تولیدات دامی
دوره ۲۴ ■ شماره ۳ ■ پاییز ۱۴۰۱
صفحه‌های ۲۷۰-۲۵۹
DOI: 10.22059/jap.2022.335575.623663

مقاله پژوهشی

مقایسه عملکرد روش‌های تجزیه مؤلفه‌های اصلی و شبکه عصبی مصنوعی در شناسایی نشانگرهای تفکیک در نژادهای مختلف اسب دنیا

سیاوش منظوری[۱]، امیرحسین خلت‌آبادی فراهانی[۲*]، محمدحسین مرادی[۲]، مهدی کاظمی بن‌چناری[۲]
۱. دانشجوی کارشناسی ارشد، گروه علوم دامی، دانشکده کشاورزی و محیط زیست، دانشگاه اراک، اراک، ایران.
۲. دانشیار، گروه علوم دامی، دانشکده کشاورزی و محیط زیست، دانشگاه اراک، اراک، ایران.
تاریخ دریافت مقاله: ۱۴۰۰/۱۰/۰۴ تاریخ پذیرش مقاله: ۱۴۰۱/۰۱/۲۸

چکیده
هدف این پژوهش مقایسه کارایی و عملکرد روش پیشرفته شبکه عصبی مصنوعی نسبت به تجزیه مؤلفه‌های اصلی در تفکیک نژادهای مختلف اسب بود. بنابراین، برای شناسایی یک زیرمجموعه از نشانگرهای SNP با بالاترین قدرت تفکیک نژادی و بررسی نحوه اختصاص حیوانات به گروه‌های نژادی خود از دو روش شبکه عصبی پرسپترون (الدن) و روش کلاسیک تجزیه مؤلفه‌های اصلی (PCA) استفاده شد. نتایج حاصل نشان داد روش شبکه عصبی (الدن) قادر است که ۳۷ نژاد اسب موردمطالعه در این پژوهش کنونی را، با زیرمجموعه کوچکی از نشانگرهای SNP (۸۰۰۰ نشانگر) و با قدرت تفکیک مشابه با تمام نشانگرهای ژنوم (صحت ۹۸ درصدی)، از همدیگر مجزا و تفکیک کند. روش انتخاب PCA تنها توانست نژادهایی که دارای خاستگاه‌های متفاوت بودند را شناسایی و تفکیک کند. با توجه به نتایج به‌دست‌آمده، روش PCA دارای خطا و ایراد بوده و برای اجرا روی داده‌های ژنومی نیاز به تغییرات و اصلاحات دارد. نتایج این پژوهش، رویکردهای عملی را در طراحی آرایه‌های اقتصادی در تفکیک نژادهای مختلف اسب ارائه می‌دهد.

کلیدواژه‌ها: آنالیز تعیین نژاد، آنالیز مؤلفه‌های اصلی، ساختار ژنتیکی، شبکه عصبی مصنوعی، نژادهای اسب.

Comparing the performance of principal component analysisand Artificial Neural Network methods in identifying the discriminating SNP(s) in different horse breeds of the world

Siavash Manzoori[1], Amir Hossein Khaltabadi Farahani[2*], Mohammad Hossein Moradi[2], Mehdi Kazemi Bon-Chenari[2]
1. M.Sc. Student, Department of Animal Science, Faculty of Agriculture and Natural Resources, Arak University, Arak, Iran
2. Associate Professor, Department of Animal Science, Faculty of Agriculture and Natural Resources, Arak University, Arak, Iran
Received: December 25, 2021 Accepted: April 17, 2022

Abstract
The aim of this research was to compare the efficiency and performance of the advanced artificial neural network method with the principal component analysis method in discriminating different horse breeds. In this study, two methods of perceptron neural network (Olden) and the principal component analysis (PCA), were used to identify a subset of SNP markers with the highest breed discrimination potential and to investigate how to assign animals to their breed groups. The results showed that the network method (Olden), is able to separate all the 37 horse breeds with a small subset of SNP markers (8,000 markers) with a same capability to all genomic markers (98% accuracy). The PCA selection method was only able to identify and separate breeds with diverse geographical originations. According to the obtained results, the PCA method is not error-free and depends upon changes and modifications to run on genomic data. The results of this study provide practical approaches in the design of economic arrays for discriminating the different horse breeds.

Keywords: Artificial neural network, Assignment analysis, Genetic structure, Horse breeds, Principal Component Analysis.

* نویسنده مسئول
Email: a-farahani@araku.ac.ir

مقدمه

ازنظر علمی اسب‌های امروزی از گونه کابالوس بوده که طبق آخرین گزارش منتشرشده در سال ۲۰۲۱، طول ژنوم این حیوان به‌طور تقریبی از ۲/۵ گیگا جفت باز (Gbp) تشکیل شده است [۱]. نشانگر SNP یکی از نشانگرهایی است که دارای محبوبیت و مزایایی هم‌چون نرخ پایین خطای تعیین ژنوتیپ و تعداد بسیار فراوان در سرتاسر ژنوم است. دانشمندان در تعیین ساختار ژنتیکی به‌دنبال شناسایی و مشخص‌کردن مبدأ افراد هستند. مزیت شناسایی و طبقه‌بندی افراد براساس ساختار ژنتیکی در زمینه‌های شناسایی پدیده مهاجرت افراد و از همه مهم‌تر مدیریت برنامه‌های حفظ نژادی و اجرای استراتژی‌های اصلاحی می‌باشد. در حال حاضر با توجه به ضعف اطلاعات شجره‌ای و محدودیت‌های رکوردبرداری صحیح در اصلاح نژاد ضرورت تخصیص نژادی بیش از پیش نمایان شده است. امروزه داده‌های آزمایشی دارای حجم وسیع‌تری نسبت به گذشته می‌باشند و به همین دلیل تجزیه وتحلیل و نیز تفسیر نتایج برای پژوهش‌گران به چالش جدی تبدیل شده است. یک راه‌حل برای رفع این مشکلات، کاهش ابعاد داده‌ها می‌باشد. هدف از روش PCA دست‌یابی به بهترین ترکیب خطی از متغیرها است که بیش‌ترین درصد تنوع و تغییر را کنترل می‌کنند [۲]. تجزیه مؤلفه‌های اصلی (PCA) به شناسایی مؤلفه‌های اصلی براساس همبستگی ژنتیکی میان حیوانات می‌پردازد که ساختار جمعیت را شناسایی نماید.

