科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月 １０ 日現在

機関番号：１１３０１
研究種目：挑戦的萌芽研究
研究期間：2012～2013
課題番号：２４６５９３１６

研究課題名（和文）循環器コホート分析における血縁構造化の影響の調査

研究課題名（英文）The confounding effect of cryptic relatedness for environmental risks of systolic blood pressure on cohort studies.

研究代表者
柴田　恭子（Shibata, Kyoko）

東北大学・大学病院・特任講師

研究者番号：９０５３５６００

交付決定額（研究期間全体）：（直接経費）　　700,000円、（間接経費）　　210,000円

研究成果の概要（和文）：地域集団サンプル1,617人を用いて、一塩基多型を用いたゲノムワイド関連解析により血縁構造を検出し（20.2%）、極めて近い血縁者を除外した集団と、除外しない集団で、表現型と環境因子との関連性の結果に違いがあるか検証した。表現型データとして血圧値、リスク環境因子としてBMI、喫煙歴、飲酒歴を用いた。交絡因子としての血縁構造化の影響が深刻な問題となる検証結果は少なかったが、遺伝率の大きさによりバイアスが係る可能性が示唆され、今後更なる詳細な検証が必要である。

研究成果の概要（英文）：We investigated the confounding effect of unadjusted cryptic relatedness (CR) among a rural cohort in the relationship between environmental risk factors (body mass index, smoking status, alcohol consumption) and systolic blood pressure (SBP). We applied the methods of population-based whole-genome association studies for the analysis of the genome-wide SNP data in 1622 subjects, and detected 20.2% CR in this cohort population. 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 coeffients between SBP and environmental risk factors, because they vary with environmental risk factors that differ across some unsuspected relatedness.

研究分野：医歯薬学

科研費の分科・細目：社会医学・公衆衛生学・健康科学

キーワード：血縁構造化

１．研究開始当初の背景

疾患発症リスクとなる環境要因を同定することが目的の伝統的なコホート研究では、対象集団の血縁構造が必ずしも十分には考慮されていない。しかし、極めて近い遺伝的背景をもつ血縁者が対象集団内に多数存在した場合には、環境因子と疾患罹患リスクの関連分析において、この遺伝的背景による誤った交絡を大きく受ける可能性がある。これまで、日本で行われてきた古典的手法によるコホート研究は、主に地方の農村部を対象とするものであったが、特にこのような地域集団のコホート研究では、交絡要因として対象集団の血縁構造化は深刻な問題となる。

近年、大規模ゲノム解析データを用いた遺伝学研究では、このような未記録の血縁構造（cryptic relatedness）の計測と調整が可能になっているが、リスク環境要因の発見を目的とする古典疫学の関連分析で同様の調整を試した例はない。

そこで、大規模ゲノムデータを用いた最新の血縁構造の計測法を、古典的コホート研究に応用し、対象集団の血縁構造化の影響を明らかにしようとする本研究課題の着想に至った。

２．研究の目的

本研究では、一塩基多型を用いたゲノムワイド関連解析のデータで血縁構造を計測し、極めて近い血縁者を除外した集団と、除外しない集団で、表現型と環境因子との関連性の違いがあるか検証し、血縁構造化の影響を明らかにする。

表現型データとして、血圧値、リスク環境因子、BMI、喫煙、アルコール摂取量を用いる。

３．研究の方法

プロジェクト前半は、一塩基多型(SNPs)を用いたゲノムワイド関連解析のデータで血縁構造を計測した。近い血縁関係にある者を除外した集団と、除外しない集団で、表現型と環境因子との関連性に違いがあるか統計解析を行った。使用するデータおよび解析ツールは、下記のとおりである。

(1) グローバルＣＯＥプログラム分子疫学コホート研究でリクルートされた約 1,600 人の DNA 検体
(2) サンプル集団約 1,600 人の表現型データとして血圧値、リスク環境因子として、BMI・喫煙
(3) 統計解析ソフトＲ、ゲノムワイド関連解析で血縁構造を計測するための解析プログラム（PLINK, EIGENSTRAT）

プロジェクト後半は、対象とする疾患の有病率、リスク環境因子の効果サイズ、サンプルサイズ、血縁構造化の程度の観点で、疫学研究における血縁構造の影響を以下の解析法により、詳細に調べた。

（１）　血縁構造の計測：グローバルＣＯＥ分子疫学コホート研究において山形県高畠地域で収集された集団サンプルを用いて、ゲノムワイド関連解析により以下のように血縁構造を計測した。

① マイナーアレル頻度＞0.05、ＳＮＰ欠測率＞0.1、ハーディ・ワインベルグ平衡検定ｐ値＞0.05 等の厳しい条件で常染色体上の高精度ＳＮＰジェノタイプを選別した。
② LD-based pruning により、強い連鎖不平衡にないＳＮＰセットを定義した。
③ 以上で定義されたＳＮＰセットを用いて以下の２つの方法を行った。

