Iranian Journal of Animal Science
Vol 48, No 4, Winter 2018 (573-584)
DOI: 10.22059/ijas.2018.248547.653596

علوم دامی ایران
دورهٔ ۴۸، شمارهٔ ۴، زمستان ۱۳۹۶ (ص ۵۸۴-۵۷۳)

ارزیابی کارایی تکنیک نمونه‌گیری تجمعی بوت‌استرپ بر صحت روش بهترین پیش‌بینی ناأریب خطی ژنومی

خبات خیرآبادی[۱]، جمال فیاضی[۲*]، هدایت‌اله روشنفکر[۳] و رستم عبداللهی آرپناهی[٤]
۱، ۲ و ۳. دانشجوی دکتری، دانشیار و استاد، گروه علوم دامی، دانشگاه علوم کشاورزی و منابع طبیعی خوزستان
٤. استادیار، گروه علوم دام و طیور، پردیس ابوریحان دانشگاه تهران
(تاریخ دریافت: ۱۳۹٦/۱۱/۱٦ - تاریخ پذیرش: ۱۳۹۷/۱/۱۳)

چکیده

به منظور افزایش صحت ارزیابی‌های روش بهترین پیش‌بینی ناأریب خطی ژنومی (GBLUP)، تکنیک نمونه‌گیری تجمعی بوت‌استرپ (بگینگ) بکار گرفته شد. بدین منظور ژنومی حاوی ۱۰۰۰۰ نشانگر تک‌نوکلئوتیدی دو آللی (SNP) با فواصل یکسان روی ۱۰ کروموزوم هریک به طول ۱۰۰ سانتی‌مورگان شبیه‌سازی شد. برای ایجاد عدم تعادل پیوستگی (LD) بین SNPها و جایگاه‌های ژنی کنترل‌کنندهٔ صفات کمی (QTL)، به مدت ۱۰۰ نسل بین ۱۰۰ فرد (۵۰ نر و ۵۰ ماده) آمیزش تصادفی صورت گرفت. در نسل ۱۰۱ (جمعیت مرجع) تعداد نمونه‌ها به ۱۰۰۰ یا ۲۰۰۰ فرد افزایش یافت و برای این افراد یک ارزش فنوتیپی شبیه‌سازی شد. سپس اثر نشانگرها در این جمعیت با استفاده از روش GBLUP و روش ترکیبی GBLUP با تکنیک بگینگ (BGBLUP) برآورد گردید. در آخر با استفاده از ضرایب رگرسیونی برآورد شده و با توجه به ژنوتیپ نشانگرها برای افراد جوان نسل‌های ۱۰۲ تا ۱۰۵ که جمعیت تائید نام دارند و فاقد فنوتیپ‌اند، ارزش‌های اصلاحی ژنومی محاسبه شد. براساس نتایج پژوهش حاضر، صحت ارزش‌های اصلاحی ژنومی روش GBLUP در همهٔ حالات از حیث عددی بالاتر از BGBLUP بوده ($p > 0.05$) و در مورد نسل اول جمعیت تائید (نسل ۱۰۲) و بدون توجه به توزیع آثار جایگزینی ژنها، با جمعیتی برابر ۱۰۰۰ (یا ۲۰۰۰) فرد در جمعیت مرجع دامنهٔ صحت ارزش‌های اصلاحی ژنومی روش GBLUP از ۰/۳۳۹±۰/۰۴۹ (۰/۴۱۲±۰/۰۴۲) برای صفت با توارث‌پذیری ۵ درصد تا ۰/۷۲۸±۰/۰۱۵ (۰/۷۸۳±۰/۰۱۵) برای صفت با توارث‌پذیری ٦۵ درصد متفاوت بود و مقادیر مشابه برای روش BGBLUP نیز به ترتیب ۰/۳۳۸±۰/۰۴۷ (۰/۴۱۱±۰/۰۴۲) و ۰/۷۲۵±۰/۰۱٦ (۰/۷۸۰±۰/۰۱۵) بود.

واژه‌های کلیدی: انتخاب ژنومی، تکنیک بگینگ، صحت ارزیابی.

Evaluation of the effectiveness of bootstrap aggregating sampling technique in the accuracy of genomic best linear unbiased prediction method

Khabat Kheirabadi[1], Jamal Fayazi[2*], Hedayatollah Roshanfekr[3] and Rostam Abdollahi-Arpanahi[4]
1, 2, 3. Ph.D. Candidate, Associate Professor and Professor, Department of Animal Science, Khuzestan Agricultural Sciences and Natural Resources University, Iran
4. Assistant Professor, Department of Animal and Poultry Science, Aburaihan Campus, University of Tehran, Iran
(Received: Feb. 5, 2018 - Accepted: Apr. 2, 2018)

ABSTRACT

In order to increase the accuracy of genomic best linear unbiased prediction method (GBLUP), bootstrap aggregating sampling (bagging) technique was applied. In this order a genome consisted of 10,000 bi-allelic single nucleotide polymorphism (SNP) over ten chromosomes, with 100 cM length each, was simulated. To generate linkage disequilibrium (LD) between SNPs and quantitative trait loci (QTL), random mating was simulated for 100 generations between 100 individuals (50 males and 50 females). Then in generation 101 (reference population) number of individuals increased to 1000 or 2000 and their phenotypes were also simulated. Then the marker effects were estimated in this population using GBLUP method or combined this method with bagging technique (BGBLUP). By using these regression coefficients and according to the genotype markers for juvenile individuals in generations 102 to 105, called validation population which had no phenotype, genomic breeding values were predicted. According to the finding of this research, the accuracies of genomic breeding values of GBLUP method were higher than those for BGBLUP ($p > 0.05$) and about the first testing set (102 generation) and regardless of QTL effects with a population of 1000 (or 2000) observations in the reference set, the range of GBLUP accuracy was 0.339±0.049 (0.412±0.042) for a trait with 0.05 heritability to 0.728±0.015 (0.783±0.015) for a trait with 0.65 heritability, whereas the accuracy of BGBLUP method were varied between 0.338±0.047 (0.411±0.042) to 0.725±0.016 (0.780±0.015).

Keywords: Accuracy of evaluation, bagging technique, genomic selection.

* Corresponding author E-mail: j_fayazi@ramin.ac.ir

مقدمه

با پیشرفت در حوزهٔ دانش بیوتکنولوژی مولکولی، امروزه اطلاعات وسیعی از کل ژنوم گونه‌های مختلف گیاهی و حیوانی به وسیلهٔ نشانگرهای چندشکلی تک‌نوکلئوتیدی (SNP)[1] در دسترس بشر قرار گرفته است. به‌طوری‌که از ترکیب آنها با اطلاعات فنوتیپی می‌توان جهت برآورد شایستگی ژنتیکی (Meuwissen *et al*., 2001) یا پیش‌بینی ارزش‌های فنوتیپی (Lee *et al*., 2008) صفات مهم و اقتصادی استفاده کرد. این روش که با عنوان انتخاب ژنومی در سال ۲۰۰۱ میلادی مطرح شد (Meuwissen *et al*., 2001)، در واقع به تصمیم‌گیری انتخاب براساس ارزش اصلاحی ژنومی اشاره دارد (Hayes *et al*., 2009). اساس روش انتخاب ژنومی، بر وجود عدم تعادل پیوستگی (LD)[2] بین نشانگرها و جایگاه ژنی صفات کمی (QTL)[3] است (Meuwissen *et al*., 2001). در این روش ابتدا با استفاده از اطلاعات ژنوتیپی و فنوتیپی مجموعه‌ای از افراد یک جمعیت مرجع اثر تمام نشانگرها برآورد شده، سپس با استفاده از این ضرایب رگرسیونی ارزش اصلاحی ژنومی افراد جوانی که تنها دارای اطلاعات ژنوتیپی‌اند (کاندیدهای انتخاب) پیش‌بینی می‌گردد. عمده مزیت روش مذکور، امکان کاهش فاصله بین نسل‌ها از طریق پیش‌بینی ارزش اصلاحی با صحت بالا در بدو تولد یا قبل از آن می‌باشد (Meuwissen *et al*., 2001; Hayes *et al*., 2009). نتایج یک مطالعهٔ شبیه‌سازی نشان داده که با به‌کارگیری روش انتخاب ژنومی، حدود ۹۲ درصد هزینهٔ شرکت‌های تجاری اصلاح‌نژاد گاوهای شیرده کاهش می‌یابد (Schaeffer, 2006). مزیت‌های منحصر به فرد این روش سبب گشته تا به سرعت مورد توجه بسیاری از کشورهای صنعتی قرار گیرد. به‌طوری‌که بعد از گذشت تنها چهار سال از آغاز استفاده گاوهای نر جوان ژنومی (فاقد دختران رکورددار در زمان استفاده) در برنامه‌های ملی اصلاح نژاد گاوهای شیری ایالات متحده آمریکا، در سال ۲۰۱۲ به ترتیب حدود ۵۱ و ۵۲ درصد نژادهای هلشتاین و جرسی ایالات مذکور را دختران این گاوها به خود اختصاص دادند (Hutchison *et al*., 2014).

