

平成28年度NGSハンズオン講習会 ChIP-seq

2016年7月28日

amelieff

本講義にあたって

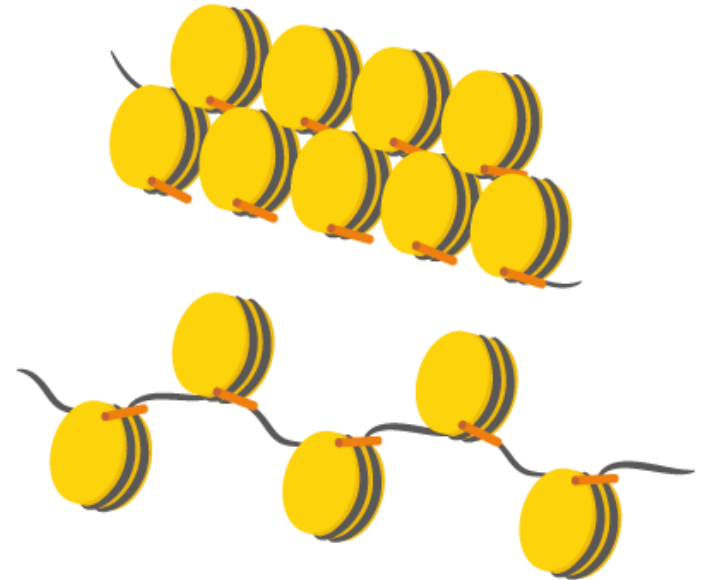
- 代表的な解析の流れを紹介します。
 - 論文でよく使用されているツールを使用します。
- コマンドを沢山実行します。
 - タイプミスが心配な方は、コマンド例がありますのでコピーして実行してください。
 - 実行が遅れてもあせらずに、課題や休憩の間に追い付いてください。

本講義の内容

- ChIP-seqとは
- ChIP-seq解析の流れ
- 公開データの取得
- クオリティコントロール
- マッピング
- ピーク検出
- ピークアノテーション
- モチーフ探索
- まとめ
- 最後に

ChIP-seqとは

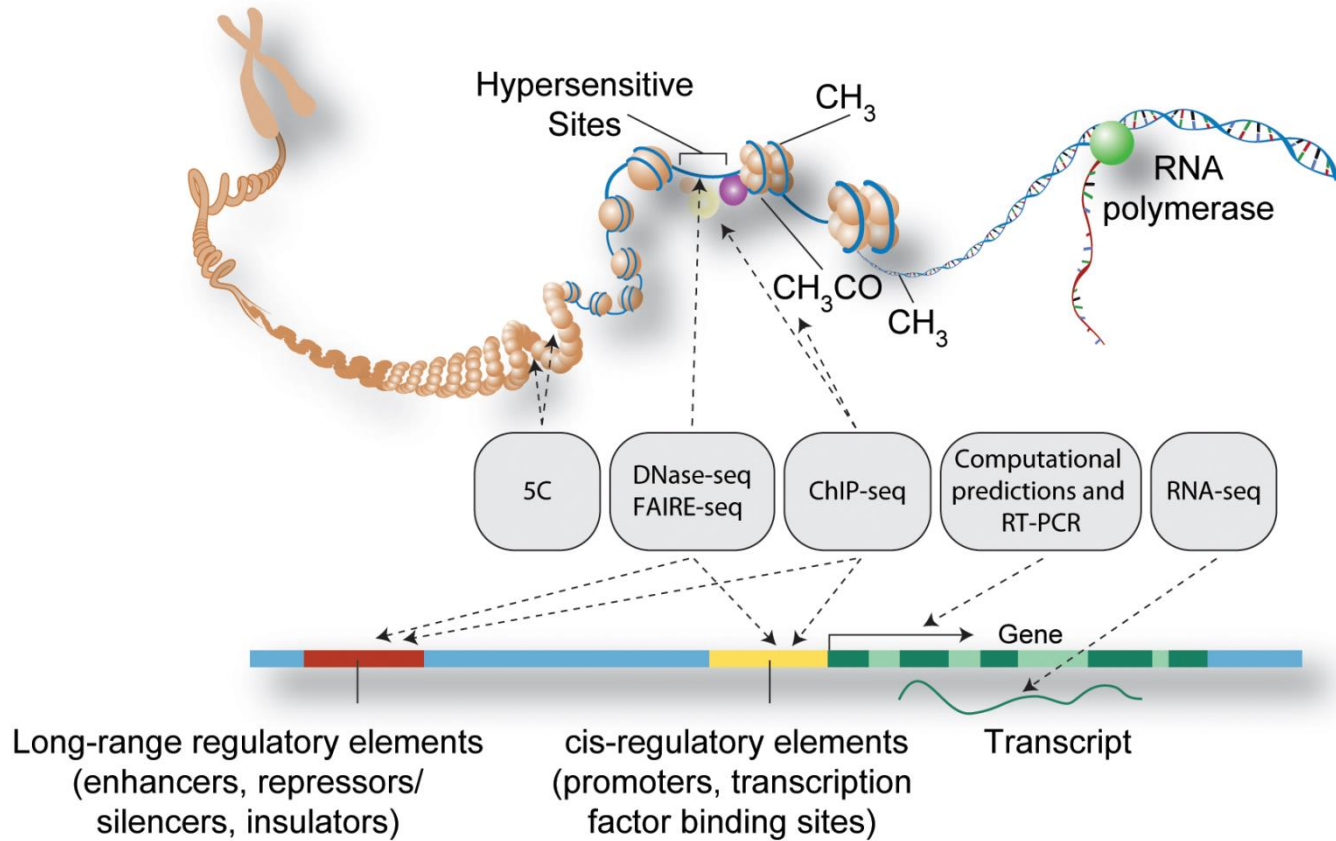
- ChIP(**Ch**romatin **I**mmuno **P**recipitation) + NGS sequencing
 - クロマチン免疫沈降により濃縮したゲノム領域をシーケンスする手法
- 主な解析対象
 - タンパクとDNAの相互作用
 - ヒストン修飾



•Licensed under CC-BY 4.0 ©Togo picture gallery by DBCLS

ChIP-seqとは

ChIP-seqで主に解析されるのは転写調節領域



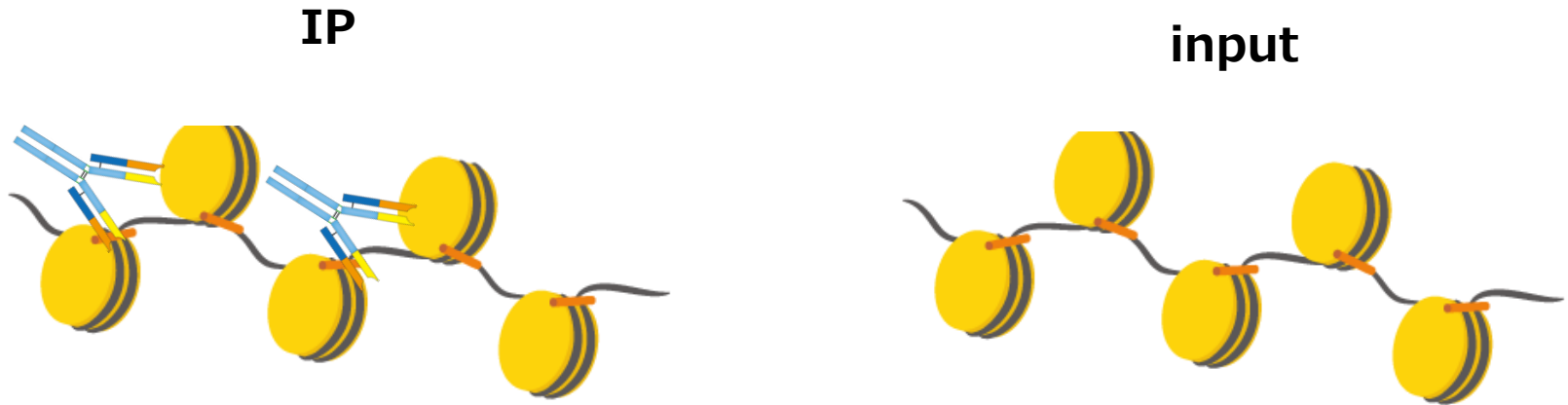
A User's Guide to the Encyclopedia of DNA Elements (ENCODE), 2011より

