

2017. 07. 18

統合化推進プログラム 2017 年度

キックオフ・ミーティング

# エピゲノム統合データベースの 開発と機能拡充

九州大学大学院・医学研究院  
発生再生医学分野・助教

**沖 真弥**

九州大学大学院・医学研究院  
医化学分野・講師

**三浦 史仁**

国立がん研究センター  
がん分子修飾制御学分野・分野長

**浜本 隆二**



## 既報の ChIP-seq データの網羅的かつ統合的データベース

[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](#)

### Funded by:

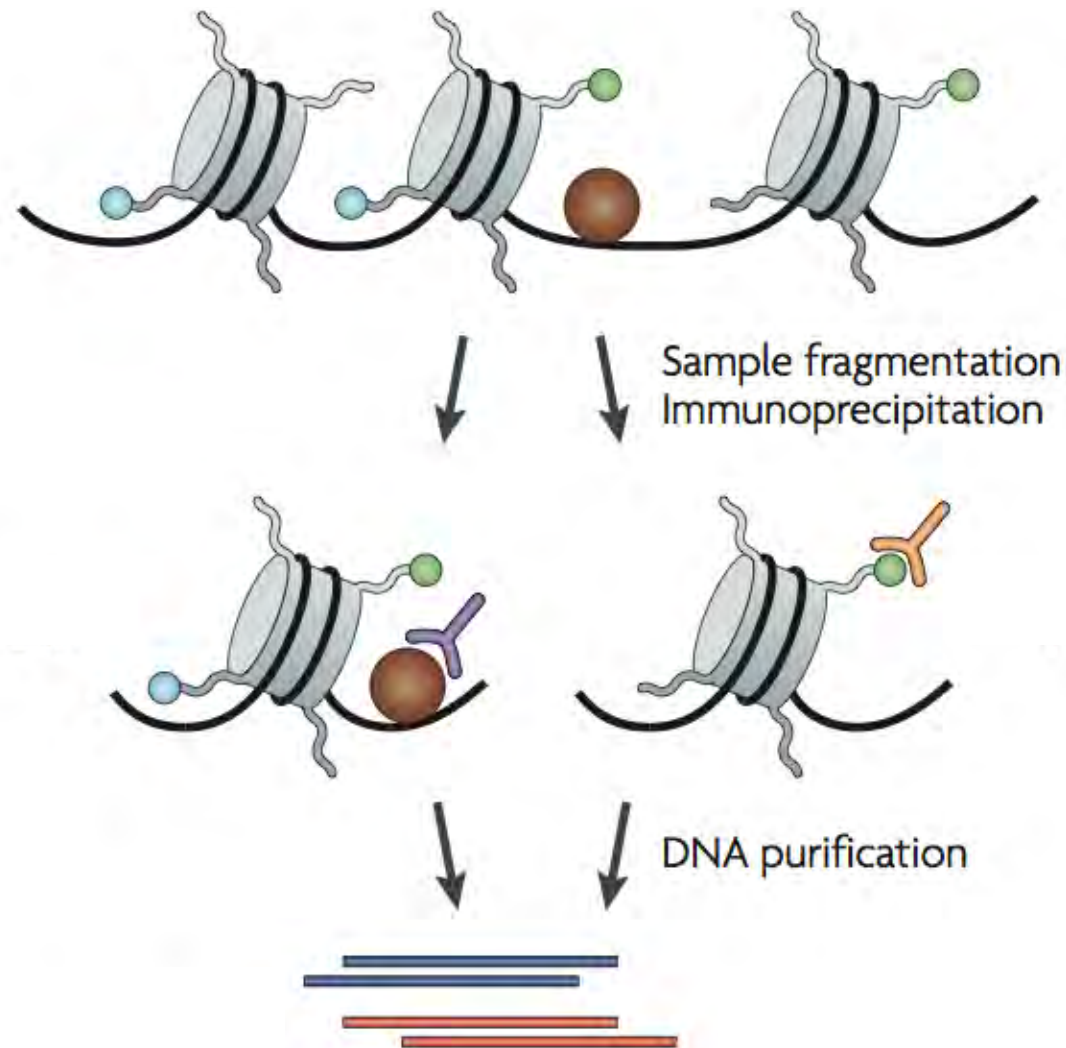
平成 27 年度 JST NBDC 統合化推進プログラム  
(統合データ解析トライアル)

### Collaborations with:

DBCLS, NBDC, DDBJ, RIKEN, 東大, 産総研, 九大

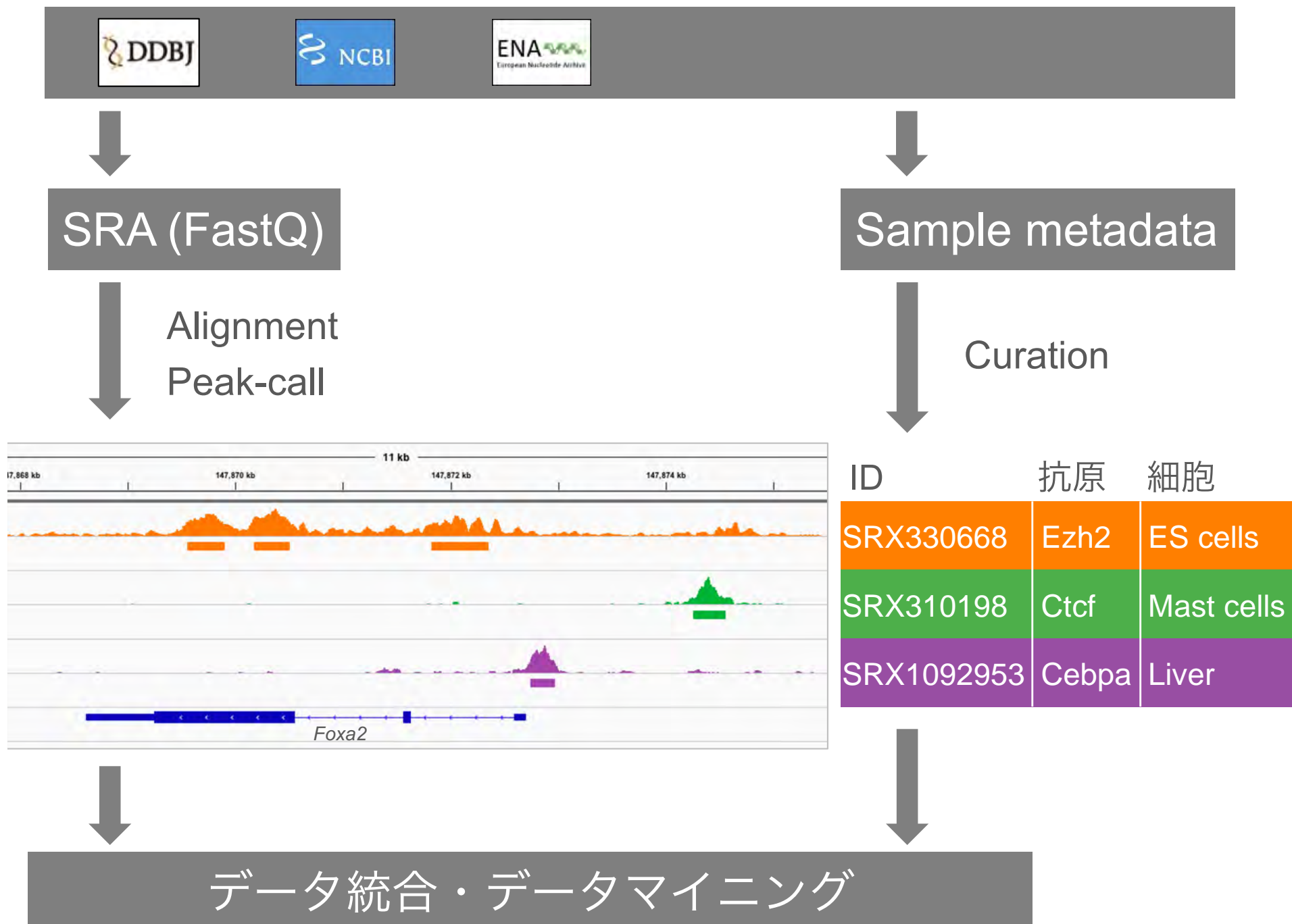
# ChIP-seq: 転写因子の結合部位を同定できる

ChIP-seq = **C**hromatin **I**mmunoprecipitation with **S**equencing



↓  
High-throughput sequencing

# データ処理・公開しているデータ



# 4つの機能

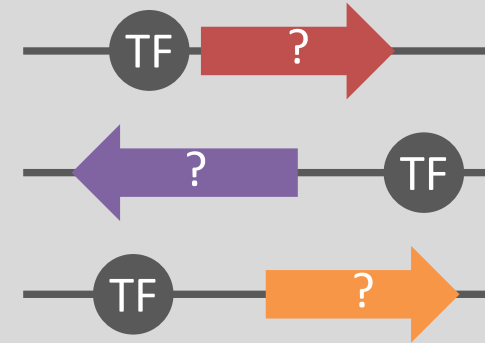
## ① Peak Browser

何がどこに結合する？



## ② Target Genes

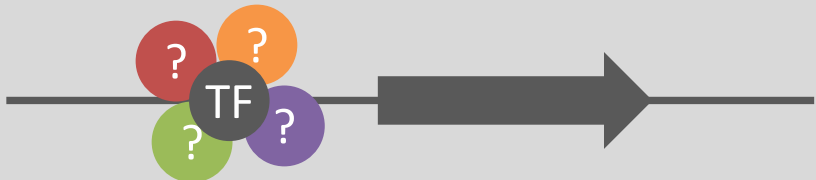
転写因子の標的遺伝子を探す。



# ChIP-Atlas

## ③ Colocalization

共局在する転写因子を探す。

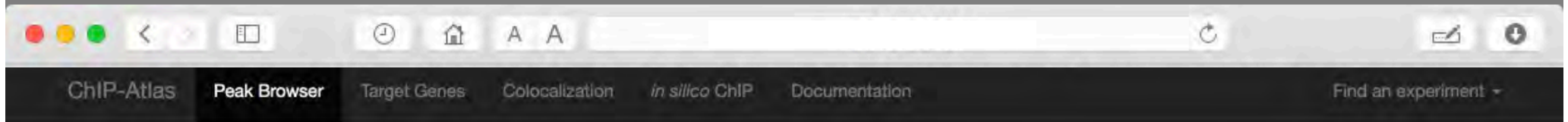


## ④ in silico ChIP

ゲノム領域への Enrichment 解析



# Peak Browser



## CHIP-Atlas - Peak Browser

Tutorial movies ▾

Visualize All Peaks from Published ChIP-Seq data.

