SPDEv3.0 使用说明

目录

Table of contents

一、SPDEv3.0 简介

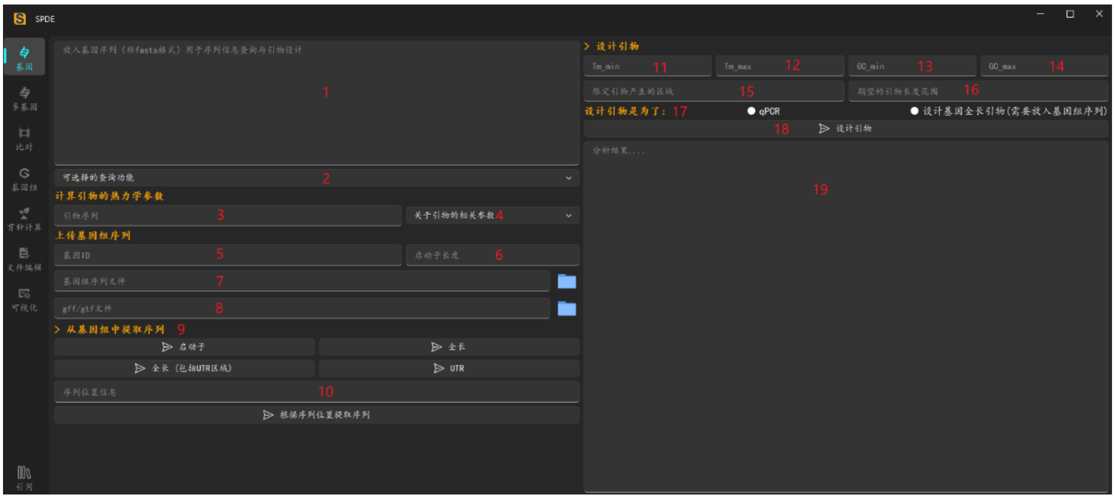
SPDEv3.0 用于分子生物学、基因组学、育种学等学科的序列提取、计算、生信分析以及结果的可视化等过程。其最突出的特点在于进一步整合了一些不需要用户决策的分析过程而进一步提高了分析的效率并进一步优化了操作的整个过程。为了使同学们可以更好地使用 SPDE，我们在软件界面提供了足够的提示信息，这些信息多以悬浮式的提示出现。我们相信 SPDE 将成为促进植物研究发展的重要工具。以下是关于软件使用的详细介绍。

请同学们注意：软件安装路径中不能有中文，其他路径也要尽量避免使用中文。同时，要尽量避免使用含有空格的路径。

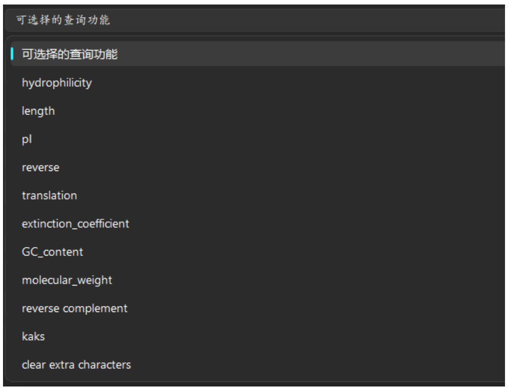
二、模块功能

1、 基因模块

该模块主要为解决分子生物实验中关于序列信息查询、提取、引物设计等需求而设计。输入方式及功能分布如下：

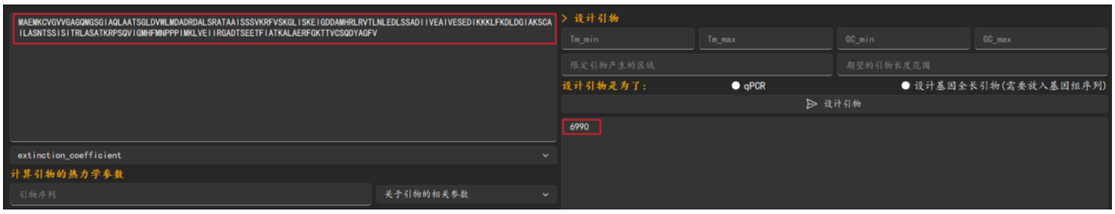


I. 当需要查询序列信息时，在 1 中放入基因序列（不需要 fasta 格式，仅序列即可）。可以直接点击 2 中功能来查询关于放入序列的一些基本信息。这些基本信息包括：



Hydrophilicity:亲水性(只用于蛋白)
length：长度
pI：等电点(只用于蛋白)
reverse：序列反向
translation：翻译(只用于 DNA)
extinction_coefficient：消光系数
GC_content：GC 含量
molecular_weight：分子量
reverse complement：反向互补序列
kaks：Ka/Ks
clear extra characters：清除多余字符

用法很简单：点开下拉框后，只需要用鼠标选择某个功能，分析的结果会自动出现在 19 这个位置：

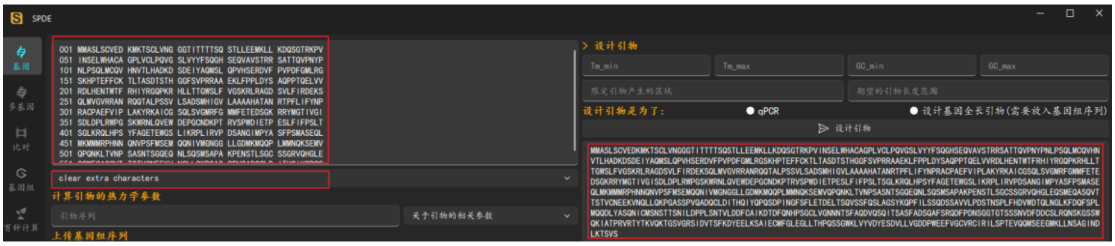


```
MAEMKCVGVVGAGQMGSGIAQLAATSGLDVWLMDADRDALSRATAAISSSVKRFVSKGLISKEIGDDAMHRLRVTLNLEDLSSADIIVEAIVESEDIKKKLFKDLDGIAKSCA
ILASNTSSISITRLASATKRPSQVIGMHFMNPPPIMKLVEIIRGADTSEETFIATKALAERFGKTTVCSQDYAGFV
```

extinction_coefficient

```
6990
```

有两个功能需要额外说明：1）kaks 功能需要用户放入两个基因的序列以便于计算，这时的输入格式是 fasta 格式；2）清除多余字符：在 NCBI 以及 TAIR 等网站，经常会看到数字和空格以及序列混合的基因序列。数字、空格等额外信息处理起来很麻烦。这个时候可以清除多余字符功能一键式去除额外信息。如下：

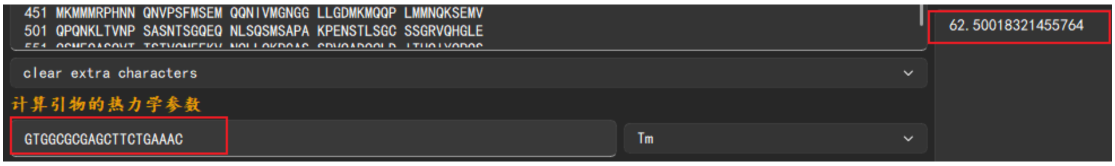


```
001 MMASLSCVED KMKTSCLVNG GGTITTTTSQ STLLEEMKLL KDQSGTRKPV
051 INSELWHACA GPLVCLPQVG SLVYYFSQGH SEQVAVSTRR SATTQVPNYP
101 NLPSQLMCQV HNVTLHADKD SDEIYAQMSL QPVHSERDVF PVPDFGMLRG
151 SKHPTEFFCK TLTASDTSTH GGFSVPRRAA EKLFPPLDYS AQPPTQELVV
201 RDLHENTWTF RHIYRGQPKR HLLTTGWSLF VGSKRLRAGD SVLFIRDEKS
251 QLMVGVRRAN RQQTALPSSV LSADSMHIGV LAAAAHATAN RTPFLIFYNP
301 RACPAEFVIP LAKYRKAICG SQLSVGMRFG MMFETEDSGK RRYMGTIVGI
351 SDLDPLRWPG SKWRNLQVEW DEPGCNDKPT RVSPWDIETP ESLFIFFPSLT
401 SGLKRQLHPS YFAGETEWGS LIKRPLIRVP DSANGIMPYA SFPSMASEQL
451 MKMMMRPHNN QNVPSFMSEM QQNIVMGNGG LLGDMKMQQP LMMNQKSEMV
501 QPQNKLTVNP SASNTSGQEQ NLSQSMSAPA KPENSTLSGC SSGRVQHGLE
```

clear extra characters

```
MMASLSCVEDKMKTSCLVNGGGTITTTTSQSTLLEEMKLLKDQSGTRKPVINSELWHACAGPLVCLPQVGSLVYYFSQGHSEQVAVSTRRSATTQVPNYPNLPSQLMCQVHN
VTLHADKDSDEIYAQMSLQPVHSERDVFPVPDFGMLRGSKHPTEFFCKTLTASDTSTHGGFSVPRRAAEKLFPPLDYSAQPPTQELVVRDLHENTWTFRHIYRGQPKRHLLT
TGWSLFVGSKRLRAGDSVLFIRDEKSQLMVGVRRANRQQTALPSSVLSADSMHIGVLAAAAHATANRTPFLIFYNPRACPAEFVIPLAKYRKAICGSQLSVGMRFGMMFETE
DSGKRRYMGTIVGISDLDPLRWPGSKWRNLQVEWDEPGCNDKPTRVSPWDIETPESLFIFFPSLTSGLKRQLHPSYFAGETEWGSLIKRPLIRVPDSANGIMPYASFPSMASE
QLMKMMNRPHNNQNVPSFMSEMQQNIVMGNGGLLGDMKMQQPLMMNQKSEMVQPQNKLTVNPSASNTSGQEQNLSQSMSAPAKPENSTLSGCSSGRVQHGLEQSMEQASQVT
TSTVCNEEKVNQLLQKPGASSPVQADQCLDITHQIYQPQSDPINGFSFLETDELTSQVSSFQSLAGSYKQPFILSSQDSSAVVLPDSTNSPLFHDVWDTQLNGLKFDQFSPL
MQQDLYASQNICMSNSTTSNILDPPLSNTVLDDFCAIKDTDFQNHPSGCLVGNNNTSFAQDVQSQITSASFADSQAFSRQDFPDNSGGTGTSSSNVDFDDCSLRQNSKGSSW
QKIATPRVRTYTKVQKTGSVGRSIDVTSFKDYEELKSAIECMFGLEGLLTHPQSSGWKLVYVVDYESDVLLVGDDPWEEFVGCVRCIRILSPTEVQQMSEEGMKLLNSAGIND
LKTSVS
```

II. 在引物的热力学计算中，用法与之前相同，当想要查询某些引物的时候，把序列直接拷贝进来，然后点击相应功能，即可输出相应计算结果：



```
451 MKMMMRPHNN QNVPSFMSEM QQNIVMGNGG LLGDMKMQQP LMMNQKSEMV
501 QPQNKLTVNP SASNTSGQEQ NLSQSMSAPA KPENSTLSGC SSGRVQHGLE
```

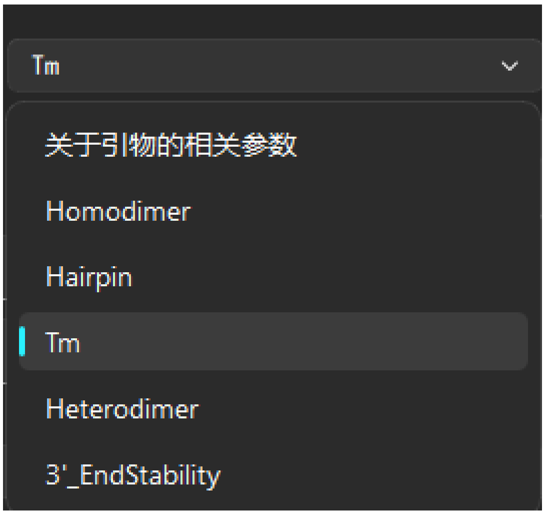
clear extra characters

计算引物的热力学参数

GTGGCGCGAGCTTCTGAAAC

Tm

```
62.50018321455764
```



可供计算的热力学参数包括：

Homodimer：同源二聚体

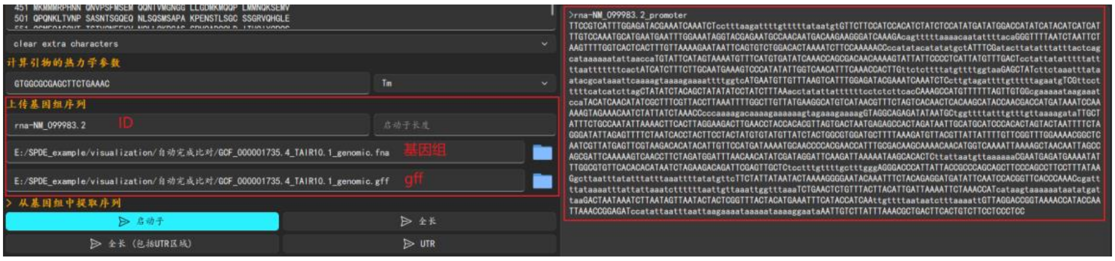
Hairpin：发夹结构

Tm：熔解温度（即 Tm 值）

Heterodimer：异源二聚体

3'_EndStability：3'端稳定性

III. 基因组的序列提取功能可以根据用户提供的基因 ID，从基因组中提取相应序列。输入文件是基因组的序列文件以及结构注释文件（GFF 和 GTF 格式）。如下：



```
451 MKMMMRPHNN QNVPSFMSEM QQNIVMGNGG LLGDMKMQQP LMMNQKSEMV
501 QPQNKLTVNP SASNTSGQEQ NLSQSMSAPA KPENSTLSGC SSGRVQHGLE
```

clear extra characters

GTGGCGCGAGCTTCTGAAAC　Tm

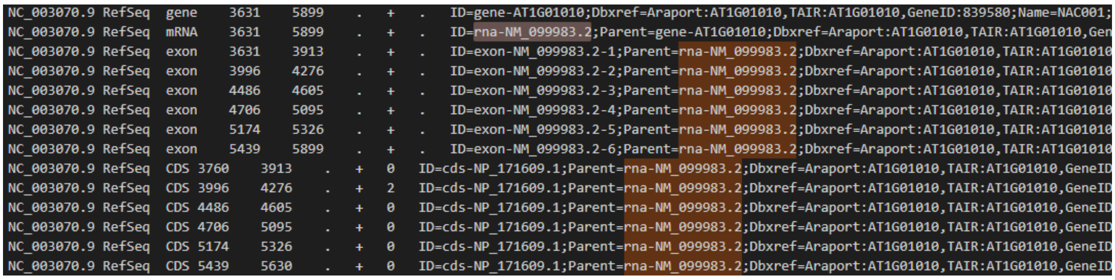
rna-NM_099983.2

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

```
>rna-NM_099983.2_promoter
TTCCGTCATTTGGAGATACGAAATCAAATCTcctttaagattttgtttttataatgtGTTCTTCCATCCACATCTATCTCCATATGATATGGACCATATCATACATCATCAT
TTGTCCAAATGCATGAATGAATTTGGAAATAGGTACGAGAATGCCAACAATGACAAGAAGGGATCAAAGAcagtttttaaaacaatattttacaGGGTTTTAATCTAATTCT
AAGTTTTGGTCACTCACTTTGTTAAAAGAATAATTCAGTGTCTGGACACTAAAATCTTCCAAAAACCcatatacatatatgctATTTCGatacttatatttatttactcag
cataaaaaatattaaccaTGTATTCATAGTAAAATGTTTCATGTGATATCAAACCAGCGACAACAAAAGTATTATTCCCCTCATTATGTTTGACTcctattatattttttatt
ttaatttttttcactATCATCTTTCTTGCAATGAAAGTCCCATATATTGGTCAACATTTCAAACCACTTGttctcttttatgttttggtaaGAGCTATcttctaaatttata
atacgcataaattcaaaagtaaaagaaaattttggtcATGAATGTTGTTTAAGTCATTTGGAGATACGAAATCAAATCTCcttgtagattttgtttttagaatgTCGttcct
ttttcatcatcttagCTATATCTACAGCTATATATCCTATCTTTAAacctatattatttttcctctcttcacCAAAGCCATGTTTTTTAGTTGTGGcgaaaaataagaaat
ccaTACATCAACATATCGCTTTCGTTACCTTAAATTTTGGCTTGTTATGAAGGCATGTCATAACGTTTCTAGTCACAACTCACAAGCATACCAACGACCATGATAAATCCAA
AAAGTAGAAACAATCTATTATCTAAACCcccaaaagacaaaagaaaaaagtagaaagaaaagGTAGGCAGAGATATAATGCtggttttatttgtttgttaaaagataTTGCT
ATTTCTGCCAATATTAAAACTTCACTTAGGAAGACTTGAACCTACCACACGTTAGTGACTAATGAGAGCCACTAGATAATTGCATGCATCCCACACTAGTACTAATTTTCTA
GGGATATTAGAGTTTTCTAATCACCTACTTCCTACTATGTGTATGTTATCTACTGGCGTGGATGCTTTTAAAGATGTTACGTTATTATTTTGTTCGGTTTGGAAAACGGCTC
AATCGTTATGAGTTCGTAAGACACATACATTGTTCCATGATAAAATGCAACCCCACGAACCATTTGCGACAAGCAAAACAACATGGTCAAAATTAAAAGCTAACAATTAGCC
AGCGATTCAAAAAGTCAACCTTCTAGATGGATTTAACAACATATCGATAGGATTCAAGATTAAAAATAAGCACACTCttattaatgttaaaaaaCGAATGAGATGAAAATAT
TTGGCGTGTTCACACACATAATCTAGAAGACAGATTCGAGTTGCTCtcctttgttttgctttgggAGGGACCCATTATTACCGCCCAGCAGCTTCCCAGCCTTCCTTTATAA
GgcttaatttatatttatttaaattttatatgttcTTCTATTATAATACTAAAAGGGGAATACAAATTTCTACAGAGGATGATATTCAATCCACGGTTCACCCAAACcgatt
ttataaaatttattattaaatctttttaattgttaaattggtttaaaTCTGAACTCTGTTTACTTACATTGATTAAAATTCTAAACCATcataagtaaaaaataatatgat
taaGACTAATAAATCTTAATAGTTAATACTACTCGGTTTACTACATGAAATTTCATACCATCAAttgttttaataatctttaaaattGTTAGGACCGGTAAAACCATACCAA
TTAAACCGGAGATccatattaatttaattaagaaaataaaaataaaaggaataAATTGTCTTATTTAAACGCTGACTTCACTGTCTTCCTCCCTCC
```

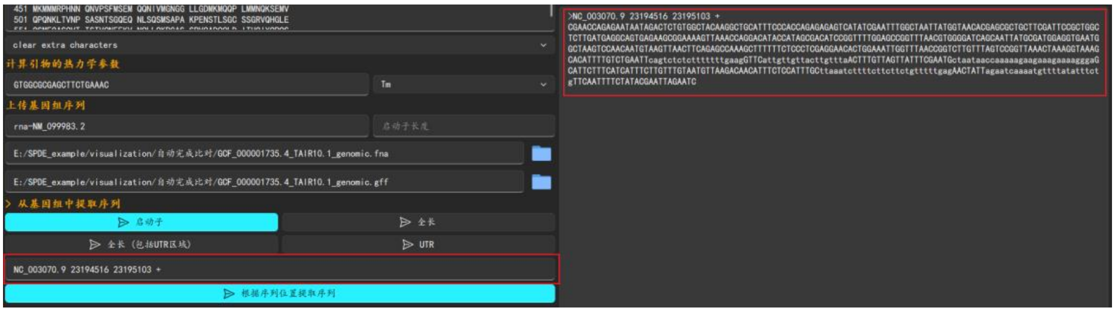
一共可以提取包括启动子、基因全长（从 ATG 到终止密码子）、基因全长（含有 UTR 区）以及 UTR 区。需要注意的是基因 ID 应该是 GFF/GTF 文件中 mRNA 对应的 ID。GFF 文件的格式案例与 ID 的具体位置如下：



```
NC_003070.9 RefSeq  gene    3631    5899    .   +   .   ID=gene-AT1G01010;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID:839580;Name=NAC001;
NC_003070.9 RefSeq  mRNA    3631    5899    .   +   .   ID=rna-NM_099983.2;Parent=gene-AT1G01010;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,Gen
NC_003070.9 RefSeq  exon    3631    3913    .   +   .   ID=exon-NM_099983.2-1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  exon    3996    4276    .   +   .   ID=exon-NM_099983.2-2;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  exon    4486    4605    .   +   .   ID=exon-NM_099983.2-3;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  exon    4706    5095    .   +   .   ID=exon-NM_099983.2-4;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  exon    5174    5326    .   +   .   ID=exon-NM_099983.2-5;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  exon    5439    5899    .   +   .   ID=exon-NM_099983.2-6;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  CDS 3760    3913    .   +   0   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
NC_003070.9 RefSeq  CDS 3996    4276    .   +   2   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
NC_003070.9 RefSeq  CDS 4486    4605    .   +   0   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
NC_003070.9 RefSeq  CDS 4706    5095    .   +   0   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
NC_003070.9 RefSeq  CDS 5174    5326    .   +   0   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
NC_003070.9 RefSeq  CDS 5439    5630    .   +   0   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
```

GFF/GTF 文件的格式很重要，请同学们在使用前根据上述示例文件，仔细核对好格式。

当然，如果没有 GFF/GTF 文件，也可以直接输入位置信息来提取文件，如下：



```
451 MKMMMRPHNN QNVPSFMSEM QQNIVMGNGG LLGDMKMQQP LMMNQKSEMV
501 QPQNKLTVNP SASNTSGQEQ NLSQSMSAPA KPENSTLSGC SSGRVQHGLE
```

clear extra characters

GTGGCGCGAGCTTCTGAAAC　Tm

rna-NM_099983.2

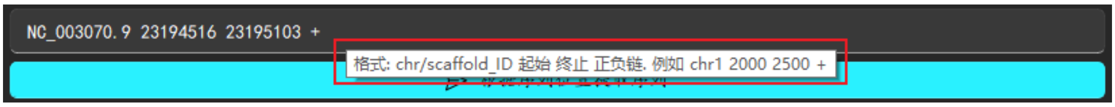
E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

NC_003070.9 23194516 23195103 +

```
>NC_003070.9 23194516 23195103 +
CGAACCAGAGAATAATAGACTCTGTGGCTACAAGGCTGCATTTCCCACCAGAGAGAGTCATATCGAATTTGGCTAATTATGGTAACACGAGCGCTGCTTCGATTCCGCTGGC
TCTTGATGAGGCAGTGAGAAGCGGAAAAAGTTAAACCAGGACATACCATAGCGACATCCGGTTTTGGAGCCGGTTTAACGTGGGGATCAGCAATTATGCGATGGAGGTGAATG
GCTAAGTCCAACAATGTAAGTTAACTTCAGAGCCAAAGCTTTTTTCTCCCTCGAGGAACACTGGAAATTGGTTTAACCGGTCTTGTTTAGTCCGGTTAAACTAAAGGTAAAG
CACATTTTGTCTGAATTcagtctctctttttttgaagGTTCattgttgttacttgtttaACTTTGTTAGTTATTTCGAATGctaataaccaaaaagaagaaagaaaagggaG
CATTCTTTCATCATTTCTTGTTTGTAATGTTAAGACAACATTTCTCCATTTGCttaaatcttttcttcttctgtttttgagAACTATTagaatcaaaatgttttatatttct
gTTCAATTTTCTATACGAATTAGAATC
```

为了使同学们更流畅地使用 SPDE，我在各功能的相应位置预留了足够的提示信息，这些多以悬浮的形式提供，可以根据这些信息来运行功能。如下：



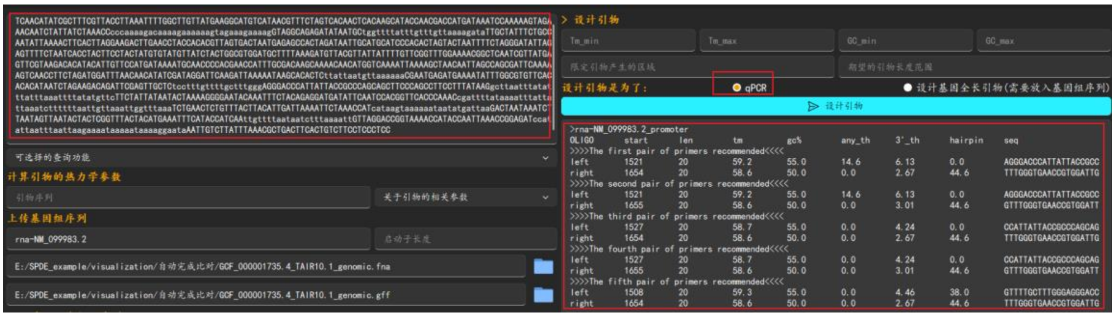
NC_003070.9 23194516 23195103 +

格式: chr/scaffold_ID 起始 终止 正负链. 例如 chr1 2000 2500 +

SPDE 提取的序列均按照正链提供。会对负链序列自动进行反向互补。

IV. 设计定量引物以及基因全长引物是分子实验中常用的两类引物。

在定量引物设计方面，我们根据经验值（包括 Tm 值（50<=Tm<=65）,40<=GC<=60）进行了默认值的设置，即当默认值满足要求时，不需要再进行输入，相应对话框空着就好。如下：



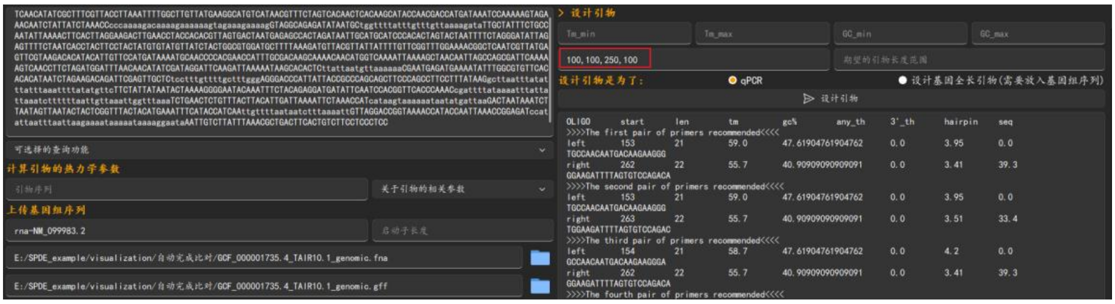
E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

设计引物是为了： ● qPCR ○ 设计基因全长引物(需要放入基因组序列)

```
>rna-NM_099983.2_promoter
OLIGO     start     len      tm      gc%      any_th    3'_th     hairpin   seq
>>>>The first pair of primers recommended<<<<
left      1521      20       59.2    55.0     14.6      6.13      0.0       AGGGACCCATTATTACCGCC
right     1654      20       58.6    50.0     0.0       2.67      44.6      TTTGGGTGAACCGTGGATTG
>>>>The second pair of primers recommended<<<<
left      1521      20       59.2    55.0     14.6      6.13      0.0       AGGGACCCATTATTACCGCC
right     1655      20       58.6    50.0     0.0       3.01      44.6      GTTTGGGTGAACCGTGGATT
>>>>The third pair of primers recommended<<<<
left      1527      20       58.7    55.0     0.0       4.24      0.0       CCATTATTACCGCCCAGCAG
right     1654      20       58.6    50.0     0.0       2.67      44.6      TTTGGGTGAACCGTGGATTG
>>>>The fourth pair of primers recommended<<<<
left      1527      20       58.7    55.0     0.0       4.24      0.0       CCATTATTACCGCCCAGCAG
right     1655      20       58.6    50.0     0.0       3.01      44.6      GTTTGGGTGAACCGTGGATT
>>>>The fifth pair of primers recommended<<<<
left      1508      20       59.3    55.0     0.0       4.46      38.0      GTTTTGCTTTGGGAGGGACC
right     1654      20       58.6    50.0     0.0       2.67      44.6      TTTGGGTGAACCGTGGATTG
```

另一个重要的设置是可以规定引物产生的区域，如下：



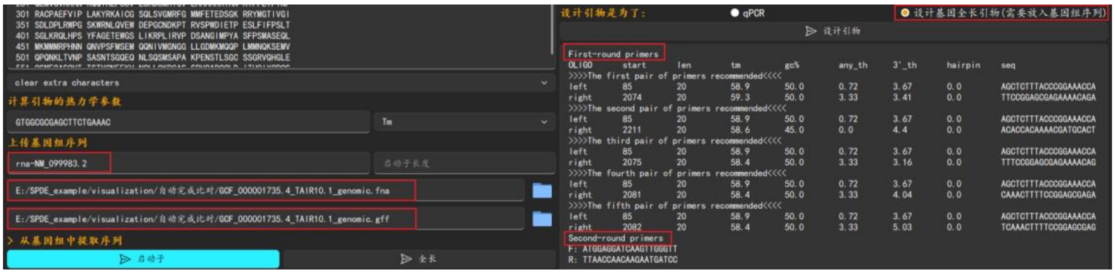
100,100,250,100

```
OLIGO     start     len      tm      gc%                 any_th    3'_th     hairpin   seq
>>>>The first pair of primers recommended<<<<
left      153       21       59.0    47.61904761904762   0.0       3.95      0.0
TGCCAACAATGACAAGAAGGG
right     262       22       55.7    40.90909090909091   0.0       3.41      39.3
GGAAGATTTTAGTGTCCAGACA
>>>>The second pair of primers recommended<<<<
left      153       21       59.0    47.61904761904762   0.0       3.95      0.0
TGCCAACAATGACAAGAAGGG
right     263       22       55.7    40.90909090909091   0.0       3.51      33.4
TGGAAGATTTTAGTGTCCAGAC
>>>>The third pair of primers recommended<<<<
left      154       21       58.7    47.61904761904762   0.0       4.2       0.0
GCCAACAATGACAAGAAGGGA
right     262       22       55.7    40.90909090909091   0.0       3.41      39.3
GGAAGATTTTAGTGTCCAGACA
>>>>The fourth pair of primers recommended<<<<
```

规则：前两个数代表左端引物的位置，其中第一个数是引物的起始位置，第二个数是一个期望的范围。比如，我想让左端引物在100到200这个范围产生，所以设置应该是100,100。<span style="color:red">需要注意的是，输入的时候一律使用**英文输入**</span>，且不做设定的话，逗号不能省略。比如，当我不想对左端引物产生区域进行限制且右端需要在300到400这个范围产生，则设置应该是,,300,100。同样的设置也适用于右端引物。

所有R端引物都已经反向互补，不需要额外操作。

在全长引物设计方面，需要考虑如何排除同源基因的干扰，因为同源基因之间一头一尾的序列通常时类似。基因的UTR区通常是特异的。在SPDEv3.0中，我们会根据基因组序列提取基因对应的UTR区域，然后在UTR区域设计引物（这时的引物被称为第一轮引物）。同学们使用第一轮引物进行扩增后，使用第一轮扩增的产物为模板，之后使用第二轮引物（即基因一头一尾的序列）进行扩增即可得到全长序列。



GTGGCGCGAGCTTCTGAAAC

rna-NM_099983.2

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

设计引物是为了： ○ qPCR ● 设计基因全长引物(需要放入基因组序列)

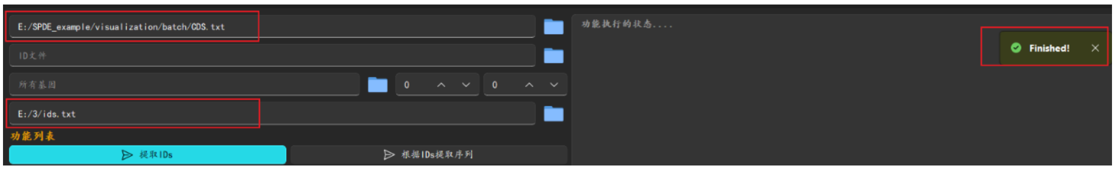
```
First-round primers
OLIGO     start     len      tm      gc%      any_th    3'_th     hairpin   seq
>>>>The first pair of primers recommended<<<<
left      85        20       58.9    50.0     0.72      3.67      0.0       AGCTCTTTACCCGGAAACCA
right     2074      20       59.3    50.0     3.33      3.41      0.0       TTCCGGAGCGAGAAAACAGA
>>>>The second pair of primers recommended<<<<
left      85        20       58.9    50.0     0.72      3.67      0.0       AGCTCTTTACCCGGAAACCA
right     2211      20       58.6    45.0     0.0       4.4       0.0       ACACCACAAAACGATGCACT
>>>>The third pair of primers recommended<<<<
left      85        20       58.9    50.0     0.72      3.67      0.0       AGCTCTTTACCCGGAAACCA
right     2075      20       58.4    50.0     3.33      3.16      0.0       TTTCCGGAGCGAGAAAACAG
>>>>The fourth pair of primers recommended<<<<
left      85        20       58.9    50.0     0.72      3.67      0.0       AGCTCTTTACCCGGAAACCA
right     2081      20       58.4    50.0     3.33      4.04      0.0       CAAACTTTTCCGGAGCGAGA
>>>>The fifth pair of primers recommended<<<<
left      85        20       58.9    50.0     0.72      3.67      0.0       AGCTCTTTACCCGGAAACCA
right     2082      20       58.4    50.0     3.33      5.03      0.0       TCAAACTTTTCCGGAGCGAG
Second-round primers
F: ATGGAGGATCAAGTTGGGTT
R: TTAACCAACAAGAATGATCC
```

2、 多基因模块

该模块主要用于批量地基因ID的提取、序列的提取、序列翻译、GC含量计算、引物设计、pI/molecular_weight/length/extinction_coefficient/hydrophilicity 等一系列参数的计算、重

复序列的移除、文件合并、最长转录本的提取、fasta 文件的格式化以及 fasta/fastq 文件的信息查询等操作。详细操作及规定如下：

提取 fasta 文件中所有的 IDs。输入的文件格式是 fasta。设置保存位置并命名后，点击按键即可完成提取。



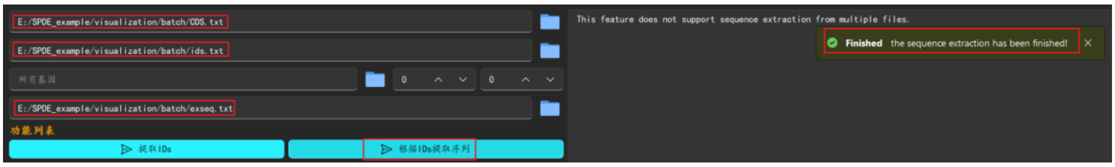
E:/SPDE_example/visualization/batch/CDS.txt

E:/3/ids.txt

功能执行的状态....

Finished!

根据 IDs 提取序列。需要准备两个文件，一个是序列文件；一个是 ID 文件。ID 文件的格式是一个 ID 一行并且应保证 ID 一定在 fasta 文件中。



E:/SPDE_example/visualization/batch/CDS.txt

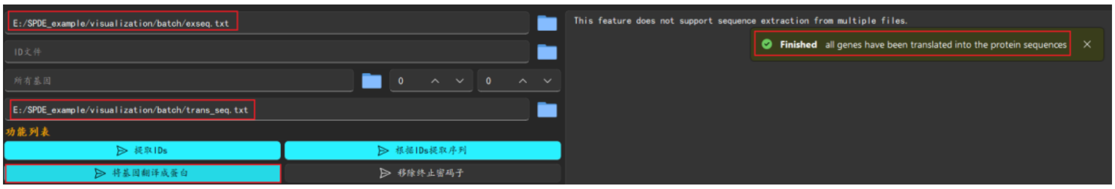
E:/SPDE_example/visualization/batch/ids.txt

E:/SPDE_example/visualization/batch/exseq.txt

This feature does not support sequence extraction from multiple files.

Finished the sequence extraction has been finished!

