原始数据和图整理

2023-12-04

LiChuang Huang

@ 立效研究院

Contents

List of Figures

List of Tables

1 摘要

下载 TCGA 原始数据，绘制 Fig 1-3 内容。

1.1 数据统计

All dowloaded data number :

616

TCGAbiolinks filtered :

557

With clinical data (vital_status) :

424

Finaly used sample number :

424

1.2 原始数据

1.2.1 文件

'TCGA LUAD raw' 数据已全部提供。

(对应文件为 ./GDCdata/TCGA-LUAD/Simple_Nucleotide_Variation/Masked_Somatic_Mutation/)

> 注：文件夹./GDCdata/TCGA-LUAD/Simple_Nucleotide_Variation/Masked_Somatic_Mutation/共
> 包含 618 个文件。
> 1. 009254a2-ea81-4c76-8044-7267a9f81364
> 2. 010ba621-7ec1-452e-9115-b7ec5f9b37d2
> 3. 0121a6c3-c7ae-4619-b07a-27bb28541478
> 4. 0182902d-bb3b-4b9d-bb2a-1b00c14e3b11
> 5. 01e89769-e754-4eca-a5a3-b1b72ac4953f
> 6. ...

1.2.2 临床数据

Table 1 (下方表格) 为表格 LUAD clinical data filtered by vital status 概览。

(对应文件为 Figure+Table/LUAD-clinical-data-filtered-by-vital-status.csv)

注：表格共有 424 行 30 列，以下预览的表格可能省略部分数据；表格含有 424 个唯一 'bcr_followup_barcode'。

Table 1: LUAD clinical data filtered by vital status

follo......1	bcr_f......2	bcr_f......3	lost_...	vital...	prima...	days_......7	days_......8	...
follo...	TCGA-...	277C6...	NO	1	Compl...	474	NA	...
follo...	TCGA-...	88DF8...	NO	1	Compl...	186	NA	...
follo...	TCGA-...	FD821...	NO	1	Compl...	889	NA	...
follo...	TCGA-...	8288A...	NO	0	Compl...	NA	999	...
follo...	TCGA-...	DF6E5...	YES	1	Compl...	119	NA	...
follo...	TCGA-...	58A4C...	NO	1	Compl...	617	NA	...
follo...	TCGA-...	5CE9E...	NO	1	Compl...	492	NA	...
follo...	TCGA-...	CD19E...	NO	1	Parti...	609	NA	...
follo...	TCGA-...	f2d5e...		0		711	711	...
follo...	TCGA-...	452c5...		1	Compl...	1938	NA	...
follo...	TCGA-...	932b0...		1		760	NA	...
follo...	TCGA-...	cca64...		0	Compl...	NA	1492	...
follo...	TCGA-...	6E245...	NO	1	Compl...	1175	NA	...
follo...	TCGA-...	9a2d2...		0		NA	593	...
follo...	TCGA-...	01b8e...		0	Progr...	NA	244	...
...	...	...	...	...	...	...	...	...

