

Hi-C解析 を知って・学んで・使う

後半・Hi-Cデータを使う

2025年1月16日
東 光一

(Koichi Higashi)
国立遺伝学研究所
khigashi@nig.ac.jp

公共Hi-Cデータを利用する（見る、探索する）

The image shows two browser windows side-by-side. The left window is the "4DN Data Portal" at <https://data.4dnucleome.org>. It features a search bar, navigation links for Data, Tools, Resources, and Help, and a user login/register button. Below the header, it displays statistics: 2265 Experiment Sets, 5274 Experiments, and 41951 Files. A prominent feature is a stacked bar chart showing the count of experiments grouped by organism and experiment type. The right window is the "4DN Visualization Workspace" at <https://data.4dnucleome.org/tools/visualization>. It has a similar header and navigation. The main content area is titled "4DN Visualization Workspace" and includes a section to "Select a Genome to start:" with buttons for GRCh38, GRCm38, and dm6. Below this, there's a description of the workspace's purpose, contact information, and a note about it being a BETA release. At the bottom, there are two HiGlass visualizations showing genomic data for GRCh38 H1-HESC and GRCm38 HF1Y cells.

4DN Data Portal

A platform to search, visualize, and download nucleomics data.

4D Nucleome Data Portal

Experiment Sets

* Only up to the top 30 terms are shown.

Group By Organism

Y Axis X Axis

Organism Legend:

- Human
- Mouse
- Fruit fly
- Chicken
- Hamster
- Zebrafish
- Green m

Hi-C

ChIP-seq

Immunofluorescence

Dilution Hi-C

multiplexed Hi-C

SPT

TSA-seq

2-SH

Multi-stage

CUT&Run

pl-Contact

Capture Hi-C

DamID-seq

ATAC-seq

sc-Hi-C

DNase Hi-C

snv-Hi-C

TCC

OphoDroplet

NBD-seq

RE-seq

single cell Hi-C

sc-RNA-seq

in situ ChIP-PET

sc-RNA-seq

AC-seq

4DN Data Collections

Hi-C Data

All Microscopy

Browse All Data

Browse By Publication

<https://data.4dnucleome.org/help>

4DN Visualization Workspace

Home > Visualization & Analysis Tools >

4DN Visualization Workspace

Select a Genome to start:

GRCh38

GRCm38

dm6

The 4DN Data Visualization Workspace allows users to create 1D or 2D genome displays. Registered users can save and share the displays they create. The Visualization Workspace is powered by HiGlass.

Please contact the DCIC at support@4dnucleome.org with any questions, suggestions, or requests. Please note that the 4DN Visualization Workspace is a BETA release with active development. Features may change without notice.

Search track and matrix files on the data portal

+ Add data

Logged in users can create new displays

Save Clear

GRCh38 H1-HESC - 4DN Hi-C 2012 model

GRCh38 HF1Y - 4DN Hi-C 2012 model

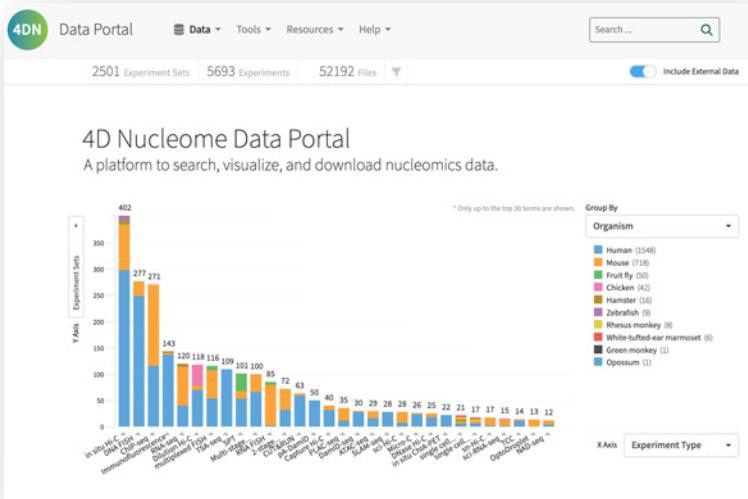
To learn more about how to use the HiGlass Visualization space in the 4DN data portal, click the image below to link out to YouTube for a 5 minute tutorial video.

https://data.4dnucleome.org/browse/?experimentset_type=replicate&type=ExperimentSetReplicate

1

Hi-C解析結果（コンタクトマップ）の公共データの入手先

- ENCODE
- 4DN data portal
- Gene Expression Omnibus



<https://data.4dnucleome.org/>

The figure shows the "Gene Expression Omnibus" homepage. It features a search bar at the top right and a "Sign in to NCBI" button. Below the header, there's a brief introduction to GEO and a "Getting Started" section with links to Overview, FAQ, About GEO DataSets, About GEO Profiles, and About GEO2R Analysis. To the right, there's a "Tools" section with links to Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation, Analyze a Study with GEO2R, and Studies with Genome Data Viewer Tracks. On the far right, there are statistics: Datasets: 4348, Series: 244054, Platforms: 26898, Samples: 7582356.

<https://www.ncbi.nlm.nih.gov/geo/>

<https://www.encodeproject.org/>

The figure shows the "Experiment Matrix" interface from the ENCODE project. At the top, there are navigation links: ENCODE, Data, Encyclopedia, Materials & Methods, Help, and a shopping cart icon. Below the header, there's a search bar with the placeholder "Enter search term(s)" and buttons for "List", "Report", "Download", and "Visualize". The main area is titled "Experiment Matrix" and contains a sidebar with filters for "ASSAY", "BIOSAMPLE", "cell line", "primary cell", "in vitro differentiated cells", and "tissue". The "ASSAY" filter shows results for "in situ Hi-C" (78), "long read scRNA-seq" (64), "Replicon-seq" (59), "PAS-seq" (40), "BruCAGE-seq" (32), "polyA minus RNA-seq" (31), "FAIRE-seq" (30), and "RNA-PET" (30). The "BIOSAMPLE" filter shows results for "mouse embryonic stem cell" (2), "GM12878" (5), "HL-60/S4" (3), "A549" (2), and "GM20431" (2). The "cell line" filter shows results for "mature B cell" (4), "mammary epithelial cell" (2), "astrocyte of the cerebellum" (1), "astrocyte of the spinal cord" (1), and "brain microvascular endothelial cell" (1). The "primary cell" filter shows results for "motor neuron" (9). The "in vitro differentiated cells" filter shows results for "gastrocnemius medialis" (4) and "transverse colon" (4). The "tissue" filter shows results for "gastrocnemius medialis" (4) and "transverse colon" (4).

コンタクトマップのデータ形式について

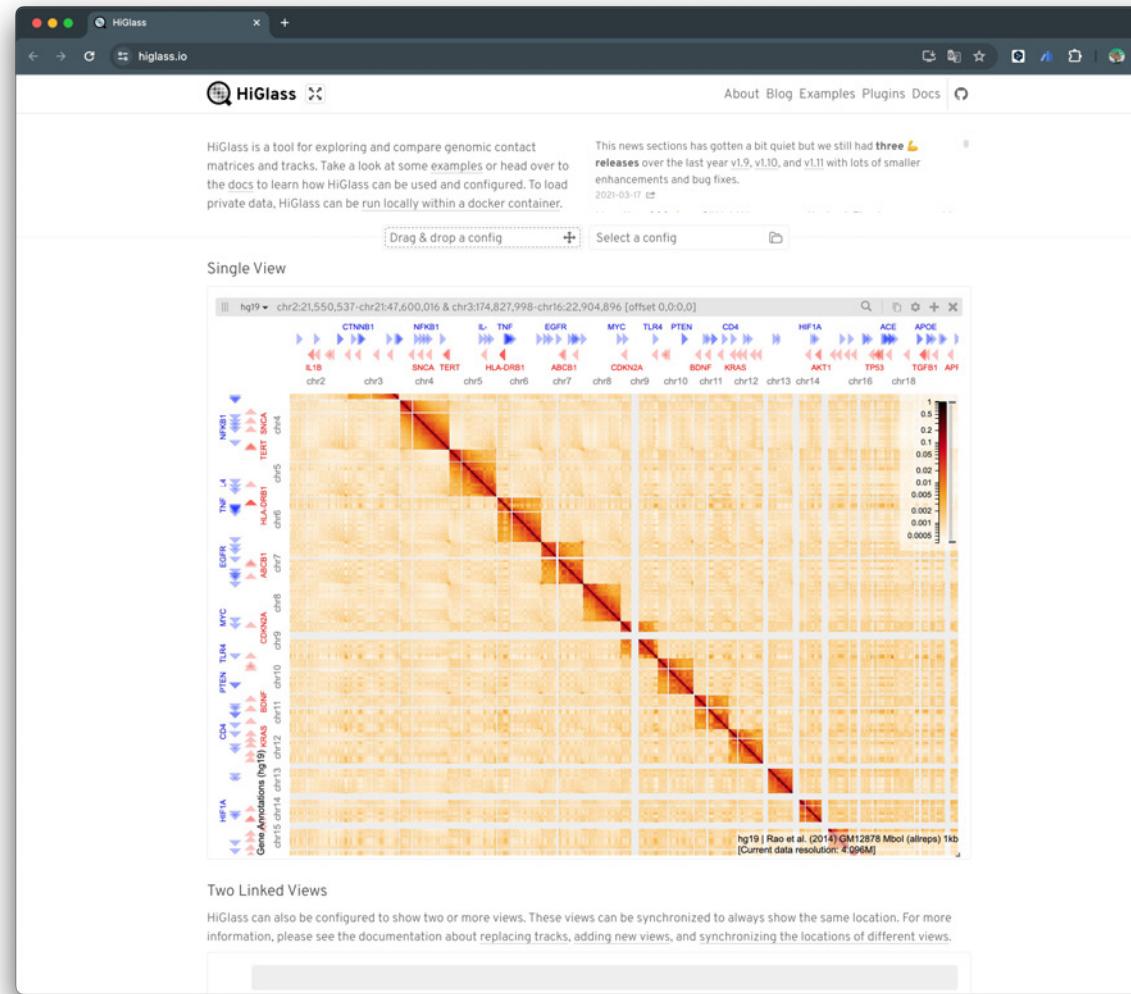
	cool (mcool) 形式	.hic (dothic) 形式	その他
概要	<ul style="list-style-type: none">cooler パッケージによって作成・操作されるコンタクトマップ形式階層的 (マルチスケール) な解像度を扱う場合は mcool (multi-resolution cool) が用いられる	<ul style="list-style-type: none">Juicer/HiC-Pro などのパイプラインで標準的に出力される形式1つのファイルで複数解像度を格納可能 (内部はバイナリ構造)	<ul style="list-style-type: none">テキスト形式 (TSV, bedGraph, matrix など)バイナリ形式 (HiCEexplorer の HDF5)独自形式 (例えばHOMER など一部ツール特有)
利点	<ul style="list-style-type: none">cooler ツールセットによる容易な操作 (クオリティフィルタやバランス等)HDF5 ベースで階層的にデータを格納するため、部分的なアクセスが高速cooler balance による正規化が容易	<ul style="list-style-type: none">Juicer Tools との連携がスムーズ1ファイルでマルチスケールなコンタクトマップを格納できるJuicebox など公式ビジュアライザとの相性が良い	軽量なテキスト形式やソフトウェア特有のフォーマットなど、用途に合わせた柔軟性
主な解析パイプライン	coolerパッケージ	<ul style="list-style-type: none">JuicerHiC-Pro (.hicへの変換が可能)	HiCEexplorerなど。
可視化	HiGlass	Juicebox	HiCEexplorerなど。

Leonid Mirnyが牽引。

Lieberman-Aidenが牽引。

HiGlass

Webアプリケーションだが、WebサーバそのもののDockerイメージを配布してくれているので、ローカルでサーバを立ち上げれば自分のコンタクトマップを描画・探索できる。
(ただ、Dockerでウェブサーバを建てるの慣れてないと少しだけ敷居が高い)



Juicebox

デスクトップ版もあるが、ウェブアプリケーションも使いやすい。
GEOのデータセットについて、巨大なコンタクトマップをダウンロードせずに確認できる。



A screenshot of the 'GEO Accession viewer' at ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1551550. It shows a detailed data entry form for a Hi-C dataset. In the center, there's a text box with Japanese text: 'GEOでFTPのURLをコピーして指定すると、' (Copy the FTP URL from GEO and specify it). Below this, a context menu is open over a table row, with the 'Copy' option highlighted.

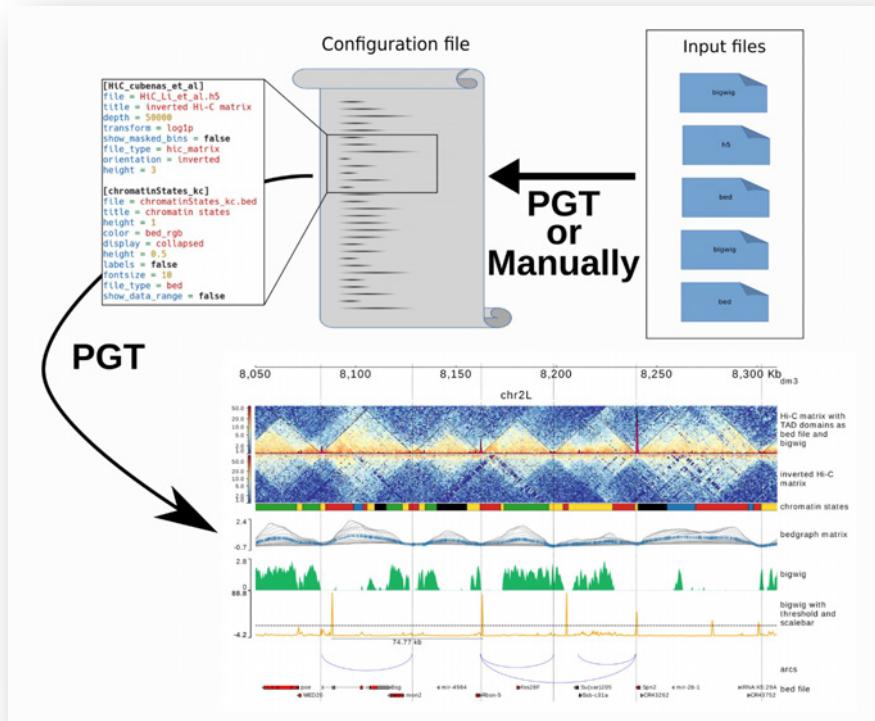
Supplementary file	Size	Download	File type/resource
GSM1551550_HIC001.hic	3.5 Gb	(ftp)(http)	HIC
GSM1551550_HIC001_30.hic	2.8 Gb	(ftp)(http)	
GSM1551550_HIC001_merged_nodups.txt.gz	3.6 Gb	(ftp)(http)	



