ライフサイエンスデータベース統合推進事業
統合化推進プログラム（統合データ解析トライアル）

研究開発課題
「ChIP-seq SRA の統合的可視化とバイオデータベースとの連携」

研究開発終了報告書

研究開発期間：　平成 27 年 5 月 15 日～平成 28 年 3 月 31 日

研究代表者：　沖　真弥
（九州大学　大学院医学研究院　助教）

§1 研究開発の概要

本研究は INSDC に SRA (Sequence Read Archive) として登録されたほとんどすべての ChIP-seq (Chromatin-Immunoprecipitation with Sequencing) データを収集し、誰もが簡単に利活用できるためのデータベース (ChIP-Atlas) を作成した。そのために4万以上の ChIP-seq 実験データをマッピングデータに変換し、ピークコールをおこなった。また、各データに用いられた抗原や細胞などのメタ情報のクレンジングをおこない、任意の属性によるデータ抽出や統合的な解析を可能にした。これにより、ユーザは興味のゲノム領域に結合するタンパク質を閲覧できるだけでなく、興味の転写因子の標的遺伝子や共局在パートナーを知ることができる。また、統合化推進プログラムで開発されたデータベースを利用し、各データに用いられた抗原や細胞などの詳細情報や入手先などが容易に得られるようにした。

§2 研究開発のねらい

ChIP-seq 法はゲノム上に結合する転写因子や修飾ヒストンなどの分布をゲノムワイドに調べることができ、遺伝子制御ネットワークを解明するための強力な実験手法である。これらは論文などに投稿する際、SRA として NCBI, DDBJ や ENA に登録される。しかし、そこに登録されるのはほとんどがシーケンス生データのみであるため、そのデータの可視化や利活用のためには複雑な計算処理が必要である。そこで本研究のねらいは、それら既報の ChIP-seq データをあらかじめ全て計算処理し、誰もが簡単に利活用できるようなデータベースを作成することである。

ChIP-seq SRA のもう一点の問題は、メタ情報の表記ゆれである。SRA として登録された ChIP-seq データには、実験対象の抗原や細胞などの情報が投稿者自身によって記述されている。しかし、その記述法に明確なガイドラインがないため、synomym や略記、さらには誤記などを含む表記ゆれが多い。そのため、任意の抗原や細胞名などの属性でデータを抽出したり統合することが極めて困難である。そこで本研究では統合化推進プログラムで開発されたデータベースやその他のリソースを利用することで、メタ情報のクレンジングをおこない、さらにはそれらに関する詳細な情報を簡単に入手できるようなプラットフォームを作成した。

§3 研究開発計画

(1) 当初の研究開発計画

■*メタデータのクレンジング*

NCBI が公開している SRA metadata より、抗原・抗体名や細胞・組織名の収集をおこない、表記のゆれを統一する。転写因子などの抗原名は Gene Nomenclature Committee の表記法に統一する。組織名は MeSH (= Medical Subject Headings) に、セルライン名は ATCC (= American Type Culture Collection) の表記法に準拠する。データクレンジングツール (Google Refine) を用いながら、手作業で表記修正する。

■ChIP-seq 統合データの作成

上記の表記ゆれの統一により、各種細胞タイプでデータ統合が可能となる。これにより、任意の細胞タイプで何がどこに結合しているかが一目でわかるような可視化データを作成する。フォーマットはインデクス付きの Bed 形式を採用し、ゲノムブラウザでスムーズな閲覧ができるようにする。

■ChIP-seq データの統合解析

類似の ChIP-seq プロファイルを示す転写因子群は複合体を形成し、発現制御機能に相乗効

果や修飾効果がもたらされる（例：ES細胞におけるOct4-Sox2-Nanogなど)。このような転写因子群を同定するために、東大・油谷研究室の仲木 竜氏の CoLo というアルゴリズム（投稿中）を使用する。これは数千の ChIP-seq データの中から類似データの組み合わせを抽出でき、共局在性の高い因子群を予測できる。

また、転写因子が標的とする遺伝子リストを作成する。これはそれぞれの転写因子について、転写開始点 ± 5 kb に結合される遺伝子を抽出し、ピークコーラ（MACS2）の統計量でランキング表示する。

■データの公開

上記までのデータを NBDC のファイルサーバに格納し、ウェブブラウザを通じて公開する。とくに表記ゆれの統一を生かし、ユーザが興味のある抗原や細胞タイプで検索できるようにする。また、それぞれの抗原や細胞タイプ、および実験番号（SRX など）に外部データベースへのリンクを設け、それらの詳細な情報が瞬時に得られるようにする。

（2）新たに追加・修正など変更した研究開発計画

■ユーザデータを解析するツール

何らかの共通性質を持つゲノム領域（BED 形式）や遺伝子リストを投稿すると、そこに結合が enrichment するタンパク質を返すようなツールを作成し、公開した。

■細胞分化や遺伝的疾患を司る転写因子の予測

上記の方法を用い、組織特異的に発現する遺伝子やエンハンサー領域、または各種疾患特異的 SNP に対し、結合が enrichment するタンパク質を同定した。早期の公開に向けて準備中。

§4 研究開発成果

■ ChIP-seq SRA データの収集と計算

NCBI の FTP サイトより、登録されているすべての SRA メタデータを収集した。その中で、LIBRARY STRATEGY が「ChIP-seq」 または「DNase-Hypersensitivity」で、なおかつ LIBRARY_SOURCE が「GENOMIC」であるものを選別した。その上で Fig. 1A に示す生物種の SRA をダウンロードし、FastQ 形式に変換後、Bowtie2 を用いて各レファレンスゲノムにマッピングした。それを BAM 形式に変換し、カバレージデータ（BigWig 形式）およびピークコールデータ（BED 形式、MACS2 を使用）を作成した。2016 年 1 月時点での実験データ数を Fig. 1A に示す。

■ *メタデータのクレンジング*

NCBI が公開している SRA metadata より、各実験に用いられた抗原・抗体名や細胞・組織名を収集し、それらすべてについて手作業で表記の修正をおこなった。表記ゆれがないように一定の基準を設けており、転写因子などの抗原名は Gene Nomenclature Committee の表記法に、修飾ヒストンは Brno nomenclature にしたがった。また、セルライン名は Yu et.al 2015 (PMID: 25877200) で提唱された統一的表記法や ATCC で採用されているそれにしたがい、組織名は MeSH の表記法に準拠した。これらの抗原名や細胞名はさらに大分類（= Class）に配属させ、データの検索や抽出を容易にできるようにした。

	変換前		変換後 (Class)	変換後 (名前)
抗原	trimethyl K4	→	Histone	H3K4me3
	OCT3/4	→	TFs and Others	POU5F1
細胞	K562	→	Blood	K-562
	ESC	→	Pluripotent stem cell	Embryonic Stem Cells

また、作業の効率化のためにデータクレンジングツール（Google Refine）用い、さらに自作の補助ツールを開発した（新たに追加された生のメタ情報表記を入力すると、過去の変換ログから正解表記がサジェストされる）。これにより、毎月更新されるSRAデータに対し、迅速なアノテーションが可能となった。2016年1月時点における各Classの数をFig. 1Bに示す。

■統合化推進プログラムで統合化されたデータベースの活用

上記のように、すべての ChIP-seq データに対して統一された表記法で抗原名と細胞名が割りふられたが、その名前だけではユーザにとって理解しがたいことが十分想定される。そこで各実験に関するメタデータや詳細な情報を web ページで閲覧できるようにした（Fig. 2、例としてSRX018625を表示）。そこではクレンジング前後のメタデータだけでなく、マップ率や総ピーク数などの解析ログが閲覧できる。さらに抗原や細胞に関する知見が得られるように外部データベースへのリンクを充実させている。抗原名についてはPosMed, PDBj, WikiGenesに、細胞名についてはATCC, MeSH, RIKEN BRCにリンクされる（下線は統合化推進プログラムで統合化されたデータベース）。またDBCLS SRAより、リードのクオリティーデータやライブラリ情報が閲覧できる。

■ChIP-seq 統合データの作成

上記のメタデータクレンジングを活かし、各種属性（抗原、細胞名）ごとにピークコールデータを統合した。後述するように web ブラウザでデータを選択、絞り込みをし、ゲノムブラウザ（IGV）でスムーズな閲覧ができるようにした。

■ChIP-seq データの統合解析

それぞれの転写因子について、転写開始点 ± N kb (N = 1, 5, or 10) に結合される遺伝子を抽出し、ピークコーラ（MACS2）の統計量でヒートカラー表示させた。また、東大・油谷研究室の仲木 竜氏の CoLo というアルゴリズム（投稿中）を使用し、任意の転写因子に対し、よく似たプロファイルを示すChIP-seqデータを抽出し、その類似度をヒートカラー表示させた。

■ユーザデータを解析するツール

ユーザによって投稿された BED ファイルや遺伝子リストを NIG supercomputer に送り、enrichment 解析をおこなったのち、ユーザに解析結果を返すためのパイプラインを構築した（DDBJの小笠原理・奥田喜広 両氏と、DBCLSの大田達郎氏との共同研究）。そこでは、投稿された2群のデータに対し、すべての（または Class 指定された）ChIP-seq ピークとのオーバーラップが計算される。その結果を統計処理し、どちらかの群と有意にオーバーラップするようなタンパク質が結果として返される。

■細胞分化や遺伝的疾患を司る転写因子の予測

上記のツールをもちい、大規模なenrichment解析をおこなった。入力データはFANTOM5より得られた組織特異的遺伝子リストや組織特異的エンハンサー（BED 形式）、および GWAS catalog より得られた疾患特異的変異部位（SNP 部位に linkage disequilibrium 領域を付け加えた BED ファイル）である。これらに対し、結合が有意に enrich するような転写因子を抽出した。さらに、そのデータをクラスター解析することにより、中でも重要な転写因子とゲノム領域を判別できるようにした。これらの解析結果は web サイトを通じて公開する予定である。

■データの公開

NBDCの畠中秀樹氏の協力により、上記までのデータをすべてNBDCのサーバに保管している。これを閲覧するためのウェブサイトを ChIP-Atlas として 2015 年 12 月に公開した（http://chip-atlas.org）。ウェブインターフェースの作成やデータベースとしての整備はDBCLSの大田氏が多大に貢献した。また、本研究の構想は大田氏と、RIKEN の塩井剛氏の助言による。

