広島大学学位請求論文

ゲノム配列と核内構造動態の解析による
多様な転写制御機構の解明

Elucidation of diverse transcriptional regulatory mechanisms
by analysis of genome sequences and intra-nuclear structural dynamics

2023 年度

広島大学大学院統合生命科学研究科

統合生命科学専攻 数理生命科学プログラム

D211108 小本 哲史

1. 主論文

ゲノム配列と核内構造動態の解析による多様な転写制御機構の解明

Elucidation of diverse transcriptional regulatory mechanisms by analysis of genome sequences and intra-nuclear structural dynamics

2. 参考論文

Komoto, T., Fujii, M., & Awazu, A. (2022). Epigenetic-structural changes in X chromosomes promote Xic pairing during early differentiation of mouse embryonic stem cells. *Biophysics and Physicobiology*, *19*, e190018.

3. 準備中論文

Komoto, T., Ikeo, K., Yaguchi, S., Yamamoto, T., Sakamoto, N., & Awazu, A. (2023). Assembly of continuous high-resolution draft genome sequence of *Hemicentrotus pulcherrimus* using long-read sequencing. *bioRxiv*, 2023-10.

目次

第Ⅰ章　序論

Ⅰ－ⅰ：DNA の発見とゲノム解析の始まり

ゲノム（genome）とは、1920 年に Winkler によって定義され、遺伝子を意味する“gene”とその総体を意味する“-ome”を組み合わせた言葉であり、「(ある生物が持つ）全ての遺伝情報」を意味するものとされている。当初ゲノムそのものに注目が集まる以前には、1869 年の Miescher による DNA の発見ののち、Avery らの実験［1］や Hershey と Chase の実験［2］から遺伝子の正体が核酸（DNA）であることが明らかになった。そして Watson と Crick による DNA の二重螺旋構造モデル［3］が提唱され、遺伝情報とは何か、核酸とはどういった組成や構造を持つのかが調べられていた。こうして DNA 分子の特徴が知られたのち、次第に DNA の配列（ヌクレオチドの並び方）とそこから生み出される機能についての研究、すなわちゲノム解析が進められていった。1976 年に RNA ウイルスであるバクテリオファージ MS2 のゲノム配列［4］が、1977 年にバクテリオファージ phiX174 のゲノム配列［5, 6］がそれぞれ解読された。さらには 1995 年に細菌で初めてインフルエンザ菌（*Haemophilus influenzae*）のゲノムが解読［7］されたのを皮切りに、1996 年に出芽酵母（*Saccharomyces cerevisiae*）のゲノム配列［8］、1998 年には線虫（*Caenorhabditis elegans*）のゲノム配列［9］、と次第に生物ゲノムも解読されていった。そして哺乳類で初となるマウスのゲノムが解読され［10］、さらには 1990 年からヒトゲノムの完全解読を目標とした「ヒトゲノム計画」も 2003 年に完了した。このようにわずか 10 年にも満たない短い期間でウイルスゲノムに始まり、ついには我々ヒトやマウスといった哺乳類のゲノムまで次々に決定されていった。しかし一方で、ゲノムを解読さえしてしまえば、タンパク質の設計図にあたる遺伝子配列から“生物を生物たらしめる情報”を紐解くことができるという期待とは裏腹に、ゲノム上には遺伝子をコードしない DNA 領域、一時期「ジャンク DNA」とも呼ばれた「非コード領域」も数多く存在することが明らかとなった。

Ⅰ－ii：DNA 配列解読技術の発展

ゲノム解析の潮流が巻き起こる以前から、DNA 配列解読（シーケンス）には主に Sanger シーケンス（ジデオキシ法シーケンスやキャピラリー電気泳動シーケンスとも呼ばれる）が用いられてきた。この技術は 1 度の反応で 1 つの DNA 断片を読むことしかできなかった。そこで生物の全ゲノム配列解析という要望に応えるために、DNA シーケンスを並列、自動化して行うための機器（シーケンサー）が開発されていった。特に Illumina 社のシーケンサーは 2005 年頃から導入され始め、それまでのシーケンサーを遥かに凌駕するスループットを生成できるため“次世代シーケンサー”とも呼ばれ、現在でも広く用いられている。しかし Illumina シーケンサーの唯一の欠点としてシーケンス可能な塩基長が短い（概ね 300 塩基程度）ことが挙げられ、この塩基長を超えるほどの反復配列がゲノム中に存在する場合には、ゲノム配列を再現する上で障害となり得る。そこで 2013 年頃から Pacific Biosciences（PacBio）社や Oxford Nanopore Technologies（ONT）社によって Illumina シーケンサーより長い塩基長で DNA を読み取ることができるシーケンサーが開発された。中でも ONT 社によって開発された MinION は、ナノスケールのタンパク質孔に DNA 分子を通過させることでイオン電流の変化を検知し DNA 配列を推定する手法を採用しており、DNA 調製法にもよるが、長ければ 10 万塩基程度まで一回でシーケンス可能である。しかし Illumina シーケンサーにおける蛍光標識を利用した一塩基伸長の検知に由来する精度の高さに比べて、MinION はシーケンスの精度が低いという欠点がある。そのため、現在でも ONT 社によってハードウェア、ソフトウェア共に定期的に改善されてはいるが、MinION を用いてゲノム配列を決定する場合には、まず Illumina シーケンサーから得られた正確な配列を用いてエラーを修正することから始めるのが一般的とされている。

Ⅰ－iii：ゲノム・エピゲノム解析の進展

DNA 配列解読技術の進展に伴って、遺伝子発現量解析や DNA-タンパク質相互作用、染色体から核内に至るまでのスケールでの DNA の配置や立体構造などに関連する多くの知見がもたらされることとなった。遺伝子発現量解析では、次世代シーケンサー登場以前から用いられていたマイクロアレイと比べると、遺伝子の事前情報を必要とせず転写産物を網羅的に取得し解析するトランスクリプトーム解析をも可能にした。さらに、従来のクロマチン免疫沈降（Chromatin immunoprecipitation, ChIP）法[11] による DNA 結合タンパク質（転写因子等）と相互作用している DNA 配列の抽

出や、Chromosome conformation capture (3C) 法 [12] のような DNA の核内での配置と構造を検出する手法においても、特定の DNA 配列に着目して解析が行われていたのに対して、次世代シーケンサーによる網羅的配列解読と併用することによりそれぞれ ChIP-seq 法 [13]、High-throughput chromosome conformation capture (Hi-C) 法 [14] といった新たな解析へと発展していった。それまでの様々な研究からも、核内における DNA の局所的な構造・エピゲノム状態及び染色体の構造・核内配置は、転写制御や複製タイミング、DNA 修復といったゲノムに関連するイベントに重要である [15-22] ことは既に知られていたが、これらの技術革新により、この生命現象をよりゲノムワイドに、より高解像に観測することが可能となり、細胞種または細胞周期特異的な制御への理解を促進することとなった [23-27]。

前節にも挙げたような近年の目覚ましい技術革新により、多くの研究者が集い、莫大な資金をかけて特定の生物のゲノムを読む時代は終わりを告げ、それぞれの研究者が興味のある生物のゲノムや、ゲノムを取り巻く情報も含めて、比較的気軽に読み解くことができるようになってきている。しかし一方で非モデル生物についてのゲノム解析については (モデル生物と比べて) 未だ発展途上であり、さらにはモデル生物についても膨大に積み上げられたデータに対して理解が追いついていない部分があるのが現状である。

Ⅰ - iv : 本論文の動機

生命の活動を司る遺伝子の転写制御には、遺伝子領域及びその周辺の制御領域を内包する塩基配列に依存した DNA の一次構造、種々の分子の結合や修飾により DNA が三次元的に折り畳まれた局所クロマチン構造、および局所クロマチン構造の生成・消滅により変化する高次染色体構造などの各階層の構造が、重要な役割を果たしている。これらを俯瞰して理解するためにはモデル生物と非モデル生物問わず、また必要に応じて、ウェット、ドライを問わず様々な手法を用いて解析する必要がある。そこで本学位論文ではそのような観点より、雌マウス胚性幹細胞における分化開始に伴う核内構造動態と、バフンウニにおける転写制御を担う特徴的な配列に着目し、多様な転写制御機構について生物種横断的に考察した。

Ⅰ－ｖ: 本論文の構成と各部の概要

第Ⅰ章 序論

ゲノム解析の歴史と、その発展に大きく貢献した次世代シーケンサーに触れ、本研究の動機について紹介した。

第Ⅱ章「マウス胚性幹細胞の初期分化過程における X 染色体のエピゲノム構造変化は Xic 対合を促進する」

雌のマウス胚性幹細胞（Embryonic stem cell : ES 細胞）が分化誘導された際、それに伴って一過的に起こる X 染色体の対合機構を考察した。

マウス ES 細胞からの細胞分化では、分化過程初期に X 染色体不活性化中心（X chromosome inactivation center：Xic）とよばれる領域間の対合が起こる [28-31]。近年この対合の形成が、X 染色体不活性化を制御する遺伝子や、核内構造、代謝、概日時計を制御する遺伝子の発現を制御している可能性が示された [31]。しかしこのような Xic 対合の胚発生中における生理的重要性が明らかになる一方で、染色体で混み合った核内において、Xic および X 染色体同士がどのように互いを認識し接近するのか、その機序は不明であった。

最近、マウス ES 細胞の分化過程の観察より、ヒストン修飾の分布や X 染色体におけるクロマチンドメイン境界の位置などのエピゲノム状態が、分化開始後 2 日で大きく変化することが見出された [23]。また近年の核内染色体構造動態の理論的研究は、染色体サイズや局所的ゲノム領域のエピゲノム状態が、核内染色体や染色体内クロマチンドメインの位置関係に影響を与える可能性を示唆している。実際、エピゲノム状態に依存してクロマチンの物理的特性が局所的に異なるとした、不均一性な高分子鎖で染色体をモデル化することで、シミュレーションにより実験で観察された核内染色体配置が再現された例も報告されている [32- 45]。よってマウス ES 細胞の分化過程における Xic 対合の促進も、X 染色体のエピゲノム状態変化に起因する物理的要因より説明される可能性がある。

そこで本研究では、マウス ES 細胞（ES-cell model）、及び Xic 対合が実際に観察される分化後 2 日目の細胞（2-day cell model）の核内染色体の粗視化分子動力学モデルを構築し、分化に伴う核内構造の変化を考察した。各染色体は Hi-C 法で得られたデータ [23] に基づいた、物理的性質の不均一性をもつ粒子鎖で記述した。ES-cell model と 2-day cell model のシミュレーションから、分化開始後の状態のモデルの方が、2 本の X 染色体が互いに接近する頻度が高くなり、従来報告されている実験結

果が定性的に再現された。さらに、Hi-C データの解析から、未分化の ES 細胞では全ての染色体の表面に非 Open クロマチン領域が広く分布する一方、分化開始後 2 日目の細胞では、X 染色体表面のみで Open クロマチン領域が広く分布することが見出された。

Open クロマチン領域ではクロマチン繊維や結合タンパク質の体積分率が非 Open クロマチン領域よりも小さいため、Open クロマチン領域は柔らかい領域であると考えられる。よって分化後 2 日目では X 染色体が常染色体と比べ顕著に柔らかくなっていると推定された。この推定に基づくと、X 染色体間の排除体積効果は他の染色体間のものよりも弱いため、X 染色体同士が隣接するとその対はコンパクトな形状になり得、その結果他の染色体に大きな空間が核内に提供されることで、系全体のエントロピーが高くなる可能性がある。そこでより簡単化したモデルで系のエントロピーを与える状態数を見積もると、系の全状態数に対する 2 本の X 染色体が隣接している場合の状態数の比が、X 染色体のみが柔らかい分化後2日目の状況で大きくなった。つまり分化に伴うエピゲノム状態変化に起因する X 染色体の物理的柔軟性の変化が、X 染色体の相互接近を駆動し得ることが見出された。

上記の効果は枯渇力とよばれる力 [46-48] や、剛性の異なるポリマーの相分離を引き起こす力 [44, 49-53] と同様のものである。本研究により、これらの力が各染色体内でのエピゲノム状態の変化を通じて調節されることで、核内構造が制御されている可能性が見出された。この知見は、細胞種や細胞周期に依存した動的な核内構造形成の機序を明らかにする、新たな考察基盤を提供する。

第Ⅲ章「ロングリードシーケンスを用いた *Hemicentrotus pulcherrimus* の連続かつ高分解能ドラフトゲノム配列の構築」

広島で採集されたバフンウニ（*Hemicentrotus pulcherrimus*）から単離したゲノム DNA を用い、バフンウニドラフトゲノムの更新を行った。

近年の脊椎動物や昆虫、植物に対する Hi-C 法等による細胞核内構造の解析は、各遺伝子の発現がエンハンサー・プロモーターループ、Topologically associated domain、A/B コンパートメントなど、様々なシスエレメントや高次構造に制御されることを明らかにしてきた [14]。 しかしこのような構造の影響の詳細な解析には、連続性の高いゲノム配列が必要である。一方従来の short-read 解析を用いたゲノムアセンブリでは、ゲノム中に散在する繰り返し配列などの存在により、多くの生物種で遺伝子領域とそのシス制御領域を含む連続性の高いドラフトゲノムを得ることは困難であった。例え

ば古くからの初期発生のモデル生物であり、核内染色体ダイナミクス、左右非相称体軸の確立、中枢神経系の起源の探索 [54-56] など様々な研究がなされているバフンウニのドラフトゲノム（HpulGenome_v1 [57]）でさえ、アリルスルファターゼ（*HpArs*）遺伝子の転写開始点から上流 2kb に位置する *Ars* インスレーターコア配列（ArsInsC）[58, 59] などのシスエレメントや、初期型ヒストンをコードする遺伝子が数十から数百のタンデムリピートを形成しているゲノム領域などが含まれていないなど、多くの課題が残されていた。

そこで本研究では、ONT 社の MinION を用いてバフンウニ精子由来 DNA の long-read 解析を行い、その long-read データと HpulGenome_v1 構築に用いられた short-read データの両方を用いたハイブリッドアセンブリにより、バフンウニドラフトゲノム配列の更新を試みた。その結果 HpulGenome_v1 と比べ、今回得られた配列では Scaffold 数が 16,251 から 2,164 へと、N50 が 143 kb から 516 kb へと、ゲノムとしての完全性（BUSCO score）が 86.1 から 96.5 へと、トランスクリプトームモデルのマッピング率が 55 % から 76 % へと、アノテーションされたモデル遺伝子数が 24,860 から 36,055 へと大幅に改善され、高いゲノム連続性と精度が実現された。また一方 10,000 以上の非アノテーション遺伝子領域も得られた。これらはノンコーディング RNA 由来遺伝子やバフンウニ特異的な遺伝子である可能性が考えられた。更にこのドラフトゲノムには、ArsInsC と 180 以上の ArsInsC 相同配列、及び初期型ヒストン遺伝子の長いタンデムリピートを内包する2つのゲノム領域が含まれていた。

本研究ではこのように、より多くの遺伝子モデルとその制御領域、及び高い配列の連続性を有するバフンウニドラフトゲノムが得られた。ここで初期型ヒストン遺伝子のタンデムリピートを含む領域が2箇所見出されたことは、先行研究でなされたバフンウニの核内蛍光イメージングの結果 [54] より整合しており、発生段階に依存し変化する Histone locus body 動態の研究の進展に貢献すると考えられる。また ArsInsC と相同な配列が 100 以上見出されたことから、脊椎動物ではほぼ CTCF 結合配列に担われていたインスレーター機能が、ウニでは多くのゲノム領域で ArsInsC やその相同配列に担われている可能性も考えられた。さらに今回得られた新規ゲノム配列は、バフンウニにおける遺伝子発現を制御する新規シスエレメントの発見とその機序解明に貢献する可能性も有し、それは広く様々な生物に存在する、多様なシス制御因子の研究を進展させると期待される。

第Ⅱ章　マウス胚性幹細胞の初期分化過程における X 染色体のエピゲノム構造変化は Xic 対合を促進する

Ⅱ－ⅰ：概要

X 染色体不活性化中心（X inactivation center：Xic）の対合は、マウスの雌胚の胚性幹細胞（Embryonic stem cell：ES 細胞）が分化する過程で観察されており、X 染色体の不活性化や概日時計の開始、核内構造の変化、代謝の活性化との関係が示唆されている。しかし、混み合った核内で X 染色体がどのようにして互いを認識し、接近するのか、その機構は明らかになっていない。我々はこの Xic 対合を駆動する機序を解明するため、まず ES 細胞および分化開始後 2 日目の細胞における核内染色体構造の特徴を、実験データの解析に基づき抽出した。その結果、初期分化過程に伴って X 染色体では、表面における Open/Closed クロマチン領域の分布が再構成されることが明らかになった。次にその知見を踏まえ、ES 細胞および分化開始後 2 日目の細胞における核内染色体の粗視化分子動力学モデルを構築した。そして、非 Open クロマチン領域の排除体積効果の方が Open クロマチン領域の排除体積効果よりも強いことなど、自然に仮定できるエピゲノム構造に依存した力学的効果を考慮し、シミュレーションを行った。その結果、X 染色体表面のエピゲノム状態の再構成が Xic 対合を促進することが見出された。これらの結果から、局所的な染色体内自身におけるエピゲノム状態によって、細胞種依存的な核内構造が駆動されることが示唆された。

Ⅱ－ⅱ：背景

X 染色体同士の対合、特に X 染色体不活性化中心（X chromosome inactivation center：Xic）と呼ばれる領域間の対合は、雌マウス胚から得られた胚性幹細胞（Embryonic stem cell：ES 細胞）が分化する過程で一過的に起こる [28–31]。そしてこの Xic の一過的な対合が、X 染色体不活性化（X chromosome inactivation：XCI）開始を誘導するとされる *Xist*/*Tsix* RNA や、*Xist* RNA の発現を活性化させる *Kdm6a* や *Kdm5c* といった遺伝子のバイアレリックな発現と関連することが、近年の研究より示唆されている [31]。さらに、*Lmnb*、*Glut8*、*Per2* など、核内構造、代謝、概日時計を制御する様々な遺伝子の発現量の変化が、Xic 間距離分布の時間的ダイナミクスと相

関していることも明らかになっている [31]。
このように Xic 対合の、胚発生中における XCI をはじめとした生理的プロセスへの重要性は明らかになる一方で、Xic および X 染色体同士がどのように互いを認識し接近するのか、その機構については不明であった。実はこれまでに、X 染色体対合に着目した研究として、XCI の促進や Xic 対合の安定化など X 染色体対合後における過程に対するモデルの提唱は、いくつかなされている [60-66]。しかしこれらが想定しているのは、核内ですでに接近した2体の X 染色体間のおよそ 100 nm 程度の小さな区画で起こる現象である [67]。しかしその前の段階で起こる、染色体で混み合った核内（例えばマウスであれば直径およそ 10 μm、40 本の染色体を含む）におけるより大きな時空間スケールの事象である、X 染色体同士の接近を説明するモデルは、これまで存在しなかった。
近年マウス ES 細胞における分化に伴う XCI の過程を観察した研究で、ヒストン修飾の分布や X 染色体におけるクロマチンドメイン境界の位置などのエピゲノム状態が、分化開始後 2 日目に大きな変化を示し、その後 X 染色体の対合が起こるとの観察結果が報告された [29-31]。またここ 10 年程で盛んとなってきた核内染色体構造動態の理論的研究では、染色体サイズや転写されたゲノム領域の比率といったゲノム自身の状態や、ヒストン修飾や特異的タンパク質の結合などのエピゲノム状態によって、核内染色体や染色体内クロマチンドメインの位置関係が駆動される可能性が示唆されている [23, 32-45]。そして例えば、ゲノムやエピゲノム状態に依存してクロマチンの物理的特性が局所的に異なるとした、不均一な高分子鎖で染色体をモデル化し数値シミュレーションを行うことで、ヒト細胞の間期 [34] やマウス桿体細胞 [35, 36] における転写活性/不活性クロマチン分布、分裂酵母の減数分裂前期の相同染色体の対合 [37] など、実験的に観察された核全体のゲノム構造が再現され、報告されている。以上のことから、マウス ES 細胞の分化過程における Xic 対合の促進も、X 染色体のエピゲノム状態変化に起因する物理的要因より説明される可能性がある。
そこで本研究では、マウス ES 細胞 (ES-cell model)、及び Xic 対合が実際に観察される分化後 2 日目の細胞 (2-day cell model) の核内染色体の粗視化分子動力学モデルを、high-throughput chromosome conformation capture (Hi-C) 法で得られたデータ [23] に基づいて構築し、分化に伴う核内構造の変化を解析した。

II－iii： 結果

II－iii－i：ES 細胞における X 染色体の A/B コンパートメント分布の不安定性と、分化過程での安定化

各染色体のエピゲノム状態を評価するために、ES 細胞および分化後 2 日目の細胞の A/B-コンパートメントプロファイルを、各細胞集団から Hi-C 法により得られた局所ゲノム領域間接触頻度行列データ（Hi-C データ）の 2 つの biological replicate（GSM3127755, GSM3127759, GSM3127756, GSM3127760）とマウスゲノムの局所的な GC 含量を用いて決定した [23]。ここで各 biological replicate において、A-コンパートメントとなる領域を Open クロマチン領域、B-コンパートメントとなる領域を Closed クロマチンと定義した [14]。

その結果、ES 細胞では、常染色体上のほぼ全ての領域が biological replicate 間で同一のコンパートメントに属する傾向が見られた。一方 ES 細胞の X 染色体上では、およそ 40 % 程度の領域が biological replicate 間で異なるコンパートメントに属することが明らかになった（Fig. 1a）。

そこで本研究では、2 つの biological replicate から得られた A/B-コンパートメントプロファイルのうち、共に A-コンパートメントまたは B-コンパートメントとなる領域をそれぞれ A 領域または B 領域と定義し、異なるコンパートメントとなる領域を M 領域と定義した。M 領域は ES 細胞の X 染色体上の広範囲に分布し、X 染色体において最も高い占有率を示した（Fig. 1b）。M 領域はその定義に基づくと、不安定なエピゲノム状態を持つ領域であることから、そのエピゲノム状態は時間的に変化しており、平均的には A 領域と B 領域の中間的な状態を示すと推測できる。

ES 細胞におけるそうした傾向とは対照的に、分化後 2 日目の細胞では、M 領域は X 染色体全体で劇的に減少し、B 領域が増加した（Fig. 1b）。一方で常染色体での A 領域、B 領域、M 領域の染色体内占有率は、X 染色体ほどの変化はなかった。

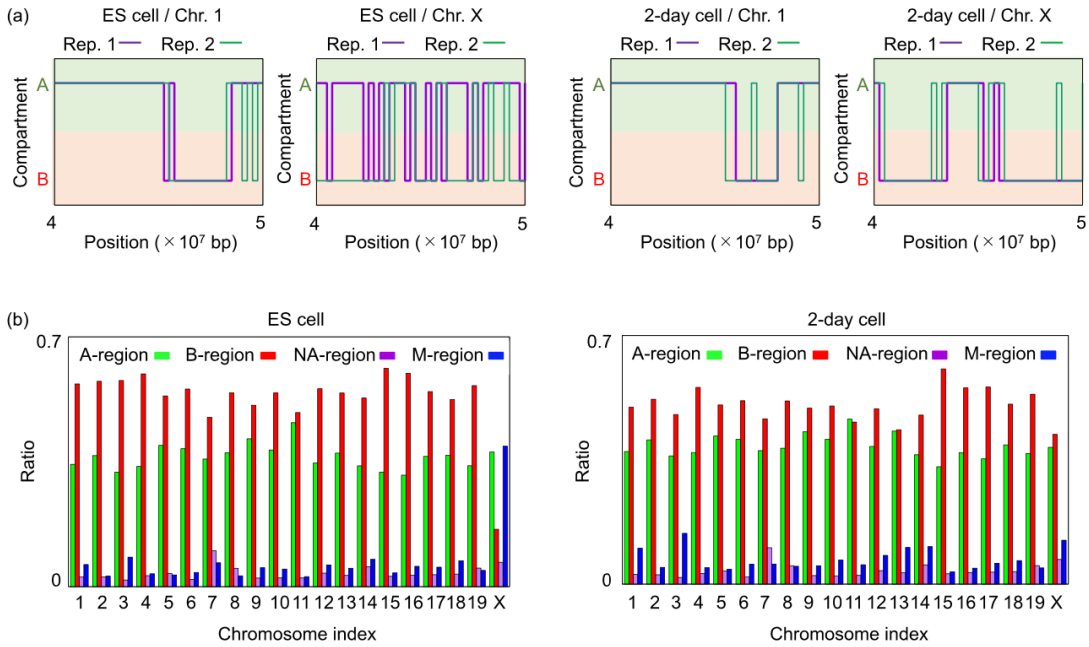

Fig. 1 ES 細胞および分化後 2 日目の細胞における常染色体および X 染色体上の A 領域、B 領域、M 領域のゲノム内分布と占有率。(a) 2 つの biological replicate における A/B-コンパートメントプロファイルと、A 領域、B 領域、M 領域の分布の例 (4×10^7 bp～5×10^7 bp、(1 番染色体および X 染色体))。M 領域は ES 細胞の X 染色体上の広範囲に分布していたが、分化後 2 日目の細胞では減少していた。(b) 各染色体上の A 領域、B 領域、M 領域の占有率。ES 細胞および分化後 2 日目の細胞ともに、常染色体上の M 領域は少ない。X 染色体上の M 領域の占有率は ES 細胞で大きく、分化後 2 日目の細胞では劇的に減少した。NA 領域は Hi-C データのない領域で、セントロメアやテロメアを含む領域に相当すると推定される。

II − iii − ii：分化過程における X 染色体上での Open/非 Open クロマチン空間分布の変化

次に、Hi-C データ［23］を用いて、ES 細胞および分化後 2 日目の細胞における各染色体の A 領域、B 領域、M 領域の空間分布を以下の手順で評価した。

(1) Hi-C データから各染色体の基本立体構造を、近年提案された手法［68-70］の一つであり PASTIS［71］にて実装されている MDS2 法を用いて、ポリマーとして推定した（II − v：手法 の項を参照）。ここで、本研究で用いた Hi-C データの bin 幅は 40 kb であるため、ポリマーを構成するモノマーは 40 kb の局所クロマチン領域に対応すると仮定した。

(2) 染色体ポリマーモデルの重心から各モノマーまでの距離（DCC）の関数として、A 領域、B 領域、M 領域のモノマー数の分布を評価し、それぞれND_A、ND_B、ND_Mとした（Fig. 2a）。ND_α は、$DCC - 0.05\,(\mu m)$ から $DCC + 0.05\,(\mu m)$ の間に存在する α 領域のモノマーの数と定義した。さらに、A 領域、B 領域、M 領域の動径分布をそれぞれ RD_A、RD_B、RD_Mとし、各染色体におけるモノマーの動径方向の確率分布 RPD を DCC の関数として評価した。ここで RD_Xと RPD はそれぞれ

$$RD_X = ND_X/4\pi DCC^2 \ (X = A, B, or\ M)$$

$$RPD = (ND_A + ND_B + ND_M)/4\pi DCC^2 N_C$$

で与えられる（ただし $N_C =$ 各染色体に含まれるモノマーの総数）。

その結果、ES 細胞と分化後 2 日目の細胞の両方において、各染色体の RPD は $DCC \geq 1.25$ において大きく減少する傾向が見られ、全染色体の平均 RPD は有意な減少を示した（Fig. 2b, Fig. S1）。したがって、$DCC \geq 1.25$ に存在するモノマーは各染色体の表面上に存在すると推測された。

そこで、$DCC \geq 1.25$ において、$RD_{B+M-A} = RD_B + RD_M - RD_A$ を DCC の関数として計算（Fig. 2c）し、各染色体表面におけるクロマチンの “closedness”（$CL = \sum_{DCC>1.2} RD_{B+M-A}$）を ES 細胞と分化後 2 日目の細胞でそれぞれ評価した（Fig. 2d）。その結果、分化後 2 日目の細胞における X 染色体のみ負の CL 値を示し、X 染色体表面に Open クロマチン領域が広く分布することが明らかになった。

一方で、ES 細胞の全ての染色体と分化後 2 日目の細胞の常染色体では、正の CL 値を示し、これらの染色体表面に Closed クロマチン領域、あるいは Closed である可能性のあるクロマチン領域（非 Open なクロマチン領域）が広く分布することが明らかになった。この傾向は、PASTIS に実装されている NMDS 法、PM1 法、PM2 法［71］を用いて各染色体の基本立体構造を推定した場合でも、定性的ではあるが全て同一の

傾向を示した（Table S2–S4）。したがって、以下の議論では、MDS2 法で構築した染色体ポリマーモデルを染色体の基本立体構造と定義する。

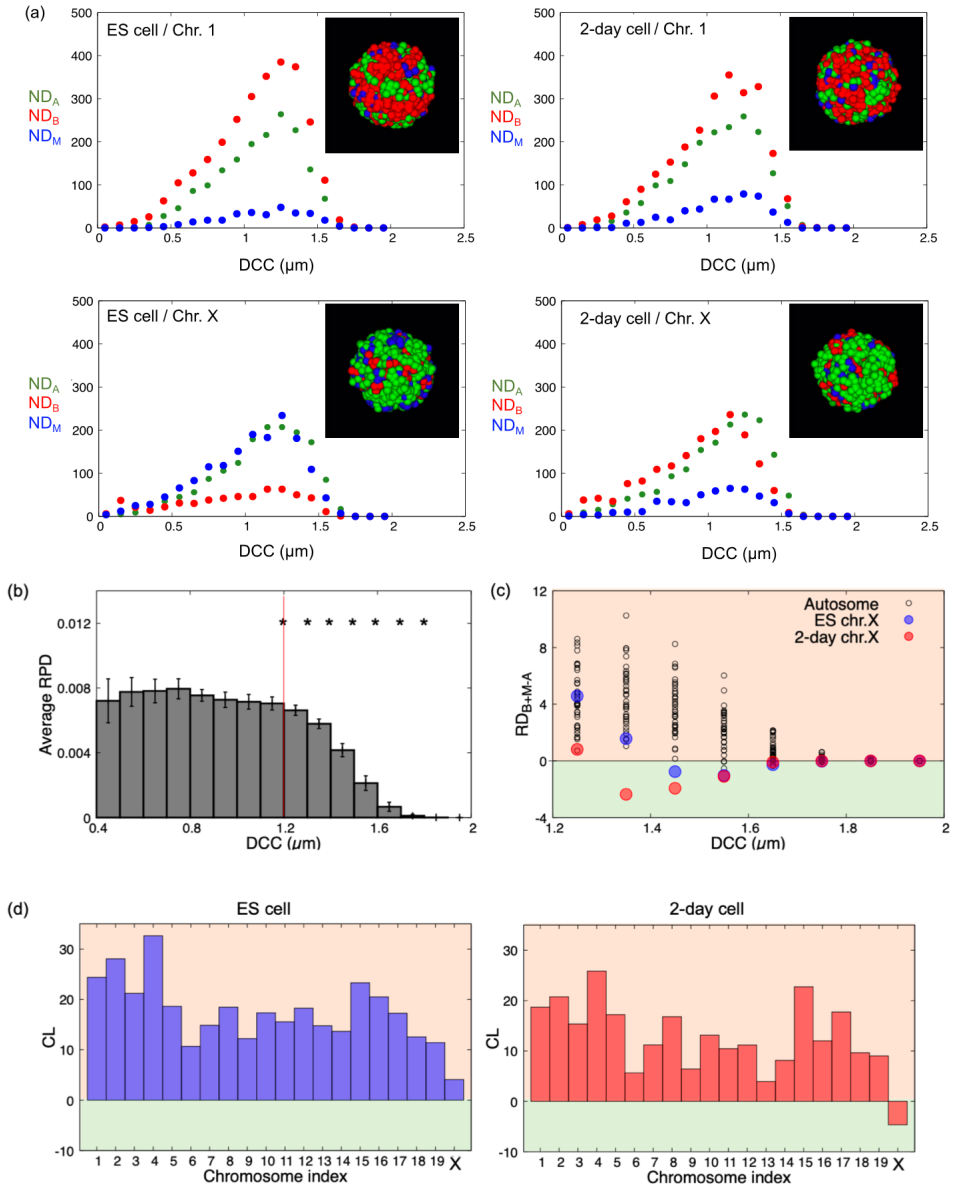

Fig. 2 Open クロマチン領域と Closed クロマチン領域の染色体内空間分布。(a) ES 細胞および分化後 2 日目の細胞における 1 番染色体および X 染色体の ND_A、ND_B、ND_M を示した。挿入図は、ある視点から撮影したスナップショットであり、染色体表面の A 領域（緑）、B 領域（赤）、M 領域（青）の典型的な分布を示す（Table S1 を参照）。(b) 全染色体の平均 RPDとその 95 % 信頼区間(エラーバーで示した)。隣り合う DCC の間の*はそれらの値における 2 つの平均 RPD 値が有意に異なることを示す（Welch の t 検定; 有意水準 α = 0.01、p 値は Fig. S1 に示した)。(c) 各染色体の $DCC \geq 1.25$ における RD_{B+M-A}。(d) 各染色体の CL値。

