نتایج بررسی‌ها نشان داد، کمترین میزان LD به‌منظور یک انتخاب

ژنگانی مناسب ۰/۲ است (Meuwissen *et al*., 2001). سطح LD یک جمعیت می‌تواند اصلاحگر را در تشخیص تراکم نشانگری (ریزتراشه‌های DNA) مناسب کمک کرده و بر درستی ارزیابی‌ها و موفقیت در انتخاب ژنگانی تأثیر مستقیم بگذارد (Solberg *et al*., 2008).

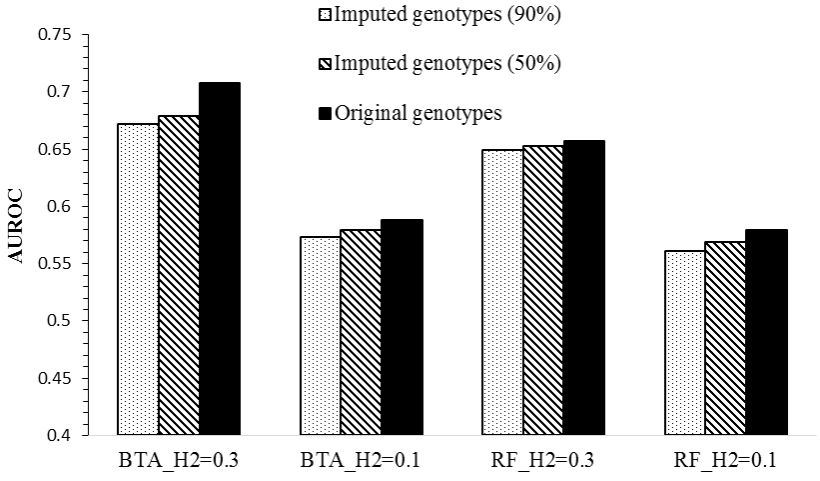
تأثیر سطح LD بر AUROC

برای ارزیابی اثر ساختار LD بر AUROC، پیش‌فرض ۳ (LD پایین) و ۴ (LD بالا) باهم مقایسه شدند (شکل ۶). میانگین LD برای پیش‌فرض ۳ و ۴ در فاصلهٔ ۰/۱ سانتی مرگان به ترتیب ۰/۱۳۵ و ۰/۲۹۵ بود. افزایش LD در مجموعهٔ داده‌ها اصلی و مستندسازی، افزایش ناچیز در AUROC روش جنگل تصادفی داشت. بااین‌حال میزان AUROC بیز آستانه‌ای A به افزایش LD حساسیت بالایی نشان داد ($p<۰/۰۵$). نتایج بررسی‌ها نشان می‌دهد، وجود LD قوی بین نشانگرها مجاور در بررسی ژنگان انسان نقش اساسی ایفا می‌کند (Ke *et al*., 2004). به‌عنوان یک اصل کلی، وجود LD بین نشانگر و QTL منبع اصلی داده‌های بوده و نقش عمده‌ای در پیش‌بینی ارزش‌های اصلاحی ژنگانی از طریق بیز A ایفا می‌کند (Sun *et al*., 2016). بنابر نتایج این تحقیق، تأثیر افزایشLD بر AUROC روش جنگل تصادفی معنی‌دار نبود و مقدار AUROC برای داده‌های با LD پایین و بالا به ترتیب دامنه‌ای بین ۰/۵۴-۰/۵۹ و ۰/۵۵-۰/۵۹ داشت (Naderi *et al*., 2016).



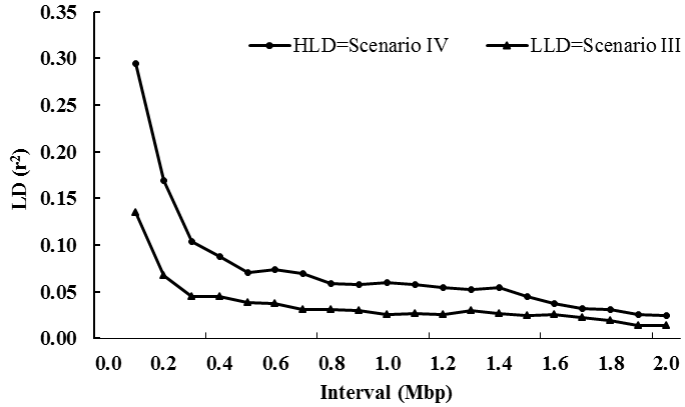
شکل ۳. تأثیر شمار متفاوت QTL بر AUROC روش جنگل تصادفی (RF) و بیز آستانه‌ای A (BTA)

Figure 3. Effect of different number of QTL on AUROC by threshold Bayes A (TBA) and random forest (RF)