شبکه‌های عصبی از جدیدترین روش‌های محاسباتی در علوم به‌شمار می‌آیند. گره‌ها در شبکه‌های عصبی واحدهای محاسباتی شناخته می‌شوند و ارتباط بین گره‌ها می‌تواند بسته به نوع عملکرد گره یک‌طرفه (ورودی به خروجی) یا دوطرفه (ورودی با پردازش، پردازش دوباره به خروجی) دیده شود. شبکه‌های عصبی برای تخمین (Estimation) و تقریب (Approximation) کارایی بسیار بالایی از خود نشان داده‌اند و در شناسایی الگوها قابلیت زیادی دارند [۳]. بیش‌تر پژوهش‌ها در زمینه استفاده از روش‌های PCA و شبکه عصبی به انسان و دام‌هایی هم‌چون گاو محدود شده است. البته الگوریتمی جدید بر پایه روش اصلی PCA ارائه شد که توانایی کار با داده‌های ژنومی را دارد [۲]. الگوریتم پیشنهادی به‌نحوی عمل می‌کرد که توانایی انتخاب نشانگرهایی را داشته که ساختار شناسایی‌شده به‌وسیله آن نشانگرها تا حد نسبتاً زیادی با ساختار شناسایی‌شده توسط کل نشانگرها مطابقت دارد. هم‌چنین به ارزیابی روش‌های کلاسیک (به‌ویژه PCA) برای انتخاب نشانگر پرداخته شد. آن‌ها گزارش کردند که روش PCA نسبت به دیگر روش‌ها عملکرد ضعیفی داشته و برای رسیدن به یک سطح مطلوب (۹۵ درصد) در آنالیز تعیین نژاد، به تعداد نشانگر بیش‌تری نسبت به دیگر روش‌ها نیاز است [٤].

تجزیه و تحلیل مؤلفه اصلی (PCA) نیز به‌تازگی به‌عنوان یک روش جایگزین برای تعیین و شناسایی نشانگرهای SNP مؤثر در ساختار جمعیت پیشنهاد شده است. این روش در گذشته در جمعیت‌های انسانی [۲] برای مشخص‌کردن ساختار جمعیت براساس داده‌های ژنوتیپی و در گاو برای شناسایی SNPهای تخصیص نژادی استفاده شده است [٥]. شبکه‌های عصبی مصنوعی روش‌هایی هستند که می‌توانند مدل‌سازی آماری غیرخطی را اجرا کنند و به‌عنوان یک راه جایگزینی برای روش‌های رگرسیونی باشند [٦]. در این پژوهش به عملکرد و کارایی دو روش آنالیز مؤلفه‌های اصلی (Principal Component Analysis - PCA) و شبکه عصبی مصنوعی (Artificial Neural Networks-ANN) در تعیین ساختار و طبقه‌بندی افراد

با رویکرد کاهش ابعاد پرداخته می‌شود. هدف از این پژوهش، تعیین مزیت استفاده از روش پیشرفته شبکه عصبی مصنوعی و مقایسه با روش کلاسیک PCA در تفکیک نژادهای مختلف اسب می‌باشد. در این پژوهش، از اطلاعات ژنومی ۳۷ نژاد اسب در سراسر دنیا استفاده شده است و سعی بر این است تا نقش مهمی در طراحی کیت‌های اقتصادی در تشخیص نژادهای مختلف اسب داشته باشد.

مواد و روش‌ها

در پژوهش حاضر، از اطلاعات ژنومی ۵۴۶۰۲ جایگاه نشانگری SNP مربوط به ۷۹۵ اسب متعلق به ۳۷ نژاد در سراسر جهان استفاده شد (جدول ۱). این اطلاعات با همکاری پروژه کنسرسیوم تنوع ژنتیکی اسب (Equine Genetic Diversity Consortium) تهیه شد که پیش‌ازاین، در پژوهشی دیگر [۷] نیز استفاده شده است که هدف آن هم‌راستا و هم‌جهت با پژوهش کنونی نبوده است.

جدول ۱. نام نژاد، شناسه آنالیزی، اندازه نمونه و فراوانی آلل کمیاب (MAF) در هر نژاد

نژاد	شناسه	تعداد	MAF[1]	نژاد	شناسه	تعداد	MAF
آخال‌تکه	AKTK	۱۹	۰/۲۳۵۱	نیوفارست	NFST	۱۵	۰/۲۱۷۰
آندولوسین	AND	۱۸	۰/۲۲۸۹	سوئدی شمال	NSWE	۱۹	۰/۲۱۰۰
عرب	ARR	۲۴	۰/۲۴۰۱	آبدره نروژی	NORF	۲۱	۰/۲۰۸۸
بلژین	BEL	۳۰	۰/۲۰۸۷	پینت	PT	۲۵	۰/۲۴۴۲
کاسپین	CSP	۱۸	۰/۲۲۳۰	پرچرون	PERC	۲۳	۰/۲۰۸۵
کلایدسدال	CLYD	۲۴	۰/۲۰۴۷	پاسو پرویی	PERU	۲۱	۰/۲۲۱۶
اکسمور	EXMR	۲۴	۰/۲۰۹۷	پاسوفینو پورتوریکو	PRPF	۲۰	۰/۲۱۸۷
فل	FELL	۲۱	۰/۲۱۲۰	کوارتر	QH	۴۰	۰/۲۴۴۵
فنلاندی	FIN	۲۷	۰/۲۰۹۴	سادل برد	SB	۲۵	۰/۲۳۳۴
ترقه فلوریدایی	FLCR	۷	۰/۲۲۹۰	شتلاند	SHET	۲۷	۰/۲۰۷۸
فرنچس‌مونتانس	FM	۱۹	۰/۲۲۰۹	شایر	SHR	۲۳	۰/۲۱۰۹
تروتر فرانسوی	FT	۱۷	۰/۲۴۲۱	استانداربرد نروژ	STBD-Nor	۲۵	۰/۲۳۷۹
هانوورین	HAN	۱۵	۰/۲۵۲۰	استانداربرد آمریکا	STBD-US	۱۵	۰/۲۴۰۹
ایسلندی	ICE	۲۵	۰/۲۱۱۸	سوئیسی خونگرم	SZWB	۱۴	۰/۲۵۱۷
لوسیتانو	LUST	۲۴	۰/۲۲۷۵	تروبرد انگستان	TB-UK	۱۹	۰/۲۷۴۸
مانگالارگا پائولیستا	MNGP	۱۵	۰/۲۲۴۲	تروبرد آمریکا	TB-US	۱۷	۰/۲۷۴۸
مارم مانا	MARM	۲۴	۰/۲۳۹۴	تووا	TUVA	۱۵	۰/۲۱۶۲
مینیاتور	MINI	۲۱	۰/۲۰۹۶	–	–	–	–
مغولی	MON	۱۹	۰/۲۱۲۵	–	–	–	–
مورگان	MOR	۴۰	۰/۲۲۵۷	–	–	–	–

۱. آماره MAF نشان‌دهنده فراوانی آلل کمیاب، برای تمام نشانگرهای SNP در هر نژاد می‌باشد.

