

تجزیه ترانسکرپتوم کلالة زعفران زراعی (*Crocus sativus* L.) با استفاده از نرم افزارهای

Trinity و SOAPdenovo

پروانه محمودی^۱، احمد معینی^{۲*}، سید مجتبی خیام نکویی^۳، محسن مردی^۴، قاسم حسینی سالکده^{۴*}

۱. دانشجوی دکتری اصلاح نباتات، دانشگاه تربیت مدرس، تهران ۲. دانشیار، گروه اصلاح نباتات و بیوتکنولوژی دانشگاه تربیت مدرس، تهران، ۳. دانشیار،

دانشکده علوم زیستی دانشگاه تربیت مدرس، تهران، ۴. دانشیاران، بخش ژنومیکس پژوهشکده بیوتکنولوژی کشاورزی، کرج

(تاریخ دریافت: ۱۳۹۳/۲/۱۱ - تاریخ تصویب: ۱۳۹۳/۶/۲۶)

Analysis of Saffron Stigma (*Crocus sativus* L.) Transcriptome Using SOAPdenovo and Trinity Assembly SoftwareP. MAHMOUDI¹, A. MOIENI^{2*}, S. M. KHAYAM NEKOEI³, M. MARDI⁴,GH. HOSSEINI SALEKDEH^{4*}

1. Ph.D. Student of plant breeding, Faculty of Agriculture, Tarbiat Modares University, Tehran, Iran.

2. Associate Professor, Faculty of Agriculture, Tarbiat Modares University, Tehran, Iran.

3. Associate Professor, Faculty of Biological Biotechnology Research Institute of Iran, Karaj, Iran.

(Received: May. 1, 2014 - Accepted: Sep. 17, 2014)

Abstract

Saffron (*Crocus sativus* L.), is the most valuable and popular spice in the world. It is a triploid species of the Iridaceae family. Its long, red, and aromatic stigma distinguishes this species from others in its family. With the rapid advances in next generation sequencing technology, RNA sequencing has risen as a cost-effective and powerful method for transcriptome study. *De novo* assembly of transcripts provides main solution to transcriptome analysis for organisms without reference genome. Precise sequencing and assembly of transcriptome reliable data are necessary for the downstream analysis. Accordingly, this work was analyzed by two of the most popular software's Trinity and SOAPdenovo for saffron stigma transcriptome to study the effective programs for transcriptome assembly. The results showed that the mean sequence length and the number of unigenes obtained by Trinity were more than SOAPdenovo (Trinity: 689, SOAPdenovo: 624). Translation of the better results produced by the appropriate assembler to protein led to a significant increase in the number of its records obtained at databases. Furthermore, these unigenes might help to identify more metabolite pathways. Assembled unigenes by Trinity had no lacking distance and were about twice the unigenes assembled by SOAPdenovo. As a general conclusion, it seems that selection of an appropriate software and its parameters is not easy without comprehension understanding of different software operation and their setting. Thus, comparison of the different softwares efficiency on the different organisms could provide some practical suggestions and choose an appropriate software.

Keywords: Saffron, Transcriptome, Assembly, Unigenes, Trinity, SOAPdenovo

چکیده

زعفران (*Crocus sativus* L.)، ارزشمندترین و محبوبترین ادویه جهانی، گونه‌ای تریپلوئید از خانواده Iridaceae می‌باشد. این گونه با کلالة‌های قرمز، بلند و معطر، از سایر گونه‌های این خانواده متمایز است. با پیشرفت‌های سریع بوجود آمده در نسل دوم توالی‌یابی، توالی‌یابی RNA (RNAseq) ابزار قدرتمندتر و ارزان‌تری در مطالعات ترانسکرپتوم شده است. گردآوری مجدد توالی‌های ترانسکرپتوم، راه حل مناسبی برای مطالعه ترانسکرپتوم موجوداتی است که توالی ژنوم آن‌ها در دسترس نیست. توالی‌یابی دقیق و مونتاژ قابل اعتماد در داده‌های ترانسکرپتوم، برای آنالیزهای پایین‌دست ضروری می‌باشد. در این مطالعه با جداسازی و تعیین توالی ترانسکرپتوم کلالة زعفران زراعی با استفاده از نسل دوم توالی‌یابی نتایج عملکرد تجزیه و تحلیل داده‌ها بوسیله دو نرم‌افزار پرکاربرد به نام‌های SOAPdenovo و Trinity مقایسه شد. نتایج این پژوهش نشان داد که میانگین طول توالی‌ها و تعداد یونی‌ژن‌های بدست آمده در Trinity بیشتر از SOAPdenovo می‌باشد (میانگین ۶۸۹ برای Trinity و ۶۲۴ برای SOAPdenovo). نتایج مونتاژ بهتر توسط مونتاژگر مناسب وقتی به پروتئین ترجمه می‌شود منجر به افزایش معنی‌داری در تعداد رکورد‌های به‌دست آمده در پایگاه‌های اطلاعاتی می‌شود و تعداد بیشتر یونی‌ژن امکان شناسایی مسیرهای بیوسنتزی متابولیت‌های مهم بیشتری را فراهم می‌کند. یونی‌ژن‌های مونتاژ شده در Trinity فاقد فاصله بوده و میانگین آن حدود دو برابر یونی‌ژن‌های مونتاژ شده بوسیله SOAPdenovo بود. به‌طور کلی انتخاب ابزار و پارامترهای مناسب برای مونتاژ ترانسکرپتوم بدون درک کاملی از عملکرد ابزارهای مختلف و تنظیمات آن‌ها کار مشکلی است. با مقایسه عملکرد نرم‌افزارهای مختلف بر روی موجودات مختلف می‌توان توصیه‌هایی در زمینه استفاده از آن‌ها ارائه داد و نرم‌افزار مناسب‌تر را انتخاب کرد.

