

Characteristics of genome with extended alphabet

Biopython course project

Kolosov Nikita
Boris Shpak
Katerina Gibitova

2019

Task

- Given the assembled reads of unknown origin with unknown “nucleotides” extract some knowledge about what we are dealing with.
- Characterise sequence from standpoint of string sequence analysis.

After assembly: 3 “chromosomes”

1)	300454	}	1
2)	300454		
3)	402937	}	2
4)	402937		
5)	300001	}	3
6)	300001		

Forward:

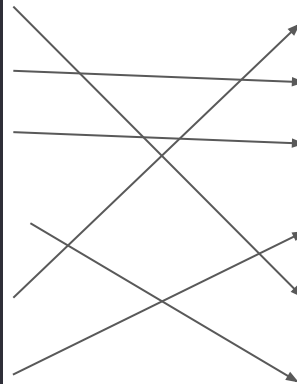
```
{ 'S': 0.1547,  
  'T': 0.1778,  
  'O': 0.1524,  
  'I': 0.1611,  
  'B': 0.1761,  
  'N': 0.1778 }
```

Reversed:

```
{ 'S': 0.1761,  
  'T': 0.1778,  
  'O': 0.1524,  
  'I': 0.1778,  
  'B': 0.1547,  
  'N': 0.1611 }
```

Forward:

```
{'S': 0.1547,  
'T': 0.1778,  
'O': 0.1524,  
'I': 0.1611,  
'B': 0.1761,  
'N': 0.1778}
```



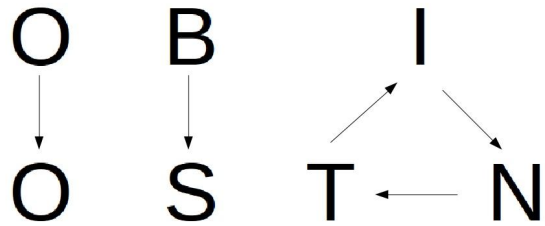
Reversed:

```
{'S': 0.1761,  
'T': 0.1778,  
'O': 0.1524,  
'I': 0.1778,  
'B': 0.1547,  
'N': 0.1611}
```

