Published in final edited form as:

Stat Appl Genet Mol Biol. ; 17(6): . doi:10.1515/sagmb-2017-0054.

A Practical Approach to Adjusting for Population Stratification in Genome-wide Association Studies: Principal Components And Propensity Scores (PCAPS)

Huaqing Zhao¹, Nandita Mitra², Peter A. Kanetsky³, Katherine L. Nathanson⁴, and Timothy R. Rebbeck⁵

¹Department of Clinical Sciences, Temple University School of Medicine, 3440 N. Broad Street, Kresge Hall East, Room 218, Philadelphia, PA 19140

²Department of Biostatistics and Epidemiology, University of Pennsylvania, 622 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104

³Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, 213 MRC CANCONT, Tampa, FL 33612

⁴Department of Medicine, University of Pennsylvania, South Pavilion, Perelman Center for Advanced Medicine, 3400 Civic Center Blvd, Philadelphia, PA 19104

⁵Division of Population Sciences, Dana Farber Cancer Institute and Department of Epidemiology, Harvard T.H. Chan School of Public Health, 1101 Dana. 450 Brookline Ave, Boston, MA 02215

Abstract

Genome-wide association studies (GWAS) are susceptible to bias due to population stratification (PS). The most widely used method to correct bias due to PS is principal components (PCs) analysis (PCA), but there is no objective method to guide which PCs to include as covariates. Often, the ten PCs with the highest eigenvalues are included to adjust for PS. This selection is arbitrary, and patterns of local linkage disequilibrium may affect PCA corrections. To address these limitations, we estimate genomic propensity scores based on all statistically significant PCs selected by the Tracy-Widom (TW) statistic. We compare a principal components and propensity scores (PCAPS) approach to PCA and EMMAX using simulated GWAS data under no, moderate, and severe PS. PCAPS reduced spurious genetic associations regardless of the degree of PS, resulting in odds ratio (OR) estimates closer to the true OR. We illustrate our PCAPS method using GWAS data from a study of testicular germ cell tumors. PCAPS provided a more conservative adjustment than PCA. Advantages of the PCAPS approach include reduction of bias compared to PCA, consistent selection of propensity scores to adjust for PS, the potential ability to handle outliers, and ease of implementation using existing software packages.

Keywords

principal components analysis; bias; propensity score; testicular germ cell tumors; Tracy-Widom statistic

Introduction

Genome-wide association studies (GWAS) are an effective approach for identifying common genetic variants associated with disease risk. GWAS have increasingly been used with very large sample sizes that combine participants from many (potentially ancestrally heterogeneous) centers, and there is an increasing trend to undertake GWAS in population admixture (i.e. stratification). Thus, GWAS associations may be biased due to population stratification (PS). The most widely used method to address PS is principal components analysis (PCA) (Price et al., 2006). PCA is a method of correction for population structure in genetic association studies on unrelated individuals. The original PCA approach was proposed by Price et al. (2006) with subsequent extension by Patterson et al. (2006). This method uses genome-wide genotype data to estimate principal components axes that can be used as covariates in subsequent association analyses. The implication of this approach is that these axes represent features of genomic ancestry that capture PS. Since the adjustment of each single nucleotide polymorphism (SNP) varies along each axis of ancestry, PCA corrects not only for false positives but also for false negatives.

Despite its utility and ease of application, there are limitations to the PCA approach. The PCA method usually recommends using the top 10 principal components (PCs) to adjust for population stratification (Price et al. 2010; Feng et al. 2009; Kang et al. 2009). However, the recommendation to use 10 PCs is arbitrary. Adjusting for more PCs than needed to sufficiently adjust for PS can harm efficiency. Using fewer PCs than needed may result in residual bias. Most importantly, patterns of local linkage disequilibrium (LD) may cause PCA to create “nuisance axes”, which may be interpreted as the existence of subpopulations that reflect localized LD phenomena rather than plausible PS (Zou et al. 2010). These axes are not related to ancestry overall, but instead end up separating individuals according to their genotype in local genomic regions. Such PCs are dominated by relatively few SNPs that map to a few small chromosomal regions with extended LD (Feng et al.; Wang et al. 2009). These SNPs with local LD are referred to as outlier SNPs, and PCA is sensitive to such outliers. Specifically, due to individual outliers (a handful of individuals with very different ancestry origin from the rest of the cohort), numerous dimensions of ancestry may appear to model a statistically significant amount of variation in the data, but in actuality they function to separate a single observation from the bulk of the data (Lee et al. 2010).

Price et al. (2006) and Bouaziz et al. (2011) provide reviews of several methods that correct for population stratification in GWAS including genomic control (Devlin and Roeder 1999), adjusted regressions and meta-analyses. Several PCA-based approaches (Zhang et al. 2013a; Zhang et al. 2013b; Liu et al. 2013; Pritchard et al. 2000) have been proposed to correct PS in GWAS. A number of other methods such as structural association (Pritchard et al.) multidimensional scaling (Li and Yu 2008), linear mixed modeling (Zhang et al. 2010),

variance component models (Kang et al. 2010), matching (Lee et al.; Luca et al. 2008; Epstein et al. 2012) and logistic mixed models (Cheng et al. 2016) have also been used to correct PS. However, these methods can be computationally intensive in GWAS (Lee et al.; Liu et al.; Pritchard et al.; Luca et al.; Epstein et al. 2012) or have similar limitations that PCA suffers from (Zhang et al. 2013a; Zhang et al. 2013b; Li and Yu; Chen et al.). Kang et al. (2010) developed the EMMAX program, taking to an expedited mixed linear model approach to correct for sample structure within human GWASs.

Recently, we developed the genomic propensity score (GPS) approach (Zhao et al. 2009; Zhao et al. 2012), which is a propensity score (Rosenbaum and Rubin 1983) based approach to correct for bias due to population stratification that allows estimation of the effect of a genotype under a wide range of genetic models (dominant, recessive, additive) while simultaneously adjusting for genetic markers along with patient and disease characteristics. The GPS reduced bias in estimated genetic effects and performed the same as or better than other existing approaches including PCA. Lin and Zeng (2011) further presented a theoretical justification for the GPS. However, one shortcoming of the GPS method is that it was developed in the context of candidate gene studies, and thus is not readily applicable to GWAS as currently formulated. To overcome the limitations of PCA and to extend the GPS for GWAS data, we propose a new approach that combines principal components and propensity scores (PCAPS) that is based on the Tracy Widom (TW) statistic (Tracy and Widom 1993; Trach and Widom 1994; Tracy and Widom 1996). This approach allows us to correct for confounding due to population stratification at each locus by adjusting for a summary covariate (the propensity score) which captures and balances differences in allele frequencies between the two comparison groups. Our goal is to obtain an unbiased estimate of the SNP association after adjusting for this confounder. The propensity score captures population stratification by summarizing the differences in all of the allele frequencies between the comparison groups. Using GWAS simulation studies under varying scenarios, we directly compare PCAPS to using PCA adjustment alone and EMMAX. We illustrate our approach in a GWAS of testicular germ cell tumors (TCGT) (Kanetsky et al. 2009).

