60 秒でわかるプレスリリース

2008 年 7 月 18 日
独立行政法人 理化学研究所

染色体上の DNA 配列の重複・欠失の組合せが世界で初めて推定可能に

-塩基配列の個人差と病気の関連を探る手法が進展 -

生命の設計図である「DNA」は、アデニン、グアニン、シトシン、チミンのたった 4 種の塩基で描かれています。この 4 種の塩基の配列をもとに、生体内のさまざまなタンパク質が生み出され、生体を構成するとともに、情報を伝達して生命のメカニズムを動かします。塩基の 1 個の違い「一塩基多型：SNP」が、がんや生活習慣病、精神疾患の危険因子になりうるとして、世界中で、病気に関わる SNP を探す研究が急ピッチで進んでいます。近年、さらに、ゲノム上の 1,000 塩基以上のまとまった塩基配列が、欠失や重複、挿入したため、遺伝子などの配列のコピー数が個人間で異なる「コピー数多型：CNV」が、病気のかかりやすさや薬剤応答に大きく影響することが注目され始めました。

理研ゲノム医科学研究センター情報解析研究チームは、この CNV の構造を、数学的に推定する手法を世界で初めて開発しました。ヒトのように、父方、母方からそれぞれ 1 本ずつ、2 本の染色体を持つ場合、染色体ごとのコピーの構成（ハプロタイプ）は実験で求めるのは困難です。今回、開発した手法は、CNV の各染色体上のコピー数と種類を、数学的に解くアルゴリズムで、ハプロタイプ頻度を推定することが可能となりました。実際に、ヨーロッパ人、アフリカ人集団から得た実験データで、2 種類の遺伝子の CNV を調べると、2 つの集団のハプロタイプ頻度の違いが 16%以上もあることがわかりました。これは、この手法が患者集団と一般集団など、集団間で差のあるハプロタイプを見つけるのに有効であることを示しており、今後、この手法を活用すると、病気にかかわるハプロタイプを見つけ出し、新たな病気診断や治療法を確立することが期待できます。