به‌طورکلی میزان موفقیت برنامه‌های انتخاب ژنومی تحت تأثیر صحت برآورد اثرات نشانگری جمعیت مرجع قرار دارد (Goddard & Hayes, 2009)، لذا در این حوزه از دانش تاکنون روش‌های آماری مختلفی جهت برآورد ضرایب رگرسیونی نشانگرهای مولکولی مطرح شده است. دامنهٔ این روش‌ها را می‌توان به‌طورکلی از بیزB که تاکید بر مؤثر بودن تنها تعداد محدودی جایگاه ژنی دارد، تا بهترین پیش‌بینی نااُریب خطی ژنومی (GBLUP) که فرض آن یکسان بودن واریانس همهٔ جایگاه‌های ژنی است محدود کرد (Meuwissen *et al*., 2001). نتایج برخی تحقیقات نشان داده که در مورد بسیاری از صفات، صحت برآورد (همبستگی بین ارزش‌های اصلاحی ژنومی و ارزش اصلاحی معمول) این روش‌ها تا حدود زیادی مشابه است (Hayes *et al*., 2009; VanRaden *et al*., 2009). به‌هرحال به دلیل سادگی مباحث آماری و پائین بودن هزینهٔ محاسباتی، روش GBLUP از مقبولیت بیشتری برخوردار می‌باشد (Goddard, 2009; Goddard & Hayes, 2009; Gianola *et al*., 2014).

تأثیرپذیری صحت پیش‌بینی‌های ژنومی از ارتباط خویشاوندی بین افراد جمعیت‌های تائید (حاوی حیوانات جوان فاقد رکورد) و مرجع کاملاً پذیرفته شده، به‌طوری‌که با کاهش نسبت افراد دارای روابط خویشاوندی زیاد بین جمعیت‌های مذکور صحت پیش‌بینی‌ها کاهش می‌یابد (Habier *et al*., 2007; Habier *et al*., 2010). روش GBLUP نیز مانند دیگر روش‌های پارامتری از این قاعده مستثنی نبوده و نتایج آن تحت تأثیر شدت روابط خویشاوندی بین جمعیت‌های مرجع و تائید قرار دارد. یک راهکار مناسب جهت افزایش ثبات صحت پیش‌بینی آثار تصادفی، استفاده از تکنیک نمونه‌گیری تجمعی بوت‌استرپ[4] است (Breiman, 1996). تکنیک مذکور که به اختصار بگینگ[5] خوانده می‌شود، با استفاده یک قاعده سادهٔ بازنمونه‌گیری می‌تواند صحت پیش‌بینی‌ها

1. Single nucleotide polymorphism
2. Linkage disequilibrium
3. Quantitative trait loci
4. Bootstrap aggregation sampling
5. Bagging

را در شرایطی که نمونه‌گیری از مجموعه داده‌های آموزشی منجر به افزایش واریانس پیش‌بینی‌گر شود بهبود بخشد (Breiman, 1996). این تکنیک آماری، در واقع فارغ از بسیاری فرضیات با ایجاد N نمونهٔ مستقل شرایط نمونه را به شرایط جامعه نزدیک و با در نظر گرفتن بسیاری از حالات مختلف تشکیل نمونه، صحت برآورد و حدود اطمینان ضرایب را افزایش می‌دهد (Efron & Tibshirani, 1993). اساس تکنیک بگینگ بدین صورت است که با استفاده از نمونه‌گیری همراه با جایگزینی n نمونهٔ تصادفی (با شانس یکسان برای انتخاب همهٔ نمونه‌ها)، یک مجموعهٔ جدید بوت‌استرپ با اندازه یکسان از مجموعهٔ اصلی استخراج و از آن برای آموزش یک مدل آماری استفاده می‌کند. این رویکرد را N بار تکرار کرده و در نهایت با استفاده از میانگین پیش‌بینی‌کننده‌های مجموعه‌های بوت‌استرپ مختلف، پیش‌بینی نهایی انجام می‌شود که نسبت به هر یک از پیش‌بینی‌گرهای مجموعه‌های مختلف واریانس پیش‌بینی و میانگین مربعات خطای کمتری دارد (Breiman, 1996). Gianola *et al.* (2014) نخستین بار در حوزهٔ انتخاب ژنومی این الگوریتم را استفاده کردند و نشان دادند که برای افزایش ثبات پیش‌بینی‌ها، تعریف ۲۵ تا ۵۰ نمونهٔ (N) بوت‌استرپ کافی است. اما در دیگر پژوهش‌های انجام شده روی اطلاعات حقیقی ژنومی، نتایج متفاوتی از به‌کارگیری این تکنیک آماری ارائه شده است. به‌طوری‌که در تحقیقات انجام شده روی صفات مربوط به رشد و تولید تخم‌مرغ جوجه‌های گوشتی (Abdollahi-Arpanahi *et al.*, 2015) و صفات تولیدی (پروتئین تولیدی) یا تولیدمثلی (شمار سلول‌های بدنی و نرخ آبستنی) گاوهای شیرده نژاد جرسی و هلشتاین آمریکایی (Mikshowsky *et al.*, 2016; Mikshowsky *et al.*, 2017)، به ترتیب موفقیت و عدم موفقیت تکنیک بگینگ در افزایش صحت پیش‌بینی‌های ژنومی روش GBLUP گزارش شده است. لذا پژوهش حاضر با شبیه‌سازی صفات با معماری‌های متفاوت ژنتیکی، به ارزیابی کارایی تکنیک بگینگ روی صحت پیش‌بینی ارزش‌های اصلاحی ژنومی روش پارامتری GBLUP می‌پردازد.

مواد و روش‌ها

شبیه‌سازی ژنوم و جمعیت

در پژوهش حاضر با استفاده از بستهٔ نرم‌افزاری *hypred* (Technow, 2013) ژنومی متشکل از ۱۰ کروموزوم با طول یکسان ۱۰۰ سانتی‌مورگان شبیه‌سازی شد. سپس روی هریک از این کروموزوم‌ها ۱۰۰۰ نشانگر SNP با فراوانی اولیه ۰/۵ و با فواصل نشانگری یکسان توزیع گردید. به منظور بررسی صفات با معماری ژنتیکی متفاوت، در مورد اثر جایگزینی QTLها از سه فرم یکنواخت (حداقل صفر و حداکثر یک)، گاما (با پارامتر شکل ۰/۴۰ و پارامتر مقیاس ۰/۶۰) و نرمال استاندارد (میانگین صفر و واریانس یک) استفاده شد و برای مقدار توارث‌پذیری صفت نیز چهار سطح مختلف (۰/۰۵، ۰/۲۵، ۰/۴۵ و ۰/۶۵) تعریف گردید. در این پژوهش، تمام صفات با فرض تأثیرپذیری از تنها ۱۰۰ QTL تعریف شدند. در واقع این QTLها همهٔ واریانس ژنتیکی صفات را توجیه می‌کردند، به‌طوری‌که ارزش اصلاحی حقیقی افراد با توجه به ژنوتیپ آنها از مجموع اثرات QTLها محاسبه می‌شد. در این پژوهش از وجود اثرات غیر افزایشی (اثرات متقابل درون و بین جایگاهی) چشم‌پوشی شده و واریانس ژنتیکی کل معادل واریانس ژنتیکی افزایشی بود.

