تحقیقات تولیدات دامی

سال نهم/شماره دوم/تابستان ۱۳۹۹ (۲۲-۱۵)

Animal Production Research

دانشگاه گیلان

مقاله پژوهشی

اثر تراکم نشانگرها و اندازه جمعیت مرجع بر صحت مستندسازی در داده شبیه‌سازی شده

یحیی محمدی*

استادیار، گروه علوم دامی، دانشکده کشاورزی، دانشگاه ایلام

(تاریخ دریافت: ۹۸/۰۸/۰۸ - تاریخ پذیرش: ۹۸/۱۱/۱۷)

چکیده

در پژوهش حاضر، اثر اندازه جمعیت مرجع و تعداد نشانگرهای چندشکلی تک نوکلئوتیدی (SNP) گم‌شده بر صحت مستندسازی (ایمپیوتیشن) مورد بررسی قرار گرفت. از نرم‌افزار QMSim برای ایجاد بانک اطلاعاتی مرجع به تعداد ۱۰۰۰ حیوان شبیه‌سازی شده استفاده شد. از داده‌های مرجع دو دسته ایجاد شد: دسته اول (A) شامل ژنوتیپ‌های اصلی حاوی داده‌های گم‌شده (تعداد ۵۲ هزار نشانگر SNP) و دسته دوم (B) با خروج داده‌های گم‌شده از مجموع داده‌ها (تعداد ۳۷ هزار نشانگر SNP) ایجاد شد. در هر دو دسته، تعداد جمعیت مرجع با ۱۰۰، ۲۵۰، ۵۰۰ و ۷۵۰ حیوان شبیه‌سازی شد. تعداد نشانگرهای SNP حذف شده به طور تصادفی و با نسبت‌های ۱۵، ۳۰، ۵۵، ۷۰ و ۹۵ درصد در هر دو دسته شبیه‌سازی شد. بر اساس همبستگی بین ارزش نشانگرهای SNP اصلی قبل از حذف و ارزش آن‌ها بعد از مستندسازی، صحت برآورد شد. نتایج مطالعه حاضر نشان داد که صحت مستندسازی تحت تأثیر اندازه جمعیت مرجع و تراکم نشانگرهای SNP گم‌شده قرار داشت. با افزایش اندازه جمعیت مرجع از ۱۰۰ به ۷۵۰ حیوان، متوسط صحت مستندسازی در هر دو دسته افزایش یافت. بیشترین میزان صحت برای جمعیت مرجع با ۷۵۰ حیوان در دامنه ۰/۸۹ تا ۰/۹۸ برای دسته A و ۰/۹۰ تا ۰/۹۹ برای دسته B مشاهده شد. به طور کلی، نتایج نشان داد که اگر اندازه جمعیت مرجع به اندازه کافی باشد، علی‌رغم تعداد زیاد نشانگر SNP گم‌شده، صحت مستندسازی تغییر زیادی نخواهد کرد.

