Iranian Journal of Animal Science
Vol 48, No 2, Summer 2017 (153-161)
DOI: 10.22059/ijas.2017.213547.653461

علوم دامی ایران
دورهٔ ۴۸، شمارهٔ ۲، تابستان ۱۳۹۶ (ص ۱۶۱-۱۵۳)

مقایسهٔ روش‌های PCA و DAPC در تجزیه‌وتحلیل ساختار جمعیتی گاومیش‌های ایران با تراشه‌های اسنیپ 90k

زهرا عزیزی[۱]، حسین مرادی شهر بابک[۲*] و محمد مرادی شهر بابک[۳]
۱. دانشجوی سابق دکتری ژنتیک و اصلاح دام، گروه علوم دامی، دانشکدهٔ کشاورزی، دانشگاه تبریز
۲ و ۳. استادیار و استاد گروه علوم دامی، پردیس کشاورزی و منابع طبیعی دانشگاه تهران، کرج
(تاریخ دریافت: ۱۳۹۵/۵/۱۸ - تاریخ پذیرش: ۱۳۹۵/۱۲/۱٤)

چکیده

اطلاع از ساختار ژنتیکی جمعیت دام‌ها در راستای اجرای بهتر برنامه‌های اصلاح نژادی و حفظ ذخایر ژنتیکی آن‌ها بسیار ارزشمند است. داده‌های ژنگانی (ژنومی) فرصتی برای حل پیچیدگی تاریخچهٔ تکاملی جمعیت‌ها و بازسازی رویدادهای تاریخی نادر، را فراهم می‌آورند. در این پژوهش برای ارزیابی ساختار جمعیتی گاومیش‌های ایران روش‌های تجزیه‌وتحلیل مؤلفه‌های اصلی (PCA) و تجزیه‌وتحلیل جداسازی مؤلفه‌های اصلی (DAPC) اجرا شد. از شمار ٤٠٤ گاومیش از سه نژاد شمالی، آذری و خوزستانی خون‌گیری شد و تعیین ژنوتیپ با تراشه‌های اسنیپ 90k توسط شرکت پادانو در کشور ایتالیا انجام شد. نتایج به‌دست‌آمده از تجزیهٔ مؤلفه‌های اصلی و تجزیهٔ جداسازی مؤلفه‌های اصلی، جداسازی سه نژاد را به‌خوبی نشان داد و هر دو تصویر آشکاری از ساختار ژنتیکی جمعیت‌های مورد بررسی را نشان دادند. در روش DAPC، برای ارزیابی شمار بهینهٔ خوشه با معیار ارزیابی BIC، K=3 بهترین نتیجه را نشان داد. نتایج اعتبارسنجی متقابل برای نگه‌داشتن شمار مؤلفهٔ اصلی بهینه برای تجزیه‌وتحلیل تشخیصی، ۵۰ مؤلفهٔ اول MSE کمتری نسبت به مؤلفه‌های دیگر داشت. در این بررسی روش DAPC احتمال عضویت افراد جمعیت‌ها را با درستی ۱۰۰ درصد پیش‌بینی کرد ولی روش PCA قادر به ارزیابی گروه‌ها نبوده و برای به دست آوردن تصویر روشن از واریانس بین جمعیت‌ها DAPC مناسب‌تر از PCA عمل می‌کند. روش DAPC در بررسی ساختار جمعیتی نسبت به روش PCA به دلیل افزایش واریانس بین گروه‌ها و کاهش واریانس درون گروه‌ها و همچنین ارائهٔ تصویر آشکاری از ساختار جمعیتی کارآمد بود و می‌تواند در کنترل کیفیت و تصحیح لایه‌بندی جمعیتی در بررسی‌های ارتباطی جایگزینی برای PCA باشد.

واژه‌های کلیدی: تراشه‌های اسنیپ، ساختار جمعیتی، گاومیش، PCA، DAPC.

Comparison of PCA and DAPC methods for analysis of Iranian Buffalo population structure using SNPchip90k data

Zahra Azizi[1*], Hossein Moradi Shahrbabak[2] and Mohammad Moradi Shahrbabak[3]
1. Former Ph.D. Student, Department of Animal Sciences, Faculty of Agricultural Sciences, University of Tabriz, Iran
2, 3. Assistant Professor and Professor, Department of Animal Sciences, University College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran
(Received: Aug. 8, 2016 - Accepted: Mar. 4, 2017)

ABSTRACT

Understanding of population genetic structure is valuable for better implementation of breeding programs and most importantly, preservation of genetic resources. Genomic data provide an opportunity to consider complex evolutionary history of populations and reconstruct rare historical events. In this research, the structure of Iranian buffalo populations was studied by using principal component analysis and discriminant analysis principal component methods. For this purpose, the number of 404 buffalos from three breeds including North, Azari and Khozestani were sampled and genotyped by SNPChip 90k from Padano Company in Italy. The results of principal component analysis and discriminant analysis principal component showed a clear picture of the genetic structure of the studied populations. Assessing the optimal number of clusters with criteria BIC, K = 3 by the DAPC method showed the best results. The result of cross-validation for retaining principal components was optimized to 50 first components that showed the lowest MSE. In this study, DAPC predicted assignment of individuals to clusters and membership probabilities with 100% accuracy. PCA method was not able to provide a group assessment and DAPC method outperformed than PCA in achieving a clear variance difference between populations. DAPC method can be applied in quality control and stratification population correction of GWAS as an alternative to the PCA because of summarizing the genetic differentiation between groups and overlooking within-group variation and providing better population structure.