将 CDS 文件翻译成蛋白序列。输入文件是 CDS 文件，设置保存位置后，点击按钮即可保存翻译的结果。



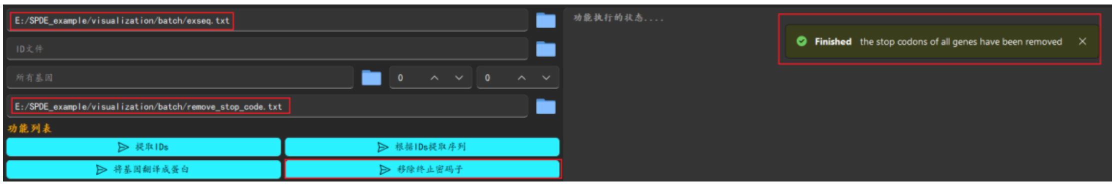
E:/SPDE_example/visualization/batch/exseq.txt

E:/SPDE_example/visualization/batch/trans_seq.txt

This feature does not support sequence extraction from multiple files.

Finished all genes have been translated into the protein sequences

批量去除终止密码子。输入文件是 CDS 文件，设置保存位置后，点击按钮即可保存结果。



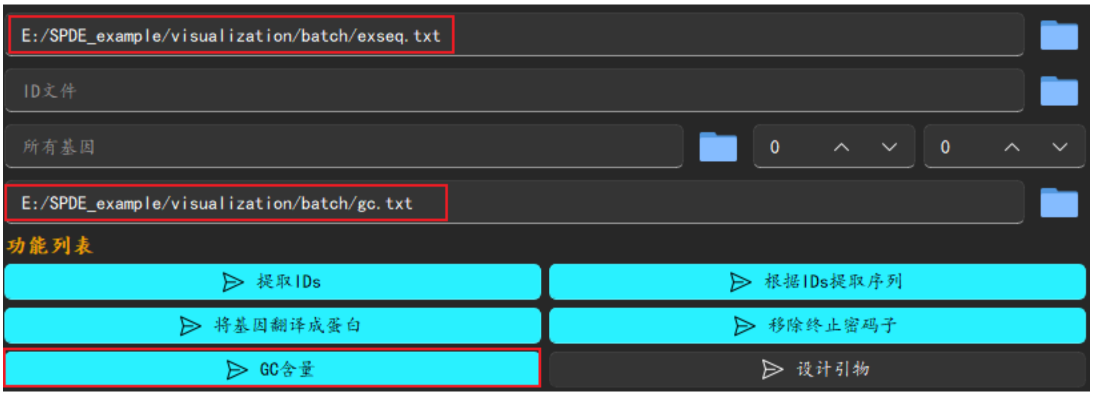
E:/SPDE_example/visualization/batch/exseq.txt

E:/SPDE_example/visualization/batch/remove_stop_code.txt

功能执行的状态....

Finished the stop codons of all genes have been removed

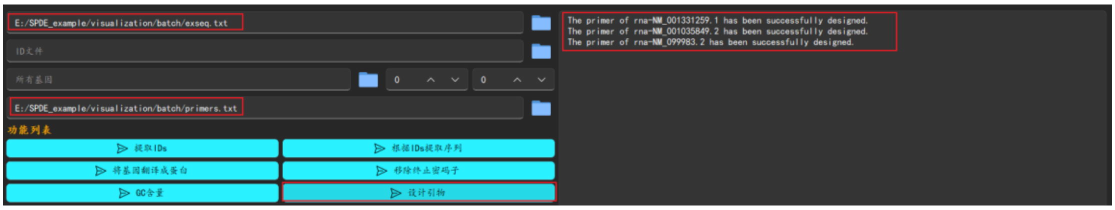
GC 含量的计算。输入文件是 CDS 文件，设置保存位置后，点击按钮即可保存结果。



E:/SPDE_example/visualization/batch/exseq.txt

E:/SPDE_example/visualization/batch/gc.txt

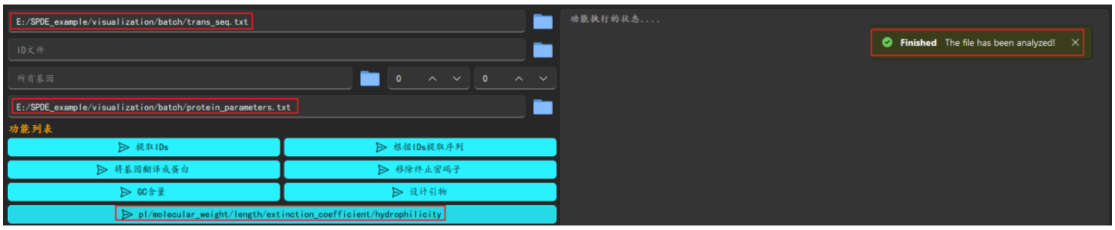
批量进行荧光定量引物的设计。输入文件是 DNA 序列文件（fasta），设置保存位置后，点击按钮即可保存结果。


E:/SPDE_example/visualization/batch/exseq.txt

E:/SPDE_example/visualization/batch/primers.txt

```
The primer of rna-NM_001331259.1 has been successfully designed.
The primer of rna-NM_001035849.2 has been successfully designed.
The primer of rna-NM_099983.2 has been successfully designed.
```

批量分析蛋白相关指标。输入文件是蛋白的序列文件（fasta），设置保存位置后，点击按钮即可保存结果。

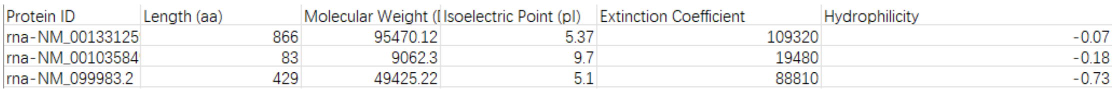

E:/SPDE_example/visualization/batch/trans_seq.txt

E:/SPDE_example/visualization/batch/protein_parameters.txt

功能执行的状态....

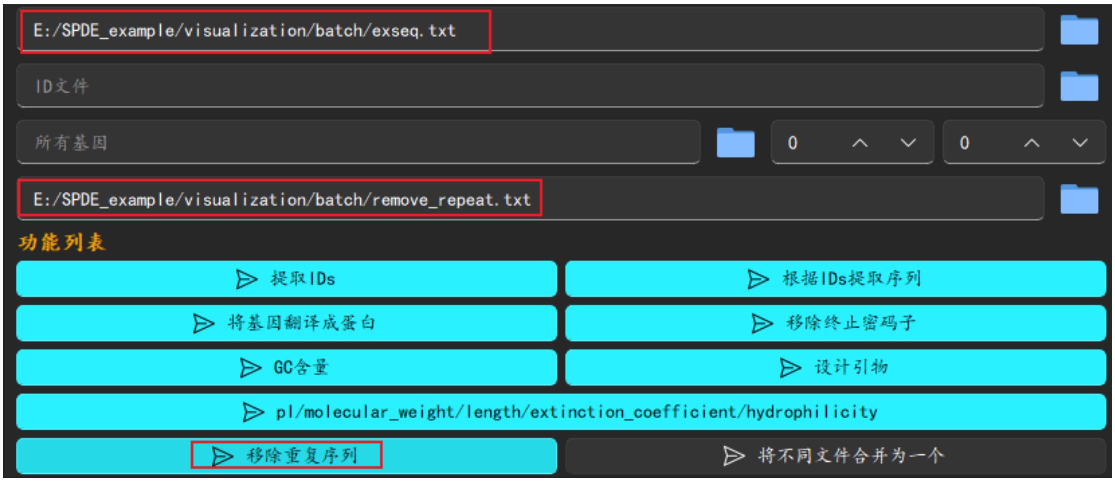
Finished The file has been analyzed!

结果：



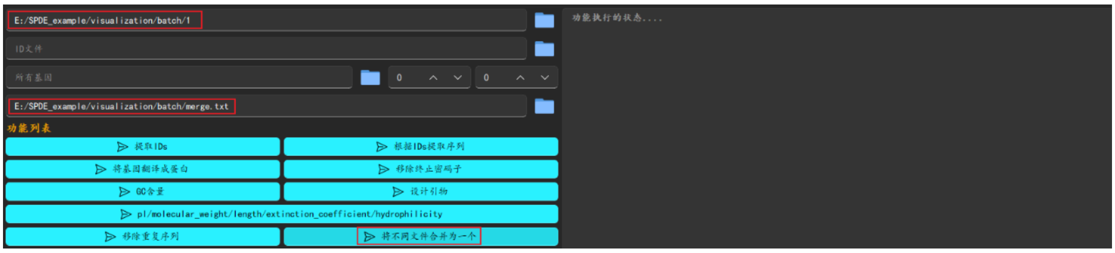
Protein ID	Length (aa)	Molecular Weight (I	Isoelectric Point (pI)	Extinction Coefficient	Hydrophilicity
rna-NM_00133125	866	95470.12	5.37	109320	-0.07
rna-NM_00103584	83	9062.3	9.7	19480	-0.18
rna-NM_099983.2	429	49425.22	5.1	88810	-0.73

去除 fasta 文件中的重复序列。在完成某些操作后可能文件会含有重复序列。比如，翻译 CDS 文件时，有些基因的 CDS 序列不同，但翻译后的蛋白序列则可能相似。


E:/SPDE_example/visualization/batch/exseq.txt

E:/SPDE_example/visualization/batch/remove_repeat.txt

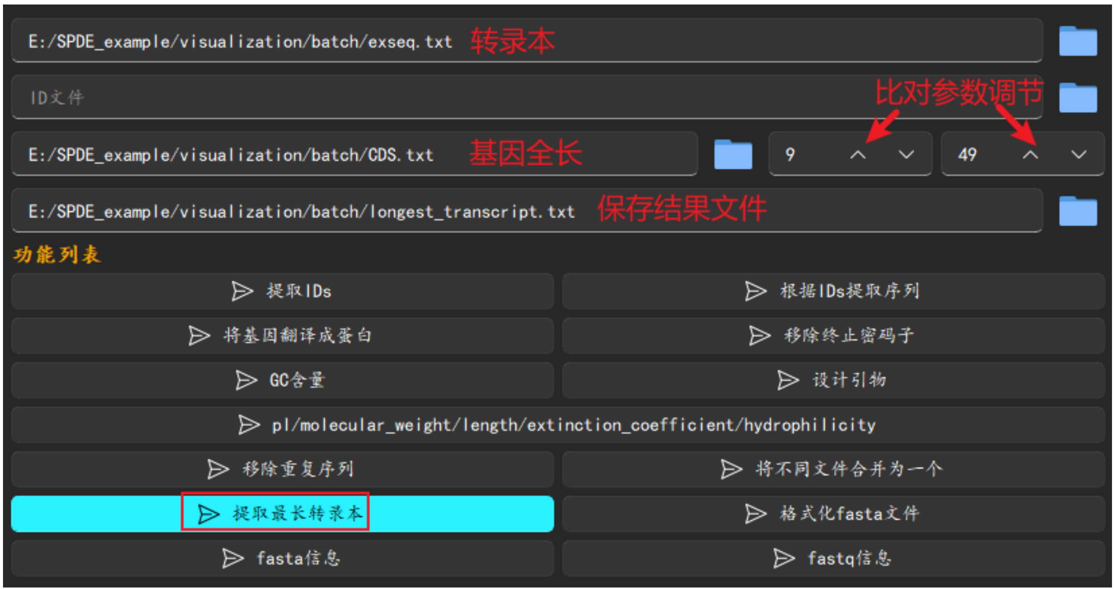
将不同文件合并为一个文件。需要将待合并的文件放入一个空文件夹中，之后将该文件夹作为输入，设置合并后文件的保存位置及名字，点击按钮即可实现不同文件的合并。


E:/SPDE_example/visualization/batch/1

E:/SPDE_example/visualization/batch/merge.txt

功能执行的状态....

提取最长转录本。需要放入两个文件：一个是转录组的拼接序列（或者基因的序列）；第二个是该物种所有基因的基因全长（可以从基因组模块中获取）。其基本过程是：SPDEv3.0会将转录本比对到基因全长文件上。之后，统计根据设置，提取比对信息并进行统计。通过统计结果，分析转录本序列主要集中在哪些基因上。最终返回分析结果。基本操作如下：



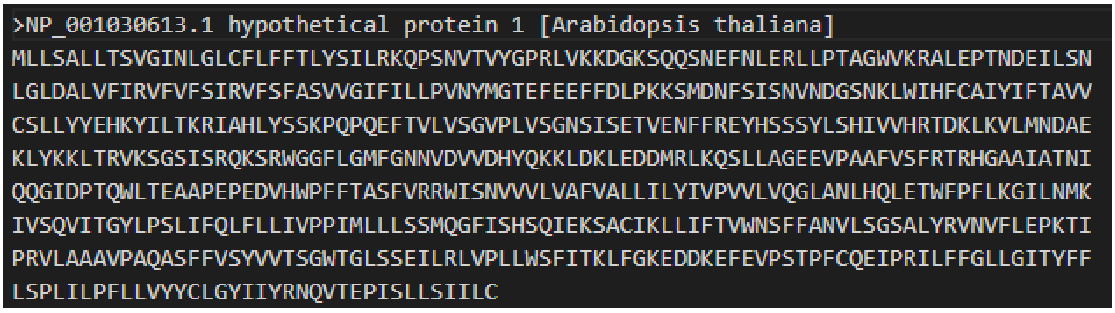
E:/SPDE_example/visualization/batch/exseq.txt 转录本

ID文件 比对参数调节

E:/SPDE_example/visualization/batch/CDS.txt 基因全长 9 49

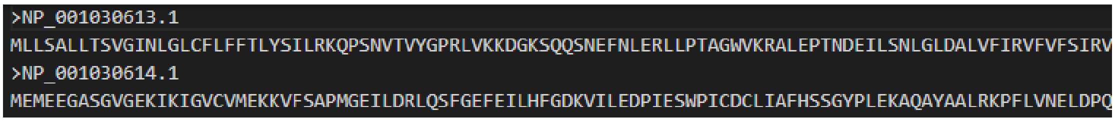
E:/SPDE_example/visualization/batch/longest_transcript.txt 保存结果文件

格式化 fasta 文件。在一些公共数据库（如 NCBI）经常会下载到一个基因分为多行的 fasta 格式的文件如下：



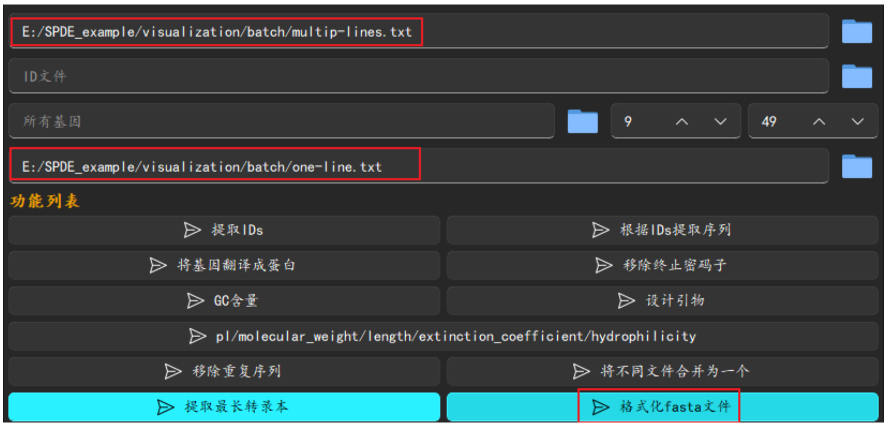
```
>NP_001030613.1 hypothetical protein 1 [Arabidopsis thaliana]
MLLSALLTSVGINLGLCFLFFTLYSILRKQPSNVTVYGPRLVKKDGKSQQSNEFNLERLLPTAGWVKRALEPTNDEILSN
LGLDALVFIRVFVFSIRVFSFASVVGIFILLPVNYMGTEFEEFFDLPKKSMDNFSISNVNDGSNKLWIHFCAIYIFTAVV
CSLLYYEHKYILTKRIAHLYSSKPQPQEFTVLVSGVPLVSGNSISETVENFFREYHSSSYLSHIVVHRTDKLKVLMNDAE
KLYKKLTRVKSGSISRQKSRWGGFLGMFGNNVDVVDHYQKKLDKLEDDMRLKQSLLAGEEVPAAFVSFRTRHGAAIATNI
QQGIDPTQWLTEAAPEPEDVHWPFFTASFVRRWISNVVVLVAFVALLILYIVPVVLVQGLANLHQLETWFPFLKGILNMK
IVSQVITGYLPSLIFQLFLLIVPPIMLLLSSMQGFISHSQIEKSACIKLLIFTVWNSFFANVLSGSALYRVNVFLEPKTI
PRVLAAAVPAQASFFVSYVVTSGWTGLSSEILRLVPLLWSFITKLFGKEDDKEFEVPSTPFCQEIPRILFFGLLGITYFF
LSPLILPFLLVYYCLGYIIYRNQVTEPISLLSIILC
```

这种格式由于回车符（换行符）存在于序列间，想要查找某些序列，都会出现问题。因此，设置了格式化 fasta 文件的功能，该功能是将不同行的 fasta 文件规整到一行中，如下：



```
>NP_001030613.1
MLLSALLTSVGINLGLCFLFFTLYSILRKQPSNVTVYGPRLVKKDGKSQQSNEFNLERLLPTAGWVKRALEPTNDEILSNLGLDALVFIRVFVFSIRV
>NP_001030614.1
MEMEEGASGVGEKIKIGVCVMEKKVFSAPMGEILDRLQSFGEFEILHFGDKVILEDPIESWPICDCLIAFHSSGYPLEKAQAYAALRKPFLVNELDPQ
```

功能设置如下：

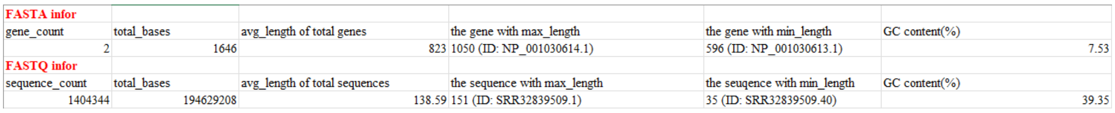

E:/SPDE_example/visualization/batch/multip-lines.txt

ID文件

所有基因 9 49

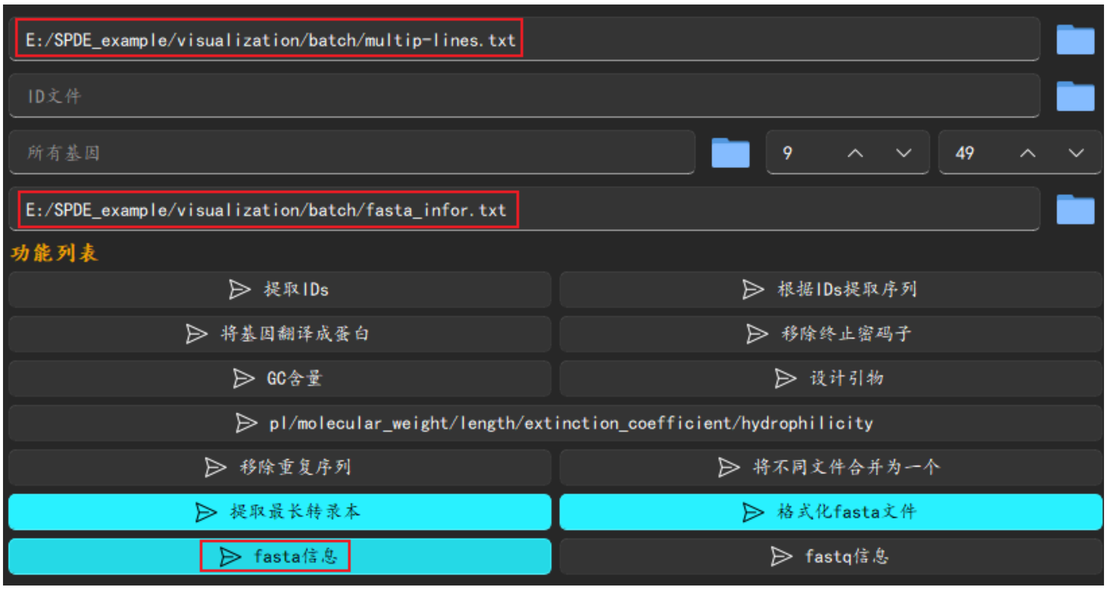
E:/SPDE_example/visualization/batch/one-line.txt

对 fasta 与 fastq 文件信息统计。统计信息如下：



FASTA infor					
gene_count	total_bases	avg_length of total genes	the gene with max_length	the gene with min_length	GC content(%)
2	1646	823	1050 (ID: NP_001030614.1)	596 (ID: NP_001030613.1)	7.53
FASTQ infor					
sequence_count	total_bases	avg_length of total sequences	the sequence with max_length	the seuqence with min_length	GC content(%)
1404344	194629208	138.59	151 (ID: SRR32839509.1)	35 (ID: SRR32839509.40)	39.35

设置如下：

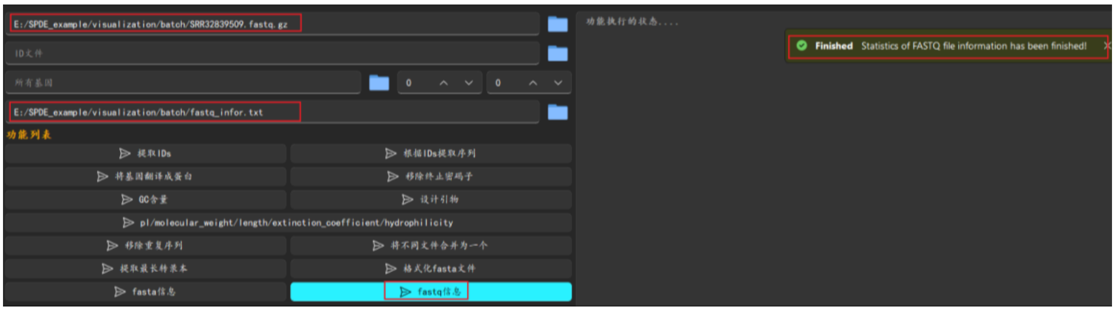

E:/SPDE_example/visualization/batch/multip-lines.txt

ID文件

所有基因 9 49

E:/SPDE_example/visualization/batch/fasta_infor.txt

Fastq 文件分析


E:/SPDE_example/visualization/batch/SRR32839509.fastq.gz

ID文件

所有基因 0 0

E:/SPDE_example/visualization/batch/fastq_infor.txt

功能执行的状态....

Finished Statistics of FASTQ file information has been finished!

请注意：为了适应日益增加的数据分析需求，这个模块中的大多数功能支持更高级别的批量处理。例如，当同学们把需要翻译的多个 CDS 文件放入一个空文件夹并将该文件夹作

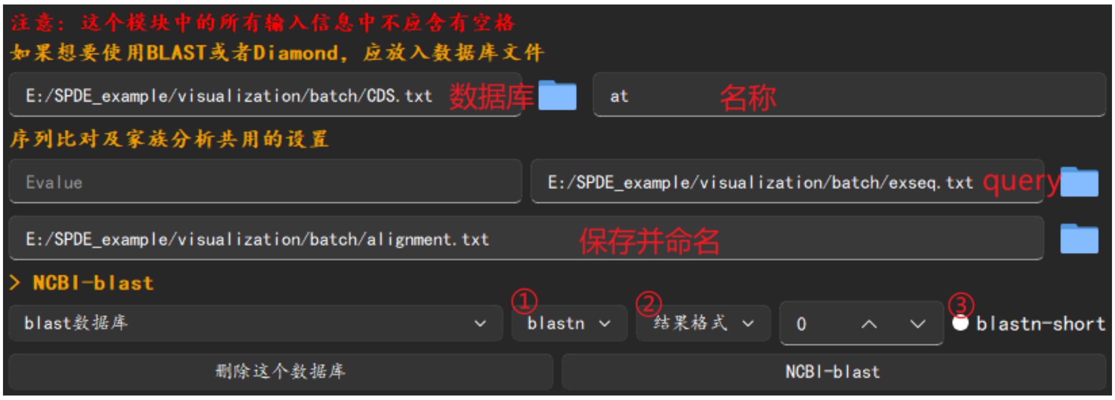
为输入，则 SDPEv3.0 会对该文件夹下每一个文件执行翻译操作。

3、 比对模块

序列比对是整个生物信息学的重要基础。在这个模块中，我们总结了目前常用的比对程序，同学们可以根据自身的需要灵活进行选择。所有操作均可以通过点击实现。

I. Diamond 和 NCBI-Blast

区别：Diamond 的运行速度高于 NCBI-Blast。但 Diamond 的缺点在于该程序只能用于蛋白比对（当然，同学们可能会查到该程序可以比对 DNA 序列，但这种比对有前提：DNA 序列必须能翻译成蛋白。换言之，它所说的 DNA 比对，在运行时还是要先翻译成蛋白，然后再比对。只是不需要同学们翻译而已）。NCBI-Blast 的优点在于可以用于比对更多类型的数据。**共同点：**在使用时，是将需求序列（query）比对库序列（reference），从而分析比对结果的。因此，需要有库文件和 query 文件。建库是必须的，尤其是第一次用这个库。在 SPDEv3.0 中，库的构建是自动的。一旦库构建完成，SPDEv3.0 在下次启动时会自动加载已经构建的库（当然，前提是同学们没有主动删除它）。因此，同一个库不需要同学们反复构建。基本设置如下：



注意：这个模块中的所有输入信息中不应含有空格

如果想要使用BLAST或者Diamond，应放入数据库文件

E:/SPDE_example/visualization/batch/CDS.txt

at

序列比对及家族分析共用的设置

Evalue

E:/SPDE_example/visualization/batch/exseq.txt

E:/SPDE_example/visualization/batch/alignment.txt

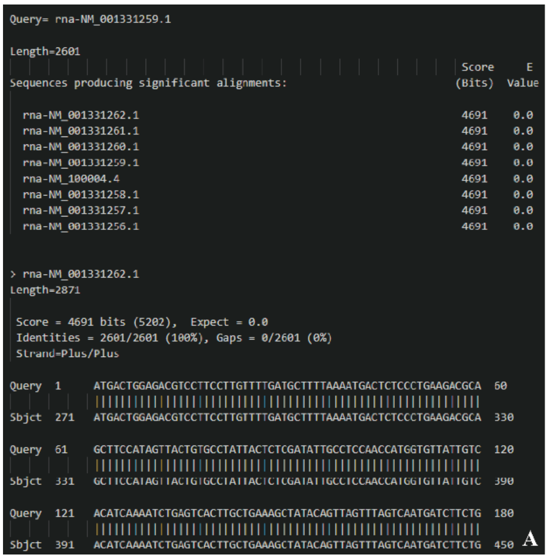
在第一次使用时，直接放入数据库文件并设定名称。加入 query 序列、保存并命名，之后点击 NCBI-blast 运行比对即可。

当再次使用且数据库已经构建，则可直接点击‘blast 数据库’来使用库进行比对。

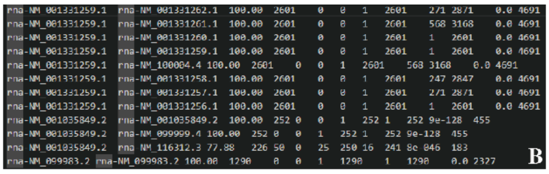
其他选项的说明：

① 是收录的几个 NCBI-blast 的比对方法。Blastn 是 DNA 间的比对；blastp 是蛋白比对；blastx 是专门用于将核酸序列翻译成蛋白质序列后进行比对；tblastn 将给定的氨基酸序列与核酸数据库中的序列（双链）按不同的阅读框进行比对，这对寻找数据库中序列没有标注的新编码区很有用；tblastx 只在特殊情况下使用，它将 DNA 被检索的序列和核酸序列数据库中的序列按不同的阅读框全部翻译成蛋白质序列，然后进行蛋白质序列比对。

② 有两个方式：normal 及 fmt6。这两种方式指的是比对结果的组织形式不同。Normal 的表现形式如下图 A，fmt6 格式如 B：



```
Query= rna-NM_001331259.1

Length=2601
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  rna-NM_001331262.1                                                  4691    0.0
  rna-NM_001331261.1                                                  4691    0.0
  rna-NM_001331260.1                                                  4691    0.0
  rna-NM_001331259.1                                                  4691    0.0
  rna-NM_100004.4                                                     4691    0.0
  rna-NM_001331258.1                                                  4691    0.0
  rna-NM_001331257.1                                                  4691    0.0
  rna-NM_001331256.1                                                  4691    0.0


> rna-NM_001331262.1
Length=2871

 Score = 4691 bits (5202),  Expect = 0.0
 Identities = 2601/2601 (100%), Gaps = 0/2601 (0%)
 Strand=Plus/Plus

Query  1     ATGACTGGAGACGTCCTTCCTTGTTTTGATGCTTTTAAAATGACTCTCCCTGAAGACGCA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  271   ATGACTGGAGACGTCCTTCCTTGTTTTGATGCTTTTAAAATGACTCTCCCTGAAGACGCA  330

Query  61    GCTTCCATAGTTACTGTGCCTATTACTCTCGATATTGCCTCCAACCATGGTGTTATTGTC  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  331   GCTTCCATAGTTACTGTGCCTATTACTCTCGATATTGCCTCCAACCATGGTGTTATTGTC  390

Query  121   ACATCAAAATCTGAGTCACTTGCTGAAAGCTATACAGTTAGTTTAGTCAATGATCTTCTG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  391   ACATCAAAATCTGAGTCACTTGCTGAAAGCTATACAGTTAGTTTAGTCAATGATCTTCTG  450
```



```
rna-NM_001331259.1  rna-NM_001331262.1  100.00  2601    0   0   1   2601    271 2871    0.0 4691
rna-NM_001331259.1  rna-NM_001331261.1  100.00  2601    0   0   1   2601    568 3168    0.0 4691
rna-NM_001331259.1  rna-NM_001331260.1  100.00  2601    0   0   1   2601    1   2601    0.0 4691
rna-NM_001331259.1  rna-NM_001331259.1  100.00  2601    0   0   1   2601    1   2601    0.0 4691
rna-NM_001331259.1  rna-NM_100004.4 100.00  2601    0   0   1   2601    568 3168    0.0 4691
rna-NM_001331259.1  rna-NM_001331258.1  100.00  2601    0   0   1   2601    247 2847    0.0 4691
rna-NM_001331259.1  rna-NM_001331257.1  100.00  2601    0   0   1   2601    271 2871    0.0 4691
rna-NM_001331259.1  rna-NM_001331256.1  100.00  2601    0   0   1   2601    1   2601    0.0 4691
rna-NM_001035849.2  rna-NM_001035849.2  100.00  252 0   0   1   252 1   252 9e-128  455
rna-NM_001035849.2  rna-NM_099999.4 100.00  252 0   0   1   252 1   252 9e-128  455
rna-NM_001035849.2  rna-NM_116312.3 77.88   226 50  0   25  250 16  241 8e-046  183
rna-NM_099983.2 rna-NM_099983.2 100.00  1290    0   0   1   1290    1   1290    0.0 2327
```

fmt6各列格式说明：

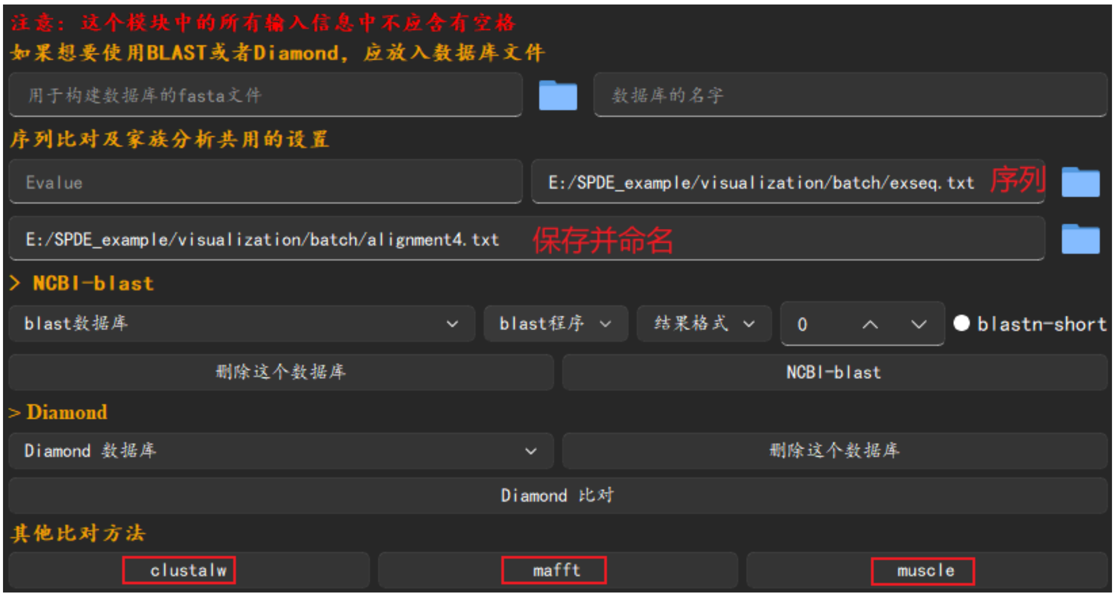
qseqid	查询序列的ID
sseqid	库序列的ID
pident	查询和库序列之间的百分比一致性
length	比对的长度
mismatch	不匹配的数量
gapopen	间隙的数量
qstart	查询序列的起始位置
qend	查询序列的结束位置
sstart	库序列的起始位置
send	库序列的结束位置
evalue	E值
bitscore	比分值

③ 选项用于将短序列比对到数据库中（换言之，没有勾选这个选项，blast 是无法比对短序列的）。

Diamond 使用方法与 NCBI-blast 类似。

II. clustalw、mafft 与 muscle

与 Diamond 和 NCBI-blast 不同，运行 clustalw、mafft 与 muscle 时，不需要进行数据库的构建和选择。它们更多是输入序列自身的比对。基本设置为：



注意：这个模块中的所有输入信息中不应含有空格

如果想要使用BLAST或者Diamond，应放入数据库文件

用于构建数据库的fasta文件

数据库的名字

序列比对及家族分析共用的设置

Evalue

E:/SPDE_example/visualization/batch/exseq.txt 序列

E:/SPDE_example/visualization/batch/alignment4.txt 保存并命名

输入的序列同样也是 fasta 格式。

III. 基因家族

SPDEv3.0 对基因家族分析做了进一步优化：收集和整理了超过 120 个家族的特征性结构域。在选择特定家族后，可以一键完成家族成员筛选、序列提取、结构域的识别与可视化过程。几个需要注意的问题：

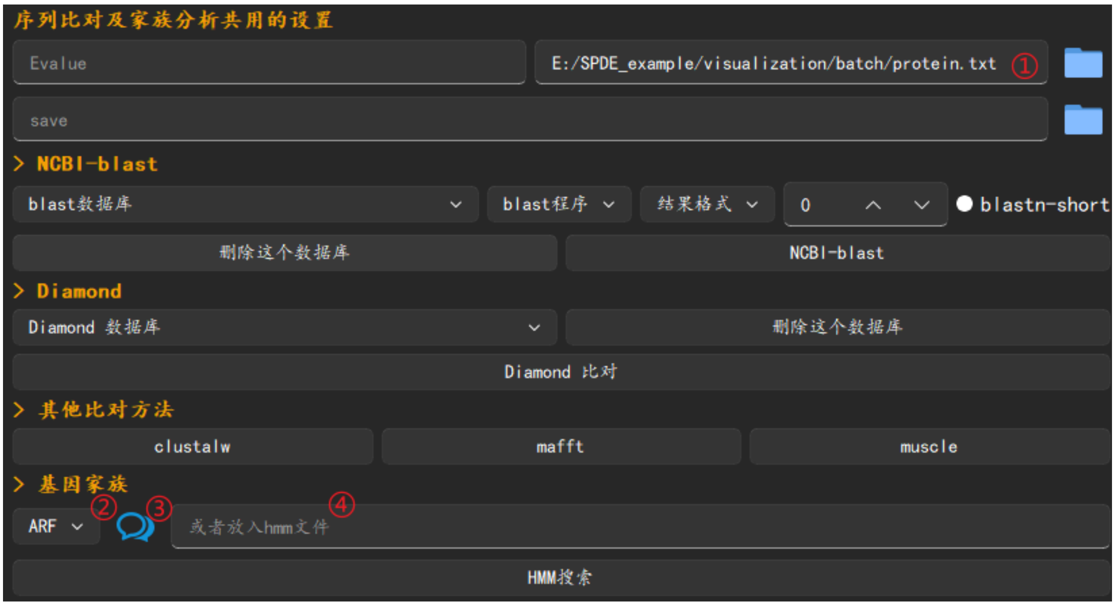
1） 输入的序列应为蛋白序列（fasta 格式）；

2） 特征性结构域的含义是：这个家族有别于其他家族的结构域。当一个家族含有多个结构域而其中的一些结构域并不是所有成员都具有的时候，这些结构域被排除在外；

3） 允许用户自行构建并输入结构域模型。构建的方法是：下载感兴趣家族的蛋白结构域的 pfam 文件，如果一个家族含有多个必须结构域时，需要将这些 pfam 文件合并成一个 pfam 文件。Pfam 下载时需要知道蛋白结构域的 pfam ID，这个 pfam ID 可以通过基因家族文章获取或者找一个一定是这个家族的蛋白序列，使用 NCBI-BLAST(https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) 来确定 ID。在明确 ID 后，可以参考我的这篇博客（https://www.jianshu.com/p/934f63b4b1a3?v=1742889666646）来下载相应的 pfam 文件。

