نشریه پژوهشهای علوم دامی ایران
جلد ۱۰، شماره ۴، زمستان ۱۳۹۷، ص. ۵۶۰-۵۵۳

Iranian Journal of Animal Science Research
Vol. 10, No. 4, Winter 2019, p. 553-560

بررسی عملکرد الگوریتم هوشمند تجزیه مقدار تکین (SVD) در بازیابی ژنوتیپ‌های از دست رفته در سناریوهای مختلف از تعداد نشانگر، اندازه جمعیت و فراوانی آلل نادر

فرهاد غفوری کسبی[۱]- علی گودرز تله جردی[۲*]

تاریخ دریافت: ۱۳۹۶/۰۵/۰۵
تاریخ پذیرش: ۱۳۹۶/۰۸/۰۲

چکیده

هدف از این تحقیق بررسی عملکرد الگوریتم هوشمند تجزیه مقدار تکین (SVD)[۳] در بازیابی ژنوتیپ‌های از دست رفته بود. به این منظور، ژنومی متشکل از ۱ کروموزوم به طول یک مورگان که بر روی آن در سناریوهای مختلف به‌ترتیب ۵۰۰، ۱۰۰۰، ۱۵۰۰، ۲۰۰۰، ۲۵۰۰ و ۳۰۰۰ نشانگر تک نوکلئوتیدی دو آللی (SNP) با فراوانی اولیه یکسان ۰/۵ توزیع شده بود برای ۱۰۰۰ فرد شبیه‌سازی شد. در ادامه جهت ایجاد فایل اطلاعات در چهارچوب اطلاعات"تعیین ژنوتیپ با توالی‌یابی" [۴](GBS) اطلاعات ژنوتیپی به ترتیب ۵٪، ۱۰٪، ۲۵٪، ۵۰٪، ۷۵٪ و ۹۰٪ از SNPهای افراد از ماتریس ژنوتیپی حذف شده و مجدداً توسط روش SVD بازیابی شدند. درصد ژنوتیپ‌های به‌درستی بازیابی شده (نسبت تعداد ژنوتیپ‌های به درستی بازیابی شده به کل ژنوتیپ‌های از دست رفته) به‌عنوان شاخصی از صحت بازیابی ژنوتیپ (*r*) در سناریوهای مختلف مورد استفاده قرار گرفت. صحت بازیابی ژنوتیپ‌های از دست رفته با استفاده از روش SVD قابل توجه بود به طوری که با افزایش درصد ژنوتیپ‌های از دست رفته تا ۵۰٪، SVD با صحتی در حدود ۸۰٪ ژنوتیپ‌های از دست رفته را بازیابی نمود. در سناریوهای ۷۵٪ و ۹۰٪ ژنوتیپ از دست رفته صحت بازیابی ژنوتیپ کاهش یافته و به ترتیب ۷۰٪ و ۴۸٪ بود. در شرایط برابر از تعداد نشانگر و درصد ژنوتیپ از دست رفته، با افزایش تعداد افراد حاضر در جمعیت از ۱۰۰۰ به ۲۰۰۰ فرد، توانایی بازیابی ژنوتیپ توسط روش SVD افزایش یافت. در یک درصد ثابت از ژنوتیپ‌های از دست رفته، با افزایش تعداد نشانگر صحت بازیابی ژنوتیپ افزایش یافت به نحوی که با افزایش تعداد نشانگر از ۵۰۰ به ۳۰۰۰ نشانگر، حدوداً ۱۰٪ به صحت بازیابی ژنوتیپ افزوده شد. یک رابطه معکوس بین میزان فراوانی آلل نادر (MAF) و *r* مشاهده شد به گونه‌ای که با افزایش MAF از ۰/۰۱ به ۰/۴۰ صحت بازیابی ژنوتیپ به میزان ۸ درصد کاهش یافت. به طور کلی نتایج این تحقیق نشان داد که الگوریتم SVD با صحت بالایی می‌تواند ژنوتیپ‌های از دست رفته را بازیابی کند به ویژه زمانی که درصد ژنوتیپ‌های از دست رفته کم باشد، اندازه جمعیت بزرگ باشد و فراوانی آلل نادر نیز پایین باشد.

واژه‌های کلیدی: الگوریتم SVD، بازیابی ژنوتیپ، SNP

مقدمه

اگرچه استفاده از اطلاعات ژنتیک مولکولی برای اهداف اصلاح نژادی سال‌ها پیش توسط نیمن-سورنسون و روبرتسون (۱۳) ارائه شده بود، اما علی‌رغم وجود دانش لازم به دلیل عدم دسترسی به اطلاعات ژنتیکی برای سالیان متمادی انجام این امر میسر نبود. در سال‌های ابتدایی قرن حاضر، پیشرفت‌های فراوان در توالی‌یابی ژنوم حیوانات اهلی که امکان شناسایی هزاران نشانگر DNA به شکل چندشکلی تک نوکلئوتیدی[۵] (SNP) را فراهم نمود زمینه را برای ارائه ایده انتخاب ژنومی فراهم کرد (۱۴). میوسن و همکاران (۱۴) در مقاله معروف خود چگونگی استفاده از ابزار ریاضی و به خصوص تئوری بیز را در پیش‌بینی اثر نشانگرها جهت برآورد ارزش‌های اصلاحی ژنومی تشریح نمودند. بعد از آن محققین دیگر نیز چگونگی استفاده از این روش در برنامه‌های مختلف اصلاح نژادی را نشان دادند (۴، ۸، ۱۶،۲۰، ۲۳ و ۲۴). در انتخاب ژنومی فرآیند انتخاب از طریق تعیین ژنوتیپ افراد برای تعداد زیادی نشانگر SNP صورت می‌گیرد که بر حسب گونه تعداد SNPها متفاوت است. برای مثال در حال حاضر از تراشه‌های متراکم SNP که در برگیرنده اطلاعات ژنوتیپی ۵۴۰۰۰ SNP (۵۴k) است به صورت تجاری برای تعیین ژنوتیپ گاوهای