Materials and Methods

Theoretical Framework

Selection of the number of significant PC's requires testing the significance of the k^{th} largest eigenvalues. Tracy and Widom (1993, 1994, 1996) laid the foundation for this approach by proposing an analytical expression of the largest eigenvalue. Johnstone (2001) later proposed using the TW statistic to select significant components. Patterson et al. demonstrated the potential of using the TW statistic with PCA to uncover population structure. Zhao et al. (2009, 2012) first developed propensity score-based approaches to correct for population stratification and Lin and Zeng provided theoretical justification for using the propensity score in a logistic regression model. PCs are bounded from 0 to 1, and so they can be used to down weight the effect of outliers (Imbens 2004). Attributed to these theoretical justifications, we propose a new approach to adjust for population stratification based on propensity scores. First, we select the number of statistically significant genotype principal components based on the TW test. We then calculate the propensity score as the likelihood

of an individual having a particular genotype given the individual's significant PCs. Finally, the propensity scores are incorporated into the association model to adjust for population stratification.

The TW distribution is the probability distribution of the normalized largest eigenvalue of a random Hermitian matrix (Dominici and Maier 2008). The TW distribution (Tracy and Widom 1993) is defined as

$$F_1(s) = \exp\left(-\frac{1}{2} \int_s^\infty q(x) dx\right) (F_2(s))^{1/2}$$

where $F_2(s) = \exp\left(-\int_s^\infty (x-s)q^2(x)dx\right)$. The function $q(x)$ is the Hastings-McLeod solution (Hastings and McLeod 1980) of a Painlevé equation of type II where $q''(s) = sq(s) + 2q(s)^3$, satisfying the boundary condition $q(s) \rightarrow \text{Ai}(s)$, $s \rightarrow \infty$; where $\text{Ai}(s)$ is the Airy function (Airy 1838).

Johnstone showed that $F_1(s)$ is the limiting distribution of the largest eigenvalue of a real sample covariance matrix. Let $C = MM^T$ where M is $m \times n$ matrix with each entry be independently and identically distributed as $\sim N(0, 1)$, and let λ_1 be the largest eigenvalue of C . Define

$$\mu(m, n) = (\sqrt{m-1} + \sqrt{n})^2 \text{ and } \sigma(m, n) = (\sqrt{m-1} + \sqrt{n}) \left(\frac{1}{\sqrt{m-1}} + \frac{1}{\sqrt{n}} \right)$$

If $\lim_{n \rightarrow \infty} \frac{n}{m} = \gamma$ where $\gamma \in (0, 1]$, then $L_1 = \frac{\lambda_1 - \mu(m, n)}{\sigma(m, n)} \sim \text{TW}_1$

where TW_1 is the TW probability density function, i.e., $\text{TW}_1 = \frac{d}{ds} F_1(s)$.

We denote G as one of the top hit SNPs from a GWAS analysis without PS correction (0, 1, 2 genotypes), D denotes disease status (0, 1 affected vs. unaffected), and X denotes a vector of statistically significant PC axes based on the TW statistic. We define the principal component and propensity score (PCAPS) to be the likelihood of an individual having a particular genotype based on that individual's covariate makeup. This can be stated explicitly as $PCAPS_i(g_i, x_i) = P(G_i | x_i)$, where $PCAPS_i(g_i, x_i)$ is the genomic propensity score for individual i calculated from that individual's x_i which represents that individual's vector of statistically significant PCs based on TW statistic, and where G_i is that individual's test-locus genotype. For example, for an additive disease risk allele a and reference allele A , $G = 2$ may represent the high-risk genotype aa , $G = 1$ may represent the heterozygote risk genotype Aa , and $G = 0$ may represent the homozygote reference genotype AA . In the current situation, the proportional odds model can be used to model the cumulative probabilities $P(G=0|X)$ and $P(G=0|X) + P(G=1|X)$ jointly as follows:

$$\log[P(G=0|X)/(1 - P(G=0|X))] = \alpha_1 + \beta_1^T X$$

$$\log [(P(G = 0|X) + P(G = 1|X))/P(G = 2|X)] = \alpha_2 + \beta_1^T X$$

where α_1 and α_2 are intercepts and β_1 is a vector of coefficients. Since $P(G = 0|X) + P(G = 1|X) + P(G = 2|X) = 1$, then we have,

$$P(G = 0|X) = \exp(\alpha_1 + \beta_1^T X) / [1 + \exp(\alpha_1 + \beta_1^T X)]$$

$$P(G = 2|X) = 1 / [1 + \exp(\alpha_2 + \beta_1^T X)]$$

$$P(G = 1|X) = 1 - P(G = 0|X) - P(G = 2|X)$$

In the subsequent text, we suppress the index i corresponding to the i^{th} individual. Assuming strong ignorability (Rosenbaum and Rubin), we can assume that given $PCAPS(g, x)$, X and G are conditionally independent, thus

$$P(X, G | PCAPS(g, x)) = P(X | PCAPS(g, x))P(G | PCAPS(g, x))$$

which balances measured PCs based on TW statistic across genotypes.

