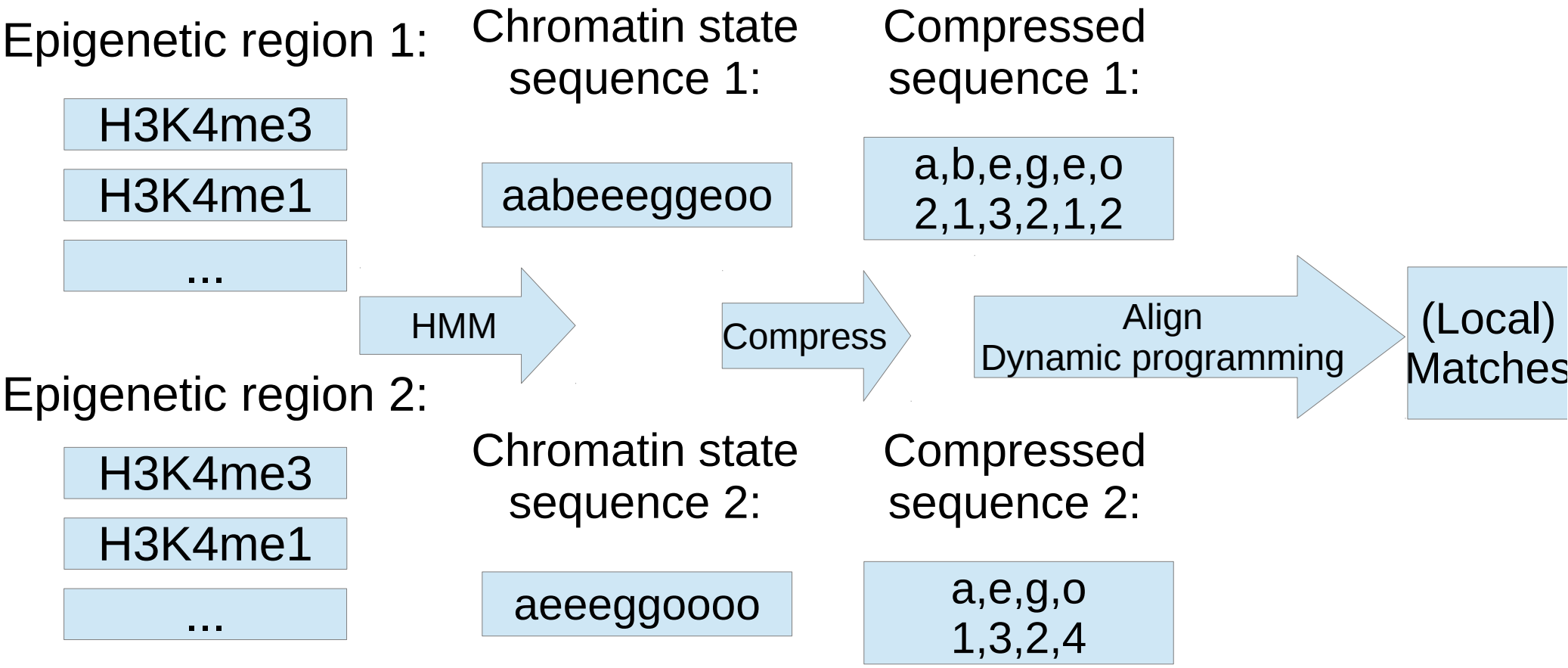
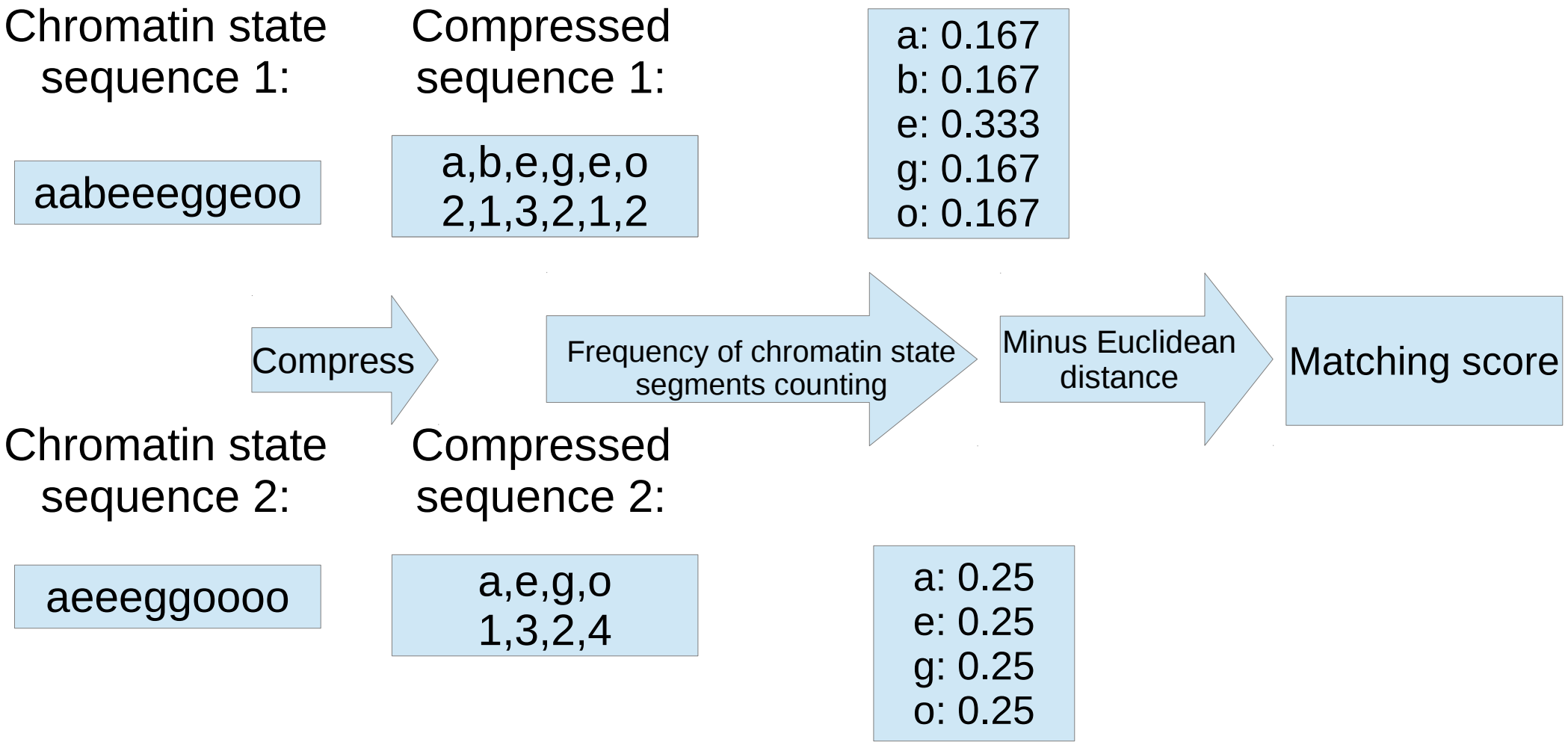
Figure 1: EpiBLAST Flow Chat

