The confounding effect of cryptic relatedness for environmental risks of systolic blood pressure on cohort studies

Kyoko Shibata, Atsushi Hozawa, Gen Tamiya, Masao Ueki, Tomohiro Nakamura, Hiroto Narimatsu, Isao Kubota, Yoshiyuki Ueno, Takeo Kato, Hidetoshi Yamashita, Akira Fukao, Takamas Kayama & Yamagata University Genomic Cohort Consortium

Abstract

The impact of cryptic relatedness (CR) on genomic association studies is well studied and known to inflate false-positive rates as reported by several groups. In contrast, conventional epidemiological studies for environmental risks, the confounding effect of CR is still uninvestigated. In this study, we investigated the confounding effect of unadjusted CR among a rural cohort in the relationship between environmental risk factors (body mass index, smoking status, alcohol consumption) and systolic blood pressure. We applied the methods of population-based whole-genome association studies for the analysis of the genome-wide single nucleotide polymorphism data in 1622 subjects, and detected 20.2% CR in this cohort population. In the case of the sample size, approximately 1000, the ratio of CR to the population was 20.2%, the population prevalence 25%, the prevalence in the CR 26%, heritability for liability 14.3% and prevalence in the subpopulation without CR 26%, the difference of estimated regression coefficient between samples with and without CR was not significant (P-value = 0.55). On the other hand, in another case with approximately >20% heritability for liability, we showed that confounding due to CR biased the estimation of exposure effects.
showed that the degree of confounding due to CR would usually be negligible. However, in contrast, they also reported on studies with sampling biases toward collecting relatives may indeed suffer from excessive rates of false positives. Typically, epidemiological designs in which individuals are ascertained nonrandomly from a closed population, the effect of the influence of close relatives might not be negligible. It is therefore important to correct or account for the confounding effect of CR in epidemiological cohort studies that have collected data from a limited or small-sized sample. However, the knowledge of the confounding effect of CR in epidemiological association studies is still unknown. Here, we examined the confounding effect of CR in the relationship between systolic blood pressure (SBP) and several environmental risk factors (body mass index [BMI], smoking status [daily smoker vs. nonsmoking], and alcohol consumption [drinking vs. abstinence]).

It is also well known that being overweight and obese increases the risk of high blood pressure (Kannel 1967; World Health Organization 2000). However, interpreting the blood pressure–BMI relationship is further complicated by data from other studies, in which there appears to be no correlation between these variables (Roche and Siervoel 1991; Spiegelman et al. 1992; Bunker et al. 1995; Gallagher et al. 1996). In this article, we examined whether the confounding effect of CR might involve in the relationship between blood pressure and BMI. Additional goals of the study were to assess the confounding effect of CR in any potential relationship between blood pressure and the risk of other environmental factors; for example, smoking and alcohol consumption. There are several studies that examined the relationship between alcohol consumption or smoking and blood pressure in a Japanese population (Kiyohara et al. 1995; Minami et al. 2002; Ohmori et al. 2002). Ohira et al. (2009) looked into the effect of habitual alcohol intake on ambulatory blood pressure among Japanese men, which was associated with increased BP in the morning. Minami et al. (1999) studied the effects of smoking cessation on blood pressure in habitual smokers. However, there were no studies which examined the confounding effect of CR in the relationship between blood pressure and the risk of environmental factors. We aim to address the question of whether CR is likely to be a serious issue for inferring epidemiological relationship between these factors using the cohort study of Takahata residents. First, using the techniques to detect and correct for unrecognized population structure in GWAS, we examined how CR was presented in the sampling. Next, we tested the assumption of parallel regressions to examine whether its confounding effect as a covariate affected on environmental risk factors in difference setting (sample size, ratio of CR to the population, prevalence in CR, the population prevalence). Then, we applied multiple regression analysis to these data with and without CR in order to examine the differences obtained in estimating the regression coefficient.

Methods

Analysis of real data

We used the genome-wide 657,366 single nucleotide polymorphism (SNP) data and SBP as a phenotype in the cohort study of Takahata. We selected BMI, variables for smoking status (1: nonsmoking vs. 2: daily smoker), and alcohol consumption (1: abstention vs. 2: drinking) as environmental risk factors, and gender (1: male; 2: female) and age as covariates. Weight, height, and SBP were measured and standardized in the Takahata cohort design. We examined the relationship between individuals genetic background by the PLINK (Purcell et al. 2007) and the EIGENSTRAT methods (Price et al. 2006). We detected relationships between subjects using identity by descent (IBD) probability as a measurement.

