مجله بیوتکنولوژی کشاورزی
علمی-پژوهشی و ISC

شناسایی واریانت‌های ژنی اسب کاسپین با استفاده از نسل جدید توالی‌یابی ژنوم با کارایی بالا

بابک عارف نژاد[۱]، حمید کهرام[۲]، محمد مرادی شهربابک[۳]، ملک شاکری[۲]، یانگ دونگ[۵]، خیائولی ژانگ[۵]، ون وانگ[۵]، قاسم حسینی سالکده*[٤]

۱-دانشجوی دکتری فیزیولوژی دام -گروه علوم دامی-پردیس کرج-دانشگاه تهران

۲-استادیار گروه علوم دامی، پردیس کرج-دانشگاه تهران

۳-دانشیار گروه علوم دامی، پردیس کرج-دانشگاه تهران

٤-استادیار گروه ژنومیکس پژوهشکده بیوتکنولوژی کشاورزی-کرج

۵- انستیتو حیات وحش کانمینگ-آکادمی علوم چین

تاریخ دریافت: ۱۳۹۲/۱۱/۰۳، تاریخ پذیرش: ۱۳۹۲/۱۲/۰۹

چکیده

تکنیک‌های نوین توالی‌یابی با کارایی بالا به عنوان رهیافت نوینی در شناسایی واریانت‌های ژنتیکی و اطلاعات عملکردی در گونه‌های بسیاری از چمانگان قرار گرفته‌اند. اسب کاسپین با توجه به ویژگی‌های منحصر به فرد ژنتیکی و فنوتیپی خود، یکی از نژادهای مهم اسب در ایران و جهان است. از این‌رو، پژوهش حاضر به شناسایی واریانت‌های ژنتیکی چندشکلی‌های تک نوکلیوتیدی، حذف و اضافه‌های کوتاه و واریانت‌های تعداد در نسخه (CNV) در اسب کاسپین و بررسی نقش آن‌ها در فرآیندها و مسیرهای بیولوژیک ویژه پرداخته است. با استفاده از توالی‌یابی با کارایی بالا، ۱۰۸Gb از DNA ژنومی سه مادیان کاسپین با میانگین عمق ٤۵/۸ توالی‌یابی شدند. این توالی‌یابی با میانگین همپوشانی ۱٤/٤۱ کم و بیش ۷٦/٤ درصد از ژنوم رفرانس اسب را پوشش داد. با استفاده از فیلترینگ سختگیرانه ۱٦٦٦۷۱۷ چندشکلی تک نوکلیوتیدی، ۳۵۸۰۲۰ حذف و اضافه‌های کوتاه و ۳۱۰۹ CNV شناسایی شدند. کلاسترینگ عملکردی واریانت‌های ژنی اسب کاسپین نشان داد که بیشتر این واریانت‌ها در ژن‌های مرتبط با فرآیندهای سیستم عصبی، تنظیم ورارسانی فرسته‌های بیوشیمیایی مرتبط با گوانیدین تری‌فسفات، موفوژنز سلولی، سازمان‌بندی اسکلت سلولی، توسعه رگی، جنبایی سلولی و انتقال غشایی روی داده است. فزون بر این، واریانت‌های ساختاری ژنوم اسب کاسپین مانند اینورژن‌ها و جابه‌جا شدگی‌ها و نیز حذف و اضافه‌های بزرگ ژنومی که در این پژوهش شناسایی شدند، می‌توانند در طراحی نشانگرهای ژنتیکی برای کارهای اصلاح نژادی و نیز بررسی‌های جمعیتی سودمند واقع شوند.

واژه‌های کلیدی: *توالی‌یابی با کارایی بالا، اسب کاسپین، واریانت ژنتیکی، مسیرهای بیولوژیک.*

*نویسنده مسئول: حمید کهرام- قاسم حسینی سالکده تلفن: ۰۲٦۳۲۲٤۸۰۸۲
Email: Hamid.kohram@yahoo.com hsalekdeh@yahoo.com

مقدمه

اسب کاسپین یکی از کهن‌ترین نژادهای اسب در دنیا است که خاستگاه آن به بیش از ۳۰۰۰ سال پیش بر می‌گردد. گفته می‌شود که این نژاد جد اولیه همه اسب‌های خون‌گرم دنیا است و همانندی زیادی با اسب‌های عرب دارد. اسب‌های کاسپین نخستین بار در سال ۱۹٦۹در شمال ایران شناخته و معرفی شدند (Firouz, 1969). لوییس فیروز پس از کشف دوباره این اسب‌ها ویژگی‌های مورفولوژیک آن‌ها را نزدیک به اسب‌های نگاره‌های پرسپولیس شرح داد. این اسب‌ها به غیر از ارتفاع بدن، همانند دیگر اسب‌ها هستند و تفاوت‌های آناتومیک ناچیزی با دیگر اسب‌ها دارند (Firouz, 1971; Firouz, 1972). بنابراین، این نژاد به علت کوتاهی قد به فرنام پونی[1] نیز شناخته شده می‌شود ولی در برابری با دیگر نژادهای پونی در جهان که اسبانی با بدنی چاق و کوتاه و اندام‌های حرکتی قوی و زمخت و بدون تناسب اندام هستند، بی‌درنگ این اسب از این نوع نژاد مجزا شده و به اسب مینیاتوری معروف شده است (Firouz, 1972).

نرخ آبستنی در این اسب‌ها کمتر از ٤٠ درصد است و گفته می‌شود که از مهمترین مشکلات تولیدمثلی آن‌ها نرخ تخمک‌ریزی پایین در آن‌ها است. در نریان‌ها نیز شمار اسپرم و جنبایی اسپرم پایین است که بازدهی پایین تولیدمثلی را در این نژاد نشان می‌دهد (Hatami-Monazah and Pandit, 1979). به علت اندازه کم جمعیت در این اسب‌ها، تنوع ژنتیکی در این توده پایین است که این امر موجب کاهش شایستگی[2] و افزایش همخونی[3] در آن‌ها شده است. از این‌رو، این اسب‌ها در دسته حیوانات در حال انقراض قرار می‌گیرند. بنابراین، سازوکارهای شناسایی گوناگونی ژنومیک (به ویژه در ژن‌های تولیدمثلی و نیز ژن‌های مرتبط با سازگاری) در بهبود شناخت بیشتر در سطح مولکولی و نیز به کارگیری روش‌های اصلاح مولکولی اهمیت ویژه‌ای در این نژاد دارند. (Shahsavarani and Rahimi-Mianji, 2012).