EpiBLAST flow chat



Baseline alignment algorithm for comparison

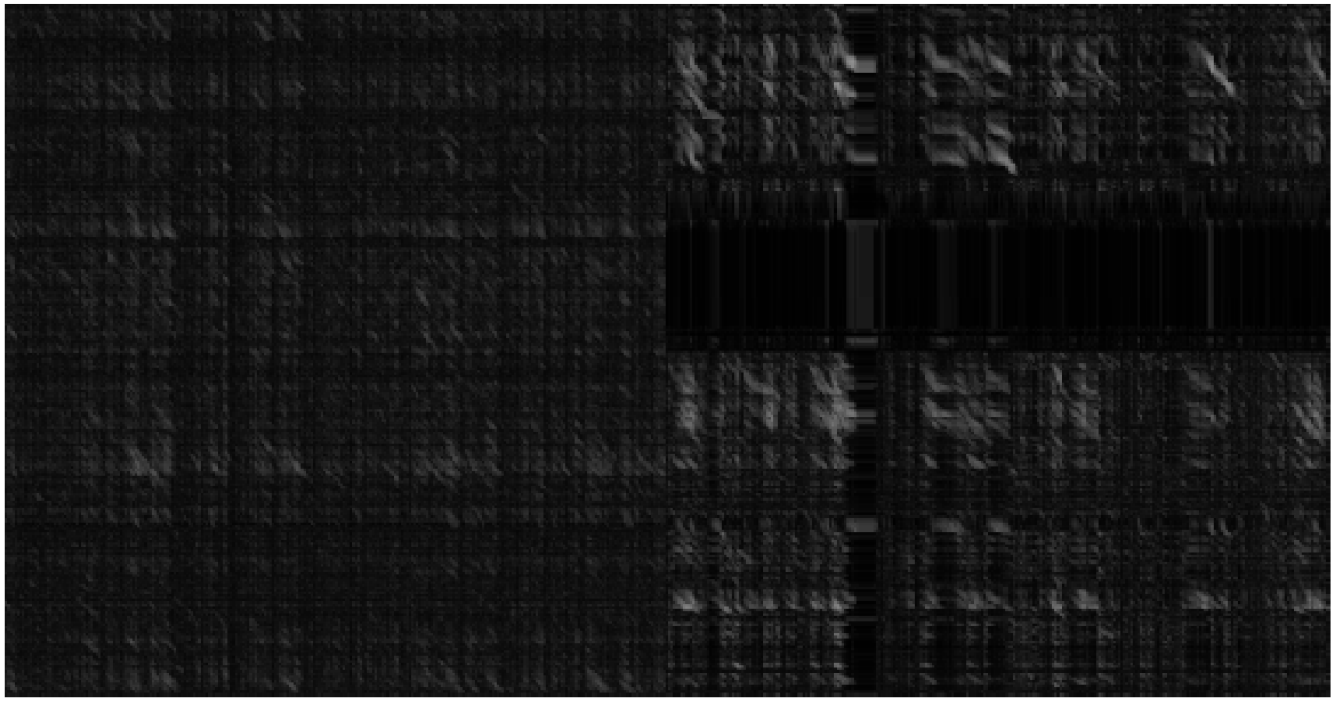


Figure? 2: Generation of fake chromosome(s)

Figure 3: Illustration of EpiBLAST results

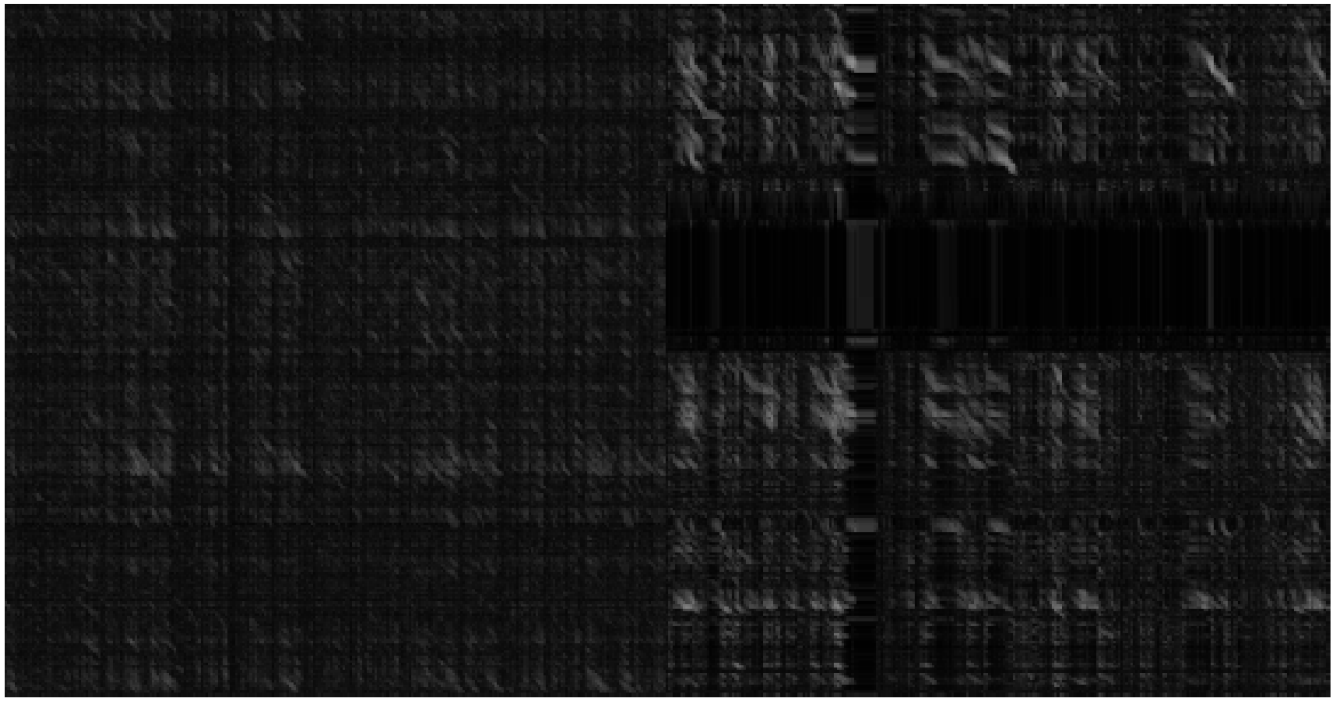
- Correlation of maximal alignment score across the samples of roadmap database.
- Image show of alignment score between chromosomes.

Visualization of Smith-Waterman Matrix



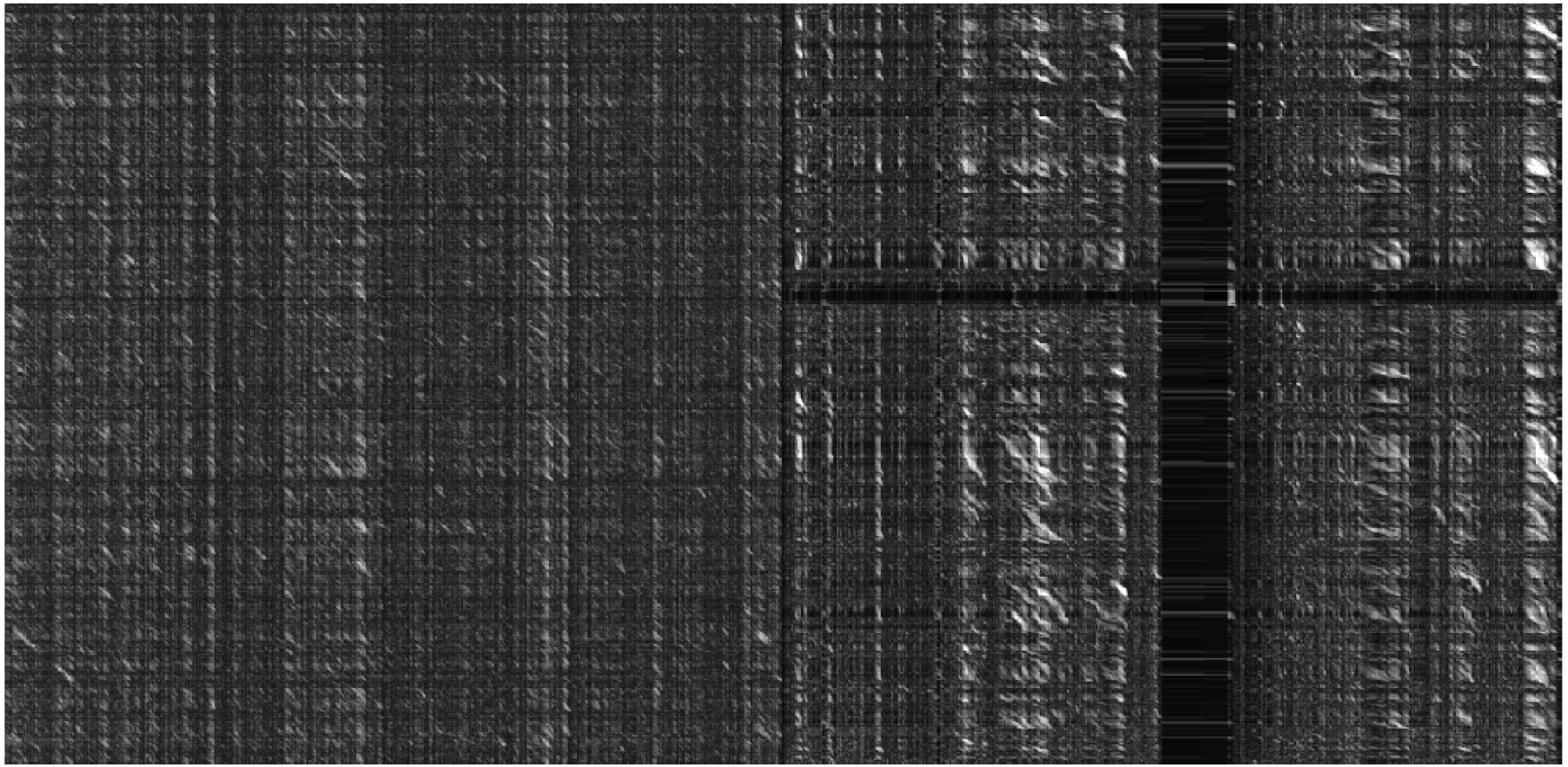
- Sample E003, chr9-chr10
- Alignment algorithm without attention score
- Fake chromosome (left), true chromosome (right)

Visualization of Smith-Waterman Matrix



- In each diagonal area of this image, pixel(i,j) is the maximal alignment score of region (i-1)*500k+1--->i*500k on sequence1 against region (j-1)*500k+1--->j*500k on sequence2.
- Here I aim to better illustrate our algorithm, it can be a part of our figures