II－iii－iii：粗視化粒子鎖モデルのシミュレーションで示された分化細胞における Xic 対合

ES 細胞（ES cell model）と分化後 2 日目の細胞（2-day cell model）における核内染色体動態モデルのシミュレーションを、粗視化粒子鎖モデルを用いて行った。

これらのモデルは、染色体の立体構造と A 領域、B 領域、M 領域の分布を表現するモデルとして構築され、以下の特徴を示す。

(1) ポリマーに沿って連続するモノマーの集団は、boundary score [23]（ II－v：手法の項を参照）と染色体ポリマーモデルに沿った適当な染色体の特徴的距離の推定の結果から、約 0.1-10 Mb の局所的なクロマチン領域を含む粒子で表現した。

(2) 各粒子の座標と半径は、粒子を構成するモノマーの重心座標と、粒子の重心から各モノマーまでの距離の標準偏差で定義した。その結果、粒子の半径は約 0.15-1 μm であった。このときの典型的な粒子サイズのスケールは、各染色体をドメインからなる鎖として記述した際に得られる染色体の持続長のスケールと同等であった。

(3) 各粒子に含まれるモノマーのエピゲノム状態に依存して、各粒子のエピゲノム状態を規定した（Fig. 3a,b）。（ II－v：手法 の項を参照）

上記の染色体粒子鎖モデルを用いて、粒子鎖の基本構造を維持するための力と、排除体積効果によって粒子に働く反発力を考慮し、核膜を半径 5 μm の球殻としたシミュレーションを行った（Fig. 3c,d）。ES cell model と 2-day cell model の両方について、異なる初期配置から 30 回のシミュレーションを行い、Xic 領域を含む 2 つの粒子間距離の時間経過と確率分布を描画した（Fig. S2、Fig 4a）。その結果、2-day cell model では ES cell model と比較して、Xic 粒子間距離が小さい状態が起こりやすく（有意に高い確率値が得られる）、Xic 粒子間距離が大きい場合が起こりにくい（有意に小さい確率値が得られる）という傾向が見られた。この結果は、2-day cell model における Xic 粒子は、ES cell model よりも有意に接近しやすいを示しており、従来報告されている実験結果 [28-31] が定性的に再現された。相同 X 染色体重心間距離の確率分布も、同様の傾向を示した（Fig. 4b）。一方、4 番と 19 番を除くほとんどの相同な常染色体間では、重心間距離の確率分布は ES cell model と 2-day cell model の間で有意な変化を示さなかった（Fig. 4c、Fig. S3）。

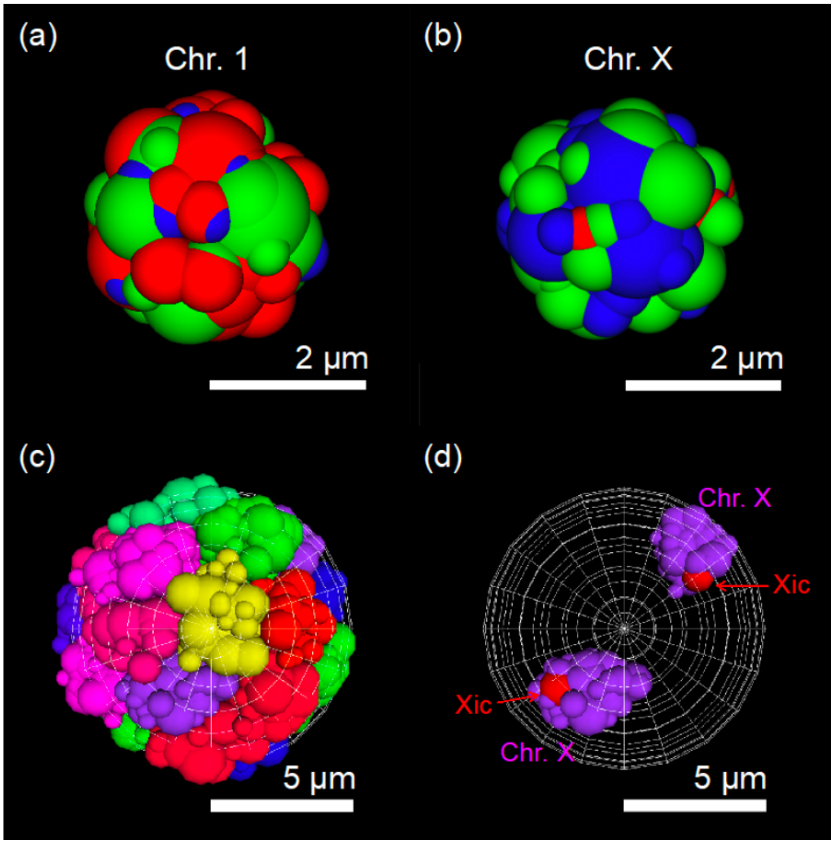

Fig. 3 粗視化粒子鎖モデルを用いたシミュレーションのスナップショット。(a,b) ES cell model における 1 番染色体 (a) と X 染色体 (b) の粒子鎖モデルのスナップショットの例 (ES cell model と 2-day cell model における他の染色体の結果は Table S5 を参照)。緑、赤、青の球はそれぞれ A 粒子、B 粒子、M 粒子を示す。(c,d) ES cell model の全染色体のスナップショット (c) と 2 本の X 染色体および Xic 粒子の位置関係 (d)。色の異なる粒子鎖はそれぞれ各染色体の粒子鎖モデルに対応する。

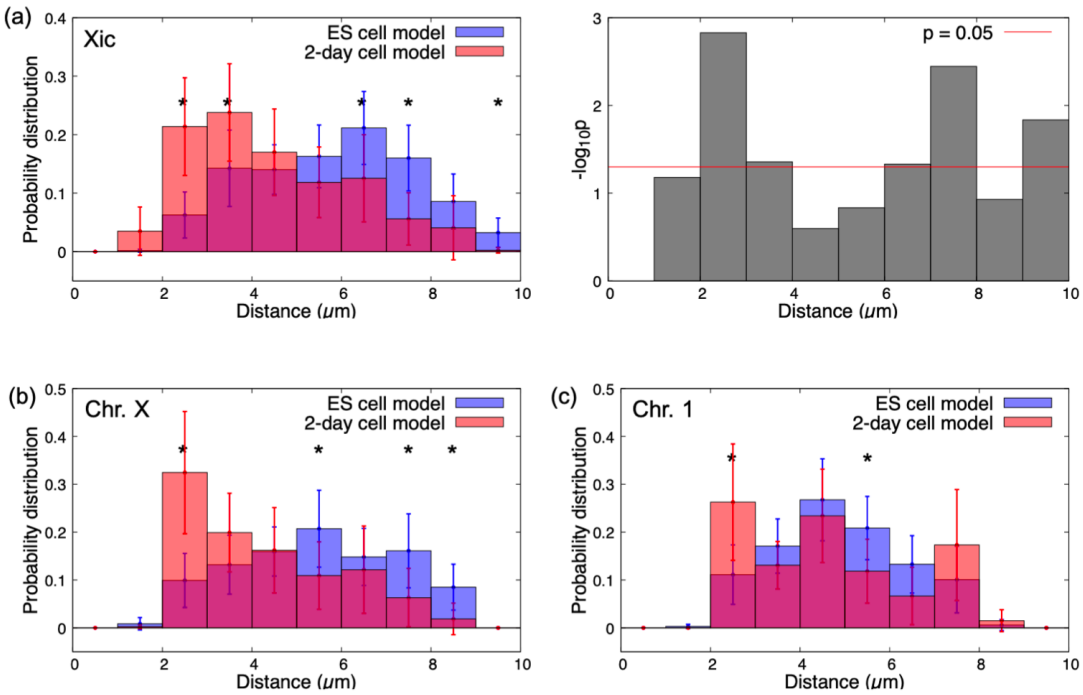

Fig. 4 Xic 粒子間および相同染色体重心間距離の確率分布と各距離区間における Welch の t 検定から得られた p 値。(a) ES cell model と 2-day cell model における Xic 粒子間距離の確率分布の平均と 95%信頼区間（エラーバーで示した）(左)、各距離区間における 2 つのモデル間での確率値の Welch の t 検定から得られた p 値(右)を描画した。(b,c) (a) の左図と同じ条件で、相同 X 染色体重心間距離 (b) と相同 1 番染色体重心間距離 (c) の確率分布を描画した。平均と 95%信頼区間は、各モデルにおける 30 回分のシミュレーション結果を用いて評価した。* の付いた距離では、2 つのモデル間の確率値の差が有意であることを示す（Welch の t 検定; 有意水準 α = 0.05; Fig. S3 と Fig. S4 も参照）。

II – iv : まとめと考察

核内染色体動態の粗視化モデルのシミュレーションから、ES cell model よりも 2-day cell model の方が、X 染色体対合が起こりやすくなることが見出された。この結果は、近年報告されている実験 [28-31] と定性的に一致した。そこで、このような核内動態を促進する駆動力について、染色体内構造とそのエピゲノム状態から推定される力学的性質に基づいて考察を行う。

まず Hi-C データを用いた解析の結果、分化過程に伴うエピゲノム状態は以下のような特徴を示した。

(1) ES 細胞の X 染色体上でのみ、不安定なエピゲノム状態を持つクロマチン領域が広範囲に分布していた。これらの不安定なエピゲノム状態を持つ領域は、潜在的な Closed クロマチン領域であるとも考えられる。また、X 染色体表面でのみ、Closed クロマチン領域および潜在的な Closed クロマチン領域（非 Open クロマチン領域）が、常染色体と比べて高い占有率を示した。

(2) 分化後 2 日目の細胞では、すべての染色体において、安定なエピゲノム状態を持つクロマチン領域がほとんどの領域を占めていた。さらに、ES 細胞では X 染色体表面に Open クロマチン領域と非 Open クロマチン領域の両方が存在したが、分化後 2 日目の細胞では X 染色体表面において Open クロマチン領域が常染色体と比べて高い占有率を示した。

(3) 常染色体では、分化過程で X 染色体のような染色体内構造におけるエピゲノム状態はほとんど変化しなかった。特に常染色体表面は Closed クロマチン領域が高い占有率を示した。

ここで、Open クロマチン領域のクロマチン繊維や結合タンパク質の体積分率は、非 Open クロマチン領域よりも小さいため、Open クロマチン領域は柔らかい領域であると考えられる。よって分化後 2 日目の細胞において X 染色体は常染色体と比べて柔らかくなっていると推定された。この推定に基づくと、X 染色体間の排除体積効果は、他の染色体間のものよりも弱いため、隣接する X 染色体同士が隣接するとその対は他のどの染色体の組よりも最もコンパクトな形状となり得る。その結果、核内において他の染色体により大きな空間が提供されることで、他の染色体が X 染色体の間に割り込むように位置している場合に比べて、2 本の X 染色体同士が隣接している場合では系全体のエントロピーがより大きくなると考えられる。そこで、柔らかいブロックと硬いブロックからなる以下の 1 次元格子ブロックモデル（Fig. 5a）を用いて、系のエントロピーを与える状態数を見積もることで説明を試みた。

(1) 端壁を持つ 9 マスの 1 次元格子内に 2 マス分の大きさのブロックを 4 個配置する。そのうち 2 個のブロックを黄色、2 個のブロックを緑色で示した。全てのブロックが硬く互いに重なることができないとしたモデルと、2 個の柔らかいブロック（緑色）のみ互いに 1 マスずつ重なることができるとしたモデル（Fig. 5a）のそれぞれで系の状態数を求めた。

(2) (1) のそれぞれのモデルにおいて、系の全状態に対する柔らかいブロックが隣接している場合の状態数の比を求めた。

その結果、全てのブロックが硬く、互いに重なることができないとしたモデルよりも、2 個の柔らかいブロックのみ互いに 1 マスずつ重なることができるとしたモデルにおいて、系の全状態に対する柔らかいブロックが隣接している場合の状態数の比が大きくなった（Fig. 5b,c）。今回は少数のブロックからなる 1 次元空間でのモデルを用いて状態数を比較したが、多数のブロックからなる 3 次元空間でのモデルへと拡張した場合でも、同様の議論が可能である。

したがって、本研究で扱った核内構造動態モデルについても同様に、X 染色体のみが柔らかい分化後 2 日目の細胞では、系の全状態数に対する 2 本の X 染色体が隣接している場合の状態数の比が大きくなると推定できる。つまりこのような状態数（エントロピー）の違いが引き起こす効果が、枯渇力として知られる力 [46-48] や、剛性の異なるポリマーの相分離を引き起こす力 [44, 49-53] と同様に、X 染色体および Xic の相互接近の駆動力となり得ると考えられる。逆に、ES 細胞における X 染色体間の排除体積効果は、他の染色体間ほど強くないが、分化後 2 日目の細胞における X 染色体間の排除体積効果と比べると十分に強いと言える。よって ES 細胞では分化後 2 日目の細胞と比べて、X 染色体同士が接近した状態の数が全体の状態の数に比べ多くならず、つまり相互接近が起こりにくくなっていると結論づけられる。

近年の Falk らの研究では、多価のヘテロクロマチン結合タンパク質を介した B-コンパートメント領域間の引力的相互作用が、染色体内および染色体間の配置を決定することが示唆されている [72] が、本モデルではそのようなタンパク質の効果は導入しなかった。本モデルにおける斥力的相互作用に加えて、このような引力的相互作用を与えれば、さらなる常染色体の凝集と、常染色体と X 染色体の分離が促進され、2-day cell model における X 染色体対合が起きやすくなると考えられる。

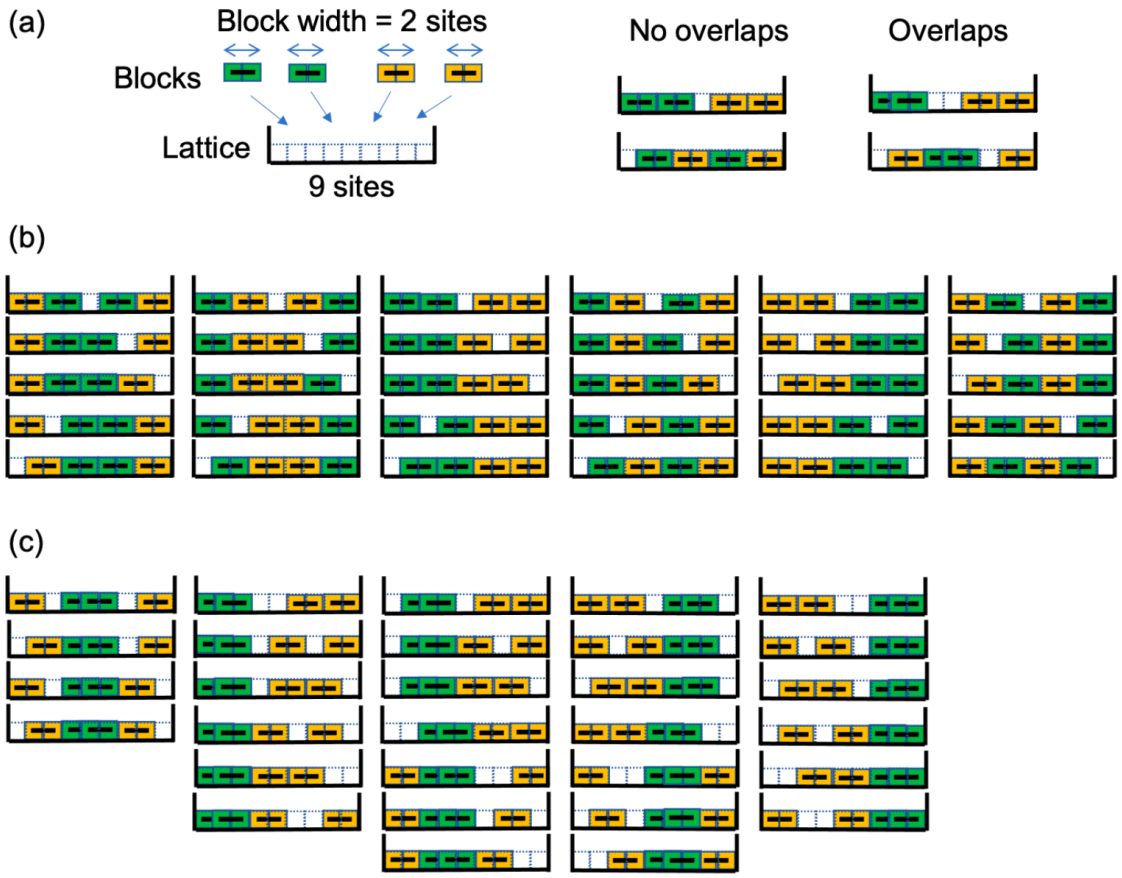

Fig. 5 2-days cell model において X 染色体が相互接近を示すメカニズムを直感的に理解するための、柔らかいブロックと硬いブロックからなる単純な 1 次元格子ブロックモデルを図示した。(a) 端壁を持つ 9 マスの 1 次元格子、2 つの黄色ブロック、2 つの緑色ブロックからなるモデル。各ブロックの幅＝2 マス、2 つの柔らかいブロックは互いに半分ずつ（1 マス）重なることができると仮定した。(b) 全てのブロックを硬いブロックと仮定した場合の全状態の図。全状態の数は 30 であり、緑色のブロック同士が接触する（重なる）状態の数は 12 であった。よって、緑色のブロック同士が接触する（重なる）確率は 12/30 = 0.4 となる。(c) 緑色ブロックと黄色ブロックをそれぞれ柔らかいブロックと硬いブロックとし、2 つの緑色ブロックが重なった場合の全状態の説明図。(b) と (c) の場合における状態の組み合わせから、2 つの緑色ブロックが重なることによって増加するポテンシャルエネルギーを U とした場合に、2 つの緑色ブロックが接触する、あるいは重なる場合の確率は $(12 + 30e^{-U/k_BT})/(30 + 30e^{-U/k_BT}) > 0.4$ となる。(k_B はボルツマン定数、T は絶対温度を示す。) この確率は $U \to 0$ の場合に $42/60 = 0.7$ に近づく。これらの事実から、2-day cell model の X 染色体のように 2 つの緑色ブロックが柔らかい場合、ES cell model の X 染色体のように緑色ブロックが硬い場合よりも、それらの接近が頻繁に起こることが示唆された。

また本研究では、*Xist* / *Tsix* RNA のような特定の分子を介した Xic 対の局所的な短距離相互作用に先行して起こる X 染色体の対合をもたらす、核内染色体動態の駆動力を探索した。そのため、これらの長鎖非コード RNA による染色体動態への影響は考慮していない。また、TAD や A/B-コンパートメントに比べて粒子を構成する領域のスケールが大きいため、Xic 周辺の局所的なクロマチン領域の詳細な構造動態は、今回のモデルでは明らかにできなかった。今回のモデルと、最近提案された Xic 相互作用の他のモデル [60-66] を組み合わせることで、Xic 対合とその安定化に関わる全ての動態を説明できるモデルが開発できるかもしれない。

さらに最近の実験から、Xic はラミン結合レセプター（Lbrs）と *Xist* RNA の相互作用によって核辺縁部に局在する傾向があり、*Xist* RNA の発現を阻害するとその傾向が弱まることが示されている [73]。今回のモデルではこのような RNA の影響を考慮していないため、実験から示されているような Xic の核辺縁部への局在は確認できなかった（Fig. S5）。今後の研究では、ES 細胞の初期分化過程における Xic の挙動全体を考慮するために、Lbrs-*Xist* 間相互作用と *Xist*-*Tsix* 駆動の Xic 間相互作用のような非自明な効果を含む修正モデルを開発する必要がある。

本研究では、ES 細胞および分化後 2 日目の細胞の Hi-C データと、その 2 つの biological replicate を解析した。ES 細胞の常染色体および分化後 2 日目の細胞の全染色体では、2 つの biological replicate 間における A/B-コンパートメントプロファイルがほとんどの領域でよく一致したことから、これらの染色体のエピゲノム状態は安定であると考えられる。一方、ES 細胞における X 染色体では、2 つの biological replicate 間における A/B-コンパートメントプロファイルがおよそ 40 % 程度の領域で一致しなかったことから、不安定なエピゲノム状態を持つと考えられる。よって X 染色体では、A/B-コンパートメントプロファイルが時間的に変化し、細胞に依存した大きな変動を示し、Hi-C データの biological replicate が増えるにつれて、ES 細胞の X 染色体では M または M 様領域がより多くの領域において観察される可能性がある。この推測は今後の研究によって検証する必要がある。

本研究における Hi-C データ解析の結果、分化後 2 日目の細胞では、X 染色体を除く染色体表面には主に非 Open クロマチン領域が分布することが示された。

この結果は、マウス胎児肝臓 [74] の様々な細胞を用いた顕微鏡観察結果である、Open クロマチン領域が染色体表面へと集積する傾向と矛盾するように思われる。この染色体内構造の差異は、ES 細胞や分化初期の細胞と、組織内における細胞の分化段階の違いに起因することが考えられる。この推論を検証するためには、今後、

様々な分化段階にある細胞の詳細な顕微鏡画像解析と比較を行う必要がある。
さらに、染色体表面の A 領域と B 領域の粘弾性を実験的に推定する必要がある。これは、高分子レオロジーの議論に基づいて、これらの領域における遺伝子座の拡散運動をライブイメージングで観察することによって推定できるかもしれない。
染色体粗視化粒子鎖モデルのシミュレーションでは、X 染色体だけでなく、4 番と 19 番染色体も、ES cell model と 2-day cell model の間で、相同染色体重心間距離の確率分布のプロファイルに大きな変化が見られた（Fig. S3）。しかし、4 番と 19 番染色体のこのような動態の機序は、X 染色体とは異なり、明らかにすることができなかった。この現象については、理論的な研究だけでなく、実験的な検証も行う必要がある。
総じて、染色体内の構造エピジェネティックな特徴とその変化は、核内における位置と移動を制御する上で重要な役割を果たしている。このような議論を応用することで、細胞種や細胞周期に依存した核内構造形成機構に関する重要な知見が得られると期待される。

II－v：手法

II－v－i：ES 細胞および分化後 2 日目の細胞の染色体における Open/非 Open クロマチン 領域の決定

Miura ら [23] によって提案された手法を用いて、Hi-C データ の 2 つの biological replicate (Gene expression omnibus ID: GSM3127755, GSM3127759, GSM3127756, GSM3127760)[23] とマウスゲノムの局所的な GC 含量から、ES 細胞および分化後 2 日目の細胞の各 biological replicate ごとに 計算された A/B-コンパートメントのプロファイルを用いた。ただし A/B-コンパートメントは Hi-C データを主成分分析することによって得られ、ヒストン修飾をはじめとしたエピジェネティックマーカーの分布と良く相関する染色体上の区画を指す [14]。A/B-コンパートメントプロファイルは Hi-C データの第一主成分ベクトルと任意のエピジェネティックマーカー分布との相関から決定され、本研究では Miura らや Ikeda らの定義 [23, 75] に従ってゲノム上の GC 含量と正の相関を持つ向きを A/B-コンパートメントプロファイルとして採用した。
以下では、A/B-コンパートメントプロファイル値が正となるゲノム上の領域を A-コンパートメント、負となる領域を B-コンパートメントと呼ぶことにする。そしてその結果に基づき、ゲノム上の各領域を以下の 4 通りのいずれかに分類した。
(1) A 領域; 2 つの biological replicate で、A-コンパートメントとなる領域

(2) B 領域; 2 つの biological replicate で、B-コンパートメントとなる領域
(3) M 領域; 2 つの biological replicate 間で異なるコンパートメントとなっている領域
(4) NA 領域; A/B-コンパートメントプロファイルにおいて対応する領域に値が存在しない領域（Hi-C リードがマップされていないことに起因する。セントロメアまたはテロメア領域に対応すると推定した。）

II－v－ii：Hi-C データとゲノム上の各領域の動径分布から得られたポリマーに基づく染色体の基本立体構造の推定

ES 細胞と分化後 2 日目の細胞においてそれぞれ平均化された Hi-C データ（https://doi.org/10.5281/zenodo.3371884）[23]から、近年提案された多次元尺度構成法（multi-dimensional scaling：MDS）を拡張した手法 [68-70] である MDS2 法を用いて、各染色体の平均的な立体構造（基本立体構造）を表すポリマー鎖を構築した。この計算には MD2 法が実装されているアプリケーションである、PASTIS [71] を使用した。ここで MDS2 法とは、染色体のモデルを以下の手順にしたがってポリマー鎖として構築する。

(1) DNA の生物物理学的モデルを仮定することにより、ゲノム上の各領域間の接触頻度を、領域間の空間的距離に変換する。
(2) 次に、領域間の空間的距離行列に最も一致する立体構造を、最適化問題を解くことによって求める。

この MDS に基づく染色体立体構造推定手法は、染色テリトリーや topologically associated domain（TAD）様構造といった大域的構造をよく再現することが知られている。

この染色体をポリマー鎖で表したモデル（染色体ポリマーモデル）では、各モノマーの空間座標が前節で解析したゲノム領域のものと対応している（Table S1）。また各モノマーは、対応するゲノム領域が属する コンパートメント に応じて A、B、M モノマーと分類される（Fig. S6a、ただし NA 領域は空間座標を計算できないため NA 領域に対応するモノマーは存在しない）。

これらの空間座標データから、各染色体ポリマーモデルにおける中心からの各モノマーまでの距離を計算し、ある距離に存在するモノマーの数（A、B、M モノマーについて、それぞれND_A、ND_B、ND_Mと表記する）を評価した（Fig. 2）。

（注：本研究では、各染色体の立体構造を、PASTIS に実装されている NMDS 法、PM1

法、PM2 法 [71] のいずれを用いて構築しても、全体的な染色体形状や各モノマーの染色体ポリマー内における分布が定性的に変わらないことを確認している（Table S2-4）。そのため、以下では MDS2 法によって推定された染色体ポリマーモデルを、粗視化粒子鎖モデルに用いる染色体の基本立体構造となるポリマーと定義し、使用する）。

II－v－iii： 各染色体の A、B、M ドメインの決定

染色体ポリマーモデルを粗視化するために、各染色体ポリマーモデルをいくつかのモノマー群（ドメイン）に分割した（Fig. S6a）。ドメインの境界となるモノマー（境界モノマー）は、Miura ら [23] によって提案された以下の手順を同一のパラメータを使用して求めた（Table S1、詳細は [23] を参照）。

(1) まず各ゲノム領域の Insulation score を計算した。
ここで Insulation score とは、Hi-C データから得られる各領域の上流 N b 内と下流 N b 内の領域間における接触頻度として計算される量であり、スコアの低い領域ほど周辺領域と区画化されていることを示す（本研究では N = 200 k とした）。

(2) ゲノム領域に沿って Insulation score を平滑化し Boundary score を計算した。
Insulation score は非常にノイジーなプロファイルとなるため、Savitzky-Goley 法および delta vector 法を用いて平滑化を行った。

(3) Boundary score が極大値を取り、かつある域値を上回る領域を境界モノマーと定義する。

本研究では、各ドメインは境界モノマー間に挟まれたモノマー群から構成されると定義した（Fig. S6a）。

次に、各ドメインのエピゲノム状態を以下の 3 通りのいずれかに分類した（Fig. S6a）。

(1) A ドメイン; ドメイン構成するモノマーのうち、50%以上が A モノマーとなるドメイン

(2) B ドメイン; ドメインを構成するモノマーのうち、50%以上が B モノマーとなるドメイン

(3) M ドメイン; 上記以外のドメイン

また、NA 領域はセントロメアもしくはテロメア領域に対応し、ヘテロクロマチンを形成すると推定できるため、ドメインのエピゲノム状態を決定する際には、NA 領域も B 領域と定義した。

また、同じエピゲノム状態となるドメインがポリマーに沿って連続している場合、それら

を同一のエピゲノム状態を持つドメイン群と定義した（Fig. S6a）。

II − v − iv：ドメイン群の定義及びドメイン群の分割による染色体ポリマーモデルの粗視化

同じエピゲノム状態となるドメインがポリマーに沿って連続している場合、それらを同一のエピゲノム状態を持つドメイン群と定義した（Fig. S6a）。そしてドメイン群が決定された上で、染色体ポリマーモデルを以下の手順でいくつかの粒子で表現する粗視化を行った。

(1) 粒子に含まれるモノマー集団がほぼ球形状の空間分布を持つ粒子群となるように各ドメイン群を分割した（Fig. S6a）。ただし分割手順は染色体ポリマーモデルごとに以下の手順で行った。

(1. a) n 番染色体ポリマーモデル（X 染色体は n = 20 とした）における特徴的なドメイン数である CNDn を求めた。CNDn は、PDND の平均の減衰率を評価して決定した。PDND は以下の式で 2 つのドメイン間のドメイン番号の差の関数として求められる。

PDND ([2 つのドメイン間のドメイン番号の差])

= [2 つのドメイン重心(それぞれに含まれるポリマーの重心)間の物理的距離] / [2 つのドメイン間のドメイン番号の差]

ここでドメイン番号とは、染色体ポリマーごとに上流からポリマーに沿って与えられた番号である。

まず上記の式から全てのドメインの組み合わせを用いて PDND を計算すると、各染色体における平均 PDND は単調減少することが示された（Fig. S7）。そこで本研究では、各染色体における PDND を exp(−[2 つのドメイン間のドメイン番号の差] / L) を用いて最小二乗法でフィッティングを行い、パラメータ L を CNDn とした。CNDn は、2 つの遺伝子座間の物理的距離が増加する場合に、それに伴う 2 遺伝子座間のドメイン数の増加が弱まる特徴的な平均距離を表し、染色体をドメインからなる鎖として記述した際に得られる持続長に近い値を持つと考えられる。

(1. b) n 番染色体の各ドメイン群を分割するクラスター数 k を、[ドメイン群に含まれるドメイン数]/CNDn の小数以下を切り捨て整数値とし、ドメイン群に含まれる全てのモノマーの 3 次元座標について k-means クラスタリングを行う。このクラスタリングにより、CNDn より多くのドメインを含むドメイン群は、

複数のクラスターに分割される。その結果、得られたクラスター中心を連結することで、染色体ポリマーモデルが内包する湾曲した形状を、少ない粒子数で表現できると考えられる。

(2) クラスタリングの結果から、各ドメイン群をいくつか粒子で表現した。

(1) で行ったクラスタリングで得られたそれぞれのクラスターが、ポリマーに沿って連続したモノマーのみで構成される場合は、そのクラスターを 1 つの粒子で表現した。そうでない場合は、そのクラスターを染色体ポリマーモデルに沿って連続してつながっているモノマーのみからなるサブクラスターに分割し、各サブクラスターを 1 つの粒子で表現した。

そして各粒子の座標を粒子に含まれるモノマー群の座標の重心で与え、半径を重心から各モノマーまでの距離の標準偏差で与えた。

(3) 各粒子のエピゲノム状態を以下の 3 通りのいずれかに分類した。

(3. a) A 粒子; 粒子に含まれるモノマーの 50 % 以上が A モノマーである粒子

(3. b) B 粒子; 粒子に含まれるモノマーの 50 % 以上が B モノマーである粒子

(3. c) M 粒子; 上記以外の粒子

(1)～(3)の手順で得られた粒子を用い、隣接する粒子をポリマーに沿って連結することにより、構造、ゲノム、エピゲノムの特徴を含む粒子鎖として、染色体の粗視化モデルが構築できる (Fig. S6b、Table S5)。ただしこのモデルでは、各染色体が最も安定な構造をとるときに、同一の鎖に属する粒子間の距離が染色体粒子鎖モデルの「基本立体構造」から得られる距離になると仮定した。

