دانشگاه علوم کشاورزی و منابع طبیعی ساری
پژوهشهای تولیدات دامی

ارزیابی صحت پیش‌بینی ژنومی در معماری‌های مختلف ژنومی صفات کمی و آستانه‌ای با جانهی داده‌های ژنومی شبیه‌سازی‌شده، توسط روش جنگل تصادفی

یوسف نادری

استادیار، گروه علوم دامی، دانشگاه آزاد اسلامی، واحد آستارا، آستارا، ایران، (نویسنده مسوول: y.naderi@iau-astara.ac.ir)

تاریخ دریافت: ۹۶/۱۰/۲۵ تاریخ پذیرش: ۹۷/۳/۲۱

چکیده

گزینش ژنومی چالشی امیدبخش برای کشف رموز ژنتیکی صفات کمی و کیفی به‌منظور بهبود رشد ژنتیکی و صحت پیش‌بینی ژنومی در اصلاح دام می‌باشد. به‌علت ناخوانا بودن نسبتی از ژنوتیپ‌ها، پیش‌بینی دقیق صحت ژنومی نیازمند برآورد این نشانگرها از طریق جانهی می‌باشد. در نتیجه هدف این تحقیق برآورد صحت جانهی و عوامل مؤثر برآن و ارزیابی صحت ژنومی روش جنگل تصادفی در برای معماری‌های مختلف ژنومی برای آنالیز صفات کمی و آستانه‌ای دودویی بود. در فاز اول، داده‌های ژنومی از طریق نرم‌افزار QMSim با سطوح متفاوت وراثت‌پذیری (۰/۰۵ و ۰/۲۵)، سطوح مختلف LD (کم و زیاد) و تراکم‌های متفاوت جایگاه‌های صفات کمی (۹۶ و ۹۶۰) و تعداد ۴۸ کروموزم شبیه‌سازی شدند. در فاز دوم، برای شبیه‌سازی شرایط واقعی، بطور تصادفی اقدام به حذف (۵۰ و ۹۰ درصد) برخی نشانگر نموده و در مرحله بعد از طریق نرم افزار Flmpute اقدام به جانهی و پیش‌بینی نقاط گم شده نموده و صحت جانهی مورد ارزیابی قرار گرفت. در فاز سوم، دادهای اصلی و جانهی با استفاده از روش جنگل تصادفی جهت ارزیابی صحت ژنومی صفات کمی و آستانه‌ای استفاده شدند. نتایج نشان داد که با افزایش سطح LD صحت جانهی بهبود می‌یابد. با افزایش نسبت حذف نشانگرها (۹۰ درصد)، اثر صحت جانهی بر صحت پیش‌بینی ژنومی پررنگتر بود. در صفات آستانه‌ای، سناریوی با حد بالای QTL، LD و وراثت‌پذیری و در صفات پیوسته، سناریوی با حد بالای LD و وراثت‌پذیری و حد پایین QTL بیشترین صحت پیش‌بینی ژنومی را در روش جنگل تصادفی به خود اختصاص دادند. به‌طور کلی عملکرد روش جنگل تصادفی در برآورد صحت ژنومی صفات آستانه‌ای نسبت به صفات کمی بهتر بود.

واژه‌های کلیدی: عدم تعادل پیوستگی، صفات آستانه‌ای، یادگیری ماشینی، جانهی، معماری ژنومی

مقدمه

هدف اصلی در برنامه‌های اصلاح دام بهبود خصوصیات ژنتیکی حیوانات در راستای بهبود صفات اقتصادی و بیشینه کردن سود است. استفاده از روش‌های مبتنی بر شجره در اصلاح دام، هرچند که تحول شگرفی در تولیدات حیوانی ایجاد کرد، اما استفاده از این روش‌ها، برای حیوانات جوان، صفات محدود به جنس و رکوردهایی که نیاز به کشتار داشتند با محدودیت‌هایی همراه بود. در نتیجه جستجو برای یافتن روش‌های کامل‌تر به‌نژادی آغاز شد. در دهه‌ی ۹۰ میلادی، پژوهش‌های زیادی با استفاده از اطلاعات DNA و انتخاب با کمک نشانگرها انجام گرفت (۴) که تا حدودی افزایش پیشرفت را به دنبال داشت، اما با توجه به اینکه شمار اندک نشانگرهای ژنتیکی توجیه کمی از واریانس ژنتیکی را در بر می‌گرفتند باعث ایجاد محدودیت در این روش شد (۱۴).

تحول چشمگیر در استفاده از اطلاعات مولکولی در اصلاح دام با ارائه مدل خاصی از MAS که با استفاده از نشانگرهای متراکم تمام ژنوم را پوشش می‌دهند، شروع شد. این روش که گزینش ژنومی نامیده می‌شود، نخستین بار در سال ۱۹۹۸ معرفی شد (۴۴) و در ادامه، روش‌ها و اصول آن توسط موویسن و همکاران (۲۴) ارایه شد. فاکتور کلیدی این روش، صحت برآورد ارزش‌های اصلاحی ژنومی می‌باشد.

در مطالعات ژنومی، هم داده‌های ژنومی واقعی و هم شبیه‌سازی شده، برای اهداف مختلفی مثل بررسی قدرت و دقت روش‌های مختلف آنالیزهای ژنتیکی و مقایسه برنامه‌های اصلاح نژادی مختلف مبتنی بر ژنوم مورد استفاده قرار می‌گیرند. داده‌های واقعی اطلاعات خاصی را منعکس می‌کند، در حالیکه داده‌های شبیه‌سازی شده به محقق اجازه می‌دهد تا جنبه‌های مختلفی مثل معماری ژنتیکی صفت (عدم تعادل پیوستگی (LD)، وراثت‌پذیری و جایگاه‌های ژنی صفات کمی)، تعداد نشانگر و درجه خویشاوندی و ارزیابی برخی از منابع تغییرات مثل رانش ژنی، که با اکثر داده‌های واقعی قابل ارزیابی نیستند، را ممکن و از این طریق در جهت پیشبرد راهکارهای داده‌های واقعی کمک گرفت (۴۸،۳۹).

اگر چه صحت بالای ارزیابی ژنومی در برآورد ارزش‌های اصلاحی مهمترین عامل در پیش‌بینی ژنومی می‌باشد با این حال جنبه‌های اقتصادی این امر را نمی‌توان نادیده گرفت. در این راستا محققان سعی بر آن دارند تا با استفاده از راهکارهایی منطقی هزینه‌های گزاف ارزیابی ژنومی کوتاه مدت را کاهش دهند. یکی از این راهکارها در ارزیابی ژنومی، جانهی (Imputation) ژنوتیپی است که به فرآیند پیش‌بینی ژنوتیپ نشانگرهای تعیین ژنوتیپ نشده در جمعیت آزمون با استفاده از الگوی تنوع هاپلوتیپی جمعیت مرجع، اطلاق می‌گردد (۸). جانهی تراشه‌های کم تراکم به تراکم بالا راهکار مساعدی در زمینه بهبود جنبه‌های اقتصادی ژنومیک بوده است. این تکنیک به محقق اجازه می‌دهد علاوه بر کاهش هزینه‌های توالی یابی، برآورد قابل قبولی از صحت پیش‌بینی ژنومی نیز حاصل شود (۴۰). اندازه مؤثر جمعیت (۴۲)، تراکم نشانگری (۹،۲۹،۴۰)، میزان LD (۲)، فراوانی آلل کمیاب (۱۷) و نوع روش استفاده شده در جانهی (۶) از مهم‌ترین عوامل مؤثر بر صحت جانهی هستند. تحقیقات در مورد گاوهای نلور برزیل (Nelore) نشان داد که تراکم نشانگرها از مهم‌ترین عوامل مؤثر بر صحت جانهی بود. جانهی