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)

Antigen

### Cell type Class

**All cell types (16138)**  
Adipocyte (120)  
Blood (4559)  
Bone (200)  
Breast (1712)  
Cardiovascular (498)  
Digestive tract (1205)  
Epidermis (431)

Cell type

### Threshold for Significance

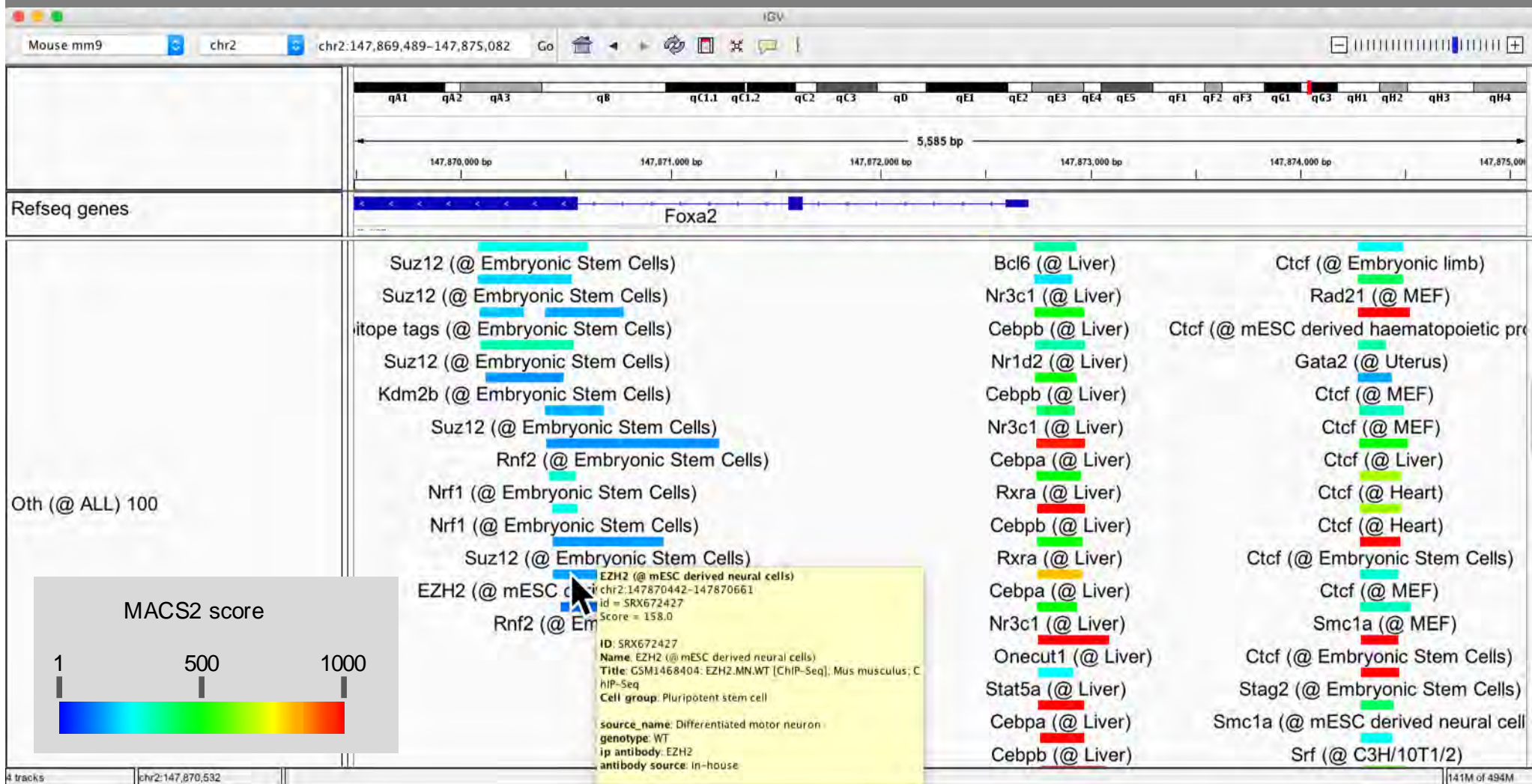
50  
100  
200  
500

View on IGV

構想：塩井 剛 (RIKEN, CLST)  
UI：大田 達郎 (DBCLS)

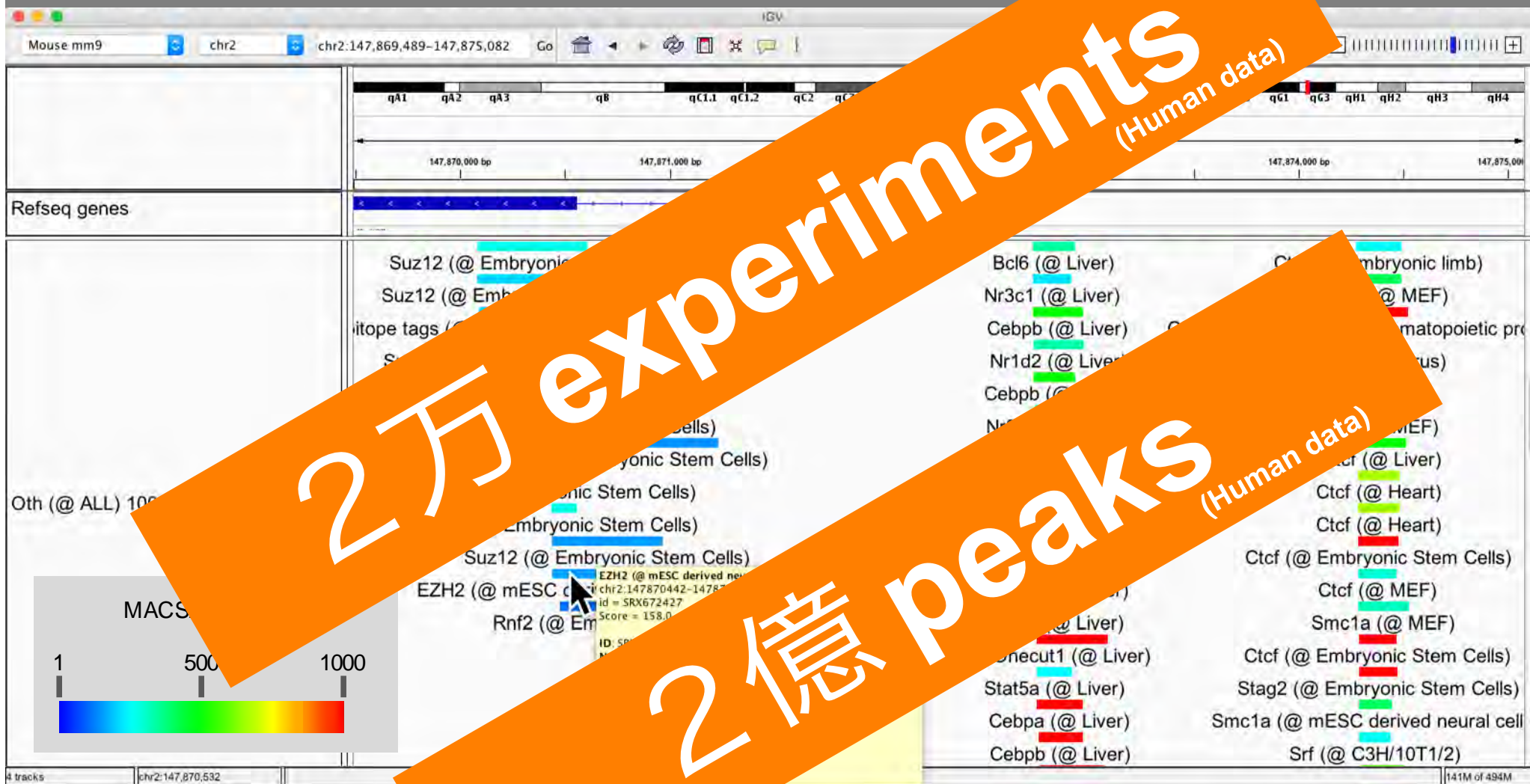


# Peak Browser



- 「なにがどこに結合するか？」がすべてわかる。
- シス調節領域の特定に応用できる。

# Peak Browser



- 「なにがどこに...」がすべてわかる。
- シス調節領域の特定に応用できる。



# Target Genes

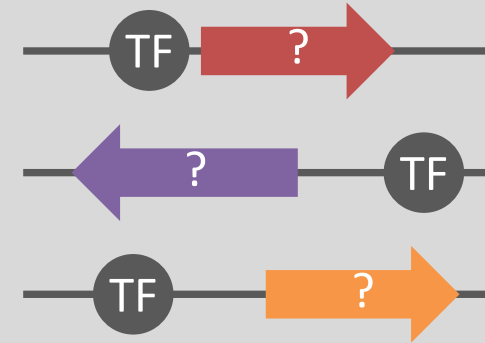
## ① Peak Browser

何がどこに結合する？



## ② Target Genes

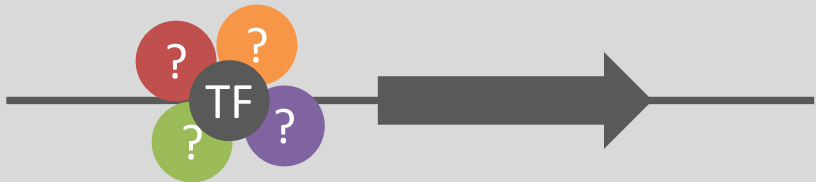
転写因子の標的遺伝子を探す。



# ChIP-Atlas

## ③ Colocalization

共局在する転写因子を探す。



## ④ in silico ChIP

ゲノム領域への Enrichment 解析



# Target Genes

[ChIP-Atlas](#)[Peak Browser](#)[Target Genes](#)[Colocalization](#)[in silico ChIP](#)[Documentation](#)[Find an experiment](#) ▾

## ChIP-Atlas - Target Genes

[Tutorial movie](#) ▾

Predict potential target genes of TFs.

