

Alignment of Epigenetic Sequences

Zhang Haowen in Jessica's lab

Sequence alignment in Biology

$$\begin{array}{l} a^\diamond = \text{AC-GG-AT} \\ b^\diamond = \text{-CCGCT-T} \end{array} \quad \text{or} \quad \begin{array}{l} a^\diamond = \text{ACGG---AT} \\ b^\diamond = \text{--CCGCT-T} \end{array}$$

$$D_w(a, b) = \min\{w(a^\diamond, b^\diamond) \mid (a^\diamond, b^\diamond) \text{ is alignment of } a \text{ and } b\}.$$

$$w(a^\diamond, b^\diamond) = \sum_{i=1}^{|a^\diamond|} w(a_i^\diamond, b_i^\diamond)$$

Smith-Waterman algorithm

$$D_{ij} = \min \begin{cases} D_{i-1,j-1} + w(a_i, b_j) & (\text{match}) \\ D_{i-1,j} + w(a_i, -) & (\text{deletion}) \\ D_{i,j-1} + w(-, b_j) & (\text{insertion}) \end{cases}$$

A A G T

	0	1	2	3	4
A	1	0	1	2	3
T	2	1	1	2	2

AAGT

--A--T

$O(mn)$ time and space

BLAST becomes an extremely successful tool

Basic local alignment search tool

[SF Altschul](#), [W Gish](#), [W Miller](#), [EW Myers](#)... - Journal of molecular ..., 1990 - Elsevier

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of ...

[Cited by 56022](#) [Related articles](#) [All 103 versions](#) [Cite](#) [Save](#)

- Label-free method, flexible in application.
- Similar to nearest neighbor search.
- Possible to handle large database search.

Example:



Source image

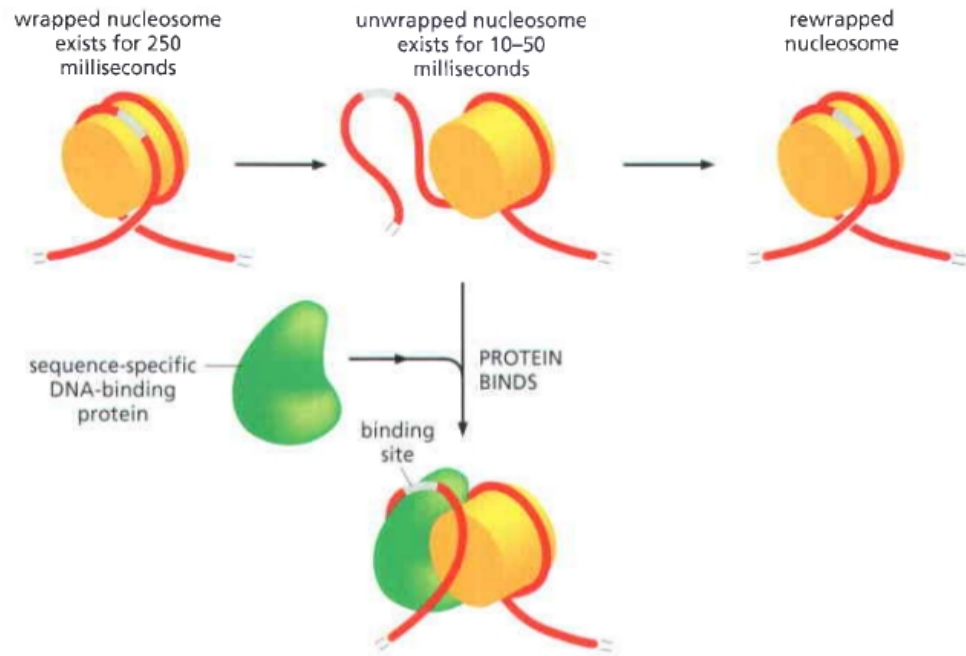


K Nearest Neighbors

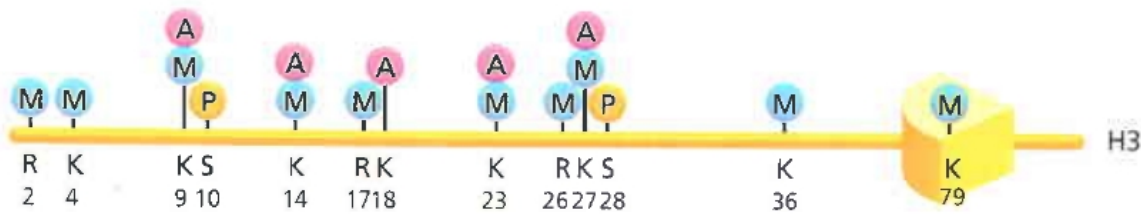
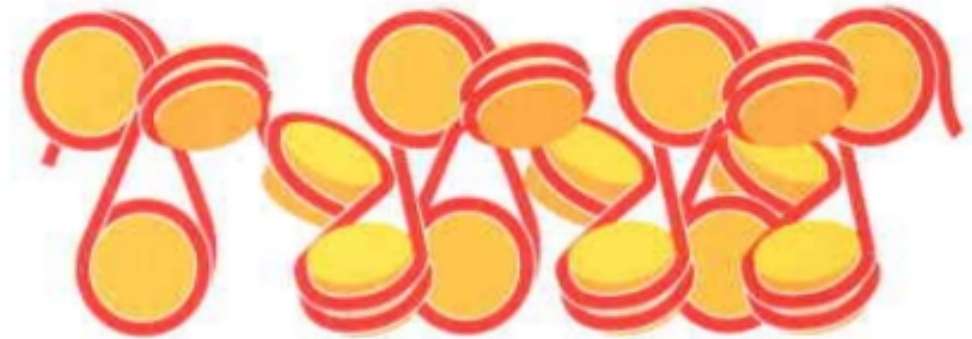
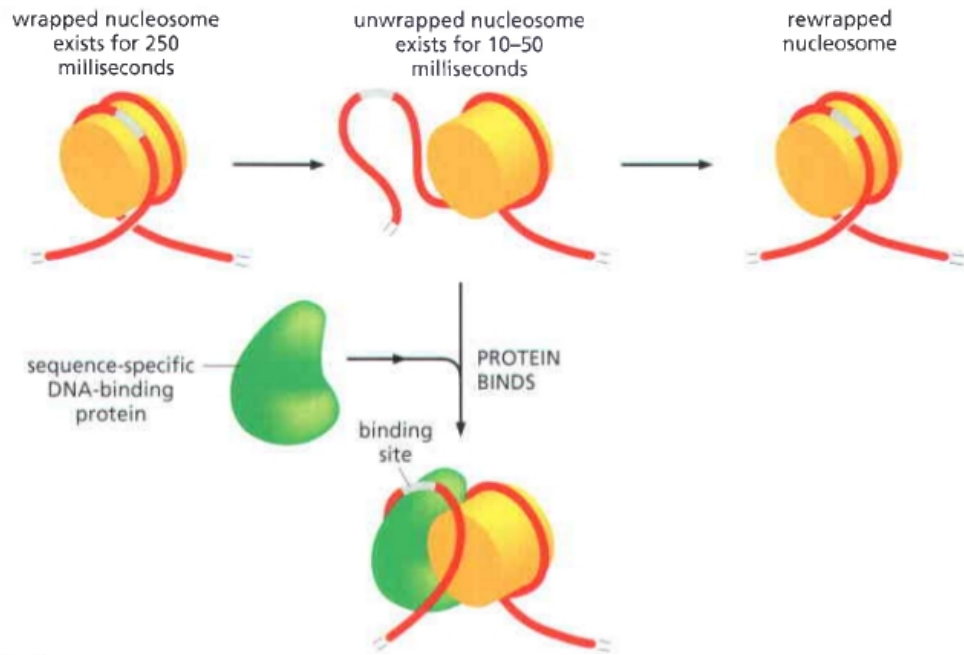
Image Segmentation



