AJACS「Hi-C解析を知って・学んで・使う」(2025年1月16日) 質疑応答
「Hi-C解析を知る」「Hi-Cデータを使う」(国立遺伝学研究所 東 光一)

No.	質問	回答
1	Hi-Cのシーケンシングで、ペアエンドの片側のアダプター配列を間違えてしまい、Read1だけで次世代シーケンサーの結果の解析をしてもらいました(つまり、片側だけで解析されたfastqが得られております)。この場合、一応Read2側も得られているのですが、Hi-C解析を行ってもよいでしょうか？	基本的には取得しなおした方がいいと思います。どのレベルで間違っちゃったのかにもよると思うんですが、そもそもプライマーがくっつけないならシーケンス反応が進行しないのでリードが生成されません。リードが得られていても、FastQCなどで見てみるとクオリティがきわめて低いかもしれません。クオリティがまともに見えても、リファレンスへのマッピング率がきわめて低いかもしれません。解析の各ステップで、まともそうな結果が得られているか確認が必要になりますが、どこまでいっても不安がつきまとうので、やっぱり取得しなおした方がいいと思います。
2	行列内の一定範囲が数値が高いので塗りつぶし上に色が濃くなるということでしょうか？	はい。コンタクトマップは単に行列データのヒートマップなので、数値が比較的高い要素がまとまったブロックが存在する、ということになります。
3	格子状に濃く見えるのはなぜでしょうか？	ゲノムがコンパートメントと呼ばれる領域に区分され、それらが交互に配置されているためです。同じコンパートメント同士(A-A, B-B)は互いの接触頻度が高く、異なるコンパートメント同士(A-B)は低いため、コンタクトマップ上で規則的な濃淡が繰り返されます。つまりコンパートメントがゲノム全体で交互に配置されているために、全ゲノムスケールの市松模様が観察されるのです。
4	インシュレーターとはなんですか？	インシュレーターとは、ゲノム上の連続した2つの領域で、それぞれ異なるエピジェネティックな制御が生じている場合に、それらの影響が互いに及ぶのを遮断する機能を持つ配列です。インシュレーターとして機能する例の1つに、CTCFタンパク質の結合配列がありますが、CTCF結合配列がすべてインシュレーターとして働くわけではありません。
5	TADとcompartmentの違いは何でしょうか？ 単にスケールの違いですか？	領域サイズのスケールというよりも、解析スケールの違いといったほうが正確です。コンパートメントはゲノム全体を次元圧縮したときに見えてくる大域的なパターンで、TADはローカルに接触頻度が高い領域が密集して形成された構造です。コンパートメントは通常、TADよりも大きなスケールで観測されますが、TAD境界とコンパートメント境界が一致する場合もあります。
6	TAD内ではそれぞれがなんとなくくっつきあってるから全体として色が濃くなるというような理解で良いのでしょうか、点というよりも幅でしょうか？	特定のピーク(ループ)や階層性が見られないTADの場合、ふたつ、解釈の可能性があると思います。 1) 個々の細胞においてその領域全体がなんとなくまとまっている可能性。特定の接触相手を持たず、領域全体としてぐちゃぐちゃっとまとまっているイメージ。 2) 個々の細胞でみると、それぞれ特定の位置でループを形成しているが、ループの位置が一定しておらず、細胞ごとに領域内のさまざまな位置関係でループが形成されている場合。 バルクHi-Cで得られる結果は実験に使用した細胞ポピュレーションの平均的な構造なので、領域の中が全体的に高い接触頻度で観測されることになります。基本的には前者の構造としてイメージされることが多いですが、後者の可能性は否定できません。近年のシングルセルHi-C関連の研究を考えると、複数の異なるループ形成のアンサンブルとしてTADパターンが平均的に観測されている可能性のほうが高いかもしれません。TAD境界は比較的安定だが、TAD内のループ位置は動的で安定しない、という議論もあります。
7	「A/Bコンパートメントだけでなく、TADも発生や分化で大きく変化しない」と聞いたことがある一方、ES細胞の発生過程でTADレベルで変化があるという論文を見かけました。細胞種間での比較解析しており、どのレベルで見るべきか悩んでいます。	安定か、動的か、については、研究/生物種/細胞型/ゲノム上の領域による、としか言えないかもしれません。ゲノム全体を俯瞰したとき、大域的に見て、種間・細胞間でTADやコンパートメント(の大半)がそれほど変化しない、という表現は間違っていないと思います。一方で、発生・分化において特定のTAD(HOXクラスタ周辺など)がダイナミックに再編成されることも観測されています。変化が激しいと見るか、安定していると見るかは、研究目的や着目している現象のスケールによります。このスケールならなんらかの変化が見えるはず、と断言できないので、関心領域についてさまざまなスケール(コンタクトマップ解像度)で調べてみるしかないと思います。解像度を上げていけばどこかの段階で必ず差は出てきてしまうものなので、それがノイズに駆動された結果ではないことをレプリケイトなどで調べることも重要です。