در مورد تمام سناریوهای این پژوهش و به منظور ایجاد عدم تعادل پیوستگی بین نشانگر و QTL، به یک جمعیت پایه با اندازهٔ مؤثر ۱۰۰ فرد (۵۰ نر و ۵۰ ماده) اجازه داده شد تا در آن برای ۱۰۰ نسل متوالی آمیزش تصادفی صورت گیرد. با فرض اینکه از هر جفت والد تنها دو فرزند حاصل گردد، اندازهٔ جمعیت در هر نسل ثابت و معادل ۱۰۰ فرد باقی ماند. به منظور بررسی تأثیرپذیری صحت پیش‌بینی‌ها از تعداد مشاهدات نیز دو حجم متفاوت برای تعداد افراد موجود در جمعیت مرجع منظور گردید، به‌طوری‌که افراد نسل ۱۰۱ به ۱۰۰۰ یا ۲۰۰۰ فرد و نسل‌های ۱۰۲ تا ۱۰۵ با نمونه‌گیری تصادفی از نسل قبل خود (بدون محدودیت تعداد فرزند به ازای هر والد) به ۲۰۰۰ فرد گسترش یافتند. افراد نسل ۱۰۱ که همهٔ اطلاعات فنوتیپی (همواره با یک توزیع نرمال) و ژنوتیپی را داشتند به عنوان جمعیت مرجع و نسل‌های

بعدی که فقط دارای اطلاعات ژنوتیپی بودند به عنوان جمعیت‌های تائید در نظر گرفته شدند. جهت اطمینان از عدم تصادفی بودن نتایج، برای تمام سناریوهای این پژوهش ۱۰ بار تکرار در نظر گرفته شد به‌طوری‌که در هریک از این تکرارها و قبل از برآورد ضرایب رگرسیونی نشانگرهایی با فرآوانی آللی نامطلوب (کمتر از ۲ یا بیشتر از ۹۸ درصد) از ماتریس مربوط به نشانگرهای ژنومی حذف شدند.

مدل آماری

GBLUP

با به‌کارگیری نرم‌افزار آماری R (نسخهٔ ۳,۴,۰) برای حل معادلات ماتریسی به روش GBLUP، که معادلهٔ آماری آن در رابطهٔ ۱ ارائه شده، ارزش اصلاحی ژنومی افراد جمعیت مرجع محاسبه شد.

$$\boldsymbol{y_i = 1\mu + \sum_{p=1}^{p} X_{ip}\beta_p + e_i} \qquad (۱)$$

در این رابطه yi فنوتیپ iاُمین فرد جمعیت مرجع، ۱ بردار واحد، μ میانگین کل (عرض از مبداء)، Xip ماتریس ضرایب ارتباط دهندهٔ مشاهدات به آثار تصادفی، β p اثر تصادفی pاُمین نشانگر و ei بردار اثرات تصادفی مانده‌ها است. فرض اساسی روش GBLUP، که مبتنی بر بهترین پیش‌بینی نااُریب خطی است (Henderson, 1975)، کوچک بودن اثرات نشانگری و یکسان بودن سهم همهٔ آنها در توجیه واریانس ژنتیکی (نرمال بودن توزیع اثرات نشانگری) است (Meuwissen *et al*., 2001; Hayes *et al*., 2009). بستهٔ نرم افزاری استفاده شده در شبیه‌سازی این پژوهش (hypred) به گونه‌ای نوشته شده که عناصر ماتریس X را برحسب ژنوتیپ فرد چنانچه برای یک نشانگر خاص هموزیگوت مغلوب (aa)، هتروزیگوت (Aa یا aA) یا هموزیگوت غالب (AA) باشد به ترتیب به صورت صفر، یک و دو کُدگذاری می‌کند (Hayes *et al*., 2009; Technow, 2013).

Bagging GBLUP (BGBLUP)

عموماً برای ایجاد یک مجموعهٔ بوت‌استرپ، به‌طور تصادفی و همراه با جایگزینی تمام اطلاعات مربوط به یکی از افراد جمعیت مرجع استخراج و این پروسه تا زمان یکسان شدن تعداد مشاهدات مجموعهٔ جدید با جمعیت مرجع ادامه می‌یابد. به دلیل استفاده از رویکرد بازنمونه‌گیری، به عبارت بهتر دادن شانس انتخاب مجدد به هر نمونه از جمعیت اصلی، ممکن است برخی از نمونه‌های جمعیت مرجع چندین بار در یک مجموعهٔ بوت‌استرپ ظاهر شوند اما برخی دیگر اصلاً انتخاب نشوند لذا علیرغم یکسان بودن تعداد مشاهدات، هر مجموعهٔ بوت‌استرپ از دیگری متفاوت خواهد بود (Efron & Tibshirani, 1993). پس از استخراج هر مجموعهٔ بوت‌استرپ، با مدل کردن روابط بین متغیر وابسته (y) و متغیرهای کمکی (X) مجموعهٔ مذکور (با استفاده از رابطهٔ ۱) ضرایب رگرسیونی اثرات نشانگری برآورد و با به‌کارگیری این ضرایب ارزش اصلاحی ژنومی تمام افراد جمعیت مرجع محاسبه می‌شود (Gianola *et al*., 2014; Abdollahi-Arpanahi *et al*., 2015). در پژوهش حاضر به منظور افزایش ثبات پیش‌بینی‌ها، با استفاده از ۵۰ مجموعهٔ بوت‌استرپ ارزش اصلاحی ژنومی جمعیت مرجع پیش‌بینی و در نهایت با میانگین‌گیری از همهٔ آنها یک ارزش اصلاحی ژنومی برای هریک از افراد جمعیت مرجع محاسبه و از آن جهت برآورد ضرایب نهایی نشانگرهای جمعیت مرجع استفاده شد (Abdollahi-Arpanahi *et al*., 2015).

پس از حل معادلات مذکور و برآورد اثرات نشانگری (β) جمعیت مرجع، ارزش اصلاحی ژنومی (GEBV) iاُمین فرد گروه‌های تائید از مجموع آثار نشانگری و با توجه به ژنوتیپ (X) آنها به صورت رابطهٔ ۲ محاسبه گردید (Hayes *et al*., 2009):

$$\boldsymbol{GEBV_i = \sum_{p=1}^{p} X_{ip}\hat{\beta}_p} \qquad (۲)$$

در این پژوهش جهت مقایسهٔ صحت ارزیابی‌های ژنومی هریک از توزیعات متفاوت آثار جایگزینی ژنی، از همبستگی پیرسون بین ارزش‌های اصلاحی حقیقی و ژنومی محاسبه شده (Correlation) و میانگین مربعات خطای (MSE) بین آنها استفاده گردید. به دلیل متفاوت بودن واریانس ارزش‌های اصلاحی ژنومی توزیع‌های

مختلف آثار ژنی، برای مقایسهٔ بین آنها تنها از معیار Correlation استفاده گردید. مقایسهٔ روش‌های آماری نیز با استفاده از آزمون t صورت گرفت.