[H. sapiens](#)[M. musculus](#)[D. melanogaster](#)[C. elegans](#)[S. cerevisiae](#)

### 1. Choose Antigen

PML  
POU2AF1  
POU2F1  
POU2F2  
POU3F2  
**POU5F1**  
PPARA  
PPARG

### 2. Choose Distance from TSS

 ±1k ±5k ±10k[View Potential Target Genes](#)[Download \(TSV\)](#)

KYUSHU UNIVERSITY

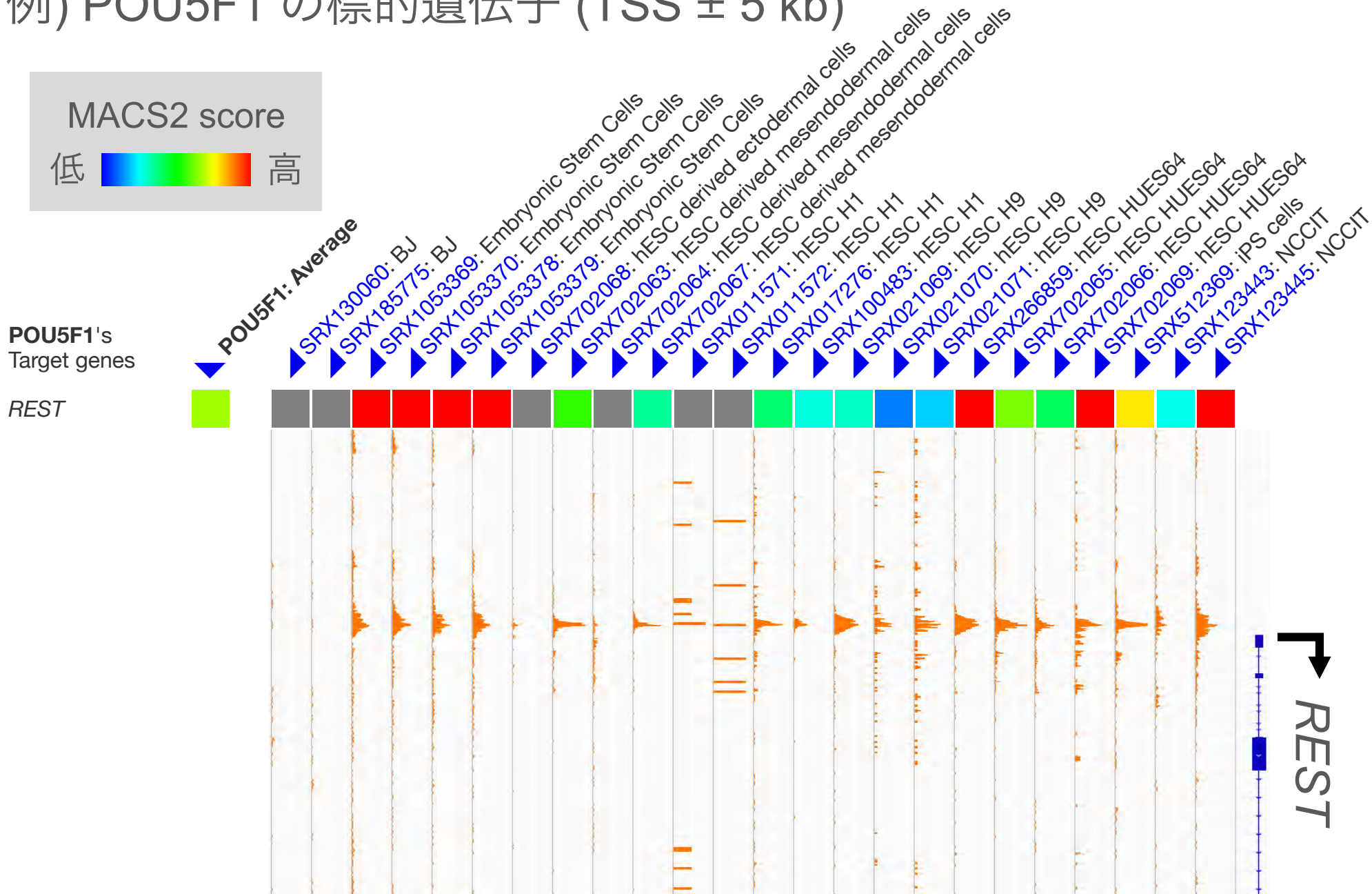
THIS WORK IS SUPPORTED BY NIG SUPERCOMPUTER SYSTEM AND NATIONAL BIOSCIENCE DATABASE CENTER.

NEED HELP? CREATE AN ISSUE ON GITHUB OR CONTACT US



# Target Genes

例) POU5F1 の標的遺伝子 (TSS  $\pm$  5 kb)



# Colocalization

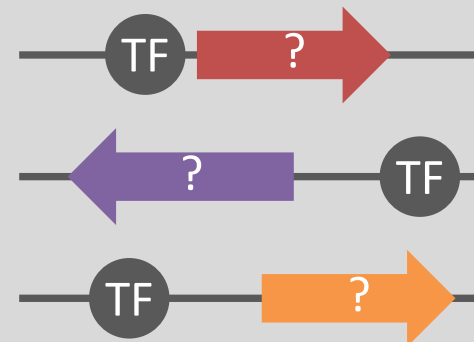
## ① Peak Browser

何がどこに結合する？



## ② Target Genes

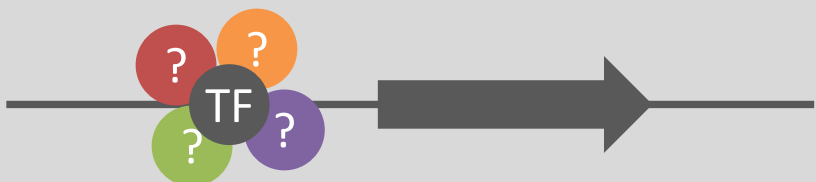
転写因子の標的遺伝子を探す。



# ChIP-Atlas

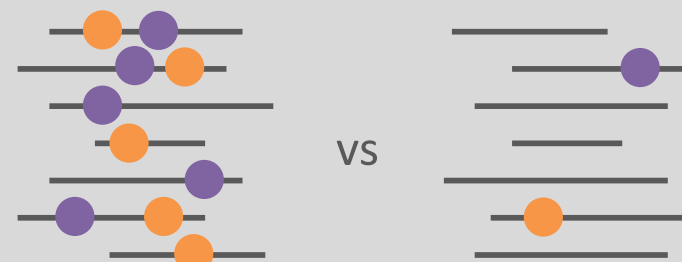
## ③ Colocalization

共局在する転写因子を探す。



## ④ in silico ChIP

ゲノム領域への Enrichment 解析



解析ツール (CoLo) :  
仲木 竜 (現 Rhelixa 社長)



# Colocalization

[ChIP-Atlas](#)[Peak Browser](#)[Target Genes](#)[Colocalization](#)[in silico ChIP](#)[Documentation](#)[Find an experiment](#) ▾

## ChIP-Atlas - Colocalization

[Tutorial movie](#) ▾

Predict colocalization partners of TFs.

[H. sapiens](#)[M. musculus](#)[D. melanogaster](#)[C. elegans](#)[S. cerevisiae](#)

### 1. Search mode

- Antigen → Cell Type  
 Cell Type → Antigen

### 2. Choose Antigen

Nacc1  
**Nanog**  
Nap111  
Nbn  
Ncapd3  
Ncapg  
Ncaph2  
Ncoa2

### 3. Choose Cell Type Class

Liver  
**Pluripotent stem cell**

[View Colocalization Data](#)[Download \(TSV\)](#)[Download \(GML\)](#)