Visualization of Smith-Waterman Matrix



- Sample E003, chr1-chr2
- Attention score: unpredictability
- Fake chromosome (left), true chromosome (right)

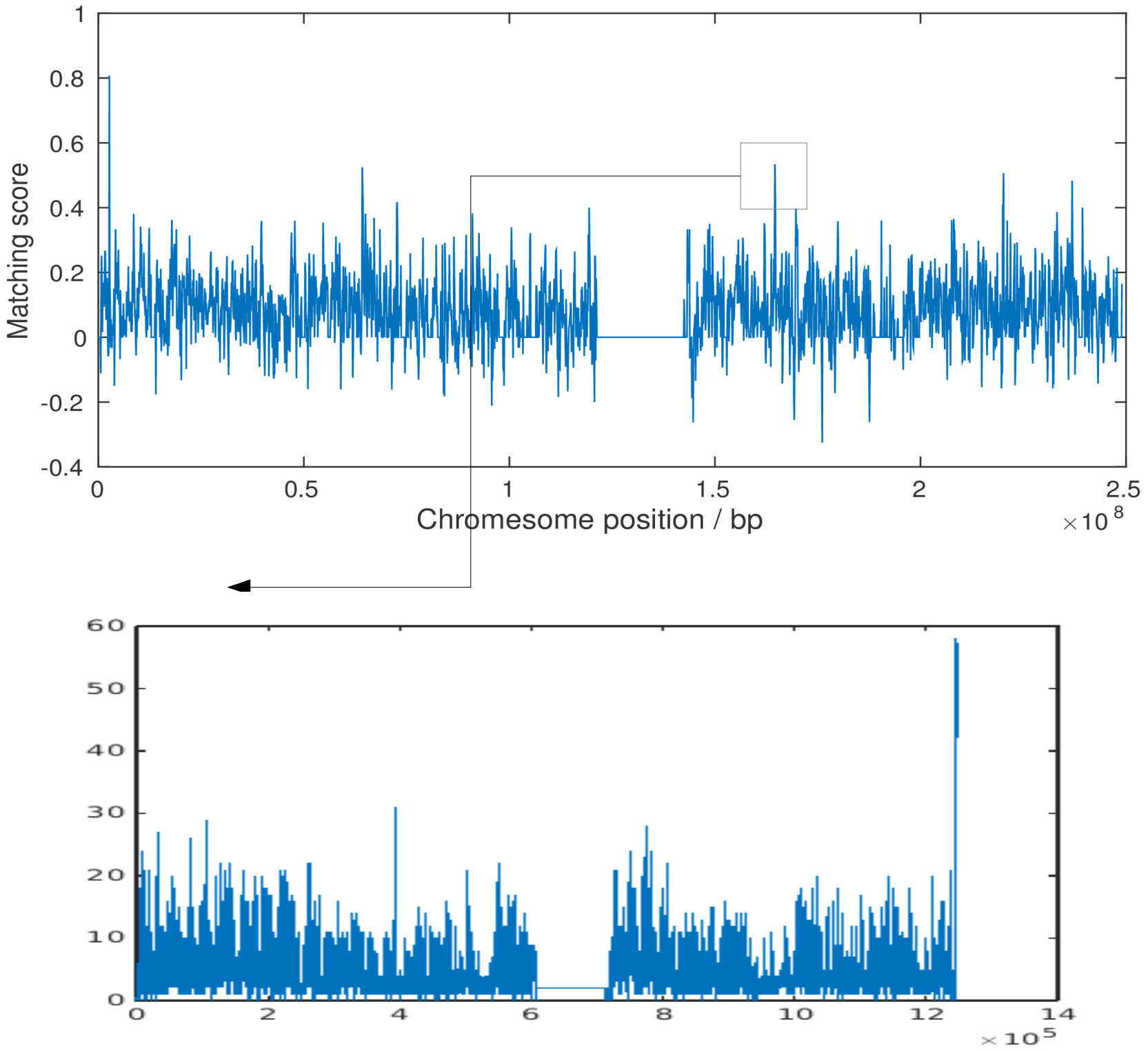
Figure 4: Horizontal Alignment

- Alignment score of one genome segment is defined as the following:

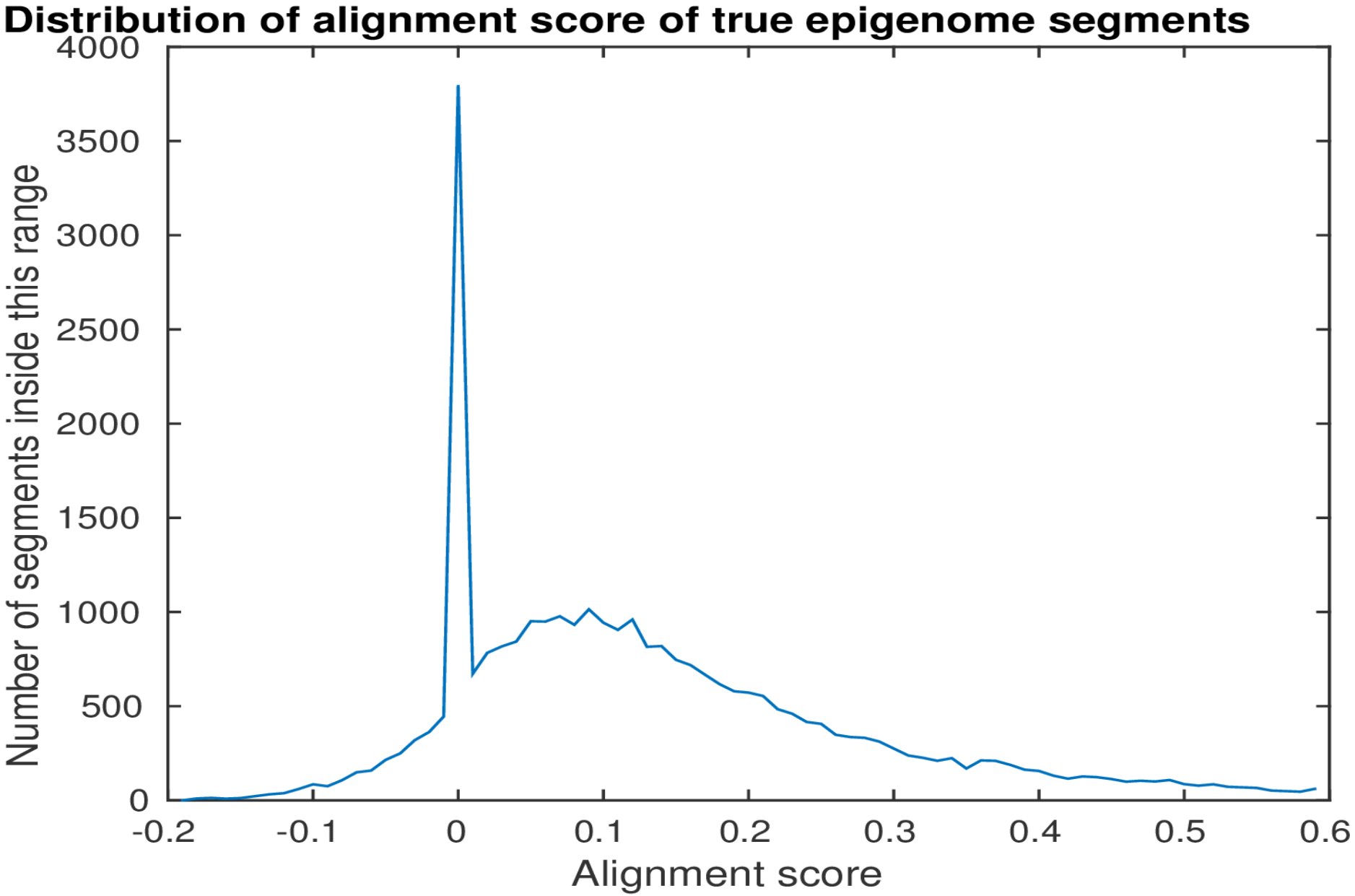
  (max(true)-max(fake)) / max(fake)

  max(true): best local alignment score between the query segment and the true epigenome.

