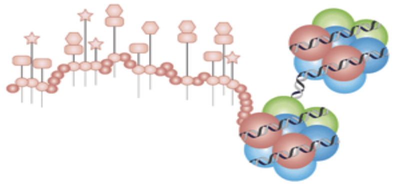
CBS MiniHack - MPRA Challenge Kickoff Talk

Jason Ernst

Professor

University of California, Los Angeles

From the epigenomic data can identify tens of thousands of candidate regulatory regions in a single cell type



From the epigenomic data can identify tens of thousands of candidate regulatory regions in a single cell type

A next challenge: test these regions and map at high resolution activating and repressive nucleotides within them

Massively Parallel Reporter Assay (MPRA)

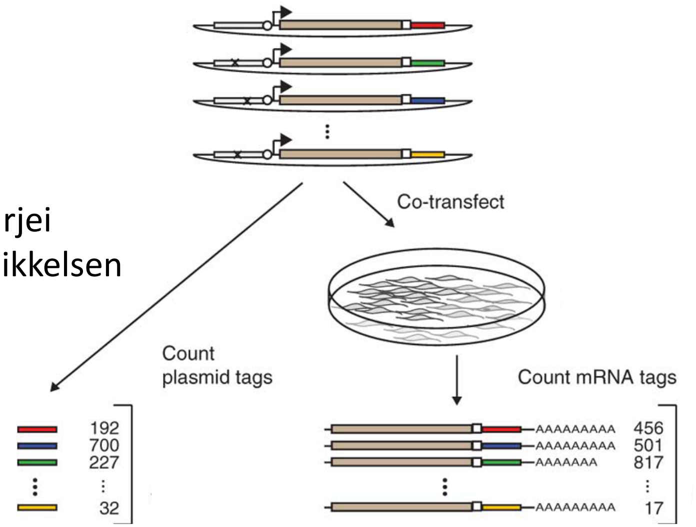
- Massively parallel reporter assay enables obtaining a quantitative readout of the ability to activate or repress gene expression for 244,000 user specified DNA sequences of 145 base pairs length

Sequence	Expression
CACCGGTGGATGTTCCCATTCAAAAGGCAAAAAAAAAGTTATATTGCTAAGTAACCTTTAACAATGAATACCATTCATTATGTATTTGCTTAATGATACAAAGGCATTGTTTGCTCTATTGCAGGTGATAACAGCAGATCAAAGA	2.22
ATTTTCAACACAGGAATTTGTTTCCCTGGTTGAGACATAAAACATGTGTCTTAATGGTTCTGCAATGTGCTGGGATTTACTCTGAAAGGATAGTGATTCACGAAGAGATGGAGTCGGGTTTCACATCCTTTGAAGTCTAGGGTTG	-0.17
GGAATGGGCCAGGGAGAGAGCCCGGACTTTACTCCGGCTGTGGTGACTCACGCCTTTGGGAAGGGCTGAGGTATTTGTCCTTGGATGTGGGGATGCTGTGGCCTCTGGTTGGGAACACAGCCTCGTTTTGTCTGCTGGCTCTGGC	3.84
GTAACGTTTTTGCCCCAGGATGCCAAAGCAAACACTGGTCCTGAACTTTGCCTCTTGAACTACATTCTTCACCATGATGCCCTTCCAAATTCCCTATAGATGTTTTCACTACTGGTGTGTACAGGAAGGAAGTTCACTTTGACTA	0.61
TGTGCTGTTGTGTAAAGCAGATTCAAGCTATTCTATTTGCTTCTTCTTATTTGTCCACTCACACAGTCACTATTTATGATGATCCTTCTATGATGAAGGCACTATTTGAGATGCCTGAAAACACAAAAGTGAGCAAGATAAACGA	0.02
TAATGCAGGTGGACCTAGTGAAGCTCTTCACCAAAATGCGTGTGTACTTGGCCCAGCGCCACACTCTGGCTTGTTGGGATCAAGGCCTGGAGGCTGGAAGGGGTAGACTGGACCCGAGGCCAGGCCGGGGCATTGCTTAACAGAA	0.32
TTTTGTTTGCATAGCTTCTCCCAGGGTCATCTTCCCACCCCTCAAGTGTGGATTTGGCCACAACACTTTGTTTAGCAATTTCCTGTCATTTTGTTTTATTTGGTCAGGCTTATTTGCATAGTTGAATGAATGCACAAAGGCCTCC	0.78
GGAGCTGGGAGCCAGAAAAGGAAGAGGACAGCCCCACCTTGCAGACCCCACCCTCTAGGAATGGATTTATGCTTCATTGTTGGTGGGAACCATAGCAACCACAGATGACCATGAGTCAGTAGATAATAAACAGGATTCTTTACAA	-0.23
GTGATCCAGGATCATGAAGCCGTGACAATGTCCATGATTGGATGTCAGGAGCAGATGATGTTAATGATTAACCGTGGGTAATGTCTACAATATGGATACAGTGCAAAAGGGATGATTCACATCCTACGTGGAATAGTGCAGGATT	0.30

.
.
.

~244,000 sequences

Tarjei Mikkelsen



Melnikov et al, *Nature Biotech* 2012; Ernst et al, *Nature Biotech* 2016

Massively Parallel Reporter Assay (MPRA)

- Massively parallel reporter assay enables obtaining a quantitative readout of the ability to activate or repress gene expression for 244,000 user specified DNA sequences of 145 base pairs length

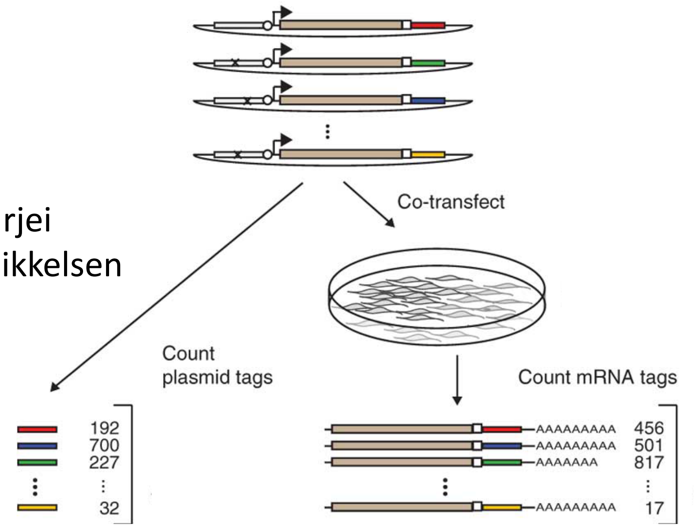
Sequence	Expression
CACCGGTGGATGTTCCCATTCAAAAGGCAAAAAAAAAGTTATATTGCTAAGTAACCTTTAACAATGAATACCATTCATTATGTATTTGCTTAATGATACAAAGGCATTGTTTGCTCTATTGCAGGTGATAACAGCAGATCAAAGA	2.22
ATTTTCAACACAGGAATTTGTTTCCCTGGTTGAGACATAAAACATGTGTCTTAATGGTTCTGCAATGTGCTGGGATTTACTCTGAAAGGATAGTGATTCACGAAGAGATGGAGTCGGGTTTCACATCCTTTGAAGTCTAGGGTTG	-0.17
GGAATGGGCCAGGGAGAGAGCCCGGACTTTACTCCGGCTGTGGTGACTCACGCCTTTGGGAAGGGCTGAGGTATTTGTCCTTGGATGTGGGGATGCTGTGGCCTCTGGTTGGGAACACAGCCTCGTTTTGTCTGCTGGCTCTGGC	3.84
GTAAC	
TGTGC	
TAATG	
TTTTG	
GGAGC	
GTGAT	

Problem: Leverage MPRA to map at close to nucleotide resolution activating and repressive bases for thousands of regions and not knowing the precise 145 base pairs to test.

.
.
.