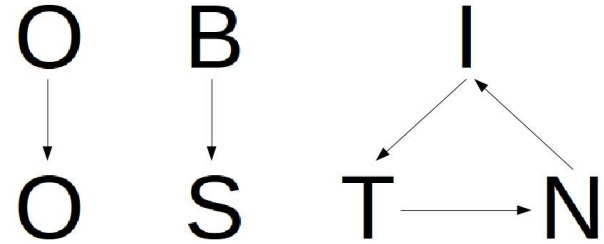
T O B S I N
↓ ↓ ↓ ↓ ↓ ↓
T O S B N I

OR

Forward

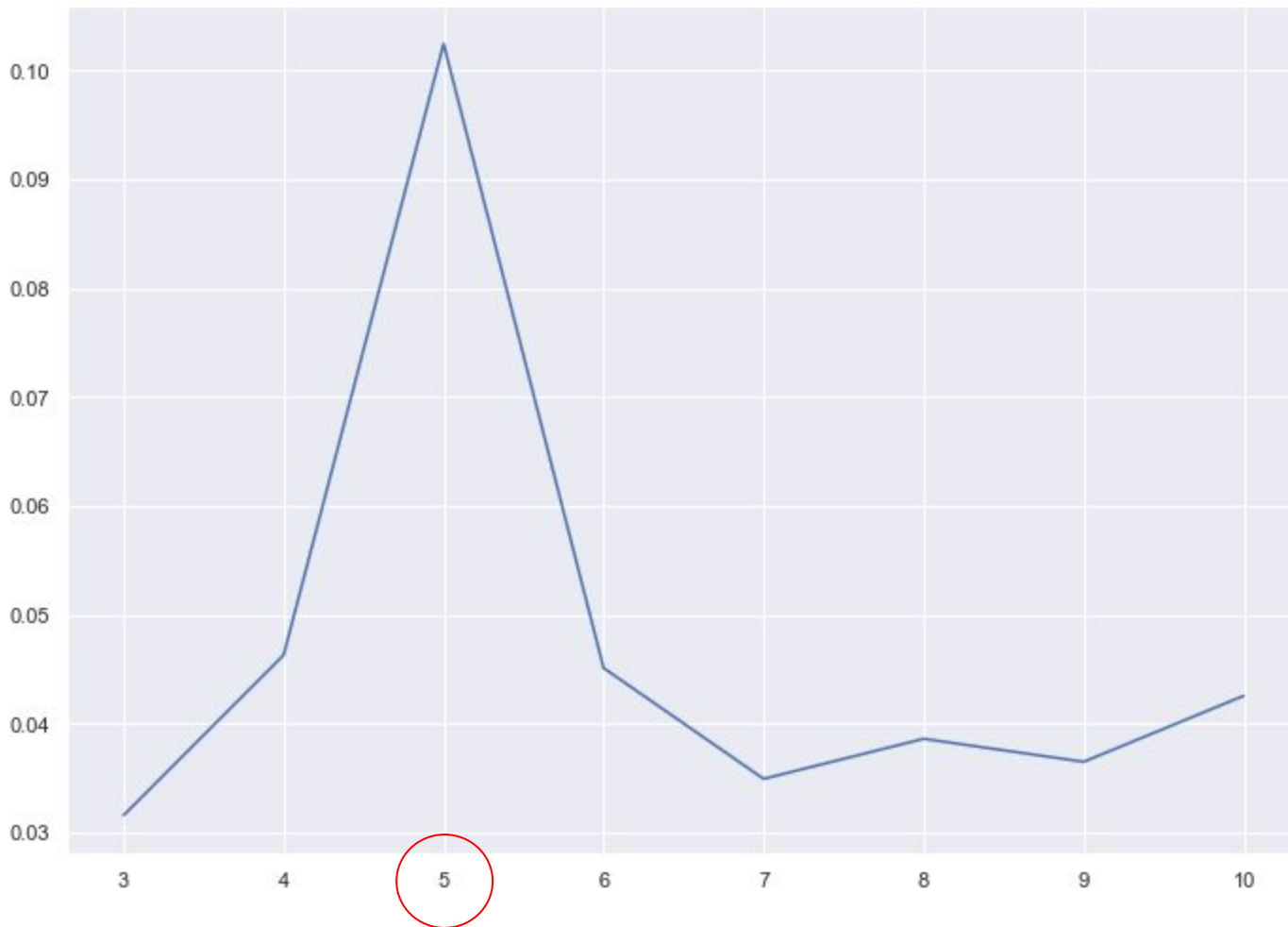


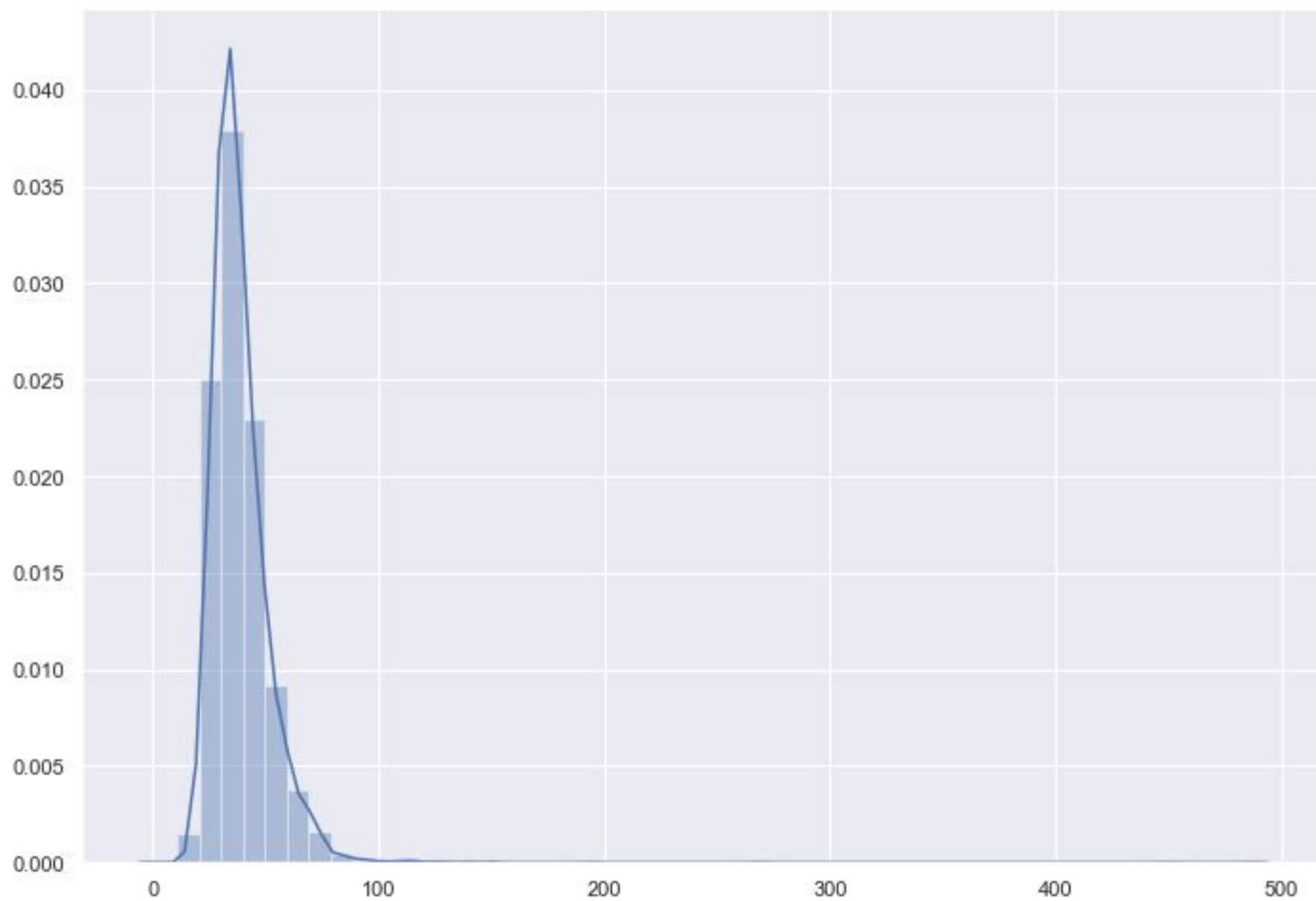
Reversed

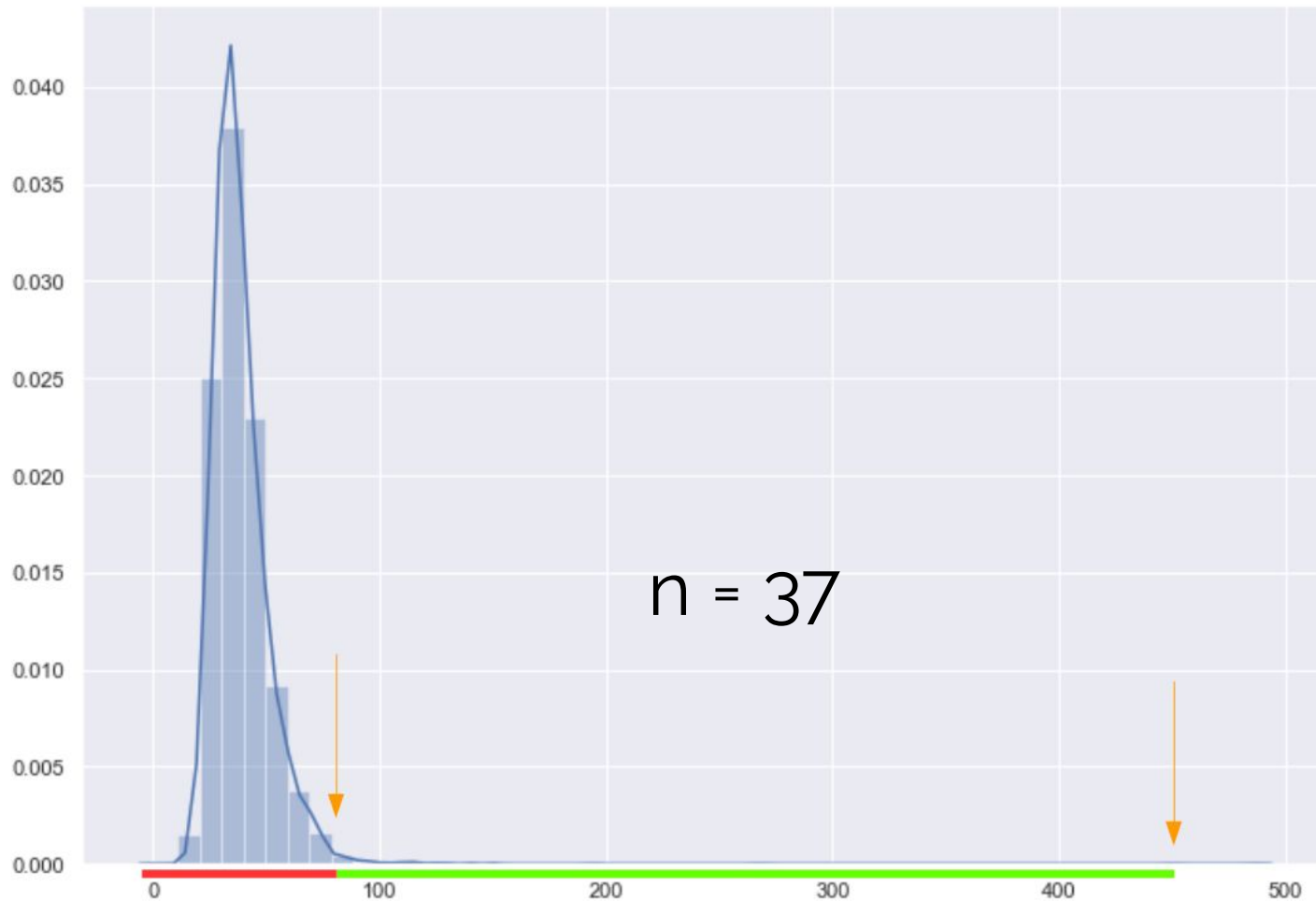


Chromosome **1**

$k = 5$





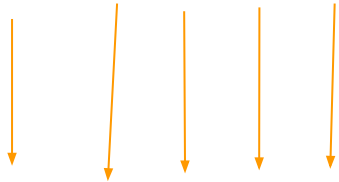


```
[('III00', 88),
 ('SISII', 88),
 ('TOOSN', 88),
 ('SBNIN', 89),
 ('IIISI', 90),
 ('NBNTT', 91),
 ('OSNTT', 91),
 ('NINIO', 93),
 ('ISIII', 94),
 ('TBTTT', 95),
 ('OTTTT', 95),
 ('IIIIIS', 95),
 ('ITTTT', 96),
 ('TTOIS', 99),
 ('IIIII', 99),
 ('OSNTN', 102),
 ('BNTTT', 106),
 ('STTTT', 108),
 ('NINIS', 108),
 ('NINIT', 112),
 ('TTOIT', 112),
 ('TTOIB', 113),
 ('IBNIN', 115),
 ('TTTTT', 115),
 ('TTOIN', 116),
 ('NINII', 121),
 ('NININ', 126),
 ('NINIB', 128),
 ('BTTTO', 129),
 ('BBNIN', 139),
 ('NBNIN', 140),
 ('NTTTO', 149),
 ('TBNIN', 151),
 ('TTOTS', 194),
 ('OOSNT', 269),
 ('TTTOI', 451)]
```

```
[('TB00I', 88),  
 ('OONNN', 88),  
 ('NNBNB', 88),  
 ('TNTSB', 89),  
 ('NBNNN', 90),  
 ('IITBO', 91),  
 ('IITST', 91),  
 ('ONTNT', 93),  
 ('NNNBN', 94),  
 ('OIIIO', 95),  
 ('IIISI', 95),  
 ('BNNNN', 95),  
 ('OIIIN', 96),  
 ('NNNNN', 99),  
 ('BNOII', 99),  
 ('TITBO', 102),  
 ('IIITS', 106),  
 ('BNTNT', 108),  
 ('OIIIB', 108),  
