نشریه پژوهشهای علوم دامی ایران
جلد ۹، شماره ۴، زمستان ۱۳۹۶، ص. ۴۹۷-۴۸۴

Iranian Journal of Animal Science Research
Vol. 9, No. 4, Winter 2018, p. 484-497

پیش‌بینی ژن‌های کاذب جدید در ژنوم مرجع گوسفند

محمدرضا بختیاری‌زاده[1*]- زهره مزدوری[2]- پدرام شاکری[3]

تاریخ دریافت: ۱۳۹۵/۰۹/۰۳
تاریخ پذیرش: ۱۳۹۶/۰۱/۲۳

چکیده

ژن‌های کاذب نسخه‌هایی از ژن اجدادی می‌باشند که به مرور زمان فعالیت آنها نسبت به ژن اولیه تغییر کرده است و در ژنوم بر اثر فرآیندهایی مانند مضاعف شدگی ژنی و همچنین رونویسی واژگون ایجاد شده‌اند. ژن‌های کاذب تا مدت‌ها به‌عنوان توالی‌های غیر عملکردی ژنوم در نظر گرفته می‌شدند. با این وجود پژوهش‌های اخیر گزارشاتی مبنی بر فعالیت زیستی این ژن‌ها ارائه داده‌اند، در نتیجه عملکردی بودن این ژن‌ها موجب افزایش حاشیه نویسی صحیح‌تر این ژن‌ها در ژنوم موجودات شده است. در پژوهش حاضر به منظور بهبود حاشیه نویسی ژنوم گوسفند، برای نخستین بار با استفاده از روش‌های محاسباتی بر پایه بررسی تشابه با استفاده از نرم‌افزار PseudoPipe، ژن‌های کاذب مرتبط با ژن‌های کدکننده پروتئین در سطح ژنوم شناسایی شدند. همچنین گروه‌های کارکردی ژن‌های والدی که ژن‌های کاذب از آنها مشتق شده‌اند با استفاده از پایگاه اینترنتی DAVID بررسی شدند. در نهایت ویژگی‌های مختلف ژن‌های کاذب کاندید جدید شناسایی شده با ژن‌های کاذب شناخته‌شده در گونه‌های انسان، موش و گاو مقایسه شدند. به طور کلی ۴۰۹۸ ژن کاذب با سطح اطمینان بالا شامل ۱۱۰۲ ژن کاذب از نوع مضاعف شده و ۲۹۹۶ از نوع پردازش شده شناسایی شدند. نتایج نشان داد که ژن‌های کاذب شناسایی شده در فرآیندهای زیستی گوناگونی مانند splicing mRNA، پیدایش ریبوزوم، اتصال rRNA، انتقال الکترون میتوکندریایی، ترجمه و غیره نقش دارند. مقایسه ویژگی‌های مختلف ژن‌های شناسایی شده با دیگر گونه‌ها نشان داد که نتایج حاصل از این پژوهش در تطابق با پژوهش‌های گذشته می‌باشد. نتایج حاصل از این پژوهش به بهبود حاشیه نویسی ژنوم گوسفند کمک خواهد کرد.

واژه‌های کلیدی: تشابه، حاشیه نویسی ژنوم، عدم تطابق، نرم افزار PseudoPipe.

مقدمه

آزمایش‌های ژنتیک مولکولی در مورد گونه گوسفند (Ovis aries) به دلیل کامل نبودن حاشیه نویسی[4] ژنوم مرجع محدود شده است (۱). پیشرفت‌های اخیر در زیست شناسی مولکولی و توسعه روش‌های آزمایشگاهی، خصوصا در توالی‌یابی ژنوم و فناوری‌های پربازده توالی‌یابی نسل بعد[5]، منجر به رشد بی‌سابقه‌ای در داده‌های زیستی شده است (۲). بنابراین دسترسی به داده‌های مولکولی بینش جدیدی به سوی آزمایش‌های ژنومی در موجوداتی که ژنوم جامع یا اطلاعات بیان ژن‌ها آنها در دسترس نیست را باز کرده است (۳)، به طوری که حاشیه نویسی ژنوم جانداران مختلف کامل‌تر شده است (۴). در سال ۱۹۷۷، یک توالی مشابه با ژن 5S rRNA از *Xenopus laevis* کشف شد که فاقد قابلیت کدشوندگی 5S rRNA عملکردی بود. این توالی دارای ۱۴ جفت باز عدم تطابق[6] با ژن 5S rRNA عملکردی بود و به‌عنوان ژن کاذب نام گرفت که اولین گزارش از ژن‌های کاذب می‌باشد (۵).

یکی از بخش‌های مهم موجود در ژنوم پستانداران، ژن‌های کاذب می‌باشند که کپی‌های غیرفعال از ژن‌های عملکردی هستند (۶). ژن‌های کاذب قطعاتی از DNA ژنومی و عضوی ازخانواده ژن اصلی می‌باشند که اگرچه شباهت توالی بالایی با ژن‌های اصلی دارند، اما غیرعملکردی بوده و فاقد توانایی کدشوندگی هستند (۵). ژن‌های کاذب در اصل به‌عنوان فسیل‌های ژنومی هستند که می‌توانند برای استنباط توالی اجدادی و تاریخچه تکاملی ژن‌هایی که امروز حضور دارند، استفاده شوند (۶). این ژن‌ها به‌عنوان جایگاه‌های ژنومی از رده خارج شده با توالی مشابه با ژن‌های عملکردی می‌باشند که به دلیل

۱- استادیار گروه علوم دام و طیور، پردیس ابوریحان، دانشگاه تهران،
۲- دانشجوی دکتری ژنتیک و اصلاح نژاد دام، گروه علوم دامی، دانشکده کشاورزی، دانشگاه تربیت مدرس، تهران،
۳- دانشجوی کارشناسی ارشد گروه علوم دام و طیور، پردیس ابوریحان، دانشگاه تهران، ایران.
(*- نویسنده مسئول: (Email: mrbakhtiari@ut.ac.ir
DOI: 10.22067/ijasr.v1i1.60511

4- Annotation
5- Next generation sequencing
6- Mismatches

جهش‌های مخرب مانند تغییر قاب کدکنندگی ژن اصلی و یا ایجاد کدون‌های پایان اشتباه، فاقد پتانسیل کدشوندگی می‌باشند (۷ و ۱۰). در حال حاضر پژوهشگران، شماری از ژن‌های کاذب را با ویژگی‌های مختلف زیستی گزارش کرده‌اند که نگرش قبلی در مورد این ژن‌ها به‌عنوان ژن‌های بدون فعالیت و زائد را تغییر داده است (۵) .

ژن‌های عملکردی مربوط به ژن‌های کاذب (ژن‌های اصلی که ژن‌های کاذب از آن منشأ گرفته‌اند) اغلب به‌عنوان ژن‌های والد شناخته می‌شوند. بر اساس ساز و کار تشکیل آنها، ژن‌های کاذب به سه دسته اصلی تقسیم‌بندی می‌شوند: ۱) ژن‌های کاذب پردازش شده[1]، که توسط رتروترنسپوزیشن[2] با رونویسی معکوس mRNA و بازگشت به DNA و درج تصادفی در ژنوم ایجاد می‌شوند، ۲) ژن‌های کاذب مضاعف شده[3] که در هنگام همانندسازی DNA و به طور ناخواسته یک کپی اضافی از یک ژن در ژنوم درج شده و این ژن‌ها ایجاد می‌شوند و ۳) ژن‌های کاذب واحد[4] که از طریق ایجاد جهش در ژن‌های کد کننده پروتئین که قبلا عملکردی بودند ایجاد می‌شوند و ژن اصلی غیرعملکردی می‌شود (۷، ۱۱ و ۱۲).

انواع مختلف ژن‌های کاذب، ویژگی‌های ژنومی متفاوتی را نشان می‌دهند. ژن‌های کاذب مضاعف ساختار ژنومی اینترون-اگزونی، مانند ژن اصلی دارند و ممکن است هنوز هم در بالادست توالی‌های تنظیمی والدهایشان حفظ شده باشند (۱۳). از آنجا که احتمالا حمل دو کپی مشابه از یک ژن عملکردی مطلوب نیست، یکی از کپی‌ها تحت فرآیند pseudogenization قرار گرفته که منجر به از دست دادن ترجمه و احتمالا رونویسی آن می‌شود (۱۴). به عبارت دیگر ژن‌های کاذب مضاعف فاقد رونویسی بوده اما برخی کنترل‌های تنظیمی بالادست ژن را از والدینشان حفظ کرده‌اند (برای مثال، نواحی اتصال فاکتور رونویسی[5] و سطوح متنوع فعالیت کروماتین) (۱۳). در مقابل، ژن‌های کاذب پردازش شده، اینترون‌هایشان را از دست داده‌اند و تنها شامل توالی اگزونی بوده و نواحی تنظیمی بالادست را حفظ کرده‌اند (۱۳). همچنین ژن‌های کاذب پردازش شده ممکن است دم پلی A را در انتهای ۳' خود نیز حفظ کرده باشند. ترکیب انواع متفاوت ژن‌های کاذب در میان ژنوم موجودات متفاوت می‌باشد (۱۵). به‌عنوان مثال در ژنوم انسان، ژن‌های کاذب پردازش شده، فراوان‌ترین نوع می‌باشند که علت آن فعالیت زیاد رتروترنسپوزون‌ها[6] در پریمات‌های (پستانداران) اجدادی ۴۰ میلیون سال پیش ذکر شده است (۱۶ و ۱۷).