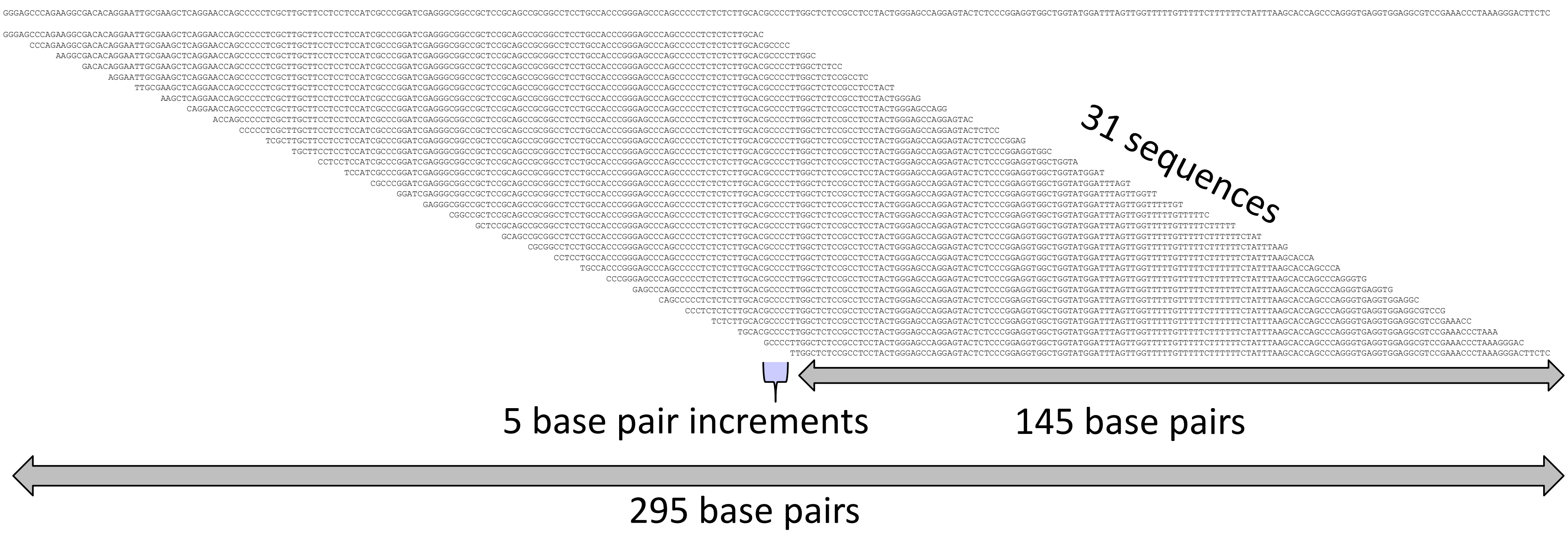
~244,000 sequences

Tarjei Mikkelsen



Melnikov et al, *Nature Biotech* 2012; Ernst et al, *Nature Biotech* 2016

Tiling Strategy



```
GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGACTTCTC

GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCAC
     CCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCC
          AAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGC
               GACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCC
                    AGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTC
                         TTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACT
                              AAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAG
                                   CAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGG
                                        ACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTAC
                                             CCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCC
                                                  TCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAG
                                                       TGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGC
                                                            CCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTA
                                                                 TCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGAT
                                                                      CGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGT
                                                                           GGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTT
                                                                                GAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGT
                                                                                     CGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTC
                                                                                          GCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTT
                                                                                               GCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTAT
                                                                                                    CGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAG
                                                                                                         CCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCA
                                                                                                              TGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCA
                                                                                                                   CCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTG
                                                                                                                        GAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTG
                                                                                                                             CAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGC
                                                                                                                                  CCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCG
                                                                                                                                       TCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACC
                                                                                                                                            TGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAA
                                                                                                                                                 GCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGAC
                                                                                                                                                      TTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGACTTCTC
```

31 sequences

5 base pair increments

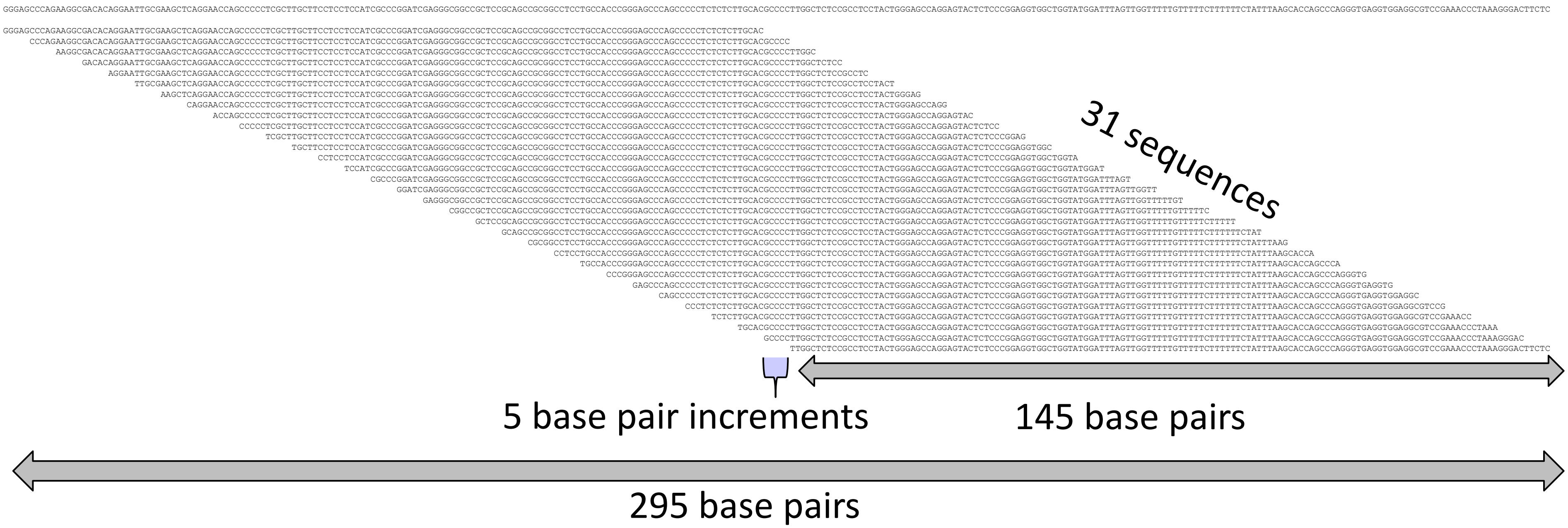
145 base pairs

295 base pairs

- ~8,000 regions can be tested in a single experiment
- Coverage of 295 base pairs for each regulatory region tested
- Information available to recover at high resolution activating and repressive bases

Ernst et al, *Nature Biotech* 2016

Tiling Strategy


GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGGACTTCTC
GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCAC
CCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCC
AAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGC
GACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCC
AGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTC
TTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACT
AAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAG
CAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGG
ACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTAC
CCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCC
TCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAG
TGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGC
CCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTA
TCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGAT
CGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGT
GGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTT
GAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGT
CGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTC
GCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTT
GCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTAT
CGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAG
CCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCA
TGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCA
CCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTG
GAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTG
CAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGC
CCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCG
TCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACC
TGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAA
GCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGGAC
TTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGGACTTCTC
31 sequences
5 base pair increments
145 base pairs
295 base pairs

- ~8,000 regions can be tested in a single experiment
- Coverage of 295 base pairs for each regulatory region tested
- Information available to recover at high resolution activating and repressive bases

Ernst et al, *Nature Biotech* 2016

Isolating Key Activating or Repressing Sequences

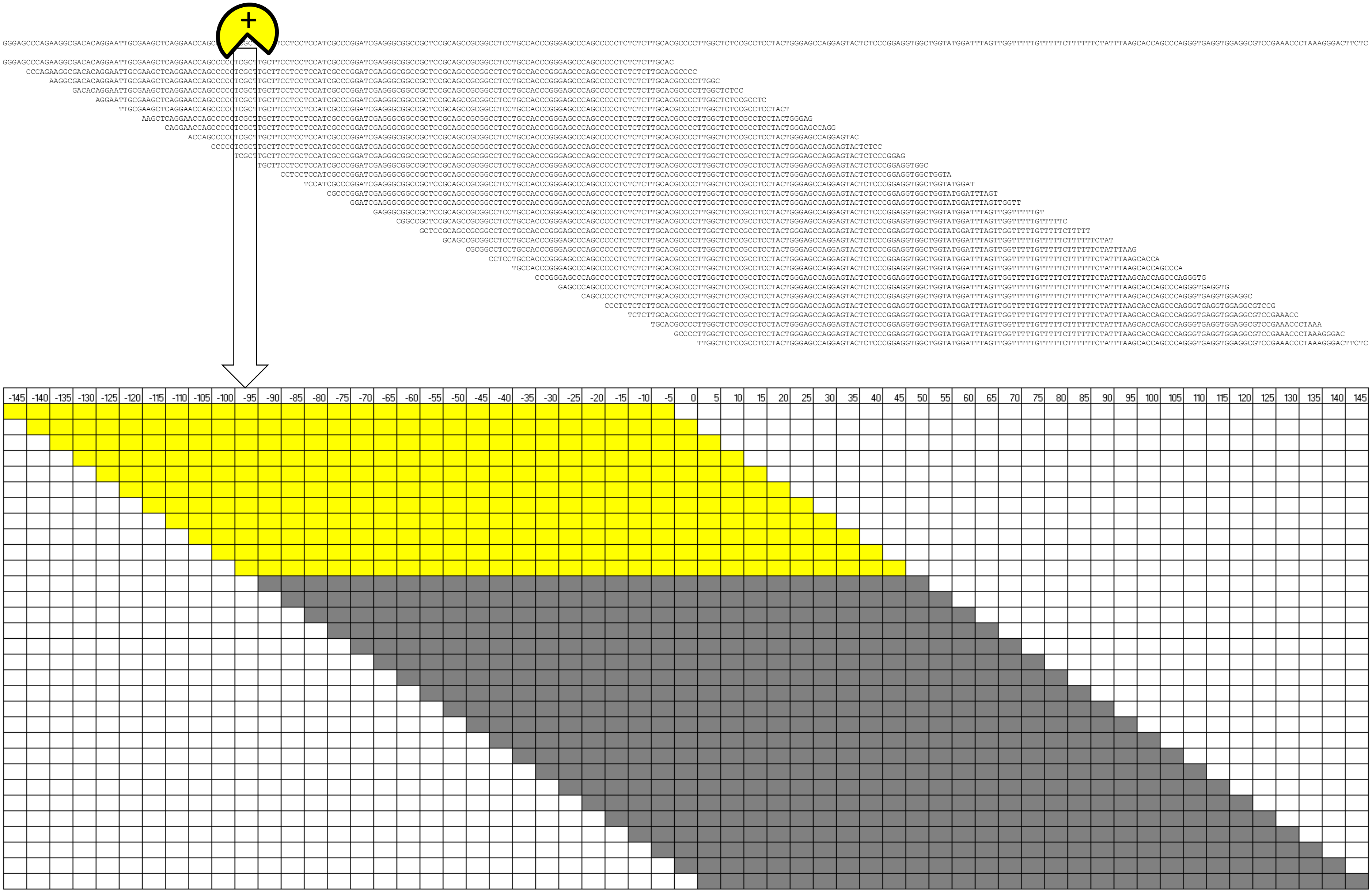
```
GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGGACTTCTC

GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCAC
     CCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCC
          AAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGC
               GACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCC
                    AGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTC
                         TTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACT
                              AAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAG
                                   CAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGG
                                        ACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTAC
                                             CCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCC
                                                  TCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAG
                                                       TGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGC
                                                            CCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTA
                                                                 TCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGAT
                                                                      CGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGT
                                                                           GGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTT
                                                                                GAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGT
                                                                                     CGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTC
                                                                                          GCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTT
                                                                                               GCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTAT
                                                                                                    CGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAG
                                                                                                         CCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCA
                                                                                                              TGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCA
                                                                                                                   CCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTG
                                                                                                                        GAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTG
                                                                                                                             CAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGC
                                                                                                                                  CCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCG
                                                                                                                                       TCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACC
                                                                                                                                            TGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAA
                                                                                                                                                 GCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGGAC
                                                                                                                                                      TTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGGACTTCTC
```

