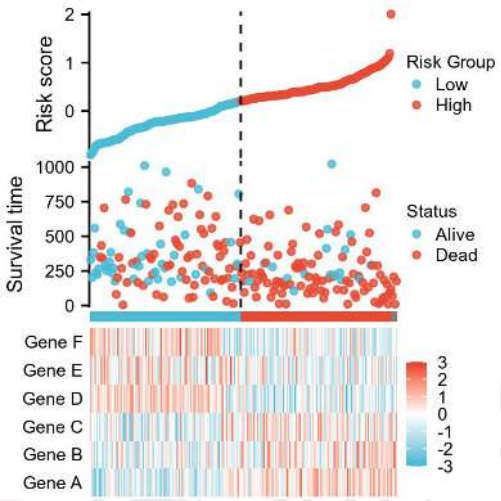
临床意义 - 风险因子图



网址：https://www.xiantao.love

更新时间：2023.03.08

目录

基本概念

- Cox 回归模型：又称为比例风险回归模型，是一种半参数回归模型。Cox 模型以生存结局和生存时间为因变量，分析众多自变量因素对生存期的影响
  - 数据要求
    - 结局建议用数字编码（0/1，1/2），其中最好用 0 代表删失或者未发生事件，1 表发生事件
    - 自变量（协变量）可以是数值或者分类变量。分类变量如果是含有等级的含义，则需要以等级资料纳入，需要设置参考组，其他组和这个参考组作对比；如果分类变量是无等级含义，一般是需要经过哑变量编码，但是经过哑变量编码后结果有可能不好解读，故无等级关系的分类变量也可以通过组合的方式形成二分类变量纳入。二分类的分类变量以等级或者非等级纳入的结果都是一致的（二分类分不分等级都一样）。数值变量可以直接以数值变量的形式纳入，亦可转换为等级资料或者二分类资料纳入
  - 条件假设：观测值独立，风险比不随时间改变（比例风险假设）。（模块内默认是满足此条件）
  - 对于回归模型的假设检验通常采用似然比检验、Wald 检验和记分检验
- Overall Survival（OS），总体生存期，指结局指标是死亡时间，这个死亡是任何原因导致的死亡都算进去，只关心是否死亡，不关心因为何种原因死亡
- Disease Free Survival（DFS），无病生存期，指经过治疗后未发现肿瘤，结局指标为疾病复发或死亡，同样不需要关心死亡原因。这一指标是临床获益的重要反映，随访时间可以缩短，因为增加了疾病复发这一节点。没有复发或没有死亡同样可以反映临床获益。这里也涉及到无疾病复发的一个定义，因此在临床资料纳入上比较困难

- PFI 无进展间隔：progression-free interval，从初次治疗的随机分组日期到疾病复发时间。（具体可以参考对应的引文）
- 中位生存时间（半数生存期）：即当累积生存率为 50%时所对应的生存时间，表示有且只有 50%的生病个体可以活过这个时间。只有当分组内最终累积生存率低于 50%才会有中位生存时间

应用场景

可视化预后模型的风险得分情况以及分组。

- 当预后模型完全由数值类型的变量（比如，分子表达）构成，可以同时绘制风险得分、风险分组、生存结局以及热图这 4 个部分
- 当预后模型不完全由数值类型的变量组成，此时不建议绘制热图部分，可以只绘制风险得分、风险分组、生存结局，以展示预后模型的风险得分的分组情况

分析流程

上传数据 ⟶ 数据验证 ⟶ 数据处理 ⟶ 风险因子可视化

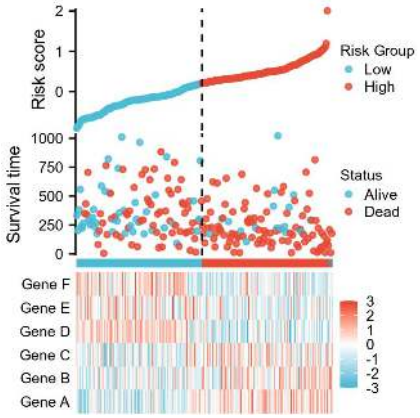
- 数据格式：
  - 第 1 列为预后结局，分类类型(不能超过 3 个分类)，可以是 0 1 编码类型
  - 第 2 列为具体时间，数值类型，必须是以天为单位，注：第 2 列(时间变量)不能都是同一个时间，并且不能出现小于 0(负数)和非数值的情况
  - 第 3 列(或者其他列)为 riskscore 值，数值类型，必须要提供，也就是说，风险因子模块需要提供 riskscore 值
  - 除了第 1、2、riskscore 这 3 列之外的所有列/变量/分子/基因都需要是数值类型的数据