گزارش شده است که ژنوم انسان دارای ۸۰۰۰-۱۲۰۰۰ ژن کاذب و ژنوم موش دارای ۵۰۰۰ ژن کاذب می‌باشد (۶). همچنین بیان شده که تقریبا ۹ درصد (۸۷۶) از ژن‌های کاذب در ژنوم انسان به طور فعال رونویسی می‌شوند که شامل ۵۳۱ ژن کاذب پردازش شده و ۳۴۵ ژن کاذب مضاعف می‌باشند (۱۳). بر این اساس این ژن‌ها در دیگر گونه‌های پستانداران (مانند گوسفند) نیز وجود داشته و نیازمند پژوهش‌هایی گسترده در این زمینه می‌باشد. یکی از روش‌های قدرتمند و با بازده بالا در رابطه با شناسایی ژن‌های کاذب استفاده از روش‌های بیوانفورماتیک و محاسباتی می‌باشد. با استفاده از این روش‌ها برخی از ژن‌های کاذب در گونه‌هایی همچون انسان (۱۸)، کرم (worm) و مگس (fly) شناسایی شده است (۱۹). با این وجود داده‌های کمی در مورد توزیع، عملکرد و انواع ژن‌های کاذب در ژنوم گوسفند وجود دارد، به طوری که تا به حال بر اساس بانک اطلاعاتی ENSEMBL (http://www.ENSEMBL.org/، نسخه ۸۷) ۲۹۰ ژن کاذب و بر اساس بانک اطلاعاتی NCBI (https://www.ncbi.nlm.nih.gov) ۲۵۷ ژن کاذب تأیید شده در ژنوم گوسفند گزارش شده است. همچنین ۳۱۶۴ ژن کاذب که با روش‌های محاسباتی در ژنوم گوسفند شناسایی شده‌اند در بانک اطلاعاتی NCBI گزارش شده است. در نتیجه شناسایی این ژن‌ها بینش جدیدی را در رابطه با مکانیسم‌های تنظیمی این ژن‌ها در ژنوم گوسفند فراهم خواهد نمود و به کامل شدن حاشیه نویسی ژنوم این حیوان کمک خواهد کرد. بنابراین هدف از پژوهش حاضر شناسایی ژن‌های کاذب (پرداش شده و مضاعف) در ژنوم مرجع گوسفند به منظور بهبود حاشیه نویسی این ژنوم می‌باشد.

مواد و روش‌ها

در پژوهش حاضر به منظور شناسایی ژن‌های کاذب از نرم افزار PseudoPipe (http://www.pseudogene.org/pseudopipe/) و بررسی دستی نتایج استفاده گردید. بدین منظور مراحل زیر انجام شد:

۱- در ابتدا توالی ژنومی گوسفند (نسخه Ovis_aries.Oar_v3.1) که مناطق تکراری آن پوشیده شده[7]، همه توالی‌های پروتئینی شناخته شده در گوسفند و مختصات‌های کروموزومی (نسخه ۷۷) ژن‌های مربوط به این پروتئین‌ها از بانک اطلاعاتی ENSEMBL استخراج گردید. مختصات کروموزومی ژن‌ها به منظور تفکیک ژن‌های کاذب کاندید از ژن‌های والدی مورد نیاز می‌باشد.

۲- همه مناطق ژنومی که با توالی‌های پروتئینی دارای مشابهت

1- Processed
2- Retrotransposition
3- Duplicated
4- Unitary
5- Transcription Factor Binding sites
6- Retrotransposons
7- Repeat-masked

بودند با استفاده از BLAST شناسایی شد (E-value $\leq 1\times10^{-4}$). با توجه به برخی محدودیت‌های روش BLAST، در نرم افزار PseudoPipe از tfasty نیز به‌عنوان یکی دیگر از نرم افزارهای قدرتمند هم‌ردیفی برای افزایش صحت استفاده شده است. این نرم افزار از مجموعه نرم‌افزاری FASTA بوده و دارای صحت بالائی در همردیفی توالی‌ها با هم می‌باشد (۱۵). در ادامه بخش‌هایی که بر اساس مختصات ژنومی با ژن‌های عملکردی همپوشانی داشتند (بیش از ۳۰ جفت باز) حذف شدند.

۳- با توجه به اینکه بر اساس الگوریتم BLAST توالی‌های بزرگ به توالی‌های کوچکتر تبدیل شده و سپس BLAST انجام می‌شود، این امکان وجود دارد که یک توالی با بیش از یک نقطه در ژنوم هم‌ردیف شود. بر همین اساس برای مناطقی از ژنوم که دارای بیش از یک توالی BLAST شده بود، توالی با E-value بیشتر حذف گردید.

۴- چون برخی از توالی‌های پروتئینی همولوگ هم بودند، برخی از توالی‌های پروتئین‌های مختلف نیز با مکان مشابهی در ژنوم هم‌ردیف شوند. برای چنین مناطقی توالی‌هایی که مرتبط با پروتئین‌های مختلف بودند و به مکان مشابهی در ژنوم هم‌ردیف شده بودند، با هم ادغام شدند. همچنین در صورت وجود دو یا چند توالی هم‌ردیف شده با ژنوم که مرتبط با یک پروتئین بوده و با فاصله کمی از هم قرار گرفته بودند، به یک مجموعه واحد (به عبارتی یک توالی) تبدیل شدند. فرض بر این است که فاصله ایجاد شده بین توالی‌های هم‌ردیف شده مربوط به یک پروتئین با ژنوم ناشی از مواردی همچون: ۱) وارد شدن برخی توالی‌های DNA اضافی در ژن کاذب طی تکامل، ۲) توالی اینترونی بین اگزون‌های مرتبط با ژن‌های کاذب مضاعف شده و ۳) وجود توالی‌های تکرار شونده موجود در ژنوم که برای هم‌ردیفی اخلال ایجاد می‌کنند، می‌باشد. با استفاده از الگوریتم موجود در PseudoPipe این مناطق بررسی گردید.

۵- در این مرحله با در نظر گرفتن آستانه‌های مختلف منشأ و نوع ژن‌های کاذب بررسی شد. در این پژوهش ژن‌های کاذب با بیش از ۴۰ درصد یکسانی در توالی‌های اسیدآمینه‌ای (توالی‌های شناخته شده در ژنوم به عنوان ژن کاذب به توالی اسیدآمینه‌ای ترجمه شدند) با پروتئین والدی به‌عنوان ژن‌های کاذب پردازش شده در نظر گرفته شدند. همچنین ژن‌های کاذب کاندید با E-value $\leq 1\times10^{-10}$ به‌عنوان ژن‌های کاذب مضاعف معرفی شدند. ژن‌های کاذب پردازش شده اینترون نداشته، توالی‌های تکراری کوچک در دو انتهای خود داشته و دارای دم پلی A می‌باشند. به همین منظور الگوریتم PseudoPipe از چنین ویژگی‌های برای تفکیک ژن‌های کاذب پردازش شده و مضاعف استفاده می‌کند. همچنین ژن‌های کاذب واحد به علت احتمال وقوع مثبت دروغین[1] بالا بررسی نشدند.

۶- به منظور اطمینان از این نکته که ژن‌های کاذب شناسایی شده در محدوده ژن‌های شناخته‌شده و اصلی موجود در ژنوم نیستند، موقعیت هر یک از ژن‌های کاذب شناسایی شده در ژنوم بررسی شد و ژن‌هایی که در فاصله کمتر از ۱۰۰۰ جفت بازی یک ژن شناخته شده بودند، حذف شدند. در این مرحله اطلاعات مرتبط با ژن‌های کاذب کاندید جدید در ژنوم گوسفند شامل موقعیت کروموزومی، توالی‌های نوکلئوتیدی، نام و توالی ژن والدی ارائه گردید.

۷- در ادامه به منظور بررسی گروه‌های کارکردی ژن‌های اصلی عملکردی که ژن‌های کاذب DUP و PSSD از آن مشتق شده‌اند، از پایگاه اینترنتی DAVID استفاده شد (۲۰). بدین منظور ژن‌های اصلی عملکردی مرتبط با ژن‌های کاذب به نرم‌افزار معرفی و عبارات معنادار ($P < ۰/۰۱$) شناسایی شد. تنها عبارات معنادار مربوط به مراحل بیولوژیکی در نظر گرفته شد.

