Kocatepe Veterinary Journal

The Detection of SSRs From The Terrier Dog Normal and Tumoral Mammary Gland Tissue EST Libraries

Zehra OMEROGLU ULU[1], Merve CELİK[2], Salih ULU[1], Nehir OZDEMIR OZGENTURK[1]*

[1]Faculty of Art and Science, Molecular Biology and Genetics, Yildiz Technical University, Istanbul, Turkey
[2]Medical Biotechnology Ph.D. Program, Institute of Health Sciences, Acıbadem Mehmet Ali Aydınlar University, 34752 İstanbul, Turkey.

ABSTRACT

Dogs share a common environment with humans and knowledge of the specific dog breed diseases is very useful in developing a model for human cancer studies. ESTs represent part of the transcribed genome of an organism and are an important resource for identifying microsatellites. Simple Sequence Repeats (SSRs), or microsatellites, which contain repetitive DNA sequences, are among the most powerful genetic markers known. The development of EST-SSRs has become a fast, efficient, and low-cost option for genomic studies. In this study, to determine SSRs from EST libraryof mammary gland tissue of the Terrier dog that has 2304 ESTs; SSRIT and IMEx software, which have web-based versions and are easily accessible, were used. SSRIT finds motifs from 2 to 10 base lengths and adjusts the minimum number of repeats by eliminating single nucleotide motifs. IMEx finds perfect and imperfect microsatellites separately. It can find motifs of different lengths from 1 to 6 and the minimum number of repeats can be set. In addition, the appropriate primer for the desired SSR region can be designed. The 2, 3, 4, 5 and 6 nucleotide motifs were found for normal tissue ESTs whereas 5 nucleotide motifs were not found for tumoral tissue ESTs.

Keywords: SSR, EST-SSR, Terrier, Cancer, Dog

***

Terrier Köpek Normal ve Tümörlü Meme Dokusu EST Kütüphanelerinden SSR'ların Belirlenmesi

ÖZ

Köpekler, insanlarla ortak bir çevreyi paylaşır ve belirli köpek ırklarının hastalıkları hakkında bilgi, insan kanser çalışmaları için bir model geliştirmede çok yararlıdır. EST'ler, bir organizmanın transkribe edilen genomunun bir parçasını temsil eder ve mikrosatellitleri tanımlamak için önemli bir kaynaktır. Tekrarlayan DNA dizilerini içeren Basit Dizi Tekrarları (SSR'ler) veya mikrosatellitler, bilinen en güçlü genetik belirteçler arasındadır. EST-SSR'lerin gelişimi, genomik çalışmalar için hızlı, verimli ve düşük maliyetli bir seçenek haline gelmiştir. Bu çalışmada, 2304 EST içeren Terrier köpeğin meme bezi dokusunun EST koleksiyonundan SSR'lerin belirlenmesi için; Web tabanlı sürümleri olan ve kolayca erişilebilen SSRIT ve IMEx yazılımları kullanılmıştır. SSRIT, 2 ila 10 baz uzunluğundaki motifleri bulur ve tekli nükleotid motiflerini ortadan kaldırarak minimum tekrar sayısını ayarlar. IMEx mükemmel ve kusurlu mikrosatellitleri ayrı ayrı bulur. Farklı uzunluklarda 1 ile 6 arasındaki motifleri bulabilir ve minimum tekrar sayısı ayarlanabilir. Ek olarak, istenen SSR bölgesi için uygun primer tasarlanabilir. Normal doku EST'leri için 2, 3, 4, 5 ve 6 nükleotidli motifler bulunurken, tümörlü doku EST'leri için 5 nükleotidli motif bulunamamıştır.

Anahtar Kelimeler: SSR, EST-SSR, Terrier, Kanser, Köpek

INTRODUCTION

It has been shown that the development of experimental technologies, as well as the development of information technologies, have been accompanied by the accumulation of different biological data (Bishop et al. 2015). The accumulation of the data in the biology has increased the need for software to automate the operations that are impossible to do manually. Developments in the field of bioinformatics have made it much easier to form large sets of biological data (Feagan et al. 2007, Collins et al. 2003). Recently, genome studies have been progressing rapidly. These genomic studies are aimed to determine when and how genes are active. Because, if the working mechanisms of the genes are understood, how these genes react to the disease can also be found (Kanehisa et al. 2003).