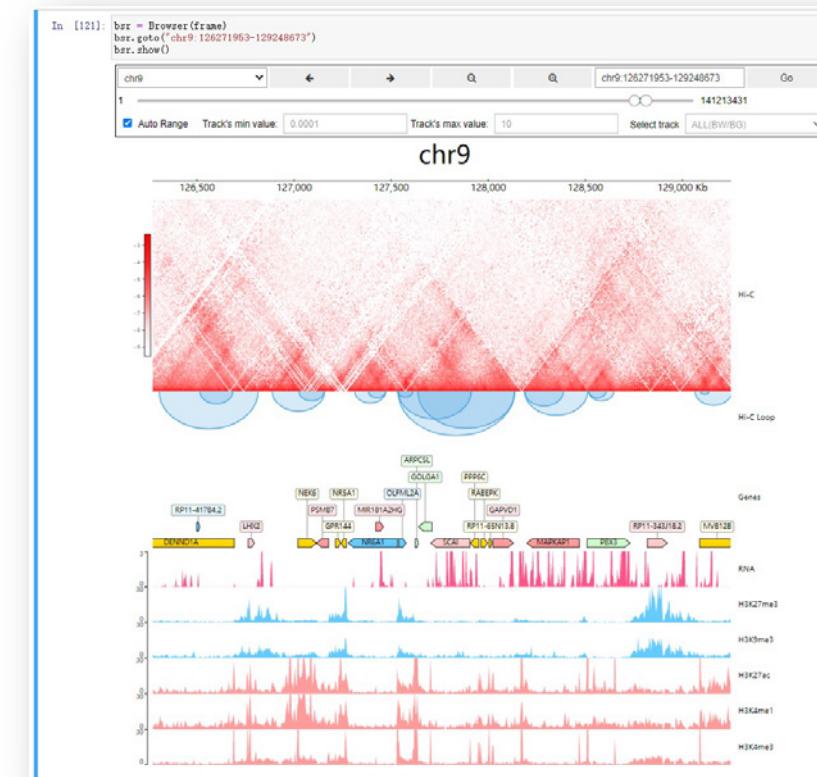
公共データをダウンロードして（あるいは自分で生成したHi-Cデータで）
ChIP-seqデータなどと比較する論文出版レベルのプロットを描く。

pyGenomeTracksなど、いくつかの可視化ツールがある。

pyGenomeTracks



CoolBox



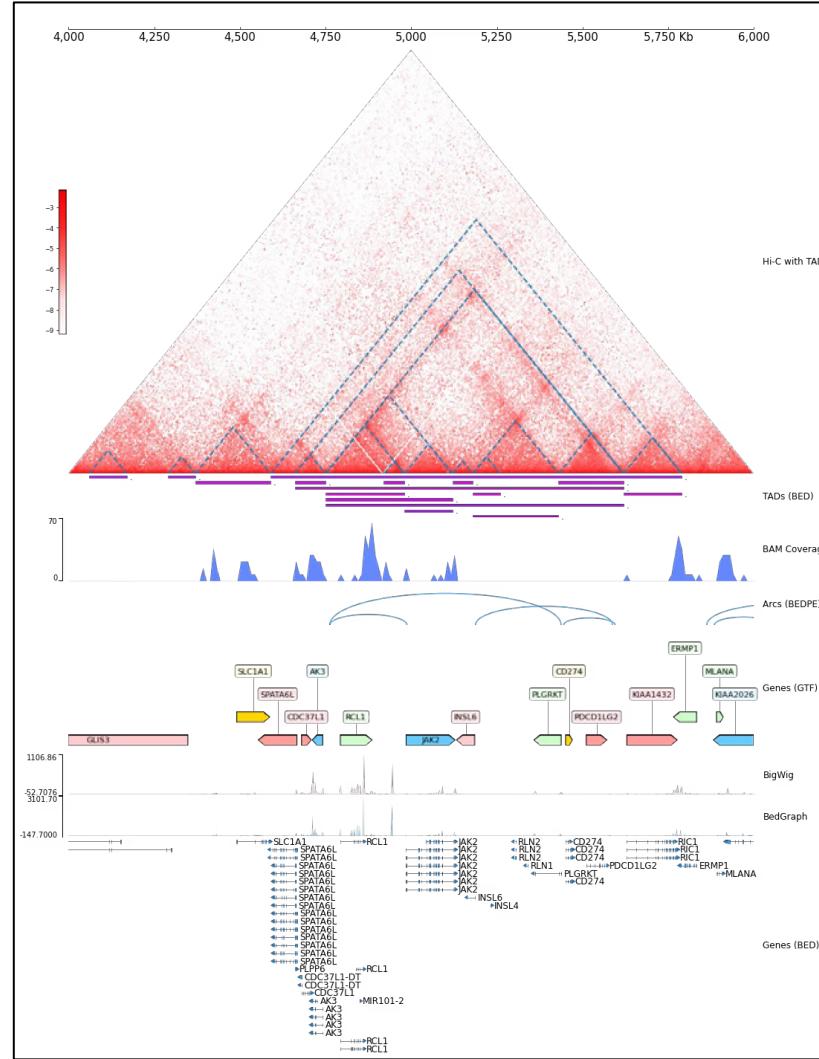
<https://qiita.com/khigashi02/items/8d68ce4fc5355083e86a>

のページで、coolboxによるコンタクトマップ+他のエピジェネ関連ゲノムトラック描画の方法を解説。

```
# ここで使うデータはCoolBoxのレポジトリにあるテストデータ
data_dir = './CoolBox/tests/test_data'
# 描画したいそれぞれのデータのパス
cool_file = f'{data_dir}/cool_chr9_4000000_6000000.mcool'
bed_file = f'{data_dir}/tad_chr9_4000000_6000000.bed'
bam_file = f'{data_dir}/bam_chr9_4000000_6000000.bam'
bedpe_file = f'{data_dir}/bedpe_chr9_4000000_6000000.bedpe'
gtf_file = f'{data_dir}/gtf_chr9_4000000_6000000.gtf'
bigwig_file = f'{data_dir}/bigwig_chr9_4000000_6000000.bw'
bedgraph_file = f'{data_dir}/bedgraph_chr9_4000000_6000000.bg'
bedannotation_file = f'{data_dir}/bed_chr9_4000000_6000000.bed'

frame = XAxis() + \
    Cool(cool_file, cmap="JuiceBoxLike", style='triangular', color_bar='vertical' \
        TADCoverage(bed_file, border_only=True, alpha=1) + \
        Title("Hi-C with TADs") + \
        Spacer(0.1) + \
        BED(bed_file, border_only=True, alpha=1) + \
        Color("#ce00ce") + \
        Title("TADs (BED)") + \
        BAMCov(bam_file) + \
        MinValue(0.0) + MaxValue(70.0) + \
        Title("BAM Coverage") + \
        Spacer(0.1) + \
        Arcs(bedpe_file, line_width=1.5) + \
        Title("Arcs (BEDPE)") + \
        GTF(gtf_file, length_ratio_thresh=0.005) + \
        TrackHeight(6) + Title("Genes (GTF)") + \
        Spacer(0.1) + \
        BigWig(bigwig_file) + \
        Title("BigWig") + \
        BedGraph(bedgraph_file) + \
        Title("BedGraph") + \
        Spacer(0.1) + \
        BED(bedannotation_file) + \
        Feature(height=10, title="Genes (BED)")

frame.plot("chr9:4000000-6000000")
```



実験医学online

スマホで読める実験医学

トップ / 実験医学 2023年6月号 / 第5回 ゲノムトラック

マネして学ぶPython論文グラフ講座

第5回 ゲノムトラック

東光一

本記事のDOI: 10.18958/7281-00024-0000502-00

今回の目的

Pythonで医学・生命科学系論文の図を描く。実際に出版されている論文のデータを使って、論文に掲載されている図を再現してみよう。第5回のテーマはゲノムトラック。ChIP-seq解析やATAC-seq解析の結果として得られるゲノム上のカバレッジプロファイル、さらにHi-C解析結果のコンタクトマップを可視化する。

シェアする

また、（有料ページになってしまふが）
実験医学でも過去にcoolboxを使った描画コードを紹介しているので、興味ある方はそちらもぜひ。

基本的には、商用のキットを使用してシークエンシングライブラリを作成した場合、そのキットの販売会社が提供するプロトコルや解析パイプラインを使用するのが安心で効率的。Dovetail Genomics社のMicro-Cの場合、以下のページにコンタクトマップ生成までの手順が紹介されている。

Welcome to Micro-C documentation

Overview

- Dovetail™ Micro-C Kit uses the Micrococcal nuclease (MNase) enzyme instead of restriction enzymes for chromatin digestion, yielding 146 bp fragments distributed frequently across the genome.

Key benefits of Micro-C:

- Sequence-independent chromatin fragmentation enables even genome-wide detection of chromatin contacts (up to 20% of the genome lacks coverage using restriction enzyme based Hi-C approaches)
- Ultra-high nucleosome-level resolution of chromatin contacts
- Highest signal-to-noise data with both enrichment of long-range informative reads and nucleosome protected fragments
- The ability to detect higher-order features, such as chromatin loops, in proximity ligation data is dependent on enriching long-range informative reads to capture chromatin interaction frequency. The increased chromosome conformation informative

Generating Contact Matrix

There are two common formats for contact maps, the [Cooler format](#) and [Hic format](#). Both are compressed and sparsed formats to avoid large storage volumes; For a given n number of bins in the genome, the size of the matrix would be n^2 , in addition, typically more than one resolution (bin size) is being used.

In this section we will guide you on how to generate both matrices types, [HiC](#) and [cool](#) based on the [.pairs file](#) that you generated in the [previous section](#) and how to visualize them.

Generating [HiC](#) contact maps using Juicer tools

Additional Dependencies

- Juicer Tools** - Download the JAR file for juicertools and place it in the same directory as this repository and name it as [juicertools.jar](#). You can find the link to the most recent version of Juicer tools [here](#) e.g.:

```
wget https://s3.amazonaws.com/hicfiles.tc4ga.com/public/juicer/juicer_tools_1.22.01.jar
mv juicer_tools_1.22.01.jar ./Micro-C/juicertools.jar
```

- Java** - If not already installed, you can install Java as follows:

```
sudo apt install default-jre
```

From [.pairs](#) to [.hic](#) contact matrix

- Juicer Tools** is used to convert [.pairs](#) file into a [HiC](#) contact matrix.
- [HiC](#) is highly compressed binary representation of the contact matrix
- Provides rapid random access to any genomic region matrix
- Stores contact matrix at 9 different resolutions (2.5M, 1M, 500K, 250K, 100K, 50K, 10K, and 5K)
- Can be programmatically manipulated using straw python API

<https://micro-c.readthedocs.io/en/latest/>

(もちろん実験やサンプルの特殊性によってケースバイケースだが)
一般的な情報解析においては、個別のツールを自前で組み合わせていくよりも、
コミュニティによって維持・管理されている「パイプライン」を利用することを推奨。

Hi-C特化パイプライン（ツールセット）：

HiC-Pro
HiCExplorer
FAN-C
など。

さらに、前処理も含めたすべての処理を、なんらかの“ワークフロー言語”で
書かれた既存のパイプラインで実行することも検討する。



1. Nextflow

概要: Nextflowは、データ駆動型のワークフローの自動化とポータビリティに特化したツールです。Groovyベースのスクリプト言語で記述されます。

特徴:

DockerやSingularityとの互換性による高い移植性。

クラウドとの統合が容易 (AWS, Google Cloud, Azure)。

複数のプラットフォーム間でスケーラブル。

2. Snakemake

概要: Pythonベースの言語で、Makefileに似た構文を使用してワークフローを定義します。

生物情報学のコミュニティに広く受け入れられています。

特徴:

明確な構文と強力なエラー処理。

多様な計算環境に対応し、ローカルマシンからクラウド、HPCまで対応。

柔軟なルール定義による複雑なデータフローの管理。

3. Common Workflow Language (CWL)

概要: 計算ワークフローを記述するための仕様。再利用可能なワークフローの記述を目指しています。

特徴:

プラットフォーム非依存、ツールの再利用が可能。

オープンスタンダードであり、広い範囲のソフトウェアとハードウェアで利用可能。

複数のバイオインフォマティクスツールと容易に統合。

Empowering bioinformatics communities with Nextflow and nf-core

Posted May 14, 2024.

Björn E. Langer, Andreia Amaral, Marie-Odile Baudement, Franziska Bonath, Mathieu Charles, Praveen Krishna Chitneedi, Emily L. Clark, Paolo Di Tommaso, Sarah Djebali, Philip A. Ewels, Sonia Eynard, James A. Fellows Yates, Daniel Fischer, Evan W. Floden, Sylvain Foissac, Gisèle Gabernet, Maxime U. Garcia, Gareth Gillard, Manu Kumar Gundappa, Cervin Guyomar, Christopher Hakkaart, Friederike Hanssen, Peter W. Harrison, Matthias Hörtelhuber, Cyril Kurylo, Christa Kühn, Sandrine Lagarrigue, Delphine Lallias, Daniel J. Macqueen, Edmund Miller, Júlia Mir-Pedrol, Gabriel Costa Monteiro Moreira, Sven Nahnsen, Harshil Patel, Alexander Peltzer, Frederique Pitel, Yulixais Ramayo-Caldas, Marcel da Câmara Ribeiro-Dantas, Dominique Rocha, Mazdak Salavati, Alexey Sokolov, Jose Espinosa-Carrasco, Cedric Notredame, the nf-core community.

doi: <https://doi.org/10.1101/2024.05.10.592912>

This article is a preprint and has not been certified by peer review [what does this mean?].

Langer, Bjorn E., et al.

"Empowering bioinformatics communities with Nextflow and nf-core."

bioRxiv (2024): 2024-05.

Download
 Print/Save
 Supplement
X Post
COVID-19
medRxiv
Subject Area

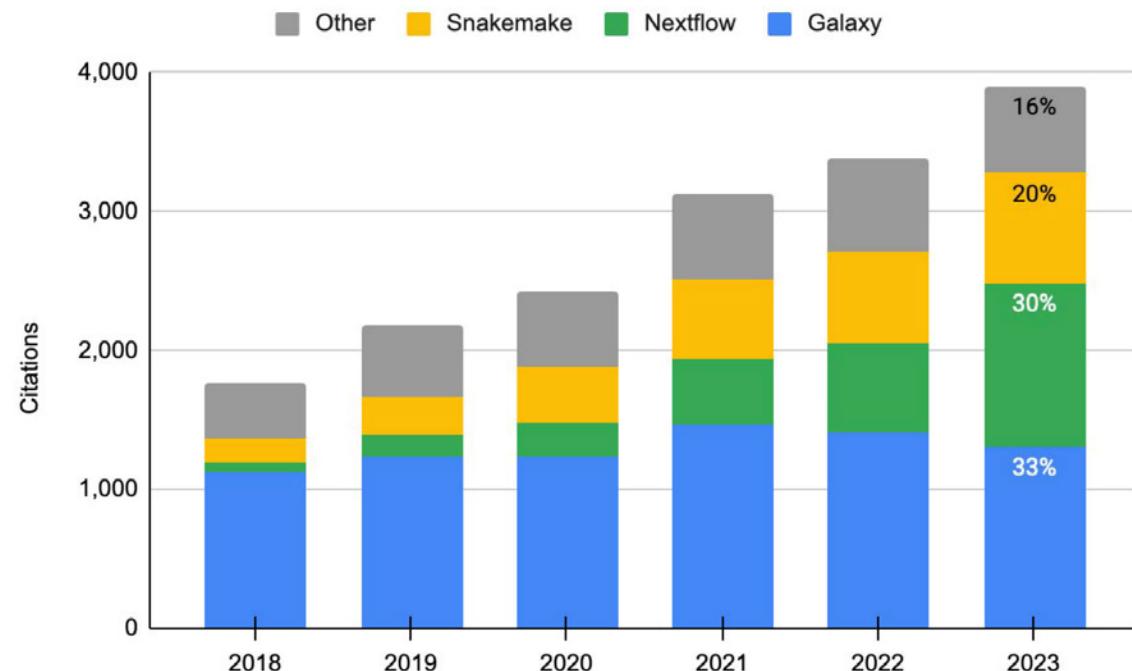


FIGURE 1: Google Scholar citation counts for bioinformatics workflow management systems. Sum of citations of the major publications of Galaxy, Nextflow, and Snakemake between 2018 and 2023 (Data in Supplementary Table 1).

ワークフロー言語はパイプラインを記述する手段を提供するが、記述方法の「標準」は提供しない。
そこで、標準の策定によってパイプライン間の相互運用性を確保した「パイプラインレジストリ」が確立されている。
Nextflowの場合、バイオインフォで著名なパイプラインレジストリが **nf-core**

nf-core

nf-co.re

Home Pipelines Resources Docs Community About Search Join nf-core

Upcoming event

Bytesize: Get your containers with "Nextflow inspect" [bytesize](#)

Phil Ewels, Seqera

May 28, 2024, 20:00-20:30

Event starts in
about 21 hours

[Event Details](#) [Export event](#)

nf-core

A community effort to collect a curated set of analysis pipelines built using Nextflow.

[VIEW PIPELINES](#)

For facilities

 Highly optimised pipelines with excellent reporting. Validated releases ensure reproducibility.

For users

 Portable, documented and easy to use workflows. Pipelines that you can trust.

For developers

 Companion templates and tools help to validate your code and simplify common tasks.

Pipelines

Browse the 107 pipelines that are currently available as part of nf-core.

