博士学位論文

ChIP-seq データベースの構築による遺伝子転写制御機構の解明

Construction of a ChIP-seq database and elucidation of gene transcription regulation mechanisms

東北大学大学院情報科学研究科
応用情報科学専攻
生命情報システム科学分野

安澤 隼人

2020 年 1 月

要旨

生命現象を理解する上で転写制御機構の解明は重要な 1 要因であり、そのための実験的手法である ChIP-seq 法により得られた公共データは年々蓄積されている一方、既存の ChIP-seq データベースは網羅性と信頼性の評価の観点から十分な情報基盤としての地位を得ていない。本研究ではこの 2 点を克服し、転写制御機構の情報提供基盤足りうる公共 ChIP-seq データベース、C4S (Comprehensive Collection and Comparison for ChIP-Seq) DB の開発を行った。

本研究では解析パイプラインを開発するに当たり、網羅性を確保すべくメタデータの自動判定を試み、結果 ENCODE Portal で公開されているデータに加えて、GEO（Gene Expression Omnibus）に登録されたデータのうち 84% のデータを自動処理する目処が立った。また ChIP-seq の大規模解析としては初めてデコイ配列を含んだリファレンスゲノムを採用し、難読領域における偽陽性の低減に効果が認められた。信頼性の観点からは、従来のクオリティコントロール（QC）手法でリード数に対するロバスト性が期待されたものの指標としての問題点が残っていた Strand cross-correlation について、初めて理論的な特性評価を行いその結果に基づいた新規 QC 手法の提案を行った。この指標はロバスト性と SN 比との高い相関を両立した指標であることを実データを用いて検証した。

これらの成果に基づき、解析パイプラインをクラウドコンピューティング環境に展開し大規模解析を行い、得られた結果を可視化するデータベース Web アプリケーション C4S DB の開発を行った。データベースは、1. 個々の QC 情報を含む実験データの可視化、2. 遺伝子周辺領域に存在するピークの可視化、3. 実験間の大域的な類似度に基づくクラスタリング、の 3 つの機能を軸として実装されており、ヒト A549 細胞の ChIP-seq データセットを用いたデモンストレーションを通して転写制御機構の解明に寄与する情報を提供することを実演する。

目次

図目次

表目次

コード目次

第 I 部
緒論

1 背景

DNA の二重螺旋構造の発見 [1] やセントラルドグマの提唱 [2] から半世紀が経過し、DNA やゲノムは生命の設計図として大衆にも認知されるようになった。2000 年初頭のヒトゲノム計画完了を経てもゲノム解析を始めとする関連分野がなお生物学の中心分野の 1 つであり続けるのは、配列情報としてのゲノム解読が完全でないこともさることながら、その設計図がいかに生物を形作る生命現象として展開されていくのかという過程が十分に解明されていない所も大きく、人類が生命の設計図を塩基配列の集合として得られようとも生命現象の総体を明らかにするまでには未だに大きな隔たりが存在する。

設計図の使い方という観点では、幹細胞の分化やクローンの個体など DNA の塩基配列がほとんど同一であるとされるにも関わらず、個体の表現型が異なるという例は広く観察される。このようなセントラルドグマでは一元的に説明できない現象はかつて環境要因による説明が試みられてきたものの、例えば細胞の分化のように一度生じた変化を継代的に維持する現象は、DNA 塩基配列の変化を伴わない遺伝情報の伝達が存在しなければ説明することが難しい。現在では、DNA はそれ自身が持つ塩基配列のみが遺伝情報の全てではなく、DNA メチル化、DNA 結合タンパクとの相互作用やそれらの化学的修飾が制御されることにより、塩基配列に含まれるどの遺伝子がいつどの程度活性あるいは抑制されるかを巧みに制御する機構が存在することが明らかになっている。このような遺伝情報を伝達・制御するシステムはエピジェネティクスと呼ばれる。

革新的なデータのハイスループット化とコストの低減をもたらした次世代シーケンシング（NGS）技術はエピジェネティクス分野にも大きな影響をもたらした。NGS の応用として確立された ChIP シーケンシング（ChIP-seq）法によって、ある DNA 結合タンパクの結合部位を一度の実験で全ゲノム領域を対象に検出することが可能となり、転写因子やヒストン修飾といった遺伝子転写制御機構に深く関わる DNA 結合タンパクの分布を明らかにする手法として現在に至るまで広く用いられるようになった。ChIP-seq 法はクロマチン免疫沈降（ChIP: Chromatin immunoprecipitation）と NGS によるシーケンシングを組み合わせた手法だが、RNA-seq のような他の NGS を応用した手法と比較すると ChIP の過程を含む点が実験手法として複雑であり、シグナル-ノイズ比（S/N）の悪化やコストの相対的増加が生じている。これらの問題は未だ根本的に解決されていない。

多くの遺伝子やタンパク質の機能が解明されるにつれてより複雑な生命現象が明らかにされつつある。転写制御も複数の転写因子やヒストン修飾の協調により達成される例も多く知られるようになり、1 つの研究において複数の制御因子を対象に実験を行い、結果を統合することで生物学的知見の発見を達成する試みが広く行われるようになった。特に ENCODE プロジェクトのような国際的なコンソーシアム主導による巨大プロジェクトは、そのような総体を明らかにする試みの最たる例である。また統合解析を推し進める手段として、複数の実験結果を比較・統合する手法の開発や先行研究の解析データを容易に再利用できる ChIP-seq データベースの構築が行われてきた。特に膨大な公開済み生データを元に、それらの解析結果を統合したデータベースは関連研

究の促進に留まらず、遺伝子転写制御機構を網羅的に明らかにし、俯瞰的な生物学的知見を提供するプラットフォームとして機能することが期待される。しかしながら、生データを格納する一次データベースの整備状況や、ChIP-seq におけるクオリティコントロール（QC）の問題を背景として、網羅性と信頼性を十分に両立した ChIP-seq の二次データベースはまだ確立されていない。

2 本研究の目的

本研究では、より多角的なデータの統合が求められつつある生命科学分野、特に転写制御を扱う ChIP-seq データを対象に、情報科学の観点からデータドリブンに生物学的な知見や研究仮説を提供するためのプラットフォームとしての ChIP-seq データベースの構築を目指す。特に、網羅性および信頼性に対する具体的な目標として次の 2 点を設定する。

1. ENCODE および GEO から入手できるヒト公共 ChIP-seq データに対して同一解析パイプラインによる再解析を実施し、比較可能な解析結果を提供する。
2. より信頼性の高い ChIP-seq データの QC 手法を確立し、全解析対象データについて QC 解析結果の提供を行う。

生物種をヒトに限定するのは、利用可能なデータが生物種の中で最も多く高い需要があること、ヒト ChIP-seq データにおいてデータベースの構築手法を確立できれば他生物種（特にヒトと並んで多数の実験が実施されている哺乳類等高等生物）への展開も容易に行えることが予想されるためである。

解析パイプラインおよびデータソースについては第 II 部で詳しく述べる。ChIP-seq データの QC に関する問題点とその解決策については第 III 部で詳しく述べる。

3 本論文の構成

第 I 部は緒論であり、本研究の背景と目的について述べている。

第 II 部では、基本的な ChIP-seq 解析の手法に触れつつ既存の解析パイプラインやデータベースの問題点を整理し、本研究の目指すデータベースの構築に必要な ChIP-seq 解析パイプラインを提案する。特に、機械的アクセスに難のあるメタデータに対応する手法の開発や、これまで試みられてこなかった ChIP-seq 解析におけるデコイ配列入りリファレンスゲノムの応用について述べる。

第 III 部では、第 II 部に関連して、ChIP-seq 実験のクオリティを評価する既存の手法に対して理論的なアプローチから知見を与えた研究について述べる。

第 IV 部では、第 III 部の成果を踏まえて新規 QC 指標を提案し、それが実際に適用可能であるか検証した結果について述べる。

第 V 部では、ここまでに得られた成果を用いて実際に解析パイプラインを構築、クラウドコンピューティング環境上で稼働させ、得られた ChIP-seq 解析結果からどのような生物学的考察が可能であるか実証を行った結果について述べる。

第 VI 部は総括であり、本研究の成果をまとめるとともに課題点や今後の展望について述べる。

第 II 部
ChIP-seq データ解析パイプラインの開発

4　導入

生命を形作る現象が DNA にコードされた遺伝情報に端を発することはセントラルドグマとして知られているが、細胞や個体の表現型が塩基配列のみによって定まるとは限らない。例えば真核生物に見られるヒストンは化学的修飾を受けることによりヌクレオソーム構造が締緩し、転写因子と呼ばれる DNA 結合タンパク質群の DNA へのアクセスを制限することで遺伝子発現を制御することが知られている。またこのような修飾は塩基配列とは独立した遺伝情報として娘細胞にも継承されることがしばしばあり、発生や細胞分化などの生命現象において重要な役割を果たしている [3]。そのため、個々の遺伝子やタンパク質の性質を研究するだけでなく、遺伝子の転写や発現がいつどのように制御されるのかを明らかにすることが生命現象をシステムとして理解する上で必要となる。

このような制御のうち、特にセントラルドグマの出発点となる転写の制御において転写因子やヒストン修飾が重要な役割を担う。これらの所在を実験的に特定する手法として ChIP-seq（クロマチン免疫沈降シーケンス）法 [4–6] が現在の主流となっている。ChIP-seq 法は転写因子やヒストン修飾の結合部位をゲノムワイドに検出することが可能であり、次世代シーケンシング（NGS）技術の発達と共に広く用いられるようになった。これらのデータは公開レポジトリの役割を持つ公共データベース（一次データベース）上に蓄積され続けている（図 1）。一次データベースに登録されるのは主に生データであるが、これを元に結合部位の一覧のような扱いやすい情報に再解析したデータから構成される二次データベースの構築も行われるようになった。二次データベースの構築は ChIP-seq データのみならず生命情報科学分野一般に盛んに行われているが、特に転写制御の研究においては、転写制御が複数の転写因子やヒストン修飾の協調により達成されることも多く、未知の転写制御機構の解明には複数の実験データを統合することが必要不可欠となりつつある。従って、ChIP-seq データを用いた二次データベースの整備は、先行研究のデータを併用することを前提とした実験計画の立案や、既存のデータのみを用いるデータ駆動型の研究を行うことを容易にし、当該分野の研究を促進させることが期待される。ChIP-seq データを用いた二次データベースの例として ReMap[7] では約 3,000 件のデータセットを元におよそ 300 種類の転写因子の情報を提供し、また再解析結果を元に転写因子共結合パターンの大規模な可視化を行った [8]。このように、可用性の高いデータの共有基盤となる二次データベースの構築は、既存の実験データの再活用を促進するだけでなく、俯瞰的・網羅的解析の足がかりとなることでこれまでにない観点からの生物学的知見を提供することができる。

しかし、既存の二次 ChIP-seq データベースは、網羅性と信頼性の観点から一次データベースに登録されたデータから十分に知見を引き出しきれているとは言えない。第一の原因として、一次データベースに登録されたメタデータ（付帯情報）の不完全性が挙げられる。例えば、ヒトゲノム内の機能的因子の網羅的解明を目標としている ENCODE プロジェクト [9] は、ChIP-seq を含む大量の NGS データが構造化されたメタデータと共に提供されており、外部のユーザーもプロジェクトで取得されたデータを包括的かつ機械的に参照することが可能である。一方、論文出

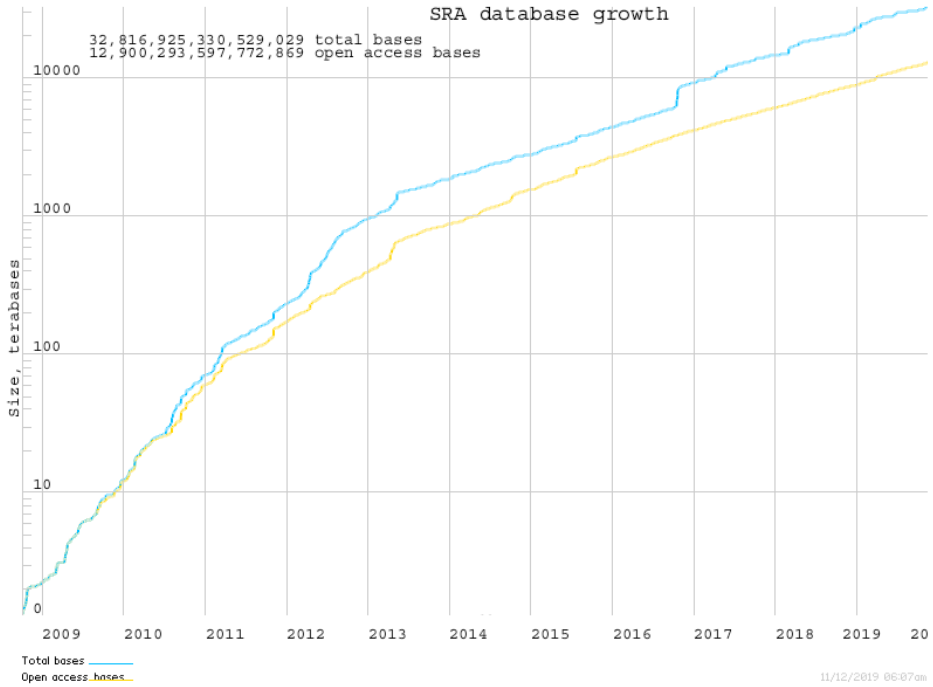


図 1: SRA に登録されたデータ量の変化

2019 年 11 月 12 日 https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi? からダウンロード

版時の NGS データセット公開レポジトリとして使用されることの多い GEO（Gene Expression Omnibus）及び SRA（Sequence Read Archive）には、ENCODE に代表される大型プロジェクトより膨大なデータが格納されており、メタデータの登録も義務付けられているものの、その登録内容の一部は構成が登録者に委ねられている場合があり、再解析に必要なメタデータが欠落している例もある。従って、コンソーシアムによる一元管理が行われているデータベースと比較すると、大部分の先行研究のデータが登録されている一次データベースはメタデータの不完全性と機械的処理を行う難易度の高さが大規模なデータ再解析を難しくしているといえる。

第二に、ChIP-seq 実験そのものの難しさに起因するクオリティーコントロール（QC）の必要性がある。ChIP-seq はサンプル調整の過程においてクロマチン免疫沈降を行うために、免疫沈降で使用される抗体の質など複合的な要因が最終的に得られる解析結果を左右しやすいことから、NGS を用いた実験手法としては比較的複雑であるといえる。このため、RNA-seq に代表される他の NGS を用いた手法と比較してノイズが多くなるため、結合部位の特定において信頼できる結果を得るためには QC が不可欠である。特に、一次データベースに登録することそのものは、登録されたデータの品質とは無関係であるため、二次データベースにおいて客観的な QC の結果を共に提供することがデータの適切な再利用を促すために必要であると考えられる。これまでに多くの QC 指標が提案され [10]、ENCODE 等のプロジェクトにおいて一定の成果を挙げているが、主要な QC 指標でも理論的裏付けが不十分であったり、解析手法等のバイアスを受けやすかったりする等、ChIP-seq 解析における QC 手法の整備そのものがまだ十分とはいえない状況にある。

表 1: 主要な二次 ChIP-seq データベースの対応実験数と QC 情報の提供状況

データベース名	データソース	対応生物種	実験数	QC 情報の提供
ReMap	GEO, ENCODE	ヒト	3,180	あり
GTRD[13]	GEO, ENCODE	ヒト, マウス	12,168	なし
Cistrome DB	GEO, ENCODE	ヒト, マウス	20,535	あり
ChIP-Atlas[11]	SRA, GEO ENCODE	ヒト, マウス等 6 種	76,217	なし

現在入手可能な ChIP-seq データとしては、GEO に約 15,000 件、SRA に約 110,000 件、ENCODE portal では約 7,200 件の実験に対するデータが公開されている。ただし、ENCODE のデータの一部は GEO にもデポジットされており、また GEO は生データの保存に SRA を使用しているためこれらのデータには重複がある。対して、現在公開されている主な二次 ChIP-seq データベースの概要を表 1 にまとめる [11]。このうち対応生物種に対する網羅性と QC 情報の提供を両立しているのは Cistrome DB[12] であるが、採用されている QC は既存の手法 [10] であり、従来の QC 手法の問題を抱えたままであることから依然として改善の余地があると言える。

5 解析パイプライン概要

本章では、研究目的を達成するための ChIP-seq 解析パイプラインの概要について、データソース・解析・品質評価の 3 点に分けて述べる。メタデータ抽出手法とデコイ配列入りリファレンスゲノムの採用については別章で述べる。図 2 は解析パイプラインの全体図である。

5.1 データソース

本節では、本研究で用いる ChIP-seq データの取得元であるデータベースやプロジェクトについて述べる。まず NGS データの主要なレポジトリである SRA・GEO とそれらの関係性について述べ、次に ENCODE プロジェクトと、それに関連するデータの取得が可能なプロジェクトについて述べる。

5.1.1 SRA

NGS データの特徴は 1 本あたり 1,000bp 以下の短い配列が大量に生成されることで、次世代シーケンシング技術の登場に伴いこのようなそれまでにない特徴を持つデータを格納する専用の公開レポジトリの需要が生じた。そのために 2007 年最初に設立されたのが NCBI（National Center for Biotechnology Information）による Short Read Archive である [14]。Short Read Archive は 2009 年に SRA（Sequence Read Archive）として改編され、NCBI, EBI（European Bioinformatics Institute）, DDBJ（DNA Data Bank of Japan）で構成される INSDC（International Nucleotide Sequence Database Collaboration）が運営している [15]。SRA は現在最大の NGS 用生データベースであり格納・公開されているデータは現在に至るまで増加し続けている（図 1）。

5.1.2 GEO

GEO（Gene Expression Omnibus）は 2000 年に NCBI によって開設された公共データベースである。当時対象としていたのは、主にマイクロアレイ解析で生じたハイスループット遺伝子発現量データであった [16]。現在では、RNA-seq・ChIP-seq などの NGS を含め多くのプラットフォーム由来のデータセットを登録できる公共データベースとして運営されている [17]。SRA との違いは生データだけではなく解析データも受け付けており、特定研究や実験に対するデータセットをより容易に検索・再利用可能にすることを主目的としている。NGS による実験の場合、解析済みデータは GEO のみに格納されるが、生データは GEO を通じて SRA に登録・保存する運用を行っている。

5.1.3 ENCODE Project

ヒトゲノムの配列が読み解かれた際の次の課題として、ヒトゲノムに含まれる配列の機能を全ゲノム規模で明らかにすることは自然な流れといえる。ENCODE (The ENCyclopedia OfDNA Elements) Project はそのような課題に取り組むために 2003 年から開始された国際的な研究プロジェクトである [18]。ENCODE Project はゲノム上の制御領域の特定や遺伝子発現量の定量などを ChIP-seq, RNA-seq のような NGS アプリケーションを主力に行っており、世界最大規模のデータセットを ENCODE Portal[19] を通じて提供している。また、ENCODE Project はヒトおよびマウスの一部の組織・細胞株を対象にしていたが、modENCODE Project[20,21], Genomics of Gene Regulation (GGR)[22], Roadmap Project[23] といった関連プロジェクトにより他のサンプルやモデル生物に研究対象が拡張されている。またこれらのデータの多くが ENCODE Portal から入手できる。

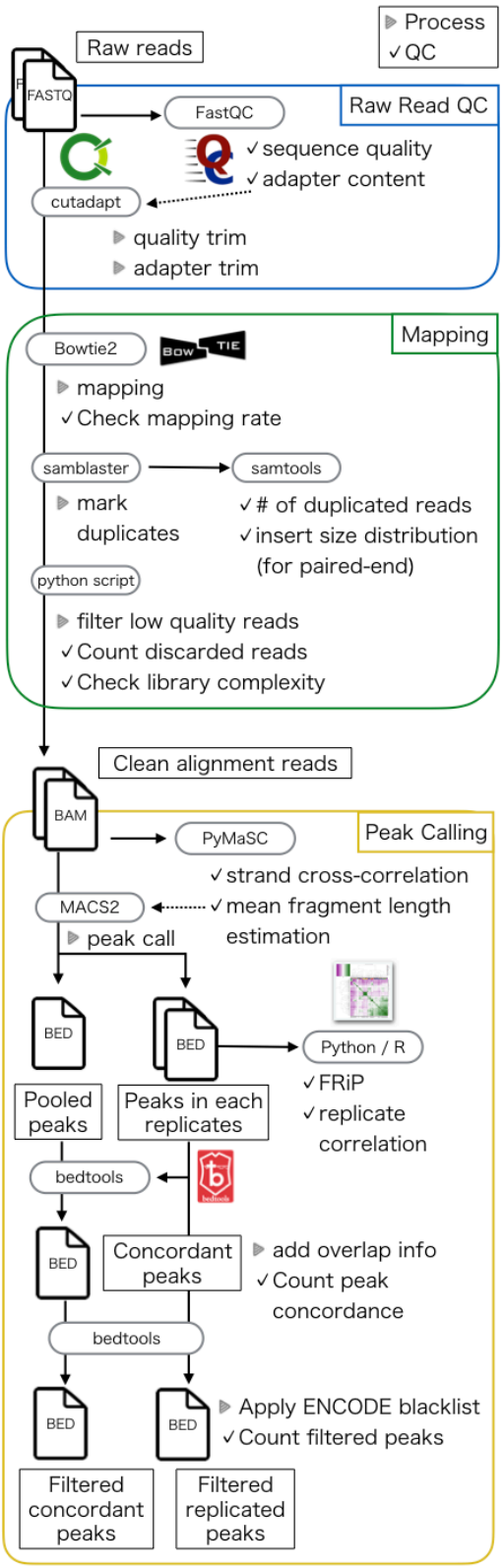


図 2: 本研究で用いる解析パイプラインの概要図

5.2 ChIP-seq データ解析

解析パイプラインに沿って各ステップでの概略を述べる。各ツールを用いる際の設定は断りのない限りデフォルトパラメータである。

1. 生データの取得 … 各データソースから FASTQ 形式の生データをダウンロードする。
2. トリミング … NGS でシーケンスされたリードにはアダプター配列など本来の DNA サンプルには含まれていない塩基配列が残る場合がある。FastQC[24] で検出されたアダプター配列は cutadapt[25] によりトリミングされる。トリミングによりリード長が 20 残基以下になったリードは破棄し、またベースコールクオリティを用いたトリミングとして、`-q 20,20` オプションを指定した。
3. マッピング … bowtie2[26] により各リードとリファレンス配列とのアライメントが行われ、各リードがゲノム上の座標にマップされる。リファレンス配列には hs37d5 を用いる（後述）。PCR などの影響で重複したリードは samblaster[27] を用いてマークし、次いで SAMTools[28] がそれらのリードやマッピングクオリティスコアの低いリードを除去した上で BAM 形式のファイルを生成する。
4. DNA 断片の平均長推定 … Single-end のサンプルではピークコールの行う際に DNA 断片の平均長を推定する必要がある。この長さを PyMaSC により精確に推定する。PyMaSC は本研究で開発したソフトウェアであり、詳細は第 III 部で述べる。
5. ピークコール … ChIP を行ったサンプルの BAM ファイルとコントロールとなるサンプルの BAM ファイルから各々リードの密度を求め、ChIP サンプルで有意にリードがエンリッチしている領域を結合部位として出力する。解析は MACS2[29] より行われ、出力形式は BED ファイルである。ChIP 実験およびコントロール実験がレプリケートを伴う場合は、レプリケートをマージした状態でのピークコール（Joint peak call）と、1 レプリケート対 1 レプリケートの全ての組み合わせでのピークコールを行う。コントロールサンプルがない場合は ChIP サンプル単独でピークコールを行う。ピークコールの対象が broad なヒストン修飾の場合は`--broad` オプションが指定される。
6. ピークの統合と Blacklist との照合 … レプリケートをマージしたピークコールで得られた領域をベースに、各々のピークが各レプリケートでも共通して検出できたかを検証しオーバーラップの情報を含めた BED ファイルを出力する。BED ファイル間の操作には bedtools[30] を用いる。Blacklist とは、ENCODE が提供している特に偽陽性が検出されやすい領域のリストである [31]。

NGS のリードを格納する FASTQ や SAM/BAM ファイルはファイルサイズが大きいため、効率的な処理を行う上で本パイプラインは極力中間ファイルを生成しないよう設計されている。例えばトリミングからマッピングまでの過程は実際には 1 ステップであり、生の FASTQ ファイルからフィルタリング済の BAM ファイルを直接生成することができる。

5.3 ChIP-seq データの品質評価

ChIP-seq では様々な要因が解析結果に影響を及ぼす可能性がある（図 3）。これらの可能性を過不足なく評価するため、ENCODE コンソーシアムが採用している QC 基準 [32] と ChIP-seq および DNase-seq の QC フレームワークである ChiLin[10] が用いている指標を元に評価する QC 指標の検討を行った（表 2）。各 QC 項目と本研究で採用した指標について概説する。

1. 不十分なサンプル収量 … ライブラリの不備により取得できるリード数が少ない場合、ChIP-seq においてはピークの検出に大きな影響を及ぼす [33]。ChiLin では具体的な基準値はない一方、ENCODE では Narrow peak の場合で 20M（million）本、Broad の場合 45M 本のリード数が推奨されている。本研究では ENCODE の基準を採用する。
2. コンタミネーション … ライブラリへの意図しない DNA の混入あるいは単にサンプルの取り違いによって不必要なリードを読んでしまう場合がある。リードのマップ率が低い場合、意図しない配列が含まれている可能性が高い。ただし、ヒトのサンプルにおけるヒト由来 DNA のコンタミネーションやヒト DNA に含まれる配列のコンタミネーションは検出することができない。
3. DNA 断片化の可否 … ChIP ターゲットにも依るが、ChIP-seq では結合部位の位置的解像度を高めるため DNA の断片長は 200bp 前後とできるだけ短く設定することが多い。そのため DNA 断片化とサイズセレクションが必須となる。実験者が意図した DNA 断片長がメタデータから得られる場合、データから平均の DNA 断片長を推定することで、このステップが正しく行われたかを評価することができる。断片長の推定には phantompeak[34] や MACS2 を用いることができるが、本研究ではより正確な推定手法 [35] を実装した PyMaSC を用いる。
4. ChIP エンリッチメント … SN 比に大きく影響するため、ChIP-seq 実験の質を左右する重要な要素である。データを元にした SN 比の評価方法としては FRiP（Fraction Reads in Peaks）が広く用いられている。本研究では加えてピークコールの手法に依存しない Strand cross-correlation を用いた指標も採用する。
5. PCR バイアス … PCR により少量の DNA のシーケンスが可能になる一方、極端に少ない DNA 量や過剰な PCR は同一の塩基配列を大量に生成し配列全体の多様性（Library complexity）が減少する。Library complexity の評価として、リードの重複率や N1 ratio, PBC といった指標 [32] があり本研究でもこれらを採用する。
6. シーケンスクオリティ … ChIP-seq 特有の QC に加えて NGS 一般のシーケンスに対する評価も必要である。本研究ではデファクトスタンダードとなっている FastQC を用いる。

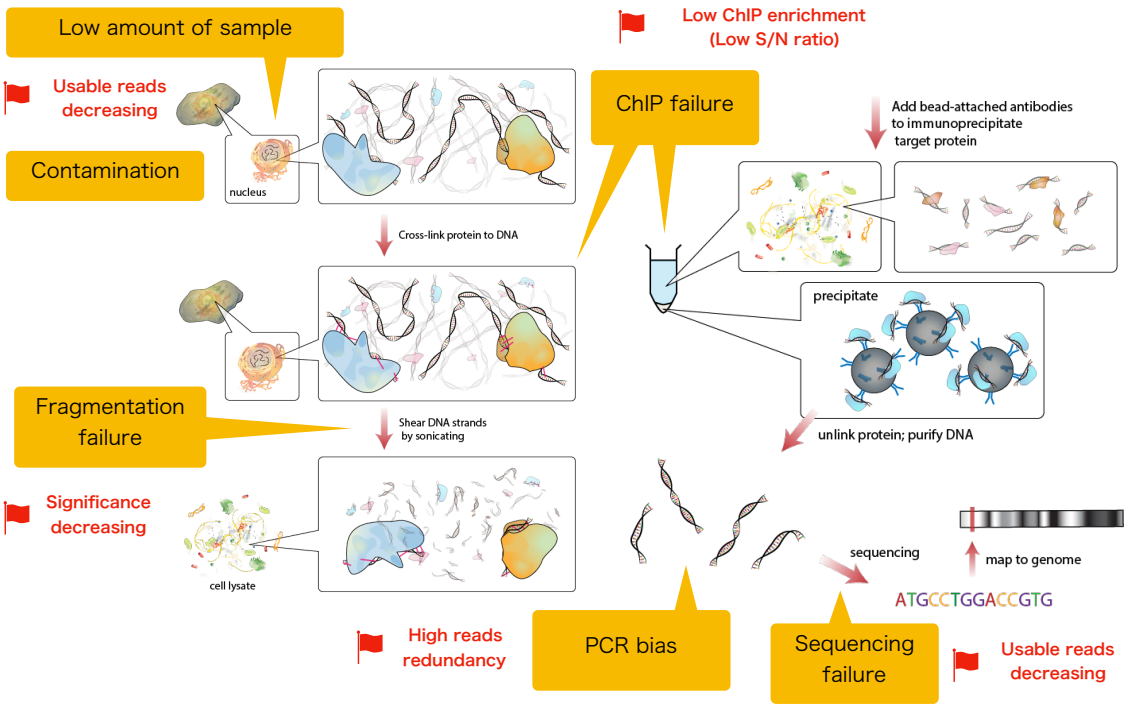


図 3: ChIP-seq の各作業段階で想定される失敗とその影響

https://en.wikipedia.org/wiki/ChIP-sequencing#/media/File:Chromatin_immunoprecipitation_sequencing.svg を元に作成

表 2: ChIP-seq 実験に対する QC 項目の比較

QC の確認項目	想定される要因例	ENCODE	ChiLin	C4S-DB
不十分なサンプル収量	ライブラリ作成時の不備	総リード数	総リード数	総リード数
コンタミネーション	環境 DNA の混入	なし	マップ率	マップ率
	サンプルの取り違い		他のリファレンスへのマップ	
DNA 断片化の可否	切断の不備	phantompeakqualtools	MACS2	PyMaSC
	サイズセレクションのミス			
ChIP エンリッチメント	cross-link の不備	FRiP	FRiP	FRiP
	抗体の失活	Strand cross-correlation		Strand cross-correlation
PCR バイアス	過剰な PCR	Library complexity	Library complexity	Library complexity
	NGS の Optical duplicates			
シーケンスクオリティ	シーケンスの失敗	なし	FastQC	FastQC

6 公共データベース SRA からのメタデータ抽出手法の開発

本節では GEO に登録されたデータを再解析を行う上での問題点、そして問題を解決するためのメタデータ抽出手法と実際に適用した結果について述べる。

6.1 現在のメタデータ整備状況

公共の NGS データのほとんどは SRA に登録されるが、本研究では GEO にも登録があるデータのみを解析対象とすることを想定している。SRA・GEO ともデータの登録時にはメタデータを共に送信する必要があるが、SRA よりも GEO の方がサンプル情報・実験手法などより詳細なメタデータが入手可能な傾向があるためである。しかしながら、GEO のメタデータに含まれる項目には記述の自由度が高く、機械的な可読性が十分に高いとはいえないため自動化処理の障害となる場合がある。NGS データのレポジトリに用いられているメタデータの例として、まずは ENCODE のある 1 実験に付与されているメタデータの一部をコード 1 に示す。

コード 1: ENCODE のメタデータ例（ENCSR000AKC JSON 形式）

```
{
    "status": "released",
    ...
    "assay_title": "ChIP-seq",
    "files": [
        {
            ...
            "replicate": {
                "experiment": {
                        ...
                        "biosample_synonyms": [
                       "A-549",
                       "A549 cell"
                    ],
                         ...
                         "assay_term_name": "ChIP-seq",
                         ...
                    "target": {
                        "status": "released",
                        "uuid": "a2cd45c9-f214-4af8-9051- ⋯
                        "name": "H3K27ac-human",
                        "title": "H3K27ac (Homo sapiens)",
                        "@id": "/targets/H3K27ac-human/",
                        "schema_version": "8",
                        "label": "H3K27ac",
                        ...
```

ENCODE の場合、特定の属性を特定のキーが一意に示し、また用いられるバリューも語彙が正確に定義されていることから、メタデータから必要な情報を抽出し、解析を自動で実行するというような実装が容易に行える [36]。

一方、GEO のサンプルに関するメタデータの例をコード 2 から 4 に示す。

コード 2: GEO のメタデータ例 1（GSE91893/GSM2323705 XML 形式）

```
<MINiML>
    <Sample iid="GSM2423705">
        <Channel position="1">
            ...
            <Characteristics tag="line">A549</Characteristics>
            <Characteristics tag="antibody">NFE2L2</Characteristics>
            ...
```

コード 3: GEO のメタデータ例 2（GSE103477/GSM3111899 XML 形式）

```
<MINiML>
    <Sample iid="GSM3111899">
        <Channel position="1">
            ...
            <Characteristics tag="cell type">monocyte derived
                macrophages (MDM)</Characteristics>
            <Characteristics tag="chip antibody">CTCF, Active Motif:
                #61311</Characteristics>
            ...
```

コード 2 と 3 はある 2 つのヒト ChIP-seq サンプルについて細胞種と ChIP ターゲットを記述した箇所を抜粋したものである。これらは XML 形式のファイルで提供されており、データの表現力としては JSON 形式と変わりない。ここで各属性は `Characteristics` タグでマークアップされているが、属性のキーとなる `tag` 属性に用いられているラベルが 2 つのサンプル間で異なっていることが分かる。このように、GEO のメタデータではキーが十分規格化されていない項目が多数存在する。また、例 1 では各項目に明確に 1 つの固有名詞が格納されているが、例 2 では細胞名が括弧付きの略称と共に記述されていたり、`chip antibody` ではターゲットのタンパク名と抗体の ID が合わせて記述されいたりと記述内容も統一されていない。

コード 4: GEO のメタデータ例 3（GSE14092/GSM353601 XML 形式）

```
<MINiML>
<Sample iid="GSM353601">
<Title>VCaP_Ethl_H3K9me3</Title>
<Accession database="GEO">GSM353601</Accession>
```

```
<Type>SRA</Type>
<Channel-Count>1</Channel-Count>
<Channel position="1">
  <Source>VCaP - Ethanol - Chromatin IP against H3K9me3</Source>
  <Organism taxid="9606">Homo sapiens</Organism>
  <Characteristics>Ethanol</Characteristics>
  <Molecule>genomic DNA</Molecule>
  <Extract-Protocol>Chromatin was fixed in …</Extract-Protocol>
</Channel>
<Description>Chromatin IP against H3K9me3</Description>
...
```

コード 4 に示した例では、細胞名や ChIP ターゲットが Characteristics タグ用いて記述されていない。また Characteristics で記述されたサンプルのトリートメントを表している Ethanol という項目については tag 属性を欠いている。細胞名や ChIP ターゲットといった情報は、Title、Source 、Description のようなタグに自然言語に近い形で記述されている。GEO のメタデータでは記述内容だけでなく、メタデータの構造についても一定していないことが分かる。

このような状態でメタデータが提供されている原因としては、必須項目を除くとメタデータの項目名や記述内容が登録者に委ねられている所が大きく、また送信されたメタデータのバリデーションを十分に行っていないためだと推察される。本研究では、公共データの大規模再解析を実施するにあたり自動化は必須であり、特にメタデータとしては生物種・サンプルソース・ChIP ターゲットが重要になる。生物種は NCBI Taxonomiy ID[37] が紐付けられるため特定は容易であるが、他 2 つについてはここまで述べてきた問題を含んでいる。ここでは、特定の生物種についてこの 2 つの情報を GEO のメタデータから抽出する手法について検討する。

6.2 抽出手法の検討

電子的に使用可能なテキストの増加に伴い、生物医学分野においても自然言語処理の技術が広く応用されるようになった。特に固有表現抽出は遺伝子・タンパク質・化学物質・医薬品・学名など分野特有のボキャブラリーが存在することからそれらに対応する最適な手法が開発されてきた[38]。ここでの問題解決においても先行研究で提案された手法が適用できる可能性がある。しかし、抗体 ID のような ChIP 特有のボキャブラリーに対応した手法はまだないこと、対象となる文章は完全なセンテンスよりはフレーズに近く単純な単語検索で十分対応可能で複雑な手法ではかえって偽陽性が高まる可能性があること、検索空間をメタデータ全体ではなく特定のタグに限定することで単純な単語のマッチングによる偽陽性をより減らせる可能性が高いことが予想されたため、本研究ではマニュアルキュレーションによる検索対象の選定と予め作成した辞書ベースの固有表現抽出を組み合わせることで対応した。

6.3 手動による検索範囲の選定

GEO に登録されたメタデータでどのようなタグが目的のメタデータの格納に使用されているか調査するため、GEO に登録された実験のうち次の条件に合致するものを検索・リスト化した。

- `Library strategy` が ChIP-Seq
- 生物種がヒト（`taxid` = 9606）
- 登録者が ENCODE コンソーシアムでない

結果、1,912 件の実験が得られ、それらに含まれる 29,498 サンプルのメタデータを取得した。次に、これらのメタデータ中にある `Characteristics` タグで `tag` 属性の値として用いられているキーのリストを作成すると計 887 種類となった。この中からサンプルソース（組織・細胞種）に関係するもの、ChIP ターゲットに関係するものをそれぞれ手動で選定した結果、各々 99 と 177 種類のタグを得た（表 3）。

