2019

Determining population stratification and subgroup effects in association studies of rare genetic variants for nicotine dependence

Ai-Ru Hsieh

Li-Shiun Chen

Ying-Ju Li

Cathy S. J. Fann

Determining population stratification and subgroup effects in association studies of rare genetic variants for nicotine dependence

Ai-Ru Hsieh[a], Li-Shiun Chen[b], Ying-Ju Li[c] and Cathy S.J. Fann[c]

Background **Rare variants (minor allele frequency < 1% or 5 %) can help researchers to deal with the confounding issue of 'missing heritability' and have a proven role in dissecting the etiology for human diseases and complex traits.**

Methods **We extended the combined multivariate and collapsing (CMC) and weighted sum statistic (WSS) methods and accounted for the effects of population stratification and subgroup effects using stratified analyses by the principal component analysis, named here as 'str-CMC' and 'str-WSS'. To evaluate the validity of the extended methods, we analyzed the Genetic Architecture of Smoking and Smoking Cessation database, which includes African Americans and European Americans genotyped on Illumina Human Omni2.5, and we compared the results with those obtained with the sequence kernel association test (SKAT) and its modification, SKAT-O that included population stratification and subgroup effect as covariates. We utilized the Cochran–Mantel–Haenszel test to check for possible differences in single nucleotide polymorphism allele frequency between subgroups within a gene. We aimed to detect rare variants and considered population stratification and subgroup effects in the genomic region containing 39 acetylcholine receptor-related genes.**

Results **The Cochran–Mantel–Haenszel test as applied to *GABRG2* ($P = 0.001$) was significant. However, *GABRG2* was detected both by str-CMC ($P = 8.04E-06$) and str-WSS ($P = 0.046$) in African Americans but not by SKAT or SKAT-O.**

Conclusions **Our results imply that if associated rare variants are only specific to a subgroup, a stratified analysis might be a better approach than a combined analysis. *Psychiatr Genet* 29:111–119 Copyright © 2019 The Author(s). Published by Wolters Kluwer Health, Inc.**

Psychiatric Genetics 2019, **29**:111–119

Keywords: acetylcholine receptor-related genes, nicotine dependence, population stratification, principal component analysis, subgroup effects, rare variant

[a]Graduate Institute of Biostatistics, China Medical University, Taichung, Taiwan, [b]Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri, USA and [c]Institute of Biomedical Sciences, Academia Sinica, Nankang, Taipei, Taiwan

Correspondence to Cathy Shen-Jang, Fann, PhD, Institute of Biomedical Sciences Academia Sinica, Office 101, Taipei, Taiwan
Tel: 886-2-27899144; fax: 886-2-27823047;
e-mail: csjfann@ibms.sinica.edu.tw

Received 4 June 2018 **Accepted** 29 March 2019

Background

Cigarette smoking is a primary risk factor for many chronic diseases (Bergen and Caporaso, 1999) including many cancers, diabetes, cardiovascular disease, and chronic lung disease (Fang *et al.*`, 2014). Recent candidate-gene association studies (Fang *et al.*, 2014) and genome-wide association studies (GWASs) (Thorgeirsson *et al.*, 2010), many of which have been reviewed by Wang and Li (2010), have searched for and, at varying levels of significance, identified common variants associated with measures of response to tobacco, tobacco consumption, nicotine dependence, nicotine metabolism, and smoking cessation.

Current smoking prevalence is similar in European Americans and African Americans (Centers for Disease and Prevention, 2008; Saccone *et al.*, 2010; Choi *et al.*, 2017). Nicotine dependence is common in both groups, with evidence of slightly lower levels of dependence in African Americans by standard measures such as cigarettes per day currently in use (Breslau *et al.*, 2001; Saccone *et al.*, 2010). Smoking cessation rates, however, are lower in African Americans compared with European Americans (Breslau *et al.*, 2001; Covey *et al.*, 2008; Saccone *et al.*, 2010; Choi *et al.*, 2017). Furthermore, there is evidence that African Americans have a higher risk of dependence at lower cigarettes-per-day levels compared with European Americans (Luo *et al.*, 2008; Saccone *et al.*, 2010). Also important are the disparities in health consequences from smoking: African Americans have higher lung cancer incidence and mortality than European Americans (Haiman *et al.*, 2006; Jemal *et al.*, 2008; Saccone *et al.*, 2010). An understanding of the genetic loci involved, and their effects and allele frequencies in diverse populations, can

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, *www.psychgenetics.com*.

0955-8829 Copyright © 2019 The Author(s). Published by Wolters Kluwer Health, Inc. DOI: 10.1097/YPG.0000000000000227

provide important clues to the risk of developing nicotine dependence across all populations. Multiple nicotinic receptor subunit genes outside of chromosome 15q25 are likely to be important in the biological processes and development of nicotine dependence, and some of these risks may be shared across diverse populations (Saccone *et al.*, 2010).

GWASs constitute an important means for identifying risk genes for complex human diseases, such as diabetes (Hindorff *et al.*, 2009; Dajani *et al.*, 2017), heart disease (Hofker *et al.*, 2014; Erdmann *et al.*, 2018), and Alzheimer's disease (Shen and Jia, 2016; Chung *et al.*, 2018), among others. Despite many successes in identifying risk alleles, most associated variants discovered through GWAS do not account for the majority of heritability estimated for these complex human diseases and traits. From a genetics perspective, by far the most studied of these complex human diseases and traits can be attributed to heritability about an estimated 60–80% of human disease (Lichtenstein *et al.*, 2009), whereas GWAS have identified only 5–10% of this heritability, leading many researchers to ponder which alleles underlie the missing heritability (Manolio *et al.*, 2009; Zuk *et al.*, 2014; Auer and Lettre, 2015). One of the approaches to deal with missing heritability is to detect new DNA variants, especially rare variants that have a relatively large impact on disease etiology, that is, those with minor allele frequency <5%, which may also contribute to complex disease (Schork *et al.*, 2009; Marian, 2012; Gusev *et al.*, 2013; Zuk *et al.*, 2014; Auer and Lettre, 2015; Ma *et al.*, 2015; Nicolae, 2016). With efforts from the 1000 Genomes Project, which sought to identify most rare genetic variants in a group of 1092 multiethnic individuals, a new generation of GWAS is being designed to enable the discovery of rare variants using next-generation sequencing data (Abecasis *et al.*, 2012; Sampson *et al.*, 2012). Hence, improved technologies for discovering rare variants provide a possible means of explaining the missing heritability.