ChIP-seqとは | input と IP

ChIP-seqでは、免疫沈降のバックグラウンドノイズを削減するため、コントロールを使用することが多い

免疫沈降（IP）を行っていないサンプルをコントロールとして使用し、検出したピークを抗体に非特異的なものとして取り除くために用いる

一般にこのコントロールを **input** と呼ぶ



•Licensed under CC-BY 4.0 ©Togo picture gallery by DBCLS

ChIP-seqとは

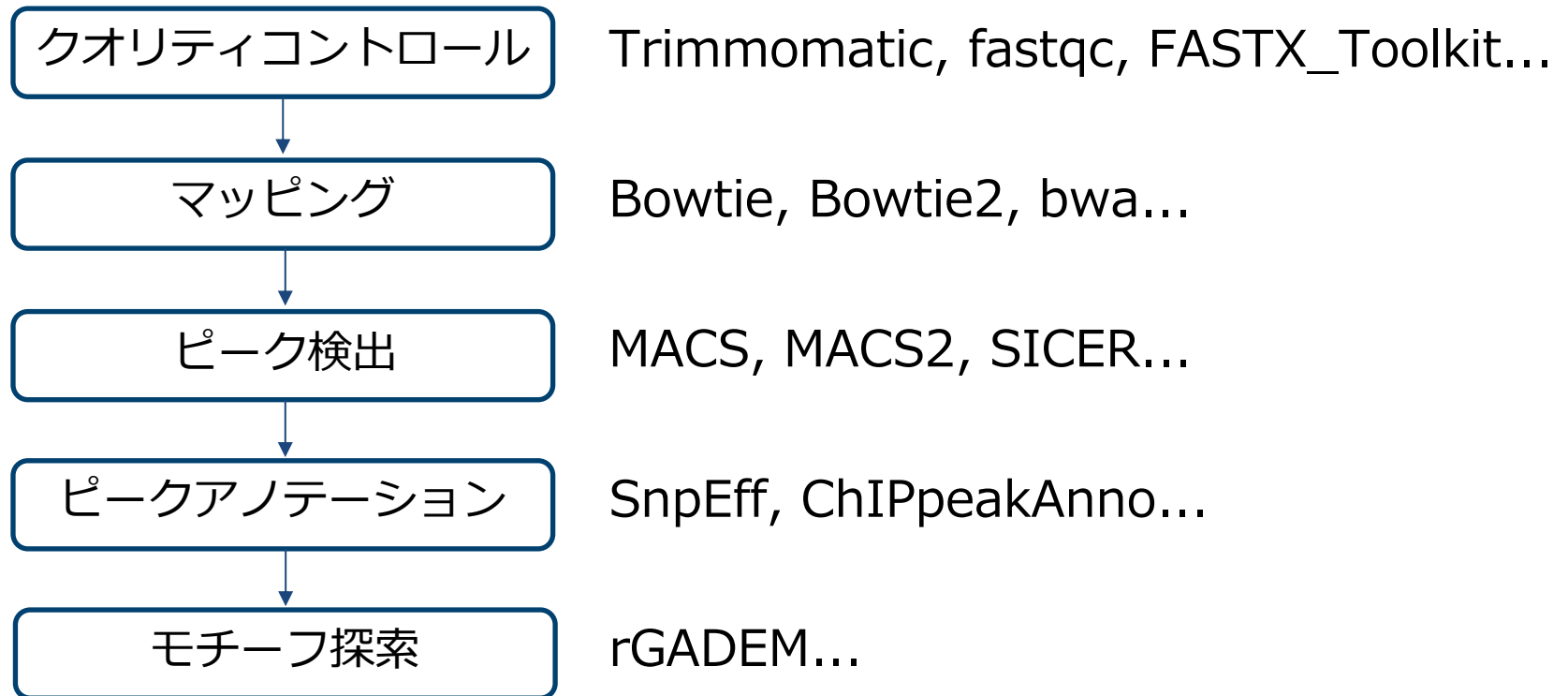
ChIP-seqでは、抗体が特異的に結合した領域をピークとして得る

—— : シーケンスリード



A User's Guide to the Encyclopedia of DNA Elements (ENCODE), 2011より

ChIP-seq解析の流れ | 代表的なソフト



- ChIP-seq解析の一般的な流れであり、全てのChIP-seqで同一の解析を行うわけではない
- 研究の目的やデータに合わせて、最適な解析を設計

クオリティコントロール

他のNGSデータ解析と同様に、解析前のクオリティコントロールを実施

■ 本日使用するソフト

- トリミング・低クオリティリードの除去
 - Trimmomatic
- クオリティチェック
 - Fastqc

■ ChIP-seqにおけるポイント

- リード長に注意する（75 bp以下など短い場合が多い）

マッピング

Reseqでも使用されるマッピングソフトがChIP-seqでよく使用される

■ 本日使用するソフト

- bowtie2
 - ギャップアラインメントに対応
 - マッピング精度が高い

■ この他に使用されるソフト

- bowtie
 - ギャップアラインメントに非対応
- bwa

ピーク検出

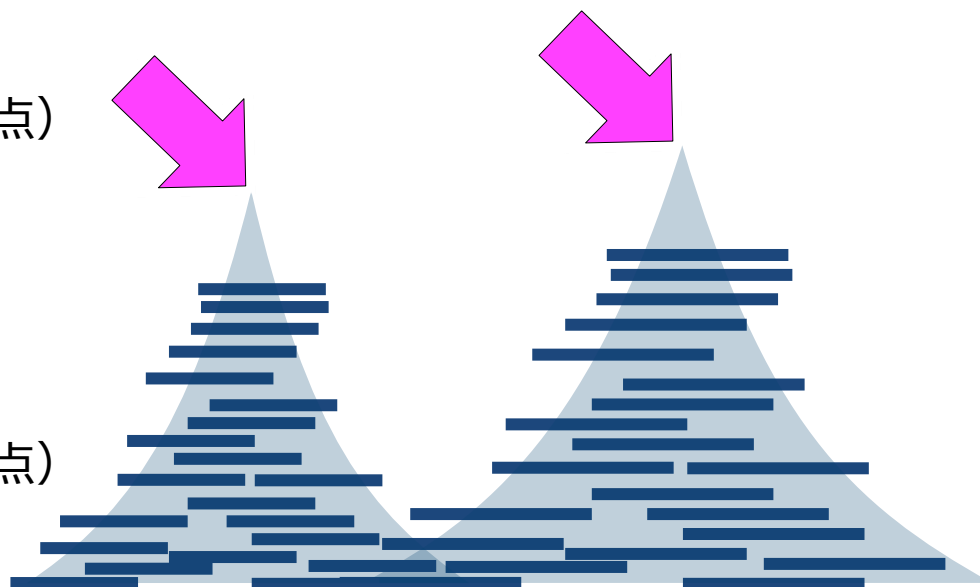
ピーク検出ソフトはIPで濃縮した領域のリードの頂点を検出する

■ 本日使用するソフト

- MACS2 (デファクトスタンダード)
 - 被引用数 2750件
(2016年7月20日時点)

■ この他に使用されるソフト

- SICER
 - 被引用数 425件
(2016年7月20日時点)
- MACS



ピークアノテーション

ピーク検出後に、ピークがゲノム上のどのような位置に存在するのかアノテーションする

■ 本日使用するソフト

- SnpEff

- 遺伝子名を付与
- 遺伝子上のドメイン（エキソン、上流など）を付与
- 様々な生物種に対応

■ この他に使用されるソフト

- ChIPpeakAnno
 - Rパッケージ

モチーフ探索

検出されたピークに共通のモチーフを探索する

モチーフは、抗体と結合する短い配列で、ピーク配列に共通して見られる

■ 本日使用するソフト

- rGADEM
- Artistic License 2.0（改変、再配布、商用可）なので利用しやすい

■ この他に使用されるソフト

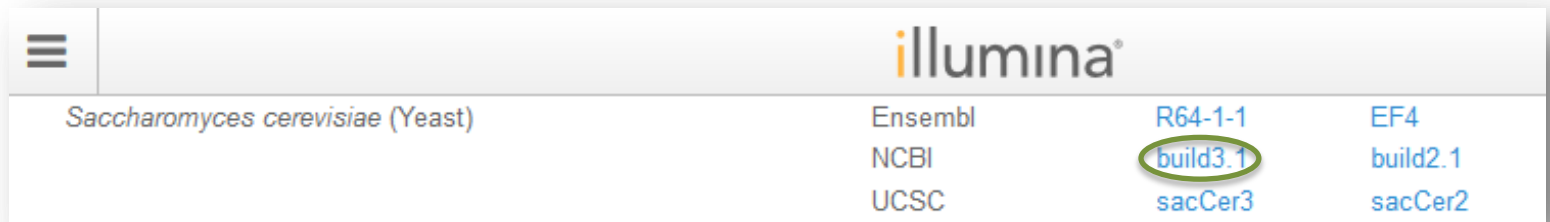
- MEME
 - 商用利用不可

公開データの取得

今回の解析に必要なデータ

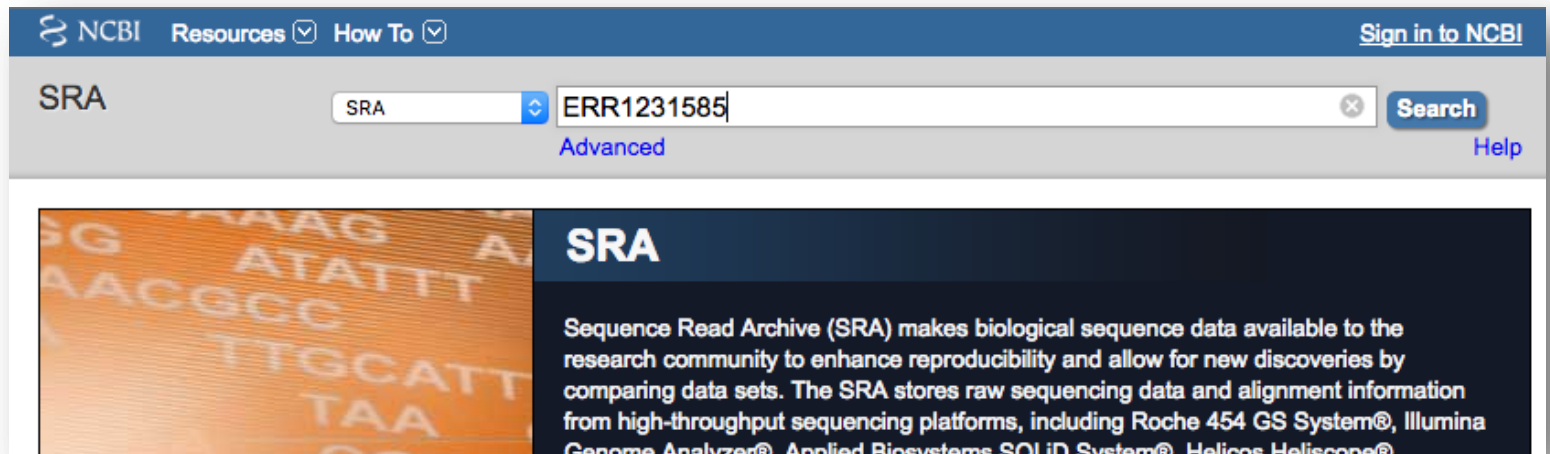
■ リファレンスゲノム (実行済み)

- http://support.illumina.com/sequencing/sequencing_software/igenome.html



	Ensembl	R64-1-1	EF4
<i>Saccharomyces cerevisiae</i> (Yeast)	NCBI	build3.1	build2.1
	UCSC	sacCer3	sacCer2

■ 解析対象のシーケンスデータ (実行済み)



NCBI Resources How To Sign in to NCBI

SRA SRA ERR1231585 Search Help

SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®

公開データの取得

酵母のリファレンスゲノムデータの取得方法

```
$ wget ftp://igenome:G3nom3s4u@usd-  
ftp.illumina.com/Saccharomyces_cerevisiae/NCBI/build3.1/Saccha  
romyces_cerevisiae_NCBI_build3.1.tar.gz  
$ tar zxvf Saccharomyces_cerevisiae_NCBI_build3.1.tar.gz
```

Saccharomyces cerevisiaeのリファレンスゲノムをイルミナのWebページからダウンロードし解凍(実行済み)

```
$ ls -l /home/ユーザー名/Desktop/amelieff/sacCer3/
```

```
          :  
-rwxr-xr-x. 1 root root 12400379  5月 23 11:09 2016 genome.fa  
-rwxr-xr-x. 1 root root      462  5月 23 11:09 2016  
genome.fa.fai  
-rwxr--r--. 1 root root   19041  5月 23 11:10 2016 mask.gtf  
-rwxr-xr-x. 1 root root  643818  5月 23 11:09 2016 refGene.txt
```

/home/ユーザー名/Desktop/amelieff/Scerevisiae/の
解凍したファイル（今回使用するもののみ）を確認

公開データの取得

fastaファイルの中身の確認

```
$ less /home/ユーザ名/Desktop/amelieff/Scerevisiae/genome.fa
```

```
>chrI  
CCACACCACACCCACACACCCACACACCACACACCACACACCACACACC  
CACACACACACATCCTAACACTACCCTAACACAGCCCTAATCTAACCCCTG  
GCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTTAC  
CCTGTCCCATTCAACCATAACCACTCCGAACCACCATCCATCCCTCTACTT  
ACTACCACTCACCCACCGTTACCCTCCAATTACCCATATCCAACCCACTG  
:
```

