

MOCCS-DB: ヒト転写因子認識配列の多様性データベース

○田原沙絵子^{1,2}、土屋貴穂^{1,3}、松澤亮輔^{1,4}、尾崎遼^{1,3,5}

1. 筑波大学医学医療系 バイオインフォマティクス研究室
2. 筑波大学医学群医学類
3. 筑波大学人工知能科学センター
4. 筑波大学大学院人間総合科学学術院
フロンティア医科学学位プログラム
5. 理化学研究所 生命機能科学研究センター
バイオインフォマティクス研究開発チーム

HOME Find TF recognizing k-mers Differential k-mer detection (Cell type) Differential k-mer detection (TF) TF-binding k-mer similarity analysis SNP-affected TFs / Cell type classes prediction
Download Documentation



What is MOCCS DB?

Four functions are available.

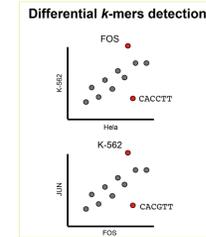
1. Find TF-binding k-mers

You can know TF-binding motifs as k-mer and MOCCS2score
You can choose Species, TFs, Cell type class.

Find TF-binding k-mers	
Target TF : JUN	
Target Cell type : K-562	
k-mer	MOCCS2score
CACGTT	33.14
CACGCG	21.56
CACATG	17.14
CACGGG	13.99
TACGTA	8.05
CATGCG	6.24
.	.

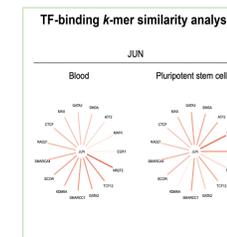
2. Differential k-mers detection

You can compare TF-binding k-mer between different Cell type Class



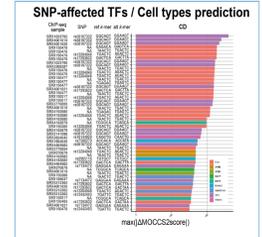
3. TF-binding k-mer similarity analysis

You can know pairs of TFs which have similar TF-recognizing k-mers
You can also see diversities of TF-recognizing k-mers among cell types



4. SNP-affected TFs / Cell types prediction

You can know phenotype-related SNP-affected TFs / Cell types.



MOCCS-DB : ヒト転写因子認識配列の多様性データベース

背景 :

転写因子は6-12bpのDNA配列を認識し、遺伝子発現調節する。

転写因子認識配列の多様性に着目したウェブデータベースは存在しない。

目的

転写因子認識配列の多様性に着目したウェブデータベースを作成する。

方法

ChIP-AtlasのChIP-seqデータを解析して得られたデータ(MOCCS profile)を用いて、4種類の機能を実装した。

MOCCS profile

ID	TF	Cell type	k-mer	MOCCS2score	q-value
SRX190276	CTCF	K-562	CCCTCG	43.5	0.0002
SRX3599305	JUN	HUVEC	GACTCA	31.2	0.0008
SRX123456	JUN	K-562	GACTCA	24.7	0.0012
SRX123456	JUN	K-562	GACTCG	13.1	0.0031
SRX123456	JUN	K-562	CACACA	6.5	0.0081
SRX123456	JUN	K-562	TACTCG	2.1	0.06

Calculate MOCCS profile similarity
(k-sim Jaccard, k-sim Pearson)

ΔMOCCS2score profile

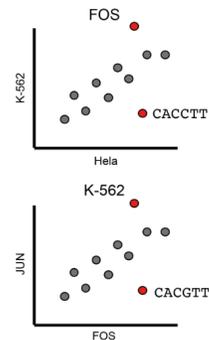
chr	position	ref k-mer	alt k-mer	ΔMOCCS2score
chr1	151829198	CACGTT	CACCTT	13.10
chr1	160887169	CACGCG	CAGCGG	9.21
chr2	144734811	CACATG	CACITG	7.12
chr3	101304898	CACGGG	CACGGG	6.70
chr3	101305102	TACGTA	TACGAA	5.57
chr20	59249248	CATGCC	CTTGCC	4.51

