启动子的分析和预测

一、软硬件平台:

1.1 硬件平台:MacBook Pro

1.2 系统平台: macOS 10.13.1

1.3 软件平台: R3.3.2、python3.5.0

1.4 数据库资源

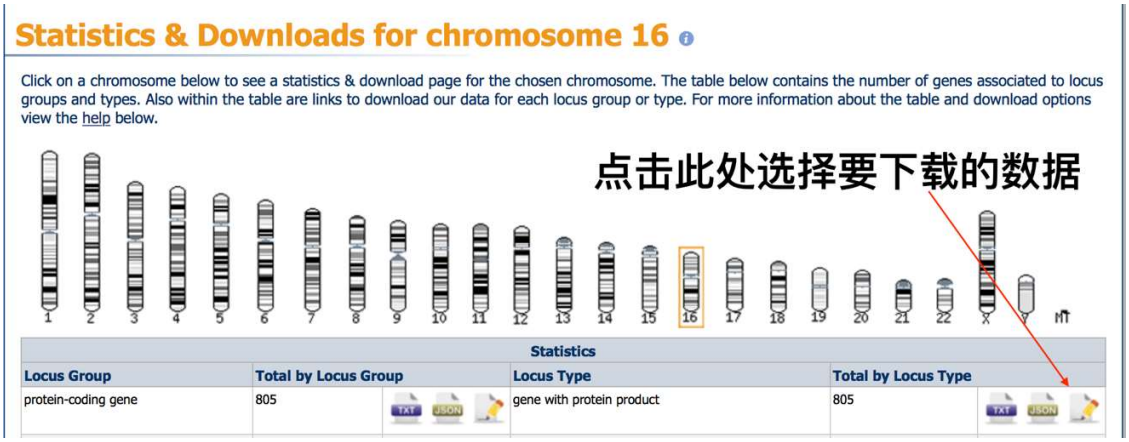
a)EPD:https://epd.vital-it.ch/index.php

b)HGNC: https://www.genenames.org

二、方法

1. 获取启动子区域序列

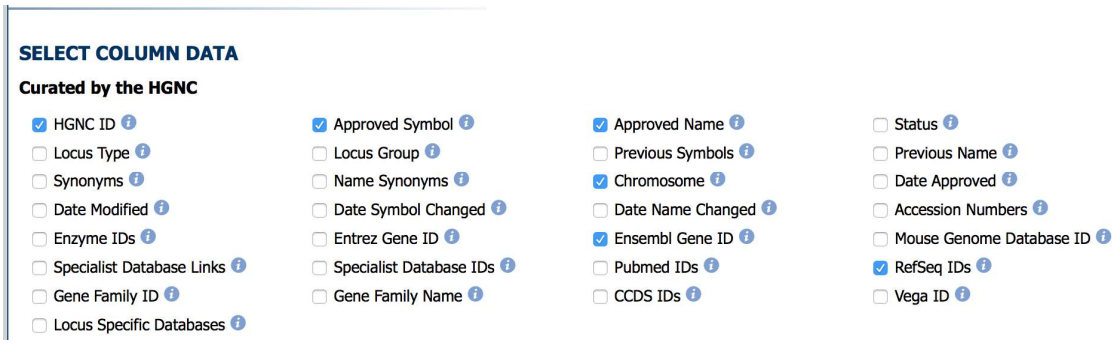
1) 登陆 HGNC 数据库，选择人类第 16 条染色题



图一、下载的数据源

2）勾选要下载的数据

包括 HGNC ID, Approved Symbol, Ensembl Gene ID, RefSeq IDs



图二、下载数据内容

3）随机提取其中 100 个基因的 Ensembl Gene ID
使用 r 语言中的 sample 函数进行随机抽取
4）在 R 语言环境中安装 biomaRt 包，利用该包下载上述 1000 个基因的上游 启动子序列(-1000bp)
通过 library(BiocInstaller)，biocLite("biomaRt")安装 biomaRt 包，通过使用 getSequence()来获取上游 1000bp 片段
详情请见代码文件夹下的 getUpstream.R 文件

```
############################## getUpstream.R ##############################
downloaddata <- read.table("/Users/fengjiarong/Desktop/download.txt",sep="\t",header=T)
new_data <- apply(downloaddata,2,as.vector)
filter_data <- new_data[which(new_data[,5]!=""),]
index <- sample(1:nrow(filter_data),100,replace=F)
select_data <- filter_data[index,]

library(biomaRt)
listMarts()
ensembl = useMart("ENSEMBL_MART_ENSEMBL")
listDatasets(ensembl)
ensembl= useDataset("hsapiens_gene_ensembl",mart=ensembl)
#listAttributes(ensembl)
ensembl=useMart("ensembl",dataset="hsapiens_gene_ensembl")

seq                                                                                    =
getSequence(id=select_data[,5],type="ensembl_gene_id",seqType="coding_gene_flank",upstrea
m=1000,mart=ensembl)
write.table(seq,"/Users/fengjiarong/Desktop/promoter.txt",sep = "\t")
```