近年提案されている大規模ゲノム構造の粗視化モデルでは、各粒子に含まれるゲノム上の領域の長さは一定であると仮定され、各染色体は一定の半径を持つ粒子の鎖で表現されてきた [76–79]。一方、本モデルでは、各粒子は局所的なクロマチン構造を反映した実験データに基づき、各粒子に含まれる長さが変化し得ると仮定した。そのため、各粒子の半径は不均一に得られ、少ない粒子数によってシミュレーションにおける計算コストを抑えつつ、これまでのモデルと比較して各染色体の形状をより適切に記述することが可能である。

ES cell model および 2-day cell model は、それぞれ ES 細胞および分化後 2 日目の細胞から得られた Hi-C に対し II − v − i から II − v − iv の操作を行うことで構成される。

II－v－v：粗視化染色体モデルにおける各粒子が従う運動方程式

ES cell model と 2-day cell model における全ての染色体（1-19 番染色体および X 染色体）のダイナミクスを、粒子間相互作用ポテンシャルと核質からのノイズの影響を受ける各粒子（Table S5）の運動の総体としてシミュレーションした。i 番目の粒子の運動は、以下に示す過減衰ランジュバン方程式に従うと仮定した。

$$\gamma_i \frac{\partial \boldsymbol{x}_i}{\partial t} = -\frac{\partial V}{\partial \boldsymbol{x}_i} + \boldsymbol{R}_i(t) \qquad \text{(Eq. 1)}$$

ここで、$\boldsymbol{x_i} = (x_i, y_i, z_i)$ はi番目の粒子の座標を、Vは系のポテンシャルエネルギーを示し、γ_i と $\boldsymbol{R}_i(t)$ はそれぞれ粘性係数と i 番目の粒子が核質から受けるランダム力であり、核質の粘性を η、i 番目の粒子の半径を r_i とした場合に $\gamma_i = 6\pi\eta r_i$ で与えられる。$\boldsymbol{R}_i(t)$ はガウシアンホワイトノイズで与えられ、$\langle \boldsymbol{R}_i(t) \rangle = 0$、$\langle R_i(t) R_j(s) \rangle = 2\gamma_i k_B T \delta_{ij} \delta(t-s)$ を満たす。ここで k_B はボルツマン定数、T は絶対温度、δ_{ij} はクロネッカーのデルタ、δ はディラックのデルタ関数を示す。

(Eq. 1) の右辺の第 1 項は、系のポテンシャルによってi 番目の粒子に働く力を示す。そのポテンシャル V は以下の式により得られる。

$$V = V_{col} + V_{chr} + V_{mem} \qquad \text{(Eq. 2)}$$

ここで、V_{col} は粒子間に働く排除体積効果のポテンシャルエネルギー、V_{chr} は染色体構造を保つためのポテンシャルエネルギー、V_{mem} 粒子を核膜内に留めるためのポテンシャルエネルギーを示す。

ポテンシャル V_{col} は以下の式で与えられる。

$$V_{col} = \sum_{\{j<i\}} \theta\left(d_{ij}^0 - (r_i + r_j)\right) \theta\left((r_i + r_j) - d_{ij}\right) \frac{k_{ij}^c}{2} \left(d_{ij} - (r_i + r_j)\right)^2 \qquad \text{(Eq. 3)}$$

ここで r_i は i 番目の粒子の半径であり、d_{ij}^0 は前節で示した i 番目の粒子と j 番目の粒子間の基本構造における距離に対応しする。$d_{ij} = |\boldsymbol{x}_i - \boldsymbol{x}_j|$ とし、θ は以下に示す Heaviside の階段関数で定義される。

$$\theta(y) = \begin{cases} 1 & y \geq 0 \\ 0 & y < 0 \end{cases} \qquad \text{(Eq. 4)}$$

d_{ij}^0 は i 番目の粒子と j 番目の粒子が同一の鎖に属する場合はそれらの基本構造における距離として与えられ、そうでない場合は無限大として与えた。$k_{ij}^c = k^c \sqrt{q_i q_j}$ は i 番目と j の粒子が互いに接触した場合に働く弾性反発力の係数であり、q_i は i 番目の粒子自身のエピゲノム状態に基づく硬さに対応する無次元パラメータとして与えられた。

ポテンシャル V_{chr} は以下の式で与えられる。

$$V_{chr} = V_{chr}^{ad} + V_{chr}^{near} \quad \text{(Eq. 5)}$$

$$V_{chr}^{ad} = \sum_{\{i<j:\,\text{adjacent}\}} \frac{k_{chr}^{ad}}{2}\left(d_{ij} - d_{ij}^{0}\right)^2 \quad \text{(Eq. 6)}$$

$$V_{chr}^{near} = \sum_{\{j<i:\text{near}\}} \theta\left(\left(r_i + r_j\right) - d_{ij}^{0}\right)\frac{k_{chr}^{near}}{2}\left(d_{ij} - d_{ij}^{0}\right)^2 \quad \text{(Eq. 7)}$$

ここで k_{chr}^{ad} および k_{chr}^{near} は染色体の基本構造を保つための弾性係数である。(Eq. 6) の Σ は i 番目の粒子と j 番目の粒子が同一の鎖に沿って連結する粒子の組についての総和を示す。(Eq. 7) の Σ は i 番目の粒子と j 番目の粒子が同一の鎖に属し、かつ基本構造において空間的に接触しているが鎖に沿って連結されていない粒子の組についての総和を示す。

ポテンシャル V_{mem} は以下の式で与えられる。

$$V_{mem} = \sum_i \theta\left(\left(r_i + d_i\right) - R\right)\frac{k_i^m}{2}\left(\left(r_i + d_i\right) - R\right)^2 \quad (8)$$

ここで $d_i = |\boldsymbol{x}_i|$ であり、R は核膜の半径、$k_i^m = k^m\sqrt{q_i}$ は i 番目の粒子が核膜に接触した場合に働く弾性反発力の係数である。

II－v－vi：シミュレーション手法

このモデルにおいてシミュレーションを行うために、ランジュバン方程式 (Eq. 1) の時間積分を、Euler–Maruyama スキームを用いて時間ステップ 10^{-5} 秒として数値的に行った。核質に関するパラメータについては $\eta = 0.64\ kg\ m^{-1}\ s^{-1}$、また $k_BT = 4.141947 \times 10^{-21}\ kg\ m^2\ s^{-2}$ $(T = 300\ K)$ とした。

A 粒子、B 粒子、M 粒子はそれぞれ対応するエピゲノム状態の定義から、Open で柔軟なクロマチン領域、Closed で高密度のクロマチン領域、中間的な特徴を持つクロマチン領域を表現している。そのため M 粒子は A 粒子よりも剛性が高く（硬く、排除体積効果が大きくなる）、かつ B 粒子よりも剛性が低い（柔らかく、排除体積効果が小さくなる）ものと仮定し、本研究では q_i は A 粒子、B 粒子、M 粒子でそれぞれ 0.1、10.0、1.0 と定義した。その他のパラメータは $k^c = 10^{-4}\ kg\ s^{-2}, k_{chr}^{ad} = 10^{-3}\ kg\ s^{-2}, k_{chr}^{near} = 10^{-5}\ kg\ s^{-2}, and\ k^m = 10^{-4}\ kg\ s^{-2}$ とした。ただし、これらのパラメータを生理学的に正確に決定することは困難である。したがって、これらの値は、2 つの局所粒子鎖が互いに透過しない程度に十分大きな値とした。また、これらのパラメータの係数部分を変えても、シミュレーション結果の定性的な特徴や今回の

研究における結論は変わらなかった。
本研究では相同染色体は同一の基本構造を持つと仮定し、各シミュレーションの初期条件として、半径 5 μm の球殻の中に全ての粒子鎖をランダムに配置した。

II − v − vii : シミュレーションデータの統計解析

Xic 領域を含む 2 つの粒子（Xic 粒子）間距離の確率分布を、過去の実験結果の報告 [28, 29] と同様に 1 μm の bin 幅で描画した。また、相同染色体重心間距離の確率分布も 1 μm の bin 幅で描画した。
ES cell model と 2-day cell model の両モデルについて、染色体粒子鎖の初期配置を変えてそれぞれ 30 回のシミュレーションを行った。各シミュレーションにおける Xic 粒子間距離や相同染色体重心間距離の確率分布は、シミュレーション開始後 18 秒から 36 秒までの粒子座標のデータを用いて評価した。両モデルとも、30 個の確率分布を用いて、各距離を示す確率の平均値と 95%信頼区間（図中のエラーバーに対応する）を計算した。各距離を示す確率の ES cell model と 2-day cell model との差の有意性の検証を Welch の t 検定（有意水準 α = 0.05）を行った。

第 III 章　ロングリードシーケンスを用いた *Hemicentrotus pulcherrimus* の連続かつ高分解能ドラフトゲノム配列の構築

III－i：概要

初期発生のモデル生物として東アジアで広く研究されているバフンウニ *Hemicentrotus pulcherrimus*（*H. pulcherrimus*, *Hp*）のドラフトゲノム配列の更新を、Oxford nanopore long-read sequencing を用いて行った。その結果、2,163 contigs、総長 626.4 Mb、 N50 = 515.7 kb となるドラフトゲノム配列が得られた。 また、このドラフトゲノム配列の BUSCO 完全性スコアとトランスクリプトームモデルマッピング率（TMMR）は、それぞれ 96.5%と 77.8%であり、既に公開されている *H. pulcherrimus* ドラフトゲノム配列である HpulGenome_v1 （16,251 scaffolds、総長 568.9 Mb、gap closing (scaffolding) に由来する不明な塩基 = 93.3 Mb、N50 = 142.6 kb、BUSCO 完全性スコア 86.1 %、TMMR = 55.4 %）に比べて、大幅に連続性と精度が改善された。また、得られたドラフトゲノム配列には、他のウニにおける遺伝子モデルとほぼ同数の 36,055 の遺伝子モデルが含まれ、さらに、初期型ヒストン遺伝子をそれぞれ 47 コピーと 34 コピー含む 2 箇所のロングタンデムリピートや、*Ars* インスレーターの種間保存コア領域である ArsInsC、及びその 185 箇所の相同配列も含まれていた。本研究で得られたドラフトゲノム配列により、*H. pulcherrimus* を用いた発生、遺伝子発現制御、核内構造動態のゲノムワイドな研究の進展が加速されると期待される。

III－ii：導入

ウニはその進化的位置づけから、初期発生や形態形成の研究によく用いられるモデル生物である。例えば、東アジアに生息する代表的なウニであるバフンウニ（*Hemicentrotus pulcherrimus*）は、核内クロマチン構造ダイナミクス、左右非対称体軸の確立、中枢神経系の起源の探索など、多細胞生物の発生システムを明らかにするために広く研究されている [54-56]。またバフンウニとアメリカムラサキウニ（*Strongylocentrotus purpuratus*）は遺伝的に近縁であり、アメリカムラサキウニに対しては、初期発生時の内中胚葉の分化を制御する遺伝子制御ネットワークなども詳

細に解析されている [80, 81]。
一般に、生命システム全体の普遍的特徴や様々な生物の特異的な活動の研究の促進には、それらの生物のゲノム配列の詳細な特徴抽出が不可欠である。例えば、ここ 10 年で急速に進展した Hi-C 法やその派生法をヒト細胞やマウスなどのモデル生物に用いた最近の研究 [14] により、各遺伝子の発現は、エンハンサー・プロモーターループ、ヌクレオソーム排他的非ループインスレーター配列 （NENLIS） [82]、topologically associated domain （TAD）、A/B コンパートメントなど、様々なシス調節要素とその高次構造によって影響を受けることが明らかになってきた [14]。 このようなシス制御構造による影響を数 kb から数 Mb の様々なスケールでの解析が可能となったのは、数 Mb 以上の連続性を持つ高精度なドラフトゲノム配列が、これらの生物で得られていたからである。
その様々な生物と同様に、ウニのドラフトゲノムも多くの種で構築が試みられている。そして 2006 年に *S. purpuratus* [83]（最新版は 2019 年に更新）、2018 年に *H. pulcherrimus* [57]、2020 年に *Lytechinus variegatus* [84]、2022 年に *Temnopleurus reevesii* [85]、2023 年に *Paracentrotus lividus* [86] のドラフトゲノムがそれぞれ公開されている。
その一方で、非モデル生物研究でまず行われる従来型の short-read を用いたゲノムアセンブリでは、転写された遺伝子領域とそのシス制御要素であるエンハンサーやインスレーターを含む連続した長い contig や scaffold を構築することは困難である。例えば、最近報告された *H. pulcherrimus* ゲノムのドラフトゲノム配列 （HpulGenome_v1）には、アリルスルファターゼ （*HpArs*） 遺伝子の約上流 2 kb に位置し、典型的なヌクレオソーム排他的ループ非形成型インスレーター配列 (NENLIS) [82] として知られる *Ars* インスレーター配列 （ArsInsC） [58, 59] が含まれていない。 さらに、HpulGenome_v1 の *HpArs* 遺伝子を含む scaffold には、もう一つの重要な制御領域である C15 エンハンサー [87, 88] も含まれていなかった。また、*H. pulcherrimus* ゲノムには初期型ヒストン H1、H2A、H2B、H3、H4 をコードする遺伝子の数十から数百のタンデムリピートが存在することが知られている [54] が、HpulGenome_v1 には対応するゲノム領域は含まれていなかった。このような結果は、short-read のみを用いたゲノムアセンブリでは、read 長を超える長さのリピート配列を含む領域によって、contig が断片化されてしまうことが、一つの大きな原因であると考えられる。
しかし上記の問題は、数十～数百 kb の read が得られる long-read シーケンサーの開発によって克服されつつある。特に Oxford Nanopore Technologies （ONT） 社によ

って開発されたMinIONを代表するlong-read sequencerでは、理論上どんな長さであってもDNAをシーケンスできるという利点はある。ただしMinIONから得られたreadには、short-readシーケンスから得られたreadの100倍以上の頻度でエラーが含まれていることが知られており、精度面での問題が懸念されている。しかし、同じ生物から得られたshort-readを用いてlong-readのエラーを修正した上でアセンブリを行うハイブリッドアセンブリを行うことで、連続的で高精度のゲノムアセンブリが可能であると考えられる。

そこで本研究では、Oxford Nanopore sequencerから得られたlong-readと、従来の*H. pulcherrimus*ドラフトゲノム配列構築に用いられたshort-readを用いたハイブリッドアセンブリにより、*H. pulcherrimus*ドラフトゲノム配列を更新した。本研究では、更新されたドラフトゲノム配列の各指標をHpulGenome_v1と比較し、近縁種との比較による遺伝子の機能アノテーションを行った。また、更新されたドラフトゲノム配列において、*HpArs*遺伝子の上流に存在するArsInsCや、初期ヒストン遺伝子の2箇所のロングタンデムリピートを同定し、HpulGenome_v1において既知のゲノム構造と矛盾していた点を克服することができた。このようなドラフトゲノム配列の改良は、*H. pulcherrimus*における遺伝子発現を制御する生理学的・構造力学的特徴に関する研究を進展させると期待される。

III－iii：結果

III－iii－i：ゲノムアセンブリの結果とドラフトゲノム配列の完全性

ONT社のMinIONを用いたシーケンシングにより、51.54Gb（17.29 M reads）のシーケンスデータが得られた。以後、MinION得られた各long readをONT-readと呼び、HpulGenome_v1のアセンブリに使用されたshort readをIllumina-readと呼ぶ（詳細はIII－v：手法　の項を参照）。ゲノムアセンブリを行うために、Illumina-readを使用したONT-readのエラー補正を行い、また短いONT-readのフィルタリングを行った結果、35.26 Gb（2.76 M reads）のONT-readが得られた（詳細はIII－v：手法　の項を参照）。これらのONT-readからRaven [89]、Flye [90]、Wtdbg2 [91]、MAECI [92]でゲノムアセンブリおよびポリッシングを行い（詳細はIII－v：手法　の項を参照）、ドラフトゲノム配列を更新した（Table 1、Table S6）。その後BLASTN [93]を用いて、ドラフトゲノム配列に含まれるcontig群のうち、*H. pulcherrimus*のミトコンドリアゲノム［NCBI Accession ID: NC_023771.1］と最も相同性が高かったcontig（contig

ID: Utg196084）を除去した。更新されたドラフトゲノム配列は 2,163 contigs、総長は 626.4 Mb、N50 は 515.7 kb（Fig. 6A、Table S6）であった。一方で、HpulGenome_v1 は、16,251 scaffolds（一部に contig を含む）、 総長は 568.9 Mb、gap closing (scaffolding) に由来する不明な塩基を 93.3 Mb 含み、N50 は 142.6 kb であった（Fig. 6B）。

BUSCO [94] を用いたゲノムの完全性評価（使用データベース; metazoa_odb10）では、更新されたドラフトゲノム配列の Complete (遺伝子のほぼ全長がゲノム配列中に存在する）スコアが 96.5 %（single copy 89.9 %、duplicated 6.6 %）を示し、HpulGenome_v1 の 86.1 %（single copy 84.8 %、duplicated 1.3 %）より高い値が得られた。また、推定トランスクリプトームモデルのマッピング率も 77.78 %と、HpulGenome_v1 の 55.35 %より高い値が得られた（Table 1)。

Table 1. 本研究で構築したドラフトゲノム配列と HpulGenome_v1 の各指標と BUSCO 評価。
* は BBtools [95] (stats.sh) を用いて計算した。

	Updated draft genome	HpulGenome_v1
Assembly size*	626.4 Mb	568.9 Mb
No. contigs*	2,163	86,128
N50 contig length*	515.7 kb	9.641 kb
No. scaffolds*	2,163	16,251
N50 scaffold length*	515.7 kb	142.6 kb
Unsure base ("N", %)* (derived from gap-closing)	0	16.41
GC-content (%)*	37.12	36.72
BUSCO completeness (%) (metazoa_odb10: 954 genes)	Complete : 96.5 Duplicated : 6.6 Fragmented : 2.1 Missing : 1.4	Complete : 86.1 Duplicated : 1.3 Fragmented : 9.9 Missing : 4.0
Mapping ratio of transcriptome models (%) (20,564 sequences)	75.65 (aligned exactly 1 time) 2.13 (aligned >1 times)	54.64 (aligned exactly 1 time) 0.71 (aligned >1 times)

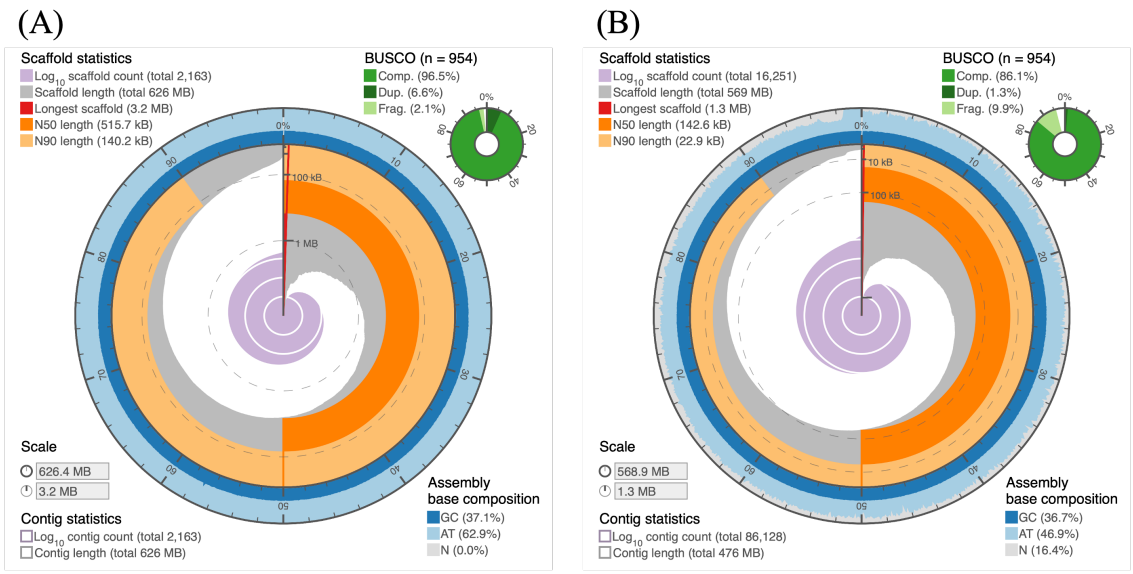

Fig. 6 本研究で得られたドラフトゲノム配列（A）と HpulGenome_v1（B）のアセンブリ指標。アセンブリ指標は assembly-stats を用いて描画した。(https://github.com/rjchallis/assembly-stats)

III − iii − ii：ドラフトゲノム配列に基づく遺伝子構造推定と近縁種とのオルソログ検索による機能アノテーション

H. pulcherrimus のトランスクリプトームデータ（DRR107783）と BRAKER2 [96] による解析の結果、更新されたドラフトゲノム配列には 46,914 個の遺伝子が存在すると推定された。短すぎる遺伝子（50 アミノ酸未満）や複数の終止コドンを含む遺伝子、ミトコンドリアゲノムにコードされている遺伝子を除外した結果、46,826 の遺伝子モデルが得られた（Table S7）。この遺伝子モデル群と、3 種のウニの遺伝子モデル群を用いた相互 BLASTP 検索を行うことにより、36,055 遺伝子モデルが他のウニにおいてホモログが見つかり、その数は *H. pulcherrimus*、*S. purpuratus*、*L. variegatus* の既知の遺伝子モデルと同等であった。ただし BLASTP 検索では e-value ≤ 1e-10 とした。その結果、相互ベストヒットな遺伝子ペアが 20,434、相互ベストヒットではないがベストヒットな遺伝子ペアが 15,621 であった（Table 2、Table S7）。

Table 2. HpBase と他のウニのタンパク質モデルとの BLASTP 検索の結果。*P. lividus* のタンパク質配列はゲノムとアノテーション情報を Zenodo (https://zenodo.org/record/7459274) からダウンロードし、gffread v0.12.7 [97] を用いて作成した。

Subject	Reciprocal best hit pair	Not reciprocal but best hit	No hit
H. pulcherrimus (HpBase, HpulGenome_v1_prot.fa)	15,652 (33.43%)	16,847 (35.98 %)	14,327 (30.60 %)
S. purpuratus (Echinobase, sp5_0_GCF_proteins.fa)	16,153 (34.50 %)	17,190 (36.71 %)	13,483 (28.79 %)
L. variegatus (Echinobase, Lvar3_0_GCF_proteins.fa)	14,802 (31.61 %)	16,613 (35.48 %)	15,411 (32.91 %)
*P. lividus**	14,958 (31.94 %)	18,388 (39.27 %)	13,480 (28.70 %)
Hp, Sp, or Lv	20,434 (43.64 %)	15,621 (33.36 %)	10,771 (23.00 %)

III – iii – iii : ロングタンデムリピートを持つ初期型ヒストン遺伝子座の検出

H. pulcherrimus ゲノムには、初期型ヒストンをコードする遺伝子が数十から数百のタンデムリピートを形成する配列が 2 copy 存在することが示唆されている [54]。しかし、一般にロングタンデムリピートは、short-read によるゲノムアセンブリでは検出するのは困難である。実際 HpulGenome_v1 では、初期型ヒストン遺伝子のロングタンデムリピートに対応するゲノム領域は検出されなかった。一方、更新されたドラフトゲノム配列では BLASTN 検索の結果、初期型ヒストン遺伝子のロングタンデムリピートが 2 つの contig において検出された。1 つの contig（contig ID：Utg198178、479,333 bp）には 47 個の high-scoring segment pair; HSP が含まれ、もう 1 つの contig（contig ID：Utg200276、127,611 bp）には 34 個の HSP が含まれていた（Table S8）。

III – iii – iv : アリルスルファターゼ遺伝子の制御配列

本研究では追加の解析として、アリルスルファターゼ（*HpArs*）遺伝子の転写制御配

列について調べた。*HpArs* 遺伝子の制御配列とその候補として以下のような特徴的配列が報告されている。

(1) プロモーター領域が転写開始点の上流 252 b から下流 38 b の領域に存在する。この配列は *HpArs* 遺伝子の発現に必要な最小領域である [87, 98]。

(2) 第 1 イントロンに C15 エンハンサーと呼ばれる、発現増強に寄与する配列が存在する [87, 88]。

(3) 分子内三重鎖構造を形成しうるポリピリミジン配列が、転写開始点の上流 2,201 b から 1,680 b の領域に存在する [99, 100]。

(4) 転写開始点の上流 2,686 b から 2,109 b の領域にインスレーター配列（*Ars* インスレーター）が存在する [59]。

(5) 転写開始点の上流 3,440 b から 3,109 b の領域に直接反復配列（direct repeat: DIR) Ars-DIR1、3,096 b から 2,592 b の領域に Ars-DIR2 が存在する [101]。

(6) 転写開始点の上流 475 b から 217 b の領域に逆位反復配列（inverted repeat: INV) Ars-INV が存在する [101]。

HpulGenome_v1 では、scaffold989 に *HpArs* 遺伝子が存在している。BLASTN 検索の結果、この scaffold にはすべてのエキソン、プロモーター、ポリピリミジン領域までの上流配列が含まれていること明らかになった。しかし、転写制御に重要な役割を果たす第 1 イントロンの C15 エンハンサーとその隣接する上流領域の *Ars* インスレーターは含まれていなかった。一方、更新されたドラフトゲノム配列では、*HpArs* 遺伝子が存在する contig（contig ID：Utg200732）には全てのエキソンおよび上記の (1)～(6) の配列全てが含まれていた。

III − iii − v : Ars-DIR1 および Ars-DIR2 と Ars-INV の相同配列とゲノム上での分布

更新されたドラフトゲノム配列において、以下の条件を満たす領域を Ars-DIR1 および Ars-DIR2 のホモログとした。

(1) ドラフトゲノム配列に対し、Ars-DIR1 および Ars-DIR2 の各反復単位（Table S10）をクエリとして BLASTN 検索を行い、クエリカバレッジ 95 % 以上かつ相同性 90 % 以上のヒットが得られる。

(2) (1) で得られたヒットのうち、隣接するヒットした領域の中心間距離が 300 b 以下となるものを連続するヒットとし、連続するヒット数をリピート数とする。

その結果、Ars-DIR1 と相同な領域が 8 箇所（全て 2 リピート)、Ars-DIR2 と相同な配列が 205 箇所（2～5 反復）見つかった（Table S10)。さらに、これらの Ars-DIR1 と

Ars-DIR2 のホモログはいずれも、最も近い遺伝子の上流領域に存在する傾向があった（Fig. 7）。

次に IRF v3.08 [102] を用いて更新されたドラフトゲノム配列上に存在する INV を検索し、以下の条件を満たすものを Ars-INV のホモログとした。

(1) 反復単位長 90-120 b かつ同一性 90%以上である。

(2) 反復単位領域に対して Ars-INV の反復単位をクエリとした BLASTN 検索を行った結果、クエリカバレッジ 90 % 以上となるヒットが得られる。

その結果 Ars-INV と相同な 9 個の INV 配列が更新ドラフトゲノム中に見つかった (Table S9)。これらの Ars-INV 相同配列のうち 7 つは遺伝子間領域に存在したが、その位置に特異的な傾向は見られなかった。

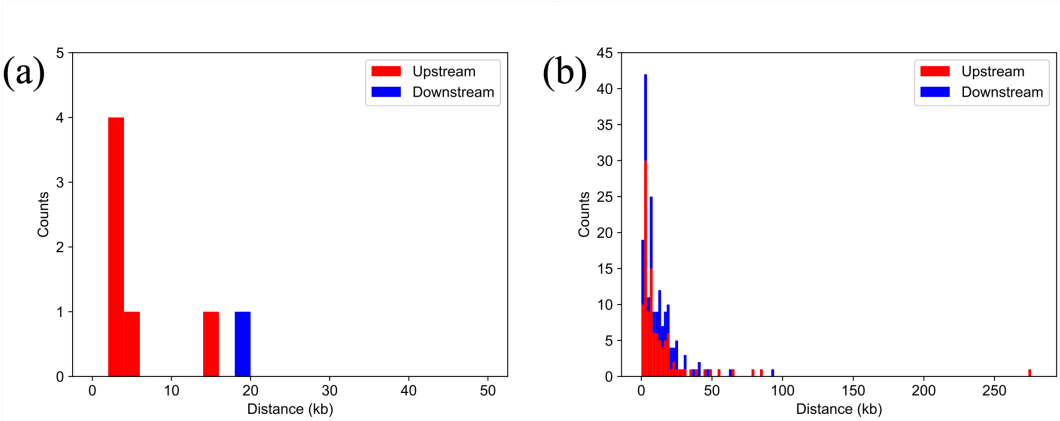

Fig. 7. DIR1 ホモログと最近接遺伝子間距離のヒストグラム（a）と、DIR2 ホモログと最近接遺伝子間距離のヒストグラム（b）。赤と青のバーの高さは、それぞれ最近接遺伝子の上流領域と下流領域における DIR ホモログの数を示す。DIR ホモログと最近接遺伝子との距離は、それらの中心間の距離と定義した。

III - iii - vi : *Ars* インスレーター及びその相同なゲノム領域の探索

HpArs 遺伝子の転写開始点から上流約 2 kb には、次の特性を持つ *Ars* インスレーターが存在する [58, 59]。

(1) *Ars* インスレーター中央部には AT リッチな領域（ArsInsC）が存在する。この領域は *S. purpuratus* の *Ars* 遺伝子上流領域にも保存されている。

(2) ArsInsC は CTCF などの結合タンパク質などを伴わず単独でインスレーター活性を示す、Nucleosome-excluding non-looping insulator sequence (NENLIS) [82] の典型例である可能性が高い。

HpulGenome_v1 では、*HpArs* 遺伝子は scaffold989 に存在していたが、ArsInsC 配列は同一 scaffold 上に存在しなかった。一方で、更新されたドラフトゲノムでは、*HpArs* 遺伝子領域の 2 kb 上流にあたる Utg200732 (46,417 to 46,236, マイナス鎖側) に ArsInsC が存在していた。BLASTN 解析の結果、ArsInsC と 90%以上の相同性を持ち、同じ長さとなる配列 (ArsInsC ホモログ) が全ゲノム中に 185 箇所見つかった (Table S11)。さらに *Ars* インスレーターと同様に、ArsInsC ホモログの上流および下流それぞれ 50 b の近傍領域は GC 含量が高い (50–80 %) 傾向を示し (Fig. 8、Fig. S8)、このうち 121 箇所の ArsInsC ホモログ (65.4 %) では、8 b 以上の G (C) –ストレッチが近傍の GC リッチ領域に存在していた (Fig. S8)。
このように、*Ars* インスレーターの配列特性はこれらの相同配列においてよく保存されており、これらの相同配列がウニゲノムにおいてインスレーターとして広く用いられていることが示唆された。

III – iii – vii : ショートタンデムリピート (short tandem repeat : STR) のゲノム上での分布
STR は遺伝子のプロモーターに濃縮している傾向があり、リピート数の変動が遺伝子発現を制御し得ることが報告されている [103, 104]。さらに、約 90 % の転写因子が STR に優先的に結合する [105] ことも報告されている。そこで、TRF v4.09 [106] を用いて、更新されたドラフトゲノムにおける STR の分布を解析した。Sawaya らのマイクロサテライトの定義 [103] に従い、長さ 12 b 以上で挿入のない 1～4 ヌクレオチドの繰り返し単位からなる反復配列を検索した。TRF の結果から、この条件を満たす STR を検索した結果、合計 111,553 個の一塩基反復、66,665 個の二塩基反復、61,983 個の三塩基反復、および 62,448 個の四塩基反復が更新ドラフトゲノム中に見つかった (Table S12)。しかし、遺伝子の近傍や翻訳開始点における STR の濃縮は観察されなかった。

