Maf Family TF Analysis Based on RNA-seq of Developmental Liver and Pancreas

Reporter: Yuwei Zhang

Group Members: Yingla Zhang Yongjun Li

Yitao Zheng Yuwei Zhang

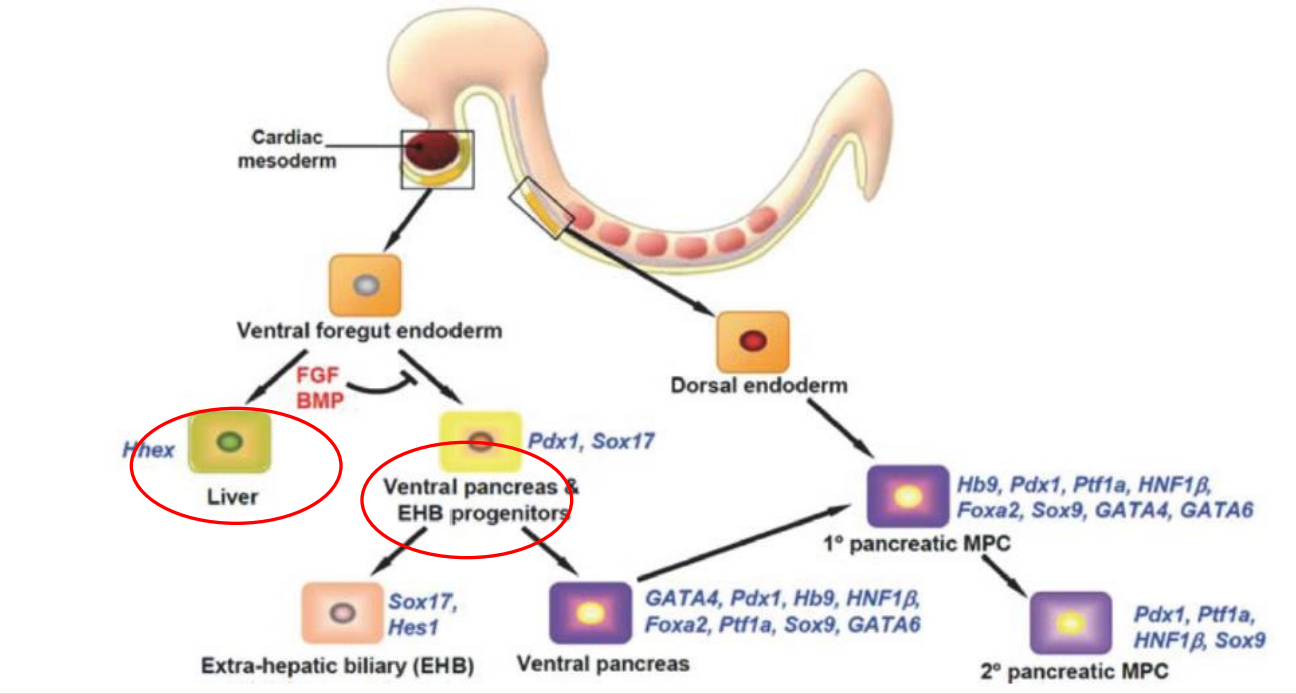
Outline

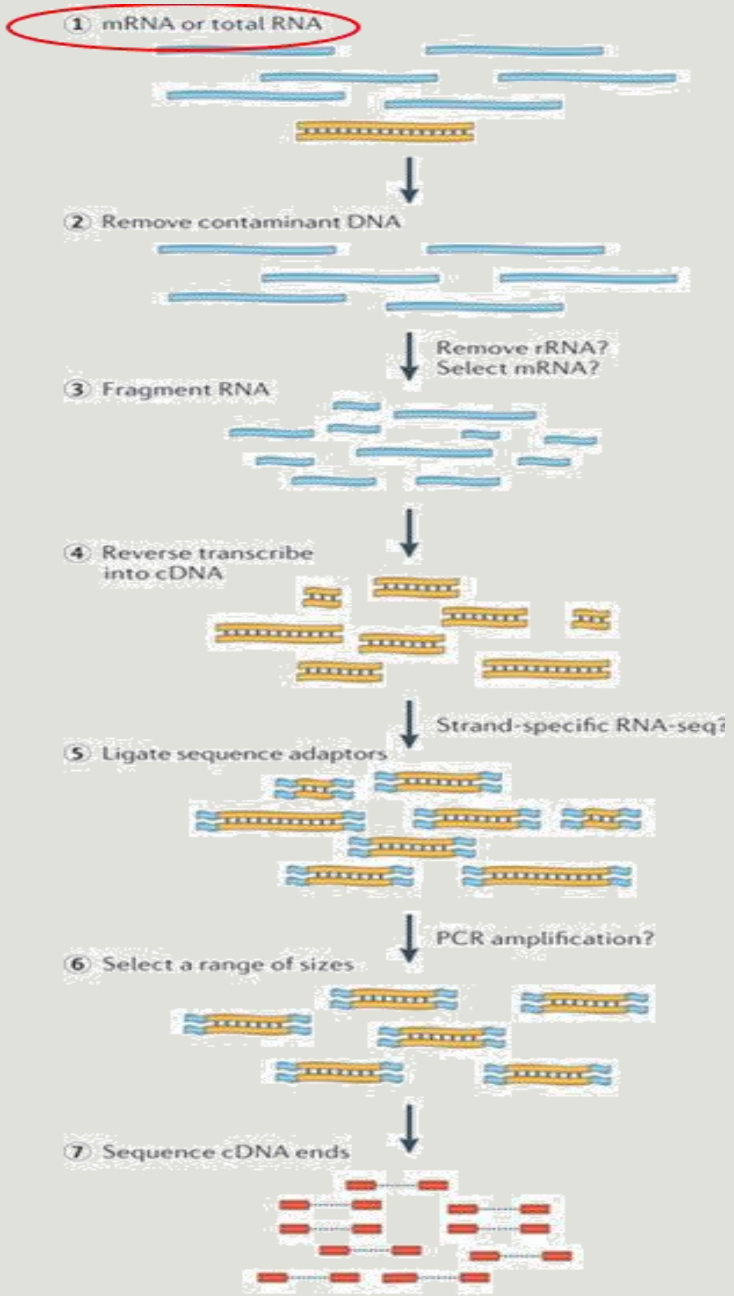
- RNA-seq of Developmental Liver and pancreas
- Data processing
- Maf family transcription factors
- The evolution & structure of Mafa &Mafb

RNA-seq of Developmental Liver and pancreas

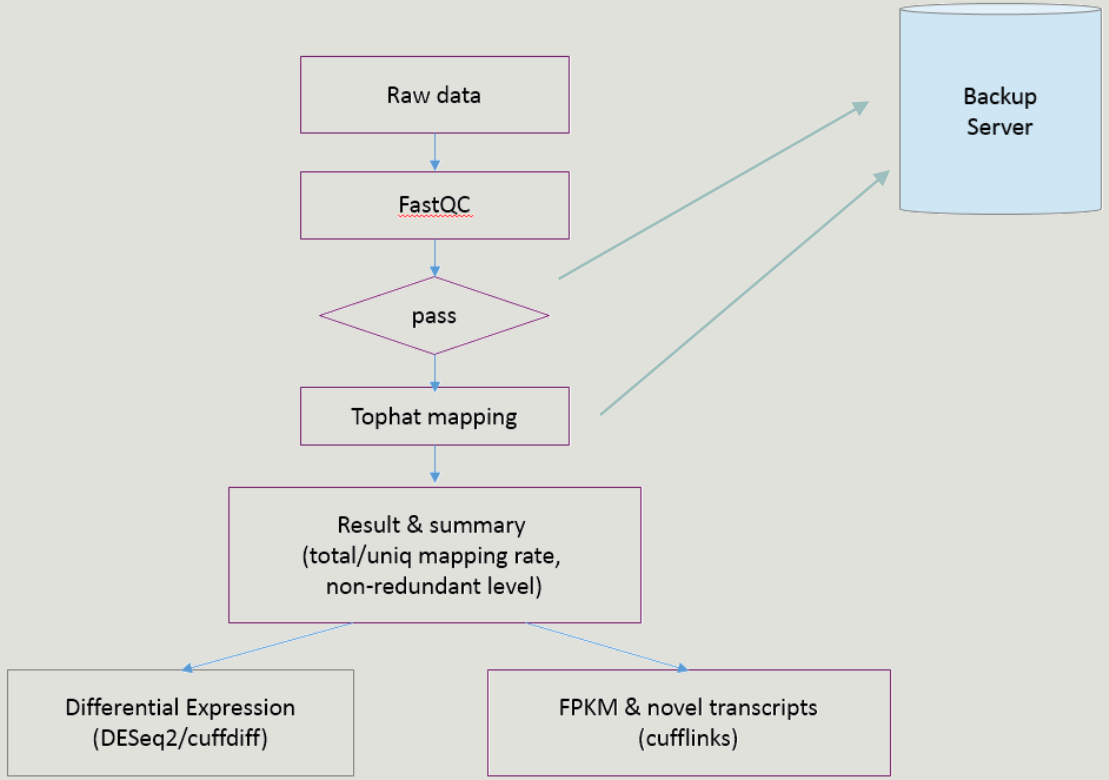
I. Liver and pancreas arise from endoderm at E10.5.