方法１：任意の２名の血縁度の推定値を両者のＳＮＰセットについて、所与のアレル頻度と遺伝子型のもとで、平均の同型的アレル類似率から推定した同祖的アレル共有確率として計算した。これをサンプルサイズｎについて、n(n-1)/2 ペアの組み合わせ全てで計算した。血縁度の推定値が 1/8 以上で 1st cousin 以上の関係にあるとみなすが、同じ血縁度の推定値を持つ異なる血縁関係のいくつかはアレル共有状態の平均から推定した。たとえば、同じ血縁度の推定値 1/2 でも、親子はアレル共有状態が常に z_1=1（IBD アレルをひとつ共有）であるが、同胞対は z_0=1/4, z_1=1/2, z_2=1/4 となることから見分けられる。明らかになった血縁関係を可能な限り、Y 染色体ＳＮＰｓやミトコンドリアハプロタイプ、HLA 型などを用いて確認・精緻化した。

方法２：上述の方法では、血縁度が離れるほど共有される下のセグメントが徐々に小さくなり、1st cousin よりも離れた血縁構造化（たとえば、異なる由来の民族集団による集団階層化など）を検出することができない。そこで、単純に個人間のユークリッド同型性距離を基準量に用いた主成分分析を行った。とくに、ここでは、各主成分（固有ベクトル）の固有値の分散を用いて Tracy-Widom 検定を行い、どの主成分の固有値も有意にならなくなるまで繰り返しサンプル除去を行った。また、有意な固有値について各個人の主成分得点を記録した。

（２）重回帰解析・共分散分析

近い血縁関係にある者を除外した集団を用いて、従属変数に血圧値（量的表現型）、独立変数にリスク環境因子（BMI、喫煙、アルコール摂取量（連続値））をあてはめ重回帰モデル解析（Ｆ検定）を行った。

また、検体約 1,600 人サンプル集団を用いて、従属変数（血圧値）、独立変数（ＢＭＩ、喫

煙、アルコール摂取量)、共変量（性別、年齢）として共分散分析を行った。

４．研究成果

地域集団サンプル 1,617 人を用いて、一塩基多型を用いたゲノムワイド関連解析により血縁構造（20.0%）を検出し（図１：血縁構造化の様子。図２：血縁構造除去後の集団の様子。)、極めて近い血縁者を除外した集団と、除外しない集団で、表現型（血圧値）と環境因子（ＢＭＩ、喫煙歴、飲酒歴）との関連性に違いがあるか、異なる遺伝率、サンプルサイズ、疾患の有病率等の条件下で検証した。その結果、下記の表１，２，３，４、５に示す。

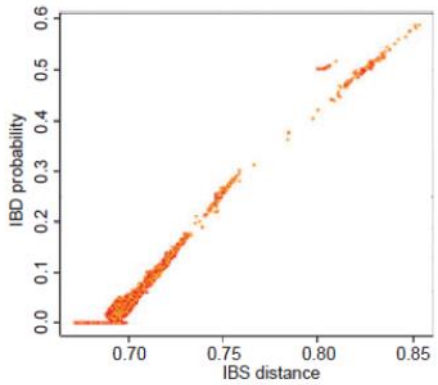

Figure 1. Plot of the relationship between total 1622 subjects with an identity by descent (IBD) probability with regard to an identity by state (IBS) distance; *y*-axis and *x*-axis describe IBD probability and IBS distance, respectively.

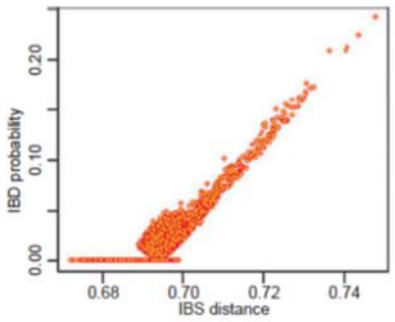

Figure 2. Plot of the relationship between 1291 subjects with an identity by descent (IBD) probability with regard to an identity by state (IBS) distance after which is removed a cryptic relatedness of 326 subjects with an IBD probability >1/4 (i.e., monozygotic twins, dizygotic twins, full-sibs, parent-offspring, half-siblings, grandparent, grandchild, aunt/uncle, and niece/nephew); *y*-axis and *x*-axis describe IBD probability and IBS distance, respectively.

Table 1. 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%.

