Genomics 94 (2009) 117–124

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

# Contrast features of CpG islands in the promoter and other regions in the dog genome

Leng Han [a,b,c], Zhongming Zhao [a,d,e,*]

[a] Department of Psychiatry and Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA 23298, USA
[b] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China
[c] Graduate School, Chinese Academy of Sciences, Beijing 100039, China
[d] Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA 23298, USA
[e] Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA

ARTICLE INFO

*Article history:*
Received 13 November 2008
Accepted 23 April 2009
Available online 3 May 2009

*Keywords:*
CpG islands
Dog
Promoter
Homologous genes
Domestication
Gene Ontology
Genome evolution
Essential genes
Housekeeping genes

ABSTRACT

The recent release of the domestic dog genome provides us with an ideal opportunity to investigate dog-specific genomic features. In this study, we performed a systematic analysis of CpG islands (CGIs), which are often considered gene markers, in the dog genome. Relative to the human and mouse genomes, the dog genome has a remarkably large number of CGIs and high CGI density, which is contributed by its noncoding sequences. Surprisingly, the dog genome has fewer CGIs associated with the promoter regions of genes than the human or the mouse. Further examination of functional features of dog–human–mouse homologous genes suggests that the dog might have undergone a faster erosion rate of promoter-associated CGIs than the human or mouse. Some genetic or genomic factors such as local recombination rate and karyotype may be related to the unique dog CGI features.



## Introduction

The dog has long been a subject of scientific curiosity because of its great diversity in both morphological (e.g., size, shape, coat color and texture) and behavioral traits [1,2]. Although the dog genome is largely similar to the human genome [3,4], it has much greater variance among its individual breeds [5]. This unique position in the mammalian phylogeny makes the dog genome suitable for evolutionary and comparative genomics studies [6,7]. Moreover, the dog represents an important model organism because it has a large catalog of disease syndromes that are more similar to the human than any other laboratory or domestic species [8,9]. Because of these important features, sequencing the dog genome (*Canis familiaris*) has been a high priority and its genome was recently completed [7]. This provides us an unprecedented opportunity to examine dog-specific features at the genome-wide level and compare it to other model genomes such as the human and mouse. As an example, a comparative genomics study suggested the euchromatic portion of the dog genome being ~18% smaller than the human genome and 6% smaller than the mouse genome, which could be explained by a lower rate of repeat insertions rather than a higher rate of deletions in the dog genome [8].

With more than twenty mammalian genomes having been sequenced thus far, a fundamental question is how the genomes have changed and what genetic factors have impacted sequence composition, size, function and complexity during the course of evolution. For instance, CpG dinucleotides are largely under-represented in most mammalian genomes, occurring only ~20–25% of their expected frequency overall [10–12]. This deficit of CpG dinucleotides is largely attributed to the high rate of deamination of methylated CpGs, which in turn accounts for approximately 80% of the total CpGs in mammalian genomes [13,14]. Conversely, CpG islands (CGIs), which are clusters of CpGs in GC-rich regions, have nearly the expected frequency of CpGs [12]. CGIs are frequently located in the 5′ region of the genes and are considered as gene markers [15,16]. Recent genome-wide investigation revealed that promoter-associated CGIs overall remained unmethylated [17], although a sizable fraction of them might be fully methylated in normal cells [17–20]. Methylation changes in promoter-associated CGIs have been found to cause transcriptional silencing and disruption of gene function [21]. In particular, many recent studies revealed that aberrant hypermethylation in promoter-associated CGIs of tumor suppressor genes may cause tumorigenesis [22]. Although CGIs have been used to estimate the number of genes in a genome [23,24], our recent study revealed large variation on the number of CGIs and their density in mammalian genomes with comparable gene number [12]. Interestingly, the dog

* Corresponding author. Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, PO Box 980126, Richmond, VA 23298-0126, USA. Fax: +1 804 828 1471.
*E-mail address:* zzhao@vcu.edu (Z. Zhao).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.04.007

genome had the largest number of CGIs and the highest CGI density among the ten mammalian genomes we studied. The number of dog CGIs was nearly 3 times that in rodent genomes [12]. It has been commonly thought that rodents might have underwent a stronger process of CpG erosion to TpGs/CpAs by *de novo* methylation and that rodent CGIs had weaker selective constraint than humans [21,23,25]. However, it remains largely unknown whether the dog genome has a relative gain of CGIs to other mammalian genomes during evolution or it has still been under similar process of erosion.

