

平成29年度 NGSハンズオン講習会 Hi-C解析

2017年9月1日
国立遺伝学研究所 東光一

本講義の内容

- Hi-C解析とは
 - Chromosome Conformation Capture の原理
 - Hi-Cで何がわかるか？コンタクトマップの見方
- Hi-C解析の流れ（実習と並行）
 - Hi-C解析のツール
 - マッピング
 - フィルタリング
 - 正規化
 - ピーク検出、TAD検出など
 - 3D構造モデリング

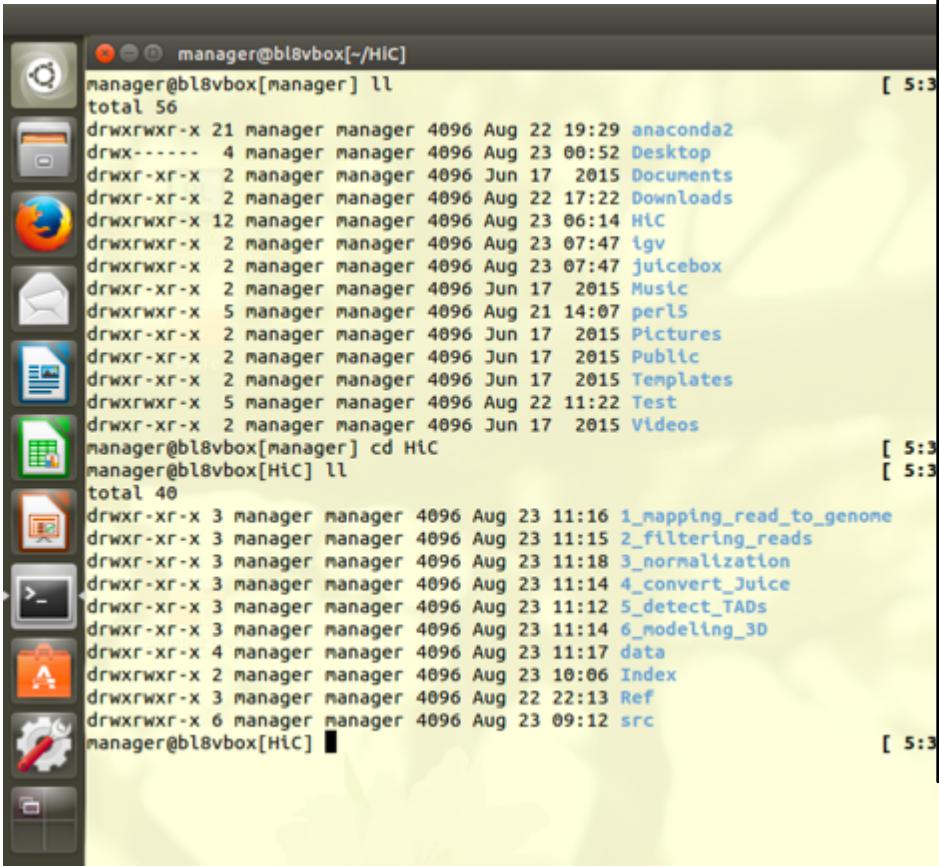
実習の内容

- 目的：
Fastqファイル（公共データ）からスタートして、
Hi-C解析論文でよく見るコンタクトマップや、
染色体3次元構造の構築までやってみる
- 複雑なコマンドは打ちません。
必要な操作はすべてpythonスクリプトにまとめてあるので、`python ○○.py` と打つだけ。

「ツールの使い方」よりも、Hi-Cデータの特徴や、データ解析で気をつけなければいけない点を理解する。

実習の内容

Bio-Linux-8.0.7_hm_kh.ova を起動。
すべて、~/HiC の中で実行する。



```
manager@bl8vbox[~/HiC]
manager@bl8vbox[manager] ll
total 56
drwxrwxr-x 21 manager manager 4096 Aug 22 19:29 anaconda2
drwx----- 4 manager manager 4096 Aug 23 00:52 Desktop
drwxr-xr-x 2 manager manager 4096 Jun 17 2015 Documents
drwxr-xr-x 2 manager manager 4096 Aug 22 17:22 Downloads
drwxrwxr-x 12 manager manager 4096 Aug 23 06:14 HiC
drwxrwxr-x 2 manager manager 4096 Aug 23 07:47 igv
drwxrwxr-x 2 manager manager 4096 Aug 23 07:47 juicebox
drwxr-xr-x 2 manager manager 4096 Jun 17 2015 Music
drwxrwxr-x 5 manager manager 4096 Aug 21 14:07 perl5
drwxr-xr-x 2 manager manager 4096 Jun 17 2015 Pictures
drwxr-xr-x 2 manager manager 4096 Jun 17 2015 Public
drwxr-xr-x 2 manager manager 4096 Jun 17 2015 Templates
drwxrwxr-x 5 manager manager 4096 Aug 22 11:22 Test
drwxr-xr-x 2 manager manager 4096 Jun 17 2015 Videos
manager@bl8vbox[manager] cd HiC
manager@bl8vbox[HiC] ll
total 40
drwxr-xr-x 3 manager manager 4096 Aug 23 11:16 1_mapping_read_to_genome
drwxr-xr-x 3 manager manager 4096 Aug 23 11:15 2_filtering_reads
drwxr-xr-x 3 manager manager 4096 Aug 23 11:18 3_normalization
drwxr-xr-x 3 manager manager 4096 Aug 23 11:14 4_convert_Juice
drwxr-xr-x 3 manager manager 4096 Aug 23 11:12 5_detect_TADs
drwxr-xr-x 3 manager manager 4096 Aug 23 11:14 6_modeling_3D
drwxr-xr-x 4 manager manager 4096 Aug 23 11:17 data
drwxrwxr-x 2 manager manager 4096 Aug 23 10:06 Index
drwxrwxr-x 3 manager manager 4096 Aug 22 22:13 Ref
drwxrwxr-x 6 manager manager 4096 Aug 23 09:12 src
manager@bl8vbox[HiC]
```

1_mapping_read_to_genome
2_filtering_reads
3_normalization
4_convert_Juice
5_detect_TADs
6_modeling_3D

解析のステップごとに、実行する
pythonスクリプトが入ったディレク
トリ（実行結果はそれぞれのResults
の中）

data

解析に使うfastqファイル

Index

ヒトゲノムのBowtie2インデックス

ref

ヒトゲノム配列（fasta）

src

使用したライブラリのソースコード

Hi-C解析とは

シーケンシングによって、ゲノム（染色体）の
3次元空間内の立体構造を明らかにする手法。

構造は機能と密接に関わっている。

(ex. エンハンサー・プロモーターループ)

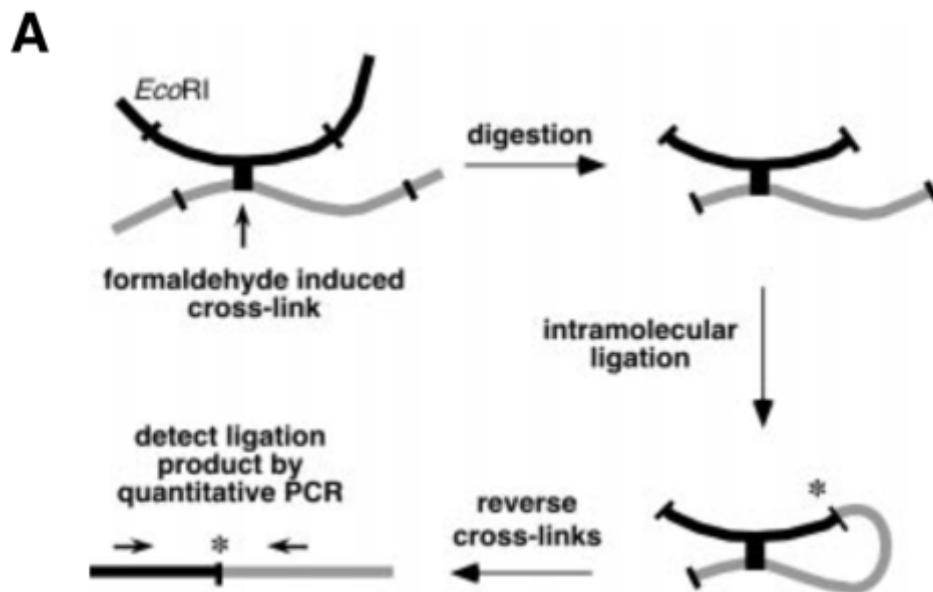
ゲノム配列そのものからはうかがい知れない染色体
の立体構造を、ゲノム配列のシーケンシングによっ
て明らかにするのがHi-C解析。

NGSの応用

- 配列そのものを知りたい
 - ゲノム
 - メタゲノム
 - Reseq
- 読み取られた配列データのパターンから、何か別のことを探りたい
 - RNA-seq
 - ChIP-seq
 - ATAC-seq
 - Hi-C
 - iRep

精度の高いリファレンスゲノムがあることが前提。

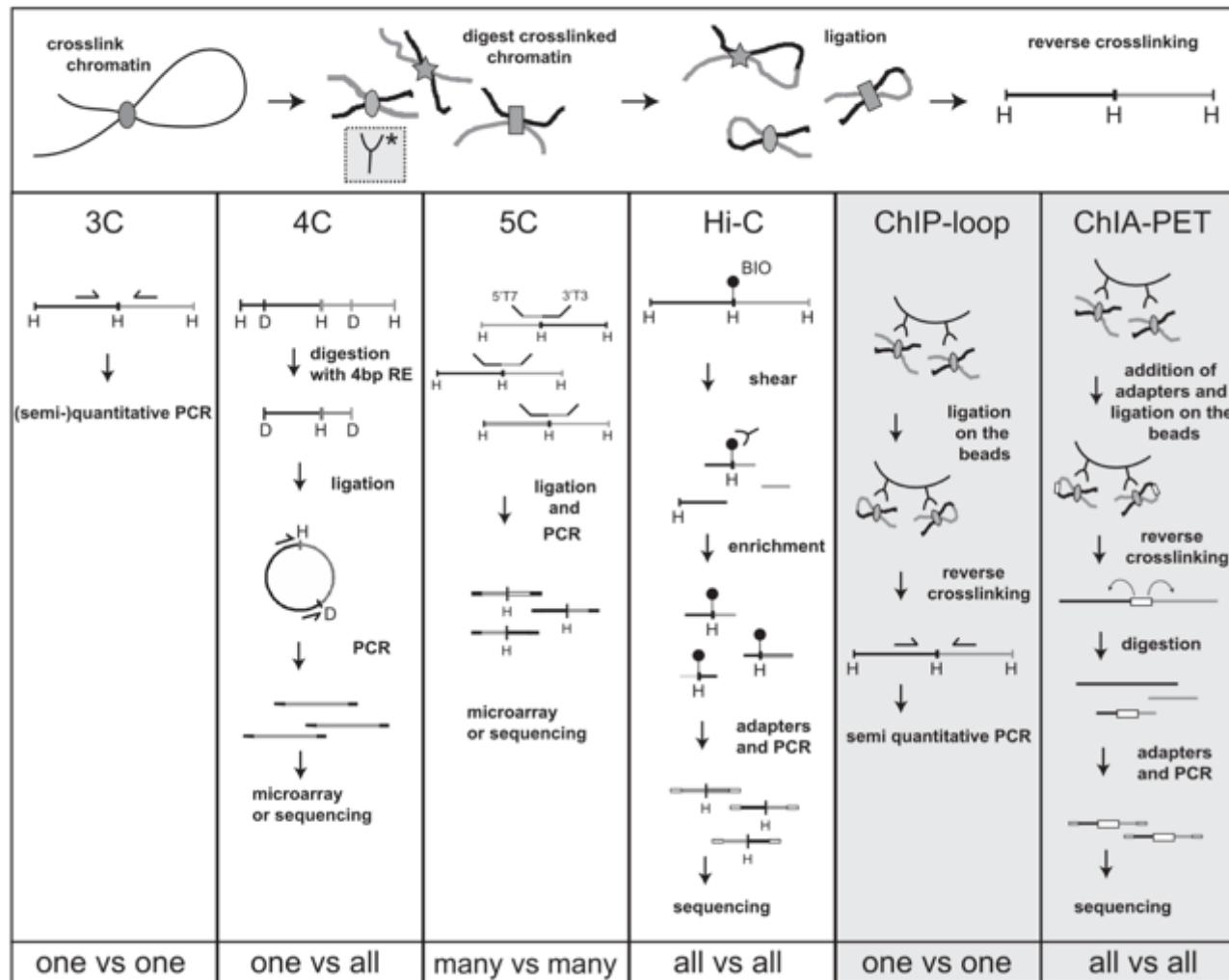
Chromosome Conformation Capture (3C)



Dekker, Job, et al. "Capturing chromosome conformation." *Science* 295.5558 (2002): 1306-1311.

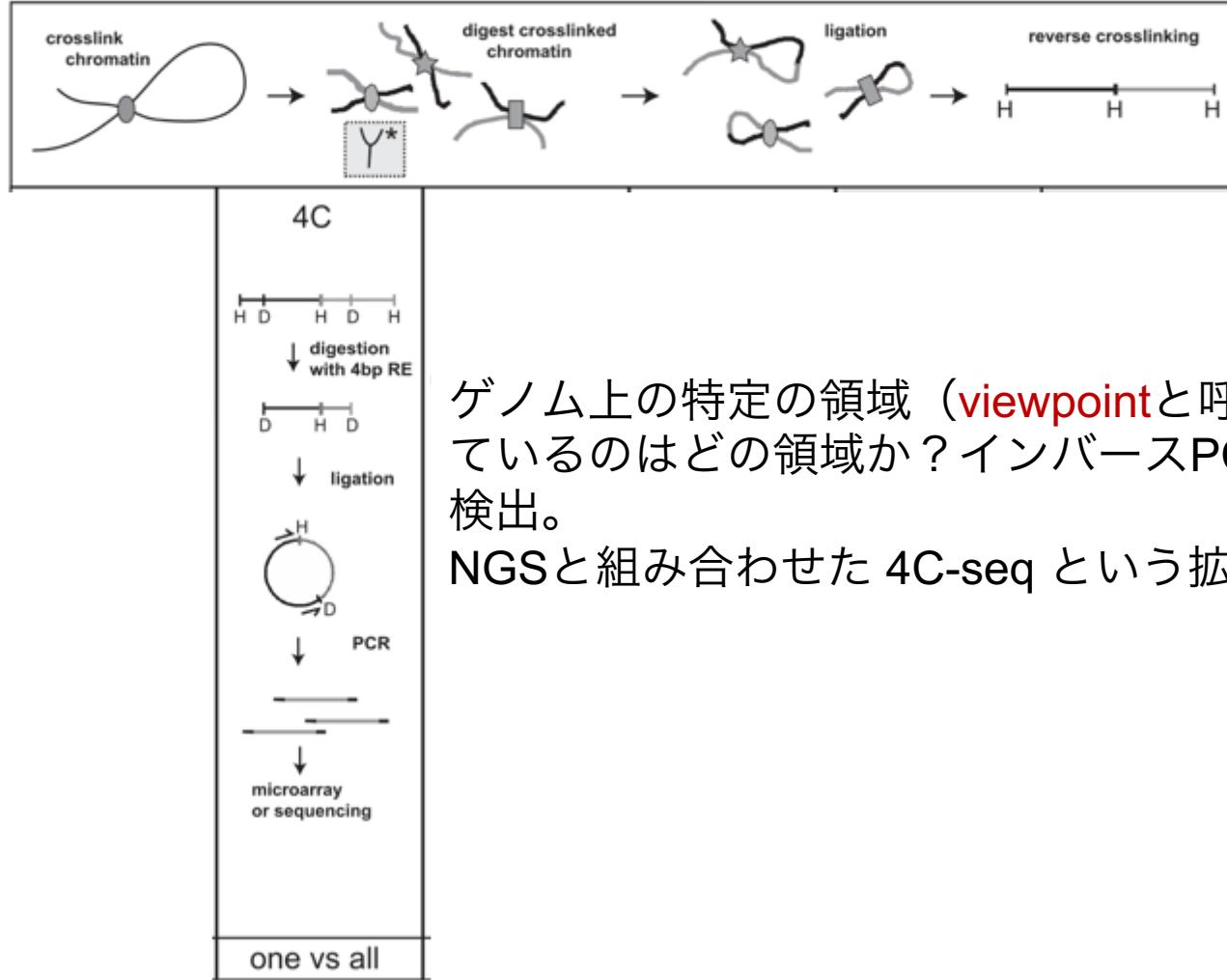
その後、次々に登場した3C-based methodすべての基礎となる手法。
鍵となるのは、DNAの空間的な近接性は、制限消化された断片間でのライゲーションの生じやすさで測れる、という考え方。

3C-based technologies



de Wit, Elzo, and Wouter de Laat. "A decade of 3C technologies: insights into nuclear organization." *Genes & development* 26.1 (2012): 11-24.

4C: Chromosome conformation capture-on-chip

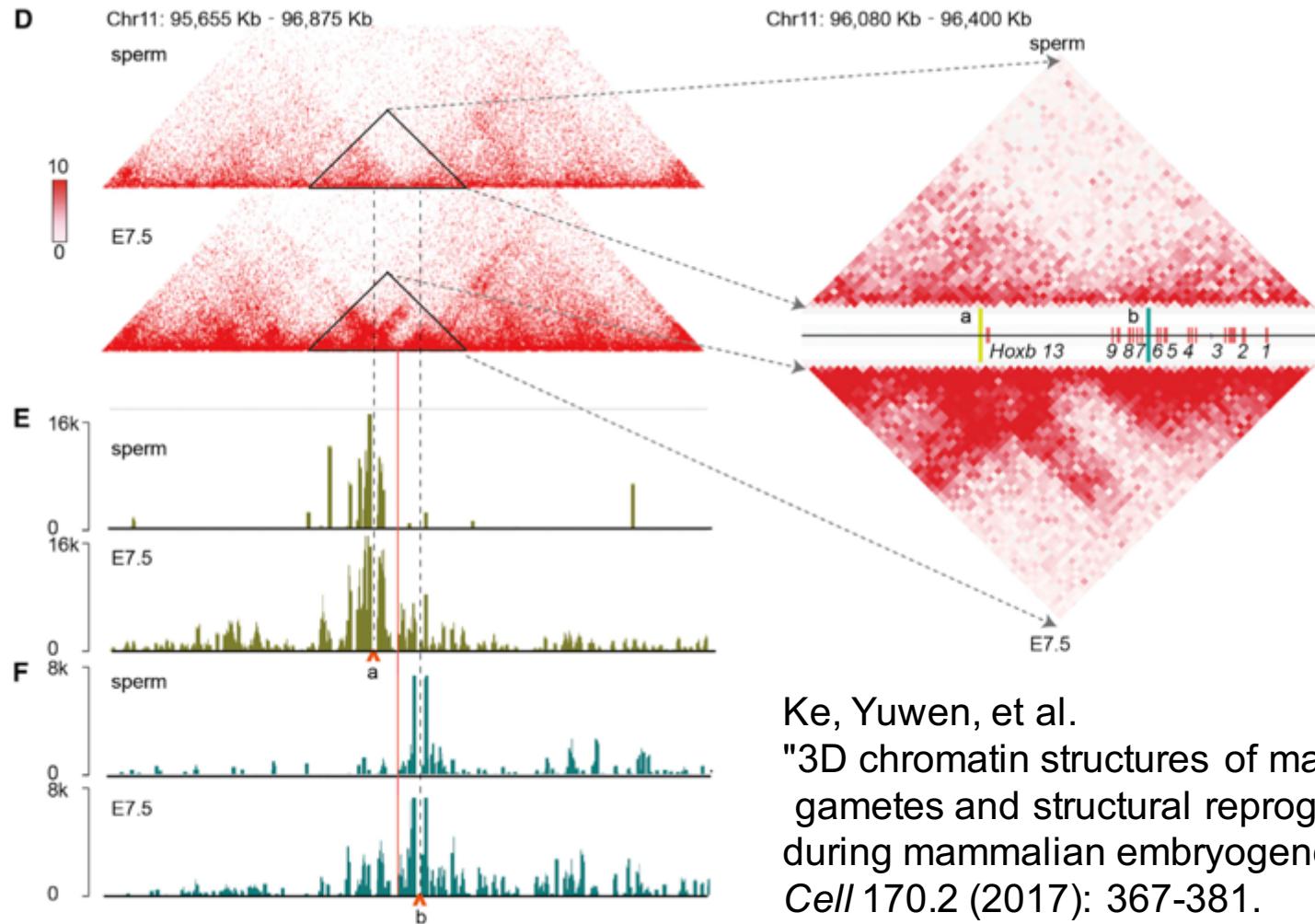


ゲノム上の特定の領域（viewpointと呼ばれる）と相互作用しているのはどの領域か？インバースPCRを利用してアレイで検出。
NGSと組み合わせた 4C-seq という拡張もある。

de Wit, Elzo, and Wouter de Laat. "A decade of 3C technologies: insights into nuclear organization." *Genes & development* 26.1 (2012): 11-24.

4C: Chromosome conformation capture-on-chip

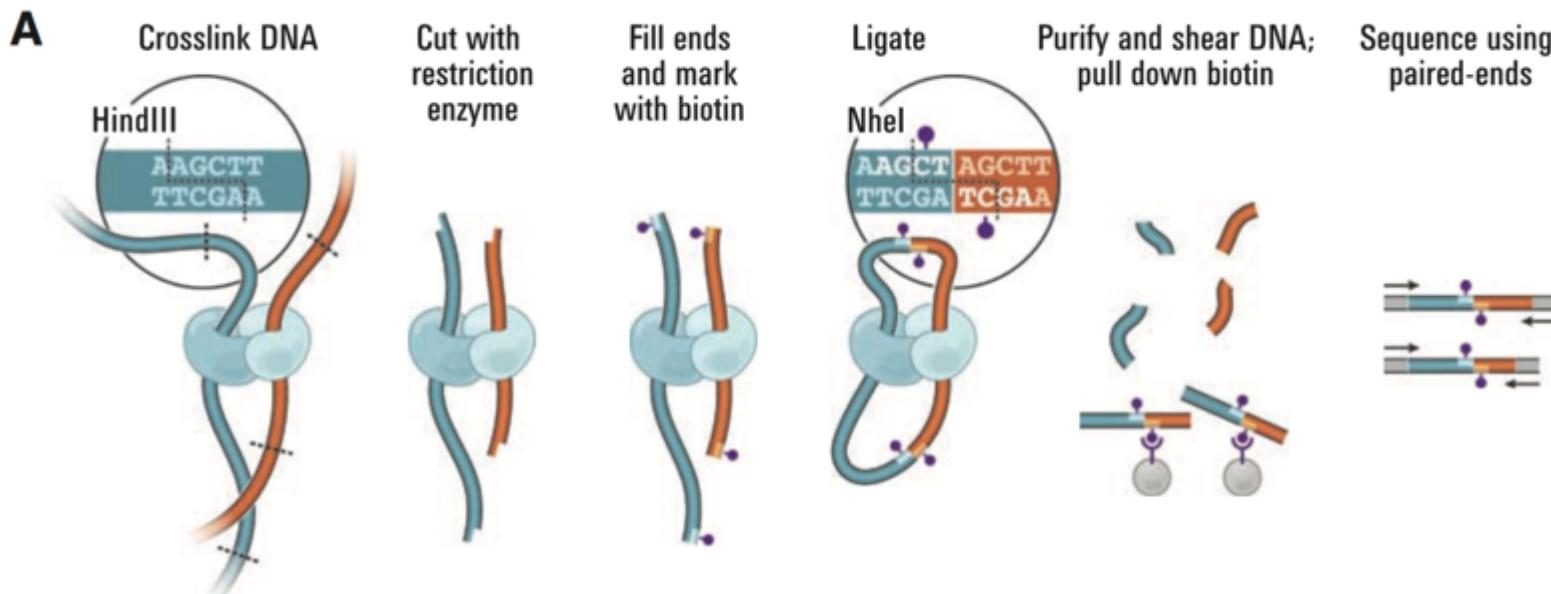
以下は4C-seqデータを使ってHi-Cデータのvalidationをした例



Hi-C

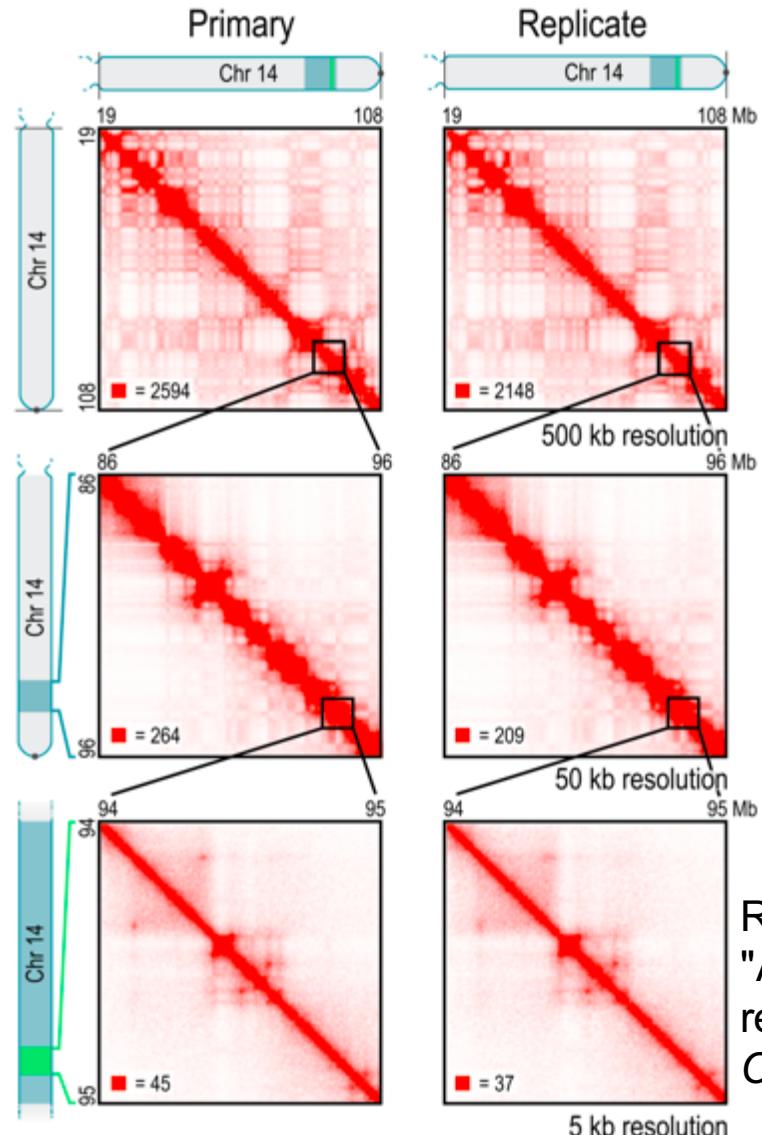
Lieberman-Aiden, Erez, et al.

"Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* 326.5950 (2009): 289-293.



全領域 vs. 全領域の近接性をすべてシーケンシングで決定してしまう。
ビオチンプルダウンで、ライゲーションジャンクションが形成されている断片のみを濃縮。
ペアエンドでシーケンスしなければ意味がない。Forward, Reverse のリードがリファレンスゲノムへマッピングされた位置を調べ、それらのゲノム上の領域がもともと空間的に近接していた、と解釈。

Hi-Cデータ解析のゴールのひとつ =コンタクトマップ（接触確率行列）の生成



左図はコンタクトマップを、ヒートマップとして可視化した図。
コンタクトマップは対称行列。

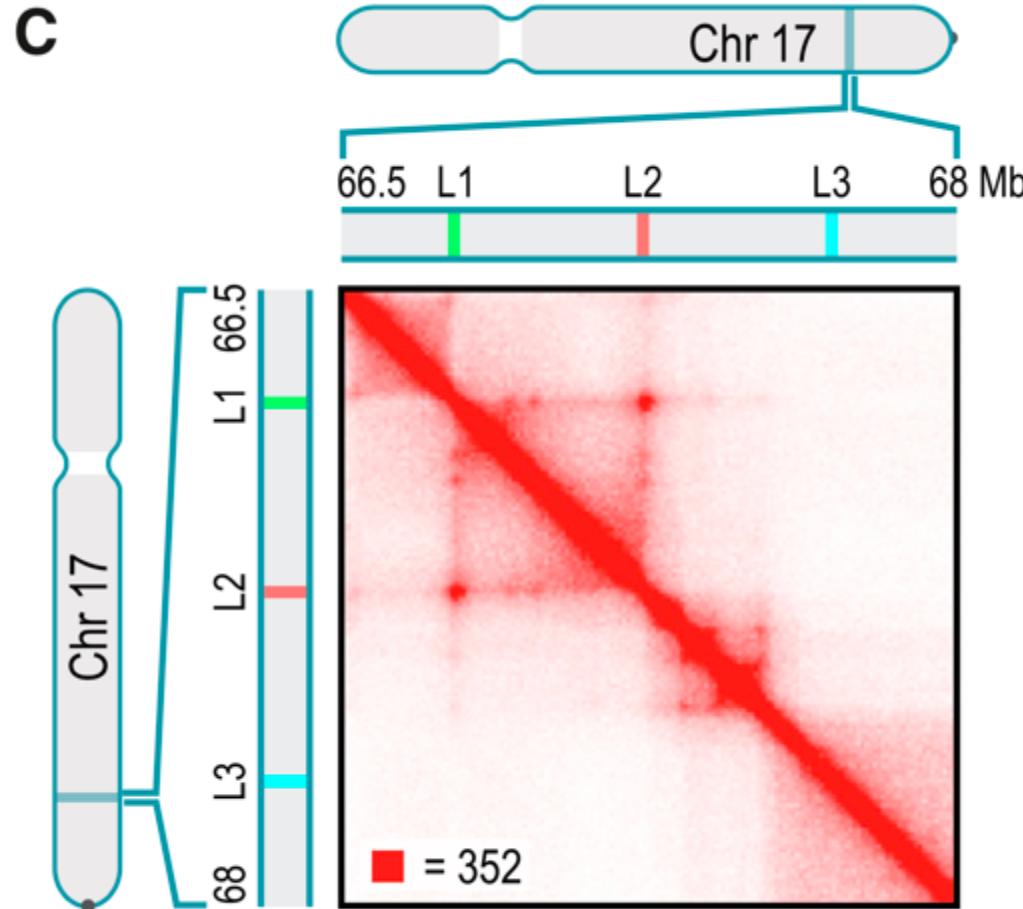
タテとヨコに同じゲノム配列を並べて、
位置 i と位置 j にマップされるペアエンドリードがあったら、行列の(i, j) の
カウントをひとつ増やす。

したがって、行列で値が大きい要素は、
その領域間でマップされるペアがたくさん見つかる、すなわち接触確率が高い領域ペアであることを意味する。

Rao, Suhas SP, et al.
"A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping"
Cell 159.7 (2014): 1665-1680.

コンタクトマップの見方

どのような3次元構造であったら、下図のようなコンタクトマップが得られるだろうか？

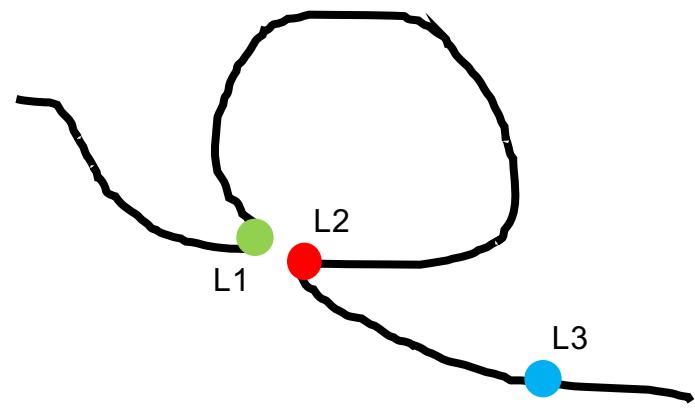
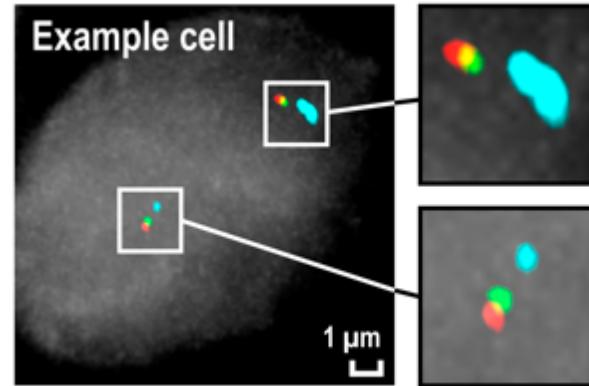
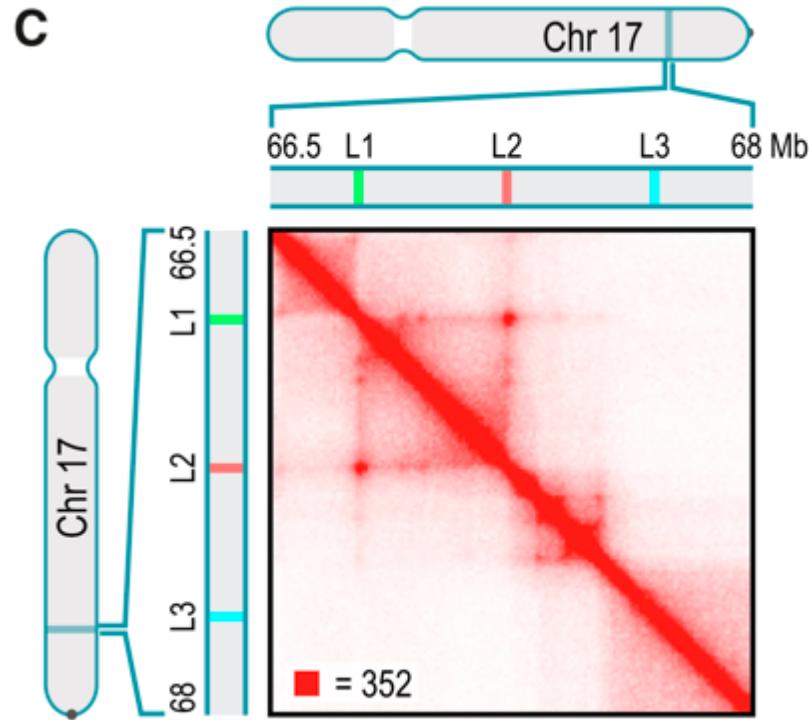


Rao, Suhas SP, et al.

"A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping"
Cell 159.7 (2014): 1665-1680.

コンタクトマップの見方

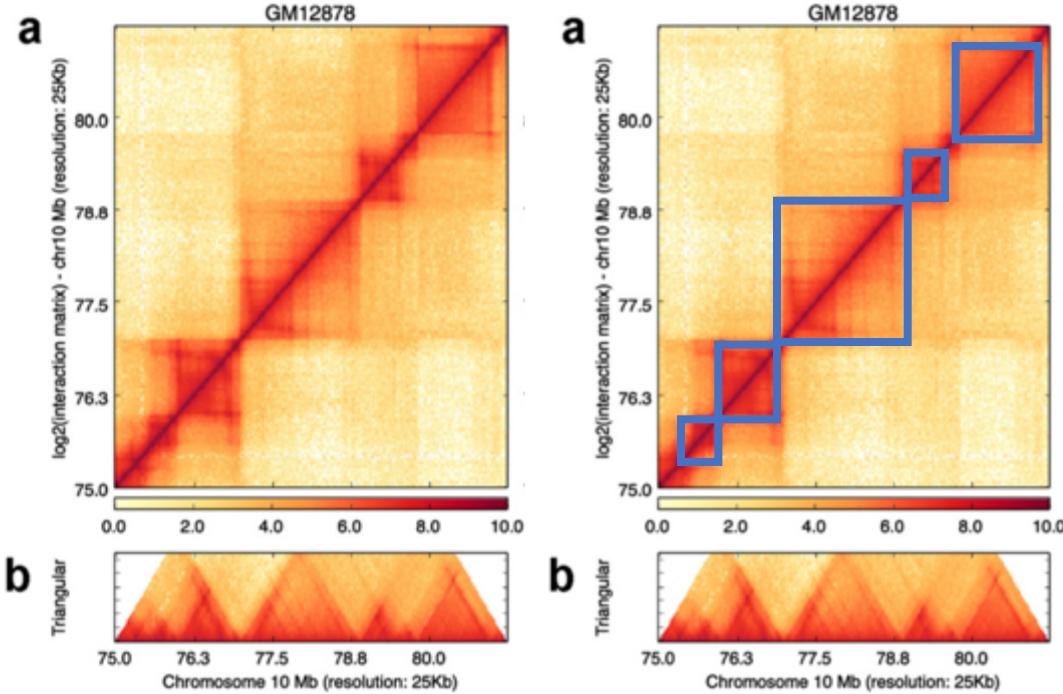
どのような3次元構造であったら、下図のようなコンタクトマップが得られるだろうか？



Rao, Suhas SP, et al.
"A 3D map of the human genome at kilobase
resolution reveals principles of chromatin looping"
Cell 159.7 (2014): 1665-1680.

コンタクトマップの見方

Topologically Associated Domains (TADs)



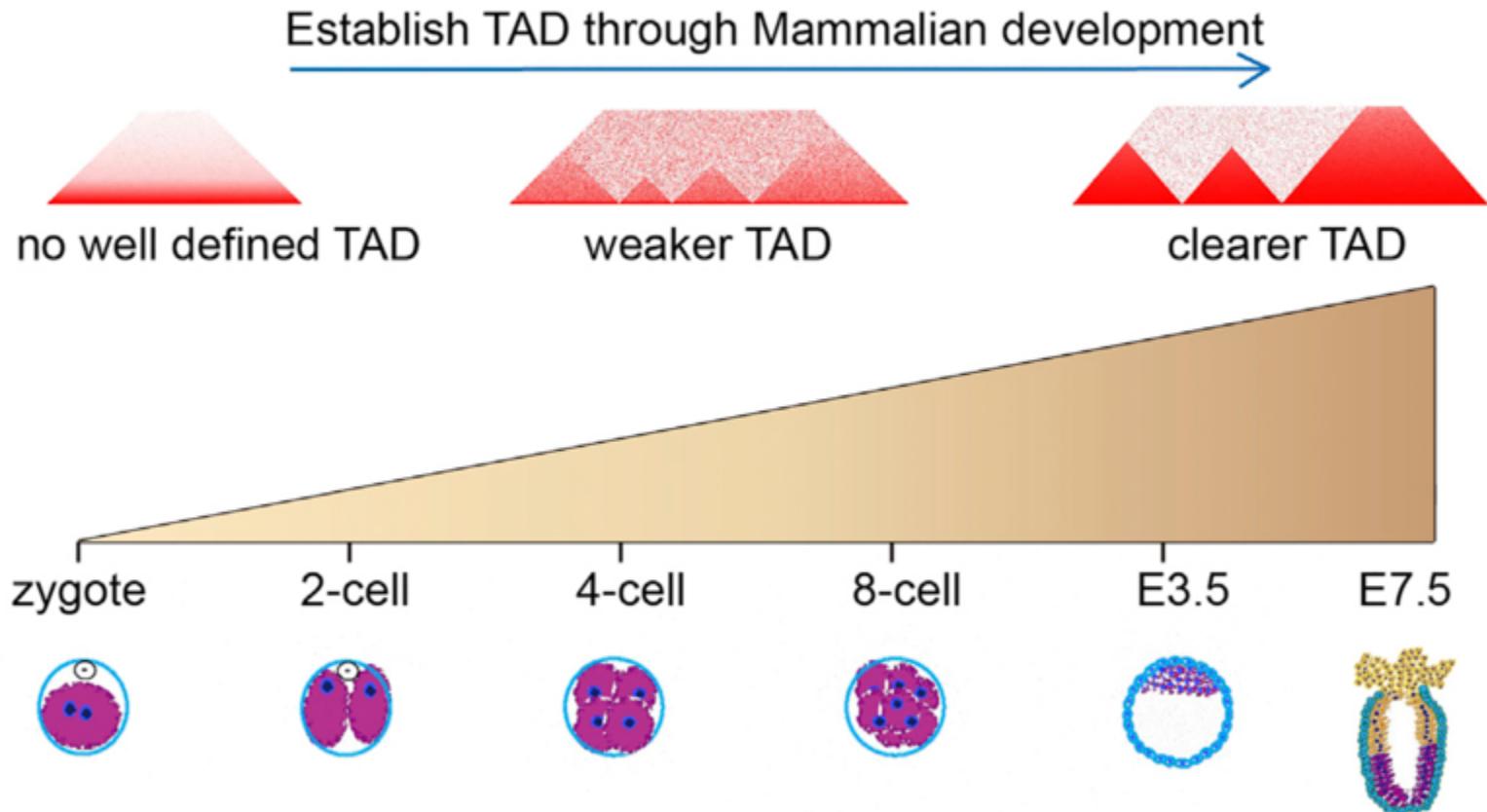
コンタクトマップの対角線上に沿ってしばしば見られる正方形のかたち。

ヒストン修飾のパターンや複製タイミングなどと強く相関している構造。

エンハンサーの影響力をひとつ
のTADの内部に隔離するイン
シュレータとしての機能を持つ？

Akdemir, Kadir Caner, and Lynda Chin.
"HiCPlotter integrates genomic data with
interaction matrices." *Genome biology* 16.1
(2015): 198.

Topologically Associated Domains (TADs)



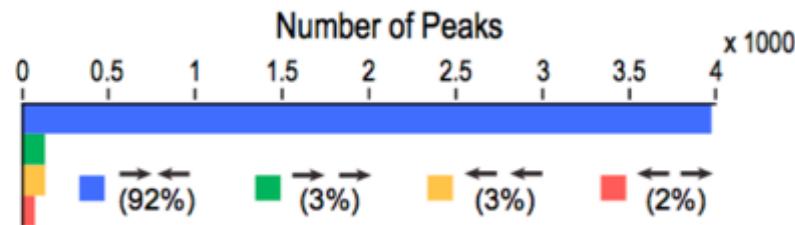
Ke, Yuwen, et al.

"3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis."

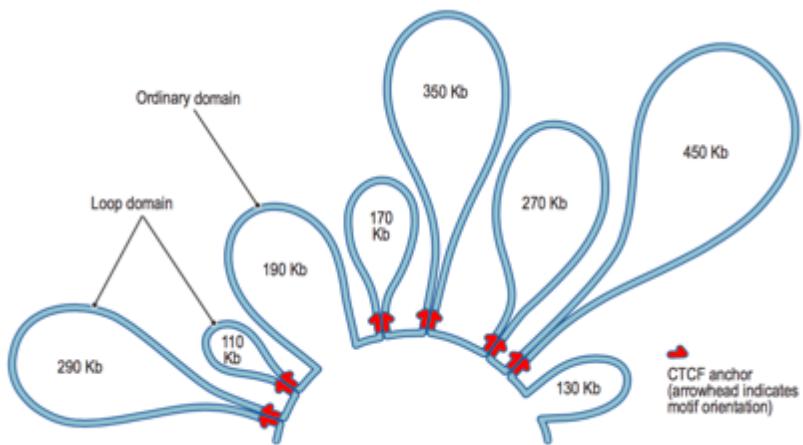
Cell 170.2 (2017): 367-381.

Topologically Associated Domains (TADs)

D

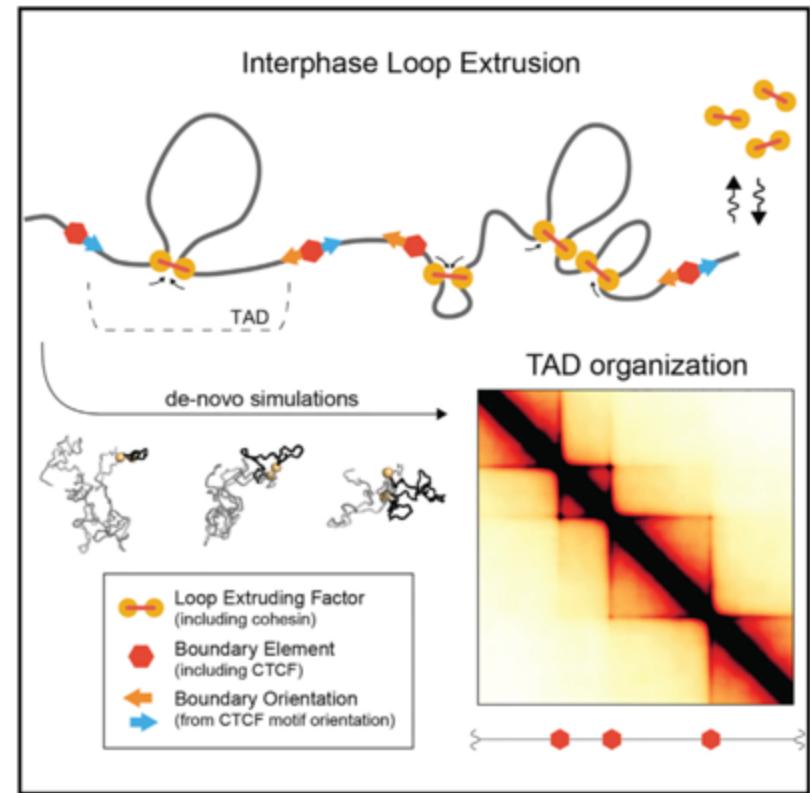


F



Rao, Suhas SP, et al.

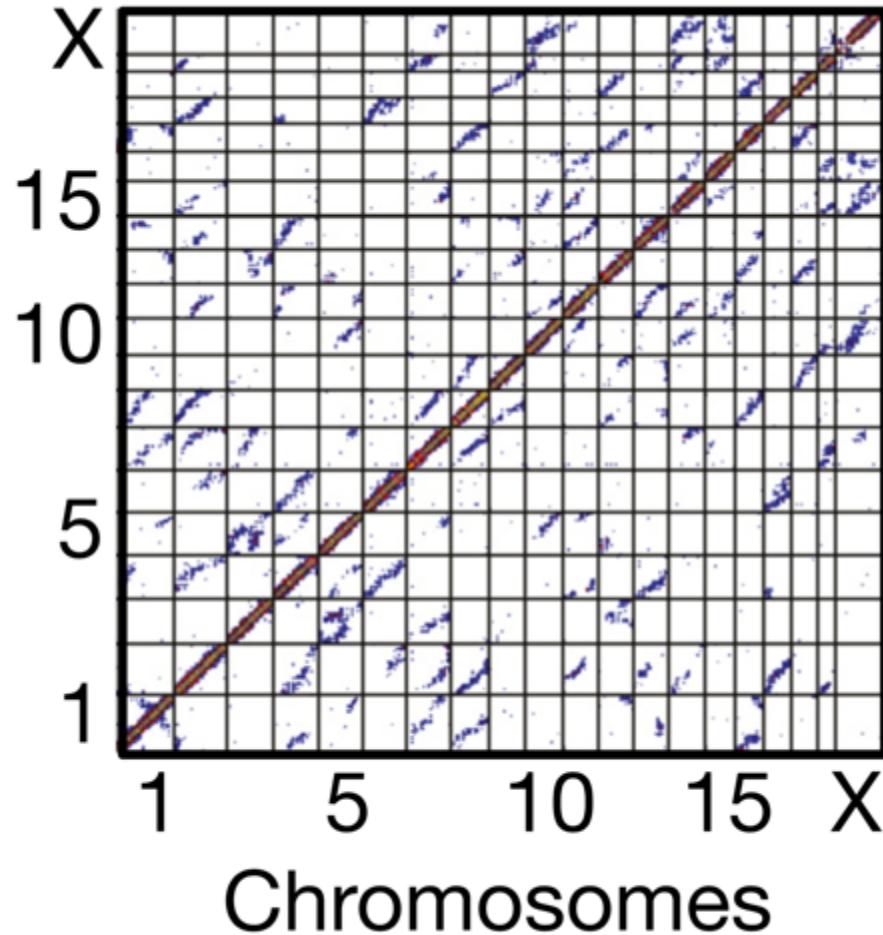
"A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping"
Cell 159.7 (2014): 1665-1680.



Fudenberg, Geoffrey, et al. "Formation of chromosomal domains by loop extrusion." Cell reports 15.9 (2016): 2038-2049.

