2023 年 2 月 6 日 版

総説

T2T 時代の全ゲノム解析を支える DNA シークエンス以外の要素：核 DNA 分子のサイズ計測とクロマチン構造の捕捉

工樂樹洋 [1,2*]、西村理 [1]、門田満隆 [1]

[1] 理化学研究所 生命機能科学研究センター 分子配列比較解析チーム
[2] 国立遺伝学研究所 分子生命史研究室

*責任著者 連絡先: skuraku@nig.ac.jp

要旨

DNA シークエンス技術の進歩により、水産育種や保全の対象となっている多様な生物種についても全ゲノム配列情報の取得が進められるようになり、その情報に基づいて集団遺伝構造を把握することや、実験動物ではアプローチできない生命現象を究める分子レベルの研究基盤を築くことが可能となった。ヒトでは、未決定領域をすべて解決したいわゆる T2T（telomere-to-telomere）グレードのゲノム情報が得られ、他のそれぞれの生物種においても、DNA 配列を完全に読取ることを視野に入れてよい時代が到来した。そのための軸となるのが、ゲノム DNA 分子を材料とするシークエンスやアセンブリであることは紛れもない事実だが、ゲノム情報の完成度について時代の流れとともに高まる要求を満たすためには、抽出した DNA 以外の材料が重要な役割を果たす。配列を染色体規模に繋ぎ上げるためには、細胞核内のクロマチンに含まれる DNA 分子の三次元構造の情報が多用されており、また、最終的に得られた配列情報の検証においては、細胞あたりの核 DNA 量、すなわちゲノムサイズとの照合が必要となるケースも少なくない。本稿では、とくにこの二点についてとくに技術的見地からの意義を議論し、筆者らによる独自の試みを紹介する。

キーワード：ゲノム、ゲノムサイズ、クロマチン、Hi-C、C 値

はじめに

この記事を執筆している 2023 年 1 月、動物の全ゲノムシークエンスに着手するなら、やはりロングリード技術の一択という状況である。具体的には、一分子リアルタイムシークエンス技術で名を馳せた 2 社、すなわち、筆者らの体験ではリード長やデータ量が比較的安定しているように思われる Pacific Biosciences の HiFi シークエンス並びに、ときには 100kb もの長さのリードが得られるという Oxford Nanopore Technologies の技術がこれにあたる。現在でも揺るぎない広い用途に用いられているショートリードシークエンス技術は、150bp などとリード長は短いものの高い精度を誇ってきた。それに対して、ロングリード技術では高いエラー率が常に問題となってきたが、近年これが大幅に改善された。低価格化もあいまって、新規ゲノム情報を取得する際の最有力技術となり、国際学術誌 Nature Methods による Method of the year 2022 にも選ばれた [1)]。

ヒトおよび生命科学の研究で頻繁に用いられる生物種のゲノムアセンブリの概要を図 1 にまとめた。ここに示したヒト以外の種のアセンブリは、現時点で多用されているものではあるが、これらがリリースされたのは、ロングリード手法が必ずしも成熟はしていない少し前の時代である。ゼブラフィッシュでは、800kbp にも上る長さのギャップ（配列未決定領域）を含む 1 万箇所以上の領域が未決定のまま残っている。他の生物種の中でもメダカのゲノムアセンブリ [2)]は、未決定領域が圧倒的に少ないが、ギャップ長が一律 1000bp となっていることがおもな理由である。ヒトでは、CHM13 という Y 染色体を持たない培養細胞を用いて、そのゲノム配列の完全解読がロングリードシークエンスを駆使してついに達成された [3)]。未決定のまま長らく残されていたテロメア、セントロメア、そしてリボソーム DNA（rDNA）遺伝子などの反復度の高い領域の配列決定を完遂しての成果であり、染色体の端から端まで、という意味で「telomere-to-telomere（T2T）」グレードの情報取得と表現された。図 1 内の比較からわかるように、ヒトのゲノムアセンブリであっても、T2T ではないものについては、この最新技術の粋を集めた T2T の完成度には遥かに届いていないというのが現実である。

どの個体から、どういった手法でゲノム DNA を抽出するか、そして、どういった性質のライブラリを調製し、そのうえでどの技術で全ゲノムシークエンスを行うか、ということだけでも悩み始めるときりがない。またさらにどのプログラムで配列のアセンブリを行うか、そして、そのパラメータをどう設定するかについても、検討の余地は大きい。本記事では、既存の多くの文献が詳説している DNA シークエンスやアセンブリの手法 [4), 5)]ではなく、アセンブリから得られたコンティグをどう完成に近づければよいか、そしてそれをどう評価すればよいかという、これまではあまり記事として紹介されていないノウハウに重点をおいて、周辺技術についてまとめてみる。