Keywords: Buffalo, DAPC, PCA, SNPChip, structure population.

* Corresponding author E-mail: hmoradis@ut.ac.ir

مقدمه

گاومیش‌های ایران به دلیل سازگاری با محیط، مقاومت در برابر بیماری‌ها، هزینه‌های نگهداری پایین و استفاده از پسماندهای کشاورزی کم‌ارزش، یکی از ذخایر ژنتیکی باارزش به شمار می‌آیند. جمعیت گاومیش‌های ایران شامل سه نژاد خوزستانی، آذری و مازندرانی است (Agricultural ministry statistics, 2012). در زمینهٔ ژنتیک و اصلاح نژاد دام، اطلاع از ساختار ژنتیکی جمعیت در راستای اجرای بهتر برنامه‌های اصلاح نژادی و حفظ ذخایر ژنتیکی بسیار ارزشمند است. از سویی دیگر ساختارهای زیر جمعیتی درون جمعیت‌های مورد بررسی باعث ایجاد اریب در بررسی‌های پویش ژنگانی (ژنومی) می‌شود (Thomas & Witte, 2002; Wacholder *et al*., 2002).

فن‌آوری توالی‌یابی نسل جدید، باعث تولید مقادیر زیادی داده برای جمعیت‌های دامی در حوزهٔ ژنتیک شده است و فرصتی برای درک پیچیدگی تاریخچهٔ تکاملی جمعیت‌ها و بازسازی رویدادهای تاریخی نادر، فراهم می‌آورند (Ekblom & Galindo, 2011). انتخاب طبیعی و مصنوعی صورت گرفته در چرخهٔ گذشت زمان، منجر به ایجاد نژادهای مختلفی شده است که از لحاظ فنوتیپی تفاوت گسترده‌ای دارند (Epps *et al*., 2013). استنتاج ساختار جمعیت از نشانگرهای ژنتیکی، در شرایط گوناگون مانند بررسی‌های ارتباطی و تکاملی، دسته‌بندی زیرگونه‌ها و تعیین بازدارنده‌های ژنتیکی سودمند است (Liu & Zhao, 2006). تجزیه‌وتحلیل ساختار جمعیتی با استفاده از حجم گستردهٔ اطلاعات نشانگری نیاز به ابزاری دارد که بتوان با استفاده از این روش‌ها به بررسی ارتباط بین جمعیت‌ها و قرار دادن افراد در جمعیت‌ها که از آن نشئت گرفته‌اند پرداخت. شناخت ساختار ژنتیکی و روابط موجود میان نژادها نه تنها ارزش حفاظتی آن‌ها را روشن‌تر می‌سازد بلکه می‌تواند برای اولویت‌بندی برنامه‌های حفاظت ژنتیکی نیز به کار گرفته شود. روش‌های پرشماری برای تعیین ساختار ژنتیکی و لایه‌بندی جمعیت وجود دارد. یکی از شیوه‌های آماری برای آزمون ارتباط بین جمعیت‌ها و اختصاص افراد به آن‌ها با استفاده از ماتریس فاصله، استفاده از تجزیه‌وتحلیل مؤلفه‌های اصلی است (Li & Yu, 2008). تجزیه‌وتحلیل مؤلفه‌های اصلی[1] (Patterson *et al*., 2006; Price *et al*., 2006) قادر به تعیین ساختار جمعیت است. تجزیه‌وتحلیل مؤلفه‌های اصلی توسط پیرسون[2] در سال ۱۹۰۱ پیشنهاد شد (Peason, 1901) پایهٔ کار این روش به این صورت است که در صفحهٔ مختصات محورهای اصلی X و Y را طوری تغییر دهیم که مسیری در فضا پیدا شود تا مؤلفه‌های اصلی مربوط به داده‌ها در طول آن مسیر قرار گیرند. هر محوری که بزرگ‌تر باشد نشان‌دهندهٔ آن است که واریانس بیشتری در میان داده‌ها در این جهت است و به همین دلیل آن‌را نخستین مؤلفهٔ اصلی گویند. هدف از تجزیهٔ مؤلفه‌های اصلی آن است که واریانس موجود در داده‌های چندمتغیره را به مؤلفه‌هایی تجزیه کند که نخستین مؤلفه تا آنجا که ممکن است علت بیشترین واریانس موجود در داده‌ها باشد. دومین مؤلفه علت بیشترین واریانس ممکن پس از مؤلفهٔ اول و الی‌آخر باشد. افزون بر این، در این روش هر مؤلفه مستقل از مؤلفه‌های دیگر است، یعنی بین هر مؤلفه و مؤلفه‌های دیگر همبستگی وجود ندارد (Jolliffe, 2002). به عبارتی هنگامی‌که شمار متغیرها زیاد است و بین این متغیرها همبستگی وجود دارد، PCA کارایی خود را نشان می‌دهد. ابزار اصلی در PCA برای تجزیه‌وتحلیل، ماتریس واریانس کواریانس و به‌عنوان روش کاهش خطی ابعاد برای تعیین ساختارهای جمعیتی استفاده می‌شود. تجزیه‌وتحلیل مؤلفهٔ اصلی ابزار استانداری در ژنتیک جمعیت است که در بررسی جمعیت‌های اروپائی و هندی استفاده شده است (Lao *et al*., 2008). این روش، روشی نافراسنجه‌ای است که از گذشته به‌عنوان یک روش کاهش خطی ابعاد برای نمایان‌سازی ساختارهای جمعیتی استفاده می‌شود (Laloë *et al*., 2007). همچنین در کنترل کیفیت در بررسی‌های ژنتیکی استفاده می‌شود. در عمل PCA به‌عنوان جایگزینی برای الگوریتم‌های خوشه‌بندی بیزی پیشنهاد شده است (Lee *et al*., 2009; Liu & Zhao, 2006; Price *et al*., 2006).