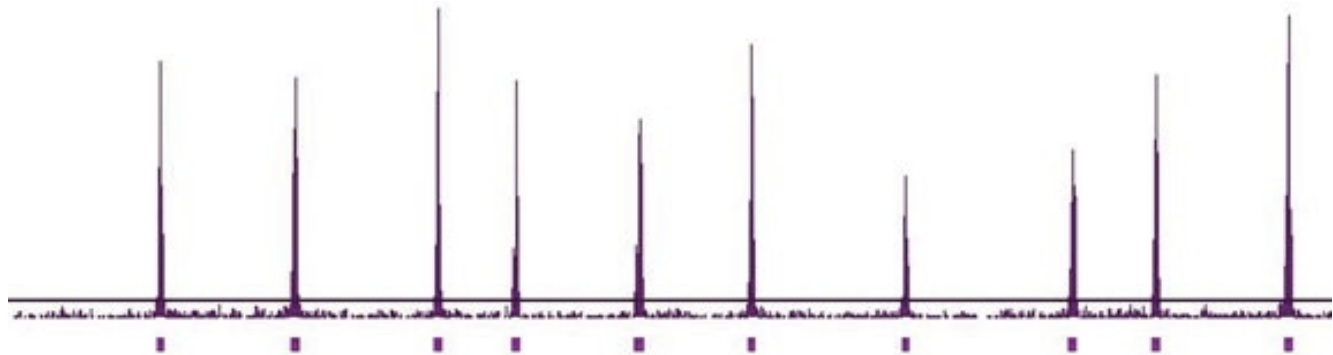
Epigenome



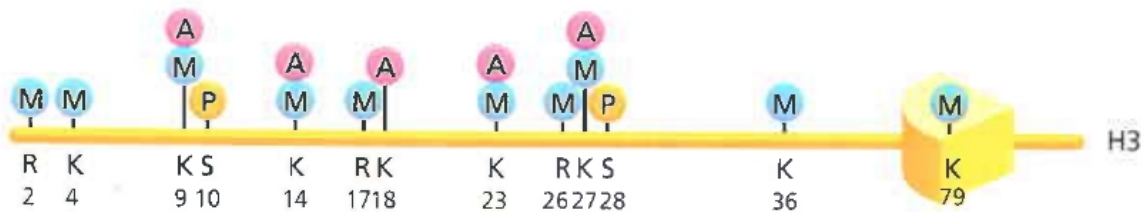
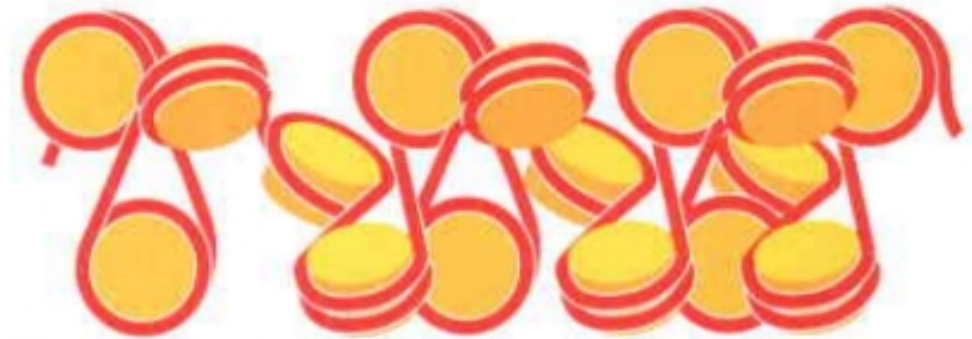
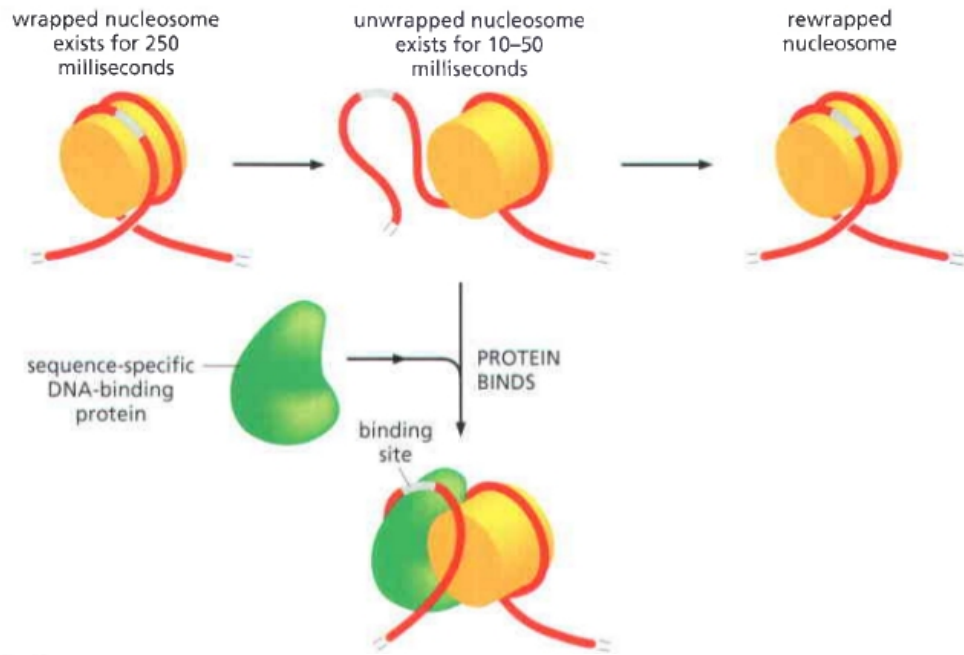
Epigenome



- Each modification corresponds to one channel.

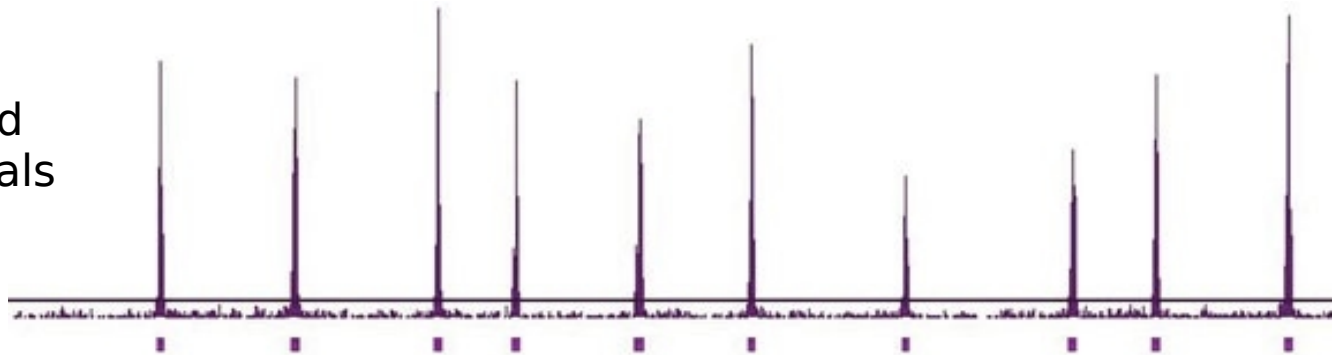


Epigenome



- Each modification corresponds to one channel.

- Epigenome could be represented by multi-channel sequence signals indexed by genome.



Multi-channel epigenetic databases

An integrated encyclopedia of DNA elements in the human genome

2012

The ENCODE Project Consortium*

1640 channels of 147 cell types

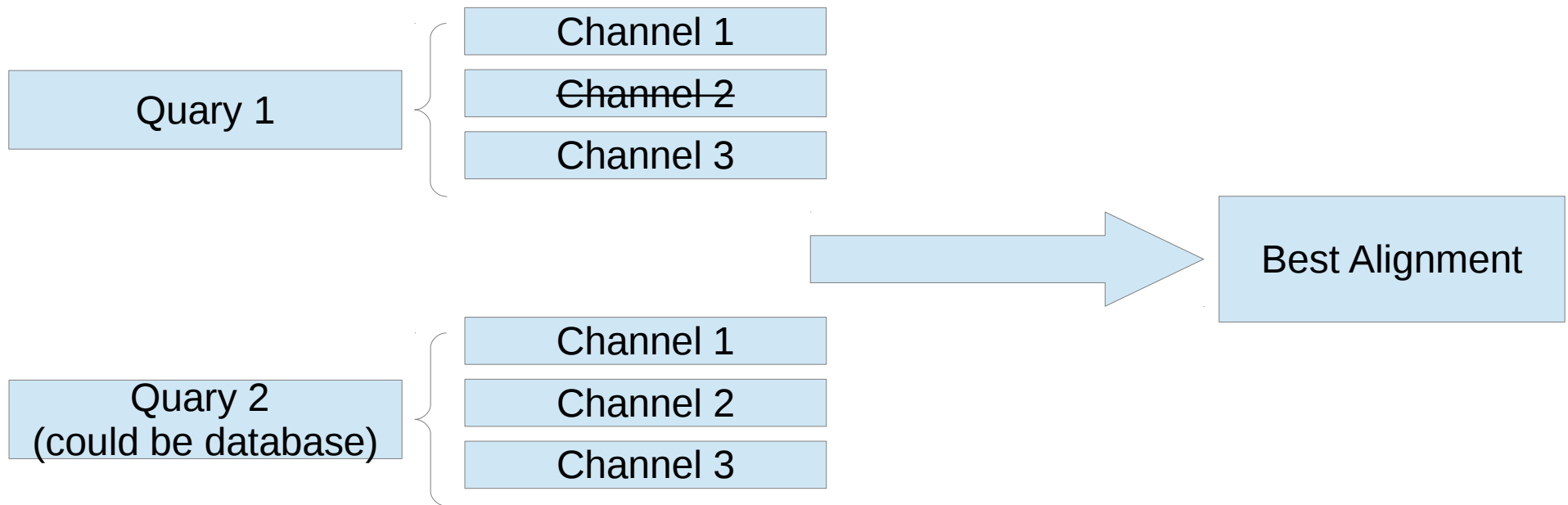
Integrative analysis of 111 reference human epigenomes

Roadmap Epigenomics Consortium†, Anshul Kundaje^{1,2,3*}, Wouter Meuleman^{1,2*}, Jason Ernst^{1,2,4*}, Misha Bilenky^{5*}, Angela Yen^{1,2}, Alireza Heravi-Moussavi⁵, Pouya Kheradpour^{1,2}, Zhizhuo Zhang^{1,2}, Jianrong Wang^{1,2}, Michael J. Ziller^{2,6}, Viren Amin⁷, John W. Whitaker⁸, Matthew D. Schultz⁹, Lucas D. Ward^{1,2}, Abhishek Sarkar^{1,2}, Gerald Quon^{1,2}, Richard S. Sandstrom¹⁰, Matthew L. Eaton^{1,2}, Yi-Chieh Wu^{1,2}, Andreas R. Pfenning^{1,2}, Xinchun Wang^{1,2,11}, Melina Claussnitzer^{1,2}, Yaping Liu^{1,2}, Cristian Coarfa⁷, R. Alan Harris⁷, Noam Shores², Charles B. Epstein², Elizabeta Gjoneska^{2,12}, Danny Leung^{8,13}, Wei Xie^{8,13}, R. David Hawkins^{8,13}, Ryan Lister⁹, Chibo Hong¹⁴, Philippe Gascard¹⁵, Andrew J. Mungall⁵, Richard Moore⁵, Eric Chuah⁵, Angela Tam⁵, Theresa K. Canfield¹⁰, R. Scott Hansen¹⁶, Rajinder Kaul¹⁶, Peter J. Sabo¹⁰, Mukul S. Bansal^{1,2,17}, Annaick Carles¹⁸, Jesse R. Dixon^{8,13}, Kai-How Farh², Soheil Feizi^{1,2}, Rosa Karlic¹⁹, Ah-Ram Kim^{1,2}, Ashwinikumar Kulkarni²⁰, Daofeng Li²¹, Rebecca Lowdon²¹, GiNell Elliott²¹, Tim R. Mercer²², Shane J. Neph¹⁰, Vitor Onuchic⁷, Paz Polak^{2,23}, Nisha Rajagopal^{8,13}, Pradipta Ray²⁰, Richard C. Sallari^{1,2}, Kyle T. Siebenthall¹⁰, Nicholas A. Sinnott-Armstrong^{1,2}, Michael Stevens^{21,42}, Robert E. Thurman¹⁰, Jie Wu^{24,25}, Bo Zhang²¹, Xin Zhou²¹, Arthur E. Beaudet²⁶, Laurie A. Boyer¹¹, Philip L. De Jager^{2,23,27}, Peggy J. Farnham²⁸, Susan J. Fisher²⁹, David Haussler³⁰, Steven J. M. Jones^{5,31,32}, Wei Li³³, Marco A. Marra^{5,32}, Michael T. McManus³⁴, Shamil Sunyaev^{2,23,27}, James A. Thomson^{35,41}, Thea D. Tlsty¹⁵, Li-Huei Tsai^{2,12}, Wei Wang⁸, Robert A. Waterland³⁶, Michael Q. Zhang^{20,37}, Lisa H. Chadwick³⁸, Bradley E. Bernstein^{2,39,40§}, Joseph F. Costello^{14§}, Joseph R. Ecker^{9§}, Martin Hirst^{5,18§}, Alexander Meissner^{2,6§}, Aleksandar Milosavljevic^{7§}, Bing Ren^{8,13§}, John A. Stamatoyannopoulos^{10§}, Ting Wang^{21§} & Manolis Kellis^{1,2§}

