p-ISSN 2615-787X e-ISSN 2615-790X
Accredited by Directorate General of Strengthening for Research and Development No: 30/E/KPT/2018

Tropical Animal Science Journal, March 2020, 43(1):25-34
DOI: https://doi.org/10.5398/tasj.2020.43.1.25
Available online at http://journal.ipb.ac.id/index.php/tasj

Application of Efficient Express Sequence Tags Information for Classification and Functional Study of Simple Sequence Repeats in Cattle Testis Tissue

M. Manavipour, A. Ehsani*, & A. A. Masoudi
Department of Animal Science, Faculty of Agriculture, Tarbiat Modares University, Tehran, P. O. Box 14115 - 336, Iran
*Corresponding author: alireza.ehsani@modares.ac.ir
(Received 14-09-2019; Revised 27-10-2019; Accepted 18-11-2019)

ABSTRACT

Genomic markers play an important role in tracing the flow of genetic causality of observable signals in animals and plants. In farm animals, the participation of male animals in the gene pool of subsequent generations are much higher than female animals and testes are the most important organs of the male reproductive system. This study was conducted to investigate simple sequence repeats (SSR) within the expressed sequence tags (ESTs) in order to classify the *Bos taurus* testis tissue's genes for their relationship and specificity with related reproductive domains. A total of 48,549 publicly available EST sequences from cattle testis tissue downloaded from GenBank database, out of which, 10,237 sequences that their library made from testis tissue were extracted and specialized as the studied sequences using several searching tools and software. Across these selective sequences, 2,039 contigs, 5,097 singletons, and 153 SSRs were detected. EST-SSRs were subsequently evaluated using GenBank and categorized based on their functions in biological systems of dairy cattle. Investigation of these motifs showed that the identified EST-SSRs can be classified into 48 types that GT in dinucleotides and GCC in trinucleotides had the highest frequency. Annotation and gene ontology analysis revealed a relationship among 54 domains with the observed SSRs. Localization and characterization of such markers can help tracing the production of amino acids coded by identified repeats as shown in this study.

Keywords: expressed sequence tags; SSR mining; conserved regions; cattle; testis tissue

INTRODUCTION

Biological and structural markers play important roles in selection processes of breeding animals and plants. These markers can speed up the breeding process and shorten the period to achieve breeding goals (Wang *et al.,* 2014). An enormous amount of genomic and gene expression data are additionally giving chance to create a new generation of molecular markers by using the novel accessible sequences (Ellis & Burke 2007). However, identification and classification of markers in terms of their biological functions, is still a challenge in genomic studies era (Qian *et al.,* 2018).

Molecular markers based on polymerase chain reaction (PCR) such as Simple Sequence Repeats (SSRs) are one of the most common markers in genomic analysis. SSRs are the tandem-repeated (about 5-50 in most cases) sequences of one to six (or more) base pairs and have random distribution within a genome. These sequences are dispersed in prokaryotic and eukaryotic genomes and can be observed in both the coding and non-coding sequences (Riar *et al.,* 2011). Investigating the simple sequence repeats (SSR) within the expressed sequence tags (ESTs) have some intrinsic benefits. EST-SSRs can be rapidly achieved by electronic sorting and are highly transferable to related taxa. Moreover, they have advantages over molecular markers including higher frequency, capable of being highly reproducible, uniformly dispersed across genome, high rates of interspecific transferability across all species/genera, and are multi allelic (Gupta & Varshney, 2000).

The SSR markers have been applied in marker-assisted selection (Kaur *et al.,* 2015), molecular mapping (Kirungu *et al.,* 2018), assessment of genetic relationships (Huson *et al.,* 2015), finding of the polymorphisms across species (Yan *et al.,* 2017), relating the phenotypes to genotypes (Kalyana Babu *et al.,* 2014), and finally as an efficient tool to link between evolutionary and population genetic studies. EST sequences are short sub-sequences of cDNA. Usually, the expressed sequence tags (ESTs) containing SSRs located in the coding regions of genome are major concern in genetic studies because of their involvement in coding of amino acids and the functions of organs or tissues (Varshney *et al.,* 2002). They may be applied in physical mapping techniques (Bhattacharjee *et al.,* 2018), determination of gene expression (Ma *et al.,* 2012), and sequence comparisons between normal and cancer tissues (Pu *et al.,* 2013). The use of EST sequences has advantages such as the rapid identification of expressed genes, identification of gene

families, phylogenetic analyses, survey of developmentally regulated genes, and examination of strain diversity (Li *et al.*, 2003).

Luckily, Public databases have brought ESTs accessible as DNA-markers for practical application. Such markers can be more helpful than SSRs from unexpressed chromosome regions (Duan *et al.*, 2013). Thus, they may deliver information to associate the complex traits phenotypes with their genetic references. The huge amount of other genomic markers such as single nucleotide polymorphisms (SNPs) cause problems in model fitting because of over-parameterization regression models in association studies (Ehsani *et al.*, 2016). The relatively small number of highly informative EST-SSRs can help preventing such backwards.

The amount and the pattern of expression of different genes and transcripts are not homogeneous across all organs and tissues. Studies showed that tissue-specific expression is a common phenomenon in live organisms (Stamatoyannopoulos 2004; Ehsani *et al.*, 2016). In other words, any given part of coding regions including ESTs may be differently expressed in different organs and tissues. This may help understanding the effects of any given gene on the performance of such organ or tissue (Janatova & Pohlreich, 2004).

Fertility traits have major concerns in animal breeding and improvement of fertility rate, using genetic tools is an important goal in this era (Muller *et al.*, 2017). Typically, the male animals have a higher genetic contribution to the next generations in the industrial mating systems due to artificial insemination and the testes are the most important organs of male animals (Garcia-Ruiz *et al.*, 2016). Studies shown that many expressed genes in the testis can modulate the fertility, survival, metabolic processes, and immune system (Djureinovic *et al.*, 2014).

