Disease Subtypes

Francesco Lescai

2023-06-08

Dataset

A hypothetical disease is quite difficult to treat due to probable subtypes of the pathology which have been difficult to identify. A study has therefore been conducted on about 8,000 patients which have been assessed for the expression of 15 genes considered critically relevant in the phenotype.

The dataset includes the following data:

gene01	gene02	gene03	gene04	gene05	gene06	gene07
570	411	0	224	424	13	105
524	438	1	244	443	13	107
607	429	1	232	473	11	100
554	373	1	235	438	13	96
545	446	3	217	465	16	87
578	397	0	235	454	16	94

gene08	gene09	gene10	gene11	gene12	gene13	gene14	gene15
47900	3	10120477	209	87793	0	1427244	27
48068	3	10115301	207	87587	0	1429141	25
47611	2	10120671	202	87777	2	1426594	23
47660	1	10116683	254	87598	0	1426647	23
48033	2	10115940	210	87988	0	1428205	31
47499	4	10115286	210	88068	0	1427269	23

Assignment

Please analyse this dataset using the most appropriate methods. Prepare a report discussing your choices step by step, and presenting a data-driven justification for the analytical decisions you made.

Provide evidence, if appropriate, of relationships of dependencies in the dataset, explaining how some of the variables might influence your findings.

Discuss in the report, where appropriate, any biological background which might support your findings.

Use the most appropriate computing environment to carry out this work, and explain the code and the choice you made in a dedicated section of the report.