فصلنامه علمی پژوهشی فیزیولوژی و تکوین جانوری
شماره پیاپی ۵۳، جلد ۱۴، شماره ۲، بهار ۱۴۰۰ صفحه۲۵ تا ۳۶
Qjaphd.sinaweb.net ISSN: ۹۸۸۰-۱۷۳۵

مطالعه تأثیر طول بلوک های هاپلوتیپی در بهبود صحت پیش بینی ژنومی به کمک روش های بیزی در گوسفند

DOR: 20.1001.1.17359880.1400.14.2.2.1

رضا سیدشریفی[۱]، فاطمه علا نوشهر[۲]، نعمت هدایت ایوریق[۱]، جمال سیف دواتی[۱]
۱- عضو هیات علمی، گروه علوم دامی، دانشکده کشاورزی، دانشگاه محقق اردبیلی. اردبیل. ایران.reza_seyedsharifi@yahoo.com
۲-دانش آموخته دکتری گروه علوم دامی، دانشکده کشاورزی دانشگاه تبریز، تبریز. ایران.

تاریخ دریافت:۱۴۰۰/۱/۱ تاریخ پذیرش: ۱۴۰۰/۱/۳۰

چکیده

زمینه و هدف: عدم تعادل پیوستگی(LD) و طرح ساختار بلوک¬های هاپلوتیپ سطح جمعیت پارامترهایی هستند که برای مدیریت مطالعات گسترده ژنومی(GWAS) و درک ماهیت رابطه غیر خطی بین فنوتیپ¬ها و ژنوتیپ ها مفید هستند. در مقایسه با چند شکلی تک نوکلئوتیدی(SNP)، استفاده از آلل های هاپلوتیپ در پیش بینی ژنومی و بهبود صحت پیش بینی کارآمدتر هستند. اما میزان افزایش صحت به چگونگی طراحی بلوک های هاپلوتیپ بستگی دارد. این مطالعه با هدف آزمون اندازه بهینه برای طول هاپلوتیپ در پیش بینی های ژنومی صورت گرفت.

روش کار: در این مطالعه آلل های هاپلوتیپ با توجه به آلل های SNP در بلوک های 125 kb، 250 kb، 500 kb و 1 Mb تعریف و آلل های هاپلوتیپ با فرکانس های کمتر از ۱، ۲/۵، ۵ یا ۱۰ درصد حذف شدند. از دو روش بیزA و بیزB برای پیش بینی اثرات ژنومیSNP ها و هاپلوتیپ ها در سه صفت با سه سطح وراثت پذیری(تولید شیر (h2=۰/۱)، وزن لاشه (h2=۰/۳) و وزن بدن در بلوغ(h2= ۰/٤٥)استفاده شد.

یافته ها: بیشترین صحت پیش¬بینی ژنومی در صفت وزن بدن در زمان بلوغ توسط روش بیزB(۰/٦٥۲) در طول بلوک هاپلوتیپی 250 kb و کمترین توسط روش بیزA در صفت تولید شیر(۰/٤۰۷) در طول بلوک هاپلوتیپی 1 Mb حاصل گردید. بلوک های هاپلوتیپی به طول 250 kb با آستانه فرکانس ۱ درصد، بالاترین میزان صحت پیش بینی ژنومی را ارائه دادند. در مقایسه دو روش بیزA و بیزB، روش بیزB صحت برآورد بالاتری هم در مدل های بر پایه SNP و هم بر پایه آلل های هاپلوتیپ ارائه داد.

نتیجه‌گیری: قرار دادن آلل های هاپلوتیپ به جای SNP ها در مدل آماری، درصورت تعریف مناسب طول هاپلوتیپ، سبب بهبود صحت پیش بینی ژنومی می شود.

واژه های کلیدی: مطالعات گسترده ژنومی، عدم تعادل پیوستگی، بلوک های هاپلوتیپ، SNP، بیزA، بیزB.

مقدمه

دسترسی به ژنوتیپ های چند شکلی تک نوکلئوتیدی(SNP) امکان تخمین ارزش های اصلاحی را با صحت بالاتر در سنین جوانی نسبت به ارزش های اصلاحی بر اساس میانگین والدین فراهم ساخته است(۲۲). پیش بینی ژنومی به طور معمول با استفاده از متغیرهایی که تراکم آلل های SNP را نشان می دهند با اتکا به عدم تعادل پیوستگی(LD) بینSNP ها و جایگاه صفات کمی(QTL) برای تخمین اثرات QTL اجرا می شود(۱۶، ۱۱). جمعیت دام ها معمولاً از تعداد زیادی برادر و خواهر ناتنی تشکیل می شود. این ساختار جمعیتی به ما امکان می دهد تا هاپلوتیپ¬ها به طور دقیق و سریع بازسازی شوند(۶). یک بلوک هاپلوتیپ ناحیه ای از ژنوم است و مجموعه ای از نشانگرهای ژنتیکی مجاور را تشکیل می-دهد(به عنوان مثال SNP ها) که آلل های فازی آن ها احتمالاً با هم به ارث می رسند. چنان چه میزان LD بین آلل های هاپلوتیپ و QTL درون بلوک هاپلوتیپی بیشتر از LD بین آلل های SNP افراد و QTL باشد، پیش بینی می شود. صحت پیش بینی های ژنومی حاصل از آلل های

