PRESS RELEASE

2022 年 11 月 2 日
理化学研究所
静岡県立総合病院
静岡県立大学

ヒトの複雑な形質に対する希少なコピー数多型の影響

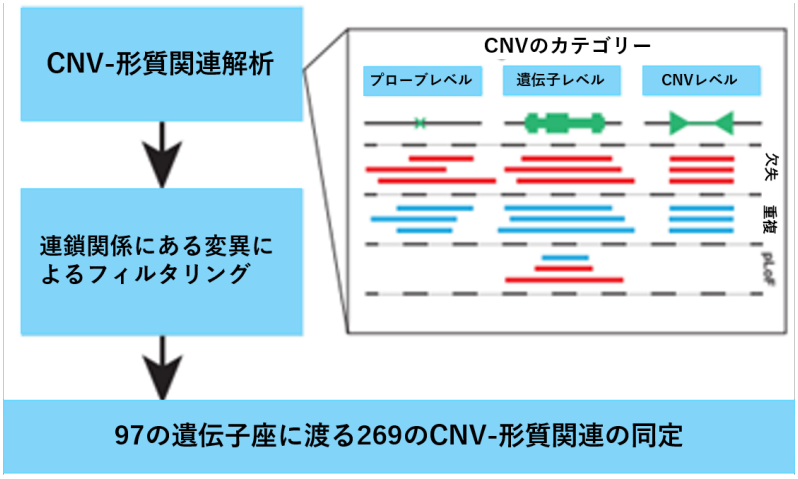
ー新たな多型検出方法の開発による成果ー

理化学研究所（理研）生命医科学研究センターゲノム解析応用研究チームの寺尾知可史チームリーダー（静岡県立総合病院免疫研究部長、静岡県立大学特任教授）らの国際共同研究チームは、ヒトゲノムに数十万個存在するといわれるコピー数多型（CNV）[1]を従来の方法よりはるかに高感度に検出する手法を開発しました。

本研究成果は、現在の技術では遺伝子型決定が困難な遺伝子領域の遺伝的多型とこれに関連する形質への影響の解明につながると期待できます。

今回、国際共同研究チームは、世界で数千万人以上の規模でデータが存在する DNA マイクロアレイ[2]（SNP[2]アレイ）のデータを用いて、祖先から継承された染色体の一部（セグメント）によって CNV を検出する「HI-CNV（Haplotype-Informed Copy-Number-Variation）」という手法を確立しました。HI-CNV では、従来法（PennCNV[3]）の 6 倍以上の CNV を検出しました。また、HI-CNV を UK バイオバンク[4]のデータに適用することで、CNV と 56 の量的形質[5]との関連を詳細に解析し、97 の遺伝子座にわたる 269 の独立した CNV-形質関連を同定することに成功し、主要な関連を日本人の結果でも確認しました。今後、HI-CNV の枠組みを全エクソームシーケンス[6]データや全ゲノムシーケンス[7]データに拡張することで、DNA マイクロアレイデータでは網羅できなかった CNV の検出が可能になります。

本研究は、科学雑誌『*Cell*』オンライン版（10 月 27 日付）に掲載されました。



97 の遺伝子座にわたる 296 の独立した CNV-形質関連を同定

背景

ヒトゲノムには、1 細胞当たり通常 2 コピーの遺伝子が存在しますが、遺伝子のコピー数には個人差があり、ある個人によっては 1 コピーのみ（欠失）あるいは 3 コピー以上（重複）となり、これを「コピー数多型（CNV）」と呼びます。CNV は、精神神経疾患を含む多くのゲノム疾患の原因となることが知られています。

CNV はタンパク質をコードする遺伝子のコード配列に直接的に影響を及ぼし、タンパク質の機能喪失を引き起こすだけでなく、遺伝子量の増大や制御要素の欠損を引き起こし、間接的にコード配列の発現量、ひいてはタンパク質の発現量に大きく影響を及ぼします。従って、CNV が「形質」に与える影響を調べることは、形質への影響力を持つ新たな変異体を発見し、複雑な形質の遺伝的構造に関する理解を深める可能性を秘めています。

