

## سرهم‌بندی ترنسکرپتوم و شناسایی EST-SSR ها در زعفران برای مطالعه تنوع ژنتیکی

محمد رضا رضائی<sup>۱</sup>، حمیدرضا شریفی<sup>۲</sup>، علیرضا سیفی<sup>۳\*</sup>

- ۱- کارشناسی ارشد بیوتکنولوژی کشاورزی، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی خراسان رضوی، سازمان تحقیقات، آموزش و ترویج کشاورزی، مشهد، ایران
- ۲- دانشیار، بخش تحقیقات علوم زراعی و باغی، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی خراسان رضوی، سازمان تحقیقات، آموزش و ترویج کشاورزی، مشهد، ایران
- ۳- استادیار، گروه بیوتکنولوژی و به نژادی گیاهان زراعی، دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران

### چکیده

گیاه زعفران ژنوم بزرگ و پیچیده‌ای دارد، در نتیجه مطالعات ژنتیکی بر روی این گیاه با محدودیت مواجه است. همچنین به دلیل تکثیر رویشی، تنوع ژنتیکی در این گیاه کم است. مارکرهای EST-SRR مزایایی از جمله هم‌غالبيت، اختصاصیت برای لوکوس خاص و پلی‌مورفیسم بالا نسبت به مارکرهای دیگر دارند. با توجه به در دسترس بودن داده‌های ترنسکرپتوم، امکان شناسایی مارکرهای EST-SSR برای مطالعات پلی‌مورفیسم در گیاه زعفران وجود دارد. جهت توسعه مارکرهای EST-SSR، داده‌های RNA-Seq گیاه زعفران از NCBI دریافت شدند. سپس کنترل کیفیت و پیرایش داده‌ها به ترتیب توسط ابزارهای FastQC و Trimmomatic انجام گرفت. با استفاده از این داده‌ها و ابزار RNA-Bloom، سرهم‌بندی ترنسکرپتوم انجام گرفت. ابزار CD-HIT-EST برای حذف ترنسکرپت‌های مشابه و تکراری، استفاده شد. کیفیت ترنسکرپتوم با استفاده از BUSCO مورد ارزیابی قرار گرفت و درصد ترنسکرپت‌های کامل، حدود ۹۰ درصد به دست آمد. پس از دستیابی به ترنسکرپتوم با کیفیت در زعفران، از نرم‌افزار MISA برای شناسایی EST-SSR ها در ترنسکرپتوم استفاده شد و توالی‌های دارای SSR شناسایی شدند. طراحی پرایمر با استفاده از نرم‌افزار Primer3 برای توالی‌های EST-SSR، انجام شد. تعداد ۳۵۴۵۹ توالی SSR شناسایی شدند و پرایمر برای آن‌ها طراحی گردید. از تعداد ۱۰ جفت پرایمر که برای تکثیر PCR روی DNA زعفران انتخاب شدند، ۷ جفت پرایمر در اندازه پیش‌بینی شده تکثیر شدند که نشان از کارایی ۷۰ درصدی مارکر دارد. این پرایمرهای طراحی شده، مارکرهای EST-SSR هستند که در مطالعات ژنتیکی زعفران بسیار کاربردی می‌باشند.

