Title: Enformer Prediction Performance Evaluation Leveraging Deep Learning to Interpret Genetic Variation in Prostate Cancer

Authors:

Waseeq Ahmad Minhas

Keywords:

Prostate Cancer, DNA, Sequence-Based, Genetic Variation, Deep Learning, Enformer Model, Genes, Mann-Whitney U Test

Abstract

Prostate cancer (PCa) is the most commonly diagnosed disease and the main cause of cancer-related deaths among men across the globe. One of the main challenges in prostate cancer biology is that the disease is genetically diverse, with genetic variants playing a key role in how it develops and progresses. This variation complicates the ability to distinguish between indolent (slow-growing) and aggressive tumors due to limited understanding of the processes responsible for tumor development. Despite significant advances in genome-wide association studies (GWAS), interpreting the roles of coding and non-coding variants in gene regulation remains a major obstacle in functional genomics. Most variants are located in non-coding regions, which make up 98% of the genome, and traditional methods for studying them are time-consuming and require extensive laboratory testing. This highlights the urgent need for computational tools that can rapidly prioritize variants based on predicted functional effects. Deep learning models like Enformer offer potential for DNA sequence-based predictions regarding the effects of variants on gene expression and chromatin states. This study aimed to assess the capability of the Enformer model to predict the functional impacts of genetic variants linked to prostate cancer. The model was used to generate numerical scores reflecting the magnitude of impact on gene expression. Variants were categorized into coding and non-coding regions and stratified by their link to known prostate cancer genes (52 commonly known genes selected for this study) versus other genes (all remaining genes in the VCF files not included in the selected prostate gene list). To investigate how effectively the model predicted gene expression scores, the “CAGE: Prostate Epithelial Cells” cell line was chosen. The Mann-Whitney U test was used to compare predicted expression scores between these groups, revealing no significant difference ($p = 0.98$ for coding, $p = 0.15$ for non-coding). However, functional enrichment analysis including GO terms, disease relevance, and pathway-level insights performed using Enrichr, ShinyGO, and g:Profiler on high-scoring genes from the “other genes” group revealed significant biological associations with prostate cancer. This supports the potential of Enformer to uncover previously unrecognized genetic contributors to prostate cancer biology. Training with prostate-specific data is needed to improve its predictive capabilities and enhance its application in precision oncology.