
“实用生物信息技术”课程小组讨论总结报告

组：G6 次：5 组长：冯睿桥 执笔：冯睿桥、凌子涵

一、 时间

2026年5月8日、5月10日

二、 方式

线上线下交流与讨论

三、 主题

TBtools 课后练习及近期学习内容交流与互助

四、 内容

- 1 Tbtools 功能解析
- 2 Tbtools 课外练习讨论分析
- 3 Tbtools 使用体验与心得

1 Tbtools 功能解析

(1) 序列处理与提取 (Sequence Toolkit)

这是 TBtools 最基础也最强大的功能之一，涵盖了几乎所有常见的 FASTA/GFF3 文件操作：

序列提取与截取：支持基于 ID 批量提取完整序列，或根据染色体坐标（如 Chr1 10000 20000）截取特定区间序列。

格式转换与统计：支持 FASTA 与表格格式互转、序列文件分割与合并、序列长度/碱基统计等。

CDS 与蛋白序列处理：可一键从基因组和 GFF3 文件中提取 CDS、mRNA、启动子（上游 2kb 等）、内含子/外显子结构，以及将 CDS 翻译成蛋白序列。

ID 管理：支持序列 ID 的批量重命名、添加前缀、简化（去除多余注释信息）等操作。

(2) 基因家族与进化分析 (Gene Family Analysis)

TBtools 为基因家族研究提供了高度集成的工作流，极大压缩了分析周期：

同源基因鉴定：内置或封装了 OrthoFinder、MCScanX、DIAMOND 等工具，可快速进行多物种同源聚类，区分直系同源与旁系同源基因。

进化树构建与美化：原生嵌入 MAFFT（多序列比对）和 IQ-TREE（最大似然法建树），并支持进化树的交互式美化与注释。

基因结构与共线性分析：可一键绘制基因内含子/外显子结构图、染色体共线性区块图，并计算同义替换率（Ks）分布。

(3) 强大的可视化绘图 (Graphics)

无需 R 语言或 Python 脚本，即可生成高质量的发表级图表：

染色体基因分布图：可视化展示基因在染色体上的物理位置、密度及成簇情况（类似“棒棒糖”图）。

热图 (Heatmap)：支持导入 RNA-seq 表达矩阵，快速绘制聚类热图，并可结合进化树进行展示。

基因组圈图 (Circos)：通过整合染色体长度、GC 含量、基因密度等数据，绘制美观的基因组 Circos 圈图。

其他图表：支持绘制 Venn 图（韦恩图）、GO/KEGG 富集分析气泡图/柱状图等。

(4) 功能注释与富集分析 (GO & KEGG)

支持导入 eggNOG-mapper 等工具的注释结果。

提供 GO 和 KEGG 富集分析功能，只需输入基因列表即可一键生成富集结果及可视化图表。

(5) 创新模块：Simple Genome Hub (简易基因组中心)