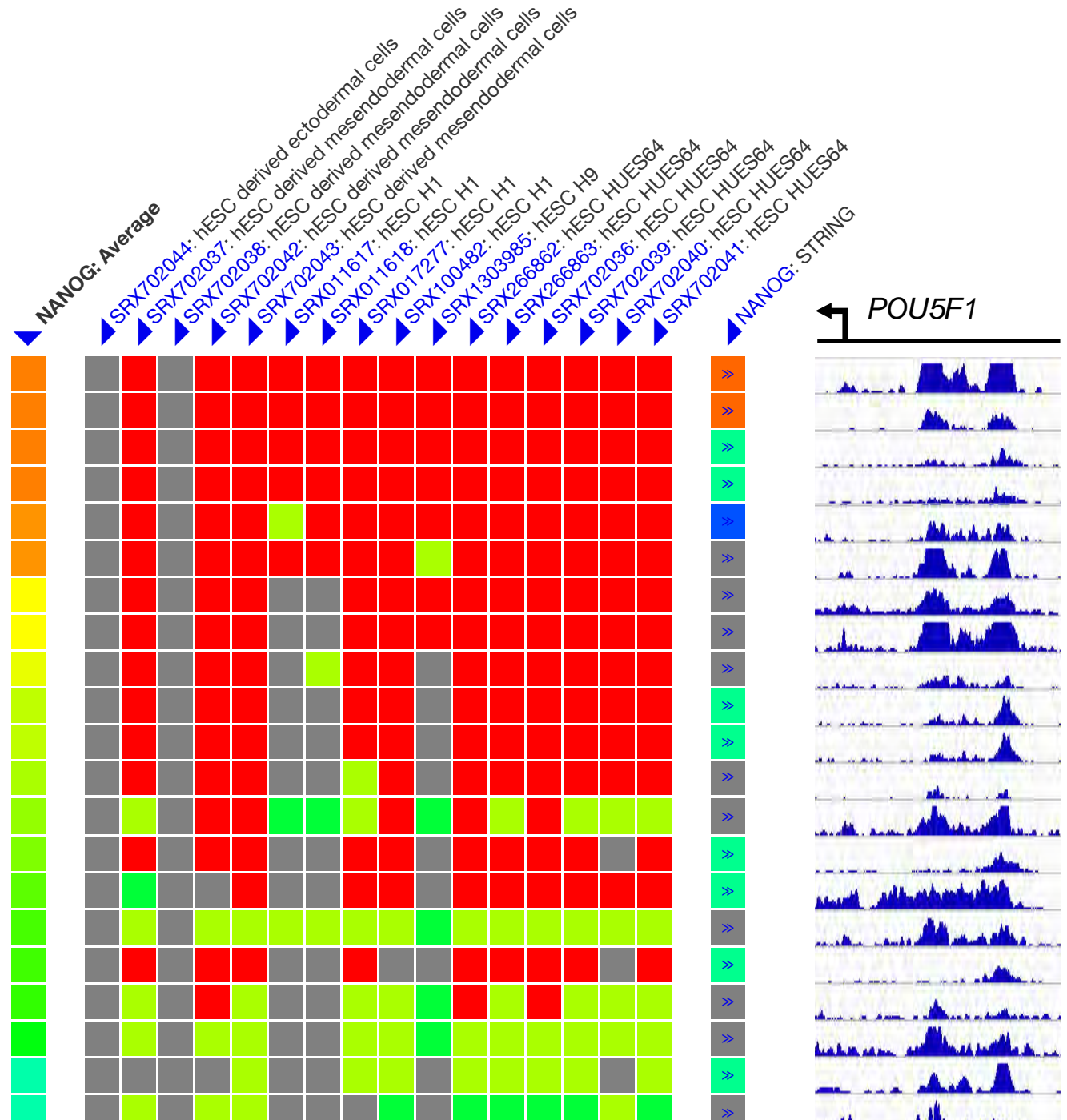
# Colocalization

類似度



(CoLo; 仲木 竜氏, unpbl)

Cell types	Exp. IDs	NANOG's Colocalization partners
iPS cells	SRX512370	SOX2
hESC H9	SRX1047413	SOX2
Embryonic Stem Cells	SRX1053379	POU5F1
Embryonic Stem Cells	SRX1053378	POU5F1
hESC H9	SRX027488	SMARCA4
hESC H1	SRX833403	CTNNB1
hESC H1	SRX833420	NIPBL
hESC H9	SRX027482	EP300
hESC HUES64	SRX702045	OTX2
hESC HUES64	SRX702069	POU5F1
hESC HUES64	SRX266859	POU5F1
hESC HUES64	SRX702132	TCF4
hESC H1	SRX833422	NIPBL
Embryonic Stem Cells	SRX1053369	POU5F1
NCCIT	SRX123445	POU5F1
hESC H1	SRX833421	NIPBL
Embryonic Stem Cells	SRX1053370	POU5F1
hESC H1	SRX833401	LEF1
hESC H1	SRX833423	NIPBL
iPS cells	SRX512369	POU5F1
hESC derived mesendoderm	SRX684516	T



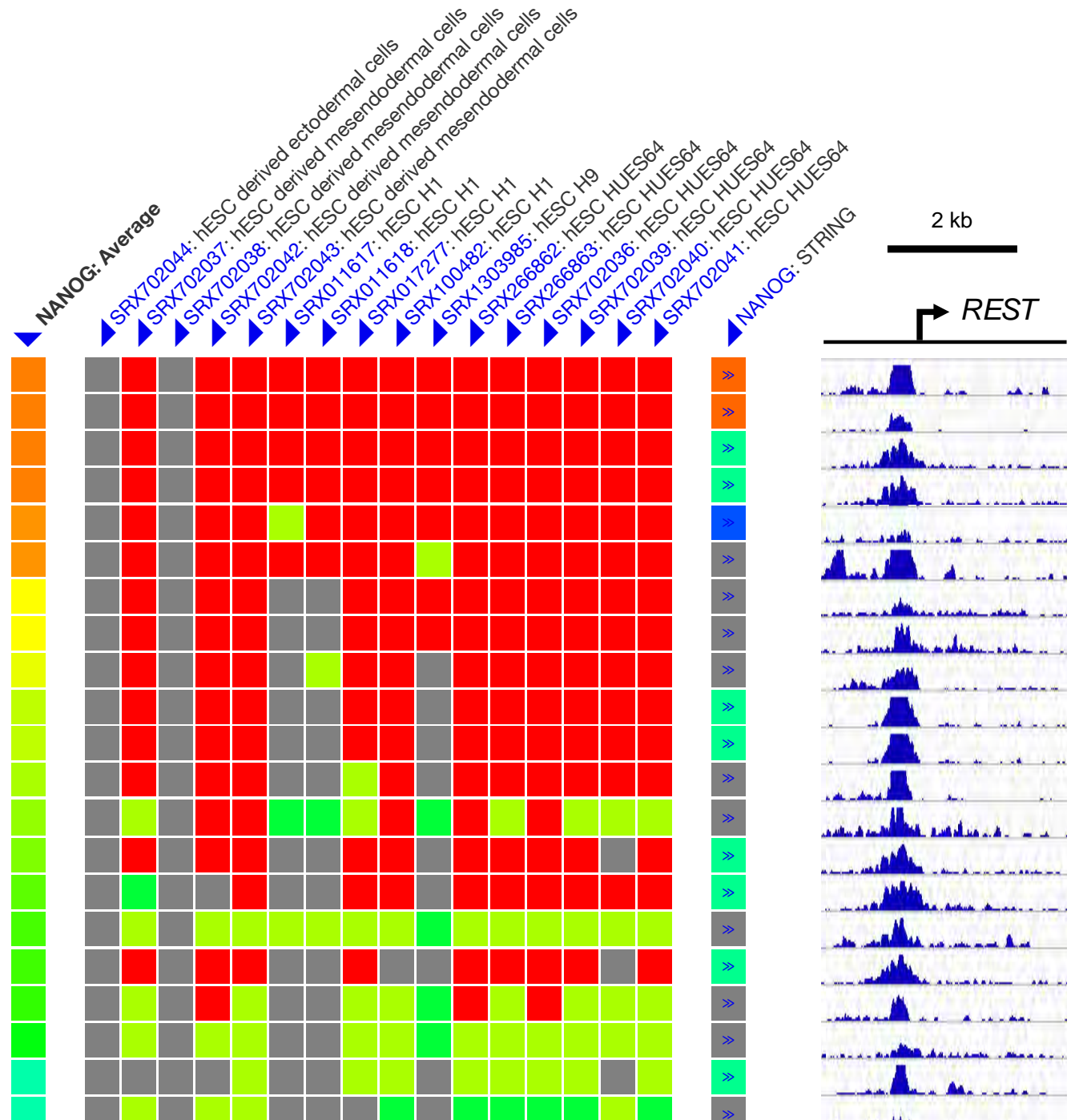
# Colocalization

類似度

低  高

(CoLo; 仲木 竜氏, unpubl)

Cell types	Exp. IDs	NANOG's Colocalization partners
iPS cells	SRX512370	SOX2
hESC H9	SRX1047413	SOX2
Embryonic Stem Cells	SRX1053379	POU5F1
Embryonic Stem Cells	SRX1053378	POU5F1
hESC H9	SRX027488	SMARCA4
hESC H1	SRX833403	CTNNB1
hESC H1	SRX833420	NIPBL
hESC H9	SRX027482	EP300
hESC HUES64	SRX702045	OTX2
hESC HUES64	SRX702069	POU5F1
hESC HUES64	SRX266859	POU5F1
hESC HUES64	SRX702132	TCF4
hESC H1	SRX833422	NIPBL
Embryonic Stem Cells	SRX1053369	POU5F1
NCCIT	SRX123445	POU5F1
hESC H1	SRX833421	NIPBL
Embryonic Stem Cells	SRX1053370	POU5F1
hESC H1	SRX833401	LEF1
hESC H1	SRX833423	NIPBL
iPS cells	SRX512369	POU5F1
hESC derived mesendoderm	SRX684516	T



# 科学技術や医療への応用性

## ■ 遺伝子制御機構の解明

Epigenome 情報が容易に理解できる

- 遺伝子制御の上下関係、制御領域の特定

## ■ 遺伝性疾患の解明

Non-coding GWAS SNP に結合する転写因子を同定した

- 発症メカニズムの分子基盤
- リスク予測、precision medicine への応用

## ■ 組織工学と再生医療

組織特異的エンハンサーに結合する転写因子を同定した

- 各種組織のマスター制御因子の特定
- ダイレクトリプログラミングへの応用

# in silico ChIP

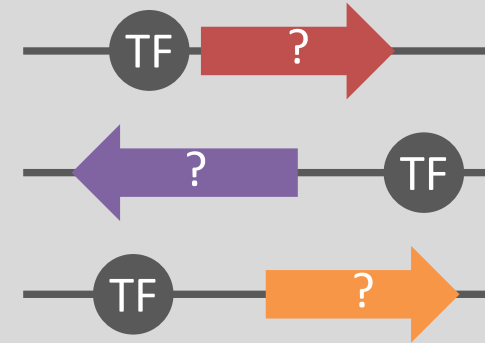
## ① Peak Browser

何がどこに結合する？



## ② Target Genes

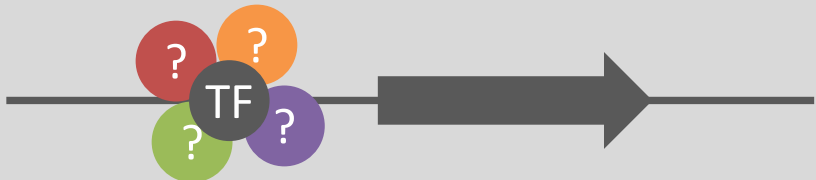
転写因子の標的遺伝子を探す。



# ChIP-Atlas

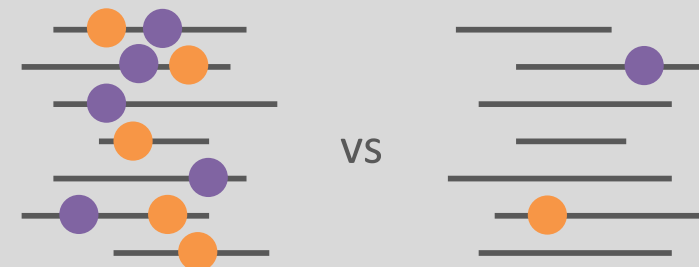
## ③ Colocalization

共局在する転写因子を探す。



## ④ in silico ChIP

ゲノム領域への Enrichment 解析





# in silico ChIP

例) 炎症性腸疾患 SNP に enrich する転写因子を探したい

## 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

- Genomic regions (BED) or sequence motif ⓘ  
 Gene list (Gene symbols) ⓘ

```
chr1 100128315 100128440
chr1 103190456 103190612
chr1 107541234 107541357
chr1 108325177 108325403
chr1 110412483 110412583
chr1 111120868 111121001
chr1 111693652 111693799
chr1 112421367 112421483
```

