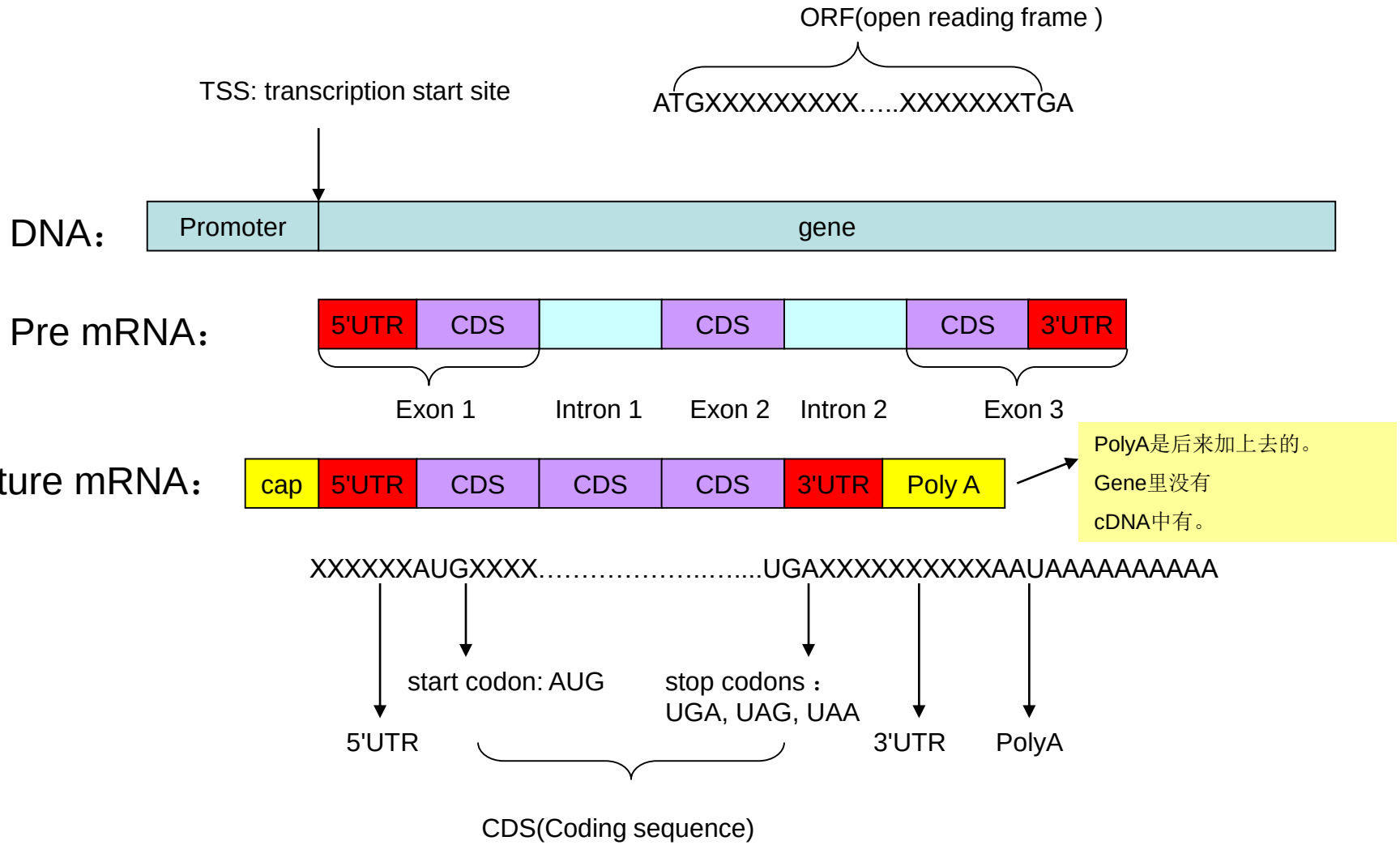
彻底搞清楚promoter, exon, intron, and UTR

- 启动子: RNA聚合酶特异性识别和结合的DNA序列。
- promoter自然不属于intron和Exon的任何一个，属于noncoding sequence。
- noncoding RNA是现在研究的热点之一。我们常见的MiRNA,SiRNA, antisense RNA tech,这些都是属于ncRNA的范围。只要你在进一步问下：这些RNA是哪里来的？你就知道部分答案，跟那些看似跟编码蛋白没有关系的DNA序列有关系。这部分DNA有个统称就junk DNA，垃圾ＤＮＡ或者冗余DNA，他们编码的RNA就属于 ncRNA.
  RNAi就是迄今最经典的ncRNA功能典范。

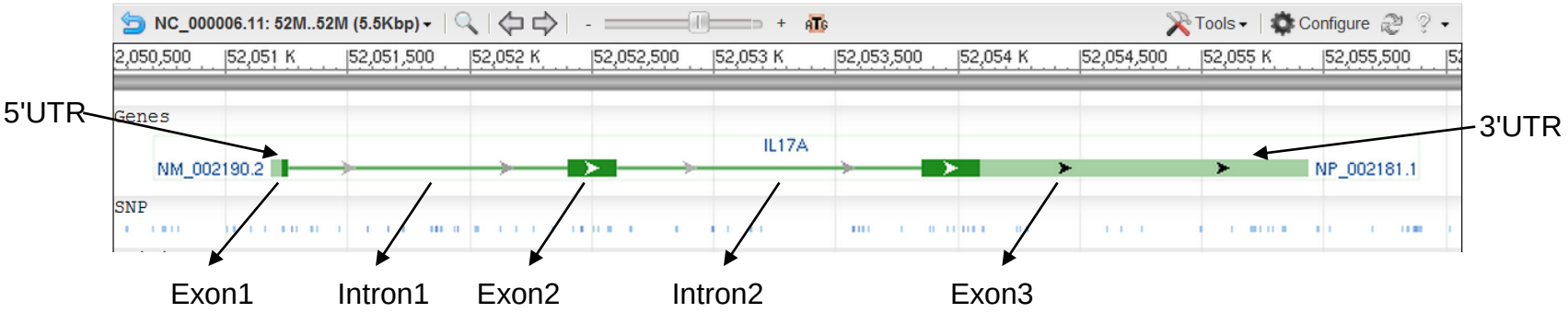
- An exon is a sequence of DNA that is expressed (transcribed) into RNA and then often, but with many noteworthy exceptions[1] , translated into protein. Adjacent exons may be separated by an intron, which is later removed from the RNA transcript via the splicing mechanism. (From Wikipedia)

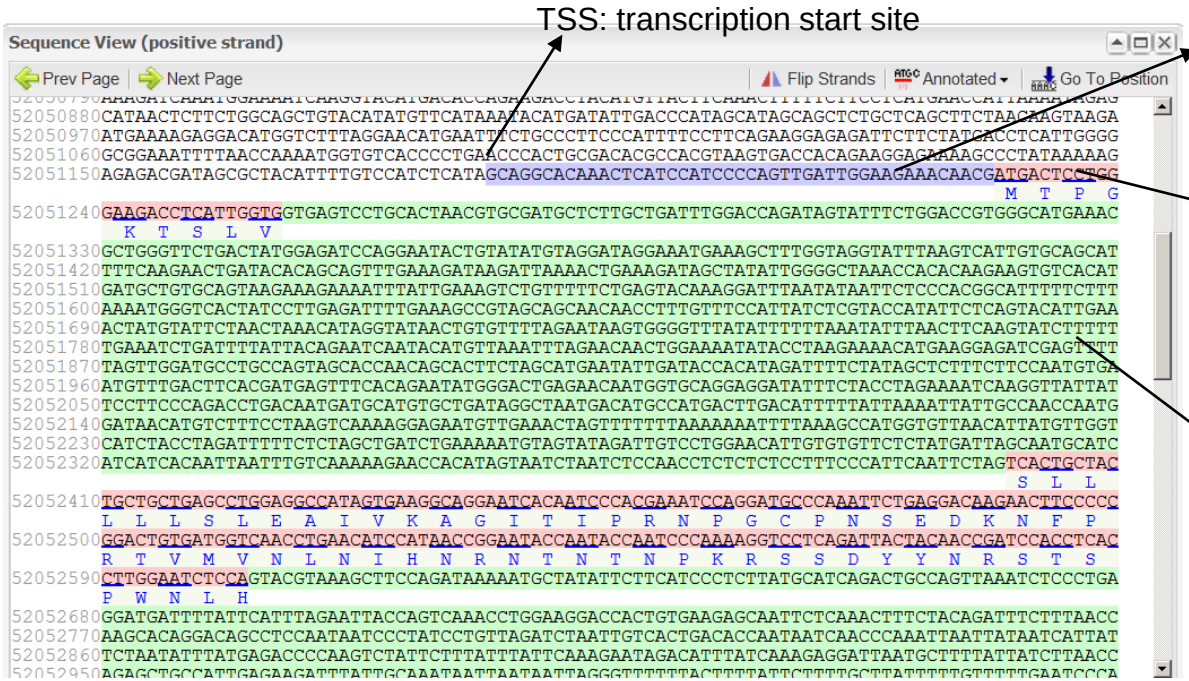
- UTR（Untranslated Regions)即非翻译区，是信使RNA（mRNA）分子两端的非编码片段。
- 5'-UTR从mRNA起点的甲基化鸟嘌呤核苷酸帽延伸至AUG起始密码子，3'-UTR从编码区末端的终止密码子延伸至多聚A尾巴（Poly-A）的末端。

转录（Transcription）是遗传信息从DNA到 RNA的转移。即以双链DNA中的一条链为模板，以ATP、CTP、GTP和UTP4种核苷三磷酸为原料，在RNA聚合酶催化下合成RNA的过程。