	A	B	C	D	E	F	G	H	I
1	event	time	RiskScore	Gene A	Gene B	Gene C	Gene D	Gene E	Gene F
2	Alive	332	-0.913509677	3.997067148	3.20613072078052	3.7767785213626	5.546214651	5.21139879	5.229296422
3	Alive	202	-0.858900329	1.515464208	3.83262379712376	2.16092950697321	1.959998574	1.170490646	5.304746358
4	Alive	382	-0.841897473	1.095569998	3.19717036372569	2.34424893563839	4.015429258	2.982293023	3.580542639
5	Alive	559	-0.771525373	-0.049302982	0.583192787404055	1.75883916486819	3.425101659	4.373648879	4.406962192
6	Alive	240	-0.716916026	0.025270008	2.61357841546149	2.41558925047676	4.050951729	5.044522003	3.581483832
7	Alive	224	-0.710835039	2.250597742	0.0980048134519782	0.36806771085768	4.093440438	3.206994958	2.883777654
8	Alive	266	-0.6889913	2.042106649	1.99627296685928	2.31732842755404	5.582997643	2.182527863	3.111634081
9	Dead	350	-0.684150417	-1.066248492	2.80821547012889	0.284791192195326	4.221040805	3.440952126	3.689060257
10	Dead	433	-0.667147561	0.518600064	2.29669218991626	1.91702682960453	3.935830867	6.378285757	3.771251738
11	Dead	340	-0.667147561	2.793486556	0.219379622906577	2.37187175958695	3.596064752	2.377363357	0.750835406
12	Dead	705	-0.661066574	0.348810513	2.43484795127442	3.64225397087211	3.917362738	6.360311094	5.522815
13	Alive	292	-0.661066574	2.282377646	1.02215731251125	0.723388659374275	3.509140021	3.858574408	3.046298344
14	Alive	252	-0.656225691	4.183276449	1.227879240783	0.201094448754223	4.858883958	5.267809201	3.048549926
15	Alive	276	-0.650144704	2.602472576	2.33842410879044	2.48200403476603	7.103662083	5.085372709	4.527221954
16	Alive	315	-0.634381952	4.683214046	2.83997346645317	2.23649223185935	4.44761887	1.746962045	4.849790596

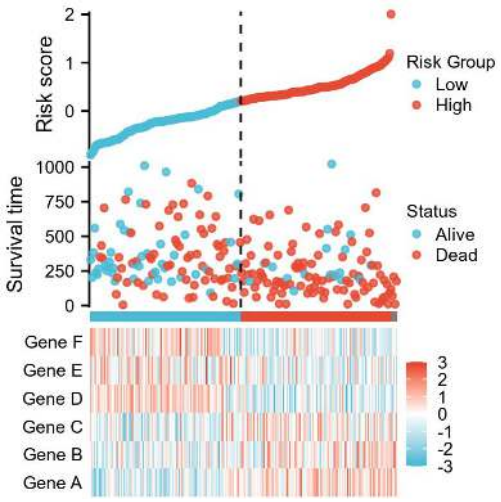
- 数据验证：对上传数据进行格式验证、数据验证等
  - 格式验证：比如第 1 列（事件）需要是分类类型数据；除了第 1 列（事件）之外的都需要是数值类型的数据
  - 数据验证：在格式验证的前提下，对不同的变量进行具体的验证，比如第 2 列（时间）不能有小于 0 的值等