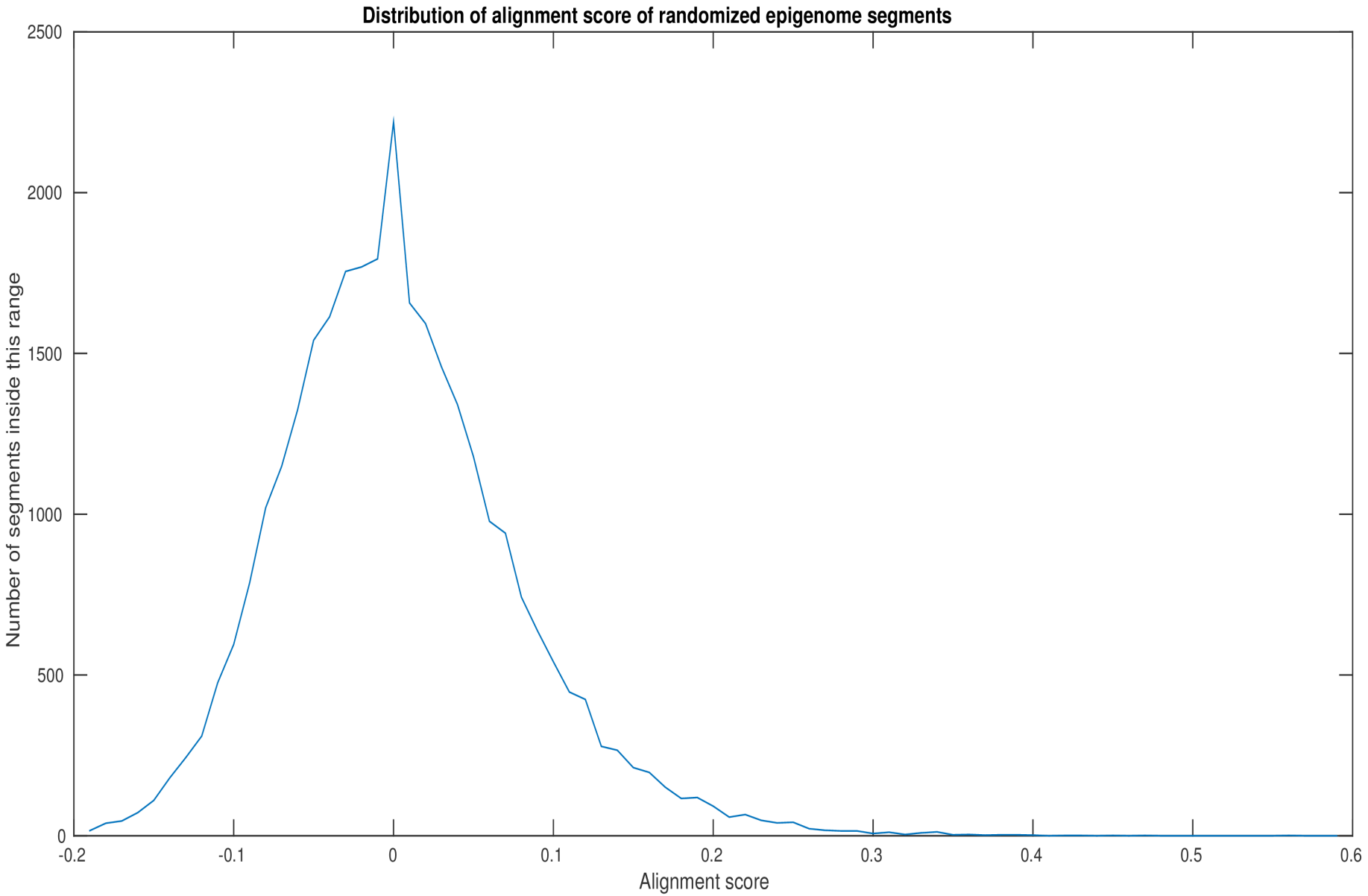
  max(fake): best local alignment score between the query segment and the randomized epigenome.
- Alignment score of peaks along true and fake chromosome.


Matching score
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
0
0.5
1
1.5
2
2.5
Chromesome position / bp
×10 8
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
×10 5

Alignment score distribution of true epigenome



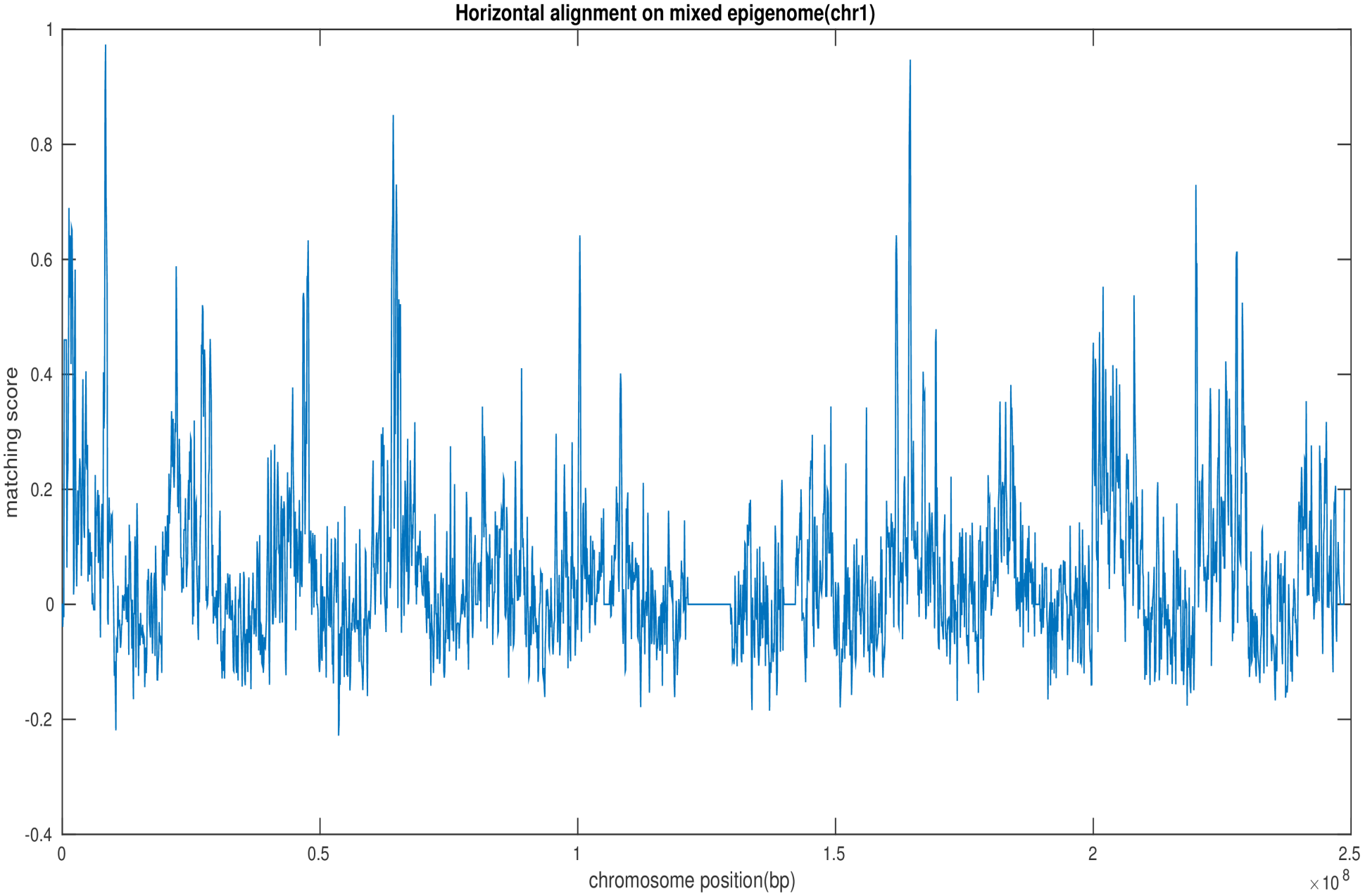
Alignment score distribution of randomized epigenome