 ('INOII', 112),  
 ('INTNT', 112),  
 ('SNOII', 113),  
 ('OIIII', 115),  
 ('TNTSN', 115),  
 ('TNOII', 116),  
 ('NNTNT', 121),  
 ('TNTNT', 126),  
 ('SNTNT', 128),  
 ('OIIIS', 129),  
 ('TNTSS', 139),  
 ('TNTST', 140),  
 ('OIIIT', 149),  
 ('TNTSI', 151),  
 ('BIOII', 194),  
 ('ITB00', 269),  
 ('NOIII', 451)]
```

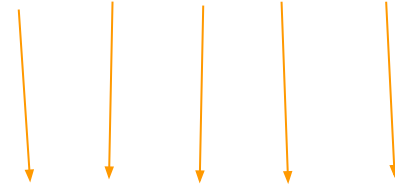
```
[('IIIOO', 88),  
 ('SISII', 88),  
 ('TOOSN', 88),  
 ('SBNIN', 89),  
 ('IIISI', 90),  
 ('NBNTT', 91),  
 ('OSNTT', 91),  
 ('NINIO', 93),  
 ('ISIII', 94),  
 ('TBT TT', 95),  
 ('OTTT0', 95),  
 ('IIII S', 95),  
 ('ITTT0', 96),  
 ('TTOIS', 99),  
 ('IIIII', 99),  
 ('OSNTN', 102),  
 ('BNTTT', 106),  
 ('STTT0', 108),  
 ('NINIS', 108),  
 ('NINIT', 112),  
 ('TTOIT', 112),  
 ('TTOIB', 113),  
 ('IBNIN', 115),  
 ('TTTT0', 115),  
 ('TTOIN', 116),  
 ('NINII', 121),  
 ('NININ', 126),  
 ('NINIB', 128),  
 ('BTTT0', 129),  
 ('BBNIN', 139),  
 ('NBNIN', 140),  
 ('NTTT0', 149),  
 ('TBNIN', 151),  
 ('TTOTS', 194),  
 ('OOSNT', 269),  
 ('TTTOI', 451)]
```

5'-NOI I I-3'

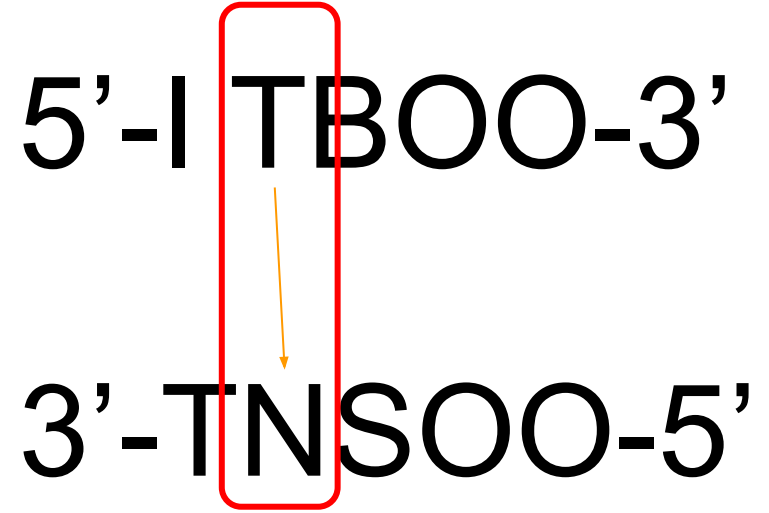
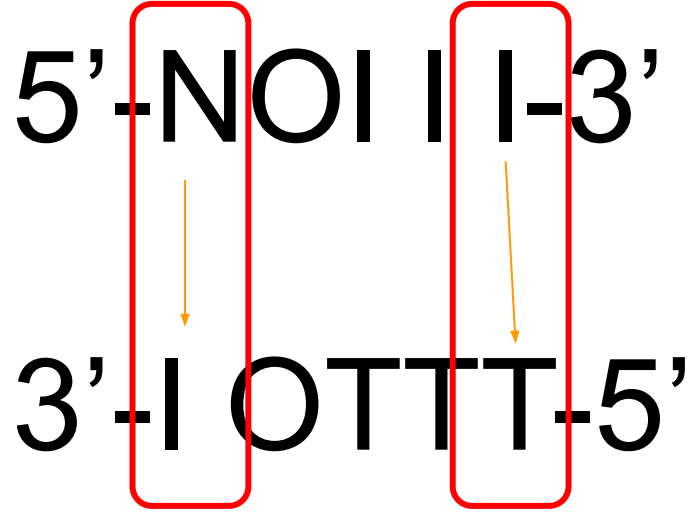


3'-I OTTT-5'

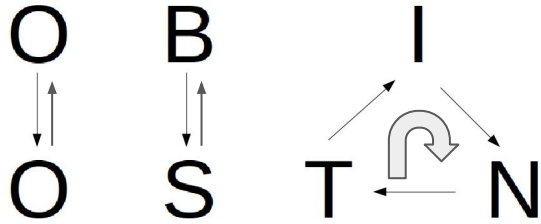
5'-I TBOO-3'



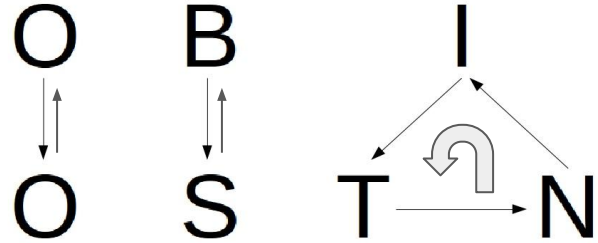
3'-TNSOO-5'