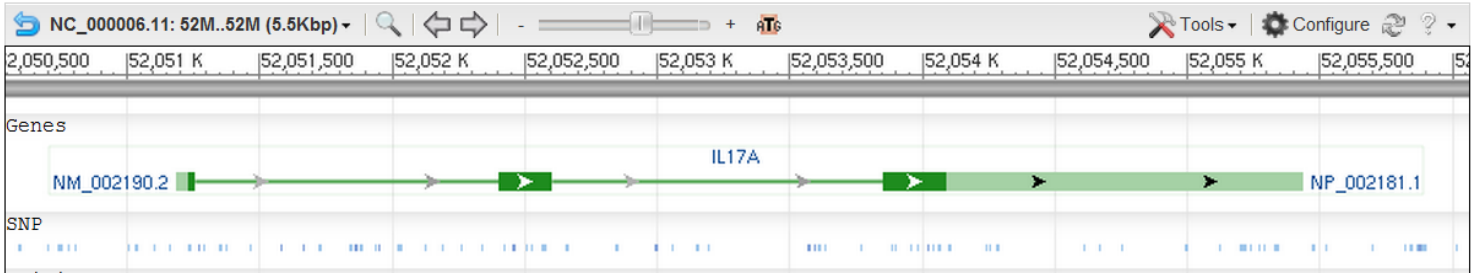

ORF(open reading frame)
TSS: transcription start site
ATGXXXXXXXXX…..XXXXXXXTGA
DNA：
Promoter
gene
Pre mRNA：
5'UTR
CDS
CDS
CDS
3'UTR
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Mature mRNA：
cap
5'UTR
CDS
CDS
CDS
3'UTR
Poly A
PolyA是后来加上去的。
Gene里没有
cDNA中有。
XXXXXXAUGXXXX……..…....UGAXXXXXXXXXXAAUAAAAAAAAA
start codon: AUG
stop codons：
UGA, UAG, UAA
5'UTR
3'UTR
PolyA
CDS(Coding sequence)
cDNA：

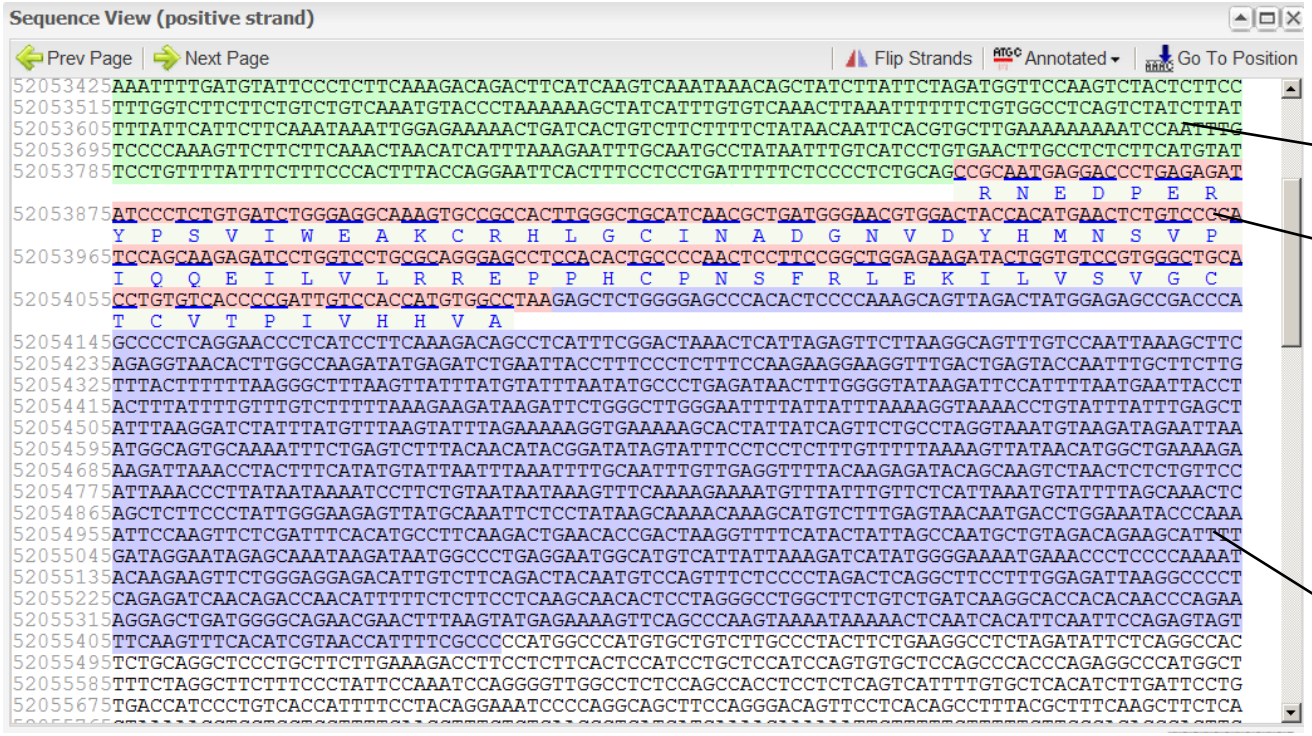
Example1
IL17A interleukin 17A[Homo sapiens]
Location : 6p12
Sequence : Chromosome: 6; NC_000006.11 (52051185..52055436)
gene 1..4252
mRNA join(1..72,1217..1419,2669..4252)
CDS join(46..72,1217..1419,2669..2906)





Example1
IL17A interleukin 17A[Homo sapiens]
Location : 6p12
Sequence : Chromosome: 6; NC_000006.11 (52051185..52055436)
gene 1..4252
mRNA join(1..72,1217..1419,2669..4252)
CDS join(46..72,1217..1419,2669..2906)





Green area: intron

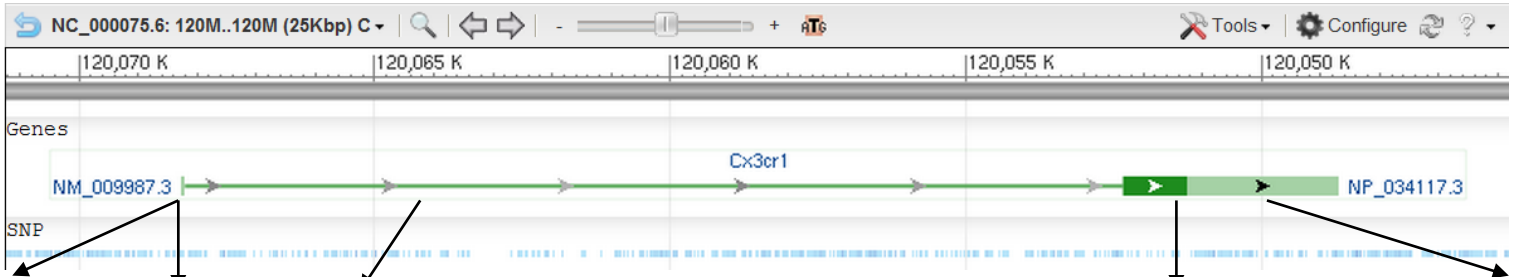
Intron2 (1420-2668)

CDS (2669-4252)

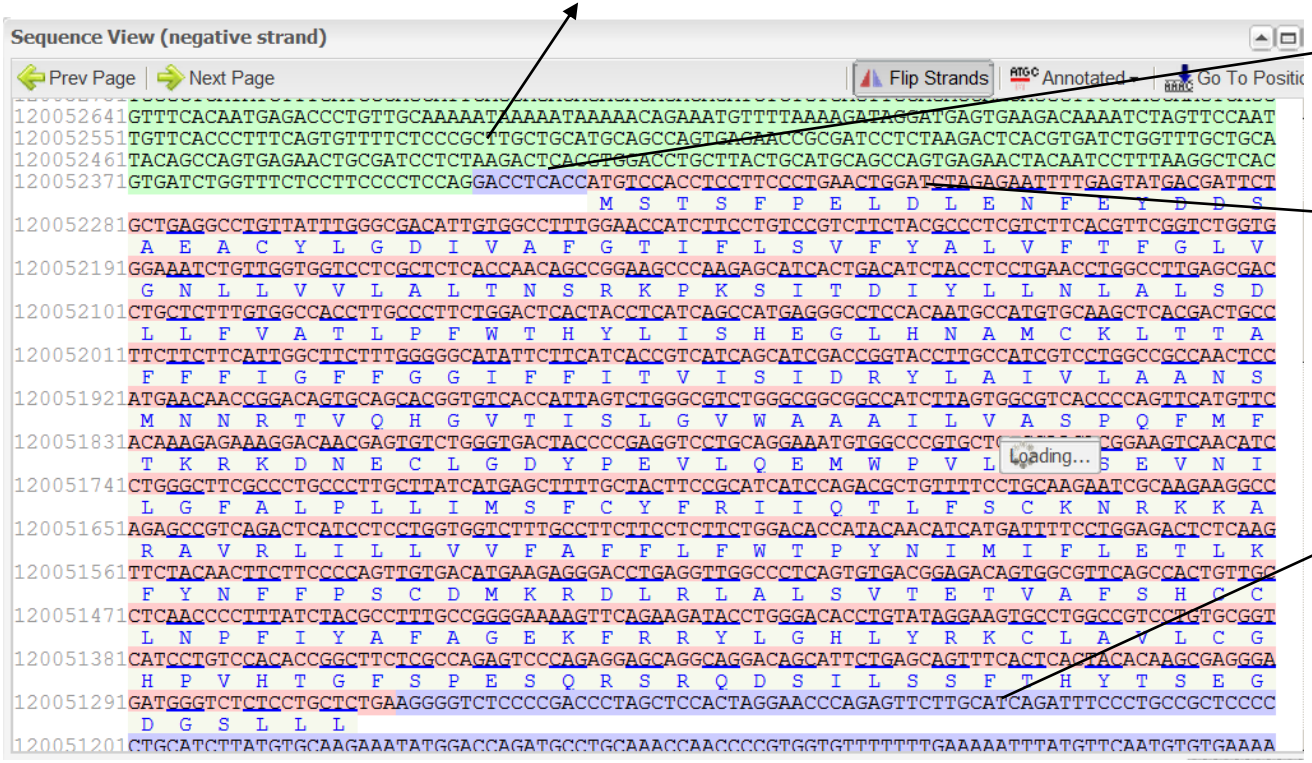
End codon: TAA

Gray area: 3'UTR (2907-4252)

