مجله بیوتکنولوژی کشاورزی

علمی-پژوهشی و ISC

ارزیابی توالی رونوشت گیاه دارویی زیره سبز (*Cuminum cyminum*) با استفاده از RNA-Seq

داریوش صادقی[1]، سید محمدمهدی مرتضویان[2]*، محمدرضا بختیاری زاده[3]

[1] دانشجوی کارشناسی ارشد اصلاح‌نباتات، پردیس ابوریحان دانشگاه تهران، تهران، ایران.

[2] دانشیار گروه علوم زراعی و اصلاح نباتات، پردیس ابوریحان دانشگاه تهران، تهران، ایران.

[3] استادیار گروه علوم دام و طیور، پردیس ابوریحان دانشگاه تهران، تهران، ایران.

تاریخ دریافت: ۱۳۹۶/۰۷/۱۸، تاریخ پذیرش: ۱۳۹۶/۱۰/۱۷

چکیده:

زیره سبز (*Cuminum cyminum* L.) گیاهی گلدار از خانواده چتریان (Apiaceae) است که بومی مناطق شرق مدیترانه تا هند می‌باشد. به‌رغم اهمیت دارویی فراوان زیره سبز، اطلاعات بسیار اندکی از ژنوم و مکانیسم‌های بیوشیمیایی ترکیبات موجود در این گیاه وجود دارد. در مطالعه حاضر جهت ارزیابی توالی رونوشت زیره سبز، از اندام گل (به دلیل تجمع بالای ترکیبات آلدئیدی و محل اصلی بیوسنتز آن)، جهت استخراج RNA از چهار نمونه و انجام آنالیزهای RNA-Seq استفاده شد. در نهایت، بیش از ۱۵۳ میلیون خوانش به طول ۵۰ باز از توالی‌یابی این نمونه‌ها حاصل شد. سرهم‌بندی خوانش‌ها به منظور ایجاد رونوشت-های بیان شده توسط نرم‌افزار Trinity (نسخه ۳ . ۱ . ۱) و با مقادیر kmer ۲۵ و ۳۲ انجام شد. بهترین سرهم‌بندی بر اساس کامل بودن توالی رونوشت‌ها با استفاده از نرم‌افزار BUSCO (نسخه ۳) شناسایی شد. بعد از سرهم‌بندی خوانش‌ها ۵۰۹۷۳ توالی ژن (کانتیگ) با میانگین طول ۷۲۵ باز و مقدار N50 برابر با ۱۱۳۶ باز به دست آمد. همچنین از این تعداد ژن، ۵۳۱۰۳ رونوشت شناسایی شد. از این تعداد، ۳۵۸۶۰ رونوشت دارای حداقل یک همولوگ در بانک اطلاعاتی Nr بودند. بیش از ۶۶/۷ درصد رونوشت‌ها حداقل دارای یک همولوگ در بانک اطلاعاتی GO (فرآیندهای بیولوژیک، عملکردهای مولکولی و اجزاء سلولی) بودند. بیشتر ژن های شناسایی شده مرتبط با تنظیم رونویسی و فعالیت های غشایی بودند. در مطالعه حاضر نخستین پروفایل توالی رونوشت در گیاه زیره گزارش شد که می‌تواند در مطالعات بعدی به منظور شناسایی ژن‌های دخیل در مسیر بیوسنتزی ترکیبات ثانویه مختلف و دیگر مطالعات ژنتیکی در این گیاه مورد استفاده قرار گیرد.

کلمات کلیدی: *رونوشت ژن، زیره سبز، گیاه دارویی، نسل دوم توالی‌یابی.*

*نویسنده مسئول: سید محمد مهدی مرتضویان تلفن: ۰۹۱۲۶۷۸۸۷۳۸ Email: Mortazavian@ut.ac.ir

مقدمه

زیره سبز با نام علمی (*Cuminum cyminum* L.) از تیره چتریان (Apiaceae) و بومی مناطق مدیترانه است که به طور گسترده در آن مناطق کشت می-شود (Sowbhagya, 2013). همچنین یکی از مهم-ترین گیاهان دارویی صادراتی برای کشورهایی نظیر ایران، هند و برخی دیگر از کشورهای آسیایی می‌باشد (Kafi, 2006). در کشور ایران، زیره سبز با حدود ۱۸ هزار هکتار، رتبه اول سطح زیر کشت در میان گیاهان دارویی را به خود اختصاص داده است (Taghavi and Eiman-Khan, 2005). این گیاه دیپلوئید (۲n=۲x=۱۴)، یکساله، دگرگرده-افشان و گلدار بوده و به عنوان مهم‌ترین گیاه دارویی اهلی در کشور ما شناخته شده است. بسیاری از نواحی کشور، مستعد کشت این گیاه دارویی مفید می‌باشند و به تدریج بر اهمیت و سطح زیر کشت آن افزوده می شود (Kafi, 2006). اصلی‌ترین ترکیبات دارویی اسانس زیره سبز کومین آلدئید، سیمن و ترپنوئیدها هستند که در بذر زیره سبز بیشترین تجمع را دارند (Thippeswamy and Naidu, 2005). حدود ۶۵-۴۰ درصد اسانس زیره سبز را کومین آلدئید تشکیل می‌دهد (Parthasarathy *et al*., 2008) که جزئی از مونوترپن‌ها محسوب می‌شود. ساخت این مونوترپن با واسطه آنزیم کلیدی لیمونن سنتاز صورت می گیرد (Mahmoud *et al*., 2004) که بیشترین بیان ژن کدکننده آن در ساقه و گل‌های بسیار کوچک (<۲mm) و کوچک (۳mm-۴) اتفاق می‌افتد (Ghanadnia *et al*., 2011). اسانس زیره سبز دارای خاصیت آنتی اکسیدانی است که تا حدود زیادی مربوط به ساختار فنلی موجود در ترکیب کومین‌آلدئید آن می‌باشد (Guenther, 1948).

درک مکانیسم‌های تنظیم بیان ژن برای ایجاد ارتباط میان ژنوتیپ و فنوتیپ امری اساسی است. سنتز و بلوغ RNAها شدیداً تحت کنترل بوده و با تشکیل یک شبکه بیان ژن، فرآیندهای بیولوژیک را هدایت می‌کند. درک عمیق اصول و مکانیسم-های حاکم بر این شبکه‌های بیان ژن برای فهم بهتر مکانیسم‌های تنظیم بیان ژن در طول مراحل نموی و پاسخ به انواع سیگنال‌های محیطی در گیاهان ضروری است. (Marguerat and Bähler, 2010). همچنین در دسترس بودن اطلاعات توالی‌های ژنوم، اصلاح‌گر را قادر به‌دستیابی به توالی ژن‌های مطلوب می‌کند. شروع استفاده از روش‌های نسل بعدی توالی‌یابی[1] به سال ۲۰۰۴ برمی‌گردد و بهبود روش‌های توالی‌یابی طی این سال‌ها منجر به تحولی در دستیابی به حجم وسیعی از اطلاعات ژنومی در موجودات مختلف شده و توانسته پاسخگوی نیازهای مختلف اصلاح‌گران در بهبود گیاهان زراعی باشد (Barabaschi *et al*., 2015). تاکنون پیش نویس ژنومی حدود ۱۰۰ گونه گیاهی

[1] Next Generation Sequencing

منتشر شده است. این دسته از روش‌ها مقرون به-صرفه، دارای تکرارپذیری بالا و انعطاف‌پذیرند (Vlk and Řepková, 2017). لازم به ذکر است، بدلیل تولید حجم انبوهی از اطلاعات، کاربرد تکنیک‌های نسل بعدی توالی‌یابی نیاز به تخصص در زمینه تجزیه و تحلیل داده‌ها و شناسایی اریبی و خطاهای موجود در طی مراحل مختلف آزمایش دارد. اریبی و خطاهای به وجود آمده در طول توالی‌یابی اثرات معنی‌داری به روی تجزیه و تحلیل‌های بیوانفورماتیک خواهند داشت و لذا روش‌های مختلفی برای بررسی و رفع این خطاها ارائه شده است. از جمله محققان دانشگاه جان هاپکینز، روشی بر مبنای K-mer به نام Rcorrector (RNA-Seq error Corrector) با هدف تصحیح خطاهای تصادفی توالی‌یابی در خوانش‌های RNA-Seq ارائه دادند (Song and Florea, 2015).

