様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２６ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０３２３９

研究課題名（和文）植物の巨大なゲノムを解読・解析する手法

研究課題名（英文）Method for sequencing and analyzing huge plant genomes

研究代表者
笠原　雅弘（Kasahara, Masahiro）

東京大学・大学院新領域創成科学研究科・准教授

研究者番号：６０３７６６０５

交付決定額（研究期間全体）：（直接経費）　　13,500,000円

研究成果の概要（和文）：ヒトゲノムの約３倍の巨大なゲノムサイズを持ち、反復配列に非常に富んでいるスギゲノムを対象として、ゲノム配列を（なるべく低いコストと手間で）高連続度・高精度に決定する方法やノウハウを開発した。決定されたスギゲノムの大部分を染色体と高精度に対応させることができ、反復配列に富むゲノム上で遺伝子の同定を行う系統的な手法を開発した。針葉樹において世界最高精度（連続度・遺伝子発見）のゲノム配列決定となった。

研究成果の学術的意義や社会的意義
本成果により、無花粉スギをはじめとして花粉症対策のために必要な研究が大きく加速された。また、スギの経済的価値が将来的に高まるように、経済的に有用な形質を持ったスギ品種をより短期間に開発するための基盤が形成された。また、スギのみならず、他の樹種に対してもゲノム配列の解読と解析をより短期間に低予算で行うことができるようになり、他種に対してもゲノム情報を活用した短期間での品種改良への道を開いた。

研究成果の概要（英文）：We developed methods and know-how for determining the genome sequence of the Japanese cedar genome, whose huge size is approximately three times that of the human genome and is extremely rich in repeat sequences, with high continuity and accuracy at the lowest possible cost and efforts. We were able to accurately correspond the majority of the Japanese cedar genome to eleven chromosomes, and developed a systematic method for identifying genes on genomes rich in repeat sequences. To our knowledge, this became the world's highest precision genome sequence determination (in terms of continuity and gene discovery) for coniferous trees.

研究分野：ゲノム情報科学

キーワード：ゲノム　ゲノムアセンブリ　ロングリード　Ｈｉ－Ｃ　遺伝子　染色体　アルゴリズム　反復配列

１．研究開始当初の背景

PacificBiosciences 社 Sequel シークエンサーや Oxford Nanopore Technologies 社 ProemthION シークエンサーなど、20 kb 以上の長鎖 DNA 配列を読み取ることのできる DNA シークエンサーが登場し、動物のゲノム配列決定は飛躍的に速く高連続度・高精度に、そして楽に行えるようになった。しかし、植物のゲノム配列は動物のゲノム配列ほどには楽に高連続度・高精度のゲノム配列を決定することができなかった。それには主に３つの理由があった。

１つ目の理由として、ゲノム研究の対象となっていた植物種にはゲノムサイズが巨大である種が多かったということがある。ゲノム研究で用いられていた動物は、ハイギョやアホロートル、イモリなどの比較的少数の例外を除いて、たいていはヒトゲノム（3 Gb）と同等かそれ以下のゲノムサイズを持っていた。しかし、植物においてはパンコムギ（17 Gb）やテーダマツ（22 Gb）などをはじめとして巨大なゲノムを持つ種も数多くゲノム研究に用いられていた。

２つ目の理由として、動物のゲノム配列（特に脊椎動物）には植物のゲノム配列に良く見られるような密集した反復配列群が比較的少なく、長鎖 DNA シークエンサーの出力断片配列が散在反復配列を乗り越えて曖昧性無くゲノム配列をアセンブルして決定できる部分が多かった。つまり、高連続度の（コンティグ長が長い）ゲノム配列を得やすかった。特に、近交系が確立されているゲノム研究の歴史の長いモデル動物は、そのゲノム配列がほとんどルーチンと言ってもよいレベルで簡単に決定できるようになっていた。これに対して、植物ではレトロトランスポゾンが連続して密集し、非反復配列（ユニーク配列）が 100 kb 以上にわたって全く現れない領域が多数見受けられた。図１に、例を示す。