ここでリストアップした `tag` の値で全サンプルのうち何割をカバーできるか検証した。結果、サンプルソースについては 91%、ChIP ターゲットのタグについては 89% のメタデータに対応できた。特に使用頻度の高い `tag` の値を表 4 および 5 に示す。サンプルの一部は複数の `Characteristics` タグを持ちうるため、割合の合計は 100% を超過することに注意されたい。

6.4 固有表現抽出を行うための辞書作成

ここでは固有表現抽出に用いる辞書の作成方法について述べる。ChIP ターゲットとしては遺伝子名とシンボルおよび抗体 ID を、サンプルソースとしてヒト由来の細胞種の名前一覧を取得した。

6.4.1 ChIP ターゲット

まず最初に遺伝子名を検出するための辞書を作成した。ヒト遺伝子名とそのシンボルは HGNC（HUGO Gene Nomenclature Committee）により管理されている [39] ため、HGNC の web サイトからヒト遺伝子の正式名称と対応する Gene Symbol の組み合わせをダウンロードした。Gene Symbol には正式なシンボルの他にエイリアスと廃止された旧称を含む。エイリアスや旧称は複数の遺伝子に対応している場合があり、このような曖昧なシンボルは辞書から除外した。また表 6 に示す単語をストップワードとして無視する。

ヒストン修飾名の検出は主に正規表現をベースに行った。コード 5 に Python の実装を示す（`search_histone_mark` 関数）。

ChIP 等に用いられる抗体には製造業者等により ID が付与されているため、抗体 ID–ターゲット名のペアを持つことで抗体 ID から ChIP ターゲットを同定することができる。抗体に関する情報を取得するため The Antibody Registry[40] から抗体情報をダウンロードした。この情報を元に前述の遺伝子辞書とヒストン修飾名の検索手法を適用し、抗体 ID–ターゲット名のペアを 4,619,069 件作成しこれらを格納した辞書を作成した。

あるサンプルが ChIP を行ったサンプルではなくコントロールサンプルである場合もある。こ

表 3: GEO に登録されたヒト ChIP-seq メタデータで用いられる重要な tag の値一覧

サンプルソースに関するもの	“Cell Line”, ”alternative cell line name”, “biomaterial_type”, “body site”, “breast cancer subtype”, “cancer”, “cancer type”, “cell”, “cell line”, “cell characteristics”, “cell description”, “cell id”, “cell identity”, “cell info.”, “cell karyotype”, “cell line”, “cell line background”, “cell line name”, “cell line origin”, “cell line source”, “cell line specificity”, “cell line specifics”, “cell line type”, “cell line/genotype”, “cell line/type”, “cell line/vendor”, “cell lineage”, “cell lines”, “cell number”, “cell organism”, “cell origin”, “cell part”, “cell passage”, “cell population”, “cell source”, “cell strain”, “cell subtype”, “cell type”, “cell type source”, “cell types”, “cell-line”, “cell-type”, “cell_line”, “cell_line/tissue”, “cell_type”, “cellline”, “cells”, “developmental stage/tissue”, “generation of cells”, “growth phase”, “histological type”, “implated cell line/type”, “line”, “line name”, “lymphoma type”, “nickname”, “organismpart”, “origin”, “original body site”, “originating cell line”, “sample common name”, “sample id”, “sorted cell type”, “source”, “source cell tyep”, “source cell type”, “source tissue”, “source type”, “source_type”, “spike-in cell line”, “stain”, “strain”, “subtype”, “t cell type”, “tiisue”, “tisssue”, “tissue”, “tissue derivation”, “tissue id”, “tissue of origin”, “tissue origin”, “tissue source”, “tissue source/type”, “tissue subtype”, “tissue type”, “tissue/cell type”, “tissue/cells”, “tissue_depot”, “tissue_type”, “tisue”, “tumor”, “tumor cell type”, “tumor id”, “tumor model”, “tumor sample”, “tumor stage”, “tumor type”, “tumor typeCell Line”, “tumour stage”
ChIP ターゲットに関するもの	“-style parameter used for homer peak calling”, “ChIP”, “ChIP Antibody”, “anitbody”, “antibodies”, “antibody”, “antibody (clone id#) used”, “antibody (antibody name, vendor, batch/lot#, catalog#)”, “antibody (catalog)”, “antibody (vender, cat#, lot#)”, “antibody antibody description”, “antibody antibodydescription”, “antibody batch/lot#”, “antibody cat#, lot#”, “antibody cat. #”, “antibody catalog”, “antibody catalog #”, “antibody catalog num”, “antibody catalog number”, “antibody catalog#”, “antibody catalog/vendor”, “antibody catalogue number”, “antibody description”, “antibody details (vendor, catalog number)”, “antibody epitope, vendor, catalog number, lot”, “antibody info”, “antibody information”, “antibody lot”, “antibody lot #”, “antibody lot num”, “antibody lot number”, “antibody lot#”, “antibody lot. #”, “antibody lot/batch number”, “antibody lot/batch number(s)”, “antibody manufacturer and catalog number”, “antibody name”, “antibody source”, “antibody target”, “antibody target description”, “antibody targetdescription”, “antibody vendor & cat. number”, “antibody vendor, cat. number”, “antibody vendor, catalog#”, “antibody vendor/catalog”, “antibody vendor/catalog#”, “antibody/-capture”, “antibody/details”, “antibody/enzyme”, “antibody_target”, “antigen”, “batch/lot#”, “beads/antibody”, “binding/selection”, “cantibody vendor/catalog#”, “cat no./lot”, “cat. #”, “catalog”, “catalog #”, “catalog number”, “catalog#”, “catalog#/vendor”, “catalog/lot#”, “catalog/vendor”, “catalogue number”, “cataog number”, “cell purification antibody catalog number”, “chip ab”, “chip anibody”, “chip anitbody”, “chip antbody”, “chip antibodies”, “chip antibody”, “chip antibody (catlogue number)”, “chip antibody (company, cat.#, clone)”, “chip antibody cat#”, “chip antibody cat. #”, “chip antibody cat. no.”, “chip antibody cat. number”, “chip antibody cat.#”, “chip antibody catalog”, “chip antibody catalog #”, “chip antibody catalog #’s”, “chip antibody catalog no.”, “chip antibody catalog number”, “chip antibody catalog#”, “chip antibody catalog/vender”, “chip antibody cataolog #”, “chip antibody details”, “chip antibody details (vendor, catalog number)”, “chip antibody host, amount, catelog number and provider”, “chip antibody info”, “chip antibody info.”, “chip antibody log #”, “chip antibody lot”, “chip antibody lot #”, “chip antibody lot # and amount used pr. chip-seq”, “chip antibody lot no.”, “chip antibody lot number”, “chip antibody lot#”, “chip antibody lot/batch #”, “chip antibody manufacturer and catalog number”, “chip antibody or biotin-streptavidin pull-down of blrp-tagged proteins”, “chip antibody ref.”, “chip antibody used for chip or oligonucleotides used for chirp (chromatin isolation by rna purification)”, “chip antibody used for chip or oligonucletides used for chirp (chromatin isolation by rna purification)”, “chip antibody vendor id”, “chip antibody vendor lot #”, “chip antibody vendor/cat. #”, “chip antibody vendor/cat.#”, “chip antibody(company, cat.#, clone)”, “chip beads/antibody”, “chip catalog number”, “chip catalog#”, “chip epitope”, “chip seq antibody”, “chip target”, “chip vendor”, “chip vendor/-catalog#”, “chip-ab”, “chip-antibody”, “chip-antibody cat. #”, “chip-antibody cat. number”, “chip-antibody lot #”, “chip-seq antibody”, “chip-seq antibody cat. #”, “chip_antibody”, “chip_antibody_catalog”, “chip_or_input”, “chip_protocol”, “chip_target”, “comment”, “control”, “control antibody”, “enriched motif ($p<0.001$) with dbd sequence similarity >65%.”, “enrichment”, “enrichment target”, “epigenetic mark”, “experiment”, “experiment type”, “experiment_type”, “expression”, “factor”, “factor chip”, “foxa1 antibody”, “foxm1 antibody chip”, “fraction”, “hgn”, “histone”, “histone mark”, “histone marks to be tested”, “histone modification”, “immunoprecipitation”, “immunprecipitation”, “input used for chip-seq peak calling”, “ip”, “ip antibody”, “library type”, “lot”, “lot #”, “lot number”, “lot#”, “lot/batch number”, “lymphoblast antibody”, “parallel chip antibody”, “primary antibody”, “procedure”, “pull-down biotap”, “pulldown”, “sample type”, “target”, “target antibody”, “target molecule”, “target protein”, “transcription factor”, “vendor/catalog#”, “vendor/catalog/lot”

コード 5: ヒストン修飾の名称を検出するアルゴリズム

```
import re
from itertools import chain

DEL = str.maketrans("[]()/", "  ", ',')

MARK_PATTERN = re.compile("H([1345]|2[AB])([.F][XZ134])?([KSTRY][1−9][0−9]∗)(ac|me[1−3]|ph|ub)".
    lower())

HISTONE_PATTERN = re.compile("Histone[−\s]\s∗(H([1345]|2[AB])(.[XZ4])?)".lower())

_HISTONE = "({0}[− ]histone)|(H([1345]|2[AB])(\.[XZ134])?[−\s]+{0})".lower()
ACCETYL_PATTERN = re.compile(_HISTONE.format("(ac|acetyl|acetylated)"))
MONOMETYL_PATTERN = re.compile(_HISTONE.format("mono[− ]?methyl"))
DIMETYL_PATTERN = re.compile(_HISTONE.format("di[− ]?methyl"))
TRIMETYL_PATTERN = re.compile(_HISTONE.format("tri[− ]?methyl"))
PHOSPHO_PATTERN = re.compile(_HISTONE.format("phospho"))
UBIQUITYL_PATTERN = re.compile(_HISTONE.format("ubiquityl"))
CROTONYL_PATTERN = re.compile(_HISTONE.format("crotonyl"))

MOD_PATTERNS = dict(
    ac=ACCETYL_PATTERN,
    me1=MONOMETYL_PATTERN,
    me2=DIMETYL_PATTERN,
    me3=TRIMETYL_PATTERN,
    ph=PHOSPHO_PATTERN,
    ub=UBIQUITYL_PATTERN,
    cr=CROTONYL_PATTERN
)

LYSINE_PATTERN = re.compile("[− ](lysine|lys|k)\s∗([1−9][0−9]∗)")
SERINE_PATTERN = re.compile("[− ](ser|s)\s∗([1−9][0−9]∗)")
THREONIE_PATTERN = re.compile("[− ](thr|t)\s∗([1−9][0−9]∗)")
ARGININE_PATTERN = re.compile("[− ](arg|r)\s∗([1−9][0−9]∗)")
TYROSINE_PATTERN = re.compile("[− ](tyr|y)\s∗([1−9][0−9]∗)")

RESIDUE_PATTERNS = dict(
    K=LYSINE_PATTERN,
    S=SERINE_PATTERN,
    T=THREONIE_PATTERN,
    R=ARGININE_PATTERN,
    Y=TYROSINE_PATTERN
)


def search_histone_mark(text):
    text = text.lower()
    mark = match_histone_mark(text)
    return mark if mark else _parse_text(text)


def match_histone_mark(text):
    match = MARK_PATTERN.search(text)
    if match:
        subunit1, subunit2, pos, mod = match.groups()
        return 'H' + subunit1.upper() + ('' if subunit2 is None else subunit2) + pos.upper() + mod


def _parse_text(text):
    text = text.rstrip().translate(DEL)

    mods = set()
    for mod, p in MOD_PATTERNS.items():
        if p.search(text):
            mods.add(mod)
    if len(mods) == 1:
        mod = mods.pop()
    else:
        return None

    foundpos = {}
    for residue, p in RESIDUE_PATTERNS.items():
        f = list(set(p.findall(text)))
        if f:
            foundpos[residue] = f
    if foundpos and sum(1 for _ in chain(∗foundpos.values())) == 1:
        residue, ((_, pos), ) = tuple(foundpos.items())[0]
    else:
        return False

    unit = set(HISTONE_PATTERN.findall(text))
    if len(unit) == 1:
        subunit, _, _ = unit.pop()
        subunit = subunit.upper()
        if ".X" in subunit or ".Z" in subunit:
            subunit = subunit.replace('.', 'F')
        return subunit + residue + pos + mod
    else:
        return False
```

表 4: サンプルソースの記述で使用頻度の高い tag の値（上位 5 つ）

tag の値	サンプル数	割合
cell line	15,053	51.03%
cell type	12,710	43.09%
tissue	4,034	13.68%
biomaterial type	1,806	6.12%
line	664	2.25%

表 5: ChIP ターゲットの記述で使用頻度の高い tag の値（上位 5 つ）

tag の値	サンプル数	割合
chip antibody	16,317	55.32%
antibody	5,905	20.02%
chip protocol	1,563	5.30%
chip antibody catalog	1,369	4.64%
chip antibody	1,351	4.58%

表 6: 遺伝子名検索時のストップワード一覧

goat 2 in of b jn ii type iga for as on aa light h tag kit ac ac2 ac5 chip tri h

表 7: コントロールサンプルの検出に用いるワード（大文字小文字問わず）

“input”, “inputdna”, “none”, “n/a”, “n.a.”, “no”, “non”, “igg”, “control”, “flag”, “background”, “--”, “mock”

れを検出するため、表 7 に示した語を含んでいた場合はコントロールサンプルとみなす。

6.4.2 サンプルソース

同じ生物種でも研究目的によりどの組織を解析するかは千差万別であり、またヒトの場合は多数の細胞株が確立されており ChIP-seq でもヒト細胞種を扱う実験は多い。細胞の名称も Gene Symbol と同じく多くの通称が存在しており、これらをユニークな名称に統一して扱う必要がある。そこで細胞種の情報を Cellosaurus[41] から取得した結果、83,151 件の細胞名とその同義語を得られた。

ヒト組織名については ENCODE で取得されたサンプルのメタデータから表 8 に示す辞書を作成した。

表 8: ヒト組織名の検出に用いる辞書

“liver”, “stomach”, “heart”, “lung”, “kidney”, “forebrain”, “midbrain”, “hindbrain”, “adrenal gland”, “spleen”, “limb”, “transverse colon”, “sigmoid colon”, “embryonic facial prominence”, “neural tube”, “upper lobe of left lung”, “small intestine”, “thyroid gland”, “gastrocnemius medialis”, “body of pancreas”, “intestine”, “heart left ventricle”, “testis”, “gastroesophageal sphincter”, “ovary”, “thymus”, “tibial nerve”, “esophagus muscularis mucosa”, “esophagus squamous epithelium”, “brain”, “placenta”, “breast epithelium”, “Peyer’s patch”, “head”, “suprapubic skin”, “chorionic villus”, “prostate gland”, “subcutaneous abdominal adipose tissue”, “skeletal muscle tissue”, “large intestine”, “lower leg skin”, “uterus”, “muscle of leg”, “muscle of arm”, “vagina”, “omental fat pad”, “pancreas”, “subcutaneous adipose tissue”, “muscle of back”, “chorion”, “layer of hippocampus”, “right atrium auricular region”, “tibial artery”, “cerebellum”, “psoas muscle”, “right lobe of liver”, “ascending aorta”, “endocrine pancreas”, “esophagus”, “spinal cord”, “thoracic aorta”, “left lung”, “parathyroid adenoma”, “trophoblast”, “heart right ventricle”, “temporal lobe”, “aorta”, “caudate nucleus”, “colonic mucosa”, “mucosa of rectum”, “placental basal plate”, “right lung”, “urinary bladder”, “digestive system”, “duodenal mucosa”, “germinal matrix”, “cingulate gyrus”, “middle frontal area 46”, “renal cortex interstitium”, “stomach smooth muscle”, “substantia nigra”, “amnion”, “angular gyrus”, “central nervous system”, “muscle layer of colon”, “muscle layer of duodenum”, “muscle of trunk”, “renal pelvis”, “adipose tissue”, “arthropod fat body”, “coronary artery”, “left kidney”, “right kidney”, “bone marrow”, “cortical plate”, “frontal cortex”, “left renal cortex interstitium”, “left renal pelvis”, “olfactory bulb”, “right cardiac atrium”, “right renal pelvis”, “mucosa of stomach”, “rectal smooth muscle tissue”, “retina”, “embryo”, “right renal cortex interstitium”, “salivary gland”, “brown adipose tissue”, “imaginal disc”, “male accessory sex gland”, “tongue”, “colon”, “occipital lobe”, “skin of body”, “diencephalon”, “gonadal fat pad”, “metanephros”, “parietal lobe”, “umbilical cord”, “arm bone”, “breast”, “camera-type eye”, “cerebellar granule layer”, “eye”, “fat pad”, “forelimb muscle”, “hindlimb muscle”, “islet of Langerhans”, “jejunum”, “leg bone”, “mole”, “pericardium”, “telencephalon”, “zone of skin”, “Ammon’s horn”, “area 11 of Brodmann”, “cerebellar cortex”, “cerebral cortex”, “cerebral cortex, layer 5”, “duodenum”, “endometrium”, “femur”, “forelimb bud”, “germinal center”, “globus pallidus”, “hindlimb bud”, “inferior parietal cortex”, “insula”, “left cardiac atrium”, “mammary gland”, “medulla oblongata”, “mesoderm”, “middle frontal gyrus”, “penis”, “pons”, “posterior cingulate cortex”, “putamen”, “rectum”, “superior temporal gyrus”, “yolk sac”

6.5 メタデータの抽出と実際のメタデータを用いたテスト

ここまで作成した辞書を用いたサンプルソースと ChIP ターゲットの検出アルゴリズムを概説する。

6.5.1 サンプルソース

1. `Characteristics` タグのうち `tag` 属性がリストアップした語句に一致する場合は、テキストにサンプルソースに該当する句が存在するかマッチングを行う。優先順位は下記の通り。複数の `Characteristics` タグに候補が存在する場合は最初にマッチしたものを返す。
   （a）細胞種名
   （b）組織名

表 9: GEO ヒト ChIP-seq メタデータから情報を抽出できたサンプルの割合

検出に成功した項目	サンプル数	割合
サンプルソースおよびターゲット	24,798	84.06%
サンプルソースのみ	2,807	9.52%
ターゲットのみ	1,736	5.89%
どちらも検出不可	261	0.88%

2. 該当しなかった場合、`Source` タグのテキストから同様にマッチングを行う。
3. 該当しなかった場合、`Title` タグのテキストから同様にマッチングを行う。
4. 該当しなかった場合、検出失敗である。

6.5.2 ChIP ターゲット

1. `Characteristics` タグのうち `tag` 属性がリストアップした語句に一致する場合は、テキストに ChIP ターゲットに該当する句が存在するかマッチングを行う。優先順位は下記の通り。複数の `Characteristics` タグに候補が存在する場合は最初にマッチしたものを返す。
   （a）遺伝子名
   （b）ヒストン修飾名
   （c）抗体 ID
   （d）Gene Symbol
2. 該当しなかった場合、`Title` タグのテキストから同様にマッチングを行う。
3. 該当しなかった場合、`Characteristics` タグのテキストを用いてコントロールサンプルに該当するか判定を行う。
4. 該当しなかった場合、`Title` タグのテキストを用いてコントロールサンプルに該当するか判定を行う。
5. 該当しなかった場合、検出失敗である。

6.5.3 検出手法のテスト

このアルゴリズムを実際に取得したメタデータに適用した結果、84% のサンプルについてメタデータを抽出することができた（表 9）。これらに偽陽性やミスアノテーションがどの程度含まれるかは今後検証する必要があるが、多くのサンプルでメタデータを半自動で抽出し、続く ChIP-seq 解析の自動化を行える目処が立ったといえる。

7 ChIP-seq 解析におけるデコイ配列の応用

本章では、1000 人ゲノムプロジェクトで導入されたデコイ配列入りヒトリファレンスゲノムの概説、ChIP-seq における応用の可能性と実際に利用した際の効果の検証について記述する。

7.1 背景

ヒトゲノム計画により最初のドラフト版ヒトゲノムが 2000 から 2001 年にかけて発表されて以降 [42,43] も、ヒトゲノムの解読とリファレンスとしての改善の努力が継続的に行われてきた。2004 年にヒトゲノム計画の解読が完了し [44]、以降は Genome Reference Consortium（GRC）が主導してヒトリファレンスゲノムの作成が行われ、現在最新のメージャーバージョンは 2013 年に公開された GRCh38（UCSC version hg38）である [45]。実用上では、アノテーション等の互換性といった観点で、1 つ前のバージョンである GRCh37（hg19）が未だに用いられていることもままある。ヒトリファレンスゲノムはゲノム解析における礎石としての役割を果たした一方、解読に用いられたヒトゲノムは民族的な偏りがあることが知られており [43]、遺伝的に距離の離れた民族のゲノム配列に対して無視できない配列や構造の差異があることも明らかになっている [46,47]。このようなギャップを埋める試みとして 2008 年に開始された 1000 人ゲノムプロジェクトでは名前の通り 1000 人以上の全ゲノムを解読することでヒトの遺伝的多様性を明らかにする計画であり [48]、2012 年には 1000 人以上のゲノム配列が公開された [49]。近年では、国際的なデファクトスタンダードとなったリファレンスゲノムに頼らず、民族ごとのリファレンスゲノムを確立する機運が高まり、様々なプロジェクトが進行中あるいは成果を発表しつつある [50–55]。

ヒトゲノム計画の完了から 20 年弱経過する現在でも、ヒトリファレンスゲノムは完全とは言えずゲノムアセンブリは研究が続いている分野であると言える。ゲノムアセンブリではまず配列をつなぎ合わせた Contig を作成し、Contig 同士を繋ぎ合わせた Scaffold を作成する。理想的には Scaffold そのものあるいはそれらの組み合わせが染色体 1 本の配列に相当しゲノムアセンブルが成功する。実際には、Scaffold の作成には成功しても、どの染色体に由来する配列か不明な場合や、染色体が判ってもその位置や向きが特定できない場合もある [56]。このようなゲノムとしての位置は不明瞭なもののゲノム中に含まれることは明らかな配列は、リシーケンシグ解析においてリファレンスの一部に含めることで、ミスマップを減らしより正確な結果を得られる可能性がある。この考えに基づき、1000 人ゲノムプロジェクトで行われたバリアントコール解析では通常の GRCh37 配列に加えてアセンブリに含まれたかった 35.4M 塩基対分の配列を追加したことでミスマップの減少を図った。これらの配列はミスマップされやすい配列を集める効果があることからデコイ配列と名付けられた [57]。

このようなデコイ配列の活用は SNP（Single Nucleotide Polymorphism, 一塩基多型）や SNV（Single Nucleotide Variant, 一塩基多様性）そしてインデル（挿入・欠損）を検出するバリアントコール解析において活用されてきたが、より正確なマッピングを行える可能性があるにもかかわらず、ChIP-seq のような他のリシーケンシングベースの手法では用いられてこなかった。ここでは、ChIP-seq 解析においてデコイ配列を含んだリファレンスゲノムを用いることでどのような効果を得られるのかを明らかにする。

7.2 テストデータの作成

7.2.1 データソース

テストデータとして ENCODE プロジェクトで公開されたヒト A549 細胞 ChIP-seq データから 119 件の実験を取得して解析した。うち 103 件が ChIP を行った実験、16 件はコントロール用の実験である。これらは 204 件のサンプルで構成されており、188 件は Single-end、16 件は Paired-end でシーケンスされていた。また ChIP ターゲットは 61 種類の転写因子と 10 種類のヒストン修飾をカバーしている。

7.2.2 解析パイプライン

解析パイプラインは 5.2 節で述べたものと同一である。

7.2.3 比較するリファレンスゲノム

デコイ配列入のリファレンスゲノムとして hs37d5（GRCh37 + デコイ配列)、比較対象として GRCh37 を用いる。リファレンス配列は 1000 人ゲノムプロジェクトのサイトで hs37d5, human_g1k_v37 として公開されている配列を利用した。これらリファレンスの概要は表 10 のとおりである。

表 10: human_g1k_v37 と hs37d5 の比較

	human_g1k_v37（GRCh37）	hs37d5（GRCh37 + デコイ配列）
染色体配列	3,095,73,981	3,095,73,981
ALT Contigs	6,110,758	6,110,758
デコイ配列	-	35,649,766

単位は bp

ALT Contigs は人種間で際の大きい領域の配列郡である。hs37d5 はデコイ配列としてアセンブリに含まれたかった配列に加え、ヒトヘルペスウイルスのゲノムを含む。

7.3 結果

7.3.1 デコイ配列によるマップ率の改善

通常のリファレンス配列に加えてデコイ配列を加えたことで、従来マップできなかったリードがデコイ配列にマップされたり、従来ユニークマップされたリードでもマルチマップすることが予想される。デコイ配列により、リードが Unmap, Unique Map, Multimap される割合がどのように変化するかを確認した。図 4 は Single-ended なサンプルの場合で図 5 は Paired-ended の場合である。予想された通り、いずれの場合も Unmapped read の割合が減少し、Multimapped read の割合が増加している。一方、Unique mapped read については他 2 つと比べると変化が小さくなっており、デコイ配列により Unique map できたリード数と Unique map だったリードが Multimapped になったリード数がおおよそ同じ割合存在することを示唆している。

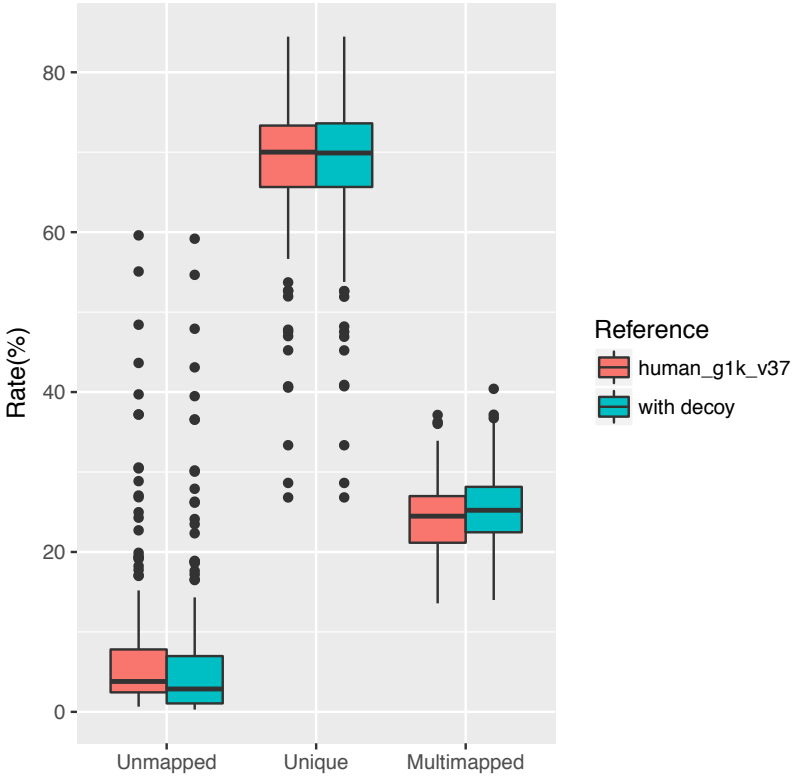


図 4: Single-end におけるデコイ配列導入によるマップ率の変化

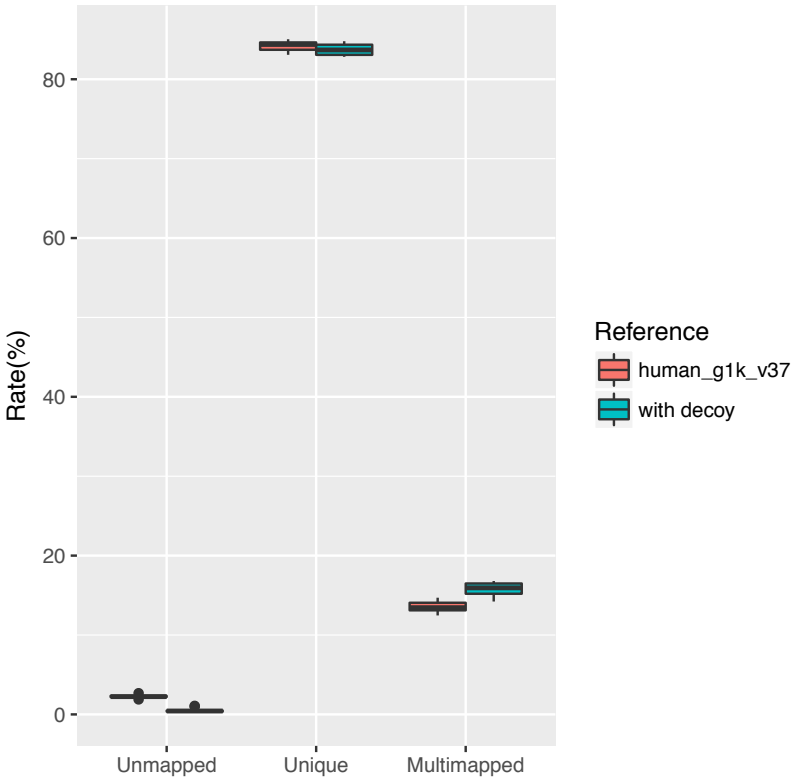


図 5: Paire-end におけるデコイ配列導入によるマップ率の変化

デコイ配列を含めることで Multimapped read が増えるということは従来不確実なリードが染色体にマップされていたということであり、デコイ配列を用いることで ChIP-seq での偽陽性を低減できる可能性がある。

7.3.2 マップ座標とクオリティスコアの変化

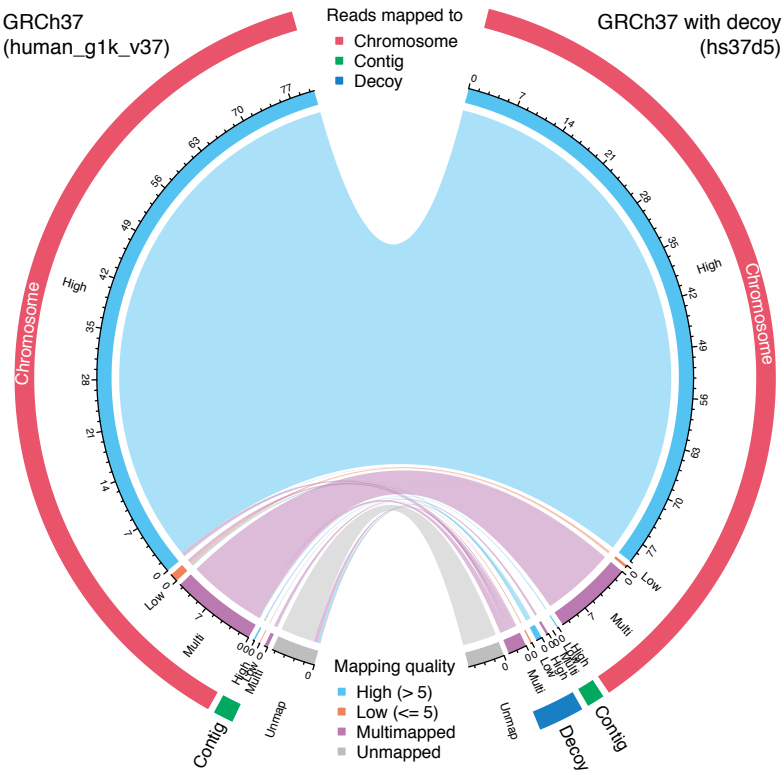
デコイ配列の効果をより詳細に解明するため、通常のリファレンスゲノムとデコイ配列入りのリファレンスゲノムを用いた場合で各リードのマップ先とマッピングクオリティスコアの変化を追跡した。図 6a から 7b はデコイ配列導入前後での各リードの変化を表した Chord Diagram である。ここでは解析で得た全てのリードを、Unique mapped（高スコア、> 5）・Unique mapped（低スコア、≤ 5）・Multimapped・Unmapped の 4 つに分類してその割合をプロットしている。まず図 6a、6b を見ると、Single-end では約 80%、Paired-end では約 90% のリードが導入前後共に染色体配列へ高いスコアでマップされていた。デコイ配列の効果に注視するため前後とも染色体配列に高いスコアでマップされたリードを抜いてプロットしたものが図 7a、7b である。Single-end, Paired-end それぞれ 3.6% と 4.7% のリードがデコイ配列に新たにマップされていた。これはデコイ配列が全リファレンスゲノムのうち 1.13% の長さでしかないことを考えると大きな影響をもたらしているといえる。特に Paired-end では染色体にユニークマップされていたリードがデコイ配列にユニークマップされていたり、マップできなかったリードの約 85% がデコイ配列にマップされるなど影響がより強くなっている。

7.3.3 局所的な効果の検証

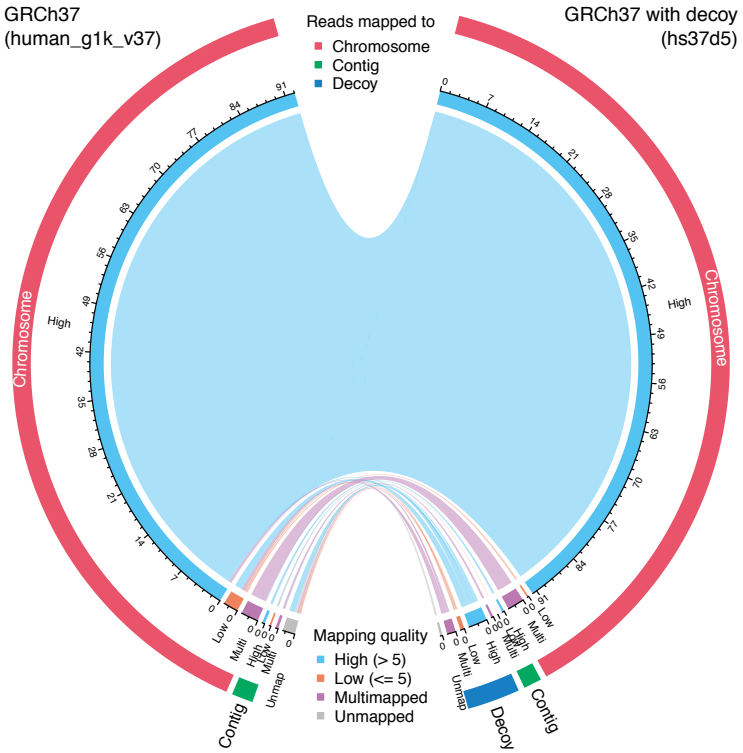
次にデコイ配列がピークコールにどのような効果を示すのか検証した。ピークコールの結果、99% のピークはデコイ配列導入前後で共通して検出された。デコイ配列が及ぼす局所的な影響について観察するため、デコイ配列の導入によりコールされなくなったピークの周辺を IGV[58] を用いて可視化した。図 8 にその一例として SP1 をターゲットにした ChIP-seq 実験を示す。各トラックの説明は表 11 にまとめた。アライメントトラックの高さは読み深度を表しており着色されている箇所はリファレンスと塩基が異なることを示している。

リファレンスが GRCh37 の場合、この領域には多くのリードがマップされている。ただしミスマッチが多いことからマップクオリティはやや低くなっていることが判る。また ChIP サンプル・コントロールサンプル共に似通ったリードの分布をしているためミスマップが疑われるが、この範囲では 1 箇所がピークとしてコールされている。一方リファレンスに hs37d5 を用いると、この領域のリードのほとんどが除外され、結果として偽陽性のピークが抑制されている。またこの領域はセントロメアに近くリピート配列が多い領域であり、ENCODE が公開しているブラックリストにもリストアップされている。

これらの結果を総合すると、リピート領域等のマッピングが難しくミスマップが発生しやすい領域にマップされたあいまいなリードがデコイ配列へのユニークマップあるいはデコイ配列とのマルチマップになった結果、染色体配列とのアライメントから取り除かれ、結果的に偽陽性の抑制につながると考えられる。

