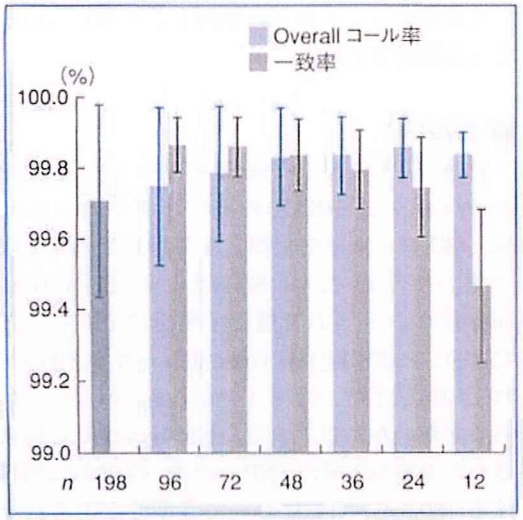


図 3　Birdseed アルゴリズムによる遺伝子型決定の精度

日本人健常者 198 検体のなかからランダムに 12 検体を選択し，その 12 検体を含む 6 つの異なるサンプルサイズ（12，24，36，48，72，96 検体）で決定した遺伝子型を，198 検体で決定した際の遺伝子型と比較した。各サンプルサイズで遺伝子型を決定した際の，その 12 検体の Overall コール率および遺伝子型一致率（concordance）を示す。

生じる偽陽性関連は解析を進めるうえで大きな障害となる。そこで，SNP Array 6.0 に搭載された全 90 万種の SNP について，できるだけ多くの SNP の遺伝子型を正確に決定する必要がある。遺伝子型決定に用いる Birdseed アルゴリズムの特性を知るために，日本人健常者 198 検体のなかからランダムに 12 検体を選択し，その 12 検体を含む 6 つの異なるサンプルサイズ（12，24，36，48，72，96 検体）で決定した遺伝子型を，198 検体で決定した際の遺伝子型と比較した[10]。

その結果，12 検体だけで遺伝子型を決定した際の Overall コール率は平均 99.84%〔信頼区間（CI）99.62～99.92〕となり，検体数が増えていくと Overall コール率は下がり，198 検体で決定した際には平均 99.71%（CI 98.07～99.89）となった（図 3）。一方，12 検体だけで決定した遺伝子型と 198 検体で決定した遺伝子型を比較した際の一致率（concordance）は，平均 99.47%（CI 98.37～99.67）と最も低く，サンプルサイズが大きくなるにつれて一致率は上昇し，96 検体で遺伝子型を決定した際の一致率は平均 99.87%（CI 99.40～99.92）となった（図 3）。この結果から，Birdseed アルゴリズムはサンプルサイズが小さくても遺伝子型を決定できるものの，高いタイピング精度を得るためには多くのサンプルを用いて遺伝子型を決定する必要があることがわかった。

Birdseed アルゴリズムによる遺伝子型決定において，サンプルサイズが大きくなるにつれて Overall コール率が低くなるという現象の原因として，QC コール率が 86%以上という閾値では十分に不良データを取り除けていないということが考えられる。日本人健常者 200 検体のタイピング結果から，QC コール率と Overall コール率との間には強い相関があることがわかったため，QC コール率の閾値をより厳しくして不良データの除去を行った。QC コール率の閾値を 95%とすると 188 検体が閾値を上回り，Overall コール率は平均 99.65%（CI 95.66～99.92）となった。しかし，厳しい閾値をパスした検体のなかに，Overall コール率がより悪くなるものがみられた。それらをさらに除外し，最終的にすべての検体で Overall コール率が上昇するまで不良データの除去を繰り返したところ，184 検体が残り，Overall コール率は平均 99.71%（CI 98.87～99.92）となった（図 2c）。

● 日本人を対象とした SNP Array 6.0 による GWAS の有用性

日本人健常者 200 検体の SNP タイピングの結果から，SNP Array 6.0 に搭載された全 909,622 SNPs のうち，約 20% に相当する 180,859 SNPs において多型性がみられなかった。また，遺伝子型が決定された SNP のなかにはタイピング精度の悪い SNP が一部含まれており，それらの SNP は偽陽性関連の原因のひとつになると考えられる。これについては，MAF，ハーディー・ワインバーグ平衡（Hardy-Weinberg equilibrium：HWE）および SNP コール率（各 SNP について，タイピングした全検体のうち遺伝子型を決定できた検体の割合）を指標と

して，タイピング精度の悪いSNPの大部分を排除することができる[11]。

われわれの解析では，MAF＞5%，HWE p 値＞0.001，SNPコール率＞95%を満たすSNPは，590,248 SNPsとなり，また，MAF＞1%，HWE p 値＞0.001，SNPコール率＞95%を満たすSNPは661,559 SNPsとなった。SNP Array 6.0で統計解析が可能であることがわかった約59万種のSNPについて，ゲノムカバー率を算出すると約75%となり，50万種のSNPを搭載したMapping 500K Arrayでのゲノムカバー率66%を上回ることが明らかになった[12]。

● SNPタイピングデータのデポジット

文部科学省の「統合データベースプロジェクト」において，われわれはSNPタイピングデータの半永続的な集約管理と研究者間の情報共有をめざして，日本人健常者のデータを登録した標準SNPデータベース，日本人健常者のコピー数多型（CNV）を登録したCNVデータベース，およびゲノムワイド関連解析のデータベース（GWAS-DB）を構築している[13]。

GWAS-DBは，研究概要，品質基準などの情報とともに，遺伝子型頻度やアリル頻度，および遺伝統計解析の結果を登録している。また，GWAS-DBはSNPだけでなくマイクロサテライトやCNVの疾患関連解析の結果も登録・閲覧することができ，エクソン情報やCNVなどの情報と遺伝統計解析の結果を重ね合わせて可視化する機能を備えている。疾患関連SNPの候補を多面的に選択できるよう，複数の機関が産出した同一疾患のデータ，および異なるプラットフォームの解析結果を比較したり，メタ解析を行ったりする機能を搭載し，専門家以外にも利用しやすいデータベースの構築をめざしている。

本研究でタイピングした日本人健常者200検体のSNP情報は，標準SNPデータベースに登録され，遺伝子型頻度やアリル頻度といった頻度情報は公開されている。また，今回タイピングした日本人健常者200検体のデータは，さまざまな多因子疾患を対象としたGWASにおいて共通のコントロール集団として用いられることが期待される。

● おわりに

日本人健常者200検体を対象としてSNP Array 6.0によるSNPタイピングを行った結果，QCコール率を指標として不良データを取り除いたうえで，48検体以上を用いてBirdseedアルゴリズムで遺伝子型を決定することにより，99.5%以上のOverallコール率，99.8%以上の遺伝子型一致率（concordance）が得られることがわかった。また，SNP Array 6.0に搭載された909,622種のSNPのうち，約20%に相当する180,859 SNPsにおいて多型性がみられないことが明らかとなった。さらに，SNPコール率，MAF，HWEなどを指標としてタイピング不良SNPの除去をすると，計590,248 SNPsがSNPコール率＞95%，MAF＞5%，HWE p 値＞0.001の条件を満たすことがわかった。この約59万種のSNPによりヒトゲノムの約75%をカバーできることから，日本人においてもSNP Array 6.0を用いたゲノムワイド関連解析が有用であることが期待される。

また，われわれはゲノムワイド関連分析によって検出された候補領域において，第一義的な疾患感受性遺伝子多型の特定に適する技術としてDigiTag2法を確立した[14,15]。現在，いくつかの多因子疾患を対象とした多施設共同研究グループと協力してゲノムワイド関連分析を進めており，さまざまな集団に共通する遺伝因子だけでなく，日本人あるいはアジア人に特徴的な遺伝因子の特定をめざしている。

文献

1) Affymetrix, Inc. [http://www.affymetrix.com/index.affx].
2) Ohnishi Y, Tanaka T, Ozaki K, et al. A high-throughput SNP typing system for genome-wide association studies. J Hum Genet 2001；46：471-7.
3) Ozaki K, Ohnishi Y, Iida A, et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 2002；32：650-4.
4) 文部科学省リーディングプロジェクト「オーダーメイド医療実現化プロジェクト」[http://www.biobankjp.