-145	-140	-135	-130	-125	-120	-115	-110	-105	-100	-95	-90	-85	-80	-75	-70	-65	-60	-55	-50	-45	-40	-35	-30	-25	-20	-15	-10	-5	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105	110	115	120	125	130	135	140	145
▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																															
	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																														
		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																													
			▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																												
				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																											
					▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																										
						▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																									
							▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																								
								▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																							
									▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																						
										▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																					
											▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																				
												▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																			
													▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																		
														▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																	
															▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																
																▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨															
																	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨														
																		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨													
																			▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨												
																				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
																					▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨										
																						▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨									
																							▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨								
																								▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨							
																									▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨						
																										▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨					
																											▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
																												▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨			
																													▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨		
																														▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

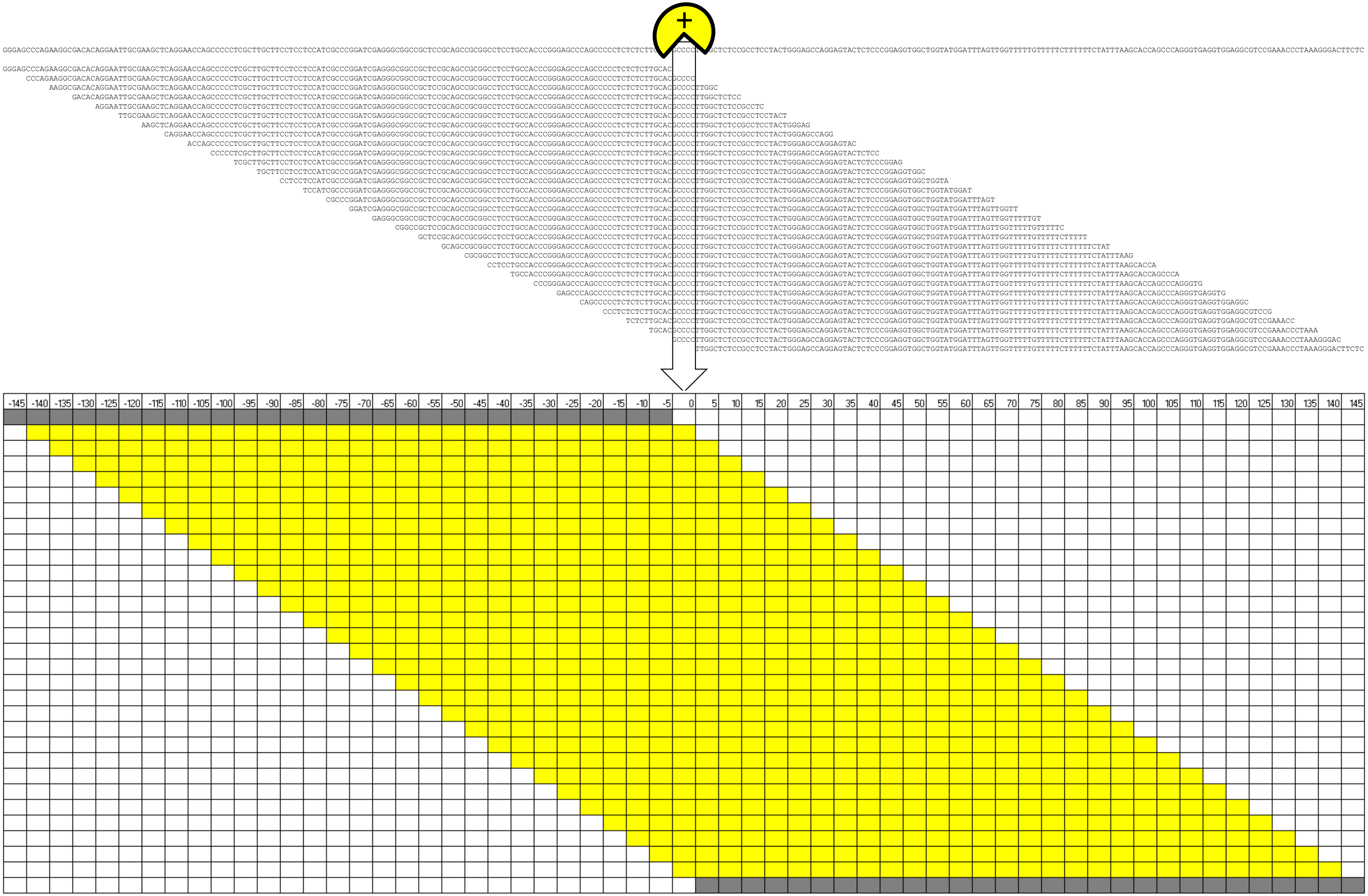
Gray – basal gene expression

Isolating Key Activating or Repressing Sequences



Gray – basal gene expression; Yellow – higher gene expression

Isolating Key Activating or Repressing Sequences



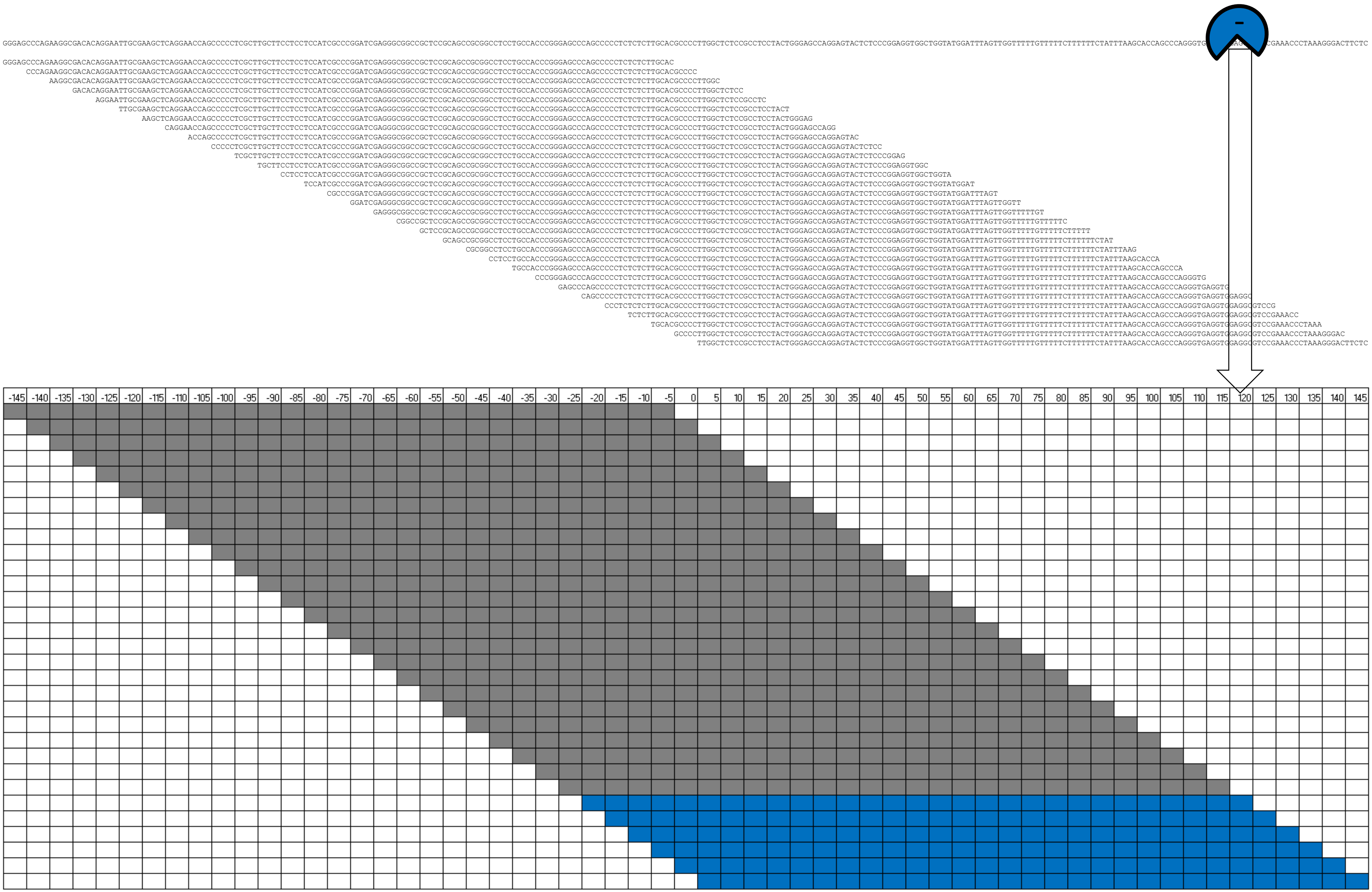
```
GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGGACTTCTC

GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCAC
     CCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCC
          AAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGC
               GACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCC
                    AGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTC
                         TTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACT
                              AAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAG
                                   CAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGG
                                        ACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTAC
                                             CCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCC
                                                  TCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAG
                                                       TGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGC
                                                            CCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTA
                                                                 TCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGAT
                                                                      CGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGT
                                                                           GGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTT
                                                                                GAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGT
                                                                                     CGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTC
                                                                                          GCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTT
                                                                                               GCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTAT
                                                                                                    CGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAG
                                                                                                         CCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCA
                                                                                                              TGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCA
                                                                                                                   CCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTG
                                                                                                                        GAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTG
                                                                                                                             CAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGC
                                                                                                                                  CCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCG
                                                                                                                                       TCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACC
                                                                                                                                            TGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAA
                                                                                                                                                 GCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGGAC
                                                                                                                                                      TTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGGACTTCTC
```