Find TF-binding k-mers

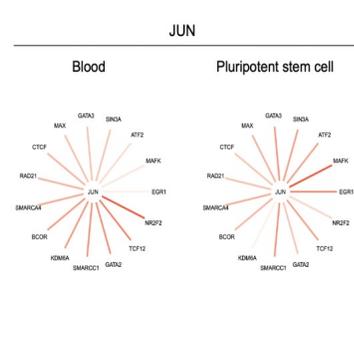
Target TF : JUN
Target Cell type : K-562

k-mer	MOCCS2score
CACGTT	33.14
CACGCG	21.56
CACATG	17.14
CACGGG	13.99
TACGTA	8.05
CATGCC	6.24

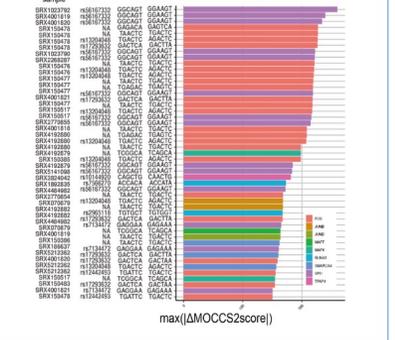
Differential k-mers detection



TF-binding k-mer similarity analysis



SNP-affected TFs / Cell types prediction



MOCCS-DB : ヒト転写因子認識配列の多様性データベース

10,000個以上のヒトChIP-seqデータを解析したデータを搭載!

～ 機能一覧 ～

- 1. Find TF-binding k -mers**
- 2. Differential k -mers detection (Cell types)
Differential k -mers detection (TFs)**
- 3. TF-binding k -mer similarity analysis**
- 4. SNP-affected TFs / Cell types prediction**

* TF = 転写因子

Find TF-binding *k*-mers

- 転写因子の認識配列が結合特異性スコア(MOCCS2score)の付随した*k*-merでわかる
- 自分のChIP-seq FASTAファイルをアップロードすると転写因子の認識配列が*k*-merでわかる

〈入力〉

STEP1: Select filters

soft

STEP2: Select Antigen

FOXA1

STEP3: Select Cell type class

Breast

STEP4: Select Cell type

ZR-75-1

* Antigen =
ChIP-seqの抗原
(=転写因子)

〈出力結果〉

ID	Antigen_class	Antigen	Cell_type_class	Cell_type	kmer	auc	count	MOCCS2score	q_value
SRX4815198	TFs and others	FOXA1	Breast	MCF-7	GTA AAC	82.5264477963052	57812	197.52111484201	0
SRX219284	TFs and others	FOXA1	Prostate	LNCAP	GTA AAC	76.0644874120173	60097	185.617841672461	0
SRX219284	TFs and others	FOXA1	Prostate	LNCAP	TAAACA	52.1671873579463	120993	180.629601417352	0
SRX2017929	TFs and others	FOXA1	Prostate	LNCAP	TAAACA	56.6036426364985	99708	177.918358065275	0
SRX1885184	TFs and others	FOXA1	Prostate	LNCAP	TAAACA	55.2421940435147	102678	176.206124141151	0
SRX219285	TFs and others	FOXA1	Prostate	LNCAP	GTA AAC	76.38893082948	52985	175.032332354363	0
SRX4815197	TFs and others	FOXA1	Breast	MCF-7	GTA AAC	82.3788038814443	45447	174.81531257168	0
SRX4815198	TFs and others	FOXA1	Breast	MCF-7	TAAACA	54.0775608451458	104692	174.174760471492	0
SRX1885181	TFs and others	FOXA1	Prostate	LNCAP	TAAACA	55.5826409739645	97252	172.543986529539	0
SRX219285	TFs and others	FOXA1	Prostate	LNCAP	TAAACA	52.913720361151	106659	172.019796947481	0

