Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Weiyan Li Date

SEX-SPECIFIC DIFFERENTIAL DNA METHYLATION RELATED TO CIGARETTE SMOKING IN AFRICAN AMERICANS

By

WEIYAN LI

MPH

EPIDEMIOLOGY

YAN V. SUN

Thesis Advisor

SEX-SPECIFIC DIFFERENTIAL DNA METHYLATION RELATED TO CIGARETTE SMOKING IN AFRICAN AMERICANS

By

WEIYAN LI

M.Sc., Emory University, 2011

B.Sc., Sichuan University, 2008

Thesis Committee Chair: Yan V. Sun, Ph.D

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Public Health

in Epidemiology

2013

Abstract

SEX-SPECIFIC DIFFERENTIAL DNA METHYLATION RELATED TO CIGARETTE SMOKING AFRICAN AMERICANS

By Weiyan Li

Tobacco smoking is a leading cause of mortality and morbidity worldwide. Differential susceptibility to tobacco smoking by sex has been observed in lung cancer patients. However, how tobacco smoking exerts its harmful health effect, especially how it may affect two sexes differently, has not been well understood. DNA methylation is one major mechanism of gene expression regulation that can be modified by environmental exposures such as cigarette smoking. Previous studies have reported smoking-related changes in DNA methylation with a loci-specific resolution. However, little investigation was done on sex-specific effect of smoking on DNA methylation on autosomal sites, and X chromosome sites were routinely excluded in epigenome-wide association studies (EWAS) due to analytical difficulties. In our study, we aim to examine sex-specific effect of tobacco smoking on DNA methylation across the entire epigenome. After adjusting for age, top 5 principal components, relatedness and multiple testing we did not detect any statistically significant sex×smoking interaction throughout autosomal CpG sites. For X chromosomal analysis, we did not detect any CpG sites that was significantly associated with current or ever smoking status after stratification by sex. However, wedetected site cg12857957 located in LZTFL2 (leucine zipper transcription factor-like 1) on chromosome 3 with marginal significance in terms of sex×current smoking status interaction (p value=4.74×10^{-6}). Although further validation is needed, our results suggest that there may be a sex-specific effect of smoking related DNA methylation changes, which may partially explain the differences in susceptibility to tobacco exposure between males and females.

Keywords: cigarette smoking; epigenome; DNA methylation; sex-specific;

SEX-SPECIFIC DIFFERENTIAL DNA METHYLATION RELATED TO CIGARETTE SMOKING IN AFRICAN AMERICANS

By

Weiyan Li

M.Sc., Emory University, 2011

B.Sc., Sichuan University, 2008

Thesis Committee Chair: Yan V. Sun, Ph.D

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology

2013

Table of Contents

Background

Introduction

Smoking is a leading cause of mortality and morbidity. However, detailed mechanism of how smoking exerts its harmful effect has not yet been fully elucidated. Better understanding the pathophysiological process between smoking and its sequelae will provide insights for better designs of intervention strategies for tobacco use-associated diseases.

Epigenetic variations of DNA, particularly DNA methylation at CpG sites, are important molecular machineries to regulate gene expression, genome stability, and DNA repair. Alterations in DNA methylation, both at the global and loci-specific level, have been observed and linked to various human diseases. In cancer, global hypomethylation of repetitive sequence LINE family member L1 has been found in lung, breast, bladder and liver tumor(1). Also, expression of oncogenes can be activated by hypomethylation at the promoter region, as observed for *SNP100* in pancreatic cancer, *SNCG* in breast and ovarian cancer and *DPP6* in melanomas(1, 2). Alterations in DNA methylation has also been linked to autoimmune diseases. Mutations in DNA (cytosine-5-)-methyltransferase 3 beta (DNMT3B) has been established as the cause of the immunodeficiency, centromeric instability and facial anomalies (ICF) syndrome. A wide range of other autoimmune diseases such as systemic lupus erythematosus (SLE)(2-4), multiple sclerosis(5) and rheumatoid arthritis(6-8)

have also shown hyper- or hypomethylation at various genetic loci that mostly located within the promoter region of functional genes.

The human epigenome interacts with the environment and affects disease susceptibility and development. One classic example is the disease concordance rate between monozygotic (MZ) twins, which ranges from 11—25% for SLE(9, 10), 12—22% for RA(11-13) and 9—31% for MS(14-17). Given that MZ twins share identical DNA sequence, one potential explanation for this discrepancy in traits is differences in epigenetic modification induced by environmental exposure. Indeed, reports have shown that MZ twins differ in epigenetic states at birth(18), and this divergence was further enhanced as they progressed into later stage of life(19, 20). However, stochastic drift in epigenetic state can also cause differences in epigenetic profile between MZ twins, and careful examination is needed to determine the proportion of these differences that are truly due to environmental exposure.

Smoking has been associated with altered DNA methylation at both global and gene-specific level. Whereas methods for assessing methylation at candidate gene loci and global DNA methylation level have been available for some time, epigenome-wide microarray technology that provides quantification of DNA methylation levels simultaneously with a locus-specific resolution has become available only recently. This study takes advantage of this new genotyping technology of methylome and builds upon previous epigenetic association study of cigarette smoking, to further investigate the sex-specific smoking-related alterations in DNA methylation in an African American population.

Previous studies conducted in the Sun group have discovered smoking-related, site-specific, differential DNA methylation in autosomal sites. Interestingly, the DNA methylation level further differentiates between males and females. More profound changes in DNA methylation at *IGF2* has been reported in male offsprings and associated with prenatal exposure to cigarette smoking (21). Sex-specific DNA methylation was reported for multiple loci in smokers with or without COPD in an Epigenome-wide association study(22). In hepatocellular carcinoma, 75 CpG sites were found differentially methylated by sex in adjacent non-tumor tissue(23). Higher DNA methylation at *GNASAS* and *INS* was found to be associated with myocardial infarction in women but not men(24). All these recent reports suggest differences in DNA methylation by sex in various disease outcomes. In addition, DNA methylation sites located on X chromosome have been excluded in previous analysis due to an analytical concern that most of these sites have different distribution of methylation between males and females. Such bimodal distribution of DNA methylation is caused by the inactivation of one of the two copies of X chromosome in females (i.e. X chromosome inactivation (XCI)). This study will further investigate in this sex-specific effect on smoking-related differential DNA methylation by (1) examining sex×smoking interaction across autosomal CpG sites; (2) examine smoking-related differential DNA methylation of X chromosomal CpG sites in both sexes.

Tobacco use associated adverse health outcomes

Tobacco smoking is a leading public health concern affecting >1 billion people worldwide and are accounted for an estimated -3 million deaths per year (25). The prolonged impact of tobacco smoking on various malignant (26-30) and non-malignant (31-33) chronic diseases has suggested a role for epigenetic reprogramming. Along with continuous effort in promotion of smoking cessation, elucidating the molecular mechanisms linking tobacco smoke exposure to diseases remains important and may provide potential targets for development of clinical interventions.

Tobacco smoking accounts for 80% lung cancer burden in men and approximately 50% in women(34). As the worldwide trend of lung cancer incidence declining in men but rising in women(34), hypothesis have been formed on a sex difference in the association of tobacco exposure and lung cancer. One possible explanation of this sex difference is acceleration of smoking-related carcinogens through estrogen receptors present in neoplastic lung tissues. This hypothesis is supported by the higher level of PAHs (polycyclic aromatic hydrocarbons)-DNA adducts observed in women(35). However, a recent study with detailed lifetime smoking data showed no support for a higher susceptibility to tobacco smoking related lung cancer in females(36).

Tobacco use associated epigenetic modifications

All somatic cells share the same genetic information in an organism. However, not all genes are expressed at all times in each cell. Epigenetics is the

study of potentially heritable mechanisms that regulates gene expression. Various mechanisms of epigenetic modifications include DNA methylation, histone modifications, nucleosome positioning and non-coding RNAs (37), among which DNA methylation has become the most widely studied form.