To better understand the genome features of the dog and their relationship with the morphological and behavior traits, we performed a systematic investigation of CGIs in the dog genome. We examined the CGIs and their distribution in different genomic regions including promoter, 3′-, genic, intronic and intergenic regions and further compared them with those in the human and mouse genomes. To understand the functional implications of CGIs, we examined promoter-associated CGIs in the genes with different expression level (e.g., housekeeping versus tissue specific genes) or functional importance (e.g., essential genes). We also examined the functional bias of genes that have likely lost CGIs in the dog lineage. This study provides detailed information of CGIs and their functional features in the dog genome and has important implications for mammalian genome evolution and gene function.

## Results

### *Distribution and features of CGIs in the dog genome*

We used Takai and Jones' algorithm [26] to identify CGIs in the dog, human and mouse genomes (see Materials and methods). Here, we first describe the distribution and features of CGIs in the dog genome. There were 58,327 CGIs in the dog genome, with an average length of 1102 bp, average GC content of 62.2%, and average $Obs_{CpG}/Exp_{CpG}$ ratio of 0.753 (Table 1). Here, $Obs_{CpG}/Exp_{CpG}$ ratio was measured by the ratio of the observed CpG dinucleotides over the expected CpGs in a sequence [16]. These dog CGIs had a total length of 64.3 Mb and accounted for 2.8% of the dog genome sequence. On average, we observed 25.2 CGIs/Mb in the dog genome; however, the standard deviation was high (±40.5 CGIs/Mb). When we examined CGIs in the non-repeat portion of the dog genome, we still found 53,102 CGIs, which accounted for 3.7% of the non-repeat portion of the dog genome (Table 1). This finding supports the assertion that Takai and Jones's algorithm can effectively exclude the short repeats, especially *Alu* repeats [26]. Correspondingly, CGI density in the non-repeat portion of the dog genome (37.9/Mb) is much higher than that (25.2/Mb) of the whole genome.

We further examined the distribution and features of CGIs on each dog chromosome. The results are shown in Table S1. The number of CGIs and CGI density varied greatly. Chromosome 1, the largest autosome, had the largest number of CGIs (3636) while chromosome 32 had the smallest number of CGIs (342). Moreover, the highest CGI density was found on chromosome 28 (42.2 CGIs/Mb), which was 4.8 times the lowest CGI density found on chromosome 32 (8.8 CGIs/Mb). As expected, we observed a trend that larger chromosomes had more CGIs (linear regression, $r=0.76$, $P=1.2\times10^{-8}$, Fig. S1A). The number of CGIs in a chromosome was significantly correlated with the number of genes in the chromosome ($r=0.86$, $P=1.9\times10^{-12}$, Fig. S1B), supporting the notion that CGIs can function as gene markers. Moreover, CGI density in a dog chromosome was highly correlated with genomic factors such as GC content ($r=0.82$, $P=6.4\times10^{-11}$, Fig. 1A) and gene density ($r=0.63$, $P=8.0\times10^{-6}$, Fig. 1B), indicating that CGIs depend on both local genomic features and gene number.