Example2:
Cx3cr1 chemokine (C-X3-C) receptor 1 [Mus musculus]
Location : 9 F4; 9
Sequence : Chromosome: 9; NC_000075.6 (120048693..120068282, complement) → 互补链上
gene 1..19590
mRNA join(1..78,15940..19590)
CDS 15949..17013



5'UTR
Exon1
Intron1
Green area: Intron1(79-15939)
Exon2
3'UTR



Gray area: 5'UTR (15940-15948) in exon2

CDS (15949-17013)

Start codon: ATG

End codon: TGA

Gray area: 3'UTR (17014-19590) in exon2

常见问题：

问：Promoter在DNA序列中是算内含子还是外显子？

答：都不是。属于noncoding sequence。

问：UTR在DNA序列中是算内含子还是外显子？

答：外显子。

问：起始密码子ATG在gene的哪个位置？

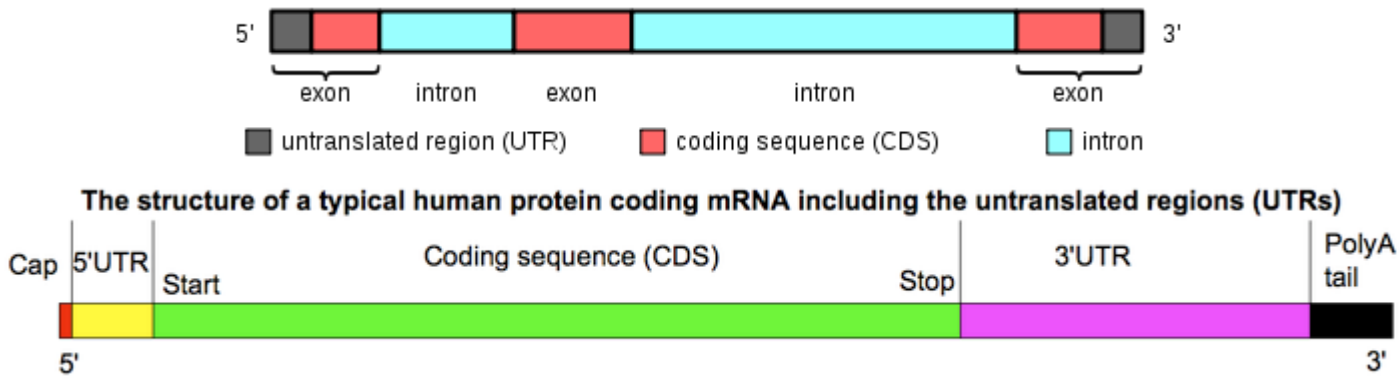
答：外显子内，CDS以ATG开始。

问：每个基因之间都有一定的间隔序列，这些基因间的间隔序列应该不属于内含子吧？

答：是的。属于noncoding sequence。

问：如何寻找启动子区域和预测转录因子结合位点？

答：见我的帖子http://www.dxy.cn/bbs/topic/23466429



目的：

- 寻找promoter区域
- 预测Transcription factor binding site

举例：

预测人基因ANKH上游2000bp启动子区域中NF-kB的结合位点

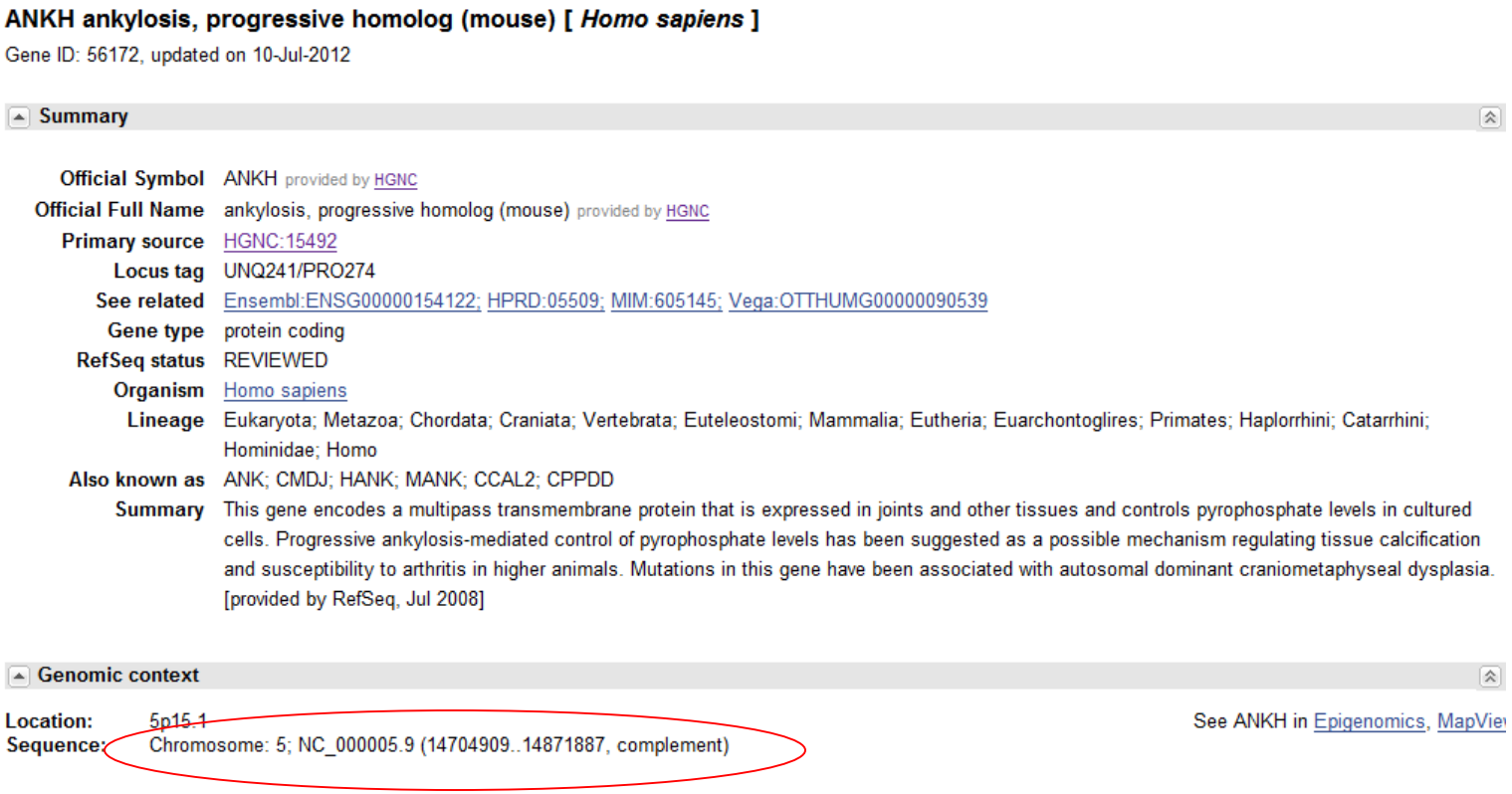
寻找promoter区域

1. 用NCBI：http://www.ncbi.nlm.nih.gov/
2. 用UCSC：http://www.genome.ucsc.edu/
3. 用Ensembl：http://www.ensembl.org/index.html
4. 用公司信息（只包含公司拥有promoter clones的信息）：http://www.genecopoeia.com/

寻找promoter区域

- NCBI ttp://www.ncbi.nlm.nih.gov/pubmed/
- 选择Gene, 输入ankh,点击search
- 选择第一项，人类Homo sapiens的ANKH
- Chromosome 5 location 14704909-14871887, complement(反义链)即-14871887 到 -14704909为基因范围

此例中选取-14873887 到-14871887 约2000bp核苷酸序列作为启动子区域



ANKH ankylosis, progressive homolog (mouse) [*Homo sapiens*]

Gene ID: 56172, updated on 10-Jul-2012

Summary

Official Symbol	ANKH provided by HGNC
Official Full Name	ankylosis, progressive homolog (mouse) provided by HGNC
Primary source	HGNC:15492
Locus tag	UNQ241/PRO274
See related	Ensembl:ENSG00000154122; HPRD:05509; MIM:605145; Vega:OTTHUMG00000090539
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	ANK; CMDJ; HANK; MANK; CCAL2; CPPDD
Summary	This gene encodes a multipass transmembrane protein that is expressed in joints and other tissues and controls pyrophosphate levels in cultured cells. Progressive ankylosis-mediated control of pyrophosphate levels has been suggested as a possible mechanism regulating tissue calcification and susceptibility to arthritis in higher animals. Mutations in this gene have been associated with autosomal dominant craniometaphyseal dysplasia. [provided by RefSeq, Jul 2008]