- 数据处理：分别对第 1 列（事件）、第 2 列（时间）、riskscore 列以及除这 3 列外所有变量进行清洗（去除掉数据中的非数值或者不符合条件的数据）

- 可视化：对处理之后的数据分别进行不同部分内容的可视化(riskscore、事件等)，如下：

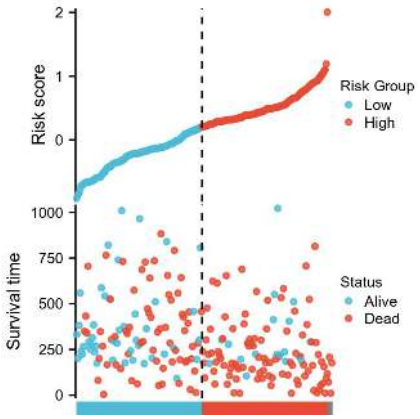


主要结果



- 图为经典风险因子图，一共包含 4 个部分：
  - 风险得分：对应上传数据中的 riskscore 列变量对每个样本的风险得分从小到大进行排序，用不同的颜色代表风险得分的高低组情况，一般是中位数分组（如果是要其他分组，需要在上传数据提供分组的列，具体看数据格式说明）
  - 生存结局：对应上传数据第 1、2 列，展示生存时间和生存结局情况，一般用点图进行展示。一般是在高风险得分组预后结局（死亡）发生多（红点多）于低风险得分组
  - 风险分组：用一个横条展示风险得分的高低分组
  - 热图：对应上传数据除了第 1、2 列和 riskscore 列之外的所有列/变量，可视化完全由数值类型（分子表达、其他纯数值类型）构建的预 后模型中的分子表达情况。一般会采取 zscore 转换。热图主要是看构建 模型的分子表达分别在高低风险得分组中的差异情况。
    - zscore 转换是热图中常用的一种对数据进行转换的方法（每个基因的表达值减去其在所有样本中的表达均值后，再除以标准差），可以减少不同分子表达值差异过大而影响整个热图的可视化效果，并且 zscore 转换保留了单个分子在样本间的差异情况

- 不绘制热图的风险因子图，一共含有三个部分：风险得分、生存结局和风险分组，如下：



- 一般用于不是完全用分子表达或者其他一类数值变量建模的预后模型展示风险因子分组的情况

数据格式

	A	B	C	D	E	F	G	H	I
1	event	time	RiskScore	Gene A	Gene B	Gene C	Gene D	Gene E	Gene F
2	Alive	332	-0.913509677	3.997067148	3.20613072078052	3.7767785213626	5.546214651	5.21139879	5.229296422
3	Alive	202	-0.858900329	1.515464208	3.83262379712376	2.16092950697321	1.959998574	1.170490646	5.304746358
4	Alive	382	-0.841897473	1.095569998	3.19717036372569	2.34424893563839	4.015429258	2.982293023	3.580542639
5	Alive	559	-0.771525373	-0.049302982	0.583192787404055	1.75883916486819	3.425101659	4.373648879	4.406962192
6	Alive	240	-0.716916026	0.025270008	2.61357841546149	2.41558925047676	4.050951729	5.044522003	3.581483832
7	Alive	224	-0.710835039	2.250597742	0.0980048134519782	0.36806771085768	4.093440438	3.206994958	2.883777654
8	Alive	266	-0.6889913	2.042106649	1.99627296685928	2.31732842755404	5.582997643	2.182527863	3.111634081
9	Dead	350	-0.684150417	-1.066248492	2.80821547012889	0.284791192195326	4.221040805	3.440952126	3.689060257
10	Dead	433	-0.667147561	0.518600064	2.29669218991626	1.91702682960453	3.935830867	6.378285757	3.771251738
11	Dead	340	-0.667147561	2.793486556	0.219379622906577	2.37187175958695	3.596064752	2.377363357	0.750835406
12	Dead	705	-0.661066574	0.348810513	2.43484795127442	3.64225397087211	3.917362738	6.360311094	5.522815
13	Alive	292	-0.661066574	2.282377646	1.02215731251125	0.723388659374275	3.509140021	3.858574408	3.046298344
14	Alive	252	-0.656225691	4.183276449	1.227879240783	0.201094448754223	4.858883958	5.267809201	3.048549926
15	Alive	276	-0.650144704	2.602472576	2.33842410879044	2.48200403476603	7.103662083	5.085372709	4.527221954
16	Alive	315	-0.634381952	4.683214046	2.83997346645317	2.23649223185935	4.44761887	1.746962045	4.849790596

