University of Tehran Press

Journal of Animal Production

Online ISSN: 2382-994X

Homepage: https://jap.ut.ac.ir/

Investigating the performance of different methods in the selection of causative SNP markers on the breed differentiation of horses

Siavash Manzoori [1] | Amir Hossein Khaltabadi Farahani[2]✉ | Mohammad Hossein Moradi[3]

1. Department of Animal Science, Faculty of Agriculture, Tarbiat Modares University, Tehran, Iran. E-mail: s.manzoori@modares.ac.ir
2. Corresponding Author, Department of Animal Science, Faculty of Agriculture and Natural Resources, Arak University, Arak, Iran. E-mail: a-farahani@araku.ac.ir
3. Department of Animal Science, Faculty of Agriculture and Natural Resources, Arak University, Arak,Iran. E-mail: moradi.hosein@gmail.com

Article Info

Article type:
Research Article

Article history:
Received: August 17, 2022
Received in revised form: March 03, 2023
Accepted: March 13, 2023
Published online: April 14, 2023

Keywords:
Comparison,
Discrimination,
Genetic structure,
Genome,
Single nucleotide polymorphism.

ABSTRACT

The present study was conducted in order to select effective markers in breed discrimination and compare the performance of SNP marker selection methods with the data of 304 animals from 14 different breeds that were genotyped using the Illumina SNP50K marker panel. Knowledge of genetic structure is very important for better understanding of genetic changes in genomic scanning studies. The information content of each biomarker is used as an index for selecting markers in reducing the size of marker panels. To estimate the information content of each marker, the following selection methods were used: F_{ST} (pairwise & global), Theta, Delta, D, G_{ST}, G'_{ST}, G''_{ST} and Principal Component Analysis. In this study, the logarithm of the likelihood ratio was used to select markers. According to the results, all selection methods for identifying markers had similar behavior and performance. The number of common markers between the methods was at least 42 markers and at most 499 SNP markers. In general, the F_{ST} statistical method required a smaller number of markers to achieve a successful assignment. G'_{ST} and G''_{ST} statistics showed poor performance with more than 350 markers to achieve 95% correct assignment. It should be noted that with only the top 60 selected markers, it is possible to achieve a success rate of more than 70%. According to the results, Wright's paired F_{ST} had better performance than other SNP selection methods. The obtained results lead to the creation of exclusive panels to identify various breeds, which have great economic importance.

Cite this article: Manzoori, S., Khaltabadi Farahani, A. H., & Moradi, M. H. (2023). Investigating the performance of different methods in the selection of causative SNP markers on the breed differentiation of horses. *Journal of Animal Production*, 25 (1), 1-11. DOI: https://doi.org/10.22059/jap.2023.347298.623703

Publisher: University of Tehran Press.
DOI: https://doi.org/10.22059/jap.2023.347298.623703

انتشارات دانشگاه تهران

تولیدات دامی

شاپا الکترونیکی: ۲۳۸۲-۹۹۴x

Homepage: https://jap.ut.ac.ir/

بررسی عملکرد روش‌های مختلف در انتخاب نشانگرهای SNP مؤثر بر تمایز نژادی اسب‌ها

سیاوش منظوری[۱] | امیرحسین خلت‌آبادی فراهانی[۲]✉ | محمدحسین مرادی[۳]

۱. گروه علوم دامی، دانشکده کشاورزی، دانشگاه تربیت مدرس، تهران، ایران. رایانامه: s.manzoori@modares.ac.ir

۲. نویسنده مسئول، گروه علوم دامی، دانشکده کشاورزی و محیط زیست، دانشگاه اراک، اراک، ایران. رایانامه: a-farahani@araku.ac.ir

۳. گروه علوم دامی، دانشکده کشاورزی و محیط زیست، دانشگاه اراک، اراک، ایران. رایانامه: moradi.hosein@gmail.com

اطلاعات مقاله

نوع مقاله: مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۱/۰۵/۲۶

تاریخ بازنگری: ۱۴۰۱/۱۲/۱۲

تاریخ پذیرش: ۱۴۰۱/۱۲/۲۲

تاریخ انتشار: ۱۴۰۲/۰۱/۲۵

کلیدواژه‌ها:
تفکیک افراد،
چندشکلی تک‌نوکلئوتیدی،
ژنوم،
ساختار ژنتیکی،
مقایسه.

چکیده

مطالعه حاضر به‌منظور انتخاب نشانگرهای مؤثر در جداسازی نژاد و مقایسه عملکرد روش‌های انتخاب نشانگر با داده‌های ۳۰۴ حیوان از ۱۴ نژاد مختلف که با استفاده از پنل نشانگر Illumina SNP50K ژنوتیپ شده بودند، انجام شد. دانش و فهم ساختار ژنتیکی برای درک بهتر تغییرات ژنتیکی در مطالعات پویش ژنومی بسیار مهم است. محتوای اطلاعاتی هر نشانگر زیستی به‌عنوان شاخصی برای انتخاب نشانگرها در کاهش اندازه پنل‌های نشانگری کاربرد دارد. برای تخمین محتوای اطلاعاتی هر نشانگر، از آماره F_{ST} رایت، آماره θ، آماره دلتا، آماره D، آماره G_{ST}، آماره G'_{ST}، آماره G''_{ST} و آنالیز مؤلفه‌های اصلی استفاده شد. در این مطالعه، از آماره لگاریتم نسبت درست‌نمایی برای انتخاب نشانگرها استفاده شد. با توجه به نتایج، همه روش‌های انتخاب برای شناسایی نشانگرها دارای رفتار و عملکرد مشابهی بودند. تعداد نشانگرهای مشترک بین روش‌ها حداقل ۴۲ نشانگر و حداکثر ۴۹۹ نشانگر SNP بود. به‌طور کلی، روش آماره F_{ST} نیازمند تعداد کم‌تری از نشانگرها برای رسیدن به انتساب موفق بود. آماره‌های G'_{ST} و G''_{ST} با بیش از ۳۵۰ نشانگر برای دستیابی به ۹۵ درصد انتساب صحیح، عملکرد ضعیفی را نشان دادند. لازم به ذکر است که تنها با ۶۰ نشانگر برتر، دستیابی به موفقیت بیش از ۷۰ درصد امکان‌پذیر است. با توجه به نتایج، F_{ST} جفتی رایت از سایر روش‌های انتخاب SNP دارای عملکرد بهتری بود. نتایج حاصله منجر به ایجاد پنل‌های انحصاری جهت شناسایی نژادهای متنوع می‌گردد که دارای اهمیت اقتصادی زیادی است.