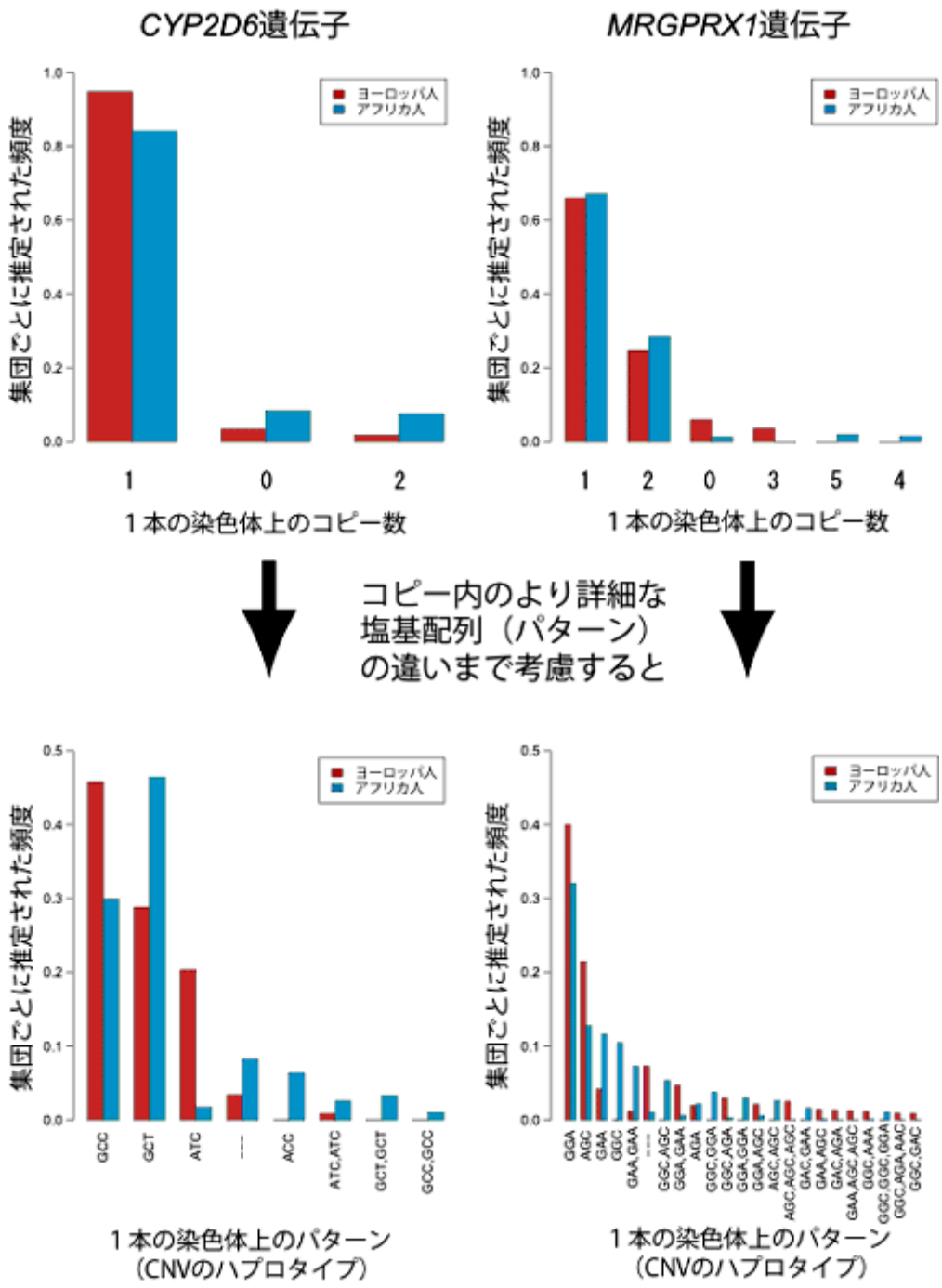

図　実際の遺伝子での観測データへの適用例

報道発表資料

2008 年 7 月 18 日
独立行政法人 理化学研究所

染色体上の DNA 配列の重複・欠失の組合せが世界で初めて推定可能に

－ 塩基配列の個人差と病気の関連を探る手法が進展 －

◇ポイント◇
- 実験データに適用し、ハプロタイプの推定が高い精度で可能に
- コピー数多型のハプロタイプ頻度が集団によって大きく異なることを発見
- 遺伝学や生物学分野でも有用な、DNA 配列変化の機構を解明する手段に

独立行政法人理化学研究所（野依良治理事長）は、がんをはじめとする病気の診断や治療に有用な情報となる、ゲノム中のコピー数多型（Copy Number Variation：CNV）※1の構造を、数学的に推定する手法を世界で初めて開発しました。理研ゲノム医科学研究センター（中村祐輔センター長）情報解析研究チームの角田達彦チームリーダー、加藤護研究員らによる成果です。

現在、病気の原因となる遺伝子の多型を探す研究が、世界中で精力的に行われていますが、その多型の 1 つとして、ゲノム上の遺伝子のコピー数が違うために病気の原因となりうるCNVが、注目を集めるようになってきました。これは、ゲノム上の 1,000 塩基以上のまとまった塩基配列が、欠失や重複・挿入し、遺伝子のコピー数が個人間で異なるようになったものです。ヒトのように、父方と母方からそれぞれ 1 本ずつ、2 本の染色体（相同染色体）を持つ場合、コピーの構成（ハプロタイプ※2）はそれらの染色体間で同じとは限りません。それぞれの染色体上でどのようなハプロタイプを持つかを実験的に観測することは極めて難しく、実際に集団のCNVのハプロタイプ頻度を求めることは非現実的でした。今回、これらを数学的に解くアルゴリズムを世界で初めて提案し、実験データからハプロタイプを推定することが可能になりました。医学的には、特定のハプロタイプを持つことによって病気が起こる場合が示唆され、この手法が、病気に関係するハプロタイプを発見するために有効となります。また、遺伝学や生物学の分野でも、CNVの生じるメカニズムや選択圧（選択と淘汰）のかかり方などを研究するために、不可欠な手法となります。