۸- تعداد و ویژگی‌های مرتبط با طول ژن‌های کاذب کاندید جدید شناسایی شده در گوسفند با ژن‌های کاذب شناخته‌شده در گونه‌های انسان، موش، گاو و گوسفند مقایسه شد. اطلاعات مربوط به توالی‌های ژن‌های کاذب گونه‌های مذکور از بانک اطلاعاتی ENSEMBL استخراج گردید. با توجه به اینکه در ژنوم انسان و موش ژن‌های کاذب به انواع پردازش شده و مضاعف تفکیک شده است، مقایسه به صورت تفکیک شده و در این سطوح انجام شد. با این وجود، در ژنوم گاو و گوسفند تنها ژن‌های کاذب پردازش شده گزارش شده است، در نتیجه، تنها ژن‌های کاذب پردازش شده در پژوهش حاضر با ژن‌های کاذب شناخته شده در این گونه‌ها مقایسه شد.

۹- در نهایت ژن‌های کاذب شناسایی شده در این پژوهش با ژن‌های کاذب گزارش شده در بانک اطلاعاتی NCBI که به با روش‌های محاسباتی شناسایی شده بودند، مقایسه شد. بدین منظور توالی‌های ژن‌های کاذب برای گوسفند از بانک اطلاعاتی NCBI استخراج شد و با استفاده از BLAST مقایسه انجام شد. تنها مقایساتی که E-value < 1e-5 داشتند، به‌عنوان ژن‌های کاذب تأیید شده در بانک اطلاعاتی NCBI در نظر گرفته شدند.

نتایج

به منظور شناسایی ژن‌های جدید ابتدا باید بر اساس ژن‌های شناخته شده مدل‌هایی را طراحی کرد و سپس از این مدل‌ها برای شناسایی ژن‌های جدید در ژنوم‌های تازه توالی یابی شده یا با حاشیه نویسی ضعیف استفاده کرد (۲۱). در پژوهش حاضر به منظور

1- False postive

شناسایی ژن‌های کاذب جدید در ژنوم گوسفند از این روش استفاده شد. با توجه به اینکه نتایج این روش به مجموعه پروتئینی اولیه بستگی دارد، انتخاب این داده‌ها بسیار مهم می‌باشد. در این پژوهش از پروتئین‌های گزارش شده مرتبط با ژنوم گوسفند از بانک اطلاعاتی ENSEMBL استفاده شد که در پژوهش‌های قبلی صحت و مناسب بودن آنها برای شناسایی ژن‌های کاذب تأیید شده است (۲۲). بدین منظور از نسخه ۷۷ این بانک اطلاعاتی پروتئین‌های شناسایی شده در ژنوم گوسفند شامل ۲۰،۰۹۱ پروتئین استخراج شد.

بعد از انجام مراحل ذکر شده در بخش مواد و روش‌ها به طور کلی ۲۷،۱۴۰ ژن کاذب که به تفکیک شامل ۴،۷۸۹ ژن کاذب پردازش شده و ۹،۲۳۸ ژن کاذب مضاعف بودند، در ژنوم گوسفند شناسایی شد. پس از حذف ژن‌های کاذبی که در محدوده ۱۰۰۰ جفت بازی ژن‌های کدکننده پروتئین شناخته شده بودند، در نهایت تعداد ۴،۰۹۸ ژن کاذب جدید که به تفکیک شامل ۲،۹۹۶ ژن کاذب پردازش شده و ۱،۱۰۲ ژن کاذب مضاعف بود، شناسایی شد (جدول ۱). همچنین در شکل یک توزیع کروموزومی ژن‌های کاذب پردازش شده و مضاعف در ژنوم گوسفند ارائه شده است.

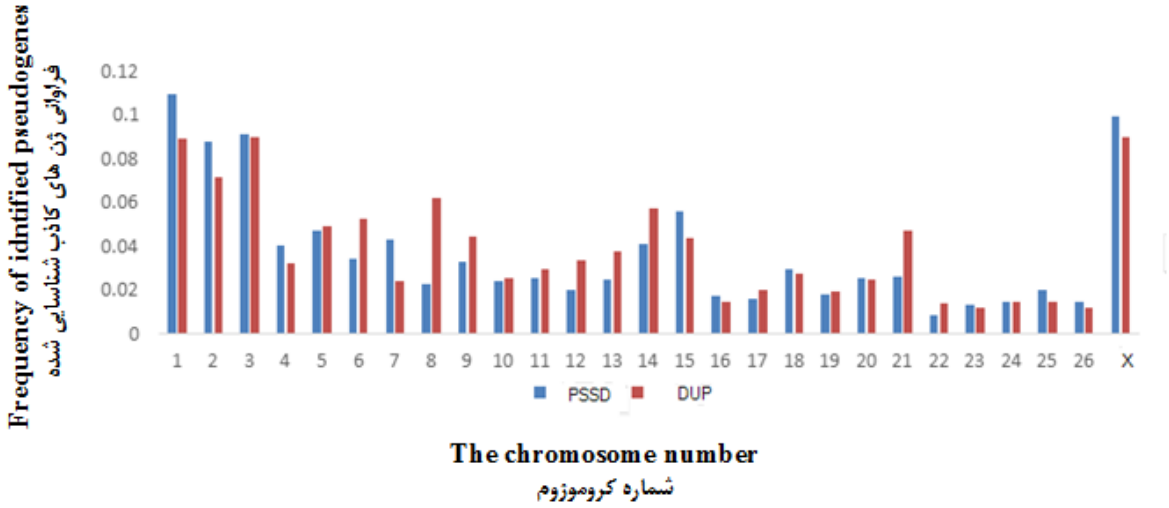
جدول ۱- خلاصه‌ای از ویژگی‌های انواع ژن‌های کاذب جدید شناسایی شده در ژنوم گوسفند

Table 1- Summary of the characteristics of thenovel identified pseudogenes in the sheep genome

	ژن‌های کاذب مضاعف خام Raw duplicated pseudogenes	ژن‌های کاذب مضاعف بعد از حذف[1] Duplicated pseudogenes after filtration[1]	ژن‌های کاذب پردازش شده خام Raw processed pseudogenes	ژن‌های کاذب پردازش شده بعد از حذف[1] Processed pseudogenes after filtration[1]
تعداد Number	9,238	1,102	4,789	2,996
میانگین طول (bp) Average length (bp)	265,565.7	4,325.22	735.596	738.893
بیشترین طول (bp) Maximum length (bp)	749,893	180,899	6,118	3,009
کمترین طول (bp) Minimum length (bp)	222	232	116	134

[1] حذف ژن‌های کاذب در محدوده کمتر از ۱۰۰۰ جفت بازی ژن‌های شناخته شده کد کننده پروتئین.

[1] Filter pseudogenes in the range of less than 1000 bp of known protein-coding genes.



شکل ۱- فراوانی انواع ژن‌های کاذب شناسایی شده در ژنوم گوسفند

PSSD: ژن کاذب پردازش شده، DUP: ژن کاذب مضاعف

Figure 1- Frequency of the identified pseudogenes in the sheep genome

PSSD: processed pseudogenes, DUP: duplicated pseudogenes.

به منظور بررسی این نکته که آیا ژن‌های والدی که ژن‌های کاذب شناسایی شده از آنها منشأ گرفته‌اند دارای عملکرد ویژه و

معناداری می‌باشند یا خیر، گروه‌های کارکردی این ژن‌ها با استفاده از پایگاه اینترنتی DAVID بررسی شد. در نتیجه بررسی گروه‌های کارکردی ۱،۱۰۲ ژن اصلی عملکردی که ژن‌های کاذب مضاعف از آنها مشتق شده، ۴۲ عبارت زیستی معنادار شناسایی شد (P<۰/۰۵) که برخی از این عبارات مرتبط با پردازش mRNA می‌باشند. همچنین در نتیجه بررسی گروه‌های کارکردی ۲،۹۹۶ ژن اصلی عملکردی که ژن‌های کاذب پردازش شده از آنها مشتق شده بود، ۱۶۷ عبارت بیولوژیکی معنادار شناسایی شد (P<۰/۰۵) که این ژن‌ها مرتبط با واژگانی همچون جزء ساختاری ریبوزوم، اتصال، زنجیره تنفسی انتقال الکترون، rRNA و غیره می‌باشند. برخی از عبارات زیستی معنادار و مهم حاصل از این تجزیه و تحلیل در جدول ۲ ارائه شده است.