[DOI: 10.29252/rap.9.20.129]

تراشه‌های ۷K به تراشه‌های با تراکم بالا (با نرخ حذف ۹۹/۱ درصد نشانگرها) صحت جانهی بالای ۰/۹۲۵ را در پی داشت. با این حال جانهی تراشه‌های ۱۵K به تراشه‌های با تراکم بالا، بهترین عملکرد را از نظر جنبه‌های اصلاحی و اقتصادی در پی داشت (۹). چن و همکاران (۱۰) به بررسی اثر تراکم‌های مختلف نشانگری و اندازه جمعیت مرجع بر صحت جانهی گاوهای نر هلشتاین در سناریوهای مختلف ژنومی پرداختند. نتایج آنها نشان داد که با کاهش تراکم نشانگرها و اندازه جمعیت مرجع، صحت جانهی به شدت کاهش یافت و جانهی تراشه‌های ۶K (در مقایسه با تراشه‌های ۳K) به ۵۰K صحت جانهی بالای ۰/۹۷ را به همراه داشت (۱۰).

فهم بهتر آنچه در انتخاب ژنومیک پیش‌بینی می‌شود به روش آنالیز آماری وابسته است. روش‌های مختلفی در گزینش ژنومی مطرح شده است. یکی از این روش‌ها یادگیری ماشینی می‌باشد (۱۱). جنگل تصادفی یکی از الگوریتم‌های یادگیری ماشینی می‌باشد که اولین بار توسط بریمن (۷) پیشنهاد شد. بعدها گنزالس ریکاردو (۱۶) از جنگل تصادفی برای آنالیز ژنومی صفات آستانه‌ای، لی و همکاران (۲۳) و نگوین و همکاران (۲۸) برای مطالعات پویش ژنومی و غفوری کسبی و همکاران (۱۲) از آن برای ارزیابی ژنومی سناریوهای شبیه سازی شده استفاده کردند. یکی از مزیت‌های کلیدی روش جنگل تصادفی توانایی آن در تجزیه و تحلیل داده‌های با ابعاد بسیار بالا می‌باشد. با این حال آشفتگی در داده‌های آموزشی از محدودیت‌های روش جنگل تصادفی می‌باشد (۱۳).

تحقیقاتی که تاکنون در مورد گزینش ژنومی در ایران و سایر کشورهای پیشرو در این زمینه انجام شده تمرکز بیشتر روی صفات کمی بوده و مطالعات کمتری (۱۲،۱۶،۲۷) در مورد صفات آستانه‌ای از جمله تولید مثلی و مقاومت به بیماری‌ها صورت گرفته است. لذا پژوهش حاضر با هدف بررسی صحت برآورد ارزش‌های اصلاحی ژنومی صفات کمی و آستانه‌ای دودویی با معماری‌های مختلف ژنتیکی شامل تعداد متفاوت QTL، سطوح متفاوت وراثت‌پذیری و LD با استفاده از روش‌های یادگیری ماشینی در دو گروه داده شبیه‌سازی شده و جانهی، انجام شده است.

مواد و روش‌ها

شبیه‌سازی ژنوم

جمعیت‌ها با استفاده از نرم افزار QMSim (۳۶) شبیه سازی شدند. در مرحله اول، برای تولید جمعیتی با LD پایین، یک جمعیت پایه با ۹۸۰۰ ماده و ۲۰۰ نر طی ۱۰۰۰ نسل شبیه سازی شد. برای تولید جمعیتی با LD بالا، پس از شبیه‌سازی جمعیت با LD پایین، تعداد افراد جمعیت از طریق ایجاد یک گلوگاه ژنتیکی (Bottleneck) به ۲۰۰ راس در نسل ۱۱۰۰ کاهش یافت. سپس در آخرین جمعیت پایه، بعد از ۱۰۰ نسل (در نسل ۱۲۰۰) تعداد افراد جمعیت به فاز اول خود یعنی ۹۸۰۰ ماده و ۲۰۰ نر برگشت داده شدند. در گام دوم، برای ایجاد جمعیت مرجع و تایید، همه افراد (۱۰۰۰۰ راس) آخرین نسل جمعیت پایه برای تولید مثل در جمعیت حاضر مورد استفاده قرار گرفتند که در این بین ۲۰۰ راس نر در نظر گرفته شد. نوع سیستم تلاقی تصادفی بود و برای ۱۰ نسل دیگر جمعیت تکثیر شد (نسل ۱۲۱۰). در طراحی جمعیت نهایی، افراد آخرین نسل (نسل ۱۲۱۰) به عنوان جمعیت تایید که این افراد اطلاعات ژنوتیپی داشته اما فاقد اطلاعات فنوتیپی بودند. همچنین افراد ۴ نسل ما قبل جمعیت تایید (نسل ۱۲۰۶ تا ۱۲۰۹) در گروه جمعیت‌های مرجع که این افراد هم اطلاعات ژنوتیپی داشته و هم ارزش‌های اصلاحی ژنومی آنها مشخص می‌باشد طبقه‌بندی شدند. شانس تلاقی در همه‌ی حیوانات برابر (در هر دوجنس) و یک فرزند برای هر زایش در نظر گرفته شد. درصد جایگزینی برای نر و ماده به ترتیب ۸۰ و ۲۰ درصد درنظر گرفته شد. انتخاب حیوانات برتر برای نسل بعد براساس ارزش اصلاحی صورت گرفت. نشانگرها به صورت دو آللی و به صورت فواصل یکسان در بین ۴۸ کروموزوم به طول ۱۰۰ سانتی مورگان توزیع شدند. به ازای هر کروموزوم ۲۰۸ نشانگر شبیه‌سازی شد. در نتیجه ۹۹۸۴ نشانگر برای پنل‌های ۱۰K شبیه‌سازی شد. دو سطح مختلف QTL (۹۶ و ۹۶۰) شبیه‌سازی شد که به صورت تصادفی در طول کروموزم‌ها توزیع شدند. نرخ جهش برای نشانگرها و QTLها در هر جایگاه و در هر نسل $۲/۵\times۱۰^{-۵}$ فرض شد (۳۷).

با توجه به اینکه از دیدگاه آماری توزیع احتمال QTLهای صفات مهم اقتصادی توسط شمار اندکی ژن‌ها دارای اثر عمده و درصد بالایی از ژن‌ها کوچک اثر هستند و این فرضیه به توزیع گاما نزدیکتر است (هایس و گودارد، ۲۰۰۱). در نتیجه توزیع احتمال QTLها، گاما فرض شد. همچنین فراوانی آللی اولیه برای نشانگرها ۰/۵در نظر گرفته شد. در هر نسل و هر جایگاه کل میزان واریانس افزایشی توسط QTL توجیح شد. دو سطح مختلف وراثت‌پذیری (۰/۰۵ و ۰/۲۵) برای هر صفت در نظر گرفته شد. خلاصه جمعیت شبیه‌سازی شده و پارامترهای بکار رفته در جدول ۱ نشان داده شده است. برای شبیه‌سازی فنوتیپ آستانه‌ای دودویی، کد صفر برای دام‌های پایین‌تر از میانگین صفت و کد یک برای دام‌های با فنوتیپ بالاتر از میانگین صفت در نظر گرفته شد. نشانگرهای با فراوانی آللی کمیاب (MAF) کمتر از ۰/۰۵ حذف شدند. برای ارزیابی مدل ۱۰ تکرار از هر سناریو در نظر گرفته شد. شکل ۱ مراحل مختلف پروسه‌ی طراحی شده در تحقیق حاضر را به صورت شماتیک نشان می‌دهد.