１．真の「ゲノムサイズ」とは？

ゲノムサイズと聞いて、塩基配列長の合計と思う方は少なくないだろう。しかし、硬骨魚類や海産無脊椎動物のような、ある程度大きく複雑なゲノムを持つ生物種について、ゲノムシークエンスによって取得した DNA 配列の長さの合計が、実際のゲノムサイズを反映することは極めて珍しく、ときに大きくかけ離れた値となることもある。この理由として、目的の種とは異なる種からのコンタミネーションによって、情報取得したいゲノム以外の配列が混入するケースが挙げられる。2 つ目の理由としては、二倍体ゲノムの場合には、父方からと母方からの DNA 配列に違いがある領域をふくむコンティグ配列（haplotig）がアセンブリによって束ねられず、両方が最終的なアセンブリに含まれるということが挙げられる。二倍体の生物個体から、父方由来と母方由来のゲノムを束ねた一倍体ゲノムを取得したい場合には、haplotig の重複を除く処理を適切に行うことが必要となる。最近広く利用されている hifiasm というアセンブリプログラム [6]はこの機能を有するが、筆者らの体験では、haplotig の重複が十分に除去されないケースもあり、この場合には、アセンブリ後に改めて haplotig を除去するためのプログラムをかけることが有効であった。3 つ目の理由としては、DNA が高次構造を形成している、などの理由で、ゲノム DNA の一部が、シークエンスへ持ち込む DNA ライブラリに含まれないというケースが考えられる。また、GC 含量による偏りは存在しないと謳われているシークエンス技術でも、筆者らの経験では、シークエンスされにくいゲノム領域はどうやら存在しており、完全な読み取りを阻む原因になる。アセンブリ配列長がゲノムサイズと一致しにくい 4 つ目の大きな理由は、反復配列の存在である。反復配列は、LINE などの散在反復配列（interspersed repeat）と、マイクロサテライトのような単純反復配列（simple tandem repeat）とに大別される [7]。ともに、反復領域に隣接した固有の配列部分の情報があれば、その反復領域を乗り越えたコンティグを形成することは可能と考えられるため、ロングリードシークエンスを利用すれば、ショートリードを用いた場合ほどはアセンブリに悪影響をおよぼさないと期待される。しかし、生物種によっては、数十 kb を超える散在反復配列も存在すること、そして、単純反復配列が頻出しがちなテロメアやセントロメアはときに数 Mb もの長さを持つことから、現代のロングリード技術を用いたとしても、配列を完全には読み取ることができないというのが現実である。ゲノム中に増幅した異なる反復配列コピーの間の塩基配列の微差と、haplotig 間で対応する DNA 分子の配列の差と、シークエンスの過程で生じるエラーとを、アセンブリプログラムは必ずしも正確に見分けることができないのである。実際に出力されたアセンブリ配列のファイルを開いてみると、まず 1 本目の配列の末端が目に飛び込んでくる。目が眩むような繰り返し配列で占められているとしたら、まさにそれが原因でその配列のさらなる延伸が阻まれた可能性がある。そして、その部分の長さが実際の長さを反映しているか、疑ってみる必要がある、ということである。ちなみに、hifiasm プログラムには、アセンブリを行うプロセスで Hi-C データを読み込んで haplotype phasing を行うための機能があるが、上記の理由に加え、Hi-C

のリンク情報が無い領域では父方・母方由来の染色体を区別することは出来ないため、期待通りに機能するかには大きな疑問がある。

以上の理由から、DNA シーケンスの結果を正当に評価するためには、できるだけ正確なゲノムサイズについての情報を、シークエンスとは別個に取得することが重要である。そのために目的とする生物種の生きた細胞を用意し、DNA の二重らせんにインターカレートするヨウ化プロピジウム（propidium iodide, PI）などで核染色をしたうえで、それが発する蛍光をフローサイトメーターで検出する手法や（イヌザメでの例[8]）、スライドに塗布した細胞の核 DNA をフォイルゲン反応により染色した後に顕微鏡下で染色画像の定量解析を行う手法（フォイルゲン光学密度測定法 Feulgen densitometry[9),10)]）が一般に用いられている。どちらも、ゲノムサイズが既知の生物種の細胞でも同時に測定を行うことで、目的とする生物種のゲノムサイズを相対的に割り出す、というものである。これまでに蓄積されたゲノムサイズの情報は、植物の場合には、Plant DNA C-values Database[11)]（https://cvalues.science.kew.org/）に、そして、動物の場合には Animal Genome Size Database（https://www.genomesize.com/）にて参照可能である。ちなみに、後者のデータベースは、同種について、ときに値の大きく異なる複数のゲノムサイズが収録されていることに注意が必要である。また、エントリ追加の申請を受け付けているにもかかわらず更新が大幅に滞っている。これまで長年にわたり蓄積されこのようなデータベースに収録されたゲノムサイズの記録の多くが、上に記した方法で取得されたものである。しかし、これらの方法には、生きた細胞を材料として要するという点や、並行して測定する参照生物の細胞も必要であるという難点があり、ゲノムサイズの正確な記録は、この手法上の制限のせいで、分類群によってはいまだ大きく限られている。

筆者らは、サメ類とエイ類を含む軟骨魚類を対象として、ゲノム情報を種横断的に解析する Squalomix（スクアロミクス）コンソーシアムの活動を進めている[12)]。対象とする生物種の中には、保護の対象となっている種や、深海性であるために生きたまま実験室に持ち込むのがほぼ不可能な種も少なくない。これらの生物種でも、DNA シークエンスの結果得られるゲノム配列を評価するためのゲノムサイズの情報を、上記の方法の制限に囚われずに取得したいと考えた。筆者らが着目したのは、以前 Wilhelm らによって提案されていた定量 PCR を用いる方法である[13)]。この方法では、ゲノム内に単一コピーとして存在する（言い換えると、複数コピー存在しない）DNA 領域を PCR によって増幅するために必要な絶対 DNA 量を測定する。この測定には、予め分子数の分かっている標準 DNA 分子をターゲット遺伝子毎に用いるが、その標準 DNA 分子や増幅鋳型となるゲノム DNA を正確に計り取ることや、ゲノムの中に存在する数万もの遺伝子の中からの相応しい単一コピー遺伝子の選定、そして、多型によるアレル間の増幅効率に差が無いプライマーの設計など、多様な要因が測定を狂わせる可能性がある。そのため、この定量 PCR を用いた方法には、かつてその信頼性に強い疑問が呈されたこともあった[14)]。とはいえ、組織試料の入手性がえてして問題となる軟骨魚類を対象とする我々にとって、生きた細胞が必要である