Dogs are affected by a large number of diseases and different dog breeds are affected by different diseases. Many dog species are particularly susceptible to inherited diseases such as cancer, heart disease, rheumatoid arthritis, autoimmune disorders, deafness and blindness (Helmen et al. 1993, Perez-Alenza et al. 2000). Dogs share a common environment with humans and they are exposed to the same carcinogens, therefore, dogs may serve as models of human diseases (Pinho et al. 2012, Antuofemo et al. 2007). Knowledge of the specific dog breed diseases is very useful in developing a model for human cancer studies. Dog is a useful model organism for human breast cancer studies due to its similarity to the human genome, genetic and morphological structure (Ozdemir Ozgenturk et al. 2018).

Researchers began collecting different types of data such as the expressed sequence-tagged sites or ESTs that would influence the nature of the sequence databases (Nagaraj et al. 2007). An EST is a short subsequence of a cDNA obtained as a result of sequence analysis. It is usually 300-500 base pairs in length and is not functional in the coding of proteins (Adams et al. 1991, Liu et al. 2013). ESTs provide a fast and inexpensive way to obtain data on the discovery of new genes, genome mapping, identification of coding regions, gene expression and regulation (Ozdemir Ozgenturk et al. 2018). ESTs represent part of the transcribed genome of an organism and are the important resource for identifying microsatellites (Li et al. 2014).

Simple Sequence Repeats (SSRs), or microsatellites, which contain repetitive DNA sequences, are among the most powerful genetic markers known. SSRs are 2-6 bp in length, short tandem-repeat sequences that are present highly in eukaryotic genomes. The repeating sequence is usually simple, consisting of two, three or four nucleotides, which can be repeated 10 to 100 times (Akagi et al. 1996, George et al. 2015,

Seyedabadi et al. 2017, Zhang et al. 2014). They have been used to study the gene mapping, genetic diversity assessing and population structure of many species because of their abundance and hypervariability throughout the genome. These markers generally show high degree polymorphism between different individuals in a species as well as between different species (Zalapa et al. 2012, Varshney et al. 2005).

SSRs could be developed from ESTs, which are especially associated with functional genes. With the high-throughput sequencing technology, the development of EST-SSRs has become a fast, efficient, and low-cost option for genomic studies (Li et al. 2014).

There are several software tools to determine SSRs in EST libraries. Many desktop and web-based tools were developed such as TRF (Tandem Repeat Finder), SSRIT (Simple Sequence Repeat Identification Tool), TROLL (Tandem Repeat Occurrence Locator), WebSat (A Web Software for MicroSatellite Marker Development), IMEx (Imperfect Microsatellite Extractor) (Benson et al. 1999, Temnykh et al. 2001, Castelo et al. 2002, Martins et al. 2009, Mudunuri et al. 2007).

SSRIT (Simple Sequence Repeat Identification Tool) was developed in 2001 by S. Temnykh, G. DeClerk, A. Lukashova and colleagues to detect SSRs in a set of 57.8 Mb size rice. SSRIT accepts the data in FASTA format. By eliminating single nucleotide motifs, it can find motifs from 2 to 10 base lengths and adjust the minimum number of repeats. It gives SSR file in table format as output. This table can give information about the motif, length, beginning and ending base number of the SSR and the length of the loaded sequence for analysis (Temnykh et al. 2001).

IMEx (Imperfect Microsatellite Extractor) was developed by SB Munduri in 2007. This program, which includes desktop and web-based versions, finds perfect and imperfect microsatellites separately. It can find motifs of different lengths from 1 to 6 and the minimum number of repeats can be set. The imperfection percentage to be detected can be adjusted. The program can analyze the data in a file, but the data to be uploaded must be in FASTA or Plain format. The output is in a table format and shows the motif, motif length, beginning and ending sequence. In addition, the appropriate primer for the desired SSR region can be designed (Martins et al. 2009).