With the passage of time, approaches for various branches of biological science began to change, but it is important how we use it in contrast to the traditional methods. Basically, the functional genomic information adapts quickly to changing conditions, including the study of molecular markers. While there are number of studies indicating that EST-SSRs could substantially be a reliable source of classification and functional studies for either plants or animals (Taheri *et al.*, 2019; Bakhtiarizadeh *et al.*, 2012b), this study showed that it can also be considered an efficient way of examining a particular tissue such as testis. In this study, the analysis of the EST-SSRs from cattle testis tissue was conducted to find the relationship among such sequences with functional domains. We tried to understand their frequency and distributions as well as to categorize them based on their types and structure to help the use of such biological markers to identify the genes and biological processes related to testis functionality. The results of this study can promote EST-SSR-based detection tool for different organs which are associated with reproductive system in the future researches, and will be useful resources for molecular breeding, genetics, and genomics. Moreover, the conservation of domains which have been found in cattle testis tissue would be truly a new resource to identify useful alleles in transcription factors, regulation of gene expression, spermatogenesis, innate immune response and the other important factors.

MATERIALS AND METHODS

Retrieving EST libraries

First of all, 48,549 EST sequences of cattle's testis tissue were downloaded from the EST database of NCBI website (http://www.ncbi.nlm.nih.gov/dbeST). To focus on the sequences that their library made from testis tissue only, we subtracted the sequences that their tissue from a pooled of several tissues including testis. This is done by looking at the "tissue type" in the FASTA format downloaded sequences into a Notepad spreadsheet. From a given accession number, the tissue type changed from testis to a "pooled" of many tissues. We removed these sequences and called the remaining sequences (10,237) as testis specific expressed sequences. After cleaning the redundant parts of vectors attached to sequences, removing short length sequences (<150) and poly A (T) tails using EST-clean software (Tae *et al.*, 2012), 2,039 contigs and 5,097 singletons were extracted using Vector NTI software (Lu & Moriyama, 2004).

Microsatellite Identification

The contigs and singletons collections were loaded into the Perl script MISA (Thiel *et al.*, 2003). The SSRs containing motifs ranging from 2 to 6 nucleotides in length were selected. The minimum repeat for motifs set to be 6 repeats for dinucleotides, 5 repeats for trinucleotides, and 4 or more repeats for tetra-, penta-, and hexa-nucleotides. The collected contigs and singletons based on the above-mentioned criteria were used for further gene ontology and functional analysis.

In order to find the functional EST-SSRs, the collected contigs and singletons submitted into GenBank non-redundant database using BlaSTX (http://blast.ncbi.nlm.nih.gov/Blast.cgi) at an E value of 1.0*10E-10 for maximum similarity. Classification of selected sequences was based on their molecular function, biological process, and cellular component by searching their names or abbreviation in the UniProt database (http://www.uniprot.org/). The chromosome regions of EST-SSRs were finally mapped using Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/) and the overall view of mapped genome for observed EST-SSRs was visualized using MapChart version 2.3 (Voorrips 2002). Furthermore, the type and the frequency of amino acids coded by the resulted functional EST-SSRs were predicted using DnaSP software (http://www.ub.edu/dnasp/).

RESULTS

Visual Classifications

Analysis of testis specific EST-SSRs using MISA (Thiel *et al.*, 2003) revealed 153 SSRs within EST sequences out of which 30.51% (n=43) were contigs and 69.49 percent (n=110) were singletons (Table 1). The

length of SSRs was ranging from 12 to 246 base pairs. Surprisingly, some of the EST sequences contained more than one microsatellites becoming imperfect microsatellites (Mudunuri & Nagarajaram, 2007). The observed SSRs were 94.11% perfect and the rest were imperfect EST-SSRs.

Di-nucleotides, tri-nucleotides, tetra-nucleotides, and hexa-nucleotides were included 40.76%, 54.14%, 4.46%, and 0.64% of motifs, respectively (Figure 1). As shown in Figure 1, there were no penta-nucleotides among the motifs. Moreover, the number of repeats ranged from 5 to 41, and trimers of 5 repeats had the highest frequency, followed by dimers of 6 and 7 repeats (Figure 2).

The SSR motifs on the basis of length were classified into two groups, class I included the repeats ranging from 10 to 20 nucleotides and class II included the repeats that have more than 20 nucleotides. The percentages for total SSRs for class I and class II were 76.88 and 23.12, respectively (Figure 3).

The results were included 48 types of various motifs in different frequencies (Table 2). The highest frequencies belonged to GT with 6.32%, GCC with 5.69%, and TG with 5.06%.

Functional Classifications

Further analysis using BLASTX showed that from 153 SSRs, 54 of them were belonged to domains that have biological functions. The classified domains and their related motifs with regard to their major role of genes in cattle testis tissue was represented in Table 3. The resulted domains were mostly categorized into spermatogenesis, energy activity, regulation of transcription and translation. Many ESTs were found to be in several categories.

Annotation and gene ontology (GO) analysis for molecular function showed that the most EST-SSR sequences were involved in protein and nucleic acid binding (Figure 4A). The translation and transcription processes had the main role in biological processes compared to the other roles (Figure 4B). The GO assignments for cellular components showed that about one third (32.14%) of SSR-containing ESTs were related to nucleus, ~16% related to membranous, ~10% related to organelles, and the rest that were ~41% were related to cytoplasm (Figure 4C).

Of the 153 sequences containing SSRs, 140 of them were attached to *Bos taurus* chromosomes using map-viewer software (http://www.ncbi.nlm.nih.gov/mapview/). Generally, the distributions of SSRs loci were indicated that the majority of markers were located on the long chromosomes. Chromosome 7 had the highest frequency of the linkage between markers and genes, unlike chromosomes 26 and 28 that had no SSR linked to any gene (Figure 5). Sequences containing motif having special function have been analyzed to predict the amino acid sequences, and the results are shown in Figure 6. The most abundant codon was CUG followed

Table 1. The list of ESTs and EST-SSRs distribution

Parameter	Number
Total number of sequences	48549
Total number of selected sequences	10237
Total number of contigs	2039
Total number of singletons	5097
Total SSR-ESTs identified	153
Total number of contigs containing ESTs	43
Total number of singletons containing ESTs	110

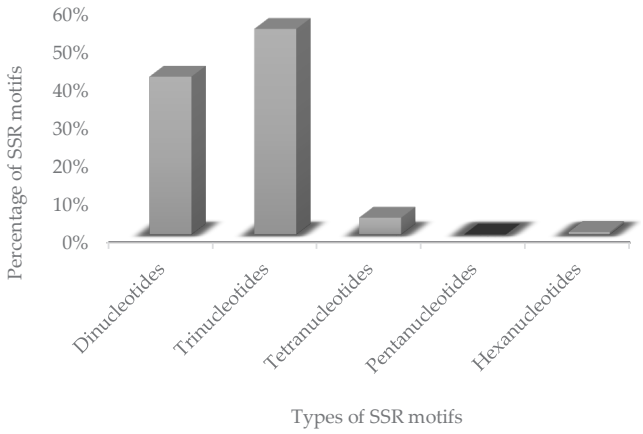

Figure 1. The distribution of dinucleotides, trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides motif type