ژنوتیپ نمونه‌ها با استفاده از آرایه‌های Illumina SNP50K BeadChip براساس دستورالعمل استاندارد شرکت ایلومینا تعیین شد. فرایند کدگذاری ژنوتیپ‌ها به‌نحوی صورت گرفت که ژنوتیپ AA با کد صفر (۰)، AB با کد یک (۱) و در نهایت ژنوتیپ BB با کد دو (۲) جایگزین شدند. مرحله کنترل کیفیت، تمامی نشانگرها با مقدار کم‌تر از ۰/۰۵ آماره MAF، نشانگرهایی با ارزش‌های کم‌تر از ۰/۹۵ آماره Sample Call Rate، SNPهایی با مقدار انحراف بیش از ۱۰e-۶ آماره تعادل هاردی- وینبرگ و در نهایت مارکرهایی با ارزش‌های کم‌تر از ۰/۹۹ آماره SNP Call Rate از آنالیز کنار گذاشته شدند. میانگین مقدار MAF در تمام نمونه‌ها ۰/۲۲۶۷ و حداقل و حداکثر MAF مشاهده‌شده به‌ترتیب در نژادهای Clydesdale و Throughbred-UK/Ire (۰/۲۰۴۷، ۰/۲۷۴۸ و ۰/۲۷۴۸) مشاهده شد.

هدف از تجزیهٔ مؤلفه‌های اصلی آن است که واریانس موجود در داده‌های چندمتغیره را به مؤلفه‌هایی تجزیه کند که نخستین مؤلفه تا آنجاکه ممکن است بیش‌ترین واریانس موجود در داده‌ها را توضیح دهد. دومین مؤلفه علت بیش‌ترین واریانس ممکن پس از مؤلفه اول باشد و این رویه الی آخر به همین منوال پیش برود. افزون بر این، در این روش هر مؤلفه مستقل از مؤلفه‌های دیگر است، یعنی بین هر مؤلفه و مؤلفه‌های دیگر همبستگی وجود ندارد. داده‌های موجود برای تفکیک نژادها از همدیگر با استفاده از روش کلاسیک آنالیز مؤلفه‌های اصلی (PCA) بررسی شدند. شاخص‌های حاصل از PCA باید از (رابطه ۱) پیروی کنند.

$$Var(Z_1) \geq Var(Z_2) \geq Var(Z_3) \geq \cdots \geq Var(Z_p) \quad (۱)$$

در این رابطه Var(Z_i)، واریانس Z_i را نشان می‌دهد. Z_iها را مؤلفه‌های اصلی می‌نامند. بعد از انجام PCA روی کل داده‌ها، نتایج حاصل از چهار تأثیرگذاری به‌دست‌آمده را برای محاسبه اهمیت نسبی هر نشانگر استفاده شد. تجزیه و تحلیل مؤلفه اصلی (PCA) با استفاده از تابع procomp در بستر نرم‌افزار R انجام شد.

در این پژوهش، از شبکه عصبی تک لایه پرسپترون و تابع فعال‌سازی سیگموئیدی و الگوریتم پس انتشار خطا برای تصحیح اوزان استفاده شد. برای محاسبه اهمیت هر متغیر ورودی به شبکه، از الگوریتم اُلدن استفاده شد که همان اوزان ارتباطی ((Olden (Connection weights) می‌باشند [۸]. این الگوریتم اهمیت نسبی هر متغیر ورودی را براساس وزن‌های سیناپتیکی طبق (رابطه ۲) حساب می‌کند [۸].

$$R_{ij} = \sum_{H=1}^{h} W_{ih} \cdot W_{hj} \quad (۲)$$

در این رابطه R_{ij}، اهمیت نسبی متغیر پیش‌بینی‌کننده x_i با توجه به نرون j در لایه خروجی است. h، نیز نمایانگر تعداد نرون در لایه پنهانی است. W_{ih} و W_{hj} به‌ترتیب اشاره بر اوزان سیناپتیکی بین نرون i ورودی و نرون مخفی h و اوزان سیناپتیکی بین نرون مخفی h و نرون خروجی j دارند.

چندین روش قابل‌قبول برای تعیین نژادهای مختلف وجود دارند. مقدار احتمال درست‌نمایی (Likelihood) نحوه اختصاص ۷۹۵ حیوان مورد آنالیز در پژوهش کنونی به گروه (نژاد) واقعی خود به‌وسیله روش [۹] محاسبه شد. با استفاده از رابطه (۳) نسبت لگاریتم درست‌نمایی (Log-Likelihood Ratios-LLR) به‌وسیله مقایسه درست‌نمایی یک فرد که به نژاد خودش انتصاب شده با درست‌نمایی آن فرد که به نژاد دیگری منصوب شده محاسبه شد.

$$LLR = log_{10}\big(T(\mathrm{g}|i_a)\big) - log_{10}\big(T(\mathrm{g}|i_b)\big) \quad (۳)$$

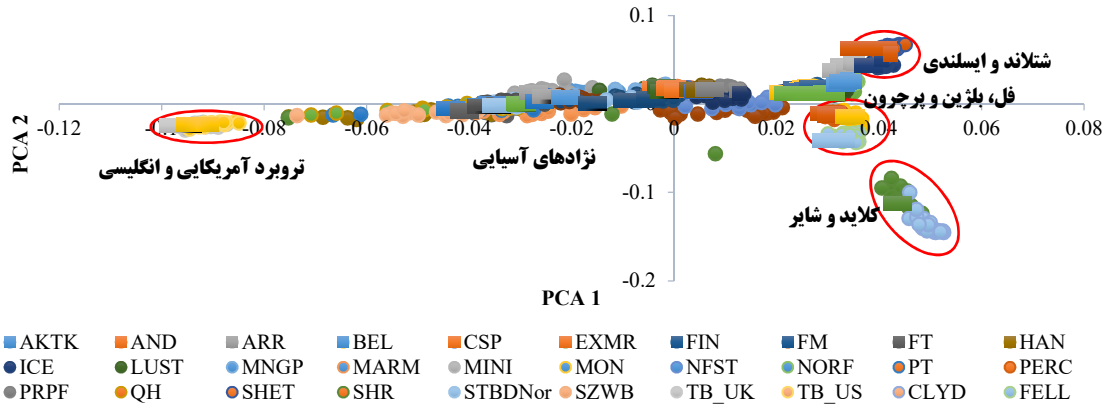
برای اطمینان بیش‌تر از صحت عملکرد روش، آستانه‌های سخت‌گیرانه متفاوتی بر روی نتایج اعمال شد. چهار آستانه مورد استفاده عبارتند از ٤ و ۳، ۲، ۱ <LLR که به‌ترتیب به این معنی هستند که یک ژنوتیپ (فرد) به احتمال ۱۰، ۱۰۰، ۱۰۰۰ و ۱۰۰۰۰ برابر در نژاد واقعی خود (تا در نژاد غیر واقعی) قرار می‌گیرد. تمامی آنالیزها

در نرم‌افزار آماری R (3.4.0) اجرا شدند [۱۰]. برای انجام آنالیزهای PCA از بسته stats و برای ساخت شبکه از بسته Neuralnet استفاده شد [۱۱]. از بسته Neural Net Tools نیز برای نمایش نتایج و درک بهتر مدل‌های شبکه عصبی استفاده شد [۱۲].