در دسترس بودن توالی‌های ژنومی دریچه‌ای بر تعیین ژنوتیپ با کارایی بالا و در مقیاس گسترده گشوده است. در آغاز، تکنولوژی‌های تعیین ژنوتیپ بر پایه ریزآرایه‌های DNA (که توانایی شناسایی SNPها را در سطح ژنوم دارند) گسترش یافتند. برکسی پوشیده نیست که این روش‌های تعیین ژنوتیپ، بازدهی شناسایی هزاران نشانگر را در یک فرآیند هیبریدسازی DNA ژنومیک با الیگونوکلیوتیدهای قرارگرفته شده روی Gene Chip بهبود بخشیده‌اند (Huang *et al*., 2009; Winzeler *et al*., 1998). هرچند که با این تکنیک، کم و بیش هدف بررسی نشانگرها در مقیاس گسترده محقق شده است، ولی همچنان روش‌های

۱-Pony

۱- Fitness
۲-Inbreeding

بر پایه ریزآرایه DNA دارای محدودیت‌های جدی‌ای هستند. برای نمونه، طراحی، تولید و نیز همه فرآیند شناسایی نشانگرها با این تکنیک دشوار، زمان‌بر و پرهزینه است (Huang *et al*., 2009).

نسل جدید تکنولوژی‌های توالی‌یابی همراه با توالی‌های در دسترس بیشماری از ژنوم‌های گوناگون، دست‌یافت تازه‌ای را برای طراحی دوباره استراتژی‌های تعیین ژنوتیپ، نقشه‌یابی ژنتیکی و نیز آنالیزهای ژنومی ارایه کرده است. تکنیک‌های جدید توالی‌یابی نه تنها کارایی و هم‌پوشانی توالی‌یابی را به گونه چشمگیری افزایش داده‌اند، بلکه امکان توالی‌یابی شمار زیادی از نمونه‌های زیستی را با استراتژی توالی‌یابی Multiplex فراهم کرده‌اند (Craig *et al*., 2008; Cronn *et al*., 2008; Huang *et al*., 2009). روی هم رفته، این تکنیک‌ها در گسترش روش‌های تعیین ژنوتیپ در مقیاس گسترده و بر پایه توالی‌یابی ژنوم پیشرفت‌های روزافزونی داشته‌اند به گونه‌ای که هم‌پوشانی، صحت و دقت در نقشه‌یابی بسیار چشمگیر است و مقایسه‌های ژنوم و نقشه‌های ژنومی در میان ارگانیسم‌ها و جمعیت‌های گوناگون سنجش‌پذیرتر است. از این رو، این تکنیک‌ها فزون بر این‌که آنالیزهای ژنتیک با هدف شناسایی واریانت‌ها در مقیاس گسترده را به گونه‌ای چشمگیر آسان و کارآمدتر می‌کنند، پاسخ‌های دقیق‌تری به پرسش‌های بیولوژیک ارایه می‌کنند (Cronn *et al*., 2008; Huang *et al*., 2009).

در جمعیت‌های گوناگون یوکاریوت‌ها، چندشکلی‌های تک‌نوکلیوتیدی (SNP)[۴] و نیز واریانت‌های CNV[۵] در ژنوم از منابع مهم واریاسیون‌های ژنتیکی و فنوتیپی هستند. فزون بر این، چندشکلی‌های INDEL[۶] نیز کم و بیش فراوانند و در بروز صفات و فنوتیپ‌های گوناگون کنش‌های معنی‌داری دارند. با این وجود، به علت دشواری و بازدهی پایین آن‌ها در پلاتفورم‌های بر پایه تکنیک‌های ریزآرایه، تاکنون کمتر شناسایی و بررسی شده‌اند؛ ولی امکان بررسی همه جانبه و نیز نقشه‌یابی همه این‌گونه از واریانت‌ها با کمک تکنیک‌های جدید توالی‌یابی و تکامل دانش نوظهور بیوانفورماتیک فراهم شده است (Shao *et al*., 2012).

تاکنون ژنوم‌های اسب‌های Quarter (Doan *et al*., 2012)، اسب عرب، Icelandic، Standardbred، Norwegian_Fjord، Thoroughbred، Przewalski (Orlando *et al*., 2013) با استفاده از این تکنیک توالی‌یابی شده‌اند. با وجود اهمیت اسب‌های ایران به ویژه اسب کاسپین در ذخایر ژنتیکی اسب هیچ‌گونه پژوهشی روی ساختار ژنوم و واریانت‌های ژنومی آن‌ها انجام نشده بود. از این رو، پروژه حاضر آغازی بر این‌گونه بررسی‌ها در اسب‌های ایرانی است. هدف اصلی این پژوهش، تعیین ساختار ژنوم و نیز

۱-Single Nucleotide Polymorphism
۲-Copy Number Variant
۳-Insertion and Deletion

شناسایی واریانت‌های رایج در ژنوم اسب کاسپین و شناسایی مسیرهای بیولوژیک مرتبط با آن‌ها با کمک تکنیک‌های نوین توالی‌یابی بود.

مواد و روش‌ها

نمونه‌گیری و استخراج DNA از سلول‌های سفید خون

سه نمونه خون از مادیان‌های کاسپین در موسسه تحقیقات خجیر (پارک ملی خجیر-تهران) گرفته شدند. نمونه‌ها در فلاسک یخ به آزمایشگاه منتقل و بی‌درنگ پس از جداسازی سلول‌های سفید از قرمز خون، DNA با روش استخراج نمکی تغییر یافته استخراج و کیفیت و کمیت آن با کمک اسپکتروفتومتری و الکتروفورز با ژل آگاروز تعیین شد. برای ساخت DNA Library و تعیین توالی به انستیتوی BGI[۷] (BGI, Shenzen, China) در چین فرستاده شدند. توالی‌یابی ژنوم اسب کاسپین با استفاده از پلاتفورم Illumina Hiseq2500 با اندازه ۳۰۰ bp و توالی‌یابی دو سویه انجام شد و داده‌های خام ژنومی که از توالی‌یابی به دست آمدند برای آنالیزهای بیوانفورماتیک پردازش شدند.

پیش‌پردازش توالی‌های کوتاه ژنومی

کنترل کیفی توالی‌های کوتاه ژنومی به دست آمده از توالی‌یابی اسب کاسپین با استفاده از *Fastqc* (Andrews, 2012) انجام شد. پس از آن، توالی‌های کوتاه دو سویه ژنومی با استفاده از *AdapterRemoval v1.2* (Lindgreen, 2012) به سه دسته توالی منفرد، کولاپس شده و دو سویه پردازش شدند. در این فرآیند، همزمان با حذف بخش‌هایی از توالی که کیفیت خوانش پایینی داشتند توالی‌های دوسویه‌ای که دستکم در ۱۱bp با هم همپوشانی داشتند، کولاپس شدند و به عنوان یک توالی منفرد درنظر گرفته شدند. در بخش‌های کولاپس شده توالی‌های کوتاه، Phred Quality Score و نوکلیوتید مربوطه بر اساس بالاترین اسکور نگه‌داشته شدند. در مواردی که هیچ همپوشانی میان دو سوی خوانش توالی‌های کوتاه پیدا نشد، توالی جداگانه پردازش و بخش‌های با کیفیت خوانش پایین، نوکلیوتیدهای N و نیز توالی‌های آداپتوری از آن‌ها حذف شدند. پس از این فرآیند، توالی‌هایی که کمتر از ۲۵ نوکلیوتید طول داشتند حذف شدند.