هاپلوتیپ نسبت به آلل هایSNP، بیشتر است. صحت پیش بینی مدل های هاپلوتیپ تحت تأثیر روش استفاده شده در تقسیم ژنوم در بلوک های هاپلوتیپی هم در سطح داده های شبیه سازی شده(۲۴،۲۵) و هم سطح داده های واقعی قرار دارد(۱۳). انتخاب طول مطلوب هاپلوتیپ به مدل سازی مناسبSNP ها و-QTL ها بستگی دارد. تغییر طول هاپلوتیپ ها می تواند در مدل سازی LD بین SNPها و QTL کمک کننده باشد. از سویی دیگر، چنان چه شباهت هاپلوتیپ ها اساس تعیین روابط خویشاوندی باشد، افزایش تعداد آلل های هاپلوتیپ منجر به کاهش روابط خویشاوندی در هر قطعه می گردد. استفاده از قطعات کوتاه هاپلوتیپ می تواند روابط خویشاوندی بسیار دور را هم در ماتریس حفظ کند، اما این امکان نیز وجود دارد که با نادیده گرفتنLD، فازهای مختلفی را بینSNPها و QTL ایجاد کند. طول مطلوب هاپلوتیپ می تواند میزان روابط خویشاوندی دور و قدیمی شجره را در برابر خطاهایLD که بینSNPها و QTLها در کل جمعیت ممکن است رخ دهد، متعادل سازد(۱۶). کیابانو و همکاران دریافتند که استفاده از آلل های هاپلوتیپ به جای SNPها، سبب افزایش صحت پیش بینی ژنومی در هنگام اجرای مدل بیزی مختلط می شود(۴). در مدل بیزA اطلاعات تمام نشانگرها به طور هم زمان وارد شده و فرض می شود که اثرات نشانگر مستقل بوده و هر نشانگر واریانس خاص خود را دارد و انتظار نمی رود که تمام مناطق ژنومی با فنوتیپ مرتبط باشند. در مدل بیزB پارامتر π تعریف شده و اثرات نشانگر از توزیع مختلطی نمونه گیری می¬شود(۱۷)، به موجب آن اثرات تقریباً π-1 نشانگر در هر تکرار از یک زنجیره مارکوف با همان فرضیات مدل بیزA نمونه برداری شده و بقیه اثرات صفر در نظر گرفته می گردند(۱۷). موویسن و همکاران در سال ۲۰۰۱ پیش-بینی کردند که اگر زمانی تکنولوژی پیشرفت نماید و هزینه های ژنوتیپ کردن کاهش یابد، می توانند تعداد زیادی نشانگر که روی کل ژنوم پراکنده شده¬اند را به طور هم زمان در حیوانات ژنوتیپ کرد و در نتیجه تمام جایگاه های مؤثر بر یک صفت را شناسایی و با برآورد اثرات هر یک از نشانگرها، ارزش اصلاحی حیوانات را تنها از روی اطلاعات نشانگری و بدون نیاز به اطلاعات فنوتیپی آن ها برآورد نمود که آن را ارزش اصلاحی ژنومی(GEBV) نامیدند(۱۷). برای برآورد ارزش های اصلاحی در انتخاب ژنومی، دو دیدگاه ارائه شده است. ابتدا فرض بر این است که تمامی SNPها بر واریانس صفت مؤثر بوده و صفت دارای مدل ژنتیکی نامحدود می باشد روش بهترین پیش بینی کننده ناریب خطی ژنومی (GBLUP) بر پایه این دیدگاه طراحی شده است. در دیدگاه دوم فرض بر آن است که تنها برخی از SNPها بر صفت تأثیر داشته و صفت دارای مدل ژنتیکی ژن های عمده اثر می باشد. بر پایه این دیدگاه برخی روش های بیزی مانند بیزB، بیزC، بیز LASSO بنا نهاده شده اند(۱۷). از عوامل مؤثر بر صحت پیش بینی ارزش های اصلاحی ژنومیک می توان به وراثت پذیری صفت، تعداد افراد و نسل های جمعیت مرجع و تأیید، نوع و تراکم نشانگرها، معماری ژنتیکی صفت کمی و روش استفاده شده در پیش بینی ارزش های اصلاحی ژنومیک اشاره کرد. در بین روش های بیزی روش بیزB و بیز LASSO در حالت توزیع گاما، تأثیرات ژنی بهترین عملکرد را نشان دادند، اما تفاوت آن ها از نظر آماری معنی دار نبود. یکی از ضعف های روش GBLUP این است که فرض می شود کلیه SNPها روی صفت مؤثر بوده و برای همه نشانگرها سهم یکسانی در پیش بینی ارزش اصلاحی در نظر گرفته می شود. در حالی که در روش های بیزی بر حسب توزیع پیشین، وزن های متفاوتی به SNPها اختصاص داده می-شود(۵). هدف از این مطالعه ارزیابی صحت، اریب و زمان محاسباتی روش های پیش بینی ژنومی بیزA و بیزB متناسب با متغیرهای آلل¬های هاپلوتیپ با طول ثابت در

مقایسه با آلل های SNP است. آلل های هاپلوتیپ با طول ۱۲۵ کیلوبایت تا ۱ مگابایت، با آستانه فراوانی آللی متفاوت، از ۱ تا ۱۰ درصد، با استفاده از مدل های بیز A و بیز B به کار گرفته شدند. هم چنین این فرضیه که آیا اندازه بهینه ای برای طول هاپلوتیپ در پیش بینی های ژنومی وجود دارد، مورد آزمون قرار گرفت.

مواد و روش ها

شبیه سازی جمعیت پایه

جمعیت پایه با استفاده از نرم افزار QMSim بر اساس روند پیش رونده در زمان شبیه سازی شدند(۱۹). ۱۰۰ دام دیپلوئید گوسفند نژاد دو منظوره گوشتی-شیری(دورست)، شامل ۵۰ نر و ۵۰ ماده، برای جمعیت پایه(نسل صفر) شبیه¬سازی گردیدند(۱۹). گامت های والدی با فرض عدم تعادل پیوستگی(LD) بر اساس نقشه یابی هالدان گامت های نوترکیب شبیه سازی و به طور تصادفی برای ایجاد یک فرد باهم ترکیب شدند(۱۲). ساختار نسل اول تا نسل ۵۰ با جفت گیری تصادفی دنبال شده تا جمعیت های عدم تعادل پیوستگی ایجاد شود. برای هر نسل،LD با استفاده از آماره r2(رابطه ۱) اندازه گیری و به عنوان میانگین LD تمام SNP ها بود. پس از جمعیت LD ده نسل دیگر(۵۱ تا ۶۰) ساخته شد(۲۰). جمعیت پایه شامل ۱۰۰۰ حیوان غیر خویشاوند(۵۰۰ نر و ۵۰۰ ماده) بود. در این مطالعه، نسل ۵۱ و ۵۲ به عنوان جمعیت مرجع و نسل های دیگر(۵۳ تا ۶۰) به عنوان جمعیت تأیید در نظر گرفته شدند.