۱- استادیار، گروه علوم دامی، دانشکده کشاورزی، دانشگاه بوعلی سینا، همدان، ایران.
۲- استادیار، گروه پاتوبیولوژی، دانشکده پیرادامپزشکی، دانشگاه بوعلی سینا، همدان، ایران.
(*- نویسنده مسئول: (Email: A.goudarz@basu.ac.ir
DOI: 10.22067/ijasr.v10i4.66389
3- Singular Value Decomposition
4- Genotype by Sequencing
5- Single Nucleotide Polymorphism

شیری استفاده می‌شود اگرچه تراشه های ۷۵۰۰۰۰ SNP (۷۵۰k) نیز در دسترس است اما به دلیل هزینه بالای تعیین ژنوتیپ با این تراشه‌ها، استفاده از آنها به پروژه‌های تحقیقاتی محدود است (۲۲).

در هنگام تعیین ژنوتیپ، معمولاً به‌طور میانگین اطلاعات ژنوتیپی ۵٪ از SNPها از دست می‌رود. لذا قبل از انجام ارزیابی ژنومی این اطلاعات از دست رفته باید به نحوی بازیابی شوند چرا که برخی از این SNPها ممکن است بزرگ اثر باشند و فقدان اطلاعات مربوط به آنها منجر به کاهش صحت ارزیابی ژنومی شود (۱۱). راه حلی که برای این مساله مورد استفاده قرار می‌گیرد بازیابی ژنوتیپ[1] (۱۲) نامیده می‌شود که مرحله‌ای ضروری و پیش نیاز ارزیابی ژنومی است. گاهی اوقات نیز بنا به دلایل اقتصادی حیوانات با تراشه‌های SNP با تراکم پایین مثلاً ۳k یا ۷k تعیین ژنوتیپ می‌شوند و جهت بهره‌برداری از اطلاعات موجود در تراشه‌های با تراکم بالاتر از SNPها مثلاً تراشه‌های ۵۴k، تراشه ۳k یا ۷k به تراشه ۵۴k بازیابی می‌شود (۲).

اخیراً به منظور توسعه انتخاب ژنومی و کاهش هزینه‌های مربوط به تعیین ژنوتیپ، استفاده از اطلاعات تعیین ژنوتیپ با توالی‌یابی و یا در اصطلاح اطلاعات GBS (۶) که یکی از پروتکل‌های خانواده نسل بعدی توالی‌یابی[2] (NGS) است مورد توجه قرار گرفته است. هزینه تعیین ژنوتیپ با استفاده از این روش به مراتب کمتر از روش‌های رایج است. استفاده از اطلاعات GBS در ارزیابی ژنومی هنوز در مراحل اولیه قرار داشته اما انتظار این است که به زودی شاهد استفاده هرچه بیشتر این اطلاعات در ارزیابی ژنومی باشیم (۵ و ۹). یکی از معایب استفاده از اطلاعات GBS این است که همه SNPها تعیین ژنوتیپ نمی‌شوند و بنابراین اطلاعات ژنوتیپی بخشی از SNPها در آنها وجود ندارد. در این روش، SNPهای فاقد اطلاعات ژنوتیپی در حیوانات مختلف مشابه نبوده و حالت تصادفی دارند (جدول ۱) و لذا با استفاده از روش‌های رایج بازیابی ژنوتیپ (۳ و ۱۶) قابل بازیابی نیستند. برخی روش‌های هوشمند که عمدتاً جزء روش‌های ناپارامتری می‌باشند برای بازیابی ژنوتیپ‌های از دست رفته در داده‌های GBS قابل استفاده هستند. در این تحقیق عملکرد یکی از این روش‌ها که در اختصار تجزیه مقدار تکین (SVD) (۲۵) نامیده می‌شود در بازیابی ژنوتیپ‌های از دست رفته مورد بررسی قرار خواهد گرفت. روش SVD جزو روش‌های توانمند در بازیابی داده‌های گم شده است به‌ویژه در زمانی‌که با فایل‌های اطلاعات دارای خطاهای ژنوتیپی مواجه هستیم، استفاده از این روش پیشنهاد می‌شود (۲۵).

مواد و روش‌ها

با استفاده از بسته نرم‌افزاری *hypred* (۲۳) ژنومی متشکل از ۱ کروموزوم به طول یک مورگان که بر روی آن در سناریوهای مختلف به‌ترتیب ۵۰۰، ۱۰۰۰، ۱۵۰۰، ۲۰۰۰، ۲۵۰۰ و ۳۰۰۰ SNP با فراوانی اولیه یکسان ۰/۵ در قالب توزیع نرمال توزیع شده بود شبیه‌سازی شد. به هر جایگاه SNP با ژنوتیپ AA کد ۲، با ژنوتیپ Aa کد ۱ و با ژنوتیپ aa کد صفر اختصاص داده شد.