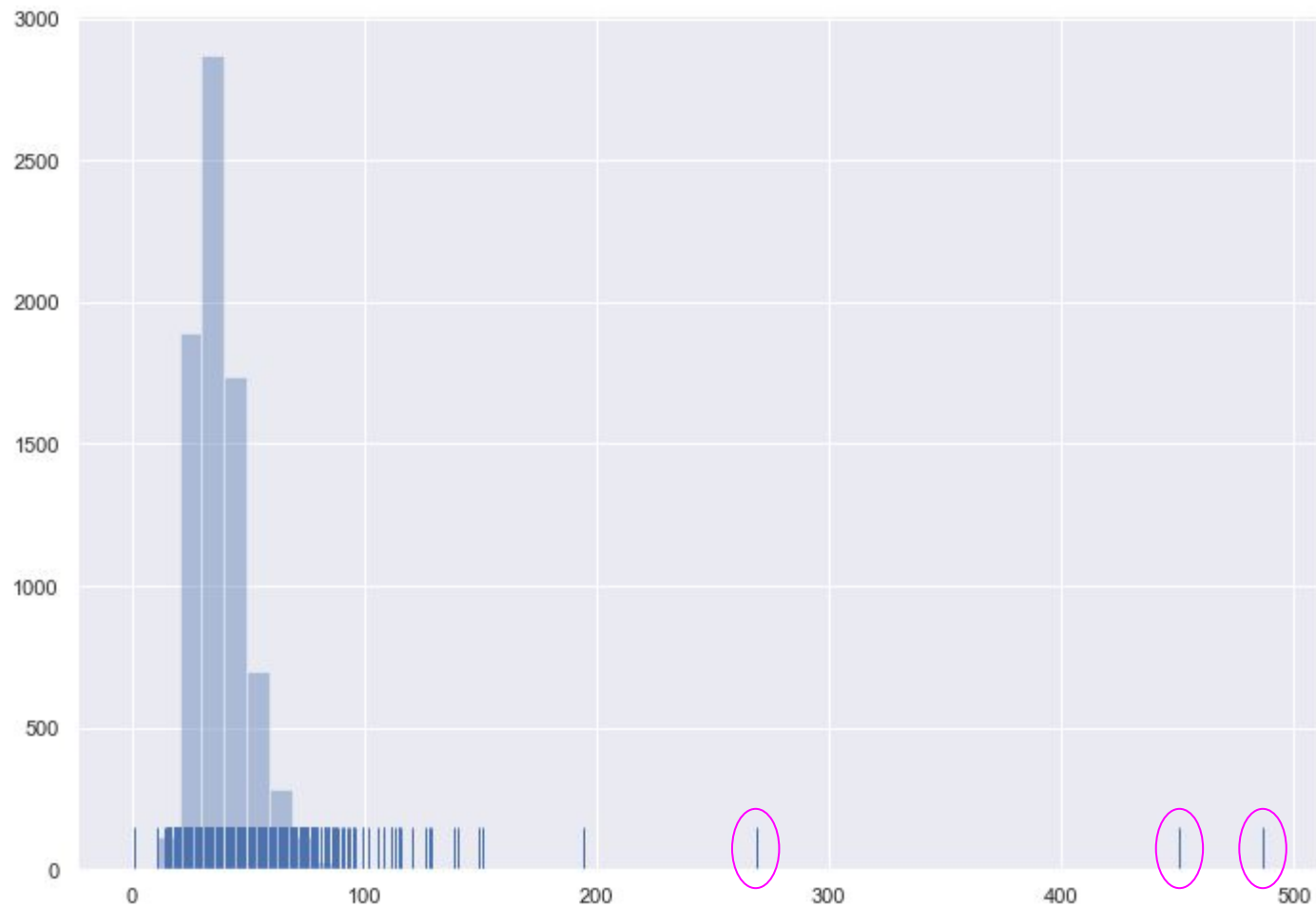
Forward



Reversed



Example: T T T O I B S N N I T S
I I I O N S B T T N I B

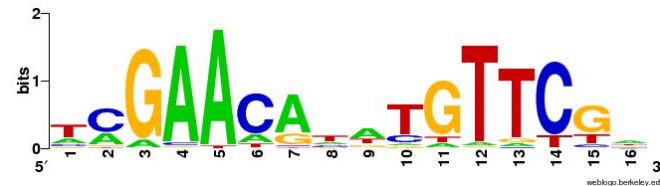
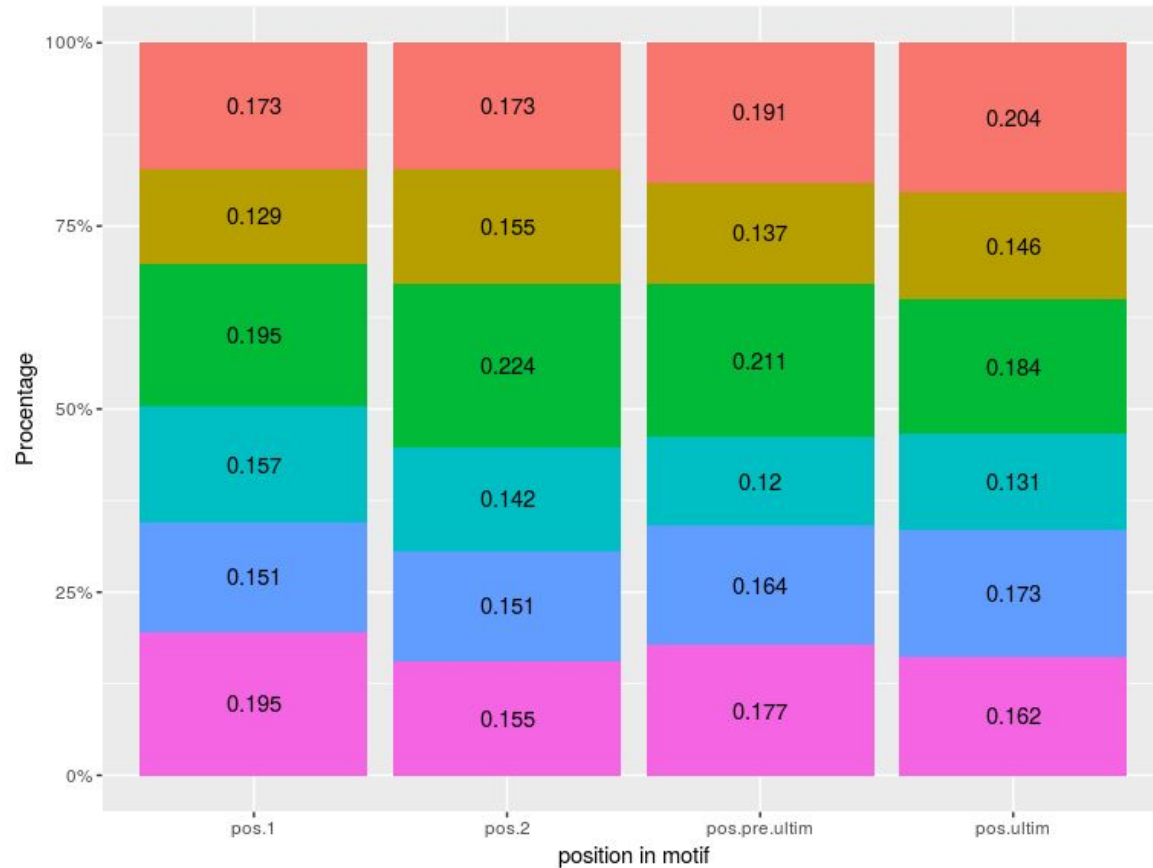