تکنولوژی RNA-Seq برای هر دو نوع موجودات مدل و غیرمدل استفاده شده است. برای موجودات غیرمدل از قبیل زیره سبز، توالی‌یابی عمیق (Deep sequencing) و به دنبال آن سرهم-بندی از ابتدا (De Novo) و خوشه‌بندی، برای دستیابی به توالی رونوشت مرجع ضروری است (Yong *et al*., 2014). طی سال‌های اخیر مطالعات بسیاری با استفاده از تکنولوژی RNA-Seq برای پی بردن به مکانیسم پیچیده انواع مقاومت به تنش-های زنده و غیرزنده و مسیر بیوسنتز ترکیبات ثانویه گیاهان دارویی و سایر گیاهان زراعی انجام شده و در حال انجام است. از جمله چنین مطالعاتی، کاربرد تکنیک RNA-Seq در گیاه دارویی (Gentiana macrophylla) است. ریشه-های خشک شده این گیاه برای درمان بیماری‌هایی چون زگیل، هپاتیت و بیماری‌های معده مورد استفاده قرار می‌گیرد. با این حال، عدم وجود اطلاعات کافی در مورد ژنتیک این گیاه، مانع از تولید ترکیبات موثره آن از طریق مهندسی ژنتیک بود. در همین راستا، کاربرد تکنیک RNA-Seq در این گیاه منجر به شناسایی ۴۲۹۱۸ تک‌ژن شد که از طریق بلاست با بانک‌های اطلاعاتی مربوط به مسیرهای بیوسنتزی نظیر بانک اطلاعاتی KEGG، تعداد ۲۳۳۹ تک‌ژن با مسیر بیوسنتزی انواع ترکیبات ثانویه در این گیاه همردیف شدند (Hua *et al*., 2014). تجزیه و تحلیل توالی رونوشت Thapsia laciniata از طریق تکنولوژی توالی‌یابی نسل بعد، به شناسایی ژن‌های جدید درگیر در بیوسنتز ترپنوئیدها منجر گردید. از مجموع ۶۶/۷۸ میلیون خوانش به دست آمده از بافت ریشه T.laciniata، ۶۴/۵۸ میلیون خوانش به ۷۶۵۶۵ کانتیگ با N50 = ۱۲۶۱ bp بازسازی شد. پس از بلاست کانتیگ‌ها از طریق بانک‌های اطلاعاتی مربوط به آنتولوژی ژن و پایگاه مسیرهای بیوسنتزی KEGG، به ترتیب تعداد ۱۷ و ۵ کانتیگ به عنوان کانتیگ‌های مسئول بیوسنتز ترپن‌ها و سزکوئی‌ترپن‌ها شناخته شدند (Drew *et al*., 2013). همچنین در سال ۲۰۱۳ توالی رونوشت گونه کرفس نیز با استفاده از این تکنولوژی مورد مطالعه قرار گرفت. پس از بازسازی خوانش‌ها در مجموع تعداد ۴۲۲۸۰ تک‌ژن تولید شد که با

استفاده از پلت‌فرم‌های مختلف از قبیل Nr، COG، GO، KEGG، همردیفی تک‌ژن‌ها مشخص شد. تعداد بسیاری SSR نیز در این گونه شناسایی و میزان تکثیر و چندشکلی ۳۱ الحاقیه از کرفس مورد بررسی قرار گرفت (Fu *et al*., 2013).

تاکنون هیچ داده توالی‌یابی از گونه‌های جنس *Cuminum* در آرشیو SRA سایت NCBI گزارش نشده است. از سوی دیگر، وجود داده‌های مولکولی بسیار اندک و ناچیز در ارتباط با زیره سبز به عنوان یکی از مهمترین گیاهان دارویی اهلی در کشور که بیشترین سطح زیر کشت را نیز به همراه دارد و دمین گیاه ادویه ای مهم در دنیاست (Bettaieb Rebey *et al*., 2012)، لزوم انجام تحقیقات ژنومیکس و ترانسکریپتومیکس را بیش از پیش ضروری می‌نماید. وجود داده‌هایی از این قبیل برای درک بهتر مسیر بیوستنزی ترکیبات ثانویه ارزشمند زیره سبز حایز اهمیت فراوان است. در پژوهش حاضر برای نخستین بار در دنیا مجموعه رونوشت‌های زیره سبز توالی‌یابی، بازسازی و از نظر عملکردی تفسیر و خوشه‌بندی گردیده است تا با کمک این اطلاعات ارزشمند مکانیسم‌های بیوسنتز این ترکیبات شناسایی گردد و راهی برای اصلاح مولکولی این گیاه و سایر گیاهان دارویی هم‌خانواده گشوده شود.

مواد و روش‌ها

بذور اکوتیپ سیوند (استان فارس) زیره سبز (*Cuminum cyminum* L.)، به مدت ۲۰ دقیقه در هیپوکلریت سدیم ۰/۲ درصد ضدعفونی (Ghanadnia *et al*., 2011) و در ادامه پنج مرتبه با آب مقطر شست و شو داده شدند. سپس به منظور حذف اثرات بازدارنده جوانه‌زنی، بذور به روی کاغذ صافی مرطوب خیسانده شدند. پس از ۴۸ ساعت، بذور در سینی‌های ۲۱۶ خانه محتوی ۷۵ درصد پیت ماس و ۲۵ درصد پرلیت به منظور تهیه نشاء کشت شدند. جهت آماده سازی گلدان‌ها مقداری سنگریزه داخل هر گلدان ریخته شد تا علاوه بر ایجاد فضای لازم جهت تبادل هوا، وزن تمامی گلدان‌ها یکسان شود. در ادامه انتقال نشاء به گلدان‌های اصلی در محیط گلخانه با میانگین دمای ۲۴ درجه سانتی‌گراد و رطوبت ۵۸ درصد انجام گرفت.