1. Principle Component Analysis
2. Pearson

تجزیهٔ تشخیصی مؤلفه‌های اصلی[1]، یک روش چند متغیره است که برای شناسایی و توصیف خوشه‌های افرادی که به‌طور ژنتیکی با هم ارتباط دارند، طراحی شده است (Jombart *et al.*, 2010). این روش هنگامی‌که گروه‌های پیشین وجود ندارند، k-means پی‌درپی و انتخاب مدل را برای استنباط ژنتیکی خوشه‌ها استفاده می‌کند. K-means شمار گروه‌هایی (k) را که تنوع بین گروه‌ها را بیشینه می‌کند پیدا می‌کند (Jombart *et al.*, 2010)DAPC . جداسازی افراد را به گروه‌های از پیش تعیین‌شده به‌دست می‌دهد و همچنین این روش احتمالاتی را برای انتساب افراد به هر گروه فراهم می‌آورد. این روش را می‌توان به داده‌های با حجم زیاد، هزاران نشانگر از هزاران فرد اعمال کرد تجزیه‌وتحلیل تشخیصی مؤلفه‌های اصلی ارزیابی چشمی از تفاوت بین جمعیت‌ها را فراهم می‌کند (Jombart *et al.*, 2010). یکی از سودمندی‌های DAPC، سازگارپذیری بالای آن است درواقع DAPC بر مدل ژنتیک جمعیت خاصی وابسته نیست و درنتیجه فرض تعادل هاردی واینبرگ و تعادل پیوستگی (لینکاژی) ندارد. به‌این‌ترتیب برای انواع موجودات صرف‌نظر از پلوئیدی بودن و نرخ نوترکیبی آن‌ها، کاربرد دارد (Jombart *et al.*, 2010). همچنین این روش محدود به دادهٔ ژنگانی (ژنومیکی) نیست و برای داده‌های کمّی دیگر مثل داده‌های ریخت‌شناسی نیز قابل‌اجرا است (Jombart & Collins, 2015). هدف این تحقیق بررسی ساختار جمعیتی گاومیش‌های ایران با استفاده از داده‌های SNPChip 90K و جداسازی افراد مناطق مختلف با مسئلهٔ یادگیری با نظارت به روش DAPC بود.

مواد و روش‌ها

نمونه‌گیری و تعیین ژنوتیپ دام‌ها

نمونه‌ها از گله‌های مردمی و گله‌های تحت نظام ثبت شجره و رکوردهای مرکز بهبود تولیدهای دامی و اصلاح نژاد گردآوری شد. در انتخاب حیوانات برای نمونه‌گیری، عامل‌های موردتوجه شامل حیوانات غیر خویشاوند و حیوانات از پراکنش‌های جغرافیایی متفاوت و با بیشترین تنوع در صفاتی مانند تولید شیر، یا ویژگی‌های ظاهری بودند. نمونه‌برداری از استان‌های آذربایجان غربی (از سه شهرستان خوی، ارومیه و مهاباد)، آذربایجان‌شرقی (از پنج شهرستان شامل تبریز، سراب، بستان‌آباد، اسکو و ایلخچی)، اردبیل (از دو شهر نمین و مشکین‌شهر)، گیلان (از هفت شهرستان ماسال، تالش، صومعه‌سرا، بندر انزلی، طاهر گوراب، رضوانشهر و اسالم) و خوزستان (شهرستان‌های دزفول، اهواز، شادگان، شوش، شوشتر و سوسنگرد) انجام گرفته بود. در کل ۴۰۸ نمونه از این استان‌ها گردآوری شده بود (جدول ۱). استخراج DNA ژنگانی از ریشهٔ مو (Alberts *et al.*, 2010) و خون با روش بهینهٔ نمکی (Grimberg *et al.*, 1989) بسته به سهولت نمونه‌گیری، انجام شد. پس از استخراج DNA در آزمایشگاه علوم دامی دانشگاه تهران و بررسی کیفیت نمونه‌های مورد نظر، نمونه‌ها برای انجام مراحل بعدی توالی‌یابی به آزمایشگاه ژنومیک مرکز تحقیقات پادانو[2] کشور ایتالیا منتقل شدند. سپس نمونه‌ها با استفاده از تراشه‌های ArrayAxiom® Buffalo Genotyping 90K مربوط به شرکت افی‌متریکس کشور ایتالیا تعیین ژنوتیپ شدند.

