



## مقایسه روش‌های مختلف آماری در انتخاب ژنومی گاوهای هلستاین

محمد تیموریان<sup>۱</sup>، محمد مهدی شریعتی<sup>۲</sup> و علی اصغر اسلمی نژاد<sup>۳</sup>

۱- دانشجوی دکتری، دانشگاه فردوسی مشهد (نویسنده مسؤول: m.teimurian@gmail.com)

۲ و ۳- استادیار و دانشیار، دانشگاه فردوسی مشهد

تاریخ دریافت: ۹۳/۳/۲۱ تاریخ پذیرش: ۹۳/۱۰/۲۰

### چکیده

انتخاب ژنومی، ارزش‌های ژنتیکی را در صفات پیچیده کمی از طریق ترکیب داده‌های ژنومی و فنوتیپی با استفاده از مدل‌های آماری پیش‌بینی می‌کند. صحت پیش‌بینی ارزش‌های ژنتیکی در جمعیت تحت انتخاب اثر زیادی بر موفقیت این روش انتخاب دارد. مدل آماری که برای برآورد اثرات نشان‌گری در جمعیت مرجع از آن استفاده می‌شود، از عوامل مؤثر بر صحت پیش‌بینی‌ها می‌باشد. عوامل مختلفی مانند تراکم نشان‌گری، جمعیت مرجع و معماری ژنتیکی صفت بر انتخاب بهترین مدل آنالیز تأثیر گذارند. در این مطالعه، گاوهای هلستاین ایران شبیه‌سازی و مدل‌های مختلف آنالیز در انتخاب ژنومی این جمعیت بررسی شدند. مدل GBLUP و چهار مدل بیزی A، B، LASSO و Cp در حالت‌های مختلفی از جمعیت مرجع، تعداد جایگاه‌های صفت کمی (QTL) و تراکم نشان‌گری مقایسه شدند. روش‌های بیزی نسبت به روش GBLUP صحت برآورد بالاتری داشتند و این برتری در حالت کم بودن تعداد ژن‌های بزرگ اثر و استفاده از اطلاعات جمعیت‌های دیگر در تشکیل جمعیت مرجع ایران بیشتر بود. برای انتخاب ژنومی در جمعیت گاوهای هلستاین ایران، به کارگیری مدل بیز B در جمعیت مرجعی شامل گاوهای نر و ماده و بهره‌گیری از اطلاعات جمعیت‌های دیگر پیشنهاد می‌گردد.

واژه‌های کلیدی: انتخاب ژنومی، روش‌های بیزی، شبیه‌سازی، هلستاین ایران

### مقدمه

طریق مجموع اثرات نشان‌گری آن‌ها حاصل می‌شود (۲۳،۸۶). اطلاعات ژنومی حیوانات منبع مکملی در کنار سایر اطلاعات پیشین می‌باشد ولی چون تعداد نشان‌گرها که مجهولات زیاد و افراد تعداد معادلات نسبت به آن کمتر در نظر گرفته می‌شود، مشکلات محاسباتی را نیز به همراه دارد (۲۲). از طرفی اثرات بعضی از این نشان‌گرها صفر یا نزدیک صفر می‌باشد و اجرای انتخاب ژنومی نیازمند توسعه و بهینه‌سازی روش‌های نوین می‌باشد. همچنین فهم بهتر آنچه در انتخاب ژنومیک پیش‌بینی می‌شود به مدل آنالیز وابسته است (۲).

روش‌های مختلفی در انتخاب ژنومی مطرح شده‌اند و عملکرد هر کدام در شرایط مختلف برای رسیدن به بهترین مدل، ارزیابی شده است. این مدل‌ها به طور کلی در دو دسته طبقه‌بندی می‌شوند: ۱- روش‌هایی که بر اساس رگرسیون مستقیم فنوتیپ به نشان‌گرها می‌باشند ۲- روش‌هایی که از ساختار واریانس-کواریانس برای رسیدن به تابعی از روابط ژنومی بهره می‌گیرند (۵). در گروه اول روش‌های رگرسیون بیزی مثل بیز A، B، Cp و LASSO وجود دارد (۱۷،۱۳). که در بعضی کشورها به صورت دو مرحله‌ای و همراه روش‌های سنتی استفاده می‌شوند. این روش‌ها در صورت زیاد بودن نشانگرها به زمان محاسباتی بالایی نیاز دارند. گروه دوم که حالت‌های مختلفی از بهترین پیش‌بینی نارایب خطی ژنومی<sup>۴</sup> را شامل می‌شود، معادل رگرسیون ریج می‌باشند (۱۴،۱۱) و از ماتریس روابط ژنومی محاسبه شده از نشانگرها (G) به جای ماتریس روابط خویشاوندی (A) استفاده می‌کنند (۲۰،۱۲).