DNA methylation occurs mostly at rare CpG dinucleotides (~1%) and tends to cluster into regions known as CpG islands(38, 39). DNA methylation at CpG sites provides extra flexibility(40-42) and complexity(43) to the human genome, and may mediate response to external exposure by regulating gene expression through hypo/hypermethylation(44-47). The global methylation pattern of genetically identical monozygotic twins shows differences at birth(18), and as the individuals age this difference grew(19, 20), reflecting different exposure history.

Widely observed persistent effect of tobacco smoking on pulmonary(48, 49), cardiovascular(50, 51) and malignant diseases(52, 53) has indicated involvement of epigenetic reprogramming in disease etiology. Particularly, alterations in DNA methylation at both global(27, 54-56) and loci-specific(57, 58) level have been associated with various diseases and are one potential mechanism mediating the harm caused by environmental exposure to tobacco smoke. The recent characterization of a human DNA methylome at single nucleotide resolution now make it possible to undertake large-scale epigenomic studies. Maternal smoking during pregnancy (PTS) has been found to be associated with decreased global DNA methylations as measured by repeated sequences including *Sat2*(59), *Alu*, and *LINE1*(60). Site-specific association with

PTS exposure has also been reported for *AXL* and *PTPRO* (60, 61)and *IGF2*(21). A recent epigenome wide scan performed on the Infinium HumanMethylation450 Beadchip identified more than 20 CpG sites mapped to 10 genes as associated with PTS after stringent Bonferroni correction(62). Recently, Breitling *et al.*(57) performed a genome-wide analysis on differential DNA methylation related to tobacco-smoking. The authors examined DNA methylation with loci-specific resolution and genome-wide coverage simultaneously by using the Illumina Infinium HumanMethylation 27 assay, and identified a CpG site located in *F2RL3* gene, which encodes thrombin protease-activated receptor-4 (*PAR4*) and has not been previously recognized in tobacco smoking research. The same group later showed in patients with stable coronary heart disease, lower methylation intensities at *F2RL3* were associated with mortality(63). Wan *et al*(58) adopted similar technology and investigated alterations in DNA methylation associated with current smoking, cumulative smoke exposure and time since smoking cessation; they not only detected *F2RL3* but also a novel locus (*GPR15*) as being significant in all three analyses from their COPD cohort. A more recent epigenome-wide association study using 450K platform also replicated the finding on *F2RL3* along with several other novel sites(64).

Recent discoveries relating tobacco smoke exposure to altered DNA methylation at specific loci has been largely restricted to Caucasian population. Given the natural occurring genetic/epigenetic variations among different ethnicity groups, whether similar phenomenon exists in other population

remains an important research question. To explore and confirm the dynamic DNA methylation modification in response to adult tobacco smoke exposure, the Sun group conducted an epigenome-wide screen for CpG sites with differential DNA methylation related to cigarette smoking in an African American population and confirmed previous findings in Caucasian populations(65). However, due to random XCI in females, DNA methylation pattern at X chromosome differs significantly by sex. Our previous studies on autosomal CpG sites showed different methylation pattern by sex. Interestingly, a recent study reported more profound prenatal smoking related differential DNA methylation at *IGF2* in male offsprings(21) compared to females. To address the sex-specific effect of cigarette smoking on DNA methylation, we examined the sex×smoking interaction for autosomal sites and performed stratified (by sex) analysis for X chromosome sites.

Methods

Study Population

The study population consists of African American individuals initially recruited from Jackson, Mississippi for the Genetic Epidemiology Network of Arteriopathy (GENOA) study. The GENOA cohort was initially ascertained through sibships in which at least two siblings were diagnosed with essential hypertension before the age of 60 and all siblings were invited to participate in the study. The main exposure variable, cigarette smoking, was collected from a self-reported questionnaire as "smoked within the past year", "not smoked within the past year", and "never smoked". Two binary variables, "current smoker" (1 for individuals smoked within the past year) and "ever smoker" (1 for individuals ever smoked more than 100 cigarettes) were derived from the original smoking status accordingly. The number of pack-years smoking was also calculated based on the number of years of smoking history and the average amount of cigarette smoking. Age, sex and other phenotypic data were collected from the physical examination and laboratory assessment at the time of the Phase II study visit. The GENOA study was approved by the Institutional Review Boards of all participating institutions. Each participant gave written informed consent.

DNA Methylation Data

Genomic DNA was extracted from stored peripheral leukocytes (PLCs) of 1,008 GENOA Phase II African American (AA) participants, bisulfite converted

and then epityped for methylation profiling of 27,578 CpG loci using the Illumina Infinium HumanMethylation27 BeadChip (Illumina, San Diego, CA). On each Illumina Infinium HumanMethylation27 BeadChip, there are 56 control probes representing a) sample independent measures of processing procedures: staining, hybridization, target removal, and DNA extension and b) sample dependent measures of bisulfite conversion, G/T mismatch, non-polymorphic and negative controls(65). The sample independent controls allow for the evaluation of the quality of the chip processing steps, while the sample dependent controls allow for the evaluation of the performance across samples. Due to poor quality of the intensity measurements of control probes, a total of 49 probes were excluded from the analysis(65). The cleaned data set included DNAm profiles of 972 AA individuals from 493 sibships.

Sites with control probe values greater than 4 standard deviations from their mean values were removed. In addition, a normalization scheme to reduce batch and chip effects by linearly regressing the methylated and non-methylated intensity signals onto the set of control probes that are orthogonal (i.e. independent) predictors of the control spot distributions(65). We also excluded sites that are located on the X and Y chromosomes for autosomal analysis. Finally, 2,984 sites that have non-specific binding probes (over 10% of the probes map to highly homologous genomic sequences at 40 or more of their base pairs) were removed(65, 66). Another 875 sites that have probes overlapping with SNPs reported in dbSNP were also excluded due to possible influence the methylation levels reported by the microarray(65). These nonspecific and polymorphic sites

were filtered out after all analyses to facilitate the interpretation of results. The data set used in the final analysis included 22,927 for autosomal analysis. For X chromosomal analysis, we selected the 1085 CpG site that mapped to the X chromosme. Also, as the linear regression to control for batch and chip effects mentioned above is not applicable to sites with bimodal distribution, we used raw beta values for analysis for X chromosome analysis.

Statistical Analysis

The signal intensities from methylated (M) and unmethylated (U) bead types were used to calculate a beta value as $\beta = M/(U + M)$ for each DNAm site in each individual. The lambda inflation factor were calculated as lambda=median[obs]/median[exp](67). We modeled DNAm status using beta values as the dependent variable, and binary variables "current smoker", "ever smoker" and "packyears" defined above as the primary independent variables in the association analysis. For autosomal analysis, we assessed sex-specific effect of smoking by adding a sex×smoking interaction term. Females were coded as 1 and males as 0. We adopted the linear mixed model to adjust for the relatedness (i.e. sibship) among our analysis cohort and included age and sex as potential confounders. Since many X chromosomal DNAm sites do not fit unimodal distribution with pooled-sex, we first stratified the data by sex and then performed similar regression analysis adjusting for age and relatedness. A Bonferroni-corrected p-value of 0.05 or False Discovery Rate (FDR) q-value of 0.05 were applied to adjust for multiple testing of 22,927 autosomal or 1085 X chromosomal DNAm sites. We calculated the top principal components of

22,927 autosomal CpG sites to adjust for potential confounders, and to control for the inflation of low p-values in the EWAS. The beta-values of each CpG site were first standardized by subtracting the mean beta-value. Using the standardized beta value, the top principal components and the corresponding eigen vectors were calculated for each individual using the princomp package in R.

All statistical analyses were performed in the R statistical environment version 2.12.1.

Results

A total of 972 AA participants from the GENOA study were included in this analysis. Among these participants, 58.3% had never smoked ("never smoker"), 29.1% had smoked but not within one year before the visit (coded as "former smoker"), and 12.6% had smoked within the past one year ("current smoker"). Males (29.3%) have higher rates of smokers, both former and current, compared to females (Table 1). Females had significantly higher BMI than males. Age is not significantly different between males and females. At the time of the visit, 70.9% of males and 75.8% of females were on antihypertensive medications.

