Original human genome might have had 25% to 35% methylated C in CpG

Fernanda Stussi, Carlos Alberto Xavier Gonçalves, Lissur Azevedo Orsine, TETSU SAKAMOTO, J. Miguel Ortega

UFMG

Abstract

Some authors suggested an effect of neighbor bases on the probability of SNPs occurrence. We built a graphical online database SNP LocAL Neighborhood and computationally over deaminate every CpG to TpG, supposing that C was methylated and increased the bias or artificially aminate fractions of TpG to simulate reversion to CpG with methylated C. Aiming to investigate comprehensively this event, we built an online database to show the pattern of bases in SNP neighborhood, available at: http://bioinfo.icb.ufmg.br/snplane/. SNP LANE comprises SNPs in Mus musculus, Homo Sapiens, Bos taurus, Rattus Norvegicus and Sus scrofa, localized in intron, CDS, 5'UTR or 3'UTR and classified by substitution type: K, M, R, Y, W or S. For each SNP class, nucleotide frequencies were calculated for the first five positions upstream and downstream surrounding the SNP. Expected baseline nucleotide frequencies for positions neighboring the SNP were estimated by randomly choosing positions in the genome and retrieving nucleotides flanking it. Two graphics are presented for each of 1200 distinct situation. In the majority of cases baseline frequency was not significantly different from observed data, indicating that the observed neighboring effect was not an influence on the mutation, but rather if T or A are more frequent downstream of C, it would seem C might be influencing the transition T/A but baseline frequency indicates that this is just an effect of non-randomness of the genome. When we deaminated all remaining C in CpG, was a small increase in bias. Simulating different percentages of amination of "CpA" and "TpG" back to CpG dinucleotides was noteworthy that bias is completely erased with 25% to 35% of amination. We do not see the neighboring nucleotide effect on these conditions. R and Y substitutions did not respond to amination, probably because amination already causes R and Y. It is suggested that dinucleotide composition produces the previously reported neighborhood bias on SNP probability. Most of this effect might be explained by deamination of C in CpG and we suggest that originally human genome would have 25% to 35% of the present CpA and TpG in the form of CpG.

Funding: Fapemig, Rede Biologia Sistêmica do Câncer