4） 在我们的 github 账号（https://github.com/simon19891216/SPDE/releases/tag/database）中放置了全部的 pfam 数据库，主要用于对家族成员蛋白结构域的全面识别。不过这个库比较大，同学们可以根据自己的情况选择是否要下载及使用。使用方法是：下载数据库；打开 SPDEv3.0 的安装位置，你可以看到一个 database/all_pfam 文件夹；将下载的文件放入该文件夹下并右键单击选择‘解压到当前文件夹’。使用的时候程序会自动调取。

基本参数设置如下：

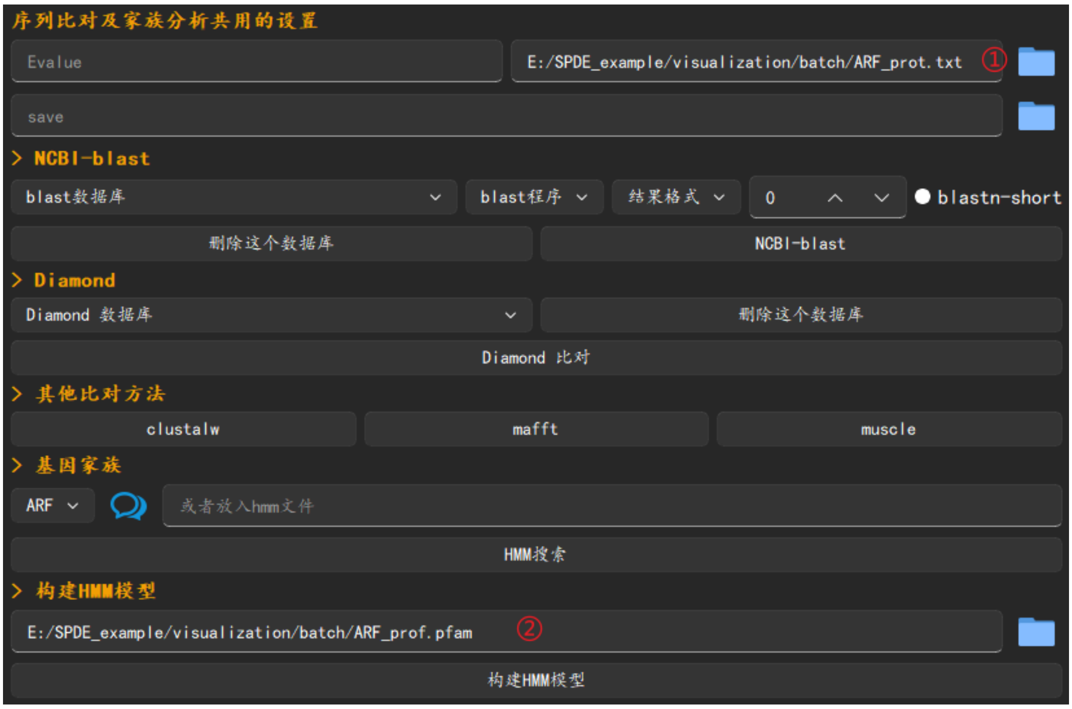


E:/SPDE_example/visualization/batch/protein.txt

① 处放入蛋白序列；②选择模型或者在④处放入自己组织的模型；③是关于该模型全称的按钮，点击会显示家族全称。结果放置在蛋白序列的同级文件夹下。

IV. 隐马尔可夫模型（HMM）的构建

如果同学们研究的家族还没有被充分研究，这个时候可能会考虑自己构建模型来进行成员的识别。我在 SPDEv3.0 中设置了 HMM 模型构建的功能，基本设置如下：



E:/SPDE_example/visualization/batch/ARF_prot.txt ①

E:/SPDE_example/visualization/batch/ARF_prof.pfam ②

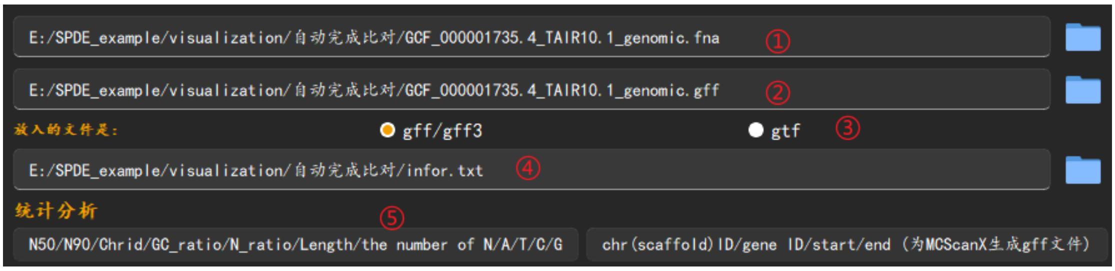
① 处放入序列；②放入保存模型的位置及命名。

需要注意的问题是放入的序列应该具有保守性的结构。需要同学们先做判断，确定有保守序列之后，在进行模型构建。

4、 基因组模块

该模块主要用于基因组层级的数据分析以及序列提取等操作，执行的主要功能是：信息查询（包括 N50，碱基数量分布，gc 含量等）；特定元素（包括 kmer，gene，mRNA，CDS，exon，gap 的位置分布（仅用于 windows 版本）等）在基因组上的分布；批量的序列提取（包括启动子、mRNA、gene、CDS、UTR 序列）以及基因组的共线性分析。基本设置如下：

I. 基因组信息查询


E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna ①

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff ②

放入的文件是： ● gff/gff3 ○ gtf ③

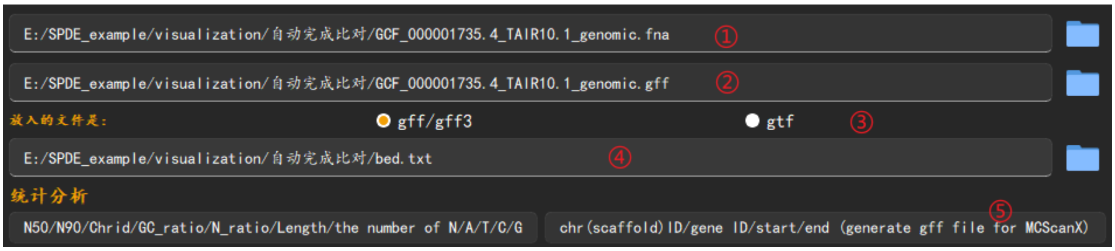
E:/SPDE_example/visualization/自动完成比对/infor.txt ④

统计分析
⑤
N50/N90/Chrid/GC_ratio/N_ratio/Length/the number of N/A/T/C/G
chr(scaffold)ID/gene ID/start/end (为MCScanX生成gff文件)

①基因组序列文件；②gff 或 gtf 注释文件，同时需要在③处对输入格式进行注明；④保存并命名；点击⑤执行功能。

II. bed 文件的生成

在不少程序中会用到 bed 文件。SPDE 生成 bed 文件的格式：chr ID, gene ID，起始位置，终止位置。



E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

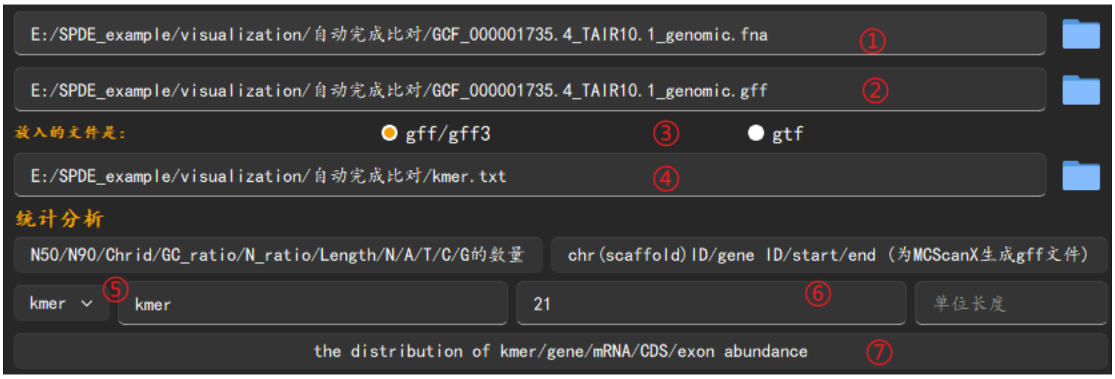
放入的文件是： ● gff/gff3 ○ gtf

E:/SPDE_example/visualization/自动完成比对/bed.txt

统计分析
N50/N90/Chrid/GC_ratio/N_ratio/Length/the number of N/A/T/C/G
chr(scaffold)ID/gene ID/start/end (generate gff file for MCScanX)
①
②
③
④
⑤

①基因组序列文件；②gff 或 gtf 注释文件，同时需要在③处对输入格式进行注明；④保存并命名；点击⑤执行功能。

III. 元素在基因组上的密度分布



E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

放入的文件是： ● gff/gff3 ○ gtf

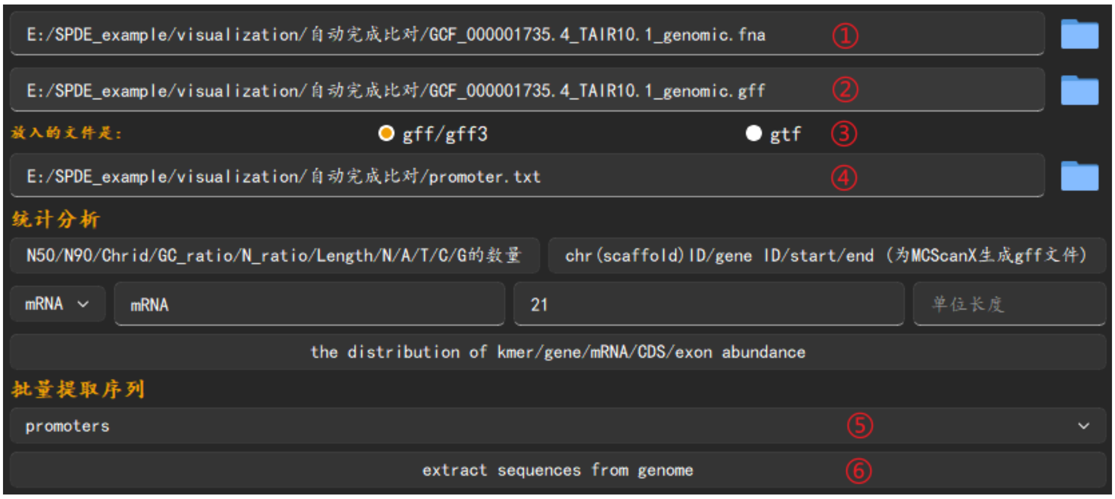
E:/SPDE_example/visualization/自动完成比对/kmer.txt

kmer | kmer | 21 | 单位长度

统计分析
N50/N90/Chrid/GC_ratio/N_ratio/Length/N/A/T/C/G的数量
chr(scaffold)ID/gene ID/start/end (为MCScanX生成gff文件)
the distribution of kmer/gene/mRNA/CDS/exon abundance
①
②
③
④
⑤
⑥
⑦

①基因组序列文件；②gff 或 gtf 注释文件，同时需要在③处对输入格式进行注明；④保存并命名；点击⑤选择类型；如果是 kmer，可以在⑥选择设定 kmer 的大小；点击⑦执行功能

IV. 批量提取序列



E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

放入的文件是： ● gff/gff3 ○ gtf

E:/SPDE_example/visualization/自动完成比对/promoter.txt

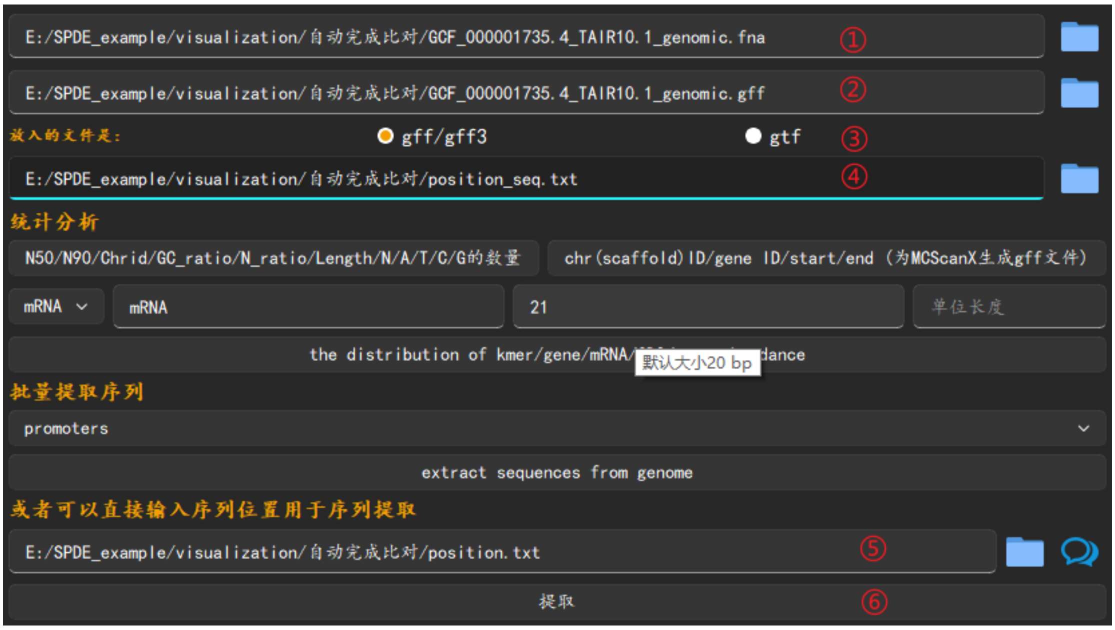
mRNA | mRNA | 21 | 单位长度

promoters

统计分析
N50/N90/Chrid/GC_ratio/N_ratio/Length/N/A/T/C/G的数量
chr(scaffold)ID/gene ID/start/end (为MCScanX生成gff文件)
the distribution of kmer/gene/mRNA/CDS/exon abundance
批量提取序列
extract sequences from genome
①
②
③
④
⑤
⑥

①基因组序列文件；②gff 或 gtf 注释文件，同时需要在③处对输入格式进行注明；④保存并命名；点击⑤选择类型；点击⑥执行功能

V. 自由的序列提取


放入的文件是：
○ gff/gff3
● gtf
统计分析
N50/N90/Chrid/GC_ratio/N_ratio/Length/N/A/T/C/G的数量
chr(scaffold)ID/gene ID/start/end（为MCScanX生成gff文件）
单位长度
the distribution of kmer/gene/mRNA/
dance
默认大小20 bp
批量提取序列
extract sequences from genome
或者可以直接输入序列位置用于序列提取
提取
①
②
③
④
⑤
⑥

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

放入的文件是：gff/gff3

E:/SPDE_example/visualization/自动完成比对/position_seq.txt

mRNA　mRNA　21

promoters

E:/SPDE_example/visualization/自动完成比对/position.txt

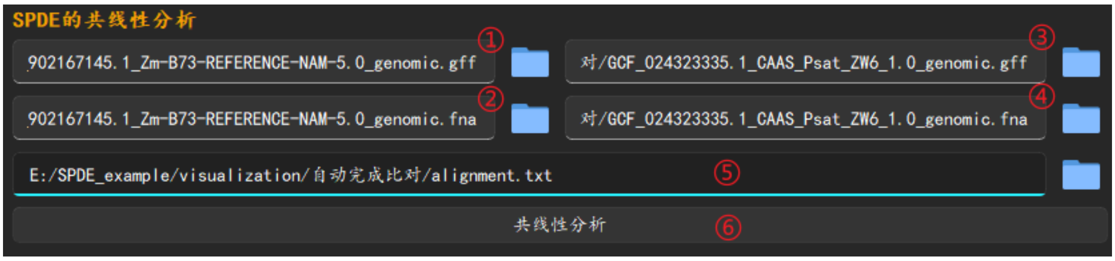
①基因组序列文件；②gff 或 gtf 注释文件，同时需要在③处对输入格式进行注明；④保存并命名；⑤输入位置文件；点击⑥执行功能

位置文件的格式是：染色体 ID，起始位置，终止位置，基因 ID，正负链

VI. SPDE 的共线性分析

以往的共线性分析需要大家准备多个文件，其中间过程复杂。在 SPDEv3.0 中，我们进一步优化了过程。我们使用基因组超过 2Gb 的物种进行测试，结果发现两个物种的共线性

分析可以在 2 分钟内完成。功能的输入本身也很简单，只需要基因组序列文件及 GFF 文件，基本设置如下图：



902167145.1_Zm-B73-REFERENCE-NAM-5.0_genomic.gff

902167145.1_Zm-B73-REFERENCE-NAM-5.0_genomic.fna

对/GCF_024323335.1_CAAS_Psat_ZW6_1.0_genomic.gff

对/GCF_024323335.1_CAAS_Psat_ZW6_1.0_genomic.fna

E:/SPDE_example/visualization/自动完成比对/alignment.txt

①和②以及③和④分别对应两个物种的 gff 文件以及基因组序列文件；⑤设置比对的保存位置；点击⑥执行功能。

我们同时为该功能配置了可视化功能（这部分将在后续介绍）。用于可视化的文件是带有‘collinearity_visualization.txt’标志的文件。

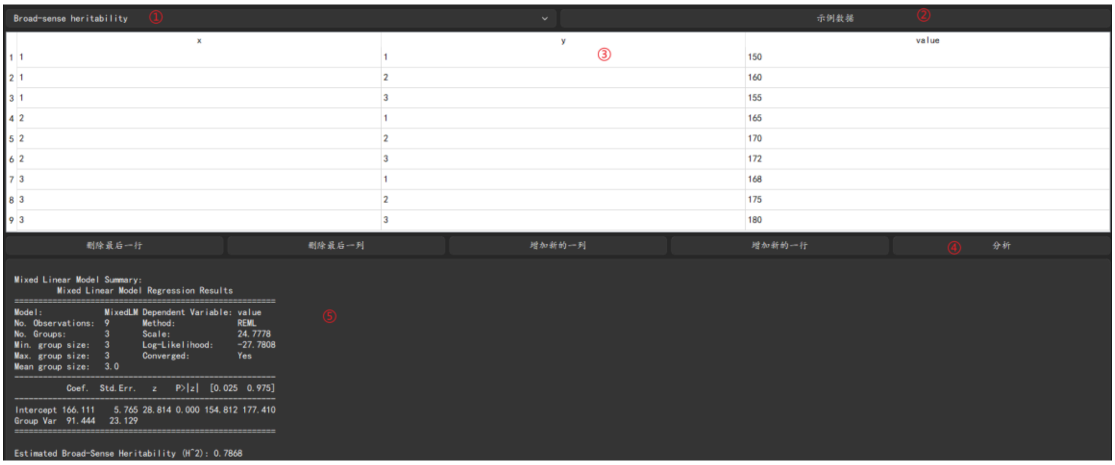
5、 育种计算模块

之前的育种学计算依赖于 SAS 或者 SPSS 等统计学软件，然而这些软件并不是专门为育种学设计，这使得整个计算功能较为复杂。在 SPDEv3.0 中，我们设计了超过 40 种育

种学方面的计算功能以高效完成育种学计算。涉及的育种学计算方法，如下表：

Analysis of variance	方差分析	Partial correlation coefficient	偏相关系数
Balanced incomplete block design	平衡不完全区组设计	Rank correlation coefficient	秩相关系数
Completely randomized design	完全随机设计	Power regression	幂回归
Latin square design	拉丁方设计	Principal component analysis	主成分分析
Split-plot design	裂区设计	Cluster analysis (Hierarchical Clustering)	聚类分析（层次聚类）
Three-way randomized block design	三因子随机区组设计	Cluster analysis (K-Means)	聚类分析（K-均值）
Two-way ANOVA	双因素方差分析	Dynamic clustering method	动态聚类方法
Two-way randomized block design	双因素随机区组设计	Trait clustering analysis	性状聚类分析
Two-way analysis of variance with equal replication	双因素方差分析（等重复）	Confidence interval estimation	置信区间估计
Two-factor variance analysis without replication	无重复双因素方差分析	Covariance analysis	协方差分析
Two-way MANOVA with equal replication	双因素多变量方差分析（等重复）	Combining ability analysis	配合力分析
One-way ANOVA	单因素方差分析	Diallel cross combining ability	双列杂交配合力分析
One-way MANOVA with equal replication	单因素多变量方差分析（等重复）	Discriminant analysis	判别分析
Orthogonal test analysis	正交试验分析	Stepwise discriminant analysis	逐步判别分析
Broad-sense heritability	广义遗传力	Homogeneity of variance test	方差齐性检验
Narrow-sense heritability	狭义遗传力	Normality test	正态性检验
Genetic correlation	遗传相关性	Multiple comparison method	多重比较方法
Heritability Estimation (Half-Sibling Model)	遗传力估算（半同胞模型）	Paired significance test	配对显著性检验
Variance components analysis	方差成分分析	Rank sum test	秩和检验
Estimated variance components	方差成分估计	Residual analysis	残差分析
Canonical correlation analysis	典型相关分析	Significance analysis	显著性分析
Multi-correlation coefficient	复相关系数	Significant difference	显著性差异

为实现高效分析，该模块对输入数据的组织形式有很高的要求。因此，在进行功能设计时，我做了严格的限制，只允许用户在示例数据的基础上，用自己的数据替换示例数据来进行分析。基本操作过程如下：



	x	y	value
1	1	1	150
2	1	2	160
3	1	3	155
4	2	1	165
5	2	2	170
6	2	3	172
7	3	1	168
8	3	2	175
9	3	3	180

```
Mixed Linear Model Summary:
         Mixed Linear Model Regression Results
==========================================================
Model:             MixedLM Dependent Variable: value
No. Observations:  9       Method:             REML
No. Groups:        3       Scale:              24.7778
Min. group size:   3       Log-Likelihood:     -27.7808
Max. group size:   3       Converged:          Yes
Mean group size:   3.0
----------------------------------------------------------
          Coef.  Std.Err.   z    P>|z|  [0.025  0.975]
----------------------------------------------------------
Intercept 166.111    5.765 28.814 0.000 154.812 177.410
Group Var  91.444   23.129
==========================================================

Estimated Broad-Sense Heritability (H^2): 0.7868
```

同学们首先通过①选择你需要的分析方法，之后点击②自动调取示例数据。如上图所示，

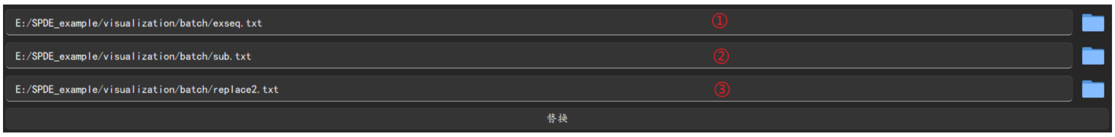
当计算广义遗传力时，只需要将你自己数据中的 x，y 以及由此产生的育种值（即 value）替换掉示例中的数据，之后点击④完成计算。计算结果展示在⑤。当然，不要求同学们的数据必须和示例数据行列数一致，可以通过设置的增减行列按钮进行行列的删除或者增加。

6、 文件编辑模块

这个模块主要用于处理一些常见的文本需求，包括了替换、提取、重新组织文件内容、转换文件格式、下载文件以及浏览大文件。以下是关于内容的具体应用：

I. 替换文件内容

在常用软件中很难一次性进行多个文本内容的替换而这种替换在生物信息学中经常用到。比如我们接触的序列文件中基因的名字没有实际的功能意义，因此需要将原始 ID 替换为具有功能意义的 ID。基本设置如下：



E:/SPDE_example/visualization/batch/exseq.txt

E:/SPDE_example/visualization/batch/sub.txt

E:/SPDE_example/visualization/batch/replace2.txt

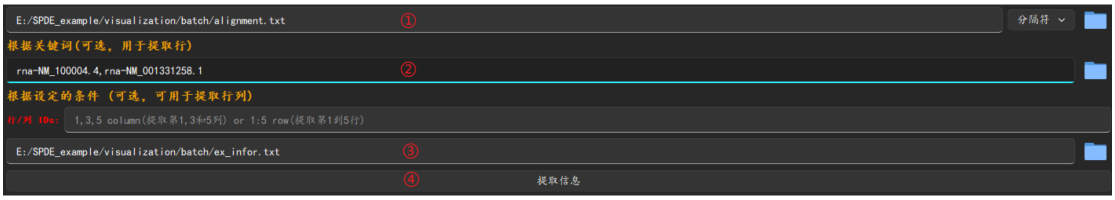
在①放入原始文件；在②放入要替换的内容。这里有两种形式：1、替换内容较多时可整理成文件，一组替换内容一行，每一组有两列，第一列是原始内容，第二列是要替换的内容，两列由空格隔开。例如，rna-NM_099983.2 ARF3。因此，有一个需要注意的点：替换前后的内容中不能有空格。2、如果替换内容不多，则格式按照 apple:orange,banana:grape,cat:dog 进行，即由逗号隔开不同组，每一组由冒号进行分列；设置保存位置并命名；按功能键执行功能。

注意，原始文件需要是文本文件而不是由 word 和 excel 产生的文件。

II. 提取需要的信息

在日常分析中，有一些文件含有大量内容而有些内容是后续分析所不需要的。因此，需要设计一定的功能来高效提取需要的内容。

1）通过关键词提取相应行（关键词间以逗号分隔，注意英文状态下输入）



E:/SPDE_example/visualization/batch/alignment.txt ①　分隔符

根据关键词(可选，用于提取行)

rna-NM_100004.4,rna-NM_001331258.1 ②

根据设定的条件（可选，可用于提取行列）

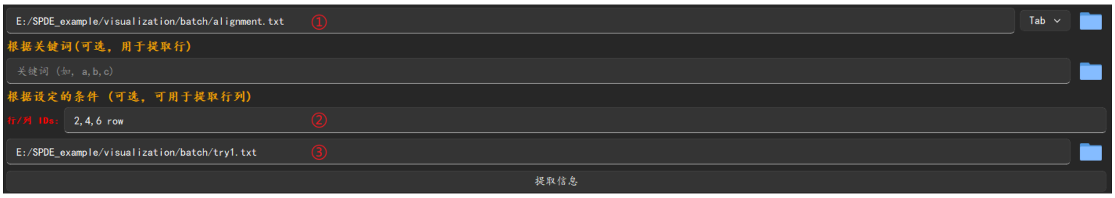
行/列 IDs: 1,3,5 column(提取第1,3和5列) or 1:5 row(提取第1到5行)

E:/SPDE_example/visualization/batch/ex_infor.txt ③

④ 提取信息

①放入文件；②放入关键词；③设置保存位置并命名；按④运行功能

2）按照行 ID 提取信息



E:/SPDE_example/visualization/batch/alignment.txt ①　Tab

根据关键词(可选，用于提取行)

关键词（如，a,b,c）

根据设定的条件（可选，可用于提取行列）

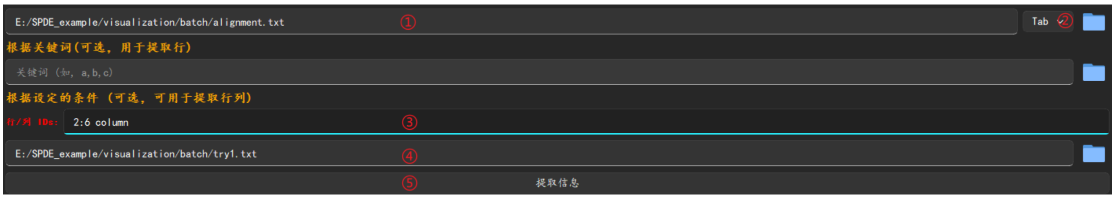
行/列 IDs: 2,4,6 row ②

E:/SPDE_example/visualization/batch/try1.txt ③

提取信息

①放入文件；②放入行 ID（不连续的行，用逗号分隔；连续的行，用冒号分隔；并使用 row 关键词来告诉软件要提取的是行；行 ID 与 row 之间用空格分隔。例如，2,4,6 row）；③设置保存位置并命名；按键运行功能。

3）按照列 ID 提取信息



E:/SPDE_example/visualization/batch/alignment.txt ①　Tab ②

根据关键词(可选，用于提取行)

关键词（如，a,b,c）

根据设定的条件（可选，可用于提取行列）

行/列 IDs: 2:6 column ③

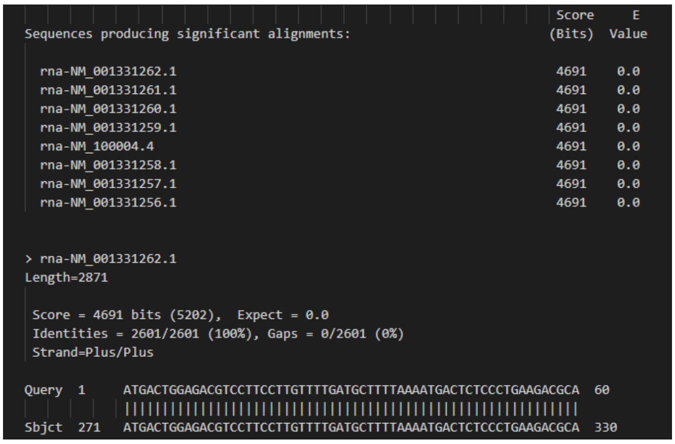
E:/SPDE_example/visualization/batch/try1.txt ④

⑤ 提取信息

①放入文件；②需要告知软件不同列的分隔符是什么（**这是与行提取最大的区别**）；③放入列 ID（规则与行 ID 的放置方式相同）；④保存并命名；按⑤执行功能。

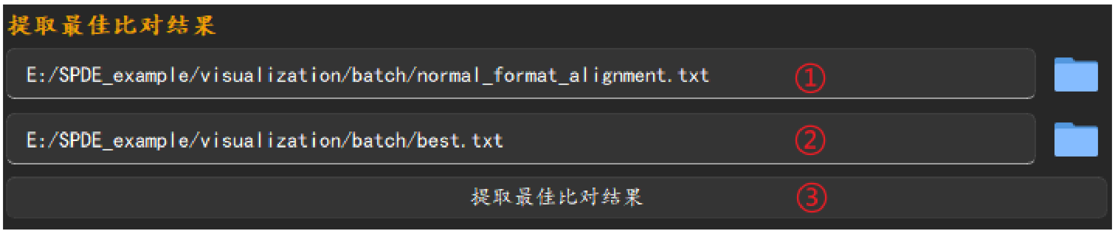
III. 提取最佳比对结果

在 NCBI-blast 的正常比对（normal 格式）中，可以显示详细的比对细节如下：



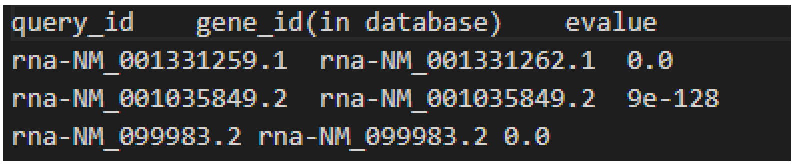
```
                                                                Score     E
Sequences producing significant alignments:                    (Bits)  Value

  rna-NM_001331262.1                                             4691    0.0
  rna-NM_001331261.1                                             4691    0.0
  rna-NM_001331260.1                                             4691    0.0
  rna-NM_001331259.1                                             4691    0.0
  rna-NM_100004.4                                                4691    0.0
  rna-NM_001331258.1                                             4691    0.0
  rna-NM_001331257.1                                             4691    0.0
  rna-NM_001331256.1                                             4691    0.0


> rna-NM_001331262.1
Length=2871

 Score = 4691 bits (5202),  Expect = 0.0
 Identities = 2601/2601 (100%), Gaps = 0/2601 (0%)
 Strand=Plus/Plus

Query  1    ATGACTGGAGACGTCCTTCCTTGTTTTGATGCTTTTAAAATGACTCTCCCTGAAGACGCA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  271  ATGACTGGAGACGTCCTTCCTTGTTTTGATGCTTTTAAAATGACTCTCCCTGAAGACGCA  330
```

但这种格式不利于我们查看最佳的比对。因此，设置了提取最佳比对结果的功能，基本设置如下：


提取最佳比对结果
E:/SPDE_example/visualization/batch/normal_format_alignment.txt ①
E:/SPDE_example/visualization/batch/best.txt ②
提取最佳比对结果 ③

① 放入文件；②保存并命名；按③执行功能。

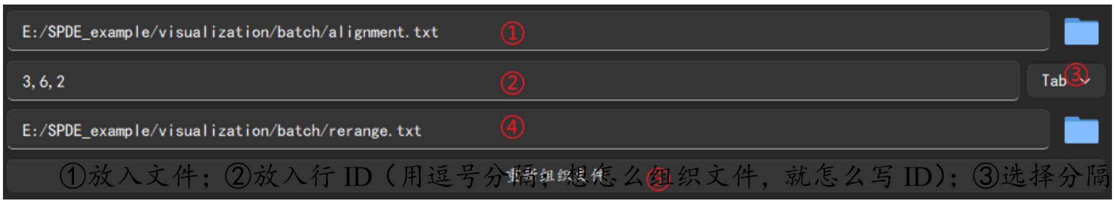
结果格式如下：



```
query_id     gene_id(in database)     evalue
rna-NM_001331259.1  rna-NM_001331262.1  0.0
rna-NM_001035849.2  rna-NM_001035849.2  9e-128
rna-NM_099983.2 rna-NM_099983.2 0.0
```

VI. 重新组织文件

不同的工具对输入文件的内容有特定的格式要求，为满足这种要求，通常需要对原始文件的内容进行重新组织。基本设置如下：

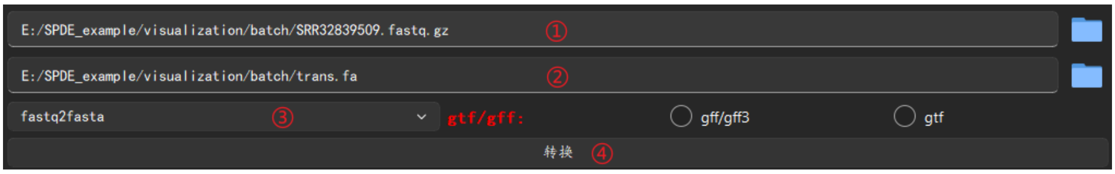

E:/SPDE_example/visualization/batch/alignment.txt ①
3,6,2 ②
Tab ③
E:/SPDE_example/visualization/batch/rerange.txt ④
重新组织文件 ⑤

①放入文件；②放入行 ID（用逗号分隔，想怎么组织文件，就怎么写 ID）；③选择分隔符；④设置保存位置并命名；按⑤执行功能。

V. 格式转换

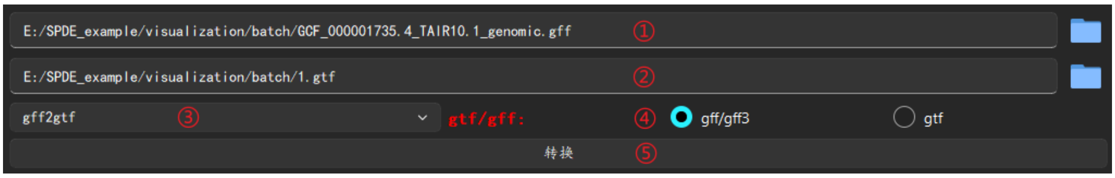
在进行生信分析时经常会涉及文件的格式转换等操作。在 SPDEv3.0 中，我们设计了五类文件的格式转换，包括：fastq 转 fasta，tree 转 nwk，sam 转 paf，gbff 转 gff 以及 gff 转 gtf。其中，当使用 gff 与 gtf 进行互相转换的时候，需要指明输入的文件格式。基本设置如下：