واژه‌های کلیدی: زعفران، ترانسکرپتوم، مونتاژ، یونی‌ژن، Trinity، SOAPdenovo

مقدمه

زعفران (*Crocus sativus* L.) ارزشمندترین و محبوبترین ادویه جهانی، گونه‌ای تریپلوئید از خانواده Iridaceae می‌باشد. این گونه با کلاله‌های قرمز، بلند و معطر، از سایر گونه‌های این خانواده متمایز است و موجب شده که این گیاه از نظر اقتصادی نیز مورد توجه قرار گیرد (Fernández and Pandalai, 2004). در طی سال‌های گذشته ایران با داشتن حدود ۷۴ هزار هکتار سطح زیرکشت و تولید ۲۵۴ تن زعفران در سال بزرگترین تولیدکننده و صادرکننده زعفران در جهان است و بیش از ۸۵ تا ۹۰ درصد تولید جهانی این محصول گرانبها به ایران اختصاص دارد. در سال ۱۳۹۲ بالغ بر ۸ تن زعفران به ارزش ۲۰۰ میلیون دلار از کشور صادر شده است (بی‌نام، ۱۳۹۲). میزان تولید زعفران در ایران در مقایسه با سایر کشورهای تولیدکننده آن در جدول ۱ نشان داده شده است.

جدول ۱- مقایسه نسبت تولید سالانه زعفران (Fernández, 2007)

ایران	۲۵۴ تن
اسپانیا	۲۵ تن
هندوستان	۸-۱۰ تن
یونان	۸ تن
مراکش	۳ تن
ایتالیا	۶۰۰ کیلوگرم
ترکیه	۱۰ کیلوگرم

نتایج پژوهش‌های مختلف نشان داده است که زعفران خواص دارویی قابل توجهی از جمله نشاط‌آوری، آرام‌بخشی، کاهش چربی و فشار خون برای انسان دارد (Wani et al., 2011). خواص عمده این گیاه در کلاله آن و به دلیل وجود سه متابولیت مهم موجود در آن شامل کروسین^۱، پیکروکروسین^۲ و

سافرانا^۳ می‌باشد. این سه متابولیت به ترتیب مسئول رنگ، مزه و بوی زعفران می‌باشند. با وجود اهمیت اقتصادی و دارویی فراوان این ادویه مهم و ارزش، اطلاعات ژنومی اندکی در مورد آن در پایگاه‌های اطلاعاتی وجود دارد. تعداد پروتئین‌های شناسایی شده از زعفران و انواع گونه‌های وحشی آن ۱۵۷۱ و تعداد ESTهای موجود در پایگاه‌های اطلاعاتی برای گونه *Crocus sativus* تعداد ۶۷۶۸ عدد می‌باشد. با توجه به اهمیت این گیاه و لزوم بررسی دقیق مسیرهای بیوسنتز متابولیت‌های مهم آن به نظر می‌رسد که دستیابی به اطلاعات دقیق‌تر و کامل‌تر ژنومی و ترانسکریپتومی ضروری می‌باشد. ترانسکریپتوم، به سری کامل از ترانس کریپت‌ها یا توالی‌های رونویسی شده در یک سلول، یک مرحله خاص از رشد و نمو و یا تحت یک شرایط فیزیولوژیک خاص گفته می‌شود (Cahais et al., 2012). آنالیز ترانسکریپتوم برای تفسیر عملکرد عناصر ژنومی و آشکار سازی اجزاء اصلی ملکولی سلول‌ها و بافت‌ها ضروری است. بر این اساس ارزیابی ترانس کریپتوم یا توالی‌یابی EST یک رویکرد مناسب برای بدست آوردن اطلاعاتی در زمینه سطوح عملکردی ژنومی برای موجودات غیر مدل می‌باشد (Liu et al., 2013).

در سال‌های اخیر، آنالیز ترانسکریپتوم به عنوان یکی از مراحل بنیادی در مطالعات بیولوژیک مطرح شده است. روش‌های مختلفی برای آنالیز ترانسکریپتوم وجود دارد، که از آن جمله می‌توان: نورترن بلات، واکنش زنجیره‌ای پلیمرز رونوشت‌برداری معکوس (RT-PCR)، ریزآرایه (Microarray) و توالی‌یابی به روش‌های سنتی نام برد (Morozova and Ma Marra, 2008). هریک از روش‌های ذکر شده سختی‌ها و مشکلات مربوط به خود را دارند، از آن جمله می‌توان به

1. Crocin
2. Picrocrocin

پایگاه ها نسبتاً اندک است. با توجه به پیشرفت‌های سریع در تکنولوژی NGS این تکنولوژی به ابزار قدرتمندی در مطالعات ترانسکریپتوم تبدیل شده است (Zhao *et al.*, 2011). نرم‌افزارهای متفاوتی برای مونتاژ قرائت‌های حاصل از توالی‌یابی خام معرفی شده‌اند که بوسیله هم‌ردیفی قرائت‌ها بر روی ژنوم رفرنس، شناسایی ژن و میزان بیان آن را انجام می‌دادند (Duan *et al.*, 2012)، اما ژنوم رفرنس برای همه گیاهان به‌ویژه گیاهانی با ژنوم بزرگ و یا پلی‌پلوئید (ژنوم‌های گیاهی) در دسترس نیست. استفاده از تئوری مونتاژ *de novo* قرائت‌های خام (گردآوری مجدد^{۱۰}) به ترانسکریپتوم، این امکان را برای محققین ایجاد کرده که توالی کامل ترانسکریپتوم را بازسازی کنند، تمام ژن‌های بیان شونده در مرحله خاصی را شناسایی و دسته‌بندی کنند، ایزوفرم‌های مختلف را از هم تفکیک کنند و میزان بیان ترانسکریپت‌ها را اندازه بگیرند. علیرغم این مزایا، استفاده از مونتاژ *de novo* بحث‌هایی را همراه داشته است و مونتاژگر باید شرایطی را که در مورد مونتاژ ژنوم وجود ندارد، حل کند. یکسان نبودن بیان برای ترانسکریپت‌های مختلف در زمان آنالیز ترانسکریپتوم، بیان ایزوفورم‌های مختلف یک ژن، بیان یک ژن بدون نواحی هم‌پوشان بر دشواری‌های این روش مونتاژ می‌افزاید (Zhao *et al.*, 2011). با توجه به این موارد، علاوه بر اهمیت روش توالی‌یابی و تکنولوژی مورد استفاده در توالی‌یابی ترانسکریپتوم، نرم‌افزارهای مورد استفاده در گردآوری (مونتاژ) قرائت‌های ایجاد شده از اهمیت بالایی برخوردار هستند. در دهه اخیر، نرم‌افزارهای متعددی برای مونتاژ معرفی شده‌اند که از میان آنها می‌توان به SOAPdenovo (Li *et al.*, 2010)، trans-ABYSS (Simpson *et al.*, 2009)، Oases (Schulz *et al.*, 2009)