1行目： コンティグ名

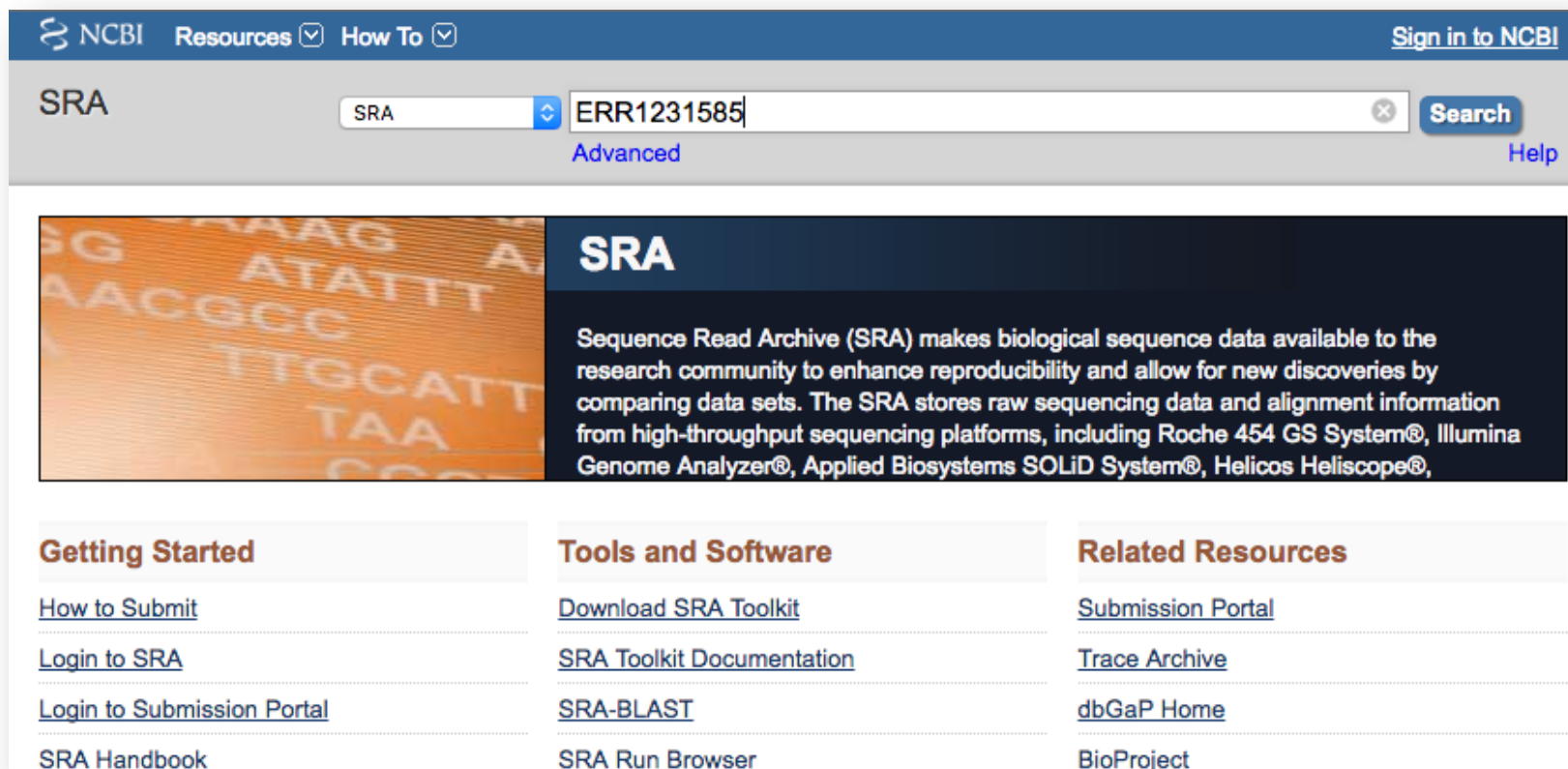
2行目以降： 実際の配列情報

※ 「q」 で閲覧を終了する

公開データの取得 | データの探し方

解析対象のシーケンスデータの取得方法

NCBI SRA (<http://www.ncbi.nlm.nih.gov/sra>) へアクセスする。



The screenshot shows the NCBI SRA website. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' dropdown menus, and a 'Sign in to NCBI' link. Below this is a search bar with 'SRA' selected in a dropdown menu, the search term 'ERR1231585', and a 'Search' button. A 'Help' link is also visible. The main content area features a large image of DNA sequence data on the left and a dark blue box with the text 'SRA' and a description of the Sequence Read Archive. Below this, there are three columns of links: 'Getting Started' (How to Submit, Login to SRA, Login to Submission Portal, SRA Handbook), 'Tools and Software' (Download SRA Toolkit, SRA Toolkit Documentation, SRA-BLAST, SRA Run Browser), and 'Related Resources' (Submission Portal, Trace Archive, dbGaP Home, BioProject).

公開データの取得 | データの探し方

論文中などから得られたアクセッション番号のERR1231585を検索する

NCBI Resources How To to NCBI

SRA SRA ERR1231585 Search Help

Advanced

SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®,

Getting Started	Tools and Software	Related Resources
How to Submit	Download SRA Toolkit	Submission Portal
Login to SRA	SRA Toolkit Documentation	Trace Archive
Login to Submission Portal	SRA-BLAST	dbGaP Home
SRA Handbook	SRA Run Browser	BioProject

公開データの取得 | データの探し方

研究情報

サンプル情報

ライブラリ情報

NCBI Resources How To Sign in to NCBI

SRA SRA ERR1231585 Search Create alert Advanced Help

Full Send to

ERX1303524: Illumina HiSeq 2500 paired end sequencing
1 ILLUMINA (Illumina HiSeq 2500) run: 2.6M spots, 262.3M bases, 138.5Mb downloads

Submitted by: WEIZMANN INSTITUE OF SCIENCE

Study: Expression Homeostasis during DNA Replication [DNA/ChIP-seq data alpha-factor]
[PRJEB12575](#) • [ERP014063](#) • [All experiments](#) • [All runs](#)
[show Abstract](#)

Sample: Saccharomyces cerevisiae; WT alpha-factor input 39
[SAMEA3860169](#) • [ERS1047303](#) • [All experiments](#) • [All runs](#)
Organism: [Saccharomyces cerevisiae](#)

Library:
Name: unspecified
Instrument: Illumina HiSeq 2500
Strategy: ChIP-Seq
Source: GENOMIC
Selection: ChIP
Layout: PAIRED

Runs: 1 run, 2.6M spots, 262.3M bases, [138.5Mb](#)

Run	# of Spots	# of Bases	Size	Published
ERR1231585	2,571,570	262.3M	138.5Mb	2016-02-10

ID: 2220688

Related information

BioProject
BioSample
Taxonomy

Search details

ERR1231585[All Fields]

Search See more...

Recent activity

Turn Off Clear

- ERR1231585 (1) SRA
- ERP014063 (47) SRA
- Grapevine powdery mildew resistance and suscepti PubMed
- Identification of loci governing

公開データの取得 | データの探し方

NCBI Resources How To Sign in to NCBI

SRA SRA ERR1231585 Search Create alert Advanced Help

All runs を選択
同じ Study でシーケンスした全てのデータを確認

ERX1303524: Illumina HiSeq 2500 paired-end sequencing
1 ILLUMINA (Illumina HiSeq 2500) run: 2,571,570 spots, 262.3M bases, 138.5Mb downloads

Submitted by: WEIZMANN INSTITUE OF SCIENCE

Study: Expression Homeostasis during DNA damage on [DNA/ChIP-seq data alpha-factor]
[PRJEB12575](#) • [ERP014063](#) • [All experiments](#) • [All runs](#)
[show Abstract](#)

Sample: Saccharomyces cerevisiae; WT alpha-factor input 39
[SAMEA3860169](#) • [ERS1047303](#) • [All experiments](#) • [All runs](#)
Organism: [Saccharomyces cerevisiae](#)

Library:
Name: unspecified
Instrument: Illumina HiSeq 2500
Strategy: ChIP-Seq
Source: GENOMIC
Selection: ChIP
Layout: PAIRED

Runs: 1 run, 2.6M spots, 262.3M bases, [138.5Mb](#)

Run	# of Spots	# of Bases	Size	Published
ERR1231585	2,571,570	262.3M	138.5Mb	2016-02-10

ID: 2220688

Related information

BioProject
BioSample
Taxonomy

Search details

ERR1231585[All Fields]

Search See more...

Recent activity

Turn Off Clear

ERR1231585 (1) SRA

ERP014063 (47) SRA

Grapevine powdery mildew resistance and suscepti PubMed

Identification of loci governing

公開データの取得 | データの探し方

SRA Run Selector でデータを検索する

NCBI SRA Run Selector Help Permalink

Search: ERP014063

Facets

- Run
- BioSample
- Sample name
- MBases
- MBytes
- Experiment
- InsertSize
- LibraryLayout
- SRA Sample
- antibody
- time

Hide common fields

Assay Type: ChIP-Seq
BioProject: [PRJEB12575](#)
Center Name: WEIZMANN INSTITUTE OF SCIENCE
Consent: public
LibrarySelection: ChIP
LibrarySource: GENOMIC
Library Name: unspecified
LoadDate: 2016-02-10
Organism: Saccharomyces cerevisiae
Platform: ILLUMINA
ReleaseDate: 2016-02-10
SRA Study: [ERP014063](#)
strain: MATa, his3?1, leu2?0, met15?0, ura3?0
time course repeat: WT (DNA content experiment)

47Runs 合計4.42GBのデータ量

	Runs	Bytes	Bases	Download
Total:	47	4.42 Gb	8.06 G	RunInfo Table Accession List
Selected:				RunInfo Table Accession List

47 Runs found

<input type="checkbox"/>	Run	BioSample	Sample name	MBases	MBytes	Experim
<input type="checkbox"/>	ERR1231588	SAMEA3860172	WT alpha-factor input 30	215	118	ERX1303
<input type="checkbox"/>	ERR1231589	SAMEA3860173	WT alpha-factor input 27	266	146	ERX1303
<input type="checkbox"/>	ERR1231590	SAMEA3860174	WT alpha-factor input 24	349	191	ERX1303
<input type="checkbox"/>	ERR1231591	SAMEA3860175	WT alpha-factor input 21	289	158	ERX1303
<input type="checkbox"/>	ERR1231592	SAMEA3860176	WT alpha-factor input 18	251	137	ERX1303
<input type="checkbox"/>	ERR1231593	SAMEA3860177	WT alpha-factor input 15	239	131	ERX1303

公開データの取得 | データの探し方

ERR1231585 (input) とERR1231597 (sample) を選択

	Runs	Bytes	Bases	Download	
Total:	47	4.42 Gb	8.06 G	RunInfo Table	Accession List
Selected:	2	255.00 Mb	464.00 M	RunInfo Table	Accession List

コントロール (input)