II. RNA-seq could detect the differential transcriptom of two organs.

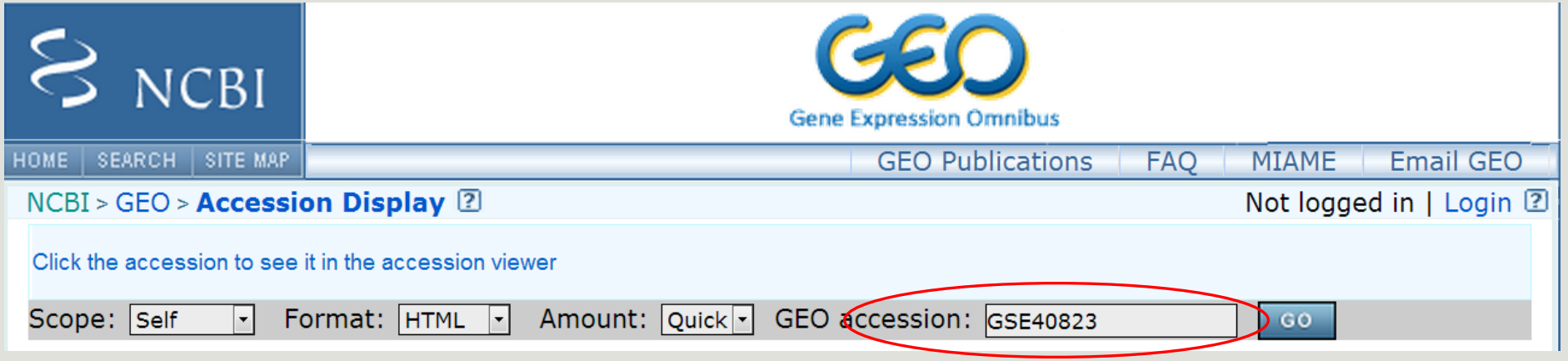




Data processing



Data Processing Step I



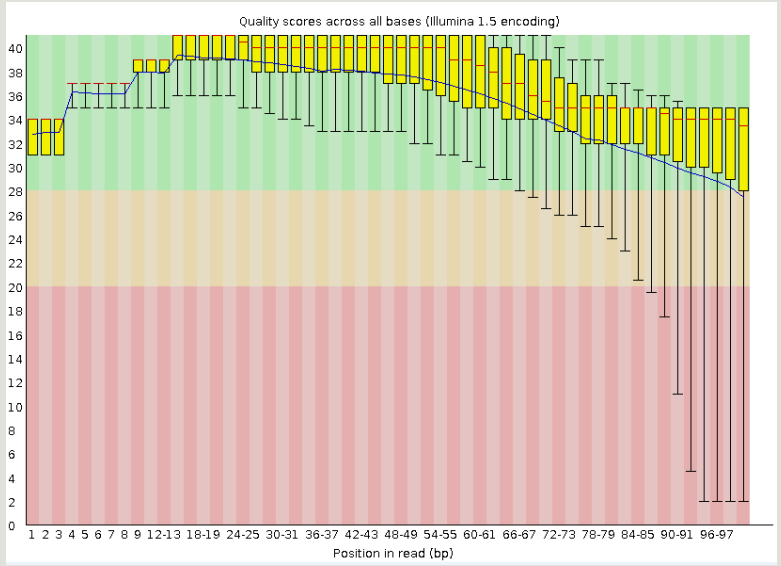
/sra/sra-instant/reads/ByExp/sra/SRX/SRX186/SRX186174/SRR567652/ 的索引

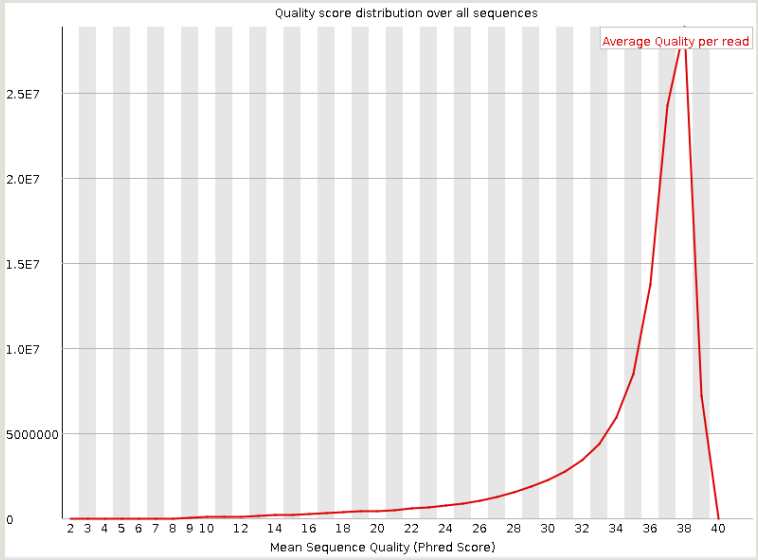
名称	大小	修改日期
[上级目录]		
SRR567652.sra	15.2 GB	12-9-12 上午12:00:00

```
1. Unzip the sra file

fastq-dump --split-3 SRR567654.sra
```

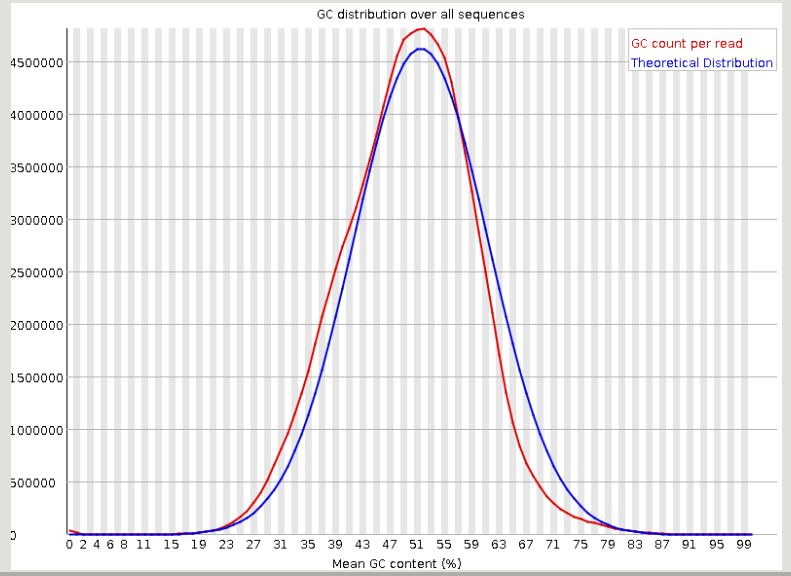
Data Processing Step II quality control

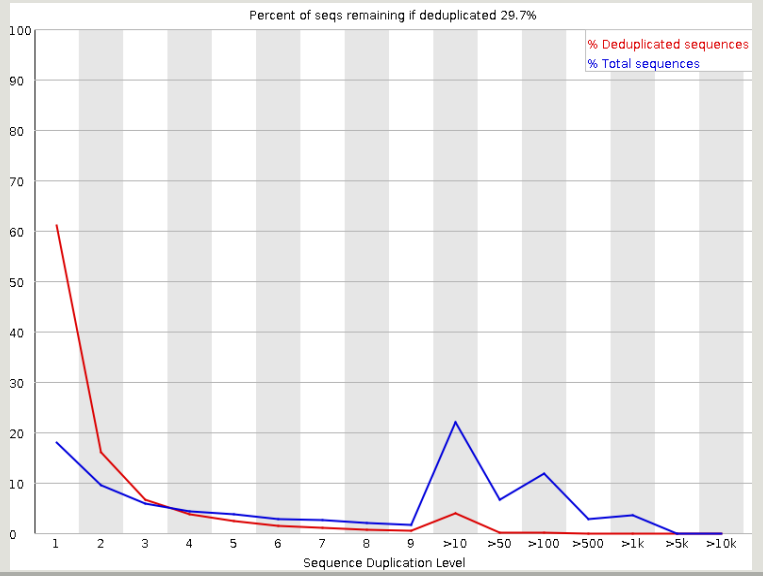
```
2. FastQC

fastqc --noextract   SRR567654_1.fastq
fastqc --noextract   SRR567654_2.fastq
```





score=-10log$_{10}$ (sequenceing error rate)





Data Processing Step III　Mapping

```
zhangyw@yp:~$ tophat2 /usr/local/data/bowtie2-index/mm9 SRR567654_1.fastq SRR567654_2.fastq

[2014-06-11 09:40:20] Beginning TopHat run (v2.0.11)
-----------------------------------------------
[2014-06-11 09:40:20] Checking for Bowtie
                  Bowtie version:        2.2.2.0
[2014-06-11 09:40:20] Checking for Samtools
                Samtools version:        0.1.18.0
[2014-06-11 09:40:20] Checking for Bowtie index files (genome)..
[2014-06-11 09:40:20] Checking for reference FASTA file
[2014-06-11 09:40:20] Generating SAM header for /usr/local/data/bowtie2-index/mm9
[2014-06-11 09:40:38] Preparing reads
```

	A	B	C	D	E	F
1	Sample	Total_read	Non-uniq	Uniq	Mapping Efficiency	Unique Mapping Efficiency
2	E10.5_DP	2.33E+8	1.79E+7	1.72E+8	81.74%	74.04%
3	E10.5_VP	2.40E+8	1.81E+7	1.77E+8	81.25%	73.75%
4	E10.5_LV	2.65E+8	2.38E+7	1.86E+8	79.25%	70.28%