هم‌ردیفی داده‌های به دست آمده از ژنوم اسب کاسپین

پس از پالایش توالی‌های کوتاه ژنوم اسب کاسپین، فرآیند نقشه‌یابی آن‌ها با ۳۱ کروموزوم اتوزوم و کروموزوم X ژنوم رفرانس اسب (http://genome.ucsc.edu) انجام شد. در این فرآیند، با استفاده از *BWA 0.5.9*[۸] (Langmead,

۱-Beijing Genomics Institute

۱-Burrow Wheeler Aligner

2002) پس از ایندکس سازی ژنوم رفرانس، همه توالی‌های کوتاه با کمک این الگوریتم و تعیین پارامترهای مناسب با ژنوم ایندکس شده، هم‌ردیف شدند. توالی‌های منفرد و کولاپس شده با استفاده از *bwa samse* و توالی‌های کوتاه دو سویه با استفاده از *bwa sampe* هم‌ردیف شدند. پس از آن توالی‌های کوتاه مضاعف شده در PCR با استفاده از *MarkDuplicates* نرم‌افزار *Picard tools* نسخه 1.99 (http://picard.sourceforge.net) حذف شدند. در پایان فایل‌های BAM که برای هر دسته از توالی‌های کوتاه به دست آمدند با استفاده از *MergeSam* با یکدیگر ادغام شدند. تک فایل BAM به دست آمده برای افزایش صحت هم‌ردیفی و نیز شناسایی واریانت‌ها، با استفاده از الگوریتم *GATK*[۹] (McKenna *et al*., 2010) و بر اساس INDELهای شناخته شده اسب دوباره هم‌ردیف شدند.

شناسایی واریانت‌های ژنوم اسب کاسپین

در این فرآیند، پس از کالیبراسیون دوباره کیفیت خوانش توالی‌های کوتاه ژنومی بر اساس کوواریت‌های Read Group، Quality Score، Cycle و Dinucleotideها برای تک فایل BAM به دست آمده از گامه‌های پیشین، واریانت‌های SNP با استفاده از الگوریتم *GATK UnifiedGenotyper* به کمک *Queue* و اسکریپت *Scala* شناسایی شدند. پس از شناسایی واریانت‌ها، SNPها با استفاده از فیلترینگ سخت‌گیرانه *GATK* برای QualByDepth کمتر از ۲، FisherStrand بیشتر از ۶۰، RMSMappingQuality کمتر از ۴۰، HaplotypeScore بیشتر از ۱۳، MappingQualityRankSumTest کمتر از ۱۲/۵ و ReadPosRankSumTest کمتر از ۸ فیلتر شدند. INDEL واریانت‌های INDEL نیز با استفاده از *Freebayes v-0.9.9* با الگوریتم بیزین در شناسایی واریانت‌ها استفاده شد (Garrison and Marth, 2012) و پس از آن برای QualByDepth کمتر از ۲، FisherStrand بیشتر از ۲۰۰ و ReadPosRankSumTest کمتر از ۲۰ فیلتر شدند. واریانت‌های CNV و INDELهای بزرگ نیز با استفاده از الگوریتم *BreakDancer* (Chen *et al*., 2009) و *CNVnator* (Abyzov *et al*., 2011) شناسایی و برای نواحی Gapهای ژنوم، تلومریک و سانترومریک فیلتر شدند.

شناسایی اثر واریانت‌ها بر کنش ژن‌ها و کلاسترینگ عملکردی

برای شناسایی اثر واریانت‌ها بر کنش ژن‌ها و نیز مکان‌یابی واریانت در بخش‌های ساختاری ژن‌ها از نرم‌افزار *snpEff* (Cingolani *et al*., 2012) و *Annovar* (Wang *et al*., 2010) استفاده شد. در این فرآیند، ابتدا واریانت‌های به دست آمده از ژنوم اسب کاسپین با واریانت‌های شناخته شده در دیگر

۱-Genome Analysis Toolkit

نژادهای اسب مقایسه و پس از آن اثر واریانت‌ها بر ژن‌ها بررسی شدند. برای بررسی Gene ontology و کلاسترینگ عملکردی ژن‌های دارای واریانت‌های با اثر بالا بر کنش ژن‌ها، از *DAVID* (Dennis Jr *et al*., 2003) استفاده شد.

نتایج

توالی‌یابی ژنوم و پیش‌پردازش داده‌های ژنومی اسب کاسپین

نتایج توالی‌یابی، نقشه‌یابی توالی‌های کوتاه و آنالیز قطعات ژنومی در جدول ١ نشان داده شده است. در فرآیند توالی‌یابی روی هم رفته، ٧١٨٨٥٩٨١٣ توالی کوتاه با ١١٩ گیگا باز به دست آمد که پس از پالایش آن‌ها بر اساس کیفیت توالی‌های کوتاه به دست آمده از توالی‌یابی مونتاژهای اتوزومی و کروموزوم X ژنوم رفرانس اسب (equCab2) نقشه‌یابی شدند. از ٦٠٦٨١٠٠١٥ توالی کوتاه به دست آمده از فرآیند پالایش کیفیت (١٠٨ گیگا باز)، ٣٦٨٢١٨٧٣٤ توالی کوتاه (٦٥/٩ گیگاباز) با ژنوم اسب هم‌ردیف شدند و میانگین همپوشانی توالی‌یابی ژنوم اسب کاسپین ١٤/٤١ و درصد همپوشانی با ژنوم رفرانس ٦٨/٤ محاسبه شد. شکل١ همپوشانی توالی‌های کوتاه را در ژنوم اسب نشان می‌دهد.