Search:

Released: 61 Under development: 34 Archived: 12 Stars: 95

rnaseq ★ 795
RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.
rna rna-seq
3.14.0 released 5 months ago

sarek ★ 343
Analysis pipeline to detect germline or somatic variants (pre-processing, variant calling and annotation) from WGS / targeted sequencing
annotation cancer gatk4 genomics germline pre-processing somatic target-panels variant-calling whole-exome-sequencing whole-genome-sequencing
3.4.2 released 20 days ago

mag ★ 182
Assembly and binning of metagenomes
annotation assembly binning long-read-sequencing metagenomes metagenomics nanopore nanopore-sequencing
3.0.0 released 14 days ago

scrnaseq ★ 172
A single-cell RNAseq pipeline for 10X genomics data
10x-genomics 10xgenomics alevin bustools celranger kallisto rna-seq single-cell star-solo
2.6.0 released 20 days ago

chipseq ★ 172
ChIP-seq peak-calling, QC and differential analysis pipeline.
chip chip-seq chromatin-immunoprecipitation macs2 peak-calling
2.0.0 released over 1 year ago

atacseq ★ 163
ATAC-seq peak-calling and QC analysis pipeline
atac-seq chromatin-accessibility
2.1.2 released 10 months ago

ampliseq ★ 157
Amplicon sequencing analysis workflow using DADA2 and QIIME2
16s 18s amplicon-sequencing edna illumina

nanoseq ★ 147
Nanopore demultiplexing, QC and alignment pipeline

methylseq ★ 132
Methylation (Bisulfite-Sequencing) analysis pipeline using Bismark or bwa-meth + MethylDackel

nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

rna rna-seq

Launch version 3.14.0

<https://github.com/nf-core/rnaseq>

→ Introduction Usage Parameters Output AWS Results ↵ Releases 3.14.0

Introduction

nf-core/rnaseq is a bioinformatics pipeline that can be used to analyse RNA sequencing data obtained from organisms with a reference genome and annotation. It takes a samplesheet and FASTQ files as input, performs quality control (QC), trimming and (pseudo-)alignment, and produces a gene expression matrix and extensive QC report.

nf-core/rnaseq

STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. QC
5. Final QC

METHOD

- Aligner: STAR, Quantification: Salmon (default)
- Aligner: STAR, Quantification: RSEM
- Aligner: HISAT2, Quantification: None
- Aligner: RSEM, Quantification: Salmon
- Pseudo-aligner: Kallisto, Quantification: Kallisto

run with

nf-core launch nf-core/rnaseq -

nf-core Nextflow Seqera Platform

subscribers stars
151 795

open issues open PRs
77 11

last release last update
5 months ago 5 months ago

get help

Ask a question on Slack

Open an issue on GitHub

contributors

nf-core/hic

Analysis of Chromosome Conformation Capture data (Hi-C)

chromosome-conformation-capture

hi-c

Pipeline summary

1. Read QC ([FastQC](#))
2. Hi-C data processing
 1. [HiC-Pro](#)
 1. Mapping using a two steps strategy to rescue reads spanning the ligation sites ([bowtie2](#))
 2. Detection of valid interaction products
 3. Duplicates removal
 4. Generate raw and normalized contact maps ([iced](#))
 3. Create genome-wide contact maps at various resolutions ([cooler](#))
 4. Contact maps normalization using balancing algorithm ([cooler](#))
 5. Export to various contact maps formats ([HiC-Pro](#) , [cooler](#))
 6. Quality controls ([HiC-Pro](#) , [HiCExplorer](#))
 7. Compartments calling ([cooltools](#))
 8. TADs calling ([HiCExplorer](#) , [cooltools](#))
 9. Quality control report ([MultiQC](#))

nextflow, dockerがインストールされた解析サーバで、以下のようなコマンドをひとつ実行するだけで、fastqから最終結果まで計算してくれる。

```
nextflow run nf-core/hic  
--input ./samplesheet.csv  
--outdir /path/to/nextflow_out  
--fasta /path/to/WholeGenomeFasta/genome.fa  
--bwt2_index /path/to/Bowtie2Index  
--chromosome_size /path/to/chrom_size.txt  
--digestion 'dpnii'  
--bin_size '5000,10000,50000,500000'  
--res_compartments '250000,500000'  
--max_cpus 16  
--max_memory '512.GB'  
-work-dir /path/to/nextflow_work  
-profile docker
```

制限酵素の設定に注意！

コンタクトマップ解像度についても、最大解像度について慎重に判断する。粗視化はあとでaggregateできるので適當でも大丈夫。

Digestion Hi-C

Parameters for protocols based on restriction enzyme

--digestion

Name of restriction enzyme to automatically set the restriction_site and ligation_site options
(hindiii, mboI, dpnII, arima)

type: string

--restriction_site

Restriction motif provided.

--ligation_site

Expected motif

--chromosome_size

Full path to file size`.

Help text

--restriction_fragments

Full path to restr

Help text

--save_reference

If generated by !

Help text

Valid Pairs Detection

Options to call significant interactions

--keep_dups

Keep duplicated reads

type: boolean

--keep_multi

Keep multi-aligned reads

--max_insert_size

Maximum fragment size to consider. Only value

--min_insert_size

Minimum fragment size to consider. Only value

--max_restriction_fragment_size

Maximum restriction fragment size to consider.

--min_restriction_fragment_size

Minimum restriction fragment size to consider.

--save_interaction_bam

Save a BAM file where all reads are flagged by

--save_pairs_intermediates

Save all types of non valid read pairs in distinct

Contact maps

Options to build Hi-C contact maps

--bin_size

Resolution to build the maps (comma separated)

type: string

default: 100000,50000

pattern: ^(\d+)(,\d+)*\$

--hicpro_maps

Generate raw and normalized contact maps with HiC-Pro

type: boolean

--ice_filter_low_count_perc

Filter low counts rows before HiC-Pro normalization

type: number

default: 0.02

--ice_filter_high_count_perc

Filter high counts rows before HiC-Pro normalization

type: integer

--ice_eps

Threshold for HiC-Pro ICE convergence

type: number

default: 0.1

--ice_max_iter

Maximum number of iteration for HiC-Pro ICE normalization

type: integer

default: 100

--res_zoomify

Maximum resolution to build mcool file

type: string

default: 5000

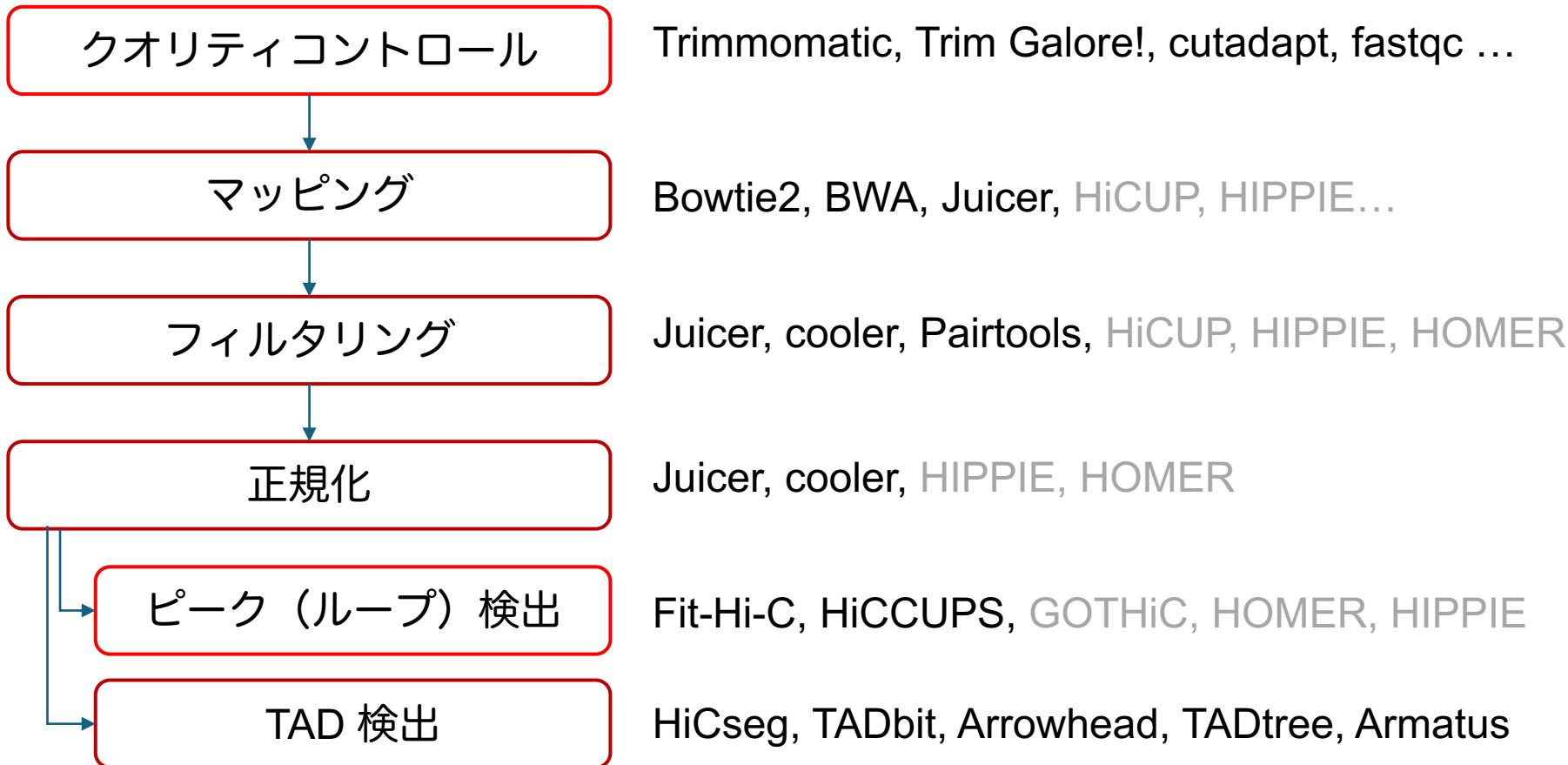
--save_raw_maps

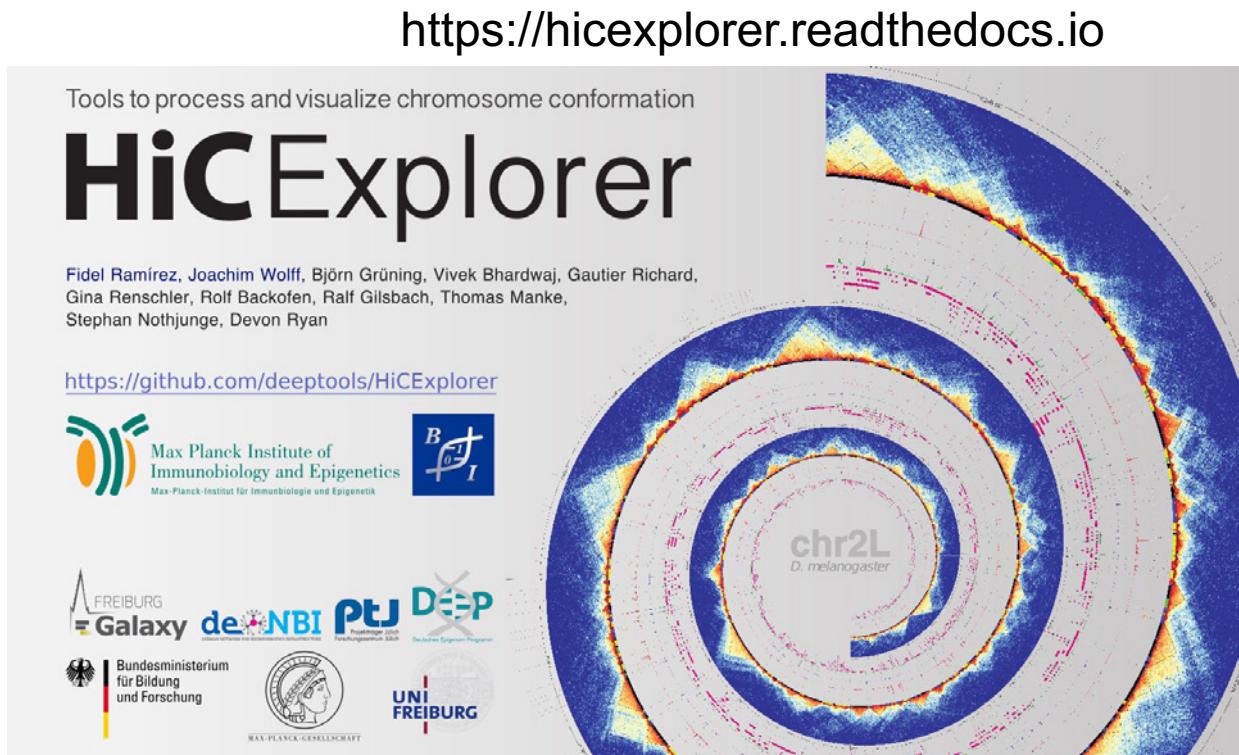
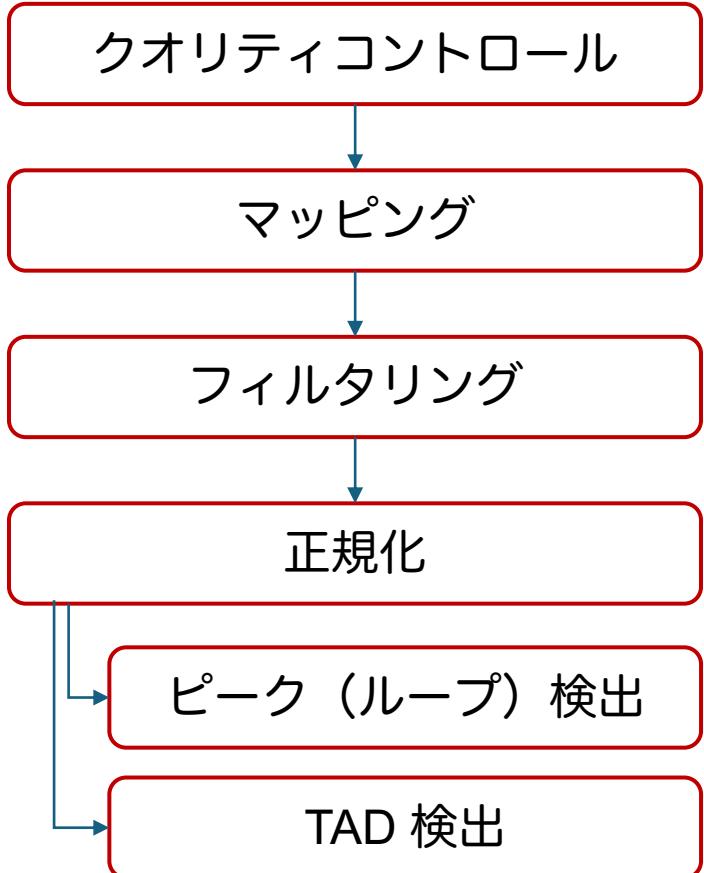
Save raw contact maps

type: boolean

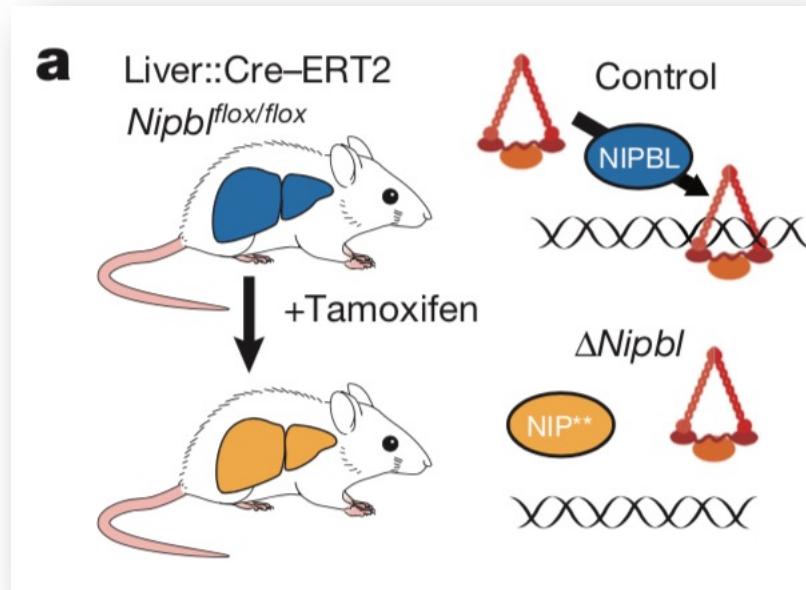
とはいえ、自分のサンプルでうまくいかなかったとき
どのパラメータをどのように調整すればいいのか検討するために
結局、標準的なパイプラインの計算の中身を理解しておくことが必要。

Hi-C解析の流れ、利用可能なツール（一部）





HiCExplorerを用いた解析例



Wibke Schwarzer, Nezar Abdennur, Anton Goloborodko, et al.
Nature (2017)

この論文では、
コヒーリングの染色体高次構造形成への影響を調べるために、
コヒーリングをクロマチンにロードする役割を持つNIPBLを
欠損させたマウスの肝細胞でHi-Cを行なっている。

