مقاله علمی پژوهشی/نشریه پژوهش‌های علوم دامی/ جلد ۳۱ شماره ۴/ سال ۱۴۰۰/ صفحات ۴۳ تا ۵۷

DOI: 10.22034/AS.2022.38215.1553

بررسی سه روش پویش کل ژنوم در شناسایی جایگاه‌های ژنی با بکارگیری داده شبیه‌سازی شده

فاطمه بیرانوند[1*]، محمد تقی بیگی نصیری[2]، مسعود شیرعلی[3] و محمود شیرعلی[3]

تاریخ دریافت: ۹۸/۱۱/۱۵ تاریخ پذیرش: ۴۰۰/۱۲/۱

[1] دانشجوی دکتری ژنتیک و اصلاح نژاد دام، گروه علوم دامی، دانشگاه رامین خوزستان

[2] استاد، گروه علوم دامی، دانشگاه رامین خوزستان

[3] دکتری ژنتیک و اصلاح نژاد دام، دانشکده کشاورزی، دانشگاه تهران

[*]مسئول مکاتبه: Email: Beiranvand_Fatemeh@yahoo.com

چکیده

زمینه مطالعاتی: شناسایی متغیرهای مؤثر بر صفات کمی در اصلاح نژاد دام اهمیت بالایی دارد. **هدف**: در این پژوهش توانایی سه روش مطالعه پویش کل ژنوم بر اساس تک- چند شکلی تک نوکلئوتیدی یا SSGWAS، مکان یابی وراثت پذیری ناحیه‌ای (RHM) و آزمون پویش سریع مبتنی بر سطح‌بندی (fastBAT) برای شناسایی متغیرهای ژنتیکی مؤثر بر صفات کمی، مورد بررسی قرار گرفته است. **روش کار**: یک جمعیت گاوی با اندازه ۴۰۴۰ رأس دام شبیه‌سازی شد. برای این جمعیت ۳ جفت کروموزم غیر جنسی با تعداد ۲۷۵۸۶ چند شکلی تک نوکلئوتیدی برای هر کروموزم در نظر گرفته شد. شبیه‌سازی در قالب ۳ سناریو با تعداد ۷۵، ۱۵۰ و ۳۰۰ جایگاه صفت کمی و با در نظر گرفتن ۱۰ تکرار برای هر سناریو انجام شد. در بررسی متغیرها ماتریس‌های روابط کل ژنوم و روابط ژنتیکی مبتنی بر شجره در مدل مورد استفاده قرار گرفت. برای مقایسه توانایی روش‌ها در شناسایی QTL ها از معیار تعداد SNP های معنی‌دار در مجاورت با QTL ها استفاده گردید. **نتایج**: از مجموع ۳۰ تکرار شبیه‌سازی شده، در روش SSGWAS تعداد ۱۶ QTL شناسایی شد که ۲ QTL دارای فراوانی آلل نادر یا MAF کوچکتر یا مساوی ۰/۱ بوده و سایر QTL ها با MAF بالاتر از ۰/۱ شناسایی شدند. در روش fastBAT ۱۰۷ ناحیه معنی‌دار شناسایی شد. در این روش تعداد ۱۲۰ QTL در ۳ سناریو شناسایی شد که تعداد ۵۲ QTL با MAF کوچکتر یا مساوی ۰/۱ شناسایی شدند. همه QTL های شناسایی شده در دو روش fastBAT و SSGWAS در روش RHM نیز شناسایی شد. در RHM تعداد ۶۱۲ ناحیه معنی‌دار شناسایی شد. در این روش تعداد ۳۱۶ QTL با MAF کوچکتر یا مساوی ۰/۱ شناسایی شد. **نتیجه‌گیری نهایی**: نتایج این بررسی نشان می‌دهد که روش RHM قابلیت بالاتری نسبت به دو روش دیگر در شناسایی QTL های مؤثر بر واریانس صفت کمی دارد.

واژگان کلیدی: جایگاه صفات کمی، چند شکلی تک نوکلئوتیدی، شبیه‌سازی، مطالعه پویش کل ژنوم

مقدمه

با توجه به پراکنش گسترده SNP[1] ها در سرتاسر ژنوم، این نشانگرها به طور وسیعی در تحقیقات اصلاح نژاد دام مورد استفاده قرار می‌گیرند. این نشانگرها غالباً برای شناسایی QTL[2] ها و تعیین واریانس ژنتیکی صفات پیچیده با استفاده از مطالعات پویش کل ژنوم (GWAS) و پیش بینی مقادیر ژنتیکی صفات اقتصادی مورد استفاده قرار می‌گیرند (ژانگ و همکاران ۲۰۱۵). به طور کلی روش‌های پویش کل ژنوم مبتنی بر دو روش SSGWAS[3] و نشانگرهای چندگانه[4] می‌باشد. روش SSGWAS قادر است تعداد زیادی از متغیرهای معمول مؤثر بر صفات کمی را شناسایی کند. با این وجود سهم بزرگی از واریانس ژنتیکی باقی می‌ماند که با این روش قابل توضیح نیست (شیرعلی و همکاران ۲۰۱۸). در واقع SSGWAS به میزان زیادی وابستگی بین SNP ها و فنوتیپ را کشف کرده است اما این SNP ها تنها بخش کوچکی از وارثت پذیری برآورد شده را تبیین می‌کنند. این یافته‌ها باعث توجه هر چه بیشتر به پایه‌های ژنتیکی صفات کمی و وراثت پذیری گمشده شده است (زیت لان و همکاران ۲۰۱۳ و هیندورف ۲۰۰۹).

وراثت پذیری یک صفت (h^2) سهمی از واریانس صفت در جمعیت است که در اثر تفاوت‌های ژنتیکی افزایشی ایجاد شده است (لی و چو ۲۰۱٤). وراثت پذیری در روش‌های کلاسیک بر اساس ساختار شجره و با استفاده از روابط کوواریانس بین خویشاوندان نزدیک برآورد می‌شود و با استفاده از این روش ممکن است پارامترها به صورت اریب برآورد شود. اریب بودن وراثت پذیری برآورد شده می‌تواند ناشی از نادیده گرفته شدن اثرات ژنتیکی نظیر اپیستازی یا غالبیت و یا اثرات محیطی در برآورد پارامتر باشد (ایوز و همکاران ۱۹۷۸، ایوانس و همکاران ۲۰۱۸ و کلر و کاونتری ۲۰۰۵). وراثت پذیری برآورد شده به وسیله SSGWAS (h^2_{GWAS}) سهمی از واریانس فنوتیپی است که در اثر متغیرهای ژنتیکی به شدت وابسته به فنوتیپ ایجاد می‌گردد (تروف و همکاران ۲۰۱۷). برآورد وراثت پذیری به عنوان یک پارامتر در تفسیر نتایج SSGWAS اهمیت بالایی دارد. در سال‌های اخیر محققان روش‌های جدیدی را برای برآورد وراثت پذیری به صورت مستقیم از تشابه SNP ها یافته‌اند که منجر به شناسایی تشابهات ژنتیکی بین افراد غیر خویشاوند شده است. پارامتر کمی که با این روش‌ها تخمین زده می‌شود بیان کننده بخشی از وراثت پذیری است که به وسیله تعدادی از SNP ها (h^2_{SNP}) ایجاد شده است (لی و چو ۲۰۱٤). با روش GREML[5] مقادیری از واریانس صفت که به وسیله SNP ها توضیح داده می‌شود، قابل برآورد است (یانگ و همکاران ۲۰۱۰). در این روش با استفاده از حداکثر درست نمائی محدود شده (REML[6]) و با ایجاد ماتریس روابط ژنتیکی (GRM[7]) با استفاده از همه SNP ها، واریانس ژنومی برآورد می‌شود. محققین با ایجاد تغییر در روش مولفه‌های واریانس GREML، وراثت پذیری را بر اساس آلل‌های مشترک بین افراد خویشاوند و غیر خویشاوند به واسطه مبدا مشترک (IBD[8]) و شباهت بین آلل‌ها (IBS[9]) برآورد کردند (زیت لان و همکاران ۲۰۱۳). در کنار هم قرار گرفتن این وراثت پذیری‌ها باعث افزایش دقت برآورد وراثت پذیری خاص خواهد شد زیرا هیچگونه فرضیه‌ای در ارتباط با وجود شباهت ژنتیکی، اثرات غالبیت و اپیستازی در این روش وجود ندارد (نولته و همکاران ۲۰۱۷ و زیت لان و همکاران ۲۰۱۳). به طور معمول در صفات کمی سهمی از واریانس فنوتیپی که به وسیله SNP ها توضیح داده می‌شود با تعداد SNP های مجاور

[1] SNP: Single Nucleotide Polymorphism
[2] QTL: Quantitative trait locus
[3] SSGWAS: Single SNP Genome-Wide Association Studies
[4] Multi - markers
[5] Genome based restricted maximum likelihood
[6] Restricted maximum likelihood
[7] Genomic relationship matrix
[8] Identity By Descent
[9] Identity By State

در نواحی ژنومی ارتباط دارد. وراثت پذیری ایجاد شده توسط این نواحی ژنومی به صورت وراثت پذیری ناحیه-ای تعریف می‌شود (زنک و همکاران ۲۰۱۷).