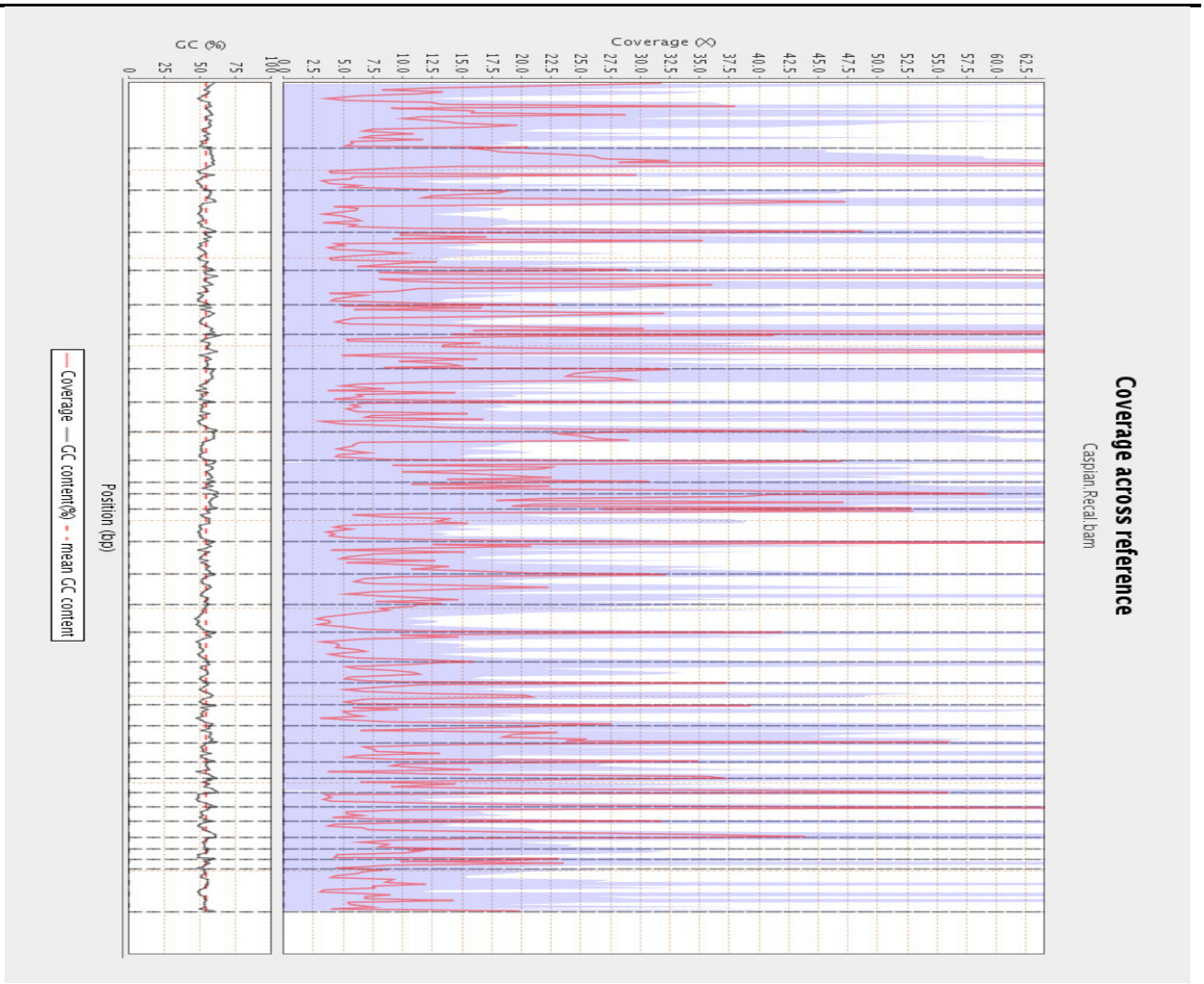
شناسایی واریانت‌های ژنتیکی

پس از هم‌ردیفی توالی‌های کوتاه به دست آمده از توالی‌یابی اسب کاسپین، واریانت‌های SNP، INDEL و CNV آنالیز شدند. SNPها با معیارهای بسیار سخت‌گیرانه الگوریتم *GATK* فیلتر شدند و کمترین میزان همپوشانی ژنومی برای شناسایی SNP، ٥X در نظر گرفته شد. روی هم رفته، ١٦٦٦٧١٧ چند شکلی تک نوکلیوتیدی (SNP) در ژنوم سه مادیان کاسپین در مقایسه با ژنوم رفرانس اسب شناسایی شد.

جدول۱- توالی‌های کوتاه تولید شده و نتایج کلی نقشه‌یابی با ژنوم رفرانس

Table1- Data summary of the Caspian horse genome resequencing and mapping reads to the reference genome.

روش Method	توال‌های کوتاه Raw Reads	اندازه داده‌های خام Raw Data Size (Gbp)	عمق Depth	خوانش‌های فیلتر شده Filtered Reads	اندازه داده‌های فیلتر شده Filtered Data Size (Gbp)	عمق پس از فیلتر Depth	خوانش‌های نقشه‌یابی شده Mapped Reads	میانگین عمق پوشش Average of Depth of Coverage	درصد ژنوم رفرانس نقشه‌یابی شده % of Reference Mapped
100 PE	718,859,813	119	50.2	606,810,015	108	45.8	368,218,734	14.41	76.4



شکل ۱- همپوشانی توالی‌های کوتاه اسب کاسپین در طول ژنوم رفرانس اسب.

Figure 1- Coverage of short reads across horse reference genome.

گوناگونی SNPها و INDELها در جدول ۲ ارایه شده‌اند. نرخ SNP برابر با ۱ چندشکلی در هر ۱٤۱۸ باز بود. شمار Transition برابر ۱۱٥٥٤۱۷ و شمار Transversion برابر ٥۱۲۹۸٦ بود که نسبت Transition/Transversion در اسب کاسپین ۲/۲٥۲۳ برآورد شد. از مقایسه چند شکلی‌های تک نوکلیوتیدی به دست آمده در این پژوهش با چند شکلی‌های موجود در پایگاه اطلاعاتی SNP (dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP) نشان داد که ۱٤٤۸۳٦٤ چند شکلی شناخته شده هستند و ۲۱۸۳٥۳ چند شکلی جدید هستند و تاکنون گزارش نشده‌اند. شمار ۱۱۲۸۲۱٦ چندشکلی تک نوکلیوتیدی در نواحی بین ژنی شناسایی شدند. چند شکلی‌های شناسایی شده در فرادست و فرودست نواحی ژنی تا ٥ کیلوباز فاصله با نواحی ژنی در نظر گرفته شدند. از همه چند شکلی‌های ژنی شناسایی شده در اسب کاسپین ۱۲۹۱۰ چند شکلی نامعنی[۱۰] و ۱۸۳۳۷ چند شکلی هم‌معنی[۱۱] شناسایی شدند. فزون بر این، ۷۳ چند شکلی بی‌معنی[۱۲] شناسایی شد. جایگزینی‌های CCA/CCG با ٥٤٦ مورد و GCT/GCC با ٥۰٦ مورد بیشترین تغییرات کدون آمینواسیدی را در ژن‌های اسب کاسپین به وجود آورده‌اند. بیشترین شمار جایگزینی‌های آمینواسیدی به آلانین/ترئونین (٤۳۸)، آلانین/والین (۳٤۸)، ایزولوسین/والین (۳۰٤) و گلوتامین/آسپاراژین (۳٥٤) اختصاص داشتند. از مجموع چند شکلی‌های شناسایی شده در نواحی ژنی ۳٦۱ چند شکلی با اثر تخریبی بالا، ۱۲۸۸۱ چند شکلی با اثر تخریبی متوسط و ۱۸٥٦۲ چند شکلی با اثر تخریبی کم در کنش ژن شناسایی شدند (جدول ۲).

