ChIP-seqの代表的なパイプラインに関する実習 (初級)

Shinpei Kawaoka (Charlie)
Group leader, Contextual Biology Group,
JST ERATO SATO Live Bio-forecasting project

Senior Researcher, Disease Dynamics group
The Thomas N. Sato BioMEC-X Laboratories
Advanced Telecommunication Research International (ATR)

注意

- 9/11でつまづいたところや、河岡のミスがあったところは、スライドを追加することによってさらに説明を足したりしています。追加分で大事なところは赤字にしてあります。質問はskawaoka@atr.jpまで。

自己紹介

- 2012年 「生殖細胞ゲノムを護る小分子RNAに関する研究」で学位を取得

- 2012年-2014年 米国コールドスプリングハーバー研究所にて、「白血病細胞におけるクロマチン制御因子の機能解析」

- 2014年4月-現在 JST ERATO佐藤ライブ予測制御プロジェクト　コンテクストバイオロジーグループ　グループリーダー/ATR 主任研究員

インフォマティクスに関わり始めたのは大学院生のころ

- 扱っていた小分子RNAの解析にどうしても次世代シーケンサーとインフォマティクス解析が必要だった

- (当時の噂で)欧米では次世代シーケンサー技術の発展に伴ってインフォマティクスのトレーニングがかなり充実してきており、扱う研究トピックによっては、「分からない」=「大損かも」という構図ができつつあった

アグリバイオの門をたたいた

はじめる前の印象

- データをとってぽちっとボタンを押せばデータが出る
- コンピュータ任せだし、間違ったりすることはない
- 方法論は完璧に確立している

はじめた後の印象
―基本的に実験と変わらない!?

- 説明書(プロトコル)を読みながら、状況に応じてad hocに対応していく必要がある

- ヒューマンエラーの入りうる余地が腐るほどある (ファイルの名前の書き間違えとか)

- 方法論を研究している方々がおり、日進月歩で開発改良が進んでいる (=現行の方法は完璧ではなく、できることとできないことがある)

今日の実習の目的

- 実験者(≒インフォとしては素人)の目線からやるエンドユーザー的ChIP-seq解析を身につける

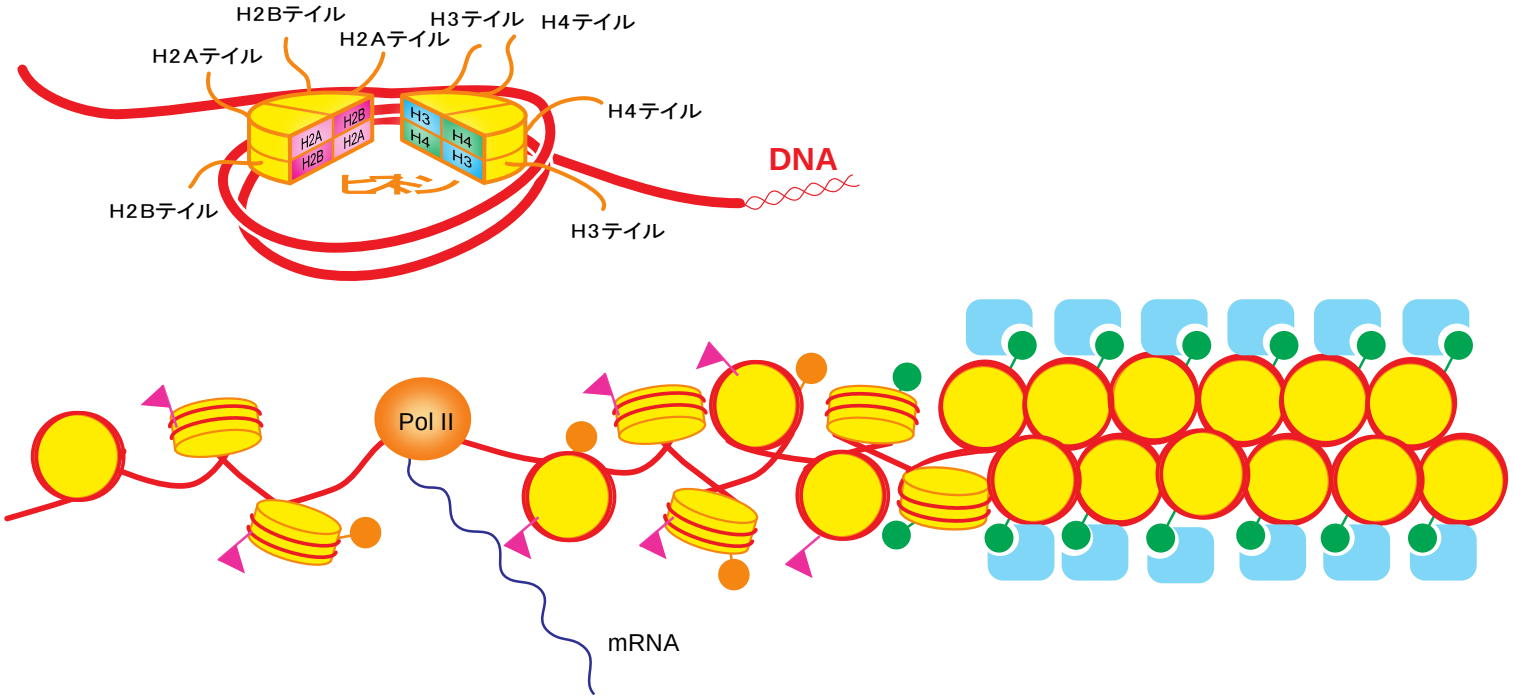
- 実験者の立場からみたChIP-seqの長所と、かゆいところ(≒ChIP-seqでは分からないところ)を理解する

目次

- そもそもChIP-seqとは
- ChIP-seqデータの見つけ方、手に入れ方
- ChIP-seqデータの見方
- いろいろ

そもそもChIP-seqとは

- 自分が興味のあるタンパク質 (ヒストン、クロマチン結合タンパク質、転写因子など)が、ゲノムにどのように局在しているかを、ある程度の解像度で調べる方法



今日の方針

- どんなときに、何のために、どういうふうにChIP-seq解析を行うのか、ということを理解してもらいたい

- というわけで、せっかくなので、実際に研究をしているつもりになってやってもらいます (とはいえもちろん、ひとのデータを使いますので、その端々に、「他人が出したデータをどのように検索し、使うか」という視点での解説を加えます)

ChIP-seq解析の流れ

- データの取得とクオリティチェック
- ゲノムへのマッピング
- 基本的な特徴付け
- エンリッチメントの定義 (Peak Call)

ChIP-seq解析の流れ (講義の後半)

- CallされたPeakの詳細な解析
- モチーフ解析
- 共局在解析

せっかくなので、問題を設定

- iPS細胞
- 山中因子 (OSKM factor)→結局全部転写調節因子
- 山中因子のゲノム上での局在を調べる

Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome

Abdenour Soufi,[1] Greg Donahue,[1] and Kenneth S. Zaret[1,*]
[1]Department of Cell and Developmental Biology, Smilow Center for Translational Research, Perelman School of Medicine, University of Pennsylvania, Building 421, Rooms 131–132, 3400 Civic Center Boulevard, Philadelphia, PA 19104-5157, USA
*Correspondence: zaret@upenn.edu
http://dx.doi.org/10.1016/j.cell.2012.09.045

SUMMARY

The ectopic expression of transcription factors can reprogram cell fate, yet it is unknown how the initial binding of factors to the genome relates functionally to the binding seen in the minority of cells that become reprogrammed. We report a map of Oct4, Sox2, Klf4, and c-Myc (O, S, K, and M) on the human genome during the first 48 hr of reprogramming fibroblasts to pluripotency. Three striking aspects of the initial chromatin binding events include an unexpected role for c-Myc in facilitating OSK chromatin engagement, the primacy of O, S, and K as pioneer factors at enhancers of genes that promote reprogramming, and megabase-scale chromatin domains spanned by H3K9me3, including many genes required for pluripotency, that prevent initial OSKM binding and impede the efficiency of reprogramming. We find diverse aspects of initial factor binding that must be overcome in the minority of cells that become reprogrammed.

chromatin and direct the binding of other transcription factors (Cirillo et al., 2002; Zaret and Carroll, 2011). Other transcription factors can access their chromatin target sites cooperatively (Adams and Workman, 1995; Filion et al., 2010). Reprogramming to pluripotency is stepwise, with morphological changes and apoptosis occurring in the first 48 hr (Samavarchi-Tehrani et al., 2010; Smith et al., 2010), changes in H3K4me2 modification occurring by one cell division (48 hr) (Koche et al., 2011), and transcriptional changes occurring at genes whose promoters were marked by H3K4me3 (Koche et al., 2011; Mah et al., 2011). Early steps involve a mesenchymal-to-epithelial transition (MET) via silencing of *Snail* genes, suppression of TGF-β signaling, and upregulating *E-cadherin* (*CDH1*) (Li et al., 2010; Mah et al., 2011). The P53 pathway promotes apoptosis and senescence, and its suppression enhances reprogramming (Hong et al., 2009; Kawamura et al., 2009; Marión et al., 2009). The timing and extent to which these genes and pathways are first targeted during reprogramming are unknown.

Genome-wide occupancy maps of embryonic stem (ES) and iPS cells have revealed that the OSK factors are central to the core pluripotency network, whereas c-Myc, bound to promoters, is part of a distinct network (Boyer et al., 2005; Chen et al., 2008; Kim et al., 2008; Sridharan et al., 2009). Mouse fibroblasts at

そもそもChIP-seqデータの見つけ方

- 誰かがやったデータを手に入れて解析する
- とりあえずUCSC genome browserを眺めてみる
- 自分でやる

ChIP-seqデータの見つけ方
論文で見つける

gramming, yet c-Myc ectopic expression, along with OSK, dramatically enhances the efficiency and the kinetics of the early steps of the process, particularly in human cells (see Introduction). Our work provides an explanation for the early role of c-Myc. We find that c-Myc facilitates, though is not absolutely required for, the initial engagement of Oct4, Sox2, and Klf4 with many chromatin sites, with no indirect effect on OSK engagement at sites lacking a c-Myc motif and lacking c-Myc binding (Figures 3C and 3D). Thus, we suggest that c-Myc has a direct mechanistic role in facilitating the action of OSK.

A striking difference between the initial OSKM network and the network seen in pluripotent cells is that, initially, Oct4, Sox2, and Klf4 bind extensively with c-Myc at distal elements of silent genes, whereas in pre-iPS and ES cells, c-Myc targets a distinct network of mainly active genes (see Introduction). Indeed, many such targets at 48 hr are for genes that promote apoptosis and senescence (Figures 2B and S3D), which is prominent within 48 hr when c-Myc is used with OSK to reprogram cells (Figure 3A). Thus, in the bulk population, many ectopic binding events contribute negatively to reprogramming. We suggest that the initial binding of ectopic transcription factors to such genes may serve as a protective mechanism to eliminate cells in which aberrant transcription factor expression has occurred, thereby preventing deleterious transdifferentiation and metaplasia.

In conclusion, the initial binding of OSKM to the fibroblast genome has provided diverse insights into the parameters that promote and impede the reprogramming process. By stepwise dissecting the mechanisms of reprogramming, starting from the initial binding of pioneer-like factors, we ultimately expect to improve the quality and efficiency of the process.

uninfected BJ control cells treated with 1 μg/ml dox for 48 hr, prior to ES culture. For ChIP of the OSKM factors in ES/iPS cells, the chromatin was obtained from hFib-iPS cells at passage 15 and human H1-ES cells (Thomson et al., 1998) at passage 35 that were expanded under feeder-free conditions for three passages (Ludwig et al., 2006).

Quantitative PCR Analysis

DNA from ChIP was amplified (ABI 7900HT Real-Time PCR System) with Power SYBR Green qPCR mix (ABI 4367659) as follows: 50°C for 4 min and 95°C for 10 min and then 45 cycles of 95°C for 15 s, 54°C for 15 s, and 72°C for 45 s. Gene targets and oligonucleotides are in Table S2. PCR specificity for each primer pair was measured by gel electrophoresis and melting curve analysis. PCR efficiency for each primer pair was set between 90% and 100% by generating a standard curve from a 5 log dilution range of input DNA (slope of ~3.2 and $R^2 > 0.98$). Threshold cycle values (Ct) used in our analysis were from three PCR replicates (SD $\leq$ 0.15) and fit within the ±8% dynamic range of the standard curve. The enrichment of ChIP DNA over input DNA was calculated by enrichment = $2^{(\text{CtChIP DNA} - \text{Ctinput DNA})}$.

Identification of OSKM Binding Sites and Computational Analysis

OSKM DNA libraries were sequenced by using an Illumina GA2 Sequencer and aligned to the human genome (NCBI v36 assembly) by using ELAND (default parameters). Nonunique sequence tags were removed from consideration. The remaining tags were used to call peaks with MACS (MFOLD = 16) and were controlled for a FDR of 0.005 (0.5%). Computational analyses were performed as described by references in the text and in the Extended Experimental Procedures.

ACCESSION NUMBERS

The GEO repository accession number for the ChIP-sequenced tags, peak assignments, and DBR locations reported in this paper is GSE36570.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2012.09.045.

GEOデータベースで検索する

NCBI Resources How To kawashin My NCBI Sign Out

GEO DataSets | GEO DataSets | Search | Advanced | Help

GEO DataSets

This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.

Getting Started

- GEO Documentation
- GEO FAQ
- About GEO DataSets
- Construct a Query
- Download Options

GEO Tools

- Submit to GEO
- Advanced Search
- DataSet Browser
- Programmatic Access
- GEO2R

More Resources

- GEO Home
- GEO Profiles
- Epigenomics
- SRA

NCBI Resources How To kawashin My NCBI Sign Out

GEO DataSets | GEO DataSets | iPS Oct4 | Search | Save search | Advanced | Help

Show additional filters

Entry type
DataSets (6)
Series (128)
Samples (402)
Platforms (3)

Organism
Select ...

Study type
Expression profiling by array

Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

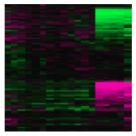
Results: 1 to 20 of 539

<< First < Prev Page 1 of 27 Next > Last >>

1. Childhood cerebral adrenoleukodystrophy patient-specific iPSC model

Analysis of induced pluripotent stem cells (iPSC) derived from childhood cerebral adrenoleukodystrophy (CCALD) patient dermal fibroblasts. Inflammatory brain demyelination is observed in CCALD patients. Results provide insight into molecular mechanisms underlying CCALD pathogenesis.

Organism: Homo sapiens
Type: Expression profiling by array, transformed count, 2 cell type, 2 disease state sets
Platform: GPL571 Series: GSE34309 18 Samples



Filters: Manage Filters

Top Organisms [Tree]
- Homo sapiens *(261)*
- Mus musculus *(260)*
- Macaca mulatta *(11)*
- Rattus norvegicus *(10)*
- Pan paniscus *(1)*

More...

Find related data

例えば7番をクリック

7. OSKM factors cooperatively engage chromatin to initiate reprogramming

(Submitter supplied) Reprogramming cells from one fate to another, using transcription factors, generates cells for research and potential therapy, yet little is known about the initial engagement of reprogramming factors with the genome. We mapped the interactions between **Oct4**, Sox2, Klf4, and c-Myc (OSKM) and the human genome during the first 48 hours of cellular reprogramming to pluripotency. Unlike that reported in ES/**iPS** cells, we find extensive overlap in the initial binding of OSKM, demonstrating that the initial regulatory network differs markedly from that in pluripotency. more...

Organism: Homo sapiens
Type: Genome binding/occupancy profiling by high throughput sequencing
Platform: GPL10999 7 Samples
Download data: GEO (BED, BW), SRA SRP011557
Series Accession: GSE36570 ID: 200036570
PubMed Full text in PMC Similar studies

8. Genome-wide analysis of histone modification, protein-DNA binding, cytosine methylation and transcriptome data in mouse and human ES cells and pig **iPS** cells

(Submitter supplied) Genome-wide analysis of histone modification (H2AZ, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3 and H3K9me3), protein-DNA binding (TAF1, P300, Pou5f1 and Nanog), cytosine methylation and transcriptome data in mouse and human ES cells and pig **iPS** cells We generated histone modification data (H2AZ, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3 and H3K9me3) and protein-DNA binding data (TAF1, P300, Pou5f1 and Nanog) using Chromatin Immunoprecipitation followed by short sequencing (**ChIP**-seq), cytosine methylation data using methylated DNA immunoprecipitation followed by sequencing (MeDIP-seq) and DNA digestion by methyl-sensitive restriction enzymes followed by sequencing (MRE-seq), transcriptome data with RNA short sequencing (RNA-seq) in human embryonic stem cells, mouse embryonic stem cells, pig induced pluripotent stem cells and mouse embryonic stem cells under activin-A-induced-differentiation.

Series GSE36570

Query DataSets for GSE36570

Status	Public on Nov 16, 2012
Title	OSKM factors cooperatively engage chromatin to initiate reprogramming
Organism	Homo sapiens
Experiment type	Genome binding/occupancy profiling by high throughput sequencing
Summary	Reprogramming cells from one fate to another, using transcription factors, generates cells for research and potential therapy, yet little is known about the initial engagement of reprogramming factors with the genome. We mapped the interactions between Oct4, Sox2, Klf4, and c-Myc (OSKM) and the human genome during the first 48 hours of cellular reprogramming to pluripotency. Unlike that reported in ES/iPS cells, we find extensive overlap in the initial binding of OSKM, demonstrating that the initial regulatory network differs markedly from that in pluripotency. OSK act as pioneer factors for c-Myc, and c-Myc enhances the engagement of OSK, including at many genes that are required for conversion to pluripotency. Distal enhancer sites in closed chromatin dominate the initial OSKM distribution. Hierarchical chromatin binding during reprogramming resembles that employed during development.
Overall design	Four chIP-seq data sets (Oct4, Sox2, Klf4, and c-Myc) are included, one lane per factor, no replicates. Also included is an input lane from the same conditions and two mock lentiviral controls (no exogenous OSKM factors) treated with Oct4 IP and c-Myc IP.

発見できるデータの実際

Platforms (1)	GPL10999 Illumina Genome Analyzer IIx (Homo sapiens)
Samples (7)	GSM896985 Oct4 ChIP-Seq at 48hrs Post-Induction
⊞ More...	GSM896986 Sox2 ChIP-Seq at 48hrs Post-Induction
	GSM896987 Klf4 ChIP-Seq at 48hrs Post-Induction

Relations

SRA	SRP011557
BioProject	PRJNA153727

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
SRP/SRP011/SRP011557		(ftp)	SRA Study
GSE36570_All_48hrs_MTFBRs.bed.gz	1.9 Mb	(ftp)(http)	BED
GSE36570_DBRs.bed.gz	2.9 Kb	(ftp)(http)	BED
GSE36570_RAW.tar	3.4 Gb	(http)(custom)	TAR (of BED, BW)

Raw data provided as supplementary file

Processed data provided as supplementary file

Processed data is available on Series record

発見できるデータのタイプ
これだけだとちょっと分かりにくい…

Index of ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP011/SRP011557/

Up to higher level directory

Name	Size	Last Modified
SRR445816		2012/08/12 午前0:00:00
SRR445817		2012/08/12 午前0:00:00
SRR445818		2012/08/12 午前0:00:00
SRR445819		2012/08/12 午前0:00:00
SRR445820		2012/08/12 午前0:00:00
SRR566752		2012/09/10 午前0:00:00
SRR566753		2012/09/10 午前0:00:00

個々のデータのほうが分かりやすい

Platforms (1) GPL10999 Illumina Genome Analyzer IIx (Homo sapiens)

Samples (7) GSM896985 Oct4 ChIP-Seq at 48hrs Post-Induction
More...
GSM896986 Sox2 ChIP-Seq at 48hrs Post-Induction
GSM896987 Klf4 ChIP-Seq at 48hrs Post-Induction

Relations

SRA SRP011557
BioProject PRJNA153727

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
SRP/SRP011/SRP011557		(ftp)	SRA Study
GSE36570_All_48hrs_MTFBRs.bed.gz	1.9 Mb	(ftp)(http)	BED
GSE36570_DBRs.bed.gz	2.9 Kb	(ftp)(http)	BED
GSE36570_RAW.tar	3.4 Gb	(http)(custom)	TAR (of BED, BW)

Raw data provided as supplementary file

Processed data provided as supplementary file

Processed data is available on Series record

Sample GSM896985

Query DataSets for GSM896985

Status	Public on Nov 16, 2012
Title	Oct4 ChIP-Seq at 48hrs Post-Induction
Sample type	SRA
Source name	BJ_Oct4 ChIP-Seq_48hr_postinduction
Organism	Homo sapiens
Characteristics	cell line: BJ cell type: foreskin fibroblast cell genotype/variation: infected with lentiviruses encoding for dox-inducible Oct4, Sox2, Klf4, and c-Myc, along with lentiviruses expressing rtTA2M2 time point: 48hrs post-induction with with 1µg/ml dox chip antibody: Oct4 chip antibody vendor: abcam chip antibody cat. #: ab19857
Treatment protocol	Lentiviral production and titration is described in the Supplemental Experimental Procedures. BJ cells at passage 10 were infected with lentiviruses encoding for dox-inducible Oct4, Sox2, Klf4, and c-Myc, along with lentiviruses expressing rtTA2M2 in the presence of 4.5 µg/ml polybrene. The expression of the OSKM factors was induced by treating the infected BJ cells with 1µg/ml dox for 48 hours. Cells were cross-linked with 1% formaldehyde for 10 minutes at room temperature. BJ cells (MOCK, not-infected) were treated with

E-mail	zaret@upenn.edu
Phone	2155735813
Organization name	University of Pennsylvania School of Medicine
Department	Cell and Developmental Biology
Lab	Zaret lab
Street address	9-131, SCTR, 3400 Civic Center Boulevard
City	Philadelphia
State/province	PA
ZIP/Postal code	19104-5157
Country	USA

Platform ID	GPL10999
Series (1)	GSE36570 OSKM factors cooperatively engage chromatin to initiate reprogramming

Relations

SRA	SRX130060
BioSample	SAMN00828870

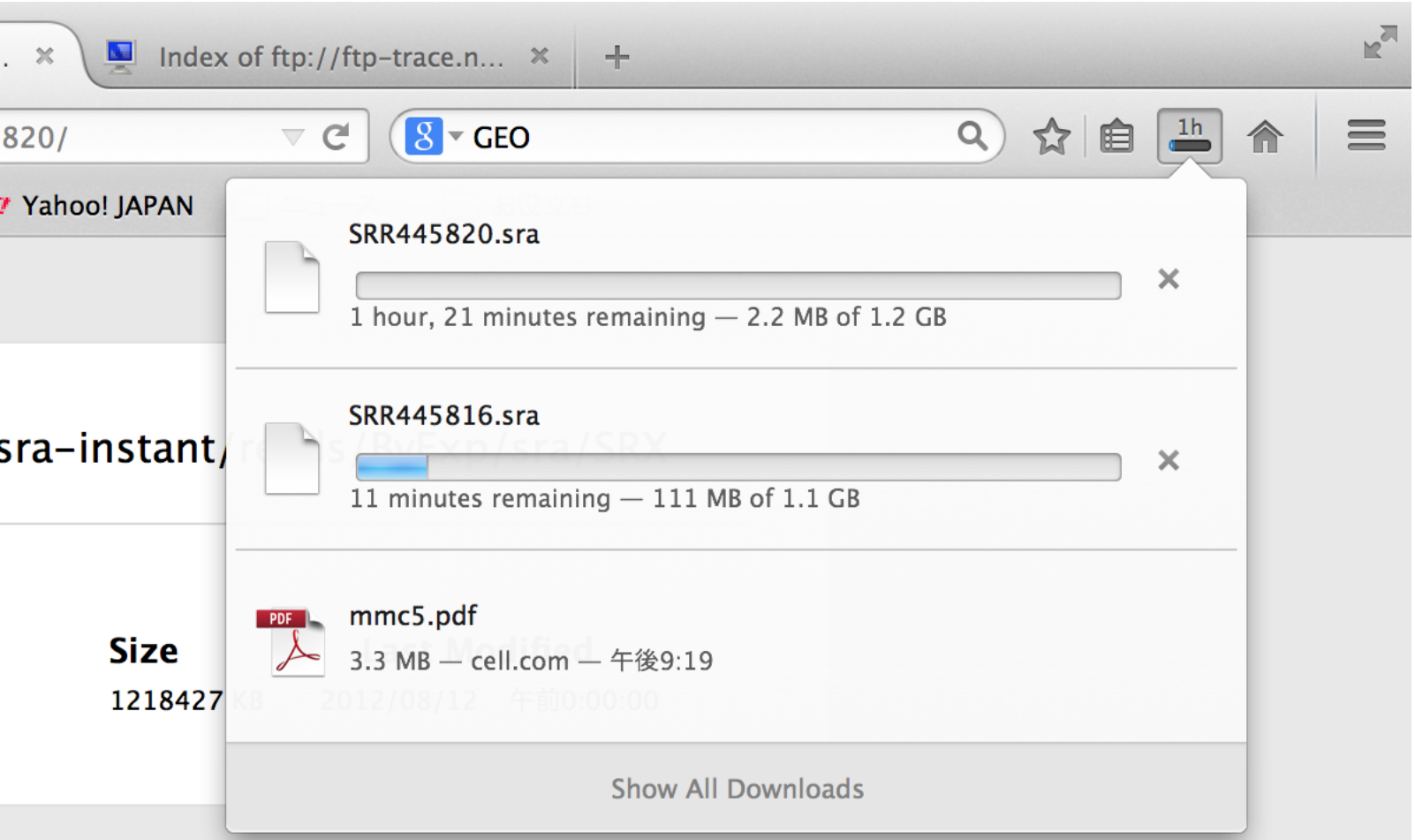
Supplementary file	Size	Download	File type/resource
SRX/SRX130/SRX130060		(ftp)	SRA Experiment
GSM896985_Oct4_48hrs.bed.gz	289.0 Mb	(ftp)(http)	BED
GSM896985_Oct4_48hrs.bw	302.0 Mb	(ftp)(http)	BW
GSM896985_Oct4_48hrs_peaks.bed.gz	476.0 Kb	(ftp)(http)	BED