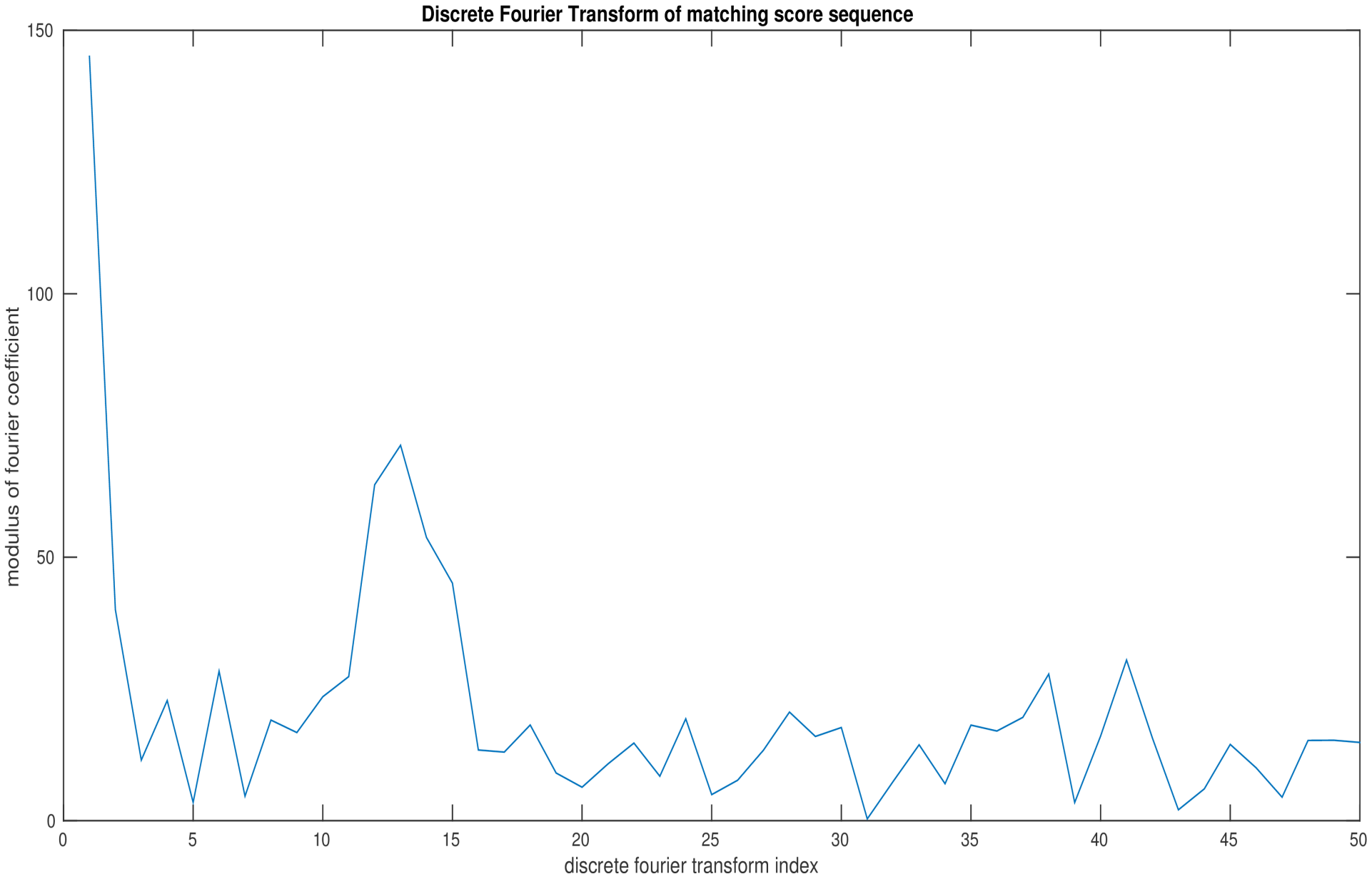
Mixed epigenome generation

- each chromosome is cut into non-overlapping contiguous w-base-pair segments.
- The segments of a chromosome are assigned alternately by the chromatin states on the corresponding area of either the true epigenome or the randomized epigenome.
- Horizontal alignment and following analysis are applied onto this mixed epigenome.

