### **Supplemental Table S15.** Optimal number of gene predictors selected by LASSO Regression.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Gene Numbera | % Devianceb | % Misclassificationc |
| SubPop 1 vs. 2,3,4d | Known Markers | 47 | 6.6 | 42 |
| DE genes | 86 | 1.4 | 43.6 |
| SubPop 2 vs. 1,3,4 | Known Markers | 52 | 29.6 | 25.5 |
| DE genes | 9 | 59 | 14.6 |
| SubPop 3 vs. 1,2,4 | Known Markers | 52 | 35.4 | 2.6 |
| DE genes | 56 | 99.7 | 3.1 |
| SubPop 4 vs. 1,2,3 | Known Markers | 8 | 89.7 | 2.4 |
| DE genes | 14 | 99.2 | 1.9 |

a Number of genes remaining in the optimal LASSO logistic regression model, which classifies cells into clusters with the lowest 10-fold cross-validation error

b Deviance was calculated by log likelihood ratio on the left-out data in the 10-fold cross-validation (in the form of -2\*(log(p)-log(y)), where log(p) is for reduced model and log(y) is for the saturated model)

c Classification error of cells into subpopulations was calculated by using data from a new validated set independent of the model training set

d 2,3,4 denotes the cells in the combined subpopulations two, three and four.