47 Runs found

Run	BioSample	Sample name	MBases	MBytes	Experiment	InsertSize	LibraryLayout	SRA #	antibody	time
<input checked="" type="checkbox"/> ERR1231585	SAMEA3860169	WT alpha-factor input 39	250	138	ERX1303524	250	PAIRED	ERS1047303	input	39
<input type="checkbox"/> ERR1231586	SAMEA3860170	WT alpha-factor input 36	282	153	ERX1303525	250	PAIRED	ERS1047304	input	36
<input type="checkbox"/> ERR1231587	SAMEA3860171	WT alpha-factor input 33	290	159	ERX1303526	250	PAIRED	ERS1047305	input	33
<input type="checkbox"/> ERR1231588	SAMEA3860172	WT alpha-factor input 30	215	118	ERX1303527	250	PAIRED	ERS1047306	input	30
<input type="checkbox"/> ERR1231589	SAMEA3860173	WT alpha-factor input 27	266	146	ERX1303528	250	PAIRED	ERS1047307	input	27
<input type="checkbox"/> ERR1231590	SAMEA3860174	WT alpha-factor input 24	349	191	ERX1303529	250	PAIRED	ERS1047308	input	24
<input type="checkbox"/> ERR1231591	SAMEA3860175	WT alpha-factor input 21	289	158	ERX1303530	250	PAIRED	ERS1047309	input	21
<input type="checkbox"/> ERR1231592	SAMEA3860176	WT alpha-factor input 18	251	137	ERX1303531	250	PAIRED	ERS1047310	input	18
<input type="checkbox"/> ERR1231593	SAMEA3860177	WT alpha-factor input 15	239	131	ERX1303532	250	PAIRED	ERS1047311	input	15
<input type="checkbox"/> ERR1231594	SAMEA3860178	WT alpha-factor input 12	332	182	ERX1303533	250	PAIRED	ERS1047312	input	12
<input type="checkbox"/> ERR1231595	SAMEA3860179	WT alpha-factor input 9	226	125	ERX1303534	250	PAIRED	ERS1047313	input	9
<input type="checkbox"/> ERR1231596	SAMEA3860180	WT alpha-factor input sync	214	118	ERX1303535	250	PAIRED	ERS1047314	input	sync
<input checked="" type="checkbox"/> ERR1231597	SAMEA3860181	WT alpha-factor H3K56ac 39	214	117	ERX1303536	250	PAIRED	ERS1047315	H3K56ac	39
<input type="checkbox"/> ERR1231598	SAMEA3860182	WT alpha-factor H3K56ac 36	233	128	ERX1303537	250	PAIRED	ERS1047316	H3K56ac	36
<input type="checkbox"/> ERR1231599	SAMEA3860183	WT alpha-factor H3K56ac 33	305	165	ERX1303538	250	PAIRED	ERS1047317	H3K56ac	33
<input type="checkbox"/> ERR1231600	SAMEA3860184	WT alpha-factor H3K56ac 30	288	159	ERX1303539	250	PAIRED	ERS1047318	H3K56ac	30
<input type="checkbox"/> ERR1231601	SAMEA3860185	WT alpha-factor H3K56ac 27	110	55	ERX1303540	250	PAIRED	ERS1047319	H3K56ac	27

サンプル (H3K56ac)

ダウンロードするデータにチェック

公開データの取得 | データの探し方

Accession List をダウンロード

	Runs	Bytes	Bases	Download	
Total:	47	4.42 Gb	8.06 G	RunInfo Table	Accession List
Selected:	2	255.00 Mb	464.00 M	RunInfo Table	Accession List



47 Runs found

Run	BioSample
<input checked="" type="checkbox"/> ERR1231585	SAMEA3860169
<input type="checkbox"/> ERR1231586	SAMEA3860170
<input type="checkbox"/> ERR1231587	SAMEA3860171
<input type="checkbox"/> ERR1231588	SAMEA3860172
<input type="checkbox"/> ERR1231589	SAMEA3860173
<input type="checkbox"/> ERR1231590	SAMEA3860174
<input type="checkbox"/> ERR1231591	SAMEA3860175
<input type="checkbox"/> ERR1231592	SAMEA3860176
<input type="checkbox"/> ERR1231593	SAMEA3860177
<input type="checkbox"/> ERR1231594	SAMEA3860178
<input type="checkbox"/> ERR1231595	SAMEA3860179
<input type="checkbox"/> ERR1231596	SAMEA3860180
<input checked="" type="checkbox"/> ERR1231597	SAMEA3860181
<input type="checkbox"/> ERR1231598	SAMEA3860182
<input type="checkbox"/> ERR1231599	SAMEA3860183
<input type="checkbox"/> ERR1231600	SAMEA3860184
<input type="checkbox"/> ERR1231601	SAMEA3860185

```
SRR_Acc_List.txt
ERR1231585
ERR1231597
```

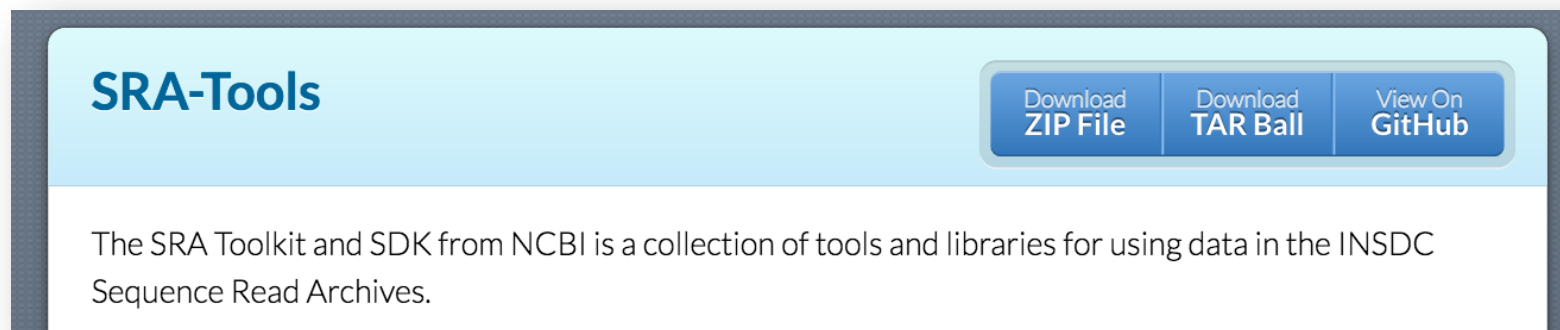
SRA Sample	antibody	time
ERS1047303	input	39
ERS1047304	input	36
ERS1047305	input	33
ERS1047306	input	30
ERS1047307	input	27
ERS1047308	input	24
ERS1047309	input	21
ERS1047310	input	18
ERS1047311	input	15
ERS1047312	input	12
ERS1047313	input	9
ERS1047314	input	sync
ERS1047315	H3K56ac	39
ERS1047316	H3K56ac	36
ERS1047317	H3K56ac	33
ERS1047318	H3K56ac	30
ERS1047319	H3K56ac	27

アクセッション番号のリスト (テキストファイル) がダウンロードされる

公開データの取得 | ダウンロード方法

SRA のダウンロードには、SRA-Toolsを使用する

- SRA-Tools (<http://ncbi.github.io/sra-tools/>)



- 主な用途【実行コマンド】
 - NCBI SRA からのデータダウンロード【**prefetch**】
 - SRA→FASTQのフォーマット変換【**fastq-dump**】

公開データの取得 | ダウンロード方法

- SRA-Toolsのインストール
 - 本日はデータを用意済みのため実施しません

```
$ wget http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.6.3/sratoolkit.2.6.3-centos\_linux64.tar.gz
$ tar xf sratoolkit.2.6.3-centos\_linux64.tar.gz
$ ln -s sratoolkit.2.6.3-centos_linux64/bin/prefetch ¥
  /usr/local/bin/
$ ln -s sratoolkit.2.6.3-centos_linux64/bin/fastq-dump ¥
  /usr/local/bin/
```

- 参考 : http://ncbi.github.io/sra-tools/install_config.html

公開データの取得 | ダウンロード方法

SRA-Tools の prefetch コマンドでまとめてSRAをダウンロード

ダウンロードしたAccession List (SRR_Acc_List.txt) を
--option-file で指定

```
$ prefetch --option-file SRR_Acc_List.txt
```

デフォルトでSRAは ~/ncbi/public/sra/ に保存される

```
$ ls ~/ncbi/public/sra/  
ERR1231585.sra  ERR1231597.sra
```

公開データの取得 | SRAの変換方法

SRA-Tools の fastq-dump を使用して SRA から FASTQ へ変換する

変換データを保存するディレクトリ(data)を作成する(実行済み)

```
$ mkdir data  
$ cd data
```

--split-files を付けてペアエンドのファイルを分割しながらFASTQに変換する
(実行済み)

どこでペアエンドかシングルエンドかを確認するのか
(次のスライドで解説)

```
$ fastq-dump ~/ncbi/public/sra/ERR1231585.sra --split-files  
$ fastq-dump ~/ncbi/public/sra/ERR1231597.sra --split-files
```

公開データの取得 | SRAの変換方法

Run selector の LibraryLayout でペアエンドであることを確認できる

	Runs	Bytes	Bases	Download							
Total:	本日使用するデータは全てペアエンド (PAIRED)										
Selected:											
47 Runs found											
	Run	BioSample	Sample name	MBases	MBytes	Experiment	Inse	.libraryLayout	SRA Sample	antibody	time
<input checked="" type="checkbox"/>	ERR1231585	SAMEA3860169	WT alpha-factor input 39	250	138	ERX1303524	250	PAIRED	ERS1047303	input	39
<input type="checkbox"/>	ERR1231586	SAMEA3860170	WT alpha-factor input 36	282	153	ERX1303525	250	PAIRED	ERS1047304	input	36
<input type="checkbox"/>	ERR1231587	SAMEA3860171	WT alpha-factor input 33	290	159	ERX1303526	250	PAIRED	ERS1047305	input	33
<input type="checkbox"/>	ERR1231588	SAMEA3860172	WT alpha-factor input 30	215	118	ERX1303527	250	PAIRED	ERS1047306	input	30
<input type="checkbox"/>	ERR1231589	SAMEA3860173	WT alpha-factor input 27	266	146	ERX1303528	250	PAIRED	ERS1047307	input	27
<input type="checkbox"/>	ERR1231590	SAMEA3860174	WT alpha-factor input 24	349	191	ERX1303529	250	PAIRED	ERS1047308	input	24
<input type="checkbox"/>	ERR1231591	SAMEA3860175	WT alpha-factor input 21	289	158	ERX1303530	250	PAIRED	ERS1047309	input	21
<input type="checkbox"/>	ERR1231592	SAMEA3860176	WT alpha-factor input 18	251	137	ERX1303531	250	PAIRED	ERS1047310	input	18
<input type="checkbox"/>	ERR1231593	SAMEA3860177	WT alpha-factor input 15	239	131	ERX1303532	250	PAIRED	ERS1047311	input	15
<input type="checkbox"/>	ERR1231594	SAMEA3860178	WT alpha-factor input 12	332	182	ERX1303533	250	PAIRED	ERS1047312	input	12
<input type="checkbox"/>	ERR1231595	SAMEA3860179	WT alpha-factor input 9	226	125	ERX1303534	250	PAIRED	ERS1047313	input	9
<input type="checkbox"/>	ERR1231596	SAMEA3860180	WT alpha-factor input sync	214	118	ERX1303535	250	PAIRED	ERS1047314	input	sync
<input checked="" type="checkbox"/>	ERR1231597	SAMEA3860181	WT alpha-factor H3K56ac 39	214	117	ERX1303536	250	PAIRED	ERS1047315	H3K56ac	39
<input type="checkbox"/>	ERR1231598	SAMEA3860182	WT alpha-factor H3K56ac 36	233	128	ERX1303537	250	PAIRED	ERS1047316	H3K56ac	36
<input type="checkbox"/>	ERR1231599	SAMEA3860183	WT alpha-factor H3K56ac 33	305	165	ERX1303538	250	PAIRED	ERS1047317	H3K56ac	33
<input type="checkbox"/>	ERR1231600	SAMEA3860184	WT alpha-factor H3K56ac 30	288	159	ERX1303539	250	PAIRED	ERS1047318	H3K56ac	30
<input type="checkbox"/>	ERR1231601	SAMEA3860185	WT alpha-factor H3K56ac 27	200	110	ERX1303540	250	PAIRED	ERS1047319	H3K56ac	27