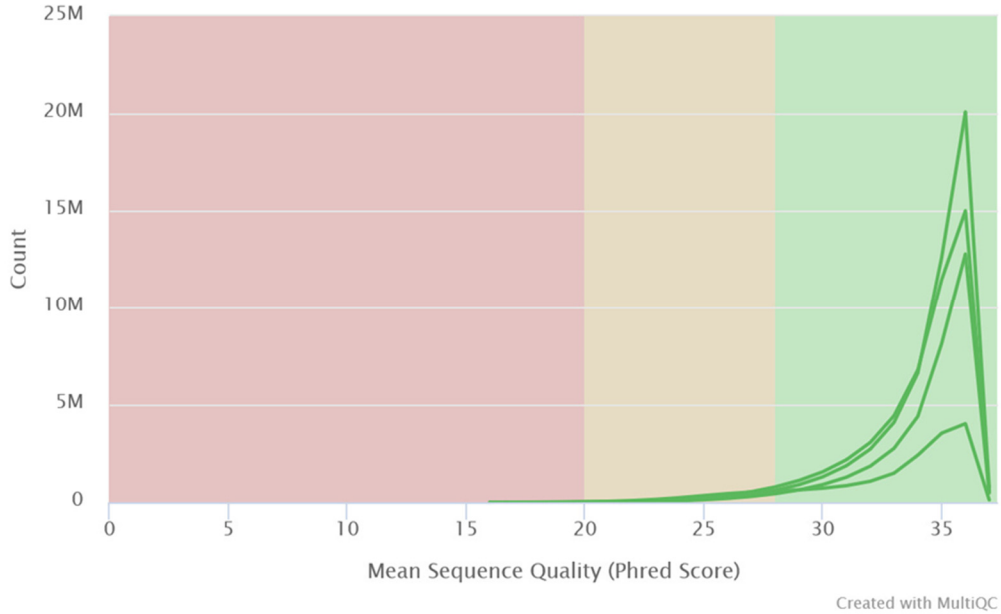
نمونه‌برداری از بافت گل گیاه، به عنوان بافت هدف (محل تشکیل بذر و تجمع متابولیت‌ها) و به صورت مخلوطی از چند بوته و به تعداد ۴ مرتبه در شرایط گلخانه در مرحله زایشی (۳۶ روز پس از کاشت) انجام شد. اندام گل، بلافاصله با ازت مایع فریز شد و برای مراحل بعدی انجام آزمایش در دمای ۸۰- درجه سانتی‌گراد نگهداری گردید. اولین مرحله در آنالیز توالی رونوشت، استخراج RNA با کیفیت و کمیت مناسب می‌باشد. استخراج RNA براساس روش پیشنهادی کیت تجاری بایوزول انجام گرفت. به منظور بررسی کیفیت RNA استخراج شده، الکتروفورز ژل آگارز انجام شد. همچنین برای اطمینان از این که مقادیر یکسانی برای ساخت کتابخانه cDNA مورد استفاده قرار می‌گیرند، کمیت RNA استخراج شده توسط دستگاه نانودراپ بررسی شد.

به منظور انجام آزمایش‌های توالی‌یابی، نمونه RNAهای استخراج شده که دارای RIN>7 (RNA Integrity Number) بودند به کمپانی BGI چین ارسال شد. نمونه‌ها به وسیله پلت‌فرم BGISEQ-500RS با استفاده از فناوری Single-end با طول قرائت ۵۰ نوکلئوتید طبق دستورالعمل کمپانی BGI توالی‌یابی شدند (جدول ۱).

ابتدا خوانش‌های خام با استفاده از نرم‌افزار Fastqc کنترل کیفیت شدند. یکی از مشکلات مرتبط با تجزیه و تحلیل داده های RNAseq اریبی و خطاهای موجود در خوانش های خام است. در همین راستا در مطالعه حاضر به منظور کاهش خطاهای احتمالی موجود در داده‌ها، از نرم‌افزار Rcorrector (نسخه ۰ . ۲) استفاده شد. در ادامه بازها و خوانش‌های با کیفیت پائین (کمتر از ۲۰) و همچنین آلودگی‌های لینکری احتمالی باقیمانده در خوانش‌ها با استفاده از نرم‌افزار Trimmomatic (نسخه 0.36) حذف شدند. به منظور سرهم‌بندی خوانش‌های خام از نرم‌افزار Trinity که بر پایه استراتژی de Bruijn's graph می‌باشد، استفاده گردید. انتخاب این نرم‌افزار بر اساس صحت بالاتر گزارش شده آن نسبت به سایر نرم‌افزارهای موجود در مطالعات قبل بود (Grabherr *et al*., 2011). جهت سرهم‌بندی خوانش‌ها و ایجاد کانتیگ‌های با کیفیت بالا دو K-mer مختلف ۲۵ و ۳۲ که در اکثر مطالعات بررسی می‌شوند، مورد تجزیه وتحلیل قرار گرفت (Blande *et al*., 2017). سپس برای ارزیابی کیفیت دو مجموعه کانتیگ بدست آمده توسط دو K-mer بیان شده، از نرم‌افزار پیشنهادی BUSCO 2.0.1 (Benchmarking Universal Single-Copy Orthologs) (Waterhouse *et al*., 2013) و سنجش N50 استفاده شد.

در نرم‌افزار BUSCO، نمونه‌برداری از صده‌ها ژنوم و گروه‌های ارتولوگ در بیش از ۹۰٪ گونه ها انجام شده است و ژن‌های ارتولوگ مربوط به شش فیلوژنی اصلی شامل مهره داران، بی‌مهرگان، جانوران چند یاخته‌ای، قارچ‌ها، یوکاریوت‌ها و ژن‌های نشانگر عمومی برای ارزیابی ژنوم‌های پروکاریوتی از طریق پایگاه OrthDB جمع‌آوری شده است. بعد از انتخاب مجموعه کانتیگی که بهتر سرهم‌بندی شده بود و جهت افزایش بازده سرهم‌بندی خوانش‌ها، کانتیگ‌های دارای بیش از ۹۵ درصد یکسانی که توسط Trinity مجزا در نظر گرفته شده بودند با استفاده از نرم‌افزار CAP3 (نسخه لینوکس) مجددا سرهم‌بندی شد.

در این تحقیق نرم افزار CAP3 کانتیگ‌هایی را که دارای بیش از ۹۵ درصد همولوژی می‌باشند به عنوان یک کانتیگ در نظر می‌گیرد. در ادامه برای بررسی کیفیت کانتیگ‌های ایجاد شده خوانش‌های مربوط به هر کدام از ۴ نمونه با توالی رونوشت حاصل از سرهم‌بندی خوانش‌ها به صورت جداگانه با استفاده از نرم‌افزار Bowtie2 (نسخه ۲ . ۳ . ۴) همردیف شدند.



شکل ۱- کیفیت خوانش‌های مربوط به ۴ نمونه پس از تصحیح با Rcorrector و انجام پیرایش.

Figure 1- Read qualities of 4 samples after correction and trimming by Rcorrector

در نهایت، همردیفی با ژنوم هویج به عنوان گیاهی هم‌خانواده با زیره سبز به منظور تعیین میزان شباهت توالی رونوشت دو گیاه جهت بررسی قابلیت انتقال نشانگرهای عمومی SSR، صورت گرفت.