**کلمات کلیدی:** نشانگر EST-SSR، تنوع ژنوتیپی، پلی‌مورفیسم زعفران، بیوانفورماتیک، داده ترنسکرپتوم

## ۱. مقدمه

گیاه زعفران به روش رویشی با استفاده از بنه تکثیر می‌شود. مطالعات مختلفی نشان داده اند که تنوع ژنتیکی در زعفران به دلیل تولید مثل غیرجنسی کم است. مارکرهای ریزماهواره مزایایی از جمله هم‌غالبيت، اختصاصیت برای لوکوس خاص و پلی‌مورفیسم بالا نسبت به مارکرهای دیگر دارند که آن‌ها را برای مطالعات تنوع ژنتیکی ایده‌آل ساخته است. شناسایی مارکرهای SSR برای مدیریت موثر ژرم‌پلاسم‌های این گیاه با ارزش که اطلاعات ژنتیکی کمی از آن در دسترس است، مورد نیاز است (Nemati et al., 2012). امکان ایجاد تعداد زیادی مارکر SRR از طریق توالی‌یابی ترنسکرپتوم و سرهم‌بندی *de novo* توالی‌ها وجود دارد. از مزایای این روش جدید می‌توان به صرفه‌جویی در هزینه و استفاده مستقیم از توالی ژن‌های بیان شده اشاره کرد که مطالعات ژنتیکی وابسته به صفات خاص را ممکن می‌کند (Chen et al., 2017). هدف از این مطالعه، ایجاد مارکرهای EST-SSR برای زعفران بر اساس سرهم‌بندی ترنسکرپتوم می‌باشد. با استفاده از این مارکرها می‌توان زعفران‌های ایرانی و خارجی را به صورت مولکولی از هم تفکیک کرد و مطالعات تنوع ژنتیکی را در این گیاه بارز از لحاظ اقتصادی انجام داد.

## ۲. مواد و روش‌ها

### ۲-۱. جمع‌آوری و کنترل کیفیت داده‌ها

داده‌های RNA-Seq زعفران از پایگاه داده NCBI دریافت شدند. این داده‌ها شامل خوانش‌هایی از مطالعات مختلف زعفران با کدهای SRR1140761، SRR8790835، SRR8790836، SRR1767298، SRR1767299، SRR1767300 و SRR1767301 می‌باشند که به ترتیب مربوط به کلاله قرمز، جوانه انتهایی در مرحله غیرگلدهی، جوانه انتهایی در مرحله گلدهی، بنه، برگ، کاسبرگ، پرچم و کلاله گیاه زعفران هستند. کنترل کیفیت داده‌ها توسط نرم‌افزار FastQC انجام گرفت و حضور توالی‌های با کیفیت کم، توالی‌های آداپتر و درصد هر نوکلئوتید در توالی مشخص شد (Andrews, 2010).

### ۲-۲. پیرایش، نرمال‌سازی و سرهم‌بندی داده‌ها

پس از تشخیص داده‌های با کیفیت کم توسط FastQC، پیرایش داده‌ها توسط نرم‌افزار Trimmomatic-0.39 انجام گرفت (Bolger et al., 2014). نرمال‌سازی داده‌های با استفاده از ابزار Bignorm انجام گرفت (Wedemeyer et al., 2017). سرهم‌بندی ترنسکرپتوم توسط ابزار RNA-Bloom انجام گرفت (Bushmanova et al., 2019).

### ۲-۳. حذف توالی‌های تکراری و بررسی کیفیت سرهم‌بندی

پس از سرهم‌بندی، برای اینکه ترنسکرپت‌های مشابه و تکراری از سرهم‌بندی حذف شوند، از ابزار CD-HIT-EST v4.8.1 استفاده شد (Fu et al., 2012). بررسی کیفیت ترنسکرپتوم توسط ابزار TrinityStats انجام گرفت (Haas et al., 2013). سپس بررسی کامل بودن ترنسکرپتوم توسط ابزار BUSCO v4.0.6 بر اساس ژن‌های مجموعه داده Embryophyta انجام شد (Seppey et al., 2019).

## ۲-۴. شناسایی EST-SSR ها در ترنسکرپتوم و طراحی پرایمر

شناسایی EST-SSR ها در ترنسکرپتوم با استفاده از ابزار MISA انجام شد (Beier et al., 2017). تنظیمات این ابزار برای این آزمایش به صورت توالی‌های ۲ نوکلئوتیدی با حداقل شش تکرار و توالی‌های ۳ تا ۶ نوکلئوتیدی با حداقل پنج تکرار انجام شد. طراحی پرایمر با استفاده از ابزار Primer3 به صورت خط فرمان در لینوکس انجام گرفت (Untergasser et al., 2012). این روش یک روش طراحی پرایمر با خروجی بالا می‌باشد و برای همه توالی‌های SSR شناسایی شده در مرحله قبل، پرایمر طراحی می‌شود. تنظیمات Primer3 به این صورت انجام گرفت که اندازه محصول PCR بین ۱۰۰ تا ۵۰۰ جفت باز، اندازه پرایمر بین ۱۸ تا ۲۳ جفت باز با اندازه مطلوب ۲۰ جفت باز، حداکثر اختلاف دمای بین جفت پرایمرها ۵ درجه سانتی‌گراد و درصد GC پرایمرها بین ۴۰ تا ۶۰ درصد تنظیم شدند. ۱۰ جفت پرایمر EST-SSR برای تکثیر محصول PCR روی DNA زعفران انتخاب شدند (جدول ۱).