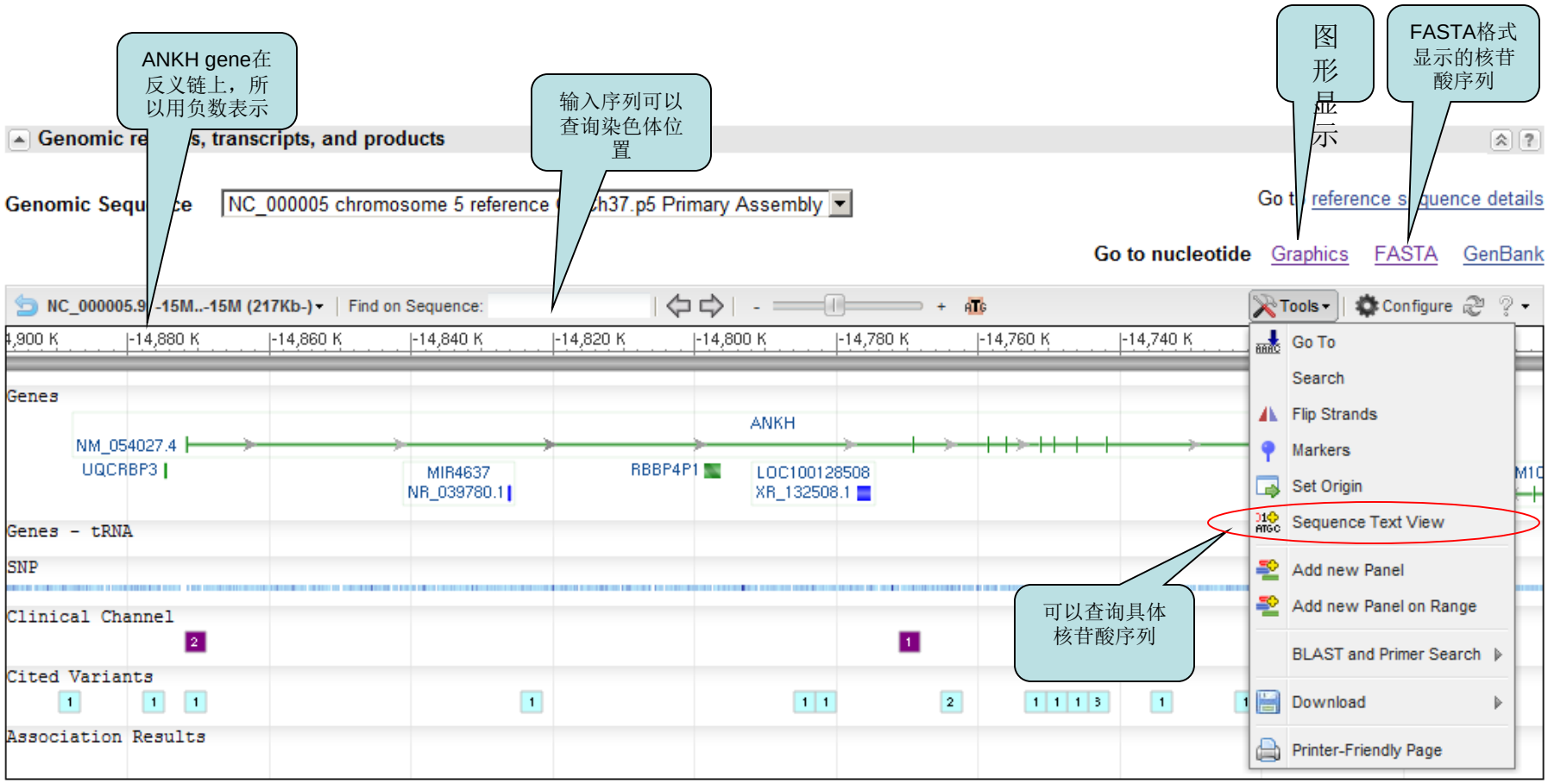
Genomic context

Location: 5p15.1
Sequence: Chromosome: 5; NC_000005.9 (14704909..14871887, complement)

See ANKH in Epigenomics, MapVie

寻找promoter区域

- 点击Graphics----Tools------Sequece Text View

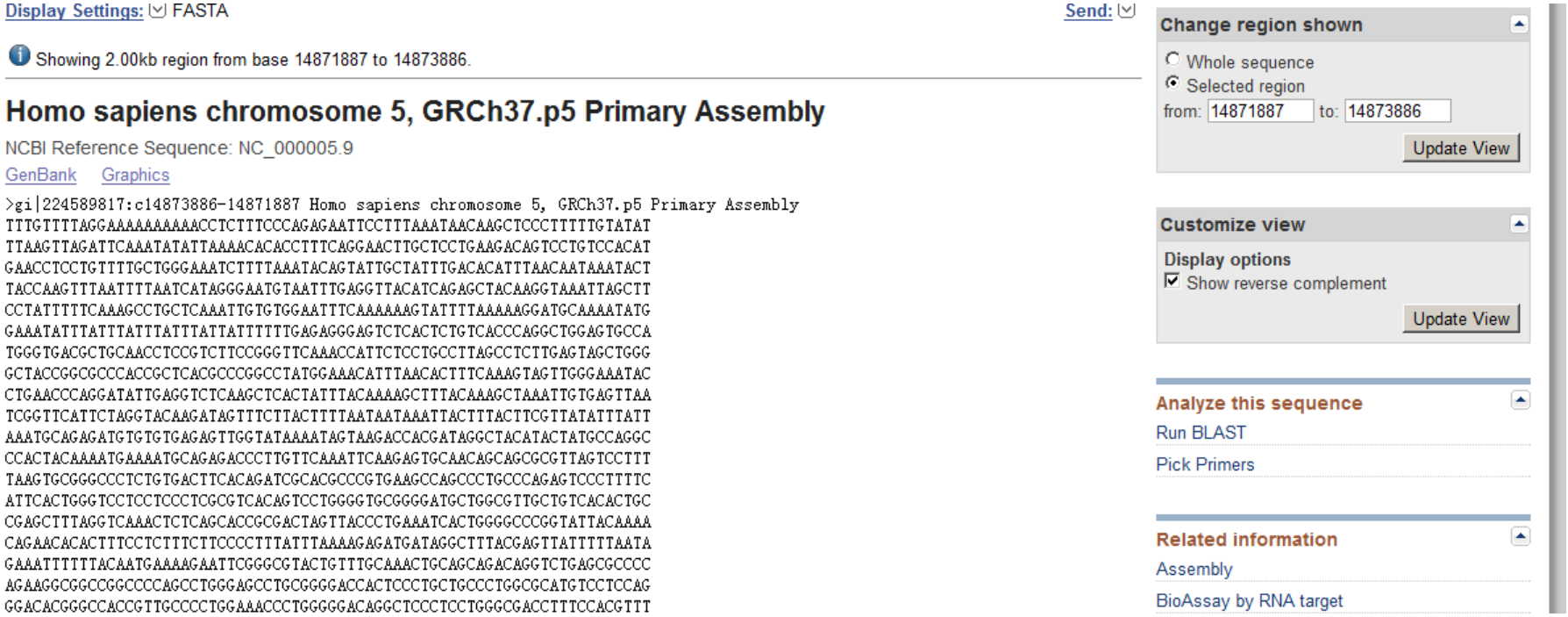


寻找promoter区域

- 点击Go To Position, 输入-14873887，点击Prev Page找到具体位置
- 复制白底黑色区域即为promoter区域。

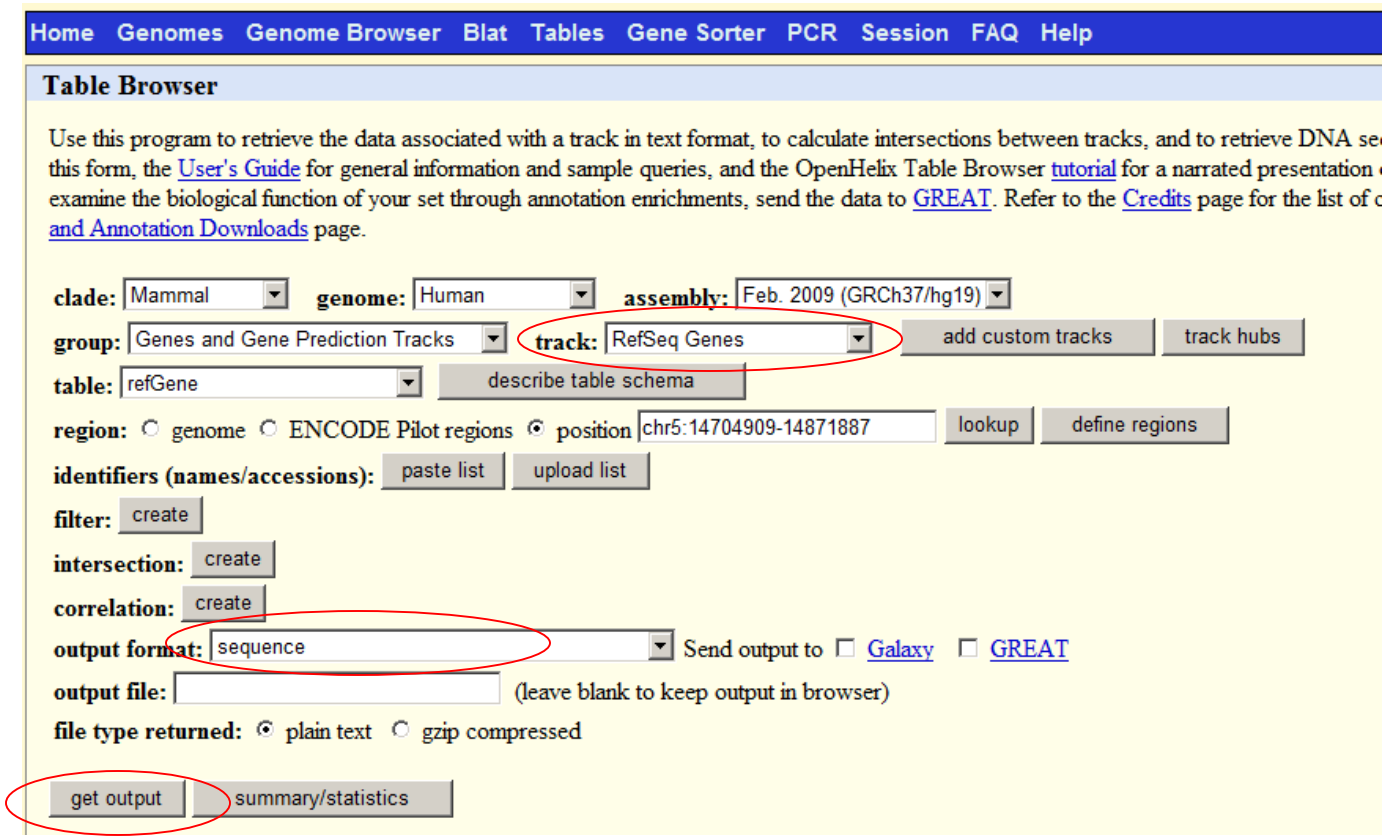
寻找promoter区域

- 在前两张幻灯片中选择FASTA
- 在右边Change region shown输入14871887到14873887
- Display options选择Show reverse complement
- 可以直接得到FASTA格式的promoter核苷酸序列（似乎有一个bp的差距，可以输入14871887到14873886 ）



