Using machine learning to identify genes of interest from epigenetic studies

Genomatix AG, Bayerstr. 85a, 80335 Munich, Germany, www.genomatix.de

Here we show how to use machine learning technologies to identify methylation sites changes induced by smoking using Infinium® HumanMethylation450K BeadChip data.

Introduction

Epigenetics has become a major part in the efforts to better understand the molecular mechanisms underlying biological processes from cell differentiation to disease manifestation and progression. While a full epigenomic study of the genetic material of a cell might yield the best results, looking at separate modifications like DNA methylation can be a reasonable first step. Combined with existing biological knowledge, one can already garner deep insight into the biology at hand and its clinical implications. Illumina®'s HumanMethylation450K BeadChip is an established platform to analyze genome-wide DNA methylation.

Machine learning technologies are usually employed to build classifiers to separate samples into two or more classes or groups depending on their associated data. There are plenty of algorithms available. In addition to the usage as a classifier for new data, the training process can also be used to identify features within the data that work as separators between the groups. Assuming that these separators mirror the underlying biological diffences between the classes, the knowledge gained will help to better understand the molecular processes at work.

Here we apply machine learning technologies to 450K data from two publicly available studies to elucidate the effect of smoking on DNA methylation and look into the associated genetic pathways.

Data sources

Methylation data was downloaded from NCBI's Gene Expression Omnibus resource (https://www.ncbi.nlm.nih.gov/geo). One study (GSE50660, Tsaprouni et al.) contained data from 464 individuals, the other one (GSE85210, Su et al.) 253 individuals.

Applying the machine learning technologies

The data downloaded from GEO was separated into a training- and a test-set containing smokers and non-smokers. Using the WEKA machine learning workbench (https://www.cs.waikato.ac.nz/ml/weka/) an attribute selection (e.g. information gain based feature selection) to reduce the number of attributes for training and testing. We then employed various of the methods available in WEKA to train a classifier for separating smokers from non-smokers. We only used methods where we would be able to assess the contribution of a feature to the classification of a sample to be able to identify relevant methylation sites. The SimpleLogistic classifier method worked nicely, we were able to classify over 90% of the test data correctly.

```
Correctly Classified      182     90.5473 %
Incorrectly Classified     19      9.4527 %

=== Confusion Matrix ===

   a   b   <-- classified as
  20   2 |   a
  17 162 |   b
```

Association with biological knowledge

The methylation site we found to contribute most to the separation of classes was cg05575921, which is associated with the aryl hydrocarbon receptor repressor (AHRR). It has already been shown to be a smoking marker in another study (Zeilinger et al.)

To better understand the underlying mechanism, we examined the transcriptional network around AHRR using our Genomatix Pathway System® (GePS) which provides networks and pathways based on literature annotation. The aryl hydrocarbon receptor repressor regulates the aryl hydrocarbon receptor, which in turn plays a role in regulating antioxidant defense in lung structural cells, where low expression of it might have an impact on the development or progression of COPD (Sarill et al.).

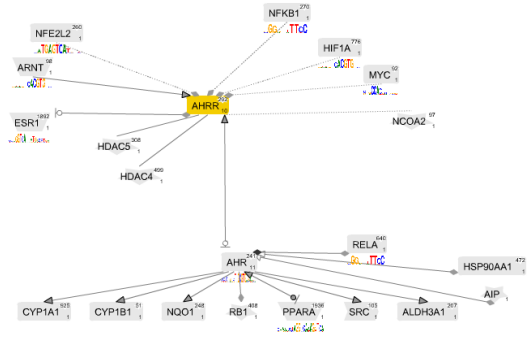

Figure 1: Association of AHRR and AHR. Higher expression of AHHR reduces expression of AHR, which might lead to the obstruction of the answer to oxidative stress in smokers. Figure created using GePS.

Things to watch out for

While the calculation of the beta values generated from the 450k BeadChips is fairly standardized, normalization sometimes seems to differ, which might affect further analyses. Looking at the datasets used in this study, we found the distribution of beta values to be significantly different. Using the wateRmelon R package we re-normalized the GSE50660 dataset for better comparability of values. You should therefore always check the distributions of your values.

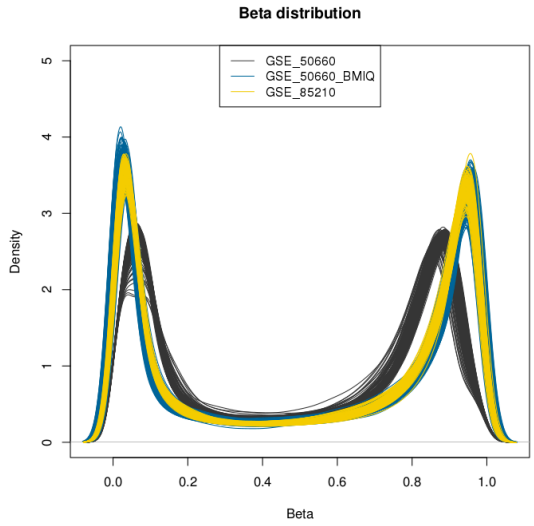

Figure 2: Density plots of beta values from the two experiments. GSE_50660 original data was significantly different from GSE_85210. Normalization with BMIQ led to more compatible datasets.

Confounders that might also skew your analysis and that you need to look out for can be simple things like gender. For example, if there are more female smokers in your study than male, methylation of the X chromosome might show up in your list of separators, although it is not related to smoking. This could be circumvented e.g. by using only women for training or leaving the X and Y chromosome aside for training and testing.

Conclusion

Machine learning approaches can be a valuable tool for the analysis of the large amount of data generated from methylation arrays. Using methods that report the contributing features to a well working classifier, it is possible to gain biological insight by using background annotation data for the selected methylation sites.

For more information on our products or services, please visit **http://www.genomatix.de**. Of course, you're also welcome to contact us at **info@genomatix.de** for any additional questions or inquiries.

References

Tsaprouni LG, Yang T-P, Bell J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9(10):1382-1396. doi:10.4161/15592294.2014.969637.

Su D, Wang X, Campbell MR, et al. Distinct Epigenetic Effects of Tobacco Smoking in Whole Blood and among Leukocyte Subtypes. Costa M, ed. PLoS ONE. 2016;11(12):e0166486. doi:10.1371/journal.pone.0166486.

Eibe F, Hall AM, Witten IH. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Zeilinger S, Kühnel B, Klopp N, et al. Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. Chen A, ed. *PLoS ONE*. 2013;8(5):e63812. doi:10.1371/journal.pone.0063812.

Sarill M, Zago M, Sheridan JA, et al. The aryl hydrocarbon receptor suppresses cigarette-smoke-induced oxidative stress in association with dioxin response element (DRE)-independent regulation of sulfiredoxin 1. Free Radic Biol Med. 2015 Dec;89:342-57. doi: 10.1016/j.freeradbiomed.2015.08.007

Teschendorff AE, Marabita F, Lechner M, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29(2):189-196. doi:10.1093/bioinformatics/bts680.

Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013;14:293. doi:10.1186/1471-2164-14-293.