A Review on the Analysis of Population Genetic Structure using Dominant Molecular Markers and Introducing the New Program STRUCTUREasy

Majid Sharifi-Tehrani*

Assistant Professor, Faculty of Basic Sciences, Department of Biology, University of Shahrekord, Shahrekord, Iran

Abstract

Analysis of population genetic structure using multilocus data, is generally performed by sequential run of the three computer programs, namely STRUCTURE, CLUMPP, and Distruct. The two latter programs lack Graphical User Interface (GUI), and the user must manually create inputs and settings files, using text editors out of the program environment. A number of facilitating programs have been developed, which are online, or dependent on statistical packages such as R. In this paper, after reviewing the aims and methods for analyzing the structure of populations using dominant molecular marker data, a new computer program, STRUCTUREasy is introduced. This program simplifies the analysis procedures, with no need for programing skills. STRUCTUREasy is an open source program with a simple user interface, running in Microsoft Office. Its application is for the extraction of Q-matrices of STRUCTURE and connecting the input files to downstream software, for providing ease of use, and more accuracy and pace, in population structure analysis using multilocus markers. This program runs in Microsoft Access 2016, and performs well in extracting data and settings between STRUCTURE, CLUMPP and Distruct software, for accuracy and pace.

Keywords: Population Genetics, Structure, Computer Program, CLUMPP, Distruct, STRUCTURE, STRUCTUREasy.

* msht.ir@gmail.com

تاکسونومی و بیوسیستماتیک، سال دهم، شماره سی و ششم، پاییز ۱۳۹۷، صفحه ۳۵- ۴۸
دریافت مقاله: ۱۳۹۷/۰۶/۱۲ پذیرش نهایی: ۱۳۹۸/۰۴/۰۵

مروری بر تحلیل ساختار جمعیت با نشانگرهای مولکولی غالب و معرفی نرم‌افزار جدید STRUCTUREasy

مجیدشریفی تهرانی *

استادیار گروه زیست‌شناسی، دانشکده علوم پایه، دانشگاه شهرکرد، شهرکرد، ایران

چکیده

تحلیل ساختار ژنتیکی جمعیت‌ها براساس نشانگرهای مولکولی عموماً با استفاده از سه برنامهٔ کامپیوتری STRUCTURE، CLUMPP و Distruct به‌طور متوالی انجام می‌شود. دو برنامهٔ CLUMPP و Distruct واسط گرافیکی کاربر ندارند و برای اجرای آنها لازم است پژوهشگر ویژگی‌های فایل‌های ورودی و خروجی و تنظیمات مربوط به انتخاب الگوریتم‌ها و شاخص‌ها را در خارج از برنامه‌های اصلی به‌طور دستی آماده کند. تاکنون نرم‌افزارهای کمکی مختلفی برای تسهیل کار نوشته شده‌اند که بیشتر به‌طور آنلاین یا وابسته به پکیج آماری R هستند. در مقالهٔ حاضر با مرور اهداف و روش‌های تحلیل ساختار جمعیت با استفاده از داده‌های نشانگرهای مولکولی دامیننت، برنامهٔ کامپیوتری جدید STRUCTUREasy که روند تحلیل‌ها را تسریع می‌کند و به آشنایی با زبان برنامه‌نویسی نیاز ندارد، معرفی می‌شود. این برنامه به‌شکل متن باز و دارای واسط گرافیکی ساده و قابل‌اجرا در محیط Microsoft Office است. کاربرد آن در استخراج ماتریس‌های Q و اتصال آنها به نرم‌افزارهای پایین‌دست و فراهم‌کردن سرعت و دقت بیشتر برای تحلیل ساختار ژنتیک جمعیت با استفاده از نشانگرهای چندلوکوسی است. این برنامهٔ کامپیوتری در محیط بانک اطلاعاتی Microsoft Access 2016 اجرا می‌شود و تمام تنظیمات و عملیات استخراج داده‌ها بین نرم‌افزارهای STRUCTURE، CLUMPP و Distruct را برای افزایش سرعت و دقت انجام می‌دهد.

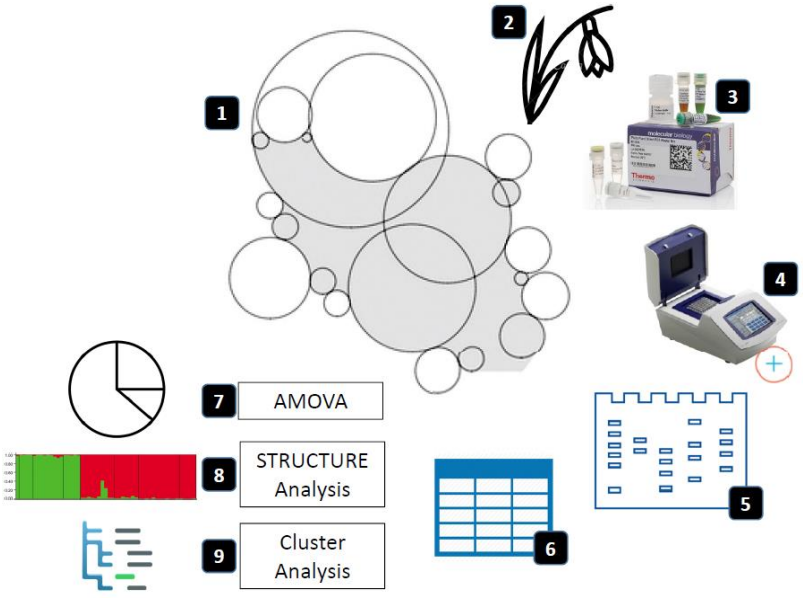
واژه‌های کلیدی: ژنتیک جمعیت، ساختار، برنامهٔ کامپیوتری، CLUMPP، Distruct، STRUCTURE، STRUCTUREasy.