نتایج و بحث

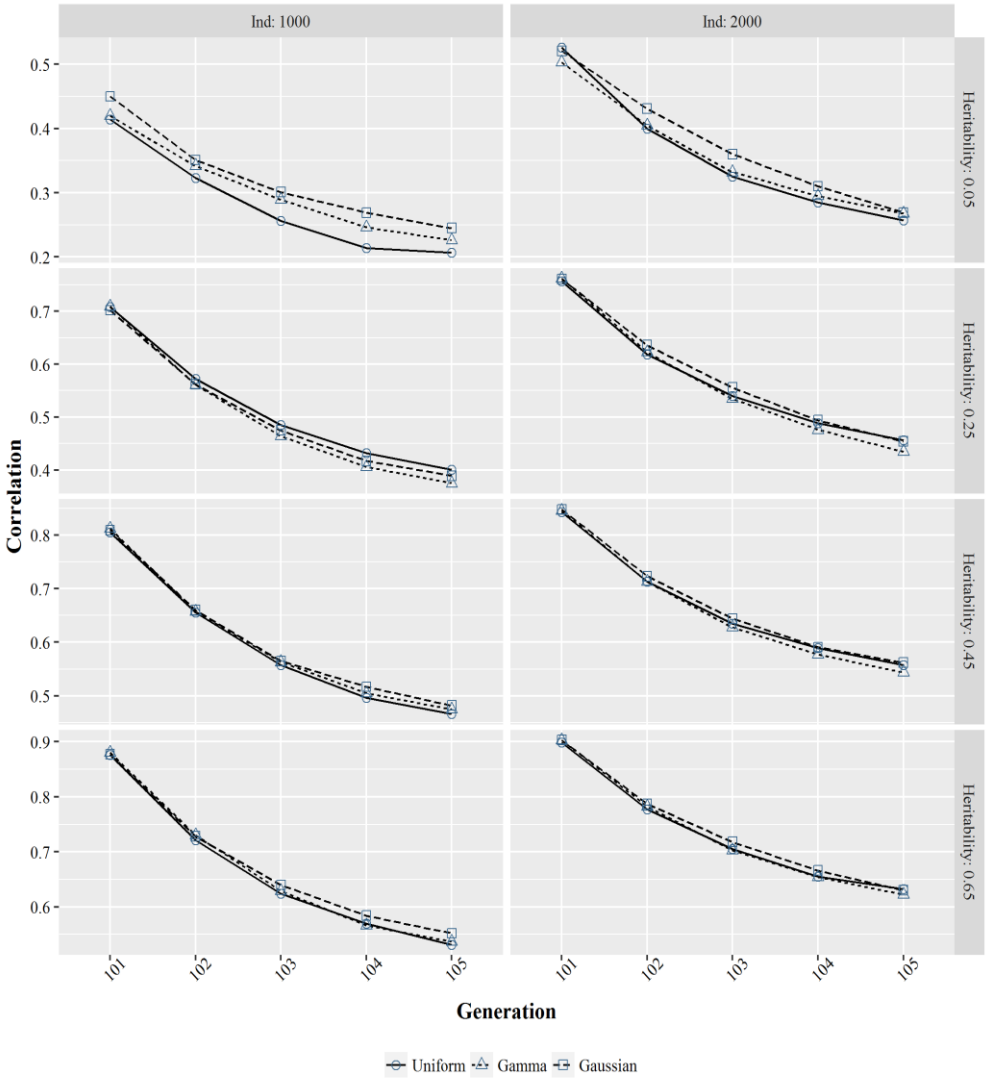
در مورد هر سه فُرم توزیع آثار جایگزینی ژن‌ها بعد از ۱۰۰ نسل آمیزش تصادفی، به دلیل ثابت ماندن اندازهٔ جمعیت میانگین آمارهٔ عدم تعادل پیوستگیr2 (Hayes *et al*., 2009) به حدود ۲۰ درصد رسید. آمارهٔ مذکور در واقع نشان‌دهندهٔ سهمی از واریانس QTLهاست که به وسیلهٔ نشانگرها توجیه می‌شود (Goddard & Hayes, 2007)، که سطح لازم آن برای موفقیت برنامه‌های انتخاب ژنومی حدود ۲۰ درصد می‌باشد (Meuwissen *et al*., 2001).

براساس نتایج پژوهش حاضر، به‌طورکلی صحت ارزیابی‌های ژنومی به ترتیب تحت اثر میزان توارث‌پذیری صفت، فاصلهٔ نسل از جمعیت مرجع، تعداد مشاهدات جمعیت مرجع و اثر جایگزینی QTLها قرار دارد (براساس مقدار F جدول آنالیز واریانس؛ نتایج نشان داده نشده است). به‌طوری‌که در هر سه فُرم توزیع آثار جایگزینی ژنی و بدون توجه به مدل آماری، میزان صحت پیش‌بینی‌های ژنومی با افزایش توارث‌پذیری صفت یا تعداد افراد جمعیت مرجع به طور معنی‌داری (p<۰/۰۰۱) بهبود یافت اما با افزایش فاصله از جمعیت مرجع به وضوح (p<۰/۰۰۱) رو به زوال بود (شکل ۱؛ به دلیل تنوع ناشی از تکرار ارزیابی‌ها، میانگین معیار ارزیابی صحت پیش‌بینی‌های ژنومی ارائه شده است).

در یک مطالعهٔ مروری، میزان LD بین SNPها و QTLها، تعداد نمونه‌های جمعیت مرجع، سطح توارث‌پذیری صفت و توزیع آثار جایگزینی ژنها به عنوان فاکتورهای مؤثر بر صحت ارزیابی‌های ژنومی مطرح شده‌اند (Hayes *et al*., 2009). در واقع اُفت صحت ارزیابی‌های ژنومی با کاهش توارث‌پذیری صفت یا افزایش فاصله از جمعیت مرجع، به ترتیب ناشی از افزایش واریانس فنوتیپی صفت (Lund *et al*., 2009) و برهم خوردن فاز LD بین نشانگرها و جایگاه‌های ژنی به‌دلیل وقوع نوترکیبی (Habier *et al*., 2007) است. به‌هرحال، در تائید نتایج پژوهش‌های پیشین مشاهده می‌شود که رابطهٔ میزان کاهش صحت ارزیابی‌ها با افزایش فاصلهٔ بین نسل‌ها به صورت غیرخطی بوده و بیشترین نرخ کاهش (نسبت به جمعیت قبلی) مربوط به نسل اول جمعیت تائید می‌باشد (Habier *et al*., 2007)؛ که با افزایش توارث‌پذیری صفت یا تعداد نمونه‌های جمعیت مرجع از شیب کاهشی آن نیز کاسته شد (p<۰/۰۱). در مورد تعداد نمونه‌های جمعیت مرجع نیز مشاهده شد که تأثیر افزایش تعداد آنها روی صحت ارزیابی‌های ژنومی بسته به معماری ژنتیکی صفت متفاوت بوده و با افزایش توارث‌پذیری، از اهمیت آن کاسته می‌شود (شکل ۱). به‌طوری‌که در مورد نسل نخُست جمعیت تائید (نسل ۱۰۲) و بدون توجه به توزیع QTLها یا مدل آماری، با افزایش تعداد مشاهدات جمعیت مرجع از ۱۰۰۰ به ۲۰۰۰ فرد برای صفات با توارث‌پذیری ۵، ۲۵، ۴۵ یا ۶۵ درصد به ترتیب ۲۲ (۰/۰۴۷±۰/۳۳۹ در مقایسه با ۰/۰۴۱±۰/۴۱۲)، ۱۱ (۰/۰۳۶±۰/۵۶۵ در مقایسه با ۰/۰۲۳±۰/۶۲۶)، ۹ (۰/۰۱۹±۰/۶۵۷ در مقایسه با ۰/۰۱۶±۰/۷۱۶) و ۸ (۰/۰۱۵±۰/۷۲۷ در مقایسه با ۰/۰۱۵±۰/۷۸۲) درصد بهبود ضریب همبستگی ارزیابی‌های ژنومی حاصل گردید. در مطالعهٔ شبیه‌سازی یک صفت پیوسته با توارث‌پذیری ۵۰ درصد، صحت ارزیابی‌های ژنومی روش BLUP در سناریوهای متفاوت ۵۰۰، ۱۰۰۰ یا ۲۲۰۰ نمونهٔ جمعیت مرجع به ترتیب ۰/۵۷۹، ۰/۶۵۹ و ۰/۷۳۲ گزارش شده است (Meuwissen *et al*., 2001).