استفاده از مواد رادیواکتیو در نورترن بلات، کارایی پایین در روش واکنش زنجیره‌ای پلیمراز رونوشت‌برداری معکوس، عدم شناسایی رونوشت‌های جدید توسط ریزآرایه و هزینه، زمان و پیچیدگی بالای توالی‌یابی سنتی اشاره کرد. در مقابل روش‌های سنتی ذکر شده، نسل جدید توالی‌یابی در طی پروژه توالی‌یابی ژنوم انسان مطرح شد. نسل دوم توالی‌یابی، به منظور فراهم کردن پوشش بالای ژنومی^۱، تفکیک‌پذیری در حد تک تک جفت بازها^۲ و رفع معایب نسل اول توالی‌یابی (توالی‌یابی سنگر)، یک روش با کارایی بالا برای آنالیز اطلاعات ترانسکریپتوم است. نسل جدید توالی‌یابی در مقایسه با روش سنتی (سنگر)^۳، قطعات توالی‌یابی شده کوتاه بسیار زیادی تولید می‌کنند که منجر به کاهش هزینه توالی‌یابی و پیچیدگی‌های آزمایشگاهی می‌شود (Morozova *et al.*, 2009). نیز برای نسل دوم توالی‌یابی معرفی و بکاربرده شده- اند که از آن جمله می‌توان به ژنوم آنالیزر^۴ (Wu *et al.*, 2010) و هایسک^۵ (Xu *et al.*, 2012) از شرکت ایلومینا^۶، جی.اس-اف.ال.ایکس^۷ از شرکت روشه^۸ (Bai *et al.*, 2011) اشاره کرد. توالی‌یابی ترانسکریپتوم با استفاده از فناوری‌های جدید در کنار یک روش گردآوری^۹ کارآمد، منابع فراوان و مناسبی برای مطالعه پروفایل بیان ژن فراهم می‌کند (Morozova and MA Marra, 2008). در سال‌های اخیر، تعداد زیادی از پایگاه‌های اطلاعاتی EST برای موجودات مدل و غیر مدل در دسترس قرار گرفته‌اند، اما سهم گیاهانی مثل زعفران در این

1. Deep coverage
2. Single base-pair resolution
3. Sanger sequencing
4. Genome analyzer
5. HiSeqTM2000
6. Illumina
7. GS-FLX
8. Roche
9. Assembly

10. *de novo* Assembly

کشیده می‌شود که حجم زیادی دارد. از معایب دیگر این مونتاژگرها این است که با افزایش تعداد قرائت‌ها خوب کار نمی‌کنند و میزان اشتباه آنها بالا می‌رود، و به همین دلیل اکثر مونتاژگرهای نسل جدید توالی‌یابی از الگوریتم دوم (de bruijn graph) استفاده می‌کنند. در این روش مونتاژ از طریق شکستن توالی به توالی‌های کوچکتر؛ که k -mer نامیده می‌شود؛ انجام می‌شود. در این روش هم‌پوشانی را به جای قرائت‌های خام اصلی بین k -mer در نظر می‌گیرند. هر دو نرم‌افزار استفاده شده برای این مطالعه از الگوریتم de bruijn graph استفاده می‌کنند.

پژوهش حاضر با هدف بررسی کارایی نرم‌افزارهای Trinity و SOAPdenovo و انتخاب مونتاژگر مناسب برای مونتاژ قرائت‌های ترانسکریپتوم کلالة زعفران انجام شده‌است. برای این تحقیق، از کلالة زعفران زراعی در مرحله آخر رشد و نمو کلالة آن (قرمز رنگ)، در اواسط آبان‌ماه نمونه‌برداری شده و RNA آن پس از استخراج با استفاده از تکنولوژی توالی‌یابی دو طرفه ایلومینا (Illumina paired-end) توالی‌یابی و سپس بررسی شده‌است. نتایج حاصل از توالی‌یابی به صورت قرائت‌های خام برای گردآوری مجدد توسط دو نرم‌افزار مذکور استفاده شد و در نهایت نتایج مونتاژ این دو نرم‌افزار مقایسه و بررسی شد. برطبق اطلاعات ما، این پژوهش، اولین بررسی کامل ترانسکریپتوم زعفران و گردآوری مجدد توالی‌های خام با استفاده از استراتژی توالی‌یابی و گردآوری مجدد با نرم‌افزار می‌باشد.

مواد و روش‌ها

مواد گیاهی و استخراج RNA

در این مطالعه برای بررسی کارایی نرم‌افزارهای مونتاژگر در آنالیز ترانسکریپتوم زعفران، از کلالة زعفران زراعی قائن (به عنوان مرغوب‌ترین زعفران ایرانی از نظر متابولیت‌های ثانویه مهم زعفران) در

Trinity (Grabherr *et al.*, 2011) و *al.*, 2012) اشاره کرد. نرم‌افزار Oases از نسخه Velvet (Zerbino and Birney, 2008) آن طراحی شده و مثل Trinity اختصاصاً برای داده‌های ترانسکریپتوم ساخته شده‌است. نرم‌افزارهای trans-ABYSS، SOAPdenovo و Velvet برای مونتاژ ژنوم ساخته شدند. در میان نرم‌افزارهای معرفی شده، Trinity و SOAPdenovo برای مونتاژ نتایج دستگاه ایلومینا بیشتر استفاده شده‌اند، و به همین دلیل در مطالعه حاضر نیز از این دو نرم‌افزار استفاده شده‌است. نرم‌افزار SOAPdenovo به طور گسترده در مطالعات مختلف استفاده شده‌است، در حالی که نرم‌افزار Trinity در سال ۲۰۱۲ معرفی شد و بر خلاف سایر نرم‌افزارهایی که برای مونتاژ قرائت‌های کوتاه ژنومی استفاده می‌شوند، به طور اختصاصی برای آنالیز ترانسکریپتوم طراحی شده‌است (Grabherr *et al.*, 2011). با توجه به تفاوت‌های ترانسکریپتوم نسبت به ژنوم، نرم‌افزارها و روش‌های گردآوری مجددی که برای آنالیز داده‌های ترانسکریپتوم استفاده می‌شوند با نرم‌افزارها و روش‌های موجود برای آنالیز داده‌های ژنوم متفاوت هستند. در توالی‌یابی RNA به‌ویژه در موجوداتی با حداقل اطلاعات ژنتیکی بزرگترین چالش در این زمینه می‌باشد. با این وجود بیشتر تلاش‌هایی که در این زمینه انجام شده بر اساس تغییراتی در فناوری‌های نرم‌افزارهای موجود برای گردآوری مجدد ژنوم بوده‌است.