Mapping quality
score=-10log$_{10}$ (mapping error rate)
non-unique mapping

Number of hits

```
yup@ngs ~/work/RNA-seq/rawData/tophat_SRR567654_2 $ samtools view accepted_hits.bam |more
SRR567654.17949297      433     chr1    3005262 0       101M    chr16   36976948        0       TGTATGCAATGGTGTCAGCATTTGGAAGCTGATTATGGGATGGATCCCTGCATATGGCAATCACTAGATGGTCCATCCTTTTGTCACAGCTCCAAATTTTG   ccccbccc`caccddddeeeeeeggg
giiiiiiiiiiiiiiihhghiiiiiiiiiihhiiiihiiiihiiiiiiihhiiggggeeeeeba_        AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU NH:i:20 CC:Z:chr10      CP:i:66424155   HI:i:0
SRR567654.51681684      401     chr1    3005262 0       101M    chr4    126741935       0       TGTATGCAATGGTGTCAGCATTTGGAAGCTGATTATGGGATGGATCCCTGCATATGGCAATCACTAGATGGTCCATCCTTTTGTCACAGCTCCAAATTTTG   cccdcccccccddcddceeeeeeggg
giiiiiiiihihhhhhhihiiiiihiiiihhfiiiiiiihiiigihiiiiiigiiggggeeeeebbb        AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU NH:i:20 CC:Z:chr10      CP:i:101688779  HI:i:0
SRR567654.49640892      321     chr1    3006356 0       101M    chr7    136493492       0       CCATATCTTCGAGGCTTTTCCCTACTTTCTCCTCTGTAAGTTTCAGTGTCTCTGGTTTTATGTGGAGTTCCTTAATCCACTTAGATTTGACCTTAGTACAA   bbbeeeeeggggggiiiiiiiiiiiii
iiiiiihhhiiiiihiihiiiiiiihiiihiiihihiiiiiihhgggggeeeeeeedcddddcbcdccc        AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU NH:i:8  CC:Z:=  CP:i:97849347   HI:i:0
SRR567654.50913589      129     chr1    3007009 0       101M    =       105435855       102428947       CGCCCATGTATTTTATATTATTTGTGACTATTGAGAAGGGTGTTGTTTCCCTAATTTCTTTCTCAGCCTGTTTATCCTTTGTGTACAGAAAGGCCATTGAC   bbbeeeeegfgggifhiii
iihghighghiihihiiiicghiiiiiiihiiiiiiiiiiiiiiiiiiihigiiiiiggggggeeeeeebdddccccccc        AS:i:-6 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:4A96       YT:Z:UU NH:i:20 CC:Z:=  CP:i:3007009    HI:i:0
SRR567654.50913589      385     chr1    3007009 0       101M    chr12   48609544        0       CGCCCATGTATTTTATATTATTTGTGACTATTGAGAAGGGTGTTGTTTCCCTAATTTCTTTCTCAGCCTGTTTATCCTTTGTGTACAGAAAGGCCATTGAC   bbbeeeeegfgggifhiiiiiiihghi
ghiihihiiiicghiiiiiiihiiiiiiiiiiiiiiiiihigiiiiiiggggggeeeeeebdddccccccc        AS:i:-6 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:4A96       YT:Z:UU NH:i:20 CC:Z:=  CP:i:3007009    HI:i:1
```

Unique mapping

```
yup@ngs ~/work/RNA-seq/rawData/tophat_SRR567654_2 $ samtools view accepted_hits.bam |awk '($5 >20)' |more
SRR567654.88689488      163     chr1    3143955 50      101M    =       3144036 182     AAAAAACTAAATTAAATTCATGTTTTAGATCCATCCTTACTTGTCATTTTTCCAGTTTAGACTAGCTTCTAGCCTTTTAACTTTATGGCAATAGTACATCA   bbbeeeeeggggggiiiiiii
iiiiiihhhiiiiiiiiihiiihiiiiiiiiiiihhhiiggfgggeeeeeecdddcdcbbcdbc        AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU NH:i:1
SRR567654.58921549      99      chr1    3144016 50      101M    =       3144024 109     CTAGCTTCTAGCCTTTTAACTTTATGGCAATAGTACATCAGAGACTGTATATTCAGACTTAGTAAAATTAGTCATTTAATAGAGTCATAATGATTTTTCTC   bbbeeeeeggggggiiiiidh
iiiiiihhiiiiiiiiiiiiiiiiiiiiiiiihhhiihihhihifhfgfgggggeeeeeeddd        AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU NH:i:1
SRR567654.58921549      147     chr1    3144024 50      101M    =       3144016 -109    TAGCCTTTTAACTTTATGGCAATAGTACATCAGAGACTGTATATTCAGACTTAGTAAAATTAGTCATTTAATAGAGTCATAATGATTTTTCTCCTTTCTTC   ddcceeeeddaggggggiii
iiiiiiiiiihhiiiiiiiiiihiiihhiiiiiiiiiiiiiiiiiiiiiiggggeeeeeb_b        AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU NH:i:1
SRR567654.57184711      99      chr1    3144026 50      101M    =       3144045 120     GCCTTTTAACTTTATGGCAATAGTACATCAGAGACTGTATATTCAGACTTAGTAAAATTAGTCATTTAATAGAGTCATAATGATTTTTCTCCTTTCTTCAG   a__eeeeggggghifgfeg
ffhhiidfghiiiiihiiihhbfgfhbhhifeggfdffgfghfh]bgfiig``dZbdecd_]b_        AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU NH:i:1
SRR567654.88689488      83      chr1    3144036 50      101M    =       3143955 -182    TTTATGGCAATAGTACATCAGAGACTGTATATTCAGACTTAGTAAAATTAGTCATTTAATAGAGTCATAATGATTTTTCTCCTTTCTTCAGTGTGACCAGC   dbddddeeeeeeeggggggg
ifhhiiihiihiiiihiiiiiiiiiiiiiiiiihihiiiiihiiiiiiiiggggeeeeeebbb        AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:101        YT:Z:UU NH:i:1
```

Data Processing Step IV　Differential expression analysis

Count the reads of each gene using htseq-count with gene structure annotation：
Ensembl mm9 （downloaded from UCSC）

Command :

```
zhangyw@yp:/home/yp/RNA-seq/tophat_SRR567654_2$ htseq-count  -o liver.reads accepted_hits.bam mm9_convert.GTF
```

Output :

```
zhangyw@yp:~$ more liver.reads
ENSMUSG00000000001      8673
ENSMUSG00000000003      0
ENSMUSG00000000028      1279
ENSMUSG00000000031      136477
ENSMUSG00000000037      224
ENSMUSG00000000049      2
ENSMUSG00000000056      2059
ENSMUSG00000000058      31
ENSMUSG00000000078      671
ENSMUSG00000000085      1064
ENSMUSG00000000088      1482
ENSMUSG00000000093      126
ENSMUSG00000000094      0
```

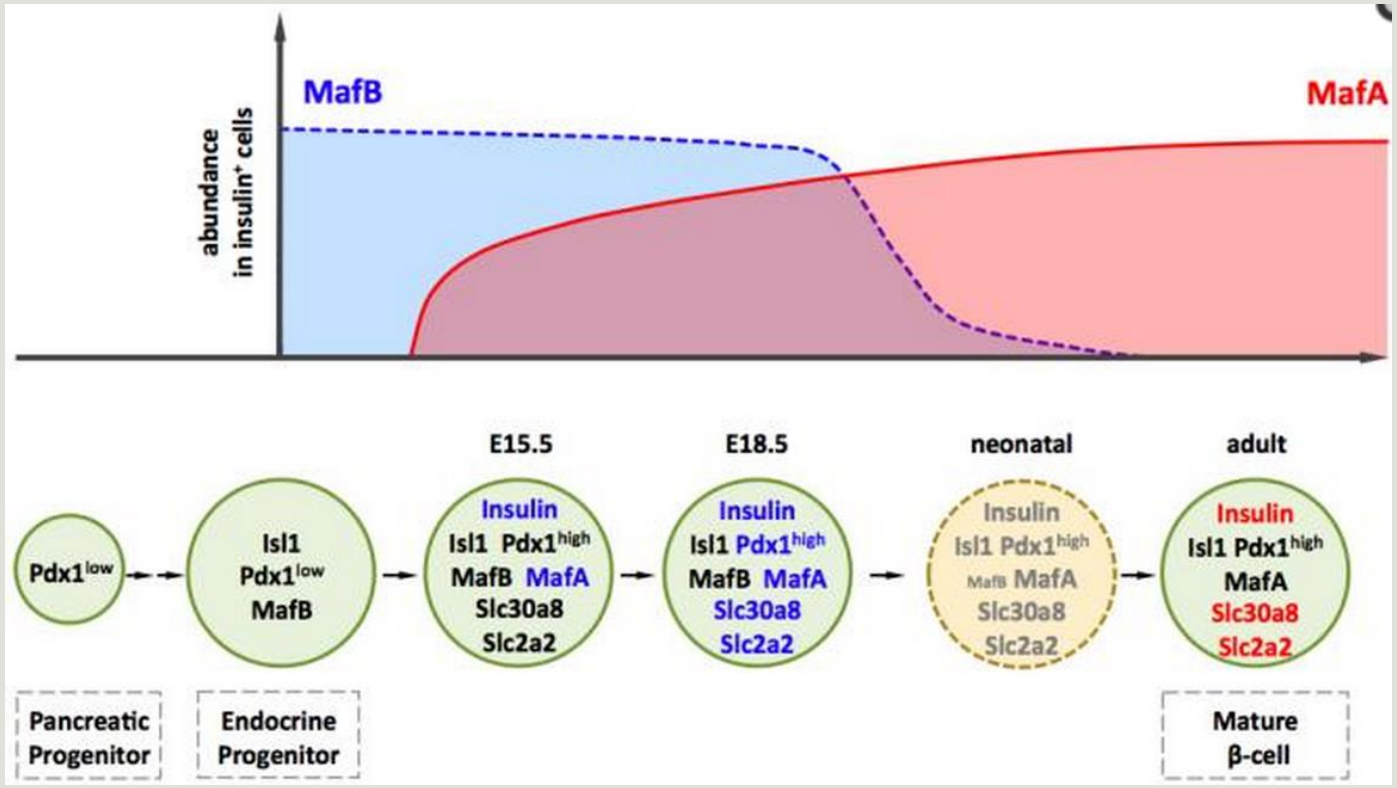
DESeq script:

```
library(DESeq2)
library("RColorBrewer")
library("gplots")

directory <- getwd()
sampleFiles <- grep("*reads",list.files(directory),value=TRUE)

#in this step, define the condition of experiments
#sampleCondition <- sub("(*).read_count","\\1",sampleFiles)

experiment<-cbind(c("Pancreas","Pancreas","Liver"),c("10.5","10.5","10.5"))
colnames(experiment)<-c("Tissue","Time")

sampleTable <- data.frame(sampleName = sampleFiles,
                          fileName = sampleFiles,
                          Tissue =experiment[,"Tissue"],Time=experiment[,"Time"])

ddsHTSeq <- DESeqDataSetFromHTSeqCount(sampleTable = sampleTable,directory = directory,design= ~ Tissue )

colData(ddsHTSeq)$Tissue <- factor(colData(ddsHTSeq)$Tissue,levels=c("Pancreas","Liver"))

dds <- DESeq(ddsHTSeq)
colnames(dds)<-c("10.5DP","10.5VP","10.5Lv")
res <- results(dds)

write.table(res,file="10.5Pancreas_Liver_diff.xls")
```