جمعیت پایه به تعداد ۱۰۰ فرد (۵۰ نر و ۵۰ ماده) شبیه‌سازی شده و اجازه داده شد تا برای ۵۰ نسل به طور تصادفی در آن آمیزش صورت بگیرد. در این حالت به طور تصادفی از هاپلوتایپ‌های پدری و مادری نمونه‌گیری شده و از آنها برای تولید نتاج استفاده گردید. از هر دو والد فقط دو فرزند ایجاد خواهد شد که در نتیجه اندازه جمعیت در طی ۵۰ نسل در تعداد ۱۰۰ فرد ثابت باقی خواهد ماند. به عبارت دیگر در طی این نسل ها اندازه موثر جمعیت (*Ne*) ۱۰۰ خواهد بود. در نسل ۵۱ اندازه جمعیت به ۱۰۰۰ و در سناریوی دیگر به ۲۰۰۰ فرد افزایش داده شد که دارای اطلاعات ژنوتیپی بوده و ماتریس ژنوتیپی بر اساس ژنوتیپ این افراد تشکیل شد. در ادامه به ترتیب ۵، ۲۰، ۵۰، ۷۰ و ۹۰ درصد از اطلاعات ژنوتیپ‌ها در افراد حذف شدند. برای بررسی تأثیر تعداد افراد، دو سناریو شامل ۱۰۰۰ و ۲۰۰۰ فرد که دارای اطلاعات ژنوتیپی برای ۱۰۰۰ نشانگر بودند بررسی شد. برای بررسی حداقل فراوانی آللی[3] (MAF)، سطوح مختلف MAF شامل ۰/۰۱، ۰/۰۵، ۰/۱۰، ۰/۲۰، ۰/۳۰ و ۰/۴۰ در نظر گرفته شد. پارامترهای استفاده شده برای شبیه‌سازی ماتریس ژنوتیپی در جدول ۲ نشان داده شده است. با استفاده از روش هوشمند SVD (۲۵)، ابتدا ماتریس حاوی اطلاعات ژنوتیپی تجزیه شده و ویژه مقدارهای آن به‌دست می‌آید و سپس از طریق آنها ژنوتیپ‌های از دست رفته بازیابی می‌شوند. ماتریس ژنوتیپی (M) به صورت زیر را در نظر بگیرید:

$$M = U\sum V^T$$

که در آن U ابعادی برابر با m×k داشته و حاوی ویژه بردارها است، V هم ماتریس حاوی ویژه مقدارها با ابعاد n×k است. بازیابی ژنوتیپ در طی مراحل زیر صورت می‌گیرد. ۱) ابتدا اطلاعات از دست رفته به‌وسیله روش[4] MNI که روشی حد واسط است بازیابی می‌شوند. روش MNI که یک روش بازیابی نسبتاً ضعیف است یک بازیابی اولیه از ژنوتیپ از دست رفته نشانگر j (jامین SNP) را به‌دست می‌دهد که به‌عنوان یک پیش برآورد برای SVD قلمداد

1- Genotype Imputation
2- Genotype by Sequencing
3- Minor Allele Frequency
4- Mean Neighbor Imputation

می‌شود. در این مرحله ماتریس M به صورت ابتدایی کامل می‌شود. ۲) ویژه مقدارهای ماتریس ژنوتیپی (M) برآورد شده و در ماتریس U ذخیره می‌شوند. ۳) برای هر نشانگر *j*، ضرایب رگرسیونی از هر ستون ماتریس U از طریق یک رگرسیون خطی چند متغیره به‌دست می‌آید (یک ضریب از هر ستون).

فرد Animal	نشانگر Marker									
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
1		0		1	2	1	0	2		0
2	2		1	0		0	0	1		1
3	1	0		2	2	0		1	0	
4		2	1		0	1	2	0	1	2
5	0		2	0		2			1	1
6	2	2	0		1		0	0	2	
7	2		1	0	2	1		2	1	0
8	1	2	0	2		1	0	0	2	

جدول ۱- شکل شماتیک از اطلاعات GBS در تعدادی از افراد ($0=A_1A_1$, $1=A_1A_2$, $2=A_2A_2$)

Table 1- Ilustration example of GBS data in some individuals ($0=A_1A_1$, $1=A_1A_2$, $2=A_2A_2$)

جدول ۲- پارامترهای مورد استفاده در شبیه‌سازی ماتریس ژنوتیپی

Table 2- Parameters used for simulation of genotypic matrix

اندازه ژنوم Genome size	1M
تعداد کروموزوم Number of chromosome	1
تعداد SNP به‌ازای هر کروموزوم Number of SNP per chromosome	500, 1000, 1500, 2500, 3000
تعداد افراد Number of individuals	1000, 2000
تعداد مؤثر جمعیت Effective population size (Ne)	100
فراوانی آلل نادر Minor allele frequency (MAF)	0.01, 0.05, 0.10, 0.20, 0.30

$$y = \hat{U} + \varepsilon$$

که در آن y یک بردار ستونی برای نشانگر *j*است، U ماتریسی با ابعاد m×k و حاوی k ویژه مقدار است، b برداری حاوی ضرایب رگرسیونی و ε میزان خطای مدل را نشان می‌دهد. فقط افرادی که فاقد ژنوتیپ از دست رفته برای نشانگر *j*ام هستند برای تخمین b استفاده می‌شوند. ۴) U و ضرایب رگرسیونی برآورد شده (b) جهت پیش‌بینی ژنوتیپ‌های از دست رفته نشانگر *j* مورد استفاده قرار می‌گیرد. ۵) با استفاده از ماتریس کامل شده مراحل ۲ تا ۴ تکرار می‌شوند تا زمانی که رابطه زیر برقرار شود:

$$\frac{[RSS_0 - RSS_1]}{RSS_1} < T$$

در رابطه فوق RSS مجموعه مربعات باقی‌مانده بین مقادیر از دست نرفته و تخمین SVDیِ آنها است. RSS_0 و RSS_1 مقادیر RSS در تکرارهای پی در پی است و T آستانه‌ای است که توسط کاربر تعیین می‌شود و تکرار تا زمانی ادامه می‌یابد تا حاصل کسر فوق از T کوچکتر شود. برای بازیابی ژنوتیپ‌ها به روش SVD از بسته نرم‌افزاری bcv (۱۷) استفاده شد. سایر تجزیه و تحلیل‌های انجام شده نیز با استفاده از توابع موجود در نرم افزار R انجام شد.

