Problem 1. (30 points)

Computations on gene means of the Golub data set.

(a) Compute the mean expression values for every gene among “ALL” patients.

(b) Compute the mean expression values for every gene among “AML” patients.

(c) Give the biological names of the three genes with the largest mean expression value among “ALL” patients.

(d) Give the biological names of the three genes with the largest mean expression value among “AML” patients.

Submit R commands that does (a)-(d). And answer directly part (c) and (d)

(a)(b): included in the R code.

(c): [1] Gdf5 gene
[2] GB DEF = Chromosome 1q subtelomeric sequence D1S553
[3] 37 kD laminin receptor precursor/p40 ribosome associated protein gene

(d): [1] GB DEF = mRNA fragment for elongation factor TU (N-terminus)
[2] GB DEF = HLA-B null allele mRNA
[3] Globin, Beta

Problem 2. (30 points)

More work on the Golub data set.

(a) Save the expression values of the first five genes (in the first five rows) for the AML patients in a csv file “AML5.csv”.

(b) Save the expression values of the first five genes for the ALL patients in a plain text file “ALL5.txt”.

(c) Compute the standard deviation of the expression values on the first patient, of the 100th to 200th genes (total 101 genes).
(d) Compute the standard deviation of the expression values of every gene, across all patients. Find the number of genes with standard deviation greater than 1.

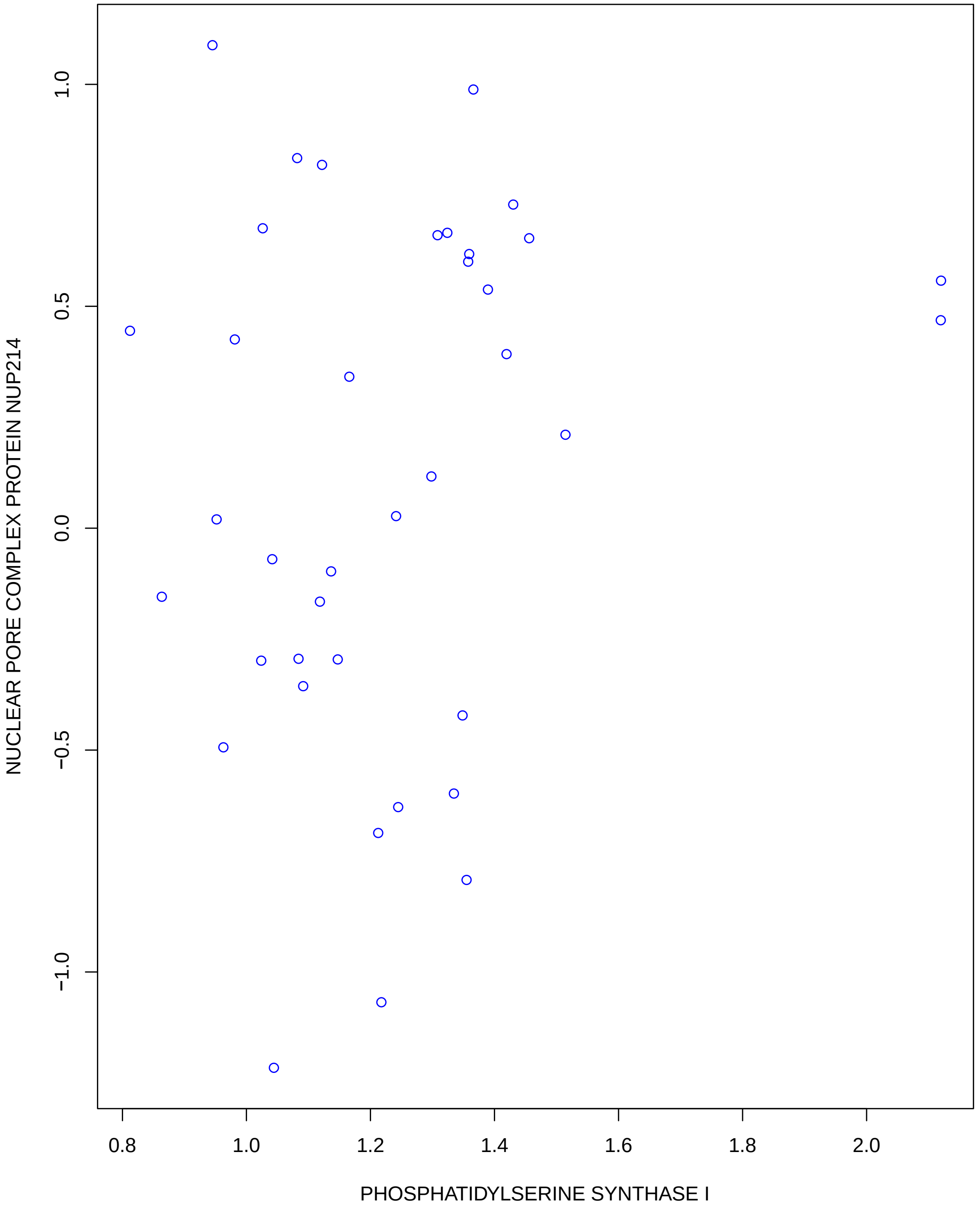
(e) Do a scatter plot of the 101th gene expressions against the 102th gene expressions, label the x-axis and the y-axis with the genes’ biological names using xlab= and ylab= control options.

