PRESS RELEASE

2023 年 5 月 17 日
理化学研究所
静岡県立総合病院
静岡県立大学

高精度の構造多型検出手法を開発

―疾患や遺伝形質に関わる構造多型や遺伝子の同定が可能に―

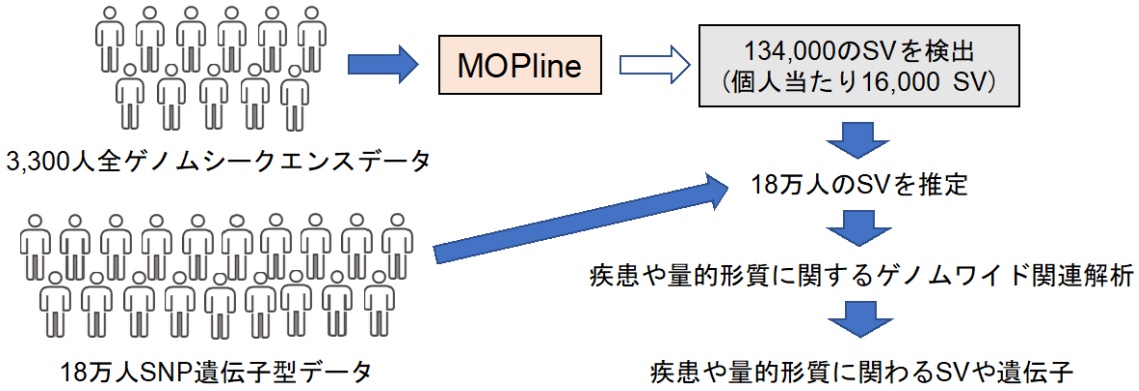
理化学研究所（理研）生命医科学研究センターゲノム解析応用研究チームの小杉俊一研究員（研究当時、現客員研究員、静岡県立総合病院リサーチサポートセンター遺伝研究部研究員）、寺尾知可史チームリーダー（静岡県立総合病院臨床研究部免疫研究部長、静岡県立大学薬学部ゲノム病態解析分野特任教授）らの共同研究グループは、全ゲノムシークエンス[1]データから「構造多型（SV）[2]」を高精度で検出する新しい手法を開発しました。

本研究成果は、これまで発見できなかった疾患や形質の原因となる遺伝子やゲノム変異の同定に貢献すると期待できます。

SV とは、個人間のゲノムの違いのうち 50 塩基対以上の長さの変異のことです。これまで SV を検出する多くのツールが開発されてきましたが、精度よく検出できる単独のツールは存在しませんでした。

今回、共同研究グループは既存の複数のツールを用いて SV を高精度に選別し、この選別過程で抜け落ちた SV を独自の遺伝子型[3]判定手法により回収する、新しい SV 検出手法「MOPline」を開発しました。MOPline を用いて、バイオバンク・ジャパン（BBJ）[4]に登録された約 3,300 人の全ゲノムシークエンスデータから約 134,000（個人当たり約 16,000）の SV を検出し、この SV と約 18 万人の BBJ データを用いて解析をしたところ、多くの疾患や量的形質[5]に SV が関わっていることが明らかになりました。

本研究は、科学雑誌『*Cell Genomics*』オンライン版（5 月 18 日付：日本時間 5 月 19 日）に掲載されます。



構造多型（SV）を用いて疾患や量的形質に関わる SV や遺伝子を究明

背景

ゲノムの「構造多型（SV）」は、50 塩基対（bp）以上の欠失[6]、挿入[7]、重複[8]、逆位[9]多型の総称であり、50bp より小さい欠失、挿入に相当する「インデル」および 1bp の塩基置換である「一塩基多型（SNV）」[10]とは区別されます。SV の出現頻度は個人当たり 1 万～2 万と、インデル（個人当たり約 70 万）や SNV（個人当たり約 400 万）に比べて低いものの、サイズが大きいために、SV に起因する個人ゲノム間の異なる塩基数は、SNV による違いの塩基数の 3～10 倍あることが示されています。

このように個人ゲノム間に大きな違いをもたらす SV は、発達障害や知的障害を含むさまざまなヒトの疾患・形質の遺伝的要因となることが近年の多くの研究から示されています[注 1、2)]。また、がんなどの体細胞変異によって引き起こされる疾患においても、SV が関わることを示す多くの研究があります[注 3、4)]。