Raw data provided as supplementary file

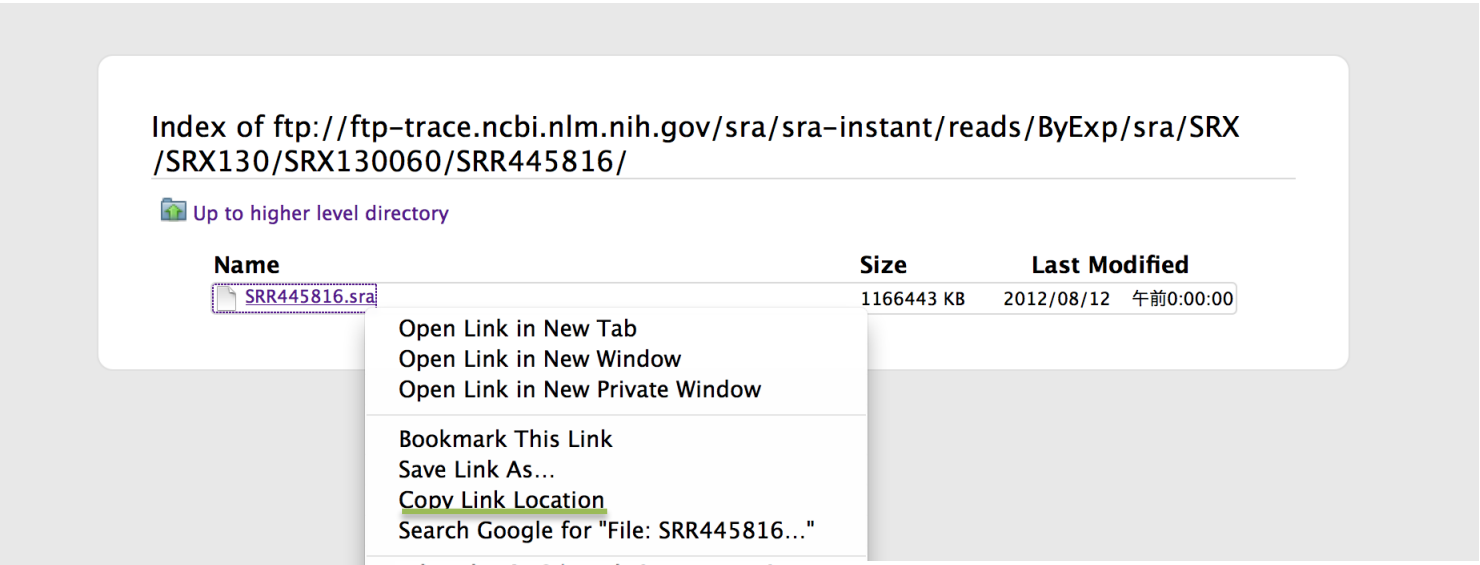
Processed data provided as supplementary file

Processed data is available on Series record

自分のPCに直接ダウンロード



作業スペース(サーバーとか)にダウンロードしたい場合




```
-bash-4.1$ cd Lecture/
-bash-4.1$ ls
-bash-4.1$ wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX%2FSRX130%2FSRX130060/SRR445816/SRR445816.sra
--2014-07-15 14:21:43--  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX%2FSRX130%2FSRX130060/SRR445816/SRR445816.sra
           => `SRR445816.sra'
Resolving ftp-trace.ncbi.nlm.nih.gov... 130.14.250.12, 2607:f220:41e:250::12
Connecting to ftp-trace.ncbi.nlm.nih.gov|130.14.250.12|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /sra/sra-instant/reads/ByExp/sra/SRX/SRX130/SRX130060/SRR445816 ... done.
==> SIZE SRR445816.sra ... 1194436637
==> PASV ... done.    ==> RETR SRR445816.sra ... done.
Length: 1194436637 (1.1G) (unauthoritative)

 8% [==>                                    ] 100,512,090  714K/s  eta 11m 5s
```

SRAって？
http://www.ncbi.nlm.nih.gov/books/NBK47537/

Overview of Input Formats

Go to:

General Considerations

The SRA is a “raw data” archive, and requires per-base quality scores for all submitted data. Thus, unlike GenBank and some other NCBI repositories, FASTA and other sequence-only formats are not sufficient for submission. FASTA can, however, be submitted as a reference sequence(s) for BAM files or as part of a FASTA/QUAL pair (see below).

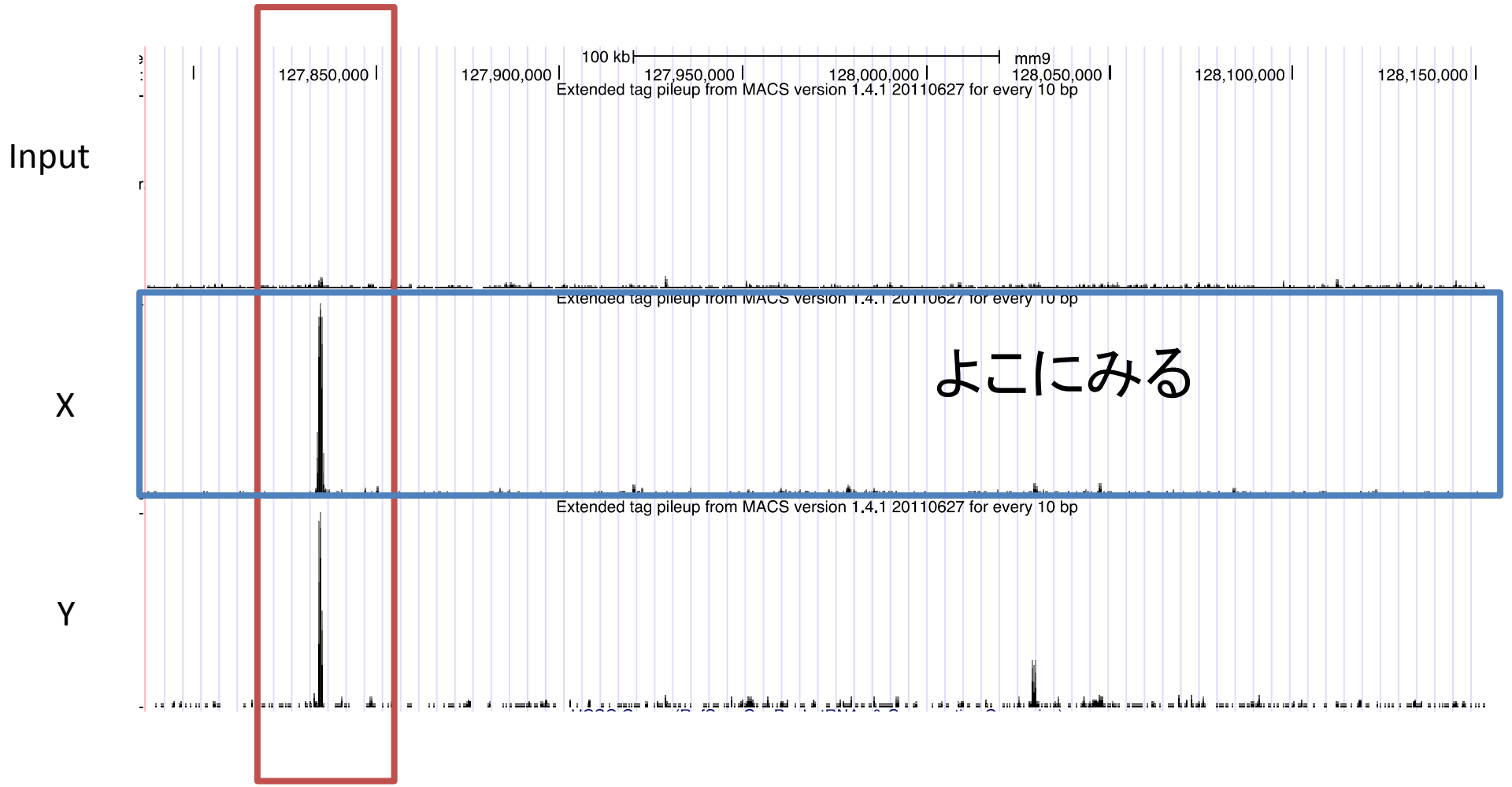
The SRA data model has transitioned from “dumps” of whole flowcell lanes or production runs into a semi-curated database of sample-specific sequencing libraries. This has implications for the types of data that we accept. Most specifically, barcoded/batch files should be split into per-sample data files (“demultiplexed”). Demultiplexing makes the sample - data linkage unambiguous in our database and should improve both the clarity and usability of submitted data. Please email sra@ncbi.nlm.nih.gov if you have specific questions about data requirements vis-à-vis samples.

Conversion to the SRA archive format (described below) is NOT required for submission. However, the SRA Toolkit can be used to “test load” your files locally if you would like to validate them prior to submission. BAM files can be evaluated with ‘bam-load’ and FASTQ files can be evaluated with ‘latf-load’ (first released in Toolkit version 2.3.5). These load utilities are effectively stand-alone and can be run by most submitters. Other SRA loading software, such as ‘sff-load’, ‘abi-load’, etc. are dependent on SRA XML documents and are only recommended for advanced users. If you elect to test load your data file(s) and encounter problems, please email sra@ncbi.nlm.nih.gov if you have questions.

Inputのデータがないが結構ありますが、Inputは重要です

Input=免疫沈降をしない(特定のタンパク質に結合しているかどうかを気にしない)で全ゲノムをシーケンスするサンプル。理想的には完全に均一な頻度でゲノムが読まれるはずだが、実際はそうはならない(ライブラリ作成の際に存在するPCRのステップなどが原因候補)。データベースでは、Input、や、whole cell lysate、のように記述されていることが多い。

ChIP-seqはたてとよこ方向でデータをみる。Inputはたてにみるのに大事。



データがあるかどうかを確認

- 分かりやすいように、作業スペースは kawaokaというフォルダにします

- mkdir kawaoka

- フォルダ内に必要なものを移動してください。 mv [ファイル] /Desktop/kawaoka

ファイルフォーマット変換
サイズが大きすぎるので今日はやりません

$ fastq-dump [対象ファイル]

```
-bash-4.1$ fastq-dump SRR445816.sra &
[1] 57805
-bash-4.1$

-bash-4.1$
-bash-4.1$ Read 58571678 spots for SRR445820.sra
Written 58571678 spots for SRR445820.sra

[1]+  終了                    fastq-dump SRR445820.sra

-bash-4.1$
-bash-4.1$ ls
SRR445816.fastq  SRR445816.sra  SRR445820.fastq  SRR445820.sra  memo
-bash-4.1$
```

シーケンスデータのクオリティをチェック

- FastQCを使ってシーケンスのクオリティをチェックすることができる
- 得られた配列の端はクオリティが低くなりがち
- 変にoverrepresentされる配列がないか？
- uni.fastqを使用する

クオリティチェックの実際

$ fastqc [対象fastqファイル]

(例) $ fastqc SRR445816.fastq
→SRR445816_fastqc.zipができる

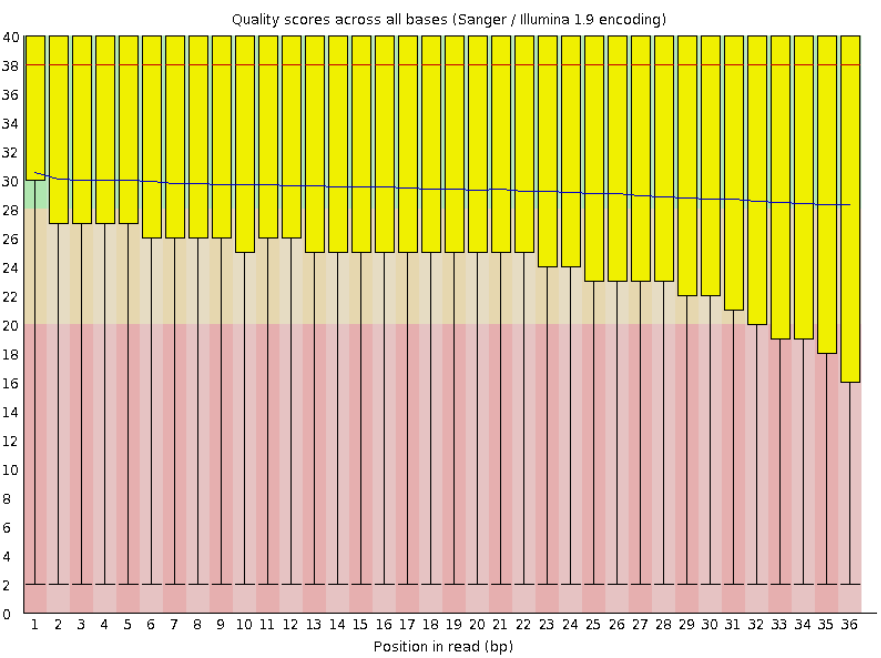
Basic Statistics

Measure	Value
Filename	SRR445816.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	56616764
Filtered Sequences	0
Sequence length	36
%GC	43

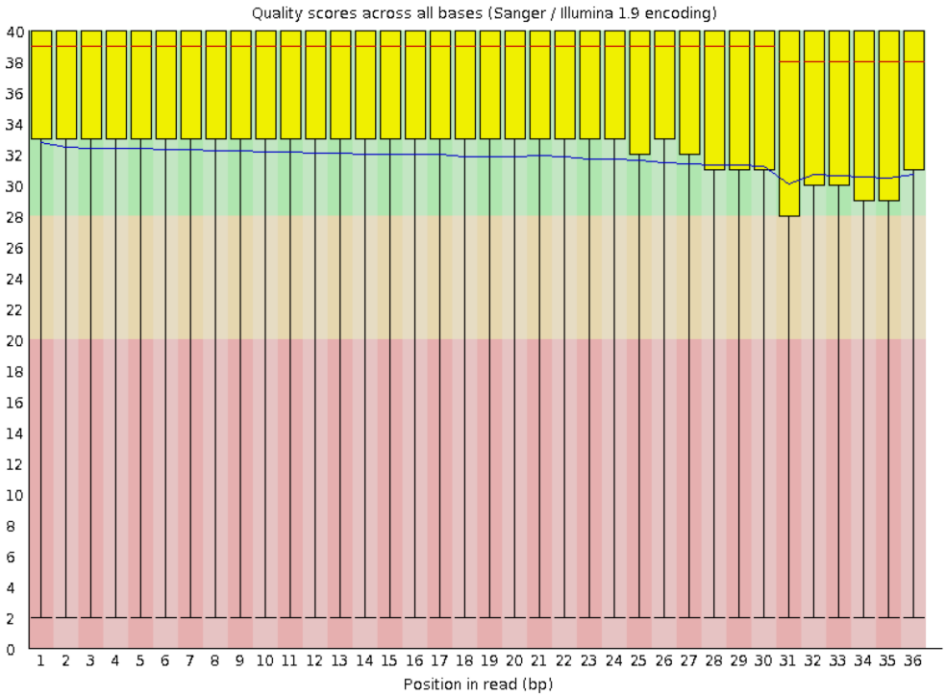
Basic Statistics

Measure	Value
Filename	uni.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	25000
Filtered Sequences	0
Sequence length	36
%GC	43

塩基ごとのクオリティ
20より低いとあまりよくない

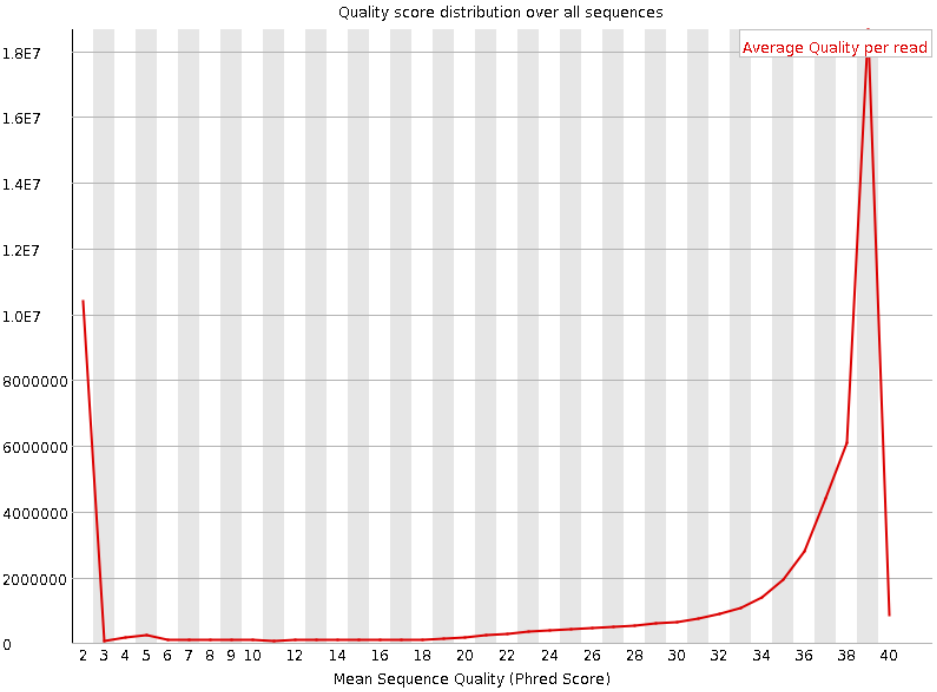


SRR445816.fastq

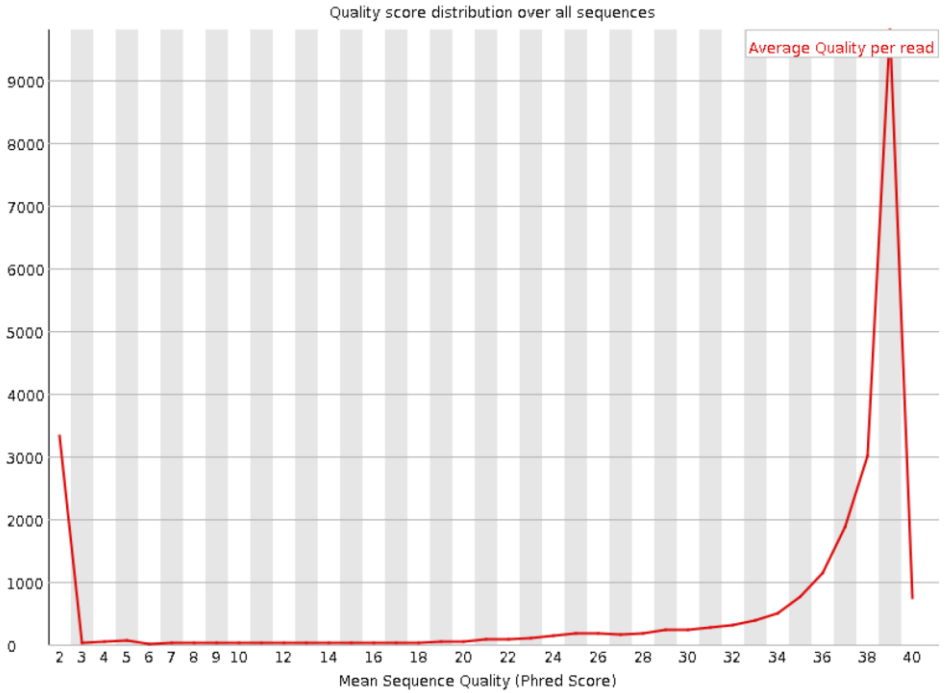


uni.fastq

塩基ごとのクオリティの分布

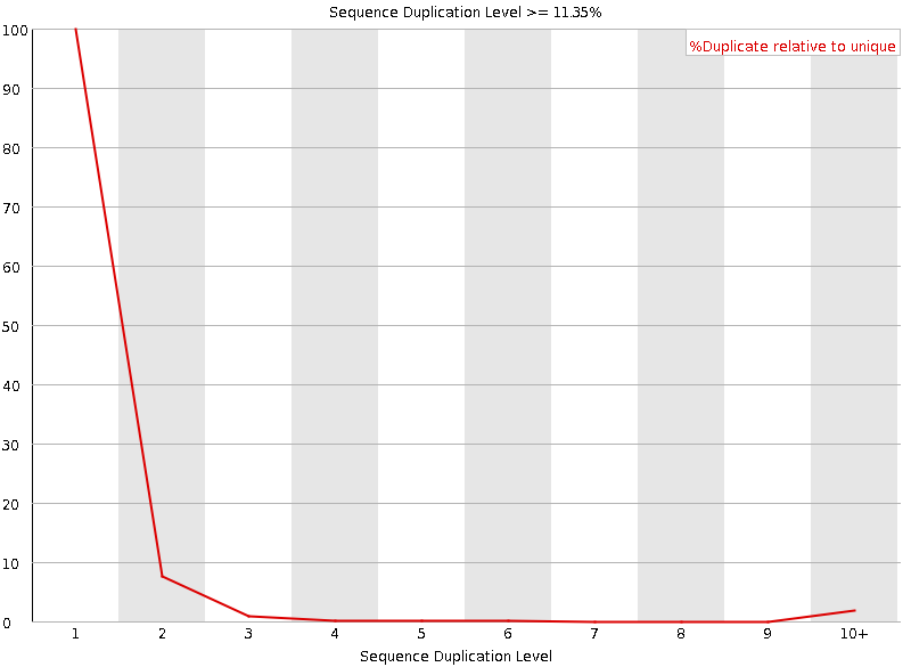


SRR445816.fastq

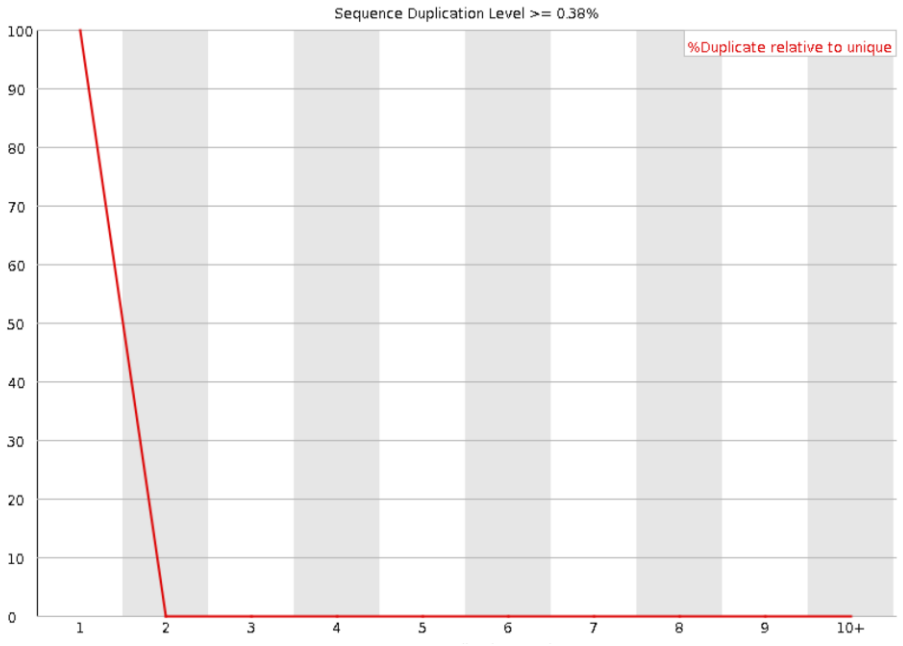


uni.fastq

Duplication level
同じ配列が何回登場したか？



SRR445816.fastq



uni.fastq

Overrepresented seuquences

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	222515	0.39301963637483767	No Hit

- 前ページと関連
- 登場回数の多すぎるシークエンスを出してくれる
- 経験談で、読めていたのがアダプターダイマーばかりだった、ということがあります

ChIP-seq解析の流れ

- データの取得とクオリティチェック
- ゲノムへのマッピング
- 基本的な特徴付け
- エンリッチメントの定義 (Peak Call)

ゲノムへのマッピング (Bowtie2)

Usage

```
bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} -S [<hit>]
```

Main arguments

`-x <bt2-idx>`	The basename of the index for the reference genome. The basename is the name of any of the index files up to but not including the final `.1.bt2` / `.rev.1.bt2` / etc. `bowtie2` looks for the specified index first in the current directory, then in the directory specified in the `BOWTIE2_INDEXES` environment variable.
`-1 <m1>`	Comma-separated list of files containing mate 1s (filename usually includes `_1`), e.g. `-1 flyA_1.fq,flyB_1.fq`. Sequences specified with this option must correspond file-for-file and read-for-read with those specified in `<m2>`. Reads may be a mix of different lengths. If `-` is specified, `bowtie2` will read the mate 1s from the "standard in" or "stdin" filehandle.
`-2 <m2>`	Comma-separated list of files containing mate 2s (filename usually includes `_2`), e.g. `-2 flyA_2.fq,flyB_2.fq`. Sequences specified with this option must correspond file-for-file and read-for-read with those specified in `<m1>`. Reads may be a mix of different lengths. If `-` is specified, `bowtie2` will read the mate 2s from the "standard in" or "stdin" filehandle.
`-U <r>`	Comma-separated list of files containing unpaired reads to be aligned, e.g. `lane1.fq,lane2.fq,lane3.fq,lane4.fq`. Reads may be a mix of different lengths. If `-` is specified, `bowtie2` gets the reads from the "standard in" or "stdin" filehandle.
`-S <hit>`	File to write SAM alignments to. By default, alignments are written to the "standard out" or "stdout" filehandle (i.e. the console).

