004 - Exploring Data - Part II

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on September 7, 2020

Summarizing relationships between two variables

Approaches for summarizing relationships between two variables vary depending on variable types...

- Two numerical variables
- Two categorical variables
- One numerical variable and one categorical variable

Two numerical variables

Two categorical variables

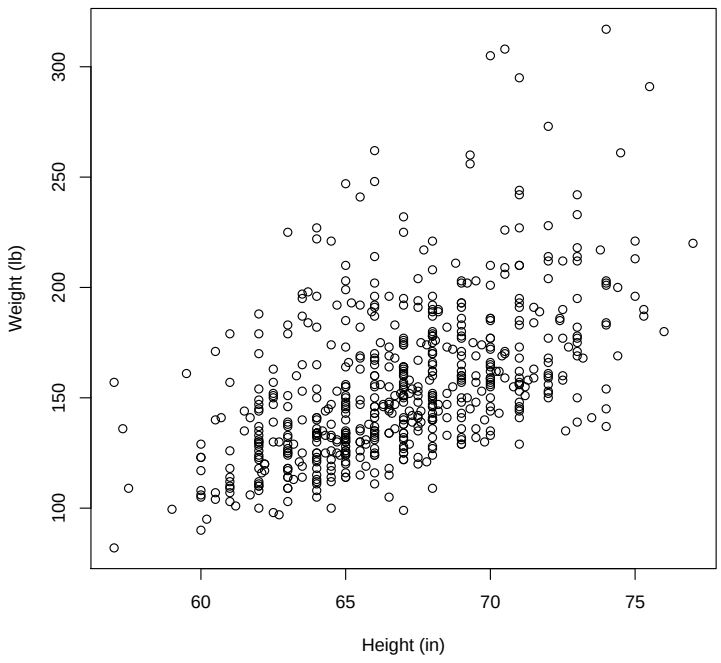
Case study: molecular cancer classification

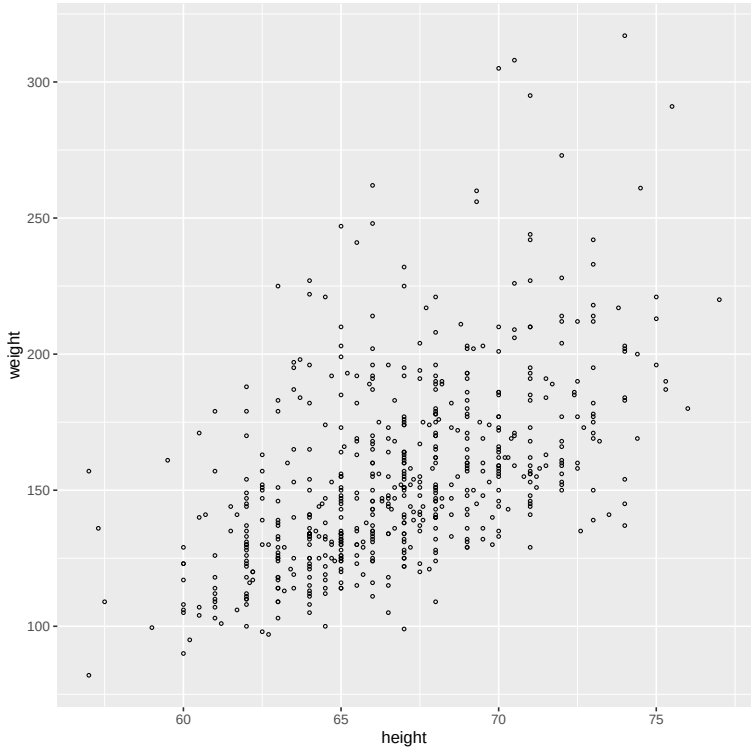
Scatterplots

```
library(ggplot2); library(oibiostat);
data(famuss)

plot(famuss$height, famuss$weight, xlab = "Height (in)", ylab = "Weight (lb)")

ggplot(data = famuss, mapping = aes(x = height, y = weight)) +
  geom_point(size = 0.8, pch = 21)
```





Correlation coefficient

- The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of x and y, and s_x and s_y are the sample standard deviations of the x and y variables, respectively.

Correlation coefficient

- The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of x and y, and s_x and s_y are the sample standard deviations of the x and y variables, respectively.

- The correlation coefficient quantifies the strength of a **linear** trend.

Correlation coefficient

- The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of x and y, and s_x and s_y are the sample standard deviations of the x and y variables, respectively.
- The correlation coefficient quantifies the strength of a **linear** trend.
- The correlation coefficient r takes on values between -1 and 1.

Correlation coefficient

- The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of x and y, and s_x and s_y are the sample standard deviations of the x and y variables, respectively.
- The correlation coefficient quantifies the strength of a **linear** trend.
- The correlation coefficient r takes on values between -1 and 1.
- The closer r is to ± 1, the stronger the linear association.