جدول ۲- برخی ازگروه‌های کارکردی مهم و معنادار حاصل از بررسی گروه‌های کارکردی ژن‌های والدی مربوط به ژن‌های کاذب جدید

Table 2- Some of the important and significant functional groups associated with parental genes of the identified pseudogenes

نوع ژن کاذب Type of pseudogene	گروه کارکردی Functional group	P-value
Duplicated pseudogenes ژن‌های کاذب مضاعف	پردازش mRNA mRNA processing	0.045
	سنتز ATP میتوکندریایی همراه با انتقال الکترون ATP synthesis coupled electron transport	1.04E-06
	انتقال الکترون میتوکندریایی، NADH به یوبیکوئینون Mitochondrial electron transport, NADH to ubiquinone	0.001
	پیرایشRNA RNA splicing	0.046
	تولید پیش‌ساز متابولیت‌ها و انرژی Generation of precursor metabolites and energy	0.001
	زنجیره تنفسی انتقال الکترون Respiratory electron transport chain	3.29E-06
Processed pseudogenes ژن‌های کاذب پردازش شده	پیدایش ریبوزوم Ribosome biogenesis	0.013
	ترجمه Translation	7.02E-19
	سنتز ATP میتوکندریایی همراه با انتقال الکترون Mitochondrial ATP synthesis coupled electron transport	7.89E-10
	انتقال الکترون میتوکندریایی، NADH به یوبیکوئینون Mitochondrial electron transport, NADH to ubiquinone	1.10E-04
	طویل شدنRNA RNA elongation	0.001
	پردازشmRNA mRNA processing	1.55E-05
	پردازشrRNA rRNA processing	0.006
	شروع رونویسی بوسیله RNA پلی‌مراز II Transcription initiation from RNA polymerase II promoter	0.036
	تولید پیش‌ساز متابولیت‌ها و انرژی Generation of precursor metabolites and energy	7.66E-10
	زنجیره تنفسی انتقال الکترون Respiratory electron transport chain	6.56E-10

در ادامه تعداد، میانگین، حداکثر و حداقل طول ژن‌های کاذب کاندید شناسایی شده در پژوهش حاضر با ژن‌های کاذب شناخته‌شده در ژنوم انسان، موش و گاو مقایسه شد. نتایج مربوط به مقایسه ژن‌های کاذب در ژنوم گوسفند با گاو در جدول ۳ و با ژنوم انسان و موش در جدول ۴ ارائه شده است. لازم به ذکر است که تعداد ژن کاذب در گونه‌های مختلف بر اساس بانک اطلاعاتی ENSEMBL

می‌باشد. همانطورکه در جدول ۴ نشان داده شده است نسبت تعداد ژن کاذب پردازش شده به مضاعف شناسایی شده در پژوهش حاضر (۲/۷) با دو ژنوم انسان (۳/۹) و موش (۲/۹) مطابقت دارد بطوریکه تعداد ژن کاذب پردازش شده در هر دو ژنوم انسان و موش بیشتر از ژن کاذب مضاعف می‌باشد. همچنین ویژگی‌های طولی مربوط به این ژن‌ها در محدوده طولی به دست آمده در دو ژنوم مذکور است. لازم به ذکر است که تطابق نتایج به دست آمده در پژوهش حاضر با ژنوم موش بیشتر می‌باشد. با توجه به اینکه تعداد ژن کاذب شناخته‌شده در ژنوم گاو و گوسفند کم بوده و به انواع مختلف نیز تفکیک نشده است، استفاده از آنها جهت مقایسه با نتایج پژوهش حاضر صحیح نمی‌باشد. همانطور که در جدول ۳ ارائه شده است ویژگی‌های طولی ژن‌های کاذب شناخته شده در دو ژنوم گاو گوسفند مطابق با ژن‌های کاذب پردازش شده است که با نتایج پژوهش حاضر نیز مطابقت دارد.

جدول ۳- مقایسه ویژگی‌های مختلف ژن های کاذب شناسایی شده در پژوهش حاضر و ژن‌های کاذب شناخته شده در گوسفند وگاو (طول بر حسب جفت باز)

Table 3- Comparison of the different characteristics of the identified pseudogenes in this study and the known pseudogenes in sheep and cattle (length in base pairs)

	ژن‌های کاذب پردازش شده (پژوهش حاضر) Processed pseudogenes (current study)	ژن‌های کاذب تائید شده در ژنوم گوسفند Known pseudogenes in the sheep genome	ژن‌های کاذب شناخته شده در ژنوم گاو Known pseudogenes in the cattle genome
تعداد Number	2,996	290	796
میانگین طول (bp) Average length (bp)	738.893	983.441	858.417
بیشترین طول (bp) Maximum length (bp)	3,009	7,403	13,419
کمترین طول (bp) Minimum length (bp)	134	98	152

جدول ٤- مقایسه ویژگی‌های مختلف ژنهای کاذب شناسایی شده در پژوهش حاضر و ژن‌های کاذب شناخته شده در انسان و موش (طول بر حسب جفت باز)

Table 4- Comparison of the different characteristics of the identified pseudogenes in this study and the known pseudogenes in human and mouse (length in base pairs)

	ژن‌های کاذب (پژوهش حاضر) Pseudogenes (current study)		ژن‌های کاذب شناخته شده در ژنوم انسان Known pseudogene in human		ژن‌های کاذب شناسایی شده در موش Known pseudogene in mouse	
	ژن کاذب مضاعف DUP	ژن کاذب پردازش شده PSSD	ژن کاذب مضاعف DUP	ژن کاذب پردازش شده PSSD	ژن کاذب مضاعف DUP	ژن کاذب پردازش شده PSSD
تعداد Number	1,102	2,996	2,661	10,270	2,344	6,781
میانگین طول (bp) Average length (bp)	4,325.22	738.893	5,416.64	789.925	4,723.97	781.77
بیشترین طول (bp) Maximum length (bp)	180,899	3,009	263,511	14,336	174,114	56,311
کمترین طول (bp) Minimum length (bp)	232	134	27	22	40	17

همچنین ویژگی‌های ژن‌های کاذب شناخته شده در پژوهش حاضر با ژن‌های کاذب پیش‌بینی شده با روش‌های محاسباتی موجود در بانک اطلاعاتی NCBI نیز مقایسه شد. با توجه به اینکه نوع ژن کاذب در این بانک اطلاعاتی گزارش نشده است همه ژن‌های کاذب و بدون در نظر گرفتن نوع آنها بررسی شد. نتایج نشان داد که میانگین، بزرگترین و کمترین طول ژن کاذب در ژن‌های کاذب پیش‌بینی شده NCBI به ترتیب برابر با ۵،۵۰۸، ۲۹۷،۳۸۱ و ۱۳۳ نوکلئوتید می‌باشد. این مقادیر به ترتیب برای ژن‌های شناسایی شده در مطالعه حاضر برابر با ۱،۷۰۴، ۱۸۰،۹۰۰ و ۱۳۵ نوکلئوتید بود. تفاوت در مقادیر مربوط به طول ژن‌ها می‌تواند ناشی از روش‌های مورد استفاده محاسباتی باشد که منجر به شناسایی برخی ژن‌های با طول متفاوت شده و این موضوع در میانگین طول بدست آمده تأثیر

گذاشته است. با این حال مقایسه صحیح‌تر زمانی می‌تواند صورت گیرد که نوع ژن‌های کاذب در دسترس باشد. با این وجود به منظور بررسی دقیق‌تر با استفاده از BLAST مقایسه بین این ژن‌ها صورت گرفت. نتایج نشان داد که ۸۹۹ ژن از ۱،۱۰۲ ژن کاذب مضاعف و ۲،۰۵۸ ژن از ۲۹،۹۶ ژن پردازش شده با ژن‌های کاذب گزارش شده در NCBI تطابق داشت (E-value < 1e-5) که بیانگر وجود تطابق نتایج مطالعه حاضر با نتایج گزارش شده در NCBI می‌باشد. همچنین در جدول ۵ برخی از ژن‌های کاذب شناسایی شده در این پژوهش به همراه طول و موقعیت کروموزومی و نوع آنها ارائه شده است.

جدول ۵- برخی از ژن‌های کاذب شناسایی شده در این پژوهش به همراه طول و موقعیت کروموزومی آنها

Table 5- Some of the identified pseudogenes in this study along with the length and chromosomal location

ژن‌های کاذب (پژوهش حاضر) Pseudogenes (current study)	نوع ژن‌های کاذب Type of pseudogenes	موقعیت کروموزومی Location of chromosome	طول کروموزومی (bp) Chromosome length (bp)	شماره کروموزوم Number of chromosome
GAPDH	PSSD	62535898 to 62536829	931	3
RPS19	PSSD	25674203 to 25674656	453	4
TPT1	PSSD	63895193 to 63895652	459	17
COX5B	PSSD	60933960 to 60934372	412	2
COX2	PSSD	72240933 to 72241466	533	3
NDUFV2	PSSD	52341595 to 52342347	752	16
ST13	PSSD	161897102 to 161898274	1172	1
COX1	DUP	60837816 to 60839145	1329	3
COIII	DUP	60836186 to 60837321	1135	3
HSPA8	DUP	86559011 to 86560981	1970	6
ND1	DUP	8885302 to 8886458	1156	9
ATP6	DUP	43980096 to 43980988	892	23
CYTB	DUP	60710432 to 60711677	1245	7
TAF9B	DUP	18623829 to 18624494	665	1