という従来法の難点は如何ともしがたい。そこで我々は、正確性が疑問視されつつも生きた細胞を必要としないという優位性が明確な定量 PCR を用いる方法の改善を図ることとした（図 2）。

２．定量 PCR によるゲノムサイズ測定：独自の試み

定量 PCR を用いる方法において、正確な測定を阻むおもな要因は上に述べた。我々は、粘性の高い長鎖 DNA を適度に断片化することに加え、DNA の吸着性の低いチューブやピペットチップ（例、エッペンドルフ社 DNA LoBind チューブ）を使用し、さらに界面活性剤を含む緩衝液を使用することでプラスチック容器表面への DNA の吸着を抑え、一定量の DNA を計り取る精度の向上を図った。測定の肝となる遺伝子の選定においては、事前に得た当該生物のゲノムあるいはトランスクリプトーム配列セットに含まれる複数の単一コピー遺伝子を使用することにより、誤差を考慮した測定結果が得られるようにするという対策を講じた。種間で共通に存在する単一コピー遺伝子の検出には、ゲノムアセンブリの完成度を評価する際に多用される BUSCO というパイプライン [15]を、以前に筆者らが綿密な吟味のうえで選んだ遺伝子セット CVG[16]（core vertebrate genes）と併せて、本目的に活用した。選定した遺伝子を増幅するためのプライマーのデザインにおいては、アレル間で多型塩基が存在する場合に、片側のアレルが増幅対象とならず定量性が乱される可能性を考慮し、プライマー配列内にそういった塩基座位が含まれないようにサンガーシーケンスにより、ターゲット領域の塩基配列を確認している。実際の研究の過程では、ゲノムシークエンスを行う前に、ゲノムサイズをまず把握したいというケースも想定できる。こういった場合には、増幅遺伝子の選定にゲノムアセンブリを使用することはできないが、代わりにトランスクリプトーム配列を利用することでほぼ同様の目的を達成できる。ただし、PCR で増幅するアンプリコン内にイントロンが含まれると、仮にイントロンが長大でないとしても、増幅効率が格段に落ちてしまう。不意にこういった状況に陥らないよう、ゲノム配列をもとにエキソン―イントロン構造が知られている他の生物種のオーソログの遺伝子構造を参考にして、一般に配列長が最も長いとされる最後尾のエキソン内にアンプリコンが含まれるよう、プライマーセットをデザインすることとした。定量 PCR を用いたゲノムサイズ測定のための一連の実験ステップについては、国際学術誌への出版のためにまとめたプロトコルの形で Squalomix コンソーシアム [12]のページ（https://github.com/Squalomix/c-value）にて公開している。

３．核型にどう近づくか？

生物種ごとの染色体の長さや本数は、染色体標本を精査することにより得られる「核型」という情報で記録されている。どういった長さの分布をもつ染色体が何本存在するかは、

DNA シークエンスによるゲノム配列決定手法が普及する以前から、多様な生物種で記録されてきた[17]。DNA 配列に依存せずに測定されたゲノムサイズと同様に、この核型もまた、全ゲノム配列情報を精査するうえで欠かせない情報となる。

ゲノムシークエンスによって得られる DNA 配列が、真のゲノムの姿を反映するわけではないことは上に述べた。真のゲノムの姿になかなか一致しないのは、全ゲノム配列の総塩基長だけでなく、個々の染色体の長さ、そして、染色体の本数についても同様である。塩基の並びのオーバーラップを利用して配列を延伸していくアセンブリ（図 3）のステップは、往々にして反復配列に阻まれ、個々の染色体のサイズにまで繋がるとは限らない。したがって、1 つの染色体の配列が何本にも分断された結果、得られるコンティグ配列の数は、対象とした生物種が本来もつ染色体の数を大きく上回るのが普通である。では、どのようにして染色体規模の配列情報を取得すればよいのか？飼育下で繁殖可能で、大量の次世代個体を得られる生物種を対象としている場合には、連鎖地図を作成することができ、その情報を使って染色体規模にゲノム配列を組み上げることが可能である。いっぽう、それ以外の生物種で染色体 DNA の全体の配列を再構築することは、複雑なゲノムの場合にはほぼ絶望的であった。ここに一石を投じたのが、Hi-C というクロマチンの相互作用に基づくオミクス情報を用いた手法である。

４．Hi-C が可能にした染色体規模のゲノム情報取得

Hi-C とは、染色体配座捕捉法（chromosome conformation capture）の一種であり、細胞の核内でタンパク質と染色体 DNA の高次複合体であるクロマチン内の DNA 領域どうしの相互作用をゲノムワイドに捕捉する方法である。もともとクロマチンの動態を調べるようなエピゲノム解析のために開発されたが[18]、そういった相互作用よりも染色体内の近位の相互作用のほうが圧倒的に高頻度なため、アセンブリによってこれ以上繋がらないというようなコンティグ同士の関係をクロマチンの近位の相互作用の情報に従って再構築することができる。クロマチン相互作用の頻度にもとづいたこのやり方で染色体の配列を組み上げていくと、結果的に他の染色体と区別することができる。一部で「Hi-C karyotyping（Hi-C による核型解析）」という言葉を使っていた研究者もいたようだが、それは言い過ぎであろう。すなわち、Hi-C スキャフォルディングなどの結果得られたアセンブリを評価するための、独立の参照情報としての、従来からの染色体標本の調製に基づく核型解析が必須であることは変わらない。我々の軟骨魚類についての解析においても、とくに重点を置いて調べている生物種については、独自に細胞培養の条件を選び出したうえで、染色体標本の作製に基づく核型解析を行い[19]、ゲノムシークエンスの参照情報としている。

