Multicollinearity analysis for FCS and FCSM features

Multicollinearity may represent a problem in multivariable regression leading to coefficients overestimation of related variables. Therefore, association between variables considered for training the models were studied by bivariate analysis (linear or logistic regression according to compared variables) in order to explore aggressors' relationship, both for FCS and FCSM training data. Additionally, bivariate analysis was complemented computing Spearman correlation scores.

In FCS training subset

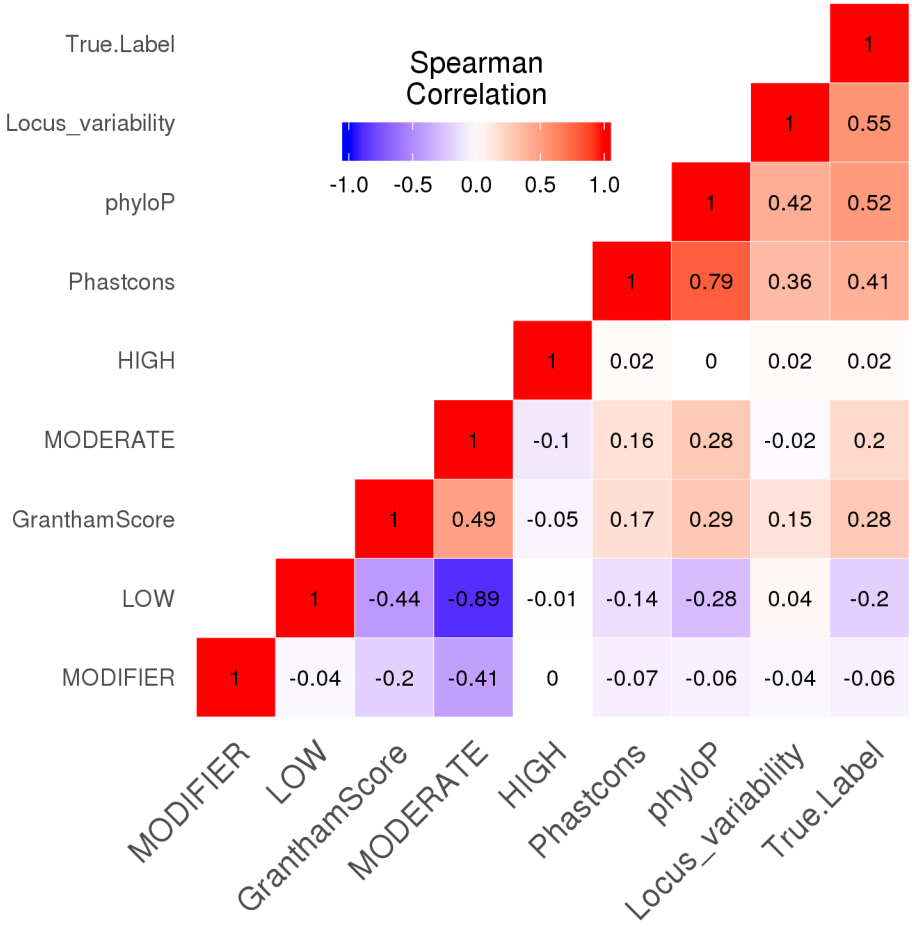

Figure 1. Heatmap for Spearman correlation rho values for FCS model's features.