-145	-140	-135	-130	-125	-120	-115	-110	-105	-100	-95	-90	-85	-80	-75	-70	-65	-60	-55	-50	-45	-40	-35	-30	-25	-20	-15	-10	-5	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105	110	115	120	125	130	135	140	145
▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																														
	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																													
		🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																												
			🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																											
				🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																										
					🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																									
						🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																								
							🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																							
								🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																						
									🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																					
										🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																				
											🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																			
												🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																		
													🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																	
														🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨																
															🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨															
																🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨														
																	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨													
																		🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨												
																			🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨											
																				🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨										
																					🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨									
																						🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨								
																							🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨							
																								🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨						
																									🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨					
																										🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨				
																											🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨			
																												🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨		
																													🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	
																														▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

Gray – basal gene expression; Yellow – higher gene expression

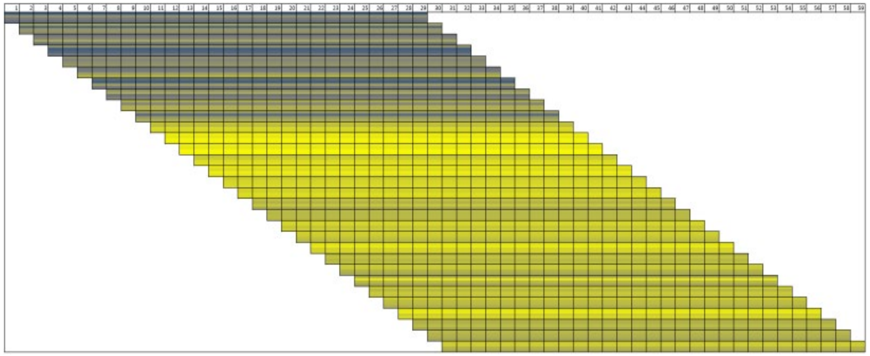
Isolating Key Activating or Repressing Sequences


-
GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTG
CGAAACCCTAAAGGACTTCTC
-145
-140
-135
-130
-125
-120
-115
-110
-105
-100
-95
-90
-85
-80
-75
-70
-65
-60
-55
-50
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145

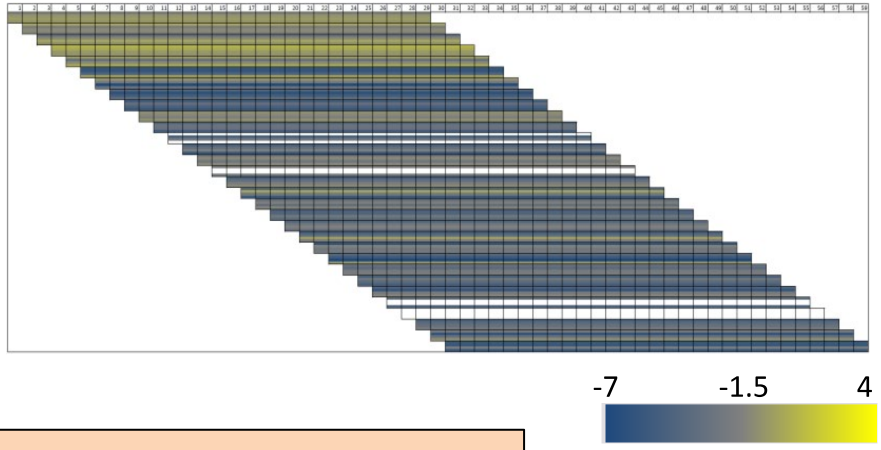
Gray – basal gene expression; Blue – lower gene expression

Example Input Data and Regulatory Activity Inferences

Input Data



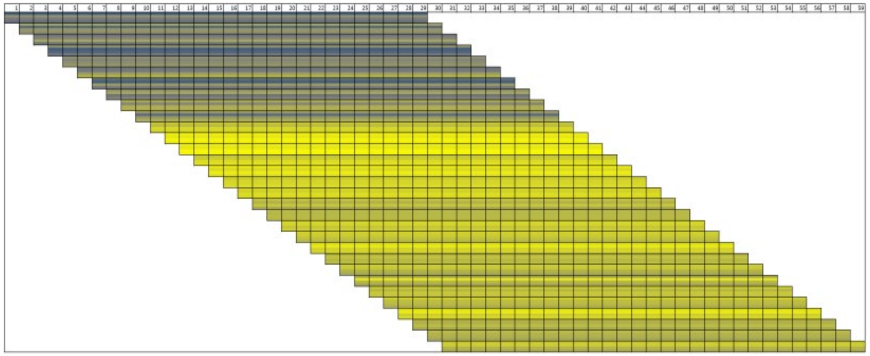
Data from four experiments shown within each tile