The analysis cohort is consisted of 972 related individuals of 493 different sibships. As shown in Table 2, the size of sibship varies from 1—10, with over 50% of participants from sibships of 2 to 3 individuals.

Sex-specific effect of smoking on autosomal DNAm sites

The main predictor variable, smoking, was characterized by three different methods: current vs. non-current smoking, ever vs. never smoking and pack-years of smoking. We first assessed sex-specific effect of smoking by looking at sex×smoking interaction for autosomal DNAm sites. For all analysis on autosomal DNA methylation sites, age and sex were both included as potential confounders as well as top 5 principal components calculated as mentioned in methods section, and a linear mixed regression model was applied to control for the relatedness among our participants. A total of 972 participants were included in all autosomal analysis.

Current smoking vs. Non-current smoking

We generated a dichotomous covariate for current smoking based on the questionnaire data. Participants who have answered "Yes" to "smoked within the past year" were coded 1 for "current smoking", while others were coded 0 regardless of whether they had recently quitted smoking (in the past year) or had never smoked. Sex as a binary variable ("1" for females) was also derived from the questionnaire. There is no missing data on either sex or current smoking status for all 972 individuals included in this analysis. Figure 1 shows the distribution of linear mixed regression P values for each autosomal DNAm locus for the sex×current smoking interaction term. One site, cg12857957 located in *LZTFL2* (leucine zipper transcription factor-like 1), deviated from the pattern. However, $P_{sex\times current\ smoking}$ was 4.74×10^{-6} and did not pass the 0.05 alpha cutoff line after stringent Bonferroni correction (Bonferroni corrected p value=0.11).

We also performed similar regressions with either top 3 or 10 principal components added into the model. Although after correcting for top 10 principal components site cg12857957 did yield a p-value for the interaction term that passed our threshold (p value=9.72×10^{-7}, Bonferroni corrected p value=0.02), the inflation factor from the top 10 principal components analysis was higher (inflation factor=1.113) compared to that of the top 5 principal components analysis (inflation factor=1.054), suggesting a more sever global inflation. Therefore, we decided to report results from the model that controlled for the top 5 principal components. Since current smoking status is considered as the most robust exposure among all smoking variables and thus treated as our main

model, we applied top 5 principal components analysis to all association analysis with other smoking variables (as discussed below) for consistency purpose, though in those particular models top 5 principal components analysis may not necessarily give the best distribution of p values or lowest inflation factor.

Because differential DNAm between sexes was suggested in previous studies smoking without the interaction term (as discussed in Background), we also specifically looked at the DNAm sites that showed top 15 associations from the previous studies to determine if there is any statistical significant interaction by sex. As shown in table 3, although for all these 15 sites DNAm was significantly associated with current smoking status in previous studies(65) none showed a significant sex by current smoking interaction model.

Ever smoking v.s Never smoking

Participants who had ever smoked more than 100 cigarettes were considered as ever smokers (coded as "1") while people who had smoked less than 100 cigarettes or never at all were considered as never smokers (coded as "0"). Figure 2 shows the distribution of linear mixed p values of sex×ever smoking interactions controlling for age, sex and top 5 principal components. No sites showed clear deviation from the pattern and no significant p values were detected after either Bonferroni or FDR correction.

Pack-years of smoking among ever smokers

To investigate whether the accumulated effect of smoking on DNA methylation differs by sex, we performed a linear mixed regression using pack-

years of smoking as the main predictor outcome and added a sex×pack-years of smoking interaction term to the model. Given we were trying to assess the accumulated effect, rather than the dose-dependent effect of smoking, all 405 ever smokers (Table 1) were included in the analysis. Figure 3a shows the distribution of linear mixed p values of sex×pack-years of smoking interaction controlling for age, sex and top 5 principal components. One DNA methylation site, cg03491181 located in *FKBP3* (FK506 binding protein 3, 25kDa), deviated from the pattern with a p value of 1.18×10^{-7} that passed the epigenome-wide cutoff point after Bonferroni correction p value=$0.05/22927=2.18\times10^{-6}$. However, after closer examination, the observed association at this particular site was mainly driven by one individual with extreme high level of both pack-years of smoking and DNAm, and exclusion of this individual leads to a null association.

Sex-specific effect of smoking on X chromosome DNAm sites

Similar to autosomal DNAm sites analysis, we treated smoking as a dichotomous covariate for the X chromosome sites analysis. Linear mixed model as described above in methods section were applied to 1085 X chromosomal DNAm sites. Participants were first stratified by sex before applying the regression model to account for different DNAm distribution of X chromosome CpG sites. In the male stratum, a total of 285 individuals were included in the analysis compared to 687 for the female stratum.

Current smoking v.s Non-current smoking

Figure 4 and 5 shows the distribution of p values of the current smoking term after controlling for age. Clear deviation from the diagonal line starts from the lower left corner of the plot, together with an inflation factor of 5.548 suggest strong inflation. Results from the males suggest more inflation compared to females. The sites with the lowest p values were cg05327750 (nominal p value: 0.0005) and cg09283007 (nominal p value: 0.0007) from males and females, respectively. However, among the 1085 X chromosome DNAm sites tested, none showed a p value that passed the Bonferroni corrected p value=$0.05/1085=4.61\times10^{-5}$.

Ever smoking v.s Never smoking

Figure 6 and 7 shows the distribution of p values of the ever-smoking term after controlling for age. In contrast to current smoking analysis where males showed inflation of a much higher magnitude compared to females, results generated from male ever smokers were deflated (inflation factor = 0.7724). Ever smoking analysis in females is still inflated (inflation factor = 1.169), though not as strongly as in previous analysis on X chromosome current smoking analysis. The sites with the lowest p values were cg04906538 (nominal p value: 0.002) and cg07897414 (nominal p value: 0.001) from males and females, respectively. Still, none of the 1085 sites yielded a p value for ever smoking status that passed the Bonferroni corrected threshold.

Discussion

To investigate whether the association between DNA methylation and cigarette smoking is significantly different by sex, we performed an epigenome-wide association study to examine sex-specific effect of cigarette smoking on DNA methylation across autosomal and X chromosomal DNAm sites. After controlling for age and top 5 principal components, no sex×smoking association passed the stringent Bonferroni corrected threshold across autosomal DNAm sites. No significant associations between X chromosome DNAm and smoking exposure was detected. However, we detected site cg12857957 located in *LZTFL1* with marginal significant p values for sex×current smoking status interaction. In the analysis of X chromosome CpG sites analysis controlled for age, inflation remains a main issue, possibly due to insufficient control of batch and chip effects.

The results for site cg12857957 across sex×current smoking models were around the significance threshold but not robust in models with various number of principal components. With top 10 principal components added to the model, cg12857957 yielded a p value for the sex×current smoking term of 9.72×10^{-7}(Bonferroni correct p value: 0.02). However, in models with top 3 and 5 principal components, the same site did not show significant interaction of DNAm and current smoking status by sex (top 3 PCA model: nominal p value=1.18×10^{-5}, Bonferroni corrected p value=0.27; top 5 PCA model: p value=4.74×10^{-6}, Bonferroni corrected p value=0.11). We performed principal components analysis (PCA) to control for undetected confounders and inflation

of low p values due to EWAS study design. However, there is no consensus process of choosing the best PCA model. We chose top 5 PCA in our report based on the low inflation factor and reasonable global distribution of p values observed from the Q-Q plot. However, this analysis strategy may not be the best correction for this specific site cg12857957 and either inflation or deflation may still leads to the observed results. Also, since this site is relatively sensitive to PCA analysis, it is possible that the association it shows is actually due to undetected confounders that PCAs are mainly dealing with. The site cg12857957 located in gene *LZTFL1* and it has not been linked it to tobacco exposure. Studies on schizophrenia showed differential methylation at cg12857957 based on data that come from peripheral blood cells and acquired by the Illumina Infinium HumanMethylation27 BeadChip array(68). The *LZTFL1* gene encodes a ubiquitously expressed protein, leucine zipper transcription factor-like 1, which localizes to the cytoplasm where it interacts with Bardet-Biedl Syndrome (BBS) proteins and affects ciliary functions by regulating protein trafficking. Although this protein is more frequently studied for BBS, it may also serve as a tumor suppressor by interacting with cytoskeleton and E-cadherin(69).