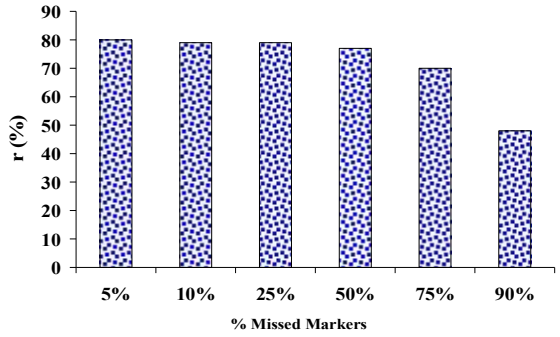
به طور مشابه با تحقیقات دیگر (۷، ۱۶ و ۲۷) درصد ژنوتیپ‌های به‌درستی پیش‌بینی شده به صورت زیر به‌عنوان شاخص صحت پیش-بینی ژنوتیپ (*r*) مورد استفاده قرار گرفت.

r= [تعداد کل ژنوتیپ از دست رفته/تعداد ژنوتیپ به درستی بازیابی شده] ×۱۰۰

نتایج و بحث

همانطور که در شکل ۱ مشاهده می‌شود صحت بازیابی ژنوتیپ‌های از دست رفته (r) با استفاده از روش SVD قابل توجه است. در سناریو ۵٪ ژنوتیپ از دست رفته، صحت بازیابی ژنوتیپ ۸۰٪ است و با افزایش درصد ژنوتیپ‌های از دست رفته تا ۵۰٪، صحت بازیابی کماکان در حدود ۸۰٪ حفظ می‌شود. با افزایش درصد ژنوتیپ‌های از دست رفته به ۷۵٪ و بعد از آن تا ۹۰٪، صحت بازیابی ژنوتیپ به ترتیب به ۷۰٪ و ۴۸٪ کاهش می‌یابد. این نتیجه منطقی است چرا که با افزایش درصد ژنوتیپ‌های از دست رفته منابع اطلاعاتی قابل بهره‌برداری توسط الگوریتم کاهش خواهد یافت. برای مثال در حالت ۹۰٪ ژنوتیپ از دست رفته، از مجموع SNP۱۰۰۰ فقط اطلاعات ژنوتیپی ۱۰۰ SNP باقی خواهد ماند در صورتی‌که در سناریوی ۵۰٪ ژنوتیپ از دست رفته، کماکان اطلاعات ژنوتیپی ۵۰۰ SNP وجود داشته که قابل استفاده توسط الگوریتم است. بری و کنی (۱) در گاوهای هلشتاین و و ورژکن و همکاران (۲۶) در جوجه‌های گوشتی با استفاده از یک الگوریتم بازیابی ژنوتیپ مبتنی بر مدل مخفی مارکوف[1] نشان دادند که هرچه تعداد SNPی که باید ژنوتیپشان بازیابی شود کمتر باشد، صحت بازیابی ژنوتیپ نیز افزایش می‌یابد چرا که درصد SNPهایی که به درستی بازیابی می‌شوند افزایش می‌یابد و به عبارتی خطای بازیابی ژنوتیپ کاهش می‌یابد. هیکی و همکاران (۱۱) درصدهای مختلف شامل ۵، ۵۰، ۷۵، ۸۷، ۹۴، ۹۸ و ۹۹ درصد از اطلاعات ژنوتیپی یک تراشه ۵۴k را حذف و سپس با استفاده از مدل مخفی مارکوف بازیابی نمودند و گزارش کردند که با افزایش حذف اطلاعات ژنوتیپی و بازیابی آن، میزان r از حدود ۱ (بازیابی ۵٪ از ژنوتیپ‌ها) به ۰/۲۰ (بازیابی ۹۹٪ از ژنوتیپ‌ها) کاهش یافت. کاهش در صحت بازیابی ژنوتیپ‌های از دست رفته به نوبه خود منجر به کاهش صحت پیش‌بینی ارزش‌های اصلاحی ژنومی خواهد شد. هیکی و همکاران (۱۱)، پی و همکاران (۱۶) و ولمن و همکاران (۲۸) گزارش نمودند که با افزایش درصد ژنوتیپ‌های بازیابی شده، صحت پیش‌بینی ارزش‌های اصلاحی ژنومی حاصل از تراشه‌های حاوی ژنوتیپ‌های بازیابی شده کاهش می‌یابد که می‌تواند به عنوان یکی از مشکلات بازیابی ژنوتیپ مطرح باشد اگرچه میزان کاهش در صحت ارزیابی ژنومی در زمانی‌که درصد ژنوتیپ‌های بازیابی شده کم باشد بسیار اندک و در حد یک درصد است. در این مطالعات، پنل‌های SNP با تراکم پایین (حداکثر ۷k و عمدتاً ۳k) به تراشه ۵۴k بازیابی شده و سپس ارزش‌های اصلاحی ژنومی در دو حالت استفاده از تراشه‌های حاوی اطلاعات بازیابی شده و تراشه‌های ۵۴k کامل پیش‌بینی شده و صحت پیش‌بینی‌ها با هم مقایسه شده‌اند. نتایج نشان داد که صحت پیش‌بینی ارزش‌های اصلاحی ژنومی با استفاده از اطلاعات حاوی ژنوتیپ‌های بازیابی شده در دامنه ۰/۴۵ (بازیابی نمودن تراشه ۰/۵k به تراشه ۵۴k) تا ۰/۹۹ (بازیابی نمودن تراشه ۷k به تراشه ۵۴k) صحت پیش‌بینی حاصل از تراشه ۵۴k قرار دارد (۲۸).