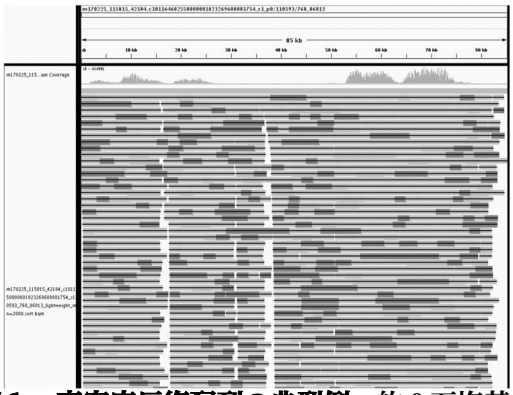

図１．高密度反復配列の典型例。約 9 万塩基対のスギゲノム断片配列（PacBio）にゲノムサイズの約 10 倍量のゲノム断片配列をアラインメントし、IGV（Robinson 2011）で表示した。非反復領域では平均 10 本しか見られないはずのアラインメントが千本以上（誌面の都合で大部分はカットされている）見られ、このゲノム部分配列全体が反復配列で占められていることが分かる。このようなクラスター領域がゲノムの高連続度のアセンブルを難しくしている。

また、１・２の理由から、植物ゲノムのアセンブリに掛かる時間は動物ゲノムと比べて圧倒的に長いことがあり、ゲノムサイズが大きい種についてはゲノムアセンブリに３ヶ月以上掛かった例もあった。我々のグループが申請以前から取り組んでいるスギゲノム解読についても、出力配列の連続度が高いと当時定評のあった Flye アセンブラを用いてアセンブリしても現実的な時間（半年程度）で計算が終わる見込みが極めて薄いことが分かっており、新たな手法の検討・開発が必要であることが示唆されていた。

２．研究の目的

本研究では、ヒトゲノムの２倍以上のゲノムサイズを持ち、大量のＬＴＲレトロトランスポゾンをはじめとした反復配列が含まれる植物種のゲノムから、長鎖 DNA シークエンサーおよび短鎖 DNA シークエンサーを用いて読み取った大量のゲノム断片配列を繋ぎ合わせアセンブルし、元のゲノム配列を高精度・高連続度で推定する方法を開発する。１つ以上の種を選定し、実際に高精度・高連続度のゲノム配列を解読する。本研究により産業上の多くの植物有用種に対して（動物種に対するゲノム解析同様に気軽に）ゲノム解析を新たに可能とすることを目指した。

３．研究の方法

本研究では、本研究では森林総合研究所上野グループ、新潟大学森口グループ、基礎生物学研究所重信グループをはじめとする共同研究グループを結成し、当時約 11 Gbp といわれていたスギゲノム概要配列を高連続度かつ高精度で解読して有用遺伝子の機能解明に資する基盤ゲノム情報を共同で構築することを目的とし、研究開始前までに取得したゲノム断片配列データも含めてデータクレンジングやアセンブリ、エラー補正などを行い。また、得られたゲノム概要配列に対して遺伝学的地図との整合性チェックを行い、精度の検証を行った。更に下流のゲノム解析のための遺伝子アノテーションを行い、反復配列に富む高連続度ゲノムを解析するための問題点やその回避手法を模索した。また、中途より国立遺伝学研究所豊田先生および東京大学鈴木穣先生の研究チームにＤＮＡシークエンシングについて支援を頂いた。

４．研究成果

まずは、初年度までに得られていた様々なシークエンサーから出力されたゲノム断片配列を繋げ合わせ、スギ概要ゲノム配列を構築するためにゲノムアセンブリの計算を行った。スギゲノムのようなヒトゲノムより遙かに大きく反復配列に富んでいるゲノムのアセンブリ手法については業界内において定まった手法が未だ存在していなかった。このため、利用可能な数多くの既存手法を比較検討し問題点を洗い出すところから研究はスタートした。ゲノムアセンブリに用いたソフトウェアは wtdbg2 (red bean)、Canu、Flye、Ra、NECAT、Shasta など、当時利用可能であったロングリードアセンブラーの６種類である。Supernova を用いたゲノムアセンブリは共同研究者が主体となり実施しており弊グループの関与が少ないため本稿では割愛する。