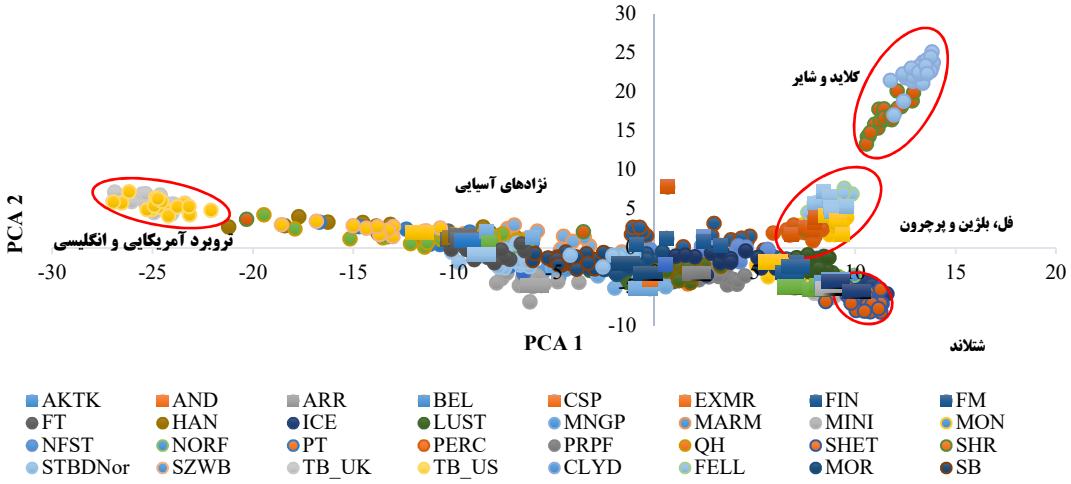
نتایج و بحث

تجزیه مؤلفه اصلی یک روش رایج در ژنتیک جمعیت‌ها، نژادها و گروه‌هاست که به بررسی ساختار در جوامع می‌پردازد و قابلیت اعمال بر داده‌هایی با ابعاد بزرگ را دارد. این روش از نظر محاسباتی بسیار سریع می‌باشد [۱۳]. به‌منظور داشتن یک نمای کلی از ساختار جمعیت نژادهای موردنظر، تجزیه‌وتحلیل مؤلفه‌های اصلی (PCA) توسط تمام نشانگرهای SNP انجام شد که مراحل کنترل کیفیت را گذرانده بودند. بعد از اجرای PCA با تمامی نشانگرها، ساختار ایجادشده به‌وسیله ۳۷ نژاد در شکل (۱) نشان داده شده است.

بیش‌ترین پراکنش مربوط به نژادهای اروپایی بود که همپوشانی و اشتراک بالایی داشتند (شکل ۱). اکثر نژادهای آسیایی و آمریکایی در مرکز شکل قرار گرفته‌اند. نتایج به‌دست‌آمده نشانگر این موضوع است که نژادهای مختلف اسب تحت تأثیر عوامل تکاملی قرار گرفته و هر نژاد برای یکسری از صفات (ژن‌ها) تخصصی شده‌اند و این مورد را می‌توان در نژادهای مربوط به منطقه اسکاندیناوی مشاهده نمود، چرا که آن‌ها جزو نژادهای اروپایی هستند اما از همه نژادهای اروپایی جدا شده و در یک منطقه قرار گرفته‌اند. این بدین معناست که بعضی از صفات اقتصادی هم‌چون باربری، سرعت، استقامت و ... تحت تأثیر این عوامل قرار گرفته‌اند و بیش‌تر تغییر در ساختار ژنتیکی نژادها مربوط به این صفات است. پس از آنالیز اولیه داده‌ها و بررسی نحوه اختصاص حیوانات به گروه‌های نژادی خود با استفاده از کل نشانگرها، ابتدا با استفاده از شبکه عصبی پرسپترون (روش اُلدن) زیرمجموعه‌ای از نشانگرهای دارای بیش‌ترین قدرت تفکیک نژادی انتخاب شدند. معیار انتخاب نشانگرها، میزان تمایز و تفکیک نژادها بعد از انتخاب و پلات PCA بود. در جایی که دیگر نژادها قابلیت متمایزترشدن و تفکیک بیش‌تر را نداشتند، عملیات انتخاب نشانگر متوقف شد. براساس نتیجه این معیار، در زیرمجموعه نشانگری انتخاب‌شده به‌وسیله روش شبکه عصبی پرسپترون (اُلدن) حدوداً ۸۰۰۰ نشانگر وجود داشت که حدود ۱۶ درصد از کل نشانگرها را شامل می‌شد (شکل ۲).



شکل ۱. خوشه‌بندی حیوانات بر پایه آنالیز مؤلفه‌های اصلی با استفاده از تمامی نشانگرها

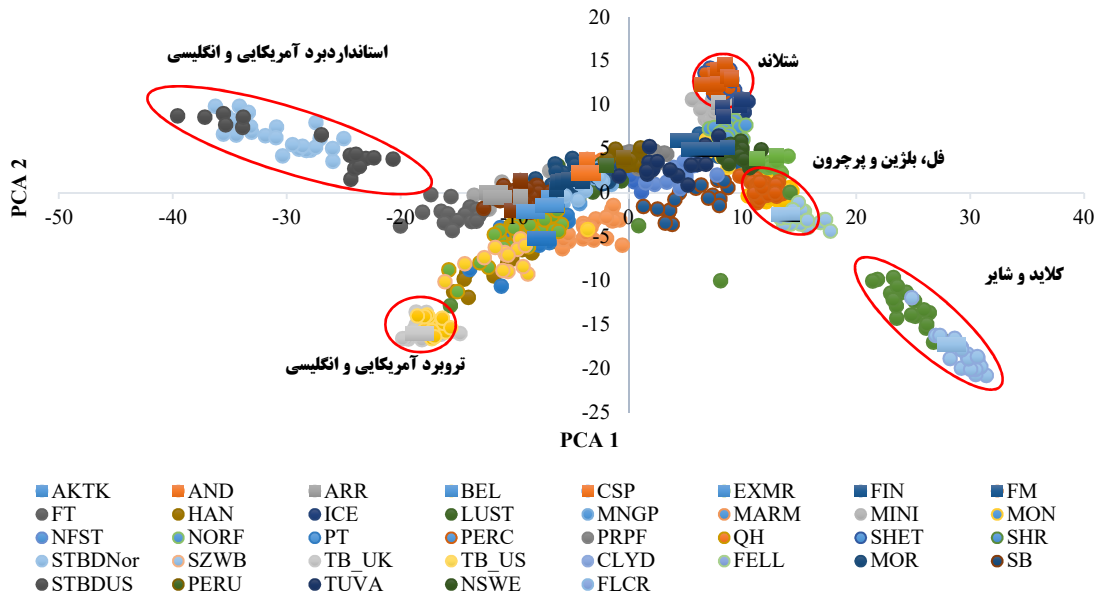


شکل ۲. پلات PCA مربوط به ۳۷ نژاد اسب حاصل از ۸۰۰۰ نشانگر SNP تشخیصی به‌دست‌آمده از روش الدن.