しかし、これまで、十分な検出力を持つフェノムワイド CNV 関連解析[8]は、バイオバンク規模のコホート（集団）で利用できる低コストの DNA マイクロアレイから検出される大きな CNV（数十 kb 以上）の検討に限定されていました。

研究手法と成果

国際共同研究チームは、バイオバンクコホート内のハプロタイプ[9]（祖先から継承された染色体）の共有を利用して、より感度の高い CNV 検出法「HI-CNV（Haplotype-Informed Copy-Number-Variation）」を開発しました。

まず、Positional Burrows-Wheeler transform（PBWT）と呼ばれるアルゴリズムを用いて、2 個体間で対立遺伝子が祖先と同じものを共有する状態の IBD（identity-by-descent）セグメントを迅速に特定し、各ゲノム位置において最も近い「haplotype neighbors」、すなわちコホート内の他のハプロタイプと最も長くマッチする IBD セグメントを特定しました（図 1）。次に、個体の遺伝データから CNV が存在する可能性に関する定量的情報と、haplotype neighbors から対応する情報を利用して、共通祖先に由来するハプロタイプ上で共有された CNV を隠れマルコフモデル[10]を用いて検出しました。

さらに、UK バイオバンクコホートで利用可能な SNP アレイの遺伝子型プローブ強度データに HI-CNV を適用するために、対立遺伝子特異的プローブ強度測定値をコピー数尤度（ゆうど）[11]に関する確率的情報に対応付ける確率的モデルを学習する方法を開発しました。CNV 内の遺伝子型プローブは、CNV 内にないプローブと比較して、特徴的な強度測定値を生成し、CNV を共有している複数の個体で一貫した偏差が観察されると、シグナルがより明確になることを利用しています。

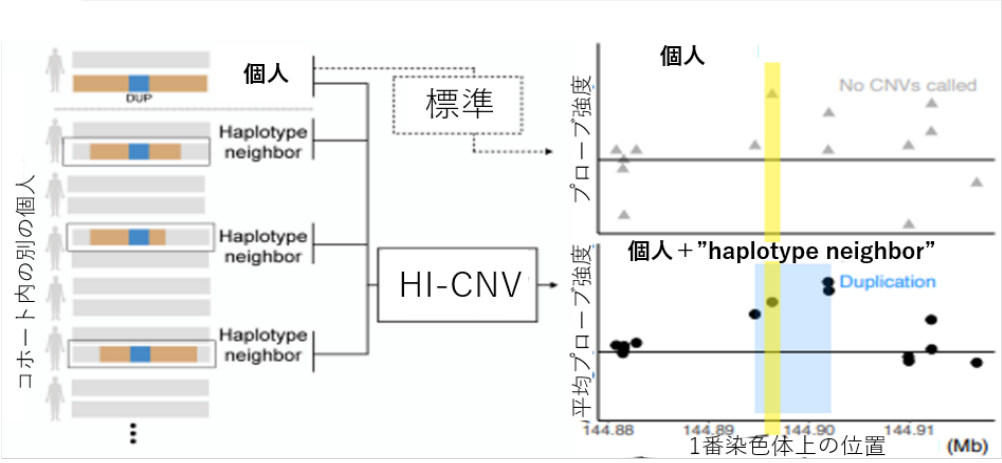


図 1 バイオバンクの SNP アレイデータから HI-CNV により CNV を検出する基本的枠組み

従来の CNV を検出する標準的な手法では、個人ごとに別個に CNV に含まれる SNP の情報（黄色の部分）を用いて CNV を同定していた。これに対し、HI-CNV は、ある個体の SNP アレイデータと長い共有ハプロタイプ（水色の部分）を持つ個体（"haplotype neighbors"）の対応するデータとを一緒に解析して、CNV の検出率を向上させる。

HI-CNV を UK バイオバンクの登録者 45 万人に適用した結果、従来法（PennCNV）の 6 倍以上の CNV を検出しました（図 2a）。43 人の参加者の全ゲノムシーケンスパイロットデータを用いた検証分析では、HI-CNV の検証率は約 91%と PennCNV と同等であり、正確性を保持したまま検出力が向上したことが確認されました（図 2b）。さらに、バイオバンク・ジャパン[12]の登録者 18 万人に HI-CNV を適用したところ検出率は約 93%と、UK バイオバンクと同様の性能であることが確認されました。これら HI-CNV の検出感度の上昇は、従来 SNP アレイデータでは検出困難でありながら全 CNV の大部分を占めていた、10kb 以下の CNV に対する検出能力の向上によるものです（図 2c）。