当時試したアセンブラーの多くは、スギゲノムのあまりのデータ量の多さに耐えきれず計算処理を完了することができなかった。用いたスーパーコンピューターのうち、メモリー量が最大のコンピューターは京都大学化学研究所スーパーコンピューターシステムの２ＴＢであり、同システムにおいて許されている最長の計算時間は３ヶ月しかなかった。よって、本稿において「計算が終わらなかった」と言う場合には３ヶ月が経過して計算が強制終了された、ないし計算の途中進行経過を見ることにより、３ヶ月以内に計算が終了する見込みが全く無いと考えられることを意味している。

最初にアセンブリを試みたゲノム断片配列データセットとしては Illumina リード（ゲノムサイズの約 30 倍）、GridION/MinION リード（ゲノムサイズの約 15 倍）、PromethION リード（ゲノムサイズの約 19 倍）、PacBio RS2 リード（ゲノムサイズの約 10 倍）であった。これらのリード配列をプラットフォーム別にアセンブリする、あるいは複数のプラットフォームから得られた配列を混ぜてアセンブリする実験を行ったところ、複数のシークエンシングプラットフォームから得られたリード配列をゲノムアセンブリの初期の段階で統合することは難しく、初期のゲノムアセンブリ処理に於いてはシークエンシングプラットフォーム別にアセンブリ処理を行い、後に統合を図る方が高連続度・高精度のゲノム概要配列を得られると考えられた。当時は必ずしもその理由は分からなかったが、研究期間終了後の後知恵で考えると、ナノポアベースのリード群は植物のゲノムＤＮＡに含まれる修飾塩基を考慮していない機械学習モデルでベースコールされていたため、ポリメラーゼベースで塩基の種類を読み取るシークエンシングプラットフォームとは大きな配列の乖離があったためではないかと推測される。

これらの実験結果を承けて、リード長が平均して最も長くコンティグ長を伸ばすために都合が良いと考えられたナノポアリード（MinION/GridION/PromethION）を基礎的なアセンブリに用いて、ナノポアリード特有のシークエンシングエラーを補正するために Illumina リードを用いて補正することを試みた。塩基補正のソフトウエアについては初年度は Pilon を用いていたが、スギゲノムのサイズと反復配列の多さを考慮すると、動作速度および分割統治的なアルゴリズムを用いていて一定距離毎のユニーク配列の存在を強く仮定しない分性能が質的に良いと考えられた Hypo にスイッチしている。

この時点で得られたスギゲノム概要配列について遺伝学的地図との比較を行い、ミスアセンブリーが十分に少ないことを確認した。また、この段階でスギゲノム断片配列においてレトロトランスポゾンのクラスターを超えるサイズのロングリードの量が不足していることが示唆されたため、PromethION シークエンサーにより追加のゲノムシークエンシングを行った。この新規データのアセンブリは研究終了時までに間に合わなかったが断片配列の統計値について報告する。　まず、Canu アセンブラーはスギゲノムのような反復配列に富んだゲノム配列を想定していないため、昨年度までに計算を行っていたがＮ５０コンティグ長がＮ５０リード長よりも短く総塩基数も 2 Gbp を下回っていたため採用を諦めた。また、Flye アセンブラー（ver 2.6, ver 2.7 の２種類を試している）は、少量のナノポアデータを用いた場合の性能検証では Flye 2.7 のみ良い結果を示したものの、手持ちの全てのナノポアロングリード配列を投入した場合には計算が終わらなかった。Ra アセンブラーはメモリーの要求量が非常に大きく計算を行うことはできなかった。不確かな見積もりではあるが、スギゲノムアセンブリにおいては４０ＴＢ近いメモリーを消費しそうであった。NECAT アセンブラーについても出力総塩基数が 10 Mbp 未満（単位は間違えていない）となるなど不満足な出力であった。Shasta アセンブラーについてはエラーによりアセンブリが完了しなかった。Shasta アセンブラーの根本的なエラーの原因は現時点では不明である。このため、当時は圧倒的優位な結果を出力した wtdbg アセンブラーを用いたゲノムアセンブリについて以下では報告する。

wtdbg アセンブラーは様々なパラメータを入力として取り、パラメータの違いにより出力コンティグの長さや精度が大きく変化する。このため、コンティグ長や精度に大きな影響を与えるパラメータについてはグリッドサーチを行ってより良いパラメータを探索した。コンティグバリデーションに必要なリード本数を指定する A/S オプション、ハッシュテーブル構築のために用いる k-mer のパラメータである e/p オプション、短い入力断片配列をフィルターする長さを指定する L オプションなど、アセンブリ結果に大きく影響を与えると思われるオプションは全て試した。また、ナノポアリードはベースコーラーのバージョンアップによって塩基の精度が大きく上がることが知られているため、全てのデータを当時最新のＧｕｐｐｙバージョン４．０．１４でコールしなおしたリードと、古いベースコーラーを用いた解析を両方とも行った。