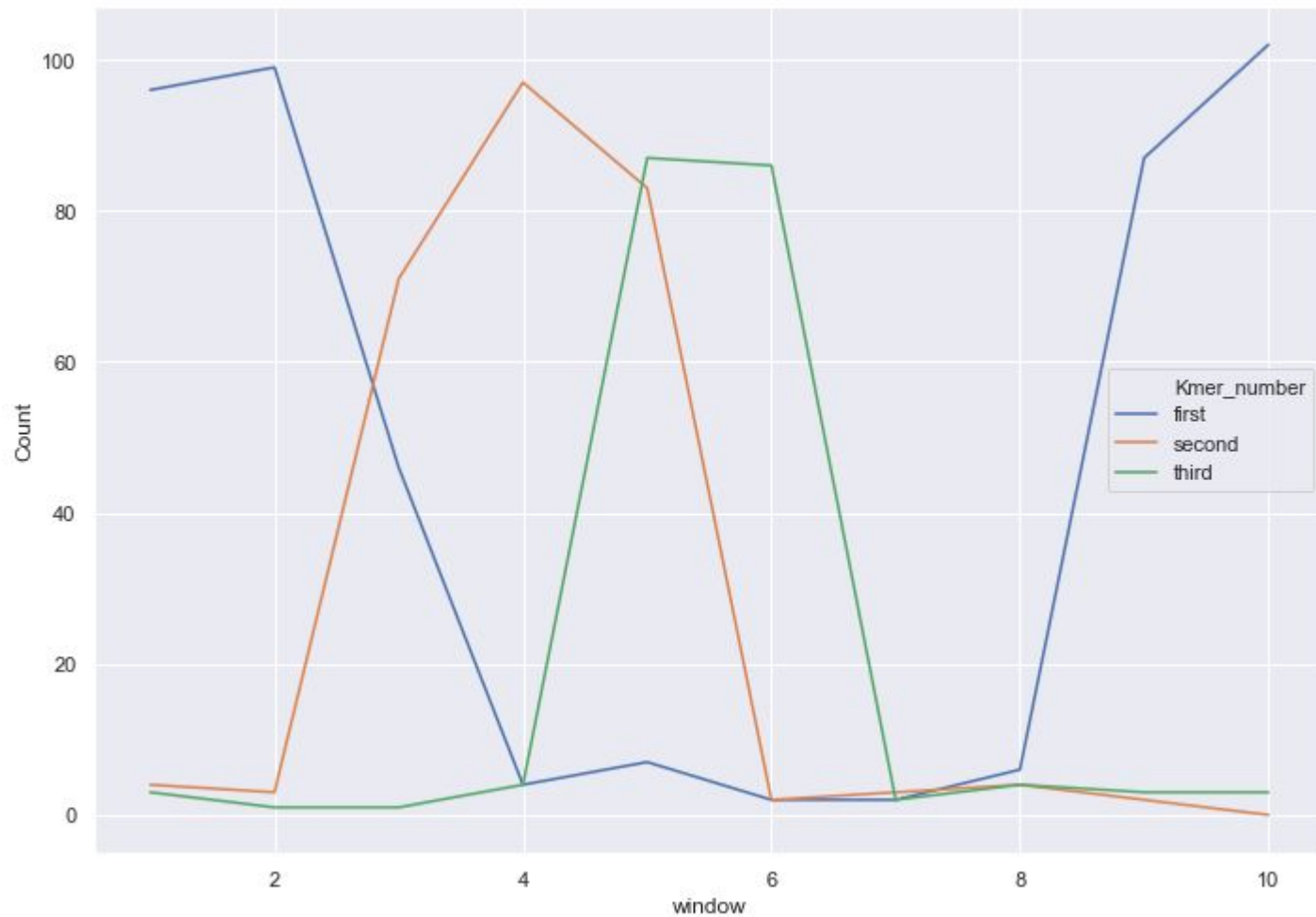
```
[('IIIOO', 88),
('SISII', 88),
('TOOSN', 88),
('SBNIN', 89),
('IIISI', 90),
('NBNTT', 91),
('OSNTT', 91),
('NINIO', 93),
('ISIII', 94),
('TBTIT', 95),
('OTTTT', 95),
('IIIIIS', 95),
('ITTTT', 96),
('TTOIS', 99),
('IIIII', 99),
('OSNTN', 102),
('BNTTT', 106),
('STTTT', 108),
('NINIS', 108),
('NINIT', 112),
('TTOIT', 112),
('TTOIB', 113),
('IBNIN', 115),
('TTTTT', 115),
('TTOIN', 116),
('NINII', 121),
('NININ', 126),
('NINIB', 128),
('BTTOO', 129),
('BBNIN', 139),
('NBNIN', 140),
('NTTTO', 149),
('TBNIN', 151),
('TTOTS', 194),
('OOSNT', 269),
('TTTOI', 451)]
```



```
[('III00', 9.710847946399747e-05),
 ('SISII', 0.00010006588585499018),
 ('IIISI', 0.00010420802794588949),
 ('IIIS', 0.00010420802794588949),
 ('ISIII', 0.0001042080279458895),
 ('IIIII', 0.00010852163047962208),
 → ('00SNT', 0.0001136123317424768),
 ('T00SN', 0.00011361233174247681),
 ('TTOIS', 0.0001200741112034278),
 ('NINIO', 0.00012510534214676386),
 ('NINIS', 0.0001269960754560493),
 ('OTTT0', 0.00013054908316174825),
 ('NINII', 0.00013225296980151914),
 ('STTT0', 0.00013252208843710274),
 ('TTOTS', 0.00013252208843710274),
 ('OSNTT', 0.0001325543370827858),
 ('OSNTN', 0.00013258659357602969),
 ('TTOIB', 0.0001367212190475519),
 ('ITTT0', 0.00013800772738187435),
 ('TTOIT', 0.00013800772738187435),
 → ('TTTOI', 0.00013800772738187435),
 ('TTOIN', 0.0001380413109327703),
 ('SBNIN', 0.0001388550594462431),
 ('NINIB', 0.00014460284633037763),
 ('IBNIN', 0.00014460284633037766),
 ('NINIT', 0.00014596351856740675),
 ('NININ', 0.00014599903812379504),
 ('BTTT0', 0.00015089498727291785),
 ('TTTT0', 0.00015231486679188688),
 ('NTTT0', 0.00015235193191989204),
 ('BBNIN', 0.0001581059631267851),
 ('TBNIN', 0.00015959369590656626),
 ('NBNIN', 0.00015963253230443205),
 ('TBT TT', 0.00017605293985528564),
 ('BNTTT', 0.00017609578153508274),
 ('NBNTT', 0.00017613863364020712)]
```

pLOGO for 'TTTOI':





