ژنتیک نوین
دوره هجدهم، شماره ۲، تابستان ۱۴۰۲
صفحه ۱۲۳ - ۱۱۱

ارزیابی ترانس‌کریپتوم گیاه دارویی سنبل‌الطیب (*Valeriana officinalis*) به‌منظور شناسایی ژن‌های دخیل در مسیر بیوسنتز ترپنوئیدها

Evaluation of the transcriptome of valerian (*Valeriana officinalis*) to identify genes involved in terpenoids biosynthesis pathway

آرش مختاری[۱]، منصور امیدی*[۱]، مرتضی ابراهیمی[۲]، هوشنگ علیزاده[۱]، احمد سبحانی[۲]

۱- به‌ترتیب دانشجوی دکتری، استاد، دانشیار، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران
۲- استادیار، پژوهشگاه بیوتکنولوژی کشاورزی، سازمان آموزش، تحقیقات و ترویج کشاورزی، کرج، ایران

Mokhtari A[1], Omidi M*[1], Ebrahimi M[2], Alizade H[1], Sobhani A[2]

1- PhD Student, Professor, Associate Professor, College of Agriculture and Natural Resources University of Tehran, Karaj, Iran
2- Assistant Professors, Agricultural Biotechnology Research Institute of Iran, Agricultural Research, Education, and Extension Organization (AREEO), Karaj, Iran

* نویسنده مسئول مکاتبات، پست الکترونیکی:momidi@ut.ac.ir

(تاریخ دریافت: ۱۴۰۱/۰۸/۲۶ - تاریخ پذیرش: ۱۴۰۱/۱۰/۱۲)

چکیده

گیاه دارویی سنبل‌الطیب منبع ترکیبات دارویی مؤثر برای درمان اعصاب، صرع و مشکلات خواب است. این ترکیبات عمدتاً از دسته ترپن‌ها (سزکوئی‌ترپنوئیدها) هستند که مسیر بیوسنتزی آن‌ها در سنبل‌الطیب تا حد کمی شناخته شده است. نظر به فقدان توالی ژنومی این گیاه، مطالعات مبتنی بر مونتاژ *de novo* ترانس‌کریپتوم حایز اهمیت است. در این مطالعه، دو ابزار Trinity و rnaSPAdes برای ساخت مونتاژ *de novo* از روی داده‌های SRR موجود در این گیاه انجام شد. در ادامه و به‌منظور تعیین مونتاژ بهینه، ابزار کمکی شامل CAP3 و CD-HIT-EST بر روی مونتاژهای اولیه اعمال شد. با توجه به نتایج ابزار سنجش کیفیت از قبیل، rnaQUST، SeqKit statistics و BUSCO، اعمال ترکیبی دو ابزار CAP3 و CD-HIT بر روی Trinity (T-CAP-CD) از لحاظ پارامترهای مختلف منجر به تولید مونتاژ بهینه شد. در مرحله انتولوژی با ابزار گیاه-ویژه، فوق‌العاده سریع و جامع Hayai-Annotation Plants، ۵۷/۷۸ درصد از ژن‌های مونتاژ T-CAP-CD در سه زیر گروه (BP, MF, CC) حاشیه‌نویسی‌شد. در بررسی مرتبط با بازسازی مسیر KEGG، تعداد ۳۰ ارتولوگ ژنی در مسیر ستون فقرات بیوسنتزی ترپنوئیدها و ۷ ارتولوگ ژنی نیز در مسیر بیوسنتز سزکوئی‌ترپنوئیدها شناسایی شد. بیشترین فراوانی خانواده‌های فاکتورهای رونویسی به‌ترتیب متعلق به bHLH (۹/۵ درصد)، NAC (۷/۳ درصد) و MYB (۶/۶ درصد) بود.

واژه‌های کلیدی
ترانس‌کریپتوم
سزکوئی‌ترپن
سنبل‌الطیب
فاکتور رونویسی

مقدمه

گیاه دارویی سنبل‌الطیب (*Valeriana officinalis*) متعلق به خانواده Caprifoliaceae و حاوی بیش از ۱۵۰ ترکیب شیمیایی است که مهمترین آن‌ها شامل والرنیک اسید و والپوتریات‌ها می‌باشد (Zamini et al. 2016). در طول سال‌ها از ریزوم سنبل‌الطیب برای درمان اعصاب، صرع و آرام‌بخشی در اضطراب عصبی استفاده شده است (Nandhini et al. 2018). والرنیک اسید متعلق به زیر گروهی از ترپنوئیدها به نام سزکوئی‌ترپنوئیدها (Pyle et al. 2012) بوده و تنها در دو جنس خویشاوند *Valeriana* و *Centranthus* و اساساً در بافت‌های ریشه و گل‌آذین تولید می‌شود. مسیر بیوسنتز والرنیک اسید به‌طور کامل شناخته نشده اما طی مسیر موالاونات (MVA) و فارنسیل پیرو فسفات (FPP) ساخته می‌شود (Wong et al. 2018). تمایل روزافزون به توسعه داروهای جدید از مواد شیمیایی یا متابولیت‌های ثانویه گیاهی با فعالیت‌های دارویی و درمانی بالقوه و نیز داشتن حداقل عوارض جانبی منجر به تحقیقات متمرکز بر روی گیاهان مورد استفاده در طب سنتی شده است (Cherukupalli et al. 2016). مواد شیمیایی موجود در گیاهان دارویی از طریق مسیرهای بیوسنتزی متنوعی از جمله مسیرهای موالونات، شیکیمیک، اسید استیک و غیره ساخته می‌شوند. در هر یک از این مسیرها، آنزیم‌های محدودکننده سرعت وجود دارد که واکنش‌های کلیدی در مسیر تبدیل پیش‌سازها به محصولات را کاتالیز می‌کنند. شناسایی این قبیل آنزیم‌ها در مسیر بیوسنتز متابولیت‌های ثانویه و همچنین بررسی ژن‌های کدکننده آن‌ها گام مهمی در راستای شناسایی و رفع گلوگاه‌های موجود در مسیرهای بیوسنتزی و متعاقباً اصلاح گیاهان دارویی خواهد بود. در این زمینه، توالی‌یابی نسل بعدی (NGS)[1] ترانس‌کریپتوم منجر به شناخت دقیق‌تر و کاوش عمیق‌تر مسیرهای ساخت متابولیت‌های ثانویه و پیشرفت در تحقیقات ژنومیکس کاربردی در حوزه گیاهان دارویی شده است (Younessi-Hamzekhanlu et al. 2022).

توالی‌یابی ترانس‌کریپتوم در گیاهان دارویی ضروری است؛ زیرا تعداد قابل توجهی از گیاهان دارویی در زمره گیاهان غیر مدل بوده و اطلاعات ژنومی و تحقیقات کافی در مورد ژن‌های عملکردی و مکانیسم‌های ژنتیکی آن‌ها در دسترس نیست. در میان گرایش‌های مختلف اومیکس، گسترده‌ترین رشته تحقیقاتی و کاربردی به ترانس کریپتومیکس تعلق دارد؛ زیرا امکان بررسی ژن‌های عملکردی و بیان افتراقی (DEG)[2] را فراهم می‌کند (Junda Guo et al. 2021). مونتاژ یا اسمبلی[3] ترانس‌کریپتوم به‌عنوان فناوری ضروری برای تجزیه و تحلیل متابولیسم ثانویه گیاهی ظهور کرده است (Tripathi et al. 2016). بسته به وجود یا عدم وجود ژنوم مرجع، مونتاژ ترانس‌کریپتوم را می‌توان به دو گروه عمده مبتنی بر توالی ژنوم مرجع و یا مونتاژ از نو (*de novo*) تقسیم نمود. چون تعداد زیادی گیاهان دارویی غیر مدل وجود دارد، مونتاژ مبتنی بر توالی مرجع با مشکل روبرو خواهد شد. از این رو، مونتاژ *de novo* تنها روش مونتاژ مناسب برای گیاهان دارویی غیر مدل است (Junda Guo et al. 2021).

