مقایسه برنامه‌های سرهمبندی و آنالیز هستی‌شناسی با استفاده از داده‌های حاصل از توالی یابی ترنسکریپتوم زرین‌گیاه (*Dracocephalum kotschyi* Boiss.)

عبدالناصر پورصلواتی[۱]، امین ابراهیمی[۲] و سجاد رشیدی‌منفرد[۱*]

[۱] تهران، دانشگاه تربیت مدرس، گروه بیوتکنولوژی کشاورزی

[۲] شاهرود، دانشگاه صنعتی شاهرود، گروه زراعت و اصلاح نباتات

تاریخ دریافت: ۹۶/۱۲/۷ تاریخ پذیرش: ۹۷/۱۰/۲۵

چکیده

با پیشرفتهای سریع در تکنولوژی توالی‌یابی نسل جدید امروزه این تکنیک به ابزاری قدرتمند و کم‌هزینه برای مطالعات در سطح ترنسکریپتوم تبدیل شده است. سرهمبندی داده‌های حاصل از توالی‌یابی نسل جدید، به صورت *de novo* باعث شکل‌گیری مسیری نوین در مطالعه و شناخت گونه‌های فاقد ژنوم مرجع گردیده است. با گسترش این تکنولوژی و افزایش روز افزون نرم‌افزارهای سرهمبندی، انتخاب مسیر و گزینش نرم‌افزار برتر برای سرهمبندی داده‌های حاصل از توالی‌یابی ترنسکریپتوم به عنوان چالشی برای زیست‌شناسان در این زمینه به شمار می‌آید. در این پژوهش برای اولین بار داده‌های حاصل از توالی‌یابی ترنسکریپتوم زرین‌گیاه با استفاده از نرم‌افزارهای Oases-velvet، SOAPdenovo-Trans، Trans-ABySS و Trinity به دو صورت مختلف با استفاده از متغیر K=25 و K=32 به منظور دستیابی به مسیر مناسب و نرم‌افزار برتر در این زمینه مورد ارزیابی و آنالیز قرار گرفت. نتایج حاصل از سرهمبندی براساس معیارهای متعددی مقایسه شده که گویای برتری Trinity و Trans-ABySS می‌باشد، در پایان خروجی حاصل از بهترین سرهمبندی به منظور بررسی فراوانی ایزوفرمهای مختلف و آنالیز هستی‌شناسی (Gene Ontology) مورد ارزیابی قرار گرفت. باتوجه به خواص دارویی و مقادیر بالای متابولیتهای ثانویه در این گیاه؛ بیشترین فراوانی مشاهده شده در بخش فرآیندهای زیستی، مربوط به فعالیتهای متابولیتی (Metabolic Process) بود.

واژه های کلیدی: Gene Ontology, Trans-ABySS, Oases-velvet, SOAPdenovo-Trans, Trinity

* نویسنده مسئول، تلفن: ۰۲۱۴۸۲۹۲۳۵۷ ، پست الکترونیکی: rashidims@modares.ac.ir

مقدمه

در سالهای اخیر آنالیز ترنسکریپتوم به عنوان یکی از مراحل بنیادی در مطالعات زیستی مطرح شده که به‌این منظور روشهای متعددی از جمله؛ نورترن بلات (Northern blot)، واکنش زنجیره ای پلیمراز رونوشت برداری معکوس (RT-PCR)، ریزآرایه‌ها (Microarray) و توالی‌یابی به روشهای سنتی را می توان نام برد (۳۰) و باتوجه به پیشرفتهای سریع در توالی‌یابی نسل جدید (Next-generation sequencing: NGS)، این مورد به ابزاری قدرتمند در آنالیز ترنسکریپتوم تبدیل شده‌است (۳۰ و ۴۶) امروزه با استفاده از تکنولوژی NGS، امکان توالی‌یابی و شناسایی طیف بسیار گسترده‌ای از ژنها در مدتی کوتاه و در میان ژنوم پیچیده و عظیم گیاهی فراهم‌آمده است (۴۰). ازسوی دیگر نرم‌افزارهای متعددی برای سرهمبندی خوانشهای خام (Raw Read) حاصل از توالی‌یابی به وسیله همردیفی خوانشها بر روی ژنوم مرجع (Reference assembly) معرفی شده که عمل شناسایی ژن را در محیط نرم‌افزاری انجام می‌دهند (۱۱)، اما ژنوم‌مرجع برای بسیاری از گونه‌ها ازجمله گونه مورد

مطالعه در این پژوهش (زرین‌گیاه) در دسترس نیست. با این وجود و با بهره‌گیری از نرم‌افزارهای به روز و معرفی تکنیکهای NGS امکان مطالعه در سطوح «امیک» (Omics) برای گونه‌های فاقد نقشه ژنومی نیز فراهم آمده (۷)، که در این حالت سرهمبندی *de novo*، این امکان را ایجاد کرده تا توالی کاملی از ترنسکریپتوم موجود بازسازی و ژنهای بیان شده در یک بافت خاص شناسایی، مقدار سنجی و دسته‌بندی گردد (۴۶).

اولین تلاشها در مطالعات نوین ترنسکریپتوم با استفاده از داده‌های RNA-Seq در سال ۲۰۰۹ به وسیله ونگ و همکاران تحت عنوان ابزاری انقلابی در مطالعه ترنسکریپتوم مطرح شد (۴۰). نرم‌افزارهای مربوط در این‌روش، برای سرهمبندی از دو الگوریتم مختلف De Bruijn graph (۴۵) و Overlap layout-consensus (۱۳) پیروی می‌کنند. در نرم‌افزارهای نسل جدید از الگوریتم De Bruijn graph استفاده شده که در این روش سرهمبندی از طریق شکستن خوانش خام اولیه، به توالیهای کوچک‌تر که *K*-mer نامیده می‌شوند انجام شده و یافتن همپوشانی میان *K*-mer ها صورت می‌گیرد (۱۴). نرم‌افزارهای مورد نظر برای این پژوهش شامل Trinity (v2.4.0) (۱۴)، SOAPdenovo-Trans (v1.03) (۴۲)، Oases-Velvet (v0.2.08) (۳۷) و Trans-ABySS (v1.5.5) (۳۴) بوده که همگی از الگوریتم De Bruijn graph پیروی کرده و با توجه به گستردگی این دست نرم‌افزارها و پارامترهای مربوط؛ انتخاب نرم‌افزار برتر و روش بهینه برای سرهمبندی داده‌ها امری ضروری است. در سال ۲۰۱۲ مقایسه‌ای میان سه ابزار مختلف سرهمبندی یعنی SOAPdenovo ، Oases و Trinity بر روی خوانشهای حاصل از سیب زمینی شیرین به منظور یافتن پوشش ژنومی قوی‌تر انجام گرفت (۳۸). همچنین بر روی خوانشهای مگس سرکه موجود در دیتابیس، مقایسه میان نرم‌افزارهای سرهمبندی صورت گرفته و نتایج حاصل از سرهمبندی با ژنهای شناخته شده این موجود مقایسه گردید (۴۶). در تحقیقی دیگر برای سرهم بندی داده‌های حاصل از توالی‌یابی با سیستم 454 برای اولین‌بار از ابزارهای مبتنی بر الگوریتم De Bruijn graph استفاده شد (۳۳). با این همه در تحقیقات انجام شده برای مقایسه نرم‌افزارها به طور معمول از داده‌های موجودات مدل استفاده شده و این در حالی‌است که این پژوهش خوانشهای زرین‌گیاه را به صورت *de novo* و بر اساس مقادیر مختلف *K* برای سرهمبندی درنظر گرفته است. بایستی به این نکته نیز توجه نمود که الگوریتمهای این برنامه‌ها درحال به روزرسانی است و مقایسه آخرین ورژن برنامه‌ها برای به دست آوردن نتیجه مطلوب ضروری می‌نماید.