### 5. Select dataset to be compared

- Random permutation of user data ⓘ  
 BED or sequence motif ⓘ

```
chr1 10488202 10488365
chr1 107973343 107973565
chr1 108293100 108293277
chr1 109370825 109371023
chr1 110319562 110319739
chr1 112298232 112298429
chr1 113351913 113352118
chr1 116524095 116524271
```

### 6. Describe datasets

User data title ⓘ

My data

Compared data title ⓘ

Control

Project title ⓘ

My project

submit

Estimated run time: 2 mins

炎症性腸疾患 SNP の  
ゲノム座標

その他の SNP の  
ゲノム座標

# 炎症性腸疾患 SNP と有意に overlap するデータ

ID	Antigen class	Antigen	Cell class	Cell	Log P-val
SRX212650	TFs and others	STAT1	Blood	Monocytes-CD14+	-26.7
SRX831873	TFs and others	STAT5B	Blood	CD8-Positive T-Lymphocytes	-21.7
SRX212648	TFs and others	STAT1	Blood	Monocytes-CD14+	-20.9
SRX1023792	TFs and others	SPI1	Blood	Macrophages	-20.2
SRX831874	TFs and others	STAT5B	Blood	CD8-Positive T-Lymphocytes	-18.9
SRX212649	TFs and others	STAT1	Blood	Monocytes-CD14+	-17.2
SRX831878	TFs and others	STAT5B	Blood	CD8-Positive T-Lymphocytes	-15.9
SRX092314	TFs and others	GATA3	Blood	Th1 Cells	-15.6
SRX831879	TFs and others	STAT5B	Blood	CD8-Positive T-Lymphocytes	-15.4
SRX1023791	TFs and others	SPI1	Blood	Macrophages	-14.9

疾患 SNP に結合する転写因子がわかる。  
疾患の原因となる細胞タイプがわかる。

# 科学技術や医療への応用性

## ■ 遺伝子制御機構の解明

Epigenome 情報が容易に理解できる

- 遺伝子制御の上下関係、制御領域の特定

## ■ 遺伝性疾患の解明

Non-coding GWAS SNP に結合する転写因子を同定した

- 発症メカニズムの分子基盤
- リスク予測、precision medicine への応用

## ■ 組織工学と再生医療

組織特異的エンハンサーに結合する転写因子を同定した

- 各種組織のマスター制御因子の特定
- ダイレクトリプログラミングへの応用

# 利用状況 (2015年12月～)

## ■ 機能的利用数

15,795 回

\*4つの機能のいずれかの利用回数(2017年1月まで)。  
単純訪問者や、botなどのアクセスは除く

年月	訪問者数	訪問回数	ページ数
2017年4月	1,171	4,118	20,656
2017年5月	1,858	6,095	50,612
2017年6月	1,637	5,717	55,772

\*2017年4月より、AWStatsで集計

## ■ 利用実績や評価



データベース部門 **第1位** \* 2015年度



データダウンロード回数

**第1位** \* 2017年5月

**第2位** \* 2017年4月

# 利用状況 (2015年12月～)

## ■被引用論文：8報

- Guo, et al. **Proc. Natl. Acad. Sci. U. S. A.** (2017).
- Ishigaki, et al. **Nat. Genet.** (2017).
- Kalender, et al. **bioRxiv** (2017).
- Kehl, et al. **Nucleic Acids Res.** (2017).
- Matsuda, et al. **Development** (2017).
- Tanegashima, et al. **Genes to Cells** (2017).  
Yevshin, et al. **Nucleic Acids Res.** (2017).
- Sugeno, et al. **Sci. Rep.** (2016).

## ■投稿中：2報

## ■投稿準備中：2報

## ■共同研究者：18名

おもに知識発見や  
仮説構築に利用



# ChIP-Atlas の特長

個々の実験データ ( $n > 50,000$ )



Curation

Alignment, Peak-call

統合解析

ユーザ



データの視覚的理解

データマイニング

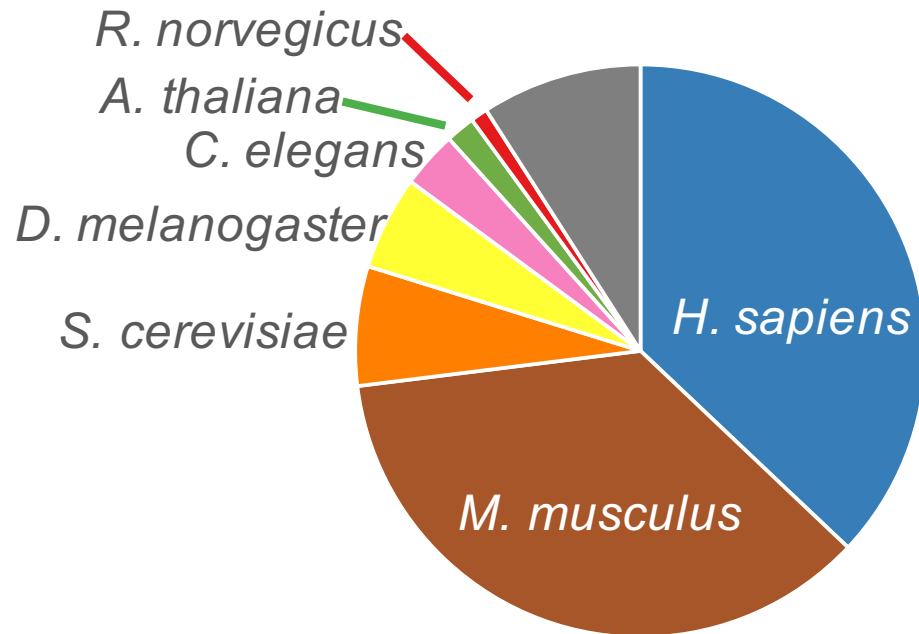
知識発見・仮説構築

- ・ 生物種の追加
- ・ Bisulfite-seq データの追加
- ・ キュレーションの効率化
- ・ 他のデータベースとの連携
- ・ 広報活動

# 生物種の追加

## ChIP-seq data

( $n = 62,672$ )

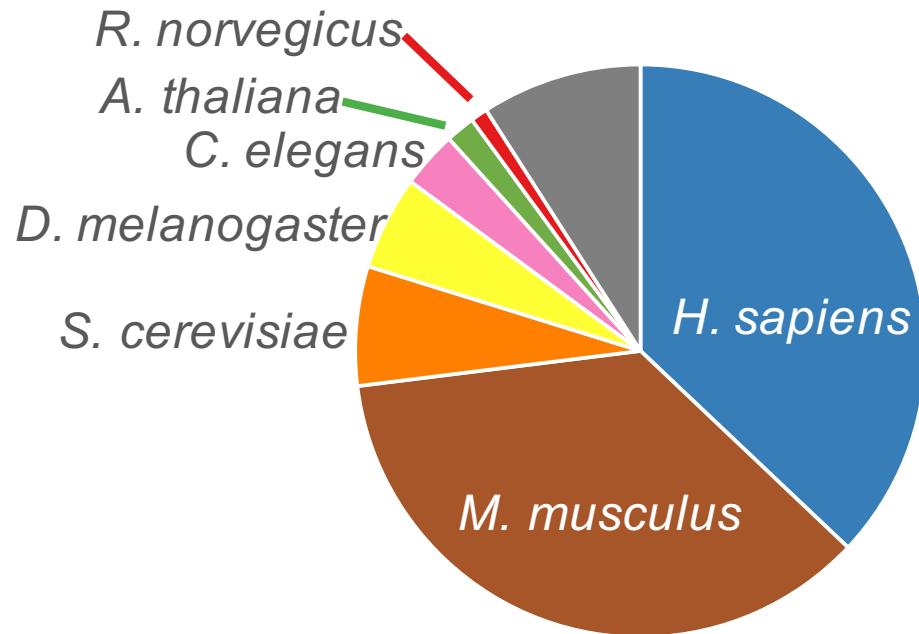


ナズナとラットのデータを追加する。

# Bisulfite-seq データの追加

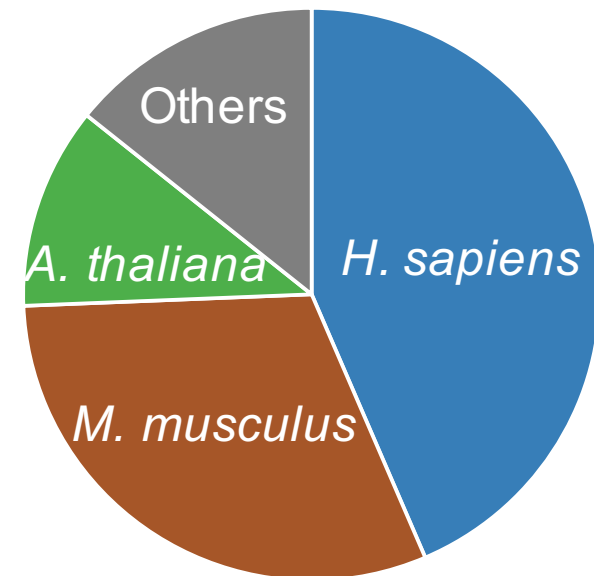
## ChIP-seq data

(n = 62,672)



## Bisulfite-seq data

(n = 19,680)