这是 TBtools 较新的“王炸”功能，极大提升了数据管理效率：

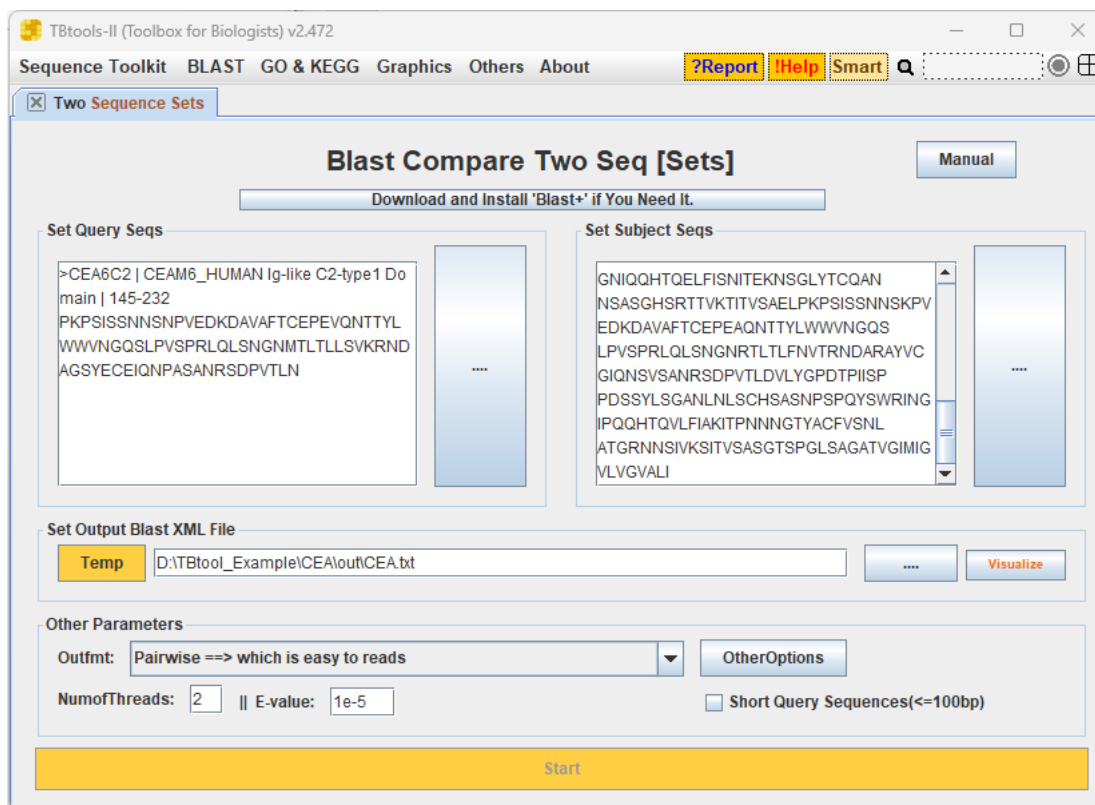
数据一站式管理：可将 CDS、蛋白序列、基因注释、RNA-seq 表达矩阵等数据统一拖入管理，像整理文件夹一样清晰。