本研究成果は、米国の科学雑誌『*American Journal of Human Genetics*』（8 月 7 日号）に掲載されるに先立ち、オンライン版（7 月 17 日付け：日本時間 7 月 18 日）に掲載されます。

1.背　景

現在、糖尿病やがんなど、さまざまな病気の原因となる一塩基多型（SNP：Single Nucleotide Polymorphism）をヒトゲノム上で網羅的に探す研究が、臨床機関と協力した研究体制で急速に進展しています。さらに、構造的多型の 1 つとして知られる、コピー数多型（CNV：Copy Number Variation）という 1,000 塩基以上に及ぶ大きさの遺伝子多型が、注目を集めるようになってきました。これはゲノム上の 1,000 塩基以上のまとまった塩基配列が、欠失や重複・挿入し、遺伝子のコピー数

が個人間で異なるようになったものです（図 1）。近年、ヒトゲノムで広範に調べた結果、CNV領域がゲノムの 1 割以上にもあたる 3 億 6,000 万塩基にも及び、普遍的で、機能的に重要な存在であることがわかってきました。CNV領域は、遺伝子全体や制御領域まで含んでいることもあり、タンパク質のコード領域の配列、遺伝子発現制御（転写量）を変化させることによって、病気へのかかりやすさなどに大きな影響を及ぼしている可能性が高いことが示唆されるようになりました。

ヒトは、父方と母方からそれぞれ 1 本ずつ、計 2 本の染色体（相同染色体）を持っています。病気へのかかりやすさなどと遺伝的多様性との関連を調べるには、1 本の染色体上でのDNA配列の構成（ハプロタイプ）を知る必要があります。しかし、1 本の染色体上のハプロタイプを実験的に観測することは難しく（図 2a）、特に、たくさんのコピーがつながったCNVが存在したり、CNV内に塩基の違いがあったりすると、観測は極めて難しくなります（図 2b）。

集団 1 人 1 人の全染色体上での遺伝子多型解析などでは、DNA配列を読み取る高速な実験技術と、解析を行うための計算アルゴリズムが必要になります。理研ゲノム医科学研究センターは、これまでインベーダーアッセイ法※3とマルチウェルプレート上の定量PCR法※4を組み合わせた、高速な実験プラットフォームを開発してきました。しかし、実験から得たデータを使ってCNV領域内でのハプロタイプを推定する方法が確立できていなかったために、網羅的、体系的かつ精密にハプロタイプを決定し、そのCNVの性質を解析したり、疾患関連解析に用いたりすることができませんでした。

2. 研究手法と成果

実験から得られるデータでは、2 本の染色体におけるCNVのコピーの総数しか観測できませんでした。研究グループは、CNV領域内のハプロタイプを推定するために、EMアルゴリズム※5という数学的手法を基本にしたCNV解析アルゴリズムを開発しました。この解析アルゴリズムは、データの部分ごとにEMアルゴリズムを適用し、結果を統合するときにもEMアルゴリズムを用いる手法を採用して、データが大規模な場合でも推論を可能にしました。この新たなアルゴリズムを適用すると、染色体ごとに分離したCNVの情報を推定することが可能になりました（図 2a）。さらに、1 本の染色体上のコピー内にも塩基配列の違いがあるような複雑なCNV構造が存在する場合でも、それらの関係を数学的に解くことが可能になりました（図 2b）。特に、コピーの数や種類の多いCNVのハプロタイプも推定できます。また、このアルゴリズムをシミュレーションデータに適用した結果、サンプルサイズが大きくなるほど、推定精度が高くなることがわかりました。

研究グループは、実際の実験データへの使用例として、このアルゴリズムを、国際HapMapプロジェクト※6で対象としたヨーロッパ人とアフリカ人、2 つの集団で調べた*CYP2D6*遺伝子や*MRGPRX1*遺伝子のCNVデータに適用しました。まず、染色体ごとのコピー数の頻度を推定すると、2 集団で数%程度の違いを見つけました（図 3 上段）。次に、CNVの細かな違いも区別した詳細な推定を行うと、それらの違いが大きく広がり、最も一般的なハプロタイプ（図 3 下段左の[GCC]と[GCT]）でも、2 集団間の頻度の違いが 16%以上にも上りました。すなわち、このようなハプロタイプの集団間の違いを論じるには、こうした詳細な推定をすることが重要で