A number of methods have been developed for identifying associations between rare variants and common diseases (Li and Leal, 2008; Madsen and Browning, 2009; Schork *et al.*, 2009). Madsen and Browning (2009) proposed a weighted sum statistic (WSS) method that assigns weights to variants according to their frequency in controls such that the variants with lower frequencies have greater weights. Li and Leal (2008) proposed the combined multivariate and collapsing (CMC) method for case-control data. Wu *et al.* (2011) proposed a sequence kernel association test (SKAT), that is, a variance-component method that aggregates individual variant-score test statistics. However, population structure and subgroups can be strong confounding factors in association studies (Pritchard *et al.*, 2000; Ziv and Burchard, 2003; Clayton *et al.*, 2005; Roeder and Luca, 2009), and thus accounting for population structure and subgroups is crucial even when seemingly homogeneous ethnic populations are sampled. To our knowledge, only a few articles have discussed rare-variant detection and considered population stratification and subgroup effects [e.g., reviewed by Moore *et al.* (2013), O'Connor *et al.* (2013), Wang *et al.* (2015), Prokopenko *et al.* (2016)] for nicotine dependence and smoking cessation studies (Saccone *et al.*, 2010). However, whether population stratification would be better than dealt with using stratified analyses or including population simply as a covariate has not been studied enough (Culverhouse *et al.*, 2011).

To achieve this goal, we evaluated the issue by considering two situations: (1) assessing the strata in separate analyses and (2) pooling data from all strata, using population as a covariate. The results from the two situations were then compared. We utilized the Cochran–Mantel–Haenszel (CMH) test to check whether the allelic distribution of single nucleotide polymorphisms (SNPs) is similar between the population stratifications/subgroups. Furthermore, we extended WSS and CMC to identify rare variants while also considering population stratification and subgroup effects using stratified analyses by principal component analysis (PCA), named here as 'str-CMC' and 'str-WSS'. To compare results obtained with the two aforementioned situations for nicotine dependence and smoking cessation studies, we analyzed a smoking cessation dataset to test for rare variants associated with nicotine dependence which was downloaded from the Database of Genotypes and Phenotypes (accession number phs000404.v1.p1). The smoking cessation dataset was from the Collaborative Genetic Study of Nicotine Dependence (COGEND; principal investigator: Laura Bierut) and the University of Wisconsin Transdisciplinary Tobacco Use Research Center (UW-TTURC; principal investigator: Timothy Baker). Evidence has recently accumulated that SNPs in the genetic region encoding the nicotinic acetylcholine receptor (nAChR) subunits α6, α5, α3, and β4 are associated with smoking and nicotine dependence (Russo *et al.*, 2011). For the smoking cessation dataset analyses, we were only interested in the acetylcholine receptor region that has been reported previously.

To evaluate the issue by considering two situations, we compared the results to those obtained with two variance-component methods, namely, SKAT (Wu *et al.*, 2011) and optimal sequencing kernel association test (SKAT-O) (Lee *et al.*, 2012), which treat both population stratification and subgroup effects in the PCA as covariates. In our results, we found ethnicity (i.e., African American and European American) was associated with the first axis of variation (PC1) arising from PCA (Supplementary Additional file 2, Supplemental digital content 2, *http://links.lww.com/PG/A221*). Our results imply that if a gene showed allele frequency differences between the two groups, it would be better to use str-CMC or str-WSS in detecting associated rare variants. By contrast, if a gene has a similar distribution of allele frequency between the two groups, this would be better

dealt with by including population stratification and subgroup effects as covariates in SKAT or SKAT-O. These results will assist researchers in identifying a biological basis for the etiology of nicotine dependence.

Materials and methods

The CMC and WSS cannot be adjusted for covariates. However, the SKAT and SKAT-O are able to adjust for covariates (Liang and Xiong, 2013). We evaluated rare variant methods for dealing with population subgroups: (1) CMC and WSS were analyzed population subgroups in population stratification analyses, that is, str-MSS and str-WSS and (2) SKAT and SKAT-O were combined data from all population subgroups, using population subgroups as a covariate. First, we calculated the first axis of variation (PC1) using the EIGENSTRAT software (Price *et al.*, 2006) to consider population stratification and subgroup effects. PCA is a linear dimensionality reduction technique used to infer continuous axes of genetic variation. Price *et al.* (2006) developed the program EIGENSTRAT to correct for population structure in association tests. It uses the top eigenvectors of the sample covariance matrix as covariates in a regression setting. Second, we performed the CMH test (Mantel and Haenszel, 1959) to assess differences in SNP allele frequencies between subgroups by the results of the PC1 arising from PCA. The CMH test was two-tailed for all analyses. To investigate the homogeneity association assumption, we used the Breslow–Day test and found no significant evidence for heterogeneity association. Third, we detected rare variants and accounted for the effects of population stratification and subgroup effects. Rare genetic variants, here defined as alleles with a frequency less than 1–5% (Wu *et al.*, 2011). For all rare variant methods, rare variants were detected within a gene, a minor allele frequency of less than 5% was used as the rare-variant criterion.

Smoking cessation data

Our analyses were based on a publicly available smoking cessation dataset from the Database of Genotypes and Phenotypes (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap) (accession number phs000404.v1.p1). Genotyping during the GWAS discovery phase used the HumanOmni2.5 BeadChip designed to analyze 2 443 179 loci. All individuals (n = 1515) in the study were from two projects: COGEND (principal investigator: Laura Bierut) and UW-TTURC (principal investigator: Timothy Baker). The individuals reported smoking at least 10 cigarettes per day. Both International Classification of Diseases 10th Revision (ICD-10) (Janca *et al.*, 1993) and Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) have separate categories for dependent and nondependent smokers. The ICD-10 and the DSM-IV are unsatisfactory and are rarely used for daily clinical care because they cannot be tailored treatments to individual needs (Helzer *et al.*, 2006; Rüther *et al.*, 2014). However, the Fagerström Test for Nicotine Dependence (FTND) measures tobacco dependencies as dimension parameters and uses continuums to indicate the severity of the dependency. Thus, the FTND has become an internationally recognized and proven method for determining tobacco dependence (Heatherton *et al.*, 1991; Rüther *et al.*, 2014). For this reason, we binned cases and controls based on the FTND when evaluating smoking cessation.

Both COGEND and UW-TTURC projects assessed nicotine dependence using the FTND (Heatherton *et al.*, 1991). The FTND is a six-item self-report measure of nicotine dependence. FTND scores on the scale range from 0 to 10 and also categorized accordingly: 2 = very low dependence; 3–4 = low dependence; 5 = moderate dependence; 6–7 = high dependence; and 8+ = very high dependence (López-Torrecillas *et al.*, 2017). For the current study, cases were defined as having a nicotine dependence if the score for this test was at least 6; all controls had a score of or less 4. To avoid potential rare variant detection biases associated with misclassification of FTND scores due to FTND breakpoints (López-Torrecillas *et al.*, 2017), our study ignored participants with an FTND score of 5 and analyzed two groups of participants with large differences in FTND score. According to this definition, there were 923 cases (135 African Americans and 788 European Americans) and 592 controls (69 African Americans and 523 European Americans).