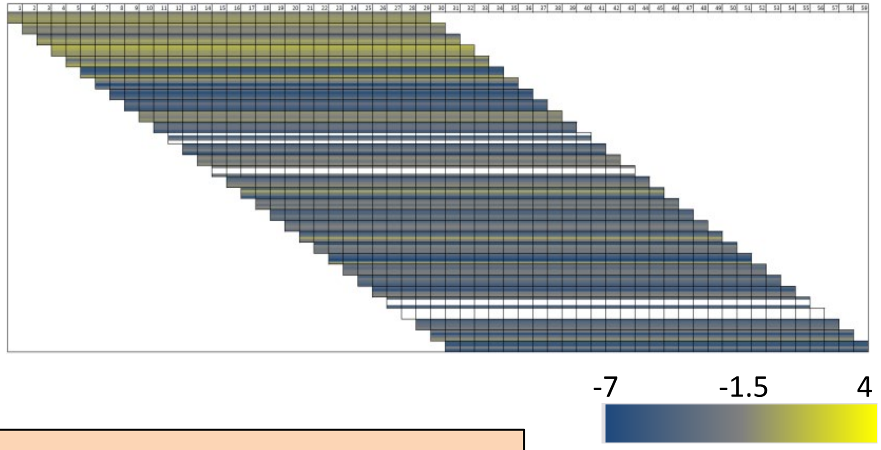


Computational probabilistic modeling - SHARPR

Example Input Data and Regulatory Activity Inferences

Input Data

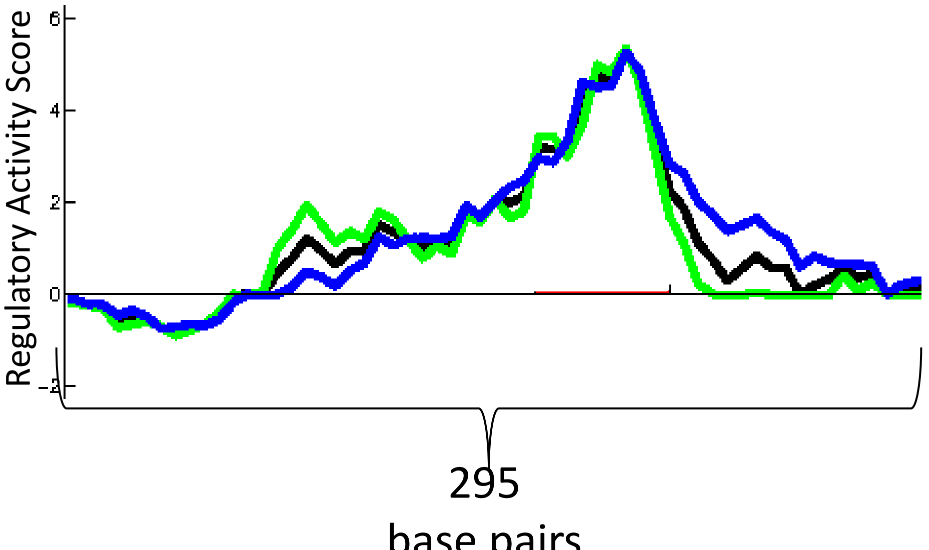


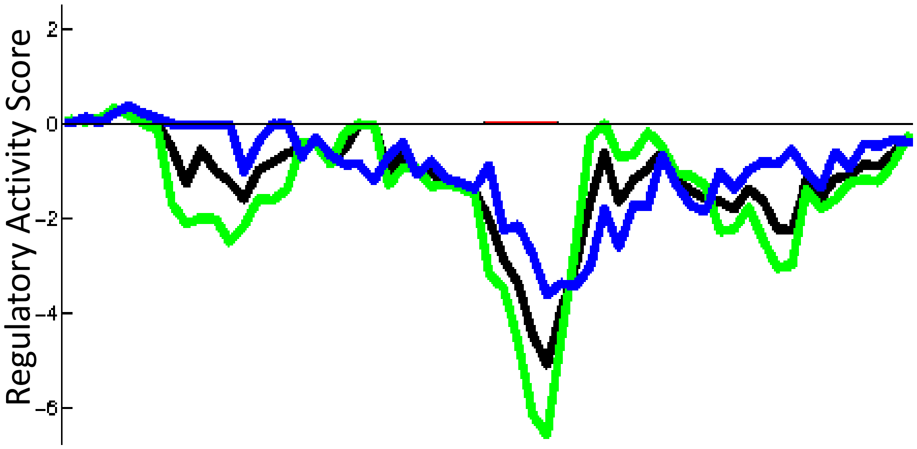


Data from four experiments shown within each tile

Computational probabilistic modeling - SHARPR

Regulatory Activity Inferences

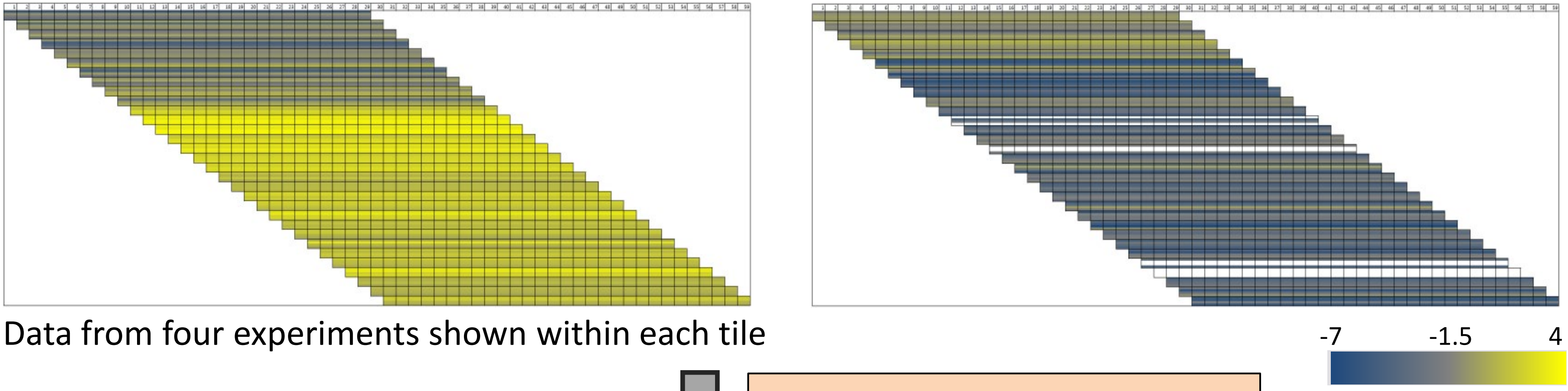




Regulatory Activity Inferences: Green – SV40 promoter only; Blue – minimal promoter only; Black combined