あることが明らかとなりました。例えば、*CYP2D6* 遺伝子内のそれぞれハプロタイプの違いは、食物中のアルカロイドなどの植物毒や、医療における薬の代謝能の違いを知るために重要となることがわかっています。今回の方法を使うと、疾患研究の場で、患者集団と一般集団間でCNVのハプロタイプ頻度の違いを検出することにより疾患や薬の作用・副作用に関連する遺伝子を新たに見つける関連解析が可能となることがわかりました。

3. 今後の期待

今回、新たなアルゴリズムを開発したことで、これまでに得てきた高速実験データから、CNV 領域内のハプロタイプを推定することができるようになりました。医学的には、特定のハプロタイプによって病気が起こる場合があり、今回の方法は病気に関係するハプロタイプを発見するために必須となる手法です。また、遺伝学や生物学の分野でも、CNV のような多型が生じるメカニズムや選択圧のかかり方などを研究するために、不可欠な手法となります。

研究グループは、今後、ゲノムワイドな高速実験技術に適した方法を提案していくとともに、実際の患者集団・一般集団の DNA データを今回の手法で解析し、病気にかかわる CNV の特定を行っていく予定です。

（問い合わせ先）
独立行政法人理化学研究所
ゲノム医科学研究センター　情報解析研究チーム
チームリーダー　角田　達彦（つのだ　たつひこ）
Tel：045-503-9556 / Fax：045-503-9555

横浜研究推進部　企画課
Tel：045-503-9117 / Fax：045-503-9113

（報道担当）
独立行政法人理化学研究所 広報室 報道担当
Tel：048-467-9272 / Fax：048-462-4715
Mail：koho@riken.jp

<補足説明>

※1 コピー数多型（CNV：Copy Number Variation）

個人ごとに、1 細胞あたりの遺伝子のコピー数が違うゲノム配列（領域）のことをいう。大きさが、1,000 塩基以上のものを対象とし、中には数 Mb（メガベース：100 万塩基）におよぶものもある。通常、人間の細胞には、遺伝子配列は父方と母方のそれぞれに由来する染色体の 1 本ずつに 1 個ずつあり、合わせて 2 個（2 コピー）存在する。それに対し、個人によっては合わせて 1 コピーしかなかったり（欠

失による)、3 コピーあったり（重複や挿入による）する。このような遺伝子配列の数の個人差（多様性）を CNV という。現在、CNV は人間のゲノムの 1 割以上を占める大規模な多型であることがわかってきており、自己免疫疾患など病気へのかかりやすさ、薬の作用・副作用、そのほかの多様性に大きくかかわっていることが示唆されている。

※2 ハプロタイプ

ハプロタイプは、1 本の染色体上にのっている DNA 配列の構成を指す。人間のように 2 倍体で相同染色体をもつとき、ハプロタイプは、一方の染色体上での各遺伝子座位の配列の組合せを指す。

※3 インベーダーアッセイ法

遺伝子多型解析法の 1 つ。もともと個人の各一塩基多型の遺伝子型を高速、高精度、安価に 判定するために考案された方法。今回の研究では定量性をより向上し、調べたい DNA 配列の存在する量の比を求めることができる。

※4 定量 PCR 法

ポリメラーゼ連鎖反応（PCR)により、調べたい DNA 配列の増幅量を経時的に測定する方法。増幅量から、もとの DNA のコピー数を定量することができる。

※5 EM アルゴリズム

観測できたデータを使い、観測できなかったものを推定する数学的手法。今回の研究では、各個人のコピーの総数が実験的に観測されるが、総数が 0 コピーの人はパターンが一意に決まるなど、パターンにはある程度の制限があり、その制限の中で個人のパターンをまず適当に仮定する。次に、この仮定によって集団となったときのパターンの分布を推定する。さらにこの結果を使って、もともとの個人のパターンを推定し直す。この時の推定は、前回より精度がよくなることが数学的に保証されている。さらに、この推定結果を使って集団のパターンの分布を、再び、よりよい精度で推定する。この一連の操作を数多く繰り返すことによって、各個人のもつパターンや集団全体でのパターンの分布を最終的に精度よく推定することができる。

