論文の内容の要旨

論文題目　虚血性心疾患発症におけるレアバリアントが果たす役割についての網羅的遺伝解析

氏名　家城　博隆

＜序文＞

虚血性心疾患（Coronary artery disease; CAD）は全世界の死因の第一位であり、日本においても患者数は増加傾向にある。医師不足や医療費の観点からも大きな社会問題になっており、CAD の病態の詳細な理解と個別化治療・予防法の開発が求められている。CAD は遺伝的背景をベースにして、喫煙、食生活、生活習慣などの環境因子が相互作用して発症に至ることがわかっており、様々な双子研究では CAD における遺伝的要因は最大 50%程度を占めると報告されている。2000 年代になり、ゲノムワイド関連解析（Genome wide association study; GWAS）が多数行われ、CAD に関連する様々な遺伝的座位が同定され、遺伝的背景が徐々に明らかになってきた。しかし、これまでに GWAS で同定された遺伝的座位の情報をすべて集めても、CAD の遺伝的要因のごく一部しか説明できないことが報告されており missing heritability と呼ばれ問題となっている。GWAS が対象とする遺伝情報はゲノム全域に存在するありふれた遺伝多型（コモンバリアント）であり、稀な多型や変異（レアバリアント）は統計的検出力が不足するために GWAS で検出することができないことが一つの原因であると言われている。近年、全ゲノムシークエンス（Whole genome sequencing; WGS）データを用いて、レアバリアントを網羅的に解析する HEAL method という機械学習手法を用いたフレームワークが提案されている。本研究では、CAD においてレアバリアントが果たす役割を明らかにするために、HEAL method を改良して日本人の心筋梗塞患者と対照群の WGS データに適用し、レアバリアントの網羅的解析を行った。

＜方法＞

心筋梗塞患者 1965 人と対照群 3984 人の血液から DNA を抽出し、Hiseq X を用いてカバレッジが 15x と 30x をターゲットに WGS を行った。GATK ソフトウェアを用いて変異を検出し、PLINK ソフトウェアを用いてバリアントレベルとサンプルレベルのクオリティコントロール（Quality Control; QC）を行った。QC により信頼度の低いバリアントやサンプル、人種の異なるサンプルや外れ値を除去し、15x のデータ 4702 人（心筋梗塞 1953 人、対照群 2949 人、総バリアント数 51,358,278 個）、30x のデータ 1024 人（心筋梗塞 200 人、対照群 824 人、総バリアント数 25,585,121 個）を得た。このデータに HEAL method を改良した手法を応用した。まず、ANNOVAR ソフトウェアを用いてバリアントにアノテーションをつけ、非同義置換変異（missense 変異、stopgain/stoploss 変異、

splicing 変異、Frameshift 変異）を抽出し、1000 人ゲノムプロジェクトの東アジア人コホートで報告がなく、gnomAD データベースの東アジア人データでのアレル頻度（Allele frequency; AF）が 0.01 以下、データ内での AF が 0.01 以下のレアバリアントを抽出した。これらのレアバリアントにバリアントの病原性を予測する REVEL スコアのアノテーションを付与し、サンプルごとにバリアントのスコアを遺伝子ごとに足し合わせることで、各サンプルにおける遺伝子ごとの mutation burden の推定値の行列を得た。機械学習を用いた学習では、15x のデータ 4702 人を訓練データとして、mutation burden のみから臨床情報を一切用いず、心筋梗塞発症の有無を推定するモデルを学習させた。機械学習モデルには L1 正則化を用いたロジスティック回帰モデルを用い、10×クロスバリデーション法を用いてハイパーパラメータを最適化したうえでモデルを学習させた。これにより機械学習モデルにおける各遺伝子の重み付けが最適化される。この遺伝子ごとの重み付けから心筋梗塞に関連する遺伝子を絞り込むことができると同時に、レアバリアントをベースとした心筋梗塞の発症予測リスクスコア（Rare variant score; RVS）を計算することができる。カバレッジ 30x の 1024 人のデータをテストデータとして、RVS による心筋梗塞の予測性能や臨床情報との関連を検討した。また GWAS のデータから作成される、コモンバリアントから心筋梗塞の発症を予測する遺伝リスクスコア（Polygenic risk score; PRS）と統合して予測性能を検討した。

＜結果＞

学習した機械学習モデルでは 73 個の遺伝子に対するが心筋梗塞と関連していた。学習した機械学習モデルにおいて最も関連度が高い遺伝子は家族性高コレステロール血症の原因遺伝子である *LDLR* 遺伝子であり、73 個の遺伝子のうち 36%（26 個）が過去の GWAS で動脈硬化と関連する表現型と有意に相関していると報告されている遺伝子であった。過去の GWAS で動脈硬化関連の表現型との相関が示唆される以上のレベルの遺伝子とノックアウトマウスにおいて動脈硬化関連の表現型が有意に観察される遺伝子を合わせると 44 個（60%）がこの中に含まれていた。この新たなフレームワークにおいて機械学習モデルは信頼度の高い遺伝子を抽出していることが推定された。テストデータにおいて RVS を計算しその予測性能を検討すると心筋梗塞を Area under the curve score（AUC）0.58（p = 0.0007）の精度で予測することができた。また RVS と患者の臨床情報の関連を比較すると LDL コレステロール値や、総ビリルビン値、アスパラギン酸トランスアミナーゼ、総コレステロール値と有意に相関していた。RVS と心血管予後について Kaplan-Meier 法で検討したところ、RVS が高い患者は有意に心血管死亡の頻度が有意に高く（p = 0.004, log-rank test）、心筋梗塞患者 200 人に絞っても同様の結果（p = 0.034, log-rank test）であった。つまり RVS は心筋梗塞を予測するだけでなく、心筋梗塞の予後にも関連していることが示唆された。また、RVS と PRS には有意な相関は認めなかったが、心筋梗塞患者においては RVS と PRS は負の相関の傾向が見られ、RVS は PRS とは別の観点から心筋梗塞を予測していることが示唆された。最後に RVS と PRS を統合した combined スコ

アを作成した結果、combined スコアによる心筋梗塞予測性能は AUC 0.662 であり、PRS からの予測スコア 0.61 よりも有意に予測性能が向上した（p = 0.005, Delong test）。

＜考察＞

機械学習を用いたフレームワークを用いて心筋梗塞におけるレアバリアントを網羅的に解析した。心筋梗塞に関連する遺伝子群を絞り込み、レアバリアントをベースとした心筋梗塞予測スコアである RVS を作成した。同定した遺伝子群には、家族性高コレステロール血症の原因遺伝子である *LDLR* を含まれ、過去に動脈硬化に関連する表現型との関連が報告されている遺伝子が多く含まれていることから、この新しいフレームワークの妥当性が示唆された。またレアバリアントをベースとした RVS はコモンバリアントをベースとする PRS との相関がないことは、PRS と RVS は遺伝情報の別の部分を見ていることを示唆し、PRS と RVS の統合により心筋梗塞予測性能が向上したことは、レアバリアントが missing heritability の一部を説明できるということを示唆するものである。今後は同定した遺伝子の機能解析、イントロン領域などのノンコーディング領域に存在するレアバリアントを含めた解析方法の確立や、このフレームワークで同定した遺伝子や RVS が日本人以外の人種の心筋梗塞にも同様に当てはまるかについてのさらなる検討が求められる。本研究の結果は、心筋梗塞発症における遺伝的背景においてレアバリアントの重要性を示唆するものである。