جدول ۱- پرایمرهای EST-SSR انتخاب شده برای تایید تکثیر محصول PCR روی DNA زعفران

پرایمر	۵' -> ۳'	T <sub>m</sub>	اندازه محصول	نوع SSR
CsSSR1	F: CTTTCATCACCTCGCTCTGC	۵۹	۴۴۹	(AAT)۸
	R: TGACAGTGACATCCCTACGG	۵۹		
CsSSR2	F: TGAATTCTTGACCATCGCC	۵۷	۴۷۰	(GA)۱۵
	R: TCCCTCATTCCGTCCTC	۵۹		
CsSSR3	F: CTCCGCAGTTTCATCCAGG	۵۹	۴۱۸	(AAAGA)۶
	R: CCACCGAACCCTACCAAAGT	۵۹		
CsSSR4	F: CTTCTTCGTCGCCACCAAAG	۵۹	۴۸۶	(TTC)۵
	R: CACTCGGCGGATACATACCA	۵۹		
CsSSR5	F: GGTTGAGGCTTGAGAGATGC	۵۹	۴۹۸	(AT)۶
	R: AGTCTCCTGGTATTCATGGT	۵۸		
CsSSR6	F: CGGGTTTCATCGTGCAAAGT	۵۷	۴۲۳	(TA)۶
	R: GAGCCGTTGAGATGTTAGC	۵۹		
CsSSR7	F: GTGTACGCTGATGTGGATGC	۵۹	۵۰۰	(TCT)۱۶
	R: CCTCTTAGCTTGTCCTTGC	۵۹		
CsSSR8	F: TGGCCGAAATGAGTTGATCC	۵۷	۴۳۷	(CA)۸
	R: ATTCGTTGCCTCACTCAGC	۵۷		
CsSSR9	F: GGCCACCATCCAATCACTG	۵۹	۴۴۹	(ACTC)۶
	R: ACAAAACCATGGACCAGGAT	۵۷		
CsSSR10	F: GGATGGTTGGTTATCTACAGCA	۵۸	۴۸۴	(TACA)۶
	R: ACTGGACACACATTGCGAGA	۵۸		

## ۲-۵. استخراج DNA و PCR

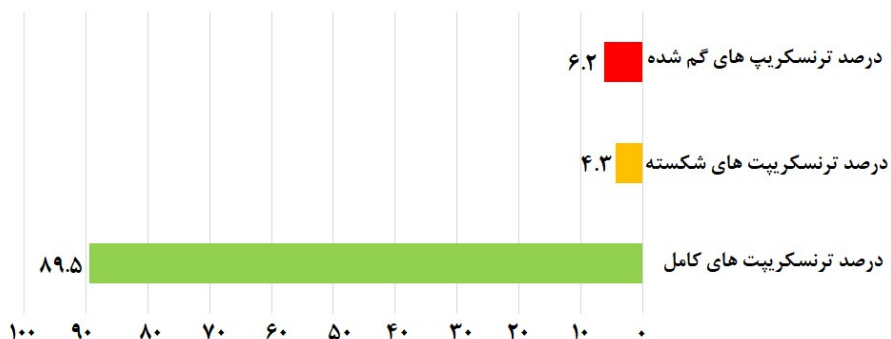
استخراج DNA به روش CTAB از بنه زعفران انجام گرفت (Doyle, 1991). PCR بر اساس چرخه دمایی جدول ۲ انجام گرفت.