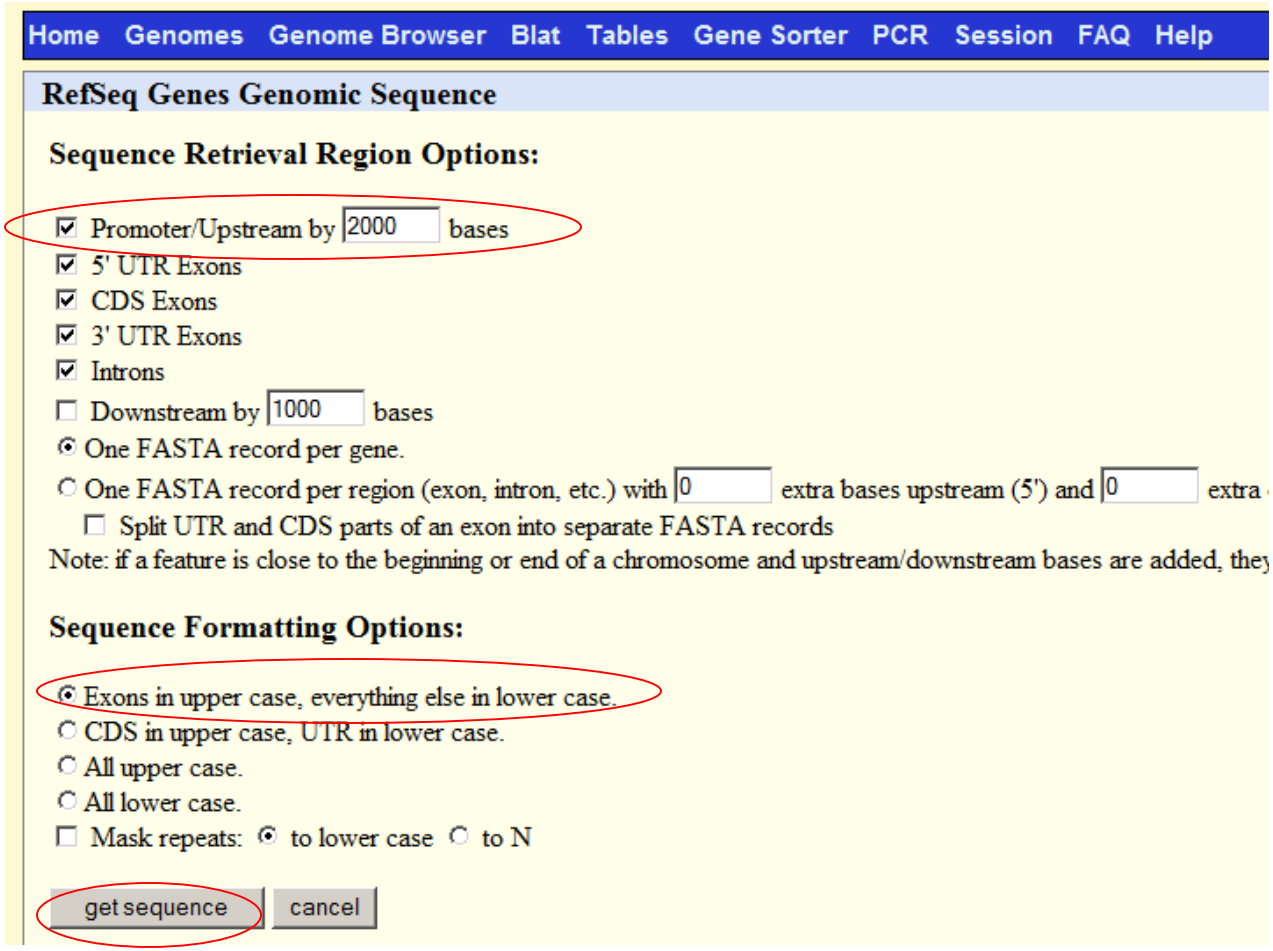
寻找promoter区域

- http://www.genome.ucsc.edu/ 选择genomes
- 在clade选择Mammal，genome选择Human，assmebly选择最新的数据库，gene中输入ANKH
- 点击Tables
- 在track中选择RefSeq Genes，在output format中选择sequence
- 点击get output。 选择genomic。



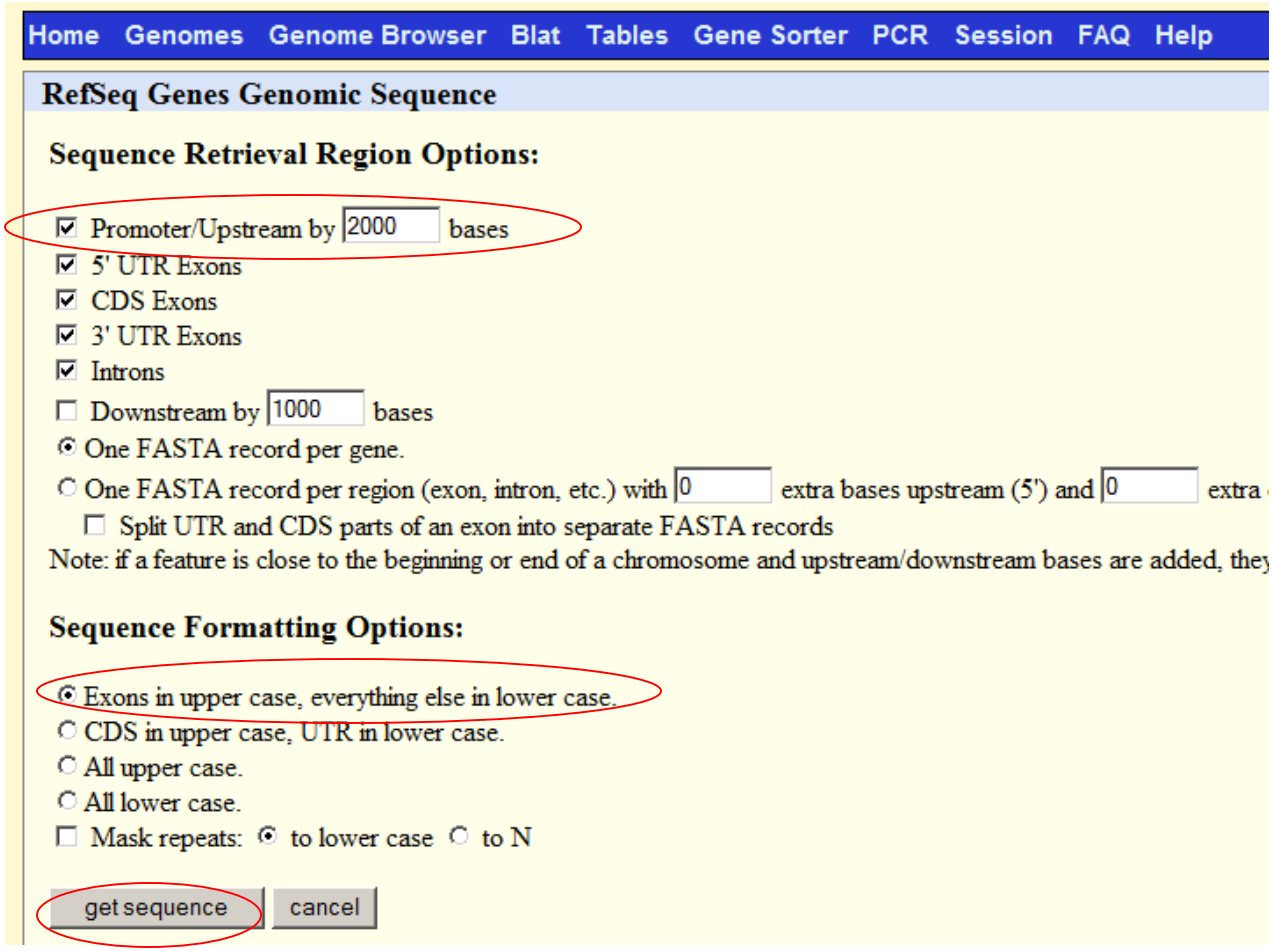
寻找promoter区域

- 选择Promoter/Upstream by 2000 bases
- Exons in upper case, everything else in lower case外显子大写，其他小写



寻找promoter区域

- 选择Promoter/Upstream by 2000 bases
- Exons in upper case, everything else in lower case外显子大写，其他小写