org/].
5) Unoki H, Takahashi A, Kawaguchi T, et al. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 2008 ; 40 : 1098-102.
6) Yasuda K, Miyake K, Horikawa Y, et al. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet 2008 ; 40 : 1092-7.
7) Miyagawa T, Kawashima M, Nishida N, et al. Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. Nat Genet 2008 ; 40 : 1324-8.
8) The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007 ; 447 : 661-78.
9) Cupples LA, Arruda HT, Benjamin EJ, et al. The Framingham Heart Study 100K SNP genome-wide association study resource : overview of 17 phenotype working group reports. BMC Med Genet 2007 ; 8 : s1.
10) Nishida N, Koike A, Tajima A, et al. Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals. BMC Genomics 2008 ; 9 : 431.
11) Miyagawa T, Nishida N, Ohashi J, et al. Appropriate data cleaning methods for genome-wide association study. J Hum Genet 2008 ; 53 : 886-93.
12) Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat Genet 2006 ; 38 : 659-62.
13) 文部科学省「統合データベースプロジェクト」[https://gwas.lifesciencedb.jp/].
14) Nishida N, Tanabe T, Hashido K, et al. DigiTag assay for multiplex single nucleotide polymorphism typing with high success rate. Anal Biochem 2005 ; 346 : 281-8.
15) Nishida N, Tanabe T, Takasu M, et al. Further development of multiplex single nucleotide polymorphism typing method, the DigiTag2 assay. Anal Biochem 2007 ; 364 : 78-85.

トランスレーショナルリサーチを支援する
遺伝子医学 MOOK 14

次世代創薬テクノロジー
実践:インシリコ創薬の最前線

別　刷

株式会社 メディカルドゥ

5．テーラーメイド医療をめざした疾患感受性遺伝子のゲノムワイド探索

西田奈央・徳永勝士

ヒトゲノム計画をはじめとするゲノム情報解析の成果として，公共のデータベースに蓄積された1100万種類を超える単一塩基多型（SNP）のうち，数十万～百万種類のSNPを同時にタイピングすることのできる手法が近年になって実用化された。われわれは，最新のプラットフォームを用いてSNPタイピングを効率的に行うためのシステムを構築し，いくつかの多因子疾患を対象としてゲノムワイド関連解析を行っている。本稿では，日本人における最新のプラットフォームの有用性を評価した結果を報告し，最後に将来の展望についても触れたい。

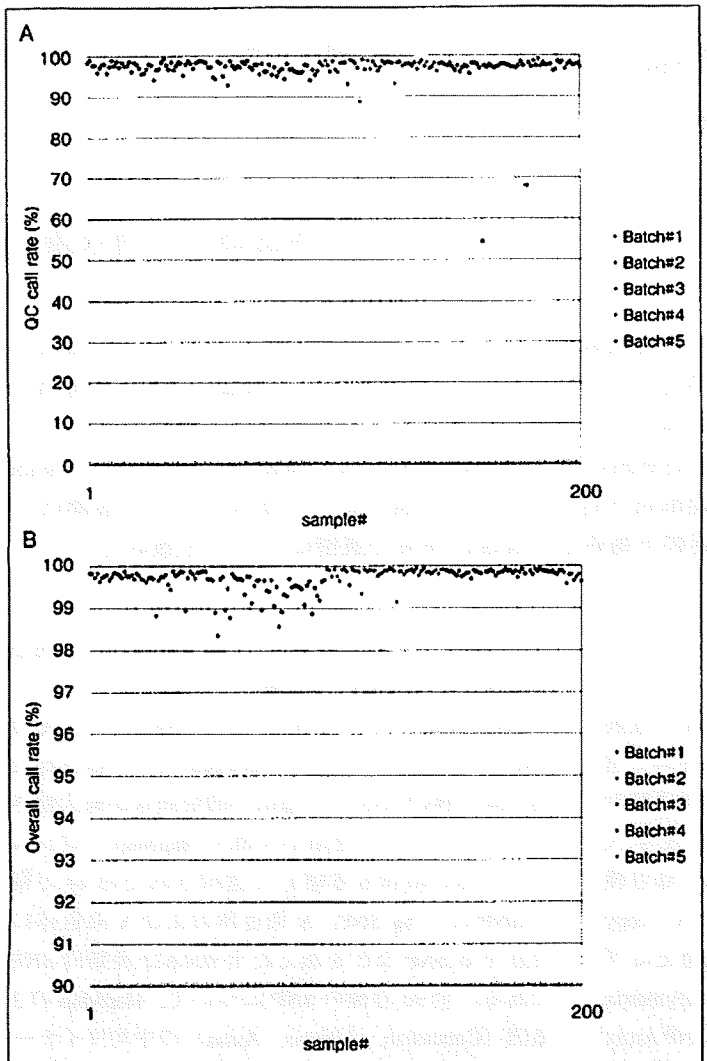
はじめに

単一塩基多型（SNP）タイピング技術の進展に伴って，ヒトの様々な多因子疾患にかかわる遺伝子を探索する戦略としてゲノムワイド関連解析（genome-wide association study：GWAS）が近年大きな注目を浴びている。2007年5月には，90万種を超えるSNP解析用プローブおよびCNV（copy number variation）解析用の94万種類を超えるプローブを搭載したキットが市販された（Affymetrix Genome-Wide Human SNP Array 6.0：以下SNP Array 6.0）[1]。関連分析の代表であるケースコントロール関連分析法は非血縁の患者群と健常対照群を対象として，疾患遺伝子と多型マーカーの連鎖不平衡（linkage disequilibrium）を検出する手法であり，これをゲノム全域にわたって適用するGWASでは数十万種類以上のSNPが必要となる。SNP Array 6.0はGWASに適したプラットフォームの1つとして考えられており，われわれはSNP Array 6.0プラットフォームを用いてSNPタイピングを効率的に行うためのシステムを構築し，いくつかの多因子疾患を対象としたGWASを実施している。

SNP Array 6.0に搭載されたSNPは，公共のSNPデータベースおよびPerlegen社に登録された約220万種のSNPから遺伝学的情報量が最大化されるように，また連鎖不平衡やHapMapプロジェクトからの情報も考慮して選択された約44万種のSNPに，Tag SNP，X染色体およびY染色体に存在するSNPなどを加えた全909,622種類のSNPである。全90万種のSNPについて，HapMapの3集団（Caucasian，African，Asian）の平均マイナーアリル頻度（MAF）は，それぞれ19.6%，20.6%，18.2%である。しかしながら，HapMapプロジェクトではわずか45検体の日本人しか解析していないため，MAFの低いSNPについては正確な頻度推定ができない。そこで，われわれはSNP Array 6.0を用いて日本人200検体のSNPタイピングを行い，日本人を対象としたGWASにおいて統計解析に用いることのできるSNP数を算出することを試みた。また，約50万種のSNPを搭載したMapping 500K

key words
SNP，ゲノムワイド関連解析，GWAS，ケースコントロール関連分析，多因子疾患，統合データベース，絞り込み，CNV

図❶　日本人健常者200検体のSNPタイピング結果



48検体を1バッチとしてタイピングを行った際のタイピング結果を示す。
A. 各検体のクオリティチェック（QC）用に用意された3022 SNPsのコール率（QC call rate）
B. 全909,622 SNPsのコール率（Overall call rate）
（グラビア頁参照）

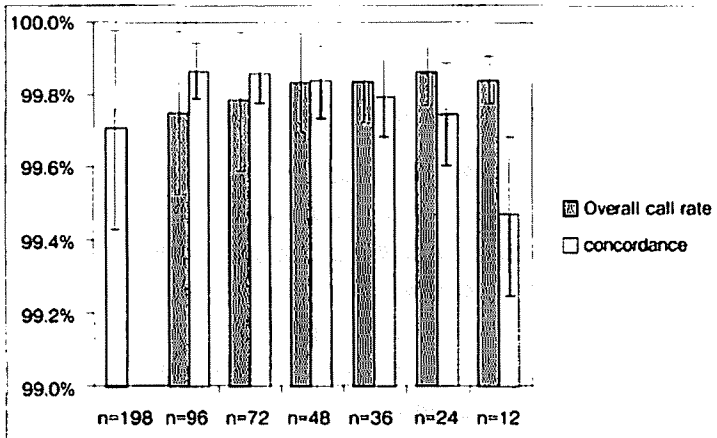
Arrayを用いたGWASの結果，有意な関連がみられた上位100種類のSNPのうちの45％がハーディー・ワインバーグ平衡からずれたという報告がある[21]。SNP Array 6.0は遺伝子型を決定するためにBirdseedアルゴリズムを用いるが，Birdseedアルゴリズムを用いた遺伝子型決定の精度を上げることが，GWASにおける擬陽性関連を効果的に排除することにつながる。そこで，日本人200検体のタイピング結果を用いて，Birdseedアルゴリズムによる正確な遺伝子型決定方法を検討した。