Table 1. Bivariate analysis between considered features for FCS in training subset.								
X / **Y**	**Locus var**	**PhastCons**	**phyloP**	**Grantham**	**High**	**Moderate**	**Modifier**	**Low**
Locus var	-	β=1.833 SE=0.024 p<0.001***	β=0.265 SE=0.003 p<0.001***	β=0.006 SE=0.0002 p<0.001***	β=1.190 SE=0.354 p<0.001***	β=-0.308 SE=0.039 p<0.001***	β=-0.667 SE=0.087 p<0.001***	β=0.519 SE=0.043 p<0.001***
PhastCons	β=0.052 SE=0.0007 p<0.001***	-	β=0.088 SE=0.0004 p<0.001***	β=0.0015 SE=0.00004 p<0.001***	β=0.223 SE=0.060 p<0.001***	β=0.241 SE=0.006 p<0.001***	β=-0.261 SE=0.015 p<0.001***	β=-0.234 SE=0.007 p<0.001***
phyloP	β=0.513 SE=0.005 p<0.001***	β=6.023 SE=0.024 p<0.001***	-	β=0.02 SE=0.0003 p<0.001***	β=0.139 SE=0.492 p=0.777	β=3.53486 SE=0.052 p<0.001***	β=-1.559 SE=0.121 p<0.001***	β=-3.932 SE=0.057 p<0.001***
Grantham	β=2.345 SE=0.081 p<0.001***	β=19.689 SE=0.480 p<0.001***	β=3.911 SE=0.057 p<0.001***	-	β=-73.8798 SE=6.876 p<0.001***	β=81 SE=0.672 p<0.001***	β=-2.945 SE=34.675 p=0.932	β=-79.523 SE=0.761 p<0.001***
High	β=0.138 SE=0.038 p<0.001***	β=1.7503 SE=0.520 p<0.001***	β=0.011 SE=0.037 p=0.777	β=-2.351 SE=35.421 p=0.947	-	β=-20.06 SE=577.993 p=0.972	β=-13.65 SE=582.971 p=0.981	β=-14.71 SE=459.117 p=0.974
Moderate	β=-0.043 SE=0.006 p<0.001***	β=1.120 SE=0.031 p<0.001***	β=0.340 SE=0.006 p<0.001***	β=7.018 SE=38.557 p=0.856	β=-15.900 SE=72.195 p=0.826	-	β=-19.092 SE=78.897 p=0.809	β=-23.526 SE=168.900 p=0.889
Modifier	β=-0.104 SE=0.014 p<0.001***	β=-1.192 SE=0.070 p<0.001***	β=-0.122 SE=0.010 p<0.001***	β=-0.261 SE=0.015 p<0.001***	β=-10.473 SE=119.029 p=0.93	β=-21.075 SE=212.632 p=0.921	-	β=-15.548 SE=168.9 p=0.927
Low	β=0.071 SE=0.0062 p<0.001***	β=-1.084 SE=0.035 p<0.001***	β=-0.403 SE=0.007 p<0.001***	β=-3.485 SE=31.182 p=0.911	β=-11.0087 SE=72.195 p=0.879	β=-23.986 SE=212.632 p=0.91	β=-14.026 SE=78.897 p=0.859	-

Features of FCS presented very weak, weak or moderate Spearman correlation values, figure 1. Spearman results suggest a strong negative correlation observed between LOW and MODERATE, but this value may reflect the fact that both are the most represented categories in variant impact and are mutually exclusive. Actually, Spearman coefficients and bivariate analysis confirmed a strong positive correlation between conservation scores. Additionally, locus variability had a moderate association degree with both conservation scores. On the other hand, locus variability and specially PhastCons were the features that better represented variant impact over canonical transcript.

In FCSM training data set

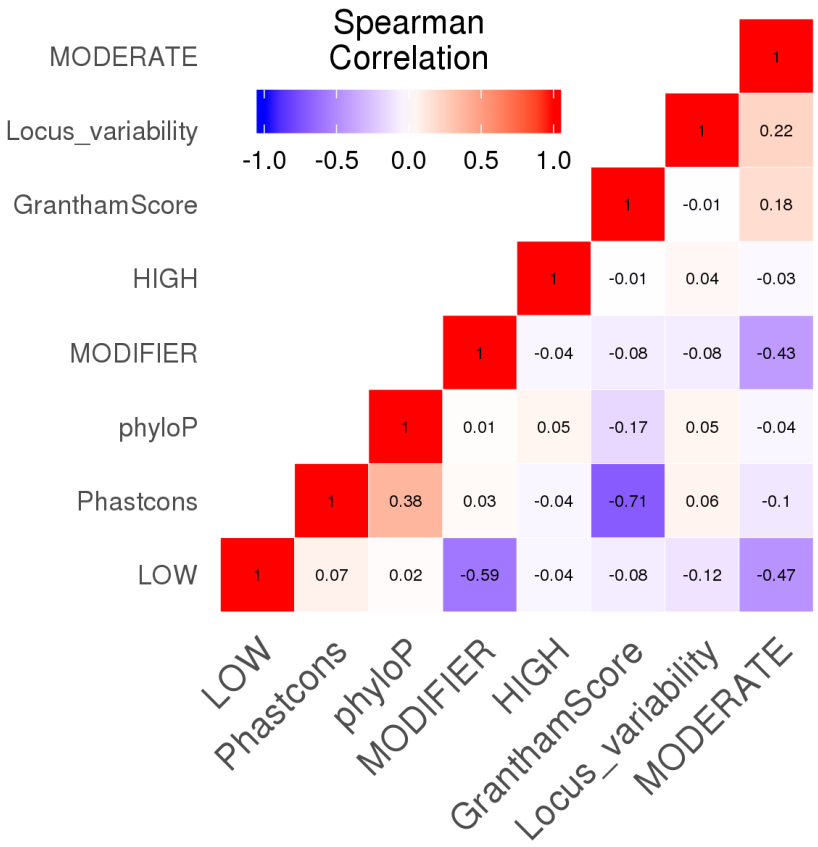

Figure 2. Heatmap for Spearman correlation rho values for FCSM model's features.