Output:

TFs	fold change(liver/pancreas)
STX1A	0.734850401
STXBP1	0.321740732
SYP	0.243882741
SYT4	0.46690488
SYT7	0.321672298
MAFA	0.418045483
MAFB	0.166969619
KCNK3	0.776610274
KCNMA1	0.27431245
KCNMB2	2.272147303
KCNN1	1.557423638
KCNN3	0.544070131
SCN1A	0.808009664

Maf family transcription factors



Yan Hang and Roland Stein

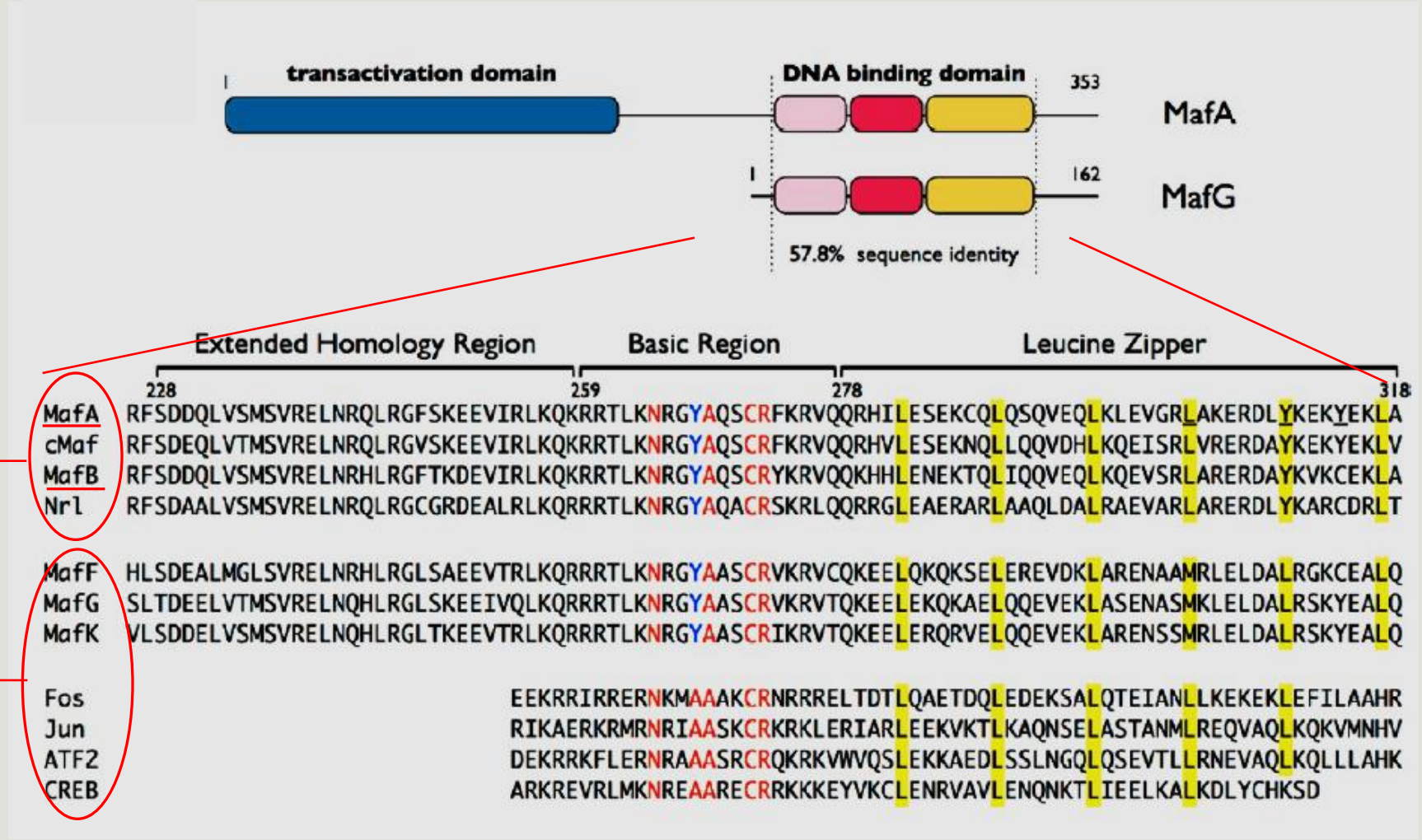
MafA and mafB is expressed in pancreatic β-cells and is essential for β-cell maturation and glucose-induced insulin expression.

Outline of MafA and MafB

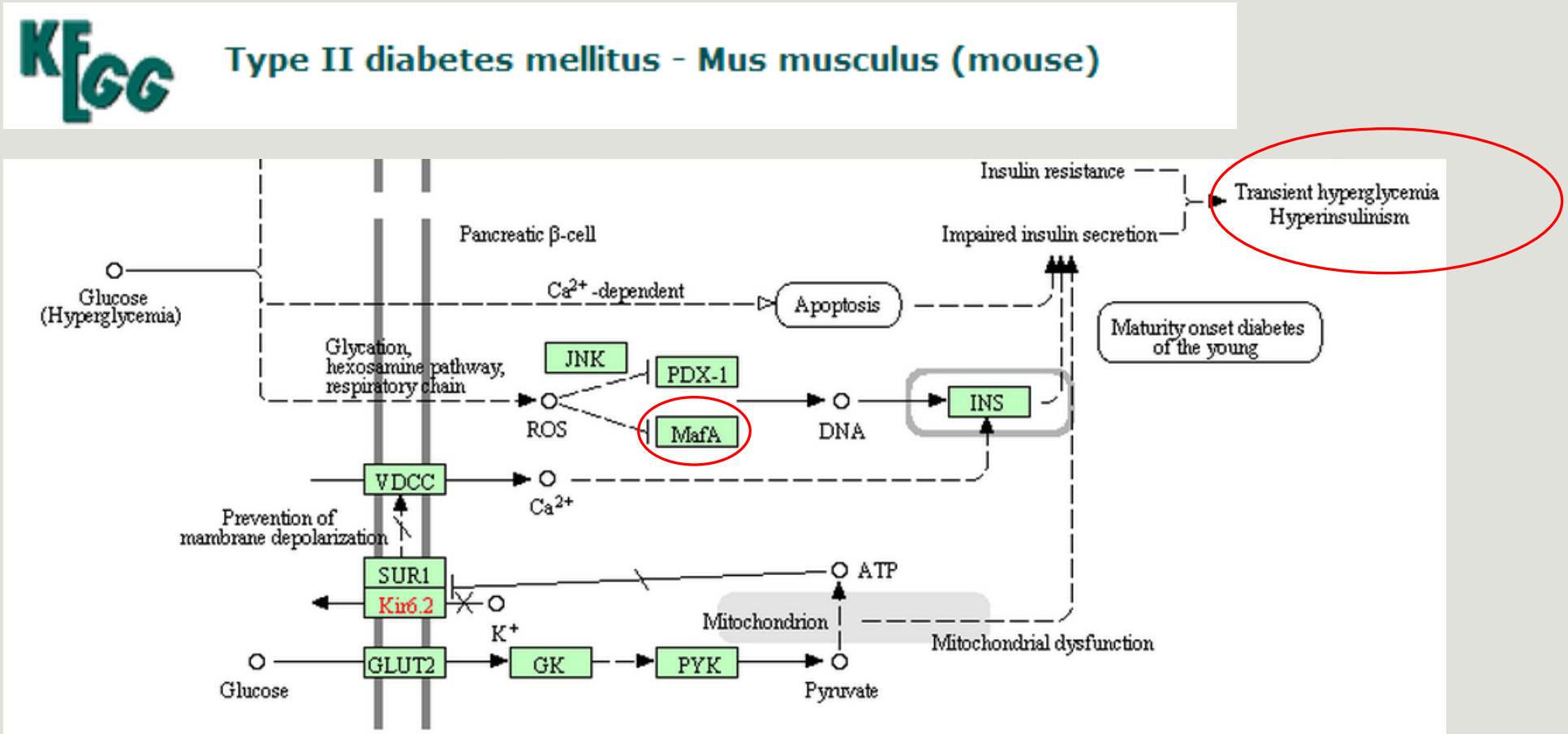
I. The maf family can be subdivided into two groups. Large mafs are proto-oncoproteins and are overexpressed in many human cancers.