Ⅰ．ゲノムワイド関連研究の動向

GWASは日本の研究者によって先駆的に行われ，これまでにいくつかのヒト多因子疾患の感受性遺伝子を特定することに成功している[3,4]。また，日本では2003年からオーダーメイド医療実現基盤を構築することを目標とした「オーダーメイド医療実現化プロジェクト」が開始され，30万人の日本人を対象とした遺伝情報解析が行われている[5]。2008年には，日本における2大プロジェクトである「オーダーメイド医療実現化プロジェクト」と「ミレニアムゲノムプロジェクト」からそれぞれ独立に2型糖尿病に関連する遺伝子である*KCNQ1*を発見したという報告がなされた[6,7]。

近年のSNP解析技術の著しい進展によって，昨年来，欧米を中心として大規模なGWASの成果が次々に報告されている。例を挙げると，WTCCC（The Wellcome Trust Case Control Consortium）は7種類のcommon diseases（双極性感情障害，冠動脈疾患，クローン病，高血圧，関節リウマチ，1型糖尿病，2型糖尿病）について，それぞれ2000人の患者とコントロールとして健常者3000人の計17000人を対象とした大規模なGWASを行った[8]。また，大規模な疫学研究として知られるFramingham Heart Studyで収集された試料のうち9000検体について，心，肺，血液，睡眠疾患に関与する遺伝子変異を探索する計画が発表され，ゲノムワイド関連解析およ

図❷ Birdseed アルゴリズムによる遺伝子型決定の精度



各サンプルサイズで遺伝子型を決定した際の，その12検体のOverallコール率および遺伝子型一致率（concordance）を示す。

びゲノムワイド連鎖解析などを行った結果が2007年にまとめて報告された[9]。

Ⅱ. SNP Array 6.0を用いたゲノムワイドSNPタイピング

1. 日本人健常者200検体を用いたSNPタイピングの結果

SNP Array 6.0によるSNPタイピングでは，1検体につき500 ngのゲノムDNAを使用する。StyI制限酵素およびNspI制限酵素を用いたゲノムDNA断片化反応において，ゲノムDNA量を250 ngとなるように調整することがSNPタイピングの精度に大きな影響を与えることがこれまでの実験から明らかとなっている[10]。日本人健常者200検体のうち195検体のゲノムDNA濃度は規定濃度（50 ng/μL）を満たしており，平均54.8 ng/μL（45.0～57.8）であったが，残る5検体は規定濃度を下回り平均41.1 ng/μL（38.2～44.5）であった。そこで，規定濃度を下回った5検体は制限酵素によるゲノムDNA断片化反応に6 μLを持ち込み，ゲノムDNAの総量が約250 ngとなるように調整してタイピングを行った。日本人健常者200検体のSNPタイピングを行った結果，QCコール率は平均97.37%となった（図❶A）。ここでのQCコール率とは，タイピングデータのクオリティを評価するために用いられる指標で，SNP Array 6.0に搭載された3,022 SNPsについてのコール率を示す。これら3,022種のSNPの遺伝子型はDMアルゴリズム（confidence score=0.17）で決定され，コール率が86%以下となった検体は解析対象から除外する。日本人健常者200検体のタイピング結果から，QCコール率が86%を下回った2検体を除外し，86%を上回った198検体についてBirdseedアルゴリズムを用いて全909,622 SNPsの遺伝子型を決定したところ，Overallコール率は平均99.58%（96.42～99.90）となった（図❶B）。

2. Birdseedアルゴリズムによる正確な遺伝子型決定方法

膨大なSNPデータを取り扱うGWASにおいて，タイピングエラーが原因で生じる擬陽性関連は解析を進めるうえで大きな障害となる。そこで，SNP Array 6.0に搭載された全90万種のSNPについて，できるだけ多くのSNPの遺伝子型を正確に決定する必要がある。遺伝子型決定に用いるBirdseedアルゴリズムの特性を知るために，日本人健常者198検体の中からランダムに12検体を選択し，その12検体を含む6つの異なるサンプルサイズ（12，24，36，48，72，96検体）で決定した遺伝子型を，198検体で決定した際の遺伝子型と比較した[10]。その結果，Overallコール率は12検体だけで遺伝子型を決定した際には平均99.84%（99.62～99.92）となり，検体数が増えていくとOverallコール率は下がり，198検体で決定した際には平均99.71%（98.07～99.89）となった（図❷）。一方，12検体だけで決定した遺伝子型と198検体で決定した遺伝子型を比較した際の一致率（concordance）は平均99.47%（98.37～99.67）と最も低く，サンプルサイズが大きくなるにつれて一致率は上昇し，96検体で遺伝子型を決定した際の一致率は平均99.87%（99.40～99.92）となった。この結果から，Birdseedアルゴリズムはサンプルサイズが小さくても遺伝子型を決定できるものの，高いタイピング精度を得るため

には多くのサンプルを用いて遺伝子型を決定する必要があることがわかった。

Birdseedアルゴリズムによる遺伝子型決定において，サンプルサイズが大きくなるにつれてOverallコール率が低くなるという現象の原因として，QCコール率86%という閾値では十分に不良データを取り除けていないということが考えられる。日本人健常者200検体のタイピング結果からQCコール率とOverallコール率との間には強い相関があることがわかったため，QCコール率の閾値をより厳しくして不良データの除去を行った。QCコール率の閾値を95%とすると188検体が閾値を上回り，Overallコール率は平均99.65%（95.66～99.92）となった。しかし，厳しい閾値をパスした検体の中にOverallコール率がより悪くなるものがみられた。それらの不良データをさらに除外し，最終的にすべての検体でOverallコール率が上昇するまで不良データの除去を繰り返したところ，184検体が残り，Overallコール率は平均99.71%（98.87～99.92）となった。

3. 日本人を対象としたSNP Array 6.0によるGWASの有用性

日本人健常者200検体のSNPタイピングの結果から，SNP Array 6.0に搭載された全909,622 SNPsのうち約20%に相当する180,859 SNPsにおいて多型性がみられなかった。また，遺伝子型が決定されたSNPの中にはタイピング精度の悪いSNPが一部含まれており，それらのSNPは擬陽性関連の原因の1つになると考えられる。これについては，マイナーアリル頻度（MAF），ハーディー・ワインバーグ平衡（HWE）およびSNPコール率（各SNPについて，タイピングした全検体のうち遺伝子型を決定できた検体の割合）を指標として，タイピング精度の悪いSNPの大部分を排除することができる[11]。われわれの解析では，MAF > 5%，HWE p値 > 0.001，SNPコール率 > 95%を満たすSNPは590,248 SNPsとなり，またMAF > 1%，HWE p値 > 0.001，SNPコール率 > 95%を満たすSNPは661,559 SNPsとなった。SNP Array 6.0で統計解析が可能であることがわかった約59万種のSNPについて，ゲノムカバー率を算出すると約75%となり，50万種のSNPを搭載したMapping 500K Arrayでのゲノムカバー率66%を上回ることがわかった[12]。

Ⅲ. SNPタイピングデータのデポジット

文部科学省の「統合データベースプロジェクト」において，われわれはSNPタイピングデータの半永続的な集約管理と研究者間の情報共有をめざして，日本人健常者のデータを登録した標準SNPデータベース，日本人健常者のコピー数多型（CNV）を登録したCNVデータベース，およびゲノムワイド関連解析のデータベース（GWAS-DB）を構築している[13]。GWAS-DBは，研究概要，品質基準などの情報とともに，遺伝子型頻度やアリル頻度，および遺伝統計解析の結果を登録している。また，GWAS-DBはSNPだけでなくマイクロサテライトやCNVの疾患関連解析の結果も登録・閲覧することができ，エクソン情報やCNVなどの情報と遺伝統計解析の結果を重ね合わせて可視化する機能を備えている。疾患関連SNPの候補を多面的に選択できるよう，複数機関が産出した同一疾患のデータ，および異なるプラットフォームの解析結果の比較や，メタ解析を行ったりする機能を搭載し，専門家以外にも利用しやすいデータベースの構築をめざしている。

