دانشگاه علوم کشاورزی و منابع طبیعی ساری
پژوهشهای تولیدات دامی

"مقاله پژوهشی"

مقایسه برخی نرم‌افزارهای مکان‌یابی در آنالیز داده‌های RNA-Seq گاو شیری

قربان الیاسی زرین‌قبایی[۱]، مصطفی صادقی[۲] و سیدرضا میرایی آشتیانی[۳]

۱- دانشجوی دکترای تخصصی، گروه علوم دامی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران، (نویسنده مسوول: Gh.elyasi@ut.ac.ir)
۲- دانشیار گروه علوم دامی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران
۳- استاد گروه علوم دامی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران
تاریخ دریافت: ۱۴۰۱/۶/۲۱ تاریخ پذیرش: ۱۴۰۱/۹/۲۷
صفحه: ۱۳۱ تا ۱۳۸

چکیده مبسوط

مقدمه و هدف: با توجه به کاربرد روزافزون تعیین توالی نسل جدید (NGS)، جهت شناسایی ژن‌های عملکردی، استفاده از الگوریتم‌ها و نرم‌افزارهای تخصصی برای انجام آنالیزهای آماری ضروری است. مکان‌یابی خوانش‌ها با ژنوم مرجع اولین و مهم‌ترین مرحله اکثر برنامه‌های آنالیز داده‌های RNA-Seq است که به‌صورت مؤثری صحت آنالیزهای پایین دستی بستگی به این مرحله دارد. لذا هدف از این تحقیق، مقایسه برخی نرم‌افزارهای مختلف هم‌ترازی داده‌های حاصل از تعیین توالی کل RNA روی ژنوم مرجع بود.

مواد و روش‌ها: از داده‌های RNA-Seq مربوط به ۵۴ رأس گاو شیری هلشتاین به‌منظور شناسایی ژن‌های مؤثر در باروری استفاده شد. تعیین کیفیت خوانش‌ها توسط نرم‌افزار FastQC و ویرایش توالی‌های کم کیفیت با استفاده از نرم‌افزار Trimmomatic انجام شد. داده‌های ویرایش شده با آخرین نسخه ژنوم مرجع گاو با استفاده از نرم‌افزارهای Bowtie2، Tophat2 و Hisat2 مکان‌یابی شدند. درصد کل خوانش‌های مکان‌یابی شده، درصد خوانش‌های مکان‌یابی شده روی یک محل در ژنوم مرجع و همچنین درصد خوانش‌های مکان‌یابی شده روی بیش از یک محل محاسبه شدند.

یافته‌ها: نتایج نشان داد که بیشترین هم‌ترازی صورت گرفته روی ژنوم مرجع گاو با استفاده از نرم‌افزار Tophat2 بوده است. از کل خوانش‌های موجود، نرم افزارهای Tophat2 و Hisat2 به ترتیب ۹۴/۱۹۷ و ۹۲/۵۲۶ درصد را روی ژنوم مرجع مکان‌یابی نمودند. نرم‌افزار Hisat2 عملکرد تخصیصی بیشتری داشت و ۸۹/۲۰۲ درصد از داده‌ها را به یک جایگاه اختصاصی مکان‌یابی کرد درصورتی‌که این پارامتر در خصوص نرم‌افزار Tophat2 به میزان ۸۷/۸۱۲ درصد از کل توالی‌ها بود. از کل توالی‌های به کار گرفته شده تنها ۳/۳۲۴ درصد توسط Hisat2 و ۶/۳۸۵ درصد توسط Tophat2 با بیش از یک جایگاه ژنوم مرجع مکان یابی شدند. نرم‌افزار Bowtie2 در مقایسه با دو نرم‌افزار دیگر عملکرد پایینی داشت.

نتیجه‌گیری: مقایسه نرم‌افزارهای هم‌ترازی داده‌های RNA-Seq روی ژنوم مرجع نشان داد که اگر چه نرم‌افزار Hisat2 بهترین نرم‌افزار برای مکان‌یابی خوانش‌ها بود بود ولی نرم افزار Tophat2 هم می‌تواند به جای آن در آنالیز داده های RNA-seq استفاده شود. در ضمن نرم‌افزار Bowtie2 در ارتباط با داده‌های RNA-Seq کارآیی چندانی ندارد.

واژه‌های کلیدی: اومیکس، بیان ژن، تعیین توالی، ژنوم مرجع و مکان‌یابی

مقدمه

خلق فناوری‌های اومیکس[۱] مانند ژنومیکس، ترانسکریپتومیکس، پروتئومیکس و متابولومیکس، آنالیز هم‌زمان تعداد زیادی از ژن‌ها، رونوشت‌ها، پروتئین‌ها و متابولیت‌ها را در بسیاری از آزمایشگاه‌ها تسهیل کرده است؛ همچنین حجم زیادی از داده‌ها در مورد ساختار سلول‌ها و رفتار آن‌ها در سطوح مختلف سلولی و شرایط محیطی مختلف تولید شده و بازسازی شبکه‌های مولکولی زیستی (به‌عنوان مثال: شبکه‌های تنظیم رونویسی، شبکه‌های اینتراکتوم، شبکه‌های متابولیک، شبکه‌های متقابل پروتئین-پروتئین) برای انجام تحلیل زیستی عمیق‌تر را فراهم ساخته است. همراه با تولید اطلاعات اومیکس، پلتفرم‌های آنالیزی به‌منظور ترسیم شبکه مولکولی زیستی توسعه پیدا کرد که بصورت ریاضی و محاسباتی، داده خام را پردازش کردند. انواع داده‌های مختلف در یک روش معنی‌دار زیست‌شناختی یکپارچه‌سازی و مرتب شده و درنهایت در یک سیستم برای بررسی درستی توصیف توابع و رفتار سلولی در کنار هم قرار داده شدند (۱۱)؛ بنابراین نیاز به یک رویکرد جامع برای رمزگشایی مقدار زیادی از داده‌های تولید شده با رویکردهای زیستی مدرن وجود دارد.