زرین‌گیاه با نام علمی *Dracocephalum kotschyi* یکی از گونه‌های ایندمیک ایران است که در شمال، غرب و مرکز ایران یافت شده (۳۲) و با نام بادرنجبویه‌دنایی نیز شناخته می‌شود (۳۱). با توجه به تعداد کم و خطر انقراض بالقوه، ضرورت حفاظت، اصلاح و اهلی کردن این گیاه بیش از هر زمانی احساس می‌شود (۱۲ و ۲۰). زرین‌گیاه در طب‌سنتی و داروسازی نیز مورد توجه بوده و تحقیقات متعددی روی آن صورت گرفته‌است (۴، ۲۱، ۲۲، ۳۶ و ۴۴). با این حال، میزان پژوهشهای بیوانفورماتیکی صورت گرفته در زمینه گیاهان ارزشمند دارویی بسیار اندک بوده و به گواه بانک اطلاعاتی NCBI به جز یک توالی rRNA و سه توالی کلروپلاستی، هیچ‌گونه فعالیت و پژوهشی در راستای آنالیزهای مولکولی بر روی زرین‌گیاه صورت نگرفته است (۴۸).

اکنون با استفاده از روشهای نوین با کارآیی بالا و بهره‌گیری از نرم‌افزارهای بیوانفورماتیکی برای تبدیل داده‌های‌خام به اطلاعات مفید و آنالیزهای In silico (۳۷) می توان اقدام به شناسایی هرچه بهتر این گیاه نمود و برای دست‌یابی مؤثر و کارآمد به این هدف، اولین گام انتخاب ابزار و نرم‌افزار مناسب با کارآیی و دقت بالاست. هدف از

این پژوهش بررسی مهمترین نرم‌افزارهای موجود در این زمینه و دستیابی به مسیری قابل اعتماد و کارآمد برای سرهمبندی خوانشها به صورت *de novo* جهت آنالیزهای پایین‌دست می‌باشد. امید است نتایج حاصل از این پژوهش بتواند در موارد مشابه راه‌گشای محققین در تحقیقات آتی قرار گیرد.

مواد و روشها

جمع‌آوری نمونه‌های گیاهی و استخراج mRNA: در این پژوهش، از زرین‌گیاه موجود در ارتفاعات رشته‌کوههای زاگرس در استان لرستان، شهرستان الشتر (با مشخصات، ارتفاع از سطح دریا: ۳۵۸۵ متر، عرض جغرافیایی: ۳۳,۹۵۵۹۹۶ و طول جغرافیایی: ۴۸,۳۲۰۰۱۱) نمونه‌های برگ جمع‌آوری و در ازت مایع به آزمایشگاه منتقل گردید. مراحل استخراج RNA باکیفیت بالا از ۱۰۰ میلی‌گرم بافت منجمد شده گیاهی با استفاده از کیت استخراج RNA کیاژن :RNeasy Plant Mini Kit (QIAGEN., Cat No. 74904) انجام گرفت. کیفیت و کمیت RNA استخراج شده توسط ژل آگارز ۱ درصد و نانودراپ بررسی شده و مجموعه RNA حاصل از استخراج جهت توالی‌یابی دوطرفه (paired end) با ۲۰میلیون خوانش به طول ۱۵۰جفت‌باز به وسیله دستگاه Illumina HiSeq^{TM} 2000 برای شرکت توپاز ارسال شد.

بررسی کیفیت نتایج توالی‌یابی، Trimming و Normalization: نتایج حاصل از توالی‌یابی شامل ۴۷ میلیون خوانش به طول ۱۵۰ جفت‌باز، در قالب فایل فست‌کیو (FASTQ) از طرف شرکت ارسال گردید. این خوانشها Trim شده بودند و توالی آداپتور از ابتدای آنها برش خورده بود. کیفیت داده‌های خام با استفاده از نرم‌افزار FastQC (Version 0.11.5; Simon Andrews) مورد سنجش قرار گرفت و برای رسیدن به صحت و دقت بالاتر در سرهمبندی، توالیهایی با کیفیت کمتر با استفاده از نرم‌افزارهای Trimmomatic (v0.36) (۶) و AfterQC (۸) حذف گردید. در ادامه عمل Normalization بر روی داده‌های خام اعمال گردید که این مرحله با استفاده از ابزار In Silico Normalization که در بسته نرم‌افزاری Trinity قرار دارد صورت گرفت. فرآیند Normalization براساس فرمول P=min(1 T/C) صورت گرفته (۱۷) که در این پژوهش پارامتر T (Max Target Coverage) در تنظیمات نرم‌افزار برابر با ۳۰ (-max_cov 30) درنظر گرفته شد. در این حالت برای خوانشهای پرتکرار حداقل ۳۰ تکرار از هر خوانش حفظ شده و مابقی حذف خواهد شد (۱۸). همچنین این مقدار توسط Haas و همکاران به عنوان مقدار بهینه در هنگام سرهمبندی خوانشها توصیه شده است (۱۷).

سرهمبندی خوانشهای حاصل از توالی‌یابی: نتایج حاصل از مراحل قبل به منظور سرهمبندی و تشکیل توالیهای ترنسکریپتوم با استفاده از نرم‌افزارهای Trinity (v2.4.0)، SOAPdenovo-Trans (v1.03)، Oases-Velvet (v0.2.08) و Trans-ABySS (v1.5.5) مورد آنالیز و ارزیابی قرار گرفت. به منظور بررسی تأثیر پارامترهای دخیل در سرهمبندی، این فرآیند با دو مقدار مختلف ۲۵ و ۳۲ برای پارامتر *K* در تمامی نرم‌افزارها (به جز نرم‌افزار Oases-Velvet که فقط متغیرهای فرد را پذیرفته و مقدار ۳۳ برای آن در نظر گرفته شد) استفاده گردید. براساس مطالعات پیشین؛ بالاترین اندازه برای N50 هنگامی ایجاد می‌شود که بازه عددی ۲۵ تا ۳۵ برای *K*-mer در نظر گرفته شده و مقادیر ۲۵ و۳۲ در اکثر مطالعات دیگر نیز مورد استفاده قرار گرفته است (۵، ۹، ۴۱ و ۴۷). همچنین مقدار ۲۵ به عنوان پیش‌فرض، در برخی از نرم‌افزارهای مورد بررسی، استفاده شده و از سوی دیگر در نرم‌افزار Trinity نیز بیشترین عدد مورد پذیرش برای متغیر *K*-mer، مقدار ۳۲ بود (۱۵). نرم‌افزارهای مورد استفاده در این پژوهش تا زمان نوشتن این مقاله، آخرین نسخه منتشر شده از سوی توسعه‌دهندگان بوده و برای بی‌اثر کردن شرایط محیطی، تمامی آنالیزها در سیستم‌عامل لینوکس و در شرایط مشابه

از نظر مشخصات سیستمی (Ram=40GB, CPU=2.4GHz×16, OS=Ubuntu Linux 16.04 LTS) صورت گرفت. در ادامه همه فایلهای خروجی سرهمبندی شده از تمامی نرم‌افزارهای مورد مطالعه، براساس حداقل طول توالی، با استفاده از ابزار فیلتر طول (fasta_filter_by_min_length) از مجموعه Trinity به مقدار ۲۰۰ جفت‌باز فیلتر گردید. مقدار ۲۰۰ جفت باز، با توجه به طول خوانشهای اولیه ۱۵۰ و مقادیر درنظر گرفته شده برای متغیر *K* (۲۵ و ۳۲)، می‌تواند مبنای مناسبی برای فیلتر کردن نتایج سرهمبندی باشد. همچنین با توجه به این نکته که در این حالت کوچکترین پروتئین حاصل از سرهمبندی، طولی در حدود ۶۶ آمینواسید خواهد داشت. به‌این ترتیب همه فایلهای خروجی از نظر حداقل طول بازسازی شده، یکسان گردید و از توالیهای کوتاه‌تر و بی‌معنی که به طور معمول حاصل از ضعف الگوریتمها و ابزارهای سرهمبندی بوده، چشم‌پوشی شد. در مرحله بعد فایلهای حاصل از سرهمبندی بر اساس پارامترهای تعداد توالی، N50، طول بیشینه و کمینه، میانگین طول توالیهای ایجاد شده، زمان مورد نیاز برای اجرا (RunTime) بررسی گردید.

همردیفی خوانشهای اولیه روی خروجیهای سرهمبندی: در ادامه خوانشهای خام اولیه (قبل از فرآیند Normalization) با استفاده از نرم‌افزار Bowtie2 (v2.3.2) (۲۳) بر روی تمامی فایلهای خروجی شامل؛ سرهمبندی با مقادیر *K*=25 و *K*=32 همچنین قبل از اعمال فیلتر ۲۰۰ و بعد از آن، همردیف گردید. به‌این ترتیب نرخ همردیفی (Alignment rate) و میزان پوشش (Coverage values) ایجاد شده میان نرم‌افزارهای مختلف مورد بررسی و مقایسه قرار گرفت.