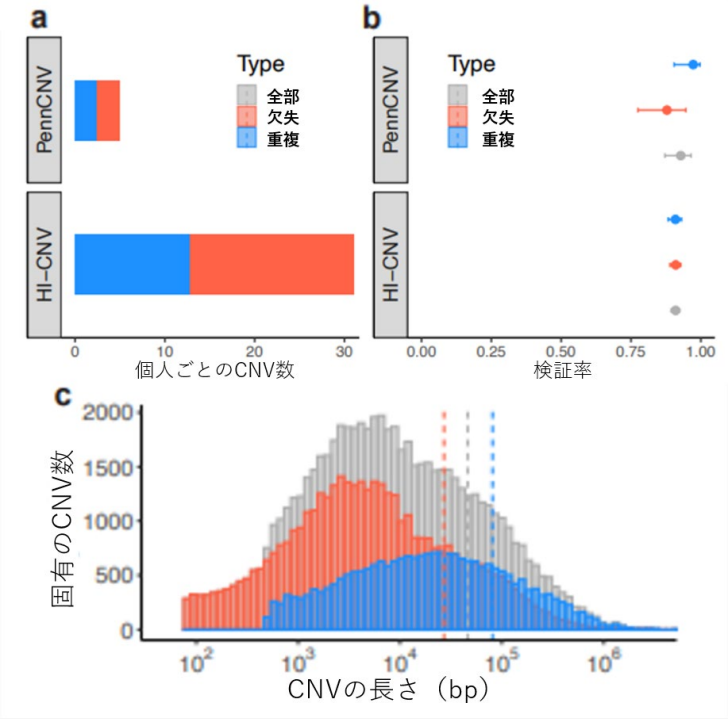


図 2 従来法（PennCNV）から大幅に検出力を改善した HI-CNV

a) HI-CNV を UK バイオバンクデータに適用すると、従来法（PennCNV）の 6 倍以上の CNV（欠失・重複）を検出した。
b) 全ゲノムシーケンスパイロットデータを用いた検証解析の結果。HI-CNV の検証率は約 91%で、PennCNV と同等であった。
c) 固有の CNV 数と CNV の長さの関係。赤、灰色、青の破線は各 CNV の長さの平均を示す。HI-CNV の検出感度は、10kb（10^4b）以下の CNV に対する検出力の向上によるものであった。

次に、検出された三つのカテゴリー（プローブレベル、遺伝子レベル、CNV レベル）の CNV と、身体測定形質、血圧、肺機能、骨密度、血球指標、血清バイオマーカーなど 56 の遺伝性定量形質の関連を調べました。線形混合モデルを用いて、UK バイオバンク登録者 45 万人について関連解析を実施し、さらにファインマッピング[13]を実施しました。その結果、97 の遺伝子座における 269 の CNV-形質関連を特定しました（図 3）。

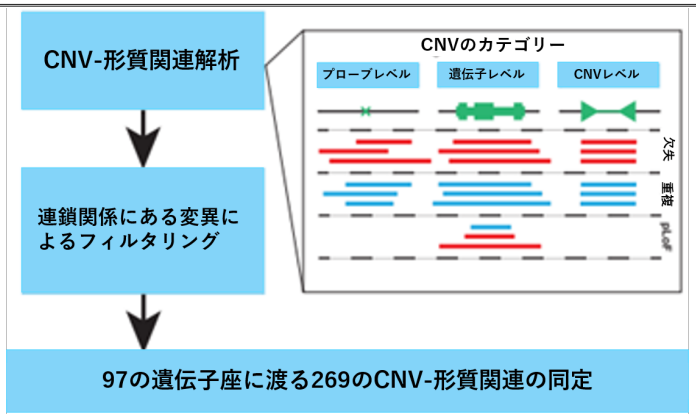


図 3 CNV-形質関連解析のパイプライン

検出された三つのカテゴリー（プローブレベル、遺伝子レベル、CNV レベル）の CNV と、身体測定形質、血圧、肺機能、骨密度など合計 56 の形質との間で関連解析を実施し、さらにファインマッピングを実施することで 97 の遺伝子座における 269 の CNV-形質関連を特定した。