استناد: منظوری، س.، خلت‌آبادی فراهانی، ا. ح. و مرادی، م. ح (۱۴۰۲). بررسی عملکرد روش‌های مختلف در انتخاب نشانگرهای SNP مؤثر بر تمایز نژادی اسب‌ها. *نشریه تولیدات دامی*، ۲۵ (۱)، ۱-۱۱. DOI: https://doi.org/10.22059/jap.2023.347298.623703

ناشر: مؤسسه انتشارات دانشگاه تهران.

۱. مقدمه

شناسایی ساختار جمعیت، قابلیت تمایز و طبقه‌بندی حیوانات یک رویکرد مفید و اساسی در مطالعات زیست‌شناسی است [۱]. شناسایی ژنتیکی یا روش‌های تخصیص مبتنی بر اطلاعات ژنوم، در فرایند استنباط اصل و انساب اسب‌ها، برای تشخیص میزان مهاجرت در بین جمعیت‌ها و برای پژوهش‌های پزشکی قانونی می‌تواند مهم‌ترین فرایند باشد [۲]. در دهه‌های اخیر، بهره‌گیری از مزایای چندشکلی‌های تک‌نوکلئوتیدی همراه با توسعه سیستم‌های محاسباتی، منجر به ساخت ابزارهای جدیدی برای درک عمیق‌تر داده‌های ژنومی شده‌اند. نشانگرهای تک‌نوکلئوتیدی با دو ویژگی عمده خطای ژنوتیپ کم‌تر و فراوانی بسیار در کل ژنوم به ابزاری مطلوب برای پژوهش‌گران تبدیل شده‌اند [۳].

برخی از اشکالات داده‌های ژنومی تعداد بالای نشانگرها، پردازش‌های محاسباتی زمان‌بر و فرایندهای پر هزینه می‌باشند. یک راه‌حل مناسب جهت مشکلات فوق‌الذکر، کاهش تعداد نشانگرها برای تجزیه و تحلیل است که با انتخاب نشانگرها براساس محتوای اطلاعاتی صورت می‌پذیرد. به‌عبارت دیگر، محتوای اطلاعاتی هر نشانگر را به‌عنوان ابزاری برای انتخاب نشانگرهای متمایزکننده در نظر می‌گیرند که آن نشانگرها موجب ایجاد بیش‌ترین تنوع هستند. روش‌های متعددی وجود دارند، که اکثر آن‌ها از فراوانی آللی به‌عنوان معیار اولیه برای تنوع ژنتیکی و تعیین اطلاعات ژنتیکی هر نشانگر استفاده می‌کنند. در علم ژنتیک و اصلاح نژاد، نشانگرهای اطلاعاتی (Informative SNPs) می‌توانند به‌عنوان ابزاری برای تفکیک بین جمعیت‌ها مورداستفاده قرار بگیرند.

روش‌هایی که در این پژوهش برای توصیف تنوع بین جمعیت‌ها در نظر گرفته شدند، شامل آماره‌های Fst، آماره تتا (θ)، آماره D، آماره‌های Gst (در مجموع سه آماره) و آماره دلتا هستند [۱۳-۴]. آماره تتا (θ) (نسخه اصلاح‌شده از F_{ST} رایت) تأثیر و اندازه محدود نمونه را در فرایند محاسباتی تصحیح می‌نماید [۶]. آماره دلتا معمولاً در پژوهش‌های انسانی استفاده می‌شود و تغییرات فراوانی آللی را در نمونه‌های موردمطالعه اندازه‌گیری می‌کند [۱۳]. در نهایت، از آنالیز مؤلفه‌های اصلی به‌عنوان یک روش جایگزین برای ارزیابی محتوای اطلاعاتی نشانگرها استفاده شد [۱۴]. هدف از این مطالعه ارزیابی روش‌های مختلف در انتخاب نشانگرهای علّی (Causative markers) و مؤثر از پنل‌های بزرگ فعلی برای تمایز نژادی در بین اسب‌های امروزی است.

۲. مواد و روش‌ها

در این پژوهش اطلاعات ۳۰۴ حیوان از ۱۴ جمعیت (نژاد)، که از نظر منطقه جغرافیایی دارای قرابت بودند، موردمطالعه قرار گرفتند [۱۵]. اطلاعات و مشخصات جمعیت‌ها (حیوانات) در جدول (۱) ارائه شده است. نمونه‌ها با استفاده از پنل نشانگری Illumina SNP50K (Illumina, San Diego, CA, USA) تعیین ژنوتیپ گردیدند. پس از فرایند کنترل کیفیت داده‌ها (MAF>۰/۰۵ و Call rate>۰/۰۵) و اعمال روش‌های انتخاب، در نهایت ۲۶۴۱۰ نشانگر برای آنالیزهای آتی باقی ماندند.

نسبت حیوانات در هر قاره جهان به‌ترتیب برای آمریکای شمالی (۲۶/۳ درصد)، آسیا (۳۱/۳ درصد) و اروپا (۴۲/۴ درصد) برآورد شد. ابتدا، اطلاعات هر نشانگر محاسبه گردید. سپس در گام دوم نشانگرها براساس این ارزش‌ها رتبه‌بندی شدند. در نهایت، از آماره نسبت لگاریتم درست‌نمایی ((log-likelihood ratio (LLR))، برای اختصاص افراد، با افزایش تجمعی تعداد نشانگرها در هر روش برآورد شد.

میزان تأثیر و اطلاعات هر نشانگر با استفاده از روش‌هایی هم‌چون آماره F_{ST} رایت (جفتی و کلی)، آماره θ (جفتی و کلی)، آماره دلتا، آماره D، آماره G_{ST}، آماره G'_{ST}، آماره G''_{ST} و آنالیز مؤلفه‌های اصلی برآورد گردید. فرض بر این است که هرچه محتوای اطلاعاتی نشانگر بیش‌تر باشد، پس دارای سهم بیش‌تری نیز در تمایز مابین جمعیت‌ها است. آنالیزها به‌وسیله کد نویسی دستی در محیط نرم‌افزار آماری R (نسخه 3.2) انجام گرفت [۱۶].

