GMQL in action to solve biological problems

Pietro Pinoli

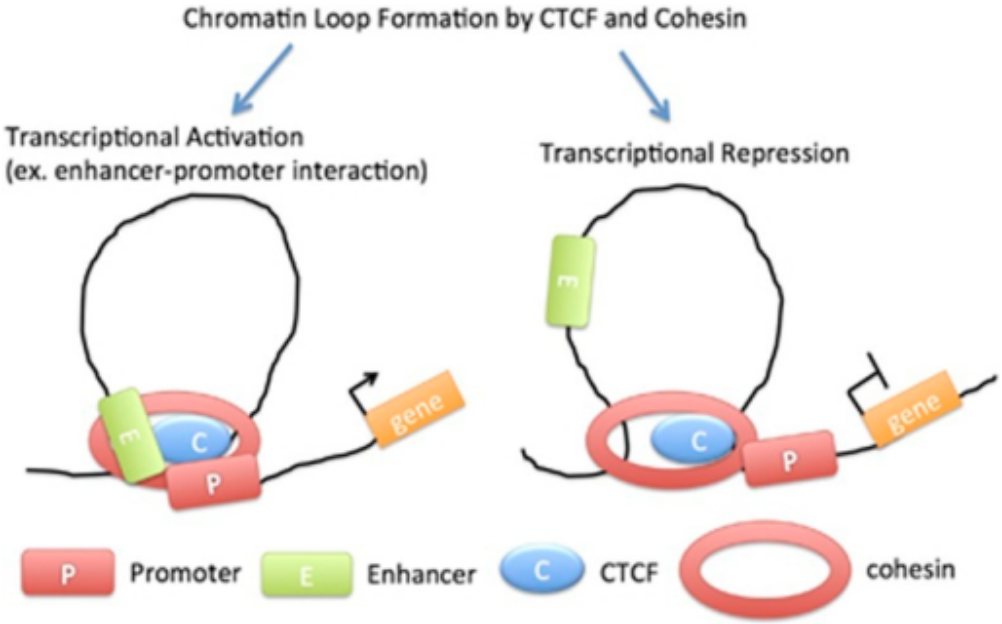
Stefano Perna

Luca Nanni

Pietro Pinoli

Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries.

Problem Statement



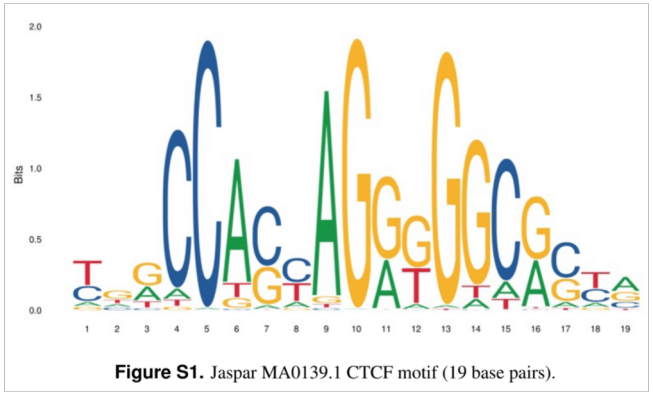
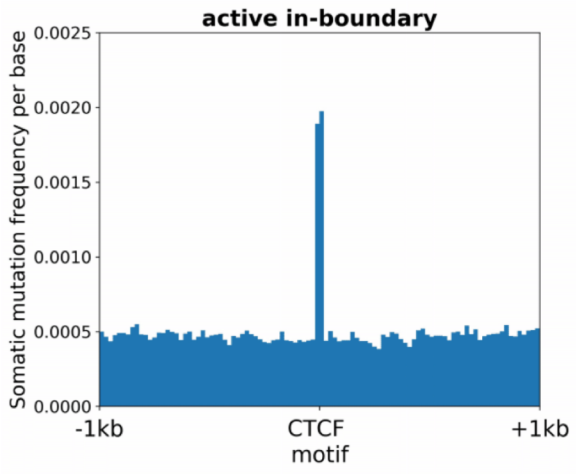
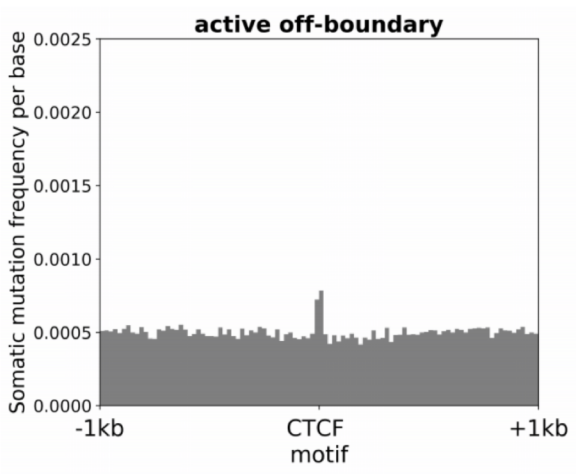


Figure S1. Jaspar MA0139.1 CTCF motif (19 base pairs).