مقدمه

نشانگرهای چندلوکوسی غالب (dominant) مانند ISSR، AFLPs، IRAP، REMAP و RAPDs ابزارهای مولکولی رایج در مطالعه‌های تنوع ژنتیکی‌اند. داده‌های نشانگرهای مولکولی برای گروه‌بندی افراد جمعیت در خوشه‌ها با استفاده از تحلیل‌های چندمتغیره از نوع خوشه‌بندی (Clustering) و نرم‌افزارهایی مانند NTSYSpc (Rohlf, 2000) به کار می‌روند. داده‌های مولکولی برای بررسی توزیع تنوع ژنتیکی درون و بین جمعیت‌ها با استفاده از تحلیل AMOVA در

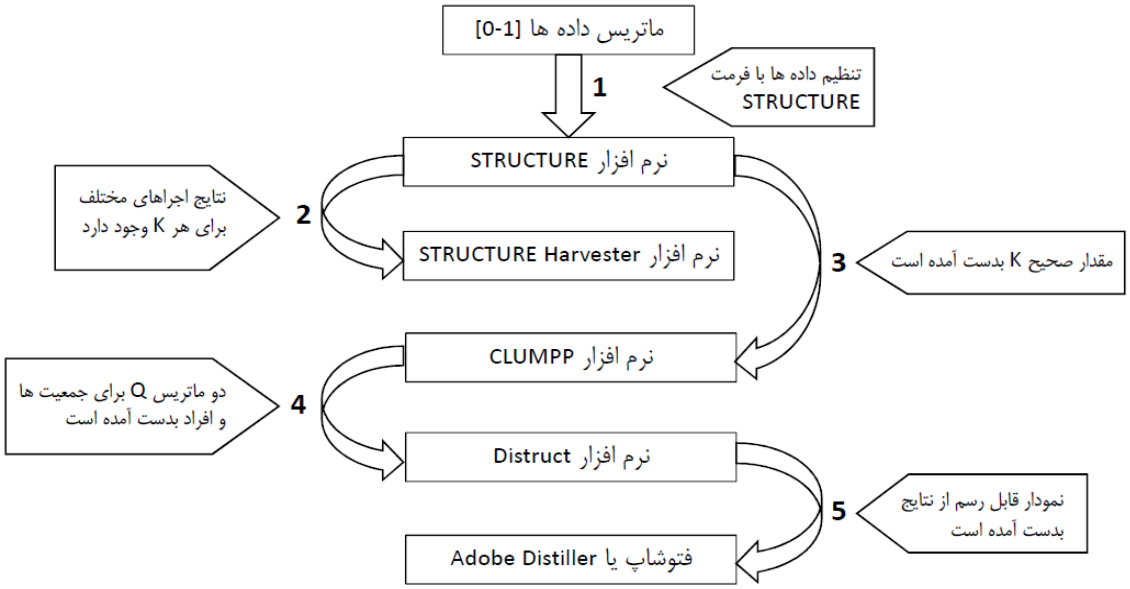
* msht.ir@gmail.com

نرم‌افزارهایی مانند GenAlEx و Arlequin نیز کاربرد دارند (Excoffier *et al.*, 2005; Peakall and Smouse, 2006). تحلیل‌های عمیق‌تر شامل آزمودن مدل‌های ژنتیکی، سیستم زادآوری جمعیت، میزان تبادل آلل‌ها بین جمعیت‌ها و تعیین ساختار ژنتیکی آنها و بررسی منشأ ژنتیکی افراد جمعیت از تعدادی خوشه (K) است. تحلیل ساختار ژنتیکی جمعیت با نرم‌افزار STRUCTURE انجام می‌شود (شکل ۱)؛ هرچند نرم‌افزارهای دیگری نیز برای این نوع تحلیل وجود دارند (Pritchard *et al.*, 2000; Falush *et al.*, 2007; Hubisz *et al.*, 2009). ساختار ژنتیکی جمعیت‌ها در تفسیر الگوهای تنوع کنونی و بررسی‌های تکاملی (Rosenberg *et al.*, 2002; Whitfield *et al.*, 2006) و بررسی مرزهای ژنتیکی جمعیت و میزان تبادل آلل‌ها در زمینه‌های مختلف ژنتیک حفاظت و بوم‌شناسی مولکولی کاربرد دارد (Manel *et al.*, 2005; Gompert and Buerkle, 2013). تحلیل STRUCTURE، یک روش خوشه‌بندی بر پایهٔ مدل ژنتیکی و نشانگرهای مستقل است که نخستین‌بار، Pritchard و همکاران آن را به کار گرفتند و Falush و همکاران آن را گسترش دادند. تحلیل STRUCTURE از روش‌های آماری موسوم به بایزین (bayesian) استفاده می‌کند؛ روش‌های بایزین، مقادیر احتمال پیشین یا توزیع داده‌ها را بر اساس حدس‌های بهینه، پیش‌از آزمایش یا مشاهده به رخدادها یا شاخص‌ها تخصیص می‌دهند و پس‌از آزمایش یا مشاهده‌های بیشتر مرتباً آنها را تصحیح می‌کنند (Pritchard *et al.*, 2000; Falush *et al.*, 2003, 2007). در تحلیل STRUCTURE، نمونه‌ها (افراد) صرف‌نظر از تعلق آنها به جمعیت‌های جغرافیایی به‌طور بدون ناظر (unsupervised) در تعدادی خوشه یا جمعیت ژنتیکی (K) دسته‌بندی می‌شوند. تمایز مفهوم جمعیت‌های جغرافیایی (جمعیت‌های پیش‌فرض) و جمعیت‌های ژنتیکی (بر اساس نتایج تحلیل داده‌های مولکولی) اهمیت دارد.



شکل ۱- تحلیل داده‌های ژنتیکی؛ ۱ و ۲. جمع‌آوری نمونه از جمعیت‌های طبیعی، ۳ و ۴. استخراج مادهٔ وراثتی و تکثیر نشانگرها با PCR، ۵ و ۶. تفکیک باندها با الکتروفورز و تشکیل ماتریس داده‌ها، ۷. تجزیه واریانس مولکولی، ۸. تحلیل ساختار ژنتیکی جمعیت، ۹. تحلیل خوشه‌ای

تحلیل ساختار ژنتیکی جمعیت‌های طبیعی با به‌کارگیری متوالی سه نرم‌افزار STRUCTURE، CLUMPP و Distruct مطابق شکل (۲) انجام می‌شود. نرم‌افزار آنلاین Structure Harvester در شکل (۲) برای تعیین تعداد خوشه‌ها (K) با استفاده از روش Evanno به کار می‌رود (Evanno *et al*., 2005; Earl, 2012). در مقالهٔ حاضر، کاربرد این نرم‌افزارها و ارتباط آنها مرور شده است.



شکل ۲- نمودار مراحل تحلیل ساختار با استفاده از نرم‌افزارهای STRUCTURE، CLUMPP و Distruct

نرم‌افزار STRUCTURE

در تحلیل STRUCTURE، هر فرد جمعیت، یک ترکیب ژنتیکی منشأیافته از تعدادی خوشه (جمعیت ژنتیکی، K) در نظر گرفته می‌شود. به‌منظور تعیین تعداد بهینهٔ خوشه‌ها (K)، مقادیر مختلف K در تحلیل STRUCTURE آزموده و هر بار، میزان تعلق هر فرد به هریک از خوشه‌ها بررسی و این مقادیر مختلف K در تکرارهای مستقل با مدل‌های ژنتیکی مختلف آزموده می‌شوند؛ درنتیجه، این نوع تحلیل‌ها وقت‌گیرند. فایل ماتریس دادهٔ ورودی به نرم‌افزار STRUCTURE، فایل متنی (text) ساده‌ایست که در آن، ستون‌های مختلف داده‌ها با Tab جدا می‌شوند. در سطر اول، دو ستون خالی و سپس نام آلل‌ها (در اینجا، باندهای هومولوگ) آورده می‌شود. از سطر دوم تا آخر ماتریس، هر سطر به یک نمونه مربوط می‌شود که در آن، ستون اول برای نام نمونه، ستون دوم برای شمارهٔ جمعیت جغرافیایی و ستون‌های بعدی شامل مقادیر یک یا صفرند که به‌ترتیب برای نشان‌دادن حضور یا غیاب آلل‌ها به کار می‌روند. ماتریس کوچک ارائه‌شده در شکل (۳)، مثالی برای ۱۰ نمونهٔ جمعیتی متعلق به چهار جمعیت جغرافیایی و حضور و غیاب ۲۰ آلل است.