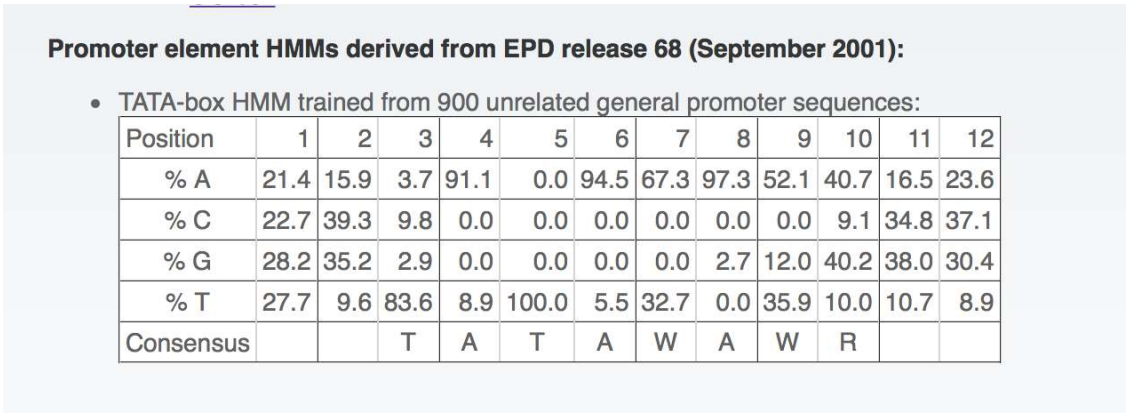
5）将获取的序列片段转化为 fasta 格式
请见代码文件夹下的 convertToFasta.py 文件

```
########################### convertToFasta.py ##############################
def convertToFasta(inputfile,outputfile):
    fr = open(inputfile,'r')
    fw = open(outputfile,'w')
    next(fr)
    for line in fr:
        sequence = line.split("\t")[0]
        id = line.split("\t")[1].strip("\n")
        fw.write(">"+id+'\n')
        fw.write(sequence+'\n')
    fr.close()
    fw.close()
convertToFasta("/Users/fengjiarong/Desktop/promoter.txt","/Users/fengjiarong/Deskto
p/promoter.fasta")
```

2. 核心启动子的鉴定识别

1）从 EPD 数据库中下载启动子相关的 DNA 元件

图三、启动子元件查找方法



Promoter element HMMs derived from EPD release 68 (September 2001):

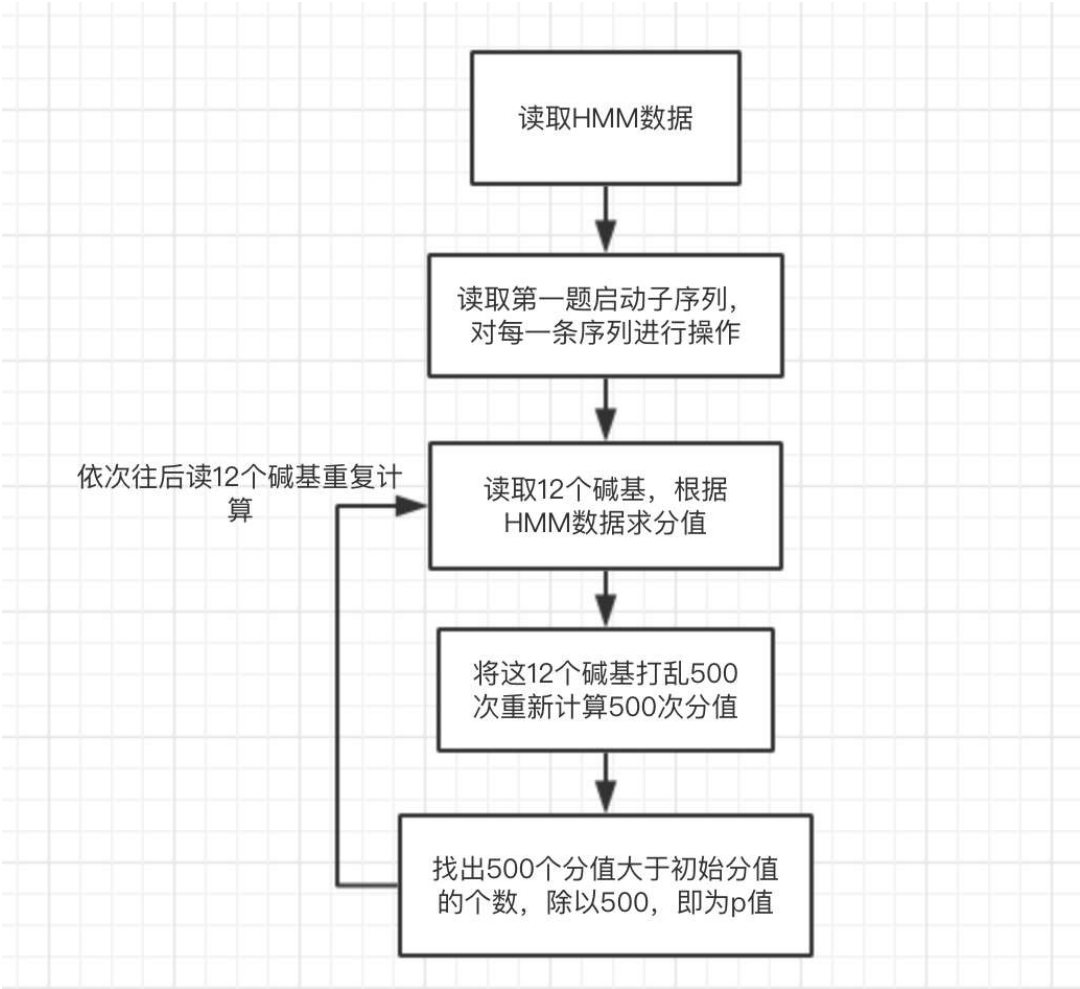
- TATA-box HMM trained from 900 unrelated general promoter sequences:

Position	1	2	3	4	5	6	7	8	9	10	11	12
% A	21.4	15.9	3.7	91.1	0.0	94.5	67.3	97.3	52.1	40.7	16.5	23.6
% C	22.7	39.3	9.8	0.0	0.0	0.0	0.0	0.0	0.0	9.1	34.8	37.1
% G	28.2	35.2	2.9	0.0	0.0	0.0	0.0	2.7	12.0	40.2	38.0	30.4
% T	27.7	9.6	83.6	8.9	100.0	5.5	32.7	0.0	35.9	10.0	10.7	8.9
Consensus			T	A	T	A	W	A	W	R		