Bowtie2の使い方チェック

```
$ bowtie2
```

コマンドのオプションなどが分からなくなったら、とりあえずコマンドをたたいてみると、いろいろな情報が出てくる

ゲノムへのマッピング (Bowtie2)

$bowtie2 –x [ゲノム] –q [fastqファイル]
-N [0 or 1] –S [.sam]

$ nohup bowtie2 –x [ゲノム]
–q [fastqファイル]
-N [0 or 1] –S [.sam] &

Multiseed heuristic

To rapidly narrow the number of possible alignments that must be considered, Bowtie 2 begins by extracting substrings ("seeds") from the read and its reverse complement and aligning them in an ungapped fashion with the help of the FM Index. This is "multiseed alignment" and it is similar to what Bowtie 1 does, except Bowtie 1 attempts to align the entire read this way.

This initial step makes Bowtie 2 much faster than it would be without such a filter, but at the expense of missing some valid alignments. For instance, it is possible for a read to have a valid overall alignment but to have no valid seed alignments because each potential seed alignment is interruped by too many mismatches or gaps.

The tradeoff between speed and sensitivity/accuracy can be adjusted by setting the seed length (`-L`), the interval between extracted

実際にマップしてみる
(nohupはoptionalです)

```
$ nohup bowtie2 –x chr17/chr17_base
–q uni.fastq
-N 0 –S uni.sam &
```

chr17のみにマップ

nohupを使用する際の注意

- 使っているマシンによって、一度に複数の計算を動かせるかどうかが異なります。ひとつの計算しかできない場合、nohupで実行した計算が終了する前に次の計算をスタートしてしまうと、最初の計算が止まってしまいます。これを防ぐためには。。。

1) nohupを使用しない
2) ターミナルに”top”とタイプし、計算中の計算のリストを呼び出すことで、計算が進行中か終わったかをチェックする

以下のところで。。。

- uni.fastqをマップしたデータではなく、トリミング前のデータを出していましたが、それが混乱を読んでしまっていたので、全て削除しました

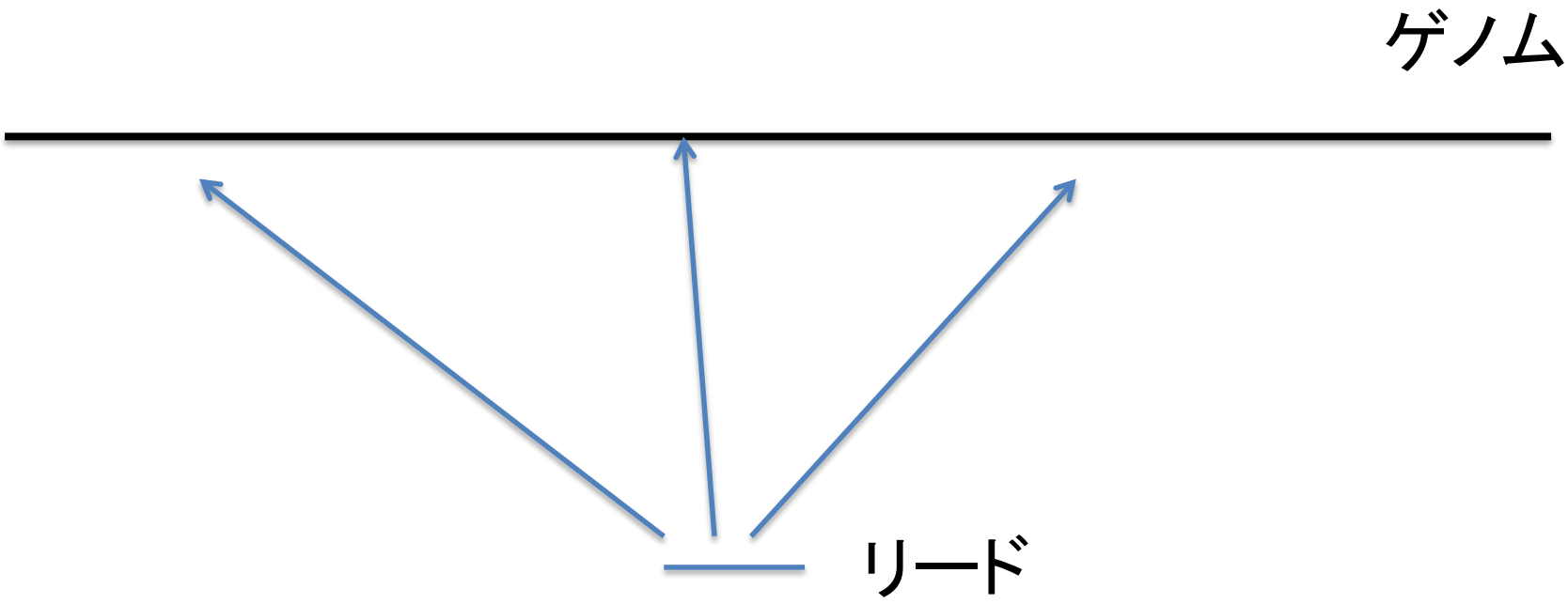
- 最初からchr17にしかマップしていなかったのにわざわざchr17にマップされたリードをgrepで抜き出すステップを作って混乱してしまったので、そのステップを削除しました

SAMファイルの理解

```
@SQ     SN:chr17        LN:81195210
@PG     ID:bowtie2      PN:bowtie2      VN:2.2.3        CL:"/home/skawaoka/seqtools/bowtie2-2.2.3/bowtie2-align-s --wrapper basic-0 -x /home/skawaoka/seqtools/hg19/chr17/chr17_base -q /home/skawaoka/Lectur
 0 -S uni.sam"
SRR445816.24975006      16      chr17   6299100 31      36M     *       0       0       AGTTAATGGGTGTAGCACACCAACATGGCACATGTA    IIIIIGIGIIIIIIIIHIIIIIIIIIFIIIIIIIII    AS:i:0  XS:i:-6 XN:i:0  XM:i:0  XO:i:
i:0  MD:Z:36 YT:Z:UU
SRR445816.24975018      16      chr17   19873930        1       36M     *       0       0       GTTCGAGACCAGCCTTGCCAACATGGTGAAACACTG    EGFBG<HHHGHEGGGDED>BGHDHHHGGFDGFHGBH    AS:i:-5 XS:i:-5 XN:i:0  XM:i:
i:0  NM:i:1  MD:Z:32C3        YT:Z:UU
SRR445816.24975025      0       chr17   19028273        0       36M     *       0       0       TGCCTGCTCTGACCTCCCAAAGTGCTGGGATTACAG    IIIIIIIIHHHHIIIIIIIIIIIIIIIDHIIIGIGHI    AS:i:-11        XS:i:-11
i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:7C0T27     YT:Z:UU
SRR445816.24975031      16      chr17   50871704        31      36M     *       0       0       TGTCACCCAGGCTGGAGTGCAGTGGCGTGATCTTAG    GIEGDIHDIIGIIIIIIHIHIIIIIEIIIHIIHIII    AS:i:0  XS:i:-5 XN:i:0  XM:i:
i:0  NM:i:0  MD:Z:36 YT:Z:UU
SRR445816.24975039      0       chr17   57227680        42      36M     *       0       0       AATTCATATTTTTTAAAAGTGATGAATCCACACTAT    IIIIIIIGIIIIIIIIIIGHIIIHIIIIIIIIIIII    AS:i:0  XN:i:0  XM:i:0  XO:i:
i:0  MD:Z:36 YT:Z:UU
SRR445816.24975062      0       chr17   33529565        1       36M     *       0       0       ATCCCTTTACCATTATGTAATGGCCTTCTTTGTCTC    HHHHHHHHHHHHHHHHGHHHHHHHHHHHGHHHHHHHH    AS:i:0  XS:i:0  XN:i:0  XM:i:
i:0  NM:i:0  MD:Z:36 YT:Z:UU
SRR445816.24975066      0       chr17   14571190        42      36M     *       0       0       AAGTTTTGTAGATGGGTAAAACTGTGATGGGACAAG    GGGEGGGGGDDHHGHHHFHHHGGGGEGGGGGHFGBHH    AS:i:0  XN:i:0  XM:i:0  XO:i:
i:0  MD:Z:36 YT:Z:UU
SRR445816.24975072      16      chr17   18721302        0       36M     *       0       0       CGCACGCACACACACACACACACACACACACCCC    EBF?E?GADGGCFEFCFADAFGGGGGFBF@@AGEFF    AS:i:-10        XS:i:-10
i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:33A1A0     YT:Z:UU
SRR445816.24975086      0       chr17   2394978 1       36M     *       0       0       GCTAACACGGTGAAACCCCGTCTCTATTAAAAATGC    IIIIIIIIIIIIIIIHIIIIIIIIHIIIIIIIIIHI    AS:i:-5 XS:i:-5 XN:i:0  XM:i:1  XO:i:
i:1  MD:Z:34A1       YT:Z:UU
SRR445816.24975089      16      chr17   9289301 1       36M     *       0       0       GGCTGGTCTTGAACTCCTAACCTCAGGTGATCCACC    GGDGGGGGGGGGGG<GGGGFGGBGGGGGGGGEEADE    AS:i:0  XS:i:0  XN:i:0  XM:i:0  XO:i:
i:0  MD:Z:36 YT:Z:UU
SRR445816.24975093      0       chr17   39202292        40      36M     *       0       0       CATGGATGAACCTGGGGGACATTACGCTCAGTGAAA    ####################################    AS:i:-6 XN:i:0  XM:i:3  XO:i:
i:3  MD:Z:24T1T1A7    YT:Z:UU
SRR445816.24975096      16      chr17   22148590        6       36M     *       0       0       TATTTCTGTGGGATCGGTGGTGATATCTCCTTTATC    ####################################    AS:i:0  XS:i:-2 XN:i:0  XM:i:
i:0  NM:i:0  MD:Z:36 YT:Z:UU
SRR445816.24975098      0       chr17   79806777        1       36M     *       0       0       CTGGGTGTGGTGGAGCACACCTGTAATCCCAGCTGC    IFIHIIIIIIHIIIIIIIIII@GIIIIIIIGIIBII    AS:i:-6 XS:i:-6 XN:i:0  XM:i:
i:0  NM:i:1  MD:Z:13T22       YT:Z:UU
SRR445816.24975099      16      chr17   49141936        0       36M     *       0       0       GAAAGAAGCCAGTCACAAAAGAATATATACTGTATG    AIIGEIGIHHHIIIHIIIIIHIIIIIIIIIIIIIII    AS:i:-18        XN:i:0  XM:i:
i:0  NM:i:3  MD:Z:22C0C5T6    YT:Z:UU
SRR445816.24975112      16      chr17   79439657        1       36M     *       0       0       ATCCTCCCACCTCAGCCTCCCAAGTAGCTGGGATTA    I>DGDGIIBIIIIIHIIIIIIIIIIIIIIIIIIII    AS:i:0  XS:i:0  XN:i:0  XM:i:
i:0  NM:i:0  MD:Z:36 YT:Z:UU
SRR445816.24975117      0       chr17   72651301        8       36M     *       0       0       TTGTGCGATCTCGGCTCACTACAACCTTCACCTCCC    DB9BBB>?>>D@:8DB?1>074;:7DD>BD>D39>;    AS:i:-13        XN:i:0  XM:i:
i:0  NM:i:3  MD:Z:1G4A5A23    YT:Z:UU
SRR445816.24975132      16      chr17   32410897        24      36M     *       0       0       TTATTTATTAATTTGTTTGAGTTCATTGTGGATTCT    ####################################    AS:i:-8 XN:i:0  XM:i:4  XO:i:
i:4  MD:Z:2T0C2G1A27 YT:Z:UU
SRR445816.24975166      0       chr17   13580924        8       36M     *       0       0       TGTAACAGAATAGAGAACCCAGAAATACACCCAAAT    HIIIIIIIHIIIIIIIIIIIIHIIHHIIIIIHIII    AS:i:-12        XN:i:0  XM:i:
i:0  NM:i:2  MD:Z:27A1G6      YT:Z:UU
SRR445816.24975167      0       chr17   53441039        0       36M     *       0       0       CGTGACCCAGTCAAGTTGATACATAAAATTAACTGT    GDGDGGGGGGHGHHBGGGGEGDGEGBGDGGDGGEGE    AS:i:-20        XN:i:0  XM:i:
```

配列情報、マップされた染色体…

Uniquely mapped readsに解析を絞ったほうが安全



ゲノムの複数箇所にリードがマップされると、
そのリードがどこからきたのか分からない！

2カ所以上にマップされたリードには“XS”が与えられている

`XS:i:<N>` Alignment score for the best-scoring alignment found other than the alignment reported. Can be negative. Can be greater than 0 in `--local` mode (but not in `--end-to-end` mode). Only present if the SAM record is for an aligned read and more than one alignment was found for the read. Note that, when the read is part of a concordantly-aligned pair, this score could be greater than `AS:i`.

$ wc uni.sam

$ 25003

$ grep “XS” uni.sam > hoge

$ wc hoge

$2515

$grep –v “XS” uni.sam > chr17_uni_uniq.sam

$wc chr17_uni_uniq.sam

$ 22488

Bowtie2のリポートをチェックする

25000 reads; of these:
  25000 (100.00%) were unpaired; of these:
    21630 (86.52%) aligned 0 times
    855 (3.42%) aligned exactly 1 time
    2515 (10.06%) aligned >1 times
13.48% overall alignment rate

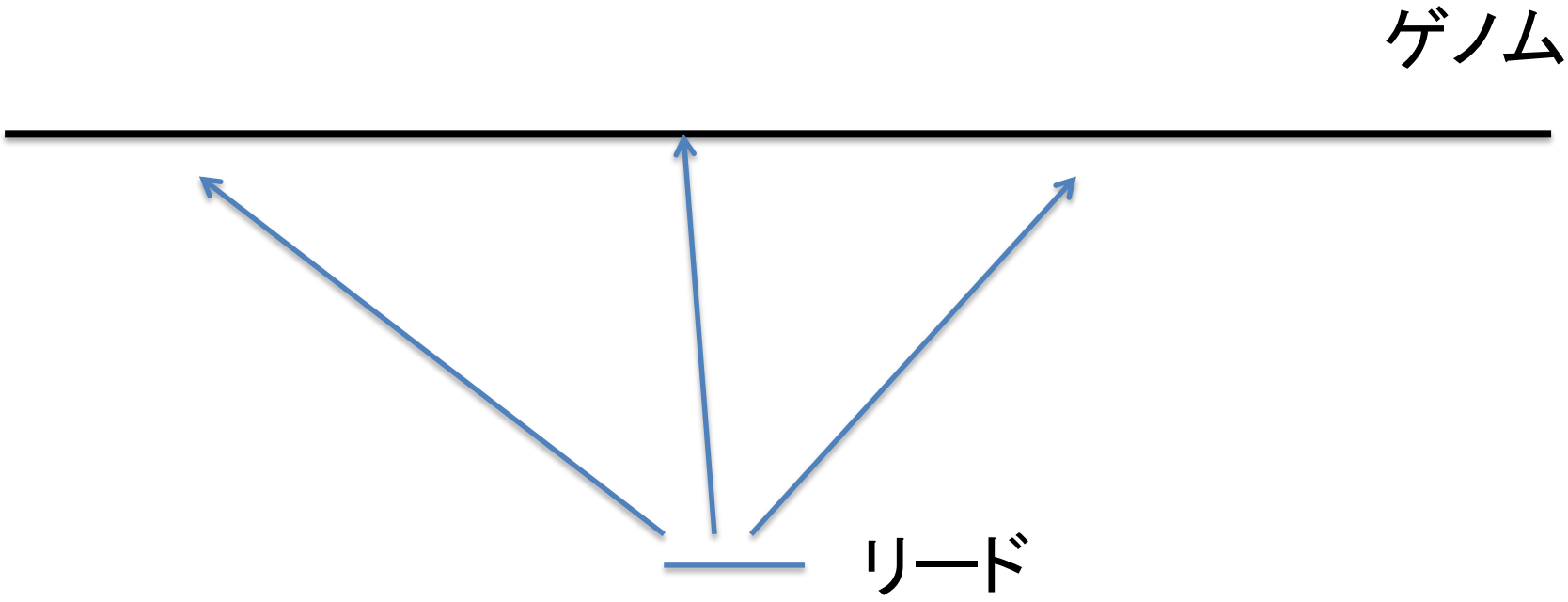
25003-22488=2515

(注1) wc uni.samで25003と3行多くなるのは、uni.samの最初の3行にbowtie2からの出力が以下のように出ているため

@HD VN:1.0 SO:unsorted
@SQ SN:chr17 LN:81195210
@PG ID:bowtie2 PN:bowtie2 VN:2.2.3

(注2) マッピング率がとても低いようにみえるのは、chr17だけにマッピングしているからです

Uniquely mapped readsに解析を絞ったほうが安全だが、一応他の方法



X箇所にマップされた場合、各箇所に1/Xをかけることにより、重みをつける

マッピングはここまで

- 以降はあらかじめこちらで用意したsamファイルを使用して実習を進めていきますので、uni.fastqのことはもう忘れてください

ChIP-seq解析の流れ

- データの取得とクオリティチェック
- ゲノムへのマッピング
- 基本的な特徴付け
- エンリッチメントの定義 (Peak Call)

SAMファイルからの作業分岐

- だいたいbamが多い。場合によってはソートしていたり、インデックス化している。

- Bedに変換しておくと、後で何かと便利ではある。

Ngsplotを使ってみよう

ngsplot

Quick mining and visualization of next-generation sequencing data by integrating genomic databases

Search projects

Project Home | Downloads | Wiki | Issues | Source

Summary | People

Project Information

Starred by 12 users
Project feeds

Code license
GNU GPL v3

Labels
R, Academic, Bioinformatics, NGS, Sequencing, ChIP-seq, RNA-seq, Visualization, Webserver, Datamining, Database

Members
shenli.sam,
shaonin...@gmail.com
1 committer
1 contributor

Featured

Wiki pages
HowAreYaxisValsCalculated
HowToCreateBed
HowToUseConfiguration
HowtoMakeThingsRight
ProgramArguments101
RunItABitFaster
SupportedGenomes
UseFurtherInfo
webngsplot

INTRODUCTION

ngs.plot is a program that allows you to easily visualize your next-generation sequencing (NGS) samples at functional genomic regions.

DNA sequencing is at the core of genomics. The NGS technology has been tremendously improved in the past few years. It can now determine more than a billion DNA sequences within a week, generating terabytes of data. Applications include but are not limited to: 1. ChIP-seq which profiles genome-wide protein-DNA interactions; 2. RNA-seq which measures the gene expression levels. It is very helpful to look at the enrichment of those sequences at various functional regions. Although a genome browser (such as the UCSC genome browser) allows a researcher to visualize these data, it limits the view to a slice of the genome. While the genome is like a huge collection of functional elements that can be classified into different categories. Each category of elements may perform distinct functions and they might further contain modules.

The signature advantage of ngs.plot is that it collects a large database of functional elements for many genomes. A user can ask for a functionally important region to be displayed in one command. It handles large sequencing data efficiently and has only modest memory requirement. For example, ngs.plot was used to draw a plot for all the genes on the mouse genome from 71GB of ChIP-seq data in 25 min, with a memory footprint of 2.7GB using 4 x 2.4GHz CPU cores. ngs.plot is also easy to use. A user only needs to create a very small text file called configuration, telling the program which samples to look at and how they should be combined with different regions, and then run the program with one command. A web-based version (integrated into Galaxy) is also available for the ones who are allergic to terminals.

Program Download Location

Since Google canceled the download function on Google code, we have to move all of our future program download files to this Google drive folder:

https://drive.google.com/folderview?id=0B1PVLadG_dCKN1liNFY0MVM1Ulk&usp=sharing

The last release on the download page is v2.08. For more recent releases, please visit the above link.

Supported Genomes

ngs.plot has an approach to install genomes on demand. It can support for any genome. All you need to do is to download an archive file and install it by yourself. The genome files can be found in this Google drive folder:

Ngsplotでできること

- ChIP-seqによって得られたリードが任意の座標の周辺にどのように分布しているかを調べることができる

- Ngsplotを入手するだけで、よく使われる座標(例えば転写開始点(TSS))などに関してはいっぱつでできるし、応用編で、Oct4結合サイトの周辺にc-Mycがどのように分布しているか、ということなども調べることができる

- インプットはインデックス化されたbamファイル

まずすること:変数のエクスポートとRパッケージのインストール

- #まず、NGSPLOT変数をエクスポートする必要があるので、ターミナルを起動して、毎回下記のコマンドを実行してください
- export NGSPLOT=/usr/local/src/ngsplot

- #NGSPLOTが依存するRパッケージをインストールする必要があります。Rを起動して、以下をコピペ
- > source("http://bioconductor.org/biocLite.R")
- > biocLite("ShortRead")
- > biocLite("BSgenome")
- > biocLite("doMC")

あらかじめ渡してあるデータセットがあることを確認してください

- chr17_Input.sam、chr17_Oct4.sam、chr17_cMyc.sam

- 練習で、ここからuniqueなリードを取り出してみましょう (注)後の解析に必要なので、3サンプル全てに関して実行してください

まずすること:ファイルの変換

- Samtoolsを使ってファイルフォーマットを変換する

I. ngs.plot.r

Use ngs.plot.r to choose a genomic region of interest and create enrichment plots of any ChIP-seq or RNA-seq samples.

```
Type "ngs.plot.r 2>&1|less" at console will give you a brief usage summary for online
reference. Here is just a truncated output:

Usage: ngs.plot.r -G genome -R region -C [cov|config]file
                  -O name [Options]

## Mandatory parameters:
  -G   Genome name. Use ngsplotdb.py list to show available genomes.
  -R   Genomic regions to plot: tss, tes, genebody, exon, cgi, enhancer, dhs or bed
  -C   Indexed bam file or a configuration file for multiplot
  -O   Name for output: multiple files will be generated
```

Bamへの変換: samtools

```
$ samtools
```

注
前の計算が終わる前に次の計算をはじめないこと！

Bamへの変換: samtools

$ samtools view –S –b [samファイル] > [bamファイル]

$ samtools view -S -b chr17_Oct4_uniq.sam > chr17_Oct4_uniq.bam

Bamファイルをソート

```
$ samtools sort
[bamファイル] […sorted]
```

```
$ samtools sort
chr17_Oct4_uniq.bam
chr17_Oct4_uniq_sorted
```

ソートしたbamファイルをインデックス化

```
$ samtools index [sorted.bam]
```

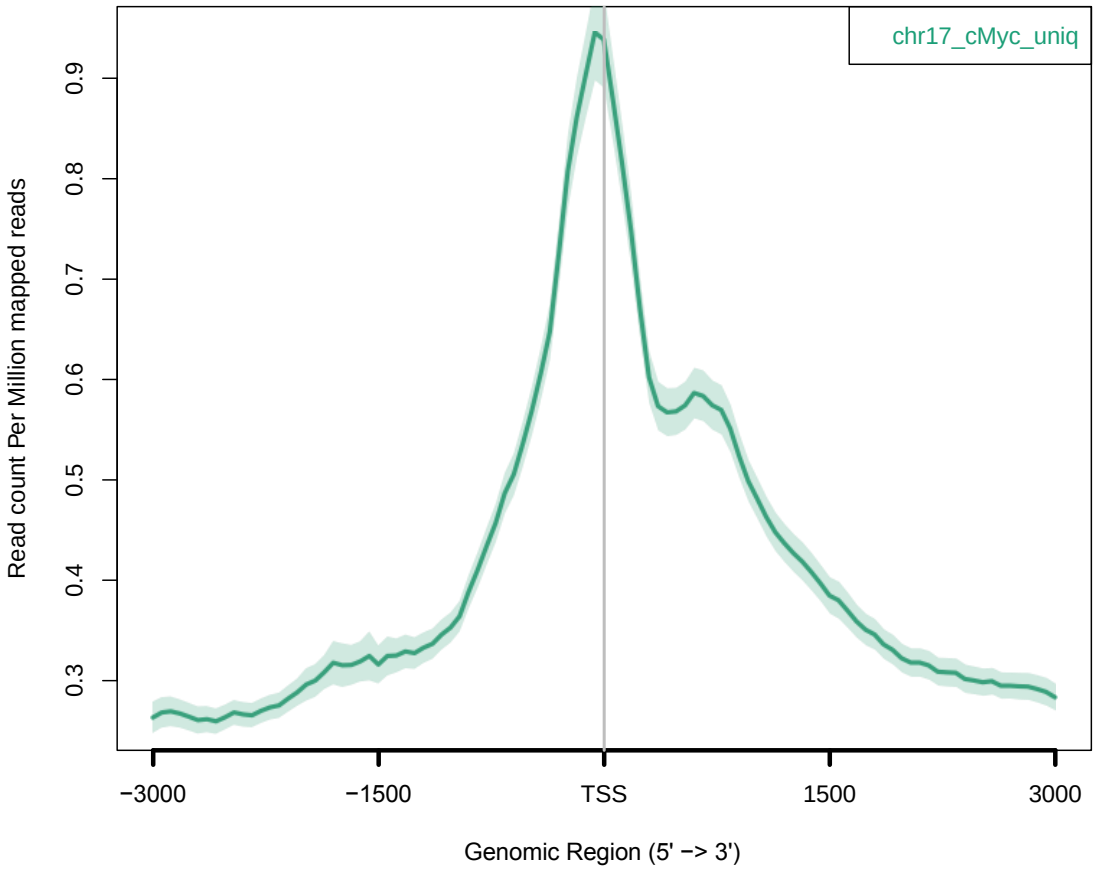
```
$ samtools index
chr17_Oct4_uniq_sorted.bam
```