اصلاح‌کنندگان دام از طریق انتخاب حیوانات برتر سعی در افزایش تولید و کاهش هزینه‌های اقتصادی دارند. اغلب صفات مهم اقتصادی دارای ماهیتی کمی و پیچیده بوده و کنترل آن‌ها از طریق تعداد زیادی ژن صورت می‌گیرد و عوامل محیطی نیز بر آن‌ها تأثیرگذار است. در گذشته انتخاب برای این صفات بر اساس پیش‌بینی ارزش‌های ارثی به روش بهترین برآورد نارایب خطی<sup>۱</sup> از طریق داده‌های فنوتیپی و شجره‌ای و به کمک مدلی که برای صفات کمی تعداد نامحدودی ژن کم اثر (اثرات پلی‌ژنی) در نظر می‌گرفت پیشرفت‌هایی به همراه داشته است (۲۱،۱۰).

امروزه با پیشرفت‌های وسیع تکنیک‌های ملکولی و شناسایی تعداد زیادی نشان‌گر ژنتیکی مخصوصاً نشان‌گرهای چندشکلی تک‌نوکلئوتیدی<sup>۲</sup>، اصلاح‌کنندگان دام به دنبال محاسبه ارزش ارثی ژنومی مستقیم از طریق مجموع اثرات نشانگرهای کل ژنوم به کمک روش انتخاب ژنومی هستند که به افزایش صحت برآوردها، کاهش فاصله نسل با انتخاب در سنین پایین‌تر (۱۱)، کاهش هم‌خونی با تأکید بر اطلاعات شخصی به جای اطلاعات خویشاوندی (۴) و در نهایت افزایش بهبود ژنتیکی منجر می‌گردد.

تئوری انتخاب ژنومی<sup>۳</sup> بر اساس پیش‌بینی اثرات جایگاه‌هایی از صفات کمی می‌باشد که در پیوستگی<sup>۴</sup> با نشانگرهای ژنتیکی در موقعیت تعادل قرار ندارند (۱۳،۸،۲). اثرات افزایشی نشانگرها با روش‌های آماری در جمعیت مرجع، شامل مجموعه‌ای از حیوانات تعیین زنوتیپ شده با داده‌های فنوتیپی آن‌ها، برآورد می‌شود و پیش‌بینی ژنومی سایر حیوانات که فقط دارای اطلاعات ژنوتیپی هستند از

1- Best Linear Unbiased Prediction (BLUP)

3- Quantitative Trait Loci (QTL)

5- Least Absolute Selection and Shrinkage operator

2- Single Nucleotide Polymorphism (SNP)

4- Linkage Disequilibrium (LD)

6- Genomic- BLUP (GBLUP)

در جمعیت‌های مختلف متغیر می‌باشد و بررسی روش‌های آماری در جمعیت‌های مختلف الزامی است (۷،۲).  
در این مطالعه، جمعیتی از گاوهای هلشتاین که بیشترین شباهت را به جمعیت واقعی هلشتاین ایران داشته باشد، شبیه‌سازی و صحت پیش‌بینی‌ها در روش‌های آماری مختلف برای جمعیت شبیه‌سازی شده در حالت‌های مختلف بررسی شده است.

### مواد و روش‌ها شبیه‌سازی

با استفاده از نرم‌افزار QMSim (۱۷) شبیه‌سازی در ۱۰ تکرار انجام شد. ابتدا برای تعادل برقرار نکردن پیوستگی بین نشانگرها و جایگاه‌های صفت کمی و برقراری تعادل موتاسیون-رانس، یک جمعیت پایه مشترک با اندازه مؤثر ۱۰۰۰ حیوان (۵۰۰ نر و ۵۰۰ ماده) ایجاد گردید. سپس ۵۰ نسل آمیزش با افزایش تدریجی تعداد به ۵۰۰۰ حیوان صورت گرفت. تعداد حیوانات هر جنس ثابت ماند و سیستم آمیزشی بر اساس ترکیب تصادفی گامت‌ها صورت گرفت.  
از آخرین نسل جمعیت پایه دو جمعیت مختلف، اولی جمعیت ایران و دومی جمعیت گاوهای خارجی، شبیه‌سازی شدند. پارامترهای به کار رفته برای شبیه‌سازی در جدول ۱ نشان داده شده است.