1.3 Figure 1-3

Figure 1 (下方图) 为图 LUAD summary of mutation 概览。

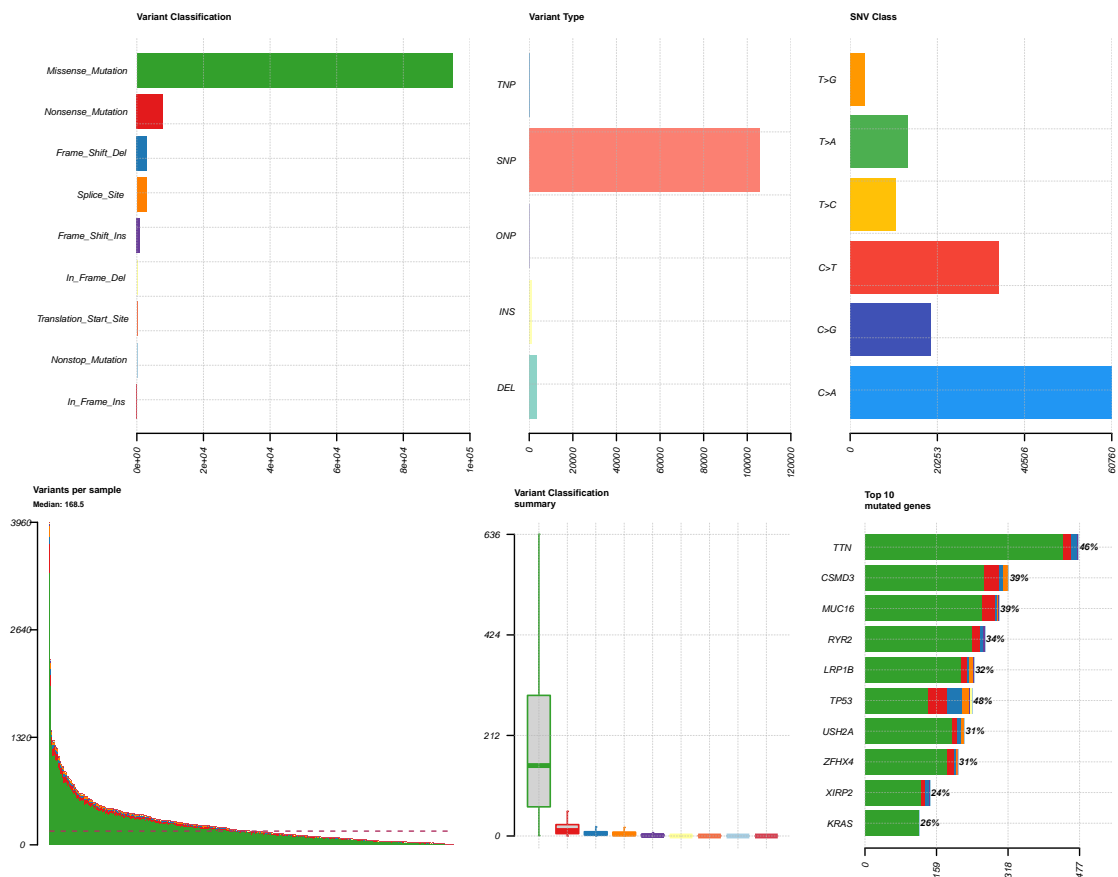**(对应文件为 `Figure+Table/LUAD-summary-of-mutation.pdf`)**

Figure 1: LUAD summary of mutation

Figure 2 (下方图) 为图 LUAD oncoplot of top genes 概览。

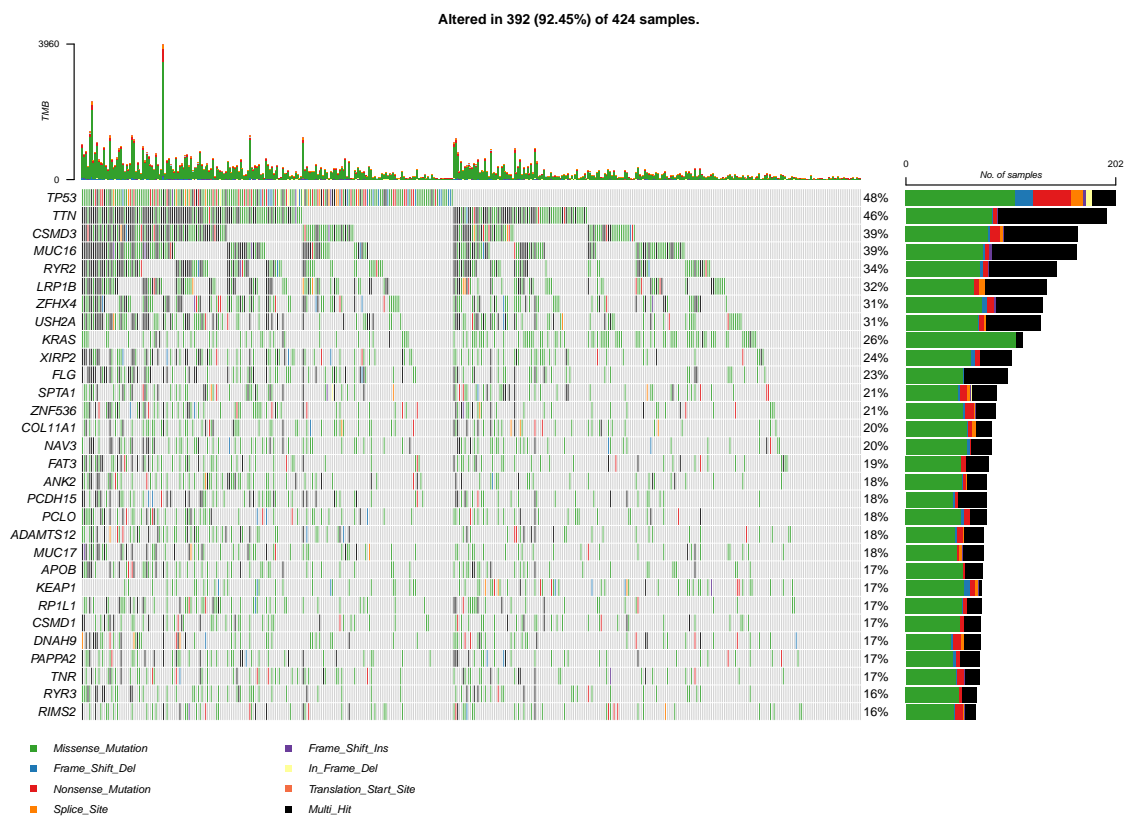**(对应文件为 Figure+Table/LUAD-oncoplot-of-top-genes.pdf)**