基因信息一键查：输入基因 ID，即可自动打包获取其序列全家桶（转录本、CDS、蛋白）及 GO/KEGG 注释信息。

交互式 BLAST：内置可视化比对界面，粘贴序列即可秒出 BLAST 结果，并能直接生成用于 PPT 展示的素材。

2 Tbtools 课外练习讨论分析

(1) 癌胚抗原家族成员双序列比对



结果分析

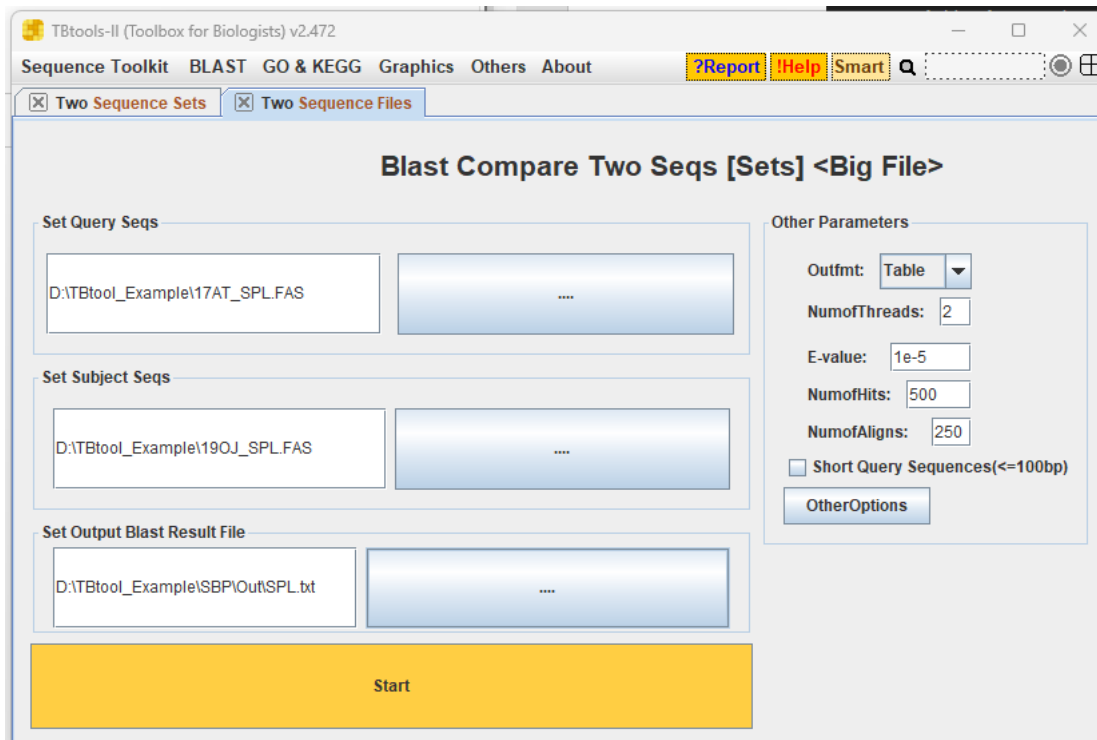
CEAM6 的恒定结构域 A 与 CEAM5 的 501-588 位点结构域相似性最高。该结构域拥有最高的 Bit Score (153) 和 Identity (85%)，且 E-value 最低 ($2e-49$)，表明其在进化上最为保守，亲缘关系最近。TBtools 的 BLAST 双序列比对与 CNCB BLAST 比较，TBtools 凭借其本地化、免配置环境的特点，能够更高效地完成 CEAM6 与 CEAM5 的精细比对，明确指出了 501-588 位点结构域为最高相似区域。相比 CNGB 的在线 BLAST，TBtools 在处理特定的小规模双序列比对及私有数据挖掘时，具有更高的操作便捷性和数据安全性。

```

71
72
73 Lambda      K          H          a          alpha
74 | 0.307      0.124      0.369      0.792      4.96
75
76 Gapped
77 Lambda      K          H          a          alpha  sigma
78 | 0.267      0.0410     0.140      1.90       42.6     43.6
79
80 Effective search space used: 44135
81
82
83 Database: TBtoolsBlastSubject42882488973265543fa
84   Posted date: May 15, 2026 5:04 PM
85   Number of letters in database: 702
86   Number of sequences in database: 1
87
88
89
90 Matrix: BLOSUM62
91 Gap Penalties: Existence: 11, Extension: 1
92 Neighboring words threshold: 11
93 Window for multiple hits: 40
94