با در نظر گرفتن فیلترینگ سخت‌گیرانه ۳٥۸۰۲۰ حذف و اضافه‌های (INDEL) کوتاه (کمتر از ۱٥bp) در ژنوم اسب کاسپین شناسایی شدند. روی هم رفته ۲۱۱۸٤۳ اضافه‌های نوکلیوتیدی و ۱٤٦۱۷۷ حذف‌های نوکلیوتیدی شناسایی شدند. بررسی اثر این حذف و اضافه‌های نوکلیوتیدی بر کنش ژن‌ها نشان داد که ٤۷٤۱ حذف و اضافه موجب تغییر الگوی نواحی رمزگردان ژنی می شوند در حالیکه ۳۱٤ حذف و اضافه اثر متوسطی بر کنش ژن هدف دارند. فزون بر این، ۱۳ حذف و اضافه موجب تغییر کدون امینواسیدی به کدون پایانی در نواحی رمزگردان ژن‌های هدف خود می‌شوند (جدول۲).

نتایج آزمون شناسایی CNV پس از تصحیح انحراف GC و فیلترینگ نواحی Gap ژنومی و تلومریک، شمار ۳۱۰۹ CNV را در ژنوم اسب کاسپین نشان داد که از این میان ۹۰۲ کم شدگی (Loss) و ۲۲۰۷ زیاد شدگی (gain) شناسایی شدند. اندازه CNVها میان ۹۰۰bp تا ۲/۸٦ Mbp

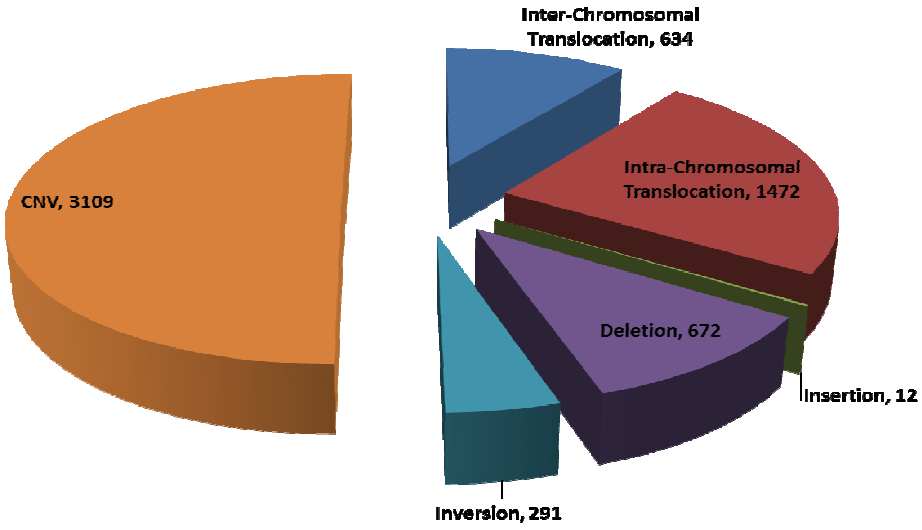
۱-Non-Synonymous
۲-Synonymous
۳-Nonsense

بود. آنالیز ژنی CNVها نشان داد که ۲۳۰۷ CNV در نواحی بین ژنی و ۸۰۲ CNV در نواحی ژنی و بیشتر در نواحی اینترونی قرار دارند. فزون بر این، ٦٣٤ جابه‌جایی بین کروموزومی، ١٤٧٢ جابه‌جایی درون‌کروموزومی، ٦٧٢ حذف‌های بزرگ، ۱۲ اضافه‌های نوکلیوتیدی بزرگ و ۲۹۱ اینورژن در ژنوم اسب کاسپین در مقایسه با ژنوم رفرانس اسب دیده شد (شکل۲).

جدول ۲- Annotation چندشکلی‌های تک نوکلیوتیدی و حذف و اضافه‌های کوچک در اسب کاسپین.

Table 2– Annotations of SNPs and INDELs within Caspian horse genome.

	ژن‌های انسمبل ENS Genes	کل چند شکلی‌های تک نوکلیوتیدی All SNPs	هموزایگوت‌ها Homozygous	هتروزایگوت‌ها Heterozygous	چندشکلی‌های جدید Novel SNPs	هموزایگوت‌ها Homozygous	هتروزایگوت‌ها Heterozygous	کل حذف و اضافه‌ها All INDELs	هموزایگوت‌ها Homozygous	هتروزایگوت‌ها Heterozygous
کل Total	26,295	1,666,717	513,252	1,153,465	218,353	99,508	118,845	358,020	176,691	181,329
نواحی بین ژنی Intergenic	NA	1,128,216	349,371	778,845	146,065	66,723	79,342	247,985	121,446	126,539
فرادست ژنی UPSTREAM	20,027	109,275	35,367	73,908	12,925	5,761	7,164	29,456	18,978	10,478
فرودست ژنی DOWNSTREAM	17,827	119,379	36,673	82,706	12,951	5,685	7,266	22,573	12,426	10,147
ژنی GENIC	15,132	698,861	213,133	485,728	73,101	32,694	40,407	113,069	53,953	53,882
اینترون INTRON	14,301	661,886	201,184	460,702	69,071	31,282	37,789	104,824	51,486	53,338
UTR_5_PRIME	1,097	1,755	792	963	84	39	45	874	804	70
UTR_3_PRIME	1,066	1,623	543	1,080	200	98	102	426	258	168
نواحی اسپلایسینگ Splice region variant	197	259	145	114	18	11	7	1154	1,102	52
اگزون EXON	9,684	32,171	10,042	22,129	3,600	1,202	2,398	5,234	.	.
واریانت‌های غیر رمزگردان Non coding exon variant	409	851	336	515	103	55	48	544	300	244
ایجاد کننده کدون آغازین Start gained	186	225	69	156	17	4	13	0	0	0
ایجاد کننده کدون پایانی Stop gained	64	73	15	58	5	1	4	13	3	10
از دست رفتگی کدون پایانی Stop lost	5	8	5	3	2	2	0	0	0	0
واریانت‌هایی که کدون پایان را تغییر نمی‌دهند Stop retained variant	8	10	2	8	1	0	1	NA	NA	NA
گروه‌های عملکردی Functional Classes										
نامترادف MISSENSE	5,578	12,910	4,525	8,385	1,128	481	647	NA	NA	NA
بی معنی NONSENSE		73	15	58	6	2	4	NA	NA	NA
هم‌معنی SILENT	7,041	18,337	5,166	13,171	1,729	743	986	NA	NA	NA
اثر بر کنش ژن Impact										
زیاد HIGH		361	177	184	28	11	17	4,741	3,814	927
کم LOW		18,562	5,236	13,326	1,748	749	999	NA	NA	NA
متوسط MODERATE		12,881	4,507	8,374	1,123	478	645	314	69	245
تغییر دهنده MODIFIER		2,040,360	629,970	1,410,390	215,935	117,505	98,430	355,032	174,638	180,394



شکل ۲- واریانت‌های ساختاری در اسب کاسپین.