图四、所选取的 HMM 模型

2）根据该 HMM 数据，编写程序对第一部分的启动子序列进行计算分析，使 用 bootstrap 抽样评估(注意随意打乱的次数问题)的方法对每个计算的片段进行 p 值计算

思路如下



图五、分析启动子元件的思路

3）根据输出的分数，绘制曲线

4）根据 p 值，提出该启动子序列，在什么位置，具有该启动子元件及其可能性大小的(一般，p<0.05 或更低)

设置筛选条件：在分数大于 0.8 的条件下，再计算子序列的 p 值，挑选 p 值<0.01 的片段

核心启动子的鉴定识别这部分代码请见代码文件夹下的 promoter_identify.py 文件

```
######################## promoter_identify.py ##############################
import pandas as pd
import math
import random
import matplotlib.pyplot as plt

#read HMM
data = pd.read_table("/Users/fengjiarong/Desktop/TATAbox HMM general promoter.txt",sep="\t")
HMM_data = []
```

```
for i in range(1,13):
    site_data = {'A':data.iloc[0,i],'C':data.iloc[1,i],'G':data.iloc[2,i],'T':data.iloc[3,i]}
    HMM_data.append(site_data)

#read promoter
promoter_data = pd.read_table("/Users/fengjiarong/Desktop/promoter.txt",index_col=False)


#calculate score
def calculate_score(seq):
    max_value = 22.193515664864993
    value=1
    for i in range(12):
        try:
            value += math.log(float(HMM_data[i][seq[i]]),10)
        except ValueError:
            return 0
    return value/max_value

#calculate pvalue
def calculate_pvalue(subseq,s):
    count = 0
    for i in range(500):
        newseq=list(subseq)
        random.shuffle(newseq)
        new_score = calculate_score(newseq)
        if new_score>s:
            count += 1
    return count/500

#取第一行的序列
seq=promoter_data.iloc[0,0]
score = []
for i in range(0,989):
    subseq = seq[i:i+12]
    s= calculate_score(subseq)
    score.append(s)

x = range(0,989)
plt.plot(x,score)
plt.show()



fw = open("/Users/fengjiarong/Desktop/mypromoter.txt",'w')
for i in range(0,100):
    id = promoter_data.iloc[i,1]
```

```
        proseq = promoter_data.iloc[i,0]
        fw.write(id+"\n")
        fw.write("%s\t%s\t%s\t%s\t%s\n"%("sequence","start","end","score","pvalue"))
        score = []
        for j in range(0,989):
                subseq = proseq[j:j+12]
                s= calculate_score(subseq)
                score.append(s)
                if(s>0.8):
                        pvalue = calculate_pvalue(subseq,s)
                        if(pvalue<0.01):
                                fw.write("%s\t%s\t%s\t%s\t%s\n"%(subseq,j,j+12,s,pvalue))
fw.close()
```

三、结果

1. 从 HGNC 数据库下载的数据

	A	B	C	D	E	F
1	HGNC ID	Approved Symbol	Approved Name	Chromosome	Ensembl Gene ID	RefSeq IDs
2	HGNC:20	AARS	alanyl-tRNA synthetase	16q22.1	ENSG00000090861	NM_001605
3	HGNC:23	ABAT	4-aminobutyrate aminotransferase	16p13.2	ENSG00000183044	NM_020686
4	HGNC:33	ABCA3	ATP binding cassette subfamily A member 3	16p13.3	ENSG00000167972	NM_001089
5	HGNC:51	ABCC1	ATP binding cassette subfamily C member 1	16p13.11	ENSG00000103222	NM_004996
6	HGNC:57	ABCC6	ATP binding cassette subfamily C member 6	16p13.11	ENSG00000091262	
7	HGNC:14639	ABCC11	ATP binding cassette subfamily C member 11	16q12.1	ENSG00000121270	NM_032583
8	HGNC:14640	ABCC12	ATP binding cassette subfamily C member 12	16q12.1	ENSG00000140798	NM_033226
9	HGNC:25070	ACD	ACD, shelterin complex subunit and telomerase recruitment factor	16q22.1	ENSG00000102977	NM_022914
10	HGNC:27288	ACSF3	acyl-CoA synthetase family member 3	16q24.3	ENSG00000176715	NM_174917
11	HGNC:18049	ACSM1	acyl-CoA synthetase medium chain family member 1	16p12.3	ENSG00000166743	NM_052956
12	HGNC:32017	ACSM2A	acyl-CoA synthetase medium chain family member 2A	16p12.3	ENSG00000183747	NM_001010845
13	HGNC:30931	ACSM2B	acyl-CoA synthetase medium chain family member 2B	16p12.3	ENSG00000066813	NM_182617
14	HGNC:10522	ACSM3	acyl-CoA synthetase medium chain family member 3	16p12.3	ENSG00000005187	NM_005622
15	HGNC:26060	ACSM5	acyl-CoA synthetase medium chain family member 5	16p12.3	ENSG00000183549	NM_017888
16	HGNC:30714	ADAD2	adenosine deaminase domain containing 2	16q24.1	ENSG00000140955	NM_139174
17	HGNC:17110	ADAMTS18	ADAM metallopeptidase with thrombospondin type 1 motif 18	16q23.1	ENSG00000140873	
18	HGNC:228	ADAT1	adenosine deaminase, tRNA specific 1	16q23.1	ENSG00000065457	NM_012091
19	HGNC:238	ADCY7	adenylate cyclase 7	16q12.1	ENSG00000121281	
20	HGNC:240	ADCY9	adenylate cyclase 9	16p13.3	ENSG00000162104	
21	HGNC:4512	ADGRG1	adhesion G protein-coupled receptor G1	16q21	ENSG00000205336	