II. Small mafs, MafF, MafG, andMafK, lack an activation domain and act as repressors or rely on dimerization partners for their transcriptional activity.

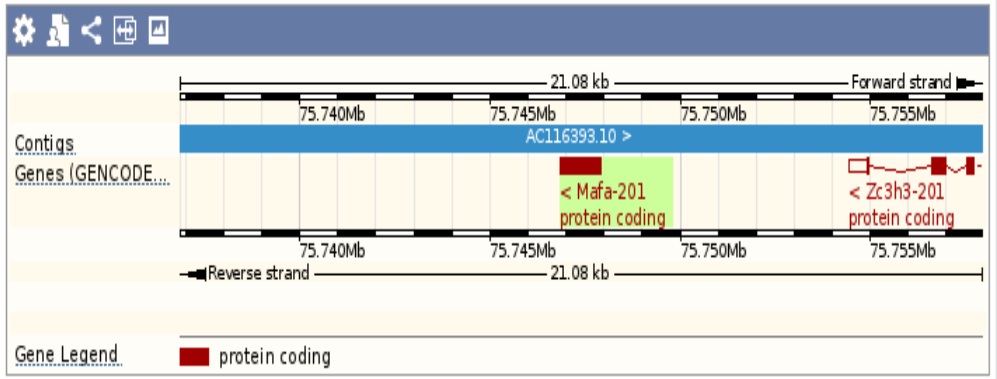
III. They have the simplest DNA-binding motif, with two long α-helices gripping DNA like a pair of chopsticks.


transactivation domain
DNA binding domain
1
353
MafA
1
162
MafG
57.8% sequence identity
Extended Homology Region
Basic Region
Leucine Zipper
228
259
278
318
MafA RFSDDQLVSMSVRELNRQLRGFSKEEVIRLKQKRRTLKNRGYAQSCRFKRVQQRHILESEKCQLQSQVEQLKLEVGRLAKERDLYKEKYEKLA
cMaf RFSDEQLVTMSVRELNRQLRGVSKEEVIRLKQKRRTLKNRGYAQSCRFKRVQQRHVLESEKNQLLQQVDHLKQEISRLVRERDAYKEKYEKLV
MafB RFSDDQLVSMSVRELNRHLRGFTKDEVIRLKQKRRTLKNRGYAQSCRYKRVQQKHHLENEKTQLIQQVEQLKQEVSRLARERDAYKVKCEKLA
Nrl RFSDAALVSMSVRELNRQLRGCGRDEALRLKQRRRTLKNRGYAQACRSKRLQQRRGLEAERARLAAQLDALRAEVARLARERDLYKARCDRLT
MafF HLSDEALMGLSVRELNRHLRGLSAEEVTRLKQRRRTLKNRGYAASCRVKRVCQKEELQKQKSELEREVDKLARENAAMRLELDALRGKCEALQ
MafG SLTDEELVTMSVRELNQHLRGLSKEEIVQLKQRRRTLKNRGYAASCRVKRVTQKEELEKQKAELQQEVEKLASENASMKLELDALRSKYEALQ
MafK VLSDDELVSMSVRELNQHLRGLTKEEVTRLKQRRRTLKNRGYAASCRIKRVTQKEELERQRVELQQEVEKLARENSSMRLELDALRSKYEALQ
Fos EEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHR
Jun RIKAERKRMRNRIAASKCRKRKLERIARLEEKVKTLKAQNSELASTANMLREQVAQLKQKVMNHV
ATF2 DEKRRKFLERNRAAASRCRQKRKVWVQSLEKKAEDLSSLNGQLQSEVTLLRNEVAQLKQLLLAHK
CREB ARKREVRLMKNREAARECRRKKKEYVKCLENRVAVLENQNKTLIEELKALKDLYCHKSD
Large maf
small maf

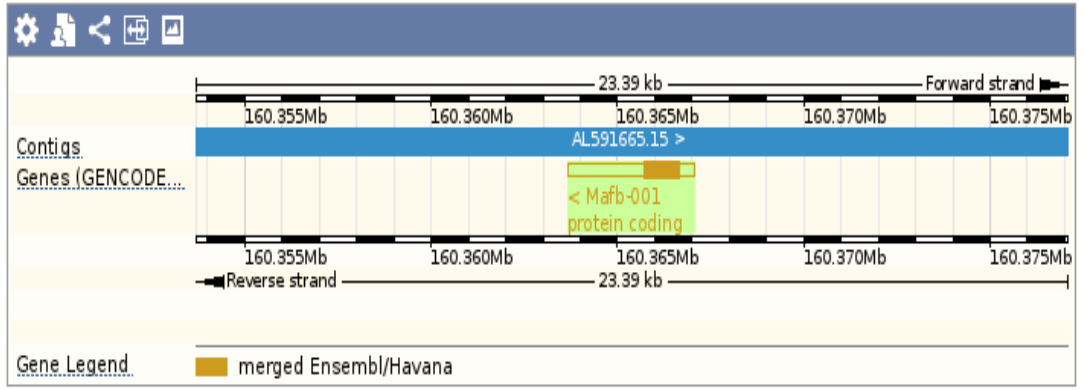
Signaling network of mafA



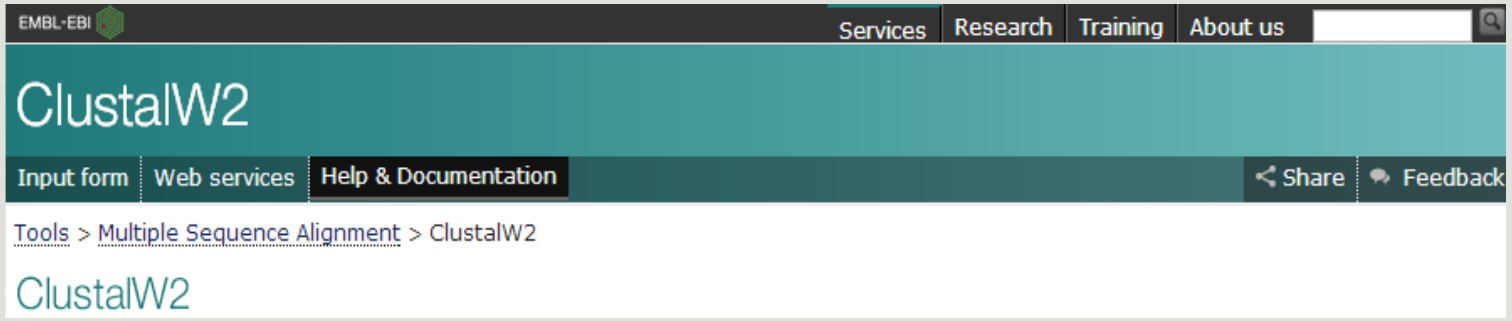
Description	v-maf musculoaponeurotic fibrosarcoma oncogene family, protein A (avian) [Source:MGI Symbol;Acc:MGI:2673307]
Location	Chromosome 15: 75,746,843-75,747,922 reverse strand.
INSDC coordinates	chromosome:GRCm38:CM001008.2:75746843:75747922:1
Transcripts	This gene has 1 transcript (splice variant) Show transcript table