امروزه توالی‌یابی، مونتاژ *de novo* و حاشیه‌نویسی[4] یک ترانس‌کریپتوم را می‌توان با سهولت در بسیاری از آزمایشگاه‌ها انجام داد. با این حال، جریان کار می‌تواند چالش برانگیز باشد. محققان این زمینه نه تنها باید با رویه‌های مربوطه آشنا باشند، بلکه باید از بین مجموعه ابزار اختصاصی برای مونتاژ *de novo* ترانس‌کریپتوم (Bushmanova et al. 2018; Chang et al. 2015; Henschel et al. 2012; Kannan et al. 2016; Liu et al. 2016; Peng et al. 2013; Robertson et al. 2010; Xie et al. 2014)، انواع مناسب را برای این منظور نیز انتخاب کنند. در مرور منابع مرتبط با تحقیقات RNA-seq، تنوع ابزار و مراحل پردازش داده‌های حاصله مشاهده می‌شود. از طرف دیگر، RNA-seq یک کار محاسباتی فشرده است و عدم آشنایی با منابع محاسباتی نیز می‌تواند مانع بزرگی باشد. در سالیان اخیر، مطالعات زیادی به مقایسه بین ابزار مونتاژ ترانس‌کریپتوم پرداخته‌اند (Chen et al. 2011; Geniza and Jaiswal 2017; Voshall and Moriyama 2018). البته، تمام این مطالعات بر یک نکته نظر مشترک دارند؛ امروزه، هیچ ابزار مونتاژ بهینه برای تمام مجموعه داده‌ها RNA-Seq وجود ندارد (Hölzer and Marz 2019). برخی ابزار مناسب

[1] Next Generation Sequencing

[2] Differential Gene Expression
[3] Assembly
[4] Annotation

رونوشت‌های ساده‌تر در پروکاریوت‌ها بوده و برخی مخصوص ایجاد ایزوفرم‌های حاصل از پیرایش متناوب (AS)[1] در یوکاریوت‌ها هستند.

بنابراین، مقایسه، انتخاب و استفاده از ابزار مختلف برای ایجاد یک مونتاژ *de novo* جامع از روی داده‌های RNA-Seq و بهره‌برداری اصولی از نتایج حاصله جهت نیل به اهداف پایین‌دست از قبیل انجام انواع هم‌ردیفی[2] و حاشیه‌نویسی ضروری به‌نظر می‌رسد. این رویکرد به‌ویژه در خصوص گیاهان دارویی غیر مدل بسیار راهگشا خواهد بود؛ زیرا می‌تواند با پیش‌بینی ژن‌های و فاکتورهای رونویسی دخیل در مسیرهای بیوسنتزی ناشناخته، اطلاعات مفیدی در اختیار محققان قرار دهد (Zhou and Zhu 2020).

اخیراً، توالی‌یابی ترانس‌کریپتوم با استفاده از فناوری‌های توالی‌یابی نسل بعدی یا توالی‌یابی RNA-Seq به طور گسترده برای توصیف ژن‌ها و عملکرد آن‌ها در بیوسنتز متابولیت‌های ثانویه استفاده شده است (Angeloni et al. 2011). به‌منظور شناسایی و درک بیشتر مسیر بیوسنتز سزکوئی‌ترپن‌ها در جنس والرین، مطالعات RNA-seq در گونه‌های *V. officinalis* (Yeo et al. 2013)، *V. jatamansi* (Shuang and Chenshu 2020) و *V. fauriei* (Park et al. 2016) گزارش شده است.

در این تحقیق، مونتاژ *de novo* با کمک دو ابزار Trinity و rnaSPAdes بر روی مجموعه داده‌های خام RNA-Seq در گیاه دارویی سنبل‌الطیب انجام شد. پس از تعیین مونتاژ بهینه، بررسی انتولوژی ژنی، گروه‌بندی خانواده‌های مهم فاکتورهای رونویسی و بازسازی مسیر متابولیکی ترپنوئیدها به‌منظور شناسایی ارتولوگ‌های ژنی مهم در این مسیر ارائه شد که می‌تواند به درک بهتر مسیر بیوسنتز سزکوئی‌ترپنوئیدها کمک نماید.

مواد و روش‌ها

تعداد ۵ مجموعه داده خام SRR شامل SRR125357، SRR125358، SRR125359، SRR343119 و SRR14294419 مربوط به سنبل‌الطیب از آرشیو نوکلئوتید اروپا[3] (ENA) دانلود شد (جدول ۱).

در ابتدا، کیفیت خوانش در SRRهای انتخابی از لحاظ حضور آداپتور، محتوی GC، محتوی باز N، توزیع طول توالی، سطوح دوپلیکیشن و غیره با کمک ابزار FastQC v0.11.9 (Andrews 2010) مورد بررسی قرار گرفت. طبق نتایج FastQC، از ابزار Trimmomatic v.0.36 (Bolger et al. 2014) به‌منظور حذف بازهای با کیفیت پایین (Q<20) و حذف توالی‌های آداپتوری استفاده شد. تنظیمات مورد استفاده در Trimmomatic شامل (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36) بود. سپس خوانش‌های حاصله با کمک نرم‌افزار Rcorrector v1.2 (Song and Florea 2015) از لحاظ خطاهای باقی‌مانده مورد غربال و تصحیح قرار گرفتند.

خوانش‌های تصحیح‌شده حاصل از مرحله قبل، توسط دو ابزار مونتاژ Trinity v.2.9.1 (Grabherr et al. 2011) با k-mer=25 و rnaSPAdes v.3.15.4 (Bushmanova et al. 2018) با حالت اتوماتیک برای انتخاب k-mer و حداقل طول کانتیگ برابر ۲۰۰ در قالب رونوشت‌های منفرد فرضی، مونتاژ شدند.

[1] Alternative Splicing
[2] Blast
[3] https://www.ebi.ac.uk/ena/browser/text-search?query=valeriana

جدول ۱- مشخصات مجموعه داده‌های خام (paired-end) SRR به‌منظور مونتاژ *de novo* ترانس‌کریپتوم سنبل‌الطیب.

SRR	Description	Base Count (Gb)
SRR125357	Illumina Genome Analyzer IIx paired end sequencing; Leaf PE RNA-Seq	4,342
SRR125358	Illumina Genome Analyzer IIx paired end sequencing; Stem PE RNA-Seq	5,032
SRR125359	Illumina Genome Analyzer IIx paired end sequencing; Stem PE RNA-Seq	4,684
SRR343119	Illumina Genome Analyzer II paired end sequencing; Transcriptome analysis of the root	6,926
SRR14294419	Illumina MiSeq paired end sequencing; Angiosperms353 Hyb-seq	0.076

به‌منظور داشتن رونوشت‌های بلندتر و نیز کاهش تزاید[1] (افزونگی) کانتیگ‌ها، نرم‌افزار CAP3 v.2.0.0 با تنظیم overlap length cutoff > 15 (40) و overlap percent identity cutoff N > 65 (90) استفاده شد. همچنین، ابزار CD-HIT-EST v.4.6.1 (Fu et al. 2012) برای حذف بیشتر توالی‌های تزایدی کوتاه و خوشه‌بندی کانتیگ‌های غیر-تزایدی (به‌عنوان unigeneهای پیش‌بینی‌شده) درون مونتاژها به‌کار رفت (جدول ۲). برای هر دو ابزار، مقادیر n (identity) و c (word size) به‌ترتیب معادل 0.95 و 8 در نظر گرفته شد. در نهایت، نواحی کدکننده (CDS) در unigeneهای مونتاژشده توسط نرم‌افزار TransDecoder v.5.5.0 (minimum protein length=100) مشخص شد.