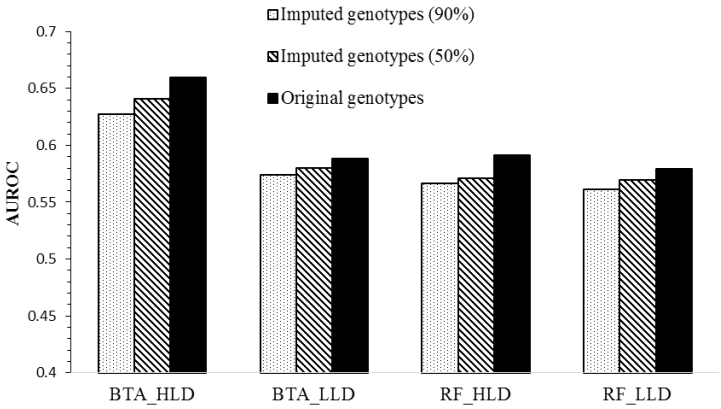
شکل ۴. تأثیر سطوح مختلف وراثت‌پذیری بر AUROC روش جنگل تصادفی (RF) و بیز آستانه‌ای A (BTA)

Figure 4. Effect of different levels of heritability on AUROC by threshold BayesA (TBA) and random forest (RF)



شکل ۵. میانگین نبود تعادل پیوستگی در فاصله‌های مختلف ژنگان

Figure 5. Mean of linkage disequilibrium in different intervals of the genome



شکل ۶. تأثیر سطح LD بر AUROC در روش جنگل تصادفی (RF) و بیز آستانه‌ای A (BTA)

Figure 6. Effect of different levels of LD on AUROC by threshold Bayes A (TBA) and random forest (RF)

نتیجه‌گیری

به‌طورکلی نسبت مستندسازی (میزان حذف نشانگرها) و ساختار LD از عامل‌های مؤثر بر درستی مستندسازی بودند. ساختار معماری ژنگانی (شمار QTL، سطح LD و وراثت‌پذیری) به همراه نسبت مستندسازی از عامل‌های مؤثر بر AUROC به‌دست‌آمده از طریق روش‌های جنگل تصادفی و بیز آستانه‌ای A در تجزیهٔ ویژگی‌های آستانه‌ای بودند. با وجود عملکرد بالای روش بیز آستانه‌ای A در برآورد AUROC، روش جنگل تصادفی حساسیت بیشتری به تغییرات QTL از خود نشان داد. بااین‌حال هنگامی‌که ویژگی‌های آستانه‌ای دارای سطوح پایین LD، وراثت‌پذیری متوسط و تحت کنترل شمار زیاد QTL قرار گیرند، روش جنگل تصادفی (به‌رغم زمان محاسبات بالا) بهترین عملکرد را نشان داد. به‌عنوان یک نتیجهٔ مهم، استفاده از داده‌های 5k (با میزان حذف ۹۰ درصدی) با میانگین درستی مستندسازی حدود ۰/۹۶۸، می‌تواند راهکار مناسبی برای کاهش هزینه‌های ارزیابی ژنگانی در نظرگرفته شود.

REFERENCES

1. Berry, D. P. & Kearney, J. F. (2011). Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. *Animal*, 5(8), 1162-1169.
2. Boison, S., Neves, H. H. d. R., O'Brien, A. P., Utsunomiya, Y. T., Carvalheiro, R., da Silva, M., Sölkner, J. & Garcia, J. F. (2014). Imputation of non-genotyped individuals using genotyped progeny in Nellore, a Bos indicus cattle breed. *Livestock Science*, 89, 166-176.
3. Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.
4. Calus, M., De Haas, Y., Pszczola, M. & Veerkamp, R. (2013). Predicted accuracy of and response to genomic selection for new traits in dairy cattle. *Animal*, 7, 183-191.