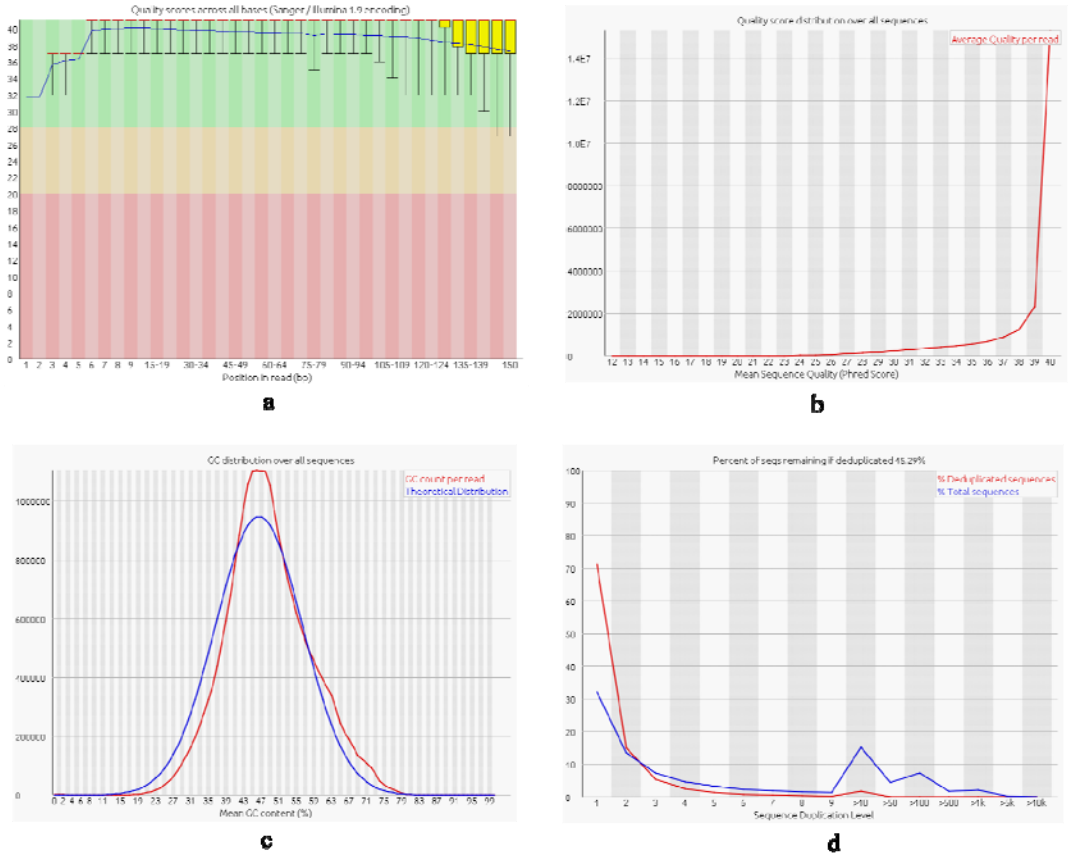
برآورد میزان فراوانی رونوشتهای حاصل از سرهمبندی: برای بررسی میزان فراوانی رونوشتهای حاصل از سرهمبندی، از ابزار RSEM (۲۴) بهره برده که این فرآیند با استفاده از همردیفی و شمارش تعداد خوانشها برای هر رونوشت با در نظر گرفتن طول رونوشت محاسبه گردید. سپس با استفاده از اسکریپت filter_low_expr_transcripts از مجموعه Trinity، به ازای هر ژن تنها یک ایزوفرم که دارای بیشترین فراوانی بود برای ادامه مسیر انتخاب شد.

مستندسازی نتایج و آنالیز GO (Gene Ontology): تمامی رونوشتهای مورد مطالعه، در ابتدا با استفاده از BLASTX (۳) در برابر پایگاه داده گیاه آرابیدوپسیس تالیانا (*Arabidopsis thaliana*) به شماره اختصاصی (taxid:3702) مورد ارزیابی قرار گرفت. هدف از این کار؛ مستندسازی رونوشتها با حساسیت بالا (1e-5) در برابر اطلاعات فراوان پروتئینی در گیاه آرابیدوپسیس تالیانا به عنوان گیاه مدل می‌باشد که در غیر این صورت امکان بررسی هستی‌شناسی در ادامه فراهم نخواهد بود، زیرا برای انجام تجزیه وتحلیل هستی‌شناسی با استفاده از ابزارهایی مانند PANTHER یا AgriGO تنها بررسی ژنهای شناخته شده و دارای شناسه GO در گیاهان محدودی از جمله آرابیدوپسیس تالیانا، برنج، کاهو، گندم و غیره... امکان‌پذیر بوده که در این میان گیاه آرابیدوپسیس تالیانا بیشترین شناسه GO منحصر به فرد را در میان سایر گیاهان به خود اختصاص داده است. این فرآیند پیش از این، در پژوهشهای مشابه برای مستندسازی و هستی‌شناسی رونوشتها صورت پذیرفته است (۱۶ و ۲۹). برای افزایش هرچه بیشتر دقت و اطمینان از نتایج حاصل از این ارزیابی مقدار E-value $\leq 10^{-5}$ در نظر گرفته شد. سپس با طراحی اسکریپت مجزایی نتایج حاصل از BLASTX به منظور دست‌یابی و جداسازی معتبرترین پاسخ برای هر رونوشت مورد بررسی قرار گرفت؛ که در طراحی این اسکریپت ابتدا تمامی نتایج حاصل از BLASTX برای هر رونوشت به طور مجزا بر اساس مقدار E-value پایین‌تر و امتیاز بالاتر مرتب گردیده و سپس بهترین نتیجه به ازای هر رونوشت در فایل خروجی نهایی ذخیره و ثبت گردید. در آخر با

استفاده از شماره اختصاصی مرتبط به هریک از نتایج، فرآیند هستی‌شناسی ژنها (Gene Ontology) در ابزار آنلاین PANTHER (۲۷) انجام شد.

نتایج

پس از دریافت خوانشهای حاصل از توالی‌یابی دوطرفه با استفاده از دستگاه Illumina HiSeq^{TM} 2000، کیفیت نتایج توالی‌یابی توسط FastQC بررسی گردید (شکل ۱) و با توجه به کیفیت قابل قبول خوانشها، سه نوکلئوتید از انتهای ′۳ برخی خوانشها با استفاده از ابزار Trimmomatic و AfterQC حذف شد.



شکل ۱- کیفیت خوانشها برای هر توالی (a)، کیفیت خوانشها برای هر نوکلئوتید (امتیاز Phred) (b)، میزان توالیهای تکراری در خوانشها (c)، مقدار GC در توالی خوانشها (d)

در ادامه باتوجه به نتایج FastQC و وجود خوانشهای تکراری فراوان در داده‌های خام؛ به منظور سهولت و سرعت بالاتر در فرآیند سرهمبندی و از سوی دیگر به علت محدودیت در منابع سیستمی برای آنالیز داده‌ها، لازم است فرآیند Normalization بر روی خوانشها اعمال شود. به این ترتیب برای فایل خام اولیه که دارای ۲۳،۵۹۹،۴۹۵ خوانش به طول ۳،۵۳۹،۹۲۴،۲۵۰ نوکلئوتید با درصد GC ۴۸ و حجم ۸/۵ گیگابایت می باشد؛ بدون از دست دادن خوانشها با تکرار کمتر از ۳۰، خوانشهای تکراری از فایل اولیه حذف شده و نتیجه حاصل از آن، ایجاد فایل نرمالایز شده نهایی با ۶،۵۲۹،۸۵۳ خوانش به طول ۹۷۹،۴۷۷،۹۵۰ نوکلئوتید و درصد GC ۴۵ به حجم ۲/۴ گیگابایت بود (تمامی فایلهای ذکر شده در قالب FASTQ قرار دارند).