Multiple regression analyses

We sampled our cohort population in a difference setting; sample size (approximately 1000, 400, or 500), the ratio of CR to the population (approximately 20%, 40%, or 50%), the prevalence in the subpopulation without CR (26%, 50%, or 76%), the prevalence in CR (26% fixed), the population prevalence (25%, 40%, or 50%), and heritability for liability (approximately 14%, 22%, or 32%) in our cohort population. Here, we estimated heritability for liability using the formula \(\frac{(x_p - x_q)}{\alpha_p}\), where \(\alpha_p\) denotes the expected proportion of alleles shared IBD (i.e., \(\rho = 2^{-R}\), where \(R\) denotes the degree of relationship), \(x_p\) denotes the difference between mean value in the subpopulation without CR and threshold, \(x_q\) denotes the difference between mean value in the subpopulation without CR and mean value in CR, and \(\alpha_p\) denotes the difference between population mean and mean value in the group of affected individuals (Yasuda 2007). In our context, \(\rho\) can be defined as the sum of expected proportion of allele shared IBD for all of the degree of relationship in our detected CR (i.e., \(R = 0, 1, 2\)). First, we tested the assumption of parallel regressions to examine whether the confounding effect of CR as a covariate affected on environmental risk factors; that is, the following test was performed,

\[ y_k = a_k + b_{ak}x_k (k = 1, 2, n = 1, 2, \ldots, 6) \]

Null hypothesis: \(b_{n1} = b_{n2}\), alternative hypothesis: \(b_{n1} \neq b_{n2}\).
where $k$ is the number of population (i.e., population with and without CR), and $n$ is the number of independent variables and covariates. Then, we applied multiple regression analysis to these data with and without CR in order to examine the differences obtained in estimating the regression coefficient.

**Cohort description**

We performed an analysis of the data collected within a closed, small prospective study, concerned with various risk factors for common diseases.

**Subject recruitment**

The Takahata cohort was established for a baseline survey in a small rural town, Takahata in Yamagata Prefecture from 2006 to 2008. The total population size has been constant, approximately 25,000 throughout this period. The Takahata cohort has become part of our large genomic cohort initiative, the Yamagata cohort, which is now ongoing in the urban prefectural capital, Yamagata City, having approximately 250,000 residents. We used genomic DNAs from 1622 individuals who completed the questionnaire for environmental exposures and informed consent for our modern prospective genomic cohort study. This cohort study was performed under the approval by the Committee on Ethics at Yamagata University and all other institutions involved.

**Genotyping**

Using genomic DNAs from the Takahata population, we carried out genotyping for 657,366 SNPs using the Infinium Assay with Human660W-Quad BeadChip (Illumina, San Diego, CA) according to the standard procedure provided by Illumina.

**Results**

Figure 1 shows the relationship between total 1622 subjects with an IBD probability with regard to an identity by distance. First step, we removed individuals for low genotyping ($P \leq 0.05$) from total 1622 sample using the PLINK method. By the PLINK and the EIGENSTRAT methods, we detected a relationship between subjects with an IBD probability $>1/4$ as a CR of 326 subjects (i.e., monozygotic twins, dizygotic twins, full-sibs, parent-offspring, half-siblings, grandparent, grandchild, aunt/uncle, and niece/nephew) in the sample of 1617 subjects. Figure 2 shows the relationship between 1291 subjects with an IBD probability with regard to an identity by distance after removed a CR of 326 subjects. Next, we removed subjects medically treated for blood pressure from samples with and without CR, respectively. Using the sample with and without CR, the sample sizes were 1039 and 829 individuals, respectively. We analyzed the data as described above. In this case, the heritability for liability was 14.3%. In the multiple regression analysis, the regression model found from the sampling data with CR was as follows:

$$
0.60 \cdot 0.5 \\
I B D \text{ probability} \\
0.70 \cdot 0.75 \\
I B S \text{ distance} \\
0.80 \cdot 0.85
$$
Note that adjusted R-squared for equation (1) was 0.14.

The regression model found from the sampling data without CR was as follows:

\[
\text{Systolic} = 82.52^* + 1.31^* \text{BMI} - 0.80 \times \text{Alcohol} - 0.66 \times \text{Smoking} + 0.38^* \times \text{Age} - 0.31^* \text{Gender} \tag{1}
\]

where “Systolic” denotes systolic blood pressure and * indicates significant, P-values <0.05. Note that adjusted R-squared for equation (1) was 0.14.

The regression model found from the sampling data for equations (3) and (4), the difference between (Table 2). By testing the assumption of parallel regression coefficient in regression modeling that the presence of CR is apparently affected by estimating regression coefficient in regression modeling (Table 1). We tested the assumption of parallel regressions for equations (1) and (2). The difference between equations (1) and (2) was not significant (F-value = 0.35, P-value = 0.55).