جانهی نشانگرها

بعد از شبیه‌سازی جمعیت با تراکم ۱۰K، با کمک برنامه نویسی در نرم‌افزار R و به‌طور تصادفی اقدام به حذف ۹۰ و ۵۰ درصد نشانگرها جمعیت تایید (نسل ۱۲۱۰) نموده و در مرحله بعدی از طریق برنامه‌ی Flmpute (۳۵) اقدام به جانهی و پیش‌بینی نقاط گم شده از طریق روابط فامیلی و الگوریتم‌های برپایه جمعیت شد و در نهایت صحت جانهی از طریق همبستگی داده‌های اصلی و جانهی برای نشانگرها مورد ارزیابی قرار خواهد گرفت.

عدم تعادل پیوستگی

سطح LD برای سناریوهای مختلف شبیه‌سازی شده با استفاده از محاسبه‌ی توان دوم ضریب همبستگی (r^2) بین

[DOI: 10.29252/rap.9.20.129]

همه‌ی جفت نشانگرهای ممکن ارزیابی گردد (۱۷):

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}$$

در این فرمول، $D = f(AB) - f(A)f(B)$ بوده و $f(AB)$، $f(A)$، $f(a)$، $f(B)$ و $f(b)$ به ترتیب، فراوانی‌های مشاهده شده هاپلوتایپ AB و آلل‌های A، a، B و b می‌باشند.

نرم‌افزار PLINK 1.9 (۳۳) برای برآورد LD بین جفت نشانگرهای مختلف در ژنوم همه حیوانات موجود در آخرین نسل مورد استفاده قرار گرفت.

روش آماری

ایده اصلی جنگل تصادفی استفاده از نمونه‌برداری پیاپی از جمعیت و بدست آوردن تقریبی از توزیع واریانس صفت می‌باشد. در داده‌های اعتبار سنجی، درختان طبقه‌بندی توسط بوت استراپینگ در آنالیز جنگل تصادفی ساخته می‌شوند. از طریق استراتژی بگینگ و انتخاب متغیر تصادفی، جنگل تصادفی باعث کاهش خطای پیش‌بینی ژنومی می‌شود.

سه پارامتر اصلی و مهمی که در جنگل تصادفی در مورد کلاسه‌بندی بایستی تنظیم شود عبارت‌اند از mtry، تعداد SNP نمونه‌برداری شده در هر بار نمونه‌گیری تصادفی، ntree یا تعداد بوت استرپ و یا تعداد درختانی که بایستی رشد کنند و معیاری برای انتخاب بهترینSNP برای تقسیم شدن هر گره است، nodesize، حداقل اندازه گره پایانی و که نشان‌دهنده‌ی تعداد مشاهدات در هر شاخه درخت است.

غفوری کسبی (۱۳) تنظیم و بهینه‌سازی پارامترهای اصلی بر عملکرد جنگل تصادفی در پیش بینی ارزش‌های اصلاحی ژنومی را از ملزومات اجرای این روش عنوان کردند. با توجه به اینکه الگوبرداری سناریوهای شبیه‌سازی تحقیق حاضر برگرفته از تحقیق نادری و همکاران (۲۶) بود، در نتیجه از پارامترهای بهینه و تنظیمی آن که بر اساس پارامتر خطای خارج از کیسه انتخاب شدند استفاده شد. در نهایت تعداد متغیر انتخاب در هر گره درخت برابر با دو–سوم تعداد نشانگرها برای صفات کمی (۶۶۰۰) و جذر تعداد نشانگرها برای صفات آستانه‌ای (۱۰۰)، تعداد درخت برابر ۲۰۰۰ و حداقل اندازه گره‌های پایانی برای داده‌های کمی و آستانه‌ای به ترتیب برابر ۵ و یک در نظر گرفته شد.

مدل کلی جنگل تصادفی به صورت زیر است.

$$\hat{f}_{rf}^{P}(x) = \frac{1}{P}\sum_{p=1}^{P} T(x;\ \Psi p)$$

در اینجا P، Ψp امین درخت و برای هر مشاهده $\hat{f}_{rf}^{P}(x)$ از طریق میانگین پیش‌بینی‌های هر درخت محاسبه می‌شود. $T(x;\ \Psi p)_1^P$ در برگیرنده مشاهدات خارج از درخت می‌باشد. سایر حیواناتی که جز این نمونه‌گیری نیستند به عنوان خارج از مجموعه شناخته شده و در اعتبارسنجی هر درخت گزینش می‌شوند.

جنگل تصادفی از مجموعه‌ای از درختان و با استفاده از n نمونه از اطلاعات افراد جمعیت مرجع ایجاد می‌شود. سپس مدل ایجاد شده در جمعیت مرجع بر جمعیت تأیید اعمال می‌شود. در ابتدا یکی از نمونه‌ها وارد هر گره از هر درخت شده و از این نمونه اطلاعات یک نشانگر برای تقسیم‌بندی افراد مورد استفاده قرار می‌گیرد و در نهایت افراد بر اساس اطلاعات ژنوتیپی خود برای نشانگرانتخاب شده دسته‌بندی می‌شوند.این عمل در گره‌های متوالی انجام می‌شود تا در نهایت به گره‌های پایانی میرسیم که در آنها حداکثر یکنواختی وجود خواهد داشت (۱۳). داده‌های ژنومی شبیه‌سازی شده از طریق بسته‌ی RanFoG (۱۶) و نرم‌افزار R مورد آنالیز قرار گرفتند.

در نهایت نتایج حاصل از داده‌ها شبیه‌سازی شده با استفاده از روش جنگل تصادفی در جمعیت تایید براورد شد و صحت پیش‌بینی ژنومی از طریق همبستگی پیرسون بین ارزش‌های اصلاحی پیش‌بینی شده (داده‌های اصلی و جانهی) و ارزش‌های اصلاحی واقعی محاسبه شد (۴۸). فرمول ضریب همبستگی پیرسون به صورت زیر می باشد.

$$r_{EBV;TBV} = \frac{\sum_{i=1}^{n}\left(\widehat{EBV} - \overline{EBV}\right)(TBV - \overline{TBV})}{\sqrt{\left(\widehat{EBV} - \overline{EBV}\right)^2}\sqrt{(TBV - \overline{TBV})^2}}$$

در اینجا $r_{EBV;TBV}$ ضریب همبستگی بین ارزش اصلاحی واقعی و پیش بینی شده، n تعداد مشاهدات، EBV برارزش اصلاحی پیش‌بینی ژنومی با استفاده از جنگل تصادفی و TBV ارزش اصلاحی واقعی (شبیه‌سازی شده) می‌باشد.