一方で、SV の構造の複雑さと大きいサイズのために、SV の検出は SNV と比較して困難です。ゲノムの多型は通常、100～150bp の短い配列（リード[11]）データをヒトの標準ゲノム配列（リファレンス配列[12]）にアライメント[13]して検出します。このリード長内に収まる SNV やインデルに対して、より大きなサイズの SV はリード内に収まらず、SV をまたいでアライメントされるリードの間接的な証拠を用いて検出しなければならないため、検出精度（検出の正確性）や検出感度（検出の効率）が低くなってしまいます。これまでに多くの SV 検出ツールが開発されてきましたが、検出結果の共通性が低いという問題があり、単一のツールで精度、感度ともに高く SV を検出できるツールは存在しませんでした[注 5)]。

注 1）Weischenfeldt J, *et al.* Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* 14, 125-38 (2013).
注 2）Marshall, C.R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* 49, 27-35 (2017).
注 3）Yi, K. *et al.* Patterns and mechanisms of structural variations in human cancer. *Exp. Mol. Med.* 50, 98 (2018).
注 4）Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534, 47-54 (2016).
注 5）Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 20, 117 (2019).

研究手法と成果

全ゲノムシークエンスデータから高い信頼度（高精度）を持つ SV を取得する方法の一つは、既存の SV 検出ツール間で共通に検出される SV（ツール間オーバーラップ SV）を選別することです。しかし共同研究グループは、必ずしもツール間オーバーラップ SV が高い精度を示すわけではないことを見いだしました。

そこで、オーバーラップ SV が高い精度を示す既存ツールの組み合わせを調べました。そして、既存の 4～9 個のツールを用いて最適なツールの組み合わせを SV タイプやサイズごとに決定するアルゴリズムを開発し、MOP（Merging Overlap calls from selected Pairs of algorithms）と名付けました（図 1 上）。

MOP を用いると、高い精度を持った SV を選別できますが、一部の SV は見

逃してしまいます。この問題を解決するために、MOP で SV が検出されなかったゲノム領域をスキャンし、SV の存在を確認する作業を行いました。この存在確認では、リードのアライメント情報を用いた独自の遺伝子型判別手法を用い、この SV の再判別手法を SMC（Supplementing Missing Calls）と名付けました（図 1 下）。そして最終的に、MOP、SMC、およびフィルタリングやアノテーション[14]機能を組み合わせた SV 検出手法、「MOPline」の開発に成功しました（図 1）。