در حال حاضر داده‌های زیستی متنوعی در پایگاه‌های بزرگ در دسترس است که شامل پروفایل بیان ژن (RNA-Seq، Microarray، EST و SAGE) و اطلاعات عملکردی ژن‌ها و پروتئین‌ها است. مرکز ملی اطلاعات بیوتکنولوژی[۲]، موسسه بیوانفورماتیک اروپا[۳] و پایگاه داده DNA ژاپن[۴] چند پایگاه اصلی تعاملی داده را تشکیل می‌دهند که در تحقیقات زیستی به‌طور گسترده‌ای مورد استفاده قرار گرفته‌اند، در این پایگاه‌های داده، اطلاعاتی در مورد تعیین توالی نوکلئوتیدها، پروتئین‌ها، ژن‌ها، ژنوم، ساختار مولکولی و بیان ژن ارائه شده است. همانند پایگاه داده توالی نوکلئوتید، پایگاه اولیه و اصلی مانند Uniport اطلاعاتی در خصوص توالی پروتئین و حاشیه‌نویسی فراهم می‌کنند و همچنین بانک اطلاعات پروتئین[۵] روی ساختارهای پروتئینی متمرکز شده است. اخیراً آشکار شده است که RNAهای غیر کد شونده[۶] از جمله میکرو RNAها در کنترل سیستم سلولی بسیار اهمیت دارند که موجب پیاده‌سازی پایگاه‌های مرتبط مانند RNAdb و miRBase شده است. پایگاه داده دائره‌المعارف ژن و ژنوم کیوتو[۷] (KEGG)، Reactome و BioCyc پایگاه‌های تعاملاتی پیشرو هستند که دارای مسیرهای واکنش‌های متابولیک و مسیرهای انتقال سیگنال می‌باشند. KEGG یک منبع دانش محور است که اطلاعاتی در خصوص ژن‌ها و پروتئین‌ها، اجزای بیوشیمیایی، واکنش‌ها و مسیرها ارائه می‌دهد. پایگاه Reactome توسط همکاری آزمایشگاه کولد اسپرینگ هاربر[۸]، موسسه بیوانفورماتیک اروپا و کنسرسیوم هستی‌شناسی ژن[۹] مدیریت می شود. این پایگاه از مشخصات دقیق (هستی‌شناسی) اجزا و واکنش‌ها استفاده می‌کند و شامل

1-Omics 2- National Center for Biotechnology Information (NCBI) 3-European Bioinformatics Institute (EMBL-EBI)
4- DNA Database of Japan (DDBJ) 5- Protein Data Bank (PDB) 6- Non-Coding RNAs (ncRNAs)
7- Kyoto Encyclopedia of Genes and Genomes (KEGG) 8- Cold Spring Harbor Laboratory 9- Gene Ontology Consortium

جزئیات مربوط به استوکیومتری، محلی سازی، مراجع پایگاه داده خارجی و غیره بوده و فرایندهایی همچون پدیده‌های متشکله پیچیده یا جابجایی مولکول‌ها را نیز پوشش می‌دهد. پایگاه مسیر دیگر با یک دامنه مشابه BioCyc است که مسیرهای مرتبط با اشرشیاکولی را پوشش داده (EcoCyc) و مسیرهای متابولیکی دیگر میکروارگانیسم‌ها (MetaCyc) و انسان (HumanCyc) را به‌خوبی پیشگویی می‌کند. شبکه مرتبط با پروتئین عملکردی (STRING) یک پایگاه داده مهم پروتئینی است که وارد بسیاری از انواع مختلف واکنش بین پروتئین‌ها در انسان و سایر موجودات مدل گردیده است. فرایندهای تنظیم ژن و شبکه تنظیم کننده ژن همانند فرایندهای متابولیک و سیگنالینگ به صورت جزئی‌نگر تحت پوشش قرار نگرفته‌اند؛ بنابراین برخی از پایگاه ها وجود دارند که اطلاعات مرتبط با مکان متصل به عوامل رونویسی را ذخیره کرده‌اند که به عنوان نمونه می‌توان به RegulonDB، TRED و Transfac اشاره کرد. فقدان مدل داده یکسان و روش دسترسی داده و پایگاه مسیرها، ادغام داده را بسیار مشکل می‌سازد. در کنار اطلاعات توپولوژیک در خصوص شبکه واکنش، داده‌های شیمیایی مانند قوانین شیمیایی و ثابت شیمیایی نیز مورد علاقه ویژه‌ای در خلق مدل ریاضی هستند که دو پایگاه داده BRENDA و SABIO-RK در این خصوص فعالیت دارند.

ژنومیکس عملکردی بر اساس اطلاعات حاصل از جنبه استاتیک ژنوم (توالی نوکلئوتید) که توسط پروژه توالی‌یابی ژنوم تولید می‌گردد ساخته می‌شود؛ اما بر جنبه دینامیک آن مانند رونویسی ژن‌ها و ترجمه متمرکز می‌گردد. با کاربرد فناوری پُر برونداد مانند آنالیز ترانسکریپتوم بر اساس ریزآرایه الیگو و یا cDNA، وظایف ژن‌ها در یک مدل کامل ژنومی مطالعه می شود که بین خلأ موجود در توالی و عملکرد زیستی ارتباط برقرار می‌کند. علیرغم اینکه سطوح mRNA همیشه منعکس کننده سطح پروتئین مرتبط با ژن نیست ولی به نظر می‌رسد سطح mRNA به‌عنوان یک پروکسی برای تنوع فنوتیپی عمل کند (۲۷). پروفایل بیان ژن می‌تواند با نقشه‌کشی QTL ترکیب شده و باعث افزایش اثربخشی آنالیز گردد.