2016 年 2 月時点までに 286 人の新規ユーザが 2,301 のクエリをおこなっている。

■ ChIP-Atlas の使用方法

ChIP-Atlas のホームページにアクセスすると、Fig. 3 のように4つの機能を利用できる（Peak Browser, Target Genes, Colocalization, in silico ChIP)。使い方をわかりやすくまとめたムービーを作成しており、そのリンクが貼られている。

(1) Peak Browser

Fig. 4A のように興味のあるデータを選択できる。すべての抗原や細胞でも可能。IGV を起動した上で、「View on IGV」ボタンをクリックすると興味のゲノム領域にどんなタンパク質が結合しているかが理解できる（Fig. 4B)。とくに、メタデータクレンジングされた抗原名と細胞名が表示され、結合強度（MACS2 score）がヒートカラー表示される。興味のピークにマウスオーバーすると詳細なメタデータが表示され（Fig. 4B 黄色の四角)、クリックするとさらに詳細な情報を閲覧できる (Fig. 2) ほか、BigWig 形式のカバレージデータも表示できる（Fig. 4B 赤色の四角)。これにより、興味の遺伝子周辺における転写調節領域を理解でき、さらにそれらを制御するタンパク質を知ることができる。

(2) Target genes

Fig. 5A のように興味のある転写因子を選択できる。ここでは例として、転写開始点 ± 5 kb において転写因子 POU5F1 が結合する遺伝子の取得をクエリとしており、Fig. 5B のような結果が表示される。列はすべての POU5F1 ChIP-seq データで、行はその標的遺伝子候補を示し、それらへの結合強度（MACS2 score）がヒートカラー表示される。これにより、ある転写因子の標的遺伝子を予測することができ、遺伝子制御ネットワークを理解するために利用できる。SRX で始まる番号は SRA に登録された実験番号を示し、それをクリックすると Fig. 2 のような詳細情報にアクセスできる。また、この表示法でユニークな点として、同一の転写因子による ChIP-seq データ同士の比較が可能である。例えば SRX011571 や SRX702068 は他の実験データとの相関がほとんどない。前者は他に比べてリード数やクオリティが低く、後者は分化によって POU5F1 が低発現した細胞を使っているためである。このように一覧して比較することで、より確かな実験データを判別できる。

(3) Colocalization

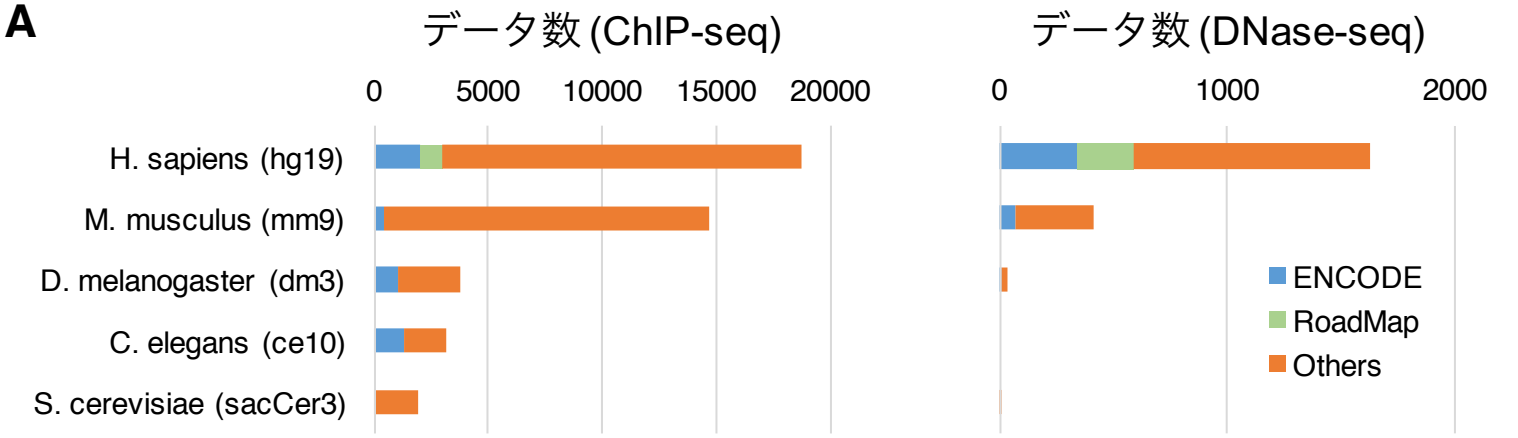
Fig. 6A のように興味のある転写因子を選択できる。ここでは例として、Pluripotent stem cell クラス内で転写因子 Nanog が共局在する因子の取得をクエリとしており、Fig. 6B のような結果が表示される。列はすべての Nanog ChIP-seq データで、行はそれらとよく似た実験データを示し、その類似度（CoLo score）がヒートカラー表示される。すでによく知られている Pou5f1 や Sox2 のほか、今までに知られていなかったような転写因子も表示されている。最右列は protein-protein interaction データベース（STRING）の情報が色づけされており、グレーのものは登録がないことを示す。転写因子はゲノム上で共局在することによって活性が強化・変化することが知られており、本機能はそのようなパートナー因子を予測するために利用できる。

(4) in silico ChIP

Fig. 7A では興味のある抗原や細胞 Class を選び、hepatocyte 特異的なエンハンサー (Data_A) とその他の FANTOM5 エンハンサー（Data_B）を BED 形式で入力している。その結果、Fig. 8 のような結果が表示される。例えば1行目は実験 ID = SRX100544 の Liver における EP300 の ChIP-seq データである。その総ピーク数 24,334 のうち 80 個が hepatocyte 特異的エンハンサー（n = 286）とオーバーラップするのに対し、その他のエンハンサー（n = 20,509）とは 1,147 個としかオーバーラップしない。Fisher の正確確率検定による P 値は $10^{-32.1}$ で、hepatocyte

特異的エンハンサーへの Fold-enrichment は 5.0 倍である。つまり、Liver における EP300 は hepatocyte 特異的エンハンサーに対し、有意に結合することが示されている。データは *P* 値で昇順ソートされており、ほかにも hepatocyte の発生やダイレクトリプログラミングに重要な因子 (HNF4A/G, FOXA1/2) などが上位にランクされている。Fig. 7B では FANTOM5 CAGE データで得られた hepatocyte 特異的遺伝子群 (Data_A) を入力し、比較対照 (Data_B) はその他の RefSeq coding gene としており、それらの転写開始点 ± 5000 bp に結合するタンパク質を比較している。その結果、Fig. 8 と同様な結果が表示されている (Fig. 9)。特筆すべきは、両者とも Fig. 7A において、検索範囲をすべての細胞 Class に設定しているにもかかわらず (青矢頭)、結果では Cell class が「Liver」のデータが上位を占めている点であり (Fig. 8, 9 青矢頭)、本手法の有効性を強く裏付けるものである。そのため本機能により、興味の遺伝子群やゲノム領域をまとめて制御するような転写因子の同定に利用できる。とくに遺伝子制御ネットワークのマスターレギュレータの理解や、ダイレクトリプログラミング因子の候補付けに利用できると期待される。

Fig.1


A
データ数 (ChIP-seq)
データ数 (DNase-seq)
0
5000
10000
15000
20000
0
1000
2000
H. sapiens (hg19)
M. musculus (mm9)
D. melanogaster (dm3)
C. elegans (ce10)
S. cerevisiae (sacCer3)
ENCODE
RoadMap
Others

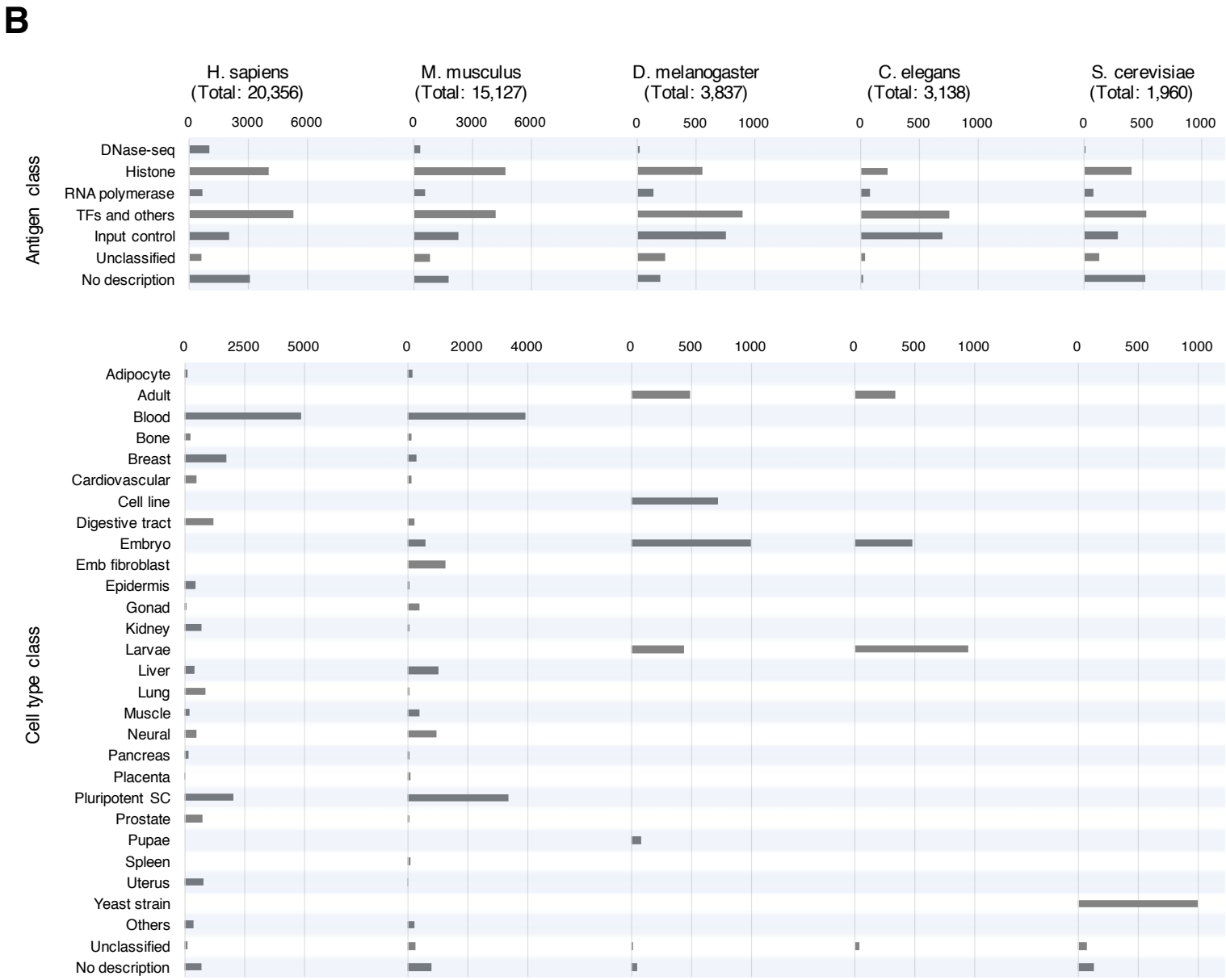

B
H. sapiens (Total: 20,356)
M. musculus (Total: 15,127)
D. melanogaster (Total: 3,837)
C. elegans (Total: 3,138)
S. cerevisiae (Total: 1,960)
Antigen class
DNase-seq
Histone
RNA polymerase
TFs and others
Input control
Unclassified
No description
Cell type class
Adipocyte
Adult
Blood
Bone
Breast
Cardiovascular
Cell line
Digestive tract
Embryo
Emb fibroblast
Epidermis
Gonad
Kidney
Larvae
Liver
Lung
Muscle
Neural
Pancreas
Placenta
Pluripotent SC
Prostate
Pupae
Spleen
Uterus
Yeast strain
Others
Unclassified
No description