روش‌های ارزیابی برای دسته‌بندی نشانگرها بر حسب سهم واریانس، یا همه نشانگرها را به صورت یک‌نواخت در نظر می‌گیرند و یا غیریک‌نواختی بین نشانگرها و سهم واریانس را لحاظ کرده و تعدادی از نشانگرها دارای سهم بالا و بقیه دارای سهم کم و حتی صفر در نظر گرفته می‌شوند و از روش‌های بیزی غیرخطی استفاده می‌کنند (۲۱،۱۶).  
در بسیاری از کشورها انتخاب ژنومی تبدیل به روش اصلی ارزیابی گاوهای شیری شده است (۲۱،۱۲) و مطالعات زیادی در زمینه‌های مختلف این روش انتخاب مانند اندازه و نوع جمعیت مرجع، تراکم نشان‌گری تراشه‌های تعیین ژنوتیپ و روش‌های آماری صورت گرفته است. به دلیل این که جمعیت مرجع مناسب در ایران وجود ندارد، هنوز این روش جای‌گزین روش‌های مرسوم نشده است. از آن‌جا که فقط حدود ۸۰ گاو نر پروف شده یا جوان در ایران وجود دارد، در صورت تعیین ژنوتیپ تمامی آن‌ها باز هم از تعداد قابل قبولی برای ایجاد جمعیت مرجع برخوردار نیستند و استفاده از گاوهای ماده و اضافه کردن اطلاعات ژنوتیپی سایر کشورها مخصوصاً کشورهایی که صادرات اسپرم بیشتری به ایران دارند اجتناب‌ناپذیر است. ترکیب اطلاعات گاوهای ماده و جمعیت‌های دیگر با جمعیت مرجع کوچک و افزایش تعداد حیوانات صحت پیش‌بینی‌ها را افزایش می‌دهد (۶).  
عملکرد هر یک از روش‌های آنالیز با توجه به معماری ژنتیکی صفت، ساختار جمعیت، تراکم نشانگری و سایر عوامل

جدول ۱- ساختار جمعیت پایه و دو جمعیت مشتق شده

| جمعیت اولیه هلشتاین ایران | جمعیت هلشتاین خارجی | ساختار جمعیت                                   |
|---------------------------|---------------------|------------------------------------------------|
| ۵۰ (۱۰۰۰)                 | ۵۰ (۱۰۰۰)           | مرحله اول: ایجاد جمعیت پایه <sup>۲</sup>       |
| ۵۰ (۵۰۰۰)                 | ۵۰ (۱۰۰۰)           | تعداد نسل‌های جمعیت پایه (تعداد حیوانات) فاز ۱ |
|                           |                     | تعداد نسل‌های جمعیت پایه (تعداد حیوانات) فاز ۲ |
|                           |                     | مرحله ۲: ایجاد جمعیت هلشتاین ایران و خارجی     |
| ۵۰۰                       | ۵۰۰                 | تعداد نر از جمعیت پایه                         |
| ۲۰۰۰                      | ۲۰۰۰                | تعداد ماده از جمعیت پایه                       |
| ۱۰                        | ۱۰                  | تعداد نسل‌ها                                   |
| ۰/۵                       | ۰/۵                 | نسبت نرها                                      |
| تصادفی                    | تصادفی              | سیستم آمیزش                                    |
| ۰/۶                       | ۰/۸                 | نرخ جایگزینی نرها                              |
| ۰/۳                       | ۰/۳                 | نرخ جایگزینی ماده‌ها                           |
| EBV یا                    | فونوتیپ یا          | نحوه انتخاب                                    |
| BLUP                      | -                   | روش برآورد ارزش ارثی                           |

بزرگ اثر، ۱۰ و ۱۰۰ در نظر گرفته شد. برای شبیه‌سازی تراکم نشانگری متوسط معادل با تراشه ۵۰ کیلو بازی، تعداد ۷۵۰۰ نشانگر دو آللی با حداقل فراوانی آللی<sup>۱</sup> بزرگ‌تر از ۰/۱ به صورت تصادفی در کل ژنوم توزیع شدند. ارزش ارثی واقعی هر حیوان از رابطه زیر محاسبه شد:

$$TBV_i = \sum_{j=1}^n b_j \cdot g_{ij}$$

که در این رابطه  $b_j$  اثر افزایشی ژنوتیپ‌های QTL،  $g_{ij}$  ژنوتیپ QTL در جایگاه  $j$ ام و  $n$  تعداد QTL‌های حیوان  $i$ ام می‌باشد. در هر دو صفت، با برابر گرفتن وراثت‌پذیری صفت و QTL، تمامی واریانس ژنتیکی به

برای شبیه‌سازی ورود اسپرم‌های وارداتی از جمعیت خارجی ۱۰۰ گاو نر به جمعیت ایران (۵۰۰ نر و ۲۰۰۰ ماده) وارد شد و این جمعیت برای ۱۰ نسل با آمیزش تصادفی شبیه‌سازی شد. پارامترهای به کار رفته در نسل‌های اخیر تقریباً مشابه شرایط واقعی با یک فرزند برای هر والد ماده در سال و ۵۰ درصد نتاج نرسورت گرفت. نرخ جای‌گزینی نرها و ماده‌ها به ترتیب ۰/۵ و ۰/۳ و انتخاب بر اساس ارزش‌های ارثی برآورد شده از روش BLUP صورت گرفت.

