American Journal of Primatology 77:34–43 (2015)

RESEARCH ARTICLE

Core Promoter Short Tandem Repeats as Evolutionary Switch Codes for Primate Speciation

MINA OHADI[1]*, ELAHEH VALIPOUR[1], SAEED GHADIMI-HADDADAN[1], PEGAH NAMDAR-ALIGOODARZI[1], ABOUZAR BAGHERI[1], ALI KOWSARI[2], MARYAM REZAZADEH[1], HOSSEIN DARVISH[3], AND SOMAYEH KAZEMINASAB[1]

[1]*Genetics Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran*
[2]*Stem Cell Research Center, Golestan University of Medical Science, Gorgan, Iran*
[3]*Department of Medical Genetics, Shahid Beheshti University of Medical Sciences, Tehran, Iran*

Alteration in gene expression levels underlies many of the phenotypic differences across species. Because of their highly mutable nature, proximity to the +1 transcription start site (TSS), and the emerging evidence of functional impact on gene expression, core promoter short tandem repeats (STRs) may be considered an ideal source of variation across species. In a genome-scale analysis of the entire *Homo sapiens* protein-coding genes, we have previously identified core promoters with at least one STR of ≥6-repeats, with possible selective advantage in this species. In the current study, we performed reverse analysis of the entire *Homo sapiens* orthologous genes in mouse in the Ensembl database, in order to identify conserved STRs that have shrunk as an evolutionary advantage to humans. Two protocols were used to minimize ascertainment bias. Firstly, two species sharing a more recent ancestor with *Homo sapiens* (i.e. *Pan troglodytes* and *Gorilla gorilla gorilla*) were also included in the study. Secondly, four non-primate species encompassing the major orders across Mammals, including Scandentia, Laurasiatheria, Afrotheria, and Xenarthra were analyzed as out-groups. We introduce STR evolutionary events specifically identical in primates (i.e. *Homo sapiens*, *Pan troglodytes*, and *Gorilla gorilla gorilla*) vs. non-primate out-groups. The average frequency of the identically shared STR motifs across those primates ranged between 0.00005 and 0.06. The identified genes are involved in important evolutionary and developmental processes, such as normal craniofacial development (*TFAP2B*), regulation of cell shape (*PALMD*), learning and long-term memory (*RGS14*), nervous system development (*GFRA2*), embryonic limb morphogenesis (*PBX2*), and forebrain development (*APAF1*). We provide evidence of core promoter STRs as evolutionary switch codes for primate speciation, and the first instance of identity-by-descent for those motifs at the interspecies level. Am. J. Primatol. 77:34–43, 2015.

Key words: short tandem repeat; core promoter; primate; non-primate; evolution

INTRODUCTION

Once considered "junk" DNA, short tandem repeats (STRs) are now believed to play a significant role in genome evolution by creating and maintaining quantitative genetic variation [Heidari et al., 2012; Iglesias et al., 2004; Jansen et al., 2012; King et al., 2006; Valipour et al., 2013]. Tandem repeats located in coding regions may increase the evolvability of proteins [Fondon & Garner, 2004; Gemayel et al., 2010; Verstrepen et al., 2005]. Non-coding genes such as microRNAs are also influenced by purifying selection in STRs [Trivedi & Hancock, 2012]. Repeats in the *cis*-regulatory regions are implicated in quantitative genetics and complex traits [Donaldson et al., 2008; Hammock & Young, 2005]. *Cis*-regulatory STRs are also suggested to be involved in

Abbreviations: SNP, single nucleotide polymorphism; STR, short tandem repeat; TF, transcription factor; TSS, transcription start site.

Contract grant sponsor: University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.

Conflicts of interest: None.

*Correspondence to: Mina Ohadi, Genetics Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran. E-mail: ohadi.mina@yahoo.com

Received 3 February 2014; revised 7 April 2014; revision accepted 16 May 2014

DOI: 10.1002/ajp.22308
Published online 5 August 2014 in Wiley Online Library (wileyonlinelibrary.com).

evolutionary processes in *Homo sapiens* [Rockman & Wray, 2002].

Vinces and co-workers have shown that genes driven by repeat-containing promoters show significantly higher rates of transcriptional divergence, and variations in repeat length result in changes in expression and local nucleosome positioning [Vinces et al., 2009]. In their experiments, replacement of tandem repeats with various sequences of the same size did not restore normal gene expression. Borel and co-workers have recently reported that in comparison with single nucleotide polymorphisms (SNPs), promoter tandem repeat variation is of more effect as causative *cis* expression quantitative trait loci for protein-coding gene expression in the case of *CSTB* gene [Borel et al., 2012]. Because of their combinatory and repeat potential, STRs may be considered a more efficient source of evolutionary codes for speciation in comparison with SNPs. STRs, by virtue of their special mutational and functional qualities have a major role in generating the genetic variation underlying adaptive evolution [Jansen et al., 2012; Kashi & King, 2006; Mohammadparast et al., 2014]. Selection could shape STRs into "tuning knobs" that facilitate evolutionary adaptation by increasing adjustability [King et al., 2006]. This possibility is consistent with numerous examples of evolutionary conservation of STRs in genes with neurological, neurodevelopmental, and embryogenic function [Bolton et al., 2013; Darvish et al., 2013; King, 2012; Valipour et al., 2013; Zarif Yeganeh et al., 2009, 2010].

Because of their proximity to the +1 transcription start site (TSS), core promoter STRs may have a *bona fide* effect on phenotypic variation across species. The majority of the *Homo sapiens* core promoter STRs are conserved across evolution [Ohadi et al., 2012]. In a similar finding, it has recently been shown that STRs near the TSSs of genes (most of them involved in development) are often highly conserved, and that distance from a STR to the nearest TSS is a good predictor of the STR conservation score [Sawaya et al., 2012]. In a recent finding by Li et al. [2012] a short genomic di-nucleotide repeat structure in the promoter of *ECE.1* constitutes a novel and functional core promoter element, coincides with *Homo sapiens* evolution, and contributes to the pathogenesis of Alzheimer's disease. This genomic region is conserved between *Homo sapiens* and *Pan troglodytes*. The *Homo sapiens SOX5*, *GABRA3*, and *MECOM* genes provide further examples of functional core promoter STRs [Heidari et al., 2012; Valipour et al., 2013].

In a genome-scale analysis of the entire human protein-coding core promoters annotated in the GeneCards database, we have recently catalogued the *Homo sapiens* core promoter STRs with exceptional repeat numbers of ≥6-repeats [Ohadi et al., 2012], which comprise less than 2% of the genes. The idea was that the STRs that have reached that length may be of selective advantage, and therefore of prime importance for a pilot research of this kind. At the top of that list, the *PAXBP1* gene contains the longest identified STR in a protein-coding core promoter. Indeed, this core promoter is a functional complex of multiple consecutive CT-STRs that has been exceptionally expanded in primates and not in any non-primate order [Mohammadparast et al., 2014]. Among primates, this STR reaches maximum length and complexity in *Homo sapiens*. PAXBP1 function is indispensible for the recruitment of Pax3 and Pax7 [Diao et al., 2012], which in turn, are involved in the development of normal craniofacial features [Liu et al., 2012; Murdoch et al., 2012; Paternoster et al., 2012], and spine morphogenesis [Guerreiro et al., 2013], properties that are differentially distinct in primates vs. non-primates.