公開データの取得 | SRAの変換方法

SRA-Tools の fastq-dump を使用して SRA から FASTQ へ変換する

変換したFASTQを確認する

```
$ ls
```

```
ERR1231585_1.fastq  ERR1231597_1.fastq  
ERR1231585_2.fastq  ERR1231597_2.fastq
```

公開データの取得 | 実習用データの作成

seqtk (<https://github.com/lh3/seqtk>) を使用し、
実習用にFASTQからデータの一部を抜粋する

seqtk のインストール (今回は実施しません)

```
$ wget https://github.com/lh3/seqtk/archive/v1.2.tar.gz
$ tar xf v1.2.tar.gz
$ cd seqtk-1.2
$ ln -s ~/src/seqtk-1.2/seqtk /usr/local/bin/
```

公開データの取得 | 実習用データの作成

seqtk を使用し、実習用にFASTQからデータの一部を抜粋する

seqtk の実行

```
$ seqtk sample -s 100 ERR1231585_1.fastq 500000 > input_1.fastq
$ seqtk sample -s 100 ERR1231585_2.fastq 500000 > input_2.fastq
$ seqtk sample -s 100 ERR1231597_1.fastq 500000 > sample_1.fastq
$ seqtk sample -s 100 ERR1231597_2.fastq 500000 > sample_2.fastq
```

-s 100: シード値を100に指定

ペアで同じシード値を使うことで、

ランダムに抽出するリードのペアを保つ事ができる

500000 : 50万リード抽出

実習パート

公開データの取得

解析対象のシーケンスデータの取得方法 (実行済み)

ダウンロード、SRA→FASTQ変換

```
$ prefetch --option-file SRR_Acc_List.txt  
$ fastq-dump ~/ncbi/public/sra/ERR12315*.sra --split-files
```

実習用の軽量なデータを作成 (実行済み)

```
$ seqtk sample -s 100 ERR1231585_1.fastq 500000 > input_1.fastq  
$ seqtk sample -s 100 ERR1231585_2.fastq 500000 > input_2.fastq  
$ seqtk sample -s 100 ERR1231597_1.fastq 500000 > sample_1.fastq  
$ seqtk sample -s 100 ERR1231597_2.fastq 500000 > sample_2.fastq
```

公開データの取得

解析対象のシーケンスデータの確認

```
$ cd /home/iu/chipseq
```

```
$ ls data
```

```
input_1.fastq.gz
```

```
input_2.fastq.gz
```

```
sample_1.fastq.gz
```

```
sample_2.fastq.gz
```

アクセス番号との対応

input_1 → ERR1231585_1.fastq.gz

input_2 → ERR1231585_2.fastq.gz

sample_1 → ERR1231597_1.fastq.gz

sample_2 → ERR1231597_2.fastq.gz

それぞれ500,000リードのデータ

クオリティコントロール | QC前の品質確認

シーケンスクオリティチェックソフトウェア FastQC の実行

```
$ mkdir fastqc_before
$ fastqc --nogroup -t 2 -o ./fastqc_before ¥
  data/input_1.fastq.gz ¥
  data/input_2.fastq.gz ¥
  data/sample_1.fastq.gz ¥
  data/sample_2.fastq.gz
$ ls fastqc_before
```

```
input_1_fastqc      input_2_fastqc.zip  sample_2_fastqc
input_1_fastqc.zip  sample_1_fastqc     sample_2_fastqc.zip
input_2_fastqc      sample_1_fastqc.zip
```

クオリティコントロール | QC前の品質確認

FastQCの結果確認 (QC前)

解析結果のhtmlファイルをブラウザ (firefox)で確認

```
$ firefox ¥  
  fastqc_before/input_1_fastqc/fastqc_report.html ¥  
  fastqc_before/input_2_fastqc/fastqc_report.html ¥  
  fastqc_before/sample_1_fastqc/fastqc_report.html ¥  
  fastqc_before/sample_2_fastqc/fastqc_report.html
```

ブラウザでタブが4つ開かれ、
クオリティチェックの解析結果が確認できる

クオリティコントロール | QC前の品質確認

fastqc summary (QC前)

input1	input2	sample1	sample2
Summary	Summary	Summary	Summary
Basic Statistics	Basic Statistics	Basic Statistics	Basic Statistics
Per base sequence quality	Per base sequence quality	Per base sequence quality	Per base sequence quality
Per sequence quality scores	Per sequence quality scores	Per sequence quality scores	Per sequence quality scores
Per base sequence content	Per base sequence content	Per base sequence content	Per base sequence content
Per base GC content	Per base GC content	Per base GC content	Per base GC content
Per sequence GC content	Per sequence GC content	Per sequence GC content	Per sequence GC content
Per base N content	Per base N content	Per base N content	Per base N content
Sequence Length Distribution	Sequence Length Distribution	Sequence Length Distribution	Sequence Length Distribution
Sequence Duplication Levels	Sequence Duplication Levels	Sequence Duplication Levels	Sequence Duplication Levels
Overrepresented sequences	Overrepresented sequences	Overrepresented sequences	Overrepresented sequences
Kmer Content	Kmer Content	Kmer Content	Kmer Content

クオリティコントロール | QC処理

今回のデータに対する処理 (Trimmomaticを用いた一括処理①)

```
$ mkdir trimmed_data
$ java -jar /usr/local/bin/trimmomatic-0.36.jar PE ¥
  -threads 2 -phred33 ¥
  data/input_1.fastq.gz ¥
  data/input_2.fastq.gz ¥
  trimmed_data/input_1_paired.fastq ¥
  trimmed_data/input_1_unpaired.fastq ¥
  trimmed_data/input_2_paired.fastq ¥
  trimmed_data/input_2_unpaired.fastq ¥
  LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:36
```

「sample～」でも同様の処理を実行

クオリティコントロール | QC処理

今回のデータに対する処理 (Trimmomaticを用いた一括処理②)

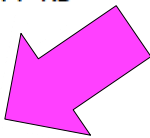
```
$ java -jar /usr/local/bin/trimmomatic-0.36.jar PE ¥  
-threads 2 -phred33 ¥  
data/sample_1.fastq.gz ¥  
data/sample_2.fastq.gz ¥  
trimmed_data/sample_1_paired.fastq ¥  
trimmed_data/sample_1_unpaired.fastq ¥  
trimmed_data/sample_2_paired.fastq ¥  
trimmed_data/sample_2_unpaired.fastq ¥  
LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:36
```

CPUのコア数に余裕があれば `-threads` の数値を大きくすることでより高速に処理することが可能

TIPS | CPUコア数を確認してみる

```
$ cat /proc/cpuinfo
```

```
processor      : 1
vendor_id     : GenuineIntel
cpu family    : 6
model         : 70
model name    : Intel(R) Core(TM) i7-4980HQ CPU @ 2.80GHz
stepping      : 1
cpu MHz       : 2793.532
cache size    : 6144 KB
physical id   : 0
siblings      : 2
core id       : 1
cpu cores     : 2
apicid        : 1
initial apicid : 1
fpu           : yes
fpu_exception : yes
cpuid level   : 13
wp            : yes
flags         : fpu vme de pse tsc msr pae mce cx8 apic sep
ll nx rdtscp lm constant tsc rep good nopl xtopology nonstop t
```



cpu cores を確認
ここでは 2 となっている

全てのコアを使用して計算しようとする

かえって遅くなる時もあるので、コマンド実行時は様子を見ながら増やす

クオリティコントロール | QC後の品質確認

FastQCの結果確認 (QC後)

```
$ mkdir fastqc_after
$ fastqc --nogroup -t 2 -o fastqc_after ¥
  trimmed_data/input_1_paired.fastq ¥
  trimmed_data/input_2_paired.fastq ¥
  trimmed_data/sample_1_paired.fastq ¥
  trimmed_data/sample_2_paired.fastq
$ firefox ¥
  fastqc_after/input_1_paired_fastqc/fastqc_report.html ¥
  fastqc_after/input_2_paired_fastqc/fastqc_report.html ¥
  fastqc_after/sample_1_paired_fastqc/fastqc_report.html ¥
  fastqc_after/sample_2_paired_fastqc/fastqc_report.html
```

クオリティコントロール | QC前の品質確認

FastQC summary (QC後)

input1

Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per base GC content](#)
- ! [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ! [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Kmer Content](#)

input2

Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per base GC content](#)
- ! [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ! [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Kmer Content](#)

sample1

Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per base GC content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Kmer Content](#)

sample2

Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per base GC content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Kmer Content](#)

クオリティコントロール | QC前の品質確認

QC前

input_1

Summary

- ✔ Basic Statistics
- ✔ Per base sequence quality
- ✔ Per sequence quality scores
- ! Per base sequence content
- ✔ Per base GC content
- ✘ Per sequence GC content
- ✔ Per base N content
- ✔ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✘ Overrepresented sequences
- ✘ Kmer Content

input2

Summary

- ✔ Basic Statistics
- ✔ Per base sequence quality
- ✔ Per sequence quality scores
- ! Per base sequence content
- ✔ Per base GC content
- ✘ Per sequence GC content
- ✔ Per base N content
- ✔ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✔ Overrepresented sequences
- ✘ Kmer Content

sample1

Summary

- ✔ Basic Statistics
- ✔ Per base sequence quality
- ✔ Per sequence quality scores
- ✘ Per base sequence content
- ! Per base GC content
- ✘ Per sequence GC content
- ✔ Per base N content
- ✔ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✘ Overrepresented sequences
- ✘ Kmer Content

sample2

Summary

- ✔ Basic Statistics
- ✔ Per base sequence quality
- ✔ Per sequence quality scores
- ✘ Per base sequence content
- ! Per base GC content
- ✘ Per sequence GC content
- ✔ Per base N content
- ✔ Sequence Length Distribution
- ✔ Sequence Duplication Levels
- ✔ Overrepresented sequences
- ✘ Kmer Content

QC後

Summary

- ✔ Basic Statistics
- ✔ Per base sequence quality
- ✔ Per sequence quality scores
- ! Per base sequence content
- ✔ Per base GC content
- ! Per sequence GC content
- ✔ Per base N content
- ! Sequence Length Distribution
- ! Sequence Duplication Levels
- ✔ Overrepresented sequences
- ✔ Kmer Content

Summary

- ✔ Basic Statistics
- ✔ Per base sequence quality
- ✔ Per sequence quality scores
- ! Per base sequence content
- ✔ Per base GC content
- ! Per sequence GC content
- ✔ Per base N content
- ! Sequence Length Distribution
- ! Sequence Duplication Levels
- ✔ Overrepresented sequences
- ✔ Kmer Content

Summary

- ✔ Basic Statistics
- ✔ Per base sequence quality
- ✔ Per sequence quality scores
- ! Per base sequence content
- ✔ Per base GC content
- ✔ Per sequence GC content
- ✔ Per base N content
- ! Sequence Length Distribution
- ✔ Sequence Duplication Levels
- ✔ Overrepresented sequences
- ✔ Kmer Content