Figure 2– Structural variants in the Caspian horse genome.

آنالیز کلاسترینگ عملکردی و Gene Ontology

آنالیز عملکردی و کلاسترینگ ژن‌های دارای واریانت‌های با اثر مخرب مسیرهای بیولوژیک مرتبط با واریانت‌های ژنتیکی را نشان می‌دهد. هرچند اطلاعات مسیرهای بیولوژیک و Gene Ontology برای ژن‌های اسب هنوز تکمیل نشده‌اند. از این رو، با استفاده از اورتولوگ‌های انسانی مرتبط با ژن‌های اسب، آنالیزهای عملکردی مرتبط با چندشکلی‌های تک‌نوکلیوتیدی بررسی شدند. از میان ۵۵۷۸ ژن دارای چند شکلی نامعنی، شمار ٥٦٦٥ ژن دارای اورتولوگ انسانی بودند و از این میان ۵۰۷۳ ژن دارای رکورد DAVID بودند. آنالیز کلاسترینگ نشان داد که چند شکلی‌های نامعنی در اسب کاسپین با FDR[۱۳] کمتر از ۱ درصد در ژن‌های مسیرهای بیولوژیک مرتبط با فرآیندهای سیستم عصبی، تنظیم ورارسانی سیگنال مرتبط با GTP، مورفوژنز سلولی، سازمان‌بندی اسکلت سلولی، توسعه رگی، جنبایی سلولی، سگنالینگ سلول–سلول، انتقال غشایی، فرآیندهای متابولیک RNAهای غیررمزگردان، تنظیم حرکت سلولی، تنظیم تولید سایتوکاین‌ها، تشخیص محرک‌ها، تنظیم ترشح، کاتابولیسم لیپیدها، تنظیم مثبت پاسخ به محرک‌ها، فرآیندهای هومیوستاتیک و تکامل مغز پیشین نقش دارند (جدول۳). فزون بر این، چند شکلی‌های حذف و اضافه با FDR کمتر از ۱ درصد، بیشتر در مسیرهای بیولوژیک مرتبط با

۱-False Discovery Rate

تنظیم فرآیند رونویسی، متابولیسم فسفر، چسبندگی سلولی، حرکت سلولی، تمایز نرون‌ها، سازمان‌بندی اسکلت سلولی، تکامل جنینی، چرخه سلولی، متابولیسم ماکرومولکول‌ها، تمایز رگی، تمایز اندام جنینی، فرآیندهای مرتبط با فیلامنت‌های حدواسط، اندوسیتوز، تنظیم حرکت و مهاجرت سلول درگیر هستند. فزون بر چندشکلی‌های تک نوکلیوتیدی نامعنی و حذف و اضافه‌های کوچک، آنالیز Gene Ontology برای ژن‌های دارای واریانت‌های با آثار بالا بر کنش و ساختار ژن نشان داد که این ژن‌ها در مسیرهای بیولوژیک مرتبط با پاسخ ایمنی و انتقال یونی نقش دارند (شکل۳).