اخیراً، توسعه فناوری جدید اومیکس باعث تولید مقادیر زیادی از اطلاعات در خصوص جنبه‌های مختلف زیست‌شناسی سلولی در سطح جهانی شده است که چالش اصلی این دوره جدید تعیین اهمیت زیستی این داده‌ها در زمینه عملکرد سلول و بافت است (۲۲). برای درک بهتر این تغییرات و عوامل کنترل کننده آن‌ها، بسیاری از مطالعات بر روی اندازه‌گیری میزان بیان ژن‌ها متمرکز شده‌اند و در سال‌های اخیر مقادیر قابل توجهی از این نوع داده‌ها تولید شده است. یکی از فناوری‌های کلیدی که این امکان را فراهم می‌کند، توالی‌یابی کل RNA است. از زمان توسعه اولیه خود در سال ۲۰۰۶ (۲۲, ۵)، RNA-seq به سرعت مرز ریزآرایه‌های بیان ژن را برای پروفایل رونویسی در مقیاس وسیعی جابجا کرده است و اکنون فناوری برتر و دلخواه در بسیاری از آزمایشگاه‌ها است. گردش کار معمول RNA-seq شامل پنج مرحله (۱۱) پیش‌پردازش داده خام، (۲۷) هم‌ترازی خوانش‌ها به ژنوم مرجع، (۲۲) بازسازی رونوشت، (۵) کمی سازی بیان و (۳۲) تجزیه و تحلیل بیان افتراقی ژن‌ها است که سه مرحله اول در تمامی آنالیزها بصورت مشترک صورت می‌گیرد. به عنوان گام اولیه، داده‌های RNA-seq ممکن است قبل از تجزیه و تحلیل داده‌ها تحت کنترل کیفیت قرار گیرند. مشابه توالی‌یابی کل ژنوم یا اگزوم، هم‌ترازی خوانش‌ها را می‌توان برای ترسیم خوانش‌ها به ژنوم مرجع یا رونوشت انجام داد. در مرحله بعد، بازسازی ترانسکریپتوم برای شناسایی تمام رونوشت‌های بیان شده در یک نمونه بر اساس اطلاعات مکان‌یابی خوانش‌ها انجام می‌شود. اگر توالی مرجع موجود وجود نداشته باشد، این روش می‌تواند مستقیماً با استفاده از رویکرد مونتاژ *de novo* انجام شود. هنگامی که همه رونوشت‌ها شناسایی شدند، فراوانی آن‌ها را می‌توان با استفاده از اطلاعات مکان‌یابی خوانش‌ها تخمین زد. در نهایت، تجزیه و تحلیل بیان افتراقی با استفاده از برنامه‌های موجود انجام می‌شود (۳۲). برنامه بیوانفورماتیک متعددی برای تجزیه و تحلیل داده RNA-seq توسعه یافته است. حتی ابزارهایی که برای یک هدف توسعه یافته‌اند بر اساس رویکردهای متمایز با استفاده از الگوریتم‌ها و مدل‌های مختلف هستند. تنوع روش این امکان را فراهم می‌کند تا تجزیه و تحلیل اطلاعات با انتخاب برنامه‌ای که بهترین تناسب را برای هر هدف خاص دارد مورد استفاده قرار گیرد. جهت هم‌ترازی خوانش‌ها و داده‌های خام دو راهبرد وجود دارد که در آن می‌توان از یک ژنوم یا رونوشت به عنوان مرجع برای مرحله هم‌ترازی خوانش‌ها استفاده کرد. رویکرد هدایت شده با ژنوم مرجع شامل دو مرحله متوالی (۱۱) تراز خوانش خام با ژنوم مرجع و (۲۷) مونتاژ خوانش‌های همپوشان برای بازسازی رونوشت‌ها است. این رویکرد زمانی سودمند است که اطلاعات حاشیه‌نویسی ژنوم مرجع شناخته شده باشد، در رویکرد مستقل از مرجع از یک الگوریتم مونتاژ *de novo* برای ایجاد مستقیم رونوشت‌های توافقی از خوانش‌های کوتاه بدون ژنوم مرجع استفاده می‌شود و زمانی مفید است که هیچ ژنوم مرجع یا رونوشت شناخته شده‌ای وجود نداشته باشد.

همان‌گونه که در بالا ذکر شد هم‌ترازی خوانش‌ها به ژنوم مرجع، از مهم‌ترین مراحل آنالیز داده‌های RNA-Seq هست که به صورت مؤثری صحت آنالیزهای پایین دستی بستگی به این مرحله دارد (۴). با توجه به گسترش روزافزون استفاده از داده‌های پُربرونداد RNA-Seq الگوریتم‌های مختلفی توسعه یافته و نرم‌افزار و بسته‌های مختلفی برای مکان‌یابی خوانش‌ها معرفی شده‌اند که بسته به هدف و نوع داده‌ها می‌توان استفاده کرد که به عنوان مثال به Cufflinks (۲۹)، Bowtie (۲۰)، Tophat (۹)، StringTie (۲۶)، RNA Star (۱۳) و HISAT2 (۱۷) نام برد؛ که هدف از این تحقیق مقایسه برخی نرم‌افزارهای هم‌ترازی در آنالیز داده‌های RNA-Seq جهت انجام آنالیزهای پایین‌دستی می‌باشد.

مواد و روش‌ها

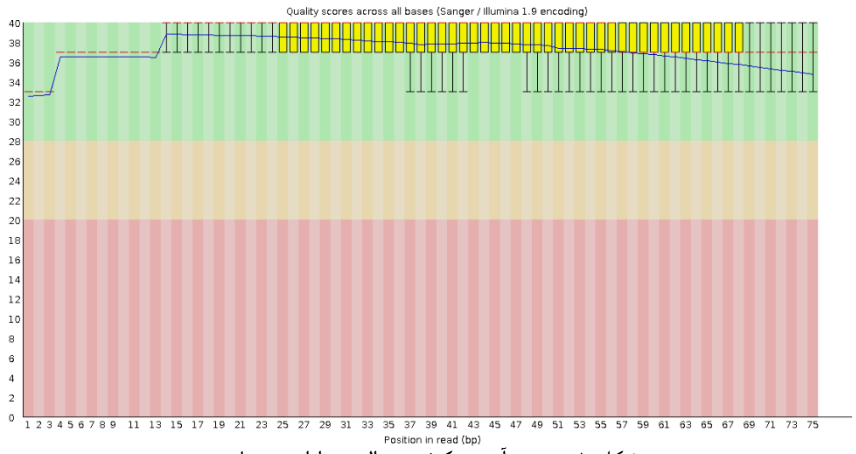
برای ایجاد داده فنوتیپی از اطلاعات ۵۴ رأس گاو هلشتاین با