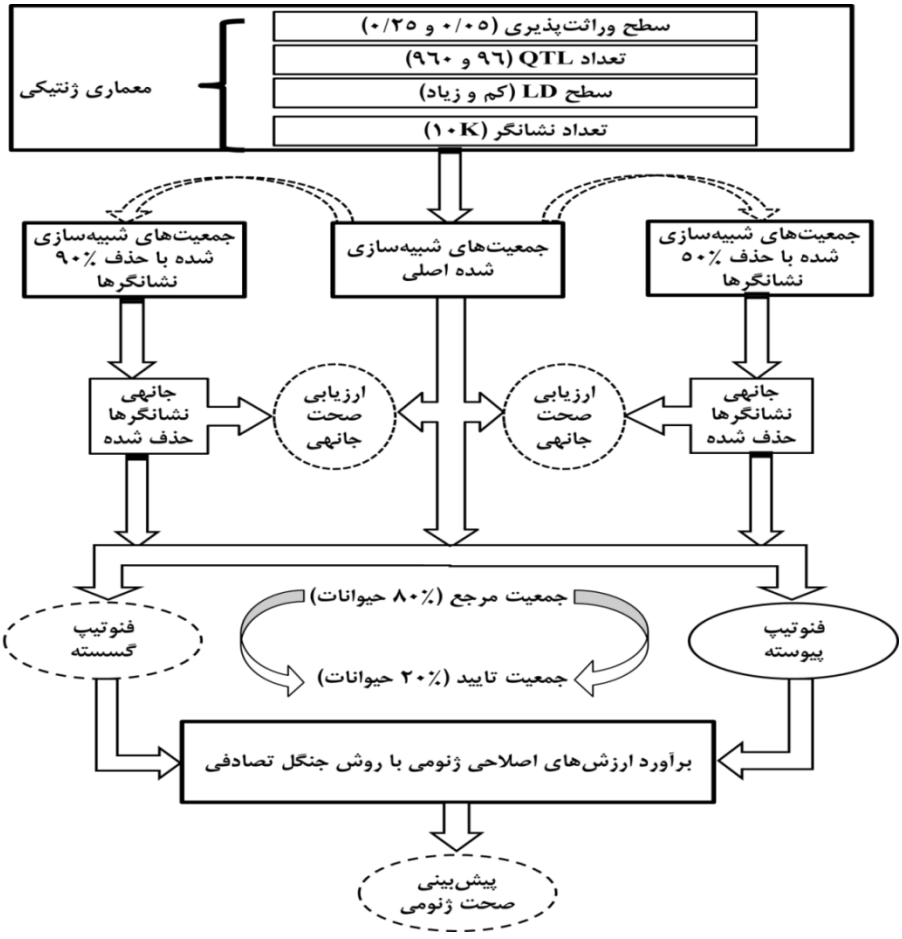
[DOI: 10.29252/rap.9.20.129]

جدول ۱- پارامترهای فرآیند شبیه‌سازی

Table 1. Parameters of the simulation process

ساختار جمعیت	LD بالا*	LD پایین
جمعیت اولیه	۱۰۰۰(۱۰۰۰۰)	۱۰۰۰(۱۰۰۰۰)
فاز اول تعداد نسل (تعداد افراد)	بله	خیر
گلوگاه	۱۱۰۰(۲۰۰)	–
فاز دوم تعداد نسل (تعداد افراد)	۱۲۰۰(۱۰۰۰۰)	–
فاز سوم تعداد نسل (تعداد افراد)	۱۰۰۰۰	
تعداد حیوانات در نسل آخر		
جمعیت اخیر	۲۰۰	
تعداد نرهای در نسل اخیر	۹۸۰۰	
تعداد ماده‌ها در نسل اخیر	۱۰	
تعداد تکثیری جمعیت اخیربعد از نسل ۱۲۰۰	افراد نسل ۱۲۰۶ تا ۱۲۰۹(۴۰۰۰۰ فرد)	
جمعیت مرجع	افراد نسل ۱۲۱۰ (۱۰۰۰۰ فرد)	
جمعیت تایید	۱	
تعداد نتایج به ازای هر زایش	۰/۵	
احتمال نر بودن نتاج	تصادفی	
انتخاب و طرح آمیزش	٪۸۰	
نرخ جایگزینی برای نرها	٪۲	
نرخ جایگزینی برای ماده‌ها	سن بالا/ ارزش اصلاحی برآوردی	
معیار حذف		
ژنوم	۴۸	
تعداد کروموزوم	۱۰۰	
طول هر کروموزوم (سانتی مورگان)	۲ یا ۲۰	
تعداد QTL به ازای هر کروموزوم	گاما (۰/۴)	
اثر آلل‌های QTL	۲۰۸	
تعداد نشانگر به ازای هر کروموزوم	$۲/۵ \times ۱۰^{-۵}$	
نرخ جهش در نشانگر و QTLها	۰/۰۵ و ۰/۲۵	
وراثت‌پذیری		

*: عدم تعادل پیوستگی



شکل ۱- طرح شماتیک استراتژی کل تحقیق

Figure 1. Schematic of the whole process

[DOI: 10.29252/rap.9.20.129]

نتایج و بحث

صحت جانهی و عدم تعادل پیوستگی

جدول ۲ میانگین عدم تعادل پیوستگی و صحت جانهی برای هریک از سناریوهای شبیه‌سازی شده از طریق همبستگی داده‌های اصلی و جانهی شده (۵۰ و ۹۰ درصد) را نشان می‌دهد. به‌طورکلی با افزایش درصد حذف نشانگرها از ۵۰ به ۹۰، صحت ایپیوتیشن کاهش یافت، و این کاهش صحت جانهی برای سناریوهای با LD پایین (۰/۲۸۳-) نسبت به سناریوهای با LD بالا (۰/۲۳۵-) مشهودتر بود. اثر عدم تعادل پیوستگی بر صحت جانهی به صورت نمودار جعبه‌ای در شکل ۲ ترسیم شده است. نتایج تجزیه واریانس نشان داد که عدم تعادل پیوستگی اثر معنی‌داری بر صحت جانهی دارد ($P<0/05$)، به‌طوری‌که میانگین صحت جانهی در سناریوهای با میانگین عدم تعادل پیوستگی بالا (LD=۰/۲۷۲) و پایین (LD=۰/۱۴۳) برای داده‌های با حذف ۵۰ درصد نشانگرها، به ترتیب ۰/۹۷۴ و ۰/۹۵۹ و برای داده‌های با حذف ۹۰ درصد نشانگرها، به ترتیب ۰/۹۵۱ و ۰/۹۳۱ بود. از چند منظر می‌توان پایین بودن صحت ژنومی حاصل از تحقیق حاضر را در مقایسه با سایر تحقیقات مورد تفسیر قرار داد. با توجه به اینکه عوامل مختلفی می‌توانند صحت ارزش‌های اصلاحی ژنومی و ارزیابی ژنومی را تحت تأثیر قرار دهد، این عوامل شامل توزیع اثرات QTL، مقدار عدم تعادل پیوستگی، نوع و تراکم مارکرها، وراثت‌پذیری، نحوه رکوردگیری، تعداد داده‌های فنوتیپی در جمعیت مرجع و فاصله زمانی (تعداد نسل) بین جمعیت مرجع و جمعیت تأیید می‌باشند (۴۳). در نتیجه هر کدام از این عوامل می‌توانند به‌طور جداگانه نتایج حاصل از تحقیقات مختلف را تحت تأثیر قرار دهد.