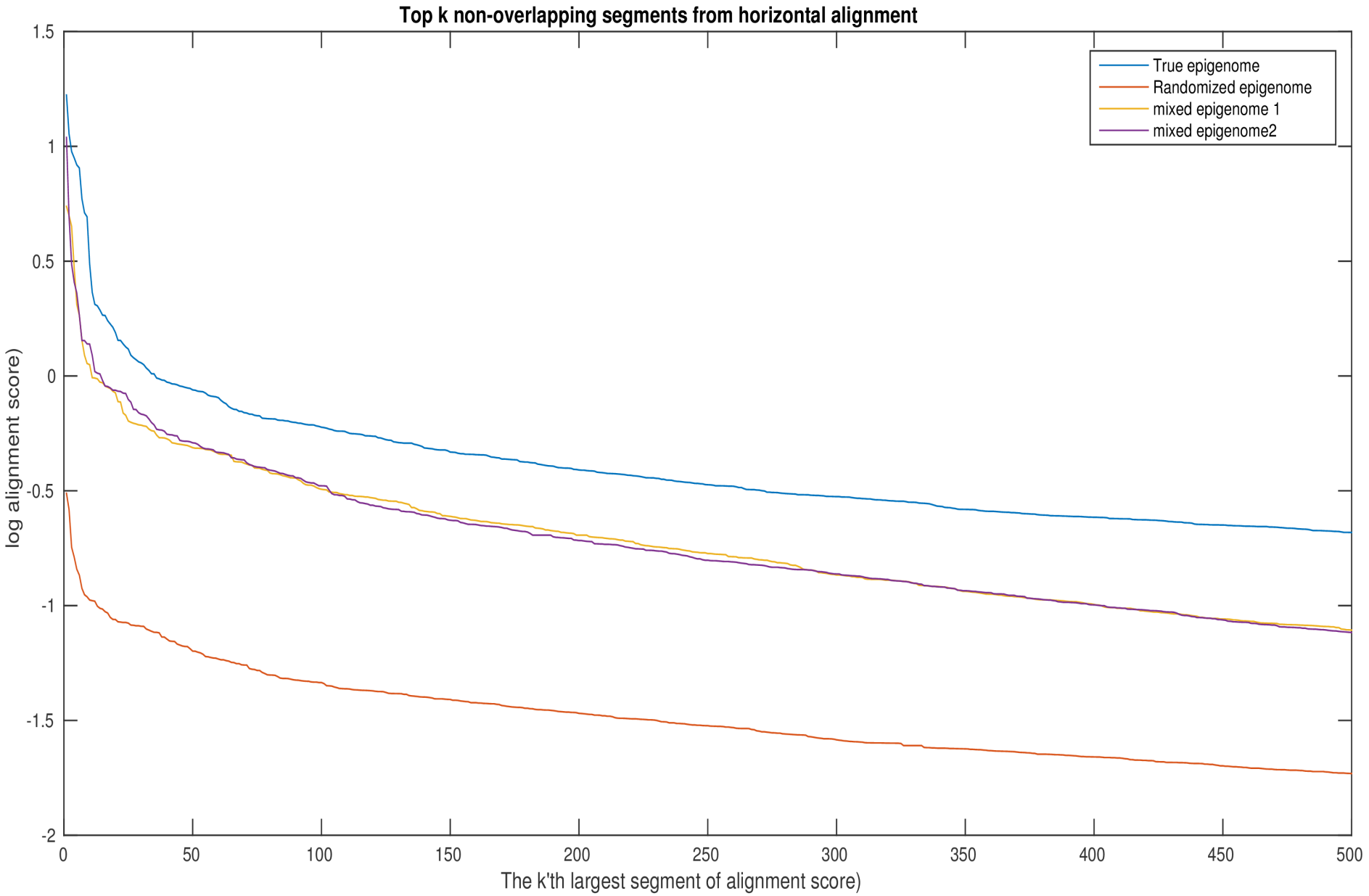
Alignment score comparison



Alignment score comparison



Alignment score comparison



Alignment score comparison

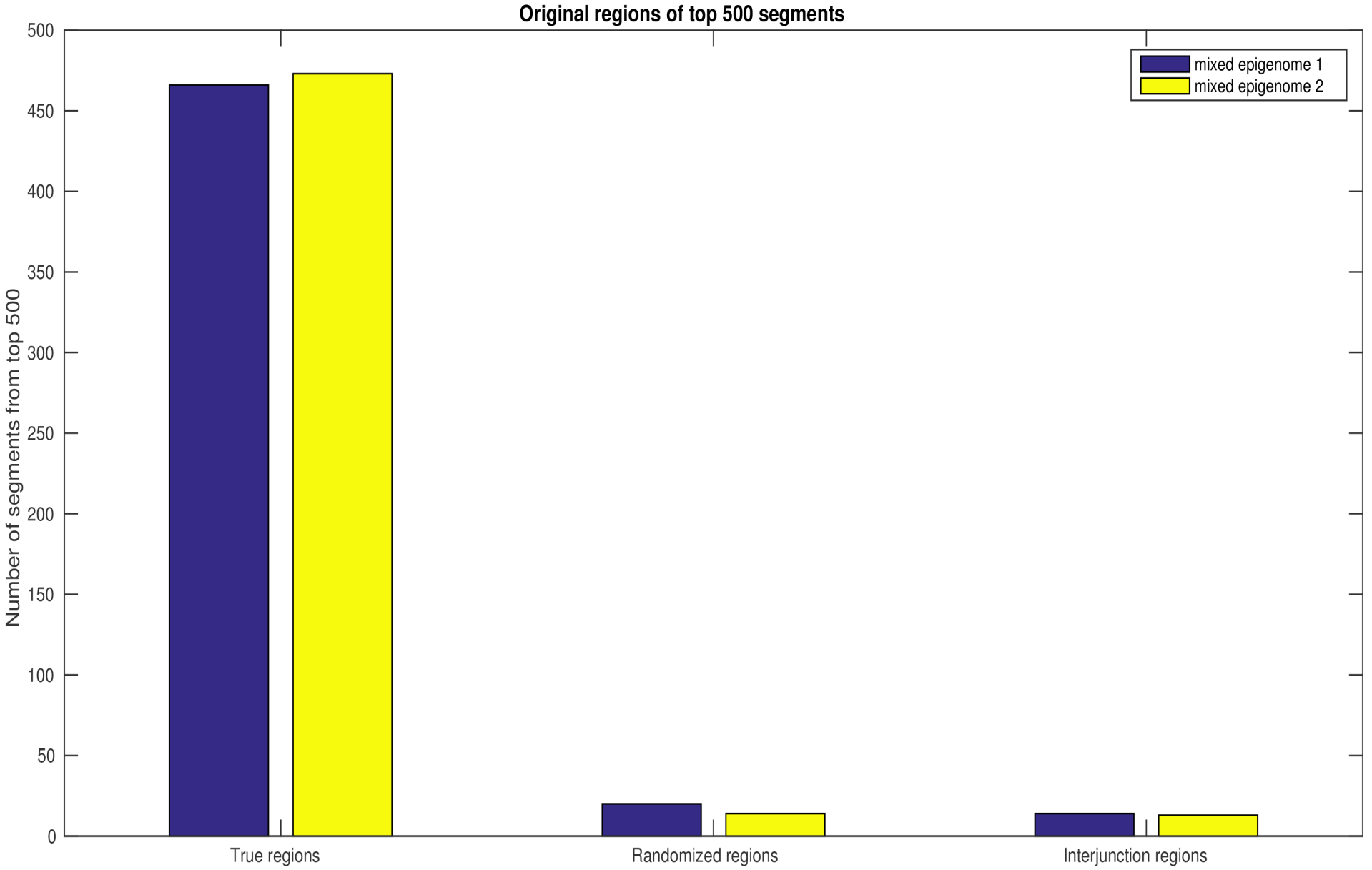


Figure 5: compare with baseline method

- E003, chr1, baseline matching score:

  best horizontal alignment score minus best alignment score to fake chromosomes

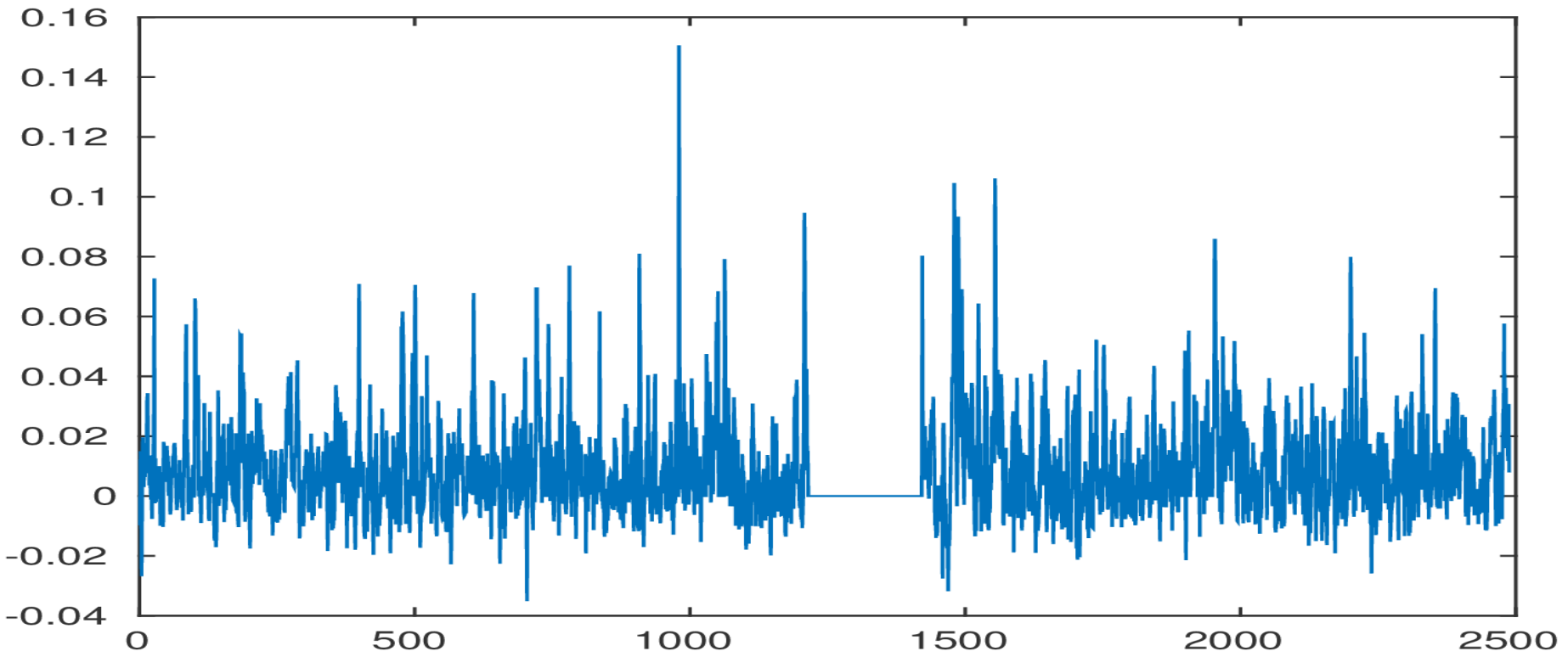


Figure 5: compare with baseline method

- E003:

  EpiBLAST(top 200 regions)

  average number of chromatin state segments: 13.565

  total gene annotation items: 3996

  Baseline(top 200 regions)

  average number of chromatin state segments: 137.835

  total gene annotation items: 1297

Figure 5: compare with naïve alignment method (tissue specific, ESC)

- Top 500 matched regions:

  EpiBLAST: 30 regions overlap with ESC specific gene

  Naïve alignment: 2 regions overlap with ESC specific gene

Figure 5: compare with naïve alignment method (tissue specific, ESC)