بحث

ژن‌های کاذب معمولا نسخه‌هایی از ژن اجدادی و اصلی می‌باشند که به مرور زمان فعالیت آنها نسبت به ژن اولیه تغییر کرده است و بر اثر فرآیندهایی مانند مضاعف شدگی ژنی و همچنین رونویسی معکوس در ژنوم ایجاد شده‌اند (۲۳). در سال‌های اخیر با توجه به پیشرفت‌های صورت گرفته در روش‌های توالی‌یابی و درک بهتر ما از ژنوم، به ژن‌های کاذب توجه بیشتری شده است. در همین زمینه اخیرا پژوهش‌ها نشان داده‌اند که در برخی موارد، ژن‌های کاذب بیان شده و می‌توانند نقش‌های تنظیمی بسیار مهمی داشته باشند (۲۴-۲۷). برخی از این ژن‌ها می‌توانند بیان ژن اصلی خود را به وسیله کاهش پایداری mRNA ژن عملکردی از طریق بیان بالایشان تنظیم کنند. به‌عنوان مثال ژن کاذب MYLKP1 که در سلول‌های سرطانی بیان بالایی دارد (۲۸) یک RNA غیرکدشونده[1] ایجاد می‌کند که بیان mRNA والدی‌اش (MYLK) را مهار می‌کند (۱۳). علاوه بر این، پژوهش‌های انجام شده روی موش و دروزوفیلا نشان داده‌اند که siRNA[2] مشتق شده از ژن‌های کاذب پردازش شده می‌توانند بیان ژن را با استفاده از مسیرهای RNA مداخله‌گر[3] تنظیم کنند (۲۵, ۲۹ و ۳۰)، بنابراین به‌عنوان siRNAs با منشأ داخلی عمل می‌کنند. علاوه بر این چنین عملکردهایی، فرض شده است که ژن‌های کاذب با تشابه توالی بالا با ژن‌های والدی، می‌توانند بیان ژن‌های والدی خود را از طریق تولید رونوشت‌های آنتی-سنس تنظیم کنند (به‌عنوان مثال تنظیم ژن Oct4) (۳۱). برخی پژوهش‌ها نیز گزارش کرده‌اند که ژن‌های کاذب می‌توانند با ژن‌های والدی خود برای اتصال به miRNAای که ژن والدی را تنظیم می‌کند، رقابت کنند. بنابراین با اتصال ژن کاذب به miRNA، بیان ژن والدی افزایش می‌یابد. برای مثال ژن کاذب PTEN1، یک سرکوبگر حیاتی تومور، بیان ژن والدی‌اش را از طریق این سازوکار تنظیم می‌کند (۲۴ و ۳۲). بر همین اساس در پژوهش حاضر با استفاده از روش‌های محاسباتی، برای اولین بار انواع ژن‌های کاذب موجود در ژنوم گوسفند به منظور کامل کردن حاشیه نویسی ژنوم مرجع گوسفند بررسی گردید.

در پژوهش حاضر ۴،۰۹۸ ژن کاذب در ژنوم گوسفند شناسایی شد. این تعداد در مقایسه با گونه‌هایی مانند انسان و موش که حاشیه نویسی ژنومشان کامل‌تر می‌باشد، کمتر می‌باشد. یکی از دلائل بسیار مهم می‌تواند کامل نبودن سرهم‌بندی و حاشیه نویسی ژنوم گوسفند باشد. لازم به ذکر است که چنین پدیده‌ای در حاشیه نویسی ژنوم انسان نیز مشاهده شده است. به طوری که تعداد ژن کاذب شناسایی شده در ژنوم انسان با کامل‌تر شدن حاشیه نویسی ژنومی آن افزایش

1- ncRNA
2- Small interfering RNA
3- RNA interfering

یافته است (۲۲). با توجه به ماهیت محاسباتی بودن روش مورد استفاده در پژوهش حاضر این امکان وجود دارد که برخی از ژن‌های کاذب شناسایی شده در پژوهش حاضر مثبت دروغین باشند. یکی از دلائل این موضوع می‌تواند حاشیه نویسی اشتباه در ژنوم گوسفند باشد. به‌عنوان مثال در ابتدا تصور بر این بود که در حدود ۱۹ هزار ژن کاذب در ژنوم انسان وجود دارد (۳۳). با این حال پژوهش‌های اخیر نشان داده است که حدود ۱۲ هزار ژن کاذب در ژنوم انسان وجود دارد. علت این کاهش در مقدار ژن کاذب در ژنوم انسان بهبود حاشیه نویسی ژنوم بیان شده است. با توجه به اینکه شناسایی ژن‌های کاذب به تعداد ژن‌های کد کننده پروتئین شناخته شده بستگی دارد در نتیجه در صورتی که با بهبود حاشیه نویسی ژنوم تعداد ژن‌های کدکننده پروتئین کاهش یابد (حذف برخی از ژن‌هایی که قبلا به اشتباه به‌عنوان ژن کدکننده شناخته شده‌اند) تعداد ژن‌های کاذب نیز کاهش خواهد یافت (۳۴). وقوع چنین حالتی در پژوهش حاضر نیز دور از انتظار نیست. با این حال بر اساس مطالعات انجام شده بهترین روش شناسایی ژن‌های کاذب روش‌های بر اساس بررسی مشابهت ژن‌های کدکننده پروتئین در ژنوم موجود مورد نظر می‌باشد (۲۲). همچنین در پژوهش حاضر پس از شناسایی ژن‌های کاذب، مکان این ژن‌ها به صورت دستی در ژنوم بررسی شد که باعث کاهش خطای مثبت دروغین و افزایش صحت نتایج گردید.

گزارش شده است که بین طول کروموزوم و تعداد ژن‌های کاذب مضاعف موجود در آن کروموزوم در انسان همبستگی مثبت وجود دارد (۱۹). در پژوهش حاضر نیز چنین رابطه‌ای مشاهده شده به طوری که توزیع کروموزومی ژن‌های کاذب شناسایی شده در این بررسی با توزیع کروموزومی ژن‌های کاذب گزارش شده انسانی در پژوهش (۱۹) تطابق داشت. همچنین، ژن‌های کاذب مضاعف به واسطه سازوکار ایجاد آنها در بیشتر مواقع در کروموزوم مشابه با ژن والدی وارد می‌شوند. بر همین اساس گزارش شده است که برخی از ژن‌های کاذب مضاعف در نزدیکی ژن‌های والدی خود در ژنوم قرار دارند (۱۹). بر همین اساس برای ژن‌های کاذب مضاعف شناسایی شده در پژوهش حاضر این موضوع بررسی شد. نتایج نشان داد که ۳۰۹ ژن کاذب مضاعف (از ۱،۱۰۲ ژن) در کروموزوم مشابه با ژن والدی خود قرار دارند. از این تعداد ۱۴۴ ژن در فاصله کمتر از ۱۰۰ هزار جفت بازی نسبت به ژن والدی قرار داشتند.

ژن‌های والدی که ژن‌های کاذب از آنها مشتق شده‌اند، برای بررسی عملکرد تنظیمی احتمالی ژن‌های کاذب و تاریخچه تکاملی آنها بسیار مهم می‌باشند (۲۲). بر همین اساس گروه‌های کارکردی ژن‌های والدی ژن‌های کاذب شناسایی شده در پژوهش حاضر بررسی شد. نتایج منجر به شناسایی به ترتیب ۴۲ و ۱۶۷ عبارت بیولوژیکی معنادار برای ۱،۱۰۲ ژن اصلی عملکردی که ژن‌های کاذب مضاعف و ۲۹۹۶ ژن اصلی عملکردی که ژن‌های کاذب پردازش شده از آنها مشتق شده‌اند شد (جدول ۲).

نتایج نشان داد که تعدادی از ژن‌های کاذب شناسایی شده مرتبط با ریبوزوم از لحاظ آماری معنادار است. شناسایی ژن‌های کاذب مرتبط با ژن‌های ریبووزومی در مطالعات قبلی اثبات شده است. به بیان دیگر دسته بزرگی از ژن‌های کاذب مرتبط با ژن‌های ریبوزومی می‌باشند (۳۵). علت وجود چنین ژن‌های کاذبی، بیان بالای این ژن‌ها بر اساس نیاز بالای سلول به آنها عنوان شده است. گزارش شده است که احتمال تبدیل ژن‌های با بیان بالا به ژن‌های کاذب به وسیله سازوکار رتروترنسپوزون‌ها بسیار بالا می‌باشد (۳۶). پروتئین‌های ریبوزومی، یکی از بزرگترین دسته ژن‌های تکرار شده و حفاظت شده را در ژنوم پستانداران تشکیل می‌دهند (۳۷). همچنین پروتئین‌های ریبوزومی نقش مهمی در سنتز پروتئین بازی کرده و در سطوح بالا بیان شده و دارای روند تکاملی بسیار آهسته‌ای می‌باشند (۳۸). برخی از ژن‌های کاذب شناسایی شده مرتبط با ژن والدی ریبوزومی در پژوهش حاضر عبارتند از RPS4X، RPL3L، RPL26L، RPS27L و RPSL24D1. همچنین پروتئین‌های ریبوزومی RPL19، RPL22، RPS17، RPS27A، RPS28، RPS29، ATP5B و TPT1 که به‌عنوان پروتئین‌های ریبوزومی در انسان گزارش شده‌اند (۳۷, ۳۹) ، در پژوهش حاضر نیز شناسایی شد. دیگر ژن کاذب گزارش شناسایی شده در پژوهش حاضر، پروتئین ریبوزومی RPS19 می‌باشد که گزارش شده است در ایجاد کم خونی دخالت دارد (۴۰). در مطالعات قبلی عنوان شده است که RPL3L یک پروتئین ریبوزومی DNA-mediated[1] می‌باشد و نسبت به ژن والدی-اش، RPL3، در برخی بافتها بیان بالاتری داشته که نشانه عملکردی بودن ژن کاذب مذکور می‌باشد (۴۱). همچنین در رابطه با ژن کاذب مضاعف RPL10L گزارش شده است که در جبران دز ژن‌های پروتئین‌های ریبوزومی وابسته به کروموزوم X دخالت دارد (۴۲).