جدول ۱. نام، شناسه، اندازه نمونه و میانگین فراوانی آلل کمیاب هر یک از نژادهای موردبررسی در مطالعه حاضر

ردیف	نژاد	شناسه	مبدأ جغرافیایی	قاره	تعداد	MAF
۱	آخال تکه	AKTK	ترکمنستان	آسیا	۱۹	۰/۲۳
۲	عرب	ARR	خاورمیانه	آسیا	۲۴	۰/۲۴
۳	بلژین	BEL	بلژیک	اروپا	۳۰	۰/۲۱
۴	کاسپین	CSP	ایران	آسیا	۱۸	۰/۲۲
۵	پونی فل	FELL	انگلستان	اروپا	۲۱	۰/۲۱
۶	فرانچز مونتاچز	FM	سوئیس	اروپا	۱۹	۰/۲۲
۷	مارم مانو	MARM	ایتالیا	اروپا	۲۴	۰/۲۴
۸	مغولی	MON	مغولستان	آسیا	۱۹	۰/۲۱
۹	مورگان	MOR	آمریکا	آمریکا	۴۰	۰/۲۲
۱۰	افجورد نروژی	NORF	نروژ	اروپا	۲۱	۰/۲۱
۱۱	سادل برد	SB	آمریکا	آمریکا	۲۵	۰/۲۳
۱۲	استاندارد برد آمریکایی	STBDUS	آمریکا	آمریکا	۱۵	۰/۲۴
۱۳	خونگرم سوئیسی	SZWB	سوئیس	اروپا	۱۴	۰/۲۵
۱۴	تیوا	TUVA	سیبری	آسیا	۱۵	۰/۲۲

آماره MAF نشان‌دهنده میانگین میزان فراوانی آلل کمیاب، برای همه نشانگرهای SNP در هر نژاد می‌باشد.

آماره F_{ST} (جفتی و کلی): آماره F_{ST} رایت به‌صورت رابطه (۱) تعریف شده است [۴].

رابطه ۱) $$F_{ST} = \frac{Var(p_A)}{\bar{p}_A(1-\bar{p}_A)}$$

که در این رابطه، $\bar{p}_A$، میانگین فراوانی آلل را در نژادهای مختلف را نشان می‌دهد و $Var(p_A)$، واریانس فراوانی آلل بین نژادها را توصیف می‌کند.

آماره تتا (θ) (جفتی و کلی): این آماره به‌عنوان نسبتی از واریانس بین جمعیت‌ها به کل واریانس تعریف شد. آماره تتا (θ) در آنالیزهایی با اندازه‌های نمونه متفاوت کاربرد دارد، چراکه اثر اندازه جمعیت را هنگام تخمین F_{ST} لحاظ می‌نماید. در مقالات پیشین توضیحات کاملی درباره نحوه تخمین آماره تتا (θ) بیان گردیده است [۶-۷].

آماره دلتا: مقدار آماره دلتا که فقط مابین جفت جمعیت‌ها استفاده می‌شود (که در این پژوهش K=14 در نظر گرفته شده است) [۹]. به‌عبارت دیگر، مقادیر دلتا نشان‌دهنده تمایز فراوانی آللی بین دو جمعیت مجزا است. مقادیر صفر و یک به‌ترتیب نشان‌دهنده عدم تمایز و تمایز کامل در بین جمعیت‌های موردمطالعه است.

آماره Jost's D: آماره D براساس دو مشکل آماره G_{ST} ابداع گشت [۹]. در گام نخست، به جای هتروزیگوسیتی، از تعداد مؤثر آلل‌ها «تنوع واقعی»، استفاده گردید. در گام دوم، آماره D در فاصله بین صفر و یک متغیر و برای اندازه‌گیری تنوع بین جمعیت است، که به‌عنوان تابعی از هتروزیگوسیتی کل و درون جمعیتی مفید می‌باشد [۱۰].

آماره G_{ST}: براساس آماره F، و مشتق‌های آن، پژوهش‌گران دیگری از هتروزیگوسیتی به‌عنوان شاخص تثبیت، آماره G_{ST}، استفاده کردند [۱۱].

آماره G'_{ST}: پس از مدتی آماره جدید G'_{ST} از آماره G_{ST} برای ایجاد یک شاخص جدید تثبیت ابداع گردید [۱۲]. مقدار آن به‌وسیله هتروزیگوت زیر جمعیت‌ها و آماره $G_{ST(max)}$ محاسبه می‌شود.

آماره G''_{ST}: زمانی‌که حجم نمونه جمعیت‌ها (k) کوچک باشد، آماره G_{ST} دچار خطای کاهش برآورد (Underestimate) می‌شود. بنابراین، شاخص دیگری از هتروزیگوسیتی $G'_{ST(Nei)}$ تعریف گردید. آماره G'_{ST} نیز زمانی که تعداد افراد جمعیت‌ها کوچک است، نیاز به اصلاح دارد چراکه ارزش‌ها را کم‌تر از حالت واقعی برآورد می‌کند. بنابراین نسخه تصحیح‌شده G'_{ST}، آماره G''_{ST}، برای اجتناب از این مشکل ارائه گردید [۱۰].

آنالیز مؤلفه‌های اصلی: الگوریتم آنالیز مؤلفه‌های اصلی مجموعه‌ای از متغیرهای اصلی را که احتمالاً دارای همبستگی هستند را به مجموعه‌ای از مقادیر تبدیل می‌کند که به این متغیرهای خطی ناهمبسته به‌دست‌آمده مؤلفه اصلی گفته می‌شود. این روش براساس همبستگی بین متغیرها کار می‌کند. بنابراین اگر متغیرها به یکدیگر وابستگی داشته باشند، آنالیز مؤلفه‌های اصلی به‌درستی عمل خواهد کرد [۱۴].