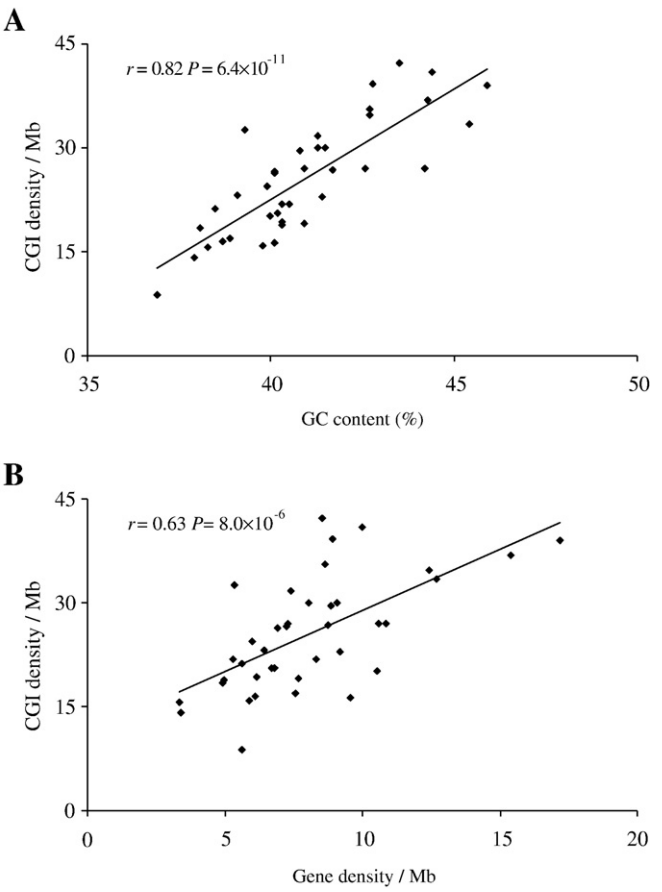


**Fig. 1.** Correlation between CGI density and genomic features on each chromosome in the dog genome. (A) CGI density versus GC content (%). (B) CGI density versus gene density (/Mb).

**Table 1**
Overview of CGIs in the whole genome and non-repeat regions.

| Species | Number of CGIs | Length (bp) | GC content (%) | $Obs_{CpG}/Exp_{CpG}$ | CGIs covered (Mb) | Genome size (Gb) | CGI density/ Mb (S.D.) |
|---|---|---|---|---|---|---|---|
| Whole genome | | | | | | | |
| Dog | 58,327 | 1102 | 62.2 | 0.753 | 64.3 (2.8%[a]) | 2.31 | 25.2 ± 40.5 |
| Human | 37,729 | 1090 | 62.0 | 0.743 | 41.1 (1.4%) | 2.85 | 13.2 ± 16.8 |
| Mouse | 21,326 | 1044 | 60.9 | 0.752 | 22.2 (0.9%) | 2.61 | 8.2 ± 8.4 |
| Non-repeat region | | | | | | | |
| Dog | 53,102 | 975 | 59.0 | 0.791 | 51.8 (3.7%) | 1.40 | 37.9 |
| Human | 28,380 | 1098 | 59.4 | 0.798 | 31.1 (2.0%) | 1.52 | 18.7 |
| Mouse | 17,109 | 1048 | 58.7 | 0.789 | 18.0 (1.2%) | 1.52 | 11.3 |

[a] Proportion of the total length of CGIs in the whole genome sequence.

### *Comparison of CGIs in the dog, human and mouse genomes*

The characteristics of CGIs in the dog genome were consistently stronger than those in the human and mouse genomes including average length (dog: 1102 bp; human: 1090 bp; and mouse: 1044 bp), average GC content (dog: 62.2%; human: 62.0%; and mouse: 60.9%) and average $Obs_{CpG}/Exp_{CpG}$ ratio (dog: 0.753; human: 0.743; and mouse: 0.752). Dog CGIs covered a larger portion (2.8%) of the dog genome than human and mouse CGIs (human: 1.4% and mouse: 0.9%) (Table 1). Interestingly, when we compared CGIs in the non-repeat portion of the three genomes, the characteristics (length, GC content and $Obs_{CpG}/Exp_{CpG}$ ratio) of dog CGIs became weaker than the corresponding ones of human CGIs (Table 1), even though the extent of decreasing the number of dog CGIs (58,327 to 53,102, 9.0%) is weaker than that of human CGIs (37,729 to 28,380, 24.8%) or mouse CGIs (21,326 to 17,109, 19.8%). Our further analysis indicated that short repeats such as SINEs may be more likely to be part of or more closely linked to the CGIs identified in the whole dog genome than in the