Analysis of mutations in p16

A number of recent studies have shown that the impact of missense mutations is more accurately predicted by consensus methods. We therefore used a strategy that assigned a growth rate for a p16 mutation based on information from six methods. Output from the methods were treated as the independent variables in a multivariate linear model, fitted to a training data set. We then used this model to predict relative cell growth rate for the variants in the challenge set.

Dataset

For training, 30 cell proliferation assay (as described in [1]) data points for p16 missense mutations at one of 25 positions were assembled from the literature and from the information provided with the challenge. Among the 30 variants, 29 were expressed in human diploid fibroblasts, and 1 was expressed in a U2-OS cell line. We excluded data from one study that used three glioma cell lines, because large variations in cell growth rate were observed for the same protein variant, and the data tended to be outliers in our model training.

Analysis of the functional impact of mutations

We used six methods to analyze the functional impact of the missense mutations: SNPs3D stability method [2], SNPs3D sequence profile method [2], Polyphen-2 [3], SIFT [4], CHASM [5], and Condel [6]. (On-line versions of SIFT and Condel were used, submitting all possible mutations in p16 to maintain the confidentiality of the data). These methods differ from each other by using different training data sets, selected features (sequence-based, or structure-based, or both), and types of classifiers.

Training

We used two multivariate linear models to fit the training data. The independent variables are scores predicted by the six methods, together with the total number of deleterious predictions and of neutral predictions. The dependent variable is the relative cell growth rate. The first linear model was fitted to the raw data, and had an intercept. The second model was adjusted so that 50% growth rate corresponded to the most neutral value for each method.

We then used R [7] to fit the linear models. Both models achieved an encouraging R-squared (0.61 and 0.91 respectively), as well as a significant P-value. We also assessed the fitted model by leave-one-out cross validation using the DAAG R-package [8]. This yielded regression residuals errors at almost same level.

Prediction

We performed predictions on the challenge set in the same way as we did on the training set, predicting the relative cell growth rate for each variant using the two fitted models. Because the cell lines used in the training set and the challenge set are different, and we observed relatively high cell growth rates in our predictions on the challenge set, we added a third submission in which predicted values are reduced by a factor of 1.47.To comply with the submission format, any prediction that is larger than 1.00 or smaller than 0.50 was adjusted back to 1.00 or 0.50. We used the residual standard errors of the corresponding models as the confidence of the predictions.

[1] Ruas, M., Brookes, S., McDonald, NQ, and Peters, G. (1999). Functional evaluation of tumour-specific variants of p16INK4a/CDKN2A: correlation with protein structure information. *Oncogene* **18**, 5423-5434.

[2] Yue, P. & Moult, J. (2006). Identification and analysis of deleterious human SNPs. *Journal of molecular biology* **356**, 1263-74.

[3] Adzhubei, IA, Schmidt, S., Peshkin, L., Ramensky, VE, Gerasimova, A., Bork, P., Kondrashov, AS, Sunyaev, SR. (2010). A method and server for predicting damaging missense mutations. *Nature Methods* **7**(4), 248-249.

[4] Kumar, P., Henikoff, S., Ng, PC. (2009). Predicting the effect of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocol* **4**(7),1073-1081.

[5] Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, VE., Kinzler, KW., Vogelstein, B., Karchin R. (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Research* **69**(16), 6660-6667.

[6] González-Pérez, A. & López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, condel. *American Journal of Human Genetics* **88**, 440-449.

[7] R Development Core Team (2011). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* **ISBN 3-900051-07-0**, URL http://www.R-project.org/.

[8] Maindonald, J. & Braun, WJ. (2013). DAAG: data analysis and graphics data and functions. *R package version 1.16.* URL http://CRAN.R-project.org/package=DAAG.