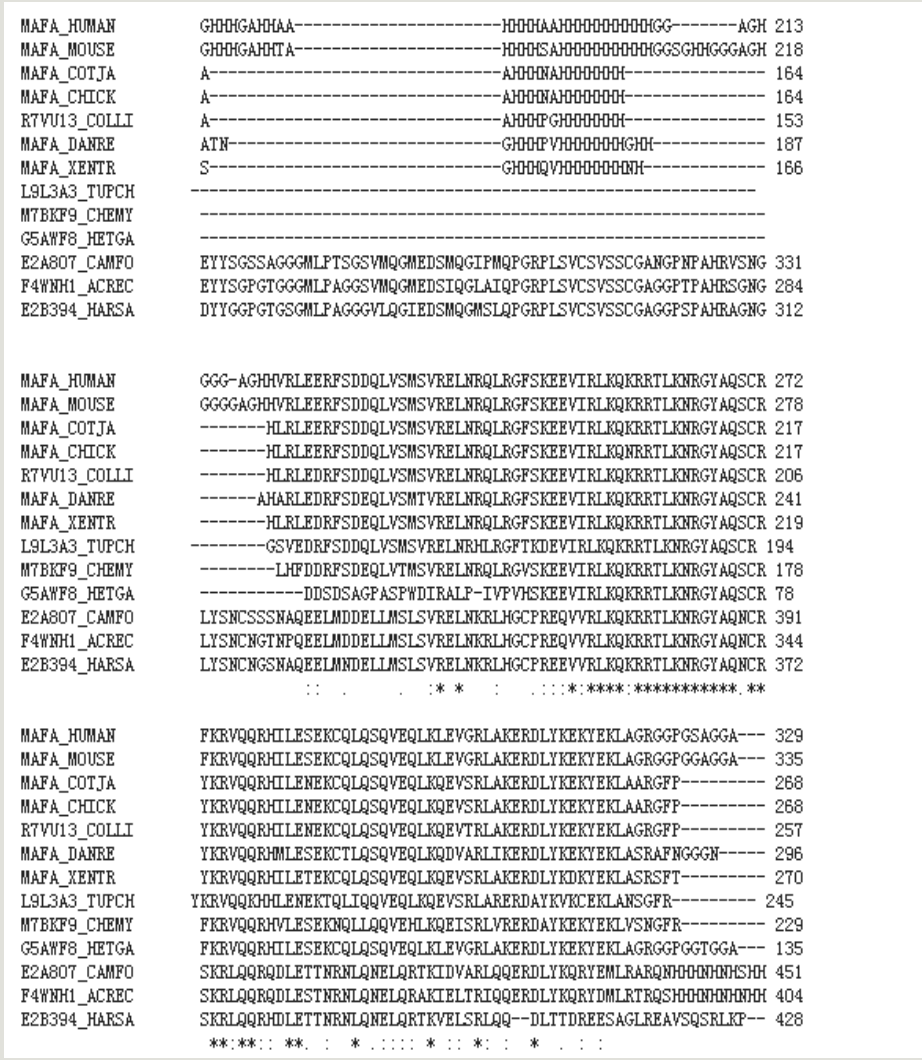


Description	v-maf musculoaponeurotic fibrosarcoma oncogene family, protein B (avian) [Source:MGI Symbol;Acc:MGI:104555]
Location	Chromosome 2: 160,363,677-160,367,065 reverse strand.
INSDC coordinates	chromosome:GRCm38:CM000995.2:160363677:160367065:1
Transcripts	This gene has 1 transcript (splice variant) Hide transcript table



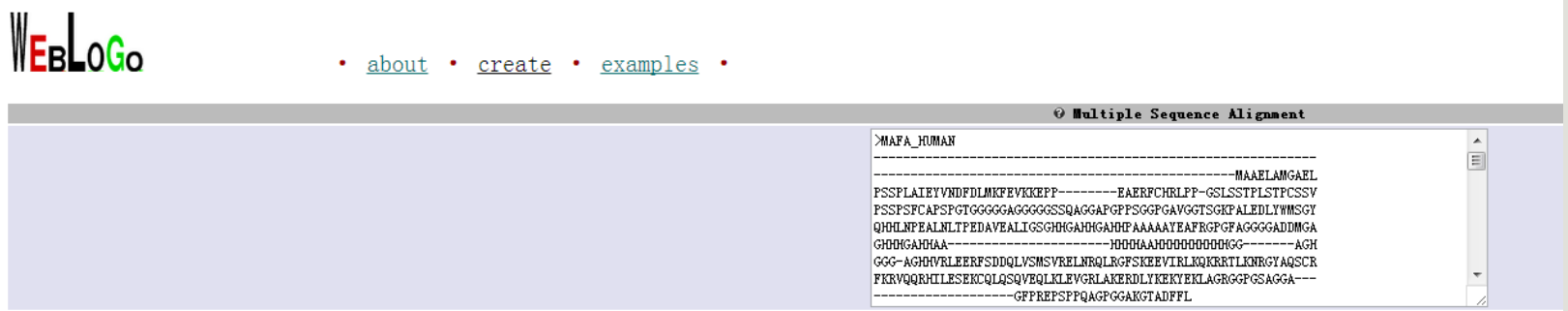
multiple sequence alignment program--- clustalW2

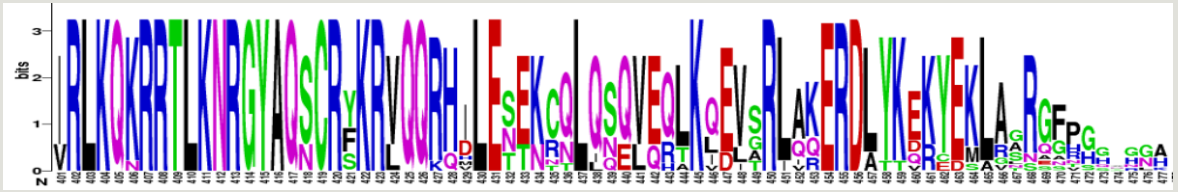




将比对结果输入weblogo中寻找sequence logo

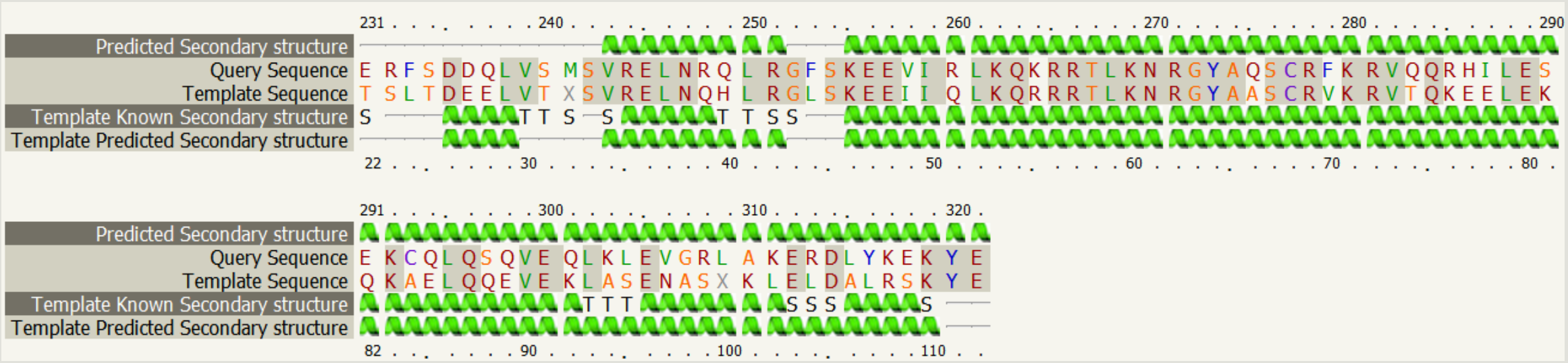
Sequence Conservation



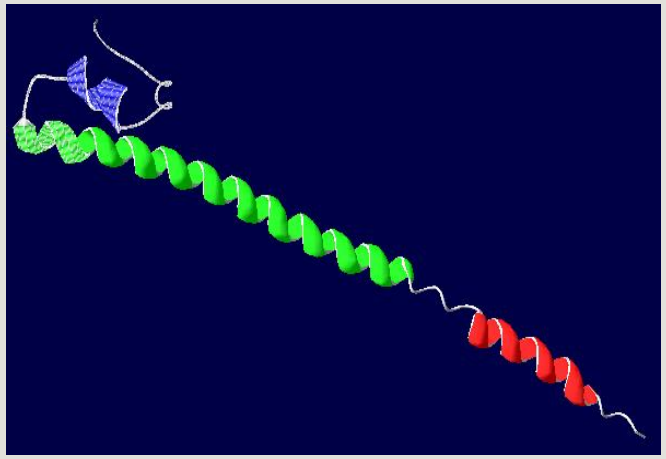


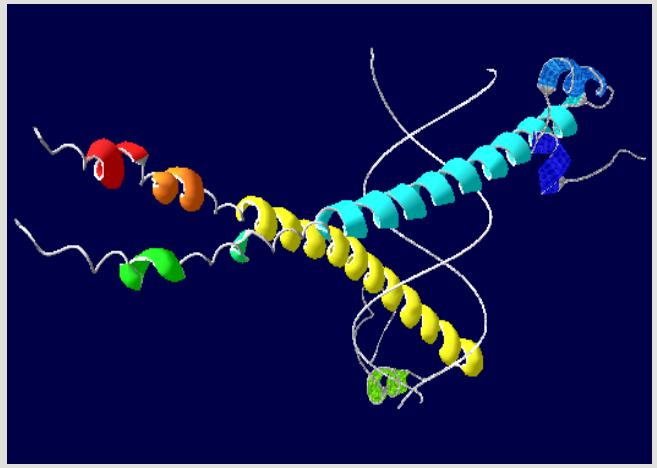
☐	Region	266 - 280	15	Interaction with DNA By similarity	
☐	Region	288 - 309	22	Leucine-zipper	

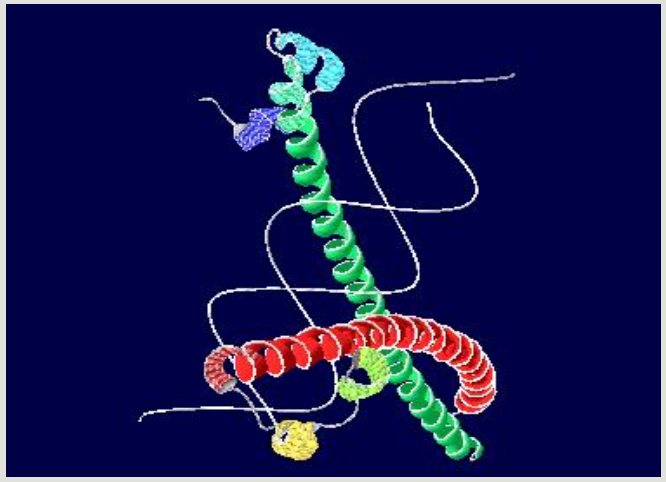
Secondary structure of MAFA_MOUSE MAFB_MOUSE