Showing 1 to 10 of 660,825 entries

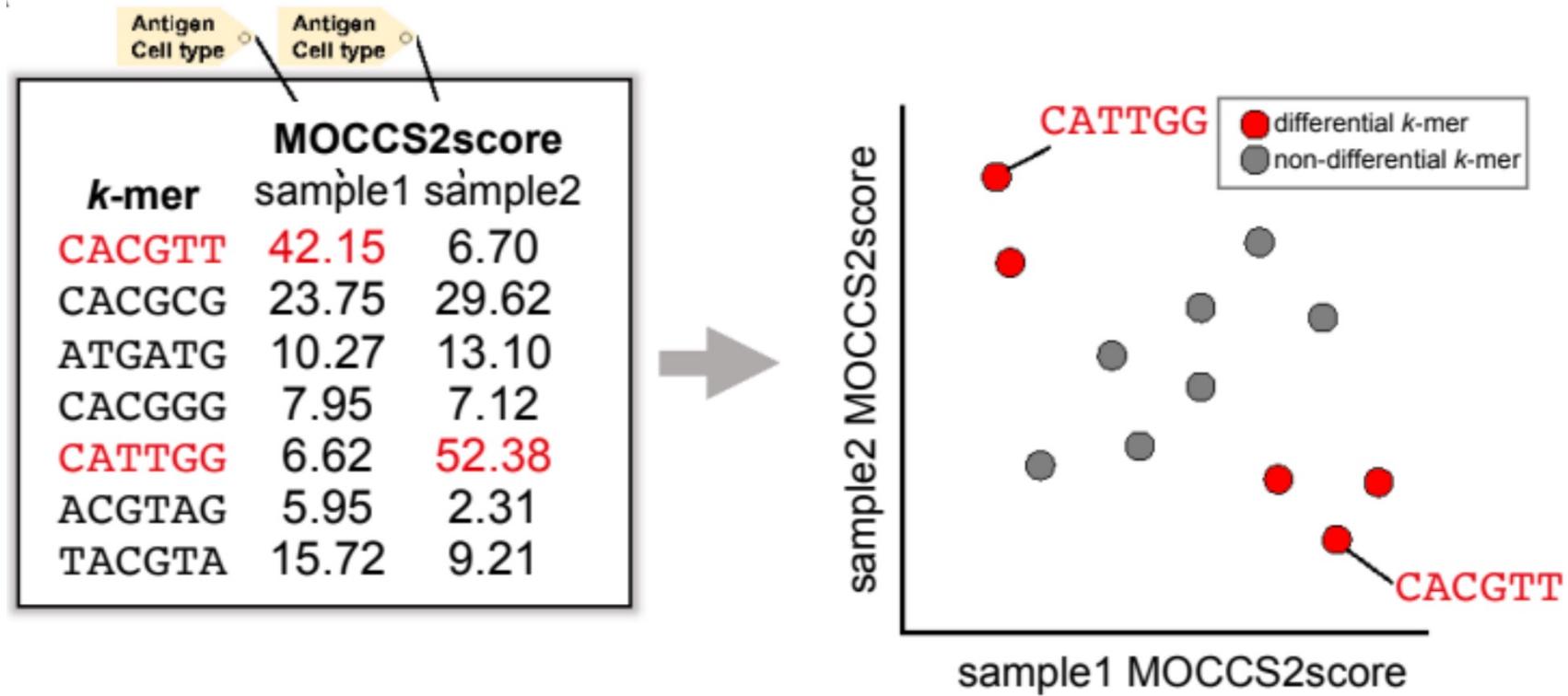
top k-mer JASPAR motif statistics

k-mer **結合特異性スコア (MOCCS2score)**

FOXA1, Breastにおける転写因子認識配列が*k*-merと結合特異性スコアでわかる！

Differential *k*-mers detection

* Differential *k*-mer : 2つのChIP-seqサンプル間で結合特異度に差がある *k*-mer



Differential *k*-mers detection (細胞型間)