We define a general class of models that specify the potential relationship among disease D , test-locus genotype G , and genetic covariates X as $f(E(D | G, X)) = \eta$, where $f(\cdot)$ is a link function, such as the logit function, that determines the relationship among D , G and X ; $E(D | G, X)$ denotes the conditional mean of D given G and X ; and η is a function of covariates, usually a linear function such that

$$\eta = \alpha + \beta G + \gamma_1 PCAPS(G = 1|X) + \gamma_2 PCAPS(G = 2|X)$$

where α and β are log odds ratios and γ_1 and γ_2 are nuisance parameters. We fit the disease prevalence model assuming a logit link to estimate the effect of risk genotype(s) β as

$$\text{logit}(D|G, PCAPS(G, X)) = \alpha + \beta G + \gamma_1 PCAPS(G = 1|X) + \gamma_2 PCAPS(G = 2|X)$$

The PCAPS Procedure

Using GWAS data, EIGENSOFT (<http://www.hsph.harvard.edu/alkes-price/software/>) (Price et al. 2006; Patterson et al.) was used to identify a number of statistically significant PC axes based on the TW method (Johnstone) described in Patterson et al. (2006). PLINK (<http://pngu.mgh.harvard.edu/~purcell/plink>) (Purcell et al. 2007) was then used to obtain uncorrected SNP effect estimates based on various association models. For all SNPs without PS correction, PCAPS scores were calculated for each SNP using the previously identified

number of significant PCs by TW test. The resulting PCAPS scores were then used as covariates in subsequent association analyses to identify potential disease associated SNPs (i.e., SNPs remaining significant after correction by PCAPS scores).

Simulation Study

We used GWAsimulator (<http://biostat.mc.vanderbilt.edu/GWAsimulator>) (Li and Li 2008) to simulate GWAS data. GWAsimulator simulates whole genome case-control samples according to user-specified multi-locus disease models. GWAsimulator can use HapMap phased genotype data as input, and the simulated data have similar linkage disequilibrium (LD) patterns as the HapMap data.

We used HapMap 3 genotypes data to generate two populations (http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010-05_phaseIII/). HapMap 3 data include 165 samples from the CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) dataset and 203 from the YRI (Yoruba in Ibadan, Nigeria) dataset. Of 165 CEU samples, 112 founders were used as input phased data for simulating population 1 by GWAsimulator. Of 203 YRI samples, 147 founders were used as input phased data for simulating population 2.

The simulation study involved the following steps. First, two populations were generated using HapMap 3 data as phased input data. The first population used Caucasian (CEU) data, and the second population used Yoruba (YRI) data as described above. To reduce the computing time, 9 chromosomes (1, 2, 3, 6, 7, 8, 17, 18, and 19) were arbitrarily selected without loss of generality. For each chromosome, one disease locus was specified with a single SNP with OR 2, 1.5, 1, 1.5, 1, 2, 2, 1, and 1.5 for the 9 chromosomes, respectively. A population of 1000 cases and 1000 controls was simulated. After generation of the two populations separately, the populations were combined to induce population stratification by varying the case:control ratio 400:600, 450:550, 500:500, 550:450 and 600:400 within each subpopulation. Common SNPs from both subpopulations were kept in the combined sample dataset. SNPs with minor allele frequency (MAF) < 0.05 were excluded. Three study samples were created to reflect different levels of population structure (Table S1).

For each sample dataset (Table S1), GWAS association analyses were conducted using PLINK. To address multiple testing issues in GWAS under a model with no PS, only SNPs with Benjamini and Hochberg false positive rate less than 0.05 were selected. The top 9 SNPs without PS correction were identified along with the pre-defined 9 SNPs. The LD pattern between the 9 SNPs and top hit SNPs were examined using pairwise r^2 . The combined top 9 hits and/or pre-defined 9 SNPs were identified for further association analyses. For each sample dataset, we then analyzed the genetic effect for each SNP identified above by obtaining p-values with 1) No adjustment; 2) PCA adjustment; 3) EMMAX; and 4) PCAPS adjustment (Table S2). The latest release of EMMAX can be downloaded at EMMAX Download Page (<http://csg.sph.umich.edu/kang/emmax/download/index.html>). The performance of the four methods were evaluated by true (false) positive (negative) rates based on identified disease SNPs vs. true disease SNPs (Table 1). We further analyzed all SNPs for the GWAS sample dataset under moderate PS (Figure 2 and Table 2). For the PCA adjustment used to compare against PCAPS, the first 10 PCs were used. We

added the simulation results when all statistically significant PCs (based on the TW statistic) were adjusted (PCA-TW) and the first 10 PCs were used to calculate PCAPS (PCAPS-10) (Tables S3). Finally, we compared estimated ORs with the pre-defined “true” ORs using no adjustment, PCA and PCAPS adjustment methods (Tables 3 and S4).

TCGT Data

We used a genome-wide association study of 292 cases with testicular germ cell tumors (TGCT) and 919 controls (Kanetsky et al.). The strongest signal was observed on chromosome 12. Of the 11 top SNPs, six were on chromosome 12. These six SNPs were all from *KITLG*. We used the top six GWAS SNPs, and treated the corresponding gene *KITLG* as a candidate gene. We then calculated PCA scores using EIGENSOFT software (Price et al. 2006; Patterson et al.), and calculated TW statistics and their associated p-values for each PC axis (Patterson et al.). Only PCs with $p < 0.05$ by the TW test were retained for subsequent analysis. We then estimated PCAPS for each individual using the PCs as predictors in a logistic model. We compared the results of the associations at *KITLG* with PCA using the top 10 PCs and PCAPS adjustment.

Results

Simulation Results

Figure 1 shows Manhattan plots under no, moderate, and severe PS. The top ranking nine SNPs (above red line) were identified under no, moderate, and severe PS. As expected, Figure 1 indicates that there were more significant associations than one would expect by chance under moderate or severe PS. Figure S1 shows Q-Q plots for the visualization of population stratification. Under no PS, p-values fit the expected distribution well. Under moderate PS, many markers have p-values that exhibit modest inflation. Finally, under severe PS, a large number of markers exhibit small p-values demonstrating severe inflation. Significant deviation from the line under moderate and severe PS in Figure S1 confirms that there were more significant associations than one would expect by chance.

Tables S2 and S3 show the p-values for the nine most significantly associated SNPs under each scenario by PS status and different adjustment methods. Among the nine top-associated SNPs under no or moderate PS, seven are either a pre-defined disease SNP or in LD with a pre-defined disease SNP (true positive); the remaining two SNPs (true negative) are potential false positive findings. Under severe PS, only three pre-defined disease SNPs are among the nine top hits SNPs. The remaining six true negative SNPs are potential false positive SNPs for all three PS-correction methods. PCAPS consistently yielded larger p-values compared to PCA for almost all of these potential false positive SNPs. This suggests that some false positive findings might be correctly identified by PCAPS, but not by PCA. If a peak SNP is a disease SNP or in LD with a disease SNP, and its p-value is less than 7.5×10^{-6} ($1/133750$ total SNPs) using a specific method, then the peak SNP is correctly identified as disease SNP by that method. Similarly, if a peak SNP is neither a disease SNP nor in LD with a disease SNP, and if its p-value is greater than 7.5×10^{-6} using a specific method, then the peak SNP is correctly identified as non-disease SNP by that method. For example, rs2305347 and rs2250054 under no PS, rs8071679 under moderate PS, and rs880295 under

severe PS were correctly identified as true negatives by PCAPS, but as false positives by PCA in Table S2. As expected, Tables S2 and S3 indicate that PCA-TW outperforms PCA-10 as more PCs are adjusted in the association models for PCA-TW.