پس از استفاده از روش شبکه عصبی (اُلدن)، شناسایی نشانگرهای تفکیک نژادی با استفاده از روش آماری PCA نیز انجام شد، که زیرمجموعه‌ای حاوی ٤٢٢٧ نشانگر SNP به‌دست آمد. در روش PCA بعضی از نژادها با خاستگاه جغرافیایی متفاوت به‌عنوان یک گروه شناسایی شدند (شکل ۳). دلیل انتخاب بیش‌تر نشانگر در روش شبکه عصبی (اُلدن)، این بود که در تعداد پایین نشانگرهای SNP نسبت به روش PCA، کیفیت و تمرکز نژادها به‌خوبی قابل تشخیص نبود.

نتایج حاصل از بررسی روش مبتنی بر شبکه عصبی نشان داد که این روش قادر است ۳۷ نژاد موردمطالعه را با زیرمجموعه کوچکی از نشانگرهای SNP و با قدرت تفکیک مشابه با کل نشانگرهای ژنوم اسب از همدیگر مجزا و تفکیک کنند. به بیانی دیگر با به‌کارگیری ۸۰۰۰ نشانگر SNP تشخیصی در روش اُلدن می‌توان ساختاری مشابه ساختار اولیه داده‌ها به‌دست آورد. اما روش PCA ساختار متفاوتی با روش‌های شبکه و تمام نشانگرها ارائه داد. با این‌حال، در دو روش آنالیزی نژادهای تروبرد انگلستان و آمریکا همانند نژادهای اسکاندیناوی (فاین، نروژی خونگرم و ایسلندی) با همپوشانی بالا و قرارگرفتن در یک ناحیه مشترک به‌عنوان یک گروه شناسایی شدند. به‌عبارت دیگر، این نژادها یا دارای جریان ژنی بالا در بین یکدیگر هستند و یا هنوز برنامه اصلاح نژادی اعمال‌شده بر روی آن‌ها تمرکز بر حفظ خلوص نژادی دارد.

تنها تفاوت در روش PCA این بود که به‌غیراز نژادهای شایر، کلاید، تروبرد (آمریکا و انگلستان) و شتلاند، که در روش شبکه به‌طور مجزا واقع شدند، نژادهای استانداردبرد نروژ و آمریکا نیز از بقیه نژادهای مورد آنالیز جدا شدند. نتایج حاصل از نمودارهای دو روش، نشان داد که به‌غیراز تفاوت‌های ساختار ژنتیکی، تفاوت‌های جغرافیایی نیز به‌وسیله روش PCA نمایان شدند [١٤]. جدایی جغرافیایی خاستگاه افراد نیز مورد بررسی قرار گرفت و گزارش شده است که به‌طورکلی، ساختار ژنتیکی و نشانگرهای تخصیصی قابلیت انتقال در بین مناطق جغرافیایی را ندارد [١٥]. این جمله به این معنی است که بعضی از جهش‌ها (ژن‌ها) مختص یک نژاد خاص بوده و حتی ممکن است آن جهش‌ها (ژن‌ها) در حیوانات هیبرید نیز بیان نشوند.



شکل ۳. پلات PCA مربوط به ۳۷ نژاد اسب حاصل از ٤٢٢٧ نشانگر SNP تشخیصی به‌دست‌آمده به روش آنالیز PCA.

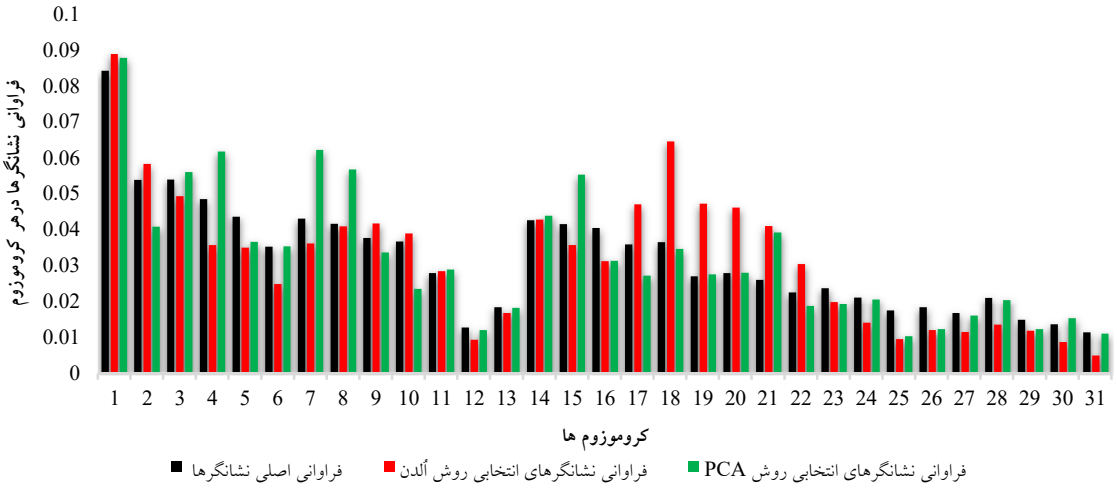
نژادهایی که در یک منطقه جغرافیایی بودند با همدیگر در یک ناحیه از نمودارهای PCA قرار گرفتند. البته لازم به ذکر است که در روند و سیر تکاملی اسب عوامل زیاد دیگری به‌غیر از جدایی جغرافیایی دخیل می‌باشند. از نظر مشارکت کروموزوم‌ها در فرایند تفکیک نژادها نیز تفاوتی بین این دو روش مشاهده شد. تعداد نشانگرهای انتخابی دو روش PCA و شبکه عصبی (الدن) در هر کروموزوم محاسبه شد که نمودار ستونی آن در شکل (٤) نمایش داده شده است.

در شکل (٤)، روش شبکه عصبی (اُلدن) رفتار متفاوتی داشته، به‌طوری‌که کروموزوم‌های ۲-۱ و ۲۲-۱۷ سهم بیش‌تری در زیرمجموعه نشانگرهای تشخیصی داشته‌اند. اما در کروموزوم شماره ۳۱، همان‌گونه که کم‌ترین تعداد نشانگر وجود دارد، روش بر پایه شبکه نیز کم‌ترین نشانگر را از این کروموزوم انتخاب نموده است. روش PCA در تمام ژنوم دارای رفتاری غیرطبیعی بوده و اکثر کروموزوم‌ها سهم بیش‌تری در انتخاب نشانگرها نسبت به روش شبکه دارد. نتیجه‌ی دیگری که از شکل (٤) می‌توان گرفت این است که در دو روش آنالیزی کروموزوم‌های ۳۱-۲٤ مشارکت کمی در فرایند انتخاب نشانگر داشتند.