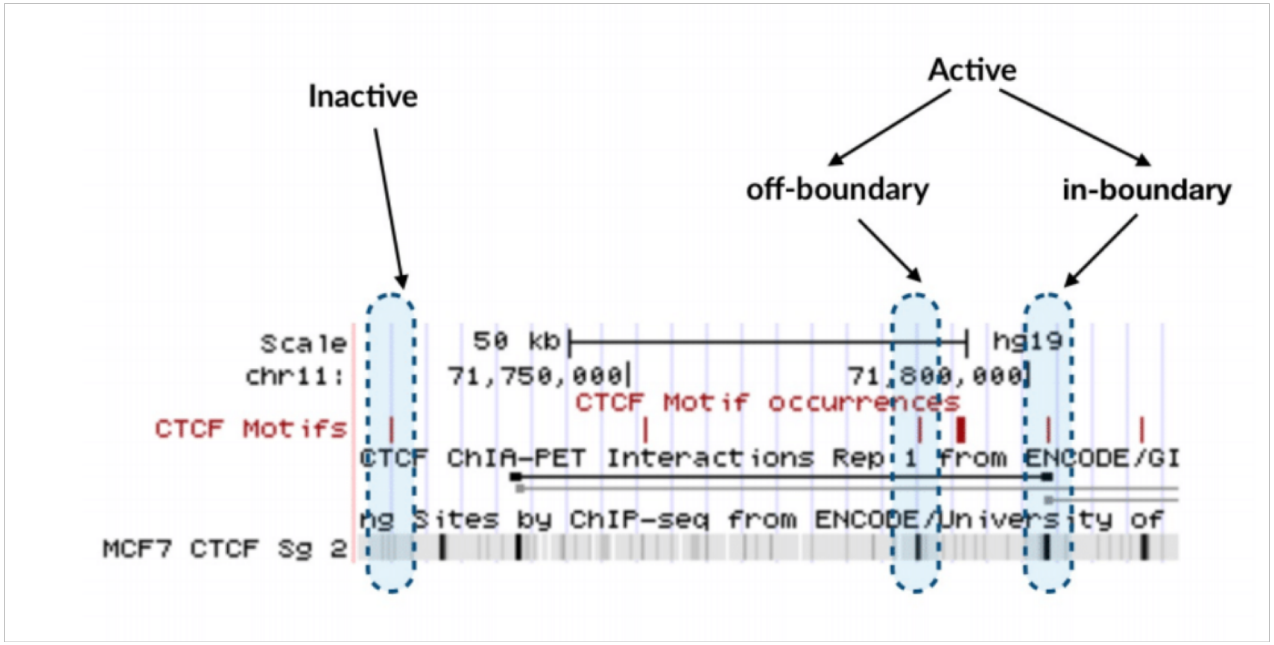


(a) hESC in-boundary, ESAD mutations



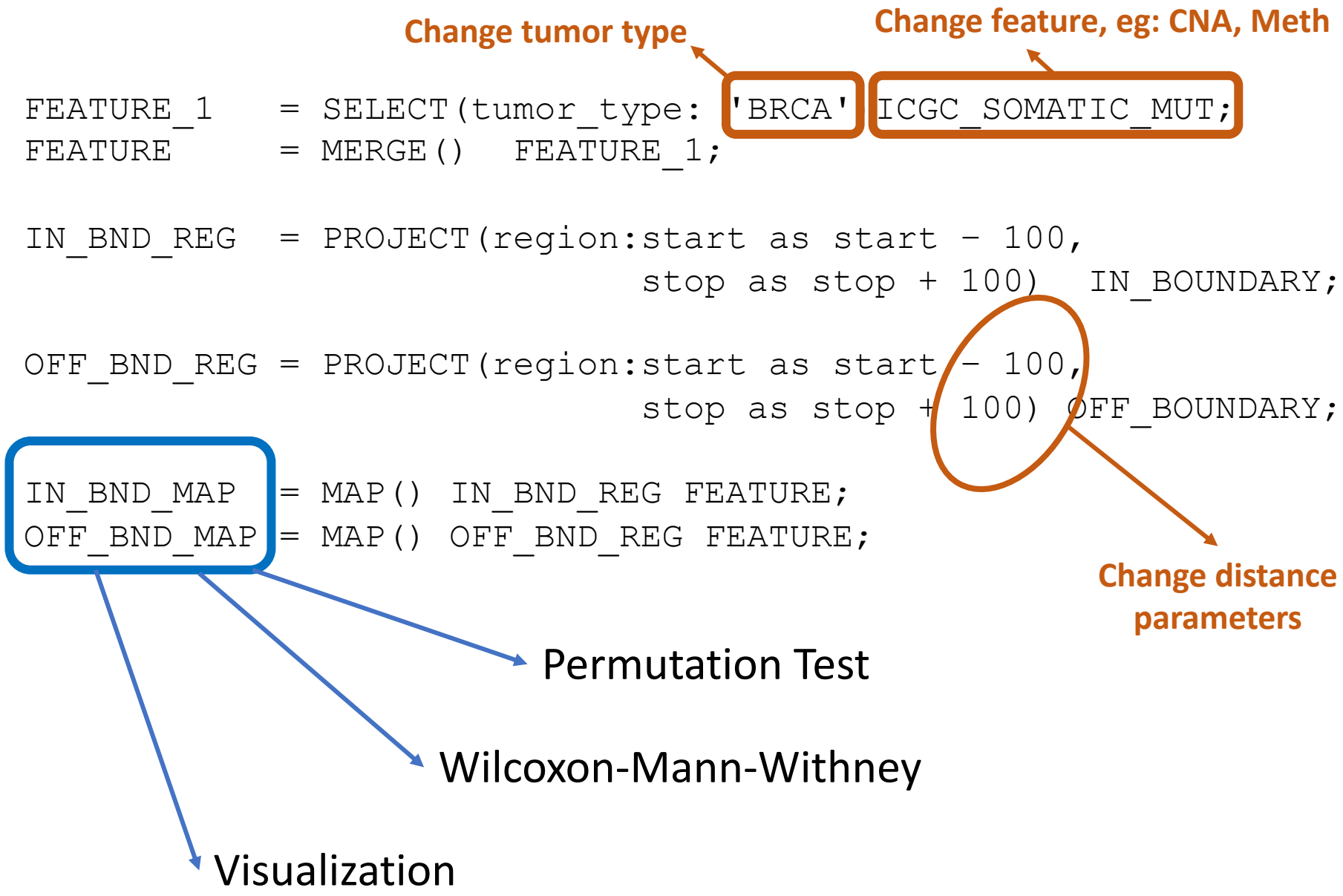
(b) hESC off-boundary, ESAD mutations

Split CTCF motifs in ***active in-boundary*** and ***active off-boundary***




```
MOTIFS         = SELECT() MOTIFS_BioStrings;
CHIP           = SELECT(cell_line == 'MCF7') ENCODE_NARROW;
ACTIVE         = JOIN(distance < 0; output: left) MOTIFS CHIP;
BOUNDARIES     = SELECT(cell_line == 'MCF7') CHIA_PET;
OVERLAP        = MAP() ACTIVE BOUNDARIES;
IN_BOUNDARY    = SELECT(region: count > 0 ) OVERLAP;
OFF_BOUNDARY   = SELECT(region: count == 0) OVERLAP;
```

Differential analysis ***in-boundary*** vs. ***off-boundary***



Results

With **~ 20 lines** of GMQL query, Sistematic analysis on:

- 3 sets of neighbourhoods
- 26 tumor types
- 3 genomic features (mutations, DNA Methylation, Copy Number Variation)