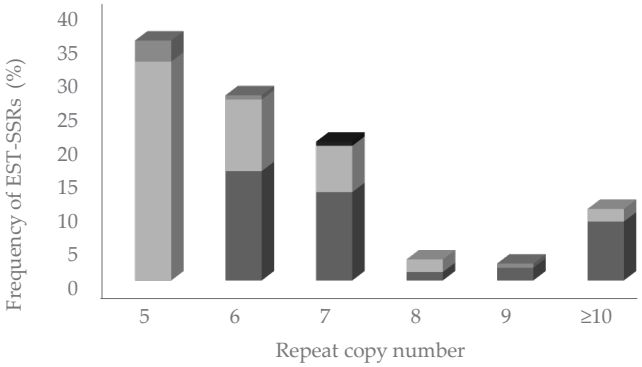

Figure 2. The frequency distribution pattern of EST-derived SSRs based on the repeat copy number for the different SSR motif types

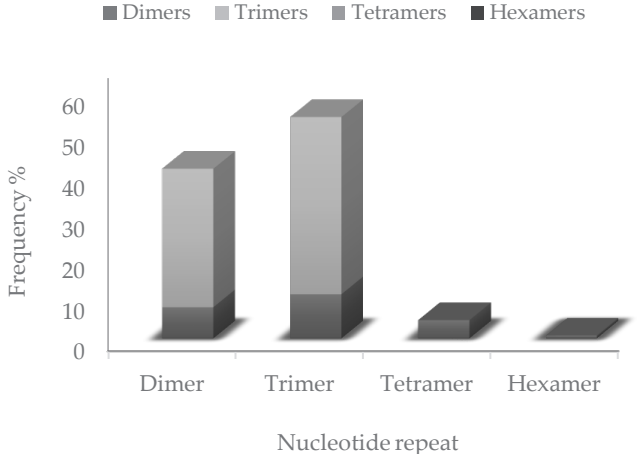

Figure 3. Frequency distribution of Class I and Class II in the total of SSRs based on the length; Class I Class II

Table 2. Various motifs in different frequencies

Motif type	Frequency (%)	Motif type	Frequency (%)	Motif type	Frequency (%)	Motif type	Frequency (%)
AG	3.79	AAG	1.89	CGG	2.53	GCT	3.16
AC	3.79	AAT	0.63	AGA	0.63	CTG	2.53
AT	5.06	GGA	0.63	GAT	0.63	CAG	1.26
GA	3.16	GGC	5.06	GAG	4.43	GAC	0.63
GC	1.89	GGT	0.63	GCG	3.16	CAAA	0.63
GT	6.32	CCA	1.89	CGC	1.89	CCGC	0.63
CA	4.43	CCG	1.89	AGC	3.16	CCTG	0.63
CG	1.26	TTA	0.63	TAT	0.63	GGCA	0.63
CT	2.53	AGG	3.16	CTC	1.26	GAGC	0.63
TA	3.16	ACC	1.26	ATG	0.63	CAGA	0.63
TG	5.06	GCC	5.69	GCA	0.63	GTGC	0.63
TC	1.26	TCC	0.63	TGC	1.89	GAGCCA	0.63

Table 3. Motif type and putative functions of EST-SSRs motifs

Type of gene	Type of motif	Molecular function	Biological process	Cellular component
Pre-B-cell leukemia transcription factor 4	(AT)7	Sequence-specific DNA binding.	Positive regulation of transcription, DNA-templated	Nucleus
Cwf Cwc 15	(GA)6	Unknown	Unknown	Unknown
TUBA1A	(CGG)6	GTPase activity, GTP binding, Structural molecule activity	Metabolic process; Cell division;	Nucleus, Cytoplasmic ribonucleoprotein
Protein C19orf66 homolog	(AGG)5(GAG)7	Uncharacterized	Uncharacterized	Uncharacterized
General transcription factor IIH subunit 5	(AGC)6	DNA binding	Cellular response to Gamma radiation, Nucleotide-excision repair, Regulation of transcription	Nucleolus
Ubiquitin-like protein fubi and ribosomal protein S30 (Fubi)	(AGA)6	Structural constituent of ribosome	translation	Ribosome
Bcl-2 like protein of testis	(CTG)5 (GC)7	Protein binding,	Apoptotic process, Negative regulation of cell proliferation, Regulation of growth rate	Nucleus
Sperm-associated antigen 7 (SPAG7)	(GT)13	Nucleic acid binding	Spermatogenesis	Nucleus
RNA-binding protein 4 (RBM4)	(GCA)5	miRNA binding, RNA binding, Zinc ion binding	Cell differentiation, Negative regulation of translational, RNA processing	Cytoplasm, Nucleolus,
Ferritin heavy chain (FTH1)	(GGT)5 (CCA)5	Iron ion binding	Cellular iron ion Immune response, Iron ion transport	Nucleus, Cytosol,
Protein disulfide-isomerase	(GAC)5	Protein disulfide Isomerase activity, Protein disulfide oxidoreductase activity	Protein folding. Oxidation-reduction process	Endoplasmic reticulum lumen
DNA-directed RNA polymerases I and III subunit RPAC2	(CCTG)5	DNA binding, DNA-directed RNA polymerase activity	Innate immune response, Regulation of gene expression	Cytosol,
MIF4G	(GGA)5	protein C-terminus binding, RNA binding,	Regulation of translation	Cytoplasm, Nucleolus
Motif of Ribosomal Protein L14	(GCT)5	Poly(A) RNA binding, structural constituent of ribosome	ribosomal large subunit biogenesis, rRNA processing, translation	Cytoplasm, membrane
Proteasome subunit alpha type-1	(ATG)5	RNA binding, endopeptidase activity	immune system process	Nucleus, Cytoplasm
Surfeit locus protein 4 (SURF4)	(GAGC)9	Preservation the structure of Golgi	positive regulation of organelle organization	Golgi membrane
SH2 domain-containing adapter protein E (SH2_SHE)	(TA)16(TA)6	SH3/SH2 adaptor activity	cell differentiation, cell proliferation, signal transduction	Cytoplasm
Alpha tubulin	(CGG)6	GTP binding, structural constituent of cytoskeleton	microtubule-based process	microtubule
RWD domain-containing protein 1	(GAT)5	Unknown	unknown	unknown