In a reverse analysis, in the current study, we screen the entire *Homo sapiens* orthologous genes in mouse, based on the Ensembl database, in order to identify core promoter STRs that have been possibly directionally contracted/evolved in the process of primate evolution. We introduce several STR evolutionary switches that are specific to primates, and may be important for the emergence and evolution of this order.

METHODS

Bioinformatics

The bioinformatics analyses were performed in two steps. In step I, the entire human protein-coding orthologous genes in mouse were analyzed for the density and repeat numbers of different classes of core promoter (−120 to +1 of TSS) STRs using the Ensembl database (http://asia.ensembl.org/index.html). This interval was screened for the presence of STRs using the Microsatellite Repeats Finder at the following link: http://biophp.org/minitools/microsatellite_repeats_finder/demo.php. STRs of ≥6-repeats were then analyzed in *Homo sapiens* and *Pan troglodytes*, to detect conserved STRs across the three species. We have previously investigated the other end of the spectrum (i.e. STRs which are exceptionally long in human) [Ohadi et al., 2012]. That study covered STRs of potential interest in mouse, which were less than the exceptional length (<6-repeats), and had been expanded in human. In Step II, we analyzed another primate species sharing an ancestor with *Homo sapiens* (i.e. *Gorilla gorilla gorilla*), and four species belonging to four major non-primate orders across mammals, including Scandentia, Laurasiatheria, Afrotheria, and Xenarthra, in order to examine specificity of STR evolutionary patterns between primates and non-primates.

The list of *Homo sapiens* genes were based on the annotated genes in the GeneCards database

(http://www.genecards.org/index.shtml). The sequences were retrieved from the Ensembl database. The following species (in parentheses) were used for the major orders across Mammals, primates (*Homo sapiens*, *Pan troglodytes*, and *Gorilla gorilla gorilla*), Rodents (Mouse), Scandentia (Tree Shrew), Laurasiatheria (Dog), Afrotheria (Elephant), and Xenarthra (Armadillo). The following datasets were used for the species analyzed in this study*: Homo sapiens*: February 2009 *Homo sapiens* high coverage assembly GRCh37 (GCA_000001405.13) from the Genome Reference Consortium; *Pan troglodytes*: Ensembl Pan troglodytes Version 2.1.4 (February 2011) of the *Pan troglodytes* genome assembly (known as Pan_troglodytes-2.1.4 or CHIMP2.1.4); *Gorilla gorilla gorilla*: Assembly gorGor3.1, December 2009, Database vs. 74.31; Mouse: the Genome assembly: GRCm38 (GCA_000001635.3); Tree shrew: Assembly tupBel1, June 2006, Database vs. 74.1, Dog: Assembly CanFam3.1 (GCA_000002285.2), Database vs. 74.31; Elephant: Assembly Loxafr3.0, July 2009, Dataset version 74.3, Armadillo: Assembly Dasnov3.0, December 2011, Dataset version 74.3.

Mouse core promoters containing at least one STR of $\geq$6-repeats were compared with the corresponding sequences (i.e. -120 to $+1$) in *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla gorilla*, and the four out-group species, following sequence alignment of the corresponding transcripts, using the ClusalW2 version 4 software (http://www.ebi.ac.uk/Tools/msa/clustalw2). ClustalW2 is a general purpose multiple sequence alignment program for nucleotides or proteins. The process of alignment was performed using the default settings (the main parameters are the gap opening penalty, and the gap extension penalty) of the software.

Core promoter STRs with $\geq$6-repeats include less than 2% of the protein-coding genes [Ohadi et al., 2012], and are hereafter designated as "exceptionally long." The "STR formula" designates classes of STRs, and their repeat numbers based on the Ensembl data sets. The term "conserved" refers to the site of occurrence, and the class of STR. The term "identical" refers to the class of STR, and the number of repeats (STR formula).

The statistical analysis for the frequency of the STR motifs was based on the average frequency of those motifs at the interval between -120 and $+1$ of TSS, based on the Ensembl database, in the primates, *H. sapiens* ($n = 19{,}000$) *Gorilla gorilla* ($n = 2{,}000$), and *Pan troglodytes*,($n = 2{,}000$), where "n" represents the number of genes screened in each species.

This research adhered to the legal requirements and protocols of the University of Social Welfare and Rehabilitation Sciences, Tehran, Iran. Also, this research adhered to the American Society of Primatologists principles for the ethical treatment of primates.

RESULTS

Step I: Conserved Core Promoter STRs Across Mouse, *Pan troglodytes*, and *Homo sapiens*

Exceptionally long mouse core promoter STRs were observed in 309 genes (Suppl. 1). In 210 of those genes, STRs or the corresponding transcripts were non-existent in *Homo sapiens*. Among the remaining genes ($n = 99$) (Table I), STRs were conserved in *Homo sapiens* and mouse, of which, in 48 genes, those STRs were also conserved in *Pan troglodytes* (Table I). Di-nucleotide STRs were among the most conserved across the three species. For example, in the gene, *GLT25D2*, CA-repeats ranging from 17 to 24 repeats were conserved in mouse, *Pan troglodytes*, and *Homo sapiens*. In the gene, *RNF130*, CG-repeats ranging from 6- to 7-repeats are conserved in the three species.

Tri-nucleotide STRs were also conserved to a lesser extent, of which combinations of GCG and GCC were among the common conserved motifs. For example, in the gene, *MBD6*, repeats of GCG were conserved across the three species.

Step IIA: Identical Evolution of Certain STRs in Primates (i.e. *Homo sapiens*, *Pan troglodytes*, and *Gorilla gorilla gorilla*) vs. Out-groups Belonging to the Major Non-primate Orders Across Mammals

Within the conserved STRs among mouse, *Pan troglodytes*, and *Homo sapiens*, in 37 genes the evolutionary pattern of STRs was identical in *Pan troglodytes* and *Homo sapiens* vs. mouse (Table II). In order to examine the specificity of this evolutionary pattern, we analyzed another primate species sharing an ancestor with humans (i.e. *Gorilla gorilla gorilla*), and four species belonging to the major non-primate orders across mammals. In 21 out of the 37 genes, STRs were identical across *Homo sapiens*, *Pan troglodytes*, and *Gorilla gorilla gorilla* vs. the non-primate out-groups (shaded areas in Table II). In the remaining 16 genes, STRs were identical, specifically between *Homo sapiens* and *Pan troglodytes* (Table II). The average frequency of the identically shared STR formulas across *Homo sapiens*, *Pan troglodytes*, and *Gorilla gorilla gorilla* ranged between 0.00005 and 0.06 (Table III). The following examples are STR evolutionary events that are identical in *Homo sapiens*, *Pan troglodytes*, and *Gorilla gorilla gorilla*, and no non-primates studied. In the gene, *STAP1*, TG/23 in mouse is split to two TG-STRs ranging from 7 to 12 repeats in the three primates. In the gene, *PALMD*, TC/29 CA/4 in mouse is shrunk to TC/3 CA/4 in the three primates. In another remarkable example, in the gene, *RGS14*, GT/20 is split into several identical GT-repeats ranging from 3- to 6-repeats in the three primates. In the gene, *NOV*, the STR formula in mouse is CA/17, which has been

TABLE I. Step I: Evolutionary Analysis of the Mouse Exceptionally Long Core Promoter STRs (≥6-Repeats) in *Pan troglodytes* and *Homo sapiens*[a]