Fig.2

SRX018625
GSM469863: HNF4a Fdomain ChIPSeq

View on IGV ▾ | View Analysis ▾ | Download ▾ | Link Out ▾

Curated Sample Data

Genome	hg19
Antigen Class	TFs and others
Antigen	HNF4A
Cell type Class	Liver
Cell type	Hep G2

Cell type information

Primary Tissue	Liver
Tissue Diagnosis	Carcinoma Hepatocellular

Attributes by Original Data Submitter

source_name	HNF4a_Fdomain_ChIPSeq
cell line	HepG2
cell type	hepatocellular carcinoma
chip antibody	HNF4a F domain

Metadata from Sequence Read Archive

Library Description

library_name	GSM469863: HNF4a_Fdomain_ChIPSeq
library_strategy	ChIP-Seq
library_source	GENOMIC
library_selection	ChIP

Platform Information

instrument_model	Illumina Genome Analyzer

External Database Query

Query antigen HNF4A | WikiGenes | PosMed | PDBj

Query cell-type Hep G2 | ATCC | MeSH | RIKEN BRC

Logs in read processing pipeline

Number of total reads	9231367
Reads aligned (%)	93.5
Duplicates removed (%)	3.8
Number of peaks	6228 (qval < 1E-05)

Sequence Quality Data from DBCLS SRA

SRR039087_fastqc

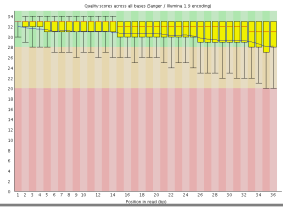
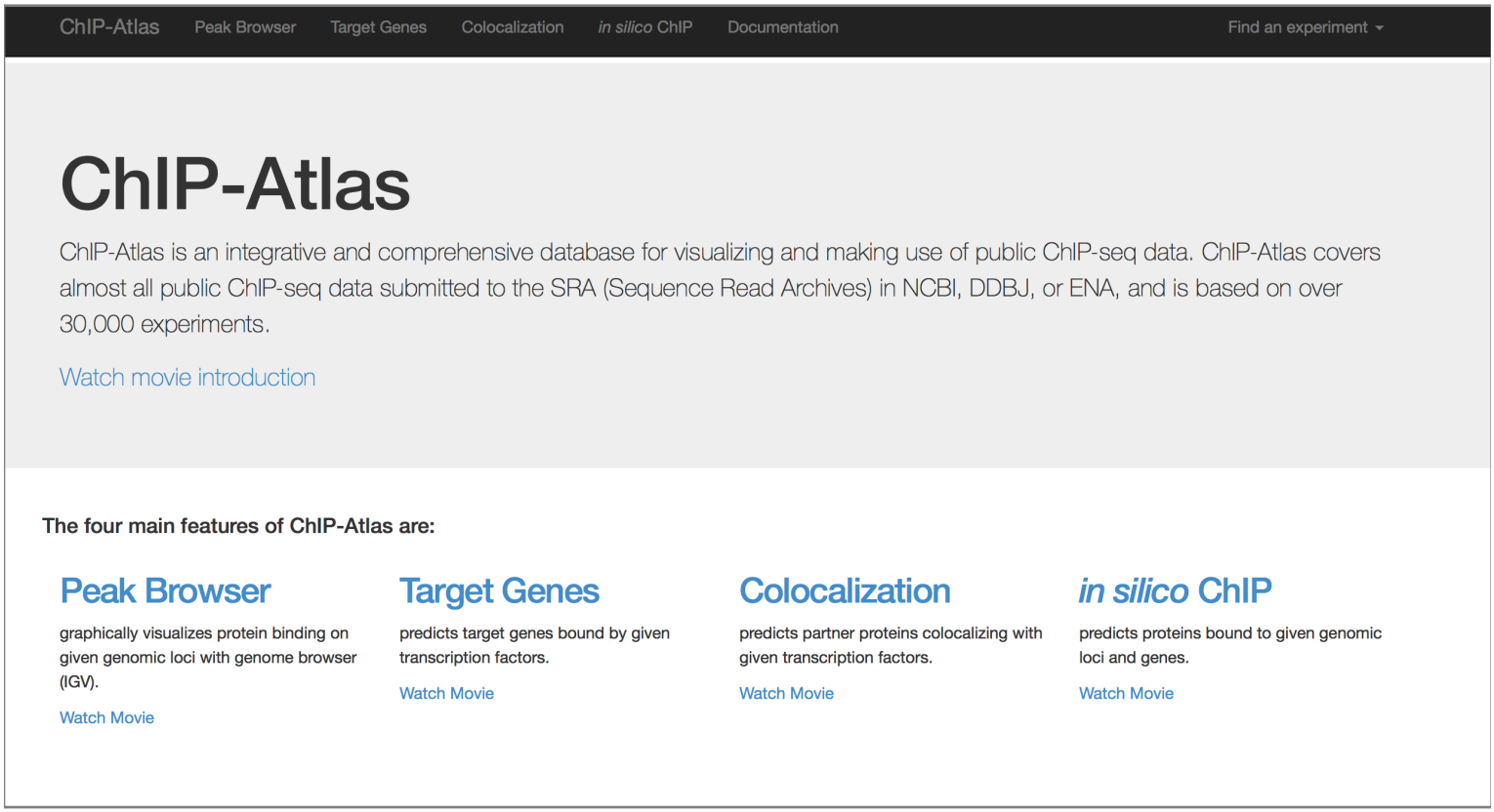


Fig.3

A


ChIP-Atlas
Peak Browser
Target Genes
Colocalization
in silico ChIP
Documentation
Find an experiment
ChIP-Atlas
ChIP-Atlas is an integrative and comprehensive database for visualizing and making use of public ChIP-seq data. ChIP-Atlas covers almost all public ChIP-seq data submitted to the SRA (Sequence Read Archives) in NCBI, DDBJ, or ENA, and is based on over 30,000 experiments.
Watch movie introduction
The four main features of ChIP-Atlas are:
Peak Browser
graphically visualizes protein binding on given genomic loci with genome browser (IGV).
Watch Movie
Target Genes
predicts target genes bound by given transcription factors.
Watch Movie
Colocalization
predicts partner proteins colocalizing with given transcription factors.
Watch Movie
in silico ChIP
predicts proteins bound to given genomic loci and genes.
Watch Movie