8	PCA analysisで得られるPC1 と PC2以上の要素が何を意味するものなのかわかりません。PC3、4、5について教えてください。	主成分分析(PCA)の数学的解釈は、Hi-C解析でも通常のPCAと同じです。コンタクトマップのPCAでは、各ゲノム領域(ビン)の接触パターンを比較し、ばらつき(分散)を最大限にとらえる軸が第一主成分(PC1)、それと直交する軸が第二主成分(PC2)として定義されます。PC1が二極化する場合、それはゲノム全体が大きく2つのクラスタ(A/Bコンパートメントなど)に分かれていることを示します。PC2は、PC1による変動を無視した場合のパターンを示し、さらに微細な構造(サブコンパートメントなど)を反映します。PC3、PC4、PC5などの高次の主成分は、上位の主成分で説明されない、より細かい特徴を捉えます(が、生物学的に解釈することは難しい場合も多いです)。これらの高次の主成分は、サブコンパートメントのクラスタリングに利用されることがあります。注意点として、PC1が必ずA/Bコンパートメントを反映するとは限りません。一部の生物種では、PC1が単に染色体の短腕・長腕の構造を反映することがあり、その場合は、コンパートメントの計算にPC2が使用されることがあります。
9	コンパートメントという結構大きな領域レベルで活性化と抑制が分かれているということは、遺伝子やエンハンサー単位で見るのは細かすぎるでしょうか？	解析したい現象のスケールによると思います。コンパートメントは、どの領域がエンハンサー・プロモーターの相互作用を活発に行っている可能性が高いか、といった情報を適用してくれますが、具体的にどのエンハンサーがどのプロモーターと相互作用しているのかはわかりません。コンパートメントは粗い状態の区分、TADは制御範囲、なので得られる情報は違います。
10	前半部43ページの上のDの図ですが、コンパートメントBは眠ったままでAは時間の推移でくっついて転写が始まるという理解で良いのでしょうか？	Bコンパートメントでも、遺伝子発現の前に凝縮状態からの緩和が生じるらしいです。
11	オスとメスの染色体でY染色体の有無の違いがありますが、どちらを使うべきでしょうか？ 意識する必要はありますか？	特に性染色体に注目した研究ではなく、議論の焦点が常染色体にあるのならば、どちらを使っても大きく影響しないと思います。もちろん性染色体が研究の主題なら意識する必要がありますが、X染色体は不活性化の影響で、Y染色体はリピート配列の影響で、一般的には解析が難しい傾向にあります。
12	同じ細胞であってもその状態は多様だとは思いますが、細胞株などでHi-Cをやる場合は細胞周期を同期させるといった工夫が必要でしょうか？	細胞周期による構造の違い、とくに有糸分裂期の構造が顕著なので、同期している方が理想的ではあります。ただ実際には、M期細胞の割合が集団中でそれほど多くないと想定できる、あるいは薬剤処理によるアーティファクトのほうが気になる、などの理由で同期させずにHi-C解析をすることも少なくありません。解析の解像度が落ちる(微細な違いが見えにくくなる)デメリットはあるものの、そのぶん大量の細胞を使ったり、大規模にシーケンシングができれば、G1からG2で比較的安定している構造はちゃんと見えます。これもやはり研究目的やサンプルの状態によります。
13	公共データのHi-Cを使う場合セルラインだと他のRNA-seqとかの照らし合わせもしやすいとなりますでしょうか？ 複数種類のデータを見比べたいという場合におすすめのデータベース、セット、ツール等お伺いすることは出来ますでしょうか？(特にがん細胞)	4DNucleomeやENCODEなどのコンソーシアムに由来するデータだと、統一的なサンプリングやパイプラインによる情報解析で複数種類のモダリティ(Hi-C、RNA-seq、ChIP-seq、ATAC-seqなど)のデータがとられているので、それらを合わせて見比べやすいと思います。また、コンタクトマップそのものではなく、TAD境界やコンパートメントなど二次的な解析で得られたデータを見るだけなら、それぞれBED形式やBW形式などで配布されているので、IGVなど広く使われるビジュアライザで他のオミクスデータと簡単に比較できます。

No.	質問	回答
14	Hi-Cの結果はコンタクトマップからゲノムの三次元構造を推測することになるかと思いますが、実際に三次元モデル化するようなパッケージはあるのでしょうか？ それとも、そのような三次元モデル化は必要とされていないのでしょうか？	いくつかツールがあります。が、そこまで一般的に広く使われているわけではありません。視覚的にわかりやすい以上のメリットが生物学者にとって見出しにくい、推定や解釈にポリマー物理に関連した知識が必要となる場合もある、そもそもバルクHi-Cの結果は多数の細胞のアンサンブルであり構造が一意に定まるとは限らない、などの理由のためだと思います。