ヒト、マウス、ナズナのデータを追加する。

# Bisulfite-seq データの追加



ChIP-seq SRA

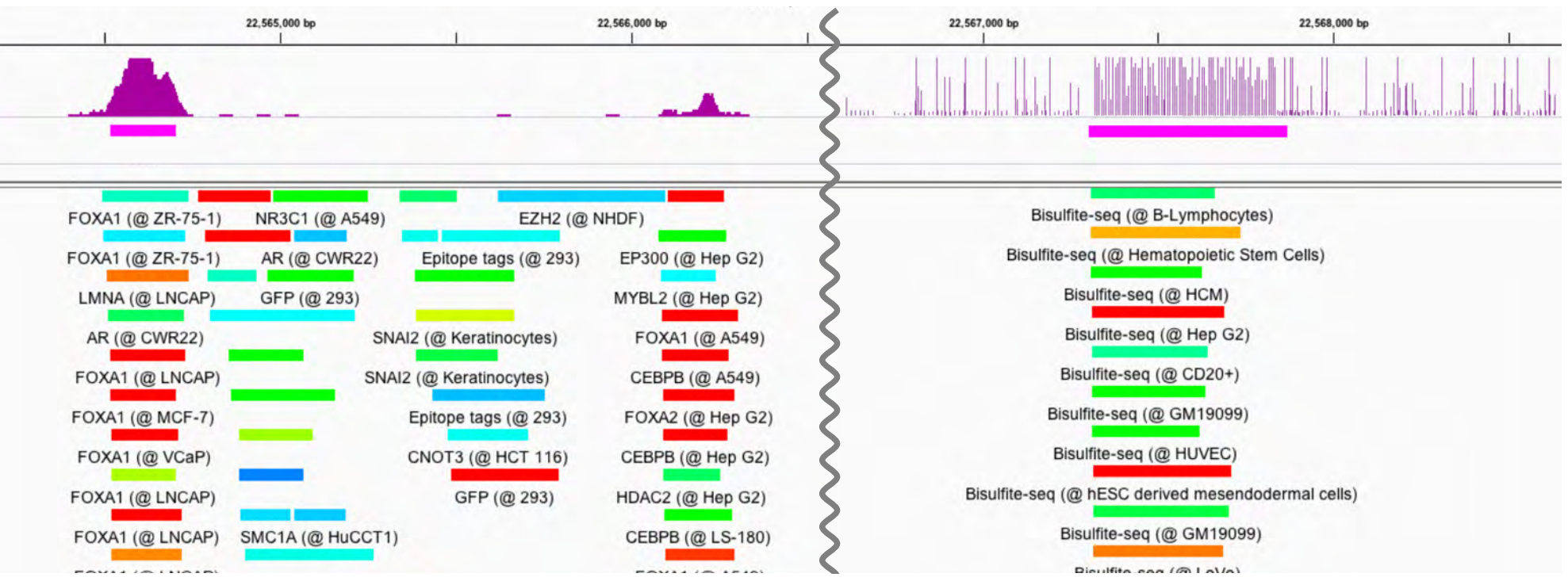
Bisulfite-seq SRA

Alignment  
Peak-call

Alignment  
Methyl-call

個別データ

統合データ





# Bisulfite-seq データの追加

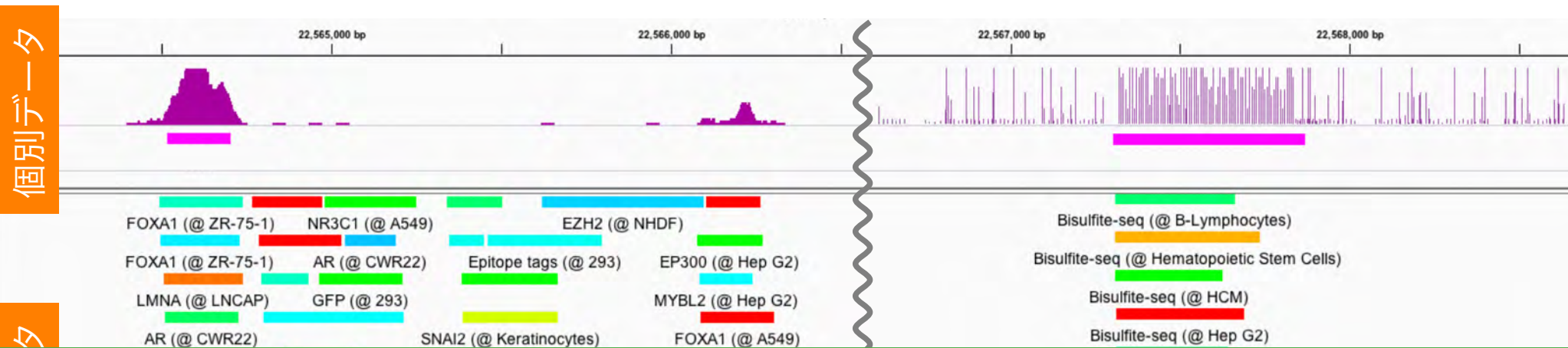


ChIP-seq SRA

Bisulfite-seq SRA

Alignment  
Peak-call

Alignment  
Methyl-call



タンパクの結合やゲノムメチル化が全てわかる。  
→ エピゲノミクス統合データベースとする。

# Bisulfite-seq データの追加：進捗状況

## アライメントツールの選定

Bisulfite-seq のリードは直接アライメントできない



# Bisulfite-seq データの追加：進捗状況

## アライメントツールの選定

Bisulfite-seq アライメントツールはたくさんある。

BatMeth	BS-Seeker	MethylCoder
Bismark	BS-Seeker2	Novoalign
Bisulfighter	B-SOLANA	Pash
BRAT-BW	GSNAP	RMAP
BSMAP	ERNE-BS5LAST	SOCS-B
BSmooth	MAQ	

Tsuji et al. *Brief Bioinform.* (2016)

### ➤ 選定とパイプライン構築を外注 (Rhelixa 社)

- 汎用度 (論文数)
- 使いやすさ (消費メモリ、実行速度)
- 再現性 (論文の図との比較)

- 生物種の追加
- Bisulfite-seq データの追加
- キュレーションの効率化
- 他のデータベースとの連携
- 広報活動

# メタ情報のキュレーション

属性	属性値
ID	SRX213809
Title	AntiGFP KD Oct4; Mus musculus; ChIP-Seq
source_name	Anti-GFP KD mESCs
strain	129S4/Svjae
phenotype	agouti
gender	male
cell type	ESC
genotype	Anti-GFP shRNA KD
chip antibody	Oct4(N-19)(sc-8628), Santa Cruz Biotechnology

細胞名 = Embryonic Stem Cells

抗原名 = Pou5f1



# メタ情報のキュレーション

## BEFORE

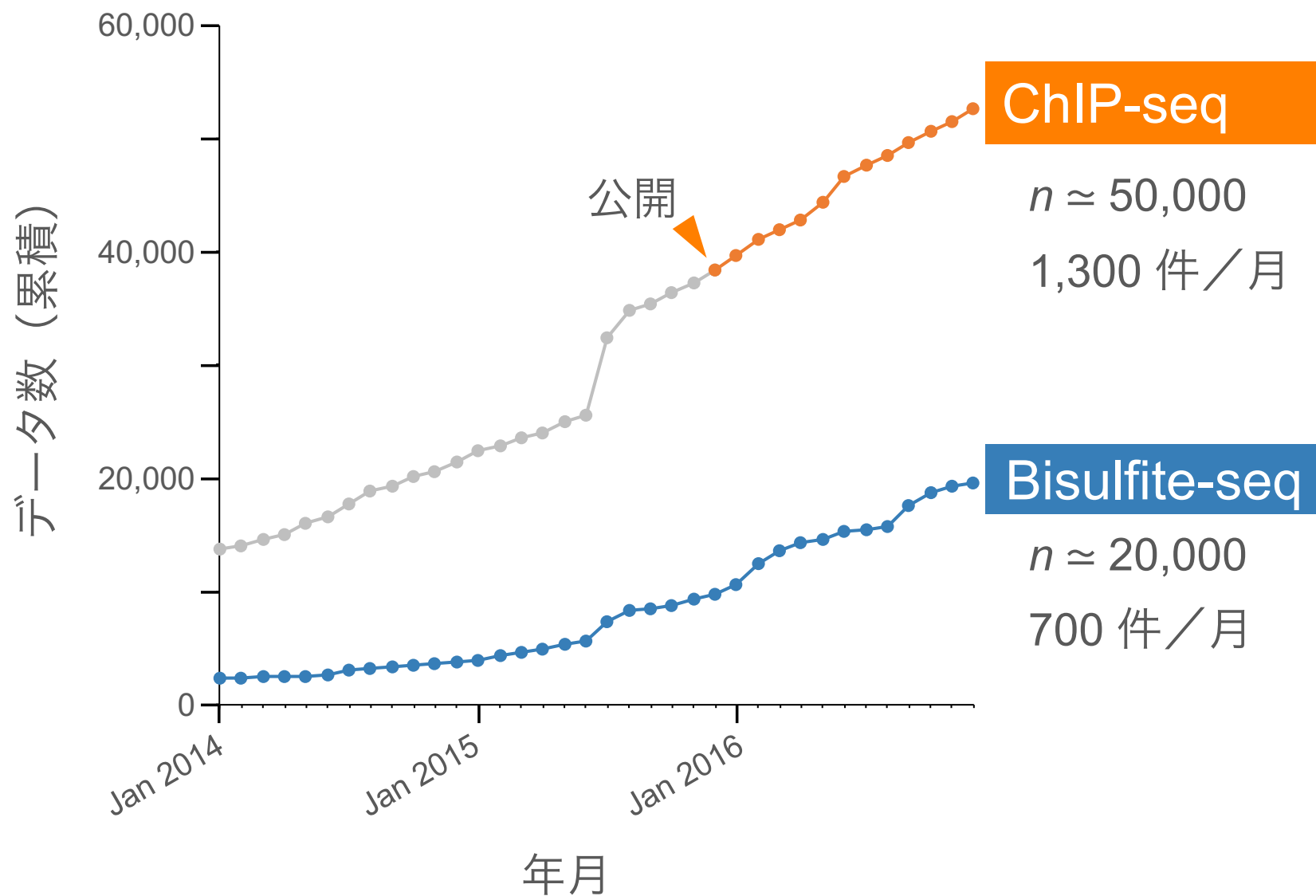
ID	Original sample metadata
ERX200492	ArrayExpress-Sex=male; ArrayExpress-CellType=B-Lymphocyte; ArrayExpress-Immunoprecipitate=CTCF; ArrayExpress-Species=Hom
SRX1024932	source_name=chronic myelogenous leukemia cell line; cell line=K562; antibody=AGO2
SRX212431	source_name=CD4+CD25+CD45RA+ expanded naive regulatory T ce donor=S030b; cell type=CD4+CD25+CD45RA+ expanded naive regul
SRX831872	source_name=Pre-activated CD8+ T cells; tissue=Peripheral blood; ce type=Pre-activated CD8+ T cells; chip antibody=STAT5B (Invitrogen,
SRX100477	source_name=HepG2; biomaterial_provider=ATCC; datatype=ChipSe datatype description=Chromatin IP Sequencing; cell
SRX530184	source_name=primary human hepatocytes; tissue=Liver; chip antibody=FXR (1:1 mixture of sc-1204x and sc-13063x, CHIP grade,
SRX159094	source_name=HuH7 cells, mitotic block and release; cell type=HuH7; cycling state=mitotic; antibody=FoxA1; vendor/catalog/lot=Abcam
SRX644410	source_name=PANC1 Pancreatic cancer cell line; antibody=ETS1 (Sa Cruz, sc-350, lot #F1312); cell line=PANC1

