Human/chimpanzee trans-specific SNPs: searching for balancing selection

ID - 224

Fumagalli Matteo[1], Pozzoli Uberto[1], Comi Giacomo P.[2], Sironi Manuela[1]

[1]Scientific Institute IRCCS E. Medea, Bioinformatic Lab, Via don L. Monza 20, Bosisio Parini (LC), Italy

[2]Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation , 20100 Milan, Italy

Motivation

Most species are monophiletic throughout most of their genome. Yet, examples of trans-specific polymorphisms have been reported in different species, including humans. Trans-specific SNPs (ts-SNPs) can be explained by three main reasons: 1) SNP survival due to random chance, 2) coincidental mutations occurred after speciation and 3) balancing selection. In particular, the role of balancing selection on human/chimpanzee coding sequence evolution has been estimated to be modest (Asthana et al., 2005). While few genome-wide attempts have been performed to detect signatures of balancing selection outside coding sequences, recent data on specific loci have indicated the role of balancing selection in the evolution of cis-acting regulators of genes involved in immune response.

The availability of extensive human and chimpanzee SNP data, as well as of genome-wide measures of nucleotide diversity and SNP allele frequency now allow the identification of regions under balancing selection.

Methods

Human and chimpanzee SNPs were retrieved from dbSNP database (build 125); SNPs positions refer to the following genomic assemblies: NCBI build 35 and NCBI build 1 version 1, respectively. Only base substitution polymorphisms were included and SNPs located at CpG sites in either species were discarded. A total of 8637604 and 1163289 SNPs were obtained for human and chimpanzee, respectively. Whole genome human-chimpanzee pairwise alignments (available through the UCSC Genome Browser, www.genome.ucsc.edu) were scanned in order to identify SNPs located at the same position and showing the same alleles in both species. Genomic annotations, including Multi-species conserved sequences (MCS), as well as data concerning human recombination rates, were retrieved from the UCSC database (tables phastCons17wayElements and snpRecombRateHapmap). Tajima's D values were obtained from the UCSC database for three populations with different descent: African, European and Chinese.

Tajima's D (Tajima, 1989) is one of the most frequently used tests to compare nucleotide diversity: regions with an excess of high-frequency variation (observed as a positive Tajima's D) are consistent with balancing selection. Still, it should be noted that Tajima's D has low power to detect selection, so that even small departures from expected (in case of neutrality) might indicate balancing selection.

All statistical analyses were performed using R (www.r-project.org).

Results

We identified a total of 1411 non-CpG ts-SNPs. Since it has been previously estimated that a shared polymorphism would survive for 4.6 million years (conservatively, the time separating human for chimpanzee), based on the number of available chimpanzee SNPs, we would expect to identify only 3 surviving SNPs. In order to verify whether the number of ts-SNPs is higher than expected from coincidental mutations we calculated the expected number of ts-SNPs on the basis of a random distribution of human/chimpanzee polymorphic sites and verified that the number of SNPs we identified is significantly higher (Binomial Test, $p < 0.0001$). Therefore a subset of ts-SNPs might be maintained by balancing selection.

ts-SNP distribution was as follows: 5 in coding regions, 46 in conserved noncoding sequences, the remaining being in introns or intergenic spacers. The frequency of ts-SNPs in conserved noncoding regions did not differ from expected.

We next wished to identify closer than expected ts-SNPs pairs: based on the distance distribution of 100 samples of randomly selected SNPs (and an empirical p value of 0.01) we determined a conservative threshold of 10 kb. We identified 24 ts-SNPs doublets or trios closer than the threshold as possibly located in regions subjected to balancing selection. Indeed, one of the doublets was located in the MHCII region (in addition to 4 more ts-SNPs), previously shown to be under balancing selection.

We next wished to verify whether Tajima's D is higher in regions surrounding ts-SNPs (or a proportion of them). Mean Tajima's D values in 10 kb surrounding ts-SNPs were significantly higher than the genome average for the 3 populations.

Also, ts-SNPs were located in regions displaying a significantly (Wilcoxon Rank Sum Test, $p<0.0001$)

higher than average polymorphism density; although a high SNP frequency might result in increased probability of retrieving a ts-SNP, an increased nucleotide diversity is expected as a result of balancing selection. The high SNP frequency surrounding ts-SNPs is not the result of strong recombination activity since recombination rates around ts-SNPs do not differ from the average (calculated on random SNP samples).
We therefore consider that some of the ts-SNPs we identified might represent a molecular mark of balancing selection and we are thus applying further tests in order to define genomic regions and specific functional elements subjected to such selective process.
Email: matteo.fumagalli@bp.lnf.it