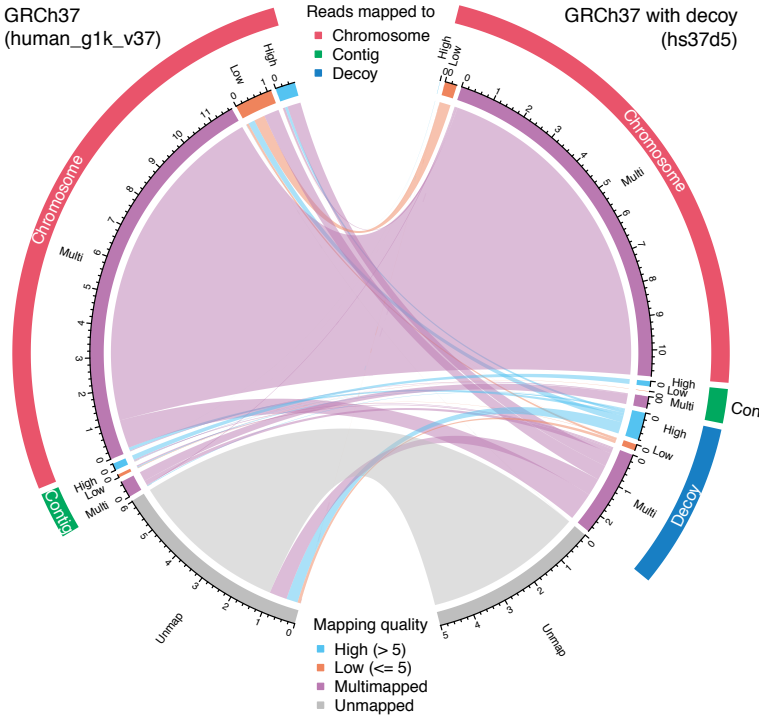


(a) Single-end における変化

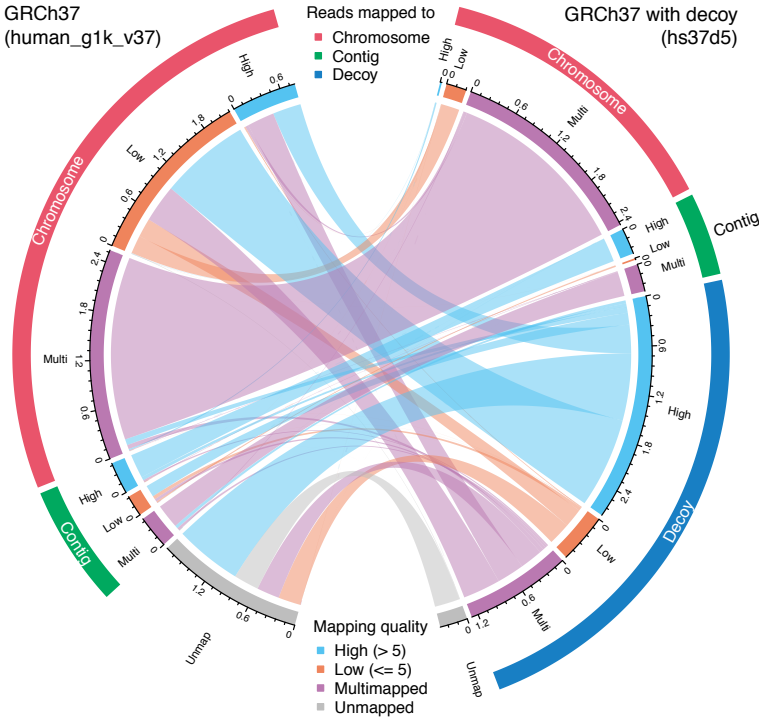


(b) Paired-end における変化

図 6: デコイ配列導入によるマップ先の変化（全体）



(a) Single-end における変化



(b) Paired-end における変化

図 7: デコイ配列導入によるマップ先の変化（高スコア→高スコアを除外）

表 11: 図 8 の各トラック概要

	ファイルの種類	リファレンス	実験 ID	説明
1	アライメント	GRCh37	ENCSR000BPE	SP1 ChIPped サンプル（ENCFF000NHM）
2	アライメント	GRCh37	ENCSR000BPD	DNA Input コントロール（ENCFF000NGH）
3	領域	GRCh37	1 vs 2	
4	アライメント	hs37d5	ENCSR000BPE	SP1 ChIPped サンプル（ENCFF000NHM）
5	アライメント	hs37d5	ENCSR000BPD	DNA Input コントロール（ENCFF000NGH）
6	領域	hs37d5	4 vs 5	
7	領域	—	—	ENCODE ブラックリスト

アライメント・・・BAM ファル

領域・・・BED ファイル

8 小括

現在も増加し続けている公共 ChIP-seq データに対し、既存の公共データベースは網羅性・信頼性の観点から十分な対応ができているとは言えない。そこで、ChIP-seq データの主要な公共データレポジトリとなっている GEO および ENCODE Portal の ChIP-seq データを効率よく処理できる解析パイプラインを開発した。本パイプラインの特色として、GEO の非正規的メタデータへの対応と種々の QC 評価、デコイ配列入りリファレンスゲノムの採用がある。これらにより既存のデータベースと比較してより多くの ChIP-seeq データに対して QC 情報を含めたデータを解析することが可能になった。また、従来バリアントコール解析にのみ用いられてきたデコイ配列について、ChIP-seq 解析における応用の可能性を検討すべく、従来のリファレンスゲノムとの比較検討を行った。結果、従来はマップ先があいまいであったリードをより正確にマップし、難読領域におけるピークの偽陽性の抑止につながることが明らかになった。これらの成果を組み込んだ解析パイプラインを用いることで、網羅性・信頼性を向上させたデータベースの開発につなげることができるといえよう。

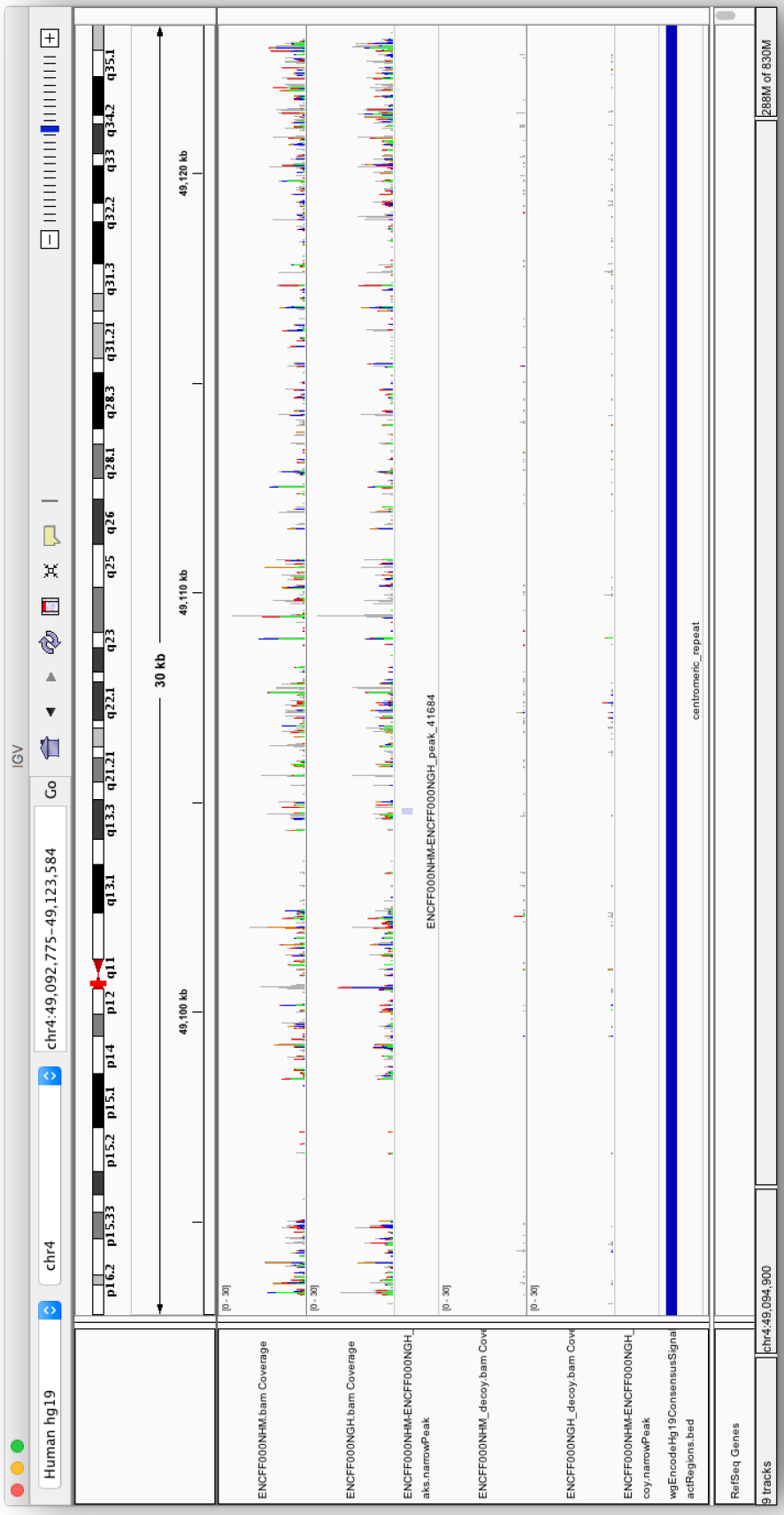


図 8: デコイ配列によりマッピングに改善が見られた領域の IGV による可視化

第 III 部
Strand cross-correlation の理論的特性評価

9 導入

NGS の発展とそれに伴うシーケンシングコストの低下により、DNA 結合タンパクのプロファイリング手法としては ChIP-sequencing 法（ChIP-seq）が主流となった。ChIP-seq を用いた研究により数多くの生物学的知見が得られてきた一方、ChIP-seq 法の実験手法としての複雑さと他の NGS を用いた実験手法と比較した際の SN 比の低さは未だに ChIP-seq 解析を難しくする原因となっている [59,60]。その上、実験手法や解析手法の差異により最終的なピークコールの結果が大きく左右することが知られている [61,62]。また、ENCODE プロジェクト [63] や ROADMAP[23] プロジェクトのような巨大プロジェクトを筆頭に ChIP-seq 解析におけるサンプル数は増加の傾向にあり、ChIP-seq のための QC 手法の重要性はより増している。

ChIP-seq 解析において各解析ステップに対してこれまで複数の QC 手法が提案されている [10]。特に、クロマチン免疫沈降の成否や検出しうるピークの数を見積もる手段として、SN 比を推測するための QC がよく用いられる。それらの中で最も一般的なのは FRiP（Fraction Reads in Peaks）と呼ばれる指標で、あるサンプルに対してピークと判定された領域に含まれるリード数の割合で表される。しかし、FRiP はピークとして判定された領域の総数（総長）に強く影響を受け [32]、そもそもピークコールは用いる手法と総取得リード数にも依存することが知られている [33,64]。したがって、ピークコール前に SN 比を評価することができればロバストで正確な QC 手法になる可能性がある。

ピークコールフリーな QC 手法としては Strand cross-correlation を用いた手法が提案されている。Strand cross-correlation は順鎖・逆鎖ごとにマップされたリードの読み深度分布を作成しその相互相関係数を計算するものである。相互相関関数としては、ピアソンの相関係数と Jaccard 係数を用いたものが提案されている [65,66]。典型的なサンプルでは、Strand cross-correlation は DNA 断片の平均長だけシフトさせた時に係数が最大値となる。またこの最大値が大きくなるほど ChIP-seq 実験としてクオリティが良いことを反映していると考えられている。Strand cross-correlation の計算はマッピングの直後に行えるためピークコールが不要で、取得したリード数に対してもロバストであることが知られている [66]。

従来 Strand cross-correlation は single-end のサンプルに対して DNA 断片の平均長を推定するために用いられてきた [29,67–69] が、ENCODE および modENCODE コンソーシアムにより Strand cross-correlation の最大値を元にした QC 指標が提案された [32]。しかし、Strand cross-correlation はシーケンシング時のリード長に相当する位置に DNA 断片長のピークとは別のピー

第 III 部の内容は下記の投稿論文を元にしている。
Hayato Anzawa, Hitoshi Yamagata and Kengo Kinoshita. Theoretical characterisation of strand cross-correlation in ChIP-seq, 30 October 2019, PREPRINT (Version 1) available at Research Square https://doi.org/10.21203/rs.2.16602/v1

ク（Phantom peak）が生じることが知られており、特に SN 比が低いサンプルにおいては本来の最大値やそのシフト長の推定に影響する場合がある [35]。この問題に対しては、Ramachandran らが 2 つの相互相関関数を提案し解決をおこなった。1 つは Naive Cross-Correlation（NCC）で、相互相関そのものはピアソンの相関係数であるが、読み深度ではなくリードの開始地点の分布を使うことと、分布のバイナリ化が従来手法と異なる点である。2 つ目は Mappability-Sensitive Cross-Correlation（MSCC）である。これは順鎖側の各塩基と、それに対応する相互相関のシフト長を加味した逆鎖側の塩基がどちらもユニークにマップ可能な領域（Doubly mappable positions）においてのみ NCC を計算するものである。MSCC は Phantom peak の影響を排除することができ、正しい断片長や最大値の推定を行うことができる。また分布がバイナリなため計算コストが比較的小さいというメリットもある。

QC 指標としてのメリットが存在する Strand cross-correlation であるが、これまで理論的な考察は行われておらず、Strand cross-correlation を元にした QC 指標とその基準値は経験的な結果に基づいて作成されてきた。そこで本研究ではより理論的根拠に基づいた QC 手法を提案することを目指し、Strand cross-correlation の特性、特に既に提案されている QC 指標の計算で用いられる Strand cross-correlation の最小値と最大値について NCC と MSCC を用いて検討する。

10　モデリングによる理論的な相互相関係数の導出

Strand cross-correlation として NCC および MSCC の理論的な考察を行うため、まず ChIP-seq のリード分布をモデリングする。次にモデルに基づきリードの観測確率を導出することで相互相関係数の最小値と最大値の理論値を期待値として計算する。

10.1　ChIP-seq におけるリード分布のモデリング

本節では ChIP-seq のリード分布のモデルを導入する。図 9 にモデルの概要図を示す。ここでは簡単のため長さ G の単一の染色体のみを含むゲノムを仮定する。複数の染色体を含むゲノムの計算方法については 11.2 節で述べる。ゲノム中には n 箇所の結合部位があるとする。各結合部位には順鎖・逆鎖側それぞれに長さ w のエンリッチ領域があり、それらは結合部位を挟んで d 離れているものとする。簡単のため、各結合部位のエンリッチ領域は $d+w$ よりも離れているものとする。すなわち別の結合部位にあるエンリッチ領域同士が重なる状況はここでは扱わない。ゲノム上の座標 $i \in \{1, 2, \ldots, G\}$ において、$f(i)$ と $g(i)$ をそれぞれ順鎖・逆鎖の位置 i でリードが存在するかどうかを示すバイナリ関数とする。ここで、関数 f および g はマップされたリードの開始地点を表すものとする。すなわち、長さ R のリードが順鎖側の位置 i から $i+R-1$ までマップされた場合、$f(i)=1$ となる。また、同じ位置の逆鎖にリードがマップされた場合は $g(b+R-1)=1$ となる。

次に、順鎖・逆鎖側には同数のリードがマップされると仮定する。

$$\sum_{i=1}^{G} f(i) = \sum_{i=1}^{G} g(i) = \frac{M_{\mathrm{u}}}{2}, \tag{1}$$

ここで M_{u} は重複するリードを抜いた後の総リード数である。ただし、実際の総リード数を M と

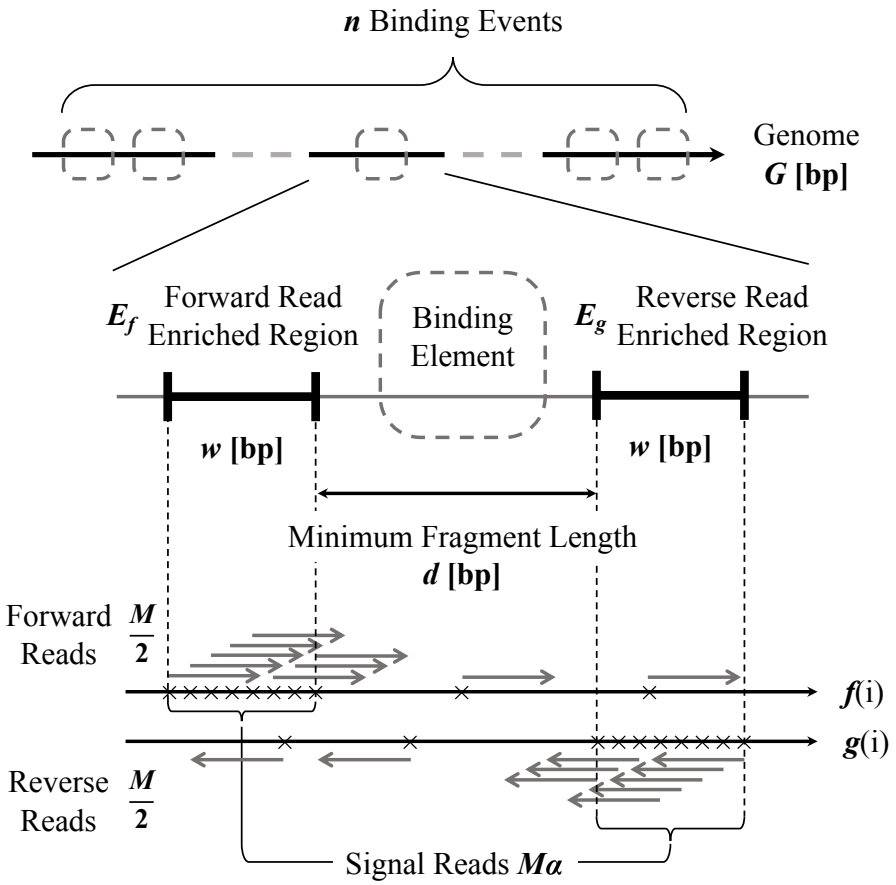


図 9: ChIP-seq のリード分布モデルの概要図

して、通常の ChIP-seq 実験では M は G より十分小さい。したがって、DNA 断片が偶然同じ位置で切断される確率は非常に低いと考えられる。そこで、必要になるまで M を M_{u} の代わりに用いる。続いて、サンプルの SN 比に応じてリードがエンリッチ領域に集中する現象をモデル化する。クロマチン免疫沈降により得られた DNA 断片のリードをシグナルリード、ノイズとなる DNA 断片をシーケンスした場合のリードをノイズリードと呼称する。これらのリードの分布を表すため、まず集合 E_f と E_g を順鎖および逆鎖側のエンリッチ領域の集合とする。ここで混合パラメータ α を導入する。α はシグナルリードとノイズリードの比を表す 0 から 1 の間の値を取るパラメータである。これらのパラメータの関係は次のような式で表される。

$$\sum_{i \in E_f} f(i) = \sum_{i \in E_g} g(i) = \frac{M}{2}\alpha \tag{2}$$

$$\sum_{i \notin E_f} f(i) = \sum_{i \notin E_g} g(i) = \frac{M}{2}(1-\alpha) \tag{3}$$

式 (2) はシグナルリードの総数を表し、式 (3) はノイズリードの総数を表す。

10.2 モデルに基づく cross-correlation の期待値の算出

関数 f と g のシフト長 x における NCC は次のように表される [35]。

$$\begin{aligned}\mathrm{NCC}(f,g)(x) &= \frac{1}{G-x}\frac{\sum_{i=1}^{G-x}(f(i)-\mu_f)(g(i+x)-\mu_g)}{\sqrt{\sigma_f\sigma_g}} \\ &\approx \frac{\frac{1}{G-x}\left(\sum_{i=1}^{G-x} f(i)g(i+x)\right)-\mu_f\mu_g}{\sqrt{\sigma_f\sigma_g}}\end{aligned} \tag{4}$$

ここで μ_f と μ_g は関数 f および g の平均、また σ_f と σ_g は分散である。順鎖・逆鎖共に同数のリード数がマップされると仮定したため、平均と分散は次のように書くことができる。

$$\mu \coloneqq \mu_f = \mu_g = \frac{M}{2G} \tag{5}$$

$$\sigma \coloneqq \sigma_f = \sigma_g = \mu(1-\mu) \tag{6}$$

式 (4) に現れる総和は $f(i)$ および $g(i+x)$ が同時に 1 になる位置 i の数え上げに等しい。そこで、そのような i の集合を D_x と置き、集合の大きさを $|D_x|$ とする。期待値 $\langle|D_x|\rangle$ を得ることができれば NCC の期待値が得られることになる。

$$\langle\mathrm{NCC}(f,g)(x)\rangle = \frac{1}{\sigma}\left(\frac{\langle|D_x|\rangle}{G-x}-\mu^2\right) \tag{7}$$

次に、$|D_x|$ を $f(i)$ と $g(i+x)$ が同時に 1 になる確率を求めることで推定する。$P_{f=1}(i)$ を f が i において 1 になる確率とし、同様に $P_{g=1}(i)$ を g についての確率とする、これらの確率を用いると、$|D_x|$ の期待値は次のように書ける。

$$\langle|D_x|\rangle = \sum_{i=1}^{G-x} P_{f=1}(i)P_{g=1}(i+x) \tag{8}$$

一様分布を仮定すると、i において順鎖のシグナルリードを観測する確率 $P_{\mathrm{S},f=1}(i)$ とノイズリードの確率 $P_{\mathrm{N},f=1}(i)$ は次のようになる。

$$P_{\mathrm{S},f=1}(i) = \begin{cases}\frac{M}{2nw}\alpha & \text{if } i\in E_f \\ 0 & \text{if } i\notin E_f\end{cases} \tag{9}$$

$$P_{\mathrm{N},f=1}(i) = \frac{M}{2G}(1-\alpha). \tag{10}$$

ここでは $M\alpha \le 2nw$ および $M(1-\alpha)\le 2G$ とする。これらの条件の妥当性については 10.2.1 で述べる。これらの確率を用いて $f(i)$ が 1 になる確率 $P_{f=1}(i)$ は次のように書ける。

$$\begin{aligned}P_{f=1}(i) &\coloneqq 1-\left(1-P_{\mathrm{S},f=1}(i)\right)\left(1-P_{\mathrm{N},f=1}(i)\right) \\ &= \begin{cases}p_{\mathrm{S}} & \text{if } i\in E_f \\ p_{\mathrm{N}} & \text{if } i\notin E_f\end{cases}\end{aligned} \tag{11}$$

ここで用いている p_{S} と p_{N} は後で展開する。ここで、式 (9) と (10) に現れるパラメータは順鎖・逆鎖に依存しないことから、$P_{g=1}(i)$ も同様に得られる。すなわち、$P_{g=1}(i)$ も同様に p_{S} と p_{N} も

用いて書き表される。

$$P_{g=1}(i) = \begin{cases} p_{\mathrm{S}} & \text{if } i \in E_g \\ p_{\mathrm{N}} & \text{if } i \notin E_g \end{cases} \tag{12}$$

したがって、$P_{f=1}(i)P_{g=1}(i+x)$ は次のように展開できる。

$$P_{f=1}(i)P_{g=1}(i+x) = \begin{cases} p_{\mathrm{S}}^2 & \text{if } i \in X_{\mathrm{SS}} \\ p_{\mathrm{S}}p_{\mathrm{N}} & \text{if } i \in X_{\mathrm{SN}} \\ p_{\mathrm{N}}p_{\mathrm{S}} & \text{if } i \in X_{\mathrm{NS}} \\ p_{\mathrm{N}}^2 & \text{if } i \in X_{\mathrm{NN}} \end{cases} \tag{13}$$

ここで

$$\begin{aligned} X_{\mathrm{SS}} &:= \{i \mid (i \in E_f) \wedge (i+x \in E_g)\} \\ X_{\mathrm{SN}} &:= \{i \mid (i \in E_f) \wedge (i+x \notin E_g)\} \\ X_{\mathrm{NS}} &:= \{i \mid (i \notin E_f) \wedge (i+x \in E_g)\} \\ X_{\mathrm{NN}} &:= \{i \mid (i \notin E_f) \wedge (i+x \notin E_g)\} \end{aligned} \tag{14}$$

である。これらの集合の大きさは x に依存し、エンリッチ領域の重なり方に応じて各々を数え上げることで得られる。最終的に $|D_x|$ は次のように得られる。

$$\langle |D_x| \rangle = |X_{\mathrm{SS}}|p_{\mathrm{S}}^2 + (|X_{\mathrm{SN}}| + |X_{\mathrm{NS}}|)p_{\mathrm{S}}p_{\mathrm{N}} + |X_{\mathrm{NN}}|p_{\mathrm{N}}^2 \tag{15}$$

10.2.1 不飽和条件と飽和条件

式（9）と式（10）においては、各リードはユニークにマップされた位置があり各々は重複していないことを暗に仮定していた。すなわち式（9）および式（10）が成り立つ条件はそれぞれ

$$\frac{M}{2}\alpha \leq nw \tag{16}$$

$$\frac{M}{2}(1-\alpha) \leq G \tag{17}$$

となる。典型的な哺乳類の ChIP-seq 実験では $M < G$ であり、式（17）は広く成り立つであろうと考えられる。一方で式（16）では、小さい nw や大きい $M\alpha$ がこの条件を破る可能性がある。すなわち $\frac{M}{2}\alpha$ が nw を上回っている状況では、エンリッチ領域においていくらかのシグナルリードが重複していることを示している。これらの状況を扱うため、式（16）が破られる場合を「飽和条件」、満たされる場合を「不飽和条件」と呼び区別する。

10.2.2 不飽和条件下の観測確率

不飽和条件における p_{S} と p_{N} は式（9）と式（10）までの導出に示した通りである。これらを展開した結果は次のようになる。

$$\text{Unsaturated Case} := \begin{cases} \mu & = \frac{M}{2G} \\ p_{\mathrm{S}} & = \frac{M}{2G}(1-\alpha) + \frac{M}{2nw}\alpha\left(1 - \frac{M}{2G}(1-\alpha)\right) \\ p_{\mathrm{N}} & = \frac{M}{2G}(1-\alpha) \end{cases} \tag{18}$$

10.2.3 飽和条件下の観測確率

飽和条件においては新たな 2 つの仮定を導入する。まず全てのエンリッチ領域にリードがマップされているとする。また M には重複したリードも含まれているため、f および g の平均を計算する際には M_{u} を代わりに用いる必要がある。

$$\text{Saturated Case} \coloneqq \begin{cases} \mu &= \frac{M_{\mathrm{u}}}{2G} \\ p_{\mathrm{S}} &= 1 \\ p_{\mathrm{N}} &= \frac{M}{2G}(1-\alpha) \end{cases} \tag{19}$$

10.2.4 不飽和条件下における期待値

最終的に、不飽和条件下における NCC の期待値は式 (4)（15）を元に p_{S}、p_{N}、μ および式 (14) で示した領域の長さを与えることで得られる。ここではまず NCC を最小および最大にする x を元に各集合の長さを与える。

もし順鎖・逆鎖側のエンリッチ領域が重複することがなければ NCC は最小になる。すなわち $0 \leq x \leq d$ の場合である。ここではそのような x を x_0 とする。このとき各集合の大きさは次のように表される。

$$\begin{aligned} |X_{\mathrm{SS}}| &= 0 \\ |X_{\mathrm{SN}}| &= nw \\ |X_{\mathrm{NS}}| &= nw \\ |X_{\mathrm{NN}}| &= G - 2nw - x_0 \end{aligned} \tag{20}$$

全てのエンリッチ領域がちょうど重なる場合、すなわち $x = d + w$ のとき明らかに NCC は最大値となる。この時の各領域の長さは次のようになる。

$$\begin{aligned} |X_{\mathrm{SS}}| &= nw \\ |X_{\mathrm{SN}}| &= 0 \\ |X_{\mathrm{NS}}| &= 0 \\ |X_{\mathrm{NN}}| &= G - nw - (d + w) \end{aligned} \tag{21}$$

不飽和条件については式（18）で示したように仮定した。式（20）に従うと、理論的な最小値は次のように計算できる。

$$\begin{aligned} \mathrm{NCC}(f,g)(x_0) &= -\frac{M\alpha\big(M(1-\alpha)^2 + (G+x_0)\alpha - 2x_0\big)}{(2G-M)(G-x_0)} \\ &\approx -\frac{M\alpha\big(M(1-\alpha)^2 + G\alpha\big)}{(2G-M)G} \\ &\approx 0 \end{aligned} \tag{22}$$

最初の近似は $x_0 \ll G$ と仮定しており通常この仮定は妥当である。もし M が G より十分小さと仮定できるならば 2 番目の近似が成り立ち、相関係数の最小値はおおよそ 0 に近似される。

最大値は式（21）で計算できる。

$$\begin{aligned}\mathrm{NCC}(f,g)(d+w) &= \frac{M\alpha^2\big(2G-(1-\alpha)M\big)^2}{4nw(2G-M)(G-(d+w))} - \frac{M\alpha\big(M(1-\alpha)^2+(G+d+w)\alpha-2(d+w)\big)}{(2G-M)(G-(d+w))} \\ &\approx \frac{M\alpha^2\big(2G-(1-\alpha)M\big)^2}{4nw(2G-M)G} - \frac{M\alpha\big(M(1-\alpha)^2+G\alpha\big)}{(2G-M)G} \\ &\approx \frac{M}{2nw}\alpha^2\end{aligned} \tag{23}$$

10.2.5 飽和条件下における期待値

飽和条件の場合については式（18）の代わりに式（19）を用いることで計算できる。まず最小値は次のような結果を得られる。

$$\begin{aligned}\mathrm{NCC}(f,g)(x_0) &= -\frac{1}{G-x_0}\frac{2nw}{M_\mathrm{N}+2nw}\frac{M_\mathrm{N}^2-2M_\mathrm{N}x_0+2nw(G-x_0)}{2G-(M_\mathrm{N}+2nw)} \\ &\approx -\frac{1}{G}\frac{2nw}{M_\mathrm{N}+2nw}\frac{2Gnw+M_\mathrm{N}^2}{2G} \\ &\approx 0\end{aligned} \tag{24}$$

ここで $M_\mathrm{N}=(1-\alpha)M_\mathrm{u}$ である。また最大値は次のとおりである。

$$\begin{aligned}\mathrm{NCC}(f,g)(d+w) &= \frac{1}{G-(d+w)}\frac{2nw}{M_\mathrm{N}+2nw}\frac{2\big(G-\frac{M_\mathrm{N}}{2}\big)^2-\big(M_\mathrm{N}^2-2M_\mathrm{N}(d+w)+2nw(G-(d+w))\big)}{2G-(M_\mathrm{N}+2nw)} \\ &\approx \frac{1}{G}\frac{2nw}{M_\mathrm{N}+2nw}\frac{2\big(G-\frac{M_\mathrm{N}}{2}\big)^2-\big(M_\mathrm{N}^2-2M_\mathrm{N}(d+w)+2nwG\big)}{2G-(M_\mathrm{N}+2nw)} \\ &\approx \frac{2nw}{M_\mathrm{N}+2nw} = \frac{nw}{\frac{M_\mathrm{u}}{2}(1-\alpha)+nw}\end{aligned} \tag{25}$$

10.2.6 MSCC を用いた場合の期待値

ここまで NCC の場合について最小値・最大値の理論値の導出について述べたが、MSCC を用いた場合はどのようになるだろうか。ここでは NCC における導出手順に沿って MSCC の場合について導出を行う。MSCC は次のように定義される [35]。

$$\mathrm{MSCC}(f,g)(x) = \frac{\frac{1}{|U^x|}\big(\sum_{i\in U^x} f(i)g(i+x)\big)-\mu_f^x\mu_g^x}{\sqrt{\sigma_f^x\sigma_g^x}} \tag{26}$$

ここで U^x はシフト長 x での Doubly mappable position を表す集合である。μ_f^x, μ_g^x, σ_f^x, σ_g^x は U^x に対応する平均および分散である。すなわち、$\mu_f^x=\sum_{i\in U^x} f(i)/|U^x|$, $\mu_g^x=\sum_{i\in U^x} g(i)/|U^x|$, $\sigma_f^x=\mu_f^x(1-\mu_f^x)$, $\sigma_g^x=\mu_g^x(1-\mu_g^x)$ である。

MSCC における状況を定式化するため、“Doubly mappable ratio” として β を導入する。これはゲノム長 G のうち Doubly mappable position の割合を示すものである。

$$\beta := \frac{|U^x|}{G} \tag{27}$$

また DNA 断片は Mappability に依らずゲノム中から等確率に得られると仮定する。すなわち Doubly mappable position の内に存在するリードは次のように書ける。

$$\sum_{i\in U^x} f(i) = \sum_{i\in U^x} g(i) = \frac{M}{2}\beta \tag{28}$$

同様に、結合部位も Doubly mappable position とは独立に分布していると仮定する。すなわち n 箇所ある結合部位のうち $n\beta$ 箇所が Doubly mappable position に含まれるとする。

$$n^x \coloneqq n\beta \tag{29}$$

$$\frac{\sum_{i\in E_f^x} f(i)}{\sum_{i\in E_f} f(i)} = \frac{\sum_{i\in E_g^x} g(i)}{\sum_{i\in E_g} g(i)} = \beta \tag{30}$$

ここで $E_f^x \coloneqq E_f \cap U^x$、$E_g^x \coloneqq E_g \cap U^x$ である。したがって、U^x 中に入るリード数は次のように書き表せる。

$$\sum_{i\in E_f^x} f(i) = \sum_{i\in E_g^x} g(i) = \frac{M}{2}\alpha\beta \tag{31}$$

$$\sum_{i\in B_f^x} f(i) = \sum_{i\in B_g^x} g(i) = \frac{M}{2}(1-\alpha)\beta \tag{32}$$

ここで $B_f^x = \{i \mid (i \notin E_f^x) \wedge (i \in U^x)\}$ であり $B_g^x = \{i \mid (i \notin E_g^x) \wedge (i \in U^x)\}$ である。よって MSCC の平均は

$$\mu_f^x = \frac{\sum_{i\in U^x} f(i)}{|U^x|} = \frac{M\beta/2}{G\beta} = \frac{M}{2G} = \mu \tag{33}$$

$$\mu_g^x = \frac{\sum_{i\in U^x} g(i)}{|U^x|} = \frac{M\beta/2}{G\beta} = \frac{M}{2G} = \mu \tag{34}$$

となるから、$\mu_f^x = \mu_g^x = \mu$ であり $\sigma_f^x = \sigma_g^x = \sigma$ となる。ここまでの結果を用いると式（26）は次のように書き直される。

$$\mathrm{MSCC}(f,g)(x) = \frac{1}{\sigma}\left(\frac{\sum_{i\in U^x} f(i)g(i+x)}{G\beta - x} - \mu^2\right) \tag{35}$$

したがって、MSCC の期待値は次のように書くことができる。

$$\langle \mathrm{MSCC}(f,g)(x)\rangle = \frac{1}{\sigma}\left(\frac{\langle |D_{U^x}|\rangle}{G\beta - x} - \mu^2\right) \tag{36}$$

ここで $D_{U^x} = \{i \mid (i \in U_x) \wedge (f(i)g(i+x) = 1)\}$ である。$|D_{U^x}|$ を推定するため、U^x 内で f および g が 1 になる確率 $P_{f=1}^x(i)$ と $P_{g=1}^x(i)$ を求める。

$$\langle |D_{U^x}|\rangle = \sum_{i\in U^x} P_{f=1}^x(i) P_{g=1}^x(i+x) \tag{37}$$

NCC の場合と同様にまず順鎖に着目する。U^x 内で順鎖のシグナルリードが観測される確率 $P_{\mathrm{S},f=1}^x(i)$ とノイズリードが観測される確率 $P_{\mathrm{N},f=1}^x(i)$ を用いると、$P_{f=1}^x(i)$ は次のように展開される。

$$P_{f=1}^x(i) = 1 - \left(1 - P_{\mathrm{S},f=1}^x(i)\right)\left(1 - P_{\mathrm{N},f=1}^x(i)\right) \coloneqq \begin{cases} p_{\mathrm{S}}^x & \text{if } i \in E_f^x \\ p_{\mathrm{N}}^x & \text{if } i \notin E_f^x \end{cases} \tag{38}$$

同様に、逆鎖についても同じ結果を得る。

$$P^x_{g=1}(i) = \begin{cases} p^x_{\mathrm{S}} & \text{if } i \in E^x_g \\ p^x_{\mathrm{N}} & \text{if } i \notin E^x_g \end{cases} \tag{39}$$

したがって、$P^x_{f=1}(i)P^x_{g=1}(i+x)$ を展開すると、

$$P^x_{f=1}(i)P^x_{g=1}(i+x) = \begin{cases} (p^x_{\mathrm{S}})^2 & \text{if } i \in X^x_{\mathrm{SS}} \\ p^x_{\mathrm{S}}p^x_{\mathrm{N}} & \text{if } i \in X^x_{\mathrm{SN}} \\ p^x_{\mathrm{N}}p^x_{\mathrm{S}} & \text{if } i \in X^x_{\mathrm{NS}} \\ (p^x_{\mathrm{N}})^2 & \text{if } i \in X^x_{\mathrm{NN}} \end{cases} \tag{40}$$

ここで

$$\begin{aligned} X^x_{\mathrm{SS}} &:= \{i \mid (i \in E^x_f) \wedge (i+x \in E^x_g)\} \\ X^x_{\mathrm{SN}} &:= \{i \mid (i \in E^x_f) \wedge (i+x \notin E^x_g)\} \\ X^x_{\mathrm{NS}} &:= \{i \mid (i \notin E^x_f) \wedge (i+x \in E^x_g)\} \\ X^x_{\mathrm{NN}} &:= \{i \mid (i \notin E^x_f) \wedge (i+x \notin E^x_g)\} \end{aligned} \tag{41}$$

である。ここで、結合部位と Doubly mappable position は独立に分布していると仮定したことから、これらの集合の大きさは式（14）で定義した集合の大きさの β 倍で近似できると仮定する。

$$\begin{aligned} |X^x_{\mathrm{SS}}| &\approx \beta |X_{\mathrm{SS}}| \\ |X^x_{\mathrm{SN}}| &\approx \beta |X_{\mathrm{SN}}| \\ |X^x_{\mathrm{NS}}| &\approx \beta |X_{\mathrm{NS}}| \\ |X^x_{\mathrm{NN}}| &\approx \beta |X_{\mathrm{NN}}| \end{aligned} \tag{42}$$

最終的に $\langle |D_{U^x}| \rangle$ は次のような形で得られる。

$$\begin{aligned} \langle |D_{U^x}| \rangle &= |X^x_{\mathrm{SS}}|(p^x_{\mathrm{S}})^2 + (|X^x_{\mathrm{SN}}| + |X^x_{\mathrm{NS}}|)p^x_{\mathrm{S}}p^x_{\mathrm{N}} + |X^x_{\mathrm{NN}}|(p^x_{\mathrm{N}})^2 \\ &\approx \beta \left(|X_{\mathrm{SS}}|(p^x_{\mathrm{S}})^2 + (|X_{\mathrm{SN}}| + |X_{\mathrm{NS}}|)p^x_{\mathrm{S}}p^x_{\mathrm{N}} + |X_{\mathrm{NN}}|(p^x_{\mathrm{N}})^2 \right) \end{aligned} \tag{43}$$