در این مطالعه برای برقراری تناسب بین اطلاعات موجود از ژنوم گاو (تعداد کروموزوم‌ها، طول کروموزوم‌ها و تعداد جایگاه‌ها) ژنومی متشکل از ۳ کروموزوم هر یک به طول ۱۰۰ سانتی‌مورگان شبیه‌سازی شد. وراثت‌پذیری صفت ۰/۲۵ و تعداد جایگاه‌های صفت کمی برای بررسی تاثیر تعداد ژن‌های

واریانس  $S_j^2$  برای اثرات نشانگری  $(S_j^2 | S_j^2) \sim N(0, S_j^2)$  فرض می‌شود که توزیع صفر و کای مربع وارون (با پارامترهای  $v$  و  $S^2$ ) به نسبت  $\pi$  و  $1-p$  برای واریانس‌ها در نظر گرفته می‌شود:

$$\begin{cases} S_j^2 = 0 & p \\ S_j^2 \sim c^{-2}(v, S) & 1-p \end{cases}$$

این مدل با استفاده از نرم‌افزار R و نمونه‌گیری‌های گیبس و متروپولیس-هاستینگز<sup>۵</sup> انجام شد. (۷،۶). با توجه به تعداد کم QTL‌ها نسبت به نشانگرها ( $\frac{1}{7500}$  و  $\frac{1}{100}$ )، پارامتر  $p$  که متغیر انتخاب است، برای صفر گرفتن اثرات نشانگرها برابر ۰/۹۵ در نظر گرفته شد.

بیز Cp: در این روش بر عکس بیز B برای نشانگرهای با اثرات مخالف صفر از یک واریانس مشترک به جای واریانس تعریف شده برای جایگاه‌ها استفاده می‌کند و بر عکس واریانس معین در GBLUP این واریانس برآورد می‌شود. این مدل در سال ۲۰۱۱ از سوی هبیر و همکاران پیشنهاد شد و نسبت  $\pi$  یک متغیر است که در مدل وارد شده و با توجه به معماری ژنتیکی صفت در هنگام آنالیزها تعیین می‌شود و به گونه‌ای تنظیم می‌شود که نشانگرهای دارای اثرات با پارامتر نامعین قابل برآورد هستند. احتمال ترکیبی نسبت  $p$  دارای توزیع پیشین یک‌نواخت می‌باشد:

$$\begin{cases} g_j = 0 & p \\ g_j | S_g^2 \sim N(0, S_g^2) & 1-p \end{cases}$$

بیز LASSO: در این روش مشکل نیاز به تعریف توزیع پیشین در بیزهای قبلی حل شده است و نیازمند اطلاعات کمتری است. این مدل مجموع قدرمطلق مقادیر ضریب رگرسیون را در مدل رگرسیونی به کار می‌برد و برآوردهای آن را می‌توان با در نظر گرفتن یک توزیع پیشین لاپلاس برای برآورد ضرایب، به صورت توزیع پیشین مدل بیزی تعبیر کرد، که توزیع لاپلاس به انقباض بیشتری برای ضرایب نشانگرها به سمت صفر نسبت به بیز A منجر می‌شود و باعث می‌شود که احتمال هم‌زمانی صفر بودن بعضی ضرایب و داشتن اثرات بالا افزایش یابد. نمونه‌گیری گیبس با نرم‌افزار R برای اجرای مدل استفاده شد (۷).

در روش‌های کاربردی برای داده‌های واقعی یا از روش پیش‌بینی خطی GBLUP و یا حالتی از مدل‌های بیزی غیرخطی مثل بیز B استفاده می‌شود. در این روش‌ها یا رگرسیون مستقیم و خطی رکوردهای فنوتیپی بر نشانگرها استفاده می‌شود و یا ارزش‌های ژنومی که تابعی از روابط ژنومی است، با استفاده از ساختار واریانس-کواریانس بین متغیرها استفاده می‌شود (۳). اغلب کشورها ارزش‌های اثری ژنومی مستقیم را با روش‌های سنتی را در روش‌های تک مرحله‌ای و دومرحله‌ای ترکیب می‌کنند.