شبیه سازی ژنوم

حیوانات با استفاده از آرایه نانویی شامل نشانگر با تراکم بالای گوسفندی(K50) برای شناسایی دقیق مکان-های ژنی مؤثر بر صفات هدف، که دارای SNP 48583 بود، ژنوتیپ شدند. این SNPها روی ۳ کروموزوم پراکنده و طول هر کروموزوم ۲۰۰ سانتی مورگان در نظر گرفته شد. برای فیلتراسیون داده های ژنوتایپینگ از معیارهای فراوانی ژنوتایپینگ نمونه ها، نرخ ژنوتایپینگ نشانگرها در هر نمونه و حداقل فراوانی آللی(MAF) استفاده شد(۱۴). در ابتدا نشانگرهایی که نرخ تعیین ژنوتیپ شده آن ها در نمونه ها کمتر از ۹۵ درصد بود شناسایی و حذف گردیدند، سپس SNP هایی با حداقل فراوانی آللی(MAF) کمتر از ۵ درصد حذف شدند. چنان‌چه فراوانی آللی نشانگرها کمتر از ۵ درصد باشد، منجر به کمتر برآورد شدن آماره r2 برای میزان عدم تعادل پیوستگی بین جفت نشانگرها می شود. QTL دو آللی با توزیع گاما(۰/۴ و ۱/۶۶) در طول ژنوم با تراکم QTL 2 در هر سانتی مورگان ایجاد شد که در مجموع QTL 500 در سطح ژنوم توزیع شد. نرخ جهش SNP ها و QTL ها $۲/۵\times۱۰^{-۵}$ به ازاء هر لوکاس در هر نسل فرض شد(۲۰). سه صفت با سه سطح وراثت پذیری پایین: ۰/۱(تولید شیر(MILK))، متوسط: ۰/۳(وزن لاشه(CARCASS)) و بالا: ۰/۴۵(وزن بدن در زمان بلوغ (MATURE)) در گوسفند شبیه سازی گردید.

عدم تعادل پیوستگی

میانگین LD با جفت SNP های مجاور و میانگین LD در سراسر کروموزوم برای هر کروموزوم برآورد شد. از نرم افزار Haploview v4.2 نیز برای شناسایی بلوک های هاپلوتیپ موجود در هر کروموزوم استفاده گردید. تنوع هاپلوتیپ به صورت $1-\sum f_i^2$ تعریف می شود که fi فراوانی iامین هاپلوتیپ را نشان می دهد. چنان چه آستانه ۹۵ درصدی اطمینانD´ بالاتر از ۰/۹۸ و حد پایین آن بالاتر از ۰/۷ باشد، دو SNP در وضعیت LD قوی در نظر گرفته می شوند. جهت محاسبه مقدار LD در بین نشانگرها از آماره r2 استفاده شد (۱۵).

رابطه ۱:

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

که در آن $D = f(AB) - f(A) \times f(B)$ ، و f (AB)، f (A)، f (a)، f (B)، f (b) به ترتیب فراوانی مشاهده شده برای هاپلوتیپ های AB، A، a، B، b می باشد.

ساختار بلوک های هاپلوتیپ

برای تشکیل فاز هاپلوتیپ از یک مدل مارکوف مخفی از نرم افزار Beagle v4.1 استفاده شد(۲). سپس بلوک های هاپلوتیپ به طور جداگانه با استفاده از نرم افزار PLINK v1.9 طبق روش Haploview v4.1 (۱) بر اساس تخمین D'برای ترکیب جفت SNPها در طول کروموزوم تعریف شدند(۳). با استفاده از مقادیر پیش فرض برای بلوک ها، یک بلوک هاپلوتیپ به عنوان منطقه‌ای تعریف می شود که ۹۵ درصد از جفت های SNP میزان LD بالایی را نشان می دهند(۸). از الگوریتم GLM و نرم افزار Plink برای تولید ماتریس هاپلوتیپی استفاده شد. هاپلوتیپ ها با چهار طول مختلف(۱۲۵ کیلو بایت، ۲۵۰ کیلو بایت، ۵۰۰ کیلو بایت و ۱ مگابایت) ساخته و آلل هاپلوتیپ¬های نادر بر اساس فراوانی آن ها در جمعیت مرجع در چهار آستانه فراوانی مختلف ۱، ۲/۵، ۵ و ۱۰ درصد حذف شدند(۱۴).

تحلیل آماری

از مدل GEBV برای محاسبه مجموع تمام اثرات نشانگر در کل ژنوم استفاده گردید(۱۷):

رابطه ۲:

$$GEBV_i = \sum_{j=1}^{l} x_j \beta_j$$

که در آن l تعداد لوکاس گسترده در ژنوم، x_j نشان دهنده ژنوتیپ فردی در jامین لوکاس با ارزش ۰، ۱ و ۲ می باشد، β_j اثر جایگزینی آلل نشانگر در jامین لوکاس است.

روش بیزA

در روش بیزA فرض می شود که تمامی SNPها دارای اثر، حتی جزئی بوده و تعدادی از SNPها در عدم تعادل پیوستگی با QTLهایی با اثرات متوسط تا بزرگ قرار دارند. اثرات SNPها از توزیع نرمال با واریانس جداگانه برای هر SNP از توزیع کای دو معکوس، نمونه گیری می‌شود. برای به کار بردن این روش از معادله فوق استفاده شد(۱۷):

رابطه ۳:

$$y = 1\mu + Xh + \sum_{j=1}^{k} Z_j \alpha_j \delta_j + e$$

که در آن y بردار N×1 از YD است، μ اثر میانگین، X ماتریس وقوع مقادیر جفت هتروزیس¬ها، h بردار اثرات هتروزیس، k تعداد متغیرهای SNP یا آلل های هاپلوتیپ، Z_j بردار N×1 شماره آللی SNPها(۰، ۱ و ۲) یا آلل های هاپلوتیپی و e بردار N×1 اثرات باقی مانده با میانگین صفر و واریانس δ_e^2 است. در این مدل توزیع شرطی در نظر گرفته شده برای تأثیرات نشانگری توزیع t به صورت:

$$p(b_j|\theta_{bj}, \sigma^2) = t(b_j|df_b, S_b)$$
$$\int N(b_j|0, \sigma_{bj}^2) x^{-2}(df_b|S_b) \partial\sigma_{bj}^2$$

که در آن dfb و Sb به ترتیب درجات آزادی و پارامتر مقیاس و (〖df〗 _b | S_b) x^(-2) توزیع کای اسکور مقیاس دار معکوس است.