Furthermore, as described in Methods, we analyzed under the following difference setting, conditional on the fixed prevalence in CR of 26%. The regression model found from the sampling data in the sample size of 400, the ratio of CR to the population 52.5%, prevalence in the population 40%, prevalence in the subpopulation without CR 55%, prevalence in CR 26%, the heritability for liability 24.2% was as follows:

\[
\text{Systolic} = 141.04^* - 0.12 \times \text{BMI} - 6.11^* \times \text{Alcohol} + 5.34^* \times \text{Smoking} - 0.04 \times \text{Age} - 1.24 \times \text{Gender} \tag{2}
\]

Note that adjusted R-squared for equation (3) was 0.25.

In contrast, the regression model found from the sampling without CR (sample size 190) was as follows:

\[
\text{Systolic} = 145.97^* + 0.04 \times \text{BMI} + 1.28^* \times \text{Alcohol} - 0.39 \times \text{Smoking} - 0.13^* \times \text{Age} - 0.92^* \times \text{Gender} \tag{3}
\]

Note that adjusted R-squared for equation (4) was 0.16.

Compared with equations (3) and (4), the results showed that the presence of CR is apparently affected by estimating regression coefficient in regression modeling (Table 2). By testing the assumption of parallel regressions for equations (3) and (4), the difference between equations (3) and (4) was significant (F-value = 41.83, P-value = 2.103e^{-10}).

The regression model found based on the sampling data in the sample size of 400, the ratio of CR to the population 52.5%, prevalence in the population 50%, prevalence in the subpopulation without CR 76.3%, prevalence in CR 26%, the heritability for liability 32.4% was as follows:

\[
\text{Systolic} = 135.37^* + 0.07 \times \text{BMI} - 12.34^* \times \text{Alcohol} + 4.15 \times \text{Smoking} + 0.17 \times \text{Age} - 1.13 \times \text{Gender} \tag{5}
\]

Note that adjusted R-squared for equation (5) was 0.24.

In contrast, the regression model found from the sampling without CR (sample size 190) was as follows:

\[
\text{Systolic} = 136.68^* + 0.06 \times \text{BMI} - 7.75^* \times \text{Alcohol} - 2.32 \times \text{Smoking} + 0.27^* \times \text{Age} + 0.11^* \times \text{Gender} \tag{6}
\]

Note that adjusted R-squared for equation (6) was 0.42. Compared with equations (5) and (6), the results showed that the presence of CR is apparently affected by estimating regression coefficient in regression modeling (Table 3). By testing the assumption of parallel regressions for equations (5) and (6), the difference between equations (5) and (6) was significant (F-value = 39.74, P-value = 5.720e^{-10}).

The regression model found from the sampling data in the sample size of 500, the ratio of CR to the population 42%, prevalence in the population 40%, prevalence in the

\begin{table}[h]
\centering
\begin{tabular}{lcccc}
\hline
 & \text{Estimated} & \text{Standard} & \text{t-value} & \text{Pr(>|t|)} \\
 & & \text{error} & & \\
\hline
\text{Intercept} & 82.52 & 5.15 & 16.03 & <2e^{-16} \\
Sample with CR$^1$ & 77.11 & 5.49 & 14.06 & <2e^{-16} \\
Sample without CR$^2$ & 1.31 & 0.14 & 9.30 & <2e^{-16} \\
\text{BMI} & 1.36 & 0.16 & 8.34 & 3.23e^{-16} \\
\text{Smoking} & -0.80 & 1.51 & -0.53 & 0.60 \\
Sample with CR & -0.42 & 1.61 & -0.26 & 0.79 \\
Sample without CR & -0.66 & 0.71 & -0.93 & 0.35 \\
\text{Alcohol consumption} & -0.02 & 1.05 & -0.02 & 0.98 \\
\text{Age} & 0.38 & 0.07 & 5.43 & 6.86e^{-08} \\
Sample with CR & 0.41 & 0.05 & 7.99 & 4.62e^{-15} \\
Sample without CR & -3.09 & 1.04 & -2.97 & 0.003 \\
\text{Gender} & -2.46 & 1.01 & -2.44 & 0.01 \\
\hline
\end{tabular}
\caption{Results of the regression coefficients between systolic blood pressure and environmental risk factors in the sample with and without cryptic relatedness (CR): sample size 1039, ratio of CR to the population 20.2%, population prevalence 25%, heritability for liability 14.3%, prevalence in the subpopulation without CR 26%, prevalence in CR 26%.
}
\end{table}

\footnote{1Size of sample with CR was 1039 subjects. From the sampling data with CR, equation (1) in Results were found. Adjusted R-squared = 0.14.}

\footnote{2Size of sample without CR was 829 subjects. From the sampling data without CR, equation (2) in Results were found. Adjusted R-squared = 0.15.}

48 © 2013 The Authors. Molecular Genetics & Genomic Medicine published by Wiley Periodicals, Inc.
The regression model found from the sampling data in the sample size of 500, the ratio of CR to the population 42%, prevalence in the population 50%, prevalence in the subpopulation without CR 67.2%, prevalence in CR 26%, the heritability for liability 31.7% was as follows:

\[
\text{Systolic} = 130.15 + 0.17 \times \text{BMI} - 16.61 \times \text{Alcohol} - 2.58 \times \text{Smoking} + 0.39 \times \text{Age} - 1.44 \\
\times \text{Gender}
\]  
(9)