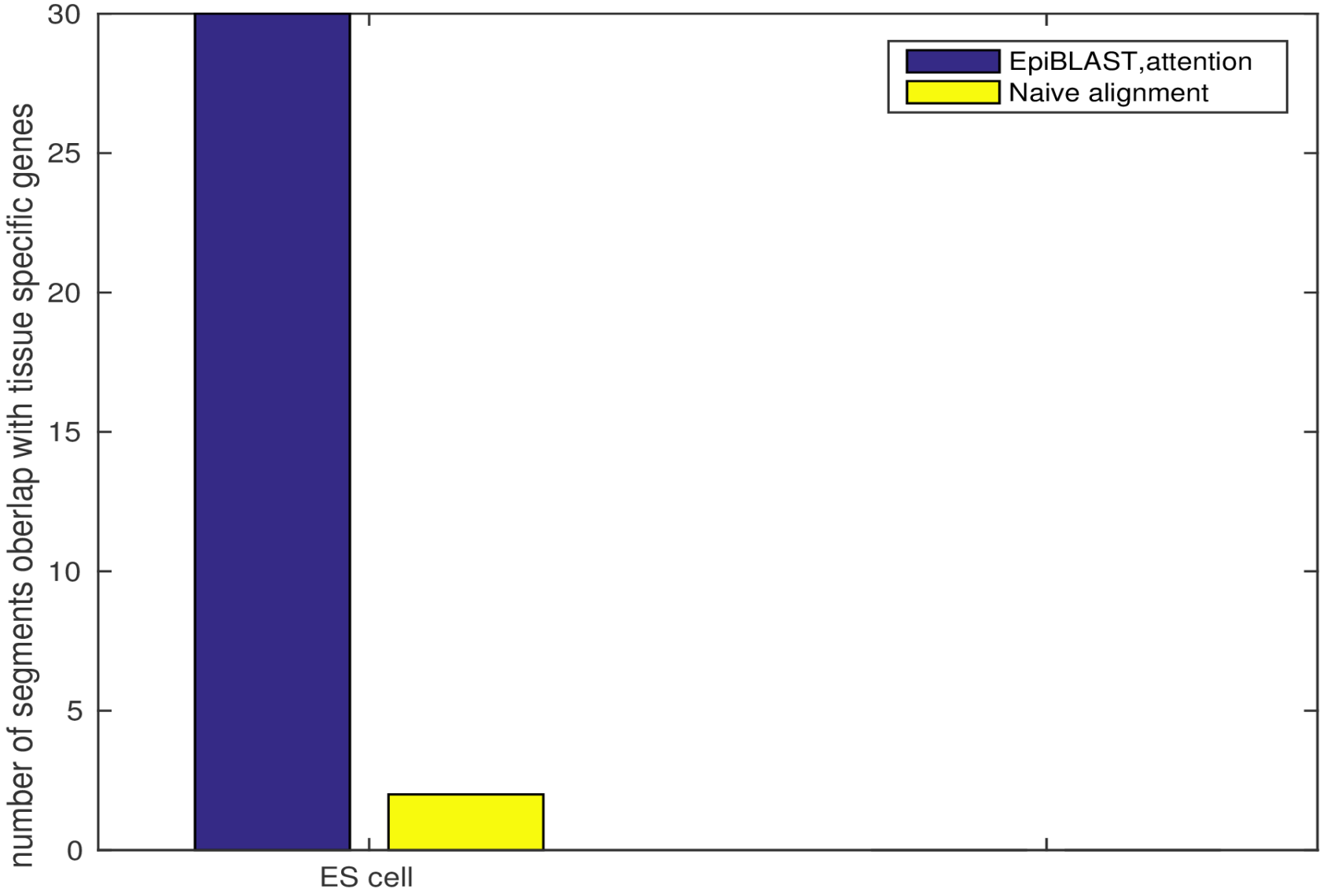
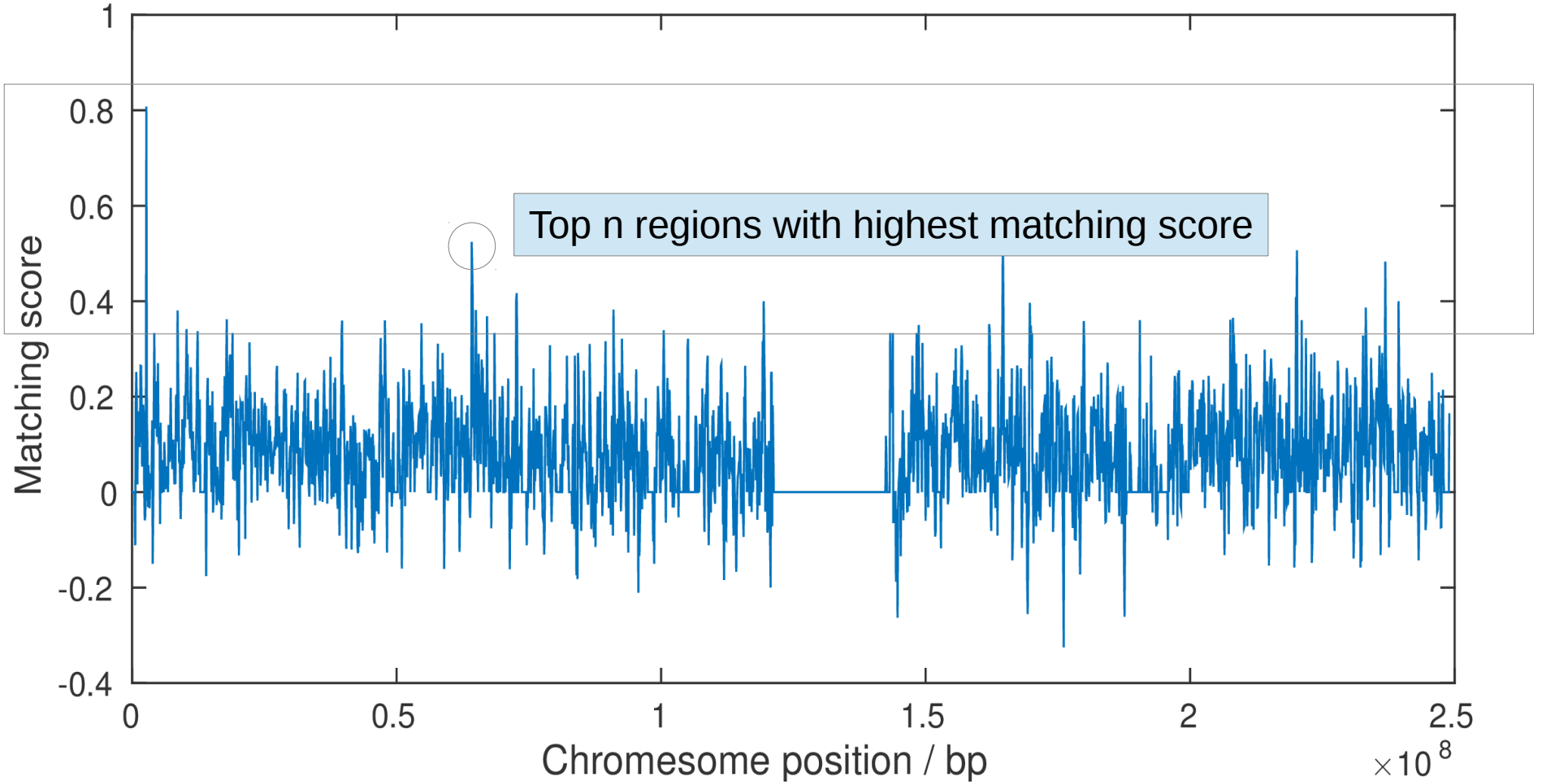


Figure 6: Motif and Case Analysis

Motif Analysis



- By reading the high matching score regions, I summarized some obvious motifs.
- The problem of this part is, it is too subjective now(by my eye). Better way(clustering)?

ab-eded Motif

- Related to activated gene.
- Consistent to existing gene annotation.
- “a” is a strong signal to the start site of a gene.

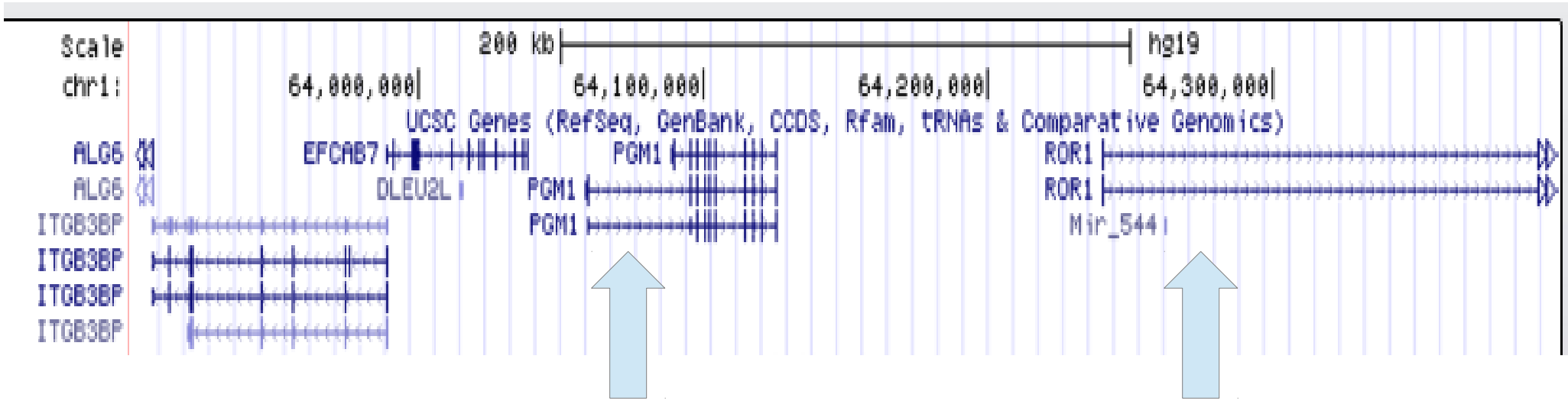
The example shown above has three ab-eded motifs.

ab-gege Motif

- This mitif is also consistent to existing gene annotation, but isn't as good as ab-eded motif.
- Similar to ab-eded motif, “a” is also a strong signal to the start site of a gene.

ab-gege Motif Example

- E003, 15 states
- chr1:63,900,001-64,400,000

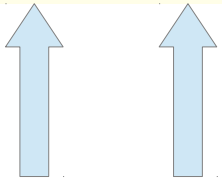


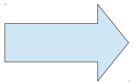
Description: Homo sapiens phosphoglucomutase 1 (PGM1), transcript variant 3, mRNA.

RefSeq Summary (NM_001172819): The protein encoded by this gene is an isozyme of phosphoglucomutase (PGM) and belongs to the phosphohexose mutase family. There are several PGM isozymes, which are encoded by different genes and catalyze the transfer of phosphate between the 1 and 6 positions of glucose. In most cell types, this PGM isozyme is predominant, representing about 90% of total PGM activity. In red cells, PGM2 is a major isozyme. This gene is highly polymorphic. Mutations in this gene cause glycogen storage disease type 14. Alternativley spliced transcript variants encoding different isoforms have been identified in this gene. [provided by RefSeq, Mar 2010].

