



**KU LEUVEN**

MASTER OF BIOINFORMATICS

---

# Applied multivariate statistical analysis

Multivariate dataset exploration: genome assembly

---

Summer 2016

*Author:*  
Cedric LOOD

*Teacher:*  
Prof. Eddie SCHREVENs



August 22, 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<sup>1</sup>.

## 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 in order to decide which variable can be used further to select the hypothesis via an ensemble method.

## 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 (43 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 longest 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

---

<sup>1</sup><https://github.com/Milt0n/MVDataExploration>

## 4 Output

### 4.1 Dataset structure

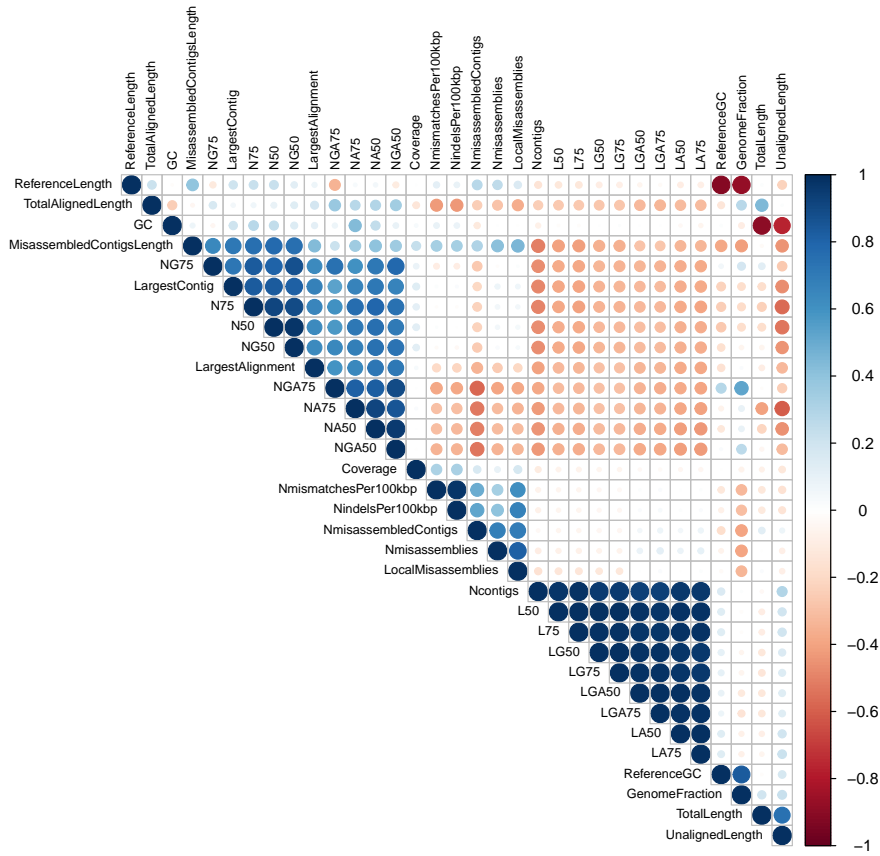


Figure 1: This graph shows the correlation between the different variables of the dataset, from a correlation of +1 (blue) to -1 (red)