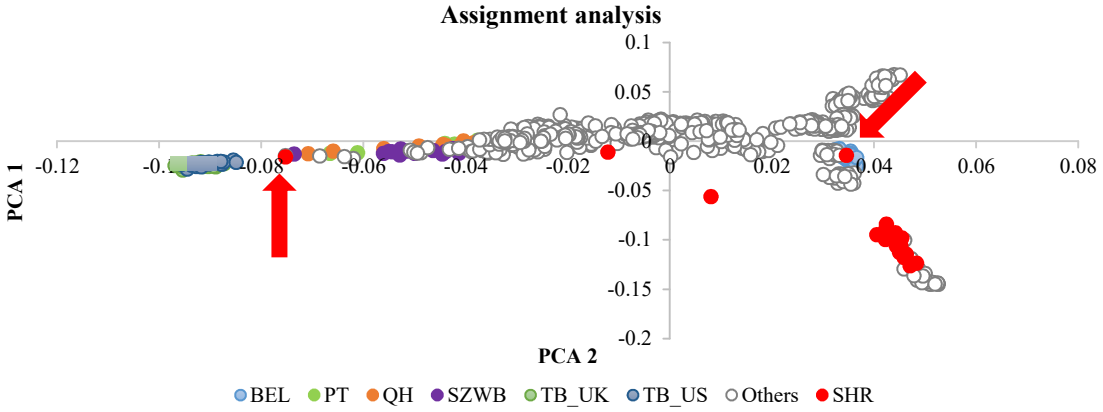
نتایج مربوط به مشارکت کروموزومی (چه روش کلاسیک و چه روش شبکه عصبی) هم‌راستا و موافق با نتایج پژوهش‌گران مختلفی می‌باشد. با توسعه یک روشی جدید، پژوهش‌گران موفق به شناسایی کامل نژادهای مختلف گاو در سطح کروموزومی شدند [١٦]. هم‌چنین با استفاده از کروموزوم‌های ۱ و ٦ ژنوم می‌توان موفق به تفکیک نژادهای مختلف گوسفند ایتالیایی شد [١٧]. در گونه گاو نیز، با استفاده از روش‌های کلاسیک می‌توان به شناسایی هاپلوتیپ‌های واقع‌شده بر روی کروموزوم ۱۹ گاو پرداخت [١٨]. پژوهش‌گران دیگر بیان داشتند که با نشانگرهای واقع‌شده بر روی کروموزوم‌های ۲، ٦ و ۲۰ تمایز گروه‌های (نژادهای) مختلف گاو به حداکثر ممکن می‌رسد [١٩]. در پژوهشی باهدف شناسایی حداقل تعداد

نشانگرهای SNP تخصیصی با قابلیت اعتماد بالا در نژادهای گوسفند بومی به ردیابی نژاد واقعی نمونه‌های ناشناخته پرداختند [۲۰]. نژاد گاو بومی Tharparkar به‌دلیل کیفیت بالای شیر و صفات مقاومتی آن بسیار در هند مشهور است. با تجزیه و تحلیل داده‌های ژنوتیپی به‌دست‌آمده از ۳۱۷ فرد موفق به شناسایی SNPهای نژادی برای انتخاب گاو نژاد Tharparkar شدند. موقعیت بیش‌تر نشانگرهای انتخابی را روی کروموزوم‌های ۲۲–۲۰ اعلام شد [۲۱].

در شکل (۵) حیواناتی از نژاد شایر که خارج از گروه نژادی خود قرار گرفته‌اند (فلش‌های قرمز رنگ) به همراه نژادهای درگیر با آن‌ها (رنگ‌های متفاوت) مشخص شده‌اند. منظور از افراد خارج از گروه، حیواناتی هستند که به یک نژاد خاص تعلق دارند ولی براساس پلات‌های PCA و شبکه عصبی خارج از گروه نژادی خویش قرار گرفته‌اند. هم‌چنین منظور از نژادهای درگیر در شکل (۵) آن نژادهایی هستند که در آنالیز تعیین نژاد به‌عنوان نژاد مبدأ این افراد خارج از گروه شناسایی و اعلام شده‌اند.



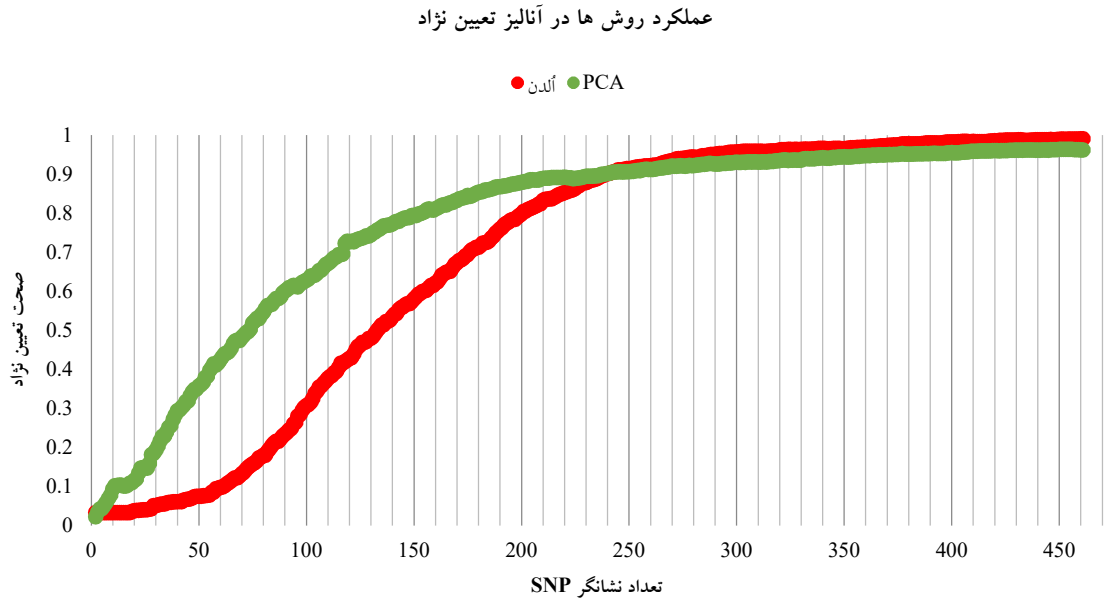
شکل ٤. فراوانی نشانگرهای انتخابی دو روش در هر کروموزوم



شکل ۵. افراد خارج از گروه نژاد شایر (با پیکان قرمز رنگ) و نژادهای درگیر با آن (رنگ‌های دیگر) مشخص شده‌اند.

در روش شبکه عصبی پرسپترون (اُلدن) یکی از افراد نژاد شایر در نژاد بلژین قرار گرفت و دیگری نیز در منطقه‌ای مابین نژادهای پینت، کوارتر و تروبرد (آمریکایی و انگلیسی) قرار گرفت. اما روش PCA به‌علت انتخاب نشانگرهایی که مختص مناطق جغرافیایی بودند، نتوانست عملکرد بالایی از خود نشان دهد (حدود ۱۰ فرد را به‌صورت نادرست، هیبرید شناسایی کرد). به‌منظور شفاف‌شدن نتایج آنالیز تعیین نژاد و برای بررسی بیش‌تر عملکرد روش‌ها، با مرتب‌کردن نشانگرهای SNP براساس ضرایب به‌دست‌آمده از هر روش، منجر به انتخاب ٤٦٠ نشانگرهای SNP شد که با صحت ۹۹ درصد قادر به تفکیک افراد براساس گروه نژادی خود بودند. در نمودار ذیل (شکل ٦) عملکرد دو روش PCA و شبکه عصبی در تفکیک حیوانات به گروه‌های نژادی خود با استفاده از زیرمجموعه‌های مختلف نشانگری نشان داده شده است.