شکم زایش ۲ به بالا که بر اساس عملکرد باروری به دو گروه با باروری زیاد (فاصله گوساله‌زایی کمتر از ۳۵۰ روز، روزهای باز کمتر از ۷۰ روز و تعداد تلقیح منجر به آبستنی کمتر از ۱/۳) و باروری کم (فاصله گوساله‌زایی بیشتر از ۴۵۰ روز، روزهای باز بیشتر از ۱۵۰ روز و تعداد تلقیح منجر به آبستنی بیشتر از ۳) تقسیم شده بودنداستفاده شد. تمام این نمونه‌ها از دام‌هایی که مدیریت یک سانی در گاوداری هلدینگ شیرین عسل در آن‌ها صورت گرفته است انتخاب شدند. پس از کشتار مجاری تولیدمثلی (بافت اندومتریوم، جسم زرد و تخمدان‌ها) بر روی یخ منتقل و بعد از ۱۵ دقیقه از کشتار قطعه قطعه شدند و در فریزر و دمای ۸۰- درجه سانتی‌گراد برای فرایندهای بعدی ذخیره شدند. حدود ۳۰ میلی‌گرم از هر بافت آسیاب شده و فوراً با بافر RLT کیت استوانه کوچک RNeasy (Qiagen, Sao Paulo, SP, Brazil) مطابق دستورالعمل شرکت سازنده مخلوط شد. استخراج RNA بر اساس روش ستونی، مطابق با دستورالعمل انجام شد. غلظت RNA تام در محصول با استفاده از روش اسپکتروفتومتری تعیین شد. یکپارچگی[۱] RNA کل استخراج شده بر پایه عدد یکپارچگی RNA در محدوده ۸/۳ الی ۸/۷ قرار گرفته بودند (۲۴). برای از بین بردن اثر شکم زایش و همچنین کاهش هزینه‌های تعیین توالی، RNA حاصل از ۳ رأس گاو با شکم‌های زایش مختلف و باروری مشابه با همدیگر ترکیب شده و در مجموع ۱۸ نمونه جهت ارسال برای توالی‌یابی RNA تهیه شد؛ که برای هر گروه با باروری کم و زیاد ۳ تکرار از هر ۳ بافت مختلف تعیین توالی شدند. RNA استخراج شده جهت تعیین توالی کل RNA به شرکت BGI[۲] چین ارسال شد. تعیین توالی RNA با استفاده از Illumina Hiseq 2500 به روش خوانش دوطرفه[۳] تولید شد که طول هر خوانش ۷۵ جفت باز بود. داده‌های خام در فرمت Fastq به حجم تقریبی ۲۹ گیگابایت و در حدود ۶۸۵ میلیون جفت باز دریافت شد و تمامی مراحل آماده‌سازی و تجزیه و تحلیل آن‌ها در پلتفرم گالاکسی نسخه ۲۲۴/۰۱ انجام گرفت (۶). سنجش کیفیت داده‌ها با استفاده از نرم‌افزار FastQC نسخه ۰/۷۳ گالاکسی صورت گرفت (۱). برای نمونه‌هایی که دارای کیفیت پایین بر اساس معیارهای ده‌گانه نرم‌افزار FastQC بودند از نرم‌افزار Trimmomatic نسخه ۰/۳۸/۰ گالاکسی جهت حذف توالی‌هایی با کیفیت پایین و همچنین حذف آداپتورهایی که در هنگام تعیین توالی به ابتدای خوانش‌ها اضافه شده بودند استفاده شد (۷). برای مکان‌یابی خوانش‌های ترانسکریپتوم با ژنوم مرجع گاو (۲۳) از نرم‌افزارهای Hisat2 نسخه ۲/۲/۱ گالاکسی (۱۷)، Tophat2 نسخه ۲/۱/۱ گالاکسی (۱۸) و Bowtie2 نسخه ۲/۴/۲ گالاکسی (۲۰) استفاده شد. ژنوم مرجع گاو نسخه Bos_taurus_UMD_3.1.1 (bostau8) نتیجه نمونه ترکیب شده خون گاوهای نر و ماده نژاد هرفورد است که با عمق خوانش 9X با استفاده از فنّاوری سانجر[۵] در تاریخ ۲۵ نوامبر ۲۰۱۴ معرفی شده است دارای ۲,۶۴۰,۱۲۳,۳۱۰ جفت باز است که بر روی ۳۰ کروموزوم تعیین نقشه شده و شامل ۳۰,۸۱۱ ژن است که از این تعداد ۲۱,۰۸۹ ژن کدکننده[۶] و ۵,۵۲۰ ژن غیر کد کننده[۷] می‌باشند (۲۳). نتایج حاصل از هم‌ترازی به‌صورت درصد کل توالی هم‌تراز شده، درصد توالی‌های هم‌تراز شده مختص به یک جایگاه و درصد توالی‌های هم‌تراز شده به بیش از یک جایگاه برای هر نرم‌افزار گزارش شده است.

نتایج و بحث

سنجش کیفیت داده‌هایی با خوانش دوطرفه به‌طور جداگانه با ۱۰ آزمون مختلف شامل: کیفیت توالی نوکلئوتیدی به ازای هر باز[۸]، کیفیت توالی نوکلئوتیدی به ازای هر کاشی[۹]، امتیاز کیفیت به ازای توالی نوکلئوتیدی[۱۰]، محتوای توالی نوکلئوتیدی به ازای هر باز[۱۱]، محتوای GC به ازای توالی نوکلئوتیدی[۱۲]، محتوای باز خوانده نشده (N) به ازای هر باز[۱۳]، توزیع طول توالی‌ها نوکلئوتیدی[۱۴]، سطوح تکراری توالی نوکلئوتیدی[۱۵]، توالی‌هایی موجود بیش‌ازحد[۱۶]، محتوای آداپتوری[۱۷] است که نتایج هر آزمون به‌طور جداگانه و به‌صورت گرافیکی ارائه می‌گردد. وضعیت قبول، رد یا سطح هشدار برای آزمون‌های ۱۰ گانه ملاک سنجش کیفیت داده‌ها می‌باشد ولی پارامتر کیفیت توالی به ازای هر باز از اهمیت بیشتری برخوردار هست؛ که به‌عنوان مثال نتیجه تعیین کیفیت یکی از تکرارهای باروری زیاد در شکل ۱ آورده شده است. با توجه به کیفیت بالای داده حاصل از تعیین توالی به روش NGS[۱۸] و با توجه به اینکه باقیمانده آداپتورهای الصاقی تعیین توالی نیز در داده خام وجود نداشت نتایج مرحله بعد از ویرایش با نرم‌افزار Trimmomatic (۷) نشان داد که تنها ۰/۸۷ درصد داده‌ها حذف و در حدود ۹۹/۱۳ درصد داده خام برای هم‌ترازی با ژنوم مرجع استفاده گردیده است این در حالی است که حذف داده با کیفیت پایین می‌تواند به‌مراتب بیشتر از این مقدار باشد (۳).