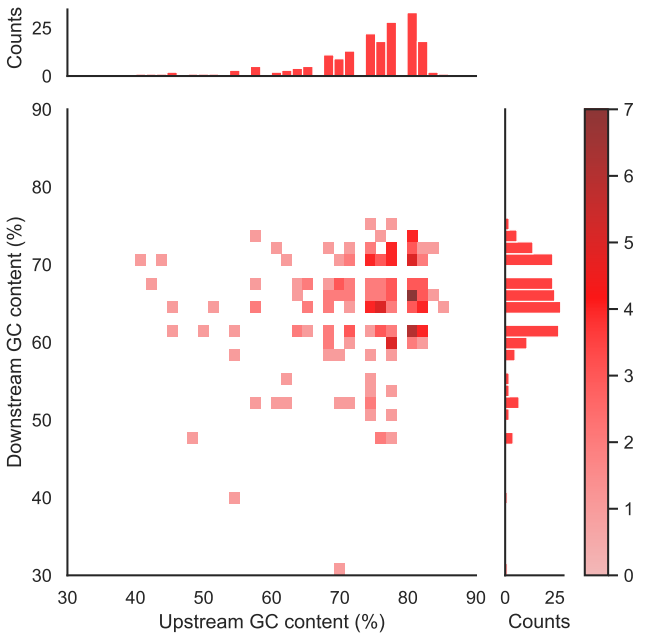

Fig. 8 ArsInsC ホモログ数のヒートマップを、上流 50 b 領域と下流 50 b 領域における各 GC 含量で示した。ヒストグラムはそれぞれ対応する軸について示したものである。

III – iv : まとめと考察

Kinjo らによって構築された HpulGenome_v1 は、静岡県下田市で採集された *H. pulcherrmius* の雄成体に由来するドラフトゲノム配列である [57]。いくつかの遺伝子配列解析の結果から、ウニゲノムは多型度が大きく、ヒトでは一塩基多型および挿入/欠失がゲノム中に 0.5 % 程度存在するのに対して、ウニでは 4〜5 % 程度存在することが報告されている [83, 107, 108]。そのため *H. pulcherrimus* におけるゲノムワイドな解析には、多型度の高さを反映した広く利用可能なゲノム配列が必要であり、本研究では広島で採集された複数個体の *H. pulcherrimus* から単離されたゲノム DNA を用いてゲノムアセンブリを行った。また、本研究では HpulGenome_v1 よりも連続的で精度の高い配列を得るために、Oxford Nanopore シーケンサーを用いた long-read シーケンスと、long-read と short-read によるハイブリッドゲノムアセンブリを行った。その結果、HpulGenome_v1 よりも BUSCO 完全性が高く、トランスクリプトームモデルのマッピング率が高いドラフトゲノム配列が得られた。

更新されたドラフトゲノム配列を用いた遺伝子モデル推定の結果、46,826 の遺伝子モデルが推定され、HpulGenome_v1 において推定された 24,860 の遺伝子モデルよりも

多くの数が得られた。このうち 36,055 の遺伝子モデルについては、近縁種ウニの相同な遺伝子モデルとの相互 BLASTP 検索による機能アノテーションを行った。
さらに、このドラフトゲノム配列上に、初期型ヒストン遺伝子のロングタンデムリピートを内包するゲノム領域が 2 箇所見出された。この結果は、バフンウニの核内蛍光イメージングの結果 [54] と一致しており、発生段階依存的な初期型ヒストン遺伝子座の相互作用および動態の解析に進展をもたらすと考えられる。
興味深いことに、このドラフトゲノム配列には ArsInsC と 185 の ArsInsC ホモログが存在していた。脊椎動物では、ゲノム内の CTCF 結合部位は遺伝子発現制御を担う典型的なインスレーター配列として知られている。しかし最近では、ウニの CTCF は分裂期から間期への移行において機能することが示唆されており [109]、ウニの間期における遺伝子発現制御には ArsInsC の他にも様々なゲノム配列由来の因子が関与している可能性が考えられる。今回更新されたドラフトゲノム配列は、*H. pulcherrimus* において遺伝子発現を制御するシスエレメントの生理学的・構造力学的研究、および広く様々な生物に存在する多様なシス制御因子の特徴及び作動機序の研究を進展させると期待される。
また、更新されたドラフトゲノムには、近縁種の遺伝子モデルにホモログの存在しない遺伝子モデルが 10,000 以上含まれていた。これらの遺伝子モデルは、ノンコーディング RNA 由来遺伝子や *H. pulcherrimus* 特異的遺伝子を含むと考えられる。今後、これらのアノテーションできなかった遺伝子モデルについては、様々な生物との遺伝子比較や実験による機能解析の結果から機能アノテーションを行う必要がある。
BUSCO を用いた解析では、HpulGenome_v1 よりも多くの重複遺伝子を持つことが示された。BUSCO の core gene set の定義に基づけば、重複 BUSCO 遺伝子の割合が低いほど妥当なゲノム配列であると言える。したがって本研究において、複数個体の DNA を用いたゲノムアセンブリを行ったことからも、本ドラフトゲノム配列は冗長な配列を含む可能性が考えられる。しかし ONT-read はエラーを比較的多く含むため、k-mer 分布解析から多型を含む領域とアセンブリに影響を与え得る領域の定量的な評価が困難であり、本研究では read のフィルタリングや contig の除去は行わなかった。本ドラフトゲノム配列上には、ATCG や N 以外で表記された塩基が 1,625 箇所存在し、多型を含むが大域的には単一の contig として表現される領域も存在すると考えられる。しかし、一般的に知られているウニの多型を含む領域の割合には及ばないため、今後は大域的アライメントや Hi-C 法による scaffolding を用いて contig 数を減らすことで、より冗長でないドラフトゲノム配列を構築する必要がある。

III－v： 手法

III－v－i： 生体の採集とゲノム DNA の抽出

広島県江田島市近海にて *H. pulcherrimus* の雄成体を採集し、精子細胞からゲノム DNA（gDNA）を抽出した。*H. pulcherrimus* の gDNA は Blood & Cell Culture DNA Midi Kit（QIAGEN）を用いて精製し、短いゲノム DNA 断片は Short Read Elimination XL Kit（Circulomics Inc.）を用いて除去した。また、Logsdon（2020）のプロトコルに従って、より長い DNA 断片を精製した。

III－v－ii： シーケンスライブラリの調製

上記の手法により得られた gDNA から Rapid Sequencing Kit（SQK-RAD004; Oxford Nanopore Technologies [ONT]）を使用してライブラリを調製した。調製したライブラリは R9.4.1 Flow Cell を用いて ONT MinION でシーケンシングを行い、生成された FAST5 ファイルを ONT MinIT（MNT-001）を用いて FASTQ ファイルへと変換した。以下では、MinION および MinIT により生成された FASTQ 形式のリードを ONT-read とする。

III－v－iii : Illumina リードおよび ONT-read の前処理

Illumina シーケンサーから得られた *H. pulcherrimus* の gDNA（DRR107784、DRR107785、DRR107786）およびトランスクリプトーム（DRR107783）の short-read を、fastp v0.22.0 [110] を用いてアダプタートリミングおよび低クオリティ read の除去を行った。また、Ratatosk v0.7.0 [111] を用いて、前処理済みの illumina short-read（合計 160.47 Gb）による ONT-read のエラー補正を行った。

III－v－iv： ゲノムアセンブリとポリッシング

Raven v1.8.1 [89]、Flye v2.8.3-b1695 [90]、Wtdbg2 v2.5 [91] を用いて、エラー補正された ONT-read（> 5 kb）から *H. pulcherrimus* の de novo ゲノムアセンブリを行った。その後 MAECI [92] を用いて単一のコンセンサス配列を作成した。また、MAECI を用いて前処理済みの illumina-read により、コンセンサス配列のポリッシングを行った。

III－v－v： ドラフトゲノム配列の評価

BUSCO v5.3 [94] を用いて得られたドラフトゲノム配列の完全性を評価した。ここで、BUSCO は対象生物が含まれる適当な動物群の core gene set（その動物群のうち 90 % 以上の生物のゲノムにおいて single copy ortholog である遺伝子のアミノ酸配列

群）データベースを指定し、その検出数で評価を行う。本研究ではデータベースとして後生動物（metazoa_odb10）を指定し BUSCO を実行した。また、hisat2 v2.2.1 [112] を用いて、HpBase における推定トランスクリプトームモデル（HpulTranscriptome_nucl.fa にクオリティスコア 30 を付与し、FASTQ 形式とした）を各ドラフトゲノム配列にマッピングし、そのマッピング率を評価した。

III－v－vi：遺伝子構造推定と機能アノテーション

ドラフトゲノム配列に基づいて以下の手順で遺伝子構造推定を行い、得られた遺伝子群の機能アノテーションを行った。

(1) トランスクリプトームから得られた前処理済みの illumina short-read を、hisat2 v2.2.1 [112] を用いてドラフトゲノム配列にマッピングし、SAMtools v1.17 [113] を用いて BAM 形式のファイル（アライメントの情報を記述したもの）を作成した。

(2) BAM ファイルのアライメント情報から、BRAKER2 v2.1.4 [96] を用いて遺伝子構造推定を行った。

(3) BLASTP v2.6.0 [93] を用いて、BRAKER2 で推定された遺伝子モデルのアミノ酸配列群と、3 種のウニの遺伝子モデル（*H. pulcherrimus* v1.0（HpBase）, *S. purpuratus* v5.0, *L. variegatus* v3.0）のアミノ酸配列群から相互ホモログ検索を行った。

III－vi：公開データ

更新されたゲノムアセンブリの配列データは、HpBase（https://cell-innovation.nig.ac.jp/Hpul/）に HpulGenome_kure_v1 という名前で掲載されている。全ゲノムシーケンスの生データは、DDBJ Sequence Read Archive（DRA）にて以下の Accession ID で公開されている（DRA017089）。

第Ⅳ章　全体のまとめ

本学位論文では、転写制御機構に関連する核内動態や、転写制御の基盤となる特徴的な配列のゲノム上分布について研究を行った。

II 章では、哺乳類のモデル生物であるマウスの ES 細胞からの分化過程において一過的に観察され、分化に伴う遺伝子発現制御への寄与が示唆されている、X 染色体対合形成［28-31］の機序について、考察した。その結果、まず Hi-C データ［23］を用いた解析から、X 染色体において対合に先駆けてエピゲノム状態が変化するだけでなく、その X 染色体内構造上の分布をも再構成されることが明らかになった。そしてエピゲノム状態がクロマチン自身の物理的柔軟性を決定するという仮定のもとで染色体構造のモデル化と核内動態シミュレーションを行うことにより、分化開始後に再構成された X 染色体内構造上のエピゲノム状態分布が、X 染色体対合を駆動し得ることが見出された。本研究により、枯渇力とよばれる力［46-48］や、剛性の異なるポリマーの相分離を引き起こす力［44, 49-53］と同様の、分子間力由来ではなく系全体の状態数由来（エントロピー由来）の力が、各染色体内でのエピゲノム状態の変化を通じて調節されることで、核内構造が制御される可能性が見出された。この知見により、細胞種や細胞周期に依存した動的な核内構造形成機序を明らかにするための考察基盤として、エピゲノム情報が示す新たな可能性を示した。

III 章では、古くからの初期発生のモデル生物であるバフンウニのドラフトゲノム配列の再構成を行い、従来公開されていたものより高い連続性と精度のものへ更新した。バフンウニでは、核内染色体ダイナミクス、左右非相称体軸の確立、中枢神経系の起源の探索［54-56］など様々な研究がなされている。しかしそのドラフトゲノム配列（HpulGenome_v1）［57］には、ArsInsC や初期型ヒストン遺伝子のロングタンデムリピートの存在といった既報のゲノム構造と矛盾する部分を含むなど、多くの問題を抱えていた。そのため本研究ではこの問題を是正するために、HpulGenome_v1 と比べ、連続的で精度の良さを持ち合わせたゲノム情報の再整備を試みた。本研究ではまず、バフンウニ精子由来 DNA の MinION による long-read 解析を行い、次いでその long-read データと HpulGenome_v1 構築に用いられた short-read データの両方を用いたハイブリッドアセンブリを行った。その結果、HpulGenome_v1 より多くの、他のウニと無矛盾な遺伝子モデルを含み、さらに ArsInsC に代表されるような転写制御領域、初期型ヒストン遺伝子のロングタンデムリピートを含む、高い連続性を持ち合わせたバフンウニドラフトゲノム配列が得られた。このドラフトゲノムでは、初期型ヒストン遺伝子のタ

ンデムリピートを含む領域が2箇所見出されたが、これはバフンウニの核内蛍光イメージング [54] との比較より妥当な結果であり、今後発生段階依存的な Histone locus body 動態の研究を進展させると考えられる。また ArsInsC ホモログが 185 程度見出されたことから、脊椎動物ではほぼ CTCF 結合配列に担われていたインスレーター機能が、ウニでは多くのゲノム領域で ArsInsC ホモログに担われている可能性も考えられた。このように今回得られた新規ゲノム配列は、バフンウニにおける遺伝子発現を制御する新規シスエレメントの同定とその作用機序解明を支えられるものであり、他の生物との比較から、旧口動物と新口動物間、新口動物間における転写制御機構の変遷を検証できる新たな可能性を示した。

Ⅱ章で扱った議論が生物種、細胞種や細胞周期普遍的に適用可能であるかどうかは、様々な場合において染色体イメージングを行うとともに、本研究と同様のモデルを構築してイメージングの結果と一致するかを検証する必要がある。さらに、Ⅲ章で扱った議論についても、まずウニにおける遺伝子発現量解析や染色体立体構造解析によって ArsInsC ホモログの詳細な機能解析を行い、さらには様々な生物における ArsInsC 様配列の探索とその機能解析も併せて検討する必要がある。これらの解析については今後の課題としたい。

本学位論文では、現象を説明するために既報の解析データからどの様な価値を創出できるか（トップダウン的なアプローチ）、また、どの様なデータを収集しなければならないか（ボトムアップ的なアプローチ）、という視点に立って研究を行った。近年のボトムアップ的なアプローチとして、2018 年から始まった地球バイオゲノムプロジェクト（Earth BioGenome Project：EBP）では「現在報告されている全ての真核生物のゲノム情報を 10 年間で解読する」という目標を掲げており、生物の進化や生態系を含む生命の核心への理解を深め、動植物の絶滅や多様性の消失に対抗するための生態系の保全などに役立てられることが期待されている。その背景として実験手法やその解析ツールにおける技術革新の影響が強く、さらには解析を自動化するパイプラインも急速に整備されており、EBP が 10 年間で実現されるかはさておき、この計画がただの夢ではない時代が我々の目の前に迫ってきている。しかし一方で、計画完了の暁には蓄積されたデータを正しく理解し、価値を創出するための膨大な作業が必要になる。そのためには既存の枠組みに囚われず、独自の視点からもデータを多角的に解釈できる研究者が求められることは明白である。本学位論文もゲノム解析技術の革新によりもたらされた成果の一つであるが、ボトムアップ的なアプローチも踏まえつつ、トップダウン的なアプローチを見直すことでゲノムデータの新たな価値を検討した。本

論文により見出された知見が生命の核心に迫る一助となることを期待している。

参考文献

[1] O. T. Avery, C. M. MacLeod, and M. McCarty, “STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES,” *J. Exp. Med.*, vol. 79, no. 2, pp. 137–158, Feb. 1944, doi: 10.1084/jem.79.2.137.

[2] A. D. Hershey and M. Chase, “INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE,” *J. Gen. Physiol.*, vol. 36, no. 1, pp. 39–56, Sep. 1952, doi: 10.1085/jgp.36.1.39.

[3] J. D. Watson and F. H. C. Crick, “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid,” *Nature*, vol. 171, no. 4356, pp. 737–738, Apr. 1953, doi: 10.1038/171737a0.

[4] W. Fiers *et al.*, “Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene,” *Nature*, vol. 260, no. 5551, pp. 500–507, Apr. 1976, doi: 10.1038/260500a0.

[5] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proc. Natl. Acad. Sci.*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977, doi: 10.1073/pnas.74.12.5463.

[6] F. Sanger *et al.*, “The nucleotide sequence of bacteriophage ϕX174,” *J. Mol. Biol.*, vol. 125, no. 2, pp. 225–246, Oct. 1978, doi: 10.1016/0022-2836(78)90346-7.

[7] R. D. Fleischmann *et al.*, “Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd,” *Science (80-.).*, vol. 269, no. 5223, pp. 496–512, Jul. 1995, doi: 10.1126/science.7542800.

[8] A. Goffeau *et al.*, “Life with 6000 genes.,” *Science*, vol. 274, no. 5287, pp. 546, 563–7, Oct. 1996, doi: 10.1126/science.274.5287.546.

[9] C. E. S. Equence, T. O. B. Iology, C. The, and S. Consortium, “Genome Sequence of the Nematode C. elegans : A Platform for Investigating Biology,” *Science (80-.).*, vol. 282, no. 5396, pp. 2012–2018, Dec. 1998, doi: 10.1126/science.282.5396.2012.

[10] Mouse Genome Consortium *et al.*, “Initial sequencing and comparative analysis of the mouse genome,” *Nature*, vol. 420, no. 6915, pp. 520–562, Dec. 2002, doi: 10.1038/nature01262.

[11] P. Collas, “The Current State of Chromatin Immunoprecipitation,” *Mol. Biotechnol.*, vol. 45, no. 1, pp. 87–100, May 2010, doi: 10.1007/s12033-009-9239-8.

[12] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, “Capturing Chromosome Conformation,” *Science (80-.).*, vol. 295, no. 5558, pp. 1306–1311, Feb. 2002, doi: 10.1126/science.1067799.

[13] S. G. Landt *et al.*, “ChIP-seq guidelines and practices of the ENCODE and modENCODE

consortia," *Genome Res.*, vol. 22, no. 9, pp. 1813–1831, Sep. 2012, doi: 10.1101/gr.136184.111.

[14] E. Lieberman-Aiden *et al.*, "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science (80-.).*, vol. 326, no. 5950, pp. 289–293, Oct. 2009, doi: 10.1126/science.1181369.

[15] T. Takizawa, K. J. Meaburn, and T. Misteli, "The Meaning of Gene Positioning," *Cell*, vol. 135, no. 1, pp. 9–13, Oct. 2008, doi: 10.1016/j.cell.2008.09.026.

[16] T. Cremer and C. Cremer, "Chromosome territories, nuclear architecture and gene regulation in mammalian cells," *Nat. Rev. Genet.*, vol. 2, no. 4, pp. 292–301, 2001, doi: 10.1038/35066075.

[17] P. Fraser and W. Bickmore, "Nuclear organization of the genome and the potential for gene regulation," *Nature*, vol. 447, no. 7143, pp. 413–417, 2007, doi: 10.1038/nature05916.

[18] S. T. Kosak and M. Groudine, "Form follows function: the genomic organization of cellular differentiation," *Genes Dev.*, vol. 18, no. 12, pp. 1371–1384, Jun. 2004, doi: 10.1101/gad.1209304.

[19] C. Lanctôt, T. Cheutin, M. Cremer, G. Cavalli, and T. Cremer, "Dynamic genome architecture in the nuclear space: Regulation of gene expression in three dimensions," *Nat. Rev. Genet.*, vol. 8, no. 2, pp. 104–115, 2007, doi: 10.1038/nrg2041.

[20] D. M. Gilbert *et al.*, "Space and Time in the Nucleus: Developmental Control of Replication Timing and Chromosome Architecture," *Cold Spring Harb. Symp. Quant. Biol.*, vol. 75, pp. 143–153, Jan. 2010, doi: 10.1101/sqb.2010.75.011.

[21] J. J. Roix, P. G. McQueen, P. J. Munson, L. A. Parada, and T. Misteli, "Spatial proximity of translocation-prone gene loci in human lymphomas," *Nat. Genet.*, vol. 34, no. 3, pp. 287–291, Jul. 2003, doi: 10.1038/ng1177.

[22] T. Misteli and E. Soutoglou, "The emerging role of nuclear architecture in DNA repair and genome maintenance," *Nat. Rev. Mol. Cell Biol.*, vol. 10, no. 4, pp. 243–254, Apr. 2009, doi: 10.1038/nrm2651.

[23] H. Miura, S. Takahashi, R. Poonperm, A. Tanigawa, S. ichiro Takebayashi, and I. Hiratani, "Single-cell DNA replication profiling identifies spatiotemporal developmental dynamics of chromosome organization," *Nat. Genet.*, vol. 51, no. 9, pp. 1356–1368, 2019, doi: 10.1038/s41588-019-0474-z.

[24] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data," *Nat. Rev. Genet.*, vol. 14, no. 6, pp. 390–403, Jun. 2013, doi: 10.1038/nrg3454.

[25] J. R. Dixon *et al.*, "Chromatin architecture reorganization during stem cell differentiation,"

Nature, vol. 518, no. 7539, pp. 331–336, 2015, doi: 10.1038/nature14222.

[26] E. de Wit *et al.*, "The pluripotent genome in three dimensions is shaped around pluripotency factors," *Nature*, vol. 501, no. 7466, pp. 227–231, Sep. 2013, doi: 10.1038/nature12420.

[27] J. E. Phillips-Cremins *et al.*, "Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment," *Cell*, vol. 153, no. 6, pp. 1281–1295, Jun. 2013, doi: 10.1016/j.cell.2013.04.053.

[28] C. P. Bacher *et al.*, "Transient colocalization of X-inactivation centres accompanies the initiation of X inactivation.," *Nat. Cell Biol.*, vol. 8, no. 3, pp. 293–9, Mar. 2006, doi: 10.1038/ncb1365.

[29] N. Xu, C. L. Tsai, and J. T. Lee, "Transient homologous chromosome pairing marks the onset of X inactivation," *Science (80-.).*, vol. 311, no. 5764, pp. 1149–1152, 2006, doi: 10.1126/science.1122984.

[30] O. Masui *et al.*, "Live-cell chromosome dynamics and outcome of X chromosome pairing events during ES cell differentiation," *Cell*, vol. 145, no. 3, pp. 447–458, 2011, doi: 10.1016/j.cell.2011.03.032.

[31] T. Pollex and E. Heard, "Nuclear positioning and pairing of X-chromosome inactivation centers are not primary determinants during initiation of random X-inactivation," *Nat. Genet.*, vol. 51, no. 2, pp. 285–295, 2019, doi: 10.1038/s41588-018-0305-7.

[32] J. A. Croft, J. M. Bridger, S. Boyle, P. Perry, P. Teague, and W. A. Bickmore, "Differences in the localization and morphology of chromosomes in the human nucleus," *J. Cell Biol.*, vol. 145, no. 6, pp. 1119–1131, 1999, doi: 10.1083/jcb.145.6.1119.

[33] F. A. Habermann *et al.*, "Arrangements of macro- and microchromosomes in chicken cells," *Chromosom. Res.*, vol. 9, no. 7, pp. 569–584, 2001, doi: 10.1023/A:1012447318535.

[34] N. Ganai, S. Sengupta, and G. I. Menon, "Chromosome positioning from activity-based segregation," *Nucleic Acids Res.*, vol. 42, no. 7, pp. 4145–4159, 2014, doi: 10.1093/nar/gkt1417.

[35] A. Awazu, "Nuclear dynamical deformation induced hetero- and euchromatin positioning," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 92, no. 3, pp. 1–5, 2015, doi: 10.1103/PhysRevE.92.032709.

[36] S. S. Lee, S. Tashiro, A. Awazu, and R. Kobayashi, "A new application of the phase-field method for understanding the mechanisms of nuclear architecture reorganization," *J. Math. Biol.*, vol. 74, no. 1–2, pp. 333–354, 2017, doi: 10.1007/s00285-016-1031-3.

[37] K. Takao *et al.*, "Torsional turning motion of chromosomes as an accelerating force to align

homologous chromosomes during meiosis," *J. Phys. Soc. Japan*, vol. 88, no. 2, pp. 1–5, 2019, doi: 10.7566/JPSJ.88.023801.

[38] S. Boyle, S. Gilchrist, J. M. Bridger, N. L. Mahy, J. A. Ellis, and W. A. Bickmore, "The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells," *Hum. Mol. Genet.*, vol. 10, no. 3, pp. 211–219, 2001, doi: 10.1093/hmg/10.3.211.

[39] H. Tanabe *et al.*, "Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 7, pp. 4424–4429, 2002, doi: 10.1073/pnas.072618599.

[40] R. Mayer, A. Brero, J. von Hase, T. Schroeder, T. Cremer, and S. Dietzel, "Common themes and cell type specific variations of higher order chromatin arrangements in the mouse," *BMC Cell Biol.*, vol. 6, pp. 1–22, 2005, doi: 10.1186/1471-2121-6-44.

[41] M. Neusser, V. Schubel, A. Koch, T. Cremer, and S. Müller, "Evolutionarily conserved, cell type and species-specific higher order chromatin arrangements in interphase nuclei of primates," *Chromosoma*, vol. 116, no. 3, pp. 307–320, 2007, doi: 10.1007/s00412-007-0099-3.

[42] I. Solovei *et al.*, "LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation," *Cell*, vol. 152, no. 3, pp. 584–598, 2013, doi: 10.1016/j.cell.2013.01.009.

[43] D. Koehler *et al.*, "Changes of higher order chromatin arrangements during major genome activation in bovine preimplantation embryos," *Exp. Cell Res.*, vol. 315, no. 12, pp. 2053–2063, 2009, doi: 10.1016/j.yexcr.2009.02.016.

[44] K. Finan, P. R. Cook, and D. Marenduzzo, "Non-specific (entropic) forces as major determinants of the structure of mammalian chromosomes," *Chromosom. Res.*, vol. 19, no. 1, pp. 53–61, 2011, doi: 10.1007/s10577-010-9150-y.

[45] I. Solovei *et al.*, "Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution," *Cell*, vol. 137, no. 2, pp. 356–368, 2009, doi: 10.1016/j.cell.2009.01.052.

[46] S. Asakura and F. Oosawa, "On interaction between two bodies immersed in a solution of macromolecules," *J. Chem. Phys.*, vol. 22, no. 7, pp. 1255–1256, 1954, doi: 10.1063/1.1740347.

[47] D. Marenduzzo, K. Finan, and P. R. Cook, "The depletion attraction: An underappreciated force driving cellular organization," *J. Cell Biol.*, vol. 175, no. 5, pp. 681–686, 2006, doi: 10.1083/jcb.200609066.

[48] F. Zosel, A. Soranno, K. J. Buholzer, D. Nettels, and B. Schuler, "Depletion interactions modulate the binding between disordered proteins in crowded environments," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 24, pp. 13480–13489, 2020, doi: 10.1073/pnas.1921617117.

[49] N. P. Adhikari, R. Auhl, and E. Straube, “Interfacial properties of flexible and semiflexible polymers,” *Macromol. Theory Simulations*, vol. 11, no. 3, pp. 315–325, 2002, doi: 10.1002/1521-3919(20020301)11:3<315::AID-MATS315>3.0.CO;2-6.

[50] S. A. Egorov, A. Milchev, A. Nikoubashman, and K. Binder, “Phase Separation and Nematic Order in Lyotropic Solutions: Two Types of Polymers with Different Stiffnesses in a Common Solvent,” *J. Phys. Chem. B*, vol. 125, no. 3, pp. 956–969, 2021, doi: 10.1021/acs.jpcb.0c10411.

[51] A. Milchev, S. A. Egorov, J. Midya, K. Binder, and A. Nikoubashman, “Entropic Unmixing in Nematic Blends of Semiflexible Polymers,” *ACS Macro Lett.*, vol. 9, no. 12, pp. 1779–1784, 2020, doi: 10.1021/acsmacrolett.0c00668.

[52] S. Fujishiro and M. Sasai, “Generation of dynamic three-dimensional genome structure through phase separation of chromatin,” *bioRxiv*, p. 2021.05.06.443035, 2021, [Online]. Available: https://www.biorxiv.org/content/10.1101/2021.05.06.443035v1%0Ahttps://www.biorxiv.org/content/10.1101/2021.05.06.443035v1.abstract

[53] S. Nakahata, T. Komoto, M. Fujii, and A. Awazu, “Mathematical model of chromosomal dynamics during DNA double strand break repair in budding yeast,” *Biophys. Physicobiology*, pp. 1–12, 2022, doi: 10.2142/biophysico.bppb-v19.0012.

[54] M. Matsushita *et al.*, “Dynamic changes in the interchromosomal interaction of early histone gene loci during development of sea urchin,” *J. Cell Sci.*, vol. 130, no. 24, pp. 4097–4107, 2017, doi: 10.1242/jcs.206862.

[55] A. Takemoto *et al.*, “Cilia play a role in breaking left-right symmetry of the sea urchin embryo,” *Genes to Cells*, vol. 21, no. 6, pp. 568–578, 2016, doi: 10.1111/gtc.12362.

[56] S. Yaguchi and H. Katow, “Expression of Tryptophan 5-hydroxylase gene during sea urchin neurogenesis and role of serotonergic nervous system in larval behavior,” *J. Comp. Neurol.*, vol. 466, no. 2, pp. 219–229, 2003, doi: 10.1002/cne.10865.

[57] S. Kinjo, M. Kiyomoto, T. Yamamoto, K. Ikeo, and S. Yaguchi, “HpBase: A genome database of a sea urchin, Hemicentrotus pulcherrimus,” *Dev. Growth Differ.*, vol. 60, no. 3, pp. 174–182, 2018, doi: 10.1111/dgd.12429.

[58] H. Takagi *et al.*, “Nucleosome exclusion from the interspecies-conserved central AT-rich region of the Ars insulator,” *J. Biochem.*, vol. 151, no. 1, pp. 75–87, 2012, doi: 10.1093/jb/mvr118.

[59] K. Akasaka *et al.*, “Upstream element of the sea urchin arylsulfatase gene serves as an insulator.,” *Cell. Mol. Biol. (Noisy-le-grand).*, vol. 45, no. 5, pp. 555–65, Jul. 1999, [Online].

Available: http://www.ncbi.nlm.nih.gov/pubmed/10512188

[60] M. Nicodemi and A. Prisco, "Symmetry-breaking model for X-chromosome inactivation.," *Phys. Rev. Lett.*, vol. 98, no. 10, p. 108104, Mar. 2007, doi: 10.1103/PhysRevLett.98.108104.

[61] M. Nicodemi and A. Prisco, "Self-assembly and DNA binding of the blocking factor in X chromosome inactivation," *PLoS Comput. Biol.*, vol. 3, no. 11, pp. 2135–2142, 2007, doi: 10.1371/journal.pcbi.0030210.

[62] A. Scialdone and M. Nicodemi, "Mechanics and dynamics of X-chromosome pairing at X inactivation," *PLoS Comput. Biol.*, vol. 4, no. 12, pp. 1–7, 2008, doi: 10.1371/journal.pcbi.1000244.

[63] A. Scialdone and M. Nicodemi, "DNA loci cross-talk through thermodynamics," *J. Biomed. Biotechnol.*, vol. 2009, 2009, doi: 10.1155/2009/516723.

[64] A. Scialdone, I. Cataudella, M. Barbieri, A. Prisco, and M. Nicodemi, "Conformation regulation of the X chromosome inactivation center: A model," *PLoS Comput. Biol.*, vol. 7, no. 10, 2011, doi: 10.1371/journal.pcbi.1002229.

[65] V. Bianco, A. Scialdone, and M. Nicodemi, "Colocalization of multiple DNA loci: A physical mechanism," *Biophys. J.*, vol. 103, no. 10, pp. 2223–2232, 2012, doi: 10.1016/j.bpj.2012.08.056.

[66] V. Mutzel *et al.*, "A symmetric toggle switch explains the onset of random X inactivation in different mammals," *Nat. Struct. Mol. Biol.*, vol. 26, no. 5, pp. 350–360, 2019, doi: 10.1038/s41594-019-0214-1.

[67] T. Cremer *et al.*, "The Interchromatin Compartment Participates in the Structural and Functional Organization of the Cell Nucleus," *BioEssays*, vol. 42, no. 2, pp. 1–18, 2020, doi: 10.1002/bies.201900132.

[68] D. Baú *et al.*, "The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules," *Nat. Struct. Mol. Biol.*, vol. 18, no. 1, pp. 107–115, 2011, doi: 10.1038/nsmb.1936.

[69] Z. Duan *et al.*, "A three-dimensional model of the yeast genome," *Nature*, vol. 465, no. 7296, pp. 363–367, 2010, doi: 10.1038/nature08973.

[70] H. Tanizawa *et al.*, "Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation," *Nucleic Acids Res.*, vol. 38, no. 22, pp. 8164–8177, 2010, doi: 10.1093/nar/gkq955.