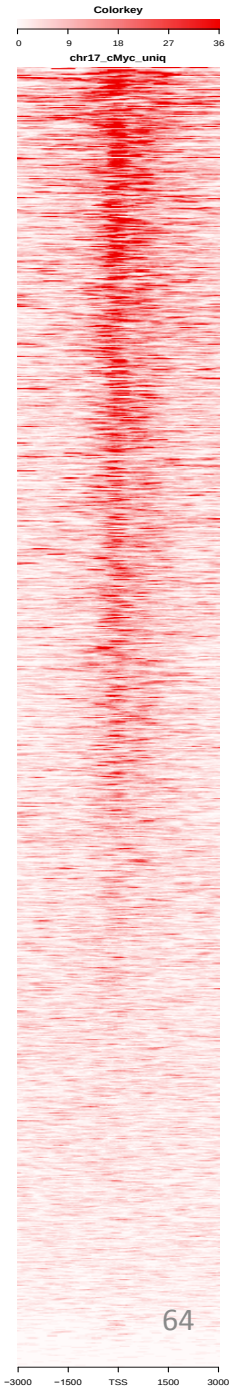
以上の作業を、Input、Oct4、cMycに対してやってみてください

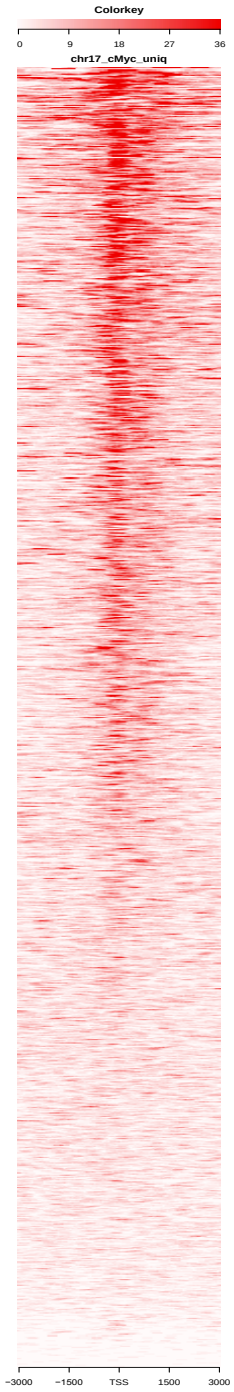
注
前の計算が終わる前に次の計算をはじめないこと！

例えばTSS周辺のcMycの分布は？

- ngs.plot.r -G hg19 -R tss -C chr17_cMyc_uniq_sorted.bam -O chr17_cMyc_uniq.tss -T chr17_cMyc_uniq -L 3000 -FL 300

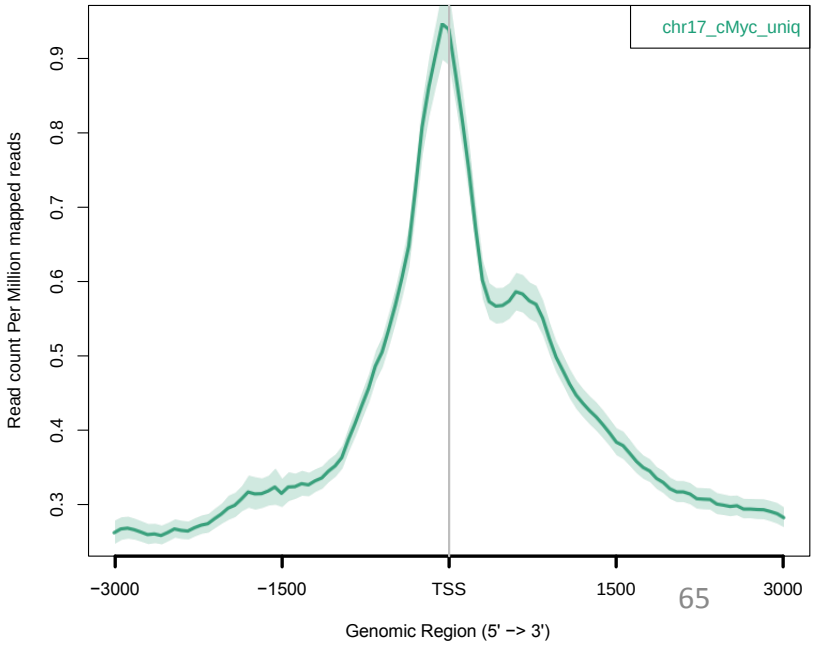






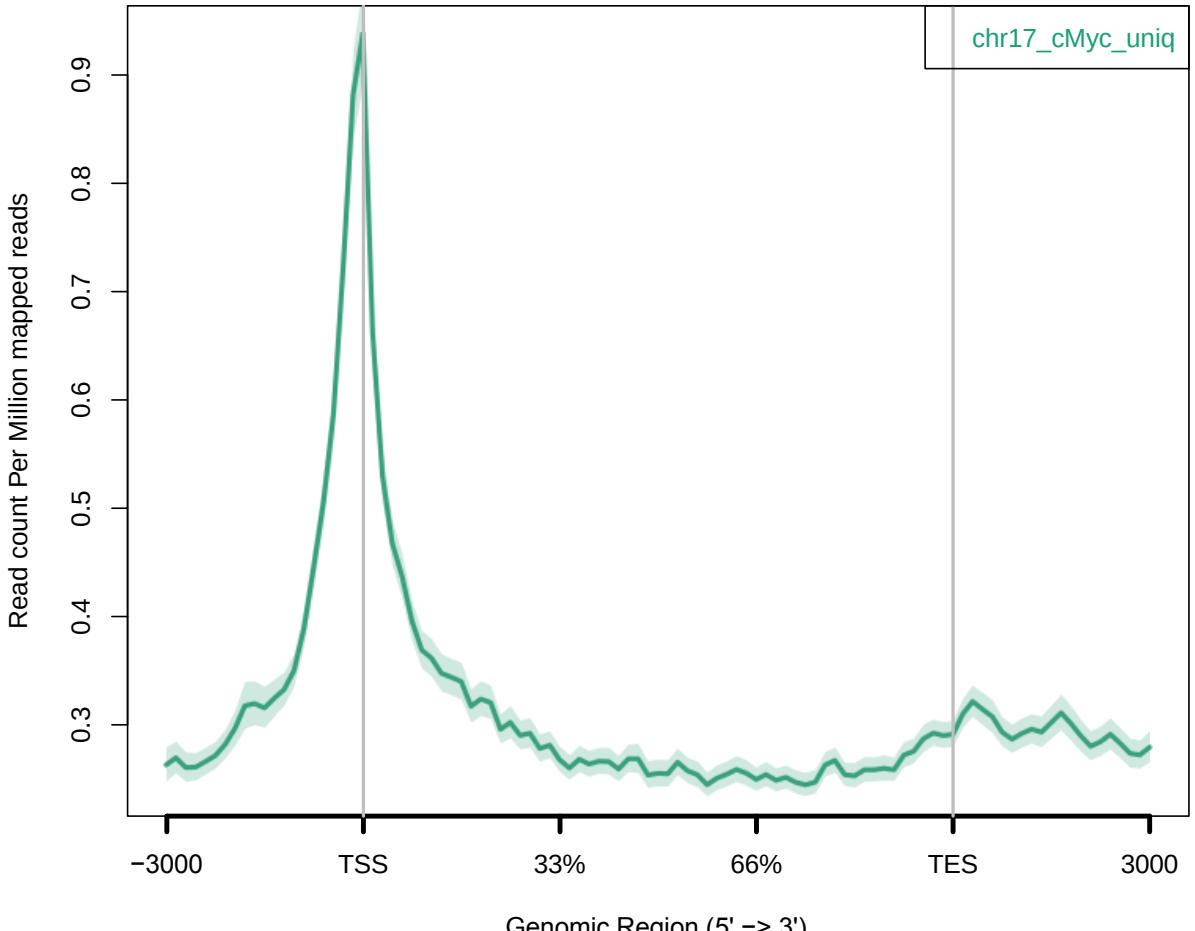
前スライドの説明

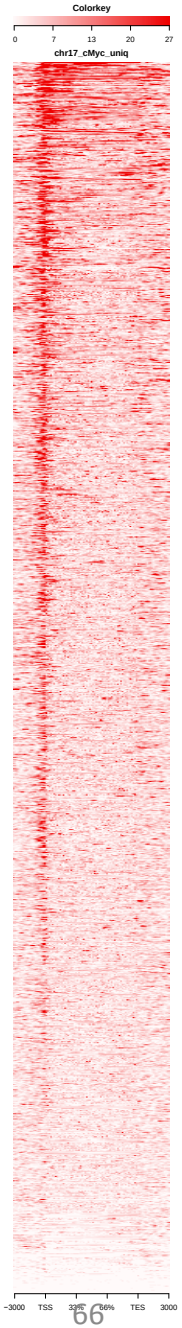
各行が観察している転写開始点に対応している。転写開始点周辺のタグの密度を赤ベースのheatmap表示したものが左図である。イメージとしては、heatmapを圧縮して、各行の平均とばらつきをとった図が右下に対応している。薄い緑がばらつきを表す。スライド68が分かりやいので参照のこと。



cMycのGenebodyなら？

- ngs.plot.r -G hg19 -R genebody -C chr17_cMyc_uniq_sorted.bam -O chr17_cMyc_uniq.genebody -T chr17_cMyc_uniq -L 3000 -FL 300





コントロールはどうする？

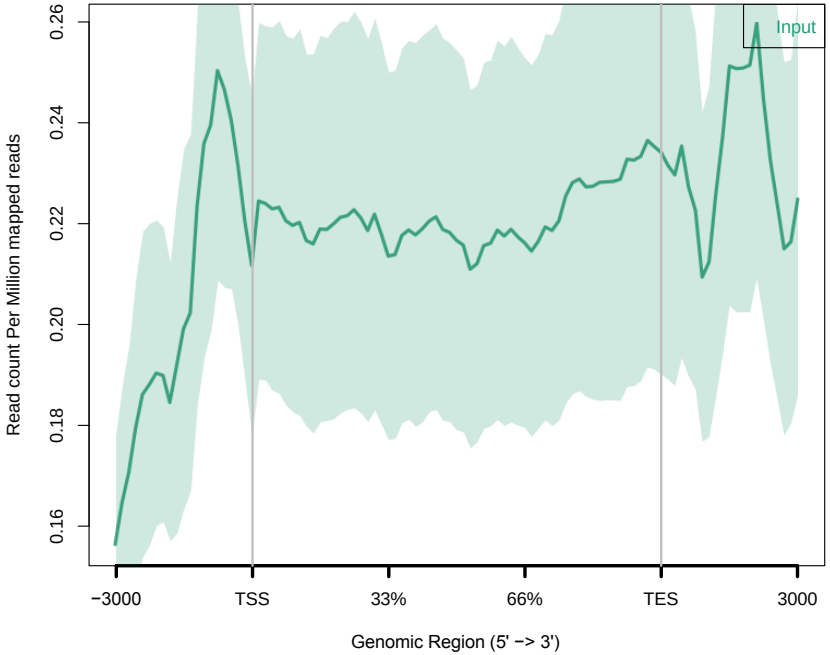
- 論文でよく見かけるChIP-seqデータのリプレゼンテーションで、よく、コントロールがないものがあります(紙面の都合で削る場合も?)

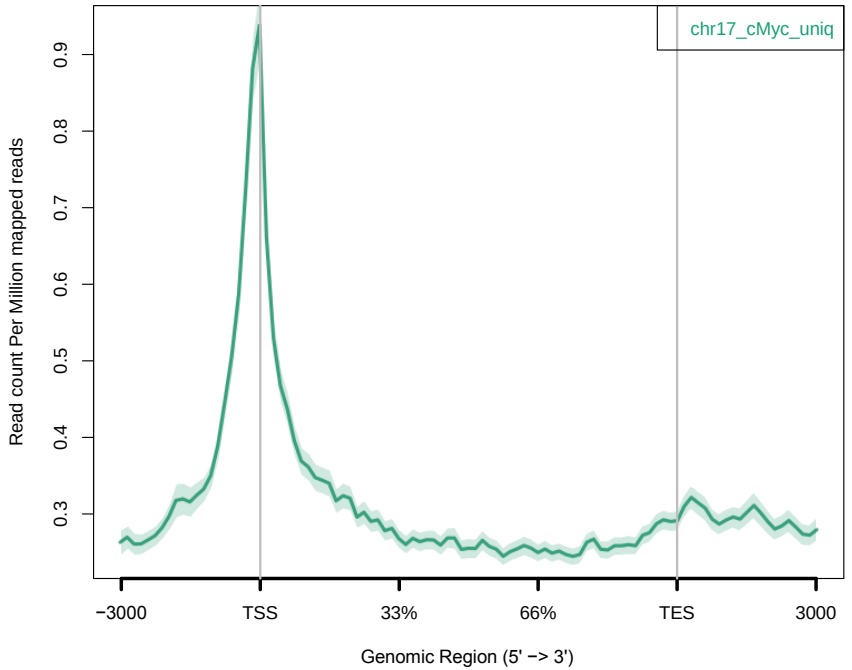
- 読み手にとっても重要な情報なので、きちんと示すことを推奨

- 今回の場合は、Input readsを使ってみる

Input genebody (てんでバラバラ)

注意すべき点は2点。ひとつが、ばらつきの大きさ。ふたつめが、縦軸のスコア。縦軸の頂点のスコアをみると、Input(左)とcMyc(右)で3倍以上の差があることが分かる。複数サンプルの重ね合わせに関してはスライド108-109を参照のこと。

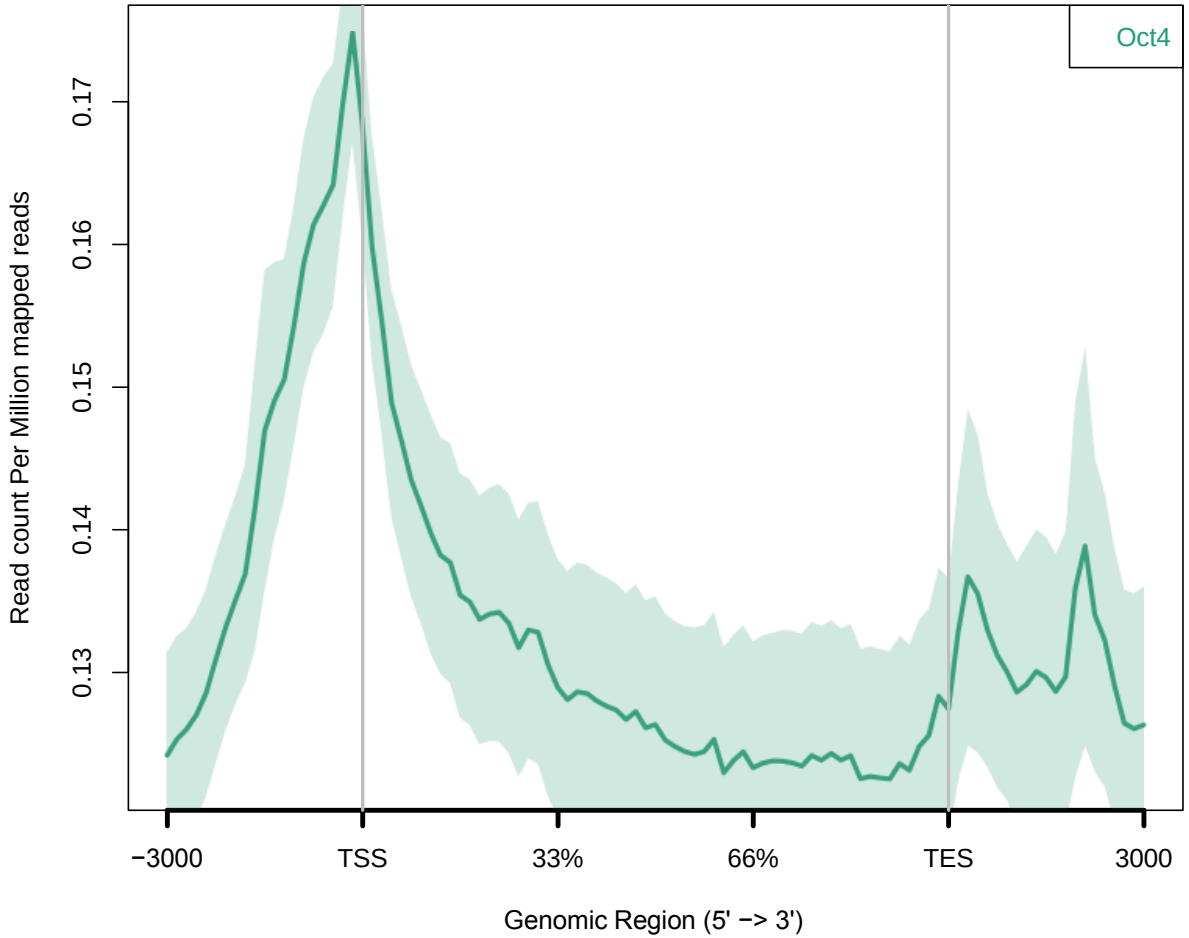




他のやつもやってみてください

- 例えばOct4のgenebodyはこんな感じです

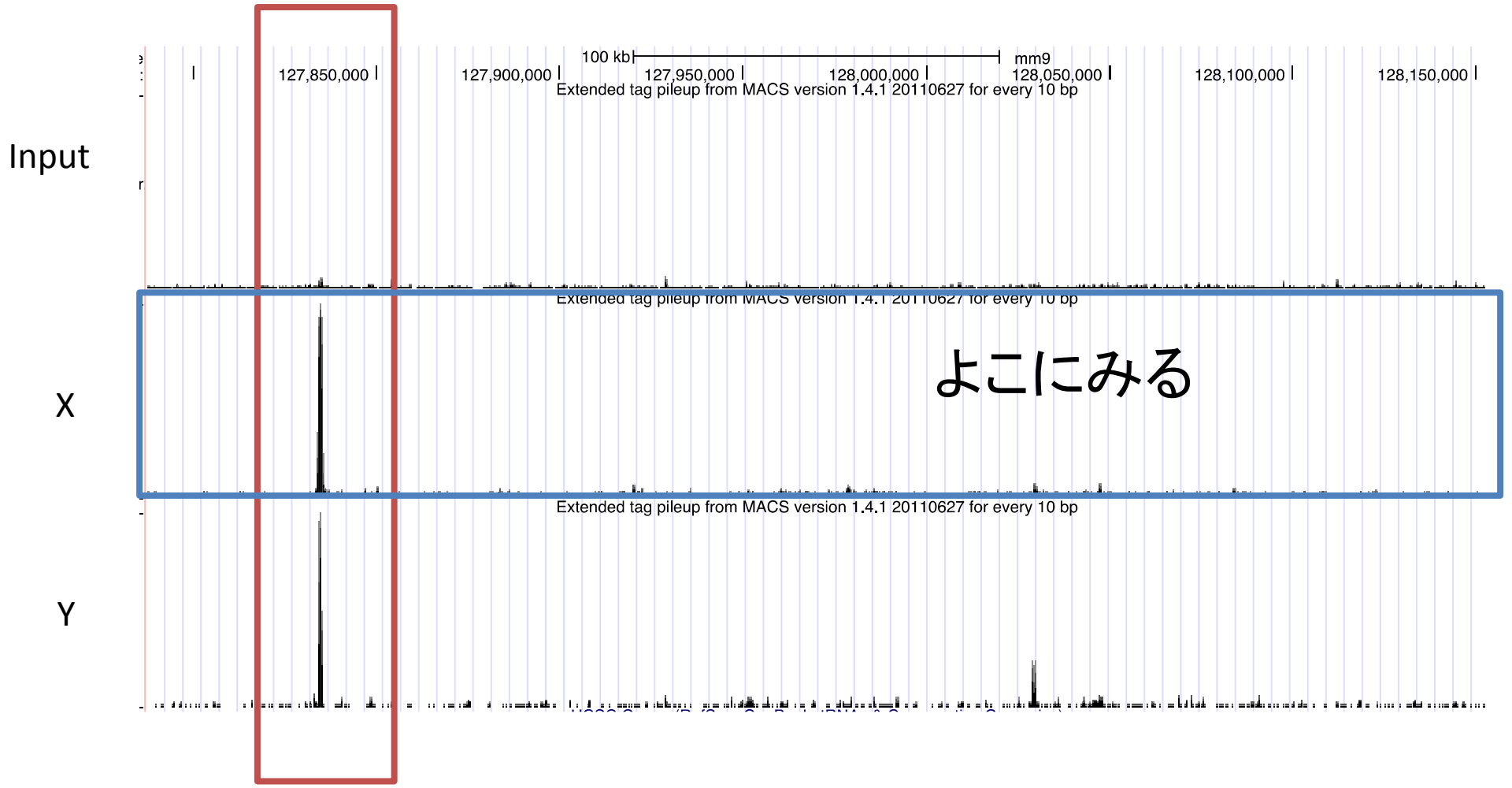
(正解コマンドは全てコマンドリストに記載)



ChIP-seq解析の流れ

- データの取得とクオリティチェック
- ゲノムへのマッピング
- 基本的な特徴付け
- エンリッチメントの定義 (Peak Call)

ピークコーリング


Input
X
Y
100 kb
mm9
127,850,000
127,900,000
127,950,000
128,000,000
128,050,000
128,100,000
128,150,000
Extended tag pileup from MACS version 1.4.1 20110627 for every 10 bp
Extended tag pileup from MACS version 1.4.1 20110627 for every 10 bp
Extended tag pileup from MACS version 1.4.1 20110627 for every 10 bp
よこにみる
たてにみる

よく使われているソフトウェア

- MACS14, Homerなど

MACS

Model-based Analysis for ChIP-Seq

Readme
Install
Download
Contributions
FAQ
ChangeLog

About

Next generation parallel sequencing technologies made chromatin immunoprecipitation followed by sequencing (ChIP-Seq) a popular strategy to study genome-wide protein-DNA interactions, while creating challenges for analysis algorithms. We present Model-based Analysis of ChIP-Seq (MACS) on short reads sequencers such as Genome Analyzer (Illumina / Solexa). MACS empirically models the length of the sequenced ChIP fragments, which tends to be shorter than sonication or library construction size estimates, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome sequence, allowing for more sensitive and robust prediction. MACS compares favorably to existing ChIP-Seq peak-finding algorithms, is publicly available c

Now, the newest version is version 1.4.2

Author

MACS is written by Yong Zhang and Tao Liu from Xiaol

Source Code

On Github

HOMER (v4.6, 3-29-2014)

Software for motif discovery and next generation sequencing analysis

HOMER (Hypergeometric Optimization of Motif EnRichment) is a suite of tools for Motif Discovery and next-gen sequencing analysis. It is a collection of command line programs for unix-style operating systems written in Perl and C++. HOMER was primarily written as a *de novo* motif discovery algorithm and is well suited for finding 8-20 bp motifs in large scale genomics data. HOMER contains many useful tools for analyzing ChIP-Seq, GRO-Seq, RNA-Seq, DNase-Seq, Hi-C and numerous other types of functional genomics sequencing data sets.

News

(3-29-2014) New version v4.6. Better super enhancer code, plus lots of other minor upgrades.

(1-27-2014) New version v4.5 with updated genome packages too - Last version did not correctly assign priority assignments in annotations (i.e. TSS > exons > introns > intergenic was not honored in last version - fixed now). Problems with some of the update scripts too. Website updates are ongoing.

(1-23-2014) Updated All Organism Packages to v1.1 - latest packages lacked the *org2*gene.tsv file (i.e. human2gene.tsv). Users that updated from the older version of HOMER probably didn't notice since the old files would have still been there - new users probably got an error.

(1-15-2014) Welcome to homer.salk.edu - new host for HOMER!

(1-15-2014) Lots of new documentation. More new documentation will probably be added to the site over the next week or so, and new PDF's will be created once that gets in good shape.