# メタ情報のキュレーション

## AFTER

ID	Antigen	Cell type Class	Cell type
ERX200492	CTCF	Blood	B-Lymphocytes
SRX1024932	AGO2	Blood	K-562
SRX212431	STAT5A	Blood	CD4-Positive T-Lymphocytes
SRX831872	STAT5B	Blood	CD8-Positive T-Lymphocytes
SRX100477	FOXA1	Liver	Hep G2
SRX530184	NR1H4	Liver	Hepatocytes
SRX159094	FOXA1	Liver	HuH-7
SRX644410	ETS1	Pancreas	PANC-1

# メタ情報のキュレーション



キュレータの雇用、養成

沖 G, 浜本 G 担当

# キュレーションの品質管理

## 転写因子名

名称	生物種
HGNC	ヒト
MGI	マウス
FlyBase	ハエ
WormBase	線虫
SGD	酵母

例)

OCT4 → POU5F1

p53 → TP53

## 細胞名

名称	概要
Yu et al 2015	Nature 誌で提唱された細胞名表記法
ATCC	非営利団体細胞バンク
MeSH	NLM が定める生命科学用語集

例)

K562, K-562 → K-562

HepG2, Hep-G2 → Hep G2

マニュアル策定・ルールの徹底

# 特徴語抽出

BEFORE	AFTER
biomaterial_provider=ATCC; cell_line= <b>K562</b> ; disease=chronic myelogenous leukem...	→ Blood K-562
Sample Description= <b>K562</b> CHIP DNA; ArrayExpress-Phenotype=Normal; ArrayExpri...	→ Blood K-562
ArrayExpress-Sex=female; ArrayExpress-CellType= <b>K562</b> ; ArrayExpress-DiseaseSt...	→ Blood K-562
datatype=ChipSeq; datatype description=Chromatin IP Sequencing; cell= <b>K562</b> ; cell ...	→ Blood K-562
antibody=ZNF263; cell type=myelogenous leukaemia; cell line= <b>K562</b> ; passage=10- ...	→ Blood K-562

## 機械学習による半自動化

- ヒューマンエラーを防ぐ
- データ爆発に備える

# キュレーションの効率化：進捗状況

## 機械学習による半自動化

- 文字 n-gram 分割
- CNN または SVM 法によるテキスト分類  
(ten-fold cross validation)

BEFORE	AFTER
Cultured K562 cells_dCas9_KRAB_HS2_CR4_H3K9me3	K-562
K562; K562_H3K122ac; human erythroleukemic cell line	K-562
K562; K562 NaBut 72hr DNase-seq; myelogenous leukemia cell line	K-562
K562; Leukemia; Leukemia	K-562
K562; Input ChIP-Seq K562 BIO	K-562
Cultured K562 cells_dCas9_KRAB_HS2_CR10_FLAG	K-562
K562 PU.1 shRNA2 DMSO DNase-seq; myelogenous leukemia cell line	K-562
chronic myeloid leukemia	K-562
K562-RNF2 ChIP-seq; K562 erythrocytic leukaemia cells (ATCC CCL-243)	K-562
K562 cells	K-562
Chromatin IP with NFYA Ab	K-562
K562 High NaBut 72hr DNase-seq; myelogenous leukemia cell line	K-562

# キュレーションの効率化：進捗状況

## 機械学習による半自動化

- 文字  $n$ -gram 分割
- CNN または SVM 法によるテキスト分類  
(ten-fold cross validation)

BEFORE	AFTER
Cultured K562 cells_dCas9_KRAB_HS2_CR4_H3K9me3	K-562
Cul	K-562
ult	K-562
ltu	K-562
tur	K-562
ure	K-562
red	K-562
K56	K-562
562	K-562
cel	K-562
ell	K-562
lls	K-562

$n = 3$  の場合



# キュレーションの効率化：進捗状況

## 正答例

INPUT	OUTPUT
K562; K562 Input Control ChIP-seq; myelogenous leukemia cell line	K-562
K562 myelogenous leukemia; N/A; N/A	K-562
K562; K562 PU.1 shRNA2 SAHA DNase-seq; myelogenous leukemia cell line	K-562
K562; K562 SAHA PU.1 ChIP-seq; myelogenous leukemia cell line	K-562
K562; chronic myelogenous leukemia; chronic myelogenous leukemia cell line	K-562
K562 cells; TBP ChIP-seq (Human K562 cells) Rep 1	K-562
K562-P300 ChIP-seq; K562 erythrocytic leukaemia cells (ATCC CCL-243)	K-562
K562-RNF2 ChIP-seq; K562 erythrocytic leukaemia cells (ATCC CCL-243)	K-562
leukemia cell line K562; K562	K-562
ATCC CCL-243; cultured K562 cells (ATCC CCL-243); K562	K-562
Mock-treated K562 erythroleukemia cells	K-562
ERS363987; K-562; mesoderm; Chronic Myelogenous Leukemia	K-562

# キュレーションの効率化：進捗状況

## 正答例

INPUT	OUTPUT
OCT4; embryonic stem cells	POU5F1
Oct4; SC-8628; Embryonic stem cells	POU5F1
POU5F1(Abcam,ab19857,GR21088-1); dEC	POU5F1
Oct4; BJ_Oct4 ChIP-Seq_48hr_mockinduction	POU5F1
Oct4; BJ_Oct4 ChIP-Seq_48hr_postinduction	POU5F1
POU5F1(Abcam,ab19857,GR21088-1); dEN	POU5F1
POU5F1(Abcam,ab19857,GR21088-1); dME	POU5F1
Oct4; sc8628; Embryonic stem cells	POU5F1

AFTER の頻出回数が 10 以上：約 9 割の正答率

10 以下：精度が極端に低い

瀬々アドバイザーとの共同研究

# キュレーションの効率化：進捗状況

## 誤答例

INPUT	OUTPUT
anti OCT3 antibody; Induced pluripotent stem cells	Input control
OCT4	Unclassified
OCT4A; embryonic carcinoma NCCIT	Unclassified
anti-Oct4 (Abcam, ab19857, GR21088-1); hES HUES6	Unclassified