Table 1 shows the performance of no adjustment, PCA, EMMAX and PCAPS methods by using the nine peak associated SNPs in GWAS simulation by PS status. Without adjustment, all non-disease SNPs were incorrectly identified as disease SNPs. Under no, moderate and severe PS, 100%, 50% and 17% non-disease SNPs, respectively, were incorrectly identified as disease SNPs by PCA. No non-disease SNPs were incorrectly identified as disease SNPs by PCAPS. As expected, the false negative rates tend to be inflated from no to moderate, and moderate to severe PS for both PCA and PCAPS. Regarding positive predictive value, 100% of PCAPS identified disease SNPs were true disease SNPs, while only 67% – 84% PCA identified disease SNPs were true disease SNPs. The accuracy rate is 78% by PCA, while PCAPS accuracies are in the range of 89% – 100%. PCAPS demonstrates advantages over PCA by consistently yielding overall low false positive rates (0% vs. 40%), high positive predictive values (100% vs. 79%) and high accuracy rates (93% vs. 78%) regardless of the level of PS. The performance of EMMAX is somewhat between the PCA and PCAPS.

Figure 2 shows Q-Q plots for the visualization of PCA and PCAPS under moderate population stratification. For PCA-TW, many markers have p-values that exhibit modest inflation. PCA-10 and PCAPS-10 p-values fit the expected distribution well. Figure 2 demonstrates that PCAPS-TW performs better than PCA-10 and PCAPS-10 with less significant associations than one would expect by chance. The performance of EMMAX was intermediate to the PCA and PCAPS results (data not shown).

Table 2 shows the performance of no adjustment, PCA, EMMAX and PCAPS methods by using all SNPs in GWAS simulation under moderate PS. Without adjustment, all non-disease SNPs were incorrectly identified as disease SNPs. Without adjustment, 99.6% non-disease SNPs were incorrectly identified as disease SNPs. With PCA-10, PCA-TW, EMMAX, and PCAPS-10 adjustment methods, approximately 40 – 50% non-disease SNPs were incorrectly identified as disease SNPs. No non-disease SNPs were incorrectly identified as disease SNPs by PCAPS-TW. The false negative rates were 0% for no adjustment and EMMAX, and approximately 10% for all other methods. Regarding positive predictive value, 100% of PCAPS identified disease SNPs were true disease SNPs, while only 57% – 67% PCA-10, PCA-TW, EMMAX or PCAPS-10 identified disease SNPs were true disease SNPs. The accuracy rate is 93% by PCAPS-TW, while PCA-10, PCA-TW, EMMAX and PCAPS-10 accuracies are in the range of 67% – 74%. PCAPS demonstrates advantages over PCA and EMMAX by consistently yielding overall low false positive rates (0% vs. >40%), high positive predictive values (100% vs. <67%) and high accuracy rates (93% vs. <74%). The performance of EMMAX is somewhat between the PCA and PCAPS.

Table 3 gives the comparison of estimated odds ratio for the six pre-defined SNPs between PCA and PCAPS under moderate PS in GWAS simulation. For all three pre-defined disease SNPs with OR=1.5 (rs1302756, rs214445, rs919276), PCAPS yielded consistently less-biased estimates. For two out of three pre-defined disease SNPs with OR=2 (rs1334997, rs12680139, rs2598435), PCAPS yielded less-biased estimates. Table S4 gives the

comparison of estimated odds ratio for the six pre-defined SNPs between PCA-TW and PCAPS-TW under moderate PS in GWAS simulation. For two out of three pre-defined disease SNPs with OR=1.5 (rs1302756, rs214445, rs919276) or with OR=2 (rs1334997, rs12680139, rs2598435), PCAPS-TW yielded less-biased estimates. When comparing width of confidence interval for odds ratio between PCA-TW and PCAPS-TW, PCAPS-TW had consistently narrower CIs and thus yielded less-biased estimates (Table S4).

Results of TCGT Data Analysis

GWAS analysis results for TGCT data were published by Kanetsky et al. (2009). The estimated genomic control inflation factor was 1.013 indicating mild PS. We conducted association analysis between TGCT and the six markers from *KITLG* gene using PCA and PCAPS. Table 4 shows the risk allele frequency and per allele odds ratio of the association between TGCT and six selected SNP markers. All estimated ORs are similar for all six SNPs regardless of the method used. The similarity of results is due to the very low-level population stratification in this study population and very high LD correlations among these SNPs. Table 4 also shows 95% confidence intervals for the estimated OR under PCA and PCAPS adjustments. Our PCAPS method resulted in more precise confidence intervals compared to the PCA method, even under conditions of low PS. This gain in efficiency may be attributed to over-adjustment for PS in the PCA approach which included 10 PCs as opposed to only 2 additional covariates that were included in the PCAPS approach.

Discussion

Propensity scores are widely used to correct for confounding effects in general observational studies. Population stratification (PS) in GWAS occurs when cases and controls have different population genetic backgrounds, which is a special case of confounding. PCA has been used to correct for PS in population based genetic association studies. Here, we use a combined approach that uses propensity scores and PCA to improve on the correction for confounding. The PCAPS values are propensity scores, which are calculated using significant PCA values by TW test. Intuitively, using propensity scores to summarize PCA values rather than using the PCAs themselves has the advantage of summarizing and capturing the variability across PCAs and capitalizes on the propensity score's balancing properties. PCAPS efficiently extends our previously developed GPS approach (Zhao et al. 2009; Zhao et al. 2012) to GWAS data to correct for bias due to population stratification. Unlike PCA, PCAPS values are calculated for each specific SNP and for each individual using TW significant PCs. Hence, this method is able to correct for PS bias and downweight the influence of outliers at both the SNP and individual levels. In contrast, PCA values are calculated based on the profile of all SNPs for a given individual and the same values are adjusted for each SNP using an arbitrarily selected identical set of 10 PCs. It may appear that using PCs to construct the stratification score is equivalent to including PCs as covariates in a model as is done in Eigenstrat (Epstein et al. 2007). However, this is not the case, as the variance estimators are different. Extensive experience with propensity scores for prospective data (Lunceford and Davidian 2004), as well as simulations performed by Epstein et al. (2007) attest to the validity of the stratification variance estimators (Allen et al.