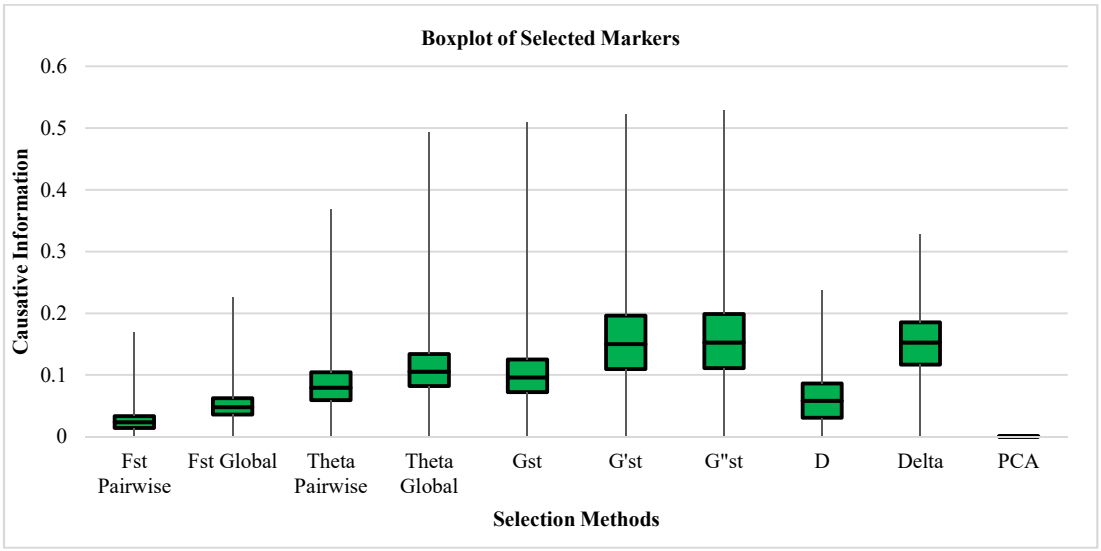
برای تخصیص ژنتیکی و آنالیز انتساب افراد، بسیاری از رویکردها از جمله کاربرد روش بیزی توسعه یافته‌اند [۲ و ۱۷]. روش LLR دارای عملکرد بالایی برای انتساب افراد است، به‌ویژه زمانی که سطوح تمایز در جمعیت مرجع بالا باشد. در این پژوهش، روش LLR (رابطه ۲) با نشانگرهای چندشکلی تک‌نوکلئوتیدی برای آنالیز انتساب اجرا گردید [۱۷].

رابطه ۲) $$LLR = \log_{10}\big(T(g|i_A)\big) - \log_{10}\big(T(g|i_B)\big)$$

نسبت لگاریتم درست‌نمایی $T(g|i_A)$ نشان‌دهنده احتمال این است که حیوان g در جمعیت i قرار دارد. اگر ارزش محاسبه‌شده نسبت لگاریتم درست‌نمایی بزرگ‌تر از پنج سطوح آستانه (صفر، یک، دو، سه و چهار) بود، در این صورت فرض می‌شود که فرد ژنوتیپ شده به‌درستی به منشأ خود اختصاص داده شده است. اگر مقدار نسبت لگاریتم درست‌نمایی کم‌تر از سطح آستانه بود، آنگاه تخصیص به‌عنوان یک شکست در نظر گرفته می‌شود.

۳. نتایج و بحث

همان‌طور که بیان گردید هدف اصلی این پژوهش مقایسه عملکرد روش‌های موجود برای شناسایی نشانگرهای متمایزکننده نژادهای اسب از پنل نشانگری 50K رایج در بازار است. با توجه به شکل (۱)، روش‌های انتخاب نشانگر و میزان اطلاعات برآوردشده به‌وسیله آن‌ها، می‌توان دریافت که روش‌ها برای محاسبه ارزش نشانگرها رویه خاصی را در نظر می‌گیرند.

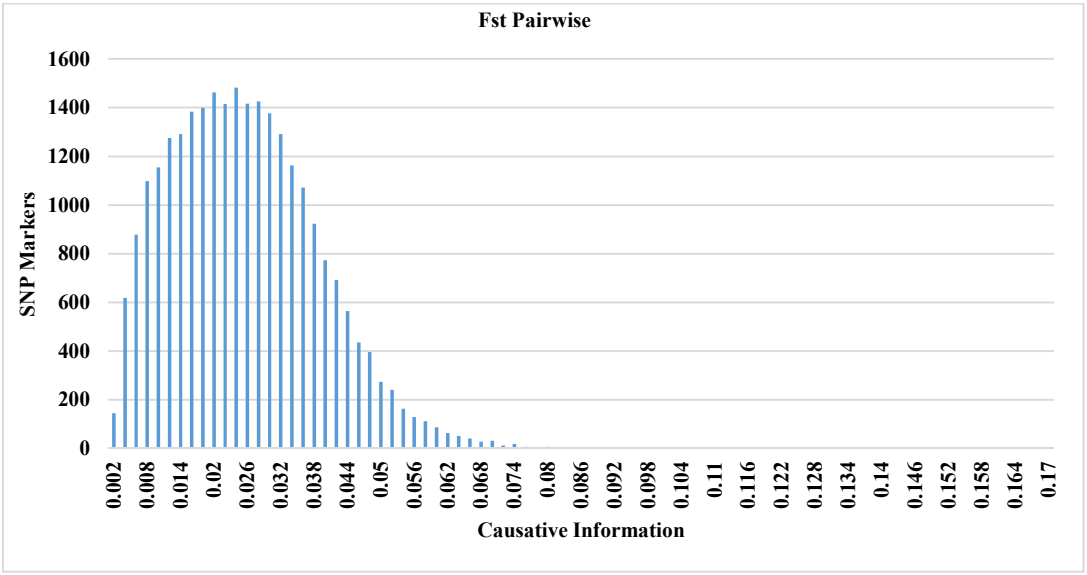


شکل ۱. نمودارهای جعبه‌ای عملکرد روش‌های مختلف در برآورد ارزش علیت و تمایز نشانگرها

در حالت کلی، همه روش‌ها دارای عملکرد مشابهی برای برآورد ارزش علیت نشانگرها بودند. به‌نحوی‌که تمامی روش‌ها برای بیش‌تر نشانگرها ارزشی نزدیک به ۰/۱ الی ۰/۲ برآورد کردند. البته لازم به ذکر است که ارزش‌های برآوردشده به‌وسیله روش مؤلفه‌های اصلی کم‌ترین مقادیر (۰/۰۳-۰/۰) را دارا بود. درحالی‌که روش‌های G'_{ST} و G''_{ST} (به‌طور تقریباً مشابه) دارای

بیش‌ترین اطلاعات برآوردشده (۰/۳۴–۰/۰) بودند. در تمامی توزیع‌ها برای اکثر روش‌های انتخاب تمایلی به سمت چپ نمودار (چوله به راست) مشاهده شد. برای درک بهتر و بیش‌تر میزان اطلاعات ژنتیکی نشانگرهای انتخاب‌شده به‌وسیله روش F_{ST} $Pairwise$ در غالب نمودار هیستوگرام در شکل (۲) نشان داده شده است.