[71] N. Varoquaux, F. Ay, W. S. Noble, and J. P. Vert, "A statistical approach for inferring the 3D structure of the genome," *Bioinformatics*, vol. 30, no. 12, pp. i26–i33, 2014, doi: 10.1093/bioinformatics/btu268.

[72] M. Falk *et al.*, "Heterochromatin drives compartmentalization of inverted and conventional nuclei," *Nature*, vol. 570, no. 7761, pp. 395–399, 2019, doi: 10.1038/s41586-019-1275-3.

[73] C.-K. Chen *et al.*, "Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing," *Science (80-.).*, vol. 354, no. 6311, pp. 468–472, Oct. 2016, doi: 10.1126/science.aae0047.

[74] M. Liu *et al.*, "Multiplexed imaging of nucleome architectures in single cells of mammalian tissue," *Nat. Commun.*, vol. 11, no. 1, pp. 1–14, 2020, doi: 10.1038/s41467-020-16732-5.

[75] T. Ikeda *et al.*, "Srf destabilizes cellular identity by suppressing cell-type-specific gene expression programs," *Nat. Commun.*, vol. 9, no. 1, 2018, doi: 10.1038/s41467-018-03748-1.

[76] S. E. Johnstone *et al.*, "Large-Scale Topological Changes Restrain Malignant Progression in Colorectal Cancer," *Cell*, vol. 182, no. 6, pp. 1474–1489.e23, Sep. 2020, doi: 10.1016/j.cell.2020.07.030.

[77] Q. Zhou *et al.*, "Spatiotemporal Dynamics of Dilute Red Blood Cell Suspensions in Low-Inertia Microchannel Flow," *Biophys. J.*, vol. 118, no. 10, pp. 2561–2573, May 2020, doi: 10.1016/j.bpj.2020.03.019.

[78] A. Lappala *et al.*, "Four-dimensional chromosome reconstruction elucidates the spatiotemporal reorganization of the mammalian X chromosome.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 42, p. e2107092118, Oct. 2021, doi: 10.1073/pnas.2107092118.

[79] E. J. Banigan, A. A. van den Berg, H. B. Brandão, J. F. Marko, and L. A. Mirny, "Chromosome organization by one-sided and two-sided loop extrusion," *Elife*, vol. 9, pp. 1–46, 2020, doi: 10.7554/eLife.53558.

[80] E. H. Davidson *et al.*, "A genomic regulatory network for development," *Science (80-.).*, vol. 295, no. 5560, pp. 1669–1678, 2002, doi: 10.1126/science.1069883.

[81] P. Oliveri and E. H. Davidson, "Gene regulatory network controlling embryonic specification in the sea urchin," *Curr. Opin. Genet. Dev.*, vol. 14, no. 4, pp. 351–360, 2004, doi: 10.1016/j.gde.2004.06.004.

[82] Y. Matsushima, N. Sakamoto, and A. Awazu, "Insulator Activities of Nucleosome-Excluding DNA Sequences without Bound Chromatin Looping Proteins," *J. Phys. Chem. B*, vol. 123, no. 5, pp. 1035–1043, 2019, doi: 10.1021/acs.jpcb.8b10518.

[83] E. Sodergren *et al.*, "The Genome of the Sea Urchin Strongylocentrotus purpuratus," *Science (80-.).*, vol. 314, no. 5801, pp. 941–952, Nov. 2006, doi: 10.1126/science.1133609.

[84] P. L. Davidson *et al.*, "Chromosomal-level genome assembly of the sea urchin Lytechinus variegatus substantially improves functional genomic analyses," *Genome Biol. Evol.*, vol. 12,

no. 7, pp. 1080–1086, 2020, doi: 10.1093/GBE/EVAA101.

[85] S. Kinjo, M. Kiyomoto, H. Suzuki, T. Yamamoto, K. Ikeo, and S. Yaguchi, "TrBase: A genome and transcriptome database of Temnopleurus reevesii," *Dev. Growth Differ.*, vol. 64, no. 4, pp. 210–218, 2022, doi: 10.1111/dgd.12780.

[86] F. Marletaz, A. Couloux, and J. Poulain, "Analysis of the P. lividus sea urchin genome highlights contrasting trends of genomic and regulatory evolution in deuterostomes," *Cell Genomics*, 2023, doi: 10.1016/j.xgen.2023.100295.

[87] Y. Iuchi, J. Morokuma, K. Akasaka, and H. Shimada, "Detection and characterization of the cis-element in the first intron of the Ars gene in the sea urchin," *Dev. Growth Differ.*, vol. 37, no. 4, pp. 373–378, Aug. 1995, doi: 10.1046/j.1440-169X.1995.t01-3-00003.x.

[88] N. Sakamoto, K. Akasaka, K. Mitsunaga-Nakatsubo, K. Takata, T. Nishitani, and H. Shimada, "Two Isoforms of Orthodenticle-Related Proteins (HpOtx) Bind to the Enhancer Element of Sea Urchin Arylsulfatase Gene," *Dev. Biol.*, vol. 181, no. 2, pp. 284–295, Jan. 1997, doi: 10.1006/dbio.1996.8455.

[89] R. Vaser and M. Šikić, "Raven: a de novo genome assembler for long reads," *bioRxiv*, p. 2020.08.07.242461, 2021, [Online]. Available: https://www.biorxiv.org/content/10.1101/2020.08.07.242461v2%0Ahttps://www.biorxiv.org/content/10.1101/2020.08.07.242461v2.abstract

[90] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, "Assembly of long, error-prone reads using repeat graphs," *Nat. Biotechnol.*, vol. 37, no. 5, pp. 540–546, 2019, doi: 10.1038/s41587-019-0072-8.

[91] J. Ruan and H. Li, "Fast and accurate long-read assembly with wtdbg2," *Nat. Methods*, vol. 17, no. 2, pp. 155–158, 2020, doi: 10.1038/s41592-019-0669-3.

[92] J. Lang, "MAECI: A pipeline for generating consensus sequence with nanopore sequencing long-read assembly and error correction," *PLoS One*, vol. 17, no. 5 May, pp. 1–9, 2022, doi: 10.1371/journal.pone.0267066.

[93] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[94] M. Manni, M. R. Berkeley, M. Seppey, and E. M. Zdobnov, "BUSCO: Assessing Genomic Data Quality and Beyond," *Curr. Protoc.*, vol. 1, no. 12, 2021, doi: 10.1002/cpz1.323.

[95] B. Bushnell, J. Rood, and E. Singer, "BBMerge – Accurate paired shotgun read merging via overlap," *PLoS One*, vol. 12, no. 10, pp. 1–15, 2017, doi: 10.1371/journal.pone.0185056.

[96] T. Brůna, K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, "BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database," *NAR Genomics Bioinforma.*, vol. 3, no. 1, pp. 1–11, 2021, doi: 10.1093/nargab/lqaa108.

[97] G. Pertea and M. Pertea, "GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]," *F1000Research*, vol. 9, no. 304, pp. 1–20, 2020, [Online]. Available: https://f1000research.com/articles/9-304/v2

[98] J. Morokuma, K. Akasaka, K. Mitsunaga-Nakatsubo, and H. Shimada, "A cis-regulatory element within the 5' flanking region of arylsulfatase gene of sea urchin, Hemicentrotus pulcherrimus," *Dev. Growth Differ.*, vol. 39, no. 4, pp. 469–476, Aug. 1997, doi: 10.1046/j.1440-169X.1997.t01-3-00008.x.

[99] T. Yamamoto, K. Akasaka, S. Irie, and H. Shimada, "A long polypyrimidine: Polypurine sequence in 5' flanking region of arylsulfatase gene of sea urchin embryo," *Int. J. Dev. Biol.*, vol. 38, no. 2, pp. 337–344, 1994.

[100] N. Sakamoto, K. Akasaka, T. Yamamoto, and H. Shimada, "A Triplex DNA Structure of the Polypyrimidine: Polypurine Stretch in the 5′ Flanking Region of the Sea Urchin Arylsulfatase Gene," *Zoolog. Sci.*, vol. 13, no. 1, pp. 105–109, Feb. 1996, doi: 10.2108/zsj.13.105.

[101] K. Akasaka *et al.*, "Corrected Structure of the 5′ Flanking Region of Arylsulfatase Gene of the Sea Urchin, Hemicentrotus pulcherrimus 5′ flanking sequence/sea urchin/arylsulfatase gene/G-string," *Dev. Growth Differ.*, vol. 36, no. 6, pp. 633–636, 1994, doi: 10.1111/j.1440-169X.1994.00633.x.

[102] P. E. Warburton, J. Giordano, F. Cheung, Y. Gelfand, and G. Benson, "Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeated that contain testes genes," *Genome Res.*, vol. 14, no. 10 A, pp. 1861–1869, 2004, doi: 10.1101/gr.2542904.

[103] S. Sawaya *et al.*, "Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements," *PLoS One*, vol. 8, no. 2, 2013, doi: 10.1371/journal.pone.0054710.

[104] M. D. Vinces, M. Legendre, M. Caldara, M. Hagihara, and K. J. Verstrepen, "Unstable tandem repeats in promoters confer transcriptional evolvability," *Science (80-.).*, vol. 324, no. 5931, pp. 1213–1216, 2009, doi: 10.1126/science.1170097.

[105] C. A. Horton *et al.*, "Short tandem repeats bind transcription factors to tune eukaryotic gene expression," *Science (80-.).*, vol. 381, no. 6664, p. 2022.05.24.493321, Sep. 2023, doi:

10.1126/science.add1250.

[106] G. Benson, “Tandem repeats finder: a program to analyze DNA sequences,” *Nucleic Acids Res.*, vol. 27, no. 2, pp. 573–580, Jan. 1999, doi: 10.1093/nar/27.2.573.

[107] R. J. Britten, A. Cetta, and E. H. Davidson, “The single-copy DNA sequence polymorphism of the sea urchin strongylocentrotus purpuratus,” *Cell*, vol. 15, no. 4, pp. 1175–1186, Dec. 1978, doi: 10.1016/0092-8674(78)90044-2.

[108] T. Yamamoto, R. Kawamoto, T. Fujii, N. Sakamoto, and T. Shibata, “DNA variations within the sea urchin Otx gene enhancer,” *FEBS Lett.*, vol. 581, no. 27, pp. 5234–5240, Nov. 2007, doi: 10.1016/j.febslet.2007.10.013.

[109] K. Watanabe *et al.*, “The crucial role of <scp>CTCF</scp> in mitotic progression during early development of sea urchin,” *Dev. Growth Differ.*, vol. 65, no. 7, pp. 395–407, Sep. 2023, doi: 10.1111/dgd.12875.

[110] S. Chen, Y. Zhou, Y. Chen, and J. Gu, “Fastp: An ultra-fast all-in-one FASTQ preprocessor,” *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, 2018, doi: 10.1093/bioinformatics/bty560.

[111] G. Holley *et al.*, “Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly.,” *Genome Biol.*, vol. 22, no. 1, p. 28, 2021, doi: 10.1186/s13059-020-02244-4.

[112] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype,” *Nat. Biotechnol.*, vol. 37, no. 8, pp. 907–915, 2019, doi: 10.1038/s41587-019-0201-4.

[113] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009, doi: 10.1093/bioinformatics/btp352.

補足情報

注） 下記 Table のうち、Table S1～S5 及び Table S11～S12 は https://github.com/Komoto-Te/doctoral_thesis/ に掲載している。

Table S1 MDS2 法を用いて推定された染色体ポリマーモデルの基本立体構造（重心を原点とした空間座標データ（μm））及び各ポリマーに対応するゲノム上の領域、エピゲノム状態、Boundary score を示す。

Table S2 NMDS 法を用いて推定された染色体ポリマーモデルの基本立体構造（重心を原点とした空間座標データ（μm））及び各ポリマーに対応するゲノム上の領域、エピゲノム状態、Boundary score を示す。

Table S3 PM1 法を用いて推定された染色体ポリマーモデルの基本立体構造（重心を原点とした空間座標データ（μm））及び各ポリマーに対応するゲノム上の領域、エピゲノム状態、Boundary score を示す。

Table S4 PM2 法を用いて推定された染色体ポリマーモデルの基本立体構造（重心を原点とした空間座標データ（μm））及び各ポリマーに対応するゲノム上の領域、エピゲノム状態、Boundary score を示す。

Table S5 染色体粗視化粒子鎖モデルの基本立体構造（重心を原点とした空間座標データ（μm））及び各粒子に対応するゲノム上の領域、エピゲノム状態、粒子半径を示す。

Table S6 本研究で得られたドラフトゲノム配列（FASTA 形式）(a) と、Raven、Flye、Wtdbg2 それぞれによるアセンブル結果の評価 (b)。

(a) “HpulGenome_kure_v1_contig.fa”
（https://cell-innovation.nig.ac.jp/cgi-bin/Hpul_public/Hpul_annot_download.cgi にてダウンロードできる）

(b)

	Raven	Flye	Wtdbg2
Assembly size*	619.2 Mb	985.5 Mb	851.0 Mb

No. contigs*	2,164	22,260	16,967
N50 contig length*	508.4kb	144.8 kb	240.3 kb
No. scaffolds*	2,164	22,175	16,967
N50 scaffold length*	508.4 kb	146.1 kb	240.3 kb
N (%)*	0	0	0
GC-content (%)*	36.91	37.36	36.94
BUSCO completeness (%) (metazoan_odb10: 954 genes)	Complete : 96.1 Duplicated : 6.5 Fragmented : 1.8 Missing : 1.8	Complete : 96.6 Duplicated : 37.5 Fragmented : 2.6 Missing : 0.8	Complete : 90.3 Duplicated : 10.4 Fragmented : 3.8 Missing : 5.9
Mapping ratio of transcriptome models (%) (20,564 sequences)	71.93 (aligned exactly 1 time) 3.12 (aligned >1 times)	69.48 (aligned exactly 1 time) 8.00 (aligned >1 times)	59.26 (aligned exactly 1 time) 1.85 (aligned >1 times)

Table S7 ドラフトゲノムから得られた遺伝子モデルとそれらの情報（それぞれ https://cell-innovation.nig.ac.jp/cgi-bin/Hpul_public/Hpul_annot_download.cgi にてダウンロードできる）。
各遺伝子モデルの塩基配列（a）とアミノ酸配列（b）（FASTA 形式）、各遺伝子の位置と属性をまとめた GTF（Gene Transfer Format）（c）、推定遺伝子モデルのアノテーション表（d）。

(a) HpulGenome_kure_v1_nucl.fa

(b) HpulGenome_kure_v1_prot.fa

© HpulGenome_kure_v1.gtf

(d) HpulGenome_kure_v1_annot.xlsx

Table S8 初期型ヒストン遺伝子のタンデムリピート探索の結果。初期型ヒストン遺伝子の塩基配列（a）、HpulGenome_v1 及び更新ドラフトゲノムにおける初期型ヒストン遺伝子と相同な領域の位置（b-d）。HpulGenome_v1（b）と更新されたドラフトゲノム（c）のそれぞれの BLASTN 検索で得られた結果と、初期型ヒストン遺伝子のロングタンデムリピートの位置（HSP が最も多かった 2 つの contig のみ表示している）（d）。

(a)

>EarlyHistone
GTCGACTTGCTAGAAGGGGTGGTGTCAAGAGGATCTCTGGTCTCATCTACGAAGAGACACGTGGTGTACTGAAGGTCTTCCTGGAGAATGTCATCCGTGATGCAGT
CACCTACTGCGAGCACGCCAAGCGAAAGACTGTCACAGCCATGGACGTTGTGTATGCACTGAAGAGACAGGGTCGTACATTGTACGGCTTCGGCGGCTAACTGTAG
CAAACCTCTCTCTCCTGGCTAGAATAACAAACGGCTCTTTTCAGAGCACCAAATAATCAAGAAAGAATACTGTTGTATGTTAATTCCGTGAAAGTAAAGAAAGAAG

AAGAAGAAATAAGCCGTCAATATCAAGCAAAACGAATAACGGCCCAAGAAGAGGCGATAGTGATTGGTATAGAAAGATGAAAGGCAAATCATGAAACAGAAATGAA
TGGATGAATGAATGAATGAATGAAATGCAAGGAGGGTGTGTCTATTCCAAAAGAAAGGCAAATTATGAAACAGAAAGCAGTAACAAATTGAGTGATAAATAGTGTA
TAGACATTGGCATTTGGAGTATGAGGATCCTTATTTATATTTTCAACTGAAATTGAAAGATAAGTATGATTTATTTACAAAAGTAAATGTAGATGCTAGTAACTAA
AAAGCCTACTCTGTGATGAAAGATGTGAAATGGATAAATGAAAAATACAAAATACATGTTGTTTTGCACAATGATAACGCGTTCATGGTTTATAATGATAATACCT
AGTATGTAATCGCACAGCGCTTTATATATATATATATATATATATATATATATATACATACATACATACATACATACATACATATATCTTATATATCTTATA
GAGGTACTCTTTTTTCATTTGCTGTAACCAAATAATTTTGAATTACATTGTCAAAGATAGTTCAATGCCCATTGATGTTATAATCATTACGTATATTATTATTGTT
TATATTATGTGAGAAGAAATGGAAAAATAAATTTGCTCTTTATAATTATATATGGAGTTGCACCACTACCACTACTATAATTATATATCTCCGAGATAATGTC
AAAAGTAGGAGTTAATCAAAACCTGTGCAAGTATGTCTTGGATAAAACGTTTCGACATGTTCCATATTCGGATATTCTAATTCCAATCAAAATCTCCATTTCATCA
ACTTTTTCATATTTCTTATTCTTTGGCGATTGAATCGAACATGGCAGAGAGATCCTATATTAGAACATAGGCCCGTATGATCTGTATGGGTTGTCACTATGTCGCC
ATCTCTAGGCAGGGGATGGTACAGACACTAATTGTGCGACGCCTACGATCAATGAAAAGATCGAGACCGAGGCTCATTTGCATAGTCGGACCGCATCATACGGATC
CGGCCCCGTGTATAAAAAGGAATGGTTCTTGCTGACCATTCACAGTATCCAAAGCATATTTGCCTGAAGTACTCGTTTCACTGCATCTTTACAGACCAGAAAACCT
CAAATCATCATGGCTCCAACAGCTCAAGTTGCTAAGAAAGGCTCCAAGAAGGCAGTCAAGGCCCCTCGGCCCAGTGGTGGCAAGAAGAGGAACAGGAAAAGGAAGG
AGAGTTATGGAATCTACATCTACAAAGTCCTCAAGCAGGTTCATCCAGATACCGGCATCTCCAGCCGGGCCATGATCATCATGAACAGCTTCGTCAACGACATCTT
CGAGCGAATTGCCGGCGAATCTTCCCGCCTCTCTCAGTACAACAAAAAGTCAACCATCAGCAGTCGCGAGATTCAGACCGCCGTCCGCCTCATTCTCCCCGGAGAG
CTGGCAAAGCACGCTGTGAGCGAGGGTACCAAGGCAGTGACGAAATACACTACCTCCAAGTAGACCGATCATATCCTGCTCTAATTGGACATACAACGGCCCTTTT
CAGGGCCACACAAATAATCAAGAAAGAAAGCATGATACCCGTTGTTTTTTGTTTTTGTTTTTTGTTTTTGTTTTTTGAATGAATGAATAAATGAATGAATGAATGA
ATGAATGAATGAATGAATGAATGAATAAATAAATAATGCGTATAATGAATTAAGTGAGACTTGCAAAAGATTACACAATACTATACAATAATATTAGCCATACATA
TTATTATTATACATCTCACTCTATAAATCTAGACAGGGGCGAAACGTGTAGTCACGATGGTTGGAAGGAAGAAGAAAATAGTCGATTACAAAAAAAAAAAAAGAAG
AGAATTAGAAGAAATACAATTCAAGGTGGCAACAAGTCCGTAATAATATACGTCCATGCATCCTTCCTTGAAAGGTTGACGAGATGCAGGTTAAGTTGATAGTTTT
TTTTTCTTTCTTTTCTCTCATGATTCATGCATCCCCACCAATCCACCTCCCCCCCCCCCCTCCCTCCCTCCCTCCCTGCCTGCCTGCCTGCCT
TCCTCACTACTCTCAGGATCAGCCGAGCAGACTCTCGGTCAAACATCTTAATCATCATCTGTTAGAATGCATATTCTCGTAATTATAATTTGGCAACACTGCGTGC
GTGAGCCGAGTTCTCGGCTTGCCATCTAACCCATGTTCAAAACGGTGGTAGCGCCACCTTTCGACACATTCTATTGACTCTGCACATACGGCATCGCCAAGCCCCG
ATCCCGTCCCGCGCTAAACAAAAGAACATCCCGGTTGGCCAATCAAGAGAGCTTTACAAACGTAGGGCGACCACACCAGGATCCTTCGCAGCTCATATAAATAGCT
GAAGATTGCCAGTGGTTTTCAATCATCCCGTCACTCGTATTTGAAGCACTGAACCTACTGTCTCTCAAGCAACTATGGCACGCACCAAGCAGACCGCTCGCAAATC
TACAGGAGGGAAGGCTCCCCGCAAGCAGCTGGCAACCAAAGCTGCCAGAAAGAGTGCCCCCGCCACTGGAGGAGTCAAGAAGCCTCATCGATACAGGCCTGGCACA
GTCGCCCTGAGAGAGATTCGCCGCTACCAGAAGAGCACTGAGCTTCTCATCCGAAAACTGCCATTCCAGCGTcTagTGCGTGAGATTGCACAGGACTTCAAGACCG
AGCTACGTTTCCAGAGTTCCGCTGTGATGGCCCTTCAAGAAGCCAGCGAGGCATACCTAGTTGGCCTCTTTGagGACACCAACCTGTGTGCCATCCACGCCAAGAG
GGTTACCATCATGCCCAAAGACATCCAGCTCGCCCGTCGAATCCGCGGAGAACGCGCCTAGAACCATCGGTACCGCATCTAGCCCATGCACCGCATACACAAACGG
CTCTTTTCAGAGCCACCACAACCCCAAGAAAGAATCACTGAGTTTTAAAATGAACACGATGCAtgAaaaTcGACGCTGGGTTTAAACTAATTATATGTATGAAATA
GATTTCAAAAAGAAAGAAAAAGAAGAAGAAGGAAAAAAAAAAAAAAAAACGACCACTGTACAATATGCGTAGAACATAATATGATACAAATGTAGAACACAATAAGTG
TTTGTTAGGAATTAAAAAGTAAGGAATATTCAAATATAAATCTATGTAATCTAGGAATGACCTCCTCAGTGTTCGTGTGCCACATGAAACAAAGCTATTGGGAATA
TGTGATGAAACATATATATTGATCTCTGAAATATAACCTAGAACCTGACACTATAGGAAATAGTCGCAGTGTACTTGGAATGGGCTATATTTTTATCACATCTTTC
AGGCGTTGTAGTCACATCCCTTTGAATTGATCTGTGCTTTTTTTTTTTTTTTTTTCATTTAAGAAACCCCATACAATCTGATCGAGTACTAATGGTGATATCAGTG
GAGAAAACGGATAAGTCGCCGCGAGATGGCACCATTCACTAATGTTGCGAAACAAGTTTAAGGAGATCCGGTCAGCTACCCGCTGATTGGACAATAGTCACAATGC
CCCCCGGCGGTCCTGCGATCCTAACACCACGTATAAATAGCCAGCAAAACGCTGCTGGACATCCATTCAAGTCAGCGAACATTGTTACGTTTTGAACTTCGCTCCC
GATTTATTCTCAACTCATCCAAAATCATCATGTCTGGCAGAGGAAAGAGTGGAAAGGCCCGCACCAAGGCAAAGACGCGCTCCTCCCGTGCAGGGCTCCAGTTTCC
AGTGGGACGTGTTCATCGGTTTCTCCGAAAGGGCAACTATGCAAAGAGGGTCGGCGGTGGAGCTCCTGTCTACATGGCCGCCGTTCTAGAGTACCTTACTGCCGAA
ATCTTGGAACTCGCAGGCAACGCTGCCCGCGACAACAAGAAATCGAGAATCATTCCACGCCACCTCCAACTCGCCGTGCGTAATGATGAAGAACTCAACAAGCTTT
TGGGTGGGGTGACGATCGCTCAAGGTGGTGTCCTGCCCAACATCCAAGCCGTGCTTCTTCCCAAGAAAACCGCTAAATCAAGCTAGATCGAGTTTGCTCCCGGCAA
ATCTTGAAACCTCAACGGCCCTTATCAGGGCCACCAATTACTCACGAAAGAATTGTATCCTTTATGAATCCCCTTCACCACTCCTCCCCTCTCTCTCTATCTCTCT
CTCTCTCTCCCTCTCTCCTCCTTAGTTTGTAACGATACACCGAAGATAAAATAAGCTTTAATAATGGTTATATATATGATTGATATATTAAATAAAGATATATACA
CCCACAGATGTTATGAACGAAGTGTGTgATTGCAGTAAAAGAAAAAGCAAGTTAATTTAAAATACATATCTACAGTTATGAGAACGGATATTAGGGTGACCAAAAT
ATTATCTTTTGAGTCAGCATTTTAATGGGAAACTATGAATATGCTATTTCCCCTGCATTTGACTGTGTTTCAAGAGGGGTGACGATCAGGGGTTGATGGTGTTATC
GCCACCTATTAAATGTGCCTTACATTTTCCCCCTTCAAATTCATATAATGCCGGTGCACATTGTACATGGTAGAAGAGGATGAAGAAAAAAGGAAGGTGAAAAGAA
AACTGAGGAGGAGAAGGTGGTGGTGGAATAAGAGCATGTATGGATGGAAAAGAAGGGAGAAAAGTTTCTGAAAACAAACACAAACTGGCACGCAGTCATGGGGGCG
GACGACCCGGGATTGTCTCCCCCCACGTACGCAACCATGCCGTATATCGATGGCGTTGCCGAGCTGATAGTTATTCGTTTTGTTAACCTCCCGACGCACCGTATAT
CAAGATGGCTGAGAAGGAAAGCTCTAAGAAGATGACTACAAAGAAGCCGGCTGCCCACCCACCGGCTGCCGAGATGGTTGCTACAGCAATCACCGAGTTGAAGGAC
CGAAATGGCTCCTCGCTGCAAGCAATAAAGAAGTATATCGCAACCAATTTCGATGTGCAGATGGACCGACAGCTGCTATTCATCAAGCGGGCCCTAAAGTCTGGCG
TGGAGAAAGGCAAACTTGTGCAGACGAAGGGGAAAGGAGCTTCGGGTTCTTTCAAGCTAAATGTGCAGGCGGCCAAGGCAGAGGCGTCGGAGAAGGCCAAGAAGGA
GAAGGAGAAGGCAAAACAGCTAGCACAGCGTGAGAAGGCCAAGGAAAAGGCTGCAGCGAAGAAGGAGAAACTGCAGAAGGCAGCCGCCgCAAAGAAAGTCAAGGCA
GCCCCCAAGAAAGCGAAGAAGCCAGTAAAGAAAACGACTGAGAAGAAAAGAGAAGAAGAAGACTCCAAAGAAGGTAACCAAGAAGCCAGCAGCCAAGAAATCAACAC
CAAAGAAGACACCCAAGAAGGCAGCCGCAAAGAAACCCAAGACTGCCAAGCCCAAGAAGCCTGCGGCAAAGAAGGCTGCAAAGTCCAAGTGATTTGCACGCCAACT
TTTCCATCCTACCAAAACGGCTCTTTTCAGAGcCACCACATACCCAAGAAAGAACACTTGTCCAAATGCACTGATGAGAAACTATGATTTTGTGTAATGAAGAAGA
AGAAGAAAGAAAGAGAGAGAAAAAAAAAggggGGGGGGAAGAGGAGCAACAAGAATgGGATGaATTATaAAAATAGATTCATGGAAAGTGATATTAGCCCAAACAG
AACTGGTTTCAGTTGAAGCCCTTTTGAAATAAAAACAAACTATATATAATACAGTCCGAAGAAAAACAAATTAGAATAAGAAATGTGAATAAGACAAAAATGAACA
TATGAATCATATACGACCAAAAAAAAAAAGAAAAAAAAAGAAGAAAAGAATAAGAAATGAAGGATGAGAGAGAATGAGGGGCTTGGGTGAGGAATGCATTCTTATT
GAATAATGTAGGTGATAACCGAtTTTTAAGTTAATGAACTTTCGTCATCCATGTGTACTGTTGGATAAGAGTCCTCAGCCTCTGCTATTTCAACTCGAGCATTCAA
CATAGGCCTCTCTAGATACATGCACGTACTGTGCTAGCGAATACTCGCCACGAGGGGGCGCACTCGAACGGGGAGTCTCCGCACTCCAGTCCCGCACACCGAATGA
TGCCGAATCTCGTCACCCAAGTCCGCAATGGTGTTACAATTGTCGCTGCAATCCGTCAGGGCATCATTCGCTTAGCGTAATATCCAGTCTACGGGATCACACAAAA
CTCGCTCTCAACTATCAATCATCACCATGTCAGGTCGAGGAAAAGGAGGAAAGGGACTCGGAAAGGGTGGTGCCAAACGTCATCGCAAGGTTCTACGAGACAACAT
CCAGGGCATCACAAAGCCTGCAATCCGTCGAC

(b)

