ABO Blood Type Classifiers

All results described here may be found, replicated, and rerun on different data using Arvados at `http://curover.se/su92l-j7d0g-swtofxa2rct8495`.

Currently, *in silico ABO* blood type classification has very few examples in the literature. These classifiers use known variants associated with the *ABO* blood type phenotype to build their predictions. For instance, BOOGIE, a predictor using SNV databases, predicts the *ABO* blood type group with 94.2% accuracy for well sequenced PGP full genome sequences [1].

75 of the 178 PGP whole genome sequences have self-reported *ABO* blood types: 32 Type O, 30 Type A, 13 Type B, and 0 Type AB. A Chi-Squared test on the 71 participants of European ancestry with reported *ABO* blood types (30 Type O, 28 Type A, and 13 Type B) did not indicate significant deviation from the expected *ABO* blood types by ethnicity [2] ($p = 0.0849$), as expected.

A Antigen Classifier

The summary of classifier parameterization results for the A antigen are in Table 1. The support vector classifier (SVC) with l1 regularization and a linear kernel predicts A antigen presence with the highest accuracy, measured using leave-one-out cross-validation, (93.3% ± 24.9%) at two non-sequential error penalties (C values), 0.01 and 3.16 (Fig. 1). The classifier with the 0.01 error penalty has one non-zero coefficient, weighting the second phase tile position 00.1c4.038c, which is in Intron 1 in the *ABO* gene (GRCh37 chr9: 136,149,787 - 136,150,036). We believe this tile position was the only *ABO* tile to be weighted due to the l1 penalty, which favors a sparse coefficient vector, and the fact that of the 28 well sequenced tile positions in the *ABO* gene, this position was the only one to have exactly 2 variants. This classifier misclassifies 5 of the training called genomes; it predicts that `hu1187FF, hu2FEC01, huC14AE1, huEBD467, huFFAD87` do not have the A antigen. The single non-zero coefficient also explains the 5 misclassifications of the training data, since called genomes may have variants resulting in the A antigen phenotype and still have the intron tile associated by the classifier with the O phenotype. Despite this classifier's drawbacks, we wish to emphasize that given 75 labeled called genomes, it selected, with no prior knowledge, out of over 2 million tile positions, one tile in the *ABO* gene. This tile provides an accuracy one percent less than the BOOGIE *ABO* classifier, which relies on SNV-blood type databases.

The classifier with the 3.16 error penalty has 30 non-zero coefficients. The

Table 1. Maximum accuracy for A antigen classifiers, measured using leave-one-out cross-validation, the optimal parameter(s) for that classifier type, and the parameters tested. The classifier type is followed by the type of kernel, and if the penalty type is available, the penalty name.

Classifier Type	Maximum accuracy (± StDev)	Parameter resulting in max acc	Parameters tested
NuSVC (linear)	0.600 ± 0.490	all	nu={0.1, 0.15, 0.2, 0.25, 0.3}
NuSVC (rbf)	0.600 ± 0.490	all	nu={0.1, 0.15, 0.2, 0.25, 0.3}
SVC (linear)	0.600 ± 0.490	all	C={0.001, 0.01, 0.01, 1, 10, 100, 1000, 10000, 100000, 1000000}
SVC (rbf)	0.600 ± 0.490	all	C={0.001, 0.01, 0.01, 1, 10, 100, 1000, 10000, 100000, 1000000}
SVC (linear, l2 penalty)	0.400 ± 0.490	all	C={0.001, 0.01, 0.01, 1, 10, 100, 1000, 10000, 100000, 1000000}
SVC (linear, l1 penalty)	0.933 ± 0.249	C = {0.01, 3.16}	C={0.001, 0.00316, 0.01, 0.0316, 0.1, 0.316, 1, 3.16, 10, 100, 1000, 10000, 100000, 1000000}