Strengths and limitations

One of the main strength in this study is that we used stringent Bonferroni correction to minimize type 1 error, which is a common concern in EWAS due to the issue of multiple testing. Also, we controlled for undetected confounders using principal component analysis. Moreover, we have 3 different set of smoking variables, each had a different mechanistic focus that allowed us to assess

current, past and accumulated smoking exposure and their association with DNA methylation respectively.

Our study examined whether the sex-specific susceptibility to tobacco exposure could be partially mediated by sex-specific susceptibility to smoking-related changes in DNA methylation. Although there has been report on sex-specific susceptibility to tobacco exposure, little investigation has been done to elucidate the mechanisms mediating this phenomenon. Higher risks of lung cancer were found for female smokers compared to their male counterpart after controlling for the level of smoking, and the differences were more profound in younger age(70, 71). Whether the differences in smoking-related risk is caused by a biologic or behavioral mechanism is still under debate. DNA methylation, as an important intermediate layer between environmental exposure and biologic consequences, has shown to be altered by smoking behavior and may also capture this differential effect of smoking between the sexes. In a recent study on prenatal exposure to cigarette smoke and methylation changes in offspring, more profound changes in DNAm were found in male offsprings at *IGF2* (insulin-like growth factor 2) DMR(21). The impact of smoking on DNAm on X chromosome may be different in men and women, as one copy of X chromosome in females undergo random inactivation and become highly methylated. However, due to different DNAm patterns in males and females, X chromosome sites are commonly excluded in previous analysis. Our study filled in the gap and looked at sex-specific effect of smoking on DNAm in both autosomal and X chromosomal sites. The lack of significant interaction by sex as reported in this

study, could be due to a relatively small sample size and small effect size that made it statistically difficult to identify potential associations without *a priori* hypothesis. Also, we have not yet performed principal components analysis or controlled for batch and chip effects for X chromosome, which could mask the true association.

Another limitation in our study is the limited statistical power due to large number of multiple testing, relatively small sample size and small potential effect size. However, this limitation is the trade-off for the hypothesis-generating study design, and is true to most other genome-wide or epigenome-wide association studies as well. Possible solutions to this issue includes performing meta-analysis to increase sample size, or including only extreme cases (heavy smoker versus never smoker) to increase the potential effect size aimed to be detected.

DNA methylation patterns are dynamic and vary significantly among different cell types. This leads to ne major concern about using peripheral leukocytes (PLCs) to obtain DNA methylation data: it may not reflect the DNA methylation pattern of target disease tissue. In our case, since we were not focusing on a particular disease and cigarette smoking has been shown to have a global health impact, PLCs serves as a good surrogate of biologic changes we look for. Also, as the PLCs are a mixture of cell population usually with distribution of each subtype unknown, changes in DNAm may reflect lineage rather than inter-individual differences(72). Traditional methods of determining cell subtype proportions such as Flow cytometry is not feasible due to both high cost and limited quantity of human biological samples. Fortunately, recent studies have

showed promising results on re-constitution of cell mixture distribution using array data(73), which can provide finer measurements of DNA methylation changes and increase the power to detect any meaningful differences.

Future directions

One analysis that can be implemented to improve the quality of our results is principal components analysis on X chromosome sites, which allows control of batch and chip effects underlying our seemly uniform ethnic group in addition to adjustment of unmeasured confounders. It would also be extremely helpful to perform validation analysis of our top association sites (sites with smallest p values) in an independent replication cohort, given that the small number of tests we would perform would require less over correction by stringent Bonferroni methods. Also, facilitated by the wide adoption of the 450K platform, replication of our findings in another cohort with data acquired from a different assay can serve as an even more convincing validation of findings in our discovery cohort of GENOA using 27K platform.

Through out this study we focused DNA methylation at CpG islands using array-based technology, which covers only a small fraction of the methylome compared to approximately 28 million CpG sites in the human genome(74, 75). Moreover, these selected sites are restricted to previously known CpG islands. In contrast, next generation sequencing based measurement of epigenetic profile provides finer resolution and more complete measurement of DNA methylation, requires no previous knowledge on DNAm sites, and can detect methylation of non-CG context. Being more cost-effective in population studies, these

sequencing-based technology will provide a more comprehensive examination of the epigenome and enable us to detect alterations in epigenetic profile that help explains smoking-related disease etiology.

Reference

1. Wilson AS, Power BE, Molloy PL. DNA hypomethylation and human diseases. *Biochim Biophys Acta* 2007;1775(1):138-62.
2. Irizarry RA, Ladd-Acosta C, Wen B, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* 2009;41(2):178-86.
3. Lu Q, Kaplan M, Ray D, et al. Demethylation of ITGAL (CD11a) regulatory sequences in systemic lupus erythematosus. *Arthritis Rheum* 2002;46(5):1282-91.
4. Kaplan MJ, Lu Q, Wu A, et al. Demethylation of promoter regulatory elements contributes to perforin overexpression in CD4+ lupus T cells. *J Immunol* 2004;172(6):3652-61.
5. Mastronardi FG, Noor A, Wood DD, et al. Peptidyl argininedeiminase 2 CpG island in multiple sclerosis white matter is hypomethylated. *J Neurosci Res* 2007;85(9):2006-16.
6. Neidhart M, Rethage J, Kuchen S, et al. Retrotransposable L1 elements expressed in rheumatoid arthritis synovial tissue: association with genomic DNA hypomethylation and influence on gene expression. *Arthritis Rheum* 2000;43(12):2634-47.
7. Nile CJ, Read RC, Akil M, et al. Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis. *Arthritis Rheum* 2008;58(9):2686-93.
8. Takami N, Osawa K, Miura Y, et al. Hypermethylated promoter region of DR3, the death receptor 3 gene, in rheumatoid arthritis synovial cells. *Arthritis Rheum* 2006;54(3):779-87.
9. Jarvinen P, Kaprio J, Makitalo R, et al. Systemic lupus erythematosus and related systemic diseases in a nationwide twin cohort: an increased prevalence of disease in MZ twins and concordance of disease features. *J Intern Med* 1992;231(1):67-72.
10. Deapen D, Escalante A, Weinrib L, et al. A revised estimate of twin concordance in systemic lupus erythematosus. *Arthritis Rheum* 1992;35(3):311-8.
11. Bellamy N, Duffy D, Martin N, et al. Rheumatoid arthritis in twins: a study of aetiopathogenesis based on the Australian Twin Registry. *Ann Rheum Dis* 1992;51(5):588-93.

12. Aho K, Koskenvuo M, Tuominen J, et al. Occurrence of rheumatoid arthritis in a nationwide series of twins. *J Rheumatol* 1986;13(5):899-902.
13. Silman AJ, MacGregor AJ, Thomson W, et al. Twin concordance rates for rheumatoid arthritis: results from a nationwide study. *Br J Rheumatol* 1993;32(10):903-7.
14. Ebers GC, Bulman DE, Sadovnick AD, et al. A population-based study of multiple sclerosis in twins. *N Engl J Med* 1986;315(26):1638-42.
15. Mumford CJ, Wood NW, Kellar-Wood H, et al. The British Isles survey of multiple sclerosis in twins. *Neurology* 1994;44(1):11-5.
16. Kinnunen E, Juntunen J, Ketonen L, et al. Genetic susceptibility to multiple sclerosis. A co-twin study of a nationwide series. *Arch Neurol* 1988;45(10):1108-11.
17. Sadovnick AD, Armstrong H, Rice GP, et al. A population-based study of multiple sclerosis in twins: update. *Ann Neurol* 1993;33(3):281-5.
18. Gordon L, Joo JH, Andronikos R, et al. Expression discordance of monozygotic twins at birth: effect of intrauterine environment and a possible mechanism for fetal programming. *Epigenetics* 2011;6(5):579-92.
19. Wong CC, Caspi A, Williams B, et al. A longitudinal study of epigenetic variation in twins. *Epigenetics* 2010;5(6):516-26.
20. Talens RP, Boomsma DI, Tobi EW, et al. Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. *FASEB J* 2010;24(9):3135-44.
21. Murphy SK, Adigun A, Huang Z, et al. Gender-specific methylation differences in relation to prenatal exposure to cigarette smoke. *Gene* 2012;494(1):36-43.
22. Genome-Wide DNA Methylation Marks Cluster By Gender In Smokers With And Without COPD In The International COPD Genetics Network. *A94 COPD: FACTORS AND INTERVENTIONS AFFECTING ADHERENCE AND OUTCOMES*: American Thoracic Society:A2256-A.
23. Shen J, Wang S, Zhang YJ, et al. Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. *Epigenetics* 2013;8(1):34-43.
24. Talens RP, Jukema JW, Trompet S, et al. Hypermethylation at loci sensitive to the prenatal environment is associated with increased incidence of myocardial infarction. *Int J Epidemiol* 2012;41(1):106-15.