Correlation coefficient

- The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of x and y, and s_x and s_y are the sample standard deviations of the x and y variables, respectively.
- The correlation coefficient quantifies the strength of a **linear** trend.
- The correlation coefficient r takes on values between -1 and 1.
- The closer r is to ± 1, the stronger the linear association.
- Two variables x and y are
  - *positively associated* if y increases as x increases ($r > 0$)
  - *negatively associated* if y decreases as x increases ($r < 0$)

Correlation in R

- Correlation between weight and height in the `famuss` dataset:

```
cor(famuss$height, famuss$weight)
## [1] 0.5308787
```

Correlation in R

- Correlation between weight and height in the `famuss` dataset:

```
cor(famuss$height, famuss$weight)
```

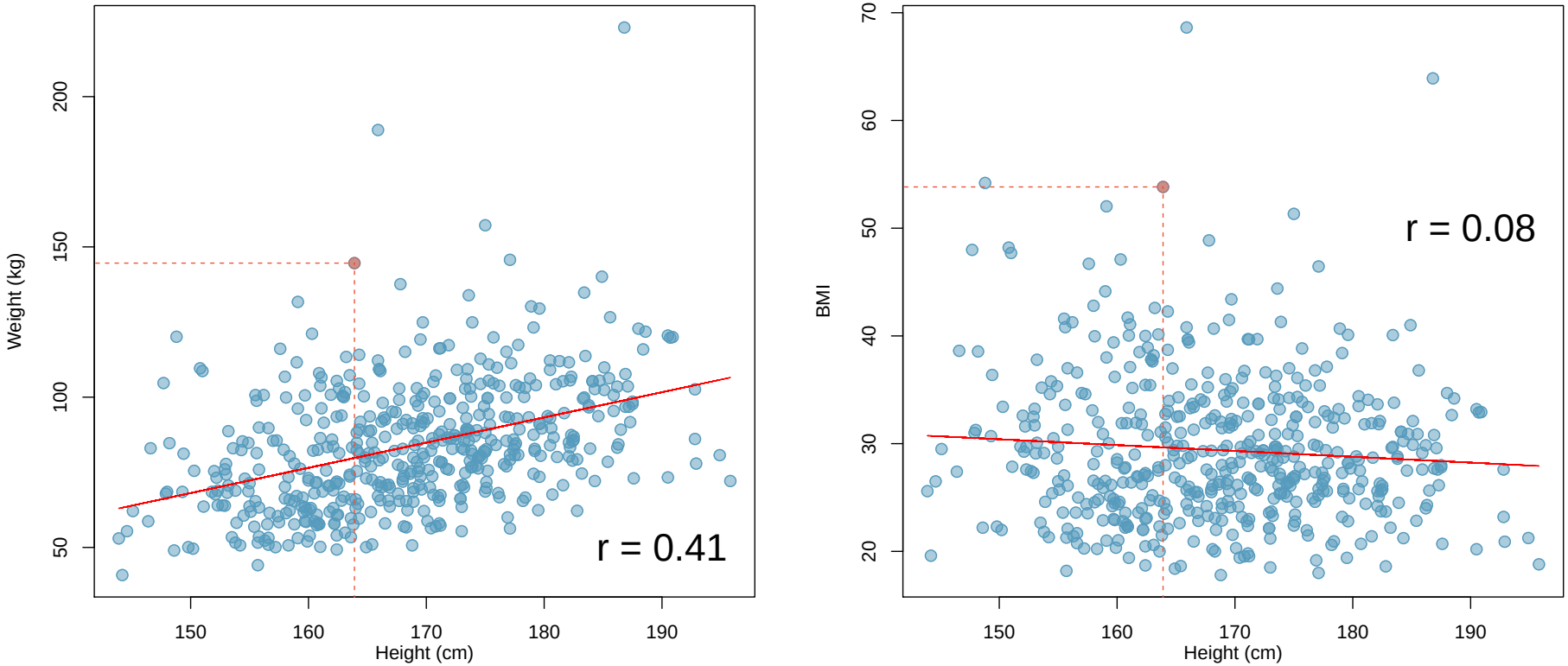
```
## [1] 0.5308787
```

- We can also obtain the correlation between `weight` and `height` from a simple linear regression:

```
summary(lm(height ~ weight, data = famuss))
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.295213   0.573200  101.70   <2e-16 ***
## weight       0.054843   0.003595   15.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.031 on 593 degrees of freedom
## Multiple R-squared: 0.2818,^^IAdjusted R-squared: 0.2806
## F-statistic: 232.7 on 1 and 593 DF,  p-value: < 2.2e-16
```