15	シングルセルHi-Cでは、細胞型と細胞状態(例: 細胞周期)とでどちらのほうが多様性が出ますか？	どういった集団について比較するかに依存します。以下、感覚的な話で、定量的に評価した研究は不勉強ながら存じません。まず、M期を含む細胞集団を考えると、細胞周期の違いが圧倒すると思います。有糸分裂による染色体凝縮の構造変化が顕著であるためです。しかしM期以外の細胞に関しては、細胞型による構造の違いがG1,S,G2の違いよりも多様であると思います。実際に多様な細胞型についてシングルセルHi-Cによって構造パターンの違いによるクラスタリングが細胞型と一致している報告があります。
16	例えばcool/hicのどちらかの形式しかなかった場合に、個人でシーケンスデータをダウンロードしなおして、もう片方のデータを自分でサーバー上等で作り直すことはツールを用意すれば可能でしょうか？	可能です。HiCExplorerにもhicConvertFormatというツールが付属しています。
17	HDF5のバイナリデータでHi-Cが提供されているのを見たのですが、これの用途は複数の解像度を1つのファイルに格納するためなどでしょうか？	たぶんそうだと思います。HiCExplorerの出力もHDF5です。cool/mcoolも実はHDF5です。なので、h5pyなど使えばmcoolから自分で特定の解像度の行列を取り出したりできます。JuicerのDotHiC(.hic)に関してはHDF5ではなくて、独自形式のバイナリで複数解像度をひとつのファイルにまとめてます。
18	実際の使用感として、nf-coreとnextflowはシームレスに併用できるでしょうか？ それともnf-coreは単体で使う方が良いでしょうか？ nf-coreは使っています。	「nf-coreのパイプラインは基本的にnextflowで記述されている。nf-coreのパイプラインを利用する場合はnextflowの使用が前提になる。」と講義中には答えたのですが、たぶんご質問の意図は、nf-coreのパイプラインと他のnextflow記述パイプラインを組み合わせて使ってもいいか、ということだったでしょうか。もちろん、単体で使うことを強制されるわけではないので、nf-coreパイプラインのnfファイルを編集すればパイプラインを拡張できますし、新たなブロックを定義したり、別のnfと接続させることも可能です。単体で使うメリットは、楽ちんってこと以外だと、論文のMethod書くときに、nf-coreのパイプライン名・バージョン・オプションを書けば済む、というのもあります。
19	scRNA-SeqにおけるCell RangerやSeurat/ScanPyに相当するデファクトスタンダードな解析ワークフローは、シングルセルHi-Cにはまだ無いですか？	デファクトと言えるようなパイプラインは無いんじゃないかと思います。scHicExplorerやSnapHicがあるみたいですが、使ったことないので使用感はよくわかりません。
20	どこがキメラの境界かわからなくて片側から張り付けるということは、張り付く幅をとって張り付けるイメージでしょうか？ 半分以上が別の配列となると、マッパーから除外されてしまいそうですが、それを許容する設定をbwaに入れるというような理解で良いのでしょうか？(not27ページの手法)	その通りです。R1/R2いずれについても、5'側の端っこでアラインメントをとります。おっしゃる通り、リード単位のグローバルアラインメントをとるとヒット扱いにならないスコアになってしまうので、お示ししたbwaではローカルアラインメントがとりやすくなるスコア設定(3'側でヒットしない連続塩基のミスマッチスコアを無視)にしています。あるいは3'側をトリムして5'側の短い領域だけでアラインメント計算したりします。
21	inter chromosomalは2割程度ならそれは実際のものという理解になるのでしょうか？ それ以上出てきた場合は誤検出という判断でしょうか？	実際に存在するコンタクトもあると思うんですが、よくわかりません。少なくとも、アーティファクトのランダムライゲーションは大部分がInter-chromosomalになるので、それとの区別がすごく難しくなります。がん細胞など構造変異が想定されるサンプルを別として、通常の細胞で染色体間のtrans contactはほとんど議論されないと思います。