Gene symbol	Mouse transcript ID	Mouse STR formula	*Pan troglodytes* transcript ID	*Pan troglodytes* STR formula	*Homo sapiens* transcript ID	*Homo sapiens* STR formula
GABRA3	ENSMUST00000055966	ga/33 ga/3 aggg/3	ENSPTRT00000041759	tc/3 ct/3 ct/3 tc/3	ENST00000370314	ag/3 ga/3 ga/19
SLC39A8	ENSMUST00000081978	ag/30 ga/4 at/3	ENSPTRT00000051039	cg/4	ENST00000356736	cg/4
CELF2	ENSMUST00000114934	ga/30 ga/3	[b]		ENST00000379261	ga/19 ag/3 ag/5
PALMD	ENSMUST00000040097	tc/29 ac/4	ENSPTRT00000001867	ct/3 ac/4	ENST00000263174	ct/3 ac/4 tg/3 gt/3
MMP9	ENSMUST00000017881	ca/24	ENSPTRT00000025282	ca/6	ENST00000372330	ac/16
GLT25D2	ENSMUST00000044311	ac/24	ENSPTRT00000044909	ac/17	ENST00000361927	ac/19
STAP1	ENSMUST00000031171	tg/23	ENSPTRT00000029999	tg/9 tg/7	ENST00000396225	tg/9 tg/7
ATP2B3	ENSMUST00000033744	cca/22 cct/3 ctcc/3 cctc/3	[b]		ENST00000349466	gc/3 cctc/3
FAM179A	ENSMUST00000097284	ct/20 ca/3 ag/3	[c]		ENST00000379558	tc/3 tc/3
RGS14	ENSMUST00000063771	gt/20 ct/3	ENSPTRT00000032499	gt/3 gt/3 gt/4 gt/6 ct/3	ENST00000408923	gt/6 gt/3 gt/5 gt/4 ct/3
IKZF4	ENSMUST00000133342	ca/20 ct/13 ct/4 ct/4 ct/4 tc/4	[b]		ENST00000431367	tc/8 ca/3
EOS	ENSMUST00000133342	ca/20 ct/13 ct/4 ct/4 ct/4 tc/4	[b]		ENST00000431367	tc/8 ca/3
RNF150	ENSMUST00000078525	ca/20 ag/3	ENSPTRT00000030603	agc/3	ENST00000306799	ac/11 ca/3 ag/3
TUSC5	ENSMUST00000062024	ag/20 ag/11	ENSPTRT00000015714	gagg/3 gc/3 ac/10	ENST00000333813	ca/18 gagg/3
C2orf82	ENSMUST00000027476	tg/18 tg/3	[c]		ENST00000331342	gt/3
FLI1	ENSMUST00000016231	ga/18	ENSPTRT00000045786	ga/10	ENST00000429175	ga/26 ga/3 ga/3
SLC6A6	ENSMUST00000032185	gt/18 gc/4 gt/3	ENSPTRT00000027410	0	ENST00000360861	tg/3 gt/3
TIAM1	ENSMUST00000002588	ag/18 gcc/6 gga/4 ac/3	[b]		ENST00000455508	gga/3 ac/3
NOV	ENSMUST00000050027	ca/17	ENSPTRT00000037987	ca/4	ENST00000259526	ca/4
ASB9	ENSMUST00000033756	gt/16	ENSPTRT00000075583	0	ENST00000380483	gt/15 ct/4
CHRM1	ENSMUST00000163785	gt/16	ENSPTRT00000007067	tc/3	ENST00000306960	tg/3 gt/3
SCRT2	ENSMUST00000064061	tc/16 ac/12 tc/4 tc/3	ENSPTRT00000067797	ct/4 ct/3	ENST00000246104	gt/6 tg/3 ct/3 tc/3
TFAP2A	ENSMUST00000021787	ac/14 tgt/4	ENSPTRT00000011439	cct/3 cct/3 cct/5 ag/3	ENST00000379608	tccc/3 cctc/3
PHF21A	ENSMUST00000111293	ag/14 ag/5 ag/3 gag/5 gtga/3	[b]		ENST00000323180	ag/11 gga/5
C2	ENSMUST00000025230	gt/12 tc/4	ENSPTRT00000033243	ca/3 tg/3	ENST00000375510	tg/3 ac/3
CHD3	ENSMUST00000092971	gag/12 gag/6 cgc/3 tgg/3 ggcg/3	ENSPTRT00000016052	gag/12 ggcg/3 gcc/3 ggt/4	ENST00000380358	gag/13 cgc/3 tgg/3 ggcg/3
MRGPRF	ENSMUST00000033386	ctc/12 ctc/3	ENSPTRT00000007446	ga/3	ENST00000320913	cct/3 ccctc/3
ARID2	ENSMUST00000096250	cgc/12 cct/3	[b]		ENST00000334344	cct/5 ccct/5
MAN1A2	ENSMUST00000008907	gag/10 agg/3	ENSPTRT00000047368	cag/3 gga/5 gga/3	ENST00000356554	gag/8 gga/3
TRIM8	ENSMUST00000026008	ctcc/9	ENSPTRT00000005517	gga/4 gc/4	ENST00000302424	gga/4 gc/4
SCMH1	ENSMUST00000000087	gcc/9 gc/4	[b]		ENST00000337495	ccgc/3 ccgc/3
SCMH1	ENSMUST00000000087	gcc/9 gc/4	[b]		ENST00000337495	ccgc/3 ccgc/3
SELL	ENSMUST00000027871	ag/9 ga/5	ENSPTRT00000003012	ag/3	ENST00000236147	ag/3
SLIT1	ENSMUST00000025993	tg/9 gt/6 cg/5 cg/3 gt/3	[c]		ENST00000266058	gc/3 gt/5 gt/7 gt/3 gc/4 ggag/3
SPO11	ENSMUST00000050442	gt/9	ENSPTRT00000025448	gt/3	ENST00000371263	gt/3
ZCCHC5	ENSMUST00000062010	ag/9	[b]		ENST00000321110	gt/3 gt/3
MAPK7	ENSMUST00000079080	gc/9	ENSPTRT00000016351	ca/3	ENST00000395602	gc/4
CILP	ENSMUST00000048762	tgc/8 ta/3	ENSPTRT00000013235	gct/3	ENST00000261883	gct/3 ta/3
PLAG1	ENSMUST00000003369	gcg/8 ggc/3	ENSPTRT00000037547	cgg/9	ENST00000316981	cgg/9
SLC3A1	ENSMUST00000024944	ca/8	ENSPTRT00000022112	ca/3	ENST00000260649	ca/3
KIAA1543	ENSMUST00000057028	acacac/7 gcgc/3 ac/3	ENSPTRT00000019101	gt/13 gt/3 gg/3	ENST00000160298	gcg/7
VKORC1L1	ENSMUST00000051758	ggc/7 gcg/3	ENSPTRT00000035573	ggc/4 ggc/4 gcg/6	ENST00000434382	ggc/4 ggc/4 gcg/6
IKZF2	ENSMUST00000027146	agg/7 gag/4 tc/4	ENSPTRT00000023889	tcc/3 tc/4	ENST00000342002	tc/4 tcc/3
PWWP2B	ENSMUST00000093993	cgg/7 ca/3 cg/3	ENSPTRT00000057219	cgg/5 cg/3	ENST00000305233	cggct/3 cg/3 cg/3
GTF2A1	ENSMUST00000021345	ggc/7 tc/3 ccct/3	ENSPTRT00000012110	gc/3	ENST00000298173	gc/3 tc/3
MBD6	ENSMUST00000026476	gcg/7 ggc/3 cgg/4	ENSPTRT00000072863	gcg/3 cgg/7	ENST00000546632	ggc/3 gcg/3 cgg4
SBSN	ENSMUST00000080518	tcc/7	[c]		ENST00000518157	tcc/3
RFTN2	ENSMUST00000027121	tc/7 tc/6 tc/3 ct/3 tc/4	ENSPTRT00000023671	ct/10 ct/7 ag/3	ENST00000295049	tc/5 ct/6 ag/3 ct/8
B3GALNT1	ENSMUST00000061826	cg/7 gc/4	[b]		ENST00000473142	gc/4 cg/3
SFRS6	ENSMUST00000130411	gc/7 gcgcc/3	[c]		ENST00000244020	gc/3 gc/7 tg/3
SH3BP1	ENSMUST00000061239	ga/7 ga/3	ENSPTRT00000047300	cg/3 ga/3 ga/3	ENST00000336738	cg/3 ga/3 ga/3
CRYM	ENSMUST00000033198	ac/7 ac/4	ENSPTRT00000014477	gc/3 ggcgg/3	ENST00000396023	gc/3 ggcgg/3
RIMKLB	ENSMUST00000068242	cg/7 gt/4 gt/3 ga/6	[b]		ENST00000357529	gcgg/3 gc/4 gcc/3 gc/3