B

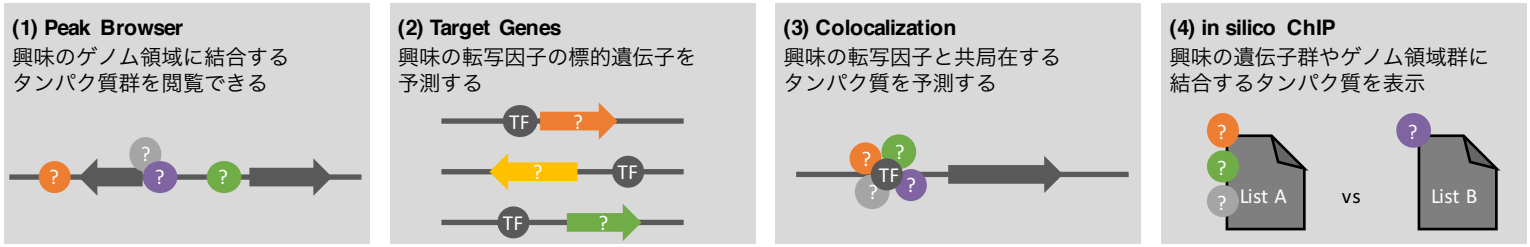
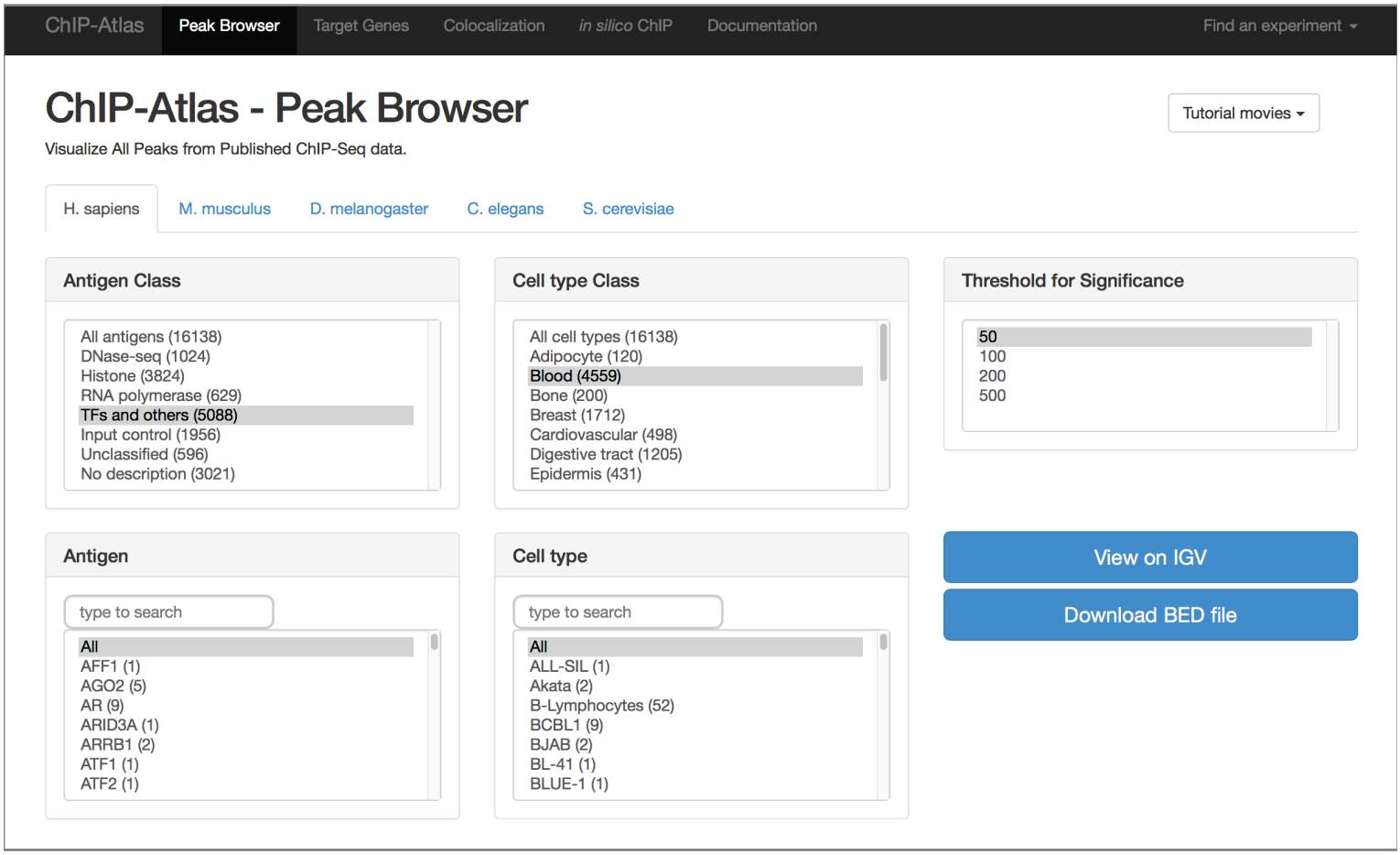
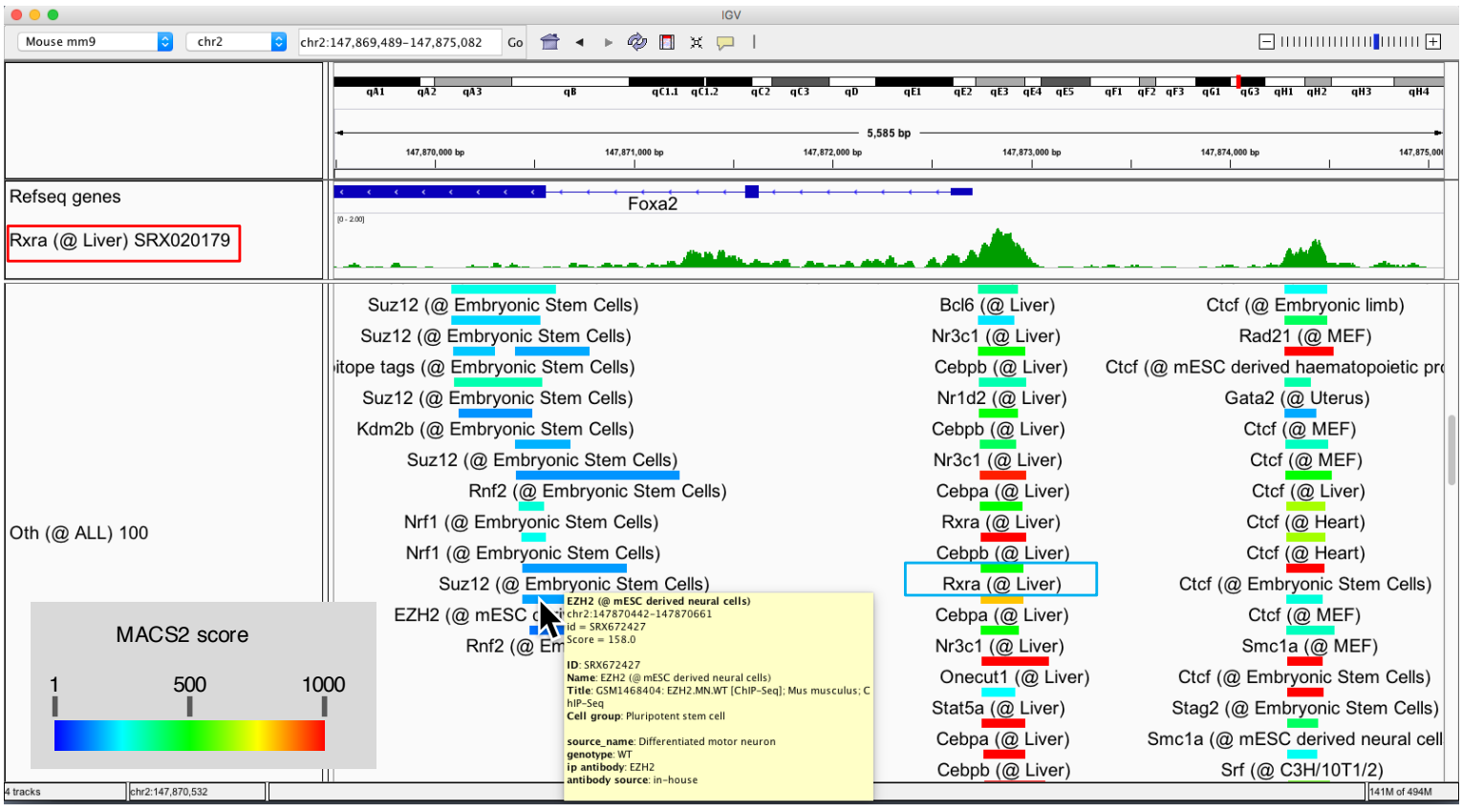

(1) Peak Browser
興味のゲノム領域に結合する
タンパク質群を閲覧できる
(2) Target Genes
興味の転写因子の標的遺伝子を
予測する
TF
(3) Colocalization
興味の転写因子と共局在する
タンパク質を予測する
TF
(4) in silico ChIP
興味の遺伝子群やゲノム領域群に
結合するタンパク質を表示
List A
vs
List B

Fig.4

A

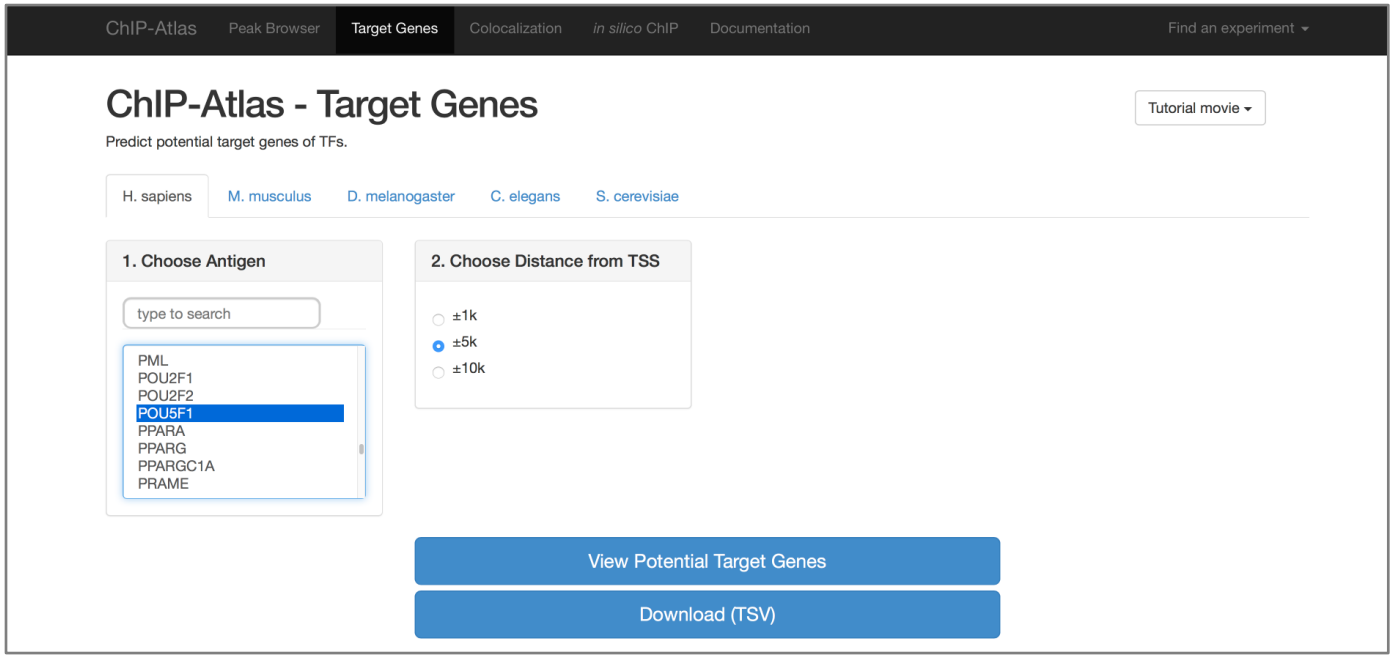

ChIP-Atlas
Peak Browser
Target Genes
Colocalization
in silico ChIP
Documentation
Find an experiment
ChIP-Atlas - Peak Browser
Visualize All Peaks from Published ChIP-Seq data.
Tutorial movies
H. sapiens
M. musculus
D. melanogaster
C. elegans
S. cerevisiae
Antigen Class
All antigens (16138)
DNase-seq (1024)
Histone (3824)
RNA polymerase (629)
TFs and others (5088)
Input control (1956)
Unclassified (596)
No description (3021)
Cell type Class
All cell types (16138)
Adipocyte (120)
Blood (4559)
Bone (200)
Breast (1712)
Cardiovascular (498)
Digestive tract (1205)
Epidermis (431)
Threshold for Significance
50
100
200
500
Antigen
type to search
All
AFF1 (1)
AGO2 (5)
AR (9)
ARID3A (1)
ARRB1 (2)
ATF1 (1)
ATF2 (1)
Cell type
type to search
All
ALL-SIL (1)
Akata (2)
B-Lymphocytes (52)
BCBL1 (9)
BJAB (2)
BL-41 (1)
BLUE-1 (1)
View on IGV
Download BED file