روش RHM[۱] برای شناسایی نواحی ژنومی کوچکی مورد استفاده قرار می‌گیرد که در واریانس فنوتیپی یک صفت سهم بزرگی دارند (ناگامینه و همکاران ۲۰۱۲). این روش با ایجاد پنجره‌هایی شامل تعداد مشخصی SNP، برای غربال ژنوم مورد استفاده قرار می‌گیرد. RHM توانایی بالایی برای شناسایی متغیرهای نادر و آلل‌های چندگانه در یک ناحیه دارد. این روش می‌تواند واریانس QTL را با استفاده از اثرات مستقل چند SNP مجاور با دقت بالایی برآورد کند. در این روش یک مدل مختلط برای برآورد مؤلفه‌های واریانس بر اساس REML مورد استفاده قرار می‌گیرد. بخشی از واریانس ژنتیکی که با روش SSGWAS قابل شناسایی نیست، با روش RHM قابل برآورد می‌باشد (اوموتو و همکاران ۲۰۱۳). برای شناسایی ژن‌های مربوط به صفات پیچیده، آزمون اثرات تجمعی مجموعه‌ای از SNP ها با استفاده از روش آزمون پویش سریع مبتنی بر مجموعه‌ای از SNP ها بسیار کارآمد است (باکشی و همکاران ۲۰۱٦). روش fastBAT[۲] برای محاسبه P-value برای یک مجموعه از SNP ها پیشنهاد شده است. داده‌های مورد استفاده در روش fastBAT خلاصه‌ای از نتایج بدست آمده از SSGWAS یا داده‌های مربوط به عدم تعادل پیوستگی (LD) مربوط به جمعیت مرجع می‌باشد (باکشی و همکاران ۲۰۱٦). در این پژوهش سعی شده است که با استفاده از اطلاعات شبیه‌سازی شده، روش‌های SSGWAS، fastBAT و RHM در شناسایی QTL های مؤثر بر صفات شبیه-سازی شده مورد بررسی و مقایسه قرار گیرد. هدف از این پژوهش مقایسه SNP ها و نواحی شناسایی شده توسط سه روش مورد مطالعه و مقایسه نتایج با QTL های شبیه‌سازی شده و بررسی و تعیین نتایج مثبت کاذب در هر روش است.

مواد و روش‌ها

شبیه‌سازی نشانگرها و جمعیت

در این پژوهش نشانگرها و جمعیت به صورت فرآیند پیشروی در زمان با استفاده از نرم افزار QMSim (سرگلزایی و شنکل ۲۰۰۹) در قالب ۳ سناریو با تعداد ۷۵، ۱۵۰ و ۳۰۰ QTL شبیه‌سازی شد. برای هر کدام از این ۳ سناریو تعداد ۳ جفت کروموزم اتوزوم با طول ۲٤۰ سانتی مورگان (طول کل ژنوم) در ۱۰ تکرار شبیه‌سازی شد. در هر جفت کروموزم ۲۷۵۸٦ SNP دو آللی در نظر گرفته شد که با توزیع یکنواخت بر روی کروموزم‌ها پراکنده شدند. اثرات آللی QTL ها بر اساس توزیع گاما با پارامتر شکل ۰/٤ شبیه‌سازی شد و موقعیت QTL ها بر روی کروموزم به صورت تصادفی در نظر گرفته شد. نرخ جهش SNP ها و QTL ها ۲/۵e−۵ در نظر گرفته شد. ساختار جمعیت شبیه‌سازی شده بر پایه ۲ قسمت مختلف ایجاد گردید: ۱) نسل‌های اولیه: این جمعیت با تعداد ۱۰۰۰ رأس دام در نسل صفر ایجاد شد. اندازه جمعیت در نسل ۱۰۰۰ برابر با ۱۰۰۰ در نظر گرفته شد و پس از ۱۲۰۰ نسل اندازه جمعیت از ۱۰۰۰ به ۲۰۰ دام کاهش یافت. روند کاهشی در اندازه جمعیت به تثبیت LD اولیه و تعادل جهش – رانش در جمعیت‌های اولیه کمک می‌کند (بریتو و همکاران ۲۰۱۱). در فاصله نسل‌های ۲۲۰۰ تا ۲۲۰۱ اندازه جمعیت ثابت در نظر گرفته شد و در نسل ۲۲۰۷ جمعیتی با اندازه ۳۱۷۰۰ دام با تعداد برابر نر و ماده شبیه‌سازی شد. ۲) جمعیت‌های ثانویه: در این مرحله تعداد ٤۰ نر و ۱۰۰۰ ماده از نسل ۲۲۰۷ در جمعیت اولیه برای آمیزش انتخاب شدند. انتخاب ماده‌ها به صورت تصادفی و انتخاب نرها بر اساس ارزش اصلاحی واقعی با شدت انتخاب بالا بود. در این جمعیت تعداد ۱ فرزند به ازای هر ماده با احتمال برابر برای نر و ماده در نظر گرفته شد و ژنوتیپ و فنوتیپ برای ۳ نسل متوالی ایجاد شد. حذف دام‌ها بر اساس ارزش اصلاحی برآورد شده پایین و نرخ جایگزینی نرها و ماده‌ها در هر

[۱] RHM: Regional Heritability Mapping

[۲] fastBAT: fast and flexible set-Based Association Test

نسل به ترتیب ٦٠و ٢٠ درصد در نظر گرفته شد. برای هر سناریو ١٠ تکرار با وراثت پذیری کل ٢٠ درصد شامل ١٠ درصد وراثت پذیری QTL و واریانس فنوتیپی ١ شبیه‌سازی شد.

کنترل کیفیت

در هر سناریو ژنوتیپ‌ها برای تعداد ٨٢٧٥٨ SNP بر روی ٣ جفت کروموزم با کدهای ٠، ٢، ٣، ٤ و ٥ ایجاد شد. در این ژنوتیپ‌ها کد ٠ هموزیگوت برای آلل ١، کد ٢ هموزیگوت برای آلل ٢، کد ٣ هتروزیگوت (آلل اول مربوط به پدر و آلل دوم مربوط به مادر)، کد ٤ هتروزیگوت (آلل اول مربوط به مادر و آلل دوم مربوط به پدر) و کد ٥ نشان دهنده SNP های تعیین ژنوتیپ نشده بود. SNP های حاصل از شبیه‌سازی به کدهای ٠ و ١ و ٢ تبدیل شدند که ٠ برای هموزیگوت مغلوب، ١ برای هتروزیگوت و ٢ برای هموزیگوت غالب بود. ژنوتیپ‌های شبیه‌سازی شده مربوط به جمعیتی با اندازه ٤٠٤٠ دام بود. برای کنترل کیفیت داده‌های ژنوتیپ شده از چند معیار متداول استفاده شد. معیار تعادل هاردی – واینبرگ ٠/٠٠١، مقدار SNP های تعیین ژنوتیپ شده برای هر فرد ٩٥ درصد، میزان ژنوتیپ‌های موجود برای هر SNP ٩٥ درصد و فراوانی آلل نادر ٥ درصد در نظر گرفته شد. در صورت پایین بودن مقادیر نسبت به این معیارها، دام و یا SNP از داده‌ها حذف گردید. سطح LD در جمعیت شبیه‌سازی شده با برآورد توان دوم ضریب همبستگی (R^2) بین هر جفت SNP ارزیابی شد:

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

در این معادله $D = f(AB) - f(A)f(B)$ و $f(AB)$، $f(A)$، $f(a)$، $f(B)$ و $f(b)$ فراوانی هاپلوتیپ‌های AB و آلل‌های A، a، B و b هستند (هیل و روبرتسون ١٩٦٨). این کار با استفاده از نرم افزار PLINK (پورسل و همکاران ٢٠٠٧) و تحت عنوان هرس[1] کردن SNP ها انجام شد. با تنظیم کردن SNP ها، زوج SNP هایی که توان دوم ضریب همبستگی آنها از آستانه تعیین شده برای LD بالاتر بود از آنالیز حذف شدند. برای این منظور پنجره‌هایی با اندازه ٥٠ SNP در طول کروموزوم در نظر گرفته شد و SNP هایی که تقریبا غیر وابسته بودند، انتخاب شدند. میزان آستانه LD (R^2) برابر با ٠/٩ در نظر گرفته شد. برای کنترل کیفیت داده‌ها از نرم افزار PLINK استفاده شد.

برآورد ماتریس‌های روابط ژنومی، روابط ناحیه‌ای ژنومی و ماتریس روابط ژنتیکی مبتنی بر شجره[2]در روش RHM

اثر افزایشی کل ژنوم با بکارگیری همه SNP ها برآورد شد و از آن برای ایجاد ماتریس روابط کل ژنوم (GRM_g) استفاده شد. همچنین اثر افزایشی ناحیه‌ای ژنومی[3]با استفاده از ماتریس روابط ناحیه‌ای ژنومی (GRM_r) برآورد شد که شامل نواحی با تعداد مشخصی SNP بود (ناگامینه و همکاران ٢٠١٢). GRM_g و GRM_r با استفاده از قرابت ژنومی[4]بین افراد جمعیت با استفاده از IBS برآورد گردید (یانگ و همکاران ٢٠١٠):

$$IBS_{ij} = \frac{1}{S}\sum_{k=1}^{S}\frac{(O_{ik}-2P_k)(O_{jk}-2P_k)}{2P_k(1-P_k)}$$

در این معادله IBS_{ij} رابطه ژنومی برآورد شده بین فرد i و فرد j، S تعداد SNP های موجود در ناحیه، O_{ik} و O_{jk} ژنوتیپ‌های i امین و j امین فرد در k امین SNP (کدهای ٠ و ١ و ٢ به ترتیب برای ژنوتیپ‌های AA و AB و BB) و P_k فراوانی آللی در k امین SNP است. برای محاسبه GRM_g و GRM_r بر اساس روابط ژنتیکی بین افراد با استفاده از SNP ها از نرم افزار GCTA نسخه 1.93.0beta (یانگ و همکاران ٢٠١١) استفاده شد. وراثت پذیری ژنومی (h_g^2) که سهمی از واریانس فنوتیپی است که توسط SNP های معمول توضیح داده می‌شود، با استفاده از GRM_g برآورد شد و از GRM_r برای

[1] Prune
[2] Pedigree base genetic relationship
[3] Regional genomic additive
[4] Genomic kinship

برآورد وراثت پذیری ناحیه‌ای (h^2_{RH}) استفاده شد (ناگامینه و همکاران ۲۰۱۲):

$$h^2_g = \frac{\sigma^2_u}{\sigma^2_p}$$

$$h^2_{RH} = \frac{\sigma^2_v}{\sigma^2_p}$$

در این روابط σ^2_u و σ^2_v و σ^2_p به ترتیب واریانس ژنتیکی افزایشی کل ژنوم، واریانس افزایشی ناحیه‌ای ژنومی و واریانس فنوتیپی هستند. σ^2_p حاصلجمع σ^2_u، σ^2_v و واریانس باقیمانده است. زیت لان و همکاران (۲۰۱۳) اثرات ژنتیکی افزایشی وابسته به ساختار شجره را به عنوان یکی از مولفه‌های واریانس مؤثر بر واریانس فنوتیپی برآورد کردند. از این مؤلفه برای برآورد h^2_{PED} استفاده می‌شود که سهمی از واریانس فنوتیپی است که توسط اثرات ژنتیکی موجود در ساختار یک خانواده توضیح داده می‌شود. در این روش ماتریس روابط خویشاوندی از GRM_g استخراج می‌شود. در این پژوهش برای محاسبه ماتریس روابط ژنتیکی مبتنی بر شجره و h^2_{PED} یک GRM_{PED} بر اساس رابطه خویشاوندی بین افراد تشکیل شد. GRM_{PED} با استفاده از بسته نرم‌افزاری Pedigree نسخه ۱/٤ (کاستر و همکاران ۲۰۱۳) مربوط به نرم افزار RStudio نسخه ۱/۱/٤٤۲ (شرکت RStudio ۲۰۱۳) محاسبه شد. h^2_{PED} بر اساس مدل بی نهایت (فیشر ۱۹۱۸) به صورت نسبت واریانس ژنتیکی افزایشی به واریانس کل $(h^2_{PED} = \frac{\sigma^2_a}{\sigma^2_p})$ محاسبه شد. برای مکان یابی وراثت پذیری و برآورد مولفه‌های واریانس از پنجره‌هایی با اندازه ۵۰ SNP با در نظر گرفتن همپوشانی ۲۵ SNP بین دو پنجره متوالی در طول ژنوم استفاده شد. برای اجرای RHM از نرم افزار REACTA (سبامانوس و همکاران ۲۰۱٤) استفاده شد. در RHM مدل مختلط بر اساس GREML بوده و برای هر ناحیه لگاریتم درستنمایی محاسبه می‌شود. در ۱۰ تکرار شبیه‌سازی شده برای هر سناریو حداکثر و حداقل تعداد پنجره‌های ایجاد شده به ترتیب ۹۵۵ و ۸۲۰ برای سناریو QTL ۷۵، ۹۰٤ و ۷۸۸ برای سناریو QTL ۱۵۰ و ۸۹٤ و ۸۱۵ پنجره برای سناریو QTL ۳۰۰ بود. تعیین معنی‌داری هر ناحیه با توجه به آزمون نسبت لگاریتم درستنمایی برآورد شده برای آن ناحیه - = LRT (2ln(logL$_0$- logL)) انجام شد. در معادله LRT، logL معادل لگاریتم درستنمائی مدل کامل و logL$_0$ معادل لگاریتم درستنمائی مدل کاهش یافته است. برای تعیین سطح معنی‌داری از تصحیح بنفرونی بر اساس تعداد نواحی ژنومی بدون همپوشانی در سطح ۵ درصد استفاده شد. مدل مختلط مورد استفاده در این روش به صورت زیر بود:

$$\mathbf{y} = 1_n \mu + \mathbf{Xu} + \mathbf{Hv} + \mathbf{Zw} + \mathbf{e}$$

در این معادله $\mathbf{y}$ بردار مقادیر فنوتیپی، $\mathbf{X}$، $\mathbf{H}$ و $\mathbf{Z}$ ماتریس‌های طرح برای اثرات تصادفی، 1_n یک بردار شامل عدد ۱ و $\boldsymbol{\mu}$ میانگین است. $\mathbf{u} \sim N(0, \mathbf{G}\sigma^2_u)$ اثر ژنتیکی افزایشی کل ژنوم، $\mathbf{v} \sim N(0, \mathbf{P}\sigma^2_v)$ اثر ژنتیکی افزایشی شجره، $\mathbf{w} \sim N(0, \mathbf{Q}\sigma^2_w)$ اثر ژنتیکی افزایشی ناحیه‌ای ژنومی و $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$ اثر باقیمانده است. ماتریس‌های $\mathbf{G}$، $\mathbf{P}$، $\mathbf{Q}$ و $\mathbf{I}$ به ترتیب ماتریس روابط ژنتیکی کل ژنوم، ماتریس روابط ژنتیکی مبتنی بر شجره، ماتریس روابط ناحیه‌ای ژنومی با استفاده از نشانگرهای هر ناحیه از ژنوم و ماتریس یکه هستند. ماتریس‌های $\mathbf{G}$، $\mathbf{P}$ و $\mathbf{Q}$ بر اساس ضرایب خویشاوندی ژنومی، خویشاوندی مبتنی بر شجره و خویشاوندی ژنومی ناحیه‌ای محاسبه شدند (اوموتو و همکاران ۲۰۱۳).