25. Peto R, Lopez AD, Boreham J, et al. Mortality from smoking worldwide. *British medical bulletin* 1996;52(1):12-21.

26. Moore LE, Pfeiffer RM, Poscablo C, et al. Genomic DNA hypomethylation as a biomarker for bladder cancer susceptibility in the Spanish Bladder Cancer Study: a case-control study. *The lancet oncology* 2008;9(4):359-66.

27. Smith IM, Mydlarz WK, Mithani SK, et al. DNA global hypomethylation in squamous cell head and neck cancer associated with smoking, alcohol consumption and stage. *International journal of cancer Journal international du cancer* 2007;121(8):1724-8.

28. Oka D, Yamashita S, Tomioka T, et al. The presence of aberrant DNA methylation in noncancerous esophageal mucosae in association with smoking history: a target for risk diagnosis and prevention of esophageal cancers. *Cancer* 2009;115(15):3412-26.

29. Tekpli X, Zienolddiny S, Skaug V, et al. DNA methylation of the CYP1A1 enhancer is associated with smoking-induced genetic alterations in human lung. *International journal of cancer Journal international du cancer* 2011.

30. Luo J, Margolis KL, Wactawski-Wende J, et al. Association of active and passive smoking with risk of breast cancer among postmenopausal women: a prospective cohort study. *BMJ* 2011;342:d1016.

31. Braber S, Henricks PA, Nijkamp FP, et al. Inflammatory changes in the airways of mice caused by cigarette smoke exposure are only partially reversed after smoking cessation. *Respiratory research* 2010;11:99.

32. Rea TD, Heckbert SR, Kaplan RC, et al. Smoking status and risk for recurrent coronary events after myocardial infarction. *Annals of internal medicine* 2002;137(6):494-500.

33. Willemse BW, ten Hacken NH, Rutgers B, et al. Effect of 1-year smoking cessation on airway inflammation in COPD and asymptomatic smokers. *The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology* 2005;26(5):835-45.

34. Jemal A, Bray F, Center MM, et al. Global cancer statistics. *CA Cancer J Clin* 2011;61(2):69-90.

35. Cote ML, Yoo W, Wenzlaff AS, et al. Tobacco and estrogen metabolic polymorphisms and risk of non-small cell lung cancer in women. *Carcinogenesis* 2009;30(4):626-35.

36. De Matteis S, Consonni D, Pesatori AC, et al. Are women who smoke at higher risk for lung cancer than men who smoke? *Am J Epidemiol* 2013;177(7):601-12.

37. Ptashne M. On the use of the word 'epigenetic'. *Curr Biol* 2007;17(7):R233-6.
38. Portela A, Esteller M. Epigenetic modifications and human disease. *Nat Biotechnol* 2010;28(10):1057-68.
39. Crews D, McLachlan JA. Epigenetics, evolution, endocrine disruption, health, and disease. *Endocrinology* 2006;147(6 Suppl):S4-10.
40. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews Genetics* 2008;9(6):465-76.
41. van der Maarel SM. Epigenetic mechanisms in health and disease. *Annals of the rheumatic diseases* 2008;67 Suppl 3:iii97-100.
42. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009;461(7265):747-53.
43. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;462(7271):315-22.
44. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. *Nature reviews Genetics* 2002;3(6):415-28.
45. Feinberg AP, Tycko B. The history of cancer epigenetics. *Nature reviews Cancer* 2004;4(2):143-53.
46. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature* 2007;447(7143):433-40.
47. Feinberg AP. Epigenetics at the epicenter of modern medicine. *JAMA : the journal of the American Medical Association* 2008;299(11):1345-50.
48. Bouloukaki I, Tsiligianni IG, Tsoumakidou M, et al. Sputum and nasal lavage lung-specific biomarkers before and after smoking cessation. *BMC pulmonary medicine* 2011;11:35.
49. Willemse BW, ten Hacken NH, Rutgers B, et al. Effect of 1-year smoking cessation on airway inflammation in COPD and asymptomatic smokers. *The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology* 2005;26(5):835-45.
50. Conen D, Everett BM, Kurth T, et al. Smoking, smoking cessation, [corrected] and risk for symptomatic peripheral artery disease in women: a cohort study. *Annals of internal medicine* 2011;154(11):719-26.

51. Kawachi I, Colditz GA, Stampfer MJ, et al. Smoking cessation and decreased risk of stroke in women. *JAMA : the journal of the American Medical Association* 1993;269(2):232-6.

52. Wilhelm-Benartzi CS, Christensen BC, Koestler DC, et al. Association of secondhand smoke exposures with DNA methylation in bladder carcinomas. *Cancer causes & control : CCC* 2011;22(8):1205-13.

53. Mani S, Szymanska K, Cuenin C, et al. DNA methylation changes associated with risk factors in tumors of the upper aerodigestive tract. *Epigenetics : official journal of the DNA Methylation Society* 2012;7(3):270-7.

54. Terry MB, Ferris JS, Pilsner R, et al. Genomic DNA methylation among women in a multiethnic New York City birth cohort. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2008;17(9):2306-10.

55. Flom JD, Ferris JS, Liao Y, et al. Prenatal smoke exposure and genomic DNA methylation in a multiethnic birth cohort. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2011;20(12):2518-23.

56. Furniss CS, Marsit CJ, Houseman EA, et al. Line region hypomethylation is associated with lifestyle and differs by human papillomavirus status in head and neck squamous cell carcinomas. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2008;17(4):966-71.

57. Breitling LP, Yang R, Korn B, et al. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *American journal of human genetics* 2011;88(4):450-7.

58. Wan ES, Qiu W, Baccarelli A, et al. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Human molecular genetics* 2012.

59. Flom JD, Ferris JS, Liao Y, et al. Prenatal smoke exposure and genomic DNA methylation in a multiethnic birth cohort. *Cancer Epidemiol Biomarkers Prev* 2011;20(12):2518-23.

60. Breton CV, Byun HM, Wenten M, et al. Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation. *Am J Respir Crit Care Med* 2009;180(5):462-7.

61. Breton CV, Salam MT, Gilliland FD. Heritability and role for the environment in DNA methylation in AXL receptor tyrosine kinase. *Epigenetics* 2011;6(7):895-8.

62. Joubert BR, Haberg SE, Nilsen RM, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect* 2012;120(10):1425-31.

63. Breitling LP, Salzmann K, Rothenbacher D, et al. Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease. *Eur Heart J* 2012;33(22):2841-8.

64. Shenker NS, Polidoro S, van Veldhoven K, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Hum Mol Genet* 2013;22(5):843-51.

65. Sun YV SA, Conneely KN, Chang Q, Li W, Binder EB, Klengel T, Cross D, Turner ST, Ressler KJ, Kardia SL. Epigenomic Association Analysis Identifies Smoking-related DNA Methylation Sites in African Americans. *Human Genetics (Accepted)* 2013.

66. Chen YA, Choufani S, Ferreira JC, et al. Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray. *Genomics* 2011;97(4):214-22.

67. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55(4):997-1004.

