SUPPLEMENTAL FILES

Supplemental Table S1 (PDF format): sequence and other features used in SNPdryad algorithm.

Supplemental Table S2 (MS EXCEL format): Pfam domains which harbor the harmful nsSNPs predicted by SNPdryad.

Supplemental Table S3 (MS EXCEL format): The dataset extracted from SNPdbe.

Supplemental Figure S1 (PDF format): multiple sequence alignment generated by SNPdryad on human Complement C3 protein (P01024). Residue Q at position 1161 (highlighted in blue box) is absolutely conserved among the orthologous sequences; a mutation to Lysine (Q->K) causes increased susceptibility to hemolytic uremic syndrome atypical type5 (AHUS5).

Supplemental Figure S2 (PDF format)**:** multiple sequence alignment generated by SIFT on human Complement C3 protein (P01024). The inclusion of protein paralogs caused the inclusion of deleterious substitutions (Q->K, Q->R) at position 1161.

Supplemental Figure S3 (PDF format)**:** multiple sequence alignment generated by PolyPhen2 on human Complement C3 protein (P01024). The inclusion of protein paralogs caused the inclusion of deleterious substitutions (Q->K, Q->R) at position 1161.

Supplemental Figure S4 (PDF format)**:** multiple sequence alignment generated by SNPdryad on Transthyretin protein (P02766). Residue I at position 104 (highlighted in blue box) is very highly conserved among the orthologous sequences; a mutation to Serine (I->S) contributes to transthyretin-related amyloidosis (AMYL-TTR).

Supplemental Figure S5 (PDF format)**:** multiple sequence alignment generated by SIFT on Transthyretin protein (P02766). The inclusion of protein paralogs caused the inclusion of deleterious substitutions (I->S) at position 104.

Supplemental Figure S6 (PDF format)**:** multiple sequence alignment generated by PolyPhen2 on Transthyretin protein (P02766). The inclusion of protein paralogs caused the inclusion of deleterious substitutions (I->S) at position 104.

Supplemental Figure S7 (PNG format)**:** Average SNPdryad prediction score for all the possible pair-wise substitutions between 20 amino acid residues on all the amino acid positions in the human proteome. The higher scores indicate higher likelihood of deleterious effect of the nsSNP. The axis aa1 represents the reference (original) amino acid residue; the axis aa2 is the variant (substitute) amino acid residue.

SUPPLEMENTAL TEXT

Parameter Settings:

For Random Forest, the number of trees is set to 100; for each tree, the maximal depth is unlimited. The number of features used in random selection is set to log2 (number of attributes used) + 1. The Sun Java random seed is set to 1.

For Naive Bayes, the implementation proposed by John and Langley is adopted \citep{John:1995:ECD:2074158.2074196}.

For Bayesian Network, the conditional probability tables are estimated directly once the structure has been learned; the K2 search algorithm is used for hill climbing \citep{Cooper:1992:BMI:145254.145259}.

For Multilayer Perceptron, back-propagation is used for learning. The learning rate is set to 0.3; the momentum applied to the weights during updating is set to 0.3. The nominal features are transformed to the corresponding binary encodings. Numeric features are normalized. The Sun Java random seed is set to 0. The maximal number of epochs is set to 500.

For AdaBoost, the M1 method is used. Decision stumps are used as the base classifiers. The number of iterations is set to 10. The Sun Java random number seed is set to 1. The weight threshold for weight pruning is set to 100.

For Support Vector Machines, both polynomial and radial basis kernels are evaluated. John Platts sequential minimal optimization algorithm is adopted for training \citep{Platt:1999:FTS:299094.299105}. The complexity parameter is set to 1. The epsilon for round-off error is set to 10 - 12. The Sun Java random number seed is set to 1. The tolerance parameter is set to 0.001.

For k-Nearest Neighbor Classifiers, different values of k, e.g. 1, 3 or 5, are evaluated.