Exploring genome-wide association between chromatin structure and gene expression: a Mixture of scalar-on-function regression coupled to random forest classification

Mathilde Bruguet*[1], Nadia Ponts†[2], David Causeur‡[3,4], and Gaël Le Trionnaire§[5]

[1]Institut Agro , Irmar UMR 6625 CNRS – L'Institut Agro Agrocampus Ouest – France
[2]Unité de recherche Mycologie et Sécurité des Aliments (MycSA) – Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – Villenave d'Ornon, France, France
[3]Agrocampus Ouest, Irmar UMR 6625 CNRS, 65 rue de Saint-Brieuc, CS 84215, 35042 Rennes cedex – Agrocampus Ouest, IRMAR – UMR CNRS 6625 – France
[4]Institut de Recherche Mathématique de Rennes (IRMAR) – Agrocampus Ouest – Campus de Beaulieu, bâtiments 22 et 23, 263 avenue du Général Leclerc, CS 7420535042 RENNES Cédex, France
[5]INRA UMR 1349, Institut de Génétique, Environnement et Protection des Plantes, Le Rheu, France – Institut National de la Recherche Agronomique : UMR1349 – France

Abstract

Epigenetic mechanisms play a crucial role in regulating gene expression by modifying chromatin structure in response to environmental changes. These modifications are highly diverse and act through a variety of mechanisms, often interacting simultaneously to regulate stress response without altering the underlying DNA sequence. However, our understanding of these mechanisms is incomplete, and the relative importance of each in regulating specific processes is poorly understood. In wheat crops, *Fusarium graminearum* - a major fungal pathogen - relies strongly on such mechanisms to modulate gene expression.
To explore the relationship between the first level of chromatin structure and transcriptional activity in *F. graminearum*, we generated high-throughput MAINE-Seq data capturing genome-wide nucleosome positioning, together with RNA-seq data for transcriptomic profiling. Our objective was to predict gene expression levels from chromatin accessibility signals using scalar-on-function regression and to develop a biologically interpretable model describing these associations. A comparative analysis of classical statistical approaches revealed limitations in capturing the functional relationships between nucleosome positioning and gene expression. To overcome these limitations, we proposed a mixture scalar-on-function regression model that accommodates heterogeneous associations across the genome.

Unlike conventional studies that rely on predefined promoter regions or signal aggregations, our approach treats each gene as a spatial functional entity, where chromatin accessibility is measured continuously along its nucleotide sequence. This framework captures the intrinsic

*Speaker
†Corresponding author: nadia.ponts@inrae.fr
‡Corresponding author: david.causeur@agrocampus-ouest.fr
§Corresponding author: gael.le-trionnaire@inrae.fr

structure of genomic signals and allows functional modeling of their regulatory effects. Prediction accuracy was improved by incorporating a classification step in which prior latent component probabilities were updated to posterior probabilities estimated through Random Forests trained on chromatin accessibility features. This hybrid approach outperformed standard machine learning algorithms and a convolutional neural network in prediction performances. The identified latent components revealed distinct, biologically meaningful patterns of association between chromatin accessibility and gene expression, offering new insights into the epigenetic regulation of pathogenicity in *F. graminearum*.

Keywords: Mixture model, functionnal data, epigenetics, scalar on function regression