For this dataset, we were only interested in the acetylcholine receptor region that has been reported previously as being candidate genes of smoking cessation (Conti *et al.*, 2008). According to Conti *et al.* (2008), several studies have identified associations of genetic regions encoding the nAChRs with nicotine dependence (Saccone *et al.*, 2007) and with smoking cessation (Berrettini *et al.*, 2007). Therefore, the nAChRs list was analyzed by Ingenuity Pathways Analysis was performed (IPA; Ingenuity H Systems, Redwood City, CA, USA; (http://www.ingenuity.com) to explore the possibility of identifying gene candidates previously reported in the literature findings from Ingenuity Knowledge Base. All possibility of identifying gene candidates searched from IPA were listed in Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A220*.

Combined multivariate and collapsing with population stratification and subgroup effects (str-CMC)

CMC (Li and Leal, 2008) aggregates multiple rare variants across a genomic region (e.g., gene, haplotype, and pathway) and analyzes them together. CMC divides markers into subgroups based on predefined criteria (e.g., allele frequency) and, within each group, marker data are collapsed into an indicator variable. The procedure we used consisted of the following four steps: (1) data were divided into subgroups by the first axis of variation (PC1) using EIGENSTRAT software (Price *et al.*, 2006); (2) markers in each gene group were classified as either rare variants

or common variants; (3) markers in each gene group are divided into subgroups on the basis of predefined criteria (e.g., allele frequencies), and within each group, marker data are collapsed into indicator variables defined for the genotype at the *i*th variant site for the *j*th individual in the case population (X_{ji}) and control population (Y_{ji}), respectively: $X_{ji}=\begin{cases}1 & \text{Genotype is AA}\\ 0 & \text{Genotype is Aa}\\ -1 & \text{Genotype is aa}\end{cases}, Y_{ji}=\begin{cases}1 & \text{Genotype is AA}\\ 0 & \text{Genotype is Aa}\\ -1 & \text{Genotype is aa}\end{cases}$ as described in Li *et al.*, 2008 (Li and Leal, 2008); and (4) Hotelling's T2 test was used to compare the groups of marker data in each k-gene group. This procedure was named as 'str-CMC'.

Weighted sum statistic with population stratification and subgroup effects (str-WSS)

Madsen and Browning (2009) described WSS, which determines a weighted rare-variant count in a genomic region (e.g., gene, haplotype, and pathway). The weights are determined according to the variance of the allele frequency estimated for cases and controls, with down-weight mutation counts in constructing the genetic score as bellow. The procedure we use consisted of the following five steps: (1) data were divided into subgroups by the PC1 using EIGENSTRAT software (Harvard University, USA; https://www.hsph.harvard.edu/alkes-price/software/) (Price *et al.*, 2006); (2) a set of markers were divided into *k* genomic regions; (3) the genetic score as described in Madsen and Browning (2009) was calculated for each gene. Madsen and Browning (2009) defined the genetic score as follows: $S_j=\sum_{i=1}^{L}\frac{I_{ij}}{\widehat{w}_i}$ where I_{ij} is the number of mutations (usually this will be the minor allele, unless common allele was reported susceptibility to disease) in variant *i* for individual *j* in a genomic region, *L* is the number of variants genotyped, and *i* = 1, 2,…, *L*. The weight, $\widehat{w}_i=\sqrt{n_i\cdot q_i(1-q_i)}$, is the estimated SD of the total number of mutations in the sample (including cases and controls), under the null hypothesis of no frequency differences between cases and controls, where $q_i=\frac{m_i^U+1}{2n_i^U+2}$, m_i^U is the number of mutant alleles observed for variant *i* in the controls, n_i^U is the number of controls genotyped for variant *i*, and n_i is the total number of individuals genotyped for variant *i* (cases and controls); (4) genetic scores were ranked for the cases and controls combined; and (5) a Wilcoxon rank sum test was used to test for association between the set of rare variants and disease status via permutation tests. This procedure was named 'str-WSS'.

Sequence kernel association test-based methods

SKAT (Wu *et al.*, 2011) and SKAT-O (Lee *et al.*, 2012) use a multiple regression model to directly correlate a phenotype with genetic variants in a genomic region (e.g., gene, haplotype, and pathway) and with covariates by the PC1 using EIGENSTRAT software (Price *et al.*, 2006).

Results

A total of 2 295 169 SNPs in chromosomes 1–22 were excluded according to the following quality-control criteria: genotype call rate < 0.95, or departure from Hardy–Weinberg equilibrium ($P < 10^{-4}$) for the control group. Missing SNPs were imputed using Beagle: University of Washington, USA; https://faculty.washington.edu/browning/beagle/beagle.html. Beagle produces a measure r^2 to estimate the squared correlation between imputed and true alleles for the marker. For quality control (QC) purpose, we excluded SNPs with r^2 less than 0.3. Finally, we used 1785 SNPs in acetylcholine receptor region that has been reported previously as being candidate genes of smoking cessation (Conti *et al.*, 2008)