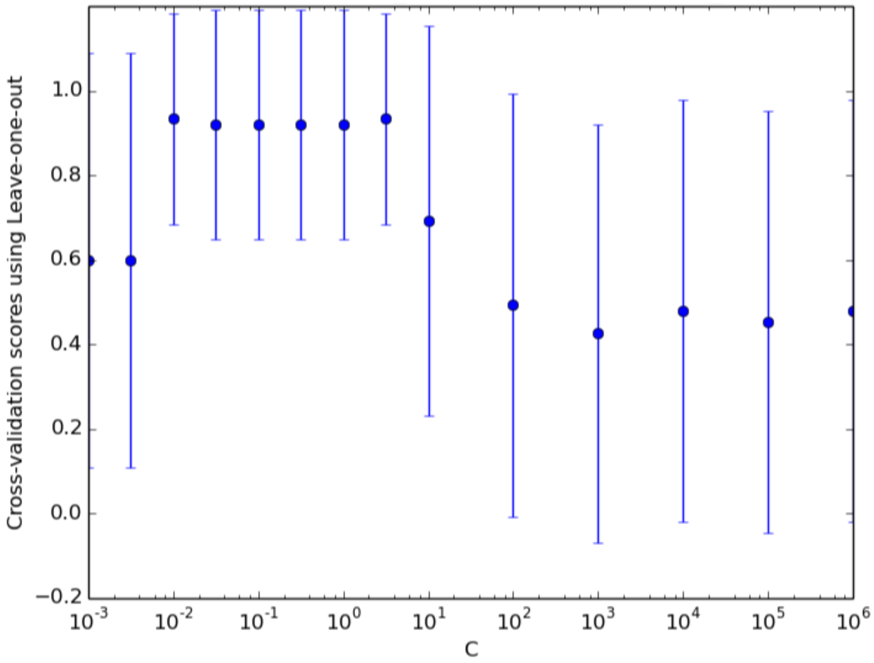

Figure 1. Classification accuracy for SVC with a linear kernel and a l1 penalty, as measured by cross-validation, as a function of the error penalty. Note two points have a maximum accuracy of 93.3%.

coefficient with the largest magnitude also weighs tile position 00.1c4.038c in the *ABO* first intron, again on the second phase. The second largest magnitude is ten percent of this magnitude and weighs tile position 00.114.2212 in chr5.q35.1. This classifier does not misclassify any of the training genomes. These classifiers produce identical predictions of A antigen presence for the 103 unlabeled called genomes (41/103 genomes are predicted to have an A antigen phenotype).

We hypothesize the error penalty of 0.01 allowed 5 misclassifications of the training data while the error penalty of 3.16 required small non-zero coefficients to exist so no members of the training data set would be misclassified. We predict the classifier with the error penalty of 3.16 overfits to the data rather than actually learning the rarer variants resulting in the A antigen phenotype in `hu1187FF, hu2FEC01, huC14AE1, huEBD467,` and `huFFAD87`. We predict training on a larger, more heterogenous training set using a less exclusive mechanism for incorporating poorly sequenced regions will widen the range of weighted tiles and increase the phenotypic accuracy.

B Antigen Classifier

The summary of the parameterization results for the B antigen are in Table 2. The nu-support vector classifier (NuSVC) and support vector classifier (SVC) with linear kernels predict B antigen presence with the highest accuracy, measured using leave-one-out cross validation (84.0% ± 36.7%). The NuSVC classifier predicts B antigen presence with this accuracy at low nu values, which restrict the number of misclassifications. The SVC classifier produces this accuracy regardless of the error penalization magnitude. We predicted B antigen presence for the 103 unlabeled called genomes for a SVC with a linear kernel and the default error penalization of 1 and a NuSVC with a nu of 0.1 and a linear kernel. Their predictions were identical: 44/103 called genomes are predicted to have a B antigen phenotype. Both classifiers had 1,786,803 non-zero coefficients (85.2% of the available tile positions). The largest coefficient magnitude was $2.11 * 10^{-5}$, 4 orders of magnitude less than the coefficient magnitudes of the A antigen classifiers. Both classifiers had three coefficients with a magnitude greater than 95% of the maximum. The largest and the third largest weigh phase A tile positions in chr14.q32.2. The second largest coefficient weighs a phase B tile position 00.1c4.0389, which is in the *ABO* first intron (GRCh37 chr9: 136,149,112 - 136,149,361). The NuSVC classifier with a linear kernel and a nu of 0.3, which had an accuracy of 82.7% ± 37.9%, had identical predictions, the same number of non-zero coefficients, the same largest coefficient magnitude, and the same 3 strongest weighed tiles. None of these classifiers misclassified any training called genomes.

Combining the predictions generated by the A antigen and B antigen classifiers with the highest accuracies, 34 of the 103 unlabeled PGP called genomes were labeled type O, 23 were labeled type A, 28 were labeled type B, and 18 were labeled AB. A Chi-Squared test of the 102 caucasian called genomes (34 predicted type O, 23 predicted type A, 28 predicted type B, and 17 predicted type AB)

Table 2. Maximum and second highest accuracy for B antigen classifiers, measured using leave-one-out cross-validation, and the parameter(s) used to obtain the reported accuracy. The classifier type is followed by the type of kernel, and if the penalty type is available, the penalty name.