Table 2. Bivariate analysis between considered features for FCSM in training data set.								
X / **Y**	**Locus var**	**PhastCons**	**phyloP**	**Grantham**	**High**	**Moderate**	**Modifier**	**Low**
Locus var	-	β=0.798 SE=0.092 p<0.001***	β=0.055 SE=0.012 p<0.001***	β=0.002555 SE=0.0012 p=0.0312*	β=0.833 SE=0.178 p<0.001***	β=0.461 SE=0.020 p<0.001***	β=-0.280 SE=0.019 p<0.001***	β=-0.11 SE=0.018 p<0.001***
PhastCons	β=0.008 SE=0.0009 p<0.001***	-	β=0.064 SE=0.001 p<0.001***	β=-0.007 SE=0.0001 p<0.001***	β=0.565 SE=0.139 p<0.001***	β=-0.021 SE=0.002 p<0.001***	β=0.005 SE=0.002 p=0.0084**	β=0.013 SE=0.002 p<0.001***
phyloP	β=0.034 SE=0.007 p<0.001***	β=3.916 SE=0.062 p<0.001***	-	β=-0.007 SE=0.0009 p<0.001***	β=-0.057 SE=0.018 p=0.0013**	β=-0.084 SE=0.016 p<0.001***	β=0.034 SE=0.015 p=0.0223*	β=0.029 SE=0.014 p=0.0447*
Grantham	β=0.153 SE=0.071 p=0.0312*	β=-41.040 SE=0.604 p<0.001***	β=-0.720 SE=0.091 p<0.001***	-	β=-0.675 SE=1.379 p=0.624	β=2.629 SE=0.157 p<0.001***	β=-1.040 SE=0.145 p<0.001***	β=-1.103 SE=0.142 p<0.001***
High	β=1.481 SE=0.293 p<0.001***	β=-2.030 SE=0.742 p=0.0062**	β=0.423 SE=0.097 p<0.001***	β=-0.866 SE=63.179 p=0.989	-	β=-16.886 SE=874.561 p=0.985	β=-17.025 SE=745.906 p=0.982	β=-17.084 SE=709.534 p=0.981
Moderate	β= 0.589 SE=0.027 p<0.001***	β=-1.834 SE=0.202 p<0.001***	β=-0.141 SE=0.029 p<0.001***	β=1.231 SE=11.587 p=0.915	β=-12.503 SE=97.752 p=0.898	-	β=-19.141 SE=166.434 p=0.908	β=-19.241 SE=158.319 p=0.903
Modifier	β=-0.291 SE=0.020 p<0.001***	β=0.578 SE=0.222 p=0.0093**	β=0.061 SE=0.027 p=0.0236*	β=-1.087 SE=13.161 p=0.934	β=-12.960 SE=97.752 p=0.895	β=-18.460 SE=118.359 p=0.876	-	β=-19.875 SE=158.319 p=0.9
Low	β=-0.115 SE=0.019 p<0.001***	β=2.114 SE=0.358 p<0.001***	β=0.052 SE=0.026 p=0.0466*	β=-1.101 SE=13.0292 p=0.933	β=-13.119 SE=97.752 p=0.893	β=-18.659 SE=118.359 p=0.875	β=-19.975 SE=166.434 p=0.904	-

Notably, bivariate analysis matched with Spearman correlation results (table 2 and figure 2, respectively). Unlike in FCS training subset (nuclear DNA), conservation scores presented very weak association between them and with locus variability, reflecting an actual difference between both genomes. Moreover, PhyloP showed a weak association with variants' impact, while PhastConst was strong and inversely related to impact over canonical transcript, in an opposite situation to FCS training subset. Focusing on bivariate analysis, locus variability was the feature that better explained variants impact.

Models training and selection

We trained four different models: a random forest, a logistic regression, a least absolute shrinkage and selection operator (LASSO) and a neural network, using 5-fold cross validation, splitting the data into 80% training and 20% evaluation set.

Random forest algorithm (both in FCS and FCSM) was trained considering four possible numbers of variables tried at each split, 2, 3, 4 and 5. Tuned up parameters for neural networks were the number of hidden units (from 1 to 10) and the weight of decay that tells how dominant the regularization term will be in the gradient computation (from 0 to 4, by intervals of 0.125).