コンタクトマップの見方

どのような3次元構造であったら、下図のようなコンタクトマップが得られるだろうか？



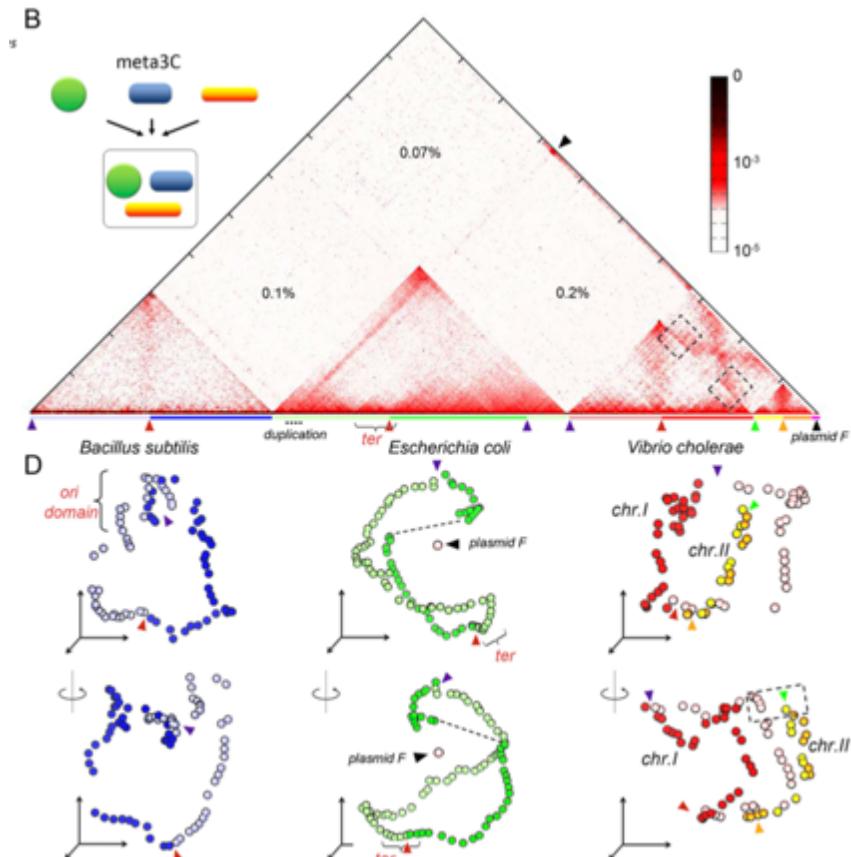
Nagano, Takashi, et al.
“Cell-cycle dynamics of
chromosomal organization
at single-cell resolution.”
Nature 547 (2017): 61–67

Hi-Cデータの応用

1. ハプロタイプフェイジング

2. ゲノムアセンブリ

3. メタゲノム (meta3C)



Marbouty, Martial, et al.

"Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms."
Elife 3 (2014): e03318.

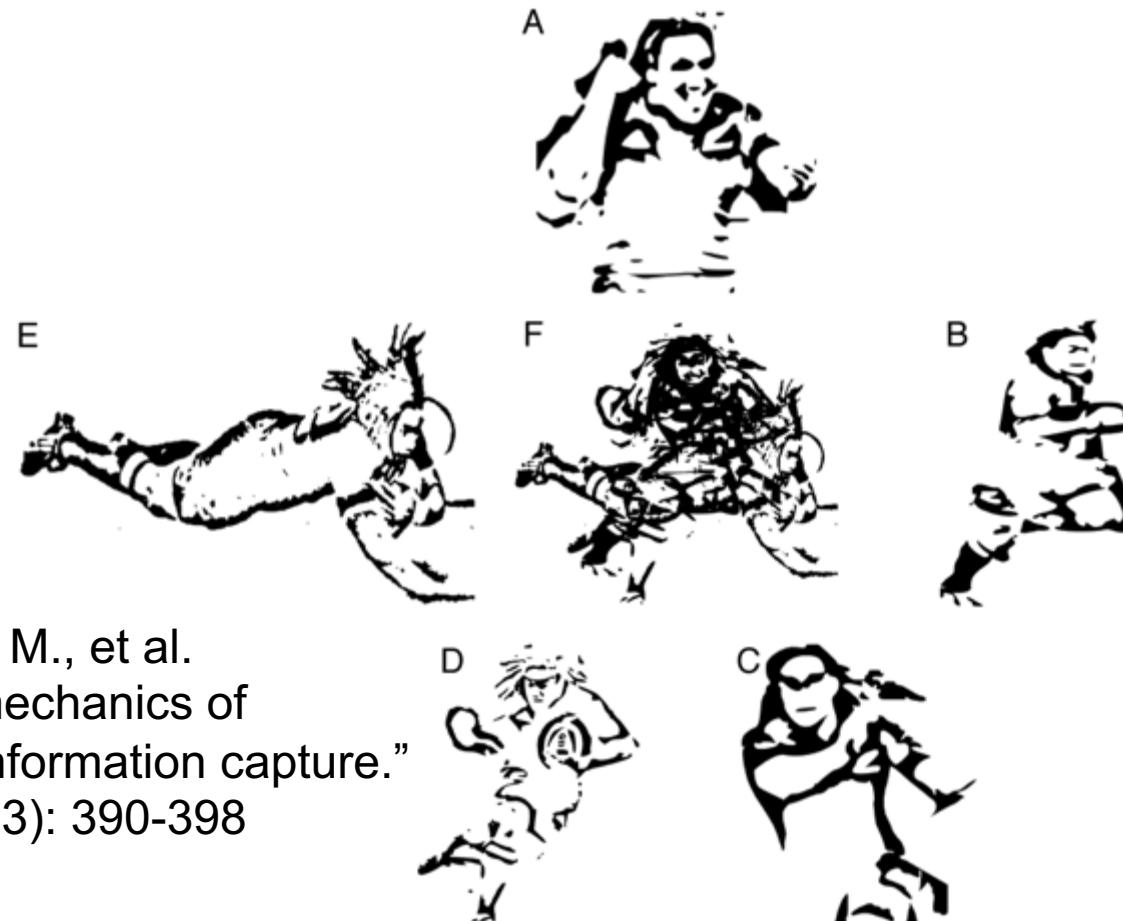
Hi-Cのデメリット①

- 高い解像度のコンタクトマップを得るためにには莫大な量のシーケンスが必要
 - コンタクトマップを構成する要素の数は、ゲノムを分割するBinのサイズ（解像度）に対して二乗で大きくなる。そのため10倍の解像度のコンタクトマップを得るためにには100倍のシーケンスが必要になる。

Hi-Cのデメリット②

- 集団の平均的構造であること (excl. single cell Hi-C)

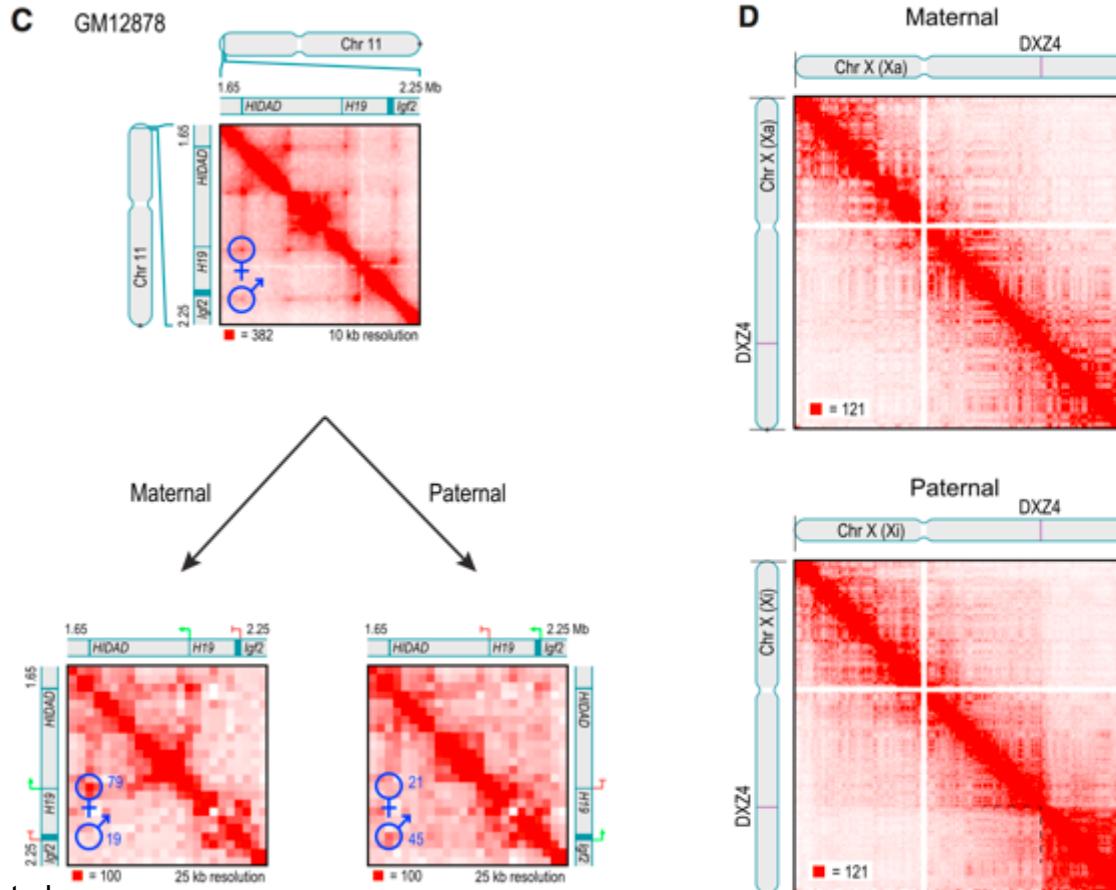
- 集団内の構造の多様性は3次元モデリングの際に「推論」するしかない



O'sullivan, Justin M., et al.
"The statistical-mechanics of
chromosome conformation capture."
Nucleus 4.5 (2013): 390-398

Hi-Cのデメリット②

- 集団の平均的構造であること (excl. single cell Hi-C)
 - 構造の多様性は相同染色体間でさえ観察される (フェイジングデータが利用できればいいが...)

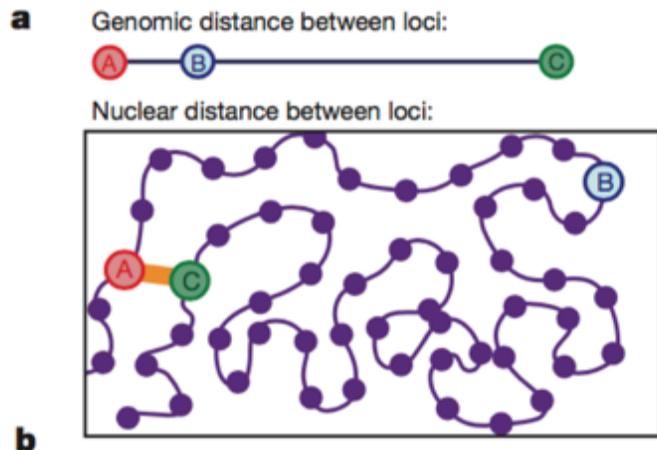


Rao, Suhas SP, et al.

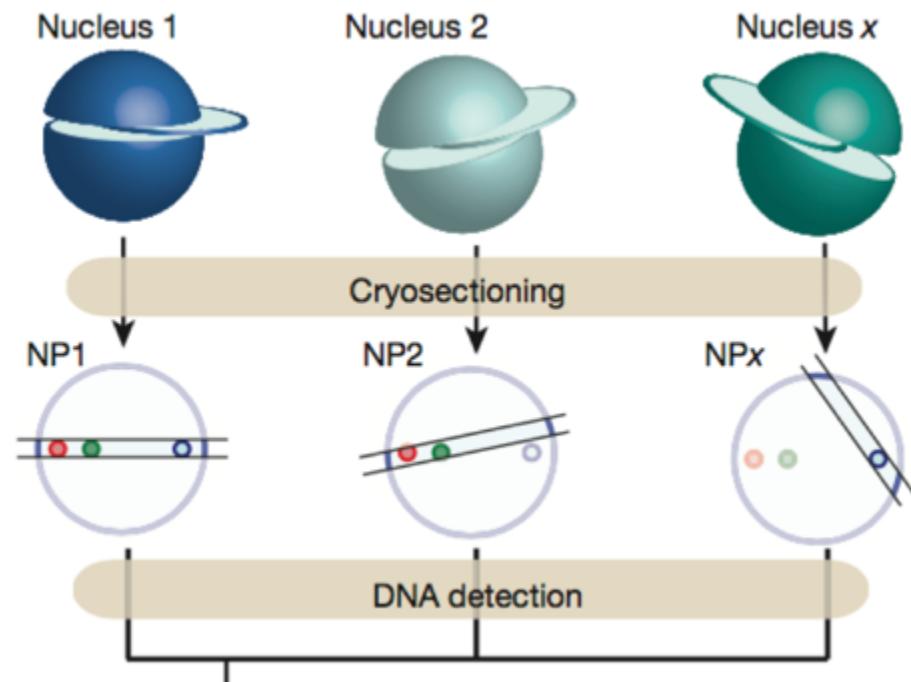
"A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping"
Cell 159.7 (2014): 1665-1680.

Hi-Cのデメリット③

- 3Cの原理的に、1対1の領域ペアの接触しか観測できない
 - 複数の領域の同時接触はデータから間接的にわかるだけ。
 - 新たな手法 Genome Architecture Mapping ならば克服できるかも。



b

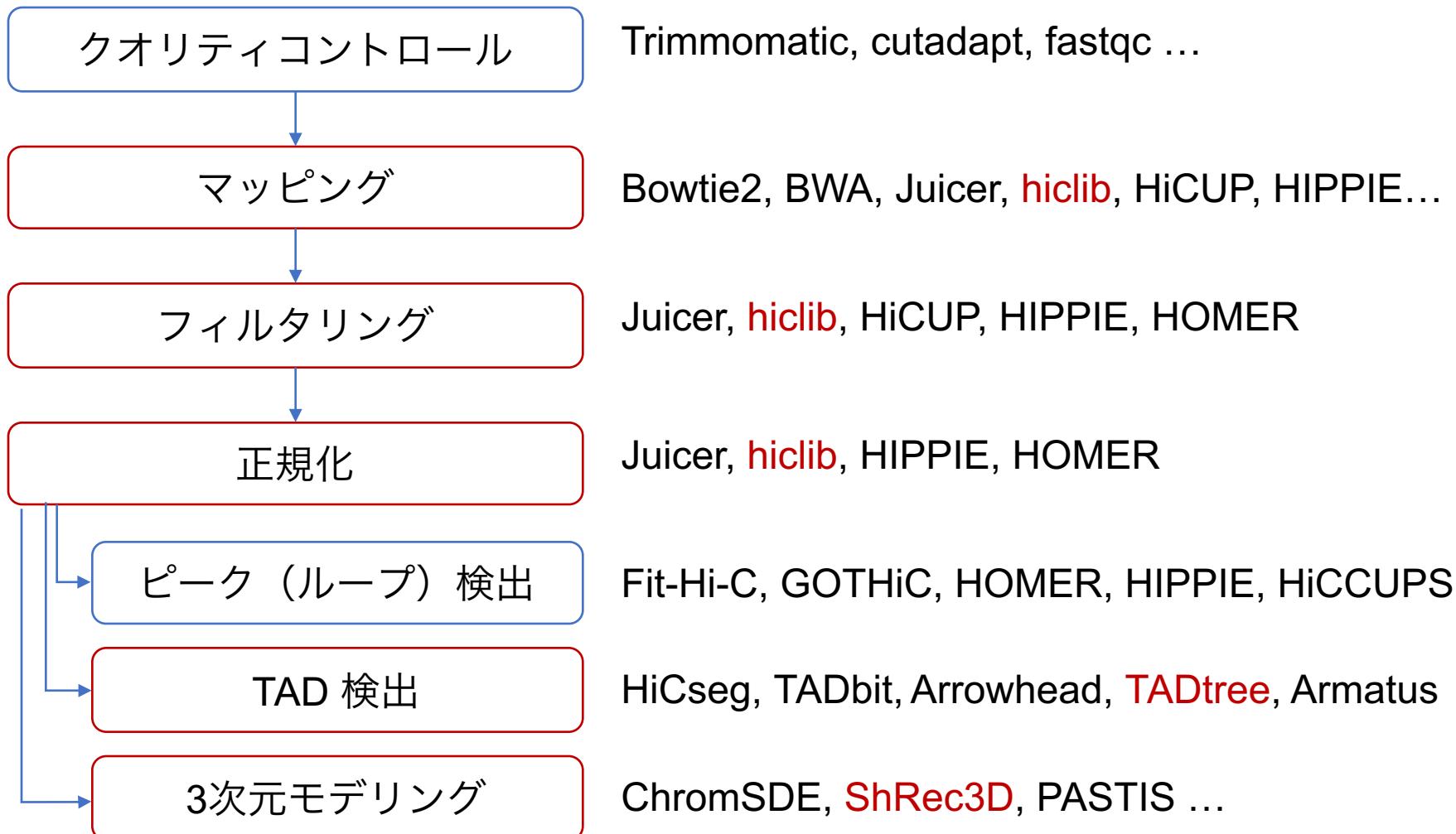


Beagrie, Robert A., et al.
"Complex multi-enhancer contacts captured by genome architecture mapping."
Nature 543.7646 (2017): 519-524.

本講義の内容

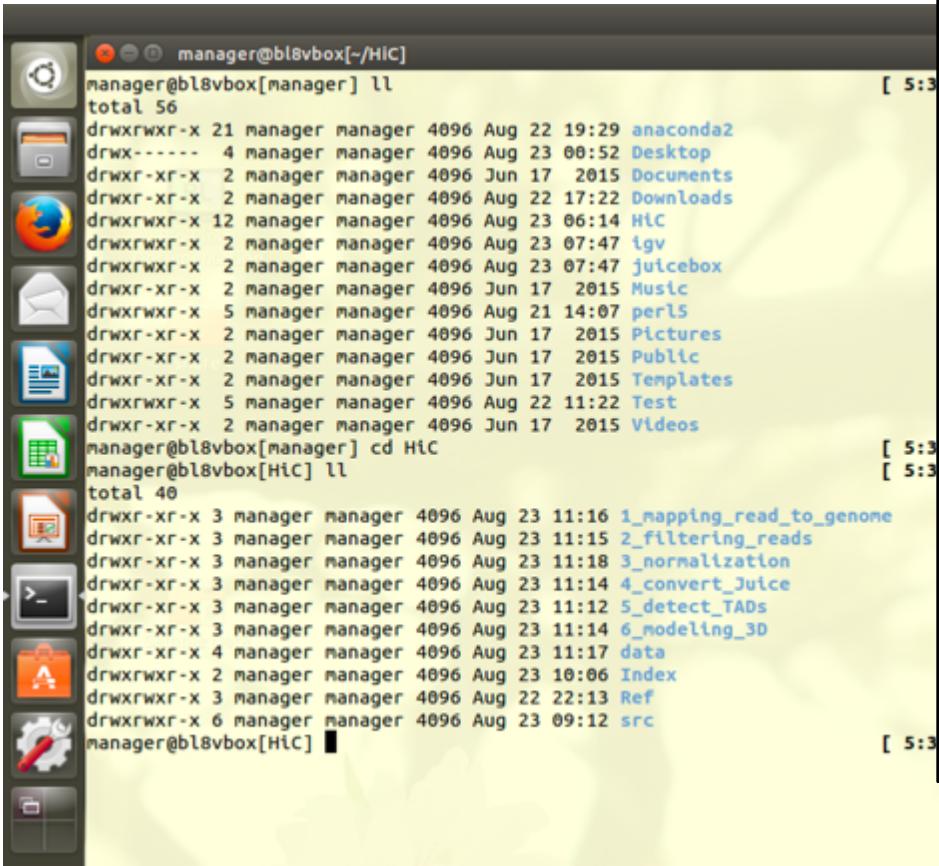
- Hi-C解析とは
 - Chromosome Conformation Capture の原理
 - Hi-Cで何がわかるか？コンタクトマップの見方
- Hi-C解析の流れ（実習と並行）
 - Hi-C解析のツール
 - マッピング
 - フィルタリング
 - 正規化
 - ピーク検出、TAD検出など
 - 3D構造モデリング

Hi-C解析の流れ、利用可能なツール（一部）



実習の内容

Bio-Linux-8.0.7_hm_kh.ova を起動。
すべて、~/HiC の中で実行する。



```
manager@bl8vbox[~/HiC]
manager@bl8vbox[manager] ll
total 56
drwxrwxr-x 21 manager manager 4096 Aug 22 19:29 anaconda2
drwx----- 4 manager manager 4096 Aug 23 00:52 Desktop
drwxr-xr-x  2 manager manager 4096 Jun 17 2015 Documents
drwxr-xr-x  2 manager manager 4096 Aug 22 17:22 Downloads
drwxrwxr-x 12 manager manager 4096 Aug 23 06:14 HiC
drwxrwxr-x  2 manager manager 4096 Aug 23 07:47 igv
drwxrwxr-x  2 manager manager 4096 Aug 23 07:47 juicebox
drwxr-xr-x  2 manager manager 4096 Jun 17 2015 Music
drwxrwxr-x  5 manager manager 4096 Aug 21 14:07 perl5
drwxr-xr-x  2 manager manager 4096 Jun 17 2015 Pictures
drwxr-xr-x  2 manager manager 4096 Jun 17 2015 Public
drwxr-xr-x  2 manager manager 4096 Jun 17 2015 Templates
drwxrwxr-x  5 manager manager 4096 Aug 22 11:22 Test
drwxr-xr-x  2 manager manager 4096 Jun 17 2015 Videos
manager@bl8vbox[manager] cd HiC
manager@bl8vbox[HiC] ll
total 40
drwxr-xr-x  3 manager manager 4096 Aug 23 11:16 1_mapping_read_to_genome
drwxr-xr-x  3 manager manager 4096 Aug 23 11:15 2_filtering_reads
drwxr-xr-x  3 manager manager 4096 Aug 23 11:18 3_normalization
drwxr-xr-x  3 manager manager 4096 Aug 23 11:14 4_convert_Juice
drwxr-xr-x  3 manager manager 4096 Aug 23 11:12 5_detect_TADs
drwxr-xr-x  3 manager manager 4096 Aug 23 11:14 6_modeling_3D
drwxr-xr-x  4 manager manager 4096 Aug 23 11:17 data
drwxrwxr-x  2 manager manager 4096 Aug 23 10:06 Index
drwxrwxr-x  3 manager manager 4096 Aug 22 22:13 Ref
drwxrwxr-x  6 manager manager 4096 Aug 23 09:12 src
manager@bl8vbox[HiC]
```

1_mapping_read_to_genome
2_filtering_reads
3_normalization
4_convert_Juice
5_detect_TADs
6_modeling_3D

解析のステップごとに、実行する
pythonスクリプトが入ったディレク
トリ（実行結果はそれぞれのResults
の中）

data

解析に使うfastqファイル

Index

ヒトゲノムのBowtie2インデックス

ref

ヒトゲノム配列（fasta）

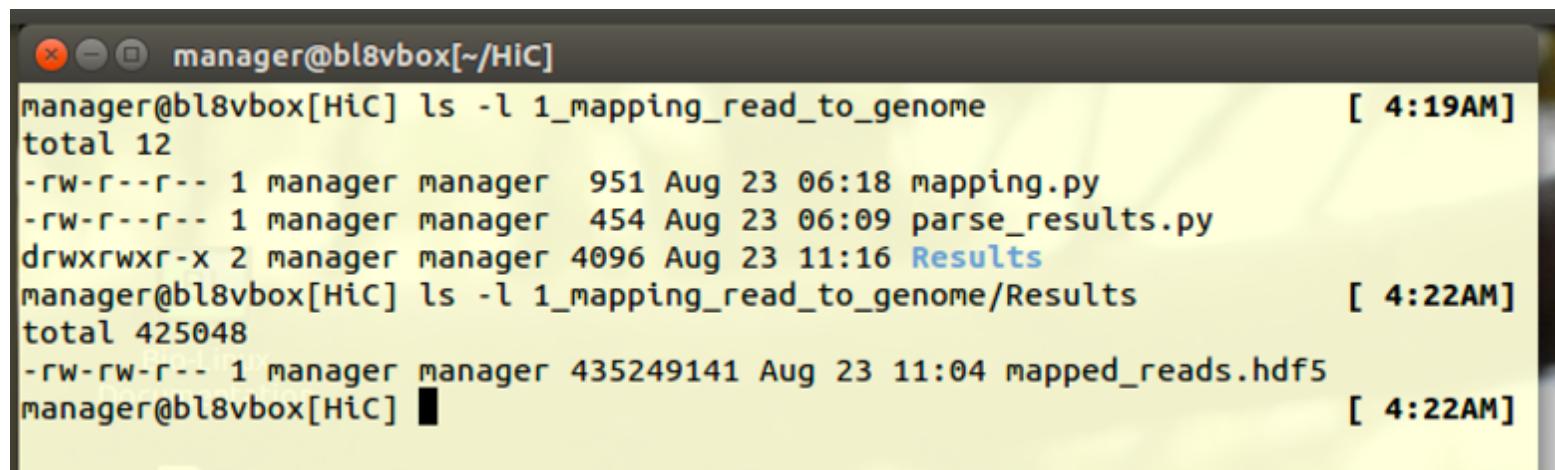
src

使用したライブラリのソースコード

うまく動かない、どうしても失敗する場合

各ステップのディレクトリの中に、それぞれ “Results” というディレクトリがあります。

その中に、そのステップで生成されるはずの正解データが入っているので、どうしても失敗する場合はそれをひとつ上の階層に mv していただければ、次のステップ以降の解析も続けられます。



```
manager@bl8vbox[~/HiC]
manager@bl8vbox[HiC] ls -l 1_mapping_read_to_genome [ 4:19AM]
total 12
-rw-r--r-- 1 manager manager 951 Aug 23 06:18 mapping.py
-rw-r--r-- 1 manager manager 454 Aug 23 06:09 parse_results.py
drwxrwxr-x 2 manager manager 4096 Aug 23 11:16 Results
manager@bl8vbox[HiC] ls -l 1_mapping_read_to_genome/Results [ 4:22AM]
total 425048
-rw-rw-r-- 1 manager manager 435249141 Aug 23 11:04 mapped_reads.hdf5
manager@bl8vbox[HiC] █
```

実習で使用するデータ

In situ Hi-CでKilobase解像度に達したヒトHi-Cのランドマーク的な論文。
100以上のサンプル、各サンプルあたり数億ペアエンド
(1サンプルのfastqでも100GB近いファイルサイズ)

Rao, Suhas SP, et al.

