GRECO TFBS Benchmarking Initiative

Selection, curation, and preparation of TF benchmarking data

Dr. Oriol Fornes
Deputy Group Leader @ Wasserman Lab
Centre for Molecular Medicine and Therapeutics
BC Children's Hospital Research Institute
University of British Columbia

oriol@cmmt.ubc.ca

@OFornes

February 18-19, 2020 (yes, the future)

Centre for Molecular Medicine and Therapeutics

Outline

- Aim
- Experiments
- Model Organisms
- TF Families
- Annotation
- Clustering

Aim

- Obtain a representative set of TFs:

  1. Supported by high-quality experimental data;
  2. From different model organisms; and
  3. From different structural families.

Experiments

- In vivo:

    1. [1]ChIP-seq (data sources: ChIP-Atlas, CistromeDB, GTRD and ReMap); and
    2. DAP-seq (plants; PMID: 27203113; SRA: SRP045296)

- In vitro:

    1. HT-SELEX (PMID: 23332764 and 28473536; SRA: ERP001824 and ERP010942);
    2. PBM (data sources: UniPROBE and CIS-BP v2.0); and
    3. SMiLE-seq (PMID: 28092692; SRA: SRP073361).

[1]Might include experiments from ChIP-Exo/Nexus

Model Organisms

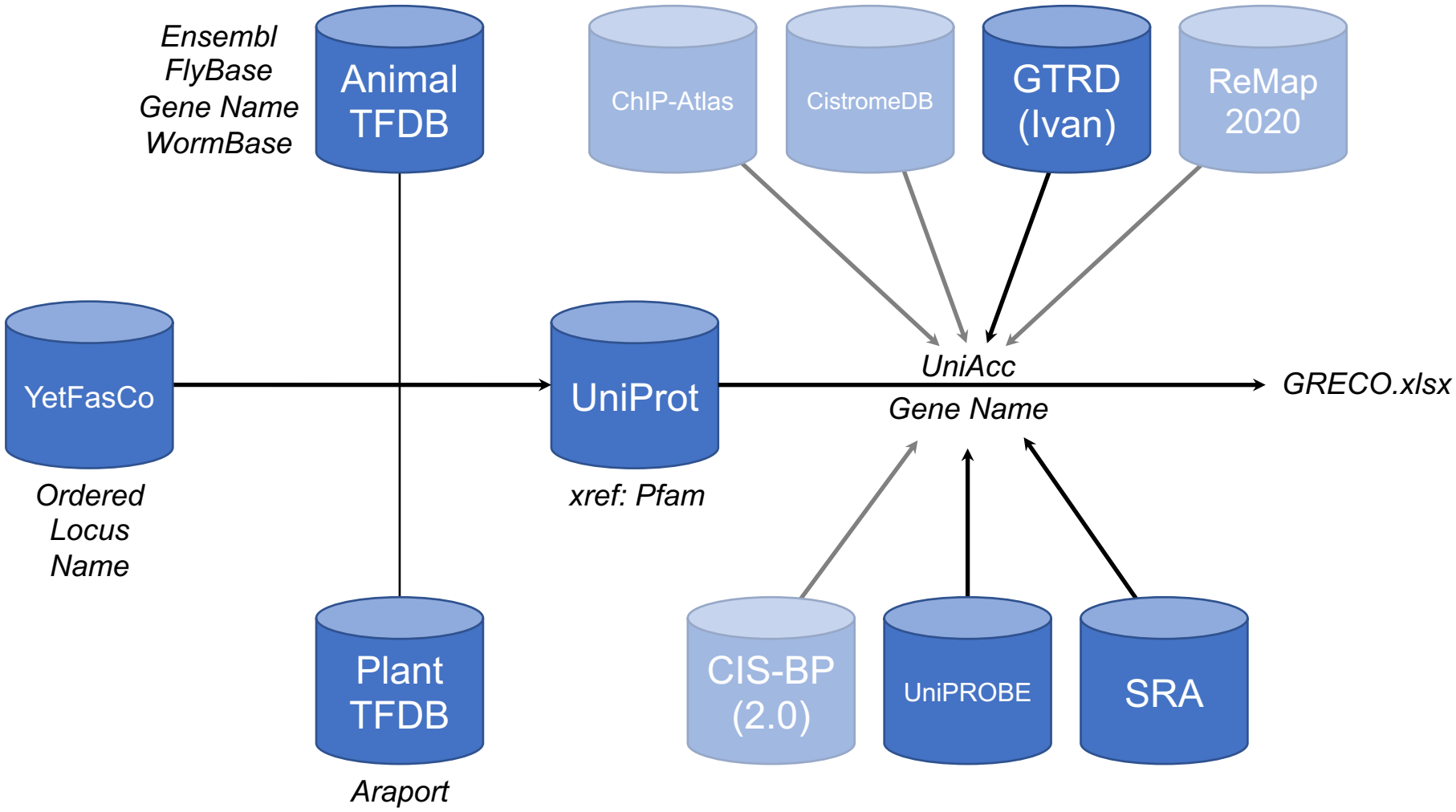
- Arabidopsis thaliana (1,717 TFs from the PlantTFDB 4.0)
- Caenorhabditis elegans (741 TFs from the AnimalTFDB 3.0)
- Drosophila melanogaster (649 TFs from the AnimalTFDB 3.0)
- Homo sapiens (1,636 TFs from the AnimalTFDB 3.0)
- Mus musculus (1,591 TFs from the AnimalTFDB 3.0)
- Saccharomyces cerevisiae (277 TFs from the YeTFaSCo)