در روش کلاسیک PCA سیر صعودی درصد صحت اختصاص حیوانات به گروه‌های نژادی خود در ۲۵۰ نشانگر اول مشاهده شد، اما افزایش تعداد نشانگرهای بعدی تأثیر چندانی در بهبود صحت آنالیز نداشت (شکل ٦). به‌طوری‌که با افزایش تعداد نشانگر SNP حتی تا ۸۰۰ نشانگر نیز باعث نشد تا صحت روش PCA به ۹۸ درصد برسد. روش PCA براساس ماتریس کوواریانس و اوزان بردار ویژه متغیرها را انتخاب می‌کند، که به‌طور مستقیم بر میزان واریانس توجیح شده توسط نشانگرهای SNP تفکیک‌کننده تأثیر می‌گذارد. هرچند روند صحت تفکیک در ابتدای آنالیز در روش PCA بالا بوده ولی در شبکه عصبی در مراحل اولیه، شبکه آموزش داده می‌شود و بالطبع صحت تفکیک آن براساس نشانگرهای اولیه انتخاب‌شده کم‌تر می‌باشد، اما در مراحل بعدی صحت و دقت شبکه نسبت به PCA افزایش پیدا می‌کند و میزان نشانگرهای انتخابی با اطلاعات کاذب را به حداقل می‌رساند.



شکل ٦. درصد صحت اختصاص حیوانات به گروه‌های نژادی خود با تعداد نشانگر SNP متفاوت با استفاده از دو روش PCA (خط سبزرنگ) و شبکه عصبی اُلدن (خط قرمز رنگ)

از طرفی همان‌طورکه در شکل (٦) مشاهده می‌شود، در روش شبکه عصبی پرسپترون (اُلدن)، نشانگر کم‌تری (۵۰۰ نشانگر کم‌تر) نسبت به روش PCA نیاز است تا به صحت ۹۵ درصد دست یافت. هرچند منحنی روش شبکه عصبی شروع ضعیفی داشت، اما بعد از ۲٤۰ نشانگر از روش کلاسیک PCA پیشی گرفت و با ادامه روند (۳۱۰ نشانگر)، میزان صحت به بالای ۹۵ درصد رسید. در جدول (۲) میزان نشانگرهای SNP موردنیاز برای رسیدن به حد مطلوب در هر روش آورده شده است. در این جدول سه حد مطلوب صحت ۹۰، ۹۵ و ۹۸ درصد و ۵ سطح آنالیزی لگاریتمی مدنظر قرار داده شده است.

جدول ۲. تعداد نشانگرهای موردنیاز برای رسیدن به سطوح مطلوب در دو روش

لگاریتم ۱۰	الدن			PCA		
	۹۰ درصد	۹۵ درصد	۹۸ درصد	۹۰ درصد	۹۵ درصد	۹۸ درصد
۰	۱۷۰	۱۹۵	۲۷۰	۱۷۱	۲۹۳	*
۱	۲۰۵	۲٤۲	۳٦۳	۲۰۳	۳۵۷	*
۲	۲۳٦	۲۸۳	۳۹٤	۲۵۳	*	*
۳	۲٦٤	۳۲۰	٤۲۲	۲۹٤	*	*
٤	۲۸۷	۳٦۷	*	۳۳۱	*	*

*** به بیش از ٤٦۰ نشانگر SNP اولیه برای حصول صحت ۱۰۰ درصد نیاز است.**

روش انتخاب PCA تنها توانست نژادهایی که دارای خاستگاه‌های متفاوت (جداگانه) بودند را شناسایی و تفکیک کند. روش PCA برای نژادهایی که در یک منطقه جغرافیایی خاص وجود دارند و احتمال تداخل ژنتیکی بین آن نژادها بسیار بالا است، دارای عملکرد بسیار خوب و قابل‌قبولی نیست. چراکه این تداخل ژنی در PCA باعث عدم تفکیک نژادها می‌شود و از نظر منطقی قابل درک می‌باشد. نتایج به‌دست‌آمده از نمودار PCA نیز بیانگر این حقیقت است. به‌عنوان مثال، می‌توان به نژادهای آسیایی اشاره نمود که دارای بیش‌ترین هم‌پوشانی با یکدیگر هستند. اما قدرت تفکیک نژادی شبکه عصبی به‌طور خاص، زمانی بروز می‌کند که نژادها از نظر جغرافیایی نزدیک به هم و یا دارای تداخل نژادی باشند که در بیش‌تر نژادهای اسب جریان ژنی فراوانی وجود دارد. این نتیجه از روش PCA هم‌راستا و همسو با نتایج پژوهش‌گرانی مانند [۲، ١٤ و ۲۲] بود. آن‌ها بیان داشتند که روش PCA قادر به شناسایی مهاجرت‌ها و جمعیت‌های ایزوله‌شده به‌وسیله عوامل مختلف هستند. با هدف بررسی ساختار و لایه‌بندی جمعیت، دو اکوتیپ آذری (استان‌های آذربایجان‌شرقی، آذربایجان‌غربی و اردبیل) و شمالی (استان گیلان) موردبررسی قرار گرفتند. با توجه به نمودارهای حاصله از تجزیه مؤلفه‌های اصلی و دیگر روش‌ها، تمامی روش‌های آنالیزی توانایی جداسازی افراد به گروه‌های اصلی را دارا بودند و احتمالاً جریان ژنتیکی نزدیکی مابین افراد این چهار استان و دو اکوتیپ شمالی و آذری وجود دارد. براساس نتایج پژوهش تعداد ۲٦ مؤلفه اصلی (PCA) اول حدود ۲۰ درصد واریانس را در این جمعیت‌ها توجیه می‌کنند که پایین بودن مقدار واریانس دلالتی بر کاهش تمایز بین دو اکوتیپ می‌باشد [۲۳].