TABLE I. Continued

Gene symbol	Mouse transcript ID	Mouse STR formula	*Pan troglodytes* transcript ID	*Pan troglodytes* STR formula	*Homo sapiens* transcript ID	*Homo sapiens* STR formula
RAB33B	ENSMUST00000054387	tc/7 cgc/3	ENSPTRT00000030587	tc/3 ga/3	ENST00000305626	ga/3 tc/3
PTAR1	ENSMUST00000099560	gc/7	ENSPTRT00000074570	cg/6 gc/3 cg/3	ENST00000340434	gc/3 cg/9
ACVR1B	ENSMUST00000000544	gc/7 ggag/5 gc/4	ENSPTRT00000009165	gc/8	ENST00000257963	cg/3 gc/4
GFRA2	ENSMUST00000022699	ag/7 ggga/3 ga/3	ENSPTRT00000037137	ga/3 ag/4	ENST00000400782	ga/3 ga/3
TFAP2B	ENSMUST00000027059	gt/7 at/4 gcgc/3	ENSPTRT00000033774	ta/3 at/4 tg/4	ENST00000393655	tg/3 ta/3 ta/3 tg/4
CHST1	ENSMUST00000065797	gc/7	[b]		ENST00000308064	gc/4
STAG2	ENSMUST00000069619	ccctc/6 cctc/3 ct/3 ccct/4	[b]		ENST00000371144	tccc/3 cctc/3 ag/3
PLEKHA5	ENSMUST00000032357	tcgc/6 cg/3 gc/3	[c]		ENST00000299275	cg/3 gc/3 gc/3 ggc/5
SETD2	ENSMUST00000153838	ggag/6 ga/3 ga/3 gcgg/3 ccag/4 gcc/3	[b]		ENST00000409792	gcc/10 cctcc/3
IQGAP2	ENSMUST00000068603	gcgg/6 cgag/4	ENSPTRT00000031522	gc/4 ggc/3	ENST00000274364	gc/4
LIMD1	ENSMUST00000026269	ccgcc/6 ccgcc/5	ENSPTRT00000027735	ac/26	ENST00000273317	ac/33
PRKCH	ENSMUST00000021527	tgc/6	ENSPTRT0000001175 2	gcg/3	ENST00000332981	gcg/3
RAP1B	ENSMUST00000064667	gcg/6	ENSPTRT00000009547	gc/3 cg/3	ENST00000250559	gc/3
PBX2	ENSMUST00000038149	cgc/6	ENSPTRT00000033293	tc/4 ct/3 cctc/3	ENST00000453487	ct/3 ct 3 cctc/3
TTLL11	ENSMUST00000028248	ggc/6 cgg/3 gcg/4	[b]		ENST00000373776	gcg/3
CASP3	ENSMUST00000093517	tgc/6	ENSPTRT00000061788	cggg/3	ENST00000308394	cggg/3
MLL5	ENSMUST00000115128	gcc/6 gggc/3 gcc/5	ENSPTRT00000036230	gcc/7 gcc/4	ENST00000334877	gcc/7 gcc/4
MEX3C	ENSMUST00000091852	ccg/6 gc/3 gc/3	ENSPTRT00000018406	gcc/3	ENST00000406189	gcc/5 gc/3
AMIGO1	ENSMUST00000106656	cgg/6 ac/3 cg/3 cgg/3 gc/5	ENSPTRT00000001979	ac/3 cg/3 cgg/3 gc/5	ENST00000369864	cg/3 cgg/3 gc/5
ANKRD13B	ENSMUST00000092892	ccg/6 ccg/3 gc/3	[c]		ENST00000394859	ccg/4
ANKRD17	ENSMUST00000081914	gcg/6 ct/4 ccct/3 ccct/3	ENSPTRT00000030072	ccct/4 gcg/5	ENST00000358602	ccct/4 gcg/6
GOLGA4	ENSMUST00000084820	cgc/6	ENSPTRT00000046942	gc/3	ENST00000361924	gc/3
ABI2	ENSMUST00000052332	agg/6 ctc/3 gag/3	ENSPTRT00000023794	cg/3 cgc/3	ENST00000261017	gc/3
KIAA0368	ENSMUST00000102889	tcc/6	ENSPTRT00000039289	ta/3	ENST00000259335	ta/3
ZBTB2	ENSMUST00000100078	gcg/6 ct/3	ENSPTRT00000034541	ct/3	ENST00000325144	tc/3 ct/3 ccg/7
GSDMD	ENSMUST00000023238	gga/6	ENSPTRT00000038201	cg/3	ENST00000262580	cg/3
TMPRSS2	ENSMUST00000000395	ctc/6 tc/3	ENSPTRT00000025975	ccgccc/3	ENST00000497881	ccgccc/3
ATXN1	ENSMUST00000167708	agg/6 ga/4 gagg/3	ENSPTRT00000065472	0	ENST00000450222	gagc/5 ga/3 gagg/4 gggag/3 gga/7
ERI1	ENSMUST00000033927	cg/6 ct/3 ccg/3	[b]		ENST00000519292	gc/3
SRCIN1	ENSMUST00000107596	ac/6 cg/3 gc/3 ac/3 cg/3 cg/3	[c]		ASMPATCHT00000002241	cg/4 cg/5 ac/3 cg/5 ac/4 ac/4
RNF130	ENSMUST00000054684	gc/6	ENSPTRT00000032588	cg/7	ENST00000522208	cg/6 ccg/3
RBMS2	ENSMUST00000092033	ct/6 tc/3	ENSPTRT00000009384	cctc/3 tc/4	ENST00000262031	cctc/3 tc/4
PRR7	ENSMUST00000046533	gc/6	[b]		ENST00000323249	cg/5 gc/4
TRPC5	ENSMUST00000040184	ga/6	[b]		ENST00000262839	tc/12 tc/18 ac/10
TSPAN5	ENSMUST00000029800	gc/6 ag/3	ENSPTRT00000030330	gc/6	ENST00000305798	cg/5 gc/3
LY9	ENSMUST00000068878	ca/6 ac/3 ca/3 ca/3	ENSPTRT00000002849	ac/5 ca/5	ENST00000368041	ac/5 ca/5
HTRA4	ENSMUST00000084031	gt/6 at/3	ENSPTRT00000037383	gt/3	ENST00000302495	gt/3
APAF1	ENSMUST00000020157	gc/6 gcg/3	ENSPTRT00000048551	gc/3 gt/4	ENST00000333991	gc/3 gt/4
FAM168B	ENSMUST00000047534	gc/6 gc/3	[c]		ENST00000409185	gc/3 gc/3
AATF	ENSMUST00000018841	gc/6	ENSPTRT00000016671	cg/3 gc/3	ENST00000225402	cg/3 gc/3
ENOX1	ENSMUST00000022589	gc/6 gc/3 cg/3	ENSPTRT00000010726	tc/3 ct/3 gc/5	ENST00000261488	gc/5 gc/3 cg/3
UCMA	ENSMUST00000027978	tc/6	ENSPTRT00000004215	cct/3 ct/4	ENST00000378681	cct/3 ct/4
XYLT1	ENSMUST00000032892	cg/6 cct/3	ENSPTRT00000014408	ggc/6 gcg/3 agg/4 ggc/7	ENST00000261381	cgg/4
TMPRSS2	ENSMUST00000000395	ctc/6 tc/3	ENSPTRT00000025975	ccgccc/3	ENST00000497881	ccgccc/3
PDE2A	ENSMUST00000166652	ct/6 ct/4 tcc/3	ENSPTRT00000007523	tc/3 gt/3 ct/5 ct/3 ct/3 ct/4 tcc/3	ENST00000334456	ct/3 ct/5 ct/3 ct/3 ct/3 ct/4 tcc/3
NHLH1	ENSMUST00000059794	gt/6 ag/3 at/4 ca/3 gt/4	ENSPTRT00000002834	gt/7 at/4	ENST00000302101	gt/6 at/4
RNF19B	ENSMUST00000168461	cgg/5	ENSPTRT00000045611	cgg/3	ENST00000356990	cgg/3