2010). In contrast, McPeck and Abney (2008) have shown a variety of situations in which Eigenstrat gives inflated type I error or has diminished power.

Ray and Basu (2017) proposed POM-PS approach based on Allen et al. stratification score, which estimates the probability of case status (disease) conditional on covariates whereas we estimate the probability of the exposure (genetic variant) conditional on covariates. The POM-PS would result in a reduction of power if this score is used for adjustment in the disease model since the score itself is based on disease. Our PCAPS approach, on the other hand, is not estimated based on disease status and thus intuitively should have more power to detect the exposure-disease association. We show in our simulations that PCAPS tends to reduce the number of false-positive findings after removing bias due to confounding even though this may reduce the power to detect a true association under certain conditions. However, consistent with the goal of GWAS to maximize true positive associations, it is preferable to avoid false-positive associations, which may lead to unnecessary follow-up or expensive validation studies.

Price et al. (2010) reviewed methods that correct for stratification while accounting for population structure as well as data that contain family structure or cryptic relatedness. A limitation of PCAPS is that it may not be able to directly account for family structure or cryptic relatedness (Voight and Pritchard 2005; Weir et al. 2006). A possible solution is to calculate principal components using SNP loadings (Patterson et al.), which measure the correlation of each SNP to a given PC in PCA. Further studies are warranted to explore how PCAPS might be able to correct for family structure and cryptic relatedness by using SNP loadings from a set of unrelated samples, either by using a different set of samples from those in the disease study or by using an unrelated subset of samples from the disease study (Zhu et al. 2008). Another approach to the study of admixed populations in family data is to calculate PCAPS based on PCs that take into account the family structure (de Andrade et al. 2015). Zhang and Pan (2015) proposed a hybrid approach that combines the advantages of PCA and LMM. The hybrid approach combines the advantages of PCA and LMM by separating out the genetic confounders and environmental confounders. Currently, the PCAPS is only proposed to capture the genetic confounders. However, PCAPS can be readily extended to the hybrid approach by using the top q PCs and random effect to calculate PCAPS. If there are known environmental confounders, these confounders can be used directly to define and calculate PCAPS.

For choosing the PCs, Li et al. (2009) proposed a PC-Finder procedure using a distance-based regression model to identify a minimal set of PCs while permitting an effective correction for PS. This procedure could be useful for removing PCs that do not appear to have “a significant effect” in reducing the level of PS. This is important when selected PCs are directly adjusted for to correct for PS, such as in the PCA approach. However, when applying PCAPS, only the propensity score is directly adjusted for to correct for PS which is calculated using many PCs. Including a few less influential PCs in the calculation of the propensity score is of less concern and has less impact (Drake 1993).

There are several advantages to using PCAPS over PCA in adjusting for population stratification. The direct consequence of using PCAPS is dramatically reduced data

dimensionality. PCAPS offers distinct advantages because it can control for numerous PCs and environmental confounders simultaneously by reducing to a single scalar variable; this greatly simplifies model building and estimation. Furthermore, it has been shown that adjusting for the estimated propensity score is better at removing bias than adjusting for the true propensity score (Cepeda et al. 2003). This is because adjusting for the estimated propensity score can remove both systematic and chance imbalances, while adjusting for the true propensity score removes only systematic imbalances. In addition, it has been shown that misspecification of an outcome model is more serious than misspecification of a propensity score model; hence, the PCAPS approach is more robust to model misspecification (Drake). Finally, it may seem at first glance that using PCs to construct the propensity scores is equivalent to including PCs as covariates in a model such as PCA. However, this is not the case, as the estimated PCAPS values are different for different loci, whereas PCA values are always the same at each locus. This provides a unique opportunity for using PCAPS for SNP-specific correction of population stratification in GWAS studies. One advantage of our approach is that “nuisance axes” values can actually be incorporated into the estimation of the SNP-specific PCAPS to capture any additional variation even if it is small since the goal of the propensity score estimation is not inference but to obtain the best probability estimates. This distinguishes PCAPS from the limitations of using PCA alone. Regarding outliers, the propensity scores are bounded between 0 and 1; hence, PCAPS will down weight the effect of outlier SNPs unlike PCAs which are not bounded.