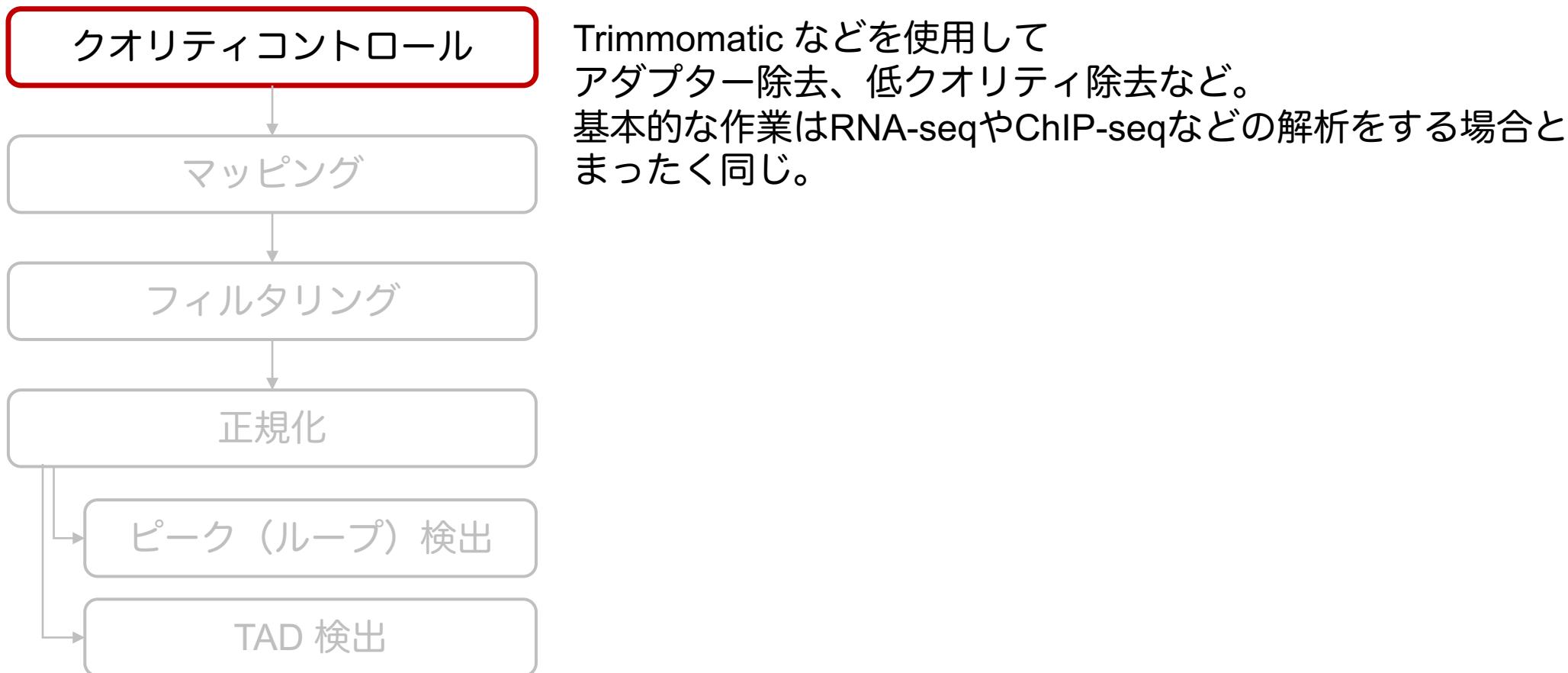
コヒーリングが染色体から外れてしまっているので、構造はだい
ぶ異なるはず。

解析例では、（ほんとはfloxedと比較すべきだけど）野生型と
 $\Delta Nipbl$ で染色体構造がどのように異なるか、それと、論文の図
を再現できるかを、公共データベースに公開されているシーク
ンスデータから自分で解析して確かめている。

以下のリンク先に
FASTQファイルからFigure作成まで全コマンドを掲載・解説。

https://github.com/khigashi1987/Hi-C_handson

Hi-C解析の流れ



Hi-C解析の流れ



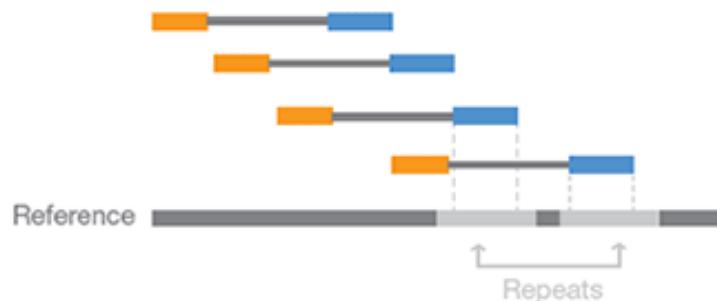
Illuminaのペアエンドシークエンス

Figure 4. Paired-End Sequencing and Alignment

Paired-End Reads



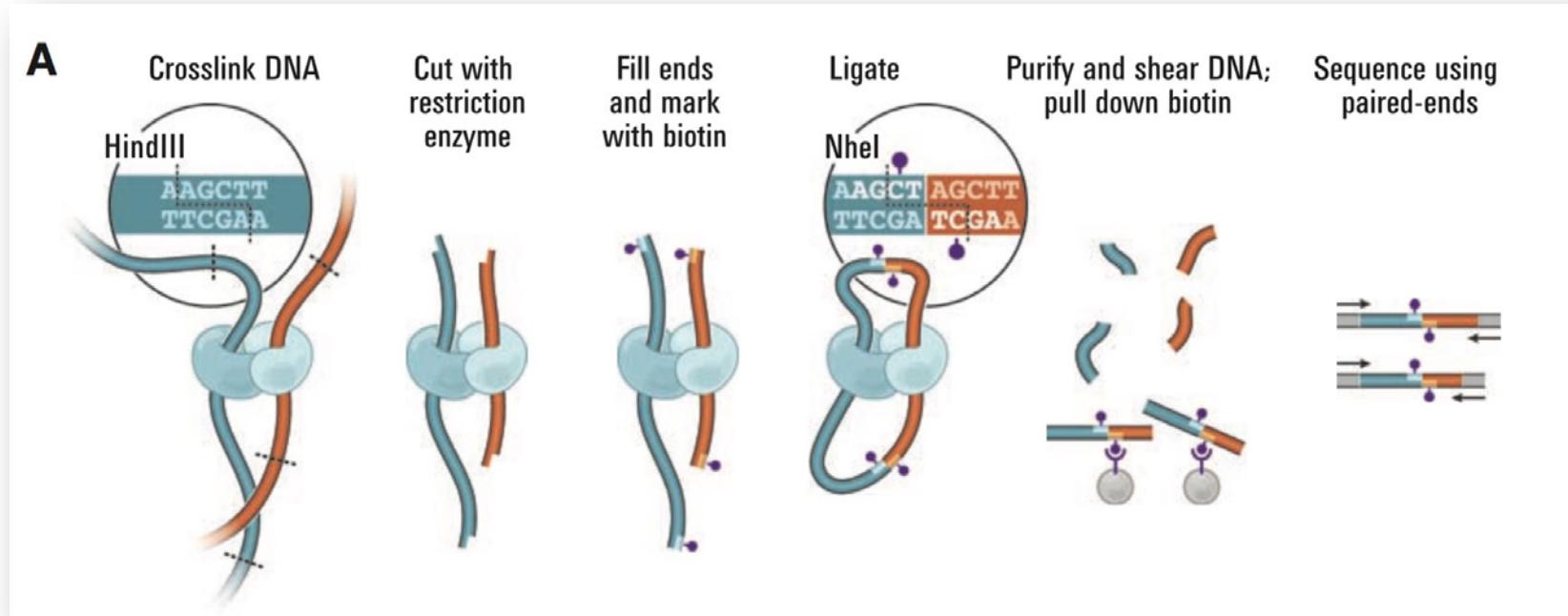
Alignment to the Reference Sequence



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>

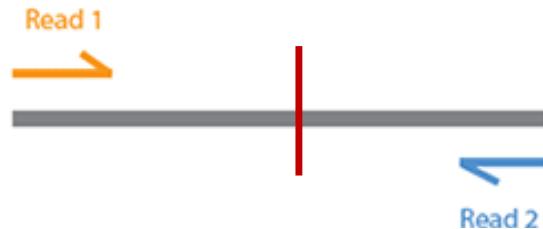
Hi-Cリードの特徴



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

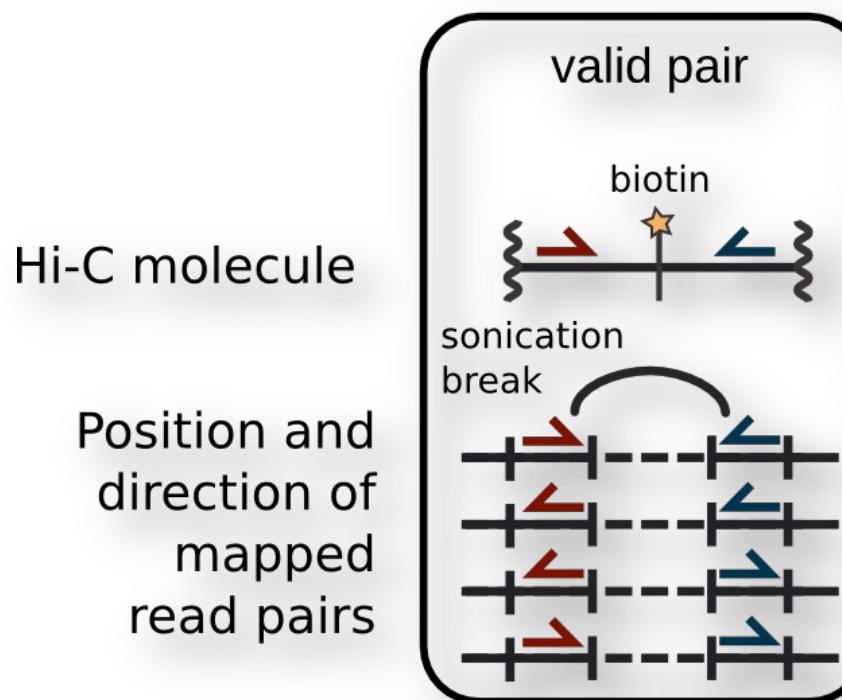


...R2がキメラリードとなっている

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

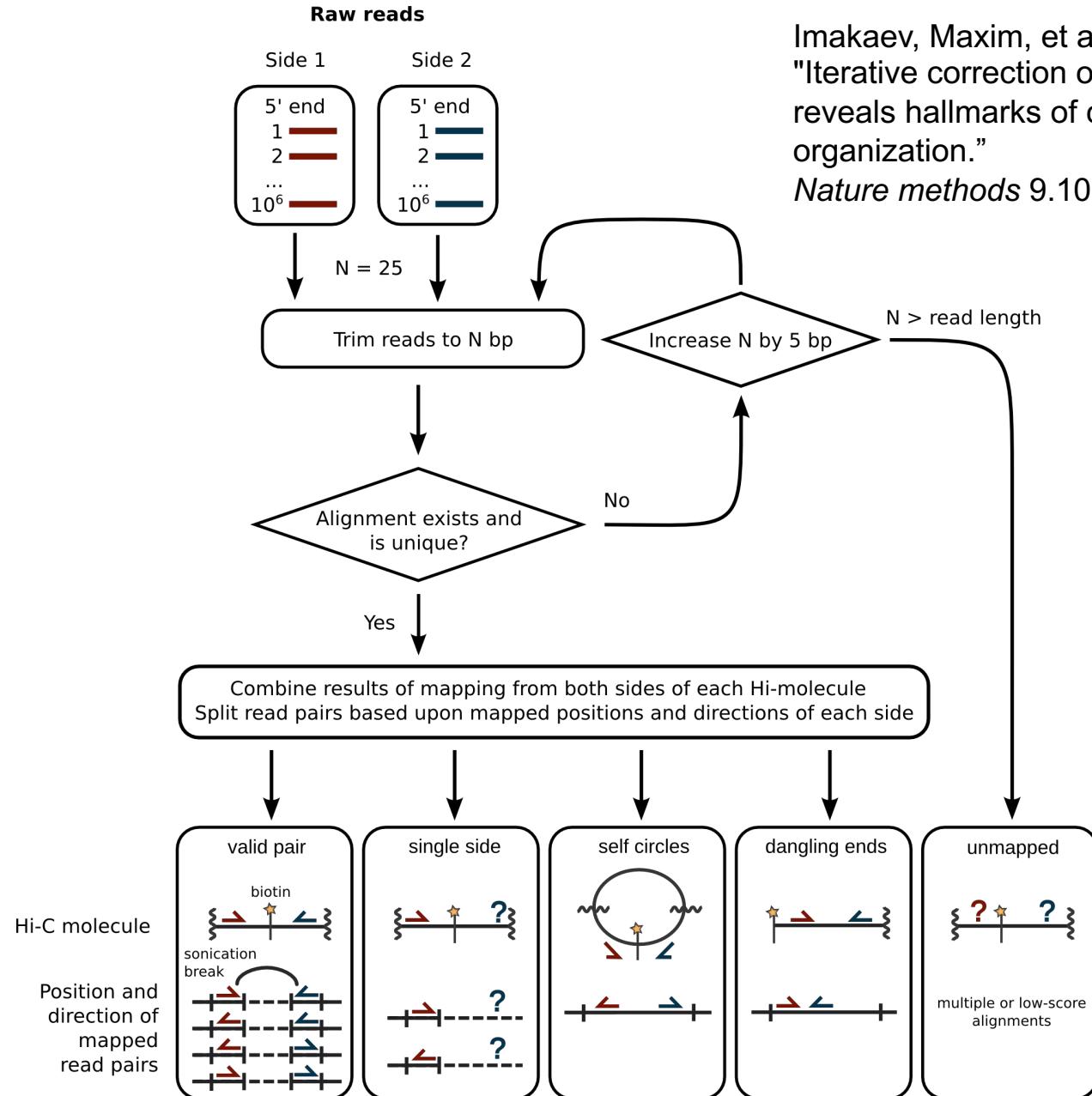


Imakaev, Maxim, et al.
"Iterative correction of Hi-C data
reveals hallmarks of chromosome
organization."
Nature methods 9.10 (2012): 999-1003.

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にまとめる。
 - マッパーによっては、ペアエンドファイルを入力するとインサートサイズを適当に仮定してすばやく検索するヒューリスティックが起動してしまう。Hi-Cデータの場合ペアはゲノム上ですごく遠い場所（あるいは別の染色体）に由来するかもしれない。なので、R1とR2は個別にマッピングする。
 - ローカルアラインメントをする。グローバルアラインメントはダメ。Hi-Cリードは本質的に ligation product で、chimeric sequence なため。
 - アラインメントのパラメータを適切に調整する。
2. Iterative alignment method

参考 : 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.