۱- RNA Integrity Number (RIN) 2- Beijing Genomics Institute 3- Paired-End 4- https://usegalaxy.org/22.01
5- Sanger 6- Protein Coding 7- None Coding 8- Per base sequence quality 9- Per tile sequence quality
10- Per sequence quality scores 11- Per base sequence content 12- Per sequence GC content 13- Per base N content
14- Sequence Length Distribution 15- Sequence Duplication Levels 16- Overrepresented sequences 17- Adapter Content
18- Next Generation Sequencing (NGS)



شکل ۱- نتیجه آزمون کیفیت توالی به ازای هر باز
Figure 1. Per base sequence quality score results

نتایج حاصل از هم‌ترازی با ژنوم مرجع نشان داد که به‌طور میانگین ۸۵/۴۷۰ درصد از توالی ترانسکریپتومی توسط نرم‌افزارهای مختلف بر روی ژنوم مرجع مکان‌یابی شدند که ۷۶/۹۶۱ درصد مربوط به اتصال و تعیین جایگاه به یک محل اختصاصی بر ژنوم مرجع هست و ۸/۵۰۹ درصد به بیش از ۱ جایگاه در ژنوم مرجع متصل شده است (جدول ۱). با توجه به نتایج درج شده در جدول مشاهده می‌گردد که نرم‌افزار Tophat2 درصد بیشتری از خوانش‌ها را (۹۴/۱۹۷) بر روی ژنوم مرجع مکان‌یابی نموده است که اختلاف معنی‌داری در سطح ۰/۰۰۱ با دو نرم‌افزار دیگر داشت که می‌تواند در آنالیز نواحی غیر کد کننده مانند miRNA مفید باشد (۴). سهم نرم‌افزار Hisat2 از درصد مکان‌یابی کل خوانش‌ها بر روی ژنوم مرجع ۹۲/۵۲۶ می‌باشد که از لحاظ عملکردی در جایگاه دوم قرار گرفت. ولی آنچه مشخص است نرم‌افزار Bowtie2 که برای هم‌ترازی داده‌های DNA بر روی ژنوم مرجع به کار می‌رود در خصوص داده‌های RNA کارآیی ندارد.

همان‌گونه که در شکل ۲ مشاهده می شود بیشترین میزان مکان‌یابی مختص به یک جایگاه ویژه بر روی ژنوم مرجع (۸۹/۲۰۲ درصد خوانش‌های کل) با استفاده از نرم‌افزار Hisat2 صورت گرفته است که هر چند اختلاف آماری با عملکرد نرم‌افزار Tophat2 (۸۷/۸۱۲ درصد از کل خوانش‌ها) ندارد، لیکن توانسته است عمل مکان‌یابی بر روی ژنوم مرجع را با تخصیص بیشتری انجام دهد که می‌تواند بر نتایج حاصل از آنالیزهای پایین‌دستی تأثیرگذار باشد. نتایج نشان داد که تنها ۳/۳۲۴ درصد از کل خوانش‌ها در نرم‌افزار Hisat2 به بیش از یک جایگاه روی نقشه ژنوم مرجع متصل شده است که این عدد در مورد Tophat2 ۶/۳۸۵ درصد بوده است، لذا ویژگی ممتاز نرم‌افزار Hisat2 در عملکرد تخصیصی آن در ارتباط با هم‌ترازی خوانش‌های حاصل از داده‌های RNA-Seq با ژنوم مرجع می‌باشد.

جدول ۱- درصد هم‌ترازی کل، هم‌ترازی به یک جایگاه اختصاصی و هم‌ترازی به بیش از یک جایگاه در نرم‌افزارهای مختلف
Table 1. The percentage of total alignment, alignment to a specific position and alignment to more than one position in different softwares

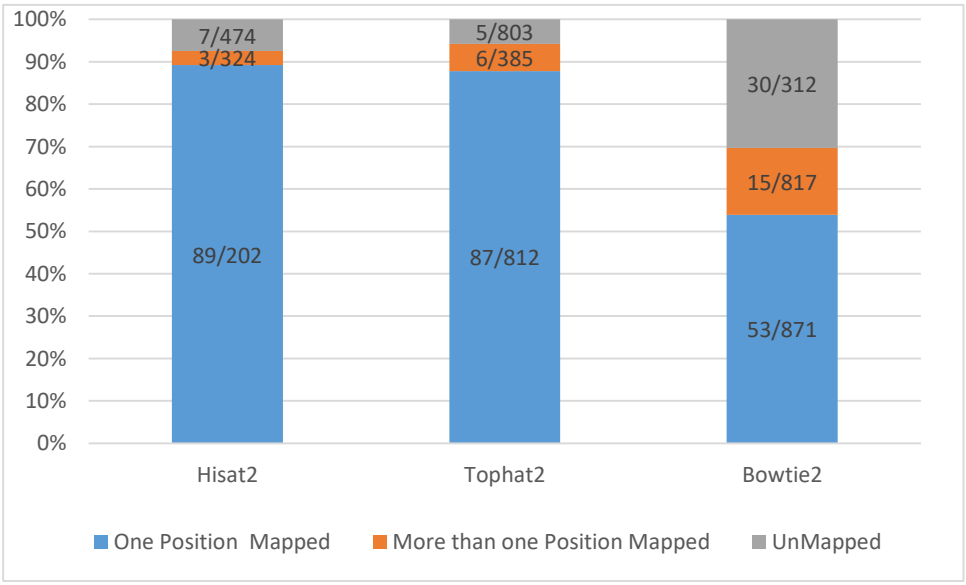
Software نرم‌افزار	Sample No تعداد نمونه	Total Mapped (%) درصد مکانیابی کل	One position Mapped (%) درصد مکانیابی به ۱ جایگاه	More than one position Mapped (%) درصد مکانیابی به بیش از ۱ جایگاه
Hisat2	18	92.526 (1.338)^{b}	89.202 (1.466)^{a}	3.324 (0.752)^{a}
Tophat2	18	94.197 (1.305)^{a}	87.812 (1.450)^{b}	6.385 (0.484)^{b}
Bowtie2	18	69.688 (1.308)^{c}	53.871 (1.045)^{c}	15.817 (1.346)^{c}
Total	54	85.470 (11.359)	76.961 (16.543)	8.509 (5.444)
SEM		0.179	0.182	0.127
Sig		0.000	0.000	0.000