B

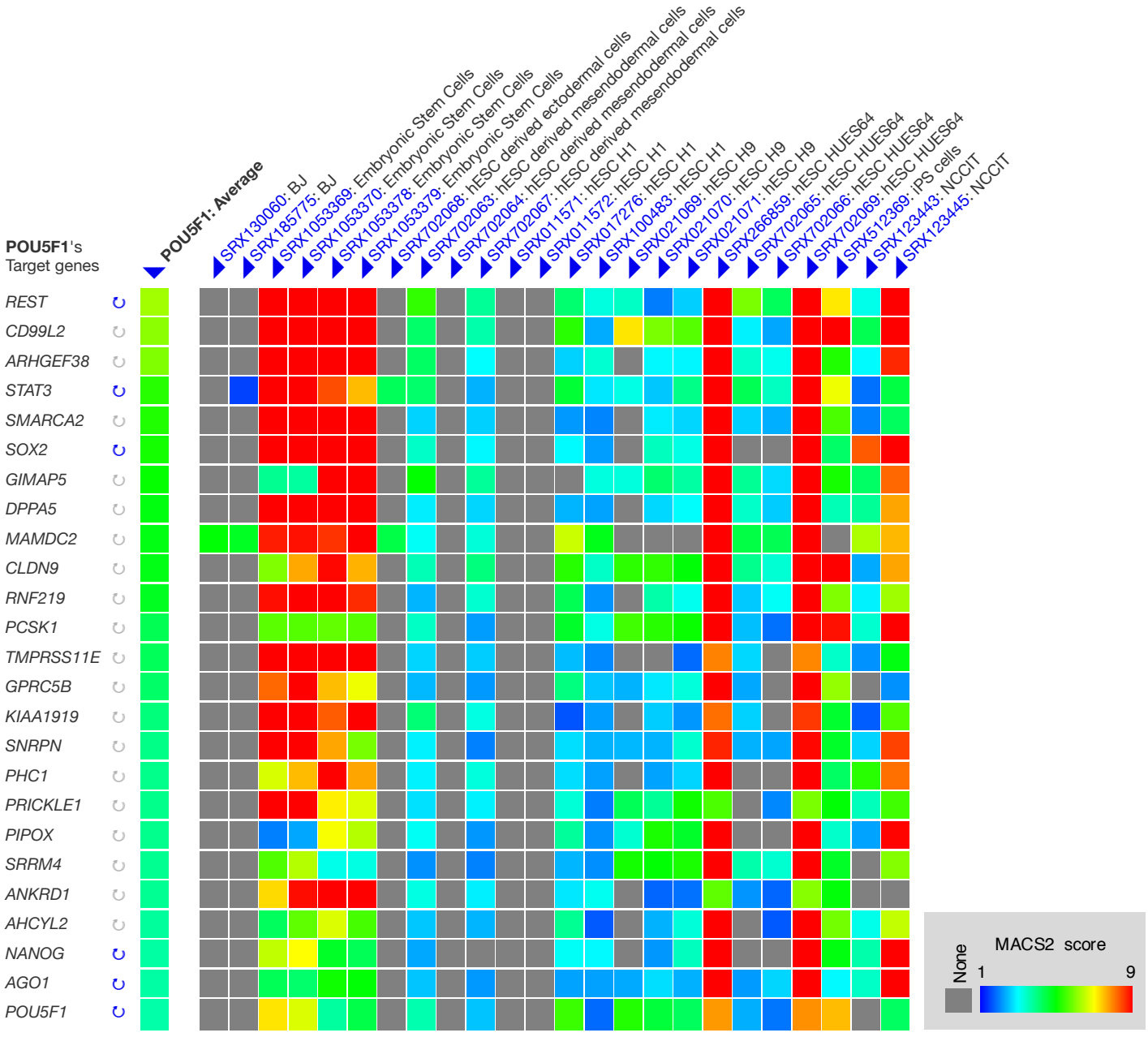

IGV
Mouse mm9
chr2
chr2:147,869,489-147,875,082
Go
5,585 bp
Refseq genes
Foxa2
Rxra (@ Liver) SRX020179
Oth (@ ALL) 100
Suz12 (@ Embryonic Stem Cells)
Suz12 (@ Embryonic Stem Cells)
Suz12 (@ Embryonic Stem Cells)
Kdm2b (@ Embryonic Stem Cells)
Suz12 (@ Embryonic Stem Cells)
Rnf2 (@ Embryonic Stem Cells)
Nrf1 (@ Embryonic Stem Cells)
Nrf1 (@ Embryonic Stem Cells)
Suz12 (@ Embryonic Stem Cells)
Bcl6 (@ Liver)
Nr3c1 (@ Liver)
Cebpb (@ Liver)
Nr1d2 (@ Liver)
Cebpb (@ Liver)
Nr3c1 (@ Liver)
Cebpa (@ Liver)
Rxra (@ Liver)
Cebpb (@ Liver)
Rxra (@ Liver)
Cebpa (@ Liver)
Nr3c1 (@ Liver)
Onecut1 (@ Liver)
Stat5a (@ Liver)
Cebpa (@ Liver)
Cebpb (@ Liver)
Ctcf (@ Embryonic limb)
Rad21 (@ MEF)
Gata2 (@ Uterus)
Ctcf (@ MEF)
Ctcf (@ MEF)
Ctcf (@ Liver)
Ctcf (@ Heart)
Ctcf (@ Heart)
Ctcf (@ Embryonic Stem Cells)
Ctcf (@ MEF)
Smc1a (@ MEF)
Ctcf (@ Embryonic Stem Cells)
Stag2 (@ Embryonic Stem Cells)
Srf (@ C3H/10T1/2)
MACS2 score
1
500
1000
EZH2 (@ mESC derived neural cells)
chr2:147870442-147870661
id = SRX672427
Score = 158.0
ID: SRX672427
Name: EZH2 (@ mESC derived neural cells)
Cell group: Pluripotent stem cell
source_name: Differentiated motor neuron
genotype: WT
ip antibody: EZH2
antibody source: in-house

A

Fig.5

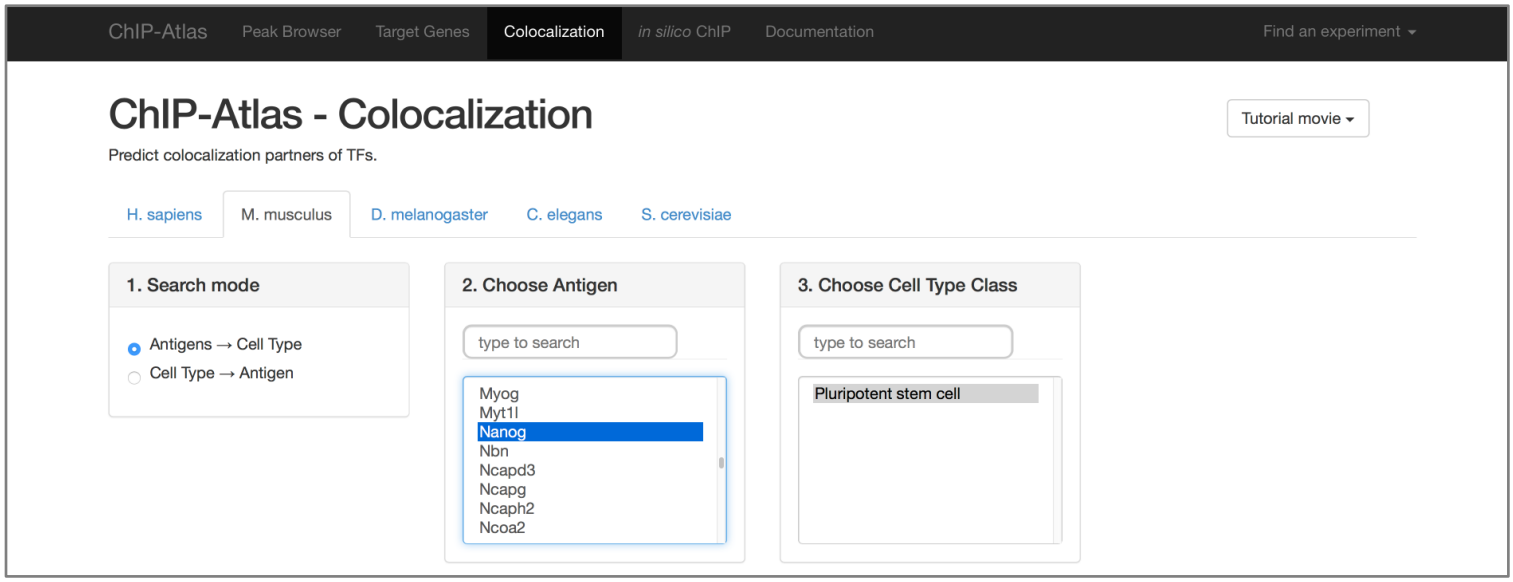

ChIP-Atlas
Peak Browser
Target Genes
Colocalization
in silico ChIP
Documentation
Find an experiment
ChIP-Atlas - Target Genes
Predict potential target genes of TFs.
Tutorial movie
H. sapiens
M. musculus
D. melanogaster
C. elegans
S. cerevisiae
1. Choose Antigen
type to search
PML
POU2F1
POU2F2
POU5F1
PPARA
PPARG
PPARGC1A
PRAME
2. Choose Distance from TSS
±1k
±5k
±10k
View Potential Target Genes
Download (TSV)