In the present study, we aimed to gain the gene expression profile of mammary gland tissue based on the total 2304 ESTs as well as results of a comparison of normal and tumoral stages. These results facilitate the use of functional genomics approaches aiming at

creating gene expression patterns of tumoral and normal mammary gland. In this study, identification of SSRs from Terrier dog EST library was performed by bioinformatics analysis.

MATERIALS and METHODS

Expressed Sequence Tags: 2304 EST, sequenced from normal and tumoral mammary gland tissue cDNA libraries prepared from Terrier dogs by Nehir Ozdemir Ozgenturk et al. were used.

SSRs Detection: It was aimed to determine the correct SSRs by comparing the two common tools to avoid mistakes caused by software weaknesses. In this study, SSRIT was used to determine SSRs in Terrier EST library that has 2304 EST. IMEx was used to control the found SSRs.

For each software, the 2304 EST sequences were divided into ten parts to search. Each piece was pasted to the search window individually.

In SSRIT software, parameters from the main page were set to maximum 6 nucleotides motifs-length and minimum 5 repeats. Output was generated and the results were tabulated.

In IMEx software, basic search mode has been selected from the main page. Parameters from the main page were set to "perfect" as the repeat type, 2-6 nucleotides as motifs-length and 5 as minimum repeat number. Output was generated the results were tabulated.

Statistical Analysis: Data were analyzed using paired t-test. A difference was regarded as significant if $p < 0.05$.

RESULTS

SSR motifs of the EST library have been determined by SSRIT software and the found SSRs have been controlled by IMEx software. Table 1 to 4 show the SSR results of each software separately in tabular form. Normal mammary gland tissue and tumoral mammary gland tissue SSR results differ from each other. The 2, 3, 4, 5 and 6 nucleotide motifs were found for normal tissue ESTs whereas 5 nucleotide motifs were not found for tumoral tissue ESTs. CT is the most common motif in 2 nucleotides motifs for both normal and tumoral EST library. There are some common results with the two software. As the number of nucleotides in the motif increased, the repeat number and the total number of the motifs found in the EST library decreased. Based on the statistical analysis of normal and tumoral SSRs, there is no statistically significant difference between samples. Tables 5 to 8 are given as supplemental tables.

Normal EST SRRs
2, 3, 4 and 6 nucleotide (nt) long motifs were counted by all two software and total number of motifs and most common motifs were reported. As shown in Table 1 and 2, CT is the most common motif in 2 nucleotides motifs for normal EST library. The most common motifs in 3 nucleotides motifs are GGA and CTG. 4 and 5 nucleotides motifs are not common for two software. TCGGGG is the most common motif in 6 nucleotides motifs for normal EST library.

Tumoral EST SSRs
2, 3, 4 and 6 nucleotides long motifs were counted by all two software and the total number of motifs and most common motifs were reported. As shown in Table 3 and 4, CT is the most common motif in 2 nucleotides motifs for tumoral EST library. The most common motif in 3 nucleotides motifs is CAG with 7 repeats number. In motifs with 4 nucleotides, the most abundant and the motifs with the longest repeats are the same and are 5 repeats of AATG, AATA, and GATG. 5 nucleotides motifs were not found for tumoral ESTs. CCGGCT is the most common motif in 6 nucleotides motifs for tumoral EST library.

Table 1. SSRIT Results for Normal ESTs (Normal EST'ler için SSRIT Sonuçları)

Motif Length	Amount of Motif Types	Total Amount of Motifs	The Most Common Motifs	Amount of The Most Common Motifs
2 nucleotides	12	102	$(CT)_{14}$	22
3 nucleotides	16	25	$(GGA)_7$ $(CTG)_9$	3
4 nucleotides	9	11	$(ATAA)_5$ $(TTTC)_{18}$	2
5 nucleotides	1	1	$(CCCCG)_6$	1
6 nucleotides	1	1	$(TCGGGG)_5$	1