در مطالعه دیگری تعداد ژن‌هایی که بر روی ژنوم مرجع در مطالعه بیان افتراقی متفاوت حاصل شده بود در نرم‌افزار STAR بیشتر از نرم‌افزار Hisat2 بود (۸). میزان مکان‌یابی خوانش‌ها به ژنوم مرجع از ۷۵/۲۷ الی ۹۲/۰۹ درصد در بیماری تنفسی گاو متغیر بوده و به‌طور متوسط تقریباً ۲۵ میلیون خوانش به‌طور منحصربه‌فرد در هر نمونه از طریق STAR هم‌تراز شده بود (۱۶). درحالی‌که مکان‌یابی توالی با ژنوم گاو در یک جایگاه اختصاصی در تمامی نمونه‌های پاراتوبرکلوزیس[1] در نمونه‌های بافتی گاو از ۹۳/۱۳–۸۱/۳۵ درصد توسط همین نرم‌افزار حاصل شده است (۲۱). در تحقیقی که روی بافت شکمبه در مراحل از شیرگیری در گاو صورت گرفته است از مجموع حدود ۴۳ میلیون خوانش، تقریباً ۹۵ درصد از خوانش‌ها با استفاده از نرم‌افزار STAR در ژنوم مرجع گاو مکان‌یابی شده بود (۸). مقایسه انجام یافته در مطالعه Raplee و همکاران (۲۸) در تحقیقات بالینی سرطان پستان نشان داد که Hisat2 مستعد خوانش‌های نادرست در مکان‌های ژنومی رتروژن[2] بوده و STAR ترازهای دقیق‌تری را برای نمونه‌های اولیه نئوپلازی[3] ایجاد کرده بود.

1- Paratuberculosis 2- Retrogene 3- Neoplasia

در تحقیق روی گاوهای دورگ حاصل از سیستانی و مونتبلیارد مکان‌یابی خوانش‌ها به ژنوم مرجع با استفاده از نرم‌افزار Tophat2 نشان داد که ۸۰/۶–۷۰/۵ درصد خوانش‌ها بر روی ژنوم مرجع مکان‌یابی شدند که درصد خوانش‌های مکان‌یابی شده بر روی یک جایگاه در ژنوم مرجع حدود ۷۸/۱–۷۲/۹ درصد بود (۲). در بررسی بر روی بافت چربی احشایی و زیر جلدی مکان‌یابی شده با ژنوم مرجع برای هر نمونه با استفاده از نرم‌افزار Tophat در مجموع ۶۵،۷۰۲،۵۶۸ خوانش برای SAT و ۷۶،۸۳۲،۴۱۰ خوانش برای VAT از داده‌های RNA-seq پس از کنترل کیفیت به دست آمد که تقریباً ۶۰٪ از قرائت‌های معتبر به ژنوم مرجع مکان‌یابی شده است (۱۴). پس از مکان‌یابی خوانش‌ها به ژنوم مرجع نوعی تمساح با استفاده از Tophat2 مشخص شد که این عملکرد از ۶۱/۲۶ درصد تا ۸۶/۸۳ درصد متفاوت بوده است (۲۵). بیشترین درصد هم‌ترازی خوانش‌ها به ژنوم مرجع توسط نرم‌افزار TopHat2در مطالعه رویان گاو حاصل شده است و نتایج نشان دادکه میزان هم‌ترازی به ژن مرجع گاوی در هر مرحله ۹۳/۱۷ تا ۹۴/۲۳ و نسبت تعداد توالی به موقعیت‌های چندگانه ژنوم از ۲/۹۹ الی ۴/۸۹ درصد بوده است (۳۱). نتایج مطالعه بهرامی (۴) نیز نشان داد که پس از نرم‌افزارهای Tophat2 و Hisat2، نرم‌افزار STAR می‌تواند جهت هم‌ترازی داده‌های ترانسکریپتوم مورد استفاده قرار گیرد و Bowtie2 کارآیی کمتری در این خصوص دارد.

غلامی طاحونه و مرادی شهربابک (۱۵) بیش از ۹۰٪ خوانش‌ها را توسط نرم‌افزار Hisat2 بر روی ژنوم مرجع گاوی مکان‌یابی نمودند. هم ترازی با Hisat2 بر روی نمونه‌های بالینی انسان که دچار افسردگی اساسی بودند نشان داد که میزان مکان‌یابی با ژنوم مرجع در گروه بیماران با اختلال افسردگی اساسی[1] (MDD) با میانگین ۹۲،۷۹، در گروه بیماران با اختلال افسردگی اساسی منجر به خودکشی[2] MDD-S ۹۲/۴۶ درصد و در گروه سالم و کنترل ۹۳/۹۷ درصد بود (۳۰). درصد کل مکان‌یابی شده با استفاده از همین نرم‌افزار به ژنوم مرجع در داده‌های حاصل از RNA-Seq تخمدان گاو بومی Xiangxi ۹۵/۰۳ بود که میزان مکان‌یابی به یک جایگاه اختصاصی ۹۲/۶۵ درصد و میزان مکان‌یابی به بیش از یک جایگاه ۲/۳۴ درصد گزارش شده است (۹). توالی‌یابی کتابخانه‌ای رونوشت‌های کبد و طحال در مرغ‌های مبتلا به سالمونلا نشان داد که از ۴۶/۹۵ میلیون خوانش ترانسکریپتوم کبد، ۴۳/۷۶ میلیون خوانش (۹۲/۸۲ درصد) مراحل کیفیت خوانش را پشت سر گذاشتند و در مجموع ۴۱/۴۶ میلیون خوانش (۹۴/۷۴ درصد) به ژنوم مرجع مرغ مکان‌یابی اختصاصی شده بودند. از ۴۱،۷۴ میلیون خوانش نمونه‌های طحال نیز، ۳۸/۶۵ میلیون خوانش (۹۲/۲۴ درصد) مرحله کنترل کیفیت را پشت سر گذاشتند و در مجموع ۳۵/۴۵ میلیون خوانش (۹۱/۷۲ درصد) با ژنوم مرجع توسط نرم‌افزار Hisat2 هم‌تراز شدند (۱۲). بهرامی (۴) نتایج مشابهی را با استفاده از داده‌های شبیه سازی شده ژنوم انسان حاصل نمود و گزارش کرد که نرم‌افزار Hisat2 دارای حساسیت بیشتری در خصوص هم‌ترازی داده‌ها بوده و از حافظه کمتری نسبت به نرم‌افزارهای دیگر استفاده می‌کند. علاوه بر این دارای دقت و صحت بالاتری برای هم‌ترازی به ژنوم مرجع می‌باشد. لذا پیشنهاد کرد که برای هم‌ترازی آنالیزهای RNA-Seq بر روی ژنوم مرجع از Hisat2 استفاده شود.



