DNA Palindromes in Human Genome

Helen Li[1], Aman Gupta[2], BS, Madhavi K. Ganapathiraju, Ph.D[1,2]
[1]Department of Biomedical Informatics, University of Pittsburgh
[2]Language Technologies Institute, Carnegie Mellon University

Summary: *A palindrome in DNA is a sequence whose one half is complement of the other half but appearing in reverse order (for example: ATCC-GGAT). Palindromic sequences can influence DNA stability and are known to be associated with diseases. Here, we present all DNA palindromes in human genome that are equal to or longer than 8 bp. There are more than 39 thousand palindromes that are longer than 50 bp.*

Introduction and Background: Palindromic sequences are believed to be distributed throughout the human genome in varying lengths and locations. Palindromes are linked to genetic disorders, such as mental retardation, X-linked recessive diseases, and many physical abnormalities caused by mutagenesis. The prevalence and function of palindromes in the human genome is not fully understood. We present our findings on the distribution of palindromes in human genome, which reveals orders of magnitude of palindromes compared to a previous study.

Methods: We had previously developed Biological Language Modeling Toolkit (v. 2) that constructs suffix array, and longest common prefix array to efficiently identify all palindromic sequences in the human genome. We employed the toolkit on UCSC Genome Browser human genome Build 38 (GRCh38/hg38) and Build 19 (GRCh37/hg19), and analyzed all palindromes and near palindromes 8 bps and longer, allowing up to four mismatches. Evidence suggests that palindrome distribution is non-uniform in gene functional regions. We defined our functional regions: exons, introns, upstream as the region -2000 from the transcription start site, intergenic region as the noncoding area between genes, 5' UTR as the region between gene transcription start site and coding region start site, 3' UTR as gene coding region end and transcription start end, and promoter as the region -200 bp from transcription start site. We computed counts in each of these regions in each chromosome.

Results and Discussion: Analysis of the palindromic lengths revealed that palindromes of length 16 are the most frequent across all chromosomes (Figure 1A). The number of palindromes was proportional to the chromosome length as well as to gene density. We found very high palindrome counts in the intronic, upstream and intergenic regions, which may be linked to transcription and replication processes and formation of functionally relevant secondary structures in pre-mRNA. In total, we found 32,976,219 palindromes in human genome Build 38 (GRCh38/hg38), with 74.99% of all palindromes being AT rich. We also found 39,501 palindromes of length greater than 50 bp, which is significantly greater than other palindrome finders that only discovered 3,500 palindromes longer than 50 bp. These longer palindromes were also predominantly AT rich (Figure 1B). High AT content in palindromes, which create palindromic AT rich repeats (PATRR), are correlated to DNA instability, breakage, and chromosomal translocation (Kato et al. (2006)). Our results are available for view as a track on the UCSC Human Genome Browser human genome Build 38 (GRCh38/hg38). We are currently analyzing genome sequences from TCGA data to understand the role of palindromes in structural alterations in DNA, as there is evidence to show that long palindromes occur frequently in human cancer cell lines in medulloblastoma (Tanaka et al. (2006)).

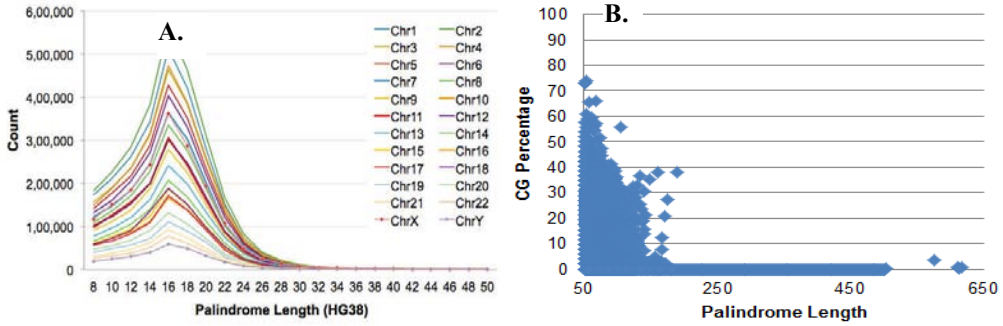

Figure 1 (A): Counts versus lengths of palindromes in each chromosome. **(B)** CG percentage versus

References

Guenthoer, J., *et al.* (2012) Assessment of palindromes as platforms for DNA amplification in breast cancer, *Genome research*, **22**, 232-245.
Tanaka, H., *et al.* (2006) Large DNA palindromes as a common form of structural chromosome aberrations in human cancers, *Human cell*, **19**, 17-23.
Kato T, Franconi CP, Sheridan MB, Hacker AM, Inagakai H, et al. (2014) Analysis of the t(3;8) of hereditary renal cell carcinoma: a palindrome-mediated translocation. Cancer Genet 207: 133-140