2015

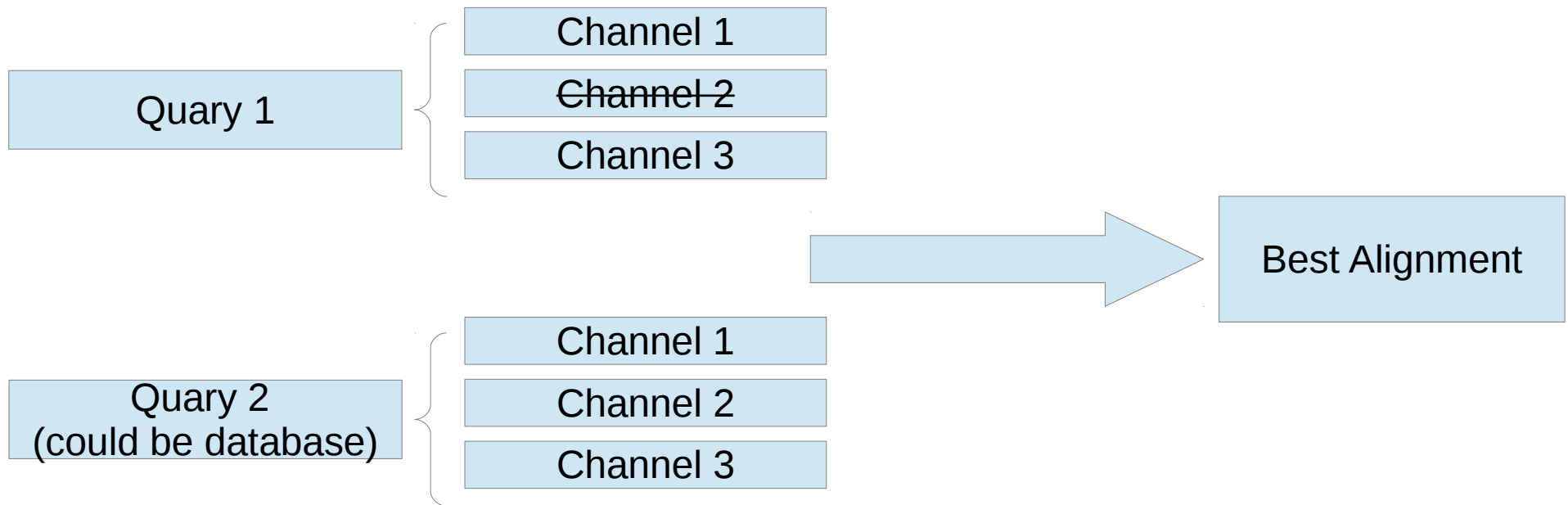
Epigenome of 111 tissues and cell types

Blast in Epigenome



- The goal is to design a multi-channel blast algorithm for epigenome.
- Should be fast to handled database search.

Blast in Epigenome



- Functional elements search and annotation.
- Study the conservation correlation between genome and epigenome.
- Looking for large scale genetic structure.

Proposed strategy

Training:

Multi-channel Epigenome Database

Model Learning

Feature sequence

Query: Unsupervised Feature Extraction

Query Sequence

Coding

Feature Sequence

Scanning

Results

Multi-channel Epigenome Database

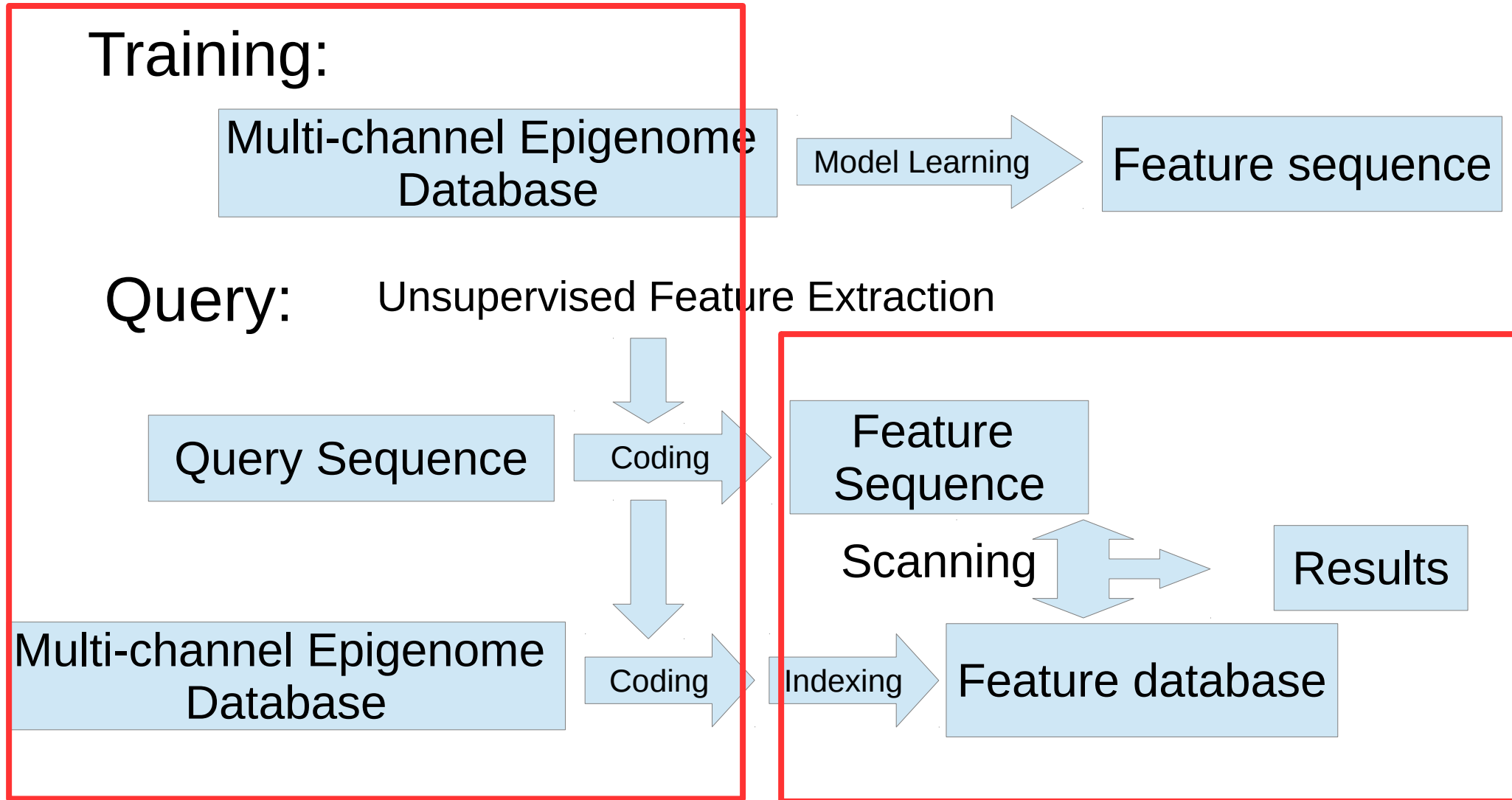
Coding

Indexing

Feature database

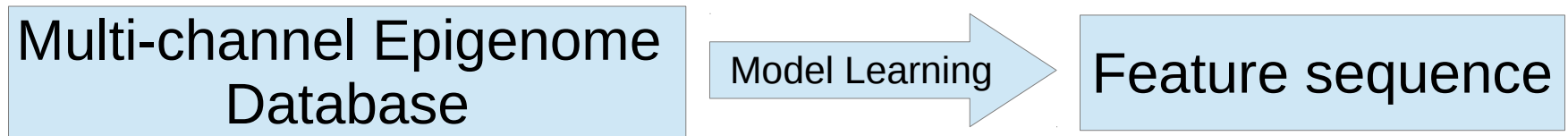
Coding

Alignment

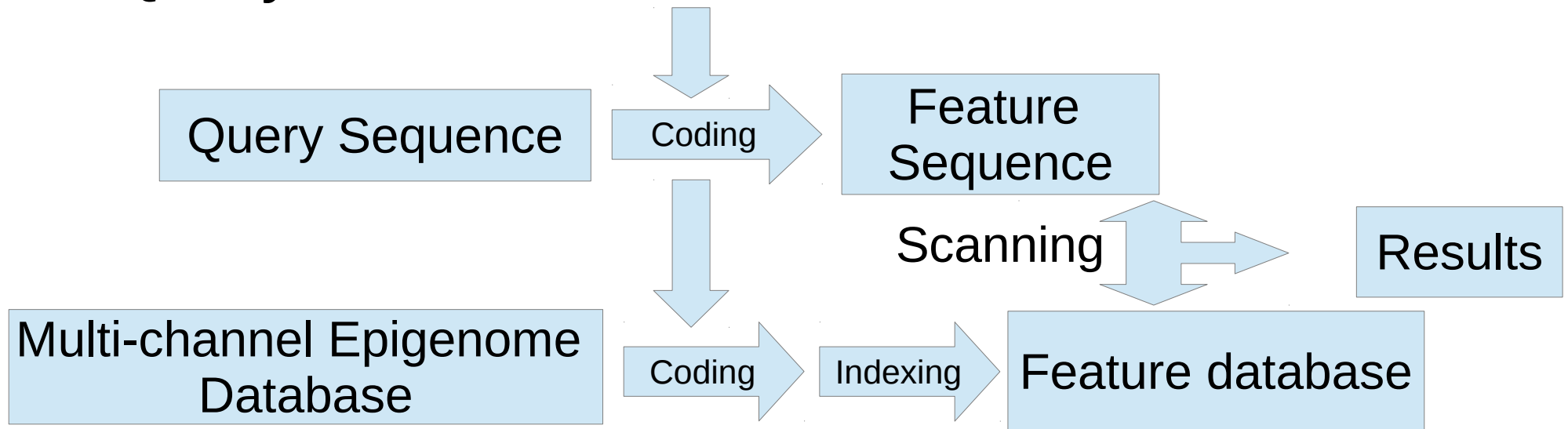


Proposed strategy

Training:



Query: Unsupervised Feature Extraction



Feature sequences are easier to handle, hence they are more feasible for large database search.