شکل ۲– نمودار انباشته داده‌های هم‌ترازی به ژنوم مرجع توسط نرم‌افزارهای Hisat2، Tophat2 و Bowtie2
Figure 2. Cumulative chart of alignment data to the reference genome by Hisat2, Tophat2 and Bowtie2 softwares

با توجه به این که پس از بررسی‌های اولیه کنترل کیفیت روی خروجی خام از توالی‌یابی، هم ترازی اولین مرحله در تجزیه و تحلیل RNA-seq است و تمام تحلیل‌های بعدی عمیقاً به این مرحله اولیه متکی است (۱۰). معمولاً،

1- Major Depressive Disorder (MDD)
2- Major Depressive Disorder Suicide (MDD-S)

قرائت‌های به‌دست‌آمده از توالی‌یابی، نقشه‌برداری می‌شوند و با یک ژنوم مرجع تراز می‌شوند که به‌ویژه به دلیل وجود قرائت‌های خروجی در اتصالات اگزون–اگزون، مستعد خطا برای داده‌های RNA-Seq است. متداول‌ترین پلتفرم‌های نرم‌افزاری موجود برای نگاشت به ژنوم مرجع (Tophat2، Hisat2 و STAR) اتصالات را شناسایی می‌کنند. این پلتفرم‌ها در سرعت محاسباتی و استفاده از حافظه و در الگوریتم‌های شان برای رسیدگی به دقت هم‌ترازی پایه و اتصال متفاوت هستند. در حال حاضر استفاده از Tophat2 کمتر شده است و به دلیل کارآمدی محاسباتی نسبی توسط Hisat2 جایگزین شده است علیرغم اینکه TopHat2 و Hisat2 بر اساس برنامه مکان‌یابی کوتاه خوانش Bowtie2 ساخته شده‌اند (۱۹).

نتیجه‌گیری کلی

نتایج حاصل از این تحقیق به‌وضوح نشان می دهد که هم‌ترازی صحیح با ژنوم مرجع نقش بزرگی در نتایج تجزیه و تحلیل بیوانفورماتیکی، به‌ویژه برای شناسایی بیان ژن‌های افتراقی دارد. این نتایج، همراه با سایر آزمون‌های مقایسه‌ای برنامه‌های مختلف (۲۸)، نشان می‌دهد که یک برنامه هم‌ترازی مشخص را نمی‌توان به‌طور عمومی برای مجموعه داده‌های RNA-seq اعمال کرد. در حالی که پیشرفت‌های قابل توجه در فناوری توالی‌یابی، طول خوانش‌های نوکلئوتید خروجی را به بیش از ۳۰۰ نوکلئوتید افزایش داده است، ممکن است دقت هم‌ترازی افزایش پیدا کند. با توجه به افزایش استفاده از RNA-seq برای تشخیص ژن‌های عملکردی، محققان زیستی باید پایه و اساس قوی‌تری در ابزارهای بیوانفورماتیک و مسیرهای تجزیه و تحلیل داده‌های RNA-seq به منظور تولید نتایج با اطمینان بیشتر منظور نمایند. اگر چه هزینه و زمان مورد نیاز برای تجزیه و تحلیل رونوشت با توسعه توالی‌یابی نسل جدید تا حد زیادی کاهش یافته است، اما نتایج حاصل از این تحقیق بر اهمیت تأثیر ابزارهای بیوانفورماتیک برای هم‌ترازی توالی ها که هدف این مطالعه بود، تأکید می‌کند. فلذا پیشنهاد می شود که جهت انجام آنالیزهای پایین دستی صحیح، داده‌های ترانسکریپتوم توسط نرم‌افزار Hisat2 با ژنوم مرجع همتراز شوند.

منابع

1. Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data.
2. Asgari Esfedan, B., G.R. Dashab, M.H. Banabazi and M. Rokouei. 2022. The effect of crossbreeding by the Montbeliard cattle on the transcriptome of the Sistani cattle. Research on Animal Production (Scientific and Research), 12(31): 134-145.
3. Attari, M., H. Moradi Shahrbabak, G. Nehzati Paghale, M. H. Banabazi and M. Hashemi. 2019. Study of differential gene expression in queen, drone and worker honey bee using RNA-seq data. Iranian Journal of Animal Science, 50(2): 103-113.
4. Bahrami, A. 2020. Which aligner software is the best for our study. Journal of Genetics and Genome Research, 7, 048.
5. Bainbridge, M.N., R.L. Warren, M. Hirst, T. Romanuik, T. Zeng, A. Go and V. Magrini. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics, 7(1): 1-11.
6. Blankenberg, D., A. Gordon, G. Von Kuster, N. Coraor, J. Taylor, A. Nekrutenko and G. Team. 2010. Manipulation of FASTQ data with Galaxy. Bioinformatics, 26(14): 1783-1785.
7. Bolger, A.M., M. Lohse and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15): 2114-2120.
8. Boschiero, C., Y. Gao, R.L. Baldwin, L. Ma, C.J. Li and G.E. Liu. 2022. Differentially CTCF-binding sites in cattle rumen tissue during weaning. International Journal of Molecular Sciences, 23(16): 9070.
9. Cheng, H., S. Ao, L. Yun, S. Weihong, L. Hong, L. Jianbo and Y. Kangle. 2022. RNA-Seq transcriptome analysis to unravel the gene expression profile of ovarian development in Xiangxi cattle. Pakistan Veterinary Journal, 42(2): 222-228.
10. Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson and X. Zhang. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology, 17(1): 1-19.
11. Covert, M.W., C.H. Schilling, I. Famili, J.S. Edwards, I.I. Goryanin, E. Selkov and B.O. Palsson. 2001. Metabolic modeling of microbial strains in silico. Trends in Biochemical Sciences, 26(3): 179-186.
12. Dar, M.A., S.M. Ahmad, B.A. Bhat, T.A. Dar, Z. Haq, B.A. Wani and N.A. Ganai. 2022. Comparative RNA-Seq analysis reveals insights in Salmonella disease resistance of chicken; and database development as resource for gene expression in poultry. Genomics, 114(5): 110475.
13. Dobin, A., C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha and T. R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1): 15-21.
14. Duan, X., Y. Liu, X. Zhang and H. Zhao. 2022. Transcriptional features of cattle visceral and subcutaneous adipose tissues: a study of RNA-seq. Open Journal of Animal Sciences, 12(3): 441-453.
15. Gholami Tahoone, M. and H. Moradi SharBabak. 2022. Differential genes expression of blood tissue related to pre-calving ketosis in holstein cow using transcriptomics data. Research on Animal Production (Scientific and Research), 13(36): 147-153 (In Persian).