سپس فرآیند سرهمبندی با استفاده از خوانشهای نرمالایز شده از مرحله قبل و نرم‌افزارهای ذکر شده صورت گرفت. تمامی نرم‌افزارهای مورد استفاده به جز Trinity همگی براساس نسخه ژنومی آنها ایجاد شده‌اند، به‌این صورت که SOAPdenovo-Trans تحت چهارچوب (Framework) SOAPdenovo، Trans-ABySS در چهارچوب ABySS و Oases با استفاده از نسخه ژنومی آن یعنی Velvet نوشته شده، در حالی که Trinity به طور اختصاصی برای سرهمبندی داده‌های حاصل از توالی‌یابی ترنسکریپتوم توسعه داده شده است. از سوی دیگر درحالی‌که تمامی این نرم‌افزارها از فرآیند ایجاد *K*-mer و الگوریتم De Bruijn graph برای سرهمبندی خوانشها استفاده می‌کنند، ولی باتوجه به گزینه‌ها و متغیرهای خود، فرآیند متفاوتی برای ایجاد فایل رونوشت نهایی و سرهمبندی *K*-mer ها درنظر می‌گیرند. در Trinity تنها یک فایل سرهمبندی به عنوان خروجی ایجاد شده و شامل ایزوفرمها و یونی ژنها می‌باشد؛ این درحالی است که سایر نرم‌افزارها علاوه بر ایجاد فایل کانتیگ (Contig) اولیه؛ قادر به ایجاد سطح بالاتری از سرهمبندی به نام اسکفولد (Scafold) براساس الگوریتم Overlap layout-consensus بوده که اگرچه ممکن است براساس نواحی همپوشان با اختصاصیت پایین تشکیل شده باشد ولی در مراحل بعد و آنالیزهای پایین‌دست می تواند مورد استفاده قرار گیرد.

در SOAPdenovo-Trans فایلهای خروجی شامل Contig و ScafSeq و نرم‌افزار Oases-Velvet نیز دارای دو فایل خروجی با نامهای Contigs و Transcripts بود. اما Trans-AbySS سه فایل با نامهای Transabyss.jn ، Transabyss.ref و Transabyss.final تولید نمود. باتوجه به حداقل طول توالیهای مختلف در فایلهای خروجی هریک از نرم‌افزارها و به منظور بی‌اثر کردن تأثیر توالیهای کوتاه در این پژوهش، فایلهای حاصل از سرهمبندی برای ادامه کار تحت تأثیر فیلتر طول به میزان ۲۰۰ جفت‌باز قرار گرفت. در نرم‌افزار Trinity، حداقل طول توالی بازسازی شده به طور پیش‌فرض ۲۰۰ جفت‌باز بوده و به این ترتیب خروجیهای سایر نرم‌افزارها با استفاده از ابزار فیلتر طول، به میزان ۲۰۰ نوکلئوتید فیلتر گردید.

در ادامه برای رسیدن به میزان صحت و پوشش خوانشها بر روی فایلهای خروجی، خوانشهای اولیه (قبل از نرمالایز شدن) بر روی تمامی فایلهای سرهمبندی شده حاصل از نرم‌افزارهای مختلف توسط Bowtie2 همردیف شده و نرخ همردیفی برای هر نرم‌افزار در هر فایل خروجی بررسی گردید (۳۹). خروجیهای مختلف تمام نرم‌افزارها؛ قبل و بعد از فیلتر ۲۰۰ که در مجموع شامل ۳۰ فایل سرهمبندی مختلف می‌باشد در جدول ۱ باهم مقایسه شده است. به این ترتیب با در نظر گرفتن *K*=25، متغیر N50 که بیانگر حداقل طول ۵۰ درصد از توالیهای سرهمبندی شده می‌باشد (۲۸) برای خروجی transcripts از نرم‌افزار Oases-Velvet (۱۹۷۶ bp) ، خروجی transabyss.ref از Trans-ABySS (۱۱۹۳bp) و همچنین خروجی Trinity (۱۵۸۵bp) در مقدار قابل قبولی قرار دارند و باتوجه به تعداد توالیهای ایجاد شده، خروجی Trinity، فایل transcripts از نرم‌افزار Oases-Velvet در وضعیت بهتری قرار دارند. همچنین در حالتی که متغیر *K*=32 در نظر گرفته شد، فایل transcripts از نرم‌افزار Oases-Velvet (۲۱۴۷bp) ، خروجی transabyss.ref از نرم‌افزار Trans-ABySS (۱۳۱۲ bp) و خروجی Trinity (۱۶۴۷ bp) بیشترین مقدار N50 را به خود اختصاص داده (جدول ۱) و در مطالعات قبلی که روی Trinity، SOAPdenovo و Trans-ABySS صورت گرفت نیز نتایج مشابهی به دست آمد؛ همچنین در اثر افزایش مقدار *K*-mer، میزان N50 به طور محسوس افزایش یافته که در این پژوهش نیز نتایج به دست آمده بیانگر این مورد است (۹ و ۲۵).

بر اساس تعداد توالی پیش‌بینی شده باتوجه به میزان N50 هریک از خروجیها، در هر دو حالت *K*=32 و *K*=25 ، فایل transabyss-final تعداد بیشتری توالی را بازسازی

نموده (*K*=25 شامل ۳۹۰۰۷۲ توالی و *K*=32 شامل ۲۸۶۰۰۴ توالی) که در مطالعه Wang و Gribskov نیز این مورد بالاترین تعداد (۱۰۷۰۸۸۷) را ایجاد نمود (۴۱). پارامتر طول توالیهای بازسازی شده در این پژوهش در حالت *K*=25 نشان‌دهنده برتری Oases-Velvet در فایل transcripts با میانگین طول ۱۰۵۱ bp و در حالت *K*=32 نیز با میانگین ۱۳۶۳ bp دارای برتری بوده ولی باتوجه به فواصل خالی (gaps) (۴۳) در توالیهای تشکیل شده توسط Oases-Velvet، این امر می‌تواند امتیازی منفی به شمار آید. در این‌صورت فایل خروجی Trinity با میانگین طول bp ۱۰۳۶ در *K*=25 و طول bp ۱۰۶۶ در *K*=32، از این نظر موفقیت بیشتری به همراه داشته (جدول ۱) و این مورد در پژوهشهای پیشین نیز مورد توجه قرار گرفته است (Zhao و همکاران میانگین طول bp ۶۰۴ و Wang و Gribskov نیز برتری Oases-velvet را برای میانگین طول بهتر به میزان bp ۱۵۹۳ برای *K*=25 و bp ۱۷۱۰ برای *K*=31 گزارش کردند) (۴۱ و ۴۷). براساس میزان منابع سیستمی، از جمله مقدار رم و زمان مورد نیاز برای اجرا، SOAPdenovo-Trans کمترین مقدار رم و سریع‌ترین زمان (۳۵ دقیقه) برای سرهمبندی را به خود اختصاص داده و در مقابل، Trinity علاوه بر مقدار بالاتر رم (یک گیگابایت به ازای هر یک میلیون خوانش (۱۹))، زمان بالاتری (۱۵ ساعت) نیز برای سرهمبندی خوانشها صرف می‌کند. البته باید به این نکته توجه نمود که کارآیی آنالیز در هر نرم‌افزار برای سرهمبندی خوانشها، ارتباط مستقیمی با دقت و صحت سرهمبندی ندارد (۴۷).

همردیف کردن خوانشهای خام اولیه بر روی خروجی هر نرم‌افزار موجب شکل‌گیری مبنای مناسبی برای مقایسه میان نتایج سرهمبندی گردیده که براساس نتایج جدول ۱ با افزایش مقدار *K* نرخ همردیفی به طور محسوس افزایش یافته و همه‌ی نرم‌افزارها به غیر از SOAPdenovo-Trans افزایش قابل توجهی را در مقدار همردیفی و تعداد خوانشهای همردیف شده نشان می‌دهند، به‌این صورت که با در نظر گرفتن *K*=25 بیشترین همردیفی مربوط به transabyss.final (۷۰/۸۷ درصد) و خروجی Trinity (۶۷/۷۱ درصد) همچنین در حالت *K*=32 نیز خروجی transabyss.final (۷۵/۴۹ درصد) و Trinity (۷۲/۹۸ درصد) بیشترین مقادیر را به خود اختصاص داده (جدول ۱) که در مطالعه سال ۲۰۱۲ روی دروزوفیلا نتایج مشابهی برای پوشش ژنومی خوانشها در برابر ژنوم مرجع موجود در پایگاه‌داده حاصل آمد (Trinity با ۷۸/۶ درصد و Trans-ABySS با ۶۴/۳ درصد) (۴۷). همچنین در مطالعه اخیر بر روی نرم‌افزارهای مطرح در این زمینه و سرهمبندی خوانشهای آرابیدوپسیس و همردیفی آنها روی ژنوم، نتیجه مشابهی به دست آمد (Trans-ABySS ۹۳/۴۴ درصد و Trinity ۹۰/۲۱ درصد) (۴۱).