مونتاژهای حاصل از دو ابزار فوق و نیز اعمال دو ابزار CAP3 و CD-HIT-EST (جدول ۲) از لحاظ آماره‌های پایه برای آزمون "دقت" و "کامل‌بودن" بررسی شدند. این آماره‌ها شامل تعداد رونوشت، متوسط طول رونوشت‌های مونتاژشده، طول کلی و طول N50 (به‌عنوان کوتاه‌ترین طول کانتیگ که ٪۵۰ از کل طول مونتاژشده را نشان می‌دهد) بود که از طریق نرم‌افزار rnaQUAST v.2.2.1 (Bushmanova et al. 2016) با min_alignment=50 و SeqKit ststistics v2.3.1 مورد بررسی قرار گرفت. به‌منظور ارزیابی "کامل‌بودن" هر یک از مونتاژها (چه تعداد از ژن‌های عمومی دارای تطابق در داده‌های ورودی هستند و آیا این تطابق‌ها تکراری، تکه‌تکه یا تمام طول هستند)، از ابزار BUSCO v.5.3.2 (Paradis and Schliep 2019) بر مبنای Viridiplantea_odb10 استفاده شد. این ابزار مجموعه‌ای از ژن‌های تک-نسخه عمومی را از پایگاه داده OrthoDB (Zdobnov et al., 2021) نگهداری می‌کند.

برای شناسایی خانواده‌های فاکتورهای رونویسی، توالی پروتئینی تمام فاکتورهای رونویسی در تمام گیاهان از پایگاه داده PlantTFDB v.5 (http://planttfdb.gao-lab.org) دانلود شد. در ادامه، Blastx (-outfmt 6 -max_target_seqs 1 -evalue 1e-5) برای جستجوی unigeneهای مونتاژشده سنبل‌الطیب در مقابل این پایگاه انجام شد.

به‌منظور حاشیه‌نویسی عملکردی ترانس‌کریپتوم، ابزار آنلاین Hayai-Annotation Plants v2.0 (http://pgdbjsnp.kazusa.or.jp/app/hayai2) به‌عنوان سیستم حاشیه‌نویسی عملکردی اختصاصی گیاهی مورد استفاده قرار گرفت. در تنظیمات این ابزار، بیشینه evalue برابر 10^{-5} و حداقل همسانی توالی برابر ۵۰ درصد در نظر گرفته شد. این ابزار از پایگاه داده KusakiDB v1.0 استفاده می‌کند.

به‌منظور تخصیص ارتولوگ‌ها و پیش‌بینی مسیرهای متابولیکی در سنبل‌الطیب، توالی آمینو اسیدی unigeneهای مونتاژ شده توسط نرم‌افزار TransDecoder وارد GhostKOALA (Kanehisa et al. 2016) شد تا حاشیه‌نویسی KO (اختصاص شماره‌های K) به داده‌های ورودی تعلق گیرد. سپس با استفاده از ابزار Kegg Mapper (Kanehisa and Sato 2020; Kanehisa et al. 2022)، حاشیه‌نویسی KO به نقشه‌های مسیر KEGG و ماژول‌های KEGG مرتبط شد.

نتایج

با استفاده از ابزار Trinity و rnaSPAdes، دو مونتاژ ترانس‌کریپتوم از روی ۵ مجموعه داده خام (SRR) ساخته شد. نتایج دو نرم‌افزار rnaQUAST و SekKit در خصوص آماره‌های مرتبط با رونوشت‌های مونتاژ شده در جدول ۲ خلاصه شده است. از لحاظ تعداد رونوشت، بیشترین مقدار مربوط به مونتاژ حاصل از ابزار Trinity بوده است. کمترین تعداد رونوشت پس از استفاده ترکیبی از دو ابزار CAP3 و CD-HIT-EST به‌ترتیب بر روی مونتاژ حاصل از ابزار rnaSPAdes (۱۲۹۰۱۸) و Trinity (۱۳۳۱٤۸) به‌دست آمد (جدول ۲). به‌طور کلی، بیشترین آماره‌های مربوطه به میانگین طول رونوشت در مونتاژ حاصل از rnaSPAdes مشاهده شد. اعمال CD-HIT-EST بر روی مونتاژ Trinity (T-CD)، کمترین میانگین را از نظر طول رونوشت‌های مونتاژشده ایجاد کرد. آماره N50 در R-CAP و T-CD به‌ترتیب در بیشینه و کمینه مقدار خود بود (جدول ۲). استفاده از ابزار CAP3 بر روی دو مونتاژ اولیه Trinity و rnaSPAsed منجر به تولید کانتیگ‌هایی با بالاترین میزان N50 و متوسط طول رونوشت شد. در بین تمام موارد، آماره Q20 برابر صفر و درصد محتوی GC دستخوش تغییرات بسیار اندکی بود. آماره Q20 نشان می‌دهد

[1] redundancy

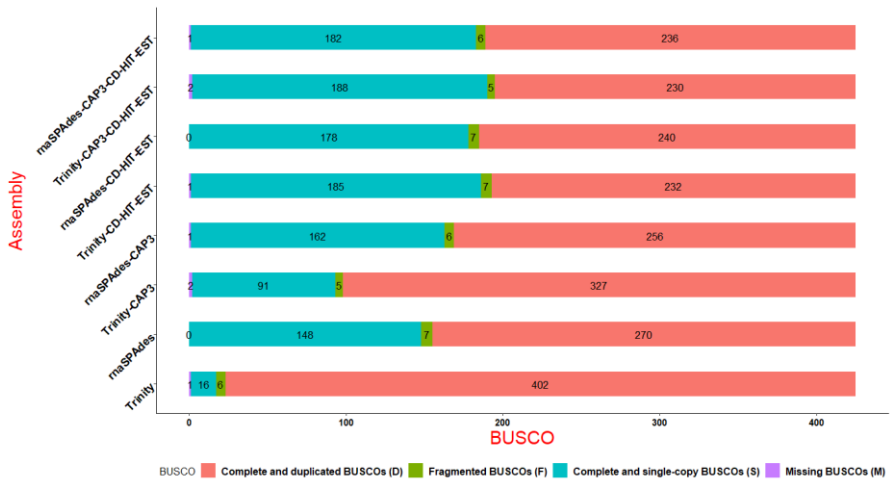
که فراخوانی باز اشتباه، ۱ در ۱۰۰ بار است. به عبارت دیگر، در هر ۱۰۰ باز که تعیین توالی شده است، ۱ باز اشتباه وجود دارد. به‌منظور بررسی پارامترهای مرتبط با کیفیت مونتاژها، تأثیر منفرد و ترکیبی دو ابزار CAP3 و CD-HIT-EST بر کیفیت مونتاژها بررسی شد. نتایج آنالیز BUSCO به‌منظور ارزیابی "کامل‌بودن" مونتاژهای *de novo* ترانس‌کریپتوم سنبل‌الطیب (شکل ۱) نشان داد که ابزار rnaSPAdes از لحاظ تولید single-copy (۱٤۸) نسبت به ابزار Trinity بهتر بوده و تعداد Duplicateهای کمتری تولید کرده و همچنین، فاقد پارامتر مفقود (Missing) بوده است. اما پس از اعمال دو ابزار CAP3 و CD-HIT، بالاترین میزان تولید single-copy در مونتاژ Trinity-CAP3-CD-HIT-EST و Trinity -CD-HIT-EST، به‌ترتیب با مقادیر ۱۸۸ و ۱۸۵ بوده است. در بین تمام موارد، اعمال CAP3 منجر به افزایش پارامتر Missing شده است (شکل ۱). بر طبق نتایج ارزیابی، مونتاژ Trinity، پس از اعمال دو ابزار CD-HIT-EST (T-CD) و CAP3 (T-CAP-CD) را می‌توان به‌عنوان مونتاژ بهینه برای ادامه مطالعات انتخاب نمود.

نتایج حاشیه‌نویسی عملکردی با ابزار آنلاین Hayai-Annotation Plants v2.0 بر روی تمام مونتاژهای مندرج در جدول ۲ نشان داد که مونتاژ T-CAP-CD با تعداد ۲۹٦۹۰ از مجموع ۱۳۳۱٤۸ رونوشت (یعنی معادل ۲۲/۲ درصد)، بیشترین میزان unigeneهای حاشیه‌نویسی‌شده را داشت. در مونتاژ T-CAP-CD، بیشترین تعداد (۳٤/٤ درصد) مربوطه به اجزای سلولی (CC) عملکردهای مولکولی (MF) بوده است.

جدول ۲- آماره‌های حاصل از رونوشت‌های مونتاژه شده سنبل‌الطیب با استفاده از دو ابزار rnaQUAST و SeqKit statistics.

