P18

Identification of conserved and polymorphic STRs for personal genomes.

Chien-Ming Chen 1, Chi-Pong Sio 1, Yu-Lun Lu 1, Hao-Teng Chang 2, Chin-Hwa Hu 3, Tun-Wen Pai 1

1 Department of Computer Science and Engineering, National Taiwan Ocean University, Taiwan

2 Graduate Institute of Molecular System Biomedicine, China Medical University, Taiwan

3 Institute of Biosciences and Biotechnology, National Taiwan Ocean University, Taiwan

Abstract

Background

Short tandem repeats (STRs) are abundant in human genomes. Numerous STRs have been experimentally proved their associations with genetic diseases, gene regulatory functions, and selected as genetic markers for evolutionary and forensic analyses. High-throughput next generation sequencers foster new cutting-edge computing techniques for genome-scale analyses, and cross-genome comparisons facilitate efficient identification of polymorphic STR markers for various applications.

Results

An automatic and efficient system for detecting human polymorphic STRs at genome-scales was proposed in this study. Assembled contigs from NGS data were aligned and calibrated according to selected reference sequences. To verify identified polymorphic STRs, human genomes from the 1000 Genomes Project were employed for comprehensive analyses, and STR markers from Combined DNA Index System (CODIS) and disease-related STR motifs were also applied as cases for evaluation. In addition, we analyzed STR variations for highly conserved homologous genes and human-unique genes. In total 477 polymorphic STRs were identified from 492 human-unique genes, among which 26 STRs were especially retrieved and clustered into three different levels for efficient comparison.

Conclusions

We have developed an on-line system which could efficiently identify polymorphic STRs and provide novel distinguishable STR biomarkers for different levels of specificity. Candidate polymorphic STRs within a personal genome could be easily retrieved and compared to the constructed STR profile through query keywords, gene names or assembled contigs.