آنالیز SSGWAS

برای داده‌های شبیه‌سازی شده آنالیزهای SSGWAS با روش MLMA[1] (یو و همکاران ۲۰۰٦) و با استفاده از نرم افزار GCTA انجام شد. در MLMA با هر بار بررسی ژنوم تنها یک SNP مورد آزمون قرار می‌گیرد (یو و همکاران ۲۰۰٦). مدل MLMA مورد استفاده در

[1] Mixed Linear Model association Analysis

GCTA به صورت زیر بوده که در این مدل اثرات SNP ها به عنوان اثرات تصادفی در نظر گرفته می‌شود (اوموتو و همکاران ۲۰۱۳):

$$\mathbf{y} = 1_n \mu + \mathbf{xv} + \mathbf{Hu} + \mathbf{Zw} + \mathbf{e}$$

در این معادله $\mathbf{y}$ بردار $n\times 1$ فنوتیپ‌ها است و n اندازه نمونه است، $\mathbf{u} \sim \mathbf{N}(0, \mathbf{G}\sigma_{\mathbf{u}}^2)$ اثر ژنتیکی افزایشی کل ژنوم، $\mathbf{v} \sim \mathbf{N}(0, \mathbf{P}\sigma_{\mathbf{v}}^2)$ اثر ژنتیکی افزایشی شجره، $\mathbf{w} \sim \mathbf{N}(0, \mathbf{Q}\sigma_{\mathbf{w}}^2)$ اثر ژنتیکی افزایشی ناحیه‌ای ژنومی و $\mathbf{e} \sim \mathbf{N}(0, \mathbf{I}\sigma_{\mathbf{e}}^2)$ اثر باقیمانده است. در این معادله بردار **xv** حاوی تک SNP هایی است که در برآورد اثر SNP ها مورد استفاده قرار می‌گیرند. برآورد اثر SNP ها در MLMA با وجود GRM_g و GRM_{PED} در مدل صورت گرفت. در SSGWAS برای تعیین سطح معنی-داری از تصحیح بنفرونی بر اساس تعداد SNP ها و برای ایجاد سطح معنی‌داری ۵ درصد استفاده شد.

آنالیز fastBAT با استفاده از نتایج SSGWAS

برای بررسی نتایج حاصل از SSGWAS با استفاده از روش fastBAT از نرم افزار GCTA استفاده شد. در این روش با فرض اینکه $z = \{z_i\}$ بردار ارقام z[۱] برای مجموعه‌ای از SNP های حاصل از SSGWAS یا آنالیز فراگیر باشد، فرض صفر عدم وجود وابستگی بین SNP ها و صفت بوده و z دارای توزیع نرمال چند متغیره است ($z \sim MVN(0, R)$) که R ماتریس همبستگی LD برای SNP ها است) (باکشی و همکاران ۲۰۱۶). SNP های تعیین ژنوتیپ شده و P-value های حاصل از SSGWAS به عنوان ورودی در روش fastBAT مورد استفاده قرار گرفت. برای این آنالیز، SNP ها در مجموعه‌هایی با طول ۵۰ و همپوشانی ۲۵ SNP و با MAF[۲] بزرگتر مساوی ۰/۰۵ برای ۳ کروموزم در نظر گرفته شد. برای هر مجموعه بالاترین مقدار P-value مشخص شد و با استفاده از تصحیح بنفرونی مجموعه‌های SNP معنی‌دار مشخص شدند.

نتایج

نتایج شبیه‌سازی

در این پژوهش ۳ سناریو شامل ۷۵، ۱۵۰ و ۳۰۰ QTL، هر کدام در ۱۰ تکرار برای ۳ کروموزم اتوزوم شبیه-سازی شد. در این شبیه‌سازی واریانس فنوتیپ کمی ۱ و وراثت پذیری کل ۲۰ درصد شامل ۱۰ درصد وراثت پذیری QTL و مابقی شامل وراثت پذیری چند ژنی در نظر گرفته شد. در ۳۰ تکرار مورد بررسی، تعداد ۸۲۷۵۸ SNP شبیه‌سازی شد که پس از کنترل کیفیت تعداد مشخصی SNP در هر سناریو برای مطالعه مورد استفاده قرار گرفت. حداقل و حداکثر تعداد SNP موجود در آنالیز پس از کنترل کیفیت به ترتیب ۱۹۶۶۲ و ۲۳۸۱۷ SNP بود. میانگین MAF برای SNP های کنترل کیفیت شذه در ۳ سناریو به ترتیب ۰/۰۷۱۴، ۰/۰۶۹۷ و ۰/۰۶۹۲ و میانگین MAF برای QTL های شبیه‌سازی شده در ۳ سناریو به ترتیب ۰/۱۰۵، ۰/۰۸۷ و ۰/۰۸۷ محاسبه گردید. برای بررسی روش‌های مورد توجه، GRM_g و GRM_{PED} با ابعادی بر اساس تعداد دام موجود در شجره تشکیل شد. دامنه عناصر قطری GRM_g از ۰/۶۹ تا ۱/۸۸ و دامنه عناصر غیر قطری آن از ۰/۴۵- تا ۱/۳۷ در ۳ سناریو بدست آمد. همچنین دامنه عناصر قطری GRM_{PED} از ۱ تا ۱/۳۷۵ و دامنه عناصر غیر قطری این ماتریس از ۰ تا ۱ در ۳ سناریو بود.

نتایج SSGWAS

نتایج MLMA بر اساس P-value در سطح ۵ درصد با استفاده از تصحیح بنفرونی تنظیم شد. موقعیت SNP های معنی‌دار نسبت به نزدیکترین QTL بر اساس فاصله فیزیکی مورد بررسی قرار گرفت و SNP های معنی‌داری که در فاصله ۵۰ SNP از QTL ها قرار داشتند، به عنوان SNP های معنی‌دار تعیین شدند. در هر سناریو تعداد QTL های شناسایی شده توسط SNP های معنی‌دار مشخص و همچنین SNP های با بالاترین میزان P-value تعیین شد. در جدول ۱ تعداد QTL های شناسایی شده