Hi-C解析の流れ

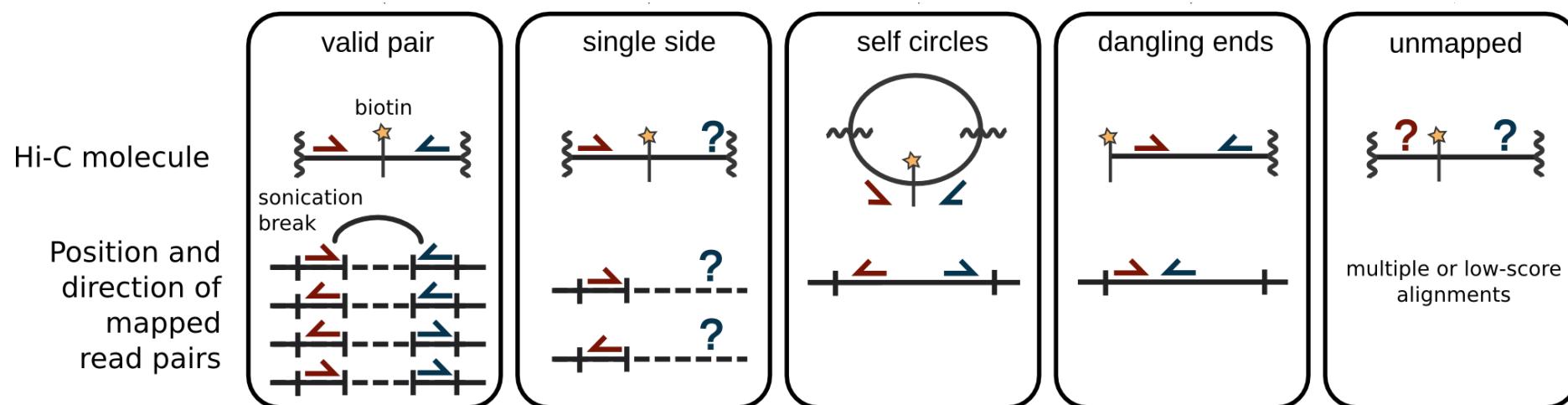


マッピングされたペア情報のフィルタリング

マッピングされたペアのすべてがHi-Cのコンタクト情報として使えるわけではない。

以下のように様々な理由で「コンタクト」を反映していないペアが存在する。

不適切なペアは、マッピングのパターン、すなわちマッピングされたStrand (= orientation) や、実験に使った制限酵素の認識配列の位置との関係などから判断することができる。



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）にはばらつきがある。
 - 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: RPKM, TPMなど遺伝子の長さによる補正

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

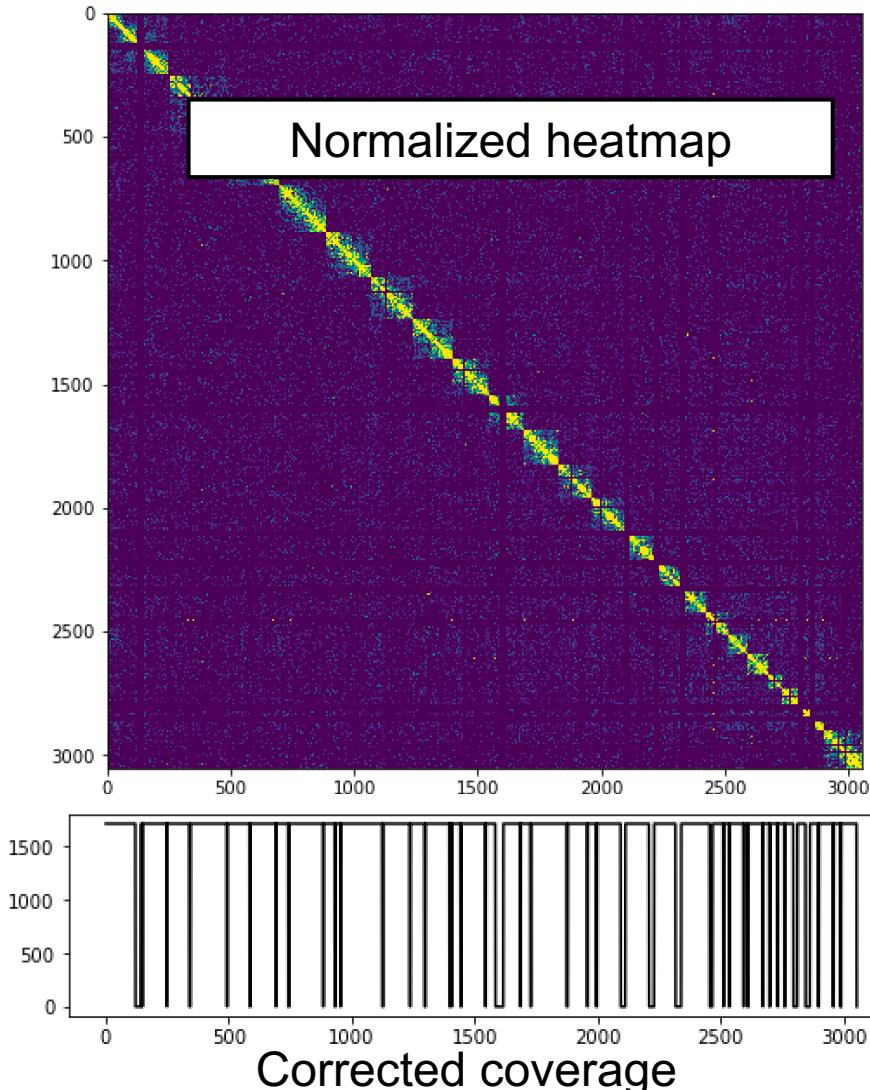
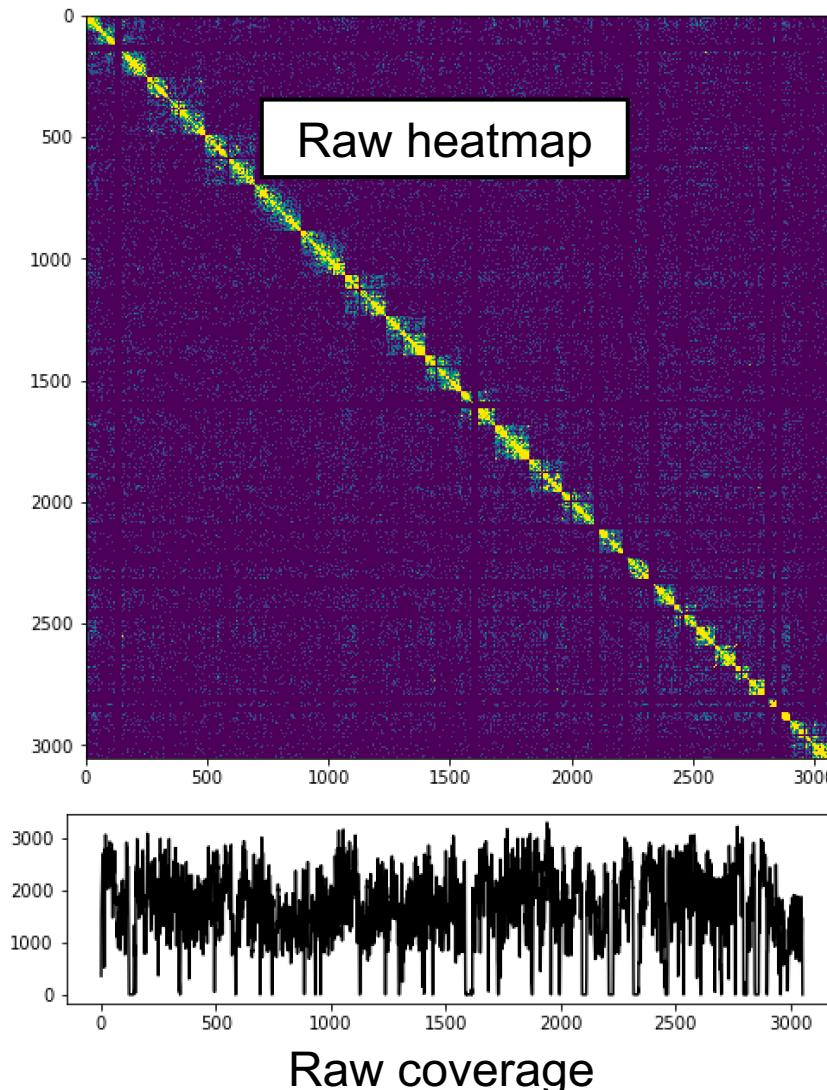
制限酵素断片長、GC含量、マッパビリティなど、バイアスを列挙（それぞれゲノム配列のみから計算可能）、領域ペアの観測確率をそれらのバイアスすべてをパラメータとした確率モデル（ポアソン、負の二項分布など）で表現し、観測値からバイアスパラメータを学習する。
Yaffe and Tanay 2011、HiCNormなど