همچنین GAPDH، ST13، TPT1 وTDGF1 از ژن‌های کاذب پردازش شده در این تحقیق می‌باشند. ژن کاذب GAPDH یکی از آنزیم‌های کلیدی در گلیکولیز می‌باشد (۴۳). گزارش شده است که در موش، خانواده چندژنی GAPDH شامل بیش از ۳۰۰ ژن کاذب می‌باشد (۴۴). ژن GAPDH یکی از ژن‌های خانه‌دار نیز می‌باشد که دارای بیان بالا در همه بافت‌هاست. در همین رابطه بیان شده است که ژن‌های خانه‌دار دارای ژن‌های کاذب بالایی می‌باشند که در توافق با پژوهش حاضر می‌باشد (۴۵). ژنST13 به‌عنوان یکی دیگر از ژن‌های کاذب پردازش شده در انسان گزارش شده است که در سرطان روده نقش دارد (۴۳). در رابطه با ژن کاذب TPT1 گزارش شده است که عامل آزاد کننده هیستامین بوده و ترجمه پروتئین‌های سرطانی در انسان را کنترل می‌کند (۴۳). همچنین ژن کاذب

1- DNA mediated ribosomal protein

TDGF1 نقش مهمی در توسعه مغز قدامی در انسان بازی می‌کند (۴۳). یکی دیگر از ژن‌های کاذب مضاعف یافت شده در پژوهش حاضر KRAS می‌باشد که در مطالعات قبلی نقش KRAS و ژن کاذبش (KRAS1P) در سرطان پروستات در انسان گزارش شده است (۲۴). ژن کاذب LALBA که در این تحقیق در نتایج هر دو نوع ژن کاذب مضاعف و پردازش شده شناسایی گردید، به‌عنوان ژن کاذب در کروموزوم پنج گاو گزارش شده است (۴۶).

گزارشات حاکی از وجود برخی ژن‌های کاذب در ژنوم پستانداران می‌باشد که پس از بیان نقش تنظیمی خود را به صورت miRNAs بازی می‌کنند. در همین راستا دو ژن کاذب Ppp4r و HDAC1 که در نتایج این پژوهش نیز حضور داشت در مطالعات پیشین به عنوان miRNAs مشتق شده از ژن‌های کاذب در اووسیت‌های موش گزارش شده است (۲۵). همچنین، مطالعات اخیر نشان داده‌اند که بسیاری از ژن‌های کاذب در بافت‌های نرمال از طریق توسعه و تمایز در پاسخ به محرک‌های مختلف بصورت RNAهای غیر کدشونده بلند (lncRNA) عمل کرده (۴۷) و در خاموشی ژن (۴۸) و سرطان (۳۶)نقش دارند. ژن‌های کاذب Rps15a،COX2،RelA، Nfkbia، Sod2 وIl6 از ژن‌های کاذب با چنین ویژگی می‌باشند که در پژوهش حاضر نیز شناسایی شده و در موش نیز گزارش شده‌اند (۴۹).

یکی از خانواده‌های بزرگ ژنی در ژنوم پستانداران، خانواده ژنی مربوط بهtRNA می‌باشد. این خانواده دارای ژن‌های کاذب فراوانی می‌باشد که به نظر می‌رسد عامل ایجاد این ژن‌ها ترانسپوزیشن وابسته به RNA است (۵۰ و ۵۱). یکی از نتایج جالب در این پژوهش معنادار شدن عبارت ترجمه می‌باشد که بیان می‌کند تعداد ژن‌های کاذب که ژن‌های والدی آنها مرتبط با tRNA و ترجمه می‌باشند، معنادار است. به‌عنوان مثال یکی از ژن‌های کاذب شناسایی شده در این تحقیق ILF3 (یک ncRNA شامل ۱۱۷ نوکلئوتید را رمز می‌کند) می‌باشد که به‌عنوان هدف Pol III گزارش شده است (۵۲). دیگر ژن مرتبط با ترجمه که در این تحقیق شناسایی شد ژن PTEN می‌باشد که در مطالعات قبلی به عنوان ژن کاذب شناسایی و نقش تنظیمی آن اثبات شده است (۵۳).

از دیگر عبارات بیولوژیکی معنادار شناسایی شده در این پژوهش mRNA splicing می‌باشد. در همین زمینه و در رابطه با ژن NF1 که در پژوهش حاضر به عنوان ژن کاذب شناسایی شد، گزارش شده است که این ژن یکی از شایع‌ترین علل ناهنجاری‌های وراثتی در انسان را با شیوع یک در ۳،۰۰۰ موجب می‌شود (۵۴). از دیگر ژن‌های کاذب شناسایی شده در این پژوهش که در گزارشات قبلی نقش آنها به‌عنوان ژن کاذب تأیید شده است می‌توان به MTIF2، NDUFB7، NDUFV2، COX5B، APOE و PDK1 اشاره کرد. ژن MTIF2 ترجمه پروتئین‌های رمزشده بوسیله ژنوم میتوکندریایی را تنظیم می‌کند (۵۵). ژن NDUFB7 که به‌عنوان ژن کاذب در انسان شناسایی شده، در زنجیره انتقال الکترون نقش دارد (۵۶). همچنین NDUFV2 به‌عنوان ژن کاذب میتوکندریایی در انسان گزارش شده است (۵۷). ژن COX5B شامل پنج اگزون و چهار اینترون بوده و نقش آن در فسفریلاسیون اکسیداتیو در انسان گزارش شده است (۵۸). نقش ژن APOE در ناهنجاری‌های متابولیسم انرژی در بیماری آلزایمر گزارش شده است (۵۹). گزارش شده است که ژن PDK1 در تنظیم متابولیکی سلول‌های بنیادی پرتوان طی بازسازی و برنامه‌ریزی مجدد نقش دارد (۶۰).

یکی از محدودیت‌های روش‌های بر پایه بررسی تشابه به منظور شناسایی ژن‌های کاذب این است که این روش‌ها بر اساس ژن‌های (و پروتئین‌ها) کد کننده پروتئین شناخته‌شده، ژن‌های کاذب را شناسایی می‌کنند. در نتیجه در صورتی که اشتباهی در حاشیه نویسی اولیه ژنوم مورد نظر وجود داشته باشد نتایج تحت تأثیر قرار خواهد گرفت. با این وجود بهترین روش شناسایی این ژن‌ها در حال حاضر استفاده از چنین روشی می‌باشد (۲۲). در این پژوهش نیز سعی شد با در نظر گرفتن بهترین حاشیه نویسی ژنوم موجود در گوسفند (بانک اطلاعاتی ENSEMBLE) و در نظر گرفتن پارامترهای سخت‌گیرانه تا حد امکان از اشتباهات احتمالی جلوگیری شود. تطابق نتایج حاصله با پژوهش‌های گذشته نیز حاکی از موفقیت روش مورد استفاده در این پژوهش می‌باشد.

به طور کلی در این پژوهش با استفاده از روش‌های محاسباتی ۴۰،۰۹۸ ژن کاذب در ژنوم گوسفند شناسایی شد. لازم به ذکر است که تعداد ژن کاذب شناسایی شده در پژوهش حاضر در مقایسه با تعداد گزارش شده در ژنوم گوسفند بسیار بالاتر می‌باشد. باید توجه داشت که ژن‌های گزارش شده در ژنوم گوسفند بر اساس تائیدات آزمایشگاهی می‌باشد. با توجه به تعداد مطالعات کم در این گونه در مقایسه با انسان و موش و همچنین سایر حیوانات اهلی مانند گاو، تعداد کم ژن کاذب تأیید شده قابل توجیه می‌باشد. در این پژوهش بر اساس روش‌های محاسباتی و با استفاده از روشی معتبر که در پروژه ENCODE استفاده شده است، اقدام به شناسایی محاسباتی این ژن‌ها در ژنوم گوسفند شد. همچنین در ادامه از پارامترهای سخت‌گیرانه دیگری برای کاهش خطای مثبت دروغین نیز استفاده شد. در مقایسه با تعداد ژن‌های کاذب گزارش شده در انسان و موش انتظار بر این است که در ژنوم سایر پستانداران مانند گوسفند نیز تعداد ژن کاذب بالا باشد. با این حال تعداد مطالعات صورت گرفته در این زمینه بسیار کم می‌باشد و کم بودن تعداد گزارش شده این ژن‌ها در ژنوم حاکی از این مطلب می‌باشد. همچنین مقایسه نتایج به دست آمده در این پژوهش با ژن‌های کاذب پیش‌بینی شده در بانک اطلاعاتی NCBI نشان داد که بیش از ۷۰ درصد این ژن‌ها مشابه می‌باشند. تطابق نتایج حاصل از این پژوهش با مطالعات قبلی بیانگر صحت و درستی روش مورد استفاده در این تحقیق می‌باشد. نتایج

حاصل از این پژوهش به حاشیه نویسی بهتر ژنوم گوسفند کمک خواهد کرد. همچنین با استفاده از چنین روش‌هایی می‌توان حاشیه نویسی ژنوم‌های موجودات مختلف را بهبود بخشید.