*$(Motif)_{repeat\ number}$

105

Table 2. IMEx Results for Normal ESTs (Normal EST'ler için IMEx Sonuçları)

Motif Length	Amount of Motif Types	Total Amount of Motifs	The Most Common Motifs	Amount of The Most Common Motifs
2 nucleotides	12	102	$(CT)_{14}$	22
3 nucleotides	16	25	$(GGA)_7$ $(CTG)_9$	3
4 nucleotides	7	10	$(TCTT)_{17}$	3
5 nucleotides	2	2	$(CGCCC)_6$ $(AACCC)_5$	1
6 nucleotides	2	2	$(TCGGGG)_5$ $(TCCTTC)_5$	1

*$(Motif)_{repeat\ number}$

Table 3. SSRIT Results for Tumoral ESTs (Tümör EST'leri için SSRIT Sonuçları)

Motif Length	Amount of Motif Types	Total Amount of Motifs	The Most Common Motifs	Amount of The Most Common Motifs
2 nucleotides	12	96	$(CT)_{15}$	19
3 nucleotides	9	17	$(CAG)_7$	4
4 nucleotides	3	3	$(AATG)_5$ $(AATA)_5$ $(GATG)_5$	1
6 nucleotides	1	2	$(CCGGCT)_5$	2

*$(Motif)_{repeat\ number}$

Table 4. IMEx Results for Tumoral ESTs (Tümör EST'leri için IMEx Sonuçları)

Motif Length	Amount of Motif Types	Total Amount of Motifs	The Most Common Motifs	Amount of The Most Common Motifs
2 nucleotides	12	96	$(CT)_{15}$	19
3 nucleotides	9	17	$(CAG)_7$	4
4 nucleotides	3	3	$(AATG)_5$ $(AATA)_5$ $(GATG)_5$	1
6 nucleotides	1	2	$(CCGGCT)_5$	2

*$(Motif)_{repeat\ number}$

Table 5. Statistical data of 2 nucleotides lenght SSRs (2 nükleotid uzunluklu SSR'ların istatiksel verileri)

2 nucleotides length	(SSRIT results)		(IMEX results)	
	Normal	Tumor	Normal	Tumor
Mean	8,5	8	8,5	8
Variance	37,545	32,545	37,364	21,636
Observations	12	12	12	12
Pearson Correlation	0,624		0,435	
Hypothesized Mean Difference	0		0	
df	11		11	
t Stat	0,337		0,296	
P(T<=t) one-tail	0,371		0,386	
t Critical one-tail	1,796		1,796	
P(T<=t) two-tail	0,743		0,773	
t Critical two-tail	2,201		2,201	

Table 6. Statistical data of 3 nucleotides lenght SSRs (3 nükleotid uzunluklu SSR'ların istatiksel verileri)

3 nucleotides length	(SSRIT results)		(IMEX results)	
	Normal	Tumor	Normal	Tumor
Mean	1,19	0,81	1,19	0,81
Variance	0,862	1,362	0,862	1,362
Observations	21	21	21	21
Pearson Correlation	-0,565		-0,565	
Hypothesized Mean Difference	0		0	
df	20		20	
t Stat	0,94		0,94	
P(T<=t) one-tail	0,179		0,179	
t Critical one-tail	1,725		1,725	
P(T<=t) two-tail	0,358		0,358	
t Critical two-tail	2,086		2,086	

Table 7. Statistical data of 4 nucleotides lenght SSRs (4 nükleotid uzunluklu SSR'ların istatiksel verileri)

4 nucleotides length	(SSRIT results)		(IMEX results)	
	Normal	Tumor	Normal	Tumor
Mean	1	0,273	1,111	0,333
Variance	0,4	0,218	0,861	0,25
Observations	11	11	9	9
Pearson Correlation	-0,677		-0,629	
Hypothesized Mean Difference	0		0	
df	10		8	
t Stat	2,39		1,793	
P(T<=t) one-tail	0,019		0,055	
t Critical one-tail	1,812		1,86	
P(T<=t) two-tail	0,038		0,111	
t Critical two-tail	2,228		2,306	