Models trained were selected according to root mean squared error (RSMSE) in train subset **(in whole training data set for FCSM)**. Selected models were evaluated in test subset **(in validation data set for FCSM)** and most accurate model, measured as the one with the highest area under the receiving operator characteristic (ROC) curve was selected as FCS or FCSM. Finally, FCS was submitted to a second validation using ClinVar validation data set.

Models were trained using caret v-6.0 (McCollum, 2009), glmnet v-2.0 (Friedman, Hastie, & Tibshirani, 2010), ranger v-0.11.2 (Wright & Ziegler, 2017) and nnet v-7.3 (Venables, W. N. & Ripley, 2002) R-packages. For Received Operative Curves performance and comparison of Areas Under the Curve pROC v-1.15.0 (Robin et al., 2011) and ROCR v-1.0 (Sing, Sander, Beerenwinkel, & Lengauer, 2005) R-packages were used.

Results for FCS:

The best random forest model, contained 500 trees and 5 variables tried at each split, for selected lasso model the minimum λ=0.00117, tuned up neural network presented an architecture of 10 units in hidden layers and decay=0.125. Random forest (AUC=0.92) outperformed all other trained models in neutral/deleterious variant classification and was selected as FCS, figure 2.

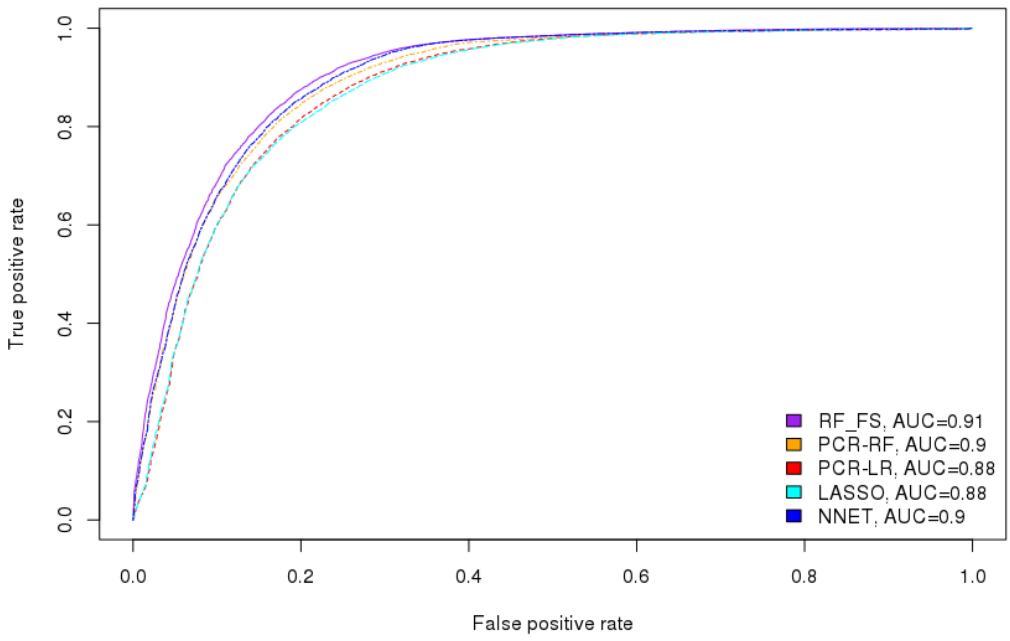

Figure 2. AUC comparison between different trained models in test subset. RF: random forest regression, LR: logistic regression; LASSO: Least absolute shrinkage and selection operator, NNET: Neural network.

Results for FCSM:

Selected Random forest model in training step presented 5 variables for splitting at each tree node and 500 trees. Tuned up lasso model had a minimum λ=0.0000613. Selected neural network consisted in a 10 units in hidden layer network and decay=0. As in FCS, random forest (AUC=0.92) outperformed all other trained models in neutral/deleterious variant classification, so was considered as FCSM, figure 3.