واژه‌های کلیدی: ارزیابی ژنومی، داده‌های گم‌شده، دام، صحت پیش‌بینی، مستندسازی

* نویسندهٔ مسئول: Mohamadi_yahya@yahoo.com

doi: 10.22124/ar.2020.14875.1466

مقدمه

در صنعت پرورش گاو شیری بسیاری از کشورهای دنیا، ارزش اصلاحی ژنومی از میانگین آثار چندشکلی‌های تک نوکلئوتیدی[1] (نشانگر SNP) محاسبه می‌شود (Hayes *et* 2009). در سال‌های اخیر استفاده از فن‌آوری مزبور به دلیل در دسترس قرار گرفتن تراشه‌های نشانگرهای SNP با تراکم بالا، سبب افزایش پیشرفت ژنتیکی در نتیجه کاهش فاصله نسل و افزایش صحت انتخاب حیوانات نر جوان شده است (Schaeffer, 2006). تمایل به استفاده از تراشه‌های با تراکم پایین به دلیل کاهش هزینه‌ها به ویژه برای کشورهای در حال توسعه در حال افزایش است. مستندسازی (ایمپیوتیشن[2]) از نشانگرهای SNP برای تراشه‌های با تراکم پایین به تراش‌های با تراکم بالا، یک راه حل برای کاهش هزینه‌ها و بدست آوردن اطلاعات ژنوتیپی حیوانات است (Zhang and Druet, 2010; Whalen *et al*., 2018). مستندسازی قادر است نقش اصلی در پیش‌بینی ارزش اصلاحی ژنومی داشته باشد. این روش قادر است تعداد زیادی از حیوانات تعیین ژنوتیپ شده را در معادلات تخمین ارزش اصلاحی ژنومی وارد نموده و در نتیجه سبب افزایش صحت پیش‌بینی‌ها شود (VanRaden *et al*., 2011). گزارش شده که با افزایش صحت مستندسازی به صورت خطی صحت پیش‌بینی ارزش اصلاحی ژنومی نیز افزایش پیدا می‌کند (Mulder *et al*., 2012). همچنین گزارش شده است که نشانگرهای SNP با صحت مستندسازی بالا به طور قابل اطمینانی اثر مثبتی در برآورد بالای صحت پیش‌بینی ارزش اصلاحی ژنومی دارند (Mulder *et al*., 2012). در مطالعه Daetwyler *et al*. (2011) گزارش شد که اگر ژنوتیپ حیوانات به طور کامل مستندسازی نشوند، ارزش اصلاحی ژنومی با صحت کمتر برآورد می‌شود. برای مستندسازی نشانگرهای SNP روش‌های متعددی وجود دارند. این گونه روش‌ها بر اساس شجره، تفرق آلل‌ها، عدم تعادل پیوستگی (LD)، اطلاعات جمعیت و ترکیب متقابل آنها قرار دارند. از بین روش‌های موجود، می‌توان روش بگل (Beagle) (Browning *et al*., 2018)، فیندهپ (Findhap) (VanRaden *et al*., 2015) و پدایمپیوت (PedImpute) (Nicolazzi *et al*., 2013) را نام برد. در بعضی موارد که ارتباط خویشاوندی قوی بین افراد وجود ندارد به اطلاعات شجره برای ایجاد مستندسازی نیاز نیست. برای ایجاد مستندسازی در این گونه جوامع، نرم‌افزار Flmpute پیشنهاد می‌شود (Sargolzaei *et al*., 2014). در مستندسازی جمعیت فرض بر این است که بین افراد جمعیت، ارتباطی وجود ندارد و ارتباط خویشاوندی به صورت غیرمستقیم است. تشابه بین افراد به کمک هاپلوتایپ‌هایی که در افراد مشترک است، بدست می‌آید (Browning and Browning, 2011). افراد با طول بلند هاپلوتیپ یکسان، ارتباط نزدیک‌تر و با طول هاپلوتیپ کوچک‌تر، ارتباط دورتری دارند. صحت مستندسازی تحت تأثیر تعداد و ترکیب افراد در جمعیت مرجع، اندازه مؤثر جمعیت و تفاوت بین تراکم ژنوتیپ‌های مرجع و مستندسازی شده قرار دارد (Sargolzaei *et al*., 2014). در مقابل، نرم‌افزار Beagle مبتنی بر یک مدل خوشه‌ای- هاپلوتیپی که هاپلوتیپ‌های موضعی را داخل خوشه‌ها قرار می‌دهد است (Browning *et al*., 2011). دقت این روش بسیار زیاد، ولی زمان محاسبات وقت‌گیر است (Sargolzaei *et al*., 2014). برای کشورهای در حال توسعه مانند ایران که گاوهای هلشتاین تعیین ژنوتیپ شده زیادی در دسترس نیست و از طرف دیگر سرمایه‌گذاری برای تعیین ژنوتیپ با تراشه‌های با تراکم بالا در گاوها نیز محدود است، افزایش صحت پیش‌بینی ژنومی با استفاده از روش مستندسازی می‌تواند راه‌گشا باشد. هدف از مطالعه حاضر، بررسی اثر اندازه جمعیت مرجع و تعداد نشانگرهای SNP از دست رفته بر صحت مستندسازی بود.