Hi-C データを取得するためのライブラリの調製は、いわゆるオミクスデータ取得のためのライブラリ調製の中で、最も複雑な部類に入り作業工程も長い（図 4）。まず、核内の DNA とタンパク質のクロスリンクを行い、そのままの状態（*in situ*）で制限酵素による

DNA の断片化、DNA 末端へのビオチン化塩基の付加、DNA 末端の再結合（ライゲーション）を行った後、DNA を抽出し、ビオチン化 DNA を回収した後にライブラリ作製を行う。最後にライブラリをペアエンドでショートリードシークエンスすることにより、染色体上は遠くに位置しながらも、クロマチンの立体構造上で近接した DNA 領域の相互作用の情報を取得することができる。塩基配列の重なりに従って DNA 配列を延伸することをアセンブリと呼ぶ一方で、配列間の塩基配列が分からないながらも、間に未決定塩基 N を挿入しながら、DNA 配列の位置関係を整列することにより配列を延伸することをスキャフォルディングと呼ぶ（図 3）。この Hi-C によるスキャフォルディングは、アカデミアの研究室で作られたプロトコルに基づいて調製されたライブラリからのシークエンスデータ取得と、スキャフォルディングのためのオープンソースプログラムとを組み合わせて、2017 年ごろには盛んに利用されるようになった [20)-22)]。それとほぼ同期して、Phase Genomics 社、Dovetail Genomics 社、そして Arima Genomics 社が解析の受託を開始するとともに、キットが販売されはじめた。キットの仕様はその後変更が重ねられてきたが、状況はそこからほとんど変わらず現在に至っている。スキャフォルディングのためのインシリコ部分も含む受託解析の費用は、往々にして、サンプルあたり 100 万円を超える。キットを利用しても、サンプルあたり 10 万円となり、それに加えてさらにシークエンス費用もかかる。必要反応数分を超える単位でのキット購入となる場合には、さらにこの費用は跳ね上がる。

5．Hi-C をより深く知り、より身近に

我々は、こういった費用面の敷居を下げつつも、植物を含む多様なサンプルに対応可能かつ良質なデータの取得を可能にする Hi-C ライブラリ調製のためのプロトコルの吟味を行ってきた。凍結組織が使用可能であることを確認したうえで、爬虫類スッポンに対して、使用する組織のチョイスや制限酵素のチョイス、そしてライブラリ作成ステップの細部を変えた 8 種類のライブラリを調製し、それらにスキャフォルディングを行うインシリコステップでの条件の変更を加えた計 22 通りの条件を比較し、結果を評価した [23)]。この比較検証からは、たとえば、複数の組織を混ぜて使用するよりも、単独の組織を材料とするほうがよい、などの実践的な知見を得ることができた。これらの経験を取り込んだ iconHi-C（アイコニック）プロトコルは、その後も改良を加え、2023 年 1 月時点での最新版バージョン 1.1 を Figshare にて公開している（https://doi.org/10.6084/m9.figshare.14669751.v1）。これまで、このプロトコルを用いて多様な生物種の Hi-C ライブラリを作成してきたが、そのうちマーモセットおよびカニクイザル [24)]、ソメワケササクレヤモリ [25)]、クサフグ [26)]、トラフザメおよびジンベエザメ [27)] について、染色体規模のゲノム情報を公表する運びとなった（図 5）。いずれも、高価なキットを使用することなく、個別に購入した試薬類を組み合わせて、サンプルあたり 5 万円以下の費用でライブラリ調製を行った。高額になりがちな Hi-C ライブラリの本番の大規模シークエンスの前に、ライブラリの質を評価するための方

法、すなわち、制限酵素消化および小規模シークエンスによるライブラリ構造の確認を行うという策を講じた。こうすることで、不適なライブラリを前もって見抜き、良好な質であると判断できたライブラリに絞ってシークエンスに費用を投入することが可能となった。その後、Hi-C から派生して、切断部位の分布に偏りが出がちな制限酵素ではなくエンドヌクレアーゼを用いた手法が導入され、Dovetail Genomics 社から Omni-C というキットが販売されている。

おわりに

本稿では、水産育種の対象となる種についても有用であろうとの期待のもと、おもにゲノムサイズ測定と Hi-C スキャフォルディングの技術に重点をおいて、我々の経験の中でとくに汎用性が高いと思われる話題をまとめた。これから挑むべき課題のほとんどは、既存のシークエンサやコンピュータプログラム自体が解決してくれる範囲を超えたところにある。全ゲノム情報取得のどのステップについても、どれだけ高価で煩雑な手法であれ、技術の原理と細胞生物学の知識に立ち返り、論理的に解きほぐして綿密に検討することが重要である。たとえば、どのような Hi-C データを取得するとゲノム配列のスキャフォルディングに最も効果的か、という疑問に対しては、近位のクロマチンコンタクトのみが高頻度に検出される分裂期（M期）の細胞[28)]がとくに有望であり、遠位のコンタクトが増えがちな血球細胞[29)]を選ぶのは賢明ではないかもしれない。現在、我々の研究室で細胞培養に成功したイヌザメ[18)]で、これを試行中である。こういったチョイスのひとつひとつが、最終的に得られるゲノムアセンブリ配列の精度と費用を大きく左右する可能性がある。