数据要求：

- 数据至少 3 列、10 行，最多支持 20 列和 5000 行数据
  - 第 1 列为预后结局，分类类型(不能超过 3 个分类)，可以是 0 1 编码类型
  - 第 2 列为具体时间，数值类型，必须是以天为单位，注：第 2 列(时间变量)不能都是同一个时间，并且不能出现小于 0(负数)和非数值的情况
  - 第 3 列(或者其他列)为 riskscore 值，数值类型，必须要提供，也就是说，风险因子模块需要提供 riskscore 值
  - 除了第 1、2、riskscore 这 3 列之外的所有列/变量/分子/基因都需要是数值类型的数据
    - 如果变量是数值变量，请以数值纳入，只要含有非数值（除空值）外，则此列有可能没有办法纳入到分析

参数说明

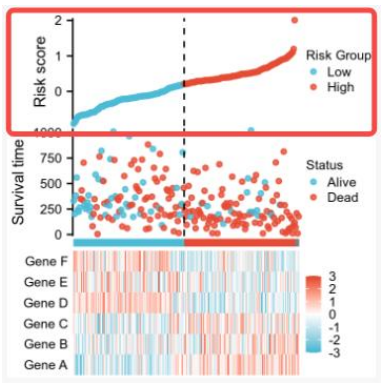
（说明：标注了颜色的为常用参数。）

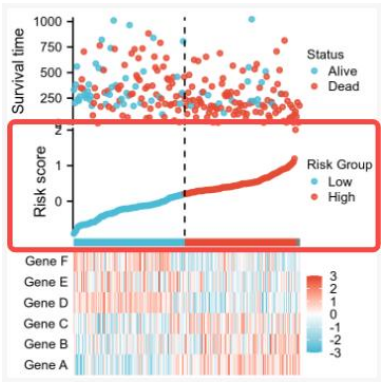
数据处理

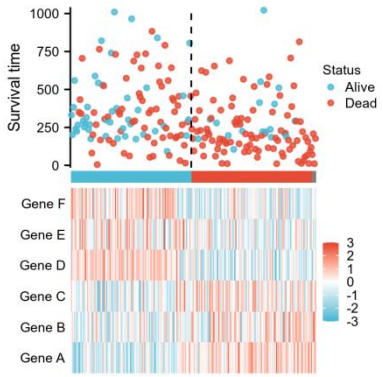
- 风险得分缺失值处理：可以选择对数据中缺失值进行处理，默认为不处理缺失

风险得分可视化

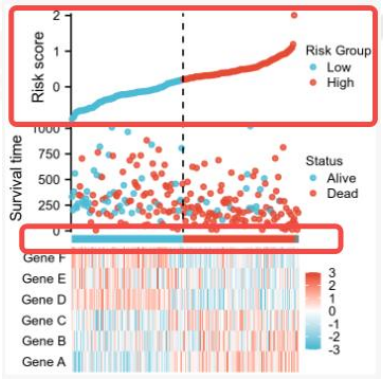
- 展示顺序：可以选择并修改风险得分部分图形在整个图形中的展示顺序，默认为 1（表示整个图形中第 1 个部分就是风险得分）还可以选择其他几个部分或者不展示（不进行风险得分部分绘制），如下：第 1 个展示顺序为 1（默认），第 2 个为展示顺序 2（调整展示顺序的时候，不管是哪一个部分都需要调整其他几个部分的位置），第 3 个为不展示