title

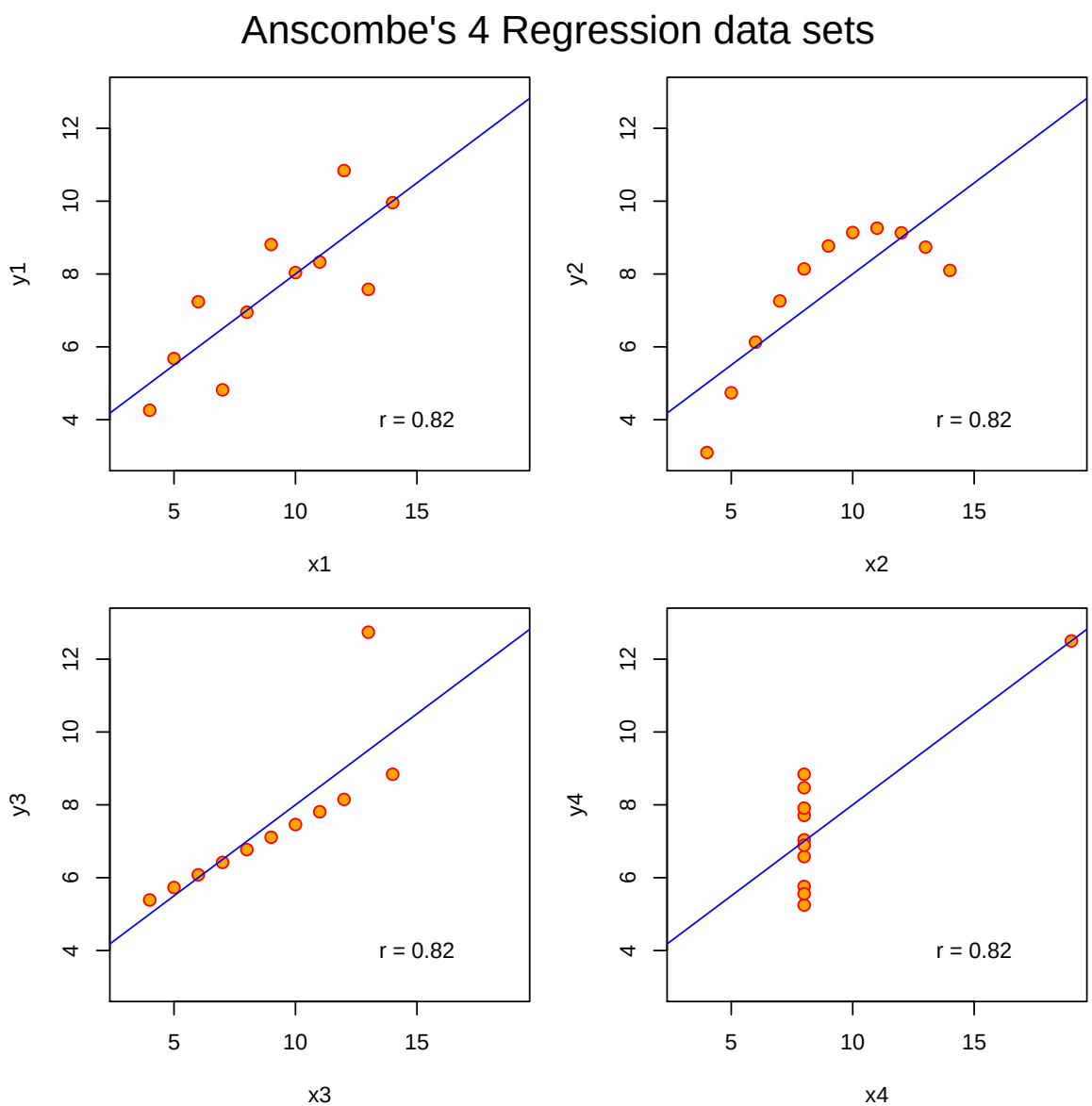


(a) (b)

Figure: (a) A scatterplot showing height versus weight from the 500 individuals in the sample from NHANES. One participant 163.9 cm tall (about 5 ft, 4 in) and weighing 144.6 kg (about 319 lb) is highlighted. (b) A scatterplot showing height versus BMI from the 500 individuals in the sample from NHANES. The same individual highlighted in (a) is marked here, with BMI 53.83. Fitted regression lines are shown in red.

Anscombe's quartet[1]

```
library(datasets)
data("anscombe")
```



Two numerical variables

Two categorical variables

Case study: molecular cancer classification

Two categorical variables

A contingency table summarizes data for two categorical variables.

```
addmargins(table(famuss$race, famuss$actn3.r577x))
```

```
##
##                CC  CT  TT Sum
##    African Am  16   6   5  27
##    Asian       21  18  16  55
##    Caucasian  125 216 126 467
##    Hispanic     4  10   9  23
##    Other        7  11   5  23
##    Sum        173 261 161 595
```