驚いたことに、塩基精度が悪いベースコーラーを用いた方が高い連続度のアセンブリを得ら

れた。総塩基数に注目すると高い精度のリードを用いた場合には総塩基数が 11 Gbp を超えており、古いベースコーラーでは 11 Gbp を下回っている程度となっている。このことからは、本ゲノムシークエンシングに用いたスギの樹にヘテロ接合性がある程度残っているため、ハプロタイプ間の違いがある領域が存在していることが考えられた。塩基の精度が悪い場合には２つのハプロタイプを区別することが難しいが、コンティグが伸びて塩基の精度が高くなると２つのハプロタイプが区別しやすくなるために、ハプロタイプあたりのリード量が実質的に半分となってしまうからである。この問題はゲノムアセンブリにおける大きな問題として旧来より知られているが、現状ではデータ量を２倍に増やす以外の簡便な解決策は知られていない。

このため、本研究ではナノポアリードの量を増やすことを目指して PromethION による追加シークエンシングを行った。シークエンシングは東京大学生命データサイエンスセンターに委託して２ラン分（２セル分）のシークエンシングを行い、合計で 154 Gbp （679 万本）の断片ゲノム配列を取得し、Ｎ５０リード長は 38kb であった。この量は従来のスギゲノム PromethION シークエンシングにおいて３ラン分のリード量であり、Oxford Nanopore Technologies 社の技術改良に伴って出力のスループットが上がってきていることが示唆される。

また、この間に現状で最も連続度が高いアセンブリ（wtdbg 2.4 ベース）に対して Illumina リード（HiSeq X Ten ２レーン分）と Hypo を用いて塩基補正を行った。この結果 BUSCO（embryophyta_odb10 データベース利用）による概要ゲノム配列の評価で、Complete 遺伝子が 48.6% から 79.2% に急上昇した。この結果から Hypo により概要ゲノム配列の塩基精度が大幅に高くなったことが確認された。また、この概要ゲノム配列と遺伝学的地図の整合性を調査した。マーカーが５つ以上含まれているコンティグを抜き出したところ、ミスアセンブリは極めて少ないことが示唆された。

その後、国立遺伝学研究所豊田グループの協力により、スギゲノムに対する PacBio HiFi リードを取得し、ゲノムアセンブリを行った。前述の理由により、この段階でもプラットフォーム別のアセンブリを行いその結果を統合する方針を採っていたため、HiFi リードは単独でアセンブルを行っていた。得られたスギゲノムの PacBio HiFi リードには最初の 12+ 塩基程度に、PacBio の公式パイプラインではうまくフィルタリングされないアダプター配列が含まれることがあり、これがアセンブリに悪い影響を与えていることを発見した。最初の 20 塩基をカットすると、アダプターの切り残しによる塩基頻度のバイアスは消失したように見えた。このため、最初の 0、20、30 塩基対をトリミングして 3 セットの HiFi リードを作成し、hifiasm 0.15.5-r350 と Flye 2.8.3 を用いてアセンブルを行った。Hifiasm は 2 日で終了し、最も N50 コンティグサイズが大きく、アダプター切り残しによる塩基頻度バイアスが観察されなかったトリムサイズ=20bp のアセンブリを用いた。最終的なコンティグの N50 サイズは 12Mb で、これまで発表されたどの針葉樹ゲノムのコンティグサイズよりも大幅に大きくなっている。また、コンティグサイズ総計はは 9.1Gbp であり、Genomescope を利用して Illumina リードから推定されたゲノムサイズ約 8.9 Gbp に極めて近かった。