Note that size of sample with CR was 500 subjects and adjusted \( R \)-squared for equation (9) was 0.21. In contrast, the regression model found from the sampling without CR (sample size 290) was as follows:

\[
\text{Systolic} = 128.16 + 0.22 \times \text{BMI} - 8.11 \times \text{Alcohol} - 2.87 \times \text{Smoking} + 0.42 \times \text{Age} - 1.29 \\
\times \text{Gender}
\]  
(10)

Note that adjusted \( R \)-squared for equation (10) was 0.24. Compared with equations (9) and (10), the results showed that the presence of CR is apparently affected by estimating regression coefficient in regression modeling (Table 4). By testing the assumption of parallel regressions for equations (7) and (8), the difference between equations (7) and (8) was significant (\( F \)-value = 24.96 \( P \)-value = 7.219e-07).

Table 2. Results of the regression coefficients between systolic blood pressure and environmental risk factors in the sample with and without cryptic relatedness (CR): sample size 400, ratio of CR to the population 52.5%, population prevalence 40%, heritability for liability 24.2%, prevalence in the subpopulation without CR 55%, prevalence in CR 26%.

| | Estimated | Standard error | r-value | Pr(>|t|) |
|---|---|---|---|---|
| Intercept | | | | |
| Sample with CR\(^1\) | 15.21 | 5.52 | 25.56 | <2e-16 |
| Sample without CR\(^2\) | 145.97 | 1.67 | 87.40 | <2e-16 |
| BMI | | | | |
| Sample with CR | -0.12 | 0.15 | -0.76 | 0.45 |
| Sample without CR | 0.04 | 0.05 | 0.79 | 0.43 |
| Alcohol consumption | | | | |
| Sample with CR | -11.11 | 1.81 | -3.37 | 0.0008 |
| Sample without CR | 1.28 | 0.56 | 2.30 | 0.022 |
| Smoking status | | | | |
| Sample with CR | 5.34 | 0.71 | 7.57 | 2.69e-13 |
| Sample without CR | -0.39 | 0.25 | -1.59 | 0.11 |
| Age | | | | |
| Sample with CR | -0.14 | 0.82 | -0.48 | 0.63 |
| Sample without CR | -0.13 | 0.22 | -5.47 | 1.44e-07 |
| Gender | | | | |
| Sample with CR | -1.74 | 1.13 | -1.10 | 0.27 |
| Sample without CR | -0.92 | 0.32 | -2.86 | 0.004 |

\(^{1}\)Size of sample with CR was 400 subjects. From the sampling data with CR, equation (3) in Results were found. Adjusted \( R \)-squared = 0.25. 
\(^{2}\)Size of sample without CR was 190 subjects. From the sampling data without CR, equation (4) in Results were found. Adjusted \( R \)-squared = 0.16.

Table 3. Results of the regression coefficients between systolic blood pressure and environmental risk factors in the sample with and without cryptic relatedness (CR): sample size 400, ratio of CR to the population 52.5%, population prevalence 50%, heritability for liability 14.3%, prevalence in the subpopulation without CR 76.3%, prevalence in CR 26%.

| | Estimated | Standard error | r-value | Pr(>|t|) |
|---|---|---|---|---|
| Intercept | | | | |
| Sample with CR\(^1\) | 153.37 | 5.77 | 23.47 | <2e-16 |
| Sample without CR\(^2\) | 136.68 | 2.09 | 65.34 | <2e-16 |
| BMI | | | | |
| Sample with CR | 0.07 | 0.16 | 0.46 | 0.64 |
| Sample without CR | 0.06 | 0.05 | 1.11 | 0.27 |
| Alcohol consumption | | | | |
| Sample with CR | -12.34 | 2.00 | -6.16 | 1.77e-09 |
| Sample without CR | -7.75 | 0.81 | -9.54 | <2e-16 |
| Smoking status | | | | |
| Sample with CR | 4.15 | 0.74 | 5.62 | 3.60e-08 |
| Sample without CR | -2.32 | 0.32 | -7.23 | 1.24e-11 |
| Age | | | | |
| Sample with CR | 7.17 | 0.99 | 7.84 | 3.56e-13 |
| Sample without CR | 2.78 | 0.22 | 7.46 | 3.89e-12 |
| Gender | | | | |
| Sample with CR | -1.13 | 1.18 | -0.95 | 0.34 |
| Sample without CR | 0.11 | 0.41 | 0.27 | 0.79 |

\(^{1}\)Size of sample with CR was 400 subjects. From the sampling data with CR, equation (5) in Results were found. Adjusted \( R \)-squared = 0.24. 
\(^{2}\)Size of sample without CR was 190 subjects. From the sampling data without CR, equation (6) in Results were found. Adjusted \( R \)-squared = 0.42.
sions for equations (9) and (10), the difference between equations (9) and (10) was significant (F-value = 35.46, P-value = 3.915e-09).