به منظور بررسی قابلیت کدکنندگی کانتیگ-های ایجاد شده همه توالی‌ها توسط نرم‌افزار Transdecoder (نسخه 3.0.1) به توالی‌های پروتئینی ترجمه شدند. به منظور یافتن همولوگ احتمالی ژن‌های شناسایی شده، توالی‌های به دست آمده در برابر بانک‌های اطلاعاتی nr (بخش پروتئینی) و UniprotKB با نرم‌افزار BLASTX بررسی شدند. همچنین توالی‌های پروتئینی ترجمه شده توسط Transdecoder نیز در برابر بانک‌های اطلاعاتی مذکور BLASTP شدند. مقدار E برای بررسی آماری 10^{-5} در نظر گرفته شد (Zhang *et al.*, 2015). اطلاعات مربوط به مسیر بیولوژیکی (موجود در بانک اطلاعاتی KEGG) و عبارات ژن آنتولوژی مرتبط با ژن‌های همولوگ شناسایی شده از بانک اطلاعاتی UniprotKB استخراج شد. برای بررسی بیشتر توالی‌های بدست آمده، این توالی‌ها در برابر بانک اطلاعاتی Rfam و miRBase نیز با استفاده از نرم‌افزار BLASTN بررسی شدند که برای بررسی معنی‌داری مقدار E، 10^{-5} در نظر گرفته شد. miRBase یک سیستم متمرکز برای اختصاص نام‌های جدید به ژن‌های miRNA فراهم می‌کند(Griffiths-Jones, 2010). همچنین توالی‌ها با استفاده از نرم‌افزار HMMscan در برابر

بانک اطلاعاتی Pfam بررسی شدند تا دمین‌های احتمالی موجود در توالی‌ها شناسایی شود.

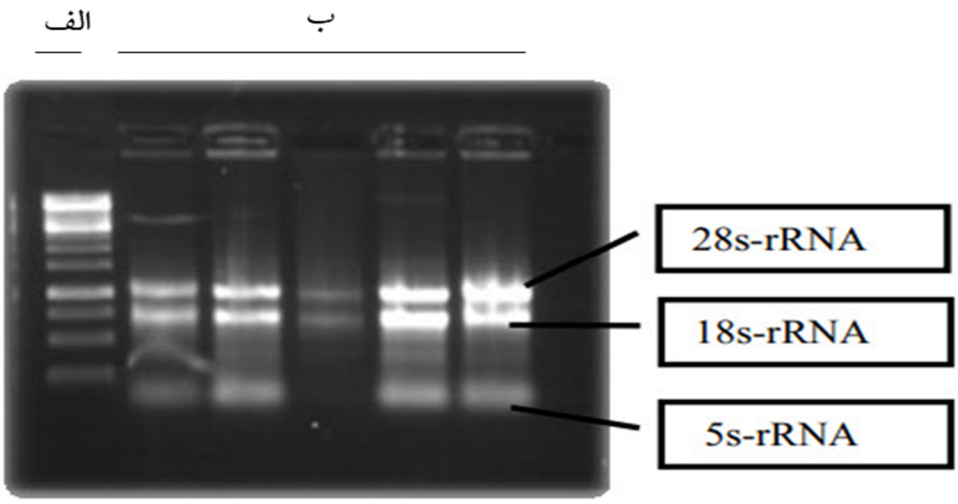
نتایج و بحث

در مطالعه حاضر با استفاده از روش سرهم‌بندی de novo اقدام به شناسایی مرجع توالی رونوشت در گیاه زیره سبز شد که گزارش در این گیاه برای نخستین بار در دنیاست.

نتایج حاصل از بررسی کیفیت و کمیت نمونه-های RNA استخراج شده در شکل (۲) نشان داده شده است. در مجموع بیش از ۱۵۲ میلیون خوانش ۵۰ نوکلئوتیدی ایجاد شد. علیرغم طول کوتاه خوانش‌ها، عمق بالای توالی‌یابی، دقت سرهم‌بندی را افزایش می‌دهد. اهمیت مقدار عمق توالی‌یابی در بهبود نتایج در مطالعات قبل ثابت شده است (Honaas *et al*., 2016; Wang and Gribskov, 2017). در فرآیند تصحیح و حذف خوانش‌های با کیفیت پایین به ترتیب ۲۸۹۱۷۵، ۲۷۴۱۰۷، ۱۲۲۸۵۸ و ۳۸۵۶۴۴ خوانش در نمونه های اول تا چهارم حذف شد (جدول ۱). در نهایت و پس از تریمینگ، بیش از ۱۵۲۶۸۰۵۸۵ خوانش حفظ و برای شناسایی ژن‌های بیان شده و ایجاد کانتیگ‌ها استفاده شد. براین اساس، دو مجموعه کانتیگ با K-mer ۲۵ و ۳۲ ایجاد شد.

نتایج حاصل از نرم‌افزار BUSCO نشان داد، از مجموع ۱۴۴۰ گروه ژنی ارتولوگ جستجو شده که در گیاهان وجود دارد، بترتیب تعداد ۹۹۳ و ۸۳۳ گروه به طور کامل در کانتیگ‌های حاصله از K-merهای ۲۵ و ۳۲ وجود دارد.

همچنین نتایج حاصل از بررسی کانتیگ‌ها نشان داد که میزان N50 در کانتیگ‌های ایجاد شده با کاربرد K-mer ۲۵ و ۳۲ به ترتیب ۱۱۳۶ و ۱۰۶۴ است. میانگین طول کانتیگ‌ها، مجموع باز-های سرهم‌بندی شده و میانگین درصد GC نیز به ترتیب، ۷۲۵ ، ۳۸۴۷۶۵۵۲ و ۴۰/۴۵ برای K-mer ۲۵ و ۷۰۹، ۳۰۶۰۶۱۳۸ و ۴۱/۸۱ برای K-mer ۳۲ بود. با توجه به اینکه کانتیگ‌های ایجاد شده توسط K-mer ۲۵ بر اساس معیارهای ذکر شده دارای کیفیت بهتری بود، این کانتیگ‌ها به عنوان سرهم‌بندی نهایی برای ادامه بررسی‌ها انتخاب شدند. در مطالعات سایر محققین، ۲۵kmer به عنوان روش سرهم بندی مناسب شناسایی شد (Aguilera *et al*., 2017; Chopra *et al*., 2014). ۵۳ درصد (۲۷۸۹۷ توالی) طول کانتیگ-های ایجاد شده بین ۲۰۰ تا ۵۰۰ نوکلئوتید بود.



شکل ۲- (الف) نشانگر وزنی 1Kb DNA Ladder RTU و (ب) نوارهای RNA دارای کیفیت‌های متنوع مربوط به استخراج‌های مختلف با استفاده از کیت بایوزول روی ژل آگارز جهت تعیین کیفیت و کمیت نمونه‌های مورد بررسی.

Figure 2- 1Kb DNA Ladder RTU (a) and RNA bands on agarose gel showing quality and quantity of different extracts using Biozol kit

جدول ۱- کمیت خوانش‌های به دست‌آمده از توالی‌یابی RNAseq در زیره سبز و تعداد خوانش‌ها قبل و بعد از پیرایش.

Table 1- Quantity of reads from RNAseq experiment in cumin and reads before and after trimming

نمونه (تکرار) Sample	طول خوانش Read length	تعداد کل خوانش‌ها Total of reads	تعداد خوانش‌های حفظ شده Survived reads	درصدGC GC%
1	50 bp	34392871	34422804	40
2	50 bp	48327034	48375158	42
3	50 bp	17261608	17277344	43
4	50 bp	52567159	52605279	41

همچنین، ۲۲ درصد (۱۱۷۹۴ توالی) و ۲۵ درصد (۱۳۱۸۷ توالی) از توالی‌ها به ترتیب طولی بین ۵۰۰ تا ۱۰۰۰ و ۱۰۰۰ تا ۳۰۰۰ نوکلئوتید داشتند. تنها برای ۲۵۵ توالی طول بزرگتر یا مساوی ۳۰۰۰ نوکلئوتید مشاهده شد.