مواد و روش‌ها

از نرم‌افزار QMSim برای ایجاد یک جمعیت در طول زمان استفاده شد (Sargolzaei and Schenkel, 2009). در ابتدا، ۱۰۰۰ نسل مجزا با ۲۰۰۰ فرد شبیه‌سازی شد. برای ایجاد LD، از نسل ۱۰۰۰ به نسل ۱۰۲۰ تعداد جمعیت به ۱۰۰۰ حیوان کاهش پیدا کرد. تعداد افراد هر جنس در این مراحل با هم برابر بوده (تعداد ۱۰۰۰ حیوان) و تلاقی بر اساس جفت شدن تصادفی گامت‌هایی که نمونه‌های تصادفی از

1. Single nucleotide polymorphism
2. Imputation

ژنگاه گامت‌های نر و ماده بودند، انجام شد. تعداد کل حیوانات تعیین ژنوتیپ شده برای ایجاد بانک اطلاعاتی مرجع[1] به تعداد ۱۰۰۰ حیوان شبیه‌سازی شد. از حیوانات موجود در بانک اطلاعاتی، دو دسته[2] ایجاد شد: دسته اول (A) شامل ژنوتیپ‌های اصلی با تعداد ۵۲ هزار نشانگر SNP که شامل داده‌های گم‌شده[3] نیز بودند، شبیه‌سازی شد و در دسته دوم (B)، داده‌های گم‌شده از مجموع داده‌ها خارج شده و تعداد ۳۷ هزار نشانگر SNP باقی ماند. هر دو دسته در نرم‌افزار FImpute مورد استفاده قرار گرفتند (Sargolzaei *et al*., 2014). دلیل استفاده از این نرم‌افزار، تصحیح سریع داده‌ها و انجام سریع محاسبات بود. ژنوتیپ-های AA، AB، BB و ژنوتیپ‌های گم‌شده به ترتیب با اعداد ۰، ۱، ۲ و ۵ جایگزین شدند. برای مستندسازی از فایل داده پیش‌فرض جهت مستندسازی جمعیت استفاده شد (Sargolzaei *et al*., 2014). برای آزمون صحت مستندسازی جمعیت، اندازه جمعیت مرجع و تعداد نشانگرهای حذف شده به عنوان دو متغیر اصلی در نظر گرفته شدند. اندازه جمعیت مرجع به طور تصادفی به چهار گروه ۱۰۰، ۲۵۰، ۵۰۰ و ۷۵۰ حیوان تعیین ژنوتیپ شده (از ۱۰۰۰ حیوان شبیه‌سازی شده) تقسیم‌بندی شد. تعداد نشانگرهای حذف شده برای دو دسته A و B به صورت تصادفی برای پنج حالت ۱۵، ۳۰، ۵۵،۷۰ و ۹۵ درصد حذف نشانگرهای SNP در نظر گرفته شد. برای هر دسته، تعداد ۲۰ فایل آزمون ایجاد شد. مستندسازی برای هر فایل با سطوح متفاوت جمعیت مرجع و نشانگرهای حذف شده به وسیله نرم‌افزار FImpute ایجاد شد (Sargolzaei *et al*., 2014). صحت پیش‌بینی بر اساس همبستگی بین ارزش نشانگرهای SNP اصلی قبل از حذف و ارزش آن بعد از مستندسازی برآورد شد (Hickey *et al*., 2011; Carvalheiro *et al*., 2014). از آنجایی که انتخاب افراد برای جمعیت مرجع و همچنین انتخاب نشانگرهای SNP حذف شده به صورت تصادفی در نظر گرفته شد، کل فرآیند به میزان ۵۰ بار تکرار شد. برای هر دسته، ۱۰۰۰ مستندسازی ایجاد شد. صحت مستندسازی به کمک رابطه زیر به عنوان متوسط صحت برای تمام تکرارها بدست آمد:

$$Average\ accuracy = \frac{\sum_{i=1}^{n} \frac{cov(SNP1,SNP2)}{\sigma SNP1 * \sigma SNP2}}{n}$$

در این رابطه، n تعداد تکرارها، SNP1 و SNP2 به ترتیب کد نشانگر SNP قبل از حذف و بعد از مستندسازی بود (Carvalheiro et al., 2014). برای مقایسه معنی‌داری دو به دو هر کدام از دسته‌های A و B از تجزیه آماری Hotelling-Williams t- test استفاده شد (Olson *et al*., 2011). مقایسه صحت پیش‌بینی در تمام راهبردها برای افزایش اندازه جمعیت مرجع به صورت جداگانه برای دسته‌های A و B و برای ۱۰ تکرار آخر به کمک نرم‌افزار SAS نسخه ۹/۴ محاسبه و مقایسه میانگین‌ها به کمک آزمون دانکن انجام شد.