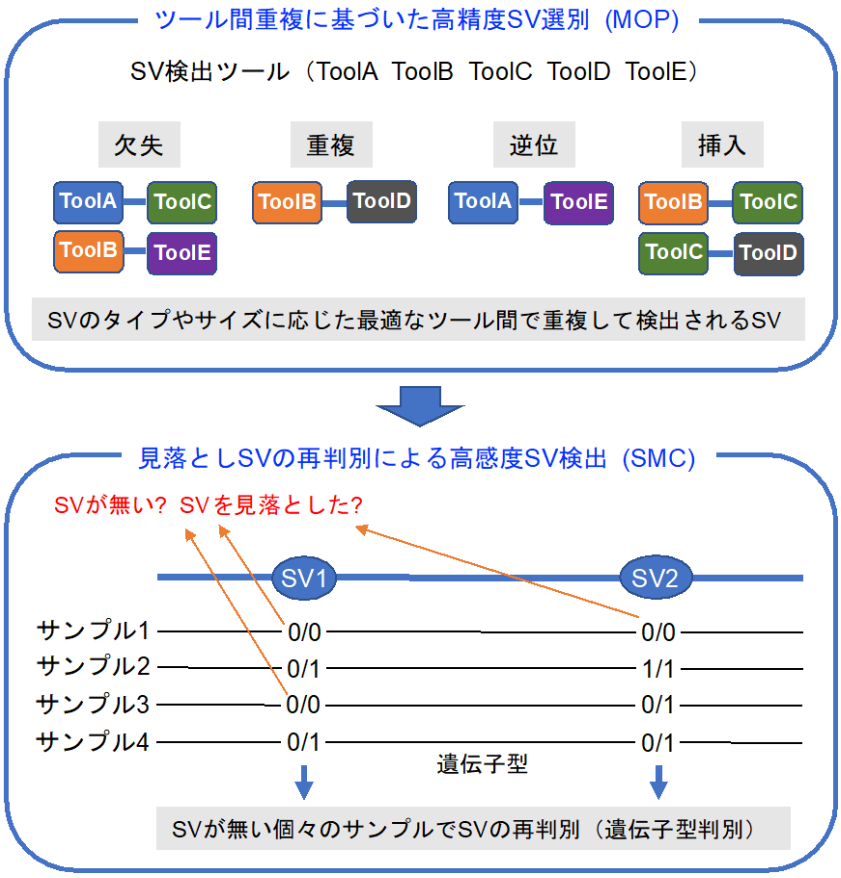


図 1　今回開発した「MOPline」のアルゴリズム

構造多型（SV）のタイプやサイズに応じて最適なツールの組み合わせを選別し、精度の高いツール間共有 SV を選別（MOP）。次に SMC アルゴリズムを用いて見落とした SV を検出し、検出感度を向上させる。

MOPline の SV 検出精度、検出感度を NA12878 などの全ゲノムシークエンスデータを用いて検証したところ、MOPline は既存のツールの精度、感度を上回っていました。さらに、複数のツールを組み合わせて SV を検出する既存のパイプライン（GATK-SV、sv-pipeline）との比較を公共データベース（1000 人ゲノムプロジェクト[15]）から取得した 100 の全ゲノムシークエンスデータを用いて行いました。その結果、MOPline の SV 検出精度は GATK-SV のものと同等ながら、真陽性 SV（特に挿入）の検出数（検出感度）に関しては GATK-SV および sv-pipeline を上回っていました（図 2）。

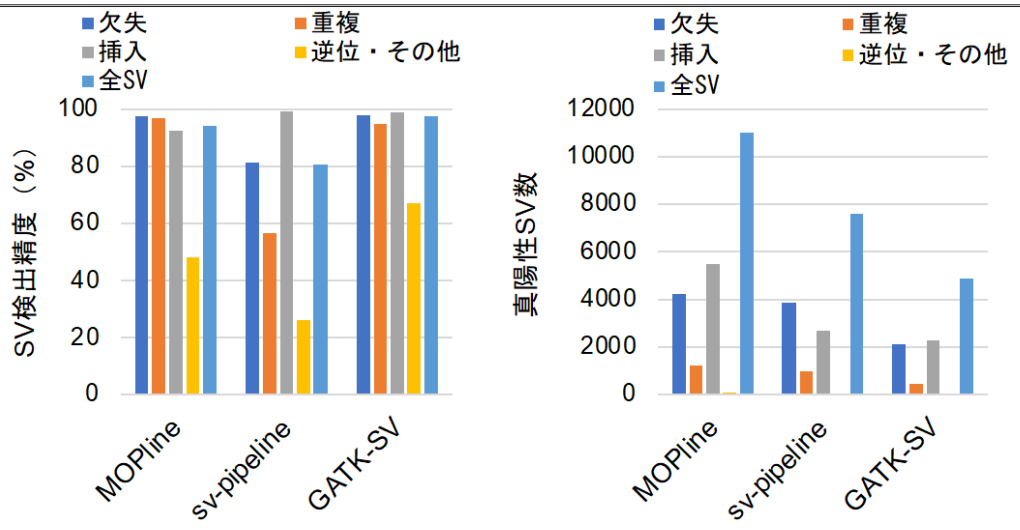


図 2　MOPline と複数ツールを用いる既存パイプラインとの性能比較

100 サンプルの SV を対象として、それぞれの方法の検出精度をロングリードデータを用いて決定した。MOPline の SV 検出精度は逆位以外で約 94%以上を示し、GATK-SV とほぼ同等だった（左）。一方、真陽性検出数については、MOPline が三つの中で最も高かった（右）。

次に、MOPline を用いて 3,258 人のバイオバンク・ジャパン（BBJ）全ゲノムシークエンスデータから SV を検出しました。その結果、約 134,000（個人当たり約 16,000）の SV が検出され、この数はこれまでの大規模 SV 研究プロジェクトで検出された個人当たりの SV 数の 1.7～3.3 倍高いものでした（図 3）。