Proposed strategy

Training:

Multi-channel Epigenome
Database

HMM

Model Learning

Feature sequence

Query:

HMM Model

Query Sequence

Coding

Feature
Sequence

Scanning

Results

Multi-channel Epigenome
Database

Coding

Indexing

Feature database

Proposed Strategy

Training:

Multi-channel Epigenome Database

HMM

Model Learning

Feature sequence

Query:

HMM Model

Query Sequence

Coding

Bag of Words Sequence

Scanning

Multi-channel Epigenome Database

Coding

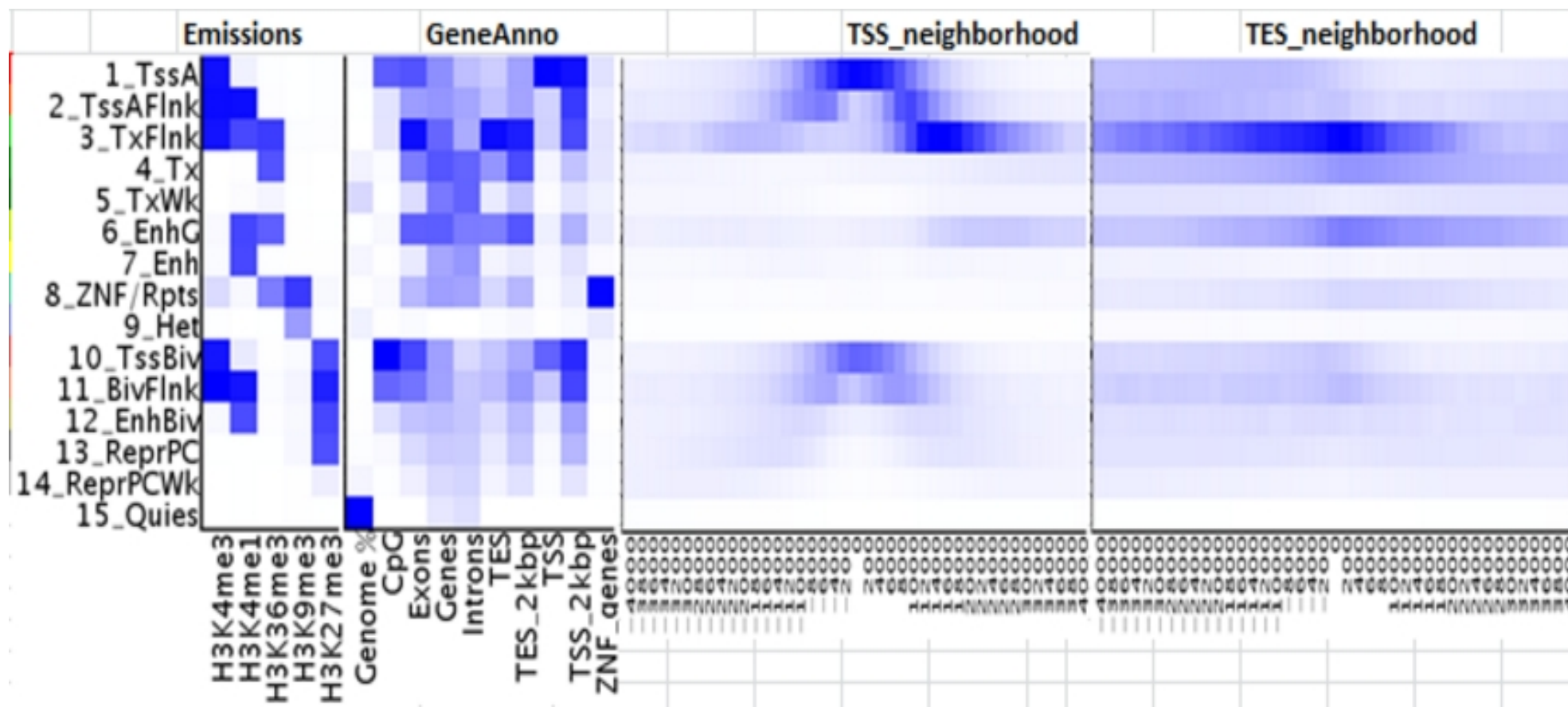
Indexing

Bag of Words Table

Results

Position based alignment V.S. Variation based alignment

ChromHMM data Illustration

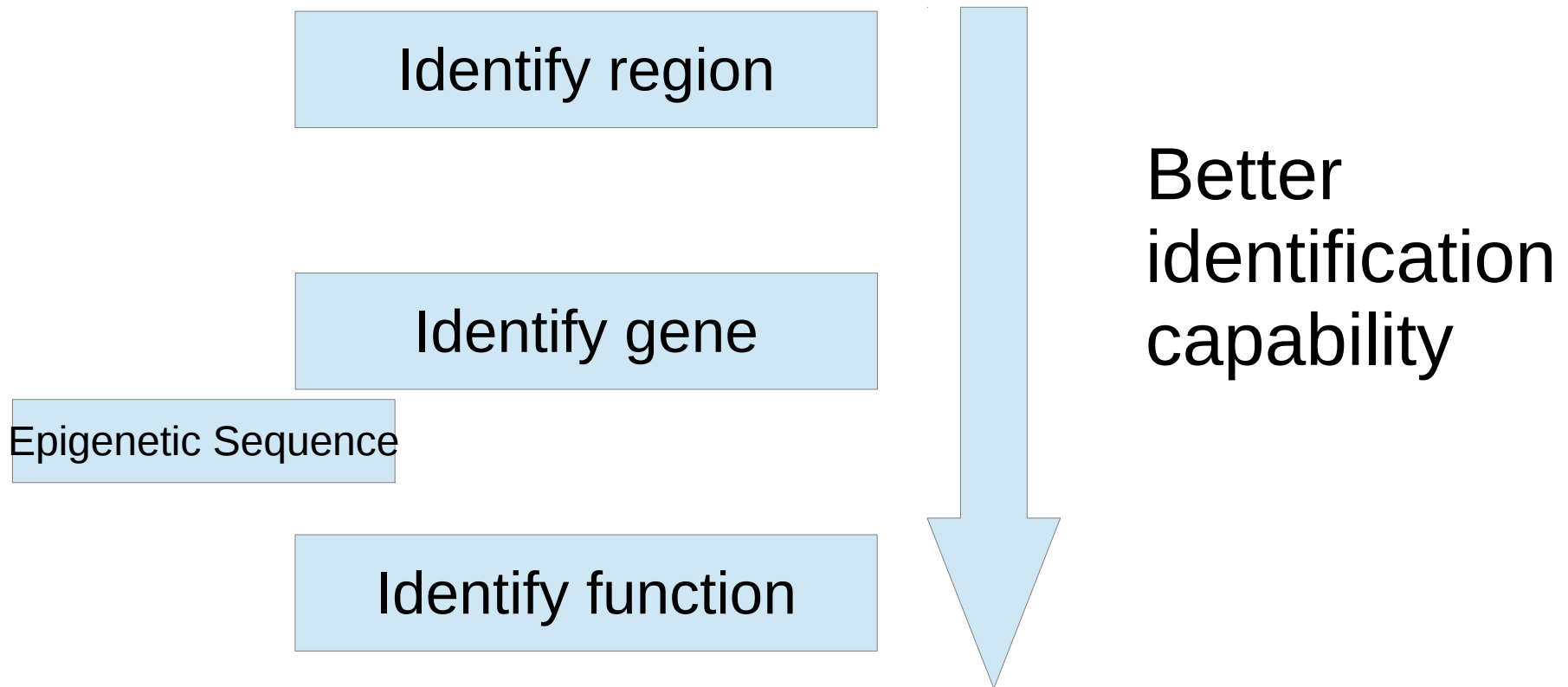


http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/annotationEnrichment_RoadmapEp_coreMarks_15State.png

ChromHMM Data (E003, chr1)

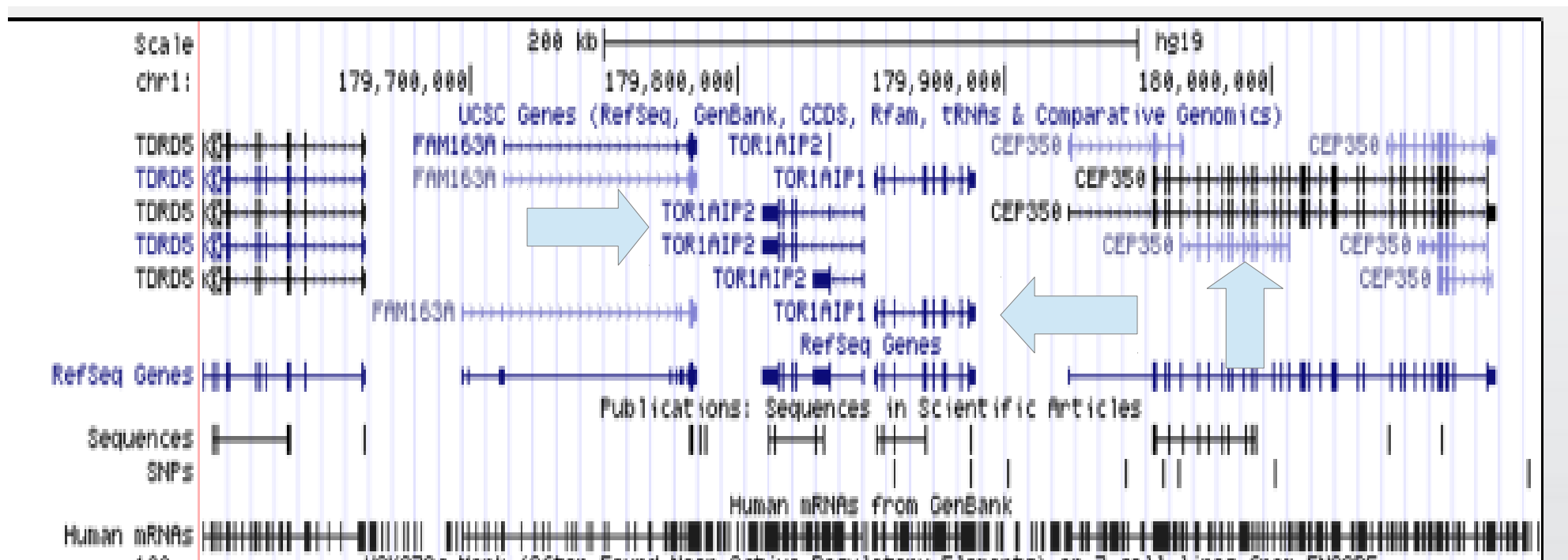
- E003chr1

3. Identification capability of epigenetic sequences



Identification capability of epigenetic sequences: Example 1

- E003, 15 states
- chr1:179600001-180100000



Human Gene TOR1AIP2 (uc001gnk.3) Description and Page Index

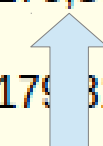
Description: Homo sapiens torsin A interacting protein 2 (TOR1AIP2), transcript variant 2, mRNA.