Statistics	T	T-CAP	T-CD	T-CAP-CD	R	R-CAP	R-CD	R-CAP-CD
Transcripts	276912	183952	168712	133148	156569	133890	140356	129018
Transcripts > 500 bp	136633	103805	69553	66819	84568	69287	70911	64772
Transcripts > 1000 bp	78406	62258	36875	37924	52301	45452	43192	41686
Ave. len. of transcripts	878.416	992.699	752.375	887.3	982.646	1018.035	925.12	974.6
Contig N50	1502	1648	1292	1510	1720	1899	1663	1820
Q20 (%)	0	0	0	0	0	0	0	20
GC (%)	40.36	40.22	40.16	40.12	40.5	40.48	40.51	40.49

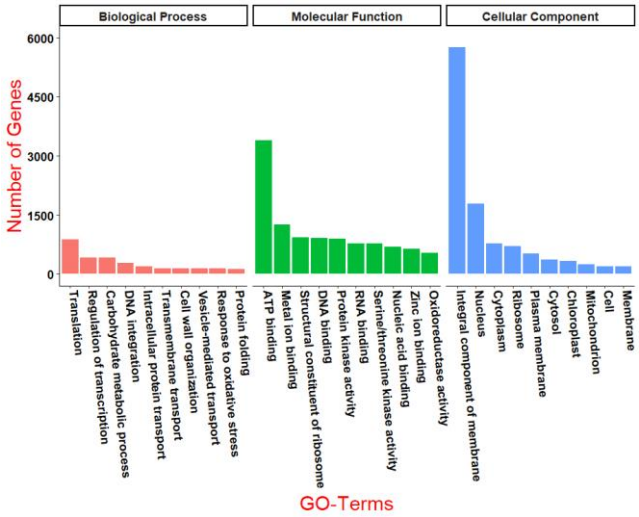
اختصارات: T/R (Trinity/ rnaSPAdes)، T/R-CAP (اعمال CAP3 روی Trinity/ rnaSPAdes)، T/R-CD (اعمال CD-HIT-EST روی Trinity/ rnaSPAdes) و T/R-CAP-CD (اعمال به‌ترتیب CAP3 و CD-HIT-EST روی Trinity/ rnaSPAdes). آماره Ave. len. of transcripts؛ متوسط طول رونوشت‌ها. آماره Contig N50؛ کوتاه‌ترین طول کانتیگ که ٥٠٪ از کل طول مونتاژ شده را نشان می‌دهد. آماره Q20؛ فراخوانی باز اشتباه برابر ۱ در ۱۰۰ است.



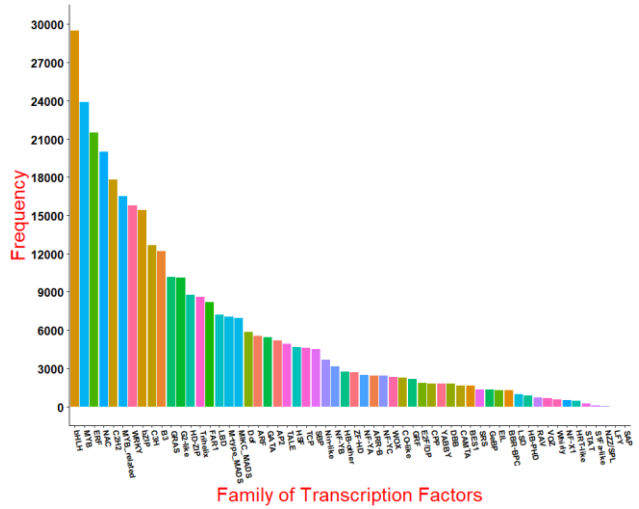
شکل ۱- نتایج ارزیابی BUSCO بر روی مونتاژ *de novo* ترانس‌کریپتوم سنبل‌الطیب با Viridiplantea_odb10.

گروه عملکردهای مولکولی (MF) و فرآیندهای زیستی (BP) به‌ترتیب با ۳۳/۹ و ۳۱/۷ درصد در رتبه دوم و سوم قرار گرفتند. در گروه (CC)، Integral component of membrane، Nucleous و Cytoplasm در اولیت بودند. در گروه (MF)، زیر گروه‌های ATP binding، Metal ion binding و Structural constituent of ribosome دارای بیشترین فراوانی ژن حاشیه‌نویسی شده بودند. در گروه (BP)، زیر گروه‌های Translation، Regulation of Transcription و Carbohydrate metabolic process بیشترین تعداد را به خود اختصاص دادند (شکل ۲؛ ۱۰ مورد برتر در هر زیر گروه انتولوژی).

طبق نتایج مرحله قبل (۳-۲)، مونتاژ Trinity پس از اعمال ترکیبی CAP3 و CD-HIT-EST تحت عنوان مونتاژ T- CAP-CD برای این بررسی انتخاب شد. در مرحله شناسایی فاکتورهای رونویسی، مونتاژ فوق در مقابل پایگاه داده PlantTFDB مورد جستجو قرار گرفت. طبق نتایج، تعداد ۱۲۱۱۸۱ عدد از unigeneهای کدکننده فاکتورهای رونویسی در قالب ۵۸ خانواده مختلف شناسایی و گروه‌بندی شدند. خانواده bHLH، MYB، ERF و NAC به‌ترتیب با فراوانی ۸/۶، ۷، ۶/۳ و ۵/۸ درصد، فراوان‌ترین خانواده‌های فاکتورهای رونویسی بودند. نادرترین خانواده مربوط به Arabidopsis STERILE APETALA (SAP) بود (شکل ۳).