TF Families

- Examples of TF family classification:

1. Pfam → TFs can be assigned multiple Pfam IDs
2. OrthoDB → Zinc fingers are grouped in a single (huge) family
3. TFClass → Great, but limited to mammals

Annotation


Ensembl
FlyBase
Gene Name
WormBase
Animal TFDB
ChIP-Atlas
CistromeDB
GTRD (Ivan)
ReMap 2020
YetFasCo
Ordered Locus Name
UniProt
xref: Pfam
UniAcc
Gene Name
GRECO.xlsx
Plant TFDB
Araport
CIS-BP (2.0)
UniPROBE
SRA

Annotation

- GTRD: 851 TFs mapped (2,040 features; 84%)
- DAP-seq: 436 TFs mapped (854 features; 81%)
- HT-SELEX: 668 (~96%) out of 697 TFs mapped
- CIS-BP: 851 (~80%) out of 1,065 TFs mapped
- UniPROBE: 537 (~93%) out of 576 TFs mapped
- SMiLE-seq: 54 (~92%) out of 59 TFs mapped

Clustering

- For each Pfam ID combination (*e.g.* Homeodomain + Pou):

  1. Make a FASTA database with all TF sequences that have assigned these Pfam IDs; and

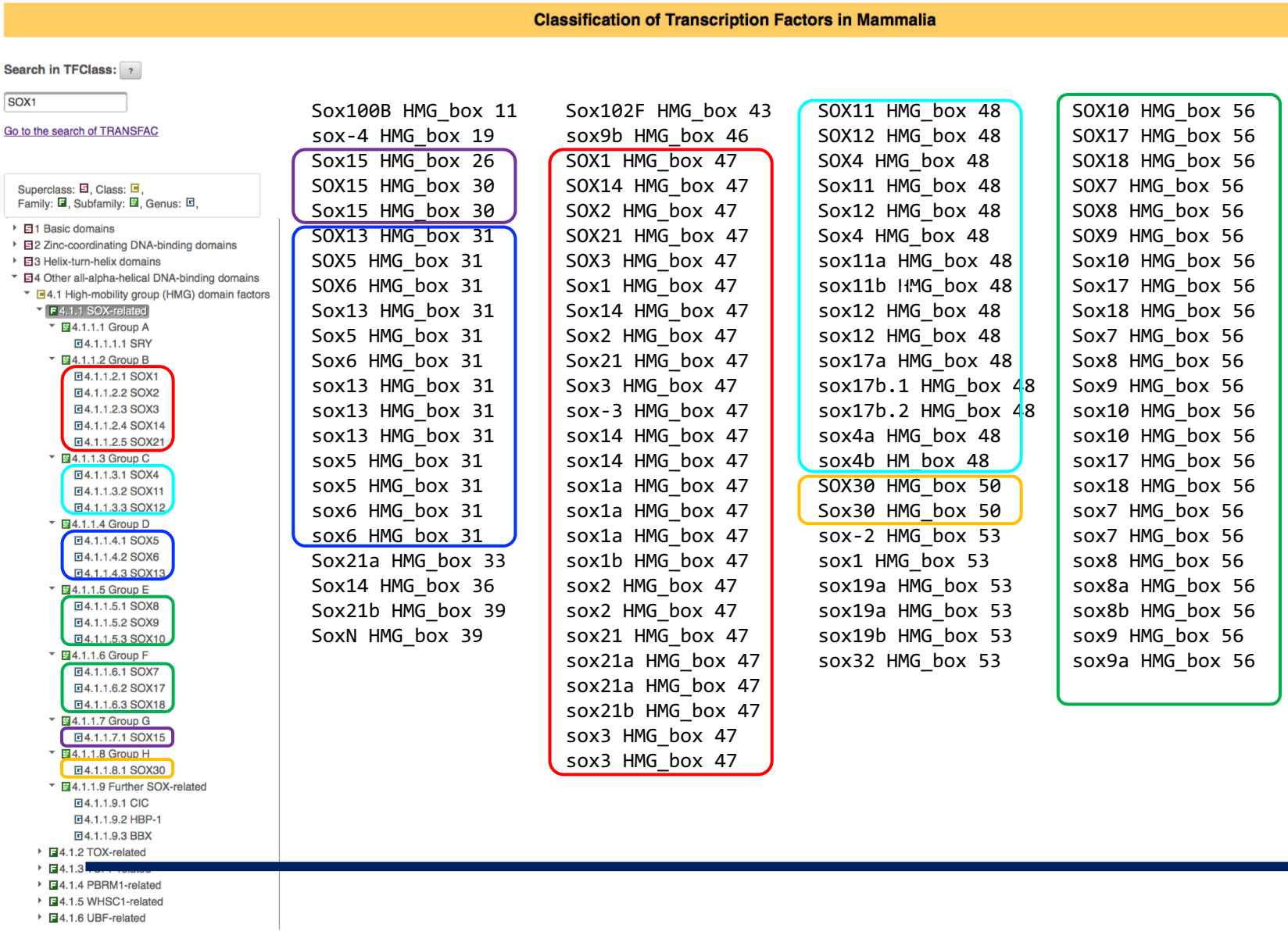
  2. Cluster sequences using MMseqs2 “easy-cluster” with options “--min-seq-id 0.3 -c 0.5”

- From the MMseqs2 user’s guide:

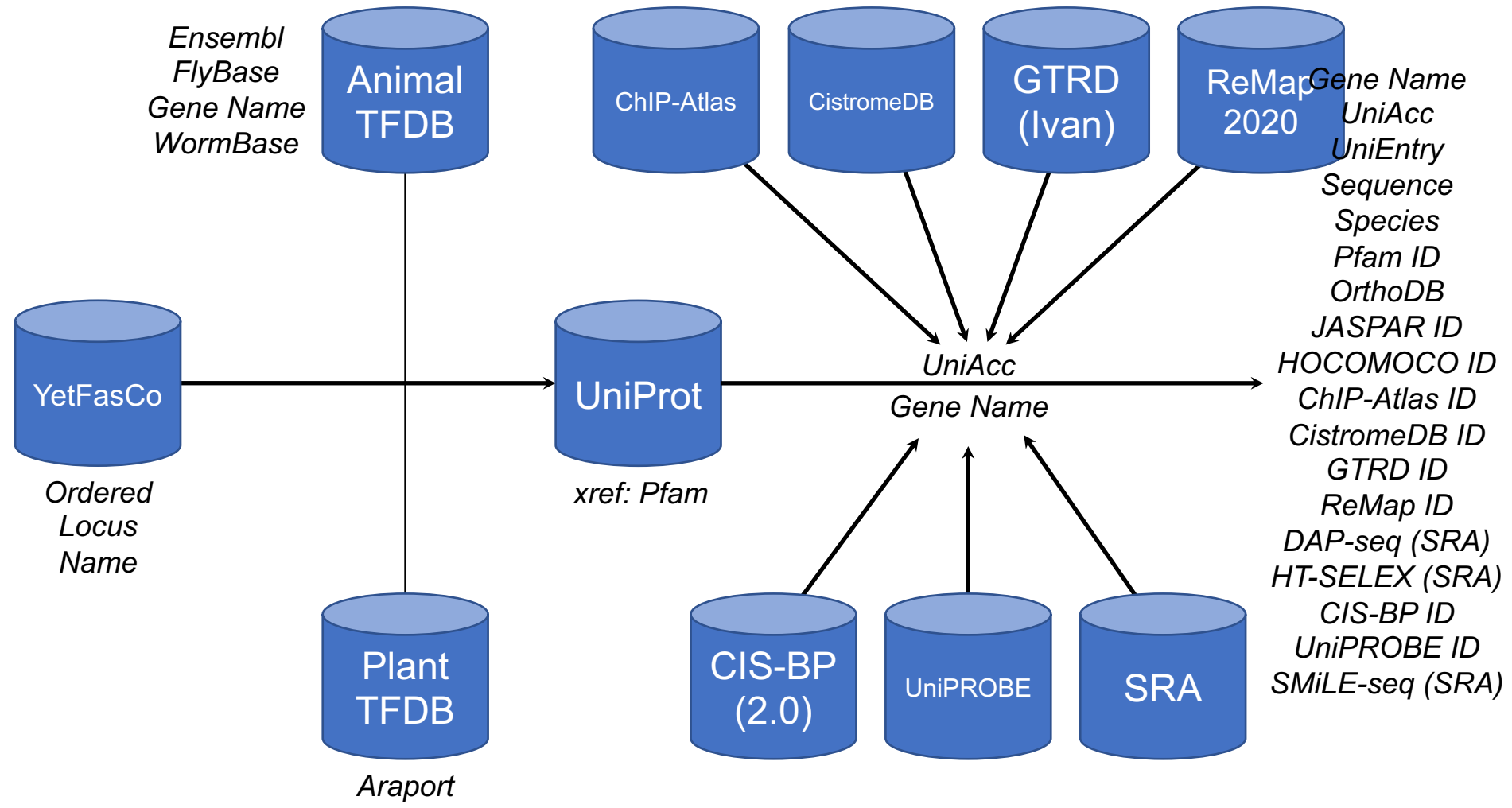
  “We are using MMseqs2 to regularly update versions of the UniProtKB database clustered down to 30% sequence similarity[1] threshold”

  [1]Most likely, authors refer to sequence identity.