در مطالعه بر روی توالی رونوشت گونه ای از رز (Pelargonium graveolens)، نیز ۵۳/۹۲ درصد از کانتیگ‌ها طول بین bp ۵۰۰ – ۲۰۰ داشتند و تنها ۲۴ توالی رونوشت دارای طول بزرگتر از ۳۰۰۰ نوکلئوتید بودند (Narnoliya *et al*., 2017) همچنین در مطالعه روی توالی رونوشت زیتون بیش از ۵۰ درصد کانتیگ ها در دامنه طول بین bp ۵۰۰ – ۲۰۰ گزارش شد (Martı, 2013). نتایج حاصل از همردیفی هر کدام از چهار نمونه با توالی رونوشت حاصل نشان داد که نمونه‌های اول تا چهارم به ترتیب، ۹۴/۵۱، ۹۲/۶۳، ۷۰/۹۲ و ۹۴/۱۵ درصد با توالی رونوشت حاصل از سرهم‌بندی De novo خوانش‌ها همردیف شدند. همچنین، در نمونه اول ۸۵/۵۶ درصد از کانتیگ‌ها دقیقا یکبار و ۸/۹۵ درصد کانتیگ‌ها بیش از یک بار با توالی رونوشت De novo خوانش‌ها همردیف شد. یکبار یا بیش از یکبار همردیفی کانتیگ‌های نمونه‌های دوم تا چهارم با توالی رونوشت De novo به ترتیب برابر با ۸۲/۱۶ و ۱۰/۴۷، ۶۳/۸۳ و ۷/۰۹، ۸۴/۲۶ و ۹/۸۹ درصد بود. خوانش‌های مربوط به هر کدام از چهار نمونه مورد بررسی با ژنوم هویج که بعنوان گیاهی هم خانواده با زیره سبز به شمار می آید همردیف شدند. نمونه‌های اول تا چهارم، هر کدام به ترتیب ۱۰/۶۰، ۱۵/۰۵، ۸/۲۱ و ۱۱/۵۲ درصد با ژنوم هویج همردیفی نشان دادند که بیانگر شباهت کم توالی رونوشت این دو گیاه با یکدیگر می‌باشد. در خانواده چتریان تنها چند نشانگرSSR عمومیِ در دسترس وجود دارد که این نشانگرها در هویج معرفی شده‌اند. در سال ۲۰۱۳ قابلیت انتقال نشانگرهای SSR از هویج به زیره سبز مورد تایید قرار گرفته است (Kumar *et al*., 2014). نشانگرهای SSR عموماً در محدودهٔ نواحی تکراری Non-coding یا اینترونی قرار دارند (Subramanian *et al*., 2003) و در مطالعه حاضر، توالی رونوشت یعنی نواحی ژنی مورد بررسی قرار گرفته است و به این ترتیب تناقض بین همخوانی نواحی SSR هویج با زیره و نتایج مطالعه حاضر قابل توجیه است. در عین حال، برخی SSRها در نواحی کدکننده ژنوم قرار دارند که می توان با مطالعات بیشتر نسبت به شناسایی این نواحی در بخش‌های همپوشان این دو ژنوم اقدام نمود. در ادامه همه توالی‌های ژنی مشاهده شده در برابر بانک‌های اطلاعاتی مختلف جستجو شدند تا همولوگ‌های احتمالی شناسایی شوند. نتایج حاصل از جستجو در برابر بانک اطلاعاتی پروتئنی Nr با دو روش BLASTP و BLASTX نشان داد که به ترتیب ۱۳۵۶۱ و ۱۵۰۶۵ کانتیگ دارای همولوگ معنادار با بیش از ۵۰ درصد یکسانی در این بانک اطلاعاتی می‌باشند. همچنین بدون در نظر گرفتن مقدار یکسانی و تنها بر اساس سطح E

معنادار (کمتر از 10^{-5}) این مقادیر به ترتیب برای BLASTP و BLASTX برابر با ۲۶۷۸۲ و ۳۹۳۴۲ توالی بود. نتایج حاصل از جستجو در برابر بانک اطلاعاتی پروتئینی UniprotKB نیز منجر به شناسایی ۲۱۶۶۹ و ۳۰۱۰۳ توالی معنادار با دو روش BLASTP و BLASTX گردید. در کل همولوگ‌های شناسایی شده در ۷۷۳ گونه مختلف مشاهده شد که بیشترین تعداد همولوگ‌های شناسایی شده در گیاه مدل Arabidopsis thaliana (با ۱۶۹۵۹ همولوگ) یافت شد. نتایج حاصل از جستجو در برابر بانک اطلاعاتی miRBase منجر به شناسایی ۲۸ توالی miRNA گردید و حاکی از این بود که توالی‌های miRNA در داده توالی‌های توالی رونوشت قابل شناسایی است. از این بین، ۱۵ توالی با ژن‌های موجود در بانک GO همردیف شد. همردیفی توالی‌های شناسایی شده به عنوان miRNA با ژن‌های کد کننده پروتئین‌ها در این بانک اطلاعاتی و همچنین طول بالای این ۱۵ توالی، حاکی از آن است که این توالی‌ها نخواهند توانست نماینده مناسبی از توالی‌های miRNA باشند و به احتمال بالا می‌توان عنوان نمود که شناسایی آن‌ها به عنوان miRNA به نوعی مثبت دروغین (False positives) خواهد بود. بنابراین می‌توان گفت تعداد ۱۳ توالی miRNA در داده‌های حاصل از توالی‌یابی رونوشت زیره سبز یافت شد. همچنین نتایج حاصل از جستجو در بانک اطلاعاتی Pfam نشان داد که ۲۱۶۵۴ توالی، حداقل دارای یک دمین پروتئینی می‌باشند.

به منظور ارزیابی بیشتر عملکردی رونوشت-های بیان شده در زیره سبز، اطلاعات مربوط به گروه‌های کارکردی (GO) و KEGG توالی‌هایی که دارای همولوگ در بانک اطلاعاتی UniprotKB بودند، استخراج و مورد بررسی بیشتر قرار گرفت. هدف بانک اطلاعاتی GO، ایجاد یک منبع اطلاعاتی واحد در راستای افزایش دانش در زمینه نقش ژن‌ها و پروتئین‌ها در تمام سلول‌های یوکاریوتی با استفاده از داده‌های تحت کنترل است (Gene *et al*., 2011). بر این اساس، تعداد ۲۵۲۹۶ ژن شناسایی شد که دارای حداقل یک عبارت معنادار متعلق به یکی از سه گروه فرآیندهای بیولوژیکی، عملکردهای مولکولی و ترکیبات سلولی بودند. طبقه‌بندی به سه گروه عملکردی با نتایج سایر مطالعات در این زمینه در توافق است (Peng *et al*., 2014). براساس نتایج به دست آمده از آزمایش‌های GO، گروه ترکیبات سلولی با معنی‌داری ۴۳/۷۷ درصد از ژن‌های به دست آمده، بزرگ ترین گروه حاوی اطلاعات (Chen *et al*., 2014) و پس از آن گروه عملکردهای مولکولی با ۴۲/۹۲ و فرآیندهای بیولوژیک با ۴۱/۱۴ درصد از کل ژن های شناسایی شده بیشترین اطلاعات را دربرگرفتند. عبارات شناسایی شده در نهایت به ۴۸ گروه عملکردی (Tang *et al*., 2014) طبقه‌بندی شدند. در گروه فرآیندهای بیولوژیکی، ژن‌های مربوط به تنظیم رونویسی (۲۵/۵۳ درصد)، ترانسپورترها (۱۹/۲۳ درصد)، رشد و توسعه (۱۷/۳۸ درصد) و فرآیندهای سلولی (۱۲/۳۱ درصد) به لحاظ فراوانی بر سایر ژن‌ها غالب بودند