روش بیزB

در این روش اکثر SNPهای موجود در منطقه ژنومی QTLای نداشته و بنابراین اثرات آن ها برابر با صفر می باشد، در حالی که تعداد اندکی از آن ها(π-1) در عدم تعادل پیوستگی با QTL قرار داشته و دارای اثر می باشند. در نتیجه اثرات غیرصفر در عدم تعادل پیوستگی بالایی با QTL قرار دارند. بیز B تحت معادله فوق اجرا شد(۱۷):

رابطه ۴:

$$y = 1\mu + Xh + \sum_{j=1}^{k} Z_j \alpha_j \delta_j + e$$

متغیرها در مدل بیزA تعریف شدند به جز α_j نشان‌دهنده اثرات صفر در مدل است.از نرم افزار GenSel v4.73R برای پیش بینی ژنومی با نهادن متغیرهای آلل SNP یا هاپلوتیپ در مدل بیزA و بیزB استفاده شد(۹). در محیط نرم افزار R برای هر تحلیل یک زنجیره مارکوف

مونت کارلو به طول ۴۱ هزار چرخه شامل ۱۰ هزار چرخه اولیه جهت گرم شدن و ۲۰۰ هزار چرخه اصلی به کار رفت که در هر ۵ دور یک بار نتایج ذخیره و در پایان نتایج، ۴۰ هزار چرخه ذخیره گردید. علاوه بر صحت و اریب مدل، تعداد اثرات تصادفی(SNP یا هاپلوتیپ) در مدل گنجانده و زمان محاسباتی نیز اندازه گیری شد.

نتایج

میانگین عدم تعادل پیوستگی

در مطالعه حاضر، میانگین LD محاسبه شده بین تمامی SNP ها ۰/۰۲۷ ± ۰/۲۰۱ (r2) به دست آمد. نتایج نشان می دهد که ۸۹ درصد مقدار LD مورد انتظار در این شبیه سازی حاصل شده است. مقدار LD مورد انتظار بر اساس مطالعه اسوید ۰/۲۱۰ گزارش شده است(۲۲).

تأثیر تراکم نشان گری بر صحت پیش بینی ژنومی

پس از کنترل کیفیت، ۲۵ رأس دام از مطالعه حذف و ۱۷۵ رأس گوسفند برای تجزیه و تحلیل باقی ماندند. علاوه بر این، SNP 1232 با نرخ حذف کمتر از ۹۵ درصد و SNP 4698 با MAF کمتر از ۰/۰۵ حذف شدند. در مجموع SNP 42653 از این فیلترهای کنترل کیفیت عبور کرده و در مجموعه داده ها حفظ شدند. تعداد SNP ها در هر بلوک هاپلوتیپی در ژنوم متفاوت بود(جدول ۲). در تمام بلوک های هاپلوتیپی حداقل تعداد SNP ها در هر بلوک هاپلوتیپی ۱ محاسبه شد. میانگین تعداد SNP ها در هر بلوک هاپلوتیپی بین ۲ تا ۳۲ و حداکثر از ۶ تا ۵۸ بود. نتایج نشان داد که با افزایش فاصله بین نشانگرها، سطح LD کاهش می یابد.

جدول ۱- خلاصه ساختار جمعیت و پارامترهای شبیه سازی

پارامتر های ژنومی	ارزش
تعداد کروموزم	۳
طول ژنوم	cM۶۰۰
تعداد QTL	۱۲۰۰
تعداد نشانگر	۴۸۵۸۳
نرخ جهش QTL	$۲/۵\times ۱۰^{-۵}$
نرخ جهش SNP	$۲/۵\times ۱۰^{-۳}$
اثر توزیع QTL	گاما (α=۱/۶۶ و β=۰/۴)
جایگاه QTL در ژنوم	تصادفی
جایگاه SNP در ژنوم	یکسان
پارامترهای صفت	ارزش
وراثت پذیری	۰/۱، ۰/۳ و ۰/۴۵
واریانس فنوتیپی	۱
صفات محدود به جنس	۰
تعداد نسل	۱۰
نوع تلاقی	تصادفی
جمعیت مرجع	همه حیوانات نسل ۵۱ و ۵۲
جمعیت تأیید	همه حیوانات نسل ۵۳ تا ۶۰

جدول ۲- میانگین و حداکثر تعداد SNPها بر اساس طول بلوک هاپلوتیپی

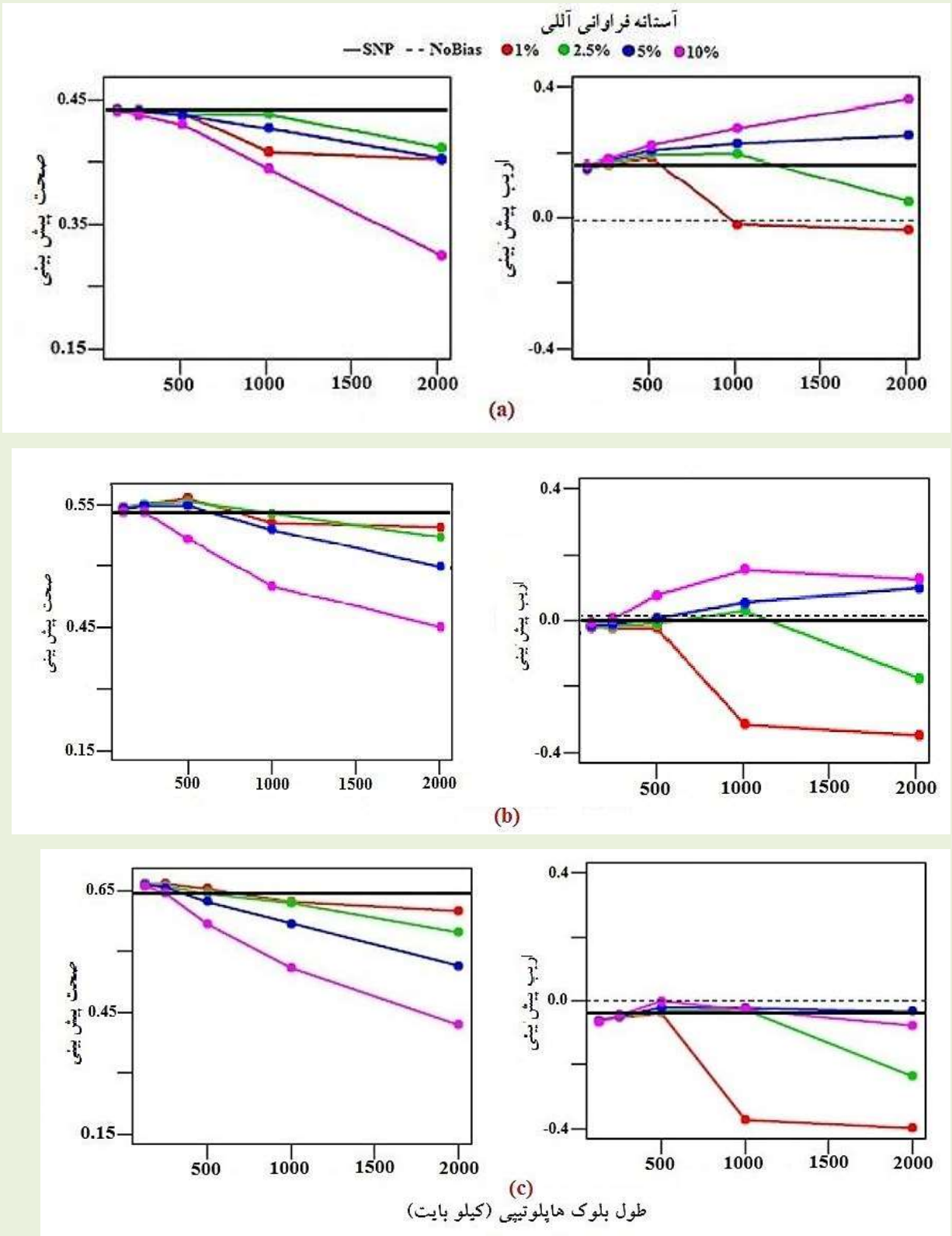
تعداد SNP در هر بلوک هاپلوتیپی		تعداد بلوک هاپلوتیپی	طول بلوک هاپلوتیپی
حداکثر	میانگین		
۶	۲	۱۸۵۲۳	kb۱۲۵
۱۱	۵	۱۱۴۲۱	kb۲۵۰
۱۸	۹	۷۳۴۱	kb۵۰۰
۳۵	۱۷	۳۴۷۷	Mb۱
۵۸	۳۲	۱۸۹۱	Mb۲