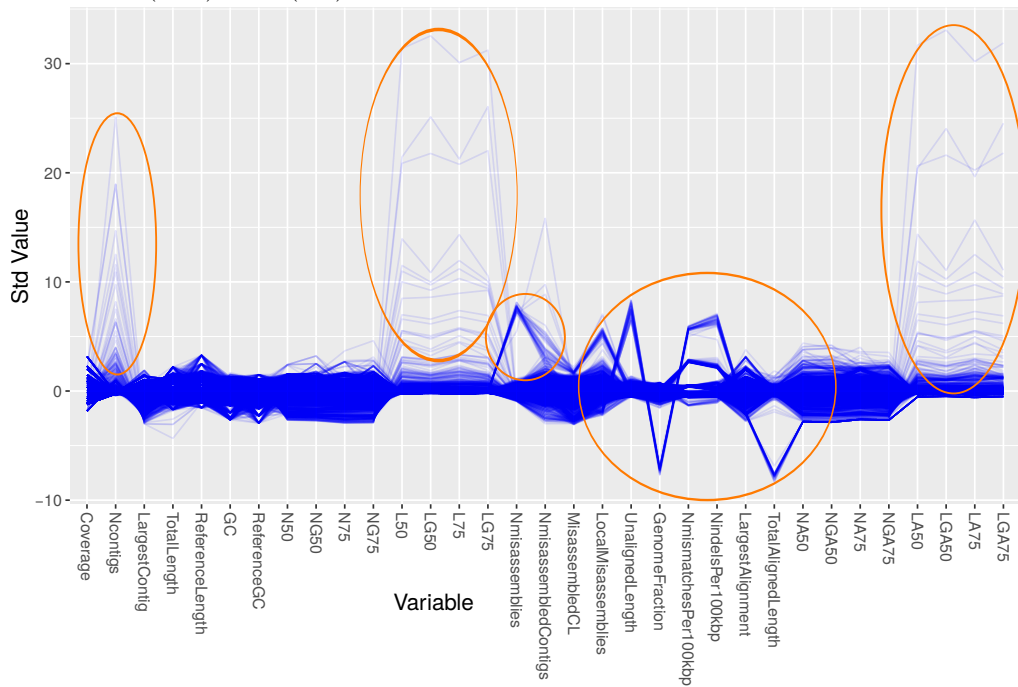
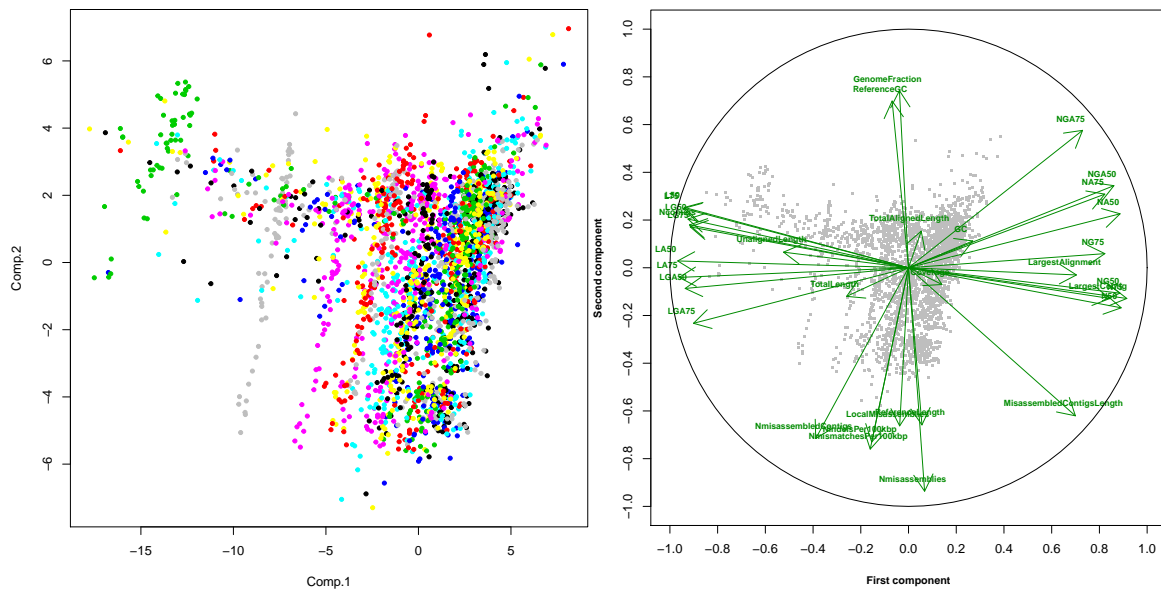
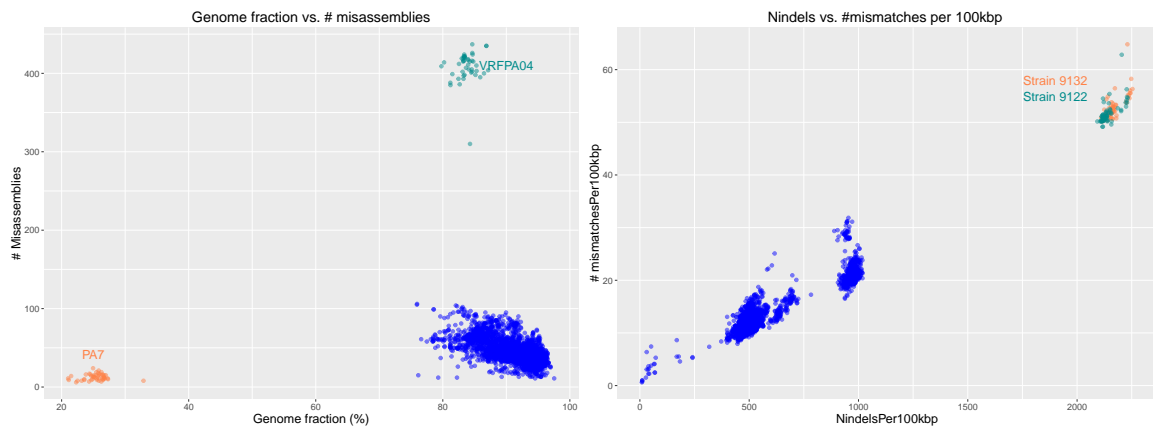