Type of gene	Type of motif	Molecular function	Biological process	Cellular component
C1 inhibitor (C1-Inh)	(CTG)6	serine-type endopeptidase inhibitor activity; protein binding	lectin pathway; degranulation; innate immune response	extracellular region
Cell adhesion molecule 1 (CADM1)	(GCG)11	PDZ domain binding, protein homodimerization activity	apoptotic process, cell-cell junction organization, Spermatogenesis	integral component of membrane, extracellular exosome
Bax inhibitor 1	(AAG)5	enzyme binding, endoribonuclease inhibitor activity, ubiquitin protein ligase binding	negative regulation of protein binding, negative regulation of apoptotic process, response to L-glutamate	Nucleus, cytoplasm
Chromo box homolog 3(CBX3)	(TA)7	enzyme binding, histone methyltransferase binding	Chromatin remodeling, negative regulation of transcription	Nucleus
Growth factor receptor-bound protein 2 (GRB2)	(CGG)6	SH3/SH2 adaptor activity, SH3 domain binding	blood coagulation, T cell costimulation	Nucleus, Cytoplasm
F-box only protein 3 (FBXO3)	(GAG)5	ubiquitin-protein transferase activity	Proteolysis	Cytoplasm, nucleoplasm
PDZ_signaling	(CCA)5	protein C-terminus binding	cell adhesion, myelination,	cytoplasm
Translation initiation factor eIF-2B subunit beta (eIF-5 eIF-2B)	(AAG)5	ATP binding, GTP binding, translation initiation factor activity	Translation, response to glucose	cytoplasm
Phosphatidylinositol phosphate kinase	(GAG)6	Enzyme activity, Producing collection of lipid messenger	cell migration	Nucleus, Cytoplasm
Protein phosphatase inhibitor 2 (IPP-2)	(CTG)5	protein serine/threonine phosphatase inhibitor	generation of precursor metabolites and energy,	unknown
Beta glucuronidase	(AAT)5	beta-glucuronidase activity, protein domain specific binding, receptor binding	carbohydrate metabolic process	Membrane, extracellular exosome
Protease associated M28 subfamily 2 (PA M28 sub2)	(GCC)7	dipeptidase activity	proteolysis	plasma membrane, vacuole
FAM174B	(CA)10	Unknown	unknown	unknown
Insulin-like growth factor binding protein (IGFBP)	(GCT)7	inhibit or stimulate the growth cells	cellular protein metabolic process; tissue regeneration; positive regulation of cell growth	extracellular region
Protein SGT1	(GGC)11	protein binding, bridging	regulation of cell cycle, protein complex assembly	ubiquitin ligase complex
Vacuolar protein sorting55 (Vps55)	(GCC)5	transporter activity	protein transport	Vps55/Vps68 complex, integral component of membrane
PGP	(GGC)5	magnesium ion binding, nucleotide phosphatase activity	carbohydrate metabolic process, dephosphorylation	Cytosol
Methionine-R-sulfoxide reductase	(CGG)7	zinc ion binding, methionine-R-sulfoxide reductase activity	protein repair, response to oxidative stress, innate immune response	Nucleus, Cytoplasm
Elongation factor like (Elf1)	(TG)6	metal ion binding, RNA polymerase II core binding	chromatin-mediated maintenance of transcription, transcription elongation from RNA polymerase II promoter	Nucleus
Augurin	(GCT)5	transmembrane proteins	cellular senescence	extracellular space
Actin-2	(GGC)5	structural constituent of cytoskeleton	response to cytokinin	Cytoplasm, nucleus
Eukaryotic initiation factor 1	(CCG)6	translation factor activity, RNA binding	response to stress, regulation of translational initiation	Nucleus, cytoplasm
CD47	(GC)7	thrombospondin receptor activity	cell adhesion, , positive regulation of cell proliferation	extracellular exosome
Sterile alpha motif of Polycomb (SAM)	(AGG)5	DNA binding, zinc ion binding	Spermatogenesis, cellular protein metabolic process, Spermatogenesis	Nucleus
Sodium/potassium-transporting ATPase subunit beta-3	(CCG)5	sodium:potassium-exchanging ATPase activity	metal ion transport	plasma membrane
78 kDa glucose-regulated protein	(TGC)5	ATPase activity, enzyme binding	blood coagulation, cellular protein metabolic process, cellular response to antibiotic	Mitochondrion, nucleus, endoplasmic reticulum
Chaperone protein DnaK	(TGC)5	ATP binding, zinc ion binding	DNA replication, response to heat	Cytoplasm, Membrane

Type of gene	Type of motif	Molecular function	Biological process	Cellular component
N-acylethanolamine-hydrolyzing acid amidase	(AGC)7	transcription factor binding	lipid metabolic process	Cytoplasm
Ras-related protein Rap-1b	(GCG)5	GTP binding,	small GTPase mediated signal transduction	cytosol
TMEM52	(AC)11	Unknown	unknown	unknown
Tektin-2	(GCC)5	ssembly or attachment of the inner dynein arm to microtubules in sperm	sperm motility, inner dynein arm assembly	Cytoplasm, nucleus
Vesicular integral-membrane protein VIP36	(GGC)5	metal ion binding, heat shock protein binding, carbohydrate binding	protein transport, retrograde vesicle-mediated transport, Golgi to ER	cell surface, Golgi membrane
Glypican-3	(CCG)5	peptidyl-dipeptidase inhibitor activity	small molecule metabolic process	Golgi lumen, extracellular exosome