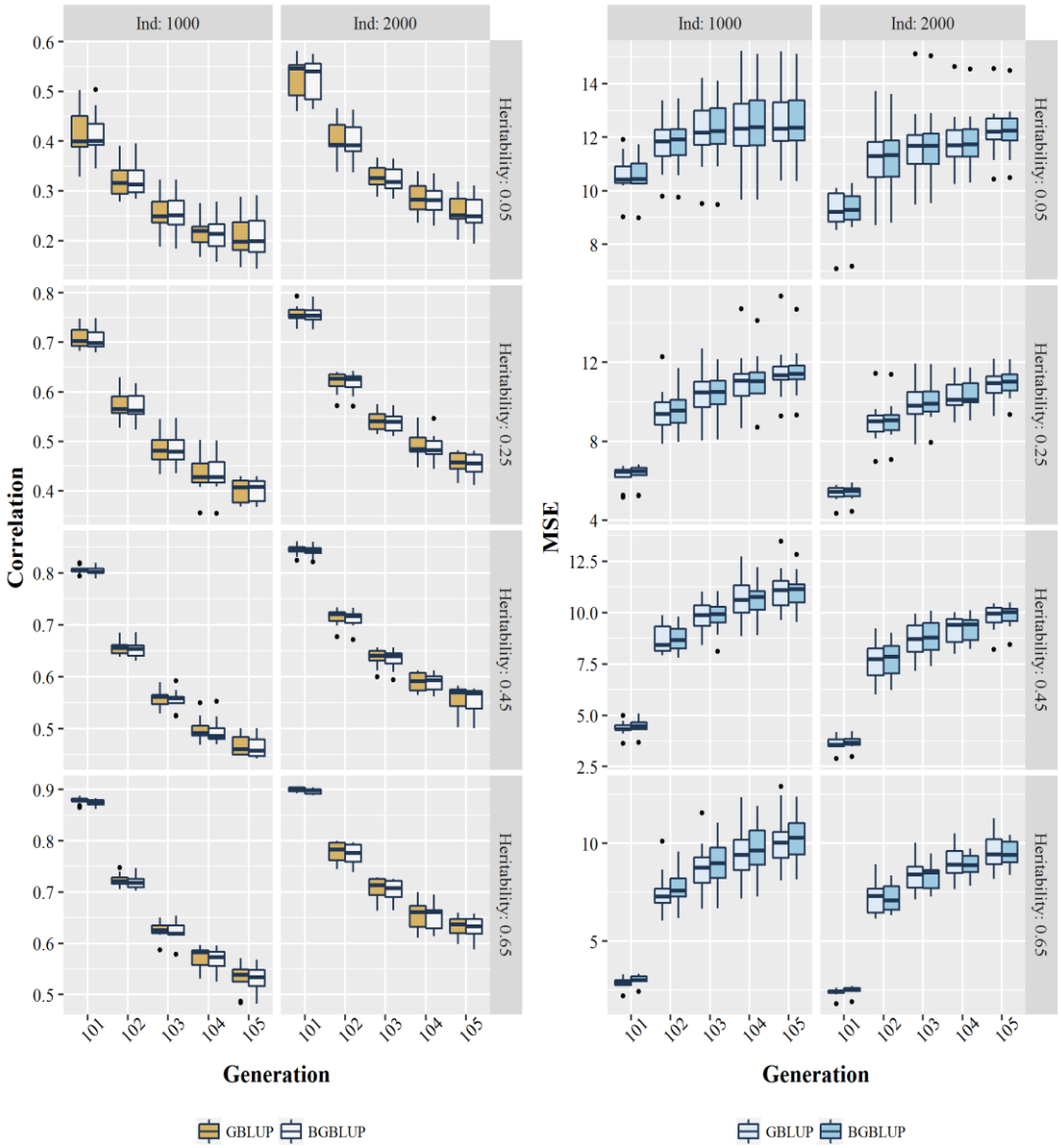
به دلیل تنوع ناشی از تکرار ارزیابی‌ها، نتایج پیش‌بینی‌های ژنومی روش‌های آماری استفاده شده به صورت نمودار جعبه‌ای و به تفکیک توزیع آثار ژنی ارائه شده است (شکل‌های ۲ تا ۴). در مورد تمام سناریوها و معماری‌های متفاوت ژنتیکی، مشاهده شد که علیرغم ارجح بودن صحت پیش‌بینی‌های ژنومی روش GBLUP بر BGBLUP تفاوت آماری معنی‌داری بین آنها وجود ندارد (p>۰/۰۵). به‌طوری‌که در مورد نسل اول جمعیت تائید (نسل ۱۰۲) و بدون توجه به توزیع آثار جایگزینی ژنها، با جمعیتی معادل ۱۰۰۰ (یا ۲۰۰۰) فرد در جمعیت مرجع صحت ارزش‌های

اصلاحی ژنومی روش GBLUP برای صفات با توارث‌پذیری ۵، ۲۵، ۴۵ و ۶۵ درصد به ترتیب ۰/۳۳۹±۰/۰۴۹ (۰/۴۱۲±۰/۰۴۲)، ۰/۵۶۶±۰/۰۳۵ (۰/۶۲۷±۰/۰۲۳)، ۰/۶۵۸±۰/۰۱۹ (۰/۷۱۸±۰/۰۱۶) و ۰/۷۲۸±۰/۰۱۵ (۰/۷۸۳±۰/۰۱۵) بود و مقادیر مشابه برای روش BGBLUP به‌ترتیب ۰/۳۳۸±۰/۰۴۷ (۰/۴۱۱±۰/۰۴۲)، ۰/۵۶۴±۰/۰۳۷ (۰/۶۲۵±۰/۰۲۳)، ۰/۶۵۶±۰/۰۱۹ (۰/۷۱۵±۰/۰۱۷) و ۰/۷۲۵±۰/۰۱۶ (۰/۷۸۰±۰/۰۱۵) محاسبه گردید. اما با مقایسهٔ میزان خطای روش‌های مذکور (میانگین مربعات خطا)، مشاهده می‌شود که کارایی تکنیک بگینگ بستگی به معماری ژنتیکی صفات و ساختارهای جمعیتی داشته و از حیث عددی ($p>0/05$) در برخی حالات (با افزایش تعداد نمونه‌های جمعیت مرجع و توارث‌پذیری صفت) منجر به بهبود صحت نتایج روش GBLUP خواهد شد (MSE؛ شکل‌های ۲ تا ۴).



شکل ۱. روند تغییر ضریب همبستگی (Correlation) ارزش‌های اصلاحی ژنومی توزیع متفاوت آثار ژنی (Uniform, Gamma, Gaussian) بین جمعیت‌های مرجع (نسل ۱۰۱) و تائید (نسل‌های ۱۰۲ تا ۱۰۵)، برای صفات با توارث‌پذیری مختلف (ردیفی از بالا به پائین به ترتیب: ۵، ۲۵، ۴۵ یا ۶۵ درصد) و تعداد متفاوت مشاهدات در جمعیت مرجع (ستونی از چپ به راست به ترتیب: ۱۰۰۰ یا ۲۰۰۰ فرد).