图六、从 HGNC 数据库下载的数据

完整文件请见过程文件夹下的 download.txt 文件

2. 获取的部分 100 个基因的上游启动子序列

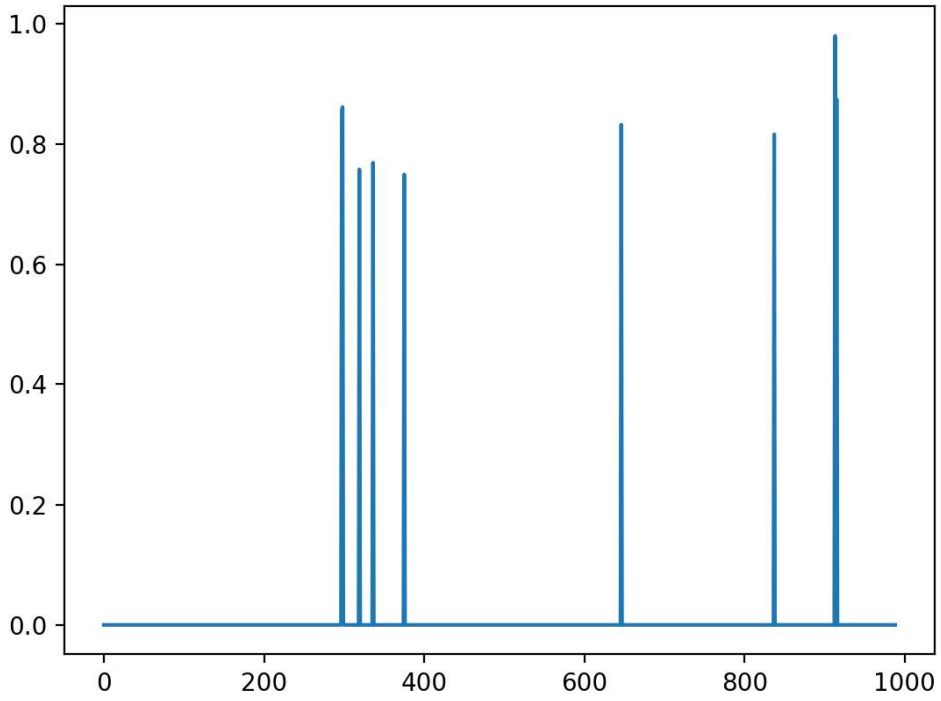
promoter.fasta

```
>"ENSG00000130656"
"AAAATTGGCCGGGCACAGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAGGCTGAGGCAGGTGGATCAGGAGTCAGGAGATCGAGACCATCCTGGCCAACATGGTGAAACCCCATCTCT
ATGAAAATACAAAAATTAGCCAGAGATGATGCCGGGTGCCTGTAATCCCAGCTACTCATGAGGCTGAGGCAGAAGAATCACTTGAACCAGGGAGTCAGAGGTTGCAGTGAGCTGAGATCGC
ACCACTGCACTCCACCCTGGGCGACAAATCGAGATTCCATCTCAAAAAAAGAAAAAAAAATTAAAAGGAATATTTGCCTCATTATGTTACAATAACTAATATGGAAAGCAATATTGCAATG
CCTATTAGCACATGACATTAGGTGAATTCTCCTTTGTCCCCGGACCTGCTGCCTCCTCCTGCTTGTCAGGGGACAGATCCAGTACATCTCCCCTCAGCGCTGGGTGGACCTAACCCTTGCT
TTCTTGGAGGAAACCCAGGAATCCAGAGACAAAGTGGAAGGGTACTGGCATGTGGTTGGGCAGGGCTGCCTGAGGTCGGTGTCAGCCGACCGTGGGGCTTGGTCCCAGGAGGCTGCTTACT
GGGCCCTGCTCCTCTGGTTTCCCCCAAGTCGTGATTCTGAAATGAATAAGGACGGTGCAGAACTGGACTACAAATGCAGGAGTGACTTCCTGGGAGGGTGGGGCCCCTATCTCTCCTAGAC
TCTGTGGTCAGACTCTGGCCAACACCCCCTGTAAGGCCACAGGAGAGGAACAGGAGTGATAGCCCCCAAACCCCAGTCCCACCAGGCCCTGAGGGCCCCTTTGTCACTGGATCTGATAAGA
AACACCACCCCTGCAGCCCCCTCCCCTCACCTGACCAATGGCCACAGCCTGGCTGGGCCCAGCTCCCTGTATATAAGGGGACCCTGGGGGCTGAGCACTACCAAGGCCAGTCCTGAGCAGG
CCCAACTCCAGTGCAGCTGCCCACCCTGCCGCC"
>"ENSG00000175311"
"TCATACGTTATTGGTGGAAATGTGAAGTAGTACAGCCACTTTGGAAAACAGTCTGGCAACTTCTGAAATTTAAATGTAAATTGCCATGCAACCCAGCAGTTTCACTCCTTGATATATACG
CAAGACACATGAAAATATGTGTCCACACAGAGACTTGTACATCAGTGCTCATAGCAGCATTATTCATCATTACTGAAAAATGGAAACAACCCAAATATTCATCATATTCATGAAACATTGA
ATGGATGAAGAACATGTGGTATGTTAGCACAGTGGAATATTATTCAGCAGTGAAAAGAAAGAAACCATTACATGCTACACATGACAAACCTCAAAAAAGTATGTGAATTGAAACAAGCTAT
ACACTAAAGACTATATATTGTAGGATTCCATTTGTCTGAAATGCCCGTAAAAAACAAATCTATAGAGACAGAAAGCAGACTGTGGTTTCCTGAGGTTGGGAGTGGAAATGGGATTGACAGC
AAATGGCATGAGGGATCTTTAAGGGCTGATGGAAACTGGATTATGATAATGGGCATGCAACTCTATAAATGTCCTAAAAATCATTGAATTGTATGCATGAAATGGGTAAATTTAATGATGT
TATACCTCAATAAAGCTGTTCTTTTTGGAAAAAACAGCTAGTGAACTAGAGTAGCTCTGCATTTTATTAAGAAATTATTTCTTACATCCTGTGCCTAGATGGCCAATATTACACTCAAGAA
ACTAATGGACATAGAAAAATTGAATGTACCTCCAAACACTCAAGATGAATACAGCAGTCTTCAATAGTCAACGTCATGGCCATTCCAGAGAGAAAAGGTCCAAGTTCAGTGCCTCGAACCT
TGTTGCCTTCCGTGAAAGGAAGCAGTTGAAATGAGCTTTGCCCCTGGTGACACTCTTCCTTGTCACTAATTAACCAGCTCTATAGGAAGGTTGCTCCGCCAACACCTTGCTCTGCCTGGAG
AGACATCTGGCCAAGTTCTGGTGAGCAGGAAAA"
```