22	normalizationの話がありましたが、二つのサンプル間でコンタクトがどのように変化したか議論したい時に、normalization方法で一般的なやり方を教えてください (以前、サンプルAとサンプルBで、そもそもマッピングされたリードの数が違ったために比較しにくかった経験があります)。	コンタクトマップの正規化はサンプルごとに実行されるので、通常、正規化後であっても、サンプル間で直接コンタクトマップの数値そのものは比較できません。おっしゃる通り、シーケンスデプスがサンプルごとに異なるためです。紹介した正規化手法によって補正されるのはあくまで、同一サンプルにおけるゲノム上の領域ごとのバイアスです。複数サンプルのコンタクトマップを直接的に扱って差分解析するツールもありますが、そこまで広く使われているわけではなく、サンプル間比較の場合はやはり、サンプルごとに個別に二次的な解析(TAD・ループ検出など)を実行して、それらの結果を比較するのが一般的だと思います。もちろんその場合でも、リファレンスゲノムのバージョンやコンタクトマップの解像度(ビンサイズ)は揃っている必要があります。
23	ピークやTADのコーラーはたくさんあるようですが、A/Bコンパートメントは手法としてはPCAだけですか？	「その通り。コンパートメントの定義自体がPCAに基づくため。ツールによって実装に違いはあれど計算はPCA」と、講義中には答えてしまったんですが、サブコンパートメントを考えると事情はちょっと違ってきます。A1/A2/B1/B2/B3など、微細なパターンの違いを捉える場合は、第一主成分以外の成分を使ったり、他のエピゲノムデータを併用することもあります。手法も、クラスタリング、HMMのモデル推定、グラフ分割、場合によっては既知のコンパートメント割り当てを教師とした機械学習モデルを使って推定する場合もあります。
24	GEOデータベースにマッピング後のデータとしてallValidPairs.txtという形式のファイルしか存在しませんでした。こちらの形式からcoolファイルなどに変換する方法はありますか？	ファイル名だけからデータ形式を推測するのは危ないかもですが、たぶんHiC-Proで計算した途中段階のファイル(コンタクトマップ構築の直前の段階)じゃないかと思います。このあたりの議論(https://github.com/nservant/HiC-Pro/issues/619)を参考にpairsに書き換えて、cooler cload すればcool形式のコンタクトマップが生成できると思います。
25	以前HiC explorerでA/Bコンパートメントの解析を行ったところ、染色体全部がAとかBになるような結果になってしまいました。これはどういう解釈すればよいですか？	本当に、巨大なスケールの凝縮が生じている、という可能性はあります。が、ありがちなパターンとして、(染色体ごとではなく)全ゲノムでコンパートメント解析を実行したことによる技術的なアーティファクト、という可能性が高い気がします。コンタクトマップの正規化がうまくいっていなかったり、染色体ごとに得られた接触頻度にばらつき・バイアスがあると、コンタクトパターンの類似性が染色体ごとのばらつきに強くひっぱられることがあります。全体的に極端に接触頻度が低い染色体が存在する場合など。そうすると、ゲノム全体の主成分分析の結果得られた第一主成分が染色体間差を捉えてしまいます。基本的には、染色体ごとに個別に計算したほうがいいと思います。その場合、染色体それぞれでどちらの符号(+/-)をA/Bに割り当てるかが変わってくる点には注意が必要です。
26	主成分解析の結果で正負の値が算出されると思いますが、その値の大きさを比較することは問題ないでしょうか？(例えば 正の値が大きければよりオープンなクロマチンであるなど)	いえ、それは危険だと思います。PC1の値は単にパターンの類似性の相対的な評価なので、アクセシビリティのような生物学的な解釈と直接的に関連するものではありません。

No.	質問	回答
27	Cre/loxPシステムなどでdeletion knockoutして、エクソンレベルで欠損した場合、WTとKDの比較などは正当に評価できるでしょうか？ WTとKDの比較評価がどのように行われていることが多いかを知りたいです。	ほとんどのゲノム領域については通常通りの比較評価(TADパターンやループの変化)が可能だと思いますが、欠損した領域周辺については注意が必要になるかもしれません。リファレンスゲノムと異なるため、deletionをまたぐようなWGSリードがコンタクトと判断されて頻繁に観測され、周辺のTAD推定などが不安定になるかもしれません。とはいえエクソンひとつや数百bp程度の小さな欠損なら技術的アーティファクトは局所的だと思うので、解釈に大きく影響しないと思います。
28	coolpup.pyでもTAD内部のintesnsity(TAD内でどれだけ相互作用しているか？)を計算できる認識でよいですか？	cooltoolsやcoolpup.pyを使った解析、つまりピークや領域のAggregate解析ですが、基本的にはその認識で正しいです。ただ、検出されたピークや領域をコンタクトマップレベルで全部まとめて平均化した表現であることに注意は必要です。
29	nf-coreをこれまで使用したことがないのですが、使い方のわかりやすいサイトや本などはありますか？	たぶんnf-coreの公式サイトを読むのが一番わかりやすいと思います。パイプラインそれぞれで使い方が違うので。