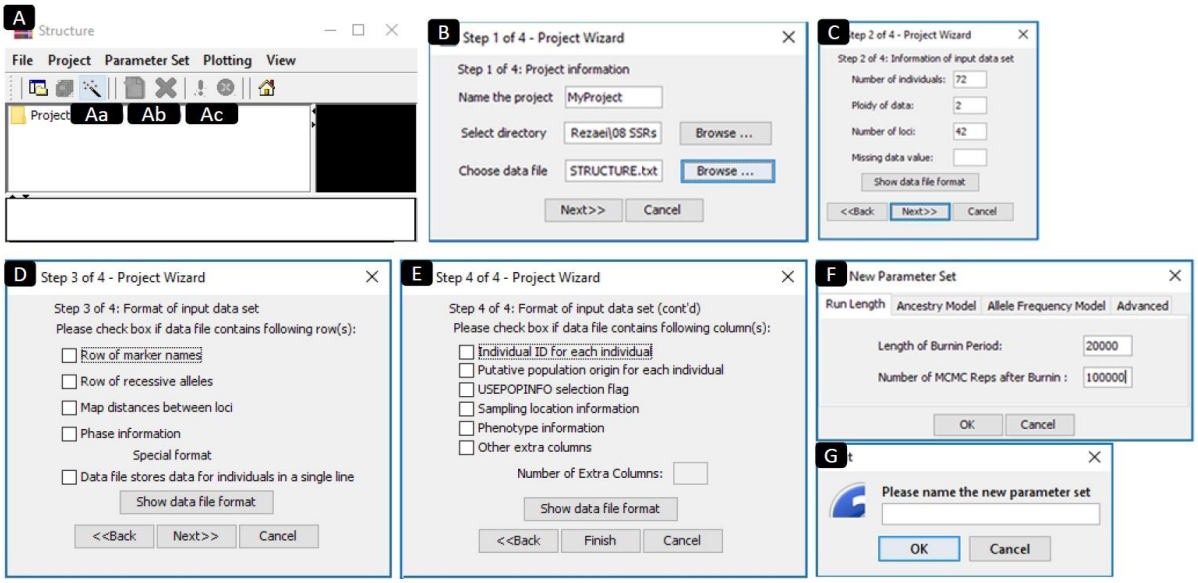
		L01	L02	L03	L04	L05	L06	L07	L08	L09	L10	L11	L12	L13	L14	L15	L16	L17	L18	L19	L20
S1	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0
S2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S3	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S4	2	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S5	2	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S6	3	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S7	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S8	3	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	1	1	1	0	0
S9	4	1	0	1	1	0	0	0	1	1	0	1	0	0	1	0	1	1	1	0	0
S10	4	1	0	1	1	0	0	0	1	1	0	1	0	0	1	0	1	1	1	0	0

شکل ۳- فرمت ماتریس ورودی نرم‌افزار STRUCTURE

پس‌از اجرای برنامهٔ STRUCTURE (شکل ۴، A)، پروژهٔ جدیدی با فرمان Aa ایجاد می‌شود؛ سپس نام پروژه، مسیر پوشه، فایل داده‌ها و تعداد افراد (نمونه‌های جمعیتی)، سطح پلوئیدی داده‌ها و تعداد آلل‌ها وارد می‌شوند (شکل ۴، B و C)؛ برای نشانگرهای دامیننت، سطح پلوئیدی داده‌ها ۱ در نظر گرفته می‌شود. چنانچه سطر اول ماتریس ورودی داده‌ها شامل نام آلل‌ها باشد، باید گزینهٔ Row of Marker Names تیک زده شود (شکل ۴، D)، چنانچه ستون اول ماتریس ورودی داده‌ها شامل نام نمونه‌ها باشد، گزینهٔ Individual ID for each individual و چنانچه ستون دوم ماتریس ورودی داده‌ها شامل شمارهٔ جمعیت جغرافیایی (منشأ نمونه‌ها) باشد، باید گزینهٔ Putative population origin of each individual تیک زده شود (شکل ۴، E).

پروژه‌ای که به این شکل تعریف می‌شود، شامل چند مجموعه شاخص (Parameter Set) است؛ با استفاده از فرمان Ab (شکل ۴) می‌توان چند مجموعه شاخص در هر پروژه ایجاد کرد. هر تحلیل با استفاده از یک مجموعه شاخص (Parameter Set) اجرا می‌شود که شامل طول پیش‌ران (Burnin period)، تعداد تکرار زنجیره‌های مارکوف (MCMC Reps در شکل ۴، F) و تعیین مدل توارث (برگهٔ Ancestry Model در شکل ۴، F) است. طول پیش‌ران (۱۰۰۰۰ یا بیشتر) عموماً کمتر از تعداد تکرار MCMC (۱۰۰۰۰۰ یا بیشتر) است و مدل توارث برای گونه‌های برون‌زادگیر، Admixture Model و برای گونه‌های درون‌زادگیر، No Admixture Model است. هر مجموعه شاخص با یک نام ذخیره می‌شود (شکل ۴، G) و برای هر مجموعه شاخص می توان مقادیر مختلف K را آزمود و این فرایند را به دفعات (تکرارهای مستقل) تکرار کرد. تحلیل با انتخاب یک مجموعه شاخص و اجرای فرمان Ac در شکل (۴) انجام می‌شود؛ تحلیل با واردکردن مقدار K آغاز و در فرایندی آماری که هزاران بار تکرارشونده است، ضریب تعلق هر فرد به خوشه‌های مختلف (۱ تا K) محاسبه می‌شود. هر تحلیل باید چند بار (R) تکرار شود. مدل انتخاب‌شده برای تحلیل STRUCTURE (برگهٔ Ancestry Model در شکل ۴، F) منشأ ژنتیکی نمونه‌ها را به‌طور مخلوط (admixture: از منشأ جمعیت‌های مختلف) و یا غیرمخلوط (no admixture: از یک جمعیت) فرض می‌کند (Pritchard *et al.*, 2000)؛ ازآنجاکه الگوریتم‌های تحلیل با استفاده از شبیه‌سازی‌های تصادفی کار می‌کنند، نتایج دقیقاً یکسانی در اجراهای مختلف تولید نمی‌شوند.