Summary

- ✔ Basic Statistics
- ✔ Per base sequence quality
- ✔ Per sequence quality scores
- ! Per base sequence content
- ✔ Per base GC content
- ✔ Per sequence GC content
- ✔ Per base N content
- ! Sequence Length Distribution
- ✔ Sequence Duplication Levels
- ✔ Overrepresented sequences
- ✔ Kmer Content

マッピング

Bowtie2によるマッピング (inputファイル)

```
$ mkdir mapping
$ bowtie2 -p 2 -x /home/iu/genome/sacCer3/Bowtie2Index/genome ¥
-1 trimmed_data/input_1_paired.fastq ¥
-2 trimmed_data/input_2_paired.fastq | ¥
samtools view -Sb - > mapping/input.bam
$ samtools sort mapping/input.bam -o mapping/input.sorted.bam
```

■ bowtie2のオプション

- p : 使用するスレッド数
- x : bowtie2で作成したゲノムファイルインデックス
- 1,-2: 入力fastqファイル

■ Samtoolsのオプション

- view: SAMもしくははBAMの中身を表示
- Sb: SAMからBAMへ変換

マッピング

Bowtie2によるマッピング (sampleファイル)

```
$ bowtie2 -p 2 -x /home/iu/genome/sacCer3/Bowtie2Index/genome ¥  
-1 trimmed_data/sample_1_paired.fastq ¥  
-2 trimmed_data/sample_2_paired.fastq | ¥  
samtools view -Sb - > mapping/sample.bam  
$ samtools sort mapping/sample.bam -o mapping/sample.sorted.bam
```

ピーク検出

MACS2によるピーク検出

```
$ macs2 callpeak ¥  
  -t mapping/sample.sorted.bam ¥  
  -c mapping/input.sorted.bam ¥  
  --outdir macs2_res ¥  
  -f BAMPE -n handson2016 -B -q 0.01 -g 1.2e+7
```

- t ターゲットサンプル (IP) のファイル
- c -tに対するコントロール (input) サンプルのファイル
- outdir 結果を出力するディレクトリ
- f -tで指定したファイルのファイル形式
BAM、SAM、BED他様々なフォーマットが指定可能
BAMPEはpaired-end readをマッピングしたbamファイル

(コマンドの説明は次スライドに続きます→)

ピーク検出

MACS2によるピーク検出

```
$ macs2 callpeak ¥  
  -t mapping/sample.sorted.bam ¥  
  -c mapping/input.sorted.bam ¥  
  --outdir macs2_res ¥  
  -f BAMPE -n handson2016 -B -q 0.01 -g 1.2e+7
```

- n 出力ファイルの接頭文字
- B フラグメントのpileup、control lambda値などをBedGraph形式で保存
- q peakcallするピークの閾値（Benjamini-HochbergによるFDRのq値）
【デフォルト 0.01】
- g 反復領域を除いたゲノムサイズ
一部のモデル生物では数字ではなく、ヒト:hs、マウス:mmなどの省略が可能

ピーク検出

MACS2によるピーク検出

```
$ ls macs2_res  
handson2016_control_lambda.bdg  handson2016_summits.bed  
handson2016_peaks.narrowPeak    handson2016_treat_pileup.bdg  
handson2016_peaks.xls
```

各出力ファイルの解説は、NGS Surfer's Wikiが参考になる

<https://cell-innovation.nig.ac.jp/wiki/tiki-index.php?page=MACS>

この後のモチーフ探索には、ピークの領域情報が記載された

handson2016_peaks.narrowPeak を用いる

ピーク検出

先頭の5行を確認

```
$ head -5 handson2016_peaks.narrowPeak
```

chrI	114052	114468	handson2016_peak_1	46	.	3.84001	8.16057
	4.66642	252					
chrII	35630	36056	handson2016_peak_2	64	.	4.27147	10.05951
	198						6.44232
chrIV	427318	427670	handson2016_peak_3	560	.	4.41420	61.01538
	56.00628	186					
chrIV	769592	769918	handson2016_peak_4	29	.	3.31637	6.20275
	2.95610	157					
chrIV	991149	991514	handson2016_peak_5	40	.	2.81939	7.45001
	4.05226	235					

ピーク検出

handson2016_peaks.narrowPeakの項目解説

列	例
1 : 染色体番号	chrI
2 : ピーク開始位置	114052
3 : ピーク終了位置	114468
4 : ピークの名前	handson2016_peak_1
5 : ピークのスコア	46
6 : スtrand	.
7 : fold-change	3.84001
8 : -log10pvalue	8.16057
9 : -log10qvalue	4.66642
10 : ピーク開始位置から頂点までの距離	252

可視化 | IGVでピークを確認する

- 検出したピークをIGVで可視化する

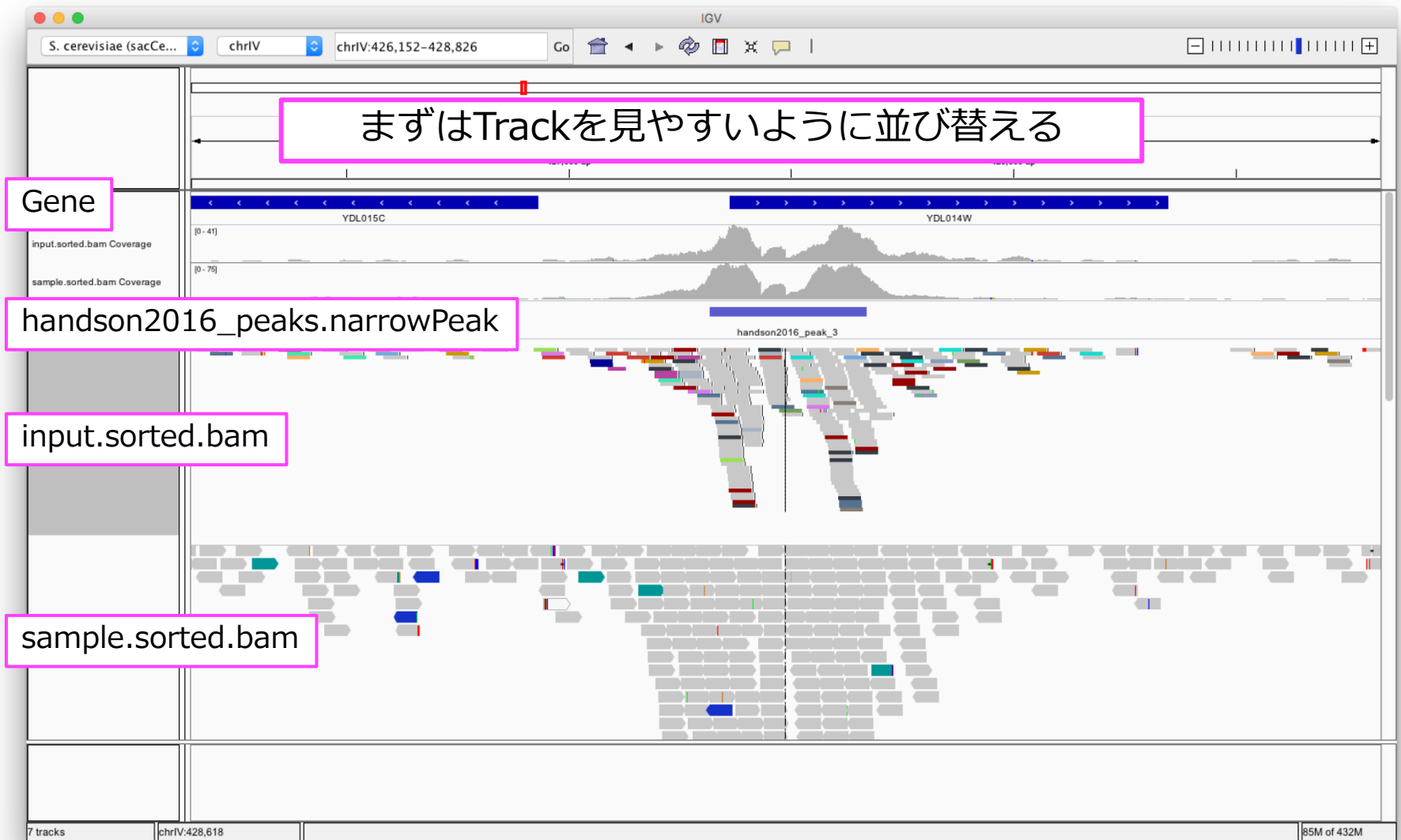
BAMファイルのインデックスを作成

```
$ cd mapping  
$ samtools index input.sorted.bam  
$ samtools index sample.sorted.bam
```