本研究でタイピングした日本人健常者200検体のSNP情報は，標準SNPデータベースに登録され，遺伝子型頻度やアリル頻度といった頻度情報は公開されている。また，今回タイピングした日本人健常者200検体のデータは，様々な多因子疾患を対象としたGWASにおいて共通のコントロール集団として用いられることが期待される。

おわりに

日本人健常者200検体を対象としてSNP Array 6.0によるSNPタイピングを行った結果，QCコール率を指標として不良データを取り除いたうえで，48検体以上を用いてBirdseedアルゴリズムで遺伝子型を決定することにより，99.5%以上のOverallコール率，99.8%以上の遺伝子型一致率が得られることがわかった。また，SNP Array 6.0に搭載さ

れた909,622種のSNPのうち，約20%に相当する180,859 SNPsにおいて多型性がみられないことが明らかとなった。さらに，SNPコール率，MAF，HWEなどを指標としてタイピング不良SNPの除去をすると，計590,248 SNPsがSNPコール率＞95%，MAF＞5%，HWE＞0.001の条件を満たすことがわかった。この約59万種のSNPによりヒトゲノムの約75%をカバーすることができることから，日本人においてもSNP Array 6.0を用いたGWASが有用であることが期待される。また，われわれはGWASによって検出された候補領域において，第一義的な疾患感受性遺伝子多型の特定（絞り込み）に適する技術としてDigiTag2法を確立した[14,15]。現在，いくつかの多因子疾患を対象とした多施設共同研究グループと協力してGWASを進めており，様々な集団に共通する遺伝因子だけでなく，日本人あるいはアジア人に特徴的な遺伝因子の特定をめざしている。

参考文献

1) http://www.affymetrix.com/index.affx
2) Hua J, et al : Bioinformatics 23, 57-63, 2007.
3) Ohnishi Y, et al : J Hum Genet 46, 471-477, 2001.
4) Ozaki K, et al : Nat Genet 32, 650-654, 2002.
5) http://www.biobankjp.org/
6) Unoki H, et al : Nat Genet, Advanced online publication, 2008.
7) Yasuda K, et al : Nat Genet, Advanced online publication, 2008.
8) The Wellcome Trust Case Control Consortium : Nature 447, 661-678, 2007.
9) Cupples LA, et al : BMC Med Genet 8, s1, 2007.
10) Nishida N, et al : BMC Genomics 9, 431, 2008.
11) Miyagawa T, et al : J Hum Genet 53, 886-893, 2008.
12) Barrett JC, Cardon LR : Nat Genet 386, 59-662, 2006.
13) https://gwas.lifesciencedb.jp/
14) Nishida N, et al : Anal Biochem 346, 281-288, 2005.
15) Nishida N, et al : Anal Biochem 364, 78-85, 2007.

参考ホームページ

- Affymetrix社
  http://www.affymetrix.com/index.affx
- 文部科学省リーディングプロジェクト「オーダーメイド医療実現化プロジェクト」
  http://www.biobankjp.org/
- 文部科学省「統合データベースプロジェクト」
  https://gwas.lifesciencedb.jp/

西田奈央

1998年　東京理科大学理工学部物理学科卒業
　　東京大学大学院総合文化研究科広域科学専攻修士課程入学
2000年　同博士課程進学
2003年　東京大学大学院医学系研究科人類遺伝学分野学術支援研究員
2004年　オリンパス株式会社入社
　　東京大学大学院医学系研究科人類遺伝学教室客員研究員
2007年　同人類遺伝学分野特任助教

特集 SNPが変える肝胆膵診療

総論

疾患関連遺伝子を探し出すためのSNP解析

西田奈央* 徳永勝士*

索引用語：単一塩基多型（SNP），SNPタイピング，ゲノムワイド関連解析，絞り込み，多因子疾患

1 はじめに

近年，ゲノムワイド関連分析法(Genome-wide association study，GWAS)により，ヒトのさまざまな多因子疾患について，その遺伝的な要因を探索する研究が日本をはじめとして世界中で行われている．このGWASは日本の研究者によって先駆的に行われ，これまでにいくつかのヒトの多因子疾患の感受性遺伝子を特定することに成功している[1~4]．また，2003年からオーダーメイド医療実現基盤を構築することを目標とした「オーダーメイド医療実現化プロジェクト」が開始され，30万人の日本人を対象とした疾患関連研究(大規模ケース・コントロール関連解析)が行われている[5]．2008年には，日本における2大プロジェクトである「オーダーメイド医療実現化プロジェクト」と「ミレニアムゲノムプロジェクト」からそれぞれ独立に2型糖尿病に関連する遺伝子であるKCNQ1を発見したという報告がなされた[6,7]．また，われわれの研究室においても，CPT1B遺伝子とCHKB遺伝子の間に存在するSNPが睡眠障害の一つであるナルコレプシーと関連していることを発見し，2008年に報告をした[8]．

ゲノムワイド関連分析法は，ゲノム全域に分布する数十万種以上のSNPについて，非血縁の患者集団と健常者集団を対象として疾患遺伝子と連鎖不平衡(linkage disequilibrium，LD)にある多型マーカーを検出する手法である[9]．ゲノムワイド関連分析によりさまざまな多因子疾患を対象とした疾患感受性遺伝子の探索が行われるようになった背景には，本稿で紹介する大規模なSNP解析技術の進展が非常に大きな役割を果たしている．従来の多くのSNPタイピング法は，個々の多型部位を含むゲノム断片を特異的にPCRで増幅した後でアリルを識別する方法であった[10~14]．これらの方法では，1,000種程度のSNPを対象としたタイピングであれば，PCRプライマーをはじめとする各種試薬にかかるコストを考えても実用可能であるといえる

Nao NISHIDA et al：Identification of susceptibility genes for multifactorial diseases by analyzing single nucleotide polymorphisms
*東京大学大学院医学系研究科人類遺伝学分野［〒113-0033 東京都文京区本郷7-3-1］

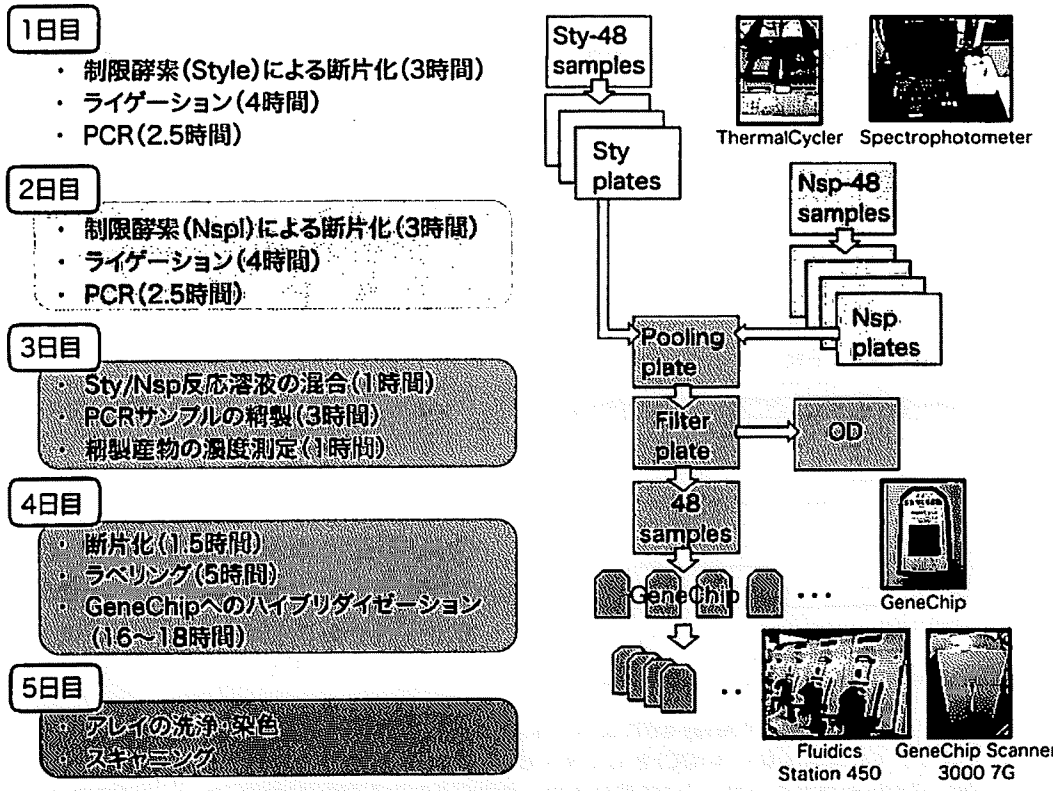


図1　SNP Array 6.0によるSNPタイピングの流れ
制限酵素(Sty I, Nsp I)による断片化反応からスキャニングまで全5日の工程でSNPタイピングが行われる．1検体につき500 ngのゲノムDNAを用いて全909,622種のSNPをタイピングすることができる．