جدول ۱. مناطق جغرافیایی و شمار حیوانات نمونه‌برداری‌شده در هر منطقه

Table 1. The geographic regions and the numbers of animals sampled in each region

Breed	Sampled provinces	Number of animal
Azari	East Azarbayjan	68
	West Azarbayjan	68
	Ardabil	54
	Guilan	71
Mazandarani	Mazandaran	27
Khozestani	Khozestan	114
	Kermashah	5

مراحل پالایش داده‌های به‌دست‌آمده از تعیین ژنوتیپ برای انجام تجزیه‌وتحلیل‌های نهایی

کنترل کیفیت اولیه روی داده‌ها توسط شرکت پادانو انجام شده بود که در نتیجه این کنترل کیفیت، چهار نمونه در جریان تعیین ژنوتیپ (دو نمونه از استان اردبیل و دو نمونه از استان گیلان) با بیش از ۵ درصد

1. Discriminant Analysis Principal Component

2. Parco Tecnologico Padano

ژنوتیپ گم‌شده حذف شده بودند (جدول۲). درمجموع شمار ۸۸۵۵ اسنیپ به دلیل MAF کمتر از ۱ درصد، ۳۳۶ اسنیپ به دلیل انحراف از تعادل هاردی-وینبرگ در سطح ۵ درصد و ۱۹ اسنیپ به خاطر موقعیت ناشناخته حذف شده بودند. در نهایت ۴۰۴ حیوان با ۶۴۷۵۰ اسنیپ، مراحل کنترل کیفیت را با MAF>0.01 و call rate>0.99 گذرانده و همهٔ اسنیپ‌های باقی‌مانده در سطح ۵ درصد در تعادل هاردی وینبرگ بودند. در این بررسی کنترل کیفیت روی داده‌های اولیهٔ ارسال‌شده توسط شرکت با استفاده از نرم‌افزار Plink، اعمال شد. بدین ترتیب که در آغاز حیوانات دارای بیش از ۵ درصد ژنوتیپ ازدست‌رفته از تجزیه‌های بعدی کنار گذاشته شد چون نمونه‌های با کیفیت پایین با احتمال بیشتری با داده‌های گم‌شده همراه هستند و منجر به افزایش خطای ژنوتیپ می‌شود (Barendse *et al*., 2009). دو عامل دست‌کم فراوانی آللی (MAF) و درصد نمونه‌هایی که برای آن نشانگر ژنوتایپ شده‌اند (Call rate) برای هر اسنیپ محاسبه شدند و اسنیپ‌هایی که دارای Call rate و MAF به ترتیب کمتر از ۹۵ و ۱ درصد بودند، حذف شدند. برای اسنیپ‌های باقی‌مانده در صورت تعادل نداشتن هاردی- واینبرگ به‌عنوان معیاری از خطای تعیین ژنوتیپ کنار گذاشته شدند (Teo *et al*., 2007). تعادل هاردی-وینبرگ در همهٔ جایگاه‌ها بررسی شد و همهٔ جایگاه‌های با P-value کمتر از $۱۰^{-۷}$، کنار گذاشته شدند. برای تعیین سطح معنی‌داری مطلوب در این آزمون از تصحیح بنفرونی ($\beta=\alpha/n$) استفاده شد (Abdi, 2007) که α مربوط به سطح معنی‌داری و n شمار جایگاه‌ها است. در مرحلهٔ نهایی، داده‌ها با نرم‌افزار plink از فرمت اللی به فرمت داده‌های ژنوتیپی با سه دسته (۰،۱ و۲) برای هر SNP، تبدیل شدند.

تجزیه‌وتحلیل‌های آماری

تجزیه‌وتحلیل PCA با تابع prcomp با بستهٔ آماری strata و تجزیه‌وتحلیل DAPC با بستهٔ آماری adegenet مربوط به نرم‌افزار R اجرا شد (https://cran.r-project.org/web/packages/adegenet/index.html).

برای شناسایی شمار بهینهٔ خوشه‌ها، k-means به‌صورت متوالی با افزایش مقادیر k، اجرا می‌شود و جواب‌های خوشه‌بندی مختلف با استفاده از معیار اطلاعات بیزی (BIC)[1] مقایسه شد. در حالت مطلوب، خوشهٔ بهینه بایستی BIC پایین‌تری داشته باشد در عمل بهترین BIC اغلب با خمیدگی در منحنی مقادیر BIC به‌عنوان تابعی از K نشان داده شد. در این روش، خوشه‌ها با تابع find.cluster به دست می‌آیند. این تابع در آغاز داده‌ها را با PCA تبدیل می‌کند، آنگاه الگوریتم k-means با افزایش مقادیر k اجرا می‌شود. برای انتخاب شمار بهینه PCهای نگه داشته شده برای تجزیه‌وتحلیل تشخیصی از اعتبارسنجی متقابل با تابع xvalDapc با ۳۰ تکرار استفاده شد.

آماده‌سازی داده‌ها برای DAPC با نظارت

برای گسترش مدل ماشین یادگیری با نظارت که داده‌ها به دو دسته آموزش و آزمایش تقسیم‌بندی می‌شوند، از روش اعتبارسنجی متقابل[2] ده باره (k fold=10) استفاده شد. در این بررسی برچسب‌گذاری یا انتخاب کلاس‌ها[3] بنا بر اطلاعات در دسترس بود که سه کلاس در تجزیه‌وتحلیل‌ها در نظر گرفته شد.

ارزیابی خطای طبقه‌بندی روش با نظارت DAPC

معیار درستی مشهورترین و عمومی‌ترین معیار محاسبهٔ کارایی الگوریتم‌های دسته‌بندی است که نشان می‌دهد دسته‌بندی طراحی‌شده چند درصد از کل مجموعه رکوردهای آزمایشی را به‌درستی دسته‌بندی کرده است و یکی از روش‌های متداول ارزیابی طبقه‌بندی، استفاده از مجموعهٔ نمونه‌های آزمون و تشکیل ماتریس خطا است که با بستهٔ caret در نرم‌افزار R اجرا شد (https://cran.r-project.org/web/packages/caret/index.html).