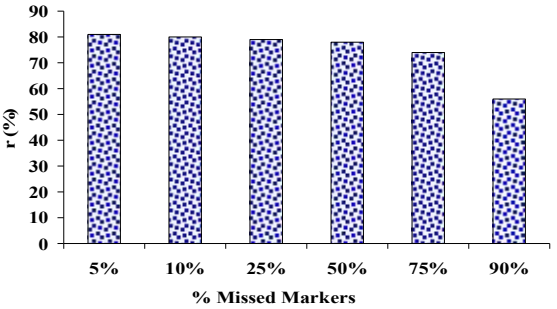
در شکل ۲، تأثیر اندازه جمعیت بر صحت بازیابی ژنوتیپ نشان داده شده است. سناریو مطالعه شده همان سناریوی شکل ۱ می‌باشد با این تفاوت که تعداد افراد حاضر در جمعیت به ۲۰۰۰ فرد افزایش پیدا کرده است و در نتیجه اطلاعات ژنوتیپی برای هر جایگاه دو برابر شده است. همان‌گونه که مشاهده می‌شود خصوصاً در سناریوهای ۷۵٪ و ۹۰٪ ژنوتیپ از دست رفته صحت بازیابی در مقایسه با شکل ۱ افزایش یافته است. پی و همکاران (۱۶) در یک مطالعه شبیه‌سازی و با استفاده از روش‌های بازیابی ژنوتیپ مبتنی بر خوشه‌بندی هاپلوتایپی، زنجیره مخفی مارکوف و مدل مخفی مارکوف گزارش کردند که در سناریوهای مختلف از تراکم نشانگری، با افزایش اندازه جمعیت از ۵۰ به ۴۵۰ صحت بازیابی ژنوتیپ به میزان ۵ درصد افزایش یافت. روشیارا و شولتز (۱۸) و روشیارا و همکاران (۱۹) با استفاده از شبیه‌سازی و همچنین اطلاعات مربوط به تراشه‌های SNP انسان تأثیر اندازه جمعیت بر صحت بازیابی ژنوتیپ را بررسی کرده و گزارش نمودند که در عمده روش‌های بازیابی ژنوتیپ، با افزایش اندازه جمعیت از ۴۰ به ۲۵۰۰ نفر صحت بازیابی ژنوتیپ افزایش می‌یابد. حیدری تبار و همکاران (۱۰) نیز افزایش در تعداد افراد را به‌عنوان یک راهکار مؤثر برای افزایش صحت بازیابی ژنوتیپ پیشنهاد دادند خصوصاً زمانی که افراد جمعیت با تراشه‌های SNP با تراکم پایین تعیین ژنوتیپ شده باشند. با افزایش اندازه جمعیت تعداد افراد دارای ژنوتیپ معلوم برای جایگاه مورد نظر افزایش می‌یابد و در نتیجه آن احتمال اینکه ژنوتیپ از دست رفته به‌درستی بازیابی شود افزایش خواهد یافت چرا که عدم قطعیت تصمیم نهایی در مورد ژنوتیپ صحیح برای جایگاه کاهش می‌یابد.



شکل ۱- صحت بازیابی ژنوتیپ (r) در درصدهای مختلف از ژنوتیپ‌های از دست رفته (جمعیت شامل ۱۰۰۰ فرد، ۱۰۰۰ نشانگر)

Figure 1- Accuracy of genotype imputation (r) in different scenarios of masked genotypes(population consists 1000 individuals each genotyped for 1000 markers)

1- Hidden Markov Model

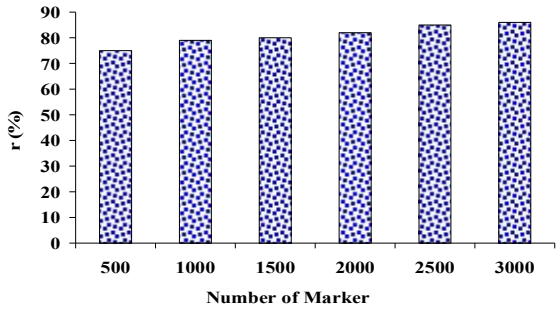


شکل ۲- تاثیر افزایش تعداد افراد از ۱۰۰۰ به ۲۰۰۰ فرد بر صحت بازیابی ژنوتیپ (r) در درصدهای مختلف از ژنوتیپ‌های از دست رفته

Figure 2- The effect of increase in number of individual from 1000 to 2000 on accuracy of genotype imputation (r) in different levels of masked genotypes

