Hausaufgabe 7

7.1

	A	B	C
G_1	22	2	3
G_2	20	1	1
G_3	18	0.5	0
G_4	19	1	5
G_5	19	2.5	1
G_6	23	1.5	7
G_7	17.5	0.5	0

1) Spalten sortieren

A_S	B_S	C_S
$17,5^7$	$0,5^3$	0^3
18^3	$0,5^2$	0^2
19^4	1^2	1^2
19^5	1^4	1^5
20^2	$1,5^6$	3^1
22^1	2^1	5^4
23^6	$2,5^5$	7^6

2) Zeilen mitteln

	M	
1	6	G_1
2	$\frac{37}{6}$	G_2
3	7	G_3
4	7	G_4
5	$\frac{49}{6}$	G_5
6	$\frac{29}{3}$	G_6
7	$\frac{65}{6}$	G_7

3) Ränge rücksubstituieren

A	B	C
$\frac{29}{3}$	$\frac{29}{3}$	$\frac{49}{6}$
$\frac{49}{6}$	7	7
$\frac{37}{6}$	6	6
7	7	$\frac{29}{3}$
7	$\frac{65}{6}$	7
$\frac{65}{6}$	$\frac{49}{6}$	$\frac{65}{6}$
6	$\frac{37}{6}$	$\frac{37}{6}$

7.2 Student t-test

A	42	39	38	60	41		
B	38	42	56	64	68	69	62

Die Nullhypothese lautet: $H_0: \mu_A = \mu_B$ mit $\alpha = 0,05$

Die Testgröße ist: $t = \frac{\bar{x}_B - \bar{x}_A}{s \cdot \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$ unter der Annahme $s_A = s_B$ und A und B normalverteilt.

Zuerst: $\bar{x}_A = \frac{1}{5}(42 + 39 + 38 + 60 + 41) = 44$

$\bar{x}_B = \frac{1}{7}(38 + 42 + 56 + 64 + 68 + 69 + 62) = 57$

Nun: $s^2 = \frac{1}{10}\left(\sum_{i=1}^{5}(x_{A,i} - \bar{x}_A)^2 + \sum_{j=1}^{7}(x_{B,j} - \bar{x}_B)^2\right) = \frac{1}{10}(330 + 926) = 125,6$

Damit $s = \sqrt{s^2} \approx 11,207$

Alle Werte eingesetzt: $t = \frac{57 - 44}{11,207\sqrt{\frac{1}{5} + \frac{1}{7}}} \approx 1,981$

Es wurden mit den beiden Mittelwerten 2 Schätzungen vorgenommen $\rightarrow (n_1 + n_2) - 2 = 10$ Freiheitsgrade

Für $f = 10$ ist für den zweiseitigen t-Test mit $\alpha = 0,05$ $t_{10;0,05} = 2,228$

Somit ist $t < t_{10;0,05}$ und die Nullhypothese wird angenommen, die Daten unterscheiden sich im gewählten Signifikanzniveau nicht ausreichend voneinander, womit die Genexpression zwischen den beiden Patientengruppen bei dieser Stichprobe im Rahmen des Konfidenzintervalls als gleich anzunehmen ist.

7.3 Multiple Testing Korrektur

Gene:	HER2	RAS	BRCA1	BRCA2	p53	TFAM	COX1
p-Value:	10^{-8}	10^{-6}	10^{-5}	10^{-4}	$10^{-3,8}$	10^{-3}	0.04

$n = 100$

Nach der Bonferroni-Korrektur ist die Signifikanzschwelle folgendermaßen anzupassen: $\alpha' = \frac{\alpha}{n}$

Mit $\alpha = 0,05$ ist $\alpha' = \frac{0,05}{100} = 5 \cdot 10^{-4}$

Eine Multiple testing Korrektur ist notwendig, da bei vielen Stichproben die Wahrscheinlichkeit für Ausreißer insgesamt steigt. Diesem Umstand muss durch eine Korrektur Sorge getragen werden.

Damit die Genexpressionsdaten weiterhin signifikant sind, muss $p < \alpha'$ So sind noch die Gene

HER2, RAS, BRCA1, BRCA2, p53 signifikant

Alle diese Gene stehen im Zusammenhang mit Krebszellen, womit davon auszugehen ist, dass hier gesundes Gewebe mit Krebszellen verglichen wurden.

Aufgabe 7.4

1. Wieviele Gene sind in ihrer Expression erhöht?

Die Differentielle Genexpressionanalyse mittels DESeq2 liefert für eine Signifikanzschwelle von 0,01 folgendes Ergebnis:

```
out of 15952 with nonzero total read count
adjusted p-value < 0.01
LFC > 0 (up)       : 141, 0.88%
LFC < 0 (down)     : 62, 0.39%
outliers [1]       : 7, 0.044%
low counts [2]     : 2165, 14%
(mean count < 9)
```

Darin ist zu erkennen, dass 141 Gene stärker exprimiert sind, als durch die angenommene Verteilung erwartet und 62 Gene weniger exprimiert sind als zu erwarten.

2. Beschreiben Sie wie sich der Volcono Plot verändert wenn Sie die Signifikanzschwelle verändern.

Im Volcanoplot wird der Logarithmus zur Basis 2 der Verhältnisse der Genexpressionen von "infected" zu "control" gegen den negativen dekadischen Logarithmus des p-Values aufgetragen. Alle Punkte, deren p-Wert kleiner (in positiven oder negativen) ist als die Signifikanzschwelle, werden dabei eingefärbt - je nach Art der Abweichung rot (für Überexprimierung) oder blau.
Dabei stellt die Grenze, ab welcher die Punkte eingefärbt werden, den negativen dekadischen Logarithmus der Signifikanzstelle dar. Verändert man diese, so verschiebt sich diese Grenze: Wird die Schwelle vergrößert werden mehr Punkte eingefärbt, wird sie verkleinert sind dies weniger.
(siehe volcano_005.pdf und volcano_001.pdf, für Signifikanzschwellen von 0,05 und 0,01)

3. Welches sind die 3 Gene die in ihrer Expression am stärksten erhöht bzw. unterdrückt sind?

Betrachtet man die Ergebnisse in "res" und vergleicht die Fold Changes "Log2FoldChange", so sind die 3 am stärksten erhöhten Gene:

1. Sobic.008G162600 mit Log2FoldChange = 5.49543
2. Sobic.006G070564 mit Log2FoldChange = 5.26315
3. Sobic.010G082600 mit log2FoldChange = 4.73206

Und die 3 am stärksten unterdrückten Gene:

1. Sobic.004G101400 mit Log2FoldChange = -4.14179
2. Sobic.009G182400 mit Log2FoldChange = -3.75887
3. Sobic.001G012200 mit Log2FoldChange = -3.68125

Sortiert wurde mit dem Befehl: res[order(res$log2FoldChange, decreasing=T{oder F})[1:3],]

4. Ab welcher Signifikanschwelle gibt es keine signifikanten Gene mehr?

Sortiert man in ähnlicher Weise die Ergebnisse nach "padj", so ergibt sich ein minimaler p-value von 4.75254e-16 beim Gen Sobic.003G079200. Setzt man das Signifikanzlevel somit auf 4,7e-16, so ergibt sich keine signifikanten Gene.
Tatsächlich tritt dies im Programm bereits bei 1e-16 auf.

```
out of 15952 with nonzero total read count
adjusted p-value < 4.7e-16
LFC > 0 (up)       : 0, 0%
LFC < 0 (down)     : 0, 0%
outliers [1]       : 7, 0.044%
low counts [2]     : 0, 0%
(mean count < 1)
```