Distinct turn-over patterns of common repeats correlate with genome size differences among cattle, dog and human

Genome size differences among cattle, dog and human

George E. Liu
Bovine Functional Genomics Laboratory, ANRI, ARS, USDA,
Building 200, Room 124B, BARC-East, 10300 Baltimore Ave,
Beltsville, Maryland 20705, USA
George.Liu@ars.usda.gov

Abstract: Optimal three-way global sequence alignments for 84 cattle clones or loci (total 11 Mb of high-quality finished genomic sequence, each larger than 50 kb) were constructed using the human and dog genome assemblies. Although unique portions of genomic sequence remained relatively constant among these three mammals, the overall size differences in cattle-dog, human-cattle, and human-dog comparisons were 10.6%, 6.2%, and 16.8% respectively, which strongly correlated with the difference between repetitive portions in dog (28.9%), cattle (39.5%) and human (45.6%). These alignments were therefore examined for the pattern, frequency, and nature of common repeats and their contribution to the genome size. This comparison indicated that distinct turn-over patterns of lineage-specific (young) or ancestral (old) repeats may account for a repeat-driven genome size change in cattle, dog and human. The smaller size of the cattle sequence relative to human is primarily due to less ancestral repeats, indicating a larger loss of them in cattle. The larger size of the cattle sequence as compared to dog is mainly due to additional lineage specific repeat sequences in cattle, suggesting a higher insertion rate and longer lineage-specific repeats in cattle. Finally, both insertions of lineage-specific repeats and retentions of ancestral repeats contributed to the larger size of the human sequence as compared to dog. Assuming that the sampled genome sequence is representative, these differences may lead to 6-16% differences of dog, cattle, human genome size, with majorities of them due to insertions and deletions of common repeats.

Keywords: common repeats, mammalian evolution, genome size

I. INTRODUCTION

It has been complicated to elucidate the mechanisms of evolution of genome size [1], although it is generally accepted that the variation in genome size is closely correlated with the amount of common repeats [2,3]. Major types of common repeats in mammals include long interspersed elements (LINEs), short interspersed elements (SINEs), long-terminal repeat (LTR) retrotransposons, DNA transposons, and others like satellites and simple repeats [4]. For example, over 46% of the human genome (2.88 Gb) is made up of common repeats with two prominent contributors: SINE/Alu (1.1×10^6 copies) and LINE/L1 (5×10^6 copies), accounting for 11% and 17% of the total sequence, respectively [5]. However, the overall activities of common repeats have been markedly declining in the hominid lineage especially for DNA transposons and LTR retrotransposons. The other published primate genomes (chimpanzee 2.82 Gb and rhesus macaque 2.87 Gb) are generally similar to the human genome with small differences in the activities of SINE/Alu, LINE/L1, and retroviral elements [6,7]. About 40% of the rodent genome sequence (mouse 2.55 Gb and rat 2.75 Gb) is identified as common repeats, which is lower than human [8,9]. Rodents have actually accumulated more young repeats than human. The apparent deficit of common repeats in rodents is not due to a higher deletion rate, but instead mostly due to a higher nucleotide substitution rate, which makes it difficult to recognize ancient repeats [8]. LINE/L1 is the most active repeats in rodents, accounting for over 25% of the rat genome. The higher accumulation of rat specific L1 copies could explain some of the size difference of the rat and mouse genomes. About 34% of the dog genome (2.33 Gb) is recognized as common repeats [10]. Its smaller size is primarily due to less lineage-specific repeats are present in dog (334 Mb) than in human (609 Mb) or mouse (954 Mb). This reflects “a lower activity of endogenous retroviral and DNA transposons, as well as the fact that the SINE element in dog is smaller than in human” [10]. Over 52% of the opossum genome (3.48 Gb) is recognized as common repeats, which is the highest in all sequenced amniotes [11]. While the total unique sequence is similar in opossum and human, LINEs are the primary reason for the relatively large opossum genome size. About 50% of the platypus genome (2.3 Gb) consists of common repeats, which is also among the highest in all sequenced metazoans [12]. In platypus, the most abundant repeats are ancient LINE/L2 (1.9 ×10^6 copies) and SINE/MIR (Mon-1, 2.75 ×10^6 copies) while DNA transposons and LTR retroelements are quite rare. On the other hand, only 11% of the chicken genome (1.05 Gb) consists of common repeats, which is markedly lower than mammals and is mainly due to low recent common repeat activity [13]. In the chicken genome, it is interesting that
LINE/CR1 comprises over 80% of all common repeats without paring with any single SINE.