图七、上游启动子序列

完整文件请见过程文件夹下的 promoter.fasta 文件

3. 分值曲线



图八、分值曲线

4. 筛选的部分启动子序列

ENSG00000130656				
sequence	start	end	score	pvalue
TGTATATAAGGG	913	925	0.9792301847610072	0.0
TATATAAGGGGA	915	927	0.8739721717057731	0.006
ENSG00000175311				
sequence	start	end	score	pvalue
TTGATATATACG	108	120	0.8792613914101766	0.008
ACTATATATTGT	371	383	0.9177857054734677	0.0
TCTTTAAGGGCT	498	510	0.8253607074582903	0.002
TCTATAAATGTC	544	556	0.9673201488218165	0.0
TCAATAAAGCTG	610	622	0.8518989248321578	0.004
ENSG00000166828				
sequence	start	end	score	pvalue
ACTATATATTAT	45	57	0.9014611331130612	0.008
GTTATTAAATCA	200	212	0.8787280417607672	0.006
AACATTAAGAGC	323	335	0.850562812619676	0.004

图九、筛选出的启动子序列

完整文件请见过程文件文件夹下的 mypromoter.txt 文件。

ENSG00000130656
sequence startend score pvalue
TGTATATAAGGG 913 925 0.9792301847610072 0.0
TATATAAGGGGA 915 927 0.8739721717057731 0.006
ENSG00000175311
sequence startend score pvalue

TTGATATATACG	108	120	0.8792613914101766	0.008
ACTATATATTGT	371	383	0.9177857054734677	0.0
TCTTTAAGGGCT	498	510	0.8253607074582903	0.002
TCTATAAATGTC	544	556	0.9673201488218165	0.0
TCAATAAAGCTG	610	622	0.8518989248321578	0.004

ENSG00000166828

sequence	start	end	score	pvalue
ACTATATATTAT	45	57	0.9014611331130612	0.008
GTTATTAAATCA	200	212	0.8787280417607672	0.006
AACATTAAGAGC	323	335	0.850562812619676	0.004

ENSG00000166152

sequence	start	end	score	pvalue
CCTTTTAATGCC	105	117	0.8853387285534868	0.0
TCTATTTATCCT	195	207	0.8636182584676708	0.0
TGAATATAGACG	472	484	0.8880099807282399	0.008
CTTATAAAAAGC	491	503	0.9681733809492686	0.0
ACCTTTTAAGGC	830	842	0.8371214030102054	0.002
CCTTTTAAGGCC	831	843	0.8638949334505719	0.0
CCTTTATGTTCA	871	883	0.8206425690954996	0.004

ENSG00000182810

sequence	start	end	score	pvalue
GGCTTAAAAGGA	353	365	0.9012869905408509	0.0

ENSG00000198156

sequence	start	end	score	pvalue
GTTTTAAATACC	37	49	0.9178950580623809	0.0
GTTATAAAAACA	50	62	0.9618452709594095	0.0
ATTATATATAGT	364	376	0.9176718475273543	0.0
GACATAAAGAGG	395	407	0.9077148270698667	0.002
TGATTAAATAGG	541	553	0.8797848979464046	0.008
GCTTTTTAGACT	590	602	0.8263227155172531	0.004
CATTTAAATGAC	694	706	0.9086779056829306	0.004
GTTTTATAGGTA	712	724	0.850154269904085	0.008
AGGTTAAAAGGG	798	810	0.8770140524331724	0.0

ENSG00000177508

sequence	start	end	score	pvalue
CGAATAAGAGGG	189	201	0.8583056647203312	0.0

ENSG00000129910

sequence	start	end	score	pvalue

ENSG00000167397

sequence	start	end	score	pvalue
GCCATTTAACCT	15	27	0.8293078950593218	0.0

ENSG00000168807

sequence	start	end	score	pvalue
TTTATATATCTC	186	198	0.8965439224058438	0.002
TCTATATATCCT	196	208	0.9192682572179941	0.0
TCCTTTTATGCT	441	453	0.805226251951874	0.0

ENSG00000135722

sequence	start	end	score	pvalue

ENSG00000196993

sequence	start	end	score	pvalue
GGATTATAGGCG	238	250	0.8426035925547214	0.002
ATTATTTATACG	890	902	0.884338242077336	0.008

ENSG00000214940

sequence	start	end	score	pvalue
TCTATTTATGGT	47	59	0.8944105413917995	0.0
TCCATTTGTAGC	73	85	0.8104951740895668	0.006
GGGATTTAAAGC	138	150	0.8622933543668166	0.0
GATTTAAAGCAG	140	152	0.8585112283981446	0.008
GTCTTATAAGGG	712	724	0.8666923290659639	0.002
CTTATAAGGGCA	714	726	0.8584818744039046	0.008