نتایج و بحث

صحت مستندسازی: بر اساس نتایج گزارش شده در جدول ۱، متوسط صحت مستندسازی به اندازه جمعیت مرجع و میزان تراکم نشانگرها وابسته بود. بیشترین میزان صحت با دامنه ۰/۸۳۲ تا ۰/۹۹۸ که شامل اطلاعات کامل نشانگرها بدون داده‌های حذفی بود در دسته B مشاهده شد. اما در دسته A که شامل نشانگرهای گم شده بود، دامنه صحت مقداری پایین‌تر از ۰/۷۶۵ تا ۰/۹۷۸ را به خود اختصاص داد. صحت پیش‌بینی ژنومی برای اندازه جمعیت مرجع در تمام راهبردها بین دسته A و B معنی‌دار برآورد شد (P<۰/۰۵). مطابق نتایج جدول ۱، با افزایش اندازه جمعیت مرجع از ۱۰۰ حیوان به ۷۵۰ حیوان در هر دو دسته، متوسط صحت افزایش معنی‌داری یافت (P<۰/۰۵)، که این نتیجه با نتایج VanRaden *et al*. (2011) و Calvalheiro *et al*. (2014) کاملاً مطابقت دارد. کمترین میزان صحت برای اندازه جمعیت مرجع ۱۰۰ حیوان و بیشترین میزان صحت برای اندازه جمعیت ۷۵۰ حیوان در هر دو دسته بدست آمد. مقدار دامنه صحت برای دسته A از ۰/۷۶۵ تا ۰/۹۷۸ و برای دسته B از ۰/۸۳۲ تا ۰/۹۹۸ مشاهده شد. گزارش شده است که با افزایش اندازه جمعیت مرجع، آثار نشانگرهای SNP با صحت و دقت بیشتر برآورد و متعاقباً صحت پیش‌بینی

1. Database
2. Dataset
3. Missing data

ژنومی افزایش می‌یابد (Elsen, 2016). در مطالعه Hong *et al*. (2017) گزارش شد که با افزایش تعداد افراد جمعیت مرجع احتمالاً مقدار LD بین نشانگر و QTL ها افزایش و در نتیجه صحت مطالعه پیش‌بینی ژنومی افزایش خواهد یافت. نتایج پژوهش حاضر در تایید مطالعات بالا بود.