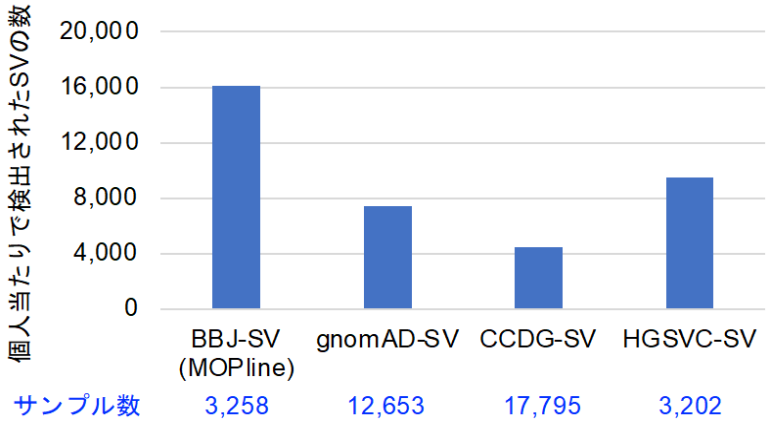


図 3　MOPline を用いて BBJ 全ゲノムシークエンスデータから検出された個人当たりの SV 数

MOPline を用いた場合と他の三つの大規模 SV 研究プロジェクトである gomoAD-SV (Collins *et al.*, *Nature* 2020)、CCDG-SV (Abel *et al.*, *Nature* 2020)、HGSVC-SV (Byrska-Bishop *et al.*, *Cell* 2022)で検出された個人あたりの SV 数の比較。MOPline を用いた場合は、他の三つのプロジェクトの 1.7～3.3 倍に達した。

この BBJ 全ゲノムシークエンスデータは、がんや認知症などのうち少なくとも一つの疾患を持つ患者からのものでした。そこで、疾患に関わる既知遺伝子のタンパク質コーディング領域と重なる SV を調べたところ、いくつかのまれに存在する SV が、疾患サンプルに特異的な既知の疾患リスク遺伝子（大腸がん、乳

がんなど）のタンパク質コード領域と重なっていることが分かりました（表 1）。

疾患	既知リスク遺伝子	遺伝子と重なる SV
大腸がん	*MLH1*	欠失 (1.2 Kb, 109 Kb)
	APC	欠失 (825 Kb, 2,760 Kb)
	MSH2	欠失 (11.2 Kb, 31 Kb)
	NTHL1	重複 (80 Kb)
乳がん	*APC*	重複 (977 Kb)
	MLH1	欠失 (4.6 Kb)
	PPM1D	重複 (179 Kb)
胃がん	*RAF1*	欠失 (12.1 Kb)
認知症	*MEF2C*	欠失 (16 Kb)

表 1 BBJ-SV データに見いだされた既知疾患リスク遺伝子コード領域と重なるまれな SV の例
4 種類の疾患に関わるまれな 12 個の SV が、既知疾患リスク遺伝子のタンパク質コード領域と重なっていた。

MOPline で検出された BBJ-SV（約 134,000）を参照パネル[16]として用い、18 万人の SNP アレイデータ[17]（SNP 遺伝子型データ）のインピュテーション[16]を行い、約 18 万人の SV を類推しました。類推した SV と約 18 万人の医療情報を用い、42 の疾患と 60 の量的形質に対するゲノムワイド関連解析（GWAS）[18]を行いました。その結果、がんなどの疾患を含む 32 形質に関して、SNP と同等もしくはより強い相関を示す 41 の SV が見いだされました。相関のあった SV のうち、8 個の SV は関連遺伝子のコード領域と重なっており、そのうちの 5 個（*MUC22*、*APOC1*、*GYPA/GYPB*、*RP11-219A15*、*FUT2* に重なる SV）は、これまでに該当形質との関連の報告が無い新しく同定された SV でした（表 2）。

疾患・形質	形質と連関	関連する遺伝子
乳がん	欠失 (30 Kb)	*APOBEC3A/B*
背丈	欠失 (1.9 Kb)	*MUC22*
赤血球数	欠失 (4.0 Kb)	*HBA1*
血小板数	欠失 (20 Kb)	*GSTM1*
LDL コレステロール	欠失 (1.6 Kb)	*APOC1*
ヘモグロビン	重複 (119 Kb)	*GYPA/GYPB*
アルブミン/グロブリン比	欠失（23 Kb）	*RP11-219A15*
血中アルカリフォスファターゼ	欠失（24 Kb）	*FUT2*

表 2　ゲノムワイド関連解析（GWAS）で見いだされた疾患・量的形質と相関する SV の例
がんなどの疾患を含む 32 形質について、SNP と同等もしくはそれ以上の強い相関を示す 41 の SV のう