※6 国際 HapMap プロジェクト

ヒトゲノム上の多型情報を臨床応用していくために不可欠な、ハプロタイプ地図を作成する国際計画。2002 年 10 月に開催した「国際 HapMap プロジェクト戦略会合」でこの地図の作成を日・米・英・中・加の協力により取り組むことが合意された。わが国からは、理研遺伝子多型センターの中村祐輔センター長が研究代表者としてプロジェクトに参加した。同プロジェクトでは、アジア人（日本人を含む)、欧米人、アフリカ人のそれぞれの人種について、各国より血液サンプルを収集するとともに、ハプロタイプ地図の作成を実施することを目標とし、達成された。プロジェクトで得られたデータをもとにした、全ゲノム規模の関連解析用実験プラットフォームが開発され、現在、多くの関連解析により疾患関連遺伝子が発見されつつある。

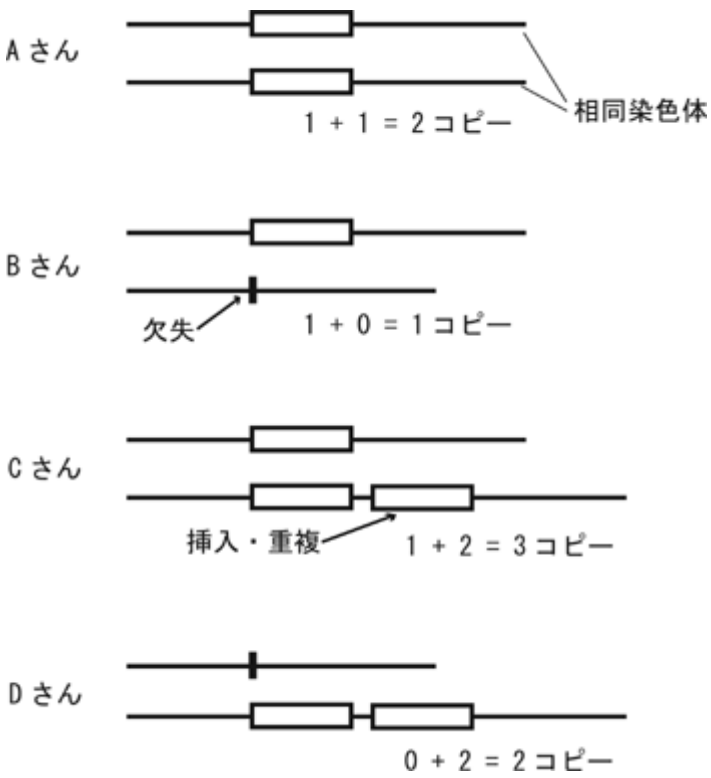

図 1　コピー数多型の基本的な概念

通常、各個人のそれぞれの染色体（例えば 1 番染色体など）には、父方と母方由来の計 2 本の染色体（相同染色体）上に遺伝子配列が 1 つ（1 コピー）ずつあり、計 2 コピーである（図中 A さんの例）。しかし、過去に起こった欠失や挿入・重複によって、各染色体上の数に、個人間で違いが見られるようになる（B さんや C さん）。これがコピー数多型である。一方の染色体上で 0 コピー、もう片方で 2 コピーの場合（D さん）、総数は A さんと同じ 2 コピーに見えるが、病気などへの影響はまったく異なる可能性があり、注目されている。