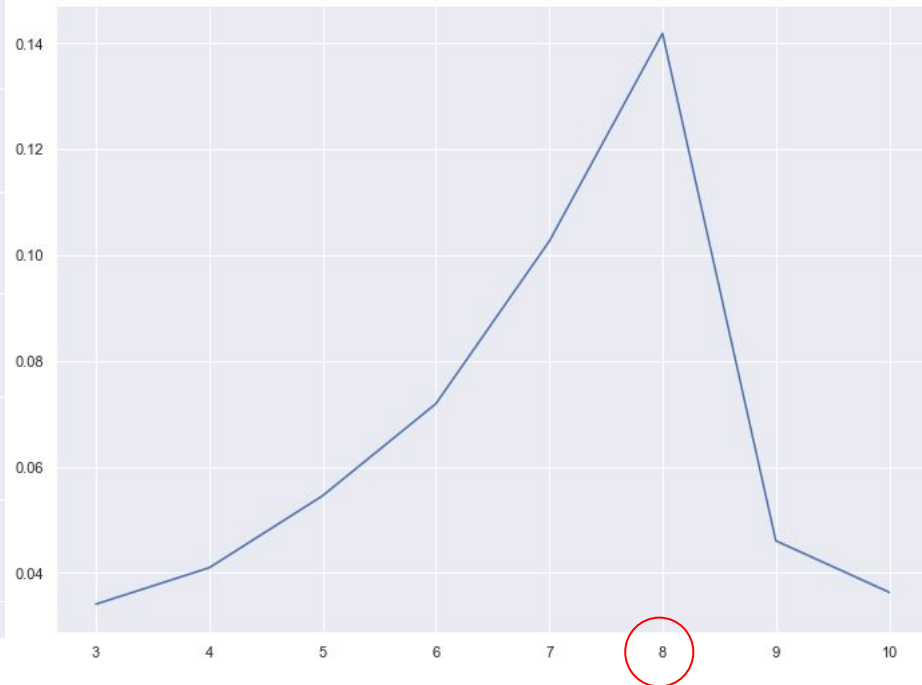
Chromosome **2** and **3**

2-nd chromosome



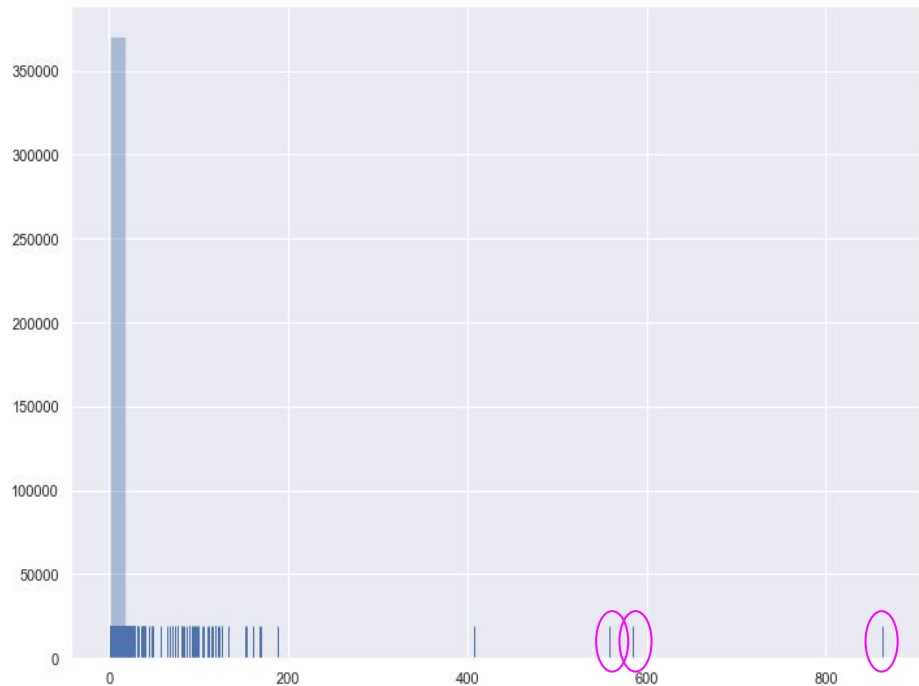
$k = 9$

3-d chromosome

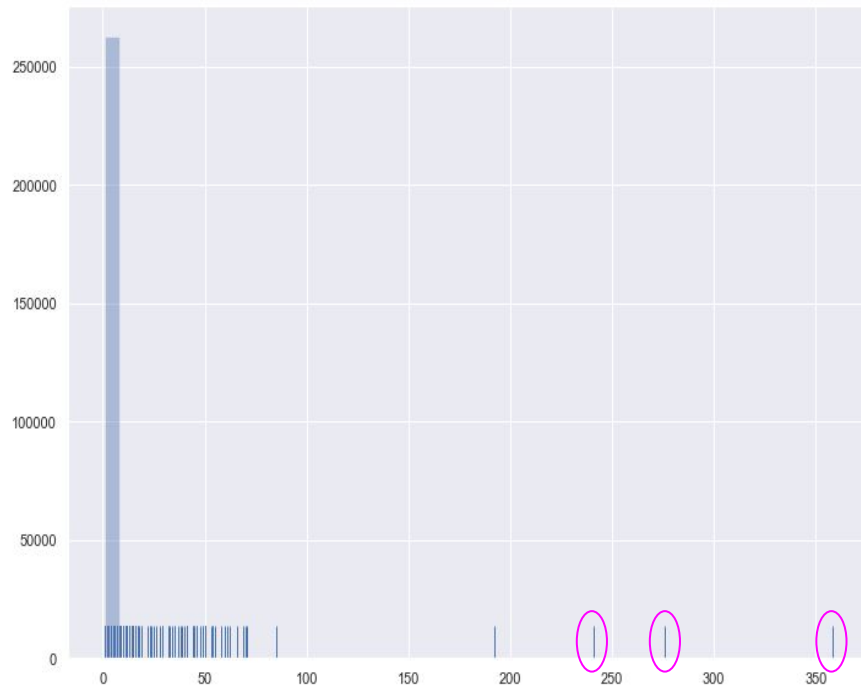


$k = 8$

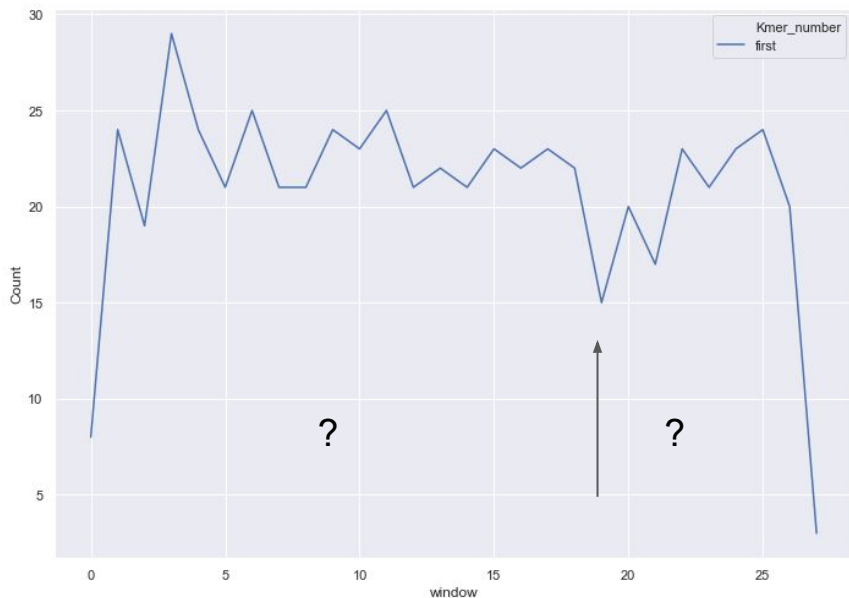
2-nd chromosome



3-d chromosome

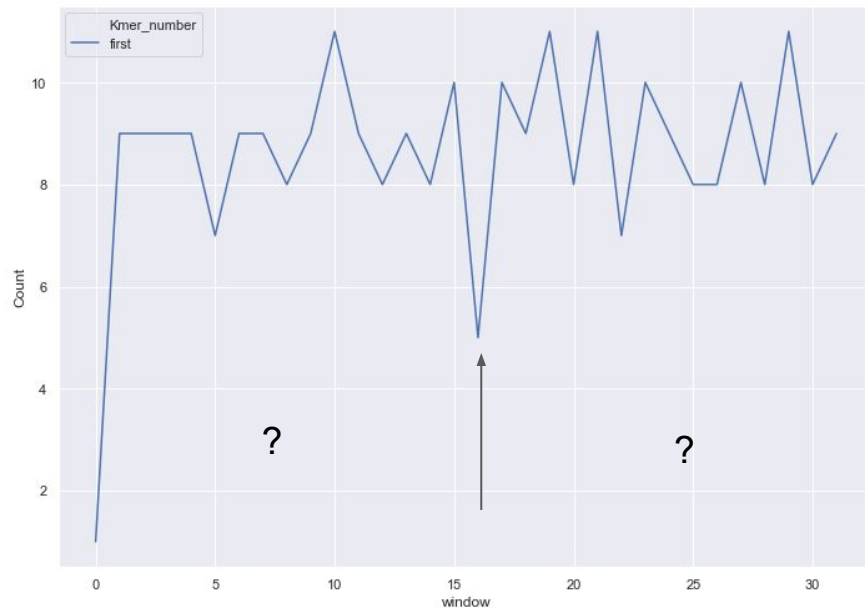