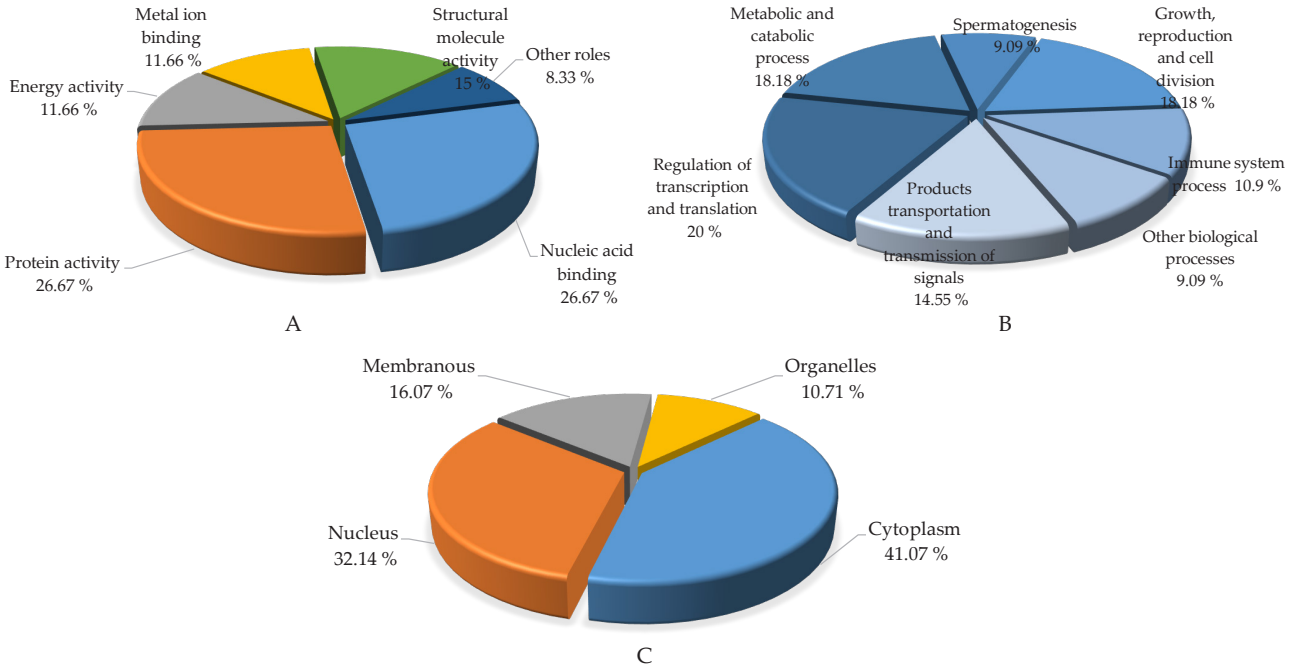

Figure 4. The frequency of functionally classified the EST-SSRs in various categories. A: Molecular functions; B: Biological processes; C: Cellular components.

by GCC and CAG, which were codes for leucine, alanine, and glutamine, respectively.

DISCUSSION

Our study showed that the freely available SSRs-containing EST sequences from EST library may be a good source of information to study functional domains, motifs distribution along genome, annotation, and gene ontology analysis. In this study, SSRs were treated as markers within EST sequences. Our method possibly helps to reveal the functional importance of the publicly available sequences of testis tissue. The study showed that the most frequent SSRs were tri-nucleotide and di-nucleotide repeats. As the length of the simple sequences increase, the number of the repeats decrease. This is possibly due to the higher rate of mutations in the longer length sequences which is naturally true and therefore they are less stable (Amos & Filipe, 2014).

The structure of SSRs may be changed by mutations, therefor the repeat of copy numbers will be changed. This transformation within microsatellite loci converts the perfect SSR to imperfect SSR (Sharma *et al.*, 2007). The investigation of EST-SSRs based on sequence types (perfect vs. imperfect) showed that perfect SSRs had higher frequency. Previous studies showed that the perfect microsatellites are less stable than imperfect microsatellites resulted by a mutation in their sequences and some of these imperfect SSRs have gene regulatory functions (Mudunuri & Nagarajaram, 2007). As an example from this study, the TA motif initially had 16 repeats and subsequently with a relatively low distance, its repetition started again with 6 more repeats. This region is one of the imperfect SSRs that is related to *SH2* domain and involves in gene regulation of intracellular signaling.

The tri-nucleotide repeats had the highest frequency, followed by dinucleotide repeats (Figure 3). It is a considerable point for inverse relationship between microsatellite length and their frequency (Molla *et al.*, 2015). In general, SSRs in Class II (sequences that were longer than 20 nucleotides) tend to be more variable