همچنین با افزایش مقدار *K*-mer، تعداد توالیهای تشکیل شده در نرم‌افزار Trinity از ۱۶۵۱۲۵ به ۱۶۵۶۹۷و در نرم‌افزار SOAPdenovo-Trans (فایل scafseq) از ۶۸۹۷۹ به ۷۶۲۲۳ عدد افزایش یافته و در دو نرم‌افزار دیگر یعنی Oases-Velvet (فایل transcripts) از ۱۲۴۸۱۳ به ۱۰۰۰۴۸و در نرم‌افزار Trans-ABySS (فایل transabyss.final) از ۳۹۰۰۷۲ به ۲۸۶۰۰۴ عدد کاهش یافته که این امر می‌تواند با الگوریتمهای متفاوت این نرم‌افزارها در سرهمبندی در ارتباط باشد. از سوی دیگر طول توالیهای بازسازی شده با افزایش مقدار *K*-mer به طور محسوس افزایش داشته و تنها در مورد SOAPdenovo-Trans با افزایش مقدار *K*-mer، میانگین طول توالیها از bp ۴۴۱ به bp ۴۰۵ کاهش یافته و باتوجه به کوتاه شدن توالیهای حاصل از سرهمبندی، این امر می‌تواند علت کاهش نرخ همردیفی در وضعیت *K*=32 در خروجی scafseq را توجیه کند. ایجاد ترنسکریپت بلندتر و بدون فاصله، نقش مهمی در تشکیل نواحی همپوشان برای آنالیزهای پایین‌دست به منظور شناسایی و پیش‌بینی ORFها (Open Reading Frame) و ژنهای این گیاه دارد (جدول ۱).

در مطالعه‌ای که با استفاده از خوانشهای شبیه‌سازی شده در کنار خوانشهای واقعی روی کروموزوم شماره۲۲ انسان جهت بررسی کارآیی چند نرم‌افزار سرهمبندی از جمله Trinity، Oases و ABySS صورت گرفت، نتایج از برتری Trinity در زمینه نرخ پوشش ژنومی و پارامترهای آماری مرتبط از جمله N50 حکایت داشته و از طرفی در مقایسه میان دو ابزار دیگر یعنی Oases و ABySS، به برتری نرم‌افزار Oases اشاره شده که علت این امر می‌تواند با به روزرسانیهای متعدد از زمان انتشار مقاله مذکور در ارتباط باشد (۱۰).

جدول ۱- مقایسه نرم‌افزارهای Trinity، SOAPdenovo-Trans، Oases-Velvet و Trans-ABySS در سرهمبندی خوانشهای حاصل از توالی‌یابی ترنسکریپتوم زرین‌گیاه

	Oases-Velvet		SOAPdenovo-Trans		Trans-ABySS			Trinity
File names	contigs	Transcripts	contig	scafSeq	Transabyss.jn	Transabyss.ref	Transabyss.final	Trinity
				K=۲۵				
N50 bp	۲۶۴ (۹۸)	۲۰۲۸ (۱۹۷۶)	۶۰۶ (۳۵۵)	۱۲۷۳ (۹۹۹)	۶۵۴ (۶۴۲)	۱۳۰۲ (۱۱۹۳)	۹۵۵ (۶۷۶)	۱۵۸۵
Sequence number	۹۹۱۰۲ (۱۳۵۴۶۲۲)	۹۴۹۱۰ (۱۲۴۸۱۳)	۸۹۳۱۸ (۳۰۴۷۷۲)	۶۸۹۷۹ (۱۴۴۰۷۵)	۵۵۳۸۲ (۶۰۰۶۳)	۵۷۰۳۸ (۹۰۳۱۹)	۱۱۲۲۴۲ (۳۹۰۰۷۲)	۱۶۵۱۲۵
Average Length bp	۲۸۲/۹ (۹۱/۹)	۱۳۳۶/۸ (۱۰۵۱)	۵۰۵/۸ (۲۱۹/۲)	۷۷۸ (۴۴۱/۲)	۵۶۷۱ (۵۳۶)	۸۳۳/۷ (۵۳۶/۱)	۶۸۴/۸ (۲۴۸)	۱۰۳۶/۸۸
Max length bp Min length	۲۴۰۰ ۲۰۰ (۴۹)	۱۵۴۲۹ ۲۰۰ (۱۰۰)	۸۵۵۲ ۲۰۰ (۲۶)	۱۵۰۳۵ ۲۰۰ (۱۰۰)	۵۲۷۶ ۲۰۰ (۴۸)	۱۴۷۳۰ ۲۰۰ (۲۶)	۱۴۷۳۰ ۲۰۰ (۲۵)	۱۳۲۵۰ ۲۰۱
Alignment Rate %	۰/۷۲ (۰/۷۹)	۵۸/۴۱ (۵۸/۴۳)	۳۹/۵۵ (۳۹/۵۹)	۴۴/۴۵ (۴۴/۴۸)	۱۵/۳۶ (۱۵/۳۸)	۶۷/۵۶ (۶۷/۵۸)	۷۰/۸۷ (۷۰/۹۱)	۶۷/۷۱
Mapped Reads	۱۷۰۸۱۵ (۱۸۵۴۵۹)	۱۳۷۸۴۶۱۴ (۱۳۷۹۰۰۶۸)	۹۳۳۲۸۰۵ (۹۳۴۴۱۸۴)	۱۰۴۹۰۸۸۹ (۱۰۴۹۶۱۴۶)	۳۶۲۴۵۹۹ (۳۶۲۸۵۰۲)	۱۵۹۴۴۱۴۸ (۱۵۹۴۹۴۰۳)	۱۶۷۲۵۵۷۰ (۱۶۷۳۳۴۵۰)	۱۵۹۷۹۱۸ ۵
RunTime	۰۰:۰۵' (۰۰:۱۳')	۰۰:۵۸' (۲:۲۰')	۰۰:۲۲' (۰۰:۲۹')	۰۰:۵۹' (۰۰:۲۹')	۰۰:۳۹' (۰۰:۴۱')	۰۰:۳۲' (۰۰:۳۳')	۰۰:۳۹' (۰۰:۴۵')	۱:۵۰'
				K=۳۲				
N50 bp	۲۶۴ (۱۲۳)	۲۱۷۰ (۲۱۴۷)	۵۸۱ (۳۳۲)	۱۲۸۸ (۹۵۵)	۷۹۱ (۷۸۴)	۱۳۶۵ (۱۳۱۲)	۱۰۵۰ (۸۳۰)	۱۶۴۷
Sequence number	۱۱۴۵۰۳ (۱۰۱۰۰۴۶)	۸۹۶۲۳ (۱۰۰۰۴۸)	۹۵۸۲۶ (۳۰۵۱۳۶)	۷۶۲۲۳ (۱۸۲۹۵۶)	۵۵۵۹۷ (۵۷۹۴۹)	۵۹۳۲۲ (۷۲۹۷۱)	۱۱۵۲۷۴ (۲۸۶۰۰۴)	۱۶۵۵۹۷
Average Length bp	۲۶۸/۶ (۱۱۷)	۱۵۰۱/۸ (۱۳۶۳)	۴۹۰/۶ (۲۳۲/۲)	۷۹۱/۱ (۴۰۵/۵)	۶۶۲/۶ (۶۴۲/۷)	۸۸۱/۷ (۷۳۹/۸)	۷۴۱/۲ (۳۵۳/۸)	۱۰۶۶/۴
Max length bp Min length	۲۵۶۲ ۲۰۰ (۶۵)	۱۵۸۵۷ ۱۲۰ (۲۰۰)	۶۸۵۸ ۲۰۰ (۳۴)	۱۵۰۷۵ ۲۰۰ (۱۰۰)	۴۸۷۷ ۲۰۰ (۷۳)	۱۶۱۵۰ ۲۰۰ (۳۳)	۱۶۱۵۰ ۲۰۰ (۳۲)	۱۲۳۳۱ ۲۰۱
Alignment Rate %	۱/۰۵ (۱/۱۴)	۶۵/۹۲ (۶۵/۹۴)	۴۰/۱۷ (۴۰/۲۲)	۴۳/۵۳ (۴۳/۵۶)	۲۱/۵۹ (۲۱/۶۰)	۷۱/۸۲ (۷۱/۸۳)	۷۵/۵۹ (۷۵/۵۲)	۷۲/۹۸
Mapped Reads	۲۴۸۷۶۸ (۲۶۹۱۹۳)	۱۵۵۵۷۲۲۴ (۱۵۵۶۱۶۴۸)	۹۴۷۹۳۱۴ (۹۴۹۱۳۶۰)	۱۰۲۷۳۸۳۷ (۱۰۲۸۰۲۴۱)	۵۰۹۴۹۹۶ (۵۰۹۷۸۶۶)	۱۶۹۴۹۰۴۴ (۱۶۹۵۱۰۹۳)	۱۷۸۱۴۸۱۷ (۱۷۸۲۲۵۱۱)	۱۷۲۲۳۸۷ ۶
RunTime	۰۰:۲۵' (۰۰:۳۶')	۱:۲۸' (۱:۲۰')	۰۰:۲۹' (۰۰:۴۸')	۰۰:۳۲' (۰۰:۳۶')	۰۰:۱۱' (۰۰:۱۶')	۰۰:۵۰' (۰۰:۵۶')	۰۰:۴۸' (۰۰:۵۰')	۲:۰۸'