[۱] z-statistics

[۲] Minor Allele Frequency

توسط روش SSGWAS همراه با دامنه MAF برای این QTL ها، دامنه و میانگین واریانس ژنتیکی تبیین شده توسط SNP ها و QTL های معنی‌دار بر حسب درصد گزارش شده است. از مجموع ۷۵۰، ۱۵۰۰ و ۳۰۰۰ QTL شبیه‌سازی شده در ۱۰ تکرار مربوط به ۳ سناریو به ترتیب ۹، ۱۱ و ۱۱ QTL در SSGWAS شناسایی شد که از این تعداد به ترتیب ۵، ۹ و ۱ QTL به صورت مثبت کاذب بودند. معیار به کار رفته در تعیین QTL های مثبت کاذب، عدم وجود QTL معنی‌دار در پنجره‌هایی با اندازه ۵۰ SNP، قبل و بعد از QTL معنی‌دار بود. برای هر تکرار پس از شناسایی SNP های معنی‌دار، واریانس ژنتیکی تبیین شده توسط این SNP ها با استفاده از رابطه $2pq\left[a+d\left(q-p\right)\right]^2$ (فالکنر و مک کی ۱۹۹۶) برآورد شد. بالاترین میزان واریانس برآورد شده (۱۴/۹ درصد) مربوط به SNP های تکرار ۳ در سناریو QTL ۳۰۰ بود. در سناریو QTL ۱۵۰، QTL شماره ۱۲۵ بر روی کروموزم ۳ در بین ۱۰ تکرار در ۲ تکرار مختلف شناسایی شد که به عنوان QTL ثابت در نظر گرفته شد. این QTL ۵/۲۸ درصد از واریانس ژنتیکی را تبیین می-کند. در سناریو QTL ۷۵، QTL های شناسایی شده ۱۲/۲۲ درصد واریانس، در سناریو QTL ۱۵۰، ۸/۴۲ درصد واریانس و در سناریو QTL ۳۰۰، QTL های شناسایی شده در مجموع ۳۰/۳۷ درصد از واریانس را تبیین کردند. واریانس تبیین شده توسط SNP ها در این سناریوها با واریانس تبیین شده توسط QTL ها برابر بود. MAF برای QTL های معنی‌دار، در ۳ سناریو در جدول شماره ۱ نشان داده شده است. در تکرار ۷ از سناریو QTL ۳۰۰ میانگین MAF مربوط به QTL ها (۰/۰۷) از میانگین MAF سناریو پایینتر بود و در سایر موارد QTL های شناسایی شده دارای MAF بالاتری نسبت به میانگین MAF در سناریوها (به ترتیب ۰/۱۰۵، ۰/۰۸۷ و ۰/۰۸۷ برای ۷۵، ۱۵۰ و ۳۰۰ QTL) بودند. با استفاده از نرم افزار GCTA وراثت پذیری صفت برآورد گردید. در این پژوهش h^2_{GWAS} در دامنه ۰/۱۴۳±۰/۰۲۴ تا ۰/۲۶۳±۰/۰۲۹ بود. در جدول شماره ۲ h^2_{GWAS} و وراثت پذیری برآورد شده مربوط به ماتریس‌های GRM_g و GRM_{PED} مربوط به تکرارهایی با حداقل و حداکثر میزان h^2_{GWAS} نشان داده شده است. همانطور که مشاهده می‌گردد، بالاترین h^2_{GWAS} مربوط به سناریو QTL ۱۵۰ و کمترین h^2_{GWAS} مربوط به سناریو QTL ۷۵ است.

Table 1- Number of detected QTLs, QTLs MAF range, range and mean of genetic variance explained by detected QTLs and SNPs in 3 scenario in SSGWAS

scenario	N_Q	QTL MAF range	SNP G_V range (%)	SNP G_V mean (%)	QTL G_V range (%)	QTL G_V mean (%)
75 QTL	4	0.15 to 0.49	0.22 to 7.55	3.06	0.22 to 7.55	3.06
150 QTL	2	0.25 to 0.47	3.14 to 5.28	4.21	3.14 to 5.28	4.21
300 QTL	10	0.069 to 0.43	2.75 to 14.9	6.08	2.75 to 14.9	6.08

N_Q: Number of QTLs, G_V: genetic variance explained.

Table 2- Maximum and minimum of h^2_{PED}, h^2_g, h^2_{GWAS} estimated by SSGWAS

scenario		$h^2_{PED} \pm SE$	$h^2_g \pm SE$	$h^2_{GWAS} \pm SE$
75 QTL	Max	0.105±0.025	0.112±0.018	0.216±0.026
	Min	0.057±0.023	0.086±0.015	0.143±0.024
150 QTL	Max	0.149±0.028	0.115±0.018	0.263±0.029
	Min	0.067±0.023	0.095±0.017	0.161±0.025
300 QTL	Max	0.12±0.027	0.108±0.018	0.227±0.028
	Min	0.095±0.025	0.08±0.015	0.174±0.026

نتایج RHM و fastBAT

با استفاده از روش RHM با پنجره‌هایی به اندازه ۵۰ SNP و با همپوشانی ۲۵ SNP، وراثت پذیری ناحیه‌ای برآورد شد. معنی‌داری هر پنجره بر اساس آزمون نسبت درستنمائی[1] یا LRT مورد ارزیابی قرار گرفت. پس از تصحیح LRT با استفاده از توزیع کای اسکور، از تصحیح بنفرونی برای تعیین سطح معنی‌داری کل ژنوم در سطح ۵ درصد استفاده شد. دامنه LRT در سناریوها از ۰/۷ تا ۶۱/۷۸ و به طور میانگین ۱/۴۰۳، ۱/۳۰۹ و ۱/۲۴۲ برای سناریوهای ۷۵، ۱۵۰ و ۳۰۰ QTL بود. با استفاده از مقادیر LRT برای هر تکرار، نواحی معنی‌دار در سطح ژنوم مشخص شد و به طور متوسط تعداد ۳۹/۲، ۳۶/۲ و ۳۸/۶ ناحیه معنی‌دار در هر تکرار به ترتیب برای ۳ سناریو ۷۵، ۱۵۰ و ۳۰۰ QTL شناسایی شد. در جدول شماره ۳ QTL های معنی‌دار (در سطح ۵ درصد) شناسایی شده و دامنه MAF مربوط به آنها نشان داده شده است. در روش RHM به ترتیب تعداد ۱۰۴، ۱۷۵ و ۳۳۳ QTL معنی‌دار برای این ۳ سناریو (مجموع ۱۰ تکرار) شناسایی شد که از این تعداد به ترتیب ۵، ۲ و ۲ QTL به صورت مثبت کاذب بودند. برای هر QTL معنی-دارترین SNP مجاور آن مشخص شد. همانطور که در جدول ۳ نشان داده شده است، SNP ها و همچنین QTL های سناریوی ۷۵ QTL به طور تقریبی بین ۷ تا ۳۹ درصد (میانگین ۱۰ تکرار) از واریانس ژنتیکی را تبیین می‌کنند. در سناریو دوم واریانس ژنتیکی تبیین شده در محدوده ۱۹ تا ۴۶ درصد (میانگین ۱۰ تکرار) واریانس و در سناریو ۳۰۰ QTL در ۴ تکرار میزان واریانس برآورد شده بالاتر از ۵۰ درصد بود و برآورد واریانس ژنتیکی بالاتر از میزان مورد انتظار بود. در ۶ تکرار دیگر این سناریو میزان واریانس بیان شده بین ۱۷ تا ۴۶ درصد (میانگین ۶ تکرار) برآورد شد. میانگین MAF مربوط به QTL های شناسایی شده در RHM به ترتیب ۰/۱۹۷، ۰/۱۷ و ۰/۱۴۴ برای ۳ سناریو بود. با استفاده از ماتریس GRM_g و GRM_{PED} و با روش GREML برآورد h_g^2 و h_{PED}^2 و h_{RH}^2 انجام شد. میانگین و انحراف معیار پارامترهای h_g^2 و h_{PED}^2 و h_{RH}^2 در ۱۰ تکرار سناریو ۷۵ QTL به ترتیب ۰/۰۹۳۹±۰/۰۰۰۶، ۰/۰۹۳۱±۰/۰۰۰۱۵ و ۰/۰۰۲۵±۰/۰۰۰۴ برآورد گردید. میانگین و انحراف معیار پارامترهای h_g^2 و h_{PED}^2 و h_{RH}^2 در ۱۰ تکرار سناریو ۱۵۰ QTL به ترتیب ۰/۰۹۵۳±۰/۰۰۰۲۳، ۰/۰±۱۰۵/۰۰۰۱۲، ۰/۰۰۱۷±۰/۰۰۰۲۶ و در ۱۰ تکرار سناریو ۳۰۰ QTL به ترتیب ۰/۰۹۹۴±۰/۰۰۰۲، ۰/۰±۱۰۱/۰۰۰۲ و ۰/۰۰۱۵±۰/۰۰۰۲ بود. در جدول شماره ۴ میزان این پارامترها برای تکرارهایی با حداقل و حداکثر میانگین h_{RH}^2 در هر سناریو نشان داده شده است. P-value برای SNP ها با روش MLMA و با وجود GRM_g و GRM_{PED} در مدل محاسبه شد. در آنالیزهای fastBAT طول هر پنجره شامل ۵۰ SNP (همپوشانی بین دو پنجره متوالی ۲۵ SNP) بود که برای هر پنجره یک SNP با بالاترین مقدار معنی‌داری و مقدار P-value مربوط به آن مشخص گردید. بعد از تصحیح بنفرونی سطوح معنی‌داری P-value برای کل ژنوم ($p <$ ۰/۰۵) برای هر تکرار مشخص شد. در غربال ژنوم با استفاده از روش fastBAT به ترتیب ۴۸، ۴۲ و ۱۷ ناحیه معنی‌دار برای سناریوهای ۷۵، ۱۵۰ و ۳۰۰ QTL شناسایی شد. QTL هایی که در این روش شناسایی شد به ترتیب ۳۴، ۴۲ و ۴۴ QTL برای ۳ سناریو بود و در هیچکدام از سناریوها QTL مثبت کاذب شناسایی نشد. در جدول شماره ۵ QTL های شناسایی شده با روش fastBAT و دامنه MAF مربوط به آنها، دامنه و میانگین واریانس ژنتیکی تبیین شده توسط SNP ها و QTL ها بر حسب درصد برای ۳ سناریو نشان داده شده است. در سناریو ۱۵۰ QTL در هر تکرار به طور میانگین ۹/۶۹ درصد واریانس تبیین می‌شود که نسبت به دو سناریو دیگر درصد واریانس تبیین شده بالاتر است. میانگین