Submit R commands that does (a)-(e). And the outputs (files for parts (a), (b), numerical answer for part (c) and (d), the figure file for part (e)).

(C）0.9174976

(D) 123

(E)


NUCLEAR PORE COMPLEX PROTEIN NUP214
1.0
0.5
0.0
−0.5
−1.0
0.8
1.0
1.2
1.4
1.6
1.8
2.0
PHOSPHATIDYLSERINE SYNTHASE I

Problem 3. (20 points)

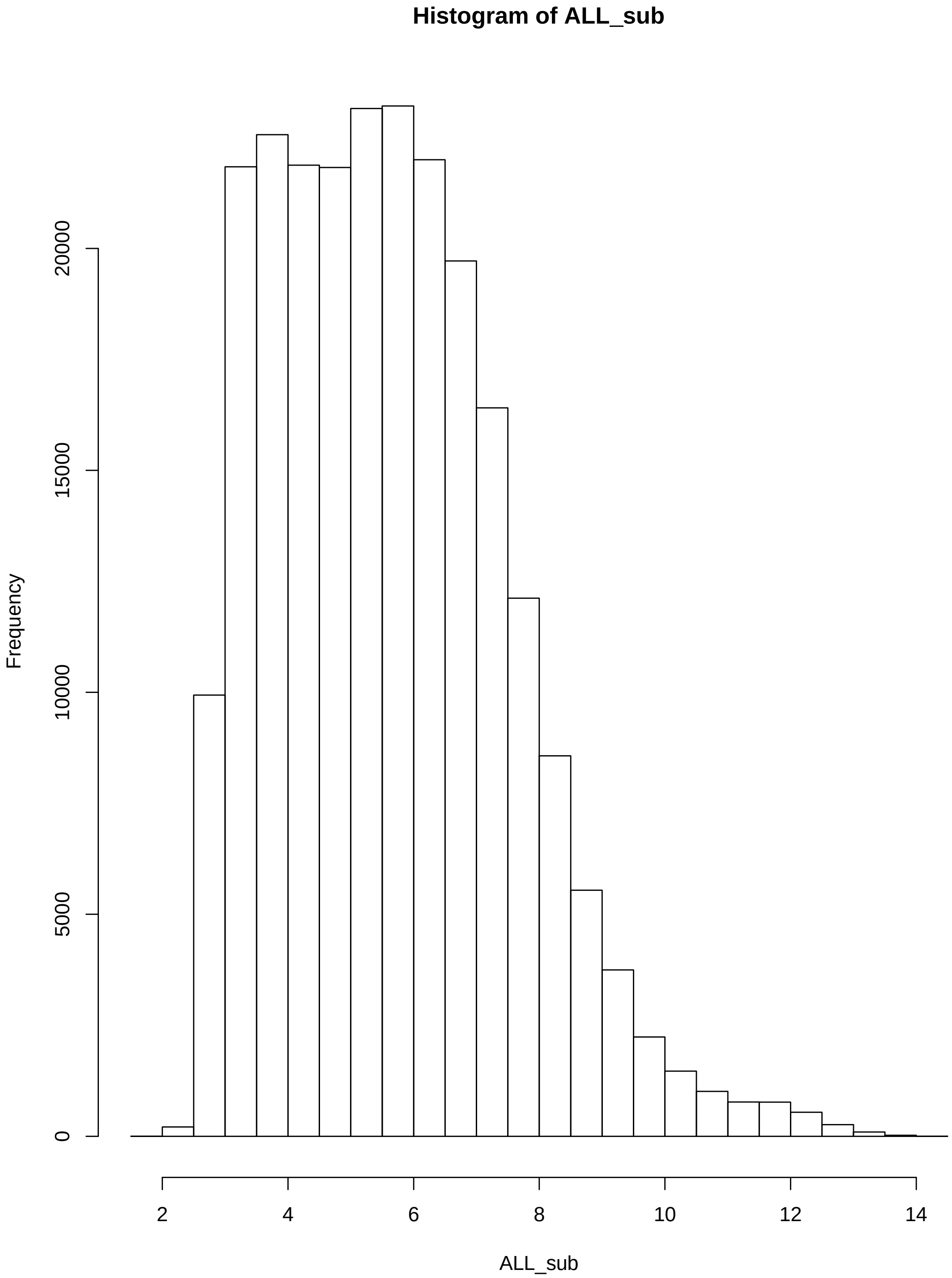
Work with the ALL data set. Load the ALL data from the ALL library and use str and openVignette() for a further orientation.