Despite the progress in the sequencing and analysis of genomes, only limited interspecies comparisons of common repeats have been performed within mammals (Human-Mouse-Rat and Human-Mouse-Dog) [8,10]. Recent sequencing effort produced a draft cattle genome assembly [14]. Similar to primates, the cattle genome (2.73 Gb) consists of ~48% of common repeats, but with many young ruminant-specific repeats such as LINE/BovB and SINE/BovA and SINE/Art2A [14]. In this current study, intraspecies comparison is applied to the cattle-dog-human alignments to assess the turn over patterns of common repeats and their contributions to the genome sizes.

II. MATERIALS AND METHODS

Optimal alignments were generated and validated as previously described [15] and they are freely accessible at http://bfgl.anri.barc.usda.gov/divergence/. Common repeats were detected using the slow search option (-s) of RepeatMasker (version 2002/07/13) with Repbase (http://www.girinst.org/, version 9.04) [4]. To rule out that cattle and dog common repeats were not efficiently masked, the insertions/deletions (indels) which were not captured by RepeatMasker were further examined. None of these indels displayed any grouping indicating that the default Repbase consensus library is sufficiently robust to identify common repeats in other mammals.

Shared ancestral repeats (AR) and lineage-specific repeats (LS) were defined as described [8,10] and determined using RepeatMasker’s DateRepeats procedure, which considers the known common repeat phylogenies based on their established classes as reported previously [4]. Permutation tests were performed at the level of the alignment as well as at the level of individual insertion/deletion events as previously described [16].

III. RESULTS

To characterize the insertions and deletions of sequence elements in the human, dog and cattle, all the nucleotides were categorized using the alignment data and RepeatMasker annotations of the insertions of repetitive elements (Fig. 1 and Table 1). Even though the length of unique portions of genomic sequence remained relatively constant (6.0-6.1 Mb), the proportions of common repeats recognizable in dog (28.9%), cattle (39.5%) and human (45.6%) were strongly correlated with the overall size differences in cattle-dog (10.6%), human-cattle (6.2%), and human-dog comparisons (16.8%), respectively. The smaller sizes of the cattle and dog genomes are primarily due to the presence of substantially less repeat sequence in cattle (4.4 Mb), dog (3.2 Mb) than in human (5.1 Mb).

The repetitive sequence was further divided into two fractions, ancestral repeats (AR) and lineage specific repeats (LS), to assess the source of these differences. Differential turn-over patterns of AR and LS repeats were found to account for the overall size differences among cattle, dog and human genomes (Fig. 2 and Table 2). The smaller size of the cattle genome relative to human is primarily due to less ancestral repeat sequences, indicating a larger loss in ancestral LINE-AR and LTR-AR repeats in cattle (Table 2). The total lengths of lineage-specific repeats were approximately the same in cattle and human, even though these elements differ in their types, frequencies and lengths.

Figure 1. Mammalian Genome Size Variation. For orthologous genomic comparisons, the length of aligned sequence and difference were considered for each comparison. Repetitive and unique portions of aligned orthologous sequences were identified by RepeatMasker. Relative percentages were calculated assuming the length of the human genome as 100%. Significance of the difference in genome size was determined by a permutation test (10,000 replicates). Asterisks over species bars indicate significant differences in overall lengths, and those between species bars indicate significant differences in either repetitive or unique lengths between two species. **, P < 0.01. Detailed results are tabulated in Table 1.

Table 1. Mammalian Genome Size Variation

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Repeats</th>
<th>Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Length (bp)</td>
<td>%</td>
<td>Length (bp)</td>
</tr>
<tr>
<td>Cattle</td>
<td>10,504,883</td>
<td>94.53</td>
<td>4,387,512</td>
</tr>
<tr>
<td>Dog</td>
<td>9,322,051</td>
<td>83.88</td>
<td>3,209,044</td>
</tr>
<tr>
<td>Human</td>
<td>11,113,231</td>
<td>100.00</td>
<td>5,069,169</td>
</tr>
<tr>
<td>Difference</td>
<td>1,182,382</td>
<td>10.64**</td>
<td>1,178,468</td>
</tr>
</tbody>
</table>