ENSG00000103126

sequence	start	end	score	pvalue
GCTTTAAAGTAG	741	753	0.8780644186311315	0.002

ENSG00000183793

sequence	start	end	score	pvalue
TCTATTTATGGT	47	59	0.8944105413917995	0.0
TCCATTTGTAGC	73	85	0.8104951740895668	0.0
GGGATTTAAAGC	138	150	0.8622933543668166	0.008
AATTTAAAGTCC	483	495	0.8734578868481554	0.004
GTCTTATAAGGG	712	724	0.8666923290659639	0.0
CTTATAAGGGCA	714	726	0.8584818744039046	0.002

ENSG00000103184

sequence	start	end	score	pvalue

ENSG00000197006

sequence	start	end	score	pvalue

ENSG00000157429

sequence	start	end	score	pvalue
GCTATAAAGATA	294	306	0.937616093649358	0.004
AAGATAAGGAGC	300	312	0.8122404464055507	0.004
ACCATTTGAAGC	620	632	0.812733543158988	0.006
GTCATATAATGG	840	852	0.8849815592795532	0.002

ENSG00000087250

sequence	start	end	score	pvalue
CCCTTATATTCA	130	142	0.848838591925425	0.004
TTCATATAAACA	138	150	0.9054226260988867	0.006
TTTATAAATAGA	649	661	0.9559287562760794	0.0

ENSG00000005187

sequence	start	end	score	pvalue
AATATATATTGC	120	132	0.9280133990188625	0.0
TCTATTAAATGC	328	340	0.9165326983129681	0.0
TGCATATGGGTG	622	634	0.813605881268677	0.002
TCTATAAAACAC	791	803	0.9540126033218923	0.0
CAAATATATGTC	918	930	0.8705837310644063	0.008

ENSG00000141002

sequence	start	end	score	pvalue

ENSG00000103024

sequence	start	end	score	pvalue
GGAATAAAGAGA	598	610	0.899251075948009	0.002
ENSG00000172366				
sequence	start	end	score	pvalue
CCCTTTAAGGGC	871	883	0.823668055358893	0.0
CCTTTAAGGGCC	872	884	0.8494006116877699	0.0
CCGTTAAGTGCG	928	940	0.8011705151505986	0.0
ENSG00000140835				
sequence	start	end	score	pvalue
CAGATAAATAGA	312	324	0.8961303755160847	0.008
ACAATAAAACAG	934	946	0.8840565450507156	0.004
TCCTTAAAGGTC	973	985	0.8584134838704977	0.002
ENSG00000179583				
sequence	start	end	score	pvalue
CCATTATAAAGG	16	28	0.8712090569354274	0.002
ATTATAAAGGGA	18	30	0.9291935799594202	0.006
TCTATAAAACAG	137	149	0.9501150538163297	0.002
ENSG00000157106				
sequence	start	end	score	pvalue
ENSG00000180269				
sequence	start	end	score	pvalue
ENSG00000103319				
sequence	start	end	score	pvalue
TGCTTAAGAAGC	662	674	0.8398867070906768	0.002
ENSG00000179776				
sequence	start	end	score	pvalue
GATATATAAAGT	119	131	0.9402326940616839	0.002
TATATAAAGTGC	121	133	0.925743371321921	0.002
CAGATAAAGGGA	531	543	0.8744446919164659	0.0
ACCATATATGGG	587	599	0.9271005626337089	0.0
CATATATGGGCG	589	601	0.8591856908064841	0.002
ATCTTAAATGCG	759	771	0.8664077571714843	0.002
ENSG00000103187				
sequence	start	end	score	pvalue
AGAATAAAGGGG	383	395	0.8985642937949496	0.0
ENSG00000166747				
sequence	start	end	score	pvalue
CTGTTATATGTC	343	355	0.8104284132708146	0.006
GTTATATGTCCC	345	357	0.8498282831962151	0.0
TGTTTTAATCTG	498	510	0.8310311350087152	0.004
TGTATTAGTTTC	689	701	0.812144457547349	0.008
AATTTAAATGGG	918	930	0.9199508939886436	0.008
TCTATTTAATGT	971	983	0.8744730612266461	0.008
ENSG00000103175				
sequence	start	end	score	pvalue
CACATAAATGCA	264	276	0.9179951118791829	0.004
GCTTTATGTGCT	327	339	0.8330303003606615	0.0
CTTTTAAATCTC	345	357	0.8612581667616631	0.004

AGCTTTTATAGG 492 504 0.8240218501554816 0.004
CTTTTATAGGCA 494 506 0.868987348062893 0.002
ATGTTATAGGGC 787 799 0.8126306120253972 0.006
ENSG00000135697
sequence startend score pvalue
TTTTTAAATAGG 283 295 0.9153688556599499 0.0
TCTTTATGTATG 351 363 0.8338813204456952 0.004
CTTATTTAGAAC 604 616 0.8533428782602295 0.004
ACAATAAGAGGA 687 699 0.8543529886815295 0.008
ENSG00000091651
sequence startend score pvalue
CACTTAAAGGAG 123 135 0.841388263391253 0.008
ENSG00000167264
sequence startend score pvalue
ENSG00000125170
sequence startend score pvalue
ENSG00000168418
sequence startend score pvalue
GGGATTAAGCGA 78 90 0.8095209992824245 0.004
TCTATAAAGTGG 260 272 0.9395534906683525 0.0
TGCATATAACCC 376 388 0.910387153659241 0.0
ENSG00000175938
sequence startend score pvalue
GCTATAAAACTC 465 477 0.9458872197640601 0.004
ENSG00000006194
sequence startend score pvalue
GGCATTTGAGGA 386 398 0.806882968753673 0.002
ENSG00000103091
sequence startend score pvalue
CAGTTAAATGCA 27 39 0.8486524952626557 0.002
TTTATTTATTTT 320 332 0.8148040826391242 0.004
AAAATAAAAAAA 796 808 0.8907069564789585 0.0
ENSG00000064270
sequence startend score pvalue
GCCATTTGAGCC 49 61 0.8161698002761929 0.0
TCGTTAAAAACC 582 594 0.8866375936203619 0.0
GGATTTTAAACC 646 658 0.8198246886960411 0.006
ENSG00000181938
sequence startend score pvalue
CATATAAATAGT 617 629 0.9428235763078562 0.0
ENSG00000155330
sequence startend score pvalue
ATTATATATTGA 369 381 0.9092877584217873 0.008
CTCATAAAATGT 422 434 0.8708224904203608 0.008
TTTATTAAATGC 479 491 0.8889516099185361 0.004
ENSG00000183044
sequence startend score pvalue
GGTTTTTAGCAG 30 42 0.8042885937514788 0.0