```

(2) 拟南芥和水稻 SBP 转录因子家族成员比对



水稻中拟南芥 SPL7_ARATH 的直系同源基因。

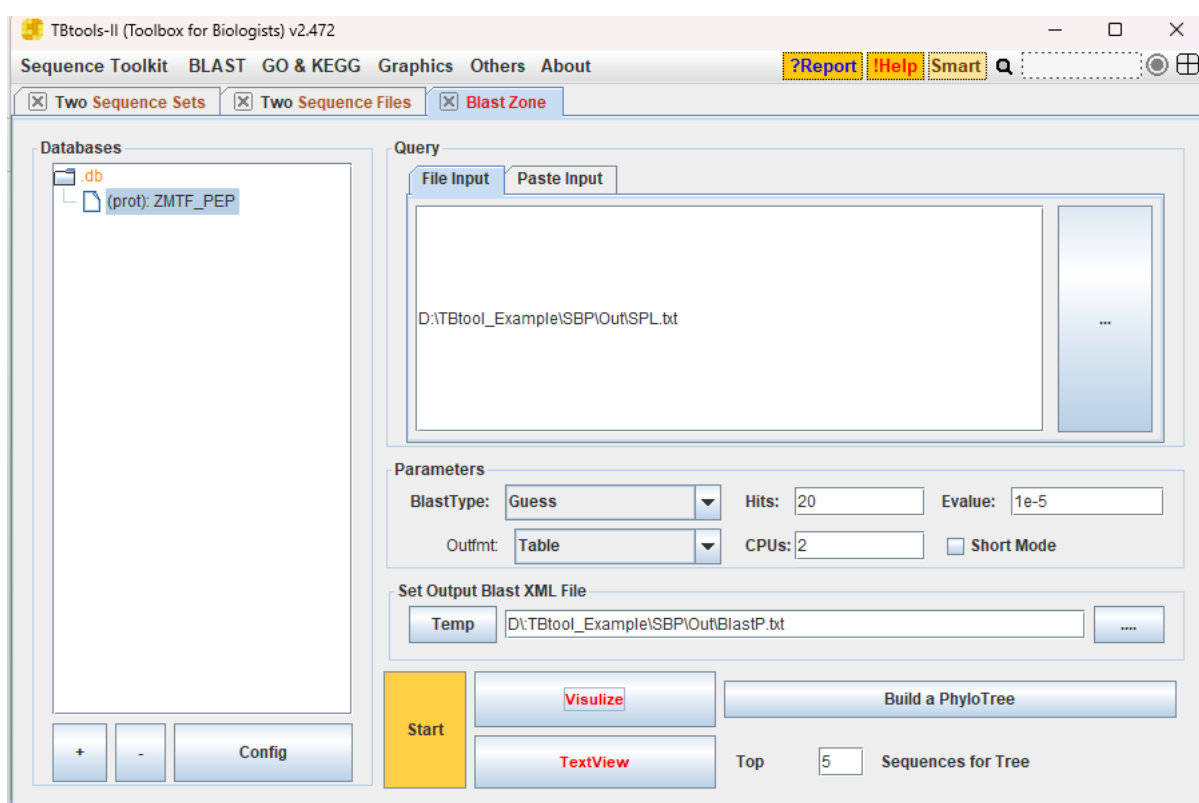
```
# BLASTP 2.12.0+
# Query: SPL7_ARATH
# Database: 19OJ_SPL.FAS.TBtoolsDB
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 21 hits found
SPL7_ARATHSPL9_ORYSJ 37.170 834 436 19 32 797 23 836 1.64e-169 502
SPL7_ARATHSPL1_ORYSJ 29.530 447 266 11 138 550 107 538 2.55e-50 181
SPL7_ARATHSPL15_ORYSJ 28.632 234 163 2 320 553 599 828 4.11e-32 124
SPL7_ARATHSPL15_ORYSJ 49.367 79 39 1 138 215 187 265 2.15e-22 93.2
SPL7_ARATHSPL6_ORYSJ 29.675 246 156 4 307 548 432 664 1.20e-26 106
SPL7_ARATHSPL6_ORYSJ 47.664 107 53 2 125 228 138 244 9.88e-25 100
SPL7_ARATHSPL14_ORYSJ 57.333 75 32 0 137 211 103 177 2.34e-25 100
SPL7_ARATHSPL2_ORYSJ 42.453 106 53 3 127 224 77 182 1.73e-23 94.4
SPL7_ARATHSPL16_ORYSJ 33.728 169 93 4 127 288 102 258 3.61e-23 94.0
SPL7_ARATHSPL7_ORYSJ 50.649 77 38 0 135 211 105 181 5.42e-23 92.0
SPL7_ARATHSPL10_ORYSJ 54.286 70 32 0 137 206 180 249 6.95e-23 92.8
SPL7_ARATHSPL18_ORYSJ 47.253 91 43 2 138 224 115 204 1.01e-22 92.8
SPL7_ARATHSPL8_ORYSJ 40.594 101 59 1 137 236 184 284 1.36e-22 91.7
SPL7_ARATHSPL19_ORYSJ 50.000 76 37 1 138 213 93 167 2.39e-22 90.1
SPL7_ARATHSPL4_ORYSJ 43.678 87 49 0 137 223 67 153 2.95e-22 87.8
SPL7_ARATHSPL5_ORYSJ 53.086 81 37 1 137 217 206 285 4.20e-22 90.5
SPL7_ARATHSPL17_ORYSJ 52.564 78 36 1 137 214 73 149 7.71e-22 89.0
SPL7_ARATHSPL3_ORYSJ 50.649 77 37 1 137 213 181 256 5.91e-21 87.0
SPL7_ARATHSPL12_ORYSJ 48.684 76 38 1 138 213 180 254 9.36e-21 86.7
SPL7_ARATHSPL13_ORYSJ 46.154 78 42 0 138 215 110 187 5.08e-20 80.1
SPL7_ARATHSPL11_ORYSJ 44.304 79 44 0 137 215 66 144 2.13e-19 80.9
```