また、NIBB 重信グループに実施された Hi-C シークエンシングの結果を利用し、Juicer と BWA を用いた Hi-C スキャッフォルディングを実施した。リードをコンティグにアライメントし、3D-DNA pipeline を用いて接続の手動修正、コンティグの順序付け、スキャフォールドへの方向付けを行った。その後、Juicebox Assembly Tools 使用して、スキャッフォルドの手動レビューを行い、明らかなミスアセンブリに対していくつかの必要な修正を行った。その後、3D-DNA pipeline を使用して、最終的な染色体長のスキャフォールドを再アセンブリした。その結果、スギゲノムの 11 本の染色体に対応する 11 個のスカフォールドを得て、約 98%の塩基を染色体に関連づけることができた（図２）。

この結果は新潟大学森口グループより提供された既存の 4 つのマッピングファミリー（F107, S1-2, S5HK7, S8HK5）から作成された連鎖地図と照らし合わされた。スキャッフォルドは（Hi-C シークエンシングにより得られる解像度の範囲では）１００%に近い精度で正しいことが推定された。少ない矛盾点についても、マッピングファミリー間を LPmerge プログラムを用いてマージした際の揺らぎやマーカー配列がマルチマップする場合などで説明は可能であった。遺伝地図はゲノムアセンブリとは完全に独立に作成されたものであり、Hi-C スキャフォールドと遺伝地図が高いレベルで一致していることは、大規模な構造的ミスアセンブリを最小限に抑えるという観点から Hi-C スキャフォールドの精度が高いことを証明するものである。マッピングされたマーカーはゲノム全体によく散らばっているので、どんな単一遺伝子座の形質も容易にマッピングされるであろうことが分かった。このため、将来的にスギの有用形質に対応する遺伝子座をマッピングする際に本スギゲノムの配列が極めて有用であることが分かった。注目すべき性質の一つとして、各染色体の中央部に組換えが抑制された長い領域が存在することが分かった。この領域はセントロメアに相当すると考えられる。遺伝子密度はこのセントロメアに向かって減少している。

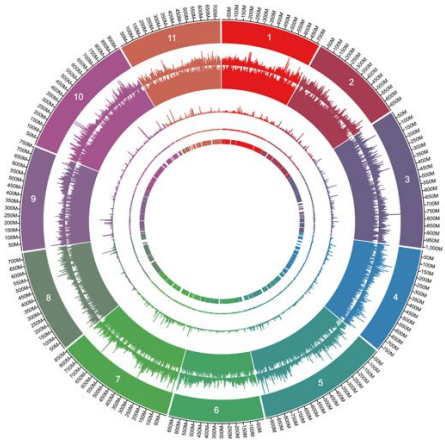

図２：*C. japonica* ゲノムの 11 本の染色体の概要。各リングは外側から染色体、遺伝子密度、繰り返し配列%、N%（コンティグ間のギャップ）、GC%、遺伝マーカーを表している。ゲノム全体で高い割合の繰り返し配列を持っていることが分かる。

また、このゲノム配列に対して RepeatModeler を用いてコンティグ内のリピート要素を同定し、リピートライブラリーを得た後、RepeatMasker によりゲノム中のリピート要素を同定した。最も感度の高いプリセットである -s オプションを使用した。RepeatMasker の検索エンジンは、最も感度の高い cross_match を使用した。他の針葉樹種と同様に、スギゲノムには繰り返し要素が豊富に存在し、ゲノムの少なくとも 83.6%が繰り返し要素と同定された。これらの繰り返し要素はゲノム上に均一に散らばっており、染色体上の位置には観察可能な大きな偏りは見られなかった。

また、スギゲノムに対して Eukan （BRAKER ベースの遺伝子アノテーションパイプライン）を用いて RNA-seq，Iso-seq，cDNA，ホモロジー(OrthoDB ベース)情報、de novo 遺伝子予測などを統合した遺伝子のアノテーションを行った。約１５万の遺伝子構造を得たが、このうち約５．５万遺伝子のみが他植物種の遺伝子と相同性が高いなど遺伝子としてのエビデンスレベルが高いと考えられたため標準遺伝子セットとした。結果、BUSCO による保存遺伝子のベンチマークを用いた遺伝子アノテーションの正確性推定において Complete 遺伝子 91%という、針葉樹では世界で最高の値を得ることができた。この結果はスギゲノムが、ゲノム配列基盤の整備されたモデル生物として将来的な針葉樹研究における重要な役割を果たせるであろうことを示している。