منابع

1- Jäger, M., C. E. Ott., J. Grünhagen., J. Hecht., H. Schell., S. Mundlos., G. N. Duda., P. N. Robinson, and J. Lienau. 2011. Composite transcriptome assembly of RNA-seq data in a sheep model for delayed bone healing. BMC Genomics, 12(1): 1-12.
2- Junker, B. H, and F. Schreiber. 2011. Analysis of biological networks. Vol. 2. John Wiley & Sons, New York.
3- Birzele, F., J. Schaub., W. Rust., C. Clemens., P. Baum., H. Kaufmann., A. Weith., T. W. Schulz, and T. Hildebrandt. 2010. Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. Nucleic Acids Research, 38(12): 3999-4010.
4- Derrien, T., R. Johnson., G. Bussotti., A. Tanzer., S. Djebali., H. Tilgner., G. Guernec., D. Martin., A. Merkel, and D. G. Knowles. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Research, 22(9): 1775-1789.
5- Chen, S. M., K. Y. Ma, and J. Zeng. 2011. Pseudogene: lessons from PCR bias, identification and resurrection. Molecular Biology Reports, 38(6): 3709-3715.
6- Zhang, Z., N. Carriero., D. Zheng., J. Karro., P. M. Harrison, and M. Gerstein. 2006. PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics, 22(12): 1437-1439.
7- Mighell, A., N. Smith., P. Robinson, and A. Markham. 2000. Vertebrate pseudogenes. FEBS Letters, 468(2-3): 109-114.
8- Harrison, P. M., N. Echols, and M. B. Gerstein. 2001. Gerstein, Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. Nucleic Acids Research, 29(3): 818-830.
9- Echols, N., P. Harrison., S. Balasubramanian., N. M. Luscombe., P. Bertone., Z. Zhang, and M. Gerstein. 2002. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucleic Acids Research, 30(11): 2515-2523.
10- Balakirev, E. S, and F. J. Ayala. 2003. Pseudogenes: are they "junk" or functional DNA?. Annual Review of Genetics, 37(1): 123-151.
11- Zhang, Z. D., A. Frankish., T. Hunt., J. Harrow, and M. Gerstein. 2010. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biology, 11(3): 1-17.
12- Harrison, P. M, and M. Gerstein. 2002. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. Journal of Molecular Biology, 318(5): 1155-1174.
13- Pei, B., C. Sisu., A. Frankish., C. Howald., L. Habegger., X. J. Mu., R. Harte., S. Balasubramanian., A. Tanzer, and M. Diekhans . 2012. The GENCODE pseudogene resource. Genome Biology, 13(9): 1-26.
14- Poliseno, L. 2014. Pseudogenes: Functions and Protocols. Humana Press, New Yok.
15- Ding, W., L. Lin., B. Chen, and J. Dai. 2006. L1 elements, processed pseudogenes and retrogenes in mammalian genomes. IUBMB Life, 58(12): 677-685.
16- Torrents, D., M. Suyama., E. Zdobnov, and P. Bork. 2003. A genome-wide survey of human pseudogenes. Genome Research, 13(12): 2559-2567.
17- Balasubramanian, S., D. Zheng., Y. J. Liu., G. Fang., A. Frankish., N. Carriero., R. Robilotto., P. Cayting, and M. Gerstein. 2009. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. Genome Biology, 10(1): 1-10.
18- Khurana, E., H. Y. Lam., C. Cheng., N. Carriero., P. Cayting, and M. B. Gerstein. 2010. Segmental duplications in the human genome reveal details of pseudogene formation. Nucleic Acids Research, 38(20): 6997-7007.
19- Sisu, C., B. Pei., J. Leng., A. Frankish., Y. Zhang., S. Balasubramanian., R. Harte., D. Wang., M. Rutenberg-Schoenberg, and W. Clark. 2014. Comparative analysis of pseudogenes across three phyla. Proceedings of the National Academy of Sciences, 111(37): 13361-13366.
20- Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols, 4(1): 44-57.
21- Brent, M. R, and R. Guigo. 2004. Recent advances in gene structure prediction. Current opinion in structural Biology, 14(3): 264-272.
22- Zheng, D, and M. B. Gerstein. 2006. A computational approach for identifying pseudogenes in the ENCODE regions. Genome Biology, 7(1): 1-10.
23- Mehraban, M., J. Jamshidi, and S. Vallian. 2014. Gene Families: Structure, Organization and Evolution. Journal of Fasa University of Medical Sciences, 4(2): 134-153. (In Persian).
24- Poliseno, L., L. Salmena., J. Zhang., B. Carver., W. J. Haveman, and P. P. Pandolfi. 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature, 465(7301): 1033-1038.

25- Oliver, H. T., A. A. Arvin., P. Stein., A. Girard., E. P. Murchison., S. Cheloufi., E. Hodges., M. Anger., R. Sachidanandam, and R. M. Schultz. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature, 453(7194): 534-538.
26- Piehler, A. P., M. Hellum., J. J. Wenzel., E. Kaminski., K. B. Haug., P. Kierulf, and W. E. Kaminski. 2008. The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference. BMC genomics, 9(1):1-13.
27- Muro, E. M., N. Mah, and M. A. Andrade-Navarro. 2011. Functional evidence of post-transcriptional regulation by pseudogenes. Biochimie, 93(11): 1916-1921.
28- Han, Y. J., S. F. Ma., G. Yourek., Y. D. Park, and J. G. Garcia. 2011. A transcribed pseudogene of MYLK promotes cell proliferation. The FASEB Journal, 25: 2305-2312.
29- Watanabe, T., Y. Totoki., A. Toyoda., M. Kaneda., S. Kuramochi-Miyagawa., Y. Obata., H. Chiba., Y. Kohara., T. Kono, and T. Nakano. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature, 453(7194): 539-543.
30- Guo, X., Z. Zhang, M. B. Gerstein and D. Zheng. 2009. Small RNAs originated from pseudogenes: cis-or trans-acting?. PLOS Computational Biology, 5(7): 1-15.
31- Hawkins, P. G., and K. V. Morris. 2010. Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. Transcription, (3): 165-175.
32- Salmena, L., A. Carracedo, and P. P. Pandolfi. 2008. Tenets of PTEN tumor suppression. Cell, 133(3): 403-414.
33- Zhang, Z., N. Carriero, and M. Gerstein. 2004. Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends in Genetics, 20(2): 62-67.
34- Clamp, M., B. Fry., M. Kamal., X. Xie., J. Cuff., M. F. Lin., M. Kellis., K. Lindblad-Toh, and E. S. Lander. 2007. Distinguishing protein-coding and noncoding genes in the human genome. Proceedings of the National Academy of Sciences, 104(49): 19428-19433.
35- Zhang, Z., P. Harrison, and M. Gerstein. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Research, 12(10): 1466-1482.
36- Kalyana-Sundaram, S., C. Kumar-Sinha., S. Shankar., D. R. Robinson., Y. M. Wu., X. Cao., I. A. Asangani., V. Kothari., J. R. Prensner, and R. J. Lonigro. 2012. Expressed pseudogenes in the transcriptional landscape of human cancers. Cell, 149(7): 1622-1634.
37- Dharia, A. P., A. Obla., M. D. Gajdosik., A. Simon, and C. E. Nelson. 2014. Tempo and mode of gene duplication in mammalian ribosomal protein evolution. Plos One, 9(11): 1-15.
38- ori, H., K. i. Higo, and S. Osawa. 1977. The rates of evolution in some ribosomal components. Journal of Molecular Evolution, 9(3): 191-201.
39- Gupta, V, and J. R. Warner. 2014. Ribosome-omics of the human ribosome. RNA, 20(7): 1004-1013.
40- Draptchinskaia, N., P. Gustavsson., B. Andersson., M. Pettersson., I. Dianzani., S. Ball., G. Tchernia., J. Klar., H. Matsson, and D. Tentler.1999. The gene encoding ribosomal protein S19 is mutated in Diamond-Blackfan anaemia. Nature Genetics, 21(2): 169-175.
41- Thorrez, L., K. Van Deun., L. C. Tranchevent., L. Van Lommel., K. Engelen., K. Marchal., Y. Moreau., I. Van Mechelen, and F. Schuit. 2008. Using ribosomal protein genes as reference: a tale of caution. Plos One, 3(3): 1-8.
42- Uechi, T., N. Maeda., T. Tanaka, and N. Kenmochi. 2002. Functional second genes generated by retrotransposition of the X-linked ribosomal protein genes. Nucleic Acids Research, 30(24): 5369-5375.
43- Zhang, Z., P. M. Harrison., Y. Liu, and M. Gerstein. 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Research, 13(12): 2541-2558.
44- Garcia-Meunier, P., M. Etienne-Julan., P. Fort., M. Piechaczyk, and F. Bonhomme. 1993. Concerted evolution in the GAPDH family of retrotransposed pseudogenes. Mammalian Genome, 4(12): 695-703.
45- Liu, Y. J., D. Zheng., S. Balasubramanian., N. Carriero., E. Khurana., R. Robilotto, and M. B. Gerstein. 2009. Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the anomalously high number of GAPDH pseudogenes highlights a recent burst of retrotrans-positional activity. BMC Genomics, 10(1): 1-12.
46- Threadgill, D. W, and J. E. Womack. 1990. Genomic analysis of the major bovine milk protein genes. Nucleic Acids Research, 18(23): 6935-6942.
47- Chang, A. L. S., P. H. Bitter., K. Qu., M. Lin., N. A. Rapicavoli, and H. Y. Chang. 2013. Rejuvenation of gene expression pattern of aged human skin by broadband light treatment: a pilot study. Journal of Investigative Dermatology, 133(2): 394-402.
48- Duret, L., C. Chureau., S. Samain., J. Weissenbach, and P. Avner. 2006. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science, 312(5780): 1653-1655.
49- Rapicavoli, N. A., K. Qu., J. Zhang., M. Mikhail., R.-M. Laberge, and H. Y. Chang. 2013. A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics. Elife, 2: e00762.
50- Deininger, P. L, and M. A. Batzer, 1993. Evolution of retroposons. Pages 157-196 in Evolutionary Biology. vol 27. M. K. Hecht., R. J. MacIntyre, and M. T. Clegg. ed. Springer, Boston, MA.
51- McBride, O. W., I. L. Pirtle, and R. M. Pirtle. 1989. Localization of three DNA segments encompassing tRNA