Two categorical variables ...

```
#row proportions
addmargins(prop.table(table(famuss$race, famuss$actn3.r577x), 1))
```

```
##
##                     CC        CT        TT       Sum
##   African Am 0.5925926 0.2222222 0.1851852 1.0000000
##   Asian      0.3818182 0.3272727 0.2909091 1.0000000
##   Caucasian  0.2676660 0.4625268 0.2698073 1.0000000
##   Hispanic   0.1739130 0.4347826 0.3913043 1.0000000
##   Other      0.3043478 0.4782609 0.2173913 1.0000000
##   Sum        1.7203376 1.9250652 1.3545972 5.0000000
```

```
#column proportions
addmargins(prop.table(table(famuss$race, famuss$actn3.r577x), 2))
```

```
##
##                     CC         CT         TT        Sum
##   African Am 0.09248555 0.02298851 0.03105590 0.14652996
##   Asian      0.12138728 0.06896552 0.09937888 0.28973168
##   Caucasian  0.72254335 0.82758621 0.78260870 2.33273826
##   Hispanic   0.02312139 0.03831418 0.05590062 0.11733618
##   Other      0.04046243 0.04214559 0.03105590 0.11366392
##   Sum        1.00000000 1.00000000 1.00000000 3.00000000
```

Two categorical variables ...

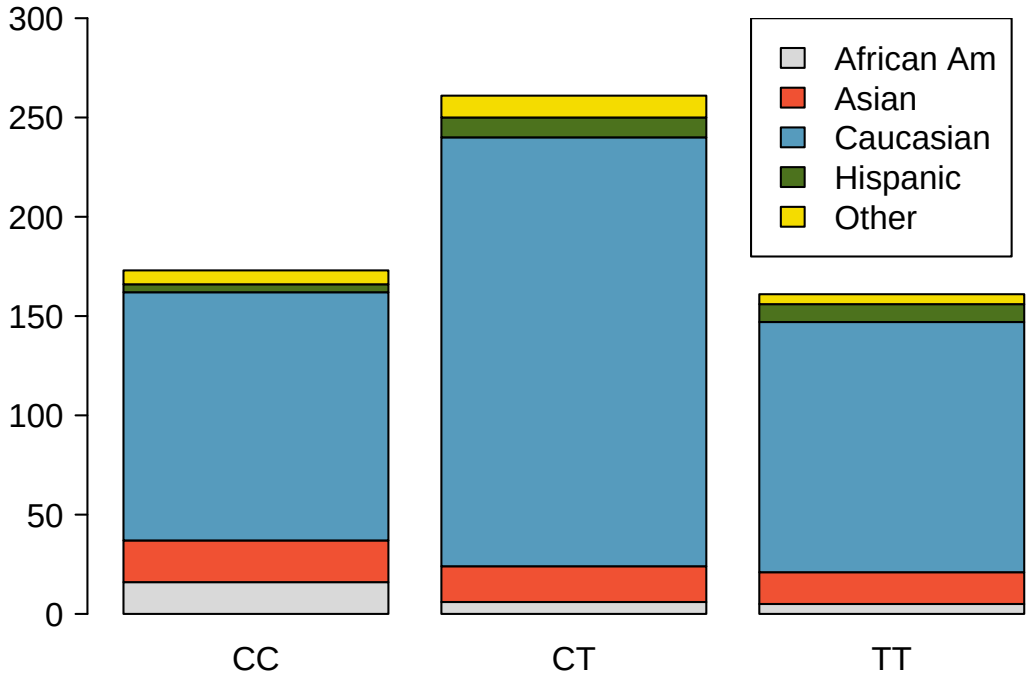


Figure: alt text

OI Biostat Figure 1.35a, segmented bar plot

Two categorical variables ...