2. Implicitにバイアスを仮定する手法

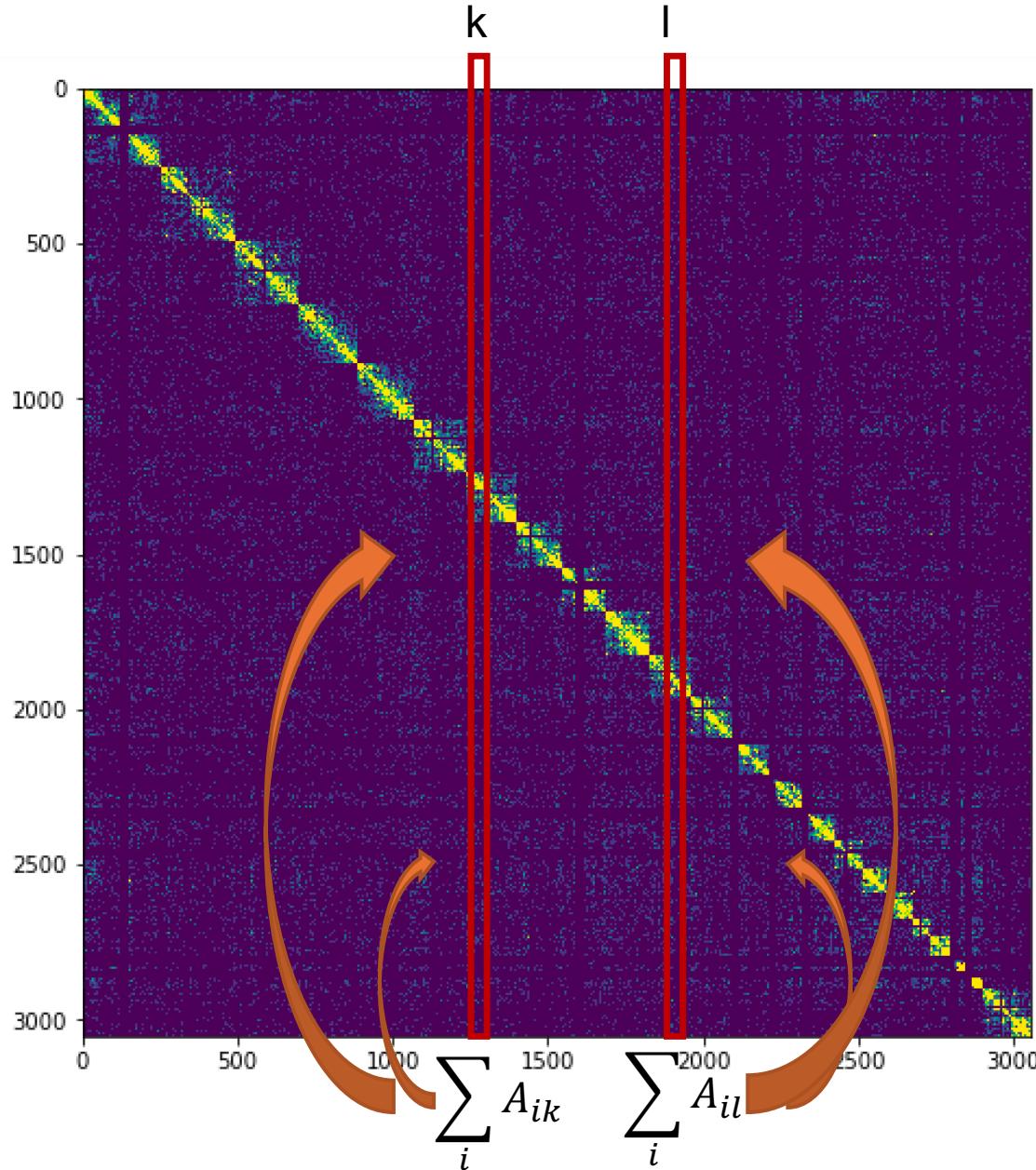
こちらの方が広く使われている。

Vanilla coverage, ICE, Knight and Ruiz 2012など

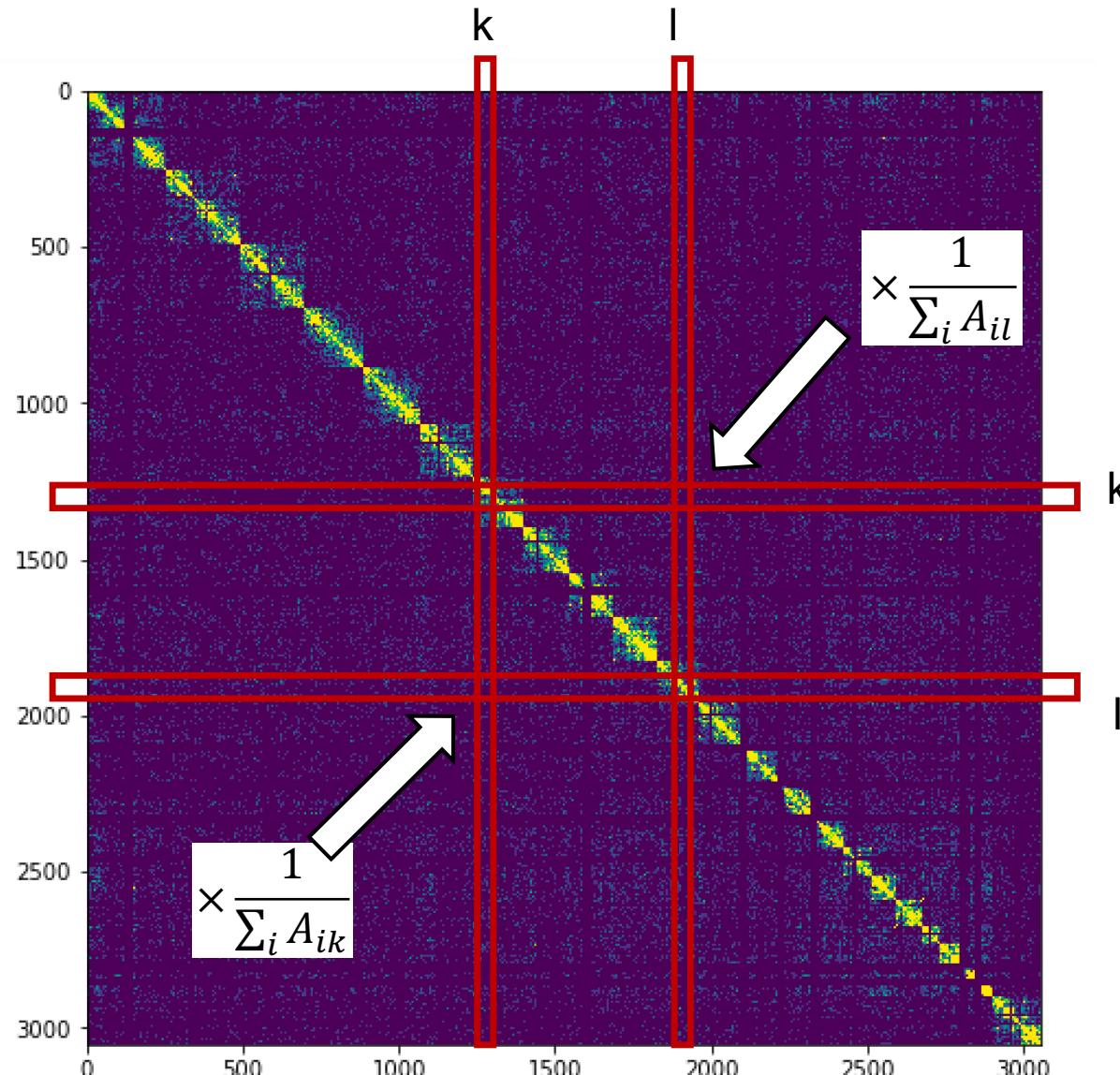
理想的な（正規化された）コンタクトマップでは、
ゲノム上のカバレッジが一定



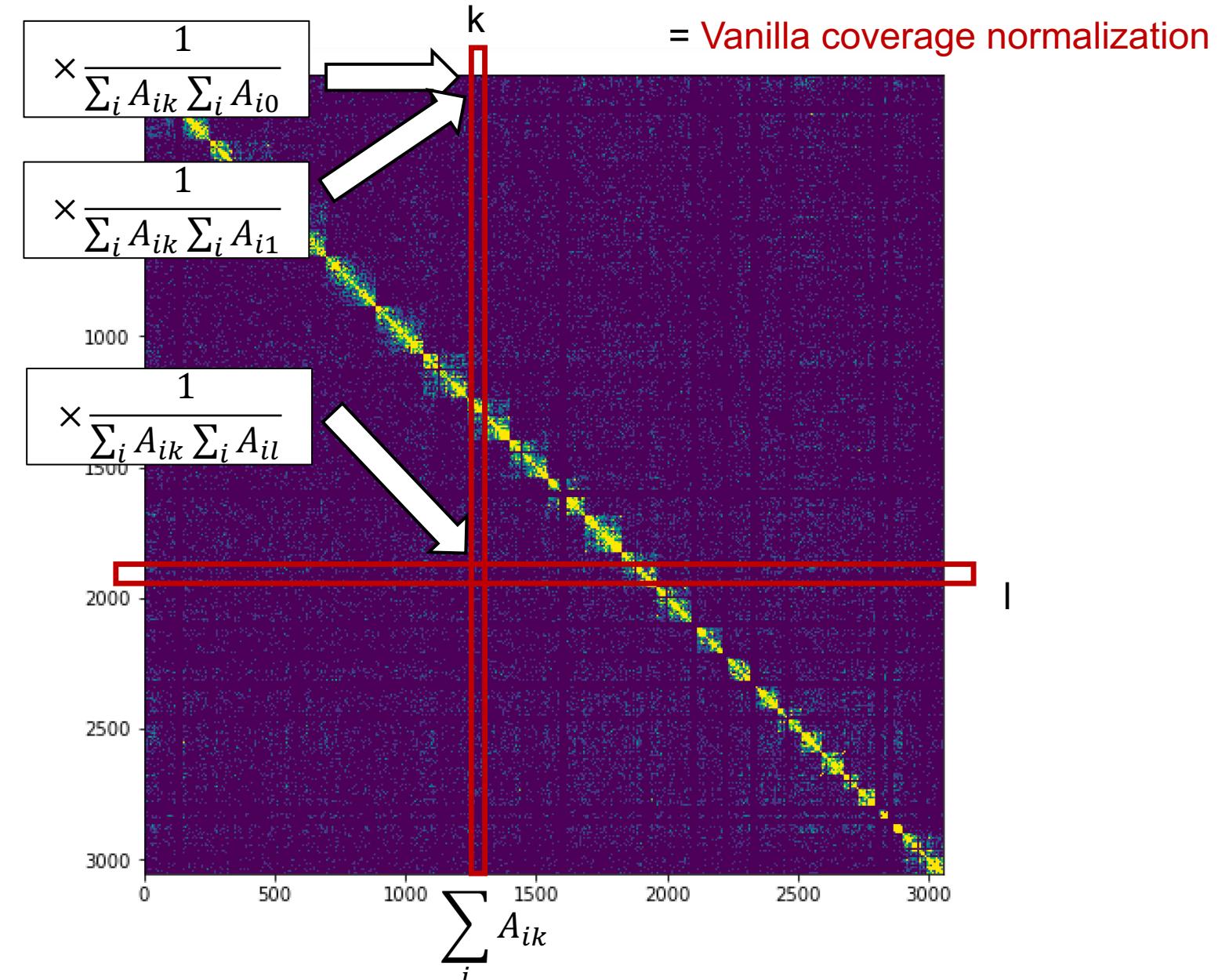
各列の和を計算して割り算すると…



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

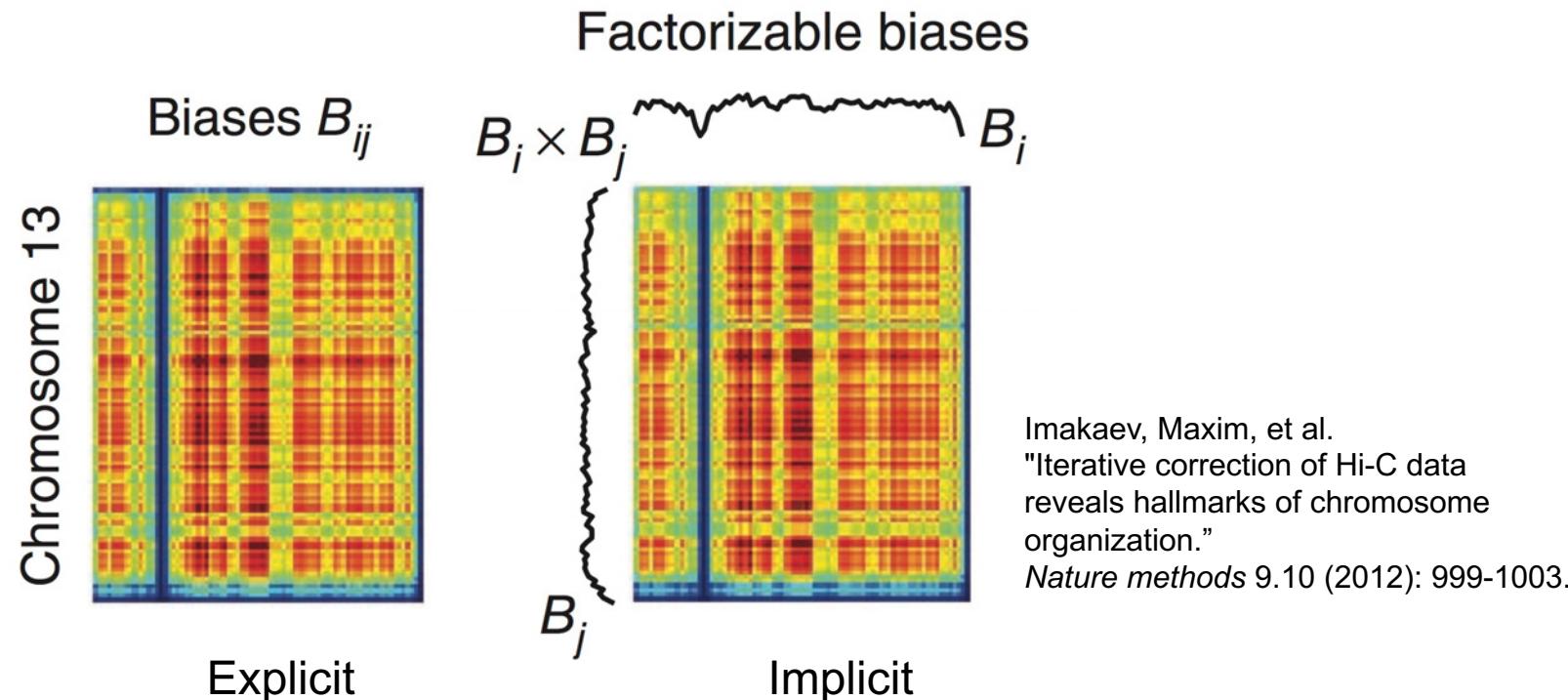


Vanilla coverage normalizationの仮定

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。



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

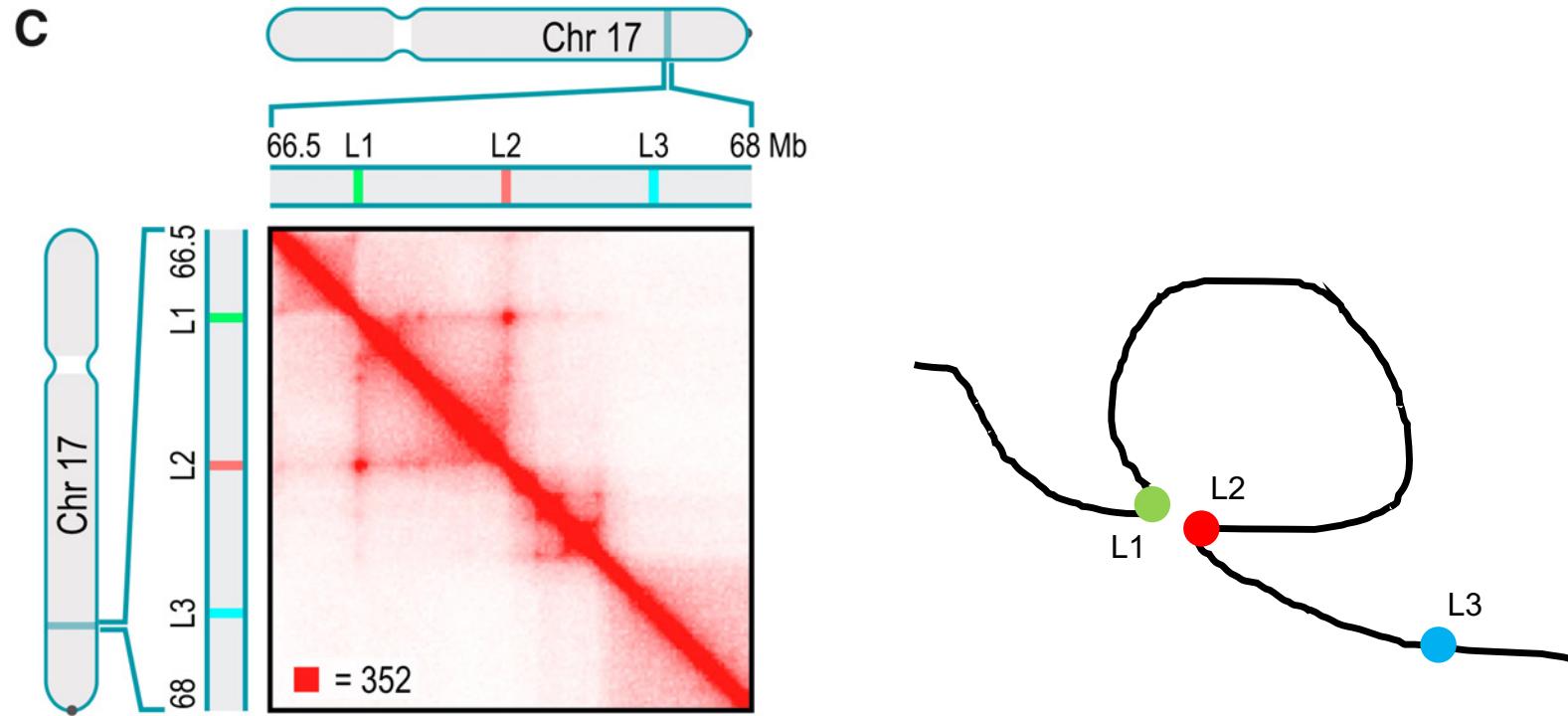
このような行列の補正手法は、“matrix balancing”と呼ばれ、歴史的に何度も再発明されてきた。

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。

Hi-C解析の流れ

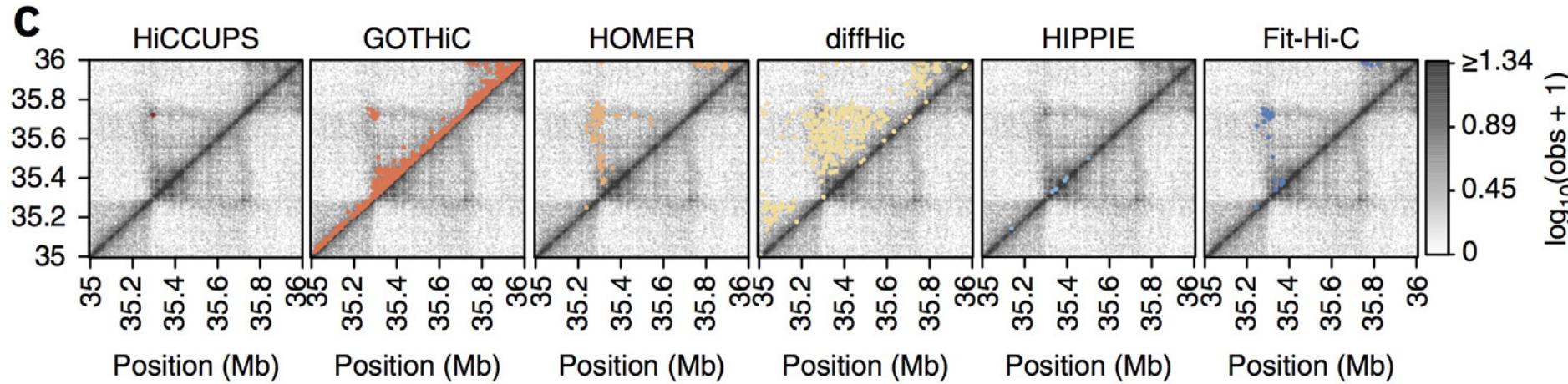


コンタクトマップ上のピーク検出 =特に相互作用の強い領域ペアを特定する



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

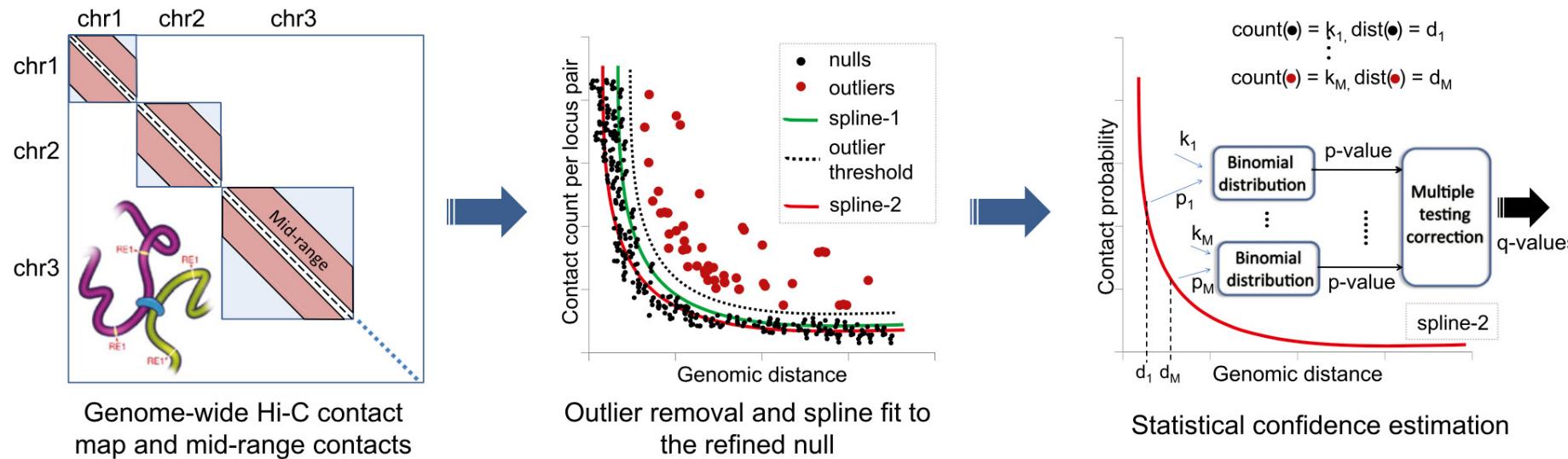


Forcato, Mattia, et al. "Comparison of computational methods for Hi-C data analysis." *Nature methods* 14.7 (2017): 679.

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。

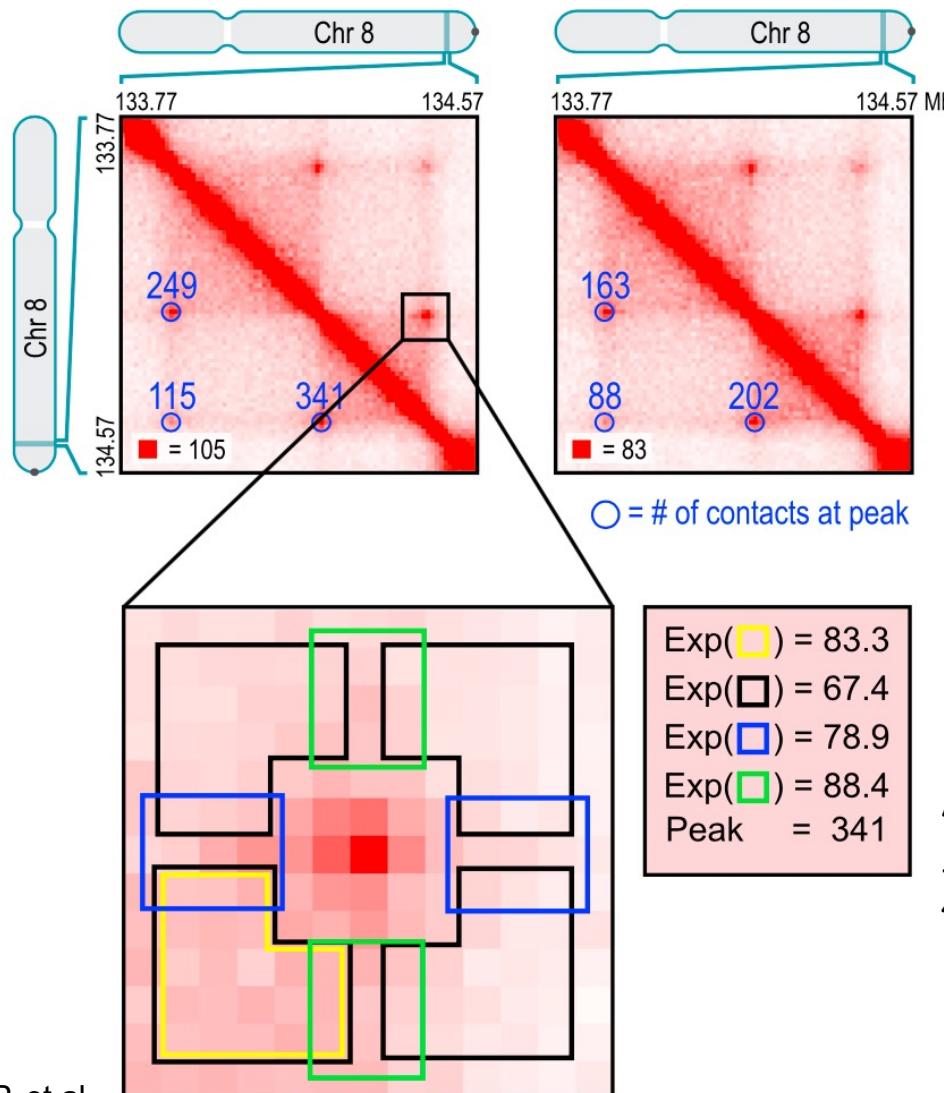
Fit-Hi-C (Global background)

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.