Example Input Data and Regulatory Activity Inferences

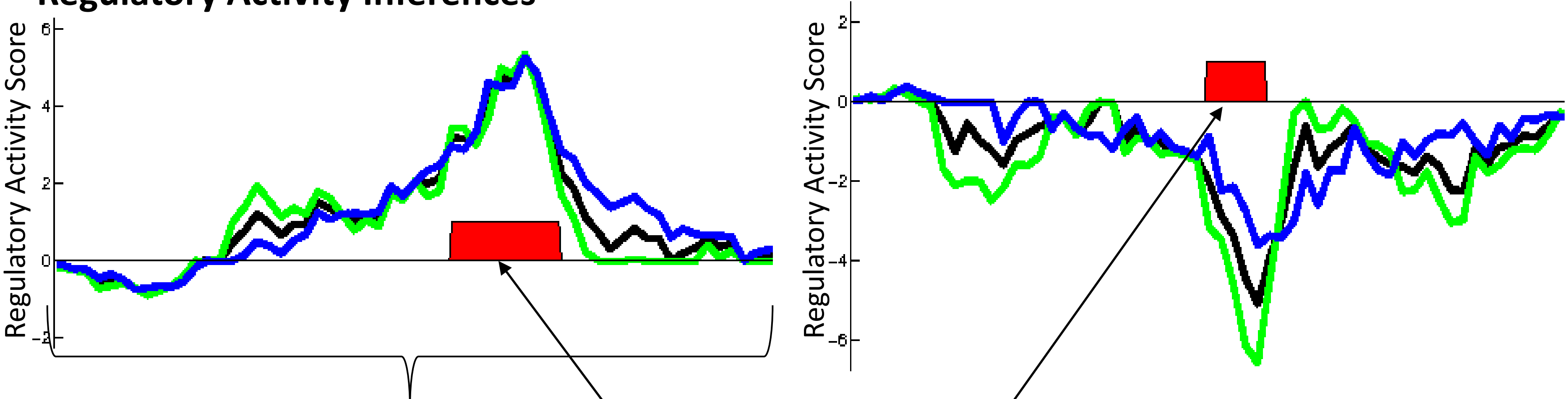
Input Data



Data from four experiments shown within each tile

Computational probabilistic modeling - SHARPR

Regulatory Activity Inferences

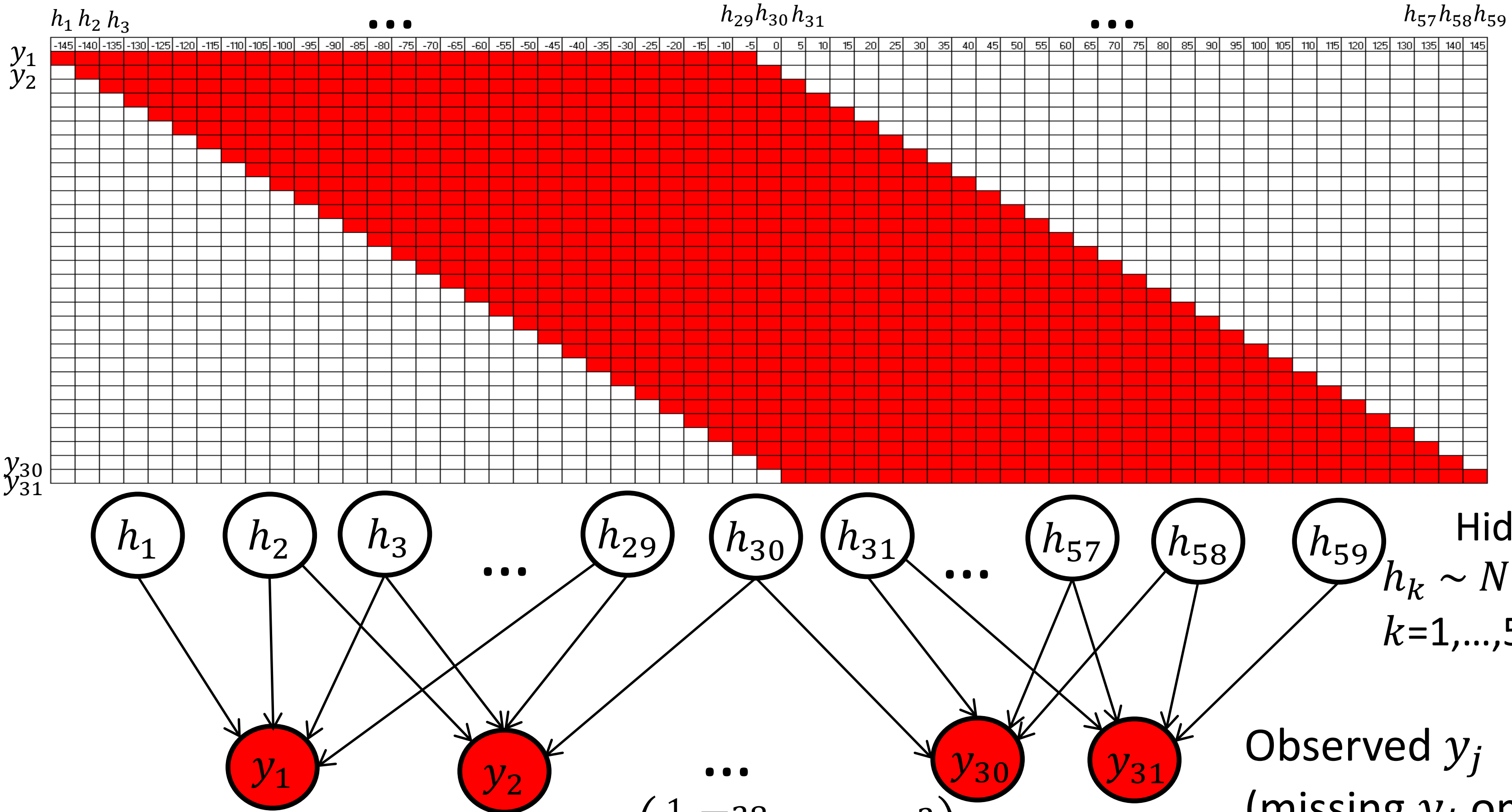


295
base pairs

Orthogonal sequence based prediction
the sequence that is unique to this strand

Regulatory Activity Inferences: Green – SV40 promoter only; Blue – minimal promoter only; Black combined
Sequence based predictions from CENTIPEDE (Pique-Regi, et al, 2011)

SHARPR - Probabilistic Model



Hidden

$h_k \sim N\left(\mu_h, \sigma_h^2\right)$

k=1,…,59

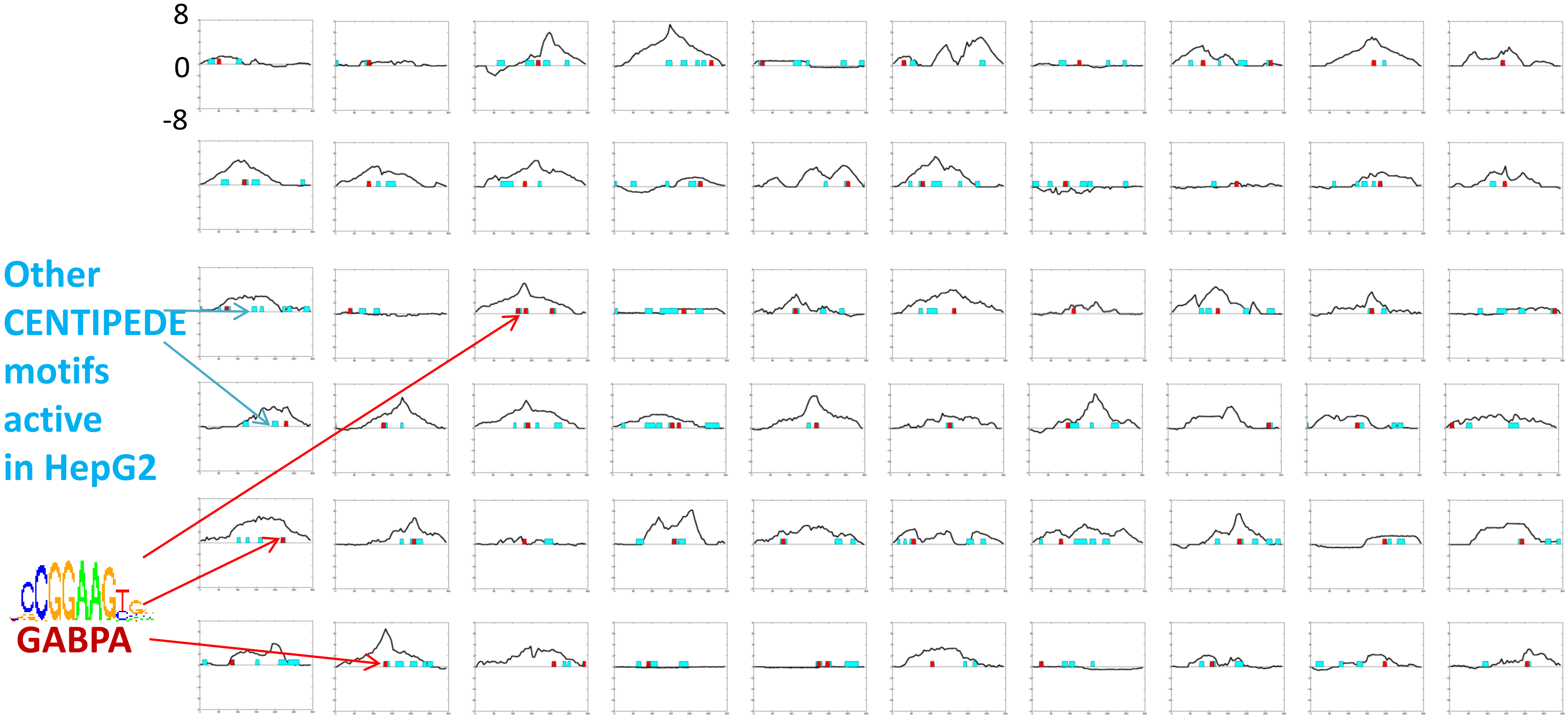
Observed y_j

(missing y_j omitted)

$y_j \sim N\left(\frac{1}{29}\sum_{i=0}^{28} h_{j+i}\,, \sigma_y^2\right)$ j=1,…,31

- μ_h and σ_y^2 - empirical mean and variances of y_j
- Hidden variables can be inferred exactly and efficiently using multivariate gaussians
- Z-score transformation on inferred hidden values, four experiments averaged
- σ_h^2 was set to both of 1 and 50, and more conservative final output was used
- Piecewise linear interpolation for base resolution predictions

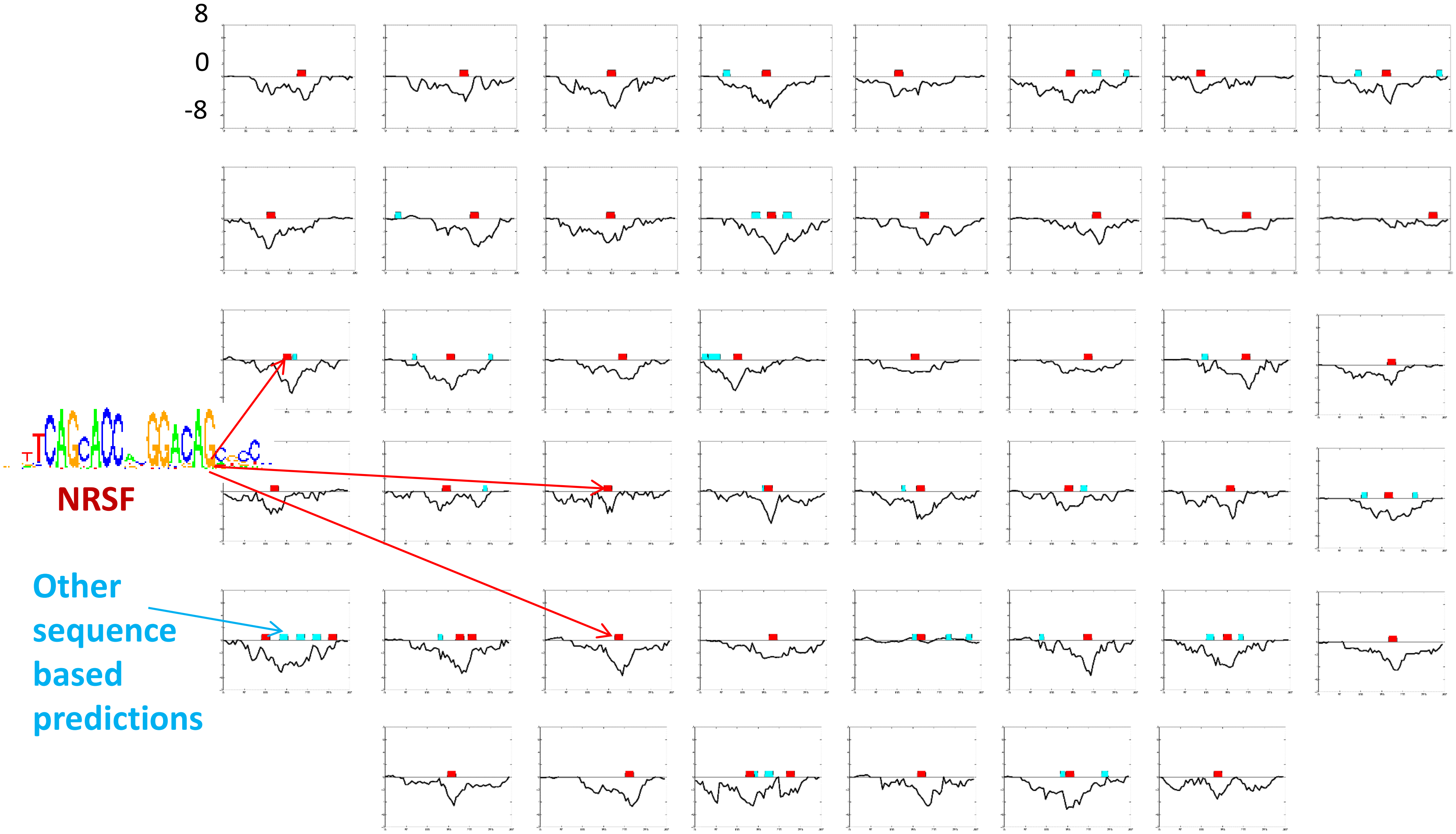
Inferred regulatory signal vs. activator motif