MACS14を使ってみる

~~$nohup macs14 –t [IP.bam] –name=[結果ファイルの名前] –c [Input.bam] –f [ファイルフォーマットの指定] –g [ゲノム] –wig&~~

```
$ nohup macs14 -t chr17_Oct4_uniq.bam --name=chr17_Oct4 -c chr17_Input_uniq.bam -f BAM -g hs --wig&
```

出てくるデータの種類

ファイルがたくさん出てきて分かりにくかったので、ここで全て明記します。Oct4のピークをコールすると、コマンドを実行したディレクトリに以下のものができます。

chr17_Oct4_MACS_wiggle (wigファイルを含む新しいディレクトリ)
chr17_Oct4_model.r
chr17_Oct4_negative_peaks.xls
chr17_Oct4_peaks.bed
chr17_Oct4_peaks.xls (後で使う)
chr17_Oct4_summits.bed (後で使う)

cMycでピークをコールすると、上記のファイルのOct4の部分がcMycにかわったファイル群が出てきます

chr17_Oct4_peaks.xls

BioLinuxのLibreOfficeで開くには、タブ区切りにチェック、コロン•空白区切りのチェックをはずしてください

エクセルなら普通にクリックで開けます

```
# This file is generated by MACS version 1.4.2 20120305
# ARGUMENTS LIST:
# name = chr17_Oct4
# format = BAM
# ChIP-seq file = chr17_Oct4_uniq.bam
# control file = chr17_Input_uniq.bam
# effective genome size = 2.70e+09
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Large dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps

# tag size is determined as 36 bps
# total tags in treatment: 921098
# tags after filtering in treatment: 869959
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.06
# total tags in control: 900922
# tags after filtering in control: 815189
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.10
# d = 63
chr     start   end     length  summit  tags    -10*log10(pvalue)       fold_enrichment FDR(%)
chr17   33      490     458     145     21      58.43   10.58   20.54
chr17   38626   39453   828     376     43      91.07   7.94    3.32
chr17   259596  260107  512     245     31      50.52   6.10    33.86
chr17   260483  261182  700     394     38      55.25   6.67    24.31
chr17   429501  430762  1262    471     66      143.12  7.38    0.33
chr17   432099  432809  711     296     29      56.47   7.90    23.16
chr17   491128  491315  188     130     15      77.31   11.90   6.82
chr17   499501  500827  1327    963     65      111.84  11.49   1.35
chr17   517039  517428  390     115     30      108.90  11.35   1.56
chr17   594217  595034  818     738     49      76.09   7.93    7.41
chr17   614406  615079  674     363     36      76.79   9.73    7.01
chr17   653439  654648  1210    770     69      75.45   5.08    7.59
chr17   654673  655139  467     298     24      63.61   10.20   15.01
```

迷子になったら

- Linuxで計算をして出力されたファイルの名前を検索してみる
- PCに存在するファイルを時系列で並べ替えて新しいファイルを探す
- 自分のいるディレクトリを検索して移動してみつける

などなど

全部が信頼できるわけではない！

- 統計値が与えられてはいるが、信頼できないピーク(例えば、再現性のないピーク)が、統計的に信頼できる、とコールされることはよくある

- これらを全部確実に除く方法はないが、いくつかの基準を設けることで(もしかすると)偽ピークを排除できうる

chr17_cMyc_peaks.xlsを例に

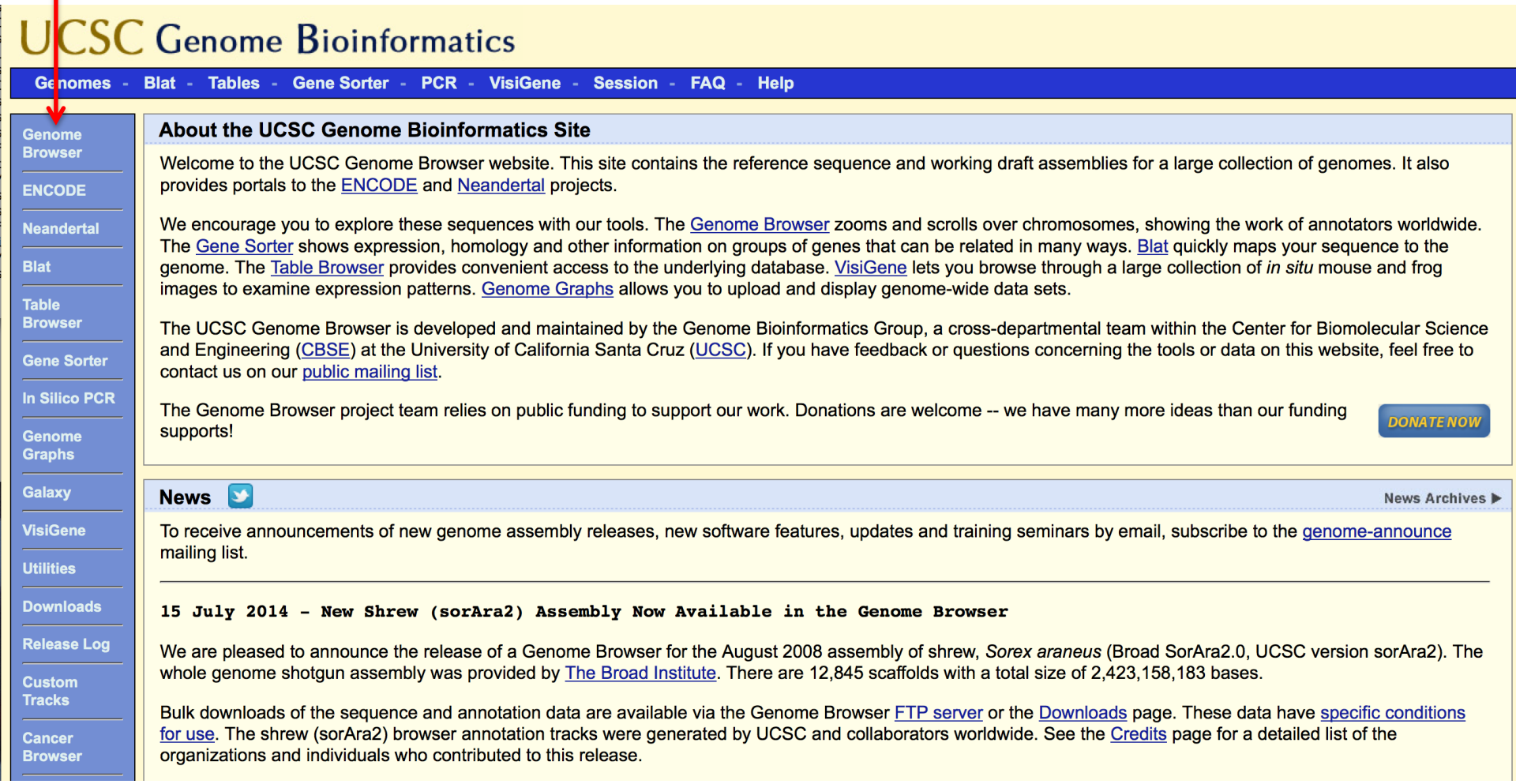
- ピークファイルのfold enrichmentとtagsを基準にピークを並べ替えてみる

Redundant rate in control: 0.10

d = 57

chr	start	end	length	summit	tags	#NAME?	fold_enrichm	FDR(%)
chr17	18472963	18473117	155	107	9	142.9	307.02	0.17
chr17	36452859	36452997	139	55	16	236.3	292.4	0
chr17	20558089	20558183	95	47	8	160.71	233.92	0.2
chr17	18747024	18747112	89	44	4	94.08	232.43	1.05
chr17	19015258	19015335	78	39	4	116.49	232.43	0.39
chr17	18585469	18585739	271	124	9	116.02	140.35	0.39
chr17	45579521	45579668	148	50	10	117.93	112.78	0.39
chr17	36631025	36631209	185	65	16	171.35	96.49	0.22
chr17	15492114	15492194	81	40	5	103.47	87.72	0.65
chr17	15752486	15752586	101	50	4	53.71	77.97	17.69
chr17	79961226	79962127	902	453	146	960.53	70.18	0
chr17	18684460	18684647	188	60	9	91.12	70.18	1.19
chr17	18312831	18312953	123	35	5	66.74	70.18	5.54
chr17	43528559	43528687	129	76	5	65.73	70.18	5.69
chr17	42147867	42148849	983	206	212	1203.24	70	0
chr17	27438437	27439184	748	551	121	644.02	66.81	0
chr17	61510584	61510837	254	144	76	634.37	55.56	0
chr17	62926001	62926137	137	100	9	92.09	52.63	1.11
chr17	15466705	15466827	123	86	4	50.63	52.63	22.22
chr17	5342003	5343033	1031	397	106	444.59	51.83	0
chr17	56326867	56327530	664	381	141	817.3	50.83	0
chr17	81009032	81009796	765	593	91	473.48	50.78	0
chr17	76210087	76210603	517	116	67	455.29	50.44	0
chr17	79935282	79935592	311	203	62	420.69	49.71	0
chr17	49230534	49232193	1660	391	363	1395.42	47.95	0

ゲノムブラウザベースのチェック
インターネットで「UCSC genome browser」と検索、genome browserをクリック



ゲノムブラウザベースのチェック

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

group	genome	assembly	position	search term	
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr1:53,971,906-54,199,877	enter position, gene symbol or search terms	submit

Click here to reset the browser user interface settings to their defaults.

track search | add custom tracks | track hubs | configure tracks and display

Human Genome Browser – hg19 assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Homo sapiens
(Graphic courtesy of CBSE)

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061:RH80175	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175, or chromosome bands

Searchの項目に、遺伝子名や染色体名、座標などを打ち込むと、好きな場所のマップを出せる

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.

group	genome	assembly	position	search term
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr1:53,971,906-54,199,877	GLIS1 submit

GLIS1 (Homo sapiens GLIS family zinc finger 1 (GLIS1), mRNA.)

Click here to reset the browser user interface settings to their defaults.

track search | add custom tracks | track hubs | configure tracks and display

Human Genome Browser – hg19 assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

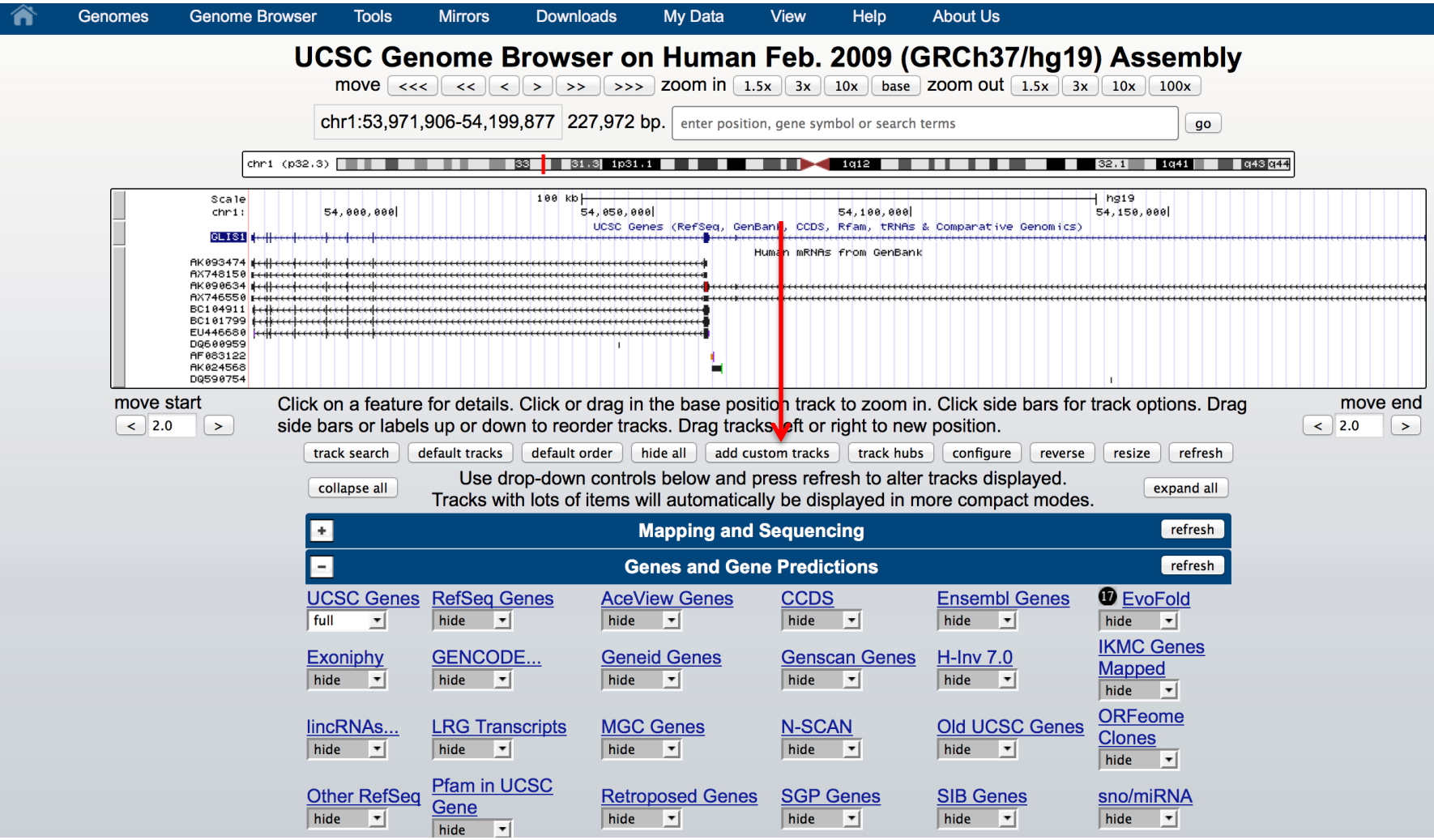
Homo sapiens
(Graphic courtesy of CBSE)

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061;RH80175 15q11;15q13 rs1042522;rs1800370	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175, or chromosome bands 15q11 to 15q13, or SNPs rs1042522 and rs1800370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.

ブラウザベースのチェック
add custom tracksからwigファイルをアップロード



カスタムトラックにwigファイルをアップロード

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Add Custom Tracks

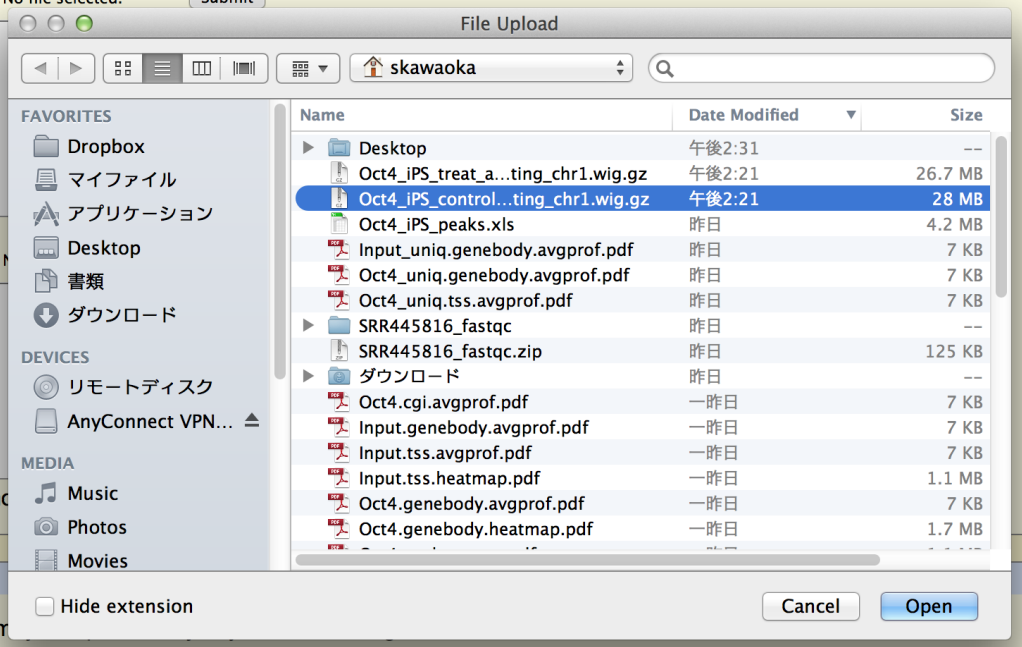
clade Mammal genome Human assembly Feb. 2009 (GRCh37/hg19)

Display your own data as custom annotation tracks in the browser. Data must be formatted in BED, bigBed, bedGraph, GFF, GTF, WIG, bigWig, MAF, BAM, BED detail, Personal Genome SNP, VCF, broadPeak, narrowPeak, or PSL formats. To configure the display, set track and browser line attributes as described in the User's Guide. Data in the bigBed, bigWig, BAM and VCF formats can be provided via only a URL or embedded in a track line in the box below. Publicly available custom tracks are listed here. Examples are here.

Paste URLs or data: Or upload: Browse... No file selected. Submit

Optional track documentation: Or upload: Browse...

Click here for an HTML document template that may be used for Genc



Loading Custom Tracks

An annotation data file in one of the supported custom track formats m

- (*Preferred*) Enter one or more URLs for custom tracks (one per line) in the data text box. The Genome Browser supports both the HTTP and FTP (passive-only) protocols.
- Click the "Browse" button directly above the URL/data text box, then choose a custom track file from your local computer, or type the pathname of the file into the "upload" text box

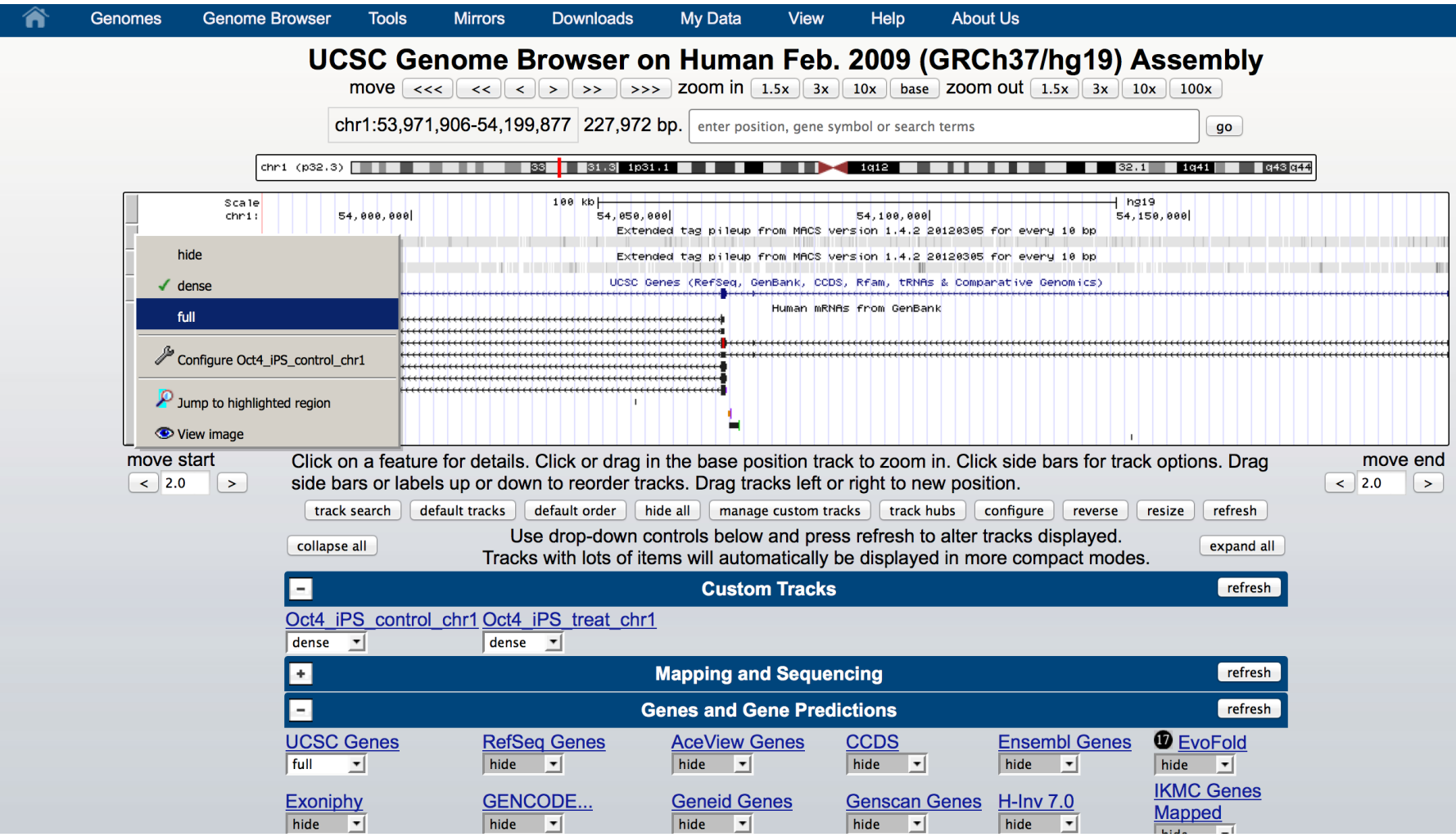
トラックしたいデータ

chr17_Oct4_MACS_wiggleフォルダの中に、controlとtreatというさらにふたつのフォルダがある。Controlに入っているのがInput、treatに入っているのが免疫沈降(IP)

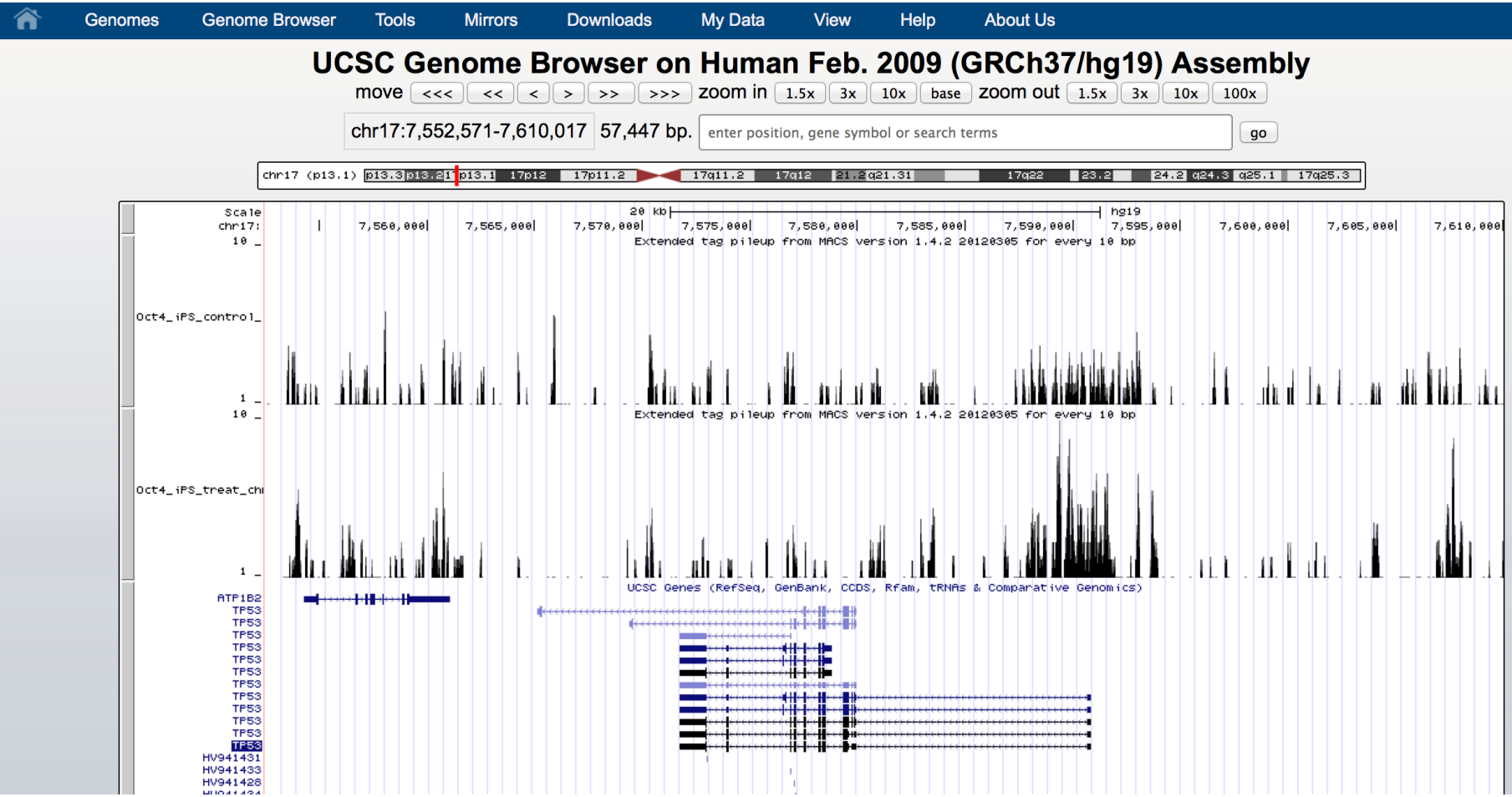
chr17_Oct4_control_afterfiting_chr17.wig.gz (これがInput)
chr17_Oct4_treat_afterfiting_chr17.wig.gz (これがIP)