が，数千から数万種を超える数のSNPをタイピングすることは困難となる．一方，近年になって多型部位特異的なPCRを行わずに大規模なSNPタイピングを行う方法が実用化された[15,16]．その一つであるAffymetrix社によって確立された方法では，まず制限酵素反応でゲノムDNAの断片化を行い，続いてそれら断片の両端にアダプター配列を付加し，まとめて増幅した後にマイクロアレイを用いたアリル特異的なハイブリダイゼーションを行う[15]．現在では，この手法を用いて90万種を超えるSNPを同時にタイピングするキットが市販されている(Affymetrix Genome-Wide Human SNP Array 6.0，以下，SNP Array 6.0)．

われわれは，ゲノムワイド関連分析で検出されたヒトの多因子疾患の疾患感受性候補領域の中から真の疾患感受性遺伝子を効率よく特定するためのSNP解析技術として，DigiTag2法を確立した[13]．DigiTag2法は，96カ所(もしくは32カ所)のSNPを同時に解析することができ，また，解析対象によらず同一のマイクロアレイを用いることができるため，専用マイクロアレイを準備する必要のない低コストのSNP解析技術である．われわれの教室に設置したヒトSNPタイピングセ

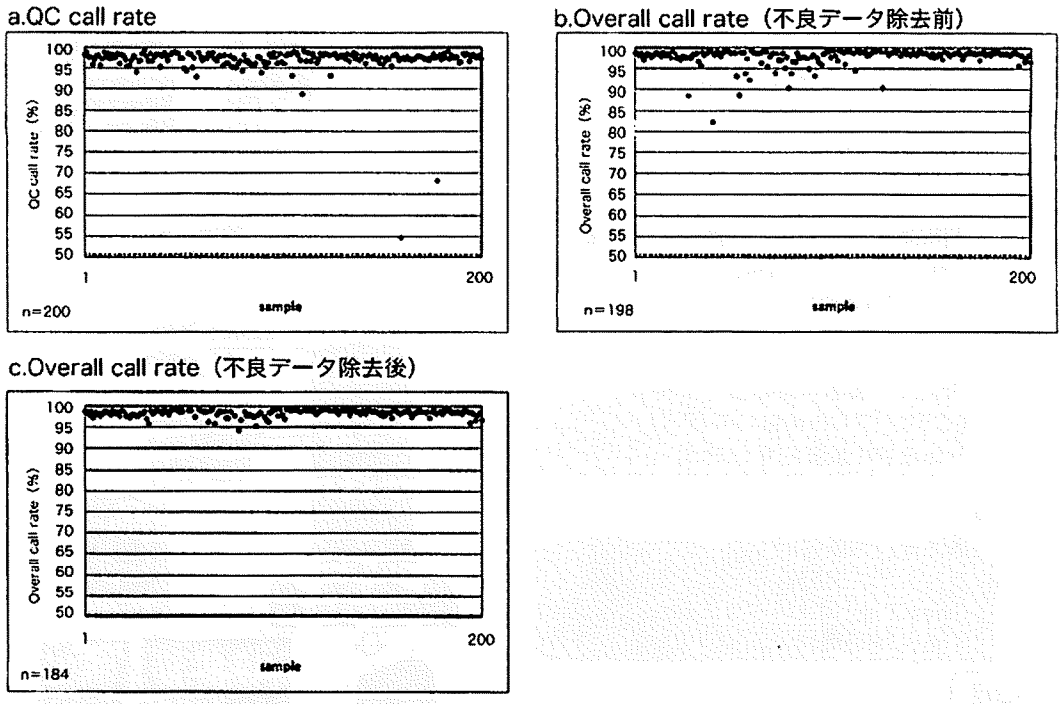


図2　SNP Array 6.0による日本人健常者200名のタイピング結果

a. クオリティーコントロール(QC)としてタイピングされた3,022SNPsのコール率を示す.
b. QC call rateが86%を上回った198検体を用いて決定された全909,622SNPsのコール率を示す.
c. QC call rateを指標として不良データを除去した後の184検体を用いて遺伝子型を決定した際の全909,622 SNPsのコール率を示す.

ンターでは，SNP Array 6.0によるSNPタイピングによりいくつかの多因子疾患についてゲノムワイド関連分析を実施し，さらにDigiTag2法を用いて疾患感受性遺伝子の特定を目的としたSNP解析を実施している.

2 ゲノムワイドSNP解析：SNP Array 6.0の技術原理

SNP Array 6.0によるSNPタイピングは，ゲノムの複雑さを低減しマイクロアレイへのハイブリダイゼーション効率を上げるための酵素反応ステップと，洗浄・染色装置(Fluidics Station 450)およびマイクロアレイ用スキャナー（GeneChip Scanner 3000 7G）を用いた検出ステップで構成される(図1)．1検体につき合計500 ngのゲノムDNAを使用して，2種類の制限酵素(Sty I，Nsp I)によるゲノムDNAの断片化を行った後，断片化したゲノムDNAの両末端にアダプター配列をライゲーション反応により付加する．アダプター配列は，続くPCRで使用されるプライマーと相同な配列を持ち，また制限酵素認識配列を突出端として持つ2本鎖DNAである．PCRでは，目的の長さを持ったDNA断片(250-1100 bp)だけが選択的に増幅される．続いて，Sty IおよびNsp Iそれぞれの PCR産物を混合した後，混合産物を精製し，DNase I制限酵素による断片化を行う．ここで，断

片化されたPCR産物は平均180 bp以下となる．最後にterminal deoxynucleotidyl transferase酵素反応により，断片化したPCR産物の末端にビオチンを導入する．

続いて，専用のマイクロアレイ(GeneChipアレイ)を用いてハイブリダイゼーションを行う．マイクロアレイに固定されるプローブは25塩基長のオリゴDNAで，SNP部位を含む塩基配列を持っている．2種類のアリルを正確に識別するために，SNP部位を25塩基長のプローブの中心に置いたプローブを基本として，SNP部位を中心から4塩基上流(+4)にずらしたプローブから4塩基下流(-4)にずらしたプローブまで7種類のプローブ(-4, -2, -1, 0, +1, +3, +4)を用意し，その中から最適な1種類のプローブを選択する．また，同一のプローブをマイクロアレイ上に3スポット用意することで，SNPタイピングデータの欠損を防ぐ工夫がなされている．

マイクロアレイへのハイブリダイゼーションが終了した後，洗浄・染色装置を用いてマイクロアレイの洗浄および蛍光染色を行う．蛍光染色は，蛍光分子で標識されたストレプトアビジンを，上述のビオチン導入されたPCR断片に結合することにより行われる．また，洗浄・染色装置内ではビオチン修飾された抗ストレプトアビジン抗体を用いてシグナルの増強が行われる．最後に蛍光染色されたマイクロアレイを専用のスキャナーで画像データとして読み取り，続いて専用のソフトウェア(Genotyping Console ver3.0 software)を用いて各SNPの遺伝子型を決定する．

複数の施設で行われたSNP Array 6.0によるSNP解析の結果から，Overall call rate（全909,622種のSNPのうち遺伝子型が決定されたSNPの割合）は平均99％以上となり，また，HapMapデータベースに登録されたタイピングデータとの遺伝子型一致率は99.7％を超えることがAffymetrix社から報告されている．また，タイピング結果が悪いことが明らかとなっている3,022種のSNPをクオリティーコントロール(QC)として用いて，QC call rate（3,022種のSNPのうち遺伝子型が決定されたSNPの割合）が86％を下回る検体を除外したうえで，全909,622種のSNPの遺伝子型は決定される．