Transcript (Including UTRs)

Position: hg19 chr1:179,809,102-179,846,941 **Size:** 37,840 **Total Exon Count:** 6 **Strand:** -

Coding Region

Position: hg19 chr1:179,815,206-179,821,800 **Size:** 6,595 **Coding Exon Count:** 4



g,g	1200,1200	179794401-179795600,179794401-179795600
e,e	5200,5200	179795601-179800800,179795601-179800800
g,g	1000,1000	179800801-179801800,179800801-179801800
e,e	400,400	179801801-179802200,179801801-179802200
g,g	200,200	179802201-179802400,179802201-179802400
e,e	12600,12600	179802401-179815000,179802401-179815000
d,d	800,800	179815001-179815800,179815001-179815800
e,e	2200,2200	179815801-179818000,179815801-179818000
d,d	2600,2600	179818001-179820600,179818001-179820600
e,e	10400,10400	179820601-179831000,179820601-179831000
d,d	1200,1200	179831001-179832200,179831001-179832200
e,e	400,400	179832201-179832600,179832201-179832600
d,d	2400,2400	179832601-179835000,179832601-179835000
e,e	7800,7800	179835001-179842800,179835001-179842800
d,d	1400,1400	179842801-179844200,179842801-179844200
e,e	1600,1600	179844201-179845800,179844201-179845800
a,a	1600,1600	179845801-179847400,179845801-179847400

Human Gene TOR1AIP1 (uc001gnq.4) Description and Page Index



Description: Homo sapiens torsin A interacting protein 1 (TOR1AIP1), transcript variant 2, mRNA.
RefSeq Summary (NM_015602): This gene encodes a type 2 integral membrane protein that binds A- and B-type lamins. The encoded protein localizes to the inner nuclear membrane and may be involved in maintaining the attachment of the nuclear membrane to the nuclear lamina during cell division. Alternate splicing results in multiple transcript variants. [provided by RefSeq, Apr 2016].

Transcript (Including UTRs)

Position: hg19 chr1:179,851,177-179,889,212 **Size:** 38,036 **Total Exon Count:** 10 **Strand:** +

Coding Region

Position: hg19 chr1:179,851,177-179,889,212 **Size:** 35,737 **Coding Exon Count:** 10

	a,a	2800,2800	179850601-179853400,179850601-179853400
	e,e	19200,19200	179853401-179872600,179853401-179872600
	d,d	1800,1800	179872601-179874400,179872601-179874400
	e,e	12400,12400	179874401-179886800,179874401-179886800
	d,d	800,800	179886801-179887600,179886801-179887600
	e,e	2600,2600	179887601-179890200,179887601-179890200
	g,g	1600,1600	179890201-179891800,179890201-179891800

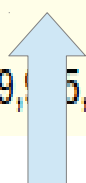
Human Gene CEP350 (uc001gnt.3) Description and Page Index

Transcript (Including UTRs)

Position: hg19 chr1:179,923,908-180,084,015 Size: 160,108 Total Exon Count: 38 Strand: +

Coding Region

Position: hg19 chr1:179,925,317-180,080,296 Size: 124,980 Coding Exon Count: 37



a,a	2800,2800	179922801-179925600,179922801-179925600
o,o	20000,20000	179925601-179945600,179925601-179945600
g,g	800,800	179945601-179946400,179945601-179946400
e,e	19200,19200	179946401-179965600,179946401-179965600
d,d	1000,1000	179965601-179966600,179965601-179966600
e,e	16400,16400	179966601-179983000,179966601-179983000
d,d	2400,2400	179983001-179985400,179983001-179985400
e,e	3600,3600	179985401-179989000,179985401-179989000
d,d	1800,1800	179989001-179990800,179989001-179990800
e,e	800,800	179990801-179991600,179990801-179991600
d,d	1200,1200	179991601-179992800,179991601-179992800
e,e	400,400	179992801-179993200,179992801-179993200
d,d	600,600	179993201-179993800,179993201-179993800
e,e	11200,11200	179993801-180005000,179993801-180005000
d,d	400,400	180005001-180005400,180005001-180005400
e,e	5400,5400	180005401-180010800,180005401-180010800
d,d	5800,5800	180010801-180016600,180010801-180016600
e,e	2400,2400	180016601-180019000,180016601-180019000
d,d	400,400	180019001-180019400,180019001-180019400
e,e	2800,2800	180019401-180022200,180019401-180022200
d,d	1600,1600	180022201-180023800,180022201-180023800
e,e	1800,1800	180023801-180025600,180023801-180025600
d,d	600,600	180025601-180026200,180025601-180026200
e,e	5200,5200	180026201-180031400,180026201-180031400
d,d	4200,4200	180031401-180035600,180031401-180035600

Human Gene CEP350 (uc001gnt.3) Description and Page Index

Transcript (Including UTRs)

Position: hg19 chr1:179,923,908-180,084,015 Size: 160,108 Total Exon Count: 38 Strand: +

Coding Region

Position: hg19 chr1:179,955,317-180,000,296 Size: 124,980 Coding Exon Count: 37



e,e	8400,8400	180035601-180044000,180035601-180044000
d,d	600,600	180044001-180044600,180044001-180044600
e,e	5000,5000	180044601-180049600,180044601-180049600
d,d	800,800	180049601-180050400,180049601-180050400
e,e	9600,9600	180050401-180060000,180050401-180060000
d,d	1400,1400	180060001-180061400,180060001-180061400
e,e	800,800	180061401-180062200,180061401-180062200
d,d	4200,4200	180062201-180066400,180062201-180066400
e,e	800,800	180066401-180067200,180066401-180067200
d,d	1000,1000	180067201-180068200,180067201-180068200
e,e	12000,12000	180068201-180080200,180068201-180080200
d,d	400,400	180080201-180080600,180080201-180080600
e,e	10400,10400	180080601-180091000,180080601-180091000
o,o	4200,4200	180091001-180095200,180091001-180095200



Negative Control: Randomized Sequences

- Gram0 randomization

Shuffle HMM state regions

aabbbc → bbbaac

- Gram1 randomization

$$P(s_{t+1} \mid s_{\leq t}) = P(s_{t+1} \mid s_t)$$

Negative Control: Randomized Sequences

- Gram0 randomization

Shuffle HMM state regions

aabbbc → bbbaac

- Gram1 randomization

$$P(s_{t+1} \mid s_{\leq t}) = P(s_{t+1} \mid s_t)$$

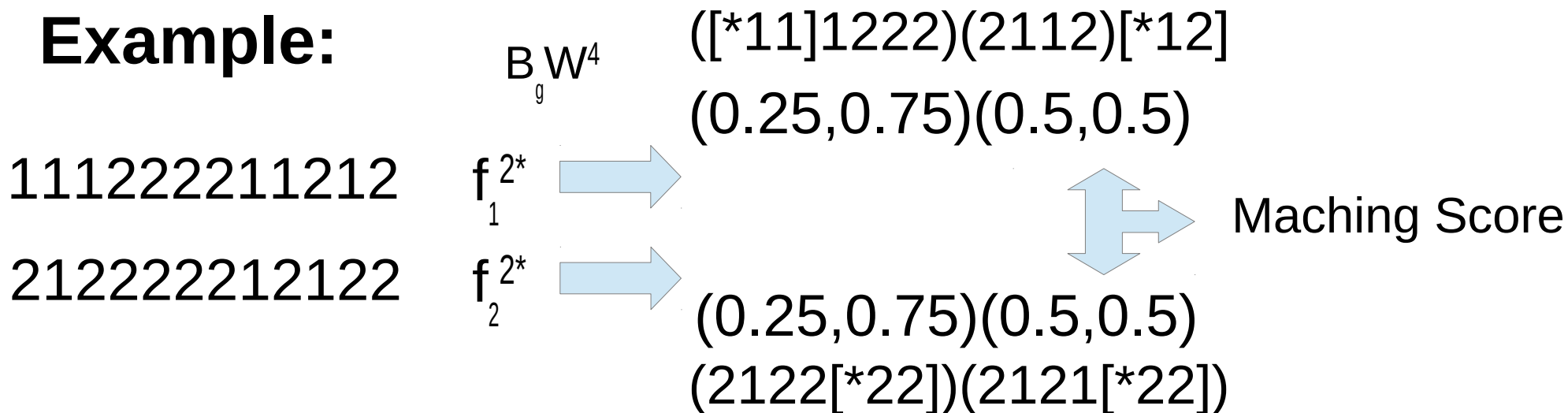
Our alignment algorithm should be able to distinct biological sequences and randomized sequences.

Position based alignment

Regional bag-of-words

- Here is an example of aligning sequences of 2 words(states). Sequences are aligned by deleting of some characters such that each sequence region have similar words(states) frequency.
- Traditional alignment is a special case of regional bag-of-words alignment

Example:



Smith-Waterman algorithm for Regional bag-of-words alignment

$$H_{BoW}(i, j) = 0, \text{ for } j < m, 0 \leq i \leq l_1;$$

$$H_{BoW}(i, j) = 0, \text{ for } i < m, 0 \leq j \leq l_2;$$

$$H_{BoW}(i, j) = \min \begin{cases} H_{BoW}(i - m, j - m) + d \left(s_1^{[i-m+1, i]}, s_2^{[j-m+1, j]} \right) & \text{(Mis)match} \\ H_{BoW}(i - k, j) + \delta & \text{Deletion in } s_1 \\ H_{BoW}(i, j - k) + \delta & \text{Deletion in } s_2 \end{cases}$$