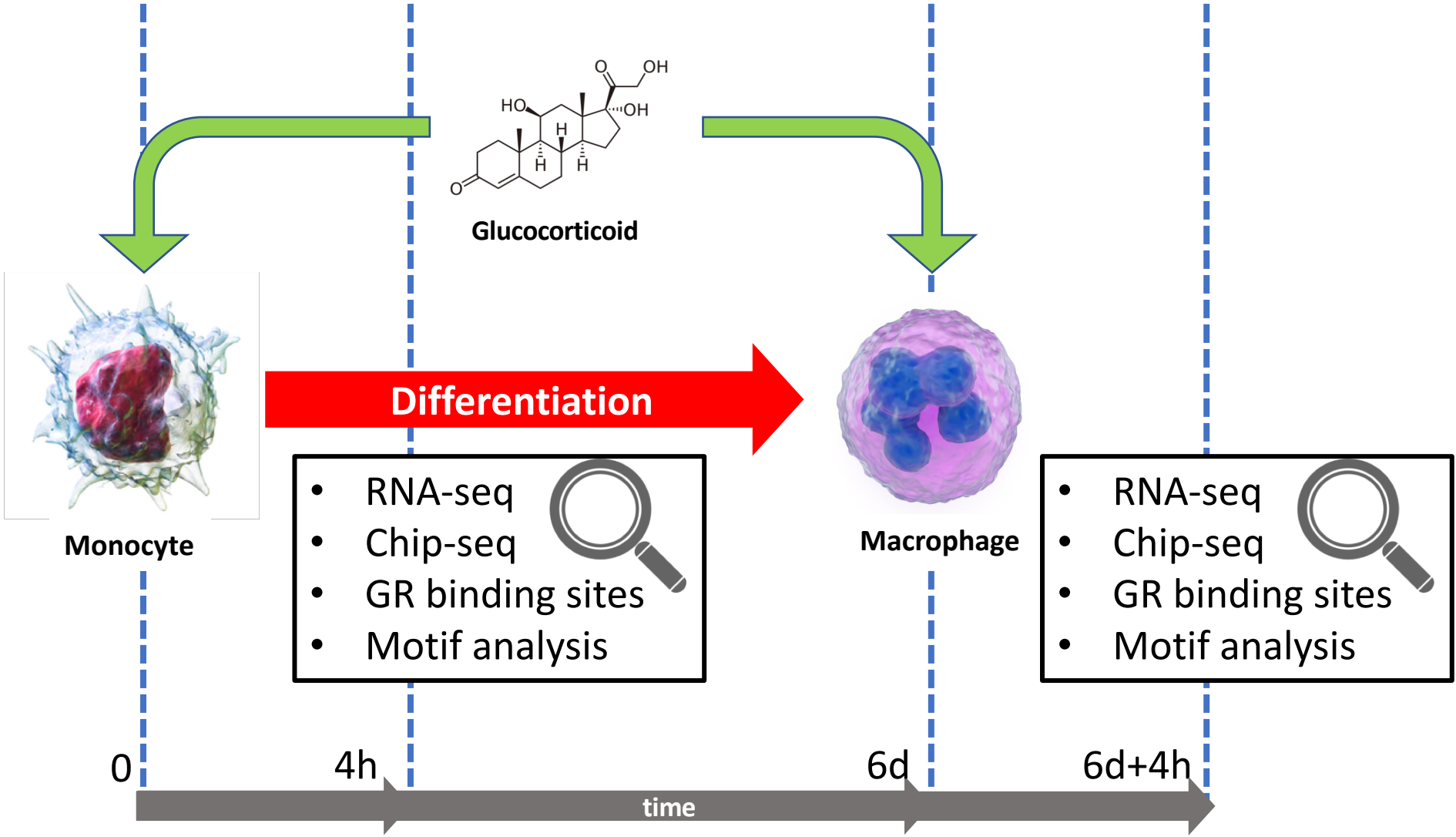
	Somatic mut.			DNA meth.			CNA		
Tumour	h	M	H	h	M	H	h	M	H
BLCA	-	-	-	N	N	N	N	N	N
BOCA	N	N	Y	-	-	-	-	-	-
BRCA	Y	Y	Y	Y	Y	Y	Y	Y	Y
BTCA	N	Y	N	-	-	-	-	-	-
COCA	N	N	N	-	-	-	-	-	-
EOPC	N	N	Y	-	-	-	-	-	-
ESAD	Y	Y	Y	-	-	-	-	-	-
GACA	Y	Y	Y	-	-	-	-	-	-
GBM	-	-	-	-	-	-	N	Y	N
HNSC	-	-	-	Y	Y	Y	N	N	N
KIRC	-	-	-	N	N	N	-	-	-
KIRP	-	-	-	-	-	-	N	N	N
LIHC	-	-	-	N	N	N	Y	Y	Y
LIRI	Y	Y	Y	-	-	-	-	-	-
LUAD	-	-	-	N	N	N	Y	Y	Y
LUSC	-	-	-	-	-	-	Y	Y	Y
MALY	N	N	N	-	-	-	-	-	-
MELA	Y	Y	Y	-	-	-	-	-	-
OV	N	Y	Y	-	-	-	Y	Y	Y
PACA	N	Y	Y	-	-	-	-	-	-
PRAD	-	-	-	Y	N	N	N	N	N
RECA	Y	Y	Y	-	-	-	-	-	-
SKCA	N	N	Y	-	-	-	-	-	-
SKCM	-	-	-	Y	Y	Y	-	-	-
THCA	-	-	-	N	N	N	N	N	N
UCEC	-	-	-	Y	Y	Y	N	N	N