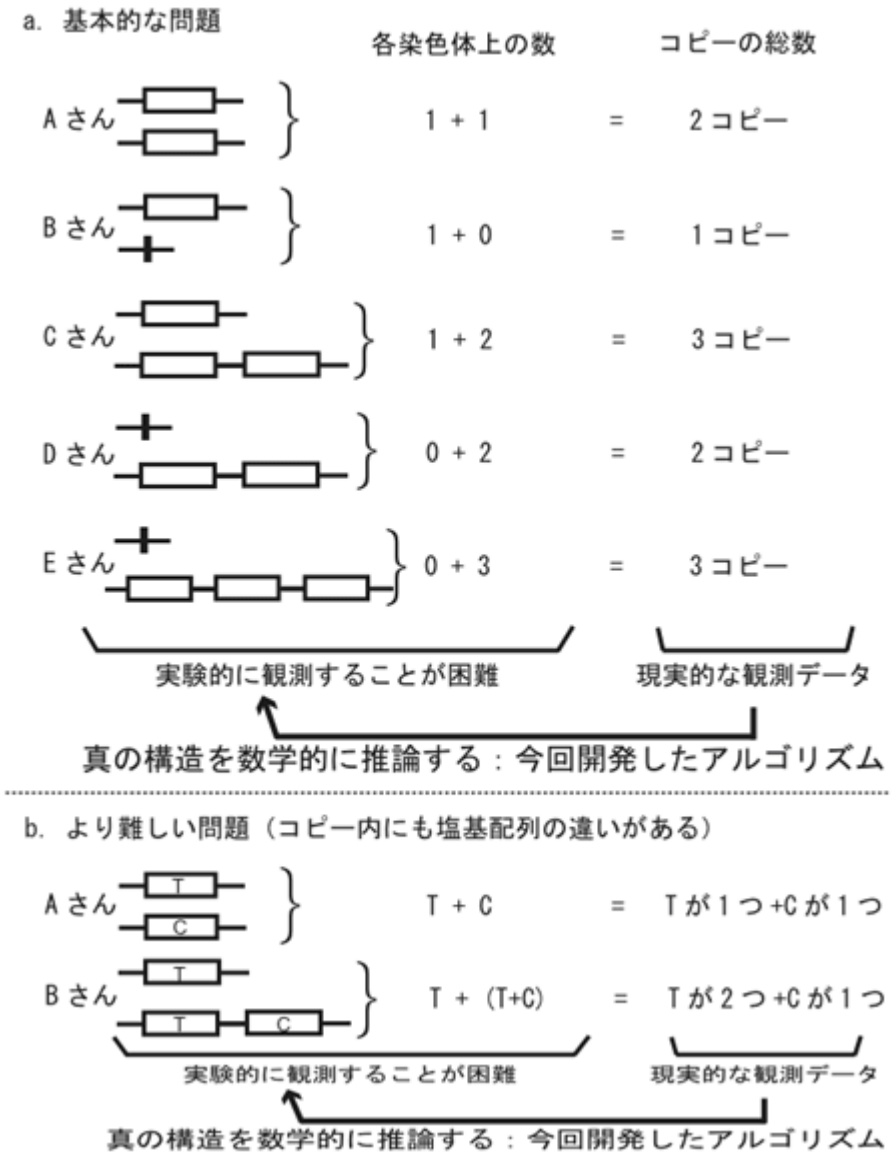

図 2　コピー数多型の問題と推論の必要性

a. 基本的な問題：各個人のコピー数多型は、父方と母方由来の計 2 本の染色体（相同染色体）を持つため、その構成は複雑で、どちらの染色体に何個ずつあるかを実験的に観測することは難しい（左）。観測できる 2 本の染色体上の総数（右）から推論するのが今回の手法。