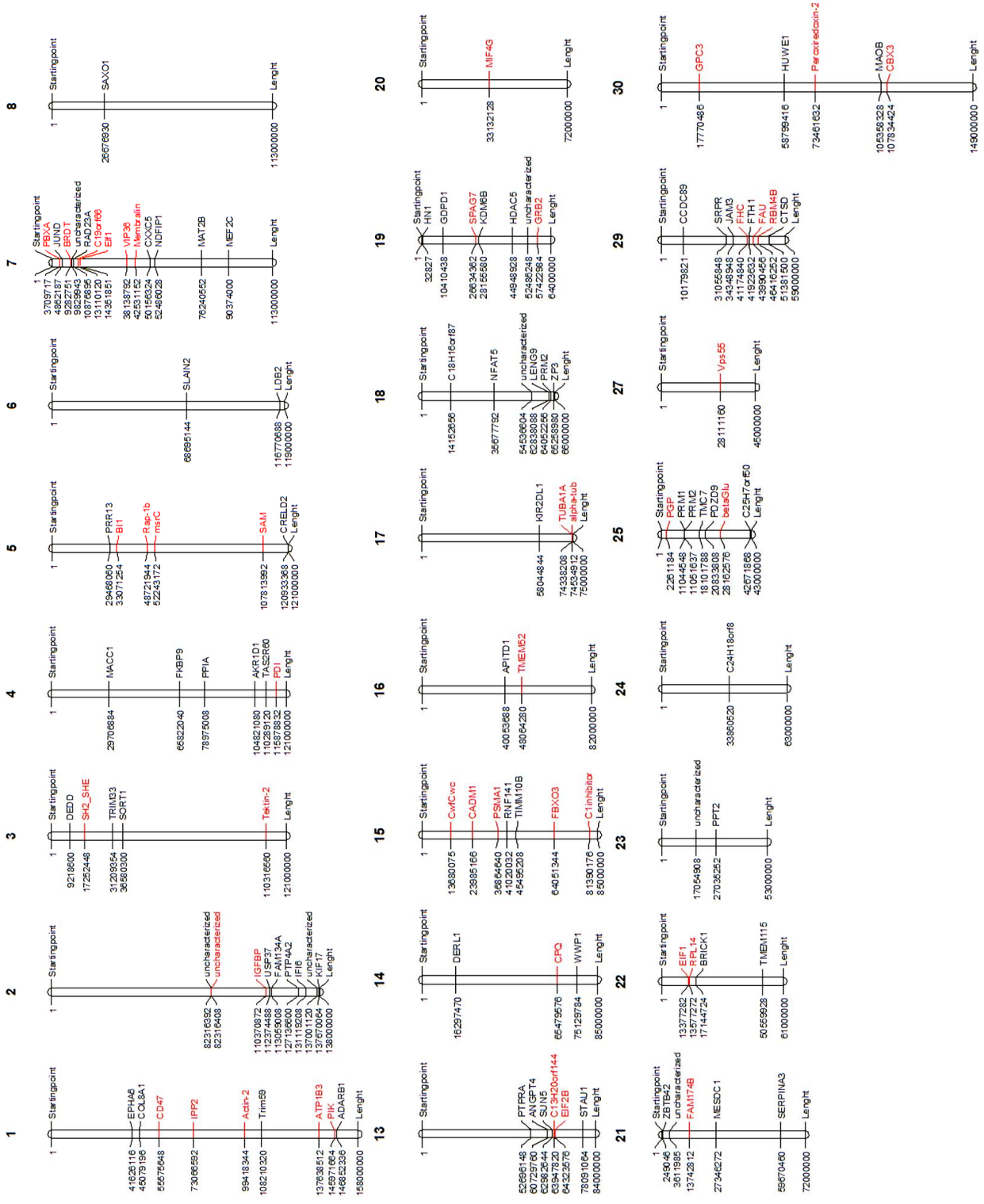

Figure 5. Chromosomal distribution of EST-SSRs identified based on linkage map. After functional analysis found the red characters were functional marker in contrast with black characters

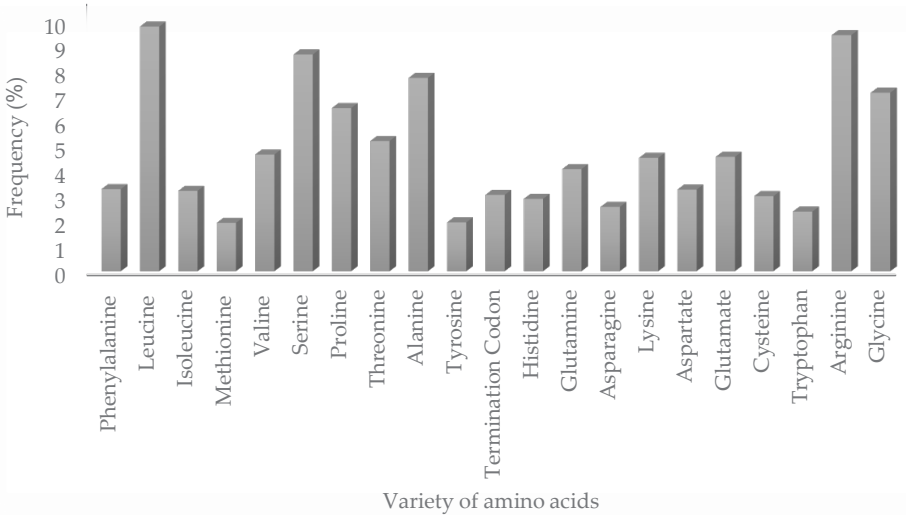

Figure 6. The frequency of amino acids within sequences containing special domains

and this class is more likely to preserve against slipped-strand abnormally (Temnykh *et al.*, 2001). Moreover, the SSRs within Class II, were more polymorphic than the SSRs of Class I, as was confirmed by the experimental data in human (Weber, 1990).

GT repeats were the most common in dimeric repeats as expected from previous studies on vertebrates (Toth *et al*., 2000) but in contrast to the other reports for cattle (Yan *et al.*, 2008), sheep (Zhang *et al.*, 2010), and chicken (Bakhtiarizadeh *et al.*, 2012a). This is unlikely to be true in plants that the AT/TA repeats have the highest frequency. In fact, this difference may be due to selection of ESTs of only one tissue (testis) in our study, whereas the reports from previous studies are based on global frequencies from all tissues.

Among the trimer repeats, GCC was the most abundant followed by GGC which was in agreement with previous studies in cattle and other mammalians (Li *et al.*, 2004). GCC codes for alanine amino acid. Abnormal frequency and distribution of this poly-alanine repeat can cause cleidocranial dysplasia that is a genetic anomaly of the unusual cellular process (Mundlos *et al.*, 1997).

Distribution pattern of codons within EST-SSRs of related domains indicated that CUG, GCC, and CAG had the higher frequencies that code leucine, alanine, and glutamine, respectively. A study on the mouse testis has been shown that CUG codon modulates in generation of Thioredoxin/Glutathione Reductase (Gerashchenko *et al*., 2010). The GCC codon by 5 repeats was linked to the Tektin-2 domain that plays an important role in sperm flagellar structure (Tanaka *et al.*, 2004). CAG has been reported as an effective codon on quantitative and qualitative features of sperm traits (Mostafa *et al.*, 2012).

SSRs usually applied as genetic markers to construct linkage maps and genetic diversity studies in non-coding regions (Blair *et al.*, 2003). So, EST-SSRs can be used to tracing the transcribed regions of genome and study the functional genes. BLASTX analysis showed that EST-derived microsatellite had the variety of categories matched to known proteins in public databases. As an example, most of the sequences that contain GT motif with E value of 1.04e-24 were matched to R3H domain. This domain is one of a group of metazoan proteins that are related to the sperm-associated antigen 7. In general, the transcription and translation factors, cell cycle and metabolic processes were the most frequent functions for the observed EST-SSRs.

The chromosomal locations of observed EST-SSRs were mapped via in silico mapping of the *Bos taurus* genome (Figure 5). As expected, there was a lack of link among testis-related genes and X chromosome, which is proved by previous studies (Moore *et al.*, 2005). There were 13 out of 153 sequences that did not attached to *Bos taurus* chromosomes. This difference was due to a low percentage of identity and query cover of EST-SSR sequences. The identified EST-SSRs loci, highlighted by red color in Figure 5 were the EST-SSRs that are known as domains and can be considered as regulating genes for reproductive traits.

CONCLUSION

Many of EST-SSRs are related to known domains in testis tissue. Polymorphisms that identified via SSRs especially the tri-nucleotides class II microsatellite repeats, may cause significant differences in their biological functions. As a result, localization and characterization of such markers can help tracing production of amino acids coded by identified repeats as shown in this study. Furthermore, with relatively fewer number of highly informative EST-SSRs compared to the other markers such as SNPs, it can help model fitting in genomic analysis and avoid over-parameterization.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interests with any financial, personal, or other relationships with other people or organization related to the material discussed in the manuscript.