(شکل ۳- الف). در گروه عملکردهای مولکولی، ژن‌های مربوط به binding (۵۴/۱۳ درصد)، فعالیت‌های پروتئینی (۲۰/۰۷ درصد)، فعالیت‌های رونویسی (۹/۹۹ درصد) و ترانسپورترها (۸/۲۲ درصد) بر سایر ژن‌ها غالب بودند (شکل ۳- ب). همچنین در گروه ترکیبات سلولی سلولی، ژن‌های مربوط به غشاء سلولی (۷۷/۷۱ درصد)، اجزاء سلول (۶۹/۸۱ درصد) و اجزاء اندامک‌ها (۳۳/۲۰ درصد) بر سایر ژن‌ها غالب بودند (شکل ۳- ج). غالب بودن گروه‌های اجزاء سلول در گروه ترکیبات سلولی و همچنین binding در گروه عملکردهای مولکولی با نتایج حاصل از سایر مطالعات نیز مشابهت دارند (Tang *et al*., 2014).

طبقه‌بندی بر اساس نتایج KEGG

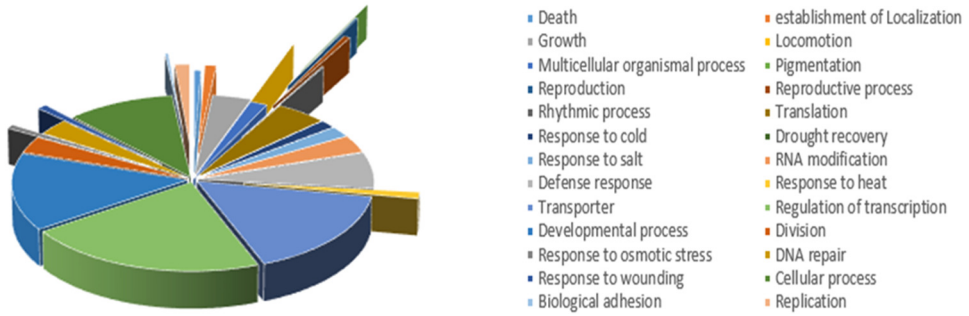
KEGG یک بانک اطلاعاتی یکپارچه و پایه‌ای متشکل از ۱۵ پایگاه داده اصلی از قبیل KEGG Gene، KEGG Genome، KEGG Phathway و ... برای تجزیه و تحلیل‌های سیستماتیک عملکرد ژن‌ها در شبکه‌های ژنی می‌باشد(Kanehisa *et al*., 2012). برای بررسی دقیق‌تر مسیرهای بیولوژیکی فعال در زیره سبز، ۵۰۹۷۳ تک‌ژن تفسیر شده توسط بلاست به ۳۴۹۲ مسیر فعال بیولوژیکی از طریق پلت‌فرم KEGG همردیف شد. این تعداد مسیر به ۵ گروه اصلی (فرآیند سلولی، پردازش اطلاعات محیط زیست، پردازش اطلاعات ژنتیکی، متابولیسم و سیستم‌های موجود زنده) طبقه‌بندی شدند. بر این اساس در ژنوم زیره سبز مسیر درگیر در متابولیسم لیپیدها با ۱۳٪ و مسیر رشد و توسعه با ۱۰٪ دارای بیشترین همردیفی با پلت‌فرم KEGG است که به ترتیب مربوط به گروه‌های اصلی متابولیسم و سیستم‌های موجود زنده می‌باشند.

این تفسیرها و طبقه بندی‌ها بعنوان یک منبع برای بررسی مسیرهای خاص از قبیل مسیر بیوسنتز متابولیت‌های ثانویه می‌باشند. فلاونوئید جزء متابولیت‌های ثانویه دارای وزن مولکولی کم است که در سیتوزول و واکوئل بسیاری از سلول‌های گیاهی تولید می‌شود و از طریق مهار پراکسیداسیون لیپیدها، فشار اکسیداتیو را در طی تنش خشکی کاهش می‌دهد (Alinian *et al*., 2016). در همین راستا و براساس طبقه‌بندی ژنوم زیره سبز از طریق پلت فرم KEGG مشخص شد که بیوسنتز فلاونوئیدها در گیاه زیره سبز در غشاء واکوئل صورت می‌گیرد و از طریق نقل و انتقال غشایی به تنش خشکی موجود در محیط پاسخ می‌دهد.

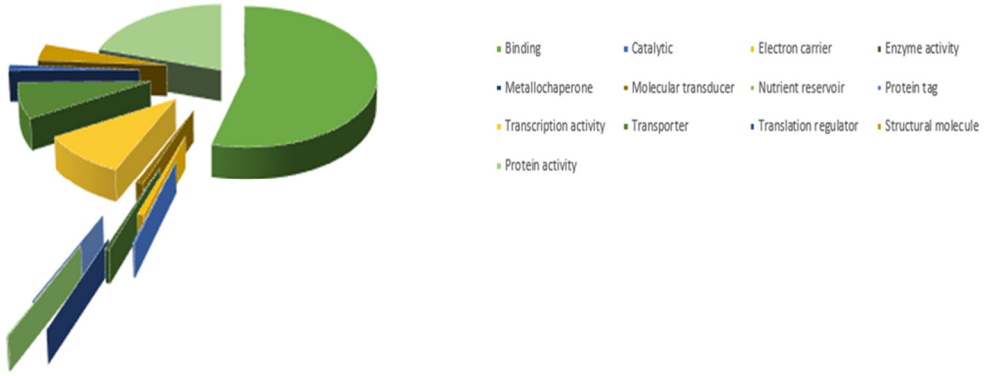
نتیجه‌گیری

بررسی‌ها نشان داده است کشور ایران در زمینه گیاه دارویی زیره‌سبز دارای مزیت نسبی صادراتی بالایی است. به علاوه، زیره سبز یکی از گیاهانی است که نیاز آبی کمی دارد و در نتیجه برای کشت در مناطق کم‌آب بسیار مساعد است (Najafi and Hasani, 2009).