60 sites containing GABPA HepG2 motifs predicted by CENTIPEDE

CENTIPEDE predictions (Pique-Regi, et al, 2011)

Inferred regulatory signal vs. repressor motif



All sites tested containing NRSF motif based predictions in HepG2

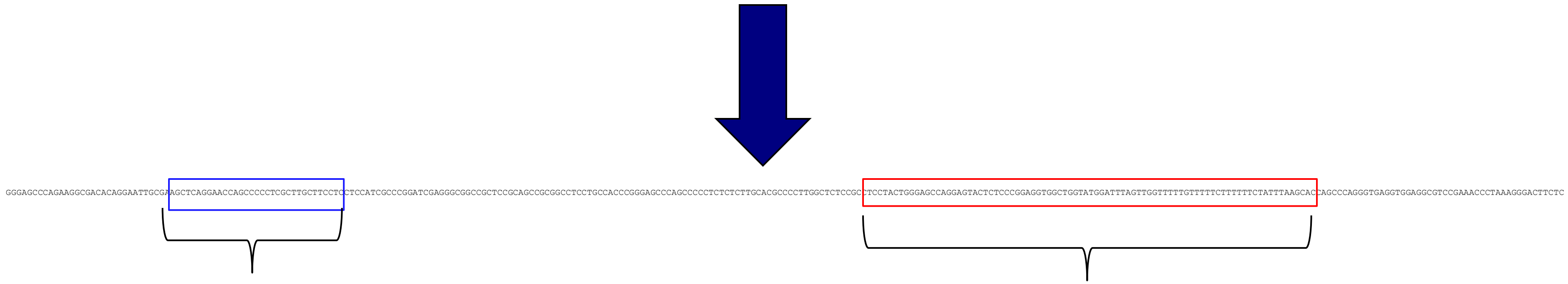
Ernst et al, *Nature Biotech* 2016

Sequence based predictions from CENTIPEDE (Pique-Regi, et al, 2011)

Mini-Hackathon CBS

- In collaboration with Computational Biologist's Society (CBS) at UCLA
- Task is to predict from DNA sequence highly activating and repressive nucleotides annotated by Sharpr-MPRA

GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGACTTCTC


GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGACTTCTC

Top repressive prediction

Top activating prediction

Why predict from DNA sequence?

- Do not have experimental data across every loci
- Not feasible to experimentally map in every individual
- If we have an effective predictive model of regulatory activity from DNA sequence, can use it to make predictions of regulatory impact of sequence variants in specific cell types

e.g. what is predicted impact of mutating an A to T?

```
TGTCGTCTTGGTTCAGCCAAGGTCACAGAGGGAGTGATAGCTTCCGCGCAGCCCTGGCTACGGACTCTGGGCATCTTTCCACTGCCCCGCTTGCGCCACC
TGTCGTCTTGGTTCAGCCAAGGTCACAGAGGGAGTGATAGCTTCCGCGCTGCCCTGGCTACGGACTCTGGGCATCTTTCCACTGCCCCGCTTGCGCCACC
```

Provided File 1

- train_MPRA.txt - contains 8000 sequences.
- One sequence per line
- For each sequence, a sequence ID, 295-bp sequence, followed by 295 columns of activity values corresponding to nucleotides in order

```
train1   AGCTCACGGGGACTAGGGCAGGGAGGCTGCGGGGATGGAAAGATC…   0.019    0.019    0.019    0.018…
train2   ACTCTCATCCCACAGAATGAGCTTTACAGTAACTTGGATCTCTAC…   -0.113   -0.113   -0.113   -0.113…
…
```

Note: do not have to use all the training data available

Provided File 2

- trainsolutions.txt
- Sample output data file that would receive a perfect score

```
R      train1310    168
R      train1310    169
R      train1310    167

…

A      train2219    193
A      train2219    194
A      train2219    195
```

- Note because of ties not a unique such solution

Provided File 3

- test_MPRA.txt

```
test1       CTCCGGACAGGTGGGTTTGACAACATCTATTTTGGTCATGCCTGGGCAGTTCTGGCTTATCCATCTACACTAGTCTCAAATGATCCTGGGATTTGCTCTTGGGTA
test2       CTGTCCCAGCCTACAGTCAGCTCAGTGCACAGCTGCCTCTTCCTGTGTGTACCTGCAGGCCCCACCTGGGCTGGAATGCTGCCTTCTTCACCACACAGAGGCGGC
test3       TCAAATCTCTTGAACTTCTCTTCCAACACCAGCTGGAGAAAGAGGCTCTCATTTTGAAGGGCCCCTGTGATTAGATTGCAACCATTTGAATAGTCTTTTTTGATT
```

- First column is ID
- Second column is 295 bp sequence
- In total 7720 sequences

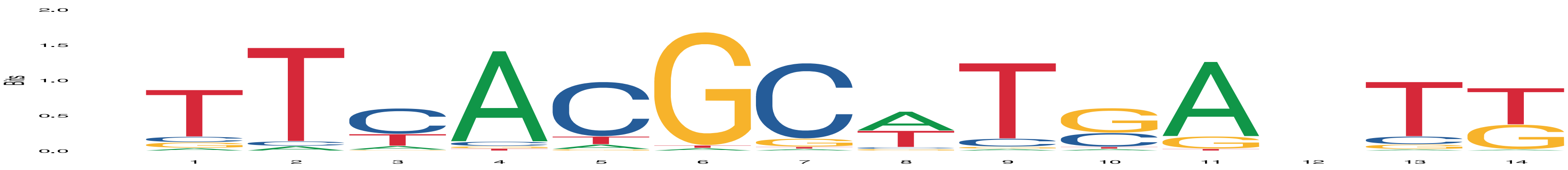
Provided File 4

- JASPAR2024_CORE_vertebrates_non-redundant_pfms_jaspar.txt

contains a library of positional weight matrices (PWMs) from JASPAR database

- Optional to use

```
>MA0004.1       Arnt
A  [      4      19       0       0       0       0 ]
C  [     16       0      20       0       0       0 ]
G  [      0       1       0      20       0      20 ]
T  [      0       0       0       0      20       0 ]
>MA0069.1       PAX6
A  [      2       2       4      39       3       1       1      21       1       2      36      11       1       1 ]
C  [      4       2      26       2      34       0      37       2       4      14       0      11       5       0 ]
G  [      4       0       1       1       1      41       4       2       1      25       6      13       3      17 ]
T  [     33      39      12       1       5       1       1      18      37       2       1       8      34      25 ]
…
```



Output

- Text file which contains your top 100,000 activating and top 50,000 repressive predictions for test sequences. One per line.

```
R        test1310        168
R        test1310        169
A        test2084        45
A        test5261        221
…
```

- First column ‘A’ for activating; ‘R’ for repressive
- Second column IDs from testing sequence
- Third column is nucleotide position in sequence where positions are indexed starting from 1
- Columns can be tab, space, or comma delimited

Competition Scoring

- The final score will be the sum of the number of top 100,000 activating base predictions that are in the top 100,000 activating bases and the number of top 50,000 repressive predictions in the top 50,000 repressive bases

Competition Rules

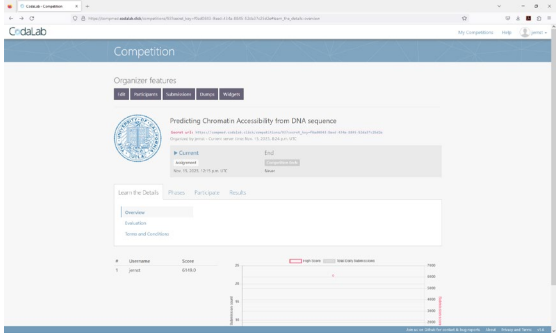
- Winners will be determined based on status of leaderboard on Friday, January 10th at 5pm
- May only use provided data files. No use of external data is allowed.
- Use of standard machine learning libraries is allowed but using existing software specifically designed for predicting from DNA sequence is not allowed
- Winning team(s) will be asked to give presentation and may be asked to submit code

Online system

- Will use CodaLab which provides leaderboard scoring
- CodaLab link: https://compmed.codalab.click/competitions/192?secret_key=409a5b0d-cfd6-4076-84aa-da6bffd958ed
- File link:

https://ucla.box.com/s/3dpi45n9fslao5uygqyjngkhval15soi

- Data files and competition system will be released with kickoff event on Monday
- Must be on UCLA network (including VPN) to access
- Will need to create a CodaLab account to submit
- Access/login issues https://uclahs.fyi/codalab-support goes to Clifford Kravit
- Follow slack for updates/clarification



Predicting Chromatin Accessibility from DNA sequence

Secret url: https://compmed.codalab.click/competitions/93?secret_key=f0ad0843-9aed-434a-8845-52da37c25d2e

Organized by jernst - Current server time: Nov. 15, 2023, 8:24 p.m. UTC

▶ Current	End
Assignment	Competition Ends
Nov. 15, 2023, 12:15 p.m. UTC	Never

#	Username	Score
1	jernst	6149.0

Submission instructions

a. Once you have your answers written in your text editor, save the file as *predictions.csv* (Required). Alternatively, convert your existing .txt file containing the answers to .csv by running the following command:

```
$ mv ./FILE.txt ./predictions.csv
```

b. Then run the following command to zip the predictions.csv into a .zip file (any name allowed)

```
$ zip -r predictions.zip predictions.csv
```

c. To submit your .zip file, go to the competition page and navigate to the **Participate** tab, then click on **Submit / View Results**. When you click the Submit button a file explorer will open up for you to select a file to upload. Then you are done. You may have to refresh the page manually after a minute for the grader's output to be calculated and tabulated on the **Results** tab/section below the submit button.