"A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping" *Cell* 159.7 (2014): 1665-1680.

The image shows a screenshot of a scientific article from the journal *Cell*. The article is titled "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping". The authors listed are Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. The article is categorized under "Article" and "Cell". The text of the article is as follows:

A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping

Suhas S.P. Rao,^{1,2,3,4,5,10} Miriam H. Huntley,^{1,2,3,4,5,10} Neva C. Durand,^{1,2,3,4} Elena K. Stamenova,^{1,2,3,4} Ivan D. Bochkov,^{1,2,3} James T. Robinson,^{1,4} Adrian L. Sanborn,^{1,2,3,5} Ido Machol,^{1,2,3} Arina D. Omer,^{1,2,3} Eric S. Lander,^{4,7,8*} and Erez Lieberman Aiden^{1,2,3,4,5*}

¹The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA
²Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
³Department of Computer Science, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA
⁴Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA
⁵School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
⁶Department of Computer Science, Stanford University, Stanford, CA 94305, USA
⁷Department of Biology, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA
⁸Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
⁹Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA
¹⁰Co-first author
Correspondence: lander@broadinstitute.org (E.S.L.), eruz@erez.com (E.L.A.)
<http://dx.doi.org/10.1016/j.cell.2014.11.021>

~/HiC/data

今回は Rao, et al. 2014 の公開データの中の
1 サンプル (Human B-Lymphocyte: GM12878) のみ、
さらに1000万リードにダウンサンプリングしたデータを扱う。

```
$cd ~/HiC/data
```

```
$ls -l
```

すると、R1, R2 それぞれの fastq ファイルがある。

~/HiC/Ref

ヒトリファレンスゲノム (hg19)

<http://hgdownload.cse.ucsc.edu/downloads.html> からダウンロード

染色体ごとのFASTAファイル

```
manager@bl8vbox[~/HiC]
```

```
manager@bl8vbox[HiC] ls -l ./Ref/hg19
total 3083696
-rw-rw-r-- 1 manager manager 138245449 Aug 22 19:57 chr10.fa
-rw-rw-r-- 1 manager manager 137706654 Aug 22 19:58 chr11.fa
-rw-rw-r-- 1 manager manager 136528940 Aug 22 19:58 chr12.fa
-rw-rw-r-- 1 manager manager 117473283 Aug 22 20:00 chr13.fa
-rw-rw-r-- 1 manager manager 109496538 Aug 22 20:00 chr14.fa
-rw-rw-r-- 1 manager manager 104582027 Aug 22 20:00 chr15.fa
-rw-rw-r-- 1 manager manager 92161856 Aug 22 20:00 chr16.fa
-rw-rw-r-- 1 manager manager 82819122 Aug 22 20:00 chr17.fa
-rw-rw-r-- 1 manager manager 79638800 Aug 23 07:22 chr18.fa
-rw-rw-r-- 1 manager manager 60311570 Aug 22 20:01 chr19.fa
-rw-rw-r-- 1 manager manager 254235640 Aug 22 19:48 chr1.fa
-rw-rw-r-- 1 manager manager 64286038 Aug 22 20:01 chr20.fa
-rw-rw-r-- 1 manager manager 49092500 Aug 22 20:01 chr21.fa
-rw-rw-r-- 1 manager manager 52330665 Aug 22 20:01 chr22.fa
-rw-rw-r-- 1 manager manager 248063367 Aug 22 19:49 chr2.fa
-rw-rw-r-- 1 manager manager 201982885 Aug 22 19:50 chr3.fa
-rw-rw-r-- 1 manager manager 194977368 Aug 22 19:50 chr4.fa
-rw-rw-r-- 1 manager manager 184533572 Aug 22 19:53 chr5.fa
-rw-rw-r-- 1 manager manager 174537375 Aug 22 19:53 chr6.fa
-rw-rw-r-- 1 manager manager 162321443 Aug 22 19:54 chr7.fa
-rw-rw-r-- 1 manager manager 149291309 Aug 22 19:55 chr8.fa
-rw-rw-r-- 1 manager manager 144037706 Aug 22 19:55 chr9.fa
-rw-rw-r-- 1 manager manager 16909 Aug 22 20:02 chrM.fa
-rw-rw-r-- 1 manager manager 158375978 Aug 22 20:02 chrX.fa
-rw-rw-r-- 1 manager manager 60561044 Aug 22 20:02 chrY.fa
-rw-rw-r-- 1 manager manager 23222 Aug 22 22:13 gap.txt
drwxrwxr-x 3 manager manager 4096 Aug 23 07:13 joblib
```

~/HiC/Index

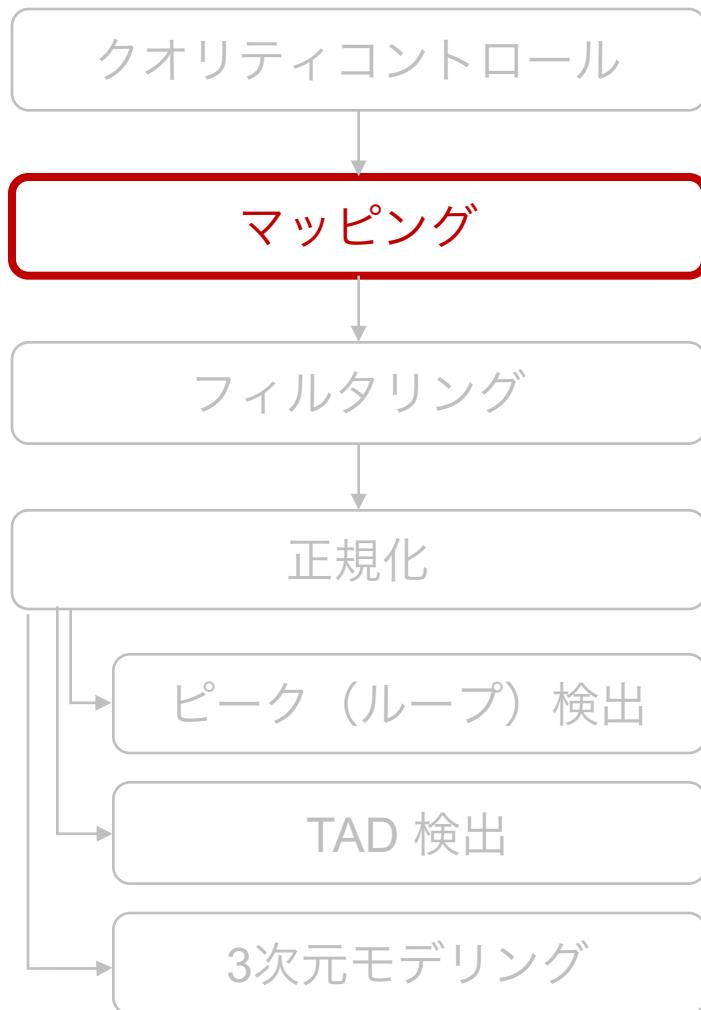
~/HiC/Ref のヒトリファレンスゲノム (hg19) を
bowtie2-build したもの (Bowtie2のインデックスファイル)

```
manager@bl8vbox[~/HiC]
manager@bl8vbox[HiC] ls -l ./Index
total 3966796
-rw-rw-r-- 1 manager manager 957980027 Aug 23 09:04 hg19.1.bt2
-rw-rw-r-- 1 manager manager 715335932 Aug 23 09:04 hg19.2.bt2
-rw-rw-r-- 1 manager manager      3284 Aug 23 08:05 hg19.3.bt2
-rw-rw-r-- 1 manager manager 715335926 Aug 23 08:05 hg19.4.bt2
-rw-rw-r-- 1 manager manager 957980027 Aug 23 09:54 hg19.rev.1.bt2
-rw-rw-r-- 1 manager manager 715335932 Aug 23 09:54 hg19.rev.2.bt2
```

Hi-C解析の流れ



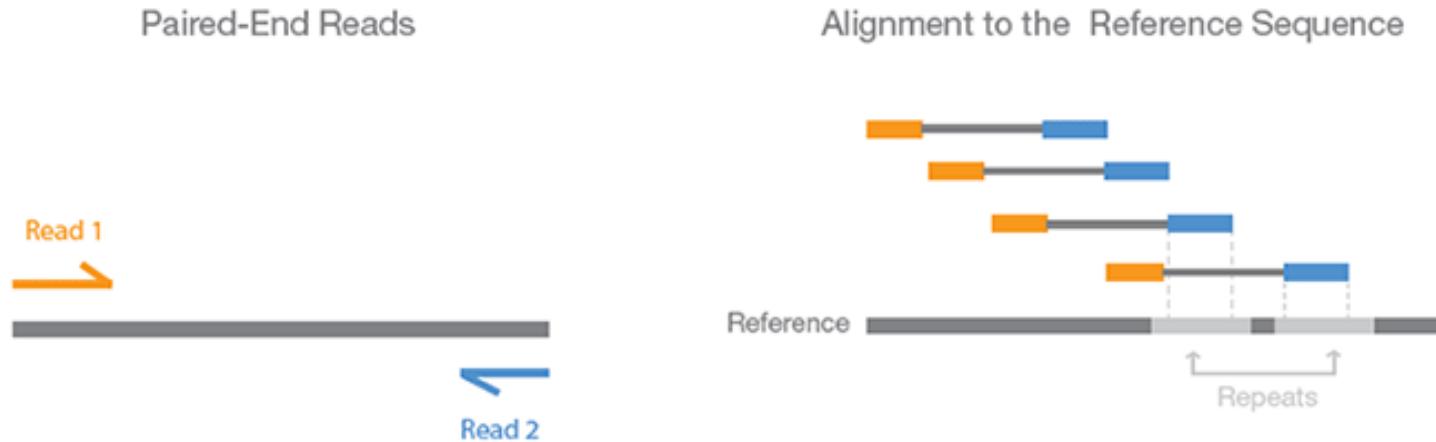
Hi-C解析の流れ



```
$cd ~/HiC/1_mapping_read_to_genome
```

Illuminaのペアエンドシーケンス

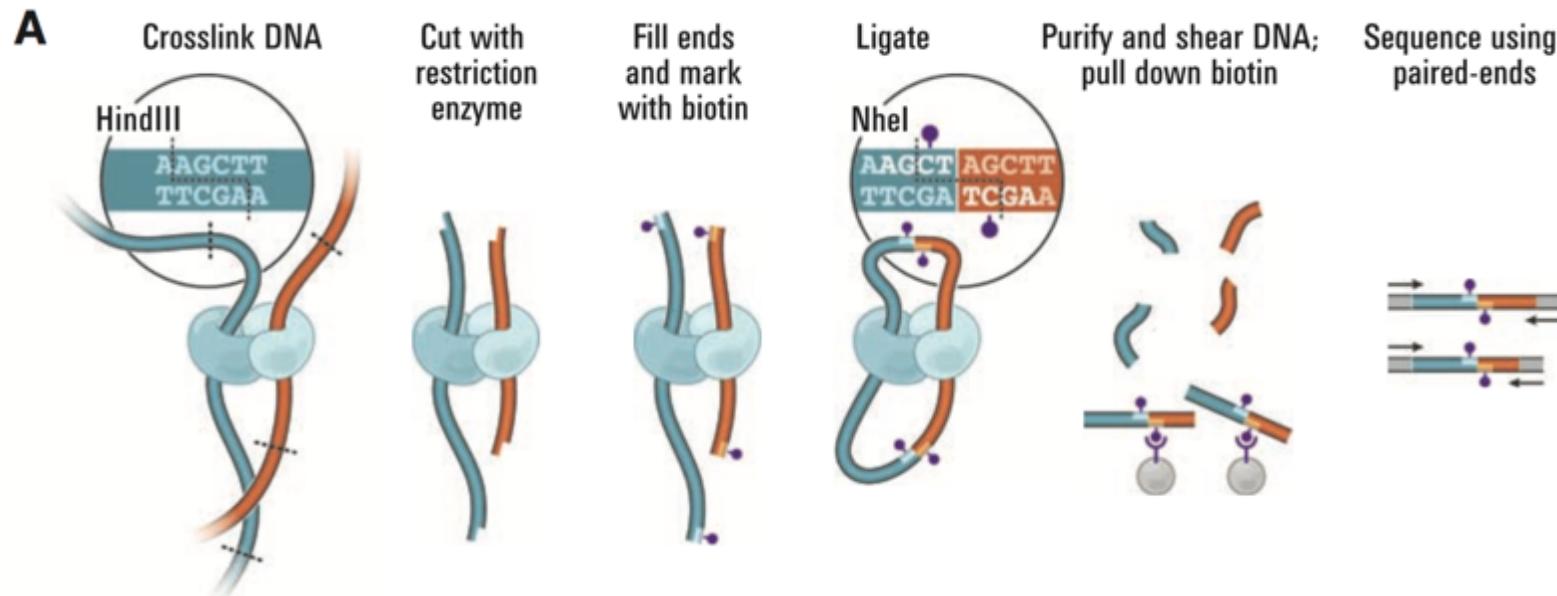
Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

<http://assets.illumina.com/content/dam/illumina-marketing/images/technology/paired-end-sequencing-figure.gif>

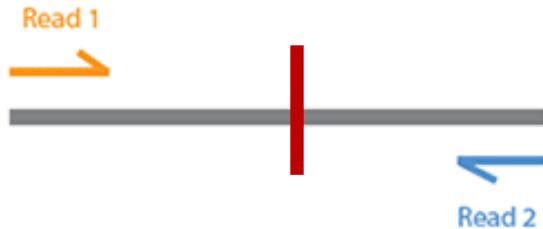
Hi-Cリードの特徴



Lieberman-Aiden, Erez, et al.
"Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* 326.5950 (2009): 289-293.

Hi-Cライブラリの特徴

ライゲーションジャンクションは、インサートのどこにでも生じうる



...R1, R2 それぞれ、リード全体が
マッピング可能



...R1がキメラリードとなっている

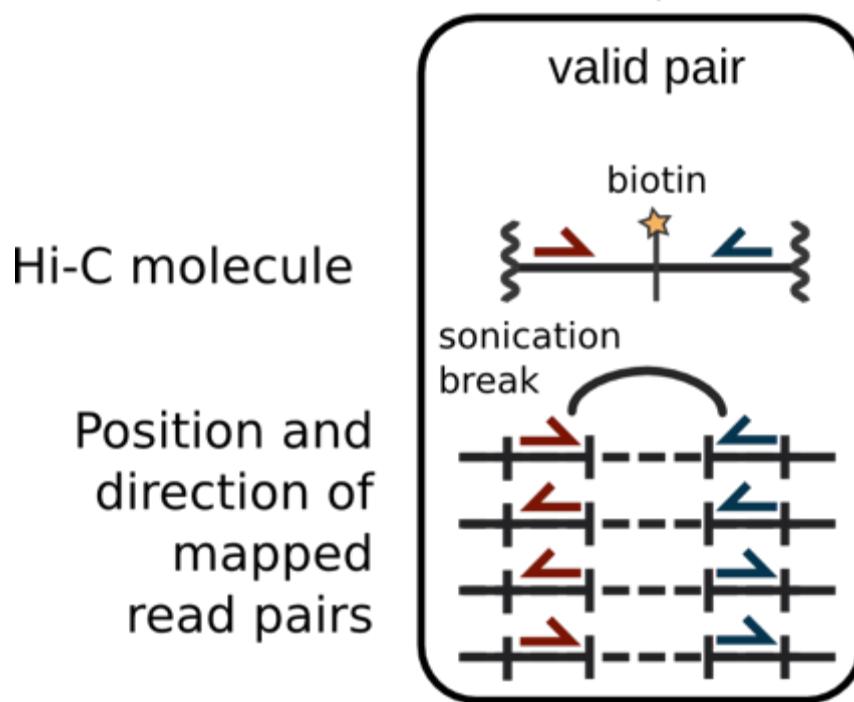


...R2がキメラリードとなっている

Hi-Cリードマッピングの際の注意点

1. キメラリードを考慮する
2. ペアのマッピング方向や、インサートサイズを仮定するようなマッピングはしない

=> R1, R2 それぞれ個別に、キメラを考慮しつつマッピングする



Imakaev, Maxim, et al.
"Iterative correction of Hi-C data
reveals hallmarks of chromosome
organization."
Nature methods 9.10 (2012): 999-1003.

マッピング戦略

1. R1, R2 個別にマッピングし、マッピング結果をパースして一対一の座標ペア情報をまとめる。

R1, R2のマッピングのパターンは以下の3通り

- I. R1, R2それぞれリード全体がマップされる

それぞれのマッピング位置間でコンタクトがあった、とみなして座標ペア情報を記録する。

- II. どちらかがキメラ

a. 一方がLocusA, LocusBのキメラ、もう一方がLocusB周辺の場合は、Locus A – B 間でコンタクトがあった、とみなして座標ペア情報を記録

b. 上記以外。破棄する。

- III. 両方キメラ

破棄する。

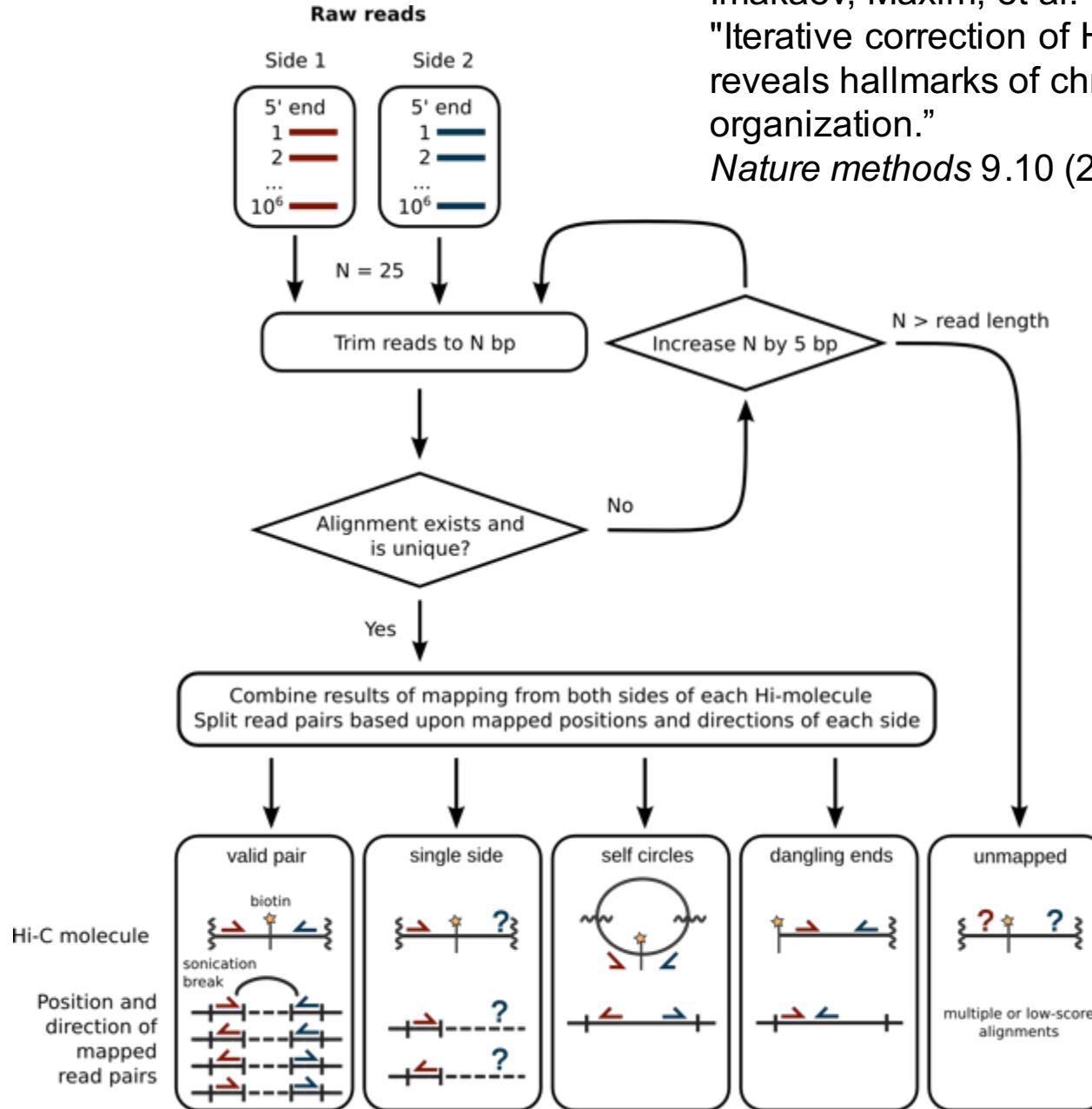
2. Iterative alignment method (今回はこっちの手法)

Iterative alignment method

Imakaev, Maxim, et al.

"Iterative correction of Hi-C data reveals hallmarks of chromosome organization."

Nature methods 9.10 (2012): 999-1003.