1）fastq 转 fasta


E:/SPDE_example/visualization/batch/SRR32839509.fastq.gz ①
E:/SPDE_example/visualization/batch/trans.fa ②
fastq2fasta ③
gtf/gff:
○ gff/gff3
○ gtf
转换 ④

①放入文件；②设置保存位置并命名文件；③选择转换类型；④按键运行功能。

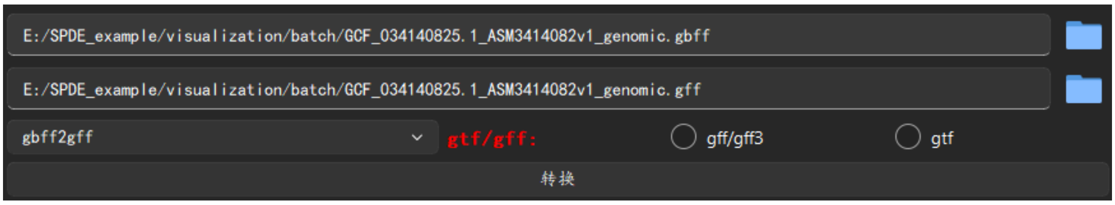
2）gff 与 gtf 间的相互转换


E:/SPDE_example/visualization/batch/GCF_000001735.4_TAIR10.1_genomic.gff ①
E:/SPDE_example/visualization/batch/1.gtf ②
gff2gtf ③
gtf/gff: ④
● gff/gff3
○ gtf
转换 ⑤

①放入文件；②设置保存位置并命名文件；③选择转换类型；④需要指明输入的文件类型；按⑤执行功能

3）GBFF 转 GFF

接下来的一些操作与之前是一致的。这些示例文件我会给同学们传到 github 账户上，请同学们在使用功能前，一定办法核对好文件格式。



E:/SPDE_example/visualization/batch/GCF_034140825.1_ASM3414082v1_genomic.gbff

E:/SPDE_example/visualization/batch/GCF_034140825.1_ASM3414082v1_genomic.gff

gbff2gff

gtf/gff: ○ gff/gff3 ○ gtf

4）树文件转 nwk 格式

能够进行转换的树文件有两类：一类是 dnd 格式，一类是 xml 格式，详细格式参考示例文件。操作同上，略。

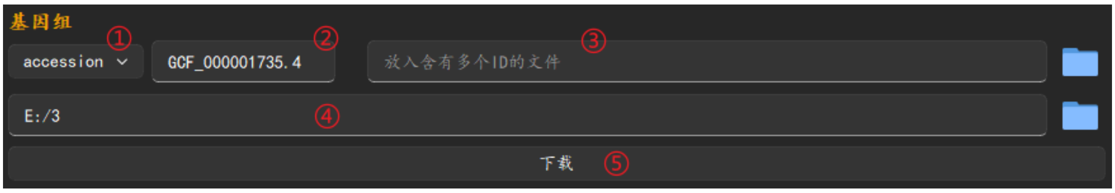
5）sam 格式转 paf 格式

Paf 格式的具体格式分布见 https://github.com/lh3/miniasm/blob/master/PAF.md。操作同上，略。

VI. 从 NCBI 下载文件

这里提供给同学们一种下载的方法。后来我测试了一下，速度不是那么快且依赖于网速。但革新的一点是可以通过将 ID 放入文件而实现批量下载。所以，同学们根据自己的需求灵活选择。

1）从 NCBI 下载基因组



基因组

accession

GCF_000001735.4

放入含有多个ID的文件

E:/3

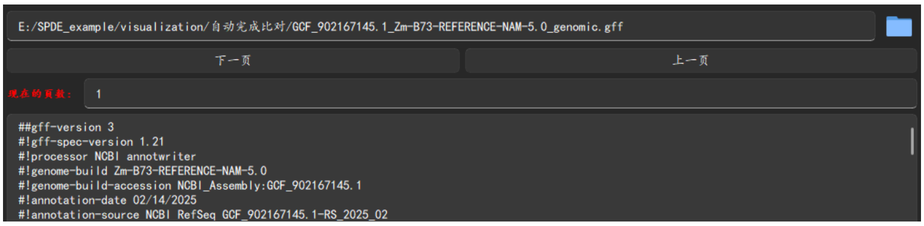
①下载的 ID 有两种类型:accession 和 taxon；②放入相应 ID；③可以考虑把不同 ID 放入一个文件（文本文件）中，一个 ID 一行，之后会根据文件中的 ID 进行自动下载；④保存的位置；按⑤执行功能

2）从 NCBI 下载基因

基本方法如上。

VII. 浏览大文件

在输入大文件后，可以通过点击上一页或下一页来进行浏览。基本设置如下：



E:/SPDE_example/visualization/自动完成比对/GCF_902167145.1_Zm-B73-REFERENCE-NAM-5.0_genomic.gff

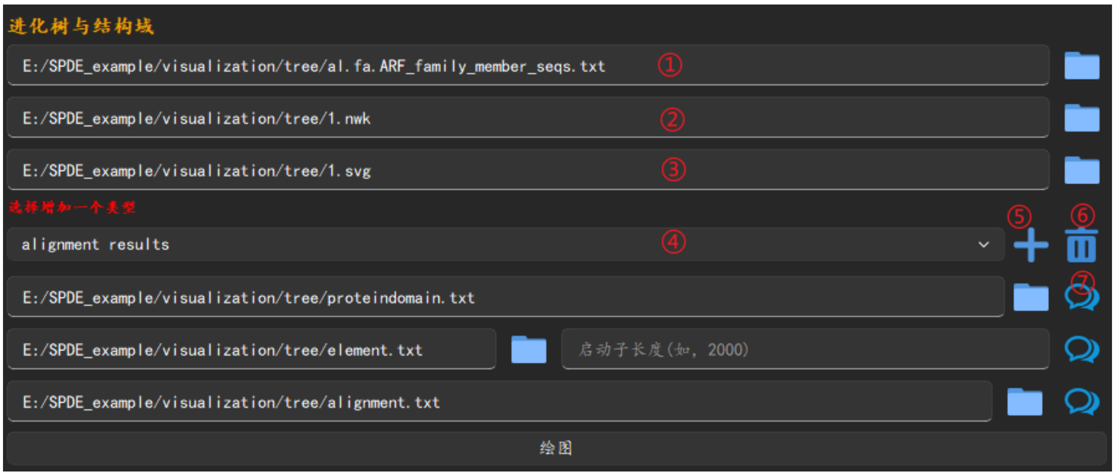
现在的页数： 1

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build Zm-B73-REFERENCE-NAM-5.0
#!genome-build-accession NCBI_Assembly:GCF_902167145.1
#!annotation-date 02/14/2025
#!annotation-source NCBI RefSeq GCF_902167145.1-RS_2025_02
```

7、　可视化模块

I. 进化树的可视化

该可视化用于将进化树与其他元素（包括蛋白结构域、启动子元件、比对结果、motif、单一元素等），同时展示在一张图中。基本设置如下：



进化树与结构域

E:/SPDE_example/visualization/tree/al.fa.ARF_family_member_seqs.txt ①

E:/SPDE_example/visualization/tree/1.nwk ②

E:/SPDE_example/visualization/tree/1.svg ③

选择增加一个类型

alignment results ④

E:/SPDE_example/visualization/tree/proteindomain.txt

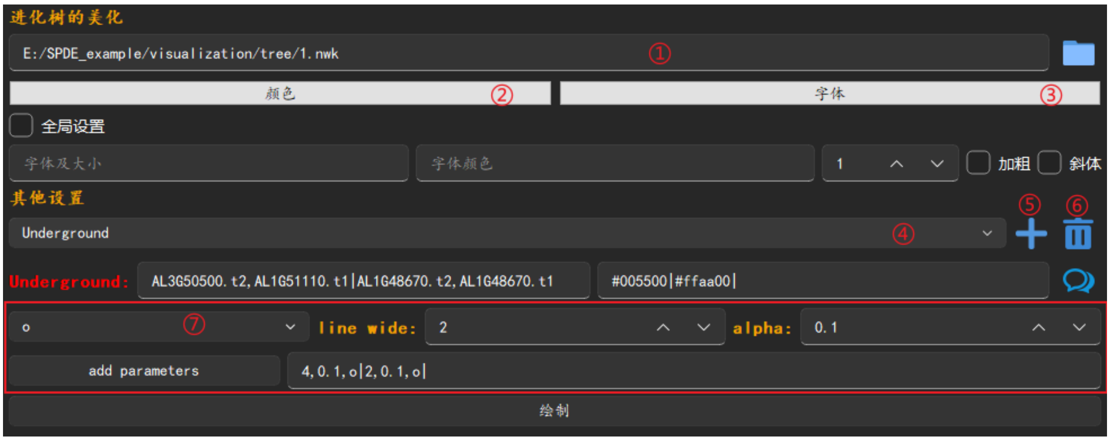
E:/SPDE_example/visualization/tree/element.txt　启动子长度(如，2000)

E:/SPDE_example/visualization/tree/alignment.txt

①基因家族的序列文件；②nwk 格式的进化树（可以参考我的这篇博客 https://www.jianshu.com/p/39f07b6f0435，使用 MEGA 软件获取）；③选择保存位置并命名；④是下拉框，可用于选择可视化的类型（如蛋白结构域），点击⑤展开相应对话框用来加入符合格式的文件；可点击⑦查看格式；点击⑥可以删除不需要的对话框（当然删除的只能是最后一个对话框）；设置完成后，点击绘图执行功能。当然如上图所示，可以将加入不同信息到进化树中。

II. 进化树的美化

输入的进化树依然是 nwk 格式，具体获取方法如上。可在进化树中添加的美化元素，包括背景、label、线（line）、marker 以及热图（heatmap）。注意，所有手动输入的内容均应在英文状态下输入。设置了较多参数，无法一一展示，只展示基本设置如下：



进化树的美化

E:/SPDE_example/visualization/tree/1.nwk ①

□ 全局设置

字体及大小　字体颜色　1　□ 加粗　□ 斜体

其他设置

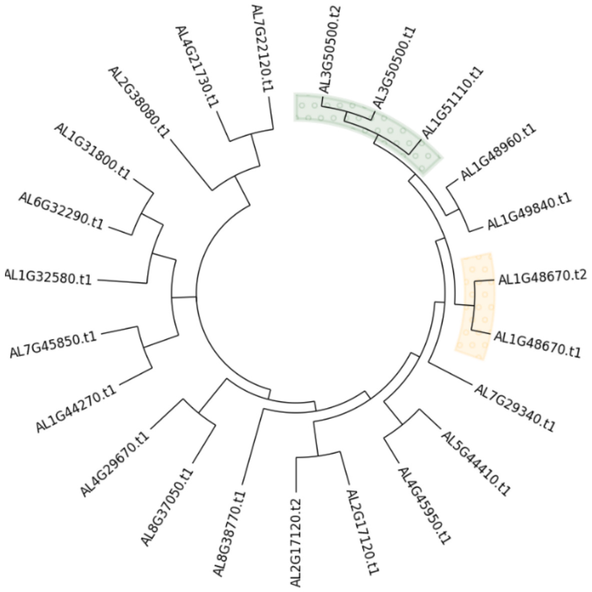
Underground ④

Underground:　AL3G50500.t2,AL1G51110.t1|AL1G48670.t2,AL1G48670.t1　#005500|#ffaa00|

o ⑦　line wide: 2　alpha: 0.1

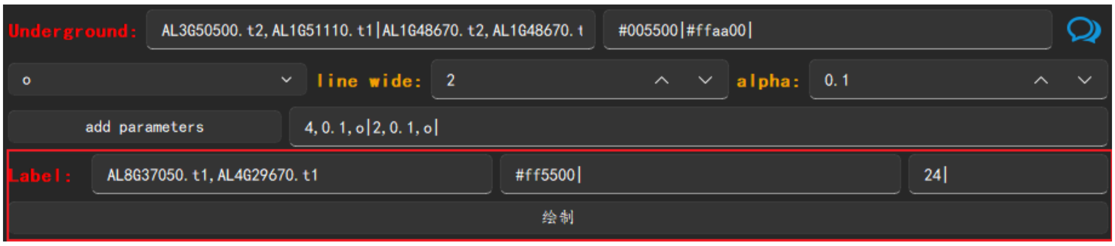
add parameters　4,0.1,o|2,0.1,o|

①输入 nwk 格式的文件；②可以通过此按钮选择颜色；③可以通过此按钮选择字体及大小；④这里主要是调出美化图层通过设置一些参数进行美化；选择类型后，点击⑤进行添加；⑥用来删除最后一个加添形式；当想要对背景做些符号标注的时候可以选择⑦所在的红框进行设置。如上图所示，它可以绘制出如下图所示的美化效果：



几个提示：当调出美化图层后，使用②进行颜色设置，所选择的颜色会自动添加。一个基本要求是：输入的 ID 有多少组（每一组利用 ‘|’ 符合分隔），就应该设置多少组颜色（当然如果有些组不想设置颜色，那就直接设置成黑色）。类似的要求见于符号添加处（即红色框处）。

对 label 的设置，如下：



Underground: AL3G50500.t2,AL1G51110.t1|AL1G48670.t2,AL1G48670.t #005500|#ffaa00|

o line wide: 2 alpha: 0.1

add parameters 4,0.1,o|2,0.1,o|

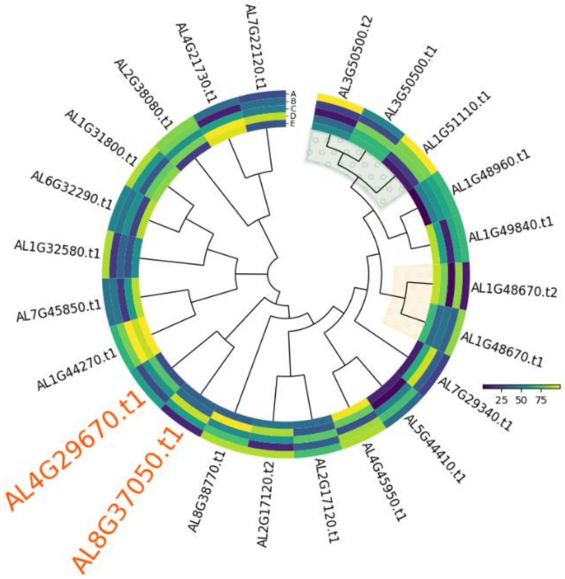
Label: AL8G37050.t1,AL4G29670.t1 #ff5500| 24|

绘制

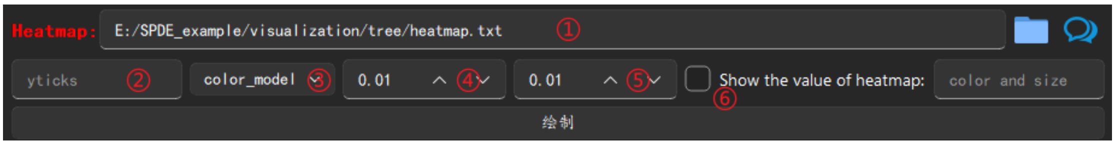
字体可以通过全局部分设置。不过，需要在“全局设置”处打勾。Label 处可以单独设置字体颜色和大小而这二者可以通过“颜色”，“字体”按钮自动添加。

其他一些美化包括 line 和 marker，没有太想要注意的地方。由于设置参数多，我就不给同学们一一展示了，在具体使用的时候，同学们可以结合示例文件充分理解各参数后，替换成自己的数据进行美化。

在进化树美化中还有一个关于 heatmap 添加的操作，可以做如下图：



关于数据的输入文件，请同学们仔细核对示例文件后操作。一个要求是数据库中有多少基因 ID 就该对应有多少列，即列数与 ID 数应该相同。基本设置：



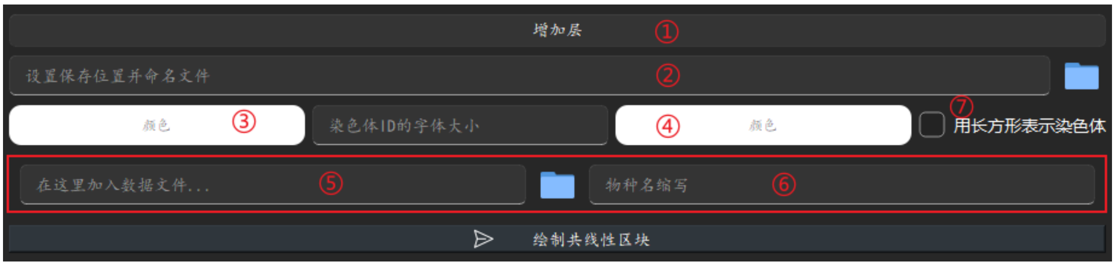
Heatmap: E:/SPDE_example/visualization/tree/heatmap.txt ①

yticks ② | color_model ③ | 0.01 ④ | 0.01 ⑤ | □ Show the value of heatmap: ⑥ | color and size

①输入数据；②数据 y 的标注（一般不需要设置）；③颜色模式；④设置 colorbar 在 x 轴的位置；⑤设置 colorbar 在 y 轴的位置；如果想要在热图上加上数字可以点选⑥，数字的颜色和大小可以在后面对话框设置。

III. 共线性绘图

该模块的输入内容可以由基因组模块得到。基本设置如下：



设置保存位置并命名文件 ②

颜色 ③ | 染色体ID的字体大小 | ④ 颜色 | ⑦ □ 用长方形表示染色体

在这里加入数据文件... ⑤ | 物种名缩写 ⑥

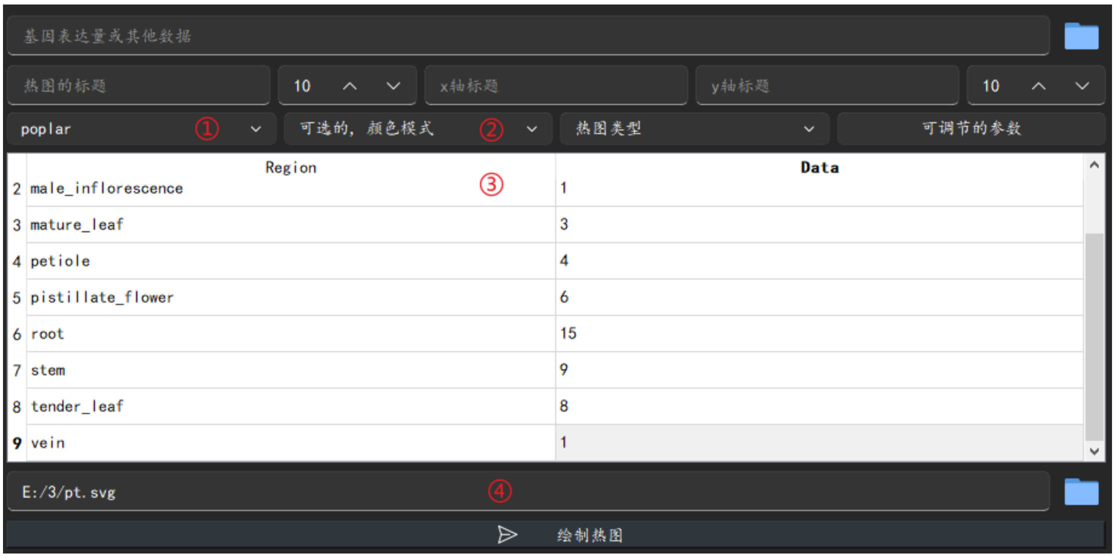
本模块支持对多个物种（大于两个）的共线性展示，点击①调出设置对话框（上图红色框）；②设置保存位置并命名文件； SDPE 会默认对每一条染色体配置不同的颜色，直到③处设置了颜色；④的颜色指的是共线性连线的颜色；点击⑤进行添加共线性分析文件；⑥这里需要设置物种名缩写。按照科研论文的要求，这里同学们应该以拉丁学名的缩写进行填写。比如，我现在想要展示拟南芥（缩写是 At）、水稻（Os）和杨树（Pt）间的共线性，则生成共线性文件时，应先用 At 与 Os 比对，然后再用 Os 与 Pt 比对。所以这里需要添加两层，第一层的物种名缩写应该是 At,Os；第二层的缩写应该是 Os,Pt。换言之，物种共线性文件以及绘图时的顺序应该是一致且连续的；当想要使用矩形表示染色体的时候可以勾选⑦。设置完

成后点击下面按钮执行功能

IV. 热图

1） 设置物种表达及地图数据可视化

这个功能主要用于展示在植物不同器官、亚细胞以及地图（包括中国地图和世界地图）上的数据分布趋势。以展示植物的各器官的数据分布为例：

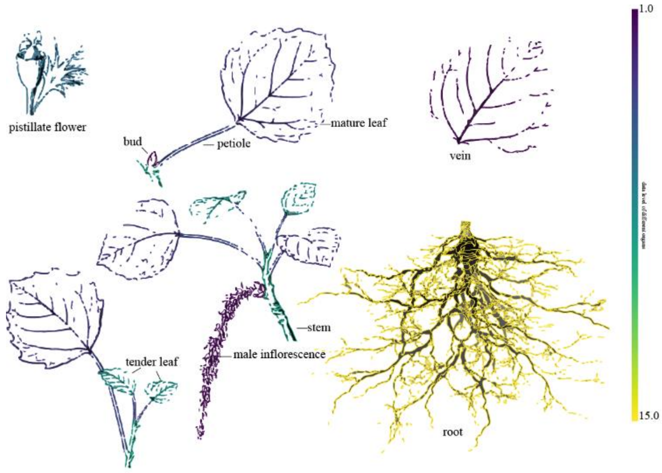


poplar

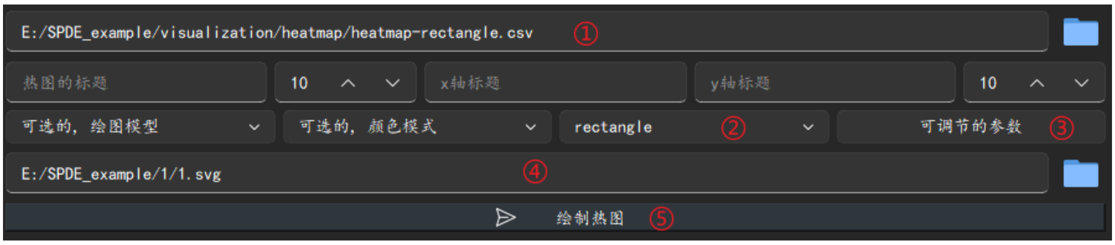
	Region	Data
2	male_inflorescence	1
3	mature_leaf	3
4	petiole	4
5	pistillate_flower	6
6	root	15
7	stem	9
8	tender_leaf	8
9	vein	1

E:/3/pt.svg

在①处选择你所需要的物种，当选定后，会自动出现一张表格，在这个表格③的 data 处添加数据（注意这里只是为了展示数据的分布趋势，所以不会有多个生物学处理这种形式。以表达量为例，同学们整理完数据后，在相应器官处放入一个平均值即可。如果没有数据，则空着或写 0）；②是可供选择的颜色模式；④设置保存位置并命名结果为文件；点击按键执行功能。如上设置，表明基因表达主要集中在杨树的根部，所以绘制以后会得到如下图所示的可视化结果。



2） 其他热图形式


E:/SPDE_example/visualization/heatmap/heatmap-rectangle.csv ①

热图的标题 | 10 | x轴标题 | y轴标题 | 10

可选的，绘图模型 | 可选的，颜色模式 | rectangle ② | 可调节的参数 ③

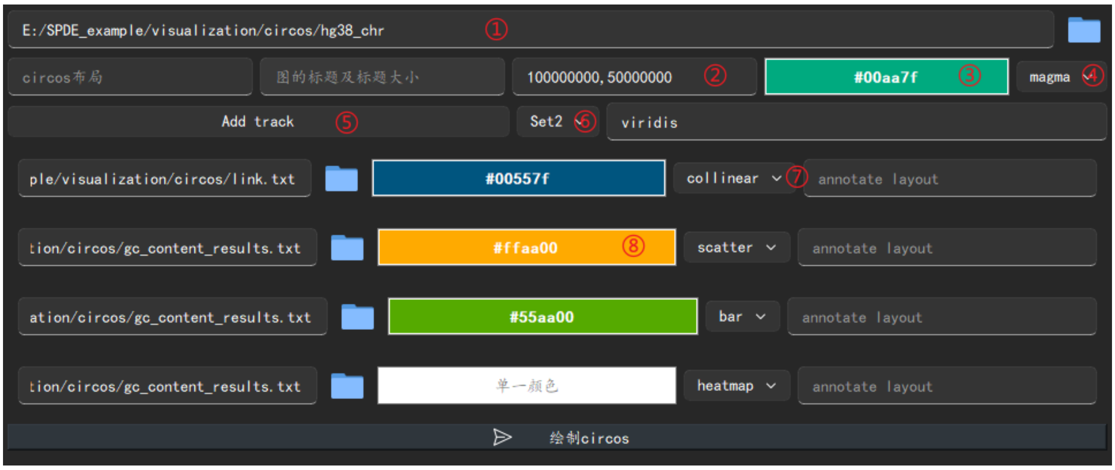
E:/SPDE_example/1/1.svg ④

绘制热图 ⑤

在①处输入（我针对每一个绘图类型给同学都做了一份示例文件，请一定按照示例文件的格式组织自身的数据）；其他参数按照提示设置就好，因此，略；在②选择绘图类型；想要调出更多调节参数，点③；在④设置保存位置并命名文件；点击⑤执行功能。

V. circos 图

基本设置如下：


E:/SPDE_example/visualization/circos/hg38_chr ①

circos布局 | 图的标题及标题大小 | 100000000, 50000000 ② | #00aa7f ③ | magma ④

Add track ⑤ | Set2 ⑥ | viridis

ple/visualization/circos/link.txt | #00557f | collinear ⑦ | annotate layout

tion/circos/gc_content_results.txt | #ffaa00 ⑧ | scatter | annotate layout

ation/circos/gc_content_results.txt | #55aa00 | bar | annotate layout

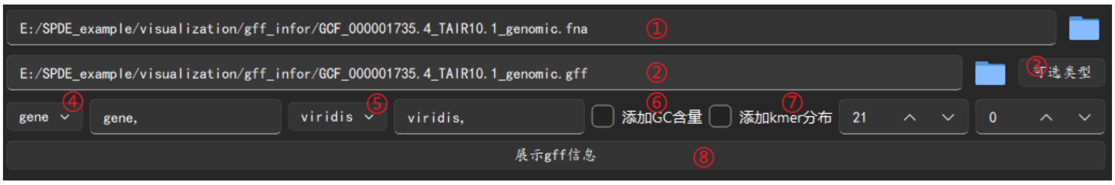
tion/circos/gc_content_results.txt | 单一颜色 | heatmap | annotate layout

绘制circos

在①放入基因组序列文件，SPDE 会根据序列文件生成 bed 文件，bed 文件是用来绘图的基础文件，所以如果没有生成终图，需要保留 bed 文件并始终保持其与基因组序列处在同一文件夹下。换言之，第一次生成 bed 文件后，以后输入①处的还是基因组序列文件，但由于 bed 文件已经存在，程序会自动识别 bed 文件而不会再次生成 bed 文件；②处的设置用于生成染色体的长度标尺的。默认值是 1Mb，单位长度 0.5Mb。如果生成的图中标尺过于密集，则可考虑增大这两个值来稀疏标尺；③是为染色体设置单一颜色；④是为染色体设置不同的颜色；circos 都是一层层的，通过点击⑤增加层；输入文件这里一共有两种类型的数据：一类用于展示共线性（即染色体间的连线），另一个是数据文件（比如 GC 含量等）。同学们在生成两类文件的时候，需要根据染色体按照从前到后排序，然后放入该模块进行可视化；点击⑦选择该层的可视化类型；当是非热图的时候可以通过点击⑧设置颜色；如果是热图，可以在⑥处选择颜色模型进行热图绘制。关于两类数据文件的格式，请参考示例数据。

VI. gff 信息展示

GFF 文件本身含有很多信息。SPDEv3.0 设置了相应功能来展示 GFF 中的信息元素。基本设置如下：



E:/SPDE_example/visualization/gff_infor/GCF_000001735.4_TAIR10.1_genomic.fna ①

E:/SPDE_example/visualization/gff_infor/GCF_000001735.4_TAIR10.1_genomic.gff ②

gene ④ | gene, | viridis ⑤ | viridis, | □ 添加GC含量 ⑥ | □ 添加kmer分布 ⑦ | 21 | 0

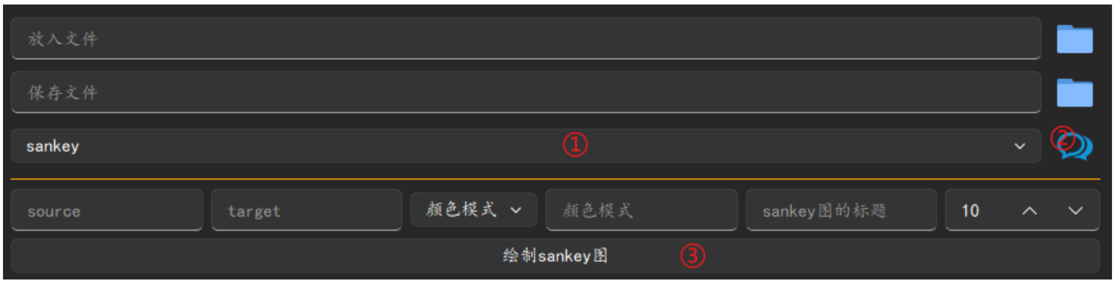
在①放入基因组序列文件；在②放入 gff 文件；之后点击③分析 gff 中含有哪些元素；分析完成后，这些信息类型会加载到④；浏览④以决定可视化哪些元素；⑤可以用来设置跟数据相关的元素并以热图形式展示 GC 含量（⑥）或者 kmer 的分布密度（⑦）。之后点击⑧运行功能。

VII. 统计绘图

该模块一共设置了柱状图、箱线图、双箱线图、地理地图、中国地理地图、主成分分析（二维）、主成分分析（三维）、散点图（双变量）、散点图（多变量）、堆叠柱状图、小提琴图（多变量）、小提琴图（单变量）。具体的使用方法与育种学计算模块相同，这里不做详细介绍了。

VIII. 其他图

除上述常用的绘图功能外，SPDEv3.0 还设置了其他的绘图功能，像桑基图、基序标志（或基序徽标）、网络图、保守基序、启动子基序、词云。基本设置：



sankey ①

放入文件并设置保存位置后，可以点击①选择绘图类型。当不明确输入文件的格式时，可点击②查看基本格式（当然，也会提供示例文件给大家）；其他的设置均提供了悬浮的提示，当鼠标停在某个元件时，会显示提示信息。点击③执行功能。

English version

I. Introduction to SPDE v3.0

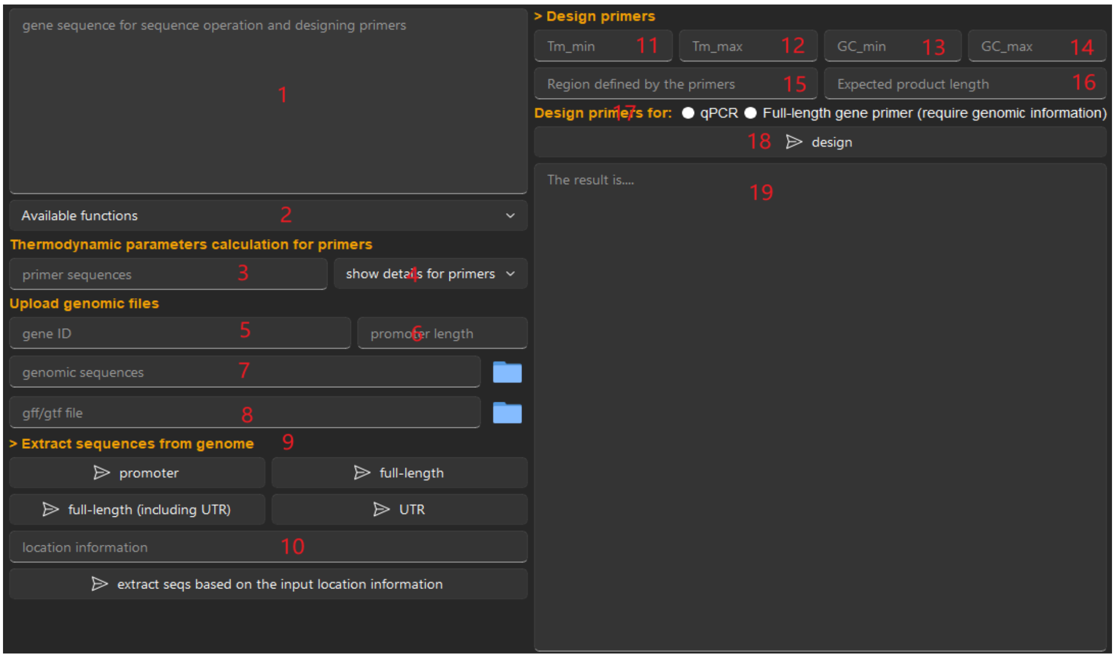
SPDE v3.0 is used in processes such as sequence extraction, calculation, bioinformatics analysis, and result visualization in disciplines like molecular biology, genomics, and breeding. Its most prominent feature lies in the further integration of some analysis processes that do not require user decisions, which further improves the analysis efficiency and optimizes the entire operation process. To enable you to use SPDE more effectively, we have provided sufficient prompt information on the software interface, most of which appears as floating tips. We believe that SPDE will become an important tool to promote the development of plant research. The following is a detailed introduction to the software's usage.

Please note: Chinese characters are not allowed in the software installation path, and it is advisable to avoid using Chinese in other paths as much as possible. Also, try to avoid paths that contain spaces.

II. Module Functions

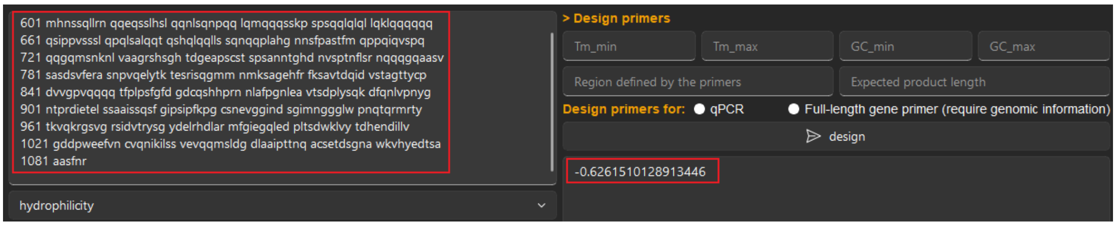
1) Gene module

This module is mainly designed to meet the requirements for sequence information query, extraction, primer design, etc. in molecular biology experiments. The input methods and function distribution are as follows:



When you need to query sequence information, enter the gene sequence (no fasta format is required, just the sequence itself) in 1. You can directly click the function in 2 to query some basic information about the entered sequence. These basic information include: Hydrophilicity, length, pI, reverse, translation, extinction_coefficient, GC_content, molecular_weight, reverse complement, kaks, clear extra characters.

The usage is very simple: after opening the dropdown box, you just need to use the mouse to select a certain function, and the analysis results will automatically appear in position 19.