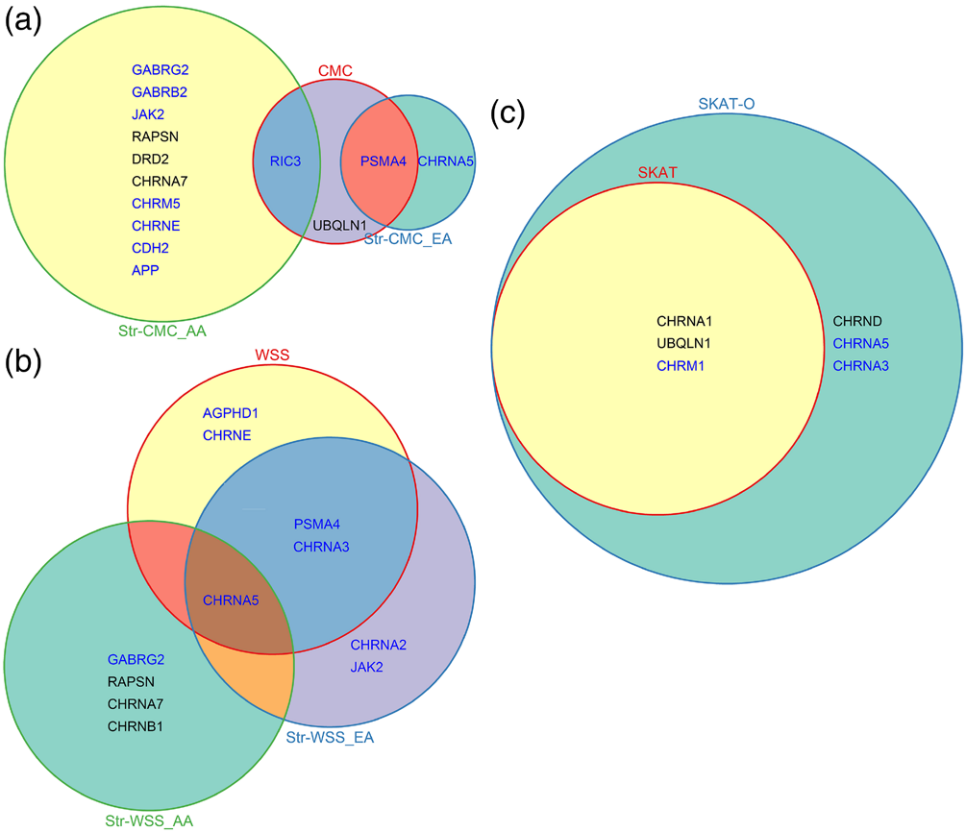
str-CMC

After using EIGENSTRAT software (Price *et al.*, 2006), we found that ethnicity was the first axis of variation (PC1) arising from PCA. Hence, we divided the data into two subgroups, that is, groups European Americans and African Americans.

The acetylcholine receptor genes *CHRNA5* (cholinergic receptor, nicotinic, alpha 5) (*P* = 0.038) and *PSMA4* (proteasome subunit alpha 4) (*P* = 0.019) of group European Americans were identified by str-CMC (Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*). *CHRNA5* is associated with risk of failure for individuals who attempt to reduce cigarette smoking (Chen *et al.*, 2014) and contributes to lung cancer susceptibility in smoking-associated nasopharyngeal carcinoma (Ji *et al.*, 2014). *PSMA4* is associated with lung cancer risk in Caucasians and African Americans (Hansen *et al.*, 2010).

The following genes of group African Americans were identified by str-CMC (Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*): *GABRB2* (gamma-aminobutyric acid A receptor, beta 2) (*P* = 2.20E-05), *GABRG2* (gamma-aminobutyric acid A receptor, gamma 2) (*P* = 8.04E-06), *JAK2* (Janus kinase 2) (p=6.58E-07), *DRD2* (dopamine receptor D2) (*P* = 2.92E-04), *RAPSN* (receptor-associated protein of the synapse) (*P* = 0.031), *RIC3* (RIC3 acetylcholine receptor chaperone) (*P* = 0.005), *CHRM5* (cholinergic receptor, muscarinic 5) (*P* = 0.03), *CHRNA7* (cholinergic receptor, nicotinic, alpha 7) (*P* = 5.49E-06), *CHRNE* (cholinergic receptor, nicotinic, epsilon) (*P* = 0.0176), *CDH2* (cadherin 2, type 1, N-cadherin) *P* < 0.00000001), and *APP* [amyloid beta (A4) precursor protein] (*P* < 0.00000001). *GABRB2* is associated with susceptibility to drug addiction (Hondebrink *et al.*, 2013), psychiatric disorders (Zhao *et al.*, 2012), and nonsmall cell lung cancer (Zhang *et al.*, 2013). *GABRG2* is also associated with epilepsy (Reinthaler *et al.*, 2015) and may contribute to the potential for suicidal behavior in schizophrenia patients with alcohol dependence or abuse (Zai *et al.*, 2014). *PTK2B* is associated with nonsmall

Fig. 1



(a) The Venn diagram of rare variants detected by str-CMC in European Americans (str-CMC_European Americans), African Americans (str-CMC_African Americans) and CMC. A color scheme of gene symbol is used to display CMH results with blue for significance, black for no significance. (b) The Venn diagram of rare variants detected by str-WSS in European Americans (str-CMC_European Americans), African Americans (str-WSS_African Americans) and WSS. (c) The Venn diagram of rare variants detected by SKAT and SKAT-O. CMC, combined multivariate and collapsing; WSS, weighted sum statistic.

cell lung cancer (Kuang *et al.*, 2013). Mutations in *JAK2*, when considered in the context of cigarette smoking status, can affect breast cancer–specific mortality (Slattery *et al.*, 2014). Although Choi *et al.* (2015) did not find a significant relationship between *DRD2* polymorphisms and success during smoking cessation therapy, *DRD2* was found to be associated with nicotine and alcohol addiction (Ma *et al.*, 2015). *RIC3* is associated with nicotinic receptor assembly, expression, and nicotine-induced receptor upregulation (Dau *et al.*, 2013). *CHRM5* may be involved in addiction to tobacco and cannabis, but not alcohol, in group European Americans (Anney *et al.*, 2007). *CHRNA7* may be involved in the development of physical dependence on nicotine (Kishioka *et al.*, 2014).

str-WSS

The following genes of group European Americans were found by str-WSS (Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*): *CHRNA2* (P = 0.038), *JAK2* (P = 0.013), *CHRNA3* (P = 0.011), *CHRNA5* (P = 0.002), and *PSMA4* (P = 0.004). *CHRNA2* is associated with nicotine dependence in groups European Americans and African Americans (Wang *et al.*, 2014) and with smoking cessation (Heitjan *et al.*, 2008). *CHRNA3* is associated with nicotine dependence (Munafò *et al.*, 2011), and *CHRNA3* polymorphisms are genetic modifiers of the risk for developing lung adenocarcinoma (He *et al.*, 2014). However, *CHRNA3* may not merely operate as a marker for the difficulty, willingness, or motivation to quit smoking (Munafò *et al.*, 2011).

The following genes were identified by str-WSS in group African Americans (Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*): *GABRG2* (P = 0.046), *RAPSN* (P = 0.011), *CHRNA5* (P = 0.039), *CHRNA7* (P = 0.016), and *CHRNB1* (P = 0.046), whereas *CHRNB1* is associated with the African Americans sample, no significant association was found in group European Americans (Lou *et al.*, 2006).

Sequence kernel association test and SKAT-O

We also observed an African American to European American difference that was the result of the first axis of

variation (PC1) arising from PCA using EIGENSTRAT software (Price *et al*., 2006). Hence, we used a subgroup effect as a covariate in SKAT and SKAT-O.