16. Jiminez, J., E. Timsit, K. Orsel, F. Van der Meer, L.L. Guan and G. Plastow. 2021. Whole-blood transcriptome analysis of feedlot cattle with and without bovine respiratory disease. Frontiers in Genetics, 12: 627623.
17. Kim, D., B. Langmead and S.L. Salzberg. 2015. HISAT: a fast spliced aligner with low memory requirements. Nature Methods, 12(4): 357-360.
18. Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley and S. L. Salzberg. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology, 14(4): 1-13.
19. Langmead, B. and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4): 357.
20. Langmead, B., C. Trapnell, M. Pop and S.L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3): 1-10.
21. Li, H., J. Huang, J. Zhang, Y. Gao, B. Han and D. Sun. 2022. Identification of alternative splicing events associated with paratuberculosis in dairy cattle using multi-tissue RNA sequencing data. Genes, 13(3): 497.
22. McGettigan, P., J. Browne, S. Carrington, M. Crowe, T. Fair, N. Forde and K. Pluta. 2016. Fertility and genomics: comparison of gene expression in contrasting reproductive tissues of female cattle. Reproduction, Fertility and Development, 28(2): 11-24.
23. Merchant, S., D.E. Wood and S.L. Salzberg. 2014. Unexpected cross-species contamination in genome sequencing projects. Peer Journal, 2: e675.
24. Mesquita, F., R. Ramos, G. Pugliesi, S. Andrade, V. Van Hoeck, A. Langbeen and H. Fukumasu. (2016). Endometrial transcriptional profiling of a bovine fertility model by next-generation sequencing. Genomics Data, 7: 26-28.
25. Nie, H., Y. Zhang, S. Duan, Y. Zhang, Y. Xu, J. Zhan and X. Wu. 2022. RNA-Sequencing Analysis of Gene-Expression profiles in the dorsal gland of alligator sinensis at different time points of embryonic and neonatal development. Life, 12(11): 1787.
26. Pertea, M., G.M. Pertea, C. M. Antonescu, T.C. Chang, J.T. Mendell and S.L. Salzberg. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology, 33(3): 290-295.
27. Preuss, T.M., M. Caceres, M.C. Oldham and D.H. Geschwind. 2004. Human brain evolution: insights from microarrays. Nature Reviews Genetics, 5(11): 850.
28. Raplee, I.D., A.V. Evsikov and C. Marín de Evsikova. 2019. Aligning the Aligners: Comparison of RNA sequencing data alignment and gene expression quantification tools for clinical breast cancer research. Journal of Personalized Medicine, 9(2): 18.
29. Trapnell, C., B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. Van Baren and L. Pachter. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology, 28(5): 511-515.
30. Verma, P. and M. Shakya. 2021. Transcriptomics and sequencing analysis of gene expression profiling for major depressive disorder. Indian Journal of Psychiatry, 63(6): 549.
31. Wang, J., J. Z. Di Fang, F. Huang, B. Liu, W. Tao, B. Cui and Q. Gao. 2022. Transcriptome analysis of cattle embryos based on single cell RNA-Seq. Pakistan Journal of Zoology, 1-8.
32. Yang, I.S. and S. Kim. 2015. Analysis of whole transcriptome sequencing data: workflow and software. Genomics and Informatics, 13(4): 119.

Comparison of some Alignment Software in the Analysis of Dairy Cows RNA-Seq Data

Ghorban Elyasi Zarringhabaie[1], Mostafa Sadeghi[2] and Seyed Reza Miraie Ashtiani[3]

1- Ph.D. Candidate, Department of Animal Science, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran, (Corresponding author: gh.elyasi@ut.ac.ir)
2- Associate Professor, Department of Animal Science, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran
3- Professor, Department of Animal Science, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran
Received: 12 September, 2022 Accepted: 18 December, 2022

Extended Abstract

Introduction and Objective: Due to the increasing use of next generation sequencing (NGS), it is necessary to use specialized algorithms and software to perform statistical analysis in order to identify functional genes. Alignment of reads with the reference genome is the first and most important step in most RNA-Seq data analysis programs, and the accuracy of downstream analyzes effectively depends on this step. Therefore, the aim of this research was to compare some different software for aligning the data obtained from total RNA sequencing on the reference genome.

Material and Methods: RNA-Seq data related to 54 Holstein dairy cows raised in industrial conditions were used to identify genes effective in fertility. The quality of the reads was determined by FastQC software and editing of low quality sequences was done using Trimmomatic software. The edited data were aligned with bovine reference genome using Bowtie2, Tophat2 and Hisat2 softwares. The total percentage of mapped reads, the percentage of mapped reads on one location in the reference genome, and the percentage of mapped reads on more than one location were calculated.

Results: The results showed that the most alignment was done on the cow reference genome using Tophat2 software. which mapped 94.197% of the From the total available reads, 94.197% and 92.526% were mapped on the reference genome by Tophat2 and Hisat2 software, respectively. The Hisat2 software had more allocation function and mapped 89.202% of the data to a specific position, while this parameter was 87.812% of the total sequences for Tophat2 software. From the total used sequences, only 3.324% and 6.385% of the sequences were aligned by Hisat2 and Tophat2 software respectively to more than one position of the reference genome. Bowtie2 software had low performance compared to other two software.

Conclusion: The comparison of RNA-Seq data alignment software on the reference genome showed that although the Hisat2 swas the best read mapping software but, Tophat2 software can also be used instead in RNA-seq data analysis. Meanwhile, Bowtie2 software is not very effective in relation to RNA-Seq data.

Keywords: Gene expression, Mapping, Omix, Reference Genome and Sequencing

[DOR: 20.1001.1.22518622.1402.14.39.14.5]
[DOI: 10.61186/rap.14.39.131]