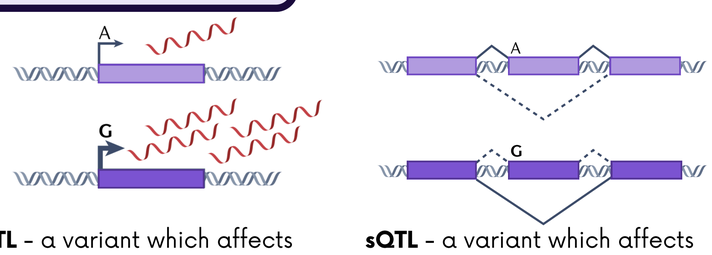
Predicting the molecular mechanisms of genetic variants

Dzvenymyra-Marta Yarish
Supervisor: Kaur Alasoo, PhD

Introduction



eQTL - a variant which affects gene expression

sQTL - a variant which affects splicing patterns

Mode of action (MoA) of a genetic variant - the specific way in which the variant influences the complex trait on a molecular level.

Drug trials for diseases with known causal genes have a higher chance of success [1]. One way to be more confident about causal genes is to prioritise sQTLs over eQTLs as drug targets. eQTL catalogue is a collection of uniformly re-computed eQTLs and sQTLs from 32 studies [2].

However, quantification methods used to distinguish between different modes of action **are not reliable** and **do not work with low-frequency alleles**.

Objective

To investigate the potential of machine learning (ML) for predicting the mode of action of genetic variants. We aim to develop **a model that can differentiate between eQTLs and sQTLs using sequence features.** The approach involves the following steps:

1. Building a dataset specifically for variant MoA analysis.
2. Validating the latest splicing and chromatin accessibility prediction models on a manually curated set of variants
3. Developing a variant mode of action prediction model that incorporates both traditional and neural features.

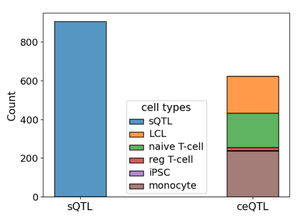
Methods

The Mode-of-Action (MoA) dataset was collected in three steps: manual labelling, chromatin accessibility (ca)QTL mapping and QTLs that affect gene expression via chromatin accessibility (ceQTLs) definition.

Manual labelling
RNA-seq coverage plots, select sQTLs

caQTLs (fine)mapping
6 datasets, 5 cell types

ceQTL definition
caQTLs ∩ eQTLs from the eQTL Catalogue



MoA model is a simple binary classifier - logistic regression or decision tree - which uses a set of classic and neural features to predict the mode of action of a variant.

Classic features

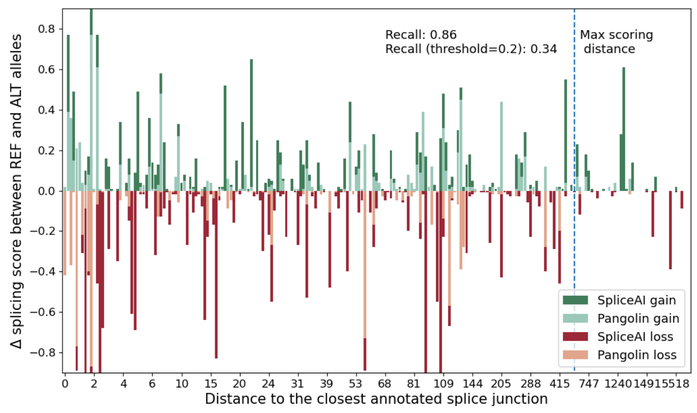
1. Binary variable indicating whether the variant is located within the gene body
2. Distance from the variant to the closest annotated splice junction
3. Number of overlaps with open chromatin regions in 5 cell types.
4. Number of overlaps with binding sites of RNA binding proteins.

Neural features

1. Splicing scores from SpliceAI [3] and Pangolin.
2. Enformer [4] SAD scores for five CAGE tracks (gene expression) and five DNASE tracks.
3. ChromBPNet [5] difference scores for five cell types.

We compared our composite MoA model **against a monolithic unified model (190M params)**, called Borzoi, which directly predicts RNA-seq coverage [6]. In the paper, Borzoi authors demonstrated its ability to distinguish between sQTLs, eQTLs and matched set of negatives. So, we used Borzoi's cell type specific RNA-seq and DNASE tracks scores to fit a classifier.

Results



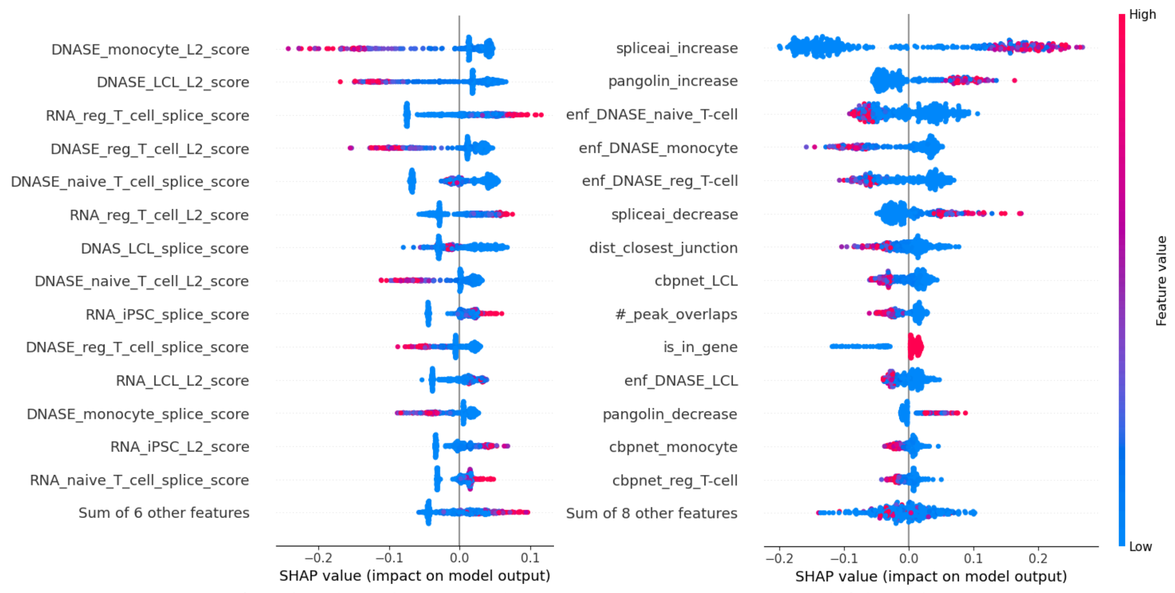
SpliceAI and Pangolin scores for hand-labelled set of sQTLs.