بیش‌تر نشانگرها مقادیر تخمینی کم تا متوسط را نشان می‌دهند و نسبت کوچکی از آن‌ها مقادیر تخمینی بالایی را نشان می‌دهند که برای تمایز جمعیت مفید است. با توجه به نتایج به‌دست‌آمده، روش F_{ST} $Pairwise$ برای اکثر نشانگرها، اطلاعات تمایز ژنتیکی کم و یا متوسطی را برآورد کرد و تنها بخش کوچکی از آن‌ها حاوی مقادیر و ارزش‌های بالای تمایز ژنتیکی بودند. البته لازم به ذکر است که تمامی روش‌های انتخاب همانند روش F_{ST} $Pairwise$، تمایلی مشابه برای برآورد ارزش‌های علیت برای نشانگرها بودند. تعداد نسبتاً بالایی از نشانگرهای مشترک در بین روش‌های انتخاب، بیانگر این موضوع است که همه روش‌ها از نظر عملکرد تا حدودی شباهت دارند. بیش‌ترین نشانگر مشترک (۴۹۹) در ۵۰۰ نشانگر اول در بین این ده روش در جدول (۲) گزارش شد.



شکل ۲. هیستوگرام روش Fst-Pairwise که محتوای اطلاعات ژنتیکی را برای نشانگرها نشان می‌دهد. (محور x معرف مقدار ارزش برآوردشده و محور y نشان‌دهنده تعداد نشانگرها می‌باشند).

جدول ۲. مقایسه‌ای بین روش‌های انتخاب نشانگر برای شناسایی SNPهای متمایزکننده

	θ Global	D	G''_{ST}	G'_{ST}	G_{ST}	PCA	θ Pairwise	Delta	F_{ST} Pairwise	F_{ST} Global
F_{ST} Global										(۲۰۵)
F_{ST} Pairwise									(۱۱۵)	۱۸۵
Delta								(۱۳۰)	۶۳	۲۲۱
θ Pairwise							(۲۱۰)	۲۴۲	۱۹۷	۴۴۵
PCA						(۱۳۳)	۹۷	۶۲	۴۲	۱۰۲
G_{ST}					(۱۹۴)	۱۴۴	۱۶۲	۸۳	۱۴۳	۱۵۷
G'_{ST}				(۲۶۰)	۳۳۱	۲۰۶	۲۰۴	۱۳۵	۱۰۵	۲۰۴
G''_{ST}			(۲۵۹)	۴۹۹	۳۳۱	۲۰۵	۲۰۴	۱۳۵	۱۰۵	۲۰۴
D		(۱۹۵)	۳۳۶	۳۳۶	۱۸۱	۱۸۳	۱۷۵	۱۴۸	۴۸	۱۷۵
θ Global	(۲۱۸)	۱۸۱	۳۲۰	۳۲۰	۴۱۶	۱۶۱	۱۶۶	۸۷	۱۵۴	۱۵۹

تعداد SNPهای مشترک نیز در هر روش در مثلث پایینی نمایش داده شده است (اعداد قطری داخل پرانتز میانگین می‌باشند).

حداقل نشانگرهای مشترک (۱۱۵ نشانگر) در بین ۵۰۰ نشانگر زیستی رده بالا مربوط به روش F_{ST} جفتی با دیگر روش‌ها بود (جدول ۲). در مورد جدول (۲)، G'_{ST} و G''_{ST} عملکرد مشابه‌ای برای فرایند انتخاب نشانگرها داشتند، زیرا هر دو دارای تعداد مشابه‌ای از نشانگرهای مشترک هستند. برای توضیح این همبستگی بالا، می‌توان استنباط کرد که دو روش G'_{ST} و G''_{ST} از نظر ریاضی دارای اشتراک فراوان بوده و همان‌گونه که در قسمت مواد و روش اشاره شد این دو روش تفاوت چندانی در محاسبات ندارند. دو روش F_{ST} رایت و آماره دلتا در این پژوهش دارای تفاوت بسیاری با یکدیگر بودند (۱۳۰ نشانگر روش Delta در مقابل ۲۰۵ نشانگر روش Fst). این درحالی است که، بنا به گفته دیگر پژوهش‌گران دو روش F_{ST} رایت و آماره Delta دارای عملکرد مشابهی هستند [۱۸].

نکته قابل‌توجه در این جدول وجود ۱۳۳ نشانگر مشترک روش آنالیز مؤلفه‌های اصلی با بقیه روش‌هاست. اگرچه مکانیسم تحلیلی آنالیز مؤلفه‌های اصلی با سایرین متفاوت می‌باشد اما قرابت این روش با روش Delta (۱۳۰ نشانگر مشترک) بیانگر وجه اشتراک مابین دو روش است. اساس آنالیز مؤلفه‌های اصلی تعیین روابط ناشناخته بین مجموعه متغیرها (نشانگرهای زیستی) توسط ساختار ماتریس کوواریانس بین فراوانی‌های آللی جایگاه‌های مختلف است. با استفاده از این تعریف، تخمین محتوای اطلاعات ژنتیکی یک نشانگر خاص به نشانگرهای دیگری بستگی دارد که در آنالیز شرکت می‌کنند، به‌ویژه زمانی‌که عدم تعادل پیوستگی میان آن‌ها بالا باشد. در آماره Delta نیز از فراوانی آللی دو جایگاه نیز استفاده می‌گردد که می‌تواند توجیه‌کننده این قرابت عملکردی باشد. در مقابل، تمام روش‌های F_{ST} و آماره Delta سطح اطلاعات هر نشانگر را مستقل از سایر نشانگرها تخمین می‌زنند. چراکه این روش‌ها از واریانس فروانی آللی یک جایگاه در هر نژاد استفاده کرده و فراوانی آللی جایگاه دیگر بر برآورد اطلاعات نشانگرها مؤثر نیست. در این پژوهش از چندشکلی‌های تک‌نوکلئوتیدی که در همه جمعیت‌ها هتروزیگوت بودند، استفاده شد. این نتیجه این واقعیت را تأیید می‌کند که چندشکلی‌های تک‌نوکلئوتیدی هتروزیگوت برای تخصیص فردی مناسب هستند، همان‌طور که قبلاً پیشنهاد شده است [۸ و ۱۸].