جدول ۲- چرخه دمایی واکنش PCR

مرحله	دما (سانتی‌گراد)	زمان (ثانیه)	تعداد
واسرشت اول	۹۴	۲۴۰	۱
واسرشت	۹۴	۳۰	۵
اتصال	۶۰ تا ۵۵	۳۰	۵
سنتز رشته	۷۲	۶۰	۵
واسرشت	۹۴	۳۰	۳۰
اتصال	۶۰	۳۰	۳۰
سنتز رشته	۷۲	۶۰	۳۰
سنتز نهایی	۷۲	۳۰۰	۱

## ۳. نتایج و بحث

نتایج سرهم‌بندی با کیفیت و جامع از ترنسکریپتوم زعفران با استفاده از داده‌های RNA-seq بافت‌های مختلف از مطالعات مختلف و توسط ابزار RNA-Bloom انجام گرفت. پس از حذف ترنسکریپت‌های تکراری با استفاده از ابزار CD-HIT-EST در نهایت ۳۳۷۰۱۱ ترنسکریپت نهایی باقی ماندند. با استفاده از TrinityStats مشخص شد که درصد GC برابر با ۴۳/۸۰ درصد و شاخص N50 برابر با ۱۳۲۳ است که نشان از کیفیت سرهم‌بندی ترنسکریپتوم دارد. درصد ترنسکریپت‌های کامل بر اساس تجزیه و تحلیل نرم افزار BUSCO، ۸۹/۵ درصد به دست آمد (شکل ۱).



شکل ۱- بررسی کیفیت ترنسکریپتوم زعفران با استفاده از ابزار BUSCO. درصد ترنسکریپت‌های کامل، شکسته و گم شده به ترتیب ۸۹/۵، ۴/۳ و ۶/۲ درصد می‌باشد.

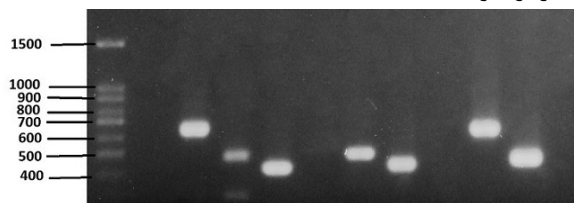
بر اساس آمار و اطلاعاتی که از نتایج شناسایی SSRها توسط نرم‌افزار MISA به دست آمد، تعداد کل SSRهای شناسایی شده ۴۴۶۶۹ و تعداد توالی‌هایی که دارای SSR هستند ۳۵۴۵۹ به دست آمد. بر این اساس، تعداد توالی‌هایی

که حاوی بیش از یک SSR بودند، ۷۱۲۹ به دست آمد. از کل تعداد SSR های شناسایی شده، SSR های دو جفت باز تعداد ۲۳۱۱۶، سه جفت باز ۱۸۸۲۶، چهار جفت باز ۱۳۰۰، پنج جفت باز ۵۴۱ و شش جفت باز ۸۸۶ به دست آمدند (جدول ۳). با استفاده از ابزار Primer3 برای ۲۴۵۲۲ توالی، پرایمر با موفقیت طراحی شد و برای ۱۵۵۱۱ توالی، پرایمر مناسبی طراحی نشد.

جدول ۳- آمار شناسایی EST-SSR در ترنسکرپتوم زعفران

تعداد	آیتم
۳۳۷۰۱۱	تعداد کل توالی های ترنسکرپتوم
۴۴۶۶۹	تعداد کل SSR های شناسایی شده
۳۵۴۵۹	تعداد کل توالی های دارای SSR
۷۱۲۹	تعداد کل توالی های دارای بیش از یک SSR
۴۶۶۵	تعداد کل SSR های ترکیبی
۲۳۱۱۶	تکرارهای دو نوکلئوتیدی
۱۸۸۲۶	تکرارهای سه نوکلئوتیدی
۱۳۰۰	تکرارهای چهار نوکلئوتیدی
۵۴۱	تکرارهای پنج نوکلئوتیدی
۸۸۶	تکرارهای شش نوکلئوتیدی

جهت تایید فرایند ایجاد مارکر EST-SSR، تعداد ۱۰ جفت پرایمر برای تکثیر PCR روی DNA زعفران انتخاب شدند. از این تعداد، محصول PCR برای ۷ جفت پرایمر در اندازه پیش بینی شده تکثیر شد (شکل ۲). نتایج، نشان دهنده کارایی ۷۰ درصدی در تکثیر مارکر دارد.



شکل ۲- تکثیر پرایمرهای EST-SSR روی ژنوم زعفران. از سمت چپ، چاهک اول لدر 100bp دنایست، چاهک دوم، کنترل و چاهک های سوم تا آخر به ترتیب از CsSSR1 تا CsSSR10 می باشند. برای CsSSR4، CsSSR7 و CsSSR10 باند مشاهده نشد.