Discussion

We detected 20.2% CR of the sample in Takahata cohort study. In our multiple regression models using sample size (N = 1000), there is no significant difference of regression coefficients in the sample with and without CR. In contrast, in the case that the population prevalence of SBP 40–50%, the prevalence in the subpopulation without CR 50–76%, the ratio of CR to the population 42–52%, sample size 400–500 and the heritability for liability 22–32%, the confounding effect of CR in the relationship between SBP and environmental risk factors is not negligible. In general, confounding is a major concern in causal studies because it results in biased estimation of exposure effects. In this respect, our study showed that confounding due to CR biased the estimation of exposure effects in the case of the heritability for liability by approximately >20%. On the other hand, although the number of predictors in the models were included enough (i.e., using independent variables for BMI, alcohol consumption, and smoking status, which were significantly correlated with blood pressure in several research groups [Kannel 1967; Minami et al. 1999; World Health Organization 2000; Ohira et al. 2009]), adjusted R-square values of our regression equation models were not high. A possible explanation of low adjusted R-square values for our models is that other independent variables due to genetic factors might contribute to the SBP phenotype. Genetic factors that confer susceptibility to hypertension were identified in several populations (Jeunemaitre et al. 1992; Hata et al. 1994; Lithton 1996; Cusi et al. 1997). Theoretically, Fisher (1918) indicated that the impact of the effect on the phenotype was evaluated by comparing variances. Falconer and Mackay (1996) showed how the phenotypic variance can be partitioned into causal components of variance using the equation $V_p = V_G + V_D$. In this sense, $V_p$ is the total phenotypic variance, $V_G$ is the total genetic variance consistent with the additive variance ($V_A$), the dominance variance ($V_D$), the interaction variance ($V_I$) and $V_E$ as the

### Table 4. Results of the regression coefficients between systolic blood pressure and environmental risk factors in the sample with and without cryptic relatedness (CR): sample size 500, ratio of CR to the population 43%, population prevalence 42%, heritability for liability 22%, prevalence in the subpopulation without CR 50%, prevalence in CR 26%.

| Variable      | Estimated | Standard Error | t-value | Pr(>|t|) |
|---------------|-----------|----------------|---------|---------|
| Intercept     | 137.18    | 4.91           | 27.92   | <2e-16  |
| Sample with CR | 141.55    | 2.99           | 47.26   | <2e-16  |
| BMI           |           |                |         |         |
| Sample with CR | 0.04      | 0.14           | 0.29    | 0.79    |
| Sample without CR | 0.09     | 0.08           | 1.11    | 0.27    |
| Alcohol consumption | | | | |
| Sample with CR | −9.47     | 1.59           | −5.94   | 5.47e-06|
| Sample without CR | −5.34    | 1.00           | −5.35   | 4.26e-07|
| Smoking status |           |                |         |         |
| Sample with CR | 3.47      | 0.64           | 5.45    | 8.10e-06|
| Sample without CR | 2.30     | 0.44           | −5.18   | 4.26e-11|
| Age           |           |                |         |         |
| Sample with CR | 0.10      | 0.07           | 1.34    | 0.18    |
| Sample without CR | 0.11     | 0.05           | 2.39    | 0.018   |
| Gender        |           |                |         |         |
| Sample with CR | −0.95     | 1.10           | −0.94   | 0.35    |
| Sample without CR | −1.10    | 0.60           | −1.83   | 0.07    |

1Size of sample with CR was 500 subjects. From the sampling data with CR, equation (7) in Results were found. Adjusted R-squared = 0.19.

### Table 5. Results of the regression coefficients between systolic blood pressure and environmental risk factors in the sample with and without cryptic relatedness (CR): sample size 500, ratio of CR to the population 43%, population prevalence 50%, heritability for liability 31.7%, prevalence in the subpopulation without CR 67.2%, prevalence in CR 26%.

| Variable      | Estimated | Standard Error | t-value | Pr(>|t|) |
|---------------|-----------|----------------|---------|---------|
| Intercept     | 130.15    | 5.51           | 23.62   | <2e-16  |
| Sample with CR | 128.16    | 3.92           | 32.73   | <2e-16  |
| BMI           |           |                |         |         |
| Sample with CR | 0.17      | 0.15           | 1.13    | 0.26    |
| Sample without CR | 0.22     | 0.10           | 2.13    | 0.03    |
| Alcohol consumption | | | | |
| Sample with CR | −16.61    | 1.67           | −9.95   | 2.99e-10|
| Sample without CR | −8.11    | 1.24           | −6.53   | 0.0003  |
| Smoking status |           |                |         |         |
| Sample with CR | 2.58      | 0.71           | 3.63    | 0.0003  |
| Sample without CR | 2.87     | 0.65           | −4.40   | 1.53e-05|
| Age           |           |                |         |         |
| Sample with CR | 0.39      | 0.08           | 4.91    | 1.21e-06|
| Sample without CR | 0.42     | 0.05           | 7.72    | 1.99e-13|
| Gender        |           |                |         |         |
| Sample with CR | −1.44     | 1.09           | −1.33   | 0.19    |
| Sample without CR | −1.29    | 0.72           | −1.79   | 0.08    |