謝辞

本記事に掲載したさまざまな技術的なノウハウにつながった研究室の活動を支えた技術スタッフ種子島千春氏および辰見香織氏に、そして、取り上げた技術についての議論のきっかけをいただいた宇野好宣博士、川口茜博士、山口和晃博士に感謝いたします。

文献

1) Method of the Year 2022: long-read sequencing. (2023) Nature Methods, 20: 1. https://doi.org/10.1038/s41592-022-01759-x
2) Ichikawa, K., S. Tomioka, Y. Suzuki, et al. (12 co-authors) (2017) Centromere evolution and CpG methylation during vertebrate speciation. Nature Communications, 8: 1833.
3) Nurk, S., S. Koren, A. Rhie, *et al*. (99 co-authors). (2022) The complete sequence of a human genome. Science, 376: 44-53.
4) Whibley, A., J.L. Kelley, S.R. Narum (2021) The changing face of genome assemblies: Guidance on achieving high-quality reference genomes. Molecular Ecology Resources,

21: 641–652.

5) 山口 和晃, 工樂 樹洋. （2020）ゲノム情報に支えられたより堅固な生命科学へ：軟骨魚のオプシンを題材として. 比較生理生化学, 37：170-179.
6) Cheng, H., G.T. Concepcion, X. Feng, *et al*. (5 co-authors) (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods, 18: 170-175.
7) 石川冬木、中山潤一 訳.（2018）「ゲノム 第 4 版」メディカルインターナショナル, 東京, pp576.
8) 門田満隆、宇野好宣、工樂樹洋 （2019）脊椎動物のゲノムサイズ（核ゲノム DNA 量）の推定. シスメックスユーザーリポート, https://shorturl.at/oBJT1
9) Michaelson, M. J., H.J. Price, J.R. Ellison, J.S. Johnston (1991) Comparison of plant DNA contents determined by Feulgen microspectrophotometry and laser flow cytometry. American Journal of Botany, 78: 183-188.
10) Hardie, D.C., T.R. Gregory, P.D. Hebert. (2002) From pixels to picograms: a beginners' guide to genome quantification by Feulgen image analysis densitometry. Journal of Histochemistry & Cytochemistry, 50: 735-49.
11) Pellicer, J., I.J. Leitch (2020) The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. New Phytology, 226: 301-305.
12) Nishimura, O., J. Rozewicki, K. Yamaguchi, *et al*. (40 co-authors) (2022) Squalomix: shark and ray genome analysis consortium and its data sharing platform. F1000Research, 11:1077.
13) Wilhelm, J., A. Pingoud, M. Hahn (2003) Real-time PCR-based method for the estimation of genome sizes. Nucleic Acids Research, 31: e56-e56.
14) Gregory, T. R., P. Nathwani, T.R. Bonnett, D.P. Huber (2013) Sizing up arthropod genomes: an evaluation of the impact of environmental variation on genome size estimates by flow cytometry and the use of qPCR as a method of estimation. Genome, 56: 505-510.
15) Manni, M., M.R. Berkeley, M. Seppey, *et al*. (5 co-authors) (2021) BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol, 38: 4647-4654.
16) 原雄一郎 (2016) どのアセンブリを使うか?: 分子系統学的観点に基づくアセンブリの評価. 日本進化学会ニュース , 17: 23-29. https://shorturl.at/ksvDZ
17) Arai, R. (2011) Fish Karyotypes. Berlin, Springer.
18) Lieberman-Aiden, E., N.L. van Berkum, L. Williams, *et al*. (18 co-authors) (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome Science, 326: 289–93.
19) Uno, Y., R. Nozu, I. Kiyatake, *et al*. (8 co-authors) (2020) Cell culture-based karyotyping of orectolobiform sharks for chromosome-scale genome analysis. Communications Biology, 3: 652.
20) Burton, J., A. Adey, R. Patwardhan. *et al*. (6 co-authors) (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature Biotechnology, 31: 1119–1125.

21) Kaplan, N., J. Dekker (2013) High-throughput genome scaffolding from in vivo DNA interaction frequency. Nature Biotechnology, 31: 1143-7.
22) Dudchenko, O., S.S. Batra, A.D. Omer, *et al*. (11 co-authors) (2017) *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science, 356: 92-95.
23) Kadota, M., O. Nishimura, H. Miura, *et al*. (6 co-authors) (2020) Multifaceted Hi-C benchmarking: what makes a difference in chromosome-scale genome scaffolding? Gigascience, 9: giz158.
24) Jayakumar, V., O. Nishimura, M. Kadota, *et al*. (17 co-authors) (2021) Chromosomal-scale *de novo* genome assemblies of cynomolgus macaque and common marmoset. Scientific Data, 8: 159.
25) Yamaguchi, K., M. Kadota, O. Nishimura, *et al*. (6 co-authors) (2021) Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies. Molecular Ecology, 30: 5923-5934.
26) Kabir, A., R. Ieda, S. Hosoya, *et al*. (18 co-authors) (2022) Repeated translocation of a supergene underlying rapid sex chromosome turnover in *Takifugu* pufferfish. Proceedings of National Academy of Sciences, U.S.A., 119: e2121469119.
27) Yamaguchi, K., Y. Uno, M. Kadota, *et al*. (9 co-authors) (2022) Elasmobranch genome sequencing reveals evolutionary trends of vertebrate karyotype organization https://doi.org/10.1101/2022.10.17.512540
28) Naumova, N., M. Imakaev, G. Fudenberg, *et al*. (7 co-authors) (2013) Organization of the mitotic chromosome. Science, 342: 948-53.
29) Ryzhkova, A., A. Taskina, A. Khabarova, *et al*. (5 co-authors) (2021) Erythrocytes 3D genome organization in vertebrates. Scientific Reports, 11: 4414.