Table 8. Statistical data of 6 nucleotides lenght SSRs (6 nükleotid uzunluklu SSR'ların istatiksel verileri)

6 nucleotides length	(SSRIT results)		(IMEX results)	
	Normal	Tumor	Normal	Tumor
Mean	0,5	1	0,667	0,667
Variance	0,5	2	0,333	1,333
Observations	2	2	3	3
Pearson Correlation	-1		-1	
Hypothesized Mean Difference	0		0	
df	1		2	
t Stat	-0,333		0	
P(T<=t) one-tail	0,398		0,5	
t Critical one-tail	6,314		2,92	
P(T<=t) two-tail	0,795		1	
t Critical two-tail	12,706		4,303	

DISCUSSION

In recent years, some studies have been published according to different dog breeds, but the information is largely dispersed and lacks functional interpretation. The aim of this paper is to present such a resource which allows the functional interpretation of the SSR data. This study and the associated database enables the integrated interpretation of SSR data from normal and tumoral mammary gland tissue of Terrier dog.

In the previous study (Ozdemir Ozgenturk et al. 2018), we have generated 2304 ESTs from a normal mammary tissue of healthy Terrier dog and a tumoral mammary tissue of another sick Terrier dog with breast cancer. To further understand the molecular mechanism and differences between tumoral and normal mammary gland tissue (Alcıgır et al. 2018), gene expression profiles were compared. This study will help to identify the genes in the mammary tissue of Terrier dog, which is used as a model organism in cancer research (Pinho et al. 2012), will be an important source for future studies and will help to understand the molecular biology of cancer.

The lack of sufficient simple sequence repeat (SSR) markers has hindered the progress of genetic selection breeding and molecular research for Terrier dog. To this end, we aimed to develop and characterize expressed sequence tag (EST)–SSRs from the mammary gland tissue of Terrier dog. Totally, potential EST-SSR markers were identified and distributed in the coding and non-coding regions. These EST-SSR markers produced in this study will greatly facilitate the Terrier dog genetic diversity and molecular breeding studies (Kaur et al. 2015).

SSRs of normal and tumoral mammary gland tissues were found with SSRIT software and controlled with IMEx software. There are differences between the data of two mammary gland tissues.

In this study, a total of 140 EST-SSRs were determined in SSRIT software for normal tissue and no significant difference was observed between them when controlled by IMEx software. A total of 118 EST-SSRs for tumoral tissue in the SSRIT software were detected and the results were observed exactly the same when controlled by the IMEx software. Results of the SSRs mined from EST libraries using SSRIT and IMEx showed dinucleotide repeats as the most abundant SSR motif type and CT is the most frequent repeats for both normal and tumoral EST libraries. Among these markers, di- and tri- repeat types were the abundant types.

Although there were some differences between the two software's work, the absence of any major differences between the results of these two programs proved the validity of SSRs and supported each other's results.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991; 252(5013): 1651–1656.

Akagi H, Yokozeki Y, Inagaki A, Fujimura T. Microsatellite DNA markers for rice chromosomes. Theor Appl Genet. 1996; 93(7): 1071-1077.

Alcıgır E.M. Anadol E. Gultıken N. Alkan Karakas K. Alkan H. Kanca H. Cxc Chemokıne Ligand 12 and G Protein-Coupled Receptor 30 Expressions In Canine Mammary

Tumors Of Mixed Origin. Kocatepe Vet J. 2018; 11(2): 104-112.

Antuofemo E, Miller MA, Pirino S, Xie J, Badve S, Mohammed SI. Spontaneous mammary intraepitelial lesions in dogs: a model of breast cancer. Cancer Epidemiol Biomarkers Prev. 2007; 16(11): 2247-2256.

Benson B. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27(2): 573-580.

Bishop OT, Adebiyi EF, Alzohairy AM, Everett D, Ghedira K, Ghouila A, Kumuthini J, Mulder NJ, Panji S, HG Patterton. Bioinformatics education—perspectives and challenges out of Africa. Brief Bioinform. 2015; 16(2): 355–364.

