Human Islet Research Network (HIRN): Alternative Splicing Events, Random Forest Model Card

Javier E. Flores

2023-01-26

Data

Inclusion levels of alternative splicing (AS) events of five different varieties (i.e. skipped exon (SE), retained intron (RI), alternative 5’ splice site (A5SS), alternative 3’ splice site (A3SS), and mutually exclusive exons (MXE)) were measured in human blood samples from two separate cohorts of patients.

Cohort 1 (Training Cohort):

- 12 healthy controls; 12 new onset type 1 diabetic (T1D) cases
- cases and controls matched on biological sex, age, and body mass index (BMI)
- 180 million reads

Cohort 2 (Testing Cohort):

- 12 healthy controls; 12 new onset type 1 diabetic (T1D) cases
- cases and controls matched on biological sex and age. BMI not recorded.
- 150 million reads

Event	Total Events (Cohort 1)	Total Events (Cohort 2)	Total Events (Shared)
Skipped exon (SE)	104590	69597	56530
Retained intron (RI)	4768	4158	4088
Alternative 5’ splice site (A5SS)	5544	4169	3919
Alternative 3’ splice site (A3SS)	8521	6374	6001
Mutually exclusive exon (MXE)	20666	12064	8332

Owners: Human Islet Research Network (HIRN)
Objective: Predict new onset T1D based on inclusion levels of AS events, and identify the most salient events for accurate prediction.

Approach

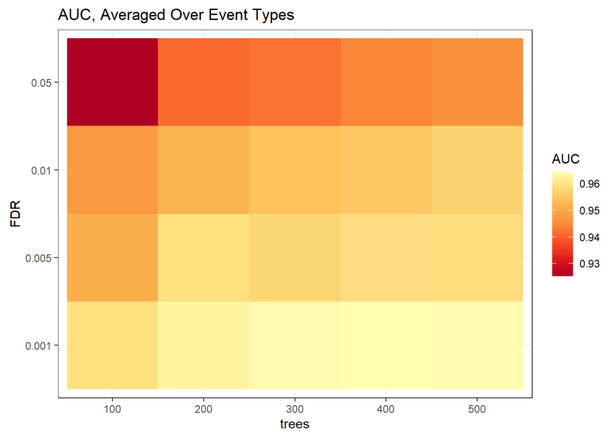
Model: Random Forest

- Implemented in R using the tidymodels and ranger packages.

Preprocessing: Event data in Cohort 1 that are missing in Cohort 2 are imputed based on the means of the Cohort 1 data.

Tuning: Grid-search

- Repeated 3-fold cross-validation with 25 repeats
- Tuned over the number of trees (100, 200, 300, 400, 500) and false discovery rate (FDR) threshold (0.05, 0.01, 0.005, 0.001)
- Other model hyperparameters (i.e. the number of randomly selected predictors and the minimal node size) were kept at software defaults
- Area-under-the-curve (AUC) was used as the selection metric



Final Model:

- 300 trees; FDR threshold of 0.001
- Evaluated on training data through repeated 3-fold cross-validation with 100 repeats
- Evaluated on (mean-imputed) testing data
- Evaluations on training and test data repeated 100 times
- AUC used as the evaluation metric

Data are accessible on DataHub. Code for data processing, model tuning, and final model fitting/evaluation is available on GitHub.

Results

Event	Event Count	AUC, Training (95% CI)	AUC, Test (95% CI)
Retained Intron (RI)	370	0.897 (0.889, 0.904)	0.869 (0.799, 0.913)
Skipped Exon (SE)	1872	0.977 (0.972, 0.981)	0.695 (0.524, 0.837)
Alternative 5’ splice site (A5SS)	179	0.969 (0.964, 0.973)	0.69 (0.583, 0.781)
Alternative 3’ splice site (A3SS)	273	0.983 (0.979, 0.986)	0.688 (0.569, 0.778)
Mutually exclusive exons (MXE)	251	0.981 (0.977, 0.985)	0.53 (0.427, 0.612)