* Differential *k*-mer : 2つのChIP-seqサンプル間で結合特異度に差がある *k*-mer

〈入力〉

異なる細胞型のChIP-seqサンプルを選択

STEP1 : Select filters
soft

STEP2 : Select Cell type1
K-562

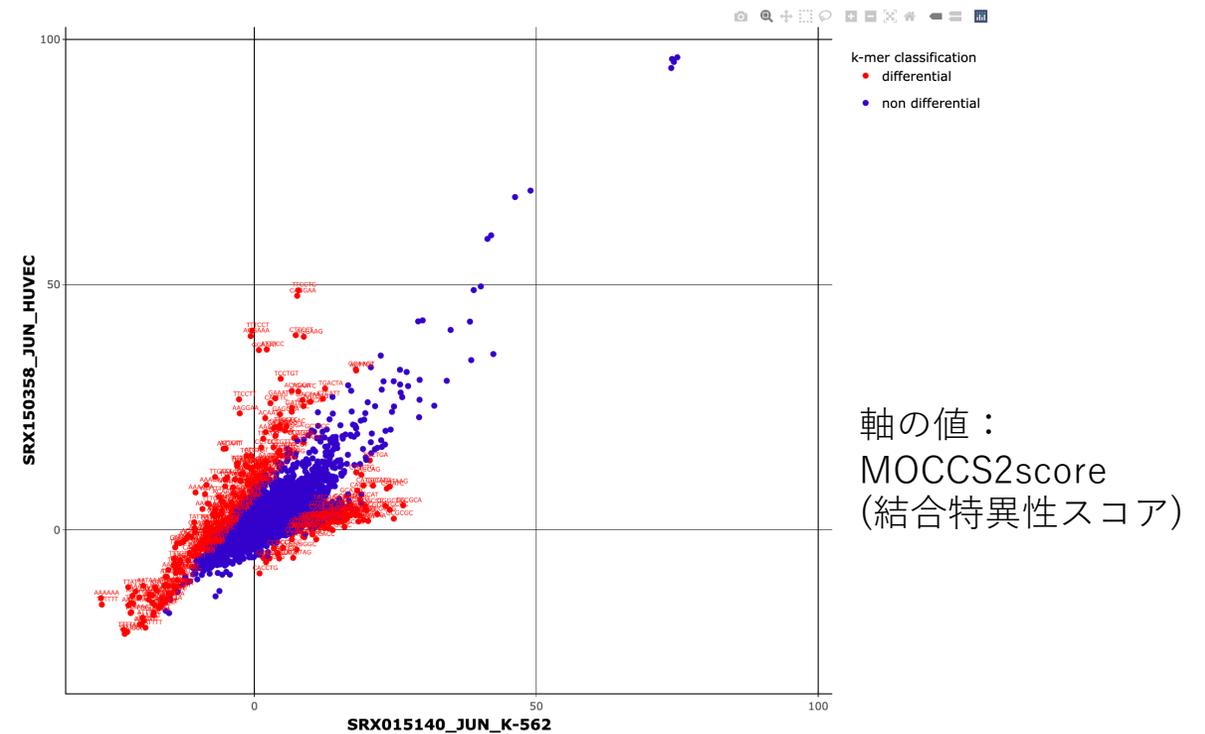
STEP3 : Select Cell type2
HUVEC

x axis
SRX015140_JUN_K-562

y axis
SRX150358_JUN_HUVEC

〈出力結果〉

異なる細胞型間のdifferential *k*-merがわかる



軸の値：
MOCCS2score
(結合特異性スコア)

細胞型間で結合特異性の異なる認識配列(*k*-mer)がわかる！

Differential *k*-mers detection (転写因子間)