نتایج تحلیل STRUCTURE برای هر ماتریس ورودی به‌شکل فایل‌های متعدد در پوشهٔ Results ذخیره می‌شوند و هر فایل حاوی دو ماتریس Q (برای جمعیت‌ها و افراد جمعیتی) است. هر ماتریس Q، ماتریسی به ابعاد C×K است که ضرایب عضویت افراد در خوشه‌ها (خوشهٔ ۱ تا K) را ذخیره می‌کند؛ C تعداد سطرهای ماتریس (تعداد افراد جمعیت یا تعداد جمعیت‌ها) و K تعداد ستون‌های ماتریس (خوشه‌ها) است. در هر سطر، جمع مقادیر مؤلفه برابر ۱ است؛ برای نمونه، تحلیل مدل ژنتیکی روی ماتریس ورودی در STRUCTURE با آزمودن K=2 تا K=10 و هریک با ۵ تکرار مستقل (R=5)، ۹۰ ماتریس Q برای افراد و جمعیت‌ها ایجاد می‌کند که در ۴۵ فایل خروجی ذخیره می‌شوند؛ از این ماتریس‌ها، ۱: برای تعیین مقدار بهینهٔ K (تعداد صحیح خوشه‌ها) و ۲: برای تعیین ضریب عضویت هر فرد در هر خوشه (۱ تا K) استفاده می‌شود. تعیین تعداد صحیح خوشه‌ها (K) به روش Evanno و گروه‌بندی ژنتیکی نمونه‌ها در خوشه‌ها با استفاده از ماتریس‌های Q متناظر با K صحیح (به تعداد R) در نرم‌افزار CLUMPP انجام می‌شود.

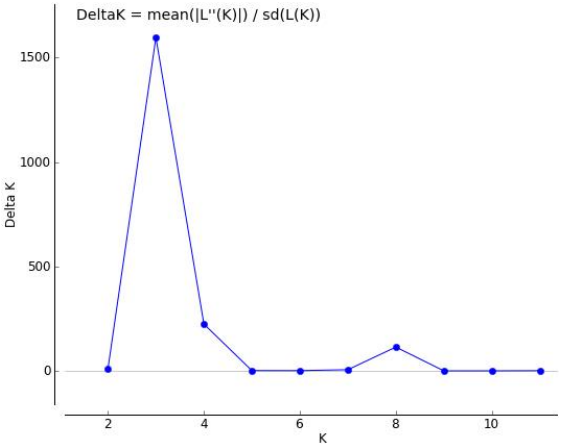


شکل ۴- مراحل اجرای نرم‌افزار STRUCTURE

تعیین تعداد خوشه‌ها (K)

مقدار K نشان‌دهندهٔ تعداد خوشه‌ها (جمعیت‌های ژنتیکی) است و ضرورتاً با تعداد جمعیت‌های جغرافیایی برابر نیست. تشخیص مقدار صحیح (بهینه) K برای هر ماتریس ورودی با استفاده از مقادیر ماتریس‌های Q و محاسبهٔ احتمال بیشینه (LnPD) و مقدار دلتای K (روش Evanno) به دست می‌آید (Evanno *et al.*, 2005). اطلاعات لازم در سطرها و ستون‌های ماتریس‌های Q در فایل‌های خروجی نرم‌افزار STRUCTURE قرار می‌گیرند. نرم‌افزار آنلاین Structure Harvester (به آدرس اینترنتی http://taylor0.biology.ucla.edu/structureHarvester) برای این محاسبه‌ها استفاده می‌شود (Earl, 2012). محتویات پوشهٔ Results حاصل از تحلیل STRUCTURE باید به‌شکل فایل ZIP فشرده و در سایت نرم‌افزار Structure Harvester آپلود شوند؛ این

نرم‌افزار دلتای K (ΔK=mean(|L''(K)|)/sd(L(K))) را محاسبه و نمودار آن را برای مقادیر مختلف K آزموده‌شده در STRUCTURE رسم و ماتریس‌های Q ورودی موردنیاز نرم‌افزار CLUMPP را استخراج می‌کند (Jakobsson and Rosenberg, 2007a, 2007b). نمودار دلتای K برازش مشخصی را در مقدار صحیح (بهینه) K نشان می‌دهد. نمونه‌ای از نمودار دلتای K که مقدار بهینهٔ K=3 را نشان می‌دهد، در شکل (۵) ارائه شده است.



شکل ۵- نمونه‌ای از نمودار دلتای K نشان‌دهندهٔ برازش منحنی در مقدار بهینهٔ K=3؛ نمونه‌ای از وب‌سایت نرم‌افزار Structure Harvester

نرم‌افزار CLUMPP: بازآرایی و تحلیل ماتریس‌های Q

ماتریس‌های Q حاصل از اجراهای مستقل (R) در مقدار بهینهٔ K باید در نرم‌افزار دیگری به نام CLUMPP بازآرایی (permutation) و تحلیل شوند تا از نتایج تقریباً برابر (حاصل از STRUCTURE)، ماتریس بهینهٔ واحد به دست آید (Jakobsson and Rosenberg, 2007a, 2017b). در این تحلیل، از R فایل خروجی STRUCTURE برای استخراج R ماتریس Q (افراد یا جمعیت‌ها) متعلق به K بهینه استفاده می‌شود. ماتریس‌های Q از فایل‌های STRUCTURE استخراج و در دو فایل با پسوندهای popfile. (برای ماتریس‌های Q جمعیت‌ها) و indfile. (برای ماتریس‌های Q نمونه‌ها) ترکیب می‌شوند؛ سپس ستون دوم ماتریس‌های Q افراد (در فایل با پسوند indfile.) به اعداد (شناسهٔ نمونه‌ها) تبدیل می‌شود.

نرم‌افزار CLUMPP، ماتریس‌های Q حاصل از R تکرار در نرم‌افزار STRUCTURE را بازآرایی (permutation) و تحلیل می‌کند و تطبیق نزدیکی از R تکرار موجود به دست می‌آورد. این نرم‌افزار از دو فایل ورودی (indfile و popfile) و یک فایل تنظیمات به نام paramfile استفاده می‌کند (Pritchard *et al.*, 2000; Jakobsson and Rosenberg, 2007b). فایل تنظیمات (paramfile) شامل اطلاعات مربوط به نوع ماتریس ورودی (افراد/جمعیت‌ها)، نام فایل‌های ورودی، تعداد افراد یا جمعیت‌ها، تعداد تکرارها، نوع الگوریتم استفاده‌شده برای محاسبه و نام فایل خروجی است (جدول ۱). نرم‌افزار CLUMPP واسط گرافیکی کاربر (GUI) ندارد. با قراردادن دو فایل ورودی و یک فایل تنظیمات در پوشهٔ برنامهٔ CLUMPP و اجرای این برنامه، نتایج به‌شکل دو فایل خروجی (ماتریس‌های Q جدید بازآرایی‌شده) ذخیره می‌شوند. این تحلیل نیز مانند STRUCTURE وقت‌گیر است.