صحت و اریب پیش بینی

صحت پیش بینی و اریب مدل بیزA و بیزB در جدول ۳ و نمودار ۱ آورده شده است. صفت وزن در بلوغ(MATURE) با وراثت پذیری بالا دارای بالاترین صحت پیش بینی نسبت به سه صفت مورد مطالعه بود، به دنبال آن صفات کیفیت لاشه(CARCASS) و تولید شیر(MILK)، مطابق با سطح وراثت پذیری خود قرار داشتند. مدل های هاپلوتیپی نسبت به مدل بر پایهSNP از صحت بالاتری برخوردار بوده و میزان اریب مشابه یا پائینی داشتند. افزایش طول بلوک هاپلوتیپی یا آستانه فراوانی آلل هاپلوتیپ سبب کاهش صحت و افزایش اریب مدل هاپلوتیپی می شود. با این حال مدل با طول بلوک هاپلوتیپی۲۵۰ کیلوبایت و فیلتر فراوانی آلل هاپلوتیپی ۱ درصد، در بین طول بلوک های مورد مطالعه بیشترین صحت را ارائه داد. افزایش سطح آستانه فراوانی آللی تأثیر منفی بیشتری بر صحت و اریب برای طول بلوک های هاپلوتیپی بلند(500 - Mb1 کیلو بایت) نسبت به بلوک های هاپلوتیپی کوتاه تر(۲۵۰ – ۱۲۵ کیلوبایت) داشت. اریب پیش بینی برای مدل هایی که از بلوک های هاپلوتیپی کوتاه تر استفاده می کنند، مشابه مدل SNP بود، در حالی که مدل هایی که از بلوک های هاپلوتیپی بلندتر استفاده می کردند، معمولاً اریب بیشتری نسبت به مدل بر پایه SNP داشتند(نمودار ۱).

جدول ۳- صحت و اریب پیش بینی ژنومی سه صفت تولید شیر، وزن لاشه و وزن بدن در بلوغ در چهار سطح بلوک هاپلوتیپ

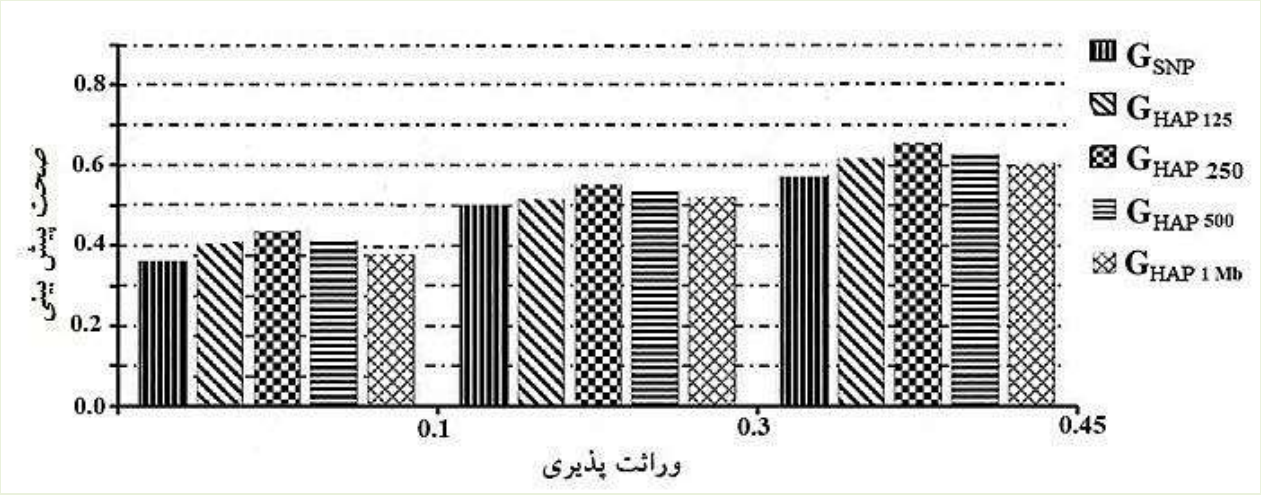
وراثت پذیری	طول بلوک هاپلوتیپ	صحت پیش بینی		اریب پیش بینی	
		BayesA	BayesB	BayesA	BayesB
تولید شیر h^2=۰/۱	۱۲۵kb	۰/۴۲۹ (۰/۰۲)	۰/۴۳۱ (۰/۰۳)	۰/۱۱۷ (۰/۰۲)	۰/۱۱۵ (۰/۰۲)
	۲۵۰kb	۰/۴۳۹ (۰/۰۲)	۰/۴۴۵ (۰/۰۲)	۰/۲۱۶ (۰/۰۱)	۰/۲۱۴ (۰/۰۱)
	۵۰۰kb	۰/۴۲۴ (۰/۰۲)	۰/۴۲۸ (۰/۰۱)	۰/۳۲۹ (۰/۰۲)	۰/۳۲۶ (۰/۰۱)
	۱Mb	۰/۴۰۷ (۰/۰۱)	۰/۴۱۳ (۰/۰۱)	۰/۴۱۵ (۰/۰۲)	۰/۴۱۱ (۰/۰۱)
وزن لاشه h^2=۰/۳	۱۲۵kb	۰/۵۳۷ (۰/۰۱)	۰/۵۴۱ (۰/۰۲)	۰/۱۱۱ (۰/۲)	۰/۰۸۹ (۰/۰۱)
	۲۵۰kb	۰/۵۴۶ (۰/۰۱)	۰/۵۵۲ (۰/۰۲)	۰/۱۱۶ (۰/۰۲)	۰/۱۱۳ (۰/۰۱)
	۵۰۰kb	۰/۵۲۹ (۰/۰۲)	۰/۵۳۵ (۰/۰۲)	۰/۲۲۲ (۰/۰۳)	۰/۲۱۸ (۰/۰۱)
	۱Mb	۰/۵۱۸ (۰/۰۲)	۰/۵۲۶ (۰/۰۱)	۰/۲۲۹ (۰/۰۱)	۰/۲۲۴ (۰/۰۱)
وزن بدن در بلوغ h^2=۰/۴۵	۱۲۵kb	۰/۶۴۱ (۰/۰۱)	۰/۶۴۷ (۰/۰۲)	-۰/۰۸۷ (۰/۰۱)	-۰/۰۶۸ (۰/۰۱)
	۲۵۰kb	۰/۶۴۶ (۰/۰۱)	۰/۶۵۲ (۰/۰۳)	-۰/۱۳۸ (۰/۰۱)	-۰/۱۱۹ (۰/۰۱)
	۵۰۰kb	۰/۶۲۱ (۰/۰۱)	۰/۶۳۴ (۰/۰۱)	-۰/۲۹۹ (۰/۰۱)	-۰/۲۸۴ (۰/۰۱)
	۱Mb	۰/۶۱۷ (۰/۰۱)	۰/۶۲۵ (۰/۰۱)	-۰/۳۸۱ (۰/۰۱)	-۰/۳۷۹ (۰/۰۳)