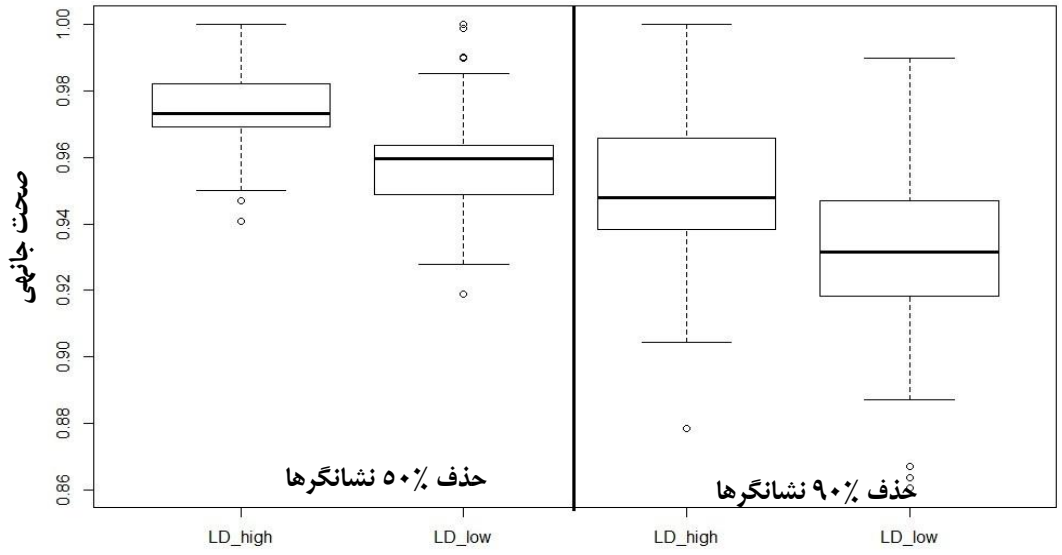
نتایج نشان داد که روش جنگل تصادفی صحت پیش‌بینی ژنومی بالاتری برای صفات آستانه‌ای در مقایسه با صفات کمی دارد که این به ماهیت ناپارامتری این روش بر می‌گردد. در نتیجه علاوه عوامل فوق الذکر، صحت ژنومی تحت تأثیر نوع صفت و مدل آماری مورد مطالعه نتایج متفاوتی را در پی خواهد داشت. با این حال نتایج حاصل از این تحقیق، مقادیر مشابهی از صحت ژنومی با سایر محققان (۲۶) هنگام استفاده از معماری‌های ژنومی یکسان از طریق روش جنگل تصادفی داشته است.

جدول۲- میانگین و انحراف استاندارد عدم تعادل پیوستگی و صحت جانهی بین ژنوتیپ‌های اصلی و جانهی شده در سناریوهای مختلف

Table 2. Mean and standard deviation (in bracket) of linkage disequilibrium and imputation accuracy between imputed and original genotypes in different scenarios

نوع سناریو	میانگین LD در ۰/۱ سانتی مورگان	صحت جانهی (۵۰٪ حذف)	صحت جانهی (۹۰٪ حذف)
سناریو ۱ (LDپایین، ۹۶QTL و وراثت پذیری ۰/۰۵)	۰/۱۴۴	۰/۹۶۳ ۰±/۰۱۵	۰/۹۳۲ ۰±/۰۱۸
سناریو ۲ (LDبالا، ۹۶QTL و وراثت پذیری ۰/۰۵)	۰/۲۸۲	۰/۹۸۰ ۰±/۰۱۳	۰/۹۵۷ ۰±/۰۱۵
سناریو ۳ (LDپایین، ۹۶QTL و وراثت پذیری ۰/۲۵)	۰/۱۴۳	۰/۹۶۰ ۰±/۰۱۵	۰/۹۳۱ ۰±/۰۱۶
سناریو ۴ (LDبالا، ۹۶QTL و وراثت پذیری ۰/۲۵)	۰/۲۶۷	۰/۹۷۳ ۰±/۰۱۳	۰/۹۵۱ ۰±/۰۱۸
سناریو ۵ (LDپایین، ۹۶۰QTL و وراثت پذیری ۰/۰۵)	۰/۱۴۱	۰/۹۵۵ ۰±/۰۱۶	۰/۹۲۶ ۰±/۰۱۸
سناریو ۶ (LDبالا، ۹۶۰QTLو وراثت پذیری ۰/۰۵)	۰/۲۶۵	۰/۹۷۰ ۰±/۰۱۱	۰/۹۴۳ ۰±/۰۱۶
سناریو ۷(LDپایین، ۹۶۰QTLو وراثت پذیری ۰/۲۵)	۰/۱۴۴	۰/۹۵۵ ۰±/۰۱۱	۰/۹۳۳ ۰±/۰۱۷
سناریو ۸(LDبالا، ۹۶۰QTL و وراثت پذیری ۰/۲۵)	۰/۲۷۲	۰/۹۷۶ ۰±/۰۱۶	۰/۹۵۳ ۰±/۰۱۸

h^2: وراثت‌پذیری، LD: عدم تعادل پیوستگی، QTL: جایگاه‌های ژنی صفات کمی



شکل ۲- نمودار جعبه‌ای، صحت جانهی برای سطوح مختلف عدم تعادل پیوستگی

Figure 2. The box-plots of imputation accuracy for the different levels of linkage disequilibrium.

[DOI: 10.29252/rap.9.20.129]

در گوسفندان رامنی، ونتورا و همکاران (۴۲) صحت جانهی از تراشه 5K به 50K را در دامنه‌ای بین ۰/۵۷۸ تا ۰/۸۵۴ گزارش کرد. دلیل این امر را می‌توان به تفاوت معماری ژنومی و همچنین کوچک بودن اندازه مؤثر جمعیت در این نژاد عنوان کرد. گزارشات دیگر حاکی از آن است که جانهی از 7K به 50K نسبت به 3K صحت جانهی بالاتری داشت و تفاوت معنی‌داری مشاهده شد. با این حال طغیانی و همکاران (۴۰) گزارش کرد که جانهی3K به تراشه‌های با تراکم بسیار بالا می‌تواند منجر به نتایج قابل قبولی از صحت داشته باشد.

سایر گزارشات در مورد ذرت بلالی (۱۸)، گاوهای هلشتاین–فریزین استرالیایی (۲۲)، گاوهای هلشتاین آلمان (۲۵)، گاوهای فلکویه (۳۰)، خوک یورکشایر (۲)، گاوهای نلور برزیل(۶)، گاوهای سیاه ژاپنی (۲۹) و گاوهای فریزین و هلشتاین (۳۱) نشان دادند که میزان LD از مهمترین فاکتورهای مؤثر بر صحت جانهی می‌باشند. این در حالی بود که نتایج هوزه و همکاران (۲۰) نشان از تأثیر جزئی میزان LD بر صحت جانهی در نژادهای گاو فرانسوی داشت. کاروالیو و همکاران (۹) صحت جانهی ژنومی را برای تراکم‌های متفاوت نشانگری در گاو نژاد نیلور مورد ارزیابی قرار داد. نتایج آن‌ها نشان داد که با نرخ حذف ۹۹/۱ درصدی نشانگرها و تبدیل تراشه به 7K بازهم صحت جانهی بالای ۰/۹۲۵ را در پی داشت. طبق گزارشات ون رادن و همکاران (۴۱) صحت جانهی در برخی از مناطق ژنوم کمتراز ۰/۶ گزارش شده است که این به ماهیت ژنوم و سطح بسیار پایین LD در این مناطق بستگی داشت.

صحت پیش‌بینی ارزش‌های اصلاحی ژنومی

جدول ۳، صحت پیش‌بینی ارزش‌های اصلاحی ژنومی روش جنگل تصادفی را در هریک از سنایوهای شبیه‌سازی شده (اصلی و جانهی) برای صفات کمی وآستانه‌ای نشان می‌دهد.

به‌طور کلی تفاوت معنی‌داری بین صحت پیش‌بینی ژنومی بین صفات آستانه‌ای و کمی مشاهده شد ($P<۰/۰۵$). افزایش در صحت ژنومی صفات آستانه‌ای نسبت به صفات کمی دامنه‌ای بین ۱۴– ۱۹درصدی داشت. این اختلاف صحت با افزایش نسبت حذف نشانگری بالاتر بود، به‌طوری که در داده‌های با حذف ۹۰٪ نشانگرها، صحت ژنومی برای صفات آستانه‌ای حدود ۰/۰۴۵ واحد نسبت به صفات کمی افزایش یافت. با افزایش درصد حذف نشانگری، صحت پیش بینی ژنومی کاهش یافت، که این کاهش برای صفات پیوسته مشهودتر بود. بیشترین و کمترین میزان صحت ژنومی به ترتیب برای صفات آستانه‌ای، سناریو ۸ و ۵ و برای صفات کمی، سناریو ۴ و ۵ بود.