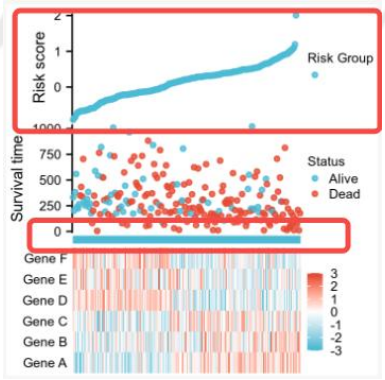




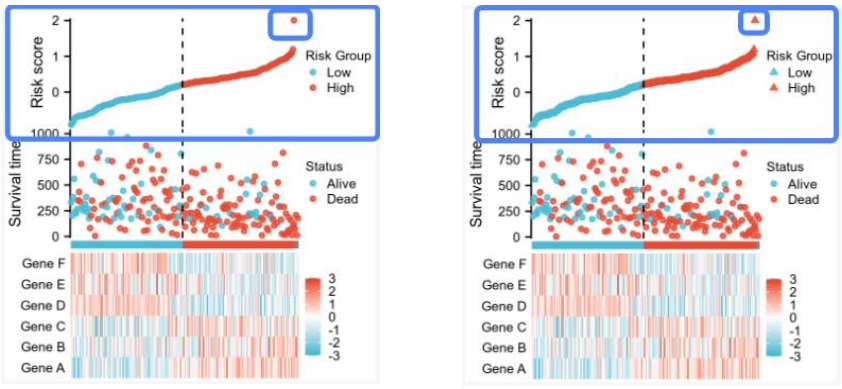


- 分组：可以选择并修改风险得分部分（第一个部分）的分组情况（就是对数据中的 riskscore 进行分组处理），默认为将 riskscore 的数据按照中位数将其分为高低两个组，同时这个分组也会同时修改第 3 个部分（风险分组）的分组情况，还可以选择根据 riskgroup（自定义上传这一列，分类类型）这一列来进行分组，或者不进行分组，如下：左侧为进行 riskscore 中位数分组，右侧为不分组：





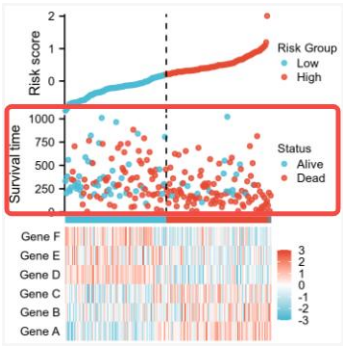
- 填充色：可以选择并修改风险得分部分图形的填充颜色
- 描边色：可以选择并修改风险得分部分图形的描边颜色
- 样式：可以选择并修改风险得分部分图形的样式，默认为圆形，还可以选择正方形、菱形、三角形、倒三角形，如下：左侧为圆形，右侧为三角形

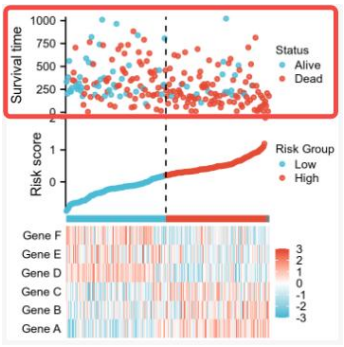


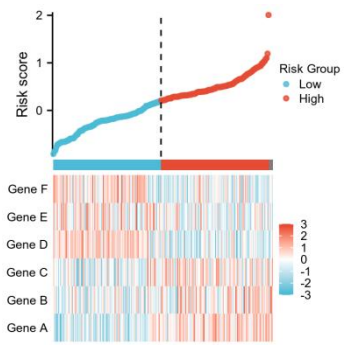
- 大小：可以修改风险得分部分图形的大小
- 不透明度：可以修改风险得分部分图形的不透明度，默认为 0.8，1 表示完全不透明，0 表示完全透明

时间部分可视化

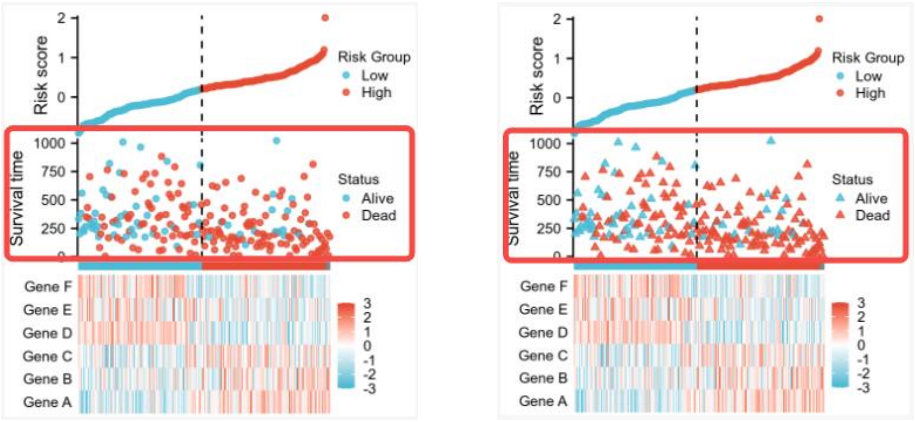
- 展示顺序：可以选择并修改时间部分图形在整个图形中的展示顺序，默认为 2（表示整个图形中第 2 个部分就是时间信息）还可以选择其他几个部分或者不展示（不进行时间部分绘制），如下：第 1 个为第 2 个部分，第 2 个为第 1 个部分（规则如风险得分部分），第 3 个为不展示