1Size of sample with CR was 500 subjects. From the sampling data with CR, equation (9) in Results were found. Adjusted R-squared = 0.21.

2Size of sample without CR was 290 subjects. From the sampling data without CR, equation (10) in Results were found. Adjusted R-squared = 0.24.
environmental variance consistent with the special environmental variance ($V_E$) that refers to the within individual variance arising from localized circumstances, and general environmental variance $V_G$ refers to the environmental variance contributing to between-individual component in origin. Note that the ratio $V_G = V_D$ is the heritability of the character. Moreover, Falconer and Mackay (1996) revealed the existence of two coefficients for genetic variances; one is the coefficient $r$ of the additive variance ($V_A$) which called the coefficient of relationship between the relatives in question, and the other is the coefficient $u$ of the dominance variance ($V_D$) which represents the probability of the relatives having the same genotype through IBD. Using these two coefficients, the total genetic variance is given by $V_G = r V_A + u V_D + 2 \text{cov}_{AD}$, where $\text{cov}_{AD}$ is the covariance of breeding values with dominance deviations (Falconer and Mackay 1996).

According to this mathematical model, it is easily understood how factors associated with genetically close relatives in a sample of ostensibly unrelated individuals contribute to the effect of phenotype; that is, the phenotype is composed of both environmental and genetic elements that contribute to the relationship between relatives. Thus, some of the differences in the estimates of regression coefficients might be because of the adjustment strategies for concomitant confounding effect of CR. Rotimi et al. (1999) examined a familial pattern of blood pressure in a population of Nigerian families and clarified that heritability of <50% for both SBP and diastolic blood pressure (DBP) reinforced the importance of the nonshared familial environmental effects. Thus, one of the approaches to select the best model for the response variable using collected cohort data from a limited or small-sized sample is that the heritability of the blood pressure phenotype might be worth considering.

Historically, the most common statistical approach for dealing with confounding in epidemiology was based on stratification. Typically, given the importance of confounding in epidemiology, statistical methods recommend the removal of significantly confounding samples. However, the resulting removal of samples with confounding factors, the sample size is reduced. As with another possible approach, we are now extensively analyzing this issue by a method incorporating principal components of a large subset of GWAS SNPs as regression covariates. This approach does not waste resources; that is we can use the entire sample. There are some similarities between the approach by Price et al. (2006) and our method; however, in contrast we examine principal components of a large subset of GWAS SNPs to adjust the confounding effect of CR. On the other hand, they examined them only to adjust the population structure. Generally, principal components typically reflect genome-wide factors attributable to the demographic history of the populations studied (Price et al. 2006). In this respect, it still remains to be clarified whether principal components reflect genome-wide factors attributable to CR. We are now extensively analyzing under what condition such approach is plausible.

In conclusion, we found a confounding effect of CR in the relationship between SBP and environmental risk factors was not negligible. In our study, we showed that heritability for liability might reflect on the estimation of regression coefficients between SBP and environmental risk factors, because they vary with environmental risk factors that differ across some unsuspected relatedness. For the genetic case-control studies, test statistics are generally inflated relative to the expectation under the association of an independent sample and without genetic association to the disease. These false positives often are attributed to CR (Devlin and Roeder 1999). Thus, more or less in any other epidemiological investigations that were performed previously, a true effect might be hidden due to confounding arising from CR. In our study, we presented a simply modeling to illustrate the effect of CR on the estimation of coefficients. The size of the CR would have a big impact on the precision of the resulting estimates of coefficients. We are now extensively analyzing this issue in different settings. Various statistical methods have been proposed to take into account confounding factors such as linear mixed-effect models (Demidenko 2004) or methods that adjust data based on a principal components analysis (Price et al. 2006). Sturmer et al. (2005) proposed a method of adjusting for multiple unmeasured confounders in a cohort study. The amount of residual confounding due to unmeasured and poorly measured covariates was important enough to qualitatively change the association between NSAID (nonsteroidal anti-inflammatory drug) use and mortality (Sturmer et al. 2005). After data collection, using these techniques in an epidemiological association study, it might be important to adjust the cryptic relative pairs based on genetic data in the relationship between environmental risk factors and phenotypes.