\$less mapping.py

```
#!/usr/bin/env python

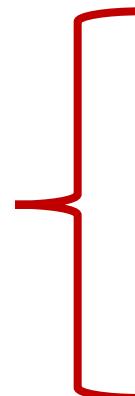
import os
import logging
from hiclib import mapping
from mirnylib import h5dict, genome
    Documentation
logging.basicConfig(level=logging.DEBUG)

if not os.path.exists('../data/tmp'):
    os.mkdir('../data/tmp')
        share

# Map the reads iteratively.
mapping.iterative_mapping(
    bowtie_path='/usr/bin/bowtie2',
    bowtie_index_path='../Index/hg19',
    fastq_path='../data/SRR1658595_10M_1.fastq',
    out_sam_path='../data/SRR1658595_10M_1.bam',
    min_seq_len=25,
    len_step=5,
    seq_start=0,
    seq_end=35,
    nthreads=2,
    temp_dir='../data/tmp',
    bowtie_flags='--very-sensitive')

mapping.iterative_mapping(
    bowtie_path='/usr/bin/bowtie2',
    bowtie_index_path='../Index/hg19',
    fastq_path='../data/SRR1658595_10M_2.fastq',
    out_sam_path='../data/SRR1658595_10M_2.bam',
    min_seq_len=25,
    len_step=5,
    seq_start=0,
    seq_end=35,
    nthreads=2,
    temp_dir='../data/tmp',
    bowtie_flags='--very-sensitive')
```

R1をマッピング



R2をマッピング



```
# Map the reads iteratively.  
mapping.iterative_mapping(  
    bowtie_path='/usr/bin/bowtie2',  
    bowtie_index_path='..../Index/hg19',  
    fastq_path='..../data/SRR1658595_10M_1.fastq',  
    out_sam_path='..../data/SRR1658595_10M_1.bam',  
    min_seq_len=25,  
    len_step=5,  
    seq_start=0,  
    seq_end=35,  
    nthreads=2,  
    temp_dir='..../data/tmp',  
    bowtie_flags='--very-sensitive')
```

初期ステップのトリミング長

次ステップで何bp延長するか

今回は、時間の関係上35bpまでで打ち切り。本当はリード全長に達するまでやる。

\$python mapping.py

```
manager@bl8vbox[1_mapping_read_to_genome] python mapping.py [12:14AM]
hello from new mapping
INFO:hiclib.mapping:Using new argument: max_len = 9999
/usr/bin/samtools
INFO:hiclib.mapping:The length of whole sequences in the file: 101
INFO:hiclib.mapping:Reading command: cat /home/manager/HiC/data/SRR1658595_10M_
.fastq.25.fastq.gz
INFO:hiclib.mapping:Mapping command: /usr/bin/bowtie2 -x /home/manager/HiC/Inde
/hg19 -q - -5 0 -3 76 -p 2 --very-sensitive
INFO:hiclib.mapping:Output editing command: awk {OFS="\t"; if ($1 ~ !/^@/) { $1
="A"; $11="g"; if ($3 ~ /\*/) $6="*"; else $6="1M"; } print}
INFO:hiclib.mapping:Output formatting command: samtools view -bs -
[samopen] SAM header is present: 25 sequences.
```

```
10000000 reads; of these:
 10000000 (100.00%) were unpaired; of these:
   454775 (4.55%) aligned 0 times
   6009135 (60.09%) aligned exactly 1 time
   3536090 (35.36%) aligned >1 times
95.45% overall alignment rate
```

第一ラウンドBowtie2結果

```
INFO:hiclib.mapping:Save the unique alignments and send the non-unique ones to th
e next iteration
/home/manager/HiC/src/mirnylab-mirnylib-a7ba48a06b92/mirnylib/systemutils.py:45
UserWarning: Please install 'pigz' parallel gzip for faster speed
  warnings.warn("Please install 'pigz' parallel gzip for faster speed")
INFO:mirnylib.systemutils:Writer created with command "[u'gzip', u'-c', u'-1']"
INFO:hiclib.mapping:4023426 non-unique reads out of 10000000 are sent the next i
teration.
INFO:hiclib.mapping:Using new argument: max_len = 9999
/usr/bin/samtools
/bin/gunzip
INFO:hiclib.mapping:The length of whole sequences in the file: 101
INFO:hiclib.mapping:Reading command: gunzip -c /home/manager/HiC/data/tmp/SRR16
8595_10M_1.fastq.25.fastq.gz
INFO:hiclib.mapping:Mapping command: /usr/bin/bowtie2 -x /home/manager/HiC/Inde
/hg19 -q - -5 0 -3 71 -p 2 --very-sensitive
INFO:hiclib.mapping:Output editing command: awk {OFS="\t"; if ($1 ~ !/^@/) { $1
="A"; $11="g"; if ($3 ~ /\*/) $6="*"; else $6="1M"; } print}
INFO:hiclib.mapping:Output formatting command: samtools view -bs -
[samopen] SAM header is present: 25 sequences.
4023426 reads; of these:
 4023426 (100.00%) were unpaired; of these:
```

第二ラウンドBowtie2結果

結果

\$ls -l ..data

```
manager@bl8vbox[1_mapping_read_to_genome] ls -l ..data [ 1:17AM]
total 6383416
drwxrwxr-x 2 manager manager 4096 Aug 23 11:17 Results
-rw-rw-r-- 1 manager manager 187913831 Aug 30 00:27 SRR1658595_10M_1.bam.25
-rw-rw-r-- 1 manager manager 76492558 Aug 30 00:36 SRR1658595_10M_1.bam.30
-rw-rw-r-- 1 manager manager 71634155 Aug 30 00:45 SRR1658595_10M_1.bam.35
-rw-r--r-- 1 manager manager 2929472122 Aug 23 06:06 SRR1658595_10M_1.fastq
-rw-rw-r-- 1 manager manager 186074628 Aug 30 00:57 SRR1658595_10M_2.bam.25
-rw-rw-r-- 1 manager manager 80347350 Aug 30 01:05 SRR1658595_10M_2.bam.30
-rw-rw-r-- 1 manager manager 75173151 Aug 30 01:15 SRR1658595_10M_2.bam.35
-rw-r--r-- 1 manager manager 2929472122 Aug 23 06:06 SRR1658595_10M_2.fastq
drwxr-xr-x 2 manager manager 4096 Aug 30 01:15 tmp
```

個別にマッピングした結果を統合する

\$less parse_results.py

```
#!/usr/bin/env python

import logging
from hiclib import mapping
from mirnylib import h5dict, genome
    Bio-Linux
logging.basicConfig(level=logging.DEBUG)

mapped_reads = h5dict.h5dict('./mapped_reads.hdf5')
genome_db = genome.Genome('../Ref/hg19', readChroms=['#', 'X'])

mapping.parse_sam(
    sam_basename1='../data/SRR1658595_10M_1.bam',
    sam_basename2='../data/SRR1658595_10M_2.bam',
    out_dict=mapped_reads,
    genome_db=genome_db,
    enzyme_name='MboI')
```

統合結果を出力するファイル
(HDF5形式)

ゲノムオブジェクト
のロード。どの染色
体を使うかを指定。

ゲノムの制限酵素消化断片にマッピングされたリードを
アサインするため、実験で用いた制限酵素を指定する。
指定できる制限酵素は、BiopythonのRestrictionクラス

個別にマッピングした結果を統合する

\$python parse_results.py

\$ls -l

```
manager@bl8vbox[1_mapping_read_to_genome] ls -l [ 1:26AM]
total 425060
-rw-rw-r-- 1 manager manager 435249141 Aug 30 01:26 mapped_reads.hdf5
-rw-r--r-- 1 manager manager      951 Aug 23 06:18 mapping.py
-rw-r--r-- 1 manager manager     454 Aug 23 06:09 parse_results.py
drwxrwxr-x 2 manager manager    4096 Aug 23 11:16 Results
```

HDF5はバイナリファイルなので、中身を見たい場合はHDFViewなどのツールを使うか、pythonのHDF5モジュールなどで開く。
少なくともどちらかのリードがマッピングされたペアについて、座標情報などが格納されている。

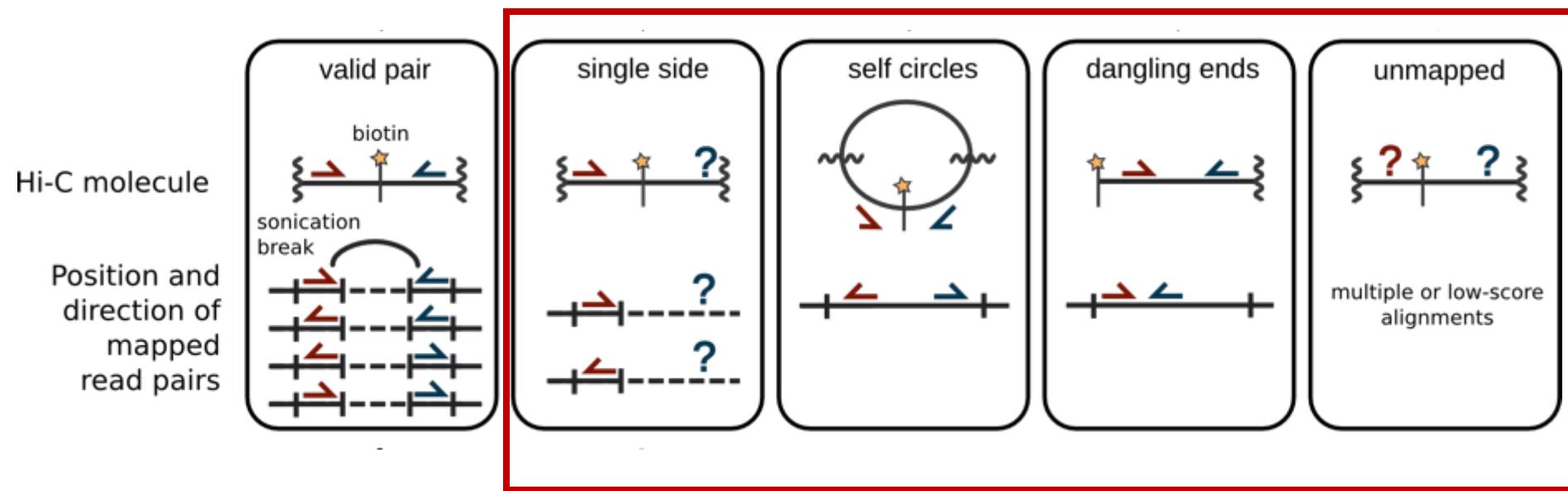
Hi-C解析の流れ



マッピングされたペア情報のフィルタリング

R1, R2 がマッピングされたすべてのペアが、Hi-Cとしての妥当な情報を持つわけではない。

以下のようなパターンでマッピングされたペアは「コンタクト」の情報を持たないため除去する。



Imakaev, Maxim, et al.
"Iterative correction of Hi-C data
reveals hallmarks of chromosome
organization."
Nature methods 9.10 (2012): 999-1003.

\$less filtering.py

```
#!/usr/bin/env python

from mirnylib import genome
from hiclib import fragmentHiC

genome_db = genome.Genome('../Ref/hg19', readChrms=['#', 'X'])
genome_db.setEnzyme('MboI')

fragments = fragmentHiC.HiCdataset(
    filename='./fragment_dataset.hdf5',
    genome=genome_db,
    maximumMoleculeLength=500,
    mode='w')

fragments.parseInputData(
    dictLike='../1_mapping_read_to_genome/mapped_reads.nats')

fragments.filterRsiteStart(offset=5)
fragments.filterDuplicates()

fragments.filterLarge()
fragments.filterExtreme(cutH=0.005, cutL=0)

fragments.saveHeatmap('./heatmap-res-1M.hdf5', resolution=1000000)
fragments.printMetadata(saveTo='./statistics.txt')
```

ゲノムデータのロード

出力ファイルの指定
maximumMoleculeLengthは、イル
ミナライブラリの断片長の情報から
設定（今回の実験では400bp）
隣接した制限断片が再びライゲー
ションした場合のペアを除去するた
めに使う

さきほど作ったHDF5ファイ
ルのロード

\$less filtering.py

```
#!/usr/bin/env python

from mirnylib import genome
from hiclib import fragmentHiC

genome_db = genome.Genome('../Ref/hg19')
genome_db.setEnzyme('MboI')

fragments = fragmentHiC.Hicdataset(
    filename='./fragment_dataset.hdf5',
    genome=genome_db,
    maximumMoleculeLength=500,
    mode='w')

fragments.parseInputData(
    dictLike='../1_mapping_read_to_sam')

fragments.filterRsiteStart(offset=5)
fragments.filterDuplicates()

fragments.filterLarge()
fragments.filterExtreme(cutH=0.005, cutL=0)

fragments.saveHeatmap('./heatmap-res-1M.hdf5', resolution=1000000)
fragments.printMetadata(saveTo='./statistics.txt')
```

追加で行うフィルタリング

filterRsiteStart():

おそらくライゲーションに失敗した
DNA断片

filterDuplicates():

ペアのどちらも同一の座標にマッピングされる2つのペアはPCR duplicateの可能性が非常に高いため、除去する

filterLarge():

10^5bp以上の制限断片にマップされるペアを除去。（リピート領域など、アセンブル精度が低い領域）

filterExtreme():

マップされるリード数がトップ0.5%の制限断片を除去。アーティファクトの可能性が高い（元論文参照）

\$less filtering.py

```
#!/usr/bin/env python

from mirnylib import genome
from hiclib import fragmentHiC

genome_db = genome.Genome('../Ref')
genome_db.setEnzyme('MboI')

fragments = fragmentHiC.HiCdataset(
    filename='./fragment_dataset',
    genome=genome_db,
    maximumMoleculeLength=500,
    mode='w')

fragments.parseInputData(
    dictLike='../1_mapping_read_t'

fragments.filterRsiteStart(offset)
fragments.filterDuplicates()

fragments.filterLarge()
fragments.filterExtreme(cutH=0.005, cutL=0)

fragments.saveHeatmap('./heatmap-res-1M.hdf5', resolution=1000000)
fragments.printMetadata(saveTo='./statistics.txt')
```

ゲノムを1MbpごとのBinに分割。
各制限断片上のマッピングされたリード数の情報から、raw read countのコンタクトマップ（ゲノム対称行列）としてデータをまとめます。

1Mbpで分割する場合、ヒトゲノムなら約3,000 × 約3,000 のサイズのマップとなる。

Binサイズの決定に定量的な基準はない。引き出したい生物学的解釈によって適切に決めるしかない。

（「高品質」なコンタクトマップの基準としてたとえば：マトリックス内の90%以上の値が非ゼロであること、かつ、80%以上で1000以上のコンタクトがあること）

フィルタリングを実行する

\$python filtering.py

\$ls -l

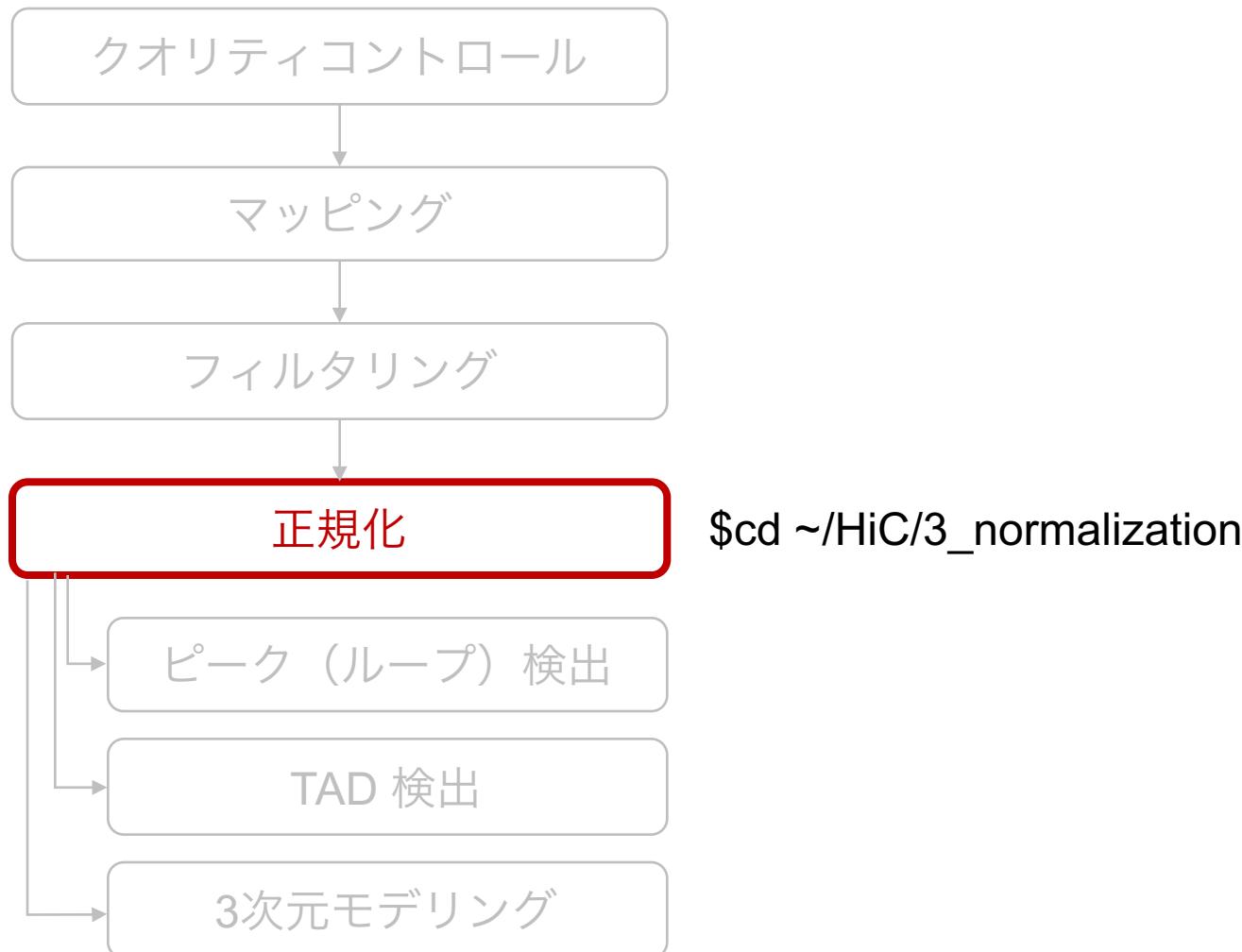
```
manager@bl8vbox[2_filtering_reads] ls -l [ 2:10AM]
total 71848
-rw-r--r-- 1 manager manager      661 Aug 23 07:32 filtering.py
-rw-rw-r-- 1 manager manager 68259820 Aug 30 01:37 fragment_dataset.hdf5
-rw-rw-r-- 1 manager manager 5298133 Aug 30 01:37 heatmap-res-1M.hdf5
drwxrwxr-x 2 manager manager     4096 Aug 23 11:15 Results
-rw-rw-r-- 1 manager manager      582 Aug 30 01:37 statistics.txt
```

filtreringを結果を見る

\$less ./statistics.txt

```
010_MappedSide1: 6465654
020_MappedSide2: 6157970
100_TotalReads: 8361390
    150_ReadsWithoutUnusedChromosomes: 8361390
        152_removedUnusedChromosomes: 0
200_totalDSReads: 4262234
    201_DS+SS: 8361390
        Bio-Linux 202_SSReadsRemoved: 4099156
        Documentation 210_sameFragmentReadsRemoved: 282673
            212_Self-Circles: 2527
            214_DandlingEnds: 269690
            216_error: 10456
        220_extraDandlingEndsRemoved: 214820
300_ValidPairs: 3764741
    310_startNearRsiteRemoved: 330988
    320_duplicatesRemoved: 3521
    340_removedLargeSmallFragments: 166480
    350_removedFromExtremeFragments: 98211
```

Hi-C解析の流れ



データ正規化の必要性

1. サンプル間で比較する場合、サンプルごとにライブラリサイズが異なる。マッピングされたリードの数（サンプルのクオリティ）も異なる。
2. ゲノムの領域ごと、さらには領域間によっても、コンタクトが観測される確率が異なる

Hi-C実験は様々なバイアスの影響で、ある領域間のペアが観測されやすかったりされにくかったりする。

- I. 制限酵素断片の長さ。両方とも長い断片の場合、両方とも短い場合、あるいは長い断片と短い断片のペアはライゲーションが起きにくい。共に中間的な長さの場合にLigationされやすい。
- II. 制限酵素断片のGC含量。シーケンシングのバイアス（読み取られやすさ）にばらつきがある。
- III. Mappability. マッピングされるリードのゲノム中の「ユニークさ」。その領域がゲノム上でユニークな塩基配列であるかに依存する。

他の実験ではどうやって正規化しているか？

ChIP-seq: INPUTのデータで割り算

RNA-seq: そもそも1サンプル単独で評価しない。サンプル間の比較。

Hi-C実験にはコントロールがないことが問題。

Hi-C正規化の方法

1. Explicitにバイアスを仮定する手法

制限酵素断片長、GC含量、マッパビリティなど、バイアスを列挙（それぞれゲノム配列のみから計算可能）、領域ペアの観測確率をそれらのバイアスすべてをパラメータとした確率モデル（ポアソン、負の二項分布など）で表現し、観測値からバイアスパラメータを学習する。

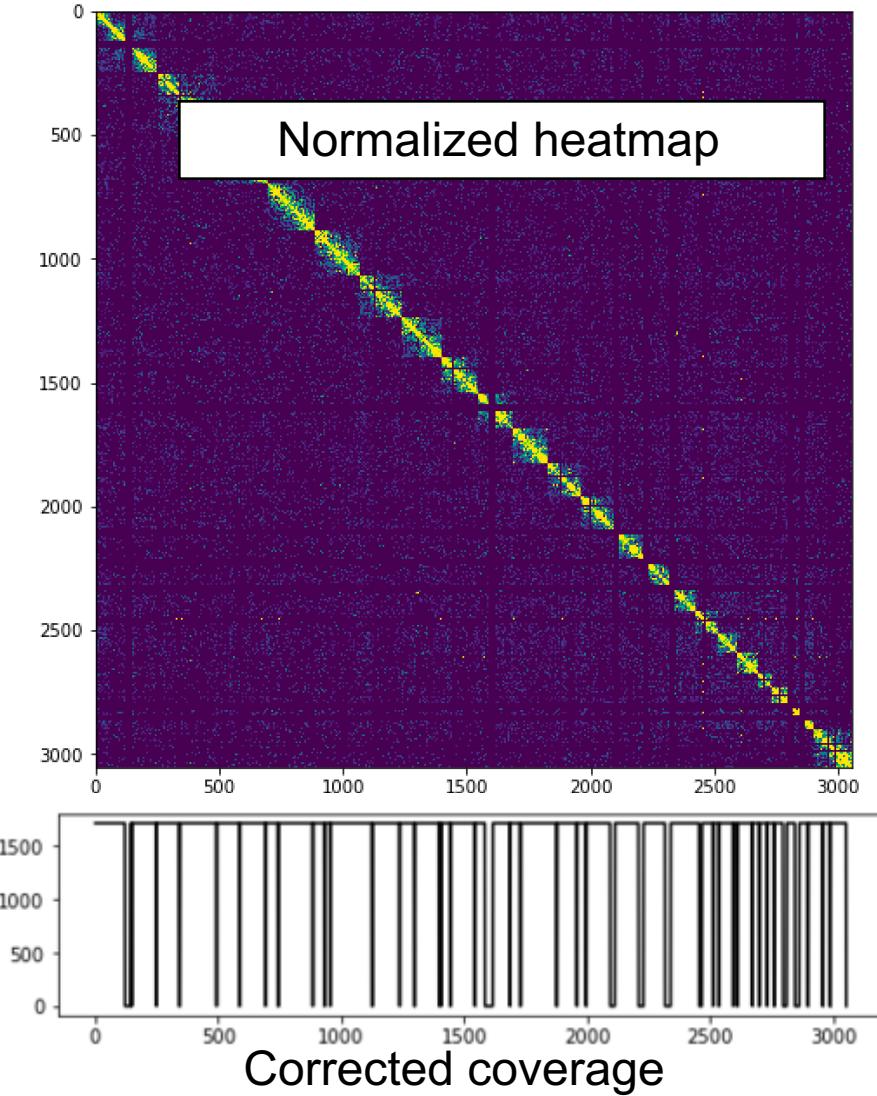
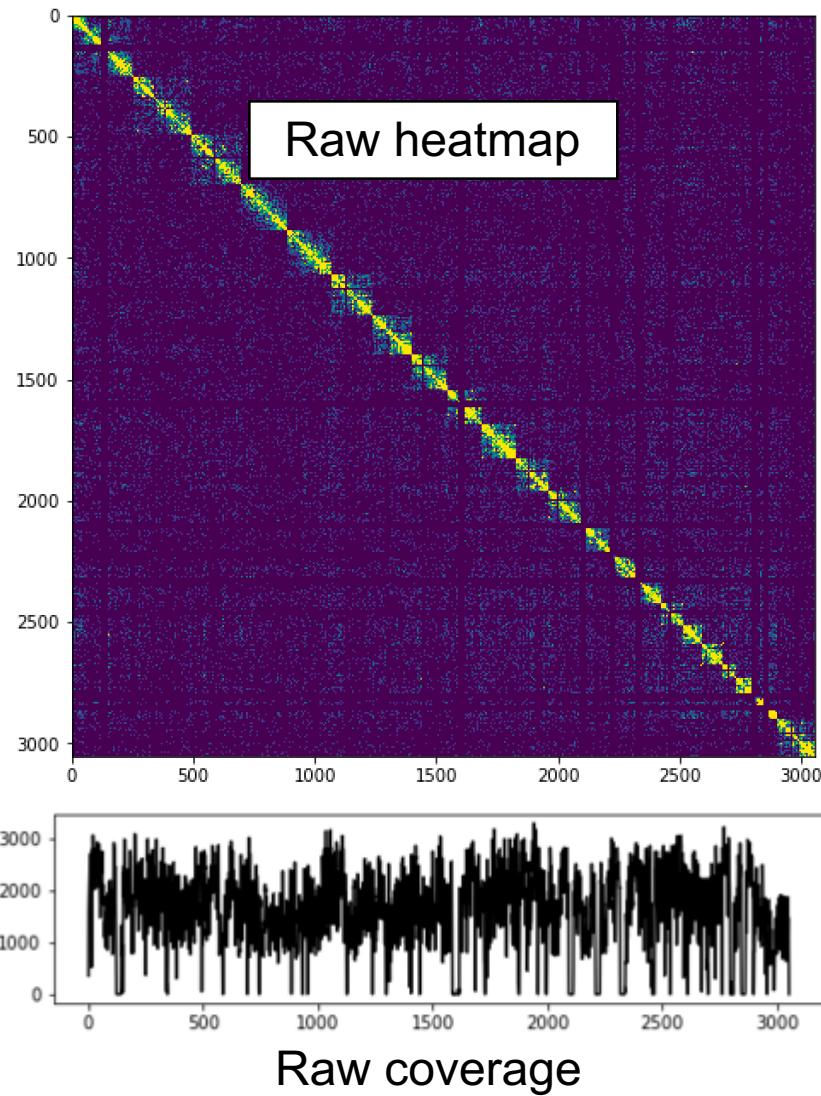
Yaffe and Tanay 2011、HiCNormなど