- 填充色：可以选择并修改时间部分图形的填充颜色
- 描边色：可以选择并修改时间部分图形的描边颜色
- 样式：可以选择并修改时间部分图形的样式，默认为圆形，还可以选择正方形、菱形、三角形、倒三角形，如下：左侧为圆形，右侧为三角形



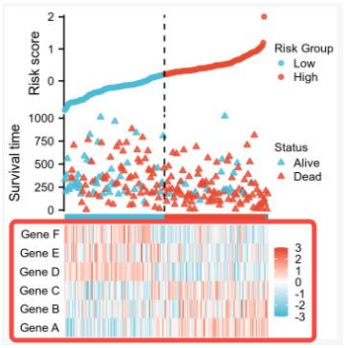
- 大小：可以选择并修改时间部分图形的大小
- 不透明度：可以选择并修改时间部分图形的不透明度，默认为 0.8，1 表示完全不透明，0 表示完全透明

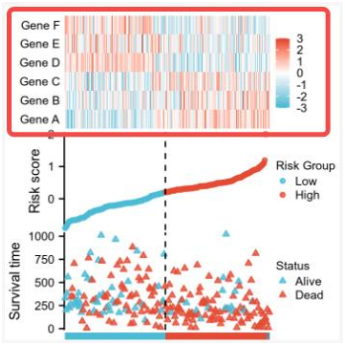
风险分组可视化

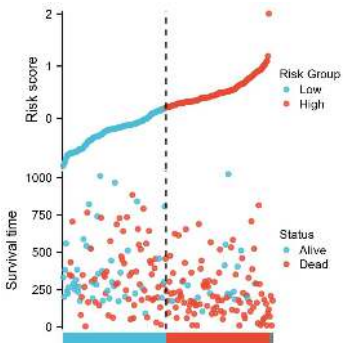
- 展示顺序：可以选择并修改风险分组部分图形在整个图形中的展示顺序，效果如上两个部分
- 颜色：可以选择并修改风险得分部分图形的颜色

热图可视化

- 展示顺序：可以选择并修改热图部分图形在整个图形中的展示顺序，默认为 4（表示整个图形中第 4 个部分就是热图）还可以选择其他几个部分或者不展示（不进行热图部分绘制），如下：第 1 个展示顺序为 4（默认），第 2 个为展示顺序 1（调整展示顺序的时候，不管是哪一个部分都需要调整其他几个部分的位置），第 3 个为不展示







- 连续变量色阶：可以选择并修改热图部分图形渐变的颜色（连续变量映射色阶的最大值和最小值对应的颜色）
- 分类变量色阶：可以选择并修改热图部分图形的颜色（分类变量映射的颜色）
- 归一化：可以选择是否对热图部分对应数据进行归一化处理，默认对列归一化(处理极值：归一化后大于 3 的值固定为 3，小于-3 固定为-3)，还可以选择归一化（默认值），和不归一化

分割线

- 是否展示：可以对风险得分以及时间对应的可视化的图添加分割线以表示分组，默认展示
- 颜色：当展示分割线的时候，可以选择并修改分割线的颜色
- 粗细：当展示分割线的时候，可以选择并修改分割线的线条粗细，默认为 0.75pt

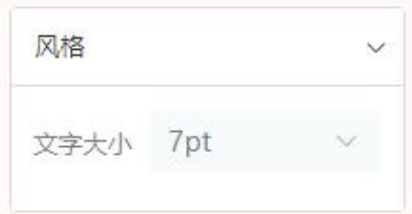
标题文本

- 大标题：大标题文本（换行可以在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]）
- 风险得分 y 轴标题：风险得分 y 轴标题（不支持换行以及其他特殊操作）
- 时间 y 轴标题：时间 y 轴标题文本（不支持换行以及其他特殊操作）