b. より難しい問題：コピー内にも塩基の違いがあるような、より複雑な問題も解くことができる。

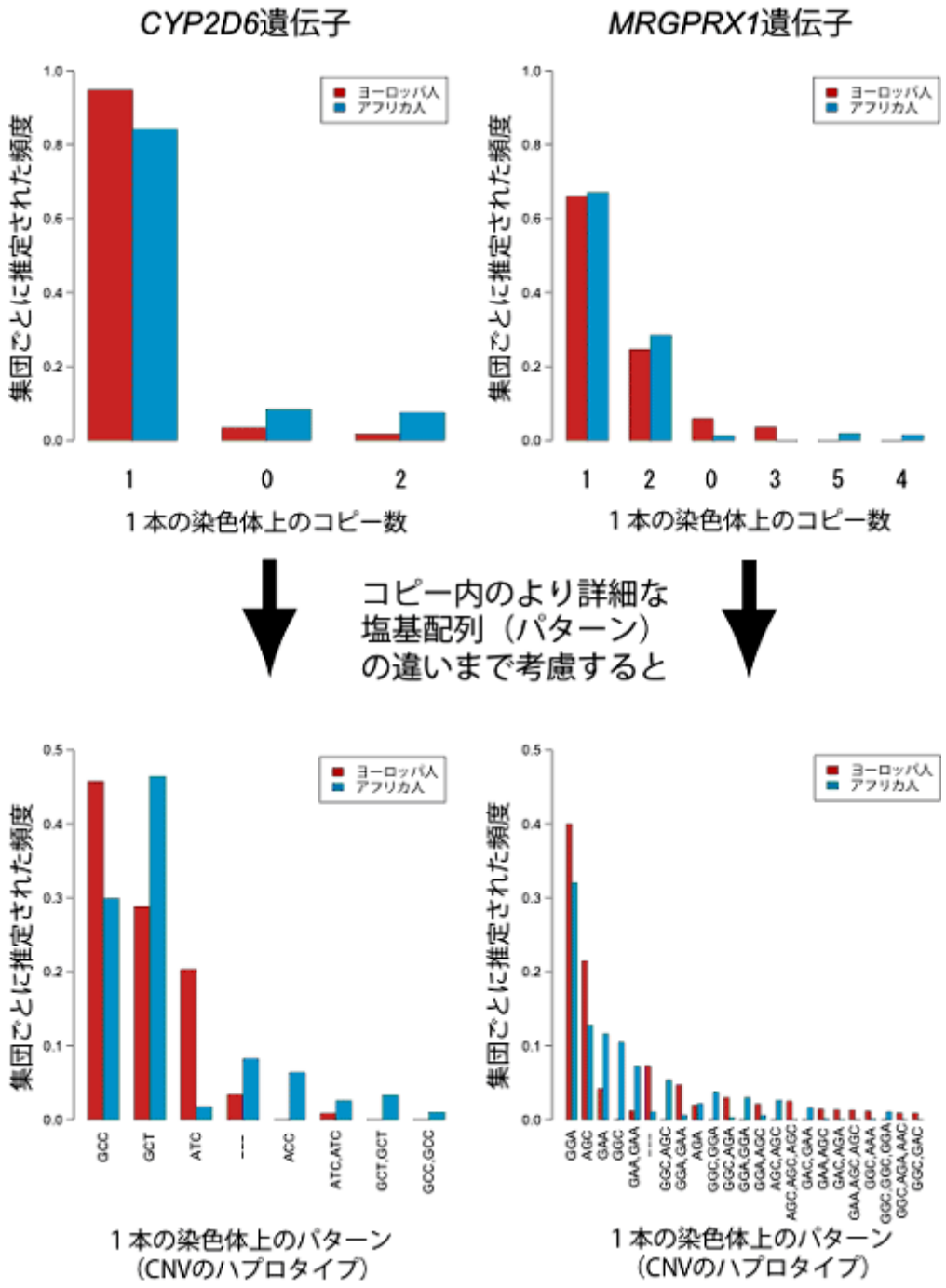

図３　実際の遺伝子での観測データへの適用例

CYP2D6 遺伝子（左）、*MRGPRX1* 遺伝子（右）上での実験データに、今回開発したアルゴリズムを適用し解いた結果。コピー数のみの基本的問題を解き、１本の染色体上のコピー数に対し、各集団内で観測される頻度をグラフにすると、多くは１コピーずつ存在しているが、のっていないもの（0 コピー）や 2 コピー以上存在しているものもある。ヨーロッパ人とアフリカ人では数%ずつ違う（上段）。コピー内の塩基の違いまで考慮し詳細に調べると（塩基の違う箇所のみで表現している）、集団間の頻度の違いがより顕著に見られる（この例では最大 16%の差）（下段）。