ACKNOWLEDGEMENT

The authors thank Dr. Mostafa Modarresi (Department of Plant Breeding and Biotechnology, TMU, Iran) for his assistance and companionship.

REFERENCES

Amos, W. & L.N.S. Filipe. 2014. Microsatellite frequencies vary with body mass and body temperature in mammals, suggesting correlated variation in mutation rate. Peer J. 2: e663. https://doi.org/10.7717/peerj.663

Bakhtiarizadeh, M. R., B. Arefnejad, E. Ebrahimie, & M. Ebrahimi. 2012a. Application of functional genomic information to develop efficient EST-SSRs for the chicken (Gallus gallus). Genet. Mol. Res. 11: 1558-74. https://doi.org/10.4238/2012.May.21.12

Bakhtiarizadeh, M.R., B. Arefnejad, E. Ebrahimie, & M. Ebrahimi. 2012b. Application of functional genomic information to develop efficient EST-SSRs for the chicken (Gallus gallus). Genet. Mol. Res. 11: 1558-1574. https://doi.org/10.4238/2012.May.21.12

Bhattacharjee, R., C. O. Nwadili, C. A. Saski, A. Paterne, B. E. Scheffler, J. Augusto, A. Lopez-Montes, J. T. Onyeka, P. L. Kumar, & R. Bandyopadhyay. 2018. An EST-SSR based genetic linkage map and identification of QTLs for anthracnose disease resistance in water yam (Dioscorea alata L.). PloS One 13: e0197717. https://doi.org/10.1371/journal.pone.0197717

Blair, M. W., F. Pedraza, H. F. Buendia, E. Gaitan-Solis, S. E. Beebe, P. Gepts, & J. Tohme. 2003. Development of a genome-wide anchored microsatellite map for common bean (Phaseolus vulgaris L.). Theor. Appl. Genet. 107: 1362-74. https://doi.org/10.1007/s00122-003-1398-6

Djureinovic, D., L. Fagerberg, B. Hallström, A. Danielsson, C. Lindskog, M. Uhlén, & F. Pontén. 2014. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. MHR: Basic science of reproductive medicine 20: 476-488. https://doi.org/10.1093/molehr/gau018

Duan, Y., P. Liu, J. Li, J. Li, & P. Chen. 2013. Immune gene discovery by expressed sequence tag (EST) analysis of hemocytes in the ridgetail white prawn Exopalaemon carinicauda. Fish Shellfish Immunol. 34: 173-82. https://doi.org/10.1016/j.fsi.2012.10.026

Ehsani, A., L. Janss, D. Pomp, & P. Sorensen. 2016. Decomposing genomic variance using information from GWA, GWE and eQTL analysis. Anim. Genet. 47: 165-73. https://doi.org/10.1111/age.12396

Ellis, J. R. & J. M. Burke. 2007. EST-SSRs as a resource for population genetic analyses. Heredity (Edinb) 99: 125-32. https://doi.org/10.1038/sj.hdy.6801001

Garcia-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-Lopez, & C. P. Van Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc. Natl. Acad. Sci. U S A 113: E3995-4004. https://doi.org/10.1073/pnas.1519061113

Gerashchenko, M. V., D. Su, & V. N. Gladyshev. 2010. CUG start codon generates thioredoxin/glutathione reductase isoforms in mouse testes. J. Biol. Chem. 285: 4595-602. https://doi.org/10.1074/jbc.M109.070532

Gupta, P.K. & R.K. Varshney. 2000. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica 113: 163-185. https://doi.org/10.1023/A:1003910819967

Huson, K.M., W. Haresign, M.J. Hegarty, T.M. Blackmore, C. Morgan, & N.R. McEwan. 2015. Assessment of genetic relationship between six populations of Welsh Mountain sheep using microsatellite markers. Notes 216: 223. https://doi.org/10.17221/8171-CJAS

Janatova, M. & P. Pohlreich. 2004. Microsatellite markers in breast cancer studies. Prague Med Rep 105: 111-8.

Kalyana Babu, B., P. K. Agrawal, D. Pandey, J. P. Jaiswal, & A. Kumar. 2014. Association mapping of agro-morphological characters among the global collection of finger millet genotypes using genomic SSR markers. Mol. Biol. Rep. 41: 5287-97. https://doi.org/10.1007/s11033-014-3400-6

Kaur, S., P. S. Panesar, M. B. Bera, & V. Kaur. 2015. Simple sequence repeat markers in genetic divergence and marker-assisted selection of rice cultivars: a review. Crit. Rev. Food Sci. Nutr. 55: 41-9. https://doi.org/10.1080/10408398.2011.646363

Kirungu, J., Y. Deng, X. Cai, R. Magwanga, Z. Zhou, X. Wang, Y. Wang, Z. Zhang, K. Wang, & F. Liu. 2018. Simple sequence repeat (SSR) genetic linkage map of D genome diploid cotton derived from an interspecific cross between Gossypium davidsonii and Gossypium klotzschianum. International Journal of Molecular Sciences 19: 204. https://doi.org/10.3390/ijms19010204

Li, L., B. P. Brunk, J. C. Kissinger, D. Pape, K. Tang, R. H. Cole, J. Martin, T. Wylie, M. Dante, S. J. Fogarty, D. K. Howe, P. Liberator, C. Diaz, J. Anderson, M. White, M. E. Jerome, E. A. Johnson, J. A. Radke, C. J. Stoeckert, Jr., R. H. Waterston, S. W. Clifton, D. S. Roos, & L. D. Sibley. 2003. Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. Genome Res. 13: 443-54. https://doi.org/10.1101/gr.693203

Li, Y. C., A. B. Korol, T. Fahima, & E. Nevo. 2004. Microsatellites within genes: structure, function, and evolution. Mol. Biol. Evol. 21: 991-1007. https://doi.org/10.1093/molbev/msh073

Lu, G. & E. N. Moriyama. 2004. Vector NTI, a balanced all-in-one sequence analysis suite. Briefings in Bioinformatics 5: 378-388. https://doi.org/10.1093/bib/5.4.378

Ma, K., G. Qiu, J. Feng, & J. Li. 2012. Transcriptome analysis of the oriental river prawn, Macrobrachium nipponense using 454 pyrosequencing for discovery of genes and markers. PLoS One 7: e39727. https://doi.org/10.1371/journal.pone.0039727