جدول ۱- تنظیمات نرم‌افزار CLUMPP (فایل paramfile)

DATATYPE	تعیین نوع ماتریس Q (صفر: افراد جمعیت، ۱: جمعیت‌ها)
INDFILE	نام فایل ورودی حاوی R ماتریس Q برای افراد جمعیت
POPFILE	نام فایل ورودی حاوی R ماتریس Q برای جمعیت‌ها
OUTFILE	نام فایل خروجی حاوی ماتریس Q بازآرایی‌شده
MISCFILE	نام فایل خروجی برای گزارش شاخص‌های استفاده‌شده در تحلیل
K	تعداد خوشه‌ها
C	تعداد نمونه‌های جمعیتی یا تعداد جمعیت‌ها (بر اساس DATATYPE)
R	تعداد تکرارهای مستقل در تحلیل STRUCTURE
M	الگوریتم استفاده‌شده (۱. full search، ۲. Greedy، ۳. LargeKGreedy)
W	وزن‌دهی بر اساس تعداد نمونه در هر جمعیت (صفر: نه، ۱: بله)
S	نوع ضریب شباهت استفاده‌شده (۱: G، ۲: G')
GREEDY_OPTION	اگر مقدار M برابر ۲ یا ۳ باشد، ترتیب ورود ماتریس‌های حاصل از اجراهای مستقل را به تحلیل تعیین می‌کند (۱: همهٔ ترتیب‌های ممکن برای اجراهای مستقل آزموده شوند، ۲: تعداد مشخصی ترتیب اجرای مستقل به‌طور تصادفی آزموده شوند، ۳: ترتیب‌های ورودی تعیین‌شده توسط پژوهشگردر فایل permutationfile آزموده شوند)
REPEATS	تعداد ترتیب‌های ورودی برای اجراهای مستقل (تکرار permutation)؛ اگر GreedyOption برابر ۲ باشد، REPEAT عدد دلخواهیست، اگر GreedyOption برابر ۳ باشد، REPEAT برابر سطرهای فایل permutationfile است، اگر GreedyOption برابر ۱ باشد، مقدار REPEAT نادیده گرفته می‌شود.
PERMUTATIONFILE	نام فایل ماتریس permutationfile، وقتی GreedyOption برابر ۳ باشد.
PRINT_PERMUTED_DATA	ذخیرهٔ ماتریس‌های Q استفاده‌شده با ستون‌های بازآرایی‌شده در فایل‌های خروجی؛ صفر: ذخیره نشوند، ۱: در یک فایل ذخیره شوند، ۲: هر ماتریس در یک فایل خروجی ذخیره شود.
PERMUTED_DATAFILE	نام فایل خروجی برای PRINT_PERMUTED_DATA، وقتی مقدار آن برابر ۱ باشد. اگر PRINT_PERMUTED_DATA برابر ۲ باشد، از پسوندهای عددی در انتهای نام فایل استفاده می‌شود.
PRINT_EVERY_PERM	ذخیرهٔ هر بازآرایی در ستون‌های ماتریس‌های Q و مقادیر H یا H' در فایل خروجی؛ صفر: ذخیره نشود، ۱: ذخیره شود. توجه: در صورت ذخیره، فایل خروجی بسیار بزرگی تولید خواهد شد.
EVERY_PERMFILE	نام فایل خروجی برای PRINT_EVERY_PERM
PRINT_RANDOM_INPUTORDER	اگر GREEDY_OPTION برابر ۲ باشد، با قراردادن این تنظیم برابر ۱، همهٔ ترتیب‌های ورودی تصادفی اجراهای مستقل در خروجی ذخیره می‌شوند. اگر این تنظیم برابر صفر باشد، ترتیب‌های ورودی ذخیره نمی‌شوند.
RANDOM_INPUTORDERFILE	نام فایل خروجی برای PRINT_RANDOM_INPUTORDER
OVERRIDE_WARNINGS	نمایش هشدارهای غیرضروری در زمان اجرای تجزیه‌وتحلیل؛ نمایش داده نشوند، ۱ نمایش داده شوند.
ORDER_BY_RUN	ترتیب خوشه‌ها در ماتریس Q بازآرایی‌شدهٔ حاصل همانند ترتیب آنها در ماتریس Q مشخصی (r) از میان تکرارهای مستقل (R) باشد. اگر این تنظیم برابر صفر قرار داده شود، وقتی M=1 باشد، ترتیب خوشه‌ها مشابه اجرای ۱ خواهد بود و اگر M برابر ۲ یا ۳ باشد، ترتیب خوشه‌ها مشابه ترتیب نخستین اجرای ورودی به CLUMPP خواهد بود. اگر این تنظیم برابر صفر نباشد، ترتیب خوشه‌ها مشابه ترتیب آنها در یکی از اجراهای تعیین‌شده توسط کاربر خواهد بود. این تنظیم برای مقایسهٔ خروجی‌های مختلف که از ترتیب‌های ورودی مختلف حاصل می‌شوند، کاربرد دارد.

نرم‌افزار Distruct: ترسیم نتایج

نتایج تحلیل CLUMPP با استفاده از برنامهٔ رسام به نام Distruct به نمودار تبدیل می‌شوند؛ این برنامه نیز واسط گرافیکی کاربر ندارد. شاخص‌های ترسیم با Distruct باید در فایل تنظیمات به نام drawparams تعریف شوند (جدول ۲). با قراردادن فایل‌های ورودی (خروجی‌های CLUMPP) و یک فایل تنظیمات به نام drawparams در پوشهٔ برنامهٔ Distruct و اجرای برنامه، نتیجه به‌شکل نمودار در فایلی با فرمت postscript (دارای پسوند فایل PS.) ذخیره می‌شود. این فرمت فایل می‌تواند با نرم‌افزار Adobe distiller یا Adobe Photoshop و ... به فرمت گرافیکی تبدیل شود (Rosenberg, 2004).