جدول ۳– میانگین و انحراف استاندارد صحت پیش‌بینی ارزش‌های اصلاحی ژنومی در داده‌های اصلی و جانهی با استفاده از روش جنگل تصادفی

Table 3. Mean and standard deviation of GEBVs accuracies by random forest method in the original and imputed SNP genotypes

شماره سناریو	صحت ژنومی صفات آستانه‌ای			صحت ژنومی صفات کمی		
	اصلی	حذف ٪۵۰ نشانگرها	حذف ٪۹۰ نشانگرها	اصلی	حذف ٪۵۰ نشانگرها	حذف ٪۹۰ نشانگرها
۱	۰/۲۴۶ ۰±/۰۱	۰/۲۱۸ ۰±/۰۲	۰/۱۹۵ ۰±/۰۲	۰/۲۱۴ ۰±/۰۱	۰/۱۹۱ ۰±/۰۳	۰/۱۶۴ ۰±/۰۲
۲	۰/۳۱۴ ۰±/۰۲	۰/۳۰۲ ۰±/۰۲	۰/۲۸۵ ۰±/۰۳	۰/۲۵۷ ۰±/۰۲	۰/۲۳۵ ۰±/۰۲	۰/۲۲۴ ۰±/۰۲
۳	۰/۳۶۱ ۰±/۰۲	۰/۳۲۵ ۰±/۰۳	۰/۲۹۴ ۰±/۰۳	۰/۲۹۴ ۰±/۰۲	۰/۲۶۴ ۰±/۰۲	۰/۲۴۱ ۰±/۰۳
۴	۰/۳۸۳ ۰±/۰۱	۰/۳۵۹ ۰±/۰۲	۰/۳۴۱ ۰±/۰۲	۰/۳۶۰ ۰±/۰۲	۰/۳۵۱ ۰±/۰۲	۰/۳۳۰ ۰±/۰۳
۵	۰/۲۲۵ ۰±/۰۲	۰/۱۹۱ ۰±/۰۲	۰/۱۷۴ ۰±/۰۲	۰/۱۸۵ ۰±/۰۲	۰/۱۵۶ ۰±/۰۲	۰/۱۲۸ ۰±/۰۲
۶	۰/۲۶۳ ۰±/۰۲	۰/۲۵۰ ۰±/۰۲	۰/۲۲۹ ۰±/۰۲	۰/۲۴۳ ۰±/۰۱	۰/۲۲۱ ۰±/۰۱	۰/۲۰۴ ۰±/۰۲
۷	۰/۳۳۰ ۰±/۰۲	۰/۲۹۱ ۰±/۰۲	۰/۲۷۵ ۰±/۰۳	۰/۲۷۶ ۰±/۰۲	۰/۲۴۷ ۰±/۰۲	۰/۲۲۲ ۰±/۰۳
۸	۰/۴۲۱ ۰±/۰۱	۰/۴۰۰ ۰±/۰۲	۰/۳۷۸ ۰±/۰۳	۰/۳۳۱ ۰±/۰۲	۰/۳۱۴ ۰±/۰۲	۰/۲۹۴ ۰±/۰۲

بررسی مدل رگرسیونی تحقیق حاضر نشان داد که صحت جانهی اثر معنی‌داری بر صحت ژنومی دارد (جدول ۴). به‌طوری که افزایش صحت جانهی، افزایش صحت پیش‌بینی ژنومی را در داده‌های جانهی به همراه داشت. اثر مثبت صحت جانهی بر صحت ژنومی، با افزایش حذف نشانگری از ۵۰ به ۹۰ درصد افزایش یافت، و این افزایش برای صفات آستانه‌ای مشهودتر بود.

جدول ۴– اثر صحت جانهی بر صحت پیش‌بینی ژنومی

Table 4. Effect of imputation accuracy on accuracy of genomic prediction

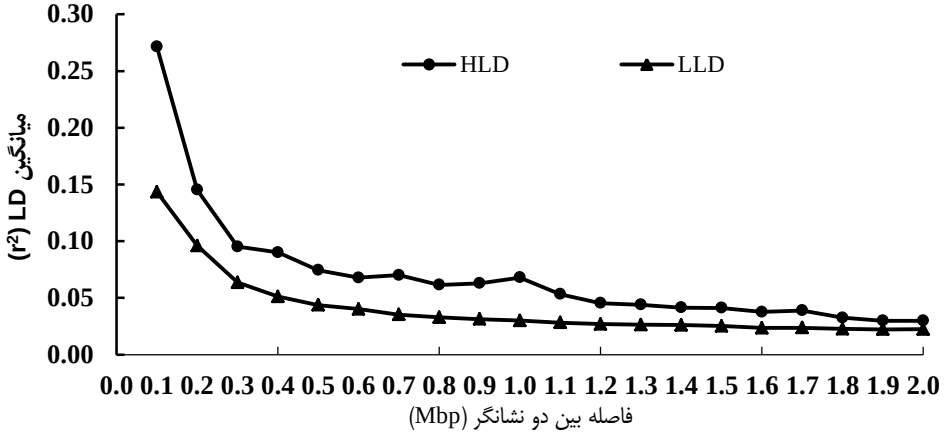
$Accuracy_{bin90\%}$= -3.42+3.93*$IA_{bin90\%}$	R^2_{adj}=0.67
$Accuracy_{con90\%}$= -3.18+3.62*$IA_{con90\%}$	R^2_{adj}=0.656
$Accuracy_{bin50\%}$= -3.98+4.42*$IA_{bin50\%}$	R^2_{adj}=0.470
$Accuracy_{con50\%}$= -3.23+3.60*$IA_{con90\%}$	R^2_{adj}=0.541

$Accuracy_{bin90\%}$ و $Accuracy_{con90\%}$ به ترتیب صحت ژنومی صفات گسسته و پیوسته برای نشانگر های جانهی با حذف ۹۰٪ و $Accuracy_{bin50\%}$ و $Accuracy_{con50\%}$ به ترتیب صحت ژنومی صفات گسسته و پیوسته برای نشانگرهای جانهی با حذف ۵۰٪ و $IA_{bin90\%}$ و $IA_{con90\%}$ به ترتیب صحت جانهی نشانگرهای جانهی با حذف ۹۰٪ و $IA_{bin50\%}$ و $IA_{con50\%}$ به ترتیب صحت جانهی نشانگرهای جانهی با حذف ۵۰٪ می‌باشند.

[DOI: 10.29252/rap.9.20.129]

اثر عدم تعادل پیوستگی بر صحت ژنومی

شکل ۳، میانگین r^2 برای فواصل مختلف ۰/۱ تا دو مگاجفت‌باز (Mbp^2) برای میانگین کروموزوم‌ها را در سناریوهای با LD پایین و LD بالا نشان می‌دهد. میانگین عدم تعادل پیوستگی برای سناریو های با LD پایین و بالا در فاصله ۰/۱ سانتی مورگان به ترتیب ۰/۱۴۳ و ۰/۲۷۱ بود. میانگین r^2 با افزایش فاصله بین نشانگرها کاهش یافت. روند نمایی کاهش LD با افزایش فاصله فیزیکی بین دو نشانگر با نتایج به دست آمده در مطالعات دیگر مطابقت دارد (۴۸،۳۹). نادری و همکاران (۲۶) عدم تعادل پیوستگی را در فاصله ۰/۰۵ سانتی مورگان در سناریوهای با LD پایین و بالا به ترتیب ۰/۲۲۴ و ۰/۴۲۵ گزارش کردند.