68. Nishioka M, Bundo M, Koike S, et al. Comprehensive DNA methylation analysis of peripheral blood cells derived from patients with first-episode schizophrenia. *J Hum Genet* 2013;58(2):91-7.

69. Wei Q, Zhou W, Wang W, et al. Tumor-suppressive functions of leucine zipper transcription factor-like 1. *Cancer Res* 2010;70(7):2942-50.

70. Brownson RC, Chang JC, Davis JR. Gender and histologic type variations in smoking-related risk of lung cancer. *Epidemiology* 1992;3(1):61-4.

71. Harris RE, Zang EA, Anderson JI, et al. Race and sex differences in lung cancer risk associated with cigarette smoking. *Int J Epidemiol* 1993;22(4):592-9.

72. Reinius LE, Acevedo N, Joerink M, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS one* 2012;7(7):e41361.

73. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* 2012;13:86.

74. Bibikova M, Le J, Barnes B, et al. Genome-wide DNA methylation profiling using Infinium(R) assay. *Epigenomics* 2009;1(1):177-200.

75. Bibikova M, Fan JB. Genome-wide DNA methylation profiling. *Wiley Interdiscip Rev Syst Biol Med* 2010;2(2):210-23.

Tables

Table 1. Characteristics of the study population			
Variable	**Male (n=285) N (%)**	**Female (n=687) N (%)**	**Total (n=972) N (%)**
Smoking			
Never	93 (32.63%)	474 (69.00%)	567 (58.33%)
Not in last year	135 (47.37%)	148 (21.54%)	283 (29.12%)
Smoker within 1	57 (20.00%)	65 (9.46%)	122 (12.55%)
Ever smoking			
Yes	192 (67.37%)	213 (31.00%)	405 (41.67%)
No	93 (32.63%)	474 (69.00%)	567 (58.33%)
Antihypertention			
Yes	202 (70.88)	521 (75.84%)	723 (74.38%)
No	83 (29.12%)	166 (24.16%)	249 (25.62%)
	Mean (SD)	**Mean (SD)**	**Mean (SD)**
Age	66.75 (7.68)	66.09 (7.56)	66.28 (7.60)
Pack-years of smoking	30.43(24.31)[a]	20.53 (17.64)[b]	25.22 (21.61)[c]
Blood pressure average			
Diastolic	80.24 (11.17)	77.54 (10.84)	78.33 (11.00)
Systolic	137.75 (20.79)	140.86 (21.62)	139.95 (21.42)
BMI	28.95 (4.83)[d]	32.06 (6.58)[e]	31.15 (6.28)[f]

a. n=213; b. n=192; c. n=405; d. Three males missing BMI data; e. One female missing BMI data; f. Four individuals missing BMI data.

Table 2. Summary table for sibship size.		
Sibship size	**Count**	**Number of**
10	1	10
7	1	7
6	4	24
5	7	35
4	31	124
3	71	213
2	181	362
1	197	197
Total	493	972

Table 3. Sex × Current smoking interaction in top 15 DNAm sites previously identified from no interaction model.

DNAm	Gene	ΔBeta	SE	t-stat	p-value
cg00353953	ZNF384	0.0051	0.0068	0.7612	0.4469
cg01500140	LIM2	-0.0036	0.0054	-0.6697	0.5034
cg03330058	ABTB1	0.0125	0.0133	0.9373	0.3491
cg03340878	OR2B6	-0.0095	0.0075	-1.2608	0.2080
cg03636183	F2RL3	0.0098	0.0124	0.7926	0.4284
cg04983977	GPR25	0.0002	0.0073	0.0260	0.9793
cg11314684	AKT3	-0.0070	0.0068	-1.0270	0.3050
cg13500388	CBFB	-0.0016	0.0084	-0.1846	0.8536
cg13633560	LRRC32	-0.0108	0.0070	-1.5445	0.1232
cg13668129	HNRPUL1	-0.0058	0.0042	-1.3647	0.1730
cg13745870	SPATA12	0.0074	0.0062	1.1996	0.2309
cg14223444	NCBP1	0.0024	0.0079	0.3020	0.7628
cg17791651	POU3F1	-0.0092	0.0143	-0.6462	0.5184
cg19859270	GPR15	-0.0026	0.0037	-0.7117	0.4770
cg26259865	LOC124220	-0.0132	0.0068	-1.9398	0.0530

Figures

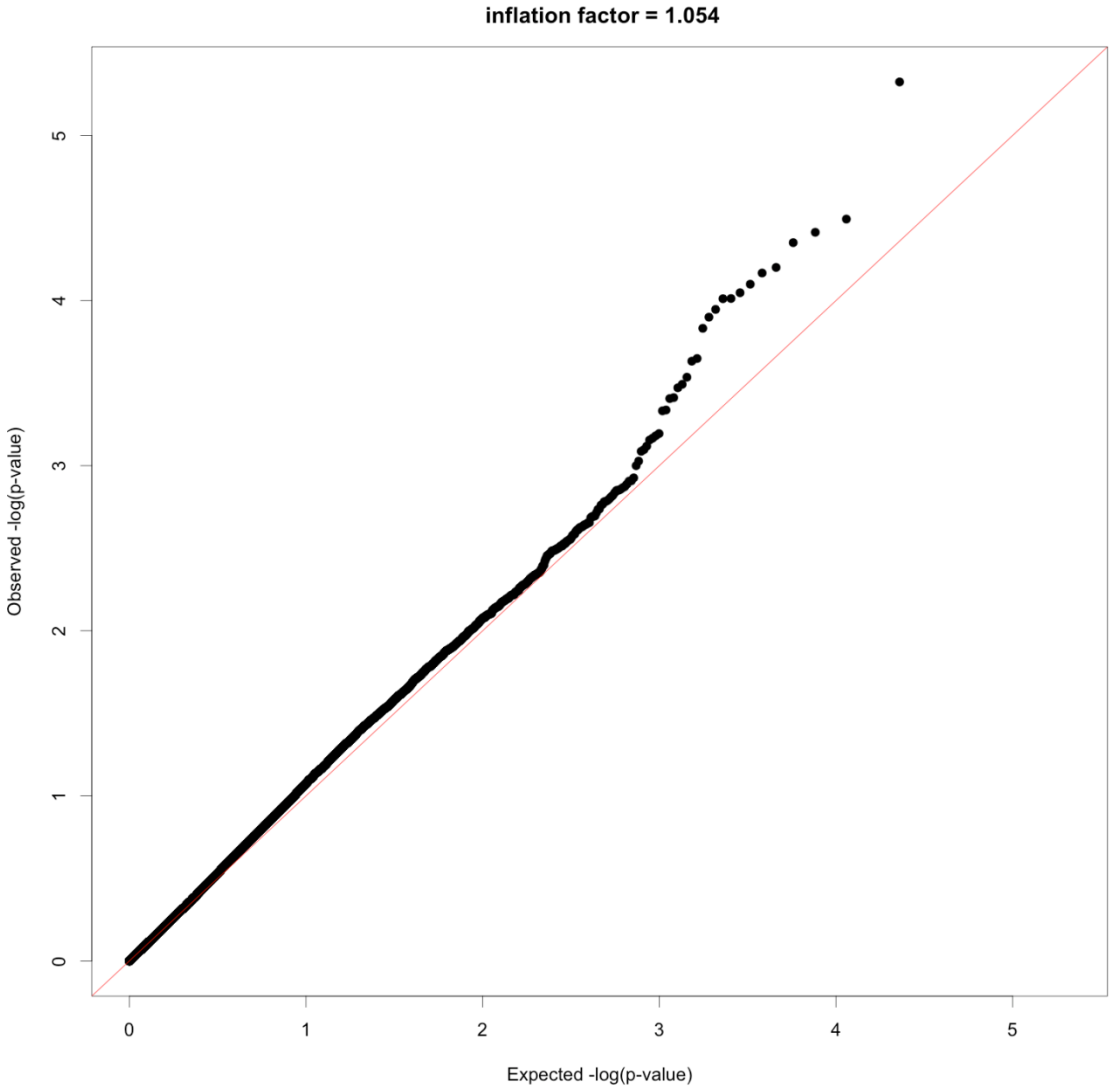

Figure 1. Linear mixed regression p values of the sex×current smoking interaction term adjusted for sex, age and top 5 principle components.