3 ゲノムワイドSNP解析：日本人健常者200検体の解析結果

SNP Array 6.0によるSNPタイピングでは，制限酵素(Sty IおよびNsp I)による断片化反応に用いるゲノムDNA量がそれぞれ250 ngとなるように調整することがSNPタイピングの精度に大きな影響を与えることがこれまでの実験結果から明らかとなっている[17]．われわれが行った日本人健常者200検体を対象としたSNP解析を例にあげると，200検体のうち195検体のゲノムDNA濃度は規定濃度である50 ng/μlを満たしており，平均54.8 ng/μlであったが，5検体は規定濃度を下回り平均41.1 ng/μlであった．そこで，規定濃度を下回った5検体は制限酵素断片化反応に6 μlを持ち込み，ゲノムDNAの総量が約250 ngとなるように調整してタイピングを行った．日本人健常者200検体のタイピング結果から，QC call rateは平均97.37％となり，また，QC call rateが86％を下回る検体は200検体のうち2検体となった(図2a)．続いて，QC call rateが86％を上回った198検体を用いて全909,622SNPsのコール率(Overall call rate)を決定したところ，平均99.58％となった(図2b)．

膨大なSNPデータを取り扱うゲノムワイド関連分析において，タイピングエラーが原

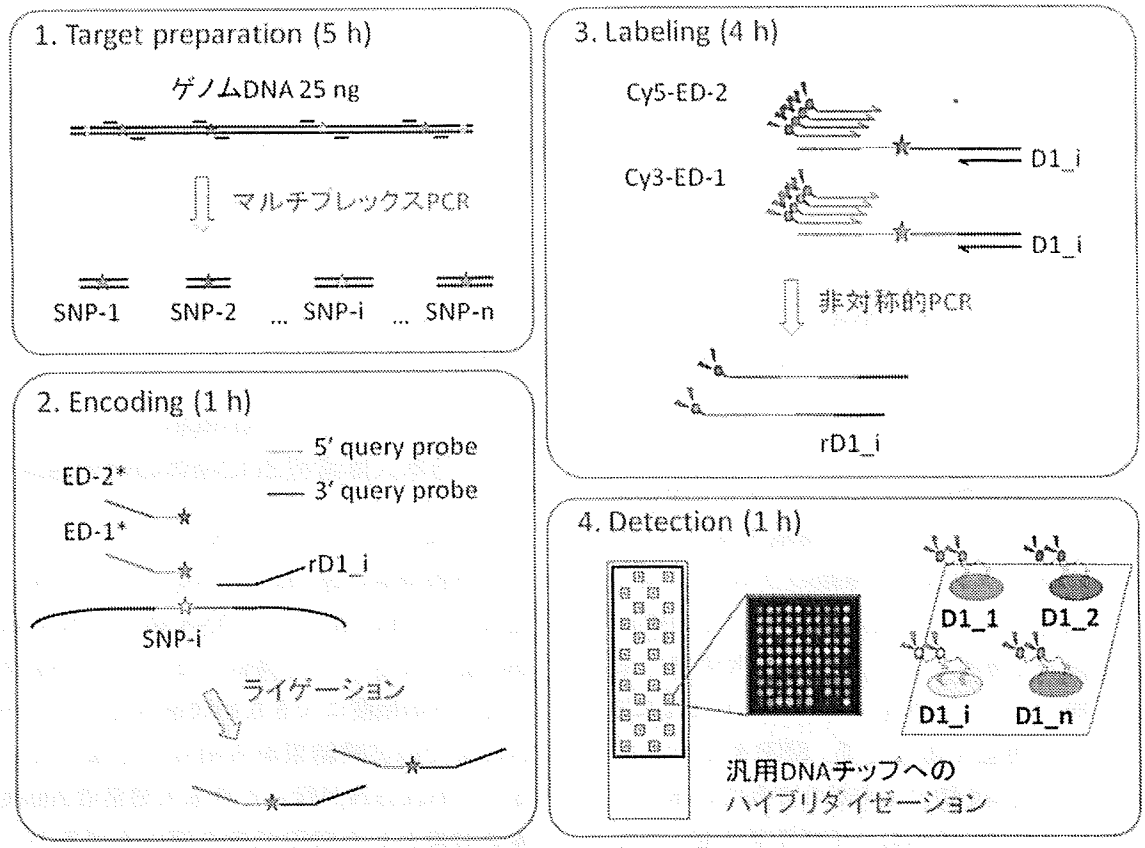


図3　DigiTag2法の概要

DigiTag2法は，ターゲット分子調製，エンコード，ラベリング，検出の4つの工程で構成される．5′ query probeにはアリルに対応して2種類のED（ED-1，ED-2）を付加し，また3′ query probeにはSNPに応じてD1（D1_i）を付加する．EDおよびD1は物理的，化学的性質が一様となるように設計した23塩基長のオリゴDNAである．配列が相補鎖である場合には配列名称の前に"r"を付けた．

因で生じる偽陽性関連は解析を進める上で大きな障害となる．そこで，SNP Array 6.0に搭載された全90万種のSNPについて，できるだけ多くのSNPの遺伝子型を正確に決定する必要がある．Affymetrix社が提供する遺伝子型決定ソフトウェアは，Birdseedアルゴリズムにより遺伝子型の決定を行う．このBirdseedアルゴリズムはサンプルサイズが小さくても遺伝子型を決定できるものの，高いタイピング精度を得るためには48検体以上を用いて遺伝子型を決定する必要があることが，これまでの解析結果から明らかとなっている[17]．また，QC call rate > 86％ではタイピング不良データを十分に取り除くことができず，タイピング不良データが混在することでOverall call rateを低下させることが明らかとなったため，われわれはQC call rate > 95％を閾値としてタイピング不良データの除去を行うこととした(図2c)．

遺伝子型が決定されたSNPの中にはタイピング精度の悪いSNPが一部含まれており，それらのSNPは偽陽性関連の原因のひとつになると考えられる．これについては，マイナーアリル頻度(MAF)，ハーディー・ワイ

ンバーグ平衡検定(Hardy-Weinberg equilibrium test，HWE)およびSNP call rate（各SNPについて，タイピングした全検体のうち遺伝子型を決定できた検体の割合)を指標として，タイピング精度の悪いSNPの大部分を排除することができる[18]．われわれの解析では，MAF > 5%，HWE p値 > 0.001，SNP call rate > 95%を満たすSNPは，590,248 SNPsとなり，この約59万種のSNPによりヒトゲノムの約75%をカバーできることから，日本人においてもSNP Array 6.0を用いたゲノムワイド関連解析が有用であることが期待される．

4 疾患感受性遺伝子特定のためのSNP解析：DigiTag2法の原理と特徴

DigiTag2法は，SNPの遺伝子型をオリゴDNAタグに変換してマルチプレックスSNPタイピング(96-plexもしくは32-plex)を行う(図3)[13]．オリゴDNAタグ(図3中，EDおよびD1)は物理的，化学的に性質が一様となるように設計した23塩基長のオリゴDNAで，オリゴDNAタグを使用することにより正確なDNA分子反応を行うことが可能となる．SNPタイピングで使用するプライマー/プローブは共通の設計水準でデザインするため，解析対象に依存しない共通の実験条件でのSNPタイピングが可能である．また，実験条件の検討が不要であることに加えて，オリゴDNAタグは解析対象となるSNPに対して自由に割り当てることができるため，結果表示に用いるDNAチップを汎用的に使用できるという特徴を持っている．

DigiTag2法はランニングコストが安いうえに，複数カ所のSNPをまとめて解析することができるため，複数のSNPを多検体で解析するのに適した技術である．本技術が他の解析技術と比較して特に優位性が高い点として，タイピング成功率(遺伝子型が決定できたSNP数/解析対象としたSNP総数)が90.72%（929/1,024SNPs)と非常に高いことが挙げられる．また，これまでにDigiTag2法を用いて，合計26,665検体以上を対象として1,100カ所以上のSNP解析を行った実績があり，DigiTag2法は高い成功率でSNPタイピングを実施できるだけでなく，96-plexまたは32-plexのいずれでも，非常に高いCall rate（平均99.53%)でSNPタイピングを行うことのできる技術である．

DigiTag2法は高い成功率でSNPタイピングを行えることから，ゲノムワイド連鎖分析あるいはゲノムワイド関連分析によって検出された候補領域における絞り込み解析を効率的に行う技術として利用されることが期待される．

5 GWASデータのデポジット

文部科学省の「統合データベースプロジェクト」において，われわれはSNPタイピングデータの半永続的な集約管理と研究者間の情報共有を目指して，日本人健常者のデータを登録した標準SNPデータベース，日本人健常者のコピー数多型(CNV)を登録したCNVデータベースおよびゲノムワイド関連解析のデータベース(GWAS-DB)を構築し，なるべく多くの研究グループのデータ登録を広くお願いしている[19]．GWAS-DBは，研究概要，品質基準などの情報と共に，遺伝子型頻度やアリル頻度および遺伝統計解析の結果を登録している．また，GWAS-DBはSNPだけでなくマイクロサテライトやCNVの疾患関連解析の結果も登録・閲覧することができ，エクソン情報やCNVなどの情報と遺伝統計解析