CATTTAAATGGT	76	88	0.8970671156459481	0.008
ATGATAAAAGGA	121	133	0.8921488947142735	0.006
AGTATATATTCA	463	475	0.9329913971652948	0.0
TTTTTAAAGTCT	679	691	0.8406985976253687	0.006
AATATATAAAGT	807	819	0.9348331288562539	0.006
TATATAAAGTCC	809	821	0.9240219517303999	0.002
AGTATAAAACCC	827	839	0.9614102333143336	0.0

ENSG00000167513

sequence	start	end	score	pvalue
CCTTTTAAAAAG	70	82	0.8743677765622043	0.008
CTTTTAAAAAGC	71	83	0.9226588087840645	0.002
GAAATAAGGAGC	530	542	0.8224073376969302	0.002

ENSG00000140832

sequence	start	end	score	pvalue

ENSG00000103313

sequence	start	end	score	pvalue
CACATATGAGCT	172	184	0.8219312371208339	0.006

ENSG00000065457

sequence	start	end	score	pvalue
CTTATTTAACCC	232	244	0.8673649360891794	0.004
CCCTTAAGTAGA	427	439	0.8220071961410723	0.008
TGGATATGAGAG	905	917	0.8269849497529518	0.008

ENSG00000102931

sequence	start	end	score	pvalue
TCTTTATGATGC	265	277	0.8423995296960288	0.002
GGGATAAGGGAC	384	396	0.8166251895464607	0.004
CTTATTTAATCC	451	463	0.8692104574701723	0.0

ENSG00000118898

sequence	start	end	score	pvalue
AGCTTTTATGAC	145	157	0.8113529388039338	0.008

ENSG00000267795

sequence	start	end	score	pvalue

ENSG00000034713

sequence	start	end	score	pvalue
CGAATAAAGGAA	248	260	0.8784390844345994	0.004
TGTTTAAATTCA	272	284	0.9066505924060638	0.002

ENSG00000102878

sequence	start	end	score	pvalue
CCTATTAAAGGC	372	384	0.9398625820966736	0.0
CTATTAAAGGCG	373	385	0.8270572782523269	0.002

ENSG00000069764

sequence	start	end	score	pvalue
GGATTATAGGCG	72	84	0.8426035925547214	0.002
GGATTATAGGTC	444	456	0.823422533245177	0.002

ENSG00000131153

sequence	start	end	score	pvalue
CCAATAAATGCT	369	381	0.8975589210230516	0.0
AAAATAAATGCG	533	545	0.9027350305137529	0.008

ENSG00000177548

sequence	start	end	score	pvalue
GACATAAATCTG	266	278	0.8750458419368417	0.004
CAGATAAAGAGC	308	320	0.8835388014678071	0.002

ENSG00000135686

sequence	start	end	score	pvalue
GGGTTAAATGAG	613	625	0.8588011834917343	0.004

ENSG00000185324

sequence	start	end	score	pvalue

ENSG00000103005

sequence	start	end	score	pvalue
GGCTTTTAGGGG	539	551	0.8077357317612928	0.002
GCGATATATGAA	724	736	0.8873928394783347	0.002

ENSG00000140990

sequence	start	end	score	pvalue

ENSG00000103056

sequence	start	end	score	pvalue

ENSG00000167964

sequence	start	end	score	pvalue

ENSG00000103496

sequence	start	end	score	pvalue
AATATAAGTTCC	200	212	0.8702719336031488	0.008
TCTATTAGTCCT	231	243	0.807598212643478	0.002
CCATTAAATGAC	300	312	0.865376558486351	0.004
AGTATTTATAGA	583	595	0.9065300484548113	0.006
TATTTATAGAGG	585	597	0.8896742054105402	0.008

ENSG00000188603

sequence	start	end	score	pvalue

ENSG00000103404

sequence	start	end	score	pvalue
CGTTTTAAGAGA	210	222	0.8548499904247561	0.006
ACTTTAAAGGTG	260	272	0.8914147393310946	0.0
AGAATTTATCCC	512	524	0.823339082040042	0.004
ACCATTAAGCTG	692	704	0.8102601517301677	0.004
TGTATTAGAGTG	786	798	0.8427601117798629	0.004

ENSG00000087245

sequence	start	end	score	pvalue
GGTATTAAAAAG	173	185	0.9219718493243038	0.006
AGGATTTAAACC	324	336	0.8551723695698658	0.002
GGATTTAAACCC	325	337	0.8046362115556623	0.004
GCATTAAATGAG	636	648	0.8657245396370875	0.004

ENSG00000171724

sequence	start	end	score	pvalue
AGAATATAGACC	23	35	0.8868580375167623	0.006
CGTATATAACCA	619	631	0.9395877721198942	0.0