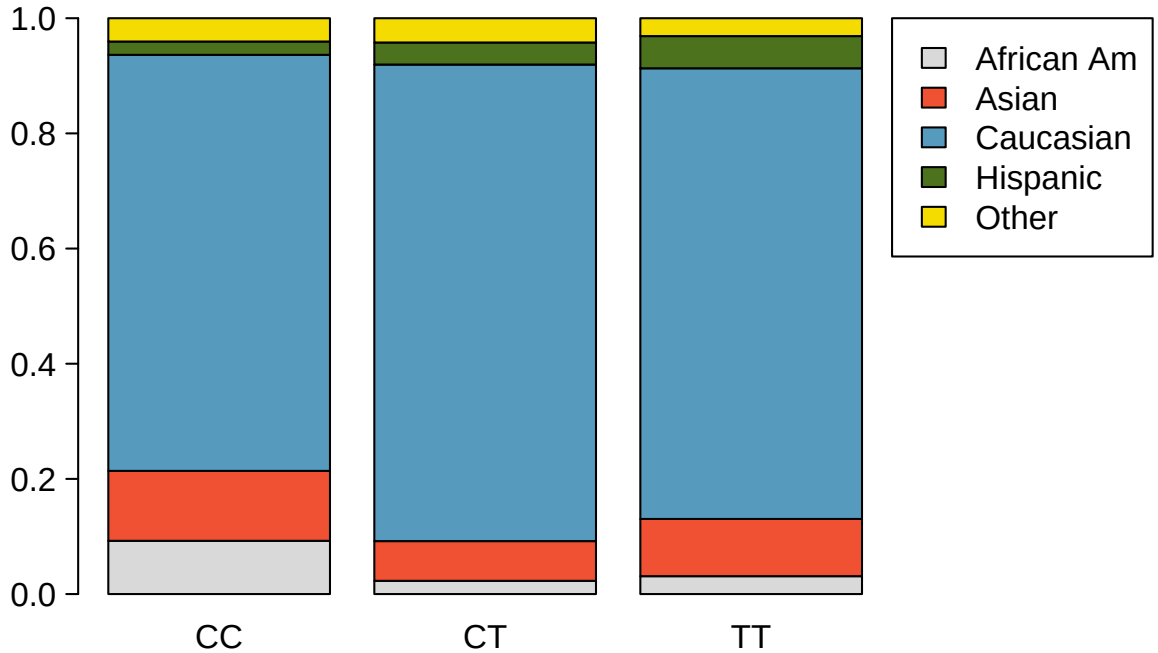
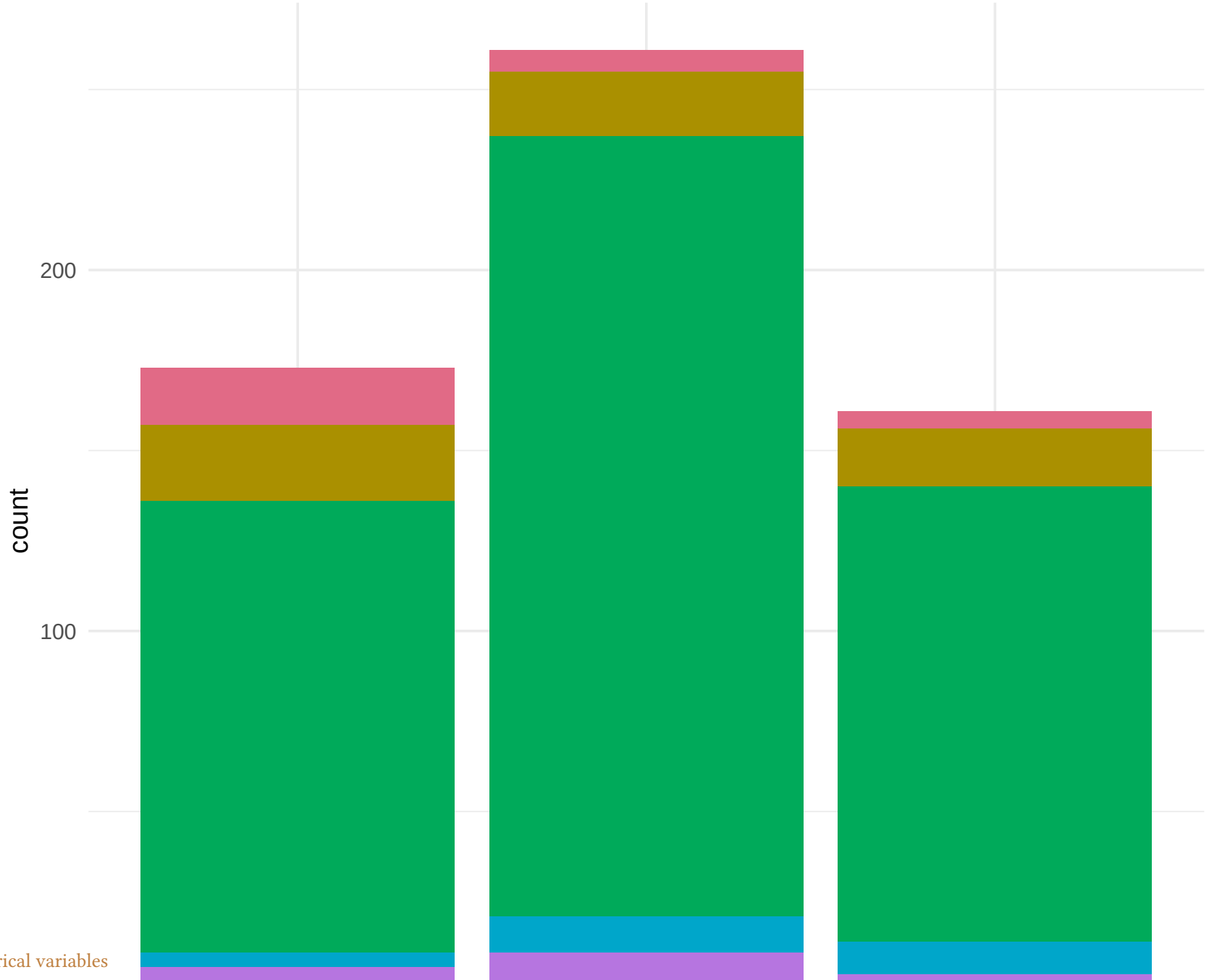


Figure: standardized segmented barplots

OI Biostat Figure 1.35b, standardized segmented bar plot

Two categorical variables



Two categorical variables ...