از طرف دیگر، با افزایش تعداد نشانگرهای SNP حذف شده، برای هر دو دسته A و B و برای تمام اندازه‌های جمعیت مرجع، صحت پیش‌بینی ژنومی کاهش یافت (جدول ۱). نتایج مطالعه حاضر در مطابقت با این فرضیه که با افزایش تعداد نشانگرهای SNP حذف شده، صحت پیش‌بینی‌ها کاهش پیدا می‌کند است (Kranjcevicova *et al*., 2019). می‌توان اشاره کرد که صحت مستندسازی قابل پیش‌بینی به ترکیب داده‌ها بستگی دارد. بیشترین میزان صحت برای جمعیت مرجع با ۷۵۰ حیوان برای دسته B که فاقد تمام اطلاعات و دارای داده گمشده بود، به مقدار ۰/۹۹۸ بدست آمد. در سطوح پایین نشانگرهای حذف شده (۱۵درصد) برای هر دو دسته در اندازه‌های متفاوت جمعیت مرجع، تفاوت چندانی در صحت پیش‌بینی مشاهده نشد. بنابراین می‌توان بیان کرد که در اندازه کم تعداد نشانگرهای SNP حذف شده، اندازه جمعیت مرجع روی صحت اثر چندان مهمی ندارد. از سوی دیگر، نتایج مطالعه حاضر نشان داد که علی‌رغم تعداد نسبتاً زیاد نشانگرهای SNP حذف شده (۹۵ درصد) با مستندسازی، صحت نسبتاً بالایی در دامنه ۰/۷۶۵ تا ۰/۸۳۲ حتی در تعداد کم حیوانات جمعیت مرجع (۱۰۰ حیوان) حاصل می‌شود. در مقایسه با پژوهش حاضر، در مطالعه Kranjcevicova *et al*. (2019)، دامنه صحت ۰/۷۸ تا ۰/۷۹ برای اندازه جمعیت مرجع برابر ۱۰۰ حیوان و نشانگرهای SNP حذف شده به میزان ۹۵ درصد گزارش شد. در مطالعه Calvalheiro *et al*. (2014)، صحت مستندسازی با تراشه با تراکم بالا (۵۰ کیلو جفت باز) به میزان ۰/۹۷ گزارش شد. در مطالعه Ventura *et al*. (2014) که مستندسازی برای تراشه با ۶ کیلو باز به ۵۰ کیلو باز اجرا شد صحت مستندسازی برای جمعیتی به تعداد ۲۳۰۰ حیوان تعیین ژنوتیپ شده برآورد شد. میزان نشانگرهای SNP گم شده برابر با ۶۵ درصد و اندازه جمعیت مرجع به میزان ۱۱، ۳۳ و ۶۵ درصد حیوانات تعیین ژنوتیپ شده فرض شد. نتایج آن‌ها نشان داد که صحت مستندسازی برای جمعیت اول، دوم و سوم به ترتیب ۰/۸۷، ۰/۹۳ و ۰/۹۴ بود. در مطالعه دیگر که مستندسازی برای تراشه با ۶ کیلو باز به ۵۰ کیلو باز اجرا شد و اندازه جمعیت مرجع شامل ۸۰ درصد افراد تعیین ژنوتیپ شده بود، میزان صحت مستندسازی برابر با ۰/۹۲ گزارش شد (Wang *et al*., 2016). برخی از محققان گزارش کردند که اگر تعداد افراد جمعیت مرجع بسیار کوچک باشد، برای افزایش صحت مستندسازی باید خویشاوندان آن جمعیت را به جمعیت مرجع اضافه نمود (Ventura *et al*., 2014). افزایش خویشاوندان بسیار نزدیک به جمعیت مرجع، صحت مستندسازی را تا سه درصد افزایش می‌دهد (Kranjcevicova *et al*., 2019). همچنین در مطالعه Ghoreishifar *et al*. (2018) که روی گاومیش آبی انجام شد، برای سه جمعیت مرجع به تعداد ۳۰، ۵۲ و ۸۰ درصد جمعیت تعیین ژنوتیپ شده، صحت مستندسازی به ترتیب ۰/۸۸، ۰/۹۳ و ۰/۹۷ بود. در مطالعه حاضر، میزان متوسط صحت مستندسازی برای هر دو دسته A و B و برای اندازه جمعیت مرجع ۲۵ درصد به ترتیب ۰/۹۱ و ۰/۹۵، برای جمعیت مرجع۵۰ درصد به ترتیب ۰/۹۴ و ۰/۹۷ و برای اندازه جمعیت مرجع ۷۵ درصد، به ترتیب ۰/۹۶ و ۰/۹۸ بدست آمد. دلیل بالا بودن متوسط صحت مستندسازی احتمالاً تعداد زیاد هاپلوتیپ‌های مشترک بین افراد است. در مطالعه‌ای گزارش شد که با افزایش تعداد نشانگرهای SNP حذف شده از ۱۰ تا ۵۰ درصد، کاهش صحت معنی‌دار نبود، ولی با افزایش میزان نشانگرهای حذفی از ۵۰ به ۷۵ و ۹۰ درصد، صحت پیش‌بینی ژنومی مستندسازی کاهش معنی-داری پیدا می‌کند (Schurz *et al*., 2019). مشابه با این نتایج، در پژوهش حاضر نیز با افزایش تعداد نشانگرهای SNP حذف شده از ۷۰ به ۹۵ درصد، کاهش معنی‌دار صحت در تمام راهبردها مشاهده شد (جدول ۱). جدول شماره ۲ تعداد نشانگرهای SNP مستندسازی شده (ایمپیوت شده) در دو دسته A و B را نشان می‌دهد. از طرف دیگر با نرم‌افزار FImpute، مستندسازی برای هر دسته فایل داده (با سطوح متفاوت جمعیت مرجع و نشانگرهای حذف شده) و تاثیر آن بر صحت مستندسازی نشانگرهای SNP بررسی شد.

