Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms

Alexej Abyzov, Shantao Li, Daniel Rhee Kim, Marghoob Mohiyuddin, Adrian Stuetz, Nicholas F. Parrish, Xinmeng Jasmine Mu, Wyatt Clark, Arif Harmanci, Ken Chen, Matthew Hurles, Jan Korbel, Hugo Y.K. Lam, Charles Lee, Mark Gerstein

Abstract

Structural variants (SVs) affect more bases in the human genome than those arising from point mutations (i.e. SNPs). While entire SVs can be thousands of bases in length, only a few bases around their breakpoints hold the most crucial information about their mechanisms of formation, which mostly fall into three broad classes: non-allelic homologous recombination (NAHR), transposable element insertion, and non-homologous (NH) mechanisms. Here we identify, classify, and analyze 8,943 breakpoints associated with deletions in 1,092 samples sequenced by the 1000 Genomes Project. We make this dataset available as a resource, representing the largest high-quality breakpoint collection to date. We found that breakpoints have more nearby SNPs and indels (on megabase scale) than the genomic average. This effect is correlated with reduced cross-species conservation, suggesting it is a consequence of relaxed selection on breakpoint regions. By investigating nucleotide substitution patterns of nearby SNPs and analyzing the correlating of breakpoints with methylation data, Hi-C interactions, and histone marks, we find that NAHR breakpoints are associated with open chromatin. We hypothesize that since replication is largely devoid of chromatin structure, some of the NAHR deletions occur without DNA replication and cell division, in embryonic and germline cells, and then are passed on to the offspring. We do not find any associated correlations with NH breakpoints; however, we do find that they are often coupled with micro-insertions. We identified template sites for over one hundred of such insertions. Surprisingly, these are located at two characteristic distances from the breakpoint and tend to be replicated later. These findings are consistent with a template switching mechanism and suggest particular spatial and temporal configurations for DNA during the switching.