2-nd chromosome



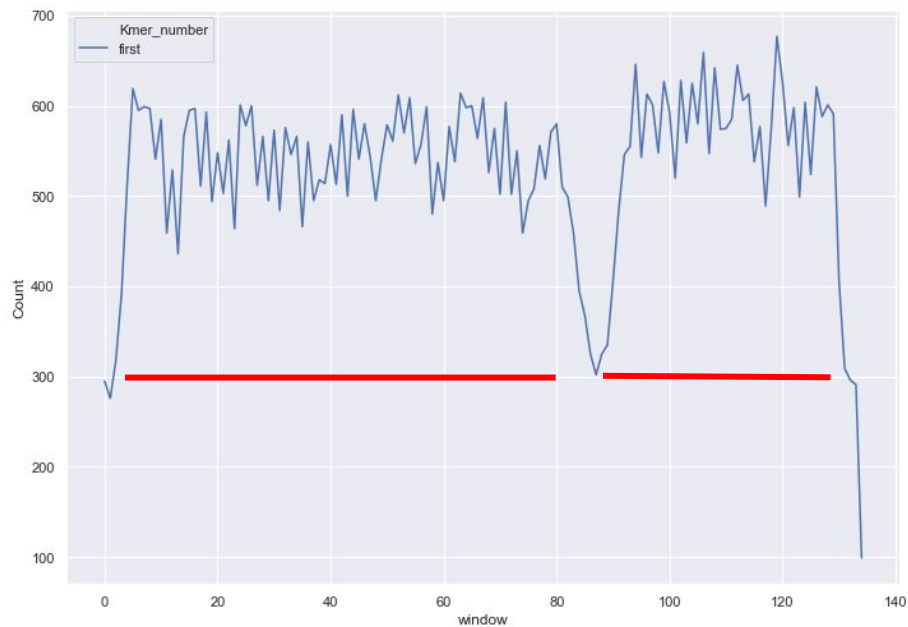
The most
frequent kmer: 'OOOTIOOOI'

3-d chromosome



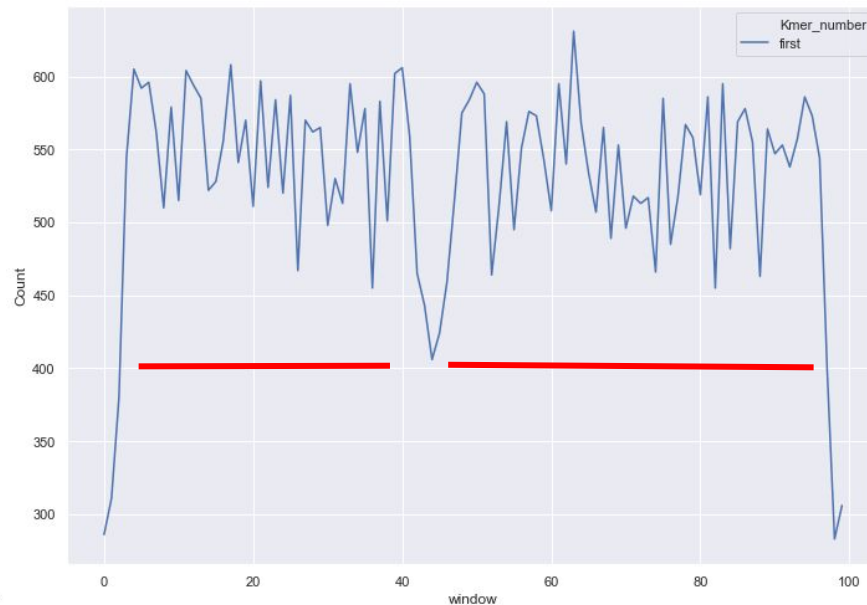
The most
frequent kmer: 'OBITNNST'

2-nd chromosome



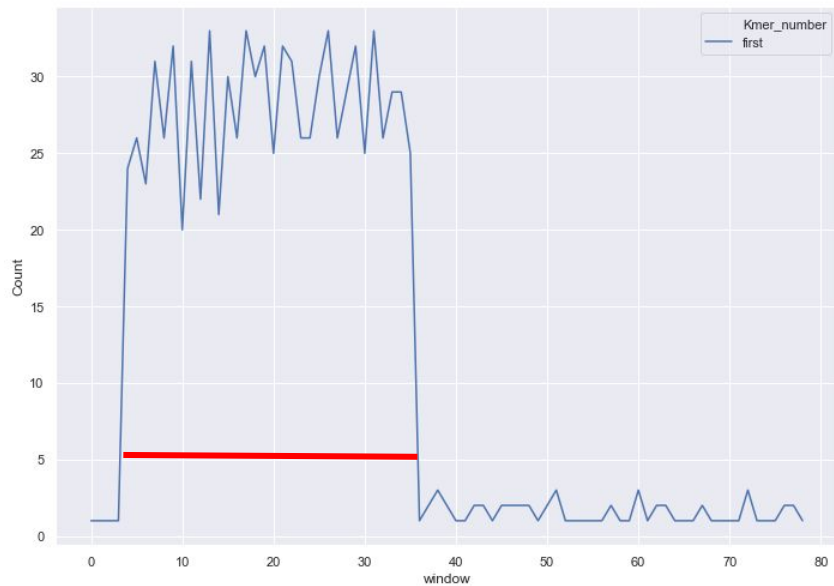
'Tl'

3-d chromosome



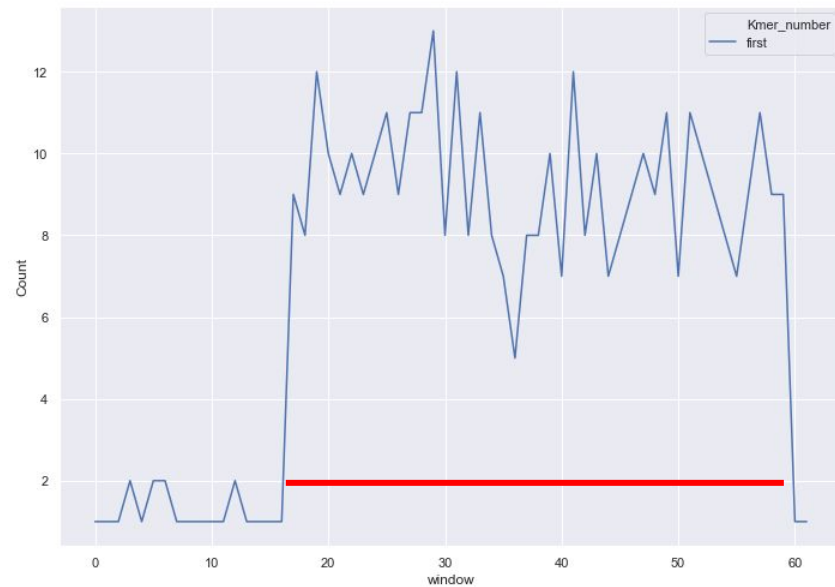
'T'