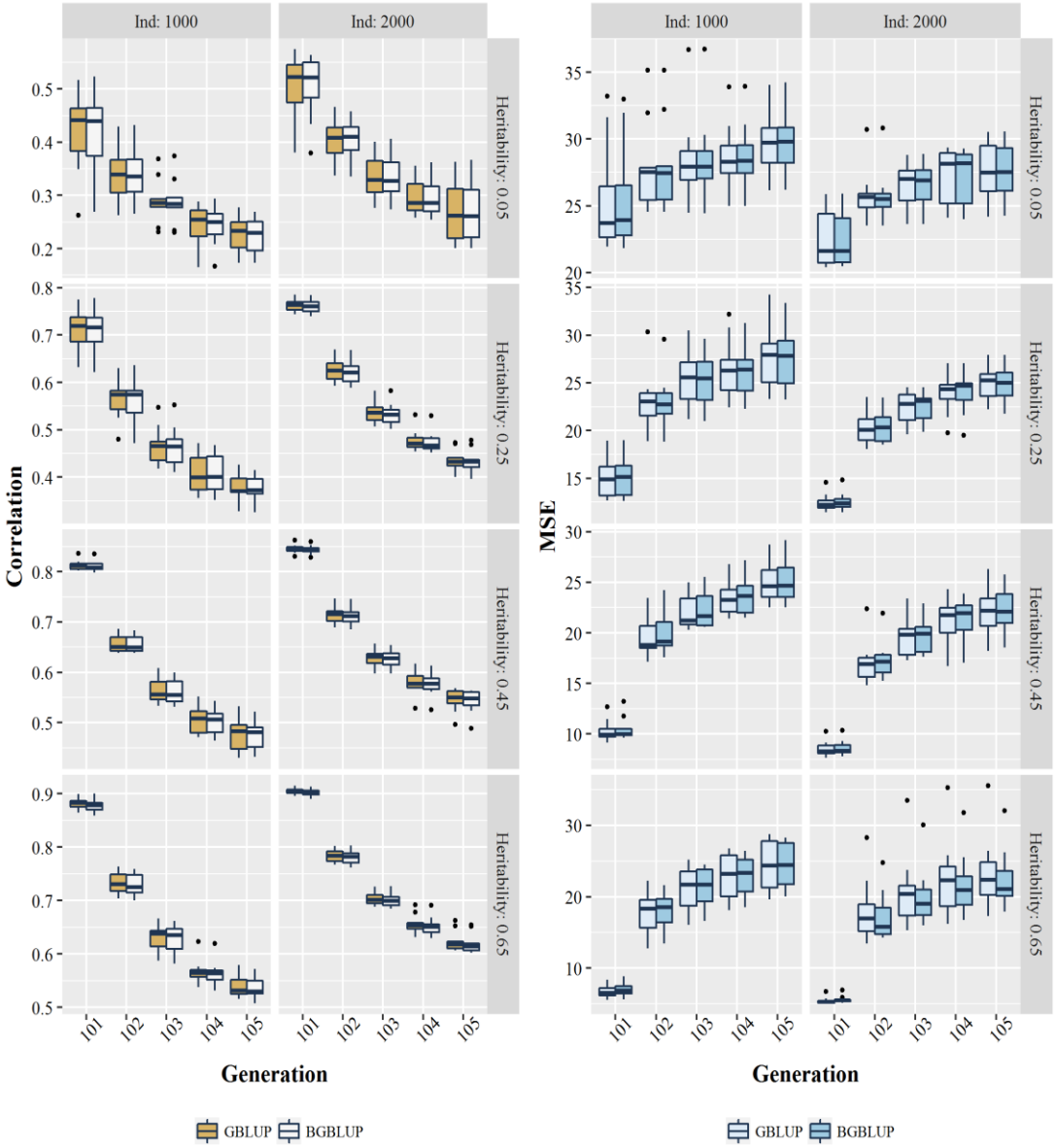
Figure 1. Correlation coefficient (Correlation) of genomic breeding values for different distribution of gene effects (Uniform, Gamma or Gaussian) in training (101 generation) and testing (102 to 105 generations) sets for traits with different heritability (top-to-bottom row respectively, 0.05, 0.25, 0.45 or 0.65) and different number of observations in the reference set (left-to-right respectively, 1000 or 2000).



شکل ۲. روند تغییر ضریب همبستگی (Correlation) و میانگین مربعات خطای (MSE) ارزش‌های اصلاحی ژنومی روش‌های GBLUP و BGBLUP بین جمعیت‌های مرجع (نسل ۱۰۱) و تائید (نسل‌های ۱۰۲ تا ۱۰۵)، با فرض یکنواخت بودن توزیع تأثیر ژنی برای صفات با توارث‌پذیری مختلف (ردیفی از بالا به پائین به ترتیب: ۵، ۲۵، ۴۵ یا ۶۵ درصد) و تعداد متفاوت مشاهدات در جمعیت مرجع (ستونی از چپ به راست به ترتیب: ۱۰۰۰ یا ۲۰۰۰ فرد). •: داده‌های پرت.

Figure 2. Correlation coefficient (Correlation) and mean square error (MSE) of genomic breeding values for GBLUP and BGBLUP methods in training (101 generation) and testing (102 to 105 generations) sets, assuming uniform distribution for traits with different heritability (top-to-bottom row respectively, 0.05, 0.25, 0.45 or 0.65) and different number of observations in the reference set (left-to-right respectively, 1000 or 2000). Outliers denoted as black dots.

به دلیل تأثیرگذاری هزینه‌های اجرایی بر میزان مقبولیت هر روش یا مدل آماری، در شکل ۵ مدت زمان محاسباتی (نتایج مربوط به یک کامپیوتر تحت سیستم عامل ویندوز با ۱۶ گیگابایت حافظهٔ رم) برازش هر دو روش آماری ارائه شده است. براساس این نتایج، تکنیک بگینگ به وضوح هزینهٔ محاسباتی روش GBLUP را افزایش داده و با افزایش تعداد مشاهدات جمعیت مرجع تأثیر آن شدت می‌یابد ($p<۰/۰۰۵$). به‌طوری‌که با دو برابر شدن تعداد نمونه‌های جمعیت مرجع (از ۱۰۰۰ به ۲۰۰۰ فرد)، متوسط زمان برآورد آثار نشانگری روش BGBLUP در هر تکرار حدود ۱۰ برابر (۸۳ در مقایسه با ۸۱۶ ثانیه) افزایش یافت ($p<۰/۰۰۵$).