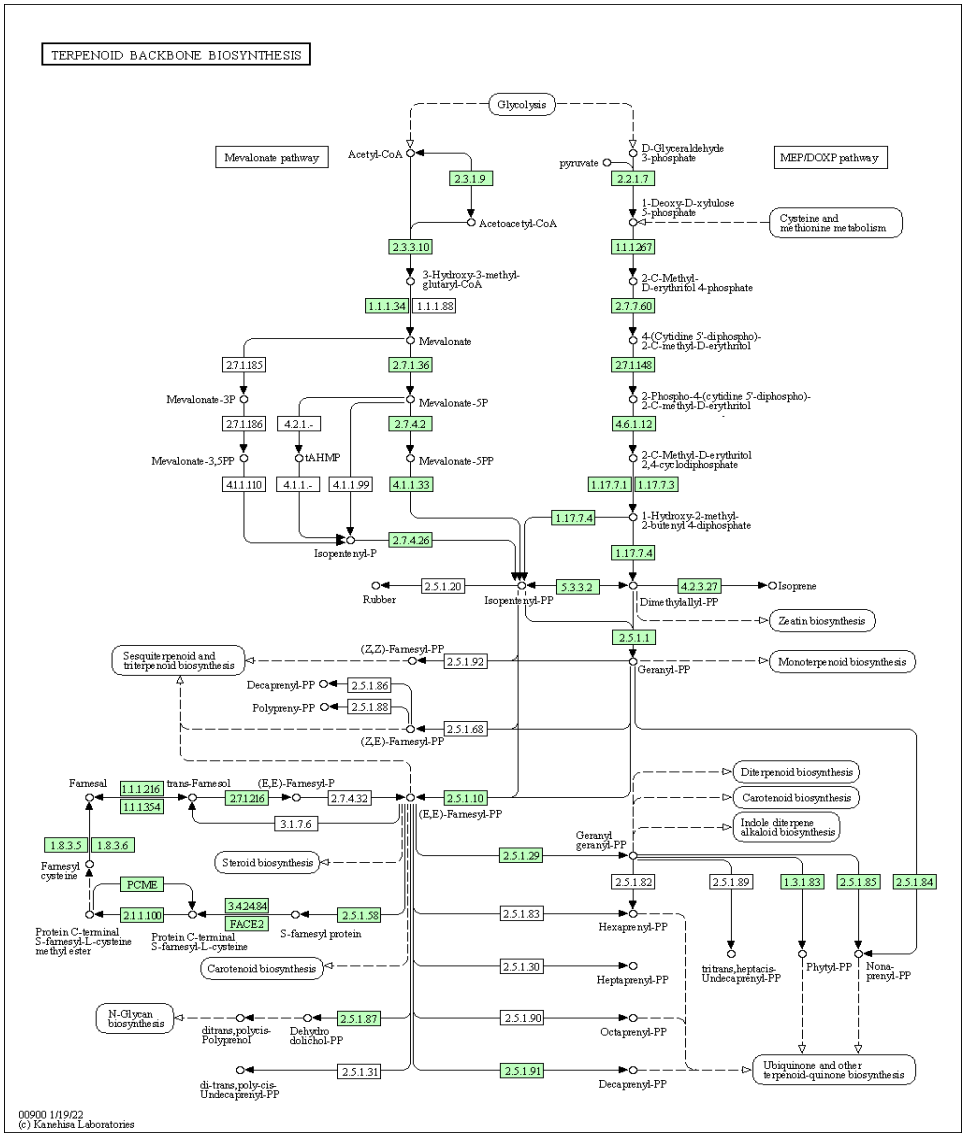


شکل ۲- ارزیابی انتولوژی ژن (GO) به‌منظور پیش‌بینی ژن‌های کدکننده پروتئین در مونتاژ *de novo* ترانس‌کریپتوم سنبل‌الطیب. فقط ۱۰ مورد برتر در هر زیر گروه انتولوژی نمایش داده شده است.

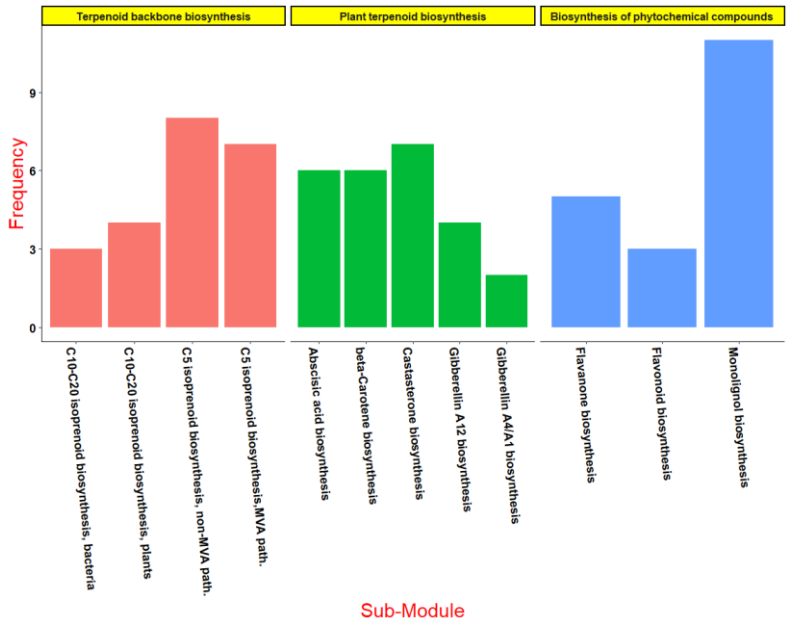


شکل ۳- توزیع فراوانی خانواده‌های فاکتورهای رونویسی بر مبنای جستجوی همولوژی بین مونتاژ T-CAP-CD و پایگاه داده PlantTFDB.(-max_target_seqs 1 Blastx (evalue 1e-5.

طبق نتایج GhosKOALA، تعداد ۲۶۵۳ (۳۵/۷ درصد) از unigeneهای ورودی، حاشیه‌نویسی KO را دریافت کردند. بیشترین فراوانی به‌ترتیب مربوط به خانواده‌های پروتئینی دخیل در پردازش اطلاعات ژنتیکی، متابولیسم کربوهیدرات و خانواده‌های پروتئینی مسئول پردازش سلولی و پیام‌رسانی بود. بر اساس نتایج KEGG MAPPER، رونوشت‌های موجود در مونتاژ T-CAP-CD در ۴۲۵ مسیر توزیع شدند (شکل ٤). در دسته متابولیسم ترپنوئیدها (map00900)، ۳۵ مورد از unigeneها به مسیر بیوسنتز ستون فقرات ترپنوئیدها تعلق داشت. همچنین، در مسیر بیوسنتز مونوترپنوئیدها (map00902)، سزکوئی-تری‌ترپنوئیدها (map00909) و دی‌ترپنوئیدها (map00904) به‌ترتیب ۱، ۷ و ۸ عدد از unigene حاشیه‌نویسی شد. ژن‌های کدکننده اسکوالن مونواکسیژناز (K0051)، فارنسیل-دی‌فسفات فارنسیل ترانسفراز (k00801)، آلفا-فارنسن سینتاز (K14173)، نرولدیول سینتاز (K14175)، جرماکرین D سینتاز (K15803)، بتا-آمیرین سینتاز (K15813) و فارنسول دهیدروژناز (K15891) در مسیر سزکوئی-تری‌ترپنوئیدها شناسایی شد. پس از مسیر بیوسنتز ترپنوئیدها، مسیر بیوسنتز فنیل پروپانوئیدها (map00940) با ۱۷ و مسیر بیوسنتز فلاوونوئیدها (map00941) با ١٤ عدد unigene، بیشترین میزان حاشیه‌نویسی KO را به خود اختصاص دادند.



شکل ٤- مسیر بیوسنتز ستون فقرات ترپنوئیدها (map00900) حاوی unigeneهای حاشیه‌نوسی‌شده (مستطیل‌های سبز) در مونتاژ T-CAP-CD با استفاده از ابزار KEGG Mapper.



شکل ۵ – توزیع فراوانی unigeneهای حاشیه‌نوسی‌شده در مونتاژ T-CAP-CD در زیر ماژول‌های مربوط به سه ماژول "بیوسنتز ستون فقرات ترپنوئید"، "بیوسنتز ترپنوئیدهای گیاهی" و "بیوسنتز ترکیبات فیتوشیمیایی" با استفاده از ابزار KEGG Mapper.

بررسی مسیر KEGG نشان داد که unigeneهای حاشیه‌نویسی‌شده به ۱۱۶ ماژول تعلق دارند. در ماژول "بیوسنتز ستون فقرات ترپنوئیدها"، زیر ماژول بیوسنتز ایزوپرنوئید C5 از مسیر غیر موالونات و در ماژول "بیوسنتز سایر متابولیت‌های ثانویه"، زیر ماژول بیوسنتز مونولیگنول به‌ترتیب با ۸ و ۱۱ زیر ماژول، در صدر بودند. کمترین تعداد زیر ماژول (۲) به زیر ماژول بیوسنتز جیبرلین GA4/GA1 تعلق داشت (شکل ۵).

بحث

ابزار متنوعی برای مونتاژ ترانس‌کریپتوم در دسترس است. همچنین، آماره‌های متعددی برای ارزیابی این ابزار توسعه یافته است. برخی از معمول‌ترین آماره‌ها شامل تعداد کانتیگ، متوسط طول کانتیگ، N50 (یا Nx) هستند (Behera et al., 2021). در خصوص آماره N50، هر چه مقدار بیشتری مشاهده شود، مبین این است که تعداد بیشتری از خوانش‌ها برای تشکیل کانتیگ‌های بزرگ‌تر، همپوشانی کرده و کیفیت بهتر مونتاژ را تأیید می‌کند (Biswal et al. 2021). باید متذکر شد که N50، تداوم یا پیوستگی کانتیگ را ارزیابی می‌کند و معیاری برای اندازه‌گیری دقت نیست (Li et al. 2014).

ابزار CD-HIT یک برنامه پرکاربرد برای خوشه‌بندی توالی‌های زیستی به‌منظور کاهش تزاید توالی و بهبود عملکرد سایر تحلیل‌های توالی است (Fu et al. 2012). کاهش در میزان رونوشت‌ها پس از اعمال CD-HIT-EST به‌علت خوشه‌بندی مونتاژهای اولیه و حذف تزاید در نیل به مونتاژ نهایی گزارش شده است (Evangelistella et al. 2017).