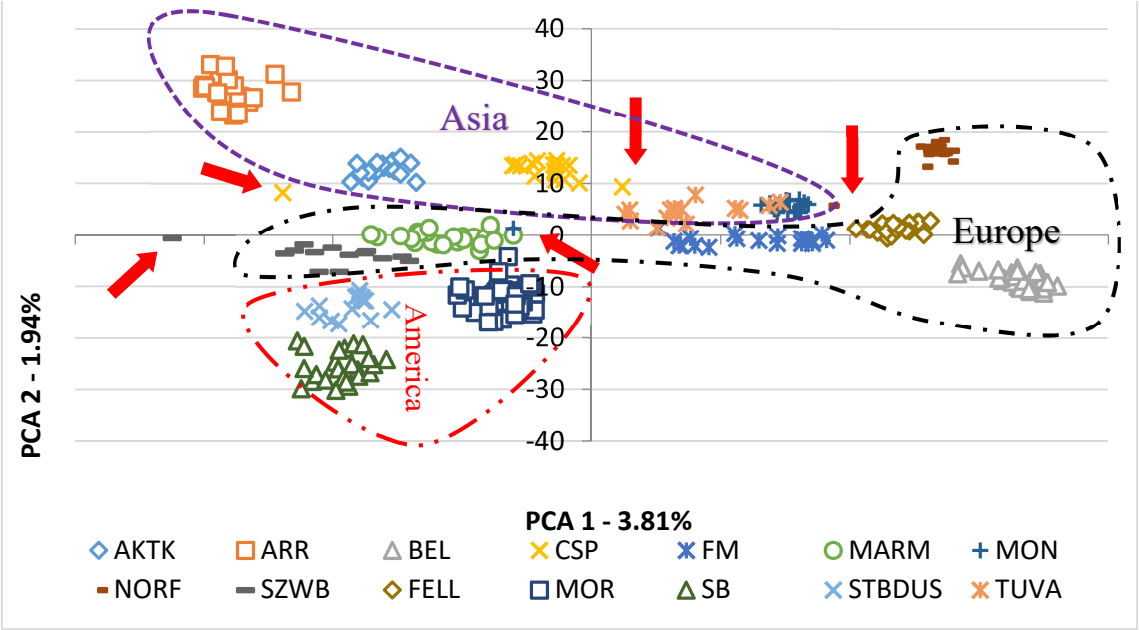
نمودار آنالیز مؤلفه‌های اصلی حاصل از ۲۶۴۱۰ نشانگر تک‌نوکلئوتیدی تمایز خوبی را در بین نژادها نشان می‌دهد (شکل ۳). ازآنجایی‌که پلات آنالیز مؤلفه‌های اصلی پنل‌های نشانگری کاهش‌یافته در هر روش دارای الگوی یکسانی برای تمایز نژادها در هنگام استفاده از همه نشانگرها (۲۶۴۱۰) بودند، از نمایش آن‌ها صرف‌نظر شد. به‌طورکلی، تعدادی از اسب‌ها در گروه (نژاد) خود قرار نگرفتند که می‌تواند به‌دلیل ترکیب ژنتیکی (هیبرید) در نسل‌های قبلی (والدین) آن‌ها باشد و یا به اشتباه در نژادی دیگر ثبت شده باشند. افراد طردشده (مجموعاً پنج مورد) با فلش‌های قرمز در شکل (۳) نشان داده شده‌اند.

در شکل (۳)، توزیع و پراکنش نژادهای متفاوت نشان داده شده است. در این شکل سه دسته اصلی که با خطوط نقطه‌چین نشان داده شده‌اند، دیده می‌شوند. بنابر نتایج حاصله، نژادهای هر قاره در یک قسمت قرار گرفته‌اند. نژادهای آسیایی (عرب، آخال‌تکه، کاسپین، تیوا و مغولی) از سمت چپ بالای نمودار تا مرکز آن قرار دارند. در میان شکل (۳) می‌توان نژادهای اروپایی را یافت. در حالی‌که سه نژاد آمریکایی در قسمت پایینی شکل به‌صورت یک خوشه مجزا قرار گرفته‌اند. نژادهایی هم‌چون تیوا و مغولی دارای همپوشانی با نژاد اروپایی فرانچ‌مونتانا هستند. مابقی نژادهای آسیایی شامل کاسپین، آخال‌تکه و عرب به‌صورت گروه‌های مجزا قرار گرفته‌اند که حاکی از تبادل ژن اندک مابین این نژادهای آسیایی است. در واقع می‌توان این‌گونه بیان کرد که، این نژادها با این‌که در یک منطقه (خاورمیانه) قرار دارند اما هنوز خلوص ژنتیکی خود را حفظ کردند.

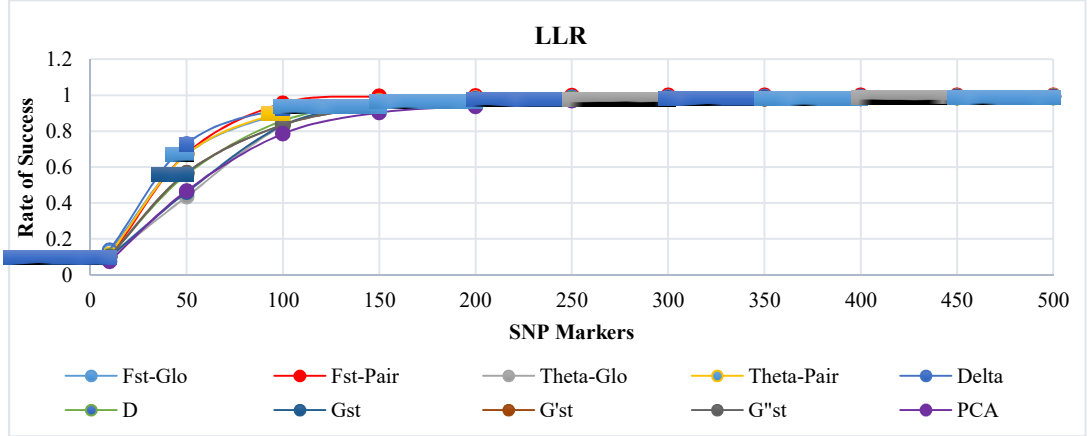
نژادهای اروپایی نیز سه نژاد BEL (بلژیک)، FELL (انگلستان) و NORF (نروژ) در یک قسمت جداگانه قرار گرفتند که حاکی از ایزوله‌بودن این نژادها حتی در قاره اروپا می‌باشد. با این وجود، نژادهای اروپایی MARM (ایتالیا) و SZWB (سوئیس) دارای همپوشانی اندکی با یکدیگر هستند. درنهایت درباره نژادهای اروپایی می‌توان این‌گونه بیان کرد که آن‌ها به‌عنوان یک رابطه بین نژادهای آسیایی و آمریکایی قرار گرفته‌اند. نژادهای آمریکایی که در قسمت پایین و

چپ نمودار قرار گرفته‌اند نیز به‌صورت جمعیت‌های ایزوله‌شده دیده می‌شوند. همان‌گونه که از شواهد پیدا است، نژادهای آمریکایی دارای کم‌ترین تبادل ژنی با دو قاره دیگر هستند.