The following genes were detected by SKAT: *CHRNA1* (P = 0.001, Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*), *UBQLN1* (P = 0.025, Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*), and *CHRM1* (P = 0.041, Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*). *CHRNA1* is associated with smoking cessation (Rose *et al*., 2010) and lung adenocarcinoma (Chang *et al*., 2013). *UBQLN1* is associated with smoking cessation (Rose *et al*., 2010) and nonsmall cell lung cancer (Shah *et al*., 2015).

The following genes were detected by SKAT-O: *CHRNA1* (P = 0.015), *CHRND* (P = 0.034), *UBQLN1* (P = 0.034), *CHRM1* (P = 0.013), *CHRNA3* (P = 0.018), and *CHRNA5* (P = 0.048). *CHRND* has been reported to be related to modify the risk for nicotine dependence associated with peer smoking (Johnson *et al*., 2010).

Comparison of the rare variant–associated results from str-CMC, str-WSS, sequence kernel association test, and SKAT-O

We applied CMH analysis to the 39 acetylcholine receptor-related genes, which revealed that 15 of these genes had differences in SNP allele frequencies between subgroups after controlling for different groups arising from PCA using EIGENSTRAT software (Price *et al*., 2006) (i.e., European Americans and African Americans) (Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*). Among these 15 genes (with different SNP allele frequencies), str-CMC found 10 genes (two in European Americans and eight in African Americans) and str-WSS detected six genes (four in European Americans, one in African Americans; one in both European Americans and African Americans). However, SKAT and SKAT-O detected only 1 genes (Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*). By contrast, we found 22 genes that did not differ with respect to SNP allele frequency between subgroups after controlling for different groups arising from PCA using EIGENSTRAT software. Of these 22 genes, str-CMC detected three genes (0 in European Americans and three in African Americans) and str-WSS detected two genes (0 in European Americans and two in African Americans). SKAT and SKAT-O also detected five genes (Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*).

For investigating differences between population stratification and combined analyses (a within test comparison), among these 15 genes (with different SNP allele frequencies) in population stratification analyses, str-CMC found 10 genes (two in European Americans and eight in African Americans). By contrast, in combined analyses, CMC only found two of them. However, str-WSS (detected six genes) and WSS (detected five genes) had comparable performance in these 15 genes.

The results indicated that for this dataset, two subgroups, that is, group European Americans and African Americans as a covariate was not an effective substitute for analyzing subgroups separately when only one of them contained an associated rare variant.

Discussion

We determined whether population stratification and subgroup effects would be better dealt with using stratified analyses or including population as a covariate. Upon comparing results from str-CMC, str-WSS, SKAT, and SKAT-O, we found that the inclusion of samples from other subgroups often introduced noise when the signal for a particular gene was strong in one of the subgroups. Without stratification analysis using CMC, in the CMH test, for example, the result for *GABRG2* was significant at P = 0.0009. However, with stratification analysis using str-CMC, the P value for *GABRG2* was P = 0.00000804 (Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*) and it was P = 0.046 for str-WSS in African Americans. On the other hand, *GABRG2* was not significant by using SKAT (P = 0.40426) and SKAT-O (P = 0.60138). In addition, *GABRG2*, *GABRB2*, *CHRNA2*, *JAK2*, *RIC3*, *AGPHD1*, *PSMA4*, *CHRNA5*, *CHRNA3*, *CHRNB4*, *CHRM5*, *CHRNE*, *CDH2*, and *APP* were significant in subsamples representing more than half of the data, and dealing with the strata in separate analyses increased the chances for detecting associated rare variants.

Despite the allele frequencies not being different according to CMH test, *CHRNA1* was not significant using CMC and WSS. However, by using SKAT and SKAT-O, the P value is borderline significant (P = 0.01, Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*).

By using these data, our results demonstrated that all rare-variant association methods considered here could yield a relatively high rate of spurious associations in the presence of fine-scale population structure. In addition, we showed that considering for the effects of population stratification and subgroup effects can confound rare variant analyses. The differences in disease risk between subgroups that generated such high spurious association rates are plausible and it is important for further interpreting rare-variant association results. For instance, there is a 2.5–10% difference in the prevalence of lung cancer among different populations of European men, although there is a less striking difference for women (Boyle and Ferlay, 2005). In our study, *GABRG2* was detected by str-CMC (P = 8.04E-06, Fig. 1; Supplementary Additional

file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*) and str-WSS ($P = 0.046$, Fig. 1; Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A221*) in African Americans. By using SKAT and SKAT-O, the *P* values for *GABRG2* are 0.404 and 0.601, respectively (Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A220*). *GABRG2* was not detected by the two tests if stratification was not considered. *GABRG2* was also associated with addiction (Klee *et al*., 2012) and alcohol use disorder (Li *et al*., 2014; Zai *et al*., 2014). In our study, *GABRG2* might affect nicotine dependence risk in African Americans.

Gene-based methods for detecting rare variants are as effective as the SNP-based methods for GWAS (Schaid *et al*., 2005; Wessel and Schork, 2006; Tzeng and Zhang, 2007). The grouping of multiple SNPs within a genomic region allows combined calculations to enhance statistical power, because rare variants include extremely sparse data so that traditional SNP-set methods for common variants might not be applicable to rare-variant detection. Some well-known rare-variant detection methods can potentially be used to combine low-frequency SNPs in GWAS. However, the number of SNPs in a gene might be important in the rare variant detection methods. For genes with larger numbers of markers, such as acetylcholine receptor-related genes, CMC and str-CMC were more likely to detect the effects compared with other methods, that is, WSS, str-WSS, SKAT, and SKAT-O. For example, in our study, we identified 262 SNPs in *APP* and 180 in *CDH2* by str-CMC in group African Americans (Supplementary Additional file 1, Supplemental digital content 1, *http://links.lww.com/PG/A220*).

Population structure and subgroups can be strong confounding factors in association studies (Pritchard *et al*., 2000; Ziv and Burchard 2003; Clayton *et al*., 2005; Roeder and Luca, 2009), so their effects need to be taken into account. To tackle this problem, we used EIGENSTRAT software (Price *et al*., 2006) that can remove some of their effects. However, the number of principal components used should depend on the distribution of the eigenvalues (Jiang and Dong, 2011) and sample sizes of subgroups.

Rare variant-detection methods can divide into two main categories: burden and variance-component tests (Bansal *et al*., 2010) such that they should complement each other for the purpose of identifying possible risk factors for nicotine dependence or other complex traits. Nicotine usage is associated with 5 million deaths per year worldwide and is considered one of the gateway drugs that lead to the use of illicit drugs. Before detecting rare variants, the CMH test can be used to determine whether there are any possible differences in SNP allele frequencies between subgroups within a gene/genomic region/haplotype/pathway. If in a gene differences in SNP allele frequencies between subgroups occur, stratification analyses such as str-CMC or str-WSS should be used in detecting rare variants. Hence, CMH should first be used in rare variant detection analysis. By contrast, when SNP allele frequencies between subgroups are similar all methods can be used directly.