در پژوهشی، ساختار جمعیتی گاومیش‌های ایران (از نژاد شمالی، آذری و خوزستانی) به‌وسیله دو روش تجزیه و تحلیل مؤلفه‌های اصلی (PCA) و تجزیه و تحلیل تشخیصی مؤلفه‌های اصلی (DAPC) ارزیابی شدند. نتایج حاصله از دو روش، منجر به جداسازی سه نژاد از یکدیگر شد و هر دو روش ساختار ژنتیکی جمعیت‌های موردبررسی را به خوبی نشان دادند [٢٤]. نتایج رویکرد PCA قادر به جداسازی نژادهای ایرانی گوسفند از یکدیگر می‌باشد، اما برای جداسازی نژادهای مغانی و قزل که نزدیک‌ترین فاصله ژنتیکی (با استفاده از روش‌های FST و Reynolds) را با یکدیگر دارند، با محدودیت‌هایی مواجه شده است [۲۰]. به

همین دلیل به‌نظر می‌رسد، روش کلاسیک آنالیز مؤلفه‌های اصلی احتیاج به اصلاحاتی دارد تا این نقص برطرف شود. این روش خالی از خطا و ایراد نبوده چرا که به‌دلیل این‌که ماحصل روش PCA یک ترکیب خطی از متغیرهای موجود در داده آزمایشی است و این روش از ماتریس کواریانس و یا ماتریس همبستگی استفاده می‌کند، این احتمال وجود دارد که دارای اریب محاسباتی باشد. مشکل دیگری هم‌چون تعداد فراوان متغیرها (p) نسبت به تعداد افراد (n) در آزمایش نیز وجود دارد (n<p). این مسئله به‌ویژه در پژوهش‌های زیست‌شناختی (میکروآرایه‌ها- Microarrays) بیش‌تر بروز می‌کند [۲۵].

در مجموع نتایج پژوهش حاضر نشان داد که روش شبکه عصبی (یادگیری ماشینی) نسبت به PCA، توانایی بیش‌تری در مواجهه با داده‌های زیستی (داده‌های نشانگری) دارد. روش شبکه عصبی توانست با حدود ۱۵ درصد از نشانگرها، کل نژادها را تفکیک نماید. که این نکته بیان‌کننده توان شبکه عصبی در انتخاب ویژگی (SNPها) می‌باشد. نتایج حاصل از روش یادگیری ماشینی می‌تواند نقش ارزشمندی در طراحی کیت‌های تشخیص نژادی اسب داشته باشد. پیشنهاد می‌شود که توانایی روش شبکه عصبی در مقابل روش‌های کلاسیک (همانند آماره‌های Fst) برای انتساب حیوانات (اسب‌های هیبرید) ناشناخته به جمعیت واقعی آن‌ها مورد ازریابی قرار بگیرد.

تشکر و قدر دانی

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

تعارض منافع

هیچ‌گونه تعارض منافع توسط نویسندگان وجود ندارد.

منابع مورد استفاده

1. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. (2020) Ensembl 2021. Nucleic Acids Research 49(D1): D884-D91.
2. 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.
3. Menhaj M (2009) Fundamentals of Neural Networks of Computational Intelligence. Tafresh Unit, Professor Hesabi Publication Center: Tehran University of Technology (Polytechnic of Tehran). (In Persian)
4. Wilkinson S, Wiener P, Archibald AL, Law A, Schnabel RD, McKay SD, et al. (2011) Evaluation of approaches for identifying population informative markers from high density SNP Chips. BMC Genetics 12(1): 45.
5. Bertolini F, Galimberti G, Schiavo G, Mastrangelo S, Di Gerlando R, Strillacci M, et al. (2018) Preselection statistics and Random Forest classification identify population informative single nucleotide polymorphisms in cosmopolitan and autochthonous cattle breeds. Animal 12(1): 12-9.
6. Tu JV (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology 49(11): 1225-31.
7. Petersen JL, Mickelson JR, Cothran EG, Andersson LS, Axelsson J, Bailey E, et al. (2013) Genetic Diversity in the Modern Horse Illustrated from Genome-Wide SNP Data. PLoS ONE 8(1): e54997.
8. Olden JD and Jackson DA (2002) Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. Ecological modelling; 154(1): 135-50.
9. Paetkau D, Calvert W, Stirling I and Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology 4(3): 347-54.
10. R. Core T (2017) R: A Language and Environment for Statistical Computing [Available from: https://www.R-project.org/.
11. Stefan Fritsch and Guenther F (2016) neuralnet: Training of Neural Networks [Available from: https://CRAN.R-project.org/package=neuralnet.

12. Beck M (2016) NeuralNetTools: Visualization and Analysis Tools for Neural Networks [Available from: https://CRAN.R-project.org/package=NeuralNetTools.
13. Hinrichs AL, Larkin EK and Suarez BK (2009) Population stratification and patterns of linkage disequilibrium. Genetic epidemiology 33(S1): S88-S92.
14. Reich D, Price AL and Patterson N (2008) Principal component analysis of genetic data. Nature Genetics 40: 491.
15. Lewis J, Abas Z, Dadousis C, Lykidis D, Paschou P and Drineas P (2011) Tracing Cattle Breeds with Principal Components Analysis Ancestry Informative SNPs. PLOS ONE 6(4): e18007.
16. Dimauro C, Cellesi M, Steri R, Gaspa G, Sorbolini S, Stella A, et al. (2013) Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes. Animal Genetics 44(4): 377-382.
17. Dimauro C, Nicoloso L, Cellesi M, Macciotta NPP, Ciani E, Moioli B, et al. (2015) Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep. Small Ruminant Research 128: 27-33.
18. Biffani S, Dimauro C, Macciotta N, Rossoni A, Stella A and Biscarini F (2015) Predicting haplotype carriers from SNP genotypes in Bos taurus through linear discriminant analysis. Genetics Selection Evolution: GSE 47(1): 4.
19. Sorbolini S, Gaspa G, Steri R, Dimauro C, Cellesi M, Stella A, et al. (2016) Use of canonical discriminant analysis to study signatures of selection in cattle. Genetics Selection Evolution: GSE 48(1): 58.
20. Moradi MH, Khaltabadi-Farahani AH, Khodaei-Motlagh M, Kazemi-Bonchenari M and McEwan J (2021) Genome-wide selection of discriminant SNP markers for breed assignment in indigenous sheep breeds. Annals of Animal Science 21(3): 807:831.
21. Kumar H, Panigrahi M, Saravanan KA, Parida S, Bhushan B, Gaur GK, et al. (2021) SNPs with intermediate minor allele frequencies facilitate accurate breed assignment of Indian Tharparkar cattle. Gene 20; 777: 145473.
22. Novembre J and Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nature genetics 40(5): 646-649.
23. Azizi Z, Rafat A, Shoja J, Moradi Shahrbabak H and Moradi Shahrbabak M (2016) Study Of Population Structure And Stratification Two Ecotypes Buffalo With Dense Single Nucleotide Polymorphism Markers Using Admixture, Mds, Pca And Gc Methods. Journal Of Agricultural Biotechnology 8(2): 53-67. (In Persian).
24. Azizi Z, Moradi Shahrbabak H and Moradi Shahrbabak M (2017) Comparison Of PCA And DAPC Methods For Analysis Of Iranian Buffalo Population Structure Using Snpchip90k Data. Iranian Journal Of Animal Science (Iranian Journal of Agricultural Sciences) 48(2): 153-161. (In Persian).
25. Ringnér M (2008) What is principal component analysis? Nature Biotechnology 26: 303-304.