不飽和条件下では $P^x_{\mathrm{S},f=1}(i)$ および $P^x_{\mathrm{N},f=1}(i)$ は次のように展開できる。

$$P^x_{\mathrm{S},f=1}(i) = \begin{cases} \dfrac{\sum_{i \in E^x_f} f(i)}{n^x w} = \dfrac{\frac{M\alpha\beta}{2}}{n\beta w} = \dfrac{M}{2nw}\alpha & \text{if } i \in E^x_g \\ 0 & \text{if } i \notin E^x_g \end{cases} \tag{44}$$

$$P^x_{\mathrm{N},f=1}(i) = \frac{\sum_{i \in E^x_f} f(i)}{|U^x|} = \frac{M(1-\alpha)\frac{\beta}{2}}{G\beta} = \frac{M}{2G}(1-\alpha) \tag{45}$$

すなわち式（9）（10）と同じ結果である。また逆鎖の場合も同様であるから、結果として $p^x_{\mathrm{S}} = p_{\mathrm{S}}$ および $p^x_{\mathrm{N}} = p_{\mathrm{N}}$ を得る。飽和条件下でも NCC の場合と同じ仮定を用いる。すなわち、$p^x_s = 1$ と $\mu = \frac{M_u}{2G}$ である。こちらの場合も結果として $p^x_{\mathrm{S}} = p_{\mathrm{S}}$ および $p^x_{\mathrm{N}} = p_{\mathrm{N}}$ を得る。よって式（43）は次のように書き換えられる。

$$\langle |D_{U^x}| \rangle \approx \beta \left(|X_{\mathrm{SS}}|p^2_{\mathrm{S}} + (|X_{\mathrm{SN}}| + |X_{\mathrm{NS}}|)p_{\mathrm{S}}p_{\mathrm{N}} + |X_{\mathrm{NN}}|p^2_{\mathrm{N}} \right) = \beta \langle |D_x| \rangle \tag{46}$$

すなわち式（36）は以下のように書き直される。

$$\langle \mathrm{MSCC}(f,g)(x) \rangle \approx \frac{1}{\sigma} \left(\frac{\beta \langle |D_x| \rangle}{G\beta - x} - \mu^2 \right) \tag{47}$$

ここで、$G\beta \gg x$ であることを利用して式（7）と比較すると、

$$\begin{aligned}\langle \mathrm{MSCC}(f,g)(x)\rangle &\approx \frac{1}{\sigma}\left(\frac{\langle |D_x|\rangle}{G}-\mu^2\right)\\ &\approx \frac{1}{\sigma}\left(\frac{\langle |D_x|\rangle}{G-x}-\mu^2\right)=\langle \mathrm{NCC}(f,g)(x)\rangle\end{aligned} \tag{48}$$

であるから、NCC と MSCC からはほとんど同じ値を得られることが期待される。

11 予測結果の実証

ここではシミュレーションデータと実データを用いた検証に先立ち、必要となるデータの準備や処理およびツールについて述べる。作成したデータの一部とツールは https://pymasc.sb.ecei.tohoku.ac.jp で公開している。

11.1 Mappability の計算

MSCC を計算するにあたり、ゲノムに対する Mappability の情報、すなわちユニークにマップ可能な領域のリストが事前情報として必要になる。本研究では、UCSC Genome Browser[70] で公開されている ENCODE の Mappability トラックと同じ手法でヒトリファレンスゲノムに対して Mappability の計算を行った。計算には GEM mappability program [71] (GEM-indexer build 1.423, GEM-mappability build 1.315, GEM-2-wig build 1.423) を用いて必要となったリード長ごとに Mappability のデータを作成した。また 2 塩基のミスマッチまで許容した。作成されたプロファイルは、ユニークマップ可能な領域（Mappability が 1）の情報のみを取り出して BigWig 形式に変換した。

11.2 実データにおける相互相関係数の統合

理論値の導出過程では単一の染色体のみを持つゲノムを仮定していたが、実際の生物種の多くは複数の染色体を持つ。各染色体ごとに Strand cross-correlation を計算することができるが、各サンプルごとに単一のプロファイルを得るためにはこれらの相関係数を統合する必要がある。本研究では、フィッシャーの Z 変換（Fisher's r-to-z transformation）[72] を用いて変換した相関係数について染色体の長さで重み付けした平均 $\hat{z_\mu}$ を計算した。この平均値を逆変換することで相関係数の統合値を得た。

11.3 大規模解析を実現するためのツールの実装

NCC と MSCC は先行研究 [35] で導入され、Ramachandran らは計算を行うツールとして MaSC を実装した。しかしながら MaSC は入力ファイルの形式が一般的ではなく形式変換の必要があり、並列処理に対応しておらず必ずしも実用的であるとは言えなかった。また、DNA 断片の平均長推定に注力しており QC に関する機能は有していない。効率的な Strand cross-correlation の計算と QC の機能を盛り込むため、MaSC アルゴリズムを Python で実装した PyMaSC を開発した。PyMaSC は一部を Cython で実装し効率的な計算が可能で並列処理をサポートしている。

また入力形式としてリードは SAM/BAM 形式、Mappability については BigWig 形式をサポートすることで効率的なファイルサイズを維持したまま計算ができるようにした。PyMaSC はオープンソースで Python パッケージの 1 つとして PyPI（the Python Package Index）研究成果発表等 - 他から入手できる。

11.4 ChIP-seq データの事前処理方法

シミュレーションで生成した ChIP-seq データと実 ChIP-seq データはどちらもヒトリファレンスゲノム hg38 に BWA（0.7.17）[73] を用いてデフォルトパラメータでマッピングを行った。マップされたリードは SAMBLASTER（0.1.24）[74] を用いて PCR duplicate をマークした。マッピングクオリティスコアが低いリード（< 6）や alternate contig や unlocalised contig にマップされたリードは SAMtools（1.3.1）[28] を用いて取り除いた。実データにおける平均断片長の推定はまず PyMaSC で MSCC を計算し、各計算結果を目視で確認して断片長を推定した。また Strand cross-correlation の最大値は、ここで得た断片長に相当するシフト長での相関係数として得た。ピークコールは MACS2（2.1.0）[29] をデフォルトパラメータのまま用いた。ただし、リードの伸長に関しては PyMaSC による推定値を用いるため`--nomodel` オプションと`--extsize` オプションを指定し、推定した断片長を与えた。また ENCODE のデータについて、各実験毎に指定されているコントロールサンプルをコントロール用プロファイルとして共に指定した。さらに加えて、ブロードピークを形成するヒストン修飾については`--broad` オプションを付加した。ただしブロードピークのサンプルについてピーク数を得る際には、MACS2 の出力のうち *gappedPeak* ファイルを用いた。

12 シミュレーションによる導出結果の検証

ChIP-seq のリード分布を表現するモデルを作成したことで、モデルに基づいてシミュレーションデータを作成することが可能になった。導出した理論的な最小値・最大値について確認する第一段階として、シミュレーションデータを用いた検証を行う。

12.1 シミュレーションの手法と条件

まずモデルに基づいたシミュレーションデータの作成方法について述べる。このシミュレーションは与えられたリファレンスゲノム配列とパラメータ $n \cdot w \cdot d \cdot M \cdot \alpha \cdot R$ に対して、リファレンスの部分文字列（FASTQ 形式）もしくはアライメント済みデータ（SAM 形式）を出力する。実装は https://github.com/ronin-gw/chipseq-simdata-generator にて公開している。

1. リファレンスゲノムを FASTA 形式のファイルとして読み込み染色体毎の配列と長さを格納する。
2. 合計が与えられた数 n になるよう、各染色体に結合部位の数をランダムに割り当てる。ただし抽選確率は染色体の長さで重み付けしたものを用いる。
3. 各染色体ごとに、割り当てられた結合部位の数だけ長さ $2w + d$ 塩基の領域を互いに重複しないようランダムに選ぶ。

表 12: シミュレーションデータ生成に用いたパラメータの組み合わせ

パラメータ	説明	設定値
n	結合部位の総数	100, 1000, 10000, 100000
w	エンリッチ領域の長さ	100
d	エンリッチ領域間の距離（最小断片長）	100
R	リード長	50
M	総取得リード数	10×10^6, 50×10^6, 100×10^6
α	SN 混合パラメータ	1.0×10^{-4}, 2.5×10^{-4}, 5.0×10^{-4}, 1.0×10^{-3}, 2.5×10^{-3}, 5.0×10^{-3}, 1.0×10^{-2}, 2.5×10^{-2}, 5.0×10^{-2}, 1.0×10^{-1}, 2.5×10^{-1}, 5.0×10^{-1}

4. 結合部位の数と同様に、合計が M になるよう各染色体にリード数を割り当てる。割り当てられたリード数のうち、シグナルリードとノイズリードは $\alpha : (1-\alpha)$ で分配される。
5. シグナルリードとノイズリードを規定数生成する。生成するに先立ち、仮想的な DNA 断片となる領域を選択する。各リードについての洗濯方法は下記の通り。
   シグナルリード　各染色体毎に設定された結合部位からランダムに選択
   ノイズリード　染色体全域から $2w + d$ 塩基の領域をランダムに選択
6. 選択された領域から部分配列を生成する。
   （a）順鎖・逆鎖どちらのリードを生成するか半々の確率で選択する。
   （b）順鎖の場合は抽選された領域の先端から w 塩基の範囲からリードの開始地点をランダムに選ぶ。逆鎖の場合は領域の後端 w 塩基からランダムに選ぶ。
   （c）選択した開始地点から R 塩基をリードとして出力する。逆鎖の場合は逆相補配列を出力する。シーケンシングクオリティ（Phred スコア）は 40（‘I’）で固定する。
   （d）ただし、部分配列に ‘N’ が含まれていた場合は配列をマップ不能とみなし出力はせず再抽選もしない。

本研究では、w・d・R については固定とし、複数の n と w の組み合わせについて α を変化させながらシミュレーションデータを作成した。用いた条件を表 12 に示す。また、各パラメータの組み合わせごとに、FASTQ 形式・BAM 形式のデータを各々 5 つずつ生成した。

12.2　シミュレーションと理論値の比較

シミュレーションで生成したデータについて PyMaSC を用いて NCC および MSCC を計算し、得られた最小値・最大値と理論的に予測された値を比較した。理論値の算出には式（22）（23）（24）（25）での第 1 近似を用いた。まず図 10 にアライメント済みのシミュレーションデータから NCC を計算した場合の結果を示す。結果はどのような n と M の組み合わせでも最大値の理論値と実測値が $M \times 10^{-11}$ までは不飽和条件・飽和条件共に非常によく一致することを示している。一方、最大値がその値より小さくなる場合と最小値については、α に依存しない一定値を取って

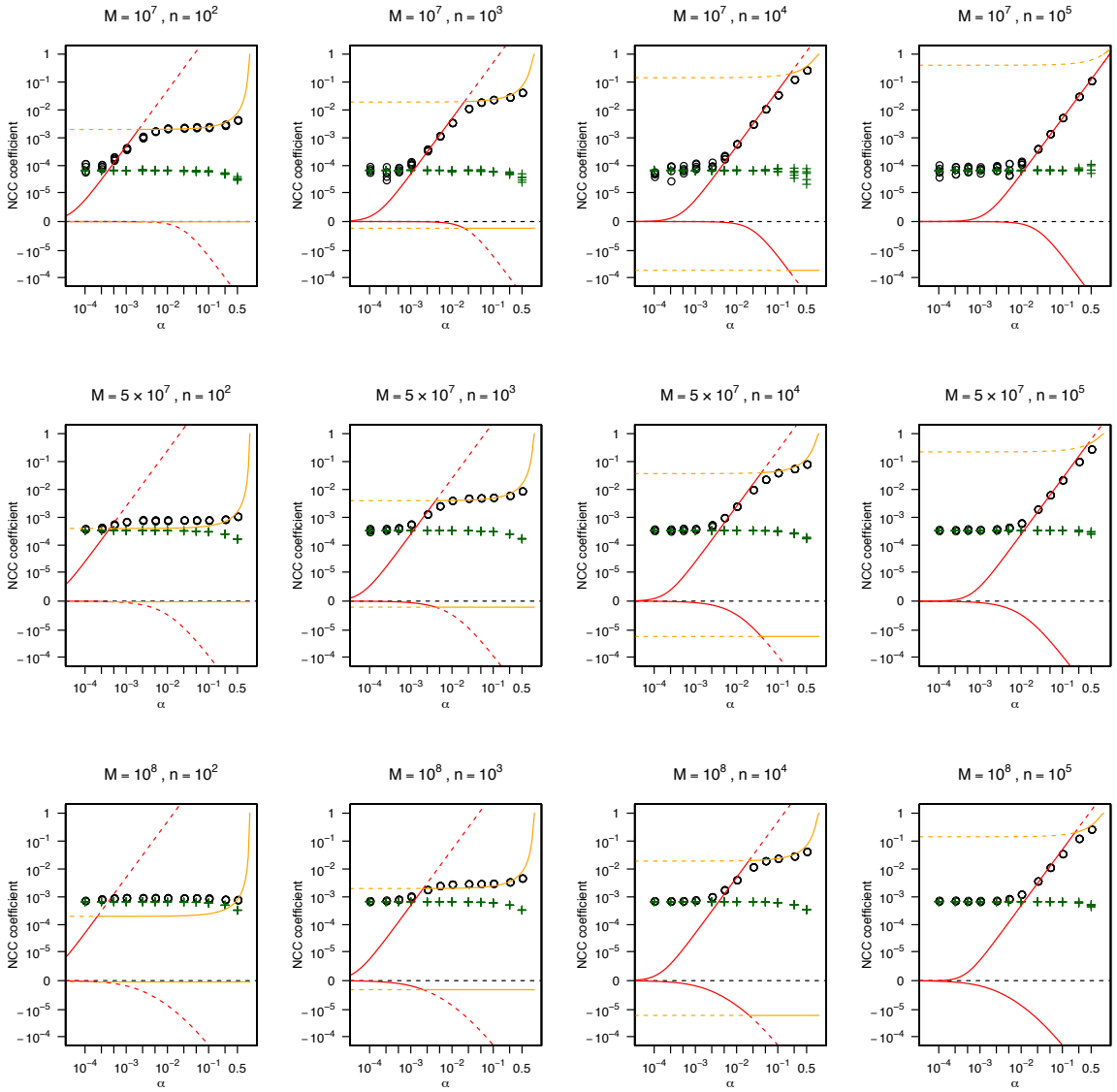


図 10: アライメント済みシミュレーションデータを用いた NCC の理論値と実測値の比較

各プロットは横軸に α、縦軸に NCC の相関係数を取った Log-modulus プロット $(y' = \mathrm{sign}(y)\log_{10}(1 + |y \times 10^5|))$。赤線は不飽和条件下、橙線は飽和条件下での理論値を示す。実線部は各条件が成り立つ範囲を表す。丸印は最大値の実測値、十字は最小値の実測値。

いる。これらは予測された理論値と一致せず、またこれらの状況下では NCC の最小値と最大値を区別できないことを示唆している。このような理論値との乖離が生じる原因としては、リファレンスゲノムの Mappability によるバイアスが考えられる。ヒトリファレンスゲノムの最新バージョンである hg38 は 2013 年に公開されたが、未だに多くの未解読（難読）領域を含んでいる [45]。図 11 は hg38 の各染色体ごとのギャップの位置を可視化したものである。これらの領域にはリードをマップすることができないため、マップされるリードの位置には偏りが生じる。またこれらの領域の長さはシフト長に比べると、テロメア・セントロメア領域のように非常に長いものもい

くつか存在するため、これらの影響をキャンセルすることができず、最小値が 0 近辺からずれた一定値を持っていると考えられる。

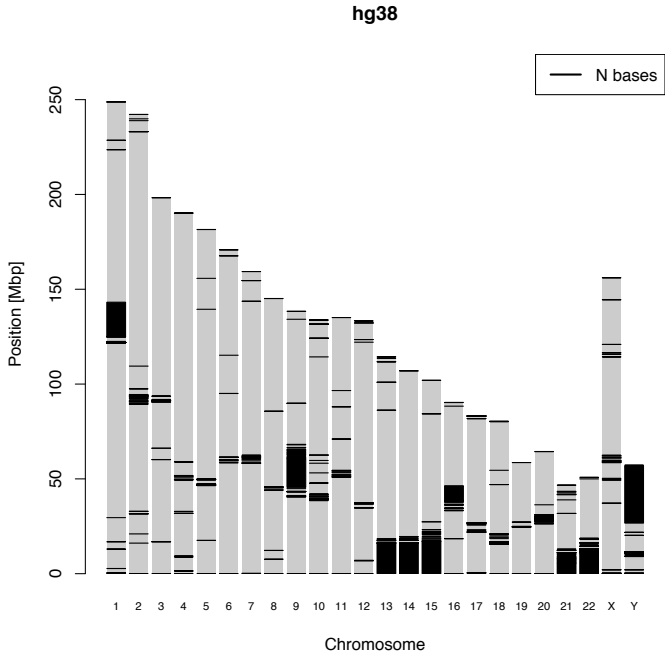


図 11: ヒトリファレンスゲノム hg38 における未読領域の分布

この問題については MSCC が解決策になる。図 10 と同様のシミュレーションデータに対して MSCC を用いて最小値・最大値を求めた場合のプロットを図 12 に示す。MSCC では最小値が 0 前後に分布しており、Mappability のバイアスを排除して Strand cross-correlation の計算を行えていることを示している。これに伴い、NCC と比較してより低い α の値でも最大値を最小値から区別することが可能である。このことから、MSCC は SN 比が非常に低いサンプルにおいても正確に Strand cross-correlation としての指標を提供できる可能性があることが判った。

これらのシミュレーションデータは部分配列の生成時にゲノム座標の情報も共に出力することで、アライメント済みのデータとして出力したものを用いてきた。そこで次の検証として、マッピングの処理が NCC ないし MSCC の計算にどのような影響を与えるかを検証するため、部分配列のみの FASTQ 形式でシミュレーションデータを出力し、マッピングによりリファレンスゲノムに貼り付け直す処理を経てから NCC・MSCC の計算を行った。まず NCC の結果を見ると（図 13)、僅かにではあるが図 10 と比較して最小値の値が上昇していることがわかる。一方、MSCC の場合は特に大きな変化を確認することができなかった（図 14)。これはゲノム中の長大な未解読領域に起因するバイアスに加えて、Mappability が低くユニークにマッピングできない領域がリードの分布をバイアスしているために生じると考えられる。これらの結果から、ゲノム構造に起因する大域的な Mappability の偏りだけでなく、配列そのものに由来する Mappability にも Strand cross-correlation は影響を受け、MSCC はこれらのバイアスをキャンセルしてプロファイルを生成できることが確認された。従って、Mappability を考慮した補正は従来提案された DNA 断片の平均長を正確に推定する用途のみならず、Strand cross-correlation そのものの正確な計算

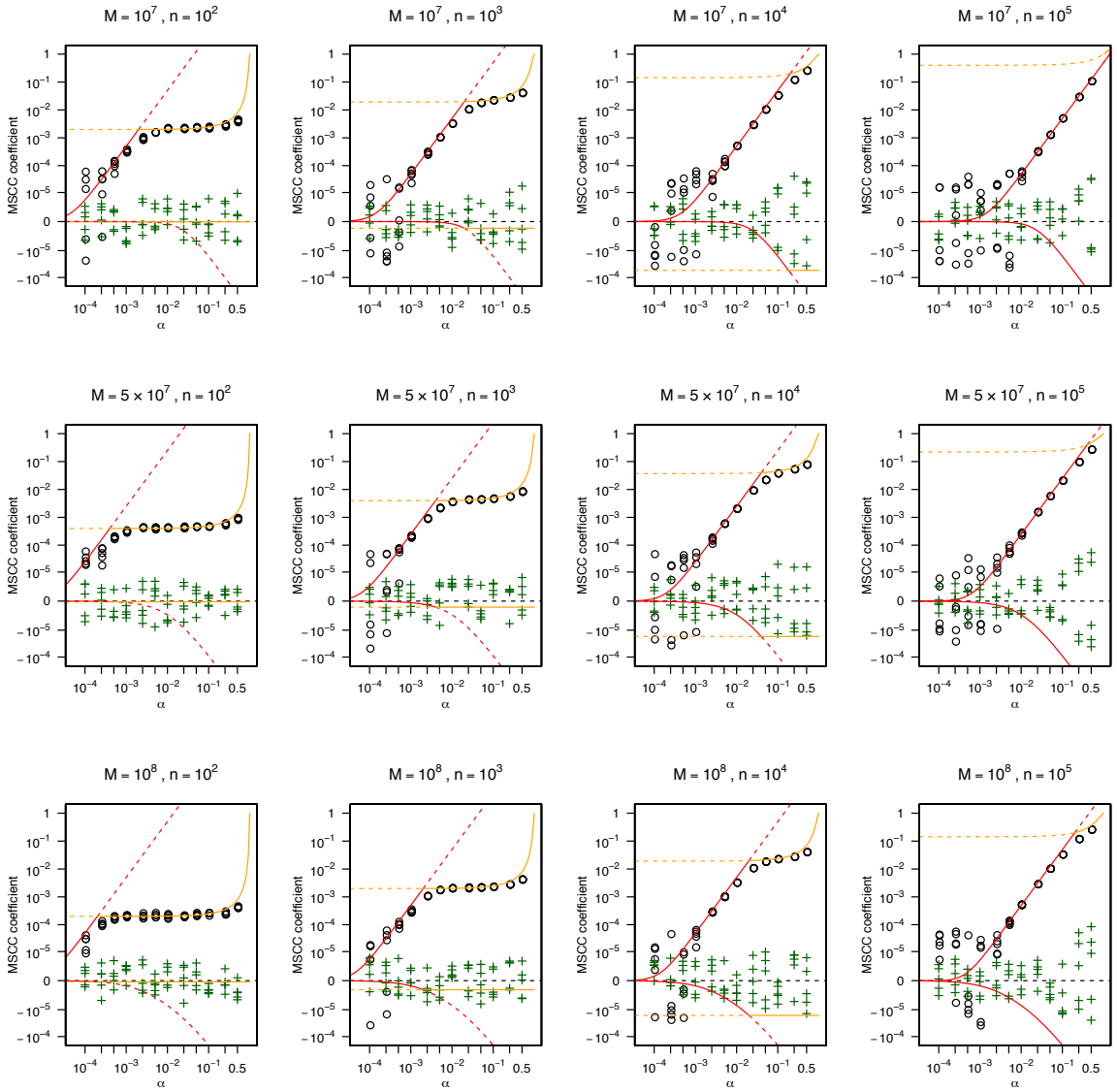


図 12: アライメント済みシミュレーションデータを用いた MSCC の理論値と実測値の比較

についても優位性があるといえる。

13 実データによる導出結果の検証

前節ではシミュレーションデータにおいて期待された理論値がよく当てはまることが確かめられたが、QC 手法の確立を最終目標とするならば実データでの実証がより重要であり、必要不可欠である。これまでの結果では、NCC および MSCC の最大値は SN パラメータ α・結合部位の総数 n・エンリッチ領域の長さ w そして総リード数 M（あるいは M_{u}）の関数として表されることが示された。これらのパラメータは、ピークコール解析の結果を用いることで推定することができる。そこで、提案したモデルと Strand cross-correlation に関する関係式が実データでも成り立つかを、Strand cross-correlation による実測値と、それとは独立したパラメータ推定を元に計算

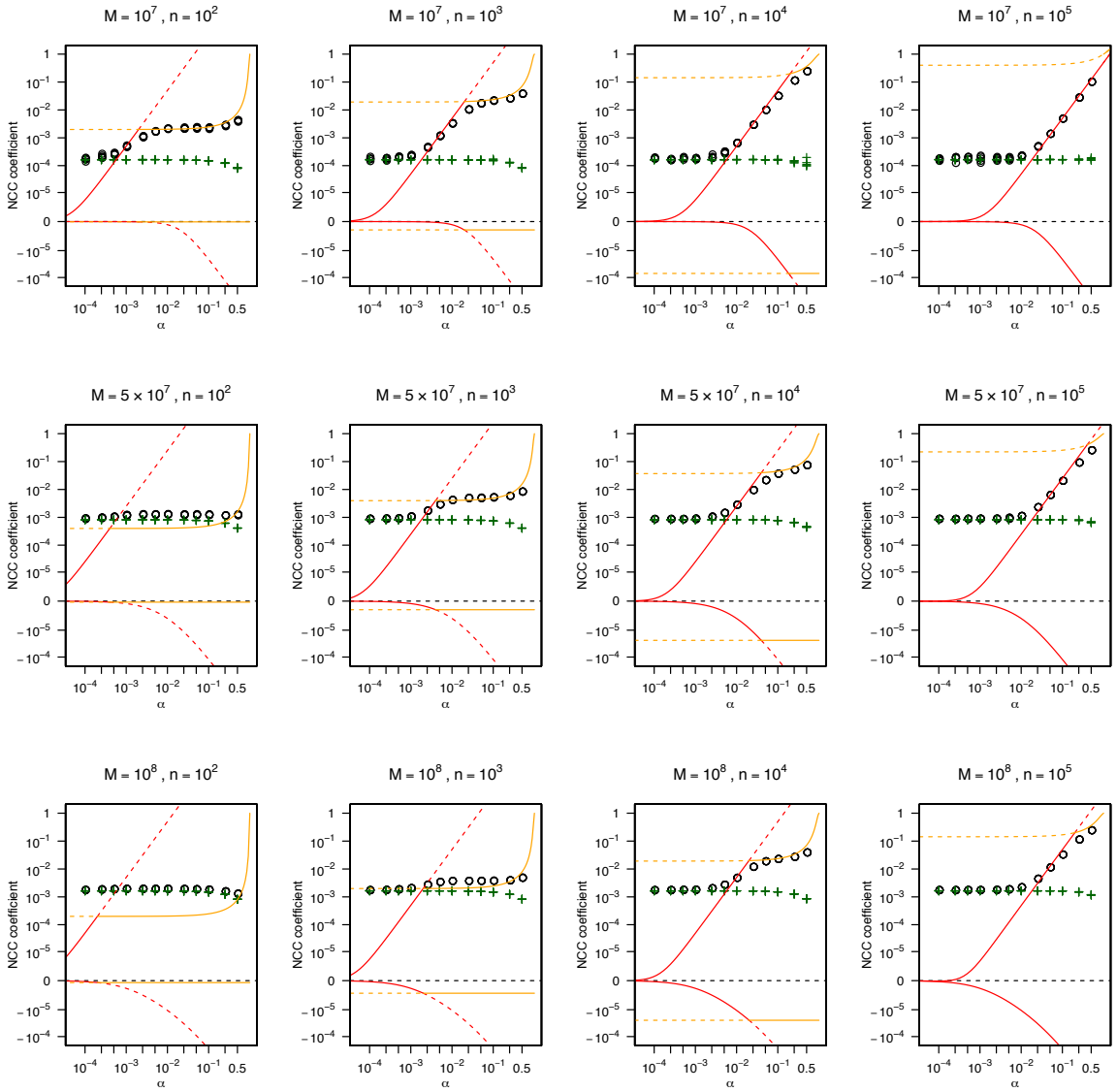


図 13: マッピングを伴うシミュレーションデータを用いた NCC の理論値と実測値の比較

した理論値とを比較することで検証を行う。

13.1 ピークコールを用いたパラメータ推定

本研究では、すべての ChIP-seq データに関して生データから単一の解析パイプラインを用いて解析を行っており、PCR duplicate やマッピングクオリティの低いリードを排除するステップを組み込んでいる。従って、M と M_{u} はクオリティフィルタリング前後の総リード数として得ることができる。ピークコールを行うと、n の推定値として得られたピークの総数を用いることができる。w と α の推定については個別に詳しく述べる。

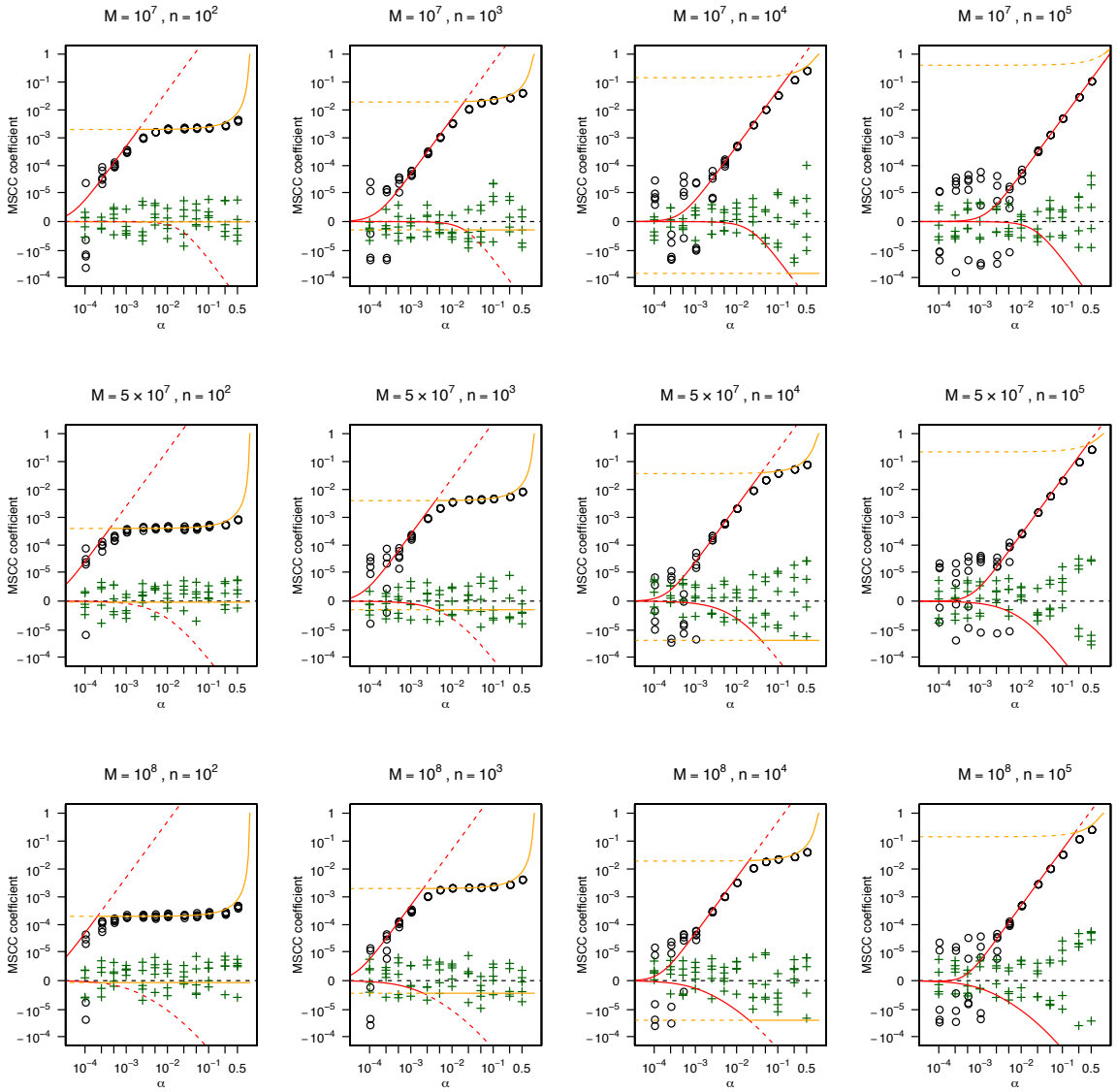


図 14: マッピングを伴うシミュレーションデータを用いた MSCC の理論値と実測値の比較

13.1.1 w の推定

MACS2 は DNA 断片の平均長を推定するために上位 1,000 箇所のピーク周辺についてリードの密度を計算する（図 15）。これを利用して、w の推定値としてこのピークの半値全幅（FWHM: Full Width at Half Maximum）を用いる。具体的には、順鎖・逆鎖の分布それぞれで FWHM を計算しその平均値を推定値として用いた。図 15 で示したプロファイルに対する実行結果を図 16 に示す。実装として R のソースコードをコード 6 に掲載する。コード中の `x`, `p`, `m` は MACS2 の出力のうち **_model.r* ファイルで定義されている。

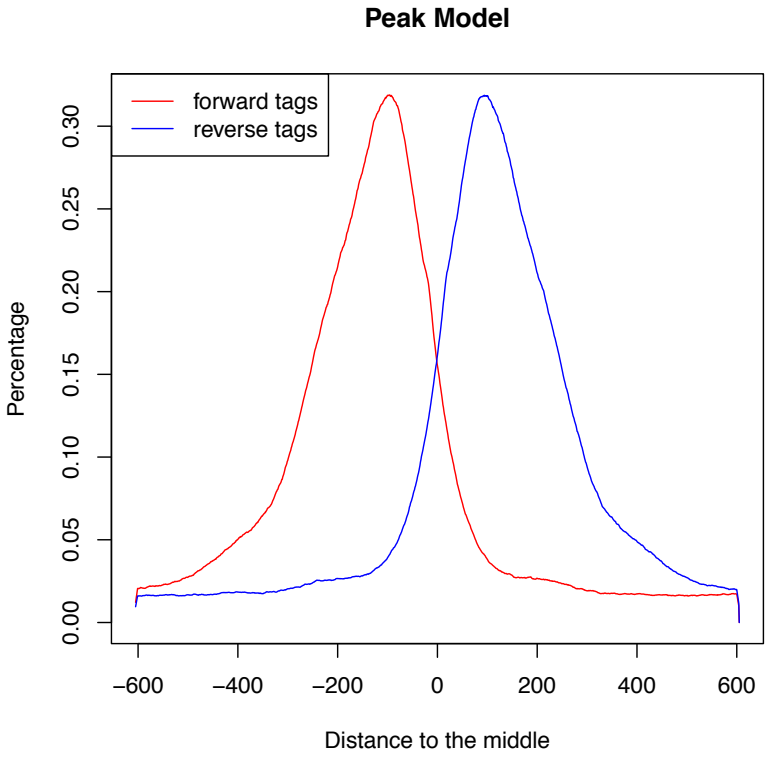


図 15: MACS2 により生成されたピーク周辺のリード密度分布

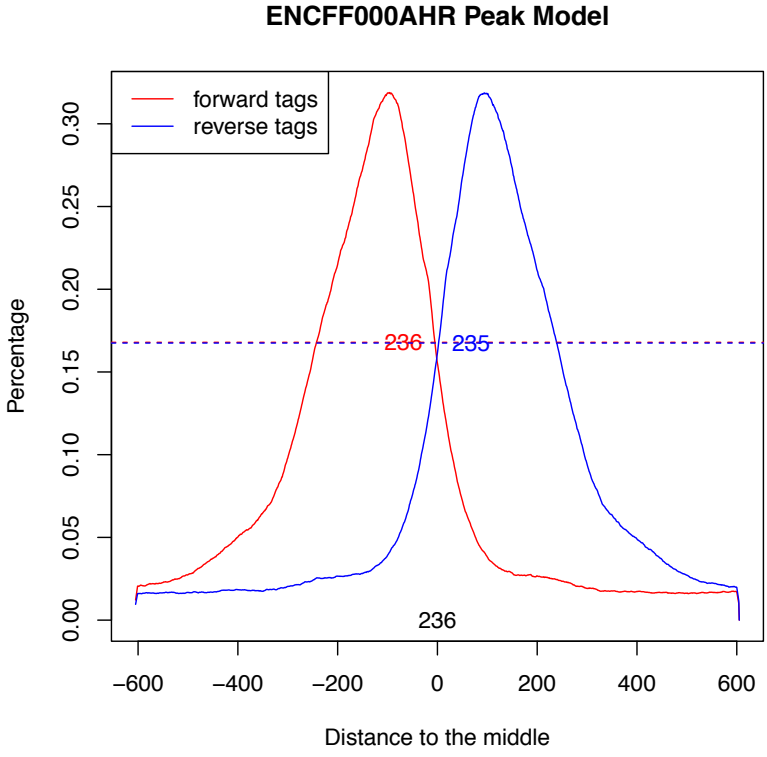


図 16: MACS2 の解析結果を用いた w の推定例

コード 6: R による w 推定アルゴリズムの実装例

```
getwidth <- function(y) {
    ly <- length(y)
    ymin <- min(median(y[1:50]), median(y[(ly-49):ly]))
    ymax <- max(y)
    xpeak <- round(mean(which(y == ymax)))

    thresh <- (ymax - ymin) * 0.5 + ymin

    rwidth <- rle(y[xpeak:ly] >= thresh)$lengths[1] - 1
    lwidth <- rle(rev(y[1:xpeak] >= thresh))$lengths[1] - 1

    return(c(thresh, x[xpeak - rwidth], x[xpeak + lwidth]))
}

pdf(commandArgs(trailingOnly=T)[1],height=6,width=6)
plot(x,p,type='l',col=c('red'),main=main,xlab='Distance to the middle',
    ylab='Percentage')
lines(x,m,col=c('blue'))
legend('topleft',c('forward tags','reverse tags'),lty=c(1,1,1),col=c('
    red','blue'))

r <- getwidth(p)
plen <- r[3] - r[2]
abline(h=r[1], col="red", lty=2)
text(mean(r[2:3]), r[1], r[3] - r[2], col="red")
r <- getwidth(m)
mlen <- r[3] - r[2]
abline(h=r[1], col="blue", lty=2)
text(mean(r[2:3]), r[1], r[3] - r[2], col="blue")

elen <- round(mean(c(plen, mlen)))
text(0, 0, elen)
cat(elen)
cat('\n')
```