269 のファインマッピングされた CNV-形質関連に関与する 97 の遺伝子座のうち、72 の遺伝子座は新しい CNV-形質関連を示しました。72 の新しい遺伝子座位の約半分である 35 座位については、標的遺伝子を同定できました。また、バイオバンク・ジャパンにおいて検出された CNV を用いて、これら新しい遺伝子座位のうち形質などの条件が適切な 14 の関連に対して再現解析[14]を行ったところ、13 の関連がその効果量について UK バイオバンクと同じ方向性を示し、ほぼ一貫した効果量が観察されました。

今回同定された新しい CNV-形質関連遺伝子座の中には、（A）近傍（500kb 以内）のどの SNP よりも強い関連を示す遺伝子座、（B）CNV が近傍の SNP とともに長い対立遺伝子系列を形成する遺伝子座、（C）推定上の標的遺伝子に新たに関与する追加遺伝子座が含まれていました。

例えば、E3 ユビキチン-タンパク質リガーゼをコードする *UHRF2* 遺伝子上の非常にまれな CNV（UK バイオバンク登録者 19 人が保有）は、身長の減少に関連していました（$P=8.2\times10^{-11}$）。*UHRF2* と身長との関連については、同じ遺伝子座の SNP はどれもゲノムワイドな有意性（$P<5.0\times10^{-8}$）を示していませんでした。また、骨形成タンパク質をコードする *BMP5* 遺伝子の上流に位置する低頻度（MAF[15]=2.2%）の欠失は骨密度増加と強く関連しており（$P=9.2\times10^{-82}$）、近傍の強い関連を示す SNP（$P=3.8\times10^{-51}$）とともに長い対立遺伝子系列を形成することが示唆されました。これらの結果は、関連する欠失 CNV をゲノムワイド関連解析（GWAS）[16]におけるファインマッピングに含めることの重要性を示しており、CNV と SNP を含む対立遺伝子系列のさらなる探索が必要であるといえます。

さらに、機能不明の遺伝子である *R3HDM4* における非常に希少な欠失は、幼弱な網状赤血球数の増加と関連していました（$P=3.5\times10^{-11}$）。この関連性は *R3HDM4* 遺伝子の PTV（proteintruncating SNP or indel variant）[17]でも確認され（$P=2.7\times10^{-7}$）、共通のイントロン（ゲノム上でタンパク質をコードしていない領域）SNP も網状赤血球数増加に対して強い有意性を示しました（$P=6.6\times10^{-}$

[86]。この欠失を詳しく調べると、エクソン（ゲノム上でタンパク質をコードしている領域）に重なる pLoF（predicted to cause loss of function）[18]欠失と、*R3HDM4* 遺伝子の第 1 イントロン内に完全に位置するイントロン欠失の両方からなることが分かりましたが、いずれも網状赤血球数の増加に関連していることが判明しました。

これらの結果は、*R3HDM4* 遺伝子のエンハンサー機能[19]が予測されるクロマチン領域[20]を含む欠失がまたがるイントロン領域が重要な制御的役割を担っていることを示唆しています。つまり、制御的 CNV が形質に大きな影響を与え、時にはコーディング配列そのものに変化を与える CNV と同じくらい強い影響を与える可能性があることを改めて示しています。

今後の期待

今回の研究では、新たな手法 HI-CNV により CNV の検出力が大幅に向上し、これに伴い 72 の遺伝子座にわたる多くの新しい CNV-形質関連を同定することができました。同定された CNV-形質関連遺伝子座には、GWAS による既知の SNP-形質関連遺伝子座も含まれており、CNV が近傍の SNP よりもより強く関連している場合や、近傍の SNP とともに形質との強い関連を示す対立遺伝子系列を形成している場合があることが示されました。今後は、これら CNV-形質関連と SNP-形質関連を組み合わせて解釈することで、より正確な形質と関連遺伝子座の関係が解明されると期待できます。