در مطالعه حاضر، استفاده از CAP3 باعث افزایش میزان N50 و متوسط طول رونوشت شده که ناشی از ادغام کانتیگ‌های همپوشان با یکدیگر در حین تشکیل scafoldها است (Hoang et al. 2018). در مطالعات مختلف، استفاده از CAP3 منجر به کاهش قابل توجه تعداد رونوشت‌های حاصل از مونتاژ شده که نتایج مطالعه حاضر را تأیید می‌کند (Duan et al. 2012; Yang and Smith 2013).

ابزار BUSCO برای کمی‌سازی "کامل‌بودن" مونتاژ ایده‌آل است؛ زیرا یافتن ژن‌ها در ژنوم و یافتن آن‌ها تنها در تک‌نسخه، از نظر تکاملی مهم است (Waterhouse et al. 2013). اصولاً، سطوح پایین تزاید، مبین کیفیت بهتر مونتاژ بوده و افزایش این مقدار نشان‌دهنده مشکلات فنی در حین مونتاژ است (Manni et al. 2021). نتایج مطالعه حاضر نشان داد که اعمال ابزار CD-HIT-EST می‌تواند به‌طور مؤثر، میزان بالایی از تزاید (Duplicated

(BUSCO(D) را در مونتاژ حاصل از ابزار Trinity کاهش دهد. پارامتر مهم دیگر در نتایج این ابزار، Missing BUSCO (M)، یا "مفقود" است. اعمال ترکیبی دو ابزار CAP3 و CD-HIT-EST و همچنین اعمال منفرد CD-HIT-EST بر روی مونتاژ Trinity از لحاظ تعداد بیشتر تک‌نسخه (Single-copy(S)) و تعداد کمتر تزاید (۲٤۰)، کیفیت بهتری در بین سایرین نشان می‌دهد. در یک مطالعه، اعمال CAP3 بر روی مونتاژ Trinity منجر به کاهش تزاید از ۳/۱ به ۲/٦ شده که در نتایج مطالعه حاضر نیز نمایان است (Yang and Smith 2013). همچنین، به‌منظور حذف بیشتر تزاید، اعمال CD-HIT-EST پس از CAP3 نیز قبلاً گزارش شده است (Feldmesser et al. 2014).

برای مرحله انتولوژی ژن از ابزار آنلاین Hayai-Annotation Plants استفاده شد. در بررسی انتولوژی ژن مشخص شد که بیشترین فراوانی تخصیص GO مربوط به گروه عملکردهای مولکولی بوده و در این گروه، زیر گروه‌های عمده شامل انواع پروتئین‌های اتصالی (به ATP، یون‌های فلزی و DNA) قرار دارند. تعداد ۳۱۳۷۹ عدد از unigeneهای مونتاژ T-CAP-CD، بدون حاشیه‌نویسی بوده که می‌تواند به دلیل انتخاب نوع پایگاه داده در مرحله بررسی همولوژی، انتخاب نوع ابزار حاشیه‌نویسی و یا کوتاهی بیش از حد unigeneها و عدم پوشش با دمین‌های پروتئینی محافظت شده (Zhou and Zhu 2020) باشد.

پس از بررسی تعیین همولوژی مشخص شد که بیشترین unigeneها، کدکننده فاکتورهای رونویسی از خانواده bHLH، MYB، ERF و NAC بودند. نقش فاکتورهای رونویسی bHLH، ERF و MYB در تنظیم متابولیسم ثانویه گیاهی و نمو سلولی اختصاصی (Chezem and Clay 2016) و نقش‌های نو ظهور خانواده فاکتورهای رونویسی NAC در گیاهان دارویی (Kumar et al. 2021) اثبات شده است. در مسیر بیوسنتز ترپنوئیدهای گیاهی، موارد فراونی از نقش فاکتورهای رونویسی در مرور منابع می‌توان یافت. برای مثال، فاکتور رونویسی bHLH (MYC2) در آرابیدوپسیس به‌طور مستقیم به پروموتر ژن‌های کدکننده سزکوئی‌ترپن‌ها متصل شده و رهاسازی سزکوئی‌ترپن‌های فرار را افزایش داده است (Hong et al. 2012). بیش‌بیان *SmMYB9b* در گیاه دارویی مریم گلی منجر به افزایش بیوسنتز ترپنوئید تانشیونین شده است (Zhang et al. 2017). فاکتور رونویسی CitERF71 ژن ترپن سنتاز *CitTPS16* را فعال می‌کند که در سنتز ژرانیول در میوه نارنج شیرین نقش دارد (Li et al. 2017). همچنین، بیش بیان *AaNAC1* در گیاه دارویی آرتمیزیا باعث افزای تولید ترپنوئیدهای آرتمزینین و دی‌هیدرو آرتمیزینیک اسید به‌ترتیب به میزان ۷۹ و ۱۵۰ درصد شده است (Lv et al. 2016). نظر به تولید مشتقات سزکوئی‌ترپن‌ها از قبیل والرنیک اسید در گیاه دارویی سنبل‌الطیب، استفاده از این نتایج می‌تواند جهت یافتن ارتولوگ‌های مهم فاکتورهای رونویسی متعدد دخیل در مسیر بیوسنتز ترپنوئیدها و استفاده از آن‌ها در برنامه‌های اصلاحی در این گیاه دارویی بسیار ارزشمند باشد.

در بررسی نتایج مسیرهای متابولیکی KEGG مشخص شد که مونتاژ T-CAP-CD از لحاظ حاشیه‌نویسی ژن‌های کدکننده مسیر بیوسنتز ستون فقرات ترپنوئیدها، غنی‌تر از سایر مسیرهای متابولیکی بوده است. با توجه به شکل ٤، بخش عمده ژن‌های کدکننده به دو مسیر مهم موالونات و MEP/DOXP تعلق دارد. با توجه به اهمیت پیش‌ساز ایزوپنتنیل پیروفسفات که به نوعی در مرکز انشعابات مسیر ترپنوئیدها قرار دارد، سه ژن کدکننده آنزیم‌های تولیدکننده این سوبسترا به نام‌های ایزوپنتنیل فسفات کیناز (ipk؛ EC: 2.7.4.26)، هیدروکسی متیل بوتنیل دی‌فسفات ریداکتاز (isPH؛ EC: 1.17.7.4) و ایزوپنتنیل دی‌فسفات دلتا ایزومراز (idi؛ EC: 5.3.3.2) شناسایی شد که می‌تواند از اهمیت بالایی برخوردار باشد. آنزیم آخر منجر به تولید دی‌متیل آلیل پیروفسفات نیز می‌شود (شکل ٤). طبق نتایج حاشیه‌نویسی KEGG، آنزیم فارنسیل دی‌فسفات سینتاز (FDPS؛ EC: 2.5.1.1, 2.5.1.10) نیز در مونتاژ T-CAP-CD شناسایی شد که با ترکیب ژرانیل دی‌فسفات (GPP) و ایزوپنتنیل دی‌فسفات، فرنسیل دی‌فسفات را به‌عنوان پیش‌ساز مسیر بیوسنتز سزکوئی/تری-ترپنوئیدها تولید می‌کند.

علاوه بر این، بررسی مسیر بیوسنتز سزکوئی/تری- ترپنوئیدها نشان داد که ۷ ژن کدکننده برای آنزیم‌های این مسیر حاشیه‌نویسی شده است. در این بین، جرماکرین D سینتاز (K15803) یا ارتولوگ آن در سنبل‌الطیب، واسطه تشکیل جرماکرین C و D با استفاده از پیش‌ساز فارنسیل پیروفسفات است (Pyle et al.