نتایج و بحث

کنترل کیفیت و پالایش داده‌های ژنگانی برای کل جمعیت‌ها

کنترل کیفیت روی ۶۴۷۵۰ اسنیپ اجرا شد که در

1. Bayesian Information Criterion
2. Cross Validation
3. Labeling

آغاز نوزده اسنیپ به دلیل موقعیت ناشناخته حذف شدند و در مراحل مختلف کنترل کیفیت روی اسنیپ‌های باقی‌مانده ۲۱ اسنیپ به دلیل انحراف از تعادل هاردی- واینبرگ از تجزیه‌وتحلیل‌های نهایی حذف شدند و در مجموع ۴۰۴ حیوان از هفت استان مختلف از سه اکوتیپ با ۶۴۷۱۰ اسنیپ وارد مرحلهٔ تجزیه‌وتحلیل نهایی شدند.

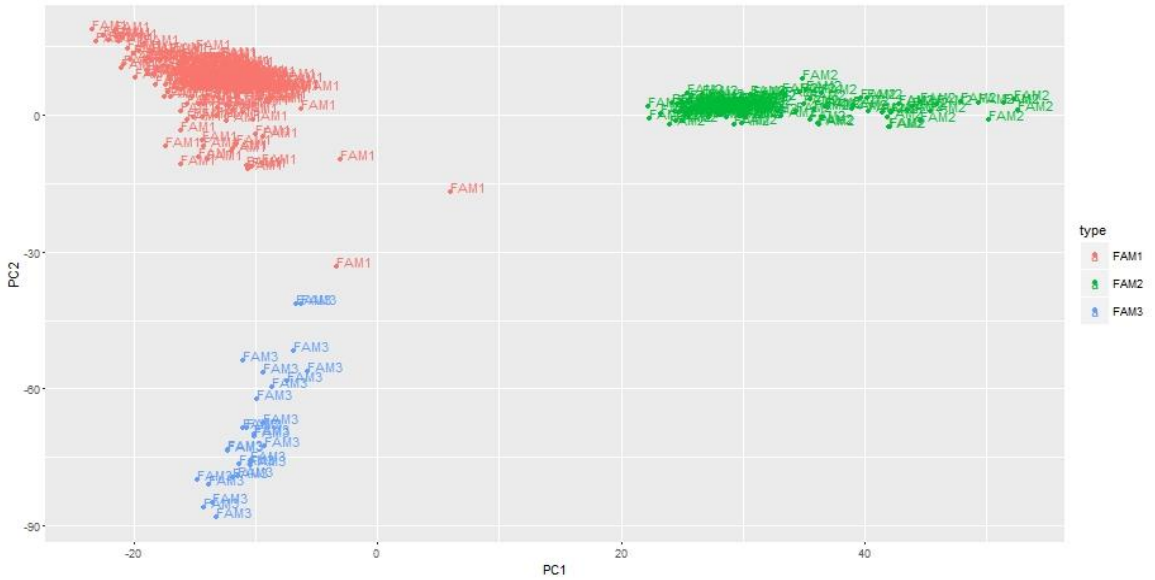
تجزیه‌وتحلیل مؤلفه‌های اصلی

برای ارزیابی تفاوت ژنتیکی میان جمعیت‌ها (سه نژاد) شکل PCA (شکل ۱) ترسیم شد که نشان‌دهندهٔ تمایز سه نژاد از همدیگر است و نژاد مازندرانی از نژاد آذری جدا است. نتایج تجزیه‌وتحلیل PCA بر پایهٔ مؤلفهٔ اول و دوم نشان داد که برای سه نژاد، دو مؤلفهٔ اول ۳/۱۱ درصد واریانس را توجیه می‌کنند و برای توجیه ۹۰ درصد واریانس بیش از ۳۱۸ مؤلفهٔ اول نیاز است.

تجزیهٔ تشخیصی مؤلفه‌های اصلی

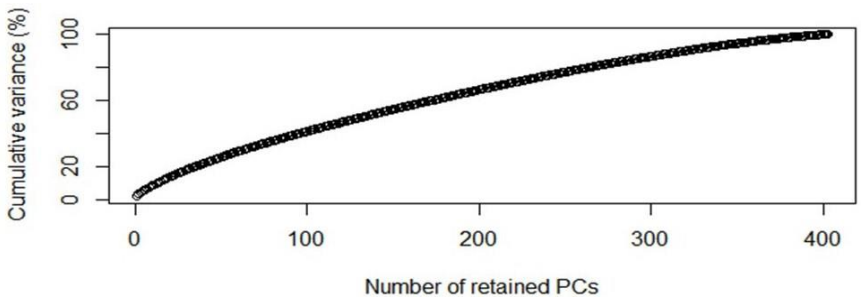
نتایج به‌دست‌آمده از تجزیه‌وتحلیل تشخیصی مؤلفه‌های اصلی نشان دادند که دو مؤلفهٔ اول و دو تابع تشخیصی ۳/۱۰۷ درصد واریانس را توجیه کردند. همچنین برای توجیه حدود ۹۹ درصد واریانس نیاز به ۳۸۸ مؤلفهٔ اول با ۳ تابع تشخیصی، است (شکل ۲).