寻找promoter区域

>hg19_refGene_NM_054027 range=chr5:14704909-14873887 5'pad=0 3'pad=0 strand=- repeatMasking=none
atttgttttaggaaaaaaaaaacctctttcccagagaattcctttaaata
acaagctccctttttgtatatttaagttagattcaaatatattaaaacac
acctttcaggaacttgctcctgaagacagtcctgtccacatgaacctcct
gttttgctgggaaatcttttaaatacagtattgctatttgacacatttaa
caataaatacttaccaagtttaattttaatcatagggaatgtaatttgag
gttacatcagagctacaaggtaaattagcttcctatttttcaaagcctgc
tcaaattgtgtggaatttcaaaaaagtattttaaaaaggatgcaaaatat
ggaaatatttatttatttatttattatttttgagagggagtctcactct
gtcacccaggctggagtgccatgggtgacgctgcaacctccgtcttccgg
gttcaaaccattctcctgccttagcctcttgagtagctggggctaccggc
gcccaccgctcacgcccggcctatggaaacatttaacactttcaaagtag
ttgggaaatacctgaacccaggatattgaggtctcaagctcactatttac
aaaagctttacaaagctaaattgtgagttaatcggttcattctaggtaca
agatagtttcttacttttaataataaattactttacttcgttatatttat
taaatgcagagatgtgtgtgagagttggtataaaatagtaagaccacgat
aggctacatactatgccaggcccactacaaaatgaaaatgcagagaccct
tgttcaaattcaagagtgcaacagcagcgcgttagtccttttaagtgcgg
gccctctgtgacttcacagatcgcacgcccgtgaagccagccctgcccag
agtccctttcattcactgggtcctcctccctcgcgtcacagtcctgggg
tgcggggatgctggcgttgctgtcacactgccgagctttaggtcaaactc
tcagcaccgcgactagttaccctgaaatcactggggcccggtattacaaa
acagaacacactttcctctttcttccccttttatttaaaagagatgatagg
ctttacgagttatttttaatagaaatttttttacaatgaaaagaattcggg
cgtactgtttgcaaactgcagcagacaggtctgagcgcccccagaaggcgg
ccggccccagcctgggagcctgcggggaccactccctgctgccctggcgc
atgtcctccagggacacgggccaccgttgcccctggaaaccctgggggac
aggctccctcctgggcgacctttccacgtttctttcgccccccccggagc
ggcacctccatacctcccgggccgctagtcccccgaagcaggccccagcc
ccggcgcaagctgcctcctccgcagtccccgcccgccgccagcccggcca
ccgccggacggcgcccttcccagccccgcgccccgctccccagccctcct
cgcgcgtccgagcccacgtcgcttgccccggctgccaagtcttccagacg
gccgggccctcccggcacagcgcgggaaccccggccccccggccccacggg
acgagggctacagggcccggccggcgagccagggcaccccggggtctcca
ggcggcccaccgccctcaccccccaccgggcgtccccaagccccccagaccg
tccccgccgcgccctcagcgccgccgccccccgggccggccgaggctccgc
ggcgagtcccgcccagcgccccggaggcgccagccccacggcccgagcgt
gcgcagcgcccccccgcggccgcgccaagcgcaggcgacggcacaggaaag
gaggccgcggcgcgcccggcccggcccccctccccagcccgccccc
cgctggcggtgtccgcgcctccgcgcccgccccctgatttcctccgggcgg
cggcggcggcggcggcgcctcgtgcgcgcggttatttatttccgggcccc
AGAGCGCTTATAATGGAGCCGCTGTCAGCAGAACCTTCTGCCGCCGCCGC
CGCCGCCGCCGTCCCTCCTCTTTTTTTCCCGGCAGATCTTTGTTGTGTG
GGAGGGCAGCAGGGATGGACTTGAGCTTGCGGATCCCCTGCTAGAGCAGC
CGCGCTCGGAGAAGGCGCCGCAGCCGCGAGGAGGAGCCGCCGCCGCCGCG
CCCGAGGCCCCGCCGCCCGCGGCCTCTGTCGGCCCGCGCCCCGCTCGCCC
CGTCGCCCCGTCGCCCCTCGCCTCCCCGCAGAGTCCCCTCGCGGCAGCAG
ATGTGTGTGGGGTCAGCCCACGGCGGGGACTATGGTGAAATTCCCGGCGC
TCACGCACTACTGGCCCCTGATCCGGTTCTTGGTGCCCCTGGGCATCACC
AACATAGCCATCGACTTCGGGGAGCAGgtaagccccggccccgcgcccca