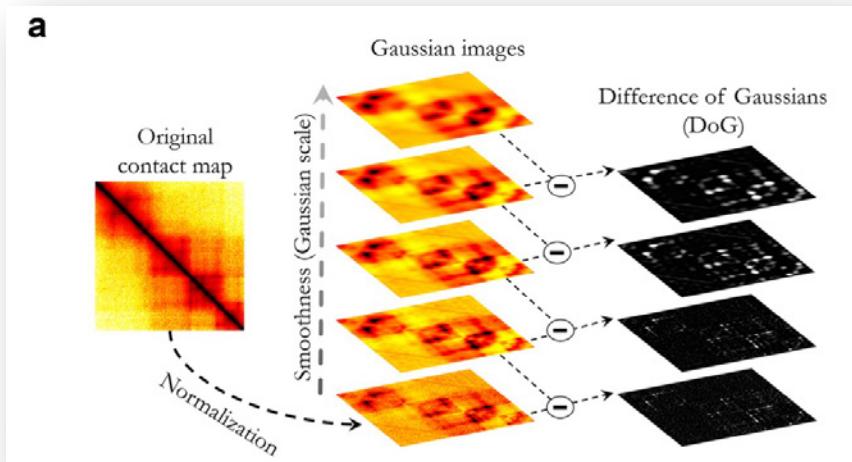
"A 3D map of the human genome at kilobase
resolution reveals principles of chromatin looping"
Cell 159.7 (2014): 1665-1680.

比較的新しいループコーラー

MUSTACHE

ローカルバックグラウンド。

画像処理におけるScale space理論に基づく特徴検出。



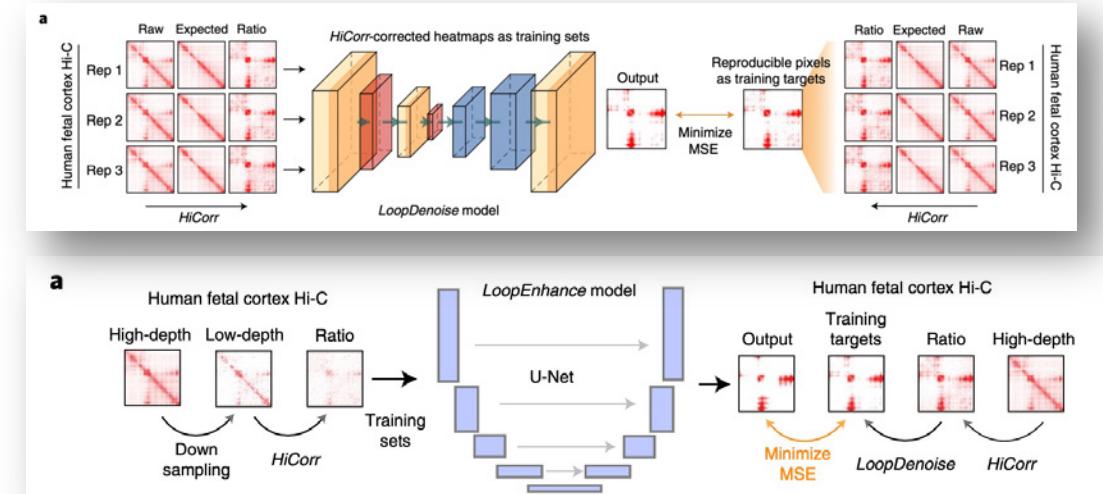
Roayaei Ardakany, Abbas, et al.

"Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation."

Genome biology 21 (2020): 1-17.

DeepLoop

ディープラーニングベース。畳み込みオートエンコーダ。
シングルセルHi-Cなどスパースなデータで有効。



Zhang, Shanshan, et al.

"DeepLoop robustly maps chromatin interactions from sparse allele-resolved or single-cell Hi-C data at kilobase resolution."

Nature genetics 54.7 (2022): 1013-1025.

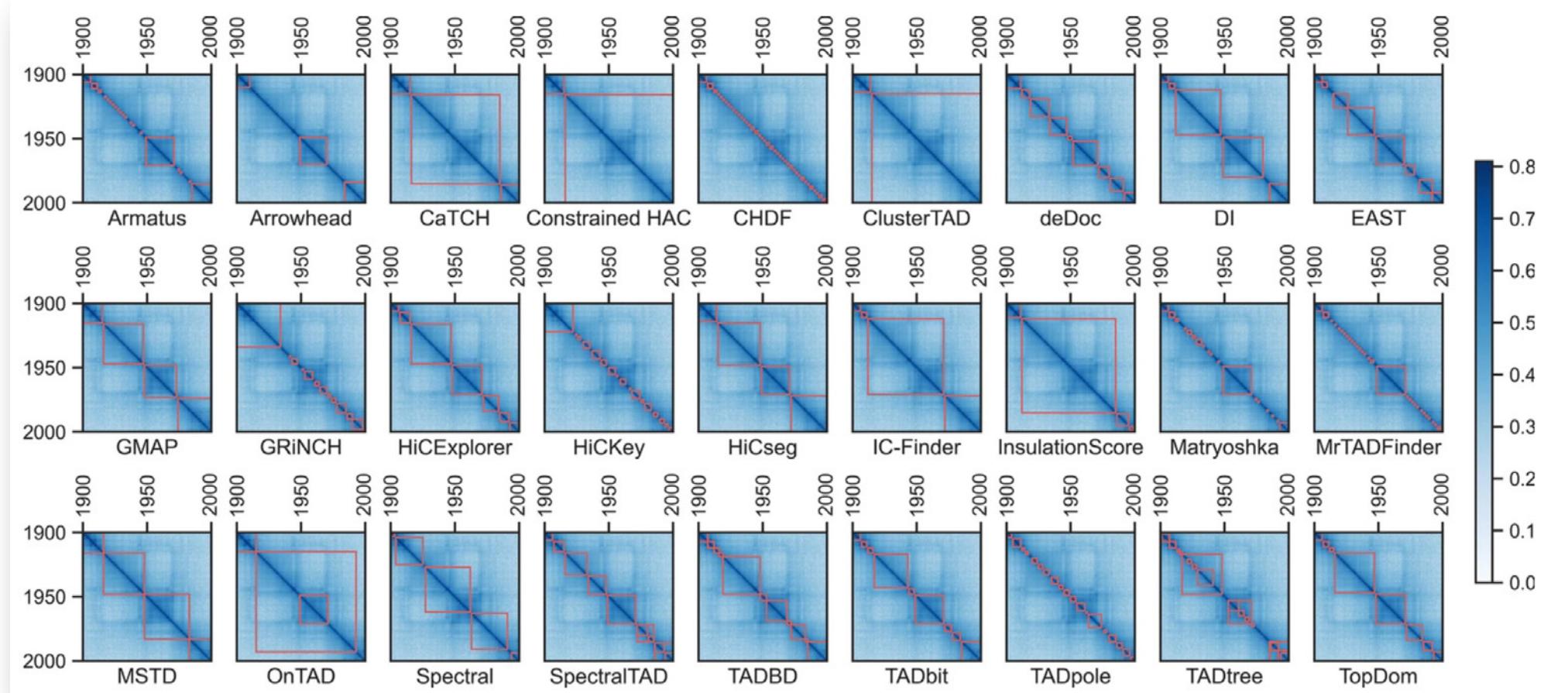
最近のベンチマーク※では、低解像度コンタクトマップについては HiCCUPSとFit-Hi-C、
高解像度 (500bp~10kbp) では MUSTACHE, DeepLoop のループコール精度 (replicates間の一貫性) が
高いと言われている。

※ Liu, Li, et al. "A comprehensive review of bioinformatics tools for chromatin loop calling."
Briefings in Bioinformatics 24.2 (2023): bbad072.

Hi-C解析の流れ



Topologically Associating Domains (TADs)の検出



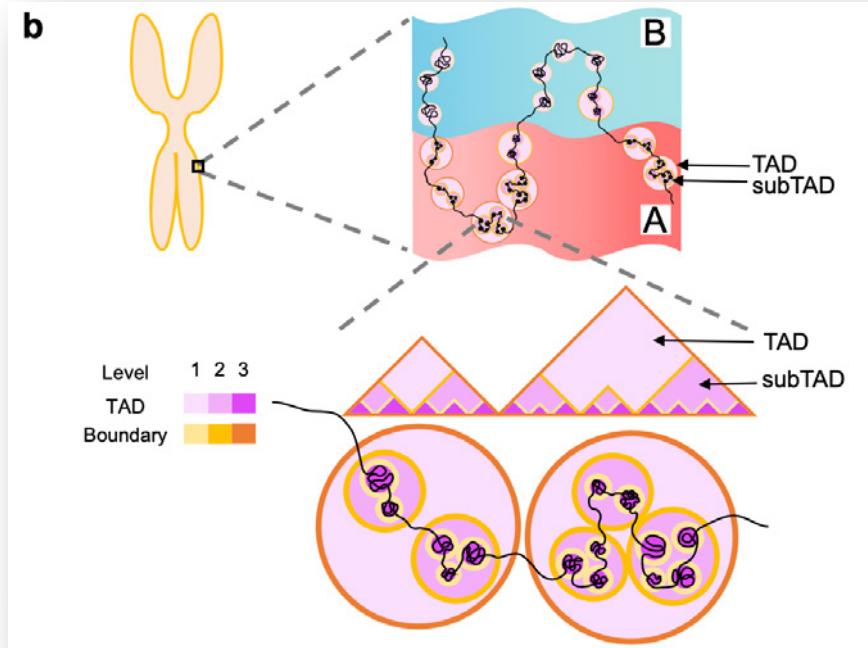
Sefer, Emre. "A comparison of topologically associating domain callers over mammals at high resolution." *BMC bioinformatics* 23.1 (2022): 127.

得られるTADのサイズや数はツールによってさまざま。

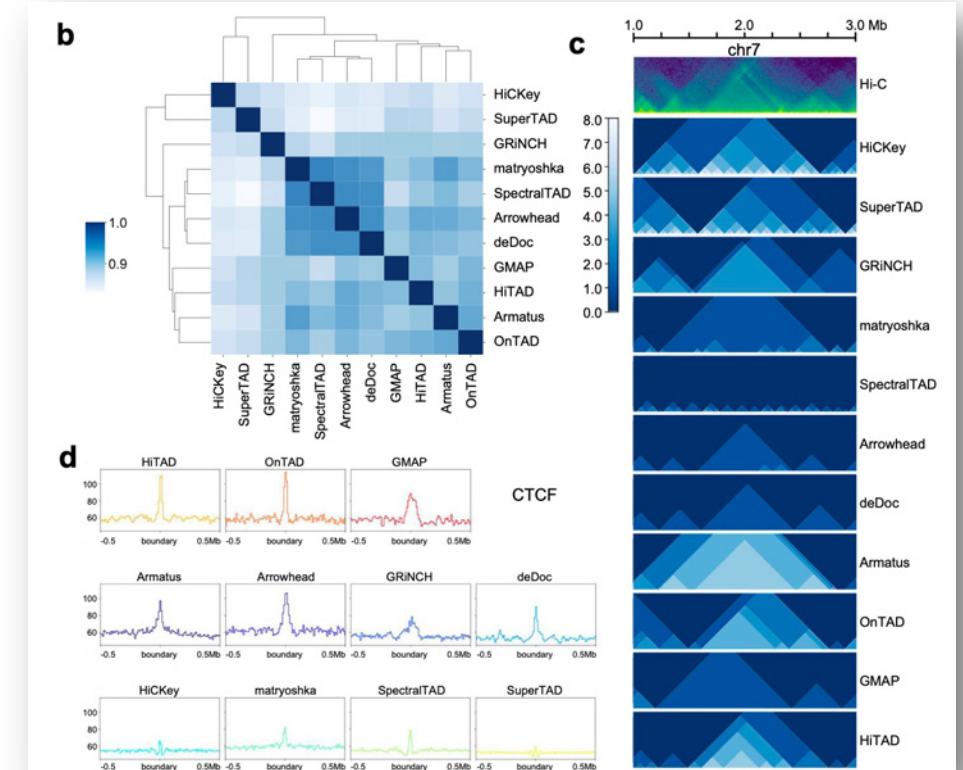
異なる解像度のコンタクトマップでも安定して同様のTADが得られるか、などを検討することが大事。

階層的TADの検出

ゲノムの一次元的な分割によるTADコードではなく、
階層的TAD (TAD, subTAD...) を検出するツールもたくさんある。



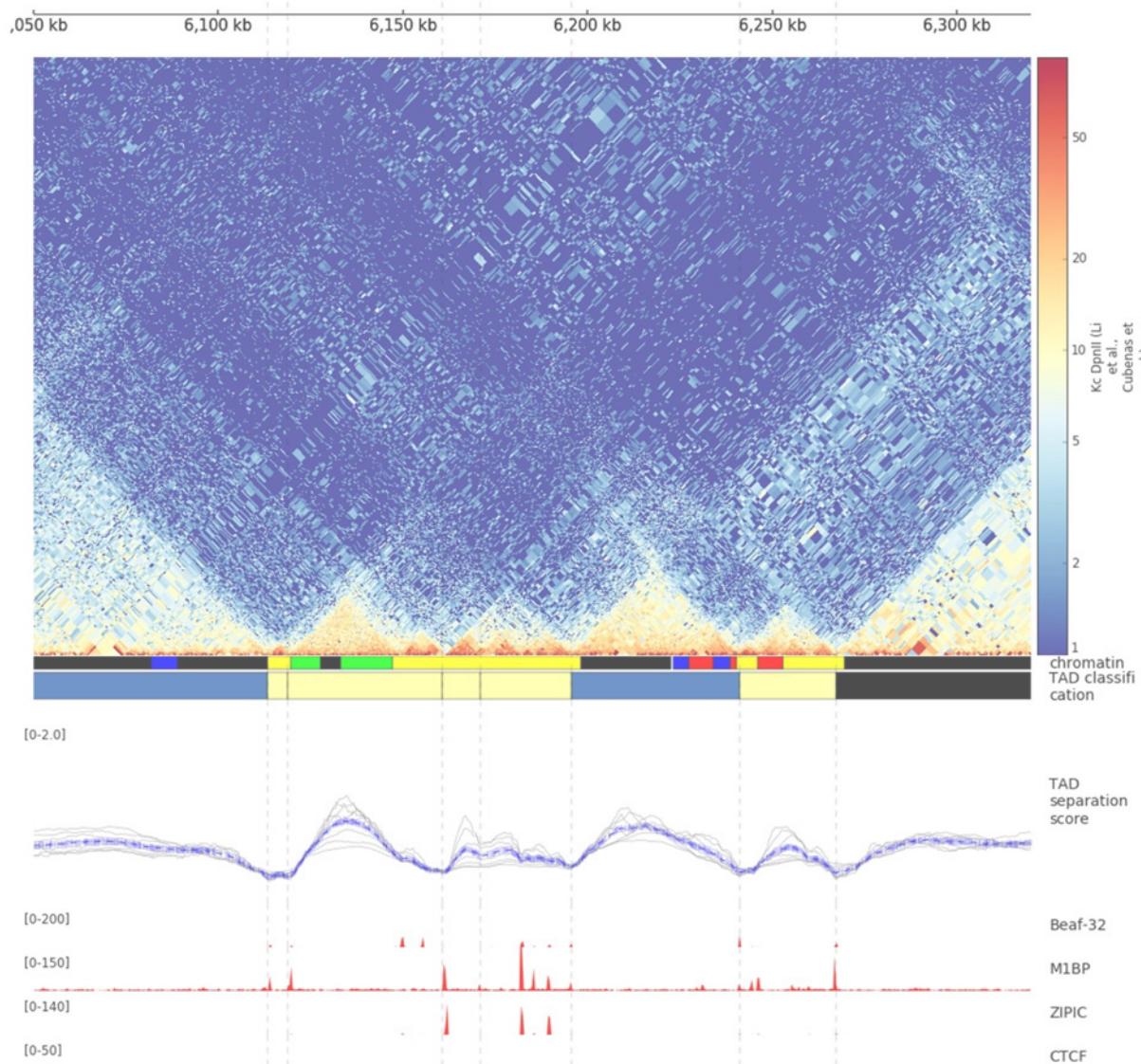
Xu, Jingxuan, et al. "A comprehensive benchmarking with interpretation and operational guidance for the hierarchy of topologically associating domains." *Nature Communications* 15.1 (2024): 4376.