genes to human chromosomes 1, 5, and 16: Proposed mechanism and significance of tRNA gene dispersion. Genomics, 5(3): 561-573.
52- Raha, D., Z. Wang., Z. Moqtaderi., L. Wu., G. Zhong., M. Gerstein., K. Struhl, and M. Snyder. 2010. Close association of RNA polymerase II and many transcription factors with Pol III genes. Proceedings of the National Academy of Sciences, 107(8): 3639-3644.
53- Johnsson, P., A. Ackley., L. Vidarsdottir., W. O. Lui., M. Corcoran., D. Grandér, and K. V. Morris. 2013. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. Nature Structural and Molecular Biology, 20(4): 440-446.
54- Ars, E., E. Serra., J. García., H. Kruyer., A. Gaona., C. Lázaro, and X. Estivill. 2000. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Human Molecular Genetics, 9(2): 237-247.
55- Overman, R. G., P. J. Enderle., J. M. Farrow., J. E. Wiley, and M. A. Farwell. 2003. The human mitochondrial translation initiation factor 2 gene (MTIF2): transcriptional analysis and identification of a pseudogene. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression, 1628(3): 195-205.
56- Emahazion, T., A. Beskow., U. Gyllensten, and A. Brookes. 1998. lntron based radiation hybrid mapping of 15 complex I genes of the human electron transport chain. Cytogenetic and Genome Research, 82(1-2): 115-119.
57- de Coo, R., P. Buddiger., H. Smeets., A. G. van Kessel., J. Morgan-Hughes., D. O. Weghuis., J. Overhauser, and B. van Oost. 1995. Molecular cloning and characterization of the active human mitochondrial NADH: ubiquinone oxidoreductase 24-kDa gene (NDUFV2) and its pseudogene. Genomics, 26(3): 461-466.
58- Lomax, M. I., C. L. Hsieh., B. T. Darras, and U. Francke. 1991. Structure of the human cytochrome c oxidase subunit Vb gene and chromosomal mapping of the coding gene and of seven pseudogenes. Genomics, 10(1): 1-9.
59- Blass, J. P., R. K. F. SHEU, and G. E. Gibson. 2000. Inherent abnormalities in energy metabolism in Alzheimer disease: interaction with cerebrovascular compromise. Annals of the New York Academy of Sciences, 903(1): 204-221.
60- Zhang, J., E. Nuebel., G. Q. Daley., C. M. Koehler, and M. A. Teitell. 2012. Metabolic regulation in pluripotent stem cells during reprogramming and self-renewal. Cell Stem Cell, 11(5): 589-595.

Iranian Journal of Animal Science Research
Vol. 9, No. 4, Winter 2018, p. 484-497

نشریه پژوهشهای علوم دامی ایران
جلد ۹، شماره ۴، زمستان ۱۳۹۶، ص. ۴۹۷-۴۸۴

Prediction of Novel Pseudogenes in Ovine Reference Genome

M. R. Bakhtiarizadeh[1*]- Z. Mozdouri [2]- P. Shakeri [3]

Received: 23-11-2016
Accepted: 17-04-2017

Introduction Pseudogenes are copies of the ancestral genes which have undergone changes that were constructed based on gene duplications and reverse transcription in the genome. They have been reported in all types of organisms ranging from bacteria to mammals. Pseudogenes increase the genetic diversity of a plethora of genes and they do so through gene conversion and recombination. Three classes of pseudogenes are known to exist: duplicated pseudogenes; processed or retrotransposed pseudogenes; and unitary or disabled pseudogenes. Pseudogenes have long been considered as nonfunctional genomic sequences. However, recent studies reported that many of them might have some form of biological activity. Recently, it has reported that pseudogenes represent a conspicuous part of the human transcriptome and proteome, as thousands of them are transcribed and hundreds are also translated. Also, it has been demonstrated that pseudogenes exert important coding-dependent and coding-independent functions that are involved in complex regulatory networks. Hence, the possibility of functionality of these genes, has increased interest in their accurate annotation. According to the best of our knowledge, there is no available report on the high-throughput pseudogene identification in sheep. Therefore, in the present study, to improve the annotation of sheep genome, we present the first genome-wide pseudogene identification for protein-coding genes using a homology-based computational approach.

Materials and Methods The pseudogene content in the sheep genome was estimated using an in-house computational annotation pipeline, named PseudoPipe. The PseudoPipe pipeline predicts pseudogenes in the genome using homology-based method (BLAST and a clustering algorithm). In the present study, repeat-masked sheep genome reference (Ovis_aries.Oar_v3.1), genome annotation gtf file (version 77) and all of the protein coding genes sequences were downloaded from ENSEMBL database. To identify pseudogenes, the sheep genome was searched in a comprehensive and consistent manner. The key steps in the pipeline involved using BLAST to rapidly cross-reference potential ‘‘parent’’ proteins against the intergenic regions of the genome and then processing the resulting ‘‘raw hits’’ such as eliminating redundant ones, clustering together neighbors, and associating and aligning clusters with a unique parent. Then, pseudogenes were classified based on a combination of criteria including homology, intron/exon structure, and existence of stop codons and frameshifts. Finally, we investigated the results manually and false positive results were removed. Also, the gene ontology (GO) of the parental genes that pseudogenes derived from them, have been investigate by DAVID software. Furthermore, different characteristics of the identified new candidate pseudogene were compared with known pseudogenes in the human, mice and cattle species.

Results and Discussion It is vital to identify pseudogenes to better understand genome annotation and disease-related molecular mechanism. Identification of pseudogenes is an ongoing effort, and there are several groups continuously working on identification of pseudogenes. The complexity of the identification of pseudogenes can be addressed by in silico analysis and using a homology-based whole genome identification approach. Here, using a computational method, we identified 4,098 high confidence pseudogenes including 1,102 duplicated and 2,996 processed pseudogenes in sheep genome. The results of the GO analysis showed that identified pseudogenes are significantly enriched in various biological processes, such as mRNA splicing, ribosome structure, binding rRNA, mitochondrial electron transport, translation and etc. Interestingly, a growing body of evidence suggests parental genes of pseudogenes roles are associated with ribosome, rRNA and translational biological processes. Detailed comparison of our results with other species showed that our results are in consistence with previous studies. For example, pseudogene distribution on the sheep chromosomes was in consistence with human and mouse genome. Moreover, it is reported that, duplicated pseudogenes are commonly found on the same chromosome as their parent genes.

Our results showed that about 28% of the identified duplicated pseudogenes were on the same chromosome

1 - Assistant Professor of Animal and Poultry Science Department, College of Aburaihan, University of Tehran, Iran,
2 - PhD Student of Animal Breeding and Genetics, Department of Animal Science, Faculty of Agriculture, University of Tarbiat Modares, Tehran, Iran,
3 - MSc. Student of Animal Breeding and Genetics, Department of Animal and Poultry Science, College of Aburaihan, University of Tehran, Iran.
(* -Corresponding Author Email: mrbakhtiari@ut.ac.ir)

with their parent genes. The results of the study will help to improve the annotation of the sheep genome. The coincidence of the results of this study with previous studies indicates accuracy of the method used in this research.

Conclusion This study, for the first time, has generated the catalog of 4,098 sheep putative pseudogenes. Our findings provide an evidence for pseudogene content in sheep which is a starting point for understanding of their regulatory mechanism. The identification of the novel pseudogenes have greatly improved the genome annotation of sheep. The results of this study will help to better annotation of sheep genome. By using such methods, we can also improve annotation genomes of various organisms.

Keywords: Genome annotation, Homology, Mismatches, PseudoPipe Software.