جدول ۲- تنظیمات نرم‌فزار Distruct (فایل drawparams)

نام فایل ورودی برای ماتریس Q جمعیت‌ها	#define INFILE_POPQ
نام فایل ورودی برای ماتریس Q افراد جمعیت	#define INFILE_INDIVQ
نام فایل ورودی برای عناوین زیر نمودار	#define INFILE_LABEL_BELOW
نام فایل ورودی برای عناوین بالای نمودار	#define INFILE_LABEL_ATOP
(اختیاری) نام فایل ورودی بازآرایی خوشه‌ها و رنگ آنها	#define INFILE_CLUST_PERM
نام فایل خروجی با فرمت PostScript	#define OUTFILE
تعداد خوشه‌ها	#define K
تعداد جمعیت‌های جغرافیایی	#define NUMPOPS
تعداد افراد جمعیتی	#define NUMINDS
چاپ‌شدن افراد جمعیت در نمودار	#define PRINT_INDIVS
چاپ‌شدن عناوین در بالای نمودار	#define PRINT_LABEL_ATOP
چاپ‌شدن عناوین در زیر نمودار	#define PRINT_LABEL_BELOW
چاپ‌شدن خطوط جداکنندهٔ افراد یا جمعیت‌ها در نمودار	#define PRINT_SEP
تنظیم ارتفاع فونت	#define FONTHEIGHT
فاصلهٔ عناوین بالا از بالای نمودار	#define DIST_ABOVE
فاصلهٔ عناوین زیر از زیر نمودار	#define DIST_BELOW
ارتفاع نمودار	#define BOXHEIGHT
عرض نمودار	#define INDIVWIDTH
جهت صفحه در چاپ	#define ORIENTATION
فاصلهٔ افقی نمودار از حاشیهٔ چپ صفحه	#define XORIGIN
فاصلهٔ عمودی نمودار از حاشیهٔ بالای صفحه	#define YORIGIN
مقیاس پیکسل‌های افقی	#define XSCALE
مقیاس پیکسل‌های عمودی	#define YSCALE
زاویهٔ چاپ عناوین بالا	#define ANGLE_LABEL_ATOP
زاویهٔ چاپ عناوین زیر	#define ANGLE_LABEL_BELOW

ضخامت خطوط حاشیه	#define LINEWIDTH_RIM
ضخامت خطوط جداکنندهٔ جمعیت‌ها	#define LINEWIDTH_SEP
ضخامت خطوط جداکنندهٔ افراد جمعیت	#define LINEWIDTH_IND
چاپ طیف خاکستری	#define GRAYSCALE
نمایش داده‌ها	#define ECHO_DATA
چاپ داده‌ها همراه با نمودار	#define REPRINT_DATA
چاپ نام فایل ورودی	#define PRINT_INFILE_NAME
انتخاب پالت رنگ	#define PRINT_COLOR_BREWER

نرم‌افزار STRUCTUREasy

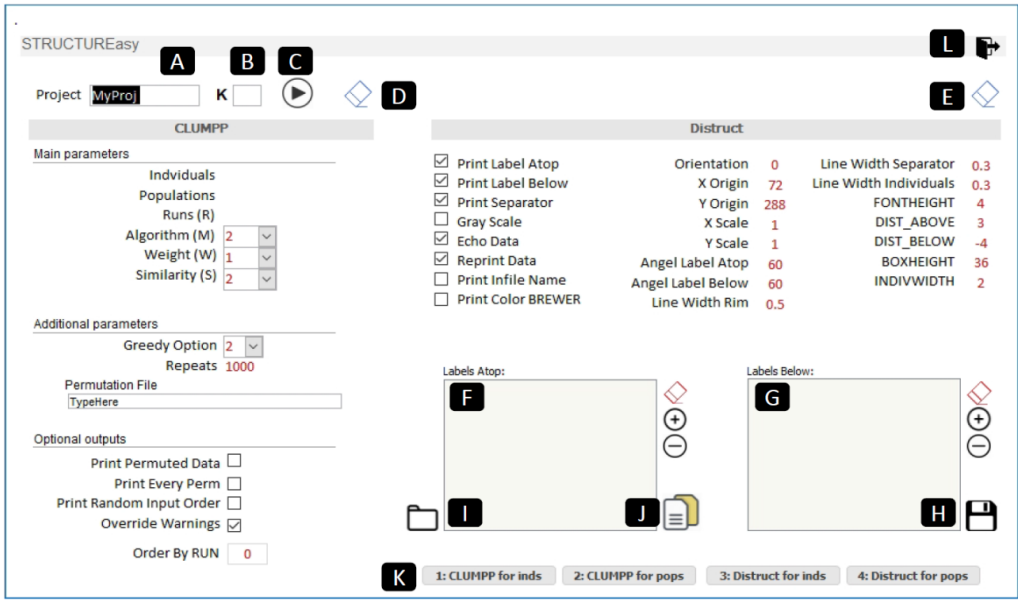
نتایج تحلیل STRUCTURE به‌شکل ماتریس‌های Q در فایل‌های خروجی متعددی ذخیره می‌شوند که استفاده از آنها را برای پژوهشگران زمینه‌های مختلف زیست‌شناسی دشوار می‌کند. نرم‌افزارهای پایین‌دست مانند CLUMPP و Distruct رابط گرافیکی ندارند و تهیهٔ فایل‌های ورودی آنها با فرمت صحیح، وقت‌گیر و دشوار است؛ به‌طوری‌که، نرم‌افزارهای کمکی متعددی برای استخراج داده‌های ماتریس‌ها و اجرای برنامه‌های CLUMPP و Distruct ارائه شده‌اند. نرم‌افزار تحت‌وب STRUCTURE Harvester عملیات استخراج ماتریس‌ها و محاسبهٔ K صحیح را انجام می‌دهد (Earl, 2012). برنامهٔ کامپیوتری CLUMPAK که وب‌سایت آن (clumpak.tau.ac.il) در دامنهٔ اینترنتی فلسطین اشغالی قرار دارد، واسط گرافیکی برای تحلیل STRUCTURE ارائه می‌دهد (Kopelman *et al.*, 2015). نرم‌افزار STRUCTURE PLOT یک نسخهٔ تحت‌وب و یک نسخهٔ عادی تحت‌پکیج آماری R برای ترسیم نتایج تحلیل CLUMPP دارد که مستلزم نصب و یادگیری مقدمات پکیج نرم‌افزاری R است (Team, 2013; Ramasamy *et al.*, 2014). برای پژوهشگران زمینه‌های بیوسیستماتیک و زیست‌شناسی حفاظت که داده‌های نشانگرهای مولکولی را در سطح جمعیت‌های طبیعی تولید می‌کنند، تحلیل کامل ساختار ژنتیکی جمعیت با CLUMPP و Distruct ممکن است تا حدی دشوار باشد؛ به‌طوری‌که، مطالعه‌های متعددی را می توان یافت که به نتایج تحلیل STRUCTURE بسنده کرده‌اند.