```
601 mhnssqllrn qqeqsslhsl qqnlsqnpqq lqmqqqsskp spsqqlqlql lqklqqqqqq
661 qsippvsssl qpqlsalqqt qshqlqqlls sqnqqplahg nnsfpastfm qppqiqvspq
721 qqgqmsnknl vaagrshsgh tdgeapscst spsanntghd nvsptnflsr nqqqgqaasv
781 sasdsvfera snpvqelytk tesrisqgmm nmksagehfr fksavtdqid vstagttycp
841 dvvgpvqqqq tfplpsfgfd gdcqshhprn nlafpgnlea vtsdplysqk dfqnlvpnyg
901 ntprdietel ssaaissqsf gipsipfkpg csnevggind sgimngggiw pnqtqrmrty
961 tkvqkrgsvg rsidvtrysg ydelrhdlar mfgiegqled pltsdwklvy tdhendillv
1021 gddpweefvn cvqnikilss vevqqmsldg dlaaipttnq acsetdsgna wkvhyedtsa
1081 aasfnr
```

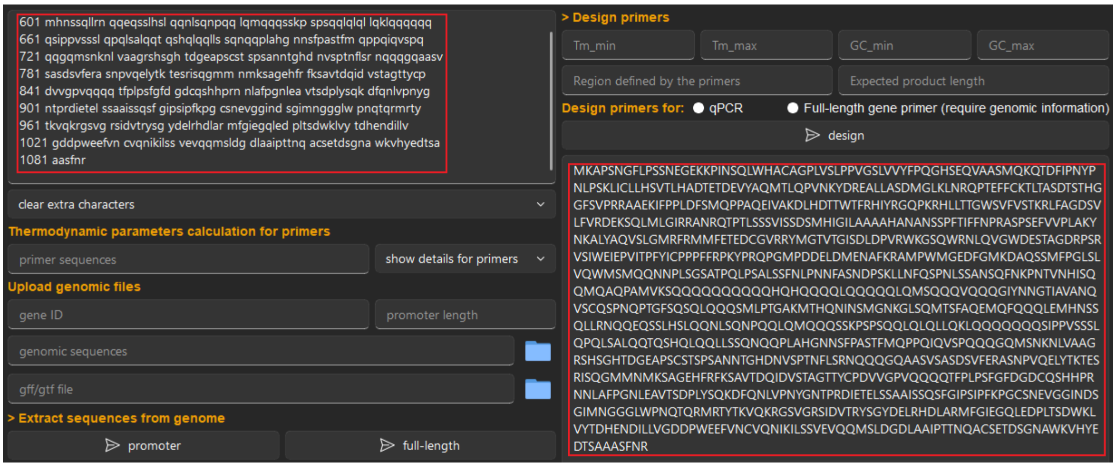
```
-0.6261510128913446
```

hydrophilicity
> Design primers
Tm_min
Tm_max
GC_min
GC_max
Region defined by the primers
Expected product length
Design primers for: ● qPCR ● Full-length gene primer (require genomic information)
design

Two functions require additional explanation:

For the kaks function, users need to input the sequences of two genes in fasta format for calculation.

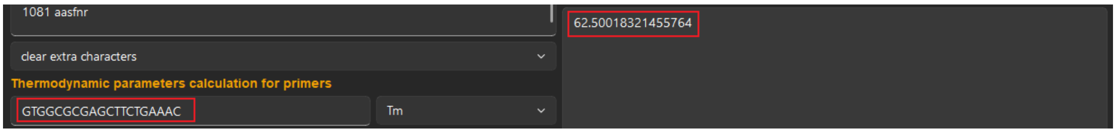
Remove extra characters: On websites such as NCBI and TAIR, gene sequences often contain a mix of numbers, spaces, and the actual sequence. It's quite troublesome to deal with such extra information like numbers and spaces. In this case, you can use the "Remove extra characters" function to eliminate this extra information with just one click, as follows:



```
601 mhnssqllrn qqeqsslhsl qqnlsqnpqq lqmqqqsskp spsqqlqlql lqklqqqqqq
661 qsippvsssl qpqlsalqqt qshqlqqlls sqnqqplahg nnsfpastfm qppqiqvspq
721 qqgqmsnknl vaagrshsgh tdgeapscst spsanntghd nvsptnflsr nqqqgqaasv
781 sasdsvfera snpvqelytk tesrisqgmm nmksagehfr fksavtdqid vstagttycp
841 dvvgpvqqqq tfplpsfgfd gdcqshhprn nlafpgnlea vtsdplysqk dfqnlvpnyg
901 ntprdietel ssaaissqsf gipsipfkpg csnevggind sgimngggiw pnqtqrmrty
961 tkvqkrgsvg rsidvtrysg ydelrhdlar mfgiegqled pltsdwklvy tdhendillv
1021 gddpweefvn cvqnikilss vevqqmsldg dlaaipttnq acsetdsgna wkvhyedtsa
1081 aasfnr
```

```
MKAPSNGFLPSSNEGEKKPINSQLWHACAGPLVSLPPVGSLVVYFPQGHSEQVAASMQKQTDFIPNYP
NLPSKLICLLHSVTLHADTETDEVYAQMTLQPVNKYDREALLASDMGLKLNRQPTEFFCKTLTASDTSTHG
GFSVPRRAAEKIFPPLDFSMQPPAQEIVAKDLHDTTWTFRHIYRGQPKRHLLTTGWSVFVSTKRLFAGDSV
LFVRDEKSQLMLGIRRANRQTPTLSSSVISSDSMHIGILAAAAHANANSSPFTIFFNPRASPSEFVVPLAKY
NKALYAQVSLGMRFRMMFETEDCGVRRYMGTVTGISDLDPVRWKGSQWRNLQVGWDESTAGDRPSR
VSIWEIEPVITPFYICPPPFFRPKYPRQPGMPDDELDMENAFKRAMPWMGEDFGMKDAQSSMFPGLSL
VQWMSMQQNNPLSGSATPQLPSALSSFNLPNNFASNDPSKLLNFQSPNLSSANSQFNKPNTVNHISQ
QMQAQPAMVKSQQQQQQQQQHQHQQQQLQQQQQLQMSQQQVQQQGIYNNGTIAVANQ
VSCQSPNQPTGFSQSQLQQQSMLPTGAKMTHQNINSMGNKGLSQMTSFAQEMQFQQQLEMHNSS
QLLRNQQEQSSLHSLQQNLSQNPQQLQMQQQSSKPSPSQQLQLQLLQKLQQQQQQQSIPPVSSSL
QPQLSALQQTQSHQLQQLLSSQNQQPLAHGNNSFPASTFMQPPQIQVSPQQQGQMSNKNLVAAG
RSHSGHTDGEAPSCSTSPSANNTGHDNVSPTNFLSRNQQQGQAASVSASDSVFERASNPVQELYTKTES
RISQGMMNMKSAGEHFRFKSAVTDQIDVSTAGTTYCPDVVGPVQQQQTFPLPSFGFDGDCQSHHPR
NNLAFPGNLEAVTSDPLYSQKDFQNLVPNYGNTPRDIETELSSAAISSQSFGIPSIPFKPGCSNEVGGINDS
GIMNGGGLWPNQTQRMRTYTKVQKRGSVGRSIDVTRYSGYDELRHDLARMFGIEGQLEDPLTSDWKL
VYTDHENDILLVGDDPWEEFVNCVQNIKILSSVEVQQMSLDGDLAAIPTTNQACSETDSGNAWKVHYE
DTSAAASFNR
```

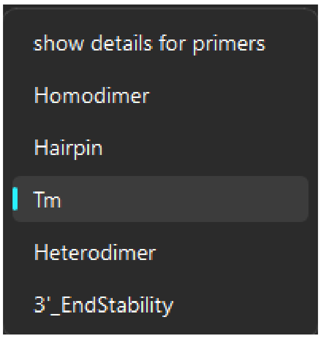
In the thermodynamic calculation of primers, the usage is the same as before. When you want to query certain primers, simply copy the sequences in and then click the corresponding function to output the corresponding calculation results.



```
1081 aasfnr
```

```
GTGGCGCGAGCTTCTGAAAC
```

```
62.50018321455764
```



The thermodynamic parameters available for calculation include:

Homodimer

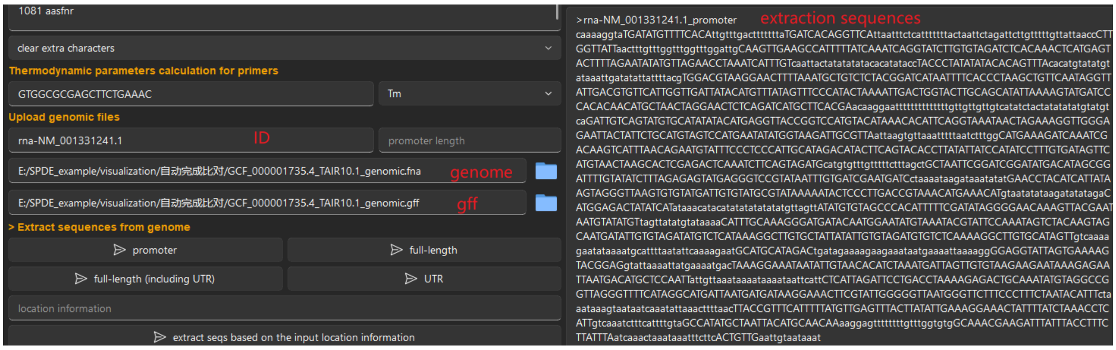
Hairpin

Tm

Heterodimer

3'_EndStability

The genome sequence extraction function can extract the corresponding sequences from the genome based on the gene IDs provided by users. The input files are the genome sequence file and the structure annotation files (in GFF and GTF formats). The details are as follows:



1081 aasfnr

clear extra characters

Thermodynamic parameters calculation for primers

GTGGCGCGAGCTTCTGAAAC | Tm

Upload genomic files

rna-NM_001331241.1 | promoter length

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

> Extract sequences from genome

promoter | full-length

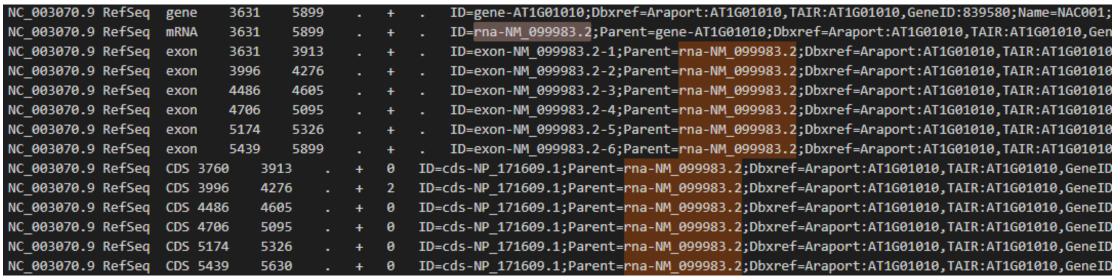
full-length (including UTR) | UTR

location information

extract seqs based on the input location information

```
>rna-NM_001331241.1_promoter
caaaaggtaTGATATGTTTTCACAttgtttgactttttttaTGATCACAGGTTCAttaatttctcattttttttactaattctagattcttgtttttgttattaaccCTT
GGTTATTaactttgtttggtttggtttggattgCAAGTTGAAGCCATTTTTATCAAATCAGGTATCTTGTGTAGATCTCACAAACTCATGAGT
ACTTTTAGAATATATGTTAGAACCTAAATCATTTGTcaattactatatatatacacatataccTACCCTATATATACACAGTTTAcacatgtatatgt
ataaattgatatattattttacgTGGACGTAAGGAACTTTTAAATGCTGTCTCTACGGATCATAATTTTCACCCTAAGCTGTTCAATAGGTT
ATTGACGTGTTCATTGGTTGATTATACATGTTTATAGTTTCCCATACTAAAATTGACTGGTACTTGCAGCATATTAAAAGTATGATCC
CACACAACATGCTAACTAGGAACTCTCAGATCATGCTTCACGAacaaggaattttttttttttttgttgttgttgtcatatctactatatatatgtatgt
caGATTGTCAGTATGTGCATATATACATGAGGTTACCGGTCCATGTACATAAACACATTCAGGTAAATAACTAGAAAGGTTGGGA
GAATTACTATTCTGCATGTAGTCCATGAATATATGGTAAGATTGCGTTAattaagtgttaaatttttaatctttggCATGAAAGATCAAATCG
ACAAGTCATTTAACAGAATGTATTTCCCTCCCATTGCATAGACATACTTCAGTACACCTTATATTATCCATATCCTTTGTGATAGTTC
ATGTAACTAAGCACTCGAGACTCAAATCTTCAGTAGATGcatgtgtttgtttttctttagctGCTAATTCGGATCGGATATGACATAGCGG
ATTTTGTATATCTTTAGAGAGTATGAGGGTCCGTATAATTTGTGATCGAATGATCctaaaataagataaatatatGAACCTACATCATTATA
AGTAGGGTTAAGTGTGTATGATTGTGTATGCGTATAAAAATACTCCCTTGACCGTAAACATGAAACATgtaatatataagatatatagaC
ATGGAGACTATATCATataaacatacatatatatatatatgttagttATATGTGTAGCCCACATTTTTCGATATAGGGGAACAAAGTTACGAAT
AATGTATATGTtagttatatgtataaaaCATTTGCAAAGGGATGATACAATGGAATATGTAAATACGTATTCCAAATAGTCTACAAGTAG
CAATGATATTGTGTAGATATGTCTCATAAAGGCTTGTGCTATTATATTGTGTAGATGTGTCTCAAAAGGCTTGTGCATAGTTgtcaaaa
gaatataaaatgcattttaatattcaaaagaatGCATGCATAGACtgatagaaaagaagaaataatgaaaattaaaaggGGAGGTATTAGTGAAAAG
TACGGAGgtattaaaattatgaaaatgacTAAAGGAAATAATATTGTAACACATCTAAATGATTAGTTGTGTAAGAAGAATAAAGAGAA
TTAATGACATGCTCCAATTattgttaaataaaataaaataattcattCTCATTAGATTCCTGACCTAAAAGAGACTGCAAATATGTAGGCCG
GTTAGGGTTTTCATAGGCATGATTAATGATGATAAGGAAACTTCGTATTGGGGGTTAATGGGTTCTTTCCCTTTCTAATACATTTcta
aataaagtaataatcaaatattaaacttttaacTTACCGTTTCATTTTTATGTTGAGTTTACTTATATTGAAAGGAAACTATTTTATCTAAACCTC
ATTgtcaaatctttcattttgtaGCCATATGCTAATTACATGCAACAAaaggagtttttttttgtttggtgtgGCAAACGAAGATTTATTTACCTTTC
TTATTTAatcaaactaaataaatttcttcACTGTTGaattgtaataaat
```

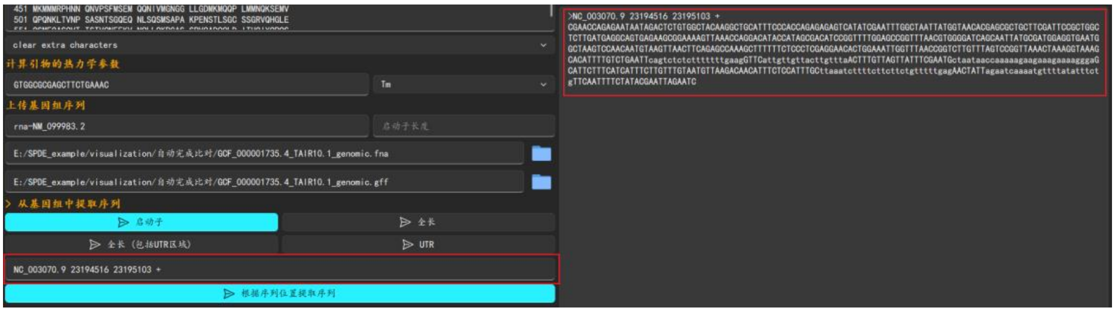
In total, it can extract sequences including promoters, full - length genes (from ATG to the stop codon), full - length genes (including UTR regions), and UTR regions. It should be noted that the gene IDs should be the IDs corresponding to mRNAs in the GFF/GTF files. Here are an example of the GFF file format and the specific location of the IDs:



```
NC_003070.9 RefSeq  gene    3631    5899    .   +   .   ID=gene-AT1G01010;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID:839580;Name=NAC001;
NC_003070.9 RefSeq  mRNA    3631    5899    .   +   .   ID=rna-NM_099983.2;Parent=gene-AT1G01010;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,Gen
NC_003070.9 RefSeq  exon    3631    3913    .   +   .   ID=exon-NM_099983.2-1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  exon    3996    4276    .   +   .   ID=exon-NM_099983.2-2;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  exon    4486    4605    .   +   .   ID=exon-NM_099983.2-3;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  exon    4706    5095    .   +   .   ID=exon-NM_099983.2-4;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  exon    5174    5326    .   +   .   ID=exon-NM_099983.2-5;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  exon    5439    5899    .   +   .   ID=exon-NM_099983.2-6;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010
NC_003070.9 RefSeq  CDS 3760    3913    .   +   0   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
NC_003070.9 RefSeq  CDS 3996    4276    .   +   2   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
NC_003070.9 RefSeq  CDS 4486    4605    .   +   0   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
NC_003070.9 RefSeq  CDS 4706    5095    .   +   0   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
NC_003070.9 RefSeq  CDS 5174    5326    .   +   0   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
NC_003070.9 RefSeq  CDS 5439    5630    .   +   0   ID=cds-NP_171609.1;Parent=rna-NM_099983.2;Dbxref=Araport:AT1G01010,TAIR:AT1G01010,GeneID
```

Please note: the format of GFF/GTF files is crucial. Please carefully check the format according to the above sample files before use.

Of course, if you don't have GFF/GTF files, you can also directly enter the location information to extract the files, as follows:



```
451 MKMMMRPHNN QNVPSFMSEM QQNIVMGNGG LLGDMKMQQP LMMNQKSEMV
501 QPQNKLTVNP SASNTSGQEQ NLSQSMSAPA KPENSTLSGC SSGRVQHGLE
```

clear extra characters

计算引物的热力学参数

GTGGCGCGAGCTTCTGAAAC | Tm

上传基因组序列

rna-NM_099983.2 | 启动子长度

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

> 从基因组中提取序列

启动子 | 全长

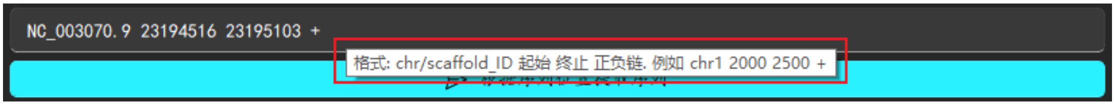
全长（包括UTR区域） | UTR

NC_003070.9 23194516 23195103 +

根据序列位置提取序列

```
>NC_003070.9 23194516 23195103 +
CGAACCAGAGAATAATAGACTCTGTGGCTACAAGGCTGCATTTCCCACCAGAGAGAGTCATATCGAATTTGGCTAATTATGGTAACACGAGCGCTGCTTCGATTCCGCTGGC
TCTTGATGAGGCAGTGAGAAGCGGAAAAGTTAAACCAGGACATACCATAGCGACATCCGGTTTTGGAGCCGGTTTAACGTGGGGATCAGCAATTATGCGATGGAGGTGAATG
GCTAAGTCCAACAATGTAAGTTAACTTCAGAGCCAAAGCTTTTTTCTCCCTCGAGGAACACTGGAAATTGGTTTAACCGGTCTTGTTTAGTCCGGTTAAACTAAAGGTAAAG
CACATTTTGTCTGAATTcagtctctcttttttgaagGTTCattgttgttacttgtttaACTTTGTTAGTTATTTCGAATGctaataaccaaaaagaagaaagaaaagggaG
CATTCTTTCATCATTTCTTGTTTGTAATGTTAAGACAACATTTCTCCATTTGCttaaatcttttcttcttctgtttttgagAACTATTagaatcaaaatgttttatatttct
gTTCAATTTTCTATACGAATTAGAATC
```

To enable you to use SPDE more smoothly, I've left sufficient prompt information at the corresponding positions of each function. Most of this information is presented in the form of floating prompts. You can refer to these prompts to run the functions. Here are the details:



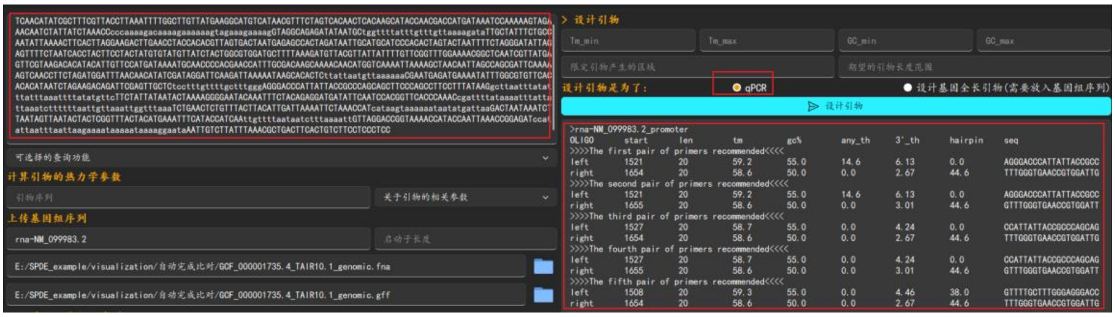
NC_003070.9 23194516 23195103 +

格式: chr/scaffold_ID 起始 终止 正负链. 例如 chr1 2000 2500 +

The sequences extracted by SPDE are all provided in the positive - strand format. Reverse complementation will be automatically performed on the negative - strand sequences.

Designing quantitative primers and full - length gene primers are two common types of primer design tasks in molecular experiments.

Regarding the design of quantitative primers, we have set default values based on empirical values (including Tm value: 50 ≤ Tm ≤ 65, 40 ≤ GC ≤ 60). That is, when the default values meet the requirements, there is no need to enter additional information. Just leave the corresponding dialog boxes empty. The details are as follows:



```
TCAACATATCGCTTTCGTTACCTTAAATTTTGGCTTGTTATGAAGGCATGTCATAACGTTTCTAGTCACAACTCACAAGCATACCAACGACCATGATAAATCCAAAAAGTAGA
AACAATCTATTATCTAAACCcccaaaagacaaaagaaaaaagtagaaagaaaagGTAGGCAGAGATATAATGCtggttttatttgtttgttaaaagataTTGCTATTTCTGCC
AATATTAAAACTTCACTTAGGAAGACTTGAACCTACCACACGTTAGTGACTAATGAGAGCCACTAGATAATTGCATGCATCCCACACTAGTACTAATTTTCTAGGGATATTAG
AGTTTTCTAATCACCTACTTCCTACTATGTGTATGTTATCTACTGGCGTGGATGCTTTTAAAGATGTTACGTTATTATTTTGTTCGGTTTGGAAAACGGCTCAATCGTTATGA
GTTCGTAAGACACATACATTGTTCCATGATAAAATGCAACCCCACGAACCATTTGCGACAAGCAAAACAACATGGTCAAAATTAAAAGCTAACAATTAGCCAGCGATTCAAAA
AGTCAACCTTCTAGATGGATTTAACAACATATCGATAGGATTCAAGATTAAAAATAAGCACACTCttattaatgttaaaaaaCGAATGAGATGAAAATATTTGGCGTGTTCAC
ACACATAATCTAGAAGACAGATTCGAGTTGCTCtcctttgttttgctttgggAGGGACCCATTATTACCGCCCAGCAGCTTCCCAGCCTTCCTTTATAAGgcttaatttatat
ttatttaaattttatatgttcTTCTATTATAATACTAAAAGGGGAATACAAATTTCTACAGAGGATGATATTCAATCCACGGTTCACCCAAACcgatttataaaatttatta
ttaaatctttttaattgttaaattggtttaaaTCTGAACTCTGTTTACTTACATTGATTAAAATTCTAAACCATcataagtaaaaaataatatgattaaGACTAATAAATCT
TAATAGTTAATACTACTCGGTTTACTACATGAAATTTCATACCATCAAttgttttaataatctttaaaattGTTAGGACCGGTAAAACCATACCAATTAAACCGGAGATccat
attaatttaattaagaaaataaaaataaaaggaataAATTGTCTTATTTAAACGCTGACTTCACTGTCTTCCTCCCTCC
```

rna-NM_099983.2

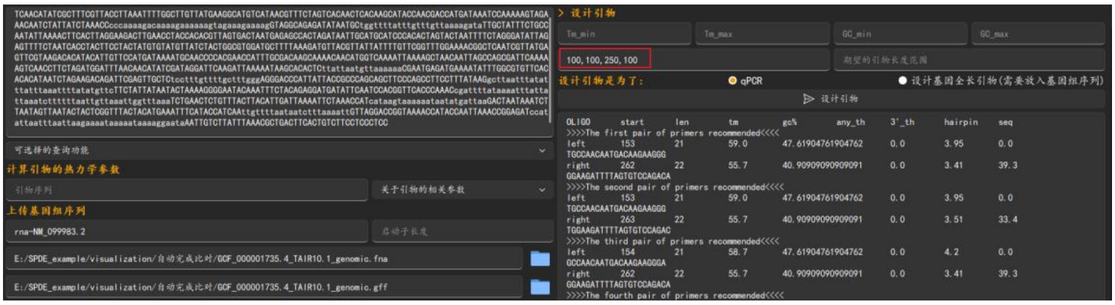
E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

设计引物是为了: ● qPCR ○ 设计基因全长引物(需要放入基因组序列)

```
>rna-NM_099983.2_promoter
OLIGO     start    len     tm      gc%     any_th   3'_th   hairpin   seq
>>>>The first pair of primers recommended<<<<
left      1521     20      59.2    55.0    14.6     6.13    0.0       AGGGACCCATTATTACCGCC
right     1654     20      58.6    50.0    0.0      2.67    44.6      TTTGGGTGAACCGTGGATTG
>>>>The second pair of primers recommended<<<<
left      1521     20      59.2    55.0    14.6     6.13    0.0       AGGGACCCATTATTACCGCC
right     1655     20      58.6    50.0    0.0      3.01    44.6      GTTTGGGTGAACCGTGGATT
>>>>The third pair of primers recommended<<<<
left      1527     20      58.7    55.0    0.0      4.24    0.0       CCATTATTACCGCCCAGCAG
right     1654     20      58.6    50.0    0.0      2.67    44.6      TTTGGGTGAACCGTGGATTG
>>>>The fourth pair of primers recommended<<<<
left      1527     20      58.7    55.0    0.0      4.24    0.0       CCATTATTACCGCCCAGCAG
right     1655     20      58.6    50.0    0.0      3.01    44.6      GTTTGGGTGAACCGTGGATT
>>>>The fifth pair of primers recommended<<<<
left      1508     20      59.3    55.0    0.0      4.46    38.0      GTTTTGCTTTGGGAGGGACC
right     1654     20      58.6    50.0    0.0      2.67    44.6      TTTGGGTGAACCGTGGATTG
```

Another important setting is that you can specify the regions where primers are generated, as follows:



```
TCAACATATCGCTTTCGTTACCTTAAATTTTGGCTTGTTATGAAGGCATGTCATAACGTTTCTAGTCACAACTCACAAGCATACCAACGACCATGATAAATCCAAAAAGTAGA
AACAATCTATTATCTAAACCcccaaaagacaaaagaaaaaagtagaaagaaaagGTAGGCAGAGATATAATGCtggttttatttgtttgttaaaagataTTGCTATTTCTGCC
AATATTAAAACTTCACTTAGGAAGACTTGAACCTACCACACGTTAGTGACTAATGAGAGCCACTAGATAATTGCATGCATCCCACACTAGTACTAATTTTCTAGGGATATTAG
AGTTTTCTAATCACCTACTTCCTACTATGTGTATGTTATCTACTGGCGTGGATGCTTTTAAAGATGTTACGTTATTATTTTGTTCGGTTTGGAAAACGGCTCAATCGTTATGA
GTTCGTAAGACACATACATTGTTCCATGATAAAATGCAACCCCACGAACCATTTGCGACAAGCAAAACAACATGGTCAAAATTAAAAGCTAACAATTAGCCAGCGATTCAAAA
AGTCAACCTTCTAGATGGATTTAACAACATATCGATAGGATTCAAGATTAAAAATAAGCACACTCttattaatgttaaaaaaCGAATGAGATGAAAATATTTGGCGTGTTCAC
ACACATAATCTAGAAGACAGATTCGAGTTGCTCtcctttgttttgctttgggAGGGACCCATTATTACCGCCCAGCAGCTTCCCAGCCTTCCTTTATAAGgcttaatttatat
ttatttaaattttatatgttcTTCTATTATAATACTAAAAGGGGAATACAAATTTCTACAGAGGATGATATTCAATCCACGGTTCACCCAAACcgatttataaaatttatta
ttaaatctttttaattgttaaattggtttaaaTCTGAACTCTGTTTACTTACATTGATTAAAATTCTAAACCATcataagtaaaaaataatatgattaaGACTAATAAATCT
TAATAGTTAATACTACTCGGTTTACTACATGAAATTTCATACCATCAAttgttttaataatctttaaaattGTTAGGACCGGTAAAACCATACCAATTAAACCGGAGATccat
attaatttaattaagaaaataaaaataaaaggaataAATTGTCTTATTTAAACGCTGACTTCACTGTCTTCCTCCCTCC
```

rna-NM_099983.2

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

100,100,250,100

设计引物是为了: ● qPCR ○ 设计基因全长引物(需要放入基因组序列)

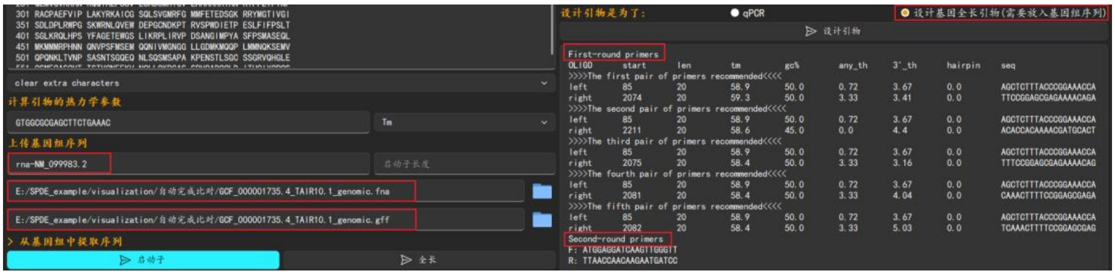
```
OLIGO     start    len     tm      gc%                any_th   3'_th   hairpin   seq
>>>>The first pair of primers recommended<<<<
left      153      21      59.0    47.61904761904762  0.0      3.95    0.0
TGCCAACAATGACAAGAAGGG
right     262      22      55.7    40.90909090909091  0.0      3.41    39.3
GGAAGATTTTAGTGTCCAGACA
>>>>The second pair of primers recommended<<<<
left      153      21      59.0    47.61904761904762  0.0      3.95    0.0
TGCCAACAATGACAAGAAGGG
right     263      22      55.7    40.90909090909091  0.0      3.51    33.4
TGGAAGATTTTAGTGTCCAGAC
>>>>The third pair of primers recommended<<<<
left      154      21      58.7    47.61904761904762  0.0      4.2     0.0
GCCAACAATGACAAGAAGGGA
right     262      22      55.7    40.90909090909091  0.0      3.41    39.3
GGAAGATTTTAGTGTCCAGACA
>>>>The fourth pair of primers recommended<<<<
```

Rules: The first two numbers represent the position of the left-end primer. The first number is the starting position of the primer, and the second number is an expected range. For example, if I want the left-end primer to be generated within the range of 100 to 200, the setting should be 100, 100. If no setting is made, the comma cannot be omitted. For example, when I don't want to restrict the generation area of the left-end primer and the right-end primer needs to be generated within the range of 300 to 400, the setting should be,, 300, 100. The same setting also applies to the right-end primer.

All R-end primers have been reverse-complemented, and no additional operation is required.

In terms of full-length primer design, it is necessary to consider how to eliminate the interference of homologous genes, because the sequences at the beginning and the end of homologous genes are usually similar. The UTR region of a gene is usually specific. In SPDE v3.0, we will extract the UTR region corresponding to the gene according to the genomic sequence, and then design primers in the UTR region (at this time, the primers are called the first-round primers).

After using the first-round primers for amplification, you can use the products of the first-round amplification as templates, and then use the second-round primers (that is, the sequences at the beginning and the end of the gene) for amplification to obtain the full-length sequence.