B

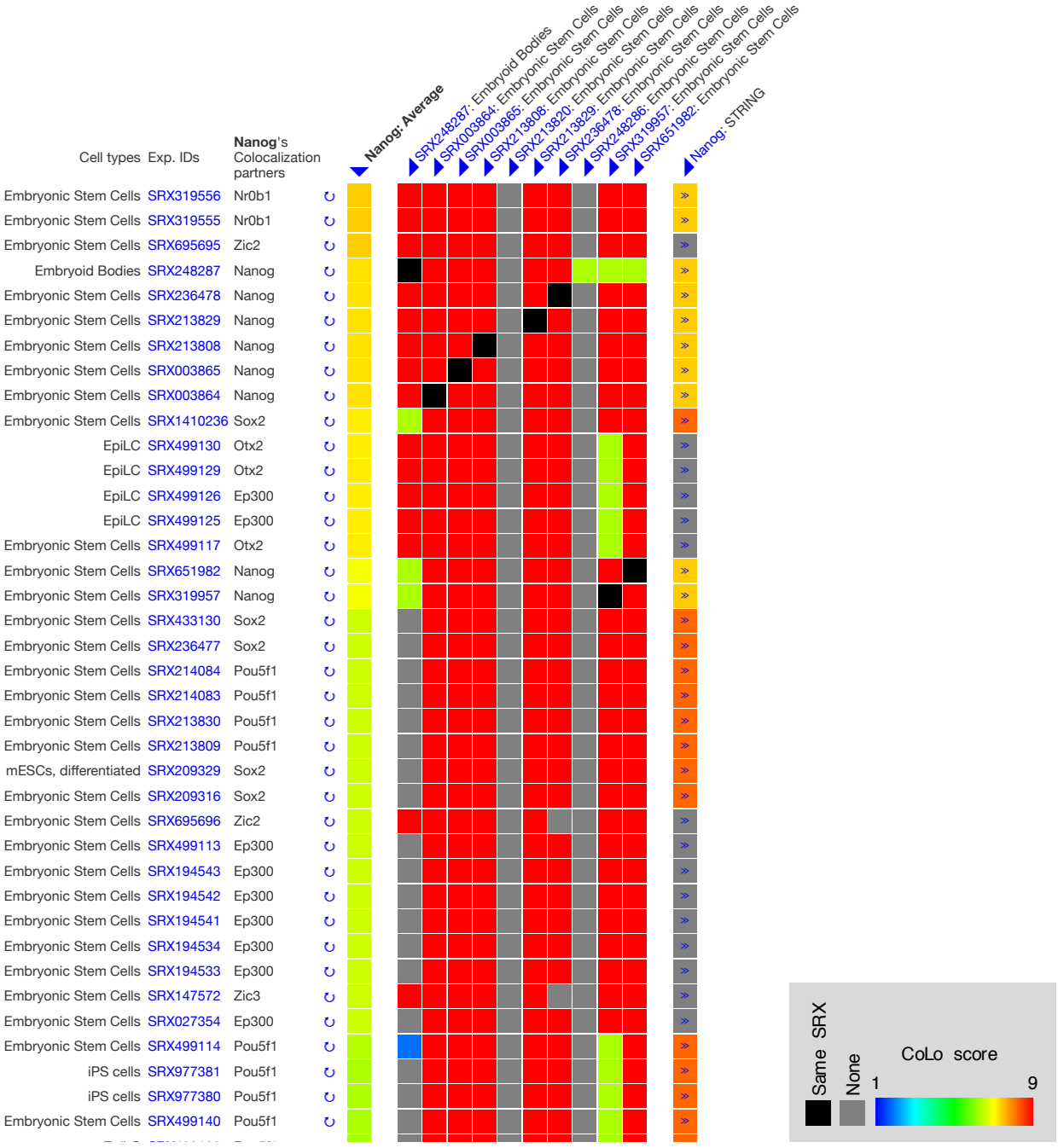

POU5F1's Target genes
POU5F1: Average
SRX130060: BJ
SRX185775: BJ
SRX1053369: Embryonic Stem Cells
SRX1053370: Embryonic Stem Cells
SRX1053378: Embryonic Stem Cells
SRX1053379: Embryonic Stem Cells
SRX702068: hESC derived ectodermal cells
SRX702063: hESC derived mesendodermal cells
SRX702064: hESC derived mesendodermal cells
SRX702067: hESC derived mesendodermal cells
SRX011571: hESC H1
SRX011572: hESC H1
SRX017276: hESC H1
SRX100483: hESC H1
SRX021069: hESC H1
SRX021070: hESC H9
SRX021071: hESC H9
SRX266859: hESC HUES64
SRX702065: hESC HUES64
SRX702066: hESC HUES64
SRX702069: hESC HUES64
SRX512369: iPS cells
SRX123443: NCCIT
SRX123445: NCCIT
REST
CD99L2
ARHGEF38
STAT3
SMARCA2
SOX2
GIMAP5
DPPA5
MAMDC2
CLDN9
RNF219
PCSK1
TMPRSS11E
GPRC5B
KIAA1919
SNRPN
PHC1
PRICKLE1
PIPOX
SRRM4
ANKRD1
AHCYL2
NANOG
AGO1
POU5F1
None
MACS2 score
1
9

A

Fig.6

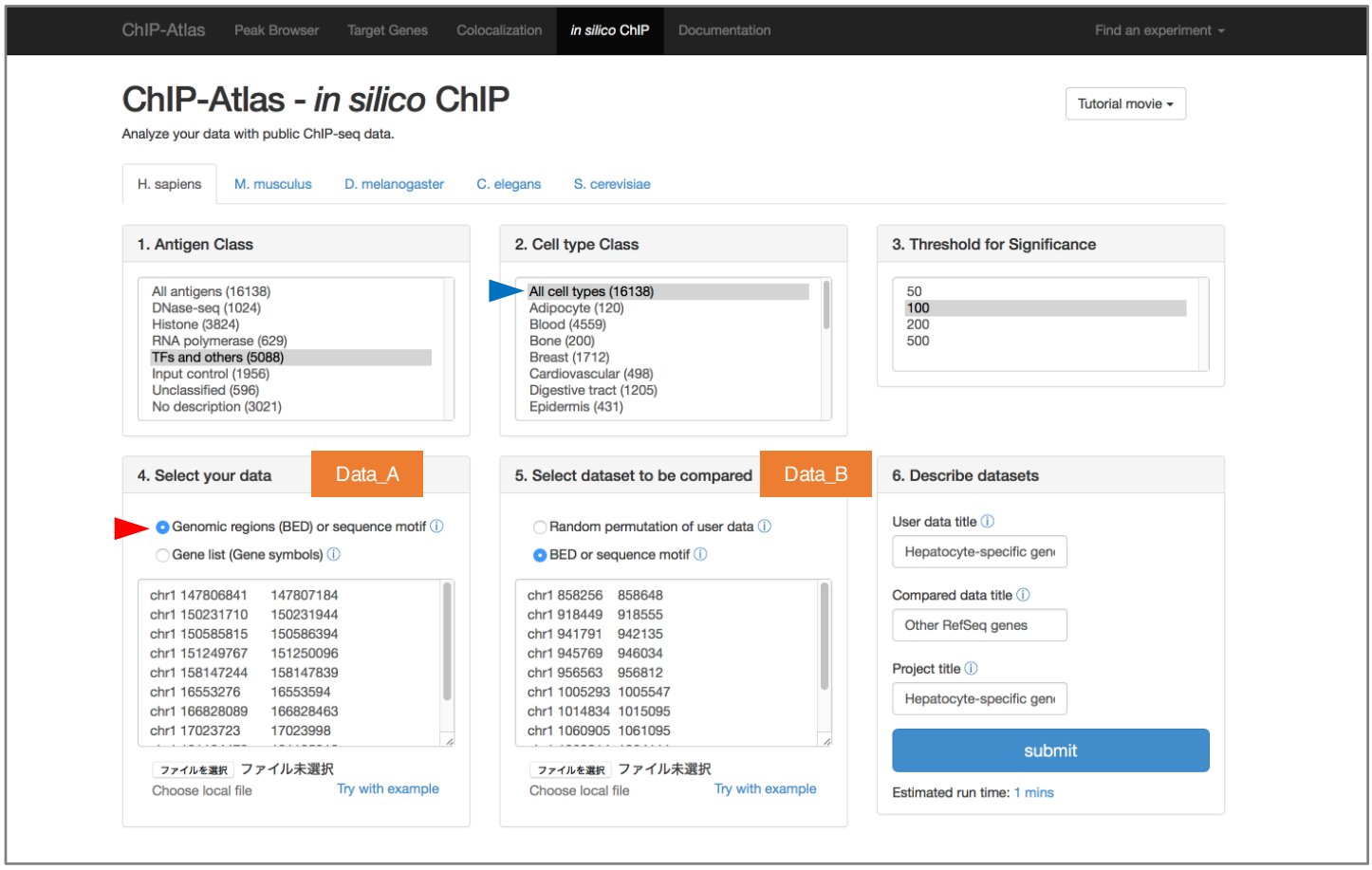

ChIP-Atlas
Peak Browser
Target Genes
Colocalization
in silico ChIP
Documentation
Find an experiment
ChIP-Atlas - Colocalization
Predict colocalization partners of TFs.
Tutorial movie
H. sapiens
M. musculus
D. melanogaster
C. elegans
S. cerevisiae
1. Search mode
Antigens → Cell Type
Cell Type → Antigen
2. Choose Antigen
type to search
Myog
Myt1l
Nanog
Nbn
Ncapd3
Ncapg
Ncaph2
Ncoa2
3. Choose Cell Type Class
type to search
Pluripotent stem cell

B


Cell types
Exp. IDs
Nanog's Colocalization partners
Nanog: Average
SRX248287: Embryoid Bodies
SRX003864: Embryonic Stem Cells
SRX003865: Embryonic Stem Cells
SRX213808: Embryonic Stem Cells
SRX213820: Embryonic Stem Cells
SRX213829: Embryonic Stem Cells
SRX236478: Embryonic Stem Cells
SRX248286: Embryonic Stem Cells
SRX319957: Embryonic Stem Cells
SRX651982: Embryonic Stem Cells
Nanog: STRING
Embryonic Stem Cells SRX319556 Nr0b1
Embryonic Stem Cells SRX319555 Nr0b1
Embryonic Stem Cells SRX695695 Zic2
Embryoid Bodies SRX248287 Nanog
Embryonic Stem Cells SRX236478 Nanog
Embryonic Stem Cells SRX213829 Nanog
Embryonic Stem Cells SRX213808 Nanog
Embryonic Stem Cells SRX003865 Nanog
Embryonic Stem Cells SRX003864 Nanog
Embryonic Stem Cells SRX1410236 Sox2
EpiLC SRX499130 Otx2
EpiLC SRX499129 Otx2
EpiLC SRX499126 Ep300
EpiLC SRX499125 Ep300
Embryonic Stem Cells SRX499117 Otx2
Embryonic Stem Cells SRX651982 Nanog
Embryonic Stem Cells SRX319957 Nanog
Embryonic Stem Cells SRX433130 Sox2
Embryonic Stem Cells SRX236477 Sox2
Embryonic Stem Cells SRX214084 Pou5f1
Embryonic Stem Cells SRX214083 Pou5f1
Embryonic Stem Cells SRX213830 Pou5f1
Embryonic Stem Cells SRX213809 Pou5f1
mESCs, differentiated SRX209329 Sox2
Embryonic Stem Cells SRX209316 Sox2
Embryonic Stem Cells SRX695696 Zic2
Embryonic Stem Cells SRX499113 Ep300
Embryonic Stem Cells SRX194543 Ep300
Embryonic Stem Cells SRX194542 Ep300
Embryonic Stem Cells SRX194541 Ep300
Embryonic Stem Cells SRX194534 Ep300
Embryonic Stem Cells SRX194533 Ep300
Embryonic Stem Cells SRX147572 Zic3
Embryonic Stem Cells SRX027354 Ep300
Embryonic Stem Cells SRX499114 Pou5f1
iPS cells SRX977381 Pou5f1
iPS cells SRX977380 Pou5f1
Embryonic Stem Cells SRX499140 Pou5f1
Same SRX
None
CoLo score
1
9