Acknowledgments

This study was supported by JSPS KAKENHI Grant Number 24659316. This study was fully supported by a Grant-in-aid from the Global Center of Excellence program of the Japan Society for the Promotion of Science, “Formation of an International Network for Education and Research of Molecular Epidemiology (Project Leader: Takamasa KAYAMA).” We thank two reviewers for insightful suggestions.

Conflict of Interest

None declared.
References

Bunker, C. H., F. A. Ukoli, K. A. Matthews, A. M. Kriska, S. L. Huston, and L. H. Kuller. 1995. Weight threshold and blood pressure in a lean black population. Hypertension 143:1203–1218.

Cusi, D., C. Barlassina, T. Azzani, G. Casari, L. Citterio, M. Devoto, et al. 1997. Polymorphisms of alpha-adducin and salt sensitivity in patients with essential hypertension. Lancet 349:1353–1357.

Demidenko, E. 2004. Mixed models: theory and applications. John Wiley and Sons, Inc., Hoboken, NJ.

Devlin, B., and K. Roeder. 1999. Genomic control for association studies. Biometrics 55:997–1004.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to quantitative genetics. Prentice Hall, London, U.K.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinburgh 52:399–433.

Gallagher, D., M. Visser, D. Sepulveda, R. N. Pierson, T. Harris, and S. B. Heymsfield. 1996. How useful is body mass index for comparison of body fatness across age, sex and ethnic groups. Am. J. Epidemiol. 143:228–239.

Hata, A., C. Namikawa, M. Sasaki, K. Sato, T. Nakamura, K. Tamura, et al. 1994. Angiotensinogen as a risk factor for essential hypertension in Japan. J. Clin. Invest. 93:1285–1287.

Jeunemaitre, X., F. Soubrier, Y. V. Kotelevtsev, R. P. Lifton, C. S. Williams, A. Charru, et al. 1992. Molecular basis of human hypertension: role of angiotensinogen. Cell 71:169–180.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, et al. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1723–1723.

Kannel, W. B. 1967. The relation of adiposity to blood pressure and development of hyper tension. The Framingham study. Ann. Intern. Med. 67:48–59.

Kiyohara, Y., I. Kato, H. Iwamoto, K. Nakayama, and M. Minami. 1999. Effects of habitual alcohol intake on ambulatory blood pressure, heart rate, and heart rate variability in Japanese men. Am. J. Hypertens. 12:575–580.

Kiyohara, Y., I. Kato, H. Iwamoto, K. Nakayama, and M. Minami, et al. 1994. Angiotensinogen as a risk factor for essential hypertension in Japan. J. Clin. Invest. 93:1285–1287.

Lifton, R. P. 1996. Molecular genetics of human blood pressure variation. Science 272:676–680.

Minami, J., T. Ishimitsu, and H. Matsuoka. 1999. Effects of smoking cessation on blood pressure and heart rate variability in habitual smokers. Hypertension 33:586–590.

Minami, J., M. Yoshii, M. Todoroki, T. Nishikimi, T. Ishimitsu, T. Fukunaga, et al. 2002. Effects of alcohol restriction on ambulatory blood pressure, heart rate, and heart rate variability in Japanese men. Am. J. Hypertens. 15:125–129.

Ohira, T., T. Tanigawa, M. Tabata, H. Imano, A. Kitamura, M. Kiyama, et al. 2009. Effects of habitual alcohol intake on ambulatory blood pressure, heart rate, and its variability among Japanese men. Hypertension 53:13–19.

Ohmori, S., Y. Kiyohara, I. Kato, M. Kubo, Y. Tanizaki, H. Iwamoto, et al. 2002. Alcohol intake and future incidence of hypertension in a general Japanese population: the Hisayama study. Alcohol. Clin. Exp. Res. 26:1010–1016.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909.

Price, A. L., N. A. Zaitlen, D. Reich, and N. Patterson. 2010. New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11:459–463.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, et al. 2007. PLINK: a toolset for whole-genome association and population based linkage analysis. Am. J. Hum. Genet. 81:559–575.

Roche, A. F., and R. M. Siervo. 1991. Measures of body composition: their relationship to blood pressure and use in epidemiologic research. Ann. Epidemiol. 1:313–320.

Rotimi, C. N., R. S. Cooper, G. Cao, O. Ogunbiyi, M. Ladipo, E. Owoaje, et al. 1999. Maximum-likelihood generalized heritability estimate for blood pressure in Nigerian families. Hypertension 33:874–878.

Spiegelman, D., R. G. Israel, C. Bouchard, and W. C. Willett. 1992. Absolute fat mass, percent body fat, and body-fat distribution: which is the real determinant of blood pressure and serum glucose? Am. J. Clin. Nutr. 55:1033–1044.

Sturmer, T., S. Schneeweiss, J. Avorn, and R. J. Glynn. 2005. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am. J. Epidemiol. 162:279–289.