在 BLASTP 分析结果中, SPL7_ORYSJ 与查询序列 SPL7_ARATH 的比对具有最高的特异性, 是其在水稻中的直系同源基因。

查询基因 (Query)	候选基因 (Subject)	同源类型	一致性 (%)	比特分 值 (Bit Score)
SPL7_ARATH	SPL7_ORYSJ	直系同源 (Ortholog)	50.649	92.0

Btools 的 BLAST 双序列比对操作具有“本地化、文件化”的显著特点。与在线 BLAST 需要繁琐的网页粘贴不同，TBtools 采用文件直读的方式，用户只需指定本地 FASTA 文件路径即可启动分析。这种设计不仅避免了长序列粘贴时的格式错误，还利用本地算力实现了离线隐私计算，特别适合在撰写论文（如 SPL 基因家族分析）时进行大批量、高频率的序列比对任务。

(3) 玉米 SBP 转录因子数据库构建和搜索



```

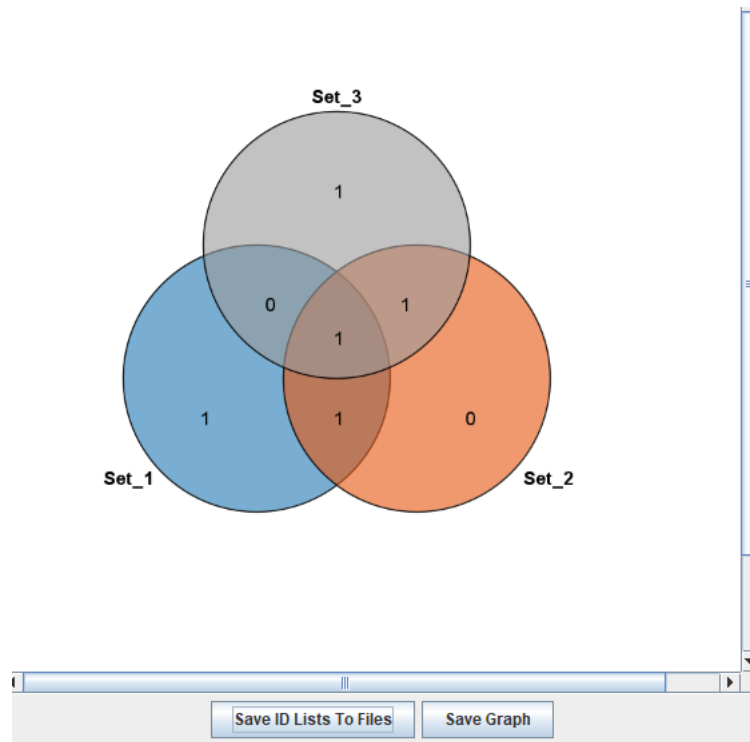
# BLASTP 2.12.0+
# Query: AT2G33810.1 Arabidopsis SBP family transcription factor, protein
sequence, 131 AA
# Database: zmtf_pep
# Fields: query acc.ver, subject acc.ver, % identity, alignment length,
mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit
score
# 4 hits found
AT2G33810.1 PTZm00608.1 45.082 122 67 0 10 131 127 248
9.03e-34 115
AT2G33810.1 PTZm00605.1 45.082 122 67 0 10 131 100 221
1.32e-33 114
AT2G33810.1 PTZm00606.1 56.757 37 16 0 54 90 180 216
1.36e-12 57.0
AT2G33810.1 PTZm00607.1 30.000 90 63 0 10 99 127 216
9.62e-12 55.1
# BLAST processed 1 queries