با توجه به تنوع معماری ژنتیکی در صفات، دو صفت یکی با تعداد کمی ژن بزرگ اثر و دیگری با تعداد زیادی ژن در نظر گرفته شد. هم‌چنین جمعیت مرجع در سه حالت مختلف متناسب با جمعیت هلشتاین ایران تشکیل شد. تراکم

QTL‌ها اختصاص یافت. اثر هر QTL از یک توزیع گاما با پارامترهای ثابت نمونه‌گیری شد.

جمعیت مرجع در سه حالت مختلف برای برآورد اثرات نشانگری تشکیل شد. در حالت اول فقط ۱۰۰۰ گاو نر از جمعیت هلشتاین ایران انتخاب شد، در حالت دوم علاوه بر این ۱۰۰۰ گاو نر، ۱۰۰۰ ماده از هلشتاین ایران به آن افزوده شد و در نهایت در حالت سوم ۱۰۰۰ گاو نر از سایر جمعیت‌ها نیز به حیوانات قبلی اضافه شد.

#### مدل ارزیابی ژنومی

ساختار کلی مدل به شکل خطی به صورت زیر می‌باشد.

$$y_i = m + \sum_j z_{ij} b_j d_j + e_i$$

که در آن  $y_i$  فنوتیپ حیوان  $i$ ام،  $m$  میانگین کل،  $z_{ij}$  یکی از ضرایب ۰، ۱ یا ۲ با توجه به ژنوتیپ نشانگر زام در حیوان  $i$ ام،  $b_j$  اثر جای‌گزینی آللی نشانگر زام،  $d_j$  متغیر شاخص صفر یا ۱ برای بودن یا نبودن نشانگر زام در برآوردها و  $e_i$  بردار مانده‌ها می‌باشد. داده‌ها با چهار مدل بیزی (بیز A، بیز B، بیز LASSO و بیز Cp) و مدل GBLUP آنالیز شدند.

روش G-BLUP: مشابه ارزیابی BLUP در مدل‌های هندرسون است با این تفاوت که ماتریس روابط ژنومی ساخته شده از اطلاعات مولکولی جایگزین ماتریس روابط شجره‌ای می‌شود و درایه مربوط به جفت حیواناتی که ژنوتیپ‌های یکسان بیشتری داشته و در نتیجه شباهت ژنومی بالاتری دارند بزرگ‌تر می‌باشد. روابط ژنومی از رابطه زیر محاسبه می‌شوند.

$$G = \frac{ZZ}{2 \sum p_i (1 - p_i)}$$

که در این رابطه  $p_i$  فراوانی آلل دوم در جایگاه  $i$ ام و  $Z$  ماتریس حاصل از تفاضل  $P$  از  $M$  می‌باشد که  $P = 2 \sum (p_i - 0.5)$  و  $M$  ماتریس کدهای ژنوتیپی به صورت ۰، ۱ و -۱ برای هموزیگوت اول، هتروزیگوت و هموزیگوت دوم می‌باشد (۲۰).

بیز A: مدل رگرسیون بیزی بر نشانگرها از سوی مويسن و همکاران (۱۶) پیشنهاد شد. همه نشانگرها دارای اثر بوده و برای اثرات نشانگری، توزیع پیشین نرمال به صورت مستقل با میانگین صفر و واریانس  $S_j^2$  فرض می‌شود  $(g_j | S_j^2) \sim N(0, S_j^2)$  که این واریانس برای نشانگرها متفاوت بوده و از یک توزیع کای مربع وارون که نوعی توزیع  $t$  می‌باشد، با درجه آزادی و پارامتر مقیاس خاص تبعیت می‌کند  $S_j^2 \sim c^{-2}(v, S)$  که این دو پارامتر ثابت فرض می‌شوند (۸). این مدل با استفاده از برنامه‌نویسی به زبان R و ۱۰۰۰ دور نمونه‌گیری گیبس<sup>۲</sup> انجام شد که ۱۰۰ دور اول آن دوره‌های سوخته<sup>۳</sup> در نظر گرفته شد (۷).

بیز B (انتخاب متغیر تصادفی بیزی):<sup>۴</sup> بخشی از نشانگرها که اطلاعاتی در مورد فنوتیپ فراهم می‌آورند در مدل قرار می‌گیرند. علی‌رغم معایب فرمولی، تقریباً قابل قبول‌ترین مدل رایج می‌باشد. ترکیبی از توزیع پیشین نرمال با میانگین صفر و

1- Minor Allele Frequency (MAF)

3- Burn-in Period

5- Metropolis-Hastings

2- Gibbs Sampling

4- Stochastic Search Variable Selection (SSVS)

## نتایج و بحث

صحت پیش‌بینی‌ها در مدل‌های مختلف آماری برای حالت‌های متفاوت از تعداد QTLها و نوع جمعیت مرجع در جدول ۲ نشان داده شده‌اند.