A

Fig .7


ChIP-Atlas
Peak Browser
Target Genes
Colocalization
in silico ChIP
Documentation
Find an experiment
ChIP-Atlas - in silico ChIP
Analyze your data with public ChIP-seq data.
Tutorial movie
H. sapiens
M. musculus
D. melanogaster
C. elegans
S. cerevisiae
1. Antigen Class
All antigens (16138)
DNase-seq (1024)
Histone (3824)
RNA polymerase (629)
TFs and others (5088)
Input control (1956)
Unclassified (596)
No description (3021)
2. Cell type Class
All cell types (16138)
Adipocyte (120)
Blood (4559)
Bone (200)
Breast (1712)
Cardiovascular (498)
Digestive tract (1205)
Epidermis (431)
3. Threshold for Significance
50
100
200
500
4. Select your data
Data_A
Genomic regions (BED) or sequence motif
Gene list (Gene symbols)
chr1 147806841 147807184
chr1 150231710 150231944
chr1 150585815 150586394
chr1 151249767 151250096
chr1 158147244 158147839
chr1 16553276 16553594
chr1 166828089 166828463
chr1 17023723 17023998
ファイルを選択 ファイル未選択
Choose local file
Try with example
5. Select dataset to be compared
Data_B
Random permutation of user data
BED or sequence motif
chr1 858256 858648
chr1 918449 918555
chr1 941791 942135
chr1 945769 946034
chr1 956563 956812
chr1 1005293 1005547
chr1 1014834 1015095
chr1 1060905 1061095
ファイルを選択 ファイル未選択
Choose local file
Try with example
6. Describe datasets
User data title
Hepatocyte-specific gen
Compared data title
Other RefSeq genes
Project title
Hepatocyte-specific gen
submit
Estimated run time: 1 mins

B

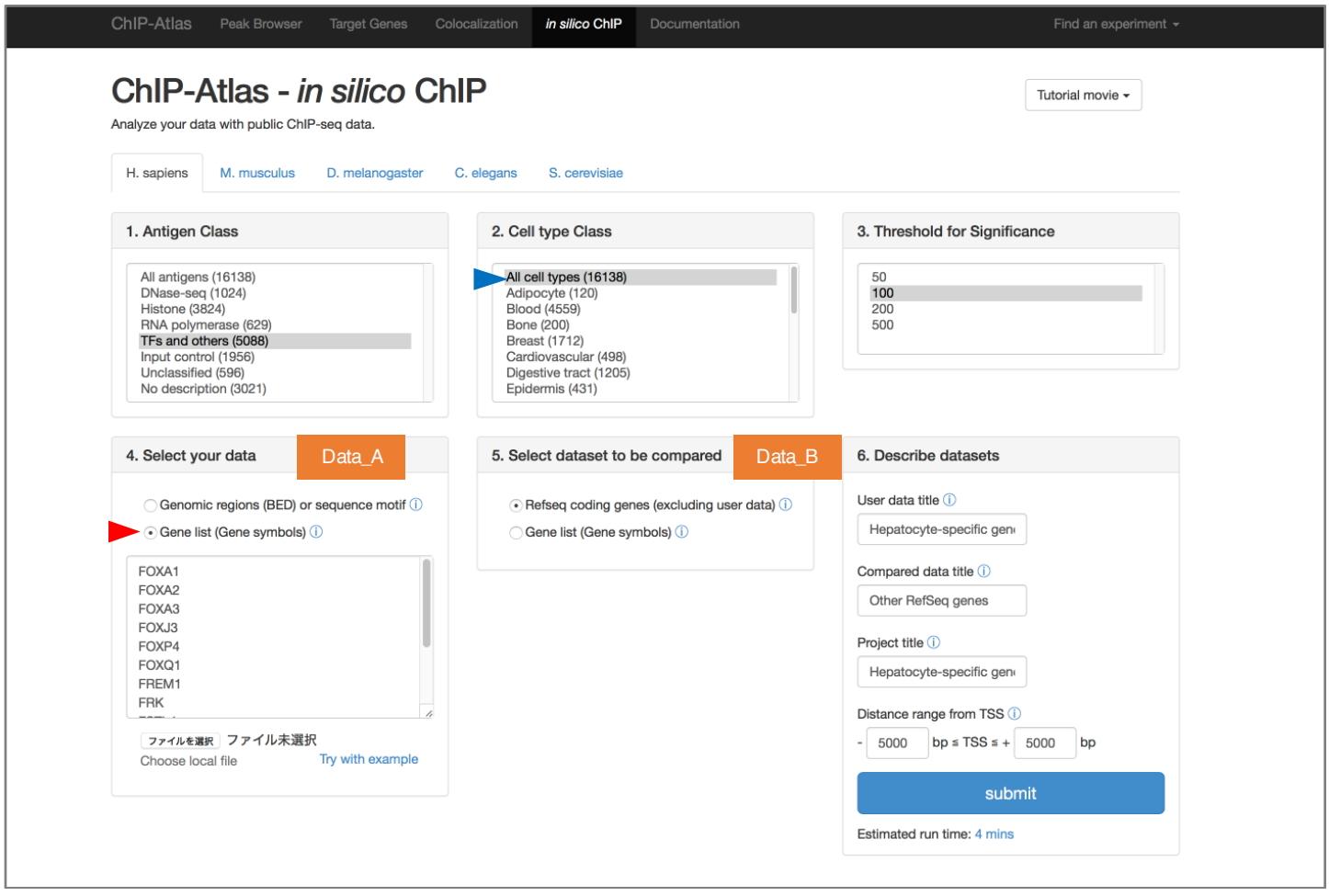

ChIP-Atlas
Peak Browser
Target Genes
Colocalization
in silico ChIP
Documentation
Find an experiment
ChIP-Atlas - in silico ChIP
Analyze your data with public ChIP-seq data.
Tutorial movie
H. sapiens
M. musculus
D. melanogaster
C. elegans
S. cerevisiae
1. Antigen Class
All antigens (16138)
DNase-seq (1024)
Histone (3824)
RNA polymerase (629)
TFs and others (5088)
Input control (1956)
Unclassified (596)
No description (3021)
2. Cell type Class
All cell types (16138)
Adipocyte (120)
Blood (4559)
Bone (200)
Breast (1712)
Cardiovascular (498)
Digestive tract (1205)
Epidermis (431)
3. Threshold for Significance
50
100
200
500
4. Select your data
Data_A
Genomic regions (BED) or sequence motif
Gene list (Gene symbols)
FOXA1
FOXA2
FOXA3
FOXJ3
FOXP4
FOXQ1
FREM1
FRK
ファイルを選択 ファイル未選択
Choose local file
Try with example
5. Select dataset to be compared
Data_B
Refseq coding genes (excluding user data)
Gene list (Gene symbols)
6. Describe datasets
User data title
Hepatocyte-specific gen
Compared data title
Other RefSeq genes
Project title
Hepatocyte-specific gen
Distance range from TSS
- 5000 bp ≤ TSS ≤ + 5000 bp
submit
Estimated run time: 4 mins

Fig.8

ChIP-Atlas / in silico ChIP

Search for proteins significantly bound to your data.

Show 100 entries

Search:

Hepatocyte vs Others

ID	Antigen class	Antigen	Cell calss	Cell	Num of peaks	Overlaps / Hepatocyte	Overlaps / Others	Log p-val	Log q-val	Fold Enrichment	FE > 1?
SRX100544	TFs and others	EP300	Liver	Hep G2	24334	80/286	1147/20509	-32.1	-28.5	5.00	TRUE
SRX190321	TFs and others	MAX	Liver	Hep G2	40220	90/286	1497/20509	-31.9	-28.5	4.31	TRUE
SRX100552	TFs and others	SP1	Liver	Hep G2	19032	64/286	755/20509	-29.7	-26.5	6.08	TRUE
SRX100449	TFs and others	HNF4G	Liver	Hep G2	15919	54/286	507/20509	-29.3	-26.2	7.64	TRUE
SRX100497	TFs and others	RXRA	Liver	Hep G2	13022	54/286	557/20509	-27.5	-24.5	6.95	TRUE
SRX100493	TFs and others	HEY1	Liver	Hep G2	26412	69/286	984/20509	-27.5	-24.5	5.03	TRUE
SRX100538	TFs and others	HDAC2	Liver	Hep G2	16071	58/286	676/20509	-27.0	-24.2	6.15	TRUE
SRX100505	TFs and others	HNF4A	Liver	Hep G2	21259	54/286	585/20509	-26.5	-23.7	6.62	TRUE
SRX190332	TFs and others	MYBL2	Liver	Hep G2	15213	55/286	637/20509	-25.7	-22.9	6.19	TRUE
SRX190264	TFs and others	CREB1	Liver	Hep G2	26690	68/286	1092/20509	-24.3	-21.5	4.47	TRUE
SRX100448	TFs and others	FOXA2	Liver	Hep G2	45130	67/286	1138/20509	-22.6	-19.9	4.22	TRUE
SRX150360	TFs and others	TBP	Liver	Hep G2	10293	38/286	327/20509	-21.7	-19.1	8.33	TRUE
SRX100506	TFs and others	FOXA1	Liver	Hep G2	50941	70/286	1295/20509	-21.7	-19.1	3.88	TRUE
SRX1165097	TFs and others	CREB1	Liver	Hep G2	21856	58/286	899/20509	-21.2	-18.6	4.63	TRUE
SRX190266	TFs and others	NR2F2	Liver	Hep G2	18201	48/286	605/20509	-20.9	-18.4	5.69	TRUE
SRX100477	TFs and others	FOXA1	Liver	Hep G2	40732	66/286	1198/20509	-20.8	-18.2	3.95	TRUE
SRX1165103	TFs and others	Epitope tags	Liver	Hep G2	14910	48/286	693/20509	-18.6	-16.1	4.97	TRUE
SRX190197	TFs and others	NFIC	Liver	Hep G2	13273	42/286	518/20509	-18.6	-16.1	5.81	TRUE
SRX150701	TFs and others	CEBPB	Liver	Hep G2	18637	52/286	849/20509	-18.0	-15.6	4.39	TRUE
SRX100545	TFs and others	JUND	Liver	Hep G2	19875	62/286	1223/20509	-17.7	-15.3	3.64	TRUE
SRX150516	TFs and others	MXI1	Liver	Hep G2	18136	47/286	729/20509	-17.1	-14.7	4.62	TRUE
SRX100412	TFs and others	TAF1	Liver	Hep G2	17725	38/286	492/20509	-16.2	-13.8	5.54	TRUE
SRX150355	TFs and others	ARID3A	Liver	Hep G2	13508	36/286	440/20509	-16.0	-13.7	5.87	TRUE
SRX026268	TFs and others	HNF4A	Digestive tract	Caco-2	12533	33/286	372/20509	-15.7	-13.3	6.36	TRUE

Showing 1 to 100 of 5,184 entries

Previous 1 2 3 4 5 … 52 Next

Fig.9

ChIP-Atlas / in silico ChIP

Search for proteins significantly bound to your data.