2012). گزینه مهم دیگر در این مسیر، آنزیم آلفا- فرنسن سینتاز بوده که به عنوان یک پروتئین حاوی دمین ترپن سینتاز می‌تواند در برنامه‌های کلونینگ و به‌منظور درک بیشتر مسیر بیوسنتز سزکوئی‌ترپن‌ها در گیاه دارویی سنبل‌الطیب مفید واقع شود. علاوه بر این، شناسایی آنزیم فارنسیل دی‌فسفات فارنسیل ترانسفراز (FDFT1؛ EC: 2.5.1.21) در مسیر سزکوئی/تری-ترپنوئیدها نیز قابل توجه است؛ زیرا به نوعی منجر به سوق پیش‌ساز کلیدی فارنسیل پیروفسفات به سمت تولید اسکوالن و مسیر رقیب تری‌ترپنوئیدها می‌شود. این یافته از این جنبه می‌تواند برای برنامه‌ریزی‌های تحقیقاتی آینده در گیاه سنبل‌الطیب مهم باشد که با استفاده از فنون خاموشی ژن کدکننده آنزیم FDFT1 می‌توان جریان بیشتری از مسیر متابولیکی را به سمت تولید سزکوئی‌ترپنوئید والرنیک اسید سوق داد. از طرف دیگر، آنزیم اسکوالن سینتاز به‌عنوان آنزیم کلیدی مسیر رقابتی دیگر یعنی؛ بیوسنتز استرول، در حاشیه‌نویسی KEGG بر روی مونتاژ ترانس‌کریپتوم سنبل‌الطیب شناسایی شد.

نتیجه‌گیری کلی

روش‌های *de novo* برای مونتاژ و بازسازی خوانش‌های کوتاه RNA بدون ژنوم مرجع به‌کار می‌روند. هر یک از این روش‌ها، مزایا و معایب خاص خود را داشته و نمی‌توان یک روش برتر از لحاظ کلیه معیارها را معرفی نمود. به‌همین علت و نظر به وفور ابزارهای جانبی برای پیش‌پردازش، استانداردسازی و استخراج نتایج مطلوب از مونتاژ مورد نظر، مطالعه حاضر با هدف تعیین دستورزی بهینه روی مونتاژ *de novo* گیاه دارویی سنبل‌الطیب انجام شد. توسعه این قبیل بررسی‌ها جهت بهینه‌سازی روش‌های بدون مرجع به‌خصوص برای گیاهان دارویی که عموماً در دسته گیاهان غیر مدل هستند، اهمیت بسیاری دارد. طبق نتایج، اعمال ترکیبی دو ابزار جانبی CAP3 و CD-HIT-EST برای نیل به یک مونتاژ ترانس‌کریپتوم بهینه در سنبل‌الطیب توصیه می‌شود. به‌خصوص در گیاهان دارویی کمتر شناخته‌شده، استفاده از ابزار بیوانفورماتیکی در جهت روشن‌سازی نقاط مبهم و ناشناخته در مسیرهای متابولیکی بسیار مهم است. تمرکز مطالعه حاضر، بررسی انتولوژی ژنی، شناسایی خانواده‌های فاکتورهای رونویسی و یافتن ارتولوگ‌های ژنی مهم در مسیر بیوسنتز ترپنوئیدها؛ به‌ویژه مسیر تولید والرنیک اسید (سزکوئی‌ترپنوئید) بود. در خصوص انتولوژی ژن، ابزار آنلاین رایگان و کاملاً اختصاصی به نام Hayai-Annotation Plants برای گیاهان با پایگاه جامع و سرعت عمل بسیار بالا مورد استفاده قرار گرفت که برای سایر مطالعات مشابه نیز توصیه می‌شود. در زمینه شناسایی فاکتورهای رونویسی، خانواده‌های عمده از قبیل bHLH، NAC و MYB دسته‌بندی شد که در تنظیم مسیر ترپنوئیدی نقش بسیار مهمی در گیاهان دیگر داشته‌اند. همچنین در بررسی مسیر KEGG، تعداد ۷ ارتولوگ در مسیر سزکوئی‌ترپنوئیدها شناسایی شد که به‌طور ویژه می‌تواند برای مطالعات آتی در این گیاه اعم از انتقال ژن بیگانه، بیش‌بیان و حتی در خصوص اسکوالن سینتاز، خاموشی ژن و جلوگیری از ورود جریان متابولیکی به مسیرهای رقیب در بیوسنتز والرنیک اسید مهم باشد. به‌طور کلی در این مطالعه، نتایج راهگشایی برای درک بهتر مسیر بیوسنتز ترکیبات ترپنوئیدی در گیاه دارویی ارزشمند اما کمتر شناخته‌شده سنبل‌الطیب ارائه شده است. به‌عنوان خط سیر تکمیلی، شناسایی فاکتورهای رونویسی ویژه و دخیل در مسیر بیوسنتز ترکیبات ترپنوئیدی از طریق بررسی شبکه‌های هم‌بیان توصیه می‌شود.

منابع

Andrews S (2010) Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. URL: https://www. bioinformatics. babraham. ac. uk/projects/fastqc.

Angeloni F, Wagemaker C, Jetten M, Op den Camp H, Janssen-Megens E, FRANCOIJS KJ, Stunnenberg H, Ouborg N (2011) *De novo* transcriptome characterization and development of genomic tools for Scabiosa *columbaria* L. using next-generation sequencing techniques. Molecular Ecology Resources 11:662-674.

Behera S, Voshall A, Moriyama E (2021) Plant transcriptome assembly: review and benchmarking. Exon Publications 109-130.

Biswal B, Jena B, Giri AK, Acharya L (2021) De novo transcriptome and tissue specific expression analysis of genes associated with biosynthesis of secondary metabolites in Operculina turpethum (L.). Scientific reports 11:1-15.
Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.
Bushmanova E, Antipov D, Lapidus A, Przhibelskiy AD (2018) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. bioRxiv. 420208. Advance online publication. Retrieved September 18:2018.
Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD (2016) rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. Bioinformatics, 32:2210-2212.
Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer CL, Huang X (2015) Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome biology 16:1-10.
Chen G, Yin K, Wang C, Shi T (2011) De novo transcriptome assembly of RNA-Seq reads with different strategies. Science China Life Sciences 54:1129-1133.
Chen X, Zhang Y, Yan H, Niu M, Xiong Y, Zhang X, Li Y, da Silva JAT, Ma G (2023) Cloning and functional analysis of 1-deoxy-d-xylulose-5-phosphate synthase (DXS) in Santalum album L. Gene 851.
Cherukupalli N, Divate M, Mittapelli SR, Khareedu VR, Vudem DR (2016) De novo assembly of leaf transcriptome in the medicinal plant Andrographis paniculata. Frontiers in plant science 7:1203.
Chezem WR, Clay NK (2016) Regulation of plant secondary metabolism and associated specialized cell development by MYBs and bHLHs. Phytochemistry 131:26-43.
Christianson DW (2018) Correction to structural and chemical biology of terpenoid cyclases. Chemical Reviews 1:11795.
Duan J, Xia C, Zhao G, Jia J, Kong X (2012) Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. BMC genomics 13:1-12.
Evangelistella C, Valentini A, Ludovisi R, Firrincieli A, Fabbrini F, Scalabrin S, Cattonaro F, Morgante M, Mugnozza GS, Keurentjes JJ (2017) De novo assembly, functional annotation, and analysis of the giant reed (Arundo donax L.) leaf transcriptome provide tools for the development of a biofuel feedstock. Biotechnology for biofuels 10:1-24.
Feldmesser E, Rosenwasser S, Vardi A, Ben-Dor S (2014) Improving transcriptome construction in non-model organisms: integrating manual and automated gene definition in Emiliania huxleyi. BMC genomics 15:1-16.
Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150-3152.
Geniza M, Jaiswal P (2017) Tools for building de novo transcriptome assembly. Current Plant Biology 11:41-45.
Ghelfi A, Shirasawa K, Hirakawa H, Isobe S (2019) Hayai-Annotation Plants: an ultra-fast and comprehensive functional gene annotation system in plants. Bioinformatics 35:4427-4429.
Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature biotechnology 2.
Guo J, Huang Z, Sun J, Cui X, Liu Y (2021) Research progress and future development trends in medicinal plant transcriptomics. Frontiers in plant science 12.
Guo J, Sun B, He H, Zhang Y, Tian H, Wang B (2021) Current understanding of bHLH transcription factors in plant abiotic stress tolerance. International Journal of Molecular Sciences 22:4921.
Gutensohn M, Orlova I, Nguyen TT, Davidovich-Rikanati R, Ferruzzi MG, Sitrit Y, Lewinsohn E, Pichersky E, Dudareva N (2013) Cytosolic monoterpene biosynthesis is supported by plastid-generated geranyl diphosphate substrate in transgenic tomato fruits. The Plant Journal, 75:351-363.
Hemmerlin A, Hoeffler J.-F, Meyer O, Tritsch D, Kagan I A, Grosdemange-Billiard C, Rohmer M, Bach TJ (2003) Cross-talk between the cytosolic mevalonate and the plastidial methylerythritol phosphate pathways in tobacco bright yellow-2 cells. Journal of Biological Chemistry 278:26666-26679.
Henschel R, Lieber M, Wu L-S, Nista PM, Haas BJ, LeDuc RD (2012) Trinity RNA-Seq assembler performance optimization. Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond.
Hoang NV, Furtado A, Thirugnanasambandam PP, Botha FC, Henry RJ (2018) De novo assembly and characterizing of the culm-derived meta-transcriptome from the polyploid sugarcane genome based on coding transcripts. Heliyon 4:e00583.
Hölzer M, Marz M (2019) De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience 8:giz039.
Hong GJ, Xue XY, Mao YB, Wang LJ, Chen XY (2012) Arabidopsis MYC2 interacts with DELLA proteins in regulating sesquiterpene synthase gene expression. The Plant cell 24:2635-2648.
Kanehisa M, Sato Y (2020) KEGG Mapper for inferring cellular functions from protein sequences. Protein science 29:28-35.
Kanehisa M, Sato Y, Kawashima M (2022) KEGG mapping tools for uncovering hidden features in biological data. Protein science 31:47-53.
Kanehisa M, Sato Y, Morishima K (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. Journal of molecular biology 428:726-73.
Kannan S, Hui J, Mazooji K, Pachter L, Tse D (2016) Shannon: an information-optimal de novo RNA-Seq assembler. bioRxiv 039230. In.
Kumar R, Das S, Mishra M, Choudhury DR, Sharma K, Kumari A, Singh R (2021) Emerging roles of NAC