必ずInputをいれること！

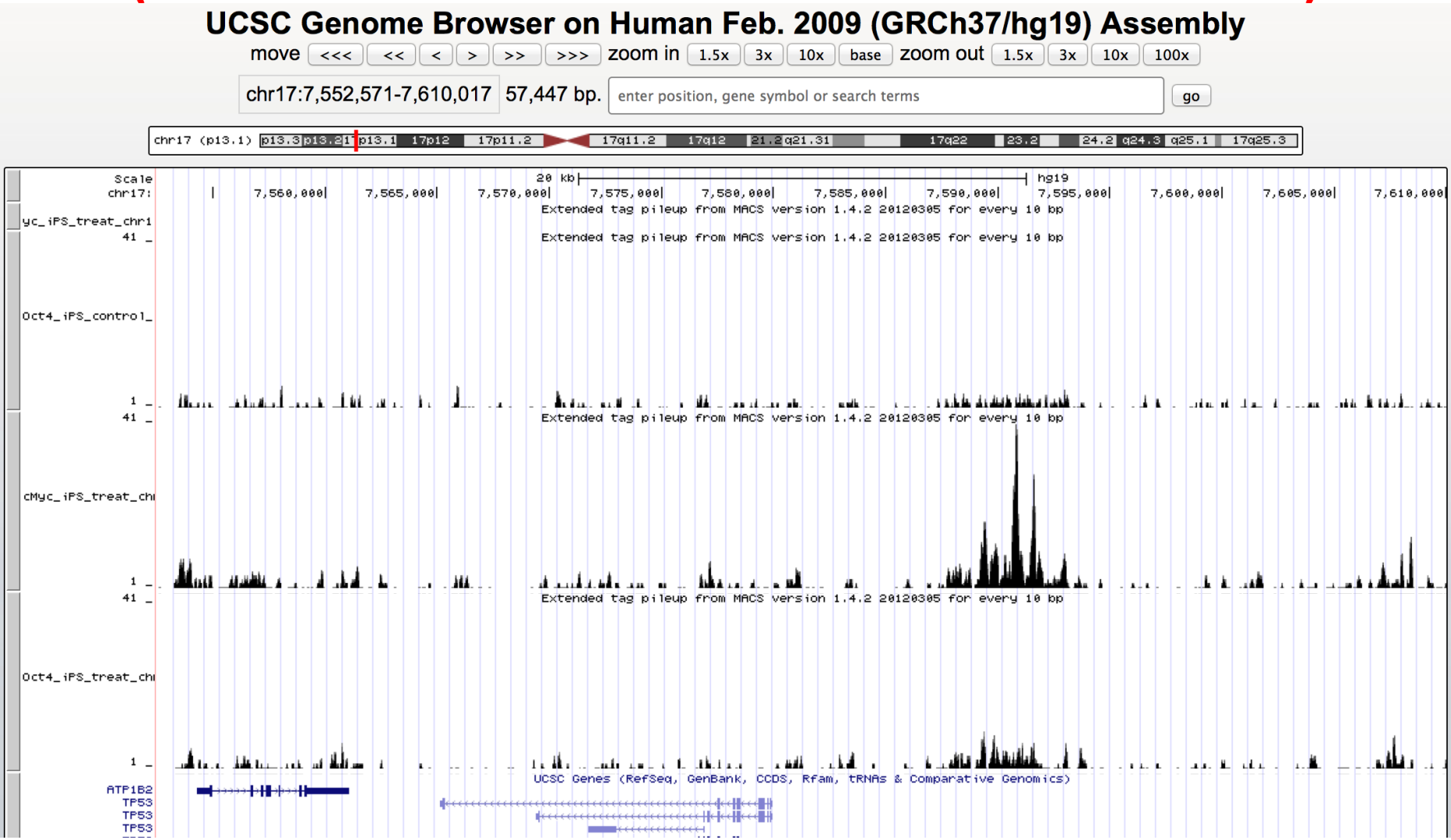
ゲノムブラウザでの「感覚的な」解析



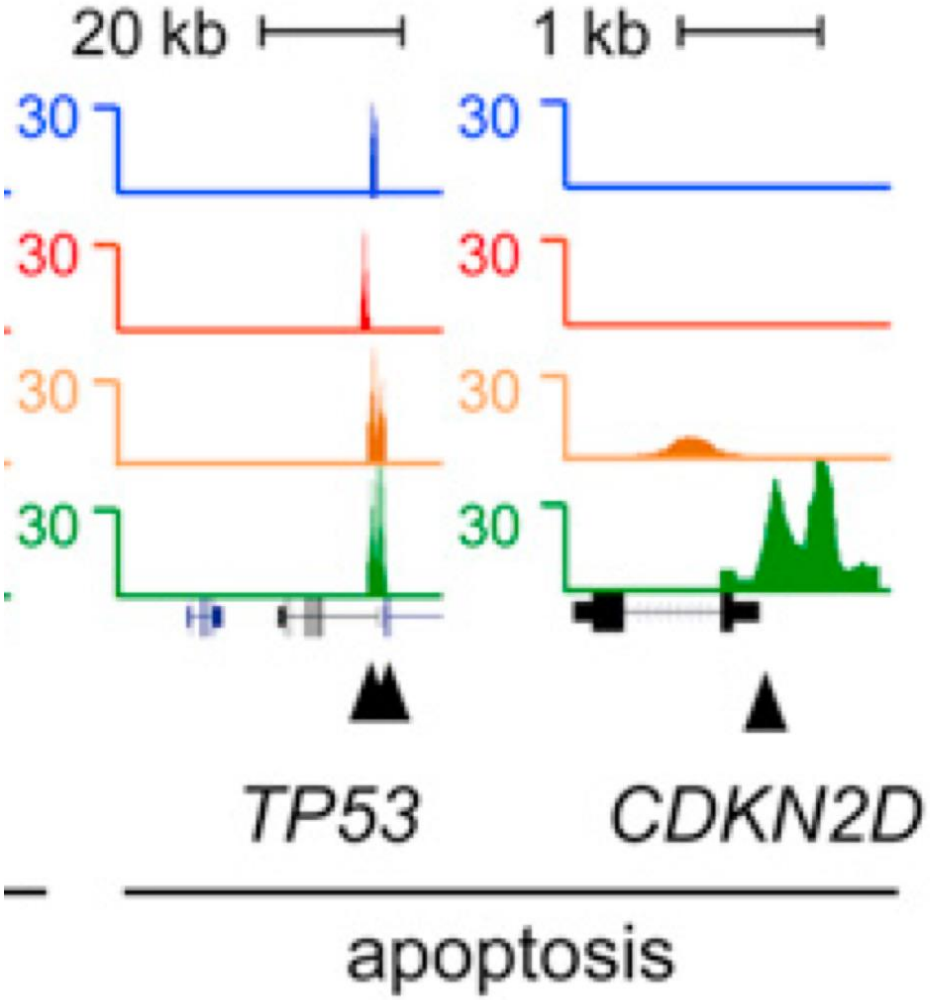
Chr17: TP53領域 Oct4
(searchのボックスにTP53と打ち込んで検索)



c-Mycも足してみる
(add custom trackからアップロード)



結構論文と違うような…?



Data representationは人それぞれ

- 自分の研究にとって重要な観察は、論文で示されているデータを鵜呑みにするのではなく、自分で解析をやりなおしてみて、感覚をつかむことが大事だと思います

Wigファイルの他に:
結構便利なbed file

- そもそも、GEOにもbedが落ちているケースが結構ある
- マッピングファイルがあれば簡単につくれる
- MACS14はbedデータもはきだしてくれる
- 何よりもサイズが軽いのでお手軽

Bedが落ちている場合

E-mail	zaret@upenn.edu
Phone	2155735813
Organization name	University of Pennsylvania School of Medicine
Department	Cell and Developmental Biology
Lab	Zaret lab
Street address	9-131, SCTR, 3400 Civic Center Boulevard
City	Philadelphia
State/province	PA
ZIP/Postal code	19104-5157
Country	USA

Platform ID	GPL10999
Series (1)	GSE36570 OSKM factors cooperatively engage chromatin to initiate reprogramming

Relations

SRA	SRX130060
BioSample	SAMN00828870

Supplementary file	Size	Download	File type/resource
SRX/SRX130/SRX130060		(ftp)	SRA Experiment
GSM896985_Oct4_48hrs.bed.gz	289.0 Mb	(ftp)(http)	BED
GSM896985_Oct4_48hrs.bw	302.0 Mb	(ftp)(http)	BW
GSM896985_Oct4_48hrs_peaks.bed.gz	476.0 Kb	(ftp)(http)	BED

Raw data provided as supplementary file

Processed data provided as supplementary file

Processed data is available on Series record

自分でBedファイルに変換

- bedtools bamtobed -i Oct4.bam > Oct4.bed

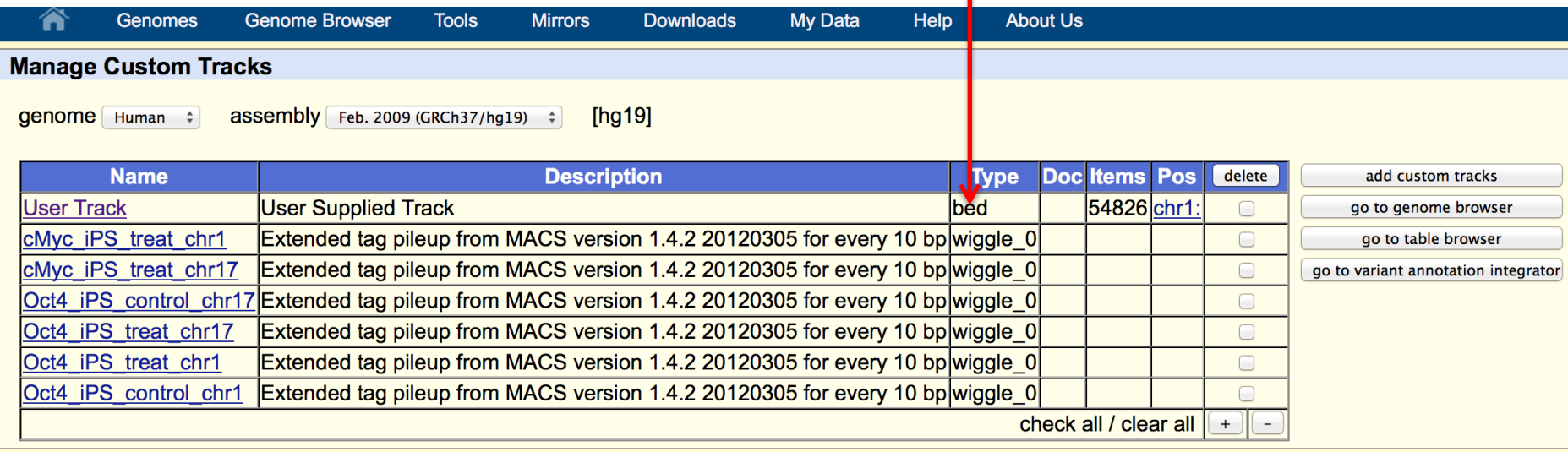
のようにすれば良いだけ

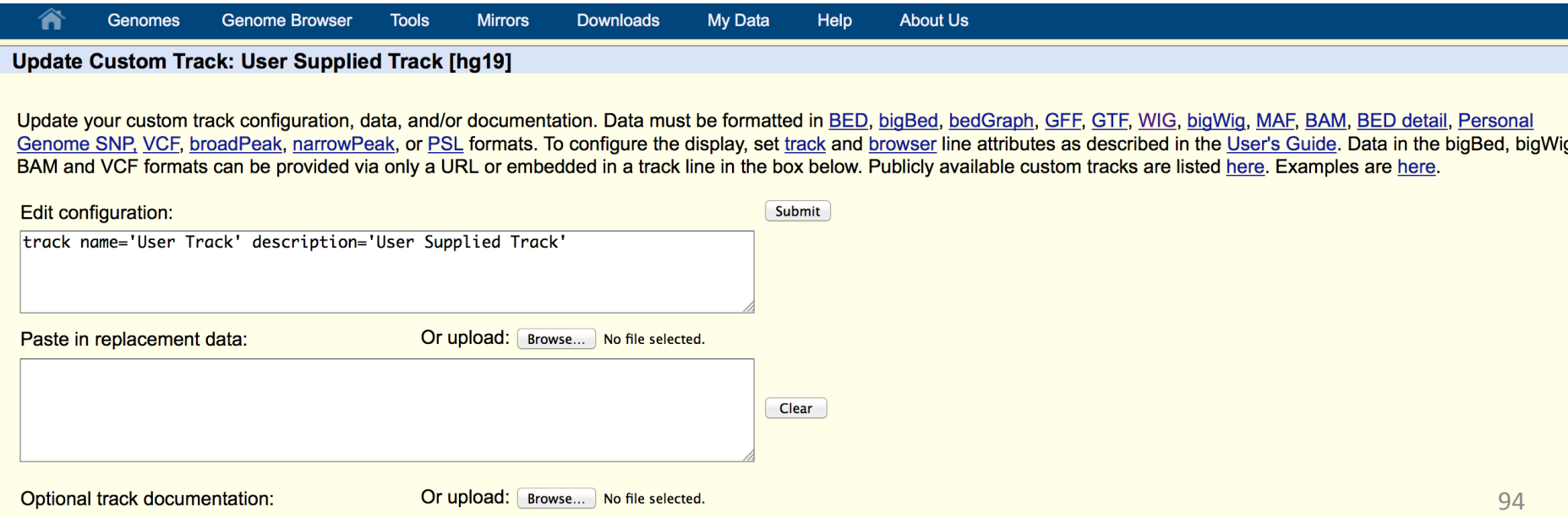
MACS14はbedも生成します

- ***peak.bed (chr17_Oct4_peaks.bed)

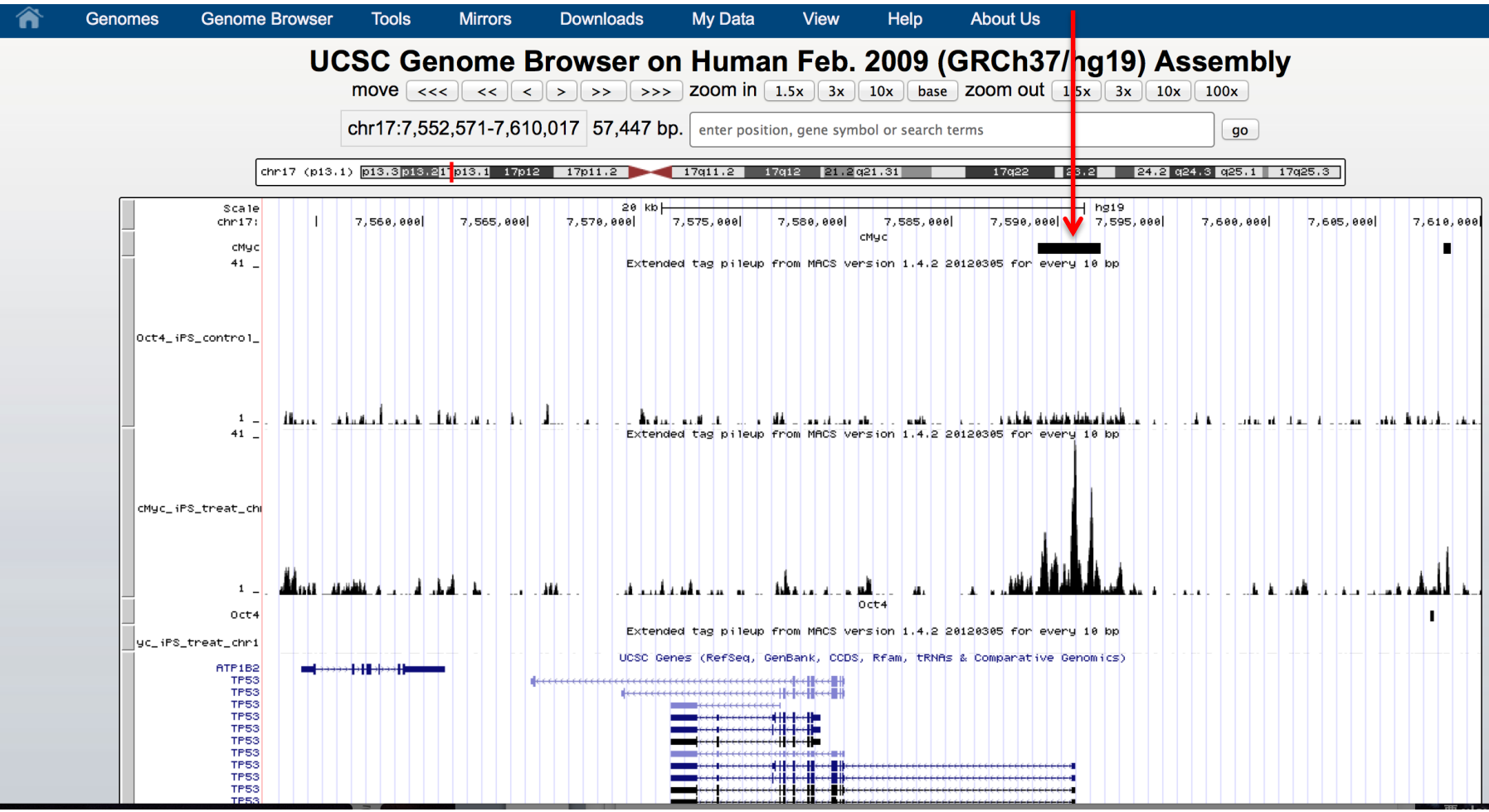
- ランキングして、ものすごく信頼性の高いものだけを選んだりしてみても良い (->その場合は、トップ100とかを恣意的に選んだファイルを作成するだけ)

ゲノムブラウザにあげてみる





ベッドファイルをみてみると

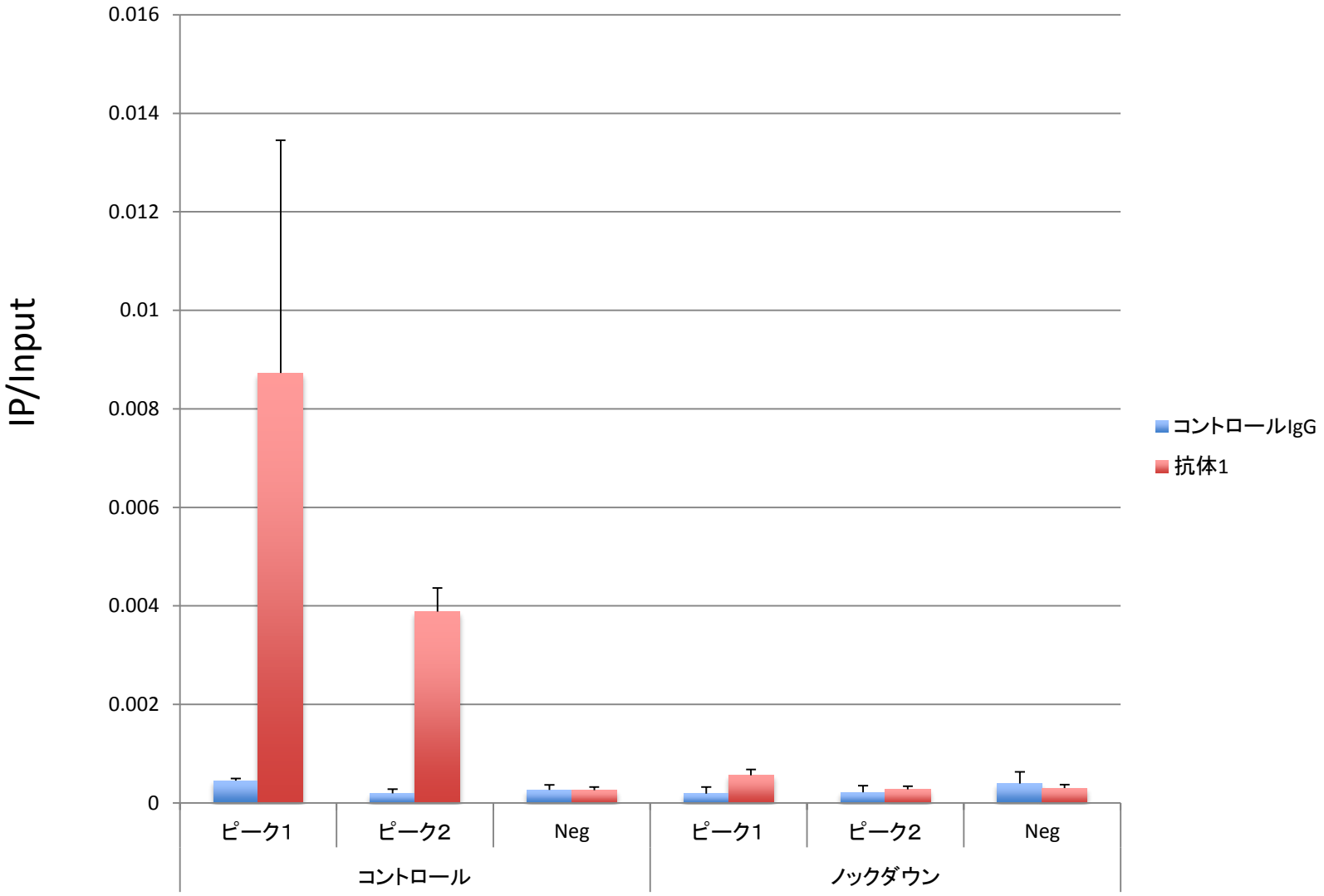


ピークとしてコールされた座標が
単にボックスとして表示される

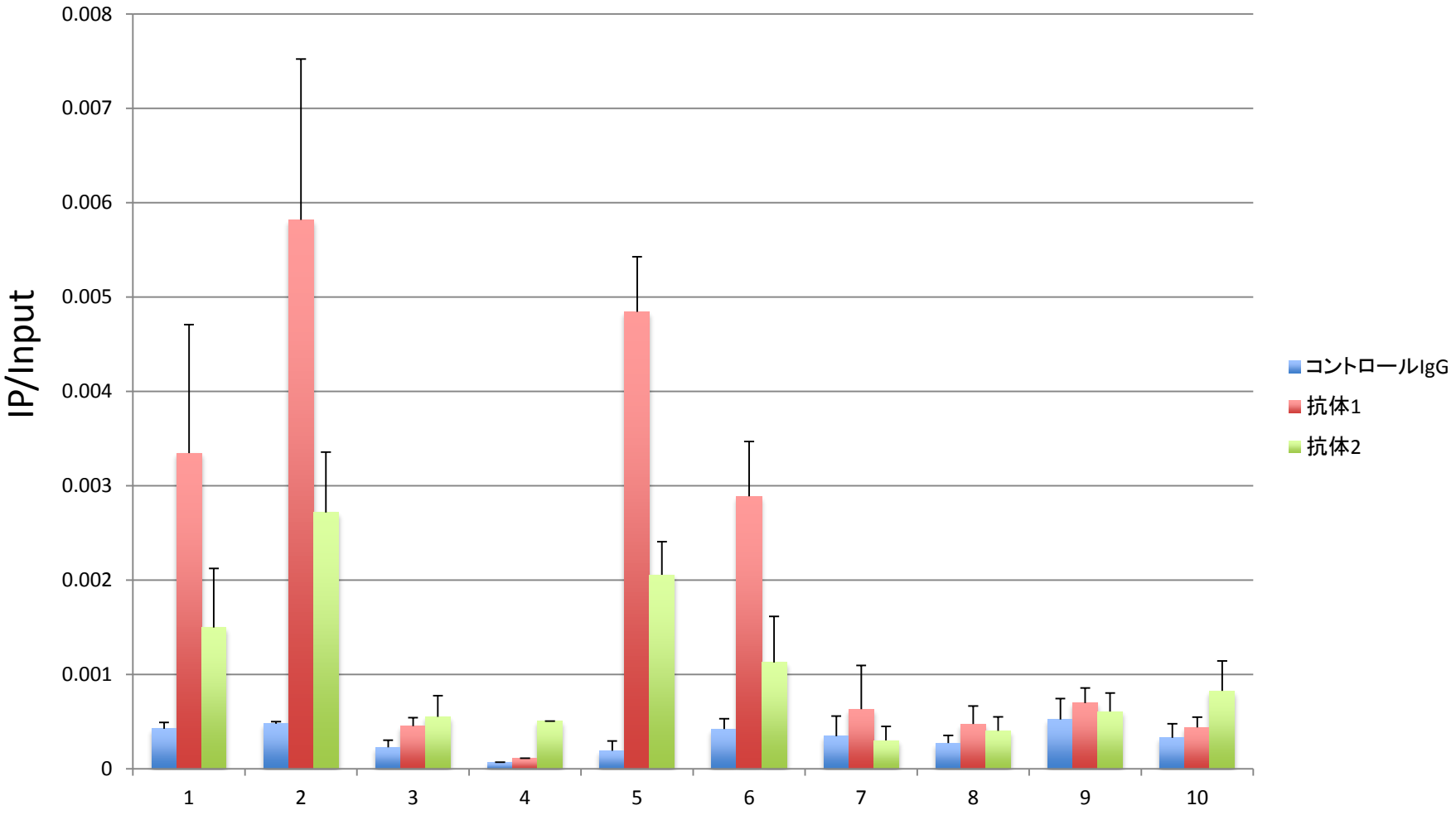
そのピーク本物？
実験屋からひとこと

- 実験で検証？ノックダウンして消えるか？
- 違う抗体？
- どの方法でも検出される？
- ENCODEプロジェクトでvalidateされた抗体は一応心配ない

ピーク検証の具体例 (ノックダウン)



ピーク検証の具体例 (違う抗体)



(さらにひとこと)
ゲノムワイドな解析とローカルな解析

- ゲノムワイドな解析→ざっくりとした全体像や、ある分子が動く、ゲノムワイドに通用するルールを抽出したい(ちょっと主観)

- ローカルな解析→特定の生命現象を説明するための「the」を探しているケースが多い(ちょっと主観)

ピークコーリングまとめ

- ピーク同定は以降の解析の基礎
- ピークコーリングで吐き出されるデータ群は役に立つ

少し発展編

- 共局在解析ー例えば、cMycとOct4はお互いにどんな関係にあるのだろうか？

- モチーフ解析ーピークとしてコールされた配列に何か特別な特徴はないだろうか？

Summit.bedとngsplotの組み合わせ

- Summitの復習 (ピークの頂点)