Hit Scaffold ID	Nb. HSPs	Score	E Value	Identity	Query Start Position	Query End Position	Query Coverage	Hit Start Position	Hit End Position	Hit coverage
scaffold1876	1	331	2E−172	95.3%	4000	4384	5.7%	54543	54158	0.4%
scaffold402	1	251	5E−128	87.6%	1420	1820	6%	86555	86956	0.2%

scaffold2002	1	237	3E-120	98%	1598	1849	3.8%	63100	63351	0.3%
scaffold678	1	210	3E-105	87.1%	2049	2419	5.5%	148598	148935	0.2%
scaffold3802	1	201	3E-100	99%	1537	1743	3.1%	30209	30415	0.6%
scaffold3838	2	199	4E-099	99%	2976	3180	3.1%	34321	34117	0.6%
scaffold3838	2	54	2E-018	100%	90	143	0.8%	33988	34041	0.2%
scaffold2842	2	169	2E-082	97.8%	6530	6710	2.7%	4907	5088	0.3%
scaffold2842	2	72	2E-028	100%	1617	1688	1.1%	44201	44130	0.1%
scaffold2214	1	163	4E-079	98.8%	1688	1856	2.5%	11909	11741	0.2%
scaffold597	2	162	2E-078	96.2%	6528	6710	2.7%	65289	65471	0.1%
scaffold597	2	52	2E-017	98.2%	1	55	0.8%	65466	65520	0%
scaffold1497	1	161	6E-078	95.3%	5774	5963	2.8%	89015	88830	0.2%
scaffold1281	2	160	2E-077	91.3%	4238	4456	3.3%	74323	74537	0.2%
scaffold1281	2	42	8E-012	97.8%	5299	5343	0.7%	32769	32813	0%
scaffold1100	1	141	7E-067	96.2%	1443	1601	2.4%	13893	13735	0.1%
scaffold2178	1	140	3E-066	98%	5235	5383	2.2%	36726	36578	0.2%
scaffold875	1	122	3E-056	95.2%	1	141	2.1%	70616	70760	0.1%
scaffold5617	3	121	1E-055	98.4%	1566	1692	1.9%	6643	6517	0.8%
scaffold5617	3	92	1E-039	98%	1596	1693	1.5%	6657	6754	0.6%
scaffold5617	3	58	1E-020	100%	5290	5347	0.9%	6815	6758	0.4%
scaffold4010	1	115	2E-052	100%	1420	1534	1.7%	31018	30904	0.3%
scaffold321	2	110	1E-049	96.7%	2945	3066	1.8%	156302	156423	0%
scaffold321	2	63	2E-023	98.5%	3126	3192	1%	156242	156307	0%
scaffold7914	1	109	4E-049	100%	1595	1703	1.6%	533	425	1.9%
scaffold487	1	108	2E-048	96.7%	5231	5350	1.8%	121854	121735	0.1%
scaffold1527	1	107	6E-048	100%	1638	1744	1.6%	80417	80523	0.1%
scaffold2906	1	91	5E-039	97%	4184	4283	1.5%	42596	42695	0.2%
scaffold228	1	89	6E-038	98.9%	2775	2867	1.4%	210391	210300	0%
scaffold2283	1	87	8E-037	97.9%	1593	1685	1.4%	1358	1265	0.1%
scaffold2508	1	86	3E-036	100%	1492	1577	1.3%	22870	22955	0.1%
scaffold616	1	86	3E-036	92.9%	3735	3846	1.7%	22400	22508	0.1%
scaffold598	2	86	3E-036	89.1%	152	280	1.9%	44330	44204	0%
scaffold598	2	56	1E-019	98.3%	72	130	0.9%	44387	44329	0%
scaffold1549	1	85	1E-035	98.9%	1665	1752	1.3%	49996	50083	0.1%
scaffold5313	1	83	1E-034	97.8%	4083	4171	1.3%	17310	17398	0.5%
scaffold1207	1	83	1E-034	97.8%	1609	1697	1.3%	1788	1876	0.1%
scaffold4826	2	81	2E-033	92.4%	4309	4413	1.6%	17470	17574	0.5%
scaffold4826	2	52	2E-017	88.1%	4231	4314	1.3%	13120	13199	0.3%
scaffold104	1	78	8E-032	88.7%	3403	3523	1.8%	87040	86922	0%
scaffold1711	1	77	3E-031	97.6%	5272	5354	1.2%	59353	59271	0.1%
scaffold936	1	76	1E-030	97.6%	1671	1752	1.2%	44365	44284	0.1%
scaffold3425	1	74	1E-029	97.5%	3525	3604	1.2%	23382	23461	0.2%
scaffold208	1	73	5E-029	98.7%	1607	1682	1.1%	170817	170892	0%
scaffold2643	1	72	2E-028	94.3%	1651	1737	1.3%	41312	41226	0.1%
scaffold50	1	71	6E-028	89.4%	1649	1752	1.5%	314552	314655	0%
scaffold851	1	70	2E-027	97.4%	3990	4065	1.1%	112796	112721	0%
scaffold843	1	70	2E-027	98.6%	3988	4060	1.1%	28558	28630	0%
scaffold417	1	70	2E-027	100%	4898	4967	1%	119487	119418	0%
scaffold2156	1	69	8E-027	100%	3524	3592	1%	61505	61573	0.1%

scaffold2037	1	67	1E-025	84.7%	2831	2954	1.8%	27315	27192	0.2%
scaffold362	1	66	4E-025	92%	3131	3217	1.3%	211040	211126	0%
scaffold83	1	66	4E-025	97.2%	1679	1750	1.1%	397396	397325	0%
scaffold2287	1	65	1E-024	95.9%	5159	5232	1.1%	24441	24368	0.1%
scaffold4985	1	62	6E-023	98.5%	5291	5355	1%	2098	2034	0.3%
scaffold591	1	61	2E-022	98.4%	3999	4062	1%	139569	139632	0%
scaffold2035	2	60	8E-022	98.4%	4066	4128	0.9%	74685	74623	0.1%
scaffold2035	2	41	3E-011	94.1%	4038	4088	0.8%	74734	74685	0.1%
scaffold546	1	60	8E-022	93.3%	3986	4060	1.1%	30652	30578	0%
scaffold2042	1	59	3E-021	98.4%	1632	1693	0.9%	44122	44183	0.1%
scaffold926	1	59	3E-021	96.9%	6529	6593	1%	37732	37796	0%
scaffold447	1	59	3E-021	87.5%	18	112	1.4%	235363	235458	0%
scaffold4409	1	58	1E-020	95.5%	4224	4290	1%	16676	16742	0.2%
scaffold4128	2	58	1E-020	98.4%	6531	6591	0.9%	4746	4686	0.2%
scaffold4128	2	46	5E-014	100%	3159	3204	0.7%	4745	4790	0.1%
scaffold1166	1	58	1E-020	96.9%	260	324	1%	11445	11382	0%
scaffold2065	1	57	4E-020	100%	4364	4420	0.8%	54815	54871	0.1%
scaffold1312	1	57	4E-020	96.8%	4232	4294	0.9%	72053	72115	0.1%
scaffold838	1	57	4E-020	96.8%	1679	1741	0.9%	101034	100972	0%
scaffold8112	1	55	5E-019	98.3%	6533	6590	0.9%	4905	4848	1.1%
scaffold3691	2	55	5E-019	94%	1675	1741	1%	7442	7508	0.2%
scaffold3691	2	43	2E-012	88.1%	1675	1741	1%	27805	27871	0.2%
scaffold1372	1	55	5E-019	98.3%	5289	5346	0.9%	93674	93617	0.1%
scaffold1146	1	55	5E-019	98.3%	4232	4289	0.9%	111611	111668	0%
scaffold377	1	55	5E-019	95.4%	3189	3252	1%	150206	150270	0%
scaffold257	1	55	5E-019	98.3%	1893	1950	0.9%	93723	93666	0%
scaffold1340	1	54	2E-018	95.2%	4220	4282	0.9%	83161	83223	0.1%
scaffold1206	1	54	2E-018	94.1%	82	147	1%	63187	63121	0.1%
scaffold306	1	54	2E-018	90.8%	5278	5351	1.1%	76013	75938	0%
scaffold2611	1	53	6E-018	92.6%	4004	4071	1%	27134	27067	0.1%
scaffold953	1	53	6E-018	93.9%	1680	1744	1%	45700	45635	0%
scaffold387	1	53	6E-018	95.2%	4000	4061	0.9%	174494	174433	0%
scaffold151	1	53	6E-018	98.2%	5295	5350	0.8%	321740	321795	0%
scaffold55	1	53	6E-018	96.6%	1622	1680	0.9%	459640	459698	0%
scaffold3551	1	52	2E-017	96.6%	4232	4289	0.9%	18806	18863	0.1%
scaffold3408	1	52	2E-017	93.8%	4226	4290	1%	35778	35715	0.2%
scaffold960	1	52	2E-017	95.1%	4232	4292	0.9%	112279	112339	0%
scaffold905	1	52	2E-017	98.2%	5293	5347	0.8%	134093	134039	0%
scaffold26	2	52	2E-017	96.6%	3027	3084	0.9%	261776	261719	0%
scaffold26	2	41	3E-011	97.7%	5045	5088	0.7%	310691	310648	0%
scaffold4677	1	51	8E-017	95%	5268	5327	0.9%	10680	10739	0.2%
scaffold2388	1	51	8E-017	95%	2926	2985	0.9%	60479	60420	0.1%
scaffold1587	1	51	8E-017	92.6%	82	147	1%	37549	37615	0.1%
scaffold574	1	51	8E-017	92.6%	82	147	1%	89316	89250	0%
scaffold431	1	51	8E-017	92.6%	82	147	1%	242305	242371	0%
scaffold141	1	51	8E-017	92.4%	5456	5521	1%	114963	114898	0%
scaffold1705	1	50	3E-016	100%	3152	3201	0.7%	78928	78879	0.1%
scaffold1378	1	50	3E-016	100%	3200	3249	0.7%	88511	88560	0%

scaffold364	1	50	3E-016	100%	345	394	0.7%	112873	112824	0%
scaffold129	1	50	3E-016	89.2%	4068	4141	1.1%	697	770	0%
scaffold7	2	50	3E-016	98.1%	4239	4291	0.8%	159986	160038	0%
scaffold7	2	47	1E-014	94.6%	4232	4287	0.8%	469107	469052	0%
scaffold2990	1	49	1E-015	94.8%	4145	4202	0.9%	38206	38149	0.1%
scaffold2890	1	49	1E-015	96.4%	4292	4346	0.8%	24277	24331	0.1%
scaffold992	1	49	1E-015	94.9%	5293	5350	0.9%	71387	71445	0%
scaffold955	1	49	1E-015	94.8%	1454	1511	0.9%	34632	34689	0%
scaffold21	1	49	1E-015	98.1%	5299	5350	0.8%	297604	297655	0%
scaffold1987	1	48	4E-015	96.3%	1648	1701	0.8%	60632	60579	0.1%
scaffold1393	1	48	4E-015	98%	4294	4344	0.8%	93148	93098	0%
scaffold717	1	48	4E-015	94.7%	5177	5233	0.8%	68824	68768	0%
scaffold135	1	48	4E-015	96.4%	5298	5350	0.8%	194509	194455	0%
scaffold4	1	48	4E-015	92.1%	3998	4060	0.9%	199068	199006	0%
scaffold1140	1	46	5E-014	93.2%	3058	3115	0.9%	45284	45226	0%
scaffold376	1	46	5E-014	100%	1648	1693	0.7%	95735	95690	0%
scaffold330	1	46	5E-014	98%	5299	5347	0.7%	240963	240915	0%
scaffold4737	1	45	2E-013	100%	3153	3197	0.7%	19775	19731	0.2%
scaffold1272	1	45	2E-013	85.9%	4094	4171	1.2%	16540	16617	0.1%
scaffold298	2	45	2E-013	97.9%	4231	4278	0.7%	216061	216014	0%
scaffold298	2	41	3E-011	94%	1467	1516	0.7%	188794	188843	0%
scaffold4906	1	44	6E-013	92.9%	5335	5390	0.8%	7016	7071	0.3%
scaffold1253	1	44	6E-013	90.3%	4004	4065	0.9%	119714	119775	0.1%
scaffold15	1	44	6E-013	97.9%	5563	5609	0.7%	376880	376926	0%
scaffold4716	1	43	2E-012	94.2%	1614	1665	0.8%	12513	12564	0.2%
scaffold4028	1	43	2E-012	97.8%	4014	4059	0.7%	12207	12162	0.1%
scaffold3681	1	43	2E-012	95.9%	4248	4296	0.7%	5728	5680	0.1%
scaffold2536	1	43	2E-012	97.8%	1661	1706	0.7%	12587	12632	0.1%
scaffold1431	1	43	2E-012	100%	4017	4059	0.6%	6447	6405	0%
scaffold941	1	43	2E-012	100%	4290	4332	0.6%	144104	144146	0%
scaffold412	1	43	2E-012	97.8%	5303	5348	0.7%	196219	196264	0%
scaffold5687	1	42	8E-012	100%	4011	4052	0.6%	1754	1795	0.3%
scaffold1917	1	42	8E-012	95.8%	4127	4174	0.7%	24217	24264	0.1%
scaffold965	1	42	8E-012	95.8%	5296	5343	0.7%	28955	28908	0%
scaffold537	1	42	8E-012	95.8%	4297	4344	0.7%	35400	35447	0%
scaffold244	1	42	8E-012	100%	1666	1707	0.6%	159396	159437	0%
scaffold100	1	42	8E-012	97.8%	2292	2336	0.7%	187639	187683	0%

(c)

Hit Contig ID	Nb. HSPs	Score	E Value	Identity	Query Start Position	Query End Position	Query Coverage	Hit Start Position	Hit End Position	Hit coverage
Utg198178	47	5429	0	97.3%	816	6710	87.9%	25238	31161	1.2%
Utg198178	47	5408	0	97.2%	816	6710	87.9%	38589	44518	1.2%
Utg198178	47	5380	0	97%	816	6710	87.9%	31914	37835	1.2%
Utg198178	47	5342	0	96.9%	816	6710	87.9%	51951	57831	1.2%
Utg198178	47	5322	0	96.7%	816	6710	87.9%	45272	51197	1.2%
Utg198178	47	3741	0	95.5%	2398	6710	64.3%	20138	24488	0.9%
Utg198178	47	2238	0	90.9%	3622	6710	46%	88738	91852	0.6%
Utg198178	47	2153	0	90%	3625	6710	46%	61420	64548	0.7%
Utg198178	47	2125	0	89.8%	3622	6710	46%	68615	71731	0.7%
Utg198178	47	2120	0	89.7%	3622	6710	46%	82053	85190	0.7%
Utg198178	47	1851	0	89.5%	3998	6710	40.4%	75717	78481	0.6%
Utg198178	47	1322	0	90.8%	4859	6710	27.6%	16005	17846	0.4%
Utg198178	47	1284	0	95.3%	820	2323	22.4%	18606	20102	0.3%
Utg198178	47	1229	0	93.9%	816	2336	22.7%	58580	60074	0.3%
Utg198178	47	938	0	90.4%	1016	2336	19.7%	86160	87501	0.3%
Utg198178	47	925	0	90.1%	1016	2336	19.7%	72697	74042	0.3%
Utg198178	47	913	0	89.6%	1016	2336	19.7%	79448	80817	0.3%
Utg198178	47	848	0	90.6%	2428	3624	17.8%	87528	88710	0.2%
Utg198178	47	846	0	90.5%	2428	3624	17.8%	60101	61291	0.2%
Utg198178	47	842	0	90.8%	2448	3624	17.5%	80858	82025	0.2%
Utg198178	47	832	0	90.2%	2428	3624	17.8%	74069	75249	0.2%
Utg198178	47	795	0	99.4%	816	1625	12.1%	93289	92480	0.2%
Utg198178	47	792	0	94.9%	2428	3364	14%	67016	67950	0.2%
Utg198178	47	747	0	98.2%	1	793	11.8%	31156	31940	0.2%
Utg198178	47	747	0	98.2%	1	793	11.8%	37830	38615	0.2%
Utg198178	47	747	0	98.2%	1	793	11.8%	44513	45298	0.2%
Utg198178	47	747	0	98.2%	1	793	11.8%	51192	51977	0.2%
Utg198178	47	733	0	97.7%	1	793	11.8%	57826	58606	0.2%
Utg198178	47	730	0	97.6%	1	793	11.8%	24483	25264	0.2%
Utg198178	47	726	0	94.8%	1016	1869	12.7%	65518	66392	0.2%
Utg198178	47	664	0	94.5%	1	798	11.9%	17841	18634	0.2%
Utg198178	47	604	0	98.4%	1	638	9.5%	91847	92477	0.1%
Utg198178	47	572	0	92.2%	1	754	11.2%	64543	65293	0.2%
Utg198178	47	572	0	92.2%	1	754	11.2%	71726	72476	0.2%
Utg198178	47	569	0	92%	1	754	11.2%	78476	79226	0.2%
Utg198178	47	564	0	91.8%	1	754	11.2%	85185	85936	0.2%
Utg198178	47	500	0	97.4%	248	793	8.1%	93801	93263	0.1%
Utg198178	47	311	3E-161	96.6%	3009	3354	5.2%	68009	68357	0.1%
Utg198178	47	241	2E-122	91.7%	3622	3942	4.8%	75277	75600	0.1%
Utg198178	47	130	1E-060	77.3%	1875	2336	6.9%	66496	66989	0.1%
Utg198178	47	100	5E-044	84.5%	832	1026	2.9%	85933	86129	0%
Utg198178	47	97	2E-042	84%	832	1026	2.9%	65290	65486	0%
Utg198178	47	97	2E-042	84%	832	1026	2.9%	72473	72669	0%
Utg198178	47	97	2E-042	84%	832	1026	2.9%	79223	79419	0%

Utg198178	47	73	5E-029	92.1%	1728	1828	1.5%	66395	66487	0%
Utg198178	47	53	7E-018	96.6%	1545	1603	0.9%	364140	364198	0%
Utg198178	47	41	3E-011	97.7%	3953	3996	0.7%	75635	75678	0%
Utg200276	34	5352	0	96.9%	816	6710	87.9%	54998	60917	4.6%
Utg200276	34	3924	0	95.7%	816	5326	67.2%	76574	81098	3.5%
Utg200276	34	3104	0	97.7%	3386	6710	49.6%	72471	75809	2.6%
Utg200276	34	3029	0	97.5%	3452	6710	48.6%	65442	68726	2.6%
Utg200276	34	2719	0	96%	3622	6710	46%	51140	54244	2.4%
Utg200276	34	2205	0	96.5%	816	3293	36.9%	61676	64135	1.9%
Utg200276	34	2167	0	90.2%	3622	6710	46%	44475	47585	2.4%
Utg200276	34	2129	0	89.8%	3622	6710	46%	37799	40923	2.4%
Utg200276	34	1873	0	96.3%	816	2946	31.8%	83290	85373	1.6%
Utg200276	34	1336	0	97.5%	5273	6710	21.4%	81087	82536	1.1%
Utg200276	34	1207	0	95.9%	2020	3397	20.5%	70637	72020	1.1%
Utg200276	34	1123	0	98.1%	816	2009	17.8%	69480	70669	0.9%
Utg200276	34	955	0	90.8%	1016	2336	19.7%	48543	49881	1%
Utg200276	34	931	0	90.3%	1016	2336	19.7%	41892	43224	1%
Utg200276	34	921	0	90%	1016	2336	19.7%	35218	36560	1.1%
Utg200276	34	907	0	91.9%	2428	3624	17.8%	49908	51112	0.9%
Utg200276	34	861	0	90.8%	2428	3624	17.8%	43251	44447	0.9%
Utg200276	34	858	0	90.8%	2428	3624	17.8%	36587	37771	0.9%
Utg200276	34	750	0	98.4%	1	793	11.8%	54239	55024	0.6%
Utg200276	34	747	0	98.2%	1	793	11.8%	68721	69506	0.6%
Utg200276	34	747	0	98.2%	1	793	11.8%	82531	83316	0.6%
Utg200276	34	735	0	97.6%	1	793	11.8%	60912	61702	0.6%
Utg200276	34	723	0	97%	1	793	11.8%	75804	76600	0.6%
Utg200276	34	566	0	91.8%	1	754	11.2%	40918	41673	0.6%
Utg200276	34	511	0	90.9%	1	706	10.5%	47580	48291	0.6%
Utg200276	34	345	3E-180	89.6%	248	754	7.6%	34492	34999	0.4%
Utg200276	34	325	4E-169	86.8%	3981	4527	8.2%	89524	90064	0.4%
Utg200276	34	100	5E-044	84.5%	832	1026	2.9%	41670	41866	0.2%
Utg200276	34	100	5E-044	84.5%	832	1026	2.9%	48321	48517	0.2%
Utg200276	34	97	2E-042	84%	832	1026	2.9%	34996	35192	0.2%
Utg200276	34	83	1E-034	88.3%	5453	5580	1.9%	92451	92578	0.1%
Utg200276	34	78	8E-032	95.6%	3321	3410	1.3%	64137	64227	0.1%
Utg200276	34	69	8E-027	94%	2688	2771	1.3%	85371	85454	0.1%
Utg200276	34	68	3E-026	94%	5576	5658	1.2%	92739	92821	0.1%
Utg198750	1	335	1E-174	98.3%	1539	1891	5.3%	520319	520671	0.1%
Utg196974	4	302	3E-156	98.7%	2775	3089	4.7%	315476	315164	0.1%
Utg196974	4	176	3E-086	98.9%	3121	3302	2.7%	315085	314903	0%
Utg196974	4	69	8E-027	98.6%	3130	3201	1.1%	205073	205002	0%
Utg196974	4	54	2E-018	89.5%	2995	3080	1.3%	315168	315092	0%
Utg198034	1	257	3E-131	88.1%	1420	1820	6%	190131	190532	0.1%
Utg200748	1	243	2E-123	98.8%	1598	1849	3.8%	115806	115555	0.1%
Utg202804	1	240	7E-122	98.4%	1598	1849	3.8%	8162	7911	0.3%
Utg197334	2	201	4E-100	81.9%	1855	2336	7.2%	79979	80438	0%
Utg197334	2	44	7E-013	90.6%	2428	2491	1%	80465	80525	0%
Utg196252	2	199	5E-099	99%	2976	3180	3.1%	91581	91377	0.1%

Utg196252	2	54	2E-018	100%	90	143	0.8%	91248	91301	0%
Utg196094	2	187	2E-092	98%	1539	1737	3%	578294	578492	0%
Utg196094	2	104	3E-046	95.1%	1770	1891	1.8%	578491	578612	0%
Utg196474	2	181	5E-089	100%	3987	4167	2.7%	1136855	1137035	0%
Utg196474	2	83	1E-034	92%	3735	3846	1.7%	1023000	1022893	0%
Utg200140	2	167	3E-081	97.3%	6528	6710	2.7%	36064	35883	0.1%
Utg200140	2	141	8E-067	98.6%	1	148	2.2%	35888	35743	0.1%
Utg197168	2	162	2E-078	96.2%	6528	6710	2.7%	283676	283494	0%
Utg197168	2	52	2E-017	98.2%	1	55	0.8%	283499	283445	0%
Utg197144	1	162	2E-078	97.7%	4313	4486	2.6%	395560	395733	0%
Utg196190	2	161	6E-078	95.3%	5774	5963	2.8%	83866	83681	0%
Utg196190	2	125	6E-058	95.9%	1	141	2.1%	260908	261052	0%
Utg196456	2	160	2E-077	91.3%	4238	4456	3.3%	268888	268674	0.1%
Utg196456	2	42	9E-012	97.8%	5299	5343	0.7%	311038	310994	0%
Utg198890	1	141	8E-067	96.2%	1443	1601	2.4%	418422	418264	0%
Utg197032	1	140	3E-066	98%	5235	5383	2.2%	44692	44544	0%
Utg202044	1	130	1E-060	90.9%	4316	4496	2.7%	114775	114595	0.2%
Utg196274	1	128	1E-059	98.5%	1608	1741	2%	80641	80507	0%
Utg200702	1	127	5E-059	88%	3788	3985	3%	87821	88020	0.1%
Utg196490	1	115	2E-052	100%	1420	1534	1.7%	16727	16841	0.1%
Utg198946	1	107	6E-048	95.9%	1623	1744	1.8%	228643	228522	0%
Utg198338	1	107	6E-048	100%	1638	1744	1.6%	278373	278479	0%
Utg198370	1	103	1E-045	100%	1594	1696	1.5%	162367	162265	0%
Utg202870	1	98	6E-043	94.9%	5161	5277	1.7%	16634	16749	0.3%
Utg196810	1	91	5E-039	97%	4184	4283	1.5%	11661	11760	0%
Utg197344	1	87	8E-037	96.9%	1590	1685	1.4%	87921	88017	0%
Utg199324	1	86	3E-036	98.9%	1743	1831	1.3%	125701	125789	0.1%
Utg196378	1	86	3E-036	89.1%	152	280	1.9%	1322104	1322230	0%
Utg201028	1	85	1E-035	100%	1668	1752	1.3%	55272	55188	0.1%
Utg196140	1	85	1E-035	95.9%	1756	1851	1.4%	1511879	1511976	0%
Utg200658	1	83	1E-034	97.8%	1609	1697	1.3%	243945	244033	0%
Utg196480	1	83	1E-034	97.8%	4083	4171	1.3%	257907	257995	0%
Utg196634	1	81	2E-033	100%	1767	1847	1.2%	887254	887174	0%
Utg200904	1	76	1E-030	97.6%	1671	1752	1.2%	104921	104840	0%
Utg201318	1	74	1E-029	97.5%	3525	3604	1.2%	225161	225240	0%
Utg199796	1	74	1E-029	100%	2994	3067	1.1%	468400	468327	0%
Utg199980	1	72	2E-028	100%	1617	1688	1.1%	269721	269650	0%
Utg198850	1	71	7E-028	98.6%	3550	3623	1.1%	23346	23419	0%
Utg201522	1	70	2E-027	98.6%	3988	4060	1.1%	13712	13784	0.1%
Utg198522	1	70	2E-027	100%	1613	1682	1%	363220	363289	0%
Utg197058	1	70	2E-027	97.4%	3990	4065	1.1%	266520	266445	0%
Utg197580	1	69	8E-027	91.4%	4190	4282	1.4%	149149	149057	0%
Utg196508	1	69	8E-027	100%	3524	3592	1%	336240	336172	0%
Utg197904	1	68	3E-026	94%	1664	1747	1.3%	218486	218568	0%
Utg197066	1	68	3E-026	88.5%	1649	1752	1.5%	179525	179628	0%
Utg197546	1	67	1E-025	98.6%	3988	4057	1%	5934	6003	0%
Utg197088	1	67	1E-025	84.7%	2831	2954	1.8%	292598	292475	0%
Utg199600	1	66	4E-025	92%	3131	3217	1.3%	14046	14132	0.1%

Utg198366	1	66	4E-025	98.6%	4304	4372	1%	300152	300220	0%
Utg197524	1	65	1E-024	95.9%	5159	5232	1.1%	594145	594218	0%
Utg196340	2	64	5E-024	97.1%	4222	4291	1%	531996	532065	0%
Utg196340	2	47	1E-014	94.6%	4232	4287	0.8%	834690	834635	0%
Utg199338	1	62	7E-023	100%	2866	2927	0.9%	113621	113682	0%
Utg196482	1	62	7E-023	100%	1631	1692	0.9%	1665887	1665826	0%
Utg199084	1	60	9E-022	93.3%	3986	4060	1.1%	523240	523314	0%
Utg198170	1	59	3E-021	87.5%	18	112	1.4%	275147	275242	0%
Utg196360	1	59	3E-021	97%	82	147	1%	132154	132090	0%
Utg199128	1	58	1E-020	96.9%	3189	3252	1%	377818	377882	0%
Utg196690	1	58	1E-020	95.5%	4224	4290	1%	285610	285676	0%
Utg199472	1	57	4E-020	96.8%	4232	4294	0.9%	49400	49462	0%
Utg196578	1	56	1E-019	94.1%	3011	3078	1%	796418	796351	0%
Utg201288	1	55	5E-019	92.9%	1675	1744	1%	24386	24317	0.1%
Utg198310	1	55	5E-019	93%	4335	4404	1%	429533	429464	0%
Utg197448	1	55	5E-019	98.3%	5289	5346	0.9%	198959	199016	0%
Utg196832	2	55	5E-019	98.3%	4232	4289	0.9%	359668	359611	0%
Utg196832	2	42	9E-012	95.8%	4012	4059	0.7%	359562	359609	0%
Utg196182	1	55	5E-019	94%	1671	1737	1%	29910	29844	0%
Utg199318	1	54	2E-018	94.1%	82	147	1%	36425	36359	0%
Utg197958	1	54	2E-018	95.2%	4220	4282	0.9%	275957	275895	0%
Utg197388	1	54	2E-018	92.8%	5278	5346	1%	256024	256092	0%
Utg196644	1	54	2E-018	96.7%	4231	4290	0.9%	188484	188425	0%
Utg198920	1	53	7E-018	98.2%	5295	5350	0.8%	94672	94727	0%
Utg198682	1	53	7E-018	93.9%	82	147	1%	96406	96470	0.1%
Utg196794	1	53	7E-018	100%	5298	5350	0.8%	270091	270143	0%
Utg196684	1	53	7E-018	93.9%	82	147	1%	89933	89997	0%
Utg196576	1	53	7E-018	96.6%	1622	1680	0.9%	214470	214412	0%
Utg198576	2	52	2E-017	96.6%	4232	4289	0.9%	729379	729436	0%
Utg198576	2	51	9E-017	92.6%	82	147	1%	1095737	1095803	0%
Utg197370	1	52	2E-017	96.6%	4232	4289	0.9%	6590	6647	0%
Utg197246	1	52	2E-017	95.1%	4232	4292	0.9%	188413	188473	0%
Utg196204	2	52	2E-017	96.6%	3027	3084	0.9%	346541	346484	0%
Utg196204	2	41	3E-011	97.7%	5045	5088	0.7%	393153	393110	0%
Utg197700	1	51	9E-017	95%	2926	2985	0.9%	417198	417139	0%
Utg197124	2	51	9E-017	92.6%	82	147	1%	1785438	1785504	0%
Utg197124	2	41	3E-011	97.7%	1650	1693	0.7%	1730618	1730661	0%
Utg196344	1	51	9E-017	95.1%	82	140	0.9%	370092	370032	0%
Utg196126	1	51	9E-017	92.6%	82	147	1%	1338959	1338893	0%
Utg198702	1	50	3E-016	89.2%	4068	4141	1.1%	565645	565718	0%
Utg198784	1	49	1E-015	98.1%	5299	5350	0.8%	147138	147189	0%
Utg196680	1	49	1E-015	94.9%	5293	5350	0.9%	537567	537509	0%
Utg196236	1	49	1E-015	94.8%	1454	1511	0.9%	361670	361727	0%
Utg196434	1	48	4E-015	98%	4294	4344	0.8%	791233	791183	0%
Utg196116	1	48	4E-015	89.9%	3992	4060	1%	562311	562379	0%
Utg197040	1	47	1E-014	100%	4156	4202	0.7%	197914	197868	0%
Utg198276	1	46	5E-014	93.2%	269	326	0.9%	103428	103371	0%
Utg198228	1	46	5E-014	100%	1648	1693	0.7%	514457	514412	0%

Utg197128	2	46	5E-014	93.2%	3058	3115	0.9%	284947	284889	0%
Utg197128	2	41	3E-011	94%	3152	3201	0.7%	411436	411387	0%
Utg196268	1	46	5E-014	98%	5299	5347	0.7%	41494	41446	0%
Utg196220	1	46	5E-014	96.2%	5296	5347	0.8%	4348	4399	0%
Utg203004	1	45	2E-013	100%	4243	4287	0.7%	20027	20071	0.2%
Utg197402	1	45	2E-013	84.1%	4097	4184	1.3%	204164	204079	0%
Utg197048	2	45	2E-013	100%	4231	4275	0.7%	137522	137566	0%
Utg197048	2	41	3E-011	94%	1467	1516	0.7%	168746	168697	0%
Utg200606	1	44	7E-013	89.2%	3993	4057	1%	107307	107243	0.1%
Utg197654	1	44	7E-013	90.3%	4004	4065	0.9%	441383	441444	0%
Utg196582	1	44	7E-013	92.9%	5335	5390	0.8%	223215	223160	0%
Utg196294	1	44	7E-013	97.9%	5563	5609	0.7%	441392	441438	0%
Utg201272	1	43	2E-012	100%	4017	4059	0.6%	22075	22033	0.1%
Utg201098	1	43	2E-012	92.7%	1614	1668	0.8%	187862	187808	0%
Utg200182	1	43	2E-012	97.8%	4014	4059	0.7%	267939	267984	0%
Utg198836	1	43	2E-012	97.8%	5303	5348	0.7%	831471	831426	0%
Utg198168	1	43	2E-012	95.9%	4248	4296	0.7%	136008	136056	0%
Utg198146	1	43	2E-012	94.2%	1614	1665	0.8%	55877	55928	0.1%
Utg196788	1	43	2E-012	100%	3158	3200	0.6%	17555	17597	0%
Utg198364	1	42	9E-012	95.8%	4297	4344	0.7%	7181	7134	0.1%
Utg196936	1	42	9E-012	97.8%	5299	5343	0.7%	307875	307919	0%
Utg196924	1	42	9E-012	95.8%	4127	4174	0.7%	249838	249791	0%
Utg196328	1	42	9E-012	95.8%	4297	4344	0.7%	36537	36490	0.1%
Utg196286	1	42	9E-012	95.8%	5296	5343	0.7%	277624	277577	0%

(d)