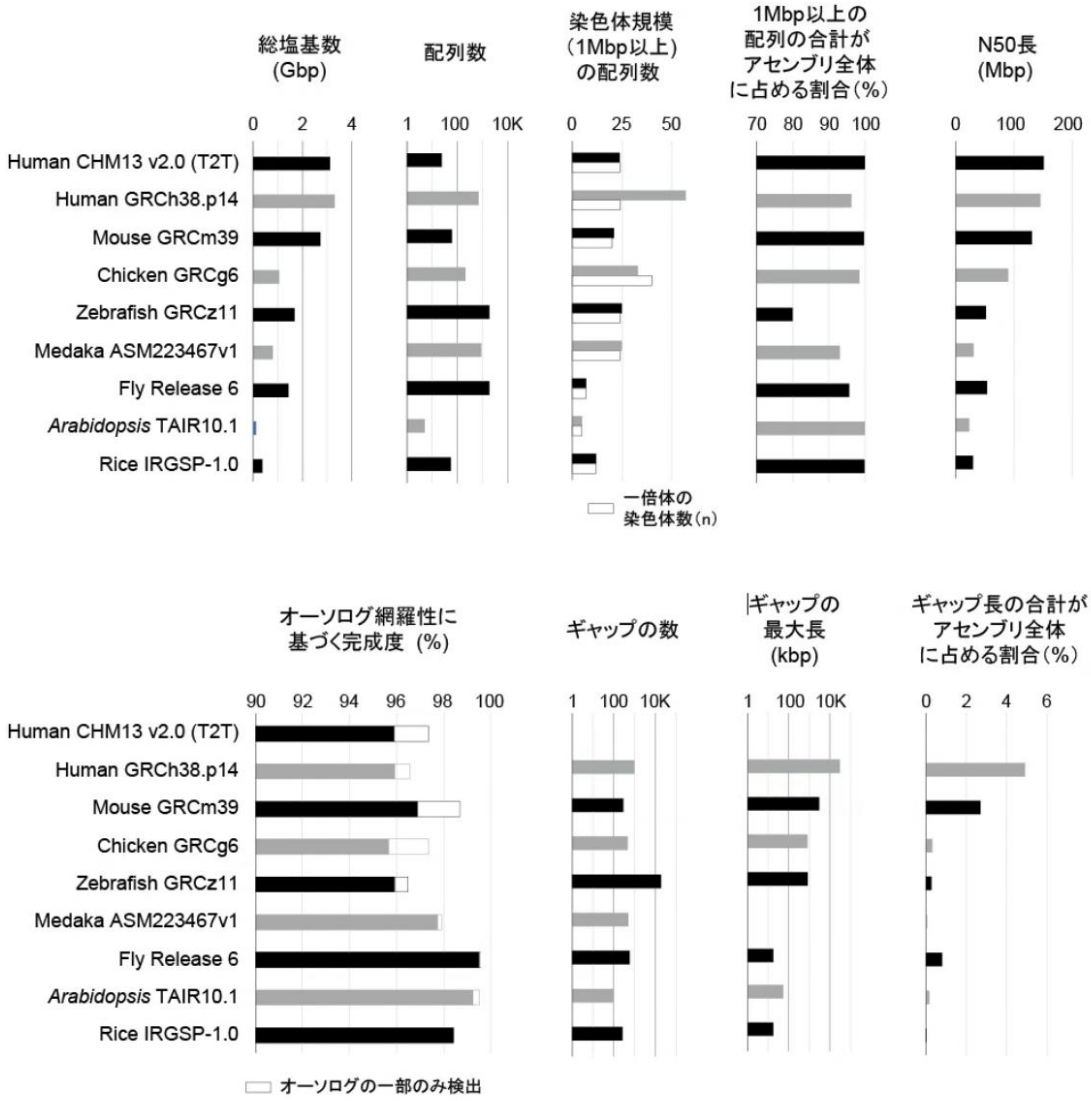

図 1. ヒトや実験生物のゲノムアセンブリの概要。染色体規模の配列数としては、1Mbp 以上の配列数の下に、核型に基づく一倍体の染色体数（白抜き）を添えた。オーソログ網羅性に基づく完成度としては、BUSCO パイプライン（バージョン 5）により個々の分類群ごとの BUSCO オーソログセットを用いた際の「Complete」の値を記載し、併せて「Fragmented」の値をオーソログの一部のみ検出された割合として白抜き部分で表示した。CHM13 v2.0 にはギャップが存在せず、そのためにギャップの最大長や全体に占める割合は記載していない。ヒトゲノム配列が「T2T」グレードであっても完成度が 100%とならないのは BUSCO パイプラインの検出感度が不十分なせいであると考えられる [25]。ヒトの CHM13 細胞は Y 染色体を持たないが、アセンブリ v2.0（NCBI, GCA_009914755.4）には別途読み取られた Y 染色体の配列が含められていることに注意（https://github.com/marbl/CHM13）。