جدول ۱- متوسط صحت مستندسازی جمعیت برای دسته‌های A و B

Table 1. Average accuracy of population imputation for datasets A and B

Deleted markers (%)	Dataset A				Dataset B			
	Size of reference population				Size of reference population			
	100	250	500	750	100	250	500	750
15	0.945±0.20^{b}	0.953±0.18^{ab}	0.976±0.23^{a}	0.978±0.20^{a}	0.976±0.20^{b}	0.989±0.23^{a}	0.993±0.26^{a}	0.998±0.27^{a}
30	0.936±0.19^{b}	0.947±0.21^{ab}	0.973±0.20^{a}	0.976±0.24^{a}	0.968±0.21^{b}	0.986±0.20^{a}	0.990±0.24^{a}	0.993±0.25^{a}
55	0.925±0.16^{b}	0.939±0.19^{ab}	0.965±0.19^{a}	0.969±0.22^{a}	0.954±0.19^{b}	0.977±0.19^{a}	0.985±0.23^{a}	0.989±0.22^{a}
70	0.918±0.16^{b}	0.922±0.17^{ab}	0.954±0.18^{a}	0.963±0.21^{a}	0.943±0.18^{b}	0.968±0.17^{a}	0.978±0.22^{a}	0.980±0.23^{a}
95	0.765±0.10^{c}	0.806±0.13^{bc}	0.832±0.18^{b}	0.895±0.19^{a}	0.785±0.20^{c}	0.832±0.17^{b}	0.887±0.21^{a}	0.901±0.24^{a}

Means within each column with different superscripts for datasets A and B are significantly different (P<0.05)

جدول ۲- تعداد نشانگرهای SNP مستندسازی شده (ایمپیوت شده) در دو دسته A و B

Table 2. Number of single nucleotide polymorphisms (SNPs) to be imputed in datasets A and B

Deleted markers (%)	Dataset A (52000)	Dataset B (37000)
15	7800	5550
30	15600	11100
55	28600	20350
70	36400	25900
95	49400	35150

در جدول ۳، تاثیر افزایش نشانگرهای حذف شده بر متوسط درصد نشانگرهای SNP مستندسازی شده صحیح در حالت-های متفاوت اندازه جمعیت مرجع بررسی شد. به طور کلی، متوسط صحت درصد نشانگرهای SNP مستندسازی صحیح برای دسته B نسبت به دسته A بیشتر برآورد شد. متوسط درصد نشانگرهای SNP صحیح، وابسته به اندازه جمعیت مرجع بود و دامنه آن برای دسته A از ۰/۷۷۳ تا ۰/۹۹۵ و برای دسته B از ۰/۷۹۷ تا ۰/۹۹۹ برآورد شد. به طور کلی، با افزایش تعداد افراد جمعیت مرجع و کاهش تعداد نشانگرهای گم‌شده SNP، این میزان افزایش یافت. در مطالعه Calvalheiro *et al*. (2014)، درصد صحت مستندسازی با تراشه با تراکم بالا و با میزان ۹۵ درصد نشانگرهای گمشده SNP به صورت ۰/۹۷ گزارش شد. در مطالعه Kranjcevicova *et al*. (2019)، با افزایش اندازه جمعیت مرجع و همچنین کاهش تعداد نشانگرهای گمشده SNP نیز درصد نشانگرهای SNP مستندسازی شده صحیح افزایش داشت، که با نتایج پژوهش حاضر مطابقت داشت. درصد نشانگرهای SNP مستندسازی شده صحیح ممکن است تحت تاثیر صحت مستندسازی نادرست یک یا دو آلل نیز قرار گیرد (Kranjcevicova *et al*., 2019). مطابق نتایج بدست آمده، متوسط درصد نشانگرهای SNP صحیح با افزایش اندازه جمعیت مرجع چندان تحت تاثیر خیلی زیاد قرار نگرفت و تفاوت نیز معنی‌دار برآورد نشد. از طرف دیگر، با افزایش تعداد نشانگرهای SNP حذف شده، متوسط درصد نشانگرهای SNP صحیح کاهش پیدا نمود و این کاهش نیز در تمام حالات مربوط به اندازه جمعیت مرجع و برای هر دو دسته معنی‌دار بدست آمد (P<۰/۰۵). با افزایش خیلی زیاد نشانگرهای SNP حذف شده از ۷۰ به ۹۵ درصد، متوسط درصد نشانگرهای SNP مستندسازی شده صحیح برای تمام حالات اندازه جمعیت مرجع و برای هر دو دسته کاهش زیادی داشت. کمترین درصد نشانگرهای صحیح برای دسته A در ۹۵ درصد نشانگرهای حذف شده به میزان ۰/۷۷۳ و در دسته B به میزان ۰/۷۹۷ بدست آمد. دلیل این که با افزایش تعداد نشانگرهای SNP حذف شده، صحت مستندسازی نشانگرهای صحیح SNP کاهش پیدا می‌کند احتمالاً تا حدودی به نرم‌افزار Beagle که بدون استفاده از روابط خویشاوندی و استفاده از شجره، مستندسازی را انجام می‌دهد مربوط می‌شود. چون صحت پیش‌بینی ژنومی با

صحت نشانگرهای مستندسازی شده صحیح ارتباط مستقیم دارد، لذا با افزایش تعداد نشانگرهای حذف شده باید کاهش صحت را انتظار داشت.