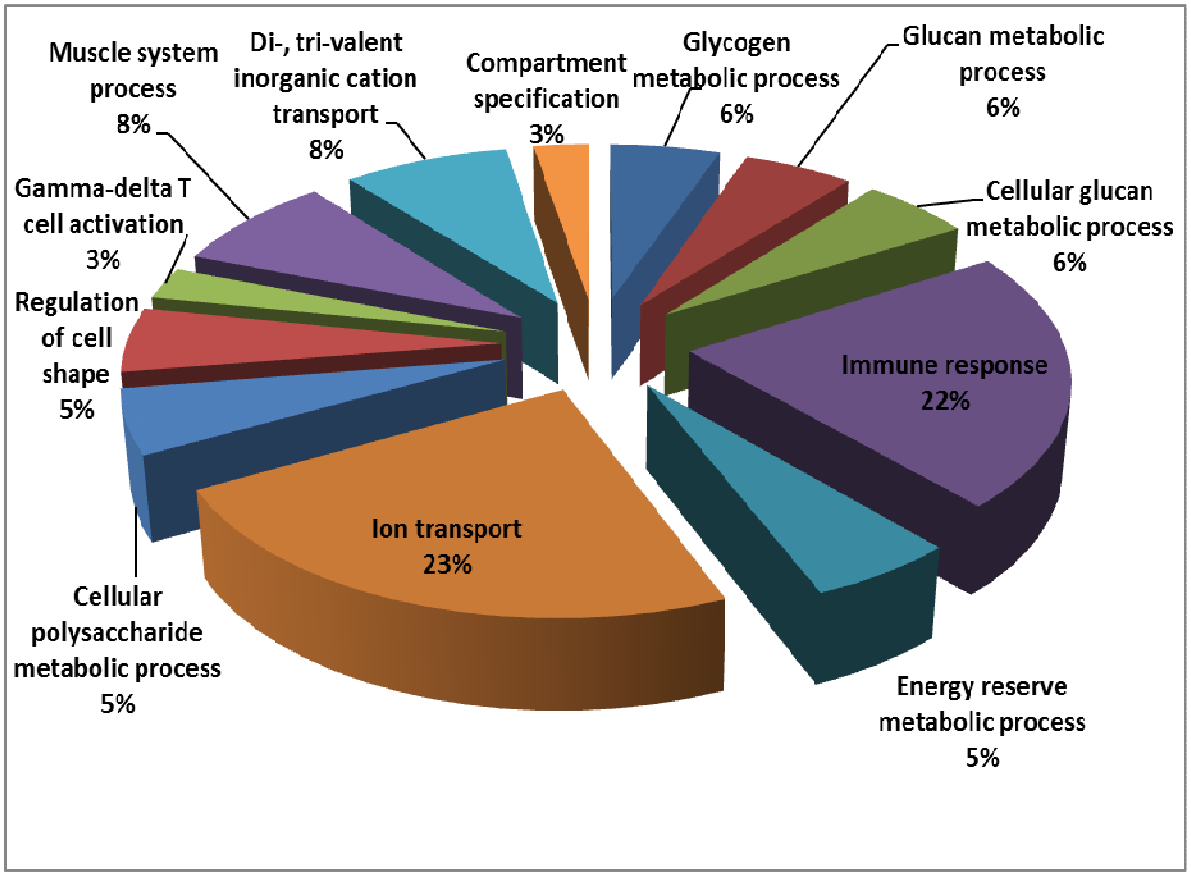
بحث

توالی‌یابی و مونتاژ ژنوم اسب یکی از دستاوردهای مهم است که کاربرد گسترده‌ای در بهبود عملکرد و سلامت حیوان و نیز درک بیشتر تفاوت‌های تکاملی و مولکولی با دیگر پستانداران دارد. تاکنون ژنوم اسب نژادهای Thoroughbred (Wade *et al*., 2009) و Quarter (Doan *et al*., 2013) عرب (Orlando *et al*., 2013)، Icelandic (Orlando *et al*., 2013)، Standardbred (Orlando *et al*., 2013)، Przewalskii (Orlando *et al*., 2013) و Norwegian Fjord (Orlando *et al*., 2013) توالی‌یابی شده و در دسترس‌اند. شمار کل واریاسیون‌های ژنتیکی که تاکنون در اسب شناسایی شده‌اند نزدیک ۳ میلیون SNP است؛ که از میان بیشتر واریاسیون‌های شناخته شده اسب (٦٤ درصد) از نریان Thoroughbred توالی‌یابی و مونتاژ شده به دست آمده است(Doan *et al*., 2012). با وجود اینکه گفته می‌شود اسب کاسپین از کهن‌ترین اسب‌های اهلی دنیاست و جد اسب‌های اورینتال خاورمیانه است ولی پژوهشی برای شناخت ساختار ژنوم و نیز واریانت‌های ژنومیک این اسب‌ها انجام نگرفته بود. پژوهش حاضر نخستین تلاش برای شناسایی ژنومیک اسب کاسپین با استفاده از تکنیک‌های نوین توالی‌یابی با کارایی بالا است که توانسته است نزدیک ۷۲ درصد از ژنوم اسب کاسپین و واریانت‌های آن را رونمایی کند. در ابتدای امر، با وجود عمق توالی‌یابی و شمار زیادی توالی کوتاه که در فرآیند توالی‌یابی اسب کاسپین به دست آمد پیش‌بینی می‌شد که بخش قابل توجهی از ژنوم اسب کاسپین پوشش داده شود ولی پس از نقشه یابی توالی‌های کوتاه با ژنوم رفرانس اسب دیده شد که ۵۵۸۵۰۷۷۸۹ نوکلیوتید از ۲۳٦۷۰٥۳٤٤۷ نوکلیوتید ژنوم هیچ هم‌پوشانی با توالی‌های کوتاه ندارند. از این رو، بررسی بیشتر ساختار ژنوم اسب کاسپین با استفاده از *De novo* Assembly توالی‌های کوتاه نقشه‌یابی نشده با ژنوم رفرانس بایسته می‌نماید.

جدول ۳- آنالیز کلاسترینگ عملکردی ژن‌های دارای واریانت‌های تک نوکلیوتیدی و واریانت‌های حذف و اضافه در اسب کاسپین.

Table 3– Functional Clustring Analysis of the genes containing SNPs and INDEL variants.

	کلاستر Cluster	اسکور Enrichment Score	عبارت فرآیند بیولوژیک Enrichment Term	تعداد Count	درصد %	Ease Source p-Value	FDR
چندشکلی تک نوکلیوتیدی نامترادف Missense SNPs	1	45.1	neurological system process	363	22.00	2.67E-68	5.01E-65
	2	17.88	regulation of small GTPase mediated signal transduction	95	5.76	3.65E-25	6.84E-22
	3	15.33	cell morphogenesis	115	6.97	9.53E-24	1.79E-20
	4	14.89	cytoskeleton organization	145	8.79	2.21E-31	4.14E-28
	5	11.95	vasculature development	78	4.73	3.71E-15	6.87E-12
	6	10.95	cell motion	123	7.45	2.32E-16	4.22E-13
	7	10.13	cell-cell signaling	136	8.24	3.55E-13	6.65E-10
	8	8.76	transmembrane transport	176	10.67	1.28E-33	2.39E-30
	9	7.82	ncRNA metabolic process	73	4.42	9.24E-15	1.73E-11
	10	7.78	steroid metabolic process	53	3.21	8.79E-08	1.65E-04
	11	7.32	defense response	118	7.15	4.11E-07	7.70E-04
	12	5.19	tRNA metabolic process	44	2.67	8.96E-12	1.68E-08
	13	4.81	regulation of cell motion	52	3.15	4.70E-08	8.82E-05
	14	4.66	regulation of cytokine production	41	2.48	1.18E-04	0.221864
	15	4.38	detection of stimulus	43	2.61	3.76E-11	7.06E-08
حذف و اضافه‌های کوتاه INDELs	1	33.23	regulation of transcription	523	32.26	4.23E-41	7.88E-38
	2	12.64	phosphorus metabolic process	195	12.03	9.34E-14	1.74E-10
	3	10.96	cell adhesion	156	9.62	4.91E-15	9.10E-12
	4	8.43	cell motion	109	6.72	1.59E-11	2.96E-08
	5	7.73	neuron differentiation	106	6.54	9.84E-13	1.83E-09
	6	7.51	cytoskeleton organization	105	6.48	1.76E-12	3.29E-09
	7	6.71	chordate embryonic development	76	4.69	2.50E-08	4.66E-05
	8	6.62	positive regulation of macromolecule metabolic process	157	9.69	2.86E-08	5.33E-05
	9	6.45	cell cycle process	125	7.71	6.08E-12	1.13E-08
	10	6.32	pattern specification process	66	4.07	1.18E-08	2.20E-05
	11	6.29	regulation of small GTPase mediated signal transduction	72	4.44	1.88E-12	3.50E-09
	12	6.18	negative regulation of macromolecule metabolic process	143	8.82	2.26E-09	4.21E-06
	13	6.12	macromolecule catabolic process	151	9.32	1.30E-09	2.41E-06
	14	5.3	blood vessel development	60	3.70	8.37E-08	1.56E-04
	15	5.08	embryonic organ development	45	2.78	6.26E-07	0.001167



شکل۳- آنالیز Gene Ontology برای ژن‌های دارای واریانت‌های تک‌نوکلیوتیدی با اثر بالا بر کنش ژن.

Figure3– Gene ontology analysis of the genes containing single nucleotide polymorphisms with high impact on function of the genes.

هرچند در پژوهش‌های مستقل (Doan *et al.*, 2012) و (Orlando *et al.*, 2013) شمار چندشکلی‌های شناسایی شده در هر نژاد از اسب‌ها بیش از ۳ میلیون بوده است ولی در این پژوهش شمار چندشکلی‌های شناسایی شده در ژنوم اسب کاسپین روی هم رفته ۱۶۶۶۷۱۷ بود. مهمترین علت این امر استفاده از الگوریتم‌های با صحت بالاتر و فیلترینگ بسیار شدیدتر روی شناسایی واریانت‌ها بوده است. علت دیگری که احتمالا در بروز تفاوت در شمار واریانت‌های شناسایی شده در اسب کاسپین با دیگر ژنوم‌های اسب در پژوهش‌های پیشین موثر بوده، ساختار منحصر به فرد ژنوم اسب کاسپین است. بی‌شک، بررسی‌های بیشتر داده‌های این پژوهش با الگوریتم‌های فایلوژنتیک ژنومیک پرده از روابط فایلوژنتیک این اسب با دیگر نژادهای اسب ارایه خواهد کرد.