[1] Likelihood Ratio Test

تقریبی MAF برای QTL های شناسایی شده به ترتیب ۰/۲٤، ۰/۲۱ و ۰/۱٦ برای سناریوهای ۷۵، ۱۵۰ و ۳۰۰ QTL است. در نتایج fastBAT در سناریوهای ۷۵ و ۳۰۰ QTL یک QTL ثابت و در سناریو ۱۵۰ QTL تعداد ٤ QTL ثابت در ٤ تکرار شناسایی شد.

Table 3- Number of detected QTLs, QTLs MAF range, range and mean of genetic variance explained by detected QTLs and SNPs in RHM

scenario	N_Q	QTL MAF range	SNP G_V range (%)	SNP G_V mean (%)	QTL G_V range (%)	QTL G_V mean (%)
75 QTL	99	0.149 to 0.237	7.26 to 39.04	22.44	7.26 to 39.04	22.44
150 QTL	173	0.118 to 0.217	0.19 to 46.16	32.71	0.19 to 46.16	32.71
300 QTL	331	0.097 to 0.159	17.27 to 46.86	48.71	17.27 to 46.86	48.71

N_Q: Number of QTLs, G_V: genetic variance explained.

Table 4- Maximum and minimum mean of h^2_{PED}, h^2_g, h^2_{RH} estimated by RHM

scenario		$h^2_{PED} \pm SE$	$h^2_g \pm SE$	$h^2_{RH} \pm SE$
75 QTL	Max	0.104±0.00012	0.114±0.0012	0.0052±0.0008
	Min	0.055±0.0002	0.096±0.0002	0.00092±0.0002
150 QTL	Max	0.13±8.98e-05	0.067±0.0004	0.0023±0.0005
	Min	0.125±0.0002	0.087±0.0002	0.0013±0.0002
300 QTL	Max	0.09±0.00017	0.108±0.0004	0.002±0.0005
	Min	0.093±0.00015	0.079±0.00014	0.0009±8.23e-05

Table 5- Number of detected QTLs, QTLs MAF range, range and mean of genetic variance explained by detected QTLs and SNPs in fastBAT

scenario	N_Q	QTL MAF range	SNP G_V range (%)	SNP G_V mean (%)	QTL G_V range (%)	QTL G_V mean (%)
75 QTL	34	0.078 to 0.456	0.3 to 19.91	7.94	0.3 to 19.91	7.94
150 QTL	42	0.125 to 0.312	3.15 to 15.97	9.69	3.15 to 15.97	9.69
300 QTL	44	0.018 to 0.233	0.0042 to 16.38	6.63	0.0042 to 16.38	6.63

N_Q: Number of QTLs, G_V: genetic variance explained.

بحث

در این پژوهش ۳ روش آماری متداول در بررسی صفات پیچیده را از نظر شناسایی QTL ها و توانایی تبیین واریانس ژنتیکی مورد ارزیابی قرار دادیم. برای این منظور داده‌های شبیه‌سازی شده در قالب ۳ سناریو با تعداد ۷۵، ۱۵۰ و ۳۰۰ QTL ایجاد شد. تعیین عناصر قطری برای ماتریس روابط ژنتیکی کل ژنوم و ماتریس روابط ژنتیکی مبتنی بر شجره انجام شد. تعیین عناصر قطری در ماتریس‌های روابط ژنتیکی در بسیاری موارد به شناسایی دام‌های نامناسب و عدم انتخاب آنها در نسل بعد کمک خواهد کرد. در توزیع و اندازه عناصر قطری ماتریس‌های GRM_g و GRM_{PED} تفاوت‌هایی وجود دارد. این تفاوت‌ها ناشی از این است که در ماتریس GRM_{PED} روابط مورد انتظار شناسایی نمی‌شود (سیمون و همکاران ۲۰۱۱). به عبارت دیگر GRM_g همواره روابط بیشتری نسبت به GRM_{PED} شناسایی می‌کند. در اینگونه موارد استفاده از فنوتیپ‌های بیشتر در داده‌های واقعی به افزایش دقت برآورد کمک خواهد

کرد (هیز و همکاران ۲۰۰۹). آنالیز SSGWAS با وجود دو ماتریس GRM_g و GRM_{PED} در مدل خطی مختلط انجام شد. در سناریوهای مورد بررسی با افزایش تعداد QTL تعداد SNP های معنی‌دار شناسایی شده با روش SSGWAS کاهش یافت. این نتایج نشان می‌دهد که با توجه به ثابت بودن وراثت پذیری، با افزایش تعداد QTL ها در ژنوم سهم هر QTL در تبیین واریانس ژنتیکی کاهش یافته است. در جدول شماره ٦ مقایسه ۳ روش از نظر تعداد QTLهای شناسایی شده، تعداد QTL های مثبت کاذب، تعداد QTL های ثابت و تعداد QTL های شناسایی شده با MAF کوچکتر یا مساوی ۰/۱ نشان داده شده است. در هر ۳ سناریو و در همه تکرارها، واریانس تبیین شده به وسیله SNP ها با واریانس تبیین شده به وسیله QTL ها برابر بود. یکسان بودن واریانس تبیین شده توسط SNP ها و QTL ها نشان می‌دهد که در صورت بالا بودن LD بین SNP ها و QTL ها روش-های SSGWAS، RHM و fastBAT توانایی بالاتری برای برآورد واریانس دارند. در SSGWAS تعداد QTL کمتری نسبت به ۲ روش دیگر شناسایی شد و حداکثر واریانس تبیین شده توسط QTL ها ١٤/٩ درصد بود. برخلاف ۲ روش دیگر، در روش SSGWAS تنها در تعدادی از تکرارها QTL های معنی‌دار شناسایی شد. همچنین در این روش درصد QTL های مثبت کاذب نسبت به ۲ روش دیگر بالاتر بود. QTL های شناسایی شده در SSGWAS در همه موارد به استثنای تکرار ۷ سناریوی QTL ۳۰۰ دارای MAF و همچنین میانگین MAF بالاتر از ۰/۱ بودند. متفاوت بودن واریانس ژنتیکی تبیین شده توسط SNP ها و همچنین QTL ها در SSGWAS و RHM می‌تواند به این دلیل باشد که در روش SSGWAS، SNP ها به عنوان اثرات ثابت در مدل قرار می‌گیرند در حالی که در RHM، SNP ها به عنوان اثرات تصادفی در نظر گرفته می‌شود بنابراین توابع درستنمائی برای دو روش متفاوت خواهد بود (ویسلر و همکاران ۲۰۱۷). در RHM تعداد QTL های شناسایی شده با MAF پایینتر از ۰/۱ نسبت به دو روش دیگر بیشتر بود که مطابق با نتایج برخی از پژوهش‌های دیگر بود (اوموتو و همکاران ۲۰۱۳). در RHM با افزایش تعداد QTL در سناریو، تعداد QTL های با MAF کمتر یا مساوی ۰/۱ افزایش یافته و به ترتیب ٤٢، ۹۱ و ۱۸۳ QTL با MAF کوچکتر یا مساوی ۰/۱ در سناریوهای ۷۵، ۱۵۰ و ۳۰۰ QTL شناسایی شد. QTL هایی که در دو روش fastBAT و SSGWAS شناسایی شدند، در RHM نیز شناسایی شدند. شیرعلی و همکاران (٢٠١٦) دو روش SSGWAS و RHM را با بکارگیری اطلاعات مربوط به غلظت لیپوپروتئین‌های خون در یک جمعیت مقایسه کردند. این محققان تائید کردند که روش RHM توانایی بالاتری در شناسایی نواحی علّی شامل نواحی حاوی متغیرهای حیاتی ژنوتیپ نشده دارد. بالا بودن توانایی شناسایی QTL ها در روش RHM نسبت به روش‌های آنالیز مبتنی بر ژن[1] در مطالعات دیگری نشان داده شده است (اوموتو و همکاران ۲۰۱۳).