if $m \leq i \leq l_1, m \leq j \leq l_2,$

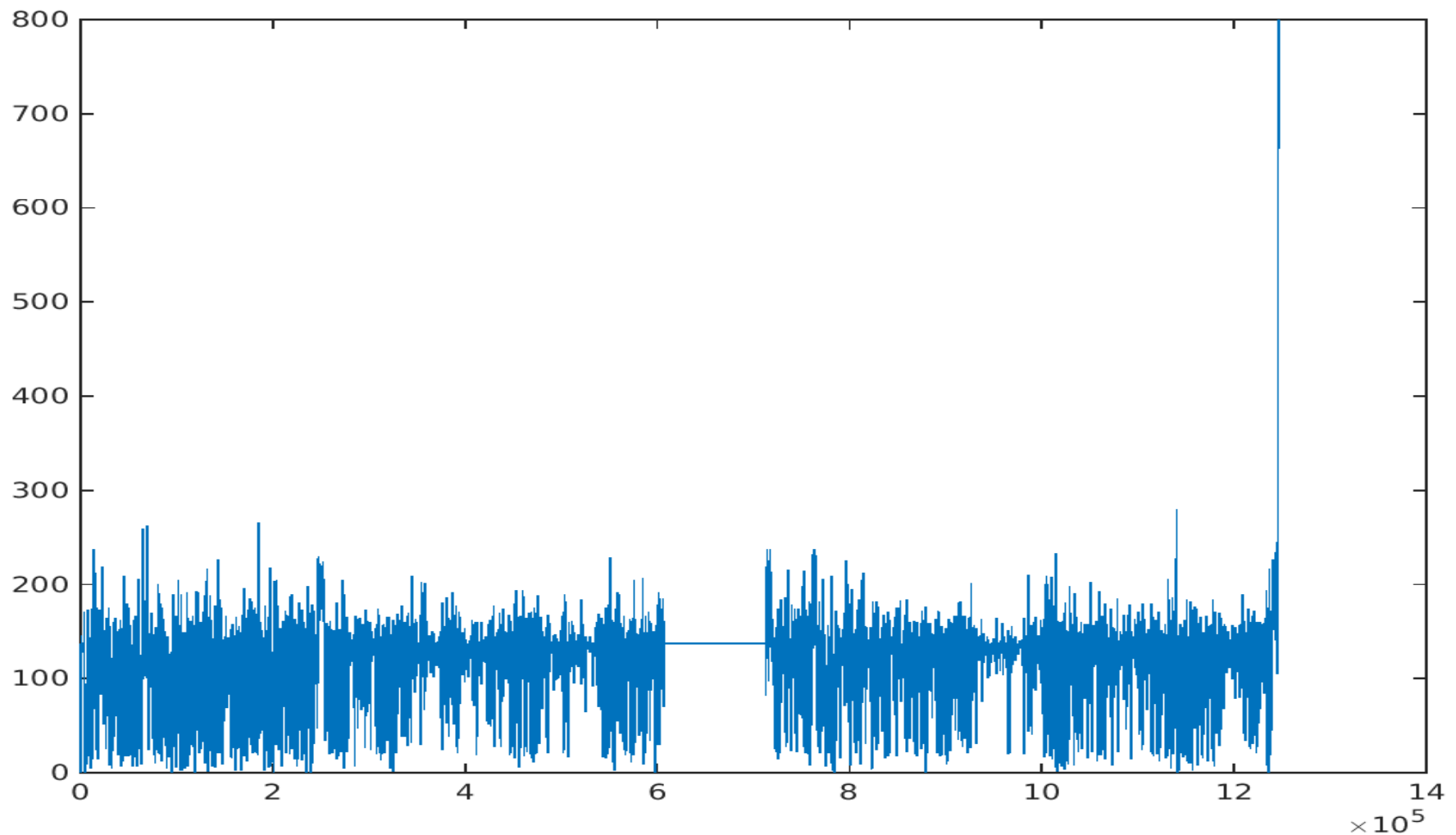
- H is the matrix of matching score.
- m is the region length word frequency is counted on.

Database query example

- A epigenome region is picked as the query sequence, database is chr 1, E003.
- Matching score along chromosome 1 is recorded.
- This region is [249053801, 249213801]

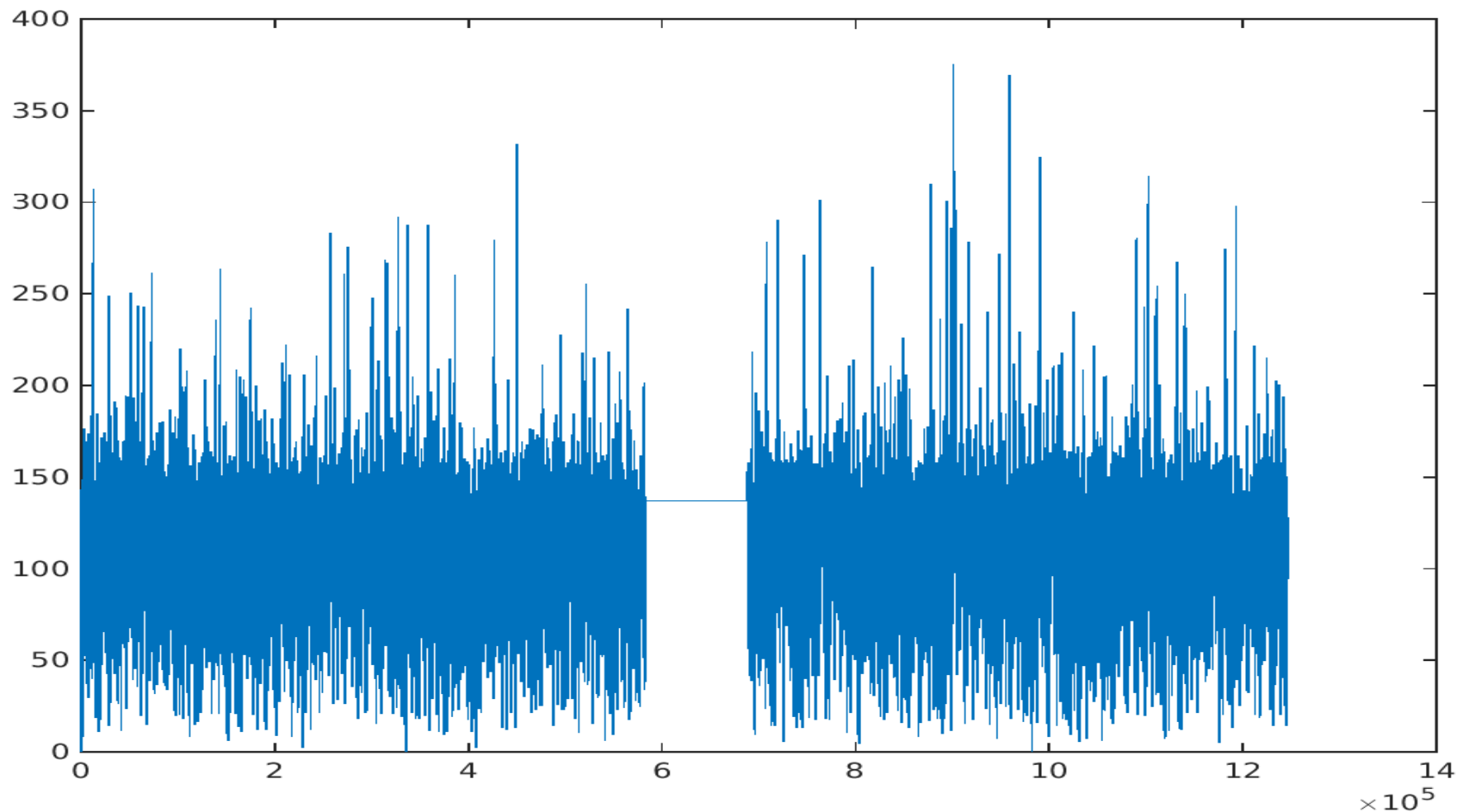
Position based alignment couldn't distinct biological and random sequences

Biological score:



Position based alignment couldn't distinct biological and random sequences

Randomized score (gram 0):

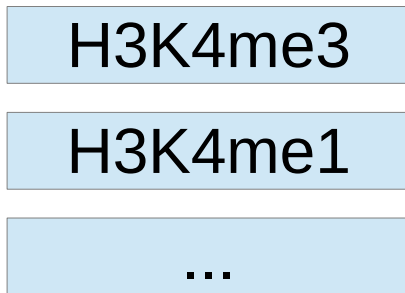


Position based alignment prefers to pair long repetitive states

- Example:

Variation based alignment

Epigenetic region 1:



Chromatin state
sequence 1:

aabeeeggeoo

Compressed
sequence 1:

a,b,e,g,e,o
2,1,3,2,1,2

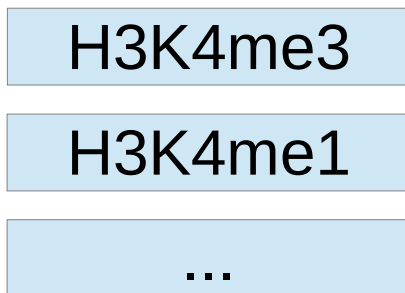
HMM

Compress

Align

(Local)
Matches

Epigenetic region 2:



Chromatin state
sequence 2:

aeeggoooo

Compressed
sequence 2:

a,e,g,o
1,3,2,4

Frequency of short combinations (15 top motifs of E003, chr1)

5	7	5	7	15	7	5	7	115
7	5	7	15	7	5	7	5	116
7	5	7	5	7	15	7	5	118
7	15	7	5	7	5	7	5	121
15	9	15	9	15	9	15	7	126
7	15	9	15	9	15	9	15	127
5	7	5	7	5	7	15	7	129
7	5	7	5	7	5	7	15	158
15	7	5	7	5	7	5	7	173
9	15	9	15	9	15	9	15	367
15	9	15	9	15	9	15	9	372
4	5	4	5	4	5	4	5	427
5	4	5	4	5	4	5	4	434
5	7	5	7	5	7	5	7	495
7	5	7	5	7	5	7	5	499

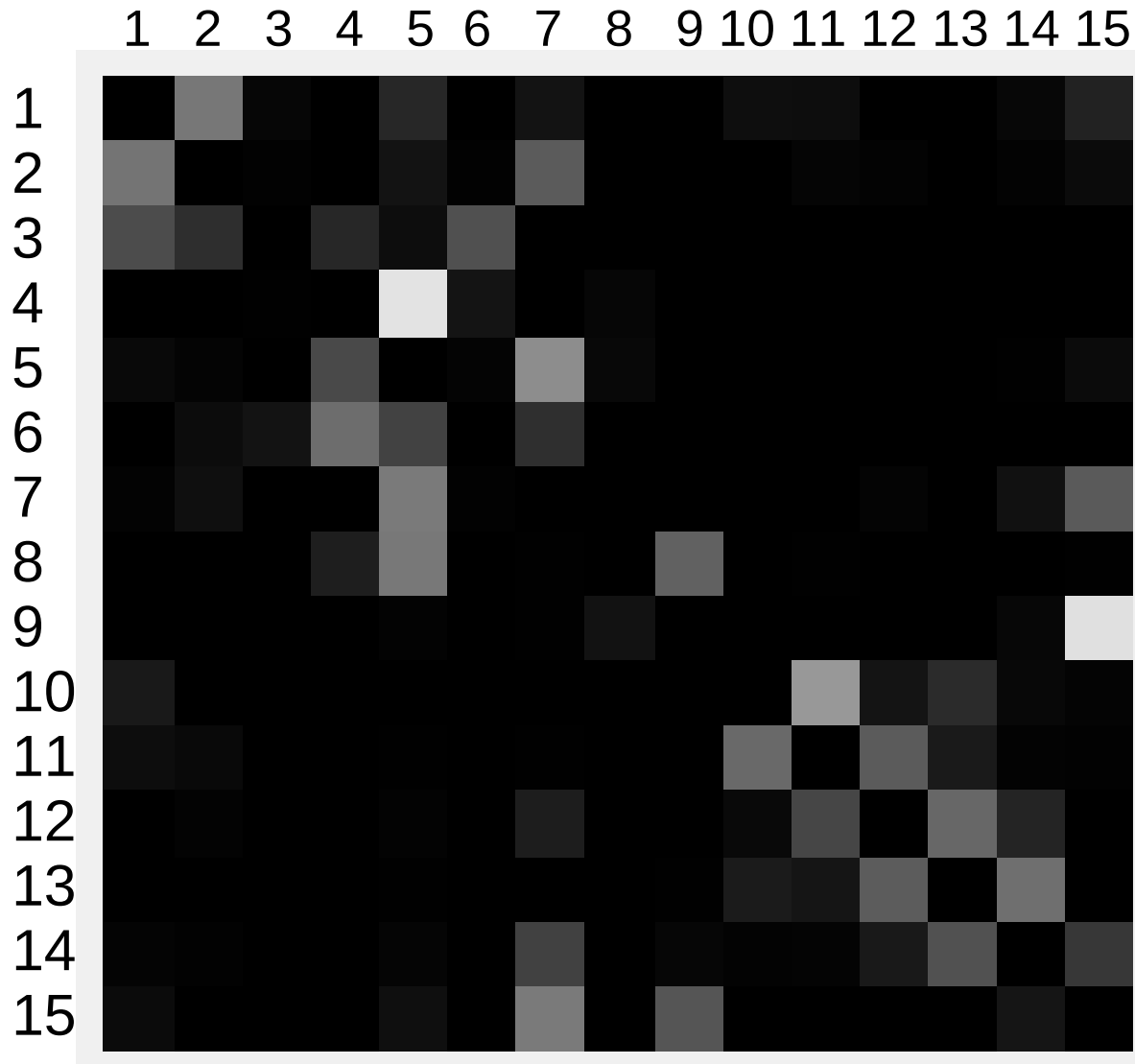
Motif frequency comparison between raw sequence and permuted sequence