Figure 2: LUAD oncoplot of top genes

Figure 3 (下方图) 为图 LUAD Variable importance 概览。
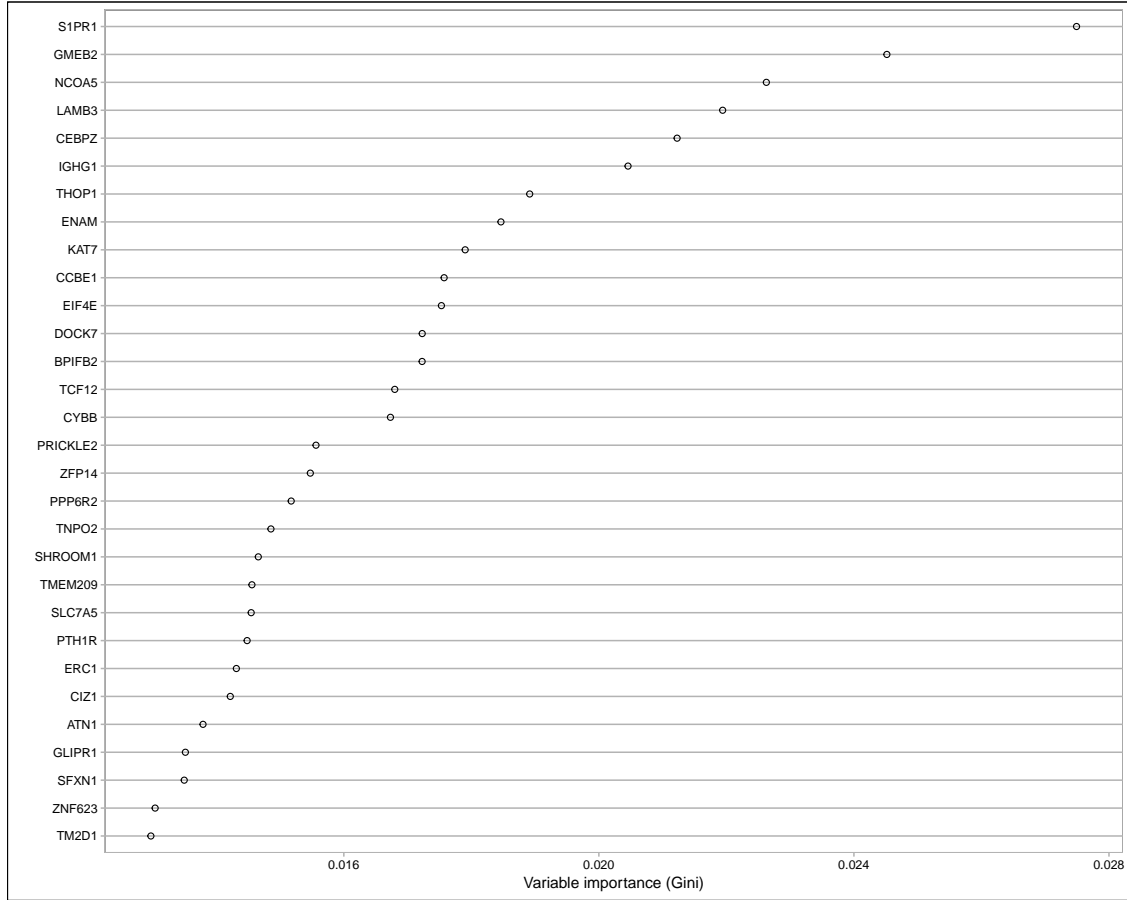
(对应文件为 Figure+Table/LUAD-Variable-importance.pdf)

Figure 3: LUAD Variable importance

2 材料和方法

2.1 方法

Mainly used method:

- Package 'maftools' used for analyzing and visualizing Mutation Annotation Format (MAF) files.[1]
- Package randomForestSRC for feature selection .
- TCGAbiolinks used for abtain TCGA dataset.[2]
- Other R packages (eg., `dplyr` and `ggplot2`) used for statistic analysis or data visualization.

Reference

1. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Research* **28**, 1747–1756 (2018).

2. Colaprico, A. *et al.* TCGAbiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research* **44**, (2015).