نرم‌افزارهای مورد استفاده برای مونتاژ قرائت‌های حاصل از توالی‌یابی از دو الگوریتم مختلف استفاده می‌کنند: overlap graph و de bruijn graph. اکثر مونتاژگرهایی که برای قرائت توالی‌های سنگر استفاده شدند از نوع اول هستند (Zeng *et al.*, 2013). در این نرم‌افزارها همه جفت‌های هم‌پوشان به صورت نمودار نشان داده می‌شوند و در این روش حتی در مورد موجودات ساده نیز هزاران نمودار

در نهایت، کتابخانه سی.دی.ان.آ بدست آمده توسط دستگاه Illumina HiSeq™ 2000 با استفاده از فناوری Paired-end طبق دستورالعمل کمپانی توالی‌یابی شد.

آنالیز نتایج توالی‌یابی

نتایج حاصل از دستگاه توالی‌یابی Illumina HiSeq™ 2000 بصورت عکس می‌باشد. بنابراین اولین مرحله در آنالیز داده‌های بدست آمده تبدیل این عکس‌ها به توالی است که از طریق فراخوانی باز انجام شد. توالی‌های بدست آمده قرائت خام^۴ نامیده می‌شوند که در قالب فستکیو^۵ ذخیره شده‌اند. در این قالب هر قرائت خام با یک @ شروع می‌شود. خطوط اول و سوم نام توالی را نشان می‌دهد که توسط دستگاه توالی‌یابی ایجاد شده است. خط دوم شامل توالی قرائت خام بدست آمده است و خط چهارم حروفی هستند که بیانگر کیفیت نوکلئوتیدهای قرائت شده در هر توالی می‌باشند. در واقع هر حرف از خط چهارم مربوط به یکی از بازهای خط دوم است. ارزش کیفی توالی‌یابی در دامنه ۲-۳۵ قرار دارد. داده خام حاوی قرائت‌های نامطلوبی^۶ است که می‌تواند در آنالیزهای بیوانفورماتیکی بعدی اثر منفی داشته باشد. بنابراین لازم است این قرائت‌های نامطلوب که شامل آدپتورها، قرائت‌هایی که بیش از ۵٪ طول خود، نوکلئوتیدهای ناشناخته دارند و قرائت‌های با کیفیت پایین هستند، حذف شوند. توالی‌های باقی‌مانده با معیار Q20 (معیاری برای سنجش کیفیت توالی‌ها)، سنجیده شده و توالی‌های خام با کیفیت مناسب برای مونتاژ استفاده شدند. از نرم‌افزارهای Trinity (trinityrnaseq.sourceforge.net) و SOAPdenovo (soap.genomics.org) برای مونتاژ این قرائت‌های خام استفاده شدند. نحوه کار با

مرحله آخر رشد و نمو کلالة آن (قرمز رنگ)، استفاده شده است. نمونه‌برداری در اوایل آبان‌ماه سال ۱۳۹۰ و در زمان گلدهی زعفران از زعفران قائن کشت شده در مزرعه تحقیقاتی پژوهشکده بیوتکنولوژی کشاورزی ایران واقع در کمال‌آباد کرج انجام شد. پس از نمونه‌برداری، کلالة‌ها بلافاصله در ازلت مایع قرار داده شدند و پس از انتقال به آزمایشگاه تا زمان استخراج RNA در دمای ۸۰°C - نگهداری شدند. استخراج RNA با روش TRIZOL (invitrogen) انجام شد. پس از استخراج RNA، تیمار با DNaseI در دمای ۳۷°C برای ۳۰ دقیقه انجام شد. کیفیت و کمیت RNA استخراج شده توسط ژل آگارز ۱٪ و نانودراپ^۱ بررسی شد. پس از تأیید کمیت و کیفیت RNA، مقدار مناسب ۳۰ ماکروگرم (یک ماکروگرم در میکرولیتر) برای توالی‌یابی استفاده شد.

ساخت کتابخانه cDNA و توالی‌یابی به وسیله فناوری Illumina HiSeq™ 2000

در توالی‌یابی، برای جداسازی mRNA های دارای دم پلی‌آ از Oligo (dT) استفاده شد. در زمان سنتز cDNA، mRNA ها توسط بافر تجزیه‌کننده به قطعات کوچک ۷۰۰-۲۰۰ نوکلئوتیدی شکسته شدند. این قطعات به عنوان الگو برای سنتز رشته اول cDNA توسط هگزامرهای تصادفی (N6) و آنزیم ترانسکریپتاز معکوس^۲ سنتز استفاده شد. دومین رشته cDNA با استفاده از RNase H و آنزیم DNA پلیمراز I ساخته شد. این قطعات کوچک cDNA خالص‌سازی شده و به منظور افزودن زنجیره پلی‌آ درون بافر EB^۳ حل شدند. سپس این قطعات کوچک به آدپتورهای توالی‌یابی متصل شده و بعد از الکتروفورز ژل آگارز، قطعات مناسب به عنوان الگو برای تکثیر در پی.سی.آر انتخاب شدند.