Our proposed PCAPS method provides an innovative and practical way of correcting population stratification in GWAS. For full GWAS data, all SNPs with p-value less than the genome-wide threshold of $5e-8$ could be used to define PCAPS. If this number is still very large, then only most biologically relevant-SNPs may be considered to define PCAPS. For each selected SNP (or all SNPs), all PCs that have significantly large genetic variation according to the TW test should be used to calculate SNP-specific PCAPS. The SNP-specific PCAPS is then added in the association model for adjustment. Our simulation studies demonstrate that PCAPS outperforms PCA and EMMAX when used to correct for bias due to population stratification in GWAS. We show that PCAPS is able to correctly identify additional false positives (i.e. true negatives) compared with PCA and EMMAX regardless of the level of PS in the study population. PCAPS also consistently yields narrower confidence intervals of OR than PCA as shown in simulations and our TGCT analysis. The advantages of our approach are the uniform selection of PCAPS scores, the ability to automatically account for the effect of outliers, and its easy implementation in standard statistical packages such as SAS, R or Stata. Since we define a general class of models that specify the potential relation among disease, genotype, and genetic covariates as $f(E(D|G, X)) = \eta$, where $f(\cdot)$ is a link function, the PCAPS approach is readily generalizable to other traits, such as continuous traits by using a generalized linear model approach. In fact, for binary traits (or other non-collapsible models such as Cox proportional hazards model) it is recommended that the PCAPS be included as an inverse-probability weight rather than a covariate adjustment. For continuous traits or collapsible models, PCAPS adjustment in the model is fine (Wan and Mitra 2016). Further studies are needed to compare the PCAPS and other methods for a continuous trait and to assess the performance of PCAPS in adjusting for PS bias in next-generation sequencing studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Airy G (1838): "On the intensity of light in the neighbourhood of a caustic," *Trans. Cambr. Phil. Soc.*, 379–402.
2. Allen A, Epstein MP and Satten GA (2010): "Score-based adjustment for confounding by population stratification in genetic association studies," *Genet Epidemiol*, 34(5), 383–385. [PubMed: 20127852]
3. Bouaziz M, Ambroise C and Guedj M (2011): "Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies," *PLoS One*, 6(12), e28845. [PubMed: 22216125]
4. Cepeda MS, Boston R, Farrar JT and Strom BL (2003): "Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders," *Am J Epidemiol*, 158(3), 280–287. [PubMed: 12882951]
5. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedón JC, Redline S, Papanicolaou GJ, Thornton TA, Laurie CC, Rice K and Lin X (2016): "Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models," *Am J Hum Genet*, 98(4), 653–666. [PubMed: 27018471]
6. de Andrade M, Ray D, Pereira AC and Soler JP (2015): "Global individual ancestry using principal components for family data," *Human Heredity*, 80(1), 1–11. [PubMed: 26159893]
7. Devlin B and Roeder K (1999): "Genomic control for association studies," *Biometrics*, 55(4), 997–1004. [PubMed: 11315092]
8. Dominici D and Maier RS (2008): *Special Functions and Orthogonal Polynomials*, American Mathematical Society
9. Drake C (1993): "Effects of misspecification of the propensity score on estimators of treatment effect," *Biometrics*, 49(4), 1231–1236.
10. Epstein MP, Allen AS and Satten GA (2007): "A simple and improved correction for population stratification in case-control studies," *Am J Hum Genet*, 80(5), 921–930. [PubMed: 17436246]
11. Epstein MP, Duncan R, Broadaway KA, He M, Allen AS, and Satten GA (2012): "Stratification-score matching improves correction for confounding by population stratification in case-control association studies," *Genet Epidemiol*, 36(3), 195–205. [PubMed: 22714934]
12. Feng Q, Abraham J, Feng T, Song Y, Elston RC and Zhu X (2009): "A method to correct for population structure using a segregation model," *BMC Proc3 (Suppl 7)*, S104. [PubMed: 20017968]
13. Hastings SP and McLeod JB (1980): "A boundary value problem associated with the second Painleve transcendent and the Korteweg-de Vries equation," *Arch. Ration. Mech. An.*, 73(1), 31–51.
14. Imbens GW (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 86(1), 4–29.
15. Johnstone IM (2001): "On the distribution of the largest eigenvalue in principal components analysis," *Ann Stat*, 29(2), 295–327.
16. Kanetsky PA, Mitra N, Vardhanabhuti S, Li M, Vaughn DJ, Letrero R, Ciosek SL, Doody DR, Smith LM, Weaver J, Albano A, Chen C, Starr JR, Rader DJ, Godein AK, Reilly MP, Hakonarson H, Schwartz SM and Nathanson KL (2009): "Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer," *Nat Genet*, 41, 811–815. [PubMed: 19483682]
17. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, Sabatti C and Eskin E (2010): "Variance component model to account for sample structure in genome-wide association studies," *Nat Gene*, 42, 348–354.

18. Kang SJ, Larkin EK, Song Y, Barnholtz-Sloan J, Baechle D, Feng T and Zhu X (2009): "Assessing the impact of global versus local ancestry in association studies," *BMC Proc* 3(Suppl 7), S107. [PubMed: 20017971]
19. Lee AB, Luca D, Klei L, Devlin B and Roeder K (2010): "Discovering genetic ancestry using spectral graph theory," *Genet Epidemiol*, 34(1), 51–59. [PubMed: 19455578]
20. Li C and Li M (2008): "GWAsimulator: a rapid whole-genome simulation program," *Bioinformatics*, 24(1), 140–142. [PubMed: 18006546]
21. Li Q, Wacholder S, Hunter DJ, Hoover RN, Chanock S, Thomas G and Yu K (2009): "Genetic background comparison using distance-based regression, with applications in population stratification evaluation and adjustment," *Genet Epidemiol*, 33(5), 432–441. [PubMed: 19140130]
22. Li Q, and Yu K (2008): "Improved correction for population stratification in genomewide association studies by identifying hidden population structures," *Genet Epidemiol*, 32(3), 215–226. [PubMed: 18161052]
23. Lin DY and Zeng D (2011): "Correcting for population stratification in genomewide association studies," *JASA*, 106(495), 997–1008. [PubMed: 22467997]
24. Liu L, Zhang D, Liu H and Arendt C (2013): "Robust methods for population stratification in genome wide association studies," *BMC Bioinformatics*, 14, 132. [PubMed: 23601181]
25. Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K and Trucco M (2008): "On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants," *Am J Hum Genet*, 82(2), 453–63. [PubMed: 18252225]
26. Lunceford JK and Davidian M (2004): "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Stat Med*, 23(19), 2937–2960. [PubMed: 15351954]
27. McPeck M and Abney M (2008): "Association testing with principal-components-based correction for population stratification," *The American Society of Human Genetics*, 11 13, 2008, Philadelphia, PA.
28. Patterson N, Price AL and Reich D (2006): "Population structure and eigenanalysis," *PLoS Genet*, 2, e190. [PubMed: 17194218]
29. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D (2006): "Principal components analysis corrects for stratification in genome-wide association studies," *Nat Genet*, 38, 904–909. [PubMed: 16862161]
30. Price AL, Zaitlen NA, Reich D and Patterson N (2010): "New approaches to population stratification in genome-wide association studies," *Nat Rev Genet*, 11(7), 459–463. [PubMed: 20548291]
31. Pritchard JK, Stephens M, Rosenberg NA and Donnelly P (2000): "Association mapping in structured populations," *Am J Hum Genet*, 67(1), 170–181. [PubMed: 10827107]
32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and Sham PC (2007): "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, 81(3), 559–575. [PubMed: 17701901]
33. Ray D and Basu S (2017): "A novel association test for multiple secondary phenotypes from a case-control GWAS," *Genet. Epidemiol*, 41(5), 413–426. [PubMed: 28393390]
34. Rosenbaum PR and Rubin DB (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41–55.
35. Tracy CA and Widom H (1993): "Level-spacing distributions and the Airy kernel," *Phys Lett B*, 305, 115–118.
36. Tracy CA and Widom H (1994): "Level-spacing distributions and the Airy kernel," *Commun Math Phys*, 159, 151–174.
37. Tracy CA and Widom H (1996): "On orthogonal and symplectic matrix ensembles," *Commun Math Phys*, 177, 727–754.
38. Voight BF and Pritchard JK (2005): "Confounding from cryptic relatedness in case-control association studies," *PLoS Genet*, 1:e32. [PubMed: 16151517]
39. Wan F and Mitra N (2016): "An evaluation of bias in propensity score adjusted non-linear regression models," *Statistical Methods in Medical Research*, 0(0), 1–17.