In our study, we cannot fully determine whether the results demonstrate the differences between population stratification versus combined analyses, or whether they reflect the differences between the statistical approaches. Because the population stratification/combined analyses and statistical approaches are fundamentally different, these methods should be considered complementary to each other when studying rare variants in various disease analyses. In our study, the small sample size for the African Americans population is a limitation, particularly when addressing a question involving low-frequency variants. It is difficult to interpret how much of the results are driven by the small numbers in the African Americans group. In future studies, we will adopt a pairwise sampling design based on Imai *et al*. (2015) to increase sample sizes.

In conclusion, we have extended the CMC and WSS methods to identify rare variants and stratify by population/subgroups while analyzing smoking cessation data. We found that including population as a covariate was not an effective substitute for analyzing the subpopulations separately when only one subpopulation contained a rare variant linked to the phenotype. The conclusion is the same as previous study findings (Culverhouse *et al*., 2011). Our results will help researchers overcome population stratification and subgroup effects when detecting rare variants. More importantly, these analyses showed that even when an identical genetic model is applied to multiple subgroups, sample size is not the only factor that determines association results. If rare causative variants are unique to a subgroup, stratified analyses might be more powerful than combined analyses although stratified analyses may entail a considerable decrease in the sample size.

Acknowledgements

We are grateful to the National Science Council and Institute of Biomedical Sciences, Academia Sinica of Taiwan and China Medical University of Taiwan for funding (MOST102-2314-B-001 -003 -MY2, CMU103-N-15 and CMU105-N-23). The authors thank Jurg Ott for editorial support of the manuscript.

This study was supported by National Science Council and Institute of Biomedical Sciences, Academia Sinica of Taiwan and China Medical University of Taiwan (MOST102-2314-B-001-003-MY2, CMU103-N-15 and CMU105-N-23).

A.-R.H. and C.S.J.F. each contributed to statistical analysis, data interpretation, and writing of the manuscript. L.-S.C. contributed to writing of the manuscript. Y.J.L. contributed to statistical analysis.

The data were downloaded from the Database of Genotypes and Phenotypes (phs000404.v1.p1). The

data were accessed after a Controlled Access Application (https://dbgap.ncbi.nlm.nih.gov/aa), and an approval process from both National Center for Biotechnology Information and Institutional Review Board at Academia Sinica (AS-IRB01-17006). The datasets analyzed during the current study are available in the Database of Genotypes and Phenotypes (phs000404.v1.p1) https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000404.v1.p1.

Conflicts of interest

There are no conflicts of interest.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, *et al.* (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**:56–65.

Anney RJ, Lotfi-Miri M, Olsson CA, Reid SC, Hemphill SA, Patton GC (2007). Variation in the gene coding for the M5 muscarinic receptor (CHRM5) influences cigarette dose but is not associated with dependence to drugs of addiction: evidence from a prospective population based cohort study of young adults. *BMC Genet* **8**:46.

Auer PL, Lettre G (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome Med* **7**:16.

Bansal V, Libiger O, Torkamani A, Schork NJ (2010). Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* **11**:773–785.

Bergen AW, Caporaso N (1999). Cigarette smoking. *J Natl Cancer Inst* **91**:1365–1375.

Berrettini WH, Wileyto EP, Epstein L, Restine S, Hawk L, Shields P, *et al.* (2007). Catechol-O-methyltransferase (COMT) gene variants predict response to bupropion therapy for tobacco dependence. *Biol Psychiatry* **61**:111–118.

Boyle P, Ferlay J (2005). Cancer incidence and mortality in europe, 2004. *Ann Oncol* **16**:481–488.

Breslau N, Johnson EO, Hiripi E, Kessler R (2001). Nicotine dependence in the united states: prevalence, trends, and smoking persistence. *Arch Gen Psychiatry* **58**:810–816.

Browning SR, Browning BL (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**:1084–1097.

Centers for Disease Control and Prevention (2008). Cigarette smoking among adults – United States, 2007. *Morb Mortal Wkly Rep* **57**:1221–1226.

Chang PM, Yeh YC, Chen TC, Wu YC, Lu PJ, Cheng HC, *et al.* (2013). High expression of CHRNA1 is associated with reduced survival in early stage lung adenocarcinoma after complete resection. *Ann Surg Oncol* **20**:3648–3654.

Chen LS, Baker TB, Piper ME, Smith SS, Gu C, Grucza RA, *et al.* (2014). Interplay of genetic risk (CHRNA5) and environmental risk (partner smoking) on cigarette smoking reduction. *Drug Alcohol Depend* **143**:36–43.

Choi HD, Shin WG (2015). Lack of association between DRD2 taq1a gene polymorphism and smoking cessation therapy: a meta-analysis. *Int J Clin Pharmacol Ther* **53**:415–421.

Choi JS, Payne TJ, Ma JZ, Li MD (2017). Relationship between personality traits and nicotine dependence in male and female smokers of African-American and European-American samples. *Front Psychiatry* **8**:122.

Chung J, Wang X, Maruyama T, Ma Y, Zhang X, Mez J, *et al.*; Alzheimer's Disease Neuroimaging Initiative (2018). Genome-wide association study of Alzheimer's disease endophenotypes at prediagnosis stages. *Alzheimers Dement* **14**:623–633.

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, *et al.* (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37**:1243–1246.

Conti DV, Lee W, Li D, Liu J, Van Den Berg D, Thomas PD, *et al.*; Pharmacogenetics of Nicotine Addiction and Treatment Consortium (2008). Nicotinic acetylcholine receptor beta2 subunit gene implicated in a systems-based candidate gene study of smoking cessation. *Hum Mol Genet* **17**:2834–2848.

Covey LS, Botello-Harbaum M, Glassman AH, Masmela J, LoDuca C, Salzman V, Fried J (2008). Smokers' response to combination bupropion, nicotine patch, and counseling treatment by race/ethnicity. *Ethn Dis* **18**:59–64.

Culverhouse RC, Hinrichs AL, Suarez BK (2011). Stratify or adjust? Dealing with multiple populations when evaluating rare variants. *BMC Proc* **5**(Suppl 9):S101.

Dajani R, Li J, Wei Z, March ME, Xia Q, Khader Y, *et al.* (2017). Genome-wide association study identifies novel type II diabetes risk loci in jordan subpopulations. *Peerj* **5**:e3618.