2. Implicitにバイアスを仮定する手法

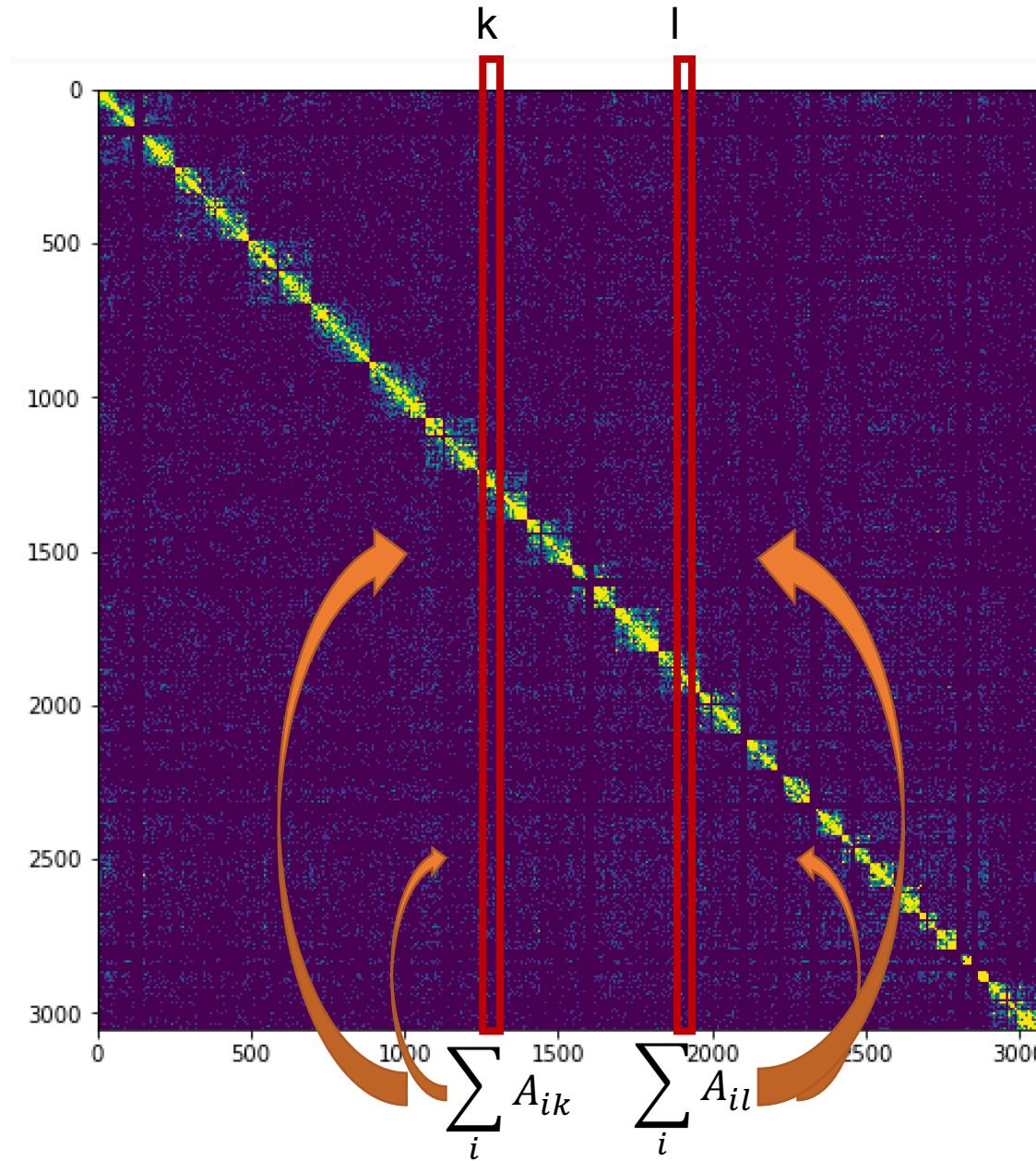
こちらの方が広く使われている。

Vanilla coverage, ICE, Knight and Ruiz 2012など

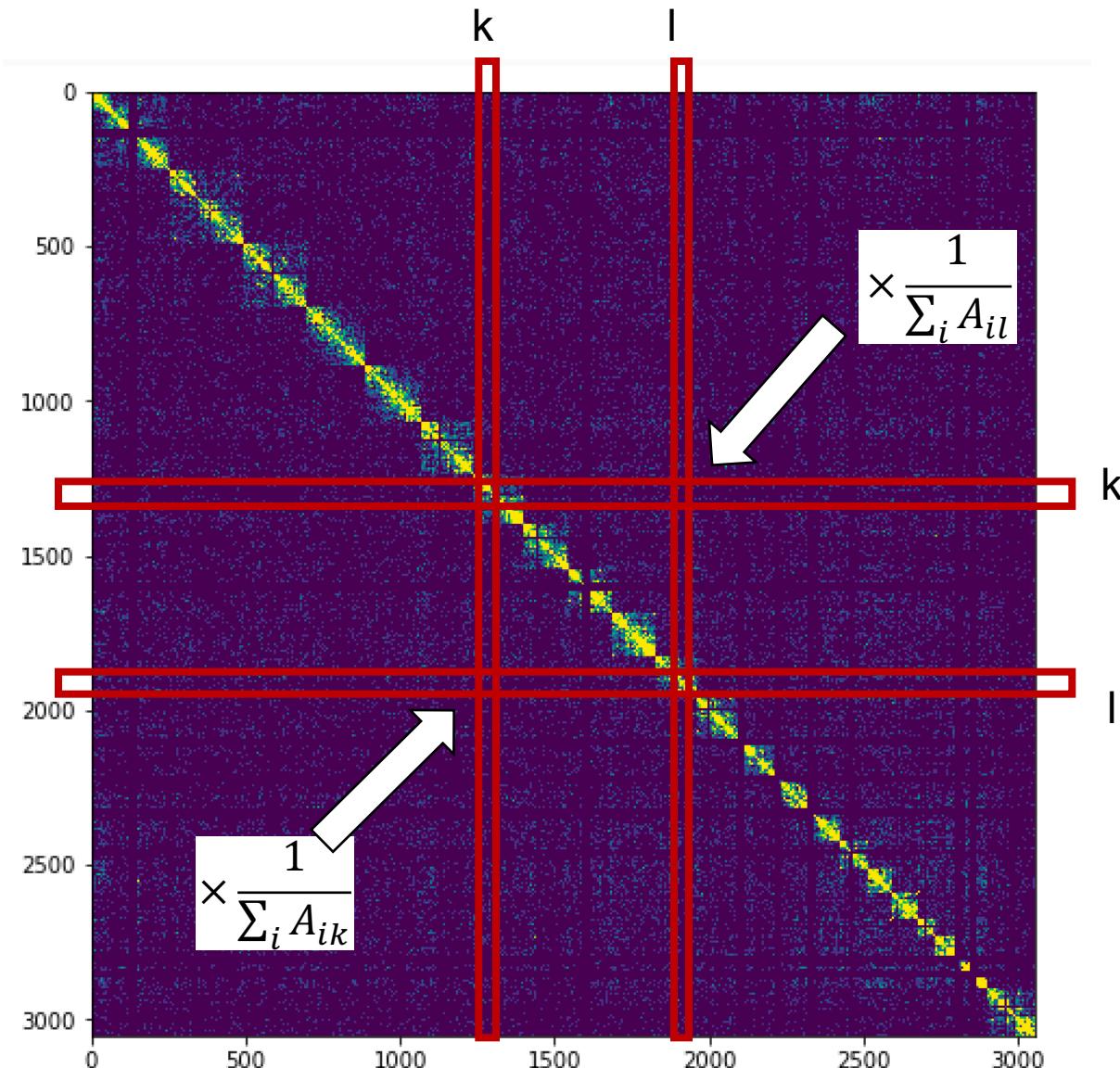
理想的な（正規化された）コンタクトマップでは、
ゲノム上のカバレッジが一定



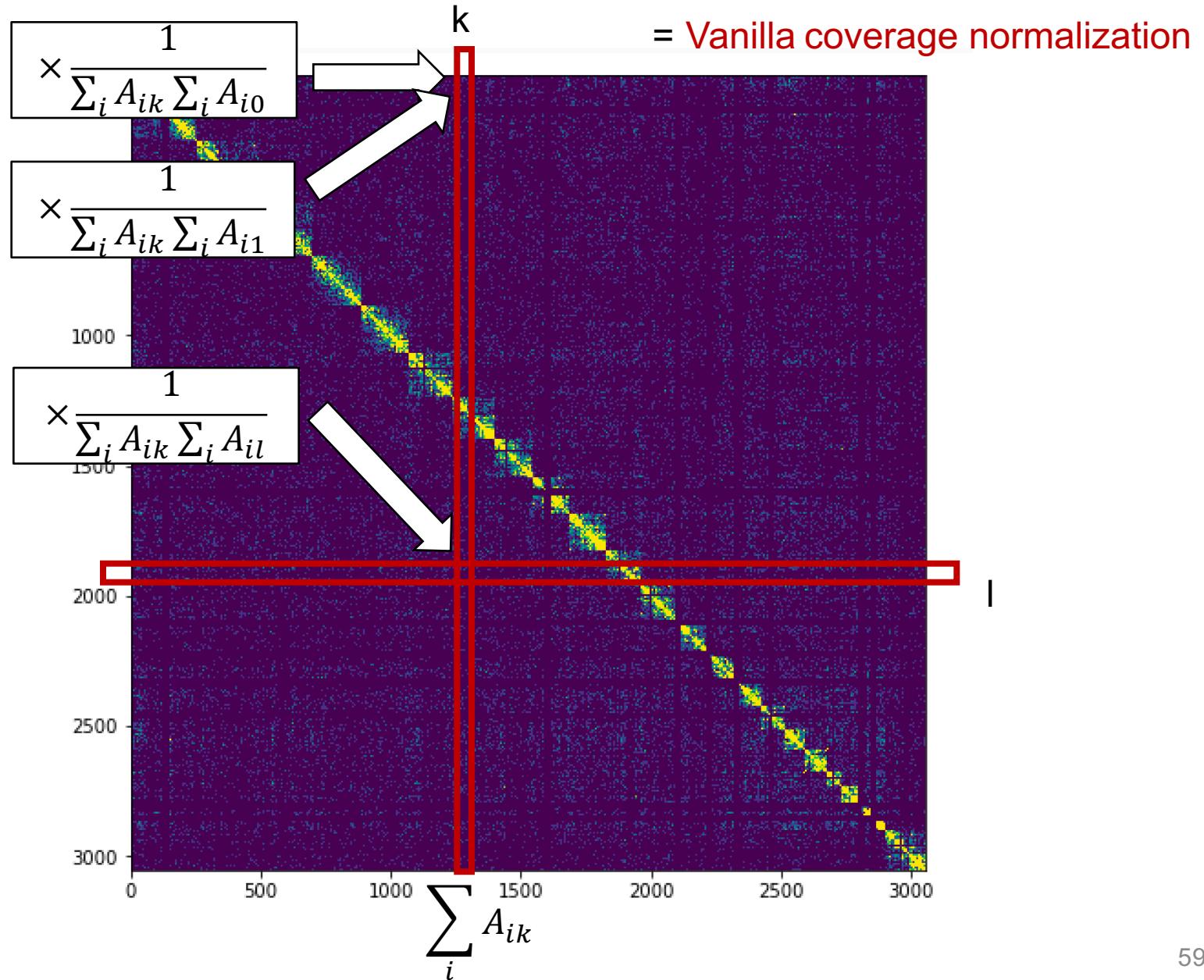
各列の和を計算して割り算すると…



行列の対称性が崩れてしまう



そこで、行の和と列の和の積で割り算する



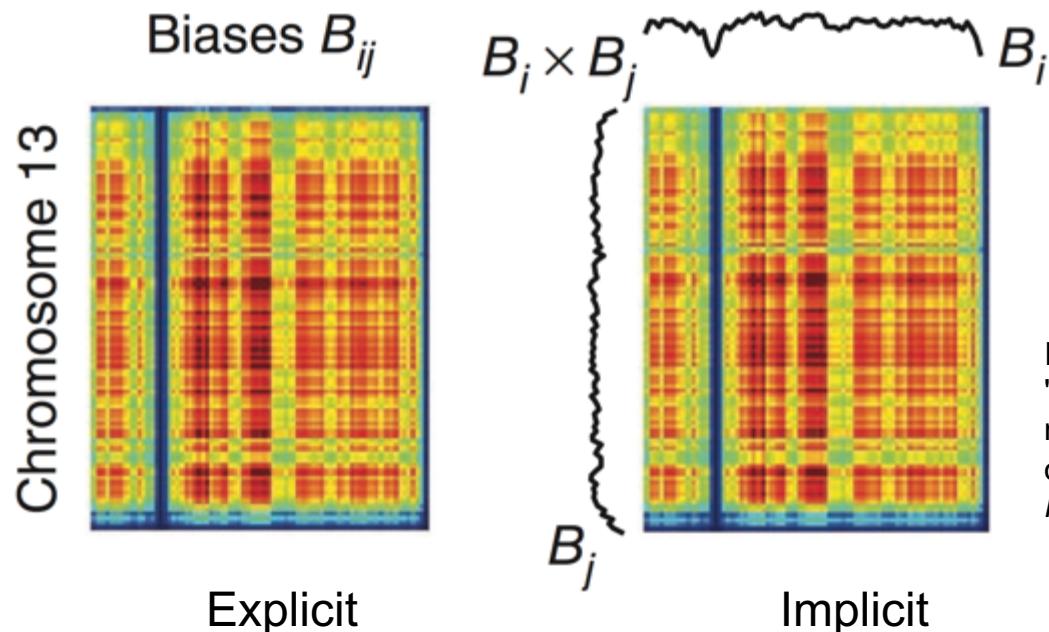
Vanilla coverage normalizationの仮定

領域 i と領域 j のペアを観測する際のバイアスは、
領域 i のバイアスと、領域 j のバイアスの積に比例する。
つまり、それぞれのバイアスが観測に独立に影響する、と仮定している。

各領域のバイアスは、GC含量やマッパビリティなど、様々な要因が重ね
合わさった結果として生じる複合的なバイアス (implicit bias)

強い仮定であるが、Explicit バイアスを仮定して推定した結果ときわめて
よく一致する。

Factorizable biases



Imakaev, Maxim, et al.
"Iterative correction of Hi-C data
reveals hallmarks of chromosome
organization."
Nature methods 9.10 (2012): 999-1003.

Iterative correction (ICE method)

単独の Vanilla coverage normalization は補正が強すぎる。
(和が非常に小さい列では、割り算結果が爆発する)

⇒ Vanilla coverage normalization を何回も適用し、行列全体が収束するまで計算する

このような行列の補正手法は、“matrix balancing”と呼ばれ、歴史的に何度も再発明されてきた。

ICEと同様の matrix balancing 手法だが、
より収束の早い Knight and Ruiz 2012 もよく使われる。

\$less normalize.py

```
#!/usr/bin/env python

import matplotlib.pyplot as plt
import numpy as np

from mirnylib import genome
from mirnylib import h5dict
from mirnylib import plotting
from hiclib import binnedData

genome_db = genome.Genome('..../Ref/hg19')
raw_heatmap = h5dict.h5dict('..../2_filtered_contacts.hdf5', mode='r')
resolution = int(raw_heatmap['resolution'])

BD = binnedData.binnedData(resolution,
                           simpleLoad='..../2_filtering_reads/heatmap-res-1M.hdf5',
                           RaO2014_10M=True)

#BD.removeDiagonal()
BD.removeBySequencedCount(0.5)
BD.removePoorRegions(cutoff=1)
BD.truncTrans(high=0.0005)
BD.iterativeCorrectWithoutSS()

BD.export('Rao2014_10M', './IC-heatmap-res-1M.hdf5')

fig = plt.figure()
plotting.plot_matrix(np.log(BD.dataDict['Rao2014_10M']+1.0))
fig.savefig('./heatmap.pdf')
```

ゲノムデータのロード

Raw read count コンタクトマップのロード

正規化に悪影響を与える可能性のあるBinを除去する。

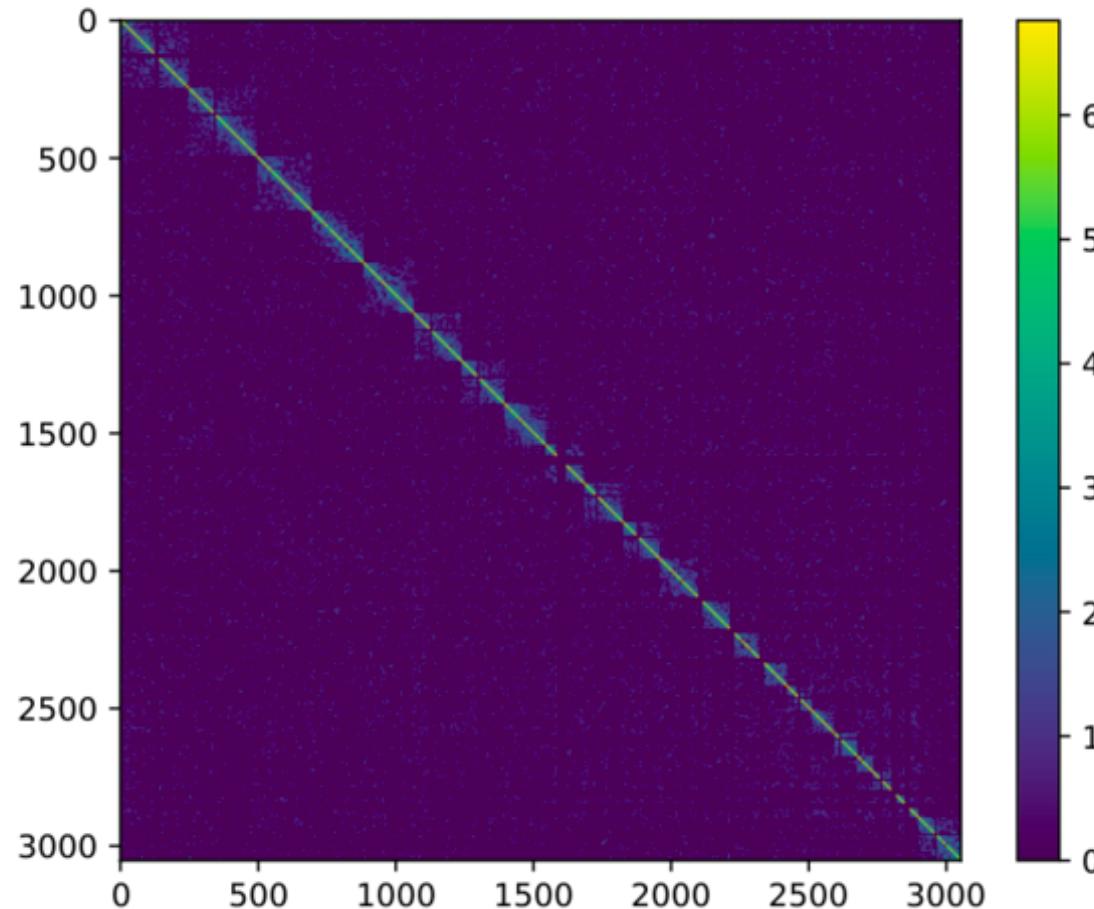
ICE正規化

正規化後のコンタクトマップを出力

正規化を実行

```
$python normalize.py
```

以下のような結果ファイル（heatmap.pdf）がで
きているはず。



あとでTAD検出、3Dモデリングに使用するため、
19番染色体の領域だけ切り出しておく。

\$python submatrix.py

\$less norm_mat.txt

JuiceBox で、コンタクトマップをインタラクティブに可視化して見る。

JuiceBoxは、独自形式で保存されたコンタクトマップデータを可視化するので、ここまで実習で生成したデータをJuiceBoxの形式に変換する必要がある。

全ゲノムを見るのはメモリ的にきついので、ここでは一番染色体だけ見る

```
$cd ~/4_convert_Juice  
$less convert_to_JuiceText.py  
$python convert_to_JuiceText.py  
$less ./forJuice.txt  
$./convert_to_JuiceHiC.sh
```