* Differential *k*-mer : 2つのChIP-seqサンプル間で結合特異度に差がある *k*-mer

〈入力〉

異なる転写因子のChIP-seqサンプルを選択

STEP1 : Select filters

soft

STEP2 : Select Antigen1

JUN

STEP3 : Select Antigen2

CTCF

x axis

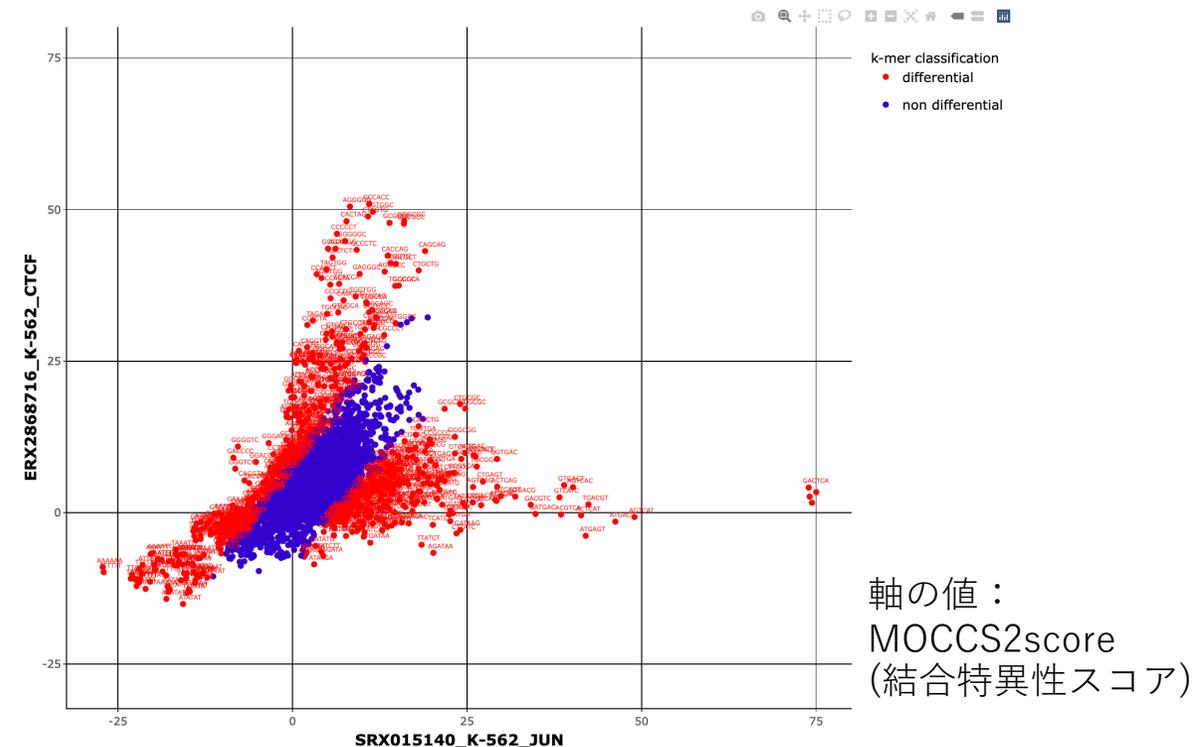
SRX015140_K-562_JUN

y axis

ERX2868716_K-562_CTCF

〈出力結果〉

異なる転写因子間のdifferential *k*-merがわかる



2つの転写因子間で、結合特異性の異なる認識配列(*k*-mer)がわかる！

TF-binding *k*-mer similarity analysis

TF-binding *k*-merの類似性を表示する。細胞型ごとの認識配列の多様性を表示する。

〈入力〉

2種類の転写因子

STEP1 : Select filters

soft

STEP2 : Select Antigen1

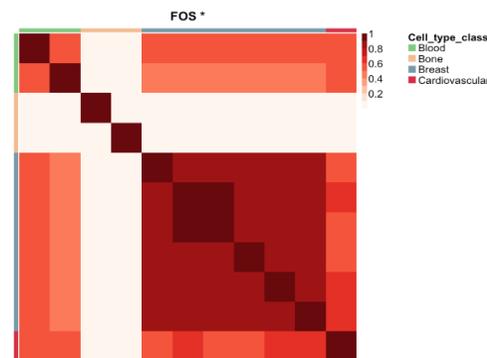
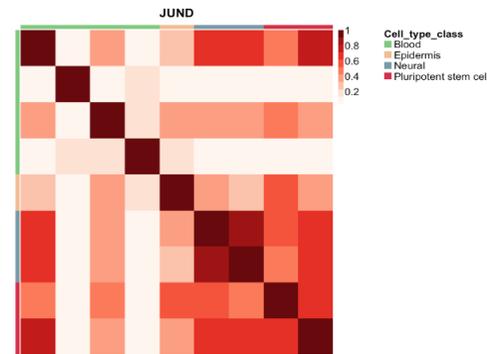
JUND

STEP3 : Select Antigen2

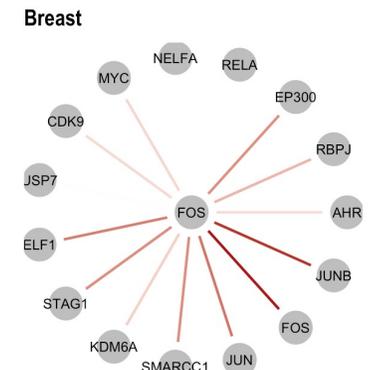
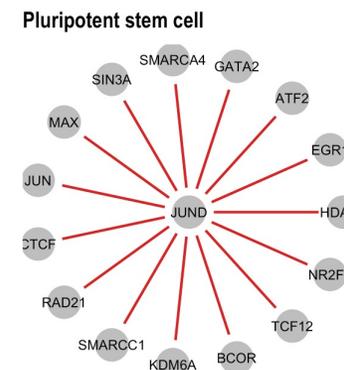
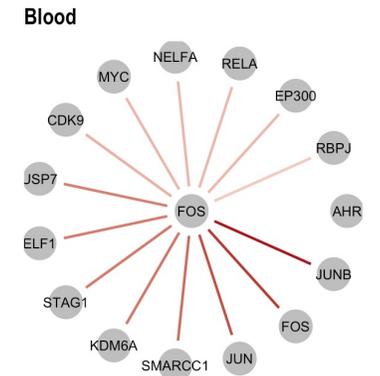
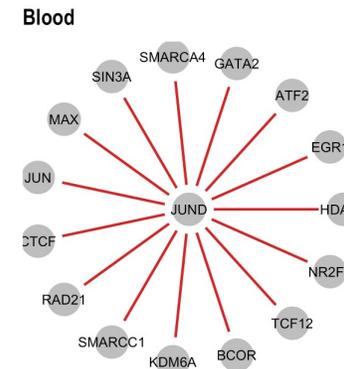
FOS