Molla, K. A., A. B. Debnath, S. A. Ganie, & T. K. Mondal. 2015. Identification and analysis of novel salt responsive candidate gene based SSRs (cgSSRs) from rice (Oryza sativa L.). BMC Plant Biol. 15: 122. https://doi.org/10.1186/s12870-015-0498-1

Moore, T., A. McLellan, F. Wynne, & P. Dockery. 2005. Explaining the X-linkage bias of placentally expressed genes. Nat. Genet. 37: 3; author reply 3-4. https://doi.org/10.1038/ng0105-3a

Mostafa, T., L. H. El-Shahid, A. A. El Azeem, O. Shaker, H. Gomaa, & H. M. Abd El Hamid. 2012. Androgen receptor-CAG repeats in infertile Egyptian men. Andrologia 44: 147-51. https://doi.org/10.1111/j.1439-0272.2010.01125.x

Mudunuri, S. B. & H. A. Nagarajaram. 2007. IMEx: Imperfect Microsatellite Extractor. Bioinformatics 23: 1181-7. https://doi.org/10.1093/bioinformatics/btm097

Muller, M. P., S. Rothammer, D. Seichter, I. Russ, D. Hinrichs, J. Tetens, G. Thaller, and I. Medugorac. 2017. Genome-wide mapping of 10 calving and fertility traits in Holstein dairy cattle with special regard to chromosome 18. J. Dairy Sci. 100: 1987-2006. https://doi.org/10.3168/jds.2016-11506

Mundlos, S., F. Otto, C. Mundlos, J. B. Mulliken, A. S. Aylsworth, S. Albright, D. Lindhout, W. G. Cole, W. Henn, J. H. Knoll, M. J. Owen, R. Mertelsmann, B. U. Zabel, & B. R. Olsen. 1997. Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. Cell 89: 773-9. https://doi.org/10.1016/S0092-8674(00)80260-3

Pu, Y., W. Wang, Y. Yang, & R. R. Alfano. 2013. Native fluorescence spectra of human cancerous and normal breast tissues analyzed with non-negative constraint methods. Appl. Opt. 52: 1293-301. https://doi.org/10.1364/AO.52.001293

Qian, F., J. Guo, Z. Jiang, & B. Shen. 2018. Translational bioinformatics for cholangiocarcinoma: opportunities and challenges. International Journal of Biological Sciences 14: 920. https://doi.org/10.7150/ijbs.24622

Riar, D. S., S. Rustgi, I. C. Burke, K. S. Gill, & J. P. Yenish. 2011. EST-SSR Development from 5 Lactuca Species and Their Use in Studying Genetic Diversity Among L. serriola Biotypes. Journal of Heredity 102: 17-28. https://doi.org/10.1093/jhered/esq103

Sharma, P. C., A. Grover, & G. Kahl. 2007. Mining microsatellites in eukaryotic genomes. Trends Biotechnol. 25: 490-8. https://doi.org/10.1016/j.tibtech.2007.07.013

Stamatoyannopoulos, J. A. 2004. The genomics of gene expression. Genomics 84: 449-57. https://doi.org/10.1016/j.ygeno.2004.05.002

Tae, H., D. Ryu, S. Sureshchandra, & J. H. Choi. 2012. ESTclean: a cleaning tool for next-gen transcriptome shotgun sequencing. BMC Bioinformatics 13: 247. https://doi.org/10.1186/1471-2105-13-247

Taheri, S., T. L. Abdullah, M. Y. Rafii, J. A. Harikrishna, S. P. O. Werbrouck, C. H. Teo, Ma. Sahebi, & P. Azizi. 2019. De novo assembly of transcriptomes, mining, and development of novel EST-SSR markers in Curcuma alismatifolia (Zingiberaceae family) through Illumina sequencing. Scientific Reports 9: 3047. https://doi.org/10.1038/s41598-019-53129-x

Tanaka, H., N. Iguchi, Y. Toyama, K. Kitamura, T. Takahashi, K. Kaseda, M. Maekawa, & Y. Nishimune. 2004. Mice deficient in the axonemal protein Tektin-t exhibit male infertility and immotile-cilium syndrome due to impaired inner arm dynein function. Mol. Cell Biol. 24: 7958-64. https://doi.org/10.1128/MCB.24.18.7958-7964.2004

Temnykh, S., G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, & S. McCouch. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11: 1441-52. https://doi.org/10.1101/gr.184001

Thiel, T., W. Michalek, R. K. Varshney, & A. Graner. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor. Appl. Genet. 106: 411-22. https://doi.org/10.1007/s00122-002-1031-0

Toth, G., Z. Gaspari, & J. Jurka. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10: 967-81. https://doi.org/10.1101/gr.10.7.967

Varshney, R. K., T. Thiel, N. Stein, P. Langridge, & A. Graner. 2002. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell. Mol. Biol. Lett. 7: 537-46.

Voorrips, R. E. 2002. MapChart: Software for the graphical presentation of linkage Maps and QTLs. Journal of Heredity 93: 77-78. https://doi.org/10.1093/jhered/93.1.77

Wang, Z., G. Yu, B. Shi, X. Wang, H. Qiang, & H. Gao. 2014. Development and characterization of simple sequence repeat (SSR) markers based on RNA-sequencing of Medicago sativa and in silico mapping onto the M. truncatula genome. PLoS One 9: e92029. https://doi.org/10.1371/journal.pone.0092029

Weber, J. L. 1990. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics 7: 524-30. https://doi.org/10.1016/0888-7543(90)90195-Z

Yan, Q., Y. Zhang, H. Li, C. Wei, L. Niu, S. Guan, S. Li, & L. Du. 2008. Identification of microsatellites in cattle unigenes. J. Genet. Genomics. 35: 261-6. https://doi.org/10.1016/S1673-8527(08)60037-5

Yan, Z., F. Wu, K. Luo, Y. Zhao, Q. Yan, Y. Zhang, Y. Wang, & J. Zhang. 2017. Cross-species transferability of EST-SSR markers developed from the transcriptome of Melilotus and their application to population genetics research. Scientific Reports 7: 17959. https://doi.org/10.1038/s41598-017-18049-8

Zhang, W., Z. Wang, Z. Zhao, X. Zeng, H. Wu, & P. Yu. 2010. Using bioinpormcotics methods to develop EST-SSR makers from sheep's ESTs. J. Anim. Vet. Adv. 9:2759-2762. https://doi.org/10.3923/javaa.2010.2759.2762