در مقالهٔ حاضر پس‌از مرور روند تحلیل ساختار ژنتیکی جمعیت (خلاصه‌شده در شکل ۲)، برنامهٔ کامپیوتری STRUCTUREasy که انجام تحلیل‌های CLUMPP و Distruct را با یک واسط گرافیکی تسهیل می‌کند، معرفی و ارائه می‌شود. برنامهٔ کامپیوتری STRUCTUREasy به زبان Visual Basic for Applications نوشته شده است و در محیط بانک اطلاعاتی Microsoft Access 2016 (از اجزای Microsoft Office) اجرا می‌شود. این برنامهٔ کامپیوتری، ماتریس‌های Q جمعیت‌ها و افراد را برای مقادیر مختلف K استخراج و به فرمت فایل ورودی نرم‌افزار CLUMPP تبدیل و دو فایل تنظیمات paramfile و drawparams را برای CLUMPP و Distruct ایجاد می‌کند؛ ازاین‌رو، اجرای فوری این برنامه‌ها بدون نیاز به استخراج داده‌ها و تهیهٔ فایل تنظیمات میسر می‌شود. به‌منظور استفاده از

STRUCTUREasy، نرم‌افزارهای CLUMPP و Distruct باید در پوشهٔ برنامه قرار داشته باشند. نسخهٔ رایگان این نرم‌افزارها در برنامهٔ STRUCTUREasy ضمیمه شده است و می‌توان آن را به‌شکل بانک اطلاعاتی از آدرس اینترنتی https://www.sku.ac.ir/File/26407/STRUCTUREasy (یا صفحهٔ اینترنتی این مقاله یا از طریق مکاتبه با نویسندهٔ این مقاله) دریافت کرد.

برنامهٔ STRUCTUREasy را با داشتن نتایج تحلیل STRUCTURE به تعداد R تکرار مستقل برای K بهینه اجرا کنید. برنامهٔ STRUCTUREasy تنها یک پنجره دارد (شکل ۶)؛ نام پروژه را در کادر A و مقدار K را در کادر B وارد کنید و سپس دکمهٔ C را کلیک و پوشهٔ نتایج STRUCTURE (Results) را انتخاب کنید. تنظیمات اصلی برنامهٔ CLUMPP شامل نوع الگوریتم محاسبه (M)، وزن‌دهی (W) و ضریب شباهت (S) به‌طور پیش‌فرض برای محاسبهٔ سریع همراه با وزن‌دهی و استفاده از ضریب شباهت 'G ارائه شده‌اند. الگوریتم محاسبه (M) به‌طور پیش‌فرض برابر ۲ (Greedy) است؛ ازاین‌رو، مقدار Greedy_Option برابر ۲ همراه با تعداد تکرار ۱۰۰۰ قرار داده شده است. در صورت تمایل تنظیمات را تغییر دهید؛ برچسب‌های بالا و پایین نمودار نهایی به‌طور پیش‌فرض در لیست‌ها F و G قرار داده شده‌اند؛ با دو بارکلیک روی هر عنوان، آن را تغییر دهید و در پایان، با فرمان H آن را ذخیره کنید. فرمان‌های H و I برچسب‌های بالا و پایین نمودار را در فایل‌های ورودی برنامهٔ Distruct ذخیره یا بازیابی می‌کنند. فرمان J عناوین لیست F را در لیست G کپی می‌کند. پس‌از اعمال تنظیمات، فرمان ۱ (CLUMPP for inds) را در بخش K و سپس برنامهٔ CLUMPP را با دو بار کلیک در پوشهٔ برنامه اجرا کنید؛ پس‌ازآن، فرمان ۲ (CLUMPP for pops) را در بخش K و برنامهٔ CLUMPP را با دو بار کلیک در پوشهٔ برنامه اجرا کنید. به‌منظور ترسیم نتایج تحلیل CLUMPP، فرمان ۳ (Distruct for inds) را در بخش K و سپس برنامهٔ Distruct را با دو بار کلیک در پوشهٔ برنامه اجرا کنید؛ سپس، فرمان ۴ (Distruct for pops) را در بخش K و پس‌ازآن، برنامهٔ Distruct را با دو بار کلیک در پوشهٔ برنامه اجرا کنید. نتیجه به‌شکل دو فایل postscript با نام پروژه و پسوند PS ذخیره و با Adobe Distiller یا Adobe Photoshop به فرمت گرافیکی تبدیل می‌شود.

استفاده از STRUCTUREasy سرعت تحلیل‌ها را افزایش می‌دهد و امکان انجام تحلیل‌های بیشتر در زمان کمتر و با حفظ تمرکز روی عملیات اصلی بدون اتلاف وقت برای استخراج داده‌ها و تهیهٔ فایل تنظیمات برنامه‌ها را میسر می‌کند. این نرم‌افزار تمام تنظیمات و پیچیدگی‌های نرم‌افزارهای CLUMPP و Distruct را به‌شکل ساده در اختیار پژوهشگر قرار می‌دهد؛ تمام تنظیمات مربوط به برنامه‌های CLUMPP و Distruct در زیربخش‌های مربوطه (شکل ۶) قرار دارند و مقادیر پیش‌فرض با فرمان‌های D و E بازنشانی می‌شوند. برنامهٔ کامپیوتری STRUCTUREasy به‌گونه‌ای طراحی شده است که فایل‌های خروجی با نام پروژه (کادر A) آغاز می‌شوند (جدول ۳) و ازاین‌رو، فایل‌های حاصل از هر پروژه قابل‌تشخیص هستند، فایل‌ها رونویسی نمی‌شوند و در نرم‌افزارهای دیگر به‌آسانی استفاده می‌شوند.



شکل ۶- برنامهٔ کامپیوتری SRTUCTUREasy

با فرض استفاده از نام ABCD برای نام پروژه (کادر A در شکل ۶)، فایل‌های نشان‌داده‌شده در جدول (۳) در پوشهٔ برنامهٔ STRUCTUREasy ایجاد می‌شوند؛ همراه با برنامهٔ STRUCTUREasy، ۶ پوشهٔ حاوی ماتریس‌های Q حاصل از تحلیل STRUCTURE روی داده‌های SSR و ISSR (کدشده به‌شکل صفر و یک) ارائه شده‌اند؛ این نمونه‌ها برای اجرای آزمایشی برنامه و آموزش استفاده می‌شوند. نمودارها و نتایج حاصل از این نمونه‌ها در مقالهٔ حاضر آورده نشده‌اند.

