KU LEUVEN

Master of bioinformatics

# Applied multivariate statistical analysis

Multivariate dataset exploration: genome assembly

Summer 2016

*Author:*
Cedric Lood

*Teacher:*
Prof. Eddie Schrevens


August 19, 2016

# 1 Context of this project

The analysis presented in this report was produced for the class of *Applied multivariate statistics* taught at KU Leuven (Winter 2015). The requirement for the class included the exploration of a multivariate dataset of our choice in order to discover its structure. The implementation was done using the R programming environment (v3.3.0), and the dataset along with the code can be found online in my github account[1].

# 2 Genome assembly: quality control metrics

The dataset consists of quality control metrics for *de novo* genome assembly [1]. It originates from an analysis performed in the early stages of my master thesis in which I investigated the genomics of a set of 47 nosocomial isolates of the bacterial species *Pseudomonas aeruginosa*.

The goal of genome assembly is to reconstruct the genome of an organism using the reads issued by a sequencer, which in my case was an *Illumina NextSeq 500*. For each of the 47 strains, multiple genome assemblies were performed using different software and approaches, each assembly giving different hypothesis for what the original sequenced genome looks like.

In this analysis, the goal is to detect outliers in order to remove them from the set of hypothesis, and explore the structure of the data.

# 3 Description of the dataset

The dataset consists of 3102 observations, as each of the 47 strains went through 67 different assemblies each, and 36 variables were surveyed. A few missing values exist in the dataset, but their amount is very limited (90 out of 3102) and can be traced back to software errors, hence they were removed prior to the analysis. An exhaustive description of the variables is available here [2], here is a summary of selected variables:

| # | name | description |
|---|---|---|
| 1 | Strain ID | Label with the ID of the strain (from 9108 to 9154) |
| 2 | Assembly | Label for the 67 assembly pipelines |
| 3 | Hybrid | Boolean value indicating the approach for the assembly |
| 4 | Coverage | Genome coverage estimation based on sequencing results |
| 5 | NContig | Total number of contigs for the assembly |
| 6 | LargestContig | Length (in base pairs) of the longuest contig |
| 7 | TotalLength | Total length of the assembled genome |
| 8 | ReferenceLength | Length of reference genome used for QC evaluation |
| 9 | GC | GC content of the assembled genome (%) |
| 10 | ReferenceGC | GC content of reference genome used for QC evaluation |
| 11 | N50 | Minimum length of contig comprising 50% of assembled genome |
| 12 | NG50 | Corrected N50 using the length of reference genome |
| 13 | N75 | Minimum length of contig comprising 75% of assembled genome |
| 15 | L50 | Number of contigs of length greater than N50 |
| 19 | Nmisassemblies | Number of misassemblies events |
| 24 | GenomeFraction | Fraction of reference genome covered by assembly |
| 29 | NA50 | Corrected N50 taking into account misassemblies |
| 30 | NGA50 | Corrected NG50 taking into account misassemblies |
| 36 | LGA75 | Corrected LG75 taking into account misassemblies |

[1] https://github.com/Milt0n/MVDataExploration

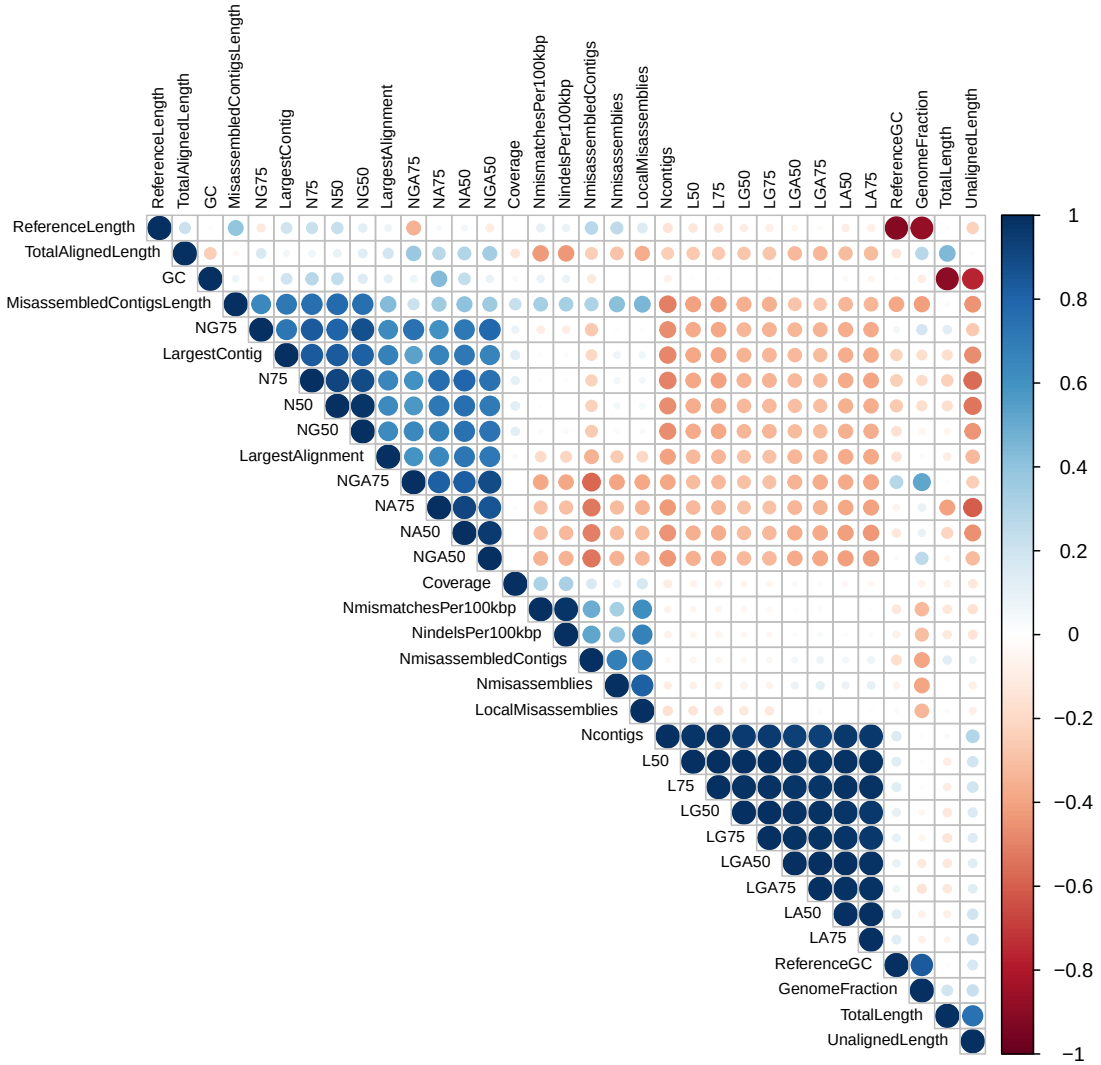


Figure 1: This graph shows the correlation between the different variables of the dataset, from a correlation of +1 (blue) to -1 (red)

# 4 R code

# 5 Output

## 5.1 Dataset structure

# References

[1] M. Baker, "De novo genome assembly: what every biologist should know," *Nature methods*, vol. 9, no. 4, p. 333, 2012.

[2] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "Quast: quality assessment tool for genome assemblies," *Bioinformatics*, p. btt086, 2013.