IGVで下記のファイルを表示

1. handson2016_peaks.narrowPeak
2. input.sorted.bam
3. sample.sorted.bam

可視化 | IGVでピークを確認する

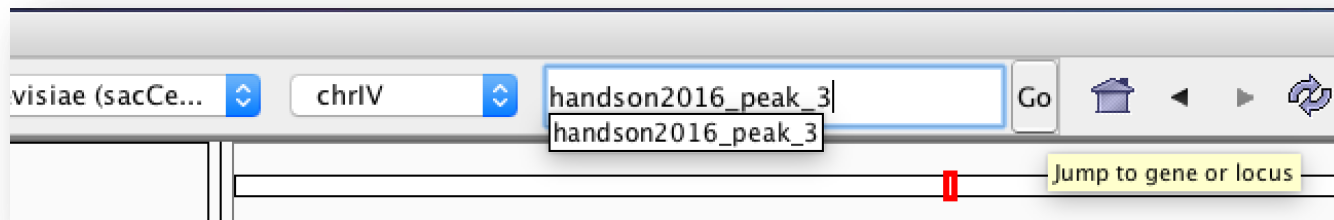


可視化 | IGVでピークを確認する

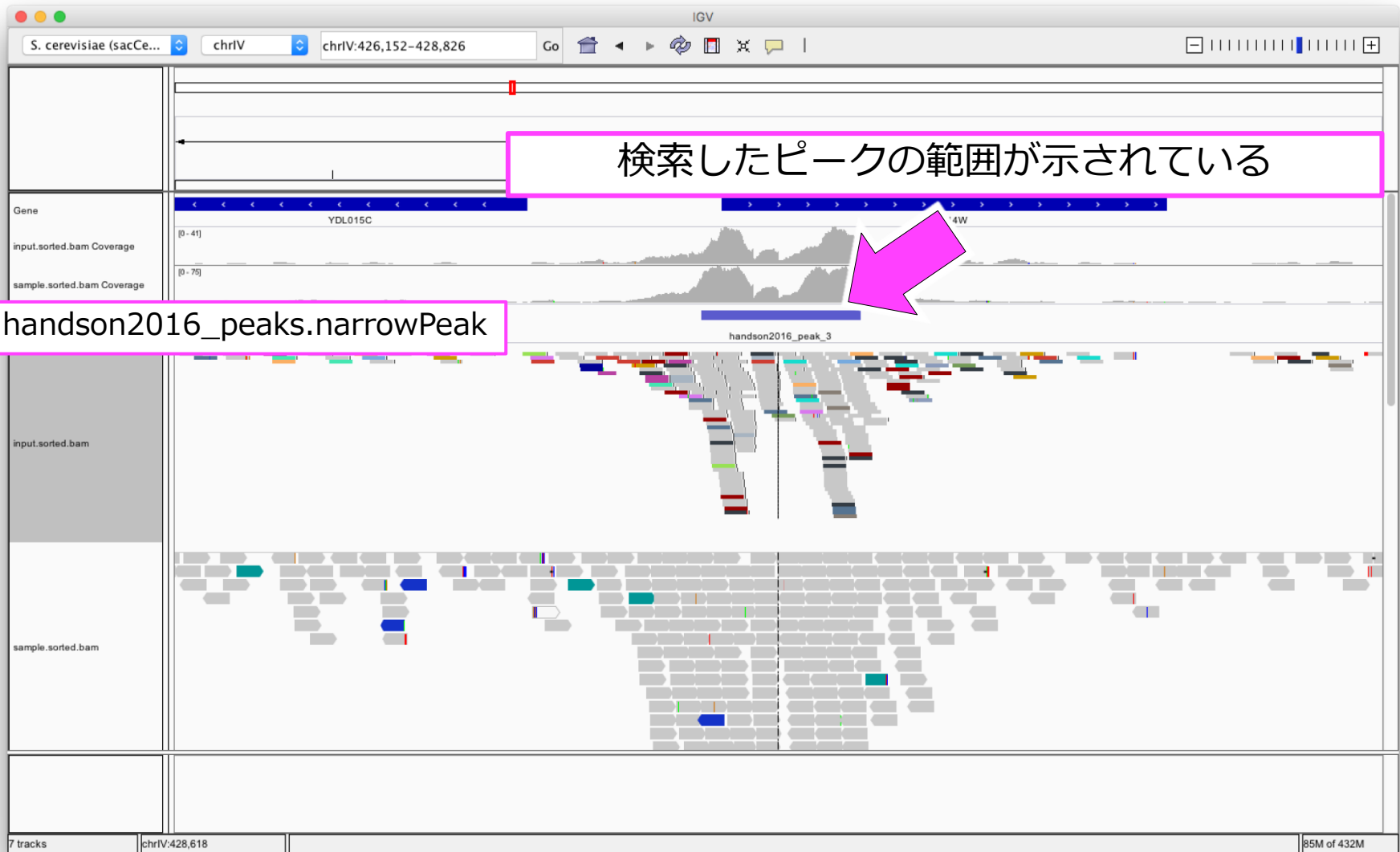
スコアの高いピークを確認

```
$ cd macs2_res  
$ cat handson2016_summits.bed | sort -k 5n
```

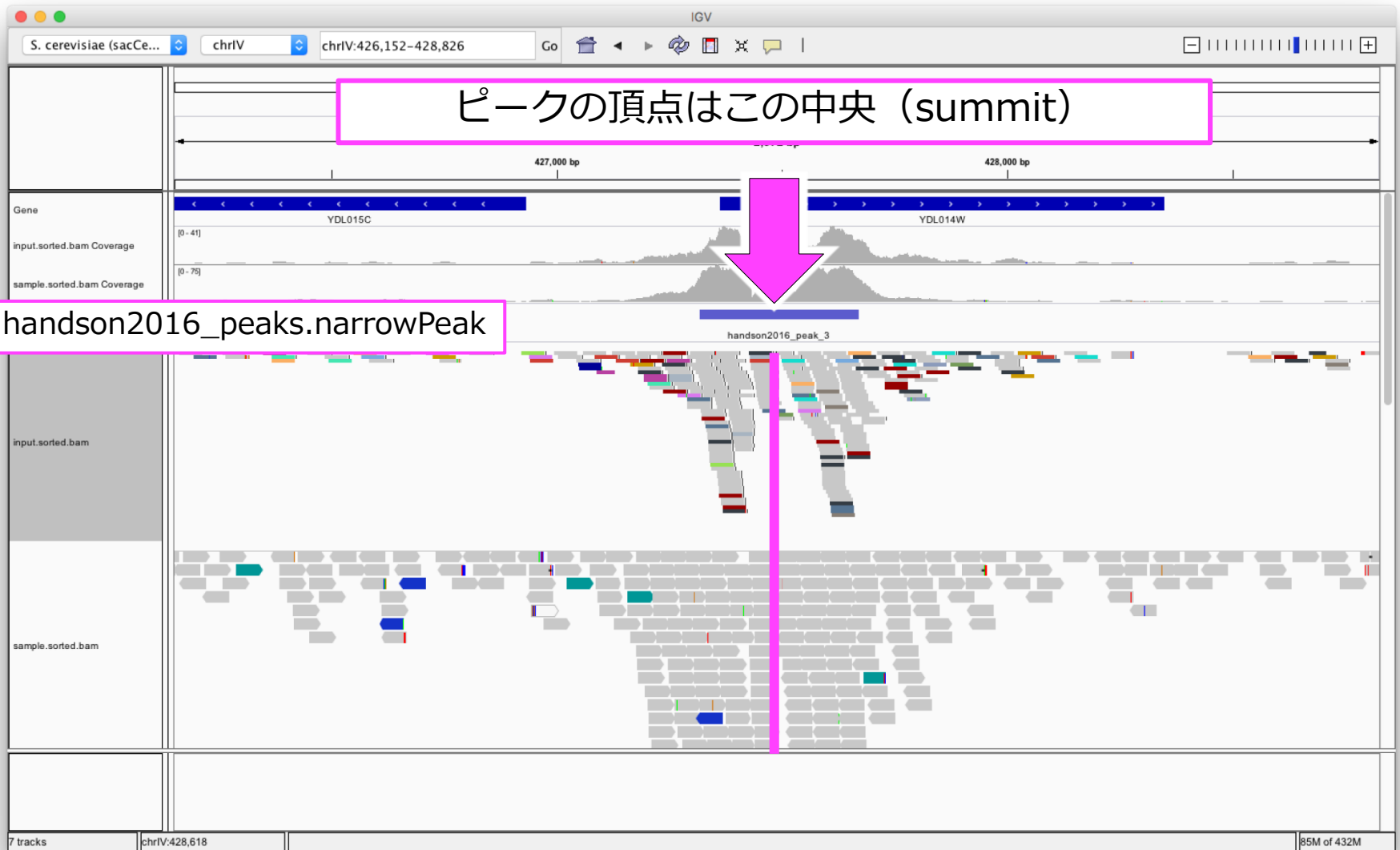
もっとも高いスコアを示したピーク名をIGVの検索窓に入れ検索



可視化 | IGVでピークを確認する



可視化 | IGVでピークを確認する



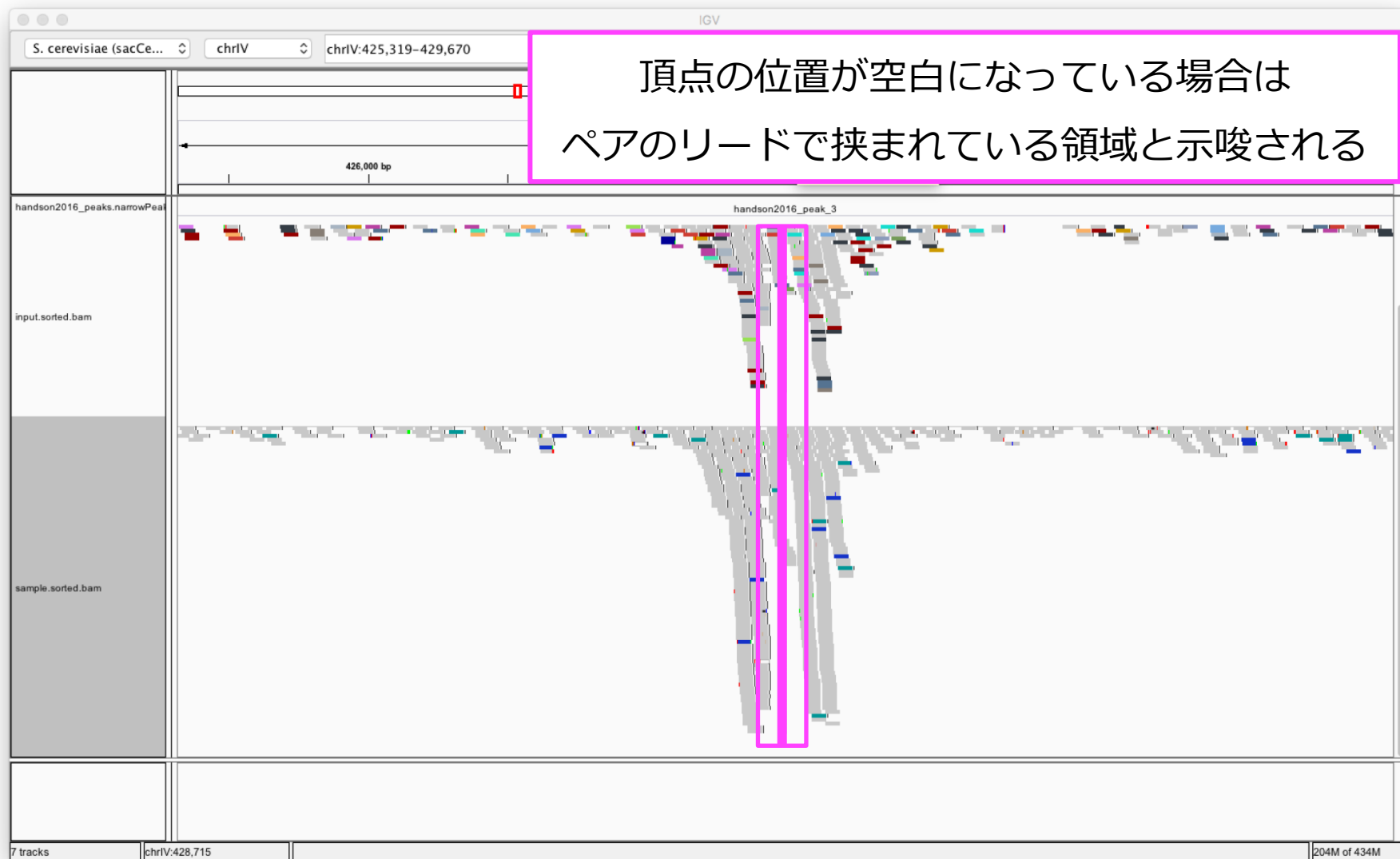
可視化 | IGVでピークを確認する

The screenshot shows the IGV interface with a context menu open over the 'sample.sorted.bam' track. The menu options are:

- sample.sorted.bam
- Rename Track...
- Copy read details to clipboard
- Group alignments by
- Sort alignments by
- Color alignments by
- Re-pack alignments
- ✓ Shade base by quality
- ✓ Show mismatched bases
- Show all bases
- View as pairs
- Go to mate
- View mate region in split screen
- Set insert size options ...
- Collapsed
- Expanded
- ✓ Squished
- Select by name...
- Clear selections
- Copy read sequence
- Blat read sequence
- Copy consensus sequence
- Sashimi Plot
- ✓ Show Coverage Track
- Show Splice Junction Track
- Hide Track
- Save image...
- Export Alignments...