```
301 RACPAEFVIP LAKYRKAICG SQLSVGMRFG MMFETEDSGK RRYMGTIVGI
351 SDLDPLRWPG SKWRNLQVEW DEPGCNDKPT RVSPWDIETP ESLFIFPSLT
401 SGLKRQLHPS YFAGETEWGS LIKRPLIRVP DSANGIMPYA SFPSMASEQL
451 MKMMNRPHNN QNVPSFMSEM QQNIVMGNGG LLGDMKMQQP LMMNQKSEMV
501 QPQNKLTVNP SASNTSGQEQ NLSQSMSAPA KPENSTLSGC SSGRVQHGLE
```

clear extra characters

计算引物的热力学参数

GTGGCGCGAGCTTCTGAAAC — Tm

上传基因组序列

rna-NM_099983.2 — 启动子长度

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

> 从基因组中提取序列

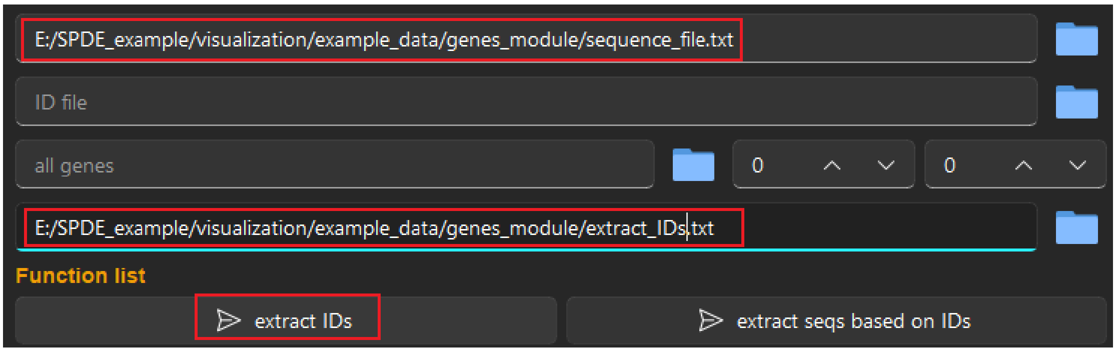
设计引物是为了: ○ qPCR ● 设计基因全长引物(需要放入基因组序列)

```
First-round primers
OLIGO      start      len      tm       gc%      any_th    3'_th     hairpin    seq
>>>>The first pair of primers recommended<<<<
left       85         20       58.9     50.0     0.72      3.67      0.0        AGCTCTTTACCCGGAAACCA
right      2074       20       59.3     50.0     3.33      3.41      0.0        TTCCGGAGCGAGAAAACAGA
>>>>The second pair of primers recommended<<<<
left       85         20       58.9     50.0     0.72      3.67      0.0        AGCTCTTTACCCGGAAACCA
right      2211       20       58.6     45.0     0.0       4.4       0.0        ACACCACAAAACGATGCACT
>>>>The third pair of primers recommended<<<<
left       85         20       58.9     50.0     0.72      3.67      0.0        AGCTCTTTACCCGGAAACCA
right      2075       20       58.4     50.0     3.33      3.16      0.0        TTTCCGGAGCGAGAAAACAG
>>>>The fourth pair of primers recommended<<<<
left       85         20       58.9     50.0     0.72      3.67      0.0        AGCTCTTTACCCGGAAACCA
right      2081       20       58.4     50.0     3.33      4.04      0.0        CAAACTTTTCCGGAGCGAGA
>>>>The fifth pair of primers recommended<<<<
left       85         20       58.9     50.0     0.72      3.67      0.0        AGCTCTTTACCCGGAAACCA
right      2082       20       58.4     50.0     3.33      5.03      0.0        TCAAACTTTTCCGGAGCGAG
Second-round primers
F: ATGGAGGATCAAGTTGGGTT
R: TTAACCAACAAGAATGATCC
```

2) Genes module

This module is mainly used for a series of operations such as batch extraction of gene IDs, sequence extraction, sequence translation, calculation of GC content, primer design, calculation of parameters like pI, molecular weight, length, extinction coefficient, and hydrophilicity, removal of repetitive sequences, file merging, extraction of the longest transcript, formatting of fasta files, and information query of fasta/fastq files. The detailed operations and regulations are as follows:

Extract all the IDs in the fasta file. The input file format is fasta. After setting the saving location and naming it, just click the button to complete the extraction.



E:/SPDE_example/visualization/example_data/genes_module/sequence_file.txt

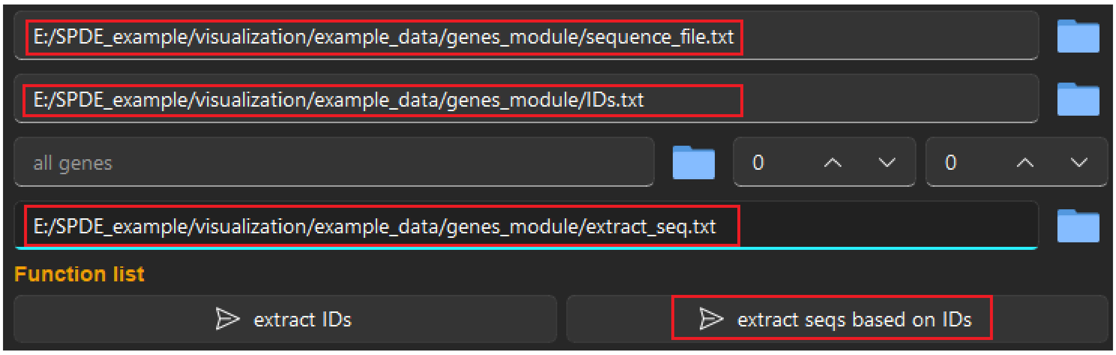
ID file

all genes — 0 — 0

E:/SPDE_example/visualization/example_data/genes_module/extract_IDs.txt

Function list

Extract sequences according to the IDs. Two files need to be prepared: one is the sequence file, and the other is the ID file. The format of the ID file is that each ID is on a separate line, and it should be ensured that the IDs are definitely present in the fasta file.



E:/SPDE_example/visualization/example_data/genes_module/sequence_file.txt

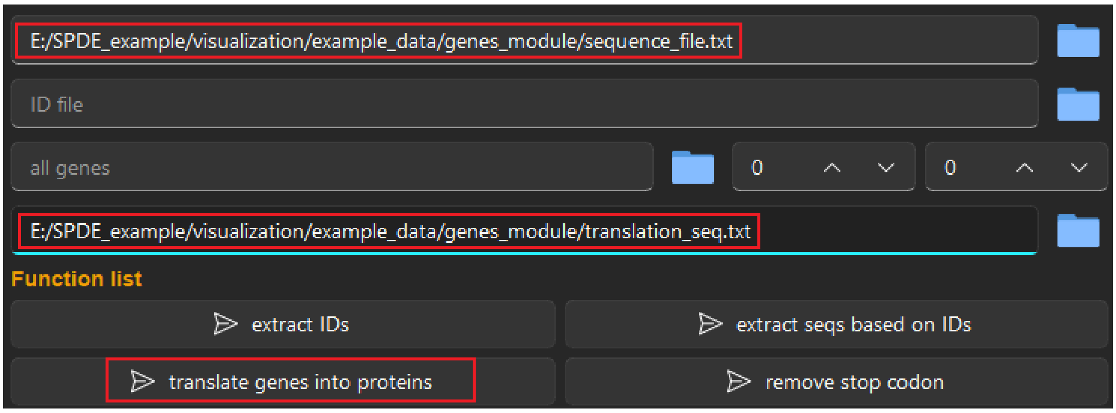
E:/SPDE_example/visualization/example_data/genes_module/IDs.txt

all genes — 0 — 0

E:/SPDE_example/visualization/example_data/genes_module/extract_seq.txt

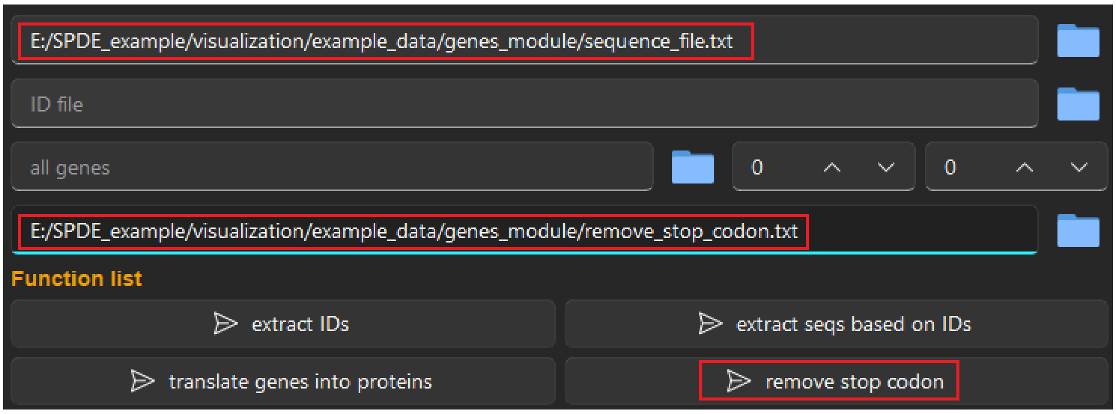
Function list

Translate the CDS file into a protein sequence. The input file is the CDS file. After setting the save location, click the button to save the translated result.


E:/SPDE_example/visualization/example_data/genes_module/sequence_file.txt

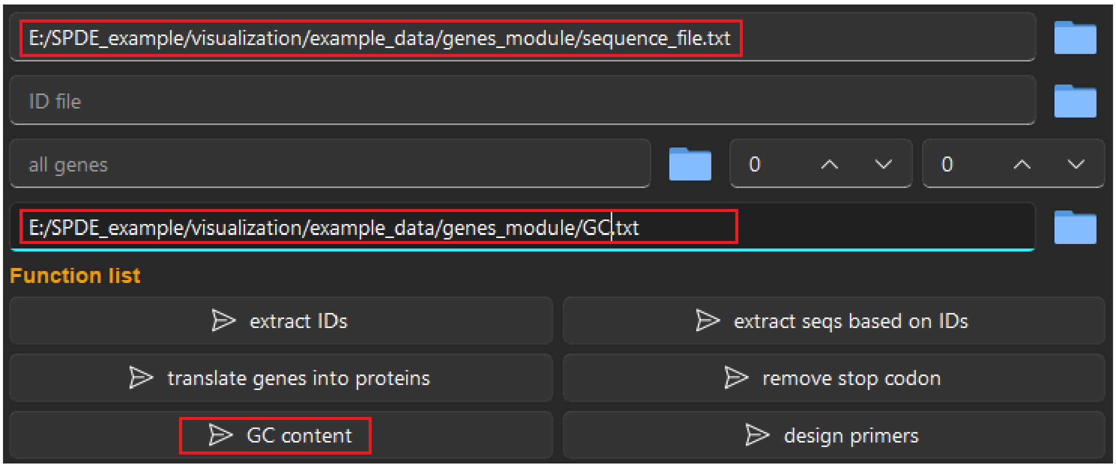
E:/SPDE_example/visualization/example_data/genes_module/translation_seq.txt

Batch remove stop codons. The input file should be in CDS format. After specifying the save location, click the button to save the result.


E:/SPDE_example/visualization/example_data/genes_module/sequence_file.txt

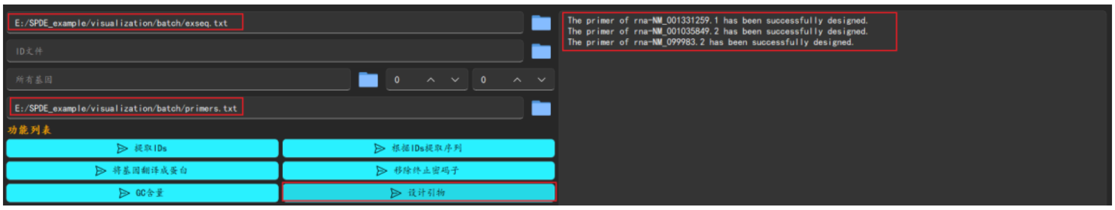
E:/SPDE_example/visualization/example_data/genes_module/remove_stop_codon.txt

Calculate the GC content. The input file should be in CDS format. After setting the save location, click the button to save the results.


E:/SPDE_example/visualization/example_data/genes_module/sequence_file.txt

E:/SPDE_example/visualization/example_data/genes_module/GC.txt

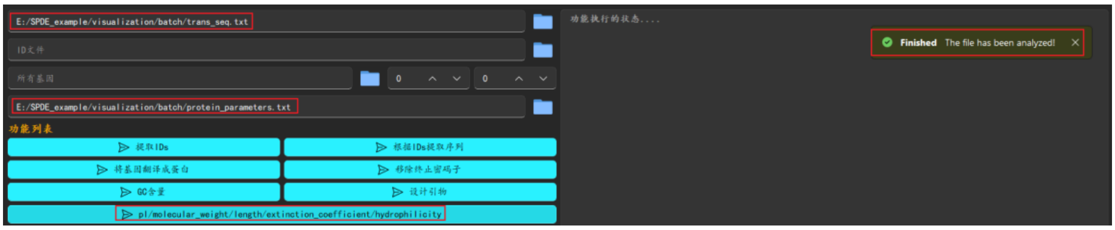
Design fluorescence quantitative primers in batches. The input file should be a DNA sequence file in FASTA format. After setting the save location, click the button to save the results.


E:/SPDE_example/visualization/batch/exseq.txt

E:/SPDE_example/visualization/batch/primers.txt

```
The primer of rna-NM_001331259.1 has been successfully designed.
The primer of rna-NM_001035849.2 has been successfully designed.
The primer of rna-NM_099983.2 has been successfully designed.
```

Conduct batch analysis of protein - related indicators. The input file is a protein sequence file in fasta format. After setting the save location, click the button to save the results.

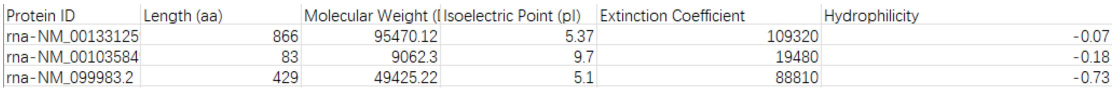

E:/SPDE_example/visualization/batch/trans_seq.txt

E:/SPDE_example/visualization/batch/protein_parameters.txt

功能执行的状态....

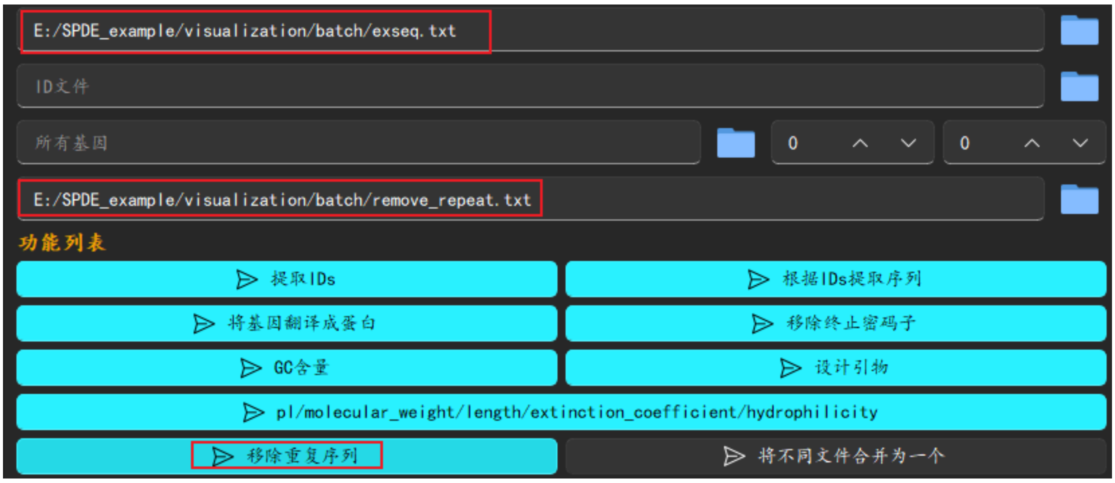
Finished The file has been analyzed!

Result：



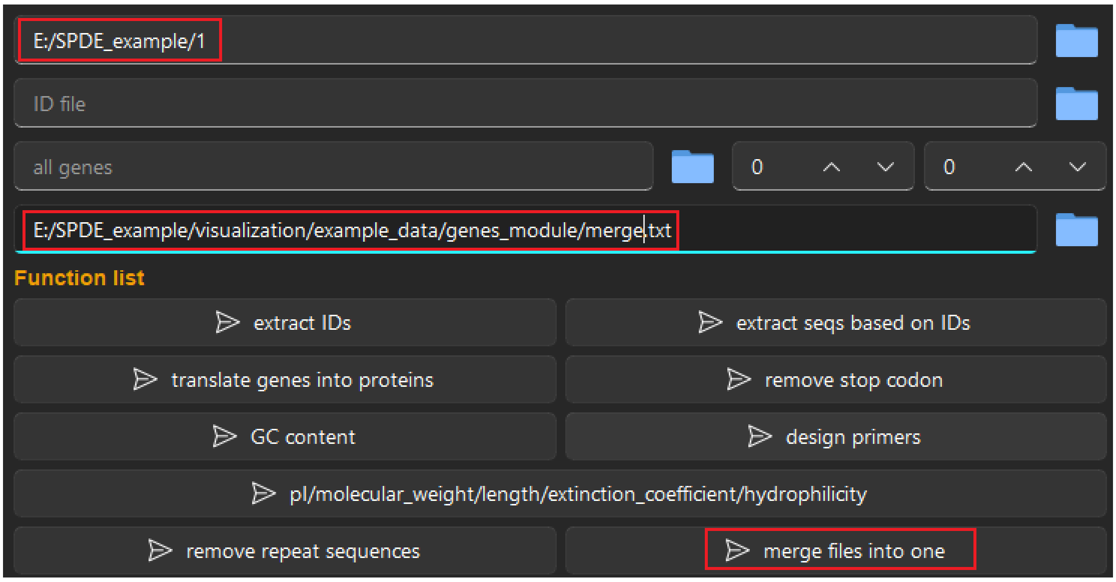
Protein ID	Length (aa)	Molecular Weight (I	Isoelectric Point (pI)	Extinction Coefficient	Hydrophilicity
rna-NM_00133125	866	95470.12	5.37	109320	-0.07
rna-NM_00103584	83	9062.3	9.7	19480	-0.18
rna-NM_099983.2	429	49425.22	5.1	88810	-0.73

Remove duplicate sequences from a FASTA file. After performing certain operations, the file may contain duplicate sequences. For example, when translating a CDS file, although the CDS sequences of some genes are different, the translated protein sequences may be similar.


E:/SPDE_example/visualization/batch/exseq.txt

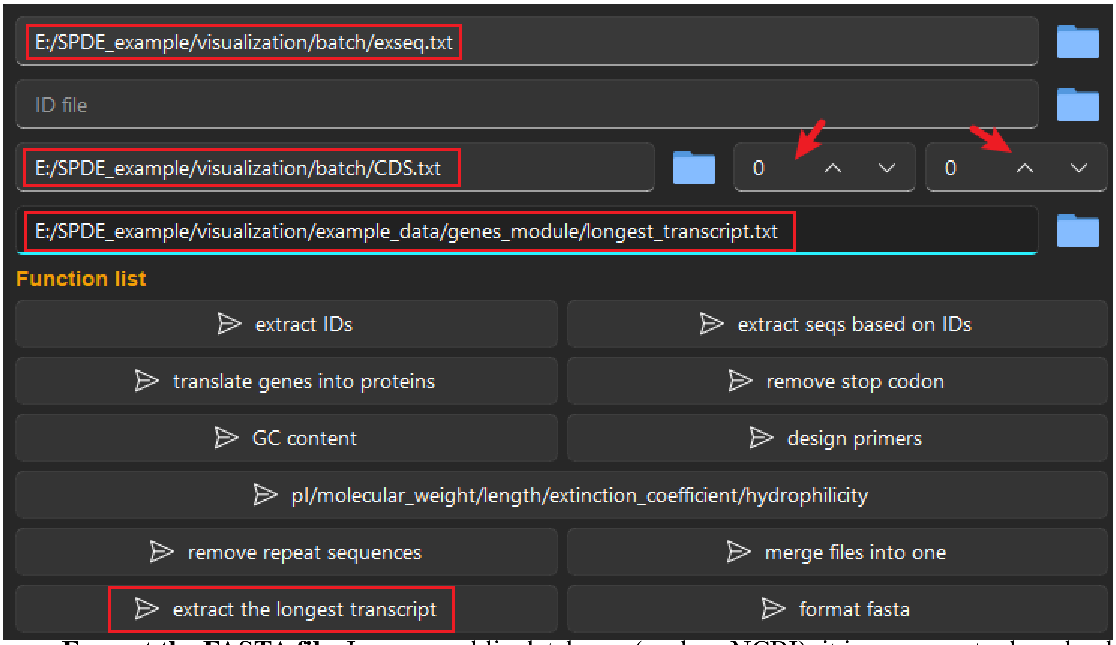
E:/SPDE_example/visualization/batch/remove_repeat.txt

Merge different files into one. You need to put the files to be merged into an empty folder. Then, use this folder as the input, set the save location and name for the merged file, and click the button to merge the different files.


E:/SPDE_example/1

E:/SPDE_example/visualization/example_data/genes_module/merge.txt

Extract the longest transcript. You need to provide two files: one is the assembled sequence of the transcriptome (or the gene sequence), and the other is the full - length sequences of all genes of the species (which can be obtained from the genome module). The basic process is as follows: SPDE v3.0 will align the transcripts to the full - length gene file. Then, according to the settings, the alignment information will be extracted and analyzed. Based on the analysis results, it will determine which genes the transcript sequences are mainly concentrated on. Finally, the analysis results will be returned. The basic operations are as follows:



E:/SPDE_example/visualization/batch/exseq.txt

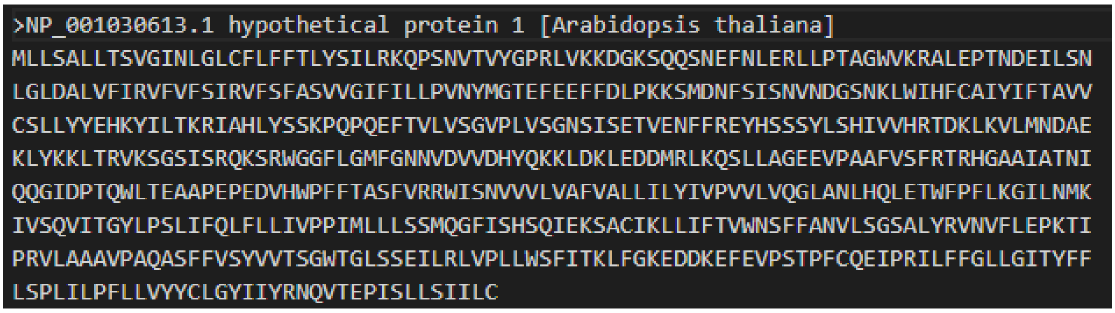
ID file

E:/SPDE_example/visualization/batch/CDS.txt 0 0

E:/SPDE_example/visualization/example_data/genes_module/longest_transcript.txt

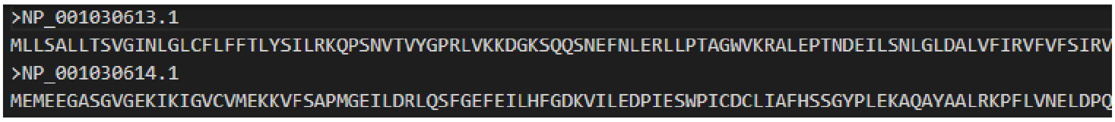
Function list
extract IDs
extract seqs based on IDs
translate genes into proteins
remove stop codon
GC content
design primers
pI/molecular_weight/length/extinction_coefficient/hydrophilicity
remove repeat sequences
merge files into one
extract the longest transcript
format fasta

Format the FASTA file. In some public databases (such as NCBI), it is common to download a FASTA file in which a single gene sequence is split into multiple lines, as shown below:



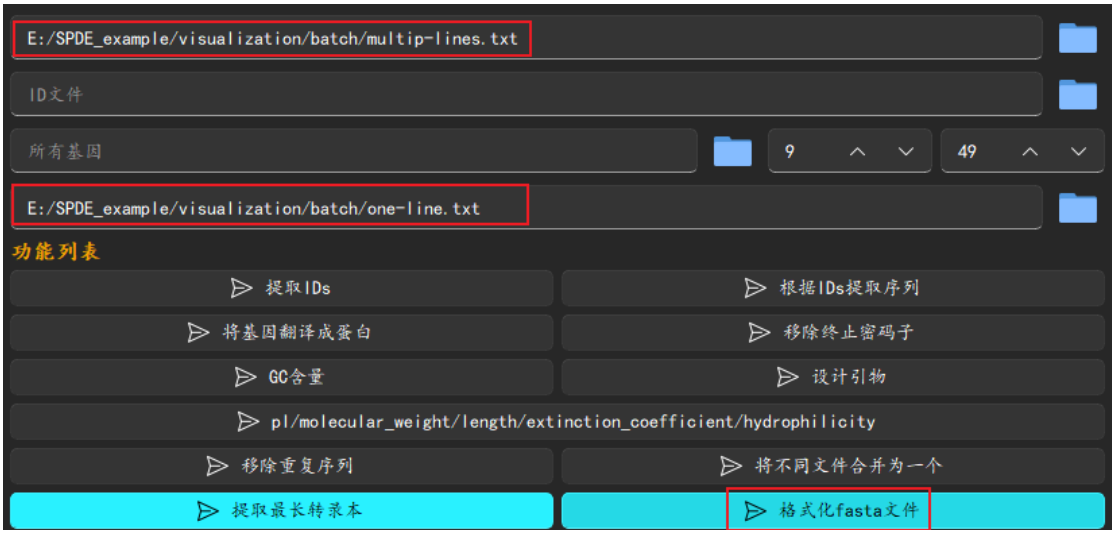
```
>NP_001030613.1 hypothetical protein 1 [Arabidopsis thaliana]
MLLSALLTSVGINLGLCFLFFTLYSILRKQPSNVTVYGPRLVKKDGKSQQSNEFNLERLLPTAGWVKRALEPTNDEILSN
LGLDALVFIRVFVFSIRVFSFASVVGIFILLPVNYMGTEFEEFFDLPKKSMDNFSISNVNDGSNKLWIHFCAIYIFTAVV
CSLLYYEHKYILTKRIAHLYSSKPQPQEFTVLVSGVPLVSGNSISETVENFFREYHSSSYLSHIVVHRTDKLKVLMNDAE
KLYKKLTRVKSGSISRQKSRWGGFLGMFGNNVDVVDHYQKKLDKLEDDMRLKQSLLAGEEVPAAFVSFRTRHGAAIATNI
QQGIDPTQWLTEAAPEPEDVHWPFFTASFVRRWISNVVVLVAFVALLILYIVPVVLVQGLANLHQLETWFPFLKGILNMK
IVSQVITGYLPSLIFQLFLLIVPPIMLLLSSMQGFISHSQIEKSACIKLLIFTVWNSFFANVLSGSALYRVNVFLEPKTI
PRVLAAAVPAQASFFVSYVVTSGWTGLSSEILRLVPLLWSFITKLFGKEDDKEFEVPSTPFCQEIPRILFFGLLGITYFF
LSPLILPFLLVYYCLGYIIYRNQVTEPISLLSIILC
```

In this format, since carriage return characters (line break characters) exist within the sequences, there will be problems when trying to search for certain sequences. Therefore, a function for formatting FASTA files has been set up. This function arranges FASTA files with sequences on different lines into a single line, as follows:



```
>NP_001030613.1
MLLSALLTSVGINLGLCFLFFTLYSILRKQPSNVTVYGPRLVKKDGKSQQSNEFNLERLLPTAGWVKRALEPTNDEILSNLGLDALVFIRVFVFSIRV
>NP_001030614.1
MEMEEGASGVGEKIKIGVCVMEKKVFSAPMGEILDRLQSFGEFEILHFGDKVILEDPIESWPICDCLIAFHSSGYPLEKAQAYAALRKPFLVNELDPQ
```

Functions as follows：


E:/SPDE_example/visualization/batch/multip-lines.txt

ID文件

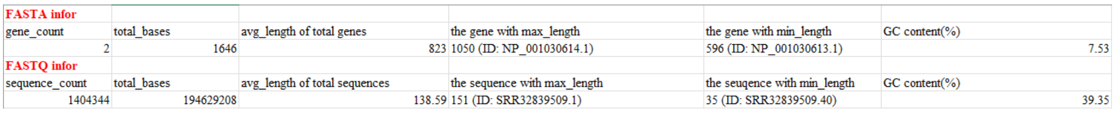
所有基因

9

49

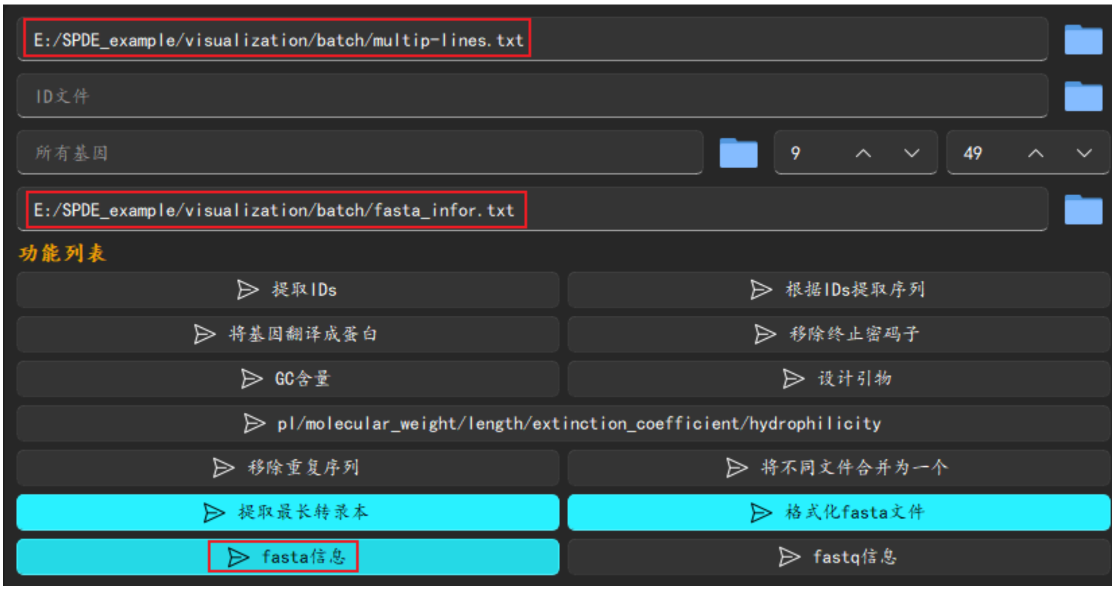
E:/SPDE_example/visualization/batch/one-line.txt

Perform information statistics on FASTA and FASTQ files. The statistics are as follows:



FASTA infor					
gene_count	total_bases	avg_length of total genes	the gene with max_length	the gene with min_length	GC content(%)
2	1646	823	1050 (ID: NP_001030614.1)	596 (ID: NP_001030613.1)	7.53
FASTQ infor					
sequence_count	total_bases	avg_length of total sequences	the sequence with max_length	the seuqence with min_length	GC content(%)
1404344	194629208	138.59	151 (ID: SRR32839509.1)	35 (ID: SRR32839509.40)	39.35

Settings as follows:


E:/SPDE_example/visualization/batch/multip-lines.txt

ID文件

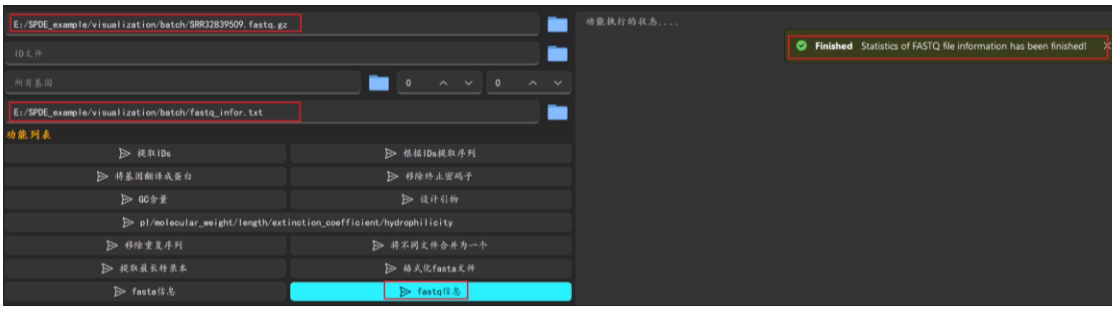
所有基因

9

49

E:/SPDE_example/visualization/batch/fasta_infor.txt

Fastq file analysis


E:/SPDE_example/visualization/batch/SRR32839509.fastq.gz

ID文件

所有基因

0

0

E:/SPDE_example/visualization/batch/fastq_infor.txt

功能执行的状态....

Finished Statistics of FASTQ file information has been finished!

Please note: To meet the increasingly growing demand for data analysis, most of the functions in this module support a higher level of batch processing. For example, when you put multiple CDS files that need to be translated into an empty folder and use this folder as the input, SDPE v3.0 will perform the translation operation on each file within that folder.

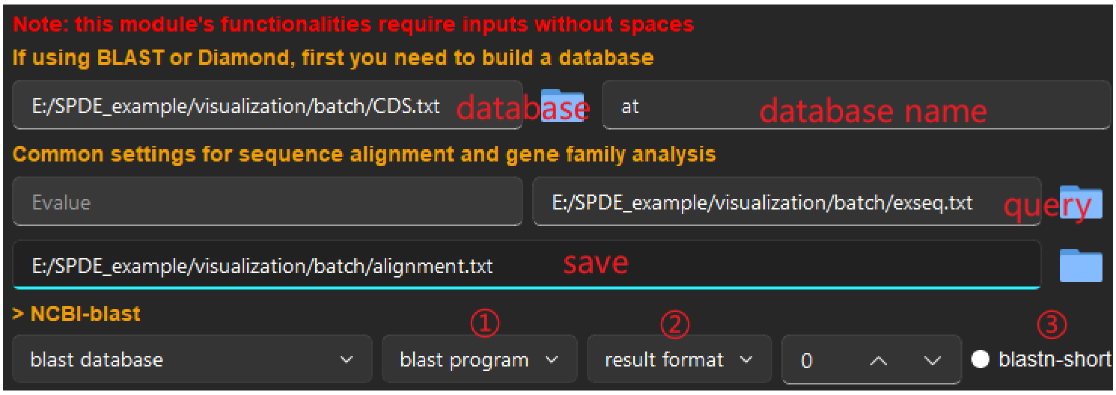
3) Alignment module

Sequence alignment is an important foundation of the entire bioinformatics field. In this module, we have summarized the commonly used alignment programs currently available, and you can flexibly choose according to their own needs. All operations can be carried out by clicking.

Diamond and NCBI-Blast

Differences: Diamond runs faster than NCBI-Blast. However, the drawback of Diamond is that this program can only be used for protein alignment. (Certainly, you may find that this program can align DNA sequences, but there is a prerequisite for this kind of alignment: the DNA sequence must be able to be translated into a protein. In other words, when it comes to the so-called DNA alignment, it still needs to be translated into a protein first during operation and then undergo alignment. It's just that you don't need to do the translation themselves). The advantage of NCBI-Blast is that it can be used to align more types of data.

Common points: When in use, the query sequence (the sequence of demand) is aligned with the reference sequence (the library sequence) to analyze the alignment results. Therefore, both a library file and a query file are required. Building a library is essential, especially when using this library for the first time. In SPDE v3.0, the library construction is automatic. Once the library is constructed, SPDE v3.0 will automatically load the constructed library when it is launched next time (of course, on the premise that you have not actively deleted it). Therefore, you do not need to repeatedly build the same library. The basic settings are as follows:



Note: this module's functionalities require inputs without spaces

If using BLAST or Diamond, first you need to build a database

E:/SPDE_example/visualization/batch/CDS.txt

at

Common settings for sequence alignment and gene family analysis

Evalue

E:/SPDE_example/visualization/batch/exseq.txt

E:/SPDE_example/visualization/batch/alignment.txt

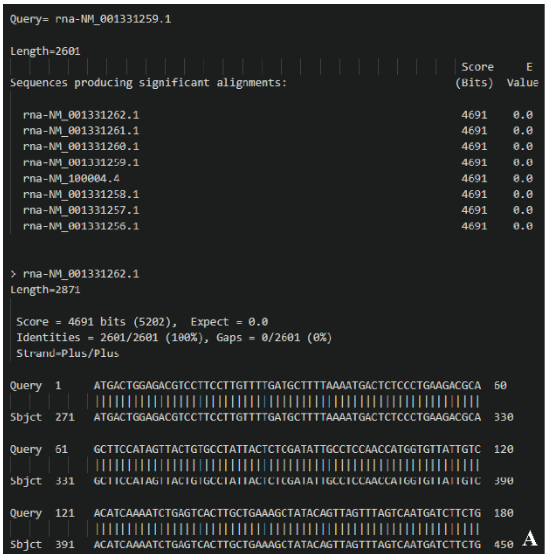
When using it for the first time, directly put in the database file and set its name. Add the query sequence, save it and give it a name, and then click the NCBI-blast button to run the alignment.

When using it again and the database has already been constructed, you can directly click the 'blast database' option to use the library for alignment.

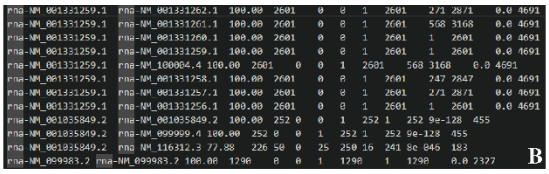
Explanations of other options:

① These are several alignment methods of NCBI-blast included. Blastn is used for the alignment between DNAs; blastp is for protein alignment; blastx is specifically used for alignment after translating a nucleic acid sequence into a protein sequence; tblastn aligns a given amino acid sequence with the sequences (double-stranded) in a nucleic acid database according to different reading frames, which is very useful for finding new coding regions that are not annotated in the sequences in the database; tblastx is only used in special cases. It translates both the DNA sequence to be searched and the sequences in the nucleic acid sequence database into protein sequences according to different reading frames, and then conducts protein sequence alignment.

② There are two modes: normal and fmt6. These two modes refer to different organizational forms of the alignment results. The manifestation of the normal mode is as shown in Figure A below, and the fmt6 format is as shown in Figure B:



```
Query= rna-NM_001331259.1

Length=2601
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  rna-NM_001331262.1                                                  4691    0.0
  rna-NM_001331261.1                                                  4691    0.0
  rna-NM_001331260.1                                                  4691    0.0
  rna-NM_001331259.1                                                  4691    0.0
  rna-NM_100004.4                                                     4691    0.0
  rna-NM_001331258.1                                                  4691    0.0
  rna-NM_001331257.1                                                  4691    0.0
  rna-NM_001331256.1                                                  4691    0.0


> rna-NM_001331262.1
Length=2871

 Score = 4691 bits (5202),  Expect = 0.0
 Identities = 2601/2601 (100%), Gaps = 0/2601 (0%)
 Strand=Plus/Plus

Query  1    ATGACTGGAGACGTCCTTCCTTGTTTTGATGCTTTTAAAATGACTCTCCCTGAAGACGCA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  271  ATGACTGGAGACGTCCTTCCTTGTTTTGATGCTTTTAAAATGACTCTCCCTGAAGACGCA  330

Query  61   GCTTCCATAGTTACTGTGCCTATTACTCTCGATATTGCCTCCAACCATGGTGTTATTGTC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  331  GCTTCCATAGTTACTGTGCCTATTACTCTCGATATTGCCTCCAACCATGGTGTTATTGTC  390

Query  121  ACATCAAAATCTGAGTCACTTGCTGAAAGCTATACAGTTAGTTTAGTCAATGATCTTCTG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  391  ACATCAAAATCTGAGTCACTTGCTGAAAGCTATACAGTTAGTTTAGTCAATGATCTTCTG  450
```