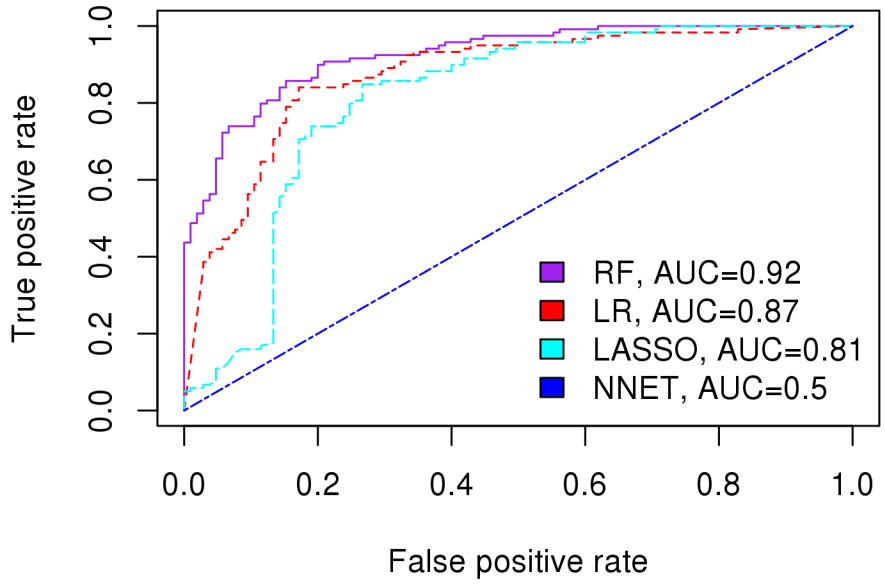

Figure 3. AUC comparison between different trained models for mtDNA. RF: random forest regression, LR: logistic regression; LASSO: Least absolute shrinkage and selection operator, NNET: Neural network.

Variable importance

Feature relative importance within selected models was evaluated by computing two different statistics, the Net Reclassification Improvement index and a D statistic computed within 2000 bootstrap re-sampling cycles as:

$$D = \frac{AUC1 - AUC2}{s}$$

where **AUC1** is the area under the curve of predictor number 1, **AUC2** is the area under the curve of predictor number 2 and **s** is the standard deviation for the difference between both values, among n prefixed cycles of bootstrap. We considered 2000 bootstrap cycles. Finally, this D statistic is compared with a normal distribution, to get the probability value. NRI was calculation PredictABLE v-1.2.2 R-package and D statistic was calculated meanwhile pROC R-package. Variable relative importance for FCS and FCSM are gathered in table 3.

Table 3. Variable importance for FCS and FCSM features measured as NRI and D statistic.

	FCS		**FCSM**	
Feature	**NRI [IC 95%] (p-value)**	**D (p-value)**	**NRI [IC 95%] (p-value)**	**D (p-value)**
Locus variability	1.4173 [1.3976 – 1.437] (p-value<0.001)	36.609 (p-value<0.001)	1.1154 [0.9136 – 1.3172] (p-value<0.001)	4.0338 (p-value<0.001)
phyloP	0.3869 [0.3666 – 0.4072] (p-value<0.001)	4.9122 (p-value<0.001)	0.1317 [-0.0267 - 0.29] (p-value=0.1031)	0.73941 (p-value=0.4597)
PhastCons	-0.0782 [-0.0925 - -0.0638] (p-value<0.00)1	3.3518 (p-value<0.001)	-0.005 [-0.1617 – 0.1516] (p-value=0.94969)	-0.037599 (p-value=0.97)
Grantham Score	0.2399 [0.221 – 0.2588] (p-value<0.001)	3.9407 (p-value<0.001)	0.0174 [-0.1455 - 0.1802] (p-value=0.83443)	0.55045 (p-value=0.582)
High impact	0.1476 [0.1362 – 0.159] (p-value<0.001)	-2.3809 (p-value=0.017)	0.0768 [-0.0463 - 0.1998] (p-value=0.22168)	0.89461 (p-value=0.371)
Moderate impact	0.0694 [0.0597 – 0.0792] (p-value<0.001)	-2.92 (p-value=0.004)	0.0779 [-0.0405 - 0.1963] (p-value=0.1973)	-0.31859 (p-value=0.75)
Modifier impact	0.1259 [0.1149 – 0.1368] (p-value<0.001)	-2.5153 (p-value=0.012)	0.0168 [-0.1009 - 0.1345] (p-value=0.77962)	0.83945 (p-value=0.4012)
Low impact	0.1162 [0.1056 – 0.1268] (p-value<0.001)	-3.0847 (p-value=0.002)	-0.0308 [-0.1623 - 0.1007] (p-value=0.64612)	-1.3973 (p-value=0.1623)

References

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20808728

McCollum, A. G. H. (2009). Building Predictive Models in R Using the caret Package. *Seminars in Orthodontics*, *15*(3), 159–160. https://doi.org/10.1053/j.sodo.2009.03.002

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), 3940–3941. https://doi.org/10.1093/bioinformatics/bti623

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Forth edit). New York: Springer.

Wright, M. N., & Ziegler, A. (2017). ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01