بررسی سرهم بندی ترنسکرپتوم با استفاده از ابزارهای TrinityStats و BUSCO نشان می دهد که سرهم بندی ترنسکرپتوم از کیفیت بیشتری نسبت به سرهم بندی های بدست آمده در مطالعات قبلی، برخوردار است. در این تحقیق، N50 برابر با ۱۳۲۳، درصد GC برابر با ۴۳/۸ و ترنسکرپت های کامل حدود ۹۰ درصد به دست آمد. در جدیدترین مطالعه که بر روی ترنسکرپتوم زعفران صورت گرفته است نیز N50 برابر با ۱۵۳۲ جفت باز و درصد GC برابر با ۴۳/۵ به دست آمده است (Hu et al., 2020). در مطالعه دیگر نیز درصد ترنسکرپت های کامل ۷۳ درصد به دست آمده است (Yue et al., 2020). این افزایش ۱۷ درصدی که در درصد ترنسکرپتوم های کامل به دست آمده

است، نشان از جامعیت و کامل بودن ترنسکریپتوم دارد. با استفاده از این ترنسکریپتوم کامل، تعداد ۳۵۴۵۹ توالی دارای SSR شناسایی شد. در مطالعه‌ای که اخیراً منتشر شده است، تعداد ۱۶۰۵۷ توالی SSR در زعفران شناسایی شده است (Jain et al., 2016). بنابراین از نظر تعداد SSR، در این آزمایش دو برابر بیشتر توالی SSR نسبت به آخرین مطالعه زعفران به دست آمد. سپس برای ۲۴۵۲۲ توالی SSR، پرایمر با موفقیت طراحی شد. این تعداد زیاد پرایمر و توالی SSR شناسایی شده، به مطالعات مختلف ژنتیکی در گیاه زعفران کمک می‌کند.

#### ۴. نتیجه‌گیری

کنترل کیفیت سرهم‌بندی انجام شده در این تحقیق نشان می‌دهد که سرهم‌بندی ترنسکریپتوم از کیفیت مطلوبی برخوردار است و می‌توان از آن برای شناسایی مارکرهای EST-SRR استفاده کرد. با استفاده از این ترنسکریپتوم، SSRهای زعفران شناسایی شدند و پرایمرها با موفقیت برای این توالی‌ها طراحی شدند. تعداد ۷ جفت پرایمر از ۱۰ جفت پرایمر انتخابی، روی ژنوم زعفران تایید شدند که نشان از کارایی بالای طراحی مارکر دارد. این تعداد زیاد توالی SSR شناسایی شده در مطالعات مختلف زعفران که ژنوم آن در دسترس نیست، بسیار کاربردی می‌باشد. شناسایی این مارکرها به تفکیک مولکولی زعفران‌های ایرانی و خارجی و مطالعات تنوع ژنتیکی در گیاه زعفران کمک خواهد کرد.

#### منابع

- Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data.
- Beier, S., Thiel, T., Münch, T., Scholz, U., & Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics*, 33(16), 2583-2585. doi:10.1093/bioinformatics/btx198
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120. doi:10.1093/bioinformatics/btu170
- Bushmanova, E., Antipov, D., Lapidus, A., & Pribelski, A. D. (2019). maSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, 8(9). doi:10.1093/gigascience/giz100
- Chen, J., Li, R., Xia, Y., Bai, G., Guo, P., Wang, Z., Zhang, H., & Siddique, K. H. M. (2017). Development of EST-SSR markers in flowering Chinese cabbage (*Brassica campestris* L. ssp. *chinensis* var. *utilis* Tsen et Lee) based on de novo transcriptomic assemblies. *PloS one*, 12(9), e0184736. doi:10.1371/journal.pone.0184736
- Doyle, J. (1991). DNA Protocols for Plants. In G. M. Hewitt, A. W. B. Johnston, & J. P. W. Young (Eds.), *Molecular Techniques in Taxonomy* (pp. 283-293). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150-3152. doi:10.1093/bioinformatics/bts565
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N., & Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), 1494-1512. doi:10.1038/nprot.2013.084