[a]Gene order is based on the length of the STRs in mouse. The STR formula designates classes of STRs and their repeat numbers based on the Ensembl data sets (See Bioinformatics section).
[b]Transcript not found.
[c]Ortholog not found.

TABLE II. Step II: Mouse Conserved Exceptionally Long Core Promoter STRs That Have Been Evolved Identically in Primates (*Homo sapiens*, *Pan troglodytes*, and *Gorilla gorilla gorilla*), vs. Non-primate Out-groups[a,b]

	Non-primates					Primates		
Gene symbol	Mouse (*Mus musculus*)	Tree Shrew (*Tupaia belangeri*)[c]	Armadillo (*Dasypus novemcinctu*)	Elephant (*Loxodoa africana*)	Dog (*Canislupus familiaris*)	Gorilla (*Gorilla gorilla gorilla*)	Chimpanzee (*Pan troglodytes*)	Human (*Homo sapiens*)
SLC39A8	ag/30 ga/4 at/3	—	cct/3 cg/4	—	—	cg/4	cg/4	cg/4
PALMD	tc/29 ac/4	—	at/3 tat/3		—	ct/3 ac/4	ct/3 ac/4	ct/3 ac/4
STAP1	tg/23	tg/20	gt/6	gt/7	gt/6	tg/12 tg/12[d]	tg/9 tg/7	tg/9 tg/7
RGS14	gt/20 ct/3	—	ga/3	—	gt/19 ct/3	gt/3 gt/4 gt/5 gt/6 gt/3	gt/3 gt/3 gt/4 gt/6 ct/3	gt/6 gt/3 gt/5 gt/4 ct/3
NOV	ca/17	—	ct/3	—	cg/3 ca/4	ca/4	ca/4	ca/4
CHD3	gag/12 gag/6 cgc/3 tgg/3 ggcg/3	—	ct/3	ag/3 ct/3	—	cg/3 gt/3	gag/12 ggcg/3	gag/13 ggcg/3
							gcc/3 tgg/3	gcc/3 tgg/3
C2	gt/12 tc/4	—	gt/3 gt/3 tg/3	tg/6 gt/5 gt/3	—	—	ca/3 tg/3	ca/3 tg/3
TRIM8	ctcc/9	tg/3	—	—	gga/4 gc/4 ct/3 gg/3 gc/3	gga/4 gc/4	gga/4 gc/4	gga/4 gc/4
SELL	ag/9 ga/5	—	aagg/3	—	agg/3 ga/3	ag/3	ag/3	ag/3
SPO11	gt/9	—	—	—	—	gt/3	gt/3	gt/3
PLAG1	gcg/8 ggc/3	—	—	—	—	aaat/3	cgg/9	cgg/9
SLC3A1	ca/8	—	—	—	—	ca/3	ca/3	ca/3
VKORC1L1	ggc/7 gcg/3	—	tt/3	—	—	ca/3 ct/4	ggc/4 ggc/4 gcg/6	ggc/4 ggc/4 gcg/6
IKZF2	agg/7 gag/4 tc/4	—	—	ag/3	—	tc/4 tcc/3	tc/4 tcc/3	tc/4 tcc/3
SH3BP1	ga/7 ga/3	—	—	—	—	gga/3	cg/3 ga/3 ga/3	cg/3 ga/3 ga/3
CRYM	ac/7 ac/4	—	—	—	—	ggcgg/3[e]	gc/3 ggcgg/3	gc/3 ggcgg/3
RAB33B	tc/7 cgc/3	—	ga/3 tc/3	cg/3 ggc/3	—	gc/3 gc/5	ga/3 tc/3	ga/3 tc/3
GFRA2	ag/7 ggga/3 ga/3	tta/3 tgc/3	ccg/3 ccg/3	ttc/3	tc/3	ga/3 ga/3	ga/3 ga/3	ga/3 ga/3
TFAP2B	gt/7 at/4 gcgc/3	—	gt/3	—	tc/3 tg/3 ta/3	ta/3 at/4 tg/4	ta/3 at/4 tg/4	tg/3 ta/3 ta/3 tg/4
PRKCH	tgc/6	—	—	—	—	gcg/3	gcg/3	gcg/3
PBX2	cgc/6	cgc/3 cg/3	—	—	—	ctg/3 tg/3	tc/4 ct/3 cctc/3	ct/3 ct/3 cctc/3
CASP3	tgc/6	—	—	—	—	cggg/3	cggg/3	cggg/3
MLL5	gcc/6 gggc/3 gcc/5	—	at/3 aa/3	—	—	gcc/10 gccg/3	gcc/7 gcc/4	gcc/7 gcc/4
ANKRD17	gcg/6 ct/4 ccct/3 ccct/3	—	cg/3	ctcc/4	—	gt/3 tc/3	ccct/4 gcg/5	ccct/4 gcg/6
GOLGA4	cgc/6	cgccgt3	—	gc/3 gc/3	gc/3	cgc/3	gc/3	gc/3
KIAA0368	tcc/6	—	—	—	ta/3	ta/3	ta/3	ta/3
GSDMD	gga/6	—	—	—	—	cg/3	cg/3	cg/3
TMPRSS2	ctc/6 tc/3	—	—	—	—	ccgccc/4	ccgccc/3	ccgccc/3
RBMS2	ct/6 tc/3	ctcc/3 tccc/3 ct/3	ct/3 ctcc/4	ctc/3	ttc/3	ct/3 tc/9 tc/9	cctc/3 tc/4	cctc/3 tc/4
LY9	ca/6 ac/3 ca/3 ca/3	gt/3	—	—	ca/5 cacg/4 ca/3 ac/4	ac/5 ca/5	ac/5 ca/5	ac/5 ca/5
HTRA4	gt/6 at/3		tc/3	—	—	gt/3	gt/3	gt/3
APAF1	gc/6 gcg/3	ta/3 ga/3	gc/4 cg/5	ta/3	ga/4	tg/3	gc/3 gt/4	gc/3 gt/4
AATF	gc/6	—	—	cg/3 ca/3 cgcgca/3	cg/3 cgcgca/3	—	cg/3 gc/3	cg/3 gc/3
UCMA	tc/6	—	—		cctc/3	cct/3 ct/4	cct/3 ct/4	cct/3 ct/4
PDE2A	ct/6 ct/4 tcc/3	—	cg/3 ggcg/4	—	—	—	tc/3 gt/3 ct/5	ct/3 ct/5 ct/3
							ct/3 ct/3 ct/4 tcc/3	ct/3 ct/3 ct/4 tcc/3
NHLH1	gt/6 ag/3 at/4 ca/3 gt/4	—	cca/3	ct/3	—	gt/6 at/4	gt/7 at/4	gt/6 at/4
RNF19B	cgg/5	—	—	—	—	—	cgg/3	cgg/3