Dau A, Komal P, Truong M, Morris G, Evans G, Nashmi R (2013). RIC-3 differentially modulates $\alpha4\beta2$ and $\alpha7$ nicotinic receptor assembly, expression, and nicotine-induced receptor upregulation. *BMC Neurosci* **14**:47.

Erdmann J, Kessler T, Munoz Venegas L, Schunkert H (2018). A decade of genome-wide association studies for coronary artery disease: the challenges ahead. *Cardiovasc Res* **114**:1241–1257.

Fang J, Wang X, He B (2014). Association between common genetic variants in the opioid pathway and smoking behaviors in chinese men. *Behav Brain Funct* **10**:2.

Gusev A, Bhatia G, Zaitlen N, Vilhjalmsson BJ, Diogo D, Stahl EA, *et al.* (2013). Quantifying missing heritability at known GWAS loci. *Plos Genet* **9**:e1003993.

Haiman CA, Stram DO, Wilkens LR, Pike MC, Kolonel LN, Henderson BE, Le Marchand L (2006). Ethnic and racial differences in the smoking-related risk of lung cancer. *N Engl J Med* **354**:333–342.

Hansen HM, Xiao Y, Rice T, Bracc PM, Wrensch MR, Sison JD, *et al.* (2010). Fine mapping of chromosome 15q25.1 lung cancer susceptibility in African-Americans. *Hum Mol Genet* **19**:3652–3661.

He P, Yang XX, He XQ, Chen J, Li FX, Gu X, *et al.* (2014). CHRNA3 polymorphism modifies lung adenocarcinoma risk in the Chinese han population. *Int J Mol Sci* **15**:5446–5457.

Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO (1991). The Fagerström test for nicotine dependence: a revision of the Fagerström tolerance questionnaire. *Br J Addict* **86**:1119–1127.

Heitjan DF, Guo M, Ray R, Wileyto EP, Epstein LH, Lerman C (2008). Identification of pharmacogenetic markers in smoking cessation therapy. *Am J Med Genet B Neuropsychiatr Genet* **147B**:712–719.

Helzer JE, van den Brink W, Guth SE (2006). Should there be both categorical and dimensional criteria for the substance use disorders in DSM-V? *Addiction* **101**(Suppl 1):17–22.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**:9362–9367.

Hofker MH, Fu J, Wijmenga C (2014). The genome revolution and its role in understanding complex diseases. *Biochim Biophys Acta* **1842**:1889–1895.

Hondebrink L, Tan S, Hermans E, van Kleef RG, Meulenbelt J, Westerink RH (2013). Additive inhibition of human $\alpha1\beta2\gamma2$ GABAA receptors by mixtures of commonly used drugs of abuse. *Neurotoxicology* **35**:23–29.

Imai A, Nakaya A, Fahiminiya S, Tétreault M, Majewski J, Sakata Y, *et al.* (2015). Beyond homozygosity mapping: family-control analysis based on hamming distance for prioritizing variants in exome sequencing. *Sci Rep* **5**:12028.

Janca A, Ustün TB, Early TS, Sartorius N (1993). The ICD-10 symptom checklist: a companion to the ICD-10 classification of mental and behavioural disorders. *Soc Psychiatry Psychiatr Epidemiol* **28**:239–242.

Jemal A, Thun MJ, Ries LA, Howe HL, Weir HK, Center MM, *et al.* (2008). Annual report to the nation on the status of cancer, 1975-2005, featuring trends in lung cancer, tobacco use, and tobacco control. *J Natl Cancer Inst* **100**:1672–1694.

Ji X, Zhang W, Gui J, Fan X, Zhang W, Li Y, *et al.* (2014). Role of a genetic variant on the 15q25.1 lung cancer susceptibility locus in smoking-associated nasopharyngeal carcinoma. *Plos One* **9**:e109036.

Jiang R, Dong J (2011). Detecting rare functional variants using a wavelet-based test on quantitative and qualitative traits. *BMC Proc* **5**(Suppl 9):S70.

Johnson EO, Chen LS, Breslau N, Hatsukami D, Robbins T, Saccone NL, *et al.* (2010). Peer smoking and the nicotinic receptor genes: an examination of genetic and environmental risks for nicotine dependence. *Addiction* **105**:2014–2022.

Kishioka S, Kiguchi N, Kobayashi Y, Saika F, Yamamoto C (2014). Development of physical dependence on nicotine and endogenous opioid system--participation of $\alpha7$ nicotinic acetylcholine receptor. *Nihon Arukoru Yakubutsu Igakkai Zasshi* **49**:227–237.

Klee EW, Schneider H, Clark KJ, Cousin MA, Ebbert JO, Hooten WM, *et al.* (2012). Zebrafish: a model for the study of addiction genetics. *Hum Genet* **131**:977–1008.

Kuang BH, Zhang MQ, Xu LH, Hu LJ, Wang HB, Zhao WF, *et al.* (2013). Proline-rich tyrosine kinase 2 and its phosphorylated form py881 are novel prognostic markers for non-small-cell lung cancer progression and patients' overall survival. *Br J Cancer* **109**:1252–1263.

Lee S, Wu MC, Lin X (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**:762–775.

Li B, Leal SM (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**:311–321.

Li D, Sulovari A, Cheng C, Zhao H, Kranzler HR, Gelernter J (2014). Association of gamma-aminobutyric acid A receptor α2 gene (GABRA2) with alcohol use disorder. *Neuropsychopharmacology* **39**:907–918.

Liang F, Xiong M (2013). Bayesian detection of causal rare variants under posterior consistency. *Plos One* **8**:e69633.

Lichtenstein P, Yip BH, Björk C, Pawitan Y, Cannon TD, Sullivan PF, Hultman CM (2009). Common genetic determinants of schizophrenia and bipolar disorder in swedish families: a population-based study. *Lancet* **373**:234–239.

López-Torrecillas F, López-Quirantes EM, Maldonado A, Albein-Urios N, Rueda MDM, Verdejo-Garcia A (2017). Decisional balance and processes of change in community-recruited with moderate-high versus mild severity of cannabis dependence. *Plos One* **12**:e0188476.

Lou XY, Ma JZ, Payne TJ, Beuten J, Crew KM, Li MD (2006). Gene-based analysis suggests association of the nicotinic acetylcholine receptor beta1 subunit (CHRNB1) and M1 muscarinic acetylcholine receptor (CHRM1) with vulnerability for nicotine dependence. *Hum Genet* **120**:381–389.

Luo Z, Alvarado GF, Hatsukami DK, Johnson EO, Bierut LJ, Breslau N (2008). Race differences in nicotine dependence in the collaborative genetic study of nicotine dependence (COGEND). *Nicotine Tob Res* **10**:1223–1230.