نتیجه‌گیری کلی

نتیجه پژوهش حاضر نشان داد که اندازه جمعیت مرجع و تعداد نشانگرهای SNP حذف شده بر صحت مستندسازی اثرگذار هستند. اگر اندازه جمعیت مرجع به میزان کافی انتخاب شود حتی در صورت درصد بالای نشانگرهای SNP حذف شده، صحت مستندسازی هنوز بالا خواهد بود. به طور کلی اندازه جمعیت مرجع کوچک می‌تواند صحت مستندسازی را کاهش دهد. از طرف دیگر، با افزایش تعداد نشانگرهای SNP حذف شده، تعداد نشانگرهای SNP مستندسازی صحیح کاهش یافته و متعاقب آن احتمالاً کاهش صحت پیش‌بینی ژنومی ایجاد می‌شود.

جدول ۳- متوسط درصد نشانگرهای SNP مستندسازی شده صحیح برای دسته‌های A و B

Table 3. Average percentage of correctly imputed SNPs for datasets A and B

	Size of reference population for dataset A				Size of reference population for dataset B			
DM◆	100	250	500	750	100	250	500	750
15	0.975±0.24^{a}	0.981±0.25^{a}	0.991±0.26^{a}	0.995±0.28^{a}	0.982±0.26^{a}	0.989±0.23^{a}	0.993±0.26^{a}	0.999±0.27^{a}
30	0.967±0.22^{a}	0.980±0.24^{a}	0.989±0.25^{a}	0.990±0.26^{a}	0.981±0.25^{a}	0.987±0.20^{a}	0.991±0.25^{a}	0.995±0.26^{a}
55	0.963±0.20^{a}	0.975±0.23^{a}	0.978±0.24^{a}	0.989±0.22^{a}	0.965±0.22^{ab}	0.980±0.21^{a}	0.987±0.24^{a}	0.990±0.24^{a}
70	0.954±0.21^{a}	0.963±0.22^{b}	0.967±0.22^{b}	0.982±0.23^{a}	0.953±0.21^{b}	0.976±0.20^{a}	0.980±0.24^{a}	0.989±0.23^{a}
95	0.773±0.19^{b}	0.823±0.19^{c}	0.844±0.21^{c}	0.901±0.22^{b}	0.797±0.21^{c}	0.856±0.20^{b}	0.890±0.19^{b}	0.923±0.21^{b}

* In each column, different superscripts indicated the significant difference between means (P<0.05).

◆ DM means Deleted Markers (%).

فهرست منابع

Browning S. and Browning B. 2011. Haplotype phasing: Existing methods and new developments. Nature Reviews Genetics, 12: 703-714.

Browning B., Zhou Y. and Browning S. 2018. A one-penny imputed genome from next-generation reference panels. The American Journal of Human Genetics, 103: 338-348.

Carvalheiro R., Boison S., Neves H., Sargolzaei M., Schenkel F., Utsunomiya Y., O'Brien A., Solkner J., McEwan J., Van Tassell C., Sonstegard T. and Garcia J. 2014. Accuracy of genotype imputation in Nelore cattle. Genetics Selection Evolution, 46, 69.

Daetwyler H., Wiggans G., Hayes B., Woolliams J. and Goddard M. 2011. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics, 189: 317-327.

Elsen J. M. 2016. Approximated prediction of genomic selection accuracy when reference and candidate populations are related. Genetics Selection Evolution, 48, 16.

Ghoreishifar S. M., Moradi- Shahrbabak H., Moradi- Shahrbabak M., Nicolazzi E. L., Williams J. L., Iamartino D. and Nejati- Javaremi A. 2018. Accuracy of imputation of single-nucleotide polymorphism marker genotypes for water buffaloes (*Bubalus bubalis*) using different reference population sizes and imputation tools. Livestock Science, 216: 174-182.