[a]Gene order is based on the length of the STRs in mouse. The STR formula designates classes of STRs and their repeat numbers based on the Ensembl data sets (See Bioinformatics section).
[b]Shaded areas represent identical STR formulas.
[c]Tree shrew represents the basal insectivore lineage whose ancestors originally gave rise to primates.
[d]This formula is shaded to mark the identical split of a long TG-STR in mouse into two TG-STRs in primates.
[e]An exceedingly rare STR, ggcgg/3, with the frequency of 0.0001, emerges in *Gorilla gorilla gorilla*, and is evolved to GGCGG/3 GC/3, in *Homo sapiens* and *Pan troglodytes*.

TABLE III. Frequency of the STR Motifs Identically Evolved in Three Primates; *Homo sapiens*, *Pan troglodytes*, and *Gorilla gorilla gorilla*

Gene symbol	STR formula shared by primates					STR motif frequency[a]					Gene ontology
SLC39A8	cg/4					0.005					Metal ion transmembrane transporter activity
PALMD	ct/3	ac/4				0.01	0.001				Regulation of cell shape
STAP1	tg/9	tg/7				0.0001	0.0005				Signal transduction
RGS14	gt/6	gt/3	gt/5	gt/4	ct/3	0.0005	0.01	0.002	0.005	0.01	Learning and long-term memory
NOV	ca/4	0.002									Regulation of cell growth and gene expression
CHD3	gag/13	cgc/3	tgg/3	ggcg/3		0.00005	0.0001	0.001	0.0001		Regulation of transcription
C2	tg/3	ac/3				0.01	0.005				Regulation of complement activation
TRIM8	gga/4	gc/4				0.001	0.02				Protein homodimerization activity
SELL	ag/3					0.01					Regulation of immune response
SPO11	gt/3					0.01					Spermatid development and female gamete generation
PLAG1	cgg/9					0.0001					Regulation of transcription
SLC3A1	ca/3					0.01					Basic amino acid transport
VKORC1L1	ggc/4	ggc/4	gcg/6			0.003	0.003	0.0002			Integral to membrane
IKZF2	tc/4	tcc/3				0.009	0.006				Positive regulation of transcription from RNA polymerase II promoter
SH3BP1	cg/3	ga/3	ga/3			0.06	0.009	0.009			Signal transduction
CRYM	ggcgg/3	gc/3				0.0001	0.06				Negative regulation of transcription from RNA polymerase II promoter
RAB33B	ga/3	tc/3				0.009	0.01				Autophagy
GFRA2	ga/3	ga/3				0.009	0.009				Nervous system development
TFAP2B	tg/3	ta/3	ta/3	tg/4		0.01	0.006	0.006	0.005		Skin development
PRKCH	gcg/3					0.0003					Positive regulation of glial cell proliferation
PBX2	ct/3	ct/3	cctc/3			0.01	0.01	0.0002			Embryonic limb morphogenesis
CASP3	cggg/3					0.0002					Heart development
MLL5	gcc/7	gcc/4				0.0005	0.005				DNA methylation
ANKRD17	ccct/4	gcg/6				0.0001	0.0001				Regulation of smooth muscle cell differentiation
GOLGA4	gc/3					0.06					Protein targeting to Golgi
KIAA0368	ta/3					0.006					May play a role in ERAD and other enhanced proteolysis
GSDMD	cg/3					0.06					Cellular response to extracellular stimulus
TMPRSS2	ccgccc/3					0.00005					Proteolysis
RBMS2	cctc/3	tc/4				0.0002	0.009				RNA processing
LY9	ac/5	ca/5				0.0003	0.001				Cell adhesion
HTRA4	gt/3					0.01					Endopeptidase activity
APAF1	gc/3	gt/4				0.06	0.005				Nervous system development, forebrain development
AATF	cg/3	gc/3				0.06	0.06				Regulation of mitotic cell cycle
UCMA	cct/3	ct/4				0.006	0.009				Regulation of osteoblast differentiation
PDE2A	ct/3	ct/5	ct/3	ct/3	ct/3	0.01	0.002	0.01	0.01	0.01	Negative regulation of transcription from RNA polymerase II promoter
	ct/4	tcc/3				0.01	0.009				
NHLH1	gt/6	at/4				0.002	0.001				Regulation of transcription
RNF19B	cgg/3					0.01					Involved in the cytolytic activity of natural killer cells and cytotoxic T-cells

[a]The average frequency of STR motifs in the −120 to +1 TSS is presented for each motif. If the motif has been repeated more than once in a STR formula, the frequency has been re-stated as well.

shrunk to CA/4 in *Homo sapiens*, *Pan troglodytes*, and *Gorilla gorilla gorilla*. In the gene, *TRIM8*, CTCC/9 is evolved to GGA/4 GC/4 in the three primates. In the gene, *SELL*, AG/9 GA/5, in mouse is shrunk to GA/3. Shrinkage of GT/9 in the gene, *SPO11*, to GT/3 in the three primates is another example of directional shrinkage of core promoter STRs. In the gene, *SLC3A1*, CA/8 in mouse is contracted to CA/3. In the gene, *IKZF2*, combinatory AGG/7 GAG/4 TC/4 is evolved to TCC/3 TC/4. In the gene, *Ly9*, the following formula in mouse, CA/6 CA/3 CA/3 CA/3, is evolved to AC/5 CA/5 in the three primates. In another remarkable observation, in the gene, *NHLH1*, the following STR formula in mouse, GT/6 AG/3 AT/4 CA/3 GT/4, is contracted to GT/6 AT/4 in the three primates. In the gene, *TMPRSS2*, the following formula in mouse, CTC/6, is evolved to CCGCCC/3. Of note, the latter evolutionary event might have risen by point mutations rather than slippage mispairing, within the CTCCTC motif in mouse, including a T- to C- and C- to G-, which gives rise to the motif, CCGCCC.