نتایج کلاسترینگ عملکردی غنی بودن واریانت‌ها در ژن‌های مربوط به توسعه نورولوژیک و نیز ادراک حسی در اسب کاسپین را به روشنی نشان داد. شاید بتوان خوی آرام و هوش این اسبان

را به تنوع ژنتیکی در این دسته از ژن‌ها ارتباط داد؛ هرچند که این بیان بیشتر پایه حدس و گمان دارد و برای روشن شدن نقش دقیق‌تر واریانت این ژن‌ها در فنوتیپ ذکر شده به بررسی‌های ژنتیک جمعیتی با استفاده از ریزآرایه‌های DNA در جمعیت اسبان کاسپین در آینده نیاز است.

از دیگر دست‌آوردهای این پژوهش، شناسایی واریانت‌های ساختاری ژنوم اسب کاسپین است که با توجه با اهمیتی که این واریانت‌ها در طراحی نشانگرهای ژنومیک در بررسی‌های ژنتیک و کارهای اصلاحی در این نژاد کهن اسب ایرانی دارد، امید است که پژوهش‌های آینده به این نژاد در حال انقراض توجهی ویژه (با رویکرد ژنتیکی نوین) پیدا کنند.

سپاسگزاری

این پژوهش با حمایت مالی آکادمی علوم چین انجام شده است. فزون بر این، آنالیزهای ژنومیک و دسترسی به منابع موردنیاز در اجرای این پژوهش با همکاری پژوهشکده رویان صورت پذیرفته است. از این رو، مجریان و همکاران مراتب سپاس و قدردانی خود را از این دو پژوهشگاه اعلام می‌دارند.

منابع

Abyzov A, Urban AE, Snyder M, Gerstein M (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research 21: 974-984.

Andrews S (2012). FASTQC. A quality control tool for high throughput sequence data. URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods 6: 677-681.

Cingolani P, Platts A, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6: 80-92.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA (2008). Identification of genetic variants using bar-coded multiplexed sequencing. Nature Methods 5: 887-893.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T (2008). Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research 36: e122-e122.

Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003). DAVID: database for annotation, visualization, and integrated discovery. Genome Biology 4: P3.

Doan R, Cohen ND, Sawyer J, Ghaffari N, Johnson CD, Dindot SV (2012). Whole-Genome sequencing and genetic variant analysis of a quarter Horse mare. BMC Genomics 13: 78.

Firouz L (1969). Conservation of a domestic breed. Biological Conservation 2: 53-54.

Firouz L (1971). Osteological and historical implication of the Caspian pony to early domestication in Iran. Proc 3rd Int Congr Agricultural Museum, Budapest: 1-5.

Firouz L (1972). The Caspian miniature horse of Iran. Field Research Projects, Florida, USA,

Garrison E, Marth G (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907.

Hatami-Monazah H, Pandit RV (1979). A cytogenetic study of the Caspian pony. Journal of Reproduction and Fertility 57: 331-333.

Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T (2009). High-throughput genotyping by whole-genome resequencing. Genome Research 19: 1068-1076.

Langmead B (2002). Aligning Short Sequencing Reads with Bowtie. Current Protocols in Bioinformatics: John Wiley & Sons, Inc.

Lindgreen S (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Research Notes 5: 337.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297-1303.

Orlando L, Ginolhac Al, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499: 74-81

Shahsavarani H, Rahimi-Mianji G (2012). Analysis of genetic diversity and estimation of inbreeding coefficient within Caspian horse population using microsatellite markers. African Journal of Biotechnology 9: 293-299

Shao H, Bellos E, Yin H, Liu X, Zou J, Li Y, Wang J, Coin LJM (2012). A population model for genotyping indels from next-generation sequence data. Nucleic acids research 41: e46-e46.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, BlÃ¶cker H, Distl O, Edgar RC, Garber M, Leeb T, Mauceli E, MacLeod JN, Penedo MCT, Raison JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della Valle G, Fryc S, GuÃ©rin G, Hasegawa T, Hill EW, Jurka J, Kiialainen A, Lindgren G, Liu J, Magnani E, Mickelson JR, Murray J, Nergadze SG, Onofrio R, Pedroni S, Piras MF, Raudsepp T, Rocchi M, RÃ¸ed KH, Ryder OA, Searle S, Skow L, Swinburne JE, SyvÃ¤nen AC, Tozaki T, Valberg SJ, Vaudin M, White JR, Zody MC, Broad Institute Genome Sequencing P, Broad Institute Whole Genome Assembly T, Lander ES, Lindblad-Toh K (2009). Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. Science 326: 865-867.

Wang K, Li M, Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research 38: e164-e164.

Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ (1998). Direct allelic variation scanning of the yeast genome. Science 281: 1194-1197.

Genic Variant Detection of Caspian Horse Using High-throughput Sequencing Technology

Arefnezhad B.[1], Kohram H.[1], Moradi Shahre-Babak M.[1], Shakeri M.[1], Dong Y.[5], Zhang X.[5], Wang W.[5], Hoseini Salekdeh Gh.[*4]

[1]Department of Animal Science, University of Tehran, Karaj, Iran.
[2] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China.
[3]Agricultural Biotechnology Research Institute of Iran (ABRII), Karaj, Iran.

Abstract

Recently, new advanced high-throughput sequencing technology as a novel tool has opened the way to study of genomic variants and functional information stored within farm animals. The Caspian horse is one of the valuable horses ever exist in the world. Hence, propose of this study was to investigate genetic variants of single nucleotide polymorphisms, insertion and deletions and copy number variations within the genome of Caspian horse and their involved biological pathways. Using high-throughput sequencing technology, we generated 108 Gb (Average depth of 45.8) of DNA sequence from three Caspian horse mares resulting in an average of 14.41X coverage and 76.4% covered with reference genome. Using a stringent filtering method, we identified 1666717 single nucleotide polymorphisms, 358020 insertion and deletions, and 3109 copy number variations. Functional clustering analysis of genic variants revealed that most of the genetic variants in the Caspian horse's genome were enriched in nervous system, GTP-related signal transduction, cellular morphogenesis, cytoskeleton organization, vascular development and cellular movement. Moreover, we have detected structural variations as like as inversion, intra- and inter-chromosomal translocations, large insertion and deletions which could be useful for marker based population genetic investigation.

Keywords, *High-throughput sequencing, Caspian horse, Genic variants, Biological pathways.*

[*]Corresponding Author: Kohram H, Hoseini Salekdeh Gh. Tel: 02632248082 Email: hamid.kohram@yahoo.com, hsalekdeh@yahoo.com