4. Raw Read
5. Fastq
6. Dirty Read

1. NanoDrop
2. Reverse Transcriptase
3. Elusion Buffer

می‌باشد. نرم‌افزار Trinity از سه نرم‌افزار مستقل به نام‌های Inchworm، Chrysalis و Butterfly به صورت متناوب برای مونتاژ قرائت‌های خام استفاده می‌کند. هر دو نرم‌افزار، ابتدا قرائت‌هایی را که در طول مشخصی همپوشانی دارند ترکیب می‌کند تا قطعات بزرگتری بسازد که این قطعات کانتیگ نامیده می‌شوند. سپس قرائت‌ها مجدداً روی کانتیگ‌ها برگردانده می‌شوند تا یونی‌ژن‌هایی ساخته شوند که از دو طرف نتوانند بزرگتر شوند. دو طرفه بودن قرائت‌ها امکان شناسایی کانتیگ‌های مرتبط به یک رونوشت و همچنین فاصله بین کانتیگ‌ها را فراهم می‌کند. با توجه به این اطلاعات، نرم‌افزار کانتیگ‌ها را به هم متصل کرده و توالی‌هایی ایجاد می‌کند که از هیچ یک از دو انتها قابل گسترش نیستند، این توالی‌ها یونیژن^۹ نامیده می‌شوند. خروجی هر دو دستگاه به صورت فایل فست‌ای^{۱۰} می‌باشد که به فایل اکسل انتقال داده شده و برای محاسبات آماری به کار می‌رود. تعداد یونی‌ژن مونتاژ شده، میانگین طول آنها، N۵۰، درصد یونی‌ژن‌های با کیفیت پایین و تعداد یونی‌ژن‌های با اندازه بلند ($\geq 1\text{kb}$) در دو نرم‌افزار بررسی و مقایسه شده‌اند.

یونی‌ژن‌های بدست آمده توسط هر دو نرم‌افزار به صورت جداگانه در پایگاه اطلاعاتی NCBI BLASTx (Nr) (www.ncbi.nlm.nih.gov)، و نرم‌افزار Blast2GO بررسی شد (Conesa et al., 2005). نتایج بررسی در پایگاه NCBI به صورت Similarity و e-Value (حد آستانه 10^{-6} برای تشابه با توالی‌های موجود در پایگاه NCBI انتخاب شد) برای هر دو نرم‌افزار بررسی و مقایسه شد. علاوه بر آن برای بررسی کارایی نتایج مونتاژ این دو نرم‌افزار، ژن‌هایی که همیشه در سلول بیان^{۱۱}

هر دو نرم‌افزار به صورت فرمان خطی^۱ و فایل ورودی به فرمت فستکیو می‌باشد. قرائت‌های خام به صورت *de novo* توسط دو نرم‌افزار با پارامترها (*k-mer*) و شرایط سیستم کامپیوتر مشابهی (رم^۲ ۳۲G، سی‌پی‌یو^۳ Core i8 ۳/۲ و سیستم عامل Linux SUSE v12) مونتاژ شدند. در نرم‌افزارهای توالی‌یابی که از الگوریتم de Bruijn استفاده می‌کنند، *K* مهمترین پارامتر می‌باشد. که به اندازه قطعات خرد شده برای مونتاژ توسط دستگاه گفته می‌شود. تکرارهای با اندازه بزرگتر از *K* نمودارها را به هم ریخته و باعث خرد شدن در کانتیگ‌ها^۴ می‌شوند. کوچک‌تر و یا بزرگ‌تر انتخاب کردن *k* مزایا و معایبی دارد، در نهایت در عمل *K* بر اساس تجربیات و مشاهدات قبلی با سری داده‌های مشابه انتخاب می‌شود (Chikhi and Medvedev, 2013). هر نرم‌افزار بر اساس یک تک *k-mer* اجرا می‌شوند. در Trinity فقط یک *k* به صورت پیش‌زمینه برای اجرای دستگاه وجود دارد ($k: 25$) و قابل تغییر نیست، اما در SOAPdenovo تعداد *k* را برای کسب بهترین نتیجه می‌توان تغییر داد^۵ و نتایج بدست آمده را با هم مقایسه کرد. خروجی در SOAPdenovo بر اساس سه مرحله مونتاژ و تولید: کانتیگ، اسکفولد^۶ و یونی‌ژن^۷ می‌باشد در حالیکه خروجی در Trinity یک مرحله کمتر داشته و شامل: کانتیگ و یونی‌ژن می‌باشد. در SOAPdenovo جای خالی^۸ در توالی‌ها دیده می‌شود اما در Trinity توالی‌ها جای خالی ندارند که این مسئله به دلیل الگوریتم متفاوت مورد استفاده در این دو نرم‌افزار

1. Command line
2. RAM
3. CPU
4. Contig
5. Multiple *K*
6. Scaffold
7. Unigene
8. Gap

9. Unigene

10. Faft A

11. Housekeeping genes

شرایط یکسان، زمان لازم برای اجرای Trinity حداقل ۲۰ برابر بیشتر از SOAPdenovo می باشد که دلیل بالاتر بودن زمان طولانی بودن مرحله اول کار نرم افزار (Inchworm) می باشد که نصف بیشتر زمان را به خود اختصاص می دهد. مدت زمان بیشتر مورد نیاز برای آنالیز و بدست آمدن نتایج در Tirnity نسبت به چندین نرم افزار دیگر در مطالعه Zhao و همکاران تایید شده بود (Zhao et al., 2011).

جدول ۲- نتایج توالی یابی کلالة زعفران زراعی توسط

فناوری Illumina HiSeq™ 2000

تعداد کل قرائت ها	52075128
تعداد کل نوکلئوتیدها (nt)	4686761520
میانگین قرائت ها	90bp
درصد GC	44.22%
درصد Q20	99.67%

مقایسه یونی ژن ها مونتاژ شده توسط دو نرم افزار مورد مطالعه

برای مقایسه دو نرم افزار، خروجی نهایی آنها به صورت یونی ژن با هم مقایسه شد. تعداد k در این نرم افزار Trinity به صورت پیش زمینه ۲۵ قرار داده شد، اما برای SOAPdenovo که قابلیت تغییر k را دارد مونتاژ با تعداد سه k مختلف نرم افزار اجرا شد و بهترین نتایج به دست آمده با نتایج Trinity مقایسه شد (جدول ۲).