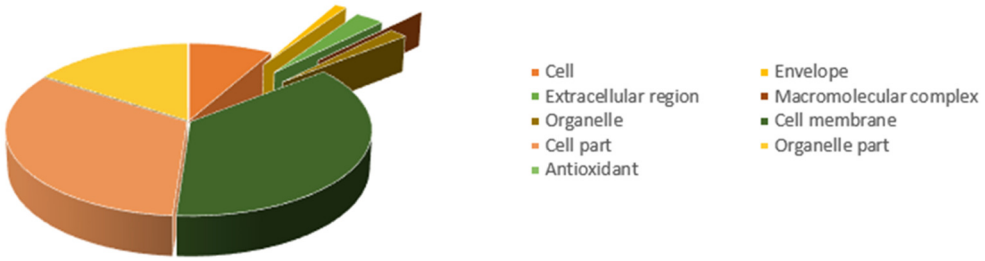
الف



ب

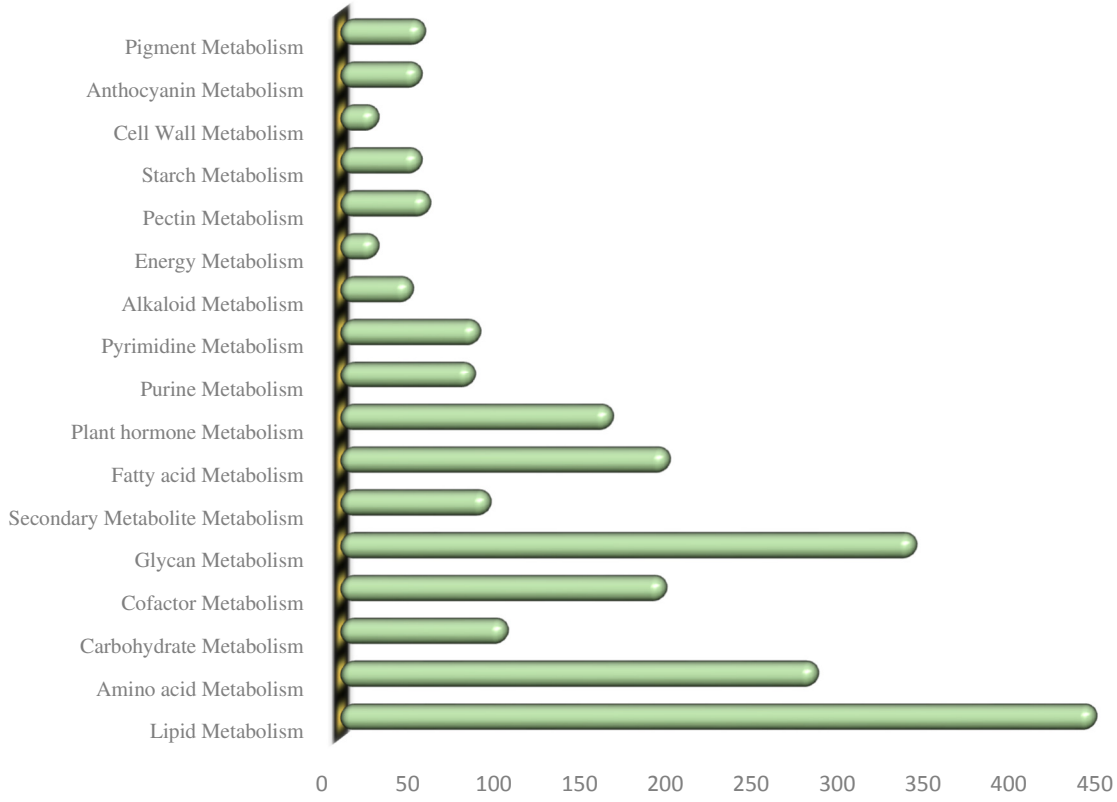


ج



شکل ۳- نمودار دایره‌ای سه گروه عمده بدست آمده از طریق بانک اطلاعاتی GO. الف) فرآیندهای بیولوژیکی. ب) عملکردهای مولکولی. ج) ترکیبات سلولی.

Figure 3- Pie diagram of three major clusters obtained from GO database. Biological process (a), molecular function (b) and cellular components (c)



شکل ۴- فراوانی مسیرهای متابولیسم ترکیبات حاصل از پلت‌فرم KEGG.

Figure 4- Metabolic pathway frequencies of compounds obtained from KEGG platform.

لذا، از طریق کاربرد تکنیک‌هایی چون RNA-Seq توان بالقوه محققین در زمینه شناسایی مسیرهای بیوسنتزی ترکیبات ثانویه و نیز شناسایی و افزایش در میزان بیان ژن‌های هدف به منظور افزایش مقاومت به تنش‌هایی نظیر خشکی، از طریق کاربرد تکنیک‌های مهندسی ژنتیک افزایش خواهد یافت. بنابراین گام اول در بهبود مقاومت گیاه به انواع تنش‌ها و افزایش میزان بیان ژن‌های درگیر در مسیرهای بیوسنتزی انواع ترکیبات ثانویه، نقشه‌یابی رونوشت خواهد بود. لذا، در این تحقیق با استفاده از تکنیک RNA-Seq کلیه اطلاعات مربوط به ژنوم و عملکردهای بیولوژیکی ژن‌های زیره سبز شناسایی و پس از انجام آزمایش‌های تکمیلی در انواع دست‌ورزی‌های ژنتیکی مورد هدف قرار خواهدگرفت.

سپاسگزاری

از حمایت مالی پارک علم و فناوری و پژوهشکده فناوری‌های همگرای دانشگاه تهران از این تحقیق قدردانی می‌گردد.

منابع

Aguilera PM, Debat HJ, Grabiele M (2017). Dataset of the transcribed 45S ribosomal RNA sequence of the tree crop "yerba mate." Data Breeds 12: 649–651.

Alinian S, Razmjoo J, Zeinali H (2016). Flavonoids , anthocynins , phenolics and essential oil produced in cumin (Cuminum cyminum L .) accessions under different irrigation regimes. Ind. Crops Production 81: 49–55.

Barabaschi D, Tondelli A, Desiderio F, Volante A, Vaccino P, Valè G, Cattivelli L (2015). Next generation breeding. Plant Sciences 242: 3–13.

Bettaieb Rebey I, Jabri-Karoui I, Hamrouni-Sellami I, Bourgou S, Limam F, Marzouk B (2012). Effect of drought on the biochemical composition and antioxidant activities of cumin (Cuminum cyminum L.) seeds. Ind. Crops Production 36: 238–245.

Blande D, Halimaa P, Tervahauta A.I, Aarts M.G.M (2017). Data Descriptor : De novo transcriptome assemblies of four accessions of the metal hyperaccumulator plant Noccaea caerulescens. Scientific Data 76: 1–9.

Chen J, Hou K, Qin P, Liu H, Yi B, Yang W, Wu W. (2014). RNA-Seq for gene identification and transcript profiling of three Stevia rebaudiana genotypes. BMC Genomics 46: 1–11.

Chopra R, Burow G, Farmer A, Mudge J, Simpson C.E, Burow M.D (2014). Comparisons of de novo transcriptome assemblers in diploid and polyploid species using peanut (Arachis spp.) RNA-Seq data. PLoS One 9: 1–16.

Drew D.P, Dueholm B, Weitzel C, Zhang Y, Sensen C.W, Simonsen H.T (2013). Transcriptome analysis of Thapsia laciniata rouy provides insights into terpenoid biosynthesis and diversity in apiaceae. International Journal of Molecular Sciences 14: 9080–9098.

Fu N, Wang Q, Shen H.L (2013). De Novo Assembly, Gene Annotation and Marker Development Using Illumina Paired-End Transcriptome Sequences in Celery (Apium graveolens L.). PLoS One 8: 1-12.