以上で、test.hic というバイナリファイル（Juice 形式のコンタクトマップを格納したファイル）ができるはず

JuiceBoxの実行

```
./execute_Juicebox.sh
```

File => Open => Local

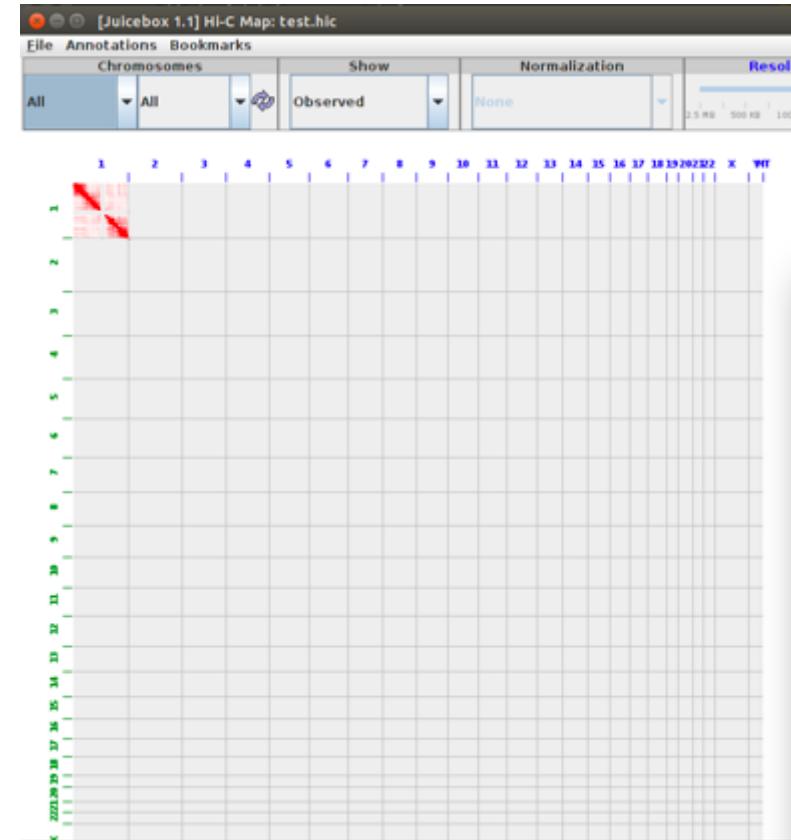
から、いま作成した test.hic を開く。

Chromosomesで拡大。

Annotationsから

ENCODEデータとの比較

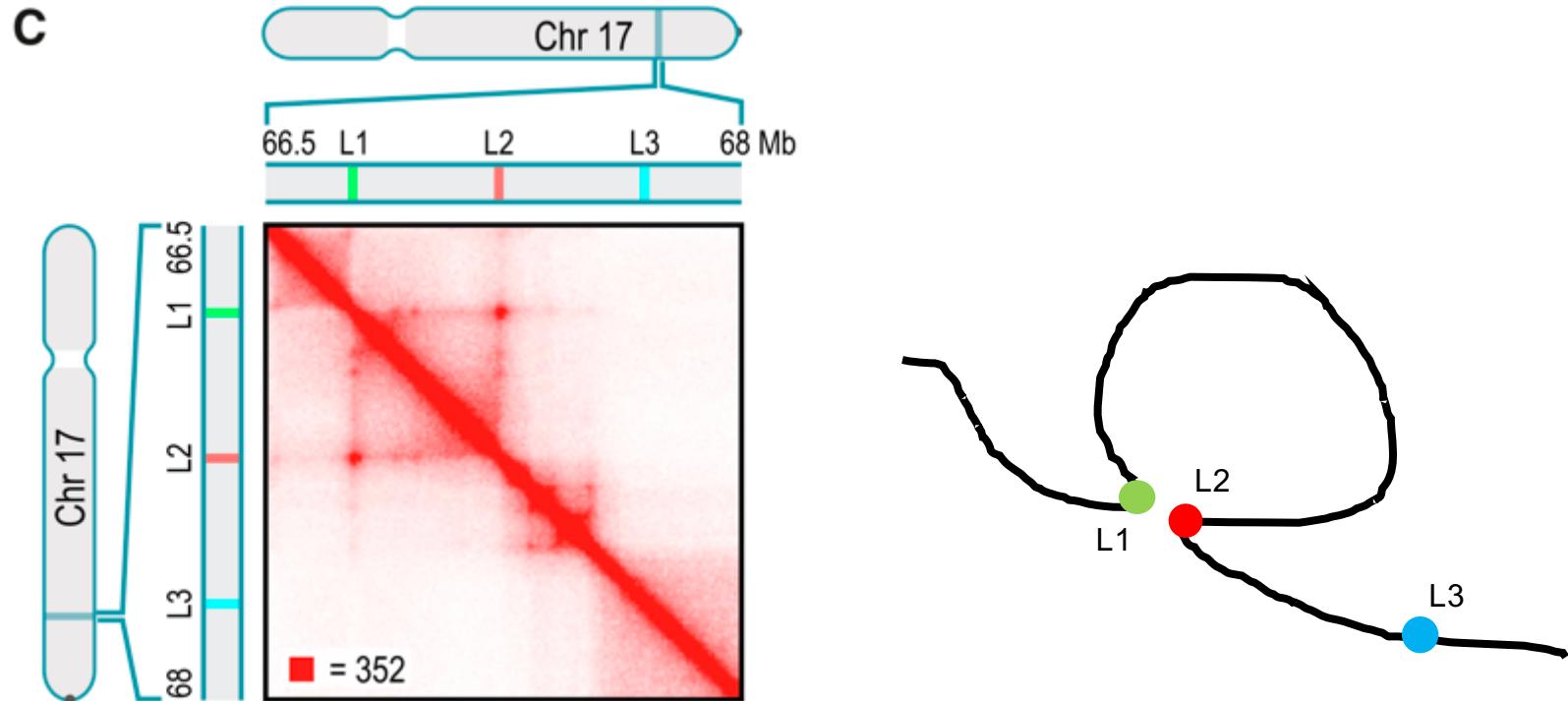
なども可能。



Hi-C解析の流れ

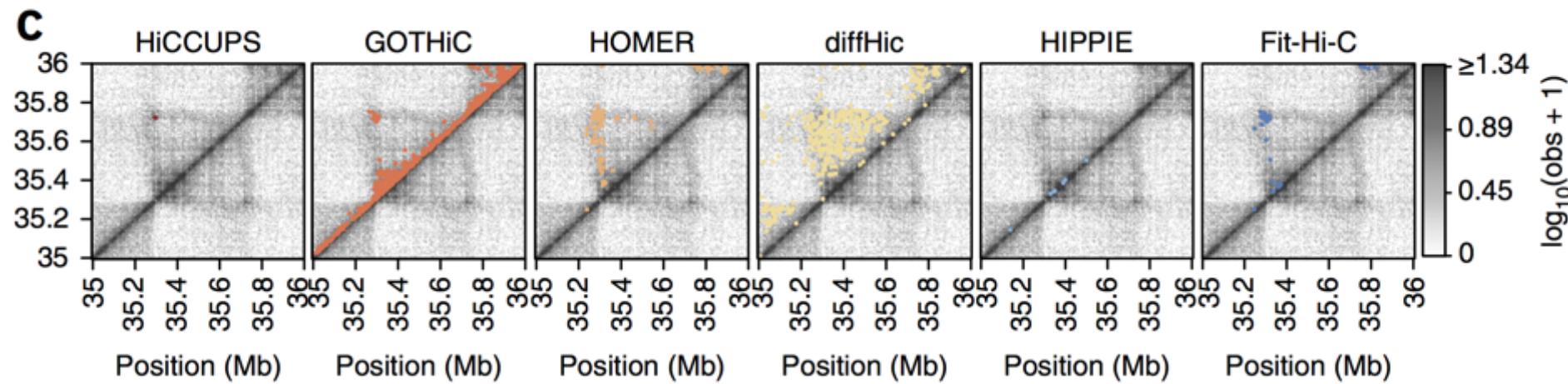


コンタクトマップ上のピーク検出 =特に相互作用の強い領域ペアを特定する



Rao, Suhas SP, et al.
"A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping"
Cell 159.7 (2014): 1665-1680.

ピーク検出手法によって、
得られるピークの数や位置は大きく異なる



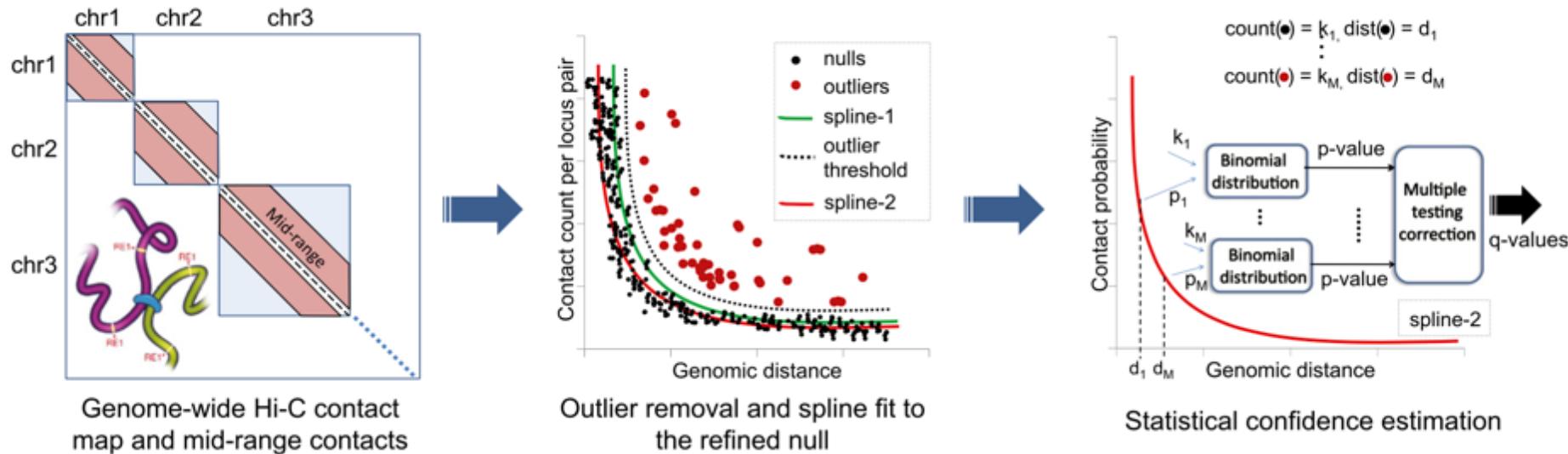
Forcato, Mattia, et al. "Comparison of computational methods for Hi-C data analysis." *Nature methods* 14.7 (2017): 679.

コンタクトマップの解像度も大きく影響する。

それぞれのピーク検出ツールが、どのように「バックグラウンド」を仮定しているかちゃんと理解することが重要。

Fit-Hi-C (Global background)

Ay, Ferhat, Timothy L. Bailey, and William Stafford Noble. "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts." *Genome research* 24.6 (2014): 999-1011.



ゲノム上の距離の関数として観測されたリードカウントをスプライン関数でモデリングする。

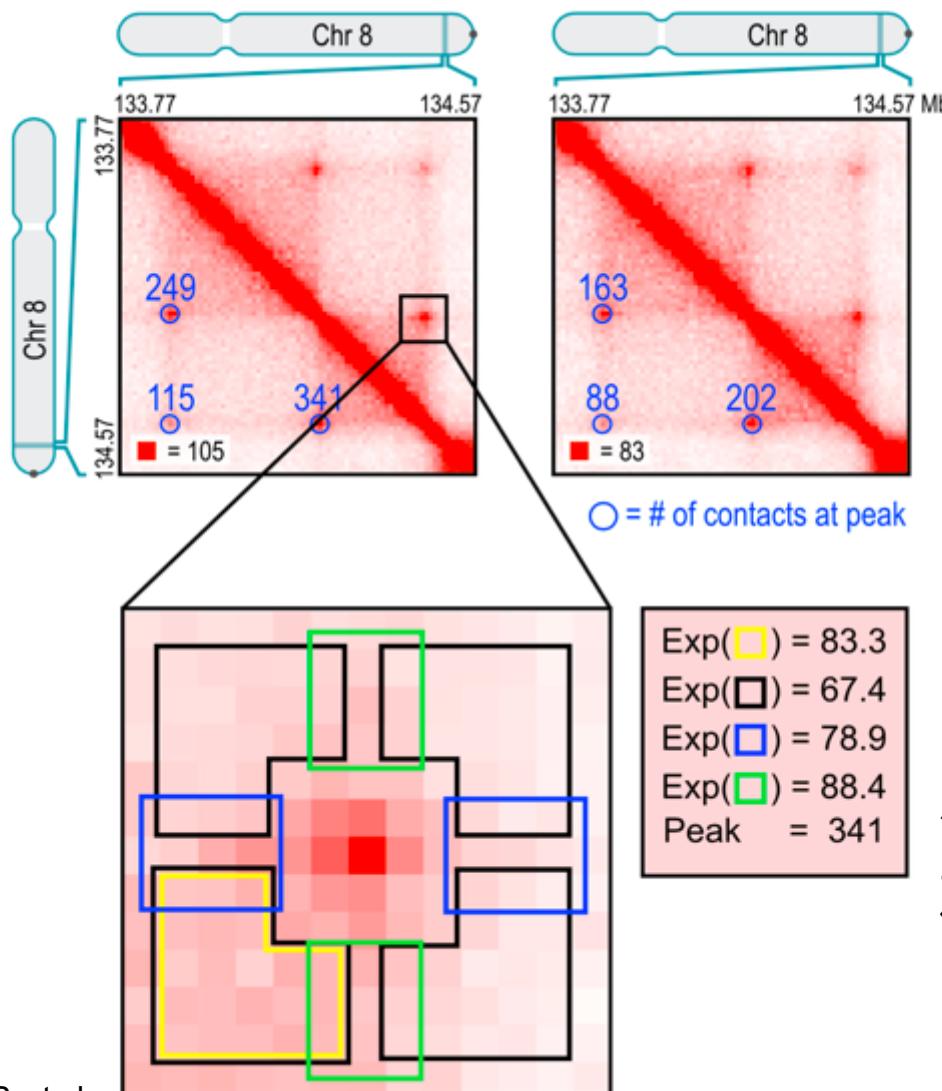
最初のスプラインは、外れ値を除去するために使われる。

その後、外れ値以外を使って、より洗練されたスプラインをモデリングする。これがヌルモデルとなる。

ヌルモデルの値（ある距離で期待されるリードカウント）を、ICE正規化手法で算出されたバイアスの値も加味して、ある距離のリードカウント観測期待値を計算する。

最後に、期待値と実際の観測値について、二項分布でp-valueを計算（多重検定補正）する。

HiCCUPS (Local background)



周辺の相互作用強度と、K&R
正規化で算出されたバイアス
値からポアソン分布のパラ
メータを計算し、ピーク位置
のp-valueを求める。

Rao, Suhas SP, et al.

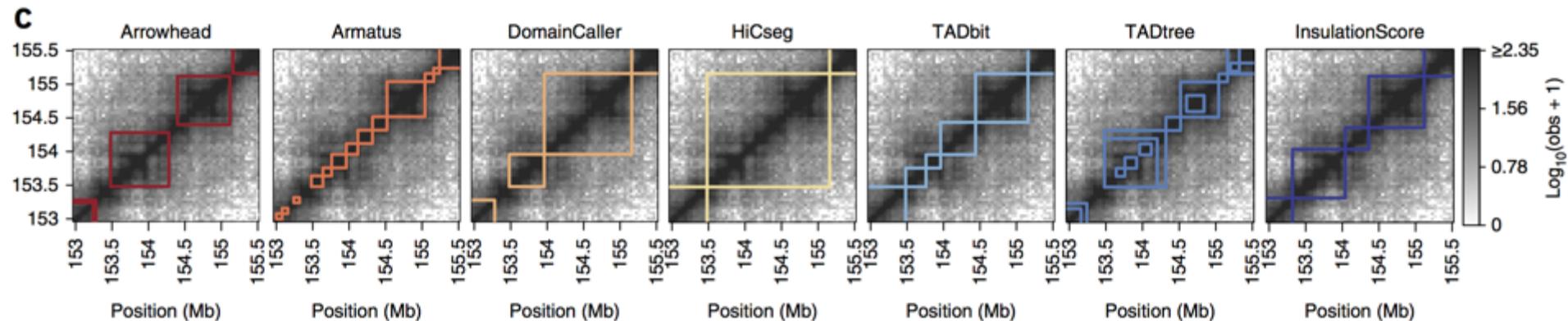
"A 3D map of the human genome at kilobase
resolution reveals principles of chromatin looping"

Cell 159.7 (2014): 1665-1680.

Hi-C解析の流れ



Topologically Associated Domains (TADs)の検出

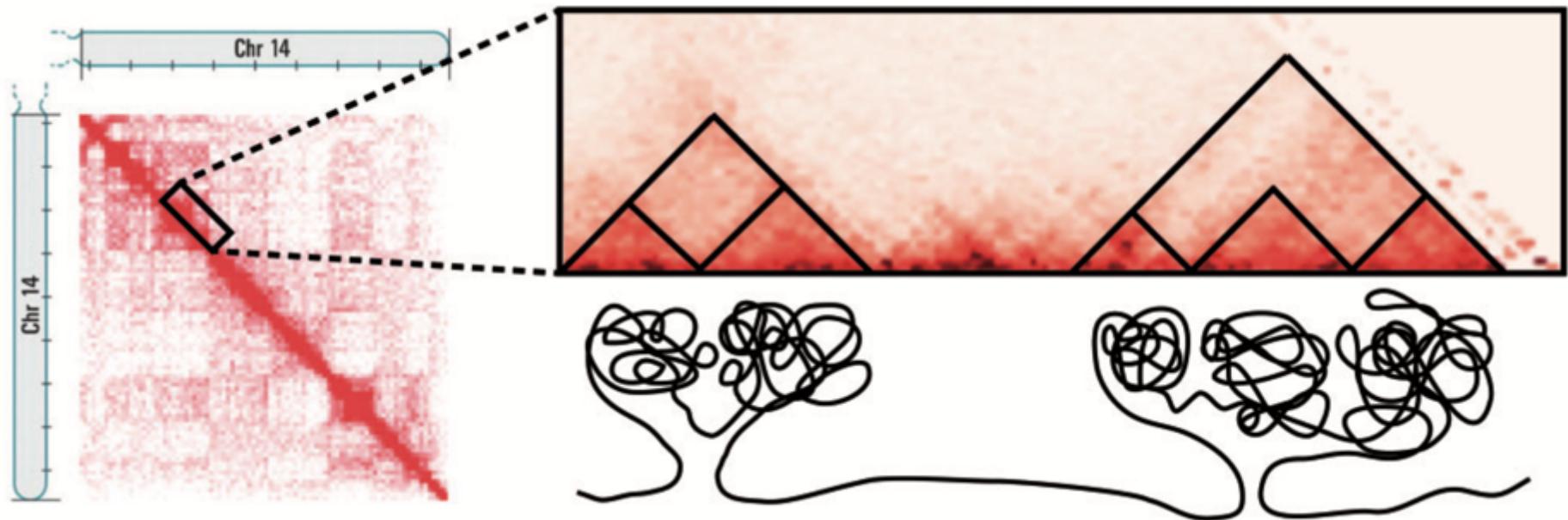


Forcato, Mattia, et al. "Comparison of computational methods for Hi-C data analysis." *Nature methods* 14.7 (2017): 679.

得られるTADのサイズや数はツールによってさまざま。
異なる解像度のコンタクトマップでも安定して同様のTADが得られるか、などを検討することが大事。

TADtree

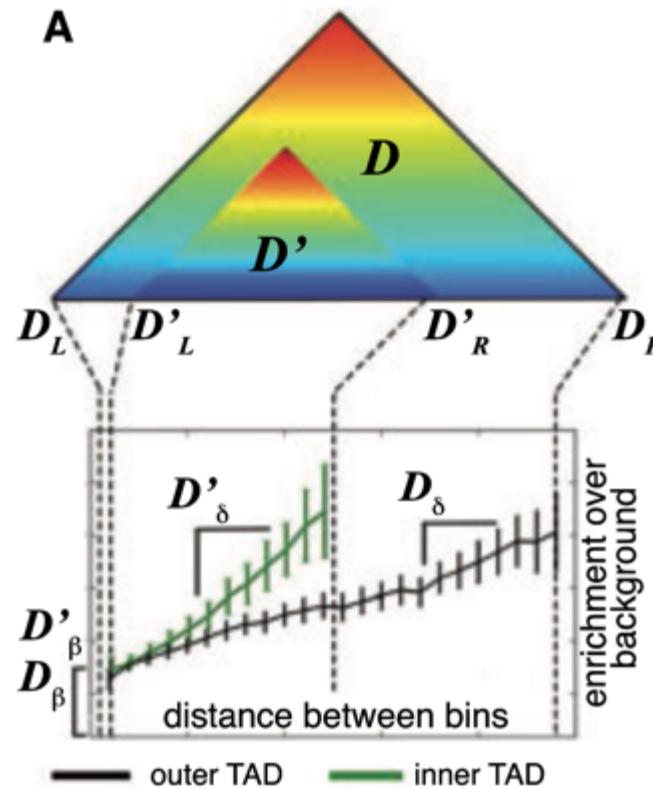
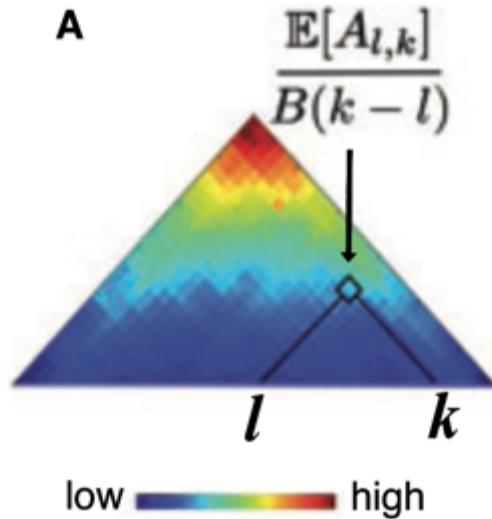
Caleb Weinreb, Benjamin J. Raphael; “Identification of hierarchical chromatin domains”, *Bioinformatics*, Volume 32, Issue 11, 1 June 2016, Pages 1601–1609



ネストしたTAD, sub-TADを検出するPythonスクリプト。
インプットは正規化したコンタクトマップ。

TADtree

Caleb Weinreb, Benjamin J. Raphael; “Identification of hierarchical chromatin domains”, *Bioinformatics*, Volume 32, Issue 11, 1 June 2016, Pages 1601–1609



TAD内のコンタクトは距離に線形に増加していくが、他のTAD内部に含まれている場合、その増加率が大きくなる、という観察結果に基づいた手法。動的計画法でベストなTAD階層構造を特定する。

```
$less ./control_file.txt
```

```
S = 50  
M = 10  
p = 3  
q = 12  
gamma = 500
```

```
contact_map_path = ../3_normalization/norm_mat.txt  
contact_map_name = chr19  
N = 100  
output_directory = ./output
```

TADの最大許容サイズ (Binいくつまでか)
高解像度データでは当然大きくすべき。

いじらないことが奨励されてるパラメータ

TAD境界に関するチューニングパラメータ。
あまりに多くのTADが同一のポイントから
生じている場合、おそらく値が大きすぎる。

検出するTADの最大数。

TADtreeを実行する。

```
$python TADtree.py ./control_file.txt
```

結果が、./output/chr19 以下に出力される。

設定したN以下の検出結果がすべて出力されるが、十分な数のTADを検出している結果が適切なため、proportion_duplicates.txt ファイルを参照して、重複したTADが検出され始めた時点の結果をチェックする。

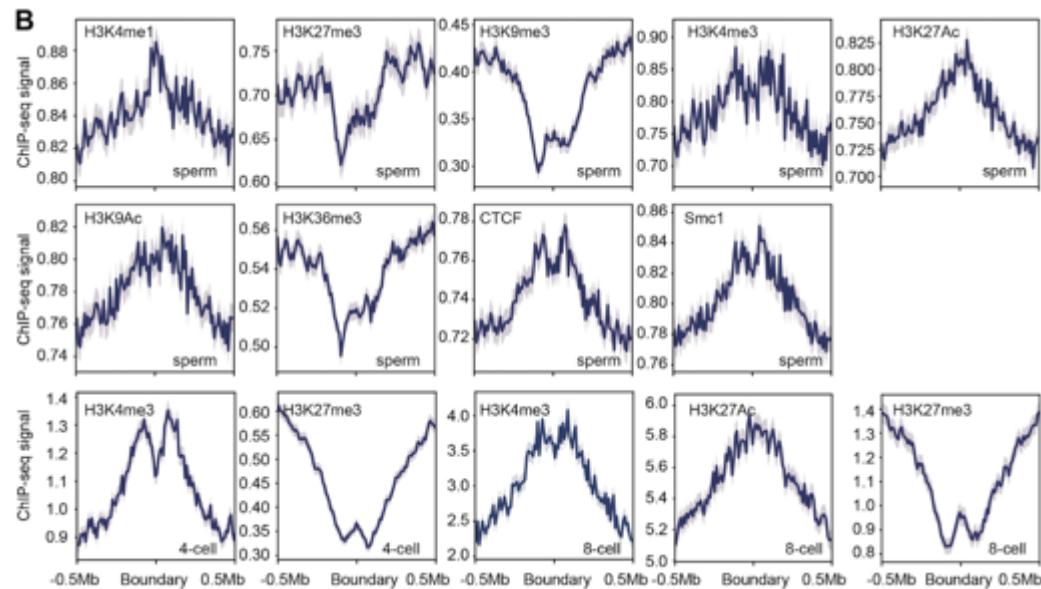
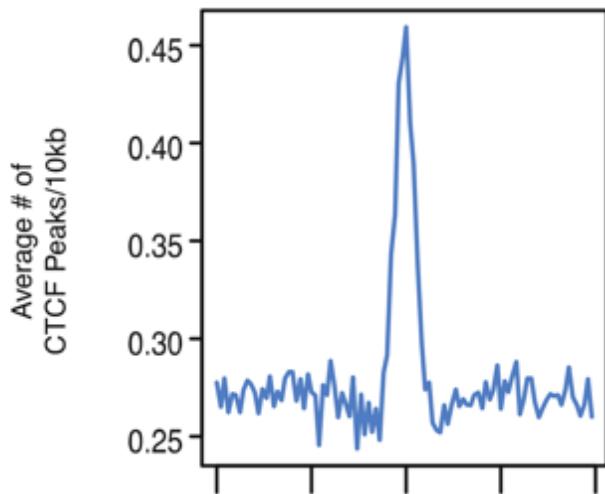
結果はBED形式で出力される（ただしポジションの数字の単位は塩基ではなくBinの数であることに注意）。

```
manager@bl8vbox[chr19] cat N15.txt
chr      start    end
chr19    4        20
chr19    6        20
chr19    9        20
chr19    13       18
chr19    20       22
chr19    22       25
chr19    22       34
chr19    28       33
chr19    34       36
chr19    37       39
chr19    40       43
chr19    40       52
chr19    52       55%
```