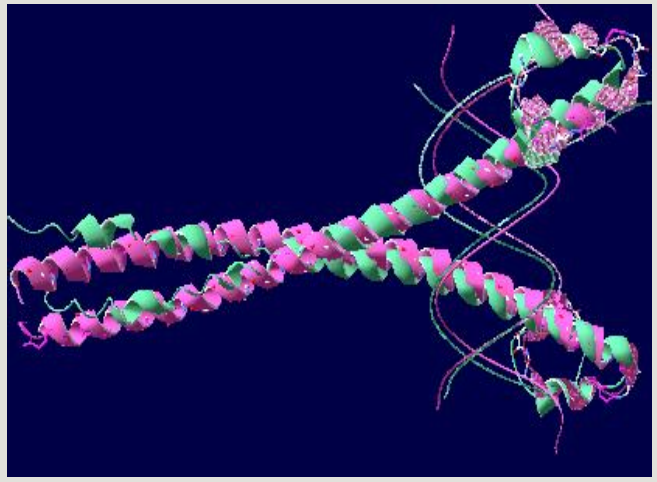


Structure of MAFA_MOUSE MAFB_MOUSE

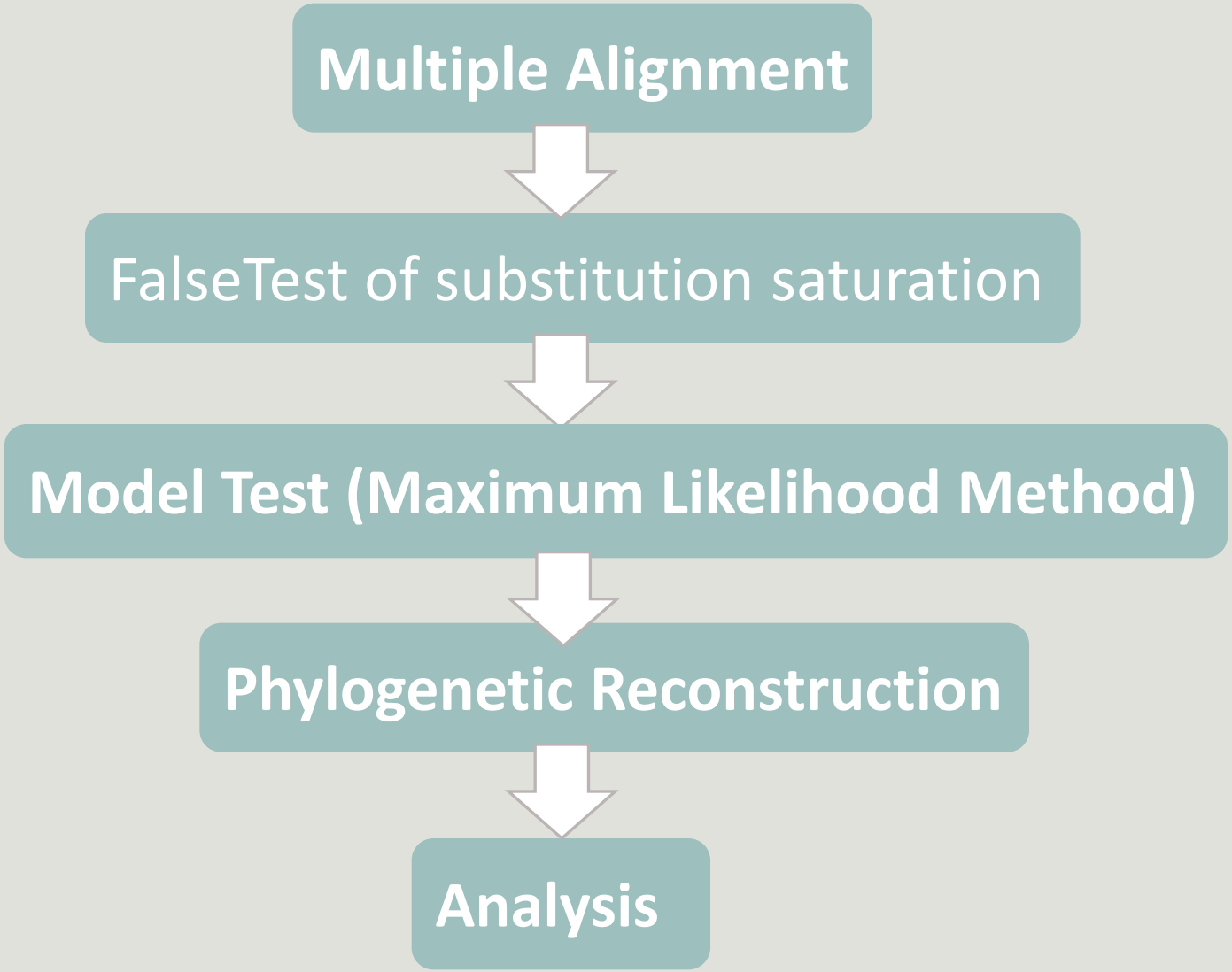








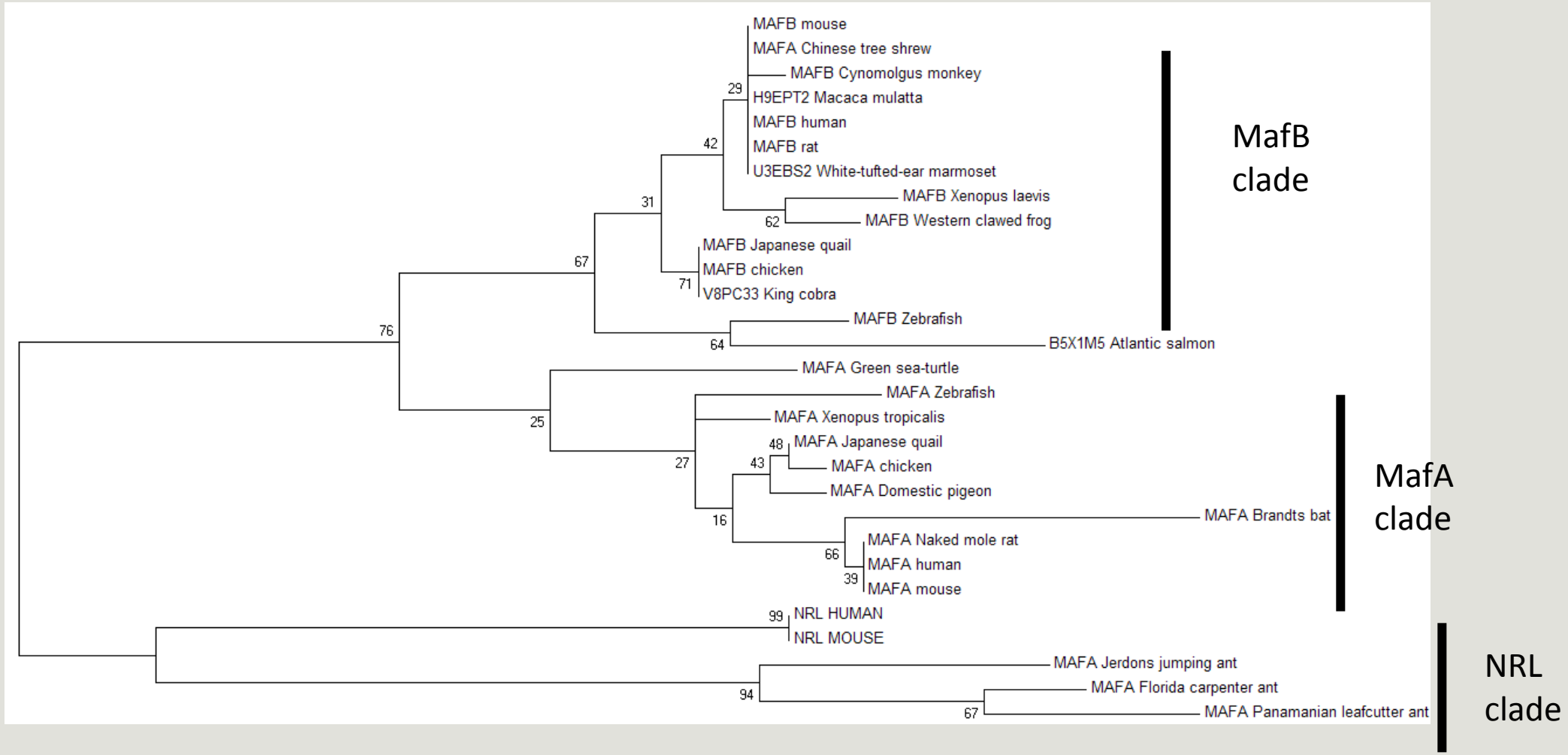
Phylogenetics Reconstruction



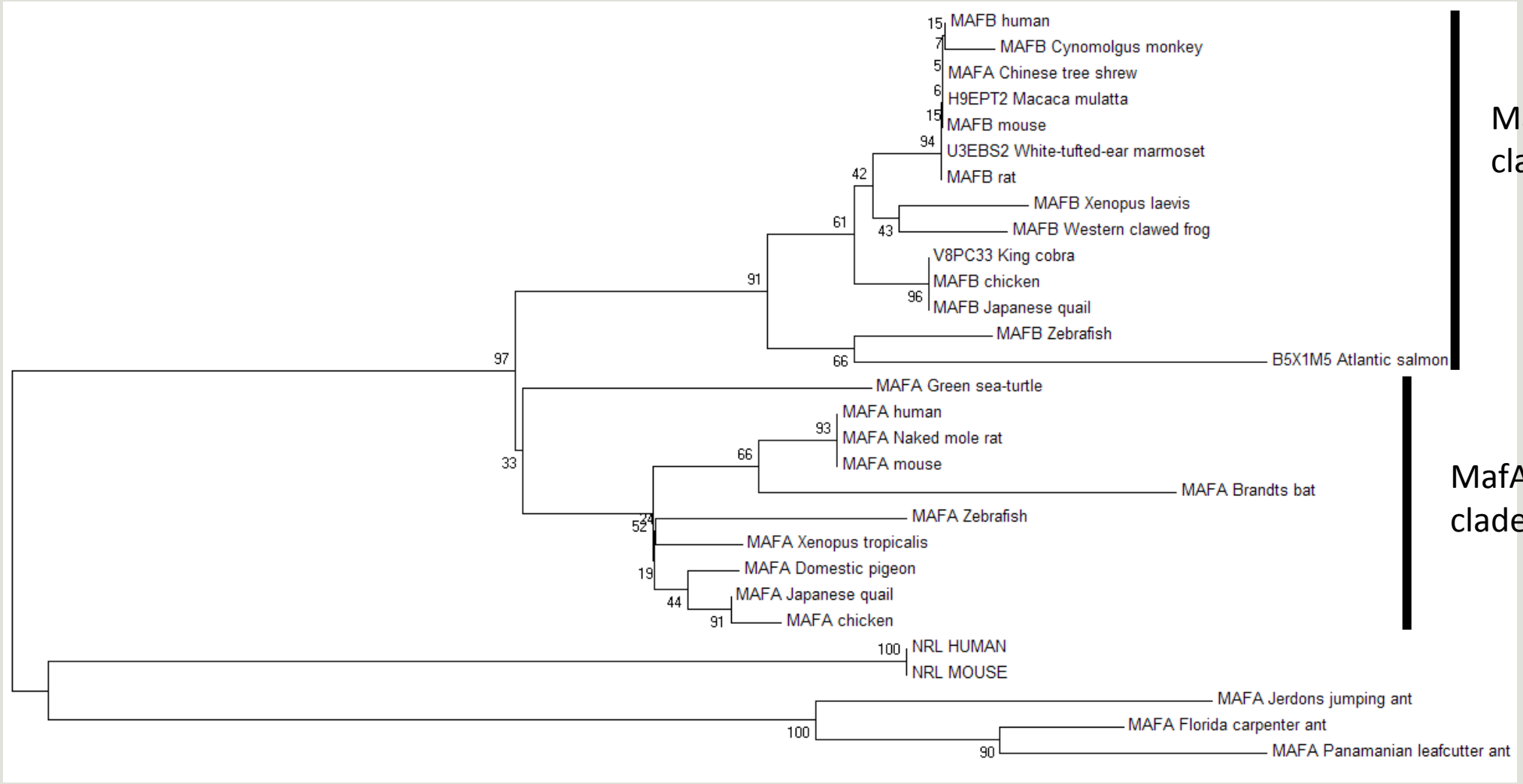
FalseTest of substitution saturation

```
FalseTest of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Analysis performed on fully resolved sites only.

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.
============================================================
Prop. invar. sites                          .0000
Mean H                                      .0000
Standard Error                              .0000
Hmax                                       1.5321
Iss                                         .0000
Iss.c                                      3.0330
T                                  3606170510843870000000000.0000
DF                                         2
Prob (Two-tailed)                           .0000
95% Lower Limit                             .0000
95% Upper Limit                             .0000

============================================================
Part II. For an extreme asymmetrical (and generally very
unlikely) tree.
============================================================
Iss.c                                     19.3750
T                                 230362535763728000000000000.0000
DF                                         2