شکل ۳- میانگین عدم تعادل پیوستگی در فواصل مختلف ژنوم
Figure 3. Mean of linkage disequilibrium in different intervals of the genome

نتایج آنالیز واریانس نشان داد که اثر LD بر صحت پیش‌بینی ژنومی ناشی از روش جنگل تصادفی معنی‌دار بود. با افزایش نسبت جانهی اثر مثبت افزایش LD بر صحت پیش بینی ژنومی در هر دو سری صفات کمی و آستانه‌ای بیشتر بود. افزایش LD در داده‌های اصلی و ایمپوت شده صفات کمی، افزایش ۲۳ تا ۳۹ درصدی و برای داده‌های آستانه‌ای افزایشی ۱۹ تا ۳۱ درصدی در صحت ژنومی را به همراه داشت، با این حال صحت ژنومی روش جنگل تصادفی در صفات آستانه‌ای برای سطوح مختلف LD بیشتر از صفات کمی بود. بیشترین میزان صحت پیش‌بینی ژنومی برای صفات آستانه‌ای، در سناریو ۸ و برای صفات کمی در سناریو ۴ بود، که در هر دو سناریو نقش ویژه LD برجسته می‌باشد. سطح بالای LD بین QTLها و نشانگرها نشان داد که نشانگرهای با تراکم زیاد سهم بالایی از واریانس ژنتیکی را به خود اختصاص می‌دهند، در نتیجه این امر منجر به عملکرد مثبت جنگل تصادفی می‌شود (۲۶). به‌طور کلی وجود LD در یک جمعیت به شدت با اندازه مؤثر جمعیت در ارتباط است (۴۵). به عنوان یک اصل کلی، وجود LD بین نشانگرو QTL منبع اصلی اطلاعات بوده، و نقش عمده‌ای در صحت پیش‌بینی ارزش‌های اصلاحی ژنومی ایفا می‌کنند (۳۸). جوناس و همکاران (۲۱) در صفات کمی و نادری و همکاران (۲۶) در صفات آستانه‌ای، وجود LD در بین نشانگرها را عاملی تأثیرگذار در بهبود صحت پیش‌بینی ارزش‌های اصلاحی ژنومی عنوان کرد.

صحت بالای جانهی با دامنه تغییرات ۰/۹۵۵- ۰/۹۸۰ (میانگین ۰/۹۶۷) و ۰/۹۲۶- ۰/۹۵۷ (میانگین ۰/۹۴۱) به ترتیب راه حل مناسبی برای برآورد SNPهای با نرخ حذف ۵۰٪ و ۹۰٪ بود. علاوه بر نرخ بالای صحت جانهی در برخی سناریوها، نقش پررنگ اثر LD بر صحت جانهی حایز اهمیت بود. اثر مثبت صحت جانهی بر صحت ژنومی در صفات آستانه‌ای نسبت به صفات کمی مشهودتر بود. به عنوان یک مسئله مهم، بکارگیری ژنوتیپ‌های جانهی با نسبت حذف ۵۰٪ (در مقایسه با حذف ۹۰٪) به جای ژنوتیپ‌های اصلی، تأثیر جزئی و غیر معنی‌داری در کاهش صحت پیش‌بینی ارزش‌های اصلاحی ژنومی داشت. در نتیجه این اثر مثبت جانهی بر صحت پیش بینی ژنومی، می‌تواند در کاهش هزینه ژنومی مؤثر باشد.

ساختار معماری ژنومی (LD و وراثت‌پذیری) و توزیع فنوتیپی صفات (کمی یا آستانه‌ای) از فاکتورهای مؤثر بر صحت پیش‌بینی ارزش‌های اصلاحی ژنومی در روش جنگل تصادفی بودند. به‌طور کلی و با درنظر گرفتن جنبه‌های مختلف معماری ژنومی و صحت جانهی، عملکرد روش جنگل تصادفی بر صحت ژنومی صفات آستانه‌ای نسبت به صفات پیوسته برتری داشت.

[DOI: 10.29252/rap.9.20.129]

منابع

1. Atefi, A., A.A. Shadparvar, and N.G. Hossein-Zadeh. 2016. Comparison of whole genome prediction accuracy across generations using parametric and semi parametric methods. Acta Scientiarum. Animal Sciences, 38(4): 447-453.
2. Badke, Y.M., R.O. Bates, C.W. Ernst, J. Fix and J.P. Steibel. 2014. Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. G3: Genes Genomes Genetics, 4(4): 623-631.
3. Baneh, H., A. Nejati-Javaremi, G. Rahimi-Mianji and M. Honarvar. 2017. Genomic evaluation of threshold traits with different genetic architecture using bayesian approaches. Research on Animal Production, 8: 149-154 (In Persian).
4. Beckmann, J. and M. Soller. 1983. Restriction fragment length polymorphismsin genetic improvement: methodologies, mapping and costs. Theoretical and Applied Genetics, 67(1): 35-43.
5. Bo, Z., J.J. Zhang, N. Hong, G. Long, G. Peng, L.Y. Xu, C. Yan, L.P. Zhang, H.J. Gao and G. Xue. 2017. Effects of marker density and minor allele frequency on genomic prediction for growth traits in Chinese Simmental beef cattle. Journal of Integrative Agriculture, 16(4): 911-920.
6. Boison, S., H.H. Neves, A.P.O. Brien, Y.T. Utsunomiya, R. Carvalheiro, M.da Silva, J. Solkner and J.F. Garcia. 2014. Imputation of non-genotyped individuals using genotyped progeny in Nellore, a Bos indicus cattle breed. Livestock Science, 166:176-189.
7. Breiman, L. 2001. Random forests. Machine learning, 45(1): 5-32.
8. Browning, S.R. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. Human genetics, 124(5): 439-450.
9. Carvalheiro, R., S.A. Boison, H.H. Neves, M. Sargolzaei, F.S. Schenkel, Y.T. Utsunomiya, A.M.P. Obrien, J. Solkner, J.C. McEwan, and C.P. Van Tassell. 2014. Accuracy of genotype imputation in Nelore cattle. Genetics Selection Evolution, 46(1): 69.
10. Chen, L., C. Li, M. Sargolzaei and F. Schenkel. 2014. Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. PLoS One, 9(7): e101544.
11. Clark, S.A., J.M. Hickey and J.H. Van der Werf. 2011. Different models of genetic variation and their effect on genomic evaluation. Genetics Selection Evolution, 43(1): 18.
12. Ghafouri-Kesbi, F., G. Rahimi-Mianji, M. Honarvar and A. Nejati-Javaremi. 2017. Predictive ability of random forests, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation. Animal Production Science, 57(2): 229-236.
13. Ghafouri-Kesbi, F., G. Rahimi-Mianji, M. Honarvar, and A. Nejati-Javaremi. 2016. Tuning and application of random forest algorithm in genomic evaluation. Research on Animal Production, 7 (13): 178-185 (In Persian).
14. Goddard, M. and B. Hayes. 2007. Genomic selection. Journal of Animal breeding and Genetics, 124(6): 323-330.
15. Goddard, M.E. and B.J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics, 10(6): 381-391.
16. Gonzalez-Recio, O. and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genetics Selection Evolution, 43(1): 7.
17. Hayes, B. and M.E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution, 33: 209.
18. Hickey, J.M., J. Crossa, R. Babu and G. de los Campos. 2012. Factors affecting the accuracy of genotypeimputation in populations from several maize breeding programs. Crop Science, 52(2): 654-663.
19. Hill, W. and A. Robertson. 1968. Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics, 38(6): 226-231.
20. Hoze, C., M.N. Fouilloux, E .Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard and P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genetics Selection Evolution, 45(1): 33.
21. Jonas, D., V. Ducrocq and P. Croiseau. 2017. The combined use of linkage disequilibrium–based haploblocks and allele frequency–based haplotype selection methods enhances genomic evaluation accuracy in dairy cattle. Journal of Dairy Science, 100(4): 2905-2908.
22. Khatkar, M.S., G. Moser, B.J. Hayes and H.W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC genomics, 13(1): 538.
23. Li, Y., J. Kijas, J. Henshall, S. Lehnert, R. McCulloch, and A. Reverter. 2014. Using random forests (RF) to prescreen candidate genes: A new prospective for GWAS. in Proc. Proc. of 10th World Congress for Genetics Applied to Livestock Production, British Columbia, Vancouver.
24. Meuwissen, T., B. Hayes and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4): 1819-1829.
25. Mulder, H., M. Calus, T. Druet and C. Schrooten. 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(2): 876-889.
26. Naderi, S., T. Yin and S. Konig. 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(9): 7261-7273.
27. Neves, H.H., R. Carvalheiro and S.A. Queiroz. 2012. A comparison of statistical methods for genomic selection in a mice population. BMC genetics, 13(1): 100.
28. Nguyen, T.T., J.Z. Huang, Q. Wu, T.T. Nguyen and M.J. Li. 2015 .Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. Page S5 in Proc. BMC genomics. BioMed Central Ltd.