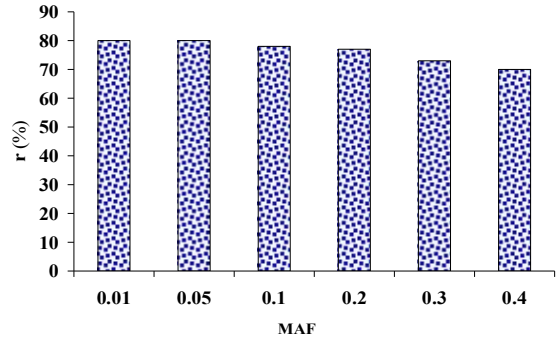
در شکل ۳، تأثیر تعداد نشانگر بر صحت بازیابی ژنوتیپ‌های از دست رفته در جمعیتی شامل ۱۰۰۰ فرد نشان داده شده است. همانگونه که مشاهده می‌شود با افزایش تعداد نشانگر صحت بازیابی ژنوتیپ افزایش می‌یابد. البته میزان افزایش صحت بازیابی ژنوتیپ به صورت خطی نیست و به عبارت دیگر با دو برابر شدن تعداد نشانگر صحت بازیابی دو برابر نمی‌شود چرا که اطلاعات مفید جهت بهره برداری توسط الگوریتم دو برابر نخواهد شد. با افزایش تعداد نشانگر از ۵۰۰ به ۳۰۰۰ نشانگر، حدوداً ۱۰٪ به صحت بازیابی ژنوتیپ افزوده شده است. نتیجه اخیر با یافته‌های هیکی و همکاران (۱۱) درتطابق است. همچنین شروتن و همکاران (۱۹) در گونه گاو نشان دادند که افزایش تعداد نشانگر منجر به افزایش صحت بازیابی ژنوتیپ خواهد شد. درضمن این محققین نشان دادند که تأثیر تعداد نشانگر بسیار بیشتر از اندازه کروموزوم بر صحت بازیابی ژنوتیپ تأثیر می‌گذارد. با توجه به اینکه نسبت حذف ژنوتیپ‌ها در همه سناریوها ثابت و به میزان ۵۰٪ بود، با افزایش تعداد نشانگر، نسبت نشانگرهای با ژنوتیپ معلوم افزایش می‌یابد و به عبارت دیگر منابع اطلاعات ژنوتیپی مورد استفاده توسط الگوریتم افزایش خواهد یافت و در نتیجه توانایی مدل جهت بازیابی ژنوتیپ‌های از دست رفته افزایش می‌یابد.

شکل ۴ نیز تأثیر سطوح مختلف MAF بر صحت بازیابی ژنوتیپ را نشان می‌دهد (نسبت حذف ژنوتیپ‌ها در همه سطوح MAF ثابت و به میزان ۵۰٪ بود). یک رابطه معکوس بین میزان MAF و r مشاهده شد به گونه‌ای که با افزایش MAF از ۰/۰۱ به ۰/۴۰ صحت بازیابی ژنوتیپ به میزان ۸ درصد کاهش می‌یابد. نتیجه اخیر نشان می‌دهد که نشانگرهای با MAF پایین با صحت بالاتری بازیابی می‌شوند. هیکی و همکاران (۱۰) و پی و همکاران (۱۶) نیز نتایج مشابهی را گزارش نمودند. لین و همکاران (۲۰۱۰) و ونگ و همکاران (۲۰۱۲) سطوح مختلف MAF را بر صحت بازیابی ژنوتیپ بررسی کرده و گزارش کردند که افزایش MAF به بیشتر از ۵٪ کاهش صحت بازیابی ژنوتیپ را در پی خواهد داشت. با افزایش MAF درصد ژنوتیپ هتروزیگوت برای جایگاه‌های مختلف افزایش می‌یابد. در حالت اخیر معمولاً در اکثر روش‌های بازیابی ژنوتیپ صحت بازیابی به دلیل کاهش عدم قطعیت تصمیم نهایی در مورد نوع ژنوتیپ برای جایگاه‌ها کاهش می‌یابد (۱۱). با کاهش تعداد نشانگرها از ۱۰۰۰ به ۵۰۰ عدد تأثیر افزایش MAF بر صحت بازیابی ژنوتیپ افزایش می‌یابد خصوصاً در سطوح بالاتر از MAF. این مسأله در شکل ۵ نشان داده شده است. ژانگ و همکاران (۳۰) گزارش نمودند که وقتی سطح MAF پایین باشد یک راه برای افزایش صحت بازیابی ژنوتیپ افزایش اندازه جمعیت و در نتیجه افزایش اندازه ماتریس ژنوتیپی است.



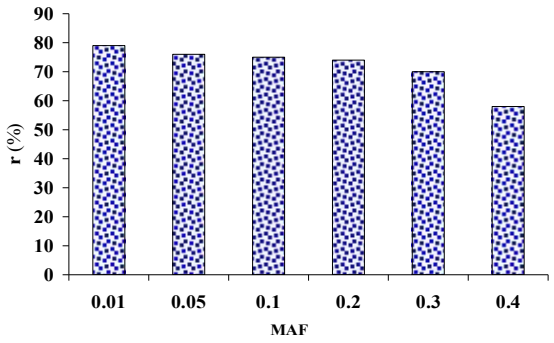
شکل ۳- تأثیر تعداد نشانگر بر صحت بازیابی ژنوتیپ (r)

Figure 3- The effect of number of markers on accuracy of genotype imputation (r)



شکل ٤- تأثیر سطوح مختلف MAF بر صحت بازیابی ژنوتیپ (r) (جمعیت شامل ۱۰۰۰ فرد، ۱۰۰۰ نشانگر)

Figure 4- The effect of different levels of MAF on accuracy of genotype imputation (r) (population consisted of 1000 individuals genotyped for 1000 markers)