نشاندگی متوسط و بالا که تراکم‌های رایج تراشه‌های تعیین ژنوتیپ هستند، شبیه‌سازی شدند و ارزش‌های ارثی ژنومی جمعیت تأیید با استفاده از برآورد اثرات نشاندگی حاصل از جمعیت مرجع در هر یک از مدل‌های آماری پیش‌بینی شدند. برای مقایسه روش‌ها از معیار صحت برآوردها به صورت مجذور هم‌بستگی بین ارزش ارثی پیش‌بینی شده و واقعی در جمعیت تأیید استفاده شد.

جدول ۲- صحت پیش‌بینی ارزش‌های ارثی در حالت‌های مختلف جمعیت مرجع و تعداد QTL  
Table2. The Accuracy of the prediction of breeding value for 3 different RF and 2 different QTL number

| تعداد QTL | جمعیت مرجع | GBLUP | بیز A | بیز B | بیز LASSO | مدل  | بیز Cp |
|-----------|------------|-------|-------|-------|-----------|------|--------|
| ۱۰        | ۱          | ۰/۲۹  | ۰/۴   | ۰/۴۷  | ۰/۵۱      | ۰/۵۴ | ۰/۵۴   |
|           | ۲          | ۰/۳۲  | ۰/۴۳  | ۰/۴۹  | ۰/۵۳      | ۰/۵۵ | ۰/۵۵   |
|           | ۳          | ۰/۳۸  | ۰/۵   | ۰/۵۴  | ۰/۵۶      | ۰/۵۹ | ۰/۵۹   |
| ۱۰۰       | ۱          | ۰/۳۲  | ۰/۳۸  | ۰/۴۲  | ۰/۴۸      | ۰/۵  | ۰/۵    |
|           | ۲          | ۰/۳۴  | ۰/۴   | ۰/۴۶  | ۰/۵۱      | ۰/۵۲ | ۰/۵۲   |
|           | ۳          | ۰/۳۹  | ۰/۴۴  | ۰/۵   | ۰/۵۳      | ۰/۵۵ | ۰/۵۵   |

جمعیت ۱: ۱۰۰۰ گاو نر ایران، جمعیت ۲: ۱۰۰۰ گاو نر + ۱۰۰۰ گاو ماده، جمعیت ۳: ۱۰۰۰ گاو نر + ۱۰۰۰ گاو ماده + ۱۰۰۰ گاو سایر جمعیت‌ها

اگرچه کم بودن روابط خویشاوندی بین افراد جمعیت مرجع با جمعیت تحت انتخاب باعث کاهش صحت پیش‌بینی‌ها می‌شود، از طرفی دیگر پائین بودن اندازه موثر در این جمعیت شبیه‌سازی شده نسبت به جمعیت‌های واقعی باعث بالا رفتن صحت برآوردها در این مطالعه شده است. روش‌های بیزی نسبت به روش GBLUP از صحت برآورد بالاتری برخوردار بود. این برتری در صفتی که تحت تأثیر تعداد کمی ژن بزرگ‌اثر قرار داشت، بیشتر بود و زمانی که صفت تحت تأثیر تعداد زیادی ژن با اثرات کم قرار داشت، کاهش یافت. روش GBLUP کمتر به معماری ژنتیکی صفت وابسته است و زمانی که صفت تحت تأثیر تعداد زیادی ژن قرار دارد، از نظر صحت پیش‌بینی‌ها تفاوت چندانی با روش‌های بیزی ندارد. در نتیجه زمانی که در انتخاب ژنومی، معماری ژنتیکی صفت از نظر تعداد و میزان اثر جایگاه‌های صفت کمی آن کاملاً مشخص نباشد، روش مناسبی خواهد بود. برتری روش بیز B نسبت به GBLUP از سوی کلارک و همکاران (۲) در یک مطالعه شبیه‌سازی گزارش شده است. کلارک و همکاران (۲) وجود ژن‌های بزرگ اثر در معماری ژنتیکی صفت را عاملی برای افزایش این برتری عنوان کرده‌اند. در بین روش‌های بیزی صحت برآوردها در روش بیز Cp و بیز LASSO نسبت به بیز B و مخصوصاً بیز A بیشتر بود. متغیر انتخاب p در مدل بیز Cp برای حالت‌های ۱۰ و ۱۰۰ مارکر به ترتیب ۰/۹۹۲ و ۰/۹۸ حاصل شد. مقادیر متفاوت این پارامتر نشان‌دهنده این مطلب است که در بیز B، آگاهی از معماری ژنتیکی صفت برای انتخاب پارامتر p می‌تواند بر عملکرد این روش مؤثر باشد. افزایش صحت روش بیز Cp برای صفت چربی شیر را کلمبانی و همکاران (۳) برای گاوهای هلشتاین فرانسه به دست آورده‌اند، اما این روش بیزی و بیز LASSO نیازمند بررسی دقیق‌تر پارامترها در مدل هستند و در نتیجه در مورد گاوهای هلشتاین ایران که اطلاعات کاملی در دسترس نیست، نیازمند مطالعات