前スライドのベンチマークでも、こっちのベンチマークでも、OnTAD ※ の評価が比較的高い。

※ An, Lin, et al. "OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries." *Genome biology* 20 (2019): 1-16.

hicexplorer.hicFindTADsのアルゴリズム



z-score変換したコンタクトマップを使う。

あるサイズの正方形でゲノムの端からスライディングウインドウを構成し、正方形内部の数値の平均値を計算する。

その値が local minimum をとる座標が、TADの境界とみなせる。

正方形を前後に少しずらした場合の内部の数値セットとwilcoxonの順位和検定を実行してp-valueを付与。
→ 多重検定補正

事前のご質問：Hi-C解析の計算コスト

計算コストはそこまで大きくない。
マウス (mm10)、5k解像度、
1億リードペア x 4サンプル、
で右のような感じ。

マッピングの部分でCPU消費、
コンタクトマップ正規化やaggregateの部分で
メモリ消費が激しくらい。
それでもせいぜい数GB程度。

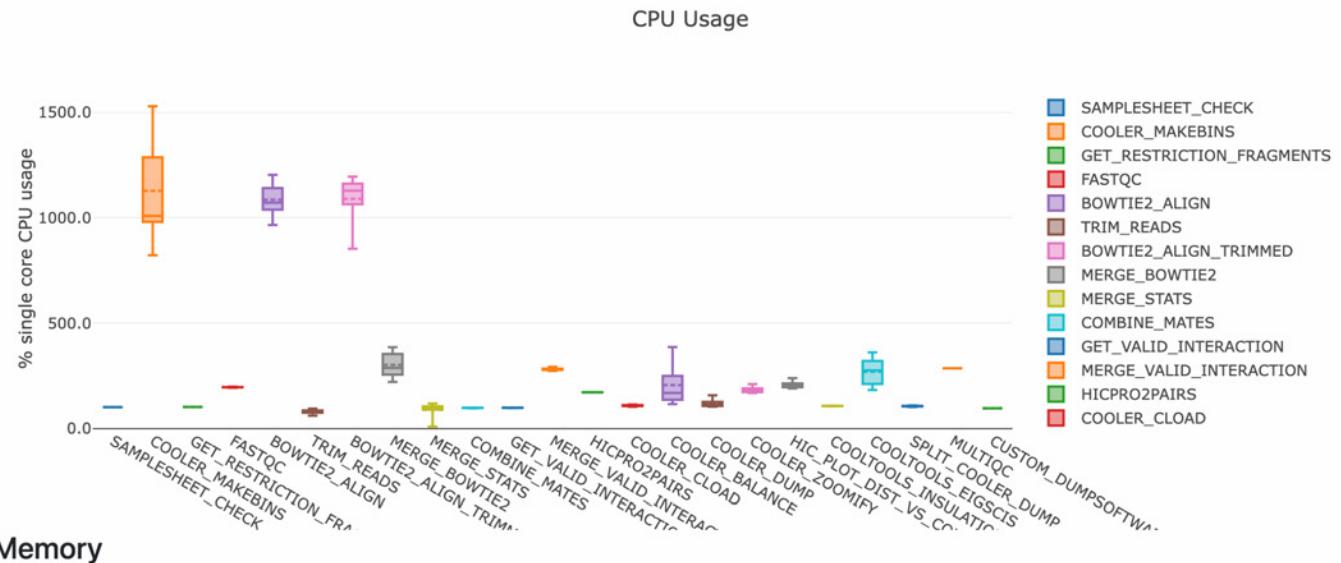
むしろストレージ消費が激しい。
途中結果の書き込みで数百GBくらい
すぐなくなるので、サンプル数がそこそそ
ある場合は数テラバイトくらいの
ストレージが必要。

Resource Usage

These plots give an overview of the distribution of resource usage for each process.

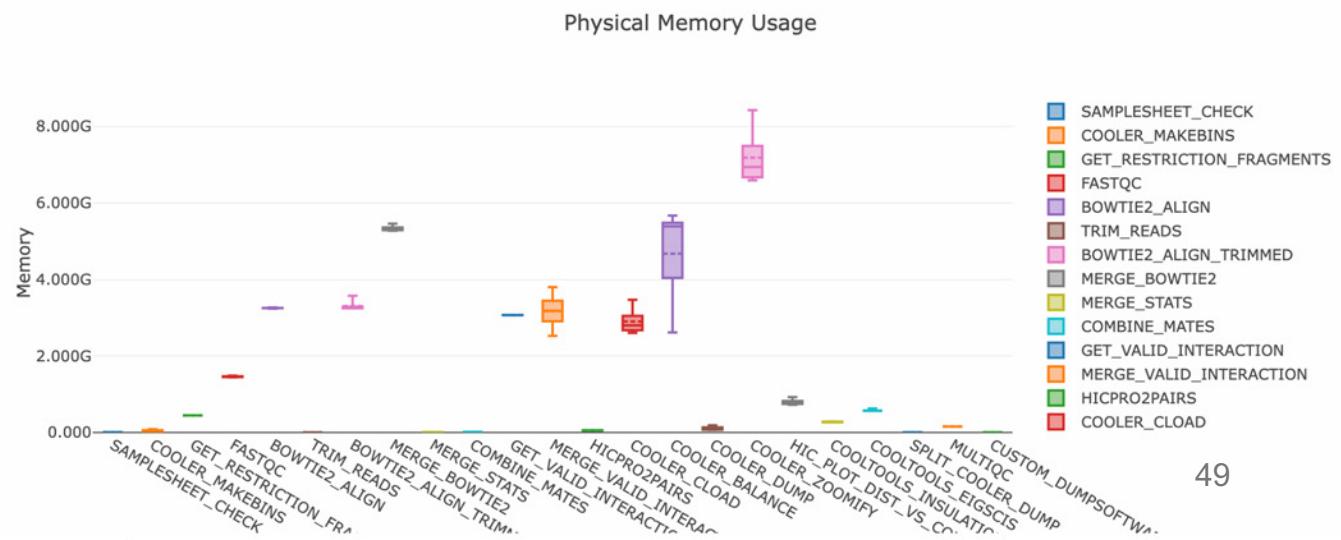
CPU

Raw Usage % Allocated



Memory

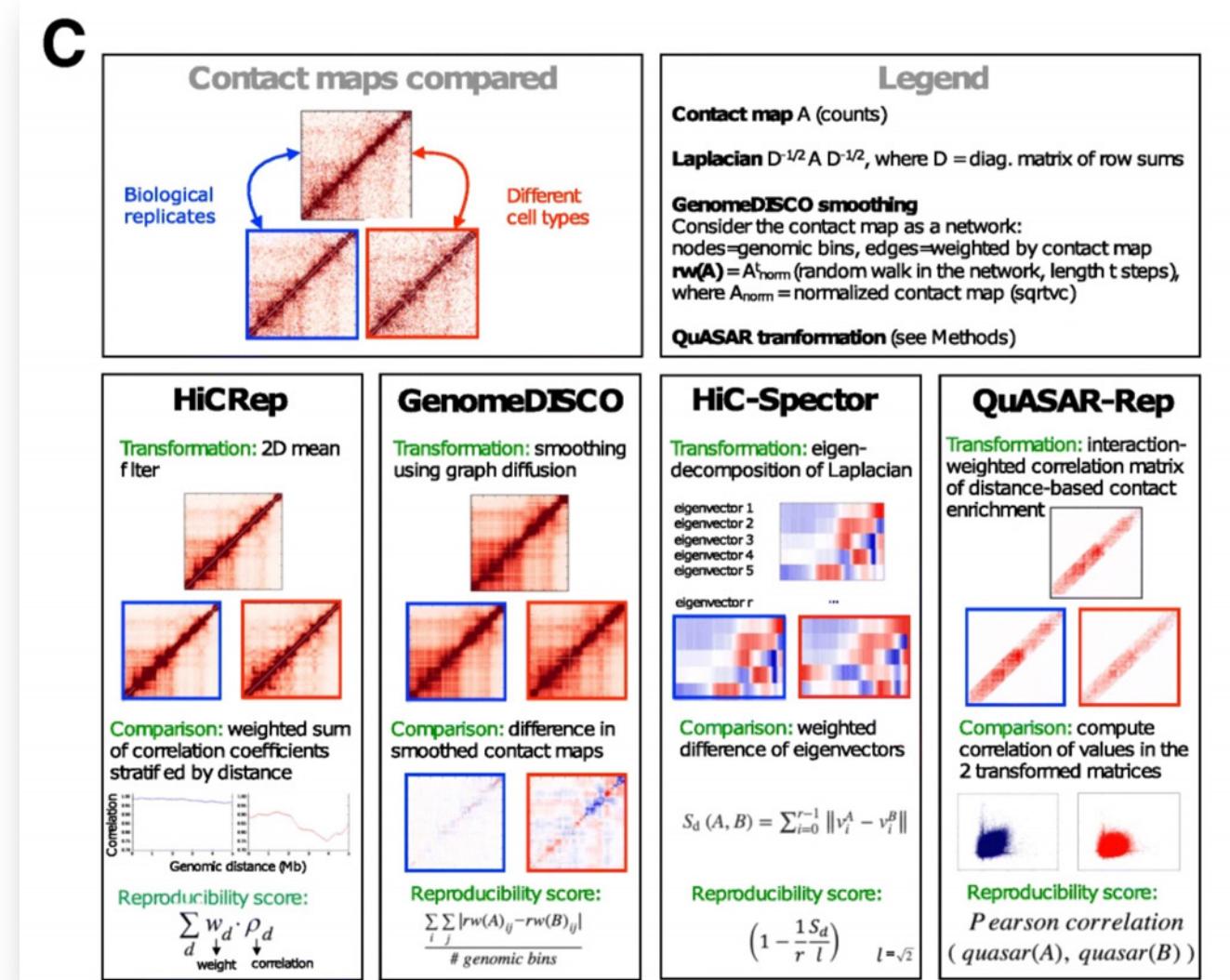
Physical (RAM) Virtual (RAM + Disk swap) % RAM Allocated



事前のご質問：どのようにしたら再現性を確認できるか。

https://github.com/kundajelab/3DChromatin_ReplicateQC

Yardımcı, Galip Gürkan, et al. "Measuring the reproducibility and quality of Hi-C data." *Genome biology* 20 (2019): 1-19.

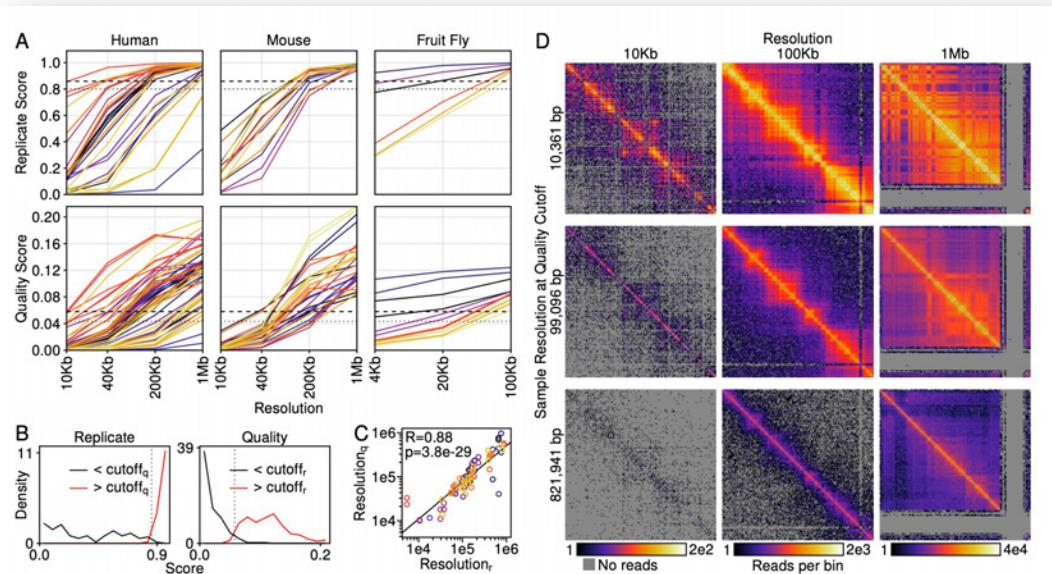


事前のご質問：「解像度」の考え方・設定の方法

アドホックに決められていることが多い。5kb, 10kb, 50kbなど。

「適切な解像度」は、以下の二点に依存する。

1. フラグメントサイズ（制限酵素、DNase/MNaseの選択）
 - 6 cutterなら約4kbp、4 cutterなら256bp程度のサイズ
 - 実際のところ、この10倍くらいのサイズが選択されることが多い
 - 6 cutterなら50k程度、4 cutterなら2k程度以上
 - Micro-Cでも最大解像度はふつう1k程度に設定
2. シーケンスデプス（妥当なコンタクトの観測数）
 - 妥当なコンタクトの総数は、実験のクオリティが高いサンプルであっても、トータルリードペア（シーケンサー出力）の50%程度の数になる。（マッピング時、コンタクトパターンによるフィルタリングのため）
 - Replicate間一貫性に基づく評価研究から、40kbp解像度の場合、2000万以上の妥当なコンタクトが必要と報告されている。
 - なので、マッピング前のトータルリードペアは4000万以上、現実としては1億リードペア、さらに高い解像度なら数億リードペアくらいあったほうがいい。



- レプリケートで一貫した二次解析結果（TAD, ループなど）が得られるか確認する
- QuASARによるシングルサンプルクオリティ評価（近傍のコンタクトパターン一貫性によるスコア化）

Sauria, Michael EG, and James Taylor. "QuASAR: quality assessment of spatial arrangement reproducibility in Hi-C data." *BioRxiv* (2017): 204438.

事前のご質問：TADやゲノムループレベルでのサンプル間比較解析を知りたい。

あまり統計的に評価されないケースが少なくないと思う。

研究者の興味のある領域のコンタクトマップのFigureを並べて、ループ検出・TAD検出ツールの結果を描画し、WT vs. KDの違いをビジュアル的に評価するだけのこと。

これまでコストの問題もあって、あまり大量のサンプルでコンタクトマップを生成できないことも多かったので、大量比較の需要が大きくななく、ツール開発が進まなかつたのかもしれない。

と思っていたが、東大の中戸先生のラボがまさにドンピシャのツールを開発してくれていた！
TogoTVで関連した解説動画も公開されているので、そちらもぜひ参照。

JOURNAL ARTICLE

HiC1Dmetrics: framework to extract various one-dimensional features from chromosome structure data

Jiankang Wang, Ryuichiro Nakato

Briefings in Bioinformatics, Volume 23, Issue 1, January 2022, bbab509, <https://doi.org/10.1093/bib/bbab509>

Published: 01 December 2021 Article history ▾

Article | Open access | Published: 19 September 2023

Context-dependent perturbations in chromatin folding and the transcriptome by cohesin and related factors

Ryuichiro Nakato , Toyonori Sakata, Jiankang Wang, Luis Augusto Ejij Nagai, Yuya Nagaoka, Gina Miku Oba, Masashige Bando & Katsuhiko Shirahige

Nature Communications 14, Article number: 5647 (2023) | Cite this article

6477 Accesses | 8 Citations | 14 Altmetric | Metrics

HiC1Dmetricsを用いたゲノム立体構造解析

バイオインフォマティクスツール・パッケージを自作する

4. HiC1Dmetricsを用いたゲノム立体構造解析 @ Bio“Pack”athon2024#10

見どころダイジェスト

00:30 1. 1次目

01:07 2. HiCを用いた立体構造解析

01:08 3. ゲノムは核内で規則的に折りたまれている

03:54 4. Hi-Cを用いた立体構造解析の原理

07:05 5. Hi-C法で観測される階層的ゲノム構造

08:33 6. トポジカルドメイン (TAD)

10:09 7. 異なる階層は異なる因子によって制御されている

11:53 8. サンプル間比較 (目視、ドグ比)

14:55 9. 多サンプル解析の場合