5	7	2	1	2	13	11	7	2	5	7	5	7	15	7	5	7	115
5	4	7	5	15	7	5	12	2	7	5	7	15	7	5	7	5	116
5	4	7	5	15	5	2	15	2	7	5	7	5	7	15	7	5	118
5	2	7	5	12	7	9	5	2	7	15	7	5	7	5	7	5	121
5	1	7	12	5	7	15	5	2	15	9	15	9	15	9	15	7	126
4	15	9	7	13	15	7	15	2	7	15	9	15	9	15	9	15	127
4	15	7	5	7	15	7	5	2	5	7	5	7	5	7	15	7	129
4	7	15	7	5	15	7	15	2	7	5	7	5	7	5	7	15	158
4	7	14	7	10	5	15	7	2	15	7	5	7	5	7	5	7	173
4	7	9	5	7	15	7	5	2	9	15	9	15	9	15	9	15	367
2	15	14	7	14	5	7	15	2	15	9	15	9	15	9	15	9	372
2	7	5	7	4	7	15	9	2	4	5	4	5	4	5	4	5	427
1	15	7	5	4	7	15	5	2	5	4	5	4	5	4	5	4	434
1	7	15	7	5	7	13	4	2	5	7	5	7	5	7	5	7	495
15	5	15	7	5	15	5	15	3	7	5	7	5	7	5	7	5	499

Shuffle

Raw sequence

Frequency baseline: 1-gram randomization



Frequency baseline: 1-gram randomization (15 top motifs)

7	5	7	5	7	5	7	5	73	5	7	5	7	15	7	5	7	115
5	7	5	4	5	7	5	7	73	7	5	7	15	7	5	7	5	116
7	5	4	5	7	5	7	5	74	7	5	7	5	7	15	7	5	118
15	9	15	9	15	7	5	7	75	7	15	7	5	7	5	7	5	121
5	4	5	7	5	4	5	7	75	15	9	15	9	15	9	15	7	126
5	4	5	4	5	7	5	7	75	7	15	9	15	9	15	9	15	127
15	9	15	9	15	9	15	7	77	5	7	5	7	5	7	15	7	129
15	9	15	7	5	4	5	7	77	7	5	7	5	7	5	7	15	158
7	5	7	5	7	5	4	5	77	15	7	5	7	5	7	5	7	173
7	15	9	15	9	15	9	15	78	9	15	9	15	9	15	9	15	367
7	5	4	5	7	5	4	5	79	15	9	15	9	15	9	15	9	372
7	5	4	5	4	5	4	5	79	4	5	4	5	4	5	4	5	427
5	7	5	7	5	4	5	7	79	5	4	5	4	5	4	5	4	434
5	4	5	7	5	7	5	7	81	5	7	5	7	5	7	5	7	495
5	4	5	4	5	4	5	7	83	7	5	7	5	7	5	7	5	499

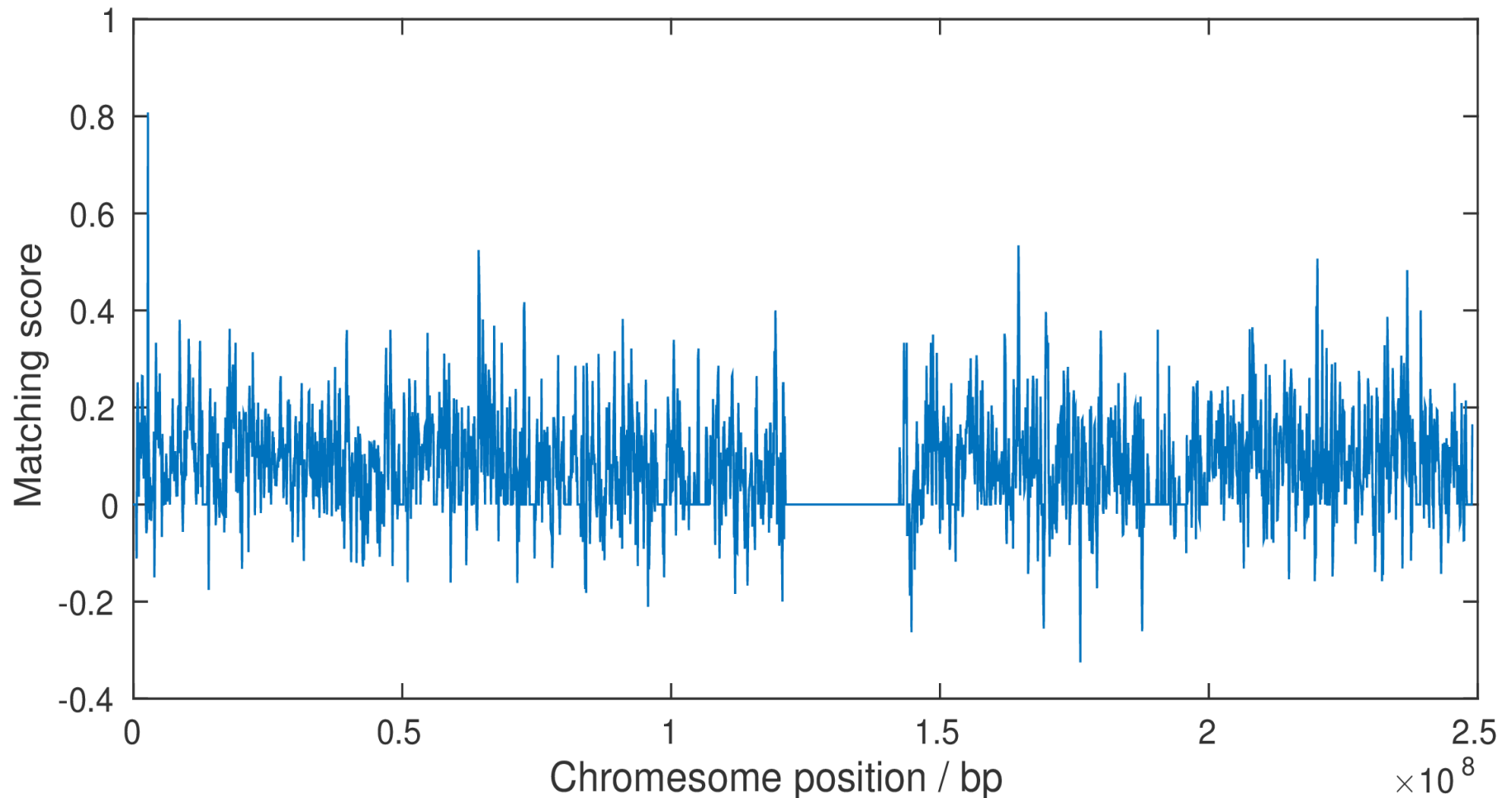
1-gram randomized sequence

Raw sequence

Fold change of top motifs

5	7	5	7	15	7	5	7	115	2.2308
7	5	7	15	7	5	7	5	116	2.0526
7	5	7	5	7	15	7	5	118	2.3800
7	15	7	5	7	5	7	5	121	1.8769
15	9	15	9	15	9	15	7	126	1.6282
7	15	9	15	9	15	9	15	127	1.6203
5	7	5	7	5	7	15	7	129	2.5490
7	5	7	5	7	5	7	15	158	3.0577
15	7	5	7	5	7	5	7	173	3.0000
9	15	9	15	9	15	9	15	367	8.3636
15	9	15	9	15	9	15	9	372	8.4773
4	5	4	5	4	5	4	5	427	10.1905
5	4	5	4	5	4	5	4	434	10.8750
5	7	5	7	5	7	5	7	495	7.4030
7	5	7	5	7	5	7	5	499	6.7568

Horizontal Alignment result



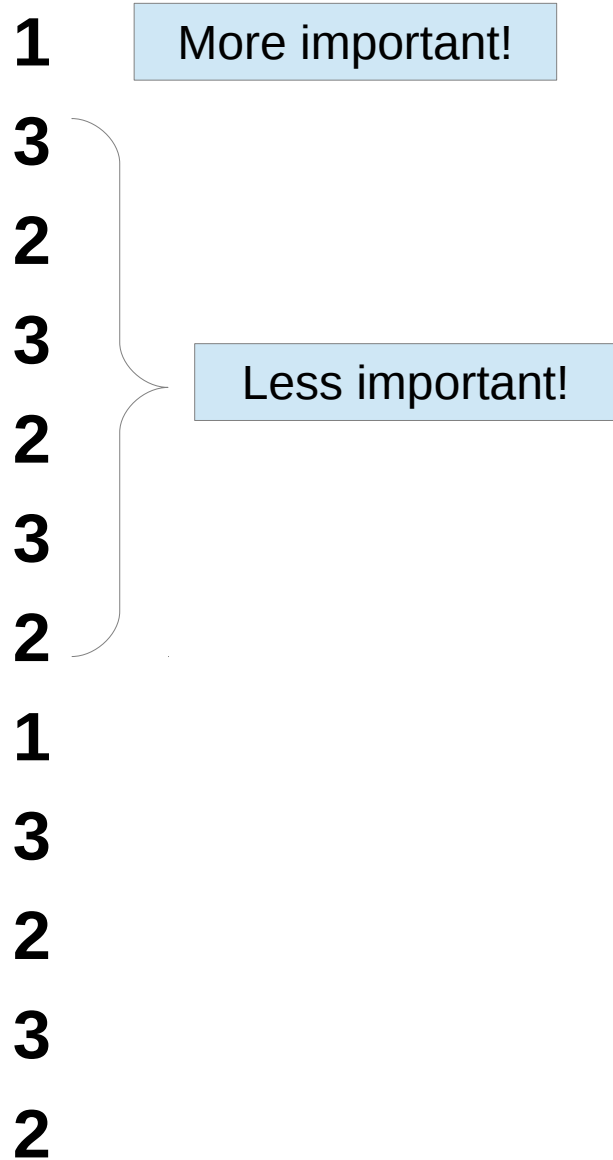
Score: ($\max_{\text{bio}} - \max_{\text{ran}}$), figure: E003,chr1

Horizontal Alignment Legend

- In this horizontal alignment example of ES cell from Roadmap database (E003). Chromatin states are predicted by ChromHMM first. Chromosome 1 was segmented into 300kb bins, adjacent bins have 200kb overlap. Each bin was then aligned to both true epigenome and fake epigenome. For each chromosome, gram-1 probability $P(s_{t+1}|s_t)$ was computed and used to generate corresponding fake chromosome. The final score of each bin is the best local matching score (except for itself) to true epigenome minus the best local matching score to fake epigenome. This figure indicate compare to fake epigenome, true epigenome has some regions share similar epigenetic pattern.