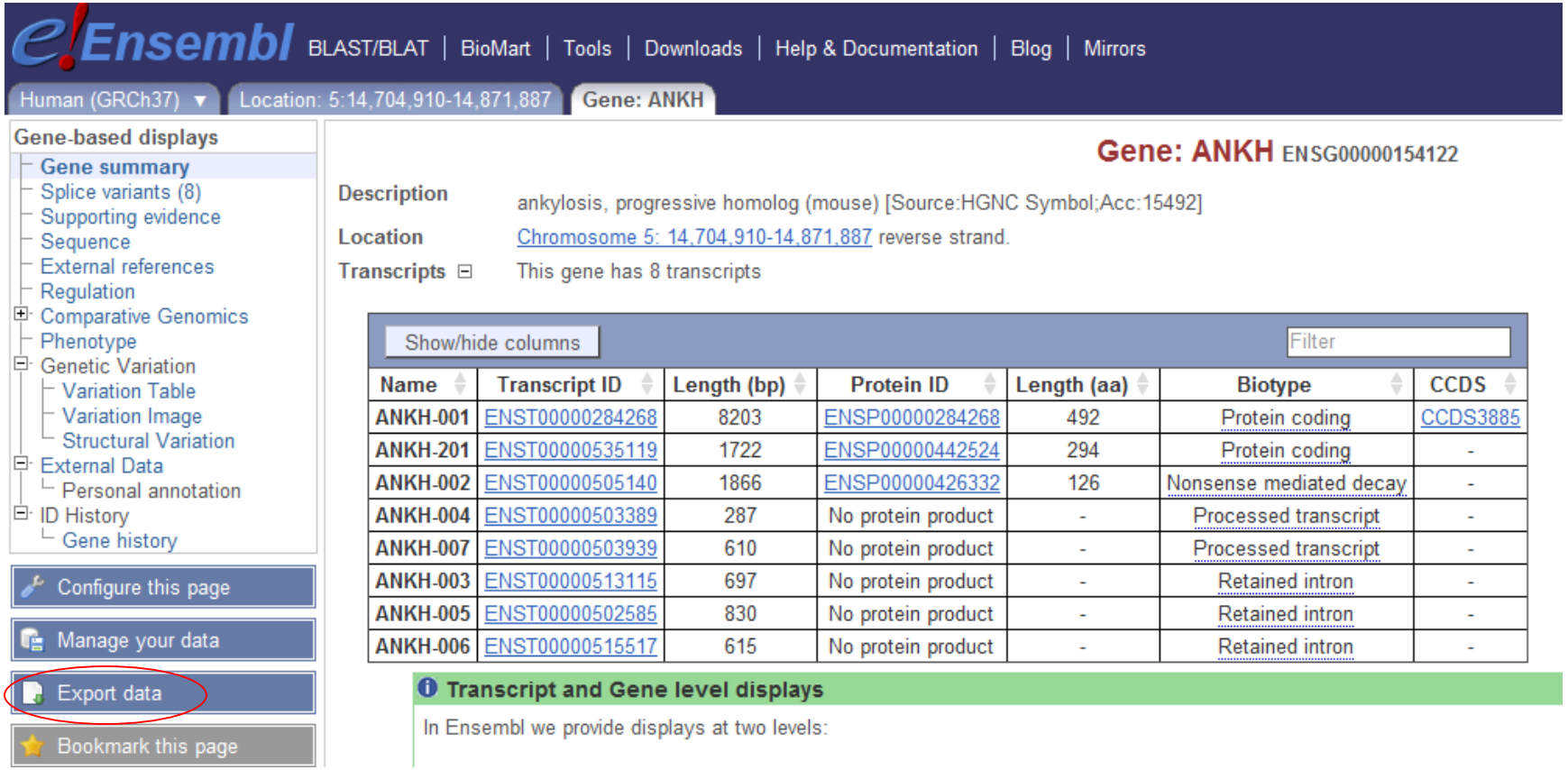
小写字母为promoter区域

大写字母为基因区域，与NCBI结果相同

ATG为CDS区起始密码子

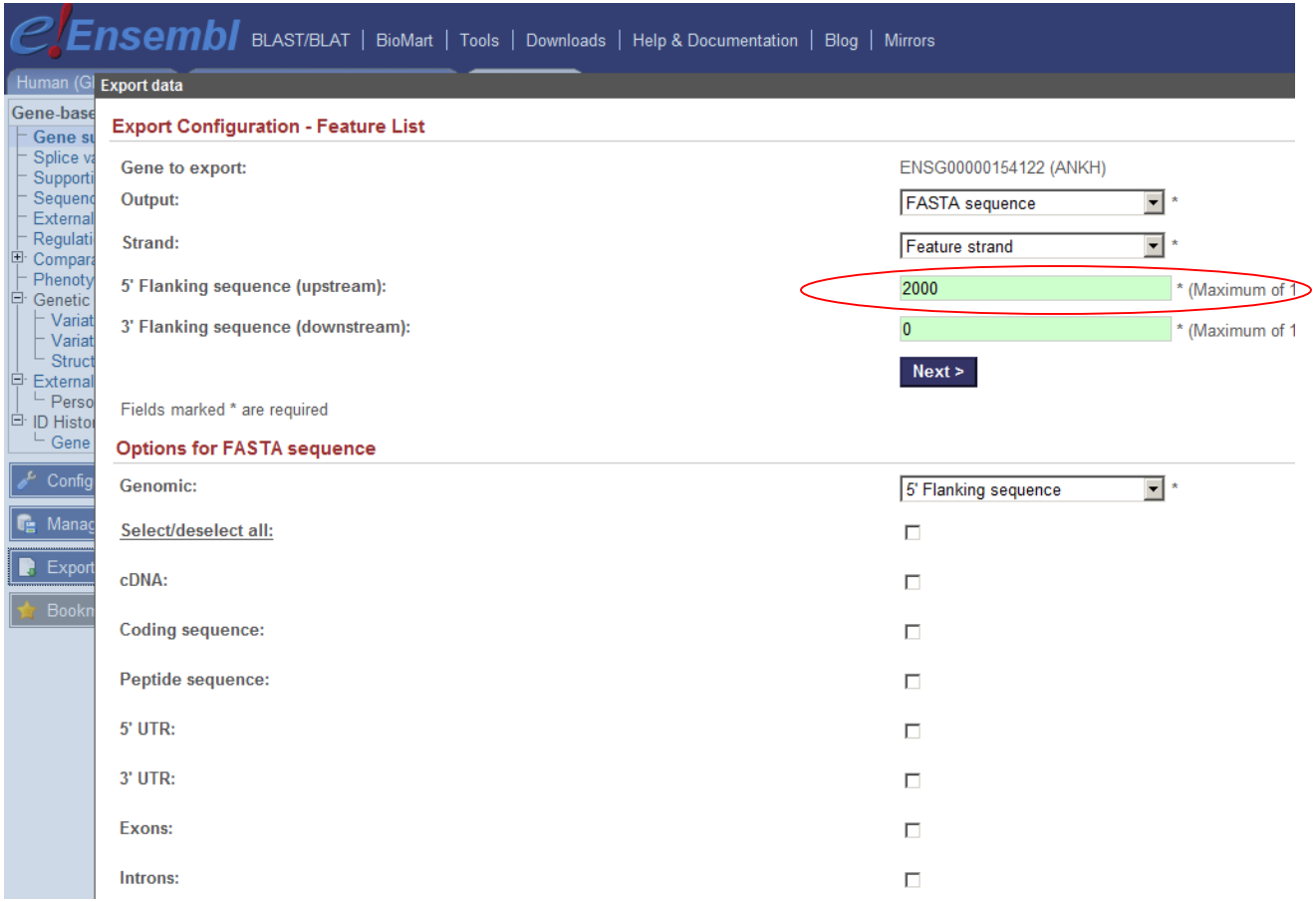
寻找promoter区域

- http://www.ensembl.org/index.html 选择human 输入 ankh
- 选择Gene，点击 GeneID ENSG00000154122
- 点击左边的Export data



寻找promoter区域

- 5 Flanking sequence 输入2000
- Options for FASTA sequence中Genomic选5 Flanking sequence，
- deselect all
- 点击Next



寻找promoter区域

- 得到2000 bases 的核苷酸序列

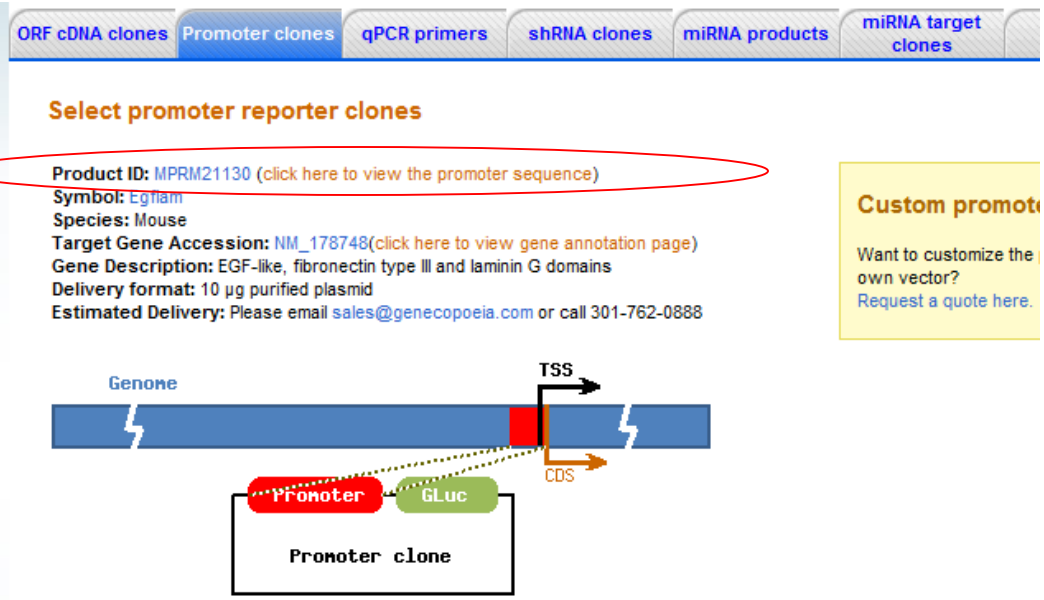
>5' Flanking sequence chromosome:GRCh37:5:14871888:14873887:-1
ATTTGTTTTAGGAAAAAAAAAACCTCTTTCCCAGAGAATTCCTTTAAATAACAAGCTCCC
TTTTTGTATATTTAAGTTAGATTCAAATATATTAAAACACACCTTTCAGGAACTTGCTCC
TGAAGACAGTCCTGTCCACATGAACCTCCTGTTTTGCTGGGAAATCTTTTAAATACAGTA
TTGCTATTTGACACATTTAACAATAAATACTTACCAAGTTTAATTTTAATCATAGGGAAT
GTAATTTGAGGTTACATCAGAGCTACAAGGTAAATTAGCTTCCTATTTTTCAAAGCCTGC
TCAAATTGTGTGGAATTTCAAAAAAGTATTTTAAAAAGGATGCAAAATATGGAAATATTT
ATTTATTTATTTATTATTTTTTGAGAGGGAGTCTCACTCTGTCACCCAGGCTGGAGTGCC
ATGGGTGACGCTGCAACCTCCGTCTTCCGGGTTCAAACCATTCTCCTGCCTTAGCCTCTT
GAGTAGCTGGGGCTACCGGCGCCCACCGCTCACGCCCGGCCTATGGAAACATTTAACACT
TTCAAAGTAGTTGGGAAATACCTGAACCCAGGATATTGAGGTCTCAAGCTCACTATTTAC
AAAAGCTTTACAAAGCTAAATTGTGAGTTAATCGGTTCATTCTAGGTACAAGATAGTTTC
TTACTTTTAATAATAAATTACTTTACTTCGTTATATTTATTAAATGCAGAGATGTGTGTG
AGAGTTGGTATAAAATAGTAAGACCACGATAGGCTACATACTATGCCAGGCCCACTACAA
AATGAAAATGCAGAGACCCTTGTTCAAATTCAAGAGTGCAACAGCAGCGCGTTAGTCCTT
TTAAGTGCGGGCCCTCTGTGACTTCACAGATCGCACGCCCGTGAAGCCAGCCCTGCCCAG
AGTCCCTTTTCATTCACTGGGTCCTCCTCCCTCGCGTCACAGTCCTGGGGTGCGGGGATG
CTGGCGTTGCTGTCACACTGCCGAGCTTTAGGTCAAACTCTCAGCACCGCGACTAGTTAC
CCTGAAATCACTGGGGCCCGGTATTACAAAACAGAACACACTTTCCTCTTTCTTCCCCTT
TATTTAAAAGAGATGATAGGCTTTACGAGTTATTTTTAATAGAAATTTTTTACAATGAAA
AGAATTCGGGCGTACTGTTTGCAAACTGCAGCAGACAGGTCTGAGCGCCCCAGAAGGCGG
CCGGCCCCAGCCTGGGAGCCTGCGGGGACCACTCCCTGCTGCCCTGGCGCATGTCCTCCA
GGGACACGGGCCACCGTTGCCCCTGGAAACCCTGGGGGACAGGCTCCCTCCTGGGCGACC
TTTCCACGTTTCTTTCGCCCCCCCCGGAGCGGCACCTCCATACCTCCCGGGCCGCTAGTC
CCCCGAAGCAGGCCCCAGCCCCGGCGCAAGCTGCCTCCTCCGCAGTCCCCGCCCGCCGCC
AGCCCGGCCACCGCCGGACGGCGCCCTTCCCAGCCCCGCGCCCCGCTCCCCAGCCCTCCT
CGCGCGTCCGAGCCCACGTCGCTTGCCCCGGCTGCCAAGTCTTCCAGACGGCCGGGCCCT
CCCGGCACAGCGCGGGAACCCCGGCCCCCGGCCCCACGGGACGAGGGCTACAGGGCCCGG
CCGGCGAGCCAGGGCACCCCGGGGTCTCCAGGCGGCCCACCGCCCTCACCCCCCACCGGC
GTCCCCAAGCCCCCAGACCGTCCCCGCCGCGCCCTCAGCGCCGCCGCCCCCGGGCCGGCC
GAGGCTCCGCGGCGAGTCCCGCCCAGCGCCCCGGAGGCGCCAGCCCCACGGCCCGAGCGT
GCGCAGCGCCCCCCGCGGCCGCGCCAAGCGCAGGCGACGGCACAGGAAAGGAGGCCGCGG
CGCGCCCGGCCCGGCCCCCTCCCCAGCCCGCCCCCGGGGCCGCTGGCGGTGTCCGCGCCT
CCGCGCCCGCCCCTGATTTCCTCCGGGCGGCGGCGGCGGCGGCGGCGCCTCGTGCGCGCG
GTTATTTATTTCCGGGCCCC