```

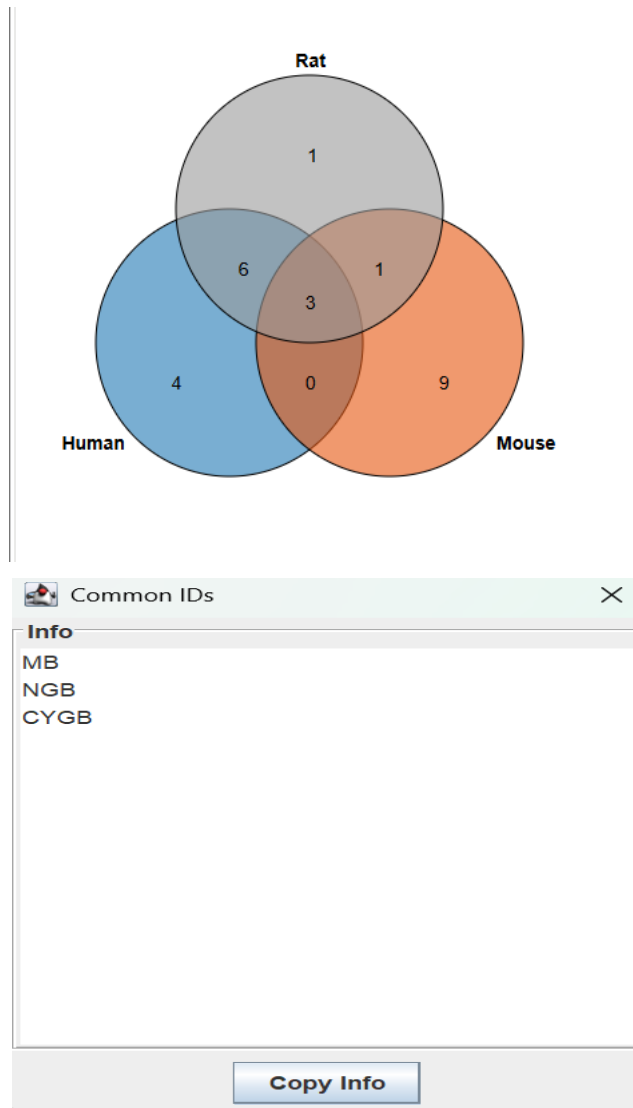
TBtools 自建本地数据库纯图形界面、零代码，操作简单，可以根据自己的需要构建数据库从而快速筛选。NCBI Blast 的在线比对基于其巨大的数据，需要通过层层筛选。建立 NCBI 本地库需要依赖大量命令行，比较复杂。

(4) 韦恩图 Venn Diagram 分析

韦恩图 (Venn Diagram) 是一种用于展示集合之间逻辑关系的图形化工具。它通过重叠的圆形 (或其他形状) 来直观地表示不同组数据之间的共有与独有部分。韦恩图使用简单的封闭曲线 (通常为圆形) 来展示集合。每个圆代表一个集合 (或一组数据)，圆的大小通常不表示数值大小 (除非是面积成比例的变体)，而是代表一个整体类别。重叠区域 (交集) 代表同时属于两个或多个集合的元素。非重叠区域 (独有部分) 代表仅属于该特定集合的元素。