تعداد یونی ژن مونتاژ شده توسط Trinity حدود دو برابر بیشتر از SOAPdenovo بود. بررسی طول کانتیگ ها و یونی ژن های مونتاژ شده توسط دو نرم افزار نشان داد که میانگین طول کانتیگ ها و یونی ژن ها در Trinity به ترتیب ۶۰۸ و ۶۸۹ و در SOAPdenovo به ترتیب ۶۰۴ و ۶۲۴ می باشد (شکل ۱ و جدول ۲). بین وجود رکورد در پایگاه های اطلاعاتی و طول یونی ژن های مونتاژ شده ارتباط مثبتی وجود دارد و در یونی ژن های بلندتر امکان

می شوند در یونی ژن های به دست آمده از هر دو نرم افزار شناسایی و از نظر درصد شباهت با توالی های مشابه موجود در پایگاه های اطلاعاتی با هم مقایسه شدند.

نتایج و بحث

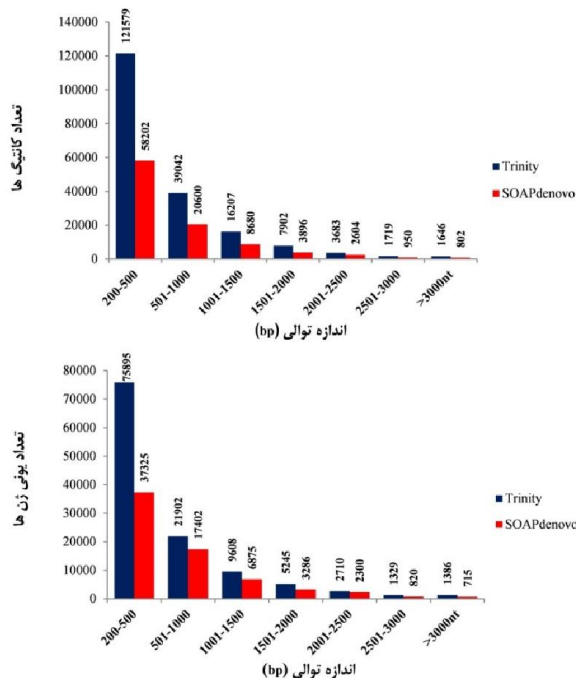
استفاده از نسل جدید توالی یابی برای بررسی ترانسکریپتوم موجودات زنده مدل و غیر مدل از سال های ۲۰۰۵ و ۲۰۰۶ پس از تجاری شدن دستگاه های شرکت های مختلف مثل Illumina SOLiD /ABI/، Science Roch/454Life و Solexa آغاز شده است و تا کنون تعداد زیادی موجود مدل و غیر مدل با استفاده از این فناوری توالی یابی شده اند. آنالیز داده های حاصل از توالی یابی ترانسکریپتوم بر اساس وجود یا عدم وجود ژنوم یا ترانسکریپتوم مرجع برای موجود زنده مورد بررسی، متفاوت خواهد بود. برای آنالیز داده های ترانسکریپتوم موجودات زنده ای که ژنوم مرجعی برای آنها وجود ندارد یا هنوز تکمیل نشده اند از روش گردآوری مجدد استفاده می شود. در این مطالعه از دو نرم افزار Trinity و SOAPdenovo برای مونتاژ قرائت های خام توالی یابی شده کلالة زعفران توسط فناوری Illumina HiSeq™ 2000 استفاده شد. این دو نرم افزارهای رایج ترین نرم افزار در مونتاژ *de novo* هستند. هر دو نرم افزار با کامپیوتر با شرایط یکسان اجرا شدند. در نهایت تعداد ۴۶۸۶،۷۶۱،۵۲۰ نوکلئوتید و ۵۲،۰۷۵،۱۲۸ قرائت خام به طول ۹۰bp بدست آمد (جدول ۱). درصد GC در این توالی یابی ۴۴/۲۲٪ بود (جدول ۲). نتایج مونتاژ داده ها نشان داد که بطور کلی برای مونتاژ حجم بیشتری از قرائت های خام در موجودات با ترانسکریپتوم بزرگتر در هر دو نرم افزار به حافظه بیشتری نیاز است، در مطالعات پیشین نیز نتیجه مشابهی بدست آمده بود (Zhao et al., 2011). مقایسه طول زمان مورد نیاز برای مونتاژ در این دو نرم افزار نشان داد که در

SOAPdenovo در گیاه بامبو نیز گزارش شده بود (Liu *et al.*, 2012). تعداد یونی‌ژن‌های بالای ۱kb، طول بزرگ‌ترین و کوچک‌ترین یونی‌ژن مونتاژ شده و تعداد یونی‌ژن منحصر به فرد مونتاژ شده توسط هر دو نرم‌افزار در جدول ۲ مقایسه شده‌اند.

وجود توالی مشابه بیشتری در پایگاه‌های اطلاعاتی دارند. علاوه بر آن بین طول یونی‌ژن‌ها و کانتیگ‌های مونتاژ شده و تعداد آن‌ها رابطه عکس وجود دارد (شکل ۱). بیشتر بودن میانگین طول توالی‌های مونتاژ شده توسط Trinity نسبت به

جدول ۲- مقایسه ویژگی‌های مونتاژ دو نرم‌افزار Trinity و SOAPdenovo

SOAPdenovo	Trinity	
68723	118075	تعداد یونی‌ژن‌ها
4686761520	4686761520	کل بازها
624	689	میانگین طول ترانسکریپت
875	951	N50
285	245	تعداد یونی‌ژن‌ها (≥ 1 kb)
42.2	42.2	درصد GC
7541	7696	طول بزرگ‌ترین یونی‌ژن (bp)
226	201	طول کوچک‌ترین یونی‌ژن (bp)
20822	66171	تعداد یونی‌ژن با تفسیر در پایگاه‌های اطلاعاتی
1725	3154	تعداد یونی‌ژن‌های منحصر به فرد



شکل ۱- مقایسه طول کانتیگ‌ها (شکل بالا) و یونی‌ژن‌های (شکل پایین) مونتاژ شده توسط Trinity و SOAPdenovo. در هر دو مورد ارتباط منفی بین تعداد کانتیگ‌ها و یونی‌ژن بدست آمده و طول آن وجود دارد و با افزایش طول کانتیگ‌ها تعداد آن‌ها کمتر می‌شود.

