CHAPTER 3

ASdeep uses a deep-learning neural network to attribute allelic expression to non-coding genetic variants

Zhenhua Zhang, Herbert T. Kruitbosch, Morris A. Swertz, and K. Joeri van der Velde
Abstract

Summary Many studies have shown that non-coding variants (NCV) affect phenotypes via regulation. Estimation of allelic difference is thus a novel way to identify underlying mechanisms through the potential regulating programmes. However, existing variant annotation tools do not incorporate allelic differences well. We therefore developed ASdeep, a tool that attributes allelic differences such as allele-specific expressions to NCV that are potentially involved in the regulation.

Availability and Implementation PyPi (https://pypi.org/project/asdeep) and GitHub (https://github.com/zhenhua-zhang/asdeep).

Introduction

Diploid organisms inherit two sets of genetic information from their parents. This introduces allelic differences that can consequently result in distinct behaviors between alleles like allele-specific expression (ASE). In ASE, the transcripts from the maternal and paternal alleles have different abundances [1]. In the past few decades, researchers have demonstrated important roles for ASE in many human phenotypes and diseases, including aging [2], tumor development [3, 4] and Parkinson’s disease [5]. Apart from well-known mechanisms such as chromatin openness and random monoallelic expression, studies have also suggested that the differential regulation caused by non-coding variants (NCV) interferes with the expression balance of allelic transcripts [6, 7, 8, 9]. Linking NCV to ASE can thus provide new insights into the regulatory mechanisms of human traits such as diseases. However, existing tools are limited to a small number of NCV due to the computational capacity of the algorithms implemented. Here, we introduce ASdeep, a command-line tool based on a state-of-the-art convolutional neural network (CNN) to predict the consequence of NCV that explicitly focuses on ASE. ASdeep uses Markov chain-Monte Carlo (MCMC) methods to estimate ASE using allelic read counts quantified from RNA sequencing data. The model can be used to predict ASE from a DNA sequence in the absence of RNA data to enable targeted follow-up transcriptomics. ASdeep can also pinpoint hot-spot genomic regions that harbor potentially causal NCV using feature-attribution methods [10], which can aid in exploring the architecture of genomic regulations.

Implementation

The ASdeep package was written in Python and built upon PyTorch, a sophisticated deep-learning framework, to work as a downstream analysis tool after quantifying reads using software like ASEReadCounter from GATK [11]. ASdeep has four main functions: (i) to infer ASE, (ii) to create the train/test database, (iii) to train the CNN model and (iv) to predict/attribute the ASE effects to NCV. To estimate the ASE (target) of a transcript, ASdeep assigns allelic read counts to each allele based on phased genotypes at the heterozygous positions, per individual. The imbalance is then estimated by Bayesian estimation using the MCMC approach, assuming the number of allelic read counts follows a Beta-Binomial distribution. Next, ASdeep transforms the DNA sequence into an image-like matrix (a two-dimensional Hilbert curve) following the same steps introduced in a previously published work [12]. However, we did not use one-hot encoding methods but rather directly used the
ASdeep first estimates the ASE effects of a transcript from allelic read counts using Bayesian estimation and MCMC. The tool then trains a CNN model with pre-built architecture on the genomic context of the transcript. Finally, ASdeep can predict ASE effects and attribute them to hot-spot genomic regions that potentially harbor variants involved in regulation of gene expression. ASdeep is also applicable to other allele-specific data, such as allele-specific methylation and allele-specific open chromatin, where allelic read counts are available.

Acknowledgments

We thank the UMCG Genomics Coordination Center, the UMCG Research IT programme and the UG Center for Information Technology and their sponsors Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL) & TarGet for storage and compute infrastructure. We thank the Biobank-Based Integrative Omics Studies (BIOS) Consortium, funded by BBMRI-NL, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO) under award number 184.021.007. The BIOS Consortium members are listed in Supplementary Data S1. We thank the Genotype-Tissue Expression (GTEx) Project, supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx).

Funding

Z.Z. received a joint scholarship from the University Groningen and China Scholarship Council (under grant number 201706350277). Z.Z., K.J.v.d.V and M.A.S. received funding from the Netherlands Organisation for Scientific Research (NWO) under VIDI grant number 917.164.455. In addition, we acknowledge support from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 779257 (Solve-RD) and 825575 (European Joint Programme on Rare Disease).

Data availability

The source code of ASDeep is publicly available on GitHub (https://github.com/zhenhua-zhang/ASdeep), and its release versions are available at PyPi (https://pypi.org/project/ASdeep).