ENSG00000169900

sequence	start	end	score	pvalue
CAAATATAGGAG	478	490	0.8537178425021507	0.008

ENSG00000184857

sequence	start	end	score	pvalue
CCCATTTAAAGC	159	171	0.8840318982212444	0.0
CCATTTAAAGCT	160	172	0.8036822118939855	0.006
CATTTAAAGCTG	161	173	0.8457902416956238	0.002

ENSG00000051523

sequence	start	end	score	pvalue
GCTATTAATAGA	653	665	0.9282099184085778	0.0

ENSG00000158486

sequence	start	end	score	pvalue
ATTTTTTAACCC	140	152	0.8206963293748446	0.008
AATATAAAAAAA	166	178	0.9517159725275322	0.0
ATTTTATATTTC	189	201	0.847825378904592	0.008
ATCATATAGGGG	329	341	0.8780756791363618	0.0
CAAATTAAAGCA	381	393	0.8505722565499475	0.006

ENSG00000172831

sequence	start	end	score	pvalue
GCTTTTTATATC	325	337	0.8526232642024496	0.006
TTTTTATATCCC	327	339	0.874107959055747	0.0

ENSG00000186187

sequence	start	end	score	pvalue

ENSG00000172382

sequence	start	end	score	pvalue
TCAATAAAATAC	324	336	0.8948491086543113	0.008
AGCATAAAACTC	902	914	0.8963833268160264	0.002

ENSG00000102904

sequence	start	end	score	pvalue

ENSG00000166676

sequence	start	end	score	pvalue

ENSG00000140750

sequence	start	end	score	pvalue

ENSG00000205220

sequence	start	end	score	pvalue

ENSG00000196470

sequence	start	end	score	pvalue
ACCATTAAGAAG	204	216	0.848048359605701	0.004
TCTTTATAAATT	464	476	0.8872756898534561	0.006
TTTATAAATTTT	466	478	0.8845783560783804	0.004
TGTTTATGGTCC	689	701	0.8097904231557278	0.0
ACGATTAAGGGA	745	757	0.835348327469006	0.002
CTAATAAAGTGC	812	824	0.8509654775629998	0.008
CTTTTTTAATCC	854	866	0.8236958853049684	0.006
ACTTTAAAAGGA	905	917	0.9399917530780248	0.0
TGTATTTATTTC	984	996	0.8681645033715416	0.0

ENSG00000186153

sequence	start	end	score	pvalue
CACATTAAGTGC	156	168	0.8242496167204881	0.004
GCTTTTTAGTCG	503	515	0.8228932979607605	0.0

ENSG00000103550

sequence	start	end	score	pvalue
GTGTTATATTCC	711	723	0.8105272107906059	0.0
GGCATTTAGGTT	786	798	0.8044124626280423	0.004
TGCTTTAAAGGA	855	867	0.8452869193021381	0.006
GCTTTAAAGGAA	856	868	0.9003350890336719	0.002

ENSG00000166592

sequence	start	end	score	pvalue
GGCTTAAATAGG	612	624	0.8991956888001679	0.002

ENSG00000167394

sequence	start	end	score	pvalue
TCCTTAAGTGGA	206	218	0.8256607658122779	0.008
GCCATTAATCCT	280	292	0.8361443079617931	0.002

ENSG00000149922

sequence	start	end	score	pvalue
GCTATAAAGGGG	929	941	0.967128905108357	0.0

ENSG00000125149

sequence	start	end	score	pvalue

ENSG00000149926

sequence	start	end	score	pvalue
CTTTTTTAATGG	155	167	0.8215197553909268	0.004

ENSG00000206053

sequence	start	end	score	pvalue
ATTATAAATATG	371	383	0.9310339067014852	0.006
TGTATAAAGCAG	396	408	0.9192273668675497	0.002
GGAATAAAAGGT	478	490	0.9086577029979296	0.0
TGTTTAAATTTC	528	540	0.8924242046455939	0.0
TGTTTTTATTGT	556	568	0.819514597215577	0.006
TGCATAAATACA	567	579	0.9376840973363867	0.0

ENSG00000129993

sequence	start	end	score	pvalue

ENSG00000140932

sequence	start	end	score	pvalue

ENSG00000182685

sequence	start	end	score	pvalue

ENSG00000169592

sequence	start	end	score	pvalue
GGGATAAATCCT	684	696	0.8658102322013896	0.0

ENSG00000196296

sequence	start	end	score	pvalue

ENSG00000103549

sequence	start	end	score	pvalue

ENSG00000059122

sequence	start	end	score	pvalue
TCAATAAGAACG	558	570	0.8628776218528272	0.002

ENSG00000140986

sequence	start	end	score	pvalue
TTCATAAAAGTA	84	96	0.8962264034760087	0.008

TTTTTATATTTT	305	317	0.8249395092242434	0.0
TGTTTTTAAGAC	363	375	0.8656385909906358	0.008

ENSG00000103507

sequence	start	end	score	pvalue
ACCATATGGGGG	355	367	0.8355124470176684	0.0

ENSG00000166816

sequence	start	end	score	pvalue
GCGTTAAAAGGT	0	12	0.8605318348061797	0.002
AGGATAAAACCT	43	55	0.8676985287824216	0.006
GCTATTTAAAAA	419	431	0.9050489331456999	0.004
ACCTTAAAGCCC	862	874	0.8473717366359568	0.002

ENSG00000090238

sequence	start	end	score	pvalue

ENSG00000140859

sequence	start	end	score	pvalue
GGAATAAAAATC	472	484	0.9120349254853942	0.006

ENSG00000278848

sequence	start	end	score	pvalue
GATATTTGGAGG	240	252	0.8097445308006849	0.006
CACTTAAATGGG	258	270	0.8791566308545744	0.004
CTGATTTAATCC	797	809	0.8034341153356492	0.006