شکل ۱- صحت و اریب پیش بینی ژنومی صفات تولید شیر(a)، وزن لاشه(b) و وزن بدن در بلوغ(c) در طول های مختلف بلوک های هاپلوتیپ و آستانه فراوانی آللی.

تأثیر سطوح مختلف وراثت پذیری بر صحت پیش-بینی ژنومی

با افزایش مقادیر وراثت پذیری تحت هر مدلی، میانگین صحت ژنومی افزایش یافت. نتایج نشان داد زمانی که وراثت پذیری یک صفت بالاتر باشد، صحت برآورد شده ارزش اصلاحی ژنومی بالاتر خواهد بود. مدل های هاپلوتیپی صحت ژنومی بیشتری نسبت به مدل SNP داشتند. بین مدل های هاپلوتیپی صحت مدل بر پایه طول بلوک هاپلوتیپی ۲۵۰ کیلو بایتی با فراوانی آللی ۱ درصد بیشتر از صحت مدل بر پایه طول بلوک هاپلوتیپی ۱۲۵ کیلو بایت با فراوانی آللی ۱ درصد بود(نمودار ۲). در این مطالعه، بهترین پیشرفت در صحت پیش بینی ژنومی با استفاده از روش بیز B با سطوح مختلف وراثت پذیری در طول های مختلف بلوک هاپلوتیپی در بین صفات مشاهده شد.



نمودار ۲- صحت پیش بینی ژنومی مدل های بر پایه SNP و هاپلوتیپ در سطوح مختلف وراثت پذیری

بحث و نتیجه گیری

مدل های هاپلوتیپ

توانایی یک مدل هاپلوتیپ برای بهبود صحت پیش بینی ژنومی به پیش فرض های قبلی مدل، روش مورد استفاده برای تعریف بلوک های هاپلوتیپی و آلل های هاپلوتیپ، تراکم SNP و تعریف جمعیت مرجع و تأیید بستگی دارد. ویلامسن و همکاران طول بهینه بلوک های هاپلوتیپ را برای صفات شبیه سازی شده با وراثت پذیری ۰/۰۲ تا ۰/۳ ارزیابی کرده و دریافتند که بلوک های هاپلوتیپی به طول ۱ سانتی مورگان نتایج بهتری را بین وراثت پذیری های مختلف ارائه می دهد(۲۵). صحت پیش بینی مدل های هاپلوتیپ که از طول بلوک هاپلوتیپ ۵۰۰ کیلوبایت یا کوتاه تر استفاده می کنند(به طور متوسط کمتر از SNP ۸ در هر بلوک هاپلوتیپ) به طور کلی دارای صحت پیش بینی بالاتری نسبت به مدل SNP هستند، به ویژه هنگامی که آلل های هاپلوتیپ با فرکانس کمتر از ۱ درصد از جمعیت مرجع حذف شوند. ویلامسن و همکاران توسط روش شبیه‌سازی نشان دادند که تعداد مطلوب SNPها در یک بلوک هاپلوتیپ به فاصله بین نشانگرها، میزان LD و ساختار جمعیت بستگی دارد. مدل-های بلوک هاپلوتیپ ۲۵۰ کیلو بایتی با فروانی آللی ۱ درصد، بهترین عملکرد را در بین چهار بلوک هاپلوتیپ بررسی شده در این مطالعه داشتند اما اجرای آن ها نسبت به مدل های SNP بسیار زمان بر بود زیرا تقریباً در این حالت، متغیرها دو برابر می شوند(۲۵). آن ها گزارش کردند که با افزایش طول بلوک هاپلوتیپ، تعداد آلل های

نادر هاپلوتیپ افزایش می یابد، اما ممکن است مشاهدات کافی برای تخمین اثرات آن ها با صحت مناسب وجود نداشته باشد. بنابراین، طراحی هاپلوتیپ ها با استفاده از اطلاعات LD ممکن است برای پیش بینی ژنومی مفیدتر از استفاده بلوک های هاپلوتیپ با طول ثابت باشد، اما از نظر محاسباتی اجرای آن ها به زمان بیشتری نیاز دارد.