Prob (Two-tailed)                           .0000

95% Lower Limit                             .0000
95% Upper Limit                             .0000

============================================================
```

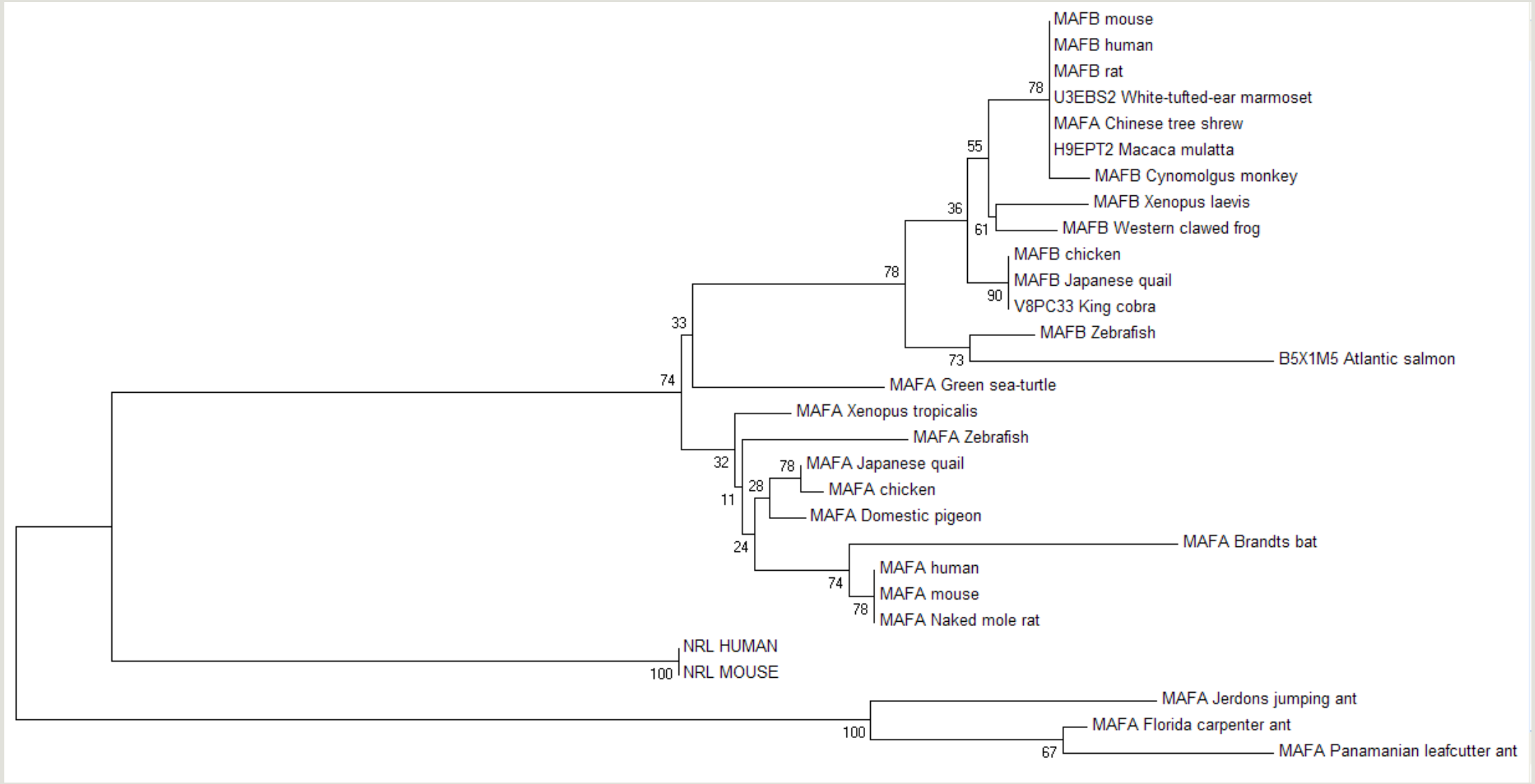
FalseTest of substitution saturation by DAMBE



Phylogenetic Maximum-parsimony tree of Maf family. Aliment of the sequence was made with ClustalX; this was used to derive phylogenetic tree with paup by the Maximum-parsimony method. Bootstrapping was carried out on 1000 replicates. Based on the tree Maf family consist of three clades.



Phylogenetic Neighbour-joining tree of Maf family. Aliment of the sequence was made with ClustalX;this was used to derive phylogenetic tree with mega5.0 by the minimum-parsimony method. Bootstrapping was carried out on 1000 replicates. Based on the tree Maf family consist of three clades.



Phylogenetic Maximum-likelihood tree of Maf family. Aliment of the sequence was made with ClustalX;this was used to derive phylogenetic tree with PHYLIP by the Maximum-likelihood method. Based on the tree Maf family consist of three clades.

Conclusion

RNA-seq could detect new genes important for development.

Maf family may have at least one replication during evolution.

Replication of maf family in Chordata may related to the origin of pancreas.

Acknowledgement

- Professor Luo
- TA Yujian Kang
- All members of our team
- All classmates of PKU14 spring ABC

Question & Answer