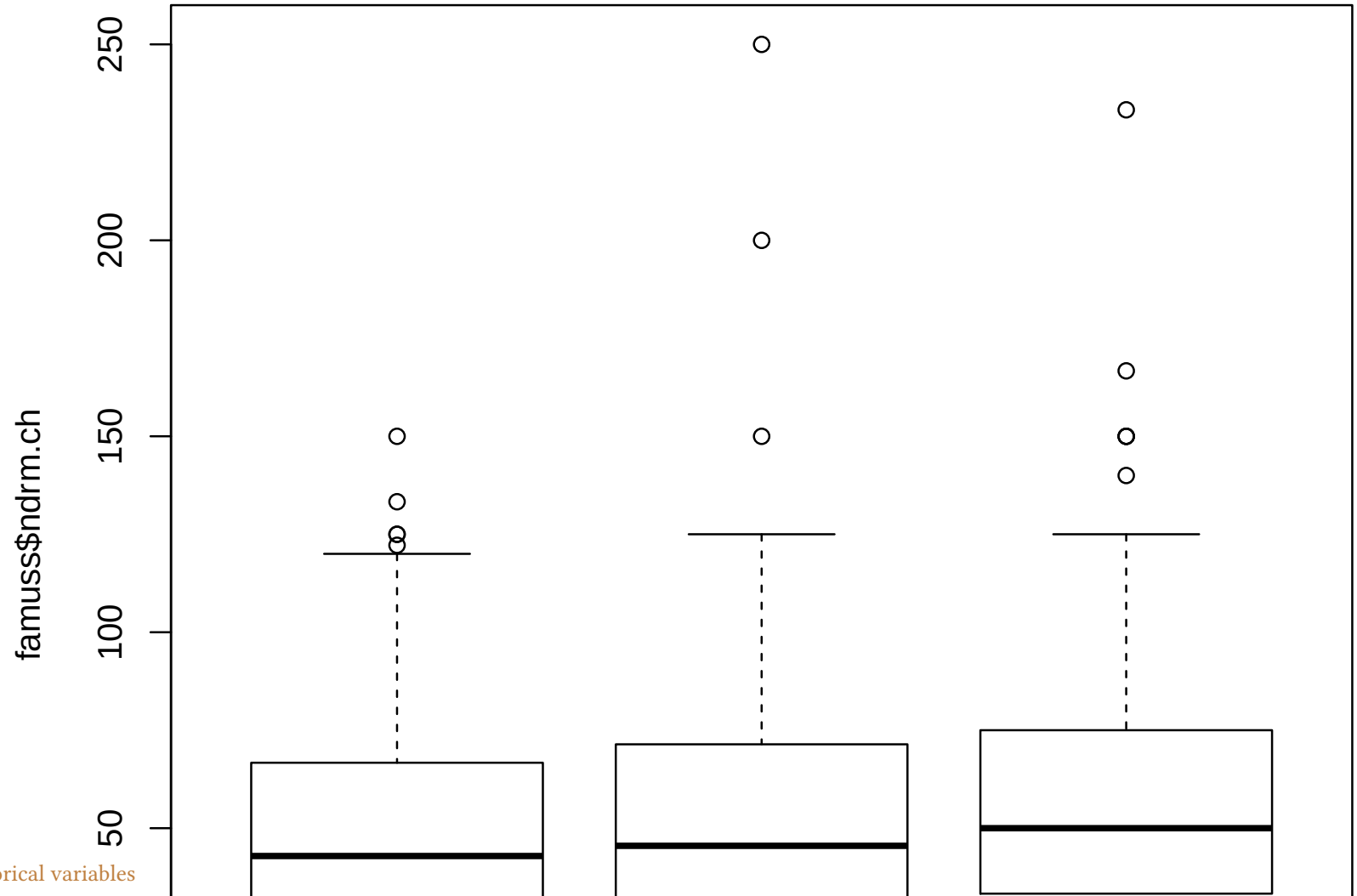
Relative risk (RR) is one way of summarizing data presented in a two-way table of study outcome by participant group.
More in Lab 1 ...

A numerical variable and a categorical variable

FAMuSS was designed to study the relationship between genotype at the location *r577x* in the gene *ACTN3* and muscle strength.
Muscle strength was assessed by the percent change in non-dominant arm strength after resistance training (`ndrm.ch`).
What visualization would be a good choice to make this comparison?

A numerical variable and a categorical variable ...

```
boxplot(famuss$ndrm.ch ~ famuss$actn3.r577x)
```



Two numerical variables

Two categorical variables

Case study: molecular cancer classification

The potential value of genomic data in cancer

The majority of cancers are diagnosed by an expert pathologist examining slides of malignant cells.
Can that be done more accurately by characterizing the genetic makeup of the malignancy?

