A model for scoring damaging mutations in the non-coding tumoral genome based on germline and tumor data

Jia Li[1], Marie-Anne Poursat[2], Stefan Michiels[3], Daniel Gautheret[1]

1Institute for Integrative Biology of the Cell , 2 Laboratoire de Mathématiques, Université Paris-Sud, Orsay France 3 Institut de Cancérologie Gustave Roussy, Villejuif, France

Daniel Gautheret[1]

daniel.gautheret@u-psud.fr

Cancer driver mutations are somatic events that promote tumor growth or metastasis. Previous computational studies have largely focused on driver mutations located in protein-coding exons that change amino acid residues with damaging effects. However, non-coding RNA (ncRNA) genes and non-coding parts of coding genes (introns, UTRs) now emerge as significant players in the regulation of gene expression and potentially in tumor progression. There is an urgent need for methods that can evaluate the effect of somatic mutations in such non-coding regions and prioritize mutations for further scrutiny. Here we develop two random forest models for predicting germline and somatic mutation constraints in any non-coding region. These models combine functional features from Encode and other genome surveys, using as response variables the mutational constraints provided by the 1000 Genome Project (germline model) and by collections of tumor whole genome sequences (somatic model). We show that each model reflects a different set of constraints acting on the normal and tumor genome and we identify the specific features (such as conserved elements and histone marks) that most contribute to these constraints. Furthermore, high scoring regions defined by each model are enriched in known disease-related mutations, indicating we can use the resulting scores as a proxy for damaging non-coding mutation. We combine both model to predict regions in ncRNAs and introns/UTRs of protein coding genes where mutations are most likely to be damaging. This system paves the way for the detection of non-coding driver genes and regulatory elements in cancer.