شکل ۵- تأثیر سطوح مختلف MAF بر صحت بازیابی ژنوتیپ (r) با استفاده از اطلاعات ۵۰۰ نشانگر (جمعیت شامل ۱۰۰۰ فرد)

Figure 5- The effect of different levels of MAF on accuracy of genotype imputation (r) using 500 markers (population consists 1000 individuals)

نتیجه‌گیری کلی

به طور کلی نتایج این تحقیق نشان داد که الگوریتم SVD با صحت بالایی می‌تواند ژنوتیپ‌های از دست رفته را بازیابی کند. زمانی که درصد ژنوتیپ‌های از دست رفته کمتر باشد توانایی این الگوریتم قابل توجه است. با افزایش درصد ژنوتیپ‌های از دست رفته، توانایی الگوریتم در بازیابی ژنوتیپ‌های از دست رفته کاهش می‌یابد که علت آن کاهش در منابع قابل بهره‌برداری توسط الگوریتم است. همچنین نتایج این تحقیق نشان داد که تحت شرایط ثابت بودن درصد ژنوتیپ‌های از دست رفته، افزایش در اندازه ماتریس ژنوتیپی از طریق افزایش در تعداد نشانگر تعیین ژنوتیپ شده و یا افزایش در تعداد افراد دارای اطلاعات ژنوتیپی، صحت بازیابی ژنوتیپ افزایش می‌یابد. با افزایش MAF صحت بازیابی ژنوتیپ کاهش یافت که میزان کاهش در شرایط تعداد کمتر نشانگر بیشتر بود. استفاده از این روش جهت بازیابی ژنوتیپ‌های از دست رفته در داده‌های واقعی حاصل از منابع مختلف می‌تواند موضوع تحقیقات دیگری باشد.

منابع

1. Berry, D. P., and J. F. Kearney. 2011. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal, 5: 1162-1169.
2. Calus, M. P. L., A. C. Bouwman, J. M. Hickey, R. F., Veerkamp, and H. A. Mulder. 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal, 21:1-11.
3. Cleveland, M. A., and J. M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. Journal of Animal Science, 91: 3583-3592.
4. Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics, 193: 347-365.
5. Donato, M., S. O. Peters Mitchell, S. E. T. Hussain, and I. G. Imumorin. 2013. Genotyping-by-Sequencing (GBS): A novel, efficient and cost-effective genotyping method for cattle using next generation sequencing. PLOS ONE, 8: e62137.
6. Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, and K. Kawamoto. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLOS ONE, 6:e19379.
7. Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science, 95:4114-4129.
8. Goddard, M. E. 2009. Genomic selection: prediction of accuracy and maximization of long term response. Genetica, 136: 245-252.
9. Gorjanc, G., M. A. Cleveland, R. D. Huston, and J. M. Hickey. 2015. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genetics Selection Evolution, 47:12.
10. Heidaritabar, M., M. P. L. Calus, A. Vereijken, A. Martien, M. Groenen, and J. W. M. Bastiaansen. 2015. Accuracy of imputation using the most common sires as reference population in layer chickens. BMC Genet. 16: 101.
11. Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Science, 52:654-663.
12. Li, Y., C. Willer, S. Sanna, and G. Abecasis. 2009. Genotype Imputation. Annual Review of Genomics and Human Genetics, 10: 387-406.
13. Lin, P., S. M. Hartz, Z. Zhang, S. F. Saccone, and J. Wang. 2010. A new statistic to evaluate imputation reliability. PLOS ONE, 5, e9697.

14. Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome wide dense marker maps. Genetics, 157: 1819-1829.
15. Neimann-Sorensen, A., and A. Robertson. 1961. The association between blood groups and several production characters in three Danish cattle breeds. Acta Agriculture Scandinavia, 11: 163-196.
16. Pei, Y. F., J. Li, L. Zhang, C. J. Papasian, and H. W. Deng. 2008. Analyses and comparison of accuracy of different genotype imputation methods. PLOS ONE, 3:e3551.
17. Perry, P. O. 2009. bcv: Cross-Validation for the SVD. R package version 1.0. Available at: http://CRAN.R-project.org/package=bcv/.18-Roshyara, N.B, and M. Scholz. 2015. Impact of genetic similarity on imputation accuracy. BMC Genetics,16:90
18. Roshyara, N. R., K. Horn, H. Kirsten, P. Ahner, and M. Scholz. 2016. Comparing performance of modern genotype imputation methods in different ethnicities. Scientific Reports, 6:34386.
19. Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics, 123: 218-223.
20. Schrooten, C., R. Dassonneville, V. Ducrocq, R. F. Brøndum, M. S. Lund, and J. Chen. 2014. Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina Bovine HD chip. Genetics Selection Evolution, 46(1):10.
21. Su, G., R. F. Brøndum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science, 95: 4657-4665.
22. Technow, F. 2013. hypred: Simulation of genomic data in applied genetics. Available at: http://cran.r-project.org/web/packages/hypred/index.html.
23. Toosi, A., R. L. Fernando, and J. C. Dekkers. 2009. Genomic selection in admixed and crossbred populations. Journal of Animal Science, 88: 32-46.
24. Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. 2001. Missing value estimation methods for DNA microarrays. Bioinformatics, 17:520-525.
25. Vereijken, A. L. J., G. A. A. Albers, and J. Visscher. 2010. Imputation of SNP genotypes in chicken using a reference panel with phased haplotypes. 9th World Conference of Genetics Applied on Livestock Production, Leipzig, Germany. Available at: http://www.wcgalp.org/system/files/proceedings/2010/imputation-snp-genotypes-chicken-using-reference-panel- phased-haplotypes.pdf
26. Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, and D. Gianola. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. Journal of Dairy Science, 93: 5423-5435.
27. Wellmann, R., S. Preuß, E. Tholen, J. Heinkel, K. Wimmers, and J. Bennewitz. 2013. Genomic selection using low density marker panels with application to a sire line in pigs. Genetics Selection Evolution, 45: 28.
28. Weng, Z., Z. Zhang, X. Ding, W. Fu, P. Ma, C. Wang, and Q. Zhang. 2012. Application of imputation methods to genomic selection in Chinese Holstein cattle. Journal of Animal Science and Biotechnology, 3:6.
29. Zhang, Z., Q. Zhang, and X. D. Ding. 2011. Advances in genomic selection in domestic animals. Chinese Science Bulletine, 56: 2655-2663.