ち、関連遺伝子のコード領域と重なる 8 個を掲載した。乳がんを除く量的形質に関わる SV のうち、ヘモグロビンの重複および血中アルカリフォスファターゼに関わる欠失はマイナスの方向（減少する）に働き、他の形質に関わる SV はプラスの方向（増加する）に作用する。

以上の研究結果は、MOPline はこれまでにない SV 検出精度と検出感度を示すツールであり、単一遺伝子疾患の原因となるまれな SV の同定を可能にするだけでなく、SV のインピュテーションを行うことで複雑な量的形質に関わる SV の同定を可能にすることを示しています。

今後の期待

本研究成果は、これまで主に SNP を用いて行われていた疾患に関わるゲノム解析を、構造多型を含めた解析に拡張させることを可能にします。

また、MOPline は、単一および数千の全ゲノムシークエンスデータを用いた SV 検出を可能とし、ヒトを含む多様な生物種の SV を検出することができます。このことから、幅広い研究分野において、これまで既存のツールでは不可能であった SV の研究が可能となると期待できます。

また、本研究で得られた日本人 3,253 人の BBJ-SV データは、これまでにない高精度な大規模データであり、SV の検証やインピュテーションのための貴重な研究資源として活用されることが期待できます。

論文情報

＜タイトル＞
Detection of trait-associated structural variations using short read sequencing
＜著者名＞
Shunichi Kosugi, Yoichiro Kamatani, Katsutoshi Harada, Kohei Tomizuka, Yukihide Momozawa, Takayuki Morisaki, The Biobank Japan Project, and Chikashi Terao
＜雑誌＞
Cell Genomics
＜DOI＞
10.1016/j.xgen.2023.100328

補足説明

[1] 全ゲノムシークエンス
次世代シークエンス技術または第三世代シークエンス技術を用いて、全ゲノム DNA を鋳型として配列を解読すること。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。構造多型の検出には、ショートリードでゲノム長の 10～30 倍、ロングリードで 10 倍以上の全ゲノムシークエンスデータを要する。

[2] 構造多型（SV）
ゲノムの個人間の違いのうち、50bp 以上の大きさの変異。変異のパターンに応じて、

欠失、挿入、重複、逆位、転座などに分類されるが、これらが混在した複雑なパターンを示す構造多型も存在する。通常、塩基対数の小さい構造多型ほど数が多いが、染色体レベルで起こる大きなサイズの構造多型も存在する。SV は Structural Variation の略。

[3] 遺伝子型
ある遺伝子座で個人が持つ遺伝子変異のこと。どちらかの親から一つの変異を受け継いでいる場合ではヘテロ遺伝子型となり、両親から同じ変異を受け継いでいる場合にはホモ遺伝子型となる。

[4] バイオバンク・ジャパン（BBJ）
日本人集団 27 万人を対象とした生体試料のバイオバンクで、東京大学医科学研究所内に設置されている。理化学研究所が取得した約 20 万人のゲノムデータを保有する。オーダーメイド医療の実現プログラムを通じて実施され、ゲノム DNA や血清サンプルを臨床情報とともに収集し、研究者へのデータ提供や分譲を行っている。

[5] 量的形質
身長や体重など連続的、量的に変化する形質。疾患の有無などの binary 形質と区別される。

[6] 欠失
構造多型のタイプの一つで、ゲノム配列の一部が失われた形態。挿入と並び、構造多型の中では最も多く存在する。

[7] 挿入
構造多型のタイプの一つで、ゲノム配列の特定の位置に別の配列が挿入された形態。挿入配列で最も多くあるものが、内在レトロ因子が挿入されたタイプで、ミトコンドリア配列やウイルスゲノム配列が挿入されたタイプもある。欠失と並び、構造多型の中では最も多く存在する。

[8] 重複
構造多型のタイプの一つで、ゲノム配列の一部の領域が重複して（2 コピー以上）挿入された形態。欠失や挿入に比べて数は少ないが、重複された領域内に遺伝子が含まれる場合、通常と異なる遺伝子発現パターンを示すことが多いため、遺伝子機能の喪失を引き起こす欠失と同様、疾患との関わりが多く報告されている。

[9] 逆位
構造多型のタイプの一つで、ゲノム配列の一部が通常と逆方向に変換されている形態。構造多型のタイプの中で最も数が少ない。

[10] 一塩基多型（SNV）
ゲノムの個人間の違いのうち、塩基配列上の 1 カ所の違い（置換）が一塩基多型と定義される。SNV は Single nucleotide variant の略。SNV のうち、ある集団内での頻度が 1%以上あるものを SNP（Single Nucleotide Polymorphism）と呼ぶ。