تصویر مربوط به خوشه‌بندی با این روش در شکل ۳ ارائه شده است. بنابر، این شکل مشخص است که سه نژاد از همدیگر متمایز بوده و تفاوت دارند.



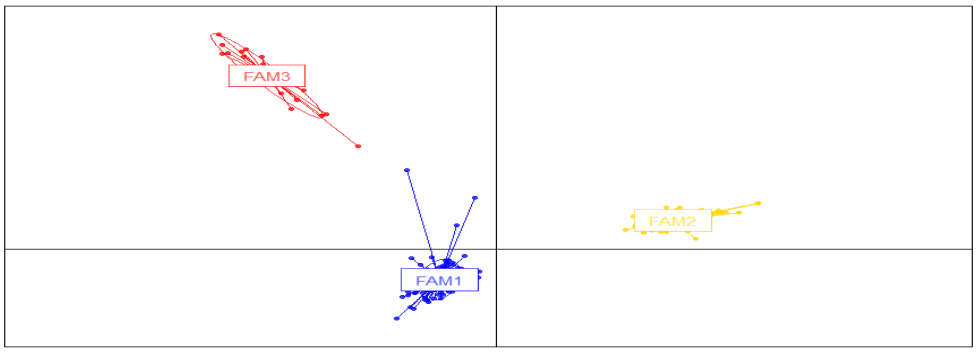
شکل۱. تجزیه‌وتحلیل PCA مربوط به سه نژاد از گاومیش‌های ایران (FAM1 تا FAM3 به ترتیب مربوط به نژادهای آذری، خوزستانی و مازندرانی هستند).

Figure 1. PCA Analysis of three Iranians Buffalos breeds ((FAM 1 to FAM 3 are related to Azari, Khuzestan and Mazandaran, respectively).



شکل ۲. واریانس تجمعی توصیف‌شده با شمار مؤلفه‌های اصلی متفاوت

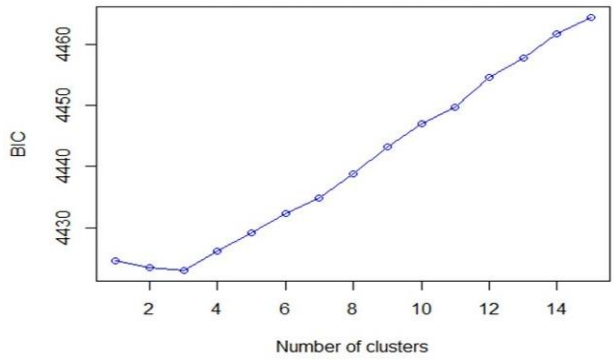
Figure 2. Cumulative variance explained by different number of principal components



شکل۳. تجزیه‌وتحلیل مؤلفه‌های اصلی تشخیصی مربوط به گاومیش‌های سه نژاد. خوشه‌ها با رنگ‌های مختلف و با FAM مشخص شده‌اند و افراد با نقطه‌ها نشان داده شده‌اند (رنگ‌های آبی، زرد و قرمز به ترتیب مربوط به نژادهای آذری، خوزستانی و شمالی هستند.)

Figure 3. DAPC Analysis of three Iranian breeds of buffalo. Clusters are marked with different colors and with FAM and individuals with dots are shown (Blue, yellow and red colors are related to Azari, Khuzestan and Mazandaran, respectively).

شکل بالا نشان می‌دهد که روش DAPC واریانس بین گروهی را بیشینه و واریانس درون گروهی را کمینه می‌کند. همچنین با توجه به تعیین خوشه‌ها و مشخص کردن مرکز هر خوشه، تصویر آشکاری از گروه‌های افراد نسبت به روش PCA نمایان می‌سازد. نتایج حاصل از اعتبارسنجی متقابل برای به‌دست آوردن شمار مؤلفه‌های اصلی که برای تجزیه‌وتحلیل نگه داشته شدند، با سی تکرار انجام شد که پنجاه مؤلفهٔ اول کمترین MSE را نشان دادند. این پنجاه مؤلفهٔ اول با دو تابع تشخیصی، ۲۵/۸۳ درصد واریانس را توجیه کردند. برای تعیین شمار بهینهٔ خوشه‌بندی کمترین میزان BIC معیار ارزیابی است که در این بررسی K=3 کمترین میزان را نشان داد (شکل ۴).



شکل۴. مقادیر BIC به ازای شمار خوشه‌ها

Figure 4. BIC values for the number of clusters

بنابر شکل ۴، کمترین میزان BIC از k=1 تا k=3 است که پس‌ازاین k روند نمودار به‌صورت افزایشی است و به‌طور آشکار نشان می‌دهد، k=3 کمترین میزان BIC را دارد و سه خوشه‌ای (k) بودن این جمعیت‌ها را تأیید می‌کند. افزون براین DAPC، احتمال عضویت هر فرد را برای گروه‌های مختلف، بر پایهٔ توابع تشخیصی بازگشتی فراهم می‌آورد که ضریب‌های به‌دست‌آمده از این روش، می‌توانند نزدیکی افراد به خوشه‌های مختلف را تفسیر کنند. به‌عبارت‌دیگر احتمال عضویت افراد، خوشه‌های ژنتیکی آشکاری را فراهم می‌آورد.