の結果を重ね合わせて可視化する機能を備えている．疾患関連SNPの候補を多面的に選択できるよう，複数の機関が産出した同一疾患のデータおよび異なるプラットフォームの解析結果を比較したり，メタ解析を行ったりする機能を搭載し，専門家以外にも利用しやすいデータベースの構築を目指している．

6 結 語

現在販売されているSNP Array 6.0は，欧米人で頻度の高いSNPが優先的に搭載されているため，日本人試料では約20％のSNPについて多型性が見られなかった．疾患感受性候補領域を最大限に検出するためにも，今後，アジア系集団に適したSNPセットを搭載したプラットフォームが作製されることを期待したい．また，SNP Array 6.0にはCNVを検出するためのプローブが搭載されているものの，CNVを解析するためのソフトウェアの解析精度にはまだ多くの問題が残っており，今後のCNV解析精度の向上が強く望まれる．いずれにせよ，ゲノムワイド多型解析情報は従来にない膨大なデータを産生することから，バイオインフォマティクスに関わる様々な研究者にとって挑戦に値する多くの課題を提供してくれるとともに，その達成によって従来にない実り豊かな成果をわれわれにもたらしてくれるに違いない．

文 献

1) Ozaki K, Ohnishi Y, Iida A et al : Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32 : 650-654, 2002
2) Tamiya G, Shinya M, Imanishi T et al : Whole genome association study of rheumatoid arthritis using 27039 microsatellites. Hum Mol Genet 14 : 2305-2321, 2005
3) The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 : 661-678, 2007
4) Cupples LA, Arruda HT, Benjamin EJ et al : The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. BMC Med Genet 8 : s1, 2007
5) 文部科学省リーディングプロジェクト「オーダーメイド医療実現化プロジェクト」[http://www.biobankjp.org/]
6) Unoki H, Takahashi A, Kawaguchi T et al : SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 40: 1098-1102, 2008
7) Yasuda K, Miyake K, Horikawa Y, et al : Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet 40: 1092-1097, 2008
8) Miyagawa T, Kawashima M, Nishida N et al : Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. Nat Genet 40 : 1324-1328, 2008
9) Ohashi J, Tokunaga K : The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. J Hum Genet 46 : 478-482, 2001
10) Holland PM, Abramson RD, Watson R et al : Detection of specific polymerase chain reaction product by utilizing the 5' → 3' exonuclease activity of Thermus aquaticus DNA polymerase. Proc Natl Acad Sci USA 88 : 7276-7280, 1991
11) Pastinen T, Kurg A, Metspalu A et al : Minisequencing: A specific tool for DNA analysis and diagnostics on oligonucleotide arrays. Genome Res 7: 606-614, 1997
12) Bannai M, Higuchi K, Akesaka T et al : Single-nucleotide-polymorphism genotyping for whole-genome-amplified samples using automated fluorescence correlation spectroscopy. Anal Biochem 327: 215-221, 2004
13) Nishida N, Tanabe T, Takasu M et al : Further development of multiplex single nucleotide polymorphism typing method, the DigiTag2 assay. Anal Biochem 364: 78-85, 2007
14) Krjutškov K, Andreson R, Mägi R et al : Development of a single tube 640-plex genotyping method for detection of nucleic acid variations on mi-

croarrays. Nucleic Acids Res 36 : e75, 2008
15) Affymetrix, Inc. [http://www.affymetrix.com/index.affx].
16) Illumina, Inc. [http://www.illumina.com/].
17) Nishida N, Koike A, Tajima A et al : Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals. BMC Genomics 9 : 431, 2008
18) Miyagawa T, Nishida N, Ohashi J et al : Appropriate data cleaning methods for genome-wide association study. J Hum Genet 53 : 886–893, 2008
19) 文部科学省「統合データベースプロジェクト」[https://gwas.lifesciencedb.jp/].

* * *

Hepatology Research 2009 doi: 10.1111/j.1872-034X.2009.00607.x

Original Article

A predictive model of response to peginterferon ribavirin in chronic hepatitis C using classification and regression tree analysis

Masayuki Kurosaki,[1] Kotaro Matsunaga,[2] Itsuko Hirayama,[1] Tomohiro Tanaka,[1] Mitsuaki Sato,[1] Yutaka Yasui,[1] Nobuharu Tamaki,[1] Takanori Hosokawa,[1] Ken Ueda,[1] Kaoru Tsuchiya,[1] Hiroyuki Nakanishi,[1] Hiroki Ikeda,[1] Jun Itakura,[1] Yuka Takahashi,[1] Yasuhiro Asahina,[1] Megumu Higaki,[4] Nobuyuki Enomoto[3] and Namiki Izumi[1]

[1]*Division of Gastroenterology and Hepatology and* [2]*Division of Pathology, Musashino Red Cross Hospital, Tokyo,* [3]*First Department of Internal Medicine, University of Yamanashi, Yamanashi, and* [4]*Department of Medical Science, Jikei Medical University, Tokyo, Japan*

Aim: Early disappearance of serum hepatitis C virus (HCV) RNA is the prerequisite for achieving sustained virological response (SVR) in peg-interferon (PEG-IFN) plus ribavirin (RBV) therapy for chronic hepatitis C. This study aimed to develop a decision tree model for the pre-treatment prediction of response.

Methods: Genotype 1b chronic hepatitis C treated with PEG-IFN alpha-2b and RBV were studied. Predictive factors of rapid or complete early virological response (RVR/cEVR) were explored in 400 consecutive patients using a recursive partitioning analysis, referred to as classification and regression tree (CART) and validated.

Results: CART analysis identified hepatic steatosis (<30%) as the first predictor of response followed by low-density-lipoprotein cholesterol (LDL-C) (≥100 mg/dL), age (<50 and <60 years), blood sugar (<120 mg/dL), and gamma-glutamyltransferase (GGT) (<40 IU/L) and built decision tree model. The model consisted of seven groups with variable response rates from low (15%) to high (77%). The reproducibility of the model was confirmed by the independent validation group ($r^2 = 0.987$). When reconstructed into three groups, the rate of RVR/cEVR was 16% for low probability group, 46% for intermediate probability group and 75% for high probability group.

Conclusions: A decision tree model that includes hepatic steatosis, LDL-C, age, blood sugar, and GGT may be useful for the prediction of response before PEG-IFN plus RBV therapy, and has the potential to support clinical decisions in selecting patients for therapy and may provide a rationale for treating metabolic factors to improve the efficacy of antiviral therapy.

Key words: data mining, decision tree, HCV, low-density-lipoprotein-cholesterol, steatosis

INTRODUCTION

COMBINATION THERAPY WITH pegylated interferon (PEG-IFN) and ribavirin (RBV) is now recognized as a standard treatment for patients with chronic hepatitis C.[1] However, the rate of sustained virological response (SVR) to 48 weeks of PEG-IFN RBV combination therapy is only 50% in patients with hepatitis C virus (HCV) genotype 1b and high HCV RNA titer, so called difficult to treat chronic hepatitis C patients.[2,3] Within this difficult to treat group, the response to treatment sometimes can be highly heterogeneous for cases which are apparently equivalent in HCV RNA titer, making the prediction of response before treatment a difficult task. It has been suggested that early virological response (EVR), defined as either undetectable HCV RNA or a 2 log drop in HCV RNA at week 12, is a reliable means to predict SVR.[2,4] More recently, it has been suggested that patients with a rapid virological response (RVR: undetectable HCV RNA at week 4) and a complete EVR (cEVR: undetectable HCV RNA at week 12)

Correspondence: Dr Namiki Izumi, Division of Gastroenterology and Hepatology, Musashino Red Cross Hospital, 1-26-1 Kyonan-cho, Musashino-shi, Tokyo 180-8610, Japan. Email: nizumi@musashino.jrc.or.jp

Received 26 May 2009; revision 25 August 2009; accepted 26 August 2009.

achieve high SVR rates, while patients with a partial EVR (pEVR: 2 log drop in HCV RNA but still detectable at week 12) have lower rates of SVR.[5] Since PEG-IFN RBV combination therapy is costly and accompanied by potential adverse effects, the ability to predict the possibility of RVR or cEVR before therapy and identifying curable patients may significantly influence the selection of patients for therapy. Moreover, identification of baseline predictors of poor response is particularly important to establish a rationale for identifying therapeutic targets to improve the efficacy of antiviral therapy.