* مقادیر داخل پرانتز بدون فیلتر ۲۰۰ برای حداقل طول توالی محاسبه شده‌اند.

براساس مشاهدات این تحقیق، Trinity و Oases-Velvet در *K*-merهای بزرگ‌تر عملکرد مناسب‌تری داشته و همان گونه که در جدول ۱ مشاهده می‌شود با افزایش مقدار *K*، میزان پوشش همردیفی بالاتری برای خروجیهای این دو نرم‌افزار ایجاد گردیده‌است. به این‌صورت که Trinity از ۶۷/۷۱ درصد به ۷۲/۹۸ درصد و Oases-Velvet از ۵۸/۴۳ درصد به ۶۵/۹۴ درصد در *K*=32 رسیده است و این درحالی است که SOAPdenovo-Trans درحالت *K*=25 نتایج مناسب‌تری تولید نمود (۴۴/۴۸ درصد و در *K*=32 نرخ همردیفی به ۴۳/۵۶ درصد کاهش یافت).

ابزار Trinity، نرم‌افزاری رایگان و متن‌باز بوده و از سه ماژول مختلف (Butterfly و Inchworm, Chrisalis) برای انجام فرآیند سرهمبندی استفاده کرده که این عمل از طریق: سرهمبندی خوانشهای توالی‌یابی به صورت رونوشتهای اولیه در Inchworm، دسته‌بندی این رونوشتها و تشکیل گراف دی‌بروین درChrisalis و درنهایت پردازش گراف به منظور گزارش تمام رونوشتها با طول کامل و ایزوفرمهای حاصل از پیرایش ثانویه در Butterfly صورت می‌گیرد. فایل نهایی حاصل از سرهمبندی در قالب FASTA (.fa) و با اندازه تقریبی ۲۴۰ مگابایت؛ شامل ۶۷۸۵۹ ژن و ۱۶۵۵۹۷ رونوشت با درصد GC ۴۲/۶۴ بود. میزان N50 نیز برای رونوشتها و ژنها به ترتیب برابر با ۱۶۴۷ و ۱۳۵۳ بوده، همچنین اطلاعات آماری بیشتر در ارتباط با ترنسکریپتوم سرهمبندی شده در جدول ۲ قابل مشاهده است که با توجه به نرخ همردیفی (Mapping) ۷۲/۹۸ (جدول ۱) و اطلاعات آماری مناسب (جدول ۲)، خروجی Trinity (*K*-mer=32) در وضعیت خوبی قرار داشته و نقطه شروع مناسبی برای آنالیزهای پایین دست می‌باشد. در ادامه با استفاده از ابزار RSEM فراوانی هریک از رونوشتها بر مبنای دو پارامتر TPM و FPKM محاسبه شد که نتایج این بررسی در فایل S1 ضمیمه شده است.

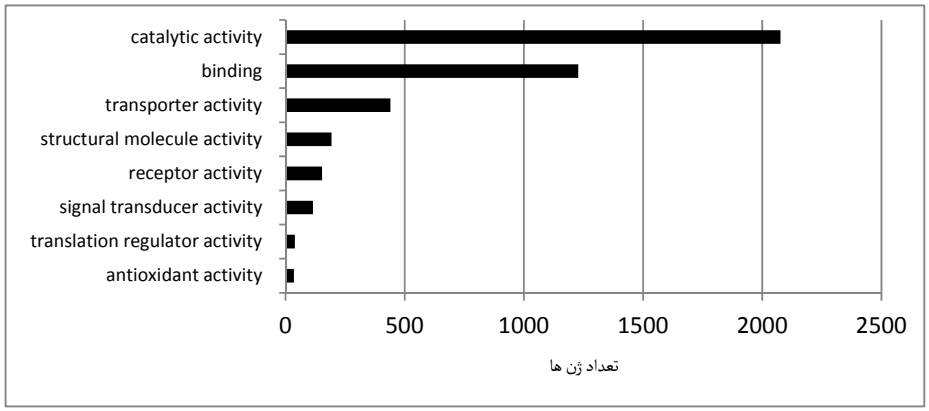
جدول ۲- اطلاعات آماری فایل خروجی سرهمبندی با ابزار Trinity

تمام ژنهای Trinity	۶۷۸۵۹
تمام رونوشتهای Trinity	۱۶۵۵۹۷
درصد GC	۴۲/۶۴
اطلاعات آماری بر اساس تمام رونوشتها	
Contig N10	۳۳۲۵ bp
Contig N20	۲۶۲۱ bp
Contig N30	۲۱۸۸ bp
Contig N40	۱۸۶۷ bp
Contig N50	۱۶۴۷ bp
طول میانه	۷۴۲ bp
طول متوسط	۱۰۳۶/۸۸ bp
تعداد بازهای سرهمبندی شده	۱۷۱۲۱۵۵۳۱
اطلاعات آماری بر اساس بلندترین ایزوفرم در هر ژن	
Contig N10	۳۲۷۵ bp
Contig N20	۲۵۲۱ bp
Contig N30	۲۰۶۵ bp
Contig N40	۱۶۹۳ bp
Contig N50	۱۳۵۳ bp
طول میانه	۴۴۶ bp
طول متوسط	۷۹۵/۶۲ bp
تعداد بازهای سرهمبندی شده	۵۳۹۸۹۹۷۲

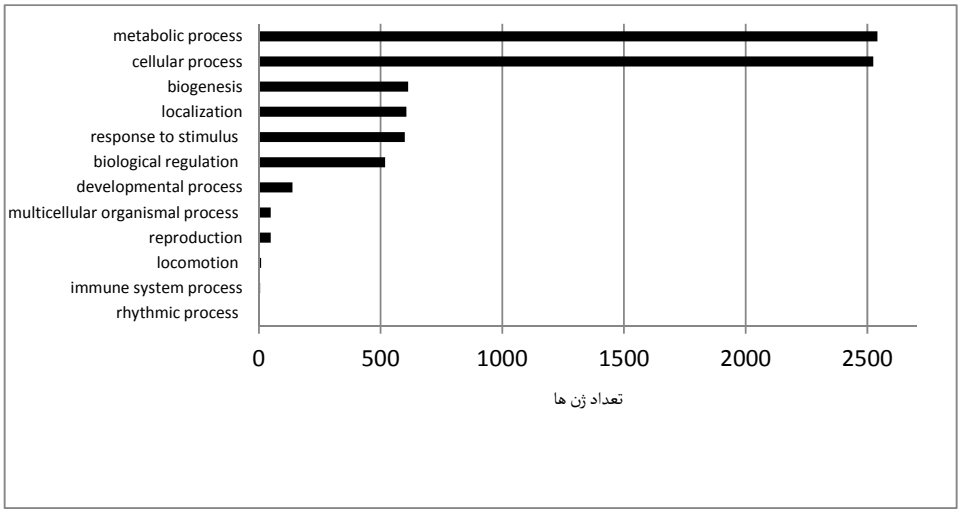
سپس به منظور جداسازی رونوشت دارای بیشترین بیان به ازای هر ژن، از ابزار filter_low_expr_transcripts استفاده کرده که در نتیجه فایل fasta با حجم ۶۵ مگابایت و دارای ۶۷۸۵۹ توالی منحصر به فرد با میزان N50 به طول bp ۱۰۲۱، میانگین bp ۶۸۳ و حداکثر طول bp ۱۱۸۲۰ برای انجام آنالیز BLASTX ایجاد شد. پس از انجام BLASTX، تعداد ۴۵۴۳۹۹ خروجی برای ۲۴۸۸۷ رونوشت از مجموع ۶۷۸۵۹ عدد رونوشت اولیه توسط فرآیند BLASTX از میان پروتئینهای گیاه آرابیدوپسیس تالیانا استخراج شد. سپس با استفاده از اسکریپت طراحی شده برای این پژوهش، معتبرترین نتایج براساس میزان امتیاز بالاتر و مقدار $E\text{-value} \leq 10^{-5}$ به ازای هر رونوشت جدا سازی گردید که شماره اختصاصی این پروتئینها در پایگاه داده آرابیدوپسیس تالیانا، در فایل S2 ضمیمه شده است. بدین ترتیب فرآیند هستی‌شناسی با استفاده از ۲۴۸۸۷ نتیجه