最後に、スギゲノム中の反復配列要素の年代を解析を行った。この解析は、植物ゲノムに多く見られるＬＴＲレトロトランスポゾンの進化の歴史を明らかとし、反復配列を含んだ植物ゲノムをアセンブルする際の指針を得る上で重要であった。我々は、Ma らの論文に従って反復配列の進化解析を行った。EDTA を実行して LTR 要素の位置を特定し、特定した LTR レトロトランスポゾンごとに 2 つの LTR コピーを並べ、２つのコピー間にみられる変異を分子時計として利用し、LTR レトロトランスポゾンの年齢を推定した。また、Redwood および Chinese pine のゲノムに対しても同様の解析を行い、N50 コンティグサイズが１メガベースを超える針葉樹ゲノムの LTR レトロトランスポゾンの年齢分布について我々の知る限り世界で最初の比較を行った。トランスポゾンの進化は種毎にユニークな足跡を辿っていたことが示唆され、また、Copia/Gypsy 両スーパーファミリーの配列についてはファミリー毎に大きく異なる挿入年代を持っていた。また、必ずしもこれらの反復配列は新しいとは限らず、詳細に検討すると古いコピーは区別可能（アセンブリ時に別々の配列であることが分かる）であることが分かった。この知見は更に大きな植物ゲノム配列をアセンブルする際にも訳に立つことが予想される。

以上により、スギゲノム配列を非常に高い連続性（N50 contig: 12.0 Mb）で構築し、Hi-C scaffolding と遺伝地図を用いて、大半（> 97%）の塩基配列を 11 本の染色体の全てに対応付けることができた。また、高密度遺伝子地図と比較することにより、染色体レベルのアセンブリの正しさを検証することができた。これにより、C. japonica の育種のための選抜マーカーを系統的な方法で短期間に開発するための遺伝情報基盤を構築することができた。また、C. japonica ゲノム配列の約 5 万遺伝子をアノテーションし、染色体レベルのアセンブリを行った針葉樹の中で、我々の知る限り最高の BUSCO スコア（91.4%の完全性）を達成した。これらのゲノムリソースは、針葉樹のモデル樹としてのスギ（C. japonica）の位置づけを大きく強化するものである。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Hasegawa Yoichi、Ueno Saneyoshi、Wei Fu-Jin、Matsumoto Asako、Uchiyama Kentaro、Ujino-Ihara Tokuko、Hakamata Tetsuji、Fujino Takeshi、Kasahara Masahiro、Bino Takahiro、Yamaguchi Katsushi、Shigenobu Shuji、Tsumura Yoshihiko、Moriguchi Yoshinari	４．巻 11
２．論文標題 Identification and genetic diversity analysis of a male-sterile gene (MS1) in Japanese cedar (Cryptomeria japonica D. Don)	５．発行年 2021年
３．雑誌名 Scientific Reports	６．最初と最後の頁 1496-1496
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-020-80688-1	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Takeshi Fujino, Katsushi Yamaguchi, Toshiyuki T. Yokoyama, Toshiya Hamanaka, Yoritaka Harazono, Hiroaki Kamada, Wataru Kobayashi, Tokuko Ujino-Ihara, Kentaro Uchiyama, Asako Matsumoto, Ayako Izuno, Yoshihiko Tsumura, Atsushi Toyoda, Shuji Shigenobu, Yoshinari Moriguchi, Saneyoshi Ueno and Masahiro Kasahara
２．発表標題 Chromosome-Level Genome Assembly of Japanese Cedar (Cryptomeria japonica D. Don)
３．学会等名 Plant and Animal Genome Conference 30 (2023)（国際学会）
４．発表年 2023年

１．発表者名 藤野 健，山口 勝司，横山 稔之，濵中 俊哉，原薗 陛正，鎌田 寛彬，小林 航，伊原 徳子，内山 憲太郎，松本 麻子，伊津野 彩子，津村 義彦，豊田 敦，重信 秀治，森口 喜成，上野 真義，笠原 雅弘
２．発表標題 スギ（Cryptomeria japonica D. Don）の全ゲノム配列の決定
３．学会等名 日本森林学会大会
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

本研究成果のデータは DDBJ および ForestGen (https://forestgen.ffpri.go.jp/jp/index.html)(森林総合研究所による公開) から公開している。

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関