40. Wang D, Sun Y, Stang P, Berlin JA, Wilcox MA and Li Q (2009): "Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: Principal-component analysis versus multidimensional scaling," *BMC Proc* 3(Suppl 7), S109. [PubMed: 20017973]
41. Weir BS, Anderson AD and Hepler AB (2006): "Genetic relatedness analysis: modern data and new challenges," *Nat Rev Genet*, 7(10), 771–780. [PubMed: 16983373]
42. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ and Tiwari HK (2010): "Mixed linear model approach adapted for genome-wide association studies," *Nat Genet*, 42, 355–360. [PubMed: 20208535]
43. Zhang Y, Guan W and Pan W (2013): "Adjustment for population stratification via principal components in association analysis of rare variants," *Genet Epidemiol*, 37(1), 99–109. [PubMed: 23065775]
44. Zhang Y and Pan W (2015): "Principal component regression and linear mixed model in associaiton analysis of structured samples: competitors or complements?," *Genet Epidemiol*, 39(3), 149–155. [PubMed: 25536929]
45. Zhang Y, Shen X and Pan W (2013): "Adjusting for population stratification in a fine scale with principal components and sequencing data," *Genet Epidemiol*, 37(8), 787–801. [PubMed: 24123217]
46. Zhao H, Rebbeck TR and Mitra N (2009): "A propensity score approach to correction for bias due to population stratification using genetic and non-genetic factors," *Genet Epidemiol*, 33(8), 679–690. [PubMed: 19353632]
47. Zhao H, Rebbeck TR and Mitra N (2012): "Analyzing genetic association studies with an extended propensity score approach," *Stat Appl Genet Mol Biol*, 11(5), ISSN (Online) 1544–6115, DOI: 10.1515/1544-6115.1790.
48. Zhu X, Li S, Cooper RS and Elston RC (2008): "A unified association analysis approach for family and unrelated samples correcting for stratificaiton," *Am J Hum Genet*, 82(2), 352–365. [PubMed: 18252216]
49. Zou F, Lee S, Knowles R and Wright FA (2010): "Quantification of population structure using correlated SNPs by shrinkage principal components," *Human Heredity*, 70(1), 9–22. [PubMed: 20413978]

Figure 1. Manhattan plot of GWAS simulation results by population stratification (PS).
X-axis: Chromosomes 1–3, 6–8, and 17–19; Y-axis: $-\log(\text{p-value})$. Red line shows the threshold to identify top 9 SNPs in each sub-figure.

Figure 2. Q-Q plot of GWAS simulation data under moderate population stratification.
 X-axis: expected $-\log_{10}$ (p-value); Y-axis: observed $-\log_{10}$ (p-value). The genomic lambda values are 0.99984, 1.06462, 0.99787 and 0.98565 for PCA-10, PCA-TW, PCAPS-10 and PCAPS-TW, respectively.

Table 1.

Performance of no adjustment, PCA, EMMAX and PCAPS evaluated using the nine most significantly associated peak SNPs in GWAS simulation under each scenario

PS	Method	Nine Peak SNPs*				False Positive		False Negative		Positive Predict		Accuracy	
		TP	FP	FN	TN	Rate (95%CI)	FP / (FP+TN)	Rate (95% CI)	FN / (FP+FN)	Value (95%CI)	TP / (TP+FP)	Rate (95% CI)	(TP+TN) / (TP+FP+FN+TN)
No	No Adjustment	7	2	0	0	100% (16 – 100%)	0%	0% (0 – 41%)	0%	78% (40 – 97%)	78%	78% (40 – 97%)	78%
	PCA	7	2	0	0	100% (16 – 100%)	0%	0% (0 – 41%)	0%	78% (40 – 97%)	78%	78% (40 – 97%)	78%
	EMMAX	7	2	0	0	100% (16 – 100%)	0%	0% (0 – 41%)	0%	78% (40 – 97%)	78%	78% (40 – 97%)	78%
	PCAPS	7	0	0	2	0% (0 – 84%)	0%	0% (0 – 41%)	0%	100% (59 – 100%)	100%	100% (66 – 100%)	100%
Moderate	No Adjustment	7	2	0	0	100% (16 – 100%)	0%	0% (0 – 41%)	0%	78% (40 – 97%)	78%	78% (40 – 97%)	78%
	PCA	6	1	1	1	50% (1 – 99%)	14%	14% (1 – 58%)	14%	84% (42 – 100%)	84%	78% (40 – 97%)	78%
	EMMAX	7	1	0	1	50% (1 – 99%)	0%	0% (0 – 41%)	0%	88% (47 – 100%)	88%	89% (52 – 100%)	89%
	PCAPS	6	0	1	2	0% (0 – 84%)	14%	14% (1 – 58%)	14%	100% (54 – 100%)	100%	89% (52 – 100%)	89%
Severe	No Adjustment	3	6	0	0	100% (54 – 100%)	0%	0% (0 – 71%)	0%	33% (7 – 70%)	33%	33% (7 – 70%)	33%
	PCA	2	1	1	5	17% (1 – 64%)	33%	33% (1 – 91%)	33%	67% (9 – 99%)	67%	78% (40 – 97%)	78%
	EMMAX	2	1	1	5	17% (1 – 64%)	33%	33% (1 – 91%)	33%	67% (9 – 99%)	67%	78% (40 – 97%)	78%
	PCAPS	2	0	1	6	0% (0 – 46%)	33%	33% (1 – 91%)	33%	100% (16 – 100%)	100%	89% (52 – 100%)	89%
Overall	No Adjustment	17	10	0	0	100% (69 – 100%)	0%	0% (0 – 20%)	0%	63% (42 – 81%)	63%	63% (42 – 81%)	63%
	PCA	15	4	2	6	40% (12 – 74%)	12%	12% (1 – 36%)	12%	79% (54 – 94%)	79%	78% (58 – 91%)	78%
	EMMAX	16	4	1	6	40% (12 – 74%)	6%	6% (1 – 29%)	6%	80% (56 – 94%)	80%	81% (62 – 94%)	81%
	PCAPS	15	1	2	10	0% (0 – 31%)	12%	12% (1 – 36%)	12%	100% (78 – 100%)	100%	93% (76 – 99%)	93%

PCA, principal components analysis correcting for the first 10 principal components; EMMAX, efficient mixed-model association expedited; PCAPS, principal components and propensity score correcting for Tracy-Widom significant principal components; SNP, single nucleotide polymorphism; PS, population stratification; CI, confidence interval. The 95% confidence intervals were calculated using the Binomial distribution.