また、今回 HI-CNV の解析対象となった SNP アレイデータによって検出された CNV は、通常ヒトゲノムに存在する数千の CNV のごく一部に過ぎません。今後、解析対象の枠組みを全エクソームシーケンスあるいは全ゲノムシーケンスデータに拡張することで、現状では遺伝子多型の決定が困難な遺伝子座における CNV の検出が可能となり、さらなる CNV-形質関連が同定されると期待できます。

論文情報

＜タイトル＞
Influences of rare copy number variation on human complex traits
＜著者名＞
Margaux L.A. Hujoel, Maxwell A. Sherman, Alison R. Barton, Ronen E. Mukamel, Vijay G. Sankaran, Chikashi Terao, Po-Ru LohMargaux L.A. Hujoel, Chikashi Terao, Po-Ru Loh
＜雑誌＞
Cell
＜DOI＞
10.1016/j.cell.2022.09.028

補足説明

[1] コピー数多型（CNV）
ヒトの DNA 配列を個々人で比較すると、塩基配列の違い（遺伝子多型）が見いだされる。遺伝子多型のうち、1kb 以上の大きさの変異をコピー数多型という。同じ配列が繰り返される重複や、連続した配列が欠損する欠失などがある。タンパク質をコードする領域に生じれば、影響はより強くなる。CNV は copy number variation の略。

[2] DNA マイクロアレイ、SNP
DNA マイクロアレイは、基板の上に、遺伝的多型（主に SNP）に相補的なプローブを搭載したビーズを高密度に配置し、数十万～数百万の遺伝的多型を検出するための分析器具。SNP は一塩基多型と呼ばれ、変異が一つの塩基に限られるものを指す。SNP は single nucleotide polymorphism の略。

[3] PennCNV
従来の DNA マイクロアレイを用いた CNV 検出アルゴリズムの代表的プログラム。アレイのプローブシグナルを基に、個人ごとに CNV の検出を行う。

[4] UK バイオバンク
英国で構築されているバイオバンクであり、50 万人規模の疾患罹患情報、臨床情報、遺伝情報などから構成される。

[5] 量的形質
ヒトの身長や体重、胸囲などのように連続的、量的に変化する形質をいう。

[6] 全エクソームシーケンス
ゲノム中のタンパク質に関する情報を含むエクソン領域のみを配列解読すること。エクソン領域の一塩基多型（SNP）や、短い塩基の挿入（ゲノム配列の特定の位置に別の配列が挿入された形態）または欠失（ゲノム配列の一部が失われた形態）などを検出できる。疾患を引き起こす変異は、エクソン領域に多く存在していることが知られており、シーケンスの対象をエクソン領域に絞ることで、全ゲノムシーケンスと比べて低コストで重要な変異を検出できる。

[7] 全ゲノムシーケンス
全ゲノム DNA を鋳型として配列解読をすること。この配列解読によって、全ゲノム長の数倍～数十倍の総塩基数に相当するショートリードまたはロングリードデータが生成される。

[8] フェノムワイド CNV 関連解析
遺伝子多型（ここでは CNV）を用いて、形質の変化と遺伝子多型の頻度差の関連を統計学的に検定する方法。多様な形質を対象として網羅的に解析するため、フェノムワイド（Phenome-wide）と呼ばれる。

[9] ハプロタイプ
生物が持っている単一の染色体上の塩基配列のことをいう。父親由来、母親由来のハプロタイプを各個人が保有している。ジェノタイプ（遺伝子型）は多型の場所におけ

る塩基配列の組み合わせのため、ハプロタイプを決定するには統計学的なデータ解析が必要になる。

[10] 隠れマルコフモデル
観測された記号系列の背後に存在する状態の遷移系列を推測するために用いられるモデル。

[11] 尤度（ゆうど）
統計学における尤もらしさの程度。目の前の観察結果を基に、観測できないパラメータを推定する際に用いられる。

[12] バイオバンク・ジャパン
オーダーメード医療実現化プロジェクトの基盤となる DNA サンプルや血清サンプルを 47 疾患（延べ約 20 万人）から収集し、臨床情報とともに保管している世界でも有数の資源バンク。