- Ngsplotは、基準点を自分で指定することができる

重ね合わせ

- 何を中心にするのか？ー例えばOct4
- 何を重ね合わせるのか？ー例えばcMyc

cMycのピークにOct4を重ねてみる

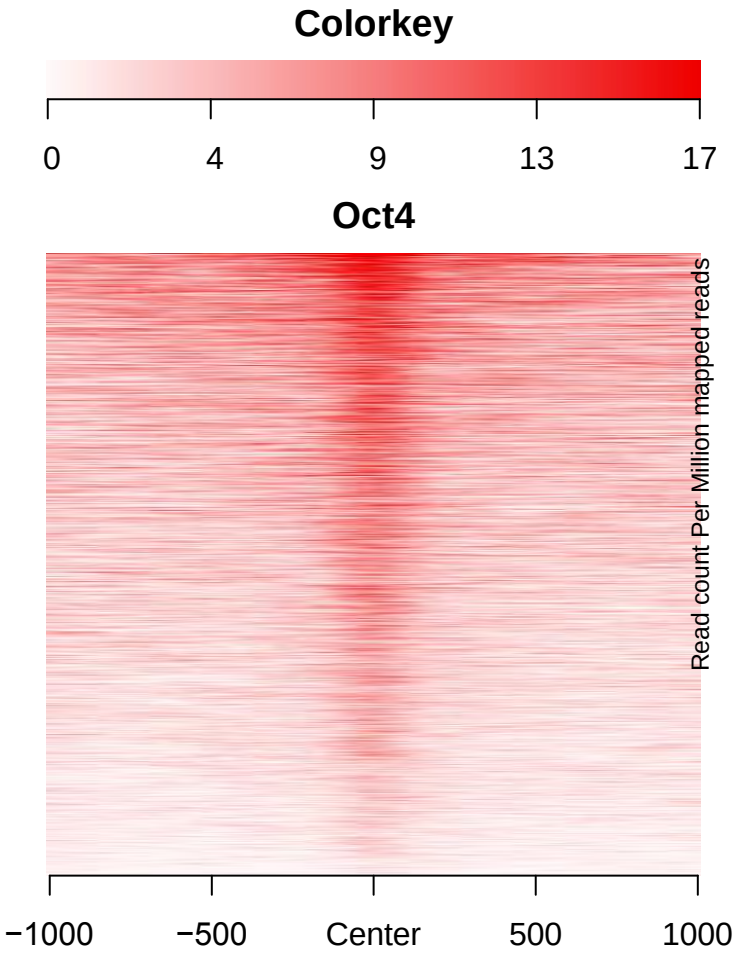
- 中心にするものーMACSでコールしたsummit.bedファイル
  (chr17_cMyc_summits.bed)

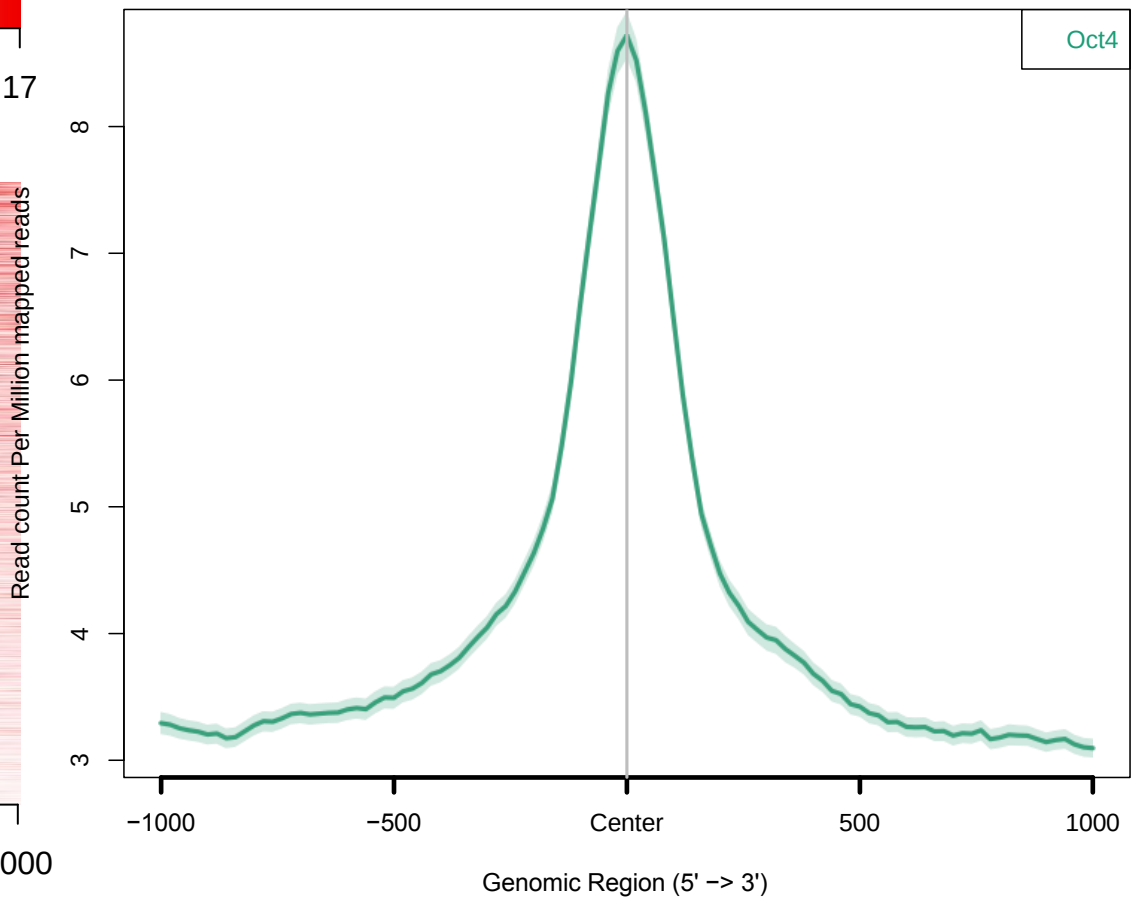
- 重ね合わせたいものー
  chr17_Oct4_uniq_sorted.bam

実際のコマンド例: ngs.plot.r

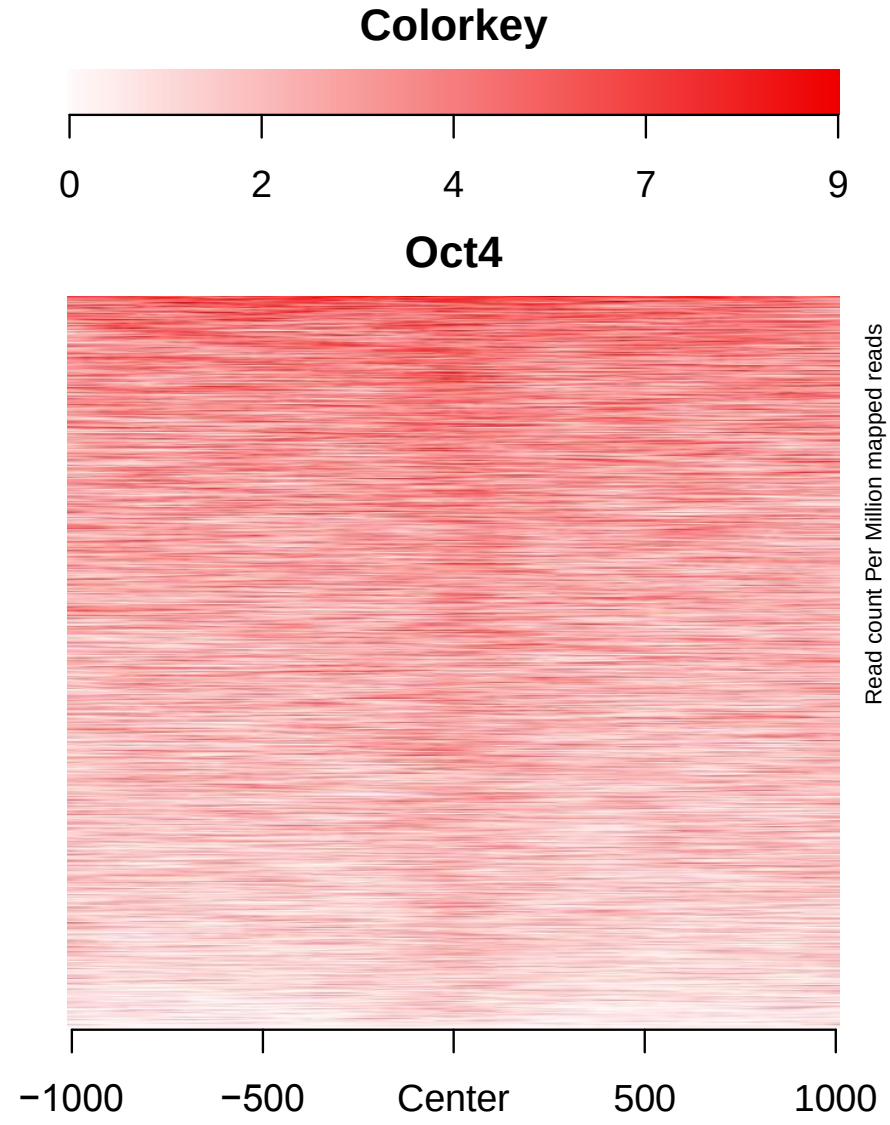
- ngs.plot.r -G hg19 -R bed -C chr17_Oct4_uniq_sorted.bam -E chr17_cMyc_summits.bed -O Oct4_cMyc_centered -T Oct4 -L 1000 -FL 150
- ngs.plot.r -G hg19 -R bed -C chr17_Input_uniq_sorted.bam -E chr17_cMyc_summits.bed -O Input_cMyc_centered -T Oct4 -L 1000 -FL 150

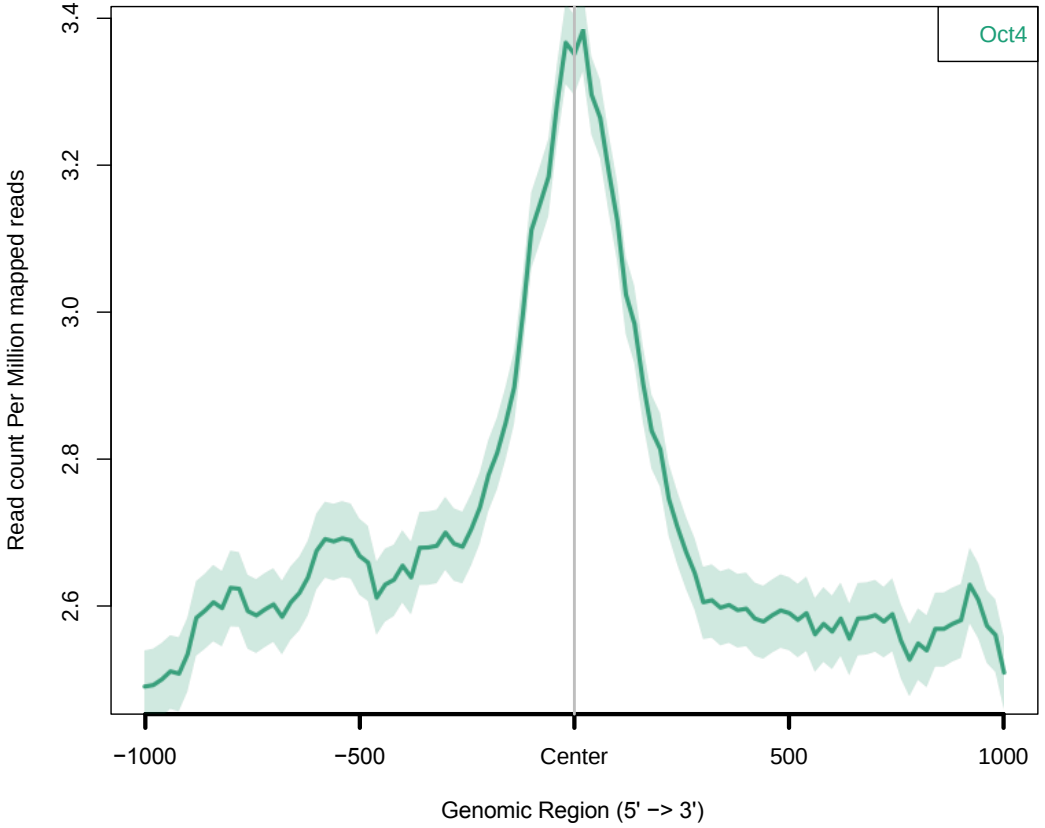
いろいろ組み合わせでやってみてください (例: cMycを中心にしたOct4)





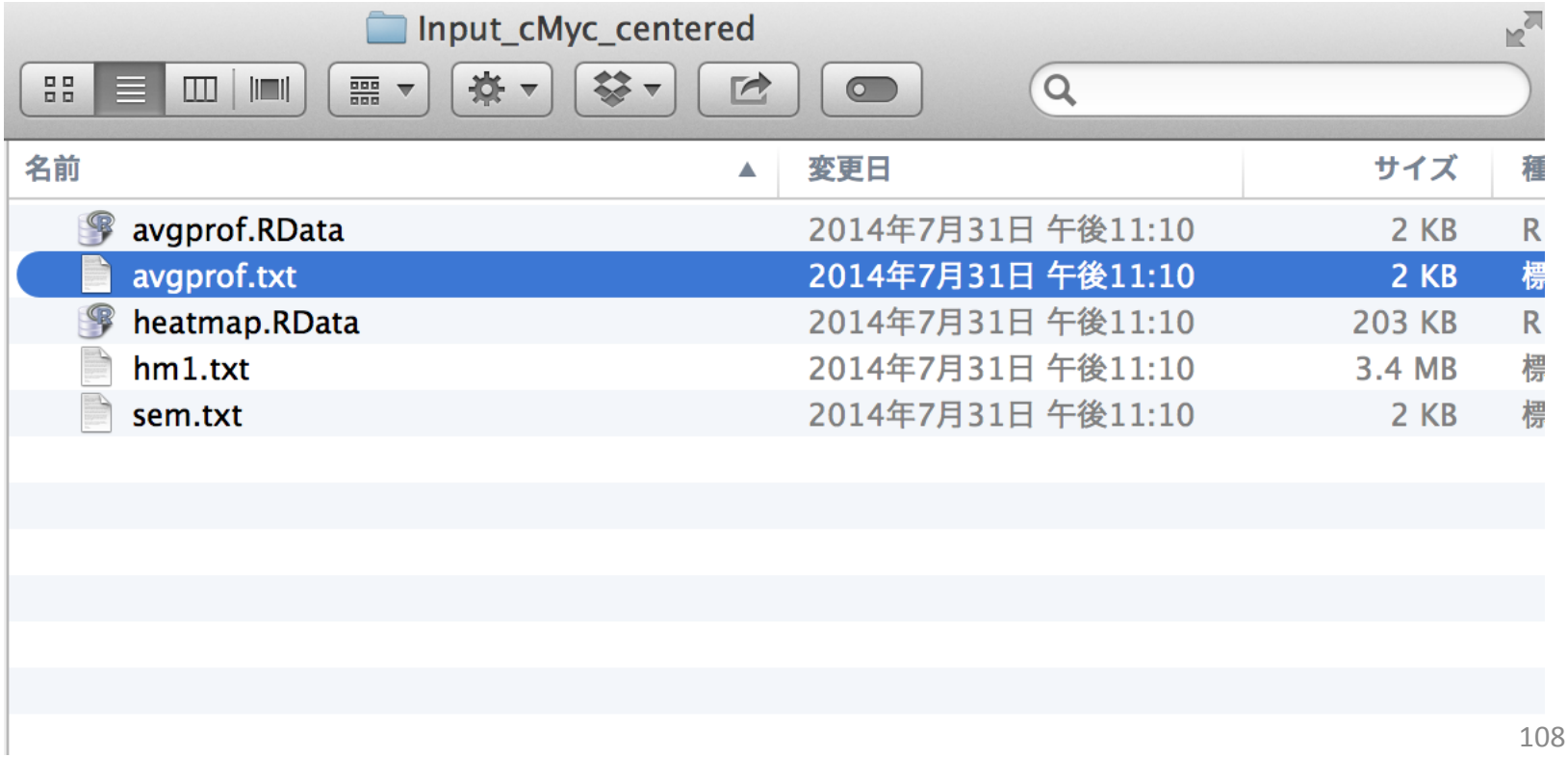
コントロールはどうする？Input!!





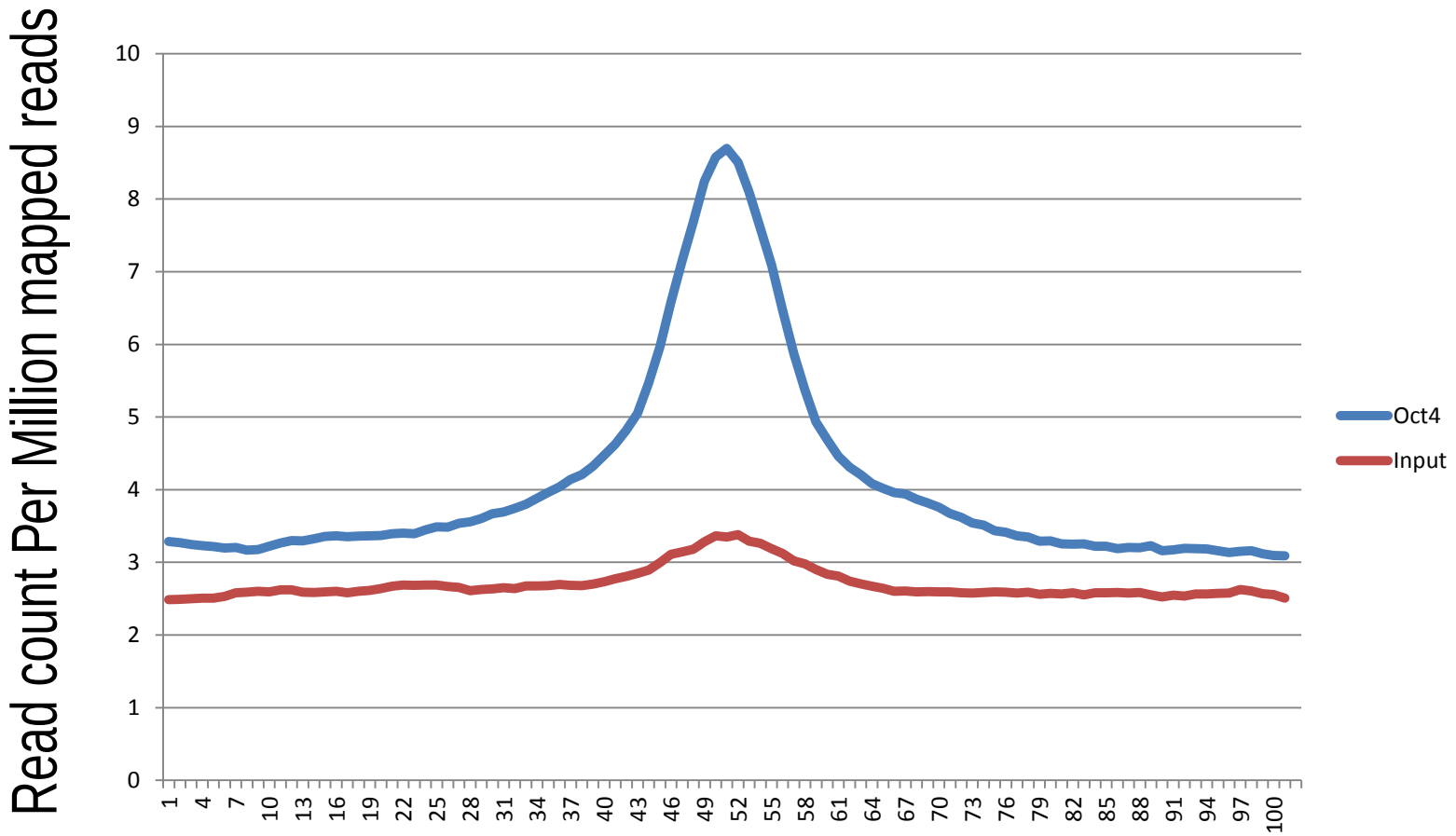
複数データをさらに重ね合わせ

- 生成されるフォルダのなかにあるavgprof.txtがなまデータなので、これをRやエクセルで読み込んで重ねれば良い



重ね合わせの実際

- 最後に重ねてみると。。。



何が分かったか？

- Oct4のまわりにはcMycがいる。では、cMycのまわりにはOct4がいるか、というのをどうやって調べれば良いか？

モチーフ解析

- 自分の調べたいタンパク質がエンリッチしているゲノム配列にどのような特徴があるかを調べたい

モチーフ解析をやってみよう

- 必要なもの: 解析したい部分のマルチファスタファイル
- 専用のソフトウェア (MEME、FIRE、などなど)
- MEMEは普通にオンラインで使えます

解析したい部分のmultifastaをつくる

- 配列がとりだせれば何でも良い
- MACSでコールしたピークファイルを使う
- blast+のblastdbcmdを使う(ネットで坊農さんが紹介していました)

blastdbcmdに必要なもの

- 染色体別のhg19のchr17の配列

- 上記をmakeblastdbでフォーマットする

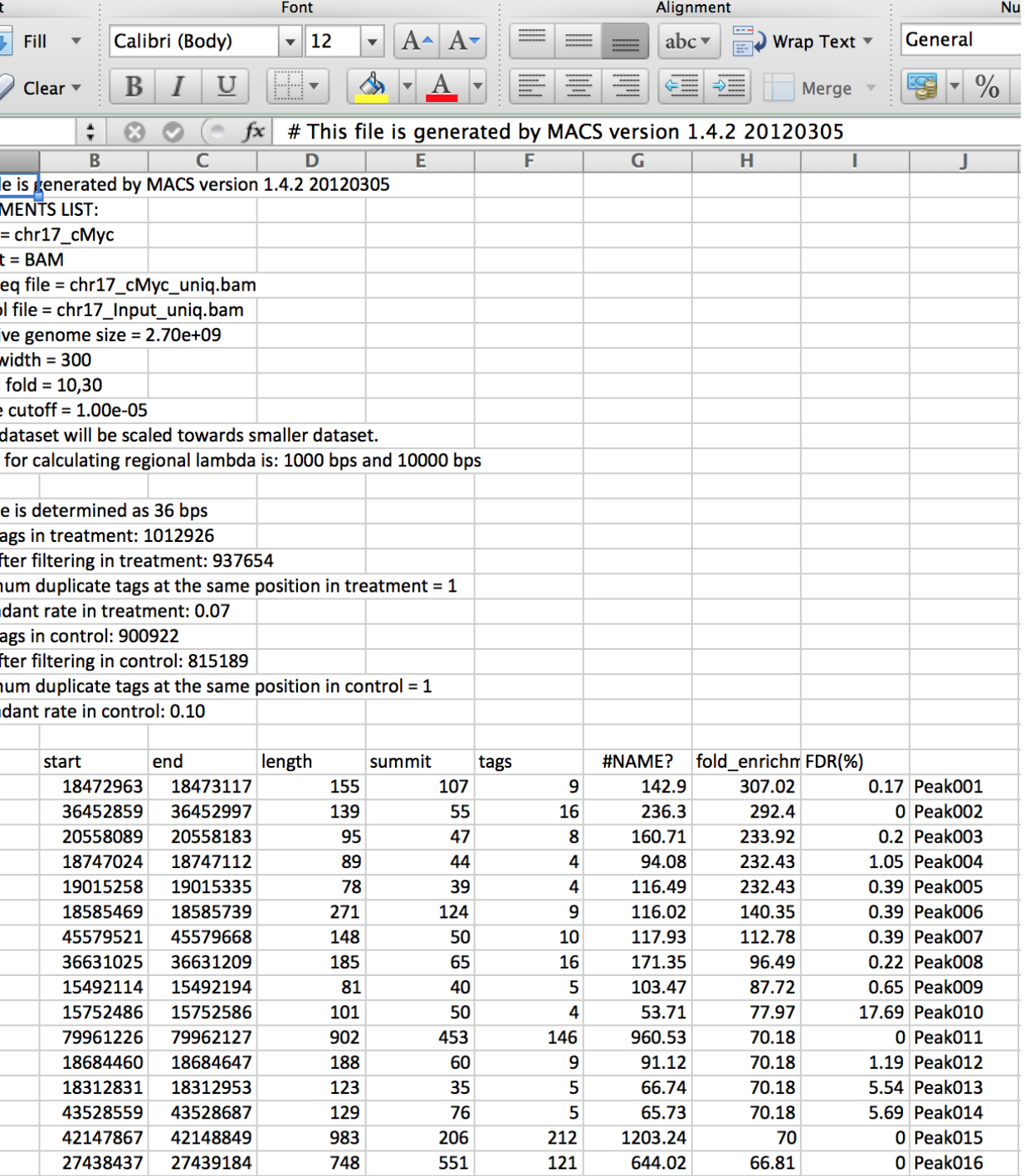
blastdbcmd

- makeblastdb -in chr17.fa -dbtype nucl -hash_index
- blastdbcmd -db chr17.fa -entry all -range 5000000-5000100
- これを自分が取ってきたピークに対して実行

考えること

- 何個まで使うのか？(つまり、どこまでをピークとして信用するのか)→とりあえずfold enrichment/tagでソートしてトップ500くらいをやってみる

- 取り出す領域はピーク全体でいいのか？→あとで考える



This file is generated by MACS version 1.4.2 20120305

	B	C	D	E	F	G	H	I	J
	le is generated by MACS version 1.4.2 20120305								
	MENTS LIST:								
	= chr17_cMyc								
	t = BAM								
	eq file = chr17_cMyc_uniq.bam								
	l file = chr17_Input_uniq.bam								
	ve genome size = 2.70e+09								
	width = 300								
	fold = 10,30								
	e cutoff = 1.00e-05								
	dataset will be scaled towards smaller dataset.								
	for calculating regional lambda is: 1000 bps and 10000 bps								
	e is determined as 36 bps								
	ags in treatment: 1012926								
	fter filtering in treatment: 937654								
	num duplicate tags at the same position in treatment = 1								
	dant rate in treatment: 0.07								
	ags in control: 900922								
	fter filtering in control: 815189								
	num duplicate tags at the same position in control = 1								
	dant rate in control: 0.10								
	start	end	length	summit	tags	#NAME?	fold_enrichm	FDR(%)	
	18472963	18473117	155	107	9	142.9	307.02	0.17	Peak001
	36452859	36452997	139	55	16	236.3	292.4	0	Peak002
	20558089	20558183	95	47	8	160.71	233.92	0.2	Peak003
	18747024	18747112	89	44	4	94.08	232.43	1.05	Peak004
	19015258	19015335	78	39	4	116.49	232.43	0.39	Peak005
	18585469	18585739	271	124	9	116.02	140.35	0.39	Peak006
	45579521	45579668	148	50	10	117.93	112.78	0.39	Peak007
	36631025	36631209	185	65	16	171.35	96.49	0.22	Peak008
	15492114	15492194	81	40	5	103.47	87.72	0.65	Peak009
	15752486	15752586	101	50	4	53.71	77.97	17.69	Peak010
	79961226	79962127	902	453	146	960.53	70.18	0	Peak011
	18684460	18684647	188	60	9	91.12	70.18	1.19	Peak012
	18312831	18312953	123	35	5	66.74	70.18	5.54	Peak013
	43528559	43528687	129	76	5	65.73	70.18	5.69	Peak014
	42147867	42148849	983	206	212	1203.24	70	0	Peak015
	27438437	27439184	748	551	121	644.02	66.81	0	Peak016
	61510584	61510837	254	144	76	634.37	55.56	0	Peak017