* Positive disease SNP is the pre-defined SNP with odds ratio 1.5 or 2, or linkage disequilibrium (r^2 0.3) with pre-defined disease SNP. If p-value is less than 7.5×10^{-6} for a peak SNP using a given method, then the peak SNP is identified as disease SNP by that method. Otherwise, the peak SNP is identified as non-disease SNP. TP – true positive, i.e. identified disease SNP is disease SNP; FP – false positive, i.e. identified disease SNP is non-disease SNP; FN – false negative, i.e. identified non-disease SNP is disease SNP; TN – true negative, i.e. identified non-disease SNP is non-disease SNP.

Table 2.

Performance of no adjustment, PCA, EMMAX and PCAPS evaluated using the significantly associated SNPs from each method under moderate population stratification in GWAS simulation

Method	Number of SNPs*				False Positive		False Negative		Positive Predict		Accuracy	
	TP	FP	FN	TN	Rate (95%CI)	FP / (FP+TN)	Rate (95% CI)	FN / (TP+FN)	Value (95%CI)	TP / (TP+FP)	Rate (95%CI)	(TP+TN) / (TP+FP+FN+TN)
No Adjustment	9	1628	0	6	99.6% (99.2 – 99.9%)		0% (0 – 34%)		0.5% (0.3 – 1%)			0.9% (0.5 – 1.5%)
PCA-10	8	6	1	6	50% (21 – 79%)		11% (0.3 – 48%)		57% (29 – 82%)			67% (43 – 85%)
PCA-TW	8	4	1	6	40% (12 – 74%)		11% (0.3 – 48%)		67% (35 – 90%)			74% (49 – 91%)
EMMAX	9	7	0	6	54% (25 – 81%)		0% (0 – 34%)		56% (30 – 80%)			68% (45 – 86%)
PCAPS-10	8	6	1	6	50% (21 – 79%)		11% (0.3 – 48%)		57% (29 – 82%)			67% (43 – 85%)
PCAPS-TW	8	0	1	6	0% (0 – 46%)		11% (0.3 – 48%)		100% (63 – 100%)			93% (68 – 99.8%)

PCA-10, principal components analysis correcting for the first 10 principal components; PCA-TW, principal components analysis correcting for Tracy-Wisdom significant principal components; EMMAX, efficient mixed-model association expedited; PCAPS-10, principal components and propensity score correcting for the first 10 principal components; PCAPS-TW, principal components and propensity score correcting for Tracy-Wisdom significant principal components; SNP, single nucleotide polymorphism; CI, confidence interval. The 95% confidence intervals were calculated using the Binomial distribution.

* Positive disease SNP is the pre-defined SNP with odds ratio 1.5 or 2, or linkage disequilibrium ($r^2 = 0.3$) with pre-defined disease SNP. If p-value is less than 7.5×10^{-6} for a SNP using a given method, then the SNP is identified as disease SNP by that method. Otherwise, the SNP is identified as non-disease SNP. Negative disease SNP is defined as pre-defined SNP with odds ratio 1 or linkage disequilibrium ($r^2 = 0.3$) with pre-defined non-disease SNP. TP – true positive, i.e. identified disease SNP is disease SNP; FP – false positive, i.e. identified disease SNP is non-disease SNP; FN – false negative, i.e. identified non-disease SNP is disease SNP; TN – true negative, i.e. identified non-disease SNP is non-disease SNP.

Table 3. Comparison of estimated odds ratio (OR) for pre-defined disease SNPs under moderate population stratification in GWAS simulation

PS	Chr	rsID	True OR	Estimated Odds Ratio (95% CI)		Absolute % Change from True OR			
				PCA	PCAPS	PCA	PCAPS	Less-biased Method	
Moderate									
2		rs13027563	1.5	1.419 (1.244 – 1.619)	1.430 (1.246 – 1.641)	5.4%	4.7%	PCAPS	
6		rs214445	1.5	1.517 (1.338 – 1.719)	1.495 (1.311 – 1.704)	1.1%	0.3%	PCAPS	
19		rs919276	1.5	1.466 (1.239 – 1.736)	1.486 (1.239 – 1.783)	2.3%	0.9%	PCAPS	
1		rs1334997	2	2.070 (1.757 – 2.445)	1.965 (1.653 – 2.331)	3.5%	1.8%	PCAPS	
8		rs12680139	2	1.965 (1.689 – 2.283)	1.908 (1.634 – 2.227)	1.8%	4.6%	PCA	
17		rs2598435	2	2.119 (1.842 – 2.439)	2.066 (1.776 – 2.410)	6.0%	3.3%	PCAPS	

PS, population stratification; Chr, chromosome; rsID, dbSNP rs number; OR, odds ratio; PCA, principal components analysis; PCAPS, principal components and propensity score.

Table 4.

Summary of association between six selected SNP markers and testicular germ cell tumor

Gene	Marker	Nonrisk/risk Allele	RAF		Per Allele OR (95% CI)		Width of CI	
			Controls	Cases	PCA	PCAPS	PCA	PCAPS
<i>KITLG</i>	rs4474514	A/G	0.802	0.899	2.33 (1.68–3.24)	2.03 (1.48–2.79)	1.56	1.31
	rs3782181	T/G	0.798	0.894	2.21 (1.61–3.05)	1.95 (1.43–2.66)	1.44	1.23
	rs1472899	T/C	0.801	0.897	2.18 (1.58–3.00)	1.92 (1.41–2.62)	1.43	1.21
	rs3782179	A/G	0.802	0.896	2.22 (1.61–3.07)	1.95 (1.43–2.67)	1.46	1.24
	rs11104952	C/A	0.802	0.896	2.22 (1.60–3.05)	1.95 (1.43–2.65)	1.45	1.23

Marker, dbSNP rs number; RAF, risk allele frequency; OR, odds ratio; CI, confidence interval; PCA, principal components analysis; PCAPS, principal components and propensity score.