[13] ファインマッピング
関連解析の手法の一つ。ある遺伝子座における、疾患に対する真の原因 SNP の数を仮定した場合に、その SNP が真に疾患の原因である確率を算出する手法。

[14] 再現解析
元の研究の基本的な発見が、研究の参加者や状況が異なる場合にも適用できるかどうかを判断するために、異なる状況や異なる被験者で、研究調査を繰り返すこと。

[15] MAF
個々のヒトゲノムを比較すると、染色体上の場所が同一であっても、遺伝子や個々の塩基配列が異なる場合がある。これらの遺伝子や塩基配列をアレルという。ある染色体上の位置において、個人により異なる塩基配列を持つ場合、集団の中で 2 番目に高いアレルの頻度を MAF という。MAF は Minor Allele Frequency の略。

[16] ゲノムワイド関連解析（GWAS）
疾患の感受性遺伝子を見つける方法の一つ。ヒトのゲノム全体を網羅する遺伝子多型を用いて、疾患を持つ群と疾患を持たない群とで遺伝子多型の頻度に差があるかどうかを統計学的に検定する方法。検定の結果得られた P 値（偶然にそのようなことが起こる確率）が低いほど相関が高いと判定できる。GWAS は、Genome-Wide Association Study の略。

[17] PTV（proteintruncating SNP or indel variant）
遺伝子のコード配列を短縮することが予測される遺伝子変異（一塩基多型や挿入欠失による変異）のことをいう。

[18] pLoF（predicted to cause loss of function）
遺伝子産物であるタンパク質の機能が低下または消失する、機能喪失型の変異（LoF）を引き起こすと推測される変異のことをいう。

[19] エンハンサー機能

非翻訳領域の中で、遠位から遺伝子の転写の可能性を高める機能領域のことをいう。

[20] クロマチン領域
核 DNA とヒストンタンパク質の複合体。DNA がヒストンタンパク質に巻き付きコンパクトに折り畳まれている。遺伝子の転写がほとんど行われないヘテロクロマチン領域と活発に転写が行われるユークロマチン領域がある。

国際共同研究チーム

理化学研究所　生命医科学研究センター
ゲノム解析応用研究チーム
チームリーダー　寺尾知可史　（テラオ・チカシ）
（静岡県立総合病院　免疫研究部長、静岡県立大学　特任教授）
ハーバード大学　ブリガムアンドウィミンズ病院（米国）
助教授　ポールー・ロウ　（Po-Ru Loh）

研究支援

本研究は、日本医療研究開発機構（AMED）難治性疾患実用化研究事業「シングルセル統合ゲノミクス解析が解き明かす強皮症の病態基盤の開発（研究代表者：寺尾知可史）」、同ゲノム医療実現推進プラットフォーム事業（先端ゲノム研究開発）「先天的/後天的構造多型に着目した免疫/精神疾患病態解明に関する研究開発（研究代表者：寺尾知可史）」、同革新的がん医療実用化研究事業「体細胞モザイクのがん発症および予後因子としての意義解明の開発（研究代表者：寺尾知可史）」の助成を受けて行われました。

発表者・機関窓口

＜発表者＞　※研究内容については発表者にお問い合わせください。
理化学研究所　生命医科学研究センター
ゲノム解析応用研究チーム
チームリーダー　寺尾知可史　（テラオ・チカシ）
（静岡県立総合病院　免疫研究部長、静岡県立大学　特任教授）
Tel: 045-503-9553
Email: chikashi.terao [at] riken.jp

寺尾　知可史

＜機関窓口＞
理化学研究所　広報室　報道担当
Tel: 050-3495-0247
Email: ex-press [at] ml.riken.jp

静岡県立総合病院　総務課
Tel: 054-247-6111　Fax: 054-247-6140
Email: sougou-soumu [at] shizuoka-pho.jp

静岡県立大学　教育研究推進部　広報・企画室
Tel: 054-264-5130
Email: koho [at] u-shizuoka-ken.ac.jp

※上記の[at]は@に置き換えてください。