B antigen classifier type	Max Acc (± StDev)	Max acc param	Next-highest acc (± StDev)	Next-highest acc param	Parameters tested
NuSVC (linear)	0.840±0.367	nu={0.1, 0.15, 0.2, 0.25}	0.827 ± 0.379	nu=0.3	nu={0.1, 0.15, 0.2, 0.25, 0.3}
NuSVC (rbf)	0.827±0.379	all	n/a	n/a	nu={0.1, 0.15, 0.2, 0.25, 0.3}
SVC (linear)	0.840±0.367	all	n/a	n/a	C={0.001, 0.01, 0.01, 1, 10, 100, 1000, 10000, 100000, 1000000}
SVC (rbf)	0.827±0.379	all	n/a	n/a	C={0.001, 0.01, 0.01, 1, 10, 100, 1000, 10000, 100000, 1000000}
SVC (linear, l2 penalty)	0.173±0.379	all	n/a	n/a	C={0.001, 0.01, 0.01, 1, 10, 100, 1000, 10000, 100000, 1000000}
SVC (linear, l1 penalty)	0.827±0.379	C=0.01	0.813 ± 0.390	C=1.0	C={0.001, 0.00316, 0.01, 0.0316, 0.1, 0.316, 1, 3.16, 10, 100, 1000, 10000, 100000, 1000000}

indicated the predicted *ABO* blood types significantly deviated from the expected *ABO* blood types by ethnicity [2] ($p < 0.00001$).

The low magnitudes of the coefficients, the high number of coefficients, and the phenotype predictions, which significantly deviates from the expected phenotypes, indicate that these classifiers are highly overfitted to our small training population. Given that the number of training sets is vastly surpassed by the number of features and that only 11 genomes have the B antigen phenotype in our labeled population, we do not consider this overfitting to be not surprising.

The support vector classifier with a linear kernel, l1 regularization, and an error penalty of 0.01, which also has an accuracy of 82.7% ± 37.9%, has no non-zero coefficients. It predicts that none of the unlabeled called genomes have the B antigen, and misclassifies the 11 training genomes that have the B antigen (`hu04DF3C, hu04F220, hu0A4518, hu687B6B, hu7A2F1D, hu8073B9, hu82436A, hu8E87A9, huA05317, huA4E2CF, huAA53E0, huB4883B,` and `huDBF9DD`). The support vector classifier with a linear kernel, a l1 regularization, and an error penalty of 1, which has an accuracy of 81.3% ± 39.0%, has 33 coefficients. The maximum coefficient, with a magnitude of 0.238, weighs a phase A tile position 00.1c3.0fd2 (GRCh37 chr9: 135,053,558 -135,053,812), 1,071,976 bases before from the *ABO* gene. Though a position in the *ABO* gene is weighed (00.1c4.0389, phase B, GRCh37 chr9: 136,149,112 - 136,149,361), it has the 22nd largest magnitude (0.00229). This classifier did not misclassify any training called genomes, and predicts B antigen presence in 12/103 unlabeled genomes.

Using the B antigen predictions generated by the support vector classifier with a linear kernel, l1 regularization, and an error penalty of 1, 55 of the 103 unlabeled PGP called genomes were labeled type O, 36 were labeled type A, 7 were labeled type B, and 5 were labeled AB. A Chi-Squared test of the 102 caucasian genomes (55 predicted type O, 35 predicted type A, 7 predicted type B, and 5 predicted type AB) indicated the predicted *ABO* blood types did not significantly deviate from the expected *ABO* blood types by ethnicity [2] ($p = 0.246002$).

We believe that adding feature selection, along with a larger and more varied training set, will increase the accuracies of our *ABO* blood type classifiers, since our current classifiers have a very large discrepancy between the number of features and the number of training sets. We are currently in the process of releasing a blood type survey to the Harvard PGP participants and adding their responses to our training set. Additionally, developing a less exclusive mechanism that incorporates poorly sequenced regions will allow the classifiers access to the underlying variants producing the *ABO* phenotype, which might also increase the accuracies of our classifiers. Finally, including known phenotypes, such as ethnicity, could strengthen the predictive accuracies of our classifiers.

References

1. Giollo M, Minervini G, Scalzotto M, Leonardi E, Ferrari C, and Tosatto SCE. BOOGIE: Predicting Blood Groups from High Throughput Sequencing

Data. PLoS One. 2015;10(4): e0124579. doi: 10.1371/journal.pone.0124579.

2. Garratty G, Glynn SA, and McEntire R. *ABO* and *RH(D)* phenotype frequencies of different racial/ethnic groups in the United States. Transfusion. 2004;44:703-706.