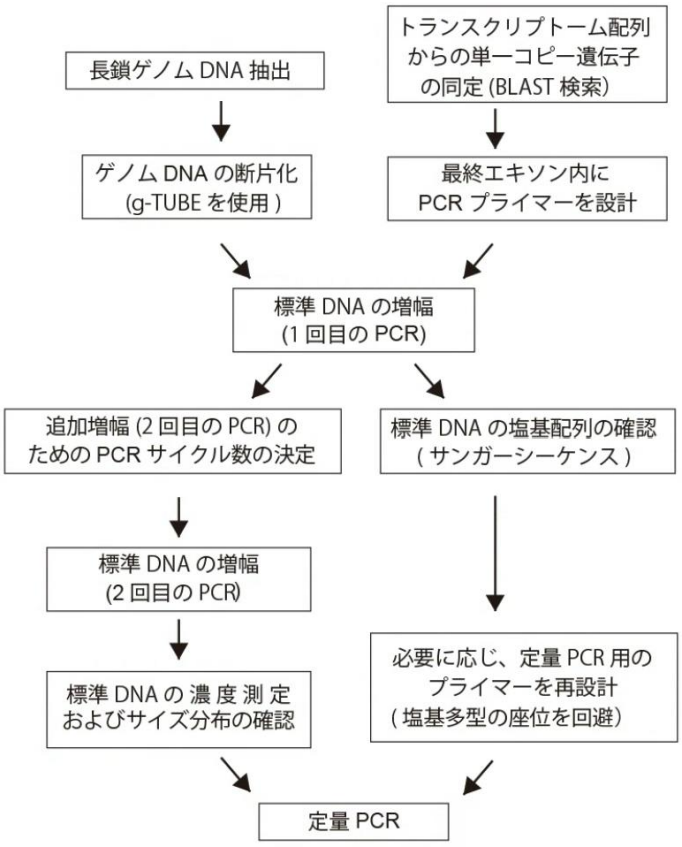

図 2. qPCR に基づくゲノムサイズ測定手法の概要。プロトコルへのアクセスや詳細は本文を参照のこと。

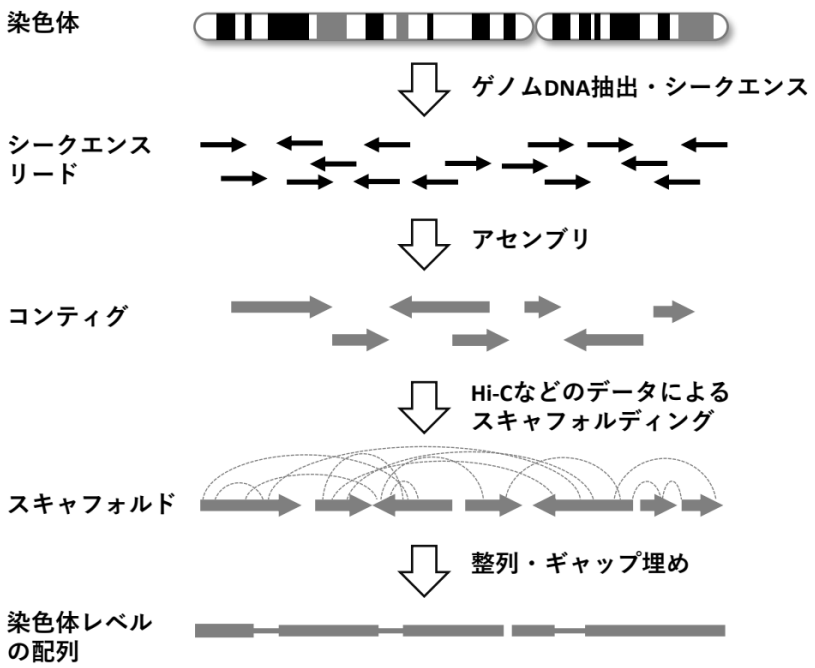

図 3. 典型的な全ゲノムシークエンスの手順。まず、配列の共通部分を頼りに繋げるアセンブリを行うが、そこで得たコンティグは染色体規模にまでは届かないため、Hi-C などの情報を追加してスキャフォルディングを行うステップが必要になるのが通例である。