[11] リード

DNA の配列決定（シークエンシング）によって得られる DNA 断片の配列情報。次世代シークエンシング技術で得られるリードは、通常 100～200bp のショートリード断片であり、ヒトゲノム解読の場合、数億～10 億本のリードを得る。第三世代シークエンシング技術では、平均 7～10Kb（7,000～10,000bp）のロングリードが得られる。

[12] リファレンス配列
ある生物種のゲノム配列として、標準ゲノム配列として公開されているもの。ヒトでは、hg19 や GRCh37 などの総塩基数約 3Gb（30 億 b）のゲノムリファレンスが公開されている。リファレンスにリードデータをアライメントすることにより、標準リファレンス配列と異なる DNA 多型が検出される。

[13] アライメント
シークエンスリードをリファレンス配列上の合致する位置に対応付けすること。通常、ショートリードは bwa などのアライメントツールを用いてアライメントし、得られたアライメントファイルを用いて構造多型を検出する。

[14] アノテーション
ゲノム上の遺伝子領域などに注釈付けを行うこと。遺伝子がどのゲノム位置にコードされているか、SV などのゲノム変異がどの遺伝子領域と重なっているかなどの注釈付けがある。

[15] 1000 人ゲノムプロジェクト
ヒトゲノムの遺伝的多様性を明らかにすることを目標として開始された国際共同研究プロジェクト。現在では 2,500 人以上のゲノム解析へ移行している。

[16] 参照パネル、インピュテーション
ヒトの場合、DNA マイクロアレイを用いて得られたデータには数十万の SNP 情報が含まれるが、実際には数百万以上の SNP が存在する。アレイに含まれない SNP やインデルを推定する手法がインピュテーションである。通常多くのサンプルの全ゲノムシークエンスデータから得られた SNP 遺伝子型情報を参照パネルとして、その SNP 遺伝子型の並び情報を基に遺伝子型推定を行う。

[17] SNP アレイデータ
SNP（Single Nucleotide Polymorphism）は一塩基多型（SNV）のうち、ある集団内での頻度が 1%以上あるものを指す。SNP アレイデータは、DNA マイクロアレイを用いて得られる SNP 遺伝子型データのこと。全ゲノムシークエンスデータを用いて得られる SNP 遺伝子型よりも少ない遺伝子型情報しか得られないが、全ゲノムシークエンスよりも安価に解析できる。

[18] ゲノムワイド関連解析（GWAS）
疾患などの特定の形質を持った集団と持たない集団との間で SNP 遺伝子型の頻度差の有意性を統計的に評価する手法。ゲノム全域にわたって一つ一つの SNP の遺伝子型を網羅的に調べる。GWAS は Genome-Wide Association Study の略。

共同研究グループ

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　臨床研究部　免疫研究部長、
　　静岡県立大学　薬学部　ゲノム病態解析分野　特任教授）
　　研究員（研究当時）　　小杉俊一　（コスギ・シュンイチ）
　　（現　客員研究員、
　　静岡県立総合病院　リサーチサポートセンター　遺伝研究部　研究員）
　　上級技師　　冨塚耕平　（トミヅカ・コウヘイ）
　　人材派遣　　原田勝利　（ハラダ・カツトシ）
　基盤技術開発研究チーム
　　チームリーダー　　桃沢幸秀　（モモザワ・ユキヒデ）
東京大学大学院　新領域創成科学研究科
　　教授　　鎌谷洋一郎（カマタニ・ヨウイチロウ）
　　（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　客員主管研究員）
　　特任研究員　　森崎隆幸　（モリサキ・タカユキ）

研究支援

本研究は、日本医療研究開発機構（AMED）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（研究開発代表者：寺尾知可史）」、同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（研究開発代表者：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（代表者：寺尾知可史）」、日本学術振興会（JSPS）科学研究費助成事業基盤研究（C）「低カバレッジロングリードを用いた効率的ゲノム構造変異同定手法の確立（研究代表者：小杉俊一）」による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　臨床研究部　免疫研究部長、
　　静岡県立大学　薬学部　ゲノム病態解析講座　特任教授）
　研究員（研究当時）　小杉俊一　（コスギ・シュンイチ）
　（現　客員研究員、
　　静岡県立総合病院　リサーチサポートセンター　遺伝研究部　研究員）

＜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。