Figure 2: Parallel coordinates plot showing strong outliers (circled in orange)



## 5 Interpretation

The dataset displays a lot of correlational structure in many of the variables, as illustrated on figure 1. The parallel coordinates plot (figure 2) allowed a very straightforward identification of many outlier clusters, which are illustrated on figure 3. A biological explanation was discovered for the presence of these clusters, and they were eliminated from the list of hypothesis.

Given Large amount of correlational structure, using PCA makes sense here to see the real dimensionality of the dataset. I used the correlation because of huge scale differences so as not to model the scales. A large amount of variability can be captured using the PC (the first 7 PC explain 91% of the variability, and 10 PC explaining 96%). The plot and biplot of the two first component (figure 4) showed an interesting line grouping structure, with the colors indicating the different strains, this grouping structure was also partially captured during attempts at hierarchical clustering. Factor analysis (code not reported here) also showed a large amount of commonality using few factors.

The analysis presented here was also combined with some other univariate exploration of the variables, such as GC content, Length of assembly, etc. and it allowed me to filter out a fair amount of outlier hypothesis. In the end, the metric NGA50 proved to be the most interesting to make my ensemble decision upon.

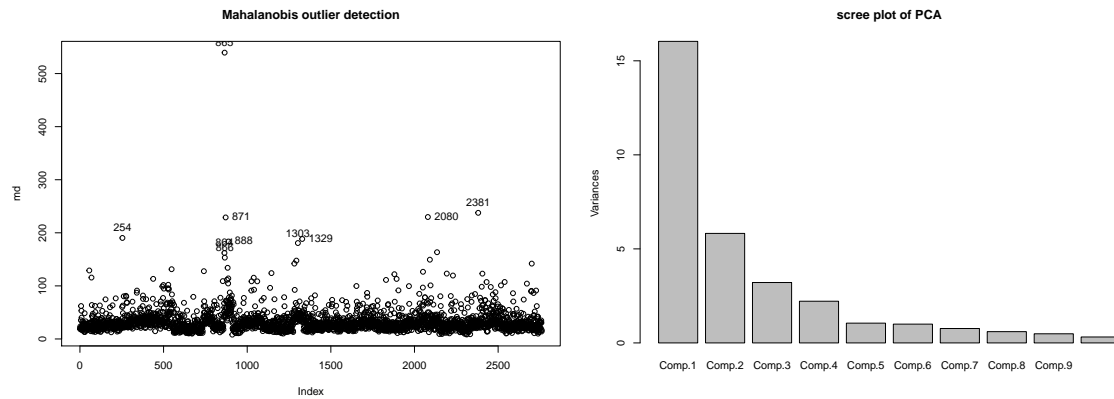


Figure 5: Left: mahalanobis plot and outliers identification, right: PCA scree plot