13.1.2 α の推定

α の推定には FRiP を用いる。FRiP はピーク中に含まれるリード数の割合であったから、本研究で提案したモデルにおいてはエンリッチ領域の内外でのリード数の比に等しい。

$$\mathrm{FRiP} = \frac{M\alpha + \frac{n(2w+d)}{G}M(1-\alpha)}{M} = \alpha + \frac{n(2w+d)}{G}(1-\alpha) \tag{49}$$

したがって α の推定値は次の式で表せられる。$n(2w+d) \ll G$ とすれば、α は FRiP で代用できる。

$$\begin{aligned}\hat{\alpha} &= \frac{\mathrm{FRiP} - \frac{\hat{n}(2\hat{w}+\hat{d})}{G}}{1 - \frac{\hat{n}(2\hat{w}+\hat{d})}{G}} \\ &\approx \mathrm{FRiP}\end{aligned} \tag{50}$$

ここでは α の推定値として FRiP を用いる。

13.2 実データを用いたテストデータセットの作成

実データとして、ENCODE portal[19] に登録されたヒト A549 細胞の Single-end ChIP-seq データを用いてデータセットを作成した。データセットには ENCODE プロジェクトと Genomics of Gene Regulation（GGR）プロジェクト [22] で取得された 790 件の ChIP サンプルと対応する 152 件のコントロールサンプルで構成される。ChIP ターゲットの内訳を表 13 に示す。ただし Broad ヒストン修飾のうち H3K9me3 については、リピート領域にエンリッチしやすく megabase スケールのエンリッチドメインを構成すること [75]、用いられた抗体の親和性が比較的低い恐れがあり [76]、特に実験・解析の難易度が高いことが予想されたため以下の解析では別項目として扱った。

表 13: テストデータに用いた ChIP-seq データの ChIP ターゲット内訳

種類	種類数	サンプル数	内訳
転写因子	59	558	（省略）
ヒストン修飾（Narrow）	5	107	H2AFZ, H3K27ac, H3K4me2, H3K4me3, H3K9ac
ヒストン修飾（Broad）	6	125	H3K27me3, H3K36me3, H3K4me1, H3K79me2, H3K9me3, H4K20me1
計	70	790	

13.3 パラメータの推定結果

図 17 に各パラメータの分布を示す。一般的に、結合部位の多い ChIP ターゲットほどより多量のリードをシーケンシングしなければ信頼できるピークコールの結果を得ることができない [33]。ENCODE および modENCODE コンソーシアムでは、転写因子の場合で 10M 本、ヒストン修飾では 50M 本以上のリード数を取得することを目指している [32]。本データセットでもその傾向が見られ、ヒストン修飾の方がより多くのリードを取得していた。ほとんどのサンプルでは、リード数がヒトゲノムサイズ（約 3Gbp）より十分少ないことが確かめられたが、リード数の多いサンプルでは 100M 本を超えており、このようなサンプルでは $M \ll G$ を仮定した近似については誤差が生じる可能性がある。

取得リード数に関連して、コールされたピーク数 $\hat{n}$ と SN パラメータ $\hat{\alpha}$ についても転写因子とヒストン修飾の間に明確な差異が見られた。取得リード数が多いほど検出しうるピーク数や FRiP

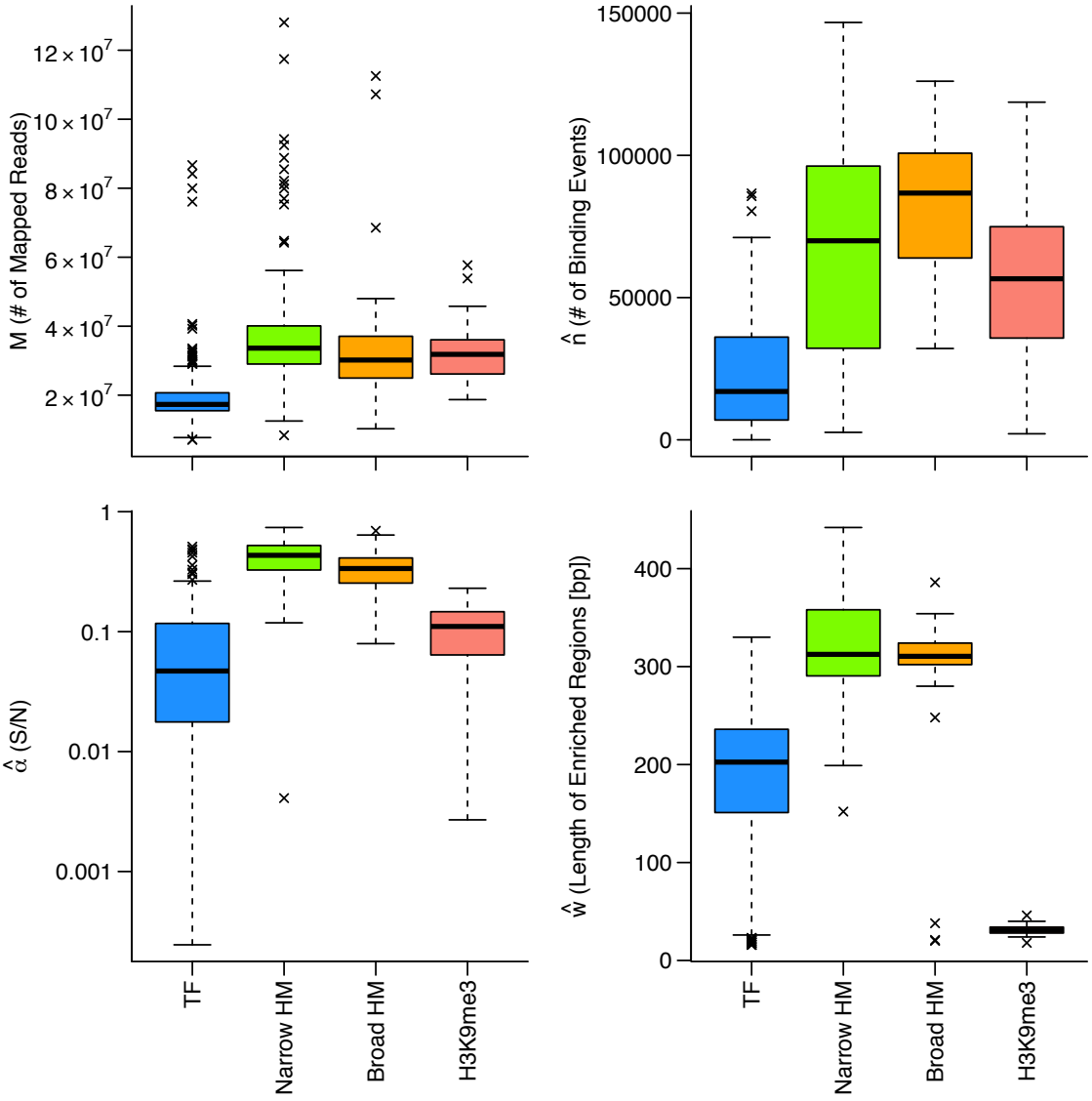


図 17: ENCODE テストデータにおける各パラメータの推定値の分布

の値が増加することは知られているため [32]、これらの関連性は妥当であると考えられる。

w の推定については H3K9me3 のみが極端に低い値を示しているが、これは MACS2 のプロファイルを用いた推定が H3K9me3 では失敗したためである（図 18）。w の長さとしては順鎖・逆鎖共に約 300bp ほどのエンリッチが見られるが、FWHM を用いた場合、中心部付近の鋭いピークしか検出できなかった。したがって、H3K9me3 については w をかなり過小推定している。

13.4 不飽和条件・飽和条件を満たす実データの確認

パラメータの組み合わせに応じて不飽和条件と飽和条件を定義したが、実際のデータはどちらの条件に従うだろうか。これを検証するため、推定したパラメータを用いて $\frac{M\alpha}{2nw}$ が 1 を下回るか上回るか（式（16）を参照）を検討した（図 19）。790 サンプルのうち 31 件のみが飽和条件下に

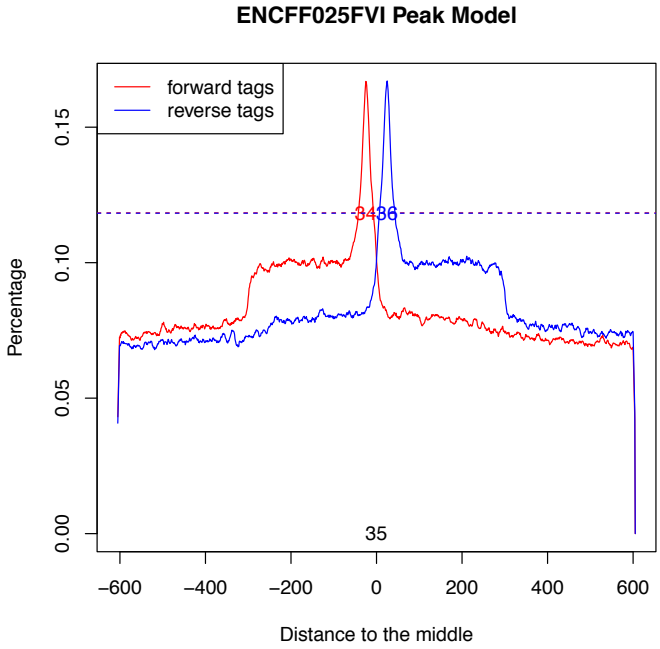


図 18: H3K9me3 サンプルでの w 推定を失敗した例

あるサンプルであった。また、飽和条件の成立と各パラメータの関係を確認すると、これらのサンプルのほとんどは非常に小さい $\hat{n}$ や w の推定の失敗に強く関係していることが判った（図 20）。これはクロマチン免疫沈降のエンリッチが十分でなかったり、推定手法に起因するものであると考えられるため、通常の ChIP-seq 実験においてはその大半が不飽和条件を満たしていると考えられる。

13.5 実データにおける NCC の最大値の理論値と実測値の比較

推定したパラメータを用いて式（23）（25）に基づく理論値を計算した結果と PyMaSC による NCC の最大値の実測値を比較した結果を図 21 に示す。転写因子については非常によく一致していた。また他の種類のターゲットについては中心から離れているもののよく相関しており、これらに見られる誤差は本研究で用いているモデルの表現力の不足よりは、ピークコールの手法やパラメータ推定の手法に起因する誤差であると考えられる。

13.6 NCC と MSCC の比較

10.2.6 において NCC と MSCC の相関係数はほとんど同じことが予想された。これを実データを用いて確かめる

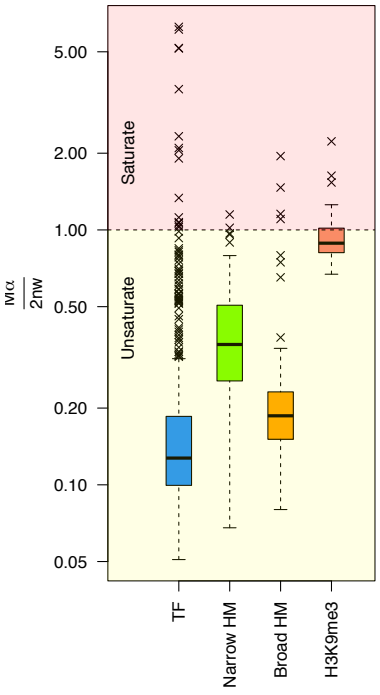


図 19: 不飽和条件・飽和条件の実データによる検討

ため PyMaSC で計算した NCC と MSCC を比較した（図 22）。結果、NCC と MSCC の最大値はよく一致することが確かめられた。

14 考察

本研究で提案したモデルが NCC と MSCC の理論値の計算に役立てられることを実証した。むろんモデルとしてはシンプルであり改善の余地がいくつか考えられる。例えば本モデルでは w や d を定数として扱っているが、本来これらはぶれが生じうるものであり何らかの分布を仮定して用いる方がモデルの表現力が向上するはずである。また、リードが丁度半々ずつ順鎖・逆鎖にマップされるという仮定も実際にはストランドバイアスが存在する事が知られている [77]。各結合部位が十分離れているという仮定も転写因子については成り立つ場合が多いが、ヒストン修飾ではピー

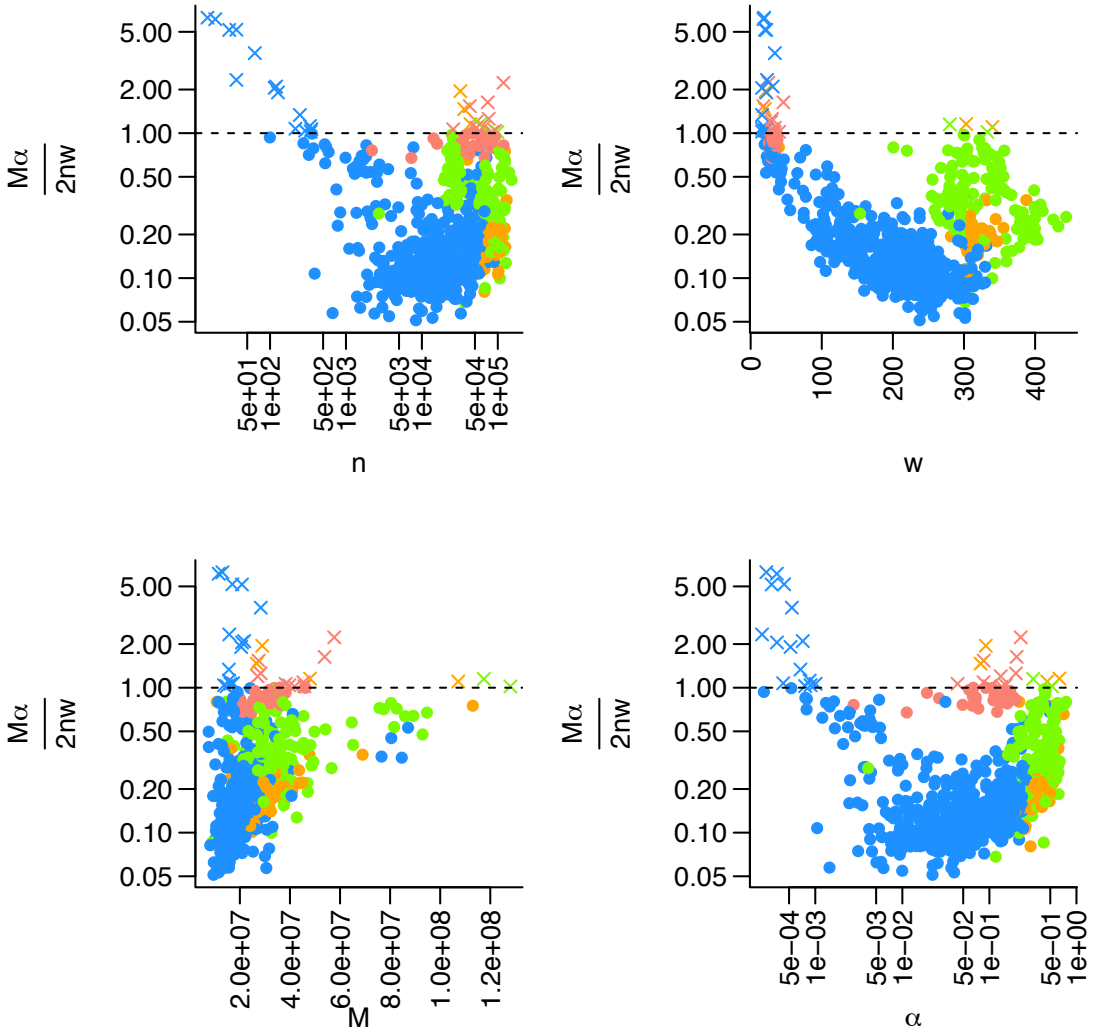


図 20: 不飽和条件・飽和条件の境界値と各パラメータの関係

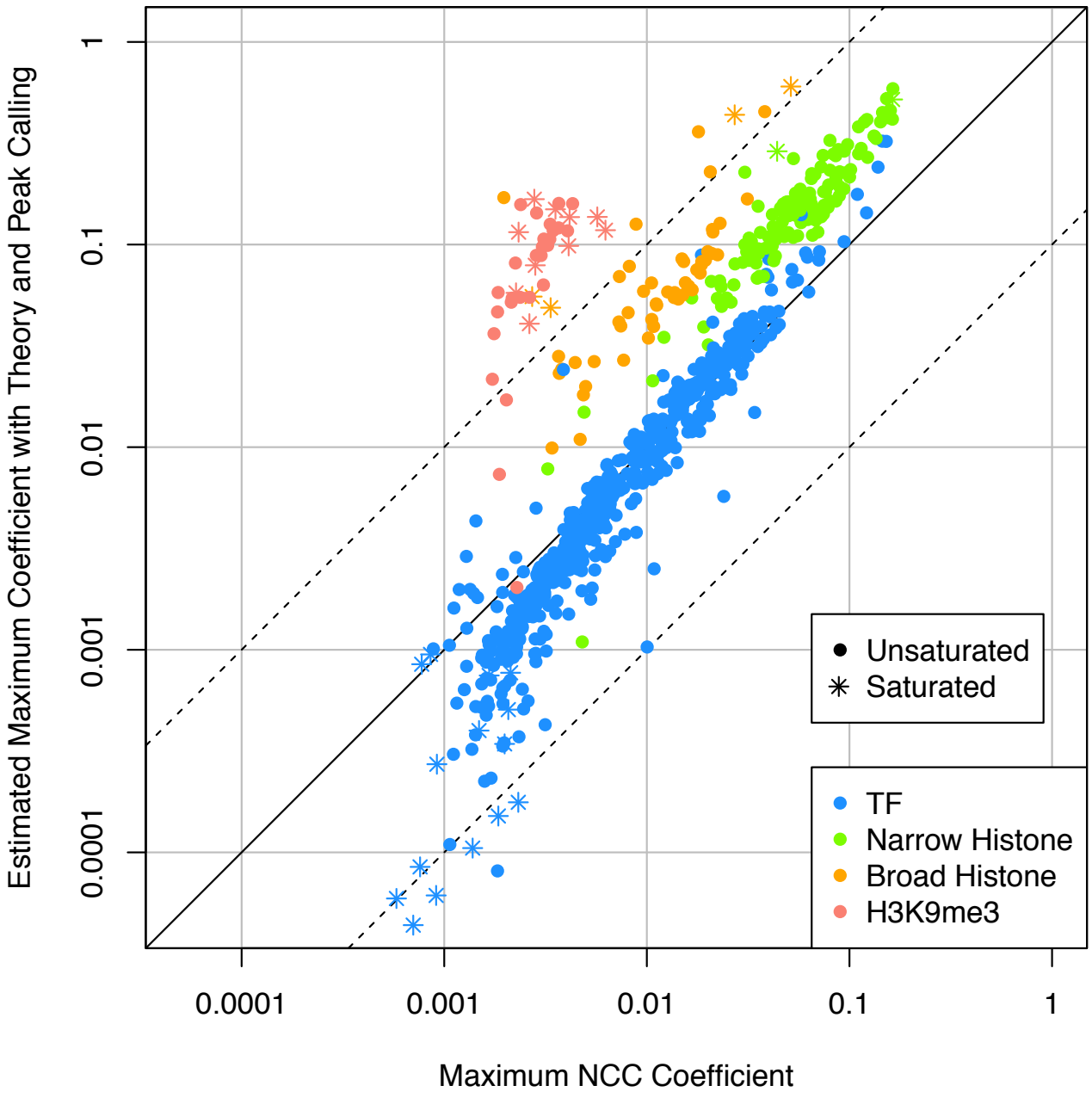


図 21: NCC の実測した最大値と推定値の比較

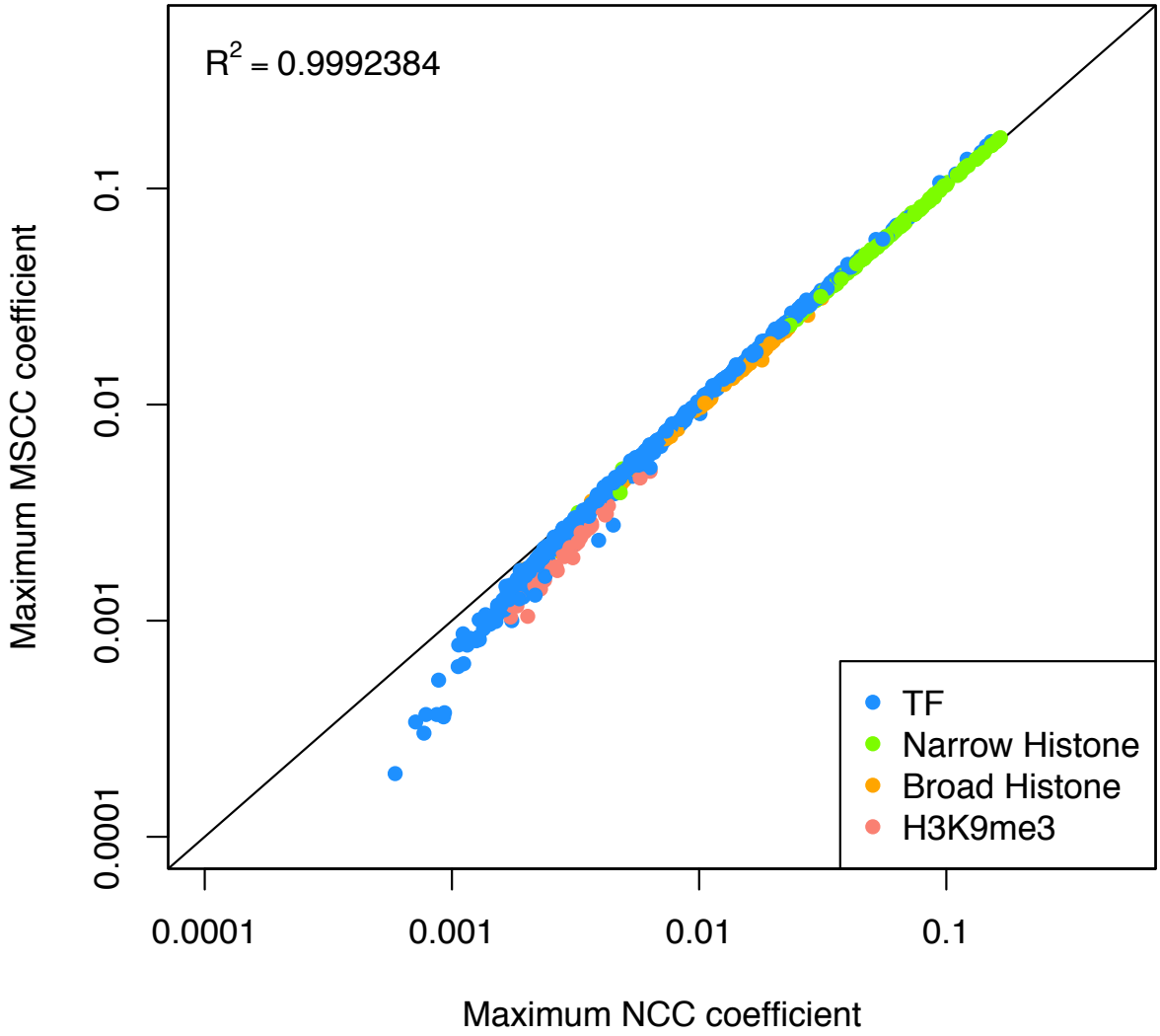


図 22: NCC の最大値と MSCC の最大値の比較

クがタンデムになる場合もあり、NCC の理論値と実測値の比較において、ヒストン修飾の場合に特に乖離が大きくなっている原因の 1 つかもしれない。このような改善を盛り込んだより詳細なモデリングとシミュレーションが Strand corss-corerlation の性質をより詳細に解明する可能性は否定できないが、理論値と実測値の比較において高い相関を示したことから、本研究で用いたモデルでも Strand cross-correlation の基本的な性質は捉えられてると期待している。

14.1 NCC および MSCC の活用方法と限界について

一般的な ChIP-seq 実験は不飽和条件を満たしていると考えられるため、NCC と MSCC の最大値は式（23）で特徴付けられると結論できる。M は既知であり、また w は Strand cross-

correlation の分散からも推定が可能だと考えられることから、相関係数の最大値を求めることで n と α の関係式をピークコール前に得られることが示された。別の視点で言えば、Strand cross-correlation を用いた場合、n と α をピークコール前に各々独立に推定することは困難かもしれないことを暗示している。

とはいえ、この関係式はどのようなサンプル間で最大値が比較可能になるかを明らかにした。例えば、ある 1 実験で複数のレプリケートを取得した場合、n や w はレプリケート間で同じ値を取ることが期待されるため、これらのサンプル間で NCC あるいは MSCC の最大値を比較することで α の大小を評価することが可能である。あるいは data-driven なアプローチとして大量のデータに対して最大値を計算し、生物種あるいはターゲットごとの基準値を策定するという方針も考えられる。

14.2 既存の Strand cross-correlation を用いた QC 指標について

これまでに提案された Strand cross-correlation を用いた QC 指標としては NSC（Normalized Strand Coefficient）が提案されている [32,66]。これは相互相関係数の最大値を最小値で割り正規化した値であった。しかしながら、本研究で得られた結果に基づくと、最小値はほとんど 0 に近い負の値であり、仮にシフト長を伸ばして値が 0 に近づくとしても最小値はシフト長に依存することになり、正規化に用いる値としてどのような効果があるか疑問が残る。従って最小値による正規化を行うよりも、最大値の値そのものか、あるいは最大値をリード数で正規化した値の方がより情報に富んでいると考えられる。

$$\frac{1}{M}\mathrm{NCC}(f,g)(d+w) = \frac{\alpha^2}{2nw} \tag{51}$$

ただし、リードの分布は様々な技術的・生物学的なバイアスを受けることが知られており [61]、シフト長を十分伸ばしても Strand cross-correlation のバックグラウンドレベルは理想的な一様分布よりも高くなる場合があることが知られている [66]。そのため、NSC で用いられている最小値が、総取得リード数やバックグラウンドレベルの一様性を表す代表値として機能しており、NSC という指標にこれらが反映されているという可能性は否定できない。また、本研究で用いた相互相関関数は NCC と MSCC であり、実際に NSC がどの程度指標として機能しているかを理論的な側面から正確に評価するためには、ピアソンの相関係数や Jaccard 係数を相互相関関数に基づいて理論的な考察を行う必要がある。

15 小括

Strand cross-correlation を用いた QC 指標は FRiP と比較してピークコールの手法に依存せず、よりロバストな QC 指標として注目されている。しかしその理論的な裏付けはなされておらず、指標の作成や基準値の設定は経験に基づいて行われてきた。結果として、これらの指標は QC 指標として具体的に何を反映しているかが不明瞭であり、適応できる範囲や限界についてもよく判っていなかった。本研究では、ChIP-seq のリード分布をモデル化し相互相関関数として NCC を用いた時の相関係数の最小値と最大値の理論値を導出した。その結果、NCC の最大値は結合部位の総数・エンリッチ領域の長さ・総リード数そして SN 混合パラメータの関数として表され

ることが示された。シミュレーションデータと実データによる検証で、この関係がよく成り立つことが確かめられた。本研究では、NCC の理論値の導出を通してその最大値が何を反映してどのような状況下で比較可能になるか、また QC 指標としての限界を明確にした。これらの成果は ChIP-seq におけるより優れた QC 指標の設定に多大な貢献をするであろう。

第 IV 部
Strand cross-correlation を用いた新規品質評価手法の提案と検証

16 導入

第 III 部において Strand cross-correlation の理論的特性を評価した結果、NCC の最大値は式（23）に従うことが明らかになった。ここで M は既知であるから、NCC あるいは MSCC の最大値を計算することで式（51）の左辺が得られる。また、相互相関係数の分散から w を推定可能なはずである。w も既知と仮定すれば、最終的に n と α の関係式を得ることができる。

$$\frac{2w}{M}\mathrm{NCC}(f,g)(d+w)=\frac{\alpha^2}{n} \tag{52}$$

ここで n と α の関係について前部のテストデータに対して推定した値をプロットしたものを図 23 に示す。本来、結合部位の総数はターゲットとサンプルに依存し、α は主に実験手法に依存することが期待されるのでこの 2 つのパラメータは理想的には独立である。しかし、ここで明確な相関関係が見られるのは n が推定値であるためである。これは α が高いほど検出できるピークの数が増加することを示しており、$\hat{n}$ を実際に検出しうるピーク数と解釈するならば妥当な関係であると言える。もし、α/n が ChIP ターゲットに依らず一定値を取ることが期待できるならば、α^2/n の大小は α の大小と一致するはずである。そこで、式（52）で得られる α^2/n を VSN（Virtual Signal-to-Noise ratio）と名付け、ピークコール前に計算可能な疑似 FRiP として提案する。

第 IV 部では VSN が実用的な指標であるかを検証するため、まず w が Strand cross-correlation から推定できるか確かめ、次に既存の Strand cross-correlation を用いた指標や FRiP との比較を行う。

17 手法

17.1 テストデータセットの作成

用いるテストデータは 13.2 節で用いた ENCODE ヒト A549 細胞の ChIP-seq データを引き続き用いる。データの前処理方法は同じく 11.4 節に準じる。

17.2 Strand Cross-Correlation と QC 指標の計算

比較対象として phantompeakqualtools（PPQT, Last updated: Feb 12, 2012）[65] および SSP（v1.1.2）[59] を用いて各々の Strand cross-correlation を計算し、NSC と RSC を求めた。ただし、これらの指標で必要となる Strand corss-correlation の最大値は PyMaSC で推定した DNA 断片の平均長を用いて得た。

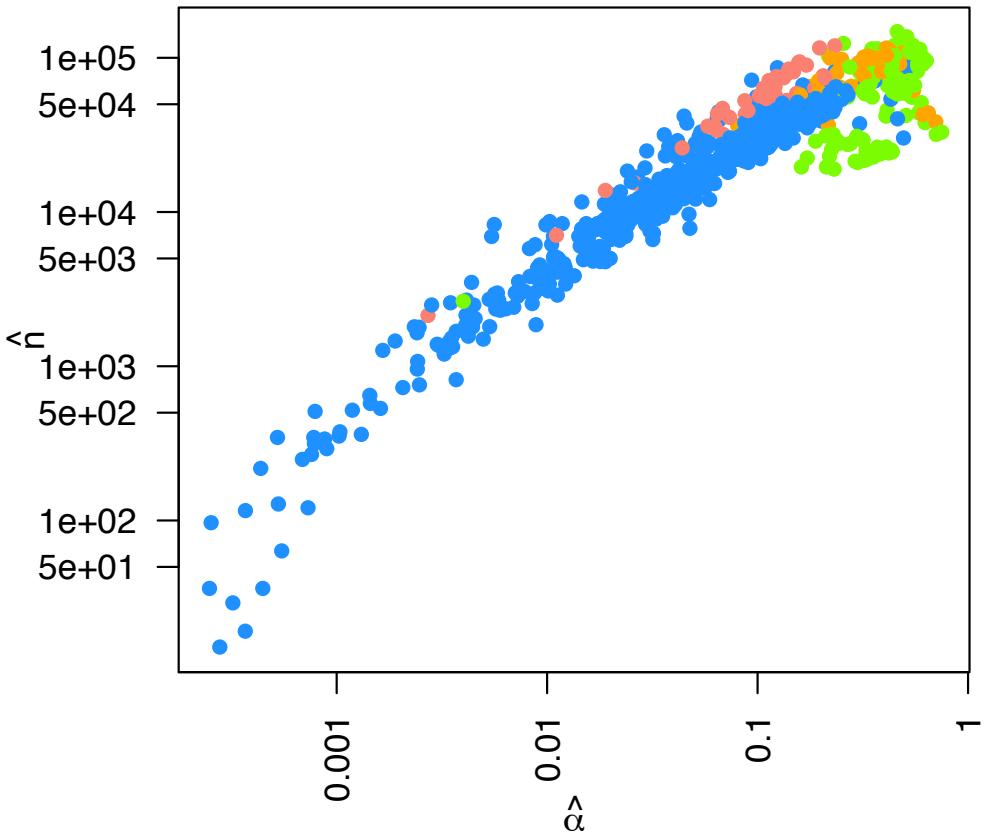


図 23: ENCODE データセットで推定した α と n の関係

18 MSCC を用いた w の推定

VSN の算出に先立ち、Strand cross-correlation を用いたエンリッチ領域の長さ w の推定について検討する。ここでは MSCC の半値全幅の 2/3 を w の推定値として用いた。MACS2 のプロファイルはピーク周辺のリード分布そのものである一方、MSCC を用いた場合はその相互相関関数を計算するため、MSCC の全値全幅は 2 倍、半値全幅は 1.5 倍になることが予想されたためである。推定した w と MACS2 のプロファイルを用いて推定した w の比較を図 24 に示す。特に転写因子のサンプルについては MACS2 での推定結果とよく一致していることが判る。また MACS2 のプロファイルで推定を失敗していた H3K9me3 などのサンプルについても PyMaSC では推定を行うことができた。

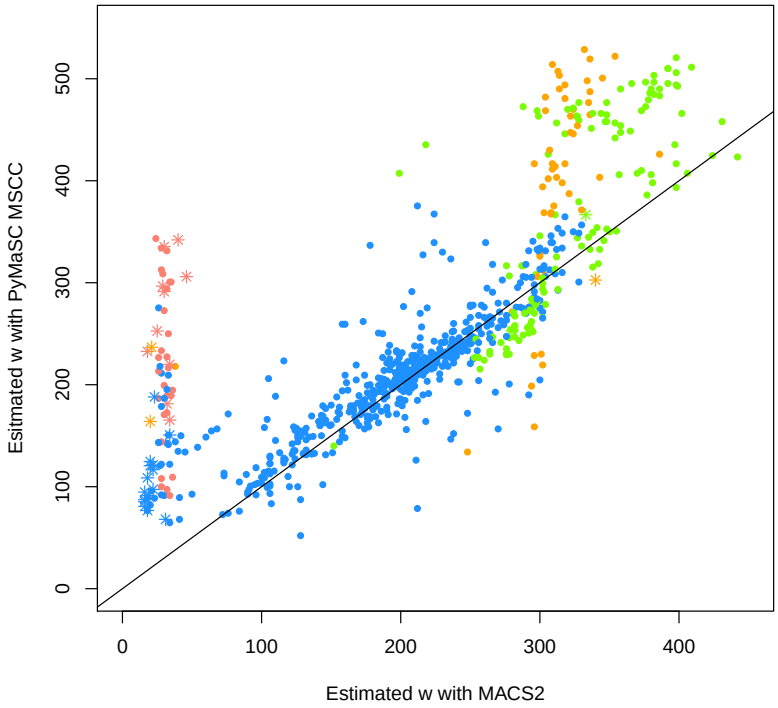


図 24: MACS2 と PyMaSC MSCC による w の推定値の比較

x 軸が MACS2、y 軸が PyMaSC で計算した MSCC に基づく w の推定値である。実線は $y = x$ を示す。

19 Strand Cross-Correlation を用いた QC 指標と FRiP の比較

ここでは既存の QC 指標および VSN と FRiP の間に整合性があるか検証する。図 25 は各指標と FRiP の比較である。PPQT および SSP の NSC は正の相関が見られるものの FRiP が低い場合の分解能が悪くまた ChIP ターゲットの種類によって傾向が異なっている。RSC では関係が線形になり NSC と比較して FRiP の値に対する分解能は良くなっているが、分散が大きくまた依然として ChIP ターゲットによる差異も観察される。VSN の場合は FRiP と正の線形の相関を保っており、Broad なヒストン修飾では差異が目立つものの全体としてはよく FRiP と対応していることが観察される。

これらの関係を定量的評価するため、FRiP と各指標のスピアマンの順位相関係数を計算した (表 14)。ただし全体の相関係数については、各ターゲットの種類がデータセット内で均一ではないため、各種類から 38 件をサンプリングした相関係数の計算を 100 回繰り返した平均値を用いた。この相関係数の分布は図 26 のようになった。VSN は比較した 5 つの指標の中で、転写因子については最も高い FRiP との相関係数を示した。また、ターゲットの種類が混在している場合でも、他の指標より高い相関係数を安定して示した。一方、ヒストン修飾については SSP NSC