تراکم SNP

افزایش تراکم SNP توانایی تمایز آلل های هاپلوتیپ با تفکیک توالی را در یک بلوک هاپلوتیپ تحت تأثیر قرار می دهد: در سطح توالی، همه آلل های هاپلوتیپ واقعی در مجموعه داده ها از نظر تئوری قادر به شناسایی هستند. در حالیکه در تراکم پایین تر، یک آلل هاپلوتیپ مشخص ممکن است نشان دهنده دو یا چند آلل هاپلوتیپ واقعی باشد. این بر توانایی یک مدل برای تخمین دقیق BV یک حیوان برای آن بلوک هاپلوتیپ تأثیر می گذارد، زیرا تأثیر آلل های هاپلوتیپ شناسایی شده علاوه بر خطای پیش بینی، میانگین وزنی اثرات آلل های اصلی هاپلوتیپ واقعی خواهد بود. تلفیق ژنوتیپ در جهش های علی در هاپلوتیپ ها امکان تخمین دقیق تری از اثرات هاپلوتیپ را در مقایسه با نداشتن جهش های علی در هاپلوتیپ فراهم کرده و توانایی تشخیص اثرات کوتاه مدت بین لوکاس-های موجود در همان بلوک هاپلوتیپ را بهبود می بخشد. بنابر این، افزایش تراکم SNP پتانسیل بهبود صحت پیش بینی ژنومی را هنگام استفاده از مدل های هاپلوتیپ دارد. با این حال، افزایش تراکم SNP باعث افزایش تعداد آلل-های هاپلوتیپ شناسایی شده می شود که سبب افزایش تعداد آلل های هاپلوتیپ نادر در یک لوکاس شده و اثر این آلل ها را به سمت صفر کاهش می دهد(۷). هنگام استفاده از تراکم SNP بالا در ایجاد هاپلوتیپ ها، تعداد متغیرهای لازم برای تخمین اثرات اغلب بیشتر از تعداد SNPها است، که سبب افزایش زمان محاسباتی می شود. تعداد متغیرهایی که نیاز به تخمین دارند، می توانند با حذف SNP ها پیش از تولید آلل های هاپلوتیپ کاهش یابند(۴). مطالعه حاضر به وضوح نشان داد که اندازه نمونه و انتخاب SNP تأثیر قابل توجهی بر تعداد بلوک ها و تعداد کل SNP های استنباط شده از یک نمونه جمعیت دارد. زمانی که افراد بیشتری را در نمونه خود گنجانده شود، هم تعداد بلوک ها و هم تعداد کل SNPها افزایش می یابد. علاوه بر این، نتایج ما نشان می دهد که برای نتیجه‌گیری معتبر در مورد تعداد بلوک ها و SNPها، بایستی تعداد نشانگر SNP بیشتری در این مناطق گنجانده شود. تعداد SNP مورد نیاز برای استنباط قابل اعتماد در ساختارهای هاپلوتیپ ممکن است تابعی از ناحیه و جمعیت تحت مطالعه باشد. سولبرگ و همکاران گزارش کردند که با افزایش تراکم نشانگر در هر مورگان، صحت ژنومی افزایش می یابد(۲۱). افزایش تعداد نشانگرها باعث افزایش LD بین ژن‌ها و نشانگرها شده و بنابراین صحت ارزیابی های ژنومی را افزایش می دهد.

اثر وراثت پذیری

در مطالعه حاضر با افزایش تراکم SNP و طول بلوک هاپلوتیپ از ۱۲۵ کیلوبایت به ۱ مگابایت صحت ژنومی افزایش می یابد. با افزایش تراکم SNP و طول بلوک هاپلوتیپ، می توان صحت مقادیر پیش بینی ژنومی را افزایش داد تا عدم تعادل پیوستگی بین نشانگرها، بلوک هاپلوتیپ و QTLها افزایش یابد. در صفات با وراثت-پذیری پایین تر، ارتباط بین فنوتیپ و ارزش ژنتیکی کمتر خواهد بود و برآورد اثرات SNP را می توان با صحت کمتری انجام داد(۱۱). نتایج این مطالعه با نتایج سولبرگ و همکاران مطابقت داشت(۲۱). نیلسون و همکاران گزارش دادند که با افزایش وراثت پذیری صفت از ۰/۲ به ۰/۴، صحت ارزیابی ژنومی حدود ۴ درصد افزایش می‌یابد. زمانی که وراثت پذیری صفت بالا باشد، ارزش فنوتیپی فردی به ارزش ژنتیکی نزدیک تر شده و در نتیجه

ارزش اصلاحی ژنومی با صحت بیشتری برآورد می-شود(۲۵، ۱۸، ۱۰).

زمان محاسبات کامپیوتری

مدل هایی که بر اساس آلل های هاپلوتیپ اجرا می-شوند معمولاً تعداد بیشتری متغیر را نسبت به مدل های بر پایه SNP استفاده کرده و از این رو زمان بیشتری برای محاسبات لازم دارند. تعداد متغیرهای تصادفی ورودی در مدل بیزA، بدون احتساب زمان طراحی مدل و فیلتر آلل هاپلوتیپ بر زمان محاسباتی تأثیر مستقیم دارد. طراحی ۲ مگابایت آلل هاپلوتیپ با آستانه فراوانی ۱۰ درصد تنها سبب تولید ۸۵۰-۷۰۰ آلل هاپلوتیپ می شود که دارای کمترین مقدار صحت پیش بینی در بین مدل ها بود. زمان محاسباتی با افزایش تعداد آلل های هاپلوتیپ افزایش می-یابد. اجرای دقیق ترین مدل برای هر سه صفت مورد مطالعه با استفاده از مدل هاپلوتیپی نسبت به مدل SNP تقریباً دو برابر زمان محاسباتی بیشتری نیاز داشت، زیرا تعداد متغیرهای آن نیز دو برابر شده است. سریع ترین مدل‌ها بسته به صفت مورد مطالعه بین ۲۰ تا ۳۰ دقیقه اجرا شد، اما این امر به شدت با کاهش صحت پیش بینی ژنومی و افزایش اریب همراه بود. بیزB در مقایسه با بیزA زمان محاسباتی کمتری داشت.آلل های هاپلوتیپ با طول ثابت در مقابل کاربرد SNPها می توانند صحت پیش بینی ژنومی را افزایش دهند. بلوک هاپلوتیپ به طول ۲۵۰ کیلوبایت با آستانه فراوانی آللی ۱ درصد منجر به بالاترین صحت در پیش بینی ژنومی شد. طول هاپلوتیپ و فیلتر بر اساس فراوانی آللی هاپلوتیپ تأثیر زیادی در صحت پیش بینی ژنومی دارد. فیلتر فراوانی آلل هاپلوتیپ بالاتر(۱۰ درصد)، تمایل به کاهش صحت پیش بینی ژنومی به ویژه هنگامی که طول هاپلو بلوک ها بزرگ تر بودند، دارد. به منظور نتیجه گیری معتبر در مورد ساختار بلوک های هاپلوتیپ، به جمعیت نسبتاً بزرگ با سطح تراکم بالای نشانگر نیاز دارد.