Luca Nanni

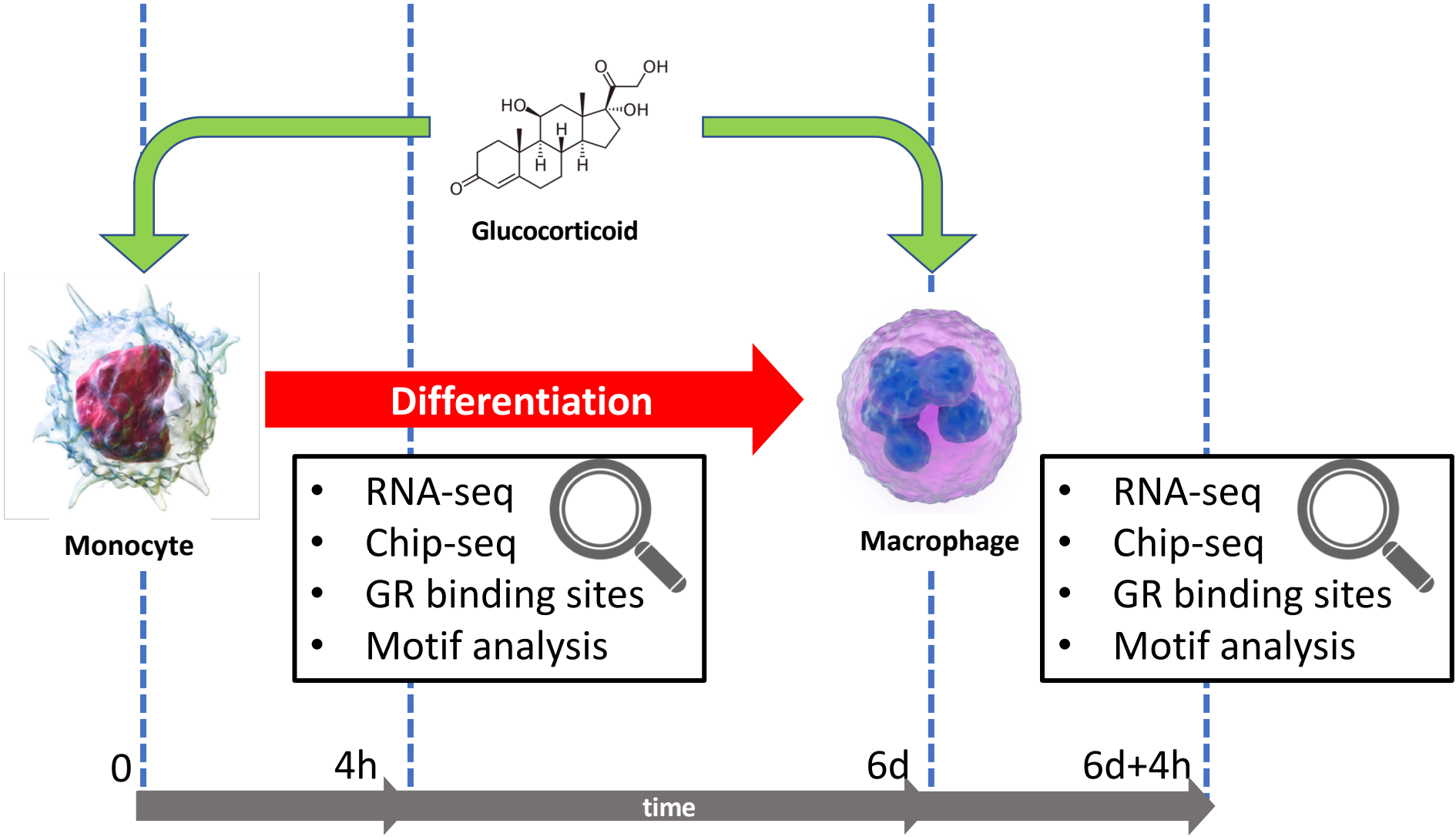
Extensive epigenomic integration of the glucocorticoid response in primary human monocytes and in vitro derived macrophages

Cheng Wang, Luca Nanni, Boris Novakovic, Wout Megchelenbrink, Tatyana Kuznetsova, Hendrik G. Stunnenberg, Stefano Ceri & Colin Logie

Experimental setup

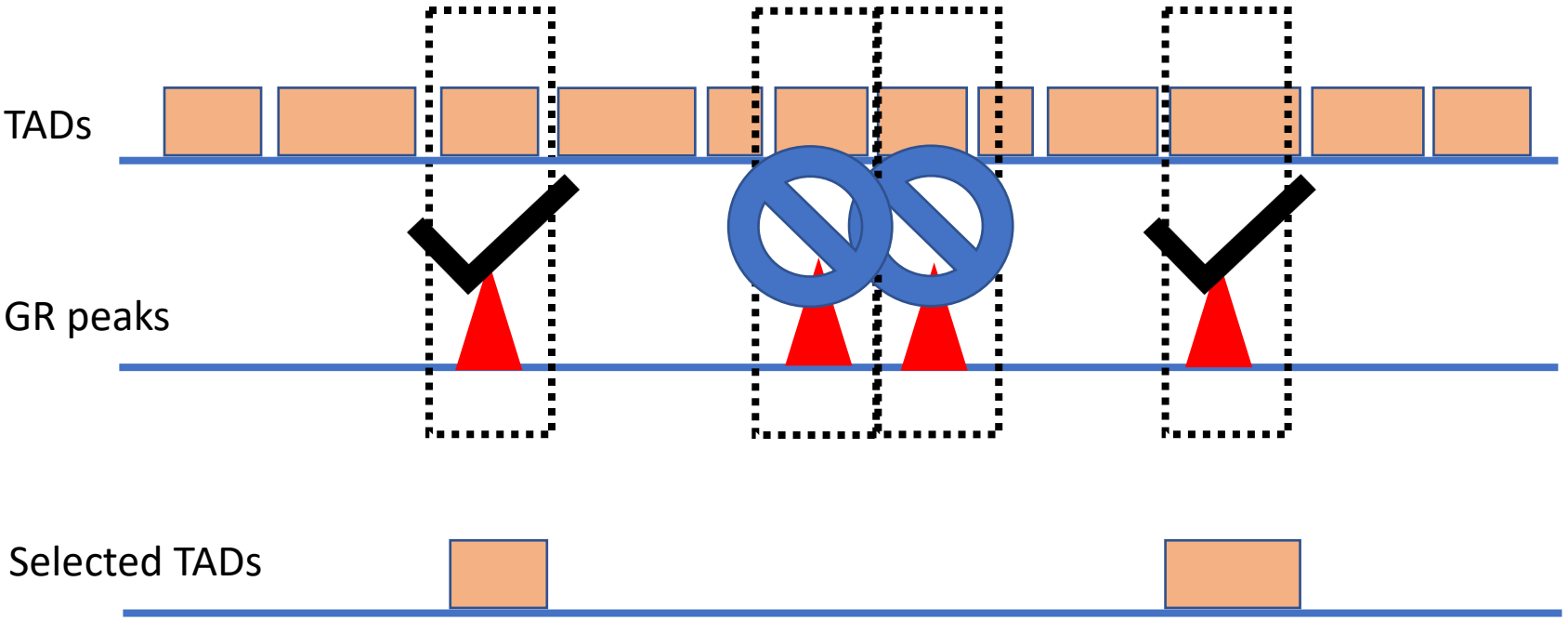


Experimental setup



Adjacency analysis

Question: *Extract the TADs bearing at least one GR peak and whose two upstream and two downstream TAD neighbors have no GR peak.*



Adjacency analysis

Question: *Extract the TADs bearing at least one GR peak and whose two upstream and two downstream TAD neighbors have no GR peak.*

```
tads = gmql.load_from_path("./tads/")

tad_tad_UP = tads.join(tads,
         [gl.UP(),gl.DLE(5e6),gl.DG(0),gl.MD(2)],
         output="LEFT", refName="LEFT",
         expName="RIGHT")

tad_tad_DOWN = tads.join(tads,
         [gl.DOWN(),gl.DLE(5e6),gl.DG(0),gl.MD(2)],
         output="LEFT", refName="LEFT",
         expName="RIGHT")
```