(a) Use exprs(ALL[,ALL$BT=="B1"] to extract the gene expressions from the patients in disease stage B1. Produce one histogram of these gene expressions in the this matrix.
(b) Compute the mean gene expressions for every gene over these B1 patients.

(c) Give the gene identifiers of the three genes with the largest mean.

Submit R commands that does (a)-(c), and answer part (c) directly.

（A）



(C) 1000_at 1001_at 1002_f_at

7.343891   4.944577   3.931296

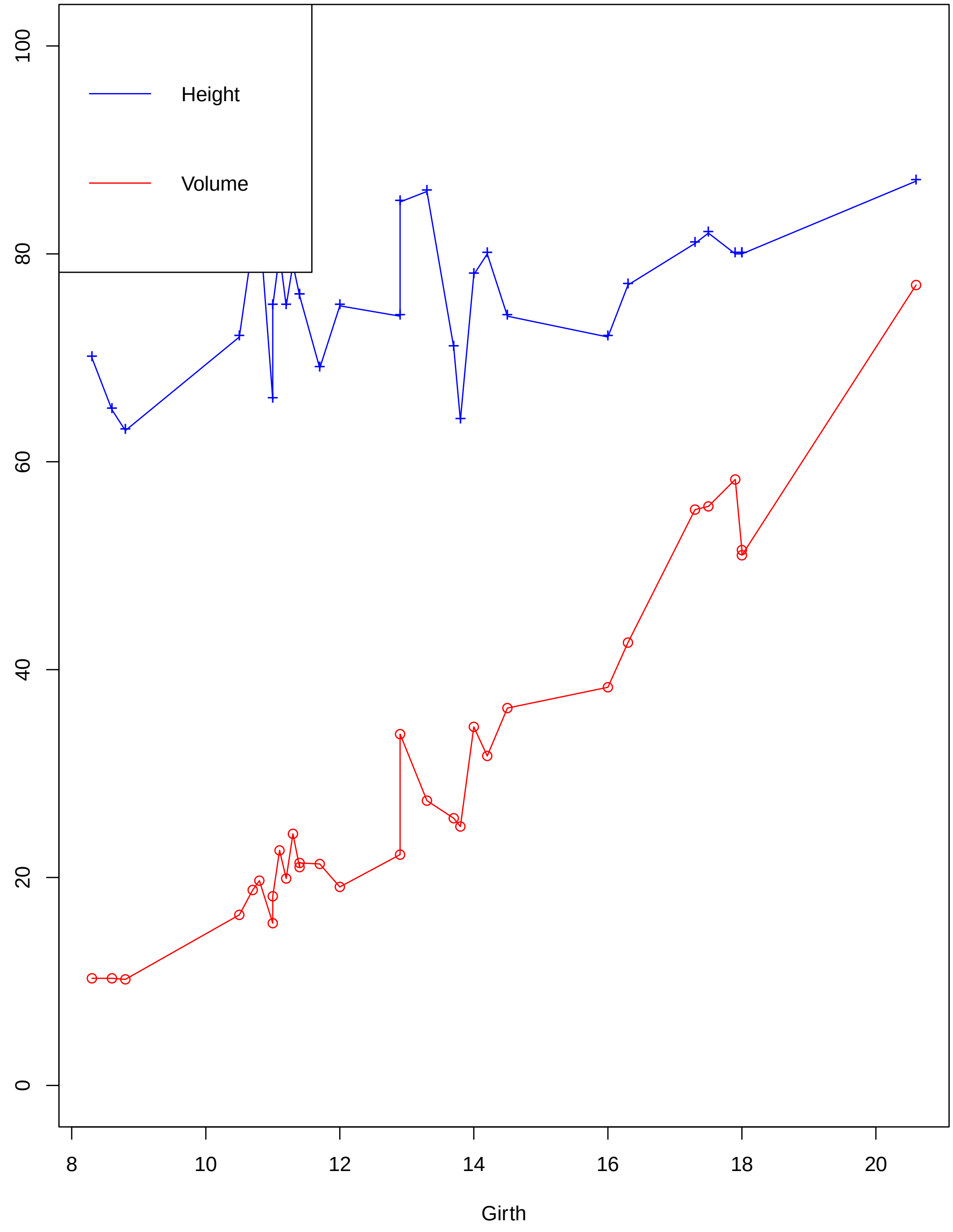
Problem 4. (20 points)

We work with the “trees” data set that comes with R.

(a) Find the type of the trees data object.
(b) Produce a figure with two overlaid scatterplots: Height versus Girth, Volume versus Girth(The Girth is on the x-axis). Do the Height plot with blue “+” symbols, and do the Volume plot with red “o” symbols. You need to learn to set the ylim= control option so that all points from the two plots can all show up on the merged figure.

Hint: you should use plot() then points() to create the overlaid two scatterplots.

（A） Numerical
（B）


Height
Volume
0
20
40
60
80
100
8
10
12
14
16
18
20
Girth