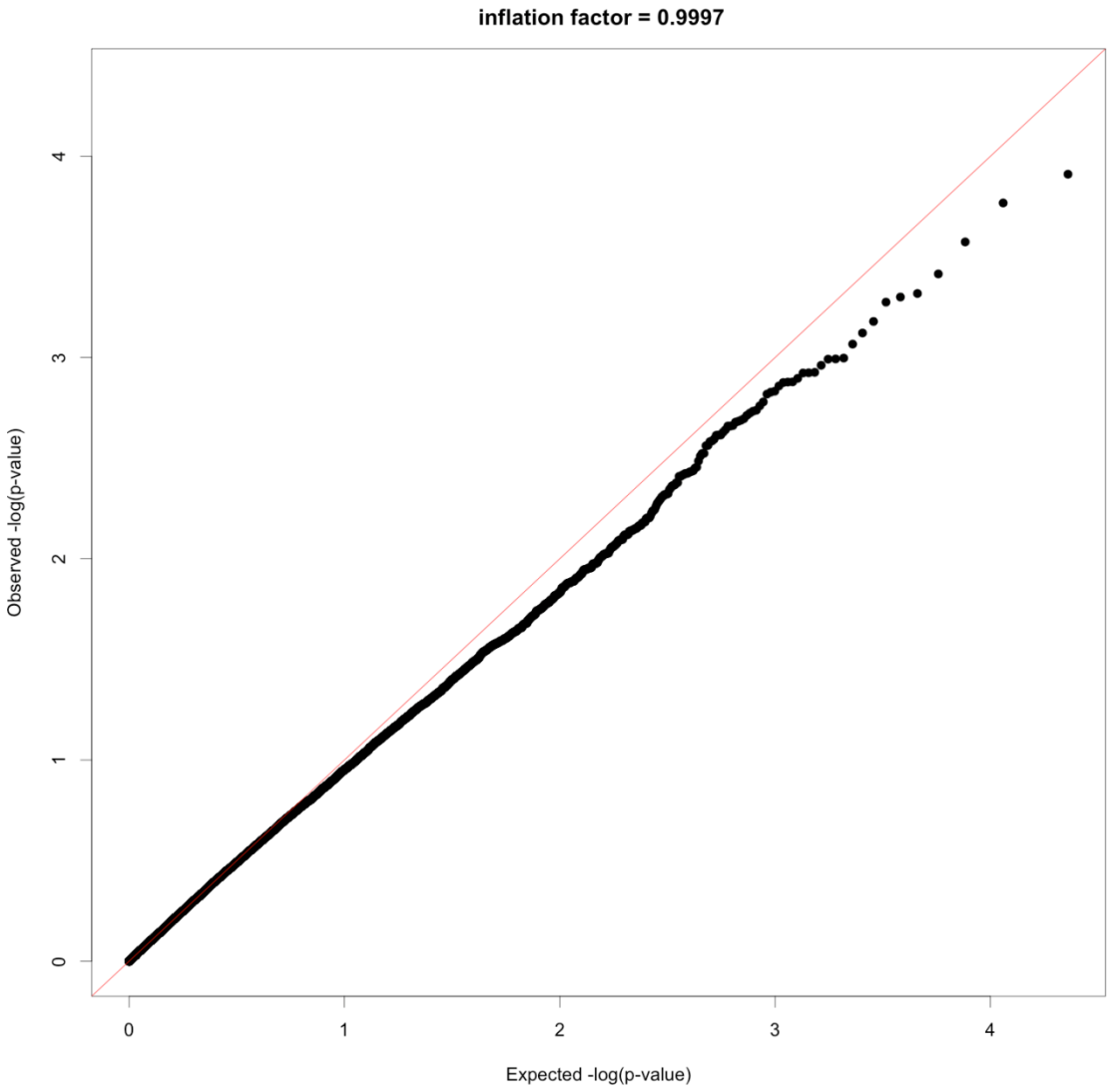

Figure 2. Linear mixed regression p values of the sex×ever smoking interaction term adjusted for sex, age and top 5 principle components.

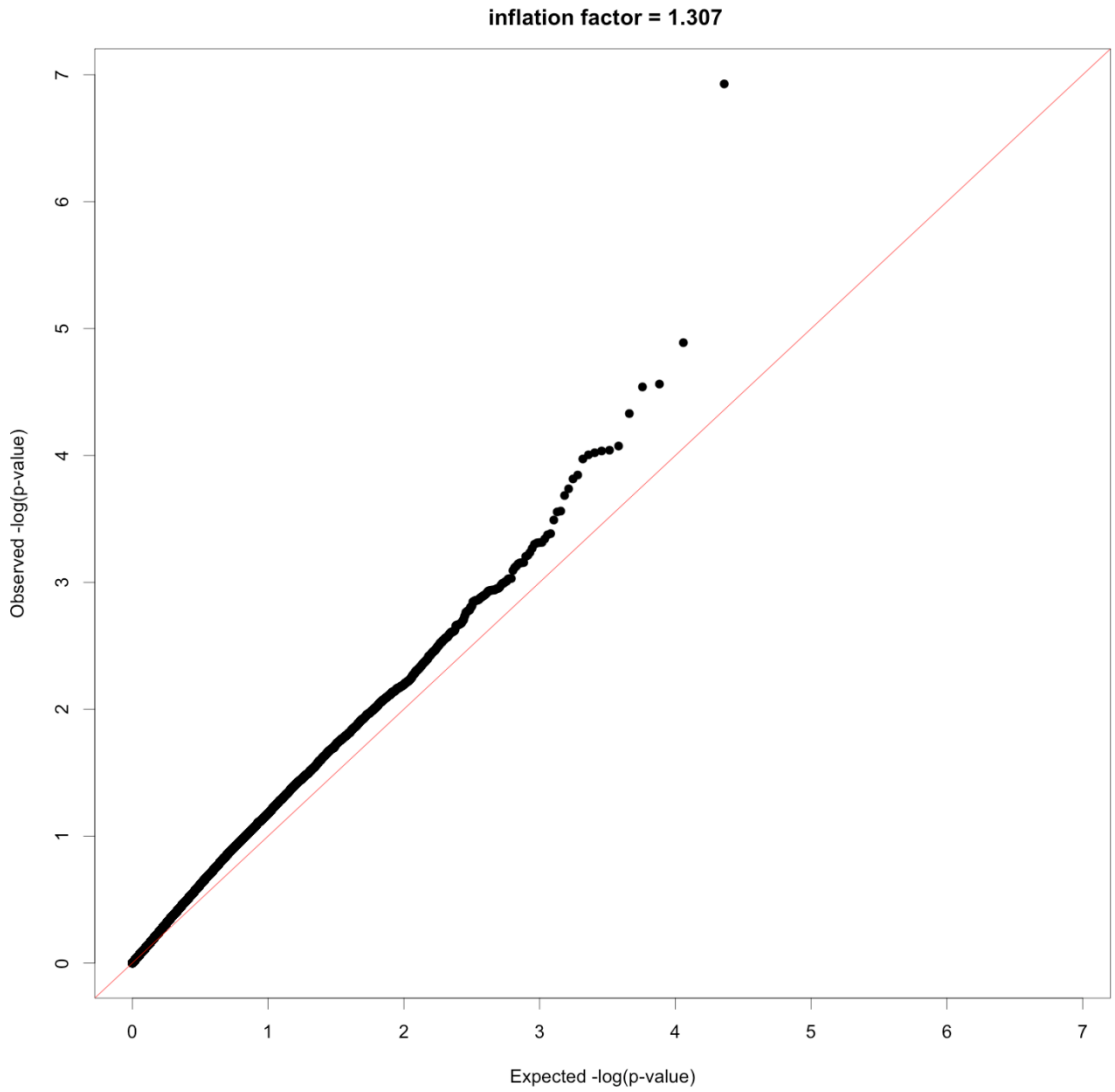

Figure 3. Linear mixed regression p values of the sex×pack-years of smoking interaction term adjusted for sex, age and top 5 principle components.