- This is perhaps the major potential of genomic characterizations of tumors.

There are many forms of childhood leukemia.

- Acute myeloblastic leukemia (AML) and acute lymphoblastic leukemia (ALL) are the most common.
- AML is a cancer of the bone marrow, where white blood cells (lymphocytes) are produced.
- ALL is a cancer of the lymphocytes and is designated as B-cell (ALLB) or T-cell (ALLT).

Prognosis of the two cancers

The probability that a child diagnosed with ALL is survives at least 5 years after the diagnosis is approximately 90%.
Approximately 65% of children diagnosed with AML survive at least 5 years.
The diagnosis of leukemia type determines the therapy that will be given to the child, and the successful treatments for ALL and AML are different.
In 1999, Todd Golub from the Dana-Farber and the Broad Institute examined the possibility of classifying leukemia through using a genetic analysis of a blood sample.

Analyzing the Golub data

We can re-analyze the Golub data using tools from graphical and numerical summaries.
Our analysis will not be identical to the Golub analysis, but will be similar in spirit.
The tools are straighforward...

- Thinking through the problem and assembling the tools is the hard part.
- The process is more important than the final recipe.

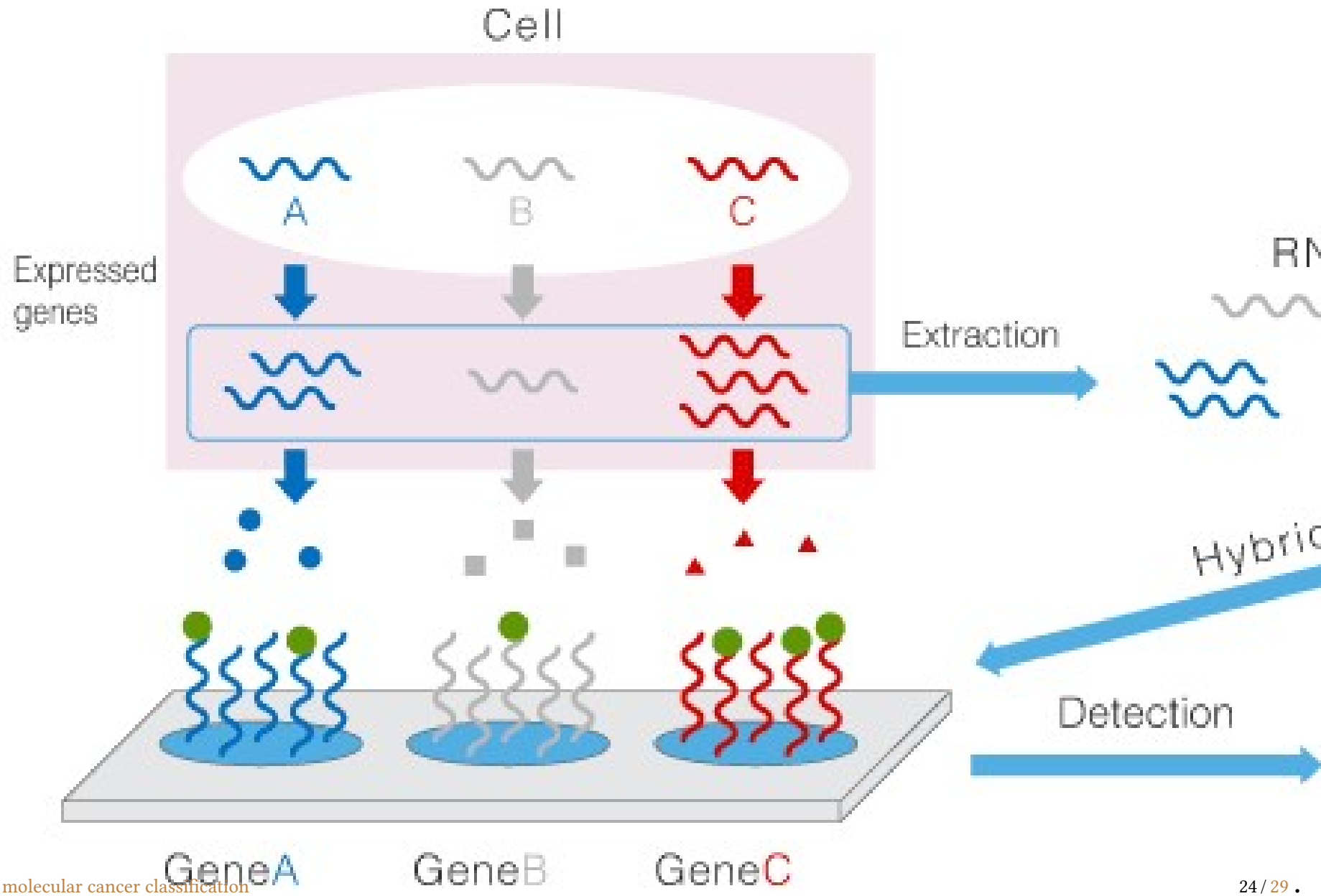
Gene expression (details in *OI Biostat*)