نهایی در ابزار PANTHER صورت گرفته و ۷۶۸۴ عدد پاسخ GO منحصر به فرد ثبت گردید که نتایج این بررسی در شکل ۲ مشاهده می‌شود. خروجی نهایی بررسی GO در فایل S3 قرار دارد.

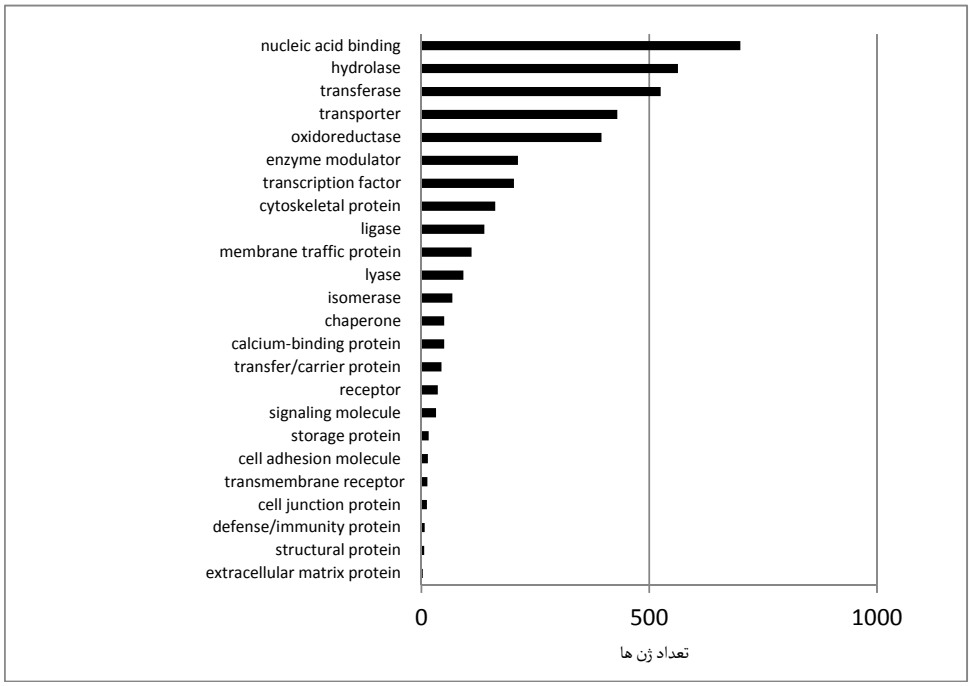
اجزای تشکیل‌دهنده سلول(Cellular Component)



فعالیتهای مولکولی (Mollecular Function)



فرآیندهای زیستی(Biological Process)



شکل۲- نتایج هستی‌شناسی (GO) ۲۴۸۸۸ عدد از رونوشتهای سرهمبندی شده در سیستم PANTHER: اجزای تشکیل‌دهنده سلول (Cellular Component)، فعالیتهای مولکولی (Mollecular Function)، فرآیندهای زیستی (Biological Process) و نتایج حاصل از دسته‌بندی پروتئینهای شناسایی شده

بحث

براساس نتایج این پژوهش خروجی scafSeq از نرم‌افزار SOAPdenovo-Trans و خروجی transcripts از نرم‌افزار Oases-Velvet همچنین در نرم‌افزار Trans-ABySS دو فایل خروجی Transabyss.ref و Transabyss.final با توجه به پوشش بهتر و میزان N50 بالاتر نسبت به سایر خروجیها در وضعیت مناسب‌تری قرار دارند. از سوی دیگر با افزایش میزان K تغییرات محسوسی در نرخ همردیفی و پوشش خوانشها ایجاد شده که K=32 می تواند مقدار مناسبی برای انجام این‌دست پژوهشها به شمار آید. در نهایت می توان خروجی نرم‌افزار Trinity و همچنین Transabyss.final از نرم‌افزار Trans-ABySS را بهترین گزینه برای انجام آنالیزهای پایین‌دست معرفی نمود. که باتوجه به سرعت بالاتر و زمان کوتاه‌تر در نرم‌افزار Trans-ABySS، این مورد می تواند امتیاز دیگری برای آن محسوب شود. با این حال خروجی Trinity تنها موردی است که علاوه بر تشکیل ترنسکریپتها، می‌تواند توالیهای ایجاد شده و مشابه به هر یونی ژن را تحت عنوان یک ایزوفرم دسته‌بندی و مشخص نماید که این عمل تنها مخصوص به این نرم‌افزار بوده و برای بررسی میزان بیان ژنها و رونوشتها و همچنین تغییرات ژنتیکی در محیط نرم‌افزاری مفید باشد. در ادامه فرآیند هستی‌شناسی در میان رونوشتهای دارای بیشترین فراوانی نشان دهنده فعالیت بالای کاتالیتیک و حضور بسیار بالای پروتئینهای دخیل در اتصال و همچنین پروتئینهای دارای فعالیتهای آنتی اکسیدانی در میان فعالیتهای مولکولی این گیاه بوده و باتوجه به دارویی بودن این گیاه و اهمیت متابولیتهای ثانویه در آن؛ از جمله متوکسی فلاون‌ها و رزمارینیک اسید (۱)، بیشترین فراوانی در بخش فرآیندهای زیستی با بیش از ۲۵۰۰ ژن (شکل ۲)، مربوط به فرآیندهای متابولیتی در این گیاه دارویی می‌باشد که با یافته‌های مشابه در گونه‌های نزدیک مطابقت دارد (۲ و ۲۶). کلاس‌بندی پروتئینهای

شناسایی شده نیز حاکی از بیان بالای پروتئینهای ترنسفراز و اکسیدو ردوکتازها بوده که با توجه به استخراج نمونه اولیه از برگ گیاه توجیه می‌گردد. با این همه؛ توسعه و گسترش نرم‌افزارها و الگوریتمهای سرهمبندی برای رسیدن به مسیر و ابزاری با قابلیت بسیار بالاتر و دقیق‌تر همچنان به عنوان یک چالش مهم در دنیای بیوانفورماتیک مطرح بوده و از سوی دیگر برای بررسی دقیق ژنها و پروفایل بیان این گیاه نیز، مستندسازی تمامی رونوشتها و بررسی میزان بیان برخی از فراوان‌ترین آنها توسط روشهای معمول آزمایشگاهی و مقایسه با روشهای نرم‌افزاری می‌تواند در یافتن مسیرهای متابولیتی با اهداف افزایش بیان متابولیت سودمند، راهگشای محققین و علاقه‌مندان قرار گیرد.

منابع

۱- ایوبی، ن.، حسینی، ب. و فتاحی، م. (۱۳۹۶). اثر القایی کیتوزان و کلشی‌سین بر تولید رزمارینیک اسید در ریشه مویین زرین گیاه (*Dracocephalum kotschyi* Boiss).مجله پژوهشهای سلولی و مولکولی، ۳۰(۱)، ۱-۱۳.