BLAST (Nr)، نشان داد که (۶۶۱۷۱) ۵۶٪ و (۲۰۸۲۲) ۳۰٪ از توالی‌های مونتاژ شده توسط

یونی‌ژن‌های به‌دست آمده از دو نرم‌افزار مورد مطالعه برای شناسایی و تفسیر در پایگاه اطلاعاتی

۴۰ تا ۸۰ درصد با توالی‌های ثبت شده موجود در پایگاه‌های اطلاعاتی داشتند (شکل ۲). همانطور که در شکل نیز مشخص است درصد بالاتری از توالی‌های مونتاژ شده با Trinity شباهت بیشتری با توالی‌های موجود در پایگاه‌های اطلاعاتی دارند.

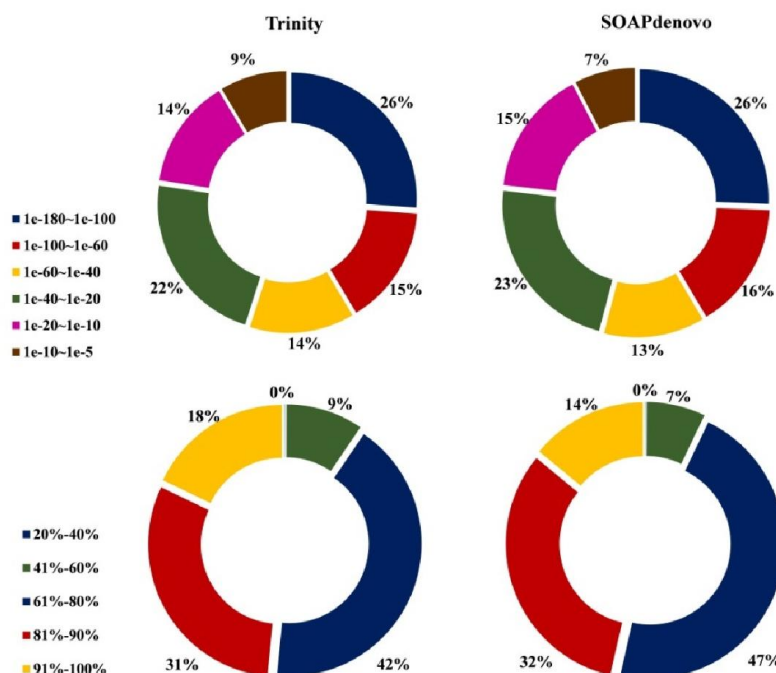
میانگین طول توالی‌های مونتاژ شده در Trinity بلندتر از SOAPdenovo می‌باشد و تعداد یونی‌ژن‌های مونتاژ شده با آن دو برابر SOAPdenovo می‌باشد، تفسیر یونی‌ژن‌ها اطلاعات مهمی از بیان و تفسیر عملکردی آن‌ها فراهم می‌کند و تولید تعداد بیشتری یونی‌ژن امکان شناسایی ژن‌های بیشتری در مسیرهای مهم شناسایی نشده در زعفران را فراهم می‌کند. این نتایج با بررسی مشابه توسط این دو نرم‌افزار روی گیاهان دیگر مطابقت دارد (Cahais *et al.*, 2012).

نتایج بررسی ژن‌های خانه‌دار مونتاژ شده توسط دو نرم‌افزار نشان داد که با وجود تعداد بیشتر یونی‌ژن‌های مونتاژ شده مربوط به ژن‌های خانه‌دار در نتایج SOAPdenovo، اما درصد کمتری (۴۱٪) از آن‌ها شباهت کامل با ORF توالی‌های مشابه در پایگاه‌های اطلاعاتی داشتند در حالی که این درصد برای Trinity بالاتر بود (۷۴٪) (تعداد یونی‌ژن‌های دارای میزان شباهت کامل / تعداد کل یونی‌ژن‌های ژن‌های خانه‌دار) (جدول ۳). در واقع توالی‌های تعداد بیشتری از این ژن‌ها در نتایج مونتاژ Trinity شباهت کاملی با توالی رفرنس این ژن‌ها در پایگاه‌های اطلاعاتی دارد. ژن‌های خانه‌دار در موجودات مختلف توالی حفاظت شده و نسبتاً ثابتی دارند، شباهت کامل تعداد بیشتری از یونی‌ژن‌های مونتاژ شده با توالی این ژن‌ها می‌تواند تاییدی بر مناسب بودن نتایج گردآوری توسط نرم‌افزارهای مورد استفاده در این مطالعه باشد.

Trinity و SOAPdenovo دارای رکورد در پایگاه‌های اطلاعاتی هستند. همچنین، تعداد یونی‌ژن‌های منحصر به فرد شناسایی شده در Trinity بیشتر از SOAPdenovo بود (جدول ۲). نتایج مونتاژ بهتر توسط مونتاژگر مناسب وقتی به پروتئین ترجمه می‌شود منجر به افزایش معنی‌داری در تعداد رکوردهای به‌دست آمده در پایگاه‌های اطلاعاتی می‌شود و علاوه بر آن با افزایش تعداد یونی‌ژن‌ها تعداد بیشتری از آن‌ها امکان دارد که دارای رکورد در پایگاه‌های اطلاعاتی باشند. تعداد بیشتری از توالی‌های مونتاژ شده توسط Trinity شباهت بالای ۸۰ درصد با توالی‌های ثبت شده در NCBI دارند. که این تعداد یونی‌ژن منحصر به فرد مونتاژ شده در Trinity در شناخت کامل مسیرهای بیوسنتز متابولیت‌های مهم ممکن است کمک نمایند (Shi *et al.*, 2011).