```
rna-NM_001331259.1  rna-NM_001331262.1  100.00  2601    0   0   1   2601    271 2871    0.0 4691
rna-NM_001331259.1  rna-NM_001331261.1  100.00  2601    0   0   1   2601    568 3168    0.0 4691
rna-NM_001331259.1  rna-NM_001331260.1  100.00  2601    0   0   1   2601    1   2601    0.0 4691
rna-NM_001331259.1  rna-NM_001331259.1  100.00  2601    0   0   1   2601    1   2601    0.0 4691
rna-NM_001331259.1  rna-NM_100004.4 100.00  2601    0   0   1   2601    568 3168    0.0 4691
rna-NM_001331259.1  rna-NM_001331258.1  100.00  2601    0   0   1   2601    247 2847    0.0 4691
rna-NM_001331259.1  rna-NM_001331257.1  100.00  2601    0   0   1   2601    271 2871    0.0 4691
rna-NM_001331259.1  rna-NM_001331256.1  100.00  2601    0   0   1   2601    1   2601    0.0 4691
rna-NM_001035849.2  rna-NM_001035849.2  100.00  252 0   0   1   252 1   252 9e-128  455
rna-NM_001035849.2  rna-NM_099999.4 100.00  252 0   0   1   252 1   252 9e-128  455
rna-NM_001035849.2  rna-NM_116312.3 77.88   226 50  0   25  250 16  241 8e-046  183
rna-NM_099983.2 rna-NM_099983.2 100.00  1290    0   0   1   1290    1   1290    0.0 2327
```

fmt6各列格式说明：

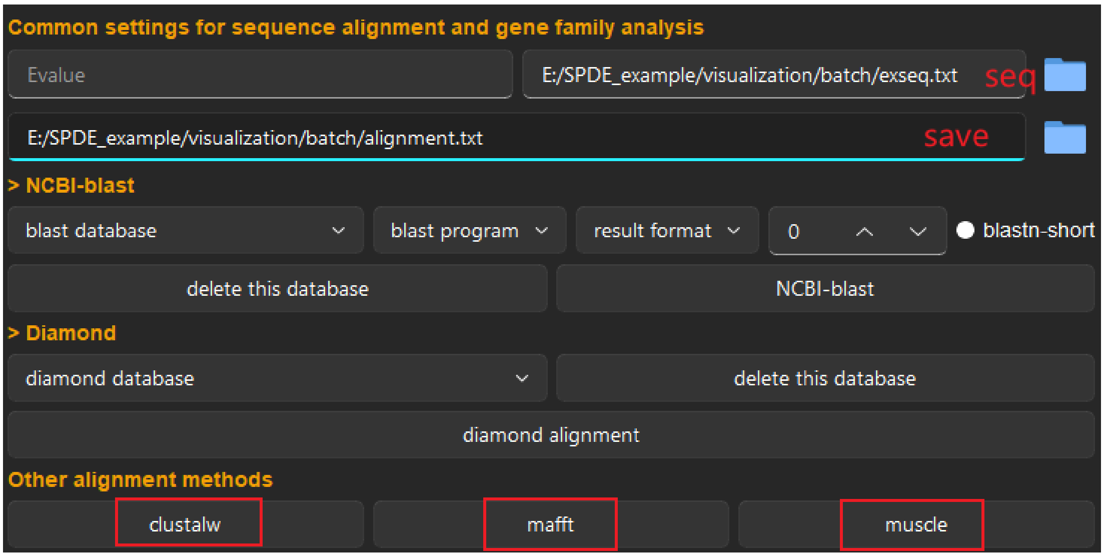
qseqid	查询序列的ID
sseqid	库序列的ID
pident	查询和库序列之间的百分比一致性
length	比对的长度
mismatch	不匹配的数量
gapopen	间隙的数量
qstart	查询序列的起始位置
qend	查询序列的结束位置
sstart	库序列的起始位置
send	库序列的结束位置
evalue	E值
bitscore	比分值

This option is used to align short sequences to the database. (In other words, if this option is not ticked, blast won't be able to align short sequences.)

The usage method of Diamond is similar to that of NCBI-blast.

clustalw、mafft and muscle

Unlike Diamond and NCBI-blast, when running ClustalW, MAFFT, and MUSCLE, there is no need to construct or select a database. They are more about aligning the input sequences with each other. The basic settings are as follows:



E:/SPDE_example/visualization/batch/exseq.txt

E:/SPDE_example/visualization/batch/alignment.txt

The input sequences are also in FASTA format.

Gene family

SPDE v3.0 has further optimized the gene family analysis: it has collected and organized the characteristic domains of over 120 families. After selecting a specific family, you can complete the processes of family member screening, sequence extraction, domain identification, and visualization with a single click. Here are several points to note:

1) The input sequences should be protein sequences in FASTA format.

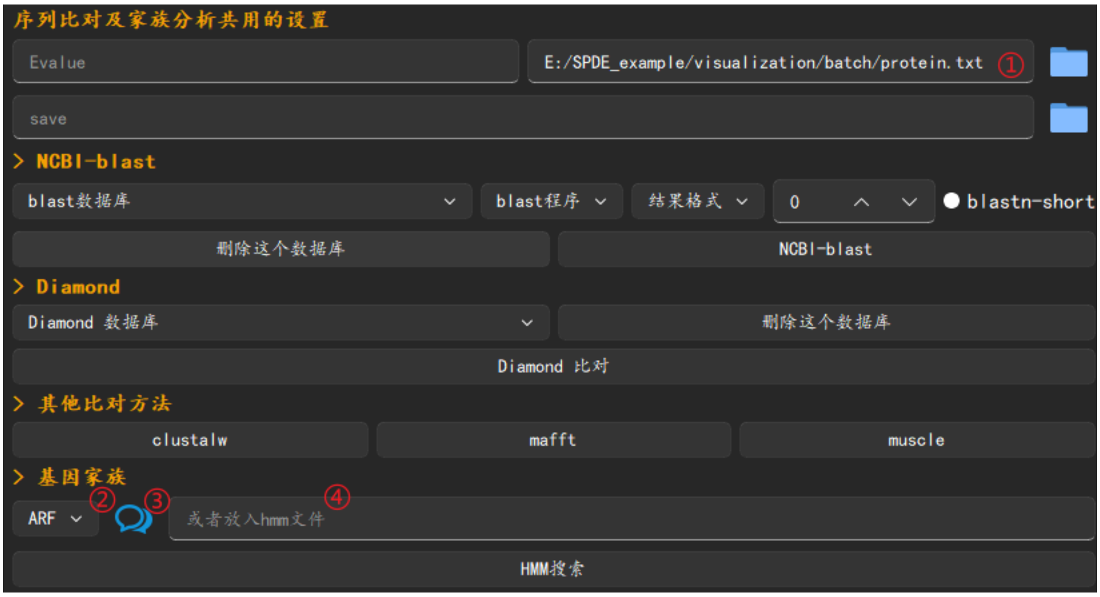
2) The characteristic domain refers to the domain that distinguishes this family from other families. When a family contains multiple domains and some of these domains are not present in all members, these domains are excluded.

3) Users are allowed to construct and input their own domain models. The construction method is as follows: download the Pfam files of the protein domains of the family you are interested in. If a family contains multiple essential domains, you need to merge these Pfam files into one. When downloading Pfam files, you need to know the Pfam ID of the protein domain. You can obtain this Pfam ID from gene family - related articles or find a protein sequence that definitely belongs to this family and use NCBI - BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM = blastp&PAGE_TYPE = BlastSearch&LINK_LOC = blasthome) to determine the ID. After clarifying the ID, you can refer to my blog (https://www.jianshu.com/p/934f63b4b1a3?v = 1742889666646) to download the corresponding Pfam file.

4) We have placed the entire Pfam database in our GitHub account (https://github.com/simon19891216/SPDE/releases/tag/database), which is mainly used for the comprehensive identification of the protein domains of family members. However, this database is relatively large. You can choose whether to download and use it according to their own circumstances. The usage method is as follows: download the database; open the installation location of SPDE v3.0, and you will see a 'database/all_pfam' folder; put the downloaded file into

this folder and right - click to select 'Extract Here'. The program will automatically retrieve it when in use.

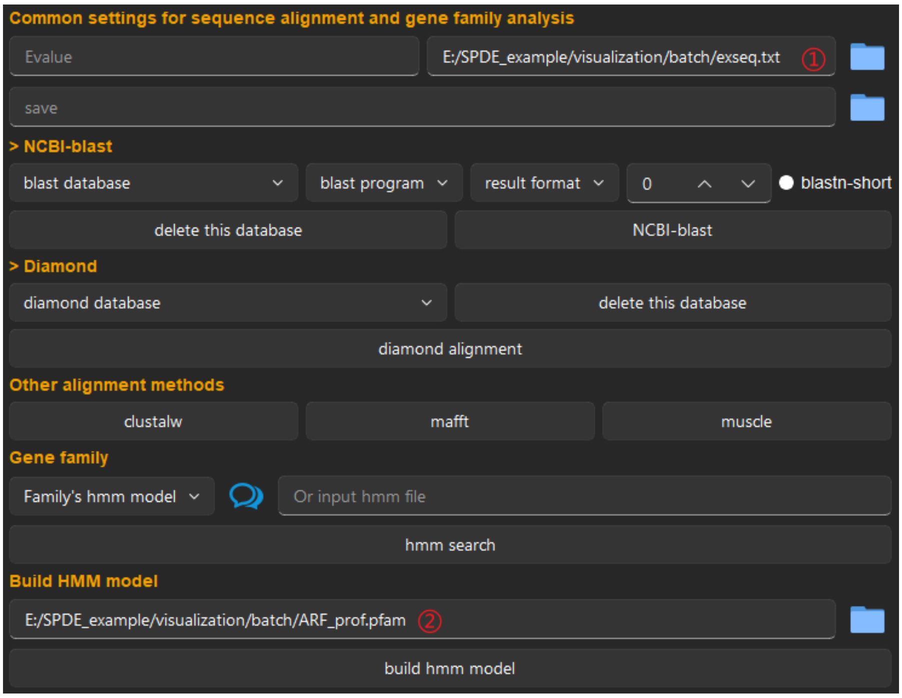
The basic parameter settings are as follows:


E:/SPDE_example/visualization/batch/protein.txt ①

Place the protein sequence in ①; select a model in ② or put your self - organized model in ④; ③ is a button for the full name of the model. Clicking it will display the full name of the family. The results will be placed in the same - level folder as the protein sequence.

Construction of Hidden Markov Model (HMM)

If the gene family you're researching hasn't been fully studied yet, you might consider building your own model to identify its members. In SPDE v3.0, I've included a function for constructing HMM models. The basic settings are as follows:


E:/SPDE_example/visualization/batch/exseq.txt ①

E:/SPDE_example/visualization/batch/ARF_prof.pfam ②

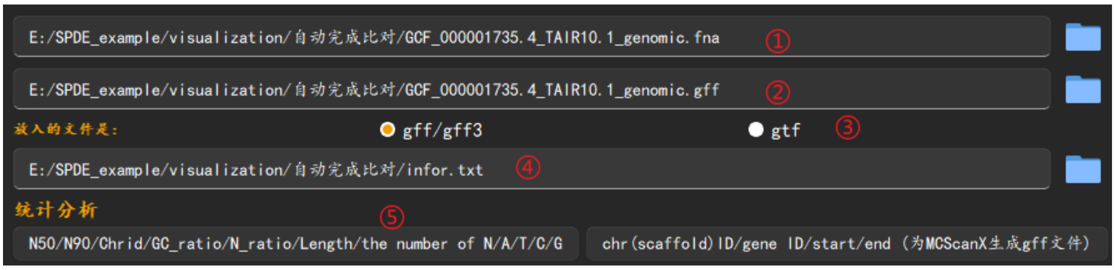
Put the sequences in ①; specify the location to save the model and name it in ②.

It should be noted that the sequences to be input should have conserved structures. you need to make a judgment first. After confirming the presence of conserved sequences, they can proceed with model construction.

4) Genome module

This module is mainly used for data analysis at the genomic level and operations such as sequence extraction. The main functions it performs are: information query (including N50, distribution of the number of bases, GC content, etc.); distribution of specific elements (including the positional distribution of kmer, gene, mRNA, CDS, exon, gap (only for the Windows version), etc.) on the genome; batch sequence extraction (including promoter, mRNA, gene, CDS, and UTR sequences); and collinearity analysis of the genome. The basic settings are as follows:

Genomic Information Query



E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

放入的文件是： ● gff/gff3 ○ gtf

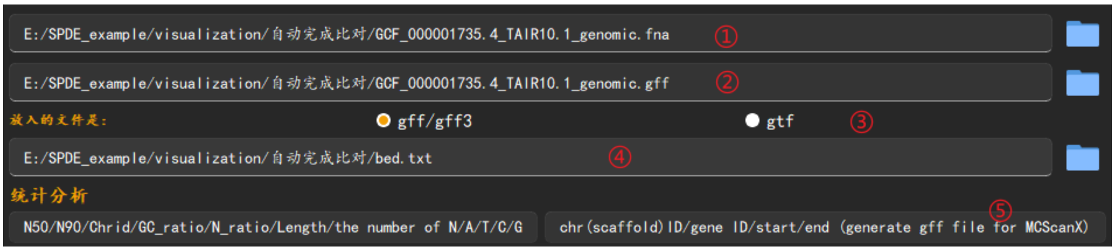
E:/SPDE_example/visualization/自动完成比对/infor.txt

①
②
③
④
统计分析
⑤
N50/N90/Chrid/GC_ratio/N_ratio/Length/the number of N/A/T/C/G
chr(scaffold)ID/gene ID/start/end (为MCScanX生成gff文件)

① Genomic sequence file; ② GFF or GTF annotation file. Meanwhile, you need to specify the input format in ③; ④ Save the file and give it a name; Click ⑤ to execute the function.

Generation of the BED File

The BED file is used in many programs. The format of the BED file generated by SPDE is: chr ID, gene ID, start position, end position.



E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

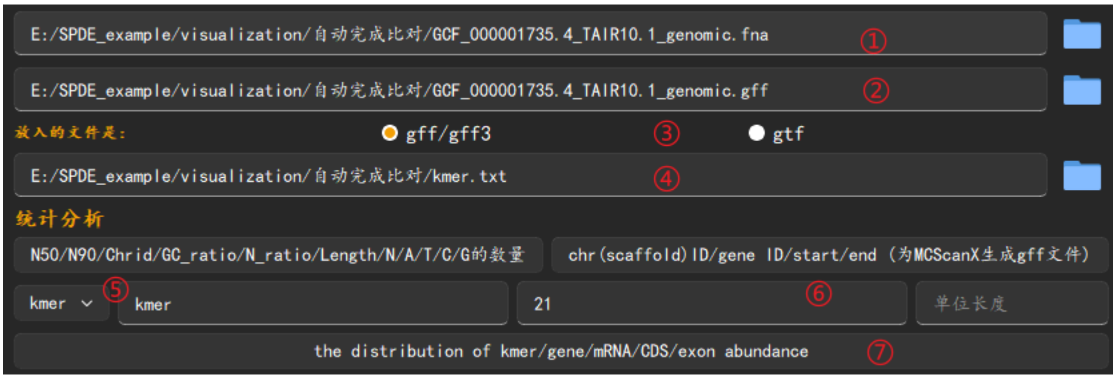
放入的文件是： ● gff/gff3 ○ gtf

E:/SPDE_example/visualization/自动完成比对/bed.txt

①
②
③
④
统计分析
N50/N90/Chrid/GC_ratio/N_ratio/Length/the number of N/A/T/C/G
⑤
chr(scaffold)ID/gene ID/start/end (generate gff file for MCScanX)

① Genomic sequence file; ② GFF or GTF annotation file. Also, it is necessary to indicate the input format at ③; ④ Save and name the file; Click ⑤ to execute the function.

The density distribution of elements on the genome



E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna ①

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff ②

载入的文件是： ● gff/gff3 ③ ○ gtf

E:/SPDE_example/visualization/自动完成比对/kmer.txt ④

统计分析

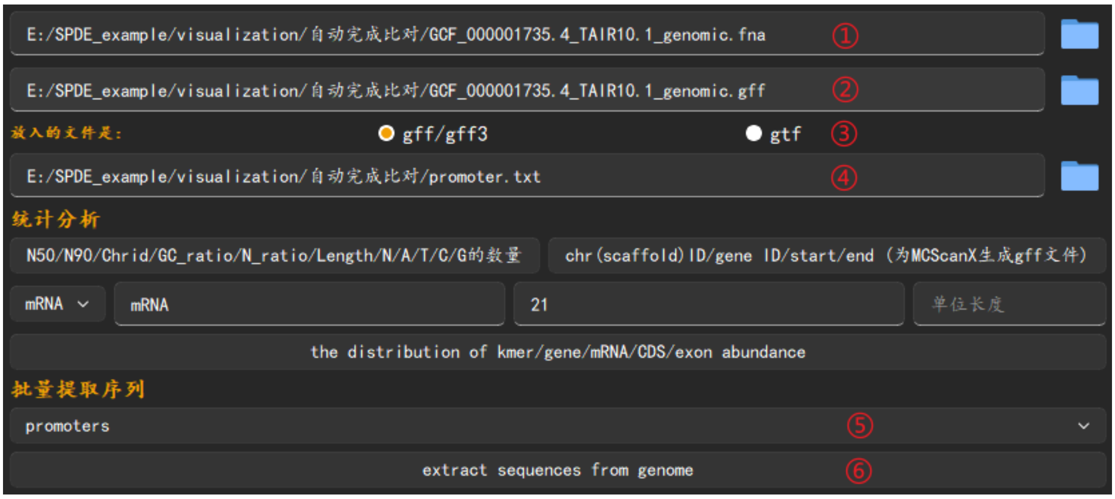
N50/N90/Chrid/GC_ratio/N_ratio/Length/N/A/T/C/G的数量 | chr(scaffold)ID/gene ID/start/end（为MCScanX生成gff文件）

kmer ⑤ | kmer | 21 ⑥ | 单位长度

the distribution of kmer/gene/mRNA/CDS/exon abundance ⑦

① Genomic sequence file; ② GFF or GTF annotation file. Meanwhile, it is necessary to specify the input format at ③; ④ Save and name the file; Click ⑤ to select the type; If it is kmer, you can set the size of the kmer by selecting options at ⑥; Click ⑦ to execute the function.

Batch sequence extraction



E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna ①

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff ②

载入的文件是： ● gff/gff3 ○ gtf ③

E:/SPDE_example/visualization/自动完成比对/promoter.txt ④

统计分析

N50/N90/Chrid/GC_ratio/N_ratio/Length/N/A/T/C/G的数量 | chr(scaffold)ID/gene ID/start/end（为MCScanX生成gff文件）

mRNA | mRNA | 21 | 单位长度

the distribution of kmer/gene/mRNA/CDS/exon abundance

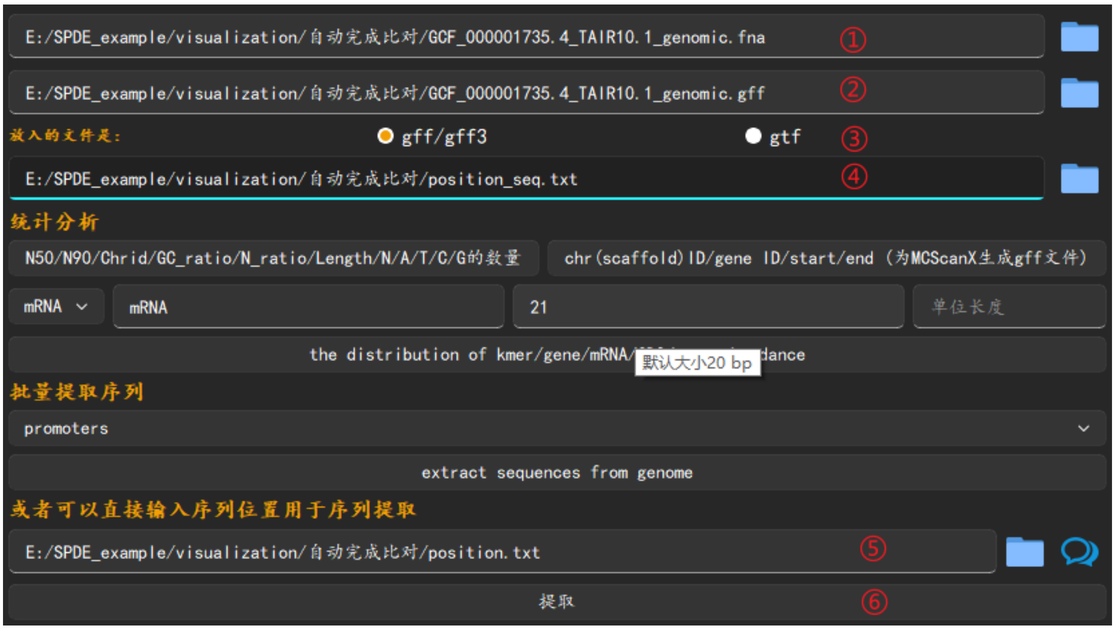
批量提取序列

promoters ⑤

extract sequences from genome ⑥

① Genomic sequence file; ② GFF or GTF annotation file. At the same time, you need to indicate the input format at ③. ④ Save the results and give them a name. Click ⑤ to select the sequence type. Click ⑥ to execute the function.

Free - form sequence extraction



E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/自动完成比对/GCF_000001735.4_TAIR10.1_genomic.gff

放入的文件是: ● gff/gff3 ○ gtf

E:/SPDE_example/visualization/自动完成比对/position_seq.txt

统计分析

mRNA | mRNA | 21 | 单位长度

默认大小20 bp

批量提取序列

promoters

或者可以直接输入序列位置用于序列提取

E:/SPDE_example/visualization/自动完成比对/position.txt

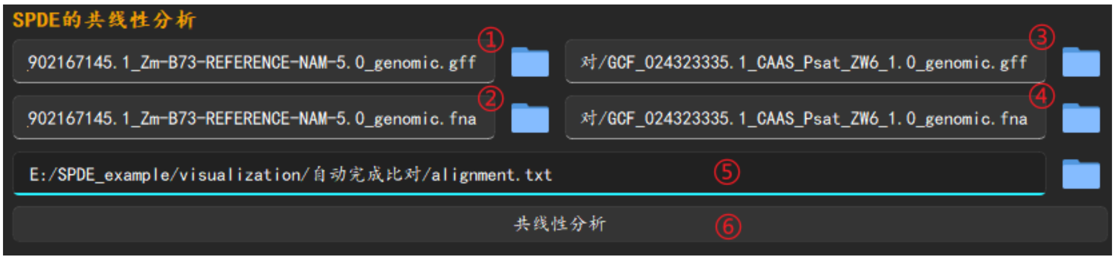
①
②
③
④
N50/N90/Chrid/GC_ratio/N_ratio/Length/N/A/T/C/G的数量
chr(scaffold)ID/gene ID/start/end (为MCScanX生成gff文件)
the distribution of kmer/gene/mRNA/...dance
extract sequences from genome
⑤
提取
⑥

① Genomic sequence file; ② GFF or GTF annotation file. Meanwhile, you need to specify the input format at ③; ④ Save and name the output; ⑤ Input the position file; Click ⑥ to execute the function.

The format of the position file is: chromosome ID, start position, end position, gene ID, positive or negative strand.

SPDE collinearity analysis

In previous collinearity analyses, users had to prepare multiple files, and the intermediate processes were quite complex. In SPDE v3.0, we have further optimized the process. We conducted tests on species with genomes larger than 2 Gb and found that the collinearity analysis between two species can be completed within 2 minutes. The input requirements for this function are also very simple. You only need the genomic sequence file and the GFF file. The basic settings are shown in the following figure:



SPDE的共线性分析

902167145.1_Zm-B73-REFERENCE-NAM-5.0_genomic.gff

902167145.1_Zm-B73-REFERENCE-NAM-5.0_genomic.fna

对/GCF_024323335.1_CAAS_Psat_ZW6_1.0_genomic.gff

对/GCF_024323335.1_CAAS_Psat_ZW6_1.0_genomic.fna

E:/SPDE_example/visualization/自动完成比对/alignment.txt

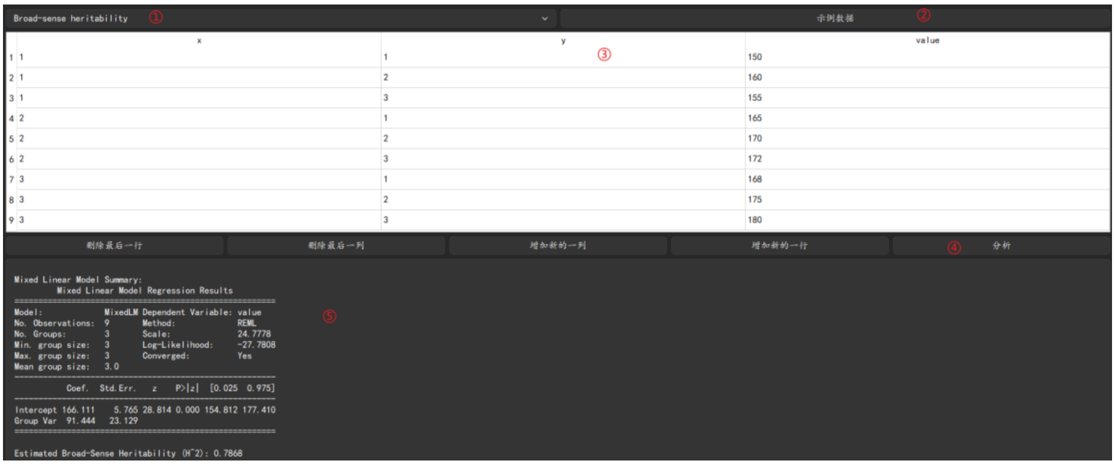
① and ②, as well as ③ and ④, correspond to the GFF files and genomic sequence files of two species respectively; ⑤ Set the storage location for the alignment results; Click ⑥ to execute the function.

We have also configured a visualization function for this feature (this part will be introduced later). The file used for visualization is the one with the flag 'collinearity_visualization.txt'.

5) Breeding module

Previous breeding science calculations relied on statistical software such as SAS or SPSS. However, these software are not specifically designed for breeding science, which makes the entire calculation function relatively complex. In SPDE v3.0, we have designed more than 40 calculation functions in the field of breeding science to efficiently complete breeding science calculations.

To achieve efficient analysis, this module has high requirements for the organizational form of the input data. Therefore, when designing the functions, I have imposed strict restrictions, allowing users to only replace the sample data with their own data based on the sample data for analysis. The basic operation process is as follows:



	x	y	value
1	1	1	150
2	1	2	160
3	1	3	155
4	2	1	165
5	2	2	170
6	2	3	172
7	3	1	168
8	3	2	175
9	3	3	180

```
Mixed Linear Model Summary:
         Mixed Linear Model Regression Results
==========================================================
Model:            MixedLM Dependent Variable: value
No. Observations: 9       Method:             REML
No. Groups:       3       Scale:              24.7778
Min. group size:  3       Log-Likelihood:     -27.7808
Max. group size:  3       Converged:          Yes
Mean group size:  3.0
----------------------------------------------------------
           Coef.  Std.Err.   z    P>|z|  [0.025  0.975]
----------------------------------------------------------
Intercept 166.111    5.765 28.814 0.000 154.812 177.410
Group Var  91.444   23.129
==========================================================

Estimated Broad-Sense Heritability (H^2): 0.7868
```

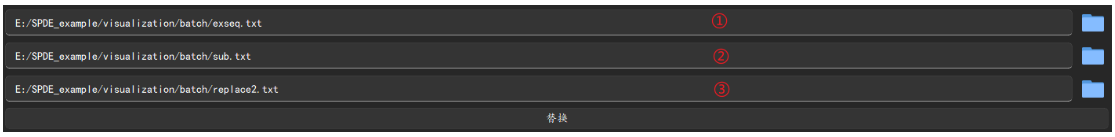
You should first select the analysis method you need through ①, and then click ② to automatically retrieve the sample data. As shown in the figure above, when calculating the broad - sense heritability, you only need to replace the data in the sample with your own data for x, y, and the resulting breeding values (i.e., value). Then click ④ to complete the calculation. The calculation results will be displayed in ⑤. Of course, it is not required that the number of rows and columns in your data must be the same as that of the sample data. You can delete or add rows and columns by using the buttons for increasing or decreasing rows and columns in the settings.

6) Edit file

This module is mainly used to handle some common text requirements, including replacing text, extracting content, reorganizing file content, converting file formats, downloading files, and browsing large files. The following are the specific applications of its functions:

Replace file content

It is difficult to replace multiple text contents at one time in commonly used software, yet this kind of replacement is frequently utilized in bioinformatics. For example, the gene names in the sequence files we deal with often have no practical functional significance. Therefore, it is necessary to replace the original IDs with IDs that carry functional meanings. The basic settings are as follows:



E:/SPDE_example/visualization/batch/exseq.txt ①

E:/SPDE_example/visualization/batch/sub.txt ②

E:/SPDE_example/visualization/batch/replace2.txt ③

Put the original file in ①; Put the content to be replaced in ②. There are two forms:

1. When there is a large amount of content to be replaced, it can be organized into a file. Each line contains a set of replacement content, and each set has two columns. The first column is the original content, and the second column is the content to be replaced. The two columns are separated by a space. For example, rna-NM_099983.2 ARF3. Therefore, it should be noted that there should be no spaces in the content before and after the replacement.

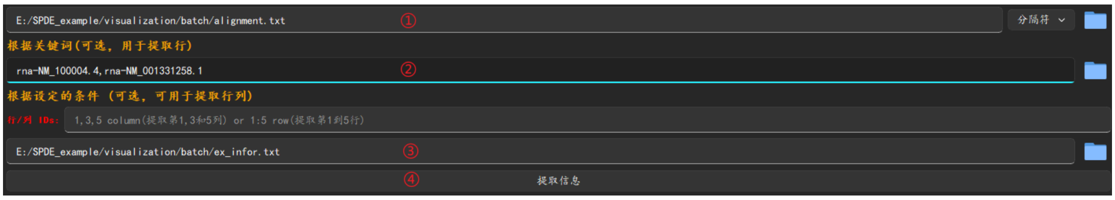
2. If there is not much content to be replaced, the format should be like apple:orange, banana:grape, cat:dog, that is, different sets are separated by commas, and each set is divided into columns by colons; Set the save location and name the file; Press the function key to execute the function.

Note that the original file must be a text file, not a file created by Word or Excel.

Extract the Required Information

In daily analysis, some files contain a large amount of content, and some of this content is not needed for subsequent analysis. Therefore, certain functions need to be designed to efficiently extract the required content.

1) Extract the corresponding rows through keywords (separate the keywords with commas).



E:/SPDE_example/visualization/batch/alignment.txt ①

分隔符

根据关键词(可选，用于提取行)

rna-NM_100004.4,rna-NM_001331258.1 ②

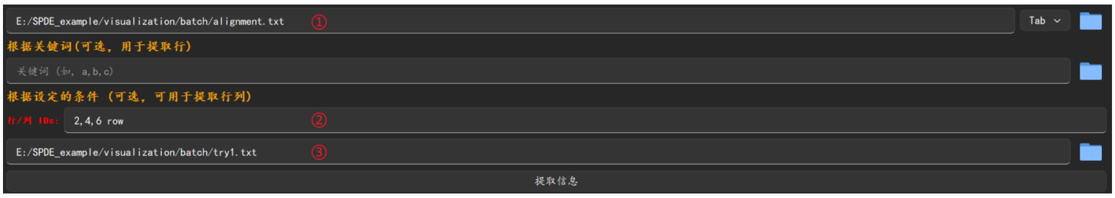
根据设定的条件 (可选，可用于提取行列)

行/列 IDs: 1,3,5 column(提取第1,3和5列) or 1:5 row(提取第1到5行)

E:/SPDE_example/visualization/batch/ex_infor.txt ③

① Put the file in the designated location, then ② enter the keywords (make sure to input them in the English character mode and separate the keywords with commas). After that, ③ set the save location and name the output file. Finally, press ④ to run the function.

2) Extract information by row ID



E:/SPDE_example/visualization/batch/alignment.txt ①

Tab

根据关键词(可选，用于提取行)

关键词 (如, a,b,c)

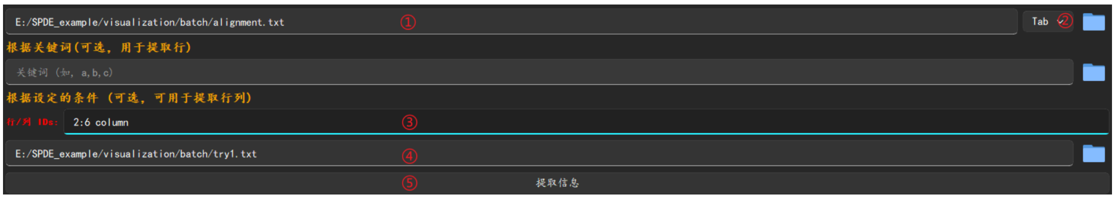
根据设定的条件 (可选，可用于提取行列)

行/列 IDs: 2,4,6 row ②

E:/SPDE_example/visualization/batch/try1.txt ③

① Place the file in the specified location. Then, ② enter the row IDs in the designated area. When dealing with non-consecutive rows, separate them with commas, and for consecutive rows, use colons for separation. Remember to use the "row" keyword to indicate to the software that you are extracting rows, and separate the row IDs from the "row" keyword with a space (for example, "2,4,6 row"). After that, ③ set the save location and name the output file. Finally, press the button to run the function.

3）Extract information according to the column ID



E:/SPDE_example/visualization/batch/alignment.txt ①

Tab ②

根据关键词(可选，用于提取行)

关键词（如，a,b,c)

根据设定的条件（可选，可用于提取行列)

行/列 IDs: 2:6 column ③

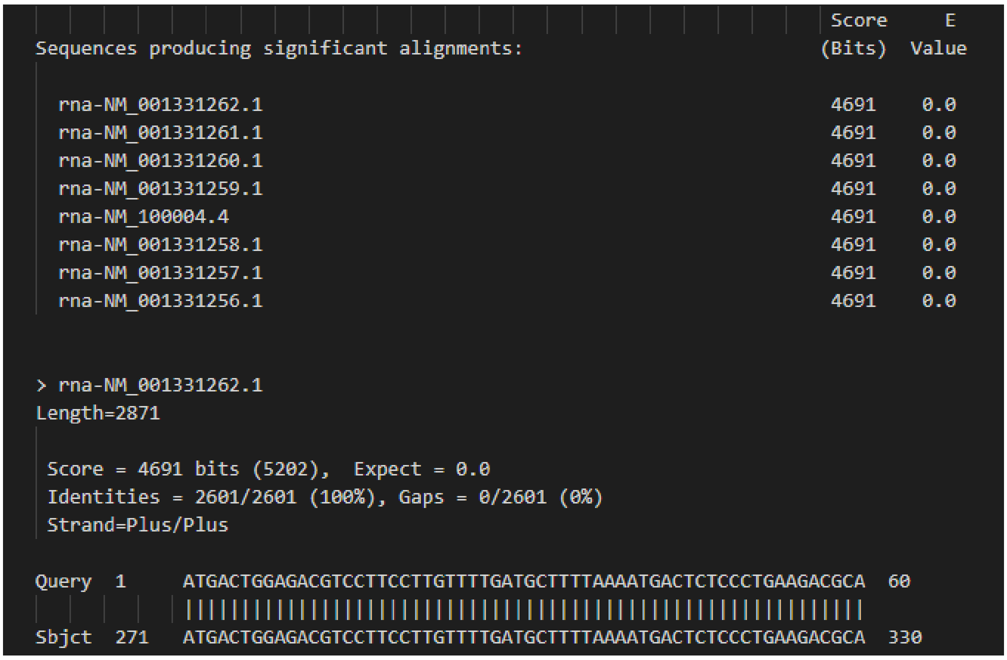
E:/SPDE_example/visualization/batch/try1.txt ④

⑤ 提取信息

First, put the file in ①. Then, a key step is that you need to inform the software what the delimiter for different columns is in ②, which is the most significant difference compared to row extraction. Next, in ③, enter the column ID following the same rules as those for placing the row ID. After that, in ④, save the file and give it a proper name. Finally, press ⑤ to execute the function.

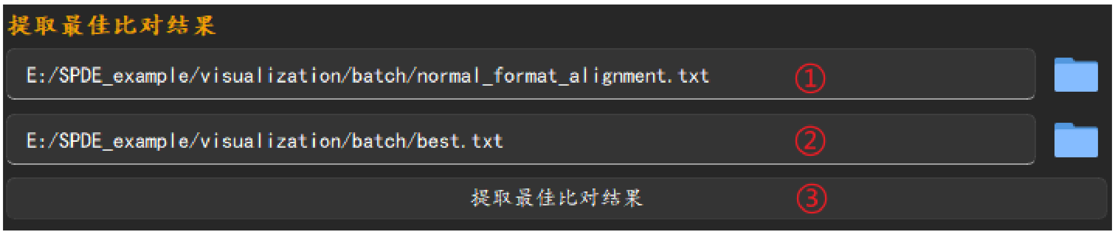
Extract the Optimal Alignment Result

In the normal alignment (in normal format) of NCBI - blast, detailed alignment details can be displayed as follows:



```
                                                                      Score     E
Sequences producing significant alignments:                           (Bits)  Value

  rna-NM_001331262.1                                                   4691    0.0
  rna-NM_001331261.1                                                   4691    0.0
  rna-NM_001331260.1                                                   4691    0.0
  rna-NM_001331259.1                                                   4691    0.0
  rna-NM_100004.4                                                      4691    0.0
  rna-NM_001331258.1                                                   4691    0.0
  rna-NM_001331257.1                                                   4691    0.0
  rna-NM_001331256.1                                                   4691    0.0