Hit Contig ID	Nb. HSPs	Score	E Value	Identity	Query Start Position	Query End Position	Query Coverage	Hit Start Position	Hit End Position	Hit coverage
Utg198178	47	5429	0	97.30%	816	6710	87.90%	25238	31161	1.20%
Utg198178	47	5408	0	97.20%	816	6710	87.90%	38589	44518	1.20%
Utg198178	47	5380	0	97%	816	6710	87.90%	31914	37835	1.20%
Utg198178	47	5342	0	96.90%	816	6710	87.90%	51951	57831	1.20%
Utg198178	47	5322	0	96.70%	816	6710	87.90%	45272	51197	1.20%
Utg198178	47	3741	0	95.50%	2398	6710	64.30%	20138	24488	0.90%
Utg198178	47	2238	0	90.90%	3622	6710	46%	88738	91852	0.60%
Utg198178	47	2153	0	90%	3625	6710	46%	61420	64548	0.70%
Utg198178	47	2125	0	89.80%	3622	6710	46%	68615	71731	0.70%
Utg198178	47	2120	0	89.70%	3622	6710	46%	82053	85190	0.70%
Utg198178	47	1851	0	89.50%	3998	6710	40.40%	75717	78481	0.60%
Utg198178	47	1322	0	90.80%	4859	6710	27.60%	16005	17846	0.40%
Utg198178	47	1284	0	95.30%	820	2323	22.40%	18606	20102	0.30%
Utg198178	47	1229	0	93.90%	816	2336	22.70%	58580	60074	0.30%
Utg198178	47	938	0	90.40%	1016	2336	19.70%	86160	87501	0.30%
Utg198178	47	925	0	90.10%	1016	2336	19.70%	72697	74042	0.30%
Utg198178	47	913	0	89.60%	1016	2336	19.70%	79448	80817	0.30%
Utg198178	47	848	0	90.60%	2428	3624	17.80%	87528	88710	0.20%
Utg198178	47	846	0	90.50%	2428	3624	17.80%	60101	61291	0.20%
Utg198178	47	842	0	90.80%	2448	3624	17.50%	80858	82025	0.20%
Utg198178	47	832	0	90.20%	2428	3624	17.80%	74069	75249	0.20%
Utg198178	47	795	0	99.40%	816	1625	12.10%	93289	92480	0.20%
Utg198178	47	792	0	94.90%	2428	3364	14%	67016	67950	0.20%
Utg198178	47	747	0	98.20%	1	793	11.80%	31156	31940	0.20%
Utg198178	47	747	0	98.20%	1	793	11.80%	37830	38615	0.20%
Utg198178	47	747	0	98.20%	1	793	11.80%	44513	45298	0.20%
Utg198178	47	747	0	98.20%	1	793	11.80%	51192	51977	0.20%
Utg198178	47	733	0	97.70%	1	793	11.80%	57826	58606	0.20%
Utg198178	47	730	0	97.60%	1	793	11.80%	24483	25264	0.20%
Utg198178	47	726	0	94.80%	1016	1869	12.70%	65518	66392	0.20%
Utg198178	47	664	0	94.50%	1	798	11.90%	17841	18634	0.20%
Utg198178	47	604	0	98.40%	1	638	9.50%	91847	92477	0.10%
Utg198178	47	572	0	92.20%	1	754	11.20%	64543	65293	0.20%
Utg198178	47	572	0	92.20%	1	754	11.20%	71726	72476	0.20%
Utg198178	47	569	0	92%	1	754	11.20%	78476	79226	0.20%
Utg198178	47	564	0	91.80%	1	754	11.20%	85185	85936	0.20%
Utg198178	47	500	0	97.40%	248	793	8.10%	93801	93263	0.10%
Utg198178	47	311	3E-161	96.60%	3009	3354	5.20%	68009	68357	0.10%
Utg198178	47	241	2E-122	91.70%	3622	3942	4.80%	75277	75600	0.10%
Utg198178	47	130	1E-60	77.30%	1875	2336	6.90%	66496	66989	0.10%
Utg198178	47	100	5E-44	84.50%	832	1026	2.90%	85933	86129	0%
Utg198178	47	97	2E-42	84%	832	1026	2.90%	65290	65486	0%
Utg198178	47	97	2E-42	84%	832	1026	2.90%	72473	72669	0%
Utg198178	47	97	2E-42	84%	832	1026	2.90%	79223	79419	0%

Utg198178	47	73	5E-29	92.10%	1728	1828	1.50%	66395	66487	0%
Utg198178	47	53	7E-18	96.60%	1545	1603	0.90%	364140	364198	0%
Utg198178	47	41	3E-11	97.70%	3953	3996	0.70%	75635	75678	0%
Utg200276	34	5352	0	96.90%	816	6710	87.90%	54998	60917	4.60%
Utg200276	34	3924	0	95.70%	816	5326	67.20%	76574	81098	3.50%
Utg200276	34	3104	0	97.70%	3386	6710	49.60%	72471	75809	2.60%
Utg200276	34	3029	0	97.50%	3452	6710	48.60%	65442	68726	2.60%
Utg200276	34	2719	0	96%	3622	6710	46%	51140	54244	2.40%
Utg200276	34	2205	0	96.50%	816	3293	36.90%	61676	64135	1.90%
Utg200276	34	2167	0	90.20%	3622	6710	46%	44475	47585	2.40%
Utg200276	34	2129	0	89.80%	3622	6710	46%	37799	40923	2.40%
Utg200276	34	1873	0	96.30%	816	2946	31.80%	83290	85373	1.60%
Utg200276	34	1336	0	97.50%	5273	6710	21.40%	81087	82536	1.10%
Utg200276	34	1207	0	95.90%	2020	3397	20.50%	70637	72020	1.10%
Utg200276	34	1123	0	98.10%	816	2009	17.80%	69480	70669	0.90%
Utg200276	34	955	0	90.80%	1016	2336	19.70%	48543	49881	1%
Utg200276	34	931	0	90.30%	1016	2336	19.70%	41892	43224	1%
Utg200276	34	921	0	90%	1016	2336	19.70%	35218	36560	1.10%
Utg200276	34	907	0	91.90%	2428	3624	17.80%	49908	51112	0.90%
Utg200276	34	861	0	90.80%	2428	3624	17.80%	43251	44447	0.90%
Utg200276	34	858	0	90.80%	2428	3624	17.80%	36587	37771	0.90%
Utg200276	34	750	0	98.40%	1	793	11.80%	54239	55024	0.60%
Utg200276	34	747	0	98.20%	1	793	11.80%	68721	69506	0.60%
Utg200276	34	747	0	98.20%	1	793	11.80%	82531	83316	0.60%
Utg200276	34	735	0	97.60%	1	793	11.80%	60912	61702	0.60%
Utg200276	34	723	0	97%	1	793	11.80%	75804	76600	0.60%
Utg200276	34	566	0	91.80%	1	754	11.20%	40918	41673	0.60%
Utg200276	34	511	0	90.90%	1	706	10.50%	47580	48291	0.60%
Utg200276	34	345	3E-180	89.60%	248	754	7.60%	34492	34999	0.40%
Utg200276	34	325	4E-169	86.80%	3981	4527	8.20%	89524	90064	0.40%
Utg200276	34	100	5E-44	84.50%	832	1026	2.90%	41670	41866	0.20%
Utg200276	34	100	5E-44	84.50%	832	1026	2.90%	48321	48517	0.20%
Utg200276	34	97	2E-42	84%	832	1026	2.90%	34996	35192	0.20%
Utg200276	34	83	1E-34	88.30%	5453	5580	1.90%	92451	92578	0.10%
Utg200276	34	78	8E-32	95.60%	3321	3410	1.30%	64137	64227	0.10%
Utg200276	34	69	8E-27	94%	2688	2771	1.30%	85371	85454	0.10%
Utg200276	34	68	3E-26	94%	5576	5658	1.20%	92739	92821	0.10%

Table S9 Ars-INV ホモログ探索の結果。Ars-INV の塩基配列 （a）、BLASTN 検索に使用した部分配列 （FASTA 形式） と、更新ドラフトゲノム中の Ars-INV ホモログの位置（c）。

(a)

>Ars-INV
ATGTCTACACAGGGCTCCCACTTAGCCGAGTTGGAGCCACGAGTTGTGGCGAGCTGGCGCGATCTTATGGCAAGCAGTTTTATCATTAAAACATATTTCTTGATAT
CACGAAATCGAAATCTTGATATCAAGAATTACCGCTGCTTTACAACGTAACAATTTTTCTTGATATCAAGAAATATGTTTAAATGATAAAACTGCTTGCCATAAGA
TCGAGCCAGCTCGCCACAACTCGAGGCTCCAACTCGGCTAAGTGGGAGCCCTGGGACATGAT

(b)

>Ars-INV_left
CACAGGGCTCCCACTTAGCCGAGTTGGAGCCACGAGTTGTGGCGAGCTGGCGCGATCTTATGGCAAGCAGTTTTATCATTAAAACATATTTCTTGATATCACGAAA

(c)

Contig ID	Left Start Position	Left End Position	Right Start Position	Right End Position	Matced Identity
Utg198086	523793	523902	523944	524053	92.73%
Utg200732	44166	44282	44331	44450	91.67%
Utg196186	657013	657111	657172	657263	84.85%
Utg196372	903929	904037	904060	904168	95.41%
Utg202116	11736	11845	11887	11996	91.82%
Utg197744	936866	936979	937024	937137	90.35%
Utg197516	34001	34116	34170	34287	86.55%
Utg196956	164236	164343	164345	164462	85.59%
Utg201764	36735	36834	36887	36988	86.27%

Table S10 Ars-DIR1 と Ars-DIR2 のホモログ探索の結果。Ars-DIR1 の各反復単位の塩基配列（a）と Ars-DIR2 の各反復単位の塩基配列（b)、 BLASTN 検索に使用した部分配列（FASTA 形式）、更新ドラフトゲノム中の Ars-DIR1 ホモログの位置（c）と更新ドラフトゲノム中の Ars-DIR2 ホモログの位置（d)。

(a)

>Ars-DIR1_1_-3440_-3329
CTGTTACGTTTCCTAAAACCATTAGAGAAAGAAGTCGATTTGTTTTACCGTTTTCGACTCTTAATCCTTGTCTGAATATCAAAAATGTATTTTGTCTGCTCCTTTTTCTTCA
>Ars-DIR1_2_-3220_-3109
CTTTTAAGTTTCATTAAATCATTAGAAAAAGAAGACGATTCGTTGTACCGTTTTGACTCGTAATCCTTGTCTGAATTTCAAAAAAATGTTTTTGTCTGCTCCTTTTTCTTCA

(b)

>Ars-DIR2_1_-3096_-2977
AGGGCCCATGGTGGAGTCGAGGGTGCGGAAACCCCTGATTCTCAGGGGTTTCAGACATTTAATTACGGTCAGATAAACATATTTGTATACCTTATTTGTGACACTGGCAGGACACTGATT

>Ars-DIR2_2_-2973_-2904
GTGAGGTCCAGTGGGCGGGGCTCCCAGAAGCCCAGGGGTTTTAAGGATTACATTACCGTCAGATTAGCAT

>Ars-DIR2_3_-2900_-2792
GTGGGGCCCAGGGGGCAGAGCCCCCGGAAGCTCAGGGGTTTAGGCATTTGATTACGGCCAGATGAACATATTTTCTTACTTTATTTGTAACACTGGCAGGACACTGATT

>Ars-DIR2_4_-2788_-2719
GTGAGGTCCAGCGGGCGGAGCCCCCAGAAAATCAACGGTTTTAGGCACTAAATTGCTGTCAGATGAGTAT

>Ars-DIR2_5_-2703_-2592
AGGGCCCTTGGTCGAGTCCCGGGAGCAGAACCCCTGTAAGCTCAGGGGTTTTTAGGCGTTTAATTACGGTCAGATGAGCAATTTTTGATACCTTAATTTGTTGTGACACTGG

(c)

Contig ID	Start	End	Strand	Repeat counts
Utg196502	922823	923180	+	2
Utg198366	231269	231612	+	2
Utg199178	431559	431933	+	2
Utg196850	10156	10527	−	2
Utg196966	820016	820366	−	2
Utg198702	549575	549908	−	2
Utg199090	419720	420076	−	2
Utg200732	47055	47386	−	2

(d)

Contig ID	Start	End	Strand	Repeat counts
Utg196090	25809	25991	+	2
Utg196090	34824	35005	+	2
Utg196094	92586	92900	+	2
Utg196168	857943	858123	+	2
Utg196180	2028889	2029069	+	2
Utg196212	224888	225066	+	2
Utg196218	432841	433220	+	3
Utg196264	47844	48268	+	3
Utg196306	170267	170448	+	2
Utg196332	8708	9214	+	3
Utg196366	388705	389022	+	2
Utg196366	551154	551450	+	2
Utg196420	857746	857938	+	2
Utg196440	124852	125353	+	4
Utg196472	49399	49773	+	3
Utg196502	923193	923695	+	4
Utg196578	391870	392063	+	2
Utg196586	354037	354412	+	3
Utg196602	911008	911310	+	2
Utg196608	52103	52407	+	2
Utg196628	54553	54858	+	2
Utg196634	2399	2701	+	2
Utg196642	2249247	2249622	+	3
Utg196662	328450	328751	+	2
Utg196724	22112	22431	+	2
Utg196724	193350	193655	+	2
Utg196742	631130	631435	+	2
Utg196790	107505	107882	+	3
Utg196848	548944	549279	+	2
Utg196848	568798	569132	+	2
Utg196850	85009	85340	+	2
Utg196874	93773	94139	+	2
Utg197006	34052	34439	+	3
Utg197012	508765	509069	+	2
Utg197026	665971	666348	+	2
Utg197100	1415777	1415969	+	2
Utg197124	200010	200343	+	3
Utg197180	656134	656437	+	2
Utg197216	553629	553939	+	2
Utg197234	366466	366646	+	2

Utg197342	20714	21011	+	2
Utg197344	9291	9471	+	2
Utg197468	267731	268195	+	3
Utg197474	19732	20106	+	3
Utg197492	30361	30666	+	2
Utg197510	173329	173659	+	3
Utg197520	67268	67536	+	2
Utg197554	11223	11528	+	2
Utg197554	101279	101597	+	2
Utg197554	192294	192614	+	3
Utg197608	15065	15440	+	3
Utg197608	749878	750181	+	2
Utg197614	55526	55846	+	3
Utg197620	14208	14711	+	4
Utg197620	143207	143510	+	2
Utg197648	127312	127491	+	2
Utg197662	859599	859781	+	2
Utg197708	445621	446006	+	2
Utg197738	41545	41726	+	2
Utg197796	381098	381600	+	3
Utg197816	144852	145375	+	4
Utg197906	401040	401342	+	2
Utg198004	335364	335879	+	4
Utg198004	480082	480586	+	3
Utg198008	272449	272957	+	3
Utg198038	163039	163231	+	2
Utg198120	125976	126280	+	2
Utg198156	110396	110895	+	3
Utg198158	297263	297639	+	3
Utg198158	326382	326686	+	2
Utg198162	85987	86291	+	2
Utg198252	263420	263796	+	3
Utg198288	682062	682367	+	2
Utg198392	543513	543693	+	2
Utg198572	934956	935256	+	2
Utg198664	12275	12652	+	3
Utg198664	450063	450364	+	2
Utg198688	27849	28160	+	2
Utg198836	566738	567236	+	3
Utg198872	30330	30705	+	3
Utg198906	34765	34945	+	2
Utg199028	62013	62341	+	2

Utg199032	467714	468009	+	2
Utg199066	80649	80949	+	2
Utg199066	110470	110774	+	3
Utg199308	90441	90744	+	2
Utg199310	6602	6979	+	3
Utg199350	118042	118223	+	2
Utg199484	101359	101673	+	2
Utg199496	231304	231680	+	3
Utg199500	30021	30524	+	3
Utg199580	786222	786526	+	2
Utg199626	82467	82770	+	2
Utg199832	117877	118182	+	2
Utg200252	352268	352571	+	2
Utg200366	97269	97573	+	2
Utg200380	54081	54404	+	2
Utg200588	38702	39005	+	2
Utg200644	30990	31367	+	3
Utg201518	37318	37510	+	2
Utg201826	326322	326512	+	2
Utg203150	46560	46741	+	2
Utg196124	196463	196764	−	2
Utg196154	737833	738136	−	2
Utg196172	144136	144317	−	2
Utg196230	124513	124817	−	3
Utg196272	96569	96945	−	2
Utg196310	818504	818881	−	3
Utg196332	17020	17323	−	2
Utg196338	435552	435733	−	2
Utg196358	960726	961230	−	3
Utg196358	1027769	1028074	−	2
Utg196360	58112	58294	−	2
Utg196382	402087	402393	−	2
Utg196420	554267	554644	−	3
Utg196444	133824	134205	−	3
Utg196450	410361	410542	−	2
Utg196458	24158	24472	−	3
Utg196460	170217	170536	−	2
Utg196470	635824	636201	−	3
Utg196484	70286	70582	−	2
Utg196626	473875	474180	−	2
Utg196656	512413	512718	−	2
Utg196706	424280	424577	−	2

Utg196730	1357845	1358026	–	2
Utg196768	619562	619742	–	2
Utg196780	815701	815882	–	2
Utg196784	195782	196297	–	3
Utg196798	98066	98247	–	2
Utg196806	176411	176726	–	2
Utg196844	111722	112098	–	2
Utg196868	82722	83027	–	2
Utg196868	810542	810848	–	2
Utg196880	1005957	1006335	–	3
Utg196904	123947	124252	–	3
Utg196914	750108	750409	–	2
Utg196920	115148	115660	–	3
Utg196946	24924	25227	–	2
Utg197008	133595	133976	–	3
Utg197016	239315	239625	–	2
Utg197020	20437	20743	–	2
Utg197026	674567	674942	–	2
Utg197030	796904	797214	–	2
Utg197042	54908	55211	–	2
Utg197048	65695	65998	–	2
Utg197050	80719	81027	–	2
Utg197054	101008	101312	–	2
Utg197116	212371	212872	–	3
Utg197176	90753	91255	–	3
Utg197176	134684	134865	–	2
Utg197244	552718	553097	–	3
Utg197460	324175	324367	–	2
Utg197468	64153	64347	–	2
Utg197488	31505	31881	–	3
Utg197558	112289	112596	–	2
Utg197574	100640	100820	–	2
Utg197584	13652	13973	–	2
Utg197600	170279	170780	–	3
Utg197624	124209	124576	–	3
Utg197662	745160	745464	–	2
Utg197740	72983	73358	–	3
Utg197790	72574	72793	–	2
Utg197816	116125	116449	–	2
Utg197866	299817	300118	–	2
Utg197914	316051	316354	–	2
Utg198056	53442	53750	–	2

Utg198166	156932	157306	–	3
Utg198188	176611	176916	–	2
Utg198206	613052	613355	–	2
Utg198228	689187	689368	–	2
Utg198438	38779	39081	–	2
Utg198468	277830	278134	–	2
Utg198502	83326	83632	–	2
Utg198506	15510	15883	–	2
Utg198566	221732	222246	–	3
Utg198688	24106	24410	–	2
Utg198788	718639	718941	–	2
Utg198814	286331	286638	–	3
Utg198852	214291	214608	–	2
Utg198890	416088	416394	–	2
Utg198948	621128	621637	–	4
Utg199066	130198	130758	–	3
Utg199070	103998	104296	–	2
Utg199090	418964	419562	–	3
Utg199128	227185	227502	–	2
Utg199176	303122	303303	–	2
Utg199340	153787	154172	–	2
Utg199364	271318	271497	–	2
Utg199372	544895	545383	–	3
Utg199464	205629	205926	–	2
Utg199472	102380	102689	–	3
Utg199522	240791	241167	–	3
Utg199530	547723	548030	–	2
Utg199604	233874	234175	–	2
Utg199640	15166	15473	–	2
Utg199714	96028	96222	–	2
Utg200054	24836	25211	–	3
Utg200188	30104	30407	–	2
Utg200390	106459	106763	–	2
Utg200588	209104	209408	–	2
Utg200732	46538	47042	–	5
Utg200778	23679	23984	–	2
Utg201068	71039	71221	–	2
Utg201962	57546	57740	–	2
Utg202012	169693	170190	–	3

Table S11: ArsInsC ホモログ探索の結果。ArsInsC の塩基配列（FASTA 形式）(a) と更新ドラフトゲノム中の ArsInsC ホモログの位置 (b)。

(a)

>HpArsInsC
ACATGTAAGCATCTCAAGAAGCATATTTCTTGCCTGGCTGTTAATTTACAAACGCATAAAAAAAATATAATTTACTAAAGAATGAGGAAAAATCTCGGGAAGTTAT
GTAATTTCAGCATTATGTGTAAACCACCGTTATGGAATAAGAAATAAACCACATTTCAATTTATTTCCCCCGAGCC

(b)

Contig ID	Start Position	End Position	Query Coverage	Identity	Score	E Value
Utg196094	93022	93204	100%	91.80%	137	3E−66
Utg196098	639474	639294	100%	91.80%	136	1E−65
Utg196124	196174	195993	100%	90.20%	128	3E−61
Utg196126	693771	693952	100%	90.70%	131	7E−63
Utg196140	966890	966709	100%	91.30%	134	1E−64
Utg196152	208820	208640	100%	92.90%	142	5E−69
Utg196168	34215	34394	100%	90.70%	130	2E−62
Utg196168	768441	768622	100%	93.40%	146	3E−71
Utg196172	143871	143690	100%	92.30%	140	7E−68
Utg196180	1114137	1113956	100%	92.30%	140	7E−68
Utg196184	42417	42238	100%	92.30%	139	2E−67
Utg196192	82809	82629	100%	92.30%	139	2E−67
Utg196212	225313	225497	100%	93.50%	147	9E−72
Utg196230	124389	124209	100%	91.20%	133	5E−64
Utg196230	186540	186361	100%	90.10%	127	1E−60
Utg196230	1016930	1017109	100%	91.80%	136	1E−65
Utg196248	779088	779268	100%	94%	148	2E−72
Utg196282	375925	376108	100%	91.90%	138	9E−67
Utg196306	170694	170874	100%	92.30%	139	2E−67
Utg196310	818272	818092	100%	92.90%	142	5E−69
Utg196310	1383796	1383978	100%	90.70%	131	7E−63
Utg196332	9336	9517	100%	92.30%	140	7E−68
Utg196332	16723	16541	100%	95.10%	155	3E−76
Utg196338	435308	435127	100%	93.40%	146	3E−71
Utg196352	184386	184203	100%	92.40%	141	2E−68
Utg196382	401983	401802	100%	92.90%	143	1E−69
Utg196392	139907	139727	100%	91.30%	133	5E−64
Utg196420	553999	553817	100%	93.50%	146	3E−71
Utg196436	1238359	1238539	100%	92.90%	142	5E−69
Utg196444	133605	133423	100%	92.30%	140	7E−68
Utg196470	33696	33516	100%	92.90%	142	5E−69
Utg196482	962685	962502	100%	92.40%	141	2E−68
Utg196490	51853	52034	100%	91.30%	134	1E−64
Utg196494	316015	315834	100%	90.20%	128	3E−61
Utg196542	421642	421462	100%	93.40%	145	1E−70

Utg196566	257172	256992	100%	90.70%	130	2E-62
Utg196578	392310	392491	100%	91.80%	137	3E-66
Utg196586	354649	354830	100%	92.90%	143	1E-69
Utg196594	919829	919647	100%	93.40%	146	3E-71
Utg196602	94202	94021	100%	94%	149	7E-73
Utg196608	52731	52912	100%	92.30%	140	7E-68
Utg196642	15067	15247	100%	91.80%	136	1E-65
Utg196642	2249975	2250157	100%	90.80%	131	7E-63
Utg196662	329032	329212	100%	91.80%	136	1E-65
Utg196666	1004743	1004561	100%	91.80%	137	3E-66
Utg196684	86678	86498	100%	92.90%	142	5E-69
Utg196706	423882	423702	100%	93.40%	145	1E-70
Utg196724	22553	22733	100%	90.70%	130	2E-62
Utg196730	1357611	1357429	100%	92.90%	143	1E-69
Utg196730	1360702	1360880	100%	91.80%	135	4E-65
Utg196756	25758	25938	100%	91.80%	136	1E-65
Utg196764	72500	72680	100%	90.70%	130	2E-62
Utg196806	80396	80578	100%	90.80%	131	7E-63
Utg196828	342667	342487	100%	93.40%	145	1E-70
Utg196844	111480	111298	100%	91.30%	134	1E-64
Utg196850	360997	361178	100%	93.40%	146	3E-71
Utg196868	82600	82419	100%	91.80%	137	3E-66
Utg196874	94371	94553	100%	93.40%	146	3E-71
Utg196898	52568	52386	100%	94%	149	7E-73
Utg196902	414167	413986	100%	92.30%	140	7E-68
Utg196904	123825	123645	100%	90.70%	130	2E-62
Utg196920	115029	114849	100%	90.80%	130	2E-62
Utg196946	47853	47672	100%	93.40%	146	3E-71
Utg196952	343413	343594	100%	92.30%	140	7E-68
Utg196966	818973	818792	100%	94.50%	152	1E-74
Utg196986	211809	211991	100%	92.40%	140	7E-68
Utg197012	509297	509478	100%	92.90%	143	1E-69
Utg197030	796719	796536	100%	91.40%	135	4E-65
Utg197042	54601	54419	100%	91.80%	137	3E-66
Utg197050	80598	80416	100%	91.30%	134	1E-64
Utg197066	254943	254761	100%	92.90%	143	1E-69
Utg197100	1416216	1416396	100%	92.90%	142	5E-69
Utg197120	38177	38358	100%	93.40%	146	3E-71
Utg197142	17933	17754	100%	92.30%	139	2E-67
Utg197142	35868	35688	100%	92.30%	139	2E-67
Utg197168	68647	68829	100%	91.30%	134	1E-64
Utg197176	90633	90452	100%	94%	149	7E-73
Utg197180	29462	29644	100%	91.80%	137	3E-66
Utg197180	656734	656917	100%	94%	150	2E-73
Utg197186	25092	24911	100%	92.40%	140	7E-68
Utg197244	552482	552300	100%	91.80%	137	3E-66
Utg197322	16711	16534	100%	90.70%	129	9E-62
Utg197334	151702	151883	100%	92.90%	143	1E-69

Utg197342	21236	21415	100%	92.90%	142	5E-69
Utg197460	320042	319860	100%	92.90%	143	1E-69
Utg197460	321309	321127	100%	93.40%	146	3E-71
Utg197460	322615	322434	100%	93.40%	146	3E-71
Utg197460	323944	323762	100%	92.90%	143	1E-69
Utg197468	64032	63852	100%	93.40%	145	1E-70
Utg197468	189725	189904	100%	92.30%	139	2E-67
Utg197468	268298	268480	100%	90.80%	131	7E-63
Utg197474	20357	20538	100%	93.40%	146	3E-71
Utg197506	4876	4695	100%	91.80%	137	3E-66
Utg197508	276069	275888	100%	92.30%	140	7E-68
Utg197570	172458	172638	100%	92.30%	139	2E-67
Utg197590	1490	1312	100%	91.80%	135	4E-65
Utg197600	170167	169985	100%	91.80%	137	3E-66
Utg197608	287233	287411	100%	90.80%	129	9E-62
Utg197608	750493	750674	100%	94%	149	7E-73
Utg197620	143810	143992	100%	93.40%	146	3E-71
Utg197624	123957	123777	100%	93.40%	145	1E-70
Utg197634	412997	412816	100%	90.70%	131	7E-63
Utg197648	55208	55027	100%	92.30%	140	7E-68
Utg197648	129330	129149	100%	90.70%	131	7E-63
Utg197662	860019	860201	100%	91.30%	134	1E-64
Utg197706	423785	423967	100%	91.80%	137	3E-66
Utg197708	446135	446315	100%	91.80%	136	1E-65
Utg197870	310935	310754	100%	92.30%	140	7E-68
Utg197906	401636	401818	100%	91.80%	137	3E-66
Utg197914	315753	315571	100%	94%	149	7E-73
Utg197930	78602	78418	100%	90.30%	129	9E-62
Utg197950	204926	205109	100%	92.90%	144	4E-70
Utg197954	62243	62423	100%	90.70%	130	2E-62
Utg198004	335996	336179	100%	93%	144	4E-70
Utg198056	53322	53142	100%	92.30%	139	2E-67
Utg198086	72400	72221	100%	93.40%	145	1E-70
Utg198156	111016	111197	100%	92.90%	143	1E-69
Utg198158	297869	298051	100%	93.40%	146	3E-71
Utg198158	326911	327094	100%	92.40%	141	2E-68
Utg198162	86608	86790	100%	90.70%	131	7E-63
Utg198166	156682	156500	100%	92.40%	140	7E-68
Utg198228	688934	688755	100%	94.50%	151	5E-74
Utg198232	79969	80153	100%	91.90%	138	9E-67
Utg198252	54608	54790	100%	92.90%	143	1E-69
Utg198374	69249	69434	100%	91.90%	139	2E-67
Utg198410	17722	17540	100%	91.80%	137	3E-66
Utg198422	85215	85393	100%	93.40%	144	4E-70
Utg198428	654576	654398	100%	91.30%	132	2E-63
Utg198468	277574	277392	100%	92.40%	140	7E-68
Utg198506	15287	15105	100%	91.80%	137	3E-66
Utg198572	44067	44241	100%	90.20%	124	5E-59

Utg198572	935541	935722	100%	92.30%	140	7E-68
Utg198578	76121	76302	100%	91.80%	137	3E-66
Utg198664	450571	450752	100%	91.30%	134	1E-64
Utg198682	93144	92963	100%	94%	149	7E-73
Utg198688	23827	23646	100%	91.80%	137	3E-66
Utg198694	66696	66876	100%	92.90%	142	5E-69
Utg198702	548833	548653	100%	92.90%	142	5E-69
Utg198774	329272	329093	100%	91.80%	136	1E-65
Utg198814	286211	286030	100%	91.30%	134	1E-64
Utg198836	567369	567550	100%	91.80%	137	3E-66
Utg198836	766421	766244	100%	90.70%	129	9E-62
Utg198872	30940	31124	100%	90.30%	129	9E-62
Utg198948	621004	620824	100%	94.50%	151	5E-74
Utg198974	185567	185747	100%	92.30%	139	2E-67
Utg199014	251298	251117	100%	92.30%	140	7E-68
Utg199090	418826	418645	100%	91.80%	137	3E-66
Utg199150	762896	762716	100%	91.30%	133	5E-64
Utg199154	435	613	100%	92.30%	138	9E-67
Utg199166	1104446	1104269	100%	91.30%	132	2E-63
Utg199300	120628	120809	100%	92.30%	140	7E-68
Utg199340	170020	170202	100%	94.50%	152	1E-74
Utg199350	17114	16929	100%	91.90%	139	2E-67
Utg199364	271077	270896	100%	91.80%	137	3E-66
Utg199372	544771	544592	100%	94%	148	2E-72
Utg199374	28086	28267	100%	92.30%	140	7E-68
Utg199422	7022	7203	100%	92.30%	140	7E-68
Utg199462	30693	30515	100%	90.20%	126	4E-60
Utg199472	102280	102098	100%	91.30%	134	1E-64
Utg199572	180745	180927	100%	93%	143	1E-69
Utg199604	233758	233577	100%	92.30%	140	7E-68
Utg199626	83070	83252	100%	94%	149	7E-73
Utg199638	171948	171766	100%	92.90%	143	1E-69
Utg199814	1518	1336	100%	91.80%	137	3E-66
Utg199884	52052	52232	100%	91.80%	136	1E-65
Utg200020	13986	14167	100%	92.90%	143	1E-69
Utg200138	113130	113309	100%	91.30%	133	5E-64
Utg200160	14443	14622	100%	90.20%	127	1E-60
Utg200172	209981	209801	100%	92.90%	142	5E-69
Utg200182	137772	137592	100%	93.40%	145	1E-70
Utg200188	29800	29619	100%	91.80%	137	3E-66
Utg200380	54524	54703	100%	90.20%	127	1E-60
Utg200654	41893	41710	100%	94%	150	2E-73
Utg200732	46417	46236	100%	100%	182	3E-91
Utg200990	157765	157584	100%	91.30%	134	1E-64
Utg201144	42084	42267	100%	90.20%	129	9E-62
Utg201496	15986	16167	100%	95.70%	158	7E-78
Utg201518	37752	37936	100%	91.40%	135	4E-65
Utg201826	326735	326916	100%	91.80%	137	3E-66

Utg202012	1091	1270	100%	90.20%	127	1E-60
Utg202012	169571	169389	100%	90.30%	128	3E-61
Utg202964	40123	39943	100%	91.30%	133	5E-64
Utg203256	5921	5741	100%	92.90%	142	5E-69
Utg203256	66781	66598	100%	90.90%	132	2E-63
Utg203350	71323	71504	100%	95.70%	158	7E-78

Table S12 更新されたドラフトゲノム配列における STR の位置。

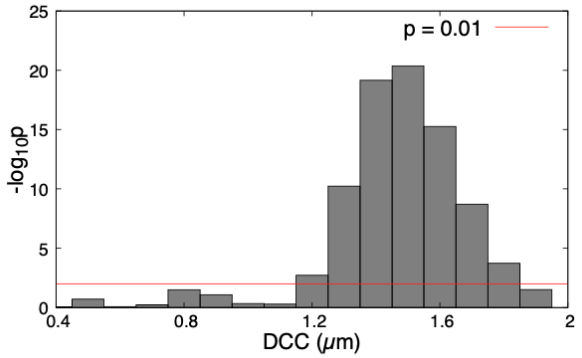

Fig. S1 RPD 値間の検定における p 値(−log10 p)の分布。隣接する 2 つの DCC 値において得られた RPD 値間の Welch の t 検定による p 値(−log10 p)の分布。