Step IIB: Identical Evolution of Certain STRs in *Homo sapiens* and *Pan troglodytes* vs. *Gorilla gorilla gorilla*, and the Out-groups Belonging to the Major Non-primate Orders Across Mammals

Within the genes in which core promoter STRs had been differentially evolved in primates vs. non-primates, we observed differential evolution of certain STRs in *Homo sapiens* and *Pan troglodytes* vs. *Gorilla gorilla gorilla* and non-primates (Table II). For example, in the gene, *CHD3*, the STR formula for multiple STRs is identical in *Homo sapiens* and *Pan troglodytes*. In the gene, *PLAG1*, the following STR formula in mouse, GCG/8 GGC/3, is evolved to GGC/9 in *Homo sapiens* and *Pan troglodytes*, only. In the gene, *SH3BP1*, GA/7 GA/3 is replaced by CG/3 GA/3 GA/3 in *Homo sapiens* and *Pan troglodytes*. The core promoter of the *AATF* gene contains a STR of CG/6 in mouse, whereas this STR is split into GC/3 GC/3, specifically in *Pan troglodytes* and *Homo sapiens*. In the gene, *MLL5*, the following formula in mouse, GCC/6 GGGC/3 GCC/5, is evolved to GCC/7 GCC/4 in *Homo sapiens* and *Pan troglodytes*. In the gene, *RNF19B*, GGC/6 in mouse is shrunk to GGC/3 in both *Homo sapiens* and *Pan troglodytes*. In the gene, *ANKRD17*, the following formula: GCG/6 CT/4 CCCT/3 CCCT/3 is evolved to CCCT/4 GCG/4, specifically in *Pan troglodytes* and *Homo sapiens*. In the gene, *VKORC1L1*, GGC/7 GCG/3 in mouse is evolved to GGC/4 GGC/4 GCG/6 specifically in *Homo sapiens* and *Pan troglodytes*.

In the gene, *CRYM*, an exceedingly rare STR, GGCGG/3, with the frequency of 0.0001, emerges in *Gorilla gorilla gorilla*, and is evolved to GGCGG/3 GC/3, in *Homo sapiens* and *Pan troglodytes*.

A human-specific STR formula was observed in the gene, *TFAP2B*, in which an extra GT/3 was observed in *Homo sapiens*, and not in *Pan troglodytes*, and *Gorilla gorilla gorilla*. This specificity occurred on a background of conserved STRs in the three primates.

DISCUSSION

Alteration in the levels of gene expression is considered to be a mechanism for phenotypic differences across species [Brawand et al., 2011; Esmaeilzadeh-Gharehdaghi et al., 2011; Mohammadparast et al., 2014; Wang & Rekaya, 2009]. We have recently reported core promoter STRs that are expanded in *Homo sapiens*, and proposed that interspecies variation in gene expression levels may, at least in part, be attributed to the core promoter STRs [Ohadi et al., 2012]. This notion is further supported by the functional role of core promoter STRs to modulate gene expression [Heidari et al., 2012; Li et al., 2012; Valipour et al., 2013]. In a reverse study, here we provide results of mouse orthologous gene core promoters that contain exceptionally long STRs, and have been conserved across evolution, shrinkage of which may be linked with the primate evolution.

In eukaryotes, the "core promoter" interval can vary from one gene to the other in respect with structure and length [Darvish et al., 2011; Moshonov et al., 2008; Sandelin et al., 2007; Vinces et al., 2009]. This interval contains the basic and crucial transcription factor (TF) binding sites for constitutive gene expression. For TATA-containing promoters (24% of the protein-coding genes) [Yang et al., 2007], this region normally includes -40 to $+1$ of TSS. For TATA-less promoters, this interval can be extended [Darvish et al., 2011; Singh et al., 2012]. As a pilot study, the interval -120 to $+1$ is chosen in the current study, which needs to be extended in the future studies.

The proximity of the core promoter STRs to the $+1$ TSS may be an evolutionary constrain on those STRs to evolve to a certain length, which may be species-specific. The idea of "exceptional expansion" of the STRs ($<2\%$ of the genes) proposes that selective/adaptive expansion of certain STRs may be linked with speciation in certain species, and not the others. STRs that have reached that length may be of more potential importance for a preliminary research of this kind. Along that notion, shrinkage of those STRs may also be linked with speciation. Data from the mouse genome have confirmed the abundance of STRs but have also revealed impressive differences with human [Mouse Genome Sequencing Consortium, 2002]. If identical search criteria are used, the mouse genome proves to be repeat-rich with two- to three-fold more STRs than human. Moreover, STRs are longer in mice than in humans.

Our current results provide core promoter STRs that are exceptionally long in mouse, and have been shrunk in primates. Surveys of variability of homologous STR loci among species may be subject to ascertainment bias for STR length where STR loci isolated in one species tend to be longer than homologous loci in related species [Vowles & Amos, 2006]. A life cycle of creation and degeneration may be considered for STRs, which undergo neutral evolution [Buschiazzo & Gemmel, 2006]. Two protocols were used in the current study to minimize ascertainment bias. Firstly, two primates sharing closer ancestors (i.e. chimpanzee and gorilla) were included in the study. Secondly, four species encompassing major orders across mammals were also included as out-group. In 37 genes, we observed identical evolution of core promoter STR formulas in primates vs. out-groups. The STR motifs identically shared were of the frequencies ranging from 0.00005 to 0.06. In 16 of the 37 genes, STRs were specifically identical between *Homo sapiens* and *Pan troglodytes*, pointing to the possible role of those STRs in the emergence of a common ancestor shared by the two species. In the remaining 21 genes, the STR formula was also identical in *Gorilla gorilla gorilla*, indicating a possible role for those STRs in the emergence of a more distant ancestor, which could have led to the common ancestor of great apes. "Specific" and "identical" co-occurrence of those STR formulas (rare frequencies of <0.01 for the majority of the shared motifs) (Table III) in three primates whose ancestors have diverged four to eight million years ago, supports an evolutionary role for those STRs as "switch codes" for primate speciation, and excludes the notion of ascertainment bias. In other words, this is the first support of identity-by-descent for STRs at the interspecies level.

Remarkably, in the gene, *TFAP2B*, an extra GT/3 was observed in *Homo sapiens*, and not in *Pan troglodytes* and *Gorilla gorilla gorilla*. This specificity occurred on a background of identical STRs in the three primates. Remarkably, mutations in the *TFAP2B* gene result in craniofacial malformations in human [Milunsky et al., 2008].

We observed three exceptions in the differential evolution of the STRs from Step II in primates vs. non-primates. The first and second exceptions were observed in the genes, *RAB33B*, and *RBMS2*, in which combination of STRs was also conserved in one non-primate species, Armadillo, and the third exception was in the gene, *KIAA0368*, in which TA/3 was conserved in one non-primate species, Dog. Gene Ontology for the genes listed (Table III) revealed important evolutionary and developmental functions for those genes, such as regulation of cell shape (*PALMD*), learning and long-term memory (*RGS14*), nervous system development (*GFRA2*), embryonic limb morphogenesis (*PBX2*), and forebrain development (*APAF1*).