> rna-NM_001331262.1
Length=2871

 Score = 4691 bits (5202),  Expect = 0.0
 Identities = 2601/2601 (100%), Gaps = 0/2601 (0%)
 Strand=Plus/Plus

Query  1     ATGACTGGAGACGTCCTTCCTTGTTTTGATGCTTTTAAAATGACTCTCCCTGAAGACGCA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  271   ATGACTGGAGACGTCCTTCCTTGTTTTGATGCTTTTAAAATGACTCTCCCTGAAGACGCA  330
```

However, this format is not conducive to our viewing of the optimal alignment. Therefore, a function for extracting the optimal alignment result has been set up, and the basic settings are as follows:



提取最佳比对结果

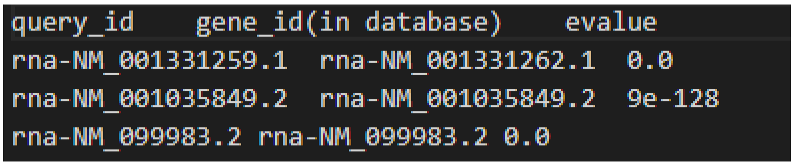
E:/SPDE_example/visualization/batch/normal_format_alignment.txt ①

E:/SPDE_example/visualization/batch/best.txt ②

提取最佳比对结果 ③

First, put the file in ①. Then, save the file and give it a name in ②. Finally, press ③ to execute the function.

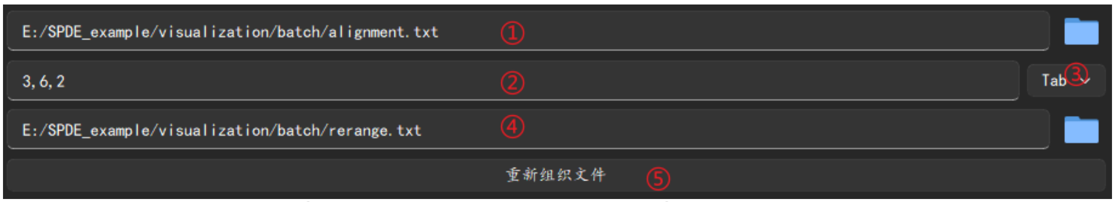
The result format:



```
query_id    gene_id(in database)    evalue
rna-NM_001331259.1  rna-NM_001331262.1  0.0
rna-NM_001035849.2  rna-NM_001035849.2  9e-128
rna-NM_099983.2 rna-NM_099983.2 0.0
```

Reorganize the file

Different tools have specific format requirements for the content of the input file. In order to meet these requirements, it is usually necessary to reorganize the content of the original file. The basic settings are as follows:



E:/SPDE_example/visualization/batch/alignment.txt

3,6,2

Tab

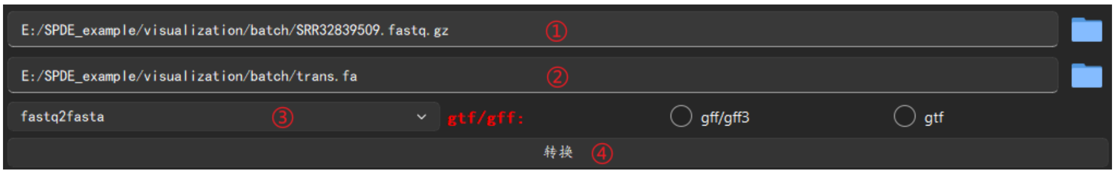
E:/SPDE_example/visualization/batch/rerange.txt

First, put the file in ①. Then, enter the row IDs in ②, separating them with commas. Write the IDs according to how you want to organize the file. Next, in ③, select the delimiter. After that, in ④, set the save location and give the file a name. Finally, press ⑤ to execute the function.

Format Conversion

When conducting bioinformatics analysis, operations such as file format conversion are often involved. In SPDEv3.0, we have designed the format conversion for five types of files, including: converting fastq to fasta, converting tree to nwk, converting sam to paf, converting gbff to gff, and converting gff to gtf. Among them, when converting between gff and gtf, it is necessary to specify the format of the input file. The basic settings are as follows:

1） fastq to fasta



E:/SPDE_example/visualization/batch/SRR32839509.fastq.gz

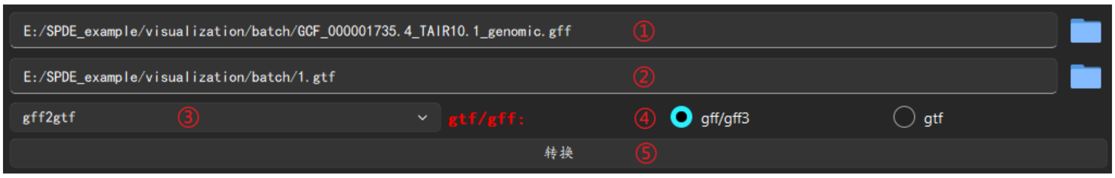
E:/SPDE_example/visualization/batch/trans.fa

fastq2fasta

gtf/gff: ○ gff/gff3 ○ gtf

First, put the file in ①. Then, set the save location and name the file in ②. Next, select the conversion type in ③. Finally, press the button to run the function in ④.

2） gff to gtf (gtf to gff)



E:/SPDE_example/visualization/batch/GCF_000001735.4_TAIR10.1_genomic.gff

E:/SPDE_example/visualization/batch/1.gtf

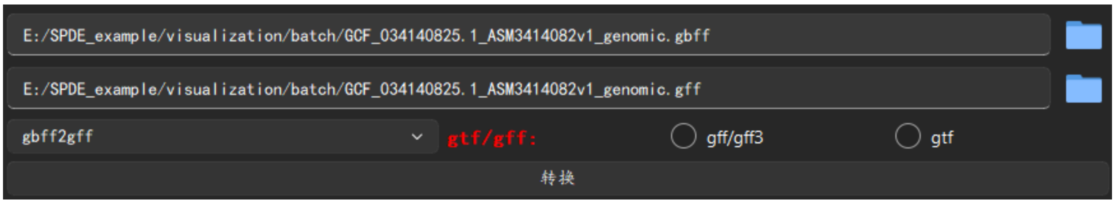
gff2gtf

gtf/gff: ● gff/gff3 ○ gtf

First, put the file in ①. Then, in ②, set the save location and name the file. Next, in ③, select the conversion type. After that, in ④, it is necessary to specify the type of the input file. Finally, press ⑤ to execute the function.

3）GBFF to GFF

The following operations are consistent with the previous ones. I will upload these sample files to my GitHub account. Please make sure to carefully check the file formats before using the functions.



E:/SPDE_example/visualization/batch/GCF_034140825.1_ASM3414082v1_genomic.gbff

E:/SPDE_example/visualization/batch/GCF_034140825.1_ASM3414082v1_genomic.gff

gbff2gff

4）file to nwk

There are two types of tree files that can be converted: one is in dnd format, and the other is in xml format. For the detailed format, please refer to the sample files. The operation is the same as above, and the details are omitted.

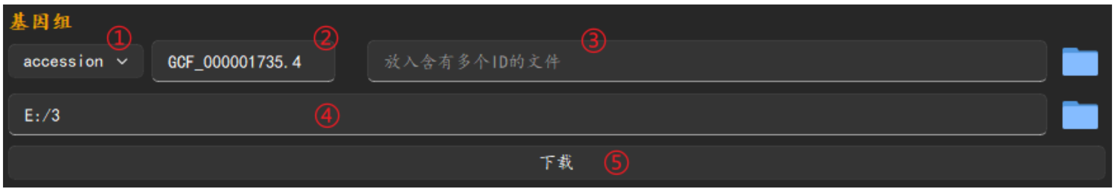
5）sam to paf

The details of Paf format can be found in https://github.com/lh3/miniasm/blob/master/PAF.md. The operation is the same as above, and the details are omitted.

Download files from NCBI

Here, I'm designing a downloading method. Later, I tested it and found that the speed isn't very fast and it depends on your internet connection. However, the innovative aspect is that you can achieve batch downloads by putting IDs into a file. So, please choose flexibly according to your own needs.

1）download genomes from NCBI



accession

GCF_000001735.4

E:/3

First, there are two types of IDs for download: accession and taxon in ①. Then, enter the corresponding IDs in ②. You can also put different IDs into a text file, with one ID per line in ③. The system will then automatically download based on the IDs in the file. Next, set the save location in ④. Finally, press ⑤ to execute the function.

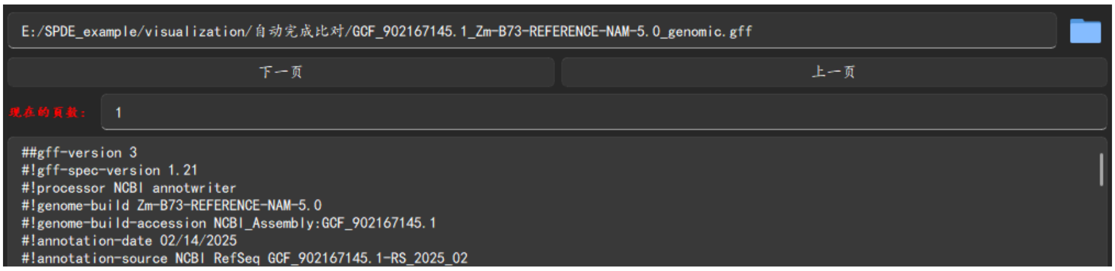
2）download genes from NCBI

The operation is the same as above.

Browse large files

After inputting a large file, you can browse through it by clicking the "Previous Page" or "Next

Page" buttons. The basic settings are as follows:



E:/SPDE_example/visualization/自动完成比对/GCF_902167145.1_Zm-B73-REFERENCE-NAM-5.0_genomic.gff

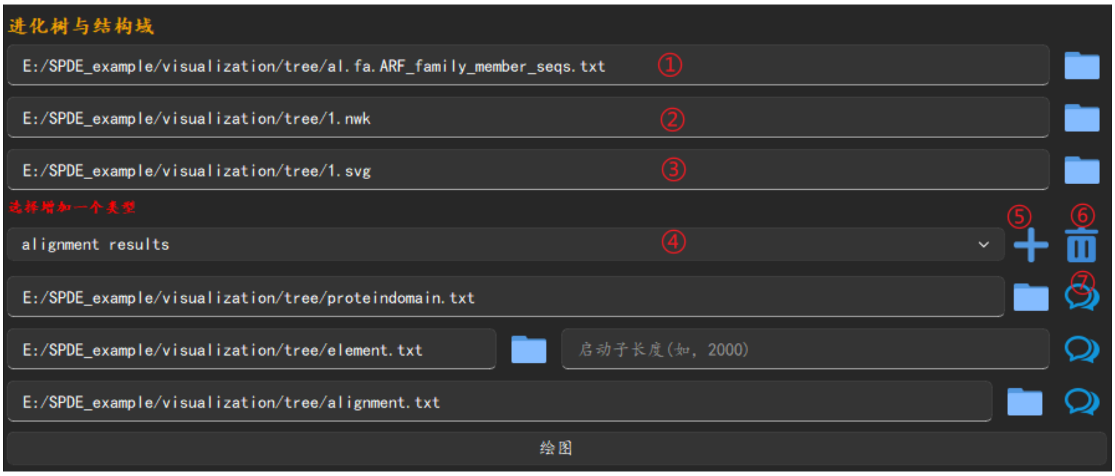
现在的頁數： 1

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build Zm-B73-REFERENCE-NAM-5.0
#!genome-build-accession NCBI_Assembly:GCF_902167145.1
#!annotation-date 02/14/2025
#!annotation-source NCBI RefSeq GCF_902167145.1-RS_2025_02
```

7) Visualization

Visualization of the Phylogenetic Tree

This visualization is used to display the phylogenetic tree together with other elements, including protein domains, promoter elements, alignment results, motifs, single elements, etc., all in a single graph. The basic settings are as follows:



进化树与结构域

E:/SPDE_example/visualization/tree/al.fa.ARF_family_member_seqs.txt

E:/SPDE_example/visualization/tree/1.nwk

E:/SPDE_example/visualization/tree/1.svg

选择增加一个类型

alignment results

E:/SPDE_example/visualization/tree/proteindomain.txt

E:/SPDE_example/visualization/tree/element.txt

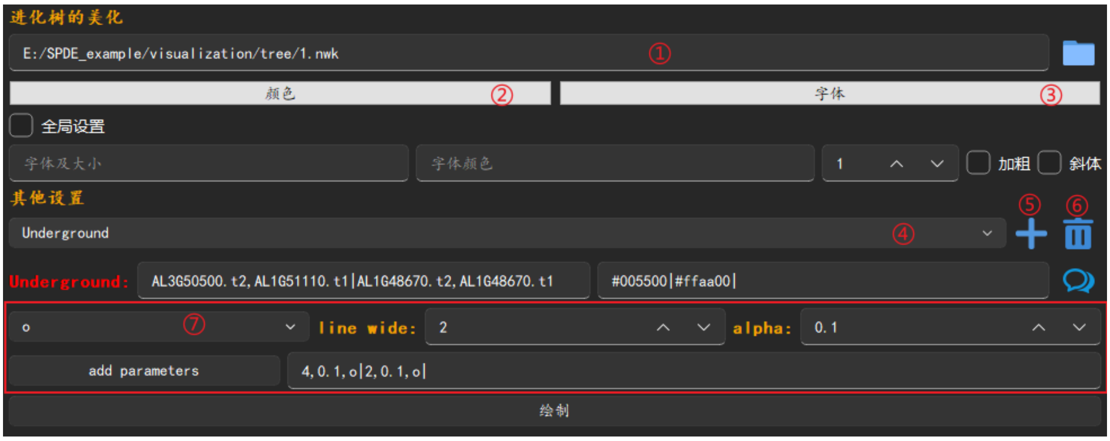
启动子长度(如, 2000)

E:/SPDE_example/visualization/tree/alignment.txt

First, prepare the sequence file of the gene family in ①. Then, get the phylogenetic tree in nwk format in ② (you can refer to my blog at https://www.jianshu.com/p/39f07b6f0435 and use the MEGA software to obtain it). Next, select the save location and name the file in ③. In ④, there is a dropdown box where you can choose the type of visualization (such as protein domains). Click ⑤ to expand the corresponding dialog box and add files in the correct format. You can click ⑦ to view the format. Click ⑥ to delete the unnecessary dialog box (note that only the last dialog box can be deleted). After completing the settings, click "Draw" to execute the function. As shown in the figure above, you can add different information to the phylogenetic tree.

Beautification of the Phylogenetic Tree

The input phylogenetic tree is still in nwk format, and the specific acquisition method is as described above. The beautification elements that can be added to the phylogenetic tree include background, label, line, marker, and heatmap. There are quite a number of parameters set, and it's impossible to display them all. Only the basic settings are shown as follows:



进化树的美化

E:/SPDE_example/visualization/tree/1.nwk ①

颜色 ② 字体 ③

□ 全局设置

字体及大小 字体颜色 1 □ 加粗 □ 斜体

其他设置

Underground ④ ⑤ ⑥

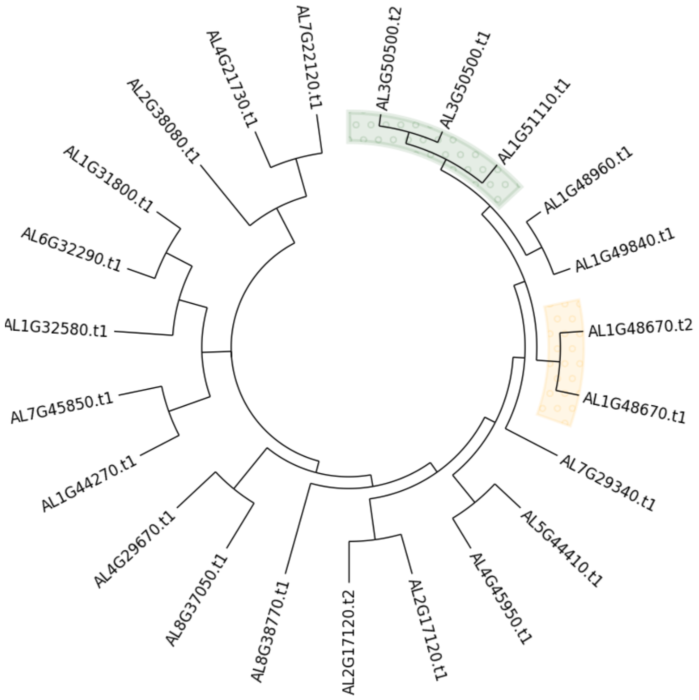
Underground: AL3G50500.t2,AL1G51110.t1|AL1G48670.t2,AL1G48670.t1 #005500|#ffaa00|

o ⑦ line wide: 2 alpha: 0.1

add parameters 4,0.1,o|2,0.1,o|

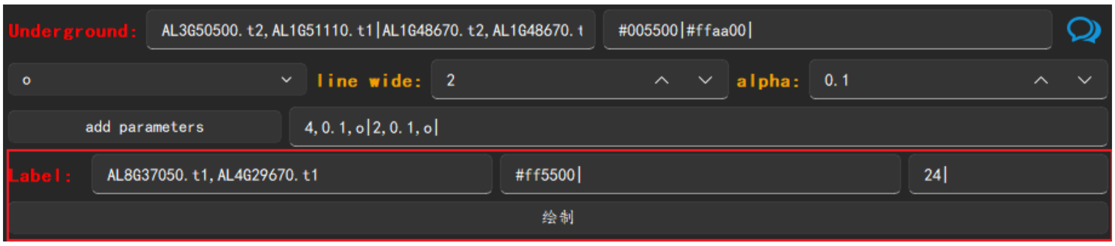
绘制

First, input the file in nwk format in ①. Then, you can select colors through this button in ②, and choose the font and its size through another button in ③. In ④, mainly bring up the beautification layer and beautify it by setting some parameters. After selecting the type, click ⑤ to add. ⑥ is used to delete the last added form. When you want to add symbol annotations to the background, you can select the red - framed area where ⑦ is located for settings. As shown in the figure above, it can create the beautification effect as shown in the following figure:



A few tips: After bringing up the beautification layer, use ② to set the colors, and the selected colors will be automatically added. A basic requirement is that the number of color groups you set should be the same as the number of ID groups (each group is separated by the '|' symbol). Of course, if you don't want to set colors for some groups, you can simply set them to black. A similar requirement applies to the symbol - adding area (that is, the red - framed area).

settings for label:



Underground: AL3G50500.t2,AL1G51110.t1|AL1G48670.t2,AL1G48670.t #005500|#ffaa00|

o line wide: 2 alpha: 0.1

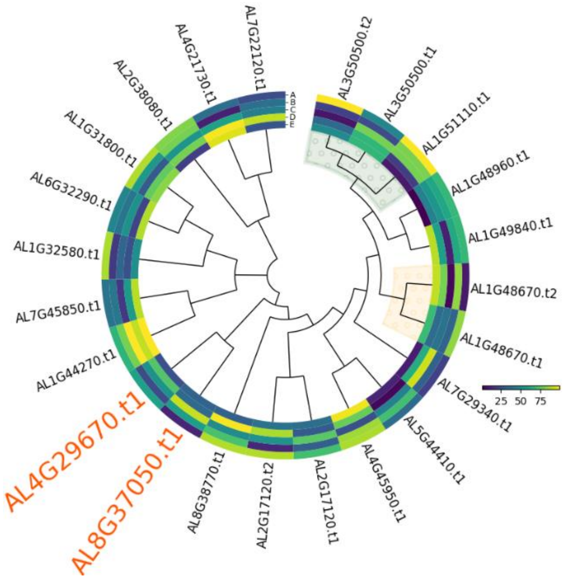
add parameters 4,0.1,o|2,0.1,o|

Label: AL8G37050.t1,AL4G29670.t1 #ff5500| 24|

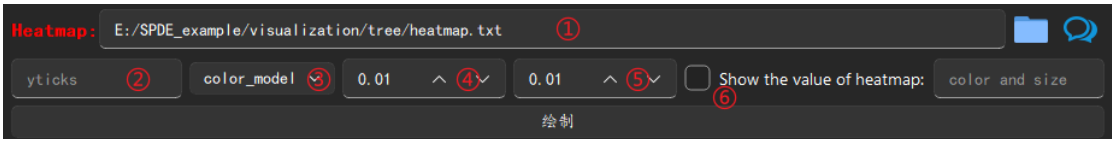
The font can be set in the global section. However, you need to check the box at "Global Settings". At the Label section, you can set the font color and size separately, and both can be automatically added via the "Color" and "Font" buttons.

For other beautification elements like lines and markers, there aren't many things to specifically note. Since there are numerous setting parameters, I won't show them all to you one by one. When using it in practice, you can refer to the sample files, fully understand each parameter, and then replace them with your own data for beautification.

There is also an operation related to adding a heatmap in the phylogenetic tree beautification, which can be done as shown in the following figure:



Regarding the input file of the data, you are requested to carefully check the sample file before performing the operation. One requirement is that the number of columns in the file should correspond to the number of gene IDs in the database, that is, the number of columns should be the same as the number of IDs. The basic settings are as follows:



Heatmap: E:/SPDE_example/visualization/tree/heatmap.txt ①

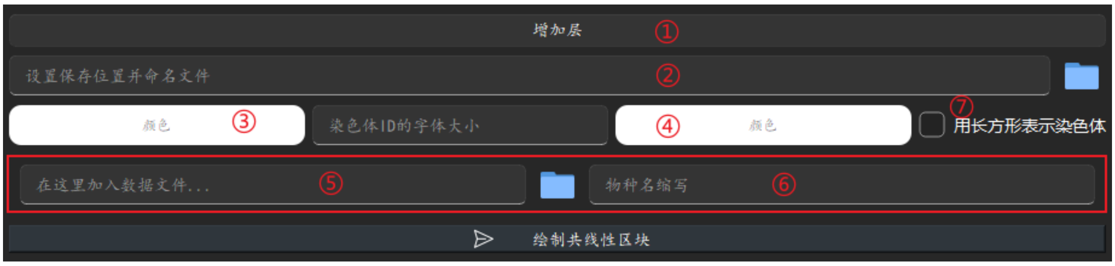
yticks ② color_model ③ 0.01 ④ 0.01 ⑤ □ Show the value of heatmap: ⑥ color and size

First, input the data in ①. Then, label the data of y (generally, there is no need for setting) in

②. Next, select the color mode in ③. After that, set the position of the colorbar on the x-axis in ④ and set the position of the colorbar on the y-axis in ⑤. If you want to add numbers to the heatmap, you can check ⑥, and the color and size of the numbers can be set in the subsequent dialog box.

Collinearity

The input for this module can be obtained from the genome module. The basic settings are as follows:

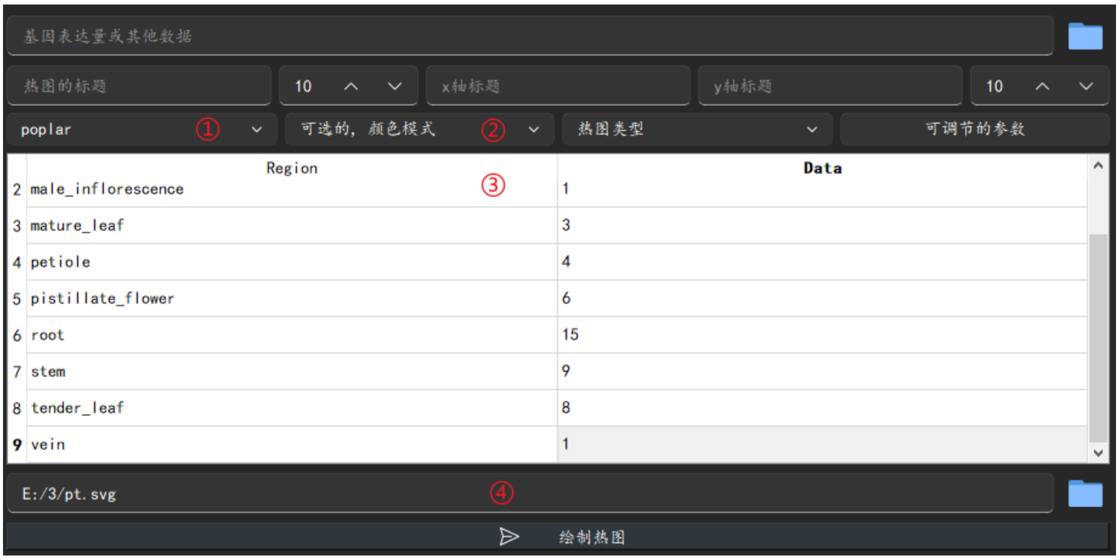


This module supports the collinearity display of multiple species (more than two). Click ① to bring up the settings dialog box (the red box in the figure above); ② Set the save location and name the file; SDPE will by default assign different colors to each chromosome until the color is set at ③; The color at ④ refers to the color of the collinearity connection line; Click ⑤ to add the collinearity analysis file; ⑥ Here, you need to set the abbreviations of the species names. According to the requirements of scientific research papers, you should fill in the abbreviations of the Latin scientific names here. For example, if I want to display the collinearity among Arabidopsis thaliana (abbreviation: At), Oryza sativa (Os), and Populus trichocarpa (Pt), when generating the collinearity file, I should first compare At with Os, and then compare Os with Pt. So, two layers need to be added here. The abbreviations of the species names in the first layer should be At and Os; the abbreviations in the second layer should be Os and Pt. In other words, the order of the species collinearity files and during the plotting should be consistent and continuous; You can check ⑦ when you want to use rectangles to represent the chromosomes. After completing the settings, click the button below to execute the function.

Heatmap

1) Set up the visualization of species expression and map data

This function is mainly used to display the data distribution trends on different plant organs, subcellular structures, as well as on maps (including the map of China and the world map). Take the display of the data distribution of various plant organs as an example:



基因表达量或其他数据

热图的标题 | 10 | x轴标题 | y轴标题 | 10

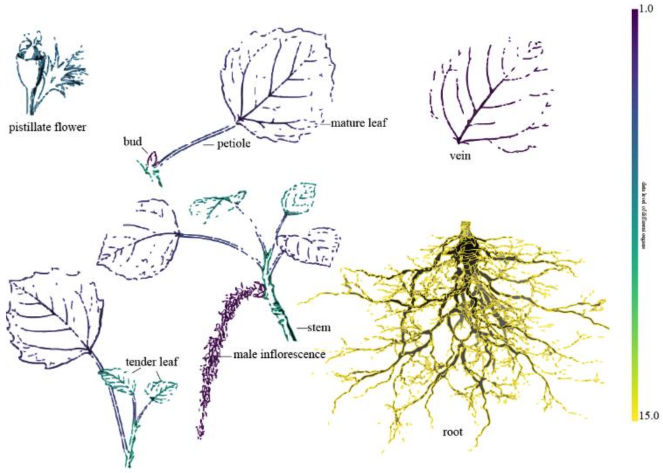
poplar ① | 可选的，颜色模式 ② | 热图类型 | 可调节的参数

	Region ③	Data
2	male_inflorescence	1
3	mature_leaf	3
4	petiole	4
5	pistillate_flower	6
6	root	15
7	stem	9
8	tender_leaf	8
9	vein	1

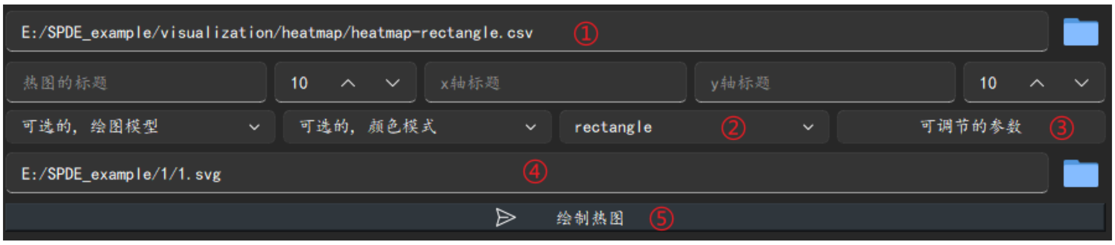
E:/3/pt.svg ④

绘制热图

Select the species you need at ①. Once selected, a table will automatically appear. Add data in the "data" section of ③ in this table. (Note that this is just to show the data distribution trend, so there won't be multiple biological treatment formats. Taking the expression level as an example, after you have organized the data, you can put an average value in the corresponding organ. If there is no data, just leave it blank or write 0). ② is the color mode that can be selected. ④ Set the save location and name the result as a file. Click the button to execute the function. With the above settings, it indicates that the gene expression is mainly concentrated in the roots of Populus trichocarpa. Therefore, after drawing, you will get the visualization result as shown in the following figure.



2) other format of heatmap



E:/SPDE_example/visualization/heatmap/heatmap-rectangle.csv ①

热图的标题 | 10 | x轴标题 | y轴标题 | 10

可选的，绘图模型 | 可选的，颜色模式 | rectangle ② | 可调节的参数 ③

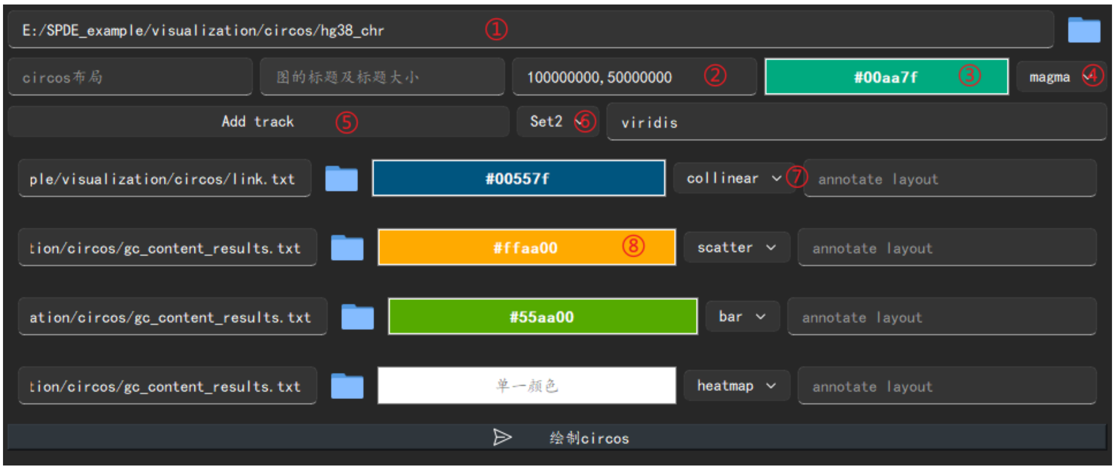
E:/SPDE_example/1/1.svg ④

绘制热图 ⑤

Enter data at ①. (I've prepared a sample file for each plotting type for you. Please organize your data strictly following the format of the sample files.) As for other parameters, you can set them according to the prompts, so I won't elaborate on them here. Select the plotting type at ②. If you want to access more adjustment parameters, click ③. Set the save location and name the file at ④. Then click ⑤ to execute the function.

CIRCOS

Basic settings as follows:



E:/SPDE_example/visualization/circos/hg38_chr

circos布局

图的标题及标题大小

100000000, 50000000

#00aa7f

magma

Set2

viridis

ple/visualization/circos/link.txt | #00557f | collinear | annotate layout

tion/circos/gc_content_results.txt | #ffaa00 | scatter | annotate layout

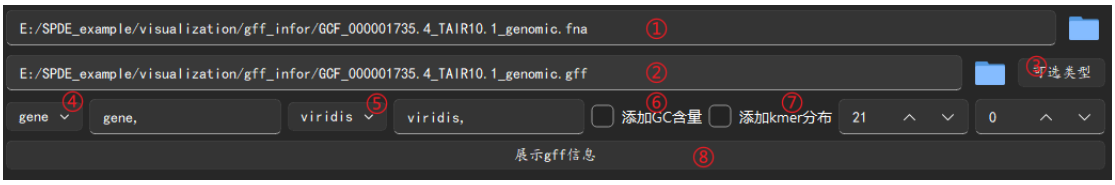
ation/circos/gc_content_results.txt | #55aa00 | bar | annotate layout

tion/circos/gc_content_results.txt | 单一颜色 | heatmap | annotate layout

Place the genome sequence file at ①. SPDE will generate a BED file based on the sequence file. The BED file is the fundamental file for plotting. So, if the final plot isn't generated, you need to keep the BED file and always ensure it's in the same folder as the genome sequence. In other words, after the BED file is generated for the first time, you still input the genome sequence file at ①. But since the BED file already exists, the program will automatically recognize it and won't generate the BED file again. The settings at ② are used to generate the length scale for the chromosomes. The default values are 1 Mb for the main scale and 0.5 Mb for the unit length. If the scales in the generated plot are too dense, you can consider increasing these two values to make the scales sparser. Set a single color for the chromosomes at ③, or set different colors for them at ④. Circos plots are composed of multiple layers. You can add a layer by clicking ⑤. There are two types of input files here: one is used to display collinearity (i.e., the connections between chromosomes), and the other is a data file (such as GC - content data). When generating these two types of files, you need to sort the chromosomes from the beginning to the end and then put them into this module for visualization. Click ⑦ to select the visualization type for this layer. When it's not a heatmap, you can set the color by clicking ⑧. If it's a heatmap, you can select a color model at ⑥ to draw the heatmap. For the formats of the two types of data files, please refer to the sample data.

Display GFF information

The GFF file itself contains a great deal of information. SPDEv3.0 has set up corresponding functions to display the information elements in the GFF file. The basic settings are as follows:



E:/SPDE_example/visualization/gff_infor/GCF_000001735.4_TAIR10.1_genomic.fna

E:/SPDE_example/visualization/gff_infor/GCF_000001735.4_TAIR10.1_genomic.gff

gene

gene,

viridis

viridis,

□ 添加GC含量 □ 添加kmer分布 21 0

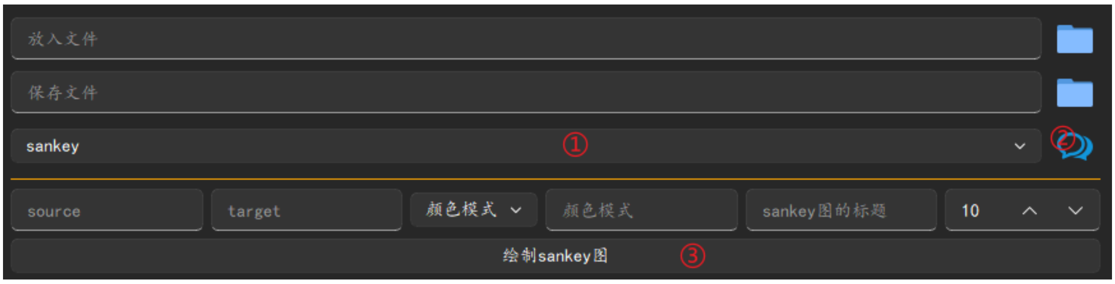
Place the genome sequence file at ① and the GFF file at ②. Then click ③ to analyze which elements are contained in the GFF file. After the analysis is completed, these information types will be loaded into ④. Browse ④ to decide which elements to visualize. At ⑤, you can set the data - related elements and display the GC content (⑥) or the distribution density of k - mers (⑦) in the form of a heatmap. Then click ⑧ to run the function.

Statistical plotting

This module has a total of settings for bar charts, box plots, double box plots, geographical maps, geographical maps of China, principal component analysis (2D), principal component analysis (3D), scatter plots (bivariate), scatter plots (multivariate), stacked bar charts, violin plots (multivariate), and violin plots (univariate). The specific usage methods are the same as those of the Breeding Science Calculation module, so detailed introductions will not be given here.

Other visualization functions

In addition to the commonly used plotting functions mentioned above, SPDEv3.0 also has other plotting functions, such as Sankey diagrams, motif logos, network diagrams, conserved motifs, promoter motifs, and word clouds. The basic settings are as follows:



放入文件

保存文件

sankey

source

target

颜色模式

颜色模式

sankey图的标题

10

After placing the file and setting the save location, you can click ① to select the plotting type. When you're not sure about the format of the input file, you can click ② to view the basic format (of course, sample files will also be provided for you). For other settings, floating prompts are available. When you hover your mouse over a certain component, a prompt message will appear. Click ③ to execute the function.