جدول ۳- فایل‌های خروجی حاصل از اجرای نرم‌افزار STRUCTUREasy با فرض استفاده از نام ABCD برای نام پروژه

	نام فایل	توضیحات
۱	ABCD.indfile	فایل ورودی برای CLUMPP؛ حاوی تعداد R ماتریس Q افراد جمعیتی برای K صحیح که از پوشهٔ نتایج STRUCTURE استخراج شده‌اند.
۲	ABCD.popfile	فایل ورودی برای CLUMPP؛ حاوی تعداد R ماتریس Q جمعیت‌ها برای K صحیح که از پوشهٔ نتایج STRUCTURE استخراج شده‌اند.
۳	paramfile	فایل تنظیمات CLUMPP؛ محتوای این فایل با اجرای فرمان ۱ و ۲ در بخش K از شکل ۶ مرتباً تغییر می‌کند تا تنظیمات لازم برای تحلیل فایل ABCD.indfile یا ABCD.popfile در آن قرار داده شود.
۴	ABCD_ind.txt	فایل خروجی CLUMMP حاصل از تحلیل فایل ABCD.indfile
۵	ABCD_pop.txt	فایل خروجی CLUMMP حاصل از تحلیل فایل ABCD.popfile
۶	drawparams	فایل تنظیمات Distruct؛ محتوای این فایل با اجرای فرمان ۳ و ۴ در بخش K از شکل ۶ تغییر می‌یابد تا تنظیمات لازم برای تحلیل فایل ABCD_ind.txt یا ABCD_pop.txt در آن قرار داده شود.
۷	ABCD_lbl_Atop.txt	فایل برچسب‌های بالای نمودار؛ این فایل توسط برنامهٔ STRUCTUREasy ایجاد می‌شود و از درون برنامه، ویرایش، ذخیره و بازیابی می‌شود.
۸	ABCD_lbl_Below.txt	فایل برچسب‌های زیر نمودار؛ این فایل توسط برنامهٔ STRUCTUREasy ایجاد می‌شود و از درون برنامه، ویرایش، ذخیره و بازیابی می‌شود.

جمع‌بندی

برنامهٔ کامپیوتری STRUCTUREasy با استخراج ماتریس‌های Q و ایجاد فایل تنظیمات نرم‌افزارهای پایین‌دست به‌طور خودکار، سهولت، سرعت و دقت بیشتری را در استفاده از نرم‌افزارهای STRUCTURE، CLUMPP و Distruct برای انجام تحلیل کامل ساختار جمعیت فراهم می‌کند؛ با استفاده از این برنامه و انتخاب تنظیمات مربوط به الگوریتم‌های سریع‌تر، سرعت انجام تحلیل‌های مقدماتی و آزمایشی افزایش می‌یابد و با حفظ تمرکز پژوهشگر روی عملیات اصلی بدون نیاز به استخراج داده‌ها و تایپ تنظیمات برنامه‌ها، تحلیل‌های بیشتر در زمان کمتر میسر و قابل‌مقایسه می‌شوند. فایل ویدئو www.sku.ac.ir/File/26412/STRUCTUREasyVideo نمونه‌ای را نشان می‌دهد که ماتریس داده‌های خام ۷۲ نمونهٔ جمعیتی و ۴۰ لوکوس SSR که به‌شکل دامیننت کد شده‌اند، با نرم‌افزار STRUCTURE تحلیل شده‌اند و برنامهٔ STRUCTUREasy تحلیل ماتریس‌های Q حاصل را با آماده‌سازی فایل‌ها و انتخاب گزینهٔ M=2 و ۱۲۰۰ تکرار انجام می‌دهد. این ویدئو با سرعت نرمال نشان داده شده است و سرعت زیاد استخراج داده‌ها در این مثال برای K=5 و انجام تجزیه‌وتحلیل‌های CLUMPP برای افراد و جمعیت‌ها و سپس ترسیم نمودارها برای هر دو، کارایی آن را نشان می‌دهد. در این نرم‌افزار، تمام تنظیمات لازم برای نرم‌افزارهای CLUMPP و Distruct در اختیار پژوهشگر قرار دارد؛ همچنین، نام‌گذاری فایل‌های ورودی و خروجی در این برنامه بر اساس نام پروژه انجام می‌شود که بازشناسی فایل‌های مربوط به تحلیل‌های مختلف را میسر می‌کند. متن کدهای این برنامهٔ کامپیوتری برای توسعه و ویرایش بیشتر در اختیار پژوهشگران قرار داده می‌شود.

سپاسگزاری

پژوهش حاضر با حمایت معاونت پژوهشی دانشگاه شهرکرد (گرنت شمارهٔ GRD1M87396) و صندوق حمایت از پژوهشگران (گرنت شمارهٔ ۹۱۰۰۳۳۵۸) انجام شده است.

منابع

Earl, D. A. (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4: 359-361.

Evanno, G., Regnaut, S. and Goudet, J. (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology 14: 2611-2620.

Excoffier, L., Laval, G. and Schneider, S. (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1: 47-50.

Falush, D., Stephens, M. and Pritchard, J. K. (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

Falush, D., Stephens, M. and Pritchard, J. K. (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. Molecular Ecology Notes 7: 574-578.

Gompert, Z. and Buerkle, C. A. (2013) Analyses of genetic ancestry enable key insights for molecular ecology. Molecular Ecology 22: 5278-5294.

Hubisz, M. J., Falush, D., Stephens, M. and Pritchard, J. K. (2009) Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources 9: 1322-1332.

Jakobsson, M. and Rosenberg N. A. (2007a) CLUMPP software and manual. University of Michigan, Ann Arbor.

Jakobsson, M. and Rosenberg, N. A. (2007b) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801-1806.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. and Mayrose, I. (2015) CLUMPP: a program for identifying clustering modes and packaging population structure inferences across K. Molecular Ecology Resources 15: 1179-1191.

Manel, S., Gaggiotti, O. E. and Waples, R. S. (2005) Assignment methods: matching biological questions with appropriate techniques. Trends in Ecology Evolution 20: 136-142.

Peakall, R. and Smouse, P. E. (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6: 288-295.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Team, R. C. (2013) R: A language and environment for statistical computing.

Ramasamy, R. K., Ramasamy, S., Bindroo, B. B. and Naik, V. G. (2014) STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. SpringerPlus 3: 431.

Rohlf, F. J. (2000) NTSYS-pc: Numerical taxonomy and multivariate analysis system, version 2.1. Exeter Software, Setauket- New York.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. and Feldman, M. W. (2002) Genetic structure of human populations. science 298: 2381-2385.

Rosenberg, N. A. (2004) Distruct: a program for the graphical display of population structure. Molecular Ecology Notes 4: 137-138.

Whitfield, C. W., Behura, S. K., Berlocher, S. H., Clark, A. G., Johnston, J. S., Sheppard, W. S., Smith, D. R., Suarez, A. V., Weaver, D. and Tsutsui, N. D. (2006) Thrice out of Africa: ancient and recent expansions of the honey bee, Apis mellifera. Science 314: 642-645.