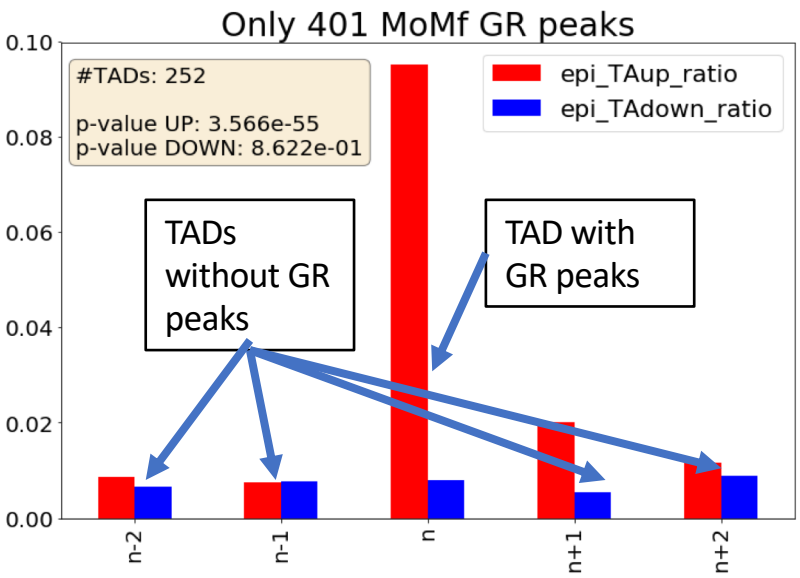
Adjacency analysis

Question: *Extract the TADs bearing at least one GR peak and whose two upstream and two downstream TAD neighbors have no GR peak.*

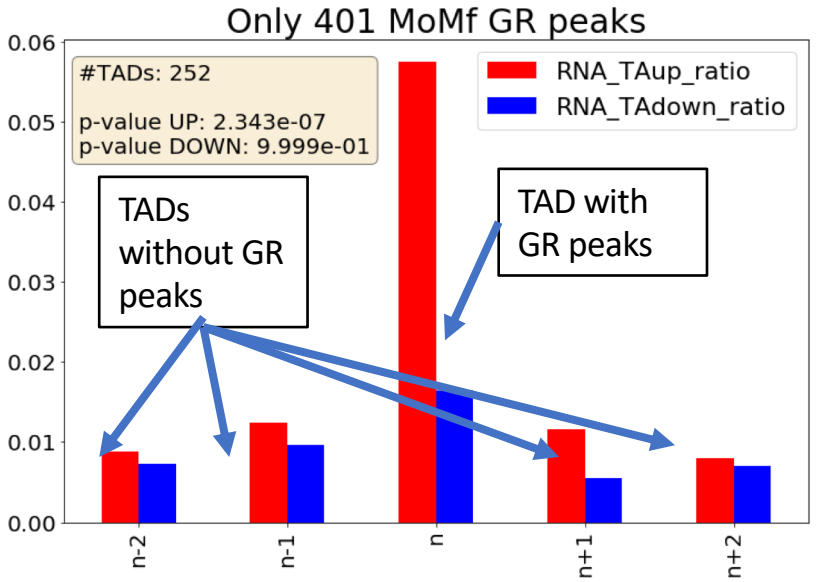
```
tad_groups = tad_tad.group(
                by=['LEFT.TAD_index'],
                agg={'near_GRs': gl.SUM("RIGHT.n_GRs")})

sel_tads = tad_groups.reg_select(
                      (tad_groups.LEFT_n_GRs > 0) &
                      (tad_groups.near_GRs == 0))

result = sel_tads.materialize()
```

Adjacency analysis

Epigenetic features



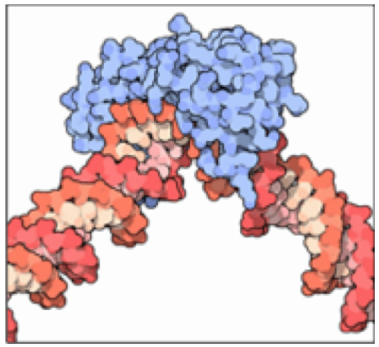
RNA features



Stefano Perna

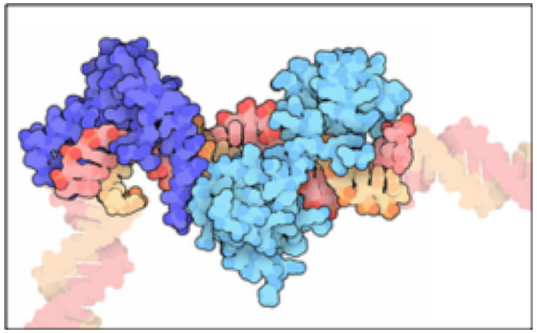
Implementing a transcription factor interaction prediction system using the Genometric Query Language.

Background



- Transcription Factors (TFs) bind and modify DNA accessibility

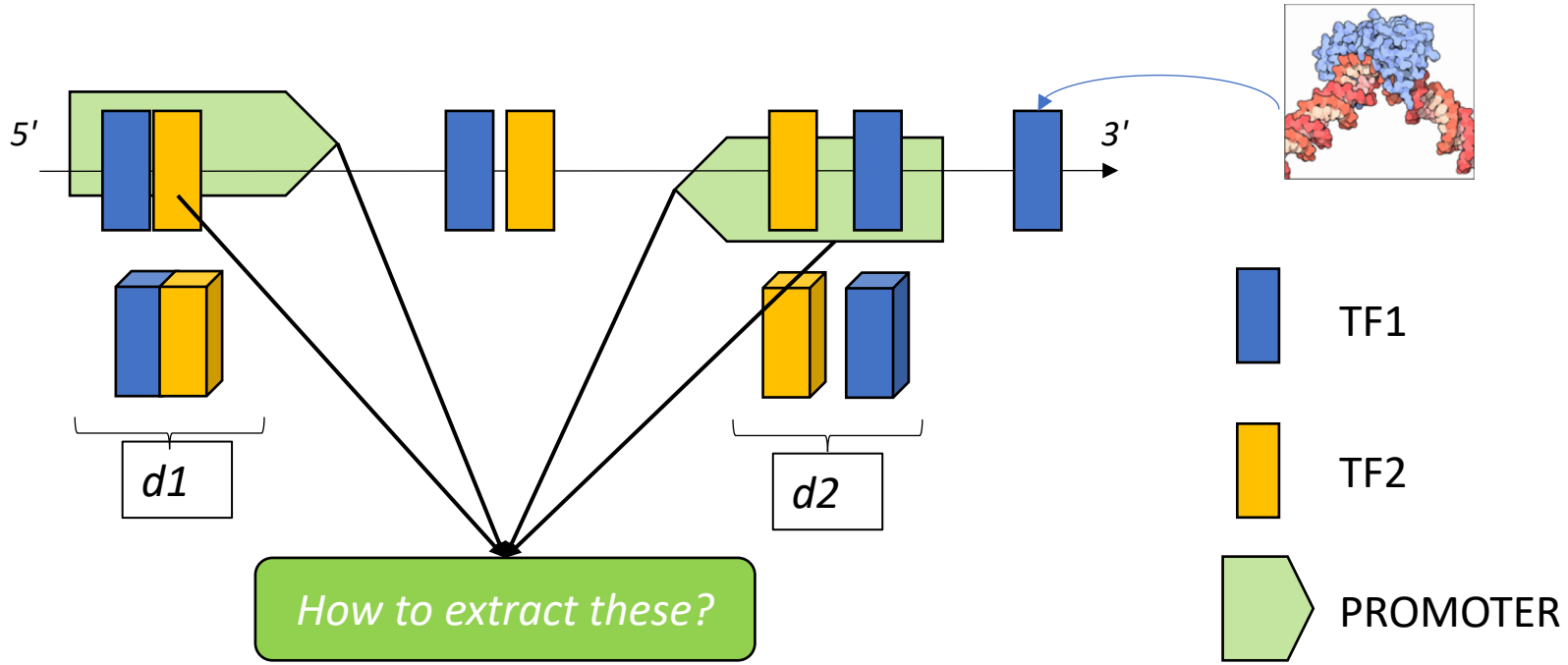
- TFs act alone or in groups
- Wet-lab investigation is combinatorial