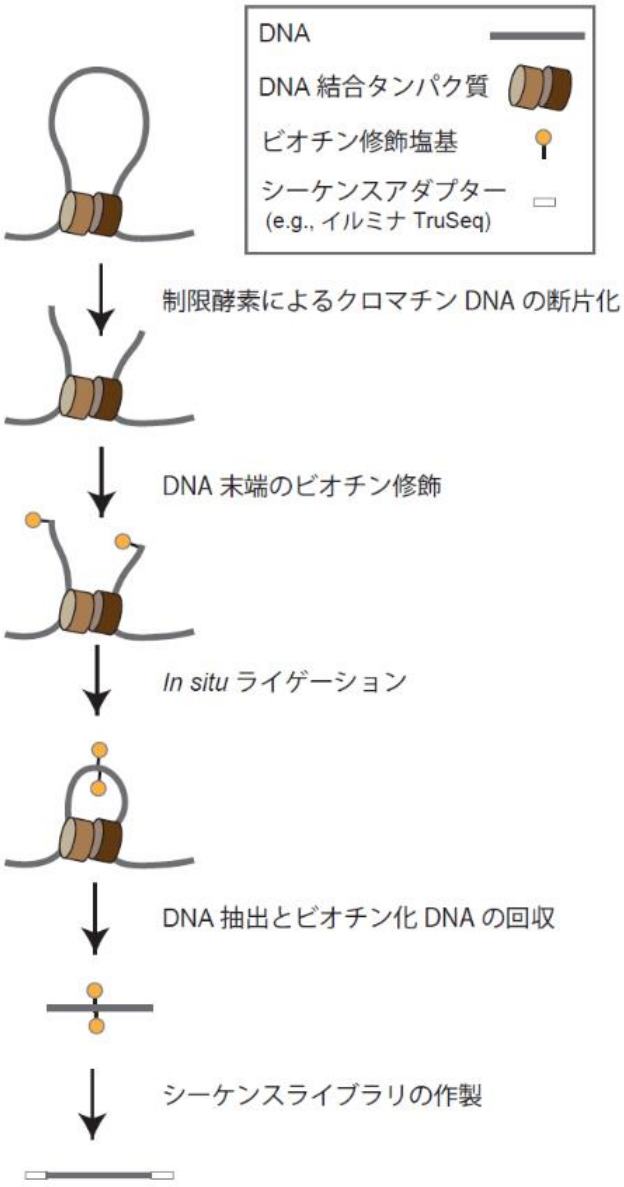

図 4. クロマチン上で近接した DNA 領域を捕捉するために用いられる Hi-C 法のライブラリ作製手順。これらのステップやインシリコデータ処理についての技術的な注意点は、既報等 [23),25)]を参照のこと。

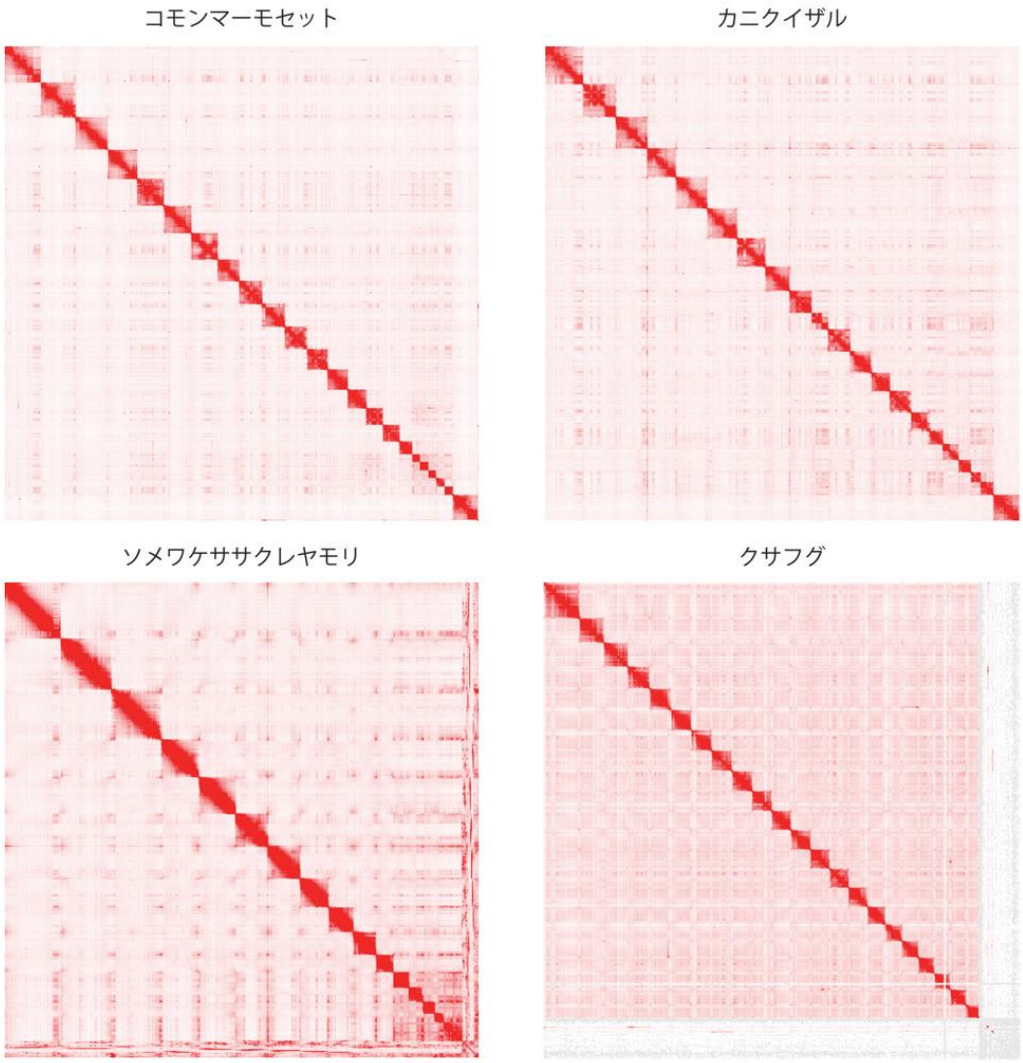

図 5. Hi-C データに基づくクロマチン相互作用のゲノムワイドプロファイル。赤色の濃淡がコンタクトの頻度に対応し、対角線上の正方形が個々の染色体を表す。それぞれの生物についての個別の研究プロジェクトの成果に基づく [24)-27)]。