پیش‌بینی داده‌های جدید برای سه نژاد

روش مؤلفه‌های اصلی تشخیصی به‌عنوان روش با نظارت نیز عمل می‌کند (Jombart *et al.*, 2010). برای مثال در DAPC ممکن است گروه‌های بیشتر افراد

شناخته شده باشند ولی برخی افراد گروه نامشخص یا ناشناخته داشته باشند. در این صورت افرادی که گروه مشخص دارند برای آموزش مدل استفاده می‌شوند و داده‌های جدید که گروه‌های نامشخص دارند بنا بر مدل آموزشی، پیش‌بینی می‌شوند (Jombart *et al.*, 2010). در این بررسی چگونگی برچسب‌گذاری بنابر نژاد بوده که نتایج به‌دست‌آمده از تجزیه‌وتحلیل سه کلاس با ده بار اعتبارسنجی متقابل درستی ۱۰۰ درصد را نشان داد.

برتری اصلی PCA همانند DAPC، توانایی آن برای شناسایی ساختارهای جمعیتی با حجم گستردهٔ دادهٔ ژنگانی با زمان محاسباتی کمتر، و نبود فرضیه‌هایی دربارهٔ مدل ژنتیک جمعیتی است. بااین‌حال، PCA بدون برخی از ویژگی‌های ضروری بررسی ساختار جمعیت‌های زیستی (بیولوژیکی) است. PCA یک ارزیابی گروهی را فراهم نمی‌سازد و نیازمند پیش‌تعریف مناسبی از خوشه‌ها برای بررسی ساختار جمعیتی است و حتی برای به دست آوردن تصویر آشکار از تنوع بین جمعیت‌ها مناسب نیست (Jombart *et al.*, 2010). تجزیه‌وتحلیل مؤلفه‌های اصلی تنوع کلی میان افراد که شامل هر دوی تنوع بین گروه‌ها و تنوع درون گروه‌ها می‌شود را خلاصه می‌کند و تصویر آشکاری از تفاوت بین گروه‌ها را نشان می‌دهد (Jombart *et al.*, 2010)که با نتایج به‌دست‌آمده از این بررسی همخوانی دارد. برای ارزیابی ارتباط بین خوشه‌های مختلف، باید یک روش مناسب بر تنوع بین گروه‌ها تمرکز کند و تنوع درون گروهی را نادیده بگیرد که تجزیهٔ تشخیصی مؤلفه‌های اصلی این کار را انجام می‌دهد.

در بررسی ساختار جمعیتی گاوهای تونسی با روش‌های مختلف از جمله روش DAPC و PCA، دو مؤلفهٔ اول ۲۰ درصد واریانس کلی را توصیف کرد و روش DAPC توانایی تشخیص بهتری بین نژادها از منشأ یکسان را داشت. در تعیین شمار بهینهٔ مؤلفه‌های اصلی نگه داشته شده نیز، شصت مؤلفهٔ اول که ۴۰ درصد واریانس را توصیف می‌کرد برای تجزیه‌وتحلیل تشخیصی حفظ شدند (Jemaa *et al.*, 2015). برتری روش DAPC نسبت به PCA در به تصویر کشیدن تفاوت بین گروهی است که تفاوت گروهی را به‌طور مناسب‌تر نشان داد. مشخصهٔ اصلی PCA توانایی آن برای شناسایی ساختارهای ژنتیکی در مجموعهٔ داده‌های بزرگ در زمان محاسباتی ناچیز و بدون هیچ فرضی دربارهٔ زمینهٔ مدل ژنتیکی جمعیت است (Jombart *et al.*, 2010). در پژوهشی برای بررسی ساختار جمعیتی با روش‌های مبتنی بر مدل و روش اکتشافی DAPC، این روش‌ها تداوم بالایی در برآورد ساختار جمعیت و استنباط احتمالات عضویت افراد به هر گروه نشان دادند (Pometti *et al.*, 2014). روش PCA به‌عنوان جایگزینی برای روش‌های خوشه‌بندی بیزی پیشنهاد شده است (Lee *et al.*, 2009; Liu & Zhao, 2006). بااین‌حال PCA بدون برخی از ویژگی‌های ضروری برای بررسی ساختار جمعیت زیستی از جمله ناتوانی در ارزیابی گروهی و نیازمند تعریف پیش‌فرض خوشه‌ها برای بررسی ساختار جمعیت است. همچنین در تصحیح لایه‌بندی جمعیتی، روش DAPC به علت اینکه واریانس بین گروهی را افزایش و واریانس درون گروهی را کاهش می‌دهد بهتر از روش PCA است (Jombart *et al.*, 2010). در استنباط و تفسیر ساختار ژنتیکی با استفاده از روش‌های پرشمار بررسی ساختار جمعیتی از جمله PCA و DAPC ، روش DAPC با توجه به برتری‌هایی همچون تفسیر بهتر زیر جمعیت‌ها و انتساب خوشه‌ها بهتر از روش PCA عمل کرد (Sethuraman, 2013). نرخ انتساب درست یا درستی روش DAPC در انتساب افراد به گروه‌ها از ۸۰ درصد تا ۹۷ درصد بسته به شمار تکرار متفاوت بود (Jombart *et al.*, 2010). در این بررسی روش DAPC در تعیین شمار بهینهٔ K بهتر از روش PCA عمل کرد و تصویر بهتری از ارتباط بین افراد نسبت به PCA ارائه داد. همچنین در انتساب افراد به گروه‌های خودشان درستی بسیار خوبی ارائه داد. با توجه به این برتری‌ها و توانایی انتساب افراد، این روش برای بررسی ساختار جمعیتی و تعیین شمار خوشه و برای کنترل کیفیت و تصحیح لایه‌بندی جمعیتی در بررسی‌های GWAS به‌جای حذف افراد بیرون از گروه‌های نژادی غیرقابل تشخیص، بهترین گزینه است.