بیشتری می‌باشند. زمانی که از اطلاعات جمعیت دیگر استفاده شد، برتری روش‌های بیزی نسبت به GBLUP افزایش یافت. افزایش صحت برآوردهای روش بیزی نسبت به روش GBLUP در جمعیت مرجع چند نژاده (هلشتاین و جرسی) از طرف بن هیز و همکاران (۱۲) گزارش شده است. برتری روش‌های بیزی نسبت به GBLUP زمانی که رابطه بین جمعیت مورد انتخاب با جمعیت مرجع کم است از سوی گائو و همکاران (۷) نیز گزارش شده است. با توجه به کمبود گاو نر پروف شده در گاو شیری ایران، استفاده از اطلاعات سایر جمعیت‌های هلشتاین که فاز آل‌های نشاندگی هم‌جهت دارند، عاملی برای افزایش صحت برآوردها خواهد بود. ترکیب اطلاعات جمعیت‌ها در پیش‌بینی ژنومی زمانی که فاز آل‌های نشاندگی در بین جمعیت‌ها حالت عکس داشته باشند، نیازمند مطالعات بیشتر می‌باشد. با این که در مطالعات شبیه‌سازی زیادی صحت روش‌های بیزی نسبت به GBLUP بیشتر بوده است (۱۳، ۱۱) اما مطالعات با داده‌های واقعی بر این موضوع اتفاق نظر ندارند. دلیل اصلی این موضوع این است که در مطالعات شبیه‌سازی معماری ژنتیکی صفات کاملاً مشابه شرایط واقعی در نظر گرفته نمی‌شود (۱۰). اگرچه صحت برآوردها در روش‌های بیزی نسبت به GBLUP مخصوصاً در هنگام استفاده از اطلاعات جمعیت‌های دیگر بیشتر است اما ایرادی که در این روش‌ها وجود دارد، این است که مسیر مشخصی برای محاسبه ارزش‌های ارثی حیوانات بدون رکورد فنوتیپی وجود ندارد (۱۲) و استفاده از روش GBLUP در صورت تصحیح نحوه محاسبه ماتریس G و بسط این ماتریس برای حالت‌های مختلف می‌تواند گزینه مناسبی در برآوردها باشد (۲۰). در روش‌های آماری بررسی شده‌ی این مطالعه از توزیع پیشین t-student استفاده شده است و برآوردهای توزیع پسین با زنجیره‌های مونت کارلوی مارکوف<sup>۱</sup> مثل نمونه‌گیری گیبس یا متروپولیس با حداکثرسازی کل توزیع پسین انجام گرفته است. روش زنجیره مارکوفی مونت کارلو برای داده‌های

1- Markov Chain Monte Carlo (MCMC)

نتایج این مطالعه نشان می‌دهد با توجه به جمعیت هلشتاین ایران، تشکیل جمعیت مرجع مناسب شامل ترکیبی از نرها و ماده‌ها در کنار استفاده از اطلاعات سایر جمعیت‌ها باعث افزایش صحت پیش‌بینی‌ها شده و برآورد اثرات نشانگری در چنین شرایطی نیازمند استفاده از مدل‌های آنالیز مناسب است که تعیین پارامترهای مدل‌های نوین، نیازمند اطلاعات کافی از معماری ژنتیکی صفات و جمعیت مورد مطالعه می‌باشد. با توجه به این محدودیت‌ها، مدل‌های GBLUP و Bیز در انتخاب ژنومی گاوهای هلشتاین ایران بهترین گزینه‌های ممکن به نظر می‌رسند و برای اجرای مدل‌های دیگر، مطالعات بیشتر داده‌های واقعی پیشنهاد می‌گردد.

با حجم بالا زمان‌بر بوده و روش‌های الگوریتم یادگیری و حداکثرسازی امید ریاضی می‌توانند با انجام مطالعات بیشتر جایگزین‌های مناسبی برای این روش‌ها باشند (در الگوریتم‌های حداکثرسازی امید ریاضی فقط به نسبتی از نشانگرها امکان برقراری فقدان تعادل پیوستگی با جایگاه‌های صفت کمی داده شده و اطلاعات پیشین توزیع اثرات نشانگری را با هم ترکیب می‌کند). در چند مقاله روش‌هایی ذکر شده‌اند که از توزیع پیشین لاپلاس استفاده کرده و برای به دست آوردن برآورد اثرات به جای استفاده از روش حداکثرسازی کل توزیع پسین از روش الگوریتم حداکثرسازی امید ریاضی<sup>۱</sup> بهره گرفته‌اند (۱۹،۱۶) که موفقیت چندانی نداشته‌اند.