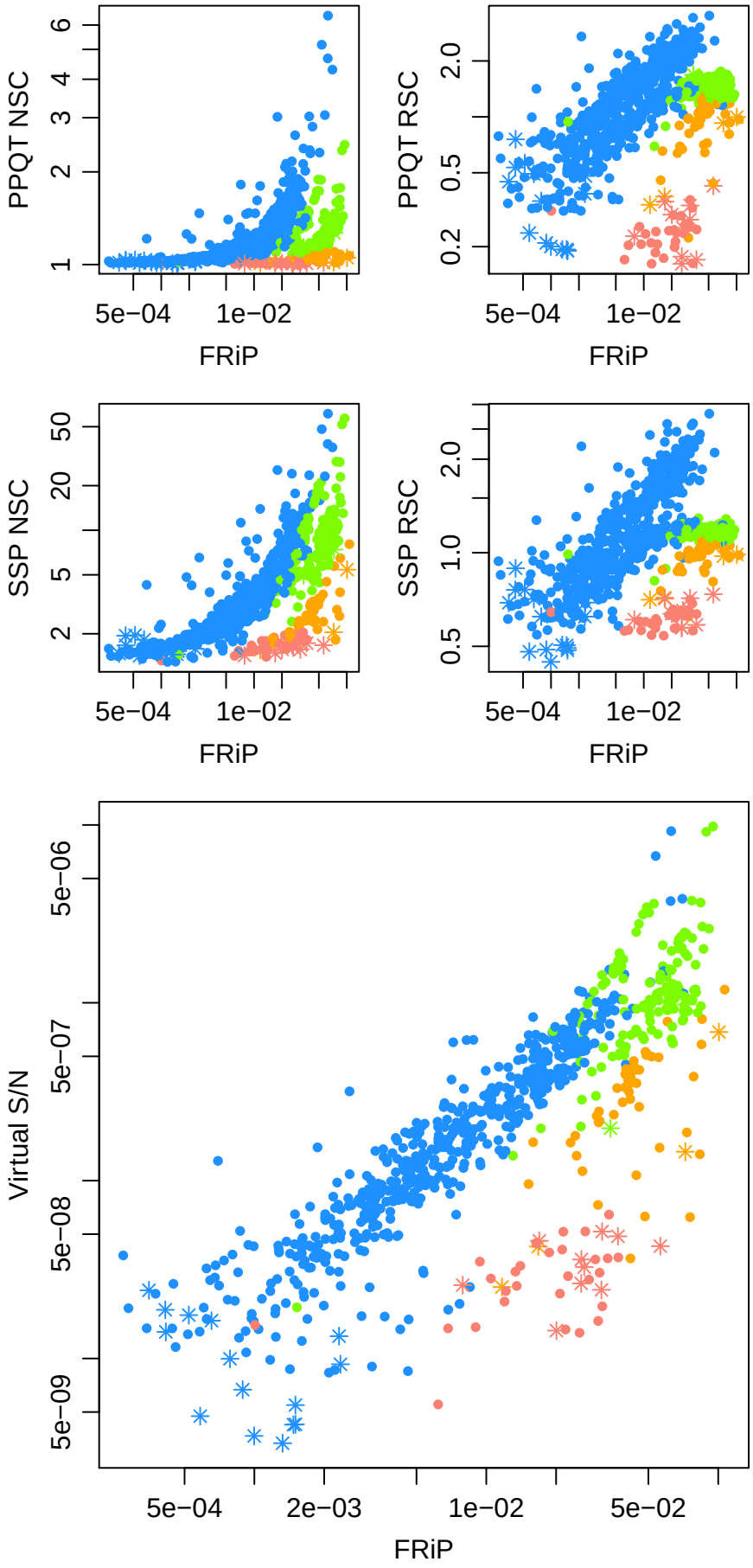


図 25: FRiP と各指標間の比較

表 14: FRiP と各指標間のスピアマンの順位相関係数

	PPQT NSC	PPQT RSC	SSP NSC	SSP RSC	VSN
全体	0.391	0.376	0.601	0.334	**0.668**
転写因子	0.915	0.835	0.926	0.824	**0.941**
ヒストン修飾（Narrow）	0.498	0.272	**0.601**	0.050	0.481
ヒストン修飾（Broad）	0.544	0.457	**0.686**	0.390	0.521
H3K9me3	0.544	0.312	**0.643**	0.477	0.454

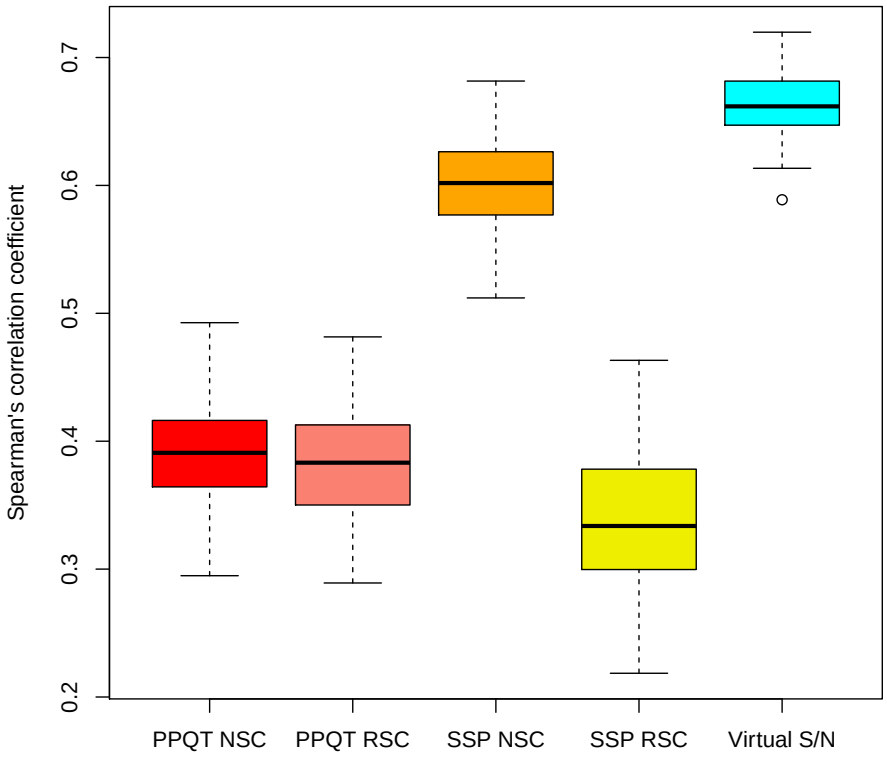


図 26: FRiP と各指標間のスピアマンの順位相関係数（100 iteration）

が最も高く、VSN はやや劣る結果となった。Narrow ヒストン修飾ではプロットが多く分けて 2 つに分離しており、個々のターゲットで比較した場合はより高い相関を示す可能性があるものの、十分なサンプル量がなかったために比較を行うことができなかった。また、MACS2 は Broad ピークの判定にやや難がある場合があり、ここで本体の値と仮定している FRiP の値についてもより検証が必要である。しかしながら、複数の種類のサンプルが混合したデータセットについても VSN はより高い相関を示した結果から、FRiP を予測する上で VSN が既存の指標と比較しても十分な性能を有していると考えられる。

20 VSN の取得リード数に対するロバスト性の検証

FRiP は取得したリード数に影響される一方、Strand cross-correlation を用いた指標はリード数に対してロバストであることが知られている。VSN はこの性質を満たすであろうか。これを検証するため、テストデータの中で特にリード数の多いサンプルを用いてダウンサンプリング解析を行った。表 15 に用いたサンプルを示す。

表 15: ダウンサンプリング解析に用いた ENCODE のデータ

実験 ID	ファイル ID	ターゲットの種類	ChIP ターゲット	総リード数
ENCSR616MOB	ENCFF587OYH	転写因子	CBX6	97,152,854
ENCSR027BPE	ENCFF127TTO	ヒストン修飾（Narrow）	H3K27ac	141,421,929
ENCSR000AUK	ENCFF000ALK	ヒストン修飾（Broad）	H3K27me3	135,411,131
ENCSR000AUN	ENCFF000AKK	ヒストン修飾（Broad）	H3K9me3	79,148,649
ENCSR000ASS	ENCFF000AHV	DNA Input コントロール	DNA input	130,433,501

これらのサンプルについて、100（100M リード以上あれば）, 50, 25, 10, 5, 2.5, 1M リードにダウンサンプリングして、各指標がどのように変化するかを検証した（図 27）。Strand Cross-Correlation を元にした指標は PPQT RSC を除くと 1M リードという極端に少ないリード数でも安定して一定の値を示していることが判る。一方 FRiP については、リード数が減少するに従い減少していく様子が確認された。VSN では、極端にリード数が少ない場合は FRiP と同様に減少が見られるものの、転写因子で 20M リード、転写因子で 45M リードに設定されている ENCODE が定めた取得リード数の目標値では、リード数が最大の時とほとんど同じ値を得ることができた。従って、一般的なリード数において VSN は他の NSC や RSC と同等のロバスト性を有していると考えられる。

21 小括

ピークコール前に SN 比を反映するであろう指標として VSN を提案した。ピークコール前に SN 比を評価できる QC 指標としては、FRiP を十分予測することができ、また既存の Strand cross-correlation を用いた指標と同じようにリード数に対するロバスト性を備えていなければならない。そこでまず、MSCC の分散から w を推定できるかを検討した。結果、MSCC の半値全幅の 1.5 倍が、MACS2 のプロファイルを用いて推定した w によく一致することを確かめた。これにより、VSN は NCC・MSCC から算出できることが明確になった。

続いて FRiP との相関をスピアマンの順位相関係数を用いて評価したところ、VSN は既存の QC 指標と比較しても FRiP との十分に高い相関を示した。また、ChIP ターゲットの種類が混在している場合も比較的高い相関を示し、VSN が他の指標よりも ChIP ターゲットの違いに対してロバストであることを発見した。さらに取得リード数に対するロバスト性を検証するため、ダウンサンプリング解析を実施した。結果、VSN は現実的なリード数の範囲で、既存の QC 指標と同

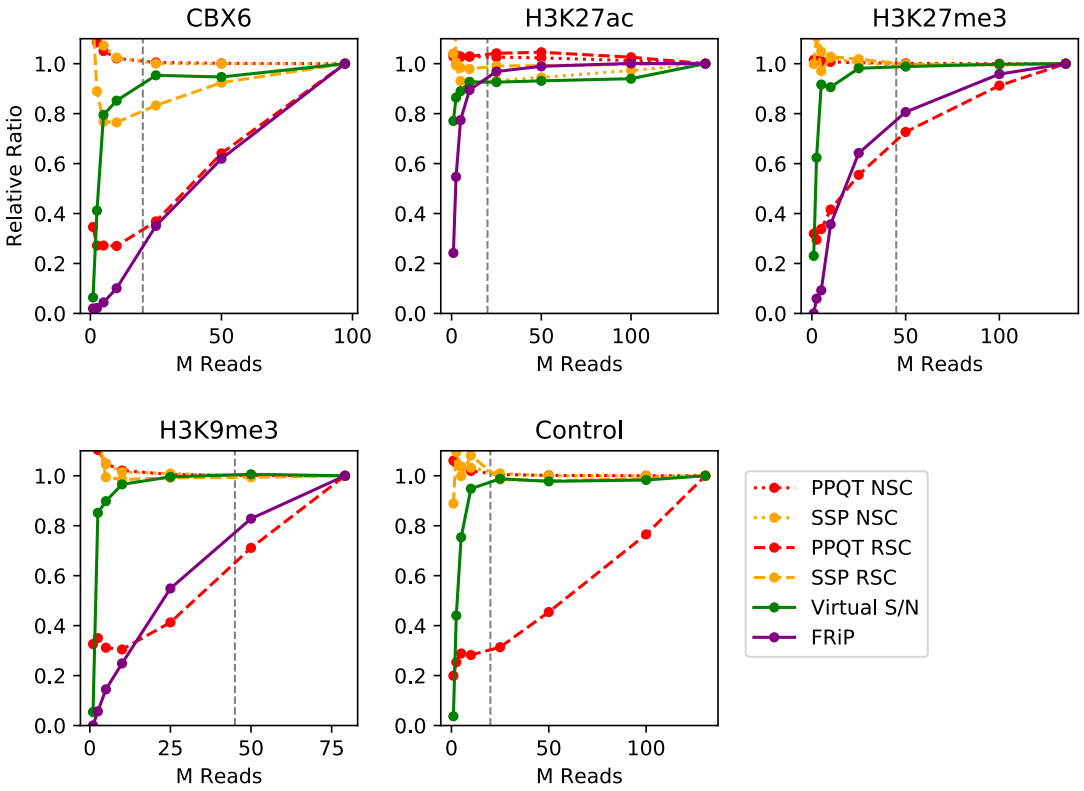


図 27: 各 QC 指標のリード数の減少に対する変化

リード数が最大の時の各指標の値を 1.0 として相対量をプロット。垂直な点線は ENCODE コンソーシアムが設定した取得リード数の目標値。

等のロバスト性を有していることが判った。

以上の結果から、本研究で提案した VSN はピークコール前に計算可能かつ FRiP と高い相関を示し、また高いロバスト性を兼ね備えた QC 指標であることが確かめられた。

第 V 部
ChIP-seq データベースの開発と転写制御解析

これまで本論文では公共 ChIP-seq データの大規模解析を実現するためのパイプラインおよび公共データを評価するための QC 手法について述べてきた。ここでは、これらの手法を実際に適用し、解析結果の可視化を行う手法について述べる。

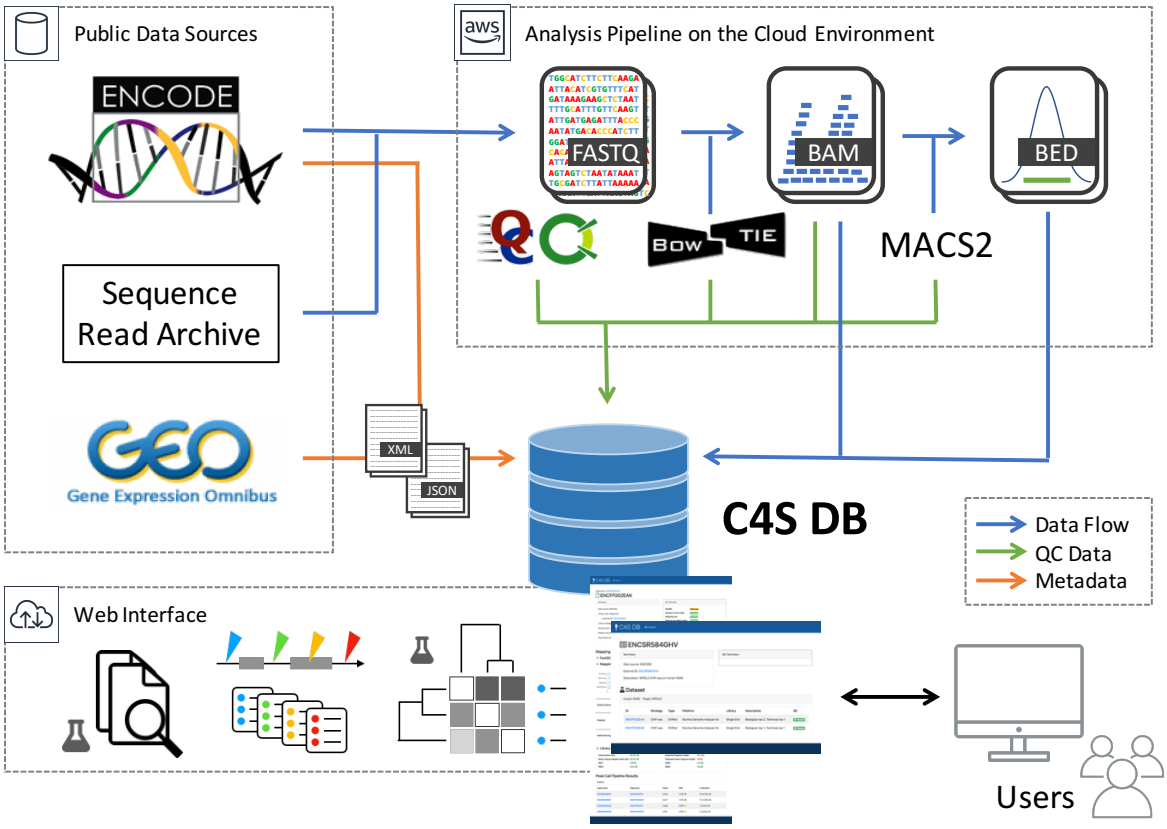


図 28: 本研究の ChIP-seq データベース構築までの概要図

22　仮想化技術とクラウドコンピューティングを用いた解析の背景

第 II 部で述べたパイプラインは主にシェルスクリプトで構成されており、一般的な Linux ディストリビューション（推奨環境は CentOS）で動作する。従って、MacOS を搭載した PC や生命科学分野で一般的に用いられる汎用スーパーコンピューターで動作させることが可能である。しかしながら、本パイプラインが用いている各ソフトウェアとその依存環境までは再現を保証することができない。これは単に動作が可能かというレベルの問題だけではなく、マイナーなバージョンの差に起因する微妙な再現性の低下などを誘発する恐れがある。

そこで本研究では、こうした解析環境の再現性・可搬性につきまとう問題を解決するため、近年学術用途でも用いられるようになったコンテナ仮想化技術とそれをサポートするクラウド計算環境を用いてパイプラインを実装し、実際の計算を行った。本章ではそれらの概要について述べる。

22.1 Docker による解析ステップのコンテナ化

以前から仮想化技術はコンピューティング分野において様々な応用例がある。従来の仮想化はあるハードウェア環境上で異なるハードウェア環境をエミュレートする技術として用いられてきた。一方、ハードウェア互換性の問題がない場面でも、OS や依存ライブラリのようなソフトウェア環境の使い分け・切り分けをする用途としても用いられるようになった。ソフトウェア環境の切り分けとして現在よく用いられている技術はコンテナ仮想化である。コンテナ仮想化はソフトウェアの実行環境のみをホストから隔離する技術で、エミュレーションの制限はあるものの仮想化時のオーバーヘッドが少なく、パフォーマンスの高い処理が可能で仮想イメージのサイズも最小限にすることができる。

コンテナ仮想化技術として現在最も一般的なのは Docker である。Docker エンジンはファイルに記述された設定に従い仮想環境を構築しコンテナイメージとして作成する。イメージはカーネルが同一で Docker が動作するならば異なる環境で実行することができ、また Docker イメージのレポジトリが構築され配布・ダウンロードが容易に行えることから仮想環境をインフラストラクチャーとして利用することが可能になった。ソフトウェアや解析環境を配布することがより容易になり、生命科学分野でも BioContainers[78] のようなレポジトリが構築されるようになった。

本研究でも、パイプラインの各ステップを Docker コンテナとして仮想化した。これにより解析の再現性を確保すると共に、将来的には解析環境の共有も視野に入れている。

22.2 AWS (Amazon Web Services) によるクラウドコンピューティング

従来、学術分野において計算資源を確保する方法としては、専用計算機を購入しオンプレミスで運用するか、スーパーコンピューターのような共有計算機を借りるのが一般的であった。前者はイニシャルコストが高く、またランニングコストや管理のコストも発生するため、研究や予算の規模によっては難しい選択肢であるといえる。また共有計算機の場合は、共同で使う上での制限や計算環境を構築する上での限界があり、計算環境として活用できるようにするためのコストが小さくない。

一方、新たな計算資源の利用形態として台頭してきたのがクラウドコンピューティングである。特に近年クラウドコンピューティングの代名詞となっている AWS (Amazon Web Services) や Microsoft Azure, Google Cloud Platform のようなサービスが従来の共有計算機と異なるのは、計算機を単にリモートで扱えるだけではなく、あらゆるハードウェア資源が仮想化され、ユーザーが必要な量を必要な時間だけ使用できるサービス（IaaS, Infrastracture as a Service）として提供されている点である。従量課金制の導入により、従来では採算の取りにくかった中・小規模な計算が容易に実行できるようになった。また前述のコンテナ仮想化を組み合わせることにより、計算をより大規模かつインフラストラクチャーの管理をせずとも実行することが容易になった。

生命情報科学分野においてもクラウドコンピューティングの活用が活発となっている。元々 Galaxy[79] のような SaaS（Software as a Service）の先駆けとも言えるツールが多く WEB 上に公開されてきたなどインターネットを活用したコンピューティングが盛んな分野であったが、ENCODE が AWS 上でデータ公開を開始 [80] したり、SRA のデータが AWS や Google Cloud

Platform でアクセス可能になる [81] など、公共データの解析がクラウドコンピューティング環境で行いやすくする流れがある。

本研究ではクラウドコンピューティング環境として AWS を選定し、Docker イメージをベースとした計算環境を展開し解析を行った。

23 解析パイプラインの AWS へのデプロイと公共 ChIP-seeq データの大規模再解析の実施

本章では Docker により仮想化した解析パイプラインが AWS 上にどのように展開し、また実際に解析を行った結果について述べる。

23.1 AWS を構成するサービスの概要

本節では今回利用した AWS のアーキテクチャの記述に先立ち、クラウドコンピューティング環境を構成する主要なサービスについて概説する。

23.1.1 仮想マシンの構成

- AMI (Amazon Machine Image) … 仮想マシン（VM）のイメージ。主要な OS や、特定のソフトウェアがプリインストールされたイメージが提供されている。
- Amazon EC2 (Elastic Compute Cloud) … AWS において VM を提供するサービスであり、各 VM はインスタンスと呼ばれる。ユーザーは需要に応じて vCPU・RAM 容量などが設定されたインスタンスタイプと起動に用いる AMI を指定し VM を作成する。
- Amazon EBS (Elastic Block Store) … EC2 インスタンスがマウントする仮想的な外部記憶装置。

23.1.2 共有データストレージ

- Amazon S3 (Simple Storage Service) … データストレージサービス。VM にマウントされるストレージは EBS が担うが、S3 はデータの保存・共有が容易でありインスタンスを必要とせず、またコストも EBS より安い。本研究では入力となる FASTQ ファイルと解析結果の一時的保存に用いた。
- Amazon EFS (Elastic File System) … EC2 用ネットワークファイルシステム（NFS）。リファレンスゲノムのインデックスのような複数の EC2 インスタンスが必要とするファイルにアクセスする場合、S3 を経由してその都度 EBS にダウンロードするよりも NFS を用いて共有する方が適している。本研究では解析パイプラインが依存するデータを置くために用いた。

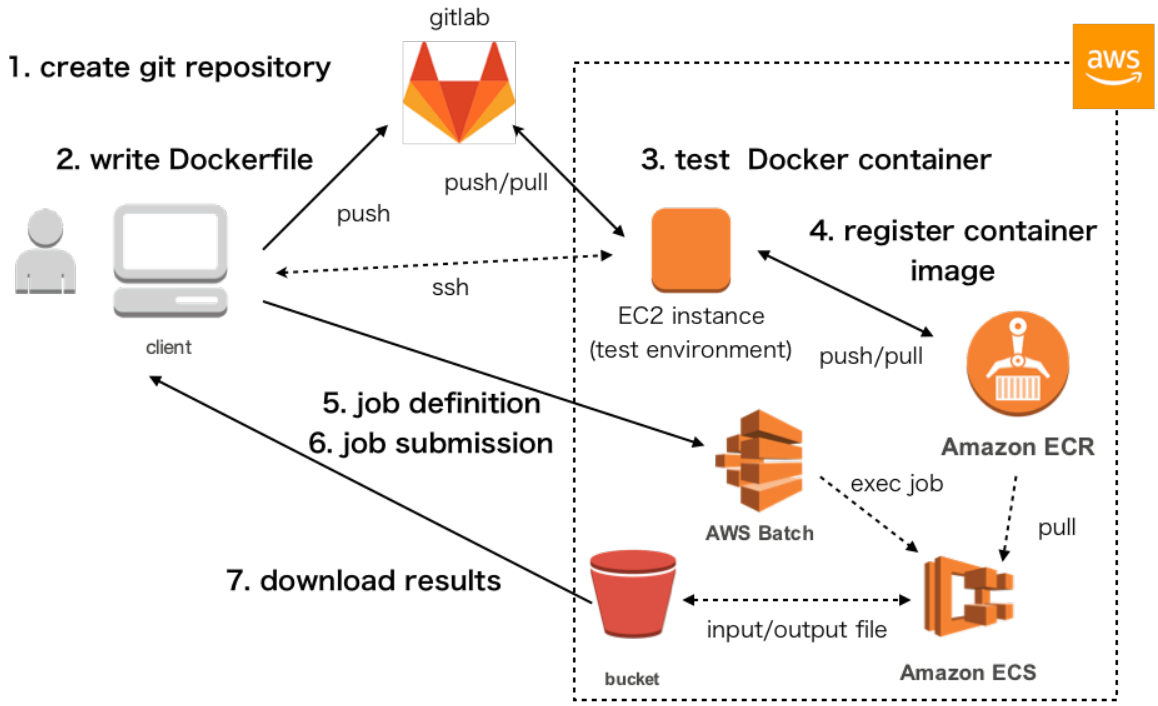


図 29: AWS を用いた Docker コンテナのテスト環境と AWS Batch によるジョブの実行

23.1.3 コンテナ実行環境

- Amazon ECR (Elastic Container Registry) … Docker コンテナレジストリ。パブリックなレジストリとしては Docker hub などがあるが、ECR はユーザーが作成した Docker イメージのためのレポジトリを提供し、AWS 内やユーザーの実行環境向けに Docker イメージを配信することを容易にする。
- Amazon ECS (Elastic Container Service) … Docker コンテナの実行・管理を提供するサービス。ECS は EC2 インスタンスの作成とクラスター化、Docker イメージの実行、自動スケーリング等を自動で行うことができる。
- AWS Batch … AWS におけるバッチジョブ実行サービス。ECS を用いることで Docker コンテナベースの環境上で大量のジョブをスケーラブルに処理できる。

23.2 コンテナ仮想化技術とクラウドコンピューティング環境による解析パイプラインの実装

AWS によるテスト環境、実行環境を用いたワークフローを図 29 に示す。AWS Batch が EC2 インスタンスの作成に使用する AMI が使用できるため、ユーザーは実際の実行環境と同一の Docker がインストール済の環境を容易に作成することができ、Docker コンテナ化を比較的容易に行うことが出来た（ステップ 3）。本研究では Docker イメージを管理する git レポジトリを作成し、研究室でホストしている gitlab を用いて管理した（ステップ 1-2）。

作成した Docker イメージは Amazon ECR に登録し、AWS 内のサービスから使用できるよう

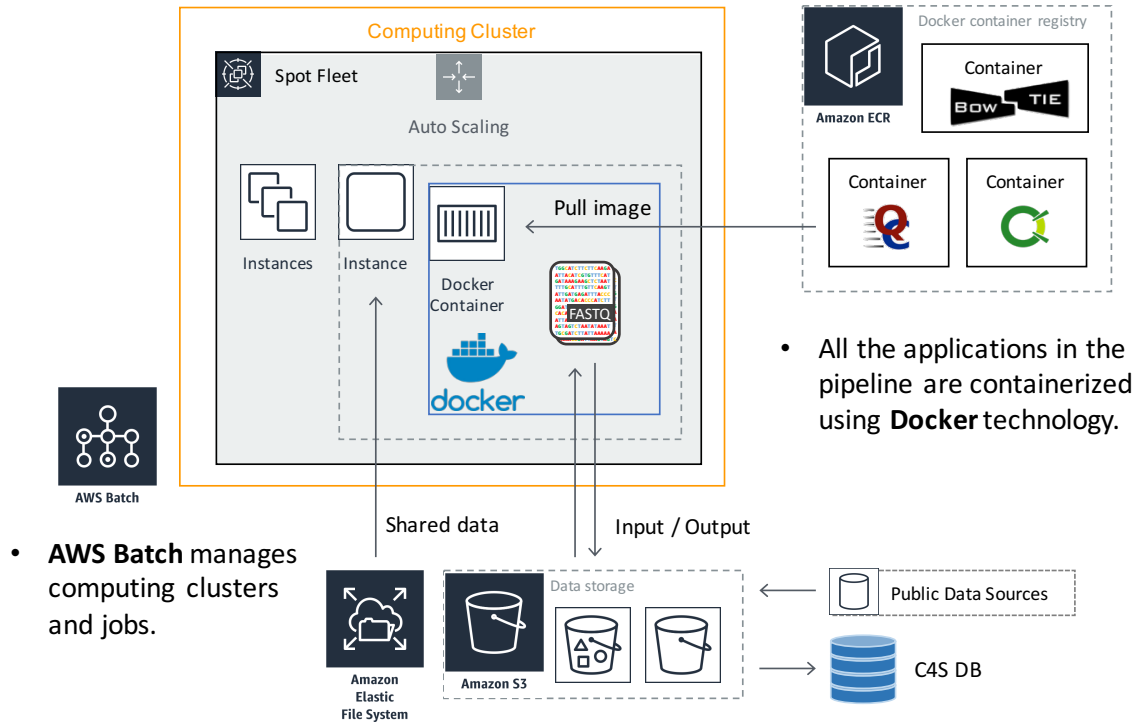


図 30: 本研究で用いた AWS Batch のアーキテクチャ

にした（ステップ 4）。続いて AWS Batch が処理に用いるジョブ定義を ECR に登録したコンテナを用いて作成し（ステップ 5）、ジョブ定義を元に必要なパラメータを合わせたジョブを発行することで解析を行える（ステップ 6）。解析結果は Amazon S3 に格納されダウンロードできるように実装した（ステップ 7）。AWS Batch のアーキテクチャを図 30 に示す。

23.3 大規模解析の実施と計算コストの評価

ここではデモンストレーションとして、ENCODE Project より A549 および K562 細胞の ChIP-seq データを解析した際の計算コストについて述べる。1 ヶ月間に計算したデータを表 31 に示す。なお、集計が 1 ヶ月単位なのは AWS の請求が月毎であったためであり、解析のスループットについては常時 1,000vCPU コア以上の計算資源を活用できたため、実際の計算は数日で完了している。また、インスタンスはアイドル状態の計算資源をおおよそ通常価格の 6 割程度で使用できるスポットインスタンスをできるだけ使うよう設定した。また、AWS では S3 へのデータの保存・データの AWS 外へのダウンロードに容量に応じた費用が掛かるが、1 回のダウンロードに必要なコストは S3 で保存する場合のおよそ 3.5 ヶ月分であり、長期的な観点では研究室で管理しているオンプレミスのファイルサーバーで管理する方がコストが安いとの判断から、解析が完了したデータはできるだけ早くダウンロードし S3 上のファイルは削除するようにした。

図 31: 集計の対象となった ChIP-seq データ

	データ量
実験数	1,081
サンプル数	2,943
入力 FASTQ の総量	4.45TB

なお、AWS 内へのデータのアップロードは課金されない。

これに対して計算費用とその内訳は図 32 のようになった。Storage は入力および解析データの保存に使用される金額であり、処理するデータ量に対しておおよそ不変である。従って大規模解析において金銭的コストを占めるのはコンピューティングの費用とデータのダウンロードに掛かる費用である。今回はこれらを合算すると約 2,461 米ドルであった。ENCODE の各実験はサンプルか少なくとも duplicate になるようデザインされており、実験数・サンプル数と比較すると、1 実験当たり約 2 ドル、1 サンプル当たり約 1 ドルという概算を得ることができた。公開されているヒト ChIP-seq 実験を 5 万件とすると、費用は約 5 万ドルとなる。AWS では大規模な並列化が可能であり計算に必要な実時間が短期間で済むこと、計算機の管理コストやランニングコストが不要であることを考慮すると、オンプレミスで計算環境を整備するよりも十分実用的であると言える。

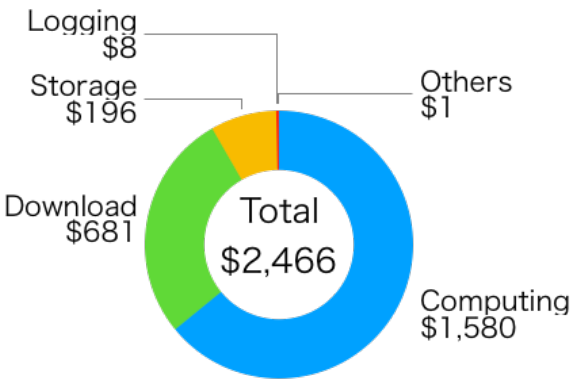


図 32: データセットに対する AWS の計算コスト

24 ChIP-seq データベースの設計と実装

実験データの大規模な解析ができたとしても 1 人の研究者が個々の解析結果全てを検討するのは困難であり、解析結果はそれを必要としている研究者がリーチできるよう公開しなければならない。特に ChIP-seq のデータは多くの要素が協調的に関与する転写制御を対象にしていることも多く、1 つの実験データにたどり着くための検索機能だけでなく、複数のデータの関係性を明らかにし、データの観察者が新たな生物学的発見をできるよう促す必要がある。これらの需要を満たすために本研究では、

- 実験データごとの解析結果の可視化
- 遺伝子周辺領域における解析結果の可視化
- 実験データ間の大域的な類似度の可視化

という 3 つの機能を軸としてデータベースの開発を行った。本章ではこれまでの成果を元に得られたデータを可視化する公共 ChIP-seq データベース、C4S（Comprehensive Collection and Comparison for ChIP-Seq）DB の設計、実装と使用例について記述する。

執筆段階で、C4S DB はテストバージョンを https://test.c4s.site にて公開している。将来的には https://c4s.site で公開される予定である。

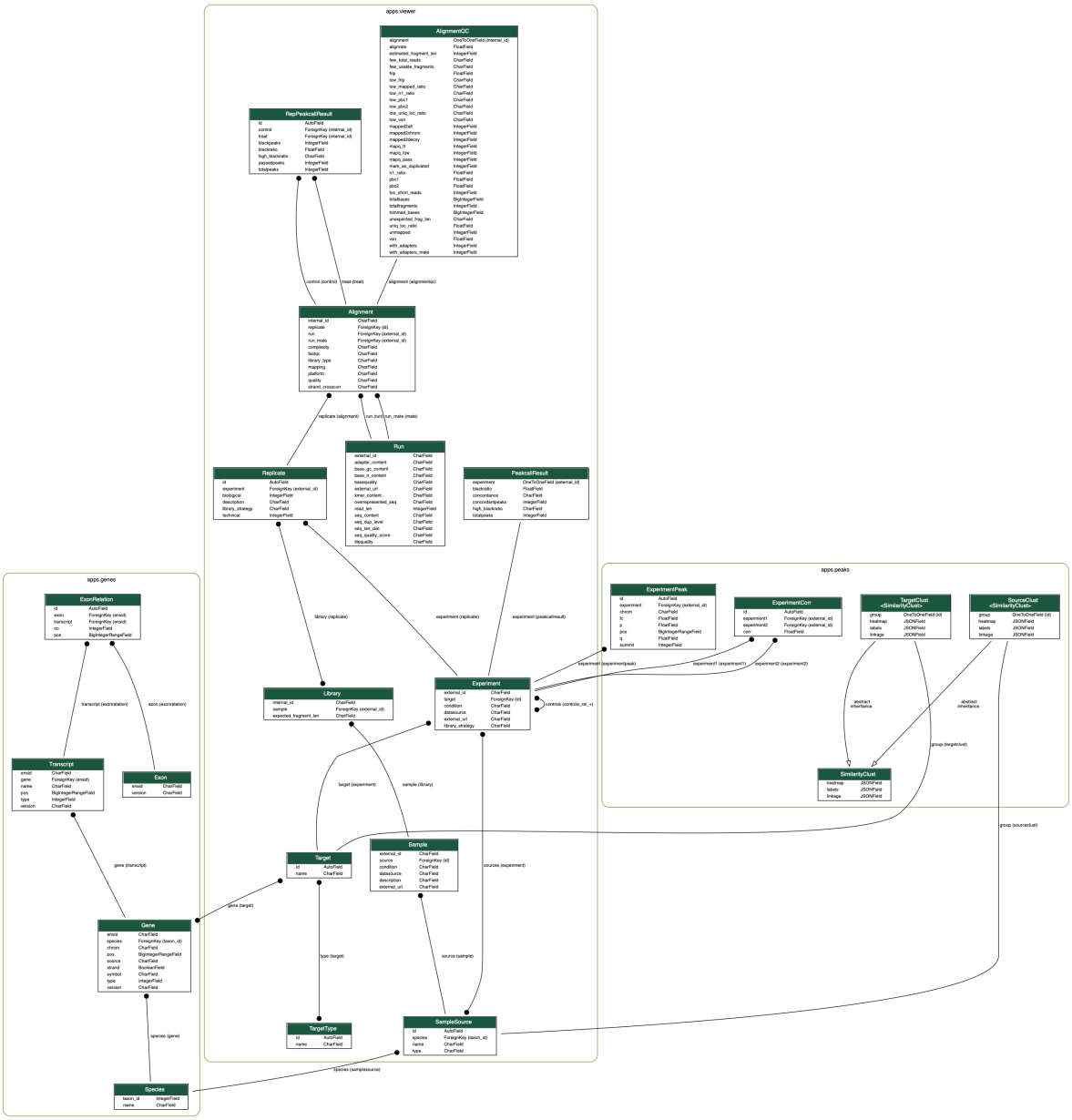


図 33: C4S DB のデータベーススキーマ

24.1 Web アプリケーションフレームワークとデータベーススキーマ

データベースの実装には Django を用いた。Django は Python で実装された Web アプリケーションフレームワークであり大規模な Web アプリケーションにも対応している。データを格納するリレーショナルデータベースは複数の管理システムに対応しているが、本研究では PostgreSQL を採用した。PostgreSQL は範囲型をサポートしており、遺伝子領域やピークの範囲など領域に対応するデータを効率よく格納、検索することができるためである。

C4S DB のデータベーススキーマを図 33 に示す。C4S DB は内部的には主に 3 つの Django アプリケーションから構成されている。GENCODE[82] の遺伝子アノテーションを格納する `genes`、

メタデータと解析結果を格納する `viewer`、解析結果のうち 1 実験に対応する Concordant なピークを格納する `peaks` である。

24.2 実験データごとの解析結果の可視化（Data Browser）

図 34 はある Run（1 つの FASTQ ファイル）の解析結果を可視化したページである。FastQC の解析結果、マッピングまでの各段階で使用可能なリード数の変化、Library complexity, Strand cross-correlation, FRiP による QC 情報と、このデータが関与するピークコールの結果について確認できる。

1 実験に対するページの例を図 35 に示す。これは ChIP を行った実験であるが、対応するコントロール実験へのリンクが表示されている。実験に含まれる Run やそれらのピークコール結果の一覧に合わせて Joint peak call の結果が表示される。右側のグラフは Joint peak call で得られた各ピークがレプリケートごとのピークコールで何回検出されたかを表しており、これによりレプリケートの一貫性を確認することができる。この情報を元に、ユーザーは左のパネルで条件を指定して Joint peak call で得られたピークをダウンロードできる。