نتیجه‌گیری کلی

در این بررسی روش DAPC در تعیین شمار بهینهٔ k بهتر عمل کرد و با توجه به در نظر گرفتن شمار زیادی PC به‌طور همزمان برای به تصویر کشیدن ساختار جامعهٔ بهتر از PCA عمل می‌کند و همچنین این روش به‌عنوان یک روش با نظارت نیز در انتساب دام‌ها کارآمد است و احتمال عضویت هر فرد را به جمعیت مورد بررسی فراهم می‌آورد که می‌توان از این اطلاعات برای کنترل اینکه آیا حیوانات به‌طور صحیحی به جمعیت‌های از پیش تعریف‌شده انتساب دارند، استفاده کرد.

REFERENCES

1. Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics*, pp. 103-107.
2. Agricultural ministry statistics. (2012).Tehran, Ministry of Agriculture, Deputy Director of Planning and Economics, Technology Center of Information and Communication. Vol 2. (in Farsi)
3. Alberts, C.C., Ribeiro-Paes, J. T., Aranda-Selverio, G., Cursino-Santos, J. R., Moreno-Cotulio, V. R., Oliveira, A. L., Porchia, B. F., Santos, W. F. & Souza, E. B. (2010). DNA extraction from hair shafts of wild Brazilian felids and canids. *Genet Mol Res*, 9(4), 2429-2435.
4. Barendse, W., Harrison, B. E., Bunch, R. J., Thomas, M. B. & Turner, L. B. (2009). Genome wide signatures of positive selection: the comparison of independent samples and the identification of regions associated to traits. *BMC Genomics*, 10(1), 178.
5. Ekblom, R. & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1), 1-15.
6. Epps, C. W., Castillo, J. A., Schmidt-Küntzel, A., du Preez, P., Stuart-Hill, G., Jago, M. & Naidoo, R. (2013). Contrasting historical and recent gene flow among African buffalo herds in the Caprivi Strip of Namibia. *Journal of Heredity*, ess142.
7. Grimberg, J., Nawoschik, S., Belluscio, L., McKee, R., Turck, A. & Eisenberg, A. (1989). A simple and efficient non-organic procedure for the isolation of genomic DNA from blood. *Nucleic Acids Research*, 17(20), 8390-8390.
8. Jemaa, S. B., Boussaha, M., Mehdi, M. B., Lee, J. H. & Lee, S. H. (2015). Genome-wide insights into population structure and genetic history of Tunisian local cattle using the illumina bovinesnp50 beadchip. *BMC Genomics*, 16(1), 1.
9. Jolliffe, I. (2002). Principal component analysis. *Wiley Online Library*.
10. Jombart, T. & Collins, C. (2015). A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0. 0. *London: Imperial College London, MRC Centre for Outbreak Analysis and Modelling*.
11. Jombart, T., Devillard, S. & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, 11(1), 94.
12. Laloë, D., Jombart, T. Dufour, A.-B. & Moazami-Goudarzi, K. (2007). Consensus genetic structuring and typological value of markers using multiple co-inertia analysis. *Genetics Selection Evolution*, 39(5), 1-23.
13. Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L. A. & Comas, D. (2008). Correlation between genetic and geographic structure in Europe. *Current Biology*, 18(16), 1241-1248.
14. Lee, C., Abdool, A., & Huang, C.-H. (2009). PCA-based population structure inference with generic clustering algorithms. *BMC bioinformatics*, 10(Suppl 1), S73.
15. Li, Q. & Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genetic epidemiology*, 32(3), 215-226.
16. Liu, N. & Zhao, H. (2006). A non-parametric approach to population structure inference using multilocus genotypes. *Human genomics*, 2(6), 1.
17. Patterson, N., Price, A. L. & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, 2(12), e190.
18. Peason, K. (1901). On lines and planes of closest fit to systems of point in space. *Philosophical Magazine*, 2, 559-572.
19. Pometti, C. L., Bessega, C. F., Saidman, B. O. & Vilardi, J. C. (2014). Analysis of genetic population structure in Acacia caven (Leguminosae, Mimosoideae), comparing one exploratory and two Bayesian-model-based methods. *Genetics and Molecular Biology*, 37(1), 64-72.
20. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904-909.

21. Sethuraman, A. (2013). *On inferring and interpreting genetic population structure-applications to conservation, and the estimation of pairwise genetic relatedness*. Ph.D. Thesis. Iowa State University, Paper 13332. U.S.
22. Teo, Y. Y., Fry, A. E., Clark, T. G., Tai, E. & Seielstad, M. (2007). On the usage of HWE for identifying genotyping errors. *Annals of Human genetics*, 71(5), 701-703.
23. Thomas, D. C. & Witte, J. S. (2002). Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiology Biomarkers & Prevention*, 11(6), 505-512.
24. Wacholder, S., Rothman, N. & Caporaso, N. (2002). Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiology Biomarkers & Prevention*, 11(6), 513-520.