Iranian Journal of Animal Science Research
Vol. 10, No. 4, Winter 2019, p. 553-560

نشریه پژوهشهای علوم دامی ایران
جلد ۱۰، شماره ۴، زمستان ۱۳۹۷، ص. ۵۶۰-۵۵۳

Studying the Performance of Intelligent Singular Value Decomposition Algorithm (SVD) in Imputation of Missing Genotypes in Different Scenarios of Number of Marker, Population Size and Minor Allele Frequency

F. Ghafouri-Kesbi[1]- A. Goudarztalejerdi[2*]

Received: 27-07-2017
Accepted: 24-10-2017

Introduction By implementing genomic selection, high accurate estimates of breeding values in newborn individuals could be obtained in the absence of phenotypic records. In genomic selection, selection decisions are based on genomic breeding values predicted from high-density SNP pannels. Dramatic advances in sequencing technologies are providing highly dimensional molecular marker information at low cost. Next generation sequencing protocols such as genotype by sequencing (GBS) technology have been suggested as an efficient and cost-effective genotyping method for genomic selection in cattle. It capable of providing acceptable marker density for genomic selection or genome-wide association studies at roughly one third of the cost of currently available genotyping technologies. However, polymorphic loci scored by GBS can contain a large proportion of missing data across samples because random fragments of the genome are sequenced at low depth, leading some loci to have zero coverage in some individuals. Most analyses require a complete dataset; therefore, marker imputation is a necessary step before GBS data can be used for most purposes such as genomic selection. Order of markers is unknown in GBS data. Therefore, an imputation method which does not require previous information about the order of the markers is needed for imputing GBS data. Nonparametric models from the machine-learning repository have been proposed as an alternative to deal with such situations. These models do not follow a particular parametric design. Several different machine-learning approaches are currently used for genotype imputation and it is important to assess the performance of diverse methodologies and identify the methods that can provide the greatest predictive accuracy in a given population. Singular value decomposition imputation (SVD is capable to impute missed markers in GBS data. The aim of this study was assessing the performance of intelligent SVD algorithm for imputation of missing genotypes.

Materials and Methods A genome consisted of one Morgan chromosome was simulated using the hypred package on which in different scenarios, respectively, 500, 1000, 1500, 2000, 2500 and 3000 SNPs with equal initial frequency of 0.5 were arrayed for 1000 individuals. Coding for each genotype with A1 and A2 alleles were 2 for A1A1, 0 for A2A2 and 1 for A1A2 or A2A1, respectively. Then, in the framework of genotyping by sequencing data (GBS), genotype information of 5%, 10%, 25%, 50%, %75 and 90% of SNPs were masked and then imputed with SVD algorithm. Imputation accuracy (r) was assessed by the percentage of genotypes imputed correctly (number of genotypes correctly imputed/total number of masked genotypes). The effect of number of genotyped individuals (1000 and 2000 individuals), number of genotyped SNPs (500, 1000, 1500, 2000, 2500 and 3000 SNP) and levels of minor allele frequency (MAF) (0.01, 0.05, 0.1, 0.2, 0.3 and 0.4) on imputation accuracy were also studied.

Results and Discussion The SVD imputation accuracy was noticeable. So by increasing the percentage of masked markers up to 50%, SVD was imputed missing genotypes with accuracy equal to 80%. In the scenarios of 70% and 90% of missing genotypes, the accuracy of imputation decreased and was 70% and 48%, respectively. In parallel to increase in the size of the population from 1000 to 2000 individuals, the imputation performance of SVD was increased, especially in the scenarios of 75% and 90% of masked genotypes. In parallel to increase in the number of markers, the imputation accuracy (r) increased in such a way that with increasing the number of markers from 500 to 3000 SNP, the accuracy of imputation increased by almost %10. An inverse relationship was observed between MAF and r in a way that by increasing MAF from 0.01 to 0.40, the accuracy of imputation decreased by 8%. In other words, markers with lower MAF were imputed with higher accuracy.

Conclusion SVD performed well regarding genotype imputation for GBS platforms in a way that missing data can be imputed with reasonable accuracy even if the level of missing data are high; up to 50%and even greater accuracies may result if number of individuals in the population is high and level of MAF of genotyped SNPs is low. Therefore, SVD can be recommended for genotype imputation in genome assisted evaluation.

Keywords: Genotype imputation, SNP, SVD algorithm

1- Department of Animal Science, Faculty of Agriculture, Bu-Ali Sina University, Hamedan, Iran
2- Department of Pathobiology, School of Paraveterinary Science, Bu-Ali Sina University, Hamadan, Iran
(*- Corresponding author email: A.goudarz@basu.ac.ir)