A pink box highlights the track name 'sample.sorted.bam' and the context menu. A text box with a pink border contains the instruction: **それぞれのトラックの名前を右クリックし Squishedを選択**

可視化 | IGVでピークを確認する



可視化 | IGVでピークを確認する

The screenshot shows the IGV interface with a track titled 'sample.sorted.bam'. A context menu is open over the track, listing various actions. The 'Expanded' option is selected. A pink box highlights the text: '今度は拡大してペアのリードを確認 Expandedを選択'. The main view shows a genomic track with a peak labeled 'handson2016_peak_3' at 4,346 bp. The track is currently in a collapsed state, but the 'Expanded' option is selected in the menu.

sample.sorted.bam

- Rename Track...
- Copy read details to clipboard
- Group alignments by
- Sort alignments by
- Color alignments by
- Re-pack alignments
- Shade base by quality
- Show mismatched bases
- Show all bases
- View as pairs
- Go to mate
- View mate region in split scr
- Set insert size options ...
- Collapsed
- Expanded
- Squished
- Select by name...
- Clear selections
- Copy read sequence
- Blat read sequence
- Copy consensus sequence
- Sashimi Plot
- Show Coverage Track
- Show Splice Junction Track
- Hide Track
- Save image...
- Export Alignments...
- Export track names...
- Remove Track

handson2016_peak_3

4,346 bp

427,000 bp 429,000 bp

Click and drag to zoom in.

今度は拡大してペアのリードを確認
Expandedを選択

7 tracks chrIV:428,715 204M of 434M

可視化 | IGVでピークを確認する



アノテーション

handson2016_summits.bedに対してsnpeffによるアノテーションを実施

アノテーション作業用のディレクトリを作成し

アノテーション前のファイルを確認

```
$ mkdir annotation  
$ cd annotation  
$ cat ../macs2_res/handson2016_summits.bed
```

アノテーション

handson2016_summits.bedに対してsnpEffによるアノテーションを実施する

```
$ java -jar /usr/local/bin/snpEff.jar eff ¥  
-csvStats stats.txt -c /usr/local/bin/snpEff.config ¥  
-i bed -o bedAnn R64-1-1.82 ¥  
../macs2_res/handson2016_summits.bed > ¥  
handson2016_summits.annotated.bed
```

eff	入力ファイルにアノテーションを行う
-csvStats	csv形式のサマリーファイルを作成する
-c	snpEffの設定ファイルを指定
-i	入力ファイルのフォーマット
-o	出力ファイルのフォーマット

(コマンドの説明は次スライドに続きます→)

アノテーション

handson2016_summits.bedに対してsnpeffによるアノテーションを実施する

```
$ mkdir annotation
$ cd annotation
$ java -jar /usr/local/bin/snpeff.jar eff ¥
  -csvStats stats.txt -c /usr/local/bin/snpeff.config ¥
  -i bed -o bedAnn R64-1-1.82 ¥
  ../macs2_res/handson2016_summits.bed > ¥
  handson2016_summits.annotated.bed
```

R64-1-1.82

アノテーションに使用するゲノムバージョン

../macs2_res/handson2016_
summits.bed

入力ファイル

アノテーション

snpEffを用いたアノテーション方法

```
$ less handson2016_summits.annotated.bed
```

```
          :  
# Chromo   Start   End   Variant;Annotation   Score  
   I       113613  114615  I:114304;EXON:ATS1  
   I       114249  114819  I:114304;GENE:YAL019W-A  
   I       109918  114918  I:114304;UPSTREAM:FUN30  
   I       113563  118563  I:114304;DOWNSTREAM:LDS1  
          :
```

検出されたピークのsummitについて、遺伝子名とその遺伝子に対してエクソン・上流・下流などの情報が付与される

モチーフ検索

R Bioconductor package 'rGADEM' を用いた *de novo* モチーフ検索①

```
$ mkdir ../motif
```

```
$ cd ../motif
```

```
$ R
```

```
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.
```

```
:
```


モチーフ検索

R Bioconductor package 'rGADEM' を用いた *de novo* モチーフ検索②

```
> library(rGADEM)
> library("BSgenome.Scerevisiae.UCSC.sacCer3")
> BED <- read.table("../macs2_res/handson2016_peaks.narrowPeak",
  header=FALSE, sep="¥t")
> BED <-
  data.frame(chr=as.factor(BED[,1]),
    start=as.numeric(BED[,2]), end=as.numeric(BED[,3]))
```

MACS2から出力されたBEDファイルを、データフレームとして読み込む

再び、 **handson2016_peaks.narrowPeak** を使用

モチーフ検索

R Bioconductor package 'rGADEM' を用いた*de novo* モチーフ検索③

```
> rgBED <- IRanges(start = BED[, 2], end = BED[, 3])
> Sequences <- RangedData(rgBED, space = BED[, 1])
> gadem <- GADEM(Sequences, verbose = 1, genome = Scerevisiae)
> pdf("motif.pdf")
> plot(gadem)
> dev.off()
> q()
```

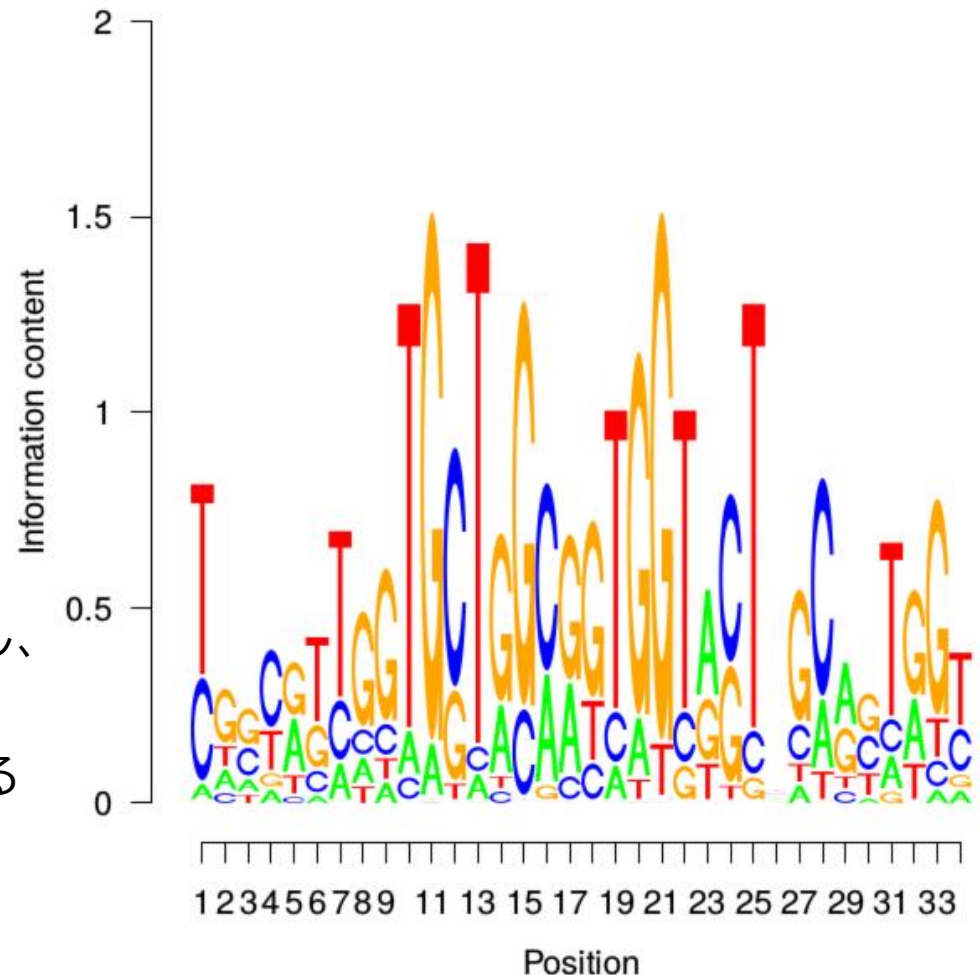
ピーク領域に頻出するモチーフを取得し、PDFにプロット

モチーフ検索

出力したモチーフを確認

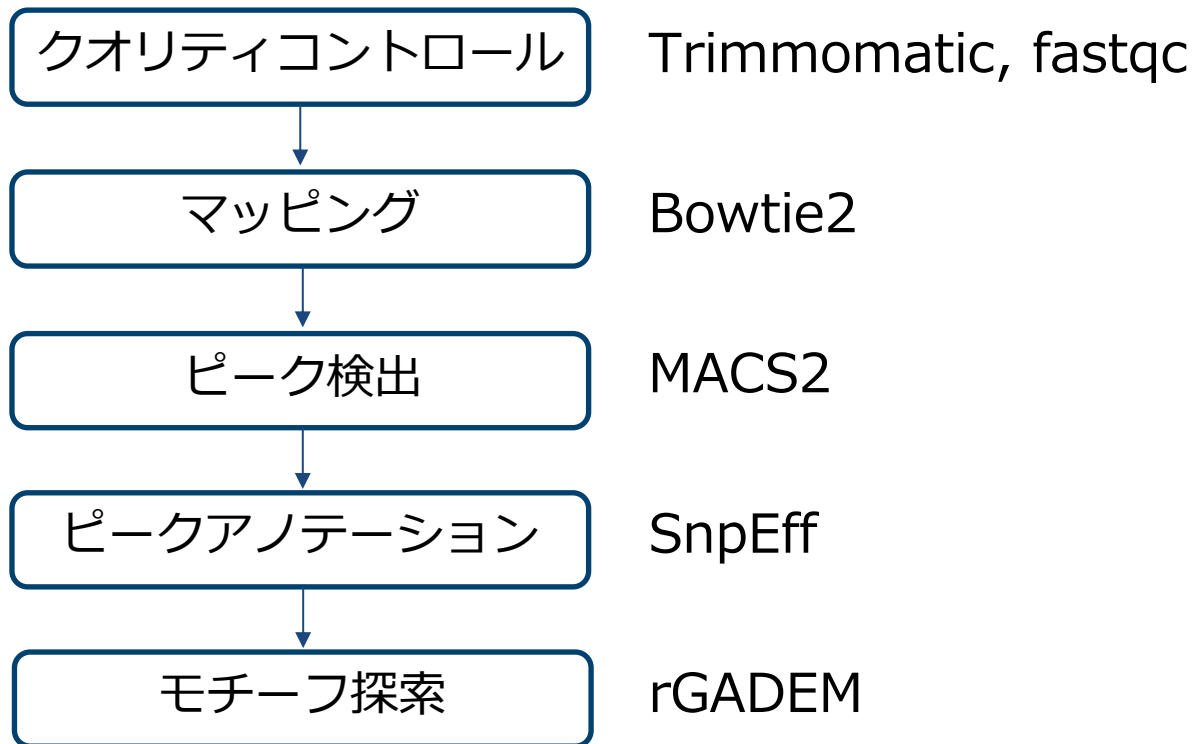
```
$ evince motif.pdf
```

この後さらに、MotIVなどを使用し、
検出したDNAモチーフが既知の
モチーフに似ているかどうか調べる
ことも可能



MotIV: <https://www.bioconductor.org/packages/release/bioc/html/MotIV.html>

まとめ | ChIP-seq解析の流れ



- ChIP-seq解析の一般的な流れであり、全てのChIP-seqで同一の解析を行うわけではない
- 研究の目的やデータに合わせて、最適な解析を設計