Ma C, Boehnke M, Lee S; GoT2D Investigators (2015). Evaluating the calibration and power of three gene-based association tests of rare variants for the X chromosome. *Genet Epidemiol* **39**:499–508.

Ma Y, Yuan W, Jiang X, Cui WY, Li MD (2015). Updated findings of the association and functional studies of DRD2/ANKK1 variants with addictions. *Mol Neurobiol* **51**:281–299.

Madsen BE, Browning SR (2009). A groupwise association test for rare mutations using a weighted sum statistic. *Plos Genet* **5**:e1000384.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, *et al.* (2009). Finding the missing heritability of complex diseases. *Nature* **461**:747–753.

Mantel N, Haenszel W (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* **22**:719–748.

Marian AJ (2012). Elements of 'missing heritability'. *Curr Opin Cardiol* **27**:197–201.

Moore CB, Wallace JR, Wolfe DJ, Frase AT, Pendergrass SA, Weiss KM, Ritchie MD (2013). Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *Plos Genet* **9**:e1003959.

Munafò MR, Johnstone EC, Walther D, Uhl GR, Murphy MF, Aveyard P (2011). CHRNA3 rs1051730 genotype and short-term smoking cessation. *Nicotine Tob Res* **13**:982–988.

Nicolae DL (2016). Association tests for rare variants. *Annu Rev Genomics Hum Genet* **17**:117–130.

O'Connor TD, Kiezun A, Bamshad M, Rich SS, Smith JD, Turner E, *et al.*; NHLBIGO Exome Sequencing Project; ESP Population Genetics, Statistical Analysis Working Group (2013). Fine-scale patterns of population stratification confound rare variant association tests. *Plos One* **8**:e65834.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**:904–909.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000). Association mapping in structured populations. *Am J Hum Genet* **67**:170–181.

Prokopenko D, Hecker J, Silverman EK, Pagano M, Nöthen MM, Dina C, *et al.* (2016). Utilizing the jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 genomes project. *Bioinformatics* **32**:1366–1372.

Reinthaler EM, Dejanovic B, Lal D, Semtner M, Merkler Y, Reinhold A, *et al.*; EuroEPINOMICS Consortium (2015). Rare variants in γ-aminobutyric acid type A receptor genes in rolandic epilepsy and related syndromes. *Ann Neurol* **77**:972–986.

Roeder K, Luca D (2009). Searching for disease susceptibility variants in structured populations. *Genomics* **93**:1–4.

Rose JE, Behm FM, Drgon T, Johnson C, Uhl GR (2010). Personalized smoking cessation: interactions between nicotine dose, dependence and quit-success genotype score. *Mol Med* **16**:247–253.

Russo P, Cesario A, Rutella S, Veronesi G, Spaggiari L, Galetta D, *et al.* (2011). Impact of genetic variability in nicotinic acetylcholine receptors on nicotine addiction and smoking cessation treatment. *Curr Med Chem* **18**: 91–112.

Rüther T, Bobes J, De Hert M, Svensson TH, Mann K, Batra A, *et al.*; European Psychiatric Association (2014). EPA guidance on tobacco dependence and strategies for smoking cessation in people with mental illness. *Eur Psychiatry* **29**:65–82.

Saccone NL, Schwantes-An TH, Wang JC, Grucza RA, Breslau N, Hatsukami D, *et al.* (2010). Multiple cholinergic nicotinic receptor genes affect nicotine dependence risk in African and European Americans. *Genes Brain Behav* **9**:741–750.

Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, *et al.* (2007). Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 snps. *Hum Mol Genet* **16**:36–49.

Sampson JN, Jacobs K, Wang Z, Yeager M, Chanock S, Chatterjee N (2012). A two-platform design for next generation genome-wide association studies. *Genet Epidemiol* **36**:400–408.

Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN (2005). Nonparametric tests of association of multiple genes with human disease. *Am J Hum Genet* **76**:780–793.

Schork NJ, Murray SS, Frazer KA, Topol EJ (2009). Common vs. Rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* **19**: 212–219.

Shah PP, Lockwood WW, Saurabh K, Kurlawala Z, Shannon SP, Waigel S, *et al.* (2015). Ubiquilin1 represses migration and epithelial-to-mesenchymal transition of human non-small cell lung cancer cells. *Oncogene* **34**: 1709–1717.

Shen L, Jia J (2016). An overview of genome-wide association studies in Alzheimer's disease. *Neurosci Bull* **32**:183–190.

Slattery ML, Lundgreen A, Hines LM, Torres-Mejia G, Wolff RK, Stern MC, John EM (2014). Genetic variation in the JAK/STAT/SOCS signaling pathway influences breast cancer-specific mortality through interaction with cigarette smoking and use of aspirin/nsaids: the breast cancer health disparities study. *Breast Cancer Res Treat* **147**:145–158.

Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, *et al.*; ENGAGE Consortium (2010). Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* **42**:448–453.

Tzeng JY, Zhang D (2007). Haplotype-based association analysis via variance-components score test. *Am J Hum Genet* **81**:927–938.

Wang J, Li MD (2010). Common and unique biological pathways associated with smoking initiation/progression, nicotine dependence, and smoking cessation. *Neuropsychopharmacology* **35**:702–719.

Wang S, D van der Vaart A, Xu Q, Seneviratne C, Pomerleau OF, Pomerleau CS, *et al.* (2014). Significant associations of CHRNA2 and CHRNA6 with nicotine dependence in European American and African American populations. *Hum Genet* **133**:575–586.

Wang X, Zhang S, Li Y, Li M, Sha Q (2015). A powerful approach to test an optimally weighted combination of rare variants in admixed populations. *Genet Epidemiol* **39**:294–305.

Wessel J, Schork NJ (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet* **79**:792–806.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**:82–93.

Zai CC, Zai GC, Tiwari AK, Manchia M, de Luca V, Shaikh SA, *et al.* (2014). Association study of GABRG2 polymorphisms with suicidal behaviour in schizophrenia patients with alcohol use disorder. *Neuropsychobiology* **69**:154–158.

Zhang X, Zhang R, Zheng Y, Shen J, Xiao D, Li J, *et al.* (2013). Expression of gamma-aminobutyric acid receptors on neoplastic growth and prediction of prognosis in non-small cell lung cancer. *J Transl Med* **11**:102.

Zhao C, Wang F, Pun FW, Mei L, Ren L, Yu Z, *et al.* (2012). Epigenetic regulation on GABRB2 isoforms expression: developmental variations and disruptions in psychotic disorders. *Schizophr Res* **134**:260–266.

Ziv E, Burchard EG (2003). Human population structure and genetic association studies. *Pharmacogenomics* **4**:431–441.

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, *et al.* (2014). Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**:E455–E464.