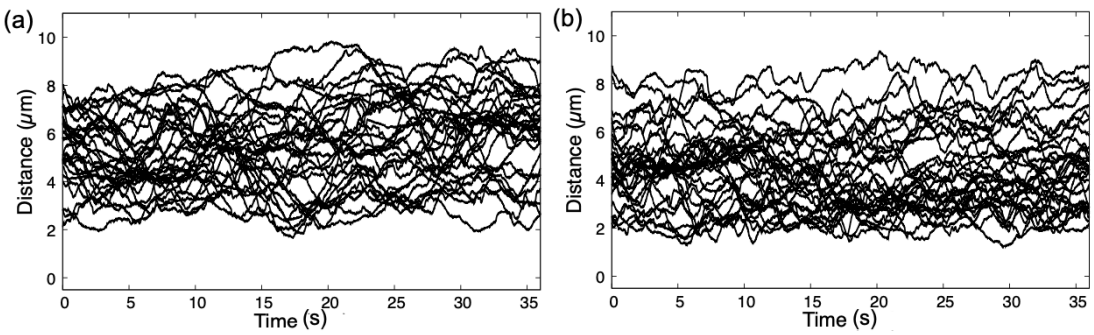

Fig. S2 Xic ペア間距離の時間経過。ES cell model(a)と 2-day cell model(b)において、30 種類の初期条件から得られた 2 つの Xic ペア間距離の時間的変化のシミュレーション結果。

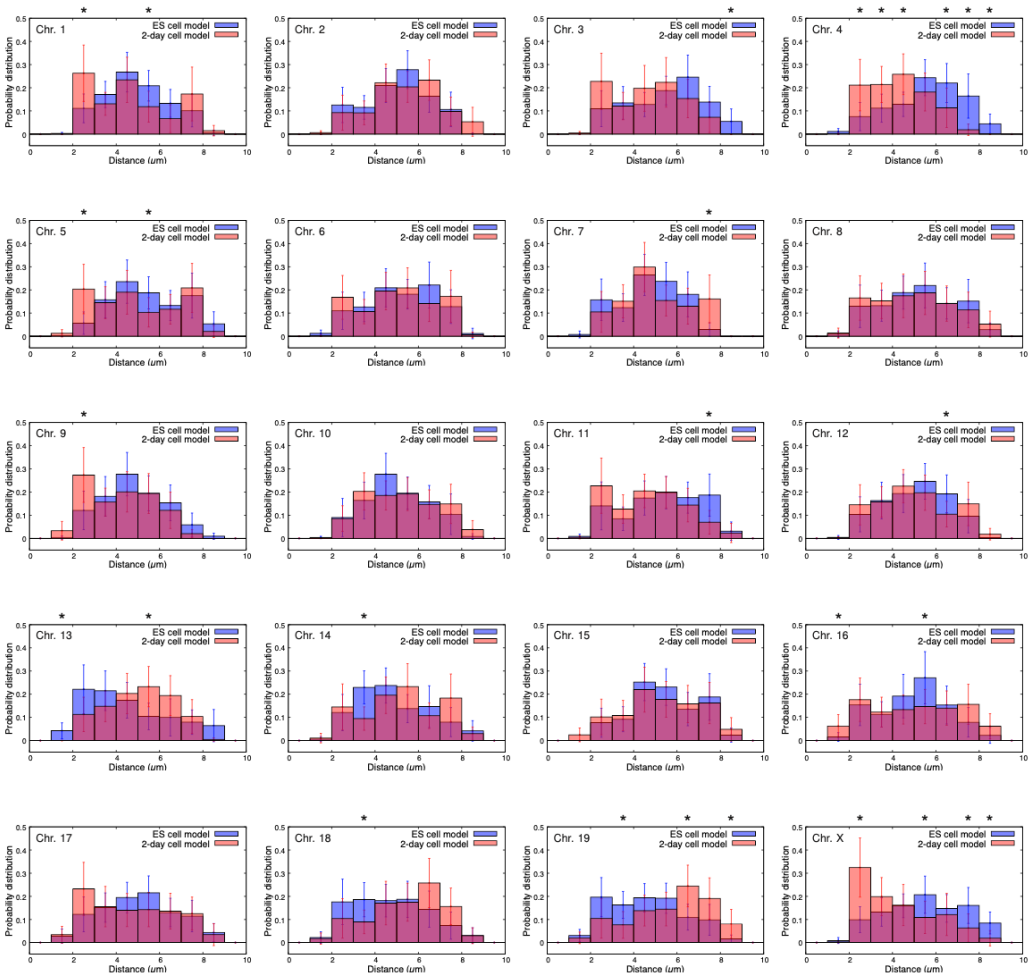

Fig. S3 相同染色体間距離の確率分布。ES cell model と 2-day cell model における相同染色体ペア間距離確率分布の、平均と 95%信頼区間（エラーバー）を描画した。平均と 95%信頼区間は各モデルの 30 サンプル分のシミュレーションの結果から計算した。*がついている距離の部分は、2 つのモデル間で確率に有意差が見られたことを示す（Welch の t 検定を用いて、有意水準 p 値<0.05 とした）。

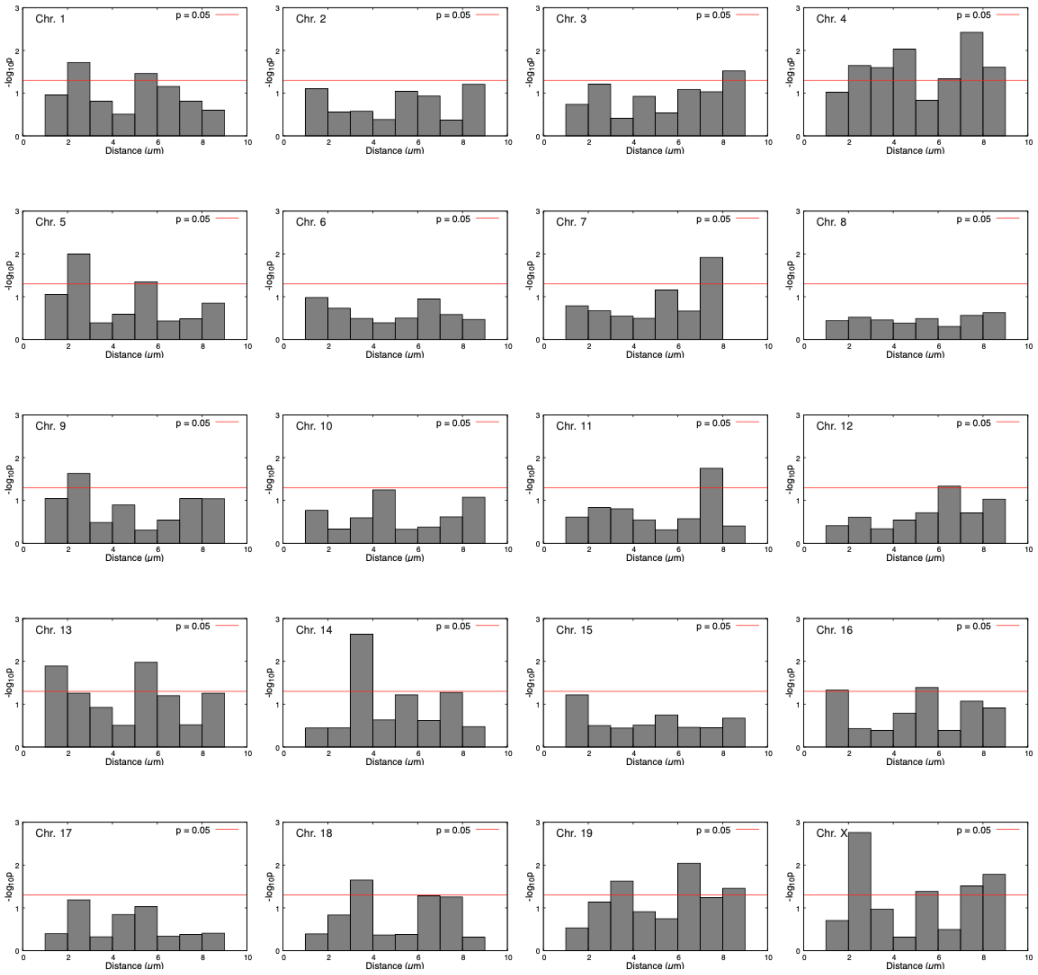

Fig. S4 相同染色体間距離の確率分布における Welch の t 検定の p 値(−log10 p)の分布。ES cell model と 2-day cell model の各距離における相同染色体間距離の確率に対して Welch の t 検定を行った。(Fig. S3)

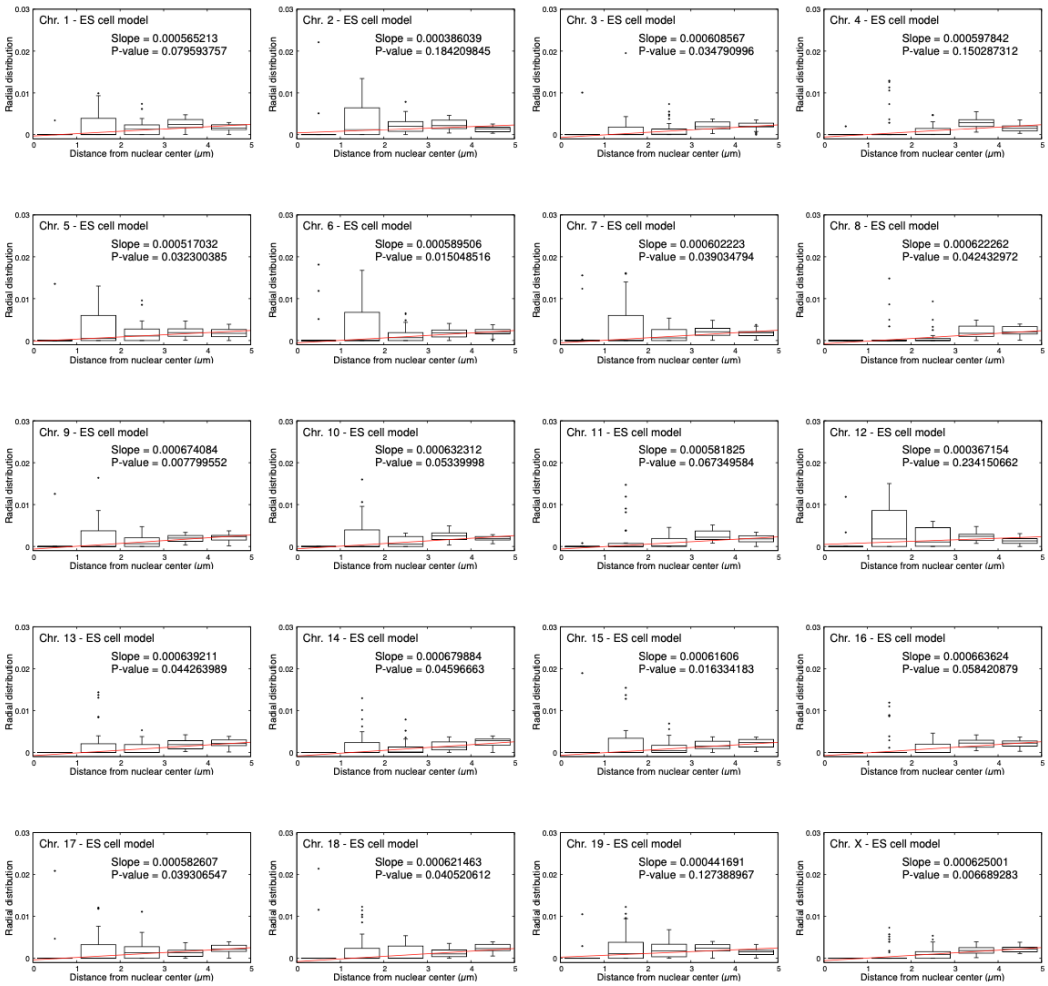

Fig. S5 各染色体の動径分布関数。ES cell model および 2-day cell model における核中心から各染色体の中心までの動径分布関数の値の箱ひげ図。動径分布関数は、[核中心から各染色体までの距離の度数分布]/(4π[核中心からの距離]^2)で定義した。ES cell model と 2-day cell model の動径分布の箱ひげ図を、それぞれ 30 回のシミュレーションのデータを用いて描画した。これらの動径分布の距離依存性の特徴は、距離と各距離の動径分布値の中央値との間の線形回帰分析による傾きと p 値で評価した。

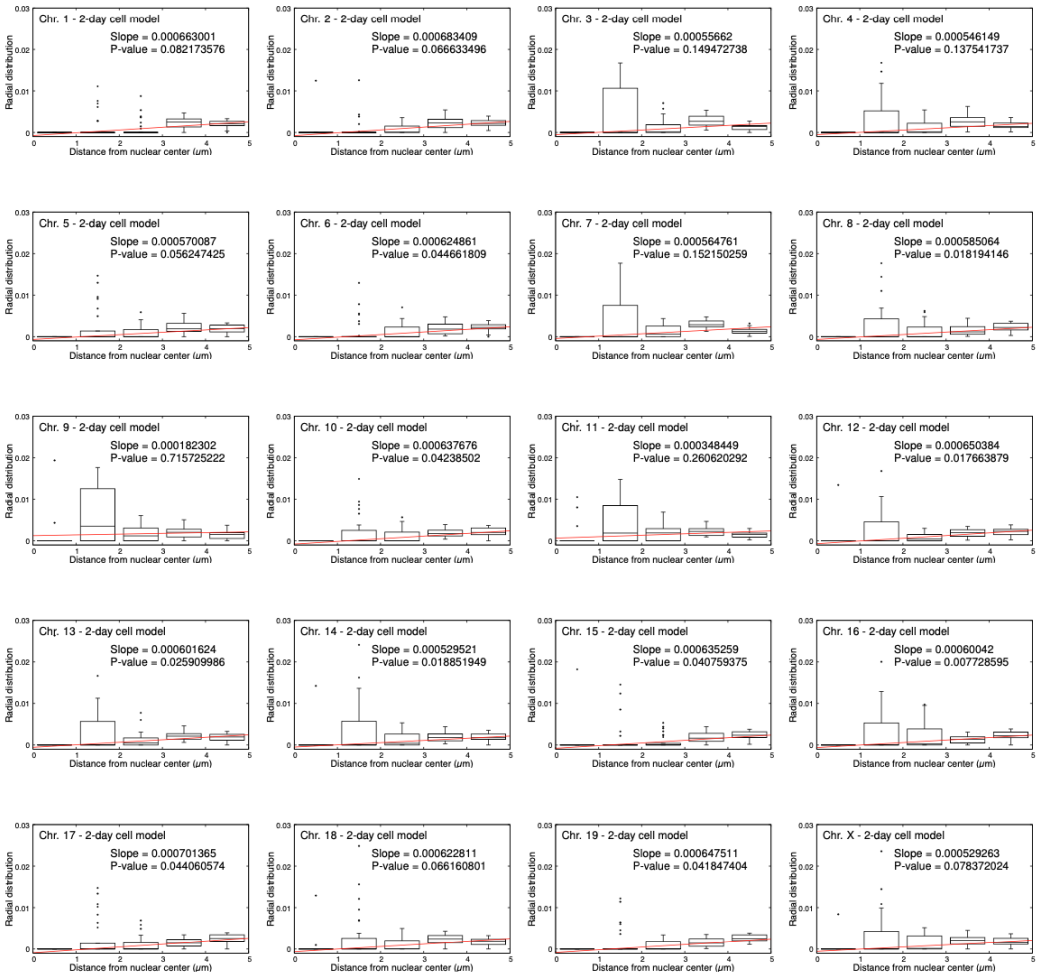

Fig. S5（続き）各染色体の動径分布関数。ES cell model および 2-day cell model における核中心から各染色体の中心までの動径分布関数の値の箱ひげ図。動径分布関数は、[核中心から各染色体までの距離の度数分布]/(4π[核中心からの距離]^2)で定義した。ES cell model と 2-day cell model の動径分布の箱ひげ図を、それぞれ 30 回のシミュレーションのデータを用いて描画した。これらの動径分布の距離依存性の特徴は、距離と各距離の動径分布値の中央値との間の線形回帰分析による傾きと p 値で評価した。

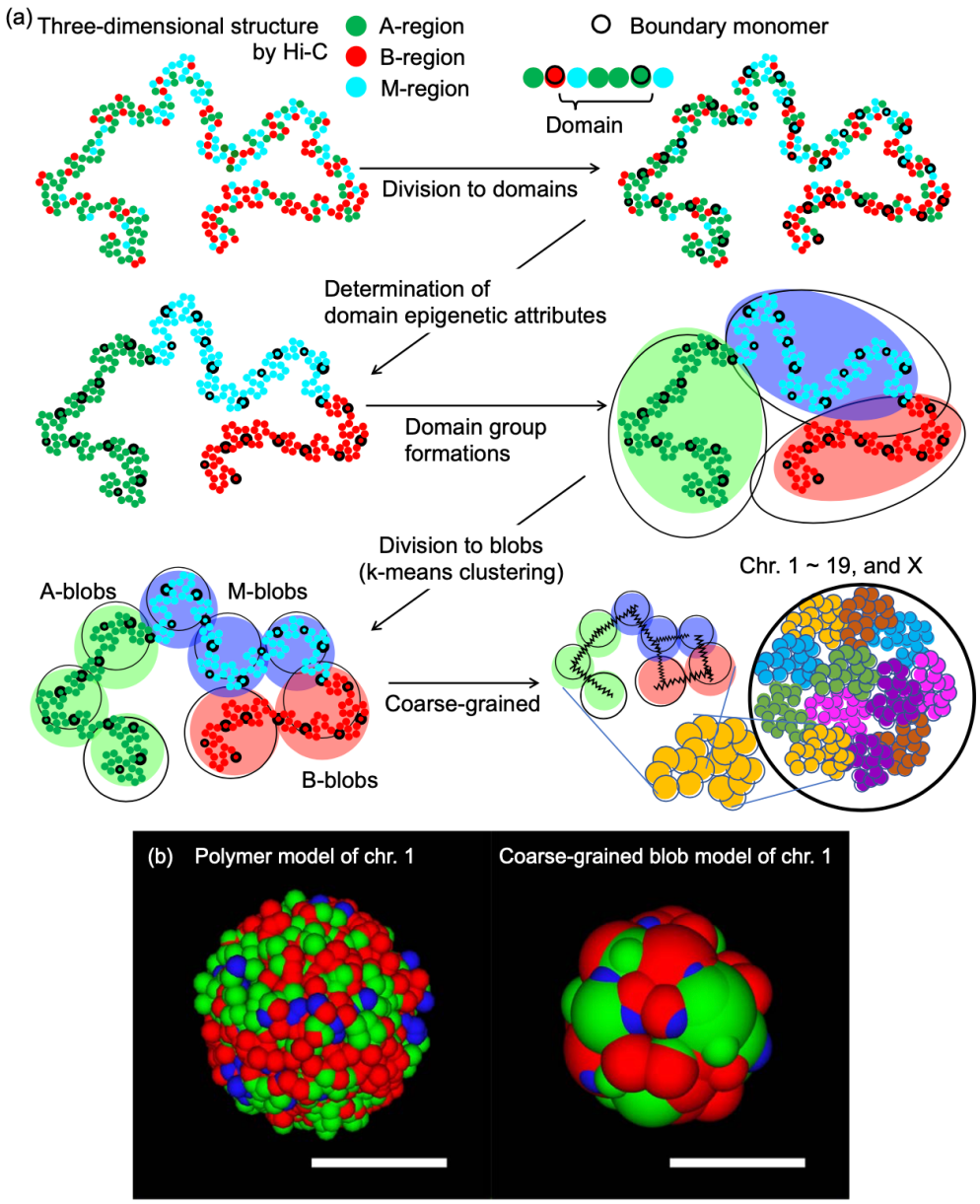

Fig. S6 染色体粗視化モデル構築手順の模式図。(a) 各ステップをより明確に示すために、仮想的な染色体ポリマーモデルを用いた。この例では、CNDn を 4 つの領域として k-means クラスタリングを行った。互いに重なり合う粒子は、それらの中心間の距離に等しい自然長を持つ弾性バネにで繋いだ（右下）。(各モデリングステップの詳細については、II – v： 手法を参照。) (b) 染色体ポリマーモデルと粗視化粒子鎖モデルの図。ES cell model 内の Chr.1 のものを例として示した。スケールバーは 2 μm を示す。

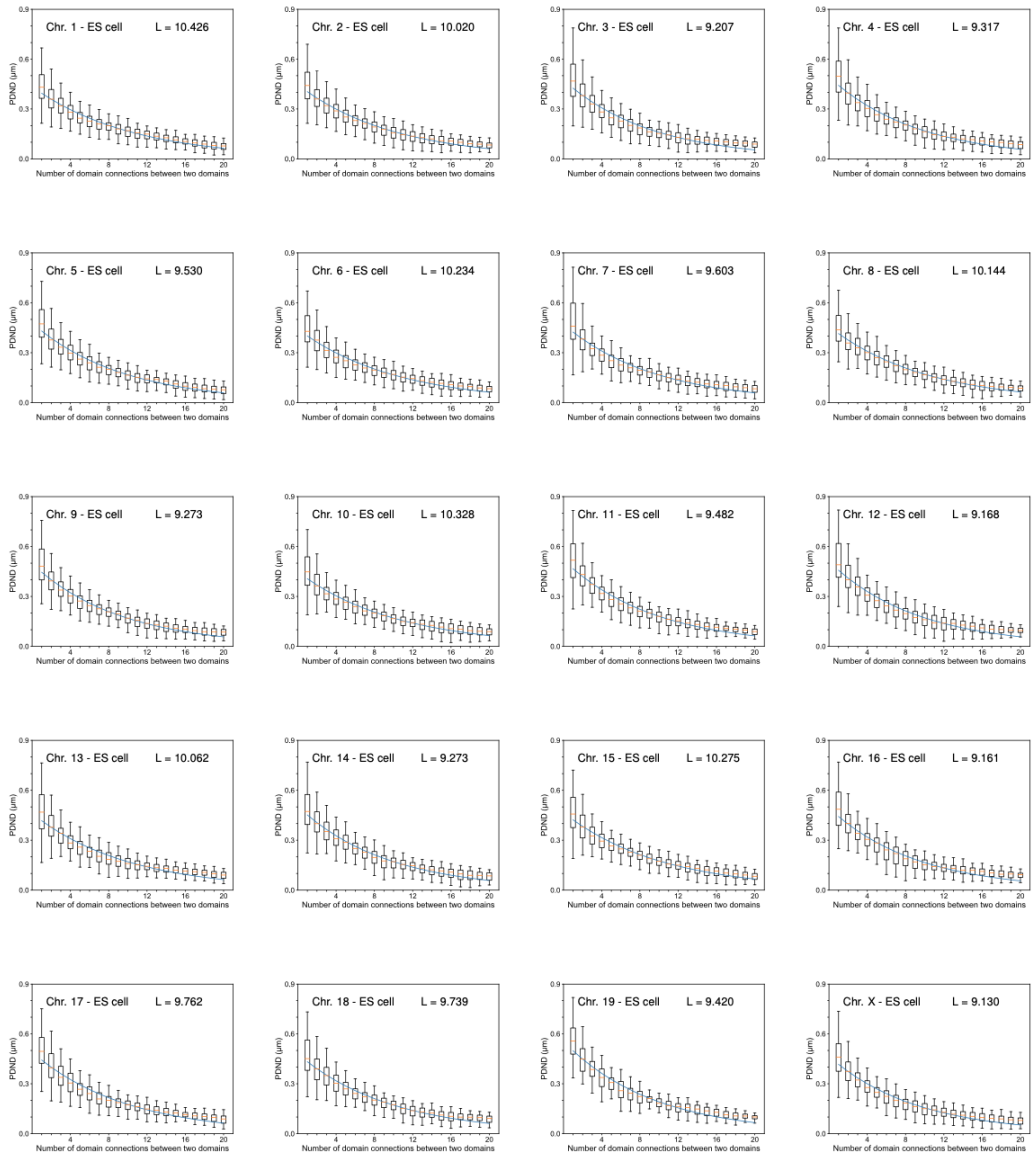

Fig. S7 染色体における特徴的なドメイン数（CNDn）を決定するための PDND：ポリマーモデルによって得られた ES 細胞と 2 日目細胞の全染色体の[ドメイン間のドメイン数]の関数としての PDND の箱ひげ図。PDND の各平均値は 95%信頼区間に含まれる PDND 値で推定した。曲線（青）は、PDND の平均値を exp(− [2 つの遺伝子座間のドメイン数] / L)としてフィッティングしたものである。

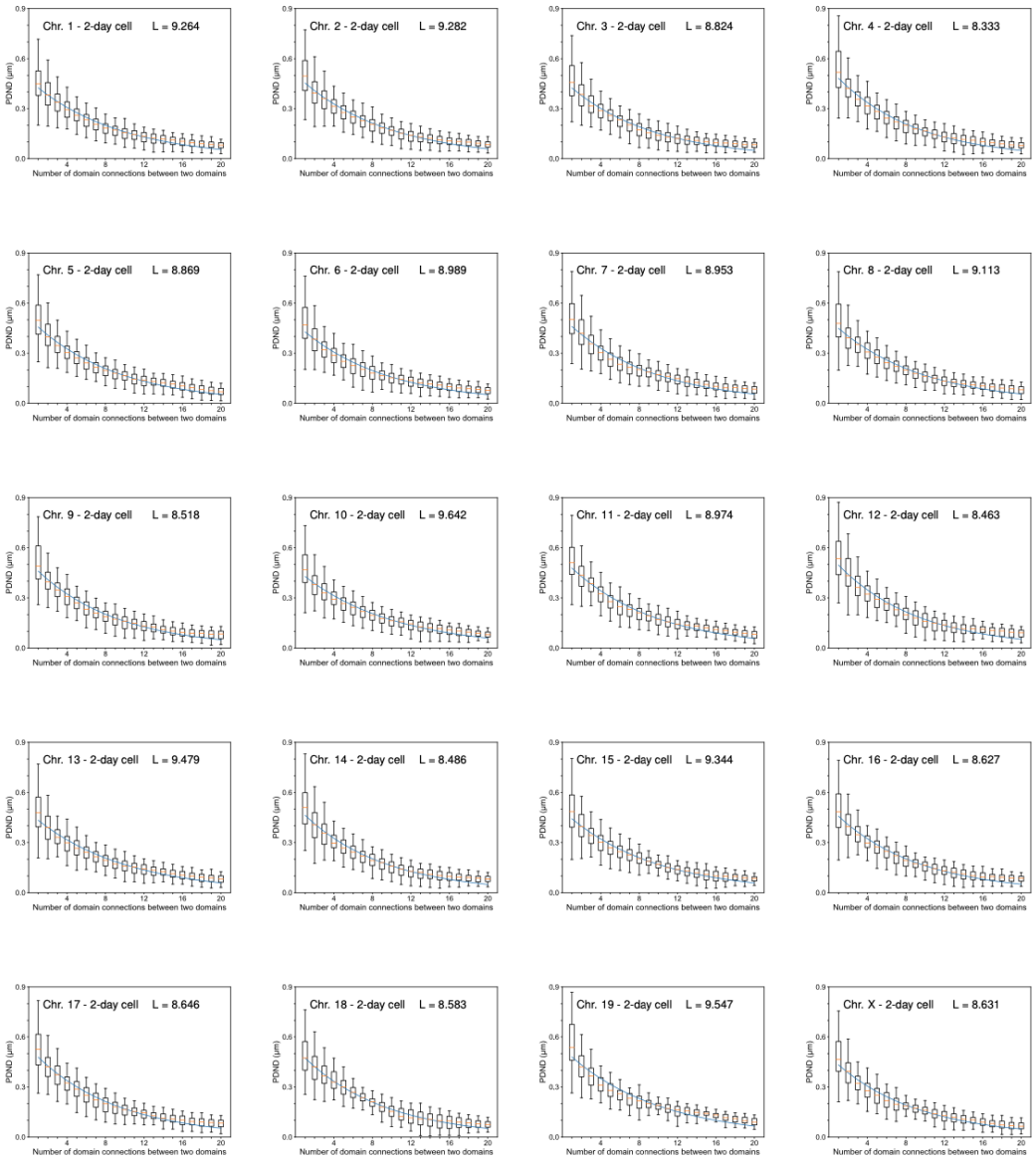

Fig. S7（続き） 染色体における特徴的なドメイン数（CNDn）を決定するための PDND：ポリマーモデルによって得られた ES 細胞と 2 日目細胞の全染色体の[ドメイン間のドメイン数]の関数としての PDND の箱ひげ図。PDND の各平均値は 95%信頼区間に含まれる PDND 値で推定した。曲線（青）は、PDND の平均値を exp(− [2 つの遺伝子座間のドメイン数] / L)としてフィッティングしたものである。

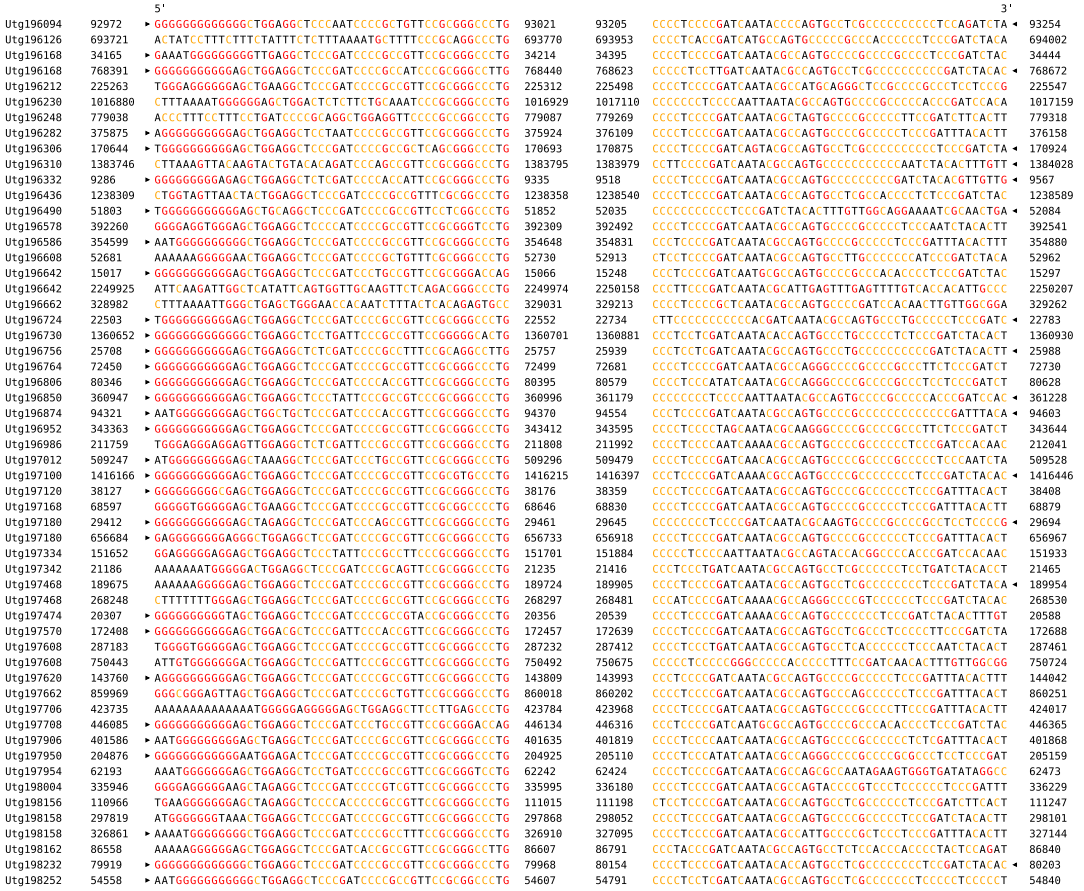

Fig. S8 185 の ArsInsC ホモログの上流（左）と下流（右）の 50bp 配列とその位置。グアニン（G）とシトシン（C）はそれぞれ赤とオレンジで着色されている。黒い矢印で示した領域は G (C)-ストレッチを含む。

Fig. S8 (続き) 185 の ArsInsC ホモログの上流（左）と下流（右）の 50bp 配列とその位置。グアニン（G）とシトシン（C）はそれぞれ赤とオレンジで着色されている。黒い矢印で示した領域は G (C) -ストレッチを含む。