nonmodel plant species1. American Journal of Botany 99:257-266.
Tripathi S, Jadaun JS, Chandra M, Sangwan NS (2016) Medicinal plant transcriptomes: the new gateways for accelerated understanding of plant secondary metabolism. Plant Genetic Resources 14:256-269.
Voshall A, Moriyama EN (2018) Next-generation transcriptome assembly: strategies and performance analysis. Bioinformatics in the era of post genomics and big data 15-36.
Wang Q, Quan S, Xiao H (2019) Towards efficient terpenoid biosynthesis: manipulating IPP and DMAPP supply. Bioresources and Bioprocessing 6:1-13.
Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic acids research 41:D358-D365.
Wong J, d'Espaux L, Dev I, van der Horst C, Keasling J (2018) De novo synthesis of the sedative valerenic acid in Saccharomyces cerevisiae. Metabolic engineering 47:94-101.
Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660-1666.
Xu ZS, Chen M, Li LC, Ma YZ (2008) Functions of the ERF transcription factor family in plants. Botany 86:969-977.
Yang D, Du X, Liang X, Han R, Liang Z, Liu Y, Liu F, Zhao J (2012) Different roles of the mevalonate and methylerythritol phosphate pathways in cell growth and tanshinone production of Salvia miltiorrhiza hairy roots. PloS one 7:e46797.
Yang Y, Smith SA (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC genomics 14:1-11.
Yeo YS, Nybo SE, Chittiboyina AG, Weerasooriya AD, Wang YH, Góngora-Castillo E, Vaillancourt B, Buell CR, DellaPenna D, Celiz MD (2013) Functional identification of valerena-1, 10-diene synthase, a terpene synthase catalyzing a unique chemical cascade in the biosynthesis of biologically active sesquiterpenes in Valeriana officinalis. Journal of Biological Chemistry 288:3163-3173.
Younessi-Hamzekhanlu M, Ozturk M, Jafarpour P, Mahna N (2022) Exploitation of next generation sequencing technologies for unraveling metabolic pathways in medicinal plants: A concise review. Industrial Crops and Products 178:114669.
Yuan X, Wang H, Cai J, Li D, Song F (2019) NAC transcription factors in plant immunity. Phytopathology Research 1:1-13.
Zamini A, Mokhtari A, Tansaz M, Zarei M (2016) Callus induction and plant regeneration of Valeriana officinalis are affected by different leaf explants and various concentrations of plant growth regulators. BioTechnologia. Journal of Biotechnology Computational Biology and Bionanotechnology 97.
Zdobnov EM, Kuznetsov D, Tegenfeldt F, Manni M, Berkeley M, Kriventseva EV (2021) OrthoDB in 2020:

transcription factor in medicinal plants: progress and prospects. 3 Biotech 11:1-14.
Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart, R, Dewey CN (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome biology 15:1-21.
Li X, Xu Y, Shen S, Yin X, Klee H, Zhang B, Chen K (2017) Transcription factor CitERF71 activates the terpene synthase gene CitTPS16 involved in the synthesis of E-geraniol in sweet orange fruit. Journal of Experimental Botany 68:4929-4938.
Liu J, Li G, Chang Z, Yu T, Liu B, McMullen R, Chen P, Huang X (2016) BinPacker: packing-based de novo transcriptome assembly from RNA-seq data. PLoS computational biology 12:e1004772.
Lv Z, Wang S, Zhang F, Chen L, Hao X, Pan Q, Fu X, Li L, Sun X, Tang K (2016) Overexpression of a novel NAC domain-containing transcription factor gene (AaNAC1) enhances the content of artemisinin and increases tolerance to drought and Botrytis cinerea in Artemisia annua. Plant and Cell Physiology 57:1961-1971.
Manni M, Berkeley MR, Seppey M, Zdobnov EM (2021) BUSCO: assessing genomic data quality and beyond. Current Protocols 1:e323.
Nandhini S, Narayanan KB, Ilango K (2018) Valeriana officinalis: A review of its traditional uses, phytochemistry and pharmacology. Asian J Pharm Clin Res 11:36-41.
Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526-528.
Park YJ, Li X, Noh SJ, Kim JK, Lim SS, Park NI, Kim S, Kim YB, Kim YO, Lee SW (2016) Transcriptome and metabolome analysis in shoot and root of Valeriana fauriei. BMC genomics 17:1-16.
Pechous SW, Whitaker BD (2004) Cloning and functional expression of an (E, E)- α -farnesene synthase cDNA from peel tissue of apple fruit. Planta 219:84-94.
Peng Y, Leung HC, Yiu SM, Lv MJ, Zhu XG, Chin FY (2013) IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics 29:i326-i334.
Pyle BW, Tran HT, Pickel B, Haslam TM, Gao Z, MacNevin G, Vederas JC, Kim SU, Ro DK (2012) Enzymatic synthesis of valerena-4, 7 (11)-diene by a unique sesquiterpene synthase from the valerian plant (Valeriana officinalis). The FEBS Journal 279:3136-3146.
Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ (2010) De novo assembly and analysis of RNA-seq data. Nature methods7:909-912.
Shuang Z, Chenshu W (2020) Deep sequencing and transcriptome analyses to identify genes involved in iridoid biosynthesis in the medicinal plant Valeriana jatamansi Jones. Notulae Botanicae Horti Agrobotanici Cluj-Napoca 48:189-199.
Song L, Florea L (2015) Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience, 4:s13742-13015-10089-y.
Strickler SR, Bombarely A, Mueller LA (2012) Designing a transcriptome next-generation sequencing project for a

evolutionary and functional annotations of orthologs. Nucleic acids research 49:D389-D393.
Zhang J, Zhou L, Zheng X, Zhang J, Yang L, Tan R, Zhao, S (2017) Overexpression of SmMYB9b enhances tanshinone concentration in Salvia miltiorrhiza hairy roots. Plant Cell Reports 36:1297-1309.
Zhang L, Jing F, Li F, Li M, Wang Y, Wang G, Sun X, Tang K (2009) Development of transgenic Artemisia annua (Chinese wormwood) plants with an enhanced content of artemisinin, an effective anti-malarial drug, by hairpin-RNA-mediated gene silencing. *Biotechnology and* Applied Biochemistry 52:199-207.
Zhou GL, Zhu P (2020) De novo transcriptome sequencing of Rhododendron molle and identification of genes involved in the biosynthesis of secondary metabolites. BMC plant biology 20:1-19.