مقایسه نتایج تفسیر یونی‌ژن‌های زعفران نشان داد که (در هر دو مونتاژگر) به ترتیب انگور قرمز، سویا^۱ و درخت صنوبر^۲ بیشترین میزان رکورد مشابه با توالی‌های زعفران را دارد. مقایسه توالی‌های از نظر e-Value نشان داد ۵۵٪ از توالی‌های مونتاژ شده توسط Trinity و SOAPdenovo دارای e-Value بالای ۴۰-e-۱ می‌باشند (شکل ۲). نتایج مقایسه میزان شباهت توالی‌ها با توالی‌های ثبت شده در پایگاه‌های اطلاعاتی (NCBI) نشان داد که در توالی‌های مونتاژ شده توسط Trinity تعداد ۲۵۸۰۶ (۳۹٪) از توالی‌ها شباهتی بین ۸۰ تا ۱۰۰ درصد، تعداد ۳۳۷۴۷ (۵۱٪) شباهتی بین ۴۰ تا ۸۰ درصد با توالی‌های موجود در پایگاه‌های اطلاعاتی دارند. در توالی‌های مونتاژ شده توسط SOAPdenovo تعداد ۲۰۶۱۰ (۳۰٪) از توالی‌ها شباهتی بین ۸۰ تا ۱۰۰ درصد، و تعداد ۴۴۶۶۹ (۶۴٪) از توالی‌ها شباهتی بین

-
1. *Vitis vinifera*
 2. *Glycine max*
 3. *Populus trichocarpa*



شکل ۲- مقایسه e-Value و Similarity در دو نرم‌افزار Trinity و SOAPdenovo. شکل‌های بالا: مقایسه e-Value در نتایج گردآوری دو نرم‌افزار برای یونی‌ژن‌های با e-Value کمتر از 10^{-5} ؛ شکل‌های پایین: مقایسه میزان شباهت (به درصد) یونی‌ژن‌ها با توالی‌های موجود در پایگاه NCBI.

جدول ۳- مقایسه ناحیه ORF ژن‌های با بیان دائم در سلول از نظر درصد تشابه با توالی ژن‌های موجود در پایگاه‌های اطلاعاتی

نام ژن	Trinity		SOAPdenovo	
	تعداد یونی‌ژن‌ها	تعداد یونی‌ژن‌ها با ORF کامل	تعداد یونی‌ژن‌ها	تعداد یونی‌ژن‌ها با ORF کامل
actin	7	6	11	7
histone	11	11	30	11
glyceraldehyde-3-phosphate dehydrogenase	2	1	4	1
phosphofructokinase	4	3	11	1
tubulin	19	12	22	13
ubiquitin	3	3	25	10
کل	46	36(74%)	103	43 (41%)

آن‌ها در گیاهان مختلف می‌توان ارائه داد.

سپاسگزاری

نویسندگان مقاله از آقایان احمد موسوی و بابک عارف‌نژاد که در انجام آنالیزهای مورد نیاز همکاری داشتند تشکر و قدردانی می‌نماید. هزینه پژوهش حاضر توسط پژوهشکده بیوتکنولوژی کشاورزی ایران تأمین شده است.

از مرحله مونتاژ توالی تا داده‌های واقعی نیازمند الگوریتم‌های مناسبی برای تولید بهترین نتایج است. در حال حاضر یک الگوریتم و راه حل ثابتی برای همه داده‌ها وجود ندارد. به‌طور کلی انتخاب ابزار و پارامترهای مناسب برای مونتاژ ترانسکریپتوم بدون درک کاملی از عملکرد ابزارهای مختلف و تنظیمات آن‌ها کار مشکلی است. با مقایسه عملکرد نرم‌افزارهای مختلف توصیه‌هایی در زمینه استفاده از

REFERENCES

- Anonymous, Government Information Center,
<http://dolat.ir/NSite/FullStory/News>
- Bai XL, Rivera-Vega P, Mamidala P, Bonello DA, Herms O (2011) Transcriptomic signatures of ash (*Fraxinus* spp.) phloem. *PloS one* 6:e16368.
- Bentley DR (2006) Whole-genome re-sequencing. *Current opinion in genetics and development* 16:545-552.
- Cahais VP, Gayral G, Tsagkogeorga J, Melo Ferreira M, Ballenghien L, Weinert Y, Chiari K, Belkhir V, Galtier N (2012) Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Molecular Ecology Resources* 12:834-845.
- Conesa AS, Götz JM, García-Gómez J, Terol M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674-3676.
- Chikhi R, Medvedev P (2012) Informed and Automated k-Mer Size Selection for Genome Assembly. Oxford University Press. 7:564-673.
- Duan JC, Xia G, Zhao J, Kong K (2012) Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. *BMC Genomics*. 13:392-404.
- Fernández JA, Pandalai S (2004) Biology, biotechnology and biomedicine of saffron. *Recent Research Developments in Plant Science*. 2:127-159.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng O (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*. 29:644-652.
- Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*. 20:265-272.
- Liu M, Qiao G, Jiang J, Yang H, Xie L, Xie J, Zhuo R (2012) Transcriptome sequencing and de novo analysis for ma bamboo (*Dendrocalamus latiflorus* Munro) using the Illumina platform. *PloS one* 7:e46766.
- Liu T, Zhu S, Tang Q, Chen P, Yu Y, Tang S (2013) De novo assembly and characterization of transcriptome using Illumina paired-end sequencing and identification of Cesa gene in ramie (*Boehmeria nivea* L. Gaud). *BMC Genomics* 14:125-136.
- Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. *Annual Review of Genomics and Human Genetics*. 10:135-151.
- Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*. 92:255-264.
- Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*. 28:1086-1092.
- Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T (2011) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics*. 12:131-150.
- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*. 19:1117-1123.
- Wani BA, Hamza AKR, Mohiddin F (2011) Saffron: A repository of medicinal properties. *J. Med. Plant Res*. 5:2131-2135.

- Wu J, Zhang Y, Zhang H, Huang H, Folta KM, Lu J (2010) Whole genome wide expression profiles of *Vitis amurens* grape responding to downy mildew by using Solexa sequencing technology. *BMC Plant Biology*. 10:234.
- Xu DL, Long H, Liang JJ, Zhang J, Chen X, Li JL, Pan ZF, Deng GB, Yu MQ (2012) De novo assembly and characterization of the root transcriptome of *Aegilops variabilis* during an interaction with the cereal cyst nematode. *BMC Genomics*. 13:133-142.
- Zeng J, Liu Y, Liu W, Liu X, Liu F, Huang P, Zhu P, Chen J, Shi M, Guo F (2013) Integration of transcriptome, proteome and metabolism data reveals the alkaloids biosynthesis in *Macleaya cordata* and *Macleaya microcarpa*. *PloS one* 8:e53409.
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18:821-829.
- Zhao QY, Wang Y, Kong YM, Luo D, Li X, an Hao P (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* 12:S2.
- Zheng X, Pan C, Diao Y, You Y, Yang C, Hu C (2013) Development of microsatellite markers by transcriptome sequencing in two species of *Amorphophallus* (Araceae). *BMC Genomics* 14:490-501.