Discovering methylation biomarkers of smoking and associated health risks

Mohammad W Hattab[1], Robin F Chan[1], Andrey A Shabalin[1], Gaurav Kumar[1], Lin Ying Xie[1], Min Zhao[1], Gerard van Grootheest[2], Karolina A Aberg[1], Brenda W Penninx[2], Edwin JCG van den Oord[1], Shaunna L Clark[1]

[1] Center for Biomarker Research and Precision Medicine, Virginia Commonwealth University, Richmond, VA, USA
[2] Department of Psychiatry, VU University Medical Center / GGZ inGeest, Amsterdam, the Netherlands

Although the adverse effects of smoking are known, the molecular events that underlie these effects remain unclear. Mounting evidence suggests a role for DNA methylation. Because methylation sites can be measured cost-effectively in easy to collect genomic DNA, they are potentially powerful biomarkers that can be used in clinical settings to improve diagnosis, prognosis and monitor response to treatment. To study the effect of smoking on the methylome and identify (panels of) methylation biomarkers for smoking and related health risks, we performed a methylome-wise association study (MWAS) in 1146 individuals from the Netherlands. To investigate all ~28 million CpGs in the human genome, we enriched for the methylated genomic fraction using methyl-CpG binding domain (MBD) protein capture followed by next generation sequencing (MBD-seq). We regressed cigarette years (number of cigarettes per day * number of years smoking) on the methylation measurements to identify individual CpG sites. As the predictive power of biomarkers may be enhanced by combining multiple associated methylation sites, we created multimarker methylation panels using elastic net regression in combination with *k*-fold cross validation to protect against over fitting and ensure unbiased estimates of predictive power.

After we used the highly conservative Bonferroni correction (threshold p-value = $1.15x10^{-08}$), we found 244 CpGs reached significance. Our top individual finding was a CpG located in *AHRR* ($p = 7.88X10^{-33}$). In addition to the *AHRR* finding, we also replicated other robust smoking-methylation associations in *F2RL3* ($p = 5.38x10^{-24}$) and 2q37.1 ($p = 1.24x10^{-32}$), and identified novel associations in *KIF5C* ($p = 4.13x10^{-14}$) and *PPP1R15A* ($p = 6.38x10^{-14}$). Preliminary results indicate that our multimarker methylation panel explains ~31.3% of the variation in cigarette years. We aim to determine if this multimarker panel predicts smoking related health risks using other health biomarkers such as measures of lung function and oxidative stress. Replication using targeted bisulfite sequencing of top findings is ongoing.