با توجه به شکل (۳)، پنج حیوان در آنالیز به‌صورت افراد پرت شناسایی شدند که با فلش‌های قرمزرنگ نشان داده شده‌اند. دلیل این وضعیت احتمالاً به‌خاطر وجود والدین آمیخته حیوانات آنالیز باشد. برای بررسی وضعیت این افراد از آنالیز انتساب استفاده شد تا میزان اختلاط ژنتیکی این حیوانات ارزیابی گردد. صحت فرایند انتساب اسب به مبدأ (نژاد) اصلی به‌صورت تجمعی با اضافه‌کردن یک نشانگر در هر نوبت، براساس کاهش محتوای اطلاعات ژنتیکی نشانگرها در هر روش محاسبه شد. در حالت کلی (میانگین مقادیر در پنج سطح)، روند افزایش میزان موفقیت تخصیص (انتساب) در بین تمام روش‌ها با افزایش تعداد نشانگرها مشابه بود (شکل ۴).



شکل ۳. نمودار آنالیز مؤلفه‌های اصلی از ۲٦٤۱۰ نشانگر نشان‌دهنده این واقعیت است، که تمایز قاره جغرافیایی مشاهده شده است و برای چند نژاد همپوشانی وجود دارد. طردشدگان در طرح PCA که با فلش‌های قرمز نشان داده شده‌اند.



شکل ٤. میزان صحت تعیین نژاد براساس ۵۰۰ نشانگر برتر (در حالت میانگین کلی پنج سطح آنالیزی)

برای همه روش‌ها، میزان موفقیت تخصیص (شناسایی) ژنتیکی حدود ۷۰ درصد با ۶۰ نشانگر اول برای ۳۰۴ ژنوتیپ (حیوان) به مبدأ خود به‌دست آمد. لازم به ذکر است که میزان موفقیت در انتساب افراد بعد از ۲۳۰ نشانگر افزایش قابل‌توجهی نمی‌یابد. روش‌های Fst جفتی (با رنگ قرمز) نسبت به دیگر روش‌ها دارای عملکردی بهتری بودند. درحالی‌که روش آنالیز مؤلفه‌های اصلی (با رنگ بنفش) ضعیف‌ترین عملکرد را از خود نشان داد.

به‌طورکلی، F_{ST} جفتی رایت دارای کم‌ترین تعداد نشانگر برای رسیدن به آستانه موفقیت است (جدول ۳). روش‌های G'_{ST} و G''_{ST} عملکرد ضعیفی با بیش از ۳۵۰ نشانگر برای دستیابی به ۹۸ درصد انتساب صحیح را داشتند. همان‌طورکه در جدول (۳) نشان داده شده است، کم‌تر از ۴۰۰ نشانگر در آنالیز انتساب ۳۰۴ حیوان موردنیاز است.

برای تخصیص صحیح ۳۰۴ ژنوتیپ، به کم‌تر از ۹۰ نشانگر برای رسیدن به میزان موفقیت قابل‌قبول بیش از ۹۰ درصد (در آستانه اول (LLR>۰)) موردنیاز است. آماره میانگین مربوط به نشانگرهای منتخب هر روش در تمامی آستانه‌های تعریف‌شده در جدول (۳) در داخل پرانتز گزارش شده‌اند. روش Fst جفتی با حدود ۹۲ نشانگر در بین تمامی روش‌ها دارای حداقل نشانگر موردنیاز برای انتساب صحیح حیوانات بود. در مقابل روش آنالیز مؤلفه‌های اصلی بیش‌ترین تعداد نشانگر (۲۰۲ نشانگر) را برای انتساب نیاز داشت. در نهایت، آماره آنالیز مؤلفه‌های اصلی، G'_{ST} و G''_{ST} به تعداد بیش‌تری از نشانگرها (بیش از ۱۷۰ نشانگر) نیاز دارند که باعث می‌شود، مجموعه انتخابی آن‌ها دارای عملکرد نسبتاً خوبی باشد.

بسیاری از مطالعات این واقعیت را نشان داده‌اند که زیرمجموعه کوچکی از چند شکل‌های تک‌نوکلئوتیدی با اطلاعات بالا برای تصحیح مؤثر فرایند طبقه‌بندی جمعیت کافی است [۱۷ و ۱۹]. افزودن نشانگرهای غیر اطلاعاتی (به‌عنوان مثال، محتوای اطلاعات ژنتیکی اندک یا نشانگرهای تک شکلی) ممکن است منجر به نتایج دارای اختلاط شود و در عین حال دقت پژوهش را کاهش دهد [۲۰ و ۲۱]. با این‌که تعداد اندکی از نشانگرها دارای حداکثر اطلاعات هستند، در واقع می‌توانند باعث بوجود آمدن مسیرهای جدید محاسباتی مانند آنالیز خوشه‌بندی بیزین شوند [۲۲]. گونه حیوان، میزان آستانه دلخواه، مقدار تمایز و ملاحظات جمعیت مهم‌ترین عوامل برای تعیین تعداد نشانگرهای موردنیاز برای آنالیز انتساب هستند.

جدول ۲. تعداد نشانگر موردنیاز هر روش آنالیزی برای تخصیص حیوانات به نژادهای اصلی

روش‌های انتخاب نشانگر	آستانه صفر		آستانه ۴	
	۹۰ درصد	۹۸ درصد	۹۰ درصد	۹۸ درصد
F_{ST} Global (۱۲۲/۳۳)	۵۳	۱۱۴	۱۲۲	۲۱۰
F_{ST} Pairwise (۹۱/۱۳)	۴۷	۸۱	۱۰۰	۱۴۴
Delta (۱۱۷/۸۰)	۳۷	۸۰	۱۱۹	۲۴۰
θ Pairwise (۱۲۱/۰۷)	۵۲	۱۰۴	۱۲۷	۱۹۹
θ Global (۱۵۱/۵۳)	۷۷	۱۴۸	۱۴۵	۳۰۰
PCA (۲۰۱/۸۰)	۸۴	۲۲۶	۱۹۰	۳۳۴
G_{ST} (۱۵۵/۶۰)	۷۶	۱۵۰	۱۵۱	۲۸۴
G'_{ST} (۱۸۲/۳۳)	۶۵	۱۵۲	۱۶۰	> ۳۵۰
G''_{ST} (۱۷۷/۸۷)	۶۵	۱۵۲	۱۶۲	> ۳۵۰
D (۱۶۳/۱۳)	۶۳	۲۰۹	۱۵۳	۳۰۱

تعداد نشانگرهای SNP موردنیاز برای به‌دست‌آوردن موفقیت ۹۰ و ۹۸ درصدی در دو آستانه برای هر روش. (اعداد داخل پرانتز، میانگین ۱۵ عدد به‌دست‌آمده در پنج آستانه هر روش است).