5. Carvalheiro, R., Boison, S .A., Neves, H. H., Sargolzaei, M., Schenkel, F. S., Utsunomiya, Y. T., O'Brien, A. M. P., Sölkner, J., McEwan, J. C. & Van Tassell, C. P. (2014). Accuracy of genotype imputation in Nelore cattle. *Genetics Selection Evolution*, 11, 44- 69.
6. Chen, L., Li, C., Sargolzaei, M. & Schenkel, F. (2014). Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. *PLoS One*, 9(8), 1-7.
7. Daetwyler, H. D., Wiggans, G. R, Hayes, B. J., Woolliams, J. A. & Goddard, M. E. (2011). Imputation of missing genotypes from sparse to high density using long-range phasing. *Genetics*, 189(1), 317-327.
8. Dekkers, J. C. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics*, 3, 22-33.
9. Felipe1, V. P. S., Okut, H., Gianola, D., Silva, M. A. & Rosa, G. J. M. (2014). Effect of genotype imputation on genome-enabled prediction of complex traits: an empirical study with mice data. *BMC Genetics*, 15(149), 1-10.
10. Goddard, M. E. & Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10, 381-391.
11. Goldstein, B. A., Hubbard, A. E., Cutler, A. & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genetics*, 1, 11-49.
12. González-Recio, O. & Forni, S. (2011). Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genetics Selection Evolution*, 43(7), 1-12.
13. Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77, 103-123.
14. Hayes, B. (2007). QTL mapping, MAS, and genomic selection. *A short-course. Animal Breeding & Genetics Department of Animal Science*. IowaState University, 1, 3-4.
15. Hayes, B. J., Bowman, P. J., Chamberlain, A. & Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*, 92, 433-443.
16. Hickey, J. M., Crossa, J., Babu, R. & de los Campos, G. (2012). Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Science*, 52, 654-663.
17. Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J., Whittaker, P., Collins, A., Morris ,A. P. & Bentley, D. (2004). The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Human Molecular Genetics*, 13, 577-588.
18. Khatkar, M. S., Moser, G., Hayes, B. J. & Raadsma, H. W. (2012). Strategies and utility of imputed SNP genotypesfor genomic analysis in dairy cattle. *BMC genomics*, 13(1), 526-538.
19. Meuwissen, T., Hayes, B. & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157, 1819-1829.
20. Montaldo, H. H. (2006). Genetic engineering applications in animal breeding. *Electronic Journal of Biotechnology*, 9(2), 157-170.
21. Muir, W. (2007). Comparison of genomic and traditional BLUP estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics*, 124, 342-355.
22. Mulder, H., Calus, M., Druet, T. & Schrooten, C. (2012). Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *Journal of Dairy Science*, 95, 876-889.
23. Naderi, S., Yin, T. & König, S. (2016). Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. *Journal of Dairy Science*, 99, 7261-7273.
24. Nejati-Javaremi, A., Smith, C. & Gibson, J. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *Journal of Animal Science*, 75, 1738-1745.
25. Nguyen, T.-T., Huang, J. Z., Wu, Q., Nguyen, T. T. & Li, M. J. (2015) Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics*, 16(5), 1-11.
26. Ogawa, S., Matsuda, H., Taniguchi, Y., Watanabe, T., Takasuga, A., Sugimoto, Y. & Iwaisaki, H. (2016). Accuracy of imputation of single nucleotide polymorphism marker genotypes from low density panels in Japanese Black cattle. *Animal Science Journal*, 87, 3-12.
27. Pausch, H., MacLeod, I. M., Fries, R., Emmerling, R., Bowman, P. J., Daetwyler, H. D. & Goddard, M. E. (2017). Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genetics Selection Evolution*, 49(24), 1-14.
28. Pimentel, E., Edel, C., Emmerling, R. & Götz, K.-U. (2015). How imputation errors bias genomic predictions. *Journal of dairy science*, 98, 4131-4138.

29. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. & Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81, 559-575.
30. Sargolzaei, M. & Schenkel, F. S. (2009). QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, 25(5), 680-681.
31. Sargolzaei, M., Chesnais, J. & Schenkel, F. (2011). FImpute-An efficient imputation algorithm for dairy cattle populations. *Journal of Dairy Science*, 94(1), 421-422.
32. Solberg, T., Sonesson, A. & Woolliams, J. (2008). Genomic selection using different marker types and densities. *Journal of Animal Science*, 86, 2447-2454.
33. Sun, X., Fernando, R. & Dekkers, J. (2016). Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. *Genetics Selection Evolution*, 48(77), 1-18.
34. Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240, 1285-1293.
35. Toghiani, S., Aggrey, S. & Rekaya, R. (2016). Multi-generational imputation of single nucleotide polymorphism marker genotypesand accuracy of genomic selection. *Animal*, 10, 1077-1085.
36. VanRaden, P., Null, D., Sargolzaei, M., Wiggans, G., Tooker, M., Cole, J., Sonstegard, T., Connor, E., Winters, M. & van Kaam, J. (2013). Genomic imputation and evaluation using high-density Holstein genotypes. *Journal of Dairy Science*, 96, 668-6678.
37. Ventura, R. V., Miller, S. P., Dodds, K. G., Auvray, B., Lee, M., Bixley, M., Clarke, S. M. & McEwan, J. C. (2016). Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. *Genetics Selection Evolution*, 48(71), 1-20.
38. Villumsen, T., Janss, L. & Lund, M. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 126, 3-13.
39. Wang, Q., Yu ,Y., Yuan, J., Zhang, X., Huang, H., Li, F. & Xiang, J. (2017). Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp Litopenaeus vannamei. *BMC genetics*, 18(45), 1-9.
40. Weigel, K. A., De Los Campos, G., Vazquez, A. I., Rosa, G. J. M., Gianola, D. & Van Tassell, C. P. (2010). Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. *Journal of Dairy Science*, 93(11), 5423–5435.
41. Yin, T., Pimentel ,E., Borstel, U. K. v. & König, S. (2014). Strategy for the simulation and analysis of longitudinal phenotypic and genomic data in the context of a temperature× humidity-dependent covariate. *Journal of Dairy Science*, 97, 2444-2454.
42. Zhang, Z., Ding, X., Liu, J., Zhang, Q. & de Koning, D. J. (2011). Accuracy of genomic prediction using low-density marker panels. *Journal of Dairy Science*, 94, 3642-3650.