همه QTL های شناسایی شده در SSGWAS به استثنای ۲ QTL (١٣٦ و ١٣٧) در روش fastBAT نیز شناسایی شد. در fastBAT تعداد QTL های شناسایی شده از SSGWAS بیشتر و از RHM کمتر بود. در این روش تعداد QTL های شناسایی شده در سناریو ۱۵۰ QTL بیشتر از دو سناریوی دیگر و تعداد QTL های شناسایی شده با MAF کوچکتر یا مساوی ۰/۱ در سناریو ۳۰۰ QTL بیشتر از دو سناریوی دیگر بود. در fastBAT برخلاف دو روش دیگر هیچ یک از QTL های شناسایی شده به صورت مثبت کاذب نبود.

مقایسه ۳ روش از نظر میزان تبیین واریانس ژنتیکی توسط QTL ها نشان داد که روش RHM با شناسایی تعداد بیشتری QTL در هر تکرار قابلیت بالاتری نسبت به ۲ روش دیگر دارد. تعداد زیادی از QTL ها و نواحی شناسایی شده با RHM توسط دو روش دیگر شناسایی

[1] Gene-based association approaches

نشدند که نشان دهنده قابلیت بالاتر این روش است. واریانس ژنتیکی تبیین شده توسط QTL های شناسایی شده در SSGWAS در دامنه ۰/۲۲ تا ۱۴/۹ درصد، در fastBAT در دامنه ۰/۰۰۴۲ تا ۱۹/۹۱ درصد و در RHM در دامنه ۷/۲۶ تا ٤٦/٨٦ درصد برآورد شد. همانطور که در جدول ۳ نشان داده شده است، در برخی از تکرارهای سناریوی QTL ۳۰۰ مقدار واریانس ژنتیکی بیش از ۵۰ درصد برآورد گردید. h^2_{GWAS} برآورد شده (جدول ۲) در این پژوهش در دامنه ۰/۱٤۳ تا ۰/۲٦۳ بود. در تکرار ۷ از سناریو QTL ۱۵۰ برآورد h^2_{GWAS} (۰/۲٦۳) بالاتر از میزان وراثت پذیری شبیه‌سازی (۰/۲) بود. مقادیر وراثت پذیری برآورد شده در دو روش SSGWAS و RHM در ۳ سناریو به صورت یکسان بود. با توجه به استفاده از GRM های یکسان در هر دو روش، انتظار می‌رود وراثت پذیری برآورد شده یکسان باشد.

Table 6- Comparison of SSGWAS, RHM and fastBAT methods based on detected QTLs

Scenario	N_Q	Method	Number of detected QTL	Number of detected false positive QTL	Number of detected stable QTL	Number of detected QTL (MAF<=0.1)
75 QTL	750	SSGWAS	9	5	-	-
		RHM	104	5	31	42
		fastbat	34	-	1	9
150 QTL	1500	SSGWAS	11	9	1	-
		RHM	175	2	49	91
		fastbat	42	-	4	18
300 QTL	3000	SSGWAS	11	1	-	2
		RHM	333	2	87	183
		fastbat	44	-	1	24

N_Q: Number of QTLs in 10 replication.

نتیجه گیری

در این پژوهش با بررسی ۳ روش SSGWAS، RHM و fastBAT مشخص گردید روش RHM توانایی بالاتری برای شناسایی QTL های معنی‌دار دارد. همچنین مشخص گردید QTL هایی که دارای MAF کوچکتر یا مساوی ۰/۱ هستند در RHM نسبت به ۲ روش دیگر بیشتر شناسایی شدند. نتایج این پژوهش نشان داد که با افزایش تعداد QTL احتمال شناسایی QTL در RHM افزایش یافت. اکثر QTL های شناسایی شده در SSGWAS در fastBAT شناسایی شدند و همه QTL های شناسایی شده در SSGWAS و fastBAT در RHM نیز شناسایی شدند. در این پژوهش مشخص گردید که با استفاده از RHM توانایی شناسایی QTL ها و نواحی معنی‌دار مؤثر بر واریانس ژنتیکی افزایش می‌یابد.

منابع مورد استفاده

Bakshi A Zhu ZH Anna AE Vinkhuyzen W Hill D McRae AF Visscher PM and Yang J, 2016. Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. Scientific Reports 6: 32894.

Brito FV Braccini Neto J Sargolzaei M Cobuci JA and Schenkel FS, 2011. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. BMC Genetics 12: 80.

Cebamanos L Gray A Stewart I and Tenesa A, 2014. Regional Heritability Advanced Complex Trait Analysis for GPU and Traditional Parallel Architectures. Bioinformatics 30(8): 1177-1179.

Coster A, 2013. https://CRAN.R-project.org/package=pedigree.

Eaves LJ Last KA Young PA and Martin NG, 1978. Model-fitting approaches to the analysis of human behaviour. Heredity (Edinb) 41: 249–320.

Evans LM Tahmasbi R Vrieze SI Abecasis GR Das S Gazal S and Keller MC, 2018. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. Nature Genetics 50(5): 737-745.

Falconer DS and Mackay TFC, 1996. Introduction to Quantitative Genetics, 4th edn. Pearson Education Limited: Harlow, UK.

Fisher RA, 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52: 399-433.

Hayes BJ Visscher PM Goddard ME, 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research 91: 47-60.

Hill WG and Robertson A, 1968. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38: 226-231.

Hindorff LA Sethupathy P Junkins HA Ramos EM Mehta JP Collins FS and Manolio TA, 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceeding of the National Academy of Sciences 106(23): 9362–9367.

Keller MC and Coventry WL, 2005. Quantifying and addressing parameter indeterminacy in the classical twin design. Twin Research and Human Genetics 8: 201–213.

Lee JJ and Chow CC, 2014. Conditions for the validity of SNP-based heritability estimation. Human Genetics 133(8): 1011–1022.

Nagamine Y Pong-Wong R Navarro P Vitart V Hayward C Rudan I Campbell H Wilson J Wild S Hicks AA Pramstaller PP Hastie N Wright AF and Haley CS, 2012. Localising loci underlying complex trait variation using Regional Genomic Relationship Mapping. PLOS ONE 7(10): e46501.

Nolte IM Jansweijer JA Riese H Asselbergs FW Harst PVD Spector TD Pinto YM Snieder H and Jamshidi Y, 2017. A Comparison of Heritability Estimates by Classical Twin Modeling and Based on Genome-Wide Genetic Relatedness for Cardiac Conduction Traits. Twin Research and Human Genetics 20(6): 489-498.

Purcell S Neale B Todd-Brown K Thomas L Ferreira MA Bender D Maller J Sklar P de Bakker PI Daly MJ and Sham PC, 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81(3): 559–575.

RStudio Inc, 2013. shiny: web application framework for R. http://CRAN.R-project.org/package=shiny.

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

Shirali M Pong-Wong R Navarro P Knott S Hayward C Vitart V Rudan I Campbell H Hastie ND Wright AF and Haley CS, 2016. Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations. Heredity (Edinb) 116: 333-338.

Shirali M Knott SA Pong-Wong R Navarro P and Haley CS, 2018. Haplotype Heritability Mapping Method Uncovers Missing Heritability of Complex Traits. Scientific Reports 8(1): 4982.

Simeone R Misztal I Aguilar I and Legarra A, 2011. Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabeled genotyped animals in a broiler chicken population. Journal of animal Breeding and Genetics 128(5): 386–393.