Table 2. Mammalian Genome Size Variation

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Repeats</th>
<th>Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Length (bp)</td>
<td>%</td>
<td>Length (bp)</td>
</tr>
<tr>
<td>Cattle</td>
<td>10,504,883</td>
<td>94.53</td>
<td>4,387,512</td>
</tr>
<tr>
<td>Dog</td>
<td>9,322,051</td>
<td>83.88</td>
<td>3,209,044</td>
</tr>
<tr>
<td>Human</td>
<td>11,113,231</td>
<td>100.00</td>
<td>5,069,169</td>
</tr>
<tr>
<td>Difference</td>
<td>1,182,382</td>
<td>10.64**</td>
<td>1,178,468</td>
</tr>
</tbody>
</table>
Cattle has more LINE-LS repeats (from its LINE/BovB elements) while human has more SINE-LS (SINE/Alu) and LTR-LS repeats. The larger size of the cattle genome as compared to dog is mainly due to additional lineage specific repeat sequences in cattle, suggesting a higher insertion rate and longer lineage-specific repeats in cattle, especially for LINE-LS (LINE/BovB) and SINE-LS (SINE/BovA and SINE/Art2A) repeats. The deletion of ancestral repeats was slightly higher in cattle than in dog; however, the contribution was minimal. Finally, both lineage-specific repeat sequences and ancestral repeat sequences contributed to the larger size of the human genome as compared to dog. This difference reflects a higher insertion activity, longer insertion elements, and lower deletion activity of ancestral repeats in human, agreeing with the earlier results [10].

This comparison of cattle, dog and human genomic sequence shows the dramatic expansions (6%–16%) in genome size (Fig. 3). In a previous study based on a sampling of DNA content, it was reported that the haploid DNA contents (i.e. c values) of these three species were significantly different ([17], cattle: 3.15 pg; dog: 2.8 pg; and human 3.5 pg, also see http://www.genomesize.com/). It was unknown, however, whether such differences were attributable to centric chromatin, which is known to be cytogenetically reduced in size (Martin 1990). This analysis of cattle, god and human data indicates that the difference is in fact euchromatic in nature and that it is almost exclusively repeat-derived (Fig. 3). All classes of younger retrotransposons (SINE, LINE, and LTR) contribute to this increase, as well as more ancient elements like DNA transposons also contribute to this increase by differential deletion. This effect was seen in both human and baboon compared to lemur [16]. Assuming a divergence of these mammals approximately 80-90 million years ago, the data would support a major increase in genome size due to the differential activities of common repeats.

In summary, rates of insertion/deletion (SINE, LINE, LTR and DNA) vary radically among different mammalian lineages. Differential turn-over patterns of lineage-specific and/or ancestral transposable elements might account for the overall size differences among cattle, dog and human genomes. Among cattle, dog and human, genome sizes were strongly correlated with the insertion and deletion patterns of transposable elements. These differences have lead to 6-16% differences of dog, cattle, human genome size, with majorities of them due to insertions and deletions of common repeats.
IV. DISCUSSION

The evolutionary mechanisms shaping genome-size include common repeats, spontaneous deletions and insertions, large rearrangements, copy number variation, and segmental duplications and many others (such as heterochromatin shrinkage and expansion). Depending on whether they affect the phenotype, the resultant genome size changes might undergo natural selection or neutral drift. The combined interplay of all these forces will finally determine the genome size. However, it has been difficult to distinguish among them and to study them individually, given the multitude of genetic and population processes involved [1].

Comparative genomics is a powerful strategy to provide valuable insights into evolutionary forces that have shaped the genome sizes. The current analysis indicates that different turn over patterns of common repeats account for the smaller genome sizes in cattle and dog as compared to human. Some of the above differences in the nature of common repeats in mammals could reflect systematic factors in their biology, whereas others may represent random fluctuations. For example, due to the higher substitution rates, common repeats in rodents might degrade to the point of being unrecognizable, thus making up a sizeable portion of unique genome sequence.

Common repeats have been extensively studied to elucidate their evolution and impacts on the genome. These include phylogenetic and comparative sequencing studies to determine their lineages and expansions. With about a dozen of mammal genome assemblies are available for comparison, it is already clear that the pattern, frequency, and nature of common repeats vary considerably from one species to another.

Besides the genome size change, common repeats can strongly influence genome architecture over evolutionary time. For example, nonallelic homologous recombination (NAHR) between two copies of common repeat family can lead to duplications, deletions, inversion and genomic rearrangements.

It is worthwhile to note that heterochromatin is well represented in eukaryotic genomes. For example, about 15% of the human genome is made up of heterochromatin (John, 1988). However, genome sequences are rarely ‘complete’ in centromeric and telomeric heterochromatin, because of the difficulties inherent in sequencing repetitive DNA. Recent studies also demonstrated that the genome size and the proportion of common repeats can vary within one species; therefore even sequencing projects can only provide partial information [2].

Because common repeats have been shown to be potent mutagens, as contributors to novel genome structure and organization, and as sources of new genic and regulatory diversity, it follows that they contribute significantly to phenotypic change and evolution. Deeper understanding of the biology of common repeats and detailed knowledge of common repeats populations in mammals should provide more insights into their evolutionary impacts.

ACKNOWLEDGMENT

The author thanks G. Wei and A. Parker for their technical support. This work was supported in part by NRI grant No. 2007-35205-17869 from USDA-CSREE and by Project 1265-31000-098-00 from USDA-ARS. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

REFERENCES