Voight, B. F., and J. K. Pritchard. 2005. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 1:e32.

World Health Organization. 2000. Obesity: preventing managing the global epidemic. Report of a WHO consultation. World Health Organ. Tech. Rep. Ser. 894:1–253.

Yasuda, N. 2007. A primer of population genetics. Shokabo, Chiyoda-ku, Japan.

Yu, J., G. Pressoir, W. H. Briqqs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208.

Appendix

The following are the contributors of Yamagata University Genomic Cohort Consortium

Steering Committee: Takamasa Kayama¹ (Chair), Hidetoshi Yamashita² (Deputy chair), Akira Fukao³, Isao Kubota⁴, Takeo Kato⁵, Chifumi Kitanaka⁶, Shinya Sato¹, Yoshiyuki Ueno¹¹

K. Shibata et al.
Respiratory, Cardiovascular, and Renal Diseases and Related Traits: Isao Kubota,5,6 Shinya Sato1,5, Tsuneo Konta6, Yoko Shibata6, Tetsu Watanabe6, Shuichi Abe6, Takuya Miyamoto6, Sumito Inoue6, Takehiko Miyashita6, Kazunobu Ichikawa6, Tetsuro Shishido6, Takanori Arimoto6, Hiroki Takahashi6, Satoshi Nishiyama6, and Ami Ikeda6

Metabolic and Neurodegenerative Diseases and Related Traits: Takeo Kato7,8, Makoto Daimon8, Toru Kawanami8, Manabu Wada8, Shigeki Arawaka8, Hidetoshi Oizumi8, Katsuro Kurokawa8, Shingi Susa8, Yuichi Katou8, Wataru Kameda8, Shingo Koyama8, Shigeru Karasawa8, Chifumi Iseki8, and Yoshimi Takahashi8

Cancer and Related Traits: Chifumi Kitanaka9,10, Yoshiyuki Ueno11

Gastrointestinal, Hepatic, and Pancreatic Diseases and Related Traits: Yoshiyuki Ueno11, Sumio Kawata11, Takafumi Saito11, Naohiko Makino11,12, Katsuo Okumoto11, Hiroaki Haga11, Takeshi Sato11, Chikako Sato11, Hisayoshi Watanabe11, Yuko Nishise11, Rika Ishii11, Akiko Matsuda12, and Tomohiro Tozawa11,12

Eye Diseases and Related Traits: Hidetoshi Yamashita2 and Kei Honma2

Cohort Establishment: Akira Fukao3,4, Hiroto Narimatsu3, Kyoko Shibata3, Akiko Miura3, Rina Inoue3, Ai Numazawa3, Kahori Kudo3, Yoko Aita3, Noriko Umezawa3, Yuko Saito3, Yumi Takahashi3, Yuka Suzuki3, Katsumi Otani3, Atsushi Hozawa3, Li Shao4, Masatusugu Orui13, Atsuko Kobayashi14, Yuka Kanoya14, Takiko Hosoya14, Ikuko Suzuki14, Mariko Otake14, Yuko Morikagi14, Akiko Sekimata14, Manami Hiraka14, Yumi Matsuda14, Chika Sato14, Yoko Takeda14, Yoko Matsunami14, Tatsuya Horie14, Shoho Sato14, Mizue Inoue14, and Kaoru Baba14

Genetic, Genomic, and Statistical Analyses: Gen Tamiya15, Masao Ueki15, and Tomohiro Nakamura15

Genotyping: Gen Tamiya15, Jamiyansure Jambaldorj16, and Satoko Araki16

DNA Extraction and Biobanking: Osamu Nakajima17,18

Database Construction: Kazuei Takahashi19 and Kazuo Goto19

Patent Control and Commercialization: Kimishige Ishizaka20

1 Yamaga University, Graduate School of Medicine, Department of Neurosurgery
2 Department of Ophthalmology
3 Advanced Molecular Epidemiology Research Institute (AMERI), Cohort Management Unit
4 Department of Public Health
5 AMERI, Respiratory and Cardiovascular Diseases Research Center
6 Department of Cardiology, Pulmonology, and Nephrology
7 AMERI, Metabolic and Degenerative Diseases Research Center
8 Department of Neurology, Hematology, Metabolism, Endocrinology and Diabetology
9 AMERI, Oncology Research Center
10 Department of Molecular Cancer Science
11 Department of Gastroenterology
12 Division of Endoscopy, Yamagata University Hospital
13 Yamaga Prefectural Tsuruoka Hospital
14 School of Nursing
15 AMERI, Genomic Information Analysis Unit
16 AMERI, Shared Laboratory
17 AMERI, Specimen Management Unit
18 Research Laboratory for Molecular Genetics
19 AMERI, Data Management Unit
20 COME center

*Present address: Hyogo Prefectural Nishinomiya Hospital.