Tropf FC, Hong Lee S, Verweij RM, Stulp G, van der Most PJ, Vlaming RD, Bakshi A, Briley DA, Rahal C Hellpap R Nyman A Iliadou AN Esko T Metspalu A Medland SE Martin NG Barban N Snieder H Robinson MR and Mills MC, 2017. Hidden heritability due to heterogeneity across seven populations. Nature human behaviour 1(10): 757–765.

Uemoto Y Pong-Wong R Navarro P Vitart V Hayward C Wilson JF Rudan I Campbell H. Hastie ND Wright AF and Haley CS, 2013. The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. Frontiers in Genetics 4: 232.

Valdisser PAMR Pereira WJ Almeida Filho JE Müller BSF Coelho GRC de Menezes IPP Vianna JPG Zucchi MI Lanna AC Coelho ASG de Oliveira JP da Cunha Moraes A Brondani C and Vianello RP, 2017. In-depth genome characterization of a Brazilian common bean core collection using DArTseq high-density SNP genotyping. BMC Genomics 18: 423.

Yang J Benyamin B McEvoy B Gordon S Henders AK Nyholt DR and Visscher PM, 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42: 565–571.

Yang J Lee SH Goddard ME, and Visscher PM, 2011. GCTA*:* a tool for genome-wide complex trait analysis. American journal of human genetics 88(1): 76–82.

Yu J Pressoir G Briggs WH Vroh Bi I Yamasaki M Doebley JF McMullen MD Gaut BS Nielsen DM Holland JB Kresovich S and Buckler ES, 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38: 203–208.

Zaitlen N Kraft P Patterson N Pasaniuc B Bhatia G Pollack S and Price AL, 2013. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLOS Genetics 9: e1003520.

Zeng Y Navarro P Fernandez-Pujals AM Hall LS Clarke TK and Thomson PA, 2017. A combined pathway and regional heritability analysis indicates NETRIN1 pathway is associated with major depressive disorder. Biological Psychiatry 81(4): 336–346.

Zhang Zh Li X Ding X Jiaqi L and Zhang Q, 2015. GPOPSIM: a simulation tool for whole-genome genetic data. BMC Genetics 16:10.

Evaluation of selected methods related to Genome-Wide Association Studies for identification of gene locus

F Beiranvand[1*], MT Beigi Nasiri[2], M Shirali[3] and M Shirali[3]

Received: February 4, 2020 Accepted: February 20, 2022

[1]PhD Sudent, Department of Animal Sciences, Ramin Agriculture and Natural Resources University, Mollasani, Khuzestan
[2]professor, Department of Animal Sciences, Ramin Agriculture and Natural Resources University, Mollasani, Khuzestan
[3]PhD in Genetics and Animal Breeding, Department of Agriculture, Tehran University

پژوهش‌های علوم دامی Animal Science Research

Journal of Animal Science/vol.31 No.4/ 2022/pp 43-57
https://animalscience.tabrizu.ac.ir

OPEN ACCESS

DOI: 10.22034/AS.2022.38215.1553

Introduction: Due to the widespread distribution of Single Nucleotide Polymorphisms (SNPs) throughout the genome, these markers are widely used in livestock breeding research. These markers have been used to predict the disease risk in human, to localize genetic variations responsible for complex traits through genome wide association study (GWAS), and to predict the genetic values of economically important traits in plant and animal breeding (Zhang et al. 2015). Mostly, whole genome scanning methods are based on two methods: Single SNP Genome-Wide Association Studies (SSGWAS) and multiple markers methods. The SSGWAS method is able to identify a large number of common variables affecting quantitative traits. However, a large proportion of the genetic variance remains to be explained (Shirali et al. 2018). In quantitative traits the proportion of phenotypic variance explained by SNPs is related to the number of adjacent SNPs in the genomic region. The heritability created by these genomic regions is defined as the regional heritability. The Regional Heritability Mapping (RHM) method is used to identify small genomic regions. This method can capture more of the missing genetic variation (Nagamine et al. 2012). In RHM, a mixed model framework based on Restricted Maximum Likelihood (REML) is used, and two variance components, one contributed by the whole genome and a second one by a specific genomic region, are fitted in the model to estimate genomic and regional heritabilities, respectively (Uemoto et al. 2013). Also fast and flexible set-Based Association Test (fastBAT) is a method that performs a fast set-based association analysis (Bakshi et al. 2016). The purpose of this study is compare SNPs and regions identified by the Genome-Wide Association methods, compare these results with the simulated Quantitative Trait Locus (QTL) and also investigate and determine the false positive results in each method.

Material and methods: In this study, markers and populations were simulated as a Forward-in-time process using QMSim software (Sargolzaei and Schenkel 2009). For this population, 27586 SNPs were counted on 3 pairs of autosomal chromosomes. Simulation was performed in 3 scenarios with 75, 150 and 300 QTL. The minimum and maximum number of SNPs in the analysis after quality control were 19662 and 23817 SNPs, respectively. For each scenario, 10 replicates were simulated, in all scenarios, heritability was 0.2 which corresponded equally to the polygenic and QTLs effects. Whole genomic relationship and pedigree base genetic relationship matrices were used in all 3 methods to estimate genetic parameters. To create the whole genomic relationships matrix, whole genomic additive effects was estimated using all SNPs. Also the additive effect of genomic regions

was estimated using the regional genomic relationship matrix. Whole genomic relationships matrix and regional genomic relationship matrix were estimated based on genetic relationships between individuals using SNPs by GCTA software (Yang et al 2011). Pedigree based genetic relationship matrix was created by the kinship relationship between individuals using pedigree package (Coster 2013) of RStudio software (RStudio Inc 2013). In this study, we considered windows containing 50 genotyped SNPs to perform RHM and to estimate variance components. Additionally, we used windows containing 25 genotyped SNPs to overlap between two consecutive windows throughout the genome. SSGWAS analysis were performed by MLMA (Yu et al. 2006) method using GCTA software. MLMA results were adjusted based on P-value at 5% significant threshold using Bonferroni correction. We used GCTA software to evaluate the results of SSGWAS using fastBAT method.

Results and discussion: For each replication after identifying significant SNPs, the genetic variance explained by these SNPs was estimated by $2pq\left[a+d\left(q-p\right)\right]^2$ equation (Faulkner & McKay 1996). In Table 1, the number of QTLs detected by the SSGWAS method, the MAF of QTLs, the range and mean of genetic variance explained by significant SNPs and QTLs are reported. For 30 replicates of simulation in SSGWAS, 16 QTLs were detected containing 2 QTLs with MAF≤0.1 and other detected QTLs with MAF≥0.1. Hundred seven significant regions were identified in fastBAT method. In this method, 120 QTLs were detected in 3 scenarios containing 52 QTLs with MAF≤0.1. All QTLs detected in the fastBAT and SSGWAS methods were also detected in the RHM method. In RHM method, 612 regions containing simulated QTLs and number of 316 QTLs with MAF≤0.1 were detected. In all replications, the variance explained by SNPs was equal to the variance explained by QTLs. In SSGWAS, less number of QTLs were detected than the other two methods and the maximum variance explained by QTLs was 14.9%. The criterion used to determine false positive QTLs was the absence of significant QTL in before and after significant windows containing QTLs. In SSGWAS method the percentage of false positive QTLs was higher than the other two methods. In fastBAT, unlike the other two methods, detected QTLs were not false positive. Number of detected QTLs, MAF range of QTLs, range and mean of genetic variance explained by detected QTLs and SNPs in fastBAT are shown in table 5. Many QTLs and regions detected by RHM method were not detected by SSGWAS and fastBAT methods. The genetic variance explained by detected QTLs in the RHM was at the range of 7.26% to 46.86% being higher than other two methods. In table 6, we have compared three methods by the number of detected QTLs, number of false positive QTLs, number of stable QTLs and the number of detected QTLs with MAF≤0.1. Correspondingly, we found that QTLs with MAF≤0.1 were more frequently detected in RHM than the other two methods.

Conclusion: In this study, we found that the potential of RHM method for identifying QTLs affecting the trait variance was higher than SSGWAS and fastBAT methods.

Key words: Quantitative traits locus, Single nucleotide polymorphism, Simulation, Whole genome scanning studies