24.3 遺伝子周辺領域における解析結果の可視化（Gene Viewer）

ChIP-seq データを観察する一般的な方法の 1 つは、特定の遺伝子に着目してその周辺のピークをゲノムブラウザ等で可視化することである。特に研究者によっては特定の遺伝子（群）に着目して研究を行う場合も多く、ユーザーが指定した遺伝子周辺の可視化は大きな需要があると考えられる。

図 36 は *TAF9B* 遺伝子の周辺（上流・下流 1,500bp を含む）を可視化した例である。ページは主にトランスクリプトの一覧、ピークを可視化したトラック、周辺の遺伝子一覧から構成されている。トランスクリプト一覧では GENCODE のアノテーションに含まれているトランスクリプトについてイントロン・エクソンを可視化し、各ピークが遺伝子のどの領域と関係がある可能性があるのか確認しやすくしている。C4S DB は指定された遺伝子に対し、遺伝子本体と上流・下流領域に重複するピークを検索し、実験ごとに 1 トラック（列）にして表示する。また、実験単位で並べるとデータが煩雑になってしまうため、各実験をさらに ChIP ターゲットもしくはサンプルソースごとにグルーピングする機能を実装した。この例では、遺伝子右端の転写開始点付近に様々な種類の制御因子がエンリッチしている様子を確認できる。また、PostgreSQL 上でのインデクシングにより約 400 万ピークが登録されたデータベースから 1,000 ピークを取得してユーザーの Web ブラウザに応答するまでを約 1.5 秒程度で処理できた。

24.4 実験データ間の大域的な類似度の可視化（Global Similarity）

Gene Viewer によって遺伝子-転写制御因子間の関係を発見することができても、制御因子同士の関係、例えば制御因子の共起について観察するのは困難である。しかしこのような関係は転写因子複合体の形成といった制御機構に対してアプローチする上で重要な情報となる。また、多数の実験間で類似度を計算し可視化することで、特定の ChIP ターゲットの実験結果にどれ位の一

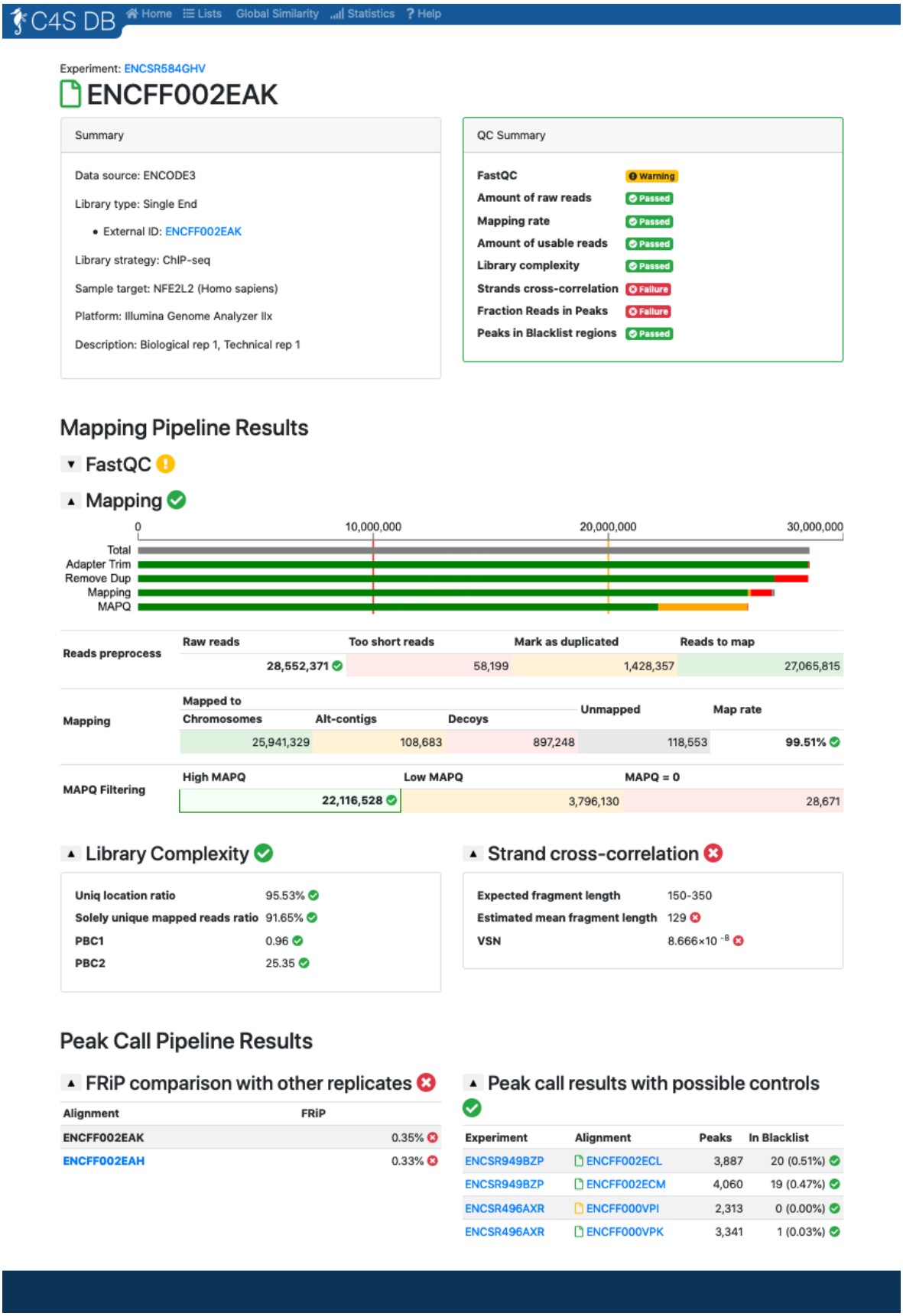


図 34: ChIP-seq データ解析結果の可視化例（1Run）

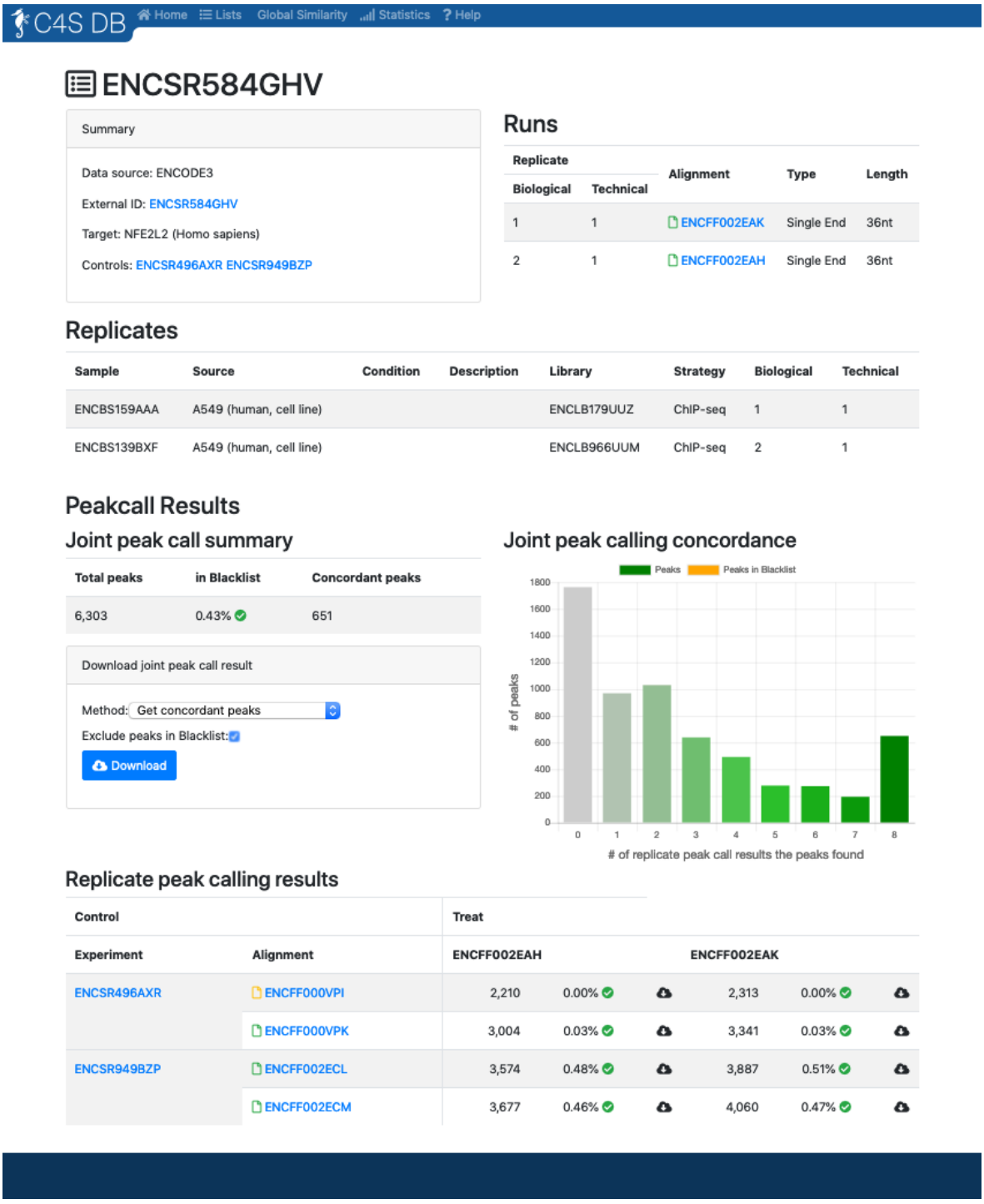


図 35: ChIP-seq データ解析結果の可視化例（1 実験）

図 36: 遺伝子と周辺領域のピークの可視化例

貫性やばらつきがあるのかや、共起・排他性を図示することで転写制御機構の全体像を提供し俯瞰的な視点からの生物学的知見をもたらせる可能性がある。

Global Similarity は実験間の大域的な類似度を元にクラスタリングを行い、実験間の関係をヒートマップで可視化する機能である（図 37）。現段階では、特定のサンプルソース（細胞株等）もしくは ChIP ターゲットに対する実験の可視化をサポートしている。この例では 283 実験の関係を巨大なヒートマップで描画しており、ユーザーは上下・左右方向にスクロールしてヒートマップを確認できる。

24.4.1 ピークを用いた実験間の類似度の計算

ここでピークを用いて実験間の類似度を計算する手法を説明する（図 38）。例えば遺伝子発現量のようなデータの場合、データの要素数（遺伝子数）が共通しているため、2 つのベクトル間で相関係数のような指標を適用することができる。しかし、ChIP-seq などで得られる結果は領域も可変であるためにそのままでは要素数が揃っておらず類似度の計算が難しい。そこで、最初に与えられた 2 つのピーク群の領域をマージした新しいピークの集合を定義する。次に、この集合に対して 2 つのデータそれぞれのピークが持つ非負のスコアをアサインする。ピークが存在しない場合はスコアは 0 とし、2 つ以上存在する場合は平均を取る。最終的に、マージしたピーク領域に対する 2 つの長さが等しいスコアのベクトルが得られるため、相関係数等の指標を用いることができる。本研究では、ピアソンの相関係数を使用した。また、クラスタリングの際は類似度を距離に変換する必要があるが、ここでは 1 – (相関係数) を距離として用いた。

25 データベースを用いた転写制御機構の解析例

最後に本データベースを用いた解析のデモンストレーションとして、ヒト肺癌細胞由来の A549 細胞株の ChIP-seq データセットに対する分析を行う。ここで示す A549 データセットは、ENCODE Project および GGR の実験 283 件から構成されている。また、GGR の実験は主にデキサメタゾンを添加した時系列の実験データが多く含まれているのが特徴である（図 39）。

25.1 Global Similarity を用いた分析

図 40 は図 38 のヒートマップ全体を描画したものである。類似度が比較的高いクラスターに注目すると、同じ制御因子をターゲットとする実験でクラスターが構成されていることが分かった。特に顕著なターゲットについては範囲とラベルを示している。

ここで示したクラスターについて、転写活性化とインシュレーター形成、グルココルチコイド受容体関連遺伝子の 2 つに分けて考察する。

25.1.1 転写活性化とインシュレーター形成

ここでは *POLR2A*, H3K4me3, H3K27ac クラスターおよび *RAD21*, *SMC3*, *CTCF* クラスターに着目する（図 41）。

POLR2A は RNA ポリメラーゼⅡ A をコードする遺伝子であり、ここではゲノム中の RNA ポリメラーゼの分布を反映していると考えられる。従って、これらの領域は遺伝子の転写開始点や

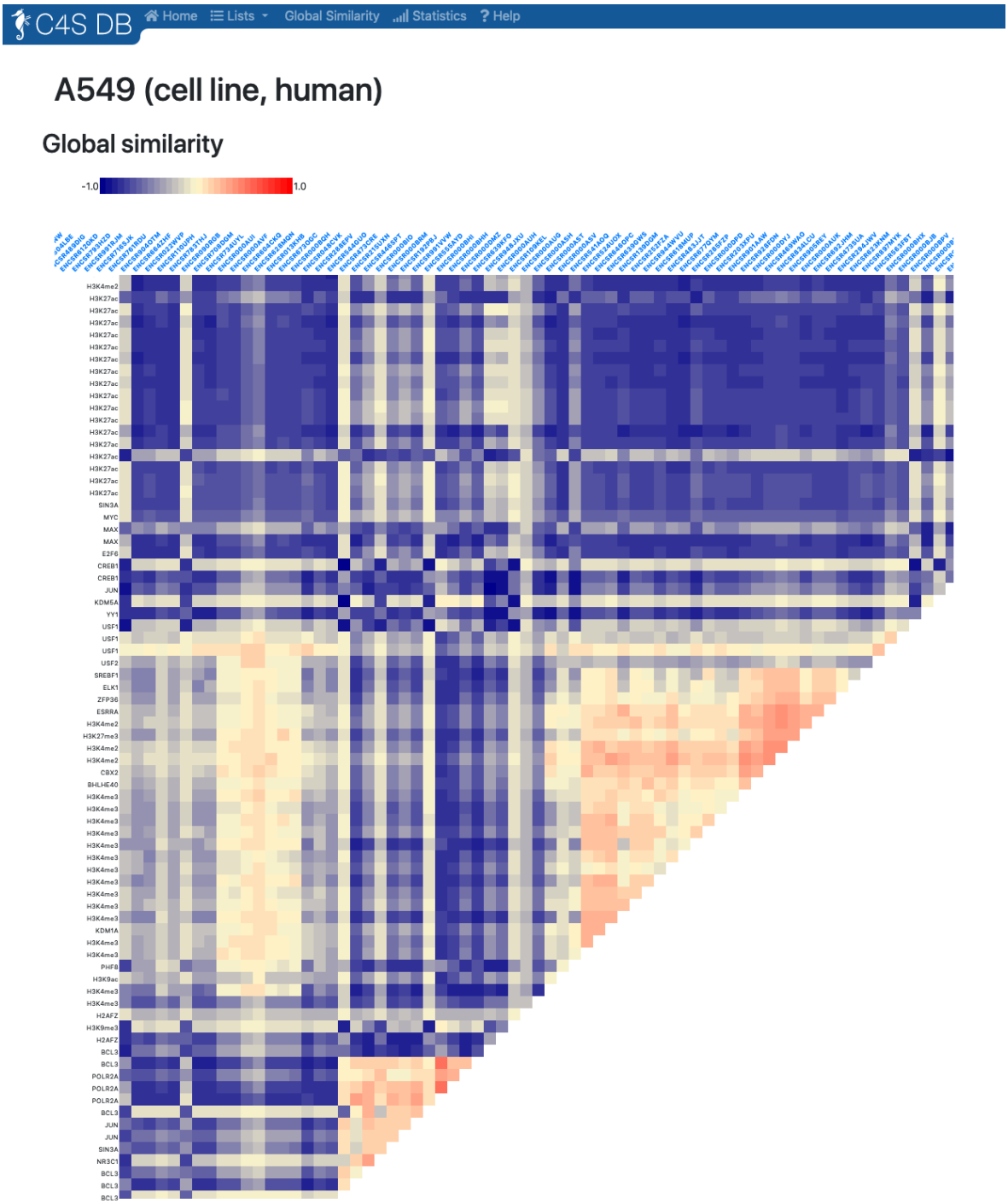


図 37: 実験間の類似度の可視化例（A549 細胞）

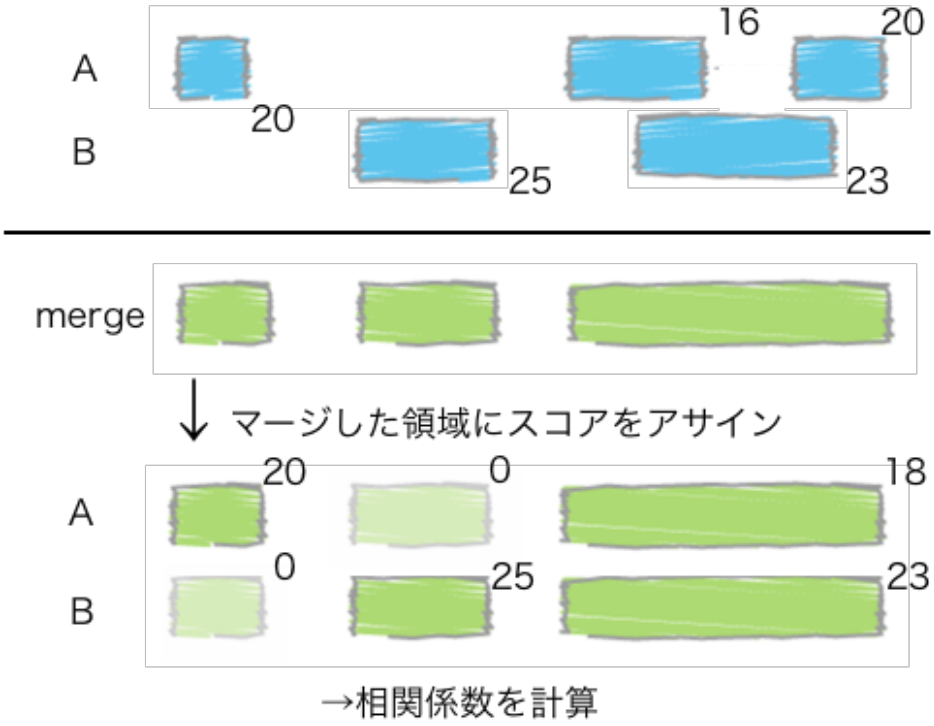


図 38: 2 つのピーク群間の類似度の計算方法

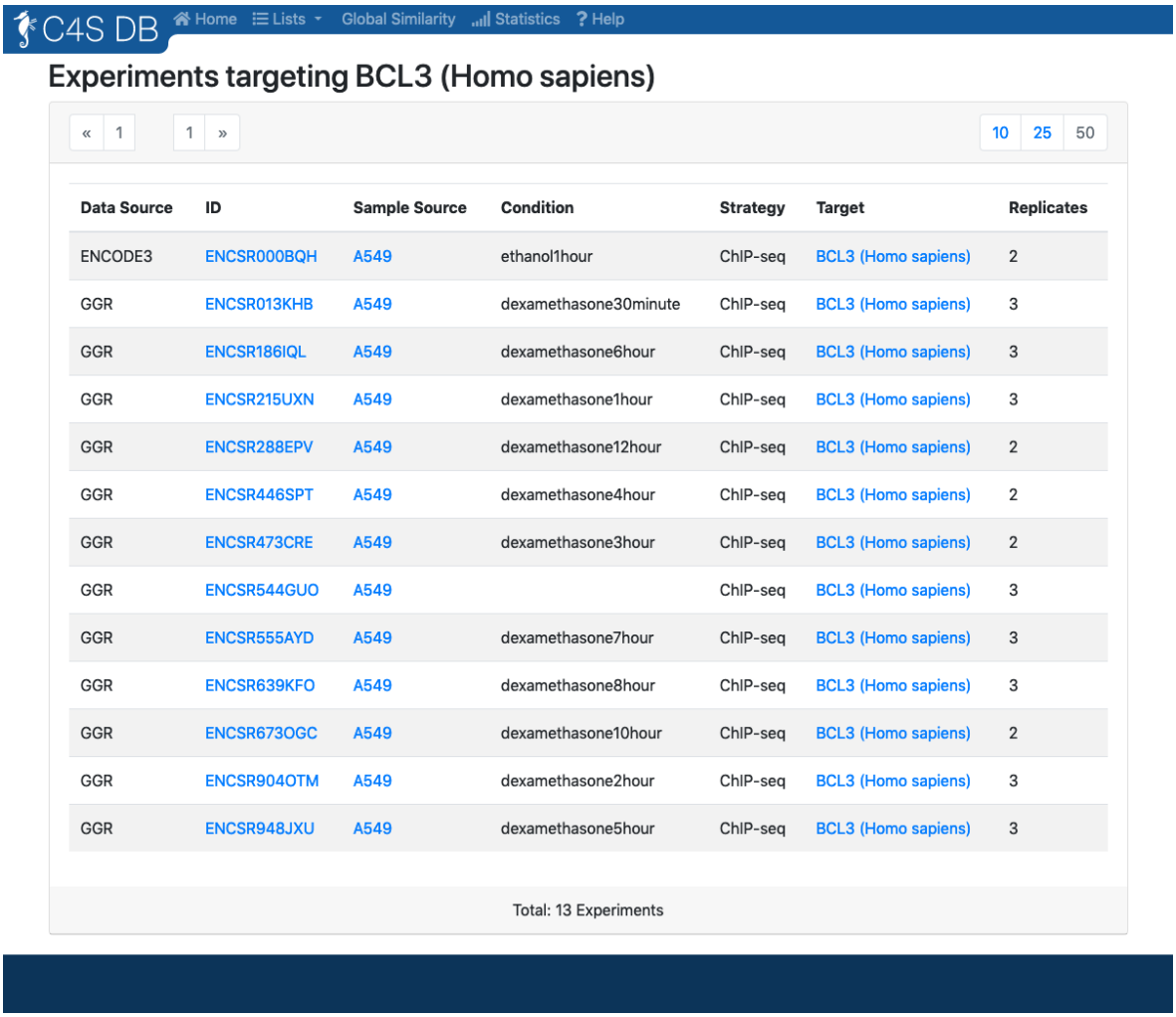
遺伝子領域に強い関係がある。*POLR2A* はヒストン修飾のうち H3K4me3 および H3K27ac と強い相関を示した。これらはプロモーター活性を示すヒストン修飾であり [84–86]、これらは遺伝子の転写活性を反映しているクラスターであると考えられる。

遺伝子発現は様々な転写制御を受けるが、時に遺伝子領域は隣接・重複している場合があり、転写制御を確実にするためにゲノム領域を分画するインシュレーターが存在する [83]。インシュレーターの形成には CCCTC-binding factor (CTCF) とコヒーシンタンパクが必須であり [87]、SMC3 と RAD21 はコヒーシンタンパクを構成する [88–90]。*RAD21*, *SMC3*, *CTCF* クラスターはインシュレーターが形成されている箇所を示していると考えられる。

また、これら 2 つのクラスター群の関係に着目すると、インシュレーターを跨いだ遺伝子の転写は起こりにくいと考えられるため、これらの結合箇所は互いに背反していると予想される。実際に図 41 ではこれらのクラスター群が交わる箇所が全体から見ても特に相関係数が低いことを表しており、この類似度マップが上記の生物学的関係をよく反映していると考えられる。

25.1.2 グルココルチコイド受容体関連遺伝子

グルココルチコイド受容体（GR）は *NR3C1* 遺伝子にコードされた核内受容体スーパーファミリーに属するリガンド依存的な転写制御因子である [91]。GR はステロイドホルモンの一種であるコルチゾールの他、人工的な GR のリガンドであるデキサメタゾン（DEX）にも強く活性化される。図 42 中で下線を引いた遺伝子は DEX 添加により誘導されることが知られている遺伝子である [92]。これらの遺伝子が含まれるクラスターがよく確認できることは、GGR に由来する実験データが実験条件特異的に強く検出されたピークによって相関が強調された可能性を示唆してお



Data Source	ID	Sample Source	Condition	Strategy	Target	Replicates
ENCODE3	ENCSR000BQH	A549	ethanol1hour	ChIP-seq	BCL3 (Homo sapiens)	2
GGR	ENCSR013KHB	A549	dexamethasone30minute	ChIP-seq	BCL3 (Homo sapiens)	3
GGR	ENCSR186IQL	A549	dexamethasone6hour	ChIP-seq	BCL3 (Homo sapiens)	3
GGR	ENCSR215UXN	A549	dexamethasone1hour	ChIP-seq	BCL3 (Homo sapiens)	3
GGR	ENCSR288EPV	A549	dexamethasone12hour	ChIP-seq	BCL3 (Homo sapiens)	2
GGR	ENCSR446SPT	A549	dexamethasone4hour	ChIP-seq	BCL3 (Homo sapiens)	2
GGR	ENCSR473CRE	A549	dexamethasone3hour	ChIP-seq	BCL3 (Homo sapiens)	2
GGR	ENCSR544GUO	A549		ChIP-seq	BCL3 (Homo sapiens)	3
GGR	ENCSR555AYD	A549	dexamethasone7hour	ChIP-seq	BCL3 (Homo sapiens)	3
GGR	ENCSR639KFO	A549	dexamethasone8hour	ChIP-seq	BCL3 (Homo sapiens)	3
GGR	ENCSR673OGC	A549	dexamethasone10hour	ChIP-seq	BCL3 (Homo sapiens)	2
GGR	ENCSR904OTM	A549	dexamethasone2hour	ChIP-seq	BCL3 (Homo sapiens)	3
GGR	ENCSR948JXU	A549	dexamethasone5hour	ChIP-seq	BCL3 (Homo sapiens)	3

図 39: A549 を対象とした実験の例（BCL3 遺伝子）

この例では、ENCODE Project（phase 3）では 1 件の実験が実施されている。GGR では DEX 添加後 12 時間までの時系列のデータが取得されている。

り、実験条件によるフィルタリング等で描画を制御できるようにすることは今後の課題と言える。

図中最下部のクラスターに含まれる *JUNB* と *FOSL2* 遺伝子はそれぞれ jun-B, FOSL2 タンパクをコードしており、これらが属する JUN ファミリーと FOS ファミリーはヘテロダイマーとなり AP-1 転写因子複合体を形成することが知られている [93]。JUN ファミリーには上から 2 番目のクラスターにある *JUN* 遺伝子がコードする c-Jun タンパクも属するが、このヒートマップでは *JUN* を対象とした実験の一部のみが *JUNB* と *FOSL2* との弱い相関を示している。また、*JUN* と共にクラスターを形成している *BCL3* は AP-1 の活性化を促すことが分かっている [94]。

GR と *CEBPB* がコードする C/EBPβ（CCAAT/enhancer-binding protein beta）ないし AP-1 複合体は共結合することが明らかになっている [95,96]。これらは凝集したクロマチンに真っ先に結合してクロマチン構造の再構成や他の転写因子のリクルートを行うことからしばしば Pioneer factor と呼ばれ [97]、GR の結合を助けていると思われる。図 42 では *NR3C1* クラスター

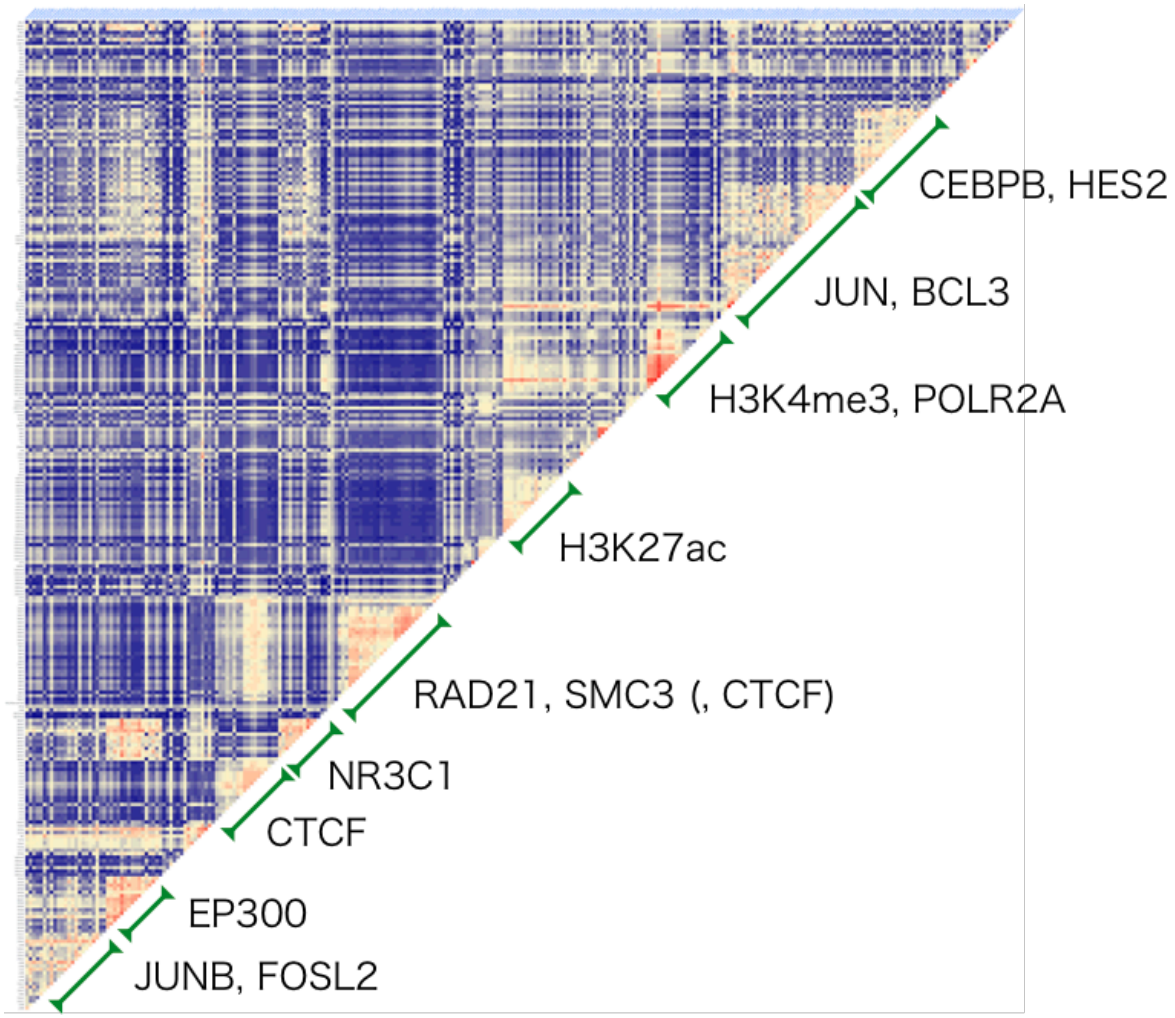


図 40: A549 ChIP-seq データセットの全実験間類似度マップ

と AP-1・C/EBPβ クラスターには強い相関が見られないものの、DEX を添加した一部の実験ではやや高い相関が見られた。GR と Pioneer factor の共結合は DEX 添加後の経過時間と関係していることを示唆しているのかもしれない。

EP300 遺伝子はヒストンアセチル化酵素複合体を形成する p300 をコードしている。p300 は転写のコアクティベーターであるが、結合部位が GR と競合していると考えられている [98]。図 42 でも *NR3C1* と *EP300* のクラスターの相関係数が高いことが確認できる。

25.2 Gene Viewer によるグルココルチコイド受容体関連遺伝子の確認

Global Similarity を用いた考察では GR とそれに関連する転写制御因子のクラスターが確認されたが、GR が DEX のレセプターとなっていることから、GR を起点として DEX 誘導性の転写因子を発現させるカスケードが存在することが予想される。そこで、Gene Viewer を用いてこれらの遺伝子周辺に GR や関連転写因子の結合部位が存在するか確認した。

まず *NR3C1* 遺伝子そのものの周辺領域を表示した様子を図 43 に示す。C/EBPβ のピークや、AP-1 を構成する c-Jun, FOSL2 のピークが確認できた。ここでは *JUNB* のピークは見られな

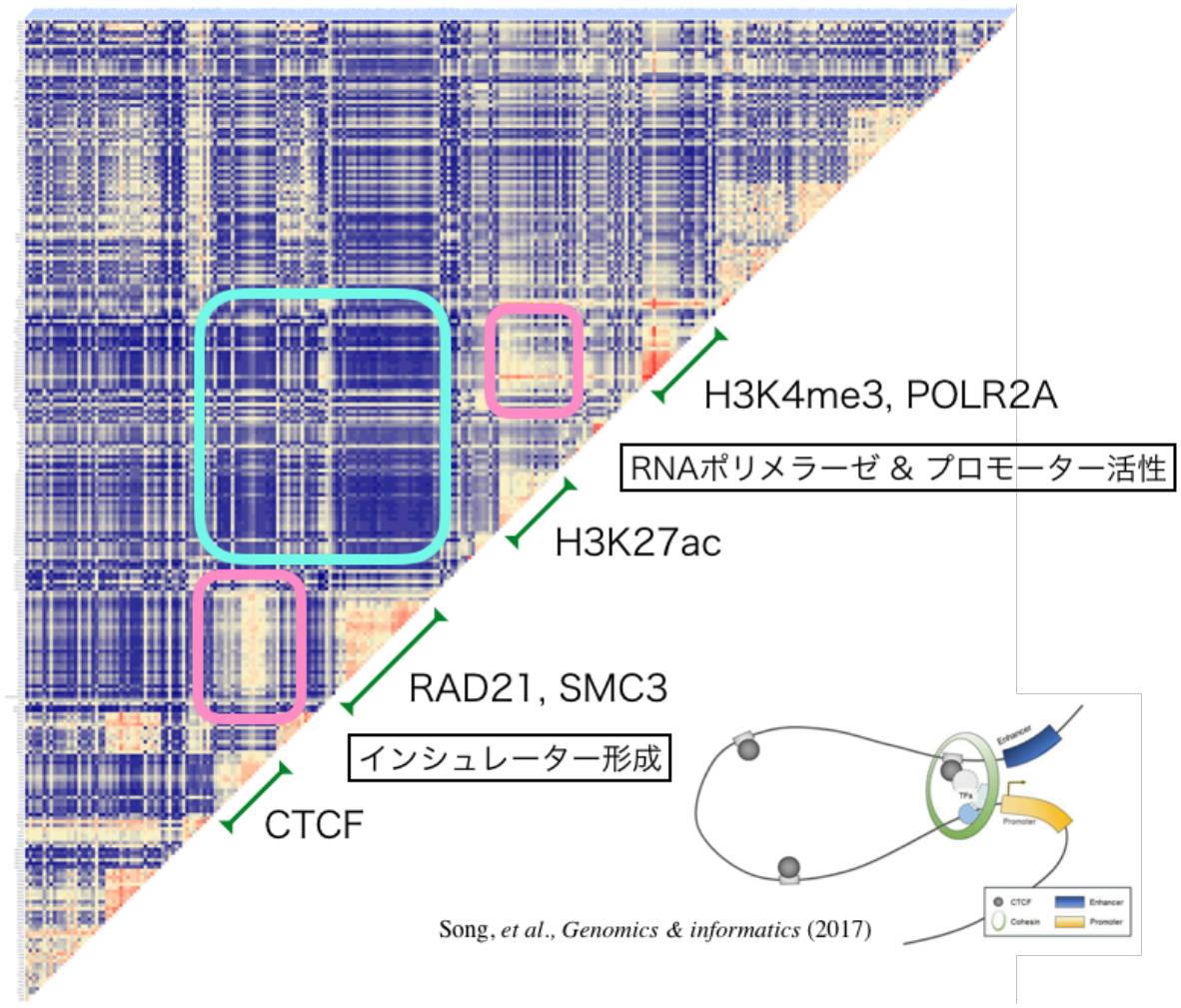


図 41: 転写活性化とインシュレーター形成に関与するクラスターとその関係
（図中のインシュレーター模式図は文献 [83] より引用）

かった。また転写活性に関わる *POLR2A* やヒストン修飾、インシュレーターを形成する *CTCF, RAD21, SMC3* のピークがトランスクリプトの転写開始点近傍に集中している様子も確認できる。興味深いことに *NR3C1* 自身のピークも含まれていた。

次に、これまでに得られた遺伝子群について互いのピークが観察できるかを確認した（図 44, 45, 46）。また、各遺伝子座におけるピークの有無を整理するため、これらの関係をネットワークで表したものを図 47 に示す。

結果、*CEBPB* や *HES2* にやや一方方向な制御関係が見られるものの、これらの遺伝子群はほとんど相互にピークが観測され、明確な一方向のカスケードは確認できなかった。これらの遺伝子の多くは、GR 自身も含めて [99]Pioneer factor や転写活性化因子として働く場合が多く、GR 特異的な制御関係が見えにくくなっていると考えられる。また、現段階では Gene Viewer が実験条件によるグルーピングに対応しておらず、実験条件の違いに起因するピークの変化が分かりにくくなっている。このことから、現段階の Gene Viewer は単一の遺伝子に対するピーク群の観察と考察には十分な機能があるものの、より体系的な制御関係の理解をもたらすためには、条件による絞り込みや比較、結合部位に対する具体的なアノテーション（上流・下流なのか、転写開始点な