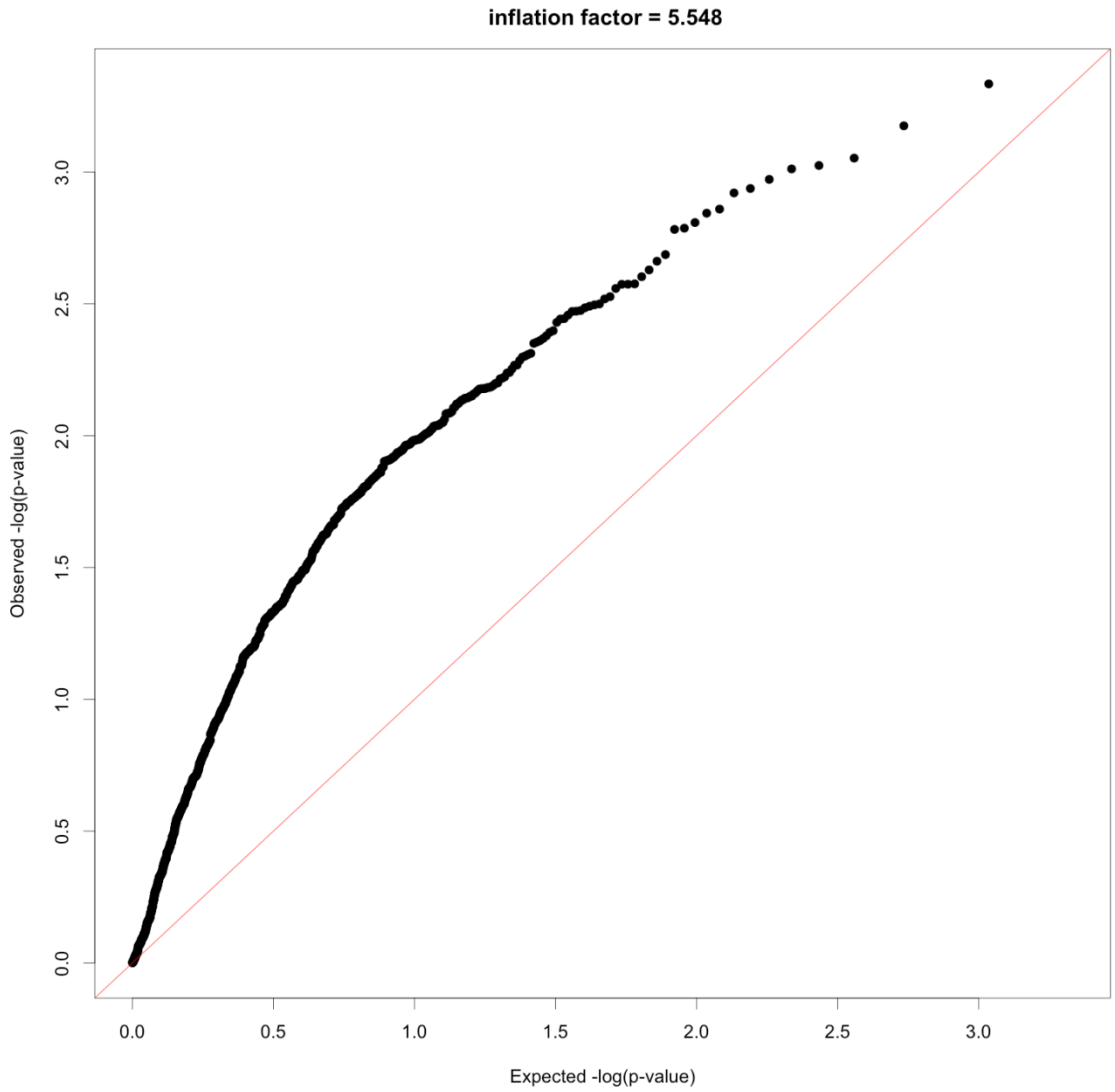

Figure 4. Linear mixed regression p values of the current smoking term for X chromosome CpG sites in males, adjusted for age.

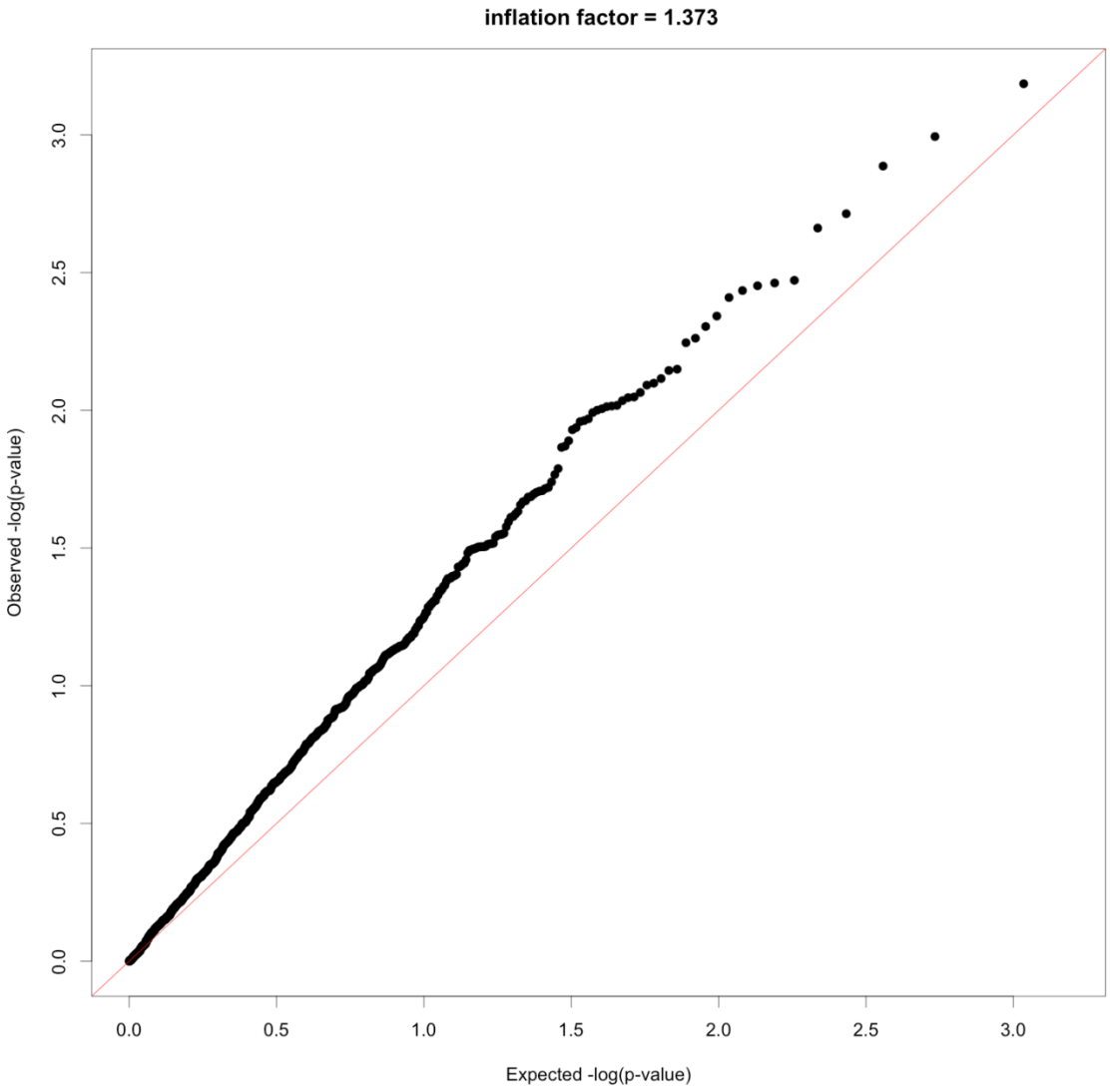

Figure 5. Linear mixed regression p values of the current smoking term for X chromosome CpG sites in females, adjusted for age.