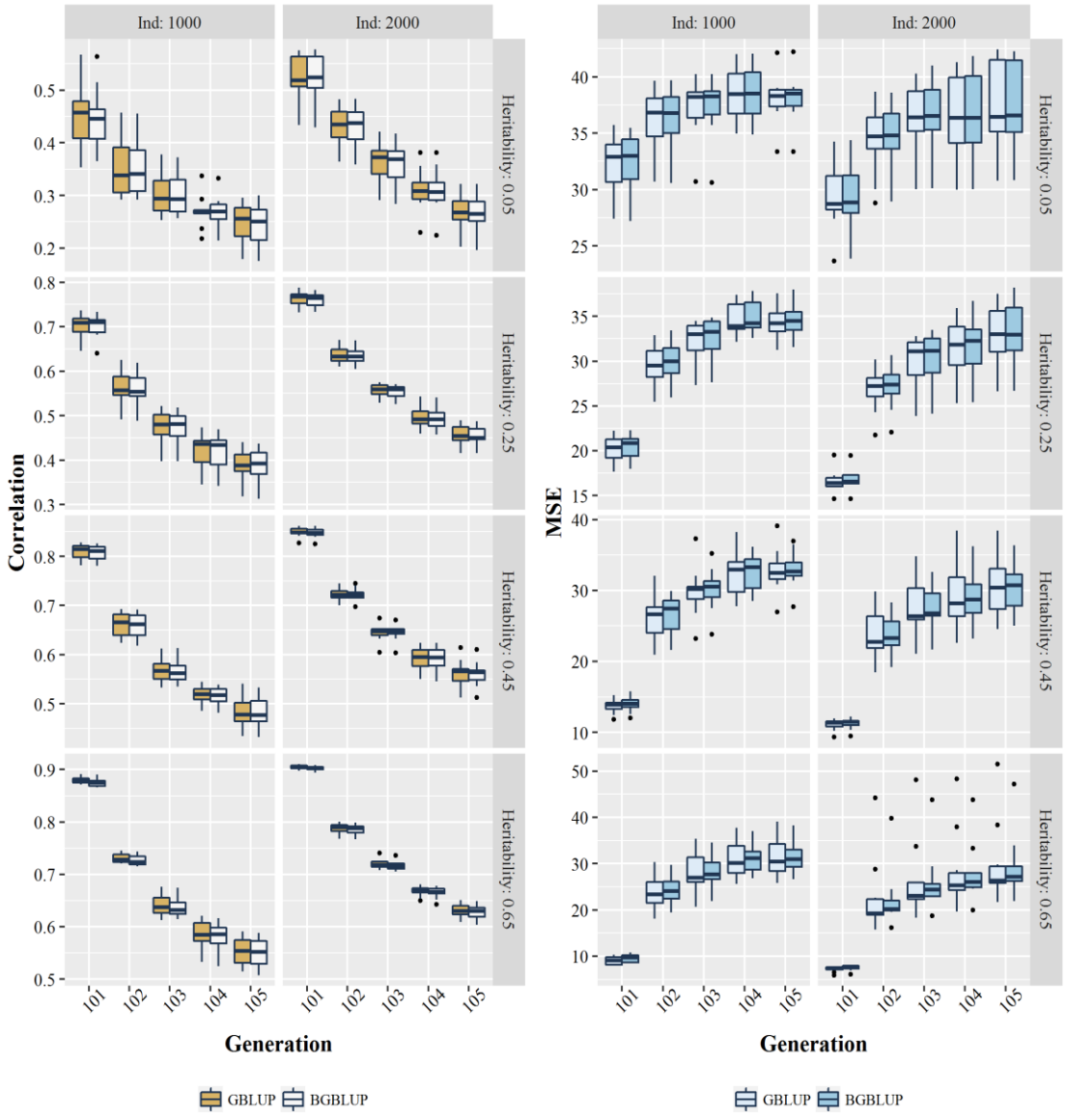


شکل ۳. روند تغییر ضریب همبستگی (Correlation) و میانگین مربعات خطای (MSE) ارزش‌های اصلاحی ژنومی روش‌های GBLUP و BGBLUP بین جمعیت‌های مرجع (نسل ۱۰۱) و تائید (نسل‌های ۱۰۲ تا ۱۰۵)، با فرض گاما بودن توزیع تأثیر ژنی برای صفات با توارث‌پذیری مختلف (ردیفی از بالا به پائین به ترتیب: ۵، ۲۵، ۴۵ یا ۶۵ درصد) و تعداد متفاوت مشاهدات در جمعیت مرجع (ستونی از چپ به راست به ترتیب: ۱۰۰۰ یا ۲۰۰۰ فرد). •: داده‌های پرت.

Figure 3. Correlation coefficient (Correlation) and mean square error (MSE) of genomic breeding values for GBLUP and BGBLUP methods in training (101 generation) and testing (102 to 105 generations) sets, assuming gamma distribution for traits with different heritability (top-to-bottom row respectively, 0.05, 0.25, 0.45 or 0.65) and different number of observations in the reference set (left-to-right respectively, 1000 or 2000). Outliers denoted as black dots.

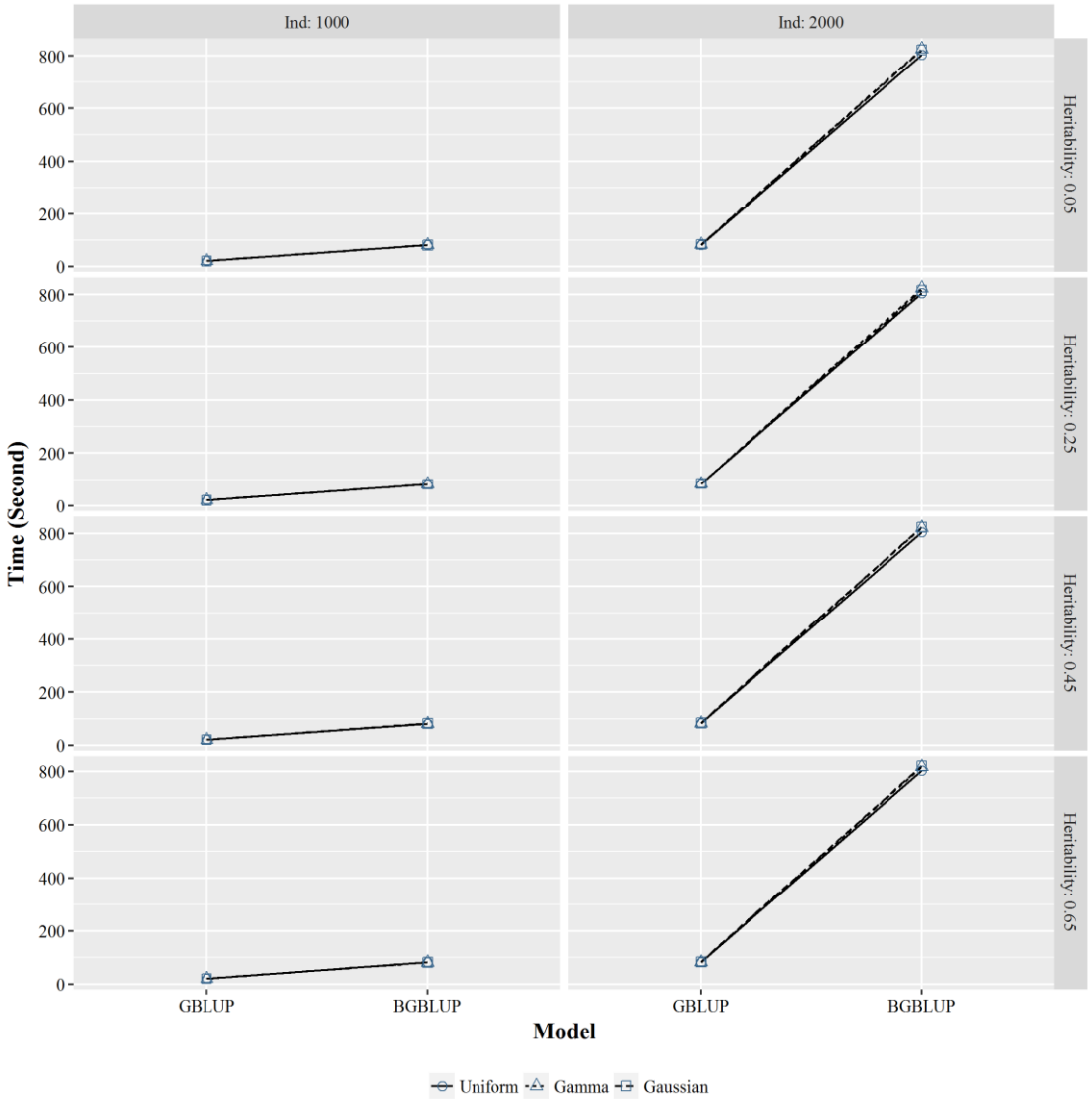
موفقیت نسبی تلفیق تکنیک Bagging در بهبود صحت پیش‌بینی‌های ژنومی روش GBLUP برای صفات رشد، لاشه و تولید تخم‌مرغ جوجه‌های گوشتی قبلاً گزارش شده است (Abdollahi-Arpanahi *et al.*, 2015). اما در مورد صفات تولیدی پروتئین و نرخ آبستنی (با توارث‌پذیری به ترتیب حدود ۳۰ و ۴ درصد؛ García-Ruiz *et al.*, 2016) گاوهای شیرده جرسی یا هلشتاین آمریکایی، موافق با نتایج این پژوهش، گزارش شده که صحت بدست آمدهٔ روش GBLUP روی تمام نمونه‌های جمعیت مرجع بهتر از

بیشینه صحت روش BGBLUP برای هر مجموعهٔ بوتاسترپ بوده است (Mikshowsky *et al.*, 2016; Mikshowsky *et al.*, 2017). به‌طوری‌که در مورد صفت پروتئین و نرخ آبستنی نژاد جرسی، برتری صحت پیش‌بینی‌های ژنومی روش GBLUP نسبت به BGBLUP به ترتیب ۲ (۰/۶۶۰ در مقایسه با ۰/۶۴۶) و ۴ (۰/۴۰۰ در مقایسه با ۰/۳۸۳) درصد بیشتر بوده است (Mikshowsky *et al.*, 2016). در مورد نژاد هلشتاین نیز مقادیر مشابه به ترتیب ۴ (۰/۶۹۰ در مقایسه با ۰/۶۶۵) و ۱۲ (۰/۵۵۷ در مقایسه با ۰/۴۹۹) درصد گزارش شده است (Mikshowsky *et al.*, 2017).