Some possible strategies

- Simple baseline approaches
- Explicit features to standard supervised classification or regression methods
- Deep learning methods directly from DNA sequences

Some possible strategies

- Simple baseline approaches
- Explicit features to standard supervised classification or regression methods
- Deep learning methods directly from DNA sequences

Simple baseline approach – GC content

- Count the number of G’s or C’s in a sequence and rank sequences in increasing or decreasing counts. Could either be the same or different choice between the two for activating and repressive.
- All bases within a sequence would be equivalently ranked.

Simple baseline approach – Center position

- Predict the most center bases in every sequence

GGGAGCCCAGAAGGCGACACAGGAATTGCGAAGCTCAGGAACCAGCCCCCTCGCTTGCTTCCTCCTCCATCGCCCGGATCGAGGGCGGCCGCTCCGCAGCCGCGGCCTCCTGCCACCCGGGAGCCCAGCCCCCTCTCTCTTGCACGCCCCTTGGCTCTCCGCCTCCTACTGGGAGCCAGGAGTACTCTCCCGGAGGTGGCTGGTATGGATTTAGTTGGTTTTTGTTTTTCTTTTTTCTATTTAAGCACCAGCCCAGGGTGAGGTGGAGGCGTCCGAAACCCTAAAGGGACTTCTC

Some possible strategies

- Simple baseline approaches
- Explicit features to standard supervised classification or regression methods
- Deep learning methods directly from DNA sequences

K-mer features

ACACCATTAGACCA

Example with *k*=2

2-mers	count
AA	0
AC	3
AG	1
AT	1
CA	3
CC	2
CG	0
CT	0
GA	1
GC	0
GG	0
GT	0
TA	1
TC	0
TG	0
TT	1

K-mer features

ACACCATTAGACCA

Example with *k*=2

2-mers	count
AA	0
AC	3
AG	1
AT	1
CA	3
CC	2
CG	0
CT	0
GA	1
GC	0
GG	0
GT	0
TA	1
TC	0
TG	0
TT	1

Positional Weight Matrix

- PWM scanning

	1	2	3	4	5	6	7
A	3/5	0	0	2/5	$\boxed{1}$	1/5	$\boxed{1}$
C	$\boxed{2/5}$	$\boxed{1}$	0	0	0	1/5	0
G	0	0	$\boxed{4/5}$	0	0	1/5	0
T	0	0	1/5	$\boxed{3/5}$	0	$\boxed{2/5}$	0

Scoring agreement of a sequence with the PWM

CCGTATA

$$\frac{2}{5} \times 1 \times \frac{4}{5} \times \frac{3}{5} \times \ 1 \ \times \frac{2}{5} \times 1 = \frac{48}{625}$$

PWM features

- May want to represent as log-ratio relative to background model e.g. probability ¼ of each nucleotide
- May want to truncate score (e.g. set to 0, log-ratio values less than 0)
- Different ways to aggregate over region (e.g. maximum or average)
- Can also consider scanning reverse complement (swap A's and T's; C's and G's; then scan in the reverse direction)

Positional information

- Could have features corresponding to where base being predicted is along the sequence and its distance to the center or ends
- Could have sequence features be the same for the entire sequence or specific to each position

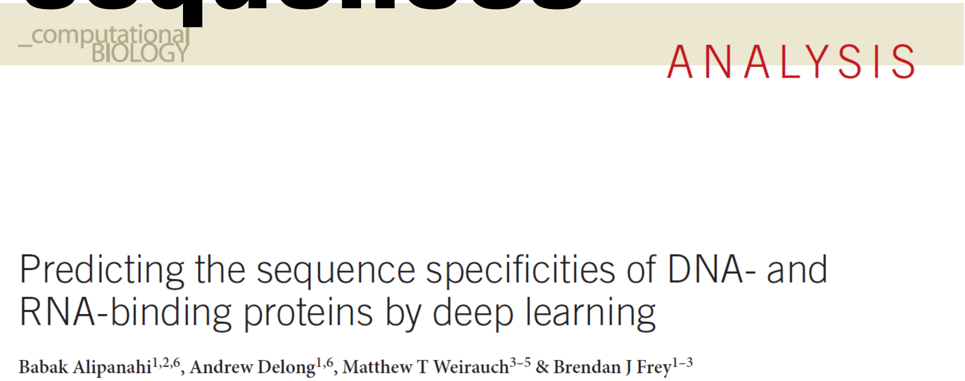
Prediction task

- Could either model task as regression or classification after discretizing
- With classification could either be model as separate classification problems for activation or repression or three -way classification
- With regression could either be modeled as single regression problem or separate regression problems for activation and repression, where low and high values are truncated respectively
- Many standard methods for regression or classification (e.g. linear/logistic regression, tree based, SVM based etc)

Some possible strategies

- Simple baseline approaches
- Explicit features to standard supervised classification or regression methods
- Deep learning methods directly from DNA sequences

Deep learning modeling of sequences



_computational BIOLOGY

ANALYSIS

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi[1,2,6], Andrew Delong[1,6], Matthew T Weirauch[3–5] & Brendan J Frey[1–3]

Nature Biotechnology, 2015

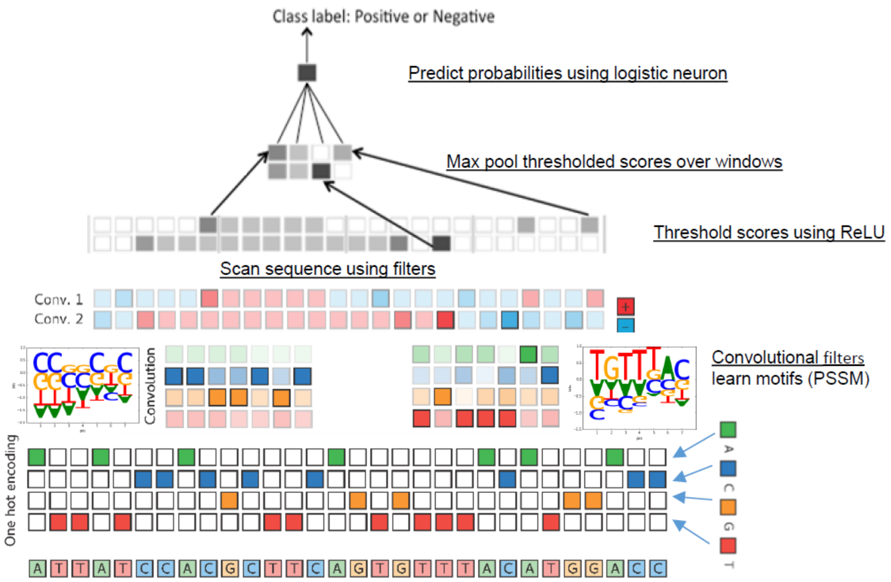

Class label: Positive or Negative
Predict probabilities using logistic neuron
Max pool thresholded scores over windows
Threshold scores using ReLU
Scan sequence using filters
Conv. 1
Conv. 2
Convolution
Convolutional filters learn motifs (PSSM)
One hot encoding
A
C
G
T
A T T A T C C A C G C T T C A G T G T T T A C A T G G A C C

Image from Anshul Kundaje

- Often give state of the art results for prediction tasks from DNA sequence
- Connection between convolutional filters and PWMs
- Could be applied de novo or could try to integrate PWM library

Final remarks

- Be creative and have fun!
- Contact: Slack or Jason Ernst (jason.ernst@ucla.edu)
- Prof. Eskin and I will be teaching “Algorithms in Computational Genomics” C122 in Winter 2025

MiniHack Timeline of Events

- Kickoff event November 25th – Franz Hall Room 1260 from 6-7PM
- Zoom office hours Dec 5th – 5pm
- Submission deadline Jan 10th – 5pm
- Presentations of winning teams TBD (likely week of Jan 13th)

Questions?

- CodaLab link: https://compmed.codalab.click/competitions/192?secret_key=409a5b0d-cfd6-4076-84aa-da6bffd958ed
- Files link:

https://ucla.box.com/s/3dpi45n9fslao5uygqyjngkhval15soi