Understanding methylation of transposable elements in A.Thaliana via interpretable machine learning

Ekaterina Antonenko*[1,2,3], Marie Dogo[2,3,4], Jeremy Cohen[2,3,4], Louna De Oliveira[5], Aurélien Petit[5], Vincent Colot[5], Chloé-Agathe Azencott[2,3,6], and Pierre Baduel[5]

[1]Centre de Bioinformatique – Mines Paris - PSL (École nationale supérieure des mines de Paris) – France

[2]Institut Curie – PSL - University – France

[3]Oncologie Computationnelle (U1331) – Institut National de la Santé et de la Recherche Médicale - INSERM – France

[4]Centre de Bioinformatique – Mines Paris - PSL (École nationale supérieure des mines de Paris) – France

[5]Institut de Biologie de l'École Normale Supérieure (IBENS) – ENS, CNRS UMR8197, Inserm U1024, Paris – France

[6]Centre de Bioinformatique – Mines Paris - PSL (École nationale supérieure des mines de Paris) – France

Abstract

Transposable elements (TEs) are present ubiquitously in genomes, and their accumulation is responsible for most of the large variations in genome size seen among eukaryotes. When mobilized, they can self-propagate through genomes, either by cut-and-paste or by copy-and-paste mechanisms. The mobilization of TEs, however, is a rare event, as it is inherently highly mutagenic. A variety of mechanisms may repress transposition, including DNA methylation as well as mutation accumulation. In plants, methylation is possible at CG, CHG, and CHH sites (where H=A, T, or C), and the gain of methylation typically prevents TEs from mobilizing.

We perform Genome-Wide Association Studies (GWAS) on Transposable Elements in Arabidopsis Thaliana, and observe that the TE presence or absence, as well as their methylation status, indeed has a strong potential to impact the expression of the nearby genes. To further understand the mechanisms underlying these associations, we model the spreading of methylation from a TE to flanking regions with interpretable machine-learning tools, such as Random Forests. Our computational results show that the CHG and CHH methylation pathways are particularly responsible for the spreading effect, together with the insertion frequency of TEs, which points to specific methylation machineries such as RNA-directed DNA methylation (RdDM). This hypothesis is further confirmed with biological experiments, and this finding illustrates the power of our approach in inferring a real epigenetic mechanism from a machine-learning model.

*Speaker

Keywords: GWAS, Transposable Elements, Epigenetics, Methylation, Machine learning, Explainability