Clustering


Classification of Transcription Factors in Mammalia
About TFClass
Search in TFClass:
SOX1
Go to the search of TRANSFAC
Superclass: , Class: , Family: , Subfamily: , Genus: ,
1 Basic domains
2 Zinc-coordinating DNA-binding domains
3 Helix-turn-helix domains
4 Other all-alpha-helical DNA-binding domains
4.1 High-mobility group (HMG) domain factors
4.1.1 SOX-related
4.1.1.1 Group A
4.1.1.1.1 SRY
4.1.1.2 Group B
4.1.1.2.1 SOX1
4.1.1.2.2 SOX2
4.1.1.2.3 SOX3
4.1.1.2.4 SOX14
4.1.1.2.5 SOX21
4.1.1.3 Group C
4.1.1.3.1 SOX4
4.1.1.3.2 SOX11
4.1.1.3.3 SOX12
4.1.1.4 Group D
4.1.1.4.1 SOX5
4.1.1.4.2 SOX6
4.1.1.4.3 SOX13
4.1.1.5 Group E
4.1.1.5.1 SOX8
4.1.1.5.2 SOX9
4.1.1.5.3 SOX10
4.1.1.6 Group F
4.1.1.6.1 SOX7
4.1.1.6.2 SOX17
4.1.1.6.3 SOX18
4.1.1.7 Group G
4.1.1.7.1 SOX15
4.1.1.8 Group H
4.1.1.8.1 SOX30
4.1.1.9 Further SOX-related
4.1.1.9.1 CIC
4.1.1.9.2 HBP-1
4.1.1.9.3 BBX
4.1.2 TOX-related
4.1.3
4.1.4 PBRM1-related
4.1.5 WHSC1-related
4.1.6 UBF-related
Sox100B HMG_box 11
sox-4 HMG_box 19
Sox15 HMG_box 26
SOX15 HMG_box 30
Sox15 HMG_box 30
SOX13 HMG_box 31
SOX5 HMG_box 31
SOX6 HMG_box 31
Sox13 HMG_box 31
Sox5 HMG_box 31
Sox6 HMG_box 31
sox13 HMG_box 31
sox13 HMG_box 31
sox13 HMG_box 31
sox5 HMG_box 31
sox5 HMG_box 31
sox6 HMG_box 31
sox6 HMG_box 31
Sox21a HMG_box 33
Sox14 HMG_box 36
Sox21b HMG_box 39
SoxN HMG_box 39
Sox102F HMG_box 43
sox9b HMG_box 46
SOX1 HMG_box 47
SOX14 HMG_box 47
SOX2 HMG_box 47
SOX21 HMG_box 47
SOX3 HMG_box 47
Sox1 HMG_box 47
Sox14 HMG_box 47
Sox2 HMG_box 47
Sox21 HMG_box 47
Sox3 HMG_box 47
sox-3 HMG_box 47
sox14 HMG_box 47
sox14 HMG_box 47
sox1a HMG_box 47
sox1a HMG_box 47
sox1a HMG_box 47
sox1b HMG_box 47
sox2 HMG_box 47
sox2 HMG_box 47
sox21 HMG_box 47
sox21a HMG_box 47
sox21a HMG_box 47
sox21b HMG_box 47
sox3 HMG_box 47
sox3 HMG_box 47
SOX11 HMG_box 48
SOX12 HMG_box 48
SOX4 HMG_box 48
Sox11 HMG_box 48
Sox12 HMG_box 48
Sox4 HMG_box 48
sox11a HMG_box 48
sox11b HMG_box 48
sox12 HMG_box 48
sox12 HMG_box 48
sox17a HMG_box 48
sox17b.1 HMG_box 48
sox17b.2 HMG_box 48
sox4a HMG_box 48
sox4b HM box 48
SOX30 HMG_box 50
Sox30 HMG_box 50
sox-2 HMG_box 53
sox1 HMG_box 53
sox19a HMG_box 53
sox19a HMG_box 53
sox19b HMG_box 53
sox32 HMG_box 53
SOX10 HMG_box 56
SOX17 HMG_box 56
SOX18 HMG_box 56
SOX7 HMG_box 56
SOX8 HMG_box 56
SOX9 HMG_box 56
Sox10 HMG_box 56
Sox17 HMG_box 56
Sox18 HMG_box 56
Sox7 HMG_box 56
Sox8 HMG_box 56
Sox9 HMG_box 56
sox10 HMG_box 56
sox10 HMG_box 56
sox17 HMG_box 56
sox18 HMG_box 56
sox7 HMG_box 56
sox7 HMG_box 56
sox8 HMG_box 56
sox8a HMG_box 56
sox8b HMG_box 56
sox9 HMG_box 56
sox9a HMG_box 56

Annotation


Ensembl
FlyBase
Gene Name
WormBase
Animal TFDB
ChIP-Atlas
CistromeDB
GTRD (Ivan)
ReMap 2020
Gene Name
UniAcc
UniEntry
Sequence
Species
Pfam ID
OrthoDB
JASPAR ID
HOCOMOCO ID
ChIP-Atlas ID
CistromeDB ID
GTRD ID
ReMap ID
DAP-seq (SRA)
HT-SELEX (SRA)
CIS-BP ID
UniPROBE ID
SMiLE-seq (SRA)
YetFasCo
Ordered Locus Name
UniProt
xref: Pfam
UniAcc
Gene Name
Plant TFDB
Araport
CIS-BP (2.0)
UniPROBE
SRA