Matching Examples

Attention Score of Epigenetic State Sequences



- We don't want to mismatch more important positions in our algorithm.
- To achieve that, each position is assigned by an “attention score”, higher attention score indicates this position is more important.
- Match of important position pairs will get higher reward, and mismatch or gap in this position will have higher penalty. The reward and penalty are functions of attention score(s).

Matching Score and Gap Score

- AS: Attention Score
- MS(i,j): Matching score if seq1(i) pairs with seq2(j).
- GS(i): Gap penalty of seq1(i)

$$MS(i, j) = \begin{cases} AS_1(i) + AS_2(j) & seq_1(i) = seq_2(j) \\ -\epsilon * (AS_1(i) + AS_2(j)) & seq_1(i) \neq seq_2(j) \end{cases}$$

$$GS_s(i) = \epsilon * AS_s(i) \quad s \in \{1, 2\}$$

Attention Score 1: log frequency

- High frequency states should have smaller attention score because they are too common to carry specific information.

$$AS_s^{lf}(i) = -\log(\text{frequency}(\text{seq}_s(i)))$$

Attention score 2: unpredictability

High predictability, low specific information, low attention score

ededede

Low predictability, high specific information, high attention score

ededade

Attention score 2: unpredictability

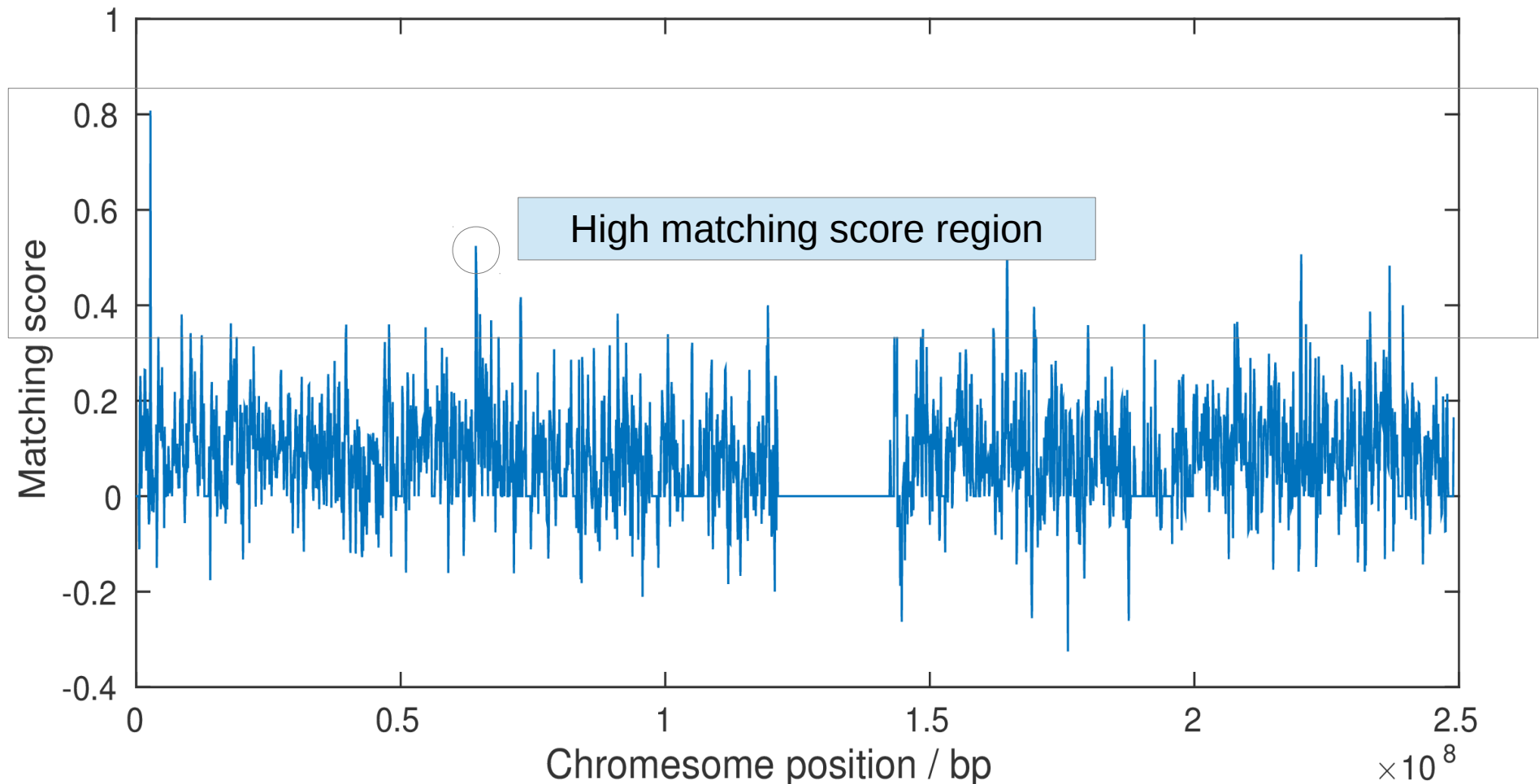
$$AS_s^{up,w}(i) = GramScore_s^w(i) * LocalRankScore_s^w(i) * LengthScore_s(i)$$

In $seq_s([i-w, i+w])$, if neither of pattern $seq_s(i-1)seq_s(i)$ nor $seq_s(i)seq_s(i+1)$ occurs elsewhere, $GramScore_s^w(i) = 1$, if both patterns occurs elsewhere, $GramScore_s^w(i) = 4$, otherwise $GramScore_s^w(i) = 2$.

For $seq_s([i-w, i+w])$, we sort the chromatin states occurred in this region by their frequency, $LocalRankScore_s^w(i) = 1 + \frac{rank(seq_s(i))-1}{|\{States \in seq_s([i-w, i+w])\}| - 1}$

If $seqlength_s(i) = 1$, $LengthScore_s(i) = 0.5$, otherwise $LengthScore_s(i) = 1$.

4. 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.

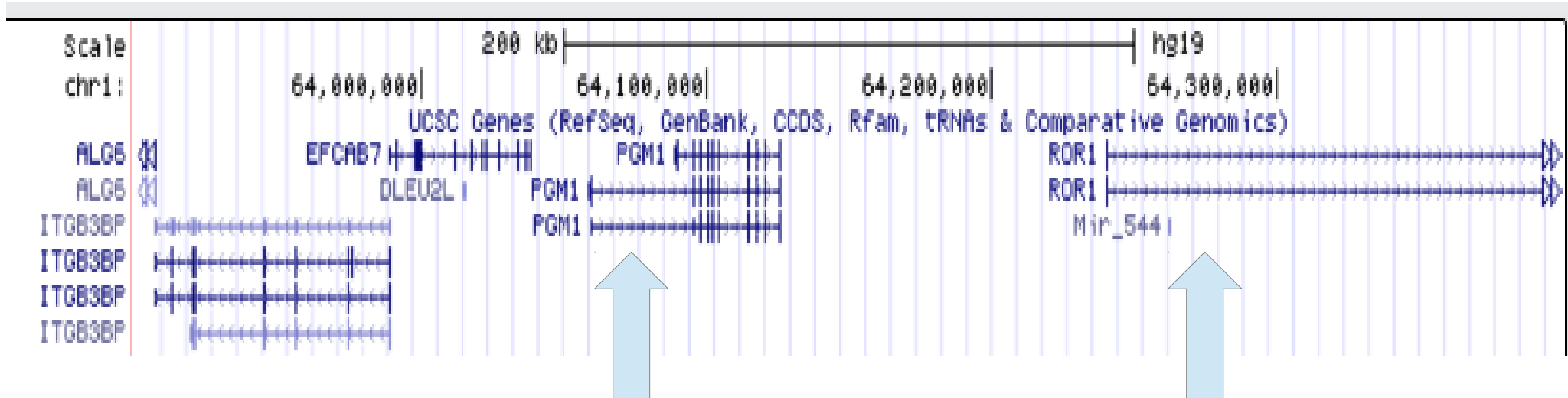
The example shown above has three ab-eded motifs.

ab-gege Motif

- This motif 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

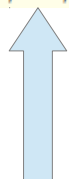
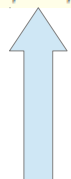


Description: Homo sapiens phosphoglucomutase 1 (PGM1), transcript variant 3, mRNA.

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. Alternatively 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

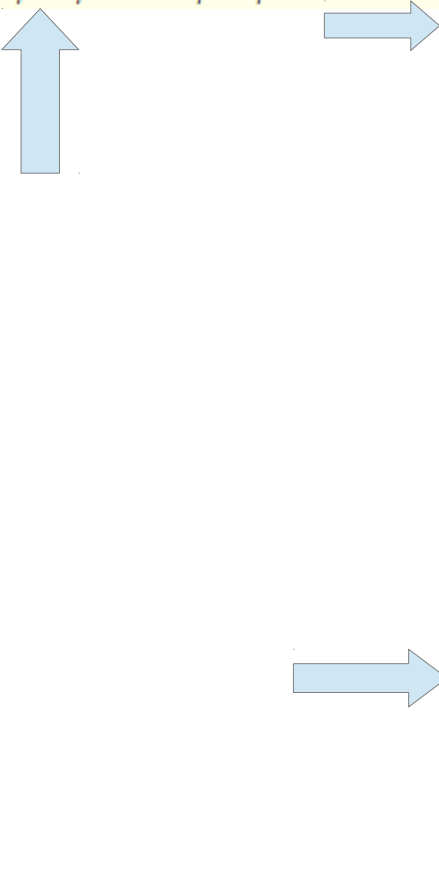
Human Gene PBX1 (uc010pku.2) Description and Page Index

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