图注

- 风险得分图注标题：可以修改风险得分图注标题
- 时间图注标题：可以修改时间图注标题
- 热图图注标题：可以修改热图图注标题

风格



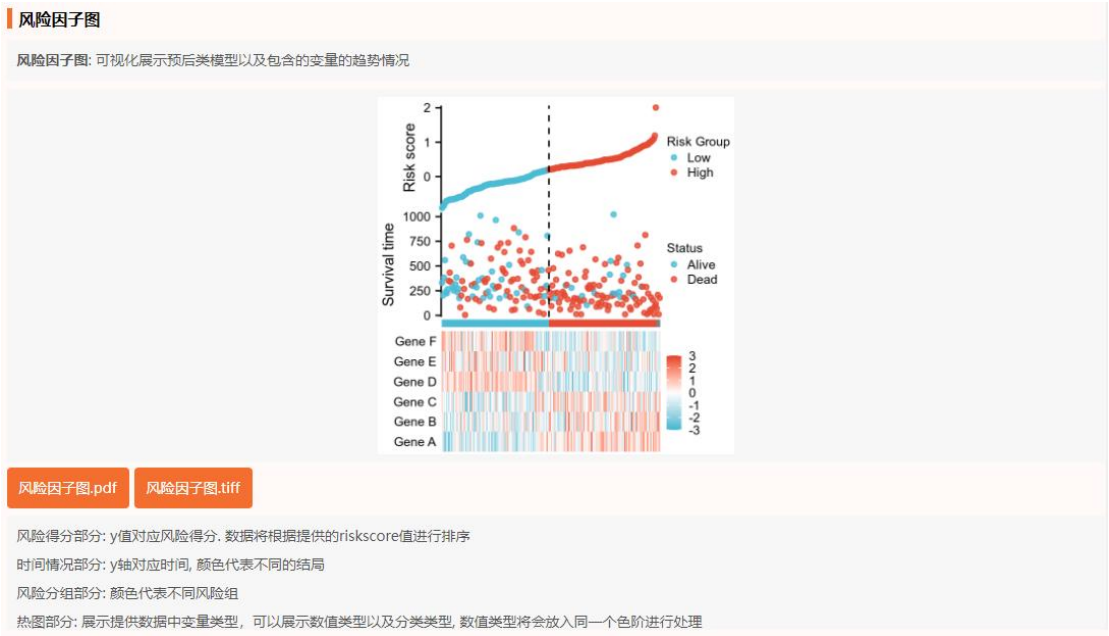
- 文字大小：控制整体文字大小，默认为 7pt

图形

- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

结果说明

主要结果



- 风险得分部分：y 值对应风险得分．数据将根据提供的 riskscore 值进行排序
- 时间情况部分：y 轴对应时间，颜色代表不同的结局
- 风险分组部分：颜色代表不同风险组
- 热图部分：展示提供数据中变量类型，可以展示数值类型以及分类类型，数值类型将会放入同一个色阶进行处理

方法学

统计分析和可视化均在R 4.2.1版本中进行

涉及的R包：ggplot2[3.3.6]

处理过程：

(1) 用ggplot2包进行风险因子图可视化

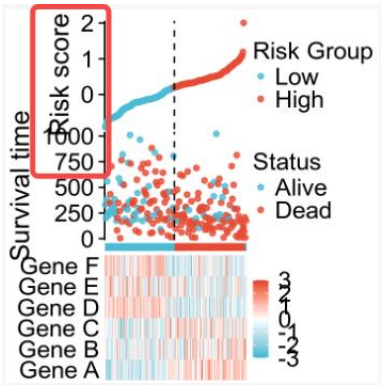
如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分（Acknowledge）致谢仙桃学术（www.xiantao.love）。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。

常见问题

1. 为什么图片中的文字会重叠?



答: 由于文字不会被压缩, 如果左侧的内容很多而图片高度不够, 就会发生重叠。解决方案可以是:

① 修改文字大小

② 增加图片高度

2. 为什么图中的结果没有垂直分割的虚线?

答: 当如果是自己的数据中包含了 Risk_Group 列, 并且右侧的参数中选择了以“Risk_Group 列”作为分组, 如果没有对数据按照 RiskScore 进行排序, 则不会有垂直分割的虚线