Perna S, Canakoglu A, Pinoli P, Ceri S, Wong LS. Implementing a transcription factor interaction prediction system using the Genometric Query Language. In Hiroshi Mamitsuka, editor, Data Mining for Systems Biology, chapter 6, pages 63-81, Springer Protocols, 2018.

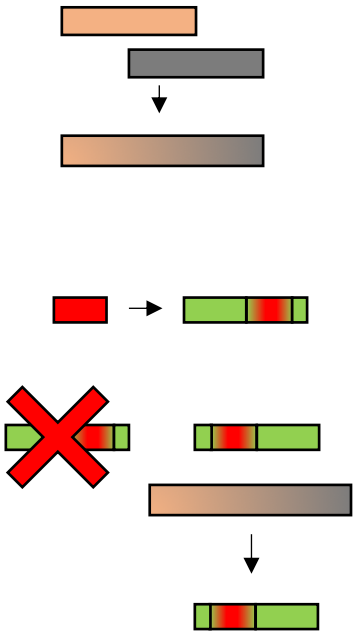
Data & Modeling

ChIP-seq experiment → Binding site information



Perna S, Canakoglu A, Pinoli P, Ceri S, Wong LS. Implementing a transcription factor interaction prediction system using the Genometric Query Language. In Hiroshi Mamitsuka, editor, Data Mining for Systems Biology, chapter 6, pages 63-81, Springer Protocols, 2018.

ETL and filtering with GMQL



Replicate merging

Region (stranded) extension

Filtering via overlap

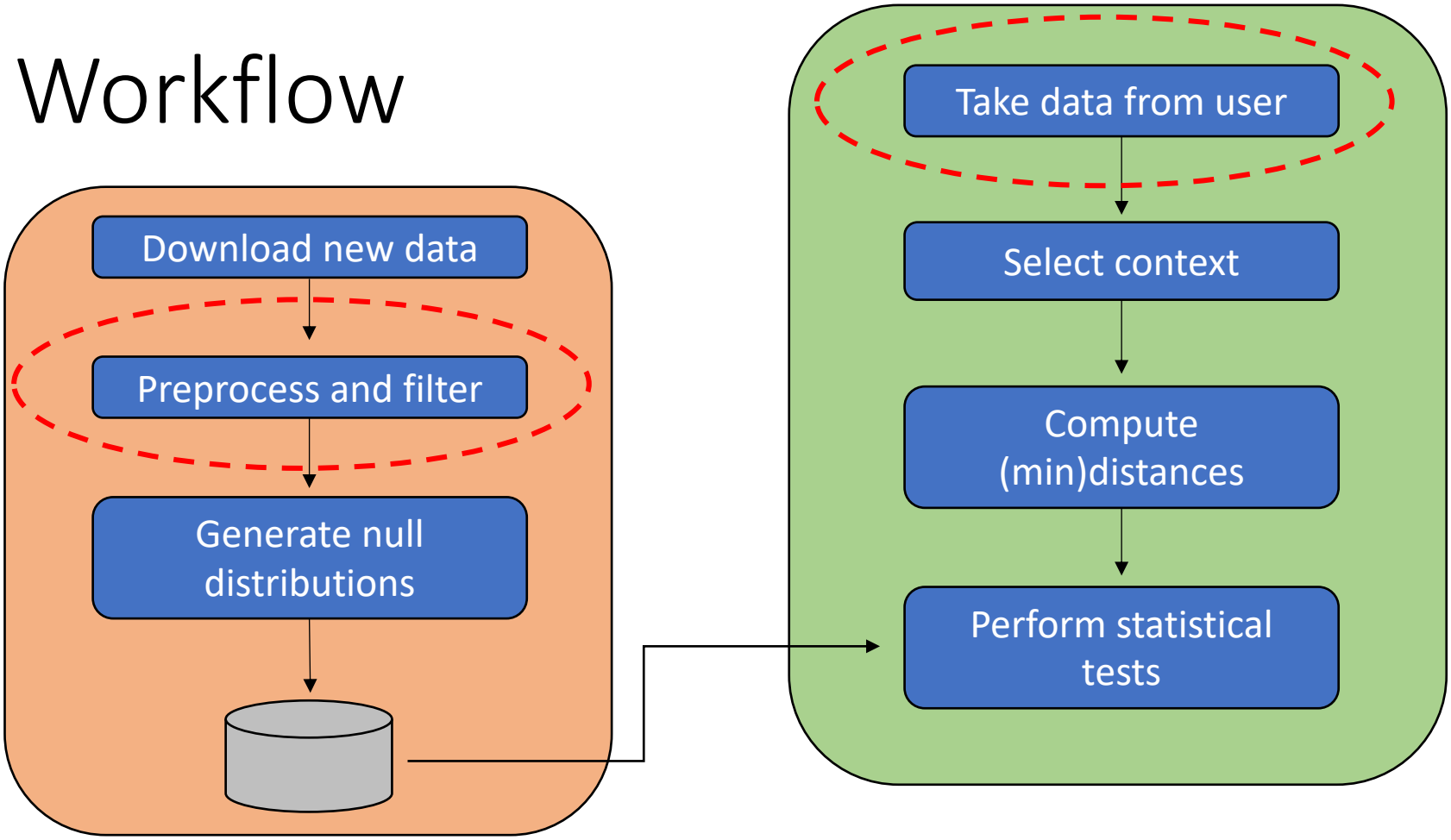
Query editor

```
1  HMs_0 = SELECT((histone_name == 'H3K9ac'
2                  OR histone_name == 'H3K4me3')
3                 AND cell == 'target_cell') ENCODE_BROADPEAK;
4  HMs = COVER(1,ANY) HMs_0;
5
6  GENE_ISOFORMS = SELECT(annotation_type == "transcript"
7                                        AND assembly=="GRCh38";
8                                 HMs: gene_type == "protein_coding"
9                                                  ) GENCODE_ANNOTATION;
10 PROMS = PROJECT(region_update: start AS start - 2000,
11                 stop AS start + 200) GENE_ISOFORMS;
12
13 HM_PROMS = MAP(count_name: hits) PROMS HMs;
14 ACTIVE_PROMS = SELECT(region: hits == 1) HM_PROMS;
15 MATERIALIZE ACTIVE_PROMS into active_proms;
16
```