AFTER の頻出回数が 10 以上：約 9 割の正答率

10 以下：精度が極端に低い

➤ Rhelixa 社との共同開発。

# キュレーションの効率化：進捗状況

## これまでのキュレーションの再チェック

- 医学的観点からの妥当性
- どのレベルまで書き下すかという問題

### 例 1) 白血病サンプル

[急性／慢性] [骨髄性／リンパ性] 白血病

+ 表面抗原？

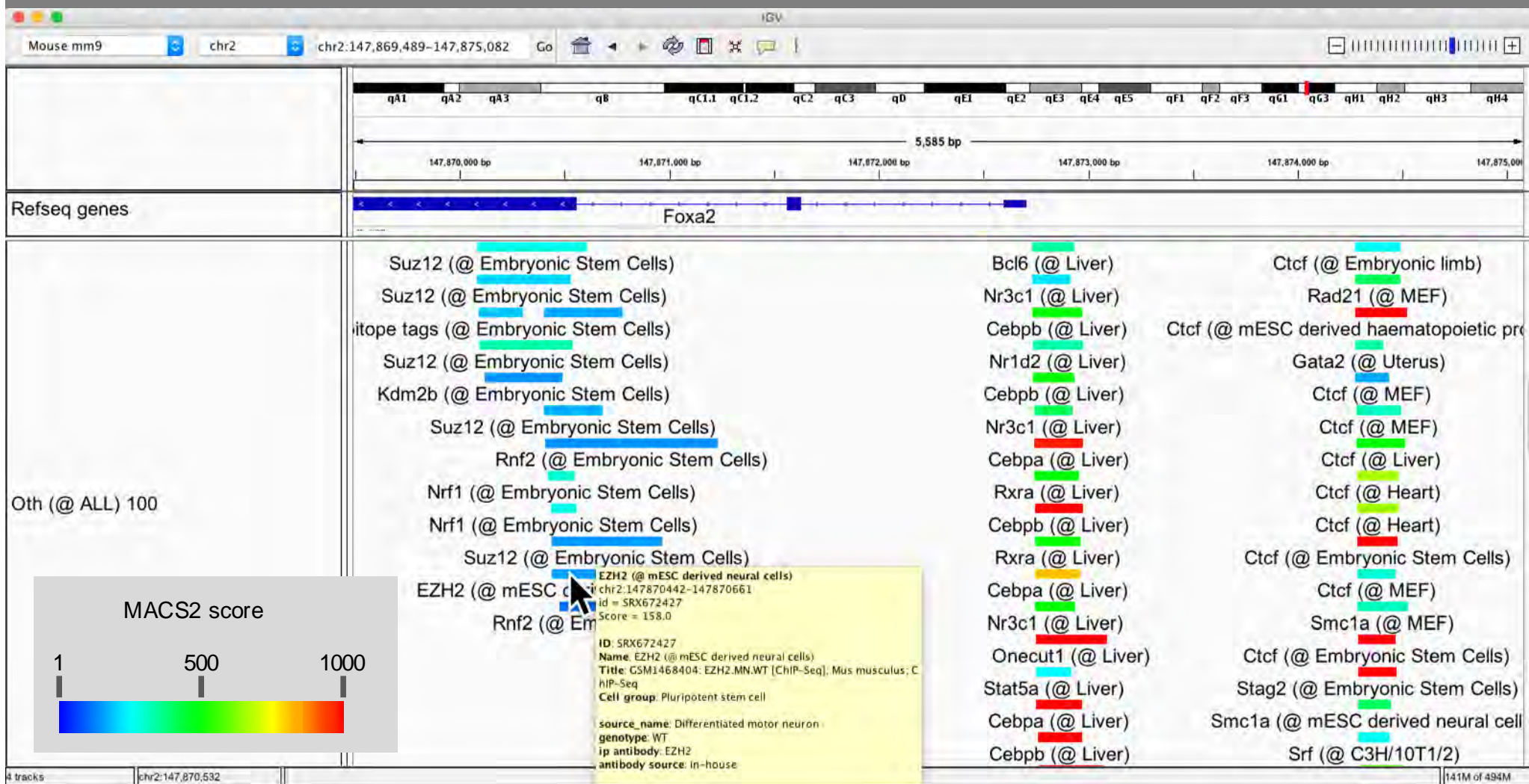
+ 採取した組織部位？

### 例 2) 乳がんサンプル

+ 硬がん／乳頭腺管がん？

+ ホルモン受容体、HER2などの陰性／陽性？

# キュレーションの効率化：進捗状況



- ・ 視認性とのトレードオフ  
(細分化するほど、視認性や検索性が低下)

浜本 G と協議中

- 生物種の追加
- Bisulfite-seq データの追加
- キュレーションの効率化
- 他のデータベースとの連携
- 広報活動

# 連携するための前提

## BEFORE

biomaterial\_provider=ATCC; cell\_line=**K562**; disease=chronic myelogenous leukem... → Blood K-562

Sample Description=**K562** CHIP DNA; ArrayExpress-Phenotype=Normal; ArrayExpri... → Blood K-562

ArrayExpress-Sex=female; ArrayExpress-CellType=**K562**; ArrayExpress-DiseaseSt... → Blood K-562

datatype=ChipSeq; datatype description=Chromatin IP Sequencing; cell=**K562**; cell ... → Blood K-562

antibody=ZNF263; cell type=myelogenous leukaemia; cell line=**K562**; passage=10- ... → Blood K-562

## AFTER

テキストデータだけでは意味不明



# 連携するための前提



Thing  
+ [entity](#)  
+ [continuant](#)  
+ [independent continuant](#)  
+ [material entity](#)  
+ [cell](#)  
+ [cell in vitro](#)  
+ [experimentally modified cell in vitro](#)  
+ [cultured cell](#)  
+ [secondary cultured cell](#)  
+ [cell line cell](#)  
+ [immortal cell line cell](#)  
+ [immortal human cell line cell](#)  
- [697 cell](#)  
- [9229 cell](#)  
- [22RV1 cell](#)  
- [2B8 cell](#)  
- [42-MG-BA cell](#)  
- [A-204 cell](#)  
- [A253 cell](#)  
- [A2780cis cell](#)  
- [A3 cell](#)  
- [A388 cell](#)  
- [A673 cell](#)  
+ [NIH:OVCAR-3 cell](#)  
[more...](#)  
- [K-562 cell](#)  
- [K-562 clone A cell](#)  
- [K-562 clone S cell](#)

K-562 ([CLO\\_0007059](#)) とは

- immortal [human cell line cell](#)  
([CLO\\_0000511](#))

- derives from patient having disease [leukemia](#)  
([DOID\\_1240](#))

- derives from [blood cell](#)  
([CL\\_0000081](#))

細胞名が controlled vocabulary に紐づくことが前提条件

# 他のデータベースとの連携

ID	Antigen	Cell	Ontology ID
SRX150420	ATF1	K-562	CLO_0007059
SRX150423	GTF3C1	K-562	CLO_0007059
SRX150436	USF2	K-562	CLO_0007059
SRX150441	NRF1	K-562	CLO_0007059
SRX150452	RPC155	K-562	CLO_0007059
SRX150465	RCOR1	K-562	CLO_0007059
SRX150472	NELFE	K-562	CLO_0007059
SRX150474	GTF2B	K-562	CLO_0007059
SRX150504	XRCC4	K-562	CLO_0007059
SRX150548	STAT2	K-562	CLO_0007059
SRX150550	STAT1	K-562	CLO_0007059
SRX150569	BRF2	K-562	CLO_0007059
SRX150574	TBP	K-562	CLO_0007059
SRX150575	TAL1	K-562	CLO_0007059
SRX150576	MXI1	K-562	CLO_0007059
SRX150578	CEBPB	K-562	CLO_0007059
SRX150580	GTF2F1	K-562	CLO_0007059
SRX150581	CHD2	K-562	CLO_0007059
SRX150583	IRF1	K-562	CLO_0007059
SRX150599	MAFF	K-562	CLO_0007059
SRX150623	HMG3	K-562	CLO_0007059
SRX150626	CCNT2	K-562	CLO_0007059
SRX150644	RFX5	K-562	CLO_0007059
SRX150647	ZNF143	K-562	CLO_0007059
SRX150653	TBL1XR1	K-562	CLO_0007059
SRX150655	BACH1	K-562	CLO_0007059

連携  
できる



K-562 (CLO\_0007059)  
を用いた実験を含む  
その他の omics DB

DBKERO (協議予定)

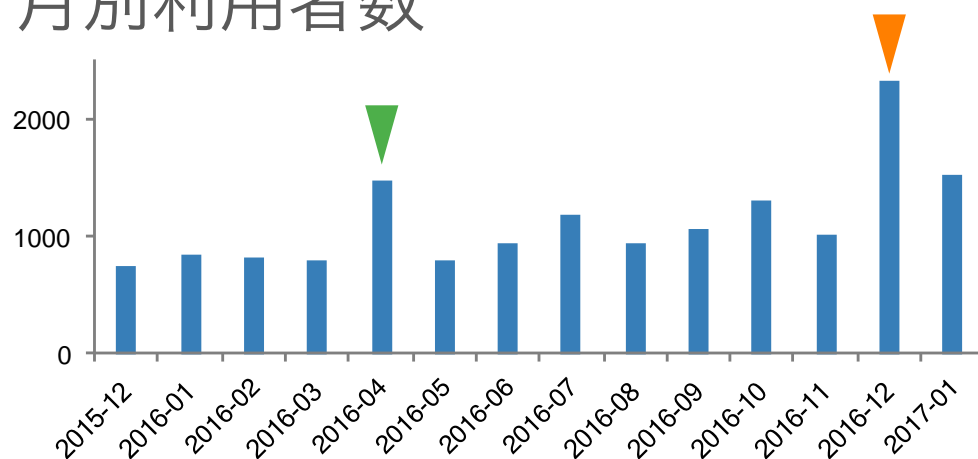
FANTOM5 (協議中)

- 生物種の追加
- Bisulfite-seq データの追加
- キュレーションの効率化
- 他のデータベースとの連携
- 広報活動

# 広報活動

会議種別	会議名
国内学会	日本分子生物学会、日本発生生物学会、
国際学会	Keystone symposium, Cold Spring Harbor Lab meeting
研究集会	遺伝研研究会、トーゴの日、「個性創発脳」領域会議 生命情報科学若手の会、NGS 現場の会、発生と代謝の会
招待セミナー	京大 CiRA、横浜理研、神戸理研、国立がんセンター (予定：東京医科歯科大、先端医療センター)
講習会	DDBJ 講習会

## 月別利用者数



# 広報活動

## 最近の活動

2017年	7月	東京大学	招待セミナー
2017年	7月	次世代生命科学の会	口頭発表
2017年	7月	個性創発脳 領域会議	ポスター
2017年	5月	日本発生生物学会	口頭発表
2017年	5月	NGS 現場の会	ポスター
2017年	3月	先端医療センター研究所	招待セミナー
2017年	3月	東京医科歯科大学	招待セミナー
2017年	2月	発生と代謝を考える会	口頭発表
2017年	2月	Keystone symposium	口頭発表

+ 論文投稿

発生・幹細胞・情報系  
コミュニティ

ニーズ ↓ ↑ 広報活動

沖 G (代表)

- 目野 (キュレーション)
- 工藤・川勝 (ナズナ)
- 江原 (自動化)

保守・管理  
解析パイプライン設計  
キュレーション

ChIP-Atlas  
+ Bisulfite-seq  
+ 生物種の追加

解析パイプライン設計

キュレーション

三浦 G (共同)  
- 荒木 (BS-seq)

ニーズ ↑ ↓ 広報活動

エピゲノム・分子生物学  
コミュニティ

浜本 G (代表)  
- 金子 (キュレーション)

ニーズ ↑ ↓ 広報活動

がん・医歯薬系  
コミュニティ

# 達成目標

## ～3年次末

- ・ラットとナズナの追加
- ・Bisulfite-seq データの追加

- ・キュレーションの半自動化
- ・オントロジー化, RDF 化

- ・現行の3倍の訪問回数  
(20万回/年)

## ～5年次末

- ・さらなる生物種の追加

- ・キュレータへの業務移行
- ・その他のDBとの連携

- ・現行の5倍の訪問回数  
(32万回/年)



# 謝辞

Web UI の作製, 様々な提言  
大田 達郎 (DBCLS)

構想, 提案  
塩井 剛 (RIKEN)

CoLo の提供  
仲木 竜 (東大)

データ可視化  
川路 英哉 (RIKEN)

計算機  
NIG supercomputer

WABI の作製  
小笠原 理  
奥田 喜広 (DDBJ)

サーバ提供  
畠中 秀樹 (NBDC)

統計解析  
瀬々 潤 (産総研)

データ考察  
目野 主税 (九大)