Dataset	r_s for variants in peaks		
	CBPNet x effect	Enformer x effect	CBPNet x Enformer
LCLs-1	**0.727**	0.623	0.725
LCLs-4	**0.802**	0.74	0.778
naive T cells	0.687	**0.742**	0.74
reg T cells	**0.838**	0.73	0.851
iPSCs	**0.75**	0.743	0.818
monocytes	**0.742**	0.651	0.743

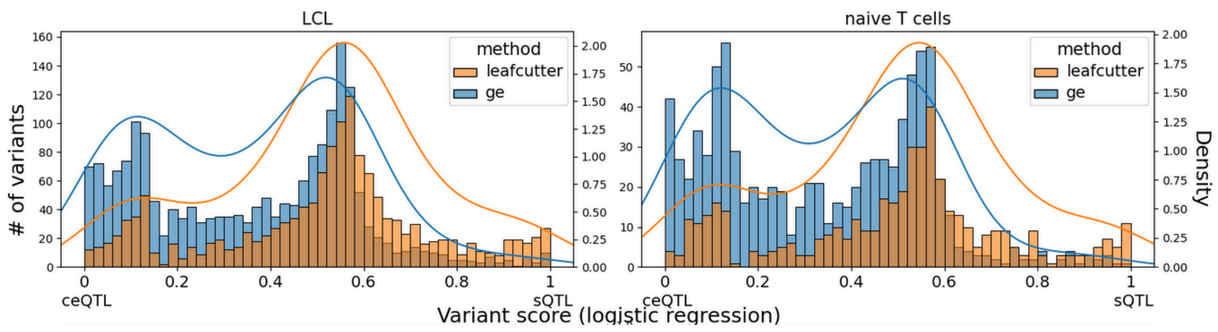
Performance of ChromBPNet and Enformer on caQTL datasets.

Cell type	Logistic Regression				Random Forest			
	Borzoi	MoA (Neural)	Borzoi*	MoA (Neural)*	Borzoi	MoA (Neural)	Borzoi*	MoA (Neural)*
All	0.71	**0.848**	-	-	0.8	**0.865**	-	-
LCL	0.637	0.814	0.604	**0.847**	0.825	0.864	0.827	**0.88**
monocytes	0.614	0.837	0.588	**0.864**	0.819	**0.885**	0.774	0.87
naive T cells	0.577	0.851	0.648	**0.873**	0.829	0.803	0.786	**0.87**

Comparison of MoA model and Borzoi-based classifiers (f1 score)
* - cell type specific features were used.



Borzoi classifier (Random Forest) | MoA model (Random Forest)



Distribution of the MoA model probabilities for eQTL catalogue QTLs.
(detected by ge - alleged eQTLs, by Leafcutter - alleged sQTLs)

Conclusion

We collected the MoA dataset, which includes two classes of molQTLs: splicing QTL and gene expression influenced by chromatin accessibility QTL. In parallel, we compared the performance of two deep learning models, Enformer and ChromBPNet, which represent two opposite approaches to predicting regulatory activity, on a set of fine-mapped chromatin activity QTLs. **ChromBPNet proved to be more precise in predicting the caQTLs effect.** Finally, we built the MoA model, combining classic genomic features and predictions of single-task deep learning models. The model demonstrated **nearly 90% accuracy in distinguishing between the two QTL classes,** compared to the 80% accuracy achieved by a classifier based on scores from a single multi-task large-scale model.

Finally, we scored the QTLs from the eQTL catalogue, detected by either gene expression or Leafcutter methods, with our model. This analysis revealed that while predictions from the MoA model more or less align with gene expression QTLs, most of the Leafcutter QTLs are not classified as sQTLs.

[1] Eric Vallabh Minikel et al. "Refining the impact of genetic evidence on clinical success". In: Nature (2024).
[2] Nurlan Kerimov et al. "A compendium of uniformly processed human gene expression and splicing quantitative trait loci". In: Nature Genetics 53 (2021), pp. 1290–1299.
[3] Kishore Jaganathan et al. "Predicting Splicing from Primary Sequence with Deep Learning". In: Cell 176 (2019), 535–548.e24.
[4]Žiga Avsec et al. "Effective gene expression prediction from sequence by integrating long-range interactions". In: Nature Methods 18 (2021), pp. 1196–1203.
[5] Anusri Pampari et al. "Bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants". URL: https://github.com/kundajelab/chrombpnet.
[6] Johannes Linder et al. "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation". In: bioRxiv (2023).

UNIVERSITY OF TARTU
Institute of Computer Science