〈出力結果〉

1. 各転写因子の細胞型ごとの認識配列多様性



2. TF間の認識配列類似性



認識配列の類似性が細胞型で変わる転写因子がわかる！

SNP-affected TFs/Cell types prediction

転写因子結合領域にあるSNPが影響するTFs / Cell typeを予測できる

* $|\Delta \text{MOCCS2score}|$ が大きいほどSNPが転写因子結合に与える影響が大きい

〈入力〉

興味のある疾患・表現型 or 手持ちのSNP list

〈出力結果〉

SNPの影響が多いと予測されたTFs / Cell types

Upload GWAS catalog

STEP1: Select filters

soft

STEP2: Select SNP list

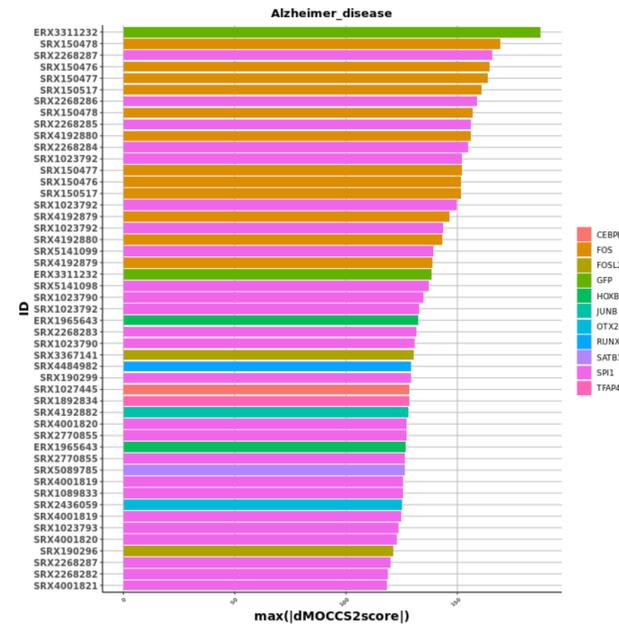
Alzheimer disease

Calculation will take time.

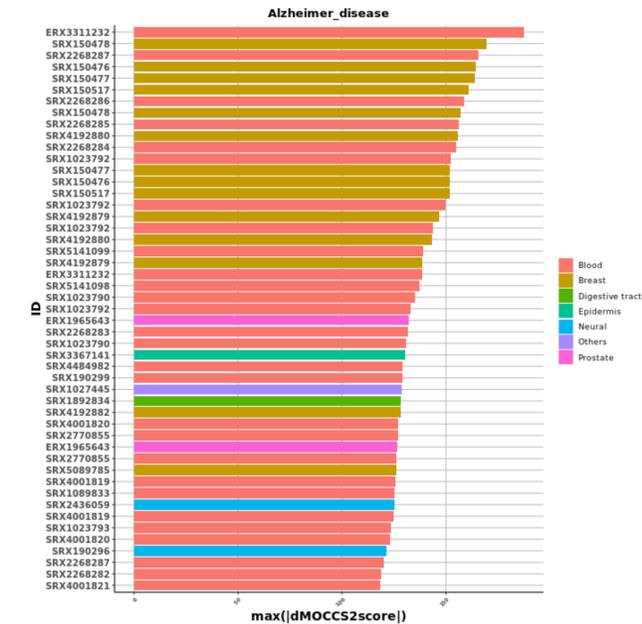
If you are ready, please push the button.

Show results

SNP-affected TFs prediction



SNP-affected Cell-type-class prediction



↑ 棒の色が転写因子ごとに塗り分けられ、 $|\Delta \text{MOCCS2score}|$ が大きい順に並び替えてある

ある疾患のSNPが影響する転写因子・細胞型が予測できる！

謝辞

- Δ MOCCS2score計算アルゴリズム方法開発
松本 拡高 博士 (長崎大学)
- ChIP-Atlasについて技術面のアドバイス
大田 達郎 博士 (DBCLS), 沖 真弥 博士 (京都大学)
- 解析技術面のアドバイス
中戸 隆一郎 博士 (東京大学)
谷澤 靖洋 博士 (NIG), 石井 学 博士 (Workflow meetup)
- 解析環境
本研究は、情報・システム研究機構 国立遺伝学研究所が有する
遺伝研スーパーコンピュータシステムを利用しました