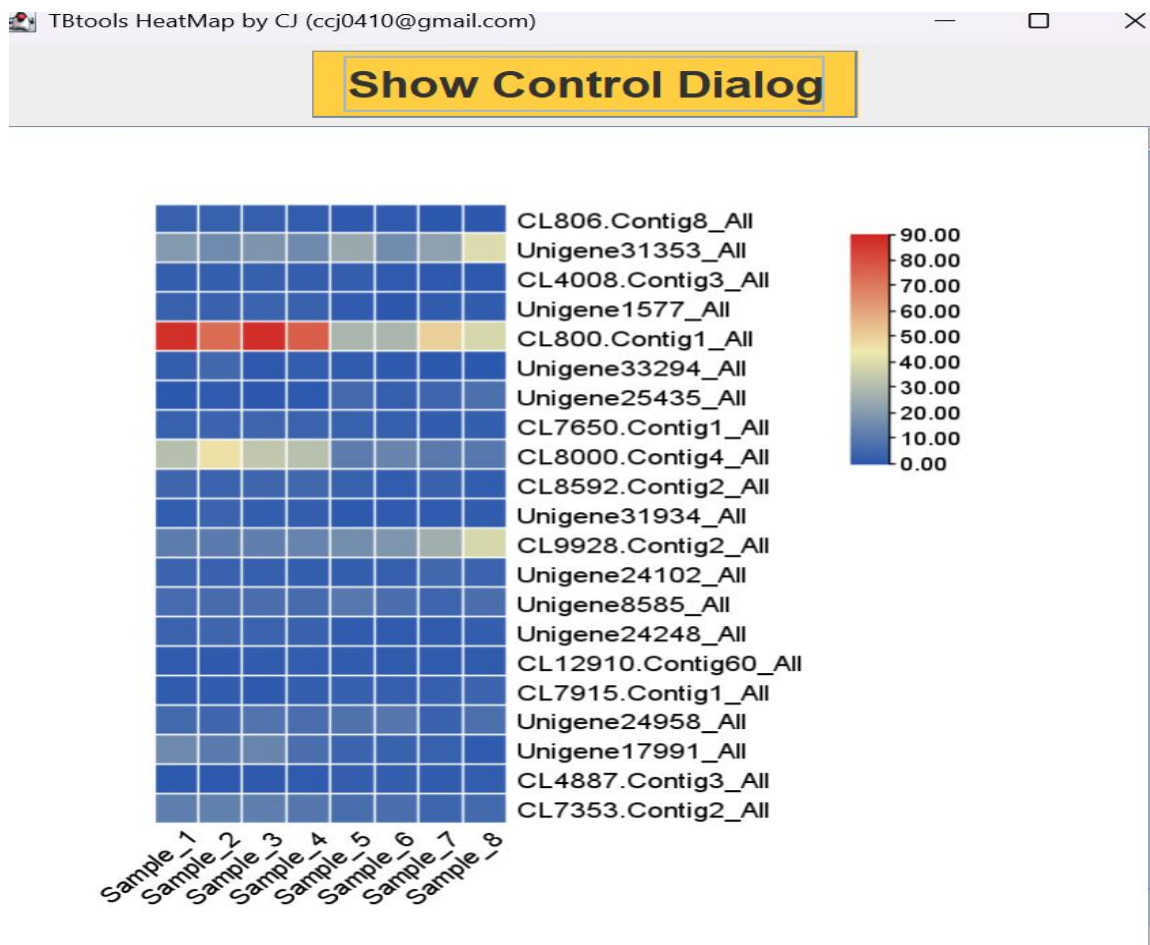
结果分析: Set1, Set2, Set3 三个数据集共有 1 个相同元素 C; Set1 和 Set2 两个数据集共有 1 个相同元素 B; Set1 和 Set3 两个数据集共有 1 个相同元素。双击图中的数字可以看到其所代表的基因。