Beyond a certain length, STRs become the binding site for TFs. For example, a minimum of 3-repeats is required for the CA-, GC-, AT-, CT, GCG, and GCC-repeats to become binding sites for TFs, where further repeats add additional TFs (http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite).

We recognize that our discovery of the instances of possible functional changes does not constitute proof, but rather provide direction for further study of the importance of the listed STRs in primate evolution, development, and disease.

REFERENCES

Bolton KA, Ross JP, Grice DM, et al. 2013. STaRRRT: a table of short tandem repeats in regulatory regions of the human genome. BMC Genomics 14:795.

Borel C, Migliavacca E, Letourneau A, et al. 2012. Tandem repeat sequence variation as causative Cis-eQTLs for protein-coding gene expression variation: the case of CSTB. Human Mutation 33:1302–1309.

Brawand D, Soumillon M, Necsulea A., et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343–348.

Buschiazzo E, Gemmel NJ. 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays 28:1040–1050.

Darvish H, Heidari A, Hosseinkhani S, et al. 2013. Biased homozygous haplotypes across the human caveolin 1 upstream purine complex in Parkinson's disease. Journal of Molecular Neuroscience 51:389–393.

Darvish H, Nabi MO, Firouzabadi SG, et al. 2011. Exceptional human core promoter nucleotide compositions. Gene 475:79–86.

Diao Y, Guo X, Li Y, et al. 2012. Pax3/7BP is a Pax7- and Pax3-binding protein that regulates the proliferation of muscle precursor cells by an epigenetic mechanism. Cell Stem Cell 11:231–241.

Donaldson ZR, Kondrashov FA, Putnam A, et al. 2008. Evolution of a behavior-linked microsatellite-containing element in the 5' flanking region of the primate AVPR1A gene. BMC Evolutionary Biology 8:180.

Esmaeilzadeh-Gharehdaghi E, Banan M, Farashi S, et al. 2011. Support for down-tuning of the calreticulin gene in the process of human evolution. Progress in Neuropsychopharmacology and Biological Psychiatry 35:1770–1773.

Fondon JW III, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. Proceedings of the National Academy of Science United States of America 101:18058.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annual Review of Genetics 44:445–477. Review.

Guerreiro I, Nunes A, Woltering JM, et al. 2013. Role of a polymorphism in a Hox/Pax-responsive enhancer in the evolution of the vertebrate spine. Proceedings of the National Academy of Science United States of America 110:10682–10686.

Hammock EAD, Young LJ. 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308:1630–1634.

Heidari A, Nariman Saleh Fam Z, Esmaeilzadeh-Gharehdaghi E, et al. 2012. Core promoter STRs: novel mechanism for inter-individual variation in gene expression in humans. Gene 492:195–198.

Iglesias AR, Kindlund E, Tammi M, Wadelius C. 2004. Some microsatellites may act as novel polymorphic cis-regulatory

elements through transcription factor binding. Gene 341:149–165.
Jansen A, Gemayel R, Verstrepen KJ. 2012. Unstable microsatellite repeats facilitate rapid evolution of coding and regulatory sequences. Genome Dynamics 7:108–125. Review.
Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. Trends in Genetics 22:253–259.
King DG. 2012. Evolution of simple sequence repeats as mutable sites. Advances in Experimental Medical Biology 769:10–25.
King DG, Trifonov EN, Kashi Y. 2006. Chapter 4. Tuning knobs in the genome: evolution of simple sequence repeats by indirect selection In: Caporale LH, editor. The implicit genome. UK: Oxford Press.
Li Y, Seidel K, Marschall P, et al. 2012. A polymorphic microsatellite repeat within the ECE-1c promoter is involved in transcriptional start site determination, human evolution, and Alzheimer's disease. Journal of Neuroscience 32:16807–16820.
Liu F, van der Lijn F, Schurmann C, Zhu G, Mallar Chakravarty M. 2012. A genome-wide association study identifies five loci influencing facial morphology in Europeans. PLoS ONE 8:e1002932.
Milunsky JM, Maher TA, Zhao G, et al. 2008. TFAP2A mutations result in branchio-oculo-facial syndrome. American Journal of Human Genetics 82:1171–1177.
Mohammadparast S, Bayat H, Biglarian A, Ohadi M. 2014. Exceptional expansion and conservation of the CT-STR complex in the core promoter of PAXBP1 in primates. American Journal of Primatology doi: 10.1002/ajp.22266.
Moshonov S, Elfakess R, Golan-Mashiach M, Sinvani H, Dikstein R. 2008. Links between core promoter and basic gene features influence gene expression. BMC Genomics 9:92.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
Murdoch M, DelConte C, García-Castro MI. 2012. Pax7 lineage contributions to the mammalian neural crest. PLoS ONE 7: e41089.
Ohadi M, Mohammadparast S, Darvish H. 2012. Evolutionary trend of exceptionally long human core promoter short tandem repeats. Gene 507:61–67.
Paternoster L, Zhurov AI, Toma AM, Kemp JP, Pourcain BS. 2012. Genome-wide association study of three-dimensional facial morphology identifies a variant in Pax3 associated with nasion position. American Journal of Human Genetics 90:478–485.
Rockman MV, Wray GA. 2002. Abundant raw material for cis-regulatory evolution in humans. Molecular Biology and Evolution 19:1991–2004.
Sandelin A, Carninci P, Lenhard B, et al. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies Nature Reviews Genetics 8:424–436.
Sawaya SM, Lennon D, Buschiazzo E, Gemmell N, Minin VN. 2012. Measuring microsatellite conservation in mammalian evolution with a phylogenetic birth-death model. Genome Biology and Evolution 4:636–647.
Singh DP, Bhargavan B, Chhunchha B, et al. 2012. Transcriptional protein Sp1 regulates LEDGF transcription by directly interacting with its cis-elements in GC-rich region of TATA-less gene promoter. PLoS ONE 7:e37012.
Trivedi S, Hancock JM. 2012. Low microsatellite frequencies in neuron and brain expressed microRNAs. Gene 508: 73–77.
Valipour E, Kowsari A, Bayat H, et al. 2013. Polymorphic core promoter GA-repeats alter gene expression of the early embryonic developmental genes. Gene 531:175–179.
Verstrepen KJ, Jansen A, Lewitter F, Fink GR. 2005. Intragenic tandem repeats generate functional variability. Nature Genetics 37:986.
Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. Science 324:1213–1216.
Vowles EJ, Amos W. 2006. Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. Molecular Biology Evolution 23:598–607.
Wang Y, Rekaya R. 2009. A comprehensive analysis of gene expression evolution between humans and mice. Evolutionary Bioinformatics Online 5:81–90.
Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E. 2007. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene 389:52–65.
Zarif Yeganeh M, Ghaffarpour M, Farhud DD, et al. 2009. Skew in the human caveolin 1 gene upstream purine complex homozygote haplotype compartment in multiple sclerosis. Journal of Neuroimmunology 216:103–107.
Zarif Yeganeh M, Mirabzadeh A, Khorram Khorshid HR, et al. 2010. Novel extreme homozygote haplotypes at the human caveolin 1 gene upstream purine complex in sporadic Alzheimer's disease. American Journal of Medical Genetics B Neuropsychiatric Genetics 153B:347–349.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.