Data mining is a method of predictive analysis which explores tremendous volumes of data to discover hidden patterns and relationships in highly complex datasets and enables the development of predictive models. The classification and regression tree (CART) analysis is a core component of the decision tree tool for data mining and predictive modeling,[6] is deployed to decision makers in various fields of business, and currently is being used in the area of biomedicine.[7-13] The results of CART analysis are presented as a decision tree, which is intuitive and facilitates the allocation of patients into subgroups by following the flow-chart form.[14] CART has been shown to be competitive with other traditional statistical techniques such as logistic regression analysis.[15]

In the present study, we used the CART analysis to explore baseline predictors of response to PEG-IFN plus RBV therapy among clinical, biochemical, virological and histological pretreatment variables and to define a pre-treatment algorithm to discriminate chronic hepatitis C patients who are likely to respond to PEG-IFN plus RBV therapy.

MATERIALS AND METHODS

Patients

A TOTAL OF 419 chronic hepatitis C patients were treated with PEG-IFN alpha-2b and RBV at Musashino Red Cross Hospital between December 2001 and December 2007. Among them, 400 patients who fulfilled the following inclusion criteria were enrolled in the present study. (i) infection by genotype 1b (ii) HCV RNA higher than 100 KIU/mL by quantitative PCR (Cobas Amplicor HCV Monitor, Roche Diagnostic systems, CA) which is usually used for the definition of high viral load in Japan (iii) lack of co-infection with hepatitis B virus or human immunodeficiency virus (iv) lack of other causes of liver disease such as autoimmune hepatitis, primary biliary cirrhosis, or alcohol intake of more than 20 g per day, and (v) having completed at least 12 weeks of therapy with an early virological response that could be evaluated. Patients received PEG-IFN alpha-2b (1.5 microgram/kg) subcutaneously every week and were administered a weight adjusted dose of RBV (600 mg for <60 kg, 800 mg for 60–80 kg, and 1000 mg for >80 kg) which is the recommended dosage in Japan. Data from two third of patients (269 patients) were used for the model building set and the remaining one third of patients (131 patients) were used as a validation set. Consent in writing was obtained from each patient and the study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the institutional review committee.

Laboratory tests

Blood samples were obtained before therapy, and at least once every month during therapy and analyzed for hematologic tests, blood chemistries, and HCV RNA. In the present study, RVR and cEVR was defined as undetectable HCV RNA by qualitative PCR with a lower detection limit of 50 IU/mL (Amplicor, Roche Diagnostic systems, CA) at week 4 and 12, respectively. SVR was defined as undetectable HCV RNA at week 24 after the completion of therapy.

Histological examination

For all patients, liver biopsy specimens were obtained before therapy and were evaluated independently by three pathologists who were blinded to the clinical details. If there was a disagreement, the scores assigned by the majority of pathologists were used for the analysis. Fibrosis and activity were scored according to the METAVIR scoring system.[16] Fibrosis was staged on a scale of 0–4: F0 (no fibrosis), F1 (mild fibrosis: portal fibrosis without septa), F2 (moderate fibrosis: few septa), F3 (severe fibrosis: numerous septa without cirrhosis) and F4 (cirrhosis). Activity of necroinflammation was graded on a scale of 0–3: A0 (no activity), A1 (mild activity), A2 (moderate activity) and A3 (severe activity). Percentage of steatosis was quantified by determining the average proportion of hepatocytes affected by steatosis and graded on a scale of 0–3: grade 0 (no steatosis), grade 1 (0–9%), grade 2 (10–29%), and grade 3 (over 30%) as we reported previously.[17]

Database for analysis

A pretreatment database of 72 variables was created containing histological findings (grade of fibrosis, activity, and steatosis), laboratory tests including the quantity of HCV RNA by Cobas Amplicor, and clinical information (age, gender, body weight, and body mass index).

The baseline characteristics and test results are listed in Table 1. The overall rate of RVR/cEVR was 43% in the model building set and 48% in the validation set. There were no significant differences in the clinical backgrounds between these two groups. Hepatitis C viral mutations, such as mutations in interferon-sensitivity determining region or core amino acid residues 70 and 91, were not included in the present analysis. The dataset of laboratory tests was based on the digitized records in this hospital. Continuous data was split into categorized data by increment of 10; For example, age was categorized into <30, 30–39, 40–49, 50–59, 60–69, and ≥70.

Statistical analysis

Based on this database, the recursive partitioning analysis algorithm referred to as CART was implemented to define meaningful subgroups of patients with respect to the possibility of achieving RVR/cEVR. The CART belongs to a family of nonparametric regression methods based on binary recursive partitioning of data. The software automatically explore the data to search for optimal split variables, builds a decision tree structure and finally classifies all subjects into particular subgroups that are homogeneous with respect to the outcome of interest.[18] During the CART analysis, first, the entire study population, and thereafter, all newly defined subgroups, were investigated at every step of the analysis to determine which variable at what cut-off point yielded the most significant division into two prognostic subgroups that were as homogeneous as possible with respect to estimates of RVR/cEVR possibilities. This algorithm uses the impurity function (Gini criterion function) for splitting.[19] A restriction was imposed on the tree construction such that terminal subgroups resulting from any given split must have at least 20 patients. The CART procedure stopped when either no additional significant variable was detected or when the sample size was below 20. The resulting final subgroups were most homogeneous with respect to the probability of achieving RVR/cEVR. For this analysis, data mining software Clementine version 12.0 (SPSS Inc, Chicago, IL) was utilized. SPSS 15.0 (SPSS Inc, Chicago, IL) was used for logistic regression analysis.

RESULTS

Factors associated with RVR/cEVR by standard statistical analysis

WE FIRST ANALYZED 72 variables by univariate and multivariate logistic regression analysis to find factors associated with RVR/cEVR (Table 2). Patients with RVR/cEVR were significantly younger than those without. Among histological findings, grade of steatosis and stage of fibrosis was significantly lower in RVR/cEVR. Among hematologic tests, hemoglobin and hematcrit was significantly higher in RVR/cEVR. Among blood chemistry tests, creatinine and low-density lipoprotein cholesterol (LDL-C) was significantly higher and gamma-glutamyltransferase (GGT), low-density-lipoprotein cholesterol (LDL-C), and blood sugar were significantly lower in RVR/cEVR. The level of HCV RNA was significantly lower in RVR/cEVR. There were no significant differences in other tests.

Multivariate logistic regression analysis was performed on age, fibrosis stage, steatosis, HCVRNA, creatinine, hemoglobin, GGT, LDL-C, and blood sugar: hematcrit was not included since it is closely associated with hemoglobin. On multivariate analysis, age, grade of steatosis, level of HCV RNA, creatinine, hemoglobin, GGT, and LDL-cholesterol remained significant whereas stage of fibrosis, hemoglobin and blood sugar were not.

The CART analysis

The CART analysis was carried out on the model building set of 269 patients using the same variables as logistic regression analysis. Figure 1 shows the resulting decision tree. The CART analysis automatically selected five predictive variables to produce a total of seven subgroups of patients. The grade of steatosis was selected as the variable of initial split with an optimal cut-off of 30%. The possibility of achieving RVR/cEVR was only 18% for patients with hepatic steatosis of 30% or more compared to 47% for patients with hepatic steatosis of less than 30%. Among patients with hepatic steatosis of less than 30%, the level of serum LDL-C, with an optimal cut-off of 100 mg/dL, was selected as the variable of second split. Patients with higher LDL-C level had the higher probability of RVR/cEVR (57% vs. 32%). Among patients with LDL-C of less than 100 mg/dL, age, with an optimal cut-off of 60, was selected as the third variable of split. Younger patients had the higher probability of RVR/cEVR (49% vs. 15%). Among patients younger than 60, the blood sugar, with an optimal cut-off of 120 mg/dL, was selected as the forth variable of split. Patients with lower blood sugar level had the higher probability of RVR/cEVR (71% vs. 31%). Among patients with hepatic steatosis of less than 30% and LDL-C of 100 mg/dL or more, age, with an optimal cut-off of 50, was selected as the third variable of split, younger being the predictor of higher RVR/cEVR probability (77% vs. 50%). Among patients older than 50,