2-nd chromosome



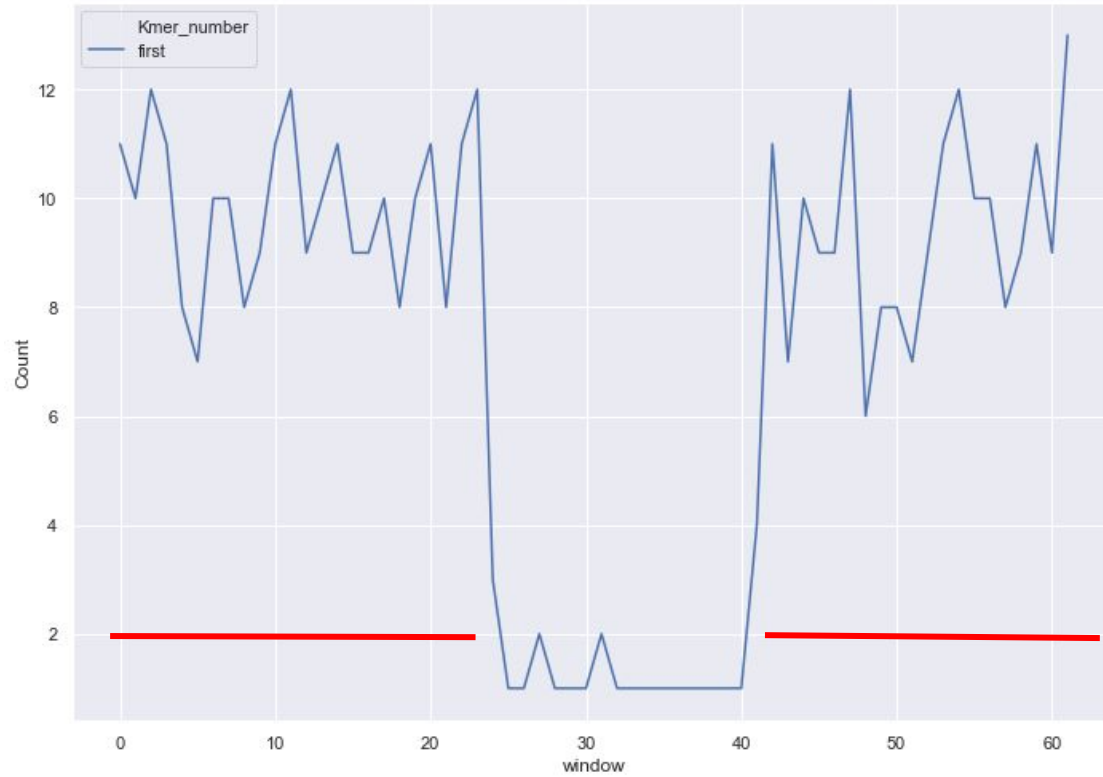
'TNNTN'

3-d chromosome



'NSTNS'

1-st chromosome



'TTTOI'

alignment?

- This fragments align better than random one
- But I do not have stats...

(527, 564)

```
TB-B-OT--BISIN--T-TOIBO-BONBIBNTTTT-BTT-TN-OISSBNTIITONOBOTBTBSIINSSNBSIISN-BBS-BTISTTS-N--BT--TBTT-T--B-NOSONBSNIO
INNINITSBIBOSNO-NBTI-TTSTO---ITTSNT-NITTSB-IBNSBB-N--IIT-T--STNOTTBSNB--TB--T---S--NO-BTB-T--N-IONN--NTT-STS-B--N-S
B-SBITIOONTBBN-N--NS-N--BSTN-----TTSIBBO-BI--BOIOIBT-----I-NIB-O-T-I--N-NBT-BNT-NIOIN--B-T-TSTNB--TBBNOBIIT-BN-TNNS
NBSBIBTTBBBIBNOBBNNNBONI-BN-TSSIONBNNN-TOBSS-ITB-I----TSTBINST-ONBSSTB-IIB-TTNIIBI----TOSTSSIOBOSNSTTIB-N-TNOS---TS-
INNOTBTONOIBNTISNB--TT-SBTT-SB-N--OBOSTB-BIONTI-N-BN-ST-IIINN---ITNSST-ONNNBB-STNN--BNNN-NTNOINIS--BB--BIOO-OSSOTN
TTNBTSOTOBOSBOOT-NBOTT-OI-ITBSBOSTSBIBIT--BSONOSI--B-TSOONT-TO---OOT-TN--TBNNSTTTTI--T-T-ITNNINSOOSIBIIISSONS-IBBBI
ITTOON-IOBO--T-----B-N-----BT-IINTO-BTOONNBS---I-BTN---B--NOSBB-IOOSBB--I--SN--NTB---NTN-BS---B-OIB-TSBB---NT-
TN-----NB--T-SI--IN--BN-TNNBNNTIT-SOTIOBBBTTO-----I-----OT-----SB-OBQB-STNNBNOBB-----NTBBNBS---NO--OBIBNTTT
-TINBTS---N--I--NN---TI--S-----BOT--INNOBBIT-TS-----TB--TT-B-SNBTN-I--BNN-I--IBOSNOTN--N-B---ONTB-S-O-IBBOT
NBNTTBN-B-B-T-OO-----B-----NSOIT-BBNSTIITTS--BNBOTIOOITOTO---N-O-TNBTO-BNNOOTI-INSOTO-B--O---S-IOOSNINIOO--
---ONO-STSTNNNTSNNNTTTTNN--I--OBBS-SNB-O---O-BB-IBOTTNBNOSN-S--BTT-ISNO--IIT-----BOT-ONI-OTITI-OT-T--OSTN--NBT-NBI
---ITNO-IOOBO-ITTNI-NSO--B--I--B-NTIOONN-IONNINTBT---OB-ST-BSSTTB---OBB--O--B--O-BN---O-T-',
'-----IIONSSOBBO---O--NO-SN--OIOBTBONSTNSINTOBNOS-NSNI-SI--N--INSOBN-BI-SBBT-B-O---N-N-BSSIBIINBBINBTNTITSNTSONIS
TBOBSOTTBSIS-NBSTNT-I--SB-N-I-N--TTNB--BTNBO-SS-N---ON--O--TB-----B---S-TBB-NB---TTSTNINB-SITB--ITSIBIN-S--B-NIO
--NIN-----O-N-SNB--BTTS--NBBI---N-SN--T-BBIB-S--SNTTII-NTOIS--O---S-BOST-NNTIIOSISNONBT-NTOBNSIO--OIN--BS-STBBONTS
-NS---I--N--BNBNSINSINTOBS-SSOONTT---BOTB-SOBO-O---SOONNISNI-TOOTSIBTNSN-TOB-TSN---NNNBBTNTST-BONT-B-OB---OBNIT--S
N-SBI--T-----O-----BO-ISBNT-S--B---ST-B--BI-BTIIIOOT-T-IN--IO---ST-OII-NTTN--NNNNT--T--I---N-TT-BONSTN-SNBSTSB
I---TB---IB-T-S-BIBTTISB--OSBTNITTO-O---OB-O---SNOB-BS-SIII--BOBI--S-BO--N-BOS---ISB---SN-N-I-I-TNBBNNBI-OIO--O--
```

diqqətinizə görə
təşəkkür edirik !