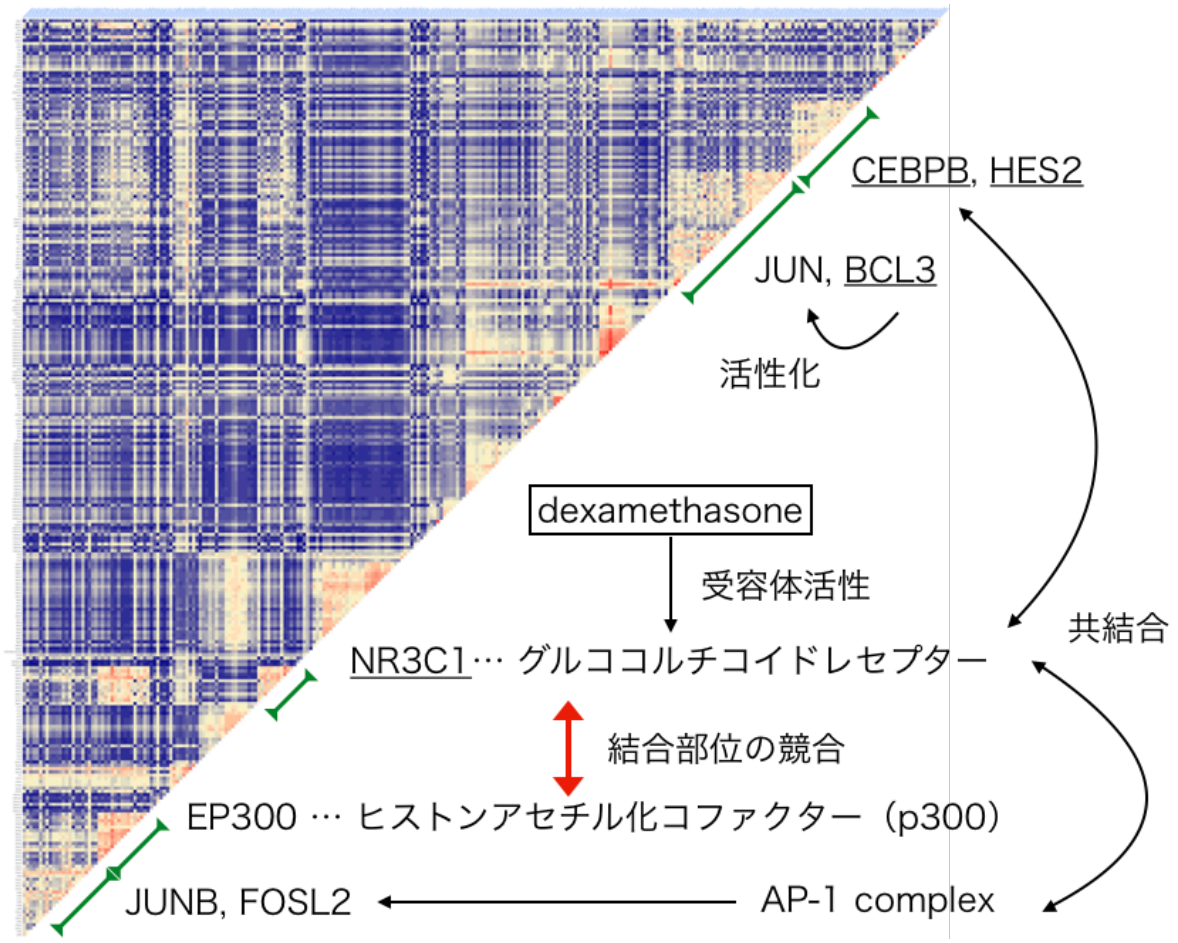


図 42: グルココルチコイド受容体に関与するクラスター群とその関係

のか等）といった機能を盛り込んだ上でネットワーク図の生成する必要があると考えられる。更に、ChIP-seq のデータのみではピークの有無が確認できるだけであり、結合した転写因子が転写を促進するのか抑制するのかといった情報を得ることは難しい。そのような制御関係の情報を得るためには、文献等によるアノテーションを盛り込んだり、RNA-seq のような発現量データ等と統合する必要があるだろう。

26 小括

第 II 部で述べた解析パイプラインを Docker によるコンテナ仮想化を用いて実装し、クラウドコンピューティング環境である AWS 上に展開した。また、実際に ENCODE Project の実験データを大規模に解析し、クラウドコンピューティングのスケーラビリティとコストパフォーマンスが確認された。得られたデータを可視化するデータベース Web アプリケーションとして C4S DB を開発し、Data Browser, Gene Viewer, Grobal Similarity の 3 機能を軸として、ChIP-seq データを多角的に検索・分析できるよう工夫した。また、デモンストレーションとして A549 細胞の ChIP-seq データセットについて考察を行った。結果、転写制御における制御因子間の関係や遺伝子単位での被制御関係について、既存の生物学的知見にも沿う関連性を見いだせることを確認で

図 43: *NR3C1* 遺伝子周辺のピークの可視化（一部のターゲットのみを図示）

きた。

この段階ではアプリケーションとしてのデータベースは必要最低限の実装しか完了しておらず、実用に堪えるためには検索機能等のさらなる実装と解析データの拡充が必要である。これらは今後改善していくと共に、日々増え続ける公共 ChIP-seq データも含めて更新していくことを目標としたい。

図 44: AP-1 関連遺伝子周辺のピーク（一部のターゲットのみを図示）

図 45: *CEBPB*, *HES2* 遺伝子周辺のピーク（一部のターゲットのみを図示）

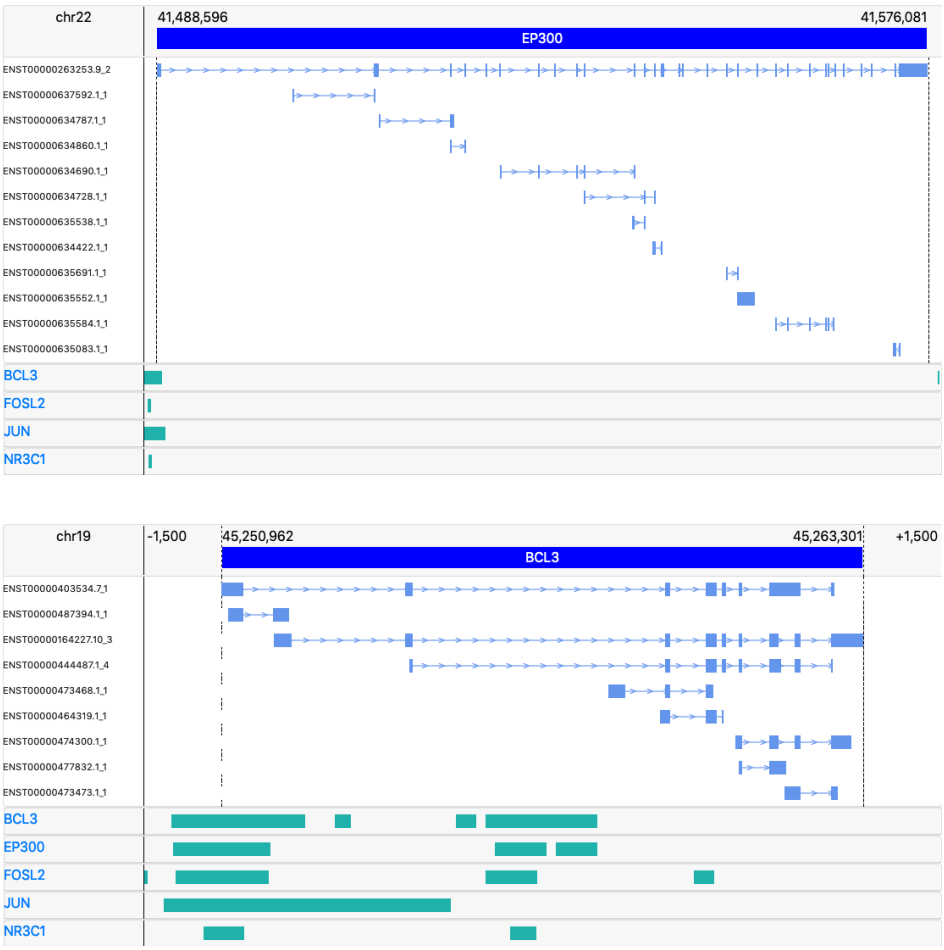


図 46: *EP300, BCL3* 遺伝子周辺のピーク（一部のターゲットのみを図示）

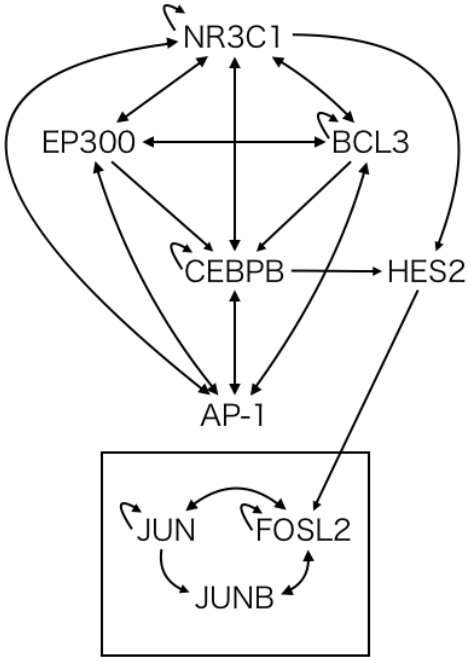


図 47: A549 細胞におけるグルココルチコイド受容体とその関連遺伝子の転写制御ネットワーク

AP-1 複合体については *FOSL2* に加えて *JUN* もしくは *JUNB* のどちらかが同時に観察された場合にエッジを張った

第 VI 部
総括

本論文では、転写制御機構の解明に寄与する公共データを用いた ChIP-seq データベースの開発を目標とした研究開発について下記の順で述べた。

1. 公共 ChIP-seq データ解析パイプラインの開発
   - FASTQ ファイルからピークコールまでを実施する解析パイプライン
   - クオリティコントロールの項目と実施方法
   - GEO のデータを自動処理するためのメタデータ解析手法の開発
   - デコイ配列入りリファレンスゲノムの ChIP-seq 解析への応用とその効果
2. Strand cross-correlation の理論的な特性評価
   - ChIP-seq リード分布のモデル化
   - NCC と MSCC を用いた理論的な最小値・最大値の導出
   - 効率的な NCC・MSCC の計算を可能にするソフトウェア PyMaSC の開発
   - シミュレーションデータと実データを用いた導出結果の検証
3. Strand cross-correlation を用いた新規指標の提案
   - 新規指標 VSN の提案
   - VSN の FRiP に対する相関とロバスト性について既存の手法との比較
4. ChIP-seq データベースの開発と転写制御解析
   - 解析パイプラインの Docker を用いたコンテナ仮想的
   - AWS を用いた解析パイプラインの展開と大規模解析
   - Django を用いたデータベース Web アプリケーション C4S DB の開発
   - A549 細胞 ChIP-seq データを用いた転写制御解析のデモンストレーション

これらの成果により、公共 ChIP-seq データの解析基盤を整えることができた。また、データベース化と公開についても現状は最小限の実装ながら軸となる機能を実現できた。以下では、データベース全般についてここまで述べなかった課題点および今後の展望について述べる。

公共 ChIP-seq データ解析パイプラインの開発

ChIP-seq に類似する実験への対応

ChIP-seq 法には、結合部位の解像度がより高い ChIP-exo 法 [100] や、クロマチンやオープンクロマチン領域の探索を目的とした MNase-seq[101], DNase-seq[102], FAIRE-seq[103], ATAC-seq[104] といった類似手法が確立されている。これらの実験から得られたデータはピークコール時のパラメーター調整等が必要になるものの、原理的には ChIP-seq 法と同じ手順で解析できる場合が多い。これらの実験に対応することで、より多くの転写制御に関する情報をデータベースに統合できるようになるだろう。

MACS2 以外のピークコーラーを用いた解析の実施

本研究で用いた解析パイプラインではピークの判定として MACS2 を使用した。MACS2 は ChIP-seq 解析において以前から用いられているデファクトスタンダードなソフトウェアの 1 つであり、MACS2 による解析結果は確実な需要が見込まれる。しかしながら、MACS2 はメンテナンスが継続されているものの初期バージョンのリリースから 9 年近くが経過しようとしており、原理的に新しいピークコールも開発されている [69]。また、MACS2 のピークコールは Narrow peak 向きの手法をベースの Broad peak のコールも行っており、数十 Kbp〜数 Mbp になることもある広い領域の判定に関しては、例えば隠れマルコフモデルを用いるステートベースの判定手法 [105,106] の方がより向いている可能性もある。今後より正確なピークの情報を提供するためにもピークコール手法のさらなる検討が必要になると考えられる。

自動検出できない GEO メタデータへの対応や SRA への拡充

GEO メタデータの自動検出手法により、現状 GEO に登録された ChIP-seq データのうち 84% のメタデータを処理することができた。とはいえ、母数が多いため 16% の実験も少ないとは言えず、また 84% の中にもいくらかの偽陽性が混ざっている可能性が高い。対策としては、より高度な固有表現抽出手法 [107] を用いる方法や、対応できないメタデータをマニュアルキュレーションできるよう UI を整備したり、間違ったアノテーションを発見したユーザーが容易に報告できるようデータベースの機能を工夫することが必要であると考えられる。

また、ラベルが信頼できるデータを十分量収集することができれば、今回の類似度マップのような手法を用いてメタデータが不十分なデータのラベルを予測するというアプローチを取ることも可能になるだろう。本研究ではまず信頼できるメタデータを有効活用するアプローチで解析を行ったが、ラベルに頼らずデータドリブンにデータセットを可視化することで想定しなかった情報が得られることも十分に有り得るため、今後検討していきたい。

ChIP-seq データベースの開発と転写制御解析

実験間類似度の計算手法の検討

本研究で実装した手法では、重複するピーク間で相関係数を計算し類似度とする手法を採用した。しかし、ピークの関係性としては転写開始点に対する、上流のエンハンサー領域のように、ある一定の距離だけ離れたピーク同士が関係する場合や、多数の因子とヘテロダイマーを形成するような転写因子では、協調する相手が自分の結合部位とどれくらい共起するかには非対称性がある。類似度の評価方法としては、例えば最近隣のピークまでの距離を考慮する手法 [108] もあるため、このような指標も組み合わせて提供することを検討する必要があるだろう。

他のデータベースやネットワークとの連携・統合

遺伝情報の変化と表現型の変化を対応付けることは、分子生物学における大きなテーマの 1 つである。セントラルドグマの起点に当たる遺伝子の転写を司るレイヤーとして、転写制御機構の情報はマルチオミクスな研究を展開する上で非常に重要であると考えられる。転写制御の情報は遺伝子発現のデータから必ずしも予測できるとは限らず、転写制御ネットワークと遺伝子発現量データがうまく一致しない例も報告されている [109]。遺伝子発現や遺伝子共発現データ [110] との統合やタンパク質間相互作用ネットワーク [111,112]、パスウェイ情報 [113,114] の導入により、さらなるデータドリブンな生物学的知見の発見を促していきたい。

引用文献

[1] WATSON JD, CRICK FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953 Apr;171(4356):737–738. Available from: https://doi.org/10.1038/171737a0.

[2] Crick FH. On protein synthesis. In: Symp Soc Exp Biol. vol. 12; 1958. p. 8.

[3] Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nature Genetics. 2003 Mar;33(3):245–254. Available from: https://doi.org/10.1038/ng1089.

[4] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell. 2007 May;129(4):823–837. Available from: https://doi.org/10.1016/j.cell.2007.05.009.

[5] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science. 2007;316(5830):1497–1502. Available from: https://science.sciencemag.org/content/316/5830/1497.

[6] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods. 2007 Aug;4(8):651–657. Available from: https://doi.org/10.1038/nmeth1068.

[7] Chèneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Research. 2017 11;46(D1):D267–D275. Available from: https://doi.org/10.1093/nar/gkx1092.

[8] Griffon A, Barbier Q, Dalino J, van Helden J, Spicuglia S, Ballester B. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Research. 2014 12;43(4):e27–e27. Available from: https://doi.org/10.1093/nar/gku1280.

[9] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep;489(7414):57–74. Available from: https://doi.org/10.1038/nature11247.

[10] Qin Q, Mei S, Wu Q, Sun H, Li L, Taing L, et al. ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. BMC Bioinformatics. 2016 Oct;17(1):404. Available from: https://doi.org/10.1186/s12859-016-1274-4.

[11] Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, et al. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO reports. 2018;19(12):e46255. Available from: https://www.embopress.org/doi/abs/10.15252/embr.201846255.

[12] Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: expanded

datasets and new tools for gene regulatory analysis. Nucleic Acids Research. 2018 11;47(D1):D729–D735. Available from: https://doi.org/10.1093/nar/gky1094.

[13] Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a database on gene transcription regulation-2019 update. Nucleic Acids Research. 2018 11;47(D1):D100–D105. Available from: https://doi.org/10.1093/nar/gky1128.

[14] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2007 11;36(suppl_1):D13–D21. Available from: https://doi.org/10.1093/nar/gkm1000.

[15] Shumway M, Cochrane G, Sugawara H. Archiving next generation sequencing data. Nucleic Acids Research. 2009 12;38(suppl_1):D870–D871. Available from: https://doi.org/10.1093/nar/gkp1078.

[16] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002 01;30(1):207–210. Available from: https://doi.org/10.1093/nar/30.1.207.

[17] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research. 2012 11;41(D1):D991–D995. Available from: https://doi.org/10.1093/nar/gks1193.

[18] The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306(5696):636–640. Available from: https://science.sciencemag.org/content/306/5696/636.

[19] Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Research. 2017 11;46(D1):D794–D801. Available from: https://doi.org/10.1093/nar/gkx1081.

[20] Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. Science. 2010;330(6012):1775–1787. Available from: https://science.sciencemag.org/content/330/6012/1775.

[21] , Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, et al. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science. 2010;330(6012):1787–1797. Available from: https://science.sciencemag.org/content/330/6012/1787.

[22] NHGRI. Genomics of Gene Regulation; 2017. Available from: https://www.genome.gov/Funded-Programs-Projects/Genomics-of-Gene-Regulation.

[23] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb;518(7539):317–330. Available from: https://doi.org/10.1038/nature14248.

[24] Andrews S, et al.. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[25] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17(1):10–12. Available from: http://journal.embnet.org/index.php/embnetjournal/article/view/200.

[26] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012 Apr;9(4):357–359. Available from: https://doi.org/10.1038/nmeth.1923.

[27] Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014 05;30(17):2503–2505. Available from: https://doi.org/10.1093/bioinformatics/btu314.

[28] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 06;25(16):2078–2079. Available from: https://doi.org/10.1093/bioinformatics/btp352.

[29] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008 Sep;9(9):R137. Available from: https://doi.org/10.1186/gb-2008-9-9-r137.

[30] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 01;26(6):841–842. Available from: https://doi.org/10.1093/bioinformatics/btq033.

[31] Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Scientific Reports. 2019 Jun;9(1):9354. Available from: https://doi.org/10.1038/s41598-019-45839-z.

[32] Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome research. 2012;22(9):1813–1831.

[33] Jung YL, Luquette LJ, Ho JWK, Ferrari F, Tolstorukov M, Minoda A, et al. Impact of sequencing depth in ChIP-seq experiments. Nucleic Acids Research. 2014 03;42(9):e74–e74. Available from: https://doi.org/10.1093/nar/gku178.

[34] Marinov GK, Kundaje A, Park PJ, Wold BJ. Large-Scale Quality Analysis of Published ChIP-seq Data. G3: Genes, Genomes, Genetics. 2014;4(2):209–223. Available from: https://www.g3journal.org/content/4/2/209.

[35] Ramachandran P, Palidwor GA, Porter CJ, Perkins TJ. MaSC: mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data. Bioinformatics. 2013 01;29(4):444–450. Available from: https://doi.org/10.1093/bioinformatics/btt001.

[36] Hong EL, Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, et al. Principles of metadata organization at the ENCODE data coordination center. Database. 2016 03;2016. Baw001. Available from: https://doi.org/10.1093/database/baw001.

[37] Federhen S. The NCBI Taxonomy database. Nucleic Acids Research. 2011 12;40(D1):D136–D143. Available from: https://doi.org/10.1093/nar/gkr1178.

[38] Wei CH, Leaman R, Lu Z. Beyond accuracy: creating interoperable and scalable text-mining web services. Bioinformatics. 2016 02;32(12):1907–1910. Available from: https://doi.org/10.1093/bioinformatics/btv760.

[39] Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Research. 2016 10;45(D1):D619–D625. Available from: https://doi.org/10.1093/nar/gkw1033.

[40] Bandrowski A, Astakhov V, Grethe J, Martone M. An antibody registry for biological sciences. Front Neuroinform Conference Abstract: 4th INCF Congress of Neuroinformatics. 2011;Available from: https://www.frontiersin.org/10.3389/conf.fninf.2011.08.00067/event_abstract.

[41] Bairoch A. The Cellosaurus, a Cell-Line Knowledge Resource. Journal of biomolecular techniques : JBT. 2018 Jul;29(2):25–38. 29805321[pmid]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29805321.

[42] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science. 2001;291(5507):1304–1351. Available from: https://science.sciencemag.org/content/291/5507/1304.

[43] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb;409(6822):860–921. Available from: https://doi.org/10.1038/35057062.

[44] Consortium IHGS. Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct;431(7011):931–945. Available from: https://doi.org/10.1038/nature03001.

[45] Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Research. 2017;27(5):849–864. Available from: http://genome.cshlp.org/content/27/5/849.abstract.

[46] Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253). Available from: https://science.sciencemag.org/content/349/6253/aab3761.

[47] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct;526(7571):75–81. Available from: https://doi.org/10.1038/nature15394.

[48] Durbin RM, Altshuler D, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct;467(7319):1061–1073. Available from: https://doi.org/10.1038/nature09534.

[49] McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov;491(7422):56–65. Available from: https://doi.org/10.

1038/nature11632.

[50] Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, et al. De novo assembly and phasing of a Korean human genome. Nature. 2016 Oct;538(7624):243–247. Available from: https://doi.org/10.1038/nature20098.

[51] Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, et al. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nature Communications. 2016 Nov;7(1):13637. Available from: https://doi.org/10.1038/ncomms13637.

[52] Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al. Long-read sequencing and de novo assembly of a Chinese genome. Nature Communications. 2016 Jun;7(1):12065. Available from: https://doi.org/10.1038/ncomms12065.

[53] Ameur A, Che H, Martin M, Bunikis I, Dahlberg J, Höijer I, et al. De Novo Assembly of Two Swedish Genomes Reveals Missing Segments from the Human GRCh38 Reference and Improves Variant Calling of Population-Scale Sequencing Data. Genes. 2018;9(10). Available from: https://www.mdpi.com/2073-4425/9/10/486.

[54] Du Z, Ma L, Qu H, Chen W, Zhang B, Lu X, et al. Whole Genome Analyses of Chinese Population and De Novo Assembly of A Northern Han Genome. Genomics, Proteomics & Bioinformatics. 2019;17(3):229 – 247. Available from: http://www.sciencedirect.com/science/article/pii/S1672022919301251.

[55] Nagasaki M, Kuroki Y, Shibata TF, Katsuoka F, Mimori T, Kawai Y, et al. Construction of JRG (Japanese reference genome) with single-molecule real-time sequencing. Human Genome Variation. 2019 Jun;6(1):27. Available from: https://doi.org/10.1038/s41439-019-0057-7.

[56] Simpson JT, Pop M. The Theory and Practice of Genome Sequence Assembly. Annual Review of Genomics and Human Genetics. 2015;16(1):153–172. PMID: 25939056. Available from: https://doi.org/10.1146/annurev-genom-090314-050032.

[57] Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014 06;30(20):2843–2851. Available from: https://doi.org/10.1093/bioinformatics/btu356.

[58] Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2012 04;14(2):178–192. Available from: https://doi.org/10.1093/bib/bbs017.

[59] Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Briefings in Bioinformatics. 2016 03;18(2):279–290. Available from: https://doi.org/10.1093/bib/bbw023.

[60] Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics. 2009 Sep;10:669. Review Article. Available from: https://doi.org/10.1038/nrg2641.

[61] Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing

methods for chromatin biology. Nature reviews Genetics. 2014 Nov;15(11):709–721. Available from: https://doi.org/10.1038/nrg3788.

[62] Zhang Q, Zeng X, Younkin S, Kawli T, Snyder MP, Keleş S. Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. BMC Bioinformatics. 2016 Feb;17(1):96. Available from: https://doi.org/10.1186/s12859-016-0957-1.

[63] Consortium EP, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.

[64] Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. Briefings in Bioinformatics. 2016 01;17(6):953–966. Available from: https://doi.org/10.1093/bib/bbv110.

[65] Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature Biotechnology. 2008 Nov;26:1351. Available from: https://doi.org/10.1038/nbt.1508.

[66] Nakato R, Shirahige K. Sensitive and robust assessment of ChIP-seq read distribution using a strand-shift profile. Bioinformatics. 2018 03;34(14):2356–2363. Available from: https://doi.org/10.1093/bioinformatics/bty137.

[67] Mammana A, Vingron M, Chung HR. Inferring nucleosome positions with their histone mark annotation from ChIP data. Bioinformatics. 2013 08;29(20):2547–2554. Available from: https://doi.org/10.1093/bioinformatics/btt449.

[68] Allhoff M, Seré K, Chauvistré H, Lin Q, Zenke M, Costa IG. Detecting differential peaks in ChIP-seq signals with ODIN. Bioinformatics. 2014 11;30(24):3467–3475. Available from: https://doi.org/10.1093/bioinformatics/btu722.

[69] Stanton KP, Jin J, Lederman RR, Weissman SM, Kluger Y. Ritornello: high fidelity control-free chromatin immunoprecipitation peak calling. Nucleic Acids Research. 2017 09;45(21):e173–e173. Available from: https://doi.org/10.1093/nar/gkx799.

[70] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome research. 2002;12(6):996–1006.

[71] Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, et al. Fast Computation and Applications of Genome Mappability. PLOS ONE. 2012 01;7(1):1–16. Available from: https://doi.org/10.1371/journal.pone.0030377.

[72] Fisher RA. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika. 1915;10(4):507–521. Available from: http://www.jstor.org/stable/2331838.

[73] Li H, Durbin R. Fast and accurate short read alignment with Burrows - Wheeler transform. Bioinformatics. 2009 05;25(14):1754–1760. Available from: https://doi.org/10.1093/bioinformatics/btp324.

[74] Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014 05;30(17):2503–2505. Available from: https:

//doi.org/10.1093/bioinformatics/btu314.

[75] Becker JS, Nicetto D, Zaret KS. H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. Trends in Genetics. 2016;32(1):29 – 41. Available from: http://www.sciencedirect.com/science/article/pii/S0168952515001961.

[76] Nishikori S, Hattori T, Fuchs SM, Yasui N, Wojcik J, Koide A, et al. Broad ranges of affinity and specificity of anti-histone antibodies revealed by a quantitative peptide immunoprecipitation assay. Journal of molecular biology. 2012 Dec;424(5):391–399. Available from: https://doi.org/10.1016/j.jmb.2012.09.022.

[77] Guo Y, Li J, Li CI, Long J, Samuels DC, Shyr Y. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics. 2012 Nov;13(1):666. Available from: https://doi.org/10.1186/1471-2164-13-666.

[78] da Veiga Leprevost F, Grüning BA, Alves Aflitos S, Röst HL, Uszkoreit J, Barsnes H, et al. BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics. 2017 03;33(16):2580–2582. Available from: https://doi.org/10.1093/bioinformatics/btx192.

[79] Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: A platform for interactive large-scale genome analysis. Genome Research. 2005;15(10):1451–1455. Available from: http://genome.cshlp.org/content/15/10/1451.abstract.

[80] ENCODE. We are thrilled to announce ENCODE as the latest AWS Public Data Set!; 2019. Available from: https://www.encodeproject.org/aws-public-dataset/.

[81] NCBI. SRA in the Cloud; 2019. Available from: https://www.ncbi.nlm.nih.gov/sra/docs/sra-cloud/.

[82] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research. 2012;22(9):1760–1774. Available from: http://genome.cshlp.org/content/22/9/1760.abstract.

[83] Song Sang-Hyun KTY. CTCF, Cohesin, and Chromatin in Human Cancer. Genomics Inform. 2017;15(4):114–122. Available from: http://genominfo.org/journal/view.php?number=493.

[84] Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, et al. Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. Cell. 2005;120(2):169 – 181. Available from: http://www.sciencedirect.com/science/article/pii/S0092867405000395.

[85] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics. 2007 Mar;39(3):311–318. Available from: https://doi.org/10.1038/ng1966.

[86] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combina-

torial patterns of histone acetylations and methylations in the human genome. Nature Genetics. 2008 Jul;40(7):897–903. Available from: https://doi.org/10.1038/ng.154.

[87] Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature. 2008 Feb;451(7180):796–801. Available from: https://doi.org/10.1038/nature06634.

[88] Michaelis C, Ciosk R, Nasmyth K. Cohesins: Chromosomal Proteins that Prevent Premature Separation of Sister Chromatids. Cell. 1997;91(1):35 – 45. Available from: http://www.sciencedirect.com/science/article/pii/S0092867401800076.

[89] Guacci V, Koshland D, Strunnikov A. A Direct Link between Sister Chromatid Cohesion and Chromosome Condensation Revealed through the Analysis of MCD1 in S. cerevisiae. Cell. 1997;91(1):47 – 57. Available from: http://www.sciencedirect.com/science/article/pii/S0092867401800088.

[90] Losada A, Hirano M, Hirano T. Identification of Xenopus SMC protein complexes required for sister chromatid cohesion. Genes & Development. 1998;12(13):1986–1997. Available from: http://genesdev.cshlp.org/content/12/13/1986.abstract.

[91] Weikum ER, Knuesel MT, Ortlund EA, Yamamoto KR. Glucocorticoid receptor control of transcription: precision and plasticity via allostery. Nature Reviews Molecular Cell Biology. 2017 Mar;18(3):159–174. Available from: https://doi.org/10.1038/nrm.2016.152.

[92] Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, Garabedian MJ, et al. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. Genome Research. 2009;19(12):2163–2171. Available from: http://genome.cshlp.org/content/19/12/2163.abstract.

[93] Hess J, Angel P, Schorpp-Kistner M. AP-1 subunits: quarrel and harmony among siblings. Journal of Cell Science. 2004;117(25):5965–5973. Available from: https://jcs.biologists.org/content/117/25/5965.

[94] Na SY, Choi JE, Kim HJ, Jhun BH, Lee YC, Lee JW. Bcl3, an I κ B Protein, Stimulates Activating Protein-1 Transactivation and Cellular Proliferation. Journal of Biological Chemistry. 1999;274(40):28491–28496. Available from: http://www.jbc.org/content/274/40/28491.abstract.

[95] Grøntved L, John S, Baek S, Liu Y, Buckley JR, Vinson C, et al. C/EBP maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements. The EMBO Journal. 2013;32(11):1568–1583. Available from: https://www.embopress.org/doi/abs/10.1038/emboj.2013.106.

[96] Biddie SC, John S, Sabo PJ, Thurman RE, Johnson TA, Schiltz RL, et al. Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding. Molecular Cell. 2011;43(1):145 – 155. Available from: http://www.sciencedirect.com/science/article/pii/S1097276511004606.

[97] Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene

expression. Genes & Development. 2011;25(21):2227–2241. Available from: http://genesdev.cshlp.org/content/25/21/2227.abstract.

[98] Almawi WY, Melemedjian OK. Molecular mechanisms of glucocorticoid antiproliferative effects: antagonism of transcription factor activity by glucocorticoid receptor. Journal of Leukocyte Biology. 2002;71(1):9–15. Available from: https://jlb.onlinelibrary.wiley.com/doi/abs/10.1189/jlb.71.1.9.

[99] John S, Sabo PJ, Johnson TA, Sung MH, Biddie SC, Lightman SL, et al. Interaction of the Glucocorticoid Receptor with the Chromatin Landscape. Molecular Cell. 2008;29(5):611 – 624. Available from: http://www.sciencedirect.com/science/article/pii/S1097276508001299.

[100] Rhee HS, Pugh BF. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell. 2011 Dec;147(6):1408–1419. Available from: https://doi.org/10.1016/j.cell.2011.11.013.

[101] Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic Regulation of Nucleosome Positioning in the Human Genome. Cell. 2008;132(5):887 – 898. Available from: http://www.sciencedirect.com/science/article/pii/S0092867408002705.

[102] Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell. 2008;132(2):311 – 322. Available from: http://www.sciencedirect.com/science/article/pii/S0092867407016133.

[103] Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, et al. A map of open chromatin in human pancreatic islets. Nature Genetics. 2010 Mar;42(3):255–259. Available from: https://doi.org/10.1038/ng.530.

[104] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods. 2013 Dec;10(12):1213–1218. Available from: https://doi.org/10.1038/nmeth.2688.

[105] Xu H, Wei CL, Lin F, Sung WK. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics. 2008 07;24(20):2344–2349. Available from: https://doi.org/10.1093/bioinformatics/btn402.

[106] Heinig M, Colomé-Tatché M, Taudt A, Rintisch C, Schafer S, Pravenec M, et al. histoneHMM: Differential analysis of histone modifications with broad genomic footprints. BMC Bioinformatics. 2015 Feb;16(1):60. Available from: https://doi.org/10.1186/s12859-015-0491-6.

[107] Wei CH, Kao HY, lu Z. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. BioMed research international. 2015 09;2015:918710.

[108] Favorov A, Mularoni L, Cope LM, Medvedeva Y, Mironov AA, Makeev VJ, et al. Exploring Massive, Genome Scale Datasets with the GenometriCorr Package. PLOS Com-

putational Biology. 2012 05;8(5):1–12. Available from: https://doi.org/10.1371/journal.pcbi.1002529.

[109] Larsen SJ, Röttger R, Schmidt HH, Baumbach J. E. coli gene regulatory networks are inconsistent with gene expression data. Nucleic Acids Research. 2018 11;47(1):85–92. Available from: https://doi.org/10.1093/nar/gky1176.

[110] Obayashi T, Kagaya Y, Aoki Y, Tadaka S, Kinoshita K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. Nucleic Acids Research. 2018 11;47(D1):D55–D62. Available from: https://doi.org/10.1093/nar/gky1155.

[111] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein - protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research. 2018 11;47(D1):D607–D613. Available from: https://doi.org/10.1093/nar/gky1131.

[112] Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Research. 2018 11;47(D1):D529–D541. Available from: https://doi.org/10.1093/nar/gky1079.

[113] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000 01;28(1):27–30. Available from: https://doi.org/10.1093/nar/28.1.27.

[114] Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Research. 2019 11;48(D1):D498–D503. Available from: https://doi.org/10.1093/nar/gkz1031.

研究成果発表等

投稿論文

- Hayato Anzawa, Hitoshi Yamagata and Kengo Kinoshita., "Theoretical characterisation of strand cross-correlation in ChIP-seq", BMC Bioinformatics （査読中）
  プレプリントは Research Square にて公開中: 30 October 2019, PREPRINT (Version 1) available at Research Square https://doi.org/10.21203/rs.2.16602/v1
- Keito Okazaki, Hayato Anzawa, Zun Liu, Nao Ota, Hiroshi Kitamura, Yoshiaki Onodera, Md. Morshedul Alam, Daisuke Matsumaru, Takuma Suzuki, Fumiki Katsuoka, Shu Takada, Ikuko Motoike, Mika Watanabe, Akira Sakurada, Yoshinori Okada, Masayuki Yamamoto, Takashi Suzuki, Kengo Kinoshita, Hiroki Sekine and Hozumi Motohashi., "Enhancer Remodeling at the NOTCH3 Locus Licenses NRF2 for the Promotion of Tumor-Initiating Activity in Non-Small Cell Lung Cancers" 投稿準備中

ポスター発表

- Hayato Anzawa, Hitoshi Yamagata and Kengo Kinoshita, "Theoretical estimation of the strand cross-correlation in ChIP-Seq data", ISMB/ECCB 2019, 2019 年 7 月, Basel (Switzerland)
- Hayato Anzawa and Kengo Kinoshita, "Effects that Decoy Sequences Bring to ChIP-Seq Analysis", International Kick-off Symposium of Graduate Program in Data Science, 2018 年 2 月, 東北大学
- 安澤 隼人, 木下 賢吾, "ChIP-Seq データのクラスタリングによる実験条件・解析手法に起因するバイアスの可視化", NGS 現場の会第五回研究会, 2017 年 4 月, 仙台
- Hayato Anzawa and Kengo Kinoshita, "Model based discrimination method of ChIPed data from control data in ChIP-seq experiment dataset", 第五回生命医薬情報学連合大会, 2016 年 9-10 月, 東京

他

- PyMaSC - Python implementation to calculate mappability-sensitive cross-correlation for fragment length estimation and quality control for ChIP-Seq. https://pypi.org/project/PyMaSC/, 最終更新日 2019 年 11 月 4 日

謝辞

本研究を遂行し本論文をまとめるにあたり、指導教官であった木下賢吾教授に深く御礼申し上げます。木下先生からは数多くのご指導・ご助言をいただきました。また大林武先生、西羽美先生をはじめとする木下・大林・西研究室の先生方、先輩・後輩も含めた学生の皆様にも御礼申し上げます。皆様のおかげで恵まれた環境で研究を行うことができました。

東北大学加齢医学研究所 加齢制御研究部門 遺伝子発現制御分野の本橋ほづみ教授をはじめとする皆様、ならびに同 腫瘍制御研究部門 分子腫瘍学研究分野の田中耕三教授をはじめとする皆様に御礼申し上げます。先生方との共同研究を通じて、様々な NGS データの解析や生物実験を行う研究者とディスカッションする貴重な機会を得ることができました。

また、本研究をサポートして頂いたキヤノンメディカルシステムズ株式会 社先端研究所 藤田博之所長ならびに山形仁フェローをはじめとする皆様に御礼申し上げます。