## 6 R Code

Some of the ggplot code has been truncated (themes and annotations). The full code is available in my github repository, together with other attempts at analysis (factor analysis, clustering):

```
1 library(ggplot2)
2 library(corrplot)
3 library(pastecs)
4 library(plotrix)
5
6 setwd(dir = "/home/sid/Dev/MVDataExploration/data/")
7 asm <- read.csv(file = "quast_all_metrics_reduced.csv", header=TRUE)
8 str(asm); dim(asm)
9
10 # basic dataset statistics, correlations
11 stat.desc(asm[,4:36], basic=TRUE, desc=TRUE)
12 asm.corr <- cor(asm[,4:36])
13 corrplot(asm.corr, type="upper", order="hclust", tl.col="black", tl.srt=90)
14
15 # Profile plot
16 asm.matrix <- as.matrix(asm[,4:36])
17 asm.matrix.std <- scale(asm.matrix, center=T, scale=T)
18 asm.melt <- melt(asm.matrix.std)
19 colnames(asm.melt) <- c("RowID", "Variable", "Value")
20
21 ggplot(asm.melt, aes(x=Variable, y=Value, group=RowID)) +
22   geom_line(colour=I("blue"), alpha=0.1)
23
24 # outliers
25 ggplot(asm, aes(GenomeFraction, Nmisassemblies)) +
26   geom_point(data=subset(asm, Assembly != "PA7" & Assembly != "VRFPA04"),
27             colour=I("blue"), alpha=0.5) +
28   geom_point(data=subset(asm, Assembly == "PA7"), colour=I("sienna1")) +
29   geom_point(data=subset(asm, Assembly == "VRFPA04"), colour=I("cyan4"))
30 ggsave("scatterplot_gfvsmis.pdf")
31
32 # subsetting (removal of the 2 pipelines PA7 and VRFPA04, and Ncontigs > 400)
33 asm.filter <- subset(asm, Assembly != "PA7" & Assembly != "VRFPA04" & Ncontigs
34   < 400)
35
36 ggplot(asm.filter, aes(NmismatchesPer100kbp, NindelsPer100kbp)) +
37   geom_point(data=subset(asm.filter, StrainID != "9122" & StrainID != "9132"),
38             colour=I("blue"), alpha=0.5) +
39   geom_point(data=subset(asm.filter, StrainID == "9122"), colour=I("sienna1")) +
40   geom_point(data=subset(asm.filter, StrainID == "9132"), colour=I("cyan4"))
41 ggsave("scatterplot_indels.pdf")
```

```

39
40 # further subsetting using mahalanobis
41 asm.filter <- subset(asm.filter , StrainID != "9122" & StrainID != "9132" &
  Ncontigs < 400)
42 md <- mahalanobis(asm.filter[,4:36], colMeans(asm.filter[,4:36]), cov(asm.
  filter[,4:36]))
43 plot(md, main="Mahalanobis outlier detection")
44 identify(md)
45 asm.filter <- asm.filter[-c(865,254,871,1329,2080,2381,888,1303,866,864),]
46
47 # PCA analysis
48 asm.pca <- princomp(asm.filter[,4:36], cor = T)
49 screeplot(asm.pca)
50 asm.pca$loadings
51 summary(asm.pca)
52 plot(asm.pca$scores[,1:2], type="p", pch=19, cex=0.7, col=asm.filter$StrainID)
53
54 # biplot
55 asm.matrix <- as.matrix(asm.filter[,4:36])
56 xm<-apply(asm.matrix,2,mean)
57 y<-sweep(asm.matrix,2,xm)
58 ss<-(t(y)%*%y)
59 s<-ss/(nrow(x)-1)
60 d<-(diag(ss))^(1/2)
61 e<-diag(d,nrow=ncol(asm.matrix),ncol=ncol(asm.matrix))
62 z<-y%*%e
63 r<-t(z)%*%z
64 q<-svd(z)
65 gfd<-((q$d[1])+(q$d[2]))/sum(q$d)
66 gfdz<-(((q$d[1])^2)+(q$d[2]^2))/sum((q$d)^2)
67 gfr<-(((q$d[1])^4)+(q$d[2]^4))/sum((q$d)^4)
68 l<-diag(q$d,nrow=ncol(asm.matrix),ncol=ncol(asm.matrix))
69 R.B<-q$u #scores matrix
70 C.B<-q$v%*%l #loadings
71 #possibility to stretch scores by a scale factor
72 scalefactor<-10
73 R.B<-q$u *scalefactor
74
75 par(mar=c(4,4,4,4),pty='s',oma=c(5,0,0,0),font=2)
76 plot(R.B[,1],R.B[,2],axes=F,xlim=c(-1,1),ylim=c(-1,1),xlab='',ylab='',cex
  =2.8, pch=".", col="grey")
77 mtext('First component',side=1,line=3,cex=.8)
78 mtext('Second component',side=2,line=3,cex=.8)
79 axis(1,at=c(-1,-.8,-.6,-.4,-.2,0,.2,.4,.6,.8,1),cex=.8)
80 axis(2,at=c(-1,-.8,-.6,-.4,-.2,0,.2,.4,.6,.8,1),cex=.8)
81 box()
82
83 text(C.B[,1] -.05,C.B[,2] +.05,as.character(dimnames(asm.matrix)[[2]]),cex=0.7,
  col="green4")
84 for (i in seq(1,nrow(C.B),by=1))
85   arrows(0,0,C.B[i,1],C.B[i,2],col="green4")
86
87 #Draw circle unit
88 draw.circle(0,0,1,border='black')

```

## 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.