Hayes B., Bowman P., Chamberlain A. and Goddard M. 2009. Invited review: Genomic selection in dairy cattle. Journal of Dairy Science, 92: 433-443.

Hickey J., Kinghorn B., Tier B., Wilson J., Dunstan N. and Van der Werf J. 2011. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genetics Selection Evolution, 43, 12.

Hong L. S, Clark S. and Van der Werf J. 2017. Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. Plos One, 21, 1-22.

Kranjčevičová A., Kašná E., Brzáková M. A, Přiby J. and Vostrý L. 2019. Impact of reference population size and marker density on accuracy of population imputation, Czech Journal of Animal Science, 64: 405-410.

Mulder H., Calus M., Druet T. and Schrooten C. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. Journal of Dairy Science, 95: 876-889.

Nicolazzi E., Biffani S. and Jansen G. 2013. Short communication: Imputing genotypes using PedImpute fast algorithm combining pedigree and population information. Journal of Dairy Science, 96: 2649-2653.

Olson K. M., VanRaden P. M., Tooker M. E. and Cooper T. A. 2011. Differences among methods to validate genomic evaluations for dairy cattle. Journal of Dairy Science, 94: 2613-2620.

Sargolzaei M., Chesnais J. and Schenkel F. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics, 15, 12.

Sargolzaei M. and Schenkel F. S. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics, 25: 680-681.

Schaeffer L. 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics, 23: 218-223.

Schurz H., Stephanie J. M., van Helden P. D., Tromp G., Hoal E. G., Kinnear C. J. and Möller M. 2019. Evaluating the accuracy of imputation methods in a five-way admixed population. Frontiers in Genetics, 10: 34.

VanRaden P., O'Connell J., Wiggans G. and Weigel K. 2011. Genomic evaluations with many more genotypes. Genetics Selection Evolution, 43, 1-10.

VanRaden P., Sun C. and O'Connell J. 2015. Fast imputation using medium or low-coverage sequence data. BMC Genetics, 16(82): 2039-2042.

Ventura R., Lu D., Schenkel F. S., Wang Z., Li C. and Miller S. P. 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. Journal of Animal Science, 92: 1433-1444.

Wang Y., Lin G., Li C. and Stothard P. 2016. Genotype imputation methods and their effects on genomic predictions in cattle. Springer Science Reviews, 4: 79-98.

Whalen A., Gorjanc G., Ros- Freixedes R. and Hickey J. 2018. Assessment of the performance of hidden Markov models for imputation in animal breeding. Genetics Selection Evolution, 50, 4-10.

Zhang Z. and Druet T. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. Journal of Dairy Science, 93: 5487-5494.

Animal Production Research

Vol. 9, No. 2, 2020 (15-22)

University of Guilan

Research paper

Impact of marker density and reference population size on accuracy of imputation in simulated data

Y. Mohammadi*

Assistant Professor, Department of Animal Science, Faculty of Agriculture, Ilam University, Ilam, Iran

(Received: 30-10-2019 – Accepted: 06-02-2020)

Abstract

In this study, effect of the reference population size and the number of missing single nucleotide polymorphisms (SNPs) on imputation accuracy was assessed. The QMSim software was used to create a reference database of 1000 simulated animals. Two datasets were created from the database reference: The first dataset (A), included original genotypes, containing the missing SNPs (52,000 SNP markers), and the second one (B) included the same genotypes without the missing data (37,000 SNP markers). In both datasets, animals were simulated for a reference population with the size of 100, 250, 500 and 750. The deleted SNPs were simulated randomly in both datasets with the proportion of 15%, 30%, 55%, 70%, and 95%. The accuracy was determined based on the correlation between the original SNP values before deletion and its values after imputation. The results of this study showed that the accuracy of the imputation was influenced by the size of reference population and density of the deleted SNP markers. By increasing the reference population size from 100 to 750 animals in both datasets, the average accuracy of the imputation was increased. The highest accuracy in the reference population of 750 animals was from 0.89 to 0.98 in dataset A and 0.90 to 0.99 in dataset B. Generally, the results showed that if the size of the reference population is sufficient, the imputation accuracy does not much change, despite large number of missing SNPs.

Keywords: Genomic evaluation, Missing data, Animal, Prediction accuracy, Imputation

*Corresponding author: Mohamadi_yahya@yahoo.com

doi: 10.22124/ar.2020.14875.1466