Mathematical modelling and Bayesian inference reveals new insights into structural variation of the human genome.

Mae Woods[1] and Chris Barnes[1]
[1] University College London

Structural variation in the human genome, in the form of deletion, insertion, inversion and translocation, occurs in both germline and somatic cells, and is observed frequently in cancer genomes. Recent accumulation in whole genome sequencing on paired tumor and non-tumor samples and the development of algorithms to estimate absolute copy number profiles, have led to a rich structural dataset from which we can further current inferences on the mutational landscape of the human genome.

It is known that the choice of repair pathway assigned to mend breaks in DNA affects the probability of structural variation. Hence, because these mutations are a direct consequence of the interplay between DNA damage and repair, the dependencies between the numbers of insertions and deletions observed might result in a deeper insight into activity of the DNA repair machinery, a group of processes that play a fundamental role in evolution.

We present a Bayesian approach to understanding the probability landscape of structural variation using approximately 2000 cancer genomes, comprising of 14 primary sites. First we hypothesize that similarities between cancers, which may be implicit by robustness of biological networks, could potentially be identifiable by analysis of structural measures. We show by statistical analysis on the data that there is a clear universal optimum in how much a chromosome can deviate from the size reported in the human reference genome. We find that after adjustment for confounding there are nontrivial dependencies between the probability of insertions, deletions and also translocations on a chromosome. This leads us to predict using a mathematical model that these highly constrained distributions on the length of a chromosome are the product of a dynamical system involving not only deletions and insertions, but also translocations. We show that this system appears to be common amongst cancer types, highlighting the importance of the information gained when multivariate correlations observed in big data are combined with the predictive power of theoretical models.

Keywords: Structural variation, Mutation processes, Evolution, Genetics.