Similar queries for TFBS filtering and TFBS-to-promoter mapping

Perna S, Canakoglu A, Pinoli P, Ceri S, Wong LS. Implementing a transcription factor interaction prediction system using the Genometric Query Language. In Hiroshi Mamitsuka, editor, Data Mining for Systems Biology, chapter 6, pages 63-81, Springer Protocols, 2018.

Workflow

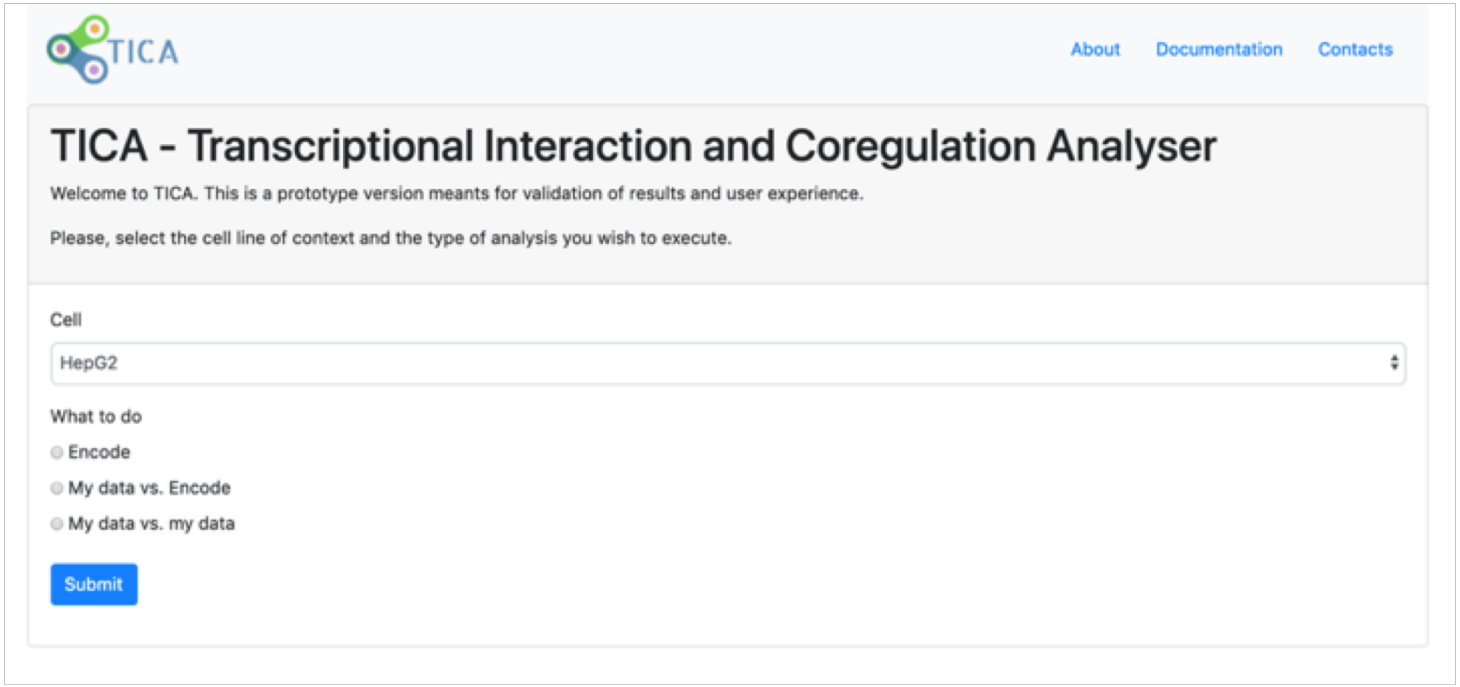


Costly, done few times

Quick, on demand

Perna S, Canakoglu A, Pinoli P, Ceri S, Wong LS. Implementing a transcription factor interaction prediction system using the Genometric Query Language. In Hiroshi Mamitsuka, editor, Data Mining for Systems Biology, chapter 6, pages 63-81, Springer Protocols, 2018.