	Estimated	Standard error	*t*-value	Pr(>|*t*|)
Intercept				
Sample with CR[1]	82.52	5.15	16.03	$<2e^{-16}$
Sample without CR[2]	77.11	5.49	14.06	$<2e^{-16}$
BMI				
Sample with CR	1.31	0.14	9.30	$<2e^{-16}$
Sample without CR	1.36	0.16	8.34	$3.23e^{-16}$
Alcohol consumption				
Sample with CR	−0.80	1.51	−0.53	0.60
Sample without CR	−0.42	1.61	−0.26	0.79
Smoking status				
Sample with CR	−0.66	0.71	−0.93	0.35
Sample without CR	−0.02	1.05	−0.02	0.98
Age				
Sample with CR	0.38	0.07	5.43	$6.86e^{-08}$
Sample without CR	0.41	0.05	7.99	$4.62e^{-15}$
Gender				
Sample with CR	−3.09	1.04	−2.97	0.003
Sample without CR	−2.46	1.01	−2.44	0.01

[1]Size of sample with CR was 1039 subjects. From the sampling data with CR, equation (1) in Results were found. Adjusted *R*-squared = 0.14.
[2]Size of sample without CR was 829 subjects. From the sampling data without CR, equation (2) in Results were found. Adjusted *R*-squared = 0.15.

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	*t*-value	Pr(>|*t*|)
Intercept				
Sample with CR[1]	141.04	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	−6.11	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	−0.040	0.082	−0.48	0.63
Sample without CR	−0.13	0.02	−5.47	$1.44e^{-07}$
Gender				
Sample with CR	−1.24	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	*t*-value	Pr(>|*t*|)
Intercept				
Sample with CR[1]	135.37	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	0.17	0.09	1.89	0.06
Sample without CR	0.27	0.03	7.84	$3.56e^{-13}$
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.

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 42%, population prevalence 40%, heritability for liability 22.1%, prevalence in the subpopulation without CR 50%, prevalence in CR 26%.

	Estimated	Standard error	*t*-value	Pr(>|*t*|)
Intercept				
Sample with CR[1]	137.18	4.91	27.92	$<2e^{-16}$
Sample without CR[2]	141.55	2.99	47.26	$<2e^{-16}$
BMI				
Sample with CR	0.04	0.14	0.29	0.77
Sample without CR	0.09	0.08	1.11	0.27
Alcohol consumption				
Sample with CR	−9.47	1.59	−5.94	$5.47e^{-09}$
Sample without CR	−5.34	1.00	−5.35	$4.26e^{-07}$
Smoking status				
Sample with CR	3.47	0.64	5.45	$8.10e^{-08}$
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 vCR	−1.10	0.60	−1.83	0.07

[1]Size of sample with CR was 500 subjects. From the sampling data with CR, equation (7) in Results were found. Adjusted *R*-squared = 0.19.
[2]Size of sample without CR was 290 subjects. From the sampling data without CR, equation (8) in Results were found. Adjusted *R*-squared = 0.16.

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 42%, population prevalence 50%, heritability for liability 31.7%, prevalence in the subpopulation without CR 67.2%, prevalence in CR 26%.

	Estimated	Standard error	*t*-value	Pr(>|*t*|)
Intercept				
Sample with CR[1]	130.15	5.51	23.62	$<2e^{-16}$
Sample without CR[2]	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	$<2e^{-16}$
Sample without CR	−8.11	1.24	−6.53	$2.99e^{-10}$
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

[1]Size of sample with CR was 500 subjects. From the sampling data with CR, equation (9) in Results were found. Adjusted *R*-squared = 0.21.
[2]Size of sample without CR was 290 subjects. From the sampling data without CR, equation (10) in Results were found. Adjusted *R*-squared = 0.24.

以上の検証結果より、交絡因子としての血縁構造化の影響が深刻な問題となるケースは少なかったが、遺伝率の大きさにより、血縁構造の影響がある可能性が示唆され、今後更なる詳細な検証が必要であることが示された。

５．主な発表論文等
（研究代表者、研究分担者
及び連携研究者には下線）

〔雑誌論文〕（計１件）

The confounding effect of cryptic relatedness for environmental risks of systolic blood pressure on cohort studies. Shibata K, Hozawa A, Tamiya G（以下、9名省略。）、Molecular Genetics & Genomic Medicine, 査読有、1(1):45-53、DOI:10.1002/mgg3.4

〔学会発表〕（計１件）

Shibata K,（以下、10名省略）、The confounding effect of cryptic relatedness for environmental risks of systolic blood pressure on cohort studies. 米国人類遺伝学会、2012年11月8日、米国サンフランシスコ

６．研究組織

(1)研究代表者
柴田　恭子（Shibata, Kyoko）
東北大学・大学病院・特任講師
研究者番号：90535600

(2)研究分担者
田宮　元（Tamiya, Gen）
東北大学・東北メディカル・メガバンク機構・教授
研究者番号： 10317745