Gene T, Consortium O, Ashburner M, Ball C.A, Blake J.A, Botstein D, Butler H, Cherry J.M, Davis A.P, Dolinski K, Dwight S.S, Eppig J.T, Harris M.A, Hill D.P, Issel-tarver L, Kasarskis A, Lewis S, Matese J.C, Richardson J.E, Rubin G.M, Sherlock G (2011). Gene Ontology : tool for the unification of biology. Nat Genet 25: 25–29.

Ghanadnia M, Hadad R, Zarrinkob F, Sharifi M (2011). Different expression of limonine synthase gene in the organs and developmental stages of Cuminum cyminum L. Iranian Journal of Medicinal and Aromatic Plants 27: 495–508.

Grabherr M.G, Haas B.J, Yassour M, Levin J.Z, Thompson D.A, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma F, Birren B.W, Nusbaum C, Lindblad-toh K, Friedman N, Regev A (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29: 644-654.

Griffiths-Jones S, (2010). MiRBase: MicroRNA sequences and annotation. Current Protocols in Bioinformatics 34: 1291–12910.

Guenther E (1948). The Essential oils. Van Nostrand Company Inc., New York.

Honaas L.A, Wafula E.K, Wickett N.J, Der J.P, Zhang Y, Edger P.P, Altman N.S, Chris Pires J, Leebens-Mack J.H, DePamphilis C.W (2016). Selecting superior de novo transcriptome assemblies: Lessons learned by leveraging the best plant genome. PLoS One 11: 1–42.

Hua W, Zheng P, He Y, Cui L, Kong W, Wang Z (2014). An insight into the genes involved in secoiridoid biosynthesis in Gentiana macrophylla by RNA-seq. Molecular Biology Reports 41: 4817–4825.

Kafi M (2006). Cumin (Cuminum Cyminum): production and processing, Science Publications.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research 40: 109–114.

Kumar S, Asamadi M.H, Fougat R.S, Sakure A.A, Mistry J.G (2014). Transferability of carrot (Daucus carota) microsatellite markers to cumin (Cuminum cyminum). International Journal of Seed Spices 4: 88–90.

Mahmoud S.S, Williams M, Croteau R (2004). Cosuppression of limonene-3-hydroxylase in peppermint promotes accumulation of limonene in the essential oil. Phytochemistry 65: 547–554.

Marguerat S, Bähler J (2010). RNA-seq : from technology to biology. Cellular and Molecular Life Sciences 67:569-579

Martı M.A (2013). De Novo Assembly and Functional Annotation of the Olive (Olea europaea) Transcriptome. DNA Research 24: 93–108.

Najafi H, Hasani Y (2009). Evaluating the relative advantage of producing, exporting and identifying target markets for cumin. Journal of Agricultural Economics Researches 1: 101–122.

Narnoliya L.K, Kaushal G, Singh S.P, Sangwan R.S (2017). De novo transcriptome analysis of rose-scented geranium provides insights into the metabolic specificity of terpene and tartaric acid biosynthesis. BMC Genomics 18: 74-83.

Peng Y, Gao X, Li R, Cao G (2014). Transcriptome sequencing and de novo analysis of Youngia japonica using the illumina platform. PLoS One 9: 1–10.

Song L, Florea L (2015). Rcorrector : efficient and accurate error correction for Illumina RNA-seq reads. Gigascience 56: 1–8.

Sowbhagya H.B (2013). R. Critical Reviews in Food Science and Nutrition 53: 1–10.

Subramanian S, Mishra R.K, Singh L (2003). Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biology 4: 9-13.

Taghavi M, Eiman-Khan N (2005). Evaluation of the effect of macroeconomic variables on Iran's medicinal plants exports. TA Journal 5: 17–36.

Tang X, Xiao Y, Lv T, Wang F, Zhu Q.H, Zheng T, Yang J (2014). High-throughput sequencing and de novo assembly of the Isatis indigotica transcriptome. PLoS One 9: 1-8.

Thippeswamy N.B, Naidu K.A (2005). Antioxidant potency of cumin varieties-cumin, black cumin and bitter cumin-on antioxidant systems. European Food Research and Technology 220: 472–476.

Vlk D, ŘEPKOVÁ J (2017). Application of Next-Generation Sequencing in Plant Breeding. Czech Journal of Genetics Plant Breeding 53:76-84.

Wang S, Gribskov M (2017). Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics 33: 327–333.

Waterhouse R.M, Tegenfeldt F, Li J, Zdobnov E.M, Kriventseva E.V (2013). OrthoDB : a hierarchical catalog of animal , fungal and bacterial orthologs. Nucleic Acids Research 41: 358–365.

Yong H, Zou Z, Kok E, Kwan B, Chow K, Nasu S, Nanzyo M, Kitashiba H, Nishio T (2014). Comparative Transcriptome Analysis of Leaves and Roots in Response to Sudden Increase in Salinity in Brassica napus by RNA-seq. Biomed Research International 42 : 1-20

Zhang W, Wei X, Meng H.L, Ma C.H, Jiang N.H, Zhang G.H, Yang S.C (2015). Transcriptomic comparison of the self-pollinated and cross-pollinated flowers of Erigeron breviscapus to analyze candidate self-incompatibility-associated genes. BMC Plant Biology 15: 248-257.

Transcriptome analysis of cumin (*Cuminum cyminum* L.) using RNA-Seq

Sadeghi D.[1], Mortazavian S.M.M.* [2], Bakhtyari Zadeh M.R.[3]

[1] M.Sc student of Plant Breeding, College of Aburaihan, University of Tehran, Tehran, Iran.
[2] Associate Professor of Plant Breeding, College of Aburaihan, University of Tehran, Tehran, Iran.
[3] Assistant Professor of Animal Science, College of Aburaihan, University of Tehran, Tehran, Iran.

Abstract

Cumin (*Cuminum cyminum* L.) is a flowering plant from Apiaceae family and native to the East Mediterranean to India. The main component of essential oil in cumin seeds is cumin aldehyde (63% of total oil). Despite the importance of the cumin derivative drugs little information is available on the genome and the molecular mechanisms involved in metabolic pathway of this plant. Transcriptomic studies have greatly contributed to better understand in metabolic pathways of medicinal plants. At the moment, the use of next-generation sequencing techniques, especially RNA-seq technique were considered as the suitable promising and most accurate methods of transcriptomic evaluation. In the present study, we report cumin transcriptome for the first time. Flower tissue was used to extract RNA from four samples for RNA-seq analysis. According to the results, more than 153000000 reads with length of 50 NT were achieved. Trinity software, using 25 and 32 K-mers, was used to assemble the reads. Selection of the best assembly was followed using BUSCO software based on the integer transcript sequences. After assembly of reads, 50973 genes with an average length of 725 NT and N50 value of 1136 NT were obtained. Moreover, 53103 transcripts were identified from all genes. From this number, 35860 transcripts had at least one homologous in Nr database. More than 66.7% of all transcripts had at least one homologous in GO database (biological process, molecular function, cellular compound). Most of the genes were related to transcriptional regulation and membrane activities. In the present study, the first transcriptome profile is reported in cumin which data can be used in subsequent studies to assess expression of genes and other genetic studies in this plant.

Keywords: *Cumin, Gene transcript, Medicinal plant, Next generation sequencing.*

* Corresponding Author: Mortazavian. S.M.M. Tel+: 09126788738 Email: Mortazavian@ut.ac.ir