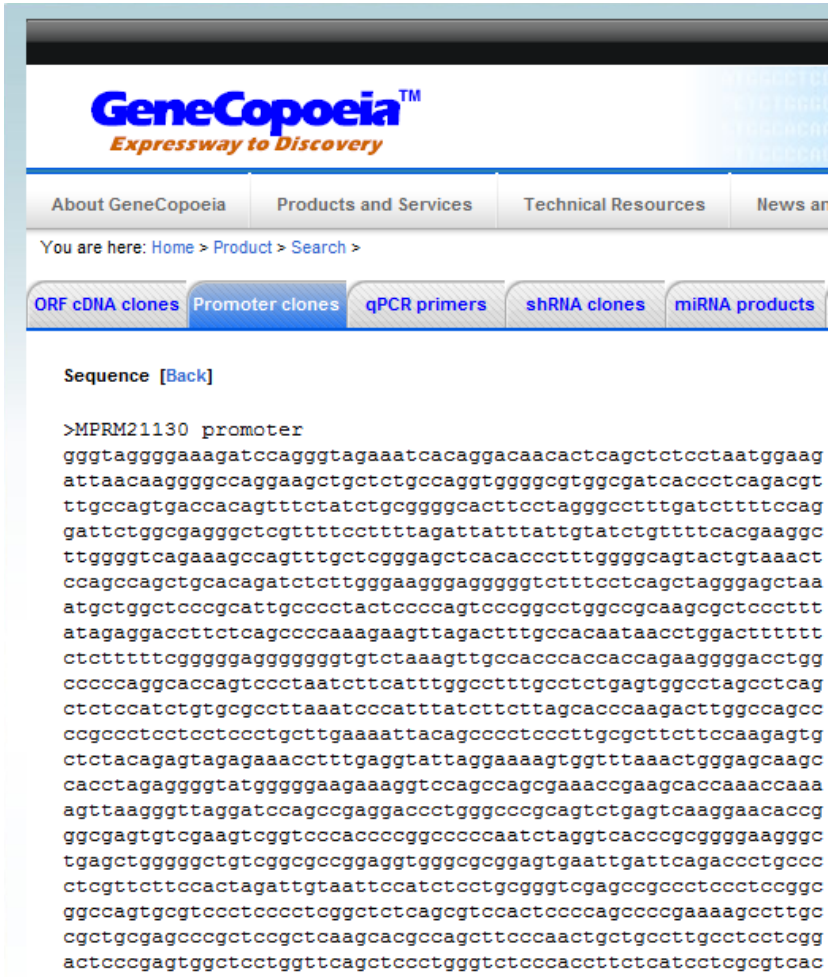
寻找promoter区域

- http://www.genecopoeia.com/
- 点击search product， 选择promoter clones，因为没有ANKH的信息，此处输入FIBRONECTIN
- 选择目的基因

寻找promoter区域

- 点击click here to view the promoter sequence
- 得到promoter信息





预测Transcription factor binding site

1. 用Jaspar http://jaspar.genereg.net/
2. 用PROMO http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
3. 用TFSEARCH（据说用的是TRANSFAC很旧的数据库）
   http://www.cbrc.jp/research/db/TFSEARCH.html
4. 用商业数据库TRANSFAC（要付费）
   http://www.gene-regulation.com/pub/databases.html/

预测Transcription factor binding site

http://jaspar.genereg.net/ 点击JASPAR CORE vetebrata

左边转录因子选择MA0061.1 NF-kappaB，右边输入ANKH启动子区域，点击SCAN

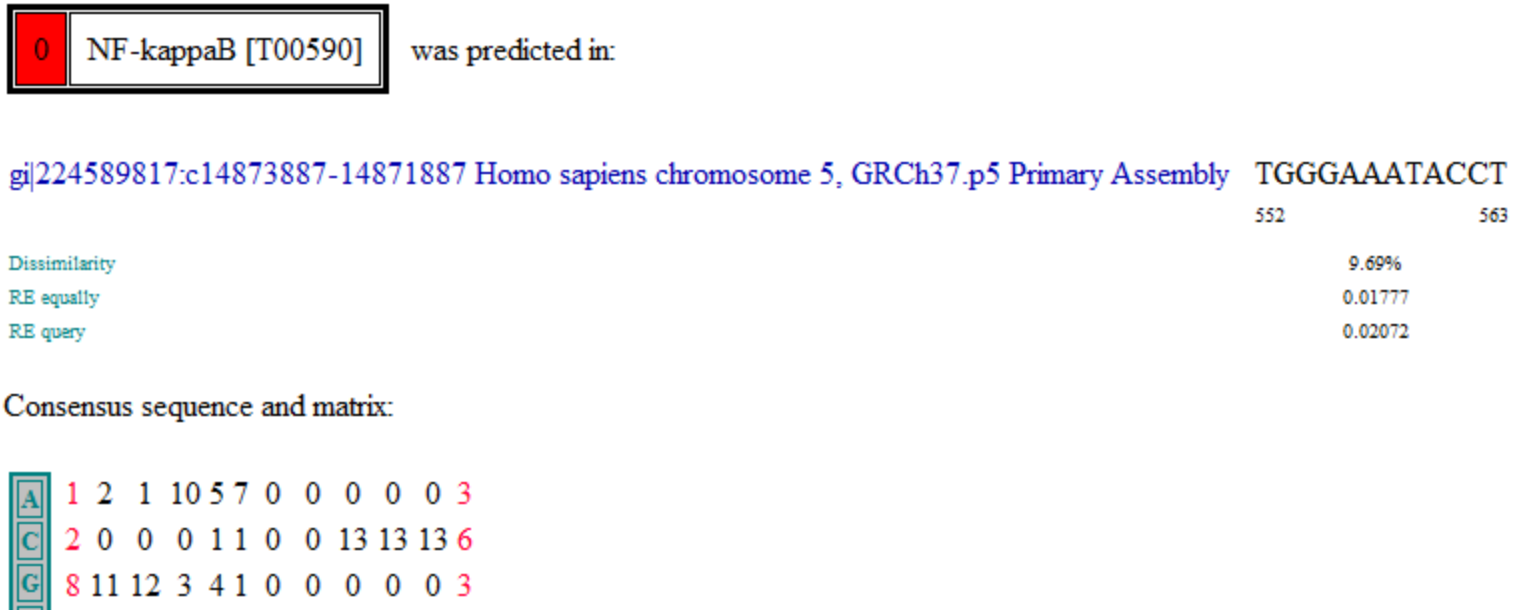
结果得到5个 Transcription factor binding site， 其中Strand -1没有特殊意义，另外三个GGGAAATACC得分最高

5 putative sites were predicted with these settings (80%) in sequence named gi|224589817:c14873887-14871887

Model ID	Model name	Score	Relative score	Start	End	Strand	predicted site sequence
MA0061.1	NF-kappaB	6.814	0.800574249111552	34	43	1	GAGAATTCCT
MA0061.1	NF-kappaB	8.606	0.843971565747803	553	562	-1	GGTATTTCCC
MA0061.1	NF-kappaB	10.772	0.896426134851665	553	562	1	GGGAAATACC
MA0061.1	NF-kappaB	6.807	0.800404728343442	1554	1563	1	GGGCCCTCCC
MA0061.1	NF-kappaB	7.528	0.817865367458808	1751	1760	-1	GGGACTCGCC

预测Transcription factor binding site

http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
Step1 selectspecies选择human
Step1 SelectFactors选择NF-kappaB [T00590]
Step2 SearchSites输入ANKH的promoter区域
结果中有一个位点TGGGAAATACCT，与JASPAR结果中得分最高的相同



预测Transcription factor binding site

http://www.cbrc.jp/research/db/TFSEARCH.html
Enter a label for the sequence: 输入基因名字ANKH
Enter your DNA sequence 输入ANKH的promoter区域
点击submit
结果中有2个NF-kap位点，其中正义链GGGAAATACC，与JASPAR结果中得分最高的相同

```
551 CGCTCACGCC CGGCCTATGG AAACATTTAA CACTTTCAAA GTAGTTGGGA entry        score
                                                --------  M00087 Ik-2   88.6
                                                   -----  M00120 dl     88.2
                                                    ----  M00054 NF-kap 88.2
                                                    <--   M00053 c-Rel  87.6
                         ----->                           M00029 HSF    86.9
                                                     ---  M00029 HSF    86.9
                                                ---------  M00109 C/EBPb 86.2
                                                    <--   M00052 NF-kap 86.1
                         ----->                           M00028 HSF    85.9
                                                     ---  M00028 HSF    85.9
                                                  -------  M00141 Lyf-1  85.7
    <---------------                                      M00033 p300   85.1

601 AATACCTGAA CCCAGGATAT TGAGGTCTCA AGCTCACTAT TTACAAAAGC entry        score
                                       -------->          M00253 cap    92.0
                  ---------->                             M00076 GATA-2 90.1
                            ----->                        M00271 AML-1a 88.7
    --->                                                  M00087 Ik-2   88.6
    ----->                                                M00120 dl     88.2
    ----->                                                M00054 NF-kap 88.2
                                              ------->    M00101 CdxA   87.9
    -------                                               M00053 c-Rel  87.6
                                              ------->    M00100 CdxA   87.2
    ->                                                    M00029 HSF    86.9
                     <--------                            M00253 cap    86.5
                            -------->                     M00211 Poly A 86.4
          ----->                                          M00029 HSF    86.3
    ---->                                                 M00109 C/EBPb 86.2
    -------                                               M00052 NF-kap 86.1
    ->                                                    M00028 HSF    85.9
              --------------->                            M00074 c-Ets- 85.8
    ->                                                    M00141 Lyf-1  85.7
                  ---------->                             M00075 GATA-1 85.7
```