شکل ۴. روند تغییر ضریب همبستگی (Correlation) و میانگین مربعات خطای (MSE) ارزش‌های اصلاحی ژنومی روش‌های GBLUP و BGBLUP بین جمعیت‌های مرجع (نسل ۱۰۱) و تائید (نسل‌های ۱۰۲ تا ۱۰۵)، با فرض نرمال بودن توزیع تأثیر ژنی برای صفات با توارث‌پذیری مختلف (ردیفی از بالا به پائین به ترتیب: ۵، ۲۵، ۴۵ یا ۶۵ درصد) و تعداد متفاوت مشاهدات در جمعیت مرجع (ستونی از چپ به راست به ترتیب: ۱۰۰۰ یا ۲۰۰۰ فرد). •: داده‌های پرت.

Figure 4. Correlation coefficient (Correlation) and mean square error (MSE) of genomic breeding values for GBLUP and BGBLUP methods in training (101 generation) and testing (102 to 105 generations) sets, assuming normal distribution for traits with different heritability (top-to-bottom row respectively, 0.05, 0.25, 0.45 or 0.65) and different number of observations in the reference set (left-to-right respectively, 1000 or 2000). Outliers denoted as black dots.



شکل ۵. متوسط هزینه محاسباتی (به ثانیه) روش‌های GBLUP و BGBLUP در هر تکرار برآورد اثرات نشانگری جمعیت مرجع برای صفات با توارث‌پذیری‌های مختلف (ردیفی از بالا به پائین به ترتیب: ۵، ۲۵، ۴۵ یا ۶۵ درصد) و توزیعات متفاوت آثار ژنی (Uniform, Gamma, Gaussian).

Figure 5. The average computational time (in second) of GBLUP and BGBLUP methods in each replicate of marker effect estimates for traits with different heritability (top-to-bottom respectively, 0.05, 0.25, 0.45 or 0.65) and different distribution of gene effects (Uniform, Gamma, Gaussian).

با پذیرش ضرورت ناپایدار بودن پیش‌بینی‌های یک روش آماری (بالا بودن واریانس برآورد اثرات متغیرهای مستقل) جهت موفقیت تکنیک Bagging در بهبود صحت پیش‌بینی‌های آن (Breiman, 1996)، می‌توان نتیجه گرفت که پیش‌بینی‌های روش GBLUP پایدار بوده و ارزش‌های اصلاحی ژنومی را با حداقل تنوع پیش‌بینی می‌کند (Mikshowsky ***et al***., 2016). لذا از یک‌سو به دلیل ناکارآمد بودن تکنیک Bagging در بهبود صحت پیش‌بینی‌های ژنومی روشGBLUP و از سوی دیگر افزایش چشم‌گیر زمان محاسباتی در برآورد آثار نشانگری، می‌توان نتیجه گرفت که حداقل در مطالعاتی با تفاوت کم بین تعداد مشاهدات و مجهولات جمعیت مرجع (در پژوهش حاضر نسبت‌های ۰/۱ و ۰/۲ تعریف شده است) استفاده از روش GBLUP ارجحیت داشته و ضرورتی برای به‌کارگیری تکنیک Bagging وجود ندارد.

نتیجه‌گیری کلی

روش پارامتری GBLUP قادر به پیش‌بینی ارزش‌های اصلاحی ژنومی با حداقل تنوع بوده و تکنیک Bagging علیرغم عدم بهبود صحت ارزیابی‌های ژنومی روش GBLUP، هزینه‌های محاسباتی آنرا نیز به شدت افزایش می‌دهد. به هرحال باید توجه داشت که تغییر نسبت افراد با تعداد بالای فرزند در جمعیت مرجع، به طور معنی‌داری می‌تواند صحت پیش‌بینی‌های ارزش اصلاحی ژنومی کاندیدهای انتخاب را تحت تأثیر قرار دهد (Habier *et al*., 2010; Mikshowsky *et al*., 2016). به‌طوری‌که وجود یا نبود آنها در جمعیت مرجع ممکن است سبب تغییر واریانس پیش‌بینی‌گرها و در نتیجه ناپایدار شدن پیش‌بینی‌های روش GBLUP یا دیگر روش‌های پارامتری (اعم از خطی یا غیرخطی) گردد. از آنجایی‌که در جمعیت‌های حقیقی دام و طیور تفاوت شایان توجهی بین خانواده‌ها از حیث تعداد و وضعیت تنی یا ناتنی بودن فرزندان آنها وجود دارد، لذا انجام مطالعات تکمیلی با در نظر گرفتن نسبت‌های مختلفی از فامیل با تعداد و روابط خویشاوندی متفاوت می‌تواند کارایی Bagging روی صحت ارزیابی‌های روش GBLUP را با اعتماد بیشتری مورد ارزیابی قرار دهد. از طرف دیگر به دلیل فرض ناصحیح این تحقیق (جهت ساده شدن انجام محاسبات آماری) در منظور نکردن دیگر بخش‌های ژنتیکی غیر افزایشی کنترل‌کنندهٔ صفات (اثرات متقابل درون و بین گروهی) و نیز به‌کارگیری صرفاً افراد حاضر در جمعیت مرجع برای تشکیل نسل اول جمعیت تائید، لازم به انجام مطالعات تکمیلی‌تری (با در نظر گرفتن عوامل ژنتیکی غیرافزایشی و استفاده از نسبت‌های مختلف والدین خارج از جمعیت مرجع در تشکیل نسل‌های تائید) است تا با اعتماد بیشتری نتایج آن تائید گردد.

REFERENCES

1. Abdollahi-Arpanahi, R., Morota, G., Valente, B.D., Kranis, A., Rosa, G. J. M. & Gianola, D. (2015). Assessment of bagging GBLUP for whole-genome prediction of broiler chicken traits. *Journal of Animal Breeding and Genetics*, 132, 218-228.
2. Breiman, L. (1996). Bagging predictors. *Machine learning*, 24, 123-140.
3. Efron, B. & Tibshirani, R. J. (1993). An introduction to the bootstrap, New York: *Chapman & Hall*.
4. Gianola, D., Weigel, K. A., Krämer, N., Stella, A. & Schön, C. C. (2014). Enhancing genome-enabled prediction by bagging genomic BLUP. *PLoS One*, 9, e91693.
5. Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetics*, 136, 245-257.
6. Goddard, M. E. & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, 124, 323-330.
7. Goddard, M. E. & Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10, 381-391.
8. García-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-López, F. J. & Van Tassell, C. P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences*, 113, E3995-E4004.
9. Habier, D., Fernando, R. & Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177, 2389-2397.
10. Habier, D., Tetens, J., Seefried, F.R., Lichtner, P. & Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution*, 42, 5.
11. Hayes, B. J., Bowman, P. J., Chamberlain, A. J. & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, 92, 433-443.
12. Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 423-447.
13. Hutchison, J., Cole, J. & Bickhart, D. (2014). Use of young bulls in the United States. *Journal of Dairy Science*, 97, 3213-3220.
14. Lee, S. H., van der Werf, J. H., Hayes, B. J., Goddard, M. E. & Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genetics*, 4, e1000231.

15. Lund, M. S., Sahana, G., de Koning, D. J., Su, G. & Carlborg, Ö. (2009). Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. *BMC proceedings*, *BioMed Central*.
16. Meuwissen, T., Hayes, B. & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157, 1819-1829.
17. Mikshowsky, A. A., Gianola, G. & Weigel, K. A. (2016). Improving reliability of genomic predictions for Jersey sires using bootstrap aggregation sampling. *Journal of Dairy Science*, 99, 3632-3645.
18. Mikshowsky, A. A., Gianola, G. & Weigel, K. A. (2017). Assessing genomic prediction accuracy for Holstein sires using bootstrap aggregation sampling and leave-one-out cross validation. *Journal of Dairy Science*, 100, 453-464.
19. Schaeffer, L. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 123, 218-223.
20. Technow, F. (2013). *R Package hypred: Simulation of genomic data in applied Genetics*. University of Hohenheim.
21. VanRaden, P., Van Tassell, V., Wiggans, G., Sonstegard, T., Schnabel, R., Taylor, J. & Schenkel, F. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science*, 92, 16-24.