Show 100 entries

Search:

Hepatocyte-specific genes vs Other RefSeq genes

ID	Antigen class	Antigen	Cell calss	Cell	Num of peaks	Overlaps / Hepatocyte-specific genes	Overlaps / Other RefSeq genes	Log p-val	Log q-val	Fold Enrichment	FE > 1?
SRX100505	TFs and others	HNF4A	Liver	Hep G2	21259	584/964	4595/17658	-105.3	-101.5	2.33	TRUE
SRX100449	TFs and others	HNF4G	Liver	Hep G2	15919	539/964	3961/17658	-104.1	-100.7	2.49	TRUE
SRX100497	TFs and others	RXRA	Liver	Hep G2	13022	429/964	3058/17658	-79.2	-75.9	2.57	TRUE
SRX100544	TFs and others	EP300	Liver	Hep G2	24334	561/964	5018/17658	-77.3	-74.1	2.05	TRUE
SRX150698	TFs and others	HNF4A	Liver	Hep G2	10069	415/964	3004/17658	-73.9	-70.9	2.53	TRUE
SRX100448	TFs and others	FOXA2	Liver	Hep G2	45130	644/964	7018/17658	-60.5	-57.5	1.68	TRUE
SRX100477	TFs and others	FOXA1	Liver	Hep G2	40732	629/964	6809/17658	-58.9	-56.0	1.69	TRUE
SRX018625	TFs and others	HNF4A	Liver	Hep G2	2654	167/964	717/17658	-50.4	-47.6	4.27	TRUE
SRX100506	TFs and others	FOXA1	Liver	Hep G2	50941	662/964	7813/17658	-49.5	-46.7	1.55	TRUE
SRX100538	TFs and others	HDAC2	Liver	Hep G2	16071	501/964	5083/17658	-47.9	-45.2	1.81	TRUE
SRX100552	TFs and others	SP1	Liver	Hep G2	19032	577/964	6352/17658	-47.8	-45.2	1.66	TRUE
SRX190234	TFs and others	CEBPB	Liver	Hep G2	12482	357/964	3125/17658	-42.5	-39.9	2.09	TRUE
SRX190331	TFs and others	TEAD4	Liver	Hep G2	10956	291/964	2280/17658	-41.2	-38.6	2.34	TRUE
SRX018626	TFs and others	HNF4A	Liver	Hep G2	1639	121/964	456/17658	-40.8	-38.2	4.86	TRUE
SRX150355	TFs and others	ARID3A	Liver	Hep G2	13508	326/964	2832/17658	-38.5	-36.0	2.11	TRUE
SRX190197	TFs and others	NFIC	Liver	Hep G2	13273	403/964	4094/17658	-34.8	-32.3	1.80	TRUE

Showing 1 to 100 of 5,158 entries

Previous 1 2 3 4 5 … 52 Next

§５ 研究開発計画に対する達成状況と将来展望

（1）達成状況

「**§３ (1)当初の研究開発計画**」の項目をすべて達成し、本データベースを 2015 年 12 月に一般公開した。また、「**§３ (2)新たに追加・修正など変更した研究開発計画**」に記載したように、ユーザデータを解析するツール（in silico ChIP）を追加したほか、それを用いた情報解析データを作成している。前者のツールはすでに公開しており、後者は解析結果を公開するための web インターフェースを作成中である。上記の追加事項により、当初目指していた研究期間内の論文投稿は果たせなかったが、本データベースの向上のためにはやむを得なかったと考えている。現在、原稿を執筆中であり、早い時期に投稿したいと考えている。

（2）ツール等の将来展望

本データベースはFig. 1に示すように、5つの生物種のすべての ChIP-seq データをカバーしており、それらはすべての ChIP-seq データのデータ数において上位 5 位を占めている。しかし、その他の生物種を追加して欲しいとの要望がすでに得られており、中でもラットや植物（とくにシロイヌナズナ）への要望が多い。将来的には解析パイプラインの改善や、メタデータクレンジングの基準策定をおこない、要望にいち早く対応できるような枠組みを整備したい。

本データベースの in silico ChIP を利用し、これまでに組織特異的遺伝子やエンハンサーに結合が enrichment するような転写因子が得られている。これらは組織特異的な遺伝子発現をまとめて司る転写因子である可能性が高く、実際に既知のマスターレギュレータやダイレクトリプログラミング因子を多く含んでいる。近いうちにこれらのデータを公開することにより、遺伝子制御のハブ因子の特定や、様々な組織へのダイレクトリプログラミング因子の候補付けに大いに貢献できるものと期待される。また GWAS データに対しても同様の解析をおこなっているが、非常に興味深いデータが得られている。たとえば炎症性大腸炎や I 型糖尿病に好発する変異部位のうち、非遺伝子領域について解析したところ、血球分化に重要な転写因子が複数得られている。一見、無関係のように思えるが、じつは両疾患の主要な原因は自己免疫疾患である（しかし詳しい原因はよくわかっていない）。ゲノムワイドな変異同定と疾患との関連はさかんに研究されているが、連鎖不平衡による変異部位や passenger mutation なども多く、主要な原因となる driver mutation を見つけ出すのが困難である。そのような現状において、in silico ChIP は転写因子の結合情報をもたらすことにより、主要な変異部位を正確に同定できると期待される。そこで得られた情報は、統合化推進プログラムで統合されたデータベース（GWAS-DB, SNP-DB など）のデータに連携できればと考えている。将来的には GWAS データだけでなく、TCGA などの大規模ながん特異的変異部位にも応用したい。また、ヒトの遺伝的多様性（1000 genome project）や種間比較（Neanderthal genome project, PhastCons）など、さまざまな応用を想定している。

§６ 研究参加者

氏名	所属	役職	研究開発項目	参加時期
○沖 真弥	九大・院・医	助教	研究遂行、統括	H27.5-H28.3

§７ 成果発表等

（1)原著論文発表 （国内（和文）誌 0 件、国際（欧文）誌 0 件）

該当なし

(2)その他の著作物(総説、書籍など)
該当なし

(3)国際学会発表及び主要な国内学会発表
① 招待講演 (国内会議 0 件、国際会議 0 件)
該当なし

② 口頭発表 (国内会議 4 件、国際会議 0 件)
1. Shinya Oki(九大), Tazro Ohta(DBCLS),Go Shioi(RIKEN), Chikara Meno(九大)、**ChIP-Atlas: Comprehensive database for visualizing all published ChIP-seq data**、48th Annual Meeting of JSDB、つくば市、20015/6/2
2. 沖 真弥(九大)、**既報の ChIP-seq データをフル活用するための統合データベース**、生命情報科学若手の会 第 7 回研究会、鶴岡市、2015/10/1
3. 沖 真弥(九大)、**ChIP-seq データベースのためのメタ情報アノテーション**、Annotathon 2015、三島、2015/11/12
4. 沖 真弥(九大), 大田 達郎(DBCLS), 塩井 剛(RIKEN), 仲木 竜(東大), 目野 主税(九大)、**既報の ChIP-seq データをフル活用するための統合データベース**、第 38 回日本分子生物学会年会、横浜、2015/12/1

③ ポスター発表 (国内会議 1 件、国際会議 1 件)
1. 沖 真弥(九大), 大田 達郎(DBCLS), 塩井 剛(RIKEN), 目野 主税(九大)、**ChIP-seq SRA を利活用するための統合データベース**、NGS 現場の会第四回研究会、つくば市、2015/7/1
2. Shinya Oki (九大), Tazro Ohta(DBCLS), Go Shioi (RIKEN), Ryo Nakaki (東大), Osamu Ogasawara (DDBJ), Yoshihiro Okuda (DDBJ), Hideki Hatanaka (NBDC), Chikara Meno (九大). **ChIP-Atlas: Comprehensive and integrative database for visualizing and mining all published ChIP-seq data. SYSTEMS BIOLOGY: GLOBAL REGULATION OF GENE EXPRESSION, Cold spring harbor laboratory**, 2016/3/17

(4)知財出願
該当なし

(5)受賞・報道等
該当なし

§9 自己評価

「§2 研究開発のねらい」の1点目、「誰もが簡単に利活用できるようなデータベース」に関してはおおむね達成できたと考えている。ユーザに利用の感想を聞いているが、大幅な改善を望むような意見は得られておらず、また、ChIP-Atlas に設置したメールフォームからもそのような意見は得られていない。2点目の「メタ情報のクレンジング」についても、今のところ大きな修正を望むような意見は得られていないため、ある程度達成できたと考えている。しかし、クレンジング方法にわりと客観的な基準を設けたとは言え、やはり自分自身の主観や知識の偏りによるバイアスがあることは否定できない。研究代表者はマウスの発生学と細胞分化を専門としているが、その他の研究分野や生物種における知識に自信がないためである。しかし、誰もが納得できるような表記法や分類法

はなく、また自分一人で決められるものでもないため、今後は多くの研究者やSRA関係者との合意形成が必要ではないかと考えている。

以上