- Hu, J., Liu, Y., Tang, X., Rao, H., Ren, C., Chen, J., Wu, Q., Jiang, Y., Geng, F., & Pei, J. (2020). Transcriptome profiling of the flowering transition in saffron (*Crocus sativus* L.). *Scientific Reports*, 10(1), 9680. doi:10.1038/s41598-020-66675-6
- Jain, M., Srivastava, P. L., Verma, M., Ghangal, R., & Garg, R. (2016). De novo transcriptome assembly and comprehensive expression profiling in *Crocus sativus* to gain insights into apocarotenoid biosynthesis. *Scientific Reports*, 6, 22456 .
- Nemati, Z., Zeinalabedini, M., Mardi, M., Pirseyediand, S. M., Marashi, S. H., & Khayam Nekoui, S. M. (2012). Isolation and characterization of a first set of polymorphic microsatellite markers in saffron, *Crocus sativus* (Iridaceae). *American Journal of Botany*, 99(9), e340-e343. doi:<https://doi.org/10.3732/ajb.1100531>
- Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In M. Kollmar (Ed.), *Gene Prediction: Methods and Protocols* (pp. 227-245). New York, NY: Springer New York.
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Research*, 40(15), e115-e115. doi:10.1093/nar/gks596
- Wedemeyer, A., Kliemann, L., Srivastav, A., Schielke, C., Reusch, T. B., & Rosenstiel, P. (2017). An improved filtering algorithm for big read datasets and its application to single-cell assembly. *BMC Bioinformatics*, 18(1), 324. doi:10.1186/s12859-017-1724-7
- Yue, J., Wang, R., Ma, X., Liu, J., Lu, X., Balaso Thakar, S., An, N., Liu, J., Xia, E., & Liu, Y. (2020). Full-length transcriptome sequencing provides insights into the evolution of apocarotenoid biosynthesis in *Crocus sativus*. *Computational and Structural Biotechnology Journal*, 18, 774-783. doi:<https://doi.org/10.1016/j.csbj.2020.03.022>

## Transcriptome assembly and EST-SSR detection in *Crocus sativus* for genetic diversity study

Mohammadreza Rezaei<sup>1</sup>, Hamidreza Sharifi<sup>2</sup>, Alireza Seifi<sup>3</sup>

- 1- MSc in Agricultural Biotechnology, Khorasan Razavi Agricultural and Natural Resources Research and Education Center, AREEO, Mashhad, Iran
- 2- Assistant Professor, Khorasan Razavi Agricultural and Natural Resources Research and Education Center, AREEO, Mashhad, Iran
- 3- Assistant Professor, Department of Biotechnology and Plant Breeding, College of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran

### Abstract

*Crocus sativus* has large and complex genome, therefore genetic studies have been limited in this plant. In addition, Genetic diversity is low due to vegetative propagation in this plant. EST-SSR markers have some priority, for example co-dominant inheritance, locus specific and highly polymorphic against all other markers. Due to the availability of transcriptome data, it is possible to develop EST-SSR markers and polymorphism studies in saffron. Development of EST-SSR markers in *C. sativus* make it possible to study genetic diversity and molecular polymorphism in different genotypes. In order to develop EST-SSR marker for *C. sativus*, we downloaded public available *C. sativus* RNA-seq data. Quality control and preprocessing of raw reads were done using FastQC and Trimmomatic tools, respectively. We performed *de novo* transcriptome assembly using RNA-Bloom. CD-HIT-EST was used in order to reduce redundancy in transcriptome assembly. The assembly quality was evaluated using the BUSCO software and completeness of transcriptome assembly was 90%. After achieving to high quality transcriptome assembly of *C. sativus*, EST-SSRs were identified by MicroSatellite identification (MISA) tool. The EST-SSRs primer were designed using Primer3. 35459 SSR-containing sequences were detected and primers pair were designed for them. Ten EST-SSR primers pair were randomly selected to amplify *C. sativus* DNA. Seven pairs of primer (70%) generated clear and reproducible bands with expected size. These EST-SSR markers can be functional and useful for *C. sativus* genetic studies.

**Keywords:** EST-SSR marker, Genotypic diversity, Saffron polymorphism, Bioinformatics, Transcriptome data