۲- میرزایی، س. (۱۳۹۶). پروفایل ترانسکریپتوم نوک شاخه و ریشه گیاه سویا. مجله پژوهشهای سلولی و مولکولی، ۳۰(۴)، ۴۹۹-۵۱۱.

3- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

4- Amirghofran Z, Azadbakht M, Karimi MH (2000) Evaluation of the immunomodulatory effects of five herbal plants. J Ethnopharmacol 72(1):167–172

5- Blande D, Halimaa P, Tervahauta AI, Aarts MGM, Kärenlampi SO (2017) *De novo* transcriptome assemblies of four accessions of the metal hyperaccumulator plant Noccaea caerulescens. Sci Data 4:160131

6- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120

7- Bräutigam A, Mullick T, Schliesky S, Weber APM (2011) Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species. J Exp Bot 62(9):3093–3102

8- Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J (2017) AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics 18(Suppl 3):80

9- Chopra R, Burow G, Farmer A, Mudge J, Simpson CE, Burow MD (2014) Comparisons of *de novo* transcriptome assemblers in diploid and polyploid species using peanut (Arachis spp.) RNA-seq data. PLoS One 9(12):e115055

10- Clarke K, Yang Y, Marsh R, Xie L, Zhang KK (2013) Comparative analysis of *de novo* transcriptome assembly. Sci China Life Sci 56(2):156

11- 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):392

12- Fattahi M, Nazeri V, Torras-Claveria L, Sefidkon F, Cusido RM, Zamani Z, Palazon J (2013) Identification and quantification of leaf surface flavonoids in wild-growing populations of *Dracocephalum kotschyi* by LC–DAD–ESI-MS. Food Chem 141(1):139–146

13- Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. Nat Methods 6:S6–S12

14- Grabherr MG, Haas BJ, Yassour M, et al (2011) Trinity: recontructing a full-length transcriptome assembly without a genome from RNA-Seq data. Nat Biotechnol 29(7):644–652

15- Grabherr M, Haas BJ, Yassour M, et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–52

16- Guo L, Allen KS, Deiulio G, Zhang Y, Madeiras AM, Wick RL, Ma L (2016) A *De Novo* - Assembly Based Data Analysis Pipeline for Plant Obligate Parasite Metatranscriptomic Studies. 7(July):1–9

17- Haas BJ, Papanicolaou A, Yassour M, et al (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity

platform for reference generation and analysis. Nat Protoc 8(8):1494–1512

18- Haas BJ, Papanicolaou A, Yassour M, et al (2013) *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc 8(8):10.1038/nprot.2013.084

19- Haas BJ, Papanicolaou A, Yassour M, et al (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8(8):1494–1512

20- Jalili A, Jamzad Z (1999) Red data book of Iran: A preliminary survey of endemic, rare and endangered plant species in Iran.

21- Javidnia K, Miri R, Kamalinejad M, Khoshneviszadeh M (2006) Constituents of the volatile oils of *Dracocephalum kotschyi* Boiss. from Iran. J Essent Oil Res 18(3):342–344

22- Kamali M, Khosroyar S, Jalilvand MR (2014) Evaluation of phenolic, flavonoids, anthocyanin contents and antioxidant capacities of different extracts of aerial parts of *Dracocephalum kotschyi*. JN Khorasan Univ Med Sci 6:627–634

23- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359

24- Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12(1):323

25- Li G, Du J, Li Y, Wu J (2015) Identification of putative olfactory genes from the oriental fruit moth grapholita molesta via an antennal transcriptome analysis. PLoS One 10(11):1–30

26- Li H, Fu Y, Sun H, Zhang Y, Lan X (2017) Transcriptomic analyses reveal biosynthetic genes related to rosmarinic acid in *Dracocephalum tanguticum*. Sci Rep 7(1):74

27- Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the panther classification system. Nat Protoc 8(8):1551–1566

28- Miller JR, Koren S, Sutton G (2010) Assembly Algorithms for Next-Generation Sequencing Data. Genomics 95(6):315–327

29- Mizrachi E, Hefer CA, Ranik M, Joubert F, Myburg AA (2010) *De novo* assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq.

30- Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics 92(5):255–264

31- Mozaffarian V (1996) A dictionary of Iranian plant names: Latin, English, Persian. Farhang Mo'aser

32- Rechinger KH (1982) Salvia in Flora Iranica, Labiatae, No. 150, edits. Rechinger KH Hedge IC, Akad. Druck Vertagsanstalt, Graz, Austria. , p 477

33- Ren X, Liu T, Dong J, Sun L, Yang J, Zhu Y, Jin Q (2012) Evaluating de Bruijn Graph Assemblers on 454 Transcriptomic Data. PLoS One. doi: 10.1371/journal.pone.0051188

34- Robertson G, Schein J, Chiu R, et al (2010) *De novo* assembly and analysis of RNA-seq data. Nat Meth 7(11):909–912

35- Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28(8):1086–1092

36- Sharafi A, Sohi HH, Azadi P, Sharafi AA (2014) Hairy root induction and plant regeneration of medicinal plant *Dracocephalum kotschyi*. Physiol Mol Biol Plants 20(2):257–262

37- Tang Q, Ma X, Mo C, Wilson IW, Song C, Zhao H, Yang Y, Fu W, Qiu D (2011) An efficient approach to finding *Siraitia grosvenorii* triterpene biosynthetic genes by RNA-seq and digital gene expression analysis. BMC Genomics 12(1):343

38- Tao X, Gu Y-H, Wang H-Y, Zheng W, Li X, Zhao C-W, Zhang Y-Z (2012) Digital gene expression analysis based on integrated *de novo* transcriptome assembly of sweet potato [*Ipomoea batatas* (L.) Lam.]. PLoS One 7(4):e36234

39- Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. Nat Biotechnol 27(5):455–457

40- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63

41- Wang S, Gribskov M (2016) Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics 215:btw625

42- Xie Y, Wu G, Tang J, et al (2014) SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics 30(12):1660–1666

43- Xu D-L, Long H, Liang J-J, Zhang J, Chen X, Li J-L, Pan Z-F, Deng G-B, Yu M-Q (2012) *De novo* assembly and characterization of the root transcriptome of Aegilops variabilis during an interaction with the cereal cyst nematode. BMC Genomics 13(1):133

44- Yaghmai MS, Taffazoli R (1988) The essential oil of *Dracocephalum kotschyi* Boiss. Flavour Fragr J 3(1):33–36

45- Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res 18(5):821–829

46- Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, Hao P (2011) Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12(14):S2

47- ZHAO L, Zachary L-R, CHEN S-Y, GUO Z-H (2012) comparing *De Novo* Transcriptome Assemblers Using Illumina RNA-Seq Reads. Plant Divers Resour 34(5):487

48- NCBI database search. https://www.ncbi.nlm.nih.gov/gquery/?term=*Dracocephalum+kotschyi*.

The Comparison of Assembly Softwares and Gene Ontology Analysis using transcriptome sequencing data from *Dracocephalum kotschyi* Boiss.

Poursalavati A.N.[1], Ebrahimi A.[2] and Rashidi monfared S.[1]

[1] Dept. of Agricultural Biotechnology, Tarbiat Modares University, Tehran, I.R. of Iran.

[2] Dept. of Agronomy, Faculty of Agriculture, Shahrood Univwrsity of Technology, Shahrood, I.R. of Iran.

Abstract

With fast advances in next generation sequencing technologies, they have become powerful and low-cost tools for transcriptome studies, Nowadays; *de novo* assembly of transcriptome data, has caused the formation of the new pathway in the study of non-model genome species. With the expansion of this technology and increasing the number of assembly softwares, choosing the best software for assembling transcriptome sequencing data has become a challenge for biologists. For the first time in this study, we used transcriptome sequencing data of *Dracocephalum kotschyi* in order to reach the appropriate parameters and superior software; so here we used Oases-velvet, SOAPdenovo-Trans, Trans-ABySS and Trinity softwares with two different values of *K* parameter; *K*=25 and *K*=32. The results of assembly by each software were compared to others in the term of several criteria. The result suggests the superiority of Trinity and Trans-ABySS softwares. Finally, the output of the best assembly was used to estimate abundance of various isoforms and Gene Ontology analysis as regards to the pharmaceutical properties of this plant and the high amount of secondary metabolites, the highest frequency of sections in the biological processes was related to the metabolic activity.

Key words: Trinity, SOAPdenovo-Trans, Oases-velvet, Trans-ABySS, Gene Ontology