### منابع

1. Brondum, R.F., G. Su, M.S. Lund, P.J. Bowman, M.E. Goddard and B.J. Hayes. 2012. Genome position specific priors for genomic prediction. *BMC Genomics*, 13: 543-557.
2. 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: 18-31.
3. Colombani, C., A. Legarra, S. Fritz, F. Guillaume, P. Croiseau, V. Ducrocq and C. Robert-Granie. 2013. Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesC $\pi$  methods for genomic selection in French Holstein and Montbéliarde breeds. *Journal of Dairy Science*, 96: 575-591.
4. Daetwyler, H.D., B. Villanueva, P. Bijma and J.A. Woolliams. 2007. Inbreeding in genome-wide selection. *Journal of Animal Breeding and Genetics*, 124: 369-376.
5. De Los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182: 375-385.
6. De Roos, A., B. Hayes and M. Goddard. 2009. Reliability of genomic predictions across multiple populations. *Genetics*, 183: 1545-1553.
7. Gao, H., G. Su, L. Janss, Y. Zhang and M.S. Lund. 2013. Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. *Journal of Dairy Science*, 96: 4678-4687.
8. Gianola, D., M. Perez-Enciso and M.A. Toro. 2003. On marker-assisted prediction of genetic value: beyond the ridge. *Genetics*, 163: 347-361.
9. Gianola, D. and J.B. Van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178: 2289-2303.
10. Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetics*, 136: 245-257.
11. Habier, D., R. Fernando and J. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177: 2389-2397.
12. Hayes, B., P. Bowman, A. Chamberlain, K. Verbyla and M. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution*, 41: 51-66.
13. Meuwissen, T., B. Hayes and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157: 1819-1829.
14. Hayes, B.J., P.M. Visscher and M.E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91: 47-60.
15. Meuwissen, T. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genetics Selection Evolution*, 41: 35-49.
16. Meuwissen, T., T.R. Solberg, R. Shepherd and J.A. Woolliams. 2009. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genetics Selection Evolution*, 41: 50-63.
17. Park, T. and G. Casella. 2008. The bayesian lasso. *The American Statistical Association*, 103: 681-686.
18. Sargolzaei, M. and F.S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, 25: 680-693.
19. Shepherd, R.K., T.H. Meuwissen and J.A. Woolliams. 2010. Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. *BMC Bioinformatics*, 11: 529-542.
20. Vanraden, P. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91: 4414-4423.
21. Vanraden, P., C. Van Tassell, G. Wiggans, T. Sonstegard, R. Schnabel, J. Taylor and F. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science*, 92: 16-24.
22. Verbyla, K.L., B.J. Hayes, P.J. Bowman and M.E. Goddard. 2009. Short Note: Accuracy of genomic selection using stochastic search variable selection in australian holstein friesian dairy cattle. *Genetic Research*, 91: 307-311.
23. Xu, S. 2003. Estimating polygenic effects using markers of the entire genome. *Genetics*, 163: 789-801.

## Comparison of Methods for the Implementation of Genomic Selection in Holstein

Mohammad Teimurian<sup>1</sup>, Mohammad Mahdi Shariati<sup>2</sup> and Ali Asghar Aslaminejad<sup>3</sup>

1- Ph.D. Student, Ferdowsi University of Mashad (Corresponding author: m.teimurian@gmail.com)

2 and 3- Assistant Professor and Associate Professor, Ferdowsi University of Mashad

Received: June 11, 2014

Accepted: January 10, 2015

### Abstract

Genomic selection combines statistical methods with genomic data to predict genetic values for complex traits. The accuracy of prediction of genetic values in selected population has a great effect on the success of this selection method. Accuracy of genomic prediction is highly dependent on the statistical model used to estimate marker effects in reference population. Various factors such as density markers, reference population and genetic architecture affect choosing the best model for analysis. The aim of this study was to evaluate methods for genomic selection in the Iranian Holstein population. Four Bayesian linear regression models, Bayes-A, Bayes-B, Bayes-Cp, Bayesian-LASSO and a genomic linear method (GBLUP) were compared using stochastic simulation across three types of reference population and a range of numbers of quantitative trait loci (QTL) in two classes of marker density. Bayesian methods had a higher accuracy than GBLUP especially when the number of loci was low and data from other population were used. Construction of reference population by bulls and cows together with data from other populations and using Bayes B is suggested for genomic selection in Iranian Holstein.

**Keywords:** Genomic selection, Iranian Holstein, Statistical methods, Simulation