TAD検出後の解析

検出されたTAD境界の位置で頻繁に見つかるモチーフや、特定のDNA結合タンパク質の結合パターンを見つける、など。

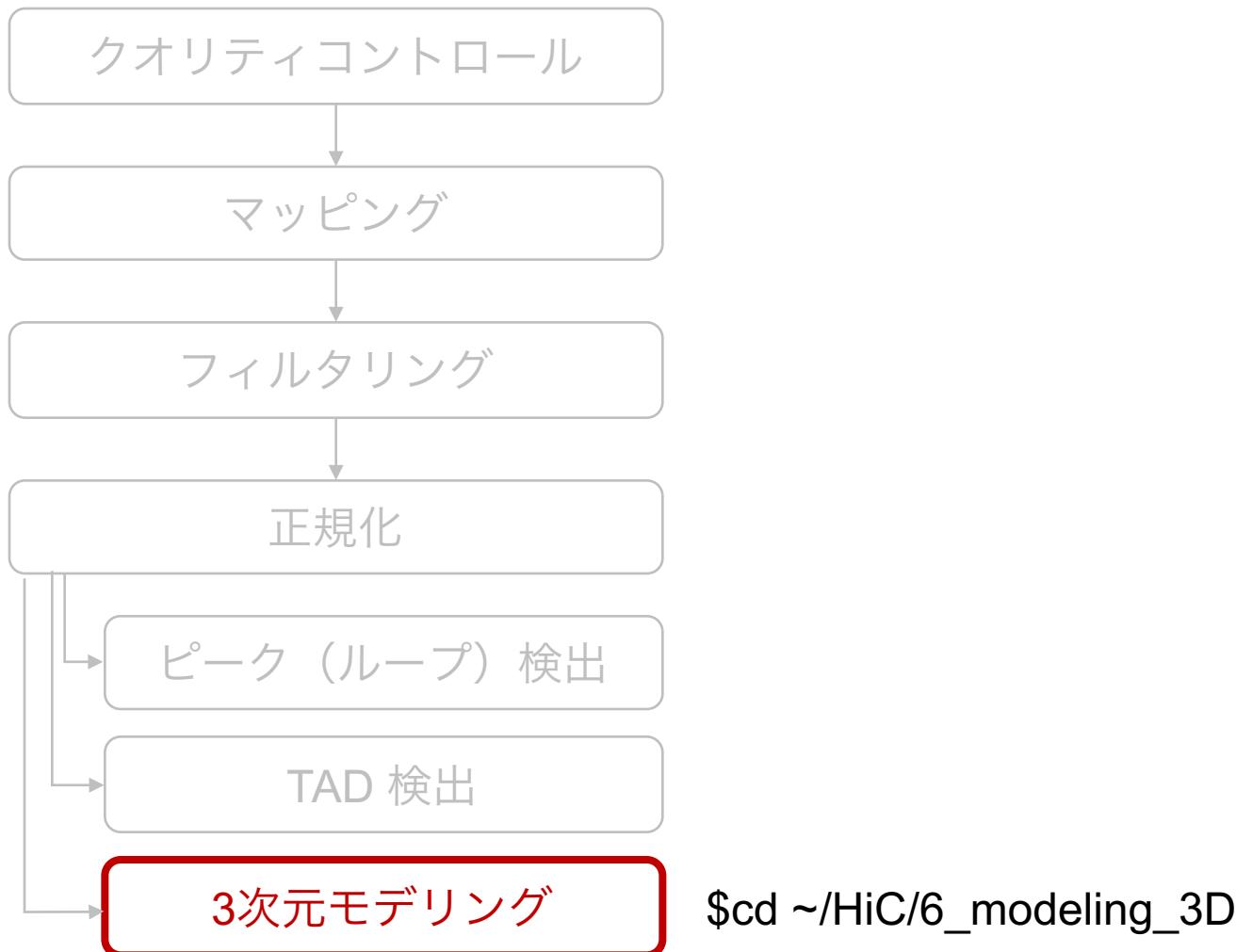
サンプル間でそれぞれTADを検出し、TADの変化と周辺の遺伝子発現の変化を比較する、など。



Forcato, Mattia, et al.
"Comparison of computational methods for Hi-C data analysis."
Nature methods 14.7 (2017): 679.

Ke, Yuwen, et al.
"3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis."
Cell 170.2 (2017): 367-381.

Hi-C解析の流れ



染色体3Dモデルの構築

Hi-C実験で得られたコンタクトマップのデータを、「空間制約」として3次元構造のモデリングに組み込む。

大きく分類して2つのアプローチがある。 (Serra, et al. 2015)

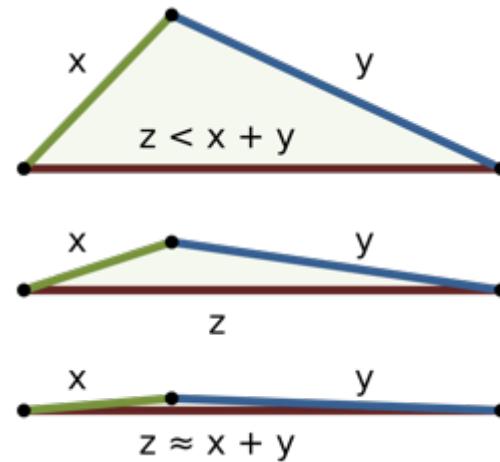
1. コンタクトマップを直接3次元構造に変換する解析的アプローチ。
コンタクトマップが単一のコンセンサス的な構造を表現していると仮定する。したがって、実際はシングルセル実験のデータによりふさわしい。
2. コンタクトマップを制約として、モンテカルロサンプリングやベイズ法でそれを満たす構造を探す最適化アプローチ。これはさらに2つのカテゴリに分類でき、
 - a. シミュレーションがそれぞれ独立に、構造の「集合」を生成する方法。得られた構造の集合が集団内多様性を表現する。
 - b. アンサンブルベースの方法。多くの構造を同時にシミュレートして、コンタクトマップ制約を満たす構造集合を探索する。

染色体3Dモデルの構築

解析的アプローチは次の2つのステップからなる。

1. コンタクトマップをユークリッド距離行列など、「距離の性質」を満たした行列データに変換する。

コンタクトマップで表現されているデータは、多くの異なる構造の「平均」であるため、必ずしも距離の性質のひとつ「三角不等式」が満たされない。

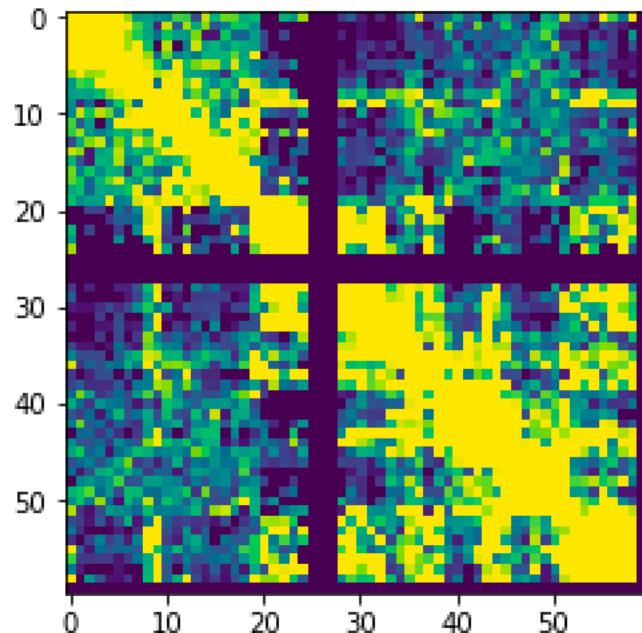


<https://ja.wikipedia.org/wiki/三角不等式>

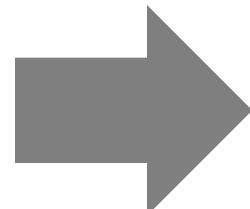
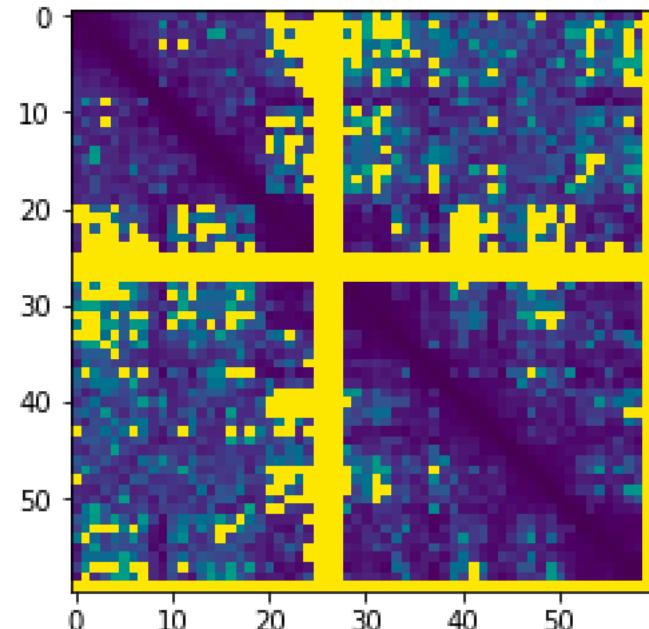
2. 距離行列を満たす直交座標系上の3次元構造をなんらかの最適化手法などで見つける。

コンタクトマップから距離行列への変換

コンタクトマップ
(値が大きいほど接触頻度が高い)



距離行列
(値が大きいほど距離が離れている)



コンタクトマップから距離行列への変換

コンタクトから距離へいかに変換するか？
単純なやり方は、単に逆数をとる。

$$D_{i,j} = \frac{1}{(A_{i,j})^\alpha}$$

$\alpha=1$ とする場合が多い（が、それは自明ではない）
しかし、この変換だと $A_{i,j} = 0$ （コンタクトが観測されなかった）場合、
距離 $D_{i,j}$ が無限大になってしまう。
スペースなコンタクトマップでは致命的。
コンタクトマップのすべての要素に適当に小さな値を足して嵩上げするか？ => ほとんどの領域間で適当に足した値の逆数が支配的になってしまふ

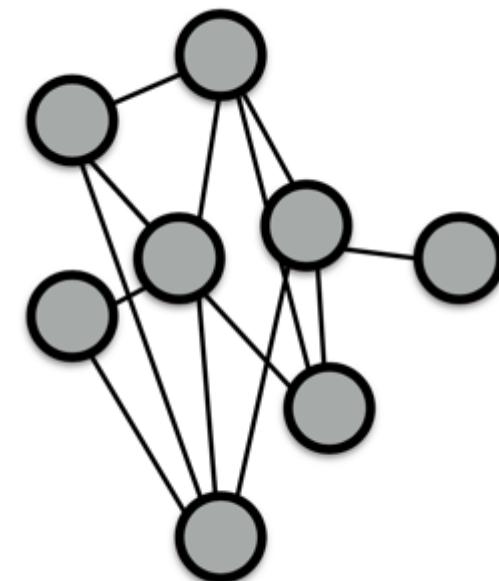
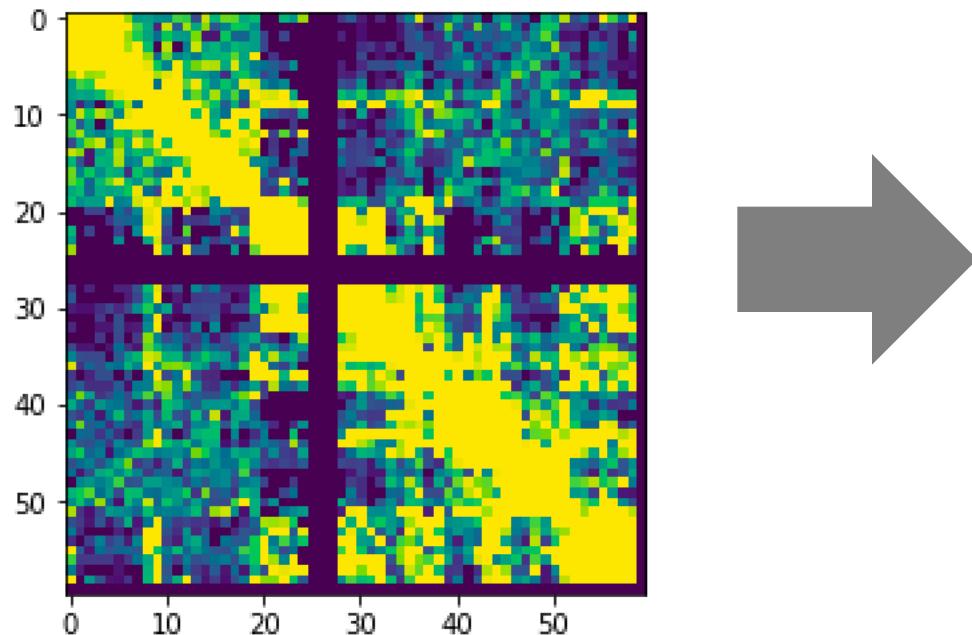
⇒ Shortest-path reconstruction

ShRec3Dの手法（Lesne, et al. 2014）。
本実習で再現（公開ツールはMATLAB）。

Shortest-path reconstruction

グラフ理論を応用した距離行列構成手法。

まず、コンタクトマップを構成するBinそれぞれをノードとし、コンタクトマップの値の逆数を重みとしたエッジを持つグラフを構成する。コンタクトマップの値がゼロの場合は、それらのノード間にエッジはひかれない。



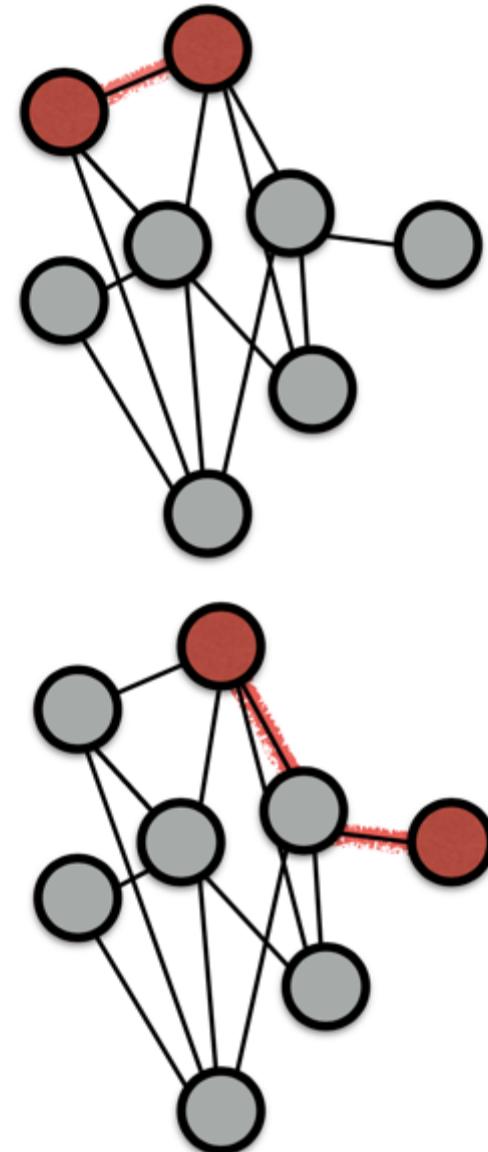
Shortest-path reconstruction

エッジが直接ひかれているノード間の距離は、単純にそのエッジの重みとする。

エッジが直接ひかれていないノード間の距離は、そのノード間の「最短パス」を構成するエッジ群の重みの和として定義する。

ノード間の最短パスはFloyd-Warshallアルゴリズム（動的計画法）で高速に検索できる。

都合のいいことに、こうして計算した「距離」は三角不等式を満たす。



コンタクトマップから距離行列への変換

```
$less ./convert_contact_to_distance.py
```

Python の NetworkXを使用してグラフ構築、最短パス検索をしている。

今回は、正規化のセクションで生成した19番染色体のコンタクトマップを使って、19番染色体の距離行列を構築してみる。

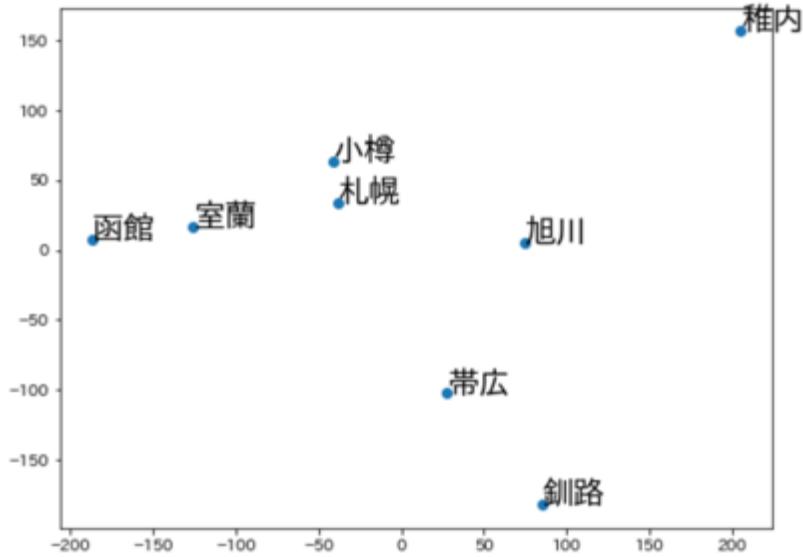
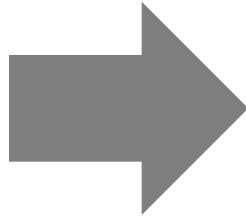
```
$python ./convert_contact_to_distance.py ../3_normalization/norm_mat.txt
```

ディレクトリ内に、dist.npy という距離行列を格納したバイナリファイルができるはず。

距離行列から3次元構造への変換

距離行列から空間構造への変換は、もっとも簡単なやり方としては、多次元尺度構成法（Multi-dimensional scaling; MDS）を使う。

	札幌	旭川	稚内	釧路	帯広	室蘭	函館	小樽
札幌	0.0	115.0	274.0	249.0	152.0	88.0	152.0	30.0
旭川	115.0	0.0	202.0	188.0	118.0	200.0	263.0	130.0
稚内	274.0	202.0	0.0	358.0	313.0	360.0	413.0	266.0
釧路	249.0	188.0	358.0	0.0	98.0	290.0	330.0	277.0
帯広	152.0	118.0	313.0	98.0	0.0	195.0	240.0	180.0
室蘭	88.0	200.0	360.0	290.0	195.0	0.0	64.0	98.0
函館	152.0	263.0	413.0	330.0	240.0	64.0	0.0	159.0
小樽	30.0	130.0	266.0	277.0	180.0	98.0	159.0	0.0



メタ16Sの論文で頻繁に出てくるPCoAもMDSの一種。
上図やメタ16S論文では距離行列から二次元平面上の配置への変換に
MDSを用いているが、本来は変換先の空間は何次元でもOK。
今回の実習では3次元。

距離行列から3次元構造への変換

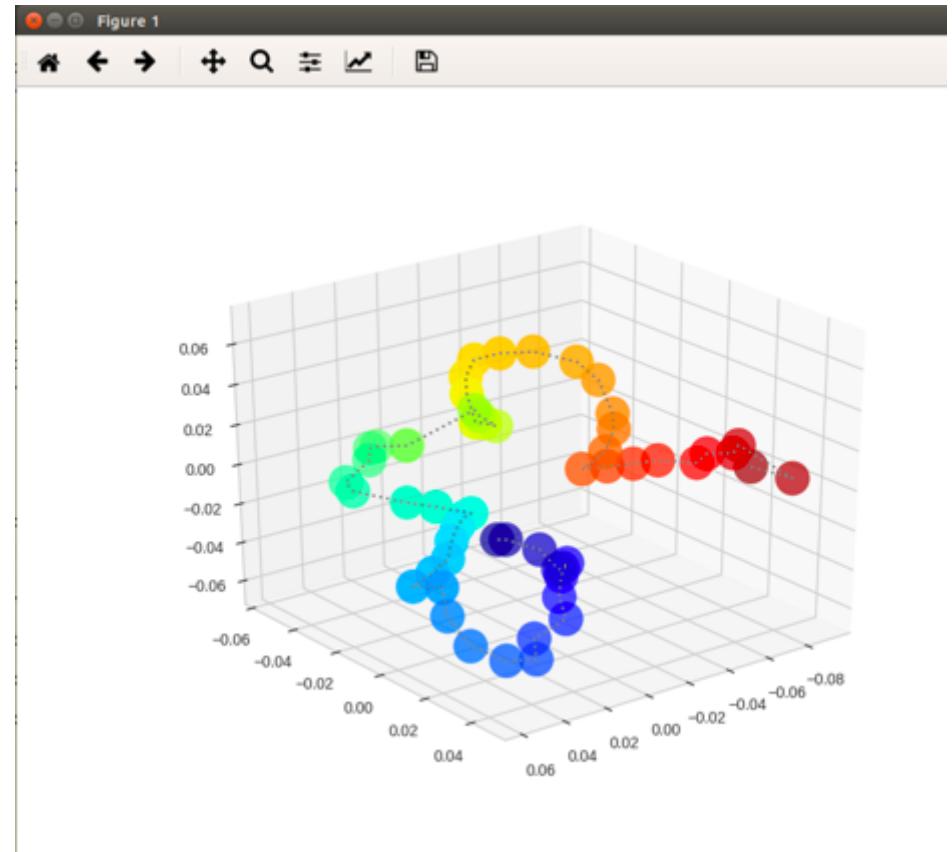
```
$less ./modeling_3d.py
```

さっき作ったdist.npy（距離行列データ）を内部でロードして、MDS計算、得られた3次元座標に基づく染色体の可視化までを行なっている。

実行方法は以下。

```
$python modeling_3d.py
```

マウスドラッグで
ぐりぐり動かすことができる。



参考文献

- Dekker, Job, et al. "Capturing chromosome conformation." *science* 295.5558 (2002): 1306-1311.
- de Wit, Elzo, and Wouter de Laat. "A decade of 3C technologies: insights into nuclear organization." *Genes & development* 26.1 (2012): 11-24.
- Ke, Yuwen, et al. "3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis." *Cell* 170.2 (2017): 367-381.
- Lieberman-Aiden, Erez, et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *science* 326.5950 (2009): 289-293.
- Rao, Suhas SP, et al. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." *Cell* 159.7 (2014): 1665-1680.
- Akdemir, Kadir Caner, and Lynda Chin. "HiCPlotter integrates genomic data with interaction matrices." *Genome biology* 16.1 (2015): 198.
- Fudenberg, Geoffrey, et al. "Formation of chromosomal domains by loop extrusion." *Cell reports* 15.9 (2016): 2038-2049.
- Nagano, Takashi, et al. "Cell-cycle dynamics of chromosomal organization at single-cell resolution." *Nature* 547 (2017): 61–67
- Marbouty, Martial, et al. "Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms." *Elife* 3 (2014): e03318.
- O'sullivan, Justin M., et al. "The statistical-mechanics of chromosome conformation capture." *Nucleus* 4.5 (2013): 390-398.
- Beagrie, Robert A., et al. "Complex multi-enhancer contacts captured by genome architecture mapping." *Nature* 543.7646 (2017): 519-524.

- Imakaev, Maxim, et al. "Iterative correction of Hi-C data reveals hallmarks of chromosome organization." *Nature methods* 9.10 (2012): 999-1003.
- Yaffe, Eitan, and Amos Tanay. "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture." *Nature genetics* 43.11 (2011): 1059-1065.
- Hu, Ming, et al. "HiCNorm: removing biases in Hi-C data via Poisson regression." *Bioinformatics* 28.23 (2012): 3131-3133.
- Knight, Philip A., and Daniel Ruiz. "A fast algorithm for matrix balancing." *IMA Journal of Numerical Analysis* 33.3 (2013): 1029-1047.
- Forcato, Mattia, et al. "Comparison of computational methods for Hi-C data analysis." *Nature methods* 14.7 (2017): 679.
- Ay, Ferhat, Timothy L. Bailey, and William Stafford Noble. "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts." *Genome research* 24.6 (2014): 999-1011.
- Caleb Weinreb, Benjamin J. Raphael; Identification of hierarchical chromatin domains, *Bioinformatics*, Volume 32, Issue 11, 1 June 2016, Pages 1601–1609
- Serra, François, et al. "Restraint-based three-dimensional modeling of genomes and genomic domains." *FEBS letters* 589.20PartA(2015): 2987-2995.
- Lesne, Annick, et al. "3D genome reconstruction from chromosomal contacts." *Nature methods* 11.11 (2014): 1141-1143.