[DOI: 10.29252/rap.9.20.129]

29. Ogawa, S., H. Matsuda, Y. Taniguchi, T. Watanabe, A. Takasuga, Y. Sugimoto and H. Iwaisaki. 2016. Accuracy of imputation of single nucleotide polymorphism marker genotypes from low-density panels in Japanese Black cattle. Animal Science Journal, 87(1): 3-12.
30. Pausch, H., B. Aigner, R. Emmerling, C. Edel, K.U. Götz and R. Fries. 2013. Imputation ofhigh-density genotypes in the Fleckvieh cattle population. Genetics Selection Evolution, 45(1): 3.
31. Pausch, H., I.M. MacLeod, R. Fries, R. Emmerling, P.J. Bowman, H.D. Daetwyler and M.E. Goddard. 2017. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genetics Selection Evolution, 49(1): 24.
32. Pimentel, E., C. Edel, R. Emmerling and K.U. Gotz. 2015. How imputation errors bias genomic predictions. Journal of dairy science, 98(6): 4131-4138.
33. Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. De Bakker and M.J. Daly. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3): 559-575.
34. Saatchi, M., J. Ward and D. Garrick. 2013. Accuracies of direct genomic breeding values in Hereford beef cattle using national or international training populations. Journal of animal science, 91(4): 1538-1551.
35. Sargolzaei, M., J. Chesnais and F. Schenkel. 2011. FImpute-An efficient imputation algorithm for dairy cattle populations. Journal of Dairy Science, 94(1): 421-422.
36. Sargolzaei, M. and F.S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics, 25(5): 680-681.
37. Solberg, T., A. Sonesson and J. Woolliams. 2008. Genomic selection using different marker types and densities. Journal of animal science, 86(10): 2447-2454.
38. Sun, X., R. Fernando and J. Dekkers. 2016. Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. Genetics Selection Evolution, 48(1): 77.
39. Thomasen, J.R. 2013. Genomic selection in small dairy cattle populations. Aarhus UniversitetAarhus University, Science and TechnologyScience and Technology, Institut for Molekylærbiologi og GenetikDepartment of Molecular Biology and Genetics, Institut for Molekylærbiologi og Genetik-Center for Kvantitativ Genetik og Genomforskning Department of Molecular Biology and Genetics-Center for Quantitative Genetics and Genomics.
40. Toghiani, S., S. Aggrey and R. Rekaya. 2016. Multi-generational imputation of single nucleotide polymorphism marker genotypes and accuracy of genomic selection. animal, 10(7): 1077-1085.
41. VanRaden, P., D. Null, M. Sargolzaei, G. Wiggans, M. Tooker, J. Cole, T. Sonstegard, E. Connor, M. Winters, and J. van Kaam. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. Journal of dairy science, 96(1): 668-678.
42. Ventura, R.V., S.P. Miller, K.G. Dodds, B. Auvray, M. Lee, M. Bixley, S.M. Clarke and J.C. McEwan. 2016. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. GeneticsSelection Evolution, 48(1): 71.
43. Villumsen, T., L. Janss and M. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. Journal of Animal Breeding and Genetics, 126(1): 3-13.
44. Visscher, P. and C. Haley. 1998. Strategies for marker-assisted selection in pig breeding programmes. Stočarstvo, 52(6): 425-434.
45. Wang, Q., Y. Yu, J. Yuan, X. Zhang, H. Huang, F. Li and J. Xiang. 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(1): 45.
46. Wang, Y., G. Lin, C. Li and P. Stothard. 2016. Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle. Springer Science Reviews, 4(2): 79-98.
47. Wientjes, Y.C., R.F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten and M.P. Calus. 2015. Empirical and deterministic accuracies of across-population genomic prediction. Genetics Selection Evolution, 47(1): 5.
48. Yin, T., E. Pimentel, U.K.V. Borstel and S. Konig. 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(4): 2444-2454.

[DOI: 10.29252/rap.9.20.129]

Evaluation of Genomic Prediction Accuracy in Different Genomic Architectures of Quantitative and Threshold Traits with the Imputation of Simulated Genomic Data using Random Forest Method

Yousef Naderi

Assistant Professor, Department of Animal Science, Astara Branch, Islamic Azad University, Astara, Iran
(Corresponding author: y.naderi@iau-astara.ac.ir)
Received: January 15, 2018 Accepted: June 11, 2018

Abstract

Genomic selection is a promising challenge for discovering genetic variants influencing quantitative and threshold traits for improving the genetic gain and accuracy of genomic prediction in animal breeding. Since a proportion of genotypes are generally uncalled, therefore, prediction of genomic accuracy requires imputation of missing genotypes. The objectives of this study were (1) to quantify imputation accuracy and to assess the factors affecting it; and (2) to evaluate the genomic accuracy of random forest (RF) algorithm to analyze binary threshold and quantitative traits. In the first phase, genomic data were simulated by QMSim software to reflect variations in heritability (h^2 = 0.05 and 0.25), number of QTL (QTL=96 and 960) and linkage disequilibrium (LD=low and high) for 48 chromosomes. In the second phase, for real condition simulating, we randomly masked markers with 50% and 90% missing rate for each scenario; afterwards, hidden markers were imputed using FImpute software, and estimated imputation accuracy. In the third phase, to estimate genomic breeding values, we applied Random forest algorithm for original (before masking a proportion of SNPs) and imputed genotypes with quantitative and quality phenotypes. The accuracy of imputation was improved with increasing level of LD. With increase a major proportion of masked markers (90%), results of current study shed light on the effects of imputation accuracy on accuracy of genomic prediction. In the scenario combining the highest heritability, LD and QTL for threshold traits and in the scenario combining the highest heritability and LD and the least QTL for quantitative traits, random forest method had the best performance of genomic accuracy. Generally, accuracy of genomic prediction for threshold traits had more precise than quantitative trait when using the random forest method.

Keywords: Linkage disequilibrium, Discrete traits, Machine learning, Imputation, Genomic architecture

[DOI: 10.29252/rap.9.20.129]