- The genetic code stored in DNA contains the information for producing the proteins that determine an organism's phenotype.
- Genes that are transcriptionally active (i.e. turned "on") are transcribed into messenger RNA (mRNA) that gets translated into proteins.
- Genes can be switched on or off, and expressed at varying levels. Variations in gene expression produce the range of physical, biochemical, and developmental differences in cells and tissues.
- Quantifying the amount of RNA produced in a cell allows for a measure of gene expression.
- The transcriptome, or expression profile, is the complete set of RNA transcripts produced by the genome in a cell or set of cells.

Microarrays (details in *OI Biostat*)

- Microarray technology is based on hybridization between two DNA strands, in which complementary nucleotide sequences specifically pair together.
- The mRNA from a sample is converted into complementary-DNA (cDNA), labeled with a fluorescent dye, and added to the microarray.
- When cDNA from the sample encounters complementary DNA probes, the two strands will hybridize, allowing the cDNA to adhere to specific spots on the slide.
- When the chip is illuminated and scanned, the intensity of fluorescence detected at each spot corresponds to the amount of bound cDNA.
- DNA microarrays do not directly quantify gene expression levels or quantity of mRNA present in a sample.
- The fluorescence intensity data only provide a relative measure of gene expression, showing which genes on the chip seem to be more or less active in relation to each other.

Microarrays



The Golub clinical data

Demographic variables described in *OI Biostat* Table 1.54:

Variable	Description
Samples	Sample or chip number. The material from each patient was examined on a separate chip and experimental run.
BM.PB	Type of patient material. BM denotes bone marrow; PB denotes a peripheral blood sample.
Gender	F for female, M for male.
Source	Hospital where the patient was treated.
tissue.mf	A variable showing the combination of type of patient material and sex of the patient. BM:f denotes bone marrow from a female patient, etc.
cancer	The type of leukemia; aml is acute myeloblastic leukemia, allB is acute lymphoblastic leukemia which started in B-cells (cells that mature into plasma cells) origin, and allT is acute lymphoblastic leukemia with T-cell origin (T-cells are a type of white blood cell).

The Golub expression data

The expression data is contained in the last 7,129 columns.
Each column is a variable with a name corresponding to the name of the probe on the microarray.
The expression levels record fluorescence intensity for each gene.

- The intensity levels have no inherent biological meaning.
- Data have been normalized to adjust for variability between the separate arrays used for each patient.

Selected variables and columns from Golub data

OI Biostat Table 1.40

Samples	Gender	cancer	AFFX-BioB-5_at	AFFX-BioB-M_at	AFFX-BioB-3_at
39	F	allB	-1363.28	-1058.59	-541.47
40	F	allB	-796.29	-1167.10	7.54
42	F	allB	-679.14	-1069.83	-690.30
47	M	allB	-1164.40	-1109.94	-990.13
48	F	allB	-1299.65	-1402.00	-1077.54

Analyzing the Golub leukemia data

We will do an analysis in class using some of the simple but surprisingly powerful ideas behind numerical and graphical summaries.
The goal of the Golub study was to develop a procedure for distinguishing between AML and ALL based only on the gene expression levels of a patient. There are two major issues to be addressed:

1. Which genes are the most informative for making a prediction?
2. What is a workable strategy for predicting leukemia type from expression data for a specific set of genes?

Starting small...

```
^^I^^I^^I^^I^^I^^I^^I##    cancer        A         B        C        D
^^I^^I^^I^^I^^I^^I^^I## 69   allB 39307.96 35232.401 41170.76 35792.79
^^I^^I^^I^^I^^I^^I^^I## 67   allT 32281.88 41432.024 59328.51 49608.14
^^I^^I^^I^^I^^I^^I^^I## 55   allB 47429.94 35568.928 56074.96 42857.78
^^I^^I^^I^^I^^I^^I^^I## 56   allB 25533.87 16983.749 28056.75 32693.92
^^I^^I^^I^^I^^I^^I^^I## 59   allB 35960.55 24191.746 27637.90 22240.75
^^I^^I^^I^^I^^I^^I^^I## 52    aml 46177.95  6189.465 12557.24 34485.41
^^I^^I^^I^^I^^I^^I^^I## 53    aml 43790.70 33661.825 38380.30 29758.25
^^I^^I^^I^^I^^I^^I^^I## 51    aml 53420.05 26109.245 31427.20 23809.70
^^I^^I^^I^^I^^I^^I^^I## 50    aml 41241.59 37589.773 47325.77 30099.36
^^I^^I^^I^^I^^I^^I^^I## 54    aml 41300.57 49198.412 66026.10 56248.62
^^I^^I^^I^^I^^I^^I^^I
```