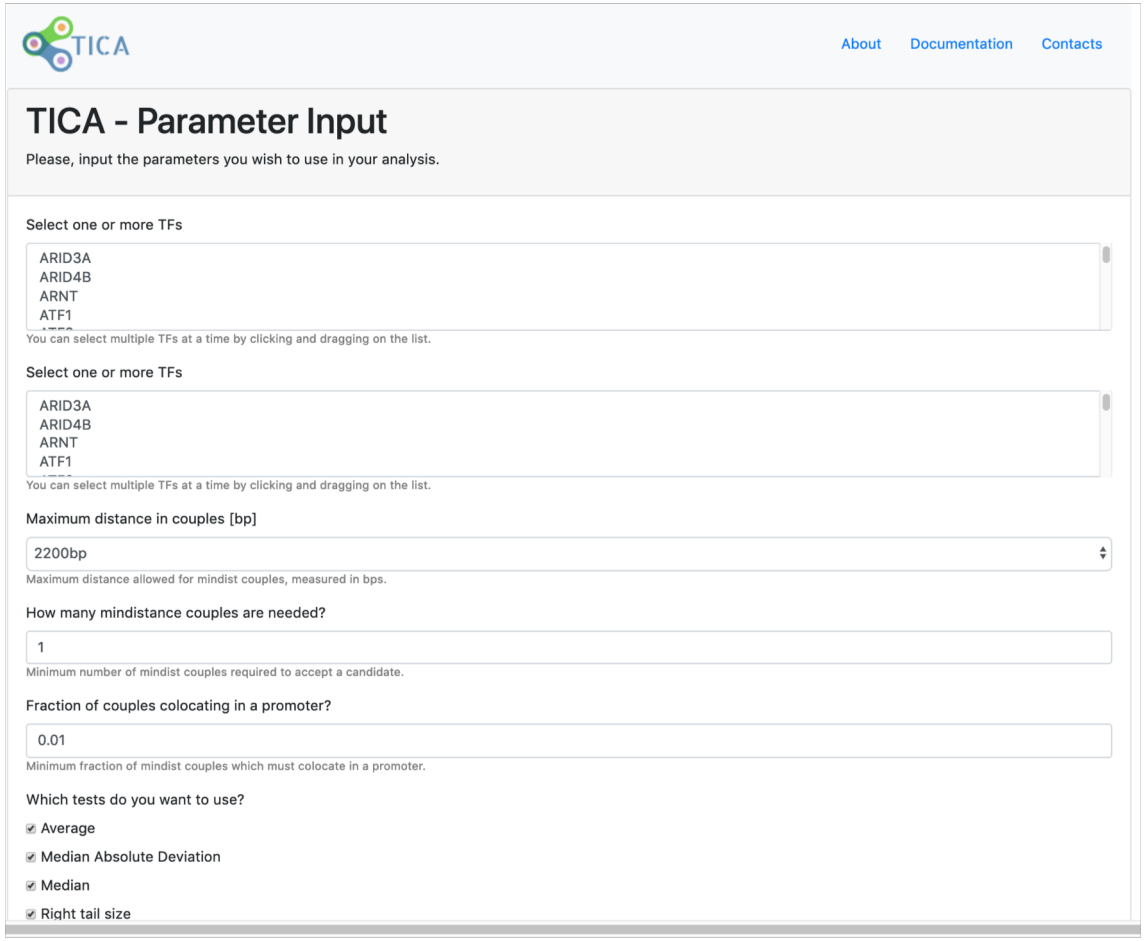
Web Service



Web service on top of (Py)GMQL queries

Perna S, Canakoglu A, Pinoli P, Ceri S, Wong LS. Implementing a transcription factor interaction prediction system using the Genometric Query Language. In Hiroshi Mamitsuka, editor, Data Mining for Systems Biology, chapter 6, pages 63-81, Springer Protocols, 2018.

Web Service



Perna S, Canakoglu A, Pinoli P, Ceri S, Wong LS. Implementing a transcription factor interaction prediction system using the Genometric Query Language. In Hiroshi Mamitsuka, editor, Data Mining for Systems Biology, chapter 6, pages 63-81, Springer Protocols, 2018.

gmql.eu/tica

TICA

About Documentation Contacts

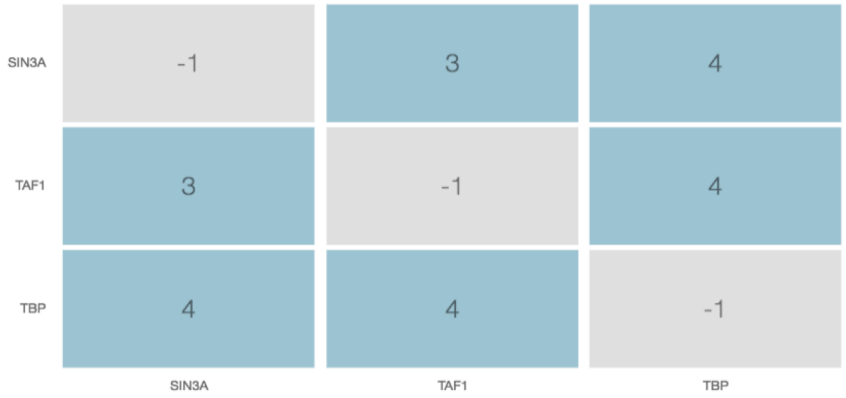
TICA - Results

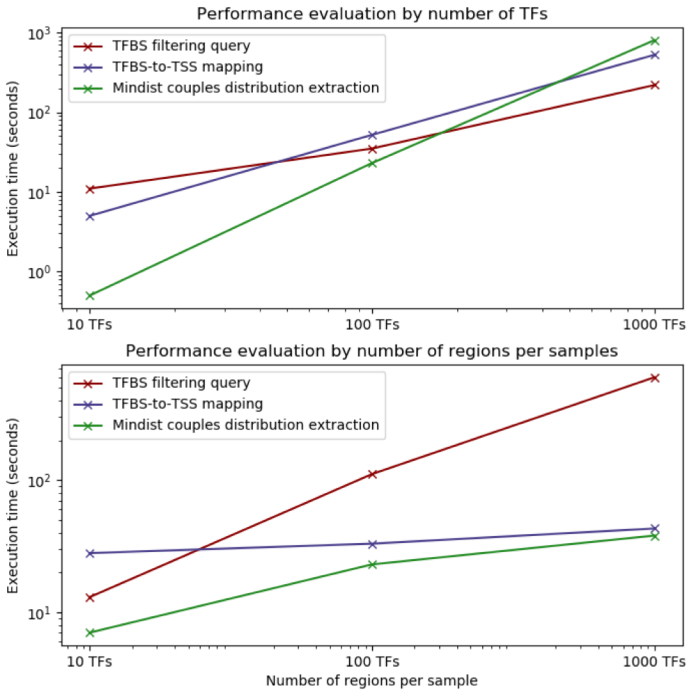
Analysis complete. Please review results below.

Testing on *SIN3A TAF1 TBP* vs. *SIN3A TAF1 TBP* on cell line hepg2 returned the following results:

Export as CSV

Name tf1	Name tf2	Couples	Couples Tss	Average	Average Passed	Median	Median Passed	Mad	Mad Passed	Tail 1000	Tail 1000 Passed
SIN3A	TAF1	104	0.654	111.644	Passed	44.0	Passed	35.0	Passed	0.019	Failed
SIN3A	TBP	878	0.61	105.196	Passed	47.0	Passed	39.0	Passed	0.003	Passed
TAF1	SIN3A	104	0.654	111.644	Passed	44.0	Passed	35.0	Passed	0.019	Failed
TAF1	TBP	1235	0.593	98.548	Passed	41.0	Passed	34.0	Passed	0.002	Passed
TBP	SIN3A	878	0.61	105.196	Passed	47.0	Passed	39.0	Passed	0.003	Passed
TBP	TAF1	1235	0.593	98.548	Passed	41.0	Passed	34.0	Passed	0.002	Passed





Perna S, Canakoglu A, Pinoli P, Ceri S, Wong LS. Implementing a transcription factor interaction prediction system using the Genometric Query Language. In Hiroshi Mamitsuka, editor, Data Mining for Systems Biology, chapter 6, pages 63-81, Springer Protocols, 2018.