Castelo AT, Martins W, Gao GR. TROLL – tandem repeat occurrence locator. Bioinformastics. 2002; 18(4): 634–636.

Collins FS, Morgan M, Patrinos A. The Human Genome Project: Lessons from Large-Scale Biology. Science. 2003; 300(5617): 286-290.

Feagan L, Rohrer J, Garret A, Amthauer H, Komp Eoh, J nson D, Hock A, Clark T, Lushington G, Minden G, Frost V. Bioinformatics process management: information flow via a computional journal. Source Code Biol Med. 2007; 2(9): 2-3.

George B, Bhatt BS, Awasthi M, George B, Singh AK. Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. Curr Genet. 2015; 61(4): 665-77.

Helmen E, Bergstrom R, Holmberg L, Spangberg IB, Hnasson K, Lindgren A. Prognostic factors in canine mammary tumors: a multivariate study of 202 consecutive cases. Vet Pathol. 1993; 30(1): 20-27.

Kanehisa M, Bork P. Bioinformatics in the post-sequence era. Nat Genet. 2003; 33: 305–310.

Kaur R, Sharma N, Raina R. Identification and functional annotation of expressed sequence tags based SSR markers of Stevia rebaudiana. Turk J Agric For. 2015; 39: 439-450.

Li RH, Lu SK, Zhang CL, Song WW, Mu CK, Wang CL. Development of polymorphic expressed sequence tag-single sequence repeat markers in the common Chinese cuttlefish, Sepiella maindroni. Genet Mol Res. 2014; 13(3): 5503-5506.

Li YF, Gao XG, Liu WD, Tian Y, Gao L, Li SL, He CB. Microsatellite markers derived from Japanese scallop (Mizuhopecten yessoensis) expressed sequence tags. Genet Mol Res.2014; 13(1): 1989-1992.

Liu D, Zhi L, Ma M, Qiao D, Wang M, Wang Y, Jin B, Li A, Liu G, Zhang Y, Song Y, Zhang H. Primarily screening and analyzing ESTs differentially expressed in rats' primary liver cancer. Chin J Cancer Res. 2013; 25(1): 71–78.

Martins WS, Lucas DCS, Neves KFS, Bertioli DJ. WebSat - A Web Software for MicroSatellite Marker Development. Bioinformation. 2009; 3(6):282-283.

Mudunuri SB, Nagarajaram HA. IMEx: Imperfect Microsatellite Extractor. Bioinformatics. 2007; 23(10): 1181-1187.

Nagaraj SH, Gasser RB, Ranganathan S. A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform. 2007; 8(1): 6–21.

Ozdemir Ozgenturk N, Omeroglu Ulu Z, Ulu S, Celik M, Tellioglu B, Yildirim F, Akis Akad I, Gurel A, Un C, **Oztabak KO.** Analysis of expressed genes in normal and tumoral mammary gland tissue of the Terrier dog. Kafkas Univ Vet Fak Derg. 2018; 24 (3): 357-364.

Perez-Alenza MD, Pena L, del Castillo N, Nieto AI. Factors influencing the incidence and prognosis of canine mammary tumors. J Small Anim Pract. 2000; 41(7): 287-291.

Pinho SS, Carvalho S, Cabral J, Reis CA, Gärtner F. Canine tumors: A spontaneous animal model of human carcinogenesis. Transl Res. 2012; 159(3): 165-172.

Seyedabadi HR, Sofla SS. Microsatellite Analysis for Parentage Verification and Genetic Characterization of the Turkmen Horse Population. Kafkas Univ Vet Fak Derg. 2017; 23 (3): 467-471.

Temnykh S, DeClerk G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 2001; 11: 1441–1452.

Varshney RK, Sigmund R, Borner A. Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. Plant Science. 2005; 168(1): 195–202.

Zalapa JE, Cuevas H, Zhu H. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am J Bot. 2012; 99(2):193-208.

Zhang J, Ma W, Song X, Lin Q, Gui JF, Mei J. Characterization and Development of EST-SSR Markers Derived from Transcriptome of Yellow Catfish. Molecules. 2014; 19(10): 16402-16415.