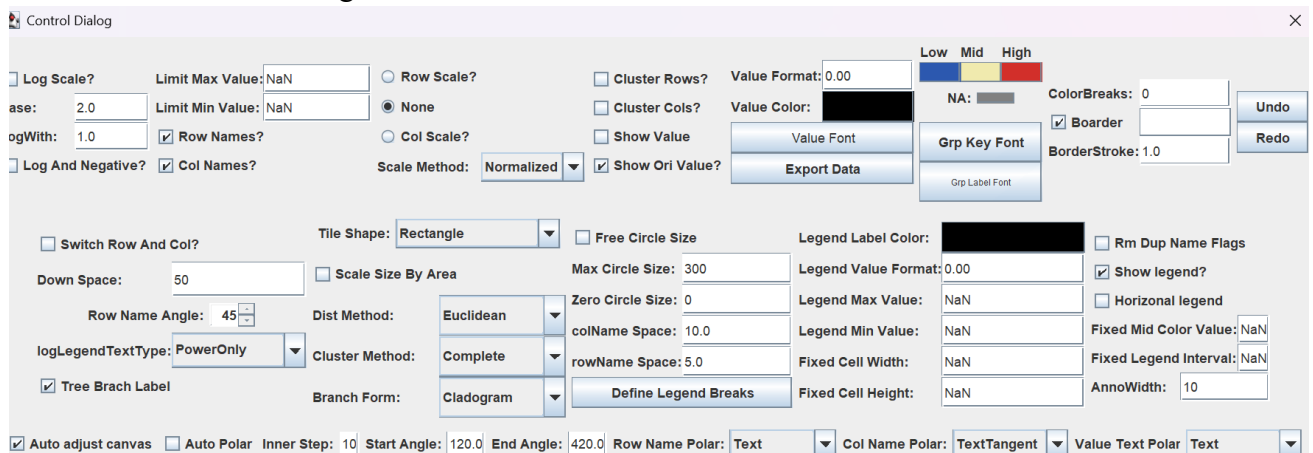
人、小鼠、大鼠珠蛋白基因家族。结果分析：三个物种均有肌红蛋白 MB，胞红蛋白 CYGB 和脑红蛋白 NGB；人和大鼠另有 6 个同名基因；小鼠和大鼠另有 1 个同名基因。

(5) 热图 Heatmap 分析

点击 Tbtools 的 Graphics 下的 Heatmap Illustrator，选择 Heatmap
Tbtools 自带的热图演示实例



点击 Show Control Dialog 可以调整热图的显示方式。



热图是一种以颜色变化（如从深蓝到红色）来直观展示二维数据矩阵中数值大小的图形。在生物信息学中，它通常用于展示基因表达量（如 RNA-seq、qPCR、芯片数据）在不同样本或处理条件下的变化模式。

行 (Rows): 通常代表基因 (Genes) 或蛋白。

列 (Columns): 通常代表样本 (Samples)、实验处理组或时间点。

颜色 (Color): 代表数值的大小（如基因表达丰度）。

为了让不同基因（有的本身表达量极高，有的极低）能在同一张图上通过颜色公平地展示变化趋势，热图分析中最常用的单位是 Z-score（标准分数）。

Z-score 表示某个数值距离平均值有多少个标准差。

$$Z = \frac{x - \mu}{\sigma}$$

Z-score = 0: 该基因在该样本中的表达量等于平均水平（通常显示为白色或中性色）。

Z-score > 0（如 +2）: 表达量显著高于平均水平（通常显示为红色，代表上调）。

Z-score < 0（如 -2）: 表达量显著低于平均水平（通常显示为蓝色/绿色，代表下调）。

使用 Z-score 作为单位，可以抹平基因间绝对表达量的巨大差异，专注于展示相对变化趋势。

颜色映射单位（Color Scale）是热图图例（Legend）上展示的单位，用于将数值映射为视觉颜色。连续型色阶如“蓝-白-红”。蓝色代表低值（Low），白色代表中间值（Medium），红色代表高值（High）。对数刻度（Log Scale）由于基因表达量跨度极大（从几到几万），热图常对原始数据进行 Log2 转换。因此，热图上的单位刻度常表现为 -2, 0, 2, 4 等对数值。

3 Tbtools 使用体验与心得

Tbtools 全程鼠标点击操作，界面布局清晰，功能分类明确，自带简易操作指引。没有编程基础也能快速完成序列比对、本地数据库构建、序列提取、多序列比对等基础实验分析，极大降低了生信数据分析的入门门槛。兼容 FASTA、GFF、GTF、BED、Excel、CSV 等生信常用文件格式，可自由实现格式互相转换，适配测序原始数据、实验室测序序列、公共数据库下载序列等各类实验数据，适配绝大多数植物、动物、微生物相关基础科研分析需求。

Tbtools 本身相对于 NCBI 等是一个软件工具，没有依托庞大数据库，因此将 NCBI 等数据库与 Tbtools 搭配使用更佳。