منابع

1.Barrett, JC., Fry, B., Maller, J., Daly, MJ. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, 21; 263–265.

2.Browning, BL., Browning, SR. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Journal of Human Genetics, 84; 210–23.

3.Chang, CC., Chow, CC., Tellier, LCAM., Vattikuti, S., Purcell, SM., Lee, JJ. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, 4; 7.

4.Cuyabano, BCD., Su, G., Rosa, GJM., Lund, MS., and Gianola, D. (2015). Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. Journal of Dairy Science, 98(10); 7351-7363.

5.De Los Campos, G., Hickey, JM., Pong-Wong, R., Daetwyler, HD., Calus, MP. (2013). Whole genome regression and prediction methods applied to plant and animal breeding. Genetics, 193; 327-45.

6.Ferdosi, MH., Henshall, J. and Tier, B. (2016). Study of the optimum haplotype length to build genomic relationship matrices. Genetics, 48(1); 75.

7.Gianola, D. (2013). Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics, 194(3); 573-596.

8.Gabriel, SB., Schaffner, SF., Nguyen, H., Moore, JM., Roy J., Blumenstiel, B. (2002). The structure of haplotype blocks in the human genome. Science, 296; 2225–9.

9.Garrick, D., Fernando, R. (2013). Implementing a QTL detection study (GWAS) using genomic prediction methodology, genome-wide association studies and genomic prediction. Springer, P; 275-298.

10. Goddard, ME. (2008). Genomic selection: prediction of accuracy and maximization of long term response. Genetics, 136(2); 245–257.

11. Habier, D., Fernando, RL., Dekkers, JCM. (2007). The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177(4); 2389-2397.

12. Haldane, JBS. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. Genetics, 8; 299-309 .

13. Hayes, BJ., Chamberlain, AJ., McPartlan, H., Macleod, I., Sethuraman, L., Goddard, ME.

(2007). Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genetics, 89(4); 215-220.
14. Hess, M., Druet, T., Hees, A., Garrick, D. (2017). Fixed length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genetics Selection Evolution, 49; 54.
15. Hill, WG., Robertson, A. (1968). Linkage disequilibrium in finite populations. Theor. Appl. Genetics, 38; 226-231.
16. Meuwissen, T., Hayes, B., Goddard, M. (2013). Accelerating improvement of livestock with genomic selection. In: H. A. Lewin and R. M. Roberts, editors, Annual Review of Animal Biosciences, Vol 1. Annual Review of Animal Biosciences No. 1. Annual Reviews, Palo Alto, p; 221-237.
17. Meuwissen, THE., Hayes, BJ., Goddard, ME. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4); 1819–1829.
18. Nielsen, HM., Sonesson, AK., Yazdi H., Meuwissen, THE. (2009). Aquaculture, 289; 259–264.
19. Sargolzaei, M., Schenkel, FS. (2009). QMSim: a large-scale genome simulator for livestock. Bioinformatics, 25; 680-681 .
20. Shirali, M., Miraei-Ashtiani, SR., Pakdel, A., Haley, C., Navarro, P., Pong-Wong, R. (2015). A comparison of the sensitivity of the BayesC and Genomic Best Linear Unbiased Prediction (GBLUP) methods of estimating genomic breeding values under different Quantitative Trait Locus (QTL) model assumptions. Iranian Journal of Applied Animal Science, 5(1); 41-46
21. Solberg, TR., Sonesson, AK., Woolliams, JA., Meuwissen, THE. (2008). Genomic selection using different marker types and densities. Journal of Animal Science, 86; 2447-2454 .
22. Sved, JA. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Bioinformatics, 2; 125-141 .
23. VanRaden, PM. (2008). Efficient methods to compute genomic predictions. Journal of Dairy Science, 91(11); 4414-4423.
24. Villumsen, TM., Janss, L. (2009). Bayesian genomic selection: the effect of haplotype length and priors. BMC Proceedings 3 Supp, l(1); S11.
25. Villumsen, TM,. Janss, L., Lund, MS. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. Journal of Animal Breeding and Genetics, 126; 3-13.

Study of the Effect of Haplotype Block Length on Improving the Accuracy of Genomic Prediction Using Bayesian Methods in Sheep

R. Seyedsharifi[1], F.Ala Noshahr [2], N. Hedayat-Evrigh[1], J. Seifdavati[1]

1. University of Mohaghegh Ardabili, Faculty of Agriculture and Natural Resources, Department of Animal Sciences, Ardabi.Iran. **reza_seyedsharifi@yahoo.com**

2. PhD graduated, Dept. of Animal Sciences, Faculty of Agricultural Sciences, University of Tabriz, Tabriz, Iran.

Received:2021.21. 3 **Accepted: 2021.19.4**

Abstract

Inroduction & Objective: Linkage disequilibrium (LD) advancement map and the specification of population-level haplotype block structures are parameters that are helpful for managing the study of the Genome wide Association (GWAS), and to comprehend the nature of non-linear relationship among phenotypes and genotype. Compared with single nucleotide polymorphisms (SNP), genomic prediction fitting haplotype alleles and improve prediction accuracy; but the increase in accuracy belong how the Haplotype block are characterized. The aim of this study was to test the optimal size for haplotype length in genomic predictions.

Material and Method: The Haplotype alleles were defined according the SNP alleles in not covering blocks 125 Kb, 250 Kb, 500 Kb, and 1 Mb. The Haplotype alleles with frequencies below 1, 2.5, 5 or 10% are eliminated. Two methods, Bayes A and Bayes B, were used to predict the genomic effects of SNPs and haplotypes. From Bayes A and B methods to predict the genomic effects of SNPs and haplotypes in three traits with three levels of heritability (milk production (h2 = 0.1), carcass weight (h2 = 0.3) and body weight in Maturity (h2 = 0.45) was used.

Results: The highest genomic prediction obtained in body weight at maturity by Bayesian method B (0.652) during 250 kb haplotypic block and the lowest by Bayesian method A in milk production (0.407) during haplotypic block 1 Mb. Haplotype blocks of 250 kb with a frequency threshold of 1% provided the highest genomic prediction accuracy. Comparing Bayes A and Bayes B methods, Bayes B method provided higher estimation accuracy in both SNP-based and haplotype allele-based models.

Conclusion: : Placing haplotype alleles instead of SNPs in the statistical model, if the haplotype length is properly defined, improves the accuracy of genomic prediction.

Keywords: GWAS, Linkage Disequilibrium, Haplotype Block, SNP, BayesA, BayesB.