Transcript (Including UTRs)

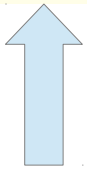
Position: hg19 chr1:64,059,480-64,125,916 **Size:** 66,437 **Total Exon Count:** 11 **Strand:** +





o,o	11000,11000	64047001-64058000,64047001-64058000
g,g	200,200	64058001-64058200,64058001-64058200
b,b	200,200	64058201-64058400,64058201-64058400
a,a	1600,1600	64058401-64060000,64058401-64060000
b,b	400,400	64060001-64060400,64060001-64060400
o,o	16400,16400	64060401-64076800,64060401-64076800
g,g	200,200	64076801-64077000,64076801-64077000
e,e	1200,1200	64077001-64078200,64077001-64078200
g,g	1000,1000	64078201-64079200,64078201-64079200
o,o	9600,9600	64079201-64088800,64079201-64088800
g,g	200,200	64088801-64089000,64088801-64089000
e,e	1800,1800	64089001-64090800,64089001-64090800
g,g	400,400	64090801-64091200,64090801-64091200
e,e	4400,4400	64091201-64095600,64091201-64095600
f,f	200,200	64095601-64095800,64095601-64095800
d,d	2400,2400	64095801-64098200,64095801-64098200
e,e	3800,3800	64098201-64102000,64098201-64102000
d,d	600,600	64102001-64102600,64102001-64102600
e,e	25200,25200	64102601-64127800,64102601-64127800
g,g	800,800	64127801-64128600,64127801-64128600
e,e	2800,2800	64128601-64131400,64128601-64131400
g,g	200,200	64131401-64131600,64131401-64131600
o,o	9400,9400	64131601-64141000,64131601-64141000
g,g	400,400	64141001-64141400,64141001-64141400

Description: Homo sapiens receptor tyrosine kinase-like orphan receptor 1 (ROR1), transcript variant 2, mRNA.
RefSeq Summary (NM_001083592): This gene encodes a receptor tyrosine kinase-like orphan receptor that modulates neurite growth in the central nervous system. The encoded protein is a glycosylated type I membrane protein that belongs to the ROR subfamily of cell surface receptors. It is a pseudokinase that lacks catalytic activity and may interact with the non-canonical Wnt signalling pathway. This gene is highly expressed during early embryonic development but expressed at very low levels in adult tissues. Increased expression of this gene is associated with B-cell chronic lymphocytic leukaemia. Alternative splicing results in multiple transcript variants encoding different isoforms. [provided by RefSeq, Jun 2012].
Transcript (Including UTRs)
Position: hg19 chr1:64,239,690-64,609,052 **Size:** 369,363 **Total Exon Count:** 7 **Strand:** +



a,a	2000,2000	64239601-64241600,64239601-64241600
b,b	200,200	64241601-64241800,64241601-64241800
a,a	200,200	64241801-64242000,64241801-64242000
b,b	600,600	64242001-64242600,64242001-64242600
g,g	400,400	64242601-64243000,64242601-64243000
b,b	200,200	64243001-64243200,64243001-64243200
g,g	23400,23400	64243201-64266600,64243201-64266600
e,e	3200,3200	64266601-64269800,64266601-64269800
g,g	1200,1200	64269801-64271000,64269801-64271000
e,e	8000,8000	64271001-64279000,64271001-64279000
h,h	3200,3200	64279001-64282200,64279001-64282200
e,e	600,600	64282201-64282800,64282201-64282800
g,g	1200,1200	64282801-64284000,64282801-64284000
e,e	400,400	64284001-64284400,64284001-64284400
g,g	200,200	64284401-64284600,64284401-64284600
e,e	2200,2200	64284601-64286800,64284601-64286800
g,g	600,600	64286801-64287400,64286801-64287400
e,e	400,400	64287401-64287800,64287401-64287800
g,g	200,200	64287801-64288000,64287801-64288000
e,e	5000,5000	64288001-64293000,64288001-64293000

ab-gege-ab-gege Gene

- As described above, the chromatin state “a” in “ab-gege” motif is strong indicator to transcription start site(tss). In most cases state “a” doesn't appear in the middle of a gene.
- ab-gege-ab-gege gene is a gene inside witch two or more “ab” segments exist, separated by “gege” repeat. Also tss is inside the first “ab” segment.
- If tss isn't inside the first “ab” segment, we call it ab-gege-ab-gege like gene.

ab-gege-ab-gege Gene

Example in E003, PBX1 is the search seed:

- PBX1:

  This gene encodes a nuclear protein that belongs to the PBX homeobox family of transcriptional factors. Studies in mice suggest that this gene may be involved in the regulation of osteogenesis, and required for skeletal patterning and programming. A chromosomal translocation, t(1;19) involving this gene and TCF3/E2A gene, is associated with pre-B-cell acute lymphoblastic leukemia. The resulting fusion protein, in which the DNA binding domain of E2A is replaced by the DNA binding domain of this protein, transforms cells by constitutively activating transcription of genes regulated by the PBX protein family. Alternatively spliced transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Mar 2011].

- JARID2:

  This gene encodes a Jumonji- and AT-rich interaction domain (ARID)-domain-containing protein. The encoded protein is a DNA-binding protein that functions as a transcriptional repressor. This protein interacts with the Polycomb repressive complex 2 (PRC2) which plays an essential role in regulating gene expression during embryonic development. This protein facilitates the recruitment of the PRC2 complex to target genes. Alternate splicing results in multiple transcript variants. Mutations in this gene are associated with chronic myeloid malignancies. [provided by RefSeq, May 2012].

ab-gege-ab-gege Gene

Example in E003:

- PHLPP1:

This gene encodes a member of the serine/threonine phosphatase family. The encoded protein promotes apoptosis by dephosphorylating and inactivating the serine/threonine kinase Akt, and functions as a tumor suppressor in multiple types of cancer. Increased expression of this gene may also play a role in obesity and type 2 diabetes by interfering with Akt-mediated insulin signaling. [provided by RefSeq, Dec 2011]. Sequence Note: This RefSeq record was created from transcript and genomic sequence data because no single transcript was available for the full length of the gene. The extent of this transcript is supported by transcript alignments and orthologous data. CCDS Note: The coding region has been updated to extend the N-terminus to one that is more supported by available conservation data and publications. There are no publicly available human transcripts that include the extended region. However, the update is supported by homologous transcript data and is consistent with the full-length 190 kDa human isoform described in the literature. This 190 kDa product, known as PHLPP1beta, has been detected in several studies, including PMIDs 17386267, 19079341, 20089132, 20819118 and 20861921. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications. ##Evidence-Data-START## Transcript exon combination :: AB011178.2, BC014927.2 [ECO:0000332] RNAseq introns :: single sample supports all introns SAMEA1965299, SAMEA1966682 [ECO:0000348] ##Evidence-Data-END##

ab-gege-ab-gege like Gene

Example in E003:

- ARID1B:

  This locus encodes an AT-rich DNA interacting domain-containing protein. The encoded protein is a component of the SWI/SNF chromatin remodeling complex and may play a role in cell-cycle activation. The protein encoded by this locus is similar to AT-rich interactive domain-containing protein 1A. These two proteins function as alternative, mutually exclusive ARID-subunits of the SWI/SNF complex. The associated complexes play opposing roles. Alternatively spliced transcript variants encoding different isoforms have been described. [provided by RefSeq, Feb 2012].

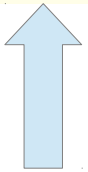
- GULP1:

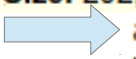
  The protein encoded by this gene is an adapter protein necessary for the engulfment of apoptotic cells by phagocytes. Several transcript variants, some protein coding and some thought not to be protein coding, have been found for this gene. [provided by RefSeq, Nov 2011].

ab-gege-ab-gege Gene: PBX1

Human Gene PBX1 (uc010pku.2) Description and Page Index

Description: Homo sapiens pre-B-cell leukemia homeobox 1 (PBX1), transcript variant 3, mRNA.
RefSeq Summary (NM_001204963): This gene encodes a nuclear protein that belongs to the PBX homeobox family of transcriptional factors. Studies in mice suggest that this gene may be involved in the regulation of osteogenesis, and required for skeletal patterning and programming. A chromosomal translocation, t(1;19) involving this gene and TCF3/E2A gene, is associated with pre-B-cell acute lymphoblastic leukemia. The resulting fusion protein, in which the DNA binding domain of E2A is replaced by the DNA binding domain of this protein, transforms cells by constitutively activating transcription of genes regulated by the PBX protein family. Alternatively spliced transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Mar 2011].
Transcript (Including UTRs)
Position: hg19 chr1:164,528,597-164,821,060 **Size:** 292,464 **Total Exon Count:** 9 **Strand:** +





a,a	5400,5400	164527601-164533000,164527601-164533000
b,b	200,200	164533001-164533200,164533001-164533200
a,a	200,200	164533201-164533400,164533201-164533400
g,g	3800,3800	164533401-164537200,164533401-164537200
f,f	200,200	164537201-164537400,164537201-164537400
e,e	400,400	164537401-164537800,164537401-164537800
f,f	200,200	164537801-164538000,164537801-164538000
e,e	400,400	164538001-164538400,164538001-164538400
d,d	1800,1800	164538401-164540200,164538401-164540200
e,e	200,200	164540201-164540400,164540201-164540400
d,d	200,200	164540401-164540600,164540401-164540600
e,e	1200,1200	164540601-164541800,164540601-164541800
g,g	800,800	164541801-164542600,164541801-164542600
e,e	400,400	164542601-164543000,164542601-164543000
g,g	1400,1400	164543001-164544400,164543001-164544400
b,b	200,200	164544401-164544600,164544401-164544600
a,a	800,800	164544601-164545400,164544601-164545400
b,b	1000,1000	164545401-164546400,164545401-164546400
g,g	2600,2600	164546401-164549000,164546401-164549000
e,e	2000,2000	164549001-164551000,164549001-164551000
g,g	4400,4400	164551001-164555400,164551001-164555400
f,f	2400,2400	164555401-164557800,164555401-164557800

Figure 7: vertical alignment