J列にファイル名を追加 する

tagや fold_enrichment で並びかえれば、上位XXX個のピーク、というふうにできる

K列に…

- =”blastdbcmd -db "&A25&".fa -entry all -range "&B25&"-"&C25&" > "&J25&".txt”

- こんな感じになる: blastdbcmd -db chr7.fa -entry all -range 1513589-1514175 > Peak001.txt

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
# This file is generated by MACS version 1.4.2 20120305															
# ARGUMENTS LIST:															
# name = chr17_cMyc															
# format = BAM															
# ChIP-seq file = chr17_cMyc_uniq.bam															
# control file = chr17_Input_uniq.bam															
# effective genome size = 2.70e+09															
# band width = 300															
# model fold = 10,30															
# pvalue cutoff = 1.00e-05															
# Large dataset will be scaled towards smaller dataset.															
# Range for calculating regional lambda is: 1000 bps and 10000 bps															
# tag size is determined as 36 bps															
# total tags in treatment: 1012926															
# tags after filtering in treatment: 937654															
# maximum duplicate tags at the same position in treatment = 1															
# Redundant rate in treatment: 0.07															
# total tags in control: 900922															
# tags after filtering in control: 815189															
# maximum duplicate tags at the same position in control = 1															
# Redundant rate in control: 0.10															
# d = 57															
chr	start	end	length	summit	tags	#NAME?	fold_enrichm	FDR(%)							
chr17	18472963	18473117	155	107	9	142.9	307.02	0.17	Peak001	blastdbcmd -db chr17.fa -entry all - range 18472963-18473117 > Peak001.txt					
chr17	36452859	36452997	139	55	16	236.3	292.4	0	Peak002						
chr17	20558089	20558183	95	47	8	160.71	233.92	0.2	Peak003						

K列をドラッグしてコピペ

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	# This file is generated by MACS version 1.4.2 20120305															
2	# ARGUMENTS LIST:															
3	# name = chr17_cMyc															
4	# format = BAM															
5	# ChIP-seq file = chr17_cMyc_uniq.bam															
6	# control file = chr17_Input_uniq.bam															
7	# effective genome size = 2.70e+09															
8	# band width = 300															
9	# model fold = 10,30															
10	# pvalue cutoff = 1.00e-05															
11	# Large dataset will be scaled towards smaller dataset.															
12	# Range for calculating regional lambda is: 1000 bps and 10000 bps															
13																
14	# tag size is determined as 36 bps															
15	# total tags in treatment: 1012926															
16	# tags after filtering in treatment: 937654															
17	# maximum duplicate tags at the same position in treatment = 1															
18	# Redundant rate in treatment: 0.07															
19	# total tags in control: 900922															
20	# tags after filtering in control: 815189															
21	# maximum duplicate tags at the same position in control = 1															
22	# Redundant rate in control: 0.10															
23	# d = 57															
24	chr	start	end	length	summit	tags	#NAME?	fold_enrichm	FDR(%)							
25	chr17	18472963	18473117	155	107	9	142.9	307.02	0.17	Peak0	blastdbcmd -db chr17.fa -entry all - range 18472963-18473117 > Peak001.txt					
26	chr17	36452859	36452997	139	55	16	236.3	292.4	0	Peak00	blastdbcmd -db chr17.fa -entry all - range 36452859-36452997 > Peak002.txt					
27	chr17	20558089	20558183	95	47	8	160.71	233.92	0.2	Peak003	blastdbcmd -db chr17.fa -entry all - range 20558089-20558183 > Peak003.txt					
28	chr17	18747024	18747112	89	44	4	94.08	232.43	1.05	Peak004	blastdbcmd -db chr17.fa -entry all - range 18747024-18747112 > Peak004.txt					
29	chr17	19015258	19015335	78	39	4	116.49	232.43	0.39	Peak005	blastdbcmd -db chr17.fa -entry all - range 19015258-19015335 > Peak005.txt					
30	chr17	18585469	18585739	271	124	9	116.02	140.35	0.39	Peak006	blastdbcmd -db chr17.fa -entry all - range 18585469-18585739 > Peak006.txt					
31	chr17	45579521	45579668	148	50	10	117.93	112.78	0.39	Peak007	blastdbcmd -db chr17.fa -entry all - range 45579521-45579668 > Peak007.txt					
32	chr17	36631025	36631209	185	65	16	171.35	96.49	0.22	Peak008	blastdbcmd -db chr17.fa -entry all - range 36631025-36631209 > Peak008.txt					
33	chr17	15492114	15492194	81	40	5	103.47	87.72	0.65	Peak009	blastdbcmd -db chr17.fa -entry all - range 15492114-15492194 > Peak009.txt					
34	chr17	15752486	15752586	101	50	4	53.71	77.97	17.69	Peak010	blastdbcmd -db chr17.fa -entry all - range 15752486-15752586 > Peak010.txt					
35	chr17	79961226	79962127	902	453	146	960.53	70.18	0	Peak011	blastdbcmd -db chr17.fa -entry all - range 79961226-79962127 > Peak011.txt					
36	chr17	18684460	18684647	188	60	9	91.12	70.18	1.19	Peak012	blastdbcmd -db chr17.fa -entry all - range 18684460-18684647 > Peak012.txt					
37	chr17	18312831	18312953	123	35	5	66.74	70.18	5.54	Peak013	blastdbcmd -db chr17.fa -entry all - range 18312831-18312953 > Peak013.txt					
38	chr17	43528559	43528687	129	76	5	65.73	70.18	5.69	Peak014	blastdbcmd -db chr17.fa -entry all - range 43528559-43528687 > Peak014.txt					
39	chr17	42147867	42148849	983	206	212	1203.24	70	0	Peak015	blastdbcmd -db chr17.fa -entry all - range 42147867-42148849 > Peak015.txt					
40	chr17	27438437	27439184	748	551	121	644.02	66.81	0	Peak016	blastdbcmd -db chr17.fa -entry all - range 27438437-27439184 > Peak016.txt					
41	chr17	61510584	61510837	254	144	76	634.37	55.56	0	Peak017	blastdbcmd -db chr17.fa -entry all - range 61510584-61510837 > Peak017.txt					

K列をコピーして、シェルスクリプトをつくる (emacsとかgeditとか何でも良い)

- vi extractseq.sh
- iを押して文章書き込みモード、編集終了はesc、セーブは:を押してからwq!
- エクセルからコピペ
- sh extractseq.sh

cat Peak* > peak_all.txt
によってファイルを結合→multifasta化

```
>gnl|BL_ORD_ID|0:67846234-67846991 chr6
AGTGAAGTCTGCACCACTGCCCTGCAGCCTGAATGACAGAGCTTCCAAATATATACATATATACACACACACATATTTCT
TTTAAAATATATATACTTCCCTCCAAACATATGTGTAAGCATGTATACATATATGCACATATGTATATGCTTGCATACAT
GTTTGCACATATGCATATGCATACATGCATGCATGTACAAATGTATATGCATGCTTACATGCACATATACATATGCACAT
ATAAATATATGCACATATGCATATGCATACATGCATGCATGTACAAATGTATATGCATGCTTACATGCACATATACATAT
GCACATATAAATATACGCACATATGCATATTCATGCATATATGTACATATGCAAGTGCATACATGCATATATGTATGCTT
ACATATGTGTGTGCATGCATGCACATATGCATATGCATACAGACATACATGTACATGTATATGCATGTACACATACATCC
ACACATGTGTATGCATGCATACATATATGTATGTATGTCTACGTGTGTATGTGTATATATGTGTACTTGCATATATGCAT
GCATATGTGTATGTATACACATATGCATGTATGTGTATGAATACATGTATGAATGCATGTGTATGTGTGTATACATATAT
GTGTGTATGTGTGTATGTCTATATGTTTATGTATGCATGTGTATATGTGTGTATGTGTGCATATGTGTATATATGTGTAT
GTATGTATACATATAAGTATATATATATTCCCCCCTCC
>gnl|BL_ORD_ID|0:88511167-88511701 chr10
GCTCCACATGCTGGATTTCTGGACTACTAGCTCCAAAGGAATTCTTGAGTAGAATCAGTGTTTAGGGATACTAAGAAGAA
TCAGACATGGGTCTTTTCATTTTTGAATGTATGTATGGGTTTTTTGTGTGTATACACACAAACATATGTACATGCATACG
TACATATATGTACATATGTGTACATGCATACTTACATATGCACACATATGCACGCATACACATGTGCACATATGCGTGCG
CACATGCACATGTGCGCACATACGCATATATGCACGTGTGCACATACGCATATGCGCACGTGTGCGCACATACGCATATG
CGCACATATGTGTGCACACATACGCATATGCGCACATATGTGTGCACACATACGCATATGCGCACATATGTGTGCACACA
TATGCATATTTGTGTGTACATATATACATATATGCATATACTTACATATGTACACAAACACCTACACACACACACAAAGT
TCAGACTAGAAATCTATTTGAATTATATTCAATAGATTATATACACACACACCCC
>gnl|BL_ORD_ID|0:84427298-84428017 chr6
TGTGTATACACACATACGTATATACGTATGTACACATATACATATATACGTATATATACACATATACACCTATATACACA
CATATACATATATGTACATATACGCACATATACATATATGTGTACATATATACACATACATATATGCACATATATACACA
CATGTACACTATATACACACATGTACACCTATGCACATATATACACACATGCACACACGTACACACACGCACACATATGT
```

Summit(ピークの頂点)だけ取り出すなら

chr	start	end	length	summit	tags	#NAME?	fold_enrichm	FDR(%)	
chr7	1513589	1514175	587	278	139	1183.26	92.59	0	
chr5	154133800	154134796	997	740	243	1798.41	71.52	0	
chr19	56151219	56152489	1271	354	264	1809.34	70.86	0	
chr15	55818703	55819145	443	166	100	826.24	68.42	0	
chr22	21983060	21984583	1524	550	280	1925.75	68.34	0	
chr16	75267757	75268793	1037	224	185	1182.99	68.07	0	
chr11	69451031	69451660	630	390	138	1049.97	67.06	0	
chr2	68589115	68589553	439	214	148	1289.82	66.62	0	
chr22	41418420	41418939	520	271	111	899.41	66.27	0	
chr10	88726173	88726879	707	306	131	951.96	64.52	0	
chr19	2391267	2391794	528	271	100	770.82	64.45	0	

[start+summit-1]±XXX base のように計算して、後は同じ (やってみてください)

Top100だけみたい場合は

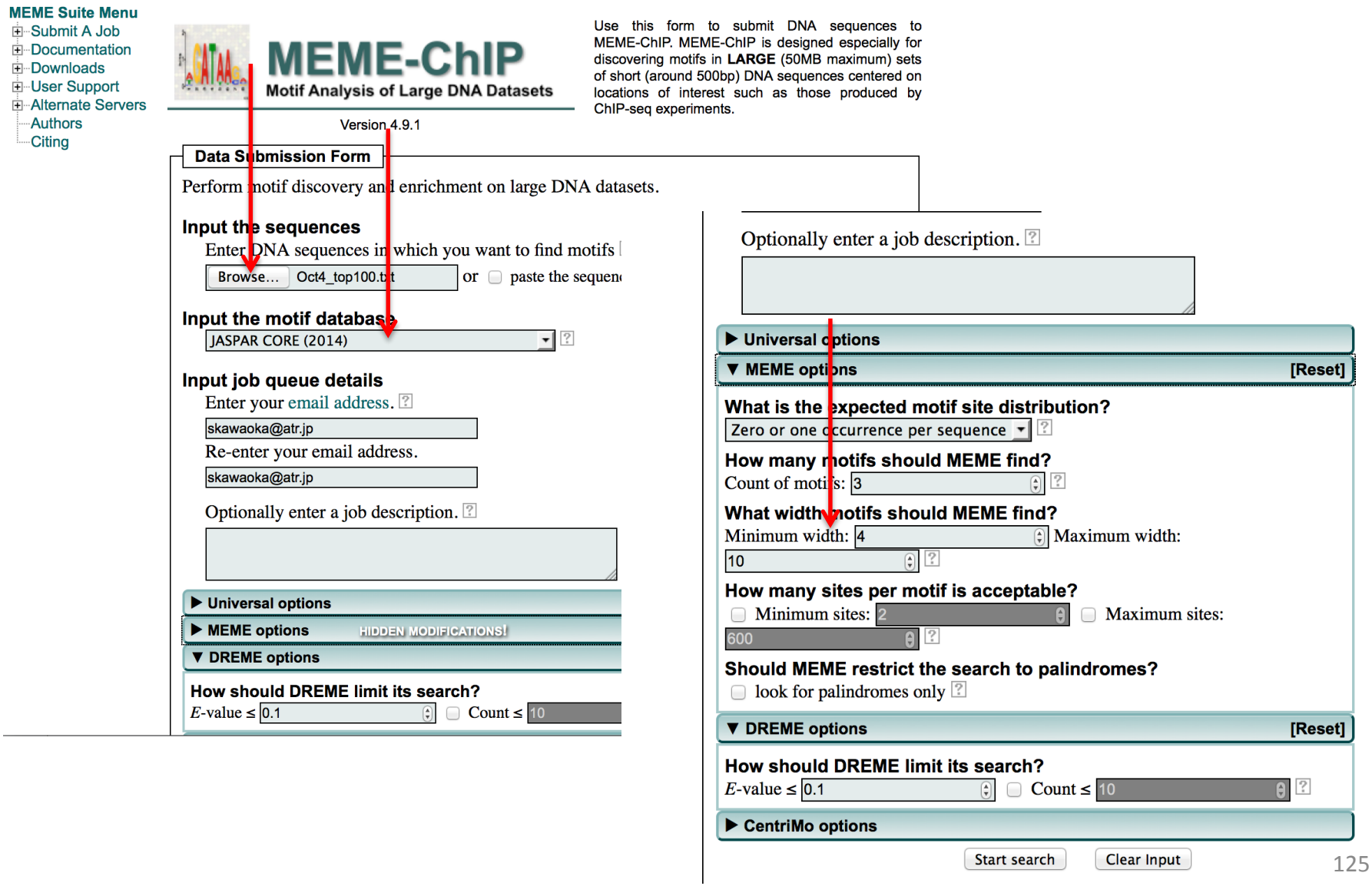
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	# This file is generated by MACS version 1.4.2 20120305															
2	# ARGUMENTS LIST:															
3	# name = chr17_cMyc															
4	# format = BAM															
5	# ChIP-seq file = chr17_cMyc_uniq.bam															
6	# control file = chr17_Input_uniq.bam															
7	# effective genome size = 2.70e+09															
8	# band width = 300															
9	# model fold = 10,30															
10	# pvalue cutoff = 1.00e-05															
11	# Large dataset will be scaled towards smaller dataset.															
12	# Range for calculating regional lambda is: 1000 bps and 10000 bps															
13																
14	# tag size is determined as 36 bps															
15	# total tags in treatment: 1012926															
16	# tags after filtering in treatment: 937654															
17	# maximum duplicate tags at the same position in treatment = 1															
18	# Redundant rate in treatment: 0.07															
19	# total tags in control: 900922															
20	# tags after filtering in control: 815189															
21	# maximum duplicate tags at the same position in control = 1															
22	# Redundant rate in control: 0.10															
23	# d = 57															
24	chr	start	end	length	summit	tags	#NAME?	fold_enrichm	FDR(%)							
25	chr17	18472963	18473117	155	107	9	142.9	307.02	0.17	Peak0	blastdbcmd -db chr17.fa -entry all - range 18472963-18473117 > Peak001.txt					
26	chr17	36452859	36452997	139	55	16	236.3	292.4	0	Peak00	blastdbcmd -db chr17.fa -entry all - range 36452859-36452997 > Peak002.txt					
27	chr17	20558089	20558183	95	47	8	160.71	233.92	0.2	Peak003	blastdbcmd -db chr17.fa -entry all - range 20558089-20558183 > Peak003.txt					
28	chr17	18747024	18747112	89	44	4	94.08	232.43	1.05	Peak004	blastdbcmd -db chr17.fa -entry all - range 18747024-18747112 > Peak004.txt					
29	chr17	19015258	19015335	78	39	4	116.49	232.43	0.39	Peak005	blastdbcmd -db chr17.fa -entry all - range 19015258-19015335 > Peak005.txt					
30	chr17	18585469	18585739	271	124	9	116.02	140.35	0.39	Peak006	blastdbcmd -db chr17.fa -entry all - range 18585469-18585739 > Peak006.txt					
31	chr17	45579521	45579668	148	50	10	117.93	112.78	0.39	Peak007	blastdbcmd -db chr17.fa -entry all - range 45579521-45579668 > Peak007.txt					
32	chr17	36631025	36631209	185	65	16	171.35	96.49	0.22	Peak008	blastdbcmd -db chr17.fa -entry all - range 36631025-36631209 > Peak008.txt					
33	chr17	15492114	15492194	81	40	5	103.47	87.72	0.65	Peak009	blastdbcmd -db chr17.fa -entry all - range 15492114-15492194 > Peak009.txt					
34	chr17	15752486	15752586	101	50	4	53.71	77.97	17.69	Peak010	blastdbcmd -db chr17.fa -entry all - range 15752486-15752586 > Peak010.txt					
35	chr17	79961226	79962127	902	453	146	960.53	70.18	0	Peak011	blastdbcmd -db chr17.fa -entry all - range 79961226-79962127 > Peak011.txt					
36	chr17	18684460	18684647	188	60	9	91.12	70.18	1.19	Peak012	blastdbcmd -db chr17.fa -entry all - range 18684460-18684647 > Peak012.txt					
37	chr17	18312831	18312953	123	35	5	66.74	70.18	5.54	Peak013	blastdbcmd -db chr17.fa -entry all - range 18312831-18312953 > Peak013.txt					
38	chr17	43528559	43528687	129	76	5	65.73	70.18	5.69	Peak014	blastdbcmd -db chr17.fa -entry all - range 43528559-43528687 > Peak014.txt					
39	chr17	42147867	42148849	983	206	212	1203.24	70	0	Peak015	blastdbcmd -db chr17.fa -entry all - range 42147867-42148849 > Peak015.txt					
40	chr17	27438437	27439184	748	551	121	644.02	66.81	0	Peak016	blastdbcmd -db chr17.fa -entry all - range 27438437-27439184 > Peak016.txt					
41	chr17	61510584	61510837	254	144	76	634.37	55.56	0	Peak017	blastdbcmd -db chr17.fa -entry all - range 61510584-61510837 > Peak017.txt					

先ほど作ったファイルをFold enrichment/tagsでソートするだけ

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	chr	start	end	length	summit	tags	#NAME?	fold_enrichm	FDR(%)							
2	chr17	18472963	18473117	155	107	9	142.9	307.02	0.17	Peak001	blastdbcmd -db chr17.fa -entry all - range 18472963-18473117 > Peak001.txt					
3	chr17	36452859	36452997	139	55	16	236.3	292.4	0	Peak002	blastdbcmd -db chr17.fa -entry all - range 36452859-36452997 > Peak002.txt					
4	chr17	20558089	20558183	95	47	8	160.71	233.92	0.2	Peak003	blastdbcmd -db chr17.fa -entry all - range 20558089-20558183 > Peak003.txt					
5	chr17	18747024	18747112	89	44	4	94.08	232.43	1.05	Peak004	blastdbcmd -db chr17.fa -entry all - range 18747024-18747112 > Peak004.txt					
6	chr17	19015258	19015335	78	39	4	116.49	232.43	0.39	Peak005	blastdbcmd -db chr17.fa -entry all - range 19015258-19015335 > Peak005.txt					
7	chr17	18585469	18585739	271	124	9	116.02	140.35	0.39	Peak006	blastdbcmd -db chr17.fa -entry all - range 18585469-18585739 > Peak006.txt					
8	chr17	45579521	45579668	148	50	10	117.93	112.78	0.39	Peak007	blastdbcmd -db chr17.fa -entry all - range 45579521-45579668 > Peak007.txt					
9	chr17	36631025	36631209	185	65	16	171.35	96.49	0.22	Peak008	blastdbcmd -db chr17.fa -entry all - range 36631025-36631209 > Peak008.txt					
10	chr17	15492114	15492194	81	40	5	103.47	87.72	0.65	Peak009	blastdbcmd -db chr17.fa -entry all - range 15492114-15492194 > Peak009.txt					
11	chr17	15752486	15752586	101	50	4	53.71	77.97	17.69	Peak010	blastdbcmd -db chr17.fa -entry all - range 15752486-15752586 > Peak010.txt					
12	chr17	79961226	79962127	902	453	146	960.53	70.18	0	Peak011	blastdbcmd -db chr17.fa -entry all - range 79961226-79962127 > Peak011.txt					
13	chr17	18684460	18684647	188	60	9	91.12	70.18	1.19	Peak012	blastdbcmd -db chr17.fa -entry all - range 18684460-18684647 > Peak012.txt					
14	chr17	18312831	18312953	123	35	5	66.74	70.18	5.54	Peak013	blastdbcmd -db chr17.fa -entry all - range 18312831-18312953 > Peak013.txt					
15	chr17	43528559	43528687	129	76	5	65.73	70.18	5.69	Peak014	blastdbcmd -db chr17.fa -entry all - range 43528559-43528687 > Peak014.txt					
16	chr17	42147867	42148849	983	206	212	1203.24	70	0	Peak015	blastdbcmd -db chr17.fa -entry all - range 42147867-42148849 > Peak015.txt					
17	chr17	27438437	27439184	748	551	121	644.02	66.81	0	Peak016	blastdbcmd -db chr17.fa -entry all - range 27438437-27439184 > Peak016.txt					
18	chr17	61510584	61510837	254	144	76	634.37	55.56	0	Peak017	blastdbcmd -db chr17.fa -entry all - range 61510584-61510837 > Peak017.txt					

どういう基準でソートするか、上位いくつのピークを使うのか、は、解析者側のチョイス

実際にモチーフ解析
モチーフサイズ→4-10



Oct4のモチーフ

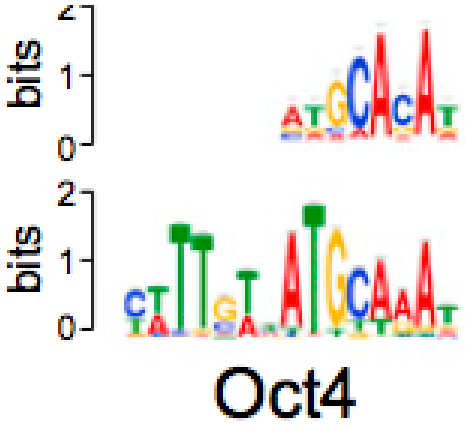
(Oct4_top100.txt(ダウンロード可))

MEME-ChIP
Motif Analysis of Large DNA Datasets

If you use MEME-ChIP in your research, please cite the following paper:
Philip Machanick and Timothy L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets", *Bioinformatics*, **27**12, 1696-1697, 2011.

MOTIFS | PROGRAMS | INPUT FILES | PROGRAM INFORMATION

The full MEME-ChIP analysis can be downloaded as a gzipped tar file from here.



Oct4

MOTIFS

The motifs found by the programs MEME, DREME and CentriMo; clustered by similarity and ordered by E-value.

Motif Found	Discovery/Enrichment Program	E-value	Known or Similar Motifs	Distribution
	MEME	1.0e-194	POU2F2 (MA0507.1)	Not Centrally Enriched
	MEME	9.2e-150	LEC2 (MA0581.1)	Not Centrally Enriched
	MEME	1.6e-125	DAF-12 (MA0538.1)	Not Centrally Enriched

cMycのモチーフ
(cMyc_top100.txt(ダウンロード可))

MEME-ChIP
Motif Analysis of Large DNA Datasets

c-Myc

If you use MEME-ChIP in your research, please cite the following paper:
Philip Machanick and Timothy L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets", *Bioinformatics*, **27**12, 1696-1697, 2011.

MOTIFS | PROGRAMS | INPUT FILES | PROGRAM INFORMATION

The full MEME-ChIP analysis can be downloaded as a gzipped tar file from here.

MOTIFS

The motifs found by the programs MEME, DREME and CentriMo; clustered by similarity and ordered by E-value.

Motif Found	Discovery/Enrichment Program	E-value	Known or Similar Motifs	Distribution
	MEME	1.2e-015	BES1 (MA0549.1) MYC::MAX (MA0059.1) PIF4 (MA0561.1)	Not Centrally Enriched
	DREME	4.6e-007	PIF4 (MA0561.1) Mycn (MA0104.3) PIF5 (MA0562.1)	Not Centrally Enriched

Reverse Complement ⇆ Show Less ↥

ChIP-seq解析で大事なのは

- 良い抗体 (言わずもがな)
- 良質のChIPed DNA (結構ダメChIPがあります)
- じゅうぶんなリード数
- 丁寧なvalidation
- 目的に応じた使い分け

(ChIP-seq以外でも大事ですが)

謝辞

門田幸二先生
孫建強氏