به‌طورکلی، پژوهش‌های پیشین با هدف انتساب ژنوتیپ‌های افراد با استفاده از نشانگرهای فعلی است و به تعداد ایده‌آل نشانگرها برای این کار اشاره‌ای نشده است. در این پژوهش‌ها، در وهله اول از نشانگرهای ریزماهواره‌ای استفاده کردند [۲۳]. هم‌چنین نشان داده شد که با استفاده از ۲۳ نشانگر ریزماهواره، بیش از ۹۰ درصد افراد را می‌توان به نژاد خود اختصاص داد [۲۳]. در وهله دوم، آن‌ها از تعداد مشخص و ثابتی از نشانگرها، که از پیش تعریف شده بودند، استفاده کردند. البته پژوهش‌گران دیگری نیز از مجموعه محدودی از نشانگرهای تک‌نوکلئوتیدی برای آنالیز تخصیص استفاده کردند [۲۴].

با استفاده از نتایج روش‌های مختلف پژوهش‌گران قادر به ایجاد پنل‌های نشانگری انحصاری جهت شناسایی نژادهای متنوع هستند. چراکه انتساب و شناسایی حیوانات آمیخته در صنعت اسب دارای اهمیتی بسیار اقتصادی است. هم‌چنین روش‌های مختلف باعث شناسایی نشانگرهای علّی می‌شوند که احتمالاً در ناحیه کدکننده ژنوم قرار دارند. این کار باعث درک بهتر مکانیسم‌های مولکولی موجود در تمایز نژادها می‌شود. علاوه بر این پیشنهاد می‌شود، که روش‌های موجود را می‌توان با سایر روش‌های انتخاب ویژگی (جنگل تصادفی و شبکه عصبی مصنوعی) ترکیب کرد تا نتایج بهتری به‌دست آورد.

۴. تشکر و قدردانی

از پروژه کنسرسیوم تنوع ژنتیکی اسب (Equine Genetic Diversity Consortium) و دکتر Petersen که داده‌های ژنوتیپی را فراهم کردند، تشکر و قدردانی می‌گردد.

۵. تعارض منافع

هیچ‌گونه تعارض منافع توسط نویسندگان وجود ندارد.

۶. منابع

1. Makina SO, Muchadeyi FC, Van Marle-Köster E, Mac-Neil MD and Maiwashe A (2014) Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. Frontiers in Genetics, 5(333) doi:10.3389/fgene.2014.00333.
2. Rannala B and Mountain JL (1997) Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences of the United States of America, 94(17): 9197-9201.
3. Weller JI, Seroussi E and Ron M (2006) Estimation of the number of genetic markers required for individual animal identification accounting for genotyping errors. Animal Genetics, 37(4): 387-389. doi:10.1111/j.1365-2052.2006.01455.x.
4. Wright S (1951) The genetical structure of populations. Annals of Eugenics, 15(1):323-354. doi:10.1111/j.1469-1809.1949.tb02451.x.
5. Holsinger KE and Weir BS (2009) Genetics in geographically structured populations: Defining, estimating and interpreting F(ST) Nature Reviews Genetics, 10(9): 639-650, doi:10.1038/nrg2611.
6. Weir BS and Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. Evolution, 38(6): 1358-1370. doi:10.2307/2408641.
7. Akey JM, Zhang G, Zhang K, Jin L and Shriver MD (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Research, 12, doi:10.1101/gr.631202.

8. Kijas JW, Townley D, Dalrymple BP, Heaton MP, Maddox JF, McGrath A and et al. (2009) A Genome Wide Survey of SNP Variation Reveals the Genetic Structure of Sheep Breeds. PLOS ONE, 4, doi:10.1371/journal.pone.0004668.
9. Jost LOU (2008) GST and its relatives do not measure differentiation. Molecular Ecology, 17(18): 4015-4026. doi:10.1111/j.1365-294X.2008.03887.x.
10. Meirmans PG and Hedrick PW (2011) Assessing population structure: FST and related measures. Molecular Ecology Resources, 11(1): 5-18. doi:10.1111/j.1755-0998.2010.02927.x.
11. Nei M and Chesser RK (1983) Estimation of fixation indices and gene diversities. Annals of Human Genetics, 47(3): 253-259. doi:10.1111/j.1469-1809.1983.tb00993.x.
12. Hedrick PW (2005) A Standardized genetic differentiation measure. Evolution, 59(8):1633-1638. doi:10.1111/j.0014-3820.2005.tb01814.x.
13. Smith MW, Lautenberger JA, Shin HD, Chretien JP, Shrestha S, Gilbert DA et al. (2001) Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. American Journal of Human Genetics, 69. doi:10.1086/323922.
14. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW et al. (2007) PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations. PLoS Genetics, 3(9):e160, doi:10.1371/journal.pgen.0030160.
15. Petersen, J. L, Mickelson, J. R, Cothran, E. G, Andersson, L. S, Axelsson, J, Bailey, E, et al. (2013). Genetic Diversity in the Modern Horse Illustrated from Genome-Wide SNP Data. PLOS ONE, 8(1):e54997. doi:10.1371/journal.pone.0054997.
16. R Core Team (2017) R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org/
17. Paetkau D, Calvert W, Stirling, I and Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology, 4(3): 347-354.
18. Rosenberg NA, Li LM, Ward R and Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics, 73, doi:10.1086/380416.
19. Lao O, van Duijn K, Kersbergen P, de Knijff P and Kayser M (2006) Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry, American Journal of Human Genetics, 78, doi:10.1086/501531.
20. Liu N, Chen L, Wang S, Oh C and Zhao H (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure, BMC Genetics, Dec 30;6, doi: 10.1186/1471-2156-6-S1-S26. PMID: 16451635; PMCID: PMC1866760.
21. Smouse PE, Spielman RS and Park MH (1982) Multiple-locus allocation of individuals to groups as a function of the genetic variatio within and differences among human populations, American Naturalist, 119, doi:10.1086/283925.
22. Falush D, Stephens M and Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies, Genetics, 164(4): 1567-1587.
23. Maudet C, Luikart G and Taberlet P (2002) Genetic diversity and assignment tests among seven French cattle breeds based on microsatellite DNA analysis, Journal of Animal Science, 80(4): 942-950. https://doi.org/10.2527/2002.804942x.
24. Negrini R, Nicoloso L, Crepaldi P, Milanesi E, Colli L, Chegdani